Modélisation d'offres d'emploi et de curriculums vitæ en vue d'un processus de présélection automatisée de candidats

Albeiro de Jesus Espinal Pulgarin

COLLEGES BRETAGNE LOIRE | SCIENCES POUR L'INGENIEUR ET LE NUMERIQUE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THÈSE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPÉRIEURE MINES-TÉLÉCOM ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE

ÉCOLE DOCTORALE N° 648
Sciences pour l'Ingénieur et le Numérique
Spécialité : Informatique

Par

Albeiro de Jesus ESPINAL PULGARIN

Modélisation d'offres d'emploi et de curriculums vitæ en vue d'un processus de présélection automatisée de candidats

Thèse présentée et soutenue à IMT Atlantique, Brest, le 11/12/2023
Unité de recherche : Lab-STICC – UMR CNRS 6285
Thèse N° : 2023IMTA0385

Rapporteurs :

Célia da COSTA PEREIRA	Maîtresse de Conférences HDR, Université Côte d'Azur
Juan-Manuel TORRES-MORENO	Maître de Conférences HDR, Université d'Avignon

Composition du Jury :

Président :	Basel SOLAIMAN	Professeur, IMT Atlantique Bretagne Pays de la Loire
Examinateurs :	Peggy CELLIER	Maîtresse de Conférences HDR, INSA Rennes
	Alexandre VOISIN	Maître de Conférences, Université de Lorraine
	Célia da COSTA PEREIRA	Maîtresse de Conférences HDR, Université Côte d'Azur
	Juan-Manuel TORRES-MORENO	Maître de Conférences HDR, Université d'Avignon
	Phillipe LENCA	Professeur, IMT Atlantique Bretagne Pays de la Loire
Dir. de thèse :	John PUENTES	Professeur, IMT Atlantique Bretagne Pays de la Loire
Co-dir. de thèse :	Yannis HARALAMBOUS	Professeur, IMT Atlantique Bretagne Pays de la Loire

Invité :
Dominique BEDART Directeur de Développement, DSI Global Services

Remerciements

Je tiens à exprimer ma profonde gratitude envers *La Vie*, qui m'a ouvert les portes du domaine fascinant de la recherche à travers la réalisation de cette thèse, et m'a permis de découvrir de nouvelles perspectives de recherche durant les moments les plus difficiles.

Un merci vibrant et sincère à la docteure Maria Luisa Piraquive de Moreno ainsi qu'aux organisations qu'elle représente avec tant de dévouement et de passion. Leur œuvre philanthropique m'a été d'un soutien moral inconditionnel tout au long de la réalisation de ma thèse, et ce, spécialement durant les moments d'incertitude causés par la pandémie.

Je souhaite étendre un remerciement profond à ma famille, fontaine incommensurable de motivation et d'affection à travers ce voyage intellectuel. Ma femme, partenaire indéfectible, m'a apporté un soutien et un amour infaillibles tant dans les vallées de difficultés que sur les sommets de chaque succès. Ma mère, qui avec son amour maternel inconditionnel a contribué à établir les fondations de ma vie et une grande partie des valeurs qui me caractérisent aujourd'hui. Et mes quatre chats (Silvestre, Kassandra, Rebecca et Totoro) qui, chaque nuit de veillée tardive, ont prétendu jouer le rôle d'assistants de recherche.

Je tiens aussi à remercier mes directeurs et encadrant de thèse (John, Yannis et Dominique) pour leur patience et leur compréhension pendant ce parcours. Un merci tout particulier également à tous les acteurs, internes et externes à mon école, qui ont enrichi mon travail par leurs retours constructifs, ainsi qu'à l'équipe chaleureuse de la bibliothèque CARÆ qui m'a accueilli à bras ouverts lors de plusieurs journées de recherche.

Mes remerciements s'adressent également à deux institutions de renom : DSI Group et IMT Atlantique. La première a déposé en moi une confiance essentielle à la concrétisation de mon travail de thèse. Quant à la seconde, grâce à son exigence incessante d'excellence, elle constitue une source d'inspiration qui stimule mon désir d'aller toujours plus loin.

Enfin, un grand merci à tous ceux et celles qui, à un moment ou un autre, ont apporté leur pierre à l'édifice de ma vie et de ma carrière. Chacune de vos contributions a été essentielle dans la construction de ce parcours académique et dans la concrétisation de cette thèse.

Contenu du Manuscrit

LISTE D'ACRONYMES

AA	Apprentissage Automatique
ACP	Analyse en Composantes Principales
ADF	Arbre de Décision Flou
AP	Apprentissage Profond
ASL	Analyse Sémantique Latente
BOW	*Bag of Words*
CAS	*Complex Adaptative Systems*
CCO	Correspondance Curriculum Vitæ - Offre d'emploi
CDI	Croyances-Désirs-Intentions
CNN	*Convolutional Neural Network*
CV	Curriculum Vitæ
EAT	Extraction Automatique de Termes
EQ	Évaluation de Qualité
FAHP	*Fuzzy Analytic Hierarchy Process*
FNN	*Feedforward Neural Network*
GRU	*Gated-recurrent Unit*
GML	Grands Modèles de Langage
IA	Intelligence Artificielle
TIC	Technologies de l'Information et de la Communication
LSTM	*Long-short Term Memory*
OE	Offre d'Emploi
PA	Projection Aléatoire
PDP	Processus de Développement du Produit
SG	Séquence Graphémique
RH	Ressources Humaines
RLC	Régression Logistique Classique
RLF	Régression Logistique Floue
RL AF	Régression Logistique avec Filtrage
RL SOF	Régression Logistique sans Filtrage
RNN	Recurrent Neural Network
TAL	Traitement Automatique de la Langue
Tfidf	*Term Frequency-inverse Document Frequency*
WR	*Weirdness Ratio*

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

Cette introduction présente une approche interdisciplinaire de l'automatisation de la présélection des candidats dans un processus de recrutement, appelée Correspondance Curriculum Vitæ - Offre d'Emploi (CCO). Nous intégrons des techniques de modélisation traditionnelles et avancées pour les offres d'emploi (OE) et les curriculums vitæ (CV). Ces techniques sont combinées avec les stratégies de prise de décision des recruteurs. Notre objectif est d'améliorer la robustesse et l'explicabilité de la CCO, en tenant compte à la fois des subtilités humaines et des incertitudes inhérentes au traitement automatisé. Les sections suivantes abordent le contexte de la thèse, les défis, les objectifs de recherche et les contributions originales.

Contexte de la thèse

Durant les années 1980, avec la montée de l'utilisation des ordinateurs dans les entreprises nord-américaines, la demande pour l'analyse et la comparaison des CV et des OE à l'aide de techniques de traitement automatique de la langue (TAL) a augmenté. Les recruteurs rencontraient des difficultés dans le processus de présélection, impliquant l'évaluation manuelle de divers critères des CV en relation avec les prérequis du poste. La nature chronophage de cette tâche entraînait des retards critiques pour les entreprises et les candidats [7].

En réponse à ces défis, le système Resumix a émergé en 1989 comme une solution révolutionnaire [7]. Il utilisait des algorithmes pour apparier et classer des milliers de CV par rapport à une OE spécifique en quelques secondes, allégeant ainsi la charge des recruteurs. Cette innovation a stimulé le développement d'un nouveau domaine de recherche, la CCO, dédié à l'affinement des méthodes pour la comparaison automatique des CV et des OE. La CCO a été depuis continuellement adaptée, notamment avec la révolution numérique remodelant son paysage.

Alors que ce domaine a considérablement progressé depuis l'avènement de Resumix, l'évolution des CV et des OE pose de nouveaux défis [8]. Les CV modernes adoptent des formats de plus en plus non structurés et graphiques [9], tandis que les OE deviennent plus succinctes, en particulier sur les plateformes de recrutement numériques [10].

Au-delà de ces enjeux liés aux formats des documents, la diversification du marché du travail en termes de professions et de compétences a rendu leur interprétation encore plus complexe [11]. Ceci a incité des recherches sur les écarts de compétences, qui englobent des divergences entre les qualifications individuelles et les demandes des employeurs. Ces disparités, plus que simplement les compétences énumérées sur les CV et les OE, reflètent un déséquilibre de compétences plus

large aggravé par des incertitudes telles que les fluctuations de l'offre et de la demande du marché [12].

Face à ces défis évolutifs et au vu du rôle pivot de la CCO dans les processus de recrutement, la présente thèse se propose de soulager la charge de travail des recruteurs dans un monde professionnel de plus en plus numérisé. Bien que cette thèse s'attache à répondre à des défis de recherche propres au domaine, il est essentiel de reconnaître aussi son contexte industriel.

En effet, le domaine de la CCO se distingue par un fort lien industriel, exigeant une approche de recherche agile ainsi que des solutions pragmatiques. Contrairement aux contextes académiques traditionnels, ce domaine nécessite une innovation dynamique pour s'aligner sur le paysage industriel en constante évolution. Il s'agit donc d'une niche spécialisée où des solutions pratiques et commercialisables sont recherchées pour relever les défis contemporains du recrutement.

Cette combinaison d'industrie et d'académie nécessite une considération flexible des problèmes. La priorité est donnée à une solution qui répond aux normes de qualité minimales plutôt qu'à la poursuite à long terme d'une méthode irréprochable. Dans cette optique, la recherche flexible de cette thèse apporte des contributions tant théoriques que pratiques. Elle s'aligne avec les besoins et les attentes industrielles tout en naviguant à travers les complexités du traitement automatique moderne des OE et des CV.

Ayant établi les origines et l'importance de la CCO pour nos travaux de recherche, il est temps de se pencher sur les complexités et les difficultés qui ont émergé dans le paysage moderne du recrutement.

Contexte du problème

Ces dernières années, le domaine de la CCO n'a pas été immunisé contre les révolutions industrielles entraînées par les avancées en apprentissage profond (AP). En conséquence, les méthodes de CCO dépendent désormais en grande partie des techniques de correspondance d'AP [13]. Bien que les méthodes de CCO basées sur l'AP offrent des avantages en matière de scalabilité et d'automatisation [13], leur opacité—souvent qualifiée de modèles "boîte noire"—a introduit des défis globaux [14].

L'émergence de la 'boîte noire' : AP dans la CCO

Pour mettre en évidence les implications des systèmes automatisés de recrutement 'boîte noire', considérons le phénomène connu sous le nom de 'Grande Démission' (*The Great Resignation*), qui a été marqué par une pénurie significative de travailleurs entre 2021 et 2022. Principalement observée aux États-Unis, cette hausse significative des démissions d'employés pendant l'ère post-pandémique pourrait avoir des répercussions mondiales. Noelle Chesley, professeure

associée de sociologie à Université de Wisconsin-Milwaukee, a souligné que la dépendance croissante à l'égard des méthodes de recrutement opaques contemporaines pourrait être un facteur contribuant à cette tendance [1], une notion appuyée par des études expérimentales récentes.

Par exemple, les biais intégrés dans ces nouvelles méthodes peuvent conduire à des distributions d'OEs biaisées, certains groupes démographiques, comme les femmes ou les candidats noirs, étant sous-représentés dans certains secteurs d'activité économique [15, 16]. De plus, les employeurs, malgré l'utilisation de systèmes d'appariement à grande vitesse, peuvent avoir du mal à trouver des candidats adaptés aux besoins spécifiques et uniques de l'entreprise [17]. En outre, certains candidats pourraient passer des mois sur des plateformes en ligne à la recherche d'emplois, pour se retrouver perplexes quant à la raison pour laquelle ils ne peuvent pas obtenir un poste, apparemment mis de côté par des algorithmes de CCO opaques [18]. Étant donné qu'un pourcentage stupéfiant de 98 % des entreprises du Fortune 500 emploient de tels systèmes de recrutement, les implications de ces biais peuvent être vastes et avoir une grande portée [19].

Au-delà des complexités des modèles d'AP dans la CCO, il est crucial de reconnaître que ceux-ci ne sont qu'une couche dans un réseau à plusieurs niveaux de défis auxquels est confronté le domaine. La question délicate des "boîtes noires" dans la CCO est située au sein d'un écosystème de recrutement plus large, qui est lui-même rempli de complexités et d'incertitudes. Pour acquérir une compréhension complète des problèmes en jeu, nous devons examiner ces couches supplémentaires de complexité qui s'étendent des obstacles technologiques aux subtilités linguistiques, et explorer comment elles aggravent les problèmes introduits par les modèles d'AP.

Couches de complexité dans les systèmes CCO actuels

Même sans les défis introduits par les modèles d'AP, le domaine de la CCO fait face à ses propres complexités inhérentes. Celles-ci vont de la nature dynamique du marché du travail [20], à la myriade de compétences professionnelles [11], aux rôles d'emploi en évolution [21], à la diversité des profils de candidats [22], jusqu'aux subtilités linguistiques des CV et des OE [23].

Le processus de recrutement, en particulier la phase de présélection ou de *screening*, sert de canal de communication où les candidats et les recruteurs échangent des informations par le biais de leurs CV et OE respectifs [24]. Au cours de cette phase, l'interaction est généralement asynchrone, la compréhension de chaque partie étant limitée au contenu du document, façonnée par des facteurs contextuels tels que les normes sociétales, la variabilité linguistique et les expériences de recrutement antérieures [25]. Ces facteurs introduisent une forme unique d'incertitude, définie comme un manque d'informations nécessaires pour une communication efficace pendant la phase de présélection [26], soulevant des problèmes d'interprétation dans les systèmes CCO automatisés et non automatisés [27].

1. https ://uwm.edu/news/automated-hiring-systems-could-be-making-the-worker-shortage-worse/

Plusieurs facteurs compliquent davantage cette phase. L'influence du contexte organisationnel du recruteur [28], l'alignement des documents avec le contexte de l'autre partie [29] ; l'extraction automatique précise et transparente des exigences essentielles des OE [30], et les éléments graphiques et non structurés dans les CV modernes [31]. En essence, malgré les avancées depuis les années 1980, les systèmes CCO restent mis au défi par les complexités qu'ils engendrent.

Ces aspects souvent négligés ont laissé un vide dans le champ d'étude. Les modèles actuels sont souvent limités sur le plan de qualité de correspondance, d'explicabilité et d'alignement avec les besoins organisationnels spécifiques.

Par conséquent, les défis sont multiples :

— assurer l'explicabilité algorithmique pour maintenir des pratiques de recrutement plus éthiques, justes et transparentes ;
— décrypter les compétences professionnelles et les rôles d'emploi dans les CV et les OE, surtout lorsque les référentiels internationaux ne correspondent pas toujours aux besoins organisationnels spécifiques ;
— extraire des informations essentielles des OE pour classer les CV sans trop dépendre de modèles non transparents et inexplicables ;
— adapter la représentation du CV et de l'OE pour répondre au contexte et aux besoins uniques des recruteurs et de chaque organisation ;
— assurer l'adaptabilité algorithmique aux changements rapides dans le paysage du recrutement, nécessitant des modèles capables d'évoluer avec les exigences constamment changeantes de ces documents ;
— naviguer dans les mises en page de plus en plus non structurées et graphiques des CV contemporains.

Compte tenu de ces multiples couches de complexité, allant des préoccupations éthiques aux défis technologiques et linguistiques, le domaine de la CCO est plus susceptible d'un examen intégral. Il existe un besoin notable de méthodologies plus holistiques pouvant accueillir les défis contemporains du processus de recrutement. Cette recherche vise à combler ces limitations, servant de pivot vers des méthodologies de CCO plus transparentes, explicables et efficaces.

Objectifs et questions de recherche

Cette recherche s'implique directement avec les défis multifacettes décrits dans la section précédente, en se concentrant particulièrement sur le problème de la sous-représentation du CV et de l'OE au sein de la phase de présélection du recrutement. L'objectif principal est de développer une méthodologie pour la représentation détaillée de ces documents dans la CCO. La recherche vise à répondre à la question suivante : quels sont les composants essentiels nécessaires à un modèle robuste pour extraire, structurer et annoter sémantiquement les informations des

OE et des CV lors d'un processus de présélection ?

Nous avons structuré cette question de recherche en quatre sous-questions, qui sont les suivantes :

— quelles considérations sont cruciales pour construire un cadre capable d'identifier les marqueurs textuels dans les OE afin d'optimiser le classement automatisé des CV des candidats ? Comment les perspectives des recruteurs et le contexte organisationnel peuvent-ils contribuer à son efficacité ?

— comment évaluer quantitativement l'incertitude associée aux termes pertinents de l'OE identifiés automatiquement, en mettant l'accent sur leur pertinence, en appliquant des modèles d'apprentissage possibilistes linéaires et non linéaires ?

— comment développer un cadre d'évaluation de la qualité qui optimise les marqueurs textuels pour une extraction précise des informations à partir des OE, en tenant compte de l'ambiguïté, de l'expertise du recruteur, de la pertinence de l'information par rapport au contexte organisationnel et d'une approche explicative ?

— quels sont les composants essentiels d'une méthodologie complète pour une représentation sensible au format contemporain des CV, et comment cette méthodologie peut-elle être conçue pour s'aligner avec le contenu grapholinguistique de ce type de document et les perspectives des recruteurs ?

Malgré quelques propositions récentes pour la modélisation algorithmique des CV et des OE [32, 33, 34, 35, 36, 37, 38], aucune méthodologie ne prend en compte de manière intégrale les aspects critiques tels que le format non structuré et graphique des documents, les stratégies de sélection des informations par les recruteurs, la représentation de l'ambiguïté dans la prise de décision, la représentation du contexte organisationnel et les incertitudes inhérentes à la phase de présélection.

Nos travaux de recherche sont basés sur l'hypothèse qu'une approche robuste pour la modélisation des CV et des OE doit analyser et intégrer les connaissances des recruteurs et les spécificités de leur contexte organisationnel, étant donné leur compréhension approfondie de la comparaison de ces documents. Nous supposons que la prise en compte du contexte organisationnel est fondamentale pour assurer une adaptation et une évolution appropriées d'une CCO optimale et contextualisée.

Approche et contributions de la thèse

Les contributions de cette thèse sont associées aux aspects méthodologiques de plusieurs domaines de recherche examinés, comme l'illustre la Figure 1. La nouveauté de notre recherche réside dans l'analyse et l'intégration des stratégies des recruteurs associées à la pertinence de l'information et à leur contexte organisationnel. De plus, cette méthodologie comprend l'optimi-

sation des marqueurs textuels liés à l'information pertinente, l'introduction de la modélisation de l'incertitude dans le processus de la CCO, et la prise en compte et la représentation du contenu grapholinguistique des CV—offrant des nouvelles perspectives à la discipline.

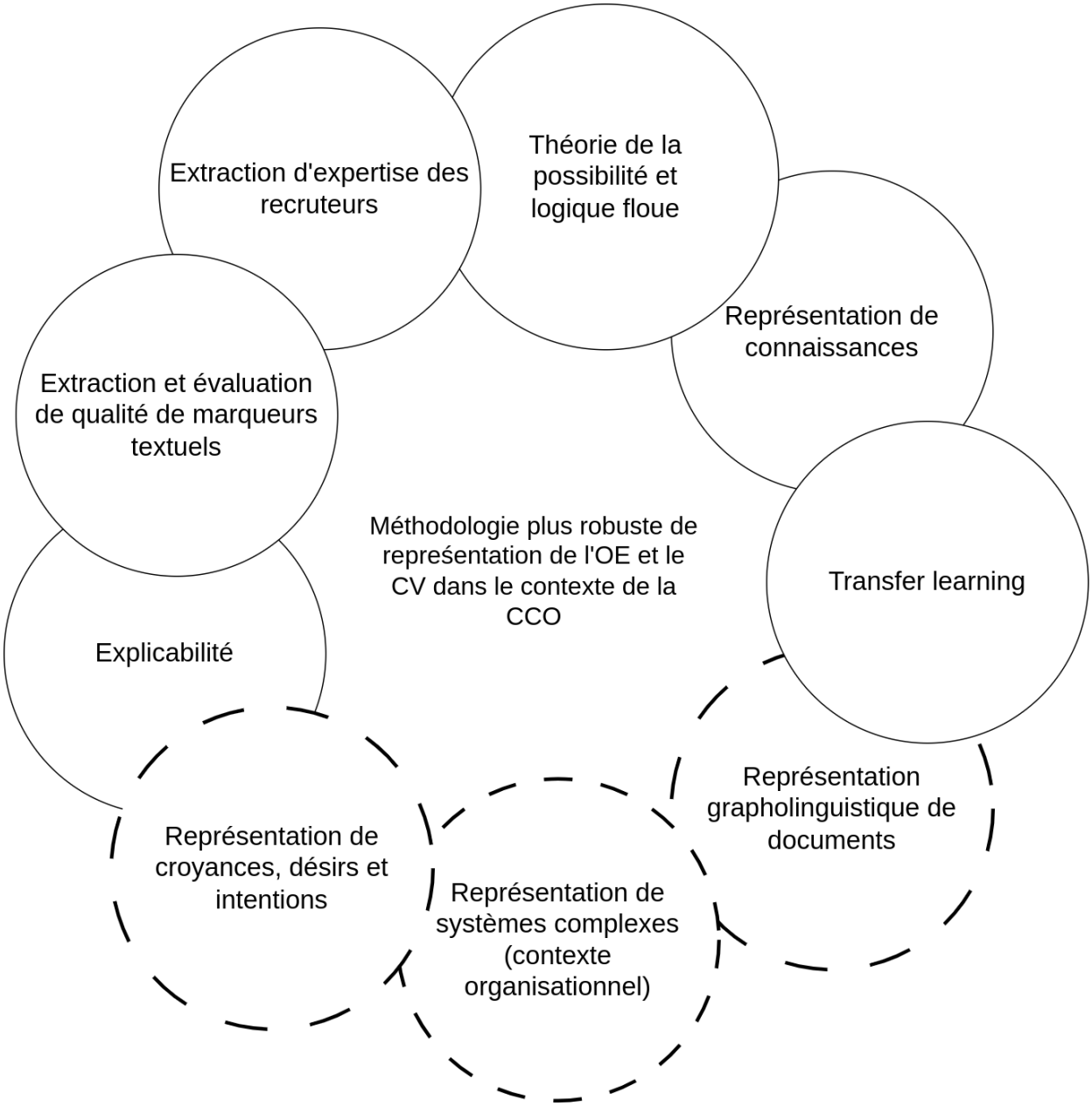


FIGURE 1 – Domaines de recherche et méthodologies étudiés dans la thèse, représentés par des cercles avec des lignes continues et discontinues respectivement.

Les contributions spécifiques de la thèse sont les suivantes.

— **Méthodologie pour étudier et intégrer l'expertise des recruteurs pour une**

meilleure extraction de l'information : la première contribution inclut le développement d'une méthodologie qui analyse les stratégies des recruteurs dans la sélection de l'information pertinente dans les OE. En identifiant les marqueurs textuels qui s'alignent avec les stratégies de sélection de l'information de ces acteurs, nous enrichissons l'extraction automatique des termes pertinents des OE en utilisant une architecture de croyances-désirs-intentions (CDI) possibiliste basée sur une ontologie. Ce processus vise à améliorer le classement des CV tout en favorisant la transparence et l'explicabilité dans le processus de sélection [39, 40].

— **Méthodologie pour l'évaluation de la qualité des marqueurs textuels orientés vers l'incertitude :** en tant que deuxième contribution, cette thèse propose une approche de modélisation de l'incertitude qui soutient l'évaluation et l'optimisation des marqueurs textuels dans les OE. Ce modèle est intégré dans l'architecture CDI, ce qui permet une adaptation dynamique aux besoins de recrutement changeants et aux contextes organisationnels [41].

— **Méthodologie pour intégrer la dimension grapholinguistique des documents :** notre recherche a mis en évidence d'importants marqueurs textuels et graphiques dans les CV, contribuant à l'optimisation des tâches d'analyse automatisée des CV comme la segmentation. En utilisant une architecture BERT (Bidirectional Encoder Representations from Transformers) minimaliste, cette contribution souligne l'importance de la dimension graphique dans les documents pour améliorer l'analyse automatisée des CV et les méthodes de CCO [42].

Positionnement spécifique des contributions par rapport aux approches traditionnelles et récentes pour la CCO

La méthodologie que nous proposons se distingue des modèles traditionnels centrés sur des paradigmes de représentation tels que les ontologies ou les modèles de sac de mots [32, 33, 34]. Elle se différencie également des approches contemporaines axées sur les modèles d'AP et les architectures transformateurs [35, 37, 36] :

— **limitations des modèles existants :** les approches traditionnelles étudiées et les modèles récents, bien qu'efficaces dans diverses expériences de la littérature sur la CCO, peuvent échouer à extraire de manière précise et transparente des informations pertinentes des OE ou à traiter les CV en alignement avec les besoins spécifiques du contexte organisationnel [24]. Pour surmonter cette limitation, d'énormes quantités de documents sont généralement nécessaires [37] ;

— **sensibilité au contexte :** les méthodes traditionnelles étudiées, dépendantes de règles prédéfinies ou de techniques de comptage de mots, manquent souvent de compréhension des significations nuancées dans les OE et les CV [43]. Les modèles basés sur des

transformateurs, bien que compétents dans la compréhension contextuelle, ont tendance à être opaques et complexes, entravant l'alignement avec les exigences organisationnelles spécifiques des recruteurs [44] ;

— **gestion de l'incertitude :** les modèles actuels manquent de fondements théoriques pour gérer l'incertitude, supposant une fiabilité uniforme des sources d'information. Cela ne tient pas compte des défis réels tels que les exigences changeantes [37], surtout dans des secteurs évolutifs comme la technologie, les biais dans la rédaction et le traitement des postes [43, 45], la variabilité dans la description des emplois par différents employeurs [46], et la complexité de l'évaluation de l'adéquation culturelle d'un candidat [47] ;

— **adaptabilité :** les méthodes traditionnelles étudiées peuvent être rigides par rapport aux exigences en évolution rapide des processus de recrutement [24], tandis que les modèles de transformateurs exigent d'importantes données pour l'affinement, posant des défis d'anonymisation des documents et d'évolution à long terme [48] ;

— **explicabilité :** la complexité des modèles de transformateurs entrave souvent la compréhension de la détermination de la pertinence de l'information dans les OE et les CV, soulevant des questions liées au recrutement contraire à l'éthique, à la discrimination et aux biais [49] ;

— **personnalisation :** les modèles traditionnels et récents étudiés font face à des difficultés pour incorporer des connaissances et un raisonnement spécifiques à l'humain [50].

En contraste, la méthodologie proposée proposée aborde ces problématiques :

— **sensibilité au contexte :** en capturant l'incertitude et la sensibilité au contexte inhérentes au raisonnement sémantique humain, la méthodologie vise à produire des résultats potentiellement plus précis et fiables ;

— **gestion de l'incertitude :** l'architecture emploie une approche possibiliste et floue pour représenter plus efficacement, du point de vue théorique, les incertitudes et les variations dans la fiabilité de l'information ;

— **adaptabilité :** la méthodologie est adaptable, avec une architecture CDI possibiliste basée sur une ontologie pour le traitement des OE permettant des ajustements basés sur des exigences en évolution rapide ;

— **explicabilité :** chaque étape de raisonnement liée à l'extraction de l'information est transparente et interprétable, favorisant la compréhension de l'utilisateur ;

— **personnalisation et flexibilité :** la méthodologie vise à représenter le raisonnement sémantique humain à travers une approche possibiliste fondée sur des ontologies. Elle intègre des modèles "boîte noire" (tels que les transformateurs) en tant que marqueurs textuels augmentés, ou règles sémantiques, ce qui permet d'évaluer l'incertitude qui leur est associée.

Étant donné ces considérations, en tirant des enseignements des approches traditionnelles et

modernes de la CCO, nous proposons une approche plus adaptable qui met davantage l'accent sur la sensibilité au contexte et l'explicabilité. Reconnaissant l'incertitude inhérente et le besoin d'adaptation continue dans le paysage de la CCO, notre méthodologie vise à refléter les attentes changeantes des recruteurs et des organisations. À mesure que l'intelligence artificielle (IA) devient de plus en plus importante et que la demande pour des algorithmes transparents et responsables augmente, notre approche privilégie la clarté, visant à favoriser une plus grande confiance parmi les parties prenantes.

Contenu du manuscrit

Ce manuscrit est composé de quatre chapitres et nous illustrons le plan à la Figure 2.

Le premier chapitre offre un aperçu de l'état de l'art dans le domaine de recherche de la thèse, en se concentrant sur les CV, les OE et la CCO pour la présélection automatisée des candidats. Il introduit également des connaissances pertinentes qui servent de fondations fondamentales pour cette thèse.

Le deuxième chapitre élucide notre méthodologie pour la modélisation des CV et des OE. Il détaille les composants de la méthodologie et leur interaction, illustrant l'approche proposée.

Le troisième chapitre examine quatre cas d'étude qui mettent en avant les applications pratiques de la méthodologie. Le premier cas d'étude aborde l'évaluation des modèles possibilistes pour la représentation de l'incertitude dans l'extraction des informations des OE, en vue d'améliorer le classement des CV. Le deuxième cas présente l'extraction CDI possibiliste d'informations essentielles à partir des OE, tandis que le troisième cas met l'accent sur l'amélioration des marqueurs textuels des OE liés à la pertinence de l'information. Le quatrième cas d'étude est centrée sur l'optimisation de la segmentation des CV. Chaque cas d'étude détaille les paramètres de l'expérimentation et les résultats tangibles obtenus.

Le quatrième chapitre s'ouvre sur une discussion sur les résultats issus de la méthodologie et des cas d'étude, et examine leur potentiel de généralisation dans la CCO. Il se conclut par un résumé des résultats de la thèse, des conclusions et des perspectives pour les orientations futures de la recherche et les développements potentiels dans le domaine.

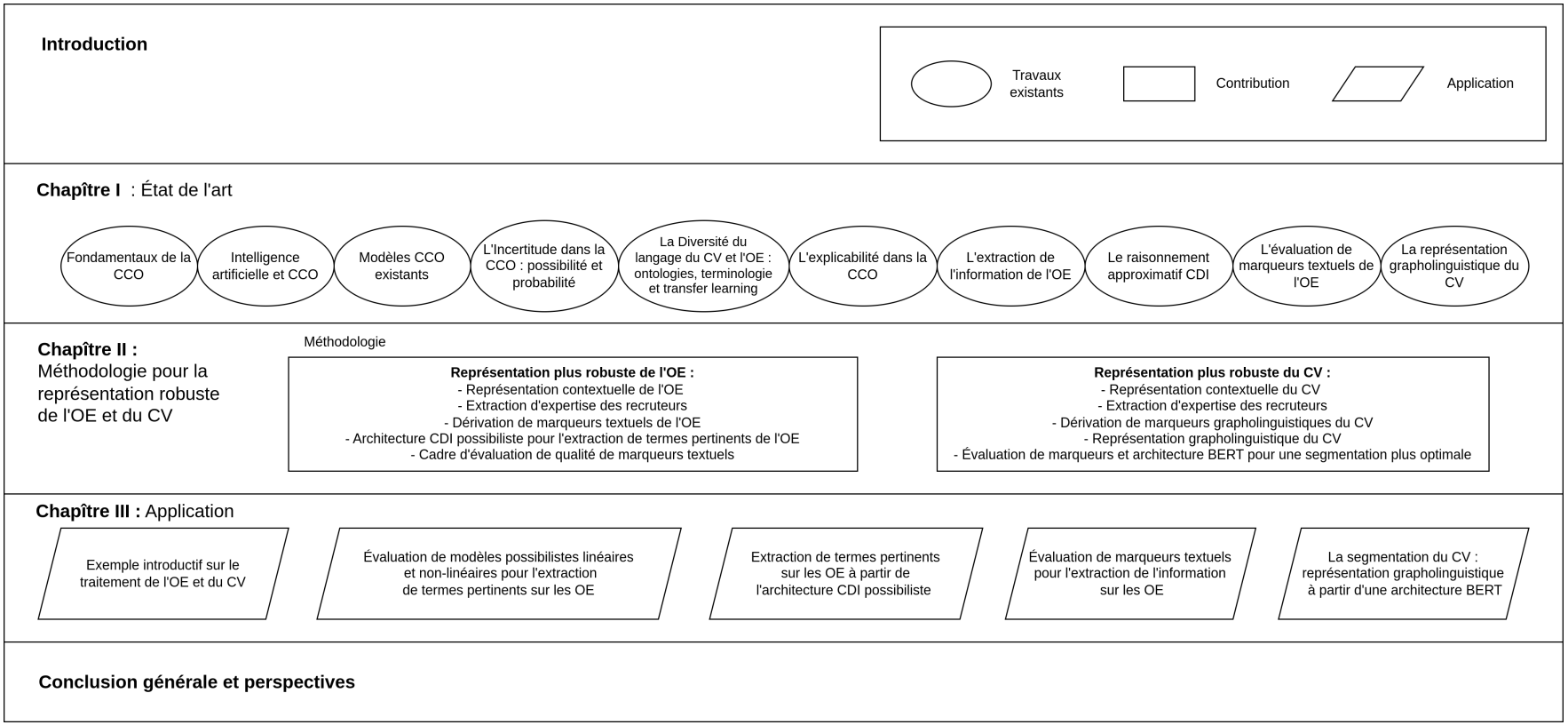


FIGURE 2 – Représentation du plan du manuscript.

Chapitre I

État de l'art

L'état de l'art présenté dans ce chapitre souligne la nature multidisciplinaire et complexe de la CCO. La CCO fait le pont entre plusieurs domaines, y compris l'apprentissage automatique (AA), le TAL, la linguistique, le comportement organisationnel, entre autres. Ce chapitre s'efforce d'offrir une exploration des méthodologies, des défis et des approches novatrices au sein de ce domaine multifacette.

La structure du chapitre est la suivante : il commence par une exploration des éléments fondamentaux de la CCO, en particulier les CV, les OE et les compétences professionnelles, fournissant une compréhension plus profonde de la phase essentielle de présélection dans le recrutement. Par la suite, l'accent est mis sur les rôles nuancés des recruteurs et le contexte organisationnel plus large au sein de la CCO. S'aventurant dans les cadres existants, le chapitre dévoile les tendances de recherche de l'état de l'art, notamment l'approche possibiliste des CDI, qui se profile comme une alternative pour redéfinir le traitement automatique des OE.

En progressant, le chapitre se lance dans une étude détaillée des méthodologies courantes, couvrant l'AA, l'AP et le TAL, avec une attention particulière portée au traitement de l'information textuelle. Il aborde les complexités inhérentes à l'incertitude dans l'automatisation de la phase de présélection, préconisant les théories de la possibilité et de la logique floue comme cadres puissants pour manœuvrer ces défis. Le chapitre éclaire également les techniques actuelles pour évaluer la qualité des marqueurs textuels, visant à améliorer l'extraction de l'information à partir des OE. Le rôle crucial du contenu graphique dans les CV est souligné, détaillant comment les stratégies modernes exploitent cette dimension pour affiner les tâches d'analyse automatisée des CV, en particulier la segmentation.

En conclusion, ce chapitre pose une base pour une plongée intégrale dans le vaste et complexe paysage de la CCO. Il confère une compréhension des méthodologies et techniques contemporaines. L'amalgame des perspectives académiques et industrielles accentue davantage la pertinence de la recherche et son influence prospective sur les défis de recrutement dans le monde réel.

I.1 Introduction et fondamentaux de la Correspondance Curriculum Vitæ - Offre d'Emploi (CCO)

Le paysage moderne du recrutement est complexe, les CV et les OE agissant comme les pierres angulaires de la phase de présélection. Cette section commence par une analyse approfondie des complexités de ces documents, mettant l'accent sur les défis posés par leur contenu souvent non structuré et incertain dans les approches actuelles de la CCO. Les sous-sections suivantes développent les rôles des compétences professionnelles, les avancées dans l'automatisation de la phase de présélection et l'influence prépondérante des recruteurs dans divers contextes organisationnels.

I.1.1 L'Offre d'Emploi (OE) en tant que type de document

Les OE, également connues sous le nom d'annonce d'emploi, constituent le principal moyen par lequel les organisations communiquent leurs postes vacants et leurs exigences aux candidats potentiels. Ces annonces jouent un rôle primaire pour attirer des candidats appropriés et leur fournir un aperçu de ce à quoi s'attendre du poste [51].

La composition d'une OE comprend généralement [51] :

— **intitulé du poste** : le titre spécifique du poste ;
— **présentation de l'entreprise** : brèves informations sur l'organisation telles que son nom, son secteur d'activité et son emplacement ;
— **description du poste** : description détaillée des tâches, responsabilités et détails de l'environnement de travail associés au poste ;
— **description du profil et/ou qualifications et compétences** : liste des compétences, expériences et exigences éducatives essentielles et souhaitables ;
— **autres informations** : cette section fournit des détails sur la rémunération, les informations contractuelles et toutes autres exigences spécifiques aux employés.

Ci-dessous un exemple d'une OE contemporaine est présenté :

Ingénieur Logiciel TAL - Systèmes Financiers & Sécurité

Description de l'entreprise : nous sommes une entreprise fintech leader axée sur les systèmes financiers et la sécurité. Opérant mondialement dans les principaux centres financiers, notre équipe diversifiée utilise des technologies avancées, l'analytique des données, et l'expertise du secteur pour créer des solutions de premier ordre. Nous servons des banques, des institutions financières, des sociétés d'investissement et des startups fintech avec des produits personnalisés, visant à remodeler la finance

avec des systèmes sécurisés et intelligents.

Description du poste : en tant qu'Ingénieur Logiciel, vous travaillerez en étroite collaboration avec nos experts financiers et notre équipe de sécurité pour développer et mettre en œuvre des systèmes financiers à grande échelle.
Vos responsabilités incluront :
- conception et développement des logiciels bancaires sécurisés avec un accent sur le TAL ;
- rencontrer des équipes interfonctionnelles pour assurer l'intégration avec d'autres applications bancaires ;
- enrichir les protocoles de sécurité avec le TAL pour la protection des données bancaires ;
- se conformer aux réglementations de l'industrie dans le secteur financier.

Description du profil :
Expérience : 3 ans et plus en développement des logiciels financiers avec un accent sur le TAL.
Compétences : programmation en Python. Connaissance approfondie du modèle BERT.
Prérequis : connaissance des réglementations financières et des protocoles de sécurité.
Détails du contrat : poste à temps plein avec salaire et avantages compétitifs.
Lieu : Metz, France.

La rédaction d'une OE efficace nécessite une considération stratégique. Elle doit être suffisamment convaincante pour attirer les bons candidats, claire dans ses exigences pour éviter les erreurs d'interprétation, et précise dans sa représentation du poste et de l'entreprise [52].

Les OE ont souvent un format non structuré [37]. Les descriptions du poste ainsi que les qualifications demandées et les autres exigences liées au poste sont généralement exprimées sous forme de textes libres, reflétant le style spécifique de l'organisation ou du recruteur [53]. Cela les rend susceptibles à la variabilité et à la subjectivité. Par exemple, une OE donnée peut indiquer "5 ans d'expérience requis", tandis qu'une autre peut dire "5 ans d'expérience préférés". Ces facteurs posent des défis dans l'extraction et l'interprétation automatiques des informations en CCO [43].

Récemment, les efforts pour automatiser le traitement des OE en utilisant le TAL et l'AA se sont multipliés [54]. Certaines approches s'appuient sur des architectures neuronales multi-couches pour estimer automatiquement les informations les plus pertinentes dans une OE [5]. D'autres méthodes se concentrent sur la représentation du document à travers un réseau de concepts correspondant aux compétences professionnelles, obtenu par des ontologies [34]. Quelle

que soit la méthode, l'objectif principal est de favoriser des systèmes qui modélisent l'appariement des emplois et pré-sélectionnent des candidats appropriés [55].

Néanmoins, des obstacles persistent. Les défis englobent la compréhension des exigences nuancées [37], le décryptage du jargon de l'industrie [46], la gestion des ambiguïtés [43, 45], et la manipulation des informations incomplètes ou partielles [56].

À cet égard, il est important de souligner que la plupart des méthodes proposées ne fournissent pas de cadre formalisé pour gérer l'incertitude et les ambiguïtés des OE. Bien que [43] ait proposé une représentation floue pour les OE, elle est insuffisante pour identifier et prioriser automatiquement les exigences essentielles du poste. De même, les modèles basés sur des transformateurs, malgré leur potentiel, fonctionnent souvent comme des "boîtes noires", laissant leurs critères de priorisation des exigences des OE peu clairs.

D'autres efforts ont visé à encapsuler les OE à travers des ontologies [54], mettant l'accent sur la représentation sans approfondir l'extraction des prérequis essentiels de la description du poste. Certains modèles tentent même d'augmenter les architectures récentes basées sur des transformateurs en injectant des caractéristiques textuelles (comme le fait si un mot apparaît dans le titre, le nombre d'années d'expérience requis, etc.) [57]. Cependant, malgré leur importance, ces caractéristiques sont souvent éclipsées par les structures neuronales complexes du modèle [57].

En résumé, cette revue de la littérature suggère des opportunités de recherche dans l'interprétation et la représentation automatisées des OE. Malgré des avancées significatives, des limitations subsistent dans le traitement et la compréhension de ces documents. Parmi les défis-clés on trouve l'extraction des exigences essentielles et nuancées des OE, l'interprétation des termes spécifiques à l'industrie, et la représentation de l'incertitude inhérente aux OE. De plus, le fonctionnement des modèles d'AP dans l'identification et la priorisation des exigences spécifiques des OE reste opaque, potentiellement en désaccord avec les besoins et les attentes des recruteurs et des organisations. Cela souligne la nécessité de méthodes d'interprétation automatisée des OE plus transparentes, efficaces et alignées sur les recruteurs.

I.1.2 Le Curriculum Vitæ (CV) en tant que type de document

Un CV est un document individualisé qui offre un aperçu de l'historique professionnel, des qualifications et des compétences d'un candidat [24]. En tant qu'outil principal utilisé par les demandeurs d'emploi pour offrir leurs compétences et leur expérience auprès des employeurs potentiels, le CV sert de pierre angulaire à la plupart des processus de recrutement [58].

Un CV comprend généralement plusieurs sections clés, disposées dans un format non structuré, qui offrent des informations sur les différentes facettes du profil d'un candidat. Celles-ci incluent généralement [55] :

— **informations de contact** : des détails personnels de base, y compris le nom, le numéro de téléphone, l'adresse électronique et parfois l'adresse résidentielle ;

— **objectif ou résumé** : une brève déclaration qui décrit les objectifs de carrière et les qualifications du candidat ;
— **expérience** : description détaillée des expériences professionnelles passées, des rôles, des responsabilités et des réalisations clés ;
— **éducation** : contexte éducatif formel, y compris les diplômes, les écoles fréquentées et les dates d'obtention du diplôme ;
— **compétences** : une liste des compétences professionnelles et techniques pertinentes pour le poste ;
— **certifications et licences** : certifications, licences ou formations supplémentaires susceptibles de renforcer les qualifications du candidat ;
— **références (dans certains pays)** : coordonnées des personnes pouvant attester des aptitudes, de l'éthique de travail et du caractère du candidat.

La Figure I.1 illustre un CV dans un format relativement ancien et structuré, tandis que la Figure I.2 présente un format plus récent et moderne, caractérisé par une structure plus personnalisée et un graphisme plus élaboré. Il est important de souligner que, pour des raisons de protection des données, les exemples de CV et OE fournis dans ce livre ne sont pas de documents réels mais des constructions synthétiques basés sur des cas réels.

Pierre DURAND
Ingénieur Logiciel - Solutions Agroalimentaires

Plus de 4 ans d'expérience en développement logiciel axé sur le secteur agroalimentaire. Expert en solutions technologiques pour améliorer la chaîne d'approvisionnement et le suivi des produits. Passionné par l'innovation et la collaboration interdisciplinaire.

Parcours profesionnel

Ingénieur Logiciel | AgriTech Innov
Jan. 2020 - Déc. 2022 | Paris

- Développement de solutions pour optimiser le traitement des requêtes clients en agroalimentaire.
- Développement d'outils de reporting pour analyser les feedbacks des consommateurs.
- Mise en place de solutions technologiques pour la traçabilité des produits.

Consultant en Transformation Numérique | GreenSolutions
Mars 2018 - Déc. 2019 | Bordeaux

- Accompagner les fintech agroalimentaires dans leur transition vers des outils numériques modernes.
- Formé des équipes sur l'adoption de nouvelles technologies.
- Optimisation des processus de production grâce à des logiciels innovants.

Développeur Junior | FoodSoft Creations
Jan. 2017 - Fév. 2018 | Nantes

- Conception d'applications pour la gestion de chaînes d'approvisionnement.
- Participation à des ateliers de brainstorming sur l'innovation agroalimentaire.
- Déployment d'applications à grande échelle.

Formation

Sept. 2017 - Sept 2019 | Lille
Master en Informatique - Tech Université
Thèse sur l'application des modèles de recherche opérationnelle dans le domaine du marketing agroalimentaire.

Sept. 2014 - Juin 2017 | Marseille
Licence en Informatique Appliquée - PolyTech
Stage de fin d'études dans une startup technologique axée sur l'agroalimentaire.

Compétences

Programmation : Python, C++, JavaScript
Outils : Git, Jira, Docker
Travail en équipe : Adaptabilité, Communication

Anglais : Courant
Espagnol : Intermédiaire

LOISIRS

Golf (Participant à des tournois locaux)
Cuisine (Spécialité en plats méditerranéens)
Voyage (Exploration des centres financiers mondiaux pour le loisir)

FIGURE I.1 – Exemple d'un CV plus ancien (format pris d'un CV de l'année 2012). Son format a tendance a être plus structuré et moins stylisé.

FIGURE I.2 – Exemple d'un CV contemporain. Son format a tendance a être moins structuré et plus stylisé.

Aujourd'hui, écrire et formater un CV peut être un art en soi, nécessitant une communication claire et concise, en mettant l'accent sur le contenu pertinent et esthétique [59]. Les CV sont également mimétiques, suivant souvent des tendances et des normes acceptées dans le domaine de la recherche d'emploi pour maximiser leur efficacité. L'objectif est de présenter un argumentaire convaincant à l'employeur potentiel, démontrant l'adéquation du candidat pour le poste, souvent sur une ou deux pages [55], une pratique typique sur le marché de l'emploi français.

Les CV sont intrinsèquement caractérisés par leur variabilité et leur subjectivité [60], dépendant souvent de la diversité des formats, des sections en texte libre, des descriptions qui esquissent le profil du candidat et des récits [61]. Par récit, nous entendons l'organisation et la présentation des informations du CV qui, ensemble, communiquent une "histoire" unique de la carrière et des compétences du candidat. Étant donné la complexité que cette hétérogénéité présente, notamment en matière d'extraction et d'interprétation automatisées, la recherche a été orientée vers l'utilisation d'un certain nombre de modèles d'IA pour l'analyse des CV [8].

L'ambition globale a été de développer des méthodes de CCO qui comprennent les CV aussi efficacement que, sinon mieux que, les recruteurs humains [24]. Cependant, plusieurs obstacles persistent, même avec des techniques de pointe. Les défis notables incluent l'intégration du contenu graphique du CV [62], l'interprétation du langage familier mais aussi le langage spécifique au domaine [61], l'identification des expériences et des compétences implicites [63] et la gestion des informations incomplètes [20].

Face à ces complexités, des études récentes ont introduit une variété de représentations de CV, y compris des n-grammes de mots, le tfidf (Term Frequency-inverse Document Frequency) [64], le sac de mots (BOW, Bag of words) [24] et le doc2vec (Doc2Vec, Document to Vector) [65]. De plus, des incorporations dérivées d'autres modèles d'AP, tels que les architectures de réseaux neuronaux convolutifs (CNN, Convolutional Neural Networks), ont été utilisées pour générer automatiquement des représentations vectorielles de textes de CV [61].

Les représentations mentionnées partagent certaines limitations :

— elles ne tiennent pas compte des aspects graphiques des CV, négligeant ainsi la valeur communicative portée par la mise en forme ;
— elles manquent de représentation de l'incertitude associée aux expressions ambiguës (par exemple, le terme "Avancé" dans "Compétences avancées en Python" peut avoir des significations variables selon la perception de chaque candidat et recruteur) ;
— elles ignorent la variabilité terminologique intrinsèque aux vocabulaires des candidats ;
— bien que les représentations centrées sur l'apprentissage profond soient computationnellement plus efficaces, elles introduisent de nouveaux défis liés au manque de transparence et à l'éthique ; leur nature de "boîte noire" obscurcit la logique de prise de décision en CCO, rendant difficile d'expliquer clairement quelles informations spécifiques du CV d'un candidat justifient sa sélection par rapport à d'autres.

D'autres approches, telles que les représentations basées sur des ontologies [34], sont également limitées en raison de l'absence d'un cadre solide pour les adapter aux connaissances spécifiques au marché et aux organisations [24]. De même, les méthodes de représentation floue [43] visent à tenir compte de la subjectivité des expressions textuelles dans les CV (par exemple, l'expression "expert en Java", qui peut être interprétée de manière variable par différents employeurs) à travers des fonctions floues, mais ne traitent pas les incertitudes liées à la variabilité terminologique et aux aspects graphiques des CV.

Enfin, diverses architectures basées sur des modèles transformateurs [36] se concentrent principalement sur le contenu textuel, supposant souvent un format de document semi-structuré ou complètement structuré. Cela pose une limitation pour une automatisation robuste de la CCO dans des contextes où les CV présentent un contenu graphique diversifié et un format non structuré.

Dans ce contexte de limitations et de divers défis, l'introduction de champs d'étude tels que la grapholinguistique, qui explore la manière dont les structures linguistiques sont représentées graphiquement à travers les systèmes d'écriture, peut potentiellement contribuer au traitement du format non structuré et graphique des CV [66]. En ce qui concerne l'évaluation et l'extraction des informations à partir de ces documents, des méthodes de prise de décision multicritères floues, telles que le Processus Analytique Hiérarchique Flou (FAHP, Fuzzy Analytic Hierarchy Process), peuvent offrir des solutions aux défis de l'analyse automatisée des CV dans la CCO [67].

À cet égard, le FAHP peut permettre une formalisation et une quantification structurées des critères d'évaluation des compétences professionnelles utilisés par les recruteurs dans les CV. Ces critères peuvent être associés au type de section du CV où la compétence professionnelle est citée, au niveau d'expertise exprimé (junior, senior, confirmé, etc.) ou aux années d'expérience. Cette approche pourrait faciliter une évaluation plus formelle, objective et cohérente des compétences qui sont souvent ambiguë et subjectives dans les CV, contribuant ainsi à renforcer la robustesse des systèmes CCO.

En résumé, bien que les CV soient indispensables dans le recrutement, automatiser leur extraction de contenu et leur interprétation demeure un défi. Malgré les avancées, des problèmes subsistent, tels que la compréhension de formats divers, la variabilité terminologique incertaine, et la gestion des informations incomplètes. Différentes techniques d'extraction existent, chacune avec ses limites. Cela présente une avenue riche pour la recherche, visant à développer des outils d'analyse de CV automatisés, plus complets et transparents pour la CCO.

I.1.3 Limitation des corpora publics dans le domaine

L'absence notable de corpus publics dans le domaine de la CCO confère une importance centrale au corpus utilisé dans cette thèse et qui sera introduit dans le Chapître III sur les applications de notre recherche. Cette pénurie de données publiques limite les comparaisons

quantitatives avec des modèles de l'état de l'art et entrave la validation de nouvelles approches dans le domaine de la CCO. Plusieurs raisons expliquent cette situation :

- **propriété privative :** la plupart des modèles de l'état de l'art dans notre domaine ne sont pas open source, limitant ainsi la possibilité de réaliser des comparaisons quantitatives approfondies ;
- **disponibilité des données :** à notre connaissance, il n'y a pas de jeux de données publics disponibles qui permettent l'extraction d'informations à partir des OE, ou l'analyse de CV non structurés récents provenant de contextes réels sans subir d'altérations drastiques ; la carence de jeux de données aux origines transparentes et aux contextes de collecte clairement définis constitue un obstacle majeur à la recherche dans ce secteur ;
- **problématiques de confidentialité** : la nature sensible des informations contenues dans les CV et les OE rend difficile le partage public de ces données, même lorsqu'elles sont anonymisées, en raison de considérations éthiques et légales de plus en plus strictes [1].

Ayant décrit les principales limitations associées à la disponibilité de corpora publics, nous poursuivons notre exploration de l'état de l'art en introduisant le rôle des compétences professionnelles dans la CCO.

I.1.4 Le rôle central des compétences professionnelles

Les compétences professionnelles sont un élément clé du marché de l'emploi, servant de pont crucial entre les capacités d'un individu (telles que présentées dans les CV) et les exigences d'un poste (détaillées dans les OE) [68, 69]. Les compétences, qui incluent les savoirs, le savoir-faire et les compétences sociales, sont mesurables, jouant un rôle non négligeable dans la carrière professionnelle de l'individu [70].

Dans le contexte de la CCO, une évaluation précise des compétences professionnelles est essentielle. Alors que les CV fournissent les compétences dont un candidat est porteur, les OE mettent en évidence les compétences qu'un employeur valorise pour le poste spécifique [71]. Assurer une identification, une représentation et une correspondance précises de ces compétences est impératif pour des procédures de CCO efficaces [71, 72].

Pour normaliser la définition et la représentation des compétences professionnelles, des ontologies spécifiques ont été établies [69]. Elles fonctionnent comme des cadres de référence internationaux. Les ontologies servent de vocabulaires formalisés de termes, souvent structurés hiérarchiquement en classes ou concepts, sous-classes et relations [34]. Elles représentent des entités au sein d'un domaine et mettent en évidence les interconnexions entre elles. Essentiellement, elles agissent comme un plan structuré pour organiser et interpréter les informations au

1. https://www.lemonde.fr/pixels/article/2023/06/22/ai-act-comment-l-union-europeenne-investit-deja-dans-des-intelligences-artificielles-a-haut-risque-pour-controler-ses-frontieres_6178669_4408996.html

sein d'un domaine de connaissance spécifique [73].

Dans la CCO, des structures semblables aux ontologies ont été utilisées pour représenter des connaissances spécifiques liées aux OE et aux dynamiques de travail. Ces structures, bien qu'elles ne soient pas toujours des ontologies complètes, offrent des définitions et des représentations standardisées des compétences professionnelles [34].

Parmi ces structures figurent le Cadre des Compétences, Qualifications et Métiers Européens (ESCO, European Skills, Competences, Qualifications and Occupations) [69], le Réseau d'Information sur les Professions (O*NET, Occupational Information Network) aux États-Unis [2], le Répertoire Opérationnel des Métiers et des Emplois (ROME) principalement utilisé en France [3], et le référentiel du Club Informatique des Grandes Entreprises Françaises (CIGREF) [4], qui est un système de référence pour les professions du numérique.

Le cadre ESCO est une initiative de l'Union Européenne qui identifie et catégorise les compétences, qualifications et métiers pertinents pour le marché du travail et la formation et l'éducation européennes [69]. Il comprend un thésaurus multilingue et fournit une structure de classification systématique, ce qui aide à comprendre et à comparer les qualifications et compétences à travers l'Union Européenne. ESCO permet une plus grande transparence et comparabilité des qualifications, améliorant ainsi l'appariement d'emploi à travers les frontières.

O*NET, en revanche, est un référentiel assez complet qui englobe un large éventail d'attributs des travailleurs et de caractéristiques des emplois [74]. Il comprend des données sur les compétences, les aptitudes, les activités de travail et les intérêts associés à diverses professions. Ces informations servent de référence, rendant le marché du travail plus transparent et permettant à ses utilisateurs de prendre des décisions de carrière plus éclairées.

ROME est une classification professionnelle française qui organise les intitulés de métier dans une structure hiérarchique basée sur les compétences et les qualifications requises pour ceux-ci [75]. Elle comprend des descriptions détaillées des professions et des compétences requises de chacune d'entre elles, fournissant au marché du travail un langage normalisé pour décrire les professions et les compétences.

Le CIGREF est une association qui fournit un référentiel pour les professions du numérique [76]. Ce référentiel décrit les compétences et les qualifications requises pour divers rôles dans l'industrie numérique, permettant aux entreprises de mieux comprendre et s'adapter au paysage des compétences numériques.

En normalisant le langage et les relations utilisées pour décrire les compétences, ces cadres de connaissances peuvent potentiellement améliorer la précision de la CCO. Ils peuvent servir de guides pour organiser des données du marché du travail diverses et complexes, telles que les OE et les CV [75]. Cependant, un écart significatif subsiste entre le langage de ces standards

2. https://www.onetonline.org/
3. https://www.pole-emploi.fr/employeur/vos-recrutements/le-rome-et-les-fiches-metiers.html
4. https://www.cigref.fr/

et le jargon utilisé dans ces derniers documents. Cet écart reflète souvent l'évolution rapide du marché [77].

Reconnaissant cet écart, il existe un domaine de recherche émergeant explorant le potentiel d'apprentissage des ontologies directement à partir de documents tels que les CV et les OE [78]. Ce domaine, connu sous le nom d'apprentissage des ontologies, vise à affiner les sources de connaissances existantes sur la base des terminologies du monde réel. En tirant parti des techniques de l'IA, telles que le TAL et l'AA, il pourrait permettre une CCO plus rapide et plus adaptative [77].

Néanmoins, la nature dynamique des compétences professionnelles, entraînée par des facteurs tels que les avancées technologiques et les besoins changeants de la société et de l'industrie, reste un défi persistant. Assurer la pertinence et l'actualité des structures de connaissances des compétences professionnelles nécessite des révisions de documents à jour, d'évaluer les métriques de qualité des ontologies, et des adaptations régulières de ces sources de connaissances [78, 79].

I.1.5 Le lien entre l'OE, le CV et la compétence professionnelle : la phase de présélection

Le processus de recrutement occupe une position-pivot dans la gestion des ressources humaines d'une organisation, visant à identifier, attirer et embaucher des candidats qui répondent aux critères techniques et socio-culturels du poste [80]. L'une des étapes-clés de ce processus est la phase de présélection, qui sert de filtre initial, évaluant les candidats par rapport aux exigences du poste établies [81, 82].

Plutôt que d'être un simple processus d'élimination basé sur les qualifications en termes d'éducation ou les années d'expérience, la phase de présélection emploie aujourd'hui une approche plus holistique. Elle vise à évaluer les candidats sur une gamme de facteurs, y compris les compétences techniques essentielles, les compétences relationnelles souhaitables — telles que les aptitudes à la communication pour les postes de management — ainsi que des attributs plus difficiles à évaluer comme l'adéquation culturelle et les aspirations professionnelles [81].

La présélection est généralement effectuée par un examen initial des CV. Dans le cadre de cette analyse, les recruteurs évaluent les notices biographiques des candidats, les ensembles de compétences, l'expérience, les qualifications en termes d'éducation et de formation et les réalisations [83]. Ils cherchent à identifier dans quelle mesure le candidat correspond au poste, et pour ce faire, ils s'appuient sur un ensemble de critères ou de marqueurs définis par la nature du poste et par les attentes de l'organisation [82, 83].

Comme cela a été précédemment noté, les récentes avancées technologiques dans le domaine de la CCO ont introduit de nouvelles perspectives aux processus de présélection. Les outils issus de la CCO assistent maintenant les utilisateurs tout particulièrement dans l'évaluation automatisée des compétences techniques [38] et l'évaluation de personnalité [84] à partir des CV

des candidats. Ces innovations ont cherché à rendre une gestion de grands volumes de candidats plus efficace [71].

Néanmoins, l'automatisation complète de la phase de présélection rencontre toujours des limitations. Des défis subsistent dans l'interprétation des résultats issus de systèmes de présélection automatisés opaques [61]. Effectuer une présélection de manière adaptée à chaque point de vue du recruteur est toujours difficile [85]. La formalisation des marqueurs textuels pour la comparaison CV-OE, marqueurs qui s'alignent à la fois sur les objectifs du recruteur et de l'organisation, demeure un problème ouvert [86].

Ces marqueurs textuels, essentiels à la formalisation mentionnée, sont des caractéristiques textuelles spécifiques dotées d'une signification/importance organisationnelle ou sociétale [87]. En tant que caractéristique textuelle, un marqueur textuel se manifeste comme un élément distinctif employé pour structurer, organiser ou mettre en relief des informations particulières dans le texte de l'OE et du CV. Il peut s'agir d'un mot, d'une phrase ou d'un ensemble d'éléments linguistiques servant à indiquer des relations logiques, structurelles ou sémantiques significatives entre les segments du document.

Les marqueurs textuels s'avèrent donc souvent indispensables à l'interprétation correcte du texte, alignée sur les attentes spécifiques d'un contexte organisationnel [87, 88]. C'est pourquoi l'absence de leur étude et formalisation affecte la qualité des CV sélectionnés dans la CCO. En effet, sans une analyse approfondie des marqueurs, l'extraction des exigences les plus pertinentes des OE, qui correspondent aux attentes organisationnelles et des recruteurs, s'en trouve limitée. Dans ce cadre, l'étude et l'interprétation adéquates de ces éléments demeurent au cœur des défis de la phase de présélection [37, 85].

D'autres problématiques émergent également lors de cette phase, tels que l'évaluation de la qualité des données [89] et de l'incertitude de la présélection automatisée [85], ainsi que la satisfaction des besoins en volume de données des modèles d'AP [57].

En résumé, bien que les technologies issues de la CCO aient fourni des outils précieux pour gérer d'importants bassins de candidats et pour évaluer des compétences spécifiques dans la phase de présélection, des défis subsistent. Ceux-ci comportent des questions de transparence, de personnalisation et d'assurance de l'alignement avec les attentes du recruteur et les objectifs organisationnels globaux. De plus, un approfondissement supplémentaire de la qualité des données, de la quantification de l'incertitude et des prérequis en données pour des modèles complexes (comme les modèles d'AP) est toujours nécessaire.

I.1.6 Les recruteurs et leur pertinence dans la phase de présélection

Les recruteurs jouent un rôle indispensable dans le processus de recrutement, en particulier dans la phase de présélection, où ils filtrent un grand nombre de candidatures pour identifier les candidats les plus prometteurs [89]. Leur expertise dans l'analyse des CV et des OE est

fondamentale pour un recrutement réussi. Cependant, cette expertise a rarement été explorée et intégrée dans les méthodes actuelles de CCO [24].

Dotés d'une compréhension nuancée des exigences du poste, les recruteurs excellent à mettre en corrélation les compétences, l'expérience et les qualifications d'un candidat avec ces demandes [90]. Leur expertise s'étend au-delà de la compréhension des exigences techniques complexes ; ils valorisent également les compétences relationnelles et d'autres facteurs intangibles (et éthiquement fragiles) comme l'adéquation culturelle au sein de l'organisation [91].

Les recruteurs possèdent la capacité d'analyse de CV complexes, même lorsqu'ils sont confrontés à des descriptions ambiguës ou à un langage technique [92]. De plus, les perspectives uniques des recruteurs leur permettent de déduire des qualifications non exprimées, telles que le potentiel de croissance ou les capacités de travail en équipe [89, 91].

Malgré leurs compétences distinctives, les recruteurs sont toujours confrontés à des défis significatifs dans les processus de présélection actuels. Le volume important de candidatures associé aux contraintes de temps peut rendre le processus de présélection redoutable [37]. De plus, au dépit de leurs meilleures intentions, les recruteurs peuvent être influencés par des biais individuels, consciemment ou inconsciemment, affectant les résultats du processus de présélection [91] et le rendant fortement affecté par des incertitudes cognitives [45].

Compte tenu de ces problèmes, une tendance croissante se dessine vers l'utilisation de l'IA pour modéliser la gestion des candidatures, intégrant plus étroitement les perspectives des recruteurs [37]. Il est important de souligner que cette relation entre les recruteurs et l'IA est mutuellement bénéfique. Les recruteurs exploitent non seulement l'IA pour rationaliser leurs flux de travail, mais contribuent également de manière significative à l'amélioration des techniques de CCO grâce à leurs compétences uniques [57].

Alors que les recruteurs ont généralement aidé à valider les méthodes de CCO existantes en fournissant des ensembles de données annotés, on trouve peu de travaux dans la littérature sur la manière d'intégrer formellement leurs perspectives [93, 91]. Une telle intégration est particulièrement nécessaire pour les tâches visant à identifier les composants-clés au sein du processus de recrutement (par exemple, l'extraction d'informations pertinentes à partir des OE) et à aligner la CCO avec un contexte organisationnel spécifique [28, 44, 24].

De plus, les études existantes, tout en étant instructives sur la CCO, négligent souvent le niveau d'expertise des recruteurs impliqués [33, 37, 24, 71]. Cela souligne la nécessité d'impliquer des recruteurs ayant une expérience directe dans la gestion des CV et OE étudiés. Une telle implication active est en passe, non seulement d'activer l'expertise des recruteurs [28], mais également de leur donner une perspective unique qui les distingue des recruteurs qui ne sont pas directement impliqués [89].

Un autre élément crucial à prendre en compte est le contexte organisationnel dans lequel les recruteurs opèrent [28]. Ce contexte influence de manière significative le processus de présélection

en façonnant les priorités, les outils et les approches que les recruteurs utilisent pour évaluer les candidats [89]. Ces influences organisationnelles font l'objet de la section suivante.

Pour conclure, la connaissance spécialisée des recruteurs est une ressource inestimable mais sous-utilisée dans l'évolution de la CCO. En intégrant leur expertise dans les systèmes pilotés par l'IA, le potentiel d'amélioration de la précision et la pertinence des processus de CCO existe, en particulier dans des tâches comme l'interprétation des CV et des OE. Reconnaître et exploiter la compréhension nuancée offerte par les recruteurs peut ouvrir la voie à des approches plus adaptatives et sensibles au contexte.

I.1.7 Le contexte organisationnel : l'environnement négligé de la CCO

Le contexte organisationnel dans la CCO est un aspect souvent sous-estimé, mais indispensable. Au-delà de l'alignement des compétences et de l'expérience, la sélection des candidats implique de trouver une résonance avec la culture unique, les normes et les objectifs stratégiques d'une organisation [28, 94]. Ce processus d'appariement est multifacette, allant au-delà de la simple infrastructure et de la hiérarchie pour englober des aspects tels que les valeurs partagées, les normes de travail, les règles non-dites et les compétences [95].

Les organisations ont des identités uniques, façonnées par des éléments culturels, structurels et opérationnels [96]. Ces identités guident la conception des rôles professionnels, la valorisation des compétences et l'exécution du travail [97]. Le défi réside dans l'intégration de ce contexte complexe dans les systèmes d'IA pour l'appariement professionnel, un domaine où les méthodologies sont encore peu explorées. Ce manque d'intégration peut potentiellement conduire à un désalignement entre les systèmes et leur contexte organisationnel, affectant le succès de la sélection des candidats au fil du temps [24].

Dans le domaine multifacette des systèmes organisationnels, la compréhension des dynamiques intrinsèques est nécessaire [98, 99]. Ces systèmes peuvent être considérés comme des systèmes adaptatifs complexes (CAS, Complex Adaptative Systems) qui évoluent continuellement [100]. Un examen détaillé inclut l'identification des acteurs-clés, des relations, des artefacts et des mécanismes sous-jacents qui constituent le tissu organisationnel [101]. La théorie causale est ici un outil stratégique, permettant la création de modèles mentaux qui décodent les relations de cause à effet sous-jacentes dans les systèmes complexes [102], tandis que le Processus de Développement du Produit (PDP) a été un outil précieux pour adapter la théorie causale à l'AA et guider l'implémentation de méthodes CCO [103]. Le PDP englobe diverses approches pour représenter les composants organisationnels, intégrant les perceptions de divers rôles tels que les managers, les développeurs et les clients [104].

De manière générale, le PDP présente encore plusieurs limites [44]. Par exemple, les équipes de développement/recherche manquent souvent de diversité, rendant potentiellement le produit résultant moins représentatif du problème global et universellement applicable. De plus, le PDP

a tendance à se concentrer étroitement sur les spécifications techniques sans prendre en compte le contexte sociétal/organisationnel plus large. Troisièmement, les théories sous-jacentes (cause-effet) qui guident le développement du produit ne sont pas toujours explicitement formalisées, entravant leur vérification et leur amélioration. Enfin, les perspectives des parties prenantes périphériques, comme les décideurs politiques, sont souvent négligées, ce qui affecte l'efficacité et l'équité du produit dans le monde réel.

Les implications du PDP deviennent évidentes lorsque ces limitations affectent le développement et la recherche des systèmes CCO. Un manque de diversité au sein de l'équipe de développement/recherche pourrait entraîner une méthode de CCO qui ne prend pas en compte de manière adéquate un large éventail d'organisations, de chercheurs d'emploi et de besoins des employeurs. L'accent étroit mis sur les aspects techniques/scientifiques peut négliger les facteurs humains et organisationnels complexes qui influencent l'appariement professionnel, tels que les changements dynamiques du processus de recrutement et l'adaptabilité des recruteurs à l'offre et à la demande de compétences sur le marché. L'absence de théories formalisées explicitement (cause-effet) pourrait conduire à des méthodes CCO fonctionnant selon des hypothèses non testées, introduisant potentiellement des biais ou des imprécisions dans la littérature. Enfin, l'ignorance des parties prenantes périphériques pourrait entraîner un système CCO déconnecté des lois du travail (par exemple, le règlement sur l'IA de l'UE [49] [5]), compromettant ainsi son équité et sa responsabilité sociale.

Compte tenu de ces limites, les communautés académique et industrielle ont lancé des efforts pour affiner et faire progresser les théories et les méthodologies sous-tendant le PDP. En conséquence, il y a eu un changement discernable vers l'incorporation de cadres plus larges et plus inclusifs qui abordent ces problèmes de front.

Ces limites ont stimulé les innovations en théorie causale, conduisant à la formation collaborative de théories causales au sein du domaine émergent de l'AA équitable (*fair machine learning*) [44]. Cela représente un écart par rapport aux méthodes traditionnelles, visant à adopter une approche plus inclusive qui s'aligne avec la nature adaptative interne et externe des organisations et le contexte sociétal des solutions d'AA.

Des outils tels que les systèmes dynamiques [105] et les diagrammes de boucle causale [106] sont des exemples prometteurs de ces concepts en action. Cependant, l'appel à des méthodes plus intégratives est fort. La méthode UNC (Universidad Nacional de Colombia – méthode de développement logiciel) [107] émerge comme une innovation significative dans ce domaine. Elle propose un cadre robuste et complet conçu pour comprendre et mieux intégrer le contexte organisationnel de la CCO. Cette méthodologie reconnaît qu'une représentation robuste du contexte d'une méthode d'AA nécessite une analyse progressive et multicouche, impliquant toutes les parties prenantes concernées et plusieurs dimensions de l'organisation (acteurs, artefacts, objectifs,

5. `https://www.artificial-intelligence-act.com/`

problèmes, processus, comportements et intérêts des acteurs, spécificités du contexte sociétal comme la dynamique du marché, entre autres). Bien qu'elle puisse ne pas offrir de mécanismes pour gérer les représentations dans le temps, elle fournit une stratégie de représentation générale unifiée sur des schémas préconceptuels, permettant l'adaptation et l'intégration d'outils externes tels que les retards temporels (time delays) [108].

En résumé, en examinant les complexités et les limites inhérentes aux méthodologies présentées, notamment le PDP, cette section établit un contexte pour faire progresser le domaine de la CCO. Les paradigmes émergeants, tels que la méthode UNC, offrent des voies prometteuses pour créer des méthodes CCO plus robustes, adaptables, inclusives et équitables en capturant le contexte organisationnel multifacette qui influence l'appariement professionnel. Cela pourrait non seulement élucider les défis qui se posent pour intégrer le paysage organisationnel, mais également fournir les bases théoriques nécessaires pour les naviguer.

I.2 Intelligence Artificielle (IA) pour un processus de CCO automatisé

Les sections suivantes fournissent un sommaire des différentes facettes de l'IA appliquée à la CCO. Elles commencent par un aperçu de l'IA pour l'appariement professionnel, offrant une comparaison détaillée des principales approches telles que l'IA symbolique, l'AA, l'AP, TAL et le raisonnement approximatif avec des modèes Croyances-Désirs-Intentions gradués en utilisant la théorie des possibilités et la logique floue. Les sous-sections suivantes se penchent sur chacune de ces méthodologies, disséquant leurs forces, leurs faiblesses et leurs applications spécifiques dans le contexte de l'appariement professionnel.

I.2.1 IA symbolique et apprentissage automatique

Aux premiers jours de l'IA, l'*IA symbolique* a jeté les bases en formalisant certains aspects des modèles de raisonnement. Entre les années 1950 et 1980, l'IA symbolique s'est concentrée sur le raisonnement logique et les systèmes à base de règles, devenant un moteur clé pour les avancées initiales dans des domaines comme le TAL [109]. Bien que ce type d'IA semble éloigné des méthodologies centrées sur les données d'aujourd'hui, l'influence de l'IA symbolique peut encore être retracée dans les techniques TAL fondamentales qui sont pertinentes dans les méthodes basées sur l'analyse du texte comme la CCO [110].

L'AA a marqué un changement significatif par rapport à ces méthodologies basées sur des règles, introduisant un paradigme axé sur les données [111]. Apparaissant initialement dans les années 1980 et gagnant en importance dans les années 1990, il a jeté les bases pour des sous-domaines avancés comme l'AP, qui influence maintenant de manière significative les techniques de TAL.

L'IA symbolique et l'AA illustrent comment les méthodologies de l'IA ont évolué, d'abord du raisonnement basé sur des règles à la reconnaissance de motifs vers des techniques plus avancées et spécialisées, telles que l'AP et le raisonnement approximatif qui seront élaborés dans les sections suivantes.

I.2.2 Apprentissage profond

L'aprentissage profond (AP) est un sous-domaine de l'AA qui se concentre sur la formation et l'utilisation de réseaux neuronaux artificiels [112]. Le terme "profond" se réfère à la présence de nombreuses couches dans le réseau, où les entrées sont traitées et transformées en représentations plus abstraites et composées [113].

L'AP trouve ses racines dans les premiers réseaux neuronaux artificiels. Alors que les premiers modèles étaient peu profonds ne comportant guère plus d'une ou deux couches, le développement de l'algorithme de rétropropagation dans les années 1980 a permis la formation de réseaux plus profonds [114]. Cependant, l'AP n'a pas beaucoup progressé jusqu'aux années 2000 en raison de contraintes de calcul et du manque de grands ensembles de données nécessaires à l'entraînement [115].

Le tournant est venu en 2012 lorsqu'un modèle CNN, AlexNet, a nettement surpassé les autres modèles dans le concours ImageNet, un concours populaire de reconnaissance d'image. Le succès d'AlexNet [116], associé à la montée des unités de traitement graphique pour le calcul et à la disponibilité de grandes données, a ouvert la voie à la révolution de l'AP.

L'AP utilise divers types de réseaux neuronaux, chacun conçu pour des types de tâches spécifiques. Les principaux incluent [113] :

- réseau de neurones à propagation avant (FNN, Feedforward Neural Network) : le type le plus simple de réseau neuronal artificiel ; lors de l'utilisation de ce type de réseau pour des prédictions (et non lors de l'apprentissage) les informations dans les FNNs circulent dans une seule direction, de la couche d'entrée, à travers une ou plusieurs couches cachées, vers la couche de sortie ;
- réseaux neuronaux convolutifs (CNNs, Convolutional Neural Networks) : ceux-ci sont principalement utilisés dans les tâches de traitement d'image ; ils utilisent des couches convolutifs qui appliquent des filtres aux données d'entrée, permettant au réseau d'apprendre automatiquement les hiérarchies spatiales des fonctionnalités ;
- réseaux neuronaux récurrents (RNNs, Recurrent Neural Network) : les RNNs sont conçus pour les données séquentielles, car ils maintiennent un état caché qui contient des informations sur les entrées passées ; des variantes telles que la mémoire à long-court terme (LSTM, Long-short Term Memory) et les réseaux d'unités récurrentes à porte (GRU, Gated-recurrent Unit) atténuent le problème de disparition du gradient dans les RNNs standard ;

— transformateurs : initialement conçus pour les tâches de traitement du langage naturel, les transformateurs utilisent des mécanismes d'attention pour peser la pertinence des points de données d'entrée ; ils ont montré un succès remarquable dans des tâches comme la traduction de langues, la génération de texte, et plus récemment même dans la reconnaissance d'image ;

— grands modèles de langage (GML) : dans le contexte du TAL, les transformateurs ont donné naissance à des Grands Modèles Linguistiques tels que GPT (Generative Pre-trained Transformer) [117], BERT [118], PaLM (Pathways Language Model) [119] et LLaMA (Large Language Model Meta AI) [120] ; ces GML ont établi de nouveaux repères dans des tâches telles que la résumé de texte, la traduction et la réponse aux questions [121] ; leur potentiel s'étend aux domaines comme la CCO, où la compréhension du contenu sémantique des OE et des CV est essentielle [121].

Les modèles d'AP ont atteint des résultats de pointe dans de nombreux domaines, tels que la vision par ordinateur, le traitement du langage naturel, la reconnaissance vocale et même les jeux. Ils alimentent des technologies comme les voitures autonomes, les assistants vocaux, les systèmes de recommandation et l'appariement professionnel [122].

Malgré ces succès, l'AP présente encore des limites importantes. Il nécessite généralement de grandes quantités de données et de ressources informatiques [123]. Ses modèles sont souvent considérés comme des 'boîtes noires' en raison de leur manque d'interprétabilité. Ils peuvent facilement surapprendre si les paramètres d'entraînement ne sont pas optimaux, mémorisant les données d'entraînement et ayant de mauvaises performances sur des données inconnues [113].

De plus, les modèles AP rencontrent des défis dans des tâches qui nécessitent un raisonnement, une connaissance du sens commun, une compréhension humaine du monde et une adaptation agile à des contextes en évolution rapide, des limitations particulièrement pertinentes dans le domaine de la CCO [50]. Ce domaine évolue continuellement, avec des recherches actives dans des domaines comme l'apprentissage auto-supervisé, apprentissage à faible nombre d'exemples (*few-shot learning*), et des *modèles émergents* qui intègrent un raisonnement symbolique [124] et approximatif avec des réseaux neuronaux [50, 125]. Il y a également un intérêt croissant pour l'IA explicable, visant à rendre les modèles d'AP plus interprétables et transparents [126], ainsi que pour l'apprentissage multi-modal afin d'intégrer plusieurs types de données lors de l'entraînement des modèles [127].

En résumé, bien que l'AP ait démontré des performances remarquables dans de nombreux domaines, il présente des limites intrinsèques pour des tâches nécessitant un raisonnement complexe et explicable, une connaissance du sens commun, une adaptation rapide à des contextes dynamiques, et une utilisation de petits ensembles de données. Ces limitations sont particulièrement manifestes en CCO. Pour surmonter ces défis, une approche qui fusionne un cadre théorique robuste pour le raisonnement explicatif avec des modèles d'AP de type "boîte grise"

s'avère pertinente. Ces modèles "boîtes grises" cherchent à équilibrer les capacités d'apprentissage des méthodes d'AP, souvent vues comme des "boîtes noires", tout en intégrant des éléments d'explicabilité, s'approchant ainsi des caractéristiques des "boîtes blanches". Dans cette optique, les architectures de transformateurs, notamment les modèles GML comme BERT, peuvent être intégrées en tant que composantes clés de ces "boîtes grises", offrant des avancées notables dans la compréhension sémantique, l'adaptabilité à des petits ensembles de données, tout en bénéficiant de méthodes d'explicabilité. Adopter cette approche pourrait non seulement bénéficier à la CCO, mais aussi stimuler les tendances de recherche émergentes en raisonnement approximatif, deep logic et en explicabilité. Même si l'AP pourrait ne pas être la solution centrale pour des applications en CCO éthiquement et légalement sensibles, il possède un potentiel significatif en tant que technologie complémentaire.

I.2.3 Traitement automatique de la langue

Le TAL est un sous-domaine de l'IA qui se concentre sur l'interaction entre les ordinateurs et les humains en utilisant le langage naturel. L'objectif est de permettre aux ordinateurs de comprendre, d'interpréter et de générer le langage humain de manière utile [128].

Les premiers travaux en TAL dans les années 1950 se sont concentrés sur la traduction automatique, notamment du russe vers l'anglais. Ces systèmes initiaux étaient basés sur des ensembles de règles artisanales. À partir des années 1980 et se poursuivant dans les années 1990, il y a eu une révolution dans le TAL avec l'introduction des algorithmes AA pour le traitement de la langue. Ceci a été suivi par le tournant statistique à la fin des années 1980 et au milieu des années 1990 [129].

Cependant, le XXI^{e} siècle a vu un changement significatif vers des approches basées sur les données. L'avènement d'Internet a fourni des quantités sans précédent de données linguistiques librement disponibles (texte et parole) pour une utilisation en TAL. Ceci, couplé avec les avancées en puissance de calcul, a conduit au développement de modèles de plus en plus sophistiqués basés sur l'AA [129].

Les méthodologies du TAL ont évolué avec le temps, passant des méthodes basées sur des règles et statistiques aux techniques d'AA et d'AP.

— Systèmes basés sur des règles [130] : ces systèmes ont été conçus pour traiter le langage en fonction d'un ensemble de règles écrites à la main. Bien que ces systèmes soient bons pour comprendre les règles spécifiques qu'ils ont été conçus pour suivre, ils ont du mal avec l'ambiguïté et la variabilité du langage humain.

— TAL statistique [131] : avec la révolution statistique à la fin des années 1980 et des années 1990, le TAL a commencé à s'appuyer sur des algorithmes AA pour apprendre automatiquement ces règles, en se servant de techniques telles que les modèles de Markov cachés, le Naïve Bayes, et les machines à vecteurs de support (SVM, Support Vector

Machines).

— TAL neuronal [132] : l'avancée la plus récente en TAL a été le passage aux réseaux neuronaux. Les RNN, LSTM, et plus récemment les modèles GML ont tous rencontré un grand succès en TAL. Les modèles GML, en particulier, en raison de leur capacité à gérer les dépendances à longue portée dans le texte, ont entraîné une série de percées en TAL.

Le TAL a un large éventail d'applications, y compris mais sans s'y limiter, la traduction automatique, l'analyse des sentiments, l'extraction d'informations, le résumé automatique, les réponses aux questions, l'étiquetage des parties du discours, la reconnaissance d'entités nommées et les chatbots. Le TAL est également la technologie clé dans les applications et interfaces basées sur la voix telles que Google Assistant, Alexa, et Siri [133].

Malgré des avancées significatives, le TAL confronte encore divers défis. L'un des problèmes les plus pressants est la compréhension du contexte, de la culture et de l'ambiguïté dans les modèles linguistiques, un défi particulièrement évident dans des modèles à la pointe de la technologie comme GPT [134, 135]. Indépendamment de cela, une autre préoccupation est la présence de biais linguistiques dans les modèles de TAL, qui peuvent perpétuer les biais sociétaux présents dans les données sur lesquelles ils sont entraînés [136]. De plus, il existe des domaines de recherche émergents qui se concentrent sur l'intégration et la normalisation des langages contrôlés, spécifiquement dans le contexte des systèmes orientés agent [137, 107]. Des recherches supplémentaires explorent également des représentations plus nuancées du langage, y compris l'utilisation d'approches possibilistes et floues [138, 139, 140].

Le paysage en évolution du TAL offre des orientations de recherche prometteuses qui sont directement pertinentes pour les défis rencontrés en CCO. L'intégration de l'AP avec le raisonnement approximatif pourrait améliorer les méthodes possibilistes basées sur des ontologies, conduisant à des représentations linguistiques plus nuancées des OE et des CV. De plus, la combinaison des paradigmes possibiliste et flou avec des modèles de TAL à la pointe de la technologie, tels que BERT, pourrait fournir une meilleure gestion des phénomènes sous-représentés comme l'ambiguïté du langage et, plus généralement, l'incertitude. Alors que le TAL lutte avec des questions telles que la compréhension contextuelle et le biais linguistique, ces défis eux-mêmes ouvrent des portes pour des recherches interdisciplinaires précieuses vers des approches CCO plus explicables et sensibles au contexte, visant à produire des représentations textuelles plus pertinentes et évolutives.

I.2.4 Raisonnement approximatif avec des Croyances-Désirs-Intentions (CDI) gradués en utilisant la théorie des possibilités et la logique floue

Le concept de raisonnement approximatif provient des principes établis par Lotfi Zadeh dans les années 1960 et 1970 dans le domaine de la logique floue et de la théorie des possibilités, développant ainsi les fondations de cette approche. C'est une dimension de l'IA qui permet la

manipulation de l'incertitude, de l'imprécision et du flou dans l'information, le langage et la représentation des connaissances [141]. Les CDI gradués, appliquant le raisonnement approximatif, désignent un paradigme où l'incertitude dans les croyances, la variabilité dans les désirs, et l'imprévisibilité des intentions sont capturées et traitées [142].

Cela est souvent combiné avec la théorie des possibilités et la logique floue [143], qui permettent de représenter et de manipuler des données et des informations qui ne sont pas précises ou complètes. La théorie des possibilités est particulièrement efficace pour gérer des informations incomplètes, tandis que la logique floue est compétente pour gérer des informations approximatives [144].

L'origine du raisonnement approximatif peut être retracée aux premiers stades de l'IA, où le besoin de modéliser un raisonnement et une prise de décision humains dans des conditions d'incertitude est devenu apparent [145].

Le modèle CDI, qui postule que les Croyances-Désirs-Intentions sont les moteurs clés du comportement intelligent, a été établi et est largement reconnu dans les domaines de la philosophie et de la psychologie cognitive [146]. Plus tard, il a été adoptée par l'IA, en particulier dans le domaine de la programmation orientée agent, vers les années 1980 [147].

L'intégration des CDI gradués avec la logique floue et la théorie des possibilités a commencé à gagner du terrain vers la fin du XX^{e} siècle, s'alignant avec le changement général vers la manipulation de l'incertitude et de l'imprécision en IA [148].

Les méthodes et techniques-clés de cette approche tournent autour de la représentation et de la manipulation des CDI gradués en utilisant des ensembles flous et des distributions de possibilité [149] :

- les ensembles flous sont utilisés pour représenter et traiter des informations vagues, imprécises ou subjectives ; par exemple, l'énoncé « Niveau d'expert en Java » d'une OE peut être représenté comme un ensemble flou où « expert » est un terme flou avec une fonction d'appartenance qui décrit le degré auquel l'expertise en « Java » est nécessaire ;
- la théorie des possibilités est utilisée pour gérer des informations incertaines ou incomplètes ; elle fournit une distribution de possibilité qui représente le degré auquel chaque valeur possible d'une variable est plausible ;
- la combinaison de ces deux théories avec le modèle CDI implique l'association des CDI sur des ensembles flous et des distributions de possibilité, permettant la manipulation de l'incertitude et du flou.

Le paradigme du raisonnement approximatif a vu des applications étendues dans le domaine de l'IA [141]. Ces applications incluent des systèmes de prise de décision [150], des systèmes de recommandation [151, 152], et plus récemment, bien qu'exclusivement dans le contexte de la logique floue, des systèmes CCO [43].

Malgré des avancées significatives, plusieurs défis subsistent dans l'intégration de cette ap-

proche avec d'autres techniques de l'IA comme l'AA et le TAL [153]. L'efficacité computationnelle est une préoccupation particulière pour les données à grande échelle ou de dimensions élevées, agissant comme un goulot d'étranglement pour les applications industrielles. La gestion des incertitudes complexes comme les données à valeurs d'intervalle ou les données flou-probabilistes est toujours problématique, nécessitant des extensions théoriques supplémentaires. Les applications émergentes dans des domaines comme le choix social (l'agrégation des préférences ou des jugements individuels en une décision collective) et l'analyse de réseau sont encore à exploiter pleinement. Des outils comme l'inférence possibiliste sont encore en développement émergent [149]. Ces problèmes soulignent le besoin de recherches continues en raisonnement approximatif.

En conclusion, à mesure que le domaine du raisonnement approximatif continue de se réinventer et de mûrir, il ouvre plusieurs voies de recherche prometteuses qui pourraient améliorer de manière significative la CCO. Par exemple, l'incorporation de cadres possibilistes pourrait offrir une base théorique solide pour gérer les incertitudes inhérentes qui caractérisent la phase de présélection de candidats. De plus, la fusion de ces cadres avec des modèles GML à la pointe de la technologie a le potentiel de relever des défis persistants, tels que la transparence, l'ambiguïté du langage et l'incertitude dans des tâches comme l'extraction d'informations et l'analyse de documents non structurés. Cette intégration pourrait imprégner le processus d'appariement professionnel d'une adaptabilité accrue, attribuable à la robustesse théorique des distributions de possibilité et des fonctions floues, ainsi qu'à une explicabilité de bout en bout.

I.3 Modèles CCO existants, leurs défis et tendances de recherche émergentes

Cette section présente une prospection dans le domaine en évolution des approches de CCO. Elle commence par un examen du rôle de l'IA dans la phase de présélection pour automatiser l'appariement des CV et des OE. En explorant le mélange dynamique de méthodologies employées dans ce domaine, la section met l'accent sur les techniques AA, les méthodes sémantiques et basées sur l'ontologie, les systèmes basés sur l'IA, et d'autres approches innovantes qui élargissent les horizons de la CCO traditionnelle. La discussion met en lumière les piliers, les défis et les tendances émergentes en matière d'appariement professionnel, jetant ainsi une base solide pour comprendre la méthodologie proposée dans cette thèse.

I.3.1 IA dans la phase de sélection pour l'automatisation de la CCO entre les OE et les CV

L'état de l'art dans le domaine de la CCO est un mélange dynamique de méthodologies, notamment :

Techniques AA

Les techniques AA sont fondamentales dans la CCO :

— **CNN :** une adaptation siamoise de CNN est utilisée pour faire correspondre les OE aux CV des candidats, comme le démontre [64] ;
— **person-job fit neural network (PJFNN) :** proposé dans [37], PJFNN projette les OE et les CV sur une représentation latente partagée, permettant des comparaisons plus efficaces ;
— **extraction de caractéristiques non supervisée :** l'utilisation de l'AA pour l'extraction de caractéristiques non supervisée, permettant de détecter des postes similaires sans recourir à des outils sémantiques supplémentaires, est mise en œuvre par [154] ;
— **techniques de clustering et de similarité :** des caractéristiques textuelles sont utilisées pour former des clusters, puis la similarité est calculée afin d'évaluer les CV, comme le fait [155] ;
— **modèles automatisés :** un modèle d'AA automatisé est introduit par [156], qui catégorise les caractéristiques des CV en vue d'un appariement plus précis ;
— **modèles AP :** tels que ceux proposés dans [157] pour prédire les détails de carrière futurs des employés, permettant des considérations d'appariement à plus long terme ;
— **systèmes de recommandation basés sur plongement de mots :** un système basé sur le plongement de mots est utilisé par [57] pour l'appariement à grande échelle des OE aux candidats, exploitant ainsi la puissance des moteurs de recommandation modernes ;
— **approche de modèle empilé :** une méthode permettant de faire correspondre les CV aux OE grâce à l'utilisation d'une hiérarchie de modèles pour améliorer la précision est présentée par [35].

Méthodes sémantiques et basées sur des ontologies

Ces méthodes exploitent la compréhension sémantique et la représentation structurée des connaissances :

— **RésuMatcher :** pour une approche de CCO plus équilibrée, [55] emploie l'étiquetage sémantique, la correspondance de motifs, des méthodes basées sur l'ontologie et l'AA ;
— **Ontology-based RésuParser (ORP) :** introduit dans [34], l'ORP analyse les CV dans des formats ontologiques, facilitant une analyse sémantique complexe ;
— **Job-onto :** en utilisant une ontologie pour représenter et rechercher des données, [54] extrait des compétences des deux côtés et évalue la correspondance à l'aide de la distance d'édition de graphe, une mesure de similarité.

Méthodes basées sur les graphes

Ces méthodes se concentrent sur des relations complexes :

— **vecteurs de phrases et graphes de termes-sujets :** un modèle présenté par [36] capture les relations en utilisant des graphes et des représentations vectorielles ;

— **réseaux d'attention à plusieurs têtes basés sur des graphes :** en utilisant des mécanismes d'attention de graphes et en tenant compte de plusieurs perspectives, [38] explore la classification des CV ;

— **représentation de graphe à double perspective :** une approche d'apprentissage qui prend en compte à la fois le côté de l'employeur et le côté du candidat pour une perspective équilibrée est proposée par [158].

Systèmes basés sur l'IA

Ces systèmes combinent d'autres techniques de l'IA pour améliorer la précision et l'interprétation :

— **interpretable person-job fit (IPJF) :** en prenant en compte les intentions des employeurs et des chercheurs d'emploi dans la procédure d'appariement, [61] ajoute une couche d'interprétabilité partielle centrée sur les intentions des utilisateurs sur les plateforme de recrutement en ligne ;

— **systèmes d'extraction floue :** un système flou destiné à la CCO dans le domaine de Technologies de l'Information et de la Communication (TIC), qui offre un processus d'appariement plus nuancé, est proposé par [43] ;

— **modèles de parcours d'apprentissage :** en se concentrant sur le développement continu, [159] fournit un parcours d'apprentissage conçu pour combler l'écart entre les OE et les CV des candidats ;

— **apprentissage à partir de cas passés :** en utilisant une forme de raisonnement basé sur les cas, [71] apprend à partir de cas résolus pour prédire de nouveaux résultats ;

— **classement sans OE :** des méthodes uniques pour le classement des CV, qui ne dépendent pas des OE ni des ressources sémantiques, sont présentées par [24] ;

— **approche InEXIT :** en se concentrant sur des attributs multivariés semi-structurés, [160] explore les relations complexes présentes dans les OE et les CV.

Défis dans le domaine de la CCO

Malgré la grande diversité des méthodes, plusieurs limitations persistent dans le domaine de l'appariement professionnel :

— **manque d'analyse complète du point de vue des recruteurs :** les méthodologies actuelles ne fournissent pas une analyse approfondie, une explication et une modélisation

de la pertinence de l'information dans les OE du point de vue unique des recruteurs ; comme ces professionnels sont souvent ceux qui scrutent et interprètent de plus près ces documents, cette absence représente une limitation significative pour comprendre et utiliser leur expertise pour un classement plus pertinent et interprétable des CV ;

— **déficience dans l'évaluation de la pertinence des marqueurs textuels :** il y a un manque notable de méthodologies qui évaluent de manière critique la pertinence des marqueurs textuels au sein des OE par rapport au contexte organisationnel spécifique ; cette absence entrave l'explicabilité des systèmes, surtout dans le domaine des modèles AP ; ces modèles restent souvent opaques en clarifiant les critères de sélection principaux, en raison de leur nature de "boîte noire" ; ce manque de transparence réduit la confiance de l'utilisateur, car il obscurcit les informations sur la logique derrière la sélection automatique des candidats ;

— **manque de prise en compte du contenu non structuré et graphique dans les CV :** les approches existantes évitent souvent les défis des CV non structurés récents en se reposant sur des formats purement textuels, standardisés ou prévisibles ; cette pratique peut conduire à une manipulation manuelle des documents, qui est chronophage et néglige le contenu significatif et les sections des candidats contenus dans des formats plus complexes ou variés ;

— **absence d'intégration du contexte organisationnel :** les méthodes actuelles n'ont pas réussi à proposer une approche qui intègre efficacement le contexte organisationnel spécifique dans lequel une méthode CCO opère ; cette omission a conduit à un désalignement notable et à une dégradation progressive de ces systèmes au fil du temps, réduisant leur pertinence et leur adaptabilité ;

— **ignorance du phénomène d'incertitude de l'information :** les méthodes de la littérature négligent souvent le phénomène complexe et omniprésent de l'incertitude de l'information ; ceci est particulièrement évident dans les OE, qui contiennent intrinsèquement des informations incomplètes ou ambiguës ; le fait de ne pas aborder cette incertitude inhibe le développement de modèles plus robustes et nuancés qui peuvent interpréter et répondre à la variabilité et à la complexité des scénarios CCO dans le monde réel.

En conclusion, l'automatisation de l'appariement professionnel implique diverses techniques, chacune avec ses avantages et ses limitations uniques. Les méthodologies existantes rencontrent souvent des difficultés en matière de pertinence de l'information, de transparence, de segmentation de documents, d'intégration du contexte organisationnel et d'incertitude de l'information.

I.3.2 Le phénomène de l'incertitude dans la CCO

L'incertitude est un élément inhérent et omniprésent de la prise de décision, provenant d'une connaissance limitée où il est impossible de décrire de manière définitive une situation ou un

résultat [161]. Dans l'appariement professionnel, l'incertitude se manifeste à travers un langage ambigu, des interprétations subjectives [45], et des données incomplètes dans les OE et les CV. Comme mentionné dans une section précédente, cette incertitude ne concerne pas seulement les informations manquantes, mais aussi l'imprécision, le flou et l'ambiguïté au sein du langage humain lui-même.

Deux principales théories ont été proposées pour modéliser l'incertitude : la théorie des probabilités et la théorie des possibilités [153, 162]. La théorie des probabilités quantifie l'incertitude numériquement comme une probabilité mesurée sur un ensemble de résultats. Cependant, elle suppose un niveau de précision rarement atteignable avec les informations ambiguës et incomplètes du CV, de l'OE et de la CCO [163].

D'autre part, la théorie des possibilités, basée sur la théorie des ensembles flous, offre un cadre plus flexible pour gérer des données imprécises et ambiguës [164]. Elle permet une représentation qualitative de l'incertitude mais aussi une représentation quantitative de celle-ci, décrivant les possibilités sur une échelle de vraisemblance. Cette approche s'est avérée mieux alignée avec les scénarios affectés par l'incertitude, comme la phase de présélection dans l'appariement professionnel, où les décisions sont souvent prises sous des contraintes d'informations incomplètes et imprécises [165].

Bien que la théorie des probabilités est très populaire [153], la théorie des possibilités pourrait être mieux adaptée à la CCO, étant donné les ambiguïtés et les imprécisions fréquentes dans ce contexte. Cette approche correspond davantage à la manière flexible et nuancée dont les recruteurs abordent la correspondance. La théorie des possibilités est d'autant plus pertinente lorsqu'il s'agit de tâches sujettes à l'incertitude ou basées sur des données incomplètes, comme l'extraction d'informations clés des OE pour optimiser le classement des CV, garantissant ainsi que les éléments essentiels sont pris en compte sans être trop restrictifs.

En synthèse, dans le domaine complexe de la CCO, où l'incertitude reste un problème significatif, les méthodes existantes sont souvent insuffisantes pour prendre en compte de manière adéquate de telles variables, conduisant à des correspondances moins pertinentes et moins interprétables. Cela met en évidence une opportunité pour des recherches supplémentaires, notamment dans l'application nuancée de la théorie des possibilités. Son potentiel s'aligne bien avec les applications de recherche émergentes telles que le raisonnement approximatif et l'extraction d'informations à partir de documents contenant des informations incomplètes ou partielles comme les OE.

I.3.3 Apprentissage automatique et apprentissage profond à travers le prisme de la CCO

Comme mentionné dans les sections précédentes, dans le paysage actuel de l'IA, les modèles AA et AP affirment leur hégémonie dans le domaine de la CCO. Cependant, leur explicabilité

et leur interprétabilité posent des défis importants, en particulier dans des domaines où ces aspects ont été sporadiquement abordés, tels que la CCO. L'importance de l'explicabilité s'étend de la construction de la confiance dans les prédictions du modèle, cruciale dans les secteurs réglementés, à la garantie de l'équité en identifiant et en rectifiant les biais du modèle.

Malgré la puissance prédictive inégalée démontrée par les modèles complexes, en particulier les réseaux de neurones profonds, leurs structures labyrinthiques ont principalement fonctionné comme des boîtes noires impénétrables. Pour démystifier ces complexités, plusieurs techniques ont été conçues :

1. Les **explications indépendantes du modèle** comme LIME (Local Interpretable Model-Agnostic Explanations) [166] perturbent les données pour décrypter les variances de prédiction et ainsi déterminer la pertinence des caractéristiques.
2. Les **cadres théoriques de la théorie des jeux**, tels que SHAP (SHapley Additive exPlanations) [167], délimitent la contribution des caractéristiques pour chaque prédiction, en élucidant leur importance.
3. En AP, des outils tels que les **cartes de saillance** [168] et la **maximisation de l'activation** [169] s'adressent principalement aux données d'image, en mettant en lumière les segments influents ou en représentant les déclencheurs des neurones.
4. Les **gradients intégrés** [170] fournissent un plan pour analyser la divergence de prédiction entre une entrée et sa base de référence, en l'attribuant à des caractéristiques distinctes.
5. La **visualisation des caractéristiques** cartographie les domaines de données à haute dimension, tandis que les **explications contrefactuelles** [171] cherchent à rationaliser les décisions du modèle en élaborant des scénarios d'entrée alternatifs.

Néanmoins, des obstacles subsistent. Il y a un compromis inhérent entre la complexité du modèle et sa transparence. La nature nuancée de la cognition humaine complique davantage la question de ce qui est universellement considéré comme "interprétable". De plus, l'intensité computationnelle de plusieurs techniques demeure une barrière persistante [171].

Une voie plausible pour contourner ces défis, dans le cas de la CCO pourrait consister à exploiter des approches de CCO fondées sur des modèles interprétables plus simples et éprouvés dans le temps :

1. Les méthodologies basées sur la **théorie des possibilités et la logique floue** [125] pourraient aider à naviguer élégamment dans l'incertitude des données, en les traduisant en des sorties linguistiques plus compréhensibles pour les parties prenantes de la CCO.
2. Les **arbres de décision flous (ADF)** [172] étendent les arbres de décision traditionnels en permettant des frontières de décision plus nuancées, capturant ainsi les subtilités du raisonnement humain.

3. La **corrélation de Pearson** [173] fournit une métrique directe pour évaluer les relations linéaires entre les variables.
4. La **régression logistique** [174], en plus de ses mérites prédictifs, élucide l'ampleur de l'influence que chaque prédicteur exerce à travers ses coefficients.
5. L'**algorithme Apriori** [175] met en lumière les schémas de données prévalents en découvrant des ensembles fréquents.

Inhérente à cette conversation est la question de savoir si des modèles comme LIME seraient même nécessaires si les modèles fondamentaux de l'approche CCO étaient intrinsèquement explicables et ne nécessitaient pas d'architecture opaque. On pourrait considérer des paradigmes fondamentaux comme le raisonnement approximatif basé sur la théorie des possibilités, qui offre une représentation transparente des incertitudes et, par sa nature même, fournit des informations directes, rendant potentiellement inutile l'utilisation de techniques d'explicabilité *a posteriori* qui ne sont toujours pas suffisantes pour fournir des explications exhaustives sur les modèles opaques.

Cependant, des complexités de modélisation peuvent encore émerger :

1. Lorsque les modèles sont imbriqués, l'interprétabilité du système global peut s'atténuer, même si les éléments individuels sont clairs.
2. Les parties prenantes peuvent désirer des explications sous des formes spécifiques que ces modèles n'offrent pas intrinsèquement.
3. Des compromis entre précision et interprétabilité pourraient survenir.
4. Il est indispensable de veiller à ce que les explications soient alignées avec les processus de données fondamentaux.
5. Des informations sur des instances de décision particulières liées à des modèles opaques pourraient être souhaitées, un domaine où des techniques comme LIME ou SHAP pourraient devenir un complément.

En conclusion, à mesure que le domaine de la CCO continue de mûrir, la relation symbiotique entre interprétabilité et performance est apparue comme un pivot pour son avancement global. Le parcours de la CCO, entrelacé avec les avancées en AA et AP, souligne l'importance primordiale de l'explicabilité des modèles. Bien que les techniques *a posteriori* (cf. LIME) offrent un répit, des approches de CCO plus robustes pourraient résider dans l'exploitation de modèles fondamentaux explicables. Ces modèles, par conception, fusionnent de manière transparente la précision avec l'intérpretabilité et pourraient même offrir un cadre pour intégrer de manière plus responsable les modèles de "boîte noire" dans la CCO.

I.3.4 Diversité linguistique dans les OE et les CV : intersection de la linguistique, de l'analyse traditionnelle et des modèles transformateurs

Comme mentionné dans une section précédente, la diversité linguistique représente un défi significatif pour le traitement des CV et des OE, surtout lorsqu'on tient compte du marché du travail mondial actuel. Plusieurs facteurs contribuent à ce phénomène, y compris les différences culturelles, le jargon spécifique à l'industrie, les choix stylistiques individuels et les dialectes régionaux [176].

D'un point de vue linguistique, la diversité peut entraîner une multitude de problèmes. Par exemple, le même terme peut être utilisé différemment dans différentes régions, ou le même concept peut être décrit en utilisant des termes différents. Il y a également souvent des cas d'ambiguïté où un terme pourrait avoir plusieurs significations en fonction du contexte, ou un titre de poste identique peut renvoyer à des rôles radicalement différents selon les industries, voire les entreprises [177].

Des approches traditionnelles, telles que l'extraction automatique de termes (EAT), ont été utilisées pour aborder ces questions [178]. L'analyse terminologique implique l'étude systématique des termes, en particulier ceux utilisés dans un domaine ou un champ spécifique. Cela comprend le traitement des relations entre différents termes, l'analyse de leur utilisation, et le développement de vocabulaires normalisés. Ce processus peut aider à clarifier la signification des termes dans un contexte donné, à réduire l'ambiguïté et à favoriser la cohérence dans l'utilisation du langage.

Cependant, l'analyse terminologique a ses limites. Elle est laborieuse, nécessite une connaissance approfondie du domaine et, par conséquent, peut ne pas suivre le rythme de l'évolution constante de l'utilisation de la langue dans différentes industries [178]. De plus, elle pourrait ne pas capturer pleinement les nuances et la complexité du langage naturel, en particulier en ce qui concerne les particularités du jargon franglais de l'informatique [179]. Par exemple, dans le domaine informatique, ce jargon pourrait utiliser le terme « **containér**isation » pour évoquer l'emploi de conteneurs (containers) Docker [6]. En français, cette expression est plus proche du mot « conteneurisation » qui est davantage associé à l'utilisation de conteneurs dans le secteur du transport de marchandises [7].

Inversement, les modèles transformateurs actuels utilisés en AA, comme BERT et GPT, offrent une solution puissante à la diversité linguistique [180]. Ces modèles analysent le contexte des mots, aidant ainsi à gérer les variantes de termes, à capturer les idiomes et même à identifier des indices subtils de sentiment humain dans le texte. Cependant, ils ont également des limites, telles que des besoins importants en données et un possible désalignement avec les besoins spécifiques du contexte organisationnel lorsqu'ils sont confrontés à un terme inconnu, phénomène

6. `https://www.docker.com/`
7. `https://www.larousse.fr/dictionnaires/francais/conteneurisation/18566`

fréquemment observé dans le domaine de la CCO.

À la lumière des défis et des approches ci-dessus, un domaine de recherche émergent qui semble prometteur est l'extraction de la terminologie et la similarité des variantes terminologiques basées sur la sémantique distributionnelle [13]. Cette approche utilise les propriétés statistiques de grands corpus de texte pour représenter les similarités contextuelles entre différentes variantes de termes, abordant ainsi les problèmes de synonymie, d'antonymie, d'hyponymie, entre autres. Ce faisant, ce type d'approche pourrait aider à identifier les termes pertinents et leurs variantes dans les CV et les OE, améliorant ainsi la précision de la représentation sémantique [181].

En conclusion, le paysage émergent de l'EAT offre des opportunités significatives pour la recherche interdisciplinaire à l'intersection de la linguistique, de l'AA et de la CCO. Étant donné les défis complexes posés par la diversité linguistique dans les CV et les OE, l'EAT a le potentiel de compléter les méthodologies existantes en gérant l'incertitude inhérente à des tâches telles que l'extraction d'informations à partir des documents. Lorsqu'il est intégré à des modèles GML comme BERT, les capacités de l'EAT pourraient être davantage améliorées, enrichissant les traitements sémantique et contextuel du texte, ainsi que des sources de connaissances telles que les ontologies de compétences professionnelles. Une telle intégration pourrait ouvrir la voie à des systèmes de CCO plus robustes, capables de s'adapter aux dynamiques en rapide évolution des marchés du travail et de l'utilisation du langage, offrant finalement une approche plus précise, évolutif et explicable de la CCO.

I.3.5 Analyse approfondie et optimisation de l'extraction d'informations à partir des OE

Comme mentionné dans les sections précédentes, l'un des principaux défis de la CCO est l'extraction d'informations pertinentes à partir des OE. Ces informations déterminent souvent les critères de sélection utilisés dans les processus de correspondance de l'emploi. Par conséquent, leur identification et extraction précises sont critiques pour l'efficacité de la méthode CCO.

Dans ce contexte, la phase de présélection en recrutement, telle qu'elle est esquissée dans ce manuscrit, est une étape où les recruteurs font correspondre les candidats au contenu d'une OE. Bien que les méthodes traditionnelles d'AA aient été employées pour automatiser cette tâche, elles présentent souvent une limitation dans l'identification précise des informations essentielles au sein des OE. Fréquemment, ces méthodes se concentrent sur des informations de fond qui ne sont pas pertinentes pour l'OE spécifique, entraînant une dégradation du classement des candidats. Ce problème est exacerbé par l'opacité des systèmes basés sur l'AP, qui compromet l'explicabilité.

Ces défis sont d'autant plus accentués lorsque l'on prend en considération deux problèmes substantiels qui persistent. Premièrement, la différence de taille entre un CV et une OE peut

induire le système en erreur, attribuant une grande pertinence aux candidats sur la base d'informations non essentielles de l'OE. Deuxièmement, les OE dépeignent souvent un profil de candidat idéal qui est peu susceptible de correspondre parfaitement à un seul CV, soulignant ainsi le besoin d'identifier avec précision les exigences professionnelles les plus primordiales [156].

Compte tenu de ces défis et limitations, il devient central d'identifier de manière plus précise et transparente les 'termes pertinents' dans les OE, surtout lorsqu'on considère les termes que les recruteurs s'attendent à trouver dans les CV des candidats idéaux. Bien que la terminologie puisse différer entre les CV et les OE, les sémantiques sous-jacentes partagent souvent un réseau de similitudes. Un cadre basé sur l'ontologie, étendu avec des modèles de sémantique distributionnelle comme BERT peut combler les écarts sémantiques, fournissant un outil robuste pour évaluer les similitudes entre ces documents [34].

Cela conduit à l'idée d'utiliser des 'marqueurs textuels'. Dans notre contexte, les marqueurs servent d'indicateurs sémantiques de pertinence au sein d'une OE et peuvent aider à automatiser l'extraction d'informations pertinentes. En tenant compte du contexte organisationnel tel qu'il est compris par les recruteurs, nous suggérons que l'efficacité des marqueurs textuels peut être optimisée, améliorant potentiellement à la fois le processus d'extraction d'informations et l'explicabilité de la CCO. Toutefois, optimiser les marqueurs textuels pour extraire les informations pertinentes des OE demeure un défi significatif [182]. Les complexités et les ambiguïtés inhérentes aux OE nécessitent encore une approche sophistiquée pour identifier avec précision les marqueurs textuels les plus pertinents pour un classement efficace des CV [45, 24].

Pour conclure, les limitations mentionnées dans cette section mettent en évidence le besoin d'une méthode robuste d'optimisation de marqueurs textuels de l'OE. Une telle méthode devrait tenir compte de facteurs clés comme la gestion des ambiguïtés, l'alignement avec l'expertise du recruteur, le maintien de la pertinence de l'information et une approche contextuelle et explicable.

I.3.6 Approches existantes d'optimisation d'extraction de l'information à partir des OE

Dans le domaine de l'optimisation de l'extraction d'informations à partir de documents, diverses méthodologies ont été mises en avant, à savoir les techniques wrapper, de filtrage, intégrées et floues [183]. Ces méthodologies existantes offrent des perspectives précieuses pour l'optimisation des caractéristiques significatives à partir du texte. Toutefois, elles présentent une limitation importante : aucune d'entre elles n'aborde explicitement la sélection optimisée de marqueurs textuels de haute qualité au sein des OE. Cette étape de sélection est pourtant critique pour la qualité de la CCO [57].

Aperçu des techniques wrapper

Les techniques wrapper impliquent des approches qui utilisent des modèles d'AA pour évaluer l'importance des sous-ensembles de caractéristiques [183]. En s'entraînant sur différents ensembles de caractéristiques et en utilisant des ensembles de données de validation, elles visent à identifier le sous-ensemble optimal. Une méthode avancée d'optimisation par essaim de particules adaptée à la sélection de caractéristiques textuelles a été mise en évidence [184]. De même, une stratégie robuste de sélection de caractéristiques qui intègre le tfidf avec le SVM-RFE (Support Vector Machines Recursive Feature Elimination) a été conçue pour des tâches d'analyse de sentiment [185]. Un algorithme d'optimisation exploitant la métaheuristique Iterated Greedy, enrichie de scores de filtre prétraités, a été introduit [186]. D'autres stratégies wrapper notables incluent les techniques génétiques, séquentielles et basées sur l'élimination [187].

Les techniques de filtrage

Au contraire, les techniques de filtrage priorisent les caractéristiques en fonction de leurs propriétés inhérentes, éliminant ainsi la dépendance à un modèle d'AA, ce qui se traduit par une efficacité accrue. Ces techniques emploient fréquemment des mesures statistiques telles que le Chi-carré [88] [188], l'indice de Gini [189], le coefficient de corrélation de Pearson [173], la précision [190], le gain d'information [191] [192], et l'entropie d'information mutuelle [193]. Certaines méthodologies, comme l'ambiguïté [172], mettent l'accent sur des mesures de prise de décision centrées sur l'humain.

Des avancées modernes ont donné naissance à des stratégies de filtrage évoluées. Le coefficient d'information maximal [194] se distingue en identifiant à la fois les relations linéaires et complexes entre les caractéristiques et la cible. Des techniques composées comme le filtre basé sur la corrélation [195] évaluent les corrélations entre caractéristiques et déterminent le meilleur sous-ensemble en utilisant un mélange de sélection arrière et de recherche séquentielle.

Aperçus des techniques intégrées

Les stratégies intégrées créent une synergie entre les atouts des techniques de filtrage et les techniques wrapper. En fusionnant la sélection de caractéristiques avec le processus AA, elles traitent l'optimisation des caractéristiques et l'entraînement du modèle. Un exemple classique est l'opérateur de rétrécissement et de sélection des moindres absolus (LASSO, Least Absolute Shrinkage and Selection Operator), qui met en évidence les caractéristiques essentielles et filtre les redondances grâce à la régularisation L1 [196]. Des techniques telles que les arbres de décision [172] (et plus largement, les forêts aléatoires) privilégient intrinsèquement les caractéristiques saillantes pour la division des données au niveau des nœuds de l'arbre. Certaines méthodologies uniques intègrent diverses techniques comme l'analyse en composantes principales (ACP),

l'analyse sémantique latente (ASL) et la projection aléatoire (PA) [197].

Exploration des techniques floues

Des développements récents ont vu l'essor des méthodes de sélection de caractéristiques floues, bénéfiques dans des scénarios marqués par des incertitudes de classification significatives. La stratégie de corrélation de caractéristiques floues intègre le gain d'information pour améliorer l'identification de la pertinence des caractéristiques [198]. La technique floue-approximative a été adoptée pour la réduction de la dimensionalité dans le texte, facilitant un traitement rapide des documents et abordant l'ambiguïté de classification [199]. Une stratégie de marqueur flou hybride sophistiquée, ancrée sur des systèmes basés sur des règles floues de type Mamdani, a montré sa supériorité sur les techniques de filtrage traditionnelles [200].

En conclusion, l'optimisation des marqueurs textuels dans les OE demeure un domaine difficile en raison des complexités et des ambiguïtés inhérentes à ces documents. Les méthodologies actuelles, bien que précieuses, négligent souvent les nuances spécifiques essentielles pour une optimisation efficace des marqueurs dans les OE. Dans le domaine de la CCO, il y a un besoin marqué d'une méthodologie qui maîtrise pleinement les subtilités des OE, en mettant l'accent sur l'évaluation de la qualité, l'intégration des perspectives des recruteurs, l'explicabilité et l'adaptabilité aux contextes organisationnels en évolution. Ces domaines mis en évidence non seulement soulignent des lacunes importantes dans le domaine, mais pointent également vers des solutions CCO évaluées plus en rigueur.

I.3.7 Aborder le format non structuré actuel des CV : une préoccupation-clé dans la CCO

Au-delà de l'optimisation de l'extraction et de l'interprétation des marqueurs textuels des OE, une autre facette critique de la CCO concerne les CV. Malgré la richesse des informations qu'ils fournissent, l'interprétation des CV est entravée par leur nature intrinsèquement non structurée et graphique. Cette caractéristique nécessite des méthodologies plus nuancées pour leur représentation et des traitements adéquats, pour une CCO plus efficace.

La tâche de structuration des CV pour une interprétation automatisée implique principalement de les segmenter en sections distinctes. De nombreuses techniques ont été développées pour faciliter cela, y compris des heuristiques orientées règles, le marquage sémantique, les plongements de mots et les modèles de transformateurs [201, 202, 34, 203, 204, 205, 206]. Le modèle BERT a également été utilisé pour améliorer les tâches de structuration des CV, y compris l'extraction d'entités [207, 208].

Malgré ces efforts, les méthodologies actuelles se concentrent principalement sur le traitement des informations textuelles trouvées dans les CV. Elles négligent souvent les éléments non structurés et graphiques inhérents à ces documents. Ce manque d'intérêt représente un obstacle

significatif, limitant la portée de la CCO et d'autres processus d'analyse automatique de CV connexes [45]. La nature hautement individualisée et pourtant mimétique des CV en fait une source cruciale d'informations personnelles et professionnelles sur les candidats [66, 209, 210].

Ces caractéristiques particulières des CV soulignent la nécessité de nouvelles approches pour gérer, représenter et incorporer leurs composants graphiques. Certains efforts ont déjà été faits pour intégrer des caractéristiques non textuelles dans BERT. Certaines méthodes ont ajouté des couches de réseaux neuronaux pour représenter la structure du document [211], tandis que d'autres ont incorporé des caractéristiques supplémentaires directement dans le texte [212]. Malheureusement, les processus de segmentation pour les CV n'ont pas encore pleinement adopté de telles représentations.

Il est intéressant de souligner que les domaines d'étude plus larges comme la grapholinguistique, qui étudie le langage naturel écrit, restent largement inexploités dans la recherche sur l'analyse automatisée des CV [66]. Malgré cela, ce domaine de connaissances offre des perspectives significatives pour améliorer ces processus.

Pour plus de clarté, définissons certains termes et concepts fondamentaux liés à l'étude du langage écrit. Les fondements de la grapholinguistique sont construits sur la notion de graphème, similaires à celle de phonème en phonologie, qui est l'unité distinctive la plus petite du texte écrit [66, p. 119]. La discipline qui étudie les graphèmes est appelée graphématique. La réalisation matérielle d'un graphème est appelée graphe [8] [66, p. 63]. L'étude scientifique des graphes est appelée graphétique.

Le langage écrit sous sa forme la plus élémentaire se manifeste comme une séquence graphémique unidimensionnelle (1-dim SG), représentée par une ligne de graphes. Cette séquence peut être horizontale comme en anglais ou verticale comme en chinois, japonais, coréen et mongole. Cependant, les contraintes spatiales sur une page de document nécessitent la division de ces séquences en segments plus petits, donnant naissance à des séquences graphémiques bidimensionnelles (2-dim SG) disposées en plusieurs lignes sur une page.

Une approche grapholinguistique va au-delà de l'analyse des graphèmes dans ces séquences pour inclure leur arrangement et les indices visuels qui les accompagnent. Ce paradigme englobe des considérations comme la distribution spatiale, la mise en page, la typographie et d'autres éléments qui ajoutent des couches supplémentaires de signification et de contexte au texte écrit [213, chap. 1]. Par conséquent, appliquer une approche grapholinguistique aux CV pourrait fournir de nouvelles perspectives sur leurs éléments graphiques, améliorant l'analyse, l'interprétation et la robustesse de la CCO [214].

En conclusion, la nature non structurée et graphique des CV présente toujours un défi notable pour leur interprétation étendue et les processus de correspondance d'emploi subséquents. Alors

8. Le terme «graphe» dans ce contexte n'est pas à confondre avec le graphe mathématique qui est un ensemble de relations binaires. En grapholinguistique, le terme «graphe» a été forgé par analogie avec le terme «phone» en phonétique (graphe/graphétique/graphème, comme phone/phonétique/phonème).

que le domaine a fait des progrès considérables dans l'interprétation textuelle, il a largement négligé les composants graphiques et le format non structuré, limitant ainsi la complétude et la précision de son analyse automatisée. Cette prise de conscience nous invite à nous concentrer sur le développement de méthodologies qui peuvent efficacement gérer, représenter et incorporer à la fois les aspects textuels et graphiques de ces documents dans la CCO.

I.4 Conclusion

L'état actuel de la CCO montre des progrès substantiels, exploitant diverses techniques telles qu'AA, l'AP, les méthodes basées sur les ontologies et les systèmes basés sur IA pour automatiser et optimiser le processus de correspondance entre le CV et l'OE. Malgré ces avancées, plusieurs défis clés demeurent :

- **pertinence des informations dans les OE** : il existe un manque de méthodologies capables d'analyser, d'expliquer et de modéliser la pertinence des informations dans les OE, surtout du point de vue des recruteurs qui, de par leur métier, possèdent une connaissance profonde de ces documents ;
- **intégration du contexte organisationnel** : l'intégration du contexte organisationnel dans les méthodes de CCO est largement négligée, ce qui conduit à un désalignement et à une dégradation progressive de ces systèmes au fil du temps ;
- **pertinence et explicabilité des marqueurs textuels** : les méthodologies existantes sont insuffisantes pour évaluer la pertinence des marqueurs textuels d'une OE par rapport au contexte organisationnel dans lequel le système de CCO fonctionne ; ceci est essentiel pour garantir la pertinence et l'alignement de la méthode avec les exigences spécifiques des organisations ;
- **approche de segmentation généralisée** : la segmentation automatique de documents, tels que les CV, demeure un problème persistant ; les systèmes actuels utilisent généralement des formats de CV prévisibles ou s'appuient sur des formats standardisés qui sont traités manuellement ;
- **incertitude de l'information dans la CCO** : la représentation du phénomène systématique d'incertitude de l'information a été largement ignorée d'un point de vue théorique et expérimental, en particulier dans les OE, qui sont généralement des documents contenant des informations incomplètes ou partielles.

En conclusion, le domaine de la CCO présente un champ riche en potentiel et en complexité. L'état actuel révèle des avancées significatives dans l'exploitation des approches basées sur l'IA. Cependant, des défis substantiels demeurent dans le traitement des données non structurées et l'incertitude de l'information. Ces limitations s'étendent à un manque de transparence, d'explicabilité, d'intégration du contexte organisationnel et à la nécessité de s'adapter aux exigences

en constante évolution au sein du processus de recrutement. De plus, les grandes quantités de données nécessaires aux approches contemporaines rendent leur mise en œuvre et leur efficacité complexes. En raison de cette exigence en termes de données, les petites et moyennes organisations, qui n'ont généralement pas accès à de tels volumes, peinent souvent à exploiter ces méthodes.

Pour relever ces défis, des approches plus équilibrées, innovantes et hybrides seront probablement la clé pour débloquer des processus de CCO plus efficaces, précis, fiables et satisfaisants. La synthèse des limitations et avancées existantes forme une base solide pour la recherche actuelle de cette thèse. Cette recherche vise à construire une approche IA plus intelligible, pragmatique et explicable pour la CCO, reflétant une synthèse des avancées technologiques et du rôle que l'humain a à jouer pour accomplir cette tâche nuancée et complexe.

Chapitre II

Méthodologie pour représenter les OE et les CV dans le contexte de la CCO

Ce chapitre introduit une méthodologie conçue pour élaborer une représentation plus complète des OE et des CV en vue de la CCO. Le modèle proposé vise non seulement à intégrer les nuances de la dynamique organisationnelle, mais aussi à aborder les incertitudes inhérentes au traitement automatisé.

Les composants clés de cette méthodologie sont :

- analyse et représentation du contexte organisationnel dans lequel les OE et les CV sont traités ;
- analyse et modélisation de l'incertitude associée au traitement automatique des OE et des CV ;
- construction d'un outil de représentation des connaissances intégré avec le savoir organisationnel ;
- extraction et représentation du texte des OE : modélisation axée sur l'analyse des stratégies de recruteurs associées à la sélection d'informations pertinentes des OE pour améliorer le classement des CV ;
- extraction et représentation du texte des CV : modélisation basée sur l'analyse des aspects grapholinguistiques des CV associés aux stratégies de lecture des recruteurs ;
- annotation sémantique des OE et des CV à partir de sources de connaissances organisationnelles.

Ces composants constituent la base d'une méthodologie de CCO améliorée, axée sur une analyse approfondie et une identification rigoureuse des exigences essentielles énoncées dans les OE, ainsi que sur le repérage des informations saillantes dans les CV des candidats.

Dans les sections suivantes, les aspects de modélisation parallèles liés au traitement des OE et des CV seront introduits. Ceux-ci incluent la représentation de l'incertitude, la représentation du contexte organisationnel, l'extraction de l'expertise des recruteurs et la construction de sources de connaissances. Ensuite, les principaux éléments spécifiques associés à la modélisation des OE

et à la modélisation grapholinguistique des CV seront présentés.

L'objectif est de fournir une perspective intégrée du processus, démontrant la capacité de notre approche à traiter la complexité de l'extraction d'informations à partir de ces documents. Chaque phase de modélisation est essentielle pour saisir le caractère distinctif de chaque OE et CV, visant à un fondement de CCO plus robuste.

II.1 Définitions contextuelles

Cette section présente les concepts plus spécifiques utilisés pour la méthodologie :

— **contexte organisationnel** : l'environnement interne où fonctionne une organisation, englobant des aspects culturels, sociaux, politiques, économiques, technologiques et physiques, ainsi que la structure organisationnelle, les ressources, les objectifs et les interactions ;

— **contexte sociétal** : plus large que le contexte organisationnel, le contexte sociétal est défini comme une couche socio-culturelle de l'environnement complexe dans lequel tous les systèmes techniques et les acteurs sociaux qui les créent et en sont affectés, existent et interagissent ;

— **incertitude** : l'absence de certitude ou la présence de doute sur un événement ou un état de choses particulier ; l'incertitude est représentée par des degrés de possibilité et de nécessité, qui sont duales ;

— **possibilité** : une mesure de la faisabilité d'un événement ou d'un état de choses ; elle est souvent représentée par une fonction d'appartenance floue qui attribue un degré de possibilité entre 0 (impossible) et 1 (complètement certain) à chaque élément d'un ensemble donné ;

— **ontologie-mère** : une ontologie conçue pour être largement réutilisable et capable de servir de base pour intégrer d'autres ontologies plus spécifiques ;

— **terminologie** : l'étude des termes et de leur utilisation ; les termes sont des syntagmes qui représentent de manière non-ambiguë les concepts et les relations d'un domaine de connaissances, à l'occurence celui de la CCO ;

— **marqueur textuel** : une caractéristique textuelle aidant les modèles AA à identifier les informations essentielles d'un texte, en tenant compte également des contextes sociétaux/organisationnels ;

— **croyance-désir-intention** : le paradigme CDI, servant à modéliser des aspects du raisonnement, trouve une application notable dans les systèmes basés sur des agents pour la représentation et la simulation des comportements humains. Ce paradigme se décompose en trois composants clés, chacun jouant un rôle distinct dans la modélisation du raisonnement et des comportements :

- **croyance (belief)** : informations que l'agent détient sur le monde. Les croyances peuvent être graduellement vraies ou fausses, et l'agent peut les mettre à jour lorsqu'il reçoit de nouvelles informations ;
- **désir (desire)** : états du monde dont l'agent planifie la réalisation ; les désirs représentent les objectifs de l'agent et ne sont pas nécessairement réalisables ;
- **intention** : actions que l'agent a décidé d'entreprendre pour atteindre ses désirs ; les intentions sont le résultat d'un processus de délibération basé sur les croyances et les désirs de l'agent, souvent structuré sous forme de règles logiques ;

— **information contextuelle de la compétence professionnelle dans le CV :** désigne des éléments-clés inscrits dans un CV, liés à la compétence professionnelle du candidat, qui captent l'intérêt des recruteurs ; ces données, telles que la section du CV où une compétence est mentionnée, les années d'expérience liées à son utilisation, ou les certifications associées, guident les recruteurs dans l'évaluation de la pertinence des compétences du candidat ; ces critères, identifiables lors des entretiens avec les recruteurs, contribuent de manière descriptive à la représentation du contexte organisationnel.

II.2 Principes de base

Nous présentons ici les principes de base de la méthodologie proposée :

— **information incomplète** : dans le contexte de la théorie des possibilités, le Principe de l'Information Incomplète postule qu'en l'absence de preuves ou de données suffisantes, aucune alternative au sein d'un ensemble de possibilités ne peut être définitivement exclue. Ce principe sert d'axiome fondamental pour établir des distributions de possibilités, en prenant en compte à la fois l'incertitude épistémique (incertitude sur la connaissance) et l'ignorance ;

— **appartenance graduelle** : la vérité et la fausseté des croyances sont représentées comme des degrés d'appartenance à des ensembles flous, permettant la modélisation de concepts imprécis, ambigus ou vagues ;

— **émergence** : les comportements complexes d'un système organisationnel émergent des interactions entre des composants plus simples, et ces comportements sont souvent imprévisibles et irréductibles aux propriétés des composants individuels ;

— **représentation du contexte organisationnel** : en accord avec le principe de l'émergence, il est essentiel de comprendre et de représenter le contexte organisationnel du traitement des CV et des OE. Reconnaître l'interaction complexe entre les éléments organisationnels est crucial, car ceux-ci entraînent des comportements émergents dans la gestion des CV et des OE. Ce point de vue favorise une conception de la CCO qui est à la fois adaptée aux besoins des organisations et suffisamment flexible pour gérer les défis

et les changements organisationnels imprévus ;

— **agence rationnelle** : le comportement intelligent des agents d'un système (par exemple, les recruteurs) découle d'un processus rationnel de sélection des actions (intentions) en fonction de leurs objectifs (désirs) et des informations disponibles (croyances) du monde ;

— **incertitude inhérente à la CCO** : les CV et les OE sont des documents aux structures complexes et incertaines. Pour comprendre et traiter la complexité associée à leur traitement automatique, il est fondamental d'identifier et de représenter les sources d'incertitude liées à leur structure et à leur contenu ;

— **fonctionnalité graphique dans les systèmes d'écriture** : les éléments d'un système d'écriture sont bien plus que de simples représentations du contenu phonologique ou sémantique. Ce sont des unités fonctionnelles qui influencent significativement le traitement cognitif, la compréhension et l'évaluation du texte écrit (par exemple, les CV). Comprendre ces éléments est essentiel pour optimiser la transmission et la réception de l'information au sein d'un système de CCO.

II.3 Description générale

Dans la Figure II.1, nous présentons le schéma général de la méthodologie proposée, visant à créer une représentation plus robuste des CV et des OE dans la CCO.

La méthodologie élaborée commence par la représentation du contexte organisationnel comme étape prioritaire pour dériver une ontologie-mère. Cette ontologie sert de base pour construire une représentation contextualisée des CV et des OE. En utilisant cette ontologie, nous effectuons des processus d'extraction d'expertise avec les recruteurs pour recueillir et exploiter leur compréhension de ce qui est le plus pertinent pendant les processus de recrutement. Cela donne lieu à des ensembles de données qui représentent les connaissances des experts du contexte organisationnel.

À ce stade, un prétraitement et une normalisation du corpus de documents étudié sont effectués, avec un accent particulier sur l'extraction de la terminologie de chaque document. Cette terminologie extraite sert alors de pierre angulaire pour dériver les marqueurs textuels du corpus. Pour les OE, les marqueurs textuels sont dérivés en fonction des stratégies des recruteurs liées à la pertinence de l'information, améliorant ainsi l'extraction automatique des informations les plus essentielles dans ces documents. Dans le cas des CV, des marqueurs grapholinguistiques sont dérivés pour optimiser une étape requise pour l'analyse automatisé du document : la segmentation.

Après avoir dérivé ces marqueurs associés aux points de vue des recruteurs, une évaluation de leur pertinence est effectuée. Pour les OE, un moteur d'inférence floue est utilisé, en employant des métriques de qualité pour évaluer la pertinence des marqueurs. Pour les CV, une régression

logistique classique (RLC) est appliquée pour évaluer la pertinence statistique des marqueurs dans l'identification des coordonnées de segmentation optimales pour le document.

Les marqueurs les plus efficaces pour les OE sont ensuite adaptés à une architecture CDI possibiliste pour extraire les informations les plus pertinentes pour le processus de CCO. Les marqueurs grapholinguistiques des CV sont intégrés dans une architecture basée sur BERT afin de la rendre sensible au format des CV, dans le but d'optimiser leur segmentation et leur annotation sémantique ultérieure.

Les résultats de ces phases permettent la dérivation d'une représentation de documents plus robuste et sensible au contexte, facilitant leur annotation sémantique pour la phase de CCO finale.

II.4 Représentation de l'incertitude

L'incertitude est un défi omniprésent dans le domaine de la CCO, en particulier lorsqu'il s'agit de CV et d'OE. La méthodologie proposée dans cette thèse vise à relever ce défi en abordant plusieurs types d'incertitude grâce à une approche multifacette. Cette section explore les types d'incertitude identifiés et les stratégies employées pour les gérer.

Incertitude linguistique

Les CV et les OE contiennent souvent des informations intégrées dans des textes en langage naturel. L'ambiguïté et l'imprécision inhérentes à ces textes peuvent entraîner une incertitude dans leur traitement automatique.

La méthodologie utilise une approche possibiliste basée sur la représentation terminologique des documents. Cette approche reconnaît et représente explicitement l'incertitude inhérente à l'analyse et à l'interprétation des textes.

Incertitude contextuelle

Cela englobe à la fois les contextes organisationnels et sociétaux, qui peuvent introduire de la variabilité et de la complexité dans le processus de recrutement.

La méthodologie intègre la représentation du contexte organisationnel et de son contexte sociétal local pour mieux comprendre les dynamiques associées au traitement des CV et des OE.

Incertitude cognitive

Cette incertitude découle de l'application de marqueurs textuels dérivés des points de vue des recruteurs sur la pertinence de l'information, capturant le degré d'incertitude inhérent à l'utilisation de ces marqueurs pour interpréter et évaluer l'information.

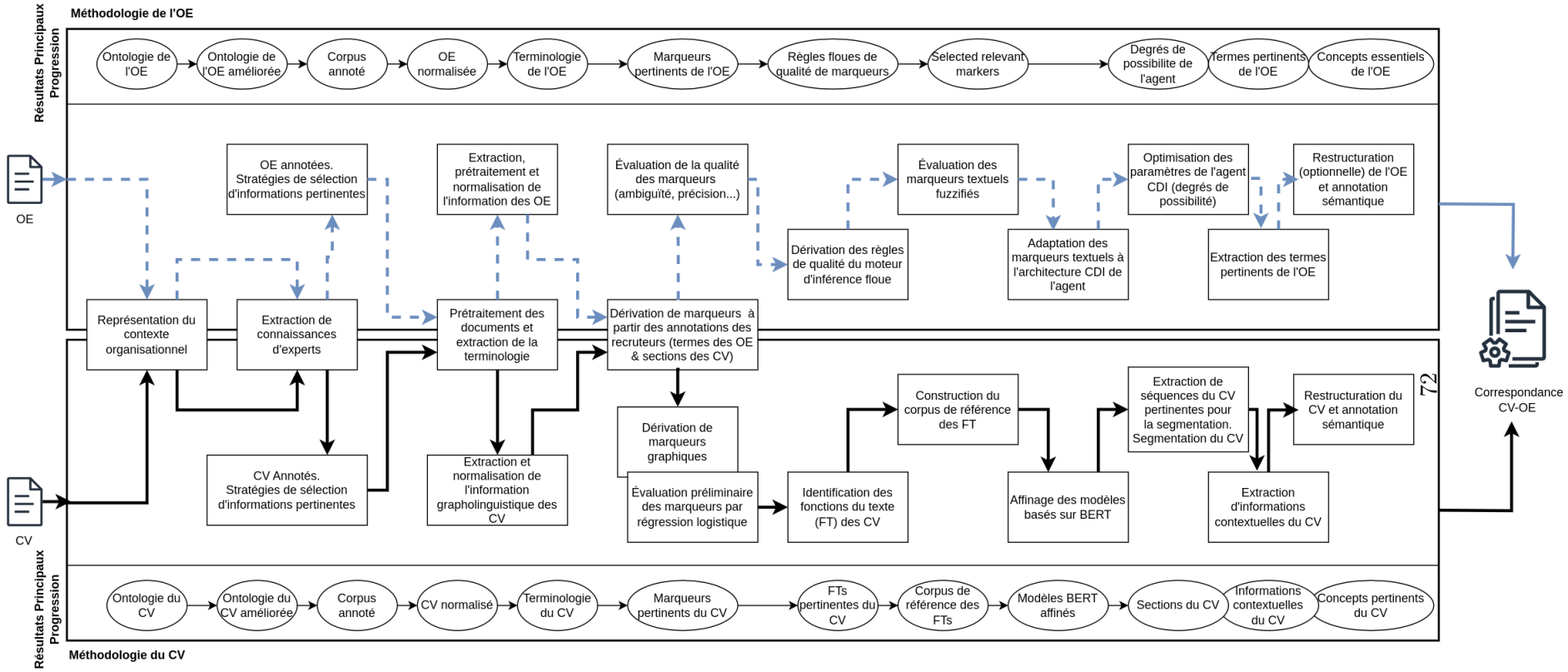


FIGURE II.1 – Schéma général de la méthodologie proposée pour une représentation plus robuste du CV et de l'OE dans le contexte de la CCO.

Une estimation de l'ambiguïté est employée pour quantifier ce type d'incertitude, capturant ainsi le degré d'incertitude inhérent à l'utilisation des marqueurs textuels.

Ambiguïté et entropie des marqueurs

Les marqueurs sont évalués en matière d'incertitude. Cette évaluation s'inscrit dans un cadre mesurant la qualité, qui prend en compte l'ambiguïté et l'entropie de l'information au sein d'un contexte organisationnel spécifique. L'incertitude est examinée à la fois qualitativement (représentation claire ou non des stratégies des recruteurs, ambiguïté) et quantitativement (degré auquel la connaissance apportée par les marqueurs réduit l'incertitude associée aux points de vue des recruteurs, entropie de l'information mutuelle).

Pour améliorer la clarté et l'interprétabilité des marqueurs textuels dans leur contexte de déploiement organisationnel, notre méthodologie intègre un moteur d'inférence de type Mamdani (voir l'annexe I). Ce moteur utilise des règles de logique floue pour définir les critères de qualité essentiels garantissant la pertinence des marqueurs.

Mécanisme de propagation de l'incertitude dans l'extraction de la pertinence de l'information

Au-delà de l'intégration de mécanismes pour capter et réduire l'incertitude, il est également important de modéliser la propagation de celle-ci pendant le traitement automatique de l'information [2]. Pour cela, un opérateur de modification de croyance (annexe E) est défini pour représenter la propagation de l'incertitude pendant l'extraction de l'information. Cet opérateur permet d'intégrer de nouvelles informations aux croyances de l'agent en fonction de leur niveau de confiance et d'incertitude associés. Ceci contribue à améliorer la flexibilité et l'explicabilité de la méthodologie, en ajoutant également un indicateur d'incertitude pour les utilisateurs qui lisent, analysent et exploitent les résultats et les données fournis de la solution CCO.

En résumé, en employant une approche multifacette englobant des méthodes possibilistes, une modélisation contextuelle, des mesures d'ambiguïté et une inférence axée sur l'incertitude, la méthodologie vise à améliorer la robustesse du traitement automatique des CV et des OE, fournissant un cadre plus large pour aborder les incertitudes inhérentes à la CCO.

II.5 Représentation du contexte organisationnel

La méthodologie proposée reconnaît que des processus comme la CCO sont intrinsèquement liés à un contexte organisationnel unique, généralement une entreprise. Ce contexte est façonné par des facteurs tels que la culture de l'entreprise, la structure organisationnelle, la stratégie commerciale et les objectifs spécifiques de chaque organisation. Par conséquent, nous proposons de caractériser ce contexte, sur la base d'une définition adaptée du contexte organisationnel et

sociétal [44]. Cette définition nous guide pour identifier les principaux acteurs des organisations, les artefacts, les actions des acteurs sur ces artefacts et les processus commerciaux impliqués. Notre approche vise à analyser et représenter le contexte organisationnel en détail, sur la base de la méthode UNC [107]. Les sous-sections suivantes détailleront la construction de cette représentation.

Réalisation d'entretiens préliminaires avec les recruteurs

Pour mieux comprendre le contexte organisationnel entourant les OE et les CV, nous réalisons des entretiens préliminaires avec les recruteurs. Ces entretiens sont essentiels pour recueillir des informations de première main sur l'organisation et ses processus de recrutement. Les recruteurs, en tant qu'acteurs centraux de ce processus, fournissent des informations précieuses sur les nuances du processus de recrutement spécifiques à leur organisation.

Ces entretiens nous aident à construire un vocabulaire commun, des schémas préconceptuels, des représentations graphiques intuitives qui nous permettent d'identifier et de modéliser des concepts-clés et leurs relations. Il s'agit d'une première étape vers la structuration du contexte organisationnel. En utilisant ces schémas, nous créons des modèles de domaine plus détaillés qui présentent une vue plus structurée des concepts.

Création de schémas préconceptuels et dérivation d'une ontologie-mère

Comme mentionné dans la section précédente, nous formulons des schémas préconceptuels et des modèles de domaine, qui servent de base au développement d'une ontologie-mère. Dans notre contexte, une ontologie-mère est une structure conceptuelle rigoureusement organisée qui regroupe un ensemble de concepts liés aux OE, aux CV et au contexte organisationnel spécifique dans lequel ces documents sont analysés. Cette ontologie-mère est centrale à notre approche, consolidant des concepts clés pour un traitement sémantique automatisé des documents adapté aux besoins spécifiques de l'organisation.

Identification des objectifs organisationnels des recruteurs

En parallèle, nous travaillons à l'identification des objectifs des recruteurs liés au cycle de vie des OE et des CV au sein des processus de recrutement. Nous faisons cela en examinant les modèles et en écoutant attentivement les recruteurs pendant les entretiens. Ces objectifs sont ensuite organisés hiérarchiquement, reflétant des objectifs organisationnels de haut niveau ainsi que des objectifs plus spécifiques liés à chaque étape du processus de recrutement.

Diagrammes de processus

Pour représenter visuellement le processus de recrutement, nous utilisons des diagrammes de processus. Ces diagrammes montrent clairement les différentes étapes du processus, les interactions entre ces étapes, des éventuels autres processus organisationnels concernés et les acteurs impliqués. En visualisant le processus, nous pouvons identifier les zones de friction ou d'inefficacité et comprendre comment chaque étape contribue à atteindre les objectifs de recrutement.

Le diagramme en arêtes de poisson

Nous utilisons un diagramme en arête de poissons pour élucider les relations entre les problèmes organisationnels liés aux documents et à leurs causes racines. Outil souvent trouvé dans la gestion de la qualité, ce diagramme aide à identifier et à comprendre les problèmes fondamentaux pour élaborer des solutions de CCO plus efficaces. Il met en lumière à la fois les symptômes et les causes fondamentales nécessitant une rectification.

Par exemple, si des retards prolongés dans le processus de recrutement affectent l'organisation, le diagramme en arêtes de poisson pourrait révéler des causes comme un processus de sélection inefficace, des pénuries de personnel, des lacunes dans la communication, des descriptions de poste peu claires ou des systèmes de CCO opaques. Une schématisation générale de cet exemple est fourni dans la Figure II.2.

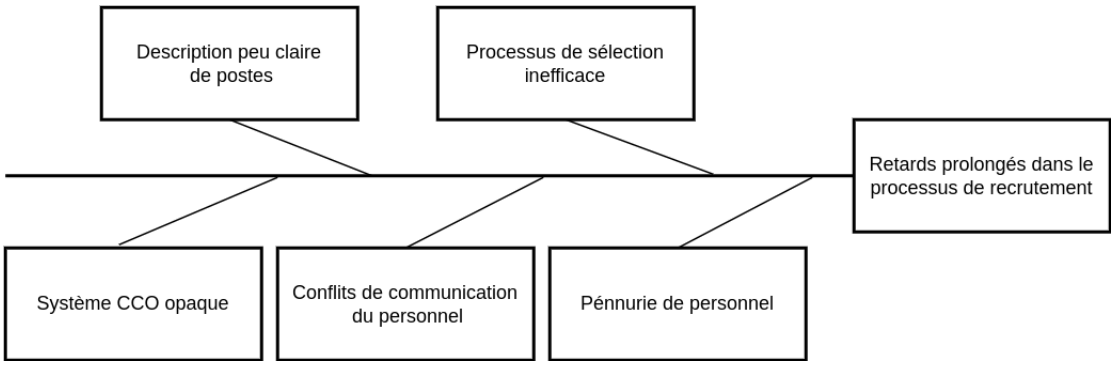


FIGURE II.2 – Schématization générale de l'exemple : sous-problème et causes associées.

Enfin, le diagramme en arêtes de poison permet de mettre en évidence les défis globaux, favorisant des solutions globales pour l'ensemble du processus de recrutement. En identifiant les causes sous-jacentes des problèmes organisationnels, nous pouvons les aborder plus efficacement et mettre en œuvre des changements ayant un impact durable pour améliorer la CCO.

Consolidation des artefacts dans un tableau explicatif

Enfin, nous consolidons toutes ces informations dans un tableau de processus explicatif, comme suggéré par Zapata [107]. Ce tableau de processus explicatif est un artefact-clé, c'est-à-dire un produit tangible résultant de la méthodologie, qui lie tous les éléments précédemment identifiés et offre une vue d'ensemble du contexte organisationnel. Dans ce tableau, chaque processus, objectif et problème est relié à ses causes et effets respectifs, créant une représentation intégrale.

Non seulement cela fournit une vue d'ensemble de l'organisation, mais cela facilite également l'identification des domaines d'intervention potentiels pour améliorer le processus de recrutement. Par exemple, si l'objectif est de réduire le temps d'embauche, le tableau peut aider à identifier des processus spécifiques qui contribuent à la durée du retard et à développer des stratégies pour les accélérer.

En conclusion, notre méthodologie met l'accent sur une compréhension approfondie du contexte organisationnel dans le recrutement. Cette consolidation, allant des acteurs aux artefacts, processus et interactions, permet une représentation détaillée du CV et de l'OE, positionnant les modèles de CCO pour des performances optimales et un alignement avec la dynamique organisationnelle.

La Figure II.3 illustre l'exemple d'une vue spécifique d'un sous-module de la représentation organisationnelle dérivée des expérimentations présentées dans le Chapitre III. Elle met principalement en évidence deux sous-processus : l'analyse, la validation, la rédaction et la communication de l'OE, ainsi que l'analyse et le stockage du CV si ce dernier présente un intérêt pour le recruteur.

Synthèse sur la représentation du contexte organisationnel

Naviguer avec succès à travers les différentes étapes de la représentation de l'organisation permet une compréhension approfondie et conceptuelle de sa structure et de ses dynamiques. Dans le cadre de cette thèse, notre représentation s'appuie sur une entreprise de conseil en TIC. Dans cette section, nous fournissons une synthèse de sa composition en tant que contexte organisationnel, en soulignant que de telles structures peuvent varier selon les organisations.

Au premier plan de la représentation du contexte organisationnel se trouvent les acteurs primaires de la CCO, notamment les recruteurs. Leurs rôles centraux et leurs actions au sein du tissu organisationnel sont détaillés, allant de tâches générales comme la collecte des CV des candidats à des procédures manuelles plus spécifiques telles que la catégorisation de ces CV en fonction de leur pertinence pour le processus de recrutement.

De plus, nous intégrons la représentation des acteurs secondaires, indirectement liés aux cycles de vie des CV et des OE. Ce groupe comprend les chefs de projet, les directeurs de développement et même les clients. En approfondissant le rôle des clients, nous clarifions les

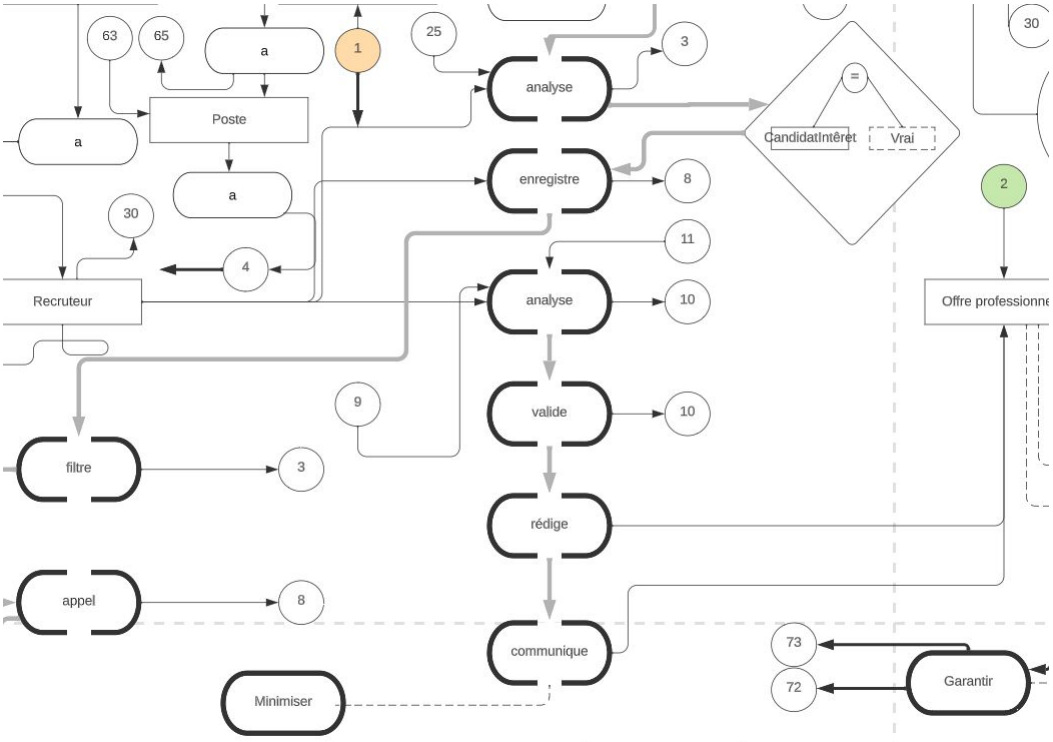


FIGURE II.3 – Vue spécifique d'un sous-module de la représentation organisationnelle, mettant en évidence les étapes de traitement d'une offre professionnelle (représentées par des ovales à ligne tiretée épaisse) et des critères de décision des recruteurs (symbolisés par un losange). Les cercles numérotés font référence à des connecteurs entre des entités éloignées sur le schéma. Les ovales à trait continu fin illustrent des attributs ou des actions, tandis que les ovales à trait continu épais symbolisent des objectifs organisationnels. La représentation du contexte organisationnel comporte plus de 96 concepts et plus de 200 relations, dérivés des expérimentations menées dans le chapitre III au sein de l'entreprise DSI Group.

actions qui sont liées à la CCO. Par exemple, un client pourrait demander des services de conseil, complétant sa demande avec le texte d'un appel d'offre. Cette offre est ensuite analysée par les recruteurs pour élaborer une OE appropriée, identifiant les caractéristiques du poste requis. De telles interactions pourraient évoluer en demandes plus critiques de la part des clients, comme des ajustements agiles au niveau de priorité d'une compétence professionnelle précédemment décrite dans une OE.

Au-delà des acteurs et de leurs actions coordonnées, nous avons identifié des attributs, leur expertise, leurs rôles spécifiques au sein de l'organisation, et leurs interactions fréquentes avec les processus associés à la CCO. Une analyse de leur comportement en réponse à des événements-clés de la CCO est également présentée. Cela comprend des stratégies lors de la rédaction de nouvelles OE, comme la compréhension initiale des concepts techniques requis, les conditions sous lesquelles les OE sont diffusées, et les événements pouvant entraîner une refonte dyna-

mique d'une offre. Par exemple, un manque de candidats appropriés pourrait nécessiter une reformulation du titre de la OE, suivi de sa republication immédiate.

Cette esquisse conceptuelle est enrichie d'informations contextuelles du CV, essentielles pour le CCO et les processus décisionnels inhérents dans l'organisation. Une grande partie de ces informations contextuelles est centrée autour de l'évaluation des CV des candidats. Cela comprend des aperçus des compétences professionnelles des candidats, tels que la durée de l'expérience mise en évidence dans la section "Expérience Professionnelle" ou les types de sections mettant en avant leur expertise. L'importance d'une compétence simplement énumérée dans la section 'Compétences' d'un CV pourrait être moindre par rapport à une compétence démontrée à travers des engagements professionnels concrets. D'autres facettes des informatiosn contextuelles du CV comprennent des mentions explicites dans le document, comme la réticence d'un candidat à se relocaliser face aux exigences de mobilité de l'OE, ou une analyse de la stature organisationnelle des entreprises où un candidat a perfectionné ses compétences professionnelles – ces entreprises sont-elles de petite, moyenne ou grande taille organisationnelle ? Sont-elles en phase avec les exigences compétitives du poste en question ?

De plus, des processus centraux, en particulier le processus de recrutement, sont intégrés de manière fluide, chaque phase étant détaillée. Des processus auxiliaires sont également représentés, renforçant la malléabilité des diverses données de la CCO aux processus alignés de l'entreprise. Ces processus couvrent un large éventail, allant de la mise en forme automatique des CV pour une présentation soignée des candidats aux clients, à des recherches de CV de candidats évalués lors de processus de recrutement déjà effectués.

Une autre partie de la représentation organisationnelle concerne les objectifs organisationnels des acteurs primaires tels que les recruteurs. Ceux-ci s'articulent autour du respect des délais de chaque phase du processus de recrutement et de la garantie que le contenu de l'OE soit adapté à une visibilité optimale sur les plateformes de recrutement en ligne. Cela nécessite une adaptation fluide, en particulier des compétences professionnelles mises en avant.

En complément des éléments mentionnés ci-dessus, la représentation du contexte organisationnel intègre également la structure et le contenu des CV et des OE, tels que perçus par les principaux acteurs qui les orchestrent. Cette structure spécifique à chaque type de document sera présentée dans les sections suivantes de ce chapitre.

En somme, la représentation du contexte organisationnel, ancrée dans les dynamiques organisationnelles internes, offre des perspectives sur les interconnexions potentielles entre acteurs, actions et objectifs au sein du processus de la CCO. Elle représente la base fondatrice de l'ontologie-mère. En explorant les actions variées des parties prenantes telles que les recruteurs et leurs interactions possibles avec des acteurs secondaires, la représentation des OE et CV peut être mieux adaptée pour correspondre aux connaissances spécifiques de chaque organisation. En plus, la représentation d'informations contextuelles du CV associées aux compétences profession-

nelles offre une perspective plus large sur l'évaluation des candidats par rapport aux pré-requis de l'OE, au-delà d'une simple comparaison des expériences professionnelles. Cela permet d'affiner la représentation et la méthodologie d'évaluation de la CCO, rendant le processus plus en phase avec les exigences techniques et organisationnelles de l'entreprise.

II.6 Extraction de l'expertise du recruteur

Sur la base de cette représentation contextualisée des documents et de l'ontologie-mère dérivée, nous menons une analyse des perceptions et des stratégies des recruteurs concernant la pertinence de l'information dans les OE pour améliorer le classement des CV.

Annotation des documents

Les recruteurs procèdent à l'annotation des informations pertinentes dans les OE en s'attendant à ce que ces éléments apparaissent dans les CV au cours des différents processus de recrutement. Alors que les recruteurs lisent et annotent les informations les plus pertinentes, chaque annotation est observée et décrite en détail.

Notre approche identifie et classe les comportements communs et transversaux des recruteurs lors de l'annotation de documents en deux catégories : les comportements explicites (tels que la sélection d'un terme) et les comportements implicites (tels que la non-sélection d'un terme, par choix ou par omission). Ces comportements d'annotation transversaux sont décrits par des règles sémantiques. L'ontologie-mère conçue à l'étape II.5 est utilisée pour modéliser les concepts et les relations nécessaires à chaque règle. Ces règles sémantiques servent de marqueurs textuels de pertinence de l'information dans les OE.

Pour donner un exemple, une règle sémantique pourrait indiquer que si un recruteur sélectionne explicitement des termes associés au secteur d'activité principal d'un poste spécifique, ces termes seront généralement considérés comme pertinents.

En ce qui concerne les CV, un objectif principal de notre méthodologie est la segmentation, une étape préliminaire mais essentielle du processus d'annotation sémantique. Nous étudions la manière dont les recruteurs traitent visuellement les CV dans le but d'extraire des marqueurs grapholinguistiques pour renforcer les processus de segmentation automatisés.

En analysant les techniques de segmentation des recruteurs, nous acquérons des connaissances sur l'importance de divers éléments de format, tels que la taille de la police, la couleur, le type de police, la mise en gras, l'italique et les tendances prédominantes pour les en-têtes de CV. Ces connaissances sont essentielles pour affiner les processus de segmentation automatiques.

La segmentation, en raison de sa nature complexe, se distingue comme l'une des tâches les plus difficiles dans l'analyse automatisée des CV. Une segmentation réussie est essentielle,

garantissant des annotations sémantiques précises, en particulier pour des sections de CV fondamentales pour la CCO, comme l'expérience professionnelle et l'éducation.

II.7 Construction et intégration de ressources ontologiques

Comme noté précédemment, la pierre angulaire de la méthodologie est une ontologie de haut niveau, appelée «ontologie-mère». Cette ontologie est davantage enrichie et renforcée par l'intégration d'autres ontologies existantes et plus spécifiques qui sont liées au contexte organisationnel étudié.

Développement de l'ontologie-mère

L'ontologie principale, basée sur l'ontologie-mère, représente à la fois les CV et les OE. L'enrichissement progressif de cette ontologie assure qu'elle représente plus efficacement les connaissances spécifiques aux domaines présentes dans les documents.

Intégration des ontologies

Nous avons fusionné les ontologies dans l'ontologie-mère en utilisant une approche hybride. Ce processus a combiné des modèles de transformateurs (BERT), une analyse des variantes terminologiques [178], et des métriques de qualité des ontologies [215].

Plus précisément, le processus d'intégration des ontologies externes a été effectué comme suit. Pour chaque paire de concepts appartenant à différentes ontologies, nous avons défini trois types possibles de relations :

- **correspondance étroite :** lorsque deux concepts ont un degré de similarité BERT supérieur à un seuil défini $\alpha \in [0, 1]$, nous considérons qu'il y a une similarité étroite entre eux ;
- **correspondance exacte :** en plus de remplir la condition de correspondance étroite, au moins une paire de concepts qui sont voisins des deux concepts principaux ont une relation de correspondance exacte ou étroite ; la Figure II.4 illustre un exemple de correspondance exacte entre deux concepts, l'un appartenant à une ontologie interne de DSI Group, et l'autre à l'ontologie internationale ESCO ;
- **pas de correspondance :** si aucune des deux relations précédentes ne peut être établie, nous supposons qu'il n'y a pas suffisamment de preuves pour conclure à une similarité étroite entre la paire de concepts.

Il convient de noter que nous avons étendu la méthode BERT pour gérer ces comparaisons. Une analyse des variantes terminologiques a été réalisée afin de déterminer si les termes associés à un concept correspondent à des variations des termes liés à un second concept (annexe G). Cette démarche contribue à affiner et à améliorer la précision du processus d'appariement, garantissant

ainsi une adaptation plus fidèle aux nuances intrinsèques de l'utilisation des termes dans divers contextes.

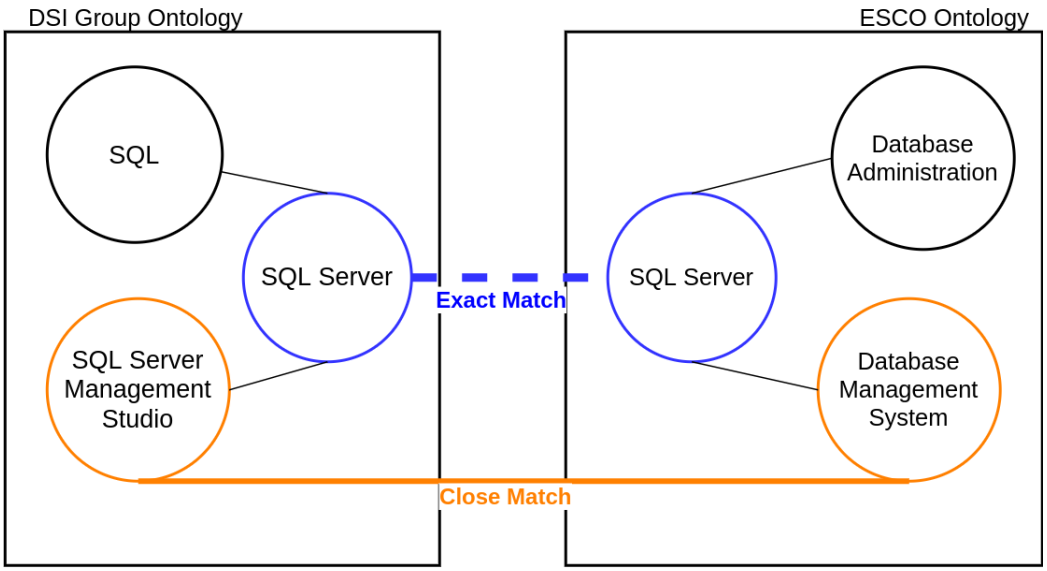


FIGURE II.4 – Exemple de correspondance exacte entre deux concepts appartenant à différentes ontologies.

De cette manière, les ontologies qui ont été intégrées sont les suivantes.

Ontologie de DSI Group

Premièrement, nous avons utilisé l'ontologie interne de DSI Group, une société de conseil française. Cette ressource englobe plus de 36 000 compétences professionnelles, exprimées à la fois en français et en anglais, couvrant un large éventail de domaines d'expertise. L'ontologie fournit une représentation complète des compétences requises dans divers contextes professionnels et sert de ressource fondamentale pour notre travail.

Ontologie ESCO

Deuxièmement, nous avons intégré l'ontologie ESCO [1]. En intégrant l'ontologie ESCO, nous pouvons tirer parti de sa large couverture des professions, compétences et qualifications pour améliorer la représentation des concepts liés au travail dans l'ontologie-mère.

1. `https://ec.europa.eu/esco/lod/static/model.html`

Cadres de compétences professionnelles

Nous avons reconstruit les ontologies qui sous-tendent les cadres de compétences professionnelles, tels que O*NET [2], CIGREF [3], et ROME [4]. Pour ce faire, nous avons utilisé la transformation de texte en triplets RDF, une approche qui facilite la transition d'un format textuel à une représentation structurée [216].

Enrichissement de l'ontologie par apprentissage ontologique

En plus de ces sources, nous avons développé une ontologie basée sur 14 000 CV anonymisés et 2 000 OE. Ce processus a été réalisé de manière semi-automatique, en utilisant des techniques avancées d'apprentissage ontologique [217]. Plus précisément, nous avons extrait et traité les sections des documents qui se concentraient sur les compétences professionnelles.

Cette approche d'apprentissage ontologique nous a permis d'identifier des concepts et des relations pertinents, enrichissant l'ontologie-mère avec les connaissances spécifiques aux domaines présentes dans le corpus. En capturant les concepts-clés et leurs associations, nous avons amélioré la capacité de l'ontologie à représenter et à interpréter les informations spécifiques du marché avec une plus grande précision.

Enrichissement, intégration et mise en œuvre de l'ontologie

Notre approche de développement ontologique est holistique, impliquant non seulement la création d'une «ontologie-mère» centrale, mais aussi son enrichissement continu, son intégration avec d'autres ontologies pertinentes et sa mise en œuvre finale dans un cadre unifié.

Processus d'enrichissement

L'ontologie-mère est continuellement affinée avec des connaissances spécifiques au domaine, des apports d'experts et des retours empiriques. Ce processus itératif assure que l'ontologie reste adaptable et reflète le domaine professionnel en évolution.

Intégration avec d'autres ontologies

Nous intégrons de nouvelles ontologies alignées, telles que celles spécifiques aux ressources humaines et aux processus de recrutement. Cette approche multi-ontologie améliore la profondeur de l'ontologie-mère, aidant à l'extraction d'informations pertinentes et à l'annotation sémantique des CV et des OE.

2. https://www.onetonline.org/
3. https://www.cigref.fr/
4. https://www.pole-emploi.fr/employeur/vos-recrutements/le-rome-et-les-fiches-metiers.html

Ontologie unifiée finale

Grâce à cette approche hybride, nous créons une ontologie intégrée qui :

— fournit une structure plus cohérente et compréhensible pour les informations extraites des OE et des CV ;

— agit comme un cadre commun pour l'annotation sémantique, transformant les données brutes en informations interprétables et exploitables alignées avec le contexte organisationnel ;

— facilite la communication entre divers composants de la méthodologie, tels que l'agent CDI possibiliste, en fournissant un vocabulaire et une structure communs.

En permettant à tous ces composants de "parler le même langage", l'ontologie unifiée vise à assurer un fonctionnement plus synergique à travers les documents étudiés et le contexte organisationnel, alignant tous les éléments vers une représentation de document plus unifiée pour une CCO plus pertinente du point de vue sémantique.

Enfin, pour plus de détails concernant l'évaluation de la qualité de l'ontologie, nous reonvoyons le lecteur à l'annexe L où les métriques de qualité utilisées sont expliquées et catégorisées conformément à la norme ISO/IEC 25012 [5].

Avantages de l'intégration ontologique

L'intégration des ontologies apporte plusieurs avantages à la méthodologie :

— **couverture améliorée :** en incorporant plusieurs ontologies, nous pouvons obtenir une couverture plus large des concepts et de leurs relations, capturant une vue plus complète du contexte organisationnel, des secteurs d'activité économique et des informations liées à l'emploi ;

— **interopérabilité sémantique :** l'intégration des ontologies assure une interopérabilité sémantique en alignant et en reliant les concepts issus de différentes ontologies, garantissant ainsi une meilleure cohérence dans la représentation des connaissances ;

— **réutilisation des connaissances existantes :** l'exploitation des ontologies existantes réduit l'effort requis pour développer une ontologie complète à partir de zéro ; nous pouvons nous appuyer sur des ressources établies et bénéficier de l'expertise qu'elles contiennent ;

— **enrichissement et adaptabilité :** le processus d'intégration facilite l'enrichissement et le raffinement continus de l'ontologie-mère, assurant sa pertinence et son adaptabilité aux besoins changeants des recruteurs et aux contextes de domaine en évolution.

5. `https://www.iso.org/fr/standard/35736.html`

II.8 Extraction d'informations à partir des OE

Dans cette section, nous présentons la méthodologie spécifique proposée pour une représentation plus robuste des OE centrée sur la pertinence de l'information.

II.8.1 Dérivation de l'ontologie OE

Sur la base de la phase précédente, une ontologie liée à l'OE a été dérivée, intégrant les concepts nécessaires pour une représentation robuste du document, tant d'un point de vue textuel que contextuel. Un aperçu de l'ontologie dérivée est présenté à la Figure II.5.

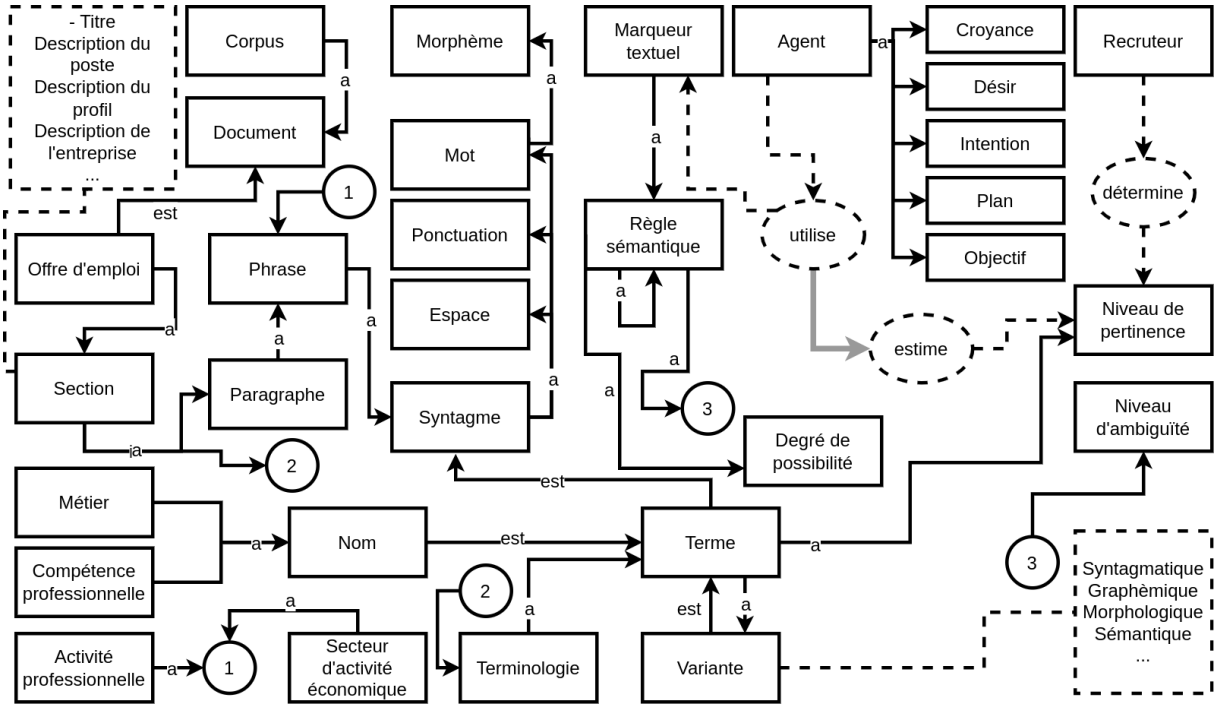


FIGURE II.5 – Vue d'ensemble de l'ontologie-mère de l'OE, dérivée de la représentation du contexte organisationnel étudié (DSI Group).

II.8.2 Prétraitement et formatage

Le prétraitement des OE est une phase préliminaire consistant à extraire et nettoyer le texte depuis divers formats numériques (PDF, Word, HTML). Cette étape élimine les éléments superflus, convertit les fichiers en format JSON pour un traitement cohérent et une compatibilité large, puis segmente et normalise le texte. La segmentation, effectuée à différents niveaux (paragraphes, phrases, etc.), est essentielle pour capturer les nuances des OE, tandis que la normalisation facilite le traitement automatisé et assure la comparabilité entre les diverses OE. Plus de détails sur le prétraitement sont fournis dans l'annexe F.

II.8.3 Évaluation de qualité du texte : l'intérprétabilité du document

Une base initiale pour le contrôle de l'incertitude commence par l'estimation préliminaire de la qualité du texte de l'OE. Dans notre contexte, nous évaluons une métrique basique mais qui peut être hautement indicative de problèmes de malformations dans le document : la métrique d'*interprétabilité du texte*. Cette métrique est définie comme la proportion du vocabulaire du document qui peut être trouvé dans un corpus de documents, comme par exemple, l'ensemble des articles de Wikipédia. Ainsi, cette métrique est définie comme :

$$\tau_{\text{interpret}} = \frac{V_{d_i,\text{wiki}}}{V_{d_i}}, \tag{II.1}$$

où $V_{d_i,\text{wiki}}$ est le nombre de mots de l'OE trouvés dans la base de données Wikipédia, et V_{d_i} est le nombre total de mots dans le document. D'autres métriques telles que la qualité syntaxique, l'orthographe, la lisibilité, entre autres, peuvent aussi être utilisées pour une évaluation plus complète.

II.8.4 Extraction de terminologie

Enfin, la représentation du texte se concentre sur les unités minimales de signification, généralement les termes, qui servent de base pour l'analyse et l'extraction d'informations à partir des OE.

Les termes, dans ce contexte, sont définis comme des classes fonctionnelles d'unités lexicales utilisées dans le discours. Ils sont identifiés à l'aide du *weirdness ratio* (WR), une métrique qui mesure la spécificité de chaque terme en comparant sa fréquence relative dans un corpus spécifique du domaine à celle dans un corpus de langue générale (pour des détails supplémentaires sur le WR, voir l'annexe G et [178, 218]).

L'identification de termes est facilitée par l'utilisation de motifs morphosyntaxiques communs, représentés par des expressions régulières. Ces motifs, identifiés dans diverses études expérimentales [178], sont adaptés à un corpus comprenant 2 000 OE et 14 000 CV. Ils sont principalement composés de phrases nominales et sont appliqués via une analyse syntaxique pour les langues française et anglaise (voir Tableau II.1 pour des exemples).

Il est important de remarquer que chaque mot dans l'OE subit un prétraitement, incluant l'étiquetage grammatical et la lemmatisation, avant l'application des motifs morphosyntaxiques. Cela permet une identification plus précise et moins redondante des termes au sein des OE.

L'étape finale consiste à analyser les variations des termes de l'OE. Sur la base de cette analyse, nous pouvons décrire les relations entre les termes simples dans le document et les termes plus complexes. En tenant compte des études précédentes [178], nous identifions les variantes terminologiques suivantes :

TABLE II.1 – Exemples de modèles morphosyntaxiques et de termes associés. Dans ces exemples, la lettre N représente un nom ou une abréviation agissant comme un nom, la lettre P représente une préposition, et la lettre A représente un adjectif.

Modèles Morphosyntaxiques des OE		
#	**Motif**	**Exemple**
1	N	ETL
2	N N	Tableau Software
3	N A	Spécifications Techniques
4	N P N	Connaissance de Stambia
5	N P N A	Cabinet de Conseil International
6	N P N N	Connaissance de Stambia ETL

— **variations de dérivation**, qui impliquent des changements morphologiques du terme ; par exemple, l'ajout de préfixes ou de suffixes à un terme de base ;
— **variations de composition**, qui impliquent la combinaison de différents termes pour former un terme plus complexe ; par exemple, la formation de termes composés tels que "spécialiste du développement logiciel" ;
— **variations de synonymie**, qui impliquent des termes ayant des significations similaires mais des formes différentes ; par exemple, les termes "développeur" et "programmeur" peuvent être considérés comme des variations synonymes ;
— **variations graphémiques** : il y a une identification systématique des termes qui diffèrent en raison d'erreurs d'orthographe ; par exemple, le terme "Mstery of the SQL Language" est une variation graphémique du terme "Mastery of the SQL Language" ;
— **variations sémantiques** : le modèle BERT permet l'identification de termes de l'OE ayant des significations étroitement liées ;
— d'autres variations liées aux différences dialectales ou aux variations stylistiques ; par exemple, l'utilisation de termes spécifiques à certaines régions ou industries ; cela inclut, par exemple, le remaniement de termes pour les compétences professionnelles et les professions, comme le remplacement de "développeur logiciel" par "artisan du code" dans une entreprise qui cherche à se démarquer en utilisant un langage plus créatif et moins formel.

Résumé

Analyser et capturer des variations terminologiques dans les OE conduit à une compréhension plus profonde de la terminologie utilisée, permettant une récupération d'informations plus précise. Reconnaître ces variations est essentiel pour des tâches telles que l'identification des compétences professionnelles requises, qui pourraient être exprimées différemment entre les recruteurs dans les OE et les candidats dans les CV. De telles variations représentent une source

significative d'incertitude pour le traitement de ces documents.

Après les étapes de prétraitement, chaque OE est transformée en un document texte propre, segmenté, normalisé et traité terminologiquement, prêt pour l'extraction d'informations. Les étapes ultérieures de notre méthodologie se concentrent sur l'identification et l'annotation des informations pertinentes.

II.8.5 Annotation des OE par les recruteurs

Une fois les OE passées par un prétraitement automatisé et une normalisation, les recruteurs les annotent. Ce processus essentiel permet aux recruteurs de mettre en évidence les informations qu'ils jugent cruciales pour des postes de travail spécifiques. De telles annotations sont inestimables pour élaborer des algorithmes et des modèles d'IA spécifiques au contexte.

Dans cette thèse, nous adoptons une approche d'annotation semi-automatique. Cela signifie que nous combinons l'intervention humaine avec l'automatisation pour obtenir des résultats d'annotation plus précis.

Définition des entités d'annotation

Avant de commencer le processus d'annotation, il est crucial de définir clairement les entités d'annotation, c'est-à-dire les types d'information que nous visons à identifier dans les OE en fonction de l'analyse précédente du contexte organisationnel. Ces entités peuvent inclure, par exemple, le titre du poste, les compétences requises, l'expérience nécessaire, le lieu de travail, le type de contrat, etc. La définition de ces entités doit être en accord avec les objectifs des recruteurs et les besoins du contexte organisationnel.

Création d'un ensemble de données annotées

Un ensemble de données annotées est essentiel pour l'entraînement, le test et la validation des algorithmes d'extraction d'informations. Cet ensemble comprend des OE sélectionnées, qui sont ensuite annotées manuellement en fonction des entités pré-définies. Des experts du domaine, en particulier des recruteurs brièvement formés aux règles d'annotation, entreprennent cette tâche.

Développement et formation de l'outil d'annotation semi-automatique

Avec un ensemble de données annotées manuellement, nous pouvons forger et affiner un outil d'annotation ou d'extraction d'informations semi-automatique. En tirant parti de l'apprentissage automatique (AA) flou avec des modèles tels que la régression logistique floue (RLF) et les arbres de décision flous (ADF), cet outil est formé pour discerner et annoter les entités recherchées dans les OE. L'efficacité de l'outil est évaluée à l'aide de métriques telles que la précision, le rappel et la F1-mesure.

Annotation semi-automatique des OE

Ce processus favorise l'utilisation optimale d'un outil d'annotation semi-automatique basé sur des modèles explicatifs de base, tels que la RLF et les ADF, pour annoter et approfondir l'analyse d'un corpus d'OE. Cependant, il convient de noter que même les outils d'annotation les plus sophistiqués ne sont pas exempts de défauts. Par conséquent, une étape de vérification et de correction manuelle doit être effectuée avec des recruteurs professionnels, ce qui peut potentiellement nécessiter plus de temps par rapport à une annotation entièrement manuelle.

Cette approche assure une amélioration substantielle de la qualité des annotations, conduisant à un corpus annoté de meilleure qualité. L'importance de cette étape ne peut être surestimée car elle sert de fondement pour la modélisation de l'extraction d'informations et d'autres opérations d'analyse ultérieures.

De plus, ce processus offre une opportunité d'acquérir des connaissances plus approfondies sur la pertinence d'une approche d'annotation et d'extraction d'informations semi-automatique. Il encourage également l'interaction constructive avec les recruteurs, qui sont invités à partager leurs commentaires sur le processus.

Mise à jour et affinement de l'outil d'annotation

L'outil d'annotation subit un affinement régulier tout au long du processus d'annotation. Les corrections manuelles sont réincorporées dans l'ensemble de données d'entraînement, et l'outil est re-entraîné, assurant son évolution et son amélioration continues.

Annotation collaborative

Pour assurer une meilleure qualité et une annotation plus homogène, nous adoptons, le cas échéant, des approches d'annotation collaborative. Dans ces approches, plusieurs annotateurs travaillent sur le même document, et leurs annotations sont ensuite comparées. Les désaccords entre les annotateurs sont analysés, ce qui conduit à une meilleure compréhension des annotations et des comportements des recruteurs.

Gestion des annotations

Une gestion efficace des annotations est essentielle pour faciliter leur utilisation ultérieure. Cela implique un stockage structuré des annotations et leur association avec les documents sources. De plus, le processus d'annotation est documenté, y compris les directives d'annotation, les décisions prises lors de la résolution des désaccords et les versions de l'outil d'annotation utilisées.

Analyse des annotations pour la dérivation de marqueurs textuels alignés avec le contexte organisationnel

Sur la base des annotations réalisées, nous analysons les stratégies utilisées par les recruteurs pour sélectionner les informations essentielles dans les OE. Pour représenter la description de chaque action d'annotation observée chez les recruteurs, nous utilisons le langage contrôlé proposé par [107]. Ce langage nous permet de représenter les actions de manière séquentielle comme des triplets $< sujet, verbe, prédicat >$.

Nous catégorisons ces actions comme actives (par exemple, <recruteur, sélectionne, terme>) ou passives (par exemple, <recruteur, évite, terme> ou <recruteur, évite, section_offre_emploi>).

Une fois que les annotations sont décrites de manière contrôlée, nous utilisons l'algorithme Apriori [175] pour identifier les sous-séquences d'actions que les recruteurs effectuent de manière constante. Ces sous-séquences d'actions décrivent des modèles comportementaux, formalisés en tant que règles sémantiques, en utilisant l'ontologie-mère décrite dans la section II.5. Les règles obtenues représentent des marqueurs de pertinence textuelle dans les OE. La Figure II.6 illustre un exemple de l'analyse des perspectives des recruteurs.

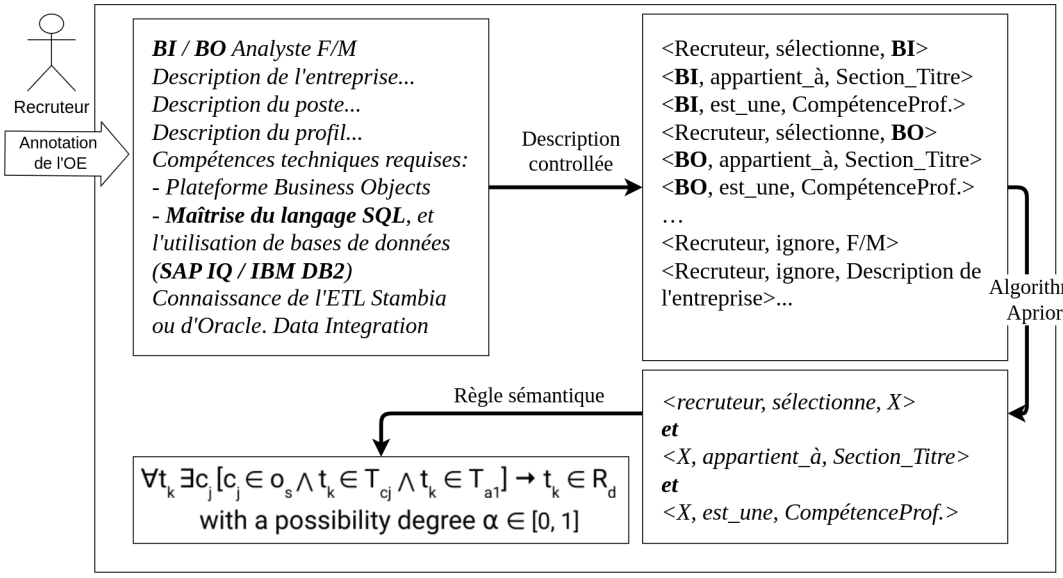


FIGURE II.6 – Processus d'analyse des points de vue des recruteurs, dans l'ordre suivant : le recruteur annote une OE, l'annotation est décrite dans un langage contrôlé, l'algorithme Apriori est utilisé pour identifier les comportements systématiques, et les règles sémantiques (marqueurs textuels) sont dérivées. Dans la règle sémantique présentée, t_k désigne un terme spécifique de l'OE, c_j représente le concept j associé à une compétence professionnelle dans l'ontologie o_s, T_{cj} correspond à un terme utilisé pour représenter le concept c_j, T_{a1} représente l'ensemble de termes du titre de l'OE, et R_d désigne l'ensemble des termes pertinents de l'offre d'emploi d.

Conclusion

En essence, l'annotation semi-automatique des OE joue un rôle central dans l'établissement d'une fondation de CCO plus robuste, facilitant l'identification d'informations pertinentes. Bien qu'elle exige une certaine supervision manuelle, elle améliore considérablement la qualité de l'annotation en tirant parti de l'expertise des recruteurs. Cette méthode approfondit également notre compréhension des comportements des recruteurs, garantissant une gestion organisée et structurée de l'annotation pour les applications ultérieures.

Ayant identifié les bases pour l'annotation des OE et l'extraction d'informations pertinentes, nous sommes prêts à introduire notre cadre pour l'extraction automatisée d'informations pertinentes des OE.

II.9 Architecture possibiliste, basée sur l'ontologie des croyances, désirs et intentions pour l'extraction d'informations des OE

Dans cette section, nous introduisons l'approche méthodologique proposée pour automatiser l'extraction d'informations des OE. Nous inspirant de l'expertise des recruteurs et des caractéristiques spécifiques du contexte organisationnel identifiées précédemment, nous nous concentrons sur l'identification des informations fondamentales contenues dans les OE.

II.9.1 Évaluation préliminaire de la pertinence d'une approche possibiliste pour l'extraction d'informations

Avant de mettre en œuvre l'approche possibiliste, nous avons mené une analyse préliminaire sur le processus d'extraction d'informations des OE dans un contexte organisationnel spécifique. Deux approches distinctes ont été examinées : probabiliste et possibiliste. En comparant les modèles classiques et flous, nous visons à évaluer et à justifier leur utilisation dans l'extraction automatique des OE. Plus précisément, nous avons évalué :

— **RLC (modèle linéaire) :** la RLC est un modèle statistique utilisé en AA et en analyse de données pour les problèmes de classification binaire ; la fonction logistique, également connue sous le nom de fonction sigmoïde, est une courbe en forme de S qui associe tout nombre réel à une valeur entre 0 et 1 :

$$p(x) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + \cdots + b_m x_m)}} \tag{II.2}$$

dans l'équation II.2, appliquée à notre contexte, $p(x)$ est la probabilité qu'un terme dans l'OE soit pertinent, x_m est le marqueur textuel ou la caractéristique m de l'OE, b_0 est la

constante de biais, et $b_1, \ldots, b_m$ sont les poids associés à l'ensemble des marqueurs ;

— **RLF (modèle linéaire) :** soit $t = \{t_1, t_2, t_3, \ldots t_m\}$ l'ensemble des termes dans l'OE ; nous postulons que ces termes peuvent être représentés comme une combinaison linéaire d'un ensemble de marqueurs textuels ou de caractéristiques $I(k)$; en utilisant l'algorithme de RLF [219], nous définissons $\mu_i \in \{C_1(\text{terme pertinent}), C_2(\text{terme non pertinent})\}$ comme l'annotation du recruteur sur le i-ème terme d'une OE ; nous estimons le paramètre $\tilde{u_i}$ à partir du rapport $\frac{\tilde{\mu_i}}{1-\tilde{\mu_i}}$; dans notre contexte, $\frac{\tilde{\mu_i}}{1-\tilde{\mu_i}}$ peut être interprété comme la possibilité qu'un terme soit pertinent par rapport à la possibilité qu'il soit non pertinent, ou vice versa ; par conséquent, le modèle est le suivant [219] :

$$\tilde{W_i} = \ln \frac{\tilde{u_i}}{1-\tilde{u_i}} = A_0 + A_1 x_{i1} + \cdots + A_n x_{in}, \quad i = 1, \ldots, m\,; \tag{II.3}$$

où $\tilde{W_i}$ est la sortie estimée qui peut être retransformée en $\tilde{u_i}$ par le principe d'extension et $A_i = (a_i,\, s_i)$ représente un nombre flou et symétrique triangulaire avec pour centre a_i et pour écart s_i ;

— **ADF (modèle non-linéaire) :** si le modèle de régression logistique floue s'avère plus pertinent que le modèle classique, la pertinence des modèles flous non linéaires tels que les ADF [172] est également évaluée.

Alors que théoriquement, une approche floue est justifiée en raison de l'incertitude inhérente de la CCO [45], ces modèles visent à fournir une évaluation pratique de ses caractéristiques dans la représentation de divers points de vue des recruteurs dans les OE.

II.9.2 Croyances, désirs et intentions dans l'architecture de l'agent

À travers diverses observations et expériences, nous avons appris que le raisonnement humain est caractérisé par une incertitude inhérente et une sensibilité remarquable au contexte dans lequel il se produit [45]. Dans le contexte de l'extraction de termes pertinents des OE, cette logique et ce principe peuvent être appliqués.

Les recruteurs font une série d'inférences pour identifier les termes pertinents lors de l'annotation des OE. En supposant une compréhension commune de la langue et des connaissances de base partagées, ce processus implique une lecture et une interprétation attentives du texte de l'OE [45].

Cette dynamique comporte un degré d'incertitude cognitive. Des termes qui peuvent paraître pertinents dans un contexte peuvent ne pas l'être dans un autre, et vice versa. De plus, chaque recruteur peut avoir sa propre interprétation et perspective, ajoutant une couche supplémentaire de complexité [91].

Face à ce défi, nous proposons la mise en œuvre d'une architecture possibiliste. Cette architecture cherche à intégrer le raisonnement sémantique humain et l'incertitude associée pour

améliorer et optimiser l'extraction automatique de termes pertinents.

Plus précisément, cette architecture est basée sur les travaux de da Costa Pereira [2] et est fondée sur ce qui a été défini comme le paradigme des CDI de l'agent. Elle est liée à la modélisation des variables en appliquant la théorie des possibilités, permettant une approche théoriquement robuste, flexible et adaptable à l'extraction d'informations des OE (Figure. II.7). Plus spécifiquement, l'approche proposée par [2] fournit une base théorique intégrée basée sur la théorie des possibilités pour la représentation et la manipulation des CDI chez les agents cognitifs. L'utilisation de distributions de possibilité distinctes permet de prendre en compte l'incertitude et la représentation progressive des croyances et des désirs. L'approche fournit également une évaluation de la cohérence logique des objectifs, en tenant compte de l'état actuel de l'agent, et en améliorant ainsi l'explicabilité du modèle. Cette approche émergente vise à offrir un paradigme enrichissant pour la modélisation des processus cognitifs dans les systèmes d'agents.

La Figure II.7 illustre comment les CDI de l'agent sont liés et interagissent pour faciliter l'extraction de termes pertinents. Dans notre contexte, l'architecture vise à permettre une prise de décision plus transparente, éclairée et précise dans la CCO. Les détails de l'architecture sont fournis dans les sections suivantes.

II.9.3 Module de croyance de l'agent

Ce module vise à représenter les diverses croyances de l'agent concernant les différents facteurs textuels et contextuels et les particularités des OE qui sont liées à la pertinence de l'information. Les croyances sont spécifiquement associées à un niveau de nécessité qui représente la mesure dans laquelle l'agent peut affirmer qu'un terme d'OE est essentiel ou non à sa représentation. Ces croyances, en raison de leur nature graduée, sont également désignées sous le terme de *croyances graduées*.

Dans cette approche possibiliste, nous définissons des niveaux de confiance et des seuils de tolérance pour les sources d'information et formulons la distribution de possibilité qui représente les croyances de l'agent. Grâce à cette structure, les croyances de l'agent peuvent être modélées et ajustées plus efficacement, en fonction des besoins spécifiques de la CCO.

Au niveau de l'architecture, le module de croyance se compose de deux sous-modules. Le premier est une distribution des niveaux de confiance τ, qui représente le niveau de confiance que l'agent attribue à ses sources ϕ. Dans notre contexte, il existe quatre catégories principales de sources : les OE, les CV, les ressources ontologiques et les marqueurs textuels de pertinence de l'information.

Nous définissons les niveaux de confiance en termes de métriques de qualité qui visent à détecter d'éventuelles incohérences, telles que des textes d'OE mal formulés, des problèmes de conformité avec l'ontologie ou des marqueurs textuels de pertinence ambigus. Chaque niveau

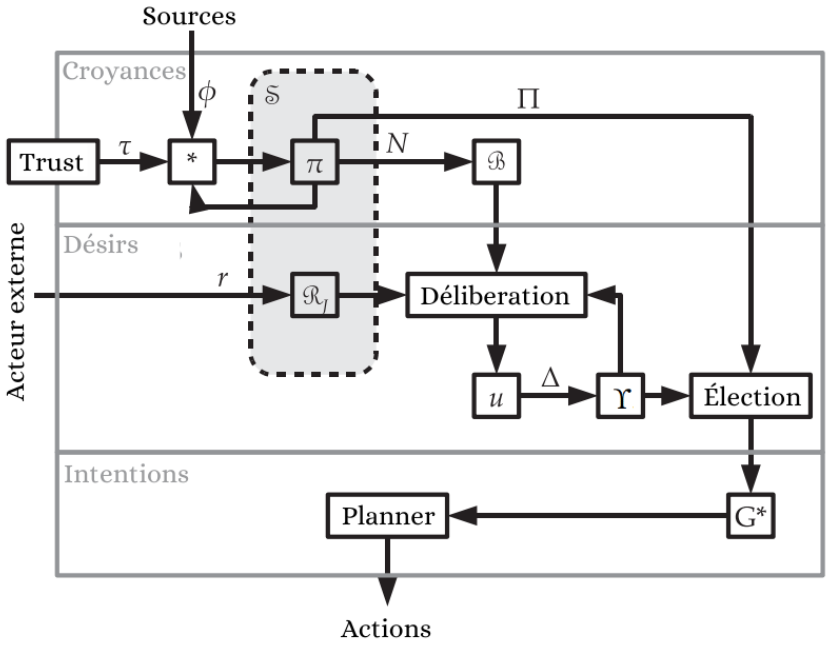


FIGURE II.7 – Structure fondamentale de l'agent CDI dynamique [1], adaptée de [2].

de confiance est associé à un seuil de tolérance β_j. Une source dont le niveau de confiance τ_j est inférieur à un seuil de tolérance donné β_j est rejetée par l'agent. Ces niveaux de tolérance peuvent être fixes ou dynamiques en fonction des conditions spécifiques de chaque métrique de qualité, de leurs interrelations et des paramètres de contexte qui leur sont associés.

Pour assurer l'exactitude et la pertinence des informations extraites, la fiabilité des sources est cruciale dans l'approche proposée. Par conséquent, la qualité et l'exhaustivité des sources d'information, ainsi que leur niveau de fiabilité, jouent un rôle crucial dans la détermination des conditions pertinentes pour les OE.

Le second sous-module du module de croyance utilise une distribution de possibilité π qui représente les croyances de l'agent $\mathcal{B}$. Chaque terme est caractérisé par une distribution de possibilité formée par les niveaux de pertinence (degrés de possibilité) fournis par chaque marqueur. Cette distribution de possibilité induit une mesure de possibilité Π [2]. Cette mesure indique dans quelle mesure un terme est pertinent. La mesure de possibilité est ensuite associée à sa mesure de nécessité duale N. Il convient de noter que les mesures de possibilité des croyances sont limitées par le niveau de confiance de leurs sources de connaissances respectives. À cette fin, un opérateur de modification de croyance $*$ est défini. Dans le cadre de cette thèse, cet opérateur est complété par un opérateur d'évaluation globale des croyances (annexe E).

Ainsi, à partir du niveau de croyance des sources d'information qui permet l'évaluation de la pertinence des termes dans un contexte dynamique, il est possible de représenter l'incertitude

inhérente dans l'interprétation de la pertinence de l'information, ce qui tend à améliorer la flexibilité, la fiabilité et l'explicabilité du système.

II.9.4 Module de désir de l'agent

Le module de croyance est complété par le module de désir, symbolisé par une distribution d'utilité (plus formellement, une distribution de possibilité). Ce module indique si l'utilisation d'un marqueur textuel de pertinence est souhaitable ou non, en tenant compte des croyances, des désirs et des règles qui génèrent ces désirs de l'agent. Les règles de génération de désirs, notées R_J, sont des conditions logiques qui expriment des relations de dépendance entre les croyances et les désirs [2].

L'évaluation de ces règles de génération de désirs fournit à l'agent la capacité d'évaluer dans quelle mesure Δ il est justifié d'utiliser chaque marqueur. Les marqueurs textuels associés à des degrés positifs deviennent des désirs justifiés, notés Υ.

Ce composant ajoute une couche supplémentaire de réflexion et de considération au processus d'extraction des termes pertinents. Il permet à l'agent non seulement d'identifier des termes pertinents, mais aussi d'évaluer et de décider du moment opportun pour utiliser des marqueurs textuels de pertinence spécifiques. Cela contribue à rendre l'extraction de termes adaptative, permettant à l'agent de s'adapter et de réagir plus efficacement à la variabilité et à la complexité du contenu (textuel) et du contexte (contextuel) des OE.

II.9.5 Module d'intention de l'agent

Le troisième composant de l'agent, appelé "intentions", délimite les actions que l'agent entreprend en fonction de ses croyances et désirs justifiés et de ses objectifs. Dans notre domaine, ces actions concernent les techniques de traitement du langage naturel essentielles pour extraire automatiquement des termes pertinents des OE.

Chaque action est déterminée par une évaluation continue des croyances et des désirs de l'agent, sur la base desquelles une séquence d'étapes concrètes à suivre est définie. Ce processus décisionnel est intrinsèquement lié aux objectifs de l'agent, qui guident et orientent ses actions.

Par exemple, si un terme donné est jugé pertinent dans une OE et que ce terme est également associé à un désir justifié, alors l'action correspondante serait d'extraire ce terme. À l'inverse, si un terme est considéré comme pertinent mais est associé à un désir non justifié, l'action appropriée pourrait être de reconsidérer sa pertinence ou de rechercher davantage d'informations (autres marqueurs textuels de pertinence de l'information).

De plus, il est important de noter que les intentions de l'agent peuvent évoluer avec le temps, en fonction des changements dans ses croyances et ses désirs. Cela signifie que l'agent vise à s'adapter et à répondre de manière flexible aux changements dans les OE et leur contexte,

améliorant ainsi la précision et la pertinence de l'extraction de termes par rapport à la dynamique du processus de recrutement.

En essence, le module d'intention est essentiel pour traduire les croyances et les désirs de l'agent en étapes d'action soutenues par un plan structuré. Ce cadre offre une base solide pour la prise de décision, permettant l'extraction fluide de termes pertinents des OE, même dans des contextes dynamiques et changeants.

II.9.6 La propagation de l'incertitude dans l'architecture : opérateurs de modification et de détermination de croyances

Dans le cadre de la mise à jour des croyances de l'agent, deux opérateurs-clés sont introduits. Le premier, appelé *opérateur de modification de croyance* [2], est activé lors de la réception de nouvelles informations (ϕ). Il ajuste la distribution de possibilité de l'agent (π) en fonction de la fiabilité de la source (τ), en tenant compte de la cohérence de ϕ avec les croyances existantes. Mathématiquement, la nouvelle distribution π' est calculée à travers trois scénarios principaux. Les deux premiers scénarios de l'opérateur permettent d'augmenter (ou laisser invariant) le degré de possibilité si l'information qui arrive confirme la réalité d'une interprétation de l'agent sur la pertinence des termes. Le troisième scénario permet de réaliser la mise à jour des croyances en cas de contradiction, en prenant en compte la fiabilité de la source d'information. Cela permet d'obtenir une croyance $B(t_i)$ sur la pertinence du terme t_i de l'OE.

Le second opérateur, appelé *mesure de croyance moyenne* (annexe E), est introduit pour évaluer globalement la pertinence d'un terme. Cette mesure agrège les niveaux de possibilité et de confiance associés à chaque marqueur lié au terme, fournissant ainsi une vue plus intégrale des croyances de l'agent. La croyance moyenne de l'agent dans la pertinence d'un terme est donnée par :

$$\overline{B}(t_i) = \frac{\sum_{i=1}^{n} \alpha_{i,t_i} * \tau_{i,t_i}}{m}. \tag{II.4}$$

où α_{i,t_i} représente le niveau de possibilité du marqueur i lié au terme t_i, τ_{i,t_i} est le niveau de confiance du marqueur i associé au terme t_i, et m est le nombre total de marqueurs intégrés dans l'architecture de l'agent.

Enfin, une métrique globale des croyances, $C(t_i)$, est introduite, combinant l'opérateur de modification de croyance et la mesure de croyance moyenne avec un facteur de pondération φ, comme défini par l'équation :

$$C(t_i) = \varphi * B(t_i) + (1 - \varphi) * \bar{B}(t_i). \tag{II.5}$$

Cette métrique offre une évaluation plus équilibrée des croyances de l'agent sur la pertinence d'un terme. Pour plus des détails sur le fonctionnement et l'application de ces opérateurs, nous renvoyons le lecteur à l'annexe E.

II.9.7 Matrice d'états de l'agent : un outil d'explicabilité

Le tableau de vérité, un outil logique listant toutes les valeurs de vérité possibles d'une proposition en fonction des valeurs de vérité de ses composants, est adapté dans notre contexte pour suivre la propagation de l'incertitude au sein de l'architecture de l'agent. Cette version adaptée du tableau de vérité, que nous nommerons matrice d'états de l'agent, est structurée de la manière suivante : ses colonnes représentent la valeur de vérité de chacune des informations reçues par l'agent, tandis que ses lignes contiennent les interprétations faites par l'agent sur la pertinence des termes, en fonction des différentes valeurs de ces informations. Ces lignes indiquent également le niveau de possibilité ou la plausibilité attribuée à chaque interprétation. Le Tableau II.2 offre une illustration générale de cette matrice.

TABLE II.2 – T_{n,I_m} correspond à la valeur de vérité de l'information ϕ_m sous l'interprétation I_m. $T_{n,r_{t_i}}$ correspond à la valeur de vérité associée à la pertinence du terme sous l'interprétation I_m. Finalement, $\pi(I_m)$ correspond au niveau de possibilité associé à l'interprétation I_m dans l'état actuel des croyances de l'agent.

#I	ϕ_1	ϕ_2	...	ϕ_m	r_{t_i}	π
1	T_{1,I_1}	T_{1,I_2}	..	T_{1,I_m}	T_{1,t_i}	$\pi(I_1)$
2	T_{2,I_1}	T_{2,I_2}	..	T_{m,I_m}	T_{2,t_i}	$\pi(I_2)$
...	..	..	..	...		..
n	T_{n,I_1}	T_{n,I_2}	..	T_{n,I_m}	T_{n,t_i}	$\pi(I_m)$

II.9.8 Formalisation du protocole pour l'extraction d'informations des OE à partir des annotations des recruteurs

Soit $U = \{t\}$ l'ensemble des termes d'une OE donné. Soit C un ensemble flou représentant les niveaux perçus de pertinence d'un terme par le recruteur dans le document. C est représenté par la fonction d'appartenance μ_C, qui associe les annotations du recruteur dans C. C se compose de deux sous-ensembles flous : le sous-ensemble C_1 des niveaux de termes pertinents et le sous-ensemble C_2 des niveaux de termes non pertinents. Ces deux catégories floues sont représentées par des fonctions triangulaires. Nous noterons R (resp. R_1, R_2), le sous-ensemble de C (resp. C_1, C_2) obtenu à partir de la fuzzification des annotations du recruteur sur l'OE.

Chaque terme t dans l'OE peut être décrit par un ensemble de marqueurs de pertinence x_k, obtenus à partir des stratégies et des points de vue des recruteurs. Nous notons $A(t) = \{x_1(t), x_2(t), \ldots, x_k(t)\}$ l'ensemble de marqueurs de pertinence associés au terme t.

Chaque marqueur de pertinence fournit un degré de possibilité pour qu'un terme soit sélectionné. L'étape suivante consiste à fuzzifier ces degrés de possibilité en appliquant une fonction d'appartenance μ_{x_k}. Bien que μ_{x_k} ait été construite de la même manière que μ_C, le codomaine spécifique de chaque marqueur x_k a été pris en compte. Nous interprétons le résultat de ce

processus de fuzzification comme une évidence E_k.

Après avoir fuzzifié x_k et les annotations du recruteur, nous évaluons la possibilité de décrire les annotations floues R sur la base de l'évidence E_k obtenue à partir de x_k. Plus précisément, nous évaluons le niveau d'ambiguïté de la règle suivante : si E_k alors R. Pour ce faire, une mesure de sous-ensemble entre l'évidence E_k et la classification de l'expert R a été proposée [172]. Dans notre contexte, en remplacement de l'opération de sous-ensemble, nous utilisons la distance de Hamming entre les ensembles flous pour refléter plus explicitement dans quelle mesure la connaissance du recruteur R s'aligne avec E_k :

$$S(E_k, R) = \sum_{t \in U} \left| \mu_{x_k}(t) - \mu_R(t) \right|. \tag{II.6}$$

Cette approche systématique permet d'évaluer le degré d'ambiguïté analytique associé à chaque marqueur de pertinence, en fonction du degré de similarité entre l'évidence obtenue à partir du marqueur et les annotations floues des recruteurs. Cela permet l'évaluation et la quantification de la précision et de l'efficacité de chaque marqueur de pertinence dans le contexte d'une OE.

La distance de Hamming mesure le degré de pertinence ou de non-pertinence d'un terme sur la base des évidences disponibles E_k. En appliquant cette distance, nous définissons la possibilité π de classer un terme comme pertinent (R_1) ou non pertinent (R_2), en relation avec les stratégies et les points de vue des recruteurs, comme suit :

$$\pi(R_i \mid E_k) = \frac{S(E_k, R_i)}{\max(S(E_k, R_1), S(E_k, R_2))}. \tag{II.7}$$

La possibilité est intrinsèquement liée au concept d'ambiguïté [172]. Sur la base de E_k, plus il est probable que nous déduisons qu'un terme est pertinent ou non pertinent, moins il y a d'ambiguïté. À partir de $\pi(R \mid E_k)$, nous estimons le niveau d'ambiguïté associé au marqueur x_k, ou de manière équivalente, à l'évidence E_k, comme suit :

$$G(E_k) = g(\pi(R|E_k)) = \sum_{i=1}^{n} (\pi_i^* - \pi_{i+1}^*) \ln(i), \tag{II.8}$$

où $\pi^* = \{\pi_1^*, \pi_2^*, \ldots, \pi_n^*\}$ est la distribution de possibilité $\pi(R \mid E_k)$ permutée et triée de telle sorte que $\pi_i^* \geq \pi_{i+1}^*$ pour $i \in \{1, \ldots, n\}$ et $\pi_{n+1}^* = 0$. Dans notre cas, $n = 2$ car nous évaluons l'ambiguïté lors de la décision de la pertinence d'un terme (R_1) ou non (R_2) en se basant sur x_k.

Cette fonction d'ambiguïté G indique dans quelle mesure il peut être déduit qu'un terme est pertinent ou non pertinent sur la base du marqueur de pertinence I_k. Une valeur de 0 signifie qu'il n'y a pas d'ambiguïté, et une valeur de $\ln(n)$ représente le niveau maximal d'ambiguïté [172].

Étant donné que $A = \{x_1, x_2, \ldots, x_k\}$ représente l'ensemble des marqueurs de pertinence,

nous nous appuyons sur les marqueurs les moins ambigus, dont le niveau d'ambiguïté est inférieur ou égal à un seuil σ, pour entraîner un ADF [172]. Nous substituons la mesure classique de l'entropie de l'information par la métrique d'ambiguïté précédemment présentée.

Comme proposé par [172], le chemin de la racine aux feuilles est transformé en règles logiques simplifiées. Ces nouvelles règles représentent des relations logiques entre les marqueurs associés à la pertinence de l'information. Les marqueurs les moins ambigus sont ajoutés comme nouvelles sources d'information dans l'architecture de l'agent, enrichissant ainsi son processus de délibération.

Le processus de construction de l'ADF est une étape importante dans l'évaluation de la pertinence des marqueurs pour automatiser la tâche de classification prédictive des termes pertinents.

II.10 Cadre d'évaluation des marqueurs textuels dans les OE

Dans cette section, nous présentons les principaux marqueurs dérivés de l'application de la méthodologie précédemment proposée. Nous introduisons également un ensemble de marqueurs extraits de l'état de l'art en extraction d'information, qui sont, en principe, indépendants du domaine selon les expérimentations conduites dans les études respectives.

Ces marqueurs ont été définis sur la base de leur pertinence et de leur capacité à caractériser efficacement les informations contenues dans les OE.

II.10.1 Définition des marqueurs

L'algorithme YAKE ! [6] a introduit le concept de marqueurs textuels, bien que de manière informelle. Dans cette section, nous établissons une définition rigoureuse du concept de marqueurs textuels, tout en maintenant la compatibilité avec les marqueurs utilisés dans l'algorithme YAKE ! et ceux utilisés par les recruteurs en pratique.

Définitions préliminaires associées aux marqueurs textuels

Comme illustré précédemment, selon [178], un terme est défini comme une classe fonctionnelle d'unités lexicales utilisées dans le discours. Dans notre contexte spécifique, les termes sont identifiés sur la base de leur *termitude (termhood)* [218], qui est quantifiée en utilisant le *weirdness ratio (WR)* [178].

L'extraction de termes est effectuée en appliquant des motifs morphosyntaxiques spécifiques à divers corpus spécialisés [178]. La plupart des termes extraits sont des syntagmes nominaux.

Soit d_i une OE appartenant à un corpus C, et $T_{d_i} = t_1, t_2, \ldots, t_n$ l'ensemble des termes dans d_i.

Soit $R_{d_i} \subseteq T_{d_i}$ l'ensemble des termes les plus pertinents dans d_i. Chaque terme $t_i \in R_{d_i}$ est considéré comme pertinent avec un degré de possibilité α_{t_i}.

Soit $A_{d_i} = a_1, a_2, \ldots, a_m$ l'ensemble des sections dans d_i (description du poste, détails du profil, etc.). Chaque section a_i peut être représentée par un sous-ensemble de termes de T_{d_i}. Un terme peut appartenir à plusieurs sections.

Soit $E_{d_i} = e_1, e_2, \ldots, e_p$ un ensemble d'adjectifs et de qualificatifs liés à un sous-ensemble de termes de T_{d_i} par des dépendances syntaxiques.

Soit $O = o_1, o_2, \ldots, o_s$ l'ontologie-mère constituée par un ensemble d'ontologies (comme présenté dans la section II.5). Soit $c_{o_s} = c_{s,1}, c_{s,2}, \ldots, c_{s,k}$ l'ensemble des concepts dans l'ontologie o_s, et $T_{c_j} = t_{j,1}, t_{j,2}, \ldots, t_{j,l}$ l'ensemble des termes qui représentent lexicalement le concept c_j dans une langue donnée.

L'ontologie-mère O, englobe une variété de concepts, y compris la structure hiérarchique des documents textuels, dont les sections, paragraphes, phrases, syntagmes, termes, mots, morphèmes, etc. Cette structure fournit un cadre pour localiser et classer les termes et expressions dans le document.

Chacun de ces éléments joue un rôle fondamental dans le processus d'extraction, permettant une extraction plus complète, flexible et détaillée des informations les plus essentielles de chaque document, en fonction des spécificités du contexte organisationnel. De plus, la formalisation des marqueurs textuels YAKE ! avec ceux dérivés des annotations de recruteurs permet une complémentarité entre les marqueurs indépendants du contexte (en principe, YAKE !) et les marqueurs dépendants du contexte spécifiques au contexte organisationnel.

II.10.2 Marqueurs textuels

Avant d'introduire la formalisation, il est essentiel de souligner la distinction entre la valeur de vérité et le degré de possibilité. Chaque marqueur formalisé possède potentiellement un ensemble de conditions ou critères logiques qui déterminent sa valeur de vérité pour un terme donné. La valeur de vérité indique si un terme est pertinent ou non en fonction des conditions du marqueur. D'autre part, le degré de possibilité est un concept distinct qui indique le degré de plausibilité et de pertinence que l'agent attribue à la valeur de vérité du marqueur dans un contexte d'application spécifique [220]. Alors que la valeur de vérité représente un état binaire (c'est-à-dire vrai ou faux, terme pertinent contre terme non pertinent), le degré de possibilité exprime le niveau de confiance ou de certitude que l'agent attribue à cette valeur de vérité, en fonction des métriques de qualité contenues dans le module de confiance de l'agent.

Avec cette distinction à l'esprit, nous pouvons maintenant introduire la formalisation de seize marqueurs textuels. Les marqueurs de 1 à 10 ont été dérivés en observant le comportement des recruteurs experts et en interagissant directement avec eux, comme décrit dans la section II.5. Les marqueurs textuels de 11 à 16 correspondent à ceux utilisés dans l'approche d'extraction de

termes YAKE !

Cette formalisation raffinée des marqueurs textuels fournit un cadre pour évaluer et interpréter la pertinence des termes dans le contexte des OE. Elle permet un processus d'extraction de termes plus informatif qui est à la fois robuste et adaptable aux variations et à la complexité du contenu des OE. Les descriptions détaillées sont présentées ci-après.

Marqueur textuel #1 : présence de compétences professionnelles ou de types de postes (métier) dans les titres

Si un terme présent dans le titre correspond à l'un des termes utilisés pour représenter des compétences professionnelles ou des types de postes, il peut être considéré comme pertinent.

Soit $a_1 \in A_{d_i}$ la section titre de d_i. Soit $t_{a1} = t_1, t_2, \ldots, t_u$ l'ensemble des termes contenus dans a_1. T_{c_j} est l'ensemble des termes représentant lexicalement une compétence professionnelle ou un concept de type de poste c_j dans l'ontologie o_s. Nous postulons que :

$$\forall t_k \; \exists c_j \; [c_j \in o_s \wedge t_k \in T_{c_j} \wedge t_k \in t_{a1}] \rightarrow t_k \in R_{d_i} \tag{II.9}$$

avec un degré de possibilité $\alpha_{t_{k,1}} \in [0, 1]$.

Ce qui signifie que si le terme t_k est présent dans T_{c_j} et également dans la section titre t_{a1}, alors nous postulons que t_k appartient potentiellement à l'ensemble des termes pertinents R_{d_i} pour l'OE d_i.

Marqueur textuel #2 : termes représentant des compétences professionnelles dans une section de description de poste ou une section de description de profil

En général, un terme utilisé pour représenter une compétence professionnelle dans une section de description de poste ou dans une section de description de profil est plus susceptible d'être choisi comme terme pertinent.

Soient s_2 et s_3 les ensembles de termes dans la section de description du poste et la section de description du profil, respectivement. Soit $t_k \in T_{d_i}$. Soit T_{c_j} l'ensemble des termes représentant une compétence professionnelle ou un concept de type de poste c_j dans l'ontologie o_s. Nous postulons que :

$$\forall t_k \; \exists c_j \; [(t_k \in s_2 \vee t_k \in s_3) \wedge t_k \in T_{c_j}] \rightarrow t_k \in R_{d_i} \tag{II.10}$$

avec un degré de possibilité $\alpha_{t_{k,2}} \in [0, 1]$.

Cette formulation stipule que si le terme t_k est présent dans la section de description du poste (s_2) ou dans la section de description du profil (s_3) et qu'il est également représenté dans T_{c_j}, alors t_k devrait faire partie de l'ensemble de termes pertinents R_{d_i} pour la OE d_i.

Marqueur textuel #3 : pertinence des sections de l'OE

Les recruteurs choisissent avec un degré de possibilité plus élevé les termes utilisés dans le titre, la description du poste ou la description du profil, contrairement aux termes associés à d'autres sections, telles que la description de l'entreprise ou les détails du contrat.

Ce marqueur n'est pas redondant avec les marqueurs #1 et #2, puisque nous n'exigeons pas que les termes soient des concepts ontologiques (compétence professionnelle ou type de poste). Soit $S = s_1 \cup s_2 \cup s_3 \subseteq T_{d_i}$, où : s_1 est l'ensemble des termes de la section titre ; s_2 est l'ensemble des termes de la section de description du poste ; et s_3 est l'ensemble des termes de la section de description du profil.

Soit $t_m \in T_{d_i} \cap S$. Ensuite, nous demandons que :

$$\forall t_m \, \forall t_n \, (t_m \in T_{d_i} \wedge t_n \notin S) \rightarrow (P(t_m \in R_{d_i}) > P(t_n \in R_{d_i})) \tag{II.11}$$

avec un degré de possibilité $\alpha_{t_{k,3}} \in [0, 1]$. $P(t_* \in R_{d_i})$ représente la possibilité que t_* soit choisi comme terme pertinent.

Ce marqueur offre une dimension supplémentaire pour l'analyse des termes pertinents, en mettant l'accent sur l'importance des sections spécifiques de l'OE. Il s'agit d'une stratégie clé pour cibler les informations les plus pertinentes et pour comprendre les priorités des recruteurs lors de la rédaction du document.

Marqueur textuel #4 : termes impliqués dans des dépendances avec des expressions de pertinence

Un terme impliqué dans une relation de dépendance syntaxique avec un syntagme de l'OE est plus susceptible d'être choisi en tant que terme pertinent.

- Soit $t_k \in T_{d_i} \cap T_{c_j}$ pour un certain c_j ;
- nous définissons une "expression pertinente" e_m comme un syntagme utilisé par le recruteur responsable de la rédaction de l'OE (par exemple, *maîtriser C#*, *bonne connaissance de l'informatique en cloud*). Supposons que e_m manifeste une dépendance syntaxique avec t_k ;
- soit t_q un adjectif qualificatif ou un modificateur de nom qui précise, renforce ou qualifie e_m. Ce modificateur est syntaxiquement dépendant de e_m ;
- dans ce contexte, si un terme t_k est associé à un modificateur t_q à travers e_m, nous postulons alors que t_k est pertinent pour cette OE.

Par conséquent, nous stipulons que :

$$\forall t_k \; \exists e_m \; \exists t_q \; (t_k \in T_{d_i} \wedge e_m \in E_{d_i} \wedge t_q \in T_{d_i} \wedge \text{est\_qualificateur\_modificateur}(t_q, e_m) \wedge \text{est\_dependant}(e_m, t_k)) \rightarrow t_k \in R_{d_i} \quad \text{(II.12)}$$

avec un degré de possibilité $\alpha_{t_{k,4}} \in [0, 1]$.

Dans ce contexte, l'expression "est_qualificateur_modificateur" identifie si un élément, spécifiquement t_q (soit un adjectif qualificatif soit un modificateur de nom), qualifie ou modifie une expression de pertinence e_m dans le texte de l'OE. D'autre part, "est_dependant" vérifie si une expression pertinente e_m manifeste une dépendance syntaxique avec un terme t_k.

Le degré de possibilité $\alpha_{t_{k,4}} \in [0, 1]$ représente la plausibilité de cette proposition pour le terme t_k. Ainsi, une valeur de $\alpha_{t_{k,4}}$ plus élevée suggère une plus grande possibilité que t_k soit pertinent pour l'OE d_i.

Marqueur textuel #5 : termes utilisés dans les traces de descriptions d'activités professionnelles

Il y a une possibilité accrue pour qu'un terme représentant un concept professionnel soit perçu comme pertinent lorsque la description d'un emploi détaille une interaction avec le concept donné.

Nous caractérisons une trace de description d'activité professionnelle comme une phrase au sein d'une annonce d'emploi qui décrit l'action d'un travailleur sur un objet. Dans cette définition, b_j représente une trace spécifique de description d'activité professionnelle trouvée dans le document d_i. Cette trace est caractérisée par l'ensemble des termes T_{b_j}. Nous exigeons que b_j contienne au minimum un verbe ou une phrase verbale et un objet dépendant. Nous supposons que les termes t_k représentant ces objets auront plus de chances d'être sélectionnés comme pertinents. Par conséquent :

$$\forall t_k \; (t_k \in T_{b_j} \wedge \text{est\_objet}(t_k, b_j)) \rightarrow t_k \in R_{d_i} \quad \text{(II.13)}$$

avec un degré de possibilité $\alpha_{t_{k,5}} \in [0, 1]$.

Dans cette formulation, "est_objet" est une fonction qui évalue si un terme donné t_k est un objet dans une trace de description d'activité professionnelle b_j. Si un terme t_k se trouve à la fois dans l'ensemble de termes de la trace T_{b_j} et est un objet dans cette trace, alors nous proposons que t_k devrait être considéré comme un terme pertinent pour cette OE.

Marqueur textuel #6 : termes représentant des compétences/activités professionnelles à haut risque

Les compétences et activités qui ont un potentiel plus élevé d'impacts négatifs significatifs sur l'activité économique d'une entreprise en raison des erreurs des employés sont considérées comme plus pertinentes.

Ce marqueur vise à attribuer une plus grande pertinence aux termes représentant des compétences ou activités professionnelles pour lesquelles une erreur de l'employé *peut nuire considérablement à l'activité économique de l'entreprise*. Une valeur de 0 signifie qu'une erreur potentielle n'aura pas d'impact significatif sur l'activité économique. En revanche, une valeur de 1 indique qu'une erreur dans cette compétence ou activité aura possiblement un effet négatif considérable.

Selon le contexte spécifique de chaque organisation, ce marqueur textuel permet à l'agent d'adapter son comportement pour répondre aux exigences particulières de chaque poste vacant.

Soit M une ontologie contenant l'ensemble des compétences et activités professionnelles pour une entreprise spécifique. M englobe une collection de concepts $c_M = c_{M,1}, c_{M,2}, \ldots, c_{M,p}$. Le recruteur attribue manuellement un niveau de risque $\epsilon_{c_{M,p}} \in [0,1]$ à chaque compétence ou activité professionnelle.

Soit t_k un terme dans une OE d_i, représentant une compétence ou une activité professionnelle dans M. Parmi les concepts liés à t_k, soit $c_{M,l}$ le concept présentant le niveau de risque maximal. Si ce niveau de risque dépasse un seuil spécifié $\beta_{c_{M,l}}$, alors t_k est choisi comme un terme pertinent et :

$$\forall t_k \; \exists c_{M,l} \; (t_k \in T_{d_i} \wedge c_{M,l} \in M \wedge t_k \in T_{c_{M,l}} \wedge \text{est\_superieur\_a}(\epsilon_{c_{M,l}}, \beta_{c_{M,l}}) \rightarrow t_k \in R_{d_i} \quad \text{(II.14)}$$

avec un degré de possibilité $\alpha_{t_{k,6}} \in [0,1]$. Ici, "est_superieur_a" est une fonction qui évalue si le niveau de risque $\epsilon_{c_{M,l}}$ d'un concept professionnel spécifique $c_{M,l}$ est supérieur à un seuil de risque spécifié $\beta_{c_{M,l}}$. Si le niveau de risque est supérieur à ce seuil, alors le terme t_k est considéré comme un terme pertinent pour la OE.

Marqueur textuel #7 : actions communiquées dans les OE de management

Il est également important pour les recruteurs d'identifier le type d'actions requises par les OE de management.

Par exemple, certains postes de management sont axés sur la gestion d'équipe, tandis que d'autres englobent des activités à responsabilité ou des tâches de développement. Cette variabilité s'explique par la diversité de tâches qui peuvent être demandées dans ce type de poste.

Soit d_i une OE de management. L'agent détecte les OE de management en se basant sur un modèle d'Allocation de Dirichlet Latente, entraîné sur 14 000 curricula vitæ. Soit t_k un terme verbal de d_i. Si t_k appartient à la trace d'une activité professionnelle f_j et est la racine de

l'arbre syntaxique de la phrase, alors c'est potentiellement un terme pertinent. Nous définissons ce marqueur comme suit :

$$\forall t_k \; \exists f_j \; (f_j \in d_i \wedge t_k \in f_j \wedge \text{est\_management}(d_i) \wedge \text{est\_verbe}(t_k) \wedge \text{est\_racine\_de}(t_k, f_j)) \rightarrow t_k \in R_{d_i} \quad \text{(II.15)}$$

avec un degré de possibilité $\alpha_{t_{k,7}} \in [0, 1]$.

Ici, "est_gestion" est une fonction qui évalue si une annonce d'emploi d_i est liée à un poste de gestion, "est_verbe" évalue si un terme t_k est un verbe, et "est_racine_de" évalue si ce terme est la racine de l'arbre syntaxique de la trace d'une activité professionnelle f_j.

Marqueur textuel #8 : similarité sémantique BERT des compétences professionnelles

Des termes spécifiques utilisés pour représenter des compétences professionnelles qui présentent une proximité sémantique (au sens de BERT) avec des termes pertinents précédemment identifiés seront considérés comme pertinents.

Soit $t_1 \in R_{d_i}$ et $t_2 \in T_{d_i}$. Nous définissons la fonction WR de spécificité $f(t)$ d'un terme comme sa fréquence relative dans un corpus spécifique C_s, divisée par sa fréquence relative dans un corpus multilingue général C_L [178].

De plus, nous définissons $g(t_1, t_2)$ comme la similarité sémantique BERT entre deux termes. Le modèle BERT peut être pré-entraîné sur des corpus de connaissances dans le domaine des compétences professionnelles ou sur des connaissances associées. Ce marqueur est formalisé comme suit :

$$\forall t_1 \; \forall t_2 \; (t_1 \in R_{d_i} \wedge g(t_1, t_2) > 0) \rightarrow t_2 \in R_{d_i} \quad \text{(II.16)}$$

avec un degré de possibilité défini par l'équation normalisée :

$$\alpha_{t_2,8} = \|(1 - \alpha_{t_1}) * g(t_1, t_2) * f(t_2))\|. \quad \text{(II.17)}$$

Ici, $g(t_1, t_2)$ représente la similarité sémantique entre deux termes t_1 et t_2, et $f(t)$ est la fonction de spécificité d'un terme. Le facteur de pondération α_{t_1}, associé au terme t_1, aide à réguler le degré résultant de pertinence attribué à t_2, en l'augmentant ou en le diminuant. L'équation II.17 définit le degré de possibilité qu'un terme t_2 soit pertinent, basé sur la similarité sémantique BERT avec un terme déjà identifié comme pertinent t_1 et la spécificité de t_2.

Marqueur textuel #9 : pertinence du secteur d'activité économique

Les termes indiquant un secteur d'activité économique requis par une OE (par exemple, finance, banque, aéronautique, etc.) seront considérés comme potentiellement pertinents.

Cela implique que :

$$\forall t_k \ (t_k \in T_{d_i} \wedge \text{est\_secteur\_requis}(t_k)) \rightarrow t_k \in R_{d_i} \tag{II.18}$$

avec un degré de possibilité $\alpha_{t_{k,9}} \in [0, 1]$.

Les secteurs d'activité économique sont reconnus en alignant les termes des OE avec les étiquettes de concepts d'activité économique, fournies par les référentiels ESCO, O*NET, ROME et CIGREF.

En bref, ce marqueur souligne l'importance d'une compréhension sectorielle dans l'interprétation des OE, car elle peut fournir un contexte précieux pour le poste et les compétences requises.

Marqueur textuel #10 : prérequis de compétences professionnelles

Les prérequis de compétences professionnelles identifiés comme pertinents ont généralement eux-mêmes une grande pertinence.

Supposons qu'il existe une *relation de prérequis* entre deux compétences professionnelles, c_1 et c_2, dans une certaine ontologie o_i. Ces relations peuvent être extraites d'ontologies comme celle de l'ESCO. Si c_2 est un prérequis pour c_1 et si c_1 est jugée pertinente (sous réserve d'un certain degré de possibilité), alors c_2 hérite du degré de possibilité de c_1.

Exprimée formellement, cela donne :

$$\forall t_1 \ \forall t_2 \ \exists c_1 \ \exists c_2 \ (c_1 \in o_i \wedge c_2 \in o_i \wedge t_1 \in T_{c_1} \wedge t_2 \in T_{c_2} \wedge \text{est\_prérequis}(c_1, c_2) \wedge t_1 \in R_{d_i}) \rightarrow t_2 \in R_{d_i} \tag{II.19}$$

avec un degré de possibilité $\alpha_{t_{k,10}} \in [0, 1]$. Ce degré de possibilité est égal au degré de possibilité de $t_1 \in R_{d_i}$. Ainsi, dans le domaine complexe des compétences professionnelles, une relation de prérequis peut ajouter une couche supplémentaire de contexte et de pertinence lors de l'évaluation des termes dans une OE. En considérant ces relations, nous pouvons mieux comprendre les exigences inhérentes à un poste et donc offrir une analyse plus précise et approfondie.

Marqueur textuel #11 : caractères en majuscules selon YAKE !

Les termes d'une OE écrits en majuscules (y compris les abréviations) ont tendance à être pertinents.

Dans notre contexte, ce marqueur YAKE ! est en corrélation avec le comportement des recruteurs, qui ont fréquemment tendance à mettre en majuscules les termes associés aux compétences professionnelles :

$$\forall t_k \ (t_k \in T_{d_i} \wedge \text{est\_en\_majuscule}(t_k)) \rightarrow t_k \in R_{d_i}. \tag{II.20}$$

Nous définissons le degré de possibilité de cette règle en fonction de l'équation de YAKE! normalisée :

$$\alpha_{t_{k,11}}(t_k) = \left\| \frac{\max(\mathrm{TF}(U(t_k), \mathrm{TF}(A(t_k))}{\ln(\mathrm{TF}(t_k))} \right\|, \tag{II.21}$$

où $\mathrm{TF}(U(t_k))$ est le nombre de fois que t_k apparaît en majuscules, $\mathrm{TF}(A(t_k))$ est le nombre d'occurrences de t_k en tant qu'acronyme (pour plus de détails, voir [6]) et $\mathrm{TF}(t_k)$ est la fréquence du terme.

Ce marqueur souligne le fait que la mise en forme textuelle, y compris l'usage des majuscules et des abréviations, peut jouer un rôle important dans la détermination de la pertinence des termes dans une OE. Il ajoute une nuance supplémentaire à notre compréhension de la manière dont les recruteurs communiquent l'importance relative des différentes compétences nécessaires pour un poste.

Marqueur textuel #12 : position du terme selon YAKE!

Les termes apparaissant au début du document ont une plus grande pertinence.

Ce marqueur émet l'hypothèse que les termes apparaissant au début du document ont tendance à être plus pertinents.

$$\forall t_k \ (t_k \in T_{d_i} \wedge \mathrm{est\_position\_marqueur\_activé}(t_k)) \rightarrow t_k \in R_{d_i}. \tag{II.22}$$

Le degré de possibilité est donné par l'équation normalisée de YAKE! suivante :

$$\alpha_{t_{12}}(t_k) = \| \ln(\ln(3 + \mathrm{Médiane}(\mathrm{Sent}(t_k)))) \|, \tag{II.23}$$

où $\mathrm{Sent}(t_k)$ est l'ensemble des positions des phrases contenant t_k.

Ce marqueur met en évidence l'importance de la position des termes dans un document pour déterminer leur pertinence. Il suggère que les termes positionnés en début de document, et donc potentiellement mis en avant par les recruteurs, sont souvent plus pertinents pour comprendre le contenu et les exigences de l'OE.

Marqueur textuel #13 : normalisation de la fréquence des termes selon YAKE!

Les termes ayant une fréquence d'utilisation plus élevée dans l'OE ont tendance à être pertinents.

Ainsi, nous avons :

$$\forall t_k (t_k \in T_{d_i} \wedge \mathrm{est\_marqueur\_de\_fréquence\_activé}(t_k)) \rightarrow t_k \in R_{d_i} \tag{II.24}$$

avec un degré de possibilité donné par l'équation normalisée suivante fournie par YAKE ! :

$$\alpha_{t_{k,13}}(t_k) = \left\| \frac{\mathrm{TF}(t_k)}{\mathrm{MoyenneTF} + \sigma} \right\|, \tag{II.25}$$

où $\mathrm{TF}(t_k)$ est le nombre d'occurrences de t_k, qui est divisé par la moyenne et l'écart-type de la fréquence.

Marqueur textuel #14 : pertinence du terme par rapport au contexte selon YAKE !

Ce marqueur YAKE ! repose sur l'hypothèse suivante : *la pertinence d'un terme candidat t diminue à mesure que le nombre de termes différents avec lesquels il co-occurre des deux côtés augmente* :

$$\forall t_k \ (t_k \in T_{d_i} \wedge \mathrm{est\_marqueur\_pertinence\_activé}(t_k)) \rightarrow t_k \in R_{d_i} \tag{II.26}$$

avec un degré de possibilité obtenu à partir de l'équation normalisée de YAKE ! :

$$\alpha_{t_{k,14}} = \left\| 1 + (DL + DR + \cdots) * \frac{\mathrm{TF}(t_k)}{\max \mathrm{TF}} \right\|, \tag{II.27}$$

où

$$DL[DR] = \frac{|A_{t,w}|}{\sum\limits_{k \in A_{t,w}} \mathrm{CoOccur}_{t,k}}. \tag{II.28}$$

$DL[DR]$ indique que l'équation s'applique à la fois pour estimer la dispersion des termes à gauche du terme t (DL) ainsi que pour estimer la dispersion des termes à droite de t (DR). $|A_{t,w}|$ représente le nombre de termes différents dans une fenêtre de taille w (les unités de w étant exprimées comme le nombre de termes de chaque côté), et TF est la fréquence du terme. Les points de suspension '...' de l'équation II.27 indiquent que d'autres éléments contextuels peuvent être intégrés.

Cette règle suggère que si un terme apparaît fréquemment aux côtés d'une variété de termes différents, il peut avoir une signification moins spécifique et donc être considéré comme moins pertinent. C'est une manière intuitive de mesurer la pertinence des termes en fonction de leur contexte d'apparition.

Marqueur textuel #15 : sentences différentes selon YAKE !

La pertinence d'un terme est positivement corrélée à la fréquence de son utilisation à travers des phrases distinctes, représentée comme suit :

$$\forall t_k \ (t_k \in T_{d_i} \wedge \mathrm{est\_marqueur\_de\_phrases\_activé}(t_k)) \rightarrow t_k \in R_{d_i} \tag{II.29}$$

avec un degré de possibilité obtenu à partir de l'équation normalisée :

$$\alpha_{t_{k,15}} = \left\| \frac{SF(t_k)}{\#\text{Sentences}} \right\|, \tag{II.30}$$

où $SF(t_k)$ est le nombre de phrases contenant t_k et #Sentences est le nombre total de phrases de d_i.

Ce marqueur textuel suggère que si un terme est utilisé fréquemment à travers différentes phrases dans une description de poste, il est probablement pertinent pour comprendre les exigences du poste. Cela reflète l'idée que les concepts clés sont souvent répétés de différentes manières pour souligner leur importance.

Marqueur textuel #16 : score global YAKE !

Nous incluons le score de pertinence global proposé par YAKE ! basé sur les marqueurs #11, #12, #13, #14, et #15 [6]. Soit $t_k \in d_i$. Un terme est considéré comme « partiellement pertinent » s'il est ainsi prédit par le score global :

$$\forall t_k \ (t_k \in T_{d_i} \wedge \text{est\_prédit\_par\_yake}(t_k)) \rightarrow t_k \in R_{d_i} \tag{II.31}$$

avec un degré de possibilité $\alpha_{t_{k,16}} \in [0, 1]$.

En combinant ces différents marqueurs, YAKE ! fournit une mesure de pertinence qui prend en compte plusieurs aspects du texte, y compris la fréquence des termes, leur position dans le texte et leur relation avec le contexte environnant. Cette mesure globale vise à fournir une évaluation robuste de la pertinence des termes pour une description de poste donnée.

II.10.3 Marqueurs supplémentaires

Dans cette section, nous introduisons des marqueurs supplémentaires qui servent à démontrer l'applicabilité du cadre proposé pour l'évaluation des marqueurs, qui sera détaillé dans la section suivante et évalué dans le cas d'étude 3 du Chapitre III. Parmi les marqueurs supplémentaires figurent des marqueurs composés (#17-20) issus de l'ADF (section II.9.8). Il s'agit de compositions logiques élaborées à partir des marqueurs dérivés du contexte et des marqueurs YAKE ! Ils ont été sélectionnés en raison de leur performance notable vis-à-vis des marqueurs individuels sous-jacents.

Par ailleurs, un marqueur obtenu par la technique ACP (#21) est également inclus. Ce marqueur, issu de l'application de l'ACP sur les marqueurs dérivés du contexte (#1-10), offre une synthèse de ces derniers. Enfin, neuf "faux" marqueurs sont inclus pour illustrer davantage l'utilité de l'approche proposée (#22-30). Ces marqueurs représentent des logiques aléatoires de pertinence de l'information, élaborées à partir de concepts mathématiques et linguistiques

variés. Ils sont présentés dans les tableaux II.3, II.4, et II.5.

II.10.4 Évaluation des marqueurs textuels

Pour que l'agent exécute des actions à la fois pertinentes et explicables, la sélection adéquate de marqueurs textuels essentiels est indispensable. Ces marqueurs, qui servent de sources de croyances pour l'architecture de l'agent, influencent directement son processus de raisonnement approximatif. Ainsi, garantir leur pertinence est nécessaire pour optimiser la prise de décision de l'agent relative à la pertinence des termes dans divers scénarios.

Notre approche d'évaluation de marqueurs commence par un ensemble de marqueurs textuels supposés être associés à la pertinence de l'information dans les OE, ainsi qu'un corpus de documents pour les évaluer. La première étape de l'évaluation consiste à estimer les corrélations entre les marqueurs étudiés afin d'approfondir leur explicabilité tout en garantissant la stabilité du processus d'évaluation dans les étapes ultérieures.

Dans la deuxième étape, nous menons une analyse de la fréquence d'activation des marqueurs sur le corpus OE, permettant une illustration graphique des marqueurs sur-activés ou sous-activés par rapport aux annotations des recruteurs.

Dans la troisième étape, nous déterminons la précision de chaque marqueur telle que mesurée dans un modèle AA traditionnel, identifiant leur capacité à classer correctement les termes pertinents et non pertinents dans les OE.

Cela sert de base pour la quatrième étape, où nous effectuons une analyse parallèle de l'ambiguïté des marqueurs et de l'entropie de l'information mutuelle. Cela nous permet d'estimer et d'analyser l'incertitude associée à chaque caractéristique à la fois d'un point de vue de l'ambiguïté et de l'entropie de l'information mutuelle.

Dans la cinquième étape, notre approche intègre une étude prospective-statistique des marqueurs en utilisant un modèle simple et hautement explicatif, comme la régression logistique, pour confirmer la qualité de leur comportement sur une plus grande échelle de données. Enfin, nous identifions les marqueurs ayant un niveau minimum d'explicabilité en associant les niveaux de possibilité fournis par les marqueurs étudiés avec les stratégies d'annotation observées des recruteurs sur les OE.

Les résultats des étapes précédentes servent de données d'entrée pour l'exécution d'un moteur d'inférence de type Mamdani (annexe I). Basé sur des règles logiques qui délimitent les critères de qualité minimale pour la pertinence dans le contexte organisationnel des recruteurs, ce moteur identifie un ensemble de marqueurs plus pertinents pour l'extraction automatique de termes pertinents dans les OE. Ces marqueurs sont ensuite implémentés sur l'architecture de l'agent CDI pour optimiser ses performances. La Figure II.8 illustre les principales étapes incluses dans la méthodologie proposée.

TABLE II.3 – Description des marqueurs textuels #17 - #21 dérivés des marqueurs contextuels (#1 - #10).

Marq. #	**Description**	**Formule**	**Degré de Possibilité**
M_{17}	Compétences professionnelles dans la section de description de poste ou de profil de l'OE, liées aux Expressions de Pertinence	Ce marqueur est défini comme la conjonction entre les marqueurs M_2 et M_4 : $\forall t_k(F_{M_2}(t_k) \wedge F_{M_4}(t_k)) \rightarrow t_k \in R_{d_i}$ où F_{M_i} représente l'ensemble des conditions logiques associées aux marqueurs M_2 et M_4.	$\alpha_{t_{k,17}}(t_k) = min(\alpha_{t_{k,2}}(t_k), \alpha_{t_{k,4}}(t_k))$
M_{18}	Termes en majuscules ou avec la première lettre en majuscule dans les sections pertinentes de l'OE	$\forall t_k(F_{M_3}(t_k) \wedge F_{M_{11}}(t_k)) \rightarrow t_k \in R_{d_i}$ où F_{M_i} représente l'ensemble des conditions logiques associées aux marqueurs M_3 et M_{11}.	$\alpha_{t_{k,18}}(t_k) = min(\alpha_{t_{k,3}}(t_k), \alpha_{t_{k,11}}(t_k))$
M_{19}	Compétences professionnelles spécifiques en majuscules ou avec la première lettre en majuscule	$\forall t_k(F_{M_2}(t_k) \wedge F_{M_8}(t_k) \wedge F_{M_{11}}(t_k)) \rightarrow t_k \in R_{d_i}$ où F_{M_i} représente l'ensemble des conditions logiques associées aux marqueurs M_2, M_8 et M_{11}.	$\alpha_{t_{k,19}}(t_k) = min(\alpha_{t_{k,2}}(t_k), \alpha_{t_{k,8}}(t_k), \alpha_{t_{k,11}}(t_k))$
M_{20}	Compétences professionnelles associées à divers motifs textuels liés aux stratégies des recruteurs sur la pertinence de l'information	$\forall t_k(F_{M_1}(t_k) \vee F_{M_2}(t_k) \vee F_{M_4}(t_k) \vee F_{M_9}(t_k)) \rightarrow t_k \in R_{d_i}$ où F_{M_i} représente l'ensemble des conditions logiques associées aux marqueurs M_1, M_2, M_4 et M_9.	$\alpha_{t_{k,20}}(t_k) = max(\alpha_{t_{k,1}}(t_k), \alpha_{t_{k,2}}(t_k), \alpha_{t_{k,4}}(t_k), \alpha_{t_{k,9}}(t_k))$
M_{21}	Méthode ACP sur les marqueurs contextuels	Nous avons mis en œuvre la méthode d'ACP sur les marqueurs dérivés du contexte ($M_1 - M_{10}$) pour les réduire à une seule dimension. Ce marqueur est défini comme : $\forall t_k(t_k \in T_{d_i} \wedge \text{est\_acp\_activé}(t_k)) \rightarrow t_k \in R_{d_i}$	$\alpha_{t_{k,21}}(t_k) = \|\|\rho_{t_k}\|\|$, où ρ_{t_k} est la valeur ACP normalisée du terme t_k

TABLE II.4 – Description des marqueurs textuels #22 - #26 correspondant à des "faux" marqueurs.

Marq. #	**Description**	**Formule**	**Degré de Possibilité**
M_{22}	Le niveau de pertinence d'un terme suit une distribution uniforme aléatoire	Ce marqueur est défini comme : $\forall t_k(t_k \in T_{d_i} \wedge \text{est\_marqueur\_aléatoire\_activé}(t_k)) \rightarrow t_k \in R_{d_i}$	$\alpha_{t_{k,22}}(t_k) = x_{t_k}$, où $x_{t_k} \sim U(0,1)$
M_{23}	Valeur absolue de la fonction cosinus évaluée à la dernière position du terme dans l'OE	Nous exécutons la requête suivante : $\forall t_k(t_k \in T_{d_i} \wedge \text{est\_marqueur\_cosin\_activé}(t_k)) \rightarrow t_k \in R_{d_i}$ Soit l_{t_k} la dernière position du terme t_k dans le texte de l'OE.	$\alpha_{t_{k,23}}(t_k) = \lvert cos(l_{t_k})\rvert$
M_{24}	Le terme contient-il un nombre pair de noms ?	Soit n_{t_k} une fonction qui retourne le nombre de noms dans un terme de l'OE. Ce marqueur est défini comme suit : $\forall t_k(t_k \in T_{d_i} \wedge est\_pair(n(t_k))) \rightarrow t_k \in R_{d_i}$	$\alpha_{t_{k,24}} \in [0,1]$
M_{25}	Au moins une des positions du terme dans le texte correspond à un nombre premier	Soit $P_{t_k} = p_{1,t_k}, p_{2,t_k}, \ldots, p_{i,t_k}$ l'ensemble des positions du terme t_k dans le texte de l'OE. Nous exécutons la requête suivante : $\forall t_k \exists p_{i,t_k}(t_k \in T_{d_i} \wedge p_{i,t_k} \in P_{t_k} \wedge est\_premier(p_{i,t_k})) \rightarrow t_k \in R_{d_i}$	$\alpha_{t_{k,25}} \in [0,1]$
M_{26}	Dans la terminologie de l'OE triée alphabétiquement, la position du terme est-elle un nombre parfait ?	Soit $B_{di} = (t_{1,*}, t_{2,*}, \ldots, t_{k,*})$ un n-uplet contenant les termes de l'OE triés alphabétiquement. Le n-uplet $P_{d_i} = (p_{t_{1,*}}, p_{t_{2,*}}, \ldots, p_{t_{k,*}})$ représente leurs positions respectives numérotées de 1 à k. Nous lançons la requête suivante : $\forall t_k \exists t_{k,*} \exists p_{t_{k,*}}(t_k \in T_{d_i} \wedge fait\_référence\_au\_meme\_terme(t_{k,*}, t_k) \wedge$ $est\_une\_position\_de(p_{t_{k,*}}, t_{k,*}) \wedge est\_nombre\_parfait(p_{t_{k,*}})) \rightarrow t_k \in R_{d_i}$	$\alpha_{t_{k,26}} \in [0,1]$

TABLE II.5 – Description des marqueurs textuels #27 - #30 correspondant à des "faux" marqueurs.

Marq. #	**Description**	**Formule**	**Degré de Possibilité**
M_{27}	Le terme contient la lettre "x"	$\forall t_k(t_k \in T_{d_i} \wedge contient(t_k, "x")) \rightarrow t_k \in R_{d_i}$	$\alpha_{t_{k,27}} \in [0,1]$
M_{28}	Le nombre de fois où le terme est utilisé pour créer de nouvelles variantes de terme	Soit $g(t_k)$ une fonction qui retourne le nombre de variantes de terme [178] associées au terme t_k. Ce marqueur est défini comme suit : $\forall t_k(t_k \in T_{d_i} \wedge est\_marqueur\_variantes\_termes\_activé(t_k)) \rightarrow t_k \in R_{d_i}$	$\alpha_{t_{k,28}}(t_k) = \|g(t_k)\|$
M_{29}	Le terme est la racine de son arbre syntaxique associé	Soit $S_{d_i} = \{s_1, s_2, \ldots, s_n\}$ l'ensemble des phrases dans l'OE. Pour chaque phrase s_i, au moins un arbre syntaxique $a_{s_i,j}$ lui est associé. Ce marqueur est défini comme suit : $\forall t_k \exists s_i \exists a_{s_i,j}(t_k \in T_{d_i} \wedge t_k \in s_i \wedge est\_racine(t_k, s_i)) \rightarrow t_k \in R_{d_i}$	$\alpha_{t_{k,29}} \in [0,1]$
M_{30}	Le terme contient au moins une lettre majuscule (n'importe où dans le terme)	$\forall t_k(t_k \in T_{d_i} \wedge contient\_lettre\_majuscule(t_k)) \rightarrow t_k \in R_{d_i}$	$\alpha_{t_{k,30}} \in [0,1]$

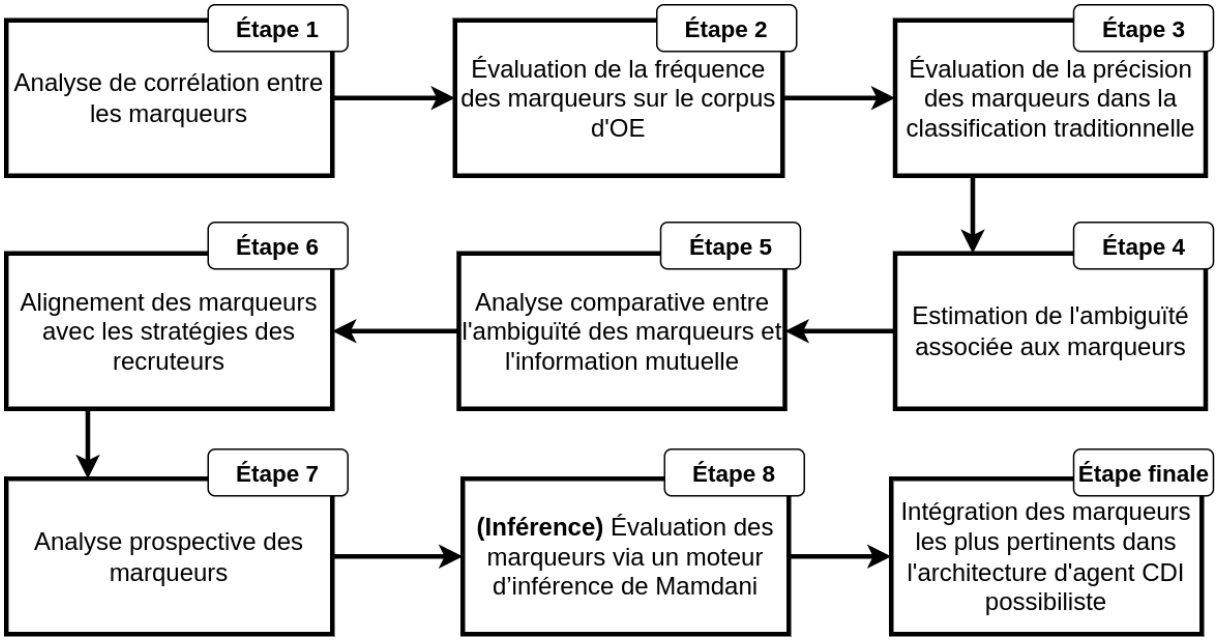


FIGURE II.8 – Description générale de l'approche d'évaluation des marqueurs textuels proposée.

II.10.5 Analyse de la corrélation entre marqueurs textuels

L'exploitation du coefficient de corrélation de Pearson permet une première approche analytique des relations linéaires potentielles entre les marqueurs étudiés et les annotations des recruteurs, agissant comme un outil statistique vers l'explicabilité.

En effet, l'importance de l'évaluation de la corrélation entre les marqueurs pour l'explicabilité réside dans sa capacité à valider et à affiner le processus d'extraction d'information. Une forte corrélation entre deux marqueurs peut suggérer un chevauchement dans la manière dont ces marqueurs identifient les informations, nécessitant une clarification pour assurer la transparence du processus. De plus, en évaluant la corrélation des marqueurs avec les annotations des recruteurs, nous alignons notre processus d'extraction sur les perceptions humaines, renforçant ainsi sa validité et son explicabilité. Enfin, cette approche basée sur la corrélation garantit que des marqueurs plus pertinents sont utilisés, optimisant ainsi le modèle qui les exploite et rendant le processus d'extraction plus compréhensible pour les utilisateurs potentiels.

Ainsi, considérons un ensemble d'OE, d_i, qui soient préalablement annotées, chacune contenant un assortiment de termes et leurs valeurs de pertinence respectives telles qu'établies par les recruteurs (R).

Considérons M_x et M_y, deux marqueurs textuels pour lesquels nous cherchons à estimer le

degré de corrélation. Le coefficient de corrélation de Pearson r_{xy} est alors défini par :

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \tag{II.32}$$

où x_i désigne le degré de possibilité offert par le marqueur M_x pour le terme t_i et y_i représente le degré de possibilité offert par le marqueur M_y pour le même terme ; $\bar{x}$ et $\bar{y}$ sont la moyenne des degrés de possibilité fournis par M_x et M_y respectivement. Enfin, s_x et s_y désignent les écarts-types des degrés de possibilité de M_x et M_y, et n représente le nombre de termes dans le corpus des OE étudiées.

À la fin de cette phase, qui nécessite une évaluation systématique des corrélations entre les marqueurs et les annotations des recruteurs, nous obtenons un ensemble de paires de marqueurs (m_i, m_j) dont le niveau de corrélation dépasse un niveau de tolérance $\omega_{r_{xy}}$. Ces couples de marqueurs facilitent une analyse plus claire et plus pertinente des marqueurs, garantissant la stabilité des modèles susceptibles de réagir sensiblement à l'utilisation de variables fortement corrélées [174], comme c'est le cas avec la régression logistique que nous appliquons dans la section II.10.11.

II.10.6 Analyse préliminaire des degrés de possibilité cumulés ou de la fréquence des marqueurs dans le corpus des OE

L'identification des marqueurs qui nécessitent une attention ou une évaluation accentuée est essentielle pour assurer la transparence et l'explicabilité du processus d'extraction. Dans ce contexte, pour identifier de manière préliminaire les marqueurs qui ont tendance à être suractivés ou sous-activés dans le corpus des OE étudié, nous estimons les degrés de possibilité cumulés ou la fréquence pour chaque marqueur évalué. Soit M_x un marqueur textuel fournissant un degré de possibilité x_i pour le terme t_i dans l'OE. Nous définissons le degré de possibilité cumulé ou la fréquence d'un marqueur comme suit :

$$l_{M_x} = \log(1 + \sum_{i=1}^{n} x_i). \tag{II.33}$$

Les marqueurs présentant des valeurs d'activation faibles par rapport aux annotations des recruteurs (R) peuvent potentiellement représenter des marqueurs très spécifiques, décrivant des détails particuliers des annotations. Inversement, ils peuvent également être des marqueurs inappropriés pour la tâche d'extraction de termes pertinents à partir des OE. Réciproquement, les marqueurs ayant des niveaux d'activation plus élevés que les annotations des recruteurs peuvent potentiellement correspondre à des marqueurs utiles pour identifier des ensembles étendus de termes qui contiennent potentiellement des termes pertinents. Ceci peut également indiquer un marqueur inapproprié pour la tâche d'extraction qui tend à détecter faussement un grand

nombre de termes comme étant pertinents.

II.10.7 Détermination du niveau de précision du marqueur pour décrire chaque classe dans un processus de classification traditionnel

La troisième étape implique l'évaluation de chaque marqueur comme un modèle AA en soi et l'évaluation de ses performances dans un processus de classification traditionnel. En supposant que le processus de classification se compose des classes $C_1, C_2, \ldots, C_n$ (dans notre contexte $n = 2$, termes pertinents et non pertinents), la précision du marqueur sur un terme donné t_i est définie comme suit :

$$p(M_x, t_i) = \begin{cases} 1 & \text{si } M_x(t_i) > 0.5 \\ 0 & \text{si } M_x(t_i) \leq 0.5. \end{cases} \tag{II.34}$$

Et la précision moyenne du marqueur sur l'ensemble de termes échantillonnés, pour décrire la classe C_n, est définie comme suit :

$$P(C_n|M_x) = \frac{\sum_{i=1}^{n} p(M_x, t_i)}{n}, \tag{II.35}$$

où n représente le nombre de termes des OE associés à la classe C_n dans le corpus d'étude.

Nous appellerons la précision d'un marqueur sur l'ensemble des termes annotés comme pertinents par les recruteurs, précision positive, et la précision d'un marqueur sur l'ensemble des termes annotés comme non pertinents, précision négative. L'interprétation de cette étape d'évaluation est faite en fonction de quatre types essentiels de marqueurs :

- *type 1 – haute précision positive et haute précision négative :* cette catégorie correspond à un marqueur idéal qui permet une description très *précise* des termes pertinents et non pertinents tels qu'annotés par les recruteurs ;
- *type 2 – haute précision positive et faible précision négative :* bien qu'une faible précision négative tende à être inacceptable dans le contexte de l'extraction de termes pour les OE (étant donné la nature concise des documents), un marqueur de ce type pourrait éventuellement servir de filtre préliminaire pour les termes pertinents ; appliqué à l'OE, il pourrait potentiellement réduire le nombre de termes non pertinents tout en garantissant que la majorité des termes pertinents sont conservés ;
- *type 3 – faible précision positive et haute précision négative :* en principe, un marqueur de cette catégorie correspond à un contre-marqueur de pertinence capable d'identifier les termes non pertinents et d'expliquer certains aspects des termes pertinents ; une autre interprétation possible est que le marqueur n'est pas vraiment pertinent et montre une haute précision négative en raison du déséquilibre des classes (beaucoup plus de termes non pertinents que de termes pertinents), ce qui est inhérent au processus de sélection des termes sur les OE ;

— *type 4 – faible précision positive et faible précision négative :* un type de marqueur qui ne fournit aucune information sur les annotations des recruteurs et est donc très probablement non pertinent.

Après avoir identifié comment les marqueurs se comportent sur les deux classes de termes dans le corpus des OE, la question suivante à examiner concerne la manière dont la distribution de la pertinence de leurs termes se compare à la distribution de la pertinence des termes observée dans les annotations des recruteurs. Pour ce faire, la mesure d'ambiguïté est appliquée à l'étape suivante.

II.10.8 Estimation de l'ambiguïté associée aux marqueurs

Dans la section II.9.8, nous nous sommes penchés sur le processus d'évaluation de l'ambiguïté, dans le but d'estimer l'incertitude cognitive liée à chaque marqueur lors de la sélection des termes par les recruteurs. S'appuyant sur le travail fondateur de Yuan [172], qui a abordé les tâches de classification comme des actions rationnelles basées sur la connaissance du décideur, nous examinons en amont les incertitudes inhérentes auxquelles les recruteurs sont confrontés lors de l'annotation des termes pertinents.

Pour rappeler rapidement :

— l'ensemble des termes dans une OE spécifique est représenté comme $U = \{t\}$;
— nous introduisons C comme un ensemble flou symbolisant les niveaux de pertinence perçus par le recruteur pour les termes dans une OE, associés à l'aide de la fonction d'appartenance μ_C ;
— cet ensemble flou est partitionné en C_1 (termes pertinents) et C_2 (termes non pertinents), tous deux définis à l'aide de fonctions triangulaires ; les sous-ensembles résultants des annotations du recruteur sont désignés par R, R_1 et R_2 ;
— chaque terme t peut être caractérisé à l'aide des marqueurs de pertinence M_k, issus des stratégies des recruteurs ; cet ensemble est exprimé comme $A(t) = \{M_1, M_2, \ldots, M_k\}$;
— ces marqueurs produisent des degrés de possibilité pour la pertinence du terme, qui sont ensuite fuzzifiés à l'aide de la fonction μ_{M_k}, conduisant à l'évidence E_k.

Une fois M_k et les annotations du recruteur fuzzifiés, nous évaluons le potentiel d'expliquer les annotations floues R à l'aide des évidences E_k de M_k. Pour plus de clarté, nous adoptons la distance de Hamming entre les ensembles flous pour évaluer dans quelle mesure R est aligné avec E_k :

$$S'(E_k, R) = \sum_{t \in U} (1 - |\mu_{M_k}(t) - \mu_R(t)|). \tag{II.36}$$

Nous avons ajouté le terme additif 1 à l'équation précédente par rapport à la définition de la section II.9.8, afin d'interpréter directement le score résultant comme élevé si le marqueur est proche des points de vue des recruteurs, sinon comme faible.

Ainsi, la distance de Hamming mesure le degré de pertinence d'un terme en fonction des évidences disponibles E_k. Cette distance peut être normalisée en fonction du nombre d'échantillons évalués avant de passer à l'étape suivante, s'il y a un déséquilibre extrême de classe. Nous nous référerons à sa normalisation par $\|S'(E_k, R)\| \in [0, 1]$.

En appliquant cette distance, nous définissons la possibilité π de classer un terme comme pertinent (R_1) ou non pertinent (R_2), par rapport aux stratégies et points de vue des recruteurs, comme :

$$\pi(R_i|E_k) = \frac{S'(E_k, R_i)}{\max(S'(E_k, R_1), S'(E_k, R_2))}. \tag{II.37}$$

Sur la base de E_k, plus il est probable de déduire qu'un terme est pertinent ou non pertinent, moins il y a d'ambiguïté. À partir de $\pi(R_i|E_k)$, nous estimons le niveau d'ambiguïté associé au marqueur M_k, ou de manière équivalente, à l'évidence E_k comme :

$$G(E_k) = g(\pi(R|E_k)) = \sum_{i=1}^{n} (\pi_i - \pi_{i+1}) \ln(i), \tag{II.38}$$

où $\pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$ est la distribution de possibilité permutée et triée $\pi(R|E_k)$ telle que $\pi_i \geq \pi_{i+1}$ pour $i \in \{1, \ldots, n\}$ et $\pi^*_{n+1} = 0$.

Nous notons que la fonction d'ambiguïté G indique le degré auquel on peut déduire qu'un terme est pertinent ou non pertinent sur la base du marqueur de pertinence M_k.

Sur cette base, nous procédons à la normalisation de l'ambiguïté des marqueurs, définie comme $N(E_k) \in [0, 1]$, et introduisons un facteur de pénalité comme suit :

$$N(E_k) = \min(\|G(E_k)\| + \gamma, 1), \tag{II.39}$$

où γ est un facteur qui vise à pénaliser les marqueurs qui ne fournissent pas d'informations claires sur l'appartenance du terme à l'ensemble des classes à l'étude (dans notre cas d'application, les classes R_1 et R_2). Cela peut être interprété comme un phénomène de manque d'information observé au niveau du marqueur, phénomène intrinsèquement associé à son incertitude dans le contexte d'application. Pour pénaliser de tels marqueurs non informatifs, γ est défini comme suit :

$$\gamma = \begin{cases} 0 & \text{si } \exists R_i(\|S'(E_k|R_i)\| > \beta) \\ 1 - \max(\|S'(E_k|R_i)\|) & \text{sinon.} \end{cases} \tag{II.40}$$

où β est un facteur de tolérance qui représente le niveau minimum à partir duquel un marqueur est considéré comme ne fournissant aucune information claire sur au moins une classe de l'ensemble des classes à l'étude.

Note sur la pénalisation du niveau d'ambiguïté

Il convient de noter qu'avec l'incorporation du facteur mentionné ci-dessus, le niveau d'ambiguïté est pénalisé inversement proportionnellement au niveau maximal d'information fourni par un marqueur par rapport à l'ensemble des classes analysées. Ceci est basé sur l'hypothèse qu'un marqueur qui échoue à fournir des informations non ambiguës sur l'une des classes étudiées, malgré des performances satisfaisantes dans un processus de classification conventionnel, peut potentiellement conduire à des prises de décision dépourvues de fondement solide, basées sur des preuves fragiles, déficientes ou un manque d'information.

Au-delà de l'analyse précédente, la mesure de l'ambiguïté permet d'estimer la justesse de la séparation entre les termes pertinents et non pertinents fournis par chacun des marqueurs du point de vue des recruteurs. Par conséquent, la mesure de l'ambiguïté peut être interprétée en termes de deux types de marqueurs :

— *type 1 – faible ambiguïté :* le marqueur textuel sépare efficacement les termes pertinents des non-pertinents, permettant une distinction facile entre eux ;

— *type 2 – forte ambiguïté :* le marqueur ne différencie pas de manière distincte les termes pertinents et non-pertinents, même s'il peut avoir de bons ou de mauvais résultats dans un processus de classification de termes traditionnel.

Par conséquent, ces marqueurs peuvent être négativement impactés lorsque les variables sur lesquelles ils sont basés subissent de légères modifications au sein d'un nouveau corpus d'OE.

Cette mesure de distance entre ensembles flous nous permet d'évaluer le niveau d'ambiguïté associé à chaque marqueur. Plus la distance de Hamming entre le marqueur et les annotations du recruteur est faible, plus le niveau d'ambiguïté est bas, indiquant une correspondance plus précise entre les annotations du recruteur et le marqueur.

II.10.9 Analyse parallèle de l'ambiguïté et de l'entropie d'information

En plus de mesurer l'ambiguïté pour estimer le niveau d'incertitude de chaque marqueur, nous évaluons simultanément l'entropie d'information mutuelle en tant que moyen d'évaluer dans quelle mesure l'incertitude peut être réduite si le marqueur est pris comme évidence pour le processus de prise de décision.

Soit M_x un marqueur et Y une fonction représentant les annotations du recruteur sur l'ensemble des termes de l'OE. L'entropie d'information mutuelle entre ces fonctions est mesurée comme suit :

$$I(Y; M_x) = H(Y) - H(Y|M_x), \tag{II.41}$$

où $H(Y)$ est l'entropie des annotations du recruteur Y, et $H(Y|M_x)$ est l'entropie conditionnelle du marqueur M_x par rapport aux annotations du recruteur Y.

L'analyse comparative de l'ambiguïté et de l'entropie peut être visuellement représentée en

identifiant quatre types de marqueurs :

- *type 1 – faible entropie et forte ambiguïté :* cela indique une faible influence du marqueur sur la distribution des niveaux de pertinence pour les termes de l'OE du point de vue des annotations du recruteur ; de plus, le niveau élevé d'ambiguïté suggère que la distinction entre les termes pertinents et non pertinents est très floue ;
- *type 2 – faible entropie et faible ambiguïté :* il y a une relation insignifiante observée entre le marqueur et le point de vue du recruteur ; cependant, le marqueur pourrait toujours être utile pour distinguer les termes pertinents des non pertinents ;
- *type 3 – forte entropie et forte ambiguïté :* ce marqueur a une relation significative avec le point de vue du recruteur mais présente une distinction très floue entre les termes pertinents et non pertinents ;
- *type 4 – forte entropie et faible ambiguïté :* c'est le type idéal de marqueur, ayant une relation hautement significative avec le point de vue du recruteur, et permettant également une distinction plus claire entre les termes pertinents et non pertinents.

Jusqu'à présent, les marqueurs ont été caractérisés en fonction de leurs corrélations, de leurs niveaux de précision dans les processus de classification traditionnels, de leur ambiguïté et de leur entropie d'information mutuelle. Nous examinerons maintenant leur niveau d'explicabilité par rapport au contexte organisationnel dans lequel ils sont utilisés.

II.10.10 Alignement des marqueurs avec les stratégies des recruteurs

Pour répondre au besoin d'explicabilité dans l'ensemble final de marqueurs sélectionnés, les hypothèses sur les marqueurs sont alignées avec les stratégies des recruteurs associées à la pertinence des termes de l'OE. Ce processus d'analyse et de contextualisation est effectué en alignant les annotations automatiques générées par les marqueurs avec les stratégies déduites des annotations des recruteurs. Cela génère un ensemble de tuples (m_n, s_m, t_j), composé du marqueur m_n, de la stratégie observée s_m du recruteur, et du type de stratégie t_j (implicite ou explicite).

À cette fin, l'algorithme Apriori (section J) est choisi pour son efficacité, sa facilité d'interprétation, sa robustesse et sa capacité à gérer de grands ensembles de données [175]. Il génère une liste de tuples fréquents facilement interprétables, un avantage significatif car l'explicabilité est cruciale dans notre cas d'application. De plus, sa sensibilité ajustable au bruit et aux valeurs aberrantes augmente son adaptabilité.

Ainsi, la pertinence d'un marqueur est déterminée par son association avec un niveau minimal d'explicabilité selon les stratégies de sélection d'information des recruteurs. Ce niveau est défini par un niveau de support ϑ basé sur la sortie de l'algorithme Apriori. Notamment, dans cette phase, l'examen du niveau de confiance n'est pas indispensable, car notre objectif réside dans la révélation d'associations récurrentes entre le marqueur et les stratégies des recruteurs. Une

exploration détaillée de leur relation prédictive sera effectuée à la phase suivante.

II.10.11 Analyse prospective des marqueurs

Approfondissant l'évaluation orientée vers l'explicabilité, il est essentiel d'évaluer la signification de chaque marqueur sur un ensemble de données plus vaste, d'un point de vue statistique. Nous utilisons la régression logistique pour déterminer l'existence d'une relation statistiquement significative entre chaque marqueur et les perceptions des recruteurs. Les corrélations établies lors de la phase initiale de notre méthodologie (voir section II.5) sont prises en compte pour assurer la stabilité [174] pendant l'entraînement de la régression.

La régression logistique permet précisément de comprendre l'impact de chaque marqueur sur la probabilité que les recruteurs considèrent un terme dans l'analyse de poste comme pertinent, tout en contrôlant simultanément les effets des autres marqueurs. Cette analyse prospective est utile pour l'amélioration continue du processus de sélection des marqueurs, tout en identifiant les marqueurs susceptibles de conserver leur pertinence dans un ensemble de données plus vaste.

II.10.12 Moteur d'inférence flou de type Mamdani pour la sélection et l'évaluation des marqueurs

Étant donné que le processus de sélection des marqueurs pertinents et non pertinents peut être hautement relatif et variable selon chaque contexte organisationnel dans lequel les OEs sont traitées, la logique floue est incorporée pour gérer l'ambiguïté et l'incertitude pendant le processus d'évaluation [172]. Cette approche est particulièrement bénéfique compte tenu de la subjectivité et de la variabilité des facteurs humains et organisationnels dans les OEs. Nous avons choisi un moteur d'inférence flou standard de type Mamdani pour la sélection des marqueurs [221]. Cette méthode qui a fait ses preuves en matière de tâches de sélection de caractéristiques [200], démontre une performance comparative malgré la complexité computationnelle [222], et offre de l'interprétabilité [223].

Dans le contexte de notre méthodologie d'évaluation, le moteur Mamdani construit une échelle de pertinence pour l'ensemble des marqueurs, en intégrant les perspectives des recruteurs. L'adoption de ce moteur permet une représentation transparente et intuitive des critères d'évaluation, rendant les décisions du système plus compréhensibles.

Les résultats de chaque étape de notre méthodologie sont traités comme suit.

Définition des ensembles d'entrée et de sortie flous pour le moteur

Pour un marqueur M_x en cours d'évaluation, le système d'inférence prend m variables d'entrée représentant des métriques de qualité (par exemple, ambiguïté, entropie, signification statistique) mesurées à travers m étapes d'évaluation dans la méthodologie de l'étude. Grâce à un

processus d'inférence flou basé sur des règles, ces entrées sont converties en variables de sortie signifiant des critères de qualité spécifiques adaptés au contexte organisationnel. Bien que notre étude de cas se concentre sur une seule variable de sortie indiquant le niveau de pertinence global de chaque marqueur, la formalisation du système d'inférence généralisé fournie ci-dessous permet d'intégrer plusieurs critères d'évaluation de la qualité des marqueurs en utilisant plusieurs variables de sortie.

Ainsi, dans notre formalisation, le système d'inférence se compose de m variables d'entrée X_i ($i = 1, \ldots, m$), qui représentent différentes métriques de qualité du marqueur textuel à l'étude. De plus, le système comprend p variables de sortie Y_k ($k = 1, \ldots, p$), qui représentent les critères de qualité ou les dimensions du marqueur à déterminer. Pour chaque métrique de qualité (ou variable d'entrée) X_i, nous définissons n_i ensembles flous d'entrée $A_{i,j}$ ($j = 1, \ldots, n_i$). Ces ensembles sont représentés par des termes linguistiques ou des catégories de qualité telles que faible, moyen-faible, moyen, moyen-élevé, etc. De même, pour chaque critère d'évaluation (ou variable de sortie) Y_k, nous définissons q_k ensembles flous de sortie ou catégories de qualité $B_{k,l}$ ($l = 1, \ldots, q_k$).

Il est important de noter qu'en représentant chaque métrique de qualité i par un nombre variable n_i d'ensembles flous ou de catégories floues, nous permettons une adaptabilité dans les niveaux de spécificité et de rigueur pour chaque métrique considérée. Cette flexibilité permet le développement de règles de qualité floues rigoureuses et complètes pour les marqueurs, adaptées afin de répondre aux exigences uniques du contexte organisationnel qui traite les OE.

Ainsi, en définissant clairement les entrées et les sorties du système, nous offrons une transparence sur les facteurs considérés lors de l'évaluation des marqueurs, ce qui améliore l'explicabilité de la méthodologie.

Création de la base de règles de qualité floue

Sur la base des métriques de qualité des marqueurs, nous définissons des règles de qualité floues, associant différentes valeurs des métriques de qualité à chaque variable de sortie représentant un critère d'évaluation de qualité.

Nous créons un ensemble de R règles «si-alors» reliant les variables d'entrée et de sortie du moteur. Chaque règle r ($r = 1, \ldots, R$) peut être définie comme suit :

$$\texttt{SI}; X_1; \text{est}; A_{1,j_1}; \texttt{ET} \ldots \texttt{ET}; X_m; \text{est}; A_{m,j_m}; \texttt{ALORS}\ Y_1; \text{est}; B_{1,l_1}; \texttt{ET} \ldots \texttt{ET}; Y_p; \text{est}; B_{p,l_p}, \quad \text{(II.42)}$$

où X_i correspond à la métrique de qualité i, A_{i,j_i} correspond à la catégorie floue j_i de la métrique i, Y_p correspond au critère d'évaluation de la qualité (variable de sortie) p, et B_{p,l_p} représente la catégorie floue l_p du critère évalué p.

Ces règles constituent la base de règles de qualité floues du système d'inférence. Le moteur

utilise ces règles pour effectuer un raisonnement et estimer le degré de réalisation pour chaque critère d'évaluation sur la base des métriques de qualité données du marqueur.

Fuzzification des scores de qualité du marqueur

Étant donné un vecteur d'entrée précis $\mathbf{x}(x_1, \ldots, x_m)$, représentant les valeurs x_i des métriques de qualité pour le marqueur textuel, nous calculons le degré d'appartenance $\mu_{A_{i,j}}(x_i)$ pour chaque catégorie de qualité $A_{i,j}$ en utilisant sa fonction d'appartenance correspondante $f_{i,j}(x_i)$:

$$\mu_{A_{i,j}}(x_i) = f_{i,j}(x_i). \tag{II.43}$$

Ici, $f_{i,j}(x_i)$ est la fonction d'appartenance pour la catégorie floue $A_{i,j}$, et x_i est la valeur d'entrée précise pour la métrique de qualité (ou variable d'entrée) X_i.

Le résultat de la fuzzification est un ensemble de degrés d'appartenance, représentant la force avec laquelle les valeurs d'entrée appartiennent à chaque catégorie floue. Ces degrés d'appartenance sont utilisés dans les étapes suivantes du processus d'inférence floue.

Évaluation des règles

Nous calculons la force d'activation $\wp_r$ pour chaque règle de qualité :

$$\wp_r = T(\mu_{A_{1,j_1}}(x_1), \ldots, \mu_{A_{m,j_m}}(x_m)). \tag{II.44}$$

Dans cette équation, $\wp_r$ est la force d'activation de la règle r, représentant le degré d'activation de la règle. T est un opérateur d'agrégation de type T-norme, la fonction minimum, qui combine les degrés d'appartenance des catégories floues antécédentes (ensembles flous d'entrée) pour une règle donnée. $\mu_{A_{i,j_i}}(x_i)$ représente le degré d'appartenance de la catégorie floue A_{i,j_i} pour la métrique de qualité X_i avec une valeur précise x_i.

Nous soulignons que la force d'activation $\wp_r$ indique à quel point une règle est pertinente ou applicable dans le contexte des valeurs d'entrée données. Elle est utilisée dans l'étape suivante du processus d'inférence floue pour pondérer la contribution de la règle à la sortie.

Évaluation des conséquences

Nous calculons l'ensemble flou de sortie tronqué B^r_{k,l_r} pour chaque règle r, sur la base de la force d'activation obtenue à l'étape précédente :

$$\mu_{B^r_{k,l_r}}(y_k) = T(\wp_r, \mu_{B_{k,l_r}}(y_k)). \tag{II.45}$$

Dans cette équation, $\mu_{B^r_{k,l_r}}(y_k)$ est le degré d'appartenance de l'ensemble flou de sortie tronqué B^r_{k,l_r} pour la règle r. $\wp_r$ est la force d'activation de la règle r, et $\mu_{B_{k,l_r}}(y_k)$ est le degré

d'appartenance de la catégorie de sortie floue originale (ou ensemble flou) B_{k,l_r}. De plus, y_k représente toute valeur dans le domaine de la variable de sortie Y_k.

De cette façon, nous modifions les ensembles flous de sortie originaux (partie conséquente de la règle) en les tronquant sur la base de la force d'activation $\wp_r$. Cette étape garantit que l'influence de la règle de qualité sur la sortie finale est proportionnelle à la force d'activation de la règle.

Agrégation des sorties

Nous combinons les ensembles flous de sortie tronqués en un seul ensemble flou de sortie agrégé B_k pour chaque critère d'évaluation de la qualité Y_k :

$$\mu_{B_k}(y_k) = S(\mu_{B^1_{k,1}}(y_k), \ldots, \mu_{B^r_{k,l_r}}(y_k), \ldots, \mu_{B^R_{k,q_k}}(y_k)). \tag{II.46}$$

Dans cette équation, $\mu_{B_k}(y_k)$ est le degré d'appartenance de l'ensemble flou de sortie agrégé B_k pour le critère d'évaluation Y_k. S est un opérateur d'agrégation de la S-norme, la fonction maximum, qui combine les degrés d'appartenance des ensembles flous de sortie tronqués de chaque règle.

En agrégeant les ensembles flous de sortie tronqués, nous obtenons un seul ensemble flou qui représente l'effet combiné de toutes les règles de qualité sur la ou les variables de sortie. Cet ensemble flou agrégé est ensuite utilisé dans l'étape de défuzzification pour calculer la valeur de sortie nette pour chaque critère d'évaluation de la qualité Y_k.

Défuzzification de la sortie

Nous calculons la valeur de sortie nette pour chaque critère d'évaluation Y_k en utilisant la méthode du centroïde [224]. Dans notre étude de cas, nous avons choisi la méthode du centroïde en fonction de critères contextuels spécifiques, y compris la précision, le temps de réponse souhaité, et les ressources informatiques disponibles :

$$y^*_k = \frac{\int y_k \mu_{B_k}(y_k) dy_k}{\int \mu_{B_k}(y_k) dy_k}. \tag{II.47}$$

Dans cette équation, y^*_k est la valeur de sortie nette pour le critère d'évaluation de la qualité (variable de sortie) Y_k. L'intégrale du numérateur calcule la somme pondérée des valeurs de la variable de sortie (y_k) multipliée par leurs degrés d'appartenance correspondants dans l'ensemble flou de sortie agrégé ($\mu_{B_k}(y_k)$). L'intégrale du dénominateur calcule la somme des degrés d'appartenance dans l'ensemble flou de sortie agrégé ($\mu_{B_k}(y_k)$). La valeur de sortie nette y^*_k est obtenue en divisant la somme pondérée (numérateur) par la somme des degrés d'appartenance (dénominateur).

Le moteur d'inférence de type Mamdani que nous proposons englobe plusieurs étapes essentielles : fuzzification, évaluation des règles de qualité floues, agrégation des règles et défuzzification. Grâce à cette structure, une valeur nette est produite pour chaque critère d'évaluation de la qualité Y_k et chaque marqueur, facilitant ainsi leur interprétation. Cette approche méthodologique s'avère particulièrement efficace dans des contextes spécifiques tels que l'analyse des OE et l'architecture de l'agent CDI possibiliste, où il est essentiel de comprendre la qualité et la pertinence des marqueurs. De plus, les informations contextuelles utilisées enrichissent la compréhension des marqueurs, optimisant ainsi les performances du système, notamment en matière de récupération d'informations sur l'emploi.

Du début à la fin du processus d'inférence, chaque étape joue un rôle clé. D'une part, elles améliorent continuellement l'extraction de l'information et l'annotation sémantique, tout en veillant à optimiser le processus de prise de décision de l'agent au fil du temps. D'autre part, elles renforcent la transparence et l'explicabilité de la méthodologie employée. Cette clarté dans le processus est essentielle pour instaurer la confiance envers les systèmes de CCO, en facilitant l'interprétation des résultats.

En conclusion, l'intégration de l'inférence floue à notre approche systématique présente des avantages potentiels pour une évaluation plus objective des marqueurs textuels de l'OE. Cette intégration vise à offrir une précision analytique plus satisfaisante, tout en conservant une adaptabilité et une flexibilité en réponse aux variabilités des contextes organisationnels.

II.11 Approche grapholinguistique pour l'analyse des CV : un focus sur le processus de segmentation

Après avoir présenté l'approche proposée pour extraire les informations les plus essentielles des OE, nous passons maintenant à la présentation de notre approche pour le traitement automatique des CV, intégrant les spécificités du contexte organisationnel et l'expertise des recruteurs. Cette partie de la méthodologie couvre en particulier le processus de segmentation automatique de ce type de documents.

Dans cette approche, il est crucial de distinguer et de structurer les différentes sections qui composent généralement un CV, telles que les informations personnelles, l'éducation, l'expérience professionnelle, les compétences, etc. Cette segmentation, réalisée automatiquement grâce à des algorithmes spécifiques, permet une meilleure analyse et une extraction plus précise des informations pertinentes.

Prétraitement et formatage des CV

Le prétraitement et formatage des CV est l'étape initiale de notre méthodologie, visant à rendre les CV plus uniformes pour faciliter l'extraction d'informations, compte tenu de la

diversité des formats de document (PDF, Word, etc.). Cette phase garantie une segmentation plus pertinente des CV en sections clés (coordonnées, expérience, éducation, compétences, entre autres), à partir d'un nettoyage préalable des données qui inclut la suppression d'en-têtes et pieds de page, la correction d'erreurs d'orthographe et l'élimination d'informations non pertinentes.

Le cadre de traitement automatique des CV se base ensuite sur une approche articulée autour de trois axes : construction d'une représentation stratifiée du contexte organisationnel dans lequel les CV sont analysés, adaptation de modèles BERT pour la segmentation grapholinguistique de ces documents avec des ensembles de données restreints, et évaluation de la pertinence des modèles BERT affinés sur de nouveaux échantillons de CV.

II.11.1 Aspects grapholinguistiques du CV

La grapholinguistique, se concentrant sur les *graphes* [66, p. 63], offre un cadre d'étude des unités minimales obtenues par opposition, les *graphèmes*, comparables aux phonèmes [66, p. 119]. La *graphématique* examine les séquences graphémiques (SG), unidimensionnelles ou bidimensionnelles (cf. aussi [213, chap. 1]).

L'approche grapholinguistique dans l'analyse des CV intègre les facettes linguistiques et typographiques du texte, améliorant ainsi la précision de l'analyse automatique. Elle présente aussi l'avantage de la flexibilité adaptative et propose une représentation enrichissante du document, des composantes fondamentales pour une CCO plus robuste.

II.11.2 Ontologie des CV

Comme nous l'avons souligné dans les sections précédentes, afin de concevoir des solutions AA plus robustes, il est essentiel d'obtenir une représentation fidèle de leur contexte sociétal et/ou organisationnel [44, 225]. Par conséquent, le premier axe de notre approche de représentation des CV commence par l'application de la méthode UNC [107]. Des entretiens préliminaires avec les recruteurs ont été menés pour identifier la représentation de CV la plus appropriée.

De manière similaire aux OE, nous construisons divers artefacts tels que des schémas préconceptuels et des modèles de domaine [107]. De plus, les objectifs des recruteurs concernant le cycle de vie des CV dans les processus de recrutement sont ensuite identifiés et organisés hiérarchiquement. Des diagrammes de processus, tels que définis par [107], illustrent les modèles de processus commerciaux liés aux CV. Un diagramme en arêtes de poisson est également utilisé pour dériver les associations entre les problèmes organisationnels concernant les CV et leurs facteurs causatifs. Enfin, en suivant [107], un tableau de processus explicatif intègre tous les diagrammes précédents pour consolider la représentation du contexte organisationnel.

Le processus mentionné ci-dessus permet une extraction directe d'une ontologie de CV adaptée aux spécificités d'une organisation donnée. Cette ontologie est un outil fondamental pour la représentation riche et approfondie de chacun des concepts qui composent le texte du CV. De

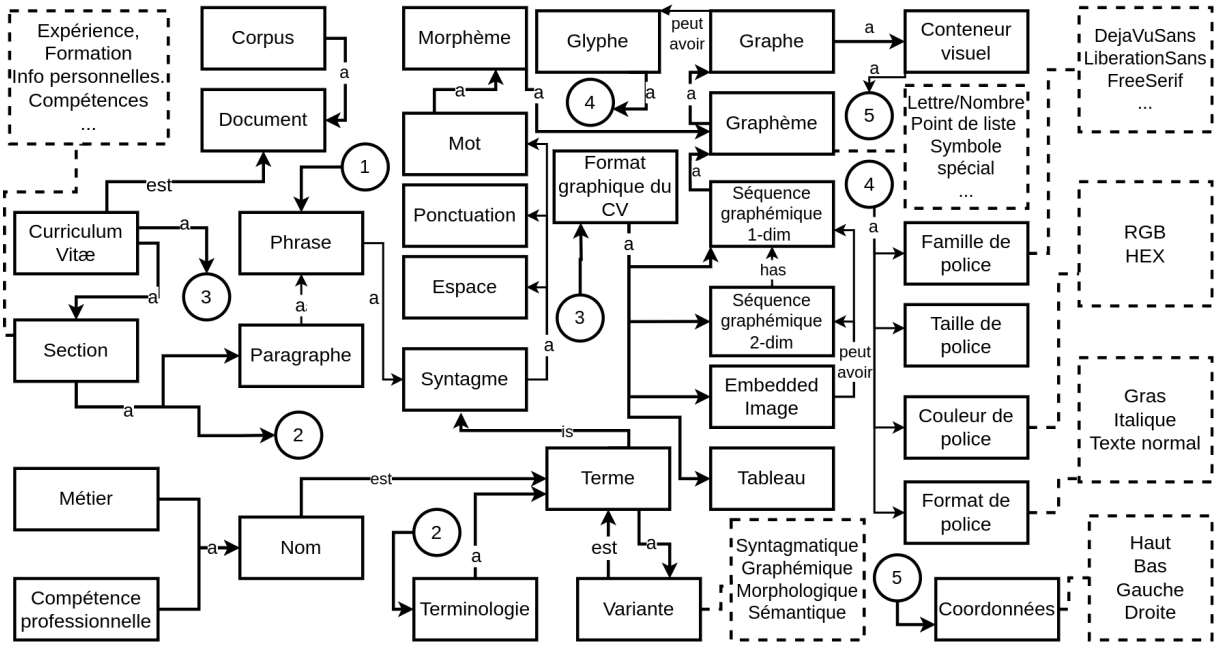


FIGURE II.9 – Vue supérieure de l'ontologie illustrant des concepts linguistiques du CV. Des cercles avec des numéros (1, 2, 3, 4 et 5) sont utilisés pour représenter les relations entre des concepts éloignés dans le diagramme.

plus, elle offre la possibilité d'annotations sémantiques supplémentaires, qui sont généralement appliquées après un processus de segmentation [208]. Dans la figure II.9, nous montrons une vue d'ensemble de l'ontologie extraite en utilisant cette approche.

II.11.3 Extraction de texte à partir des CV

Nous effectuons une extraction de texte sensible au format à partir du CV. En utilisant une stratégie d'analyse syntaxique paresseuse, les caractères Unicode et leurs propriétés graphétiques (coordonnées et styles de police) sont identifiés dans les données PDF du CV. Ensuite, nous appliquons un algorithme de mise en page pour extraire les SG.

II.11.4 Extraction de la terminologie

Après avoir identifié les SG, nous extrayons la terminologie à partir du CV en utilisant l'approche proposée par [178], comme cela a été fait pour les OE. Les termes dans le CV sont identifiés en estimant leur « termitude (termhood) » [218] en utilisant le mesure de weirdness ratio [178]. De plus, nous identifions les compositions syntagmatiques, les variantes morphologiques, graphémiques et sémantiques des termes [178].

II.11.5 Analyse des annotations des recruteurs

Après avoir extrait la terminologie spécifique aux CV, le deuxième axe de notre approche commence par analyser comment les recruteurs segmentent manuellement les CV. Un panel de recruteurs segmente manuellement un ensemble de CV $D = \{d_1, \ldots, d_M\}$. Pour chaque CV m, nous obtenons un tuple de sections annotées $S_m = (s_{m,1}, \ldots, s_{m,l})$. Notons que les sections annotées peuvent varier d'un CV à l'autre. Nous générons ensuite des segmentations de documents automatiques qui capturent les modèles de lisibilité [226] utilisés par les recruteurs pour identifier les sections de CV.

Pour ce faire, nous représentons d'abord chaque SG dans un CV annoté m en termes de ses caractéristiques visuelles, telles que la famille de police, le corps de la police, la couleur de la police et le style gras/maigre, italique/droit. Nous regroupons les SG en fonction de ces caractéristiques pour obtenir un ensemble de clusters $C_m = \{C_{m,1}, C_{m,2}, \ldots, C_{m,N}\}$. Pour chaque cluster $C_{m,n}$, nous obtenons les coordonnées spatiales de ses SG constitutives dans le CV m correspondant et les utilisons comme points de partition pour segmenter automatiquement le CV uniquement en fonction de ses caractéristiques graphiques ou visuelles.

Ainsi, nous obtenons un tuple de sections hypothétiques de CV $H_{C_{m,n}} = (h_{C_{m,n},1}, \ldots, h_{C_{m,n},k})$. Nous appliquons ensuite une métrique de similarité conçue pour mesurer le niveau de ressemblance entre deux segmentations, à savoir la similarité 'S' [227]. Cette mesure permet de calculer la similarité entre la segmentation manuelle du recruteur S_m et chaque segmentation hypothétique $H_{C_{m,n}}$, dans le but d'identifier des segmentations de CV de plus en plus pertinentes en fonction du point de vue des recruteurs sur le *format*.

Dans la métrique de similarité 'S', la segmentation entière ainsi que les segments individuels sont conceptualisés comme ayant une masse. La similarité entre deux segmentations est quantifiée comme la proportion de frontières qui restent non transformées lorsque les segmentations sont comparées en utilisant la distance d'édition, où cette dernière fonctionne comme un mécanisme de pénalité. Cette pénalité est mise à l'échelle en fonction de la taille de la segmentation, faisant de 'S' une fonction symétrique qui exprime la similarité en pourcentage.

Ce mécanisme permet à 'S' d'être applicable à des segmentations de diverses granularités et unités (par exemple, paragraphes, phrases, syntagmes). Le processus implique de convertir une segmentation en une séquence de valeurs de masse de segment, puis de transformer des paires de segmentations en séquences parallèles d'ensembles de frontières. La distance d'édition entre ces séquences est ensuite calculée, et cette valeur est utilisée pour soustraire des pénalités pour chaque opération d'édition du nombre de frontières potentielles dans une segmentation. Le résultat est normalisé par le nombre total de frontières potentielles, produisant une valeur de similarité allant de 0 à 1, où 0 indique aucune similarité et 1 représente des segmentations identiques.

Ce processus sert d'analyse initiale qui nous permet de comprendre les liens entre les ca-

ractéristiques graphiques du CV et les segmentations manuelles effectuées par les recruteurs. Cependant, au-delà de cela, il est intéressant d'identifier les fonctions du texte au sein du CV qui sont le plus fortement associées à ces caractéristiques graphiques. Une telle connaissance pourrait être exploitée pour améliorer le processus de segmentation du document.

Par conséquent, après avoir identifié le cluster de segmentation optimal $C_{m,n}$ pour chaque CV m d'un point de vue visuel, nous décrivons les SG de ces clusters en termes de fonctions du texte ou fonctions textuelles (FT), telles que les titres de section, les titres de sous-section, les titres de liste, les noms de compétences professionnelles, etc. Nous conservons cette description sous forme de triplets <Séquence Graphémique du Cluster X, a, Fonction Textuelle Y>, obtenant un ensemble de triplets descriptifs T_m pour chaque CV. Cette description vise à unifier les propriétés graphiques et linguistiques du CV, fournissant une compréhension plus profonde des relations entre son contenu *grapholinguistique* et les segmentations manuelles des recruteurs.

Finalement, nous appliquons l'algorithme Apriori [175] à l'ensemble T de tous les triplets de CV extraits, et nous identifions l'ensemble des FT les plus fréquentes et pertinentes $\mathrm{FT} = \{\mathrm{FT}_1, \ldots, \mathrm{FT}_R\}$, permettant de décrire les segmentations effectuées par les recruteurs sensibles au format. Ces FT servent de moyen pour identifier des coordonnées de segmentation de CV plus optimales.

II.11.6 Construction semi-supervisée d'un corpus de référence

À ce stade, nous construisons progressivement un corpus de référence $\mathfrak{G}r$ destiné à représenter la vérité de terrain pour chaque FT pertinente FT_r où $r = 1, \ldots, R$. À partir de chaque CV m, nous extrayons automatiquement des SG en 1D et en 2D correspondant à des instances de FT_r qui sont familières aux recruteurs (par exemple, les titres de section du CV, les sous-titres, les noms de listes, etc.). Par la suite, à travers un processus d'ingénierie des caractéristiques, nous identifions les marqueurs graphiques GM_j et les marqueurs textuels TM_i permettant de représenter FT_r tel que $\mathrm{FT}_r = \{\mathrm{GM}_{1,r}, \ldots, \mathrm{GM}_{j,r}, \mathrm{TM}_{1,r}, \ldots, \mathrm{TM}_{i,r}\}$. En d'autres termes, nous identifions les marqueurs qui modélisent le mieux les FT étroitement associées aux segmentations manuelles réalisées par les recruteurs. Dans le Tableau II.6, nous présentons les marqueurs de format et textuels qui ont été identifiés comme étant potentiellement associés à la FT "Titre de section". La Figure. II.10 illustre différents exemples de titres de section dans des CV modernes.

Nous évaluons la signification statistique de ces marqueurs à travers des modèles de régression logistique (optimisés en utilisant l'estimation du maximum de vraisemblance). Dans notre cas d'application, la RLC est avantageuse en raison de son interprétabilité éprouvée, de son efficacité à identifier des caractéristiques (ou marqueurs) pertinents par rapport à une variable de prédiction, et de sa robustesse contre le sur-ajustement dans de petits ensembles de données [228].

Suivant cette approche, une fois que nous avons identifié les marqueurs qui représentent de

TABLE II.6 – Marqueurs associés à la FT "Section de titre". GM désigne les marqueurs graphiques (de format) et TM les marqueurs textuels.

Marqueur	Nom	Description
GM_1	Taille de police	Utilisée pour indiquer la taille de la police dans le CV.
GM_2	Famille de police	Type de police utilisé.
GM_3	Couleur	Mesure de la distance entre la couleur de la SG et la couleur de police la plus fréquente du CV.
GM_4	Gras	Indication si le texte est en gras.
GM_5	Italique	Indication si le texte est en italique.
TM_1	Capitalisé	Indication si le début de chaque mot du texte est en majuscule.
TM_2	Tout en majuscules	Indication si tout le texte est en majuscules.
TM_3	Variante du titre du terme	Utilisation de différentes variantes d'un terme.
TM_4	Fréquence dans les titres de CV	Fréquence agrégée des mots de la SG dans le corpus de référence de titres, chaque fréquence de mot étant pénalisée par un facteur $\sigma(-p)$, où p est la position du mot dans la SG et σ est la fonction sigmoïde.
TM_5	Fréquence dans les phrases courantes de CV	Fréquence agrégée des mots de la SG dans les échantillons négatifs du corpus de référence.

FIGURE II.10 – Exemples du titre de section "Expériences" extrait de CV français contemporains. La diversité des couleurs, la famille/taille de la police, la police en gras et les variantes de termes sont quelques-unes des caractéristiques qui font qu'un titre de CV moderne se démarque.

manière optimale chaque FT_r, chaque SG dans un CV m est représentée en termes des marqueurs liés à FT_r. À partir de cette représentation, un second processus de clustering est effectué pour chaque CV afin d'identifier automatiquement et de manière exhaustive toutes les séquences liées

à FT_r.

De plus, nous évaluons dans quelle mesure chaque cluster résultant représente un ensemble de véritables instances de FT_r en calculant un degré moyen de possibilité. Pour ce faire, nous définissons une fonction d'appartenance f qui associe des tuples de marqueurs de FT_r à l'intervalle [0,1], exprimant le degré de possibilité qu'une SG spécifique dans le CV corresponde à de véritables instances de FT_r. Les clusters ayant un degré moyen de possibilité dépassant un seuil β_r sont sélectionnées comme instances plus fiables de FT_r et deviennent partie du corpus de référence $\mathfrak{G}_r$.

Étant donné que ces instances sont concises, elles peuvent être validées manuellement pour assurer la qualité du processus de réglage fin. Dans l'étape finale, en utilisant une approche de sous-échantillonnage basée sur le regroupement [229], des échantillons négatifs sont également extraits des CV pour compléter le corpus.

Cette approche nous permet d'établir un corpus de référence composé de CV réels, dûment annotés et regroupés selon les FT. Ces regroupements fournissent un moyen efficace de distinguer les différentes sections et informations dans un CV, rendant le processus de segmentation et d'analyse des documents plus robuste et précis. Par ailleurs, l'importance accordée à l'intégration des analyses graphiques et textuelles lors du traitement des CV contribue à garantir une meilleure robustesse de la CCO.

II.11.7 Formatage des séquences, affinage des modèles basés sur BERT, et segmentation

Dans la phase suivante de notre approche, chaque SG dans le corpus de référence $\mathfrak{G}_r$ est formatée par rapport à la FT_r correspondante. Cette opération est réalisée en concaténant [212] le texte de la SG avec les marqueurs graphiques/format et textuels de FT_r, qui ont été préalablement fuzzifiés. Le processus de fuzzification est effectué par rapport à cinq catégories floues (ou variables linguistiques) représentant des gammes de valeurs, qui sont modélisées par des fonctions triangulaires standard. Pour améliorer la capacité de BERT à interpréter le sens des marqueurs numériques flous, nous utilisons des noms de catégorie contrastés (par exemple, très petit pour 0 contre très grand pour 1). En transformant les marqueurs numériques en variables linguistiques, la fuzzification simplifie la complexité de la segmentation sensible au format, qui dans notre approche repose sur l'extraction des connaissances des recruteurs. Cette connaissance extraite est naturellement affectée par des phénomènes tels que les informations incomplètes et les incertitudes cognitives [45] que la fuzzification aide à surmonter [230].

Sur la base du corpus de référence FT_r ainsi formaté, nous exploitons la version distillée du modèle BERT multilingue [231], dans le but d'affiner les classificateurs de séquence BERT pour prédire si les SG de CV sont des "Instances fiables de FT_r" ou des "Instances non fiables de FT_r". Plus précisément, nous utilisons la version *distilbert-base-multilingual-cased*, qui convient

aux environnements de production et aux tâches de classification BERT [231]. La Figure II.11 présente l'architecture globale de notre approche pour représenter et classer les SG de CV.

Après l'affinage des modèles basés sur BERT, nous automatisons la tâche de segmentation des CV. Pour un CV m donné, nous extrayons son texte et sa terminologie pour identifier les SG associées aux FT les plus pertinentes. Les coordonnées spatiales de ces SG deviennent des coordonnées de référence pour la segmentation et servent de coordonnées initiales pour chaque section du CV. Ensuite, les SG restantes du CV sont attribuées à l'une de ces sections en fonction de trois critères essentiels. Tout d'abord, nous nous assurons que la SG est à une distance minimale de la coordonnée initiale de la section. Deuxièmement, nous vérifions que la séquence n'est pas spatialement au-dessus de la coordonnée initiale de la section. Enfin, nous confirmons que la SG et la coordonnée initiale de la section se trouvent dans la même colonne du CV. En suivant ces étapes, nous pouvons segmenter un CV en ses différentes sections en exploitant les FT les plus pertinentes identifiées tout au long de la méthodologie proposée. Un workflow illustrant ce processus est présenté dans la figure II.12.

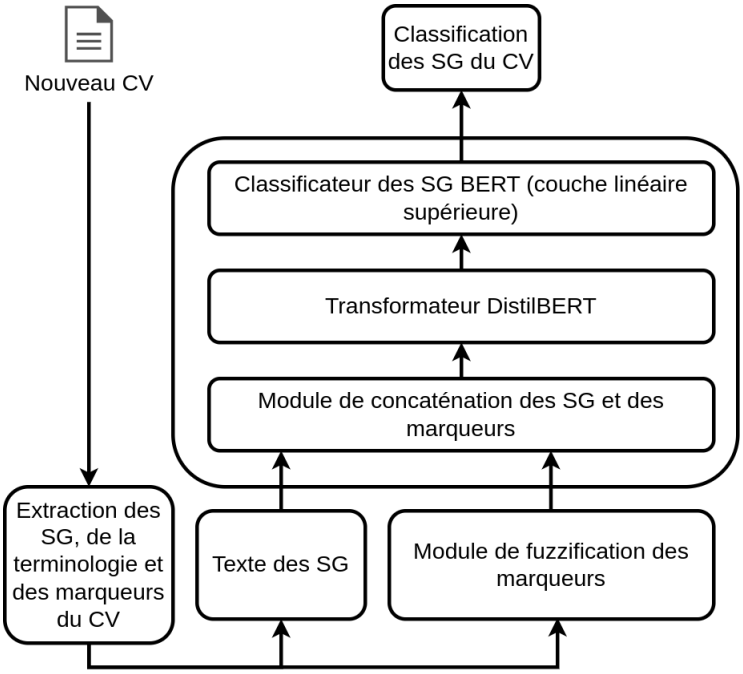


FIGURE II.11 – Architecture générale pour le traitement des SG.

II.11.8 Extraction d'informations contextuelles du CV et annotation sémantique

Semblable aux OE, nous postulons qu'un processus d'annotation semi-automatique peut servir à isoler les concepts les plus essentiels dans les CV. Nous reconnaissons que l'identification

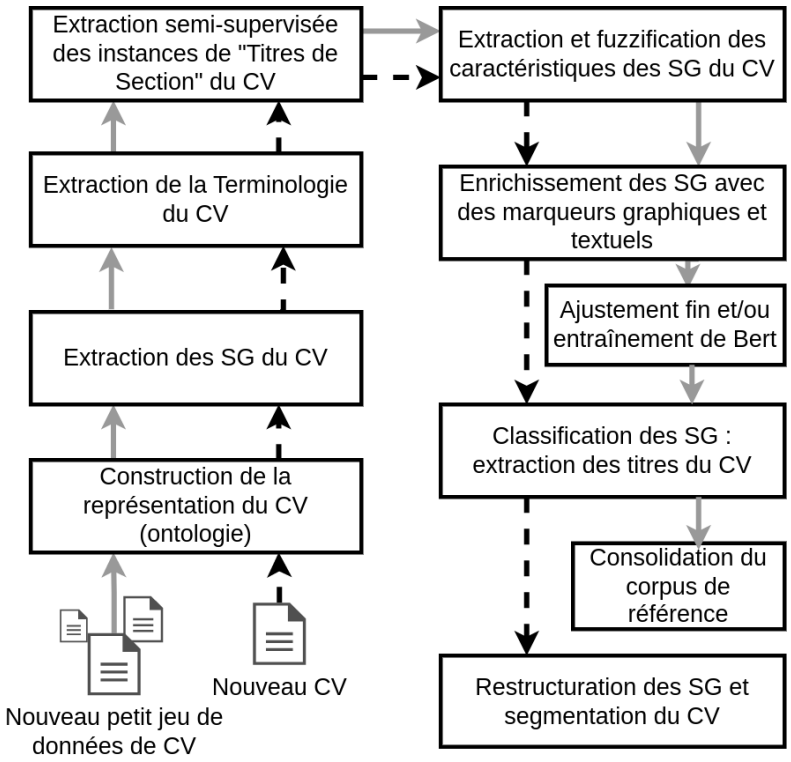


FIGURE II.12 – Flux de travail dérivé en appliquant l'approche proposée. Les flèches noires en pointillés illustrent comment de nouveaux échantillons de CV sont traités sur la base d'un modèle BERT ajusté pour l'extraction des titres de CV. Les flèches grises pleines indiquent comment de nouveaux petits ensembles de données de CV peuvent être exploités pour ajuster le classificateur BERT.

de marqueurs textuels, destinée à extraire ces éléments les plus significatifs du CV, devrait être une démarche collaborative. Cette approche devrait idéalement solliciter non seulement l'expertise des recruteurs, mais aussi prendre en compte les perspectives des candidats tout en explorant minutieusement leurs divers contextes sociétaux, professionnels et académiques.

Cependant, étant donné la complexité inhérente à une telle formalisation pour le CV, qui requerrait une disponibilité accrue tant de la part des recruteurs que des candidats, nous choisissons de limiter l'extraction des informations les plus pertinentes des CV dans notre recherche. Notre méthodologie se concentre donc sur l'extraction d'informations contextuelles du CV relatives à la compétence professionnelle qui intéressent les recruteurs. Ces informations, telles que les années d'expérience liées à une compétence ou le type d'expériences professionnelles associées (technique, fonctionnelle, management), sont indiquées par les recruteurs lors de l'élaboration de la représentation du contexte organisationnel.

Ces éléments sont extraits suite à la phase de segmentation. Le traitement automatique du CV culmine avec une annotation sémantique. Cette dernière facilite la différenciation des entités

jugées pertinentes pour le processus de CCO. Parmi ces entités, on peut citer les sections relatives à l'expérience professionnelle, à l'éducation, ainsi que les principales entités associées, telles que les intitulés de postes, les dates-clés, les compétences acquises lors de chaque expérience, les diplômes obtenus, les universités fréquentées, etc.

II.11.9 Vérification de la qualité de l'annotation

Assurer la qualité de l'annotation humaine peut être une étape pertinente pour garantir la cohérence et la fiabilité des annotations effectuées sur les CV. Dans notre méthodologie, nous suggérons un accord entre annotateurs lorsque cela est possible pour évaluer la qualité des annotations.

Lorsque plusieurs recruteurs participent à l'annotation d'un même CV, ce qui se produit souvent lors d'un processus de recrutement collaboratif, il est possible d'évaluer le niveau d'accord entre leurs évaluations à l'aide de mesures statistiques. En cas de divergence notable entre leurs points de vue, un mécanisme de résolution peut être mis en œuvre. Ce dernier pourrait inclure des discussions entre les recruteurs ou même la participation d'un recruteur supplémentaire pour la médiation.

II.12 La phase de la Correspondance Curriculum Vitæ - Offre d'Emploi

Cette section introduit l'approche proposée pour la Correspondance Curriculum Vitæ - Offre d'Emploi (CCO) en se basant sur la représentation de l'OE et du CV.

II.12.1 Définitions et notations

Nous commençons par introduire les concepts-clés et les notations nécessaires pour formaliser notre approche.

Compréhension et utilisation des informations contextuelles dans les CV pour l'évaluation des candidats

Le concept d'informations contextuelles du CV fait référence à un ensemble d'éléments pertinents pour les recruteurs, explicitement mentionnés dans le CV et liés à la compétence professionnelle du candidat, qui peuvent être identifiés lors de la phase de représentation du contexte organisationnel via des entretiens (section II.5).

Les informations contextuelles du CV, comme la section mentionnant une compétence, les années d'expérience sur cette compétence, les certifications associées, la mobilisation de la compétence dans des contextes techniques ou managériaux, ainsi que le niveau d'expertise du can-

didat exprimé dans son CV (expert, junior, confirmé, etc.), constituent des critères essentiels pour les recruteurs lors de l'évaluation du profile d'un candidat.

En synthèse, le concept d'informations contextuelles du CV regroupe des critères de décision-clés pour les recruteurs, explicitement mentionnés dans le document. Ces critères essentiels, peuvent être identifiés durant la phase de représentation du contexte organisationnel (section II.5) et facilitent l'évaluation de la pertinence des compétences énoncées dans le CV.

Ensembles

À ce stade, nous définissons les principaux ensembles nécessaires pour la modélisation finale de la CCO :

- soit $G_{CV} = (\mathcal{C}_{CV}, \gamma_{CV})$ le graphe sémantique d'un CV obtenu grâce à l'ontologie, où $\mathcal{C}_{CV}$ est l'ensemble des concepts représentant les entités sémantiques dans le document et γ_{CV} est l'ensemble des arêtes représentant les relations entre ces entités ; de même, soit $G_O = (\mathcal{C}_{OE}, \gamma_O)$ le graphe sémantique d'une OE ; $\mathcal{C}_{OE}$ est l'ensemble des concepts les plus essentiels de ce dernier document, identifiés après le processus d'extraction d'information ; cette représentation est centrée sur le concept de compétence professionnelle en raison de son importance primordiale dans le processus d'évaluation de la pertinence des candidats ;
- soit $\mathcal{D}$ l'ensemble de dimensions d'évaluation de la similarité entre le CV et l'OE, pertinentes pour le contexte organisationnel ; trois dimensions ont été principalement explorées dans cette thèse : les similarités syntaxique, ontologique et distributionnelle. La dimension syntaxique évalue la similarité entre les compétences mentionnées dans le CV et celles dans l'OE, en se basant sur la similarité des chaînes de caractères des termes qui représentent ces compétences ; la similarité ontologique mesure la similarité entre les concepts du point de vue de l'ontologie organisationnelle, évaluant ainsi la distance conceptuelle entre eux au sein de cette base de connaissances ; la similarité distributionnelle utilise des modèles transformateurs, tels que BERT, pour évaluer la similarité entre les compétences professionnelles des deux documents, en prenant en compte une représentation plus large de la langue et des contextes d'utilisation de chaque compétence, sur des corpus de données textuelles plus universels [231] ; chaque contexte organisationnel pourrait avoir un intérêt à évaluer d'autres dimensions de similarité (e.g. la dimension pragmatique) ;
- soit $\mathcal{M}$ l'ensemble d'informations contextuelles du CV relatives à la compétence professionnelle, pertinentes pour les recruteurs.

Fonctions de similarité

L'estimation de la similarité entre les compétences de l'OE et du CV est centrale dans l'approche proposée. Nous introduisons ici les fonctions qui accomplissent cette tâche pour les différentes dimensions de similarité :

— $S_d(C_{OE_i}, C_{CV_j})$: le score de similarité entre la compétence C_{OE_i} de l'OE et la compétence C_{CV_j} du CV pour la dimension d.
— $S_{\text{syntaxique}}(C_{OE_i}, C_{CV_j})$: mesure la similarité syntaxique (distance de Levenshtein) entre les termes des compétences ; cette mesure est utilisée pour identifier dans le texte du CV d'un candidat les éléments explicites qui correspondent aux compétences essentielles de l'OE ;
— $S_{\text{ontologique}}(C_{OE_i}, C_{CV_j})$: évalue la similarité basée sur les connaissances de l'ontologie organisationnelle (mesure basée sur le chemin), capturant des relations organisationnelles plus spécifiques que celles présentes dans le texte ;
— $S_{\text{distributionnelle}}(C_{OE_i}, C_{CV_j})$: mesure la similarité sémantique en utilisant des modèles de langage avancés comme BERT pour mieux capturer le contexte et les nuances de la langue.

Ambiguïté

La mesure d'ambiguïté de chaque concept essentiel de l'OE, déjà présentée dans cette méthodologie (section II.9.8), est reprise à cette phase.

Soit $A(C_{OE_i})$ le niveau d'ambiguïté pour la compétence essentielle C_{OE_i} extraite de l'OE.

II.12.2 Formalisation

Nous procédons maintenant à la description des étapes d'évaluation de la correspondance entre un CV et une OE en utilisant les définitions et notations précédentes.

Calcul des scores de similarité pour chaque dimension

À ce stade, la similarité entre chaque compétence identifiée dans le CV et l'OE est évaluée selon les différentes dimensions de similarité, comme illustré ci-dessous :

$$\mathcal{E}_d(C_{OE_i}, C_{CV_j}) = \begin{cases} \text{Distance de Levenshtein normalisée} & \text{si } d = \text{syntaxique} \\ \text{Mesure basée sur le chemin dans l'ontologie} & \text{si } d = \text{ontologique} \\ \text{Similarité cosinus des embeddings} & \text{si } d = \text{distributionnelle.} \end{cases} \tag{II.48}$$

Ce qui précède permet d'obtenir un score global de similarité, défini comme la somme pondérée de chacune des similarités :

$$\begin{aligned} S_d(C_{OE_i}, C_{CV_j}) = w_{\text{syntax}} * \mathcal{E}_{\text{syntax}}(C_{OE_i}, C_{CV_j}) \\ + w_{\text{ontol}} * \mathcal{E}_{\text{ontol}}(C_{OE_i}, C_{CV_j}) + w_{\text{distrib}} * \mathcal{E}_{\text{distrib}}(C_{OE_i}, C_{CV_j}), \end{aligned} \tag{II.49}$$

où w_d est le poids associé à la dimension d de similarité. Ces poids peuvent être déterminés en fonction de l'importance relative ou de la confiance attribuée à chaque dimension dans le contexte organisationnel. Par exemple, si la similarité syntaxique est considérée comme moins importante que la similarité ontologique, son poids pourrait être inférieur.

Nous introduisons également un facteur de tolérance β_S, qui indique le seuil à partir duquel le niveau de similarité dans la dimension d devient significatif pour la compétence du CV en comparaison avec celle de l'OE. Ce seuil est ajusté en fonction des exigences spécifiques du contexte organisationnel.

Ajustement des scores avec l'ambiguïté

L'ambiguïté peut influencer la fiabilité des scores de similarité. Nous ajustons donc les scores de similarité en fonction de l'ambiguïté :

$$S'_d(C_{OE_i}, C_{CV_j}) = (1 - A(C_{OE_i})) \times S_d(C_{OE_i}, C_{CV_j}). \tag{II.50}$$

Il est important de souligner que le facteur de fiabilité $\rho = (1 - A(C_{OE_i}))$ peut être étendu afin d'intégrer des facteurs additionnels de qualité associés à l'extraction de l'information sur les CVs.

Ajustement des scores avec les informations contextuelles du CV

À cette étape, les scores de similarité sont ajustés en fonction des informations contextuelles associées à chaque compétence du CV. Cette ajustement peut soit amplifier, soit diminuer le score initial, selon le contexte dans lequel la compétence est utilisée ou exprimée dans le document :

$$S''_d(C_{OE_i}, C_{CV_j}) = S'_d(C_{OE_i}, C_{CV_j}) \times F(C_{CV_j}), \tag{II.51}$$

où $F(C_{CV_j})$ est une fonction multi-critère qui permet de représenter à quel niveau la compétence professionnelle CV_j est pénalisée ou favorisée à cause de ses informations contextuelles associées.

La fonction multi-critère $F(C_{CV_j})$

La fonction multi-critère $F(C_{CV_j})$ est composée de sous-fonctions qui représentent des informations contextuelles pouvant amplifier ou diminuer l'intérêt d'un recruteur pour une compétence professionnelle mentionnée par un candidat dans son CV. Ainsi, cette fonction peut être représentée par un ensemble de fonctions $f_{m_1}(C_{CV_j}), f_{m_2}(C_{CV_j}), \ldots, f_{m_n}(C_{CV_j})$, modélisant respectivement les informations contextuelles $m_1, m_2, \ldots, m_n$ du CV.

Ces informations contextuelles peuvent présenter divers types de relations, ce qui est propre au raisonnement flexible et éventuellement relatif des recruteurs. Parmi ces types de relations,

nous pouvons trouver :

- *relation complémentaire :* certaines informations contextuelles du CV peuvent se compléter mutuellement pour fournir une vision plus complète de la compétence du candidat ; par exemple, les années d'expérience et les certifications peuvent ensemble offrir une image plus robuste de la qualification d'un candidat dans une compétence spécifique ;
- *relation de dépendance :* une information contextuelle peut dépendre d'une autre ; par exemple, la pertinence de la mobilisation d'une compétence dans un contexte de management peut dépendre du nombre d'années pendant lesquelles cette compétence a été mobilisée dans un contexte technique ;
- *relation d'exclusion :* certaines informations contextuelles peuvent être exclusives et ne pas coexister pour une compétence donnée ; si elles apparaissent ensemble dans un CV, cela peut indiquer une potentielle contradiction de la part du candidat ; exemple : une information contextuelle indique que le candidat possède des compétences en gestion d'équipes de développement de logiciels tandis qu'une autre information contextuelle indique qu'il n'a pas d'expérience technique en tant que développeur de logiciels ;
- *relation de redondance :* des informations contextuelles peuvent transmettre des informations similaires ou redondantes ; par exemple, une information contextuelle indique que le candidat a plus de 10 ans d'expérience dans une compétence tandis qu'une autre indique qu'il a un niveau senior d'expertise.

L'identification des relations entre les informations contextuelles du CV peut conduire à l'identification de séquences et/ou de hiérarchies de critères de décision inhérents au raisonnement des recruteurs. L'interaction entre celles-ci et leur importance relative peut être déterminée à partir d'un processus analytique hiérarchique flou pour la modélisation multicritère (annexe K).

II.12.3 Introduction au processus analytique hiérarchique flou pour la modélisation multi-critère

Compte tenu du scénario présenté, il est crucial d'adopter une approche floue pour saisir les incertitudes et imprécisions associées aux points de vue des recruteurs. Le Processus Analytique Hiérarchique Flou (FAHP, Fuzzy Analytic Hierarchy Process) offre un cadre structuré pour déduire l'importance de chaque critère de la fonction multi-critère $F(C_{CV_j})$ par rapport aux compétences professionnelles mentionnées dans les CV. Les étapes principales sont les suivantes.

1. **Création de la hiérarchie du processus de décision :**
   - la première étape est fondamentale pour structurer le processus décisionnel en hiérarchisant les objectifs, critères et sous-critères (ou alternatives) ; elle permet d'organiser et de clarifier les éléments clés du processus de décision, facilitant ainsi l'analyse subséquente ;

— niveau supérieur (objectif) : déterminer la pertinence d'une compétence du CV à partir de ses informations contextuelles associées ;
— niveau intermédiaire (critères) : compétences professionnelles essentielles extraites de l'OE ;
— niveau inférieur (sous-critères ou alternatives) : informations contextuelles du CV influençant la perception des recruteurs sur la pertinence de la compétence citée par le candidat.

2. **Comparaison floue par paires :**
— la comparaison floue par paires permet de capturer les nuances et les incertitudes dans les jugements des recruteurs ; elle permet d'exprimer des préférences relatives entre les informations contextuelles du CV de manière qualitative, offrant ainsi une représentation plus fidèle et flexible des opinions des recruteurs ;
— pour chaque compétence essentielle de l'OE, une comparaison par paires floue est effectuée entre les informations contextuelles du CV associées ; les recruteurs expriment, via des termes linguistiques (p. ex. "beaucoup plus important", "moins important"), la mesure dans laquelle une information contextuelle du CV est plus importante qu'une autre.

3. **Construction d'une matrice de comparaison floue :**
— cette étape transforme les jugements linguistiques en nombres flous, facilitant ainsi les calculs et analyses ultérieurs ; elle permet également de visualiser et de comprendre comment chaque information contextuelle du CV contribue à la pertinence des compétences professionnelles essentielles.
— pour chaque compétence essentielle de l'OE, des matrices de comparaison floues sont créées, associant et évaluant les informations contextuelles du CV ;
— ainsi, chaque matrice met en évidence comment chacune des informations contextuelles contribue à la pertinence de la compétence essentielle de l'OE, mentionnée dans le CV.

4. **Calcul des poids flous :**
— des poids flous pour chaque information contextuelle du CV sont déterminés en calculant la moyenne géométrique de chaque ligne dans les matrices de comparaison floues ; ils sont ensuite normalisés ;
— le calcul des poids flous est essentiel pour quantifier l'importance relative de chaque information contextuelle du CV ; il offre une base solide pour l'analyse et l'interprétation des résultats, en fournissant des indicateurs quantitatifs de l'importance des différents éléments.

5. **Défuzzification :**
— une défuzzification (méthode du centroïde) est effectuée pour obtenir des poids nets

et comparables, représentant l'influence relative de chaque information contextuelle du CV sur la compétence professionnelle concernée ;
- la défuzzification transforme les nombres flous en valeurs scalaires ; elle facilite la comparaison et l'analyse des résultats, en fournissant des valeurs claires et compréhensibles qui peuvent être facilement interprétées et utilisées dans le processus de décision ;
- ce processus est répété pour chaque compétence professionnelle essentielle identifiée dans l'OE, permettant ainsi d'obtenir une matrice de poids claire et justifiée.

6. **Vérification de consistance :**
   - la consistance des matrices de comparaison floues est assurée et ajustée selon les perspectives des recruteurs ;
   - cette étape assure que les matrices de comparaison floues sont cohérentes et reflètent fidèlement les opinions des recruteurs ; elle contribue à la fiabilité et à la validité des résultats obtenus.
7. **Analyse :**
   - les poids défuzzifiés associés à chaque information contextuelle sont analysés en relation avec chaque compétence essentielle de l'OE ; cette analyse permet de déterminer le degré d'influence que ces informations du CV, exercent sur la perception de l'importance des compétences par les recruteurs.
   - cela offre des aperçus de valeur sur l'influence relative des différentes informations contextuelles du CV sur la perception des compétences par les recruteurs.

Suite à l'application des étapes méthodologiques détaillées ci-dessus, une modélisation simplifiée peut être envisagée pour illustrer concrètement comment les informations contextuelles du CV interagissent et peuvent être intégrées dans l'évaluation des compétences. Par exemple, une telle modélisation pourrait identifier une relation de conjonction entre les différentes informations contextuelles du CV, conduisant à une fonction multi-critère minimaliste comme la suivante :

$$F(C_{CV_j}) = ||t_1 \times f_1(C_{CV_j}) + t_2 \times f_2(C_{CV_j}) + ... + t_m \times f_m(C_{CV_j})|| \tag{II.52}$$

où t_m sont les poids qui reflètent l'importance relative de chaque information contextuelle du CV du point de vue des recruteurs.

Agrégation possibiliste pour chaque compétence

Ainsi, nous procédons à une agrégation possibiliste de l'ensemble de compétences essentielles de l'OE par rapport aux compétences associées du candidat. Considérant les scores de similarité comme une évidence que le candidat possède la compétence professionnelle spécifique, nous

définissons deux mesures. Le niveau d'évidence maximum qui permet de conclure que le candidat possède la compétence professionnelle et le niveau d'évidence minimum selon toutes les évidences disponibles. Ainsi, pour chaque compétence C_{OE_i} de l'OE, le niveau d'évidence de la compétence dans le CV du candidat est défini comme :

$$\Pi(C_{OE_i}, \mathcal{G}_{CV}) = \max_{C_{CV_j} \in \mathcal{G}_{CV}} S''_d(C_{OE_i}, C_{CV_j}), \tag{II.53}$$

où $\mathcal{G}_{CV}$ représente l'ensemble de compétences du CV avec un niveau de similarité supérieur au facteur de tolérance β_S. L'opérateur Π représente le niveau maximum de possibilité sous lequel on considère que le candidat possède la compétence C_{OE_i}. De plus, le niveau de possibilité minimum sous lequel on considère que le candidat possède la compétence professionnelle est défini par l'opérateur de possibilité garantie :

$$\triangle(C_{OE_i}, \mathcal{G}_{CV}) = \min_{C_{CV_j} \in \mathcal{G}_{CV}} S''_d(C_{OE_i}, C_{CV_j}), \tag{II.54}$$

où $\triangle$ est le niveau minimum ou garanti de possibilité sous lequel nous considèrons que le CV du candidat correspond à la compétence C_{OE_i}.

Calcul du score global de possibilité maximale et de possibilité minimale garantie pour le CV du candidat

Ainsi, on estime globalement à quel niveau de possibilité le candidat peut être considéré comme possédant la compétence essentielle mentionnée dans l'OE, ainsi que le niveau minimum global de possibilité qui justifie sa pertinence :

$$\Pi_{\text{global}} = \frac{1}{|\mathcal{E}_{OE}|} \sum_{C_{OE_i} \in \mathcal{E}_{OE}} \Pi(C_{OE_i}, \mathcal{G}_{CV}) \tag{II.55}$$

$$\triangle_{\text{global}} = \frac{1}{|\mathcal{E}_{OE}|} \sum_{C_{OE_i} \in \mathcal{E}_{OE}} \triangle(C_{OE_i}, \mathcal{G}_{CV}), \tag{II.56}$$

où $|\mathcal{E}_{OE}|$ est le nombre de compétences essentielles extraites de l'OE.

Ici, Π_{global} est le score global de possibilité pour le CV, basé sur le meilleur scénario possible de correspondance pour chaque compétence, et $\triangle_{\text{global}}$ est le score global garanti pour le CV, basé sur le pire scénario possible de correspondance pour chaque compétence.

Ces deux scores combinés offrent une perspective plus globale sur la pertinence du CV par rapport à l'OE. Ils peuvent être utilisés soit individuellement, soit en combinaison, pour classer et évaluer les CVs en fonction de leur adéquation avec l'OE.

Fusion des scores globaux

Enfin, afin de produire des classements comparatifs de CVs, un score global de pertinence est généré pour chaque CV. Soient γ_π et $\gamma_\triangle$ les poids associés aux mesures de possibilité maximale et garantie, respectivement.

Le score global fusionné pour un CV est donné par :

$$\text{Score}_{\text{global}} = \gamma_\pi \times \Pi_{\text{global}} + \gamma_\triangle \times \triangle_{\text{global}}. \tag{II.57}$$

Le score $\text{Score}_{\text{global}}$ représente une combinaison des mesures de possibilité maximale et minimale pondérée par leur importance relative (comme défini par γ_π et $\gamma_\triangle$). Ces poids peuvent être déterminés empiriquement pour optimiser la performance du système dans un contexte d'application spécifique.

En conclusion, la formalisation présentée décompose le processus de CCO en plusieurs étapes. Elle ajuste les scores de similarité en fonction de l'ambiguïté et des informations contextuelles du CV relatives aux compétences, puis agrège ces scores pour obtenir une évaluation globale plus alignée sur les avis des recruteurs. Cette approche offre une vue synthétique et explicative de la manière dont chaque élément (ambiguïté, compétences essentielles de l'OE, informations contextuelles du CV, différentes dimensions de similarité) influence l'évaluation finale de la pertinence d'un CV par rapport à une OE.

II.12.4 Résumé de l'approche de CCO proposée

Dans ce chapitre, nous avons introduit notre approche pour la CCO basée sur les représentations de l'OE et du CV décrites précédemment. Nous avons détaillé les éléments fondamentaux de cette approche, y compris les ensembles, les fonctions de similarité et les mécanismes d'ajustement et d'agrégation de scores. Ensuite, nous avons proposé une formalisation détaillée de notre approche, décomposant le processus d'évaluation en étapes distinctes et expliquant comment chaque élément contribue à l'évaluation finale de la pertinence d'un CV.

L'approche proposé a plusieurs implications pour le processus de la CCO :

— **réduction de l'espace de recherche :** les traitements proposés permettent d'éliminer les termes non pertinents des OE, réduisant ainsi l'espace de recherche pour les correspondances potentielles. Au lieu d'une correspondance qui part d'un espace entre tous les termes possibles des deux documents, nous ciblons un sous-ensemble plus restreint, pertinent et interprétable. Cela non seulement facilite le processus d'appariement en réduisant la complexité computationnelle, mais aussi améliore la qualité des correspondances en écartant les termes généraux et bruyants des documents ;

— **augmentation de la précision :** la segmentation plus robuste des CVs pour identifier les sections les plus importantes, comme l'expérience professionnelle, signifie qu'il est pos-

sible d'attribuer une priorité ou un poids différent à différentes sections du CV selon les perspectives de pertinence de l'information des recruteurs ; par exemple, les compétences ou les qualifications mentionnées dans la section "expérience professionnelle" peuvent avoir plus de poids ou d'importance que celles mentionnées dans d'autres sections moins contextualisées comme la section de “Compétences” ; cela contribue à augmenter la précision de la correspondance en donnant la priorité aux informations qui sont plus d'intérêt pour l'organisation ;

— **meilleure interprétabilité :** en se concentrant sur les concepts-clés, les résultats de la correspondance deviennent plus interprétables ; il est plus facile pour les recruteurs de comprendre pourquoi une certaine correspondance a été faite ; cela permet d'identifier des possibles inconsistances logiques du processus de correspondance automatique ;

— **gestion d'ambiguité et informations contextuelles du CV :** l'analyse sémantique préliminaire et la segmentation des CVs peuvent également aider à réduire l'incertitude inhérente à la correspondance automatique ; le contexte fourni par la structure du CV peut aider à mieux déterminer la pertinence de l'information pour l'appariement ;

— **compatibilité avec les approches DL** : l'approche proposée, ainsi que les différentes informations de valeur qu'elle permet de dériver, peuvent être exploitées par les modèles de DL, afin de mieux guider les premières phases d'entraînement ; ceci pourrait potentiellement aider à éviter l'utilisation opaque et incertaine de termes non pertinents du CV et de l'OE lors de l'optimisation automatique de l'architecture neuronale ;

— **évolution :** si les OE évoluent ou si de nouveaux formats de CV apparaissent, le processus d'analyse et de segmentation peut nécessiter des ajustements ; en ayant une structure bien plus formelle et consistente avec le contexte organisationnel, il est possible d'intégrer plus facilement les changements.

II.13 Explicabilité dans la méthodologie présentée

La méthodologie proposée est fondée sur des modèles explicables adaptés aux spécificités du contexte organisationnel. Ces modèles permettent de comprendre en profondeur les facteurs influençant le processus de sélection des CV du début à la fin.

Premièrement, l'extraction des termes essentiels des OE aide à identifier les informations les plus significatives de ces documents. Les informations extraites sont interprétées grâce aux marqueurs textuels qui soulignent leur pertinence. Notamment, ces marqueurs ne sont pas des règles sémantiques accessibles uniquement aux utilisateurs techniques. Ils peuvent être décrits en langage naturel, compréhensible même par des recruteurs sans connaissance technique de la CCO.

Deuxièmement, le cadre d'évaluation de la qualité des marqueurs textuels des OE fournit

plus de clarté au niveau du comportement des marqueurs dans des contextes organisationnels. Il met en lumière des aspects tels que leur capacité à définir clairement la pertinence de l'information du point de vue des recruteurs (mesure d'ambiguïté) ou l'étendue de l'information qu'ils représentent par rapport à l'extraction des détails essentiels de tels documents (entropie mutuelle de l'information). Cela est renforcé par d'autres éléments d'explicabilité, comme la visualisation des corrélations et l'association de catégories de qualité pour chaque métrique. Cela permet de décrire chaque métrique à travers de termes linguistiques (qualité élevée, moyenne, faible) pour les utilisateurs non techniques. Tout cela est applicable même aux marqueurs reposant sur des modèles DL opaques (e.g. les *transformers*), comme le marqueur M_8 dérivé du contexte organisationnel.

Troisièmement, concernant l'architecture BERT proposée pour la segmentation des CV, il est essentiel de noter qu'elle intègre des marqueurs grapholinguistiques. Lorsqu'ils sont directement incorporés dans le texte des séquences traitées dans l'architecture proposée, il devient possible de discerner les caractéristiques spécifiques sur lesquelles le modèle a choisi de segmenter le CV. Cela ajoute une couche d'explicabilité à l'architecture BERT, dont les caractéristiques et leur importance sont aussi clarifiées à l'aide d'un modèle de régression logistique interprétable.

Enfin, la méthodologie vise à renforcer son explicabilité dans les dimensions de représentation et de réduction de l'incertitude. Elle permet d'associer des niveaux d'ambiguïté aux termes extraits des OE selon les points de vue des recruteurs. Elle permet d'attribuer des niveaux de confiance à chaque source d'information, et elle fournit un cadre pour propager l'incertitude liée à ces sources tout au long du processus d'extraction des termes essentiels des OE.

II.14 Conclusion

Cette thèse aborde le défi crucial de déterminer les composants essentiels du traitement pour une extraction robuste de l'information, une structuration et une annotation sémantique des CV et des OE dans le contexte de CCO. Notre objectif était de définir une approche optimisée de la présélection des candidats dans le recrutement.

Pour relever ce défi, nous introduisons une méthodologie qui se concentre principalement sur les OE mais qui est également adaptable aux CV. Cette méthodologie est délimitée en cinq étapes clés :

- développement et intégration d'ontologies de ressources dérivées de la représentation du contexte organisationnel des recruteurs ;
- construction de représentations des OE et des CV, en tenant compte des perspectives des recruteurs ;
- établissement d'une architecture possibiliste ancrée dans des croyances, désirs et intentions pour l'extraction d'informations pertinentes à partir des OE, visant à améliorer le

classement des CV ;
— évaluation de la qualité des marqueurs textuels pertinents dans les OE ;
— introduction d'une architecture BERT basée sur la grapholinguistique pour affiner le processus de segmentation des CV, améliorant ainsi l'extraction des informations pertinentes pour la CCO, habituellement localisées dans les sections d'expérience professionnelle et d'éducation.

Au-delà de cette méthodologie, il convient de souligner deux éléments distinctifs de notre approche dans la CCO. Premièrement, elle intègre le principe théorique de l'incertitude, offrant une perspective originale dans le domaine. Deuxièmement, elle préconise une approche grapholinguistique, mettant l'accent sur la dimension graphique des documents, en particulier les CV.

Chapitre III

Application de la méthode

Dans ce chapitre, nous présenterons quatre études de cas menées dans le cadre de l'application de la méthodologie proposée. Tout d'abord, nous fournirons un exemple basique d'extraction de l'information afin d'illustrer le paradigme possibiliste CDI proposé dans la thèse. Les documents présentés dans cet exemple correspondent à des documents synthétiques basés sur des scénarios réels.

Pour le premier cas d'étude, nous mettrons en évidence les résultats les plus significatifs obtenus en comparant les approches probabiliste et possibiliste. Nous mettrons un accent particulier sur deux modèles possibilistes : la Régression Logistique Floue (RLF, modèle linéaire) et l'Arbre de Décision Flou (ADF, modèle non-linéaire).

Après avoir évalué le potentiel des modèles flous pour une modélisation robuste des documents, nous introduirons un deuxième cas d'étude. Celui-ci sera centré sur les résultats obtenus grâce à l'application de la méthodologie proposée, visant à construire une représentation de l'OE plus robuste. Cette approche se concentre sur l'extraction des informations les plus essentielles de ce document sur la base de l'architecture d'agent CDI possibiliste basée sur des ontologies et des ADF.

Dans le troisième cas d'étude, nous aborderons l'évaluation et l'optimisation de la qualité des marqueurs textuels dans les OE au sein d'un contexte organisationnel spécifique, en utilisant une approche d'inférence floue. Cette optimisation a pour objectif d'améliorer les performances de l'agent CDI dans la tâche d'extraction d'informations à partir des OE.

Enfin, nous présenterons les résultats et la discussion correspondante relatifs à l'application de la méthodologie conçue pour la représentation et l'analyse automatique des CV. Ce quatrième cas d'étude, enrichi par une approche floue de représentation et des modèles basés sur BERT, s'appuie sur la représentation grapholinguistique. Notre travail expérimental s'est principalement concentré sur l'optimisation d'une tâche cruciale pour toute analyse automatisée et approfondie du CV : la segmentation et l'identification des sections du document.

Les expériences précédentes font partie d'un ensemble d'expériences progressives qui ont permis d'appliquer et d'évaluer les principales hypothèses et dimensions proposées dans la méthodologie en vue d'optimiser la représentation du CV et de l'OE pour la CCO.

III.1 Description des corpora d'OE et de CV utilisés pour les différentes expérimentations

Les expérimentations de cette thèse reposent sur l'analyse approfondie de trois corpus distincts, regroupant à la fois des OE et des CV.

Contenu du corpus :

- **corpus CP1** : il met en correspondance 98 OE et 473 CV ; utilisé progressivement lors des expérimentations 1, 2 et 3 ;
- **corpus CP2** : il contient 303 CV et est utilisé pour l'évaluation de la segmentation des CV ; utilisé progressivement lors de l'expérimentation 4 ;
- **corpus CP3** : il rassemble 14.000 CV et 2.000 OE non appariés ; ces derniers sont utilisés pour des tâches d'ajustement fin, notamment pour l'adaptation des ontologies à la connaissance spécifique du marché.

Si la majorité des documents sont rédigés en français, ils intègrent souvent des termes anglais propres au domaine des TIC (par exemple pour le langage de programmation «Go»).

Provenance :

- **source** : les documents proviennent des processus de recrutement de la société de consulting DSI Group ;
- **période de collecte** :
  - **corpus CP1** : collecté progressivement entre 2020 et 2022 ;
  - **corpus CP2** : collecté entre 2021 et 2022 ;
  - **corpus CP3** : collecté entre 2014 et 2019.
- **taille des documents** :
  - **CV** : en moyenne, ils font une page et demi (A4, Arial 12), avec des longueurs minimale d'une page et maximale de cinq pages ;
  - **OE** : elles font en moyenne une page, avec des longueurs minimale d'une demi-page et maximale de deux pages.

Représentativité :

- **secteur d'activité** : les documents sont spécifiques au secteur des TIC ;
- **niveaux de postes** : ils couvrent des rôles techniques, managériaux et fonctionnels ;
- **diversité géographique** : l'ensemble des CV et OE sont spécifiques au marché français.

Traitement préalable :

Les documents sont disponibles dans une grande diversité de formats (PDF, Microsoft Word, texte brut, HTML, ...) sans traitement préalable.

Considérations éthiques et légales :

Le jeu de données a été élaboré en respectant strictement les directives sur la protection des données. Il intègre des techniques d'anonymisation des documents.

Accessibilité :

En raison de la sensibilité des informations contenues dans les documents, les corpus ne sont pas diffusés.

III.2 Exemple introductif de la méthodologie proposée

Dans cette section, nous présentons un bref exemple d'extraction d'information sur les OE selon la méthodologie proposée afin d'illustrer le fonctionnement de base de l'architecture CDI complexe. De même, nous illustrons brièvement l'approche proposée pour la représentation grapholinguistique des CV. Les fondements théoriques pour comprendre plus en profondeur les exemples et le fonctionnement de l'agent CDI se trouvent dans l'annexe A.

III.2.1 Example d'extraction d'informations sur les OE

Le traitement automatique des OE commence par le pré-traitement et la segmentation du document. Généralement, dans certains contextes d'application spécifiques, les OE peuvent suivre un format relativement prédéterminé, qui peut essentiellement être décomposé en les sections suivantes : titre, description de l'entreprise, description du poste offert, description du candidat recherché, et autres informations telles que les détails du contrat, la localisation, le salaire, etc. La qualité de la segmentation peut être mesurée à travers la proportion de sections correctement spécifiées.

De plus, comme une OE représente en soi une source d'information, la qualité de son texte peut être évaluée à l'aide d'une ou plusieurs métriques de qualité. Dans cet exemple, supposons qu'il y a une seule métrique de qualité utilisée, consistant en la proportion du vocabulaire de l'OE trouvé dans la base de données Wikipédia (section II.8.3). Supposons que l'OE introduite dans le chapître II, désignée par d_i, ait un taux de qualité de $\tau_{\text{interpret}} = 0,95$. Supposons également que l'agent rejette les sources d'information ayant un taux inférieur à $\beta_{\text{interpret}} = 0,85$. Ainsi, l'agent traite cette OE où le recruteur, en charge du processus de recrutement, a annoté (souligné) 6 termes pertinents que nous présentons dans le texte qui suit :

Ingénieur Logiciel **TAL** - Systèmes Financiers & Sécurité

Description de l'entreprise: Nous sommes une entreprise fintech leader axée sur les systèmes financiers et la sécurité. Opérant mondialement dans les principaux centres financiers, notre équipe diversifiée utilise des technologies avancées, l'analytique des données, et l'expertise du secteur pour créer des solutions de premier ordre. Nous servons des banques, des institutions financières, des sociétés d'investissement et des startups fintech avec des produits personnalisés, visant à remodeler la finance avec des systèmes sécurisés et intelligents.

Description du poste: En tant qu'Ingénieur Logiciel, vous travaillerez en étroite collaboration avec nos experts financiers et notre équipe de sécurité pour développer et mettre en œuvre des systèmes financiers à grande échelle. Vos responsabilités incluront :

- Conception et **développement des logiciels bancaires sécurisés** avec un accent sur le TAL.
- Rencontrer des **équipes interfonctionnelles** pour assurer l'intégration avec d'autres applications bancaires.
- Enrichir les **protocoles de sécurité** avec le TAL pour la protection des données bancaires.
- Se conformer aux réglementations de l'industrie dans le secteur financier.

Description du profil :
Expérience : 3 ans et plus en développement des logiciels financiers avec un accent sur le TAL.
Compétences : Programmation en Python. Connaissance approfondie du **modèle BERT**.
Prérequis : Connaissance des **réglementations financières** et des protocoles de sécurité.
Détails du contrat : Poste à temps plein avec salaire et avantages compétitifs.
Lieu : Metz, France.

L'interprétation de l'OE exemple

L'interprétation de l'OE précédente est présentée du point de vue de l'ensemble de marqueurs textuels (M_1-M_{10}) dérivés des stratégies des recruteurs :

- M_1 **:** le titre « Ingénieur logiciel TAL - systèmes financiers & sécurité » met distinctement en évidence des compétences professionnelles (telles que TAL, Systèmes Financiers et Sécurité) et des types d'emploi (Ingénieur Logiciel) ; cela résonne avec la prémisse que les titres encapsulent souvent les responsabilités ou compétences principales, renforçant leur pertinence ;
- M_2 **:** les termes au sein des descriptions du poste et du profil mettent en avant des compétences professionnelles particulières, comme « applications bancaires », « diriger des équipes interfonctionnelles » et « Python » ; ces expressions sont alignées avec les tâches et qualifications clés, affirmant leur importance dans la représentation des compétences attendues ;
- M_3 **:** la plupart des termes pertinents résident dans le titre, la description du poste et la description du profil ; cela reflète la stratégie de recrutement courante qui consiste à mettre en évidence les informations essentielles dans ces sections, en laissant d'autres détails pour un contexte supplémentaire ;
- M_4 **:** des expressions comme « familiarité », « connaissance de » ou « connaissance approfondie » véhiculent des dépendances syntaxiques, renforçant le caractère entrelacé des rôles et compétences et accentuant leur importance pour le poste ; cela s'applique à des compétences comme Python, le modèle BERT et les réglementations financières ;
- M_5 **:** les interactions avec les concepts professionnels, manifestes dans des déclarations comme « travailler en étroite collaboration avec nos experts financiers et notre équipe de sécurité », symbolisent l'esprit collaboratif du rôle, soulignant la valeur du travail d'équipe ;
- M_6 **:** des termes comme « développement des logiciels bancaires sécurisés » et « protocoles de sécurité » sont cruciaux car des erreurs pourraient affecter profondément l'entreprise ; leur inclusion accentue la nécessité de précision et de maîtrise dans ces domaines ;
- M_8 **:** la référence aux « modèles BERT » confirme que la compréhension de modèles TAL spécifiques est essentielle ; cette connaissance, lorsqu'elle est connectée à d'autres termes mettant en avant les tâches en TAL (Python, Ingénieur Logiciel TAL...), confirme le besoin d'une expertise approfondie dans ce domaine, validant sa pertinence ;
- M_9 **:** les mentions au secteur de finances soulignent explicitement les facettes économiques du rôle, mettant en évidence l'importance de la familiarité avec le secteur de l'activité financière ;
- M_{10} **:** dans cette OE, des prérequis sont explicitement décrits, avec des termes comme « réglementations financières » et « protocoles de sécurité », suggérant les connaissances

de base requises, indiquant ainsi leur pertinence.

Aperçu sur l'intégration des croyances de l'agent

L'analyse de l'OE est réalisée par l'agent sur la base de la terminologie du document. Fondamentalement, des termes simples (composés d'un seul mot comme "sécurité") et des termes complexes (composés de plusieurs mots, comme "protocoles de sécurité") sont extraits. C'est sur cet ensemble de termes que l'agent effectue l'analyse pour extraire les termes les plus pertinents. Pour cette analyse, l'agent construit des croyances qui, dans le contexte de cette thèse, sont représentées comme des marqueurs textuels associés à la pertinence de l'information. Ces marqueurs sont composés d'un ensemble de propositions atomiques liées à la pertinence des termes qui ont déjà été illustrées au chapitre II.

Nous utiliserons la matrice d'états de l'agent introduite dans la section II.9.7 pour illustrer la mise à jour des croyances de l'agent. Nous commencerons par utiliser une table de vérité élémentaire et, à travers un exemple, nous l'étendrons en utilisant la distribution de possibilité afin d'expliciter les croyances graduées de l'agent sur la pertinence des termes.

Illustrons donc ce processus en supposant initialement que l'agent dispose d'un seul marqueur textuel, le marqueur M_1. Ce marqueur est principalement composé de deux propositions atomiques : le terme appartient à l'ensemble de termes qui représentent une compétence professionnelle ou un type de travail de l'ontologie-mère s_1, et le terme appartient au titre de l'OE s_2. Ainsi, le marqueur M_1 est fondamentalement constitué par les expressions atomiques $A_{M_1} = \{s_1, s_2\}$.

Définissons également une proposition atomique supplémentaire qui représente la vérité r_{t_i} sur la pertinence d'un terme t_i de l'OE (cf. "Sécurité"). Ce qui précède nous permet de construire le tableau de vérité suivant, servant de base pour suivre les croyances de l'agent :

s_1	s_2	r_{t_i}
1	1	1
1	1	0
1	0	1
1	0	0
0	1	1
0	1	0
0	0	1
0	0	0

Le tableau de vérité ci-dessus permet de représenter toutes les possibilités ($2^3 = 8$) (mondes possibles) auxquelles l'agent peut être confronté lorsqu'il juge de la pertinence du terme t_i. Il s'agit, en essence, d'un moyen complet d'explicabilité concernant le processus de prise de décision

de l'agent.

Supposons maintenant que dans le contexte organisationnel où l'OE est traitée, la relation entre les propositions atomiques des marqueurs et la pertinence d'un terme est représentée par une implication matérielle. Avant de poursuivre et d'entrer plus dans les détails, clarifions certains points.

Premièrement, soit ϕ_{M_1} une expression définie pour simplifier notre représentation. Cette expression, $\phi_{M_1} = s_1 \wedge s_2$, représente la conjonction (c'est-à-dire l'intersection ou la condition "ET") entre les propositions atomiques du marqueur M_1. En d'autres termes, ϕ_{M_1} est vrai (ou égal à 1) si et seulement si s_1 et s_2 sont toutes deux vraies.

Deuxièmement, introduisons l'expression de l'implication matérielle utilisée. L'implication matérielle, notée $A \rightarrow B$, est une façon de relier deux propositions. Elle est vraie sauf dans le cas où A est vrai et B est faux. Autrement dit, si nous avons une implication $\phi_{M_1} \rightarrow r_{t_i}$, celle-ci sera fausse uniquement lorsque ϕ_{M_1} est vrai et r_{t_i} est faux.

Avec ces clarifications, considérons maintenant le tableau suivant simplifié pour illustrer les mondes possibles et les interprétations (# I) que l'agent peut faire :

$\#I$	ϕ_{M_1}	r_{t_i}	$\phi_{M_1} \rightarrow r_{t_i}$
1	1	1	1
2	1	0	0
3	0	1	1
4	0	0	1

Chaque ligne du tableau représente une combinaison possible de valeurs de vérité pour ϕ_{M_1} et r_{t_i}, et la colonne $\phi_{M_1} \rightarrow r_{t_i}$ montre le résultat de l'implication matérielle pour chaque combinaison. L'explication de chaque interprétation est la suivante :

— *interprétation 1 :* si les conditions du marqueur sont vraies, et que le terme est pertinent, l'implication est vraie ; il s'agit de l'interprétation centrale qui permet l'association des degrés de possibilité fournis par les marqueurs avec la pertinence des termes ;

— *interprétation 2 :* si les conditions du marqueur sont vraies, et que le terme n'est pas pertinent, nous considérons que l'implication est invalide ;

— *interprétation 3 :* si les conditions du marqueur ne sont pas remplies, et que le terme est pertinent, l'agent choisira de croire que le fait que les conditions du marqueur ne soient pas remplies n'est pas incompatible avec le fait que le terme soit pertinent ; en général, les marqueurs textuels sont des indicateurs de pertinence, pas des vérités absolues ; en d'autres termes, le fait que le marqueur ne considère pas le terme comme pertinent n'exclut pas que le terme puisse être pertinent ; il peut y avoir d'autres preuves ou marqueurs qui soutiennent la pertinence du terme ; par conséquent, l'implication est vraie ;

— *interprétation 4 :* si les conditions du marqueur ne sont pas remplies et que le terme n'est

pas pertinent, il n'y a pas de conflit entre le marqueur et la pertinence du terme, rendant l'implication vraie ; autrement dit, puisque le marqueur ne considère pas le terme comme pertinent, et que le terme n'est effectivement pas pertinent, il n'y a pas de conflit entre l'évaluation du marqueur et la pertinence réelle du terme.

Les interprétations vraies (1, 3 et 4) de l'implication matérielle conduisent essentiellement à la préservation des croyances actuelles de l'agent sur la pertinence du terme, tandis que l'interprétation invalide (numéro 2) conduit à une réduction de la croyance de l'agent sur l'influence possible du marqueur sur la pertinence du terme.

Comme on peut le voir dans le tableau ci-dessus, les propositions atomiques et la valeur de vérité de l'implication sont fournies en valeurs binaires. Cependant, dans le contexte des OE, la présence d'incertitude nécessite une représentation plus flexible. Dans notre contexte, cela se concentre sur ce que l'on appelle les croyances graduées de l'agent. Pour rendre l'exemple précédent gradué, représentant le phénomène d'incertitude, introduisons la distribution de possibilité π associée aux croyances de l'agent :

$\#I$	ϕ_{M_1}	r_{t_i}	$\phi_{M_1} \rightarrow r_{t_i}$	π
1	1	1	1	**1**
2	1	0	0	**1**
3	0	1	1	**1**
4	0	0	1	**1**

Dans le tableau précédent, la nouvelle modification est affichée en gras. Nous attribuons initialement une valeur de 1 à toutes les interprétations, supposant qu'initialement, l'agent ignore complètement la pertinence du terme (voir annexe A).

Par la suite, nous illustrerons comment les informations fournies par les marqueurs sont reçues par l'agent et adoptées comme croyances. Supposons qu'en tant que première information, l'agent reçoit la notification que le marqueur ϕ_{M_1} permet d'expliquer la pertinence des termes, et que l'implication $\phi_{M_1} \rightarrow r_{t_i}$ est associée à un taux de fiabilité de 0,9. Dans le contexte de cette thèse, la mesure analytique de l'ambiguïté est principalement utilisée à cette fin. L'intégration de cette information dans la base de croyances de l'agent est effectuée en utilisant l'opérateur de mise à jour des croyances (section II.9.3) :

$$\pi'(I) = \begin{cases} \frac{\pi(I)}{\Pi([\phi])} & \text{si } I \models \phi \text{ et } B(\neg\phi) < 1 \\ 1, & \text{si } I \models \phi \text{ et } B(\neg\phi) = 1 \\ \min(\pi(I), 1-\tau) & \text{si } I \not\models \phi. \end{cases}$$

L'opérateur ci-dessus peut être résumé comme suit. Pour chaque interprétation associée à de nouvelles informations entrantes, trois scénarios sont évalués :

— *scénario 1* ($I \models \phi \, et \, B(\neg\phi) < 1$) *:* si sous l'interprétation I, l'information entrante ϕ (une expression logique) est considérée comme vraie, et que l'agent ne croit pas le contraire, alors la croyance actuelle sur cette interprétation est consolidée, normalisée par le degré maximal auquel l'agent croit en la nouvelle information ;
— *scénario 2* ($I \models \phi \, et \, B(\neg\phi) = 1$) *:* si sous l'interprétation I, l'information entrante ϕ est considérée comme vraie, mais que $B(\neg\phi) = 1$, alors la priorité sera donnée à l'information entrante ; en d'autres termes, l'agent consolidera les croyances liées à l'information la plus récente ;
— *scénario 3* ($I \not\models \phi$) *:* si l'interprétation I est contre la validité de la nouvelle information, alors la croyance associée à l'interprétation I sera éventuellement réduite par un facteur $1 - \tau$, qui peut être interprété comme le taux d'incertitude associé à la nouvelle information.

Pour illustrer le fonctionnement de cet opérateur, revenons à l'OE précédente, en supposant que l'agent n'a qu'un seul marqueur textuel. Rappelons que la nouvelle information arrivant est : $\phi_{M_1} \rightarrow r_{t_i}$ avec un niveau de certitude de 0,9. En appliquant l'opérateur précédent, le résultat suivant est obtenu :

$\#I$	ϕ_{M_1}	r_{t_i}	$\phi_{M_1} \rightarrow r_{t_i}$	π
1	1	1	1	1
2	**1**	**0**	**0**	**0,1**
3	0	1	1	1
4	0	0	1	1

Si nous évaluons la croyance de l'agent dans l'implication $\psi : \phi_{M_1} \rightarrow r_{t_i}$, nous avons $B(\psi) = N([\psi]) = \inf_{I \not\models \psi}(1 - \pi(I)) = 0,90$. Cela signifie que l'agent croit à la vérité de cette implication avec une nécessité de 0,9 et une incertitude associée de 0,1. Cette incertitude exprime que la même implication pourrait être invalide sous certaines circonstances qui ne peuvent pas être contrôlées en raison de la présence d'incertitude dans la source d'information respective.

Maintenant, supposons qu'un nouveau marqueur textuel soit ajouté à l'architecture de l'agent, par exemple, le marqueur M_{13}, un marqueur YAKE ! indépendant du contexte et associé à la fréquence du terme dans l'OE. L'introduction de ce nouveau marqueur augmente les possibilités que l'agent doit considérer à 2^3. Soit $\phi_{M_{13}}$ l'expression logique qui représente les relations entre les conditions logiques (c'est-à-dire, les propositions atomiques) associées au marqueur M_{13}. Désormais, le tableau de vérité est le suivant :

$\#I$	ϕ_{M_1}	$\phi_{M_{13}}$	r_{t_i}	π
1	1	1	1	1
2	1	1	0	0,1
3	1	0	1	1
4	1	0	0	0,1
5	0	1	1	1
6	0	1	0	1
7	0	0	1	1
8	0	0	0	1

Supposons que dans le contexte organisationnel de gestion de ce marqueur, il est identifié que l'implication $\phi_{M_{13}} \rightarrow r_{t_i}$ est associée à un niveau de confiance de 0,3. Ainsi, l'agent reçoit la nouvelle information, ce qui conduit à la mise à jour de l'ensemble d'interprétations concernées :

$\#I$	ϕ_{M_1}	$\phi_{M_{13}}$	r_{t_i}	π
1	1	1	1	1
2	1	1	0	0,1
3	1	0	1	1
4	1	0	0	0,1
5	0	1	1	1
6	**0**	**1**	**0**	**0,7**
7	0	0	1	1
8	0	0	0	1

Comme on peut le voir, l'interprétation 6 a été mise à jour à $1-0,3$, le niveau d'incertitude de l'implication précédente, donc sa possibilité n'est pas écartée. Dans le même cas, l'interprétation 2, bien qu'étant fausse, n'est pas modifiée, car son niveau de possibilité associé est inférieur au niveau d'incertitude de l'implication associée à M_1.

Après la mise à jour des croyances, il peut être vérifié que $B(\phi_{M_1} \rightarrow r_{t_i}) = N([\phi_{M_1} \rightarrow r_{t_i}]) = 0,9$ et que $B(\phi_{M_{13}} \rightarrow r_{t_i}) = N([\phi_{M_{13}} \rightarrow r_{t_i}]) = 0,3$, reflétant l'intégration des informations des marqueurs comme nouvelles croyances de l'agent. Suivant cette logique, le niveau de croyance dans l'implication basée sur les deux marqueurs $\varrho = \phi_{M_1} \rightarrow r_{t_i} \wedge \phi_{M_2} \rightarrow r_{t_i}$ peut être analysé : $B(\varrho) = N([\varrho]) = 0,3$, ce qui est cohérent avec le théorème de la théorie de la possibilité $N(\phi \wedge \sigma) = min(N(\phi), N(\sigma))$.

Maintenant, supposons qu'après avoir exécuté avec succès le marqueur M_{13}, nous recevions une nouvelle information indiquant que, selon l'équation respective de YAKE !, le terme pourrait être jugé pertinent par un recruteur avec un niveau de possibilité de 0,5. Cette valeur de possibilité peut être obtenue après un processus de fuzzification de la fréquence (section II.9.8).

L'agent adopte cette nouvelle croyance sur $\phi_{M_{13}}$ de la manière suivante en utilisant l'opérateur de croyances :

$\#I$	ϕ_{M_1}	$\phi_{M_{13}}$	r_{t_i}	π
1	1	1	1	1
2	1	1	0	0,1
3	**1**	**0**	**1**	**0,5**
4	1	0	0	0,1
5	0	1	1	1
6	0	1	0	0,7
7	**0**	**0**	**1**	**0,5**
8	**0**	**0**	**0**	**0,5**

Enfin, supposons qu'une nouvelle information arrive indiquant que le marqueur ϕ_{M_1} est vrai avec un niveau de possibilité de 0,95 (par exemple, compte tenu du taux de qualité du document analysé). En mettant à jour l'ensemble des interprétations en conséquence, nous obtenons :

$\#I$	ϕ_{M_1}	$\phi_{M_{13}}$	r_{t_i}	π
1	1	1	1	1
2	1	1	0	0,1
3	1	0	1	0,5
4	1	0	0	0,1
5	**0**	**1**	**1**	**0.05**
6	**0**	**1**	**0**	**0,05**
7	0	0	1	0,05
8	0	0	0	0,05

Après que l'agent ait satisfait ses désirs (exécuter les marqueurs dès que possible pour une nouvelle OE reçue), il procède à satisfaire son désir d'estimer sa croyance globale sur la pertinence du terme. En évaluant $B(r)$, on obtient qu'elle a une valeur de 0,9. C'est-à-dire que l'agent détermine que le terme pourrait être sélectionné par un recruteur comme pertinent, avec un niveau de possibilité de 0,9. Notez que cette valeur est fortement influencée par le marqueur M_1 en raison de son haut niveau de fiabilité, reflétant le principe de l'opérateur de modification de croyances qui privilégie les sources d'information les plus fiables [2].

De cette manière, on peut estimer la croyance globale de l'agent sur la pertinence du terme (annexe E). La croyance moyenne est donnée par :

$$\bar{B}(r) = \frac{1 * 0,9 + 0,5 * 0,3}{2} = 0,525$$

Par conséquent, la croyance globale de l'agent sur la pertinence du terme est déterminée par

l'équation suivante :

$$C(r) = \varphi * B(r) + (1 - \varphi) * \bar{B}(r) = \varphi * 0,900 + (1 - \varphi) * 0,525 \quad \text{(III.1)}$$

En accordant une importance égale aux deux perspectives avec $\varphi = 0,5$, on obtient :

$$C(r) = 0,5 * 0,9 + (1 - 0,5) * 0,525 = 0,45 + 0,26 = 0,71$$

Ainsi, en utilisant la mesure globale, l'agent détermine la pertinence du terme avec un niveau de possibilité de 0,71, une valeur plus équilibrée qui prend en compte à la fois l'influence des croyances les plus solides (niveau de possibilité de 0,9) et l'ensemble des croyances, y compris les plus faibles (niveau de possibilité de 0,525).

Il est important de souligner que le mécanisme fondamental illustré ici s'étend à l'utilisation de n marqueurs textuels associés à la pertinence de l'information. Comme on peut le voir dans cet exemple, la qualité des marqueurs joue un rôle crucial dans la mise à jour de croyances, ce qui a motivé également l'étude de leur qualité [41].

L'extraction de termes pertinents sur l'OE exemple, les résultats de modèles de l'état de l'art et modèles émergents du type chatbot

Sur la base des marqueurs dérivés du contexte (M_1-M_{10}), l'ensemble des termes identifiés comme pertinents par l'agent CDI est affiché dans le tableau III.1 (colonne *CDI Agent*). Ces résultats sont consolidés à travers les métriques de précision, rappel et F1-mesure dans le tableau III.2 (ligne *CDI Agent*). Pour l'agent CDI, les métriques d'ambiguïté de l'expérimentation 3 ont été utilisées (section III.5.3), ces métriques servant de points de référence pour les niveaux de confiance.

Bien que des chatbots tels que ChatGPT n'aient pas été inclus dans l'ensemble des expérimentations menées dans cette thèse, principalement parce qu'ils ont commencé à devenir progressivement disponibles au public en décembre 2022, et de manière plus ouverte en mars 2023, vers la fin de la phase de recherche de cette thèse, nous avons choisi de les inclure dans une très courte expérimentation sur l'OE d'exemple.

Nous leur avons fourni l'OE précédente, en demandant : "Quels sont les 12 termes les plus pertinents de l'annonce d'emploi suivante, classés du plus au moins pertinent ?". Les évaluations de ChatGPT [1], Bard [2], et Bing [3] ont été réalisées le 7/09/23. Une demande de suivi, utilisant la même requête, a été faite à ChatGPT le 12/09/23. Les résultats de Claude 2 [4] ont été exclus en raison de ses performances comparativement plus faibles par rapport aux autres modèles.

1. `https://chat.openai.com/`
2. `https://bard.google.com/`
3. `https://bing.com/`
4. `https://claude.ai/`

Un cas similaire s'est produit avec le modèle open source LLaMA [5], qui a inclus dans le top 12 des termes tels que "Ingénieur Logiciel", "Banques", "Institution financière", "Société d'investissement", "Startup fintech" et "Produits personnalisés".

Finalement, une dernière requête a été effectuée sur le modèle ChatGPT le 15/09/23, mais elle n'a pas été affichée dans le tableau en raison de l'espace disponible. Cette requête a fourni un classement bien plus éloigné de l'avis du recruteur. Dans l'ordre de pertinence, les termes extraits sont les suivants : "Ingénieur Logiciel", "TAL", "Systèmes Financiers & Sécurité", "entreprise fintech", "systèmes financiers", "sécurité", "analytique des données", "expertise du secteur", "banques", "institutions financières", "sociétés d'investissement" et "startups fintech"."

5. https://huggingface.co/blog/llama2

TABLE III.1 – Comparaison effectuée sur les termes extraits par plusieurs modèles : ChatGPT, Bard et Bing. Des comparaisons ont également été menées pour les modèles tfidf, YAKE ! et l'agent CDI.

	tfidf	ChatGPT 7/09/23	ChatGPT 12/09/23	Bard 12/09/23	Bing 12/09/23	YAKE !		CDI Agent		
Pos.	Term	Term	Term	Term	Term	Term	Sco.	Term	Sco.	Amb.
1	Bancaire	Ingénieur Logiciel	Ingénieur Logiciel TAL	Ingénieur logiciel	Ingénieur Logiciel	Ingénieur Logiciel TAL	1.00	Protocoles de Sécurité	1.000	0.30
2	Financier	TAL	Systèmes Financiers	TAL	TAL	Entreprise fintech leader	0.99	Python	0.884	0.24
3	Sécurité	Systèmes Financiers	Sécurité	Systèmes financiers	Systèmes Financiers	Fintech leader axée	0.99	Modèle BERT	0.884	0.24
4	Entreprise	Sécurité	Développement des logiciels bancaires sécurisés	Sécurité	Sécurité	Systèmes Financiers	0.81	Développement des logiciels bancaires sécurisés	0.832	0.23
5	Équipe	Fintech	TAL	Développement	Développement des logiciels bancaires sécurisés	Entreprise	0.56	TAL	0.818	0.29
6	Industrie	Développement de logiciels bancaires sécurisés	Équipes interfonctionnelles	Mise en œuvre	Accent sur le TAL	Leader axée	0.56	Réglementations Financières	0.767	0.19
7	Accent	Intégration	Protocoles de sécurité	Collaboration	Équipes interfonctionnelles	Financiers	0.40	Systèmes Financiers à Grande Échelle	0.748	0.26
8	Régulation	Protocoles de sécurité	Réglementations de l'industrie	Experts	Intégration	Ingénieur Logiciel	0.30	Intégration	0.746	0.27
9	Logiciel	Réglementations de l'industrie financière	Expérience en développement des logiciels financiers	Équipe	Protocoles de sécurité	TAL	0.17	Protection des Données Bancaires	0.653	0.31
10	Fintech	Metz	Python	Protocole	Protection des données bancaires	Systèmes	0.17	Équipes interfonctionnelles	0.611	0.31
11	Sécurisé	Développement de logiciels financiers	Modèle BERT	Réglementation	Poste à temps plein avec salaire et avantages compétitifs	Sécurité	0.05	Experts financiers	0.436	0.42
12	Leading	Python	Réglementations financières	Python	Réglementations de l'industrie dans le secteur financier	Entreprise fintech	∼0.0	Ingénieur Logiciel	0.422	0.45

TABLE III.2 – Résultats de précision, rappel et F1-mesure pour chaque modèle évalué sur l'exemple d'OE.

Model	Précision@6	Rappel@6	F1-mesure@6	Précision@12	Rappel@12	F1-mesure@12
tfidf	0.00	0.00	0.00	0.00	0.00	0.00
ChatGPT 7/09/23	0.33	0.33	0.33	**0.50**	0.50	0.50
ChatGPT 12/09/23	0.50	0.50	0.50	**0.50**	**1.00**	**0.67**
Bard 12/09/23	0.17	0.17	0.17	0.08	0.17	0.11
Bing 12/09/23	0.33	0.33	0.33	0.33	0.67	0.44
YAKE!	0.00	0.00	0.00	0.08	0.17	0.11
CDI Agent	**0.83**	**0.83**	**0.83**	**0.50**	**1.00**	**0.67**

Discussion des résultats : évaluation qualitative

Avant de discuter brièvement ces résultats, il est important de souligner que la discussion à leur sujet est strictement limitée à l'OE de l'exemple. De plus, cette discussion s'inscrit dans le cadre de la thèse qui envisage dans ses perspectives une étude extensive et approfondie sur les modèles émergents de type chatbot. Après avoir apporté cette clarification, une brève analyse des résultats obtenus pour cette OE est fournie.

Les modèles tfidf et YAKE! semblent se démarquer dans l'extraction de termes plus courts du document. Cependant, les termes qu'ils identifient ont tendance à être génériques et ne correspondent souvent pas à la perspective du recruteur.

Le classement du modèle Bard penche vers des termes courts et non spécifiques. Néanmoins, ils sont sémantiquement plus en phase avec le point de vue des recruteurs. Par exemple, des termes tels que "Équipe", "Protocole" et "Réglementation" sont possiblement associés à "Équipes Interfonctionnelles", "Protocole de Sécurité" et "Réglementations Financières" telles que notées par le recruteur.

Il est intéressant de noter que les prédictions du modèle ChatGPT ont montré des variations significatives en l'espace de huit jours. Cette variation pourrait s'expliquer par la nature intrinsèquement variable de ce modèle dans la génération de langage, introduisant ainsi un élément d'incertitude. Celle-ci pourrait entraîner un écart significatif par rapport à la perspective du recruteur, comme cela a été le cas dans l'exemple de l'OE présentée. Parmi les différentes prédictions faites par ce modèle, il convient de mentionner celle du 12/09/23, qui semble être en ligne avec celle de l'agent CDI, en couvrant tous les termes annotés par le recruteur.

En ce qui concerne le modèle Bing, il montre un comportement similaire aux modèles ChatGPT, se concentrant sur le contenu et la structure plutôt que sur les termes du titre de l'OE. Comme ChatGPT, il présente certaines redondances terminologiques, comme les termes "TAL" et "Accent sur le TAL". Il est également à noter qu'il introduit des termes non mis en évidence par d'autres modèles, des termes liés aux avantages salariaux et aux avantages compétitifs du poste. À cet égard, le classement de Bing semble mettre l'accent sur les termes qui

pourraient être plus pertinents pour les candidats.

En revanche, le modèle de l'agent CDI semble se caractériser par sa spécificité des termes, une caractéristique découlant de l'analyse terminologique préalable. La majorité de ses termes les mieux classés coïncident avec ceux mis en évidence par les recruteurs, suggérant son alignement avec les stratégies du recruteur. Une divergence notable est le rang élevé de "Python", mis en évidence dans le texte de l'OE mais négligé par le recruteur. Ceci pourrait être dû à divers facteurs tels que son évidence perçue ou des croyances dictées par le marché sur certaines compétences.

Enfin, bien que nous avons principalement visé à optimiser l'extraction de termes, nous n'avons pas approfondi l'optimisation des positions de termes ou évalué l'écart entre leurs valeurs de pertinence et les perceptions du recruteur. Malgré cela, l'échelle de pertinence de l'agent CDI semble plus uniformément répartie que celle de YAKE ! dans cet exemple. Les valeurs d'ambiguïté, mettant en évidence l'absence de certitude absolue sur la pertinence d'un terme et la relativité de la prise de décision, soulignent l'incertitude inhérente à ce domaine. Cette incertitude suggère que différents recruteurs pourraient interpréter la même OE différemment, surtout sans un contexte organisationnel partagé.

Discussion des résultats : évaluation quantitative

Nous avons également mené des évaluations quantitatives pour l'OE exemple. Pour ces évaluations, nous avons utilisé des métriques traditionnelles en recherche d'information (Information Retrieval, IR) : précision@N, rappel@N, et F1-mesure@N. Dans ce contexte, @N représente le nombre de termes qu'un recruteur a annotés comme pertinents dans l'OE. Par conséquent, ces métriques sont définies comme suit :

— **Précision@N :** la précision à N évalue la proportion de termes correctement prédits par le modèle parmi les N termes identifiés comme pertinents ; précision@N est normalisée par le nombre total des N premiers termes identifiés par le modèle. Elle est formellement définie par :

$$\text{Précision@}N = \frac{|\{\text{T}\} \cap \{\text{R}\}|}{N} \tag{III.2}$$

où $\{T\}$ est l'ensemble des termes identifiés par le modèle parmi les N premiers et $\{R\}$ est l'ensemble des termes annotés par les recruteurs comme étant pertinents ;

— **Rappel@N :** le rappel au rang N mesure la proportion de termes véritablement pertinents que le modèle parvient à identifier parmi les N premiers ; rappel@N est normalisé par le nombre total de termes pertinents dans l'OE, noté $|\{R\}|$ et annoté par les recruteurs. Il est formellement défini par :

$$\text{Rappel@}N = \frac{|\{\text{T}\} \cap \{\text{R}\}|}{|\{\text{R}\}|}\,; \tag{III.3}$$

— **F1-mesure@N :** la F1-mesure au rang N fournit une moyenne harmonique de la précision et du rappel au rang N. Il est formellement défini par :

$$\text{F1-mesure@}N = 2 \cdot \frac{\text{Précision@}N \cdot \text{Rappel@}N}{\text{Précision@}N + \text{Rappel@}N}. \tag{III.4}$$

Avec ces définitions à l'esprit, nous procédons à discuter la Tableau III.2.

Le modèle de l'Agent CDI a obtenu une F1-mesure@6 de 0.83 pour les 6 premiers termes. Cette F1-mesure suggère un équilibre entre la précision et le rappel pour l'exemple traité, indiquant que les prédictions du modèle coïncident avec une majorité des termes mis en évidence par le recruteur.

Pour un ensemble de termes plus étendu (top @12), la version de ChatGPT du 12/09/23 et le modèle de l'Agent CDI semblent alignés avec l'avis du recruteur, obtenant chacun une F1-mesure@12 de 0.67. Ces résultats laissent penser que, sur un ensemble plus étendu de termes issus de cette OE, ChatGPT pourrait se rapprocher de manière significative des attentes des recruteurs.

Une autre observation significative concerne les performances variables entre les différentes versions du modèle ChatGPT. Les termes prédits ainsi que leur alignement avec les perspectives des recruteurs varient. Ce comportement pourrait illustrer les incertitudes associées aux systèmes "boîte noire", en dépit de performances notables dans l'extraction de termes pertinents sur l'exemple donné.

En conclusion, les modèles "Agent CDI" et "ChatGPT 12/09/23" semblent être mieux adaptés au traitement automatisé de l'exemple en question, surtout du point de vue des recruteurs. Toutefois, ces deux modèles présentent des différences notables. D'une part, l'Agent CDI semble performant pour l'identification des 6 termes les plus pertinents, probablement grâce à une meilleure adaptation au contexte des recruteurs. D'autre part, bien que ChatGPT soit conçu pour des tâches d'extraction d'information, il s'avère pertinent dans cet exercice spécifique. Le premier repose sur des modèles explicables mais computationnellement plus complexes, tandis que le deuxième offre des performances élevées, au prix d'une explicabilité réduite et d'une incertitude inhérente non maîtrisée.

Naviguer entre ces deux approches représente un défi stimulant pour les chercheurs et les praticiens, invitant à une réflexion approfondie sur les compromis entre performance, explicabilité, adaptabilité et responsabilité sociale des modèles ML dans le contexte évolutif de la CCO.

III.2.2 Exemple de segmentation des CV pour l'extraction d'informations

Le processus de traitement automatique des CV, se concentrant spécifiquement sur leur segmentation, peut être élucidé à travers l'exemple suivant, expliquant son fonctionnement global.

Initialement, nous avons identifié à la fois des caractéristiques graphiques et textuelles qui

facilitent la description des séquences graphémiques les plus cruciales pour une segmentation automatique, s'alignant avec les segmentations manuelles effectuées par des recruteurs.

D'emblée, il convient de souligner que le CV existe sous un format non structuré, divisé en deux ou trois colonnes (dans notre cas d'application). Cette disposition présente des défis pour les méthodes d'extraction traditionnelles, rendant le CV relativement difficile à traiter.

Lors d'une inspection visuelle du CV de la Figure III.1, il devient évident que les sections titrées peuvent être distinguées par leur couleur, la taille de la police et la présence de mots-clés couramment associés aux CV.

Toutes ces caractéristiques peuvent être exploitées par une architecture basée sur BERT, facilitant une analyse grapholinguistique du document non structuré. Cette analyse vise à identifier des séquences graphémiques clés (Compétences, Expériences, Formation...) qui sont essentielles pour le problème d'analyse automatique des CV dans le contexte de la CCO. Dans nos expériences, l'accent est mis sur la segmentation du document.

L'identification automatique de telles séquences graphémiques, illustrée par les rectangles rouges, se convertit en coordonnées de segmentation optimales. Ces coordonnées s'alignent potentiellement avec le type de segmentation attendue par les recruteurs pour des tâches spécifiques d'analyse automatique de ce type de document.

Suite à ces exemples introductifs, nous procédons maintenant à la présentation des quatre principaux cas d'étude analysés au cours de la thèse, qui ont été réalisés au sein du département RH de DSI Group [6].

6. DSI Group est une entreprise de conseil française opérant dans le secteur des technologies de l'information et de la communication, https ://www.group-dsi.com/.

FIGURE III.1 – Example de CV. Les sections concernées par la segmentation se trouvent en rouge.

III.3 Premier cas d'étude : évaluation de la performance des méthodes floues linéaires vs non linéaires

Les modèles flous, avec leur capacité intrinsèque à gérer l'incertitude, sont idéalement positionnés pour capturer de manière appropriée les points de vue des recruteurs lors de l'extraction de termes à partir des annonces de poste. Dans nos expériences préliminaires nous avons comparé des modèles flous (RLF, ADF) et non flous (RLC). Ces deux types ont montré des capacités prédictives comparables, comme le montre la littérature existante. Cependant, les modèles flous présentent un avantage distinct en offrant une classification plus nuancée, flexible et compréhensible, surtout compte tenu de l'incertitude intrinsèque de cette tâche.

Pour approfondir le potentiel des modèles flous, nous avons recentré notre attention sur deux modèles représentatifs : la RLF, choisie pour sa modélisation linéaire intuitive, incarnant la famille des modèles flous linéaires ; et les ADF, qui, avec leur capacité à imiter des processus complexes de prise de décision humaine, se tiennent comme référence de la famille des modèles flous non linéaires.

Lorsqu'ils sont adaptés à des études de cas impliquant des jeux de données limités, ces modèles non seulement fournissent une base théorique solide, mais amplifient également l'explicabilité et le détail de la représentation, surtout en comparaison avec leurs homologues non flous. La polyvalence des modèles flous peut potentiellement améliorer le classement des CV dans les contextes organisationnels en favorisant une méthode plus raffinée et adaptative pour extraire des informations des OEs.

III.3.1 Configuration de l'expérimentation

Notre processus expérimental s'est initialement concentré sur la représentation du contexte organisationnel entourant les OEs, en s'appuyant sur des entretiens menés avec cinq recruteurs du département RH de DSI Group. Cette procédure nous a permis de construire l'ontologie déjà présentée dans la section II.5. Par la suite, nous avons sollicité le recruteur A, le responsable du département, pour décrire les exigences les plus saillantes de cinq OEs sous sa responsabilité. Dans le contexte du recrutement d'un candidat, les exigences pertinentes sont celles qui ne tolèrent aucune flexibilité.

En exploitant la représentation du contexte organisationnel et les stratégies de l'expert A pour sélectionner les informations cruciales dans chaque OE, nous avons dérivé des marqueurs textuels de pertinence de l'information. En général, les termes annotés étaient liés aux compétences professionnelles et, dans une moindre mesure, à la localisation, à la disponibilité, aux conditions de contrat, entre autres. Une fois ces marqueurs textuels dérivés selon les conclusions du recruteur A, nous avons invité quatre autres recruteurs (B, C, D et E) à déterminer si la stratégie déduite du comportement du recruteur A était généralisable ou non. Le processus

d'évaluation a été mis en œuvre de la manière suivante :

- les recruteurs B, C, D et E ont annoté les OEs qu'ils avaient gérés. Cela a résulté en un total de 25 documents annotés appartenant au corpus CP1 (section III.1) ; en moyenne, chaque OE contenait 100 termes, parmi lesquels entre 4 et 10 termes étaient annotés comme pertinents. Un premier ensemble de données de 2 501 termes a ainsi été généré ;
- pour former les modèles flous, un second ensemble de données a été généré à l'aide d'un algorithme de sous-échantillonnage aléatoire stratifié [229, 232] ; un ensemble de données de 500 termes a été obtenu, dont 35 % étaient pertinents et 65 % non pertinents ;
- les modèles non flous (RLC), flous linéaires (RLF) et flous non linéaires (ADF) ont été entraînés sur 70 % du second ensemble de données et testés sur les 30 % restants ; nous avons utilisé un échantillonnage stratifié pour garantir la proportion de termes pertinents et non pertinents dans chaque ensemble de données. De plus, nous avons examiné la fiabilité des modèles résultants en utilisant une validation croisée stratifiée à 10 blocs ;
- les deux modèles flous ont été comparés à des approches d'extraction de termes de l'état de l'art ; pour chaque OE annoté, nous avons évalué la pertinence de chaque modèle, basée sur les métriques de précision@K, rappel@K, et F1-mesure@K (où N représente le nombre de termes annotés par le recruteur). Ainsi, nous avons évalué les K termes prédits en haut de la liste par chaque méthode qui sont pertinents.
- les évaluations des modèles ont été effectuées avec les termes restants du premier ensemble de données, après avoir exclu les termes du second ensemble de données utilisé pour l'entraînement ; la procédure d'entraînement visait à obtenir le meilleur modèle en évitant le surapprentissage et en garantissant une variance maximale des échantillons d'entraînement ; enfin, la procédure d'évaluation pour mesurer les métriques de précision@K, rappel@K, F1-mesure@K visait à confronter les modèles entraînés à un contexte beaucoup plus réaliste avec une quantité importante de termes non pertinents ;
- concernant la mise en œuvre du marqueur M_8, nous avons utilisé un modèle SBERT (Sentence Bidirectional Encoder Representations from Transformers) pré-entraîné sur le corpus Wikipedia [233] pour extraire la signification sémantique des termes complexes ; ce modèle a été ensuite affiné sur la base des référentiels de compétences professionnelles, incluant CIGREF, e-CF [7], c2i [8] et ROME ; les concepts du titre sont principalement utilisés comme termes potentiellement pertinents pour l'identification de concepts associés, spécifiques et pertinents dans le corps de l'OE ;
- il convient de noter que dans cette expérience préliminaire, l'ADF complet a été utilisé, sans effectuer des opérations supplémentaires pour minimiser les règles de décision apprises [172].

7. `https://esco.ec.europa.eu/fr/node/193`
8. `https://www.education.gouv.fr/node/284519`

III.3.2 Exemple d'OE annotée

Ci-dessous (exemple 1), nous présentons un résumé d'un exemple d'OE annotée (avec les termes pertinents en gras) par le recruteur B. De plus, le tableau III.3 montre un exemple de termes extraits de cette OE à l'aide des modèles flous (processus de classification des termes entre pertinents ou non pertinents).

Exemple 1. *OE annotée avec les termes pertinents en gras :*

BI / BO *Analyste H/F*
Description de l'entreprise... (elle contient 121 mots)
Description du poste... (elle contient 89 mots)
Description du profil... (elle contient 69 mots)
Vous détenez un diplôme en ingénierie informatique. Vous possédez des compétences techniques telles que :
- Plateforme Business Objects - ***Maîtrise du langage SQL****, et l'utilisation de bases de données (****SAP IQ / IBM DB2****)*
Connaissance de l'ETL Stambia ou d'Oracle. Data Integration serait appréciée
Bonnes compétences relationnelles, dynamisme, esprit de synthèse, proactivité,
et esprit d'équipe sont des qualités qui vous caractérisent.
Expérience professionnelle : Minimum 2 ans. Lieu du poste : Metz-57. Géolocalisable : Oui.

Le tableau III.3 présente les 5 principaux termes prédits par la RLF et les modèles ADF sur l'exemple d'OE, ainsi que les scores de pertinence de chaque terme, avec des intervalles associés et des niveaux d'ambiguïté. Certains termes prédits (comme DSI et Activité Commerciale) font partie des sections de description de l'entreprise. Dans ce cas, tant au niveau syntaxique que sémantique, le modèle d'arbre de décision prédit des termes étroitement liés à ceux annotés par le recruteur.

Table III.3 – Top N=5 termes qui ont été prédits par la RLF et l'ADF.

	RLF			**ADF**		
#	**Terme**	**Score**	**Intervalle**	**Terme**	**Ambiguïté %**	**Score**
1	DSI	0.98	0.02	BI	9	0.97
2	Maîtrise du langage SQL	0.93	0.09	BO	9	0.97
3	Activité de l'entreprise	0.91	0.15	Maîtrise du langage SQL	16	0.87
4	BI	0.87	0.16	SAP IQ	28	0.71
5	SAP IQ	0.87	0.16	Compétence technique	25	0.69

III.3.3 Mise en œuvre expérimentale

Le tableau III.4 met en évidence les valeurs des coefficients attribuées à chaque marqueur textuel, selon les modèles obtenus. Nous avons entraîné une RLC comme modèle non flou, dans le but d'inclure un modèle complémentaire bien connu et explicable, avec une approche probabiliste. L'évaluation de l'ambiguïté des marqueurs textuels à l'aide de l'ADF met en lumière des aspects intéressants de la manière dont les termes pertinents sont identifiés.

Par exemple, une ambiguïté relativement faible est observée pour les marqueurs M_1, M_{12} et M_{16}, signifiant ainsi que : les recruteurs ont tendance à considérer comme pertinents les termes figurant dans les intitulés de postes (selon M_1) ; les termes situés entre le début et la moitié du document sont généralement jugés relativement pertinents (conformément à M_{12}), ce qui pourrait être influencé par la présence de la section de description de l'entreprise au début de certaines OE ; enfin, en raison des caractéristiques propres à YAKE !, qui sont en principe indépendantes du contexte, bien que M_{16} prédise parfois des termes non pertinents comme pertinents, son comportement prédictif global tend à identifier correctement la majorité des termes non pertinents, tout en reconnaissant également certains termes pertinents, une tendance qui contribue à son niveau d'ambiguïté observé.

TABLE III.4 – Évaluation individuelle orientée floue des 16 marqueurs textuels extraits en appliquant la RLC, la RLF, et l'ADF. Coef. : coefficients de RLC, SE : erreurs standard de RLC, Coef. A : centre du nombre flou triangulaire, Coef. S : étendue du nombre flou triangulaire.

	RLC			**RLF**		**ADF**
Marqueur Textuel	Coef.	SE	p-value	Coef. A	Coef. S	Ambig. % [40]
M_1	1.18	0.67	0.078	0.33	<0.001	12
M_2	4.02	0.52	< 0.001	3.40	<0.001	40
M_3	2.66	0.81	< 0.001	1.23	<0.001	26
M_4	1.66	0.52	0.002	1.00	<0.001	17
M_5	2.30	0.56	< 0.001	1.61	<0.001	18
M_6	1.48	0.65	0.023	0.03	<0.001	9
M_7	-0.41	0.63	0.512	0.63	<0.001	8
M_8	1.81	0.53	< 0.001	1.08	<0.001	13
M_9	-0.30	0.66	0.647	0.71	<0.001	8
M_{10}	1.02	0.68	0.132	0.26	<0.001	8
M_{11}	1.09	0.45	0.015	0.81	<0.001	39
M_{12}	-0.56	0.26	0.029	-0.85	<0.001	19
M_{13}	-0.27	0.63	-0.436	0.68	<0.001	31
M_{14}	0.12	0.10	0.246	-0.02	<0.001	20
M_{15}	3.87	2.73	0.160	1.71	<0.001	35
M_{16}	1.86	0.91	0.041	0.41	<0.001	5
Intercept	-4.51	0.86	< 0.001	-2.48	0.730	

Le Tableau III.5 met en évidence les résultats détaillés de nos expériences. Nous avons adopté deux approches spécifiques pour mener tous nos tests : la RLF et l'ADF. Pour l'entraînement de chaque modèle, nous avons utilisé des marqueurs textuels de l'état de l'art [E], ainsi que des marqueurs textuels orientés contexte que nous avons proposés [R] (M_1–M_{10}). Nous avons également tenté une fusion des deux procédures d'extraction de marqueurs textuels [R+E] (M_1–M_{16}).

L'observation des mesures indique une performance remarquable de l'ADF par rapport à la RLF et à l'algorithme YAKE ! De plus, nous avons examiné les algorithmes présentés par [4] et [5], qui se sont avérés moins efficaces que YAKE ! L'ADF a obtenu les meilleurs résultats dans l'extraction avec une performance maximale pour le rappel@N de 53% et pour le rappel@2N de 78%. Il convient de souligner que les marqueurs textuels de l'état de l'art ont été adaptés au contexte spécifique de la CCO pendant le processus d'entraînement.

TABLE III.5 – Résultats de précision, rappel, et F1-mesure de chaque méthode testée sur 25 OE. RLF ; ADF ; [E] : marqueurs textuels de l'état de l'art ; [R] : marqueurs textuels proposés basés sur le contexte ; [R+E] : combinaison des marqueurs textuels de l'état de l'art et des marqueurs textuels basés sur le contexte.

Métrique /Modèle	**YAKE !**	**RLF[E]**	**ADF[E]**	**RLF[R]**	**ADF[R]**	**RLF[R+E]**	**ADF[R+E]**
Précision@N, Rappel@N et F1-mesure@N [a]	0.10	0.16	0.19	0.24	0.38	0.41	**0.53**
Rappel@2N	0.25	0.33	0.40	0.42	0.57	0.62	**0.78**
Précision@2N	0.12	0.16	0.20	0.21	0.28	0.31	**0.39**
F1-mesure@2N	0.16	0.22	0.27	0.28	0.37	0.41	**0.52**

a. Rappel@N, Précision@N et F1-mesure@N sont équivalents à N.

III.3.4 Discussion

L'estimation de l'incertitude est essentielle pour améliorer l'identification des termes pertinents extraits automatiquement des OE. Cette première expérimentation suggère une analyse des métriques de possibilité et d'incertitude, afin d'évaluer la pertinence des marqueurs textuels identifiés.

La RLC présente une valeur R^2 de 0,64, ce qui indique un ajustement relativement solide [234]. Cette valeur a été utilisée comme un indicateur pratique, bien que non décisif (en raison de l'incertitude des données), révélant dans quelle mesure l'introduction de marqueurs orientés contexte a contribué à mieux décrire les points de vue des recruteurs sur ce qui est pertinent dans les OE, d'un point de vue statistique. De plus, notre hypothèse selon laquelle un modèle probabiliste des annotations des recruteurs n'était pas suffisamment adapté est susceptible d'être

confirmée par les valeurs-p de la RLC. Selon les coefficients de la RLF, les indicateurs orientés recruteur, M_2, M_3, M_4, M_5 et M_8 semblent être des marqueurs contextuels plus pertinents.

Concernant le coefficient de coordonnée à l'origine de la RLF en appliquant le principe d'extension [219], la *possibilité* de prédire un terme comme hautement pertinent est centrée sur 8% si toutes ses valeurs de marqueur textuel sont nulles, ce qui est une hypothèse plus pertinente en raison de l'incertitude des points de vue des recruteurs. Le coefficient de coordonnée à l'origine du modèle RLC donne une *probabilité* centrée sur 1% à la place, indiquant que même si toutes les variables de régression sont nulles, il existe un niveau d'incertitude qui n'est toujours pas décrit, associé aux points de vue des recruteurs sur la pertinence de l'information.

Les modèles flous appliqués semblent mieux adaptés pour gérer l'information considérablement incertaine [45] communiquée par les recruteurs. Étant donné les preuves de la plus grande adaptabilité et flexibilité des modèles flous, en passant à l'analyse des modèles flous non linéaires, l'ADF montre de meilleures performances. Cela suggère son alignement possible avec les stratégies et décisions prises par les recruteurs. Ceci est soutenu par le fait que la F1-mesure de l'ADF était meilleur en utilisant uniquement des marqueurs dépendants du contexte, des marqueurs indépendants du contexte et les deux types de marqueurs combinés. Spécifiquement, nous avons observé que plusieurs règles de décision produites après l'entraînement de l'ADF correspondent à des comportements précédemment observés chez les recruteurs. La règle suivante en est un exemple :

Si la possibilité est élevée (0.75-1.00) qu'un terme dans le titre représente une compétence professionnelle ou un métier (M_1), et si la possibilité est également élevée (0.75-1.00) qu'il représente une compétence professionnelle mentionnée dans les sections de description de poste ou de profil (M_2), alors il est plus possible que ce terme soit pertinent.

À partir de ce modèle flou non linéaire, nous avons également observé que certains marqueurs indépendants du domaine sont corrélés au contexte des OE. Par exemple, le marqueur TM_{11} est associé au comportement des recruteurs pour capitaliser les termes représentant les compétences professionnelles, qui sont généralement pertinents pour les OE. Malgré son importance, un tel marqueur pourrait également être ambigu (39%), ce qui est cohérent car la capitalisation n'implique pas nécessairement l'importance.

Enfin, il est crucial de souligner que l'approche floue offre généralement des prédictions plus flexibles et interprétables du point de vue de l'incertitude inhérente au contexte, ce qui est bénéfique compte tenu des perspectives des recruteurs, qui peuvent varier considérablement. Ceci est manifeste non seulement dans les intervalles de prédiction sur les termes des OE étudiés par les modèles flous, mais aussi dans l'interprétation de l'interception des modèles flous, comme cela a été illustré précédemment.

Les résultats acquis lors de cette première phase expérimentale suggèrent une adaptabilité de l'approche possibiliste au sein du contexte organisationnel d'application. Cette approche permet

l'ajustement des fonctions d'appartenance floue et des paramètres de ces fonctions, en vue de concevoir un modèle plus représentatif de l'incertitude et des caractéristiques spécifiques du contexte organisationnel dédié au traitement automatique des OE.

III.4 Deuxième cas d'étude : extraction de termes pertinents et annotation sémantique des OE à partir d'une approche possibiliste

La mise en évidence de l'adaptabilité des approches possibilistes à la problématique spécifique de l'extraction de l'information la plus pertinente des OE nous encourage à approfondir une modélisation plus sophistiquée de l'extraction de termes essentiels dans les OE. Pour ce faire, nous adoptons une approche posibiliste capable d'intégrer directement des sources de connaissance et des mesures d'incertitude.

C'est dans cette perspective que s'inscrit la deuxième expérimentation centrée sur la mise en œuvre de marqueurs textuels dérivés des annotations des recruteurs. Cette mise en œuvre est réalisée en utilisant l'architecture CDI possibiliste qui tire parti de l'ontologie organisationnelle et des ADF pour optimiser l'extraction des termes les plus pertinents.

III.4.1 Configuration de l'expérimentation

L'évaluation de l'architecture CDI posibiliste basée sur des ontologies a été effectuée au sein de DSI Group. Au total, quatre recruteurs, désignés par A, B, C et D dans cette étude, ont participé à l'expérimentation. Il est important de rappeler que les annotations du recruteur A, qui est le directeur du département RH, ont servi à élaborer les marqueurs contextuels M_1 à M_{10}.

Il est important de remarquer que ces recruteurs ont consacré plusieurs jours, voire des semaines, à diriger des processus de recrutement et à gérer les OE associées. Ils ont interagi quotidiennement avec des experts en gestion et en technique. En somme, ils ont développé une compréhension approfondie des besoins fondamentaux de leurs OE, acquérant ainsi des niveaux élevés de connaissances contextuelles.

Ainsi, dans le cadre de cette expérimentation, nous avons évalué la validité et la reproductibilité des stratégies du recruteur A sur 20 processus de recrutement gérés par les recruteurs B, C et D. Ces processus font partie de l'ensemble de cas étudiés lors de la première expérimentation. Le processus d'évaluation a été exécuté comme suit :

— nous avons demandé aux recruteurs B, C et D d'annoter les informations les plus essentielles exprimées dans toutes les OE associées à leurs processus de recrutement ;

— nous avons comparé leurs annotations aux termes les plus pertinents prédits par l'agent CDI dérivé des stratégies et des points de vue du recruteur A ;
— il est important de souligner que, contrairement à la première expérimentation, dans ce cas, l'intégralité de l'ADF n'est pas utilisée, car elle tend à être hautement complexe, frisant l'inexplicabilité ; par conséquent, les marqueurs moins ambigus et simplifiés identifiés par l'arbre de décision sont utilisés par l'agent, comme proposé dans la section II.9.8 ;
— nous avons comparé les performances de l'agent proposé à celles des méthodes d'extraction de termes de l'état de l'art ; nous avons utilisé les métriques précision@K, rappel@K, et F1-mesure@K pour évaluer l'adéquation de chaque modèle, en prédisant les N et $2N$ termes les plus pertinents de chaque OE annotée (où N représente le nombre de termes annotés par le recruteur sur chaque document).

Bien que théoriquement, notre approche devrait être comparée à d'autres méthodes spécifiquement conçues pour les OE, nous avons écarté cet objectif pour deux raisons. À notre connaissance, il n'existe actuellement aucune méthode automatisée et open-source pour extraire spécifiquement des informations pertinentes des OE. De plus, il n'existe pas de corpus public d'OE annotées pour l'extraction d'informations pertinentes. Pour ces raisons, nous avons comparé notre approche aux performances de méthodes d'extraction de termes de l'état de l'art indépendantes du domaine. Dans leurs articles respectifs, ces approches ont été évaluées sur des documents académiques annotés, des textes d'actualité et des articles scientifiques [6]. Par conséquent, c'est la première fois qu'ils sont évalués sur des OE.

Le tableau III.6 présente les résultats de nos expérimentations. Comme l'indiquent les métriques, les résultats de notre algorithme sont nettement meilleurs que ceux des quatre autres algorithmes. En particulier, par rapport aux meilleurs résultats de ces algorithmes (la plupart provenant de YAKE!), les améliorations de notre approche varient de 15% à 29%, atteignant 56% pour rappel@2N en termes de performance la plus élevée.

TABLE III.6 – Résultats expérimentaux. Niveaux de précision, de rappel, et de F1-mesure de chaque méthode sur 20 OE en utilisant RAKE [3], FRAKE [4], l'approche BERT topics [5], YAKE! [6], et notre Agent CDI.

Métrique/Méthode	RAKE	FRAKE	BERT	YAKE!	Notre Agent
Rappel@N, Précision@N F1-mesure@N	0.02	0.09	0.17	0.10	**0.38**
Rappel@2N	0.08	0.17	0.20	0.27	**0.56**
Précision@2N	0.04	0.08	0.10	0.13	**0.28**
F1-mesure@2N	0.05	0.11	0.14	0.18	**0.38**

Concernant la performance de l'agent CDI, trois aspects de son comportement ont été essentiels pour atteindre et maintenir un F1-mesure supérieur à @N et @2N. Tout d'abord, les marqueurs orientés contexte associés à de faibles niveaux d'ambiguïté influencent considérable-

ment le processus d'extraction de l'agent. Par exemple, les marqueurs M_1, M_4 et M_5 étaient généralement fortement indicatifs en eux-mêmes pour déterminer si un terme (selon les conditions de chaque marqueur) est pertinent ou non. Deuxièmement, les marqueurs orientés contexte présentant des niveaux d'ambiguïté moyens tendent à réduire leur incertitude associée lorsqu'ils sont utilisés conjointement avec d'autres marqueurs. Troisièmement, nous avons observé que l'inclusion de mesures d'incertitude pour déterminer les degrés de confiance permet à l'agent de contrôler l'ambiguïté relative associée à certains marqueurs tels que M_{12}.

Enfin, nos résultats reflètent la complexité d'extraire des termes pertinents de documents courts tels que les OEs. Les travaux sur YAKE! avaient déjà souligné ce problème sur des corpus composés de documents de courte longueur. Plusieurs corpus ont été impliqués dans leurs travaux, par exemple WWW, KDD et pak2018 [6]. Dans notre cas, YAKE! présente un comportement similaire (18%) par rapport à son maximum rapporté de 17,2% pour ces ensembles de documents de courte longueur.

D'autre part, une F1-mesure@2N maximale de 5% et 14% pour RAKE (n-gramme) et BERT (embeddings) respectivement pourrait être attribué à l'inadéquation des représentations de documents pour reproduire les concepts d'un contexte organisationnel. En effet, l'approche n-gramme a tendance à sous-représenter les termes, tandis que l'approche AP peut ne pas s'aligner avec la connaissance inhérente d'un contexte organisationnel donné, même après un processus d'affinage. Ce dernier défi peut devenir particulièrement prononcé lorsqu'une organisation n'a accès qu'à de petits ensembles de données. Il semble également que l'algorithme FRAKE soit limité par la courte longueur des OEs, étant donné qu'il est centré sur des marqueurs basés sur des graphes statistiques.

Selon ces éléments, avec une F1-mesure maximal de 38%, la performance de notre approche possibiliste illustre que les marqueurs orientés contexte peuvent compléter l'extraction d'informations pertinentes dans les OEs, intégrant des connaissances contextuelles organisationnelles spécifiques du point de vue d'un recruteur.

III.4.2 Discussion

Nous proposons une approche basée sur une ontologie multi-sources et un cadre possibiliste, utilisé dans une architecture d'agent CDI dynamique, pour analyser des règles simples qui définissent 16 marqueurs textuels. Étant donné le manque actuel d'approches spécifiques pour cette tâche, nous avons comparé notre méthode aux techniques les plus proches, c'est-à-dire aux approches d'extraction automatique de termes pour l'indexation de documents.

En rappel, parmi les 16 marqueurs textuels, 10 ont été définis sur la base des stratégies et points de vue d'un recruteur expert, tandis que les 6 restants proviennent de l'algorithme YAKE! Les résultats montrent que l'agent CDI défini est plus performant que les algorithmes d'extraction automatique de termes pour extraire des informations pertinentes des OEs. Les

performances obtenues suggèrent que la représentation du contexte organisationnel des OEs, en termes de stratégies et points de vue des recruteurs, est susceptible d'améliorer l'identification des informations pertinentes, au-delà des termes pertinents.

De plus, nous avons observé que notre mise en œuvre est susceptible de détecter des ensembles de termes pertinents plus cohérents que d'autres approches. Étant donné que l'agent extrait des termes suivant une analyse terminologique [218, 178], il détecte des termes complexes et spécifiques, qui sont souvent choisis comme pertinents par les recruteurs. En outre, nous avons constaté que certaines représentations textuelles, comme celle proposée par RAKE [3], sous-représentent les termes complexes, réduisant considérablement la prédictibilité.

Un examen des résultats révèle que les mesures de performance de tous les modèles, sauf pour le rappel@2N de notre algorithme, sont en dessous de la performance de référence attendue. Cela peut s'expliquer par le fait que les marqueurs statistiques sur lesquels reposent les travaux de [6, 3, 4] sont insuffisants pour déterminer de manière adéquate la pertinence des termes dans les OEs. Nos résultats montrent que les marqueurs de pertinence qualitatifs basés sur le contexte sont essentiels pour obtenir une meilleure F1-mesure, étant donné qu'ils sont indépendants de la taille de l'OE. Par exemple, il n'est pas rare de trouver des OEs où la taille de la description de l'entreprise est importante par rapport à la description du poste. En conséquence, les marqueurs statistiques ont tendance à donner un score plus élevé aux termes de la première section qui ne sont pas souvent nécessaires pour décrire les exigences essentielles du poste.

À la lumière des résultats et observations énoncés précédemment, il convient d'apporter une précision supplémentaire qui s'est avérée pertinente au cours de notre étude. Après avoir effectué une analyse combinatoire des marqueurs textuels, nous avons constaté que l'agent obtient une meilleure F1-mesure en exécutant les marqueurs textuels de la manière suivante. La première étape consiste à générer une population de croyances sur la pertinence des termes, en utilisant des marqueurs primaires ayant des niveaux d'ambiguïté plus faibles. Par exemple, le marqueur M_1 basé sur le contenu des titres ou le marqueur M_6 axé sur l'impact financier des compétences/activités professionnelles. Ensuite, des marqueurs secondaires avec de faibles niveaux d'ambiguïté (comme M_{10}) peuvent être appliqués afin de renforcer la population actuelle de croyances sur la pertinence des termes.

III.5 Troisième cas d'étude : évaluation des marqueurs textuels de l'OE

Dans cette section, nous présentons la principale expérimentation liée à la partie de la méthodologie concernant l'évaluation et l'optimisation des marqueurs textuels pour l'extraction d'informations des OEs vis-à-vis d'améliorer le classement des CV.

L'évaluation de la qualité des marqueurs textuels, tant nouveaux qu'existants, s'effectue selon

diverses dimensions. Cela englobe l'analyse de l'ambiguïté des marqueurs lors de la distinction entre termes pertinents et non pertinents, l'examen de l'alignement des marqueurs avec les perspectives des recruteurs tout en assurant une explicabilité minimale, ainsi que l'évaluation des relations de redondance entre les marqueurs et la quantification de l'information qu'ils véhiculent, entre autres critères.

Cette évaluation multidimensionnelle vise à optimiser la qualité des croyances reçues par l'agent CDI, permettant ainsi des interprétations mieux alignées sur les perspectives des recruteurs quant à la pertinence des termes présents dans les OE. L'expérimentation comprend les quatre étapes suivantes.

1. *Annotation des termes dans les OEs* : dix recruteurs ont annoté les exigences essentielles de 73 OEs issues du corpus CP1, lesquelles n'avaient pas été utilisées lors des deux premiers cas d'étude. 580 termes pertinents et 15 487 termes non pertinents ont été identifiés.
2. *Division de l'ensemble de données collecté en partitions ou blocs d'entraînement et de test* : la technique de validation croisée à 10 blocs a été employée, avec les métriques pour l'évaluation de la pertinence du marqueur (ambiguïté, entropie, etc.) étant évaluées pour chaque bloc, et leur moyenne rapportée. De plus, puisque la métrique d'ambiguïté peut être très sensible au déséquilibre des classes, un processus d'équilibrage des classes est effectué à l'intérieur de chaque bloc pour l'estimer plus rigoureusement, en garantissant que 35% des termes sont pertinents et 65% ne le sont pas [229].
3. *Sélection de marqueurs textuels* : nous avons examiné les performances de 30 marqueurs textuels (M_1–M_{30}, section II.10.2) pour modéliser les annotations des recruteurs.
4. *Implémentation de l'architecture CDI* : les marqueurs identifiés comme étant les plus pertinents après le processus d'évaluation ont été introduits dans l'architecture d'un agent CDI possibiliste ; leur pertinence a été davantage confirmée grâce à un processus d'extraction automatique en deux étapes des termes pertinents. Initialement, ce processus a été appliqué aux blocs du jeu de données de test (14 OEs par bloc) du corpus de l'étude actuelle. Par la suite, il a été appliqué au corpus du deuxième cas d'étude (section III.4), composé de 20 OEs.

III.5.1 Marqueurs textuels évalués

Comme mentionné précédemment, nous avons évalué 30 marqueurs textuels, qui comprenaient des marqueurs dérivés du contexte (M_1–M_{10}), des marqueurs YAKE ! (M_{11}–M_{16}), des marqueurs dérivés d'un ADF (M_{17}–M_{20}, voir section II.9.8), et un marqueur obtenu via la méthode ACP pour la réduction de dimension (M_{21}). De plus, nous avons testé neuf faux marqueurs à des fins de comparaison (M_{22}–M_{30}), composés de règles logiques qui, en principe, n'ont pas

de pertinence significative dans notre domaine.

III.5.2 Paramètres du moteur d'inférence Mamdani

Les détails de la mise en œuvre du moteur d'inférence de type Mamdani (annexe I) comprennent :

— *calcul des degrés de possibilité des marqueurs* : étant donné que l'objectif de cette expérience ne se concentre pas sur l'estimation des degrés de possibilité du marqueur, une approche simplifiée a été utilisée pour évaluer les métriques de qualité ; pour les marqueurs sans formule de degré de possibilité spécifique $\alpha_{t_{k,i}}$, leurs degrés de possibilité étaient supposés égaux à leurs valeurs de vérité ;

— *catégories floues* : un minimum de 2 catégories de qualité a été sélectionné pour les métriques de qualité les moins pertinentes dans le contexte organisationnel, et un maximum de 7 catégories de qualité pour les métriques plus critiques (par exemple, ambiguïté et précision négative) afin de les modéliser avec des règles plus détaillées et rigoureuses ;

— *fonctions d'appartenance* : nous avons utilisé des fonctions d'appartenance triangulaires standard pour toutes les catégories de qualité ;

— *critère d'évaluation* : nous avons évalué la qualité globale de chaque marqueur textuel en fonction de leurs métriques.

III.5.3 Résultats expérimentaux

Cette section détaille les résultats de l'évaluation à chaque étape de notre cadre d'évaluation de la qualité. Les figures présentées montrent l'applicabilité et l'utilité des métriques de qualité dans le contexte organisationnel spécifique. Elles révèlent la pertinence des marqueurs, indiquant visuellement comment la stratégie de modélisation pour chaque métrique et les règles floues associées peuvent être ajustées lorsqu'elles sont introduites dans le système d'inférence.

1. *Identification des corrélations :* la Figure. III.2 illustre les niveaux de corrélation de Pearson entre chaque marqueur et la variable R, qui désigne les annotations ou points de vue des recruteurs.
2. *Évaluation de la fréquence des marqueurs :* la Figure. III.3 présente le logarithme des niveaux de possibilité accumulés ou de la fréquence de chaque marqueur.
3. *Évaluation de la précision des marqueurs sur les termes pertinents et non pertinents :* la Figure. III.4 montre les résultats de l'évaluation de la précision de chaque marqueur sur les termes pertinents (axe des x) et non pertinents (axe des y).
4. *Estimation de l'ambiguïté :* les niveaux d'ambiguïté estimés de chacun des marqueurs évalués sont illustrés à la Figure. III.5.

FIGURE III.2 – Coefficients de corrélation de Pearson entre les marqueurs (M_x) et les annotations des recruteurs (R). Des couleurs plus foncées indiquent des niveaux de corrélation plus élevés.

5. *Comparaison de l'ambiguïté des marqueurs et de l'entropie d'information mutuelle :* la Figure. III.6 affiche un graphique comparatif composé de la dimension d'ambiguïté normalisée (axe des y) et de la dimension d'entropie d'information mutuelle normalisée (axe des x) de chaque marqueur.

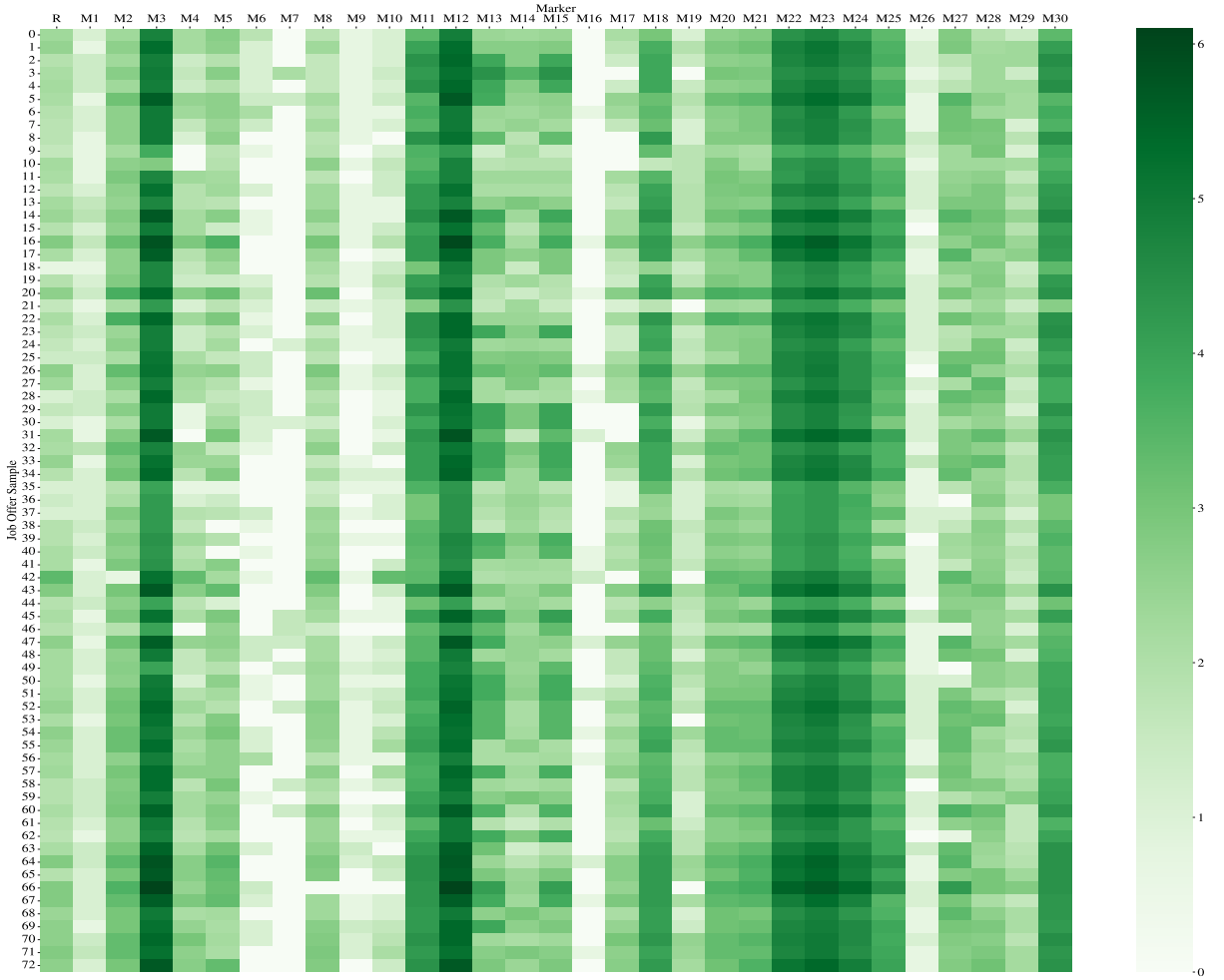


FIGURE III.3 – Logarithme des degrés de possibilité accumulés des marqueurs textuels (M_x) et des annotations des recruteurs (R). Les marqueurs sur-activés sont représentés par une couleur plus foncée et les marqueurs sous-activés par une couleur plus claire.

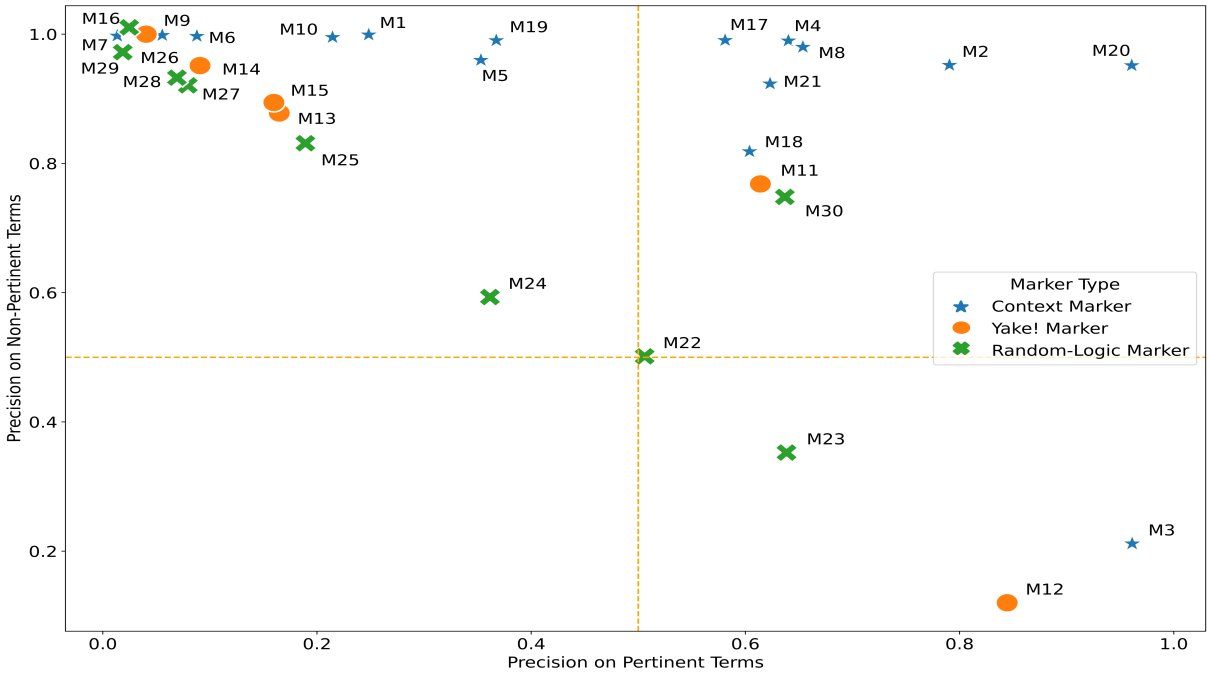


FIGURE III.4 – Précision des marqueurs sur les termes pertinents et non pertinents. Étoiles bleues pour les marqueurs dérivés du contexte, cercles orange pour les marqueurs YAKE!, et croix vertes pour les faux marqueurs.

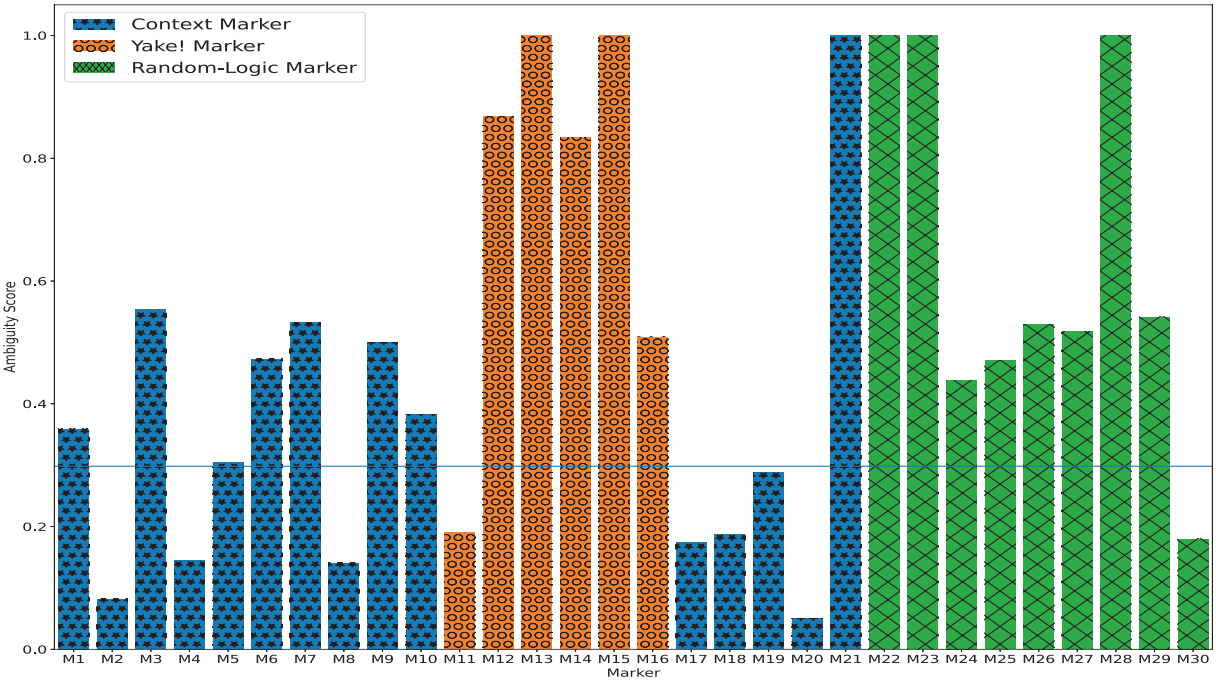


FIGURE III.5 – Degrés d'ambiguïté estimés des marqueurs textuels. Les marqueurs dérivés du contexte sont indiqués en bleu avec des étoiles, les marqueurs YAKE! en orange avec des cercles, et les faux marqueurs en vert avec des croix.

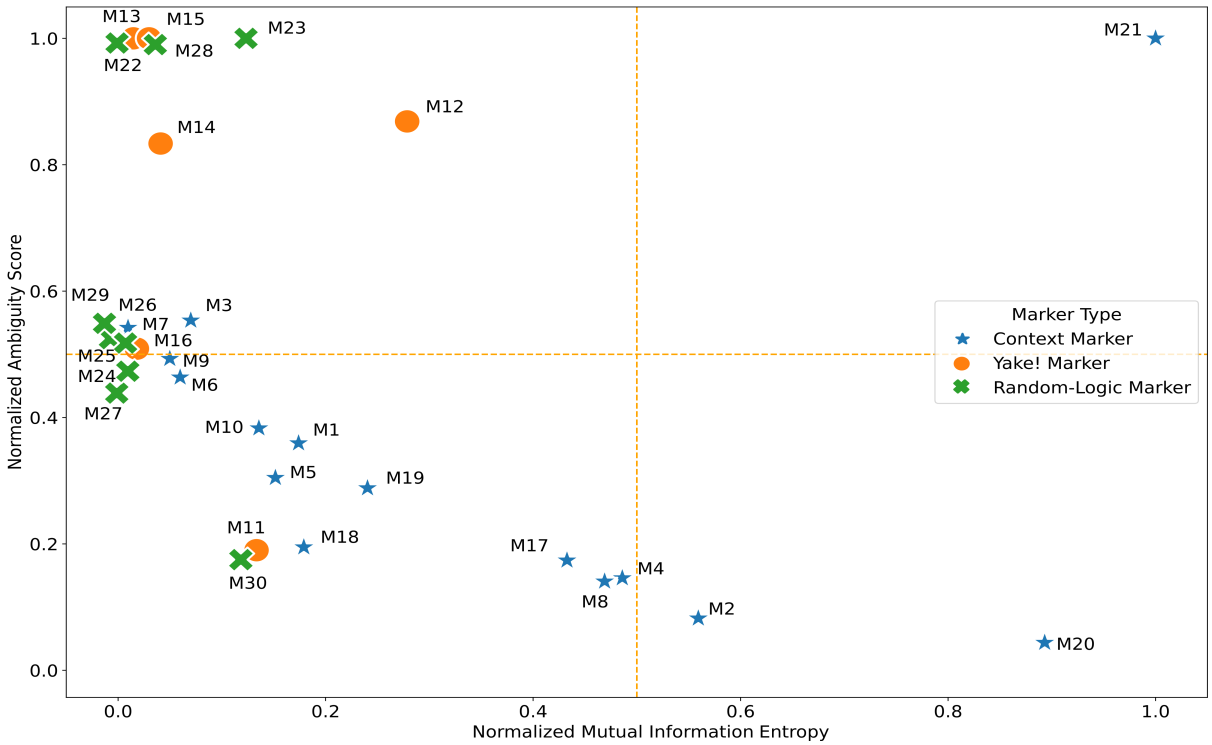


FIGURE III.6 – Ambiguïté des marqueurs textuels et entropie d'information mutuelle. Étoiles bleues pour les marqueurs dérivés du contexte, cercles oranges pour les marqueurs YAKE!, et croix vertes pour les faux marqueurs.

6. *Analyse prospective :* l'analyse prospective, réalisée en entraînant un modèle RLC, est présentée dans le Tableau III.7 (dans la colonne $P?$). Elle montre si les marqueurs textuels ont été détectés comme très significatifs à travers les blocs du processus de validation croisée à 10 blocs, avec un niveau de confiance de 95% (test de Wald).
7. *Alignement des marqueurs textuels par rapport aux stratégies des recruteurs :* le Tableau III.7 présente également (dans la colonne intitulée "Rec ?") les marqueurs à la fois dépendants et indépendants du contexte qui ont montré une association minimale avec les stratégies des recruteurs.
8. *Résultats complets de qualité :* le Tableau III.7 fournit également les résultats complets de qualité à chaque étape d'évaluation. La colonne EQ (évaluation de qualité) représente le résultat de défuzzification du moteur d'inférence, tandis que la colonne Pertinence signifie la catégorie floue correspondante à la pertinence de chaque marqueur. De plus, le Tableau III.8 présente un exemple d'une règle de qualité floue complexe dérivée du cas d'application.
9. *Performance de l'agent CDI. Évaluation 1 :* le Tableau III.9 présente les résultats de performance de l'agent CDI sur les blocs de l'ensemble de données de test dans la tâche

d'extraction de termes pertinents des OEs, en utilisant les marqueurs textuels les plus optimalement identifiés. Il est important de souligner que la performance des techniques évaluées de l'état de l'art sur cet ensemble de données n'était pas supérieure aux résultats que nous avions déjà rapportés dans le deuxième cas d'étude.

10. *Performance de l'agent CDI. Évaluation 2 :* enfin, le tableau III.10 fournit les résultats de performance sur le corpus d'OEs du deuxième cas d'étude, illustrant les résultats de performance avant et après la mise en œuvre de notre méthodologie d'évaluation des marqueurs textuels proposée. Pour obtenir ces résultats, nous avons utilisé des métriques classiques en recherche d'information : précision@N, rappel@N et F1-mesure@N.

Après cette présentation des résultats, nous procédons à la discussion de ces derniers dans la section suivante.

TABLE III.7 – Résultats de l'évaluation des marqueurs textuels sur le corpus étudié. Les métriques d'évaluation appliquées sont Pearson (coefficient de corrélation de Pearson), Prec+ (précision sur les termes pertinents), Prec- (précision sur les termes non pertinents), Amb. (ambiguïté normalisée du marqueur textuel), MI (information mutuelle normalisée du marqueur), P ? (indiquant si le marqueur a été identifié comme statistiquement significatif à travers les blocs de l'ensemble de données) et Rec ? (indiquant si le marqueur est minimement associé aux stratégies des recruteurs). La colonne EQ (Évaluation de Qualité) correspond au score du moteur d'inférence Mamdani obtenu en utilisant la méthode de défuzzification du centroïde. La colonne Pertinence indique la catégorie floue correspondant au score EQ. L'écart type de chaque métrique de marqueur évalué variait entre approximativement 0 et 0.03, à l'exception du marqueur 24, qui présentait un écart type de 0.21 pour la métrique d'ambiguïté.

Marqueur	Pearson	Prec+	Prec-	Amb.	MI	P ?	Rec ?	EQ	Pertinence
M_1	0.47	0.25	0.99	0.36	0.17	Yes	Yes	0.50	Moyenne
M_2	0.54	0.79	0.95	0.08	0.56	Yes	Yes	0.82	***Élevée***
M_3	0.08	0.96	0.21	0.55	0.07	No	Yes	0.50	Moyenne
M_4	0.66	0.64	0.99	0.15	0.49	Yes	Yes	0.84	***Élevée***
M_5	0.27	0.35	0.96	0.3	0.15	Yes	Yes	0.50	Moyenne
M_6	0.2	0.09	0.99	0.47	0.05	Yes	Yes	0.50	Moyenne
M_7	0.05	0.01	0.99	0.53	0.01	No	Yes	0.50	Moyenne
M_8	0.59	0.65	0.98	0.14	0.47	Yes	Yes	0.84	***Élevée***
M_9	0.15	0.05	0.99	0.5	0.03	Yes	Yes	0.50	Moyenne
M_{10}	0.36	0.21	0.99	0.38	0.14	Yes	Yes	0.50	Moyenne
M_{11}	0.17	0.61	0.77	0.19	0.13	Yes	Yes	0.68	Moyenne-Élevée
M_{12}	-0.04	0.84	0.12	0.87	0.28	No	Yes	0.25	Faible-Moyenne
M_{13}	0.06	0.16	0.89	1.00	0.01	No	No	0.17	Faible
M_{14}	0.08	0.09	0.95	0.83	0.04	No	No	0.24	Faible
M_{15}	0.07	0.16	0.89	1.00	0.03	No	No	0.17	Faible
M_{16}	0.2	0.04	0.99	0.51	0.02	No	No	0.21	Faible
M_{17}	0.63	0.58	0.99	0.17	0.43	Yes	Yes	0.81	***Élevée***
M_{18}	0.21	0.6	0.82	0.19	0.18	Yes	Yes	0.74	Moyenne-Élevée
M_{19}	0.46	0.37	0.99	0.29	0.24	Yes	Yes	0.59	Moyenne-Élevée
M_{20}	0.64	0.96	0.95	0.05	0.88	Yes	Yes	0.83	***Élevée***
M_{21}	0.31	0.62	0.92	1.00	1.00	Yes	Yes	0.25	Faible-Moyenne
M_{22}	0.00	0.51	0.5	1.00	0.01	No	No	0.16	Faible
M_{23}	0.00	0.64	0.35	1.00	0.12	No	No	0.18	Faible
M_{24}	-0.02	0.36	0.59	0.44	0.00	No	No	0.19	Faible
M_{25}	0.01	0.19	0.83	0.47	0.00	No	No	0.17	Faible
M_{26}	0.03	0.02	0.99	0.53	0.01	No	No	0.19	Faible
M_{27}	0.00	0.08	0.92	0.52	0.01	No	No	0.19	Faible
M_{28}	0.00	0.07	0.93	1.00	0.03	No	No	0.17	Faible
M_{29}	-0.01	0.02	0.97	0.54	0.00	No	No	0.17	Faible
M_{30}	0.17	0.64	0.75	0.18	0.13	Yes	Yes	0.70	Moyenne-Élevée

TABLE III.8 – Exemple d'une règle floue pour l'évaluation de qualité (EQ) du marqueur sur l'échelle par ordre croissant de qualité : très faible, faible, moyenne-faible, moyenne, élevé et très élevé.

IF	Association aux stratégies des recruteurs	**N'EST PAS**	FAIBLE
ET	Ambiguïté	**EST**	(TRÈS FAIBLE **OU** FAIBLE **OU** MOYENNE-FAIBLE)
ET	Récurrence de la valeur-p	**N'EST PAS**	FAIBLE
ET	Précision négative	**EST**	(ÉLEVÉE **OU** TRÈS ÉLEVÉE)
ET	Précision positive	**N'EST PAS**	FAIBLE
ET	Information mutuelle	**N'EST PAS**	FAIBLE
ET	Pearson correlation	**N'EST PAS**	FAIBLE
ALORS	EQ du marqueur	**EST**	MOYENNE-ÉLEVÉE

TABLE III.9 – Résultats expérimentaux sur les blocs de test du jeu de données OE. Moyenne et écart-type des niveaux de précision@N, rappel@N, et F1-mesure@N de l'agent CDI.

Métrique/Méthode	Valeur Moyenne	Écart Type
Rappel@N, Précision@N F1-mesure@N	0.56	0.028
Rappel@2N	0.67	0.038
Précision@2N	0.34	0.019
F1-mesure@2N	0.45	0.026

TABLE III.10 – Résultats expérimentaux sur le corpus du deuxième cas d'étude. Niveaux de précision, de rappel et de F1-mesure de chaque modèle sur le corpus précédent de 20 OE en appliquant RAKE [3], FRAKE [4], l'approche thématique BERT [5], YAKE! [6], ainsi que notre agent, tant dans le scénario du deuxième cas d'étude (Avant) que suite à l'application de la méthodologie d'évaluation des marqueurs textuels (Après).

Métrique/Méthode	RAKE	FRAKE	BERT	YAKE!	Avant	Après
Rappel@N, Précision@N F1-mesure@N	0.02	0.09	0.17	0.10	0.38	*0.55*
Rappel@2N	0.08	0.17	0.20	0.27	0.56	*0.64*
Précision@2N	0.04	0.08	0.10	0.13	0.28	*0.32*
F1-mesure@2N	0.05	0.11	0.14	0.18	0.38	*0.43*

III.5.4 Discussion

L'analyse des niveaux de possibilité accumulés (Figure. III.3) délimite deux groupes distincts de marqueurs : le groupe à haute fréquence d'activation (M_3, M_{12}, et M_{23}) et le groupe à faible fréquence d'activation (M_6, M_7, M_9, et M_{16}). Les marqueurs à haute fréquence pourraient être utiles pour identifier les sections de documents riches en termes pertinents, tandis que les marqueurs à faible fréquence pourraient détecter des termes pertinents dans des contextes spécifiques. Par exemple, le marqueur M_7 s'avère particulièrement efficace pour examiner les OE dans le secteur de management.

L'évaluation des corrélations (Figure. III.2) révèle des corrélations significatives entre les marqueurs YAKE! M_{13}, M_{14}, et M_{15}, en raison de leur dépendance à la fréquence des termes dans le texte. L'influence mutuelle du marqueur logique aléatoire M_{30} et du marqueur YAKE! M_{11} est notable, due à leur concentration sur la détection des termes en majuscules. De même, les marqueurs contextuels M_2, M_4, et M_8 manifestent des corrélations positives, attribuées à leur accent sur les compétences professionnelles. Comprendre ces corrélations peut offrir des perspectives sur l'interrelation des marqueurs et leur efficacité.

D'autre part, du point de vue de cette même métrique, les marqueurs composites M_{17}-M_{20}, dérivés de la simplification d'un ADF, correspondent fortement aux marqueurs contextuels qui ont formé leur base, soulignant l'accent des recruteurs sur les compétences professionnelles. À l'inverse, les faux marqueurs M_{22}-M_{29} ne correspondent guère aux perspectives des recruteurs. Cette différence souligne la valeur de l'analyse de corrélation pour comprendre la pertinence du marqueur.

En termes de précision (Figure. III.4), les marqueurs contextuels (M_1-M_{10}) et le marqueur composite M_{20} affichent une performance élevée pour identifier les termes non pertinents dans les OE, un attribut essentiel étant donné l'ensemble de termes hautement pertinents, généralement limités dans ces documents. Des marqueurs dérivés du contexte comme M_2, M_4, M_8, et les marqueurs composites M_{17}, M_{20}, et M_{21} démontrent une haute précision pour les termes pertinents et non pertinents. Cependant, M_3 et M_{12}, bien qu'efficaces pour reconnaître les termes pertinents, affichent une précision plus faible pour les termes non pertinents. Ce phénomène résulte probablement de leur classification large des termes dans des sections de document spécifiques comme potentiellement pertinents.

Le marqueur aléatoire M_{22}, qui maintient une précision de 50% pour chaque classe de termes, est notable. Cependant, les faux marqueurs, qui sont plus restrictifs, ont tendance à afficher une précision plus élevée pour les termes non pertinents et plus basse pour les pertinents, influencés par la forte prévalence de termes non pertinents dans le corpus. Cela peut introduire une incertitude dans le paramètre d'évaluation. Nous concluons l'analyse de la précision, tout en soulignant que la précision des marqueurs contextuels, du marqueur YAKE! M_{11}, et du marqueur logique aléatoire M_{30} dans la représentation des termes pertinents et non pertinents au sein des OE est

relativement significative.

L'analyse de l'ambiguïté (Figure III.5) montre une faible ambiguïté dans les marqueurs M_2, M_4, M_5, et M_8. Cependant, la plupart des marqueurs indépendants du contexte et de logique aléatoire démontrent une haute ambiguïté. Notamment, le marqueur M_{21} présente une ambiguïté considérable malgré une précision satisfaisante dans un processus de classification traditionnelle, compliquant potentiellement la classification des termes et augmentant l'ambiguïté dans les processus de raisonnement approximatifs comme celui effectué par l'agent CDI.

Lors de l'évaluation de l'entropie d'information mutuelle (Figure III.6), les marqueurs M_2, M_4, M_8, M_{17}, et M_{20}, et dans une moindre mesure M_1, M_{19}, M_{12}, M_{10}, et M_5, se caractérisent par une faible ambiguïté et une entropie d'information mutuelle relativement élevée. À l'inverse, le marqueur aléatoire M_{22} et les marqueurs indépendants du contexte comme M_{13} et M_{15}, présentent une haute ambiguïté et une faible entropie d'information. La plupart des faux marqueurs montrent également une entropie d'information mutuelle minimale et des niveaux d'ambiguïté élevés, confirmant leur inadéquation.

Dans l'analyse prospective des marqueurs (Tableau III.7), les marqueurs contextuels M_1, M_2, M_4, M_5, M_6, M_8, M_9, M_{10}, et les marqueurs YAKE ! M_{11} et M_{12} montrent des associations systématiques significatives avec les annotations des recruteurs sur les blocs de l'ensemble de données évalué. Certaines stratégies associées à ces marqueurs incluent :

- l'incorporation d'un nom représentatif métier dans le titre de l'OE ainsi que des compétences pertinentes (M_1) ;
- la sélection de termes représentant des compétences professionnelles situées dans des sections de document pertinentes (M_2) ;
- l'association d'expressions de pertinence avec des compétences importantes (M_4) ;
- fournir des détails sur le type d'interaction requis de la part du travailleur concernant des termes pertinents (M_5) ;
- utiliser des termes hautement pertinents (par exemple, dans le titre du document) pour désigner d'autres termes qui sont sémantiquement liés et d'une grande pertinence pour le poste de travail (M_8) ;
- accentuer l'importance d'un terme s'il correspond à un secteur économique pertinent de l'OE ou, plus particulièrement, à une compétence professionnelle essentielle (M_{20}).

Ce dernier marqueur indique, dans une plus grande mesure, que les compétences professionnelles sont généralement mises en évidence par divers moyens, y compris leur incorporation dans le titre du poste, leur placement dans les sections les plus pertinentes de l'OE (comme la description du poste), ou leur amplification à travers des expressions de pertinence explicites ou implicites. Un terme est probablement considéré comme potentiellement non pertinent s'il ne répond à aucun de ces critères.

Dans l'évaluation des relations des marqueurs avec le contexte organisationnel, l'analyse des

marqueurs tels que YAKE !'s M_{11} et M_{30} (logique aléatoire) met l'accent sur leurs limites, en particulier pour discerner les termes pertinents. Bien que les recruteurs écrivent souvent des noms de compétences professionnelles en majuscules ou avec des lettres capitales, ces marqueurs indépendants du contexte classent toujours à tort des termes non pertinents comme pertinents. Cette mauvaise classification inclut des termes tels que ceux qui suivent immédiatement un point, les noms d'entreprises, les lieux géographiques et les titres de listes. Ces erreurs signalent le besoin d'affiner ces marqueurs et représentent des opportunités d'amélioration.

Par ailleurs, le système d'inférence floue de type Mamdani met en évidence l'utilité des marqueurs M_2, M_4, M_8, M_{17}, et M_{20} dans l'extraction d'informations des OE. Ces marqueurs ont entraîné une augmentation de 5% de la F1-mesure par rapport au deuxième cas d'étude, comme illustré dans le Tableau III.10. Cette haute performance reste relativement constante à travers les blocs de l'ensemble de données de test. La plupart des marqueurs dérivés du contexte, tels que M_1, M_5, M_6, M_7, M_8, M_9, M_{10}, M_{18} et M_{19} ont été classés comme relativement pertinents en raison de leur signification statistique, explicabilité, haute précision sur les termes non pertinents, et niveaux d'ambiguïté moyenne-faible. Même si M_3 a été considérablement pénalisé pour sa précision diminuée, il peut jouer un rôle essentiel dans le filtrage initial de termes pertinents. À l'inverse, M_{21} présente de faibles performances, expliquées par sa séparation très floue entre les termes pertinents et non pertinents. Les marqueurs restants (M_{13}, M_{14}, M_{15}, M_{16}, et la plupart des faux marqueurs) ont été considérés comme faiblement pertinents.

Il convient de noter que les marqueurs identifiés comme hautement et modérément pertinents sont associés aux stratégies de sélection des recruteurs. Certains de ces marqueurs, notamment M_1, M_6, M_7, M_9, M_{10} et M_{19}, ont une haute précision pour les termes non pertinents. Malgré leur faible précision positive et leur ambiguïté relative dans le corpus étudié, leur signification statistique suggère une pertinence potentielle plus élevée dans d'autres corpus d'études d'OE. L'intégration de ces marqueurs dans le processus d'extraction d'informations de l'agent a amélioré sa F1-mesure@2N de 0,7-1% dans les blocs de l'ensemble de données de test et le corpus du deuxième cas d'étude. Ainsi, tout en améliorant l'explicabilité de l'agent sans impacter négativement sa performance, ces marqueurs offrent une gamme d'informations plus complète, basée sur une diversité accrue d'hypothèses sur la pertinence de l'information.

Pour conclure cette discussion, il est important de souligner que tout au long de l'expérimentation, la variabilité de chaque métrique de marqueur évaluée à travers les blocs de l'ensemble de données, mesurée par l'écart type, est restée comparativement faible. La valeur observée la plus basse était proche de zéro, tandis que la limite supérieure se situait autour de 0,03. Ceci est principalement attribuable au faible ratio de termes pertinents dans chaque OE. En conséquence, les marqueurs affichent généralement des niveaux de performance constants - soit bons, moyens, ou faibles - à travers toutes les métriques de qualité évaluées pour chaque bloc. Cependant, le marqueur M_{24} s'éloigne de cette tendance avec un écart type de 0,21 sur la métrique d'ambiguïté,

qui peut être attribué à des pics d'ambiguïté prononcés dans les OE où les recruteurs favorisent un plus grand nombre de termes simples (à un seul mot) comme pertinents.

III.6 Quatrième cas d'étude : segmentation grapholinguistique et annotation sémantique des CV

Dans cette dernière section, nous présentons notre expérimentation sur la construction de représentations de CV en utilisant une approche grapholinguistique. Ce processus est centré sur l'optimisation de la segmentation des documents.

III.6.1 Configuration de l'expérimentation

Nous avons réalisé une évaluation du cadre proposé en vue d'une représentation, segmentation, et analyse automatisée plus robuste des CV.

Des recruteurs et leurs assistants ont annoté 870 sections de 150 CV du corpus CP2 (section III.1), sélectionnés au hasard parmi certains processus de recrutement. Nous nommerons désormais cet ensemble de CV corpus CP2.1. Nous avons constaté que la fonction textuelle (FT) la plus pertinente pour décrire leurs segmentations manuelles était le "Titre de Section" (similitude de segmentation moyenne de $S = 63\%$). Nous avons donc dérivé cinq marqueurs de format et cinq marqueurs textuels associés à cette FT et les avons évalués en formant des modèles RLC.

Ensuite, nous avons adapté six modèles pour modéliser la FT : un classificateur RLC et cinq classificateurs de séquences BERT. Nous avons utilisé 70% le corpus CP2.1 pour l'entraînement et les 30% restants pour l'évaluation. Les résultats ont été validés en utilisant une validation croisée stratifiée à 10 blocs. Par la suite, avec les recruteurs, nous avons testé la validité des modèles adaptés en utilisant les 153 CV annotés restants du corpus CP2, contenant 923 sections. Dorénavant, nous désignerons cet ensemble restant de CV sous le nom de corpus CP2.2.

Notez que comme notre cadre vise à exploiter de petits ensembles de données, nous avons utilisé un paramétrage spécifique de BERT pour éviter le surapprentissage. Tout d'abord, le processus d'apprentissage est analysé à l'échelle des étapes plutôt que des époques, avec un arrêt anticipé. Deuxièmement, nous utilisons l'optimiseur ADAM avec une décroissance du poids de 0,01. Troisièmement, chaque modèle est formé à un maximum de 3 époques, avec un taux d'apprentissage de 2e-5, une décroissance linéaire, et une taille de lot de 16. Enfin, en raison des problèmes connus d'instabilité lors de l'affinage des modèles BERT [235], chaque modèle est exécuté 20 fois et la moyenne est rapportée.

III.6.2 Évaluation des marqueurs graphiques et textuels de la section de titre

Comme nous avons réalisé que les marqueurs graphiques n'étaient pas suffisants pour représenter les titres de section dans les CV moins stylisés, nous avons complété ces marqueurs avec des marqueurs textuels. Ensuite, la RLC est utilisée pour estimer la signification statistique des marqueurs par rapport à la FT "Titre de Section". Ce modèle a été initialement entraîné sous trois scénarios différents :

1. Pour évaluer préliminairement la signification statistique, 30 CV ont été initialement sélectionnés du corpus CP2.1 par échantillonnage aléatoire, et chacune des séquences graphémiques (SG) des documents a été représenté par les marqueurs associés à la FT. Cette représentation a été utilisée pour former une RLC optimisée par validation croisée stratifiée à 10 blocs (35% des SG sont des titres, tandis que les 65% restantes ne le sont pas). Un premier ensemble de marqueurs significatifs a été identifié.

2. Par la suite, une deuxième RLC a été entraînée sur l'intégralité du corpus CP2.1 en utilisant une procédure analogue.

3. Une troisième RLC a été appliquée au corpus CP2.1 en filtrant les instances de FT, réduisant considérablement le nombre de SG ne correspondant pas aux vrais titres.

Le Tableau III.11 montre les résultats des deuxième et troisième modèles entraînés.

TABLE III.11 – Évaluation de la signification statistique des marqueurs graphiques et textuels dans l'identification des titres de CV : RL SOF (Régression Logistique sans filtrage) évaluée sur des instances de titres, soit 17300 SGs avec 870 SGs correspondant à de vrais titres ; RL AF (Régression Logistique avec filtrage) évaluée sur l'approche d'extraction des instances de titres proposée dans l'étude actuelle afin de réduire les échantillons négatifs, spécifiquement 2485 SGs contenant 870 vrais titres). Les valeurs-p ont été obtenues à l'aide du test-z (test de Wald).

	RL SOF			**RL AF**		
	Coefficients	**Erreur Std**	**valeur-p**	**Coefficients**	**Erreur Std**	**valeur-p**
GM_1 (Taille de police)	4.61	0.68	<0.001	4.98	3.81	<0.001
GM_2 (Famille de police)	−0.39	0.78	0.620	−1.18	0.85	0.210
GM_3 (Couleur)	1.86	0.34	<0.001	2.30	0.94	<0.001
GM_4 (Gras)	0.72	0.31	0.019	0.60	0.41	0.120
GM_5 (Italique)	−0.10	0.86	0.910	−0.22	0.40	0.770
TM_1 (Capitalisé)	0.20	0.38	0.600	0.42	0.76	0.360
TM_2 (Tout en majuscules)	1.52	0.34	<0.001	1.58	0.46	<0.001
TM_3 (Variantes de titre)	2.98	0.70	<0.001	4.11	0.42	<0.001
TM_4 (Fréq. dans les titres de CV)	5.63	0.59	<0.001	6.99	1.08	<0.001
TM_5 (Fréq. dans les phrases courantes)	23.82	3.14	<0.001	24.80	0.88	<0.001
Intercept	−30.39	3.17	<0.001	−30.22	3.82	<0.001
R^2	0.72			0.76		

III.6.3 Évaluation de la segmentation

Nous avons évalué un total de 6 modèles pour la segmentation des CV en utilisant différents types d'enrichissements de séquences. Ces modèles sont présentés et rapportés car ils représentent diverses approches de modélisation, allant d'une analyse purement basée sur le texte du CV à une approche hybride qui incorpore à la fois des caractéristiques textuelles et graphiques pour effectuer le processus de segmentation.

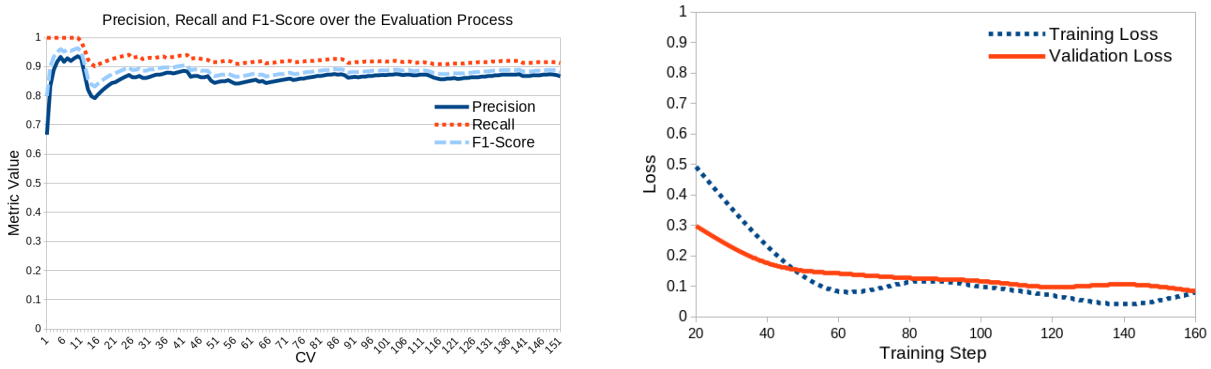
Le modèle de base est une RLC, qui est utile pour estimer le potentiel des marqueurs dérivés pour décrire la FT "Titre de Section". Nous avons également affiné cinq modèles de classificateur de séquences BERT. Le premier modèle (**BERT** WM, sans marqueurs) reçoit en entrée le texte des SGs. Le deuxième modèle (**BERT**+AM, avec tous les marqueurs) est affiné en enrichissant les SGs avec tous les marqueurs dérivés. Nous avons également formé un modèle basé sur les SGs formatés avec les marqueurs graphiques les plus significatifs (**BERT**+GM), et un autre en enrichissant uniquement les SGs avec les marqueurs textuels les plus significatifs (**BERT**+TM). Ensuite, nous avons affiné un modèle en enrichissant chaque SG avec les marqueurs graphiques et textuels les plus significatifs (**BERT** TM+GM).

Nous avons évalué les performances des modèles optimisés dans la tâche de segmentation des CV inconnus du corpus CP2.2. Nous avons utilisé les métriques rappel, précision et F1-mesure pour déterminer les performances des modèles dans la segmentation des sections de CV.

Le Tableau III.12 fournit les résultats de nos expériences. Nous montrons dans la Figure. III.7.a comment le meilleur modèle (**BERT** TM+GM) a évolué au cours de l'évaluation sur le corpus CP2.2. Dans la Figure III.7.b, nous illustrons la perte de validation et d'entraînement du modèle correspondant.

TABLE III.12 – Précision, rappel et résultats F1-mesure pour chaque modèle évalué sur les échantillons de test du corpus CP2.1 (45 CVs) et les échantillons du corpus CP2.2 (153 CVs).

Model	**LR**	**BERT** WM	**BERT**+AM	**BERT**+TM	**BERT**+GM	**BERT** TM+GM
Training Dataset						
Rappel	0.88	**0.97**	0.93	0.94	0.92	0.95
Précision	0.88	0.92	0.91	0.91	0.92	**0.93**
F1-mesure	0.88	**0.95**	0.92	0.93	0.92	0.94
Test Dataset						
Rappel	0.86	0.88	0.76	0.85	0.88	**0.91**
Précision	0.84	0.85	0.73	0.82	0.86	**0.87**
F1-mesure	0.85	0.86	0.74	0.84	0.87	**0.89**



(a) (b)

FIGURE III.7 – Performance du meilleur modèle **BERT** TM+GM pendant la phase d'évaluation. (a) évolution de la performance sur le corpus CP2.2. (b) perte d'entraînement (training loss) et de validation (validation loss).

III.6.4 Discussion

Les modèles de RLC utilisés pour estimer la signification des marqueurs ont montré que les marqueurs GM_1 (Taille de la police), GM_3 (Couleur de la police), GM_4 (Gras), TM_2 (En majuscules), TM_3 (Variante de titre de terme), TM_4 (Fréquence dans les titres de CV) et TM_5 (Fréquence dans les phrases courantes de CV) étaient les plus significatifs avec un niveau de confiance moyen de 95%. Cela illustre comment des marqueurs graphiques et textuels potentiellement significatifs peuvent être dérivés des segmentations des recruteurs avec une perspective grapholinguistique. Cette idée est soutenue par la valeur R^2 des deux régressions, qui reflète un ajustement significatif même en utilisant un ensemble de données moins équilibré (RL SOF). Ces résultats montrent également le potentiel de la fonction textuelle "Titre de section", qui est plus simple à modéliser et très pertinente pour identifier la structure des CV modernes.

En particulier, les marqueurs graphiques GM_1, GM_3, ainsi que GM_4 ont été identifiés comme les plus significatifs, reflétant la tendance actuelle des candidats à ajouter des titres avec de forts contrastes de couleur (GM_3), de tailles plus grandes (GM_1), et des styles spéciaux tels que le gras (GM_4) ou, dans une moindre mesure, l'italique (GM_5). Même si un titre de CV peut être écrit avec une police différente (GM_2), cela n'est pas fréquent, selon nos résultats.

Concernant les marqueurs textuels, le marqueur TM_2 révèle une tendance constante des candidats à écrire les titres de CV en lettres majuscules. En examinant le marqueur TM_3 centré sur l'identification des variantes terminologiques des titres, nous avons trouvé une indication statistique montrant qu'une partie des candidats fournit des termes ambigus, ce qui peut fortement dégrader la performance des méthodes d'analyse automatique des CV. Par conséquent, une analyse terminologique pourrait être utilisée pour travailler sur de telles sources d'incerti-

tude. De plus, nous avons observé que le marqueur TM_4 fournit une distinction plus claire entre les SG correspondant aux titres. En revanche, le marqueur TM_5 permet d'identifier ceux (plus nombreux) qui ne correspondent pas aux titres du point de vue du contenu textuel du CV.

En ce qui concerne la performance des modèles pour segmenter les CV, la RLC a montré un comportement relativement stable dans le corpus CP2.2 de CV. Cependant, elle n'a pas été aussi performante que les modèles basés sur BERT. Cela illustre la robustesse des marqueurs graphiques et textuels dérivés associés à la fonction textuelle "Titre de section".

Dans le cas des modèles basés sur BERT, nous soulignons d'abord qu'il est possible de les entraîner et affiner en minimisant des phénomènes tels que le surapprentissage ou le sous-apprentissage, qui se produisent lorsqu'il y a de petits ensembles de CV ou des paramètres d'entraînement non optimaux. Deuxièmement, nous avons constaté que le modèle **BERT** WM, qui utilise uniquement le texte des SG, peut s'adapter beaucoup mieux à l'ensemble de données d'entraînement, ce qui pourrait s'expliquer par la petite taille des titres de CV, représentant une complexité réduite. Cependant, la performance de ce modèle diminue lorsqu'il est évalué sur le corpus CP2.2 de CV inconnus. Nous avons observé que, en raison de la grande variabilité terminologique des titres d'un CV à l'autre, plusieurs nouveaux titres inconnus n'étaient pas correctement interprétés par un tel modèle.

En ce qui concerne les modèles dont les SG ont été formatés avec les marqueurs, nous avons observé que l'enrichissement textuel avec les marqueurs graphiques (**BERT**+GM) rendait le modèle basé sur BERT légèrement plus robuste pour les CV inconnus, lui permettant de produire des prédictions correctes sur des titres très rares mal gérés par le modèle uniquement textuel. Cependant, lorsque des CV moins stylisés sont traités, le modèle introduit de nouveaux types d'erreurs car l'algorithme se concentre sur des caractéristiques graphiques qui ne sont pas pertinentes pour ces types de CVs.

D'autre part, de meilleurs résultats ont été obtenus par rapport au modèle non formaté (**BERT** WM) et graphique (**BERT**+GM), grâce au modèle **BERT** TM+GM, qui combine les marqueurs graphiques et textuels les plus significatifs. Ce modèle capture de manière plus pertinente les motifs intrinsèques entre le texte des titres, les marqueurs graphiques et les marqueurs textuels, atteignant ainsi la meilleure performance. Enfin, le modèle **BERT**+AM a obtenu de bons résultats sur l'ensemble de données d'entraînement mais pas sur le corpus CP2.2, car trop de marqueurs sont intégrés dans le texte de chaque SG, rendant la tâche de classification beaucoup plus compliquée, et la taille de l'ensemble de données insuffisante pour effectuer un processus d'affinage général.

III.7 Conclusions

Les expériences du premier cas d'étude ont mis en avant l'application de modèles d'apprentissage automatique possibilistes linéaires et non linéaires tels que la RLF et les ADF pour évaluer l'incertitude dans les termes d'OE identifiés automatiquement. Ces modèles possibilistes ont démontré une adaptabilité et une flexibilité supérieures dans la gestion des incertitudes inhérentes des recruteurs, l'ADF correspondant particulièrement bien aux stratégies des recruteurs. La capacité de l'approche possibiliste à fournir des prédictions plus flexibles et interprétables souligne son potentiel en tant paradigme pertinent dans le traitement automatique des OE, reflétant avec précision les complexités nuancées des perspectives des recruteurs.

Dans le deuxième cas d'étude, un cadre possibiliste basé sur une ontologie a été proposé pour identifier les marqueurs textuels dans les OE, améliorant ainsi le classement des CVs. En synthétisant le contexte organisationnel, en extrayant l'ontologie de l'OE, et en employant un modèle d'agent possibiliste CDI dynamique, le cadre a permis d'obtenir des meilleurs résultats par rapport aux méthodes traditionnelles pour extraire des informations pertinentes des OE. Les résultats ont montré que les marqueurs de pertinence qualitative dérivés du contexte étaient essentiels et qu'une mise en œuvre stratégique des marqueurs textuels conduisait à de meilleurs résultats. Ces informations enrichissent considérablement notre compréhension du classement automatique des CV, soulignant la nécessité de la représentation du contexte et de l'incertitude dans le processus de recrutement.

Le troisième cas d'étude a évalué le cadre pour affiner les marqueurs textuels afin d'extraire des informations des OE de manière plus optimale, en accord avec l'expertise des recruteurs et des mesures d'ambiguïté et d'entropie de l'information. L'identification de marqueurs spécifiques à haute et basse fréquence pour cibler des termes pertinents et la précision des marqueurs contextuels dans leur classification étaient notables. L'analyse de corrélation, couplée à un système d'inférence flou de type Mamdani, a davantage mis en évidence les interrelations, la pertinence et l'utilité de ces marqueurs, les alignant sur les stratégies de recrutement et améliorant la précision de l'extraction de l'information. Malgré quelques limitations observées dans certains marqueurs, les résultats montrent une avancée substantielle vers une méthodologie axée sur le contexte, avec des fondements théoriques plus transparents pour la récupération d'informations dans les OE. Cette approche offre des performances relativement consistantes sur différents ensembles de données.

Le dernier cas d'étude, centrée sur les CV, a réussi à identifier les composants essentiels d'une méthodologie de représentation de CV sensible au format, en accord avec le contenu grapholinguistique des CV et les perspectives des recruteurs. Le cadre proposé a souligné l'importance de différents marqueurs graphiques et textuels dans les CV récents, illustrant le potentiel d'affinage des modèles basés sur BERT pour obtenir une segmentation améliorée. Ces résultats indiquent l'importance d'intégrer à la fois les styles graphiques et le contenu textuel dans l'analyse auto-

matisée des CV.

En résumé, les études de cas soulignent des composants fondamentaux pour une méthodologie de CCO plus robuste. Les modèles possibilistes permettent une gestion nuancée des incertitudes dans les OE. Le cadre possibiliste CDI basé sur une ontologie contribue à améliorer l'extraction d'informations de l'OE, améliorant ainsi le classement des CVs. De plus, l'utilisation du système d'inférence flou de type Mamdani souligne l'importance d'une approche globale et flou pour évaluer la qualité des marqueurs textuels, intégrant les perspectives des recruteurs et l'évaluation de l'incertitude. Enfin, l'intégration des contenus graphiques et textuels des CV indique une direction pour des modèles CCO toujours plus performants dans le processus de présélection, permettant une adaptation constante face à la révolution numérique.

Conclusion générale et perspectives

Dans cette section, nous présentons un résumé du problème, du travail réalisé, des discussions, des conclusions et des perspectives.

Énoncé du problème

L'interrelation entre les RH et l'IA, particulièrement dans l'appariement automatique des OE et des CV, incarne une mutation significative dans le domaine du recrutement. Néanmoins, un défi persistant réside dans la sous-représentation des OE et des CV durant la phase de présélection. Ce constat nous amène à la question centrale de cette recherche : comment élaborer une méthodologie permettant une représentation plus robuste des OE et des CV dans le contexte de la CCO ?

Un sous-problème initial réside dans la construction d'un cadre d'identification des marqueurs textuels dans les OE. Ce cadre vise à optimiser le classement automatisé des CV en tenant compte des perspectives des recruteurs et du contexte organisationnel. Une sous-question de recherche se pose alors : quelles sont les considérations essentielles pour élaborer un tel cadre et comment peuvent-elles contribuer à son efficacité ?

De plus, l'impact non contrôlé de l'incertitude inhérente au domaine de la CCO, et en particulier celle associée à l'extraction automatique de termes pertinents des OE, nécessite une représentation quantitative. L'accent de cette problématique est mis sur la pertinence des termes de l'OE dans le processus de présélection, et cela à travers l'application de modèles d'apprentissage possibilistes, tant linéaires que non linéaires. La question qui se pose est donc la suivante : comment peut-on quantifier efficacement cette incertitude tout en garantissant la pertinence des termes identifiés ?

Par ailleurs, le développement d'un cadre d'évaluation de la qualité des marqueurs textuels intégrant l'évaluation de l'incertitude est nécessaire afin d'améliorer la qualité de l'extraction de l'information dans ces documents. Ce cadre doit permettre une extraction plus précise des informations en considérant l'ambiguïté, l'expertise du recruteur, l'alignement de l'information par rapport au contexte organisationnel, et une approche explicable. Quels sont alors les composants essentiels de ce cadre et comment peuvent-ils être intégrés efficacement au processus d'extraction de l'information sur les OE ?

Enfin, la thèse s'attache à définir les composants essentiels d'une méthodologie sensible au format actuel des CV. Cette méthodologie doit s'aligner avec le contenu grapholinguistique des CV et les perspectives des recruteurs. Il est donc essentiel d'identifier ces composants et de comprendre comment concevoir une méthodologie qui répond aux exigences des formats contemporains de ce type de document.

En réponse à ces interrogations, la thèse présente une approche innovante de CCO. Cette méthodologie propose une représentation plus robuste tant des OE que des CV, facilitant ainsi leur analyse automatisée lors du processus de mise en correspondance. Les modèles et méthodologies utilisés dans cette recherche reposent sur des bases théoriques à la fois solides et plus transparentes. Cela permet de proposer des modèles de CCO qui, non seulement sont plus explicites et compréhensibles, mais sont également capables de naviguer à travers les défis multifactoriels du domaine. Ils sont également en mesure d'intégrer des structures "boîte noire", telles que les GML, en adoptant un paradigme orienté vers l'évaluation et la gestion des incertitudes inhérentes, facilitant ainsi la détection de biais, discriminations et autres conséquences issues de processus de CCO opaques.

Travaux réalisés

Au cours de cette thèse, plusieurs efforts de recherche ont été entrepris pour améliorer la CCO. Voici un résumé du travail réalisé :

Objectif 1 - évaluation quantitative de l'incertitude associée aux termes pertinents de l'OE : en mettant l'accent sur la mise en œuvre des modèles d'apprentissage possibilistes linéaires et non linéaires pour évaluer quantitativement l'incertitude associée aux termes pertinents de l'OE identifiés automatiquement, nous avons étudié l'utilisation d'estimations basées sur l'incertitude. Cette étude visait à évaluer l'efficacité de différents marqueurs textuels dans l'extraction de termes pertinents des OE, en utilisant deux approches différentes : la RLF et les ADF. Cette approche combinée nous a permis d'améliorer significativement les indicateurs de performance par rapport aux techniques existantes d'extraction de termes [39].

Objectif 2 - construction d'un cadre optimal pour l'identification des marqueurs textuels de l'OE : nous avons élaboré un cadre pour l'identification et la dérivation de marqueurs textuels dans les OE, intégrant des considérations essentielles telles que les perspectives des recruteurs et le contexte organisationnel. Possibiliste, axé sur le paradigme CDI et la représentation du contexte, ce cadre avait pour objectif non seulement d'identifier, mais également d'extraire des informations pertinentes des OE afin d'optimiser le classement automatisé des CV des candidats. En considérant une ontologie organisationnelle et en appliquant des mesures d'incertitude, ce cadre a permis d'améliorer les performances dans l'extraction de termes pertinents des OE [40].

Objectif 3 - développement d'un cadre d'évaluation de qualité pour optimiser les marqueurs textuels de l'OE :

nous avons introduit un cadre basé sur l'inférence floue pour optimiser l'utilisation et l'évaluation des marqueurs textuels dans les OE. Ce cadre, qui vise à maximiser l'efficacité des marqueurs textuels, prend en compte l'ambiguïté, l'expertise du recruteur, la pertinence de l'information par rapport au contexte organisationnel, et la quantité d'informations véhiculées par les marqueurs, tout en intégrant une approche explicative. En évaluant l'ambiguïté des marqueurs et leur pertinence continue dans un contexte organisationnel, il a démontré une amélioration des performances dans l'extraction de termes pertinents des OE [41].

Objectif 4 - élaboration d'une méthodologie pour la représentation du CV sensible au format :

nous avons élaboré et validé une méthodologie plus intégrale, sensible au format contemporain des CV, pour l'extraction, la transformation, et l'application du contenu grapholinguistique dans l'analyse automatisée du contenu. Cette méthodologie, qui est alignée avec le contenu grapholinguistique des documents et tient compte des perspectives des recruteurs, s'appuie sur un cadre développé qui est basé sur BERT pour la segmentation de ces documents, démontrant une performance améliorée lors des tests [42].

Discussion

Nos travaux de recherche ont entrepris une exploration approfondie avec pour principale motivation de proposer une représentation plus robuste des CV et des OE, spécifiquement dans le contexte de la CCO. En exploitant des approches traditionnelles et avancées, nous avons proposé une approche qui permet d'affiner la représentation de ces documents et les processus d'extraction d'informations concernées. Cette méthodologie contribue à traiter quelques-unes des limitations existantes dans le domaine de la CCO et vise à poser les bases pour des recherches futures sur des méthodes plus transparentes, adaptables et sensibles au contexte.

Les avantages de l'approche proposée incluent un contrôle accru sur la transparence et l'explicabilité du processus de la CCO du début à la fin. Ce contrôle est fondamental dans les modèles actuels de CCO, facilitant l'explication de la sélection ou du rejet d'un candidat, et aidant ainsi à mieux détecter et contrôler les biais et la discrimination. Un autre avantage est l'adaptabilité de l'approche à des contextes organisationnels spécifiques, permettant l'intégration de caractéristiques organisationnelles uniques à travers les ontologies et des processus d'extraction terminologique adaptés aux besoins organisationnels. Cette adaptabilité est encore enrichie par l'intégration de modèles transparents, tels qu'une architecture CDI possibiliste, qui incorpore des aspects du raisonnement humain nécessaires dans les contextes de CCO. De plus, l'approche proposée est adaptable à des langues telles que les langues romanes, dans lesquelles

des approches utilisées comme l'extraction automatique de termes sont hautement développées dans la littérature.

En lien avec l'explicabilité et l'adaptabilité mentionnées précédemment, un autre avantage de l'approche proposée concerne l'évaluation de la qualité, comme illustré par le processus d'évaluation des marqueurs. Notre approche permet l'identification et le suivi de la pertinence des divers traitements automatiques liés à la CCO, notamment ceux associés à l'extraction de l'information sur les OE. La pertinence de ces traitements est évaluée selon les critères de pertinence de l'information, associés aux avis des recruteurs, dans le contexte organisationnel spécifique en question, et ce, à l'aide de métriques de qualité dédiées. L'évaluation mise en œuvre englobe des analyses floues qualitatives et quantitatives. L'explicabilité de ces analyses est renforcée par l'intégration de catégories linguistiques floues aux métriques de qualité. Ces catégories, directement liées à la pertinence de l'information, sont appuyées par des outils de visualisation.

D'un autre côté, il est essentiel de reconnaître les défis rencontrés au cours de cette recherche. L'accent de cette thèse a été mis sur le développement et l'évaluation de méthodologies innovantes dans un contexte organisationnel, plutôt que sur des comparaisons quantitatives approfondies avec les modèles existants, en raison du manque de corpus disponibles dans le domaine.

Une limitation de l'approche proposée pour l'évaluation de la qualité réside dans sa dépendance envers un ensemble spécifique de données, ce qui pourrait potentiellement biaiser son interprétabilité lors de l'application à de nouveaux ensembles de documents. Bien que l'intégration de métriques statistiques, telles que celles associées à la régression logistique, atténue partiellement cette limitation en permettant l'évaluation du comportement des traitements automatiques sur des volumes de données plus importants, des audits réguliers de pertinence de l'évaluation de qualité s'avèrent indispensables. Ces contrôles sont essentiels pour apporter les ajustements nécessaires et assurer ainsi que le cadre d'évaluation de qualité demeure aligné sur les critères de pertinence de l'information évolutifs de l'organisation.

Par ailleurs, il convient de noter d'autres points d'attention dans notre contexte d'application. L'accès à l'expertise des recruteurs de l'organisation étudiée a été modéré en raison de leurs divers engagements professionnels, ce qui a limité l'approfondissement de l'analyse de leurs processus cognitifs associés à la pertinence de l'information. Il est aussi important de remarquer que la disponibilité des CV et des OE appariés a été limitée, en partie en raison de l'absence de corpus publics annotés et de l'acquisition progressive de ces documents au sein de la PME analysée.

Il convient également de souligner que, pour des raisons logistiques de cette organisation, diverses approches ont commencé à être mises en œuvre à l'échelle industrielle peu avant la fin de cette thèse. Par conséquent, la validation industrielle de la pertinence des CV sélectionnés et des candidats recrutés à partir des différents résultats de la méthodologie est un domaine de recherche et de publication envisageable.

Concernant les traitements automatiques proposés, plusieurs axes d'amélioration sont identifiés. Ceux-ci incluent l'optimisation de la performance de l'extraction terminologique sans perdre en explicabilité et en adaptabilité par rapport à l'approche utilisée dans cette thèse. Il y a également un besoin à étudier et à suivre l'évolution des variantes terminologiques dans chaque processus de recrutement, afin d'identifier automatiquement des variantes correspondant à de nouveaux concepts non couverts par les ontologies utilisées par le système. De plus, il est nécessaire d'approfondir l'intégration des ontologies avec les GML, permettant à ces derniers de mieux s'adapter aux connaissances spécifiques et organisationnelles sur les compétences professionnelles.

En matière de gestion de l'incertitude, une limitation de la thèse actuelle pourrait être que la modélisation proposée ne permette pas d'estimer et de contrôler pleinement l'incertitude associée aux GML. Le problème des incohérences générées par ces modèles est encore un domaine de recherche émergent ; par conséquent, l'estimation de leur incertitude peut être difficile à expliquer, comme observé dans un bref exemple d'OE dans ce manuscrit. L'importance d'évaluer l'incertitude associée à ces modèles contemporains est devenue évidente, tout comme il est essentiel d'expliquer et de comprendre cette incertitude.

De plus, il existe une limitation dans l'estimation des niveaux de possibilité associés aux marqueurs textuels dérivés des stratégies des recruteurs pour optimiser les tâches de la CCO. Étant donné l'absence de méthodologies dans ce domaine, et même au sein du domaine de la théorie de la possibilité, il est encore nécessaire d'approfondir sur la formalisation de la relation entre ces marqueurs textuels, les points de vue relatifs des recruteurs et l'estimation des niveaux de possibilité associés. Cela permettrait de positionner plus clairement chaque marqueur par rapport aux variables de la CCO qui peuvent impacter négativement leur performance.

Par ailleurs, l'approche possibiliste a été élaborée pour aborder l'incertitude associée à la CCO. Toutefois, elle pourrait ne pas refléter de manière exhaustive les subtilités et les perspectives variées des recruteurs. Plusieurs facteurs entrent en jeu : la dimension contextuelle des termes utilisés, mais aussi les interprétations individuelles des recruteurs, façonnées par l'environnement spécifique de leur organisation. Ces éléments pourraient entraver la flexibilité et l'applicabilité universelle de la méthodologie.

Additionnellement, une autre limitation est que les résultats obtenus lors de diverses expériences n'ont pas atteint un comportement parfait dans l'extraction des termes pertinents des OE. Deux facteurs fondamentaux contribuent à cet aspect : d'abord, les expériences ont montré que l'incertitude cognitive des recruteurs joue un rôle central. Cela est évident lorsque, par exemple, les recruteurs incluent diverses compétences professionnelles dans le titre d'une OE mais n'en sélectionnent qu'une comme la plus pertinente, impactant significativement la performance des marqueurs concernés par les compétences mentionnées dans le titre. Deuxièmement, le potentiel d'amélioration des marqueurs proposés indique un besoin d'étudier des corpus plus

étendus pour identifier éventuellement des marqueurs textuels plus robustes que ceux dérivés du petit ensemble de données étudié.

En ce qui concerne l'analyse des perspectives des recruteurs, la méthodologie se concentre fortement sur l'extraction de l'expertise de ces acteurs. Cela pourrait conduire à l'introduction de biais liés aux inclinaisons individuelles. Une surveillance et un raffinement continus des différentes méthodes d'analyse automatique sont essentiels pour minimiser ces types de biais et garantir l'alignement avec les besoins du contexte organisationnel spécifique.

En complément, l'accent mis sur la représentation du contexte organisationnel au sein de la CCO, tout en visant un ajustement sur mesure, introduit certaines limitations. Des potentielles erreurs d'interprétation ou de compréhension de ce contexte peuvent affecter la précision des modèles dérivés. Faire face à ces défis exige un processus analytique rigoureux et intensif pour saisir avec précision le contexte organisationnel, garantissant ainsi la minimisation des biais.

Bien que l'accent de la méthodologie soit mis sur la compréhension du contexte organisationnel et des perspectives des recruteurs, la variabilité entre les contextes d'entreprise peut engendrer des complexités. L'architecture dynamique et sensible au contexte de la méthodologie montre un potentiel d'adaptabilité, suggérant des avantages possibles pour les organisations. Cependant, pour une adoption industrielle généralisée, un raffinement et une simplification supplémentaires sont essentiels pour assurer une scalabilité et une reproductibilité optimales de la méthodologie.

Dans ce même sens, bien que la méthodologie soit particulièrement conçue pour les organisations de petite et moyenne envergure, sa dépendance à des jeux de données de taille réduite pose un défi. Elle peut être sensible aux restrictions des données disponibles, ce qui pourrait rendre sa stabilité sensible, surtout lorsque le volume des données augmente de manière significative. Cette caractéristique devient d'autant plus cruciale lorsqu'il s'agit de traiter des ensembles de documents plus vastes, comme des CV présentant un contenu graphique plus sophistiqué. Il est donc impératif d'adopter une surveillance continue afin d'ajuster l'approche aux besoins évolutifs du contexte de CCO.

D'un point de vue computationnel, la conception complexe de la méthodologie, intégrant de multiples composants et gérant d'énormes données pour la construction des diverses représentations plus explicatives, présente des défis en termes de temps de traitement et de problèmes potentiels de passage à l'échelle, en particulier par rapport aux modèles transformateurs contemporains. Cependant, il convient de noter que cette méthodologie cherche d'établir une base solide pour intégrer les dynamiques de recrutement. Ancrée dans un paradigme axé sur les croyances graduées et la représentation des systèmes complexes, elle offre un niveau de détail et d'explicabilité encore indispensable dans les modèles plus modernes.

En conclusion, bien que l'accent de la méthodologie soit mis sur la compréhension du contexte organisationnel et des perspectives des recruteurs, la variabilité entre les contextes d'entreprise

peut engendrer des complexités. L'architecture dynamique et sensible au contexte de la méthodologie montre un potentiel d'adaptabilité, laissant entrevoir des avantages possibles pour les organisations. Néanmoins, pour une adoption industrielle généralisée et immédiate, un raffinement et une simplification supplémentaires sont essentiels pour garantir une scalabilité et la reproductibilité plus optimale de la méthodologie.

Conclusions

Le domaine de la CCO, en particulier l'extraction, la structuration et l'annotation sémantique des CV et des OE, a constamment suscité l'attention académique et industrielle. Cette recherche, bien que distincte dans son approche, met en lumière les incertitudes inhérentes au processus et les avantages potentiels de l'intégration des perspectives des recruteurs et de leur contexte organisationnel lors du traitement de ces documents.

Plus précisément, cette thèse a exploré la possibilité d'augmenter l'estimation basée sur l'incertitude pour une représentation nuancée des OE et des CV. Au lieu de s'appuyer principalement sur la similitude sémantique basée sur AP, l'approche préconise l'utilisation de l'estimation de l'incertitude pour évaluer les marqueurs textuels, dans le but d'améliorer l'extraction d'informations pertinentes des OE et éventuellement d'affiner le classement des CV. Les résultats préliminaires indiquent de possibles améliorations des métriques F1-mesure et de rappel, suggérant des avancées potentielles dans la CCO.

Une contribution notable de cette thèse est l'introduction d'un cadre possibiliste axé sur le contexte. Contrairement à certaines méthodes précédentes qui ne tiennent pas compte de la représentation et de l'intégration du contexte organisationnel spécifique, le cadre proposé tente d'extraire des informations potentiellement pertinentes des OE en intégrant une représentation du contexte et des points de vue des recruteurs dans les documents. Cette approche axée sur le contexte pourrait avoir le potentiel d'améliorer les mesures de performance et éventuellement d'optimiser l'alignement de la CCO avec les besoins organisationnels.

La thèse a également exploré un cadre basé sur l'inférence possibiliste pour optimiser la sélection des marqueurs textuels dans les OE. Cette méthode estime de manière provisoire l'ambiguïté des marqueurs, évalue les informations qu'ils pourraient transmettre et évalue leur pertinence potentielle dans un contexte organisationnel. La sélection des marqueurs textuels pour les OE étant souvent un domaine nécessitant plus d'attention dans notre domaine, le cadre basé sur l'inférence floue proposé pourrait améliorer la précision de l'extraction de termes dans les OE, améliorant ainsi le classement des CV, et l'explicabilité des modèles sous-jacents.

Une autre exploration notable de cette thèse concerne l'extraction du contenu grapholinguistique des CV. À notre connaissance, les recherches antérieures n'ont pas exploité la riche mise en forme des CV. En réponse, cette thèse met en avant la valeur potentielle de cette mise en forme

et introduit un cadre provisoire sensible au format et basé sur BERT pour la segmentation des CV. Cette approche pourrait offrir une perspective plus complète sur la segmentation des CV, comme le suggèrent les résultats des F1-mesures de nos expérimentations.

De plus, cette thèse a tenté de mettre en lumière l'explicabilité de la CCO à travers les modèles explicables proposés. Alors que les approches d'AP contemporaines pourraient ne pas répondre pleinement à cela, les modèles de cette thèse s'efforcent de fournir non seulement une précision améliorée, mais aussi une explicabilité CCO plus claire de bout en bout. Ceci est réalisé grâce à l'intégration provisoire de marqueurs graphiques et textuels contextualisés et à une architecture de raisonnement approximatif proposée qui pourrait aider à mieux comprendre les différentes étapes de la CCO.

En conclusion, cette recherche doctorale a exploré plusieurs domaines de recherche au sein du champ de la CCO. Bien que les méthodologies et cadres introduits montrent des promesses, ils soulignent le besoin continu de validation et de raffinement, compte tenu de la nature dynamique du marché du travail et des évolutions technologiques. Le travail est un signe de collaboration interdisciplinaire, tissant des aperçus de la linguistique, des sciences des données et des ressources humaines. Cependant, les implications pratiques de ces explorations appellent à un engagement continu en matière d'adaptabilité et de raffinement dans d'autres scénarios réels.

Perspectives

Notre exploration pluridisciplinaire du domaine de la CCO a mis en lumière de multiples voies méritant une investigation approfondie. La robustesse empirique de nos indices et tendances souligne l'impératif d'intégrer une représentation de la couche de contexte organisationnel dans les paradigmes de l'apprentissage automatique de la CCO, en particulier lors de l'évaluation des marqueurs textuels. Cette intégration non seulement augmente la saillance de ces marqueurs, mais prépare également le terrain pour des recherches analytiques plus nuancées.

Il serait bénéfique que les futurs travaux de recherche collaborent étroitement avec les recruteurs. Cette collaboration permettrait de mieux comprendre leurs critères de sélection des candidats et leurs mécanismes décisionnels durant le processus de recrutement. Un tel effort enrichirait et approfondirait les stratégies d'évaluation de qualité utilisées dans les traitements automatiques de CCO, comme ceux proposés dans cette thèse, et viserait à améliorer la transparence et la compréhension intuitive de ces traitements.

Pour atteindre cet objectif, une méthodologie IHM (interface humain machine) formelle et bien structurée pourrait être élaborée et mise en œuvre pour les recruteurs. Cette initiative faciliterait l'analyse et l'interprétation des données issues des traitements automatiques et clarifierait les mécanismes de la CCO. Elle fournirait également les outils nécessaires pour une sélection de candidats plus transparente et alignée avec les besoins organisationnels. L'intégration de cette

méthodologie renforcerait également la cohérence et la rigueur des critères d'évaluation de qualité associés aux traitements automatiques de la CCO. À long terme, cela pourrait conduire à l'établissement d'un cadre normatif et de meilleures pratiques standardisées, facilitant ainsi des processus de recrutement automatisés, transparents et équitables dans divers secteurs et organisations.

Tout en consolidant la formalisation de pratiques de recrutement, il est nécessaire d'approfondir d'autres aspects comme la formalisation des informations contextuelles du CV liées aux compétences professionnelles. Dans cette optique, une extension essentielle de notre recherche serait d'élargir l'analyse et la formalisation des informations contextuelles du CV utilisées dans l'approche de CCO proposée, en les conceptualisant comme des marqueurs textuels de ce document. Cette démarche permettrait une extraction d'informations du CV plus sensible au contexte. Comme nous l'avons fait pour les OE, cette approche tiendrait compte non seulement du contexte organisationnel des recruteurs – qui devraient participer de manière plus active à la formalisation – mais également des perspectives uniques des candidats et des contextes sociétaux spécifiques dans lesquels ils évoluent. Reconnaître la nature multifacette des expériences et des antécédents des candidats, et comprendre comment ceux-ci se croisent avec les normes et attentes sociétales, peut considérablement renforcer la transparence, l'équité, et la pertinence des processus d'extraction d'information du CV employés dans la CCO.

Il sera également indispensable d'approfondir l'étude et formalisation des interactions entre les marqueurs textuels de l'OE et du CV, comprenant ces deux types de marqueurs comme des représentations des informations les plus essentielles communiquées dans chacun des documents. Ces marqueurs, en tant que reflets des éléments constitutifs de ces documents, jouent un rôle central dans une communication efficace et pertinente des qualifications, compétences et exigences. Un approfondissement dans l'étude et la représentation de la diversité des concepts traités dans les deux documents sera nécessaire, ainsi que dans l'examen de leurs interrelations dans le cadre de la CCO. Cette démarche permettra non seulement une meilleure compréhension des dynamiques textuelles et sémantiques en jeu mais aussi favorisera le développement de méthodologies plus précises et adaptées pour l'analyse et l'interprétation automatisées dans le domaine du recrutement.

Par ailleurs, approfondir notre compréhension des processus cognitifs des recruteurs s'avère nécessaire, en adoptant une approche pluridisciplinaire, en particulier du point de vue de la psychologie cognitive. Cette exploration permettrait de mieux comprendre les processus cognitifs des recruteurs, notamment ceux associés à la sélection de l'information, les mécanismes intrinsèques à la CCO ainsi que le phénomène d'incertitude qui lui est inhérent.

En s'aventurant plus profondément dans des corpus étendus, nous sommes incités à étudier l'ubiquité des marqueurs textuels des documents à travers des milieux organisationnels hétérogènes. De telles explorations conduisent à des questions sur l'adaptabilité des mesures d'incerti-

tude aux nuances organisationnelles et les ramifications potentielles des métamorphoses organisationnelles sur l'évaluation de ces marqueurs. Dans le même esprit, il est pertinent d'explorer l'adaptabilité à d'autres secteurs d'activité économique, différents du secteur des Technologies de l'Information et de la Communication, tels que le secteur médical ou la restauration. Dans cette optique, l'évaluation de la qualité des marqueurs textuels émerge également comme une ligne d'approfondissement, avec un accent particulier sur l'évaluation de la sensibilité et de la fiabilité du système d'inférence floue proposé vis-à-vis de la diversité des contextes organisationnels.

De plus, il est impératif d'examiner minutieusement les valeurs de pertinence des termes produits par les modèles. Cette analyse vise à déterminer dans quelle mesure la différence de niveaux de pertinence entre deux termes, mesurée sur une échelle de référence, concorde avec la perception et les avis des recruteurs. Une exploration complémentaire pourrait consister à comparer les annotations des experts du domaine de recrutement avec celles des non-experts. Cette démarche permettrait de déterminer si les connaissances capturées par les marqueurs textuels dérivés lors de cette thèse correspondent à des connaissances générales, exploitables de manière transversale dans le domaine de la CCO, ou à des expertises très spécialisées.

En outre, fusionner l'approche basée sur l'ontologie CDI possibiliste avec les techniques d'AP contemporaines, comme les GML, présente une avenue prometteuse pour transformer le domaine de la CCO. La méthode basée sur l'ontologie, ancrée dans le paradigme possibiliste, offre une représentation structurée de la connaissance. Elle vise à capturer habilement les incertitudes et ambiguïtés inhérentes qui sont omniprésentes dans le domaine de la CCO. D'un autre côté, les GML, avec leurs vastes dépôts de connaissances et leur prouesse avancée en AP, peuvent fournir une compréhension sémantique plus profonde et une interprétation contextuelle du contenu textuel.

Dans le même sens, une investigation plus approfondie du phénomène des incohérences générées par les GML devient nécessaire. Cette recherche supplémentaire serait essentielle pour développer une modélisation de la CCO plus robuste, fiable et explicable, capable de gérer plus efficacement les erreurs et les biais potentiels introduits par la nature opaque de ces modèles.

Alors que les GML restent en quelque sorte une "boîte noire" dont l'explicabilité complète défie même les principales entreprises du domaine, les explications de surface qu'ils offrent en langage naturel pourraient potentiellement être modélisées sous le paradigme de croyances graduées proposé dans cette thèse. Cela permettrait d'encapsuler ces modèles dans une "boîte grise" de niveau supérieur, représentant leurs croyances, désirs et intentions manifestes sur les réponses qu'ils fournissent. Cette approche pourrait permettre d'exploiter plus en profondeur le potentiel de la théorie de la possibilité pour estimer, expliquer et suivre plus rigoureusement l'incertitude inhérente à ces modèles.

Par ailleurs, dans le contexte d'un paysage d'emploi de plus en plus mondialisé, l'exigence de déchiffrer les subtilités des CV et des OE à travers les clivages linguistiques et culturels devient

primordiale. Cela nécessite le développement de modèles compétents pour traduire et aligner fluidement les profils à travers de nombreux spectres culturels et linguistiques.

Les synergies interdisciplinaires, fusionnant l'acuité des scientifiques des données, des linguistes, des spécialistes des RH, des éthiciens et des psychologues organisationnels, promettent de produire des solutions plus holistiques et impactantes. La scalabilité des modèles, garantissant une mise en correspondance de profil en temps réel sans compromettre la fidélité, couplée à des considérations éthiques mettant en avant les droits des candidats, reste centrale comme une nécessité. En ce qui concerne la scalabilité, une voie d'exploration prioritaire est l'optimisation de la révision des croyances. En effet, d'un point de vue théorique, la révision des croyances d'un agent CDI peut devenir extrêmement difficile, constituant un problème NP-complet. Toutefois, des expérimentations existantes suggèrent qu'avec des structures de données optimales, ce processus peut atteindre une performance similaire à celle d'un algorithme de complexité linéaire [236].

En conclusion, alors que nous naviguons dans la complexité de la CCO, un principe directeur reste immuable : la technologie, dans sa capacité avancée, devrait compléter de manière synergique la nature en constante évolution du comportement et de l'expertise humaine, qui est intrinsèquement influencée par des facteurs psychologiques (perceptions, attention, mémoire, contexte organisationnel, etc.) et les incertitudes qu'ils apportent. Cela a le potentiel de façonner un paradigme de recrutement qui est plus responsable, équitable, transparent, explicable et robuste.

Liste de publications

[1] A. Espinal, Y. Haralambous, D. Bedart, and J. Puentes, "Fuzzy-oriented terminological analysis to extract job offer information relevant to candidate ranking," *Information Technology in Industry*, vol. 10, no. 1, pp. 9–20, 2022. [Online]. Available : https://it-in-industry.com/issue/archive/papers/122.html

[2] ——, "An ontology-based possibilistic framework for extracting relevant terms from job advertisements," in *Proceedings of the 14th International Joint Conference on Computational Intelligence (IJCCI 2022) - FCTA. Best Paper Award 2022 Edition*, INSTICC. SciTePress, 2022, pp. 163–174.

[3] ——, "Uncertainty-oriented textual marker selection for extracting relevant terms from job offers," in *8th International Conference on Artificial Intelligence and Fuzzy Logic System (AIFZ 2022)*, CS & IT - CSCP. CSITY, 2022, pp. 1–16.

[4] ——, "A format-sensitive bert-based approach to resume segmentation," in *33rd Conference of Open Innovations Association (FRUCT)*. IEEE, 2023, pp. 30–37.

[5] ——, "A quality assessment framework for information extraction in job advertisements," *SN Computer Science*, vol. 4, no. 6, pp. 787–807, Oct 2023. [Online]. Available : https://doi.org/10.1007/s42979-023-02247-5

BIBLIOGRAPHIE

[1] A. Espinal, Y. Haralambous, D. Bedart, and J. Puentes, "An ontology-based possibilistic framework for extracting relevant terms from job advertisements," in *Proceedings of the 14th International Joint Conference on Computational Intelligence – FCTA*, INSTICC. Setúbal : SciTePress, 2022, pp. 163–174.

[2] C. Da Costa Pereira and A. G. B. Tettamanzi, "An integrated possibilistic framework for goal generation in cognitive agents," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems : Volume 1*, 2010, pp. 1239—1246.

[3] S. Rose, D. Engel, N. Cramer, and W. Cowley, *Automatic Keyword Extraction from Individual Documents*. New York : John Wiley & Sons, 2010, ch. 1, pp. 1–20.

[4] A. Zehtab-Salmasi, M.-R. Feizi-Derakhshi, and M.-A. Balafar, "FRAKE : Fusional Real-time Automatic Keyword Extraction," 2021, arXiv 2104.04830.

[5] R. Dagli, A. M. Shaikh, H. Mahdi, and S. Nanivadekar, "Job Descriptions Keyword Extraction using Attention based Deep Learning Models with BERT," in *3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, jun 2021, pp. 1–6.

[6] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, "Yake! collection-independent automatic keyword extractor," in *Advances in Information Retrieval*, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Cham : Springer International Publishing, 2018, pp. 806–810.

[7] L. Tokuda, "Computers assist humans in human resources," in *Proceedings of the The Second Conference on Innovative Applications of Artificial Intelligence*, ser. IAAI '90. AAAI Press, 1990, p. 179–188.

[8] S. Rojas-Galeano, J. Posada, and E. Ordoñez, "A bibliometric perspective on ai research for job-résumé matching," *The Scientific World Journal*, vol. 2022, pp. 1–15, 2022.

[9] R. Krum, *Cool infographics : Effective communication with data visualization and design*. John Wiley & Sons, 2013, pp. 174–174.

[10] A. Mahjoub and P. M. Kruyen, “Efficient recruitment with effective job advertisement : an exploratory literature review and research agenda,” *International Journal of Organization Theory & Behavior*, vol. 24, no. 2, pp. 107–125, 2021.

[11] E. . Cedefop, *Skills forecast : trends and challenges to 2030.* Luxembourg : Publications Office. Cedefop reference series, 2018, no. 108, pp. 1–140.

[12] G. Brunello and P. Wruuck, “Skill shortages and skill mismatch : A review of the literature,” *Journal of Economic Surveys*, vol. 35, 04 2021.

[13] Y. Wang, B. Daille, and N. Hathout, “Exploring terminological relations between multi-word terms in distributional semantic models,” *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 2023. [Online]. Available : https://www.jbe-platform.com/content/journals/10.1075/term.21053.wan

[14] A. Köchling and M. C. Wehner, “Discriminated by an algorithm : a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development,” *Business Research*, vol. 13, no. 3, pp. 795–848, November 2020. [Online]. Available : https://ideas.repec.org/a/spr/busres/v13y2020i3d10.1007_s40685-020-00134-w.html

[15] D. Hangartner, D. Kopp, and M. Siegenthaler, “Monitoring hiring discrimination through online recruitment platforms,” *Nature*, vol. 589, no. 7843, pp. 572–576, 2021.

[16] P. Kline, E. K. Rose, and C. R. Walters, “Systemic Discrimination Among Large U.S. Employers*,” *The Quarterly Journal of Economics*, vol. 137, no. 4, pp. 1963–2036, 06 2022. [Online]. Available : https://doi.org/10.1093/qje/qjac024

[17] R. E. Hall and M. Kudlyak, “The unemployed with jobs and without jobs,” *Labour Economics*, vol. 79, p. 102244, 2022. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0927537122001348

[18] S. Smythe, A. Grotlüschen, and K. Buddeberg, “The automated literacies of e-recruitment and online services,” *Studies in the Education of Adults*, vol. 53, no. 1, pp. 4–22, 2021.

[19] J. Sánchez-Monedero, L. Dencik, and L. Edwards, “What does it mean to ’solve’ the problem of discrimination in hiring ? social, technical and legal perspectives from the uk on automated hiring systems,” ser. FAT* ’20. New York, NY, USA : Association for Computing Machinery, 2020, p. 458–468. [Online]. Available : https://doi.org/10.1145/3351095.3372849

[20] J. Teixeira da Silva, J. Dobránszki, A. Alkhatib, and P. Tsigaris, “Curriculum vitae : Challenges and potential solutions,” vol. 8, pp. 109–127, 12 2020.

[21] G. Bosio and A. Cristini, *Is the Nature of Jobs Changing ? The Role of Technological Progress and Structural Change in the Labour Market.* Cham : Springer International Publishing, 2018, pp. 15–41. [Online]. Available : https://doi.org/10.1007/978-3-319-90548-8_2

[22] P. Castells, N. Hurley, and S. Vargas, *Novelty and Diversity in Recommender Systems.* New York, NY : Springer US, 2022, pp. 603–646. [Online]. Available : https://doi.org/10.1007/978-1-0716-2197-4_16

[23] J. Martinez-Gil, A. Paoletti, and M. Pichler, "A novel approach for learning how to automatically match job offers and candidate profiles," *Information Systems Frontiers*, vol. 22, 12 2020.

[24] L. A. Cabrera-Diego, M. El-Béze, J. M. Torres-Moreno, and B. Durette, "Ranking résumés automatically using only résumés : A method free of job offers," *Expert Systems with Applications*, vol. 123, pp. 91–107, jun 2019.

[25] M. Kaya and T. Bogers, "Understanding recruiters' information seeking behavior in talent search," in *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, ser. CHIIR '23. New York, NY, USA : Association for Computing Machinery, 2023, p. 14–23. [Online]. Available : https://doi.org/10.1145/3576840.3578311

[26] J. J. C. Raupp, *Uncertainty.* John Wiley & Sons, Ltd, 2018, pp. 1–9. [Online]. Available : https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119010722.iesc0196

[27] M. Ganesan, S. Antony, and E. George, "Dimensions of job advertisement as signals for achieving job seeker's application intention," *Journal of Management Development*, vol. 37, 05 2018.

[28] J. A. Breaugh, "Employee Recruitment," *Annual Review of Psychology*, vol. 64, pp. 389–416, 2013.

[29] T.-Y. Tung, S. Kobus, and D. Gündüz, "Context-aware effective communications," *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pp. 334–339, 2021. [Online]. Available : https://api.semanticscholar.org/CorpusID:247230496

[30] I. Khaouja, I. Kassou, and M. Ghogho, "A survey on skill identification from online job ads," *IEEE Access*, vol. 9, pp. 118 134–118 153, 2021.

[31] R. Haddad and E. Mercier-Laurent, "Curriculum vitae (cvs) evaluation using machine learning approach," in *Artificial Intelligence for Knowledge Management*, E. Mercier-Laurent, M. Ö. Kayalica, and M. L. Owoc, Eds. Cham : Springer International Publishing, 2021, pp. 48–65.

[32] A. Singh, C. Rose, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla, "Prospect : A system for screening candidates for recruitment," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM '10. New York, NY, USA : Association for Computing Machinery, 2010, p. 659–668. [Online]. Available : https://doi.org/10.1145/1871437.1871523

[33] R. Kessler, N. Béchet, M. Roche, J.-M. Torres-Moreno, and M. El-Bèze, "A hybrid approach to managing job offers and candidates," *Information Processing & Management*, vol. 48, no. 6, pp. 1124–1135, 2012. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0306457312000416

[34] D. Çelik, "Towards a semantic-based information extraction system for matching résumés to job openings," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 24, no. 1, pp. 141–159, 2016.

[35] P. Xu and D. Barbosa, "Matching résumés to job descriptions with stacked models," in *Advances in Artificial Intelligence*, E. Bagheri and J. C. Cheung, Eds. Cham : Springer International Publishing, 2018, pp. 304–309.

[36] X. Wang, Z. Jiang, L. Peng, and Z. Lu, "A deep-learning-inspired person-job matching model based on sentence vectors and subject-term graphs," *Complex.*, vol. 2021, jan 2021. [Online]. Available : https://doi.org/10.1155/2021/6206288

[37] C. Zhu, H. Zhu, H. Xiong, C. Ma, F. Xie, P. Ding, and P. Li, "Person-job fit : Adapting the right talent for the right job with joint representation learning," *ACM Trans. Manage. Inf. Syst.*, vol. 9, no. 3, sep 2018. [Online]. Available : https://doi.org/10.1145/3234465

[38] T.-T.-Q. Trinh, Y.-C. Chung, and R. Kuo, "A domain adaptation approach for resume classification using graph attention networks and natural language processing," *Know.-Based Syst.*, vol. 266, no. C, apr 2023. [Online]. Available : https://doi.org/10.1016/j.knosys.2023.110364

[39] A. Espinal, Y. Haralambous, D. Bedart, and J. Puentes, "Uncertainty-oriented textual marker selection for extracting relevant terms from job offers," in *8th International Conference on Artificial Intelligence and Fuzzy Logic System (AIFZ 2022)*, CS & IT - CSCP. CSITY, 2022, pp. 1–16.

[40] ——, "An ontology-based possibilistic framework for extracting relevant terms from job advertisements," in *Proceedings of the 14th International Joint Conference on Computational Intelligence (IJCCI 2022) - FCTA. Best Paper Award 2022 Edition*, INSTICC. SciTePress, 2022, pp. 163–174.

[41] ——, “A quality assessment framework for information extraction in job advertisements,” *SN Computer Science*, vol. 4, no. 6, pp. 787–807, Oct 2023. [Online]. Available : https://doi.org/10.1007/s42979-023-02247-5

[42] ——, “A format-sensitive bert-based approach to resume segmentation,” in *2023 33rd Conference of Open Innovations Association (FRUCT)*. IEEE, 2023, pp. 30–37.

[43] A. Habous and E. H. Nfaoui, “A fuzzy logic and ontology-based approach for improving the cv and job offer matching in recruitment process,” *Int. J. Metadata Semant. Ontologies*, vol. 15, no. 2, p. 104–120, jan 2021. [Online]. Available : https://doi.org/10.1504/ijmso.2021.120278

[44] D. Martin Jr., V. Prabhakaran, J. Kuhlberg, A. Smart, and W. S. Isaac, “Extending the Machine Learning Abstraction Boundary : A Complex Systems Approach to Incorporate Societal Context,” 2020, arXiv 2006.09663.

[45] E. Pavlick and T. Kwiatkowski, “Inherent disagreements in human textual inferences,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 677–694, 2019.

[46] M. Pejic-Bach, T. Bertoncel, M. Meško, and Živko Krstić, “Text mining of industry 4.0 job advertisements,” *International Journal of Information Management*, vol. 50, pp. 416–431, 2020. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0268401218313677

[47] V. Simon, N. Rabin, and H. C.-B. Gal, “Utilizing data driven methods to identify gender bias in linkedin profiles,” *Information Processing & Management*, vol. 60, no. 5, p. 103423, 2023.

[48] J. A. Teixeira da Silva, J. Dobránszki, A. Al-Khatib, and P. Tsigaris, “Curriculum vitae : challenges and potential solutions,” *KOME : An International Journal of Pure Communication Inquiry*, vol. 8, no. 2, pp. 109–127, 2020.

[49] L. Floridi, M. Holweg, M. Taddeo, J. Amaya Silva, J. Mökander, and Y. Wen, “Capai-a procedure for conducting conformity assessment of ai systems in line with the eu artificial intelligence act,” *Available at SSRN 4064091*, 2022.

[50] T. Tran, V. Le, H. Le, and T. M. Le, “From deep learning to deep reasoning,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD ’21. New York, NY, USA : Association for Computing Machinery, 2021, p. 4076–4077. [Online]. Available : https://doi.org/10.1145/3447548.3470803

[51] J. Kim and P. Angnakoon, “Research using job advertisements : A methodological assessment,” *Library & Information Science Research*, vol. 38, no. 4, pp. 327–335, 2016. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S074081881630322X

[52] A. Rafaeli, “Sense-making of employment : on whether and why people read employment advertising,” *Journal of Organizational Behavior*, vol. 27, pp. 747–770, 2006. [Online]. Available : https://api.semanticscholar.org/CorpusID:145454469

[53] M. M. Harris and L. S. Fink, “A field study of applicant reactions to employment opportunities : Does the recruiter make a difference ?” *Personnel Psychology*, vol. 40, pp. 765–784, 1987. [Online]. Available : https://api.semanticscholar.org/CorpusID:144386939

[54] T. T. Phan, V. Q. Pham, H. D. Nguyen, A. T. Huynh, D. A. Tran, and V. T. Pham, “Ontology-based resume searching system for job applicants in information technology,” in *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, H. Fujita, A. Selamat, J. C.-W. Lin, and M. Ali, Eds. Cham : Springer International Publishing, 2021, pp. 261–273.

[55] S. Guo, F. Alamudun, and T. Hammond, “Résumatcher : A personalized résumé-job matching system,” *Expert Systems with Applications*, vol. 60, pp. 169–182, 2016. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0957417416301798

[56] P. van Esch, J. S. Black, and J. Ferolie, “Marketing ai recruitment : The next phase in job application and selection,” *Computers in Human Behavior*, vol. 90, pp. 215–222, 2019. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0747563218304497

[57] J. Zhao, J. Wang, M. Sigdel, B. Zhang, P. Hoang, M. Liu, and M. Korayem, “Embedding-based Recommender System for Job to Candidate Matching on Scale,” 2021, arXiv 2107.00221v1.

[58] R. A. Moore and A. Reeves, “The job market’s first steps : using research tools to simplify the process,” *PS : Political Science &Amp ; Politics*, vol. 44, pp. 385–391, 2011.

[59] M. Gottlieb, S. B. Promes, and W. C. Coates, “A guide to creating a high-quality curriculum vitae,” *AEM Education and Training*, vol. 5, pp. 1–5, 2021. [Online]. Available : https://api.semanticscholar.org/CorpusID:244904831

[60] B. Macfarlane, “The cv as a symbol of the changing nature of academic life : performativity, prestige and self-presentation,” *Studies in Higher Education*, vol. 45, pp. 796 – 807, 2018.

[61] R. Le, W. Hu, Y. Song, T. Zhang, D. Zhao, and R. Yan, "Towards effective and interpretable person-job fitting," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA : Association for Computing Machinery, 2019, p. 1883–1892. [Online]. Available : https://doi.org/10.1145/3357384.3357949

[62] H. Sajid, J. Kanwal, S. U. R. Bhatti, S. A. Qureshi, A. Basharat, S. Hussain, and K. U. Khan, "Resume parsing framework for e-recruitment," in *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2022, pp. 1–8.

[63] N. Goyal, J. Kalra, C. Sharma, R. Mutharaju, N. Sachdeva, and P. Kumaraguru, "JobXMLC : EXtreme multi-label classification of job skills with graph neural networks," in *Findings of the Association for Computational Linguistics : EACL 2023*. Dubrovnik, Croatia : Association for Computational Linguistics, May 2023, pp. 2181–2191. [Online]. Available : https://aclanthology.org/2023.findings-eacl.163

[64] S. Maheshwary and H. Misra, "Matching resumes to jobs via deep siamese network," in *Companion Proceedings of the The Web Conference 2018*, ser. WWW '18. Republic and Canton of Geneva, CHE : International World Wide Web Conferences Steering Committee, 2018, p. 87–88. [Online]. Available : https://doi.org/10.1145/3184558.3186942

[65] S. Schnitzer, D. Reis, W. Alkhatib, C. Rensing, and R. Steinmetz, "Preselection of documents for personalized recommendations of job postings based on word embeddings," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC '19. New York, NY, USA : Association for Computing Machinery, 2019, p. 1683–1686. [Online]. Available : https://doi.org/10.1145/3297280.3297602

[66] D. Meletis and C. Dürscheid, *Writing Systems and Their Use. An Overview of Grapholinguistics*. Berlin/Boston : De Gruyter, 2022.

[67] Y. Liu, C. M. Eckert, and C. Earl, "A review of fuzzy ahp methods for decision-making with subjective judgements," *Expert Systems with Applications*, vol. 161, p. 113738, 2020.

[68] D. J. Deming and L. B. Kahn, "Skill requirements across firms and labor markets : Evidence from job postings for professionals," *Journal of Labor Economics*, vol. 36, pp. S337 – S369, 2017. [Online]. Available : https://api.semanticscholar.org/CorpusID:168725780

[69] M. Le Vrang, A. Papantoniou, E. Pauwels, P. Fannes, D. Vandensteen, and J. De Smedt, "ESCO : Boosting job matching in Europe with semantic interoperability," *Computer*, vol. 47, no. 10, pp. 57–64, 2014.

[70] F. Ștefanica, S. Abele, F. Walker, and R. Nickolaus, *Modeling, Measurement, and Development of Professional Competence in Industrial-Technical Professions*. Cham : Springer International Publishing, 2017, pp. 843–861. [Online]. Available : https://doi.org/10.1007/978-3-319-41713-4_39

[71] J. M. Gil, A. L. Paoletti, and M. Pichler, "A novel approach for learning how to automatically match job offers and candidate profiles," *Information Systems Frontiers*, pp. 1–10, 2016. [Online]. Available : https://api.semanticscholar.org/CorpusID:10566090

[72] A. Barducci, S. Iannaccone, V. La Gatta, V. Moscato, G. Sperlì, and S. Zavota, "An end-to-end framework for information extraction from italian resumes," *Expert Systems with Applications*, vol. 210, p. 118487, 2022. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S095741742201572X

[73] I. Jurisica, J. Mylopoulos, and E. S. K. Yu, "Ontologies for knowledge management : An information systems perspective," *Knowledge and Information Systems*, vol. 6, pp. 380–401, 2004. [Online]. Available : https://api.semanticscholar.org/CorpusID:1436526

[74] P. D. Converse, F. L. Oswald, M. A. Gillespie, K. A. Field, and E. B. Bizot, "Matching individuals to occupations using abilities and the o*net : Issues and an application in career guidance," *Personnel Psychology*, vol. 57, pp. 451–487, 2004. [Online]. Available : https://api.semanticscholar.org/CorpusID:144846920

[75] D. Alfonso-Hermelo, P. Langlais, and L. Bourg, "Automatically learning a human-resource ontology from professional social-network data," in *Advances in Artificial Intelligence*, M.-J. Meurs and F. Rudzicz, Eds. Cham : Springer International Publishing, 2019, pp. 132–145.

[76] D. Wilson, D. Leahy, and D. Dolan, "The european e-competence framework : past, present and future," *IADIS International Journal on Computer Science and Information Systems*, vol. 10, no. 1, pp. 1–13, 2015.

[77] A. González-Eras and J. Aguilar, "Determination of professional competencies using an alignment algorithm of academic profiles and job advertisements, based on competence thesauri and similarity measures," *International Journal of Artificial Intelligence in Education*, vol. 29, pp. 536 – 567, 2019. [Online]. Available : https://api.semanticscholar.org/CorpusID:201815499

[78] J. Vrolijk, S. T. Mol, C. Weber, M. Tavakoli, G. Kismihók, and M. Pelucchi, "Ontojob : Automated ontology learning from labor market data," *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pp. 195–200, 2022. [Online]. Available : https://api.semanticscholar.org/CorpusID:247618033

[79] R. Lourdusamy and A. John, "A review on metrics for ontology evaluation," *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pp. 1415–1421, 2018. [Online]. Available : https://api.semanticscholar.org/CorpusID:49539690

[80] N. Palaniappan and B. S. Arasu, "Impact of effective recruitment in business organization : A brief literature review," *Imperial journal of interdisciplinary research*, vol. 3, 2017. [Online]. Available : https://api.semanticscholar.org/CorpusID:54821489

[81] S. Mehta, R. Pimplikar, A. Singh, L. R. Varshney, and K. Visweswariah, "Efficient multifaceted screening of job applicants," in *Proceedings of the 16th International Conference on Extending Database Technology*, ser. EDBT '13. New York, NY, USA : Association for Computing Machinery, 2013, p. 661–671. [Online]. Available : https://doi.org/10.1145/2452376.2452453

[82] F. Huang and P. Cappelli, "Applicant screening and performance-related outcomes," *The American Economic Review*, vol. 100, pp. 214–218, 2010. [Online]. Available : https://api.semanticscholar.org/CorpusID:51754320

[83] E. Faliagka, L. A. Iliadis, I. Karydis, M. Rigou, S. Sioutas, A. Tsakalidis, and G. Tzimas, "On-line consistent ranking on e-recruitment : seeking the truth behind a well-formed cv," *Artificial Intelligence Review*, vol. 42, pp. 515 – 528, 2013. [Online]. Available : https://api.semanticscholar.org/CorpusID:254239458

[84] G. Sudha, S. K. K, S. J. S, N. D, S. S, and K. G., "Personality prediction through cv analysis using machine learning algorithms for automated e-recruitment process," *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, pp. 617–622, 2021. [Online]. Available : https://api.semanticscholar.org/CorpusID:246945487

[85] M. Lashkari and J. Cheng, ""finding the magic sauce" : Exploring perspectives of recruiters and job seekers on recruitment bias and automated tools," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA : Association for Computing Machinery, 2023. [Online]. Available : https://doi.org/10.1145/3544548.3581548

[86] Y. Kino, H. Kuroki, T. Machida, N. Furuya, and K. Takano, "Text analysis for job matching quality improvement," *Procedia Computer Science*, vol. 112, pp. 1523–1530, 2017, knowledge-Based and Intelligent Information & Engineering Systems : Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S1877050917313984

[87] D. Jurafsky and J. H. Martin, *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, N.J. : Pearson Prentice Hall, 2009.

[88] C. D. Manning, *An introduction to information retrieval.* Cambridge : Cambridge University Press, 2009.

[89] H. J. Walker and A. S. Hinojosa, "15 recruitment : The role of job advertisements," *The Oxford handbook of recruitment*, p. 269, 2013.

[90] M. S. Cole, R. S. Rubin, H. S. Feild, and W. F. Giles, "Recruiters' perceptions and use of applicant résumé information : Screening the recent graduate," *Applied Psychology*, vol. 56, no. 2, pp. 319–343, 2007.

[91] M. S. Cole, H. S. Feild, W. F. Giles, and S. G. Harris, "Recruiters' inferences of applicant personality based on resume screening : Do paper people have a personality ?" *Journal of Business and Psychology*, vol. 24, pp. 5–18, 2009. [Online]. Available : https://api.semanticscholar.org/CorpusID:37460611

[92] C. Zhang, H. Wang, and Y. Wu, "Resumevis," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, pp. 1 – 25, 2017. [Online]. Available : https://api.semanticscholar.org/CorpusID:9011

[93] K. Shinkawa, K. Saito, M. Kobayashi, and A. Hiyama, "Towards extracting recruiters' tacit knowledge based on interactions with a job matching system," in *Human Aspects of IT for the Aged Population. Applications, Services and Contexts*, J. Zhou and G. Salvendy, Eds. Cham : Springer International Publishing, 2017, pp. 557–568.

[94] G. V. Hoye, A. M. Saks, F. Lievens, and B. Weijters, "Development and test of an integrative model of job search behaviour," *European Journal of Work and Organizational Psychology*, vol. 24, no. 4, pp. 544–559, 2015.

[95] T. A. Judge and D. M. Cable, "Applicant personality, organizational culture, and organization attraction," *Personnel Psychology*, vol. 50, pp. 359–394, 1997. [Online]. Available : https://api.semanticscholar.org/CorpusID:55328544

[96] S. L. Brickson, "Organizational identity orientation : The genesis of the role of the firm and distinct forms of social value," *Academy of Management Review*, vol. 32, pp. 864–888, 2007. [Online]. Available : https://api.semanticscholar.org/CorpusID:145306546

[97] J. Johari and K. K. Yahya, "Linking organizational structure, job characteristics, and job performance constructs : a proposed framework," *International Journal of Biometrics*,

vol. 4, p. 145, 2009. [Online]. Available : https://api.semanticscholar.org/CorpusID:44862317

[98] J. Olenick and C. R. Dishop, "Clarifying dynamics for organizational research and interventions : A diversity example," *Organizational Psychology Review*, vol. 12, pp. 365 – 386, 2022. [Online]. Available : https://api.semanticscholar.org/CorpusID:250463078

[99] C. Bratianu, "A holistic view of the organizational knowledge dynamics," *HOLISTICA – Journal of Business and Public Administration*, vol. 9, pp. 22 – 7, 2018. [Online]. Available : https://api.semanticscholar.org/CorpusID:67760249

[100] J. D. McDowall, *An Overview of Complex Adaptive Systems.* Berkeley, CA : Apress, 2019, pp. 13–34. [Online]. Available : https://doi.org/10.1007/978-1-4842-4306-0_2

[101] R. Axelrod and M. D. Cohen, *Harnessing complexity.* Basic books, 2008.

[102] J. Ladyman, J. Lambert, and K. Wiesner, "What is a complex system ?" *European Journal for Philosophy of Science*, vol. 3, pp. 33 – 67, 2012. [Online]. Available : https://api.semanticscholar.org/CorpusID:256068817

[103] A. Sihvonen and K. Pajunen, "Causal complexity of new product development processes : a mechanism-based approach," *Innovation*, vol. 21, pp. 253 – 273, 2018. [Online]. Available : https://api.semanticscholar.org/CorpusID:53382351

[104] K. Santos, E. Loures, F. Piechnicki, and O. Canciglieri, "Opportunities assessment of product development process in industry 4.0," *Procedia Manufacturing*, vol. 11, pp. 1358–1365, 2017, 27th International Conference on Flexible Automation and Intelligent Manufacturing, FAIM2017, 27-30 June 2017, Modena, Italy. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S2351978917304730

[105] X. Zhang, M. M. Khalili, and M. Liu, "Long-term impacts of fair machine learning," *Ergonomics in Design : The Quarterly of Human Factors Applications*, vol. 28, pp. 11 – 7, 2019. [Online]. Available : https://api.semanticscholar.org/CorpusID:208840409

[106] B. K. Bala, F. M. Arshad, and K. M. Noh, *Causal Loop Diagrams.* Singapore : Springer Singapore, 2017, pp. 37–51. [Online]. Available : https://doi.org/10.1007/978-981-10-2045-2_3

[107] C. M. Zapata Jaramillo and F. Arango Isaza, "The UNC-method : a problem-based software development method," *Ingeniería e Investigación*, vol. 29, pp. 69–75, 2009.

[108] B. Rahman, "Time-delay systems : An overview," *Nonlinear Phenomena in Complex Systems*, vol. 23, no. 07, 2020.

[109] M. Garnelo and M. Shanahan, "Reconciling deep learning with symbolic artificial intelligence : representing objects and relations," *Current Opinion in Behavioral Sciences*, vol. 29, pp. 17–23, 2019, artificial Intelligence. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S2352154618301943

[110] K. Hamilton, A. Nayak, B. Bozic, and L. Longo, "Is neuro-symbolic ai meeting its promise in natural language processing ? a structured review," *ArXiv*, vol. abs/2202.12205, 2022. [Online]. Available : https://api.semanticscholar.org/CorpusID:247084226

[111] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, 2020.

[112] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification : a comprehensive review," *ACM computing surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[113] M. A. Nielsen, *Neural networks and deep learning*. Determination press San Francisco, CA, USA, 2015, vol. 25.

[114] P. J. Werbos, *The roots of backpropagation : from ordered derivatives to neural networks and political forecasting*. John Wiley & Sons, 1994, vol. 1.

[115] J. Schmidhuber, "Deep learning in neural networks : An overview," *Neural networks : the official journal of the International Neural Network Society*, vol. 61, pp. 85–117, 2014. [Online]. Available : https://api.semanticscholar.org/CorpusID:11715509

[116] W. Yun-Zhou, Z. Min-Ling, C. Lei, Z. Peng, L. Hao-Nan, and X. Bo-Lang, "Realization of tree and grass recognition based on alexnet," *2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pp. 617–621, 2022. [Online]. Available : https://api.semanticscholar.org/CorpusID:256034168

[117] L. Floridi and M. Chiriatti, "Gpt-3 : Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, pp. 681–694, 2020.

[118] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert : Pre-training of deep bidirectional transformers for language understanding," 2019.

[119] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito,

D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm : Scaling language modeling with pathways," 2022.

[120] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama : Open and efficient foundation language models," 2023.

[121] M. Mars, "From word embeddings to pre-trained language models : A state-of-the-art walkthrough," *Applied Sciences*, vol. 12, no. 17, p. 8805, 2022.

[122] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 2, pp. 604–624, 2020.

[123] M. A. Bansal, D. R. Sharma, and D. M. Kathuria, "A systematic review on data scarcity problem in deep learning : Solution and applications," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1 – 29, 2022. [Online]. Available : https://api.semanticscholar.org/CorpusID:245772675

[124] G. Marra, F. Giannini, M. Diligenti, and M. Gori, "Integrating learning and reasoning with deep logic models," in *Machine Learning and Knowledge Discovery in Databases*, U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet, Eds. Cham : Springer International Publishing, 2020, pp. 517–532.

[125] D. Dubois and H. Prade, "From possibilistic rule-based systems to machine learning - a discussion paper," in *Scalable Uncertainty Management*, J. Davis and K. Tabia, Eds. Cham : Springer International Publishing, 2020, pp. 35–51.

[126] G. Ras, N. Xie, M. Van Gerven, and D. Doran, "Explainable deep learning : A field guide for the uninitiated," *Journal of Artificial Intelligence Research*, vol. 73, pp. 329–396, 2022.

[127] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, and A. Jabbar, "A review on methods and applications in multimodal deep learning," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2s, pp. 1–41, 2023.

[128] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, pp. 261 – 266, 2015. [Online]. Available : https://api.semanticscholar.org/CorpusID:8420436

[129] K. S. Jones, *Natural Language Processing : A Historical Review.* Dordrecht : Springer Netherlands, 1994, pp. 3–16. [Online]. Available : https://doi.org/10.1007/978-0-585-35958-8_1

[130] A.-H. Dediu, J. M. Matos, and C. Martín-Vide, "Natural language processing, moving from rules to data," in *Theory and Applications of Models of Computation*, T. Gopal, G. Jäger, and S. Steila, Eds. Cham : Springer International Publishing, 2017, pp. 24–38.

[131] G. Weikum, "Foundations of statistical natural language processing," *SIGMOD Rec.*, vol. 31, no. 3, p. 37–38, sep 2002. [Online]. Available : https://doi.org/10.1145/601858.601867

[132] Y. Goldberg, *Neural network methods for natural language processing.* Springer Nature, 2022.

[133] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing : State of the art, current trends and challenges," *Multimedia tools and applications*, vol. 82, no. 3, pp. 3713–3744, 2023.

[134] D. Hovy and D. Yang, "The importance of modeling social factors of language : Theory and practice," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies.* Online : Association for Computational Linguistics, Jun. 2021, pp. 588–602. [Online]. Available : https://aclanthology.org/2021.naacl-main.49

[135] A. Liu, Z. Wu, J. Michael, A. Suhr, P. West, A. Koller, S. Swayamdipta, N. A. Smith, and Y. Choi, "We're afraid language models aren't modeling ambiguity," 2023, arXiv 2304.14399.

[136] E. Balkir, S. Kiritchenko, I. Nejadgholi, and K. Fraser, "Challenges in applying explainability methods to improve the fairness of NLP models," in *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022).* Seattle, U.S.A. : Association for Computational Linguistics, Jul. 2022, pp. 80–92. [Online]. Available : https://aclanthology.org/2022.trustnlp-1.8

[137] R. Pieraccini, *Natural Language Understanding in Socially Interactive Agents*, 1st ed. New York, NY, USA : Association for Computing Machinery, 2021, p. 147–172. [Online]. Available : https://doi.org/10.1145/3477322.3477328

[138] N. Marín, G. Rivas-Gervilla, and D. Sánchez, "Fuzzy logic for vagueness management in referring expression generation," in *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation.* Santiago de Compostela,

Spain : Association for Computational Lingustics, Sep. 2020, pp. 71–76. [Online]. Available : https://aclanthology.org/2020.intellang-1.8

[139] P. Kapustin and M. Kapustin, "Semantic parsing with fuzzy meaning representations," in *Proceedings of the Second International Workshop on Designing Meaning Representations*. Barcelona Spain (online) : Association for Computational Linguistics, Dec. 2020, pp. 78–89. [Online]. Available : https://aclanthology.org/2020.dmr-1.8

[140] F. Bobillo, J. Bosque-Gil, J. Gracia, and M. Lanau-Coronas, "Fuzzy lemon : Making lexical semantic relations more juicy," in *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*. Marseille, France : European Language Resources Association, Jun. 2022, pp. 45–51. [Online]. Available : https://aclanthology.org/2022.ldl-1.6

[141] H.-N. L. Teodorescu, "On the meaning of approximate reasoning - an unassuming subsidiary to lotfi zadeh's paper dedicated to the memory of grigore moisil -," *Int. J. Comput. Commun. Control*, vol. 6, pp. 577–580, 2011. [Online]. Available : https://api.semanticscholar.org/CorpusID:56141868

[142] D. Dubois and H. Prade, "Toward multiple-agent extensions of possibilistic logic," in *2007 IEEE International Fuzzy Systems Conference*. IEEE, 2007, pp. 1–6.

[143] ——, "Fuzzy sets in approximate reasoning, part 1 : inference with possibility distributions," *Fuzzy Sets and Systems*, vol. 100, pp. 73–132, 1999. [Online]. Available : https://api.semanticscholar.org/CorpusID:122967379

[144] B. Bouchon-Meunier, D. Dubois, L. Godo, and H. Prade, *Fuzzy Sets and Possibility Theory in Approximate and Plausible Reasoning*. Boston, MA : Springer US, 1999, pp. 15–190. [Online]. Available : https://doi.org/10.1007/978-1-4615-5243-7_2

[145] L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning—i," *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975. [Online]. Available : https://www.sciencedirect.com/science/article/pii/0020025575900365

[146] M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge, "The belief-desire-intention model of agency," in *Intelligent Agents V : Agents Theories, Architectures, and Languages : 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*. Springer, 1999, pp. 1–10.

[147] M. E. Bratman, D. J. Israel, and M. E. Pollack, "Plans and resource-bounded practical reasoning," *Computational intelligence*, vol. 4, no. 3, pp. 349–355, 1988.

[148] D. Dubois, E. Lorini, and H. Prade, "Nonmonotonic desires - a possibility theory viewpoint," in *DARe@ECAI*, 2014. [Online]. Available : https://api.semanticscholar.org/CorpusID:14408039

[149] C. Da Costa Pereira and A. G. B. Tettamanzi, "An integrated possibilistic framework for goal generation in cognitive agents," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems : Volume 1*, 2010, pp. 1239—1246.

[150] M. Vanegas-Hernandez, C. da Costa Pereira, D. Moreno, G. Fusco, A. G. B. Tettamanzi, M. Riveill, and J. T. Hernández, "A new urban segregation-growth coupled model using a belief-desire-intention possibilistic framework," in *Proceedings of the International Conference on Web Intelligence*, ser. WI '17. New York, NY, USA : Association for Computing Machinery, 2017, p. 340–347. [Online]. Available : https://doi.org/10.1145/3106426.3106486

[151] A. B. Othmane, A. Tettamanzi, S. Villata, N. L. Thanh, and M. Buffa, "An agent-based architecture for personalized recommendations," in *Agents and Artificial Intelligence*, J. van den Herik and J. Filipe, Eds. Cham : Springer International Publishing, 2017, pp. 96–113.

[152] K. Bauters, K. McAreavey, W. Liu, J. Hong, L. Godo, and C. Sierra, "Managing different sources of uncertainty in a bdi framework in a principled way with tractable fragments," *Journal of Artificial Intelligence Research*, vol. 58, pp. 731–775, 2017.

[153] D. Dubois and H. Prade, "Possibility theory and its applications : Where do we stand ?" *Springer handbook of computational intelligence*, pp. 31–60, 2015.

[154] Y. Lin, H. Lei, P. C. Addo, and X. Li, "Machine learned resume-job matching solution," 2016.

[155] G. Sridevi and S. K. Suganthi, "Ai based suitability measurement and prediction between job description and job seeker profiles," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100109, 2022.

[156] P. K. Roy, S. S. Chowdhary, and R. Bhatia, "A machine learning approach for automation of resume recommendation system," *Procedia Computer Science*, vol. 167, pp. 2318–2327, 2020.

[157] Y. Deng, H. Lei, X. Li, and Y. Lin, "An improved deep neural network model for job matching," *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 106–112, 2018. [Online]. Available : https://api.semanticscholar.org/CorpusID:49538770

[158] C. Yang, Y. Hou, Y. Song, T. Zhang, J.-R. Wen, and W. X. Zhao, "Modeling two-way selection preference for person-job fit," in *Proceedings of the 16th ACM Conference on Recommender Systems*, ser. RecSys '22. New York, NY, USA : Association for Computing Machinery, 2022, p. 102–112. [Online]. Available : https://doi.org/10.1145/3523227.3546752

[159] A. Qodad, A. El Kenz, A. Benyoussef, and M. El Yadari, "An adaptive learning system based on a matching jobs and resumes engine," in *Proceedings of the 4th International Conference on Big Data and Internet of Things*, ser. BDIoT'19. New York, NY, USA : Association for Computing Machinery, 2020. [Online]. Available : https://doi.org/10.1145/3372938.3373009

[160] T. Shao, C. Song, J. Zheng, F. Cai, H. Chen *et al.*, "Exploring internal and external interactions for semi-structured multivariate attributes in job-resume matching," *International Journal of Intelligent Systems*, vol. 2023, 2023.

[161] D. Dubois and H. Prade, "An overview of the asymmetric bipolar representation of positive and negative information in possibility theory," *Fuzzy sets and Systems*, vol. 160, no. 10, pp. 1355–1366, 2009.

[162] A. V. Zubyuk, "A new approach to specificity in possibility theory : Decision-making point of view," *Fuzzy Sets Syst.*, vol. 364, pp. 76–95, 2019. [Online]. Available : https://api.semanticscholar.org/CorpusID:85518753

[163] D. Hose and M. Hanss, "A universal approach to imprecise probabilities in possibility theory," *International Journal of Approximate Reasoning*, vol. 133, pp. 133–158, 2021. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0888613X21000438

[164] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy sets and systems*, vol. 1, no. 1, pp. 3–28, 1978.

[165] A. A. Alola, M. Tunay, and V. U. Alola, "Analysis of possibility theory for reasoning under uncertainty," *International Journal of Statistics and Probability*, vol. 2, p. 12, 2013. [Online]. Available : https://api.semanticscholar.org/CorpusID:55001379

[166] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 525–541, 2021.

[167] E. Albini, J. Long, D. Dervovic, and D. Magazzeni, "Counterfactual shapley additive explanations," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability,*

and Transparency, ser. FAccT '22. New York, NY, USA : Association for Computing Machinery, 2022, p. 1054–1070. [Online]. Available : https://doi.org/10.1145/3531146.3533168

[168] R. Levin, M. Shu, E. Borgnia, F. Huang, M. Goldblum, and T. Goldstein, "Where do models go wrong ? parameter-space saliency maps for explainability," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15 602–15 615, 2022.

[169] X.-H. Li, C. C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue *et al.*, "A survey of data-driven and knowledge-aware explainable ai," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 29–49, 2020.

[170] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 9269–9278. [Online]. Available : https://proceedings.mlr.press/v119/sundararajan20b.html

[171] R. Guidotti, "Counterfactual explanations and how to find them : literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

[172] Y. Yuan and M. J. Shaw, "Induction of fuzzy decision trees," *Fuzzy Sets and Systems*, vol. 69, no. 2, pp. 125–139, 1995.

[173] I. M. Nasir, M. A. Khan, M. Yasmin, J. H. Shah, M. Gabryel, R. Scherer, and R. Damaševičius, "Pearson correlation-based feature selection for document classification using balanced training," *Sensors*, vol. 20, no. 23, p. 6793, 2020.

[174] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, 3rd ed. New York : John Wiley & Sons, 2013, vol. 398.

[175] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94, vol. 1215. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1994, p. 487–499.

[176] S. Jarvis, "Capturing the diversity in lexical diversity." *Language Learning*, vol. 63, pp. 87–106, 2013. [Online]. Available : https://api.semanticscholar.org/CorpusID:142760241

[177] G. Lupyan and R. Dale, "Language structure is partly determined by social structure," *PloS one*, vol. 5, no. 1, p. e8559, 2010.

[178] D. Cram and B. Daille, "Terminology extraction with term variant detection," in *Proceedings of ACL-2016 system demonstrations*, 2016, pp. 13–18.

[179] A. Rigouts Terryn, V. Hoste, and E. Lefever, "In no uncertain terms : a dataset for monolingual and multilingual automatic term extraction from comparable corpora," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 385–418, 2020.

[180] N. Xu, T. Gui, R. Ma, Q. Zhang, J. Ye, M. Zhang, and X. Huang, "Cross-linguistic syntactic difference in multilingual BERT : How good is it and how does it affect transfer ?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates : Association for Computational Linguistics, Dec. 2022, pp. 8073–8092. [Online]. Available : https://aclanthology.org/2022.emnlp-main.552

[181] A. Hazem, M. Bouhandi, F. Boudin, and B. Daille, "Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France : European Language Resources Association, Jun. 2022, pp. 648–662. [Online]. Available : https://aclanthology.org/2022.lrec-1.68

[182] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2015, pp. 1200–1205.

[183] J. T. Pintas, L. A. F. Fernandes, and A. C. B. Garcia, "Feature selection methods for text classification : a systematic literature review," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 6149–6200, Feb. 2021. [Online]. Available : https://doi.org/10.1007/s10462-021-09970-6

[184] Y. Lu, M. Liang, Z. Ye, and L. Cao, "Improved particle swarm optimization algorithm and its application in text feature selection," *Applied Soft Computing*, vol. 35, pp. 629–636, 2015.

[185] N. S. M. Nafis and S. Awang, "An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification," *IEEE Access*, vol. 9, pp. 52 177–52 192, 2021.

[186] O. Gokalp, E. Tasci, and A. Ugur, "A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification," *Expert Systems with Applications*, vol. 146, p. 113176, 2020.

[187] V. Feofanov, E. Devijver, and M.-R. Amini, "Wrapper feature selection with partially labeled data," *Applied Intelligence*, vol. 52, no. 11, pp. 12 316–12 329, 2022.

[188] S. D. Sarkar and S. Goswami, "Empirical study on filter based feature selection methods for text classification," *International Journal of Computer Applications*, vol. 81, no. 6, pp. 38–43, 11 2013.

[189] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," *Expert Systems with Applications*, vol. 33, no. 1, pp. 1–5, 2007. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S095741740600114X

[190] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML '97 : Proceedings of the Fourteenth International Conference on Machine Learning*, vol. 97, 1997, pp. 412–420.

[191] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Information Processing & Management*, vol. 42, no. 1, pp. 155–165, 2006. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0306457304000962

[192] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[193] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[194] J. B. Kinney and G. S. Atwal, "Equitability, mutual information, and the maximal information coefficient," *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3354–3359, 2014.

[195] P. Kumbhar and M. Mali, "A survey on feature selection techniques and classification algorithms for efficient text classification," *International Journal of Science and Research*, vol. 5, no. 5, p. 1267–1275, 2016.

[196] S. Hara and T. Maehara, "Enumerate lasso solutions for feature selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[197] N. Zainuddin, A. Selamat, and R. Ibrahim, "Hybrid sentiment classification on twitter aspect-based sentiment analysis," *Applied Intelligence*, vol. 48, pp. 1218–1232, 2018.

[198] Q. Li, L. He, and X. Lin, "Dimension reduction based on categorical fuzzy correlation degree for document categorization," in *2013 IEEE International Conference on Granular Computing (GrC)*, 2013, pp. 186–190.

[199] Z. Zuo, J. Li, P. Anderson, L. Yang, and N. Naik, "Grooming detection using fuzzy-rough feature selection and text classification," in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018, pp. 1–8.

[200] F. Aghaeipoor and M. M. Javidi, "A hybrid fuzzy feature selection algorithm for high-dimensional regression problems : An mrmr-based framework," *Expert Systems with Applications*, vol. 162, p. 113859, 2020. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0957417420306692

[201] V. Bhatia, P. Rawat, A. Kumar, and R. R. Shah, "End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT," 2019, arXiv :1910.03089.

[202] H. Sajid, J. Kanwal, S. U. R. Bhatti, S. A. Qureshi, A. Basharat, S. Hussain, and K. U. Khan, "Resume parsing framework for e-recruitment," in *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2022, pp. 1–8.

[203] V. S. Kumaran and A. Sankar, "Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping expert," *International Journal of Metadata, Semantics and Ontologies*, vol. 8, pp. 56–64, 05 2013.

[204] M. Tikhonova and A. Gavrishchuk, "NLP methods for automatic candidate's CV segmentation," in *2019 International Conference on Engineering and Telecommunication (EnT)*, 2019, pp. 1–5.

[205] C. Ayishathahira, C. Sreejith, and C. Raseek, "Combination of neural networks and conditional random fields for efficient resume parsing," in *2018 International CET Conference on Control, Communication, and Computing*, 2018, pp. 388–393.

[206] J. Liu, Y. Shen, Y. Zhang, and S. krishnamoorthy, "Resume parsing based on multi-label classification using neural network models," in *Proceedings of the 6th International Conference on Big Data and Computing*, ser. ICBDC '21. New York, NY, USA : Association for Computing Machinery, 2021, p. 177–185.

[207] X. Li, H. Shu, Y. Zhai, and Z. Lin, "A method for resume information extraction using bert-bilstm-crf," in *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, 2021, pp. 1437–1442.

[208] A. Barducci, S. Iannaccone, V. La Gatta, V. Moscato, G. Sperlì, and S. Zavota, "An end-to-end framework for information extraction from italian resumes," *Expert Systems with Applications*, vol. 210, p. 118487, 2022.

[209] Y. Haralambous, *Fonts & Encodings. From Advanced Typography to Unicode and Everything in Between.* Sebastopol, CA : O'Reilly, 2007.

[210] Y. Haralambous and M. Dürst, "Unicode from a linguistic point of view," in *Proceedings of Graphemics in the 21st Century, Brest 2018*, Y. Haralambous, Ed. Brest : Fluxus Editions, 2019, pp. 167–183.

[211] M. Trabelsi, Z. Chen, S. Zhang, B. D. Davison, and J. Heflin, "Strubert : Structure-aware bert for table search and matching," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 442–451.

[212] K. Gu and A. Budhkar, "A package for learning on tabular and text data with transformers," in *Proceedings of the Third Workshop on Multimodal Artificial Intelligence.* Mexico City, Mexico : Association for Computational Linguistics, 2021, pp. 69–73.

[213] Y. Haralambous, "Des graphèmes à la langue et à la connaissance," Ph.D. dissertation, Université de Bretagne Occidentale, 2020, https://hal.science/tel-02986651.

[214] F. Coulmas, *Writing systems : An introduction to their linguistic analysis.* Cambridge University Press, 2003.

[215] S. Mc Gurk, C. Abela, and J. Debattista, "Towards Ontology Quality Assessment," 2017, http://ceur-ws.org/Vol-1824/ldq_paper_2.pdf.

[216] M. Somodevilla García, D. Vilariño Ayala, I. Pineda, M. Somodevilla García, D. Vilariño Ayala, and I. Pineda, "An Overview of Ontology Learning Tasks," *Computación y Sistemas*, vol. 22, no. 1, pp. 137–146, 2018.

[217] D. Alfonso-Hermelo, P. Langlais, and L. Bourg, "Automatically Learning a Human-Resource Ontology from Professional Social-Network Data," in *Canadian AI 2019 : Advances in Artificial Intelligence*, ser. LNAI, vol. 11489. Springer, 2019, pp. 132–145.

[218] K. T. Frantzi, S. Ananiadou, and J. Tsujii, "The C-value/NC-value Method of Automatic Recognition for Multi-word Terms," *Research and Advanced Technology for Digital Libraries*, vol. 1513, pp. 585 – 604, 03 2002.

[219] S. Pourahmad, S. M. T. Ayatollahi, S. M. Taheri, and Z. H. Agahi, "Fuzzy logistic regression based on the least squares approach with application in clinical studies," *Computers and Mathematics with Applications*, vol. 62, no. 9, pp. 3353–3365, nov 2011.

[220] T. Denœux, D. Dubois, and H. Prade, *Representations of Uncertainty in Artificial Intelligence : Probability and Possibility.* Cham : Springer International Publishing, 2020, pp. 69–117. [Online]. Available : https://doi.org/10.1007/978-3-030-06164-7_3

[221] E. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1–13, 1975. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0020737375800022

[222] M. Blej and M. Azizi, "Comparison of mamdani-type and sugeno-type fuzzy inference systems for fuzzy real time scheduling," *International Journal of Applied Engineering Research*, vol. 11, no. 22, pp. 11 071–11 075, 2016.

[223] R.-E. Precup and H. Hellendoorn, "A survey on industrial applications of fuzzy control," *Computers in Industry*, vol. 62, no. 3, pp. 213–226, 2011. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0166361510001363

[224] S. Chakraverty, D. M. Sahoo, and N. R. Mahato, *Defuzzification.* Singapore : Springer Singapore, 2019, pp. 117–127. [Online]. Available : https://doi.org/10.1007/978-981-13-7430-2_7

[225] D. Hovy and D. Yang, "The importance of modeling social factors of language : Theory and practice," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, 2021, pp. 588–602.

[226] L. Rello, M. Pielot, and M.-C. Marcos, "Make it big ! the effect of font size and line spacing on online readability," in *CHI '16 : Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA : Association for Computing Machinery, 2016, p. 3637–3648.

[227] C. Fournier and D. Inkpen, "Segmentation similarity and agreement," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies.* Montréal, Canada : Association for Computational Linguistics, Jun. 2012, pp. 152–161.

[228] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, pp. 673–705, Nov 2019. [Online]. Available : https://doi.org/10.1007/s10458-019-09408-y

[229] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409-410, pp. 17–26, 2017. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0020025517307235

[230] S. Thaker and V. Nagori, "Analysis of fuzzification process in fuzzy expert system," *Procedia Computer Science*, vol. 132, pp. 1308–1316, 2018, international Conference on Computational Intelligence and Data Science. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S1877050918307798

[231] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert : smaller, faster, cheaper and lighter," in *NeurIPS EMC^2 Workshop*, 2019, https ://arxiv.org/abs/1910.01108.

[232] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost : Improving classification performance when training data is skewed," in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[233] N. Reimers and I. Gurevych, "Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019, pp. 3982–3992.

[234] D. McFadden *et al.*, *Conditional logit analysis of qualitative choice behavior*. Institute of Urban and Regional Development, University of California, 1973.

[235] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi, "Revisiting Few-sample BERT Fine-tuning," in *International Conference on Learning Representations*, 2021, https ://arxiv.org/abs/2006.05987.

[236] C. d. C. Pereira and A. G. Tettamanzi, "A syntactic possibilistic belief change operator for cognitive agents," in *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 2, 2011, pp. 38–45.

[237] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[238] M. B. van Riemsdijk, *Cognitive agent programming : A semantic approach, Ph.D. dissertation*. Utrecht University, 2006.

[239] P. Gärdenfors, "Belief revision : A vade-mecum," in *International Workshop on Meta-Programming in Logic*. Springer, 1992, pp. 1–10.

[240] G. V. Caprara, L. Di Giunta, C. Pastorelli, and N. Eisenberg, "Mastery of negative affect : a hierarchical model of emotional self-efficacy beliefs." *Psychological Assessment*, vol. 25, no. 1, p. 105, 2013.

[241] J. Lang, "Conditional desires and utilities : an alternative logical approach to qualitative decision theory." in *ECAI*. PITMAN, 1996, pp. 318–322.

[242] W. Pedrycz, "Why triangular membership functions ?" *Fuzzy sets and Systems*, vol. 64, no. 1, pp. 21–30, 1994.

ANNEXES

A Cadre théorique de la possibilité

La théorie de la possibilité [142], basée sur les principes de la théorie des ensembles flous [237], offre une approche structurée pour représenter et quantifier l'incertitude. Au lieu d'une valeur déterministe pour une variable, cette théorie capture sa plage de valeurs possibles à l'aide d'une fonction d'appartenance, appelée "distribution de possibilité", notée π. Cette distribution est comprise entre 0 et 1.

Lorsqu'il existe une valeur telle que $\pi(v_0) = 1$ dans la distribution, on dit alors que cette distribution est normalisée.

Mesures de possibilité et de nécessité

Chaque distribution de possibilité, π, donne lieu à deux mesures interdépendantes : une mesure de possibilité (Π) et sa dualité, la mesure de nécessité (N). Ces deux mesures opèrent sur un ensemble défini A et sont représentées comme suit :

$$\Pi(A) = \max_{s \in A} \pi(s) \tag{5}$$

$$N(A) = 1 - \Pi(\bar{A}) = \min_{s \in \bar{A}} \{1 - \pi(s)\}. \tag{6}$$

Un autre dérivé de la distribution de possibilité est la mesure de "possibilité garantie" [161].

Mesure de possibilité garantie

Pour une distribution de possibilité donnée π et un ensemble spécifique A, la mesure de possibilité garantie, représentée par Δ, est définie comme :

$$\Delta(A) = \min_{s \in A} \pi(s). \tag{7}$$

Lors de la normalisation de la distribution de possibilité sur un domaine fini Ω, elle adhère aux propriétés suivantes pour tous les sous-ensembles A, B dans Ω :

— $\Pi(A \cup B) = \max\{\Pi(A), \Pi(B)\}$: ceci met en évidence la propriété distributive de la mesure de possibilité sur les opérations d'union ;

— $\Pi(A \cap B) \leq \min\{\Pi(A), \Pi(B)\}$: ceci souligne la nature limitée de la mesure de possibilité sur les opérations d'intersection ;

— $\Pi(\emptyset) = N(\emptyset) = 0$; $\Pi(\Omega) = N(\Omega) = 1$: mettant en évidence les extrémités des mesures de possibilité et de nécessité.
— ... et ainsi de suite, pour d'autres propriétés (cf. $\Pi(A) = 1 - N(\overline{A})$ (dualité), $N(A) \leq \Pi(A)$ (consistance)...).

Un aperçu mathématique crucial dérivé de ces propriétés est $\max\{\Pi(A), \Pi(\bar{A})\} = 1$, soulignant la dualité inhérente à la théorie de la possibilité. Dans les situations d'incertitude absolue concernant A, à la fois $\Pi(A)$ et $\Pi(\bar{A})$ égalent 1, dénotant une incertitude maximale.

B Langage propositionnel et le cadre agent de croyance-désir-intention possibiliste

Dans cette annexe, les fondations du cadre CDI possibiliste sont présentées. Ce cadre a été initialement proposé par [2], et il a été adapté pour l'automatisation des informations les plus pertinentes des OE.

Langage propositionnel et ses interprétations

Dans le domaine des agents cognitifs, un langage propositionnel classique sert d'outil puissant pour encapsuler et manipuler l'information.

Construction du langage

Considérons A comme un ensemble fini de propositions atomiques. Nous définissons L comme étant le langage propositionnel englobant satisfaisant les critères suivants :
— $A \cup \{\top, \bot\}$ est un sous-ensemble de L ;
— pour chaque paire de formules ϕ, ψ appartenant à L, l'ensemble L inclut également la négation $\neg\phi$, la conjonction $\phi \wedge \psi$, et la disjonction $\phi \vee \psi$;
— au-delà de ces connecteurs logiques de base, on peut introduire des connecteurs auxiliaires qui agissent comme des raccourcis pour diverses combinaisons dans L ; par exemple, l'implication $\phi \supset \psi$ peut être représentée comme $\neg\phi \vee \psi$.

Nous utilisons Ω pour représenter l'ensemble de toutes les interprétations possibles sur A, mathématiquement exprimé comme :

$$\Omega = \{0, 1\}^A. \tag{8}$$

Une interprétation I dans Ω est caractérisée comme une association :

$$I : A \to \{0, 1\}. \tag{9}$$

Ceci attribue une valeur de vérité p_I à chaque proposition atomique p dans A. Par extension,

cette cartographie attribue une valeur de vérité ϕ_I à toutes les formules ϕ dans L.

Définition : L'ensemble modèle d'une formule ϕ dans L est noté $[\phi]$, et est défini comme :

$$[\phi] = \{I \in \Omega : I \models \phi\}. \tag{10}$$

Pour un sous-ensemble de formules S qui fait partie de L, nous pouvons étendre la définition ci-dessus à :

$$[S] = \{I \in \Omega : \forall \phi \in S, I \models \phi\} = \bigcap_{\phi \in S} [\phi]. \tag{11}$$

Cette définition souligne l'universalité des interprétations qui satisfont chaque formule dans l'ensemble S.

C Codage des croyances et désirs chez les agents cognitifs

Les croyances et les désirs servent de piliers cognitifs pour les agents, capturant des situations couvrant le passé, le présent et le futur. Ces éléments essentiels sont quantifiés en utilisant deux distributions de possibilité distinctes : π (pour les croyances) et u (pour les désirs). Une distinction essentielle surgit ici : alors que π est normalisée (reflétant l'hypothèse selon laquelle les croyances d'un agent doivent être cohérentes), u n'est pas strictement normalisée, tenant compte de l'éventuelle incohérence des désirs.

La genèse des désirs peut être retracée aux règles de génération de désir. Ces règles peuvent être perçues comme une adaptation possibiliste du concept de "règles d'adoption de désir" de van Riemsdijk [238, 2].

Règles de génération de désir

Une règle de génération de désir, notée R, peut être formulée comme :

$$\beta_R, \psi_R \Rightarrow_D^+ \phi, \tag{12}$$

où β_R, ψ_R, et ϕ sont tous des éléments de L. La formule à droite de R est représentée par l'expression $\text{rhs}(R)$.

Si la règle est inconditionnelle, elle prend la forme :

$$\alpha \Rightarrow_D^+ \phi \tag{13}$$

avec α résidant dans l'intervalle $(0, 1]$.

En interprétant une règle de génération de désir conditionnelle, on déduit : "Si un agent croit fermement en β_R et désire ψ_R, il aura un désir pour chaque scénario où ϕ est vrai." Dans le

contexte de l'utilité qualitative, cela se traduit par : "L'agent attribue une utilité qualitative à chaque monde satisfaisant ϕ qui est au moins aussi élevée que sa croyance en β_R et son désir pour ψ_R." Pour une règle inconditionnelle, la notion sous-jacente est que l'agent attribue une utilité qualitative d'au moins α à chaque monde où $I \models \phi$.

Finalement, les croyances dans ce cadre sont encapsulées par des degrés de nécessité provenant d'une distribution de possibilité normalisée :

$$\pi : \Omega \to [0, 1]. \tag{14}$$

Ici, π délimite une hiérarchie de plausibilité parmi les états possibles des affaires. Plus précisément, le degré de possibilité d'une interprétation I est représenté par $\pi(I)$.

D Croyances évaluées chez les agents cognitifs

Dans le paysage des agents cognitifs, les croyances ne sont pas simplement des entités binaires ; elles possèdent une gradation, reflétant l'intensité ou la conviction de cette croyance.

Définition : quantification de l'intensité de la croyance

Étant donné la mesure de nécessité N influencée par π, et une formule ϕ, l'intensité avec laquelle l'agent croit en ϕ est représentée comme :

$$B(\phi) = N([\phi]) = 1 - \max_{I \not\models \phi}\{\pi(I)\}. \tag{15}$$

Cette métrique fournit un aperçu quantitatif de la conviction de l'agent envers une croyance spécifique. Plus la valeur de $B(\phi)$ est élevée, plus la croyance de l'agent en ϕ est forte.

Décodage de l'état de l'agent

L'état d'un agent est une structure composite, englobant à la fois ses croyances et les règles régissant la transition des croyances aux désirs.

Définition : caractérisation de l'état de l'agent

La constitution d'un agent peut être détaillée de manière exhaustive par le tuple $S = \langle \pi, R_J \rangle$, où :

— π représente une distribution de possibilité qui donne lieu aux croyances évaluées de l'agent, notées B ;

— R_J est une collection de règles de génération de désir. En tandem avec B, ces règles formulent une attribution d'utilité qualitative, u.

Cette bifurcation souligne que, tandis que les croyances servent de fondement à la cognition de l'agent, les règles relient ces croyances aux désirs de l'agent, dessinant un tableau explicatif de l'état cognitif de l'agent.

E Dynamique de mise à jour des croyances

À mesure que les agents cognitifs naviguent dans des environnements changeants, ils rencontrent inévitablement de nouvelles informations. La manière dont ils adaptent leur système de croyances face à une telle information nouvelle, surtout lorsqu'elle provient d'une source ayant des degrés de fiabilité variables, est essentielle pour maintenir la cohérence et la consistance.

Définition : opérateur de modification de croyance

Lorsqu'un agent est confronté à une nouvelle information ϕ provenant de l'ensemble de langage L, et que cette information émane d'une source ayant une fiabilité τ (variant de 0 à 1), la distribution de possibilité π de l'agent subit une mise à jour. La nouvelle distribution de possibilité, π', qui donne naissance à l'ensemble de croyances mis à jour B', est calculée à partir de la distribution de possibilité originale π (pertinente pour l'ensemble de croyances initial B) de la manière suivante. Pour chaque interprétation I :

$$\pi'(I) = \begin{cases} \frac{\pi(I)}{\Pi([\![\phi]\!])} & \text{si } I \models \phi \text{ et } B(\neg\phi) < 1; \\ 1 & \text{si } I \models \phi \text{ et } B(\neg\phi) = 1; \\ \min\{\pi(I), (1-\tau)\} & \text{si } I \not\models \phi. \end{cases} \tag{16}$$

Cette équation met en avant trois scénarios :

— *scénario 1* : si la nouvelle information ϕ est cohérente avec les croyances de l'agent et que l'agent ne croit pas pleinement à la négation de ϕ, la distribution de possibilité est ajustée proportionnellement ;

— *scénario 2* : si l'agent croit pleinement à la négation de ϕ mais que la nouvelle information soutient ϕ, l'agent adopte sans réserve ϕ comme nouvelle croyance ;

— *scénario 3* : si la nouvelle information ϕ contredit les croyances de l'agent, l'agent ajuste sa distribution de possibilité en fonction de la fiabilité τ de la source.

L'équation 16 souligne la flexibilité de l'opérateur de modification de croyance. Il ne se contente pas de prendre en compte l'introduction de nouvelles croyances, mais facilite également la révision des croyances conflictuelles. Si une contradiction surgit entre une nouvelle information et une croyance fermement établie, l'agent, en utilisant l'opérateur susmentionné, privilégiera

les informations les plus récentes, abandonnant l'ancienne croyance pour rétablir la cohérence.

Fait intéressant, cet opérateur a été prouvé [2] pour s'aligner sur une interprétation possibiliste des postulats de rationalité de révision AGM K^*1 à K^*8 [239, 2]. Cette adhérence garantit que l'opérateur maintient une approche de révision de croyance plus rationnelle et solide.

Définition : croyance moyenne de l'agent dans la pertinence d'un terme

Un mécanisme appliqué de manière cohérente est introduit pour évaluer les croyances globales d'un agent concernant la pertinence de termes spécifiques.

S'inspirant des fondements de la psychologie cognitive [240], il est suggéré que les croyances humaines pourraient présenter une structure hiérarchique. Une telle organisation implique que les agents cognitifs peuvent posséder des croyances primaires, qui sont potentiellement étayées par des croyances secondaires ou de niveau inférieur dans cette hiérarchie. Sur cette base, nous introduisons le concept de croyance moyenne d'un agent sur la pertinence d'un terme. Cette croyance intègre diverses perspectives de chaque marqueur textuel tout en tenant compte de leurs niveaux respectifs d'incertitude (par exemple, l'ambiguïté).

Définissons la croyance moyenne de l'agent comme suit :

— soient $\alpha_{1,t_i}, \alpha_{2,t_i}, ..., \alpha_{n,t_i}$ les niveaux de possibilité associés à n marqueurs liés au terme t_i, chaque niveau de possibilité du marqueur étant supérieur à zéro en ce qui concerne la pertinence du terme ;

— soient $\tau_{1,t_i}, \tau_{2,t_i}, ..., \tau_{n,t_i}$ les niveaux de confiance de chaque marqueur correspondant ;

— soit m le nombre total de marqueurs intégrés dans l'architecture de l'agent.

Par conséquent, la croyance moyenne de l'agent dans la pertinence d'un terme est :

$$\overline{B}(t_i) = \frac{\sum_{i=1}^{n} \alpha_{i,t_i} * \tau_{i,t_i}}{m}. \tag{17}$$

Cette approche se distingue de—et complète—l'opérateur de modification de croyance introduit précédemment par [2]. Tandis que ce dernier met l'accent sur la dominance des croyances les plus fortes sur les plus faibles, notre mécanisme se concentre sur une perspective globale en moyennant les croyances, en tenant compte même des contributions les plus faibles.

Ces deux stratégies d'évaluation des croyances sont complémentaires :

— *profondeur vs. largeur :* l'opérateur de changement de croyance explore en profondeur comment un agent devrait ajuster ses croyances à la lumière de nouvelles informations tout en tenant compte de la fiabilité de la source. En revanche, la mesure de croyance moyenne offre une vue panoramique, agrégeant les perspectives de tous les marqueurs ;

— *application :* l'opérateur de changement de croyance est réactif, déclenché par l'introduction de nouvelles données. D'autre part, la mesure de croyance moyenne est réfléchie, visant à fournir une vision englobante du système de croyances de l'agent ;

— *fiabilité vs. incertitude :* l'opérateur de changement de croyance s'appuie sur la fiabilité lors de l'évaluation de nouvelles informations. À l'inverse, la mesure de croyance moyenne équilibre à la fois la fiabilité et l'incertitude inhérente (via τ_i), offrant un aperçu plus détaillé de l'état des croyances de l'agent.

Définition : croyance globale de l'agent sur la pertinence d'un terme

Sur la base de la discussion précédente, nous présentons une métrique globale des croyances de l'agent :

— laissons $B(t_i)$ désigner la croyance de l'agent dans la pertinence du terme t_i, dérivée à l'aide de l'opérateur de modification de croyance après traitement de l'information des marqueurs textuels ; ce composant reflète la pertinence du terme du point de vue des marqueurs les plus fiables dans le cadre de l'agent ;

— laissons $\overline{B}(t_i)$ être la croyance moyenne de l'agent sur la pertinence du terme, dérivée à l'aide du mécanisme d'évaluation de la croyance moyenne de l'agent ; ce composant englobe le point de vue de l'agent, même lorsqu'il est influencé par des croyances plus faibles.

Ainsi, nous définissons la mesure de croyance globale de l'agent concernant la pertinence d'un terme comme :

$$C(t_i) = w * B(t_i) + (1 - w) * \bar{B}(t_i). \tag{18}$$

Dans le contexte de l'évaluation de la croyance d'un agent concernant la pertinence des termes, l'équation suggère l'importance de combiner à la fois la profondeur des croyances les plus fiables et le large consensus des croyances. L'expression $C(t_i) = w * B(t_i) + (1 - w) * \bar{B}(t_i)$ sert précisément cet objectif. Ici, $B(t_i)$ représente les croyances fiables de l'agent dérivée des marqueurs les plus fiables, capturant la profondeur de la croyance de l'agent. D'un autre côté, $\overline{B}(t_i)$ désigne une croyance moyenne qui intègre les perspectives de tous les marqueurs, même ceux ayant des contributions plus faibles, capturant ainsi la largeur. Le facteur de pondération w permet d'équilibrer ces deux perspectives.

En utilisant cette combinaison pondérée, l'équation garantit que la croyance globale résultante, $C(t_i)$, n'est ni trop influencée par quelques croyances fortes, ni diluée par la multitude de croyances moyennes. Elle fournit une compréhension plus équilibrée, complète et nuancée de la croyance de l'agent sur la pertinence d'un terme, rendant l'équation utile pour améliorer le classement des termes des OE.

Dynamique de mise à jour des désirs

Alors que les croyances forment la base fondamentale de la compréhension d'un agent cognitif, les désirs représentent les aspirations de l'agent. Ces désirs, contrairement aux croyances, peuvent

parfois être non cohérents, reflétant la complexité des souhaits de l'agent.

Définition : quantifier les désirs justifiés

Étant donné une affectation d'utilité qualitative u (qui est formellement représentée comme une distribution de possibilité), l'intensité avec laquelle un agent désire une formule ϕ dans L est formulée comme :

$$J(\phi) = \Delta([\phi]) = \min_{I \models \phi} u(I). \tag{7}$$

L'intuition sous-jacente ici est qu'un désir est considéré comme "justifié" lorsque chaque monde concevable qui réalise le désir est jugé favorable par l'agent. Interpréter $J(\phi)$ comme un degré d'appartenance cristallise l'ensemble flou J, qui englobe le bassin de désirs justifiés de l'agent.

Cependant, comment un agent détermine-t-il son affectation d'utilité qualitative u ?

Définition : activation des règles de génération de désir

Considérez une règle de génération de désir R de la forme $\beta_R, \psi_R \Rightarrow_D^+ \phi$. L'intensité d'activation, ou le degré auquel R est "activé", noté $\mathrm{Deg}(R)$, est défini comme :

$$\mathrm{Deg}(R) = \min\{B(\beta_R), J(\psi_R)\}. \tag{19}$$

Pour les règles qui sont inconditionnelles, à savoir de la forme $\alpha_R \Rightarrow_D^+ \phi$, le degré d'activation se simplifie à :

$$\mathrm{Deg}(R) = \alpha_R. \tag{20}$$

Ce cadre suppose que les intensités de croyance et de désir sont commensurables. Cette hypothèse ouvre la voie à une juxtaposition directe des intensités de croyance et de désir, soulignant l'interconnexion de ce qu'un agent croit et de ce qu'il cherche à réaliser.

Pour délimiter davantage la mécanique, soit R_I^J le sous-ensemble de R_J où seules les règles dont le côté droit est vrai dans le monde I sont conservées. Pour un état donné $S = \langle \pi, R_J \rangle$, un algorithme séquentiel est utilisé pour calculer l'affectation d'utilité qualitative correspondante, u [2].

Intentions et objectifs : distiller les désirs en aspirations réalisables

Alors que les désirs encapsulent les souhaits d'un agent, les objectifs représentent des cibles réalisables. Une distinction clé émerge : les désirs peuvent souvent abriter des contradictions, mais les objectifs, étant réalisables, doivent être élagués pour former un sous-ensemble cohérent

de désirs. Cela nous amène à l'interprétation possibiliste de la spécificité du désir, une notion initiée par Lang [241]. Ce concept joue un rôle crucial dans la sélection des objectifs.

Proposition : spécificité dans les désirs

Considérez un ensemble de formules $S \subseteq L$. Soit ϕ_1 et ϕ_2 deux éléments de S tels que ϕ_1 est plus spécifique que ϕ_2, noté $[\phi_1] \subseteq [\phi_2]$. Alors, leurs intensités de désir respectives satisfont :

$$J(\phi_2) \leq J(\phi_1). \tag{21}$$

Corollaire : équivalence dans les désirs

Si deux désirs ϕ_1 et ϕ_2 sont équivalents, c'est-à-dire $\phi_1 \equiv \phi_2$, alors leur intensité de désir est également identique :

$$J(\phi_1) = J(\phi_2). \tag{22}$$

Proposition : désirs composés

Le désir d'un agent pour une proposition composée, $\phi_1 \wedge \phi_2$, ne dicte pas nécessairement qu'il désire les composants individuels ϕ_1 ou ϕ_2. Mathématiquement :

$$J(\phi_1 \wedge \phi_2) > 0 \nRightarrow J(\phi_1) > 0. \tag{23}$$

Cette proposition souligne la complexité des désirs : un désir composite ne garantit pas un désir pour ses constituants individuels.

F Prétraitement de l'OE

La phase initiale du traitement automatisé des OE est le prétraitement des documents. Cette étape consiste à extraire du texte à partir de divers formats numériques – PDF, Word, ou HTML – et à éliminer des éléments superflus tels que les en-têtes et les pieds de page. Les sous-sections ci-dessous approfondissent les détails.

Conversion de documents

Les OE existent en plusieurs formats tels que PDF, Word et HTML. Pour assurer un traitement cohérent, ces fichiers sont convertis en un format unifié. JSON est le format de choix pour sa large compatibilité avec les outils d'extraction et sa capacité à inclure des métadonnées. Il est également propice aux plateformes d'analyse de texte à grande échelle comme la pile Elastic.

Nettoyage de documents

La phase suivante consiste à purger le document du "bruit" – des éléments tels que les entêtes, les pieds de page, les numéros de page et les publicités qui peuvent entraver un traitement efficace. Cette étape est centrale pour préserver l'essence de chaque OE, en particulier lorsqu'il s'agit de petits ensembles de données, comme dans cette étude.

Segmentation de documents

Après le nettoyage, le texte est divisé en segments plus petits et gérables. La segmentation se produit à divers niveaux – paragraphes, phrases, syntagmes, mots – en fonction de l'ontologie utilisée pour la représentation. Une segmentation fine est essentielle pour capturer les nuances de chaque OE sans compliquer excessivement le processus.

Normalisation de documents

Enfin, les documents segmentés sont normalisés. Ils sont structurés selon le schéma défini par l'ontologie qui inclut des sections spécifiques comme le titre du poste, la description du poste et les compétences requises. Le but est d'optimiser le traitement automatisé à grande échelle. Bien que cette étude se concentre sur la qualité plutôt que sur la quantité, la normalisation assure la comparabilité entre les OE diverses.

G Extraction de terminologie

Dans le domaine de la CCO, comprendre les subtilités du langage utilisé dans les OE et les CV est primordial. Le lexique de ces documents est rempli de jargons spécifiques à l'industrie, d'acronymes et de termes spécialisés, présentant des défis pour garantir des correspondances précises et significatives. De telles complexités soulignent la nécessité de cadres analytiques sophistiqués. Un cadre particulièrement prometteur pour relever ce défi est l'extraction automatique de termes (EAT), un domaine qui explore ces documents pour identifier et mettre en évidence systématiquement les termes spécifiques au domaine. Ce faisant, il promet une représentation plus robuste, favorisant un alignement plus pertinent entre les OE et les profils professionnels.

Le domaine de la CCO est fréquemment confronté à des variations terminologiques. Par exemple, des descripteurs tels que "développeur front-end", "ingénieur frontend" et "développeur UI" représentent souvent des rôles analogues. Sans un mécanisme efficace pour naviguer dans ces nuances linguistiques, des correspondances potentielles peuvent être négligées, entraînant des inefficacités dans le processus d'appariement de postes. L'extraction terminologique sert à atténuer cela, en regroupant des variantes terminologiques, instaurant ainsi une couche de

normalisation. Cela garantit que même les termes très techniques ou de niche sont alignés de manière plus pertinente avec leurs concepts respectifs.

Cependant, les mérites de l'extraction terminologique vont au-delà de la simple gestion des complexités linguistiques. À l'ère actuelle, où les marchés du travail évoluent rapidement, les ontologies traditionnelles, souvent utilisées dans l'analyse sémantique, peuvent devenir obsolètes. De plus, bien que les modèles basés sur les transformateurs présentent une solution innovante, leur déploiement reste difficile en raison des complexités de la recherche en cours [181]. L'absence d'un processus d'extraction terminologique raffiné laisse les modèles existants de CCO et d'analyse automatique de CV et d'OE susceptibles d'erreurs, en particulier lorsqu'ils sont confrontés aux variations linguistiques individuelles des candidats, des recruteurs et des organisations. De telles variations pourraient introduire des termes ou des concepts qui échappent même à des modèles sophistiqués, en particulier ceux qui dépendent fortement de corpus étendus pour l'affinage, comme les grands modèles du langage.

Les résultats d'études obtenus dans cette thèse soutiennent le rôle vital de l'extraction terminologique. Non seulement elle améliore la précision des algorithmes de CCO, mais son absence amplifie également les vulnérabilités du système, en particulier avec la terminologie en constante évolution utilisée par les recruteurs et les candidats. À la lumière de ces évidences, il devient évident que l'extraction terminologique n'est pas simplement un cadre auxiliaire. Au contraire, elle se présente comme un pilier fondamental dans le processus d'extraction et d'analyse des informations des OE et des CV.

En conclusion, à mesure que le domaine de la CCO continue d'évoluer et de faire face à des défis multifacettes, le rôle de l'extraction terminologique devient de plus en plus essentiel. Elle offre non seulement de la précision, mais aussi une profondeur de compréhension, se révélant indispensable à chaque étape du processus d'appariement de postes.

L'approche d'extraction de la terminologie

Dans le domaine de l'appariement de postes entre OE et CV, la terminologie joue un rôle crucial. Les spécificités linguistiques et techniques inhérentes à ce domaine nécessitent une approche méticuleuse pour l'extraction de termes. Dans le contexte de cette recherche, nous avons choisi une méthodologie basée sur des règles pour l'extraction terminologique. Ce choix est fondé sur l'observation que les termes au sein des OE et des CV adhèrent souvent à des structures courtes et prévisibles. Par exemple, des phrases comme "développeur Java" ou "spécialiste du marketing numérique" présentent des structures récurrentes dans de tels documents. De plus, cette approche s'avère particulièrement bénéfique dans des contextes avec de plus petits ensembles de données et où la transparence dans le processus d'extraction est primordiale.

Suivant la méthodologie proposée par [178], l'approche se concentre sur la détermination

de la *spécificité terminologique* (le degré auquel une unité linguistique fonctionne comme un terme dans un domaine spécifique) et de la *cohésion terminologique* (la mesure dans laquelle une séquence de mots forme une unité cohésive et significative) des termes potentiels à travers deux étapes cruciales :

— **repérage** : à cette phase, des unités semblables à des termes sont détectées dans les textes, en se concentrant principalement sur les sous-ensembles de phrases nominales ; les phrases nominales sont des séquences de mots qui ont un nom comme mot principal, éventuellement accompagné de modificateurs ; dans une OE recherchant un "ingénieur logiciel avec expérience en Python", les unités "ingénieur logiciel" et "expérience en Python" seraient identifiées comme des termes potentiels.

— **filtrage et classement** : par la suite, ces unités subissent un raffinement et une organisation ; les termes moins fiables sont exclus, et les candidats restants sont classés en fonction de critères tels que leur cohésion et leur pertinence contextuelle.

La procédure commence par un pipeline TAL complet qui englobe la tokenisation (le processus de conversion d'une séquence de texte en tokens ou mots individuels), l'étiquetage morpho-syntaxique (attribution de libellés de type de mot tels que nom, verbe, adjectif, etc.), la lemmatisation (réduction des mots à leur forme de base ou de dictionnaire) et la stemmisation (réduction des mots à leur forme racine). Une caractéristique distinctive de la méthodologie est le composant conçu pour reconnaître les variantes de termes, conduisant à une précision et une richesse accrues de l'extraction terminologique.

Composant de repérage de termes multi-mots

Ce composant est chargé d'identifier les termes multi-mots et leurs variations à l'aide de UIMA Tokens Regex. Il se concentre sur les annotations qui apparaissent séquentiellement dans le texte, résultant en un moteur de reconnaissance de complexité linéaire [178]. Des exemples tels que "gestion de projet" ou "analyste système" seraient efficacement identifiés par ce composant en raison de leur structure composite couramment trouvée dans des documents liés à l'emploi.

Syntaxe des tokens regex UIMA

UIMA Tokens Regex utilise une syntaxe influencée par Stanford TokensRegex [178], articulée à travers une grammaire ANTLR4 [9]. Ce cadre introduit trois types de correspondance et facilite la création de règles sur des séquences d'annotations UIMA, chacune distinguée par des quantificateurs spécifiques. Par exemple, dans le contexte de la CCO, reconnaître des phrases comme "full-time position" ou "3-year experience" impliquerait une telle syntaxe regex.

9. `https://github.com/antlr/antlr4`

Application d'extraction de terminologie

Une règle illustrative extrait les Termes Multi Mots (TMM) — des séquences comme "développeur logiciel senior" ou "spécialiste du marketing numérique". Ce système regex permet des pré-définitions de correspondance pour plus de clarté et de réutilisabilité. Par exemple, en définissant N pour les noms et A pour les adjectifs, le système peut repérer efficacement des termes comme "compétences avancées en Python" ou "codage avancé". De plus, le système intègre un filtrage lexical et contextuel, visant à réaliser une extraction plus raffinée qui peut différencier, par exemple, "Java" le langage de programmation et "Java" l'île géographique.

Regroupement de variantes de termes

La méthodologie est équipée pour regrouper les termes en fonction de critères syntaxiques et morphologiques prédéfinis, en utilisant la syntaxe YAML. Dans le domaine de la CCO, cela signifie reconnaître que des phrases comme "développeur front-end" et "ingénieur frontend" sont des variantes sémantiques d'un concept similaire.

Syntaxe pour le regroupement de variantes

Une règle de variante établit les critères pour associer des candidats de termes. Elle englobe un nom de règle, des motifs spécifiques pour les termes source et cible, et une expression logique booléenne. Cette 'règle', traitée par un moteur Groovy, est élaborée dans une syntaxe Groovy valide. Par exemple, une telle règle pourrait discerner la similarité linguistique entre "spécialiste RH" et "Spécialiste des Ressources Humaines", les regroupant en raison de leur champ sémantique partagé.

Regroupement de termes basé sur des caractéristiques

La méthode utilise à la fois les caractéristiques de lemme et de racine des mots pour regrouper les termes. Par défaut, le lemme est utilisé, ce qui signifie que des conditions comme ‘s[0] == t[0]‘ sont interprétées comme comparant les lemmes des termes. La règle "S-PI-NN-P" illustre cela en liant des termes comme "management of operations" and "operational management" basés sur des racines partagées.

Variantes morphologiques

La morphologie, l'étude de la structure des mots, joue un rôle essentiel dans l'extraction de termes. En utilisant l'outil Compost, l'approche identifie si un terme, encapsulé dans une seule unité graphique, se qualifie de terme simple (à mot unique (TMU)) ou complexe (composé de plusieurs mots (MWT)). Pour les composés, Compost offre une séparation optimale pour

délimiter ses composants. Cette distinction aide à reconnaître et à regrouper des termes morphologiquement liés. Par exemple, dans le domaine de la CCO, des termes tels que "market-driven" and "market drive" seraient reconnus comme étant liés.

Moteur et regroupement de variantes

Étant donné la complexité de $O(n^2)$ associée au regroupement de variantes de termes, le système a recours à la pré-indexation comme moyen computationnel. Chaque candidat au terme subit une indexation basée sur des paires de ses lemmes à mot unique. En opérant uniquement sur des termes partageant des clés d'indexation identiques, le système garantit un traitement rationalisé et efficace, crucial pour de vastes ensembles de données comme les OE et les CV.

Grammaires de langage et classement par pertinence terminologique

L'extraction terminologique possède des règles de repérage MWT et des règles de regroupement de variantes adaptées pour l'anglais et le français.

La pertinence terminologique fait référence à la probabilité qu'une séquence linguistique soit un terme dans un domaine spécifique. Dans cette approche, les termes sont classés en fonction de leur spécificité terminologique, évaluée à l'aide du weirdness ratio (WR). Le WR est une métrique comparant la proéminence relative d'un terme dans un corpus spécifique à un domaine (comme les OE) à sa prévalence dans un corpus générique de la langue. La référence linguistique générale ici comprend des journaux de la collection CLEF [178], garantissant une perspective linguistique large pour évaluer l'unicité des termes dans le corpus spécifique de domaines des OE et des CV.

H Ensembles flous triangulaires

Cette annexe présente des concepts flous clés appliqués dans la méthodologie proposée.

Ensemble flou triangulaire

Un ensemble flou triangulaire (EFT) [242] A dans l'univers du discours X est défini par trois paramètres : a, b, et c, tels que $a \leq b \leq c$. La fonction d'appartenance $\mu_A(x)$ de cet EFT est donnée par :

$$\mu_A(x) = \begin{cases} \frac{x-a}{b-a} & \text{pour } a \leq x < b \\ \frac{c-x}{c-b} & \text{pour } b \leq x \leq c \\ 0 & \text{sinon.} \end{cases} \tag{24}$$

Ici,

– a et c sont les "pieds" du triangle, représentant respectivement les limites inférieure et supérieure du support.

– b est le "sommet" du triangle, représentant le point où la valeur d'appartenance est maximale (c'est-à-dire, 1). Dans la Figure IV.8, un exemple de fonctions triangulaires standards est présenté.

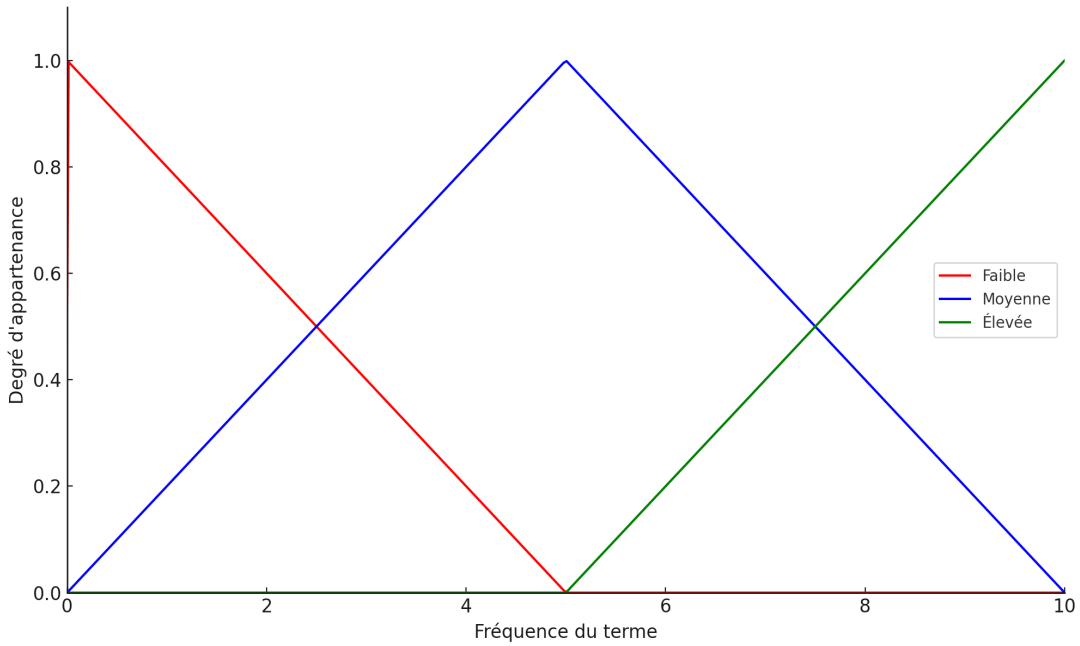


FIGURE IV.8 – Fonctions triangulaires standards utilisées pour représenter trois catégories floues de pertinence d'un terme, en considérant uniquement la fréquence comme évidence de pertinence : faible, moyenne et élevée.

Forme standard de l'EFT

La forme standard d'un EFT est un cas spécifique où le triangle est symétrique, impliquant $b-a=c-b$. Ainsi, b devient le point médian entre a et c, rendant la représentation plus simple et plus intuitive.

Justification de l'utilisation de l'EFT standard

— **Simplicité :** l'EFT standard a une forme symétrique, ce qui le rend plus facile à visualiser, à comprendre et à utiliser ; les calculs deviennent plus simples en raison de la symétrie inhérente ;

— **polyvalence :** malgré sa simplicité, l'EFT standard est suffisamment polyvalent pour représenter une large gamme d'informations incertaines ; en ajustant a, b, et c, différents

niveaux d'incertitude peuvent être capturés ;

— **interprétabilité linguistique :** les ensembles flous triangulaires, en particulier les standards, correspondent bien aux variables linguistiques (comme "faible", "moyen", "élevé") ; cela les rend adaptés aux scénarios où l'interprétation humaine est essentielle ;

— **efficacité computationnelle :** dans des scénarios où l'objectif principal est de démontrer l'utilité de la logique floue plutôt que de trouver la fonction floue optimale, l'EFT standard offre un équilibre entre précision et efficacité computationnelle ;

— **gestion de l'incertitude :** l'essence même des ensembles flous est de gérer l'incertitude, le flou et l'imprécision ; l'EFT standard, en raison de sa simplicité et de sa polyvalence, peut représenter un large éventail de scénarios incertains sans la complexité computationnelle de fonctions floues plus complexes ;

— **fondation pour des ensembles plus complexes :** l'EFT sert de fondation ; des ensembles flous plus complexes, comme les ensembles trapézoïdaux ou gaussiens, peuvent être mieux compris lorsqu'on maîtrise une fonction floue fondamentale comme la triangulaire.

Bien qu'il existe diverses formes d'ensembles flous (comme trapézoïdaux, gaussiens, etc.), l'ensemble flou triangulaire, en particulier dans sa forme standard, offre un mélange harmonieux de simplicité et de capacité. Il sert d'outil efficace pour gérer l'incertitude, ce qui en fait un choix populaire dans de nombreuses applications, des systèmes de contrôle aux processus de prise de décision.

I Moteur d'inférence Mamdani

Le moteur d'inférence Mamdani, également connu sous le nom de méthode d'inférence Max-Min, est une approche largement utilisée dans les systèmes de contrôle flous. Il fonctionne sur le principe d'évaluation des règles dans une base de règles pour déterminer la sortie d'un système flou.

Notation :

— soit R_i la i-ième règle de la base de règles ;

— A_i et B_i sont les ensembles flous associés à l'antécédent et au conséquent de R_i, respectivement ;

— x est l'entrée du système flou ;

— $\mu_{A_i}(x)$ est la fonction d'appartenance de A_i évaluée à x.

Évaluation de la règle : pour chaque règle R_i dans la base de règles :

$$R_i : \text{SI } x \text{ EST } A_i \text{ ALORS } y \text{ EST } B_i. \tag{25}$$

La force ou le poids d'activation de la règle R_i est donné par :

$$w_i = \mu_{A_i}(x) \tag{26}$$

Agrégation : l'ensemble flou de sortie agrégé B' est obtenu en combinant les sorties de toutes les règles :

$$\mu_{B'}(y) = \bigvee_i (w_i \wedge \mu_{B_i}(y)), \tag{27}$$

où $\bigvee$ désigne l'opérateur maximum et $\wedge$ désigne l'opérateur minimum.

Défuzzification : la sortie nette finale y^* du système est obtenue en défuzzifiant l'ensemble flou de sortie agrégé B'.

En résumé, le Moteur d'Inférence Mamdani consiste à évaluer les règles de la base de règles pour déterminer les forces d'activation, à agréger les résultats pour obtenir un ensemble flou de sortie, puis à défuzzifier cet ensemble pour obtenir une sortie nette.

J L'algorithme Apriori

Objectif : trouver tous les ensembles d'actions d'annotation (combinaisons d'actions) qui ont un support (fréquence d'occurrence) supérieur à un seuil spécifié α dans un ensemble d'OE donné.

Notation :

— I = ensemble de toutes les actions décrites sous forme de triplets = $\{i_1, i_2, \ldots, i_m\}$. À partir de maintenant, nous ne ferons plus référence aux « actions » mais aux triplets.

— T = ensemble de toutes les OE = $\{t_1, t_2, \ldots, t_n\}$.

— Chaque OE t_j est un sous-ensemble de I.

Algorithme :

1. **Initialisation** :
   — Commencez avec $k = 1$
   — L_1 = ensemble des 1-ensembles de triplets fréquents (OE qui atteignent le seuil de support minimum α)
2. **Boucle Principale** :
   — Tant que L_k n'est pas vide :
      — Générez C_{k+1} : Ensembles de triplets candidats de taille $k+1$ à partir des k-ensembles de triplets fréquents dans L_k en joignant les k-ensembles de triplets avec eux-mêmes.
      — Pour chaque OE t dans T :
         — C_t = sous-ensembles de t qui sont candidats
         — Pour chaque candidat c dans C_t :

— $count[c] = count[c] + 1$
— $L_{k+1} = \{c \in C_{k+1} | count[c] \geq \alpha \times n\}$
— $k = k + 1$

3. **Résultat** :
— Retournez $\bigcup_k L_k$

Génération de candidats :
— Pour générer C_{k+1} à partir de L_k :
— Pour chaque paire d'ensembles de triplets l_1, l_2 dans L_k :
— Si les $k-1$ premiers triplets de l_1 sont les mêmes que les $k-1$ premiers triplets de l_2, alors :
— Fusionnez l_1 et l_2 pour former un ensemble de triplets de $k+1$.

Élagage :
— Après avoir généré C_{k+1}, éliminez les candidats qui ont un sous-ensemble de triplets de k qui n'est pas dans L_k.

L'algorithme Apriori exploite la propriété selon laquelle tous les sous-ensembles non vides d'un ensemble d'triplets fréquent doivent également être fréquents. Cette propriété est utilisée pour élaguer l'espace de recherche, rendant l'algorithme plus efficace.

L'algorithme continue jusqu'à ce qu'aucun ensemble de triplets fréquent ne puisse être trouvé. Le résultat est une collection de tous les ensembles de triplets (d'actions d'annotation des recruteurs) qui ont un support supérieur à α.

K Le processus analytique hiérarchique flou

La méthode Processus Analytique Hiérarchique Flou (FAHP, Fuzzy Analytic Hierarchy Process) est une extension de la méthode AHP classique qui permet de gérer l'incertitude et l'imprécision dans les jugements des décideurs [67].

Définition des ensembles

Soit $C = \{c_1, c_2, \ldots, c_n\}$ l'ensemble des compétences essentielles de l'OE que nous appellerons "critères", et $A = \{a_1, a_2, \ldots, a_m\}$ un ensemble d'informations contextuelles du CV que nous appellerons "sous-critères" ou "alternatives".

Matrice de comparaison par paires floue

Pour chaque critère $c_i \in C$, une matrice de comparaison par paires floue $F_i = [f_{ijk}]$ de dimension $m \times m$ est construite, où f_{ijk} est un nombre flou représentant la préférence de l'alternative a_j par rapport à a_k selon le critère c_i.

Vecteur de poids flou

Le vecteur de poids flou $W_i = [w_{ij}]$ pour chaque critère c_i est calculé en utilisant une méthode appropriée, comme la méthode de la moyenne géométrique floue. Chaque élément w_{ij} représente le poids de l'alternative a_j selon le critère c_i.

La moyenne géométrique est utilisée en raison de sa consistance multiplicative dans les matrices de comparaison. Cette technique agrège efficacement les nombres flous, réduit l'impact des valeurs extrêmes et maintient la transitivité des préférences floues. De plus, elle est invariante à l'échelle, ce qui simplifie la gestion de critères divers et normalise les données, assurant ainsi des comparaisons équitables entre les différentes informations contextuelles du CV.

Matrice de décision floue

La matrice de décision floue $F = [f_{ij}]$ de dimension $n \times m$ est construite, où chaque élément f_{ij} est un nombre flou représentant l'importance de l'alternative a_j par rapport au critère c_i.

Défuzzification et score global

La défuzzification est effectuée pour convertir les nombres flous en valeurs scalaires. Le score global pour chaque alternative est ensuite calculé en agrégeant les valeurs scalaires obtenues après la défuzzification. Ceci peut être réalisé en utilisant un opérateur d'agrégation approprié, comme l'opérateur de somme ou moyenne pondérée.

En conclusion, la méthode FAHP offre une approche robuste et flexible pour la prise de décision multicritère en présence d'incertitude et d'imprécision. En analysant les perspectives des recruteurs, cette méthode facilite l'examen des interactions entre les informations contextuelles du CV et les compétences essentielles mentionnées dans l'OE. Ceci permet de mesurer le degré d'influence de ces informations sur la pertinence des compétences professionnelles avancées par le candidat.

L Gestion de qualité de l'ontologie

La qualité de l'ontologie mère et ses composants est évaluée à travers des métriques de qualité d'ontologie. Ces métriques sont catégorisées selon la norme de qualité de données ISO/IEC 25012 [215], qui comprend plusieurs catégories :

- **Précision** :
  - IAC1 : domaine ou étendue incorrects - le domaine ou l'étendue d'une propriété est défini de manière incorrecte.
  - IAC2 : utilisation incorrecte de propriété - une propriété est utilisée d'une manière qui contredit sa définition.

- IAC3 : caractéristiques de propriété incorrectes - les caractéristiques d'une propriété (fonctionnelle, inverse fonctionnelle, symétrique, transitive) sont définies de manière incorrecte.
- IAC4 : hiérarchie de classes incorrecte - la hiérarchie des classes est définie de manière incorrecte.
- IAC5 : hiérarchie de propriétés incorrecte - la hiérarchie des propriétés est définie de manière incorrecte.
- IAC6 : type d'individu incorrect - un individu est typé de manière incorrecte.

— **Exhaustivité** :
- ICM1 : propriété manquante - une propriété qui devrait être dans l'ontologie est manquante.
- ICM2 : classe manquante - une classe qui devrait être dans l'ontologie est manquante.
- ICM3 : individu manquant - un individu qui devrait être dans l'ontologie est manquant.
- ICM4 : domaine ou étendue manquants - le domaine ou l'étendue d'une propriété est manquant.
- ICM5 : caractéristiques de propriété manquantes - les caractéristiques d'une propriété (fonctionnelle, inverse fonctionnelle, symétrique, transitive) sont manquantes.
- ICM6 : hiérarchie de classes manquante - la hiérarchie des classes est manquante.
- ICM7 : hiérarchie de propriétés manquante - la hiérarchie des propriétés est manquante.

— **Cohérence** :
- ICC1 : domaine ou étendue incohérents - le domaine ou l'étendue d'une propriété est incohérent.
- ICC2 : utilisation incohérente de propriété - une propriété est utilisée de manière incohérente.
- ICC3 : caractéristiques de propriété incohérentes - Les caractéristiques d'une propriété (fonctionnelle, inverse fonctionnelle, symétrique, transitive) sont incohérentes.
- ICC4 : hiérarchie de classes incohérente - la hiérarchie des classes est incohérente.
- ICC5 : hiérarchie de propriétés incohérente - la hiérarchie des propriétés est incohérente.
- ICC6 : type d'individu incohérent - un individu est typé de manière incohérente.

— **Actualité** :
- ICCU1 : propriété obsolète - une propriété est obsolète.
- ICCU2 : classe obsolète - une classe est obsolète.
- ICCU3 : individu obsolète - un individu est obsolète.

— **Conformité** :

- ISCO1 : utilisation non standard des constructions de langage - les constructions de langage sont utilisées de manière non standard.
- ISCO2 : utilisation non standard des annotations - les annotations sont utilisées de manière non standard.

— **Compréhensibilité** :

- ISU1 : annotations manquantes - des annotations qui devraient être dans l'ontologie sont manquantes.
- ISU2 : regroupement de propriétés - les propriétés sont regroupées au lieu d'être correctement séparées.
- ISU3 : utilisation de différentes conventions de nommage - différentes conventions de nommage sont utilisées dans l'ontologie.

— **Disponibilité** :

- SA1 : le degré selon lequel l'ontologie est facilement disponible en ligne sous un usage spécifique.

COLLEGES SCIENCES
BRETAGNE POUR L'INGENIEUR
LOIRE ET LE NUMERIQUE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Modélisation d'offres d'emploi et de curriculums vitæ en vue d'un processus de présélection automatisée de candidats

Mot clés : Correspondance curriculum vitae - offre d'emploi, Présélection de candidats en recrutement, Modélisation curriculum vitae - offre d'emploi, Incertitude dans la présélection automatique de candidats, Contextualisation de la correspondance curriculum vitae - offre d'emploi, Intérprétabilité de la correspondance curriculum vitae - offre d'emploi

Résumé : Dans le contexte de la présélection des candidats lors du recrutement, cette thèse explore la modélisation des offres d'emploi et des curriculums vitae en vue d'une correspondance automatique plus robuste entre ces documents. Malgré les avancées significatives engendrées par l'intelligence artificielle pour l'appariement automatisé, plusieurs défis subsistent, tels que la gestion des données non structurées des CV, la prise en compte de l'incertitude lors de l'extraction des informations des offres d'emploi, la nécessité de garantir la transparence et l'interprétabilité des méthodes employées, et la considération des avis et du contexte organisationnel des recruteurs.

Pour adresser ces enjeux, nous introduisons une méthodologie orientée à l'intégration des stratégies des recruteurs, la représentation de l'incertitude et l'interprétabilité des traitements automatiques de la langue. L'approche proposée combine l'extraction d'informations saillantes et l'annotation sémantique des offres d'emploi grâce à un modèle possibiliste basé sur des ontologies. Elle intègre également les grands modèles de langage, tels que BERT, ainsi que la représentation grapholinguistique du texte pour une analyse approfondie des curriculums vitae. Mise en œuvre au sein du département de recrutement de DSI Group, une société française de conseil en Technologies de l'Information et de la Communication, cette méthodologie a révélé une capacité accrue à identifier des informations essentielles des offres d'emploi et à segmenter de manière efficiente les curriculums vitae contemporains. Cela contribue à la construction de processus de correspondance automatique plus robustes et sensibles au contexte.

Notre travail met en lumière les défis et opportunités contemporains de la correspondance entre ces documents, et suggère des solutions novatrices intégrant des méthodes traditionnelles et récentes. Ces perspectives peuvent intéresser chercheurs, professionnels du recrutement et décideurs.

Title: Modeling of Job Advertisements and Resumes for an Automated Pre-selection Process of Candidates.

Keywords: Resume - job advertisement matching, Candidate preselection in recruitment, Resume - job advertisement modeling, Uncertainty in automated candidate preselection, Contextualization of resume - job advertisement matching, Interpretability of resume - job advertisement matching

Abstract: In the context of pre-selecting candidates during recruitment, this thesis explores the modeling of job advertisements and resumes with the objective of defining a more robust automatic matching between these documents. Despite the significant advancements brought about by artificial intelligence for automated matching, several challenges remain. These include managing unstructured data from resumes, accounting for uncertainty during the extraction of information from job advertisements, ensuring the transparency and interpretability of the methods used, and considering the opinions and organizational context of recruiters.

To address these challenges, we introduce a methodology oriented towards integrating recruiter strategies, representing uncertainty, and interpreting automatic language processing. The proposed approach combines the extraction of salient information and the semantic annotation of job advertisements through a possibilistic model based on ontologies. It also incorporates major language models, such as BERT, as well as grapholinguistic representation of text for an in-depth analysis of resumes. Implemented within the recruitment department of DSI Group, a French consulting company in Information and Communication Technologies, this methodology has demonstrated an increased ability to identify essential information from job advertisements and efficiently segment contemporary resumes. This contributes to the construction of more robust and context-sensitive automatic matching processes.

Our work highlights the contemporary challenges and opportunities of matching these documents and suggests innovative solutions integrating traditional and recent methods. These insights may be of interest to researchers, recruitment professionals, and decision-makers.