Recherche d'Information Possibiliste: De la Désambiguïsation et la Reformulation de Requêtes vers la Fiabilité de l'Information Recherchée

Bilel Elayeb

Université de la Manouba

Ecole Nationale des Sciences de l'Informatique

Ecole doctorale STICODE

Mémoire de Recherche

Présenté en vue de l´obtention d'une

Habilitation Universitaire

de l'Ecole Nationale des Sciences de l'Informatique

(Spécialité Informatique)

par

Bilel Elayeb

Recherche d'Information Possibiliste:

De la Désambiguïsation et la Reformulation de Requêtes vers la Fiabilité de l'Information Recherchée

---

Soutenue le 15/12/2018 devant le jury composé de :

Président : Mme. Henda Hajjami Ben Ghézala, Professeur, ENSI, Université de la Manouba.

Rapporteurs : Mme. Chiraz Latiri, Maître de Conférences, ISAMM, Université de la Manouba.

Mme. Lynda Tamine, Professeur, Université Paul Sabatier, Toulouse, France.

Examinateurs : Mme. Rim Faiz, Professeur, IHEC, Université de Carthage.

Mme. Narjès Bellamine Ben Saoud, Professeur, ENSI, Université de la Manouba.

A ma femme Myriam

A mes enfants Issra, Yassmine et Adam

A mes parents et ma grand-mère

A mes frères et sœurs

A tous ceux que j'aime…

Remerciements

Je tiens à remercier très sincèrement les membres du jury qui m'ont honoré d'avoir accepté d'évaluer ce travail. En particulier, je remercie :

- Mme. Henda Hajjami Ben Ghézala, Professeur à l'Ecole Nationale des Sciences de l'Informatique (ENSI, Université de la Manouba), d'avoir accepté de présider le jury de mon habilitation universitaire.
- Mme. Chiraz Latiri, Maître de Conférences à l'Institut Supérieur des Arts Multimédia de la Manouba (ISAMM, Université de la Manouba) et Mme. Lynda Tamine, Professeur à l'Université Paul Sabatier de Toulouse (France), pour l'intérêt qu'elles ont porté à mes travaux de recherche en acceptant de rapporter mon mémoire d'habilitation. La qualité de mon manuscrit a été considérablement améliorée grâce à leurs remarques constructives et pertinentes.
- Mme. Rim Faiz, Professeur à l'Institut des Hautes Etudes Commerciales de Carthage (IHEC, Université de Carthage) d'avoir accepté d'examiner mon travail et faire partie du jury.
- Mme. Narjès Bellamine Ben Saoud, Professeur et Directrice de l'ENSI, pour l'honneur qu'elle m'a fait en acceptant d'être examinateur de mon habilitation. Je la remercie vivement pour la confiance qu'elle a toujours témoignée à mon égard depuis neuf ans de travail de recherches postdoctorales. Je profite de cette occasion pour apprécier l'occasion qu'elle m'a fourni pour co-encadrer des nombreux projets des Mastères et des Thèses qui ont contribué à l'existence et au succès de ce travail. Qu'elle trouve ici le fruit de mes efforts comme témoignage de mon très grand respect.

Je remercie également les membres de l'Ecole Doctorale de l'Ecole Nationale des Sciences de l'Informatique de l'université de la Manouba pour le temps qu'ils ont consacré à l'évaluation de mon dossier d'habilitation universitaire.

Mes vifs et sincères remerciements s'adressent à Monsieur Fabrice Evrard, Maître de Conférences à l'ENSEEIHT à Toulouse (France) pour tout le soutien qu'il m'a apporté durant dix ans de collaboration scientifiques avec lui (2005-2014). C'est grâce à son soutien et ses précieux conseils que ce travail de recherche a pu progresser au cours des années et a pu voir le jour. Je remercie Fabrice d'avoir accueillir mes étudiants chercheurs au sein de son équipe de recherche à l'Institut de Recherche en Informatique de Toulouse (IRIT) en France.

Je voudrai aussi exprimer mes reconnaissances et remerciements à Monsieur Ibrahim Bounhas, Maître-Assistant en Informatique à l'Institut Supérieur de Documentation (ISD) de l'université de la Manouba pour ses précieuses contributions et collaborations scientifiques. Je saisis cette occasion pour apprécier sa compétence et sa rigueur scientifique.

Mes profonds remerciements vont également à Monsieur Mohamed Ben Ahmed, Professeur Emérite à l'Ecole Nationale des Sciences de l'Informatique de l'université de la Manouba pour ses précieux conseils et encouragements continus.

Je tiens aussi à exprimer ma profonde gratitude à Monsieur Yahya Slimani, Professeur à l'Institut Supérieur des Arts Multimédia de la Manouba. Je lui adresse toute ma gratitude pour son soutien, ses précieux conseils et ses encouragements durant mon parcours postdoctoral.

Je tiens à remercier mes collègues professeurs et chercheurs français et tunisiens qui ont collaboré avec notre équipe de recherche. Je pense particulièrement au Professeur Mathieu Roche, Directeur de Recherche CIRAD et membre du laboratoire LIRMM à l'université de Montpellier (France), au Professeur Kamel Smaili de l'université de Lorraine (France), au Professeur Kais Haddar de l'Université de Sfax, ainsi qu'au Professeur Lamia Labed Jilani de l'ISG de Tunis pour leurs précieuses remarques pertinentes et constructives lors de l'évaluation des travaux de nos thésards.

Mes sincères remerciements s'adressent également à mes collègues et étudiants chercheurs du laboratoire RIADI, particulièrement les membres de ma petite équipe de recherche. Je pense particulièrement aux Dr. Oussama Ben Khiroun, Dr. Raja Ayed, Mme Wiem Ben Romdhane, Mlle Amina Chouigui, Mme Wiem Lahbib et Mme Nadia Soudani. C'est grâce à leurs efforts, leurs disponibilités et leurs compétences scientifiques que nous avons pu partager et discuter plusieurs idées scientifiques au cours de nos nombreuses réunions de travail. Je les remercie pour la confiance qu'ils m'ont accordée pour codiriger et réussir nos travaux de recherche.

Enfin, je remercie vivement ma femme Dr. Myriam Bounhas, mes deux filles Issra et Yassmine et mon fils Adam pour leur patience et de m'avoir encouragé et toléré mes absences continues et répétitives. J'espère que ce mémoire soit une récompense de leurs sacrifices. Je n'oublie jamais à remercier et saluer fortement les membres de ma grande famille : ma mère, mon père, mes frères et sœurs, ma grand-mère, ma tante, mes oncles, ma belle-mère ainsi que Monsieur Said Bounhas pour leur soutien moral constant et leurs encouragements.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

1. Contexte général et objectifs de recherche

Ce travail présente une synthèse de nos travaux de recherche accomplis durant les neuf dernières années de notre parcours universitaire au sein du laboratoire RIADI[1] auprès de l'Ecole Nationale des Sciences de l'Informatique (ENSI) en collaboration avec des structures de recherche aussi bien tunisiennes que françaises telles que : LISI[2], IRIT[3], LIPN[4], etc. Ce projet s'inscrit dans le domaine de la recherche d'information qui recouvre plusieurs domaines tels que l'extraction de connaissances, le Traitement Automatique des Langues (TAL) et l'étude et l'évaluation de la fiabilité de l'information recherchée. Nous partons dans ce projet des travaux réalisés au cours de notre thèse de doctorat ayant pour objet "**SARIPOD** : **S**ystème multi-**A**gent de **R**echerche **I**ntelligente **PO**ssibiliste de **D**ocuments Web". Cette thèse a été préparée en cotutelle entre l'Institut Nationale Polytechnique de Toulouse (INPT) en France et l'Ecole Nationale des Sciences de l'Informatique de l'Université de la Manouba en Tunisie et soutenue à Toulouse en juin 2009 [Elayeb, 2009]. En effet, les travaux de notre thèse ont été focalisés sur la RI monolingue et plus particulièrement le français. Depuis, et pour notre projet d'habilitation, nous avons étendu nos problématiques à la RI translinguistique. Outre les langues néolatine et anglo-saxonne (français et anglais), nous nous sommes intéressés à impliquer la langue arabe en RI. C'est un vrai défi de la recherche puisque la langue arabe est extrêmement ambiguë et difficile à traiter par comparaison aux autres langues. De plus, les travaux de RI et TAL dédiés à l'arabe sont souvent insuffisants par rapports aux travaux de recherche pour les autres langues et sont en évolution continue.

Un processus classique de RI implique plusieurs acteurs à savoir la requête de l'utilisateur, la collection de documents, ainsi que les diverses phases permettant d'avoir des résultats satisfaisant les besoins de l'utilisateur. Ces étapes sont principalement : (i) l'analyse et l'indexation, (ii) les modélisations de la requête et des documents, (iii) l'appariement des deux modèles de requête et de documents, et (iv) l'évaluation et la rétroaction. Néanmoins, le résultat retourné par n'importe quel système de recherche d'information (SRI) ne peut pas satisfaire les besoins de l'utilisateur si ce dernier ne présente pas clairement et explicitement sa requête de départ. Notre objectif/motivation derrière ce projet d'habilitation est de répondre à un problème de type : "Etant donné une requête utilisateur, comment un SRI pourra retourner un ensemble de documents le plus **pertinent** et le plus **fiable** possible ? ". En fait, ce problème pourra être synthétisé sous forme des **trois verrous scientifiques** dont les travaux connexes du domaine de RI ont essayé de lever. Ces verrous proviennent à la fois de la requête utilisateur elle-même ainsi que de l'ensemble de documents pertinents retournés par le SRI comme suit :

A. *Problème de requête ambiguë:* une requête ambiguë contient des termes qui ont plusieurs sens possibles dans un dictionnaire. Ainsi, le SRI retournera tous les documents contenant tous les sens possibles des termes de cette requête. Uniquement un seul sens (à identifier étant donné le contexte de la requête) est pertinent pour l'utilisateur, et les documents liés à ce sens sont ceux qui répondent au besoin de l'utilisateur. Le reste des documents liés aux autres sens sont non

[1] **R**echerche en génie log**I**ciel, **A**pplications distribuées, systèmes **D**écisionnels et **I**magerie intelligente.
[2] **L**aboratoire d'**I**nformatique pour les **S**ystèmes **I**ndustriels, INSAT, Tunisie.
[3] **I**nstitut de **R**echerche en **I**nformatique de **T**oulouse, France.
[4] **L**aboratoire d'**I**nformatique de **P**aris **N**ord, France.

désiré par l'utilisateur. D'où l'importance de la résolution du problème d'ambigüité de la requête avant de lancer le processus de recherche. Ceci diminuera le nombre de documents non-pertinents et augmentera davantage la précision du SRI.

B. *Problème de requête de contexte limité:* une requête contenant un nombre limité de termes suggéré par l'utilisateur est souvent évaluée de contexte limité. C'est le cas lorsque l'utilisateur du SRI souffre de connaissances limitées de son domaine de recherche et/ou de difficultés linguistiques pour bien exprimer son besoin d'information. En conséquence, le SRI retournera un nombre limité de documents pertinents correspondants au contexte limité de la requête suggérée. D'où l'importance de la résolution du problème de contexte limité de la requête avant de commencer l'étape de la recherche. Ainsi, l'expansion sémantique du contexte de la requête permet au SRI d'élargir l'ensemble de documents pertinents, et en conséquence d'améliorer la précision du résultat retourné. Par ailleurs, les deux problèmes d'ambigüité et de contexte limité de la requête ne sont pas indépendants, car le processus d'expansion du contexte de la requête pourra injecter des nouveaux termes ambigus dans la requête reformulée. Ainsi, il faut résoudre le problème d'ambigüité de la requête avant et/ou après son expansion.

C. *Problème d'identification du degré de fiabilité des documents pertinents retournés:* l'ensemble de documents pertinents correspondants aux ensembles de requêtes correctement désambiguïsées et reformulées pourront souffrir du problème de degrés de fiabilités. En fait, ces documents jugés pertinents sont issus des différents sources ayant des différents niveaux de fiabilité (fiable, suspect, non fiable). Par exemple, un document jugé pertinent par le SRI, mais provenant d'une source d'information non fiable, ne pourra pas convaincre l'utilisateur. D'où l'importance de la résolution du problème d'identification et d'évaluation de la fiabilité de l'information recherchée afin de mieux satisfaire le besoin de l'utilisateur.

D'abord, nous nous sommes intéressés d'emblée à la résolution de l'ambiguïté sémantique des termes de la requête afin d'améliorer la performance globale de RI en réduisant l'effet de bruit. Malgré leurs multiples avantages, les dictionnaires traditionnels souffrent de l'absence d'informations précises utiles pour la désambiguïsation. En outre, les méthodes d'apprentissage souffrent des limites de couverture des corpus sémantiquement étiquetés. Pour ces multiples raisons, l'utilité d'une nouvelle ressource linguistique externe s'impose. Une ressource qui puisse combiner les dictionnaires traditionnels et les corpus étiquetés afin de profiter de leur double avantage. Cette ressource sera utile pour l'identification des sens corrects de mots selon leur contexte.

D'autre part, l'ambiguïté morphologique est l'une des formes d'ambiguïté les plus critiques en langue arabe. Elle est détectée quand l'analyse fournit, à un mot donné, plusieurs valeurs pour certains attributs morphologiques non-conformes au contexte de ce mot. En fait, l'analyse morphologique d'un mot a pour objectif d'identifier les valeurs d'un grand nombre de caractéristiques ou d'attributs morphologiques d'un mot donné, comme la catégorie grammaticale, le genre, le nombre, etc. Ainsi, une approche pour la désambiguïsation morphologique arabe est nécessaire pour faire face à l'ambiguïté des mots non-voyellés. La désambiguïsation consiste à attribuer la valeur exacte d'un attribut morphologique parmi celles proposées par l'analyseur. De nombreux travaux utilisent des approches de classification pour résoudre la tâche de désambiguïsation morphologique. Ceci nous a encouragés à proposer de nouvelles approches pour la désambiguïsation morphologique arabe basée sur les techniques de classification.

Ensuite, nous nous focalisons sur l'expansion sémantique de requêtes utilisant une ressource linguistique externe. En fait, les corpus ont une couverture linguistique limitée par rapport aux autres ressources telles que les thésaurus, les ontologies et les dictionnaires. Bien que construits à partir des

indexes de documents pour englober les sens de mots qui sont clairement distinguées, les thésaurus souffrent des problèmes de couverture et d'ambiguïté. Alors que les ontologies exigent des outils sophistiqués utiles pour l'extraction et l'organisation des connaissances et elles sont les plus coûteuses en termes de temps nécessaire pour leurs constructions. De plus, les ontologies ne sont pas disponibles pour tous les domaines et toutes les langues. Quant aux dictionnaires, ils sont construits pour couvrir l'ensemble de la langue et représentent naturellement les significations et les relations entre les mots. En outre, ils sont considérés comme les ressources les plus exhaustives étant donné leurs disponibilités pour toutes les langues. Par ailleurs, les dictionnaires multilingues s'avèrent utiles en RI multi- et translinguistique[5] [Elayeb et Bounhas, 2016]. Ainsi, les dictionnaires sont les ressources les plus génériques qui sont les plus aptes pour soutenir le processus d'expansion de requêtes dans les SRI intelligents. Pour tous ses arguments, nous proposons et comparons des approches d'expansion sémantique de requêtes basées sur un dictionnaire.

Malgré leurs performances, les techniques d'expansion sémantique de requêtes souffrent du problème d'ambiguïté des termes de la requête à reformuler provoquant des bruits dans les documents retournés. Comme les techniques de désambiguïsation sémantique de requêtes permettent de pallier au problème d'ambiguïté, nous étudions l'impact de la désambiguïsation sémantique de requêtes sur leurs expansions.

Enfin, nous nous sommes intéressés à l'étude, l'identification et l'évaluation de la fiabilité de l'information retournée par le SRI. Nous nous intéressons particulièrement à la fiabilité des textes arabes. Historiquement, les savants arabes se sont intéressés au problème de la fiabilité durant les nombreux siècles passés. Alors qu'aujourd'hui ce problème résulte de l'expansion de l'Internet englobant un grand nombre de documents venant de plusieurs sources différentes. En fait, les savants arabes ont focalisé leurs efforts sur l'assurance de la transmission fiable des discours historiques et des événements entre les personnes. Pour cela, ils ont suivi une méthodologie dédiée englobant un ensemble de règles de transmission de l'information et d'étude de sa fiabilité. Nous signalons que la crédibilité des acteurs participant à la production ou à la transmission de l'information ainsi que leurs relations sont les principaux facteurs utiles pour l'évaluation de la fiabilité.

Ainsi, la démarche de cette méthodologie s'articule sur les étapes suivantes : D'abord, identifier l'ensemble des acteurs de la narration ainsi que leurs relations. Ensuite, évaluer la réputation de chaque acteur selon sa biographie. Enfin, attribuer une classe de fiabilité à chaque narration. Cependant, les experts du domaine ont attribué manuellement un degré de confiance à chaque narrateur afin de juger sa réputation. En partant de notre intuition que cette méthodologie pourra pallier au problème d'étude et d'évaluation de la fiabilité de l'information, nous proposons un ensemble d'outils qui mettent en œuvre ces règles. Notre objectif étant de réduire les efforts humains en automatisant les étapes de la méthodologie de traitement des narrations arabes afin d'évaluer leurs fiabilités. Sur le plan théorique, nous exploitons la méthodologie de narration arabe afin de proposer un modèle générique pour l'évaluation des données de fiabilité. Ce modèle s'appuie sur les pratiques liées à la narration arabe et propose aussi des guides théoriques nécessaires pour étudier la fiabilité des autres types de textes arabes à savoir classiques ou modernes.

[5] Un SRI translinguistique (ou par croisement de langues) retourne des documents écrits dans une langue différente de celle des requêtes. Il nécessite une étape de traduction des documents ou de la requête. Dans un SRI multilingue, en plus du module de traduction, il faudrait un autre module de combinaison des résultats dans les différentes langues pour aboutir à une seule liste fusionnée (la façon traditionnelle de présenter les résultats de recherche). C'est ce problème de fusion qui augmente la complexité de la RI multilingue, en plus de la difficulté de la traduction.

2. Positionnement : Vers la RI possibiliste fiable

Un SRI est confié de sélectionner à partir d'une collection de documents, ceux qui sont capables de satisfaire un besoin d'information de l'utilisateur exprimé par une requête. La plupart des SRI existants calculent un score de pertinence entre l'index de la requête et celui du document. Ce score de pertinence est souvent connu par le modèle d'appariement de RI. Généralement, ces modèles sont fondés sur des notions algébriques, logiques, probabilistes et statistiques.

2.1. Insuffisances des modèles existants de RI

Dans la littérature, les modèles existants de RI mesurent la pertinence d'un document soit via un score d'appariement entre la requête et le document, soit à travers une probabilité de pertinence d'un document étant donné une requête. Ces modèles sont distingués selon les trois différentes modélisations de la pertinence [Dominich, 2001] comme suit :

- La pertinence est modélisée par la similarité entre la requête et le document [Salton, 1971 ; Singhal et al., 1997].
- La pertinence est vue comme une variable aléatoire binaire. Des modèles probabilistes calculent la probabilité de pertinence des documents vis-à-vis une requête. Ces modèles eux-mêmes peuvent être distingués en deux sous familles: les modèles basés sur la génération de requêtes [Maron, 1961 ; Robertson et al., 1982 ; Fuhr, 1992] ; où les poids des termes dépendent de facteurs liés au jugement de l'utilisateur, et les modèles basés sur la génération de documents [Robertson et Jones, 1976 ; Rijsbergen, 1977 ; Fuhr, 1992] ; où les poids des termes d'indexation dépendent de la répartition des termes dans les classes des documents pertinents et non-pertinents. Mais, quel que soit le type de modèle, il est difficile d'estimer les poids des termes. D'autre part, les modèles probabilistes calculent une valeur unique, donnant l'événement (la pertinence) et son contraire.
- La pertinence est modélisée par l'incertitude sur la déduction de la requête à partir du document. Ces modèles sont fondés sur des probabilités calculant l'incertitude liées aux inférences [van Rijsbergen, 1986 ; Turtle et Croft, 1991]. Dans ces modèles, la représentation de la requête ou des documents ne dépend pas du cadre des inférences. Autrement dit, et au contraire de la pertinence, les termes ne sont pas déduits à partir de la requête ou des documents.

En fait, quelle que soit la définition de la pertinence ou la sémantique proposée pour la représentation de la requête ou du document, les modèles décrites ci-dessous possèdent des caractéristiques générales pareilles telles que :

- La plupart d'entre eux modélisent la requête et les documents par une liste de mots clés pondérés.
- Ces poids sont considérés comme des données certaines.
- Le score de pertinence est calculé à partir des poids des termes de la requête et ceux des documents.
- Seuls les termes de la requête communs à ceux des documents sont considérés.
- Les termes de la requête absents des documents ne sont pas pris en compte.
- Le document est pertinent ou non-pertinent à un certain degré.

D'abord, les travaux de recherche existants sur la notion de la pertinence [Saracevic, 1975, 1996 ; Rijsbergen, 1979 ; Schamber, 1990 ; Harter, 1992 ; Froehlich et Eisenberg, 1994, Mizzaro, 1997 ;

Borlund et Ingwersen, 1998 ; Kekäläinen et Järvelin, 2002 ; Boughanem et Brini, 2003 ; Brini et Boughanem, 2003 ; Boughanem et al., 2009] ont confirmé qu'il est difficile de trouver une définition précise de la pertinence, étant donné que cette notion dépend de la perception de l'utilisateur, a un caractère multidimensionnelle et dynamique. En conséquence, il s'est avéré qu'il est difficile de couvrir la totalité de la sémantique de la pertinence via une valeur unique, car ceci ne peut pas être certain ni fiable.

Ensuite, ces modèles de RI affectent un poids unique à chaque terme existant dans les documents. Ce poids mesure le degré de représentativité du terme du contenu d'un document donné. Ces poids sont dans la majorité des cas calculés par la combinaison de l'exhaustivité (TF) et la spécificité (IDF). Cependant, ces deux facteurs sont définis sur des données appuyant sur des échelles différentes et induisent implicitement des incertitudes et des imprécisions non suffisamment traitées par les modèles existants de RI [Boughanem et al., 2009].

Enfin, lors du calcul des scores de la pertinence, ces modèles ne considèrent que les termes de la requête communs à ceux des documents. Toutefois, certains termes de la requête sont parfois aptes à sélectionner une partie des documents de la collection. En conséquence leur ignorance lors du calcul des scores de pertinence des documents ne les contenant pas pose le verrou de l'incertitude dans ce calcul.

Etant donné le cadre possibiliste sur lequel s'appuient nos travaux dans ce mémoire, la section 2.2 présente brièvement les fondements de cette théorie en comparant ses contributions par rapport à un cadre probabiliste classique. Nous nous focalisons principalement sur la RI ainsi que la classification possibiliste. Dans la section 2.3, nous détaillons le processus d'évaluation de la fiabilité.

2.2. Concepts de base de la théorie des possibilités

Introduite par [Zadeh, 1978] et développée par plusieurs auteurs tels que [Dubois et Prade, 1988], la théorie des possibilités traite l'incertitude dans l'intervalle [0,1] appelée échelle possibiliste. Cette section rappelle les éléments de base de cette théorie à savoir les distributions de possibilité, les mesures de nécessité et de possibilité et les réseaux possibilistes (voir les sections 2.2.1, 2.2.2 et 2.2.3 respectivement). Pour plus de détail voir [Dubois et Prade, 1987 ; 1994 ; 1998 ; 2006 ; 2009]. Cette théorie a été utilisée comme méthode de classification (voir section 2.2.4) et comme modèle d'appariement dans les SRI (voir section 2.2.5). Ces traitements de base vont nous servir dans plusieurs phases de notre processus de recherche d'information possibiliste.

Notre choix est justifié par les résultats obtenus dans des recherches récentes qui ont appliqué la théorie des possibilités. Par exemple, [Brini et al., 2004] ont développé le premier SRI possibiliste et démontré ses performances par rapport aux autres modèles de RI. Ce modèle a été ensuite développé par d'autres chercheurs tels que [Elayeb, 2009 ; Bounhas et al., 2010, 2011ab, Elayeb et al., 2011, 2015ab]. La théorie des possibilités permet aussi de pallier aux problèmes d'imprécision, d'incertitude et de manque de données dans les attributs des instances lors de la classification. Par exemple, [Jenhani et al., 2008 ; Haouari et al., 2009 ; Bounhas et al., 2013 ; 2014] ont développé un classifieur possibilistes qui tient compte de ces phénomènes. Par rapport à notre problématique, l'évaluation de la qualité ou de la fiabilité de l'information est souvent modélisée comme un problème de classification. Par exemple, [Stvilia et al., 2007] ont utilisé l'algorithme C4.5 basé sur les arbres de décision. Cette tâche nécessite des données exhaustives sur les sources d'information. La collecte de telles informations n'est pas toujours évidente. Il devient donc nécessaire de proposer des modèles qui permettent de classifier même en cas de données manquantes, imprécises ou incertaines.

2.2.1. Distribution de possibilité

La théorie des possibilités est basée sur les distributions de possibilité. Étant donné un univers de discours $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$, un concept fondamental dénoté par π correspond à une fonction qui associe à chaque élément ω_i une valeur dans un ensemble limité et linéairement ordonné (L, <). Cette valeur est appelée degré de possibilité et encode les connaissances du monde réel. Par convention, $\pi(\omega_i) = 1$ signifie qu'il est parfaitement possible que ω_i soit du monde réel. Alors que $\pi(\omega_i) = 0$ signifie que ω_i est impossible. La flexibilité est modélisée en permettant de donner un degré dans l'intervalle]0, 1[. Dans la théorie des possibilités, les cas extrêmes sont modélisés par :

- *Connaissance complète :* $\exists \omega_i \in \Omega, \pi(\omega_i) = 1 \; et \; \forall \omega_j \neq \omega_i, \pi(\omega_j) = 0$
- *Ignorance partielle :* $\forall \omega_i \in A \subseteq \Omega, \pi(\omega_i) = 1, \forall \omega_i \notin A, \pi(\omega_i) = 0$
- *Ignorance totale :* $\forall \omega_i \in \Omega, \pi(\omega_i) = 1$

2.2.2. Les mesures de possibilité et de nécessité

Une distribution de possibilité π sur Ω permet d'évaluer les événements en termes de leur plausibilité et leur certitude en utilisant deux mesures duales appelées respectivement *Possibilité* et *Nécessité*. Étant donnée une distribution de possibilité π sur un univers de discours Ω, les valeurs de possibilité et de nécessité évaluent chaque événement $A \subseteq 2^{\Omega}$ comme suit :

- La possibilité : $\Pi(A) = max_{\omega \in A} \pi(\omega)$
- La nécessité : $N(A) = min_{\omega \notin A}(1 - \pi(\omega)) = 1 - \Pi(\bar{A})$

$\Pi(A)$ évalue à quel niveau A est consistant avec nos connaissances représentées par π, alors que $N(A)$ évalue à quel degré A est certain selon nos connaissances. La taille du gap entre $N(A)$ et $\Pi(A)$ évalue le taux d'ignorance sur A [Dubois et Prade, 1987].

2.2.3. Les réseaux possibilistes

Un graphe possibiliste [Benferhat et al., 1999 ; 2002] orienté sur un ensemble de variables V est caractérisé par une composante *qualitative* et une composante *numérique* (quantitative). La première est un graphe acyclique orienté. La structure du graphe représente l'ensemble des variables ainsi que l'ensemble des relations d'indépendance. La seconde composante quantifie les liens du graphe en utilisant des distributions de possibilité conditionnelles de chaque nœud dans le contexte de ses parents. Ces distributions de possibilité doivent vérifier la contrainte de normalisation. Pour chaque variable V :

- Si V est un nœud racine et $Dom(V)$ le domaine de V, la possibilité *a priori* de V doit satisfaire : $max_{v \in Dom(V)} \Pi(v) = 1$
- Si V n'est pas un nœud racine, la distribution conditionnelle de V dans le contexte de ses parents doit satisfaire : $max_{v \in Dom(V)} \Pi(v|u_V) = 1; \; u_V \in Dom(U_V)$

Avec: $Dom(V)$: le domaine de V ; U_V: l'ensemble des parents de V ; $Dom(U_V)$: le domaine des parents de V.

Dans le chapitre 1, nous proposons une approche possibiliste de désambiguïsation sémantique basée sur un réseau possibiliste. Ce dernier relie les sens des mots (S_i) aux mots d'une phrase polysémique $ph = (t_1, t_2, \ldots, t_T)$, qui représentent son contexte. Dans ce cas: $v_i = t_i$; $u_V = S_i$; $Dom(V) = \{t_1, t_2, \ldots, t_T\}$; et $Dom(U_V) = \{S_1, S_2, \ldots, S_N\}$.

Les réseaux et distribution possibilistes peuvent être interprétés d'une manière quantitative en utilisant l'opérateur *produit* ou d'une manière qualitative en utilisant l'opérateur *minimum*.

2.2.3.1. Les réseaux possibilistes à base de produit

Un réseau possibiliste basé sur le produit est un réseau possibiliste où les possibilités conditionnelles sont obtenues par l'opérateur produit. La distribution de possibilité de ces réseaux, notée par π_p, est obtenue par la règle de chaînage :

$$\pi_p(V_1, \dots, V_N) = \prod_{i=1}^{N} \Pi(V_i | U_{V_i})$$

2.2.3.2. Les réseaux possibilistes à base de minimum

L'opérateur minimum (*MIN*) est utilisé pour obtenir les possibilités conditionnelles dans un tel réseau. La formule suivante permet de calculer une distribution de possibilité π_M dans un réseau à base de minimum:

$$\pi_M(V_1, \dots, V_N) = \underset{i=1,\dots,N}{MIN} \Pi(V_i | U_{V_i})$$

2.2.4. Les classifieurs possibilistes

Plusieurs méthodes de classification permettent de prédire la classe d'une instance en fonction de ses attributs. Les principales méthodes basées sur l'apprentissage automatique sont les arbres de décisions, les réseaux de neurones, les K-voisins les plus proches et les réseaux Bayésiens. Ces derniers constituent les classifieurs les plus efficaces. Cependant, les réseaux Bayésiens naïfs, comme les approches probabilistes, affrontent des problèmes si les données sont imparfaites.

Plusieurs théories de l'incertitude ont été proposées pour traiter les données incertaines et imprécises. Nous citons la théorie de l'évidence [Shafer, 1976], la théorie des ensembles flous [Zadeh, 1965] et la théorie des possibilités [Dubois et Prade, 1987]. L'utilisation des réseaux possibilistes est encouragée par sa simplicité et la performance dans le traitement des données imparfaites [Haouari et al., 2009]. Dans ce cas, le graphe relie les attributs A_i aux classes possibles C_i. Les poids des arcs sont estimés dans l'étape d'apprentissage.

Par ailleurs, les applications de fouille de données et d'apprentissage automatique ont largement profité des techniques de classification. Durant la phase d'apprentissage, le classifieur est entraîné sur un ensemble d'apprentissage formé par un ensemble d'instances. Chaque instance est formée d'un ensemble d'attributs dont la classe est connue a priori. Durant la phase du test, le classifieur est appelé à prédire la valeur de la classe des nouvelles instances de test en utilisant les valeurs connues de leurs attributs. Les classifieurs possibilistes ont été introduits par [Jenhani et al., 2008 ; Haouari et al., 2009], puis développés par [Bounhas et al., 2013, 2014]. Ces classifieurs ont été fondé sur l'idée du classifieur Bayésien naïf de [Friedman et al., 1997]. Ce genre de classifieurs possibilistes ont résolu le problème des données imparfaites. Autrement dit, ils ont été efficaces dans les cas où les instances de l'ensemble d'apprentissages contiennent des attributs et/ou des classes imparfaites (possédant plusieurs valeurs possibles). En fait, l'imperfection est concrétisée par un ensemble normalisé aléatoire de distributions des possibilités définies sur des valeurs possibles d'attributs et de classes.

Nous avons profité des classifieurs possibilistes afin d'identifier les valeurs exactes des attributs morphologiques des mots arabes ambigus. En effet, le processus de désambiguïsation morphologique des textes arabes requière une phase d'apprentissage et une phase de test. Les travaux de la littérature dans ce domaine ont confirmé que les mots arabes voyellés sont moins ambigus que ceux non-voyellés. En se basant sur cette hypothèse, nous proposons de résoudre le verrou de l'ambiguïté morphologique des textes arabes via des classifieurs possibilistes qui apprennent à partir des textes voyellés et testent sur des textes non-voyellés. En effet, nous avons utilisé des analyseurs morphologiques arabes [Ayed, 2017] afin de construire les instances de

l'ensemble d'apprentissage. Toutefois, ces analyseurs peuvent souffrir des deux problèmes d'imprécision et d'incertitude en cas où plusieurs valeurs possibles peuvent être obtenu pour le même attribut morphologique d'un mot arabe voyellé. Par exemple, le mot arabe voyellé "أَشْرَفُ" pourra se référer à l'adjectif "qui a plus d'honneur" ou au nom propre "Ashraf". En conséquence, le classifieur apprenant à partir de ce genre de mots arabes voyellés génèrera des instances imprécises. Tandis que, la théorie des possibilités tient compte de l'imprécision et de l'incertitude des instances de classification afin de pallier aux verrous posés par les classifieurs existants. Ces derniers exigent des corpus annotés et/ou des connaissances collectées manuellement. Notre objectif s'étend jusqu'à l'automatisation totale de la tâche de désambiguïsation morphologique des textes arabes en profitant des atouts des classifieurs possibilistes.

La première étape, dans les méthodes de classification, consiste à élaborer des règles de classification à partir des connaissances disponibles (phase d'apprentissage). Par la suite, il sera crucial de tester les règles fournies sur d'autres connaissances (phase de test).

Phase d'apprentissage : Nous considérons les cas où l'ensemble d'apprentissage *Tr* est imparfait (i.e. pourra contenir des attributs imparfaits). Autrement dit, au lieu d'avoir une valeur d'attribut exacte, nous avons une distribution de possibilité sur un ensemble possible *S* de valeurs candidates appartenant au domaine de l'attribut. En fait, α est un facteur d'incertitude dans [0, 1] qui reflète le degré de confiance de l'expert que l'ensemble *S* contient la valeur réelle de l'attribut. Les éléments de l'ensemble *S* sont supposés équipossible d'être la valeur réel de l'attribut. En conséquence, Haouari et al. (2009) ont assigné à chacun d'eux une distribution de possibilité $\beta = 1/n$, où β exprime l'imprécision relative à l'attribut, et *n* désigne le cardinal de *S*. De plus, les auteurs ont pondéré le moyen possibiliste pour prendre en compte les facteurs β et α.

Le moyen possibiliste $\pi_{Pm}(a_j|c_i)$ d'une valeur d'attribut a_j étant donnée la classe c_i est donnée par la formule suivante [Haouari et al., 2009]:

$$\pi_{Pm}(a_j|c_i) = mean_{Tr(a_j,c_i)}(\beta_j * \alpha_j * \pi(I_k|c_i))$$

Avec : - $\pi(I_k|c_i)$ est la distribution de possibilité des valeurs d'attribut (I_k tel que $a_j \in I_k$) de la $k^{ième}$ instance étant donné la classe c_i.
- $mean_{Tr(a_j,c_i)}$ indique la moyenne sur l'ensemble d'apprentissage *Tr*.

Phase de test : Dans le cas d'un algorithme de propagation à base de *produit*, une instance donnée est affectée à la classe la plus plausible *c**:

$$c^* = \underset{c_i}{\text{argmax}}(\prod_{j=1}^{m} \pi_{Pm}(a_j|c_i))$$

Tandis que, dans le cas d'un algorithme de propagation à base de *minimum*, nous avons :

$$c^* = \underset{c_i}{\text{argmax}}(MIN_{j=1}^{m} \pi_{Pm}(a_j|c_i))$$

En fait, nous avons profité de ces deux classifieurs possibilistes, à base de *produit* et à base de *minimum*, afin d'identifier la classe de fiabilité d'une chaîne de narration arabe (chapitre 5). Par contre, le classifieur possibiliste à base de *produit* nous a servi dans notre processus de désambiguïsation morphologique des textes arabes (chapitre 2). En fait, ce dernier utilise uniquement la mesure de possibilité qui n'évalue pas le pouvoir discriminant des valeurs d'un attribut. Néanmoins, certaines valeurs d'un attribut donné pourront avoir plus d'effet dans l'identification de la classe correcte. De plus, la théorie des possibilités est capable de modéliser cet effet à travers la mesure de nécessité.

Ainsi, nous avons proposé l'implication de la mesure de nécessité dans l'extension de notre classifieur possibiliste de base [Ayed et al., 2012ab] inspiré des travaux de [Haouari et al., 2009].

2.2.5. Le modèle possibiliste de RI

Les réseaux possibilistes sont employés comme un modèle d'appariement dans les SRI [Boughanem et al., 2009 ; Elayeb et al., 2009]. Dans ce cas, ils relient les termes (t_i) aux documents (D_j). La relation entre les termes d'une requête et les documents sont quantifiées par les mesures de possibilité et de nécessité. Le processus de recherche retourne les documents plausiblement ou nécessairement pertinents à un utilisateur. Un SRI possibiliste est capable de générer des propositions du genre:

- Il est plausible à un certain degré que le document d_i constitue une bonne réponse pour la requête Q,
- Il est nécessaire ou certain à un degré donné que le document d_i soit pertinent pour la requête Q,
- Le document d_i est plus pertinent que d_j ou un ensemble $\{d_i, d_j\}$ répond mieux à la requête qu'un autre ensemble $\{d_k, d_l\}$.

Le modèle possibiliste suppose qu'il est difficile de traduire la notion de pertinence ayant un caractère vague et imprécis avec une seule mesure de probabilité. En fait, la théorie des possibilités suggère des mesures duales modélisant l'incertitude liée à l'information d'une manière flexible et différente de la théorie des probabilités. La mesure de possibilité tend à travers le premier type de proposition à éliminer les documents non-pertinents. Dans la deuxième, la mesure de nécessité renforce notre croyance envers les documents pertinents. Cela permet d'organiser les documents selon un ordre de pertinence exprimé par la troisième proposition. Par contre, la théorie des probabilités permet uniquement de mesurer la certitude d'un événement et de son contraire. En outre, les modèles probabiliste de RI ne tiennent pas compte des termes de la requête qui sont absents dans les documents lors du calcul des scores de pertinence. Face à ces limites restrictives, [Prade et Testemale, 1987] ont proposé l'usage de la théorie des possibilités en RI. Brini et Boughanem (2003) ont présenté une première application de ce modèle qui a été ensuite développé par [Boughanem et al., 2009 ; Elayeb, 2009] afin de tenir compte de la structure des documents et des préférences entre les termes de la requête.

D'autre part, le modèle possibiliste de RI considère la requête comme l'information la plus sûre disponible pour le SRI. De plus, l'absence d'un terme de la requête d'un document donné pénalise le score de pertinence de ce document en fonction de l'importance du terme dans la collection. Ainsi, l'absence d'un terme de la requête dans la représentation des documents est une information à ne pas ignorer lors du calcul de la pertinence [Boughanem et al., 2009]. En fait, notre SRI possibiliste est à base d'une propagation qui considère tous les termes de la requête qu'ils soient présents ou absents dans les documents.

Ce modèle permet de calculer un score de ressemblance entre une requête et un document comme suit : La requête Q est composée par des termes qui représentent des contraintes. Nous considérons le cas général où ces termes sont pondérés (par exemple selon les préférences de l'utilisateur) :

$Q = \{(t_1, \omega_1); (t_2, \omega_2); \dots; (t_m, \omega_m)\}$; Où ω_i représente le poids du terme t_i.

Le degré de pertinence possibiliste (*DPP*) d'un document D_j étant donné la requête Q est calculé par les deux mesures de possibilité (Π) et de nécessité (N). Le modèle de base est inspiré de [Boughanem et al., 2009 ; Elayeb et al., 2009] où $\Pi(D_j|Q)$ est proportionnelle à :

$$\Pi'(D_j|Q) = \pi(t_1|D_j) * \omega_1 * \dots * \pi(t_m|D_j) * \omega_m$$

Les distributions de possibilité $\pi(t_i|D_j)$ sont estimées par les fréquences de chaque terme t_i dans chaque document D_j. Nous avons donc :

$\Pi'(D_j|Q) = nFreqt_{1j} * \omega_1 * \ldots * nFreqt_{mj} * \omega_m$

Avec : - $nFreqt_{ij} = \frac{Freqt_{ij}}{maxFreqt_{ij}}$: La fréquence normalisée du terme t_i dans le document D_j.

- $Freqt_{ij} = \frac{le\ nombre\ d'occurrences\ du\ terme\ t_i\ dans\ le\ document\ D_j}{nombre\ de\ termes\ dans\ le\ document\ D_j}$

- $maxFreqt_{ij}$: La fréquence maximale.

La nécessité de D_j pour la requête Q notée $N(D_j|Q)$ est calculée comme suit :

$N(D_j|Q) = 1 - \Pi(\neg D_j|Q)$

Où : $\Pi(\neg D_j|Q) = \frac{\Pi(Q|\neg D_j) * \Pi(\neg D_j)}{\Pi(Q)}$

De la même manière $\Pi(\neg D_j|Q)$ est proportionnelle à :

$\Pi'(\neg D_j|Q) = \pi(t_1|\neg D_j) * \ldots * \pi(t_m|\neg D_j)$

Ce qui peut être exprimé par :

$\Pi'(\neg D_j|Q) = \left(1 - \frac{\phi_{1j}}{\omega_1}\right) * \ldots * \left(1 - \frac{\phi_{mj}}{\omega_m}\right)$

Avec : $\phi_{ij} = Log_{10}\left(\frac{|D|}{nD_i}\right) * (nFreqt_{ij})$

Où : - $|D|$ est le nombre de documents de la collection.

- nD_i est le nombre de documents de la collection contenant t_i (i.e. $Freq_{ij} > 0$).

Le degré de pertinence possibiliste (*DPP*) de D_j est souvent calculé comme la somme des deux mesures de possibilité et de nécessité :

$DPP(D_j|Q) = \Pi(D_j|Q) + N(D_j|Q)$

Les documents retournés par le SRI possibiliste sont triés selon un ordre décroissant de leurs degrés des pertinences possibilistes.

Parmi les objectifs ultimes de notre travail est d'appliquer ce modèle possibiliste de RI, non seulement en utilisant un ensemble de requêtes et une collection de documents comme dans le cas de travaux de [Brini et al., 2003, 2004 ; Boughanem et al., 2009 ; Elayeb, 2009], mais aussi en désambiguïsation sémantique des termes d'une requête ambiguë (cf. chapitre 1), en expansion sémantique des termes d'une requête à reformuler (cf. chapitre 3) et en reconnaissance des identités des personnes d'une chaîne de narrateurs arabes afin d'identifier le degré de fiabilité de cette chaîne (cf. chapitre 5).

Généralement, les modèles existants pour la désambiguïsation et l'expansion de requêtes ainsi que ceux dédiés à l'identification du degré de fiabilité des documents pertinents retournés sont dans la plupart des cas des modèles d'apprentissage et d'appariement probabilistes. Dans ce contexte, les données traitées sont souvent manquantes, incertaines et imprécises pour lesquelles la théorie de probabilité s'avère limitée. Afin de pallier à nos problèmes ci-dessus décrits, nous avons profité de la théorie des possibilités qui est naturellement conçue pour ce type d'applications. Elle permet d'exprimer l'ignorance et de tenir compte de l'imprécision et de l'incertitude au même temps. En effet:

D'abord, dans le cas de la *désambiguïsation sémantique de requêtes*, le cas d'imprécision provient du fait que pour chaque terme ambigu de la requête correspond plusieurs sens possibles (existant dans le dictionnaire). C'est pour cette raison que nous avons évalué la pertinence du sens d'un mot polysémique vis-à-vis de son contexte, donné sous forme d'une phrase, en utilisant deux types de pertinences: la pertinence plausible (possible) et la pertinence nécessaire. La pertinence possible permet de rejeter les sens non-pertinents, alors que la pertinence nécessaire permet de renforcer la pertinence des sens de mots restants, qui n'ont pas été rejetés par la possibilité. D'autre part, dans le cas de *la désambiguïsation morphologique* des textes arabe, la désambiguïsation consiste à identifier la valeur exacte d'un attribut morphologique parmi celles proposées par l'analyseur. Le cas d'imprécision est vécu lorsque l'analyseur morphologique fournit plus qu'une seule valeur d'attributs morphologiques. Les classifieurs possibilistes que nous avons proposé visent à résoudre le problème d'ambigüité morphologique des textes arabes. La tâche de désambiguïsation morphologique consiste à accorder à un mot ambigu les valeurs d'attributs morphologiques appropriées. Les résultats d'analyse morphologique, donnés par les mots voyellés, sont généralement moins ambigus que ceux donnés par les mots non-voyellés. Ainsi, nous avons proposé d'apprendre à partir des textes voyellés et de tester sur des textes non-voyellés.

Ensuite, dans le cas d'*expansion sémantique de requêtes*, le cas d'imprécision vient du fait que plusieurs mots possibles peuvent être injectés dans la requête de l'utilisateur afin d'élargir son contexte. C'est pour cette raison que nous avons modélisé le problème d'expansion sémantique de requêtes via une double mesure: la mesure de possibilité est utile pour éliminer les mots non sémantiquement proches de termes de la requête originelle, alors que la mesure de la nécessité renforcera la pertinence des mots restants. Le processus d'expansion retourne à l'utilisateur les mots plausiblement ou nécessairement pertinents.

Enfin, dans le cas de l'*identification du degré de fiabilité des documents pertinents retournés*, l'étape de reconnaissance des identités des narrateurs est modélisée en tant qu'un SRI possibiliste. Les noms extraits d'une chaîne sont les requêtes alors que les biographies des personnes (stockées dans une base) sont les documents recherchés. Comme tout SRI possibiliste, nous avons modélisé les liens de dépendance entre les éléments de la requête et les personnes à travers un réseau possibiliste et quantifié ces liens par les deux mesures de possibilité et de nécessité. Les personnes retrouvées sont celles qui sont possiblement ou nécessairement pertinentes étant donné les noms de la chaîne. D'autre part, nous avons identifié la classe de la fiabilité d'une chaîne via des classifieurs possibilistes, en analogie avec notre motivation, décrite ci-dessus, pour l'utilisation de ce genre de classifieur.

2.3. Les métriques d'évaluation des SRI

Un SRI est qualifié performant s'il fournit un ensemble de documents le plus proche possible des réponses idéales que l'utilisateur souhaite obtenir pour satisfaire sa requête. Autrement dit, plus ces deux ensembles sont proches, mieux c'est pour le SRI. Cette comparaison de réponses du SRI aux réponses idéales est assurée via les métriques suivantes :

$$Rappel = \frac{Nombre\ de\ documents\ pertinents\ séléctionnés}{Nombre\ total\ de\ documents\ pertinents}$$

$$Précision = \frac{Nombre\ de\ documents\ pertinents\ séléctionnés}{Nombre\ total\ de\ documents\ séléctionnés}$$

$$F - mesure = \frac{(1+\beta^2) \times précision \times rappel}{(\beta^2 \times précision) + rappel}$$

Le paramètre β permet de pondérer la précision ou le rappel, il est égal généralement à la valeur 1.

Nous calculons aussi la Précision Moyenne, notée par *Pr.Moy* (en anglais *Mean Average Precision* (*MAP*)) et la Précision Exacte, notée par *Pr.Ex* ou *R-Précision* (en anglais *R-Precision*) selon les formules suivantes :

$$Pr.Moy = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{|rel_j|}\sum_{r=1}^{N_j} P(r) \times isRel(r)$$

Avec : $P(r) = \frac{Nombre\ de\ documents\ pertinents\ trouvés\ au\ rang\ r\ ou\ moins}{r}$

Où : - $|Q|$: le nombre total de requêtes.
- $|rel_j|$: le nombre de documents pertinents pour la requête *j* dans toute la collection.
- N_j : le nombre de documents retournés par la requête *j*.
- *isRel(r)* : une fonction binaire indiquant si le résultat au rang *r* est un document pertinent (1) ou pas (0).

$$Pr.Ex = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Précision(\{D_{kj}\})$$

Avec : $Précision(\{D_{kj}\})$ correspond à la précision des *k*-premiers résultats de la requête *j*, où *k* désigne le nombre de documents pertinents associés à la requête *j* dans la collection.

2.4. Le besoin d'évaluation de la fiabilité de l'information

Les besoins des utilisateurs des systèmes de recherche d'information (SRI) deviennent de plus en plus diversifiés et exigeants en conséquence des changements vécus par le secteur de la demande d'information. En fait, les évaluations de ces SRI dépendent des besoins exprimés par ses utilisateurs. La notion de pertinence de l'information intègre de plus en plus de dimensions couvrant tous les critères de la qualité de l'information (QI). D'où la naissance de la pertinence thématique où l'utilisateur s'intéresse à un thème ou à un détail spécifique. Par ailleurs, la fiabilité de l'information est l'un des critères les plus importants dans tout système d'information. Ce critère n'est pas récent vu son existence depuis longtemps dans diverses disciplines telles que les sciences de l'histoire et de la religion, les sciences de l'informatique (évaluation des pages Web et des articles de journaux, l'analyse des biographies, etc.). Notons ici que les Hadiths sont les seuls textes arabes ayant été sujet d'étude dans ce domaine. Dans [Bounhas, 2012] nous avons résumé et synthétisé notre étude sur les méthodologies et les applications dans ce domaine. Nous avons détaillé la notion de pertinence et ses dimensions.

En effet, ces exemples d'applications de la fiabilité montrent l'impact de la fiabilité dans l'évaluation de l'information. A partir de ces applications et des travaux dans le domaine de la qualité de l'information [Naumann et Rolker, 2000 ; Knight et Burn, 2005 ; Stvilia et al., 2007 ; Stvilia, 2008] le processus d'évaluation de la fiabilité se résume dans les cinq étapes suivantes:

1. Pour chaque domaine et application donnés, nous sélectionnons dans la littérature les critères d'évaluation appropriés.
2. Identifier, pour chaque critère, la méthode d'évaluation à utiliser. Dans ce cadre, nous nous basons sur les travaux de Naumann et Rolker (2000) qui ont spécifié certaines techniques basées sur l'analyse du contenu et le jugement par les experts. Par ailleurs, un agent peut décider la réputation d'un autre en se basant sur les transactions qu'il a eu avec lui. En effet, la réputation d'un agent augmente proportionnellement avec l'augmentation du nombre des transactions réussies effectuées avec lui. Notre décision se base parfois sur des informations fournies par un témoin qui a déjà évalué l'agent en question. Enfin, la réputation des agents est déterminée en fonction des relations et des rôles joués par ces derniers. Régulièrement,

nous avons tendance à croire les agents avec qui nous sommes socialement proches ou qui ont des rôles spécifiques (par exemple, les agents qui appartiennent à une autorité gouvernementale).

3. Définir des scores/métriques pour chaque critère. Selon Naumann et Rolker (2000), le score doit être précis, pratique et connu par l'utilisateur. En plus, son calcul ne doit pas être ni complexe ni coûteux.
4. Chercher/définir une méthode ou une formule d'agrégation des critères dans une seule mesure. Dans cette étape, nous pouvons appliquer les méthodes d'agrégation proposées par [Stvilia et al., 2007].
5. Développer des mécanismes de visualisation adaptés à l'utilisateur qui tiennent compte en particulier de son expertise. En fait, l'évaluation de la fiabilité ne se limite pas à un calcul d'indices mais s'étend sur une stimulation du processus cognitif de l'utilisateur par des mécanismes de visualisation et d'interaction.

Par ailleurs, les approches existantes peuvent être classées selon l'architecture ou la localisation des informations sur la réputation [Noorian et Ulieru, 2010]. Dans une première approche, un seul agent possède ces informations (par exemple Yahoo Internet Life). Cette centralisation est examinée du fait qu'elle contredit le caractère dynamique et ouvert du Web. Dans un système ouvert et dynamique voire à large échelle, il est difficile d'établir un accord sur un seul agent de recommandation. La deuxième approche considère que tout agent peut jouer le rôle de recommandation. C'est le cas du SRI multicritère de [Da Costa Pereira et Pasi, 2007]. Le processus d'établissement de la confiance pourra être ralenti à cause de l'absence d'une autorité de contrôle et la décentralisation complète. Toutefois, des agents non crédibles peuvent publier des jugements non fiables sur d'autres en profitant de la sensibilité de ces systèmes aux intrusions. La dernière approche hybride a profité des avantages des deux approches précédentes en chargeant un ensemble déterminé d'agents pour fournir des recommandations.

Nous commençons par la première étape en identifiant les dimensions de la fiabilité de l'information généralement utilisées. Cependant, les autres étapes nécessitent une étude plus détaillée du domaine et de l'application visés. En effet, nous distinguons trois dimensions comme suit [Bounhas et al., 2010] :

- *Autorité*: définie comme l'ensemble des indicateurs qui prouvent (ou qui peuvent être exploités pour vérifier) la crédibilité des acteurs. Par exemple, pour évaluer la fiabilité d'un site, nous devons vérifier l'existence d'information tels que les noms des auteurs, leurs affiliations, les textes de copyright, etc.

- *Objectivité*: définie comme le degré auquel l'information n'est pas biaisée, préjugée ou partielle [Knight et Burn, 2005]. L'objectivité d'une information est analysée en identifiant sa nature (un fait ou une opinion) et son objectif (travail de recherche, discours politique, publicité commerciale, etc.).

- *Vérifiabilité*: elle interprète l'existence d'éléments intrinsèques ou extrinsèques qui aident à vérifier la fiabilité en termes d'autorité et d'objectivité.

Par ailleurs, Chen et al., (2007) ont ajouté à ses trois dimensions le moyen de transmission. Par exemple, ils ont confirmé que la transmission orale est moins fiable que le format papier. Ils ont affirmé que le degré de fiabilité de l'information est proportionnel à sa rationalité. Divers travaux ont mentionné l'importance du flux pour la fiabilité dans le domaine du e-learning [Romney et Romney,

2005] et en médicine [Parker et al., 2006]. D'autres études ont utilisé, en plus, des critères liés au processus d'édition des documents [Chevalier et al., 2010].

Les méthodologies et les applications dans le domaine de l'évaluation de la fiabilité sont considérées comme un critère important de la pertinence de l'information. En fait, nous avons pu mixer les méthodologies classiques des sciences de l'histoire et de Hadith avec les développements modernes dans les sciences de l'informatique, ce qui a donné naissance à une démarche générique, malgré la diversité des domaines d'application. Par ailleurs, nous pouvons dire que l'évaluation de la fiabilité de l'information est un processus complémentaire de toutes les étapes de la démarche de RI, en partant de la désambiguïsation et l'expansion de requêtes jusqu'à la vérification de la fiabilité des documents retournés. De plus, nous signalons la sensibilité de la précision des métriques, des indices de fiabilité et de l'ergonomie de leur affichage. En effet, il faut parcourir toutes les étapes d'étude et d'analyse qui permettent d'identifier et de structurer les informations nécessaires au calcul de ses indices. Ces étapes sont aussi nécessaires pour évaluer les autres critères de pertinence tels que la pertinence thématique.

3. Contributions de recherche

Nos contributions de recherche seront détaillées dans les prochains chapitres et peuvent se résumer sous forme des **deux axes de contributions majeurs** qui touchent d'une part la requête de l'utilisateur, et d'autre part, l'ensemble de documents pertinents retournés par le SRI :

1. Le **premier axe** de contribution est consacré à la **désambiguïsation et l'expansion de requêtes**. Cet axe est lui-même pourra être subdivisé en quatre sous-contributions majeures : La première est dédiée à la désambiguïsation sémantique de textes français. La seconde s'intéresse à la désambiguïsation morphologique de textes arabes. La troisième se focalise sur l'expansion sémantique de textes français. Enfin, la quatrième évalue l'impact de la désambiguïsation sémantique de requêtes sur leurs expansions. Ces contributions sont résumées comme suit :

    a. **Nouvelles approches de désambiguïsation sémantique de requêtes :** Nous avons proposé et comparé deux nouvelles approches une possibiliste et une probabiliste pour la désambiguïsation sémantique monolingue. Ces deux approches exploitent une nouvelle ressource linguistique externe que nous appelons "dictionnaire sémantique de contextes (DSC)", obtenu suite à une combinaison d'un dictionnaire traditionnel avec un corpus étiqueté. Ce DSC permet de soutenir l'apprentissage automatique dans notre plate-forme de désambiguïsation sémantique. Ces deux approches sont évaluées dans la désambiguïsation des textes français. D'abord, notre première approche possibiliste exploite les réseaux possibilistes afin de profiter d'une double mesure de pertinence. D'une part, la *pertinence possible* permet d'écarter les sens non-pertinents d'un mot à désambiguïser. D'autre part, la *pertinence nécessaire* permet de renforcer la pertinence des sens restants non-éliminés par la possibilité. Ensuite, notre approche probabiliste utilise et étend une *distance sémantique probabiliste* existante afin de calculer un score de similarités entre les mots en exploitant un graphe sémantique d'un dictionnaire traditionnel et le DSC. Enfin, ces deux approches ont été évaluées et comparées aux systèmes existants de désambiguïsation monolingue français en utilisant la collection de test ROMANSEVAL. Nos résultats possibilistes ont prouvé des améliorations encourageantes en termes de taux de désambiguïsation de mots français [Ben Khiroun et al., 2012 ; Elayeb et al., 2015a]. Ces résultats confirment la contribution de la théorie des possibilités comme un outil de traitement de l'imprécision dans les systèmes de désambiguïsation sémantique.

b. **Nouvelles approches de classification possibiliste pour la désambiguïsation morphologique des textes arabes :** Nous avons examiné la tâche de désambiguïsation des attributs morphologiques des textes arabes non-voyellés comme une tâche de classification. Pour ce faire, nous avons proposé un modèle de classifieur possibiliste tenant compte des données imprécises dans les deux phases d'apprentissage et de test. D'abord, nous avons proposé et testé trois classifieurs possibilistes : un premier à base de mesure de possibilité, un second à base de la nécessité et un troisième à base de la somme des deux mesures. Puis, nous avons affecté des pondérations aux attributs dans l'ensemble d'apprentissage afin de discriminer l'effet de chacun d'entre eux. Ensuite, nous avons étudié l'impact de la possibilité lexicale dans le modèle de repondération. Enfin, nous avons expérimenté et comparé nos trois premiers classifieurs (sans repondération et impliquant la possibilité lexicale) aux classifieurs non-possibilistes existants en utilisant deux différents types de corpus (classique et moderne) à savoir le corpus du Hadith[6] et le Treebank arabe. En outre, nous avons testé aussi à quel degré cette approche discriminative améliorait le taux de désambiguïsation et extrairait les relations des dépendances entre les attributs. Nos résultats [Bounhas et al., 2015bc] ont montré l'efficacité de notre classifieur possibiliste discriminatif à base de la somme de possibilité et de nécessité et impliquant un modèle de repondération et une possibilité lexicale par rapport aux autres classifieurs en termes de taux de désambiguïsation de 14 attributs morphologiques [Ayed et al., 2012ab, 2014ab].

c. **Nouvelles approches d'expansion sémantique de requêtes :** Nous nous concentrons sur les techniques d'expansion sémantique de requêtes afin de contribuer à l'amélioration de l'efficacité de la recherche. Ces techniques ont exploité un dictionnaire comme ressource linguistique externe. Dans notre cas, nous avons exploité le dictionnaire français "Le Grand Robert". D'abord, nous avons modélisé la structure de ce dictionnaire par le biais d'un Réseau Petits-Mondes Hiérarchiques (RPMH) afin d'énumérer les circuits entre les nœuds du graphe RPMH utiles pour le calcul d'un score de proximité sémantique entre les termes. Cette approche a été initiée dans [Elayeb et al., 2009] et a été limitée uniquement à un RPMH des verbes, alors que nous étendons ici la couverture en tenant compte de toutes les catégories grammaticales tels que les adjectifs, les verbes, les noms et les adverbes. Puis, nous avons proposé une seconde approche d'expansion sémantique de requêtes à base des réseaux possibilistes afin de profiter d'une double mesure de proximité sémantique entre les articles d'un dictionnaire et les termes de la requête à reformuler : la *pertinence possible* permet de rejeter les articles du dictionnaire non-sémantiquement proches des termes de la requête, et la *pertinence nécessaire* permettant de renforcer la pertinence des articles restants non-éliminés par la possibilité. Ensuite, nous avons combiné ces deux approches en utilisant deux nouveaux scores d'agrégation dont un à base de *somme* et l'autre à base de *produit*. Nous avons bénéficié aussi de notre technique existante [Elayeb et al., 2009] de repondération des termes de la requête dans le modèle d'appariement possibiliste afin d'améliorer davantage le processus d'expansion. Nous avons testé nos approches via la collection de test "LeMonde94", faisant partie du standard CLEF-2003. Nos résultats [Ben Khiroun et al., 2011 ; Elayeb et al., 2011] ont

[6] Le texte de Hadith est tout ce qui est rapporté du Prophète Mohamed (Paix et Bénédiction soient Sur Lui) comme paroles, actions, acquiescements, ou caractéristiques (physiques, traits de caractères, etc.). Mais contrairement au Coran, la question de l'authenticité d'un Hadith se pose avant qu'il ne soit une source de législation.

prouvé la contribution de ces nouvelles approches hybrides concrétisées par le nombre de requêtes améliorées.

d. **Contribution de la désambiguïsation sémantique de requêtes à leurs expansions :** Notre objectif s'articule ici sur l'étude de l'impact de la désambiguïsation sémantique de requêtes sur leurs expansions dans la cadre d'un SRI possibiliste. Pour cela, nous avons exploité des graphes de cooccurrence modélisés par des réseaux possibilistes. En conséquence, notre modèle de jugement de pertinence a bénéficié de la double mesure de pertinence fournie par la théorie des possibilités. Afin de profiter des avantages de nos deux approches de désambiguïsation et d'expansion sémantique de requêtes, nous avons reformulé des requêtes de la collection CLEF-2003 contenant des termes ambigus existant dans le standard ROMANSEVAL. Nous avons commencé par désambiguïser ces requêtes avant leurs expansions. Nos résultats [Ben Khiroun et al., 2014a ; Elayeb et al., 2015b] ont confirmé l'impact positif de la désambiguïsation de requêtes sur leurs expansions en se basant sur les indicateurs standards de rappel/précision. En outre, l'approche possibiliste a dépassé les performances d'une approche probabiliste à base de circuits.

2. Le **Second axe** de contribution s'intéresse à l'identification et l'évaluation du **degré de fiabilité** des documents pertinents retournés par le SRI. Dans cet axe nous nous sommes limités à la fiabilité des textes arabe comme suit :

   e. **Nouvelle approche possibiliste d'identification et d'évaluation de la fiabilité de l'information :** Nous avons exploité la méthodologie des sciences de Hadiths afin de proposer une architecture dédiée pour l'étude de la fiabilité des narrations arabes. Nous partons de l'idée que la fiabilité d'une narration dépend de la crédibilité de ses narrateurs. D'abord, nous avons proposé des grammaires utiles pour l'analyse des noms propres ainsi que les chaînes de narrateurs des narrations arabes. Ensuite, ces grammaires ont été exploitées par notre outil intelligent de reconnaissance de l'identité. Ce dernier a été modélisé comme un système de recherche d'information (SRI) possibiliste assurant l'appariement des noms propres existant dans les chaînes de narrateurs (requêtes) aux biographies des personnes correspondantes (documents). Enfin, les méta-données disponibles dans les biographies sont utilisées pour analyser les chaînes de narrateurs. Cette étape soutient l'utilisateur dans l'identification des origines de manque de fiabilité. En outre, nous considérons la tâche d'identification de la fiabilité d'une narration arabe comme un phénomène de classification possibiliste. Nous avons expérimenté notre approche sur trois domaines du corpus de Hadith à savoir le *mariage*, les *boissons* et la *purification*. Nos résultats [Bounhas et al., 2010 ; 2015a] enregistrés pour les entités nommées et la reconnaissance de l'identité ont confirmé l'efficacité de l'approche possibiliste pour les métriques d'évaluation rappel, précision et F-mesure.

4. Organisation du mémoire

Nous structurons la suite de ce mémoire en cinq chapitres synthétisants nos contributions. Chaque chapitre englobe une introduction, une synthèse des travaux de l'état de l'art ainsi que nos contributions afin de résoudre le verrou posé. Nous concluons à la fin de chaque chapitre par un bilan de nos apports et nous suggérons quelques perspectives de recherche.

Le chapitre 1 décrit nos deux approches possibiliste et probabiliste pour la désambiguïsation sémantique monolingue. Ces deux approches sont à base d'une nouvelle ressource linguistique externe à savoir le dictionnaire sémantique des contextes (DSC) bénéficiant à la fois des connaissances extraites à partir d'un dictionnaire traditionnel et des informations contextuelles/distributionnelles apprises à partir d'un corpus. Nous avons prouvé que l'utilisation de tel dictionnaire améliore davantage les résultats du calcul de sens d'un mot selon son contexte. Il permet aussi d'analyser les résultats de désambiguïsation en représentant des informations explicitement contextuelles. En conséquence, nos approches ont dépassé les performances des systèmes existants de désambiguïsation monolingue français.

Le chapitre 2 discute la contribution des nouvelles approches pour la désambiguïsation morphologique arabe basées sur la classification possibiliste. Notre objectif étant de profiter des textes arabes voyellés afin d'apprendre des dépendances morphologiques. Puis, nous avons testé plusieurs alternatives d'un classifieur possibiliste sur des textes arabes non-voyellés afin de prouver son efficacité dans le traitement des données imprécises dans les deux phases d'apprentissage et de test. Ensuite, ces classifieurs ont été étendus vers de nouvelles alternatives impliquant d'une part, un modèle de repondération, et d'autre part, une possibilité lexicale afin d'améliorer davantage le taux de désambiguïsation.

Le chapitre 3 met en valeur des approches d'expansion sémantique de requêtes. D'abord, nous avons modélisé le processus d'expansion via un graphe de réseaux petits-mondes hiérarchique (RPMH) d'un dictionnaire afin de calculer un score de proximité sémantique entre les mots en exploitant les circuits entre eux. Puis, nous avons modélisé le processus d'expansion à l'aide d'un réseau possibiliste reliant les termes de la requête à reformuler aux articles d'un dictionnaire afin de calculer un score possibiliste de proximité sémantique entre eux. Ensuite, nous avons profité de notre approche existante de repondération des termes de la requête dans le modèle d'appariement possibiliste afin d'améliorer davantage le processus d'expansion. Enfin, des agrégations de ces scores de proximité ont été suggérées et testées.

Le chapitre 4 étudie l'impact de la désambiguïsation sémantique de requêtes sur leurs expansions dans un SRI possibiliste. Nous avons proposé deux approches dont une pour la désambiguïsation et une pour l'expansion de requêtes basées sur l'analyse de corpus en utilisant des graphes de cooccurrence modélisés par des réseaux possibilistes. Nous avons profité de la double mesure de pertinence à savoir plausible et nécessaire afin de modéliser notre approche de jugement de pertinence. Nous avons comparé davantage l'efficacité de l'approche possibiliste par rapport à une approche probabiliste à base de dénombrement de circuits.

Le chapitre 5 s'intéresse à la méthodologie de la narration arabe comme un ensemble de principes pour l'évaluation de la fiabilité de l'information. Nous avons proposé une architecture dédiée à l'étude de la fiabilité des narrations arabes. Cette architecture nécessite des grammaires utiles pour l'analyse des noms propres ainsi que les chaînes de narrateurs. Nous avons modélisé le processus de reconnaissance de l'identité comme une tâche de recherche d'information possibiliste assurant l'appariement entre les noms trouvés dans les chaînes de narrateurs aux biographies des personnes correspondantes. En outre, les méta-données disponibles dans les biographies ont été utiles dans la phase d'analyse des chaînes permettant de soutenir l'utilisateur dans l'identification des sources de manque de fiabilité. Enfin, nous avons exploité un classifieur possibiliste permettant d'identifier la classe de fiabilité de chaque narration.

La conclusion résume le bilan global des différentes contributions ainsi que les principales perspectives de recherche.

Chapitre 1 :
Désambiguïsation sémantique de requêtes

1. Introduction

L'une des principales caractéristiques des langues naturelles réside dans le fait qu'un mot, une expression ou une phrase peut avoir plusieurs sens différents. Il s'agit du problème de l'ambiguïté sémantique qui reste un défi majeur pour tous les systèmes de recherche d'information (SRI) ou de traduction automatique. Ce problème oblige les chercheurs à développer des outils pour la compréhension du langage naturel. Certains auteurs [Vidhu Bhala et Abirami, 2012; Nguyen et Ock, 2013] distinguent plusieurs types d'ambiguïtés tels que la polysémie et l'homonymie, mais nous considérons généralement les mots qui ont la même orthographe et des sens différents, quel que soit le degré de proximité entre ces sens. Ces cas peuvent biaiser les résultats de tout système de traitement automatique du langage naturel (TALN). Cependant, il est nécessaire d'identifier, dans un premier temps, le sens exact d'un mot polysémique en utilisant une technique appelée la désambiguïsation sémantique (en anglais : *word sense disambiguation* (WSD)). Elle est définie comme la capacité à identifier automatiquement les sens corrects des mots suivant leurs contextes [Navigli, 2009 ; Elayeb, 2018]. Cette tâche est importante dans de nombreux domaines tels que la reconnaissance optique de caractères (OCR), la lexicographie, la reconnaissance de la parole, la compréhension du langage naturel, l'analyse et la catégorisation du contenu, la recherche d'information et la traduction automatique [Ide et Véronis, 1998; Yarowsky, 2000; Yarowsky et al., 2001; Chan et al., 2007].

Le problème de la désambiguïsation sémantique a été traité dans de nombreux travaux de recherche de diverses manières. Cependant, il a toujours été considéré comme une tâche difficile dans le domaine de TALN car il nécessite des ressources lexicales énormes telles que des corpus étiquetés, des dictionnaires, des réseaux sémantiques et/ou des ontologies [Navigli, 2009; Nguyen et Ock, 2013]. Néanmoins, il n'existe pas encore une solution suffisante qui répond aux besoins de l'utilisateur quand il est confronté à des problèmes d'ambiguïté dans les tâches de recherche d'information ou de traduction automatique. En effet, les nombreuses recherches dans ce domaine ont été fondées sur l'idée principale suivante : les relations entre une occurrence d'un mot et son contexte seront maximisés par le sens le plus probable de cette occurrence [Zhou et Han, 2005; Navigli, 2009].

Les approches, les résultats et les discussions actuelles aident à identifier et à constituer les grandes lignes du problème de la désambiguïsation sémantique et à planifier les prochaines tâches qui peuvent être effectuées pour améliorer ce domaine de recherche. Dans ce cadre, nous cherchons à améliorer les systèmes de traitement sémantique en proposant de nouveaux modèles et méthodes. Dans ce chapitre, nous proposons d'utiliser les réseaux possibilistes d'une part, et les graphes probabilistes sémantiques d'autre part comme un moyen de représenter le sens pour la désambiguïsation automatique. En fait, de nombreux types d'informations peuvent être représentés grâce à des graphes et leurs arêtes tels que les relations de synonymie, antonymie et hyperonymie. Par conséquent, l'étude des relations existantes entre les entrées d'un dictionnaire peut être réduite à une étude d'un graphe visant à exploiter des réseaux de mots.

La majorité des travaux sur la désambiguïsation sémantique est basée sur des dictionnaires traditionnels ou d'autres ressources tel que WordNet [Barque et Chaumartin, 2008], qui n'est pas très différent en termes d'organisation des sens. Le problème est que les dictionnaires traditionnels ont

été conçus pour un usage humain plutôt que pour le TALN. En fait, ces dictionnaires souffrent du manque d'informations précises utiles pour la désambiguïsation ainsi que leurs couvertures limitées. Une difficulté inhérente audit problème réside dans le manque de corpus sémantiquement étiquetés utile pour l'étape d'apprentissage [Audibert, 2002]. Même si ces corpus sont disponibles, l'existence de bruit et la dispersion des connaissances nécessaires pour la désambiguïsation rendent cette tâche largement difficile. Pour ces raisons, il est nécessaire de définir de nouveaux types de structures qui peuvent être formés et utilisés pour représenter des connaissances utiles pour la désambiguïsation. Nous profitons d'un graphe sémantique contextuel qui sera utile pour l'apprentissage et la mise-à-jour au cours du processus de désambiguïsation. Le mécanisme d'apprentissage devrait être en mesure d'acquérir de nombreux types de liens sémantiques entre un mot polysémique et ses définitions dictionnairiques contribuant à sa désambiguïsation. Le dictionnaire sémantique de contextes (DSC) est basé sur cette idée en assurant l'apprentissage automatique dans une plate-forme de désambiguïsation sémantique. Ainsi, nous combinons les connaissances extraites des dictionnaires traditionnels avec les dépendances contextuelles tirées à partir d'un corpus.

En fait, les approches de désambiguïsation ont besoin de modèles d'apprentissage et d'appariement assurant les calculs des scores des similitudes (ou de pertinence) entre les sens des mots polysémiques et leurs contextes. Les modèles existants pour la désambiguïsation sont basés sur des données pauvres, incertaines et imprécises et utilisent des modèles d'apprentissage et d'appariement probabilistes (par exemple [Loupy, 2000; Yuret et Yatbaz, 2010; Nguyen et Ock, 2013]), alors que la théorie des possibilités est naturellement conçue pour ce type d'applications. Elle permet d'exprimer l'ignorance et de tenir compte de l'imprécision et de l'incertitude dans le même temps. Par exemple, dans nos travaux récents [Ayed et al., 2012ab ; 2014ab], nous avons proposé une approche possibiliste pour la désambiguïsation morphologique automatique de textes arabes. Nous avons montré aussi la contribution des modèles possibilistes par rapport à ceux probabilistes. Notre contribution dans ce chapitre consiste à proposer un modèle possibiliste pour la désambiguïsation sémantique monolingue appliquée à la langue française. En effet, nous évaluons la pertinence du sens d'un mot polysémique vis-à-vis de son contexte donné sous forme d'une phrase en utilisant deux types de pertinences: la *pertinence plausible* et la *pertinence nécessaire*. D'autre part, le problème de désambiguïsation doit être modélisé du point de vue dynamique. En effet, le calcul dynamique du sens dans un espace sémantique consiste à spécifier des contraintes sur chacun des points de cet espace. Il permet d'obtenir des relations sémantiques entre les mots. A partir de ces relations, nous pouvons calculer les distances sémantiques entre un mot polysémique et ses définitions mentionnées dans un dictionnaire traditionnel, étant donné ses informations contextuelles.

Nous proposons dans ce chapitre une nouvelle approche possibiliste pour la désambiguïsation automatique monolingue que nous évaluons par la suite. En fait, en dépit de leurs avantages, les dictionnaires traditionnels souffrent de l'absence d'informations précises utiles pour la désambiguïsation. En outre, les méthodes d'apprentissage souffrent des limites de couverture des corpus sémantiquement étiquetés. Pour ces multiples raisons, l'utilité d'un dictionnaire sémantique de contextes (DSC) s'impose pour améliorer l'apprentissage automatique dans une plate-forme de désambiguïsation sémantique. Nous proposons une nouvelle approche possibiliste qui combine les dictionnaires traditionnels et un corpus étiqueté afin de construire un DSC et identifier le sens d'un mot en utilisant un modèle d'appariement possibiliste. D'autre part, nous présentons et nous évaluons une deuxième nouvelle approche probabiliste pour la désambiguïsation automatique monolingue. Cette approche utilise et étend une distance sémantique probabiliste existante pour calculer les similarités entre les mots en exploitant un graphe sémantique d'un dictionnaire traditionnel et le DSC. L'évaluation et la comparaison de ces deux approches sont faites sur la collection de test ROMANSEVAL. Nous comparons davantage nos résultats à des systèmes

existants de désambiguïsation monolingues français. Les expérimentations ont montré une amélioration encourageante en termes de taux de désambiguïsation de mots français [Elayeb et al., 2015a]. Ces résultats révèlent la contribution de la théorie des possibilités comme un moyen de traiter l'imprécision dans les systèmes d'information. En fait, cette approche possibiliste de désambiguïsation sémantique est exploitée dans une phase de désambiguïsation de requêtes et contribue davantage à l'amélioration de leurs expansions dans un SRI intelligent que nous détaillons dans le chapitre 4.

Le présent chapitre est organisé comme suit : Nous discutons dans la section 2 les problèmes posés par les principales approches de désambiguïsation sémantique monolingue ainsi que nos contributions pour résoudre certaines limites. Le dictionnaire sémantique de contexte sera détaillé dans la section 3. Les deux approches de désambiguïsation sémantique possibiliste et probabiliste sont détaillées respectivement dans les sections 4 et 5. La section 6 récapitule les expérimentations réalisées ainsi qu'un bilan comparatif de ses deux approches avec les travaux existants. La section 7 résume nos contributions relatives à la désambiguïsation ainsi que nos perspectives de recherche.

2. Synthèse des approches existantes de désambiguïsation sémantique

Les approches existantes de désambiguïsation sémantique sont classées dans la littérature selon deux critères : le premier est lié à la source de connaissance utilisée, alors que le second s'intéresse à sa manière de structuration [Vidhu Bhala, 2012]. En fait, les performances de ces approches sont sensibles à ces critères. Nous nous intéressons principalement aux quatre types d'approches, à savoir : (i) les approches basées sur la modélisation des connaissances ou de raisonnement ; (ii) les approches supervisées ; (iii) les approches non supervisées ; et (iv) les approches hybrides utilisant à la fois des connaissances extraites à partir de ressources lexicales (comme les dictionnaires traditionnels) et des informations contextuelles/distributionnelles apprises à partir d'un corpus.

Les approches à base de connaissance utilisent des modèles de l'intelligence artificielle, à savoir symboliques/cognitifs ou connexionnistes basés sur les réseaux de neurones, afin de modéliser la compréhension humaine de la langue naturelle [Audibert, 2003]. En fait, les bases de connaissances particulières utilisées par ce genre d'approche ne permettent pas une couverture suffisante de la langue en question. De plus, ces volumes importants des connaissances sont collectés et traités manuellement. D'autre part, Ponzetto et Navigli (2010) ont confirmé que les approches à base de connaissances peuvent, à un certain niveau, remplacé les approches supervisées. En effet, les approches à base de connaissances profitent uniquement des informations provenant des bases de connaissances lexicales et elles sont dispensées de chercher d'autres informations issues des corpus.

Par ailleurs, plusieurs approches à base de connaissances ont profité de WordNet comme ressource linguistique [Sinha et Mihalcea, 2007 ; Navigli et Lapata, 2010 ; Ponzetto et Navigli, 2010 ; Agirre et al., 2014]. Par contre, les tendances de la communauté scientifique du domaine encouragent aujourd'hui les chercheurs à collecter et tester d'autres ressources linguistiques loin de WordNet.

Les méthodes supervisées [Navigli, 2009 ; Ponzetto et Navigli, 2010] de WSD ont progressivement perdu leurs popularités malgré l'efficacité de leurs algorithmes de désambiguïsation. En fait, ce genre de méthode exige une phase de réapprentissage à base de données annotées avant qu'elle sera adoptée à d'autres langues [Panchenko et al., 2017]. De plus, la réutilisation des modèles d'une langue à une autre détériore davantage le taux de classification [Khapra et al., 2009]. Au contraire aux approches supervisées, les approches non supervisées [Duque et al., 2015 ; Koppula et al., 2017] sont à base de techniques d'apprentissage automatique utilisant des corpus non annotés, sans connaissance au préalable [Nasiruddin, 2013]. En outre, les approches non supervisées sont à base

de techniques de discrimination ou d'induction de sens afin d'identifier automatiquement les sens corrects en utilisant des corpus non étiquetés.

Certes, construire des corpus étiquetés pour l'apprentissage des algorithmes de désambiguïsation sémantique est une tâche difficile et nécessite beaucoup de temps [Agirre et Martinez, 2000; Agirre et Edmonds, 2006; Nguyen et Ock, 2013]. Si nous supposons que ces corpus existent, extraire des connaissances utiles pour la désambiguïsation lexicale est une tâche difficile pour plusieurs raisons telles que : (i) le problème de bruit dans le corpus ; (ii) les informations pertinentes nécessaires à la désambiguïsation sont distribuées dans le corpus ; et (iii) les mots polysémiques ont peu d'occurrences par rapport à la taille du corpus [Audibert, 2003]. Cependant, les dictionnaires représentent des réseaux d'associations riches entre les mots et un ensemble de catégories sémantiques potentiellement exploitables pour le TALN. Ils contiennent moins de bruit et il est facile d'extraire les sens de la majorité des mots polysémiques. Le problème est que ces dictionnaires ont été faits pour l'utilisation humaine et ne conviennent pas pour des traitements automatiques. De plus, ils souffrent du manque d'informations précises utiles pour la désambiguïsation. En outre, l'incohérence des lexicographes des dictionnaires est devenue une limite bien connue [Kilgarriff, 1994]. Un dictionnaire est également lié à une période de l'histoire, donc contenant certains sens qui ne sont pas nécessairement utilisés dans d'autres périodes.

Par ailleurs, les approches hybrides combinent les deux approches tant à base de dictionnaires, qu'à base de corpus afin de profiter de leurs avantages et pallier à certaines de leurs limites. Cependant, nous serons limités à des approches basées sur les dictionnaires parce que les autres types de ressources (telles que les thésaurus et ontologies) ne sont pas disponibles pour toutes les langues et ne couvrent pas nécessairement une langue donnée. Dans une approche hybride utilisant un dictionnaire, le rôle principal de ce dernier est : (i) fournir tous les sens possibles des mots; et (ii) fournir une définition moins bruitée de chaque sens. Ensuite, il est facile d'extraire et d'exploiter ces définitions afin d'en tirer des connaissances hors-contexte. Quant au rôle du corpus, il se résume à : (i) filtrer les sens des mots en ne conservant que les sens vraiment utilisés; et (ii) fournir des connaissances contextuelles ou distributionnelles utiles pour la désambiguïsation.

Approche de WSD	**Avantages & inconvénients de l'approche**
A base de connaissance	👍 Les approches sont à base des algorithmes de haute-précision 👎 Les résultats de désambiguïsation sont sensibles aux degrés des couvertures des ressources linguistiques utilisées. 👎 Existence de divergences structurelles et du contenu entre les ressources choisies qui ne sont pas aussi disponibles pour toutes les langues.
Supervisé	👍 Titulaire des meilleurs taux de désambiguïsation sémantique. 👎 Difficulté d'avoir un ensemble d'apprentissage annoté couvrant tout le lexique d'une langue donnée.
Non Supervisé	👍 Exempté de ressources linguistiques externes (thésaurus, dictionnaires, etc.) et de corpus sémantiquement annotés. 👎 Les approches sont à base des algorithmes compliqués ayant des performances inférieurs aux deux approches précédentes.
Hybride	👍 Bénéficie de double avantage issu des approches supervisées et non supervisées. 👎 Lors de l'hybridation, la mise en place des approches combinées semble complexe avec un choix de pondération aléatoire ou heuristique.

Tableau 1.1 : Synthèse des approches de désambiguïsation sémantique

Récemment, nous avons montré dans [Elayeb, 2018] que le choix d'une approche de désambiguïsation sémantique en dépit d'autres dépend principalement de ressources linguistiques

utilisées lors de la phase d'entrainement. Ainsi, les approches combinant plusieurs ressources à la fois ont réussi à dépasser plusieurs lacunes. Le tableau 1.1 [Ben Khiroun, 2018] synthétise une étude comparative des approches existantes de désambiguïsation sémantique.

En fait, les approches hybrides [Lafourcade, 2007 ; Barathi et Valli, 2010 ; Yuret et Yatbaz, 2010 ; Jimeno-Yepes, 2011 ; Lafourcade et Brun, 2017] n'ont pas été encore massivement étudiées. Cependant, il est probable que ce type d'approche puisse fournir de meilleurs résultats. C'est pour cette raison que nous proposons dans ce chapitre une approche hybride basée sur un dictionnaire pour la désambiguïsation sémantique monolingue. Nous exploitons un corpus étiqueté pour extraire des connaissances contextuelles, qui modélisent les liens de cooccurrence et les dépendances entre les mots (contextes) et leurs sens qui sont vraiment utilisés. Les sens sont extraits d'un dictionnaire traditionnel de haute couverture. Selon nos connaissances, aucune des méthodes actuelles n'a traité d'une manière exhaustive le problème de l'organisation du lexique. Pour résoudre ce problème, nous définissons et nous utilisons un dictionnaire sémantique des contextes (DSC) qui stocke les connaissances extraites à la fois du corpus et du dictionnaire (cf. section 3). De plus, ces approches de désambiguïsation consistent à calculer les similitudes entre les sens des mots et leurs contextes afin d'identifier le "meilleur" sens. Les approches existantes utilisent des distances probabilistes, alors que la théorie des possibilités fournit un cadre novateur pour une telle application, mais qui n'a pas encore été appliquée pour le domaine de la désambiguïsation sémantique. En effet, les méthodes existantes basées sur les calculs des similitudes ne cherchent pas à représenter les distances sémantiques entre les sens et ne gèrent pas correctement l'organisation des sens obtenus.

En outre, plusieurs travaux de recherche ont tenté de résoudre le problème de la polysémie au niveau du dictionnaire. Gaume (2006) a utilisé un dictionnaire comme source d'information pour découvrir les relations entre les éléments lexicaux. Son travail est basé sur un algorithme qui calcule la distance sémantique entre les mots du dictionnaire en tenant compte de la topologie complète du dictionnaire, ce qui lui confère une plus grande robustesse. Cet algorithme permet de désambiguïser des mots polysémiques dans les définitions du dictionnaire. Il a testé cette approche sur la désambiguïsation des définitions des dictionnaires eux-mêmes. Il a proposé la méthode PROX détaillée ci-dessous dans la section 5.

Le modèle que nous proposons est soutenu par un espace sémantique où les différents sens d'un mot sont organisés. Selon la classification de Vidhu Bhala et Abirami (2012), nous utilisons une représentation sémantique structurale impliquant des relations sémantiques entre les mots et les sens. Le calcul du sens d'une phrase est un processus dynamique au cours duquel les sens des différents mots sont mutuellement influencés, ce qui mène simultanément à la détermination du sens de chaque mot et à l'identification du sens global de la phrase. D'une part, nous utilisons les réseaux possibilistes pour calculer la distance entre le contexte et un sens donné. D'autre part, nous proposons une approche probabiliste exploitant la distance probabiliste de Gaume (2006) afin de le généraliser au point d'obtenir une méthode dynamique de calcul de sens. Nous proposons également de calculer un taux d'ambiguïté préliminaire de chaque phrase polysémique.

3. Modélisation du dictionnaire sémantique des contextes (DSC)

Selon Vidhu Bhala et Abirami (2012), la robustesse des connaissances disponibles sous forme de corpus ou de dictionnaires est l'un des paramètres les plus importants qui régissent les orientations de la recherche en désambiguïsation sémantique. En fait, il est nécessaire de modéliser cette connaissance pour résoudre le problème de la polysémie. Dans notre cas, nous avons utilisé des graphes comme une solution générique pour modéliser les connaissances requises pour la

désambiguïsation. Nous appelons cette nouvelle source linguistique un dictionnaire sémantique des contextes (DSC). Pour construire et représenter de manière automatique le graphe $G = (S; A)$ associé à un mot, il est nécessaire de définir l'ensemble des nœuds S ainsi que l'ensemble des arêtes A. Ces nœuds et arêtes sont générés à la fois à partir du corpus et du dictionnaire traditionnel dans l'étape d'apprentissage.

3.1. L'ensemble de nœuds

Les mots polysémiques (nœuds du graphe), leurs contextes, ainsi que les entrées du dictionnaire traditionnel sont les entrées du dictionnaire sémantique des contextes. Nous définissons ci-après les paramètres suivants par :

- ***Polysémie*(*ph*) :** Ensemble des mots polysémiques dans la phrase *ph* : $p_1, p_2, ..., p_k$
- ***Significatif*(*ph*) :** Ensemble des mots significatifs et non polysémiques constituant la phrase *ph* : $m_1, m_2, ..., m_i$.
- ***Contexte*(*p*, *ph*) :** $\{Significatif\ (ph) \cup Polysémie(ph)\} \setminus \{p\}$; avec p est un mot significatif $\in ph$.
- ***Contexte*(*ph*) :** $\{Significatif(ph) \cup Polysémie(ph)\}$.
- ***Définition*(m_i) :** Ensemble des définitions des mots significatifs m_i dans le dictionnaire traditionnel d_{mi} si le mot m_i n'est pas polysémique ; et $p^1_i, p^2_i, ..., p^a_i$ si le mot m_i est polysémique. Cet ensemble contient uniquement les sens qui sont réellement utilisés dans la phase d'apprentissage (i.e. dans le DSC).

3.2. L'ensemble d'arêtes

Il existe plusieurs types de réseaux lexicaux, selon la nature des relations sémantiques qui définissent les arêtes du graphe (nœuds représentant les lexèmes de la langue). Les trois principaux types de relations utilisées dans notre DSC sont :

- **Relations syntagmatiques**, ou de cooccurrence; nous construisons une arête entre deux mots s'ils cooccurrent dans le même contexte dans le corpus. En effet, nous nous sommes inspirés des approches qui apprennent à partir des dépendances entre les mots apparaissant dans un contexte donné et les sens représentant un mot [Yuret et Yatbaz, 2010], sauf que nous apprenons à partir de corpus étiquetés. Ces relations sont formalisées comme suit:

$$\forall m_i, m_j \in contexte(ph)\ si\ i \neq j \rightarrow <m_j, m_i> \in A$$

- **Relations paradigmatiques**, en particulier de synonymie; nous construisons un graphe dans lequel deux nœuds sont reliés par une arête si les mots correspondants maintiennent une relation synonymique [Ploux et Victorri, 1998]. Autrement dit, si ses nœuds partagent des mots dans leurs définitions dictionnairiques:

$$\forall m_i, m_j, si\ \{Définition(m_i) \cap Définition(m_j)\} \neq \varnothing \rightarrow <m_j, m_i> \in A$$

- **Relations de proximité sémantique**; ils sont des relations moins spécifiques qui prennent en compte à la fois l'axe paradigmatique et l'axe syntagmatique. Comme dans Véronis et Ide (1990), nous avons construit un graphe lexical du dictionnaire traditionnel. Cela permet de créer des liens entre les mots et les sens indépendamment des contextes [Yuret et Yatbaz, 2010]. En effet, nous construisons une arête entre deux mots m_i et m_j si m_j apparaît dans la définition de m_i. Cela peut être formalisé comme suit:

$$\forall\ m_i \in Polysémie(ph), \forall\ m_j \in Définition(m_i) \rightarrow <m_j, m_i> \in A$$

Ces arêtes sont pondérées selon des formules qui seront présentées dans les sections suivantes et illustrées par des exemples dans [Elayeb et al., 2015a].

4. Approche possibiliste de désambiguïsation sémantique

Le processus de désambiguïsation est également considéré comme une tâche de classification où nous avons deux étapes d'apprentissage et de test. Dans l'étape d'apprentissage, nous apprenons à partir de dépendances entre les sens des mots et leurs contextes. Ceci peut être réalisé dans les corpus étiquetés (apprentissage à base des jugements) conduisant à une approche semi-automatique. Nous pouvons également pondérer ces dépendances directement à partir d'un dictionnaire traditionnel (apprentissage à base de dictionnaire), ce qui peut être considéré comme une approche automatique [Ben Khiroun et al., 2012]. Dans ce cas, nous avons besoin d'organiser toutes les instances de manière à améliorer le taux de classification. En fait, nous proposons de trier les instances selon leurs taux d'ambiguïté (cf. section 4.2). Dans l'étape de test, la distance entre le contexte d'une occurrence d'un mot et ses sens est calculée afin de sélectionner le meilleur sens. Par conséquent, nous présentons dans la suite nos formules mathématiques relatives aux calculs du Degré de Pertinence Possibiliste (DPP) ainsi que le taux d'ambiguïté d'une phrase polysémique. Des exemples de ces calculs sont détaillés dans [Elayeb et al., 2015a].

4.1. Le Degré de Pertinence Possibiliste (DPP)

Supposons que nous ayons un seul mot polysémique dans une phrase *ph*. Notons $DPP(S_i|ph)$ le degré de pertinence possibiliste d'un sens S_i étant donné la phrase polysémique *ph*. Considérons que *ph* est composée de termes $ph = (t_1, t_2, \ldots, t_T)$. Nous évaluons la pertinence d'un sens S_i étant donné *ph* par un modèle d'appariement possibiliste de RI utilisée par [Boughanem et al., 2009; Elayeb, 2009; Elayeb et al., 2009 ; 2011]. Dans le cas de RI, le but est de calculer un score de correspondance entre une requête et un document.

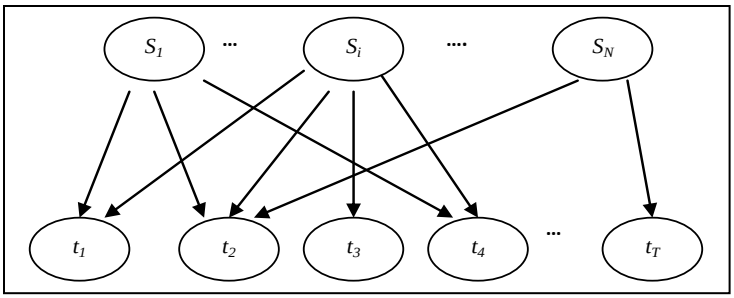

Figure 1.1 : Réseau possibiliste de l'approche de désambiguïsation sémantique

Dans le cas de désambiguïsation sémantique, la pertinence d'un sens étant donné une phrase polysémique est modélisée par une double mesure. La *pertinence possible* permettant de rejeter les sens non-pertinents, et la *pertinence nécessaire* permettant de renforcer la pertinence des sens de mots restants, qui n'ont pas été rejetés par la possibilité. Dans notre cas, le réseau possibiliste relie le sens S_i aux mots d'une phrase polysémique $ph = (t_1, t_2, \ldots, t_T)$ tel le cas présenté dans la figure 1.1. La pertinence de chaque sens S_i étant donné une phrase polysémique ph_i est calculée comme suit :

Selon le modèle d'appariement dans [Elayeb et al., 2009 ; 2011], la possibilité $\Pi(S_j|ph)$ est proportionnelle à :

$$\Pi'(S_j|ph) = \pi(t_1|S_j) * \ldots * \pi(t_T|S_j) = nFreqt_{1j} * \ldots * nFreqt_{Tj} \quad (1.1)$$

Avec : - $nFreqt_{ij} = \frac{Freqt_{ij}}{maxFreqt_{ij}}$: La fréquence normalisée du terme t_i dans le sens S_j.

- $Freqt_{ij} = \frac{le\ nombre\ d'occurrences\ du\ terme\ t_i\ dans\ le\ sens\ S_j}{nombre\ de\ termes\ dans\ le\ sens\ S_j}$

- $maxFreqt_{ij}$: La fréquence maximale.

La nécessité de restituer un sens pertinent S_j pour la phrase *ph*, notée $N(S_j|ph)$, est donnée par :

$$N(S_j|ph) = 1 - \Pi(\neg S_j|ph) \quad (1.2)$$

Avec:

$$\Pi(\neg S_j|ph) = \frac{\Pi(ph|\neg S_j)*\Pi(\neg S_j)}{\Pi(ph)} \quad (1.3)$$

De même $\Pi(\neg S_j|ph)$ est proportionnelle à :

$$\Pi'(\neg S_j|ph) = \pi(t_1|\neg S_j) * \dots * \pi(t_T|\neg S_j) \quad (1.4)$$

Ce numérateur peut être exprimé par :

$$\Pi'(\neg S_j|ph) = (1 - \phi S_{1j}) * \dots * (1 - \phi S_{Tj}) \quad (1.5)$$

Avec:

$$\phi S_{ij} = Log_{10}\left(\frac{nCS}{nS_i}\right) * (nFreqt_{ij}) \quad (1.6)$$

Avec: - *nCS* : Nombre de sens du mot dans le dictionnaire traditionnel.

- nS_i : Nombre de sens du mot contenant le terme t_i. Ceci inclut uniquement les sens qui sont dans DSC et ne couvre pas tous les sens du t_i existant dans le dictionnaire traditionnel.

Nous définissons le Degré de Pertinence Possibiliste (*DPP*) de chaque sens S_j étant donné une phrase polysémique *ph* par la formule suivante :

$$DPP(S_j|ph) = \Pi(S_j|ph) + N(S_j|ph) \quad (1.7)$$

Les sens préférés sont ceux qui ont une valeur $DPP(S_j|ph)$ élevée. Ce score nous servira dans le calcul du taux d'ambiguïté de chaque phrase polysémique détaillé dans la section suivante.

4.2. Le taux d'ambiguïté d'une phrase polysémique

Une phrase est considérée comme ayant un taux d'ambiguïté élevé si les sens correspondants aux mots ambigus dans la phrase ont une signification similaire et/ou ne correspondent pas au contexte de la phrase. Nous calculons le taux d'ambiguïté d'une phrase polysémique *ph* en utilisant les valeurs de la possibilité et de nécessité comme suit : (i) nous indexons tous les sens possibles d'un mot ambigu; (ii) nous utilisons l'index de chaque sens comme une requête; (iii) nous évaluons la pertinence de la phrase polysémique *ph* étant donné cette requête; et (iv) cette phrase est considérée comme très ambiguë si elle est pertinente pour de nombreux sens ou si elle ne l'est pour aucun sens. Par conséquent, le taux d'ambiguïté est inversement proportionnel à la valeur de l'écart type. Il est calculé par la formule (1.8).

$$Taux\ d'ambiguïté(ph) = 1 - \sigma(ph) \quad (1.8)$$

Avec : $\sigma(ph)$ représente l'écart type du score $DPP(S_j|ph)$ correspondant à chaque sens d'un mot ambigu contenu dans la phrase polysémique *ph*. Il est calculé par la formule (1.9).

$$\sigma(ph) = \sqrt{\left(\frac{1}{N}\right) * \sum_j (DPP(S_j|ph) - S)^2} \quad (1.9)$$

Où S est la moyenne des scores $DPP(S_j|ph)$ et N le nombre de sens possibles dans le dictionnaire traditionnel.

En fait, le calcul du taux d'ambiguïté de chaque phrase polysémique de la collection de test nous a permis de tester et comparer plusieurs méthodes d'apprentissage du DSC. Suite à chaque test, nous avons généré un DSC qui servira comme un sous-ensemble d'apprentissage lors de l'évaluation des phrases polysémiques du corpus de test. Dans [Ben Khiroun et al., 2012], nous avons proposé deux méthodes d'apprentissage du DSC, une à base de *dictionnaire* et l'autre à base des *jugements*.

Etant donné une phrase polysémique *ph* englobant le mot polysémique *w*, nous joignons les termes de *ph* avec le sens correct du *w*. En effet, ce sens correct de *w* peut être sélectionné via les connaissances contextuelles de cooccurrence issues d'un dictionnaire. Dans ce cas, on parle d'*apprentissage à base de dictionnaire*. Comme il peut être sélectionné via les annotations présentes dans un corpus. Dans ce cas, on parle d'*apprentissage à base des jugements*.

L'algorithme 1 résume les étapes du processus d'apprentissage à base de dictionnaire. Dans cette méthode, nous avons utilisé la règle 20/80 de la façon suivante : D'abord, nous trions les phrases annotés avec des sens selon un ordre croissant (respectivement décroissant) de leurs taux d'ambiguïtés. Ensuite, nous générons le DSC à partir de 80% des phrases les moins (respectivement les plus) ambiguës. Enfin, le 20% des phrases restantes seront évaluées en utilisant le DSC obtenu comme nouvelle ressource linguistique de désambiguïsation.

Algorithme 1 : Apprentissage à base de dictionnaire [Ben Khiroun, 2018]

Entrées : phrases ambiguës.
Sorties : phrases annotées avec des sens et triées selon le taux d'ambiguïté.
Variables : w_i : mot ; S_i, S_{max}: sens ;

```
1 début
2 |   pour chaque phrase ambiguë faire
3 |   |   pour chaque mot ambigu wi faire
4 |   |   |   calculer DPP pour chaque sens Si ∈ dictionnaire.
5 |   |   |   associer au mot wi le sens Smax ayant le plus grand DPP.
6 |   |   fin
7 |   fin
8 |   pour tous les phrases ambiguës annotées avec des sens faire
9 |   | trier en ordre croissant/décroissant les phrases selon le taux d'ambiguïté.
10|   fin
11 fin
```

Par ailleurs, la génération du DSC via la méthode d'apprentissage à base des jugements nécessite le recours à une technique de validation croisée [Kohavi, 1995]. En effet, le DSC est obtenu à partir de 90% des phrases aléatoirement sélectionnées issues de la collection ROMANSEVAL, alors que le 10% restantes sont utiles pour les tests. Nous avons itérés ce processus (9+1) fois en utilisant les 60 mots ambigus du ROMANSEVAL.

5. Approche probabiliste de désambiguïsation sémantique

Nous proposons dans cette section une nouvelle approche probabiliste basée sur une version généralisée de la méthode PROX, initialement proposée par [Gaume et al., 2004]. Nous présentons

le calcul sémantique dans la suite. Un exemple détaillé du calcul est dans [Elayeb et al., 2015a]. La modélisation du problème de désambiguïsation par graphe sémantique nécessite des représentations mathématiques et des traitements des données permettant de mesurer le sens d'un mot par rapport à ses définitions mentionnées dans le dictionnaire traditionnel. Pour ce faire, nous avons transformé le graphe en une matrice de Markov dont les états sont les nœuds du graphe et les arêtes sont les transitions possibles. Nous générons d'abord à partir du graphe du DSC une matrice de transition ou d'adjacence (cf. section 5.1), qui est ensuite transformée en une matrice de Markov (cf. section 5.2). Enfin, nous identifions les sens corrects des mots polysémiques en appliquant l'algorithme de désambiguïsation basé sur la proximité, que nous détaillons dans la section 5.3.

5.1. Construction de la matrice d'adjacence

Nous générons la matrice d'adjacence à partir du graphe $G = <S, A>$. On note $[G]$ la matrice carrée $n \times n$ telle que pour tout $r, s \in S$, $[G]_{r,s} = |<s, r>|$ si $(r, s) \in A$, et $[G]_{r,s} = 0$ si $(r, s) \notin A$. Nous appelons $[G]$ la matrice de transition de G. Puisque G n'est pas orienté, donc $[G]$ est une matrice symétrique. En plus, G est un graphe réflexif ; d'où $\forall r \in S$, $[G]_{r,r} = 1$.

5.2. Construction de la matrice de Markov

Nous générons la matrice de Markov à partir de la matrice d'adjacence. Notons $[\hat{G}]$ la matrice de Markov correspondante au graphe $G = <S, A>$ et définie par :

$$\forall r, s \in S, \left[\hat{G}\right]_{r,s} = \frac{[G]_{r,s}}{\sum_{x \in S} [G]_{r,x}} \tag{1.10}$$

5.3. L'algorithme de désambiguïsation basé sur la proximité

Nous présentons d'abord le principe de l'algorithme basé sur la proximité. Puis, nous présentons une nouvelle version de cette méthode qui est appliquée à la désambiguïsation sémantique.

5.3.1. L'algorithme basé sur la proximité

Cette méthode a été proposée par Gaume et al. (2004). Il s'agit d'une méthode stochastique utilisée pour étudier la structure d'un graphe de dictionnaire à l'aide de chaînes de Markov. Nous rappelons que les chaînes de Markov ont atteint des résultats satisfaisants en désambiguïsation sémantique (par exemple [Loupy, 2000]). Le principe de la méthode consiste à transformer un graphe en une chaîne de Markov dont les états sont les sommets du graphe en question, et ses arêtes sont les transitions possibles : une particule, partant à l'instant $t = 0$ d'un sommet s_0, se déplace en un pas sur s_1 l'un des voisins de s_0 sélectionné aléatoirement ; la particule se déplace alors de nouveau en un pas sur s_2, l'un des voisins de s_1 sélectionné aléatoirement, etc. Si au t-ième pas la particule est sur le sommet s_t elle se déplace alors en un pas sur le sommet s_{t+1} qui est sélectionné aléatoirement parmi tous les voisins équiprobables de s_t. Selon Gaume (2004), une trajectoire $s_1, s_2, ..., s_t, ...$ ainsi sélectionnée est une "balade" aléatoire sur le graphe, et ce sont les dynamiques de ce trajectoires qui donnent les propriétés structurelles des graphes étudiés.

Gaume et al. (2004) ont défini $PROX(G, i, s, r)$ comme étant la probabilité que la particule en partant du sommet r à l'instant $t = 0$, soit sur le sommet s à l'instant $t = i$. Ainsi, $PROX(G, i, s, r) = [\hat{G}^i]_{r,s}$; où A^i est la matrice A multipliée i fois par elle-même.

5.3.2. Le calcul dynamique de sens

Nous proposons une méthode de calcul dynamique du sens d'un mot étant donné son contexte, en exploitant le graphe sémantique et en calculant la distance sémantique. Notre approche est basée essentiellement sur le principe de PROX [Gaume et al., 2004]. Cette méthode représente ainsi une mesure de similarité entre les sommets d'un graphe en calculant la distance sémantique entre un mot et ses définitions, ce qui permet d'envisager une exploitation originale et novatrice des graphes sémantiques. Nous définissons la tâche comme suit :

On considère un lemme m_i comme mot polysémique. Nous notons :

- m_i est un nœud du graphe G.
- $Définition(m_i) = p^1_i, p^2_i, ..., p^a_i$.
- $G^\infty = \lim_{t\to\infty}\left[\hat{G}^t\right]$; d'où G^∞ est un vecteur de $\mathbb{R}^a$.
- $f_\infty(r,s) = \lim_{t\to\infty} PROX(G,t,s,r)$. La fonction f_∞ indique la proximité sémantique entre les mots polysémiques et leurs définitions dans le DSC.

Ainsi, nous avons la propriété suivante :

***Propriété* :**

Puisque le graphe G est réflexif[7] et fortement connexe[8] alors :

$\forall a \in S, \lim_{t\to\infty} PROX(G,t,r,s) = \lim_{t\to\infty} PROX(G,i,a,r)$; Cela signifie que la probabilité, pour un temps t assez long, d'atteindre un sommet s ne dépend pas du sommet de départ (s ou a).

On dit que a est fortement lié à d^j si est seulement si : $\forall j \in \mathbb{N} \setminus \{i\}, f_\infty(a, d^i) > f_\infty(a, d^j)$. Et dans ce cas, le résultat de désambiguïsation du mot polysémie a est d^i.

Le mot β portant de l'information vérifie les propriétés suivantes :

- $\beta \in contexte(a, ph)$;
- $\forall \delta \in contexte(a, ph) \setminus \{\beta\}, f_\infty(a, \beta) = max_\delta(f_\infty(a, \delta))$; avec β est la définition sémantique de a.

Dans ce cas, β est la définition sémantique de a dans le dictionnaire sémantique des contextes.

6. Résultats expérimentaux

Cette section présente la collection de test ROMANSEVAL (cf. section 6.1) et les scénarios expérimentaux utilisés dans nos expériences (cf. section 6.2). Pour améliorer notre évaluation, nous avons réalisé deux types d'évaluation à l'étape d'apprentissage qui ont été détaillées et discutées dans [Ben Khiroun et al., 2012]. Un premier apprentissage à base des jugements, et un second à base d'un dictionnaire. Ce faisant, nous analysons, interprétons et comparons dans la section 6.3 la performance des différents tests effectués des deux approches possibiliste et probabiliste.

[7] Un graphe $G = (V, E)$ est dit réflexif lorsque : $\forall r \in V, (r, r) \in E$.

[8] Soit $G = (V, E)$ un graphe. Nous dirons que G est fortement connexe si et seulement si : $\forall r, s \in V$, il existe un chemin c de longueur finie dans G dont r est l'origine et s l'arrivée.

6.1. La collection de test ROMANSEVAL

Nous avons utilisé dans nos expériences la collection de test ROMANSEVAL, qui fournit les outils nécessaires pour la désambiguïsation sémantique, y compris: (1) un ensemble de documents (publié par le Journal Officiel de la Commission Européenne); et (2) une liste de phrases de test, y compris les mots ambigus. L'ensemble des documents est constitué de textes parallèles en 9 langues extraits du Journal Officiel de la Commission Européenne (série C, 1993). Les textes (au nombre de plusieurs milliers) sont constitués de questions écrites sur un large éventail de sujets et les réponses correspondantes de la Commission Européenne. La taille totale du corpus est d'environ 10.2 millions de mots (environ 1.1 million de mots par langue), qui ont été recueillis et préparés dans les projets MULTEXT-MLCC [Segond, 2000]. Ces textes ont été préparés afin d'obtenir une collection de test. Le corpus a été découpé en mots étiquetés avec des étiquettes catégoriques afin de distinguer les noms (N), les adjectifs (A) et les verbes (V). Ensuite, les 600 mots les plus fréquents (200 N, 200 A, 200 V) ont été extraits selon leurs contextes d'apparition. Ces mots ont été annotés en parallèle par 6 étudiants en linguistique, en conformité avec les sens du dictionnaire français "Le Petit Larousse". Chaque occurrence de mot peut avoir une ou plusieurs étiquettes de sens ou pas. Après cette première étape, les 60 mots les plus polysémiques ont été conservés (20 N, 20 A, 20 V) et leurs occurrences ont été étiquetés en 3624 contextes.

6.2. Les scénarios expérimentaux

Afin d'effectuer la tâche d'apprentissage, nous avons construit le dictionnaire sémantique de contextes au format XML. Pour alimenter le DSC, nous appliquons généralement la technique de validation croisée [Kohavi, 1995] dans nos scénarios expérimentaux, sauf pour l'apprentissage à base de dictionnaire détaillé dans [Ben Khiroun et al., 2012]. Dans cette méthode de validation, 90% des phrases, sélectionnées aléatoirement à partir de la collection ROMANSEVAL, sont utilisées pour la phase d'apprentissage du DSC et les 10% restants sont utilisées pour les tests. Cet essai est répété 10 fois pour tous les 60 mots dans chacun des dix passages. Le taux de précision (La moyenne *Kappa*) est calculé sur les 9+1 combinaisons. Nous avons organisé le DSC en trois fichiers selon la catégorie grammaticale (Adjectifs, Noms et Verbes). Notons que les mêmes fichiers du DSC sont utilisés dans les deux approches possibiliste et probabiliste. Les phrases sont lemmatisées à l'aide de l'outil *TreeTagger*[9] pour la langue française.

Ainsi, nous avons proposé trois manières différentes d'apprentissage pour la génération du DSC, à savoir : (i) un apprentissage à base de dictionnaire par ambiguïté décroissante ; (ii) un apprentissage à base de dictionnaire par ambiguïté croissante ; et (iii) un apprentissage à base des jugements.

Nous avons comparé ces trois méthodes de génération du DSC en utilisant le taux d'accord calculé via la formule suivante [Segond, 2000] :

$$Accord = \frac{\left|\left\{S_i \in \Delta,\ où\ S_i^{système} = S_i^{juges}\right\}\right|}{|\{S_i \in \Delta\}|} \quad (1.11)$$

Où : - Δ représente l'ensemble de sens, jugés par les annotateurs, correspondant aux phrases de test.
- $S_i^{système}$ représente le sens sélectionné par le système suite au calcul du DPP.
- S_i^{juges} représente le sens proposé par les juges.

En outre, nous calculons le taux de précision pour chaque mot en utilisant la métrique "*Kappa*" [Cohen, 1968 ; Eugenio, 2000]. En effet, le coefficient *Kappa* est fondé sur l'accord observé entre des

[9]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

jugements qualitatifs ou non, résulte de la somme d'une composante "aléatoire" et d'une composante d'accord "véritable ". De plus, le coefficient *Kappa* propose de chiffrer l'intensité ou la qualité de l'accord réel entre des jugements qualitatifs appariés. Il exprime une différence relative entre la proportion d'accord observée ($P_{\text{observée}}$) et la proportion d'accord aléatoire ($P_{\text{aléatoire}}$)[Viera et Garrett, 2005].

$$Kappa = \frac{P_{\text{observée}} - P_{\text{aléatoire}}}{1 - P_{\text{aléatoire}}} \quad (1.12)$$

Lors du calcul du degré *Kappa* d'accord entre deux jugements, nous considérons les deux jugements donnés par le système ainsi que par les annotateurs. En effet, ces deux types de jugements sont résumés dans le tableau 1.2 [Ben Khiroun, 2018] comme suit :

		Jugement Système				
	Sens	1	2	…	m	Total
Jugement Annotateur	1	n_{11}	n_{12}	...	n_{1m}	L_1
	2	n_{21}	n_{22}	...	n_{2m}	L_2
	…	...	...	...	...	...
	m	n_{m1}	n_{m2}	...	n_{mm}	L_m
	Total	C_1	C_2	...	C_m	N

Tableau 1.2 : Matrice des deux jugements de sens et *m* sens possibles

La proportion des mots classés dans la diagonale de cette matrice représente la proportion d'accord observée ($P_{\text{observée}}$) calculée via la formule (1.13). Tandis que, la proportion d'accord aléatoire ($P_{\text{aléatoire}}$) est donnée par la formule (1.14) :

$$P_{observée} = \frac{\sum_{i=1}^{m} n_{ii}}{N} \quad (1.13)$$

$$P_{aléatoire} = \frac{\sum_{i=1}^{m} L_i \times C_i}{N^2} \quad (1.14)$$

Avec :

- n_{ij} est le nombre de cas jugés en relation avec les sens i par les annotateurs et en relation avec le sens j par le système. Les deux jugements sont en accord total si $i = j$ et en désaccord si $i \neq j$.
- N est le nombre de cas à désambiguïser pour un mot donné.
- m est le nombre de sens possibles existants dans un dictionnaire.

La mesure *Kappa* tient compte de l'accord produit par hasard et est considéré comme une valeur raffinée. Landis et Koch (1977) ont suggéré des significations (ou interprétations) des valeurs de *Kappa* résumés dans le tableau 1.3 [Ben Khiroun, 2018].

Kappa	**Significations (interprétations)**
< 0	Désaccord
[0,00 - 0,20]	Accord très faible
[0,21 - 0,40]	Accord faible
[0,41 - 0,60]	Accord modéré
[0,61 - 0,80]	Accord fort
[0,81 - 1,00]	Accord presque parfait

Tableau 1.3 : Signification des valeurs *Kappa* de Cohen

En utilisant cette métrique, la tâche d'évaluation doit être effectuée avec précaution, car l'évaluation humaine de sens est difficile à juger. En fait, Véronis (2003) a mentionné que la pertinence cognitive n'avait jamais été nécessaire. Dans l'une de ses expériences, Véronis a prouvé que l'homme possède

une faible précision lors de l'association d'un sens d'un dictionnaire à l'apparition d'un mot dans un communiqué. D'autre part, Edmonds et Hirst (2002) ont été parmi les rares auteurs qui ont remarqué que l'apparition d'un mot peut avoir plusieurs sens possibles, sans être en mesure de les distinguer. Ce phénomène, appelé *indétermination*, est cependant largement lié à l'expressivité de la langue.

Dans le but de renforcer l'évaluation de nos approches et d'assurer des comparaisons objectives avec les travaux de [Segond, 2000], nous avons utilisé d'autres métriques d'évaluations telles que le *Rappel*, la *Précision* et le *F-Mesure*. En fait, ces métriques sont principalement recommandées dans l'évaluation des SRI. Dans notre cas, ces métriques sont calculées comme suit :

$$Rappel = \frac{Les\ sens\ récupérés}{Total\ des\ sens\ en\ référence} \quad (1.15)$$

$$Précision = \frac{Les\ sens\ récupérés}{Total\ des\ sens\ proposés} \quad (1.16)$$

$$F - Mesure = \frac{2*Rappel*Précision}{(Rappel+Précision)} \quad (1.17)$$

6.3. Résultats et étude comparative

Nous comparons dans la section 6.3.1 les deux méthodes d'apprentissage du DSC, à base de dictionnaire et à base des jugements en utilisant la métrique *Accord*. Dans la section 6.3.2, nous évaluons et nous comparons nos résultats en utilisant la métrique *Kappa*.

6.3.1. Résultats en utilisant la métrique *Accord*

Les résultats obtenus dans les figures 1.2 et 1.3 montrent que l'accord moyen est élevé lorsqu'il s'agit d'un mot fréquent dans le corpus et possède quelques sens possibles. En effet, nous signalons que les verbes sont les moins fréquents dans le corpus, et en conséquence ils sont les plus ambigus. Par contre, les noms sont les plus fréquents dans le corpus ROMANSEVAL ; c'est pour cette raison qu'ils sont moins ambigus [Ben Khiroun et al., 2012].

Généralement, les scores des *Accords* moyens dépendent des natures des corpus utilisés. Notons ici que les textes de ROMANSEVAL s'intéressent particulièrement aux domaines politiques et économiques discutés au Parlement Européen. A titre d'exemple, le nom « *constitution* » possède six sens possibles, à savoir : *constitution*, *mise en place*, *incorporation*, *règle*, *habitude* et *code*. Il est doté d'une valeur faible du score d'*Accord* moyen en comparaison avec les autres noms. Par contre, le nom « *économie* » possède quatre sens différents, à savoir : *économie*, *finances*, *épargne* et *élevage*.

Par ailleurs, les résultats obtenus dans les figures 1.2(def) montrent que la méthode d'apprentissage du DSC à base de dictionnaire doit commencer par le traitement des phrases les plus ambiguës suivant un ordre décroissant de leurs taux d'ambiguïtés.

La figure 1.3 récapitule et compare les résultats des *Accords* moyens pour les trois catégories grammaticales (Adjectifs, Noms et Verbes) en utilisant les deux méthodes d'apprentissage à base de jugements et à base de dictionnaire. En effet, la méthode d'apprentissage du DSC à base de jugements semble meilleure que celle à base de dictionnaire, étant donné qu'elle a profité des connaissances provenant des évaluations manuelles des juges dans la phase de l'étiquetage du ROMANSEVAL. Par contre, la méthode d'apprentissage du DSC à base de dictionnaire est semi-automatique. En conséquence, elle pourra être prometteuse en cas d'absence d'une annotation sémantique du corpus utilisé.

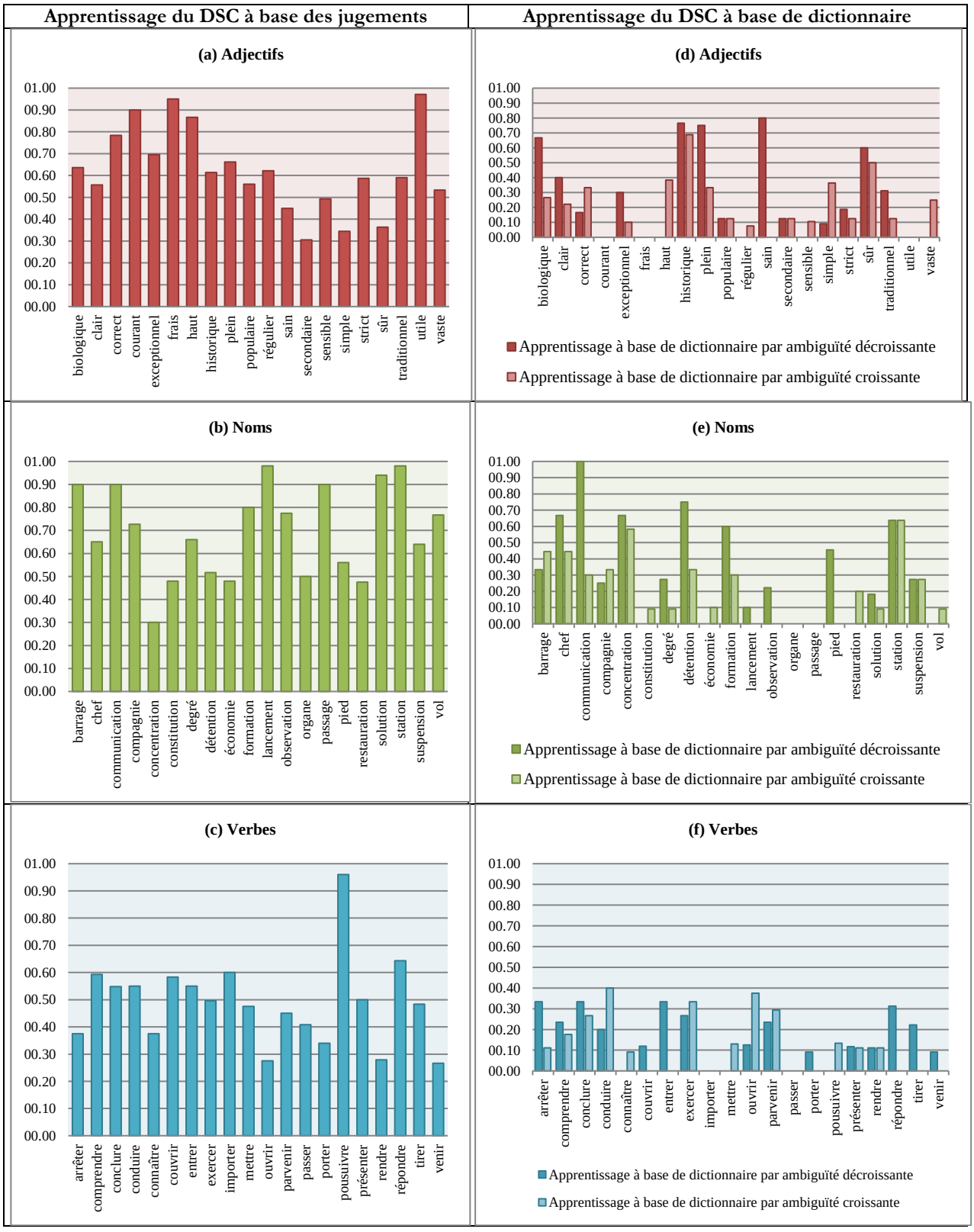

Figure 1.2 : Comparaison des méthodes d'apprentissage du DSC

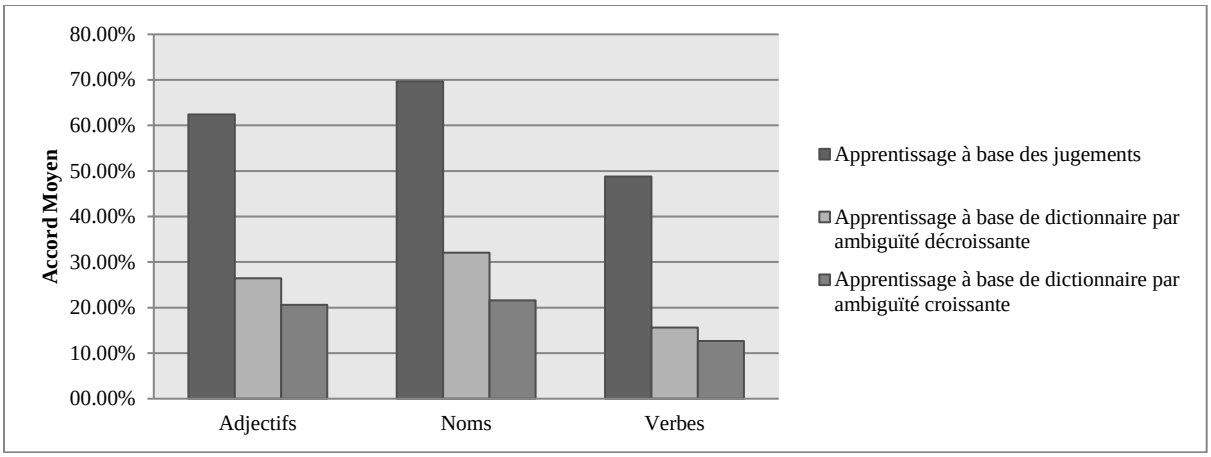

Figure 1.3 : L'Accord moyen des méthodes d'apprentissage du DSC

6.3.2. Résultats en utilisant la métrique *Kappa*

La métrique *Kappa* est la statistique la plus couramment utilisée à cet effet. Un *Kappa* de 1 indique une concordance parfaite, tandis qu'un *Kappa* de 0 indique un accord équivalent au hasard [Viera et Garrett, 2005]. Une étude comparative, en fonction de la métrique *Kappa*, est détaillée dans les figures 1.4, 1.5 et 1.6. Les résultats dans ces figures comparent nos deux approches de désambiguïsation possibiliste (POSS) et probabiliste (PROBA) aux résultats du système XEROX; un système utilisant la même collection de test ROMANSEVAL pour la désambiguïsation monolingue française [Segond, 2000].

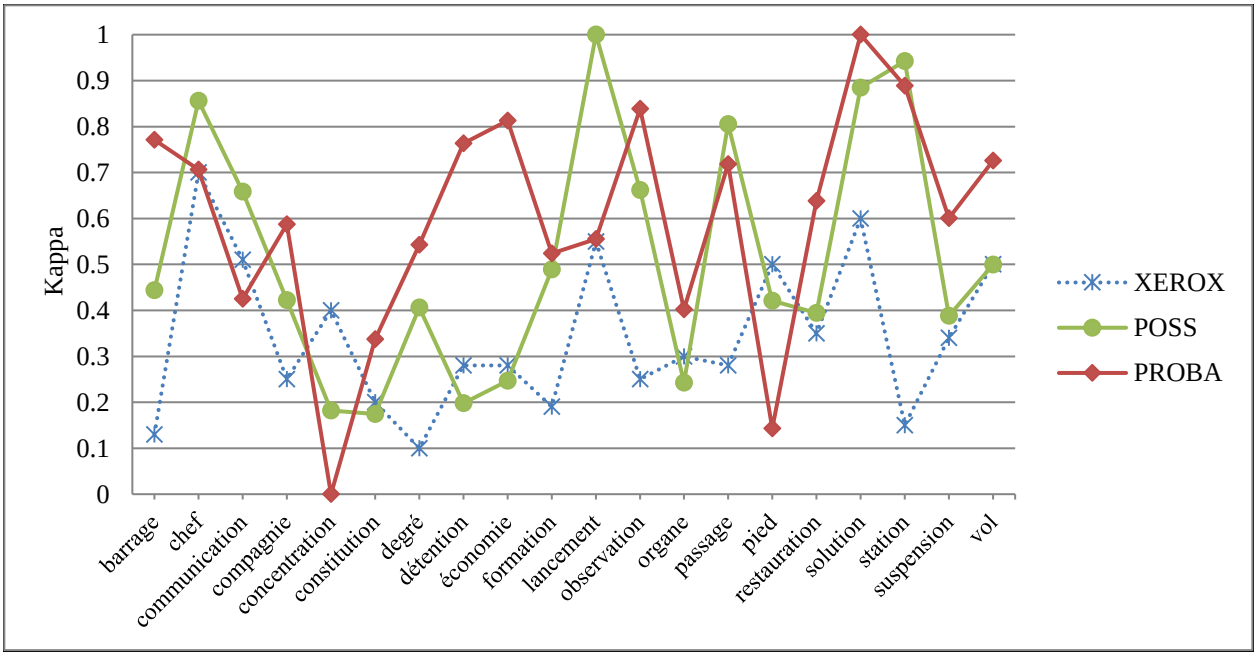

Figure 1.4 : Comparaison des *Kappa* pour la désambiguïsation des Noms

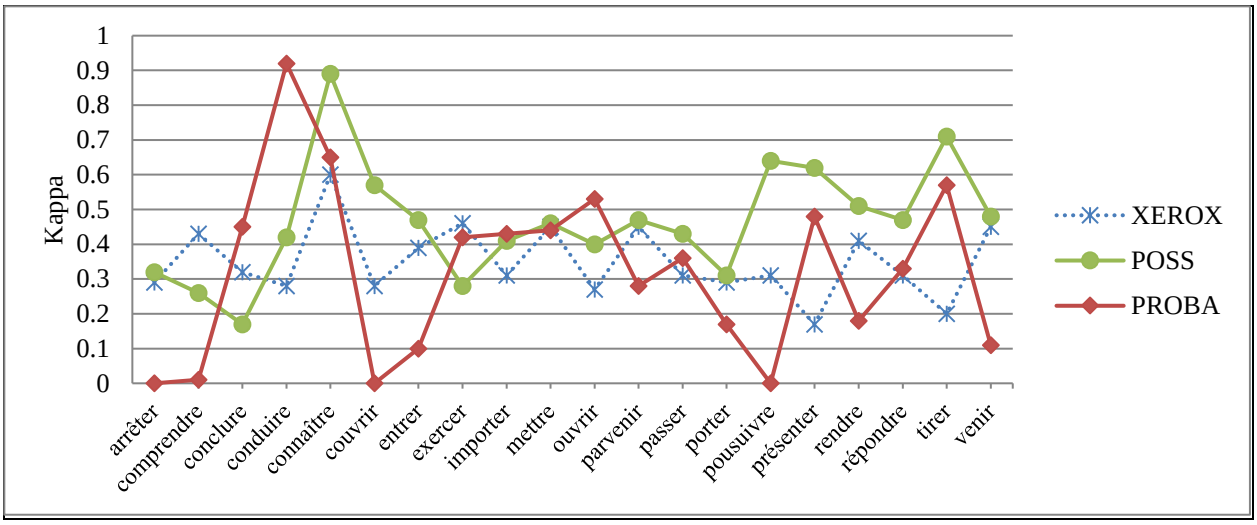

Figure 1.5 : Comparaison des *Kappa* pour la désambiguïsation des Verbes

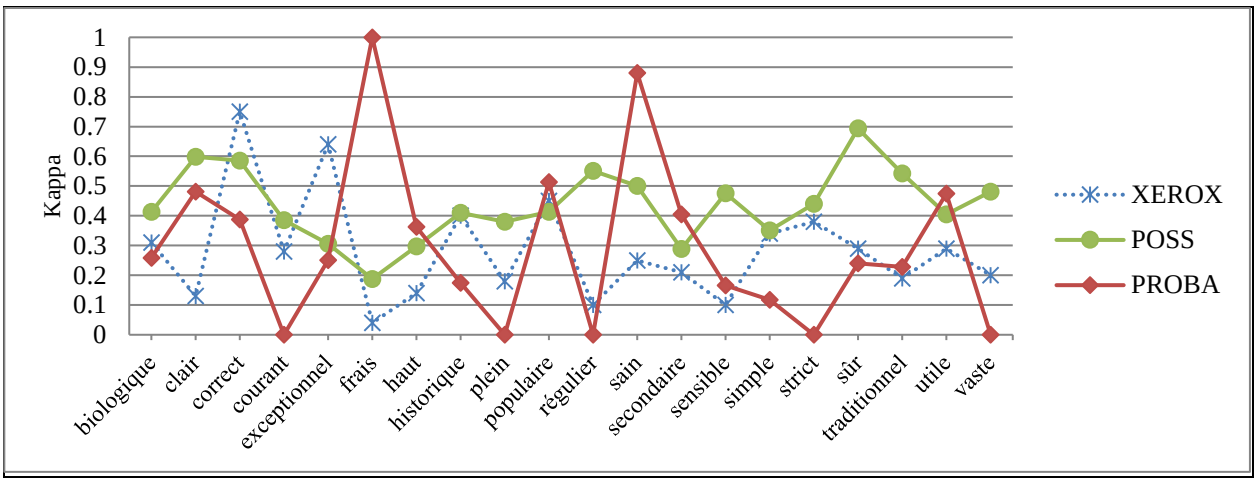

Figure 1.6 : Comparaison des *Kappa* pour la désambiguïsation des Adjectifs

Dans la figure 1.4, nous remarquons une valeur de *Kappa* inférieure à 0.2. Ceci est considéré comme un "*léger accord*", selon l'échelle de [Krippendorff, 1980]. D'autres mots tels que "*concentration*", "*couvrir*", "*vaste*" ont des valeurs nulles de *Kappa*. Dans les figures 1.7 et 1.8, nous comparons les performances de nos deux approches possibiliste et probabiliste avec cinq autres systèmes de désambiguïsation monolingue participant à l'exercice français [Segond, 2000]. En fait, pour avoir une étude comparative objective de nos approches, nous nous sommes limités à ces cinq systèmes de désambiguïsation monolingue datés de l'année 2000. Toutefois, et selon nos connaissances, ces systèmes sont les plus récents dans les littératures qui s'intéressent à la langue française et qui ont été évalués en utilisant la même collection de test ROMANSEVAL. Par exemple, nous ne pouvons pas se comparer aux travaux de [Brun et al., 2001], parce qu'ils ont utilisé la collection "LeMonde94" pour évaluer leurs résultats. Les systèmes étudiés dans [Segond, 2000] ont été développés respectivement par l'**EPFL** (Ecole Polytechnique Fédérale de Lausanne), **IRISA** (Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes), **LIA-BERTIN** (Laboratoire d'Informatique, Université d'Avignon, et BERTIN, Paris), et **XEROX** (Centre de recherche Xerox Europe, Grenoble). Une étude comparative entre ces systèmes est disponible dans [Segond, 2000].

La figure 1.7 montre les valeurs de la métrique *Kappa* de ces cinq systèmes ainsi que nos deux approches POSS et PROBA.

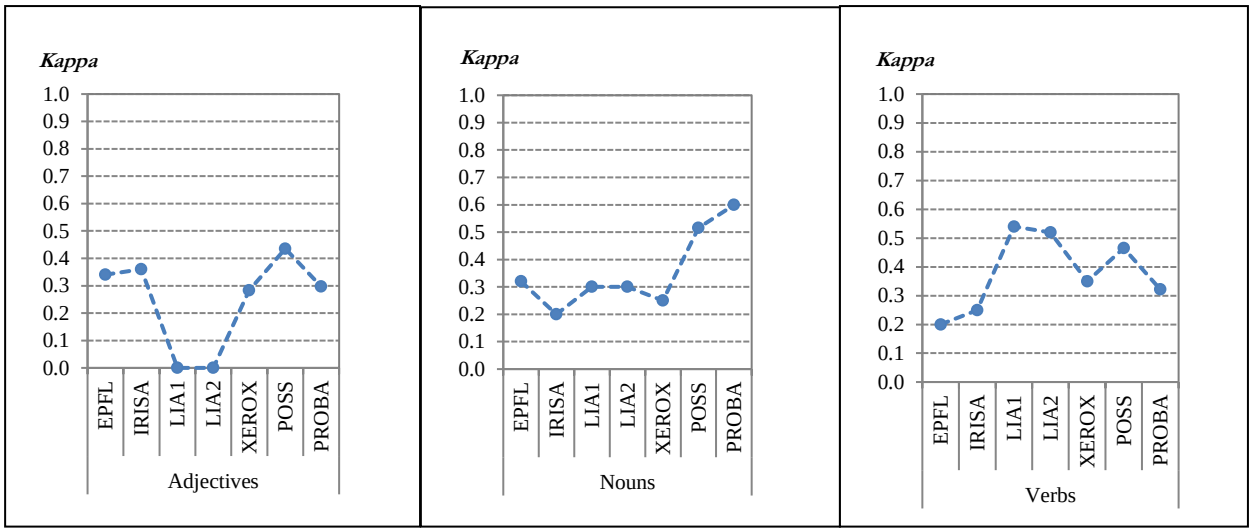

Figure 1.7 : Comparaison des *Kappa* des cinq systèmes par catégorie grammaticale

Pour les adjectifs, l'approche POSS est meilleure que tous les autres systèmes y compris l'approche PROBA. Mais, pour les noms l'approche PROBA semble meilleure que toutes les autres y compris l'approche POSS, qui est aussi plus performante que les cinq autres systèmes. Pour les verbes, l'approche POSS est meilleure que l'EPFL, IRISA, PROBA et XEROX, mais légèrement moins performante que LIA1 et LIA2.

Si nous focalisons notre attention sur les résultats globaux pour toutes les catégories grammaticales (cf. figure 1.8), notre approche possibiliste (POSS) se distingue des cinq autres systèmes ainsi que de l'approche PROBA pour les valeurs *Kappa* (EPFL : 0.29; IRISA : 0.27 ; LIA1 : 0.28; LIA2 : 0.27; XEROX : 0.29; PROBA : 0.41). En fait, l'accord entre notre système et d'autres juges n'est pas par hasard selon une valeur *Kappa* modérée (POSS : 0.47).

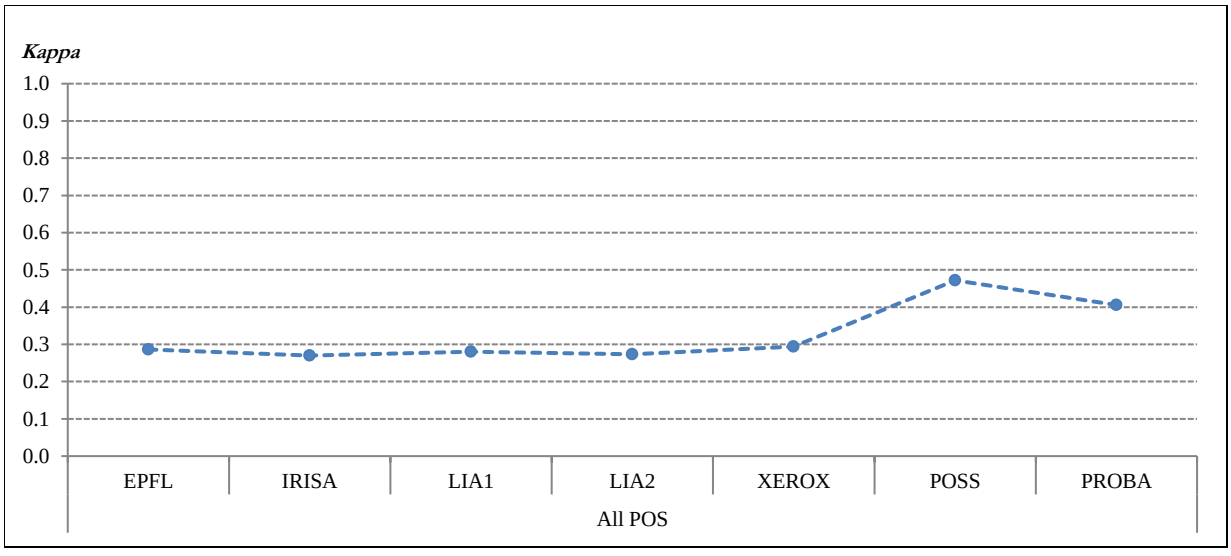

Figure 1.8 : Comparaison des *Kappa* des résultats globaux

Notons ici que le désaccord entre les juges humains qui ont préparé les sens étiquetés de ROMANSEVAL est si important selon [Véronis, 1998]: *Kappa* qu'il se situe entre 0.92 (pour le nom "détention") et 0.007 (pour l'adjectif "correcte"). En d'autres termes, il n'y a plus d'accord que le hasard pour certains mots. Si les annotateurs humains ne sont pas d'accord à propos de plusieurs mots, il semble que les systèmes, qui produisent aléatoirement des étiquettes de sens pour ces mots, doivent être considérés comme satisfaisants. Ce phénomène est bien remarqué dans le domaine de la désambiguïsation en raison du fait que les annotateurs humains ont également tendance à être en désaccord [Vidhu Bhala et Abirami, 2012].

Afin de renforcer notre étude comparative entre l'approche possibiliste et les deux approches, probabiliste et Xerox, en termes de résultats de la moyenne *Kappa*, nous utilisons le test de Wilcoxon (*Matched-Pairs Signed-Ranks Test*) proposé par [Demsar, 2006]. En effet, ce test est une alternative non-paramétrique de *t*-test qui nous permet de comparer deux approches (POSS vs. Xerox et POSS vs. PROBA) pour chaque catégorie grammaticale (adjectif, nom et verbe), et pour toutes les catégories grammaticales ensemble (All POS). Les valeurs indiquées (*p-valeur*) sont calculées en comparant les moyennes *Kappa* de l'approche possibiliste à leurs correspondants des deux autres approches probabiliste et Xerox.

Les résultats de la comparaison du tableau 1.4 [Elayeb et al., 2015a] montrent que l'approche possibiliste est toujours nettement meilleure que l'approche de Xerox pour chaque catégorie grammaticale (*p-valeur* < 0.05). De plus, l'approche POSS est fortement meilleure que Xerox pour toutes les catégories grammaticales (All POS) ; parce que la valeur de $p = 0.000007$ est très faible comparée à 0.05. En outre, l'approche POSS semble être meilleure que PROBA pour les adjectifs et les verbes, mais ce n'est pas la même chose pour les noms ($p = 0.184992 > 0.05$) [Véronis, 1998]. Cependant, les résultats sont presque significatifs lorsque l'on compare l'approche POSS avec l'approche PROBA pour toutes les catégories grammaticales ($p = 0.052847$). Ces résultats sont certainement influencés par les effets non significatifs des noms.

		POSS vs. XEROX	**POSS vs. PROBA**
Kappa	**Adjectifs**	0.004550	0.033340
	Noms	0.011220	0.184992
	Verbes	0.016852	0.019569
	Toutes les catégories grammaticales (All POS)	0.000007	0.052847

Tableau 1.4 : Les résultats de *p-valeur* pour le test de Wilcoxon

En fait, l'évaluation quantitative détaillée ci-dessus de nos deux approches POSS et PROBA en utilisant la métrique *Kappa* doit encore être renforcée en termes d'utilisation d'autres métriques d'évaluation telles que le rappel, la précision et la F-mesure; généralement recommandé pour la RI. En effet, la combinaison de rappel et précision a été initialement utilisée comme étant les principaux paramètres d'évaluation de performance dans les exercices *SensEval/SemEval*.

Le tableau 1.5 [Elayeb et al., 2015a] donne une étude comparative de nos résultats en utilisant rappel, précision et F1 pour les adjectifs (A), les noms (N) et les verbes (V) entre nos deux approches et les cinq systèmes de désambiguïsation monolingue existants dans [Segond, 2000]. En outre, le tableau 1.6 donne les résultats pour les deux approches POSS et PROBA pour toutes les catégories grammaticales (All POS).

	Adjectifs (A)			Noms (N)			Verbes (V)		
	Rappel	**Précision**	**F1**	**Rappel**	**Précision**	**F1**	**Rappel**	**Précision**	**F1**
EPFL	0.54	0.56	0.549	0.51	0.52	0.514	0.40	0.39	0.394
IRISA	0.69	0.61	**0.647**	0.55	0.48	0.512	0.29	0.28	0.284
LIA1	0.00	0.00	-	0.75	0.64	**0.690**	0.88	0.71	0.785
LIA2	0.00	0.00	-	0.76	0.63	0.688	0.89	0.72	**0.796**
Xerox	0.56	0.48	0.516	0.45	0.43	0.439	0.31	0.29	0.299
POSS	0.54	0.57	**0.554**	0.63	0.66	**0.644**	0.44	0.47	**0.454**
PROBA	0.49	0.53	0.509	0.59	0.63	0.609	0.40	0.44	0.419

Tableau 1.5 : Résultats des Rappel, Précision et F1 pour A, N et V

Les résultats du tableau 1.5 montrent que l'IRISA effectue les meilleurs scores pour les adjectifs, juste avant l'approche possibiliste. Cependant, les systèmes de LIA1 et LIA2 n'étaient pas efficaces pour les adjectifs avec des scores nuls. Par contre, LIA1 et LIA2 ont les meilleurs scores pour les noms et les verbes, juste avant l'approche possibiliste, qui semble être meilleure que les quatre autres systèmes.

	POSS			PROBA		
	Rappel	**Précision**	**F1**	**Rappel**	**Précision**	**F1**
Toutes les catégories grammaticales (All POS)	0.543	0.570	0.556	0.507	0.546	0.526

Tableau 1.6 : Résultats des Rappel, Précision et F1 pour toutes les catégories grammaticales

Sur un autre versant, et en utilisant le rappel, la précision et la F1, l'approche possibiliste effectue de meilleurs résultats pour toutes les catégories grammaticales (All POS) que celle probabiliste (cf. tableau 1.6). En fait, la métrique *Kappa* peut ne pas refléter seulement la performance réelle de l'approche de désambiguïsation; car il est utile dans la normalisation de la précision, en corrigeant le résultat de l'accord estimé avec le classifieur idéal par hasard [Cohn, 2003]. Ces métriques d'évaluation se complètent les unes les autres afin de parvenir à une évaluation objective des approches de désambiguïsation. C'est pourquoi nous proposons d'utiliser tous ces paramètres dans l'évaluation de nos deux méthodes de désambiguïsation.

7. Bilan des contributions et perspectives

Nous avons présenté dans ce chapitre l'essentiel de nos contributions dans le domaine de la désambiguïsation sémantique monolingue, qui est considérée comme l'une des tâches les plus difficiles dans le domaine du traitement sémantique [Navigli, 2009 ; Elayeb, 2018]. En effet, nous avons proposé deux approches de désambiguïsation sémantique monolingue une possibiliste et une probabiliste que nous avons appliquées aux textes français pour les évaluer et les comparer. L'originalité de ces approches consiste à combiner les dictionnaires traditionnels et un corpus étiqueté pour construire et mettre à jour un dictionnaire sémantique des contextes (DSC). Tout d'abord, nous avons utilisé un réseau possibiliste pour quantifier la pertinence d'un sens de mot ambigu étant donné une phrase polysémique. Cette pertinence est modélisée par une double mesure. Si la *pertinence possible* permet de rejeter les sens non-pertinents d'un mot ambigu, la *pertinence nécessaire* permet de renforcer la pertinence des sens restants non-éliminés par la possibilité. Par la suite, nous avons exploité une *distance probabiliste* existante pour proposer une nouvelle approche probabiliste de désambiguïsation sémantique monolingue. Cette approche calcule une distance sémantique entre les mots du dictionnaire en prenant en compte la topologie complète de ce dernier, considéré comme un graphe sémantique sur ses entrées.

Afin d'évaluer et comparer nos deux approches avec les systèmes existants de désambiguïsation monolingues, nous avons effectué des expériences sur la collection de test ROMANSEVAL. Nous avons également résumé et discuté les résultats des différents tests effectués selon une analyse détaillée et globale. Dans nos premières expériences, nous avons exploité la métrique *Accord* afin de comparer deux méthodes d'apprentissage du DSC, une à base de jugements des experts et l'autre à base d'un dictionnaire. Tandis que, nos secondes expériences, utilisant la métrique *Kappa*, ont montré une amélioration encourageante en termes de taux de précision de désambiguïsation de mots français. L'approche possibiliste de désambiguïsation (POSS) donne de meilleurs résultats que Xerox pour les adjectifs, les noms, les verbes ainsi que toutes les catégories grammaticales (All POS). Mais, l'approche probabiliste de désambiguïsation (PROBA) semble être meilleure que l'approche possibiliste et Xerox pour les noms. Cela peut expliquer pourquoi certains chercheurs ont tenté de développer des approches spécifiques à des catégories grammaticales et des corpus étiquetés pour la désambiguïsation (par exemple, l'approche de [Brown et al., 2011] pour les verbes). Pour toutes les catégories grammaticales et selon les expériences effectuées, la possibiliste semble être meilleure que toutes les autres approches. Ces résultats révèlent la contribution de la théorie des possibilités, car elle a fourni de bons taux de précision dans cette première expérience. Des expériences utilisant rappel, précision et F-mesure ont confirmé que l'approche possibiliste effectue globalement de meilleurs résultats que Xerox et l'approche PROBA pour les adjectifs, les noms, les verbes ainsi que toutes les catégories grammaticales. Elle semble également meilleure que tous les autres systèmes de désambiguïsation monolingue existant dans [Segond, 2000].

En outre, l'approche possibiliste est plus fine que celle probabiliste et les autres systèmes de désambiguïsation monolingue existant dans la littérature et utilisant le même standard de test ROMANSEVAL. Cela s'explique par le fait que les deux mesures de possibilité et de nécessité augmentent davantage la pertinence des sens corrects, en pénalisant au même temps les scores de ceux qui restent. En fait, la pénalisation et l'augmentation des scores sont proportionnelles à la capacité des mots polysémiques à faire la distinction entre les différents sens de la collection [Elayeb et al., 2015a]. Cependant, le modèle de correspondance possibiliste ne prend pas en compte les relations entre les mots du contexte dans le réseau possibiliste utilisé. De nombreuses solutions sont à étudier telles que : (i) Améliorer ce modèle afin de tenir compte de ces types de relations; ou (ii) Combiner ce modèle avec une distance sémantique telle que PROX. Nous avons déjà fait une première expérience dans [Elayeb et al., 2011] pour combiner les scores possibilistes avec ceux à base de dénombrement des circuits dans un graphe de termes afin de profiter des avantages procurés par chacun d'entre eux.

D'autre part, le dictionnaire sémantique des contextes a été utilisé comme une seconde entrée avec le dictionnaire traditionnel pour nos deux approches de désambiguïsation. Nous avons montré que l'utilisation du DSC améliore les résultats du calcul de sens de mot dans un contexte donné, et permet par conséquence d'obtenir de bonnes performances en les comparants aux systèmes existants de désambiguïsation monolingue. Il permet également d'analyser les résultats de désambiguïsation ; car il représente des informations explicitement contextuelles. Ainsi, nous sommes en mesure de comprendre l'origine des décisions effectuées par les systèmes de désambiguïsation.

Afin de diminuer la taille de l'espace sémantique et par conséquent améliorer le temps de réponse de notre plate-forme de désambiguïsation, nous pensons proposer une fonction permettant de déterminer l'espace sémantique d'un mot polysémique (c'est-à-dire la taille de la fenêtre contextuelle) de manière dynamique selon le DSC. Une autre perspective est liée à l'intégration de cette plate-forme de désambiguïsation dans certains systèmes possibilistes de recherche intelligente

d'information (SRI) tels que SARIPOD [Elayeb, 2009], SPORSER [Elayeb et al., 2011] et SPORT [Ben Romdhane et al., 2013]. En fait, en s'inspirant de plusieurs travaux récents qui ont montré le rôle de la désambiguïsation de requêtes en RI [Soto et al., 2008; Barathi et Valli, 2010], nous avons évalué l'impact du processus de désambiguïsation de requêtes sur leurs expansions dans un SRI intelligent [Ben Khiroun et al., 2014a]. En effet, nous avons profité de nos deux approches de désambiguïsation monolingue détaillées dans ce chapitre afin d'améliorer l'expansion de requêtes et en conséquence de la performance globale du SRI. D'autre part, l'objectif à court terme de notre travail est d'améliorer la performance d'un système de recherche d'information translinguistique en introduisant une étape de désambiguïsation de requêtes et de documents dans un contexte translinguistique [Elayeb et Bounhas, 2016 ; Ben Romdhane et al., 2017 ; Elayeb et al., 2018]. Ainsi, ce travail pourra être étendu à d'autres langues telles que l'arabe et l'anglais. Pour l'anglais, nous pouvons utiliser les standards de tests disponibles gratuitement dans [Koeling et al., 2005]. Néanmoins, la tâche est plus complexe pour la désambiguïsation de l'arabe, car nous avons besoin de structurer les dictionnaires arabes bruts ainsi que de construire et/ou chercher un standard de test pertinent pour la langue arabe. Certains travaux connexes [Zouaghi et al., 2012 ; Khemakhem et al., 2013] révèlent des solutions prometteuses pour structurer les dictionnaires arabes utiles à la tâche de désambiguïsation. De plus, nos outils et structures de données sont des composants réutilisables qui peuvent être intégrés dans d'autres domaines tels que l'extraction d'information, la traduction automatique, l'analyse du contenu, le traitement et l'organisation de la terminologie, la lexicographie et les applications du Web sémantique. Dans le chapitre suivant, nous présentons une évaluation des approches de classification possibiliste pour la désambiguïsation morphologique des textes arabes.

Chapitre 2 :
Désambiguïsation morphologique des textes arabes

1. Introduction

De nombreux mots arabes possèdent la même forme orthographique. Ceci est dû à la richesse morphologique de cette langue [Diab et al., 2004]. En effet, l'omission des voyelles courtes peut générer plus de 12 interprétations morphologiques d'un mot donné [Habash et Rambow, 2007]. Par conséquent, la forme d'ambiguïté la plus relevée en arabe réside dans la dimension morphologique. Un mot peut être ambigu à l'égard de sa structure interne. Le traitement morphologique porte sur le morphème qui constitue l'unité élémentaire discernable. L'analyse morphologique d'un mot a pour rôle de déterminer les valeurs d'un grand nombre de caractéristiques ou d'attributs morphologiques d'une entité lexicale (un mot), comme la catégorie grammaticale (nom, verbe, etc.), le genre, le nombre, etc. En fait, un mot non-voyellé peut conduire à de nombreuses solutions morphologiques. Par exemple, le mot وقف (wqf), en dehors du contexte, peut être interprété comme وَقَفَ (waqafa, "il s'est levé") ou وَقْفٌ (waqfun, "cession") ou encore وَقِفْ (waqif, "et lève-toi"), où ce mot est une concaténation de la conjonction و (wa "et") avec le verbe قِفْ "se lever" qui est conjugué à l'impératif. Malgré leur importance, les voyelles courtes ne sont utilisées que dans les textes religieux (Coran, Hadith) et les manuels didactiques, contrairement aux textes modernes trouvés dans les journaux et dans les livres [Ayed, 2017 ; Elayeb, 2018].

L'ambiguïté morphologique se manifeste lorsque l'analyse associe, à une unité lexicale, plusieurs informations non-conformes au contexte du mot, autrement dit quand l'analyse fournit plusieurs valeurs pour certains attributs morphologiques [Hajic, 2000]. Par ailleurs, une approche pour la désambiguïsation morphologique arabe est nécessaire pour faire face à l'ambiguïté des mots non-voyellés. La désambiguïsation consiste, donc, à identifier la valeur exacte d'un attribut morphologique parmi celles proposées par l'analyseur. De nombreux travaux utilisent des approches de classification pour résoudre la tâche morphologique de désambiguïsation [Roth et al., 2008].

La désambiguïsation morphologique d'un mot arabe consiste à identifier l'analyse morphologique appropriée correspondante à ce mot. Dans ce chapitre, nous présentons trois modèles de désambiguïsation morphologique de textes arabes non-voyellés basés sur la classification possibiliste. Cette approche traite les données imprécises dans les phases d'apprentissage et de test, étant donné que notre modèle apprend à partir de données non étiquetées. Pour discriminer l'effet de chaque attribut, nous affectons des pondérations aux attributs dans l'ensemble d'apprentissage.

Nous testons nos approches sur deux corpus, à savoir le corpus du Hadith et le Treebank arabe. Ces corpus contiennent des données de types différents classiques et modernes. D'abord, nous comparons nos modèles avec des classifieurs non-possibilistes concurrents. Pour ce faire, nous transformons la structure des ensembles d'apprentissage et de test pour remédier au problème d'imperfection des données. Ensuite, nous présentons les résultats concernant tous les attributs morphologiques afin de déterminer à quel degré cette approche discriminative améliore le taux de désambiguïsation et extrait les relations des dépendances entre les attributs. Les résultats révèlent la contribution de la théorie des possibilités pour résoudre les ambiguïtés dans les applications réelles. Nous comparons également les taux de réussite dans les deux cas des textes arabes modernes et classiques. Enfin, nous évaluons l'impact de la probabilité lexicale sur la désambiguïsation morphologique.

Le présent chapitre est organisé comme suit : Tout d'abord, dans la section 2, nous présentons brièvement un état de l'art sur la désambiguïsation morphologique arabe. Nos approches pour la désambiguïsation morphologique possibiliste sont détaillées dans la section 3. Les résultats expérimentaux sont présentés et discutés dans la section 4. Nous concluons, dans la section 5 et nous proposons quelques pistes pour d'éventuelles recherches futures.

2. Synthèse des approches existantes de désambiguïsation morphologique

Plusieurs travaux conduisent la désambiguïsation des mots arabes d'un texte à l'identification de leurs catégories grammaticales (POS : *part-of-speech*). La désambiguïsation de POS est le fait de déterminer la catégorie grammaticale d'un mot étant donné son utilisation dans un contexte particulier. Elle peut, également, être considérée comme un problème de classification : l'ensemble des valeurs de POS présentent les classes et une méthode de classification est utilisée pour attribuer, à chaque occurrence d'un mot (analyse d'un mot), une classe sur la base de la certitude du contexte. L'une des étapes importantes dans la désambiguïsation est la sélection de la méthode de classification. Des méthodes de classification automatique supervisée ont été appliquées. Elles utilisent des techniques d'apprentissage pour apprendre un classifieur à partir des ensembles d'apprentissage annotés (les valeurs de la classe POS sont identifiées). Dans la littérature, les approches de désambiguïsation, se répartissent en trois catégories. Principalement, ces approches sont : les approches à base de règles, les approches statistiques et les approches hybrides qui combinent les deux dernières [Ayed, 2017 ; Elayeb, 2018] (cf. Figure 2.1).

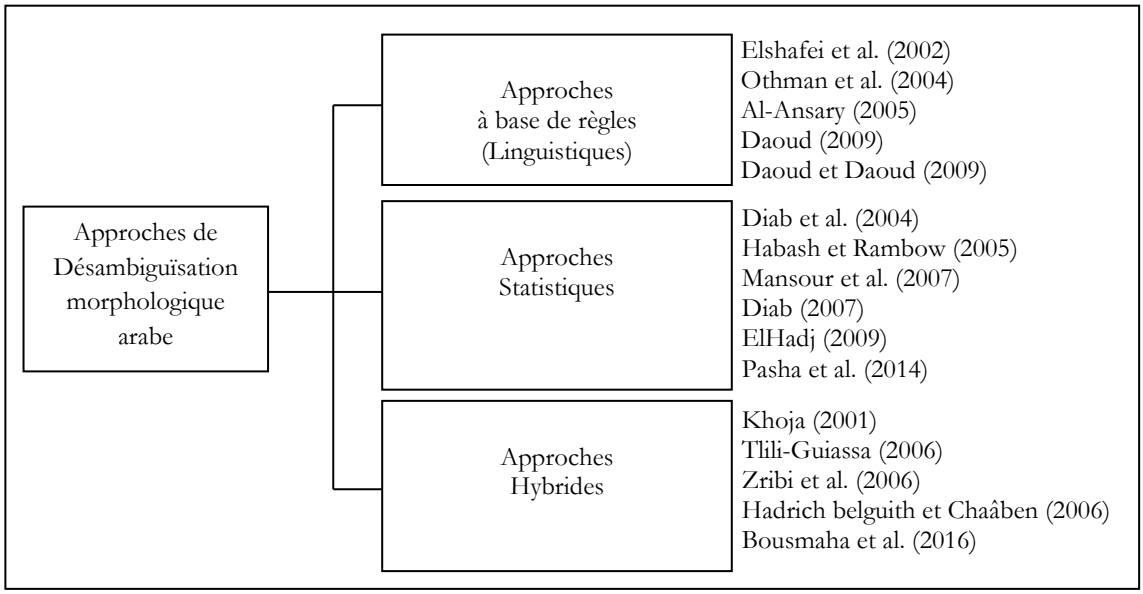

Figure 2.1 : Approches existantes de désambiguïsation morphologique arabe

Les approches à base de règles sont encore dites linguistiques. Elles utilisent une base de connaissances des règles écrites par des linguistes permettant d'attribuer des étiquettes aux différentes catégories morphologiques [Othman et al., 2004 ; Daoud, 2009]. Nous parlons principalement des heuristiques, des règles contextuelles et des règles non-contextuelles [Elshafei et al., 2002]. Notons que ces règles ont été classées dans [Al-Ansary, 2005] en catégories logiques, grammaticales et structurales. En outre, Daoud et Daoud (2009) ont suggéré un type particulier d'analyseurs connus par En-Converters (EnCo) développés via le langage de programmation à base

de règles UNL (Universal Networking Language). En effet, les auteurs ont profité des combinaisons des dépendances contextuelles syntaxiques et morphologiques afin de définir un ensemble de règles de désambiguïsation. Toutefois, il s'est avéré qu'il est difficile d'évaluer leur analyseur en termes de précision, couverture et réutilisabilité à cause de l'absence de la validation expérimental de l'outil proposé. Tandis que, Othman et al. (2004) ont profité de l'analyse syntaxique complète pour réussir leur approche de désambiguïsation morphologique.

Les approches statistiques forment des modèles d'apprentissage à partir des corpus annotés. Elles incorporent des méthodes de classification telles que les modèles de Markov cachés, SVM, etc. pour calculer des taux de probabilité de chaque valeur résultante d'une catégorie grammaticale d'un mot. Un modèle peut être utilisé pour classer automatiquement les autres textes en se référant aux taux déjà calculés. Diab et al. (2004) ont développé un classifieur morphologique utilisant SVM. Ils ont entrainé et testé le classifieur sur un Treebank arabe de 4000 phrases d'apprentissage et 100 phrases de test. Habash et Rambow (2005) ont utilisé SVM en se basant sur des informations fournies à partir d'un analyseur morphologique. D'abord, les auteurs ont proposé le désambiguïseur morphologique arabe MADA (atteint 86% de précision). Ensuite, les deux outils MADA [Habash et Rambow, 2005] et AMIRA [Diab, 2007] ont été combinés par [Pasha et al., 2014] pour donner naissance à un nouveau outil d'analyse et de désambiguïsation morphologiques connu par MADAMIRA.

Par ailleurs, Mansour et al. (2007) ont combiné les probabilités calculées sur des ensembles d'apprentissage arabes et hébreux pour classer les catégories grammaticales des mots des textes arabes. Ils ont utilisé les mêmes paramètres de test de [Diab et al., 2004]. Quelques travaux de recherches existants comprennent les modèles de Markov cachés (HMM). Par exemple, ElHadj et al. (2009) ont présenté un système d'étiquetage grammatical qui combine l'analyse morphologique et le modèle de Markov. L'étiqueteur se base sur la structure de la phrase arabe. Dans un premier lieu, le texte est entièrement analysé morphologiquement pour réduire le nombre de valeurs possibles de POS. Dans un second lieu, le modèle statistique (HMM), fondé sur la structure de la phrase arabe, est utilisé pour attribuer à chaque mot la valeur exacte de sa catégorie grammaticale. ElHadj et al. (2009) ont utilisé leur propre corpus annoté qui est composé de vieux livres arabes. Le total des mots dans ce corpus est environ 21000 mots.

Une approche hybride combine les informations statistiques avec les règles linguistiques dans l'objectif de réussir la désambiguïsation morphologique des textes arabes. Khoja (2001) a proposé et testé une approche hybride à base de l'algorithme de Viterbi [Forney, 1973; Fettweis et Meyr, 1991]. Cette méthode est à base du calcul de deux probabilités sur un corpus annoté composé de 50000 mots : (i) une probabilité contextuelle, qui est la probabilité d'une étiquette à suivre une autre, et (ii) une probabilité lexicale, qui est la probabilité qu'un mot ait une certaine valeur d'un attribut morphologique spécifié. Une liste de règles grammaticales a été générée à partir de ces statistiques afin d'atteindre plus de 90% de précision.

Par ailleurs, Zribi et al. (2006) ont mixé l'approche à base de règles avec un étiqueteur trigramme HMM [Collins, 2002]. L'apprentissage du classifieur trigramme a été fait sur des textes contenant 6000 mots. Des règles heuristiques ont été appliquées pour sélectionner une analyse parmi les résultats suggérés. Tlili-Guiassa (2006) a suggéré une méthode qui analyse les affixes grammaticaux et flexionnels et les règles grammaticales en se basant sur l'approche MBL (*Memory-based learning*) [Lin et al., 1994]. L'objectif étant de classer une collection de textes éducatifs et coraniques.

Hadrich Belguith et Chaâben (2006) ont proposé et testé une approche d'analyse et de désambiguïsation morphologiques des textes arabes. Malgré cette approche est à base de règles, elle

est considérée comme une approche statistique. Les auteurs ont proposé un scénario de cinq étape afin d'assurer la désambiguïsation morphologique. D'abord, une étape de segmentation du texte est suivie d'une phase de prétraitement morphologique permettant de supprimer les clitiques en utilisant une liste prédéfini. Puis, une étape d'analyse affixale permettant d'identifier les affixes et les racines des mots. Ensuite, une analyse morphologique utilisant l'outil MORPH2 [Hadrich Belguith et Chaâben, 2006]. Enfin, une phase d'identification des groupes des mots en utilisant un ensemble de règles et un lexique. Mais, parmi ses limites, cette approche permet de d'assurer le calcul des attributs morphologique de chaque mot arabe en utilisant un lexique réduit. D'autre part, Bousmaha et al. (2016) ont suggéré et testé une approche hybride de désambiguïsation combinant une décision multicritère avec une approche linguistique. L'approche est basée sur la sélection des diacritiques à différent niveaux d'analyse et pourra être utilisé dans la désambiguïsation morpho-lexicale. Les auteurs ont profité de l'analyseur morphologique en ligne et ils ont enregistré des résultats ayant plus que 80% de F-mesure.

Hoceini et al. (2011) ont confirmé que les outils de désambiguïsation linguistiques sont plus rapides et plus efficaces et fiables que les outils statistiques. En fait, les deux approches statistiques et hybrides exigent une phase d'apprentissage dans l'objectif d'apprendre les paramètres requis pour la désambiguïsation. L'approche hybride est considérée comme la plus efficace et cohérente en termes d'analyse, car elle combine les deux approches et tire profit de leurs avantages. Notons que les statistiques calculées pour l'apprentissage sont appliquées à n'importe quel domaine de test. Tandis que, l'approche linguistique, qui n'exige que de l'intervention manuelle d'un linguiste, définit un ensemble de règles particulières à un domaine spécifique.

Les analyseurs et désambiguïseurs morphologiques arabes de la littérature ne tiennent pas compte des données incertaines et/ou imprécises. Tandis que, la théorie des possibilités est dédiée à résoudre ces deux problèmes d'incertitude et d'imprécision des données utilisées. C'est pour ces raisons que nous proposons dans ce chapitre une approche d'analyse et de désambiguïsation morphologiques des textes arabes à base des classifieurs possibilistes, et qui tient compte de l'incertitude et de l'imprécision des instances de classification afin de pallier aux limites des classifieurs existants. D'autre part, la plupart des désambiguïseurs morphologiques arabes ne traitent que la catégorie grammaticale (POS). Les travaux récents [Habash et al., 2009 ; Ayed et al., 2012b, 2014ab ; Ayed, 2017] définissent 14 attributs qui décrivent les caractéristiques morphologiques d'un mot. Nous étendons dans ce chapitre la classification à ces 14 attributs morphologiques.

3. Approches possibilistes de désambiguïsation morphologique

Nous proposons des approches de désambiguïsation morphologique des textes arabes basées sur la théorie des possibilités. Plusieurs travaux utilisent les approches de classification pour résoudre l'ambiguïté morphologique [Habash et Rambow, 2005]. Un mot est considéré ambigu si l'analyseur morphologique fournit plus d'une seule solution pour ses attributs morphologiques. La classification assigne une classe à une instance de test donnée. La tâche de désambiguïsation consiste donc à accorder à un mot ambigu les valeurs des attributs morphologiques appropriées. Elle est divisée en deux grandes phases qui sont l'apprentissage et le test. Les résultats d'analyse morphologique donnés par les mots voyellés sont généralement moins ambigus que ceux donnés par les mots non-voyellés. Ainsi, nous proposons d'apprendre à partir des textes voyellés et de tester sur des textes non-voyellés.

Pour ce faire, nous commençons par définir l'ensemble d'apprentissage. Cet ensemble est constitué d'une liste d'instances qui sont caractérisées par des attributs avec des valeurs de classes connues. Par

conséquent, pour résoudre l'ambiguïté de la catégorie grammaticale (par exemple), nous déterminons d'abord les attributs appropriés qui décrivent chaque instance. En nous inspirant de la technique de classification Yamcha [Diab et al., 2004], nous estimons qu'un attribut morphologique d'un mot est fortement lié à celui des mots qui le précèdent ainsi que ceux qui le suivent. Nous définissons une fenêtre qui contrôle le nombre de mots (avant et après) considérés comme des attributs décrivant la classe d'une instance. Dans des approches existantes, la taille de la fenêtre est 2 [Habash et Rambow, 2005]. Notre modèle applique une fenêtre avec une taille quelconque. Pour classer la catégorie grammaticale (POS : *part-of-speech*) d'un mot particulier, si la fenêtre est de 2, nous définissons les attributs POS-2, POS-1, POS+1 et POS+2 [Bounhas et al., 2015bc]. Ils indiquent, respectivement, les catégories grammaticales des deux mots précédents et des deux mots suivants. POS peut être décrit par l'ensemble des autres attributs morphologiques, en plus du POS. Nous pouvons utiliser, par exemple, les attributs genre-2, genre-1, nombre+1, nombre+2, et ainsi de suite. La valeur de la classe est la catégorie grammaticale du mot courant. A cet effet, nous identifions, pour chaque mot d'un texte voyellé, 14 attributs morphologiques qui sont *POS, conjonction, particule, déterminant, pronom, personne, voix, aspect, genre, nombre, cas, préposition, mode* et *adjectif*. Ces attributs sont calculés par l'analyseur morphologique *Aramorph* [Ayed et al., 2012b]. Ayant l'exemple de la phrase suivante :

الرَازِي وَالْبَغْدَادِي دَرَسَا عُلُومَ الطِّبِ (Al-Razi et Al-Bagdadi ont étudié les sciences de la médecine), nous déterminons, l'instance du tableau 2.1, associée au mot « دَرَسَا (ont étudié) ». Pour cette instance, la classe est la catégorie grammaticale (POS) et les attributs utilisés sont les catégories grammaticales des 2 mots adjacents.

POS-2	POS-1	POS+1	POS+2	POS
NOM_PROPRE	NOM_PROPRE	NOM	NOM	VERBE

Tableau 2.1 : L'instance d'apprentissage reliée au mot « دَرَسَا »

L'analyse morphologique d'un mot est fournie indépendamment de son contexte. Dans un texte arabe, même les mots voyellés peuvent donner une analyse morphologique ambiguë. La forme voyellée «إِبْنِ » fournit des valeurs de l'attribut POS à savoir un verbe (tu construis) et un nom (fils de). Par conséquent, les instances d'apprentissage peuvent fournir des informations incomplètes. Ces informations sont dites imprécises lorsque les attributs et/ou la classe donnent plus d'une seule valeur.

Nous pouvons affirmer clairement que le contexte nécessaire pour lever l'ambiguïté d'un mot donné est lui-même ambigu ce qui est considéré comme un cas d'imprécision. En effet, la théorie des probabilités est incapable de traiter un tel type de données imprécises, alors que la théorie des possibilités s'applique naturellement à ces problèmes. Nous proposons des modèles d'apprentissage et de test (classification) basés sur la théorie des possibilités.

3.1. L'apprentissage possibiliste des attributs morphologiques

Dans la phase d'apprentissage, nous formons une classification pour chaque attribut morphologique. Autrement dit, nous instaurons un ensemble d'apprentissage pour chaque attribut morphologique. Nous obtenons globalement 14 ensembles. Chacun est décrit par les attributs AM $\pm$ *i* où AM forme la totalité des attributs morphologiques et *i* constitue la taille de la fenêtre. Si cette taille est égale à 2, nous obtenons 56 (14x4) attributs d'apprentissage. A chaque mot voyellé est liée une instance décrite par les valeurs de ces 56 attributs dont la classe est reconnue. Cette classe est l'attribut morphologique associé à l'ensemble d'apprentissage.

Nous devons prendre en compte le fait que les attributs et/ou les classes des instances de classification sont imprécis ; autrement dit, ils ont plus d'une seule valeur possible. L'imprécision est gérée par des distributions de possibilités désignées par π. Soit T un ensemble de données d'apprentissage et I_k l'ensemble des valeurs des attributs de l'instance k. On note également A_j le $j^{ième}$ attribut de cet ensemble et a_{jL} une valeur possible d'A_j. Nous nous inspirons des travaux de [Haouari et al., 2009] et le modèle de recherche d'information possibiliste développé par [Bounhas et al., 2011b] pour calculer la fréquence normalisée d'une valeur d'un attribut a_{jL} pour une classe c_i comme suit [Bounhas et al., 2015bc] :

$$Freq\left(a_{jL}, c_i\right) = \frac{Occ(a_{jL}, c_i)}{Max_{L=1}^{|A_j|} Occ(a_{jL}, c_i)} \quad (2.1)$$

$\mathrm{Occ}(\mathrm{a_{jL}}, \mathrm{c_i})$ indique le nombre d'occurrences de la classe c_i avec la valeur a_{jL}; c'est-à-dire le nombre d'instances dont la classe est égale à c_i et la valeur a_{jL} est une valeur possible de l'attribut A_j. $|A_j|$ est le nombre de valeurs possibles d'A_j. Nous utilisons l'opérateur *Max* pour obtenir les fréquences normalisées [Bounhas et al., 2011bc]. La somme de toutes les fréquences associées à une classe c_i n'est pas égale à 1 ce qui est l'une des principales hypothèses de la théorie des possibilités afin de traiter des données imparfaites. Dans le cas de l'imperfection des données, le nombre d'occurrences d'une valeur d'un attribut est flou. Nous introduisons une mesure β_{jk} appelée le taux de l'imprécision de l'attribut A_j dans l'instance I_k [Haouari et al., 2009]. Le nombre d'occurrences est calculé suivant la formule (2.2) [Bounhas et al., 2015bc] :

$$Occ\left(a_{jL}, c_i\right) = \sum_{k=1}^{|T|} \beta_{jk} * \emptyset_{ijkL} \quad (2.2)$$

Le taux $\beta_{jk} = \frac{1}{|A_{jk}| * |C_k|}$; où $|A_{jk}|$ représente le nombre de valeurs de l'attribut A_j dans l'instance I_k et $|C_k|$ le nombre de classes possibles de I_k. Si l'instance est parfaite, alors $\beta_{jk} = 1$. Si dans une instance donnée, un attribut possède deux valeurs et la classe n'a qu'une seule valeur, alors le taux de l'imprécision est égal à 0.5. $\emptyset_{ijkL}$ est égale à 1 si la valeur a_{jL} appartient aux valeurs possibles de A_j dans l'instance I_k, et la classe c_i appartient aux valeurs de classes de I_k ; et 0 ailleurs.

Les fréquences normalisées sont calculées pour la totalité des instances des différents ensembles d'apprentissage. Elles traduisent les distributions de possibilités de chaque attribut par rapport à une classe.

3.2. La classification possibiliste des attributs morphologiques

La classification des 14 attributs morphologiques consiste à désambiguïser chaque mot non-voyellé en lui associant les valeurs correctes et précises de ces attributs. Pour ce faire, nous commençons par préparer les instances de l'ensemble de test. En effet, chaque instance décrit un mot non-voyellé d'un texte par des attributs de classification qui représentent les mêmes attributs d'apprentissage ; c'est-à-dire $AM \pm i$. La classe de l'instance est la valeur correcte à identifier de l'attribut morphologique. Le tableau 2.2 décrit une instance de test dont l'attribut morphologique à classer est le POS. Pour simplifier la représentation de l'instance, nous nous contentons de 4 attributs de classification à savoir DET-2, POS-1, CONJONCTION-1 et POS+2. Elle est réellement décrite par les 56 attributs. Cette instance est imprécise puisqu'elle donne deux valeurs possibles de l'attribut POS-1.

DETERMINANT-2	POS-1	CONJONCTION-1	POS+2	…	POS
DET	{VERBE; NOM_PROPRE}	NCONJ	NOM	…	?

Tableau 2.2 : Un exemple d'une instance de test imprécise

Nous calculons la possibilité de chaque classe c_i par rapport à une instance imparfaite I_k ayant m attributs. Cette mesure s'inspire du classifieur possibiliste de [Haouari et al., 2009]. La mesure de possibilité est le produit des fréquences de tous les attributs calculés par rapport à l'ensemble d'apprentissage. Cependant, un facteur spécifique est ajouté pour les attributs imprécis. Ce facteur est le taux de l'imprécision β_{jk}. Par exemple, si un attribut a quatre valeurs possibles, nous calculons le produit des fréquences de ces quatre valeurs et nous introduisons le taux β_{jk} égal à ¼. Ainsi, la mesure de possibilité est donnée par la formule (2.3) [Bounhas et al., 2015bc].

$$\Pi(c_i|I_k) = \prod_{j=1}^{m} \prod_{L=1}^{|A_{jk}|} Freq\,(a_{jL}, c_i) * \beta_{jk} \quad (2.3)$$

$$socre(c_i|I_k) = \Pi(c_i|I_k) \quad (2.4)$$

En se référant à l'instance du tableau 2.2, si la classe POS possède trois valeurs possibles ; i.e. NOM, VERBE et NOM_PROPRE, alors trois mesures de possibilités sont à calculer par rapport à cette instance.

Ces mesures sont : $\Pi(\text{POS} = \text{NOM}|I_k)$, $\Pi(\text{POS} = \text{VERBE}|I_k)$ et $\Pi(\text{POS} = \text{NOM\_PROPRE}|I_k)$. Pour déterminer la mesure $\Pi(\text{POS} = \text{NOM}|I_k)$ les fréquences nécessaires sont *Freq*(DETERMINANT-2 = DET, POS = NOM), *Freq*(POS-1 = VERBE, POS = NOM), *Freq*(POS-1 = NOM_PROPRE, POS = NOM), etc. Ces fréquences sont calculées dans la phase d'apprentissage.

3.3. Le classifieur possibiliste discriminatif

Notre classifieur possibiliste de base, inspiré des travaux de [Haouari et al., 2009], a été défini dans [Ayed et al., 2012a] et n'évalue pas le pouvoir discriminant des valeurs d'un attribut, car il utilise uniquement la mesure de possibilité (formule (2.3)). Cependant, nous pouvons découvrir que certaines valeurs, d'un attribut donné, ont un plus grand impact dans la résolution de la bonne classe. La théorie des possibilités modélise cet effet par la mesure de nécessité. Elle détermine le degré auquel on attend l'occurrence d'un événement [Elayeb et al., 2009]. C'est pour cette raison que plusieurs travaux existants [Bounhas et al, 2011b; Elayeb, 2009; Elayeb et al., 2009 ; 2011 ; 2015ab] ont utilisé cette mesure pour évaluer le pouvoir discriminant des termes de la requête dans la sélection des documents pertinents ; c'est-à-dire les termes qui n'existent que dans quelques documents ont un impact plus important dans l'évaluation de la pertinence.

Cette mesure est donnée par la formule (2.5) [Bounhas et al., 2015b] :

$$N(c_i|I_k) = 1 - \prod_{j=1}^{m} \prod_{L=1}^{|A_{jk}|} \left(1 - \frac{\lambda_{ijL} * Freq\,(a_{jL}, c_i)}{\beta_{jk}}\right) \quad (2.5)$$

Où : $\lambda_{ijL} = log_{10}(\frac{P}{nC_{jL}})$

Avec : P est le nombre de classes possibles et nC_{jL} est le nombre de classes ayant une fréquence non nulle avec la valeur a_{jL} ou en d'autres termes $Freq(a_{jL}, c_i) > 0$.

La nécessité peut être directement utilisée pour la classification comme suit :

$$\text{socre}(c_i|I_k) = N(c_i|I_k) \quad (2.6)$$

Ainsi, nous combinons les deux mesures de possibilité et de nécessité :

$$\text{socre}(c_i|I_k) = \Pi(c_i|I_k) + N(c_i|I_k) \quad (2.7)$$

Dans nos expériences, nous comparons les trois alternatives adoptant respectivement les formules (2.4), (2.6) et (2.7) afin d'identifier la meilleure classe. Dans tous les cas, la meilleure classe de l'instance I_k est titulaire du plus grand score parmi toutes les classes :

$$c^* = \arg\max_{c_i}(\text{score}\ (c_i|I_k)) \quad (2.8)$$

3.4. Le modèle de repondération

Le modèle de repondération permet d'attribuer des poids absolus aux attributs de classification. Les outils de désambiguïsation, comme MADA [Habash et Rambow, 2005; 2007; Roth et al., 2008; Alkuhlani et al., 2013], implémentent de tels modèles pour évaluer relativement ces attributs et améliorer davantage les taux de classification. Ceci permet de réduire l'espace contextuel nécessaire pour désambiguïser un attribut donné et en conséquence simplifier le processus de désambiguïsation.

Dans notre cas, nous calculons le *gain informationnel* basé sur l'entropie [Blansché, 2006], qui a été utilisé dans de nombreux domaines de l'intelligence artificielle comme la sécurité dans les services en ligne [Krause et Horvitz, 2008], le clustering [Liping et Huang, 2007] et la classification [Yue, 2012]. L'utilisation de l'entropie dans la classification est justifiée par sa capacité à partitionner l'espace en des régions de décision non-chevauchantes pour la classification. C'est exactement ce qui est nécessaire pour la désambiguïsation morphologique dans le cas des différentes valeurs d'une même classe (par exemple : nom, verbe, etc. pour l'attribut POS) en permettant de partitionner les mots en des ensembles complètement séparés. En outre, Lee et al. (2001) affirment que "*la procédure de sélection de l'attribut basée sur l'entropie ne réduit pas seulement la dimension d'un problème, mais aussi défausse les attributs à bruit corrompu, redondant et sans importance*". L'entropie peut être utilisée pour pondérer les attributs. Cependant, nous étions obligés d'adapter les formules pour le cas imparfait. La principale modification consiste à introduire des facteurs permettant d'attribuer des poids inférieurs aux attributs ou classes qui sont imparfaits (i.e. ayant plusieurs valeurs possibles). Ce changement est une contrainte imposée par le type d'application, à savoir la tâche de désambiguïsation. Les exemples dans [Bounhas et al., 2015b] expliquent et montrent l'importance de ces facteurs.

Soit S un ensemble donné d'instances. Si nous avons P classes, alors S sera segmentée en P sous-ensemble $S= \{S_1,\ldots, S_P\}$; où les instances de chaque S_i appartiennent à la même classe.

L'entropie d'information de S, dans le cas idéal, est donnée par [Blansché, 2006] :

$$I(S) = -\sum_{i=1}^{P} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \quad (2.9)$$

Nous supposons que l'attribut A_j possède v valeurs différentes. L'entropie requis pour classer les instances de tous les sous-ensembles de T, selon l'attribut A_j, est donnée par :

$$E(A_j) = \sum_{L=1}^{v} \frac{\left|\{I_k \in T | A_j = a_{jL}\}\right|}{|T|} * I(I_k \in T\ |A_j = a_{jL}) \quad (2.10)$$

Dans le cas imprécis, il n'est pas possible de calculer directement $\left|\{I_k \in T | A_j = a_{jL}\}\right|$. Il est également nécessaire d'adapter la formule (2.10) afin de calculer $I(I_k \in T\ |A_j = a_{jL})$. En effet, nous ne pouvons

pas attribuer une valeur binaire à $A_j = a_{jL}$, parce que chaque attribut peut contenir plusieurs valeurs possibles qui sont équipondérées.

Nous procédons comme suit :

$$E(A_j) = \sum_{L=1}^{v} \frac{F_{jL}}{|T|} * I'(I_k \in T \,|A_j = a_{jL}) \quad (2.11)$$

Avec: F_{jL} est un nombre flou des instances ayant la valeur a_{jL}. Elle est donnée par:

$$F_{jL} = \sum_{k=1}^{|T|} \frac{1}{|\mathrm{A_{jk}}|} * \emptyset_{\mathrm{jkL}} \quad (2.12)$$

Avec $\emptyset_{\mathrm{jkL}}$ est égale à 1 si la valeur a_j existe dans les valeurs possibles de A_j dans l'instance I_k; et 0 ailleurs. Encore, nous divisions par $|\mathrm{A_{jk}}|$ afin de diminuer le poids de cas imprécis. Nous avons aussi:

$$I'(I_k \in T \,|A_j = a_{jL}) = -\sum_{i=1}^{P} \frac{Occ(a_{jL}, c_i)}{F_{jL}} log_2 \frac{Occ(a_{jL}, c_i)}{F_{jL}} \quad (2.13)$$

Avec $Occ(a_{jL}, c_i)$ est calculée selon la formule (2.2). Le gain informationnel [Blansché, 2006] de l'attribut A_j est calculé comme suit :

$$Gain\,(A_j) = I'(T) - E(A_j) \quad (2.14)$$

$I'(T)$ est calculé comme suit :

$$I'(T) = -\sum_{i=1}^{P} \frac{Occ(c_i)}{|T|} log_2 \frac{Occ(c_i)}{|T|} \quad (2.15)$$

Avec $Occ(c_i)$ est un nombre flou des instances ayant la classe c_i. Elle est donnée par :

$$Occ(c_i) = \sum_{k=1}^{|T|} \frac{1}{|\mathrm{C_k}|} * \emptyset_{\mathrm{ik}} \quad (2.16)$$

Avec $\emptyset_{\mathrm{ik}}$ est égale à 1 si la valeur de c_i existe dans les classes possibles de l'instance I_k; et 0 ailleurs.

Enfin, nous introduisons le gain informationnel dans les mesures de possibilité et de nécessité en modifiant les formules (2.3) et (2.5) de la manière suivante :

$$\Pi(\mathrm{c_i}|\mathrm{I_k}) = \prod_{\mathrm{j=1}}^{\mathrm{m}} \prod_{\mathrm{L=1}}^{|\mathrm{A_{jk}}|} Freq\,(a_{jL}, c_i) * \beta_{\mathrm{jk}} * Gain\,(A_j) \quad (2.17)$$

$$N(c_i|\mathrm{I_k}) = 1 - \prod_{j=1}^{m} \prod_{\mathrm{L=1}}^{|\mathrm{A_{jk}}|} \left(1 - \frac{\varphi_{ijL}}{\beta_{\mathrm{jk}} * Gain\,(A_j)}\right) \quad (2.18)$$

Où : $\varphi_{ijL} = log_{10}(\frac{P}{nC_{jL}}) * Freq\,(a_{jL}, c_i)$

3.5. La possibilité lexicale

Chaque instance unique des ensembles d'apprentissage et de test décrit les valeurs des attributs d'un mot. Ainsi, le classifieur possibiliste ne tient pas compte de la valeur du mot pour déterminer son analyse précise (classe). Par conséquent, l'implication de valeur du mot devient cruciale dans le calcul du score possibiliste (cf. formule (2.8)). Nous introduisons le score lexical *score*(c_i|w_k) inspiré de [Jurafsky et Martin, 2009]. Cette mesure calcule le degré de dépendance d'un mot w_i avec une classe particulière c_i dans l'ensemble d'apprentissage. Si w_k est le mot de l'instance de test I_k, alors la possibilité lexicale répond à la question : si nous nous attendions à ce que c_i soit la classe de I_k, quelle

serait donc la possibilité pour que le mot soit w_i? De même, ce score peut être calculé de trois manières différentes en utilisant la possibilité, la nécessité ou la somme des deux mesures.

La classe choisie correspond à la valeur c^*. La meilleure classe de l'instance I_k est celle ayant le plus grand score parmi toutes les classes [Bounhas et al., 2015b].

$$c^* = arg\max_{c_i}(score\ (c_i|I_k) * score(c_i|w_k)) \quad (2.19)$$

Le score lexical $score(c_i|w_k)$ est positif dans le cas de la coexistence d'un mot w_k avec une classe particulière c_i dans l'ensemble d'apprentissage. Sinon, la valeur de c^* de la formule (2.19) sera zéro. Ainsi, cette mesure ne sera considérée que lorsque cette condition de coexistence est remplie. Si le score lexical $score(c_i|w_k)$ est égale à zéro, nous ignorons sa valeur et nous calculons c^* en utilisant la formule (2.8).

3.6. La classification non-possibiliste des attributs morphologiques

Nous visons à comparer les résultats de la classification possibiliste avec les résultats donnés par des classifieurs non-possibilistes afin de désambiguïser les attributs morphologiques. Ces classifieurs ne traitent pas les données imparfaites. Par conséquent, nous proposons de transformer la structure des données des ensembles d'apprentissage et de test afin de les préparer pour qu'elles soient utilisées par des classifieurs non-possibilistes. Les nouveaux attributs doivent donner des informations précises. Pour ce faire, nous commençons par présenter un ensemble de données imparfaites. La figure 2.2(a) donne un exemple d'un ensemble d'apprentissage.

POS-1	CONJONCTION+1	POS
NOM	NCONJ	{NOM_PROPRE ; VERBE}
{NOM ; VERBE}	CONJ	NOM

(a) Instances imparfaites

POS-1_NOM	POS-1_VERBE	CONJONCTION+1_CONJ	CONJONCTION+1_NCONJ	POS
Oui	Non	Non	Oui	{NOM_PROPRE; VERBE}
Oui	Oui	Oui	Non	NOM

(b) Instances dont les attributs sont précis et les classes sont incertaines

POS-1_NOM	POS-1_VERBE	CONJONCTION+1_CONJ	CONJONCTION+1_NCONJ	POS
Oui	Non	Non	Oui	NOM_PROPRE
Oui	Non	Non	Oui	VERBE
Oui	Oui	Oui	Non	NOM
Oui	Oui	Oui	Non	NOM

(c) Instances parfaites

Figure 2.2 : Transformation des instances imparfaites en des instances parfaites

Nous supposons que la classe à désambiguïser est POS et que les attributs utilisés, pour l'apprentissage et le test, sont POS-1 et CONJONCTION+1. Cet ensemble est composé de deux instances. La première instance est imprécise, car elle fournit deux valeurs possibles de la classe (NOM_PROPRE et VERBE). La deuxième instance est imprécise puisqu'elle fournit deux valeurs de l'attribut POS-1 (NOM, VERBE). Nous transformons la structure de données afin d'obtenir un ensemble parfait sans perdre les informations qui s'y trouvent. Pour résoudre le problème de l'imprécision, nous désignons les valeurs de l'attribut A par $A_i = \{a_1, a_2, ..., a_n\}$. Nous constituons de

nouveaux attributs. En effet, nous associons l'attribut A à chacune de ses valeurs a_i pour former des attributs notés "A_a_i". Ainsi, l'attribut POS-1 a deux valeurs possibles (NOM et VERBE) dans l'ensemble des données de la figure 2.2. Nous obtenons donc deux attributs POS-1_NOM et POS-1_VERBE. Nous accordons, aux nouveaux attributs, des valeurs binaires (Oui ou Non). Pour une instance donnée, si a_i appartient à l'une des valeurs de l'attribut A, alors l'attribut "A_a_i" est égal à "Oui". A partir des données de la figure 2.2(a), nous formons un nouvel ensemble de données précises de figure 2.2(b).

Pour résoudre l'imprécision des classes, nous décomposons une instance en plusieurs parties ayant chacune une seule valeur de la classe. Si une instance possède n valeurs possibles de la classe $\{c_1, c_2, \ldots, c_n\}$, alors nous obtenons n instances dont les valeurs des attributs sont similaires. Nous associons à chaque instance une valeur c_i.

Les instances dont la classe est précise (ayant une seule valeur) seront dupliquées afin d'augmenter leurs poids dans le calcul des scores de classification. La figure 2.2(c) présente un ensemble de données parfaites générées à partir des instances de la figure 2.2(a). Pour désambiguïser des textes non-voyellés moyennant les approches non-possibilistes, nous utilisons les méthodes SVM, le modèle Bayésien naïf et les arbres de décision. Nous alignons les données au format d'entrée du software *Weka*[10]. Cet outil fournit des algorithmes d'apprentissage automatique et donne leurs résultats de classification. Nous utilisons *Weka* pour classer les attributs morphologiques selon les modèles SVM, les arbres de décision et le modèle Bayésien naïf.

4. Résultats expérimentations

Dans [Ayed et al., 2012ab], nous avons présenté nos résultats préliminaires qui ont prouvé un taux de réussite élevé pour notre classifieur possibiliste de base. Ils nous ont permis d'étudier l'indépendance du domaine de notre classifieur. D'abord, nous décrivons dans cette section les corpus utilisés pour nos expérimentations. Ensuite, nous présentons la méthode d'évaluation et les résultats expérimentaux mettant en évidence les aspects de classification possibiliste et non-possibiliste. Enfin, nous évaluons aussi le classifieur possibiliste discriminatif, le modèle de repondération ainsi que l'effet de l'implication de la possibilité lexicale.

4.1. Les collections de test

L'objectif principal de notre approche est d'acquérir des dépendances morphologiques à partir des textes voyellés et de tester sur des textes non-voyellés. En outre, nous considérons les textes arabes classiques, qui ont été ignorés dans des travaux connexes précédents. Par conséquent, nous utilisons une collection d'histoires arabes "Hadiths" qui a fait le sujet de plusieurs travaux [Bounhas et al., 2010 ; 2011ab ; 2015ab ; Harrag et al., 2013 ; Ayed et al., 2012 ; 2014ab ; Ben Khiroun et al., 2014c], etc. Les Hadiths parlent de toutes les préoccupations du monde réel et couvrent des connaissances communes et universelles. Pour justifier notre choix, nous estimons que le corpus de Hadiths est l'un des rares corpus arabes voyellés. Il contient environ 1400 livres voyellés de Hadith, chacun comporte des milliers d'histoires arabes. Les six livres les plus reconnus comprennent plus de 2.5 millions de mots et plus de 95000 fragments (titres et paragraphes). Par ailleurs, ce corpus est bien structuré et les titres des chapitres et des sous-chapitres représentent des informations contextuelles pertinentes pour désambiguïser des textes [Bounhas et al., 2011ab]. Parmi les textes du corpus de Hadiths, nous utilisons six livres encyclopédiques, regroupés par thèmes, qui sont : Sahih Al-Bukhari (صحيح البخاري),

[10]http://weka.wikispaces.com/

Sahih Muslim (صحيح مسلم), Sunan Abi Dawud (سنن أبي داود), Sunan Ettermidhi (سنن الترمذي), Sunan Ibn Majah (سنن ابن ماجة) et Sunan Annasaii (سنن النسائي) [Ayed et al., 2012a].

En effet, nous limitons nos expériences aux trois sous-collections correspondantes aux trois domaines d'intérêt, à savoir (cf. Tableau 2.3) : الأشربة (Al>$rbp "boissons"), الزواج (AlzwAj "mariage") and الطهارة (AlThArp "purification") [Ayed et al., 2012a]. Nous menons nos expérimentations également sur le corpus Treebank arabe (ATB part 2 v2.0) [Maamouri et al., 2009]. Il s'agit d'un corpus de textes arabes non-voyellés qui a été produit par *Linguistic Data Consortium* (LDC). Ce corpus comprend plus de 500 articles du journal égyptien Al Oumma. Il contient environ 144K de mots annotés.

	Boissons	**Mariage**	**Purification**	**Total**
Sahih Al-Bukhari	02766	11521	11016	25303
Sahih Muslim	09117	06693	05063	20873
Sunan Abi Dawud	02672	05780	15319	23771
Sunan Ettermidhi	01835	05910	09291	17036
Sunan Ibn Majah	01748	06539	13179	21466
Sunan Annasaii	06703	08741	12554	27998
Nombre total de mots	**24841**	**45184**	**66422**	**136447**

Tableau 2.3 : Statistiques de nombre de mots dans les trois sous-collections des six livres

L'annotation indique la catégorie grammaticale de chaque mot. Les corpus utilisés présentent deux types de textes modernes et classiques. Pour pouvoir apprendre les dépendances morphologiques du Hadith, nous passons par l'analyseur morphologique des textes voyellés *Aramorph*. Cet analyseur nous fournit les valeurs des 14 attributs morphologiques. L'annotation du corpus Treebank arabe nous donne uniquement les valeurs de l'attribut POS. Le test (ou la classification) se fait directement sur les textes non-voyellés de Treebank arabe. Quant aux textes de Hadith, une étape d'élimination des voyelles courtes est indispensable pour pouvoir tester sur des textes non-voyellés. Pour évaluer les résultats des classifications possibilistes et non-possibilistes, nous utilisons la méthode de la validation croisée [Kohavi, 1995]. En effet, nous formons 10 itérations pour chaque texte du corpus : 90% d'un texte voyellé est utilisé pour l'apprentissage et 10% de mots de ce texte seront classés après avoir éliminé leurs voyelles courtes.

4.2. Evaluation de classifications possibiliste et non-possibiliste

Pour classer les 14 attributs morphologiques, nous procédons comme suit : Tout d'abord, nous analysons les textes voyellés de Hadith et nous sauvegardons les solutions morphologiques de chaque attribut. Nous formons, pour tout attribut morphologique *A*, un ensemble d'apprentissage. A chaque mot voyellé est associée une instance. Les instances de cet ensemble sont décrites par les attributs $AM \pm i$ (voir section 3.1) et la classe est l'attribut morphologique *A*. Nous aurons 14 ensembles d'apprentissage. Nous supprimons, par la suite, les voyelles courtes des mêmes textes. Nous formons de la même manière des ensembles de test décrits par les mêmes attributs que les ensembles d'apprentissage. Les valeurs de classes de leurs instances sont non reconnues (ambiguës). Elles constituent les attributs morphologiques à classer. Ensuite, nous désambiguïsons chaque mot de ces textes par le biais de nos trois modèles de classification possibiliste. Pour ce faire, nous calculons les mesures de possibilité et de nécessité en se référant aux fréquences calculées par rapport aux ensembles d'apprentissage (voir section 3). Nous comparons les résultats obtenus avec ceux donnés par les mots voyellés. Pour classer les 14 attributs morphologiques en utilisant les classifieurs non-possibilistes, nous utilisons les mêmes structures des instances d'apprentissage et de test.

Les approches non-possibilistes ne supportent pas l'imperfection des données. Nous les transformons en des données parfaites (voir section 3.6) et nous les adaptons au format d'entrée de l'outil *Weka* pour qu'elles soient appliquées sur des algorithmes de classification de SVM, Arbres de décision et les classifieurs Bayésiens Naïfs. Le tableau 2.4 présente les taux de désambiguïsation des 14 attributs morphologiques.

Les expérimentations prouvent que les classifieurs possibilistes donnent de meilleurs taux de désambiguïsation par rapport aux classifieurs SVM, Bayésien Naïf et les arbres de décision. Ils en résultent des moyennes de plus de 80% d'instances non-voyellés correctement classées. Certains attributs morphologiques donnent les mêmes résultats de classification.

Attribut morphologique	**Classifieur Bayésien Naïf**	**Arbres de décision**	**Classifieur SVM**	**Classifieur possibiliste utilisant Π**	**Classifieur possibiliste utilisant *N***	**Classifieur possibiliste utilisant (Π + N)**
POS	88.62 %	89.58 %	89.98 %	91.58 %	90.17 %	90.45 %
ADJECTIF	96.51 %	96.51 %	96.51 %	97.58 %	96.86 %	97.63 %
ASPECT	71.20 %	71.20 %	71.20 %	81.78 %	86.20 %	86.16 %
CAS	56.12 %	56.12 %	56.12 %	68.40 %	68.76 %	76.55 %
CONJONCTION	83.03 %	83.03 %	83.03 %	95.04 %	88.66 %	90.79 %
DETERMINANT	64.12 %	64.16 %	64.12 %	95.25 %	95.92 %	96.13 %
GENRE	57.15 %	57.15 %	57.15 %	93.23 %	90.45 %	93.78 %
MODE	99.32 %	99.32 %	99.32 %	99.93 %	99.96 %	99.96 %
NOMBRE	85.18 %	85.18 %	85.18 %	95.30 %	87.00 %	93.25 %
PARTICULE	96.65 %	96.65 %	96.65 %	96.91 %	98.87 %	98.87 %
PERSONNE	60.22 %	60.22 %	60.22 %	66.27 %	65.07 %	66.88 %
PREPOSITION	82.87 %	82.87 %	82.87 %	88.60 %	90.20 %	95.70 %
VOIX	71.21 %	71.21 %	71.21 %	78.75 %	78.80 %	79.05 %
PRONOM	55.02 %	55.84 %	56.88 %	59.10 %	59.56 %	58.79 %
Moyenne	**76.23 %**	**76.36 %**	**76.46 %**	**86.27 %**	**85.46 %**	**87.43 %**

Tableau 2.4 : Les taux de désambiguïsation des attributs morphologiques en utilisant les classifieurs possibilistes et non-possibilistes dans le corpus du Hadith

Ceci peut être expliqué par le fait que les attributs morphologiques associés fournissent peu de nombres de valeurs de classe (ne dépassant pas 6 chacune). D'un autre côté, l'attribut « PRONOM », par exemple, offre environ 64 valeurs de la classe qui peut générer des résultats distincts pour les différents classifieurs. Parmi les classifieurs possibilistes, nous remarquons que le modèle qui assemble les mesures de possibilité et de nécessité (Π+N) fournit de meilleurs résultats (87.43%). Ceci confirme la capacité du modèle possibiliste à traiter les données imprécises, sachant que les textes arabes ont un taux d'ambiguïté élevé.

4.3. Evaluation du classifieur possibiliste discriminatif avec modèle de repondération

Nos travaux antérieurs [Ayed et al., 2012ab] ont expérimenté notre classifieur possibiliste basé uniquement sur la mesure de possibilité afin de déterminer la classe la plus plausible. Il attribue un poids uniforme pour tous les attributs associés à la procédure de classification. Nous nous focalisons dans cette section sur notre classifieur possibiliste discriminatif ainsi que le modèle de repondération.

Nous pouvons, également, extraire les relations de dépendance entre POS et les autres attributs. Nous calculons le gain informationnel (IG_j) pour chaque attribut (A_j) étant donné la classe POS. Ces poids attestent que le POS d'un mot donné est plus lié à la POS des mots précédents et suivants par rapport aux autres attributs. Nous affirmons que, pour désambiguïser le POS, nous avons besoin d'impliquer le POS, le PRONOM et le DETERMINANT des mots précédents et suivants. Ces attributs possèdent les gains informationnels moyens les plus élevés. Cette hypothèse peut être linguistiquement prouvée.

Par ailleurs, nous calculons les trois premières valeurs de gain informationnel de chaque attribut. Ces valeurs sont affectées aux attributs les plus discriminants de chaque attribut morphologique. Nous remarquons que les attributs les plus liés sont ceux des mots adjacents (précédents et suivants). En outre, nous pouvons remarquer que tous les attributs dépendent du POS, car au moins un attribut est associé avec le POS apparaît dans la liste des plus hautes valeurs de gain informationnel. Ce fait est normal, parce que le POS détermine la catégorie grammaticale du mot dont dépendent étroitement les autres attributs. Ainsi, la relation étroite et communautaire entre les attributs POS, PRONOM et DETERMINANT. En fait, les déterminants ne sont applicables que pour les noms, et le type de pronom dépend du POS (verbe ou nom).

Attribut morphologique	**Sans repondération**					**Avec repondération**					
	Π	***N***		**Π + N**		**Π**		***N***		**Π + N**	
	DR	**DR**	**DB**	**DR**	**DB**	**DR**	**DB**	**DR**	**DB**	**DR**	**DB**
POS	90.95%	90.34%	-0.61%	90.34%	-0.61%	91.16%	**+0.21%**	90.52%	-0.43%	90.93%	-0.02%
CONJONCTION	87.92%	81.72%	-6.20%	82.74%	-5.18%	88.98%	**+1.06%**	81.67%	-6.25%	91.07%	**+3.15%**
PARTICULE	96.91%	98.87%	**+1.96%**	98.87%	**+1.96%**	98.46%	**+1.55%**	98.87%	**+1.96%**	98.87%	**+1.96%**
DETERMINANT	94.95%	95.12%	**+0.17%**	95.33%	**+0.38%**	95.12%	**+0.17%**	94.90%	-0.05%	96.93%	**+1.98%**
PRONOM	59.10%	59.56%	**+0.46%**	58.79%	-0.31%	59.56%	**+0.46%**	59.56%	**+0.46%**	59.56%	**+0.46%**
PERSONNE	65.21%	64.91%	-0.30%	65.22%	**+0.01%**	64.91%	-0.30%	64.63%	-0.58%	65.28%	**+0.07%**
VOIX	78.75%	78.80%	**+0.05%**	79.05%	**+0.30%**	79.16%	**+0.41%**	78.81%	**+0.06%**	79.11%	**+0.36%**
ASPECT	76.49%	76.89%	**+0.40%**	79.19%	**+2.70%**	77.27%	**+0.78%**	76.91%	**+0.42%**	81.30%	**+4.81%**
GENRE	92.11%	89.55%	-2.56%	93.66%	**+1.55%**	93.74%	**+1.63%**	95.05%	**+2.94%**	95.62%	**+3.51%**
NOMBRE	91.25%	86.56%	-4.69%	90.91%	-0.34%	90.78%	-0.47%	90.21%	-1.04%	92.41%	**+1.16%**
CAS	59.49%	59.57%	**+0.08%**	63.36%	**+3.87%**	59.61%	**+0.12%**	59.63%	**+0.14%**	63.52%	**+4.03%**
PREPOSITION	85.61%	85.57%	-0.04%	85.80%	**+0.19%**	85.70%	**+0.09%**	85.70%	**+0.09%**	85.80%	**+0.19%**
MODE	99.93%	99.96%	**+0.03%**	99.96%	**+0.03%**	99.96%	**+0.03%**	99.93%	00.00%	99.96%	**+0.03%**
ADJECTIF	97.58%	96.86%	-0.72%	97.63%	**+0.05%**	97.71%	**+0.13%**	97.88%	**+0.30%**	99.00%	**+1.42%**
Moyenne	84.02%	83.16%	-0.86%	84.35%	**+0.33%**	84.44%	**+0.42%**	83.88%	-0.14%	85.67%	**+1.65%**

Tableau 2.5 : Taux de réussite moyenne de désambiguïsation des attributs morphologiques de différentes combinaisons de classifieurs

Enfin, nous présentons dans le tableau 2.5 les taux de désambiguïsation (DR) de tous les attributs en utilisant les six combinaisons des classifieurs à base de Π, N et (Π+N), sans et avec repondération, justifiant l'importance de l'implication du gain informationnel dans les mesures de possibilité et de nécessité (voir formules (2.17) et (2.18)). Nous calculons aussi la différence entre les résultats des cinq derniers modèles avec le classifieur de base (DB). Par exemple, si nous passons du cas « sans repondération » de notre classifieur de base (Π) au classifieur utilisant uniquement la mesure de nécessité (N), nous perdons 0.61% du taux de désambiguïsation de l'attribut POS. Les expériences ont montré que la plupart des attributs sont étroitement interdépendants. Ils fournissent souvent plus de 70% des taux de désambiguïsation. Pour la majorité des attributs, les taux de réussite, donnés par le classifieur (Π), sont meilleurs que ceux donnés par le classifieur (N). Mais, le classifieur (Π+N) avec puis sans repondération fournit les meilleurs résultats. Cela peut s'expliquer par le phénomène d'ordre relativement aléatoire des mots dans la phrase [Bounhas et al., 2015b]. En conséquence, les valeurs d'un attribut donné (attributs des mots précédents et suivants) semblent être distribuées de façon égale sur les valeurs de la classe (attribut du mot courant).

L'utilisation de gain informationnel a augmenté les taux de désambiguïsation pour les trois mesures, sauf pour certaines valeurs de possibilité et/ou de la nécessité de certains attributs (par exemple N-

CONJONCTION, N-DETERMINANT, N-PERSONNE, Π-PERSONNE, Π-NOMBRE et N-MODE). Cette détérioration des taux est mineure et ne dépasse pas 0.47%. Les moyennes globales des taux de désambiguïsation de tous les attributs donnent des valeurs élevées lorsque nous impliquons le gain informationnel, ce qui montre l'utilité de ce modèle pour la repondération des attributs.

Pour comparer les six classifieurs en termes de taux de désambiguïsation (DR), nous utilisons le test de Wilcoxon (*Matched-Pairs Signed-Ranks Test*) proposé par [Demsar, 2006] et utilisé par Bounhas et al. (2013). Il s'agit d'un test statistique non-paramétrique permettant de comparer nos classifieurs deux-à-deux sur plusieurs attributs. En effet, nous comparant le classifieur (Π+N) avec repondération aux cinq autres classifieurs restants. Les résultats de la comparaison, donnés dans le tableau 2.6, montrent que le classifieur (Π+N) avec repondération est toujours nettement meilleur (valeur de *p-value* < 0.05) que les cinq autres classifieurs pour tous les attributs.

	***vs.* (Π+N) Sans repondération**	***vs.* (Π) Sans repondération**	***vs.* (N) Sans repondération**	***vs.* (Π) Avec repondération**	***vs.* (N) Avec repondération**
(Π+N) Avec repondération	*p = 0.003330*	*p = 0.001225*	*p = 0.003346*	*p = 0.005077*	*p = 0.002218*

Tableau 2.6 : Les *p-valeurs* de test de Wilcoxon

4.4. Etude de l'indépendance du domaine

Nous montrons l'indépendance du domaine de nos modèles possibilistes. Pour ce faire, nous menons nos expérimentations sur le corpus Treebank arabe rassemblant les textes de journaux. Ce corpus donne les résultats de désambiguïsation de l'attribut POS. A cet effet, les instances des ensembles d'apprentissage et test seront décrites par les attributs POS-2, POS-1, POS+1 et POS+2 qui représentent respectivement les catégories grammaticales des deux mots qui suivent et des deux mots qui précédent le mot courant. Le tableau 2.7 présente les taux de désambiguïsation de l'attribut POS pour les deux corpus Hadith et Treebank arabe donnés par les six classifieurs.

Nous obtenons des résultats proches avec des taux élevés. Ces résultats révèlent que l'approche de désambiguïsation possibiliste est indépendante du domaine et de type du texte. Elle fournit des taux raisonnables (plus de 80%) pour les textes de journaux ainsi que pour les textes de Hadith. Cependant, il y a une différence d'environ 7% entre les deux corpus. Comme les tailles des deux corpus sont presque égales, nous pouvons expliquer ce fait par la nature de l'analyseur morphologique (i.e. *Aramorph*) dont le lexique est plutôt classique. Ainsi, cet outil est incapable d'analyser certaines entrées modernes.

Classifieur / Corpus	**Bayésien Naïf**	**Arbres de décision**	**SVM**	**Possibiliste utilisant Π**	**Possibiliste utilisant N**	**Possibiliste utilisant (Π+ N)**
Hadith	88.62 %	89.58 %	89.98 %	91.58 %	90.17%	90.45%
Treebank arabe	80.98%	81.85%	81.77%	84.23%	83.26%	83.35%

Tableau 2.7 : Classification possibiliste et non-possibiliste de l'attribut POS pour le Hadith et le Treebank arabe

D'autre part, le tableau 2.8 présente les taux de désambiguïsation de l'attribut POS dans les trois domaines "Boissons", "Mariage" et "Purification" en utilisant notre classifieur possibiliste de base (utilisant la possibilité). Nous avons utilisé seulement le POS des deux mots précédents et suivants. Le taux moyen de désambiguïsation de l'attribut POS pour les trois domaines était d'environ 88.38%.

	Boissons	**Mariage**	**Purification**	**Moyenne**
Taux de désambiguïsation des Noms	98.07%	98.33%	98.83%	98.41%
Taux de désambiguïsation des Verbes	98.47%	96.82%	95.25%	96.84%
Taux global de désambiguïsation	89.47%	87.73%	87.96%	88.38%

Tableau 2.8 : Taux de désambiguïsation de l'attribut POS pour les trois domaines

Nous avons obtenu des résultats similaires pour les trois domaines. Par conséquent, nous montrons que les approches basées sur le contexte (le classifieur possibiliste) peuvent être utilisés pour n'importe quel domaine ou type de texte. Pour renforcer notre évaluation, nous effectuons des expériences en utilisant le corpus Treebank arabe (Part 2).

	Hadith	**Treebank arabe**
Taux de désambiguïsation des Noms	98.41%	94.78%
Taux de désambiguïsation des Verbes	96.84%	88.71%
Taux global de désambiguïsation	88.38%	83.35%

Tableau 2.9 : Taux de désambiguïsation de l'attribut POS pour le Hadith et le Treebank arabe

Le tableau 2.9 présente les taux de désambiguïsation de l'attribut POS en utilisant les deux corpus Hadith et Treebank arabe. Nous obtenons des résultats proches des taux élevés. Nous utilisons le classifieur (ΠI+N) avec repondération, étant donné qu'il a donné les meilleurs résultats dans la plupart des cas. Ces résultats révèlent que l'approche de désambiguïsation possibiliste est réutilisable indépendamment des domaines et des types de texte, car il a fourni des taux de désambiguïsation (plus de 80%) sur des textes de presse (Treebank arabe) ainsi que des textes religieux (Hadiths). Cependant, il y a une différence d'environ 5% dans les deux taux globaux des deux corpus. En effet, et étant donné que les tailles de ces deux corpus sont presque égales, nous pouvons expliquer cette différence par le fait que le corpus du Hadith contient des expressions récurrentes, qui existent à la fois dans les deux ensembles d'apprentissage et de test (par exemple la phrase "صلى الله عليه وسلم": Paix et Bénédiction soient Sur Lui).

4.5. Impact de la possibilité lexicale dans le modèle de repondération

Nous calculons la possibilité lexicale de chaque mot dans l'ensemble de test. Le tableau 2.10 donne les taux de désambiguïsation des attributs morphologiques pour les six classifieurs. La classe choisie correspond à celle qui a la plus grande valeur de la formule (2.19). En comparant ces résultats à ceux du tableau 2.6, nous confirmons que la possibilité lexicale améliore les taux de désambiguïsation de tous les attributs morphologiques.

De même, le classifieur (ΠI+N) avec repondération donne, en général, les meilleurs résultats de taux de désambiguïsation. Par conséquent, nous estimons que la possibilité lexicale améliore la désambiguïsation des attributs morphologiques en fournissant un taux d'amélioration moyen de l'ordre de 3%. En effet, cette amélioration est expliquée par l'existence de quelques mots avec leurs valeurs correctes de classe dans l'ensemble d'apprentissage. Cependant, cette contrainte n'est pas toujours satisfaite. Dans le cas où la possibilité lexicale est égale à zéro, le classifieur possibiliste discriminatif est dispensé du calcul de cette possibilité lexicale en utilisant la formule (2.8).

Attribut morphologique	Sans repondération					Avec repondération					
	Π	***N***		Π + **N**		Π		***N***		Π + **N**	
	DR	**DR**	**DB**	**DR**	**DB**	**DR**	**DB**	**DR**	**DB**	**DR**	**DB**
POS	91.58%	90.17%	-1.41%	90.45%	-1.13%	91.72%	**+0.14%**	90.70%	-0.88%	91.75%	**+0.17%**
CONJONCTION	95.04%	88.66%	-6.38%	90.79%	-4.25%	95.05%	**+0.01%**	86.41%	-8.63%	97.55%	**+2.51%**
PARTICULE	96.91%	98.87%	**+1.96%**	98.87%	**+1.96%**	98.46%	**+1.55%**	98.87%	**+1.96%**	98.87%	**+1.96%**
DETERMINANT	95.25%	95.92%	**+0.67%**	96.13%	**+0.88%**	95.92%	**+0.67%**	95.02%	-0.23%	97.86%	**+2.61%**
PRONOM	59.10%	59.56%	**+0.46%**	58.79%	-0.31%	59.56%	**+0.46%**	59.56%	**+0.46%**	59.56%	**+0.46%**
PERSONNE	66.27%	65.07%	-1.20%	66.88%	**+0.61%**	65.07%	-1.20%	64.82%	-1.54%	67.00%	**+0.73%**
VOIX	78.75%	78.80%	**+0.05%**	79.05%	**+0.30%**	79.16%	**+0.41%**	78.81%	**+0.06%**	79.11%	**+0.36%**
ASPECT	81.78%	86.20%	**+4.42%**	86.16%	**+4.38%**	86.44%	**+4.66%**	86.24%	**+4.46%**	88.76%	**+6.98%**
GENRE	93.23%	90.45%	-2.78%	93.78%	**+0.55%**	94.00%	**+0.77%**	96.78%	**+3.55%**	97.23%	**+4.00%**
NOMBE	95.30%	87.00%	-8.50%	93.25%	-2.05%	93.25%	-2.05%	91.22%	-4.08%	96.38%	**+1.08%**
CAS	68.40%	68.76%	**+0.36%**	76.55%	**+8.15%**	70.15%	**+1.75%**	70.46%	**+2.06%**	76.62%	**+8.22%**
PREPOSITION	88.60%	90.20%	**+1.60%**	95.70%	**+7.10%**	88.74%	**+0.14%**	90.42%	**+1.82%**	95.70%	**+7.10%**
MODE	99.93%	99.96%	**+0.03%**	99.96%	**+0.03%**	99.96%	**+0.03%**	99.93%	00.00%	99.96%	**+0.03%**
ADJECTIF	97.58%	96.86%	-0.72%	97.63%	**+0.05%**	97.71%	**+0.13%**	97.88%	**+0.30%**	99.00%	**+1.42%**
Moyenne	86.27%	85.46%	-0.82%	87.43%	**+1.16%**	86.80%	**+0.53%**	86.22%	-0.05%	88.95%	**+2.69%**

Tableau 2.10 : Taux de désambiguïsation impliquant la possibilité lexicale

5. Bilan des contributions et perspectives

Nous avons présenté dans ce chapitre de nouvelles approches possibilistes de désambiguïsation des attributs morphologiques des textes arabes non-voyellés. La désambiguïsation est considérée comme une tâche de classification. A cet égard, nous avons défini un classifieur possibiliste pour apprendre et tester des données imprécises. D'abord, nous avons établi trois modèles de classification qui calculent respectivement les mesures de possibilité, de nécessité et la somme de ces deux mesures. Nous avons effectué une étude comparative de ces trois modèles de classification possibiliste avec des classifieurs non-possibilistes pour désambiguïser 14 attributs morphologiques en utilisant une technique de validation croisée. En comparant les résultats des différents classifieurs, nous avons conclu que la théorie des possibilités a donné de meilleurs taux de désambiguïsation quand elle combine les mesures de nécessité et de possibilité.

Ensuite, nous avons présenté et étudié un classifieur possibiliste discriminatif avec modèle de repondération dédié pour désambiguïser les attributs morphologiques des textes arabes. L'apprentissage de ce classifieur a été fait sur des textes voyellés et testé sur des textes non-voyellés. La première phase de cette approche (l'apprentissage) a généré les solutions morphologiques de textes voyellés en utilisant l'outil *AraMorph*. La deuxième phase (test) a déterminé, pour les textes non-voyellés, les instances correspondantes décrites par les mêmes attributs utilisés dans l'ensemble d'apprentissage. Les mots non-voyellés de ces textes ont, généralement, plus d'une valeur pour chaque attribut morphologique. Pour désambiguïser l'attribut d'un mot ambigu, nous nous sommes basés sur l'ensemble d'apprentissage de cet attribut. Nous désambiguïsons ces attributs en identifiant la classe qui correspond à la plus-haute mesure de possibilité et/ou de nécessité calculée sur l'ensemble d'apprentissage spécifique. En effet, nous avons défini trois modèles de classification basés sur ces différentes mesures (sans et avec repondération) pour déceler la classe précise. Ces mesures sont: la possibilité, la nécessité et la somme de possibilité et de nécessité. Pour calculer

l'impact de chaque attribut dans la désambiguïsation morphologique, nous avons intégré dans ces mesures le gain informationnel, basé sur l'entropie, acquérant ainsi six classifieurs au total.

Nous avons effectué les expérimentations de ces six classifieurs possibilistes susmentionnés en utilisant une technique de validation croisée sur les 14 attributs morphologiques. Le meilleur taux de désambiguïsation est assuré grâce au classifieur possibiliste avec repondération, en utilisant la somme de deux mesures de possibilité et de nécessité. En outre, nous avons discerné également les relations de dépendance entre les différents attributs. En effet, nous avons précisé pour chaque attribut les gains informationnels des attributs qui ont présenté les poids discriminatoires les plus élevés. Nous avons conclu que l'attribut POS est impliqué dans la désambiguïsation de tous les attributs morphologiques. Nous devons également noter que seuls les textes religieux (Coran et Hadiths) et certains textes didactiques pour enfants sont entièrement voyellés. La grande taille de ces collections permet un apprentissage précis. Basé sur la même idée, nous avons étendu notre travail en testant le corpus Treebank qui représente un autre type de texte (des textes modernes des articles de journaux). En effet, nos expériences ont révélé que nos classifieurs sont réutilisables indépendamment du domaine et du type de texte.

Malgré ces résultats encourageants, nous avons remarqué que notre approche n'arrive pas à désambiguïser intégralement la totalité des attributs morphologiques. Cela peut s'expliquer par un phénomène linguistique connu en langue arabe qui se traduit par un ordre relativement aléatoire des mots dans la phrase [Keskes et al., 2013] et également par l'incapacité de désambiguïser les particules qui ont un taux d'ambiguïté élevé, même dans les textes voyellés. Comme perspectives, nous envisageons de faire face à ces problèmes en adoptant l'une des deux alternatives. D'une part, nous pouvons agrandir l'ensemble d'apprentissage. D'autre part, l'intégration d'une analyse linguistique manuelle dans la phase d'apprentissage permettra de filtrer les mots vides et de minimiser le taux d'ambiguïté résultant. Cependant, nous essaierons de réduire le taux d'intervention afin d'éviter de traiter tout l'ensemble d'apprentissage à la main. Nous visons aussi à intégrer notre approche dans un système de recherche d'information (SRI) arabe qui traite des textes voyellés et non-voyellés, en introduisant une phase primitive de désambiguïsation de requêtes et de documents. A cette étape, nous pouvons renoncer à la désambiguïsation des particules car elles sont considérées comme des mots vides et ne sont pas utilisés dans l'indexation. En outre, les attributs morphologiques calculés par nos outils sont utiles même pour d'autres niveaux d'analyse à savoir syntaxiques et sémantiques [Bounhas et Slimani, 2009a ; Bounhas et al., 2011a]. En outre, nous programmons une comparaison avec l'outil de désambiguïsation morphologique MADA. En effet, et afin d'assurer une comparaison objective de ces deux outils, ces derniers devraient être entraînés et testés sur le même corpus. Enfin, notre contribution dans ce chapitre constitue une tentative pour traiter l'imprécision dans un cas d'application réelle. Nous avons introduit de nouveaux facteurs à la classification possibiliste, à savoir l'implication de la mesure de nécessité ainsi que le gain informationnel qui nécessite plus d'investigation dans d'autres domaines.

Chapitre 3 :
Expansion sémantique de requêtes

1. Introduction

Le développement quasi exponentiel de la connaissance humaine répartie sur des champs d'intérêt variés conduit à la génération d'une grande masse d'informations de plus en plus difficile à gérer et à entretenir. Dans cet environnement à grande échelle, caractérisé à la fois par le grand nombre d'utilisateurs et l'immense masse de données, il devient essentiel de concevoir et de développer des outils permettant un accès efficace et organisé. Il est crucial de développer des interfaces automatisées qui permettent de formuler/reformuler et satisfaire les besoins d'information des utilisateurs. Par ailleurs, la Recherche d'Information (RI) est une branche de traitement de données qui s'intéresse à l'acquisition, l'organisation, le stockage et la recherche de documents de différents formats. Nous avons besoin de systèmes de recherche d'information (SRI) qui constituent des outils informatiques visant à capitaliser l'information et à localiser les documents pertinents. Compte tenu de l'obligation d'information exprimée sous forme de requête, la pertinence est quantifiée selon un modèle de correspondance entre les termes de la requête et les documents. Quelles que soient les sémantiques données aux représentations des objets (documents ou requête) ou la définition de la pertinence, ces modèles ont un comportement général identique. La majorité d'entre eux représente les documents et les requêtes par des listes de mots-clés pondérés. Par conséquent, à partir de la notion de requête/réponse, la pertinence du résultat donné par un SRI dépend principalement de la requête. Toutefois, l'utilisateur est souvent incapable de donner quelques mots-clés qui décrivent explicitement et clairement son besoin intentionnel, ce qui peut détériorer la qualité des résultats attendus.

L'expansion de requêtes est l'une des stratégies mises en œuvre dans les SRI pour améliorer leur performance et mieux satisfaire les utilisateurs. Elle consiste à renforcer la requête de l'utilisateur en ajoutant de nouveaux termes à sa requête originelle pour mieux exprimer son besoin. En fait, il existe dans la littérature deux approches principales d'expansion de requêtes : l'expansion automatique de requêtes (EAR) et l'expansion interactive de requêtes (EIR) [Ruthven, 2003]. Notons que l'expansion automatique est plus simple pour l'utilisateur, mais limite ses performances ; car la participation de l'utilisateur dans le processus d'expansion a été ignorée. En impliquant les usagers, l'expansion interactive est plus complexe pour l'utilisateur, mais ce dernier sera en face de plusieurs problèmes tels que les requêtes ambiguës. En outre, les résultats d'un SRI pourront être erronés soit par le retour d'un nombre limité de documents pertinents (faible rappel) soit par la récupération d'un grand nombre de documents non-pertinents (faible précision). Historiquement, l'expansion automatique a un meilleur rappel par rapport à l'expansion interactive [Vélez et al., 1997]. Néanmoins, l'expansion automatique diminue fréquemment la précision lorsque les termes utilisés pour élargir le contexte d'une requête changent souvent le sens de cette dernière. Le problème est que les utilisateurs ne considèrent généralement qu'un nombre limité des premiers résultats retournés [Jansen et McNeese, 2005], ce qui rend la précision indispensable à la performance de la recherche. En revanche, l'expansion interactive a équilibré la précision et le rappel conduisant à une amélioration de performance de la recherche.

Cependant, comme l'expansion automatique, la précision de l'expansion interactive a besoin d'être améliorée. En fait, des approches ont commencé à améliorer la précision en intégrant des connaissances sémantiques [Crabtree, 2009]. Ceci peut être réalisé par différentes techniques telles

que l'analyse et la classification de corpus [Nie et al., 1997; Claveau et al., 2004], la reformulation par réinjection (ou rétroaction) de la pertinence (en anglais *Relevance Feedback* (RF)) et l'intégration des ressources linguistiques externes (dictionnaires, thésaurus, ontologies, etc.). Nous nous concentrons dans nos approches sur cette dernière technique d'expansion interactive de requêtes à l'aide d'un dictionnaire, comme proposé dans [Elayeb, 2009]. Plusieurs expériences d'expansion de requêtes ont été menées par exemple en utilisant des bases de données lexicales telles que WordNet dans les SRI anglais [Voorhees, 1994; Smeaton, 1997].

Notre travail est impliqué dans les stratégies d'expansion sémantique de requêtes (ESR) basées sur les ressources linguistiques externes. Dans notre cas, nous avons exploité le dictionnaire français "Le Grand Robert". Nous présentons dans ce chapitre une double contribution. Tout d'abord, nous profitons d'une approche existante de [Elayeb et al., 2009] modélisant le dictionnaire sous forme d'un graphe permettant le calcul des similarités entre les termes de la requête en exploitant les circuits existants entre eux. La seconde contribution s'intéresse à une nouvelle approche possibiliste d'expansion sémantique de requêtes. En fait, nous nous sommes inspirés de la théorie des possibilités en profitant d'une double mesure de pertinence reliée à la possibilité et à la nécessité entre les articles du dictionnaire et les termes de la requête. De plus, nous combinons ces deux approches en utilisant deux méthodes d'agrégation différentes. Nous profitons également d'une approche existante de repondération des termes de la requête [Elayeb et al., 2009 ; 2011] dans le modèle d'appariement possibiliste pour améliorer le processus d'expansion. Afin d'évaluer et comparer nos approches, nous avons effectué des expérimentations sur la collection de test "LeMonde94".

Le présent chapitre est structuré comme suit : La section 2 constitue une revue de la littérature dans le domaine d'expansion sémantique de requêtes. Dans les sections 3 et 4, nous présentons respectivement nos deux approches à base de circuits et possibiliste pour l'expansion sémantique de requêtes. Ses deux approches sont basées respectivement sur les réseaux petits-mondes hiérarchiques (RPMH) et les réseaux possibilistes (RP) [Elayeb, 2009]. La section 5 est consacrée à l'extension de notre modèle de base d'ESR, en détaillant nos agrégations des scores possibilistes et à base de circuits ainsi que le processus de repondération de termes de la requête dans le modèle d'appariement possibiliste. L'ensemble des expérimentations, les analyses des résultats et les études comparatives feront l'objet de la section 6. La section 7 discute nos apports et résultats. Nous concluons dans la section 8, par présenter le bilan de nos contributions et proposer les principales orientations pour les futures recherches.

2. Synthèse des approches existantes d'expansion sémantique de requêtes

Carpineto et Romano (2012) ont suggéré une taxonomie des approches d'expansion automatique de requêtes (cf. Figure 3.1 [Ben Khiroun, 2018]). En fait, les approches globales, à savoir les approches basées sur l'analyse de corpus ainsi que les approches basées sur des ressources linguistiques externes (dictionnaire, ontologie, thésaurus, etc.) sont les plus populaires. Nous nous focalisons dans la suite sur ce genre de méthodes. Une discussion plus détaillée est disponible dans [Ben Khiroun, 2018].

Le processus d'expansion sémantique de requêtes consiste à chercher des termes sémantiquement proches des termes de la requête proposée par l'utilisateur, en utilisant des ressources linguistiques externes [Picton et al., 2008]. Plusieurs travaux de la littérature [Fang, 2008 ; Zhang et al., 2009 ; Agirre et al., 2010] ont profité de WordNet pour étendre le contexte des requêtes en RI. En effet, Fang (2008) a obtenu des résultats prometteurs lors du test de son approche d'expansion de requêtes utilisant WordNet dans un contexte de recherche axiomatique. La méthode consiste à sélectionner

les termes qui chevauchent entre les glossaires des termes de la requête et les glossaires de WordNet. Au contraire, Zhang et al. (2009) ont profité de WordNet afin de désambiguïser les termes des requêtes ambiguës avant de procéder à leurs expansions. Pal et al. (2014) ont suggéré une approche mixte utilisant WordNet en considérant sa relation sémantique avec la requête, sa distribution et son association statistique avec les termes de la requête.

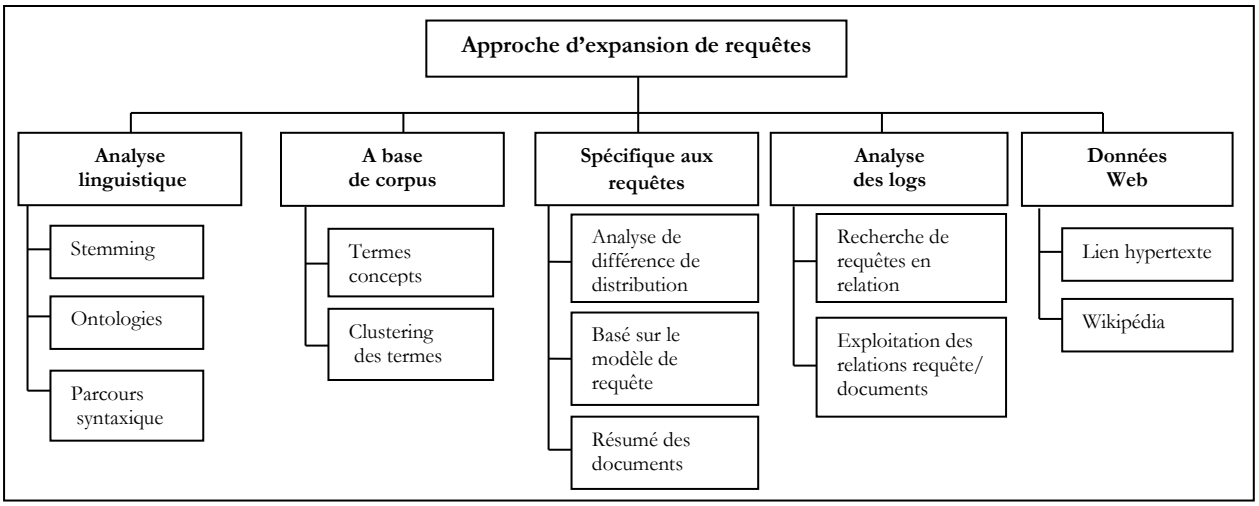

Figure 3.1 : Taxonomie des approches d'expansion automatique de requêtes

Par ailleurs, certains auteurs [Hersh et al., 2003 ; Fu et al., 2005 ; Nilsson et al., 2006 ; Bhogal et al., 2007] ont supporté l'utilisation des ontologies de domaine en expansion sémantique de requêtes. En fait, l'utilisation des ontologies générales ainsi que des bases lexicales indépendantes du domaine (WordNet) en ESR peut poser le problème d'ambiguïté des nouveaux termes injectés dans la requête. Dramé et al. (2014) ont utilisé le thésaurus MeSH en expansion de requêtes dans le domaine d'accès aux informations médicales. Lv et al. (2015) ont testé une approche de modélisation linguistique pour la recherche de microblog dans Twitter. L'expansion de requête a profité des termes de connaissances issues de Freebase (projet de collecte et de connexion des connaissances issues du Web). Récemment, Zingla et al. (2018) ont proposé une approche hybride d'expansion de requêtes en combinant des ressources linguistiques externes avec des règles associatives. Leurs résultats sont prometteurs en testant l'approche en recherche de microblog sur twitter (en utilisant Robust 2004), ainsi qu'en contextualisation des tweets (en utilisant INEX 2014).

D'autre part, les approches basées sur l'analyse de corpus ont profité des associations des termes dans la collection dans le but d'injecter des termes voisins dans la requête, soit via des méthodes de classification automatique de document [Chifu et Mothe, 2014], soit en calculant des liens contextuels entre les termes [Diaz et al., 2016 ; Ermakova et Mothe, 2016 ; Kuzi et al., 2016]. Par exemple, l'approche de Kuzi et al. (2016) a utilisé Word2Vec afin de générer les termes sémantiquement proches des termes de la requête. De plus, Diaz et al. (2016) ont confirmé que les modèles de plongements des mots (*word embedding*) tels que GloVe et Word2Vec, utilisés dans l'apprentissage global, sont moins efficaces que les approches spécifiques aux requêtes.

Il existe deux familles d'approches qui ont profité des journaux de recherche (fichier log) dans leurs processus d'ESR. La première famille a utilisé les relations existantes entre les requêtes et les résultats de recherche afin de générer un contexte supplémentaire. Par exemple, l'extraction des termes à partir des résultats cliqués [Cui et al., 2003] ou la recherche de requêtes associées aux mêmes documents [Billerbeck et al., 2003]. Par contre, la seconde famille considère les requêtes

comme des documents afin d'extraire les caractéristiques de celles liées à la requête originelle, avec ou sans exploitation des résultats de recherche associés [Jones et al., 2006 ; Yin et al., 2009].

Les approches d'expansion de requêtes utilisant un corpus local sont gênés par plusieurs lacunes telles que : (i) difficulté de trouver des termes d'expansion correspondant aux termes de la requête inexistants dans le corpus ; (ii) les relations entre les termes sont limitées. Plusieurs approches [Milne et Witten, 2008 ; Almasri et al., 2014 ; Gan et Hong, 2015] ont convergé vers l'exploitation des articles Wikipédia, riche en information sémantique, afin de pallier aux limites des approches à base de corpus.

Les différentes techniques d'expansion sémantique de requêtes (ESR) incluent les approches à base de réinjection de la pertinence (*Relevance Feedback*), les modèles de connaissances dépendantes et indépendantes de corpus et les approches à base des ressources linguistiques externes. En fait, les ressources externes telles que les dictionnaires, les thésaurus et les ontologies ont une couverture linguistique plus élevé par rapport aux corpus. Mais, la construction d'ontologie est coûteuse en termes de temps et nécessite des outils sophistiqués pour extraire et organiser les connaissances. En outre, les ontologies ne sont pas disponibles pour toutes les langues et tous les domaines. Les thésaurus sont construits à partir des indexes de documents. Par conséquent, ils souffrent de problèmes d'ambiguïté et de couverture étant donné que tous les sens de mots sont représentés et clairement distingués. Cependant, les dictionnaires sont des ressources exhaustives qui sont construites pour couvrir l'ensemble du langage et représentent naturellement les significations et les relations entre les mots. De plus, ces dictionnaires sont disponibles pour toutes les langues. Enfin, les dictionnaires multilingues peuvent être utilisés en recherche d'information translinguistique (RIT) [Elayeb et Bounhas, 2016]. Ainsi, les dictionnaires sont les ressources les plus génériques qui puissent être utilisées pour soutenir les SRI en particulier dans le processus d'expansion de requêtes [Elayeb et al., 2011].

D'autre part, les approches d'ESR basées sur l'analyse distributionnelle ou LSA (*Latent Semantic Analysis*) sont sensibles à la qualité et au degré de couverture du corpus utilisé. En fait, la tâche de génération de tous les sens possibles d'un mot polysémique donné semble être difficile à achever. En outre, il n'est pas garanti que le corpus inclue toutes les relations possibles entre les mots qui sont nécessaires pour l'étude quantitative. En cas d'existence de toutes ces relations, il n'est pas aussi garanti qu'elles soient convenablement traitées avec l'analyse distributionnelle ou LSA ou toute autre technique utilisant l'analyse de corpus.

Audeh (2014) a montré que l'injection automatique des nouveaux termes sémantiquement proches de termes de la requête originelle peut causer d'autres problèmes, indépendamment de la méthode d'ESR utilisée. Ces nouveaux défis sont particulièrement :

- *Choix de l'approche d'ESR :* La disponibilité de ressources linguistiques, la nature de stockage nécessaire ainsi que le temps de calcul sont parmi les critères du choix d'une approche d'ESR. Mais, les priorités de l'utilisateur dans un contexte donné peuvent impliquer d'autres critères du choix. Par exemple, une approche performante d'ESR dans le domaine médicale doit améliorer le facteur Rappel, alors que c'est la Précision qu'il faut considérer pour une approche d'ESR applicable dans la recherche Web. De plus, la nature de la collection de documents est aussi un facteur critique lors du choix de la méthode d'expansion. Par exemple, les approches d'ESR qui dépendent de statistique sur la collection préfèrent des collections statiques contenant des documents stables, au contraire aux collections dynamiques susceptibles d'être modifiée à tout moment (par exemple les pages Web en ligne) [Audeh, 2014].

- *Choix des paramètres des techniques de rétroaction de pertinence :* Parmi les paramètres à fixer par l'utilisateur lors de son utilisation de l'une des techniques de rétroaction de pertinence, nous citons : le nombre de terme à extraire, le nombre de documents pertinents retournés, etc. Plusieurs travaux de recherche (par exemple [Montgomery et al., 2004 ; Ksentini et al., 2016 ; Romberg, 2017]), utilisant cette technique d'ESR, ont prouvé que la manipulation de ces paramètres est parmi le défi majeur de la RI. Car le bon réglage de ces paramètres pour une collection de test donnée ne restera pas forcément efficace en testant sur d'autres collections. Même dans la même collection, deux requêtes distinctes sont parfois paramétrées des deux façons différentes.
- *Dérivation de la requête :* Cet évènement se produit lorsque le processus d'ESR conduit à une altération de l'objectif initial du besoin d'information proposé par l'utilisateur. C'est le cas lorsque les documents retournés par les nouveaux termes injectés sont plutôt pertinents par rapport à un autre sens loin de celui désiré par l'utilisateur au départ.

Par ailleurs, les données sur les quelles sont basées les approches d'expansion de requêtes dans la littérature sont pauvres, imprécises et incertaines, alors que la théorie des possibilités est naturellement appropriée pour ce type d'application. Elle permet d'exprimer les phénomènes d'ignorance, d'imprécision et d'incertitude [Boughanem et al., 2009]. En effet, la théorie des possibilités définit deux types de pertinence. D'une part, la *pertinence plausible*, quantifiée par la possibilité, permet d'éliminer les termes non sémantiquement similaires aux termes de la requête originelle (ceux non-pertinents). D'autre part, la *pertinence nécessaire*, quantifiée par la nécessité, contribue à l'amélioration de nos croyances en termes restants non-éliminés par la mesure de possibilité (i.e. les termes sémantiquement proches utiles pour l'expansion) [Elayeb et al., 2011].

Dans ce contexte, nous proposons, comparons et combinons dans ce chapitre deux approches d'expansion sémantique de requêtes basées sur un dictionnaire afin de les exploiter pour améliorer la performance d'un SRI possibiliste.

3. Approche d'ESR à base de circuits

Dans [Elayeb, 2009] nous avons étudié le problème d'expansion sémantique de requêtes (ESR) et son impact sur un SRI possibiliste intelligent. Notre méthode a été basée sur le calcul du nombre de circuits entre les termes nœuds d'un graphe généré à partir d'un dictionnaire considéré comme des réseaux petits-mondes hiérarchiques (RPMH). Avant de présenter la distance sémantique à base de circuits (cf. section 3.2), nous rappelons brièvement les RPMH dans la section 3.1. Un exemple de calcul pour l'expansion de la requête à base de circuits est détaillé dans [Elayeb et al., 2011].

3.1. Les Réseaux Petits-Mondes Hiérarchiques (RPMH)

Les RPMH ont été définis pour exploiter les caractéristiques statistiques de graphes. Ils ont d'abord été proposés par Watts et Strogatz (1998) dans le domaine de l'analyse de réseau social. Dans ces réseaux, il a été remarqué que la plupart des nœuds ont peu de relations entre eux ce qui permet de constituer de petits-mondes. En fait, la caractéristique principale de ces graphes est leur capacité à regrouper les nœuds proches. Cela a encouragé de nombreux auteurs à utiliser ces graphes pour modéliser, classer et grouper les termes ou les mots [Newman, 2003; Gaume et al., 2004; Elayeb et al, 2009; Elayeb, 2009].

D'autre part, Gaume et al. (2004) ont montré les différences entre les graphes aléatoires (nœuds sont donnés, les arêtes ou arcs sont tirés au hasard), les graphes réguliers et les RPMH. Les

caractéristiques des RPMH par rapport à ces types de graphes (un haut taux de clustering et les chemins sont courts) permettent de regrouper des nœuds selon les circuits qui les unissent. Dans notre cas, nous avons transformé le dictionnaire français "Le Grand Robert" en un RPMH en considérant qu'il pourrait représenter un graphe caractérisé par une concentration de relations (arêtes) entre tous les mots français (nœuds) ayant la même signification. En fait, nous sommes partis de l'idée suivante : Il existe un arc entre un nœud terme t_i et un nœud terme t_j si et seulement si t_j apparaît dans la définition dictionnairique de t_i comme synonyme.

L'utilisation de ce type de réseau dans nos systèmes est justifiée par plusieurs arguments. En premier lieu, il s'agit d'un outil flexible qui permet d'analyser les connaissances pour en insérer d'autres. La flexibilité vient de la théorie des graphes, riche en algorithmes de manipulation, ce qui répond à notre besoin de personnalisation et d'adaptation. En deuxième lieu, les RPMH peuvent être utilisés pour divers types de connaissances qu'elles soient sémantiques ou sociales. Ils sont aussi génériques du point de vue source de données. Par exemple, nous pourrons facilement extraire les termes d'un réseau à partir d'un thésaurus au lieu d'un dictionnaire. En troisième lieu, la caractéristique classificatoire est fondamentale dans le modèle que nous proposons étant donné qu'elle permet à l'utilisateur de comprendre la structure de son espace informationnel et donc de l'appréhender.

3.2. La proximité sémantique à base de circuits

Ce qui fait la force des RPMH est leur caractéristique classificatoire. Cette caractéristique permet de découvrir des *clusters* de nœuds. Nous avons proposé dans [Elayeb, 2009] de regrouper les termes d'un dictionnaire structuré sous forme d'un graphe RPMH en utilisant le nombre de circuits comme distance. En partant d'un dictionnaire qui représente la langue, un graphe de termes est constitué. Deux termes t_i et t_j sont liés si l'un d'eux apparaît dans la définition de l'autre. Dans ce graphe, les mots du dictionnaire maintiennent des relations qui font parfois des circuits. Pour un terme t_i donné de la requête initiale Q^{old}, nous avons calculé un score de proximité sémantique (en termes du nombre de circuits) du terme t_i avec tout autre terme t_j, selon la formule suivante [Elayeb, 2009] :

$$Proximité\ Sémantique(t_i, t_j) = \frac{Nombre\ de\ circuits(t_i,t_j)}{Nombre\ maximum\ de\ circuits\ dans\ le\ graphe} \quad (3.1)$$

Avec : *Nombre de circuits*(t_i, t_j) représente le nombre de circuits existants, en partant du nœud t_i et passant par le nœud t_j dans le RPMH de dictionnaire (i.e. $t_i \rightarrow \ldots \rightarrow t_j \rightarrow \ldots \rightarrow t_i$). Le processus d'expansion consiste à enrichir Q^{old} par des termes sémantiquement proches afin de générer la nouvelle requête reformulée Q^{new}. En fait, une de nos contributions, en comparaison avec nos travaux dans [Elayeb, 2009], réside dans la pleine exploitation du dictionnaire français "Le Grand Robert". Contrairement à Elayeb qui a été limité dans ses tests à un sous-ensemble de verbes de ce dictionnaire en raison des limites des ressources de l'ordinateur, nous considérons ici, en plus des verbes, toutes les catégories grammaticales de mots, à savoir les adverbes, les noms et les adjectifs. Par ailleurs, la longueur maximale du circuit est l'un des paramètres importants dans cette distance. En fait, plus le circuit est long plus il y a de chances de mélanger différents groupes de sens. Cependant, tenir compte uniquement des circuits trop courts pourra causer le groupement de termes, liés au même hyperonyme, en différents groupes. Pour plus de détails sur le principe de regroupement, [Elayeb, 2009] a précisé que la longueur maximale de circuit que nous puissions prendre en compte est d'environ 4 arcs.

4. Approche possibiliste d'ESR

Notre approche est basée sur les réseaux possibilistes. Elle permet d'identifier les termes les plus appropriés pour être ajoutés à la requête initiale. Un exemple de calcul des scores de proximité possibiliste est donné dans [Elayeb et al., 2011]. Les termes ajoutés sont pondérés de nouveau selon le modèle de repondération présenté dans la section 5.1.

Notre modèle est inspiré des travaux de [Brini et al., 2004] qui ont proposé un modèle d'appariement possibiliste quantitatif pour la recherche d'information. Ce modèle a ensuite été étendu par [Elayeb et al., 2009; Elayeb, 2009] vers un cadre possibiliste qualitatif. Nous exploitons la connaissance extraite du dictionnaire français "Le Grand Robert" pour proposer nos nouveaux réseaux possibilistes pour l'expansion sémantique de requêtes. En effet, le dictionnaire présente un ensemble d'entrées qui sont des termes ayant des définitions que nous appelons articles dans le reste de ce chapitre. Chaque article est indexé par un ensemble de termes qui apparaissent dans sa définition. Les articles et leurs termes d'indexation sont représentés par les réseaux possibilistes naïfs comme le montre la figure 3.2. Une relation de dépendance existe entre un terme et un article [Elayeb et al., 2011].

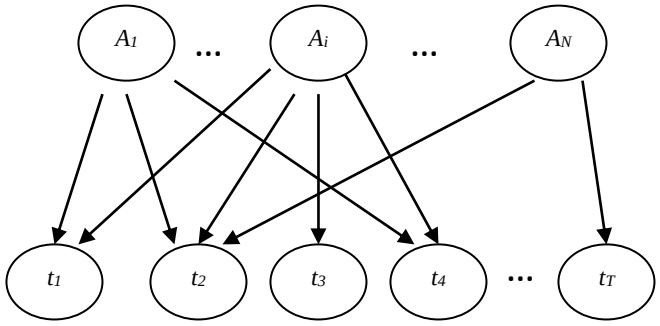

Figure 3.2 : Architecture générale de l'approche possibiliste d'ESR

Un arc entre un nœud article A_i et un nœud terme t_j reflète la possibilité et la nécessité que t_j soit représentatif (ou non) de A_i. La mesure de possibilité est utile pour éliminer les articles non sémantiquement proches, alors que la mesure de nécessité renforcera la pertinence des articles restants. Le processus d'expansion retourne à l'utilisateur les articles plausibles ou nécessairement pertinents. Cette dépendance entre t_j et A_i est calculée en fonction de la fréquence de t_j dans la définition de A_i ainsi que sa fréquence dans l'ensemble d'articles de dictionnaire.

Etant donné une requête Q_j composée de t_j (ou plusieurs termes), notre modèle possibiliste devrait être capable de répondre à des propositions du type :

- Est-il plausible à un certain degré que l'article A_i constitue une bonne réponse à la requête Q_j?
- Est-il nécessaire (ou certain dans le sens possibiliste), que l'article A_i réponde à la requête Q_j?
- L'article A_i est-il préférable qu'un autre article soit ajouté à la requête Q_j?

La première question est destinée à retirer de la réponse les articles faiblement plausibles. La seconde se concentre sur les articles qui sont réellement pertinents. Le dernier type de proposition suggère que la liste ordonnée des articles, en réponse à un besoin utilisateur, peut être traitée d'une manière qualitative et que des approches ordinales pourraient être utilisées dans la représentation des articles et des requêtes. L'utilisation de la théorie des probabilités dans la définition de la pertinence d'une requête donnée ne tient pas compte de notre connaissance limitée à propos de la pertinence d'un

article ; car elle ne tient pas compte des caractères imprécis et vagues qui sont intrinsèques à la pertinence [Brini et al., 2004].

La pertinence quantitative de chaque article de dictionnaire (A_j), étant donné la requête $Q = (t_1, t_2, \ldots, t_T)$, est calculée de la manière suivante :

Selon [Elayeb et al., 2009], le degré de possibilité $\Pi(A_j|Q)$ est proportionnel à :

$$\Pi'(A_j|Q) = \pi(t_1|A_j) * \ldots * \pi(t_T|A_j) = nFreqt_{1j} * \ldots * nFreqt_{Tj} \quad (3.2)$$

Avec : - $nFreqt_{ij} = \frac{Freqt_{ij}}{maxFreqt_{ij}}$: La fréquence normalisée du terme t_i dans l'article de dictionnaire A_j.

- $Freqt_{ij} = \frac{le\ nombre\ d'occurrences\ du\ terme\ t_i\ dans\ l'article\ A_j}{nombre\ de\ termes\ dans\ l'article\ A_j}$

- $maxFreqt_{ij}$: La fréquence maximale.

La certitude de restituer un article pertinent A_j pour une requête, notée $N(A_j|Q)$, est donnée par :

$$N(A_j|Q) = 1 - \Pi(\neg A_j|Q) \quad (3.3)$$

Avec:

$$\Pi(\neg A_j|Q) = \frac{\Pi(Q|\neg A_j) * \Pi(\neg A_j)}{\Pi(Q)} \quad (3.4)$$

De même $\Pi(\neg A_j|Q)$ est proportionnelle à :

$$\Pi'(\neg A_j|Q) = \pi(t_1|\neg A_j) * \ldots * \pi(t_T|\neg A_j) \quad (3.5)$$

Ce numérateur peut être exprimé par :

$$\Pi'(\neg A_j|Q) = (1 - \phi A_{1j}) * \ldots * (1 - \phi A_{Tj}) \quad (3.6)$$

Avec:

$$\phi A_{ij} = Log_{10}\left(\frac{nCA}{nA_i}\right) * (nFreqt_{ij}) \quad (3.7)$$

Où : - nCA est le nombre total d'articles dans le dictionnaire.

- nA_i est le nombre d'articles de dictionnaire contenant le terme t_i.

Nous définissons le Degré de Pertinence Possibiliste (DPP) de chaque article A_i de dictionnaire étant donné une requête Q par :

$$DPP(A_j|Q) = \Pi(A_j|Q) + N(A_j|Q) \quad (3.8)$$

Les articles préférés sont ceux qui ont une valeur $DPP(A_j|Q)$ élevée. En effet, ces articles sont sémantiquement proches des termes de la requête Q^{old} et sont utiles pour son expansion.

5. Extension des approches d'ESR

Nous proposons dans cette section une extension de nos approches d'ESR. D'abord, nous présentons brièvement notre modèle de repondération des termes de la requête (cf. section 5.1). Ensuite, nous détaillons deux nouvelles agrégations des scores possibilistes et à base de dénombrement de circuits (cf. section 5.2).

5.1. Repondération des termes de la requête

Nous profitons ici de notre approche de repondération des termes de la requête qui a amélioré davantage notre approche possibiliste d'appariement [Elayeb, 2009; Elayeb et al., 2009]. Dans une première étape du processus d'expansion, l'utilisateur choisit pour chaque terme de la requête initiale Q^{old}, un nombre de termes sémantiquement proches à ajouter afin de générer la nouvelle requête reformulée Q^{new}. Nous définissons le nouveau poids de chaque terme t_i (w_i) étant donné la requête reformulée Q^{new} de la façon suivante :

$$\omega_i = \left[\frac{nTP_{t_i}(Q^{new})}{nT(Q^{old})} + 1\right] * Freq_{t_i}(Q^{new}) \quad (3.9)$$

Avec: - $nTP_{t_i}(Q^{new})$: le nombre de termes sémantiquement proches choisi pour t_i dans Q^{new}.

- $nT(Q^{old})$: le nombre de termes dans Q^{old}.

- $Freq_{t_i}(Q^{new})$: le nombre d'occurrences de t_i dans Q^{new}.

Ces nouveaux poids des termes sont insérés dans le modèle possibiliste d'appariement tel que nous l'avons détaillé dans l'introduction générale (cf. section 2.2.5) [Elayeb, 2009; Elayeb et al., 2009]. Ainsi, en considérant ces nouveaux poids, la formule (3.2) sera (3.10) et la formule (3.6) sera (3.11) :

$$\Pi'(A_j|Q) = \pi(t_1|A_j) * \omega_1 * \ldots * \pi(t_T|A_j) * \omega_T = nFreqt_{1j} * \omega_1 * \ldots * nFreqt_{Tj} * \omega_T \quad (3.10)$$

$$\Pi'(\neg A_j|Q) = \left(1 - \frac{\phi_{A_1 j}}{\omega_1}\right) * \ldots * \left(1 - \frac{\phi_{A_T j}}{\omega_T}\right) \quad (3.11)$$

5.2. Agrégations des scores possibiliste et à base de circuits pour l'ESR

Notre objectif dans cette étape est de proposer deux nouvelles approches à base d'agrégations de scores possibilistes et à base de circuits pour l'ESR. Nos heuristiques ici commencent à partir des premiers tests qui prouvent l'efficacité de l'approche à base de circuits ayant un taux d'amélioration plus considérable. Ainsi, nous proposons d'améliorer les résultats fournis par l'approche possibiliste en intégrant les connaissances provenant de circuits. En effet, le processus d'expansion possibiliste traite les termes de la requête dans son ensemble. Néanmoins, l'approche à base de circuits s'applique indépendamment de chaque terme de la requête. Par conséquent, le seul moyen d'agréger les deux approches consiste à renforcer le score possibiliste de chaque terme par son score à base de circuits. Un exemple de calcul est détaillé dans [Elayeb et al., 2011].

Nous définissons le Degré de Pertinence Hybride (DPH), relié à une requête $Q = (t_1, t_2, \ldots, t_T)$, par les formules d'agrégations suivantes.

(i) Agrégation à base de Somme:

$$DPH_{Somme}(A_i|Q) = DPP(A_i|Q) + MSC(A_i, Q) \quad (3.12)$$

(ii) Agrégation à base de Produit:

$$DPH_{Produit}(A_i|Q) = DPP(A_i|Q) * MSC(A_i, Q) \quad (3.13)$$

Avec $MSC(A_i, Q)$ est la Moyenne des Scores à base de Circuits d'un article A_i. Elle est calculée par la formule (3.14). Tandis que, $DPP(A_i|Q)$ est le Degré de Pertinence Possibiliste de chaque article A_i de dictionnaire étant donné une requête Q.

$$MSC(A_i, Q) = \frac{\sum_j Proximité\ Sémantique(A_i, t_j)}{T} \quad (3.14)$$

Avec :

- $Proximité\ Sémantique(A_i, t_j) = \frac{Nombre\ de\ circuits(A_i, t_j)}{Nombre\ maximum\ de\ circuits\ dans\ le\ graphe}$ (3.15)

- *T* est le nombre de termes dans la requête *Q*.

6. Résultats expérimentaux

Dans les sections qui suivent, nous présentons la collection de test utilisée dans nos expériences (cf. section 6.1). Pour améliorer notre évaluation, nous avons réalisé deux types d'évaluation. Nous analysons nos résultats à l'échelle globale et à l'échelle locale (cf. sections 6.2 et 6.3 respectivement).

6.1. La collection de test "LeMonde94"

Nous avons utilisé dans nos expériences une série du standard de test CLEF-2003. Ce standard fournit des outils nécessaires à l'évaluation des SRI sur des grands corpus, y compris un ensemble de documents, un ensemble de requêtes et la liste des documents pertinents pour chaque requête. Chaque requête est représentée, sous le format XML, par un titre, une description et une narration. Dans nos expériences, nous avons utilisé les titres comme requêtes pour tester nos approches. La collection nommée "LeMonde94" est un sous-ensemble de CLEF. Elle comprend des articles du journal français "Le Monde". Cette collection se compose de 44013 documents et 40 requêtes de test, le tout formant 154 Mo.

6.2. Analyse globale des résultats

Cette section résume et discute la performance globale des différents tests effectués. Le tableau 3.1 présente les principales expériences ainsi que les scores d'évaluation pour chacune. La première colonne représente l'identifiant de l'expérience (*Run id*) possédant le format suivant : [P|C|SA|PA][1..4][t|f]. Avec: (i) [P|C|SA|PA]: *P* réfère à l'expansion possibiliste, *C* réfère à l'expansion à base de circuits, *SA* désigne une expansion utilisant une agrégation à base de *Somme*, et *PA* désigne une expansion utilisant une agrégation à base de *Produit*; (ii) [1..4]: un nombre entre 1 et 4 relié au nombre de termes d'expansion pour chaque mot-clé de la requête initiale; (iii) [t|f]: *t* (resp. *f*) désigne l'application (resp. non application) de la repondération de termes après expansion de la requête. Les deux dernières colonnes représentent la précision moyenne *MAP* (*Mean Average Precision*) pour chaque requête ainsi que la précision exacte (*R-Precision*), qui est la précision au rang *R* ; où *R* est le nombre total de documents pertinents. La première ligne présente les premiers essais sans expansion [Elayeb et al., 2011].

Sans repondération, l'application de l'expansion de requête montre que la performance de toutes les approches diminue légèrement pour tous les tests s'il y a ajout de nouveaux termes. Cependant, avec repondération, seulement la méthode d'expansion utilisant l'agrégation à base de *Produit* présente des résultats légèrement meilleurs que les autres approches d'expansion lorsque nous ajoutons de nouveaux termes. Par conséquent, nous concluons que la combinaison des deux approches possibiliste et à base de circuits contribue à l'amélioration de la performance globale du SRI. Néanmoins, les résultats d'expansion ne permettent pas d'atteindre des améliorations très considérables. En fait, l'application de l'expansion de requêtes génère parfois du bruit dans les résultats de recherche ce qui dégrade en conséquence les valeurs des précisions.

		Run id	MAP	R-Precision
		topics2000	0.2358	0.2298
Expansion Possibiliste	**Sans repondération**	P1f	0.1681	0.178
		P2f	0.1497	0.1545
		P3f	0.1421	0.1475
		P4f	0.1341	0.143
	Avec repondération	P1t	0.1526	0.1389
		P2t	0.1031	0.1072
		P3t	0.0782	0.0769
		P4t	0.0683	0.0656
Expansion à base de Circuits	**Sans repondération**	C1f	0.1792	0.1887
		C2f	0.1644	0.1675
		C3f	0.1522	0.1523
		C4f	0.1426	0.1445
	Avec repondération	C1t	0.1714	0.1861
		C2t	0.1457	0.1432
		C3t	0.1272	0.1142
		C4t	0.1067	0.1036
Agrégation à base de Somme	**Sans repondération**	SA1f	0.1697	0.1623
		SA2f	0.1641	0.1596
		SA3f	0.1583	0.1506
		SA4f	0.1565	0.1481
	Avec repondération	SA1t	0.1862	0.1903
		SA2t	0.1854	0.1856
		SA3t	0.1742	0.1647
		SA4t	0.1739	0.1725
Agrégation à base de Produit	**Sans repondération**	PA1f	0.1730	0.1562
		PA2f	0.1690	0.1618
		PA3f	0.1669	0.1604
		PA4f	0.1655	0.1604
	Avec repondération	PA1t	0.1863	0.1817
		PA2t	0.1937	0.1859
		PA3t	0.1964	0.1896
		PA4t	0.1988	0.1873

Tableau 3.1 : Résultats des tests effectués au niveau de l'analyse globale

Au-delà de cette performance globale décevante, nous avons procédé à une analyse plus détaillée en examinant le résultat de chaque requête à part. Nous concluons que la dégradation générale des résultats est générée par des requêtes qui sont la plupart du temps mal interprétées à cause de la nature linguistique des mots. Par exemple, dans la requête "Le syndrome de la guerre du Golfe", le terme "golfe" est censé faire référence à la guerre en Irak au Moyen-Orient. Toutefois, dans le processus d'expansion, il a été interprété comme un nom ordinaire (non pas un nom propre). Les mots trouvés dans le dictionnaire comme "fleuve", "mer" et "aplanissement" sont dotés des hauts scores des proximités sémantiques, et en conséquence ils ont été ajoutés à la requête lors de son expansion. Ces erreurs d'interprétations grammaticales similaires augmentent sensiblement le bruit dans les résultats globaux des documents retournés par le SRI ce qui dégrade par la suite sa précision.

6.3. Analyse locale des résultats

Notre objectif dans cette section est d'étudier les améliorations de précision achevées suite au processus d'expansion de requêtes. En outre, nous évaluons nos approches d'agrégation en termes de nombre de requêtes dont les précisions moyennes (MAP) ont été améliorées. Le tableau 3.2 compare nos approches en fournissant le nombre de requêtes améliorées par chaque approche ainsi que le pourcentage d'amélioration correspondant, qui est calculé comme suit :

$$Pourcentage\ d'amélioration = \left[\frac{MAP(Q^{new}) - MAP(Q^{old})}{MAP(Q^{old})}\right] * 100 \quad (3.16)$$

Dans cette formule, le $MAP(Q^{old})$ et le $MAP(Q^{new})$ correspondent respectivement à la précision moyenne des deux requêtes originelle (*old*) et reformulée (*new*).

En analysons les résultats obtenus dans le tableau 3.2, nous constatons que la contribution de l'expansion possibiliste sans repondération est la moins efficace. Par contre, l'application de la repondération des termes dans les requêtes reformulées a permis d'augmenter significativement le taux d'amélioration de la méthode d'expansion possibiliste comme le montre la figure 3.3. En fait, le processus de repondération améliore l'interprétation des termes dans les requêtes reformulées. Le nouveau poids d'un terme sera de plus en plus important lorsque celui-ci est enrichi d'un grand nombre de mots sémantiquement similaires dans la requête reformulée; car il est considéré comme le plus significatif pour l'utilisateur.

Nous constatons également que le nombre de termes à ajouter pour générer la requête reformulée est un facteur important dans toutes les expériences. Cependant, nous n'avons pas pu confirmer que le nombre optimal de termes d'expansion est d'environ 3 comme il est montré dans les figures 3.3 et 3.4. D'une part, nous sommes limités à quatre termes dans toutes les expériences en raison de la limitation des ressources informatiques (en particulier pour la méthode d'expansion à base de circuits qui utilise des ressources excessives lors du calcul des circuits). D'autre part, les taux d'amélioration introduits dans le tableau 3.2 sont liés à des sous-ensembles de requêtes (la cardinalité d'un sous-ensemble est présenté par la colonne "Nombre de requêtes améliorées").

		Sans repondération		**Avec repondération**	
	Nombre de termes à ajouter	**Nombre de requêtes ameliorées**	**Pourcentage d'amélioration**	**Nombre de requêtes ameliorées**	**Pourcentage d'amélioration**
Expansion Possibiliste	1 terme	3	5.72%	4	49.09%
	2 termes	4	1.56%	3	93.93%
	3 termes	2	1.29%	2	97.89%
	4 termes	3	0.6%	2	96.03%
Expansion à base de Circuits	1 terme	4	52.5%	4	47.63%
	2 termes	2	98.85%	2	98.64%
	3 termes	2	98.5%	2	97.79%
	4 termes	2	84.52%	2	85.2%
Agrégation à base de Somme	1 terme	3	31.67%	8	23.48%
	2 termes	3	13.8%	9	17.82%
	3 termes	3	14.49%	11	31.6%
	4 termes	3	6.7%	10	27.43%
Agrégation à base de Produit	1 terme	2	75.15%	8	20.94%
	2 termes	3	13.95%	8	22.71%
	3 termes	2	18.45%	9	28.5%
	4 termes	2	15.59%	9	25.19%

Tableau 3.2 : Résultats des tests effectués au niveau de l'analyse locale

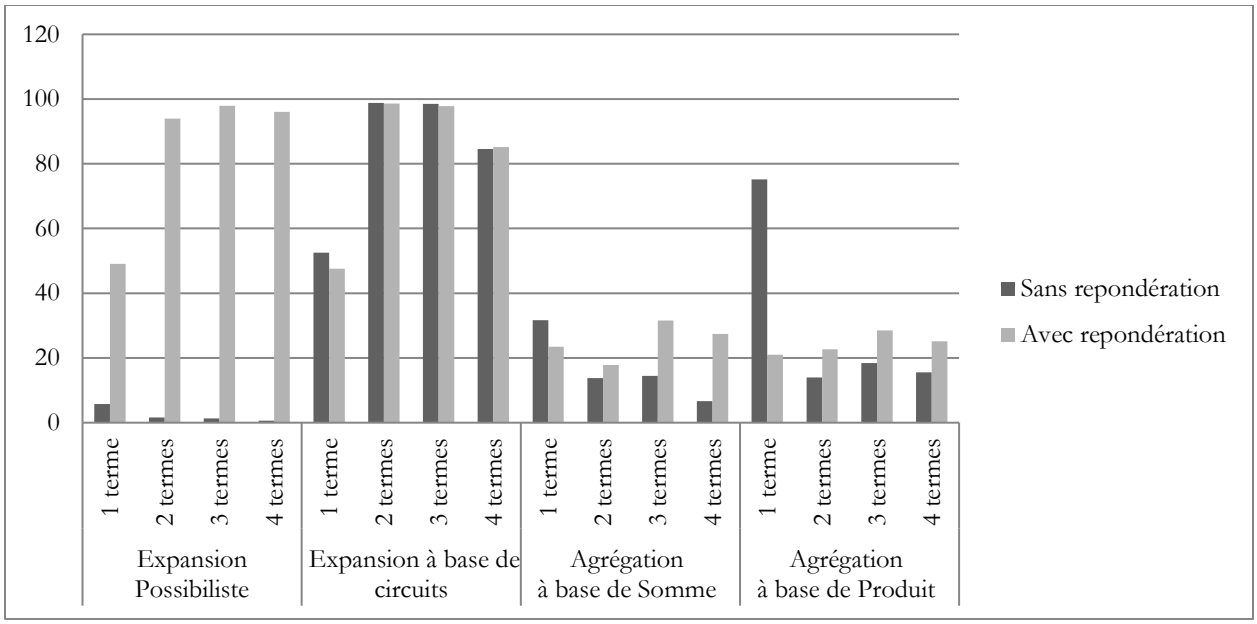

Figure 3.3 : Les taux moyens d'amélioration par méthode d'expansion

Par ailleurs, les résultats de l'expansion à base de circuits sont approximativement similaires avec ou sans repondération à cause de la nature de la formule de proximité sémantique utilisée dans le processus d'expansion. Cela prouve l'efficacité de l'approche possibiliste d'expansion qui concrétise l'importance de la repondération dans le SRI possibiliste proposé.

Le tableau 3.2 présente également la contribution de l'application des méthodes d'agrégation à base de somme et produit. La figure 3.4 récapitule le nombre de requêtes améliorées par chaque méthode d'expansion. En effet, sans l'agrégation, de 2 à 4 requêtes sont améliorées. Alors que lorsque nous regroupons les deux approches, nous arrivons à améliorer de 8 à 11 requêtes. Cela montre que nos deux approches d'expansion sont complémentaires chacune essaie de corriger les lacunes de l'autre.

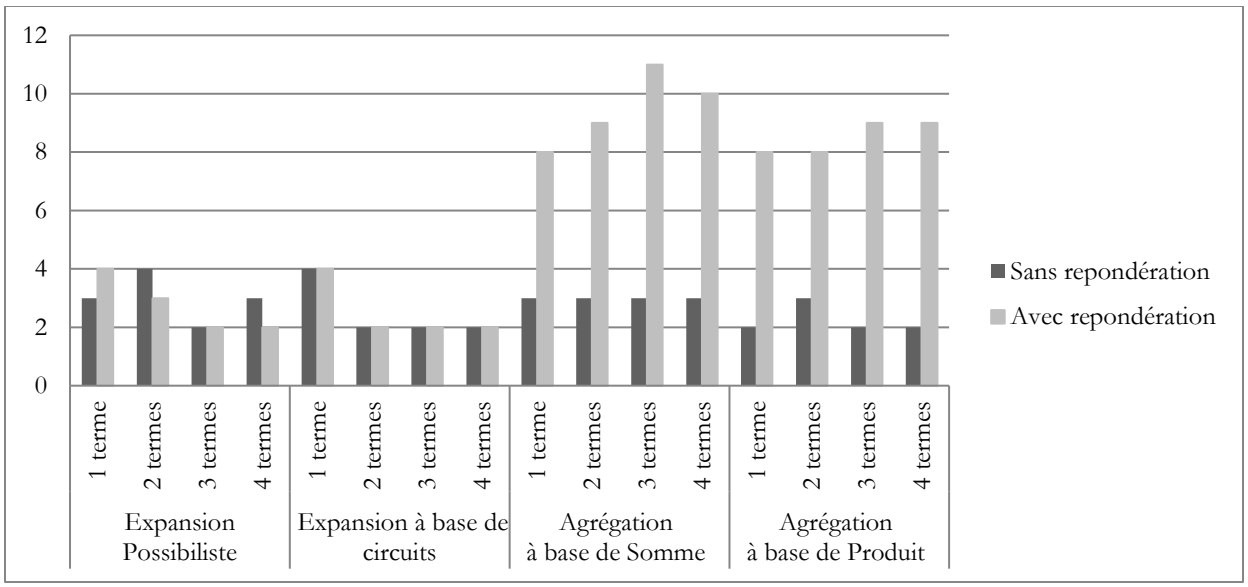

Figure 3.4 : Nombre de requêtes améliorées par méthode d'expansion

Cependant, les taux moyens d'amélioration, assurés par ces deux méthodes d'agrégations des scores, restent encore faibles, par comparaison au taux moyens d'amélioration de l'approche possibiliste (avec repondération) et au taux moyens d'amélioration de l'approche à base de circuits (avec et sans repondération). Cela est dû à la réduction du nombre de termes dans les requêtes de test. En effet, un nombre réduit de termes à ajouter pour élargir le contexte de la requête stimule davantage le taux d'ambiguïté de cette requête et renforce la sélection aléatoire des termes sémantiquement proches.

7. Discussion

Dans ce chapitre, nous avons proposé et évalué différentes approches d'expansion de requêtes basées sur un dictionnaire. Notre SRI possibiliste utilisant ces approches profite des dépendances existant entre les termes de la requête originelle et les définitions des articles d'un dictionnaire afin d'achever le processus d'expansion de requêtes. D'une part, et dans l'approche possibiliste d'expansion de requêtes, nous avons quantifié ces relations par deux mesures: la possibilité et la nécessité. Le même modèle possibiliste a été utilisé pour relier les termes de la requête reformulée aux documents de la collection dans le processus d'appariement. En fait, le degré de possibilité permet d'éliminer les documents non-pertinents (resp. articles), alors que la nécessité permet de renforcer la pertinence des documents (resp. articles) restants non-éliminés par la possibilité. Les documents récupérés (resp. articles) sont ceux qui sont nécessaires voir même pertinents étant donné une requête utilisateur. La requête utilisateur est considérée comme une nouvelle information à se propager dans un réseau possibiliste. D'autre part, les deux mesures de possibilité et de nécessité sont plus efficaces que le facteur IDF, puisque la distribution de termes dans le document (resp. article) ne dépend pas seulement de la présence ou de l'absence des termes dans les documents (resp. article) de la collection (comme IDF), mais aussi de la distribution de leur densité dans les documents de la collection (resp. articles). Ainsi, ces mesures se sont avérées efficaces pour la discrimination négative, comparée notamment à IDF.

En outre, si l'approche de base tient compte de l'aspect quantitatif, notre système permet de l'étendre au cadre qualitatif possibiliste, par l'introduction de poids (préférences) aux termes de la requête. L'intégration de poids dans les formules de possibilité et de nécessité permet d'augmenter les scores de pertinence possibiliste des documents contenant ces termes dans le but de pénaliser les scores de pertinence des documents ne les contenant pas. La pénalisation et l'augmentation des scores sont proportionnelles au pouvoir des termes à discriminer entre les documents de la collection. De plus, ces poids permettent de restaurer les documents classés par ordre de préférence de pertinence. Il est possible dans ce cas d'évaluer à quel point un document d_1 (resp. un article a_1) est préférable à un document d_2 (resp. un article a_2) ou de mesurer la préférence d'un document d_1 (resp. un article a_1) par rapport à un ensemble de documents $\{d_3, d_4\}$ (resp. un ensemble d'articles $\{a_3, a_4)\}$.

Par ailleurs, la contribution des réseaux possibilistes (RP) est renforcée par la distance à base de circuits, calculée à partir de graphe de dictionnaire considéré comme un RPMH. Cette distance s'est avérée utile seulement pour proposer des termes sémantiquement proches à ajouter à la requête originelle. En fait, nous avons proposé de combiner les deux scores afin d'augmenter le nombre de requêtes améliorées.

Dans le contexte général de RI et plus spécifiquement les SRI possibilistes, nous résumons nos contributions comme suit : D'une part, notre travail est dédié au domaine de RI où la définition de la pertinence nécessite plus d'optimisation dans la recherche d'information. D'autre part, SPORSER est une contribution dans le domaine de la modélisation des informations par la structuration de

l'espace d'information en deux réseaux possibilistes (RP). Le premier RP modélise les dépendances entre les termes (mots-clés de la requête originelle et les termes entrées d'un dictionnaire) et permet l'expansion de la requête. Ce type de relations n'a pas été pris en compte dans le modèle de base de [Boughanem et al., 2009]. Le second RP modélise les dépendances entre les termes de la requête reformulée et les documents de la collection. En effet, ces deux réseaux sont mixés pour traduire la pertinence des documents étant donné une requête reformulée.

Cependant, l'objectif principal est centré sur les difficultés d'améliorer les résultats de la recherche d'information indépendamment de la collection de test. A titre de comparaison, Lioma et al. (2005) ont employé la même collection "LeMonde94" dans le but d'étudier la contribution de technique de rétroaction de pertinence (*Relevance Feedback*) en RI français. Ils ont conclu que la pauvre performance de l'expansion de requêtes peut être liée à la collection de test elle-même [He and Ounis, 2009]. En outre, Picton et al. (2008) ont présenté une étude sur les méthodes d'expansion sémantique de requêtes basées sur l'analyse de la collection de documents comme une ressource linguistique pour le processus d'expansion. Ils ont réalisé un programme d'analyse distributionnelle dérivé du système UPERY [Bourigault, 2002] et ils l'ont appliqué à une série d'articles du journal "LeMonde" entre 1991 et 2000. Ils ont également proposé une nouvelle approche combinant cette méthode avec une technique de pseudo-réinjection de pertinence. Ce mixage a donné naissance à une nouvelle méthode d'expansion nommée "rétroaction distributionnelle".

Toutefois, les travaux de Picton sur l'expansion distributionnelle ont conduit à une diminution globale dans la performance du système en raison de l'existence de certaines catégories de mots qui devraient être évitées dans le processus d'expansion (cas des adjectifs, des noms particuliers et entités nommées). Si nous comparons nos résultats en utilisant la collection "LeMonde94" à ceux de Picton, nous confirmons que les performances des méthodes d'expansion de requêtes dépendent directement des requêtes de tests utilisés. Les termes polysémiques insérés dans la requête reformulée peuvent provoquer des bruits dans les résultats de recherche. Pour réduire ce problème, nous avons intégré un mécanisme de désambiguïsation sémantique de requêtes afin de résoudre le problème de termes ambigus dans le processus d'expansion [Ben Khiroun et al., 2014a ; Elayeb et al., 2015b]. D'autre part, nous prévoyons de résoudre le problème lié au nombre optimal de termes à ajouter par l'utilisateur à sa requête originelle en proposant une technique d'expansion automatique de requêtes telle que proposé dans [Ogilvie et al., 2009 ; Latiri et al., 2012].

8. Bilan des contributions et perspectives

Nous récapitulons nos approches d'expansion sémantique de requêtes dans le système SPORSER basé sur une ressource linguistique externe à savoir le dictionnaire français "Le Grand Robert". La première approche modélise la structure de ce dictionnaire par le biais d'un Réseau Petits-Mondes Hiérarchiques (RPMH). Nous avons énuméré les circuits entre les nœuds du graphe RPMH pour calculer un score de proximité sémantique entre les termes. En fait, nous avons étendu notre approche initiée par [Elayeb et al., 2009] et limitée seulement aux verbes français à toutes les catégories grammaticales, à savoir les adverbes, les noms et les adjectifs.

Nous avons introduit une deuxième approche sémantique basée sur les réseaux possibilistes (RP). Cette méthode affine la recherche de nouveaux termes pour l'expansion sémantique de requêtes. En effet, elle prend en compte une double mesure de proximité sémantique entre les articles d'un dictionnaire et les termes de la requête à reformuler. En fait, nous cherchons les articles d'un dictionnaire possiblement et nécessairement pertinents pour l'expansion de la requête initiale de l'utilisateur. Ensuite, nous proposons et nous comparons deux nouvelles approches d'expansions de

requêtes combinant les deux techniques possibiliste et à base de circuits. La contribution de ces nouvelles approches hybrides est confirmée par le nombre de requêtes améliorées. En outre, SPORSER constitue un système adaptatif de recherche d'information guidant l'utilisateur au cours de processus d'expansion de requêtes. Il propose un cadre interactif qui permet de visualiser la structure du dictionnaire sous forme de graphe RPMH et permet à l'utilisateur de sélectionner et d'élargir les termes de sa requête.

Nous avons utilisé la collection de test "LeMonde94" pour évaluer nos approches. Cela a conduit à une amélioration partielle des résultats de certaines requêtes de test à l'aide du modèle d'appariement possibiliste. Ces améliorations, non confirmées à l'échelle globale de l'analyse, prouvent que la performance de toute approche d'expansion sémantique de requêtes dépend de la nature des requêtes de test dans la collection. Nous prévoyons d'étudier d'autres alternatives adaptant avec la collection de test "LeMonde94". En fait, SPORSER accepte n'importe quel standard de test organisant des requêtes et des documents au format XML. Basé sur la plate-forme Terrier, son architecture est extensible et offre un haut-niveau d'abstraction. En outre, SPORSER est implémenté en Java selon le modèle MVC (*Model-View-Controller*) ce qui améliore davantage sa portabilité et son interopérabilité au même temps. Nous étudions dans le chapitre suivant l'impact de la désambiguïsation possibiliste de requêtes sur leurs expansions dans les SRI intelligents.

Chapitre 4 : Impact de la désambiguïsation possibiliste de requêtes sur leurs expansions

1. Introduction

Les systèmes de recherche d'information (SRI) souffrent aujourd'hui de nombreux problèmes, notamment ceux liés aux requêtes des utilisateurs. Ces derniers expriment leurs besoins sous forme de requêtes courtes qui peuvent aussi contenir des termes ambigus. Par conséquent, les résultats du SRI peuvent inclure plusieurs documents non-pertinents (bruit) en raison du contexte limité fourni par ces requêtes. Ce bruit diminue l'efficacité de la recherche et ouvre les portes sur deux problèmes à résoudre tels que la désambiguïsation sémantique de requêtes (DSR) et l'expansion sémantique de requêtes (ESR) afin d'améliorer les résultats de recherche.

Le processus de désambiguïsation sémantique de requêtes [Krovetz, 1997; Paskalis et Khodra, 2011; Zhong et Ng, 2012] est basé sur la désambiguïsation de leurs termes ambigus étant donné le contexte entièrement fourni par la requête. En effet, la désambiguïsation consiste à sélectionner le sens convenable d'un mot étant donné son contexte [Navigli, 2009 ; Elayeb 2018]. Ce problème reste majeur dans le domaine du traitement automatique du langage naturel (TALN) et a une grande influence dans plusieurs applications connexes telles que la recherche d'information mono-, multi- et translinguistique, l'extraction d'information, la traduction automatique, l'analyse du contenu, le traitement et l'analyse de texte, la lexicographie et les applications du Web sémantique.

Récemment, le domaine de désambiguïsation sémantique a été principalement amélioré grâce aux compétitions *Senseval* et *SemEval*. Par exemple, Chan et al. (2007) et Carpuat et Wu (2007) ont confirmé que l'efficacité des systèmes de traduction automatique a été considérablement améliorée grâce à l'incorporation d'une tâche de désambiguïsation appuyant le processus de traduction. Cependant, dans le domaine de la recherche d'information, la tâche de désambiguïsation sémantique a montré également son importance, soit au niveau de la requête soit à celui du document : (i) les termes de requête peuvent avoir des sens étroitement liés à d'autres mots qui n'existent pas dans la requête. Par conséquent, le rappel peut être amélioré si l'on tient compte de ses liens sémantiques entre les mots ; et (ii) les termes de requêtes et de documents peuvent avoir plusieurs sens qui diminuent la précision de recherche [Chifu et Ionescu, 2012]. En fait, la sélection du sens correct pour les termes de requêtes et de documents peut améliorer considérablement la précision de recherche en diminuant le bruit dans les résultats de documents retournés.

En général, les systèmes de désambiguïsation soutiennent les SRI en identifiant les sens appropriés des termes des requêtes et des documents au cours du processus de recherche. D'une part, l'étape d'analyse et d'indexation de requêtes est améliorée par l'identification du sens correct de chaque terme étant donné son contexte. D'autre part, les sens correct des termes des documents doivent également être identifiés afin de les indexer convenablement compte tenu de leur contexte. En conséquence, les deux tâches de désambiguïsation des termes de requêtes et de documents doivent être effectuées avant de commencer le processus de recherche. Néanmoins, cette conclusion n'a pas été approuvée dans certains travaux de recherche tels que [Krovetz et Croft, 1992; Voorhees, 1993], où l'efficacité de la recherche ne peut pas être améliorée en dépit de l'intégration d'un système de désambiguïsation dans leur SRI. Au contraire, d'autres travaux tels que [Schütze et Pedersen, 1995; Gonzalo et al., 1998; Stokoe et al., 2003; Kim et al., 2004; Liu et al., 2005; Zhong et Ng, 2012] ont

justifié l'amélioration des performances globales de leurs SRI grâce à l'intégration des systèmes de désambiguïsation.

L'expansion sémantique de requêtes [Elayeb et al, 2011; Carpineto et Romano, 2012] est le processus de reformulation de l'ensemble des termes de la requête originelle en lui ajoutant quelques autres termes sémantiquement proches afin d'identifier et d'enrichir le contexte de la recherche. Cette technique vise à renforcer l'efficacité de la recherche dans les SRI. En cas de moteur de recherche Web, l'expansion de requête comprend l'évaluation de termes de la requête d'origine de l'utilisateur ainsi que l'expansion de ses termes afin de récupérer le maximum de documents pertinents. En fait, l'expansion de requêtes implique de nombreuses autres techniques telles que : (i) la repondération des termes de la requête originelle et reformulée, (ii) la sélection de la racine de chaque terme de la requête en vue d'identifier les différentes formes morphologiques des termes, (iii) l'identification des erreurs d'orthographe par la recherche automatique des formes corrigées, proposées dans les résultats de la recherche, et (iv) la recherche des synonymes des termes de la requête originelle afin d'enrichir son contexte.

Toutefois, la tâche d'expansion de requêtes peut reformuler la requête initiale en ajoutant certains termes ambigus. Ce problème ne peut être résolu qu'avec un processus de désambiguïsation de requêtes. Cette relation de dépendance entre les deux tâches prouve la nécessité de combiner les deux dans le but d'améliorer la performance globale de la recherche. Dans [Elayeb et al., 2011] et [Ben Khiroun et al., 2012] nous avons proposé respectivement des approches d'expansion et de désambiguïsation sémantiques de requêtes fondées sur les réseaux possibilistes en utilisant les dictionnaires comme ressources linguistiques externes. Nous proposons dans ce chapitre une méthode combinant les deux approches de désambiguïsation et d'expansion de requêtes en utilisant les réseaux possibilistes et appliqués sur un graphe de cooccurrence [Ben Khiroun et al., 2014a]. Nous avons également testé les réseaux possibilistes pour améliorer les résultats de RI, en étudiant de nombreuses combinaisons de scénarios de désambiguïsation, d'expansion et de réinjection (rétroaction) de pertinence. Nos expériences sont effectuées sur la collection de test ROMANSEVAL pour le processus de la désambiguïsation, et la collection CLEF-2003 pour le processus d'expansion. Les résultats montrent l'impact positif de la désambiguïsation de requêtes sur leurs expansions basées sur les indicateurs standards de rappel/précision. Nous comparons davantage nos résultats possibilistes à ceux obtenus par une seconde approche à base de dénombrement de circuits.

Dans ce chapitre, nous proposons, évaluons et comparons une nouvelle approche possibiliste d'expansion de requêtes utilisant la désambiguïsation sémantique à base d'un graphe de cooccurrence. D'abord, nous présentons dans la section 2 certains travaux connexes. Puis, la section 3 est consacrée à l'approche possibiliste combinant la DSR et l'ESR. Ensuite, l'ensemble d'expérimentations, leurs résultats et interprétations font l'objet de la section 4. Enfin, la section 5 conclut ce chapitre par l'évaluation de notre travail et la proposition des orientations pour des futures recherches.

2. Synthèse des approches existantes combinant la DSR et l'ESR en RI

Dans cette revue de littérature, nous étudions les approches existantes combinant les deux techniques de désambiguïsation et d'expansion de requêtes et leurs impacts sur la performance des SRI. En effet, plusieurs approches dans la littérature ont étudié l'impact de la DSR sur la performance de RI en utilisant des connaissances à partir de thésaurus. En effet, certaines méthodes à base de thesaurus ont accompli des améliorations en matière d'efficacité de RI en élargissant les

termes désambiguïsés de la requête par des synonymes ainsi que d'autres informations issues de WordNet [Voorhees, 1994; Liu et al., 2004; Liu et al., 2005; Fang, 2008]. En outre, l'expansion de sémantique de requêtes a également bénéficié des connaissances de WordNet, qui a enregistré des améliorations dans la performance du SRI [Cao et al., 2005; Agirre et al., 2010].

D'autre part, Pinto et Pérez-sanjulián (2008) ont exploité WordNet comme ressource linguistique externe à la fois pour la DSR et l'ESR. Ils ont signalé la nécessité d'une étape de désambiguïsation de requête au cours de son expansion afin d'augmenter les performances de la RI. Les résultats expérimentaux, obtenus en utilisant des requêtes courtes et longues de la collection de texte TREC-8, ont confirmé que l'ESR appliquée sur les requêtes courtes et longues ne suffit pas pour accroître l'efficacité de RI. Par contre, l'identification du sens approprié de chaque terme ambigu de la requête, en utilisant un ensemble de synonymes extraits de WordNet, peut principalement contribuer à améliorer les performances de RI. Par conséquent, l'efficacité de la recherche a été significativement améliorée tant pour les requêtes courtes que longues.

Par ailleurs, Paskalis et Khodra (2011) ont proposé, testé et évalué plusieurs scénarios de RI en utilisant la DSR et l'ESR, la lemmatisation et une technique de réinjection de pertinence. Pour la tâche de DSR, les auteurs ont utilisé une version étendue de l'algorithme de Lesk [Banerjee et Pedersen, 2002] afin d'identifier le sens exact de chaque terme de la requête et de documents. Pour la tâche d'ESR, ils ont exploité tout d'abord un thésaurus, à base de cooccurrence, construit automatiquement à partir de la collection de documents. Ensuite, ils ont profité d'une technique de pseudo-réinjection de pertinence en utilisant un ensemble de meilleurs documents pertinents afin d'en extraire les termes les plus représentatifs. Ces termes sont enfin injectés dans la requête initiale [Manning et al., 2008] pour améliorer le processus d'expansion.

D'autre part, les deux compétitions SemEval-2010 et 2013 ont enregistré que les approches de DSR et/ou d'ESR à base des graphes sont parmi les meilleures en terme d'amélioration de performance de la RI [Duque et al., 2015]. Certaines de ces approches sont à base des algorithmes existants et/ou étendus [Mihalcea, 2005 ; Navigli et Lapata, 2010 ; Agirre et al., 2014]. Ces algorithmes ont exploité les structures des graphes afin de profiter des relations de cooccurrence issues de l'analyse de corpus. Par exemple, l'objectif de l'algorithme *HyperLex* suggéré par [Véronis, 2004] est de générer un graphe de cooccurrence pour toutes les paires de mots qui cooccurrent dans le contexte d'un mot donné en utilisant un corpus. En effet, le graphe en question possède les caractéristiques d'un Réseaux Petits-Mondes Hiérarchique (RPMH) [Elayeb, 2009]. Ce réseau de termes fortement connectés a été exploité pour lever le verrou de désambiguïsation sémantique.

L'algorithme *HyperLex* a été comparé par [Agirre et al., 2006] à une version adapté de l'algorithme *PageRank* [Brin et Page, 1998] afin d'étudier l'impact des graphes sur la désambiguïsation des noms. Les résultats ont montré que les performances de ces deux algorithmes sont proches, malgré que *PageRank* utilise moins de paramètres d'optimisation par rapport à *HyperLex*. En outre, Silberer et Ponzetto (2010) ont profité des travaux de [Véronis, 2004] et [Agirre et al., 2006] dans l'objectif de proposer une approche de désambiguïsation à base de graphe de cooccurrence généré à partir d'un corpus parallèles multilingues. D'abord, ces auteurs ont appliqué les algorithmes des graphes dédiés initialement à la désambiguïsation sémantique monolingue. Ensuite, ils ont exploité l'algorithme d'arbre couvrant minimal (*Minimum Spanning Tree*) afin d'exécuter le processus de désambiguïsation des mots.

Récemment, Duque et al. (2015) ont proposé et testé une approche combinant un dictionnaire bilingue avec un graphe de cooccurrence dans le but de désambiguïser les traductions dans un contexte multilingue. Les algorithmes suggérés ont profité de plusieurs ressources telles que : (i) le

poids ou l'importance de chaque nœud du graphe utilisé ; (ii) des groupes des mots (sous-graphes) ayant des sens proches ; et (iii) les distances sémantiques entre les mots nœuds du graphe. Les résultats expérimentaux, utilisant les standards de test de SemEval-2010 et 2013 ont montré la performance de l'approche non supervisé à base de graphe. En fait, cette approche a considéré le document comme une information cohérente. Au contraire, les approches concurrentes ont considéré des fenêtres de taille particulière afin de générer le contexte et exécuter les algorithmes des cooccurrences.

Les approches existantes combinant la DSR avec l'ESR en RI ont utilisé différentes types des ressources linguistiques externes tel que WordNet. Cependant, ces approches sont fondées sur des données pauvres, incertaines et imprécises, alors que la théorie des possibilités est dédiée naturellement à ce genre d'application. Sur la base des avantages prévus par la théorie des possibilités, nous proposons dans la suite une approche possibiliste utilisant à la fois la DSR ainsi que l'ESR en profitant des atouts des graphes de cooccurrence extraits à partir de corpus. Notre objectif est d'évaluer l'impact de la désambiguïsation possibiliste de requêtes sur leurs expansions.

3. Approche possibiliste combinant la DSR et l'ESR

Notre approche combine la DSR, l'ESR et la pseudo-réinjection de pertinence. Pour les deux premières tâches, nous avons besoin de calculer la similarité entre les termes de requêtes (dans le cas de l'expansion) ou entre les termes et leurs sens possibles (dans le cas de la désambiguïsation). Dans cette approche, nous avons opté pour les graphes de cooccurrence extraits à partir de corpus afin de modéliser les liens et les similarités contextuels. Néanmoins, notre calcul de la similarité est suffisamment générique de sorte à ce qu'il soit utilisé par d'autres types de graphes (par exemple, des graphes de dictionnaire tel que nous l'avons utilisé dans [Elayeb et al., 2011]).

3.1. Représentation des connaissances à base des graphes

Notre approche est basée sur les réseaux possibilistes pour la DSR et l'ESR. En fait, nous avons généré le graphe de cooccurrence en considérant que deux nœuds termes sont liés s'ils existent dans la même phrase. Les arêtes sont bi-orientées et pondérées par la fréquence normalisée de cooccurrence des termes connexes. D'autre part, les mots ambigus sont liés avec leurs sens appropriés dans le dictionnaire. Considérons les différentes composantes comme suit:

- T : l'ensemble de termes dans le corpus.
- S : l'ensemble de sens dans le dictionnaire.
- Un nœud t_i est relié à un nœud t_j si t_i et t_j cooccurrent dans la même phrase ; avec $\{t_i, t_j \in T\}$.
- Un nœud t_i est relié à un nœud s_j si t_i est un terme ambigu et s_j représente un sens de t_i; avec $\{t_i \in T\}$ et $\{s_j \in S\}$.

3.2. Similarité possibiliste à base de graphe

Pour calculer les similarités entre les termes dans les deux processus de DSR et d'ESR, nous avons adapté le modèle possibiliste de [Elayeb et al., 2011] aux graphes de cooccurrence. En effet, le modèle proposé permet de calculer les deux scores de possibilité $\Pi(n_j|Q)$ et de nécessité $N(n_j|Q)$ de chaque nœud n_j du graphe de cooccurrence étant donné la requête $Q = (t_1, t_2, \ldots, t_T)$. La possibilité permet de rejeter les nœuds termes non-pertinents du graphe. Ces nœuds ne sont pas

sémantiquement proches du contexte de la requête et sont inutiles aussi bien pour sa désambiguïsation que pour son expansion. Par contre, la nécessité permet de renforcer la pertinence des nœuds termes restants non-éliminés par la possibilité. Ces deux mesures sont calculées de la façon suivante :

Selon [Elayeb et al., 2009], le degré de possibilité $\Pi(n_j|Q)$ est proportionnel à :

$$\Pi'(n_j|Q) = \pi(t_1|n_j) * \dots * \pi(t_T|n_j) = nFreqt_{1j} * \dots * nFreqt_{Tj} \quad (4.1)$$

Avec : - $nFreqt_{ij} = \frac{Freqt_{ij}}{maxFreqt_{ij}}$: La fréquence normalisée du terme t_i dans le graphe de cooccurrence.

- $Freqt_{ij}$: C'est le poids de l'arrête reliant les nœuds t_i et n_j (i.e. c'est le nombre de fois où les deux nœuds cooccurrent).

- $maxFreqt_{ij}$: le nombre maximale de fois où deux nœuds du graphe cooccurrent.

La certitude de restituer un nœud n_j du graphe de cooccurrence pour une requête, notée $N(n_j|Q)$, est donnée par :

$$N(n_j|Q) = 1 - \Pi(\neg n_j|Q) \quad (4.2)$$

Avec:

$$\Pi(\neg n_j|Q) = \frac{\Pi(Q|\neg n_j) * \Pi(\neg n_j)}{\Pi(Q)} \quad (4.3)$$

De même $\Pi(\neg n_j|Q)$ est proportionnelle à :

$$\Pi'(\neg n_j|Q) = \pi(t_1|\neg n_j) * \dots * \pi(t_T|\neg n_j) \quad (4.4)$$

Ce numérateur peut être exprimé par :

$$\Pi'(\neg n_j|Q) = (1 - \phi n_{1j}) * \dots * (1 - \phi n_{Tj}) \quad (4.5)$$

Avec:

$$\phi n_{ij} = Log_{10}\left(\frac{nCN}{nN_i}\right) * (nFreqt_{ij}) \quad (4.6)$$

Où : - nCN est le nombre total de nœuds reliés aux termes de la requête dans le graphe de cooccurrence.

- nN_i est le nombre de nœuds reliés au terme t_i.

L'utilisation de la fonction *Log* (comme dans TF-IDF) permet de calculer le pouvoir discriminant des termes de la requête. Ainsi, nous sélectionnons les nœuds du graphe qui sont les plus proches des éléments les plus discriminants de l'information contextuelle représentée dans la requête.

Nous définissons le degré de pertinence possibiliste (*DPP*) pour chaque nœud n_j du graphe de cooccurrence étant donné la requête $Q = (t_1, t_2, \dots, t_T)$ par :

$$DPP(n_j|Q) = \Pi(n_j|Q) + N(n_j|Q) \quad (4.7)$$

Les nœuds du graphe de cooccurrence préférés sont ceux qui ont une valeur $DPP(n_j|Q)$ élevée.

3.3. Processus d'ESR utilisant la DSR

Le processus d'ESR utilisant la DSR est récapitulé dans la figure 4.1. Il englobe les différentes ressources utilisées dans les tâches de DSR, d'ESR et de pseudo-réinjection de pertinence (PRP). A

partir d'une requête initiale, le module d'ESR est exécuté pour générer une requête reformulée. Dans le cas des termes ambigus, le module de DSR est utilisé avant l'application d'ESR. Ainsi, le meilleur nœud sens ayant le plus grand score possibiliste est choisi et les termes existants dans sa définition sont utilisés pour reformuler la requête initiale.

Pour les deux processus de DSR et d'ESR, le graphe de cooccurrence est utilisé pour le calcul des scores des pertinences possibilistes des termes. L'appariement entre la requête reformulée et la collection de documents permet d'obtenir un ensemble de documents résultats de la recherche. Une pseudo-réinjection de pertinence (PRP) est appliquée à la fin du processus en sélectionnant les termes les plus importants de l'ensemble des documents en tête du résultat retourné. L'ensemble du processus peut être réitéré. Afin d'effectuer la pseudo-réinjection de pertinence basée sur l'ensemble de documents retournés, nous avons utilisé la méthode Bo1 (Bose-Einstein 1) disponible dans la plate-forme de recherche d'information Terrier. Toutefois, nous avons choisi dans nos expériences les paramètres suivants : le nombre de termes à ajouter à une requête est fixé à 10 et le nombre de documents les mieux classés, à partir desquels les termes du processus PRP sont extraits, est limité à 3 documents [Elayeb et al., 2015b].

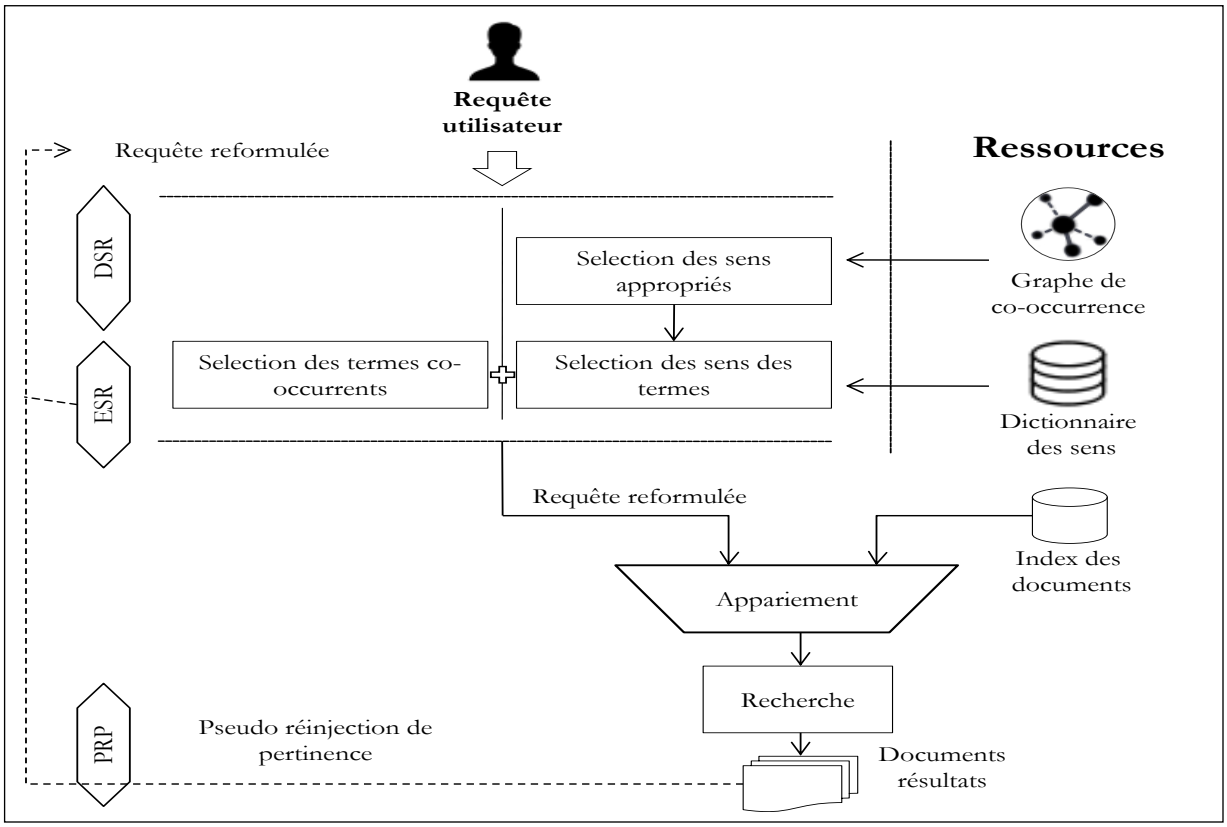

Figure 4.1 : Processus d'ESR utilisant la DSR

4. Résultats expérimentaux

Afin d'étudier l'impact de la DSR sur l'ESR en langue française, nous avons utilisé deux collections de test pour expérimenter notre approche à savoir ROMANSEVAL (cf. chapitre 1) et CLEF-2003 (cf. chapitre 3). Dans toutes nos expériences, nous nous sommes concentrés uniquement sur les requêtes de la collection de test CLEF-2003 qui contiennent des termes ambigus inclus dans la collection de test ROMANSEVAL.

4.1. Scénarios des expériences

Le sous-ensemble de requêtes utilisées pour nos expériences est composé de 15 requêtes contenant des mots ambigus de la collection de test ROMANSEVAL. Dans une première étape, nous avons étudié dans la section 4.2 l'impact de l'ESR, comme un processus séparé, sur la performance de RI. Ensuite, la DSR est expérimentée à part dans la section 4.3 afin d'évaluer le processus de désambiguïsation de termes de la requête. L'impact du processus de DSR sur l'ESR est expérimenté dans la section 4.4. Dans la section 4.5, nous comparons nos résultats possibilistes par rapport à une approche à base de circuits.

Nous avons utilisé la plate-forme expérimentale Terrier pour évaluer notre système. Deux métriques d'évaluation sont utilisées: (1) La précision mesurée par le ratio de documents pertinents retournés au nombre de documents trouvés, et (2) Le rappel présentant le rapport de documents pertinents retournés au nombre de documents pertinents dans la collection. Dans ce chapitre, nos expériences sont limitées au modèle d'appariement Okapi (BM25) disponible dans la plate-forme Terrier. Mais, nous prévoyons d'expérimenter notre approche via le modèle d'appariement possibiliste proposé par [Elayeb et al., 2009] afin de comparer les résultats à ceux obtenus par Okapi. En fait, notre objectif est de montrer que notre approche est générique et indépendante du modèle d'appariement utilisé.

4.2. Evaluation de l'approche possibiliste d'ESR

Nous comparons dans le tableau 4.1 des différents scénarios d'ESR basés sur le graphe de cooccurrence possibiliste (CooQE) généré à partir de la collection de test ROMANSEVAL. Ogilvie et al. (2009) ont étudié le nombre de termes à ajouter dans l'ESR automatique par huit systèmes de RI. Leurs résultats ont montré que le nombre de termes d'expansion qui optimise la précision moyenne (MAP) varie considérablement selon les systèmes et l'ensemble de requêtes de test. Pour de nombreuse requêtes, au moins dix termes d'expansion fournissent la meilleure précision moyenne dans les expériences de [Ogilvie et al., 2009]. Cette hypothèse est étudiée pour la langue française comme suit : Le nombre de termes d'expansion dans le tableau 4.1 a été modifié à partir de *N div 4* termes jusqu'à *N* termes ; où *N* représente le nombre de termes dans la requête initiale. Ces chiffres, pour les termes d'expansion, sont choisis en fonction de la longueur de la partie narrative des requêtes de test (plus de 10 termes). Toutefois, l'application du processus d'ESR sur ces requêtes longues, comme détaillé par [Pinto et Pérez-sanjulián, 2008], peut produire des résultats imprécis et non interprétables. Ainsi, nous nous sommes limités à un quart des termes de la requête comme scénario minimum pour avoir des résultats significatifs d'expansion.

Méthode	Nombre de termes d'ESR	MAP	R-précision
Baseline	-	0.5487	0.5174
CooQE	N	0.4180	0.4043
	N div 2	0.4700	0.4633
	N div 4	0.5083	0.4742

Tableau 4.1 : Résultats d'expansion sémantique de requêtes

Les deux dernières colonnes du tableau 4.1 [Elayeb et al., 2015b] présentent la mesure de la précision moyenne (MAP), qui est la moyenne des scores de précision moyenne pour chaque requête, et la précision exacte (R-précision), qui est la précision au rang *R* ; où *R* est le nombre total de documents pertinents [Manning et al., 2008]. Les résultats de Baseline, appliqués sur les requêtes initiales sans référence à l'ESR, sont également présentés dans ce tableau.

Nos résultats du tableau 4.1 montrent une dégradation des performances de RI lors de l'application du processus d'ESR proportionnellement au nombre de termes d'expansion, pour les deux métriques MAP et R-précision à la fois. Selon les courbes de rappel-précision présentées dans la figure 4.2, les résultats pour les trois scénarios d'ESR ne sont pas satisfaisants en comparaison avec les résultats de Baseline. Cependant, nous pouvons confirmer que l'ESR (principalement pour le scénario *N div 4*) est meilleure que la Baseline pour les taux de rappel élevés (proche de 1) et faibles (entre 0 et 0.1).

Ces résultats sont affectés par l'ambiguïté des requêtes et la difficulté de sélectionner les sens corrects pour les termes ambigus. En fait, plus la requête est longue plus la performance de RI est détériorée [Elayeb et al., 2015b].

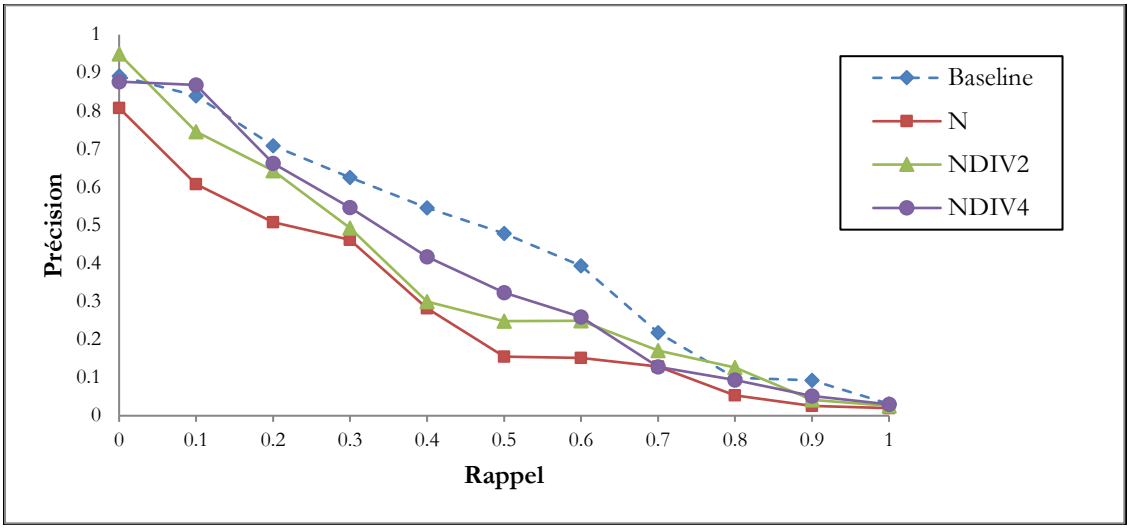

Figure 4.2 : Courbes de rappel-précision pour l'évaluation de l'ESR

4.3. Evaluation de l'approche possibiliste de DSR

Dans cette section, nous expérimentons l'efficacité de la DSR en utilisant l'approche possibiliste décrite dans la section 3. Nous ne considérons que le sens ayant le meilleur score de DPP selon le calcul basé sur les graphes de cooccurrence possibiliste. Ensuite, nous avons procédé à une évaluation des experts pour la pertinence du sens choisi en fonction de la requête initiale et étiquetée par trois degrés de pertinence : 1 (pertinent), 0 (partiellement pertinent) ou -1 (non-pertinent). Après avoir appliqué notre processus de DSR sur les 15 requêtes de tests ambiguës, nous avons identifié 5 sens pertinents et 4 sens non-pertinents. Cette évaluation a été réalisée manuellement à cause de l'absence de l'étiquetage des mots des contextes ambigus dans la collection ROMANSEVAL selon les requêtes de la collection CLEF-2003.

4.4. Combinaison des deux approches possibilistes de DSR et d'ESR

Cette série d'expérimentations consiste à appliquer la DSR avant l'ESR. Cette tâche peut aider à sélectionner les meilleurs sens des mots ambigus avant d'appliquer le processus d'ESR visant à réduire le bruit. Par conséquent, les termes composant le sens sélectionné sont injectés dans la requête et un processus d'ESR est ensuite appliqué (expérience *WSD_QE*). Nous avons également appliqué la technique de pseudo-réinjection de pertinence (RF) dans nos expériences à la fin des deux tâches de désambiguïsation et d'expansion (expérience *WSD_QE_RF*). Pour toutes les

requêtes reformulées dans la figure 4.3 (ajout de *N* termes), figure 4.4 (ajout de *N div 2* termes) et la figure 4.5 (ajout de *N div 4* termes), la combinaison de la DSR et l'ESR possibiliste a réalisé une amélioration de performance en comparaison avec les résultats d'ESR sans DSR. Néanmoins, les résultats des deux expériences *WSD_QE* et *QE* sont au-dessous du niveau de référence (*Baseline*). Cependant, lors de la combinaison de DSR, ESR et pseudo-réinjection de pertinence (expérience *WSD_QE_RF*), nous observons une meilleure performance de RI en particulier pour un nombre limité de termes d'expansion (cf. figure 4.4 et figure 4.5).

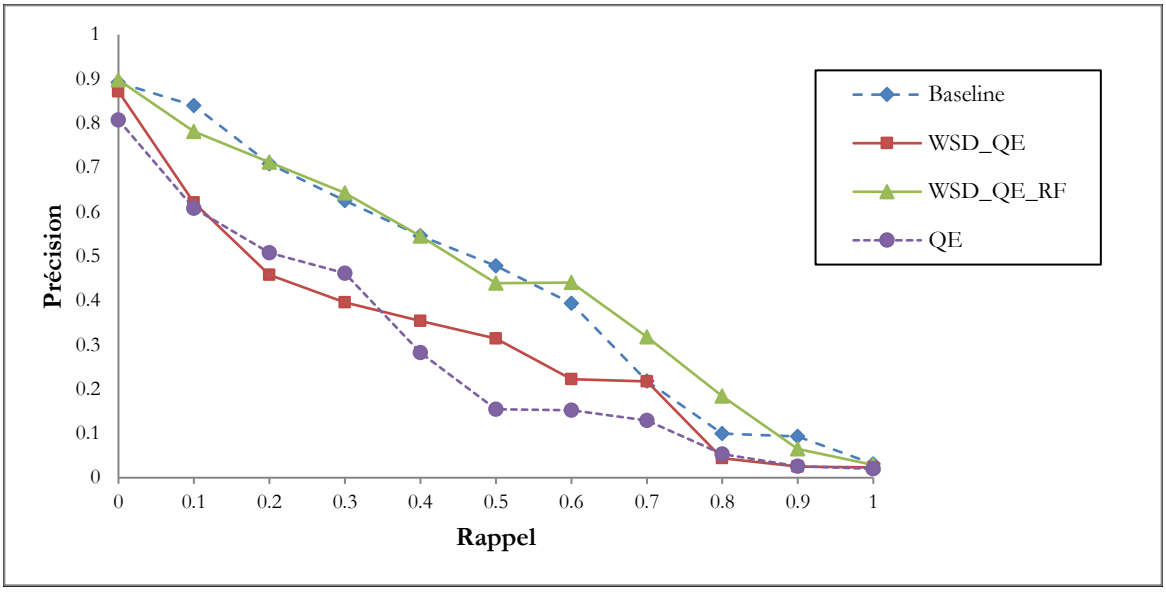

Figure 4.3 : Courbes de rappel-précision en ajoutant *N* termes avec et sans DSR

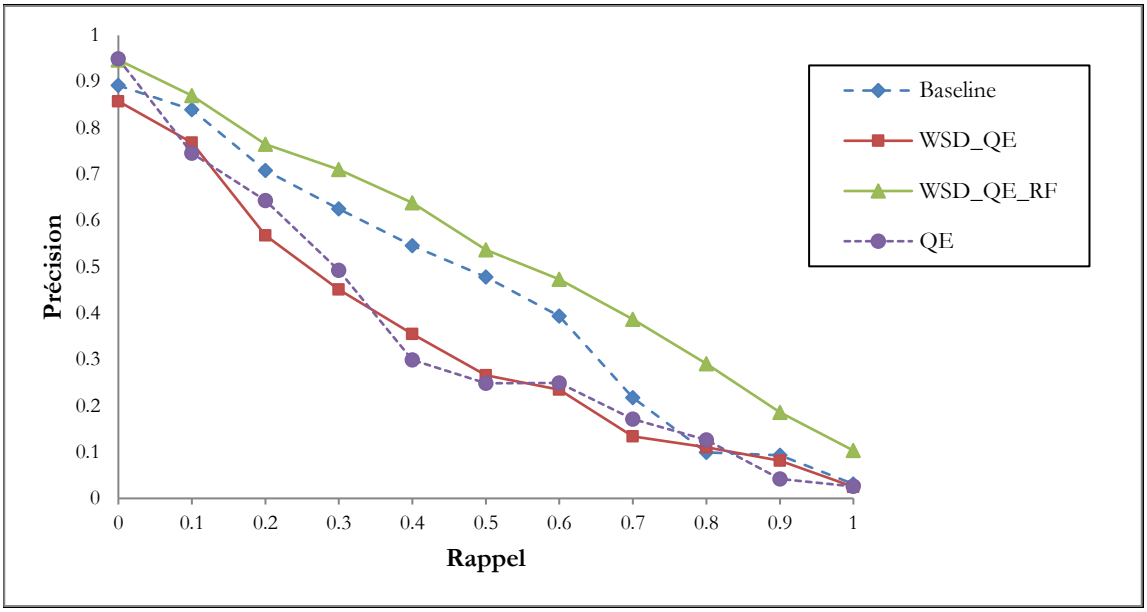

Figure 4.4 : Courbes de rappel-précision en ajoutant (*N div 2*) termes avec et sans DSR

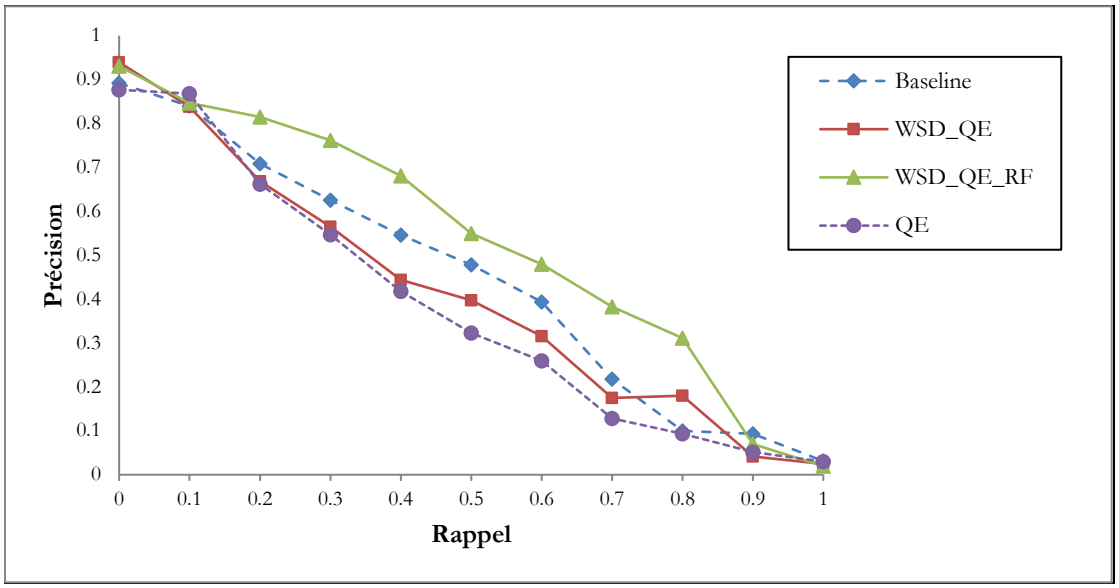

Figure 4.5 : Courbes de rappel-précision en ajoutant (*N div 4*) termes avec et sans DSR

Selon les trois scénarios, nous pouvons confirmer l'impact positif de la DSR sur l'ESR principalement pour les niveaux de rappel initial (<10%). La combinaison de la pseudo-réinjection de pertinence avec la DSR et l'ESR a également contribué à l'amélioration de la performance globale de RI. Le même impact positif de réinjection de pertinence a été observé par [Paskalis et Khodra, 2011]. Nous confirmons également les interprétations des travaux de [Pinto et Pérez-Sanjulian, 2008] qui ont étudié la performance de RI selon les requêtes courtes et longues qui peuvent générer un bruit tout en appliquant un processus d'ESR.

4.5. Comparaison à une approche à base de circuits

Nous avons détaillé dans le chapitre 3 notre approche d'ESR à base de dénombrement de circuits à partir d'un graphe de dictionnaire afin de calculer un score de proximité sémantique entre les termes nœuds du graphe. Nous rappelons ici la formule de calcul de ce score :

$$Proximité\ Sémantique(t_i, t_j) = \frac{Nombre\ de\ circuits(t_i,t_j)}{Nombre\ maximum\ de\ circuits\ dans\ le\ graphe} \quad (4.8)$$

Avec : *Nombre de circuits*(t_i, t_j) représente le nombre de circuits existants en partant du nœud t_i et passant par le nœud t_j dans le graphe du dictionnaire (i.e. $t_i \rightarrow \ldots \rightarrow t_j \rightarrow \ldots \rightarrow t_i$).

Pour le processus de DSR à base de circuits, nous définissons un score de proximité sémantique entre un sens S_i correspondant à un terme t_k de la requête Q par la formule suivante [Elayeb et al., 2015b]:

$$Proximité\ Sémantique(S_i, Q) = \sum_{s_{ij} \in S_i} \sum_{t_k \in Q} Proximité\ Sémantique(s_{ij}, t_k) \quad (4.9)$$

Le tableau 4.2 récapitule nos résultats et les essais effectués pour les deux approches possibiliste et à base de circuits. En effet, nous avons effectués les deux scénarios suivants : (i) application du processus d'ESR unique (expériences : *Poss_QE* et *Circuit_QE*); et (ii) application du processus de DSR avant l'ESR (expériences : *Poss_QD&E* et *Circuit_QD&E*). L'expérience *Baseline* correspond aux essais de la requête originelle sans les deux processus de DSR et d'ESR.

Nous remarquons à partir des résultats du tableau 4.2, que l'application du processus d'ESR détériore les performances globales de RI pour tous les tests effectués et pour les deux approches possibiliste et à base de circuits. L'approche possibiliste d'ESR (*Poss_QE*) a effectué une légère amélioration, en termes des deux métriques MAP et R-précision, par rapport à l'approche d'ESR à base de circuits (*Circuit_QE*). En outre, l'approche possibiliste (*Poss_QD&E*) semble être meilleure que celle à base de circuits (*Circuit_QD&E*) dans les cas d'application du processus de DSR avant l'ESR, qui a contribué davantage à l'amélioration des résultats de la recherche. En fait, cette performance globale faible d'ESR (avec et sans DSR) par rapport à l'essai *Baseline* pourra s'expliquer par la génération de bruit dans les résultats de recherche. Pour résoudre ce problème, nous avons limité le nombre de termes d'ESR au quart de la longueur *N* de la requête (*N div 4*) [Elayeb et al., 2015b].

		MAP	**R-précision**
Approche possibiliste	Poss_QE	0.5083	0.4742
	Poss_QD&E	**0.5124**	**0.4760**
Approche à base de circuits	Circuit_QE	0.4920	0.4633
	Circuit_QD&E	0.5071	0.4642
Baseline		0.5487	0.5174

Tableau 4.2 : Résultats de comparaison de deux approches possibiliste et à base de circuits

Nous superposons dans la figure 4.6 les courbes de rappel-précision des expériences effectuées toute en introduisant deux nouveaux essais impliquant un processus de pseudo-réinjection de pertinence (expériences : *Poss_QD&E_RF* et *Circuit_QD&E_RF*) après avoir réalisé les deux processus de DSR et d'ESR [Elayeb et al., 2015b].

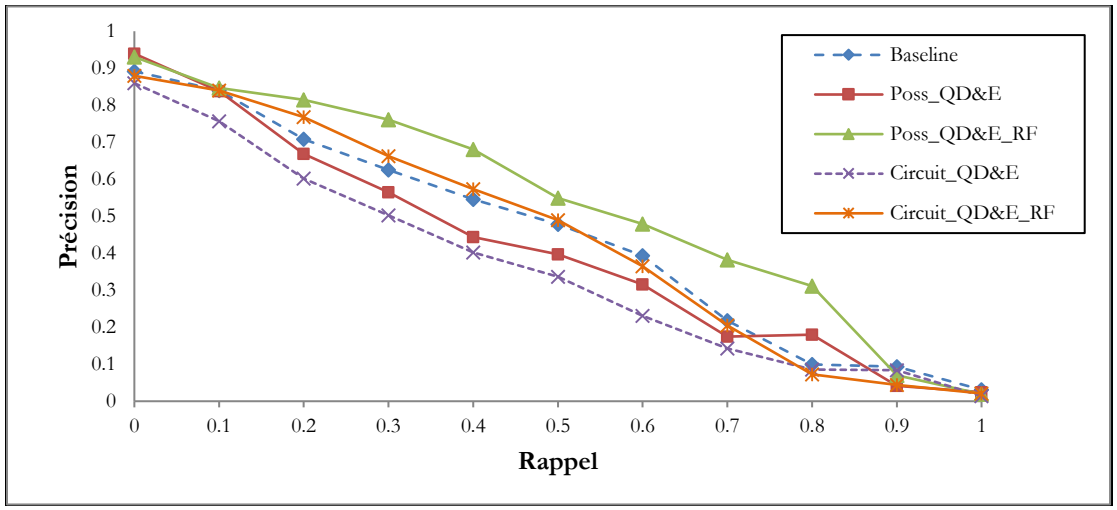

Figure 4.6 : Courbes de rappel-précision de l'étude comparative de différents tests

Si nous nous focalisons sur le scénario du test "*Poss_QD&E*", nous remarquons que l'ESR combiné avec la DSR dépasse la performance du Baseline pour des niveaux de rappel élevés. De plus, et pour les deux approches possibiliste et à base de circuits, l'application du processus de pseudo-réinjection de pertinence (RF) améliore davantage la performance globale de la recherche. Mais, la performance de l'approche possibiliste dépasse celle à base de circuits. En effet, l'approche possibiliste affine la recherche de nouveaux termes (respectivement les sens) pour l'ESR (respectivement pour la DSR) en profitant d'une double mesure de pertinence (possible et nécessaire) entre les nœuds du graphe de cooccurrence.

5. Bilan des contributions et perspectives

Nous avons présenté dans ce chapitre une approche possibiliste pour étudier l'impact de la désambiguïsation sémantique de requêtes (DSR) sur leurs expansions (ESR). L'approche a été appliquée pour la langue française pour vérifier de nombreux scénarios de traitement de la requête, mais elle est également applicable à d'autres langues. Dans un premier temps, nous avons préparé un graphe de cooccurrence à partir de la collection de documents. Ensuite, cette ressource a été utilisée dans la sélection de sens/termes utiles à la fois pour la DSR et l'ESR. Nos résultats ont confirmé que la DSR est une étape nécessaire dans le processus RI afin de surmonter le problème de l'ambiguïté des termes de la requête avant son expansion. En outre, la pseudo-réinjection de pertinence permet d'améliorer davantage la performance de l'approche combinant la DSR et l'ESR. Toutefois, le rendement du processus d'ESR est détérioré lors de l'utilisation de nombreux termes d'expansion. Ce phénomène est interprété par l'effet du bruit causé par les connaissances issues du graphe de cooccurrence.

Afin de prouver l'efficacité de l'approche possibiliste, nous avons comparé cette dernière à une approche à base de circuits. Nos résultats ont montré que l'approche possibiliste est plus efficace dans l'amélioration de la performance globale de RI. Cette approche semble plus fine grâce à son exploitation de la double mesure de pertinence (possible et nécessaire) ce qui permet d'accroître les scores des sens/termes pertinents en pénalisant les scores des sens/termes restants.

A court terme, nous envisageons de comparer l'utilisation des connaissances issues de l'extraction de documents (graphe de cooccurrence) à d'autres ressources linguistiques externes telles que les dictionnaires, les thésaurus, les ontologies, etc. Nous visons à étudier également l'efficacité de nos approches dans un cadre translinguistique en utilisant les standards usuels des compétitions *SensEval* et *SemEval*. Enfin, nos algorithmes de traitement de requêtes fondées sur les graphes ont été implémentés d'une manière générique indépendamment de la langue ; ce qui nous encourage à tester nos approches sur d'autres langues telles que l'anglais et l'arabe.

Chapitre 5 :
Etude et évaluation de la fiabilité de l'information recherchée

1. Introduction

Le grand nombre de fournisseurs d'information sur Internet ainsi que l'énorme quantité d'information disponible ont donné naissance au sujet de fiabilité de l'information. La validation de l'information, tâche confiée aux auteurs, aux éditeurs et aux bibliothécaires, est désormais effectuée par l'utilisateur [Vignaux, 2005]. Dans de nombreuses situations, ce dernier est incapable d'identifier la source d'information ou de juger de sa crédibilité, en particulier, lorsque de nombreux acteurs participent à sa production et/ou sa transmission. C'est le cas de certains forums où beaucoup d'informations sont échangées sans aucun moyen d'identification de leurs sources ou de leur acheminement. Cette situation soulève des doutes. La cause de ce problème est l'absence d'un organisme de réglementation qui surveille la fiabilité de ce qui se publie sur Internet. La question est alors: comment évaluer la fiabilité de l'information recherchée ?

Au cours des dernières années, Zacklad (2007) a introduit le Web socio-sémantique dont le but est d'*étudier les interactions sociales et la façon dont elles ont conduit lors de la création de représentations des connaissances explicites et sémantiquement riches* [Gruber, 2006]. En ce qui concerne la fiabilité, Zacklad (2007) a mentionné que l'identification de l'auteur est nécessaire pour la compréhension, l'interprétation et l'exploitation de documents. Un utilisateur ne peut pas profiter d'un document s'il n'a pas confiance en son auteur. En outre, l'utilisateur doit avoir la même confiance vis-à-vis des acteurs qui ont transmis ces informations. Ainsi, il devient clair que l'actualité n'est pas le seul critère de jugement de pertinence [Xu et Chen, 2006]. Selon Da Costa Pereira et Pasi (2007), la pertinence d'un document dépend de la notion de fiabilité qui est dépendante de "l'auteur", "la source" et "l'utilisateur". C'est pourquoi nous pensons que l'identification des acteurs (producteurs et émetteurs) et l'étude de leurs biographies sont indispensables avant toute lecture de documents.

Bien que le problème de la fiabilité de l'information résulte de l'expansion de l'Internet, les savants arabes se sont intéressés à ce sujet pendant de nombreux siècles. Leurs efforts ont été concentrés sur l'assurance de la transmission fiable des événements et des discours historiques entre les individus. Ils ont établi un ensemble de règles de transmission de l'information et d'étude de la fiabilité que nous appelons la "méthodologie de la narration arabe". Une description de cette méthodologie est présentée par [Lucas, 2002]. Selon cette méthodologie, la fiabilité de l'information est liée à la crédibilité (ou à la réputation) des acteurs qui ont participé à sa production ou à sa transmission et les relations qui existent entre eux. Pour évaluer la fiabilité d'une narration, la première étape est donc l'identification de ses acteurs et de leurs relations. La deuxième étape consiste à étudier la biographie de chaque acteur pour déterminer sa réputation. Dans le passé, ce travail a été fait manuellement. La réputation des acteurs a été jugée par les experts du domaine, qui attribuent un degré de confiance à chaque narrateur.

La méthodologie de la narration arabe fournit des solutions au problème de la fiabilité de l'information. La fiabilité d'une narration dépend de la crédibilité de ses narrateurs. Pour assurer la vérification de la fiabilité, les noms des narrateurs, explicitement cités à l'entête de la narration, constituent sa chaîne de narrateurs. Les narrations ont été signalées d'une génération à l'autre pour assurer la transmission fiable de la connaissance historique. Nous présentons dans ce chapitre un ensemble d'outils basés sur la méthodologie de la narration arabe. Nous commençons par présenter cette méthodologie comme un ensemble de principes pour l'évaluation de la fiabilité de

l'information. Ensuite, nous détaillons une architecture conçue pour appuyer l'étude de la fiabilité des narrations arabes. En effet, nous avons développé des grammaires pour l'analyse des noms propres ainsi que les chaînes de narrateurs des narrations arabe. Pour cela, nous proposons un outil intelligent de reconnaissance de l'identité reliant les noms trouvés dans les chaînes de narrateurs aux biographies des personnes correspondantes stockées dans une base de données[11]. Nous modélisons cette étape comme une tâche de recherche d'information possibiliste. Enfin, les chaînes sont analysées à travers les méta-données disponibles dans les biographies afin d'aider l'utilisateur à identifier les sources qui dénotent le manque de fiabilité. Nous proposons d'identifier la classe de fiabilité d'une narration via un classifieur possibiliste. Les résultats obtenus pour les entités nommées et la reconnaissance de l'identité ont été satisfaisants et confirment les objectifs fixés pour les métriques d'évaluation rappel, précision et F-mesure [Bounhas et al., 2010 ; 2015a]. En outre, les outils développés sont des composants réutilisables qui peuvent être utilisés pour étudier la fiabilité des autres types de textes arabes.

Ce chapitre est structuré comme suit : Dans la section 2, nous discutons le problème de la fiabilité de l'information tel qu'il est introduit dans la littérature récente et dans la méthodologie des narrations arabes. Puis, nous étudions dans la section 3 les exigences de la spécification d'une approche d'étude de la fiabilité des narrations arabes assistée par ordinateur. L'architecture de cette approche est présentée dans la section 4. L'analyse automatique des livres de Hadith est décrite dans la section 5. La section 6 détaille la reconnaissance des identités arabes. Dans la section 7, nous validons les différentes étapes de notre approche par une suite de résultats expérimentaux. Enfin, nous concluons notre travail, la section 8, et nous proposons quelques pistes de futures recherches.

2. La fiabilité dans la méthodologie de la narration arabe

La méthodologie de la narration arabe a été fondée selon les principes solides de la fiabilité. Basée sur l'étude de l'identité et de comportement, cette méthodologie fournit une base pour juger la réputation des participants à la transmission ainsi que leurs sources d'information. En effet, une narration arabe rapporte les discours, les actions ou les titres associés à une personne. Dans une narration arabe, les histoires ont été transmises entre les générations par des personnes qui sont appelées narrateurs. Etant donné que ces narrations rapportent des événements historiques importants, les experts du domaine ont établi des règles strictes pour la transmission de l'histoire.

Tout d'abord, pour assurer la dimension "*autorité*", le narrateur est obligé de citer la liste des personnes dont il a obtenu son histoire ou sa narration. Ainsi, chaque narration arabe est précédée par une chaîne de narrateurs. Ensuite, quand un narrateur (le Sheikh) communique une narration à son disciple (son élève), il utilise des verbes qui indiquent comment il a obtenu cette narration de son prédécesseur (son Sheikh) que nous appelons le mode de transmission. La narration est donc précédée d'une information exhaustive sur la circulation de l'information qui évalue sa *vérifiabilité*.

Pour être acceptée, une narration doit avoir une chaîne de narrateurs composée de personnes crédibles. De plus, il ne devrait y avoir aucun écart temporel et/ou géographique entre les deux narrateurs successifs, ce qui signifie qu'ils doivent avoir vécu dans la même période et s'être rencontré. La crédibilité de l'information est ensuite évaluée par l'étude de la biographie ou le comportement du narrateur. Nous devons mentionner que l'étude du comportement est effectuée par des chercheurs spécialisés reconnus comme des experts du domaine. Ils représentent les autorités de contrôle ou organismes de régulation chargés de juger la crédibilité des narrateurs.

[11] http://www.arbdownload.com/2009/04/29/gu-sz-zbnpnl.html

Enfin, l'histoire doit être exempte de biais, ce qui signifie que les narrateurs ne devraient pas avoir de raisons politiques ou théologiques pour falsifier la narration. Ensuite, la narration est transmise à partir d'un narrateur à l'autre sans changement. Pour être sûr de l'*objectivité* de la narration, les différentes versions de cette même narration (rapportée par différents narrateurs) sont comparées et les anomalies sont identifiées. Ainsi, nous concluons que toutes les dimensions et les exigences de la fiabilité sont prises en compte dans la méthodologie de la narration arabe. Nous remarquons également que la chaîne de transmission est la composante la plus importante pour l'étude de la fiabilité de telle narration.

3. Spécification de l'approche

Notre objectif est de développer une application, qui accepte en entrée une chaîne de narrateurs et décide de sa fiabilité en attribuant une classe de fiabilité. Nous commençons par déterminer les méthodes d'évaluation appropriées pour notre cas (cf. section 3.1). Etant donné que les noms propres arabes sont les principaux composants de la chaîne, nous présentons dans la section 3.2, les composants de noms propres arabes. Dans la section 3.3, nous étudions l'ensemble de la structure des chaînes de narrateurs définissant ainsi les principaux concepts de notre application.

3.1. Méthodes d'évaluation des dimensions de fiabilité de la narration

En se basant sur notre étude détaillée à la section 2, nous identifions quatre dimensions requises pour évaluer la fiabilité des narrations arabes à savoir l'*autorité*, l'*objectivité*, la *vérifiabilité* et la *fiabilité de transmission*. Les méthodes d'évaluation et les paramètres/outils nécessaires pour chaque dimension sont résumés dans le tableau 5.1 [Bounhas et al., 2015a].

Dimension	Méthode d'évaluation	Paramètres/outils
Autorité	Entrées de l'expert	Jugements des experts à propos de la crédibilité des narrateurs.
Objectivité	Analyse du contenu	Comparaison de versions.
Vérifiabilité	Analyse de la structure	- Analyse des chaînes de narrateurs et reconnaissance des entités nommées. - Reconnaissance de l'identité.
Fiabilité de transmission	- Analyse de la structure - Entrées de l'expert	

Tableau 5.1 : Méthodes, paramètres et outils d'évaluation des dimensions de la fiabilité des Hadiths

En effet, notre modèle profite des méta-données riches contenant des évaluations faites par des experts du domaine sur la réputation de chaque narrateur. Les travaux existants utilisent des méta-données souffrant de manque de connaissances suffisantes sur le comportement des acteurs. Par exemple, Stvilia [Stvilia et al., 2007 ; Stvilia, 2008] a proposé un cadre général pour l'évaluation de la qualité de l'information, mais n'a pas expliqué comment juger la réputation d'un participant à une chaîne de narration. Pour les articles de Wikipédia, l'auteur n'a considéré que si l'éditeur d'un article est inscrit ou non. D'autre part, Lynch (2001) a mentionné que l'identité de la source d'information n'est pas suffisante pour sa fiabilité. Une étude du comportement doit être effectuée afin d'évaluer la source.

Certaines approches délèguent l'évaluation de la fiabilité (ou certaines de ses dimensions) à l'utilisateur final [Richardson et al., 2003 ; Da Costa Pereira et Pasi, 2007]. Bien que cette approche considère l'opinion de l'utilisateur final, l'évaluation des sources d'information est une tâche difficile. D'autre part, l'évaluation de la fiabilité pourra être effectuée par des experts du domaine qui soutiennent l'activité de l'utilisateur final.

Ainsi, nous préférons les méthodes d'évaluation basées sur l'avis de l'expert car les experts du domaine qui évaluent les narrateurs sont des chercheurs reconnus. En outre, ils sont en mesure d'évaluer le comportement et la crédibilité des narrateurs, étant donné qu'ils vivaient à peu près qu'en même période. L'analyse du contenu est également une méthode précise, car la structure de la chaîne de narrateurs suit souvent des conventions qui permettent d'identifier ses sous-éléments. Cependant, notre travail dans ce chapitre est limité à des chaînes de narrateurs. L'analyse du contenu de la narration nécessite des outils linguistiques sophistiqués (par exemple, les analyseurs syntaxiques et sémantiques) qui ne sont pas disponibles pour la langue arabe. En outre, les évaluations de la crédibilité faites par les chercheurs prennent souvent en compte l'objectivité des narrateurs.

3.2. Structure d'un nom propre arabe

Un nom propre arabe est un terme composé mais dont la structure diffère de celle des syntagmes dont nous avons fait la typologie dans [Bounhas, 2012]. En outre, la structure d'un nom propre arabe diffère complètement de celle d'un nom propre dans une autre langue. En effet, comme c'est reporté dans [Shaalan et Raza, 2007 ; Bounhas et al., 2010], un nom propre arabe est une combinaison des éléments suivants :

- Le **prénom** (الإسم) : un nom propre personnel attribué à la naissance (Par exemple "Adam"). Dans certains cas, il est composé par le mot "عبد" suivi de l'un des noms de Dieu comme "الله" (Allah) ou une autre vertu comme "العزى" (alozza).
- La **konia** (الكنية) : généralement c'est une référence attribuée au premier fils de la personne en utilisant le terme "أبو" (père de) ou "أم" (mère de). Par exemple : "أبو علي" (père de Ali) est la konia d'un homme dont le premier fils s'appelle "علي" (Ali). Dans d'autres cas, il est attribué pour d'autres raisons.
- Le **Nasab** (النسب) : indique les antécédents de la personne en utilisant le terme "ابن" (fils de) ou "بنت" (fille de). Par exemple, une personne nommée "آدم" (Adam) et dont le père s'appelle "أحمد" (Ahmed) est référencé par " آدم بن أحمد" (Adam fils de Ahmed).
- Le **laqab** (اللقب) : une description, souvent religieuse, d'une personne qui indique par exemple l'une de ses qualités. Exemple : "الرشيد" (sensé ou rationnel).
- La **nisba** (النسبة) : un nom dérivé de la tribu, la profession, le lieu de résidence ou de naissance ou de l'affiliation religieuse. Exemples : "النجار" (Al-Najjar: le menuisier), "التونسي" (Al-Tounsi : le Tunisien).

D'autres expressions structurées des noms propres des narrateurs de notre base de données. Par exemple, le terme "مولى" précède la lige de la personne qui est aussi un nom propre en arabe. Nous présentons dans la suite de ce chapitre plus de détails sur ces expressions dans la grammaire du nom propre arabe.

3.3. Structure des chaînes de narrateurs

Une chaîne de narrateurs est typiquement composée de verbes indiquant la manière de transmission (voir section 3.3.1) et de noms de personnes (voir section 3.3.2). Avec cette simple définition, l'analyse d'une telle chaîne est relativement simple. Cependant, nous remarquons que le narrateur est libre d'ajouter des expressions ou des commentaires en rapportant un Hadith. La chaîne n'est donc pas une liste de noms propres et de verbes mais possède une structure complexe qui peut contenir différents types d'informations (voir section 3.3.3).

3.3.1. La manière de transmission

L'utilisation de ces verbes dans les chaînes de narrateurs affecte leur structure particulièrement au niveau des noms des narrateurs. Selon le verbe et/ou les prépositions et leurs positions, ces noms changent de mode [Bounhas et al., 2010].

Prenons l'exemple suivant : "عن **أحمدٍ** حدثنا **صالحٌ** أن **جابراً** أخبره".

Qui peut être traduit comme suit : "Selon Ahmed, Saleh lui a dit que Jeber l'a informé".

Dans cet exemple, la chaîne de narrateurs est composée de trois personnes: "**أحمد**" (Ahmed), "**صالح**" (Saleh) et "**جابر**" (Jeber). Etant donné que la préposition "عن" (selon) précède le premier nom, la voyelle courte "اً" lui est ajouté. Le deuxième nom est en mode nominatif puisqu'il représente le sujet d'une phrase verbale dont le verbe est "حدثنا" (il nous a dit). Dans le dernier cas, le verbe "أخبر" (informer) vient après le nom du narrateur qui est en mode accusatif et prend à sa fin une lettre supplémentaire et une voyelle courte ("اً").

3.3.2. Les noms de narrateurs

Dans une chaîne, un narrateur peut être référencé par plusieurs expressions correspondant à une ou plusieurs composantes de son nom. Ceci implique que la même personne peut être référencée de plusieurs manières différentes ce qui complique son identification. Par exemple, plusieurs personnes ont la *konya* "أبو علي" (abou Ali) parce que le nom Ali est largement utilisé et donc ambigu. Ces ambiguïtés peuvent être résolues si la personne est référencée en même temps par d'autres composantes de son nom.

Dans certains cas, les narrateurs sont référencés sans aucune composante de leurs noms. C'est le cas quand un narrateur indique qu'il a reçu le Hadith de l'un de ses proches. Par exemple, un narrateur peut rapporter qu'il a reçu un Hadith de son grand-père comme suit : "حدثني جدي" (Mon grand-père m'a dit). Dans certains autres cas, les relations sociales sont combinées avec les noms. Par exemple, quelqu'un pourra rapporter comme suit : "حدثني أخي أحمد" (mon frère Ahmed m'a dit). En plus des liens de parenté, d'autres types de relations peuvent être invoquées. Par exemple, un narrateur peut rapporter qu'il a reçu le Hadith d'un ami. Enfin, un narrateur peut citer deux ou plusieurs de ces Sheikhs en utilisant les conjonctions "و" (et) et "أو" (ou). La première est utilisée si le narrateur a reçu le Hadith de deux ou plusieurs personnes à la fois et la deuxième est utilisée quand il est douteux.

3.3.3. Les informations supplémentaires dans les chaînes de narrateurs

Les chaînes de narrateurs peuvent contenir plusieurs types d'informations autres que les noms de narrateurs et les verbes indiquant la manière de transmission. Nous pouvons citer les principaux types comme suit :

- Les expressions spécifiant le cadre spatio-temporel ou décrivant la situation lors de la transmission du Hadith.
- La description du narrateur utilisée par exemple pour décrire ou confirmer sa crédibilité.
- Le caractère "ح" qui indique que la chaîne est composée de deux sous-chaînes ce qui signifie que le narrateur a reçu le Hadith de deux sources différentes. Dans l'exemple suivant, le narrateur a reçu l'histoire de deux personnes ("يعقوب بن إبراهيم" (yakoub fils d'Ibrahim) et "آدم" (Adam)) qui l'ont reçu de deux personnes différentes ("بن علية" (son of Olaya) et "شعبة" (Cho'ba)).

حدثنا **يعقوب بن إبراهيم** قال حدثنا **بن علية** عن **عبد العزيز بن صهيب** عن **أنس ح** وحدثنا **آدم** قال حدثنا **شعبة** عن **قتادة** عن **أنس**

Qui peut être traduit comme suit :

"**Yakoub** fils d'**Ibrahim** nous a dit que le fils d'**Olaya** lui a dit selon **Abdelaziz** fils de **Sohayb** selon **Anas *H*** et **Adam** nous a dit que **Cho'ba** lui a dit selon **Katada** Selon **Anas**"

- Autres commentaires : phrases reliées au contenu du Hadith.

Sur la base de notre étude des méthodes d'évaluation et la structure des chaînes, nous présentons dans la section suivante l'architecture de notre approche.

4. Processus d'évaluation de la fiabilité d'une chaîne de narrateurs

Notre objectif est de développer des outils permettant l'analyse des chaînes de narrateurs de narrations arabes afin d'appliquer la méthode de fiabilité de l'information. Cependant, ces outils doivent être génériques pour traiter d'autres types de textes. Comme le montre la figure 5.1, l'architecture proposée est composée de cinq composants, qui mettent en œuvre trois étapes distinctes :

Tout d'abord, nous avons développé un processus de reconnaissance des entités nommées capable d'analyser une chaîne de narrateurs ou un nom propre arabe en générant sa structure logique au format XML. Puis, nous avons utilisé deux analyseurs de structures différentes pour indexer les chaînes de narrateurs et les noms propres arabes (respectivement). Ensuite, nous avons identifié les personnes qui ont rapporté une narration en faisant correspondre les indexes de noms trouvés dans la chaîne de narrateurs et les indexes de noms dans la base de données de biographies en utilisant les réseaux possibilistes. Enfin, nous avons évalué la fiabilité de la chaîne via notre outil possibiliste de cartographie de chaînes de narrateurs. Il s'agit d'un analyseur de chaînes chargé de calculer des scores possibilistes correspondant à des critères de fiabilité, qui identifie la classe de fiabilité et apporte des causes légères de manque de fiabilité/suspect dans la chaîne. L'ensemble du processus prend en entrée des méta-données des narrateurs. Nous devons également, dans les étapes intermédiaires, modéliser les noms et les chaînes.

5. Analyse automatique des livres de Hadith

Les noms des narrateurs, les Hadiths et les titres des thèmes sont analysés en utilisant des grammaires hors contexte apprises d'une manière semi-automatique. Tout d'abord, nous avons commencé par l'analyse des noms des narrateurs existant dans la base des biographies. Ensuite, les noms des personnes dont les titres et les Hadiths ont été reconnus en utilisant la grammaire apprise. D'autre part, et en utilisant les 400 premiers Hadiths de livres Sahih Al-Bukhari (صحيح البخاري), Sahih Muslim (صحيح مسلم), Sunan Abi Dawud (سنن أبي داود), Sunan Ibn Majah (سنن ابن ماجة), Bounhas et Slimani (2009b) ont proposé une évaluation expérimentale réalisée sur 1600 Hadiths extraits de ces quatre livres. En effet, les auteurs ont utilisé 20% des Hadiths dans la phase du test, et le reste (1280 Hadiths) dans la phase d'apprentissage. Ils ont appris une suite de grammaires qui correspond aux éléments mis en gras dans la DTD illustrant la structure d'un livre de Hadith de la figure 5.2.

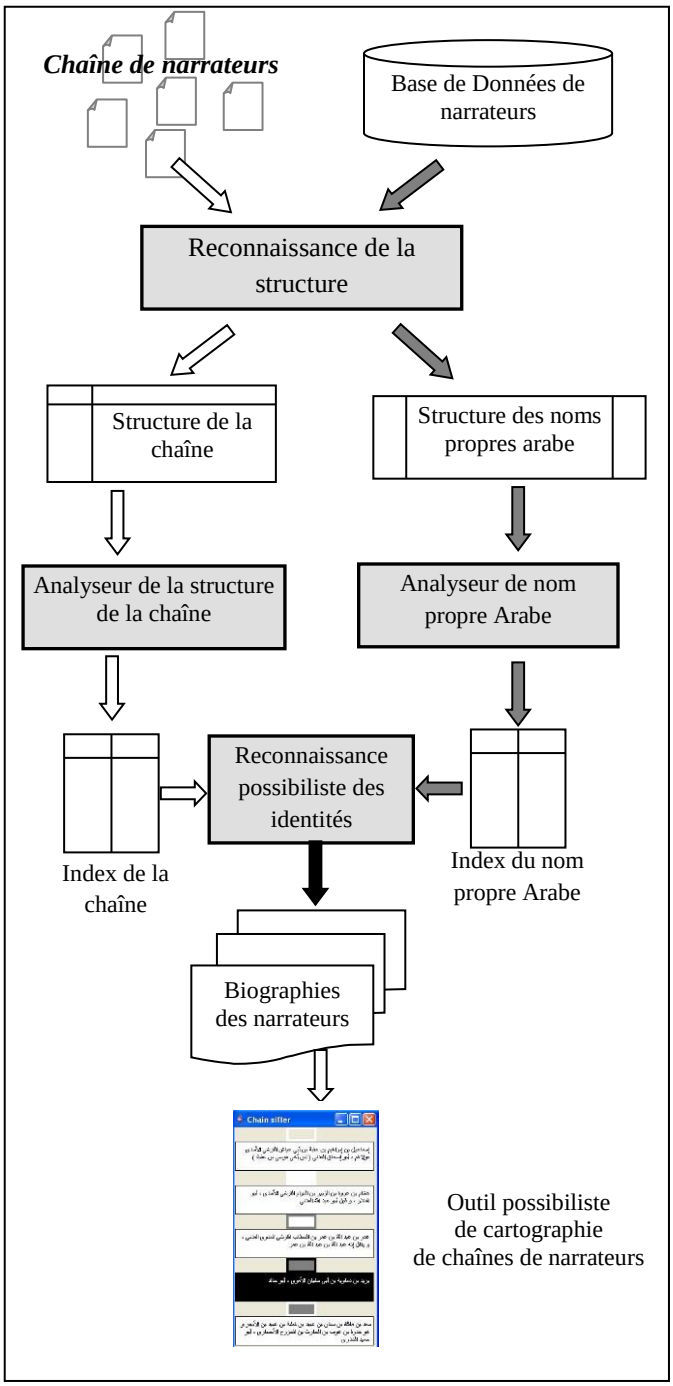

Figure 5.1 : Processus d'évaluation de la fiabilité d'une chaîne de narrateurs

Dans [Bounhas, 2012], nous avons présenté d'une manière détaillée les grammaires que nous avons obtenues. Nous tenons ici à mentionner que :

- La grammaire des acteurs englobe les différentes composantes d'un nom propre arabe. Elle tient compte aussi du nom du maître lorsqu'il est cité dans le nom de la personne.
- La grammaire des chaînes de narrateurs englobe les différentes configurations des références des narrateurs et des verbes de transmission. Elle modélise les différents types de références

dont le cas où le narrateur est référencé par une relation à une autre personne, ou à son prédécesseur dans la chaîne.

- Chacune des grammaires utilisées engendre un résultat au format XML, qui est utilisé par d'autres grammaires selon l'ordre de priorité. De plus, et en utilisant la dernière grammaire dans la table de priorité, le résultat final d'un analyseur micro-logique est structuré au format XML aussi.

```
<!ELEMENT LivreHadith (Theme+) >
<!ELEMENT Theme (Titre, (Verset, Theme | Hadith)+)>
<!ELEMENT Titre (Texte, Interpretation*)>
<!ELEMENT Hadith (Chaîne | Contenu | Commentaire | IndicationVersion | Verset)+>
<!ELEMENT Commentaire (CommentaireActeur | CommentaireFiabiliteHdith | Interpretation)>…
```

Figure 5.2 : DTD illustrant la structure d'un livre de Hadith

Ces grammaires sont utilisées par deux analyseurs micro-logiques. Les tableaux 5.1 et 5.2 illustrent respectivement les tables de priorité de ces deux analyseurs. En fait, le premier permet l'analyse des titres des thèmes, alors que le second permet l'analyse d'un Hadith avec toutes ses composantes.

Nom de la grammaire	Priorité	Type d'analyse
Verset	1	Partielle
Acteur	2	Partielle
Titre	3	Complète

Tableau 5.1 : Table de priorités de l'analyseur des titres des thèmes

Nom de la grammaire	Priorité	Type d'analyse
Verset	1	Partielle
Acteur	2	Partielle
Chaîne	3	Partielle
Indication_Version	4	Partielle
Commentaire_Acteur	5	Partielle
Commentaire_Fiabilite_Hadith	7	Partielle
Hadith	8	Complète

Tableau 5.2 : Table de priorités de l'analyseur des Hadiths

D'abord, nous identifions les versets coraniques puis les acteurs, qui sont les éléments les plus fins en termes de granularité. Nous percevons aussi que la grammaire "Acteur" et "Interprétation" sont communes aux deux analyseurs, ce qui confirme les possibilités de leurs réutilisations. En outre, nous avons utilisé les métriques d'évaluation rappel, précision et F-mesure pour évaluer nos résultats d'analyse. En fait, nous ne considérons valide dans cette évaluation que les éléments qui ont été identifiés et correctement analysés. Nous synthétisons dans le tableau 5.3 nos résultats obtenus par type de fragment.

Il est à noter que les chaînes et les noms suivent une structure plus ou moins régulière et les cas exceptionnels sont rares. Au contraire, l'identification des indications de versions, des commentaires et des interprétations sont plus délicates étant donné que les experts s'expriment librement en commentant les Hadiths. C'est pour ces raisons que les taux de reconnaissance des acteurs et des chaînes sont supérieurs à ceux obtenus pour les autres éléments. Le tableau 5.4 présente une étude comparative des approches existantes de reconnaissance des entités nommées par rapport à nos travaux.

Type de fragments	Rappel	Précision	F-mesure
Versets	100.00%	100.00%	100.00%
Acteurs	98.95%	97.24%	98.09%
Chaînes	97.96%	95.66%	96.79%
Indications de versions	94.54%	93.01%	93.77%
Commentaires	84.29%	85.51%	84.89%
Total	**98.43%**	**96.63%**	**97.52%**

Tableau 5.3 : Résultats d'expérimentation de l'analyseur des Hadiths

Année	Approche	Précision	Rappel	F-mesure
2005	[Zitouni et al., 2005]	75.30 %	70.20 %	72.70 %
2007	[Shaalan et Raza, 2007]	85.50 %	89.00 %	87.50 %
2009	[Shaalan et Raza, 2009]	86.30 %	89.20 %	87.70 %
2010	**Notre approche [Bounhas et al., 2010]**	**98.95 %**	**97.24 %**	**98.09 %**

Tableau 5.4 : Étude comparative des approches de reconnaissance des entités nommées

Nous remarquons que nos résultats prouvent une amélioration significative des performances. Ces approches ont utilisé des corpus constitués essentiellement d'articles de magazines et de journaux. En fait, les chaînes de narrateurs ont une structure plus régulière, ce qui explique l'amélioration réalisée. Toutefois, les travaux existants se limitent uniquement à l'identification des entités nommées. Alors que dans notre approche la reconnaissance des identités est facilitée grâce aux relations sociales achevées par la tâche d'analyse de structure de chaque entité. A notre connaissance, aucune des approches existantes n'a considéré la reconnaissance des entités nommées de cette manière.

6. Reconnaissance des identités

Dans [Bounhas et al., 2010], nous avons modélisé l'étape de reconnaissance des identités comme un SRI dont la requête est un nom extrait d'une chaîne et les documents sont les biographies des personnes stockées dans une base. Pour cela, nous avons proposé un modèle d'indexation pour la requête et pour les documents. Nous commençons d'abord par présenter le modèle d'indexation des noms propres arabes (cf. section 6.1) puis celui de chaînes de narrateurs (cf. sections 6.2). Notre SRI exploite notre modèle d'appariement possibiliste [Elayeb et al., 2009], qui permet d'évaluer chaque personne de la base étant donné un nom de la chaîne (cf. section 6.3). D'autre part, et étant donné l'ambiguïté des noms arabes, cette étape peut engendrer plusieurs personnes qui ont le même score possibiliste. En conséquence, une étape de filtrage (cf. section 6.4) s'avère nécessaire afin de résoudre ce problème d'ambiguïté.

6.1. Le modèle d'indexation des noms propres arabes

L'extraction de chaque nom propre et son indexation selon le modèle de la figure 5.3 résulte d'un parcours du code XML généré par l'analyseur micro-logique [Bounhas et al., 2010].

$$\text{nom} = (it_1, it_2, \ldots, it_n)$$
$$it_i = (c_i, v_i)+$$

Figure 5.3 : Modèle d'indexation des noms propres arabes

Notons it_1, it_2,…,it_n l'ensemble d'items constituant un nom propre arabe. Désignons par (c_i, v_i) les couples clé-valeur composant chaque item. En effet, les valeurs correspondent aux composantes d'un nom

propre arabe comme détaillé dans la section 3.2. Les clés désignent des symboles qui indiquent le type de chaque composante. Nous récapitulons dans le tableau 5.5 [Bounhas et al., 2010] toutes les valeurs et les désignations des clés. Dans ce modèle, les clés P_1,…, P_n correspondent aux noms des antécédents de la personne. P_1 correspond au père, P_2 au grand-père et ainsi de suite.

Composante	Clé
Le prénom (الإسم)	N
La *konia* (الكنية)	K
Le *laqab* (اللقب)	L
La *nisba* (النسبة)	B
Le *nasab* (النسب)	$P_1 \ldots P_n$
Le prénom du maître	MN
La *konia* du maître	MK
Le *laqab* du maître	ML
Le *nisba* du maître	MB
La *nasab* du maître	$MP_1 \ldots MP_n$

Tableau 5.5 : Composantes du modèle d'indexation des noms propres arabes

Considérons, comme exemple d'indexation, le nom arabe suivant:

وهب بن عبد الله و يقال **ابن وهب أبو جحيفة السوائى** يقال له **وهب الخير** وهو مولى **علي بن صالح**

Qui peut être traduit comme suit : "Wehb fils d'Abd Allah ou fils de Wehb Abou Jahifa Al-sawai appelé Wehb Al-kheyr allié d'Ali fils de Salah".

Son index est représenté par la figure 5.4 selon le modèle d'indexation défini ci-dessus.

(N,وهب)
(P_1, عبد الله) (P_1, وهب)
(K,أبو جحيفة)
(B,السوائى)
(L,وهب الخير)
(MN,علي)
(MP_1, صالح)

Figure 5.4 : Exemple d'index d'un nom propre arabe

Etant donné le doute à propos du père de la personne, nous remarquons que le second item de cet index contient deux paires : (P_1, عبد الله) et (P_1, وهب) [Bounhas et al., 2010].

6.2. Le modèle d'indexation des chaînes de narrateurs

Le modèle de la figure 5.5 [Bounhas et al., 2010] récapitule l'analyse de la structure d'une chaîne de narrateurs.

```
chain = (riwaya | separateur)+
riwaya = (tahamoul ?, Rawi)
rawi = (nom1?, Relation?, nom2?)
```

Figure 5.5 : Modèle d'indexation des chaînes de narrateurs

En effet,"*riwaya*" et "*separateur*" sont les deux types d'items qui composent une chaîne de narrateurs. Pour chaque item du premier type, nous calculons les deux attributs suivants :

- *Tahamoul* : la manière de transmission.
- *Rawi* : une référence à un narrateur qui contient un ou deux noms propres (indexés selon le modèle de la figure 5.3) et éventuellement une relation sociale comme décrit dans la section 3.3.2.

Prenons l'exemple d'indexation de la chaîne de narrateurs suivante illustrée par la figure 5.6 [Bounhas et al., 2010] :

حدثنا هشام **عن** أبيه **عن** مريم بنت أبي الرشيد

Qui peut être traduite comme suit : "Hichem **nous a dit**, **selon** son père, **selon** Myriam fille d'Abou Al-Rachid".

chain = riwaya1, riwaya2, riwaya3
riwaya1 = tahamoul: (sama3 : حدثنا)
name1: (N, هشام)
riwaya2 = tahamoul: (An: عن)
relation: (أبيه, Abouh)
riwaya3 = tahamoul: (An: عن)
name1: (N, مريم)
(P_1, بنت أبي الرشيد)

Figure 5.6 : Exemple d'index d'une chaîne de narrateurs

Dans cet exemple, riwaya1, riwaya2 et riwaya3 sont les trois éléments de type "riwaya" qui composent cette chaîne. Le premier riwaya1 possède deux attributs: "*tahamoul* : (sama3 : حدثنا)" qui signifie que le premier narrateur a communiqué l'histoire oralement et "(N, هشام)" qui indique le nom du narrateur. La manière de transmission du deuxième et du troisième composant est "*tahamoul* : (An: عن)" qui signifie "Selon". Le deuxième narrateur est référencé par une relation de filiation au précédent narrateur "(أبيه, Abouh)" (son père). Le dernier narrateur est une femme référencée par son prénom "(N, مريم)" et son père "(P1, أبي الرشيد)".

6.3. Le modèle d'appariement

Ce modèle permet de calculer le degré de similitude entre deux noms arabes dont le premier apparaît dans une chaîne et le deuxième est stockée dans la base des biographies. Notre solution permet de retrouver la biographie adéquate même si le nom est ambigu ou si des erreurs d'analyse ont eu lieu dans les étapes précédentes. Par exemple, le nom "وهب" (Wehb) peut être utilisé comme *laqab* ou comme prénom. Etant donné que l'outil de reconnaissance des entités nommées attribue un seul label à chaque lexème. Ceci implique que la clé "L" peut être remplacée par "N" et vice versa. Une autre ambiguïté concerne le nom du père. En effet, un narrateur peut être référencé par son prénom et le nom de son grand-père. Dans ce cas la clé "P_2" remplace la clé "P_1". Par exemple, le nom " آدم بن وهب" (Adam fils de Wehb) peut être interprété comme "Adam dont le père est Wehb" ou "Adam dont le grand-père est Wehb". Enfin, une personne peut hériter le *nisba* de son maître. La clé "B" peut donc remplacer la clé "MB". Nous modélisions ce problème par les tables de correspondance illustrées par les tableaux 5.6, 5.7 et 5.8 [Bounhas et al., 2010].

Les matrices doivent être lues de la gauche vers la droite. Ainsi, une cellule dont le fond est gris signifie que la clé en ligne peut être remplacée par la clé en colonne. Par exemple, dans la deuxième matrice la clé P_1 peut être remplacée par P_2. Par contre P_2 ne peut pas être remplacée par P_1.

	N	K	L	B	MN	MK	ML	MB
N								
K								
L								
B								
MN								
MK								
ML								
MB								

Tableau 5.6 : Matrice de correspondance (noms des personnes et des maîtres)

	P_1	P_2	...	P_i	P_{i+1}	...	P_{n-1}	P_n
P_1								
P_2								
...								
P_i								
P_{i+1}								
...								
P_{n-1}								
P_n								

Tableau 5.7 : Matrice de correspondance (clés des pères)

	MP_1	MP_2	...	MP_i	MP_{i+1}	...	MP_{n-1}	MP_n
MP_1								
MP_2								
...								
MP_i								
MP_{i+1}								
...								
MP_{n-1}								
MP_n								

Tableau 5.8 : Matrice de correspondance (clés des pères du maître)

Notons *Qname* un nom qui apparait dans une chaîne et $person_j$ une personne de la base. Nous indexons *Qname* et $person_j$ par un ensemble d'items conformément au modèle d'indexation de la figure 5.3. Nous avons alors :

Qname = (itQ_1, itQ_2,…,itQ_m) et $person_j$ = (itP_1, itP_2,…, itP_p)

Comme tout SRI possibiliste, notre outil encode des liens de dépendance entre les items de la requête et les personnes à travers un réseau possibiliste et quantifie ces liens par les deux mesures de possibilité et de nécessité [Bounhas et al., 2010]. Les personnes retrouvées sont celles qui sont possiblement ou nécessairement pertinentes étant donné le nom de la chaîne. Sachant que les items de la requête ne sont pas pondérés, la pertinence d'une personne ($person_j$) de la base, étant donné un nom *Qname*, est calculée comme suit :

L'expression $\Pi(person_j|Qname)$ est proportionnelle à :

$$\Pi'(person_j|Qname) = \pi(itQ_1|person_j) * \dots * \pi(itQ_m|person_j) = Freq(itQ_{1j}) * \dots * Freq(itQ_{mj}) \quad (5.1)$$

Dans cette formule, $Freq(itQ_{ij})$ est la fréquence de l'item numéro *i* de *Qname* (itQ_i) dans le nom de la personne numéro *j* de la base. Elle est calculée comme suit :

$$Freq(itQ_{ij}) = \begin{cases} 1 & \text{S'il existe un item } itP_k \text{ dans } person_j \text{ ayant la même clé et la même valeur que } itQ_i \\ 0.5 & \text{S'il existe un item } itP_k \text{ dans } person_j \text{ ayant la même valeur que } itQ_i \text{, et la clé de } itQ_i \text{ est différente de (mais peut être remplacée par) celle de } itP_k \\ 0 & \text{Dans les autres cas.} \end{cases} \quad (5.2)$$

En effet, la fréquence est fixée à 0.5 s'il y a une ambiguïté, car les deux items ne sont pas exactement équivalents.

La nécessité de retourner une personne $person_j$ pour un nom *Qname*, notée N($person_j|Qname$), est donnée par :

$$N(person_j|Qname) = 1 - \Pi(\neg person_j|Qname) \quad (5.3)$$

Avec:

$$\Pi(\neg person_j|Qname) = \frac{\Pi(Qname|\neg person_j)*\Pi(\neg person_j)}{\Pi(Qname)} \quad (5.4)$$

De même $\Pi(\neg person_j|Qname)$ est proportionnelle à :

$$\Pi'(\neg person_j|Qname) = \pi(itQ_1|\neg person_j) * \ldots * \pi(itQ_m|\neg person_j) \quad (5.5)$$

Ce numérateur peut être exprimé par :

$$\Pi'(\neg person_j|Qname) = (1 - \phi person_{1j}) * \ldots * (1 - \phi person_{mj}) \quad (5.6)$$

Avec:

$$\phi person_{ij} = Log_{10}\left(\frac{nDP}{nP_i}\right) * Freq(itQ_{ij}) \quad (5.7)$$

Où : - *nDP* est le nombre de personnes dans la base.

- nP_ireprésente le nombre de personnes dans la base dont $Freq(itQ_{ij})$ n'est pas nulle.

Le degré de pertinence possibiliste (*DPP*) n'est autre que la somme des deux scores de possibilité et de nécessité :

$$DPP(person_j|Qname) = \Pi(person_j|Qname) + N(person_j|Qname) \quad (5.8)$$

Les personnes sélectionnées sont ceux qui ont une valeur $DPP(person_j|Qname)$ élevée.

Nous considérons aussi le cas où le narrateur est référencé par une relation sociale avec une autre personne. Pour illustrer ce cas, nous prenons comme exemple le cas de la relation "père" : Quand un père transmet un Hadith à son fils, l'identification du premier requiert la reconnaissance du deuxième. Si le fils est identifié, nous procédons comme suit : (i) Générer l'index du fils à partir de la base, (ii) Générer l'index du père qui constitue une requête, et (iii) Appliquer le calcul possibiliste pour la nouvelle requête. En fait, le traitement de la relation "oncle" se fait d'une manière similaire. Quand un narrateur *A* transmet un Hadith au fils de son frère *B*, nous considérons que *A* doit avoir le même grand-père que *B*. Ce processus est appelé reformulation sociale de requêtes [Bounhas et al., 2010].

6.4. La fonction de filtrage

L'étape d'appariement génère, pour chaque narrateur dans la chaîne, plusieurs candidats pondérés par leurs scores de *DPP*. Mais, le traitement ne s'achève pas à ce stade. En effet, nous engendrons les chemins possibles entre les candidats de tous les narrateurs de la chaîne. Considérons, par exemple, une chaîne composée de deux narrateurs *A* et *B*. Supposons aussi que l'appariement génère deux candidats A_1 et A_2 pour *A* et deux autres B_1 et B_2 pour *B*. Nous avons donc quatre chemins possibles. Nous procédons à une étape de désambiguïsation permettant la sélection des chemins et des candidats valides, sachant que le nombre de chemins est combinatoire. Pour cela, nous calculons, pour chaque chemin, le nombre de liens valides. En fait, un lien entre deux narrateurs est dit valide s'il correspond à une relation "Sheikh-disciple" de la base. Nous choisissons le chemin ayant le nombre maximum de liens valides.

6.5. Les résultats d'évaluation

Dans [Bounhas et al., 2010], nous avons reporté des résultats d'évaluation qui concernent 200 Hadiths du livre Sahih Al-Bukhari (صحيح البخاري). Nous avons évalué les résultats de la reconnaissance des identités des narrateurs en utilisant les métriques rappel, précision et F-mesure comme illustré par le tableau 5.9.

Précision	80.88%
Rappel	98.97%
F-mesure	89.01%

Tableau 5.9 : Résultats de la reconnaissance des identités

En effet, nous avons réussi à identifier exactement les narrateurs dans 89.54% des cas. Dans 9.44% des cas, notre outil a retourné une liste qui contient la bonne personne. Il a échoué à retrouver la personne dans 1.02% des cas.

7. Évaluation de la fiabilité des Hadiths

Les Hadiths préalablement analysés et dont les chaîne sont indexées passent à la dernière étape d'évaluation de la fiabilité. En effet, la *crédibilité des narrateurs*, la *continuité de la chaîne* et la *fiabilité de transmission* sont les trois principaux facteurs de la méthodologie des sciences de Hadith. Dans notre évaluation, nous exploitons la théorie des possibilités pour classer le Hadith dans l'une de trois classes possibles: fiable (F), non fiable (NF) et suspect (S).

En premier lieu, nous définissons les distributions de possibilité de ces trois classes par rapport aux trois attributs en se basant sur la théorie des experts de Hadith (cf. sections 7.1, 7.2 et 7.3). En second lieu, nous utilisons une agrégation à base de minimum ou à base de produit (cf. section 7.4) afin d'attribuer un score possibiliste à chaque classe de fiabilité. Enfin, nous confrontons les résultats d'évaluation des Hadiths par rapport aux décisions des savants (cf. section 7.5). Par ailleurs, notre outil d'évaluation des Hadiths est doté d'un affichage graphique permettant de découvrir les sources de suspect ou de (non) fiabilité [Bounhas et al., 2010 ; 2015a].

7.1. La crédibilité des narrateurs

Le tableau 5.10 récapitule la distribution de possibilité des trois classes selon le critère (ℓ) de crédibilité des narrateurs.

c	$\pi(c\|F)$	$\pi(c\|S)$	$\pi(c\|NF)$
[1..4]	0	0	1
[5..9]	1/6	4/6	1/6
[10..12]	3/6	2/6	1/6

Tableau 5.10 : Distribution de possibilité selon la crédibilité des narrateurs

Selon ses distributions des possibilités, la chaîne est considérée non fiable si elle contient une seule personne non crédible (degré entre 1 et 4). En effet, si un narrateur crédible *A* transmet une histoire à un narrateur non crédible *B* alors la narration de ce dernier est considérée inacceptable. Néanmoins, ceci n'influence pas notre confiance en *A*. Ainsi, la crédibilité d'un ou de plusieurs narrateurs d'une chaîne n'implique pas que cette dernière est fiable, car nous devons vérifier les autres critères tels que la continuité de la chaîne et la fiabilité de transmission. Pour la classe suspect (degré ente 5 et 9), nous attribuons une forte possibilité que la chaîne soit suspecte (4/6), mais nous supposons qu'il est possible qu'elle soit fiable ou non à un certain degré (1/6).

7.2. La continuité de la chaîne

Nous calculons la continuité d'une chaîne de narrateurs en utilisant les trois critères suivants : (i) La relation sociale (RS), qui traduit l'existence d'une relation de type "Sheikh-disciple" ou de parenté entre deux narrateurs successifs, (ii) Le gap temporel (GT), et (iii) Le gap géographique (GG). En effet, nous utilisons la base des biographies afin d'identifier les relations sociales entre les narrateurs. Leurs dates de naissance et de décès sont utiles pour calculer les gaps temporels entre eux. Cependant, si la date de décès (respectivement la date de naissance) est inconnue, nous la remplaçons par la date de décès la plus récentes (respectivement la date de naissance la plus ancienne) de la génération du narrateur. Ainsi, l'attribut continuité de la chaîne (*cc*) peut donc prendre l'une de quatre valeurs suivantes :

- *Oui* : il existe un gap temporel et aucune donnée n'est manquante.
- *Oui-manquant* : il existe un gap temporel et certaines données sont manquantes.
- *Non* : il n'existe pas de gap temporel et aucune donnée n'est manquante.
- *Non-manquant* : il n'existe pas de gap temporel et certaines données sont manquantes.

Par ailleurs, le gap géographique entre deux narrateurs est calculé de la façon suivante. Soit v_1 (respectivement v_2) un vecteur composé des informations suivantes sur le premier narrateur (respectivement le deuxième) : le lieu de naissance, le lieu de décès et la valeur du composant *nisba* de son nom. Le gap géographique peut prendre l'une des trois valeurs suivantes :

- *Oui* : les deux vecteurs v_1 et v_2 n'ont aucun élément en commun et ne sont pas nuls.
- *Non* : les deux vecteurs v_1 et v_2 ont au moins un élément en commun et ne sont pas nuls.
- *Inconnu* : l'un des vecteurs v_1 ou v_2 est nul.

Nous utilisons la valeur minimale de tous les liens de la chaîne pour les trois paramètres *RS, GT* et *GG*. En fait, s'il existe une relation sociale entre deux narrateurs, nous considérons que la distribution de possibilité est indépendante des deux autres paramètres (*GT* et *GG*). Au contraire, le gap temporel sera doté de la plus grande importance. De ce fait, nous estimons que deux narrateurs pourraient se rencontrer s'ils vivaient dans la même période et même s'ils n'étaient dans le même endroit. Le tableau 5.11 récapitule nos calculs de la continuité de la chaîne (*cc*), alors que le tableau 5.12 illustre les distributions des possibilités selon ce critère.

RS	**Gap temporel et géographique**			
	GT \ **GG**	Oui	Inconnu	Non
Non	Oui	1	2	3
	Oui-manquant	4	5	6
	Non-manquant	7	8	9
	Non	10	11	12
Oui	13			

Tableau 5.11 : Valeurs du critère de continuité selon la relation sociale, le gap temporel et le gap géographique

Nous remarquons que la valeur *cc* =13, indiquant qu'il existe une relation sociale entre les deux narrateurs, est l'unique valeur du critère de continuité qui permet d'assurer la fiabilité. De plus, nous considérons comme non fiables tous les cas où il existe un gap temporel entre narrateurs. Enfin, la chaîne est considérée suspecte si les narrateurs ont vécu dans la même période mais n'ont pas eu de relation sociale.

cc	π(cc|F)	π(cc|S)	π(cc|NF)
[1..6]	0	0	1
[7..12]	1/6	4/6	1/6
13	3/6	2/6	1/6

Tableau 5.12 : Distribution de possibilité selon le critère de continuité

7.3. La fiabilité de transmission

Nous rappelons que les experts arabes sont ceux qui ont identifié et évalué les différentes manières de transmission des Hadiths. Ces manières sont numérotées de 1 à 8 selon le même ordre du tableau 5.13 [Bounhas, 2012 ; Bounhas et al., 2015a].

Manière de transmission	Verbes
L'audition (السماع)	سمعت فلان (J'ai entendu x) حدثني فلان (x m'a dit)
La lecture au Sheikh (القراءة)	قرأت على فلان (J'ai appris de x) أخبرني فلان (x m'a informé)
La permission (الإجازة)	أجاز لي فلان (x m'a autorisé) أنبأني فلان (x m'a annoncé)
Transmission main en main (المناولة)	ناولني فلان (x m'a donné)
Par écrit (الكتابة)	كتب إلي فلان (x m'a écrit)
Par notification (الإعلام)	أعلمني فلان (x m'a mis au courant)
Par recommandation (التوصية)	أوصى إلي فلان (x m'a recommandé)
Par découverte (الوجادة)	وجدت بخط فلان (J'ai trouvé écrit par x)

Tableau 5.13 : Les manières de transmission du Hadith

Les calculs des distributions des possibilités selon le critère de fiabilité transmission (*FT*) sont récapitulés dans le tableau 5.14 [Bounhas et al., 2010]. En fait, nous utilisons dans ce tableau la manière de transmission la moins fiable de toute la chaîne.

FT	π(FT|F)	π(FT|S)	π(FT|NF)
[6..8]	0	0	1
[4..5]	1/6	4/6	1/6
[1..3]	3/6	2/6	1/6

Tableau 5.14 : Distribution de possibilité selon le critère de fiabilité de transmission

7.4. Identification de la classe de fiabilité

Nous agrégeons les trois critères précédemment calculés (*c*, *cc*, *FT*) afin de calculer un score global pour chaque classe de fiabilité (c_i). Etant donné que nous utilisons des réseaux possibilistes, nous comparons à ce niveau les deux scores possibilistes à base de minimum ($Score_{min}$) et à base de produit ($Score_{prod}$) qui sont donnés respectivement par les formules suivantes :

$$Score_{min}(c_i) = min\{\pi(c|c_i), \pi(cc|c_i), \pi(FT|c_i)\} \quad (5.9)$$

$$Score_{prod}(c_i) = \pi(c|c_i) * \pi(cc|c_i) * \pi(FT|c_i) \quad (5.10)$$

Dans les deux cas, nous choisissons la classe (c^*) titulaire du plus grand score [Bounhas et al., 2010] :

$$c^* = argmax_{c_i}(Score(c_i)) \quad (5.11)$$

7.5. Expérimentation et évaluation

Cette section présente le corpus de test (cf. section 7.5.1) ainsi que les scénarios expérimentaux utilisés dans nos expériences (cf. section 7.5.2).

7.5.1. Le corpus de test

Notre corpus de test est constitué de trois domaines à savoir : "les boissons" (الأشربة) contenant 32320 mots, "le mariage" (الزواج) contenant 53752 mots, et "la purification" (الطهارة) contenant 107058 mots. Ce corpus contient au total 193130 mots. En fait, la généricité et la forte existence de ces trois domaines dans les différents livres de Hadith constituent les deux principaux critères qui justifient notre choix. De plus, la taille de cet échantillon est comparable à certains corpus utilisés dans d'autres travaux dans le domaine. Par exemple, l'analyseur morphologique MADA a été testé dans un corpus composé de 51 K-mots. L'étiqueteur grammatical de [Diab et al., 2004] a été testé uniquement sur 400 phrases. Toutefois, l'évaluation manuelle du résultat d'un analyseur morphologique ou d'un étiqueteur grammatical est une tâche fastidieuse et coûteuse en termes de temps. Cependant, des corpus plus larges peuvent être utilisés dans l'évaluation des approches qui ne nécessitent pas une analyse complète. Par exemple, [Boulaknadel et al., 2008] ont évalué leur travaux dans un corpus qui contient 475148 mots.

7.5.2. Résultats de test

Nous présentons dans cette section les résultats des expérimentations de notre classifieur possibiliste de calcul de la fiabilité. Notre but consiste à confronter les jugements de ce classifieur par rapport aux décisions des savants. Pour cela, nous utilisons les Hadiths des trois domaines du corpus décrit ci-dessus. Les moyennes des scores des trois classes de fiabilité dans les six livres, pour les algorithmes à base de minimum et à base de produit, sont présentées par le tableau 5.15 [Bounhas et al., 2015a]. Pour la classe fiable, nous remarquons que le livre le plus authentique est Sahih Al-Bukhari (صحيح البخاري) titulaire du meilleur score (95.31%), puis nous trouvons le livre Sahih Muslim (صحيح مسلم) titulaire du second score (91.84%). Ainsi, nos résultats sont conformes avec la réalité.

	Fiable		Suspect		Non Fiable	
Livre	**Minimum**	**Produit**	**Minimum**	**Produit**	**Minimum**	**Produit**
Sahih Al-Bukhari (صحيح البخاري)	95.31%	95.90%	69.79%	36.23%	34.90%	3.94%
Sahih Muslim (صحيح مسلم)	91.84%	90.18%	72.11%	45.43%	36.05%	4.28%
Sunan Abi Dawud (سنن أبي داود)	79.37%	80.46%	80.42%	58.67%	40.21%	4.66%
Sunan Ibn Majah (سنن ابن ماجة)	77.78%	79.07%	81.48%	60.91%	40.74%	4.74%
Sunan Annasaii (سنن النسائي)	91.33%	91.75%	72.00%	40.89%	36.00%	4.03%
Sunan Ettermidhi (سنن الترمذي)	82.43%	82.69%	71.17%	45.15%	40.99%	9.44%

Tableau 5.15 : Moyennes des scores attribués pour les trois classes de fiabilité selon les deux algorithmes à base de minimum et à base de produit

Nous analysons ses attributs un par un afin de mieux comprendre plus précisément nos résultats. En fait, et étant donné que les six savants ont toujours utilisé des manières fiables de transmission, le critère de fiabilité de transmission n'a aucun effet dans les Hadiths que nous avons examiné. Les deux autres critères à savoir la crédibilité (ϵ) et la continuité (α) sont analysés dans le tableau 5.16 tout en donnant les valeurs moyennes et minimales dans chaque livre.

Nous remarquons que les deux livres authentiques Sahih Al-Bukhari (صحيح البخاري) et Sahih Muslim (صحيح مسلم) sont titulaires des meilleures scores pour ces deux critères. En effet, la classe minimale de leurs narrateurs est 8, alors que nous trouvons des narrateurs issus des classes 5 et 6 dans les

autres livres. Ceci confirme que les deux savants Al-Bukhari (البخاري) et Muslim (مسلم) sont les plus exigeants quant aux deux critères de crédibilité et de continuité. Ce qui montre encore une fois que nos résultats correspondent à la réalité. Par exemple, le savant Al-Bukhari (البخاري) a une valeur idéale pour le critère de continuité à savoir 13. Ceci nous rappelle que ce savant exige que le disciple doive nécessairement rencontrer son Sheikh pour que ses narrations soient acceptées.

Livre	Moyenne (c)	Minimum (c)	Moyenne (cc)	Minimum (cc)
Sahih Al-Bukhari (صحيح البخاري)	9.70	8	13.00	13
Sahih Muslim (صحيح مسلم)	9.41	8	12.94	8
Sunan Abi Dawud (سنن أبي داود)	8.78	5	12.73	7
Sunan Ibn Majah (سنن ابن ماجة)	8.22	5	12.67	7
Sunan Annasaii (سنن النسائي)	9.40	6	12.73	7
Sunan Ettermidhi (سنن الترمذي)	8.97	5	12.70	7

Tableau 5.16 : Valeurs moyennes et minimales des critères de fiabilité dans les six livres

Par ailleurs, nous comparons les résultats de notre système par rapport aux décisions des savants afin d'avoir une évaluation globale. Nous présentons dans le tableau 5.17 [Bounhas et al., 2015a], pour chaque classe de fiabilité, le pourcentage de Hadiths de la base de test, le pourcentage des Hadiths qui ont été jugés fiables (F), suspects (S) et non fiables (NF). Notons que les classes rares dont nous ne pouvons interpréter les résultats sont affichées dans les dernières lignes de ce tableau avec un fond gris. De plus, la majorité des Hadiths (95.02%) sont réellement fiables avec des degrés différents, ce qui montre l'importance des six livres en tant que source de Hadith. Toutefois, certains narrateurs de ces livres ont un degré de crédibilité entre 5 et 9. En conséquence, notre système attribue la classe "Suspect" à un pourcentage important de Hadiths fiables, malgré que les savants traitent ces narrateurs d'une manière sélective en acceptant certains de leurs Hadiths et en rejettent d'autres. D'autre part, l'existence de ce genre de narrateurs est confirmée si nous nous focalisons sur la classe de Hadith "Faible" dont 10% des Hadiths sont non fiables et 70% ont été classés comme suspects. Ainsi, les narrateurs suspects n'existent pas uniquement dans les Hadiths non fiables, mais aussi dans ceux qui sont fiables, ce qui prouve à la fois l'expertise des savants du Hadith et la difficulté d'automatisation de leur méthodologie.

Selon la moyenne des scores attribués à la classe "fiable", nous trions les classes de fiabilité dans la colonne numéro 1 du tableau 5.17. En fait, nous remarquons que l'ordre obtenu correspond bien à la réalité. Par exemple, et selon la méthodologie des sciences de Hadith, la classe "حسن صحيح" regroupe les Hadiths dont le degré de fiabilité est entre "صحيح" et "حسن". De plus, les savants ont étudié exclusivement les chaînes de narrateurs tout en ignorant les contenus des Hadiths dans les deux classes "صحيح الإسناد" et "حسن الاسناد". C'est pour cette raison qu'ils sont moins fiables que ceux dont nous avons étudié les contenus à savoir les Hadiths des classes "صحيح" et "حسن صحيح". Ce résultat est approuvé par notre système. Notons aussi que nos deux algorithmes possibilistes à base de minimum et à base de produit conduisent au choix de la même classe dans tous les Hadiths que nous avons examinés. De plus, si nous comparons les scores attribués à la même classe pour le même Hadith, nous remarquons que l'algorithme à base de minimum s'avère plus exigeant. En effet, et selon le tableau 5.15, les scores des deux classes "non fiable" et "suspect" diminuent, alors que celui de la classe "fiable" augmente. Ainsi, l'algorithme à base de produit semble le plus réaliste étant donné que la majorité des Hadiths examinés sont fiables. Néanmoins, la généralisation de ces résultats requiert une évaluation sur un échantillon plus grand de Hadiths englobant d'autres livres moins authentiques.

Classe de fiabilité	%	% Fiable	% Suspect	% Non Fiable
صحيح (authentique)	84.33%	78.76%	21.24%	0.00%
حسن صحيح (bon authentique)	1.74%	71.43%	28.57%	0.00%
صحيح الإسناد (chaîne authentique)	3.48%	64.29%	35.71%	0.00%
حسن الاسناد (chaîne bonne)	1.00%	50.00%	50.00%	0.00%
مسكوت عنه (inconnu)	1.49%	33.33%	66.67%	0.00%
ضعيف (Faible)	4.98%	20.00%	70.00%	10.00%
حسن (Bon)	0.75%	66.67%	33.33%	0.00%
صحيح لغيره (Authentique en vertu d'autres Hadiths)	0.75%	33.33%	66.67%	0.00%
صحيح الإسناد مقطوع (la chaîne est authentique mais le contenu est assigné à un disciple)	0.50%	100.00%	0.00%	0.00%
حسن الإسناد مقطوع (la chaîne est bonne mais le contenu est assigné à un disciple)	0.25%	100.00%	0.00%	0.00%
حسن صحيح الإسناد (bon avec chaîne authentique)	0.25%	100.00%	0.00%	0.00%
صحيح الإسناد مدرج (la chaîne est authentique mais certaines expressions ont été ajoutées au contenu)	0.25%	100.00%	0.00%	0.00%

Tableau 5.17 : Comparaison des résultats du système par rapport aux décisions des savants

Par ailleurs, [Ghazizadeh et al., 2008] sont arrivé à identifier correctement la bonne classe dans 94% des cas sans expliquer l'étape d'évaluation, alors que notre taux est de l'ordre de 73.75%. Ce taux semble faible par rapport au premier pour diverses raisons. D'abord, notre algorithme a attribué la classe "suspect" au lieu de la classe "fiable" ou "non fiable" dans 25.25% des cas. Cependant, et grâce à l'affichage graphique, l'utilisateur de notre système peut résoudre le problème et prendre la bonne décision étant donné que la classe "suspect" a été définie afin de mettre l'accent sur les cas douteux. Ensuite, notre algorithme s'est uniquement trompé dans la classification de 1% des cas. Enfin, une comparaison objective avec les travaux de [Ghazizadeh et al., 2008] nécessite l'utilisation de la même collection de test et des mêmes métriques et scénarios d'évaluation.

8. Bilan des contributions et perspectives

Dans ce chapitre, nous avons énuméré les dimensions de la fiabilité de l'information sur la base des travaux de recherche récents dans ce domaine. Comme premier résultat, il a été constaté que la tradition arabe de la narration (ou ce que nous appelons "la méthodologie des sciences de Hadith") répond à ces exigences et donne un modèle de haute-qualité pour l'évaluation des données de fiabilité. Ainsi, nous avons présenté cette méthodologie comme une solution au problème de la fiabilité de l'information. En fait, nous avons évalué la fiabilité de l'information par l'identification exacte des acteurs impliqués dans le processus d'information ainsi que leurs relations. Étudier les biographies de ces acteurs permet aussi d'évaluer la fiabilité de l'information.

Nous avons proposé et évalué une architecture basée sur la théorie des possibilités, qui appuie l'étude de la fiabilité des narrations arabes. En fait, nous avons utilisé un dispositif de reconnaissance des entités nommées pour extraire et analyser les noms propres de personnes de chaînes de narrateurs des narrations arabes. Notre outil possibiliste de reconnaissance de l'identité permet de chercher, pour chaque nom propre dans la chaîne, les personnes candidates dans notre base de données en utilisant des fonctions d'appariement et de filtrage. L'évaluation de la fiabilité a été modélisée comme étant un problème de classification. En effet, nous décidons si la chaîne est "fiable", "suspect" ou "non fiable". Pour résoudre le problème de la classification des chaînes à partir de données imparfaites, nous avons utilisé une approche possibiliste basée sur un classifieur fondé sur les réseaux possibilistes naïfs. Ce classifieur est la contrepartie possibiliste du classifieur Bayésien

naïf [Stvilia et al., 2007]. Ainsi, la deuxième contribution de ce travail est une suite d'outils composées de : (i) un dispositif générique de reconnaissance des entités nommées arabes, (ii) un nouvel algorithme pour la reconnaissance de l'identité arabe (un concept qui n'a pas été étudié auparavant), et (iii) un classifieur possibiliste qui détermine la classe de fiabilité d'une chaîne de narration. Nous avons montré que cette architecture peut être généralisée à d'autres applications et/ou d'autres types de textes.

Même si les outils que nous avons développés ont obtenu de bons taux de réussite, un travail supplémentaire doit être effectué pour améliorer les résultats obtenus. En fait, nous n'avons considéré pour la reconnaissance de l'identité que deux types de relations qui sont les relations père et oncle. Nous prévoyons d'examiner plusieurs types de relations dans cette étape. Nous pouvons aussi améliorer notre analyse des critères de fiabilité. En plus de l'analyse de l'information de la chaîne de narrateurs, nous prévoyons de développer un nouvel outil, qui analyse le contenu de la narration et compare ses versions afin de découvrir d'autres types d'anomalies. Cela aidera à découvrir si un narrateur a ajouté ou supprimé des parties de la narration et si les modifications apportées sont justifiées. En effet, un narrateur peut tout simplement raconter la même narration différemment en choisissant des mots différents qui n'ont pas été utilisés par son prédécesseur, bien que tous les deux soient d'accord sur l'événement et sa façon d'interprétation. Dans d'autres cas, les changements peuvent corrompre la narration si le narrateur est influencé par son interprétation ou son point de vue sur l'événement. Il s'agit de la dimension historique que nous devons prendre en compte. En fait, selon la période et l'évolution de la société, un narrateur ne rapporte pas nécessairement un événement de la même manière qu'un autre ne lui donne la même importance. Mais, la découverte de telles anomalies nécessite des mécanismes d'analyse plus avancés permettant d'évaluer la cohérence logique d'un ensemble de versions du même Hadith. Par ailleurs, il serait important d'identifier, d'une manière plus adéquate, la classe de fiabilité en sélectionnant les classes reconnues dans la méthodologie du Hadith. Toutefois, l'annotation manuelle des distributions de possibilité dans ce cas s'avère une tâche difficile. Ainsi, il faudra profiter des Hadiths déjà évalués pour procéder à une étape d'apprentissage.

Cependant, notre travail dans ce chapitre a été limité à des paramètres calculés à partir des chaînes et des méta-données des narrateurs. Autrement dit, le score obtenu reflète si une chaîne est fiable ou non. En effet, d'autres paramètres liés au contenu de l'histoire et au contexte de la narration doivent être pris en compte afin d'attribuer une note globale de fiabilité pour chaque narration. Néanmoins, nous croyons qu'il est possible d'étudier la fiabilité des autres types de textes arabes avec cette architecture. En outre, les outils développés sont des composants réutilisables qui peuvent être intégrés dans d'autres applications de traitement automatique de la langue (TAL), de la traduction automatique ou de la recherche d'information (RI).

Conclusion et perspectives

Jusqu'à nos jours, les SRI intelligents existants font face à plusieurs défis liés à la désambiguïsation de requêtes, à leurs expansions et à l'évaluation de la fiabilité de l'information recherchée. Nous nous sommes focalisés dans notre projet d'habilitation sur la RI appliquée aux deux langues arabe et française. En effet, nous avons constaté qu'afin d'extraire les connaissances et d'indexer les documents, plusieurs travaux dans la RI arabe ont été effectués à base d'heuristiques ou d'approches statistiques [Boulaknadel, 2006 ; Boulaknadel et al., 2008]. Par ailleurs, Rodriguez et al. (2008) ont proposé des systèmes d'organisation de connaissances arabes, à savoir le WordNet arabe, en profitant des ressources existantes dans d'autres langues. Toutefois, le problème d'ambiguïté des textes arabes a été ignoré dans ces travaux à cause de l'absence d'une analyse morphosyntaxique complète. En fait, la solution consiste à utiliser l'analyse superficielle exigeant moins de ressources que l'analyse complète. De plus, les systèmes d'extraction de connaissances et les SRI arabes requièrent des informations pertinentes issues d'une étape d'analyse de textes arabes. Cette dernière nécessite des corpus d'apprentissage et des outils sophistiqués capables de traiter ces textes quelle que soit leur période. Cependant, les systèmes existants souffrent encore du problème d'ambiguïté des mots et des expressions résultant de leurs exploitations des informations insuffisantes.

Les travaux de recherche relatifs aux systèmes d'extraction de connaissances et les SRI arabes ont donné naissance à plusieurs contributions dispersées sur divers niveaux d'analyse. En effet, plusieurs outils d'analyse et de désambiguïsation morphologiques des textes arabes ne sont pas suffisamment exploités dans les systèmes d'extraction de connaissances et de RI. De plus, les SRI arabes actuels considèrent la notion de pertinence comme un concept monodimensionnel. D'autre part, et malgré la naissance du Web socio-sémantique qui tient compte de double besoins sociaux et sémantiques, la plupart des travaux de recherche se sont intéressées uniquement à l'axe sémantique. En outre, le corpus de Hadith a été utilisé dans l'évaluation de plusieurs SRI sémantiques, mais ces derniers ont ignoré la dimension de fiabilité, bien qu'il existe des techniques d'évaluation automatique de la fiabilité des Hadiths.

Par ailleurs, les nouvelles orientations des SRI récents convergent vers des techniques plus avancées qui anticipent la démarche classique de "requête - liste de résultats". Désormais, la nouvelle RI est considérée comme un scénario d'enquête qui implique plusieurs critères. De plus, la nécessité d'une vue globalisante et détaillée de l'espace informationnel s'impose aujourd'hui, elle pourrait être assurée grâce à la structuration et la présentation de cet espace d'une façon plus appréhensible. En conséquence, l'accès aux ressources documentaires devient de plus en plus personnalisé grâce à l'exploitation d'une variété de mécanismes de visualisation et d'interaction. En fait, nous avons essayé d'impliquer l'utilisateur dans les différentes étapes du processus de RI, tout en intégrant les différentes tâches d'extraction, de représentation et d'accès à la connaissance dans un seul processus de cartographie socio-sémantique [Bounhas, 2012].

En effet, nous avons proposé un processus de RI qui tient compte de la richesse et des spécificités des deux langues et civilisations arabes et françaises. Cette nouvelle orientation répond davantage aux nouveaux besoins de la RI moderne. Notre objectif consiste à préparer le terrain pour l'intégration de ces deux langues (particulièrement l'arabe) dans les systèmes d'ingénierie des connaissances. Pour cela, nous avons proposé des approches d'analyse, de désambiguïsation et d'expansion des textes français et arabes. En outre, il s'est avéré nécessaire de considérer à la fois les deux aspects sémantiques et sociaux dans les divers axes de l'analyse. A ce niveau, nous avons

renforcé la nécessité de l'évaluation de la fiabilité comme critère principal de la pertinence de l'information. En fait, nous avons tenu compte de ce critère en exploitant la méthodologie des sciences du Hadith pour l'évaluation de la fiabilité. Par ailleurs, cette méthodologie tient compte des critères reconnus dans la littérature liée à la qualité de l'information [Naumann et Rolker, 2000]. De plus, les nouvelles visions du Web socio-sémantique s'articulent sur le concept de confiance, alors que cette méthodologie est parfaitement cohérente avec ces visions.

1. Choix principaux

Nous avons utilisé dans nos évaluations, d'une part, le standard ROMANSEVAL pour la désambiguïsation de requêtes françaises, et d'autre part, la collection CLEF-2003, particulièrement "LeMonde94", pour leurs expansions. Ces deux collections s'avèrent nécessaires afin de constituer ensemble un standard pour l'évaluation de l'impact de la désambiguïsation de requêtes sur leurs expansions dans les SRI intelligents. En outre, nous avons utilisé deux types de textes modernes et classiques à savoir le corpus Treebank arabe (ATB part 2 v2.0) et le corpus de Hadith afin d'évaluer nos approches de classification possibiliste pour la désambiguïsation des textes arabes. En fait, nous avons confirmé l'indépendance du domaine de nos modèles possibilistes en menant nos expérimentations sur le corpus Treebank arabe rassemblant les textes de journaux.

Par ailleurs, les caractéristiques des livres du Hadith justifient leurs choix comme cas d'application dans l'évaluation de la fiabilité de l'information en langue arabe. En effet, le processus d'évaluation de la fiabilité est composé de plusieurs étapes adéquates à la structure des livres du Hadith. Cette structure est primordiale dans l'étape de production des documents dans les livres du Hadith. Ces derniers documentent toutes les transactions sémiotiques de transfert et d'interprétation des informations. De plus, nous avons modélisé les connaissances, grâce à cette structure, afin d'accomplir une recherche précise et personnalisée de l'information. La taille des livres de Hadith, leur richesse et leur organisation par thème permettent, d'une part, de développer et d'évaluer des méthodes d'extraction de connaissances et de RI multicritères, et d'autre part, de tenir compte des pratiques des utilisateurs dans leurs accès multipoints de vue. En fait, la sévère méthodologie des sciences de Hadith pour l'évaluation de la fiabilité de l'information a donné naissance à ce fond documentaire riche en thèmes et en connaissances socio-sémantiques. Ainsi, et grâce à ces nombreux avantages, le corpus de Hadith a été ciblé par plusieurs travaux de recherche en informatique tels que [Al-Muhtaseb et al., 2009 ; Harrag et al., 2009 ; Alkhatib, 2010 ; Yuso et al., 2010 ; Bounhas et al., 2011ab ; 2015ab ; Bounhas, 2012 ; Ayed et al., 2012ab ; Ayed, 2017].

En outre, nous avons gardé la même organisation des livres de Hadith en termes des thèmes tels qu'ils sont présentés dans les livres des savants arabes collecteurs de Hadiths. En fait, nous avons proposé dans [Bounhas et al., 2010 ; 2011ab ; 2015ab] des outils d'organisation et d'évaluation automatique des connaissances dans le but d'améliorer les mécanismes d'accès aux Hadiths. Ces outils sont exploités parallèlement à une recherche arborescente dans les cartes de thèmes [Bounhas, 2012]. En réalité, nous avons étendu nos travaux dans [Elayeb et al., 2009], modélisant le processus de RI d'une manière novatrice, en mixant deux types de réseaux à savoir les Réseaux Petits-Mondes Hiérarchiques (RPMH) et les réseaux possibilistes (RP). En effet, et dans le but d'avoir une vue globalisante des connaissances ainsi que d'éliciter les liens implicites, nous avons exploité les RPMH comme outil d'organisation des connaissances. En conséquence, nous avons réussi à représenter n'importe quelle dimension de notre espace informationnel grâce à la flexibilité et au caractère générique des RPMH. Ensuite, nous avons profité des réseaux possibilistes afin de lier les différentes dimensions d'un tel espace. Enfin, plusieurs travaux récents dans le domaine de la classification possibiliste [Haouari et al., 2009 ; Bounhas et al., 2013 ; 2014] et de RI possibiliste [Boughanem et

al., 2009, Elayeb et al., 2011, 2015a] ont prouvé l'efficacité de la théorie des possibilités comme outil de modélisation. En fait, cette théorie s'intéresse à la fois aux deux cadres quantitatif ou qualitatif. De plus, elle prend en considération les phénomènes d'imperfection dans les données telles que l'incomplétude, l'imprécision et l'incertitude.

2. Bilan et apports

Nos récents travaux sur la recherche d'information s'articulent autour des **deux axes de contributions majeurs** qui touchent d'une part la requête de l'utilisateur, et d'autre part, l'ensemble de documents pertinents retournés par le SRI.

Pour les travaux du **premier axe**, nous nous sommes focalisés sur la **désambiguïsation et l'expansion de requêtes**. Cet axe pourra être partagé en quatre sous-contributions majeures : (i) la désambiguïsation sémantique de textes français ; (ii) la désambiguïsation morphologique de textes arabes. (iii) l'expansion sémantique de textes français ; et (iv) l'évaluation de l'impact de la désambiguïsation sémantique de requêtes sur leurs expansions. Nous résumons dans la suite ces quatre contributions :

D'abord, nous avons exploité et comparé deux approches de désambiguïsation sémantique monolingue : possibiliste et probabiliste. En effet, ces deux approches ont profité à la fois d'un dictionnaire traditionnel ainsi que d'un corpus étiqueté afin de donner naissance à une nouvelle ressource linguistique externe que nous appelons "dictionnaire sémantique des contextes (DSC)". Dans l'approche possibiliste, nous avons utilisé un réseau possibiliste afin de quantifier la pertinence d'un sens de mot ambigu étant donné une phrase polysémique. Cette pertinence est modélisée par une double mesure, à savoir la pertinence possible dont le but est d'éliminer les sens non-pertinents d'un mot ambigu, et la pertinence nécessaire renforçant la pertinence des sens restants non-rejetés par la possibilité. Dans l'approche probabiliste, nous avons utilisé et étendu une distance probabiliste existante afin de calculer un score sémantique, entre les mots du dictionnaire, en prenant en compte la topologie complète de ce dernier vu comme un graphe sémantique sur ses entrées.

Ces deux approches sont évaluées et comparées via la collection de test ROMANSEVAL et selon une double analyse détaillée et globale. Dans nos premières expériences, nous avons utilisé les deux métriques d'évaluation *Accord* et *Kappa*. En effet l'*Accord* a été exploité lors de la comparaison des deux méthodes d'apprentissage du DSC à base de jugements et à base d'un dictionnaire. Par contre, *Kappa* a été utilisé dans le cadre d'un bilan comparatif avec les systèmes concurrents de désambiguïsation monolingue, afin de prouver nos contributions encourageantes en termes de taux de précision de désambiguïsation de mots français. En effet, notre désambiguïsation possibiliste a dépassé la performance du système Xerox pour les adjectifs, les noms, les verbes ainsi que toutes les catégories grammaticales. Alors que notre désambiguïsation probabiliste a réussi à désambiguïser les noms mieux que les systèmes possibiliste et Xerox. Globalement, et si nous considérons toutes les catégories grammaticales (All POS), l'approche possibiliste est titulaire du meilleur taux de précision en la comparant avec ses concurrents. Dans nos deuxièmes expériences, nous avons utilisé les métriques d'évaluation rappel, précision et F-mesure afin de diversifier et élargir nos scénarios d'expérimentation et de prouver, en conséquence, les performances de nos approches indépendamment de la métrique utilisée. En effet, notre approche possibiliste reste globalement meilleure que celle probabiliste et que Xerox dans la désambiguïsation des adjectifs, des noms, des verbes ainsi que toutes les catégories grammaticales. En outre, le système possibiliste est aussi globalement meilleur que tous les autres systèmes de désambiguïsation monolingue français existants dans [Segond, 2000]. Ces résultats prouvent la contribution de la théorie des possibilités comme un

moyen de traiter l'imprécision dans les systèmes d'information, particulièrement les systèmes de désambiguïsation sémantique de requêtes.

Puis, nous avons considéré la tâche de désambiguïsation des attributs morphologiques des textes arabes non-voyellés comme une tâche de classification. Pour cela, nous avons proposé des classifieurs possibilistes assurant l'apprentissage et le test à partir des données imprécises. D'abord, nous avons testé trois classifieurs possibilistes respectivement à base des mesures de possibilité, de nécessité et de leur somme. Ces trois classifieurs ont été comparé aux classifieurs non-possibilistes existants en termes de taux de désambiguïsation de 14 attributs morphologiques. Ensuite, ces trois classifieurs ont été étendus à trois autres impliquant un modèle de repondération et une possibilité lexicale. Nos résultats, effectués sur deux types différents de corpus classiques et modernes à savoir le corpus du Hadith et le Treebank arabe, ont prouvé l'efficacité de notre classifieur possibiliste discriminatif à base de la somme de possibilité et de nécessité et impliquant un modèle de repondération et une possibilité lexicale, par rapport à ses concurrents.

Ensuite, nous avons proposé, évalué et comparé des approches d'expansion sémantique de requêtes utilisant le dictionnaire français "Le Grand Robert" comme ressource linguistique externe. La première approche à base de circuits a exploité le graphe du Réseau Petits-Mondes Hiérarchiques (RPMH) afin de modéliser la structure de ce dictionnaire. En fait, nous avons profité d'un nouveau score de proximité sémantique entre les termes nœuds du graphe RPMH en fonction de circuits énumérés entre eux. En réalité, nous avons étendu nos travaux dans [Elayeb et al., 2009], qui ont été limités uniquement au graphe RPMH des verbes français, pour exploiter un graphe englobant toutes les catégories grammaticales à savoir les verbes, les adverbes, les noms et les adjectifs. La deuxième approche a profité des réseaux possibilistes (RP) pour la modélisation de la structure du même dictionnaire. En effet, cette approche tient compte d'une double mesure de pertinence afin de définir un nouveau score possibiliste de proximité sémantique entre les articles d'un dictionnaire et les termes de la requête à reformuler. Autrement dit, le processus d'expansion consiste à chercher les articles du dictionnaire qui sont possiblement et nécessairement proches des termes de la requête originelle à reformuler. Si la pertinence possible permet de rejeter les articles non-pertinents (non-sémantiquement proches), la pertinence nécessaire permet de renforcer les scores des articles restants non-éliminés par la possibilité. Cette technique possibiliste semble plus fine que celle à base de circuits dans la recherche de nouveaux termes d'expansion sémantique de requêtes. En outre, nous avons proposé deux nouvelles approches combinant les scores, des proximités sémantiques, issus des deux premières techniques afin de donner naissance à deux scores agrégés dont un à base de *somme* et l'autre à base de *produit*. En fait, la performance de ces nouvelles approches hybrides est prouvée par le nombre de requêtes améliorées.

Enfin, nous avons évalué l'impact de la désambiguïsation sémantique possibiliste de requêtes (DSR) sur leurs expansions (ESR). D'abord, nous avons généré un graphe de cooccurrence à partir de la collection de documents "LeMonde94" faisant partie du standard CLEF-2003. Ensuite, ce graphe a été exploité afin de sélectionner les sens/termes utiles à la fois pour les deux processus de DSR et d'ESR. En effet, nous nous sommes focalisés dans nos tests sur les requêtes du standard CLEF-2003 contenant des termes ambigus existant dans le standard ROMANSEVAL. Nos résultats ont confirmé l'importance de l'étape de la désambiguïsation sémantique possibiliste des termes de la requête originelle avant son expansion possibiliste. Enfin, nous avons profité de la technique de pseudo-réinjection de pertinence afin d'améliorer davantage la performance de notre approche combinant la DSR et l'ESR. Cette approche possibiliste a prouvé aussi son efficacité devant une approche à base de dénombrement de circuits. Néanmoins, l'injection d'un grand nombre de termes

d'expansion détériore l'efficacité du processus d'ESR. Cette détérioration est expliquée par l'effet du bruit résultant des connaissances issues du graphe de cooccurrence.

Nous nous sommes intéressés dans le **deuxième axe majeur** à la détermination de doubles besoins sociaux et sémantiques des utilisateurs afin d'**identifier et d'évaluer la dimension fiabilité**. Nous partons de l'idée que l'organisation sociale des utilisateurs et leurs besoins ont beaucoup d'impact sur les pratiques des utilisateurs et les mécanismes que le système doit fournir. En conséquence, nous avons commencé notre processus d'analyse par une étude sociale. Cette étude identifie d'une part, les outils d'analyse nécessaires, et d'autre part, le niveau de granularité lors de la segmentation des documents. De plus, nous avons exploité une analyse micro-logique à base des grammaires hors contexte [Bounhas et Slimani, 2009b] dans l'objectif d'appuyer la réutilisation de nos outils d'analyse. En fait, le traitement de chaque type de fragment de document à part, ainsi que la simplification de l'apprentissage semi-automatique des règles de ces grammaires permettent ensemble de réduire la complexité des textes arabes.

Par ailleurs, nous avons proposé une approche de désambiguïsation morphosyntaxique arabe basée sur la structure des documents du Hadith [Bounhas et al., 2011b]. En effet, le contexte sémantique nécessaire pour la désambiguïsation est constitué de titres des thèmes de Hadiths. D'une part, nous avons réussi à réaliser une évaluation qualitative des pertinences des termes au domaine grâce à leur pondération en fonction de leurs positions dans la structure du document. D'autre part, nous avons réussi aussi à réunir les deux tâches de l'évaluation de la pertinence au domaine et de la désambiguïsation en une seule étape. Ainsi, nos contributions s'articulent sur trois niveaux : (i) nous avons prouvé l'interdépendance des différents niveaux d'analyse, (ii) le mixage des tâches nous a permis d'accélérer le processus d'analyse, et (iii) nous avons confirmé l'impact positif de l'analyse de la structure du document sur le processus de désambiguïsation morphosyntaxique. En outre, dans [Bounhas et al., 2011a] nous avons regroupé les termes arabes d'une manière cohérente grâce à une analyse distributionnelle basée sur l'exploitation du réseau de dépendances syntaxiques. Suite à cette analyse, nous avons interprété séparément les relations syntaxiques arabes possédant des diverses sémantiques.

Le processus d'identification et d'évaluation de la fiabilité a été démarré par le lancement des étapes préliminaires à savoir la reconnaissance des entités nommées et des identités des personnes. Ainsi, notre SRI intelligent implique l'axe social ignoré par la majorité des SRI monocritères. En fait, l'extraction de la structure de chaque entité nommée au format XML ainsi que la représentation explicite de leurs relations sociales sont les fruits de l'exploitation des grammaires hors contexte. De plus, le problème d'ambiguïté des noms arabes a été résolu grâce au réseau social exploité par notre outil de reconnaissance de l'identité. Cet outil, vu comme un SRI social assurant le calcul automatique de la classe de fiabilité est doté d'une interface graphique renforçant l'analyse de la fiabilité.

Du point de vue environnement, nous avons implémenté en Java une boite à outils générique qui traite la structure, la morphologie, la syntaxe et les entités nommées dans les documents arabes. Ces outils, réutilisables, pourront soutenir les plates-formes d'ingénierie de connaissances, d'ingénierie ontologique, de recherche d'information, de traitement automatique de langue et de traduction automatique. En outre, les opérations de base sur les graphes RPMH et RP tels que le filtrage, la transformation et le clustering facilitent d'une part, à l'utilisateur de mieux comprendre et manipuler son espace informationnel, et d'autre part, elles permettent au système d'automatiser certaines étapes de RI, telles que la désambiguïsation et l'expansion de requêtes [Elayeb et al., 2011 ; 2015a]. L'interactivité avec l'utilisateur permet d'identifier ses contraintes utiles à l'évaluation de l'information. Enfin, nous avons proposé un modèle d'appariement possibiliste multicritère

complétant nos objectifs de proposition d'une plate-forme de RI socio-sémantique intelligente et multicritère [Bounhas, 2012 ; Ayed, 2017].

Toutefois, nos travaux en langue arabe souffrent du manque de standards d'évaluation. En conséquence, nous étions obligés de construire manuellement des listes de référence particulièrement au niveau sémantique. Les limites de ces listes nous ont poussé à renforcer l'évaluation de nos approches via une double validation à savoir automatique (assurée par le système) et manuelle (assurée par l'expert) [Bounhas et al., 2011ab].

3. Perspectives de recherche

Les travaux actuels du domaine de la désambiguïsation sémantique s'orientent vers le cadre translinguistique depuis les compétitions de SemEval-2010 et 2013. En effet, les raisons de cette migration d'un contexte mono-linguistique vers un cadre translinguistique sont argumentées dans [Lefever et Hoste, 2013]. Ainsi, nous programmons d'intégrer nos deux approches des désambiguïsations sémantiques en français et en arabe dans un SRI translinguistique afin d'appuyer le processus de désambiguïsation de requêtes [Ben Romdhane et al., 2017 ; Elayeb et al., 2018 ; Ben Khiroun et al., 2018]. D'autre part, et étant donné la généricité et l'indépendance de la langue, notre approche possibiliste de désambiguïsation du français pourra être étendue et adaptée aux autres langues telles que l'anglais et l'arabe. Pour l'anglais, nous pouvons profiter des collections de test disponibles dans [Koeling et al., 2005]. Cependant, cette extension semble plus délicate pour le cas de la désambiguïsation arabe là où nous avons besoin, non seulement d'un standard de test pertinent pour la langue arabe, mais aussi d'une structuration des dictionnaires arabes bruts. En fait, certains travaux [Zouaghi et al., 2012 ; Khemakhem et al., 2013] ont proposé des solutions encourageantes pour structurer les dictionnaires arabes utiles à la tâche de désambiguïsation arabe. En outre, nous espérons pouvoir exploiter nos outils et structures de données réutilisables dans d'autres domaines tels que l'extraction d'information, la classification des textes [Chouigui et al., 2017], la traduction automatique [Ben Romdhane et al., 2017 ; Ben Khiroun et al., 2018 ; Ben Khiroun, 2018 ; Elayeb et al., 2018 ; Elayeb, 2018], l'analyse du contenu, le traitement et l'organisation de la terminologie, la lexicographie et les applications du Web sémantique.

Par ailleurs, nous avons mentionné que nos approches de classification possibiliste pour la désambiguïsation des textes arabes n'ont pas réussi à désambiguïser intégralement la totalité des attributs morphologiques arabes. Cette incapacité est causée, d'une part, de l'ordre relativement aléatoire des mots dans la phrase, et d'autre part, des particules qui ont un taux d'ambiguïté élevé, même dans les textes voyellés. Afin de résoudre ce problème, nous proposons soit d'élargir l'ensemble d'apprentissage, soit de procéder à une analyse linguistique manuelle dans l'étape d'apprentissage permettant de se débarrasser des mots vides et de minimiser, en conséquence, l'échec de leurs désambiguïsations. Au même temps, nous souhaitons limiter l'intervention de l'utilisateur afin d'éviter le caractère manuel du traitement de l'ensemble d'apprentissage. Nous espérons aussi exploiter notre approche dans des phases de désambiguïsation des requêtes et des documents arabes, et en conséquence améliorer les performances globales des SRI qui traitent des textes voyellés et non-voyellés. Enfin, les attributs morphologiques calculés par nos outils peuvent être exploités dans d'autres niveaux d'analyses syntaxique et sémantique.

En outre, nous avons étudié l'impact de la désambiguïsation possibiliste de requêtes sur leurs expansions dans les SRI intelligents. D'abord, nous visons à comparer notre approche, combinant la désambiguïsation et l'expansion de requêtes à base d'un graphe de cooccurrence, à d'autres approches utilisant différentes ressources linguistiques externes telles que les dictionnaires, les

thésaurus [Ben Khiroun et al., 2018] et les ontologies. Ensuite, nous souhaitons profiter des standards de test des exercices SensEval/SemEval afin d'étendre nos approches vers un cadre translinguistique. Enfin, la généricité de nos composants de traitement des requêtes fondés sur les graphes ainsi que leur indépendance de la langue, nous encourage à étendre et adapter nos approches à d'autres langues à savoir l'anglais et l'arabe.

Dans notre approche possibiliste d'identification et d'évaluation de la fiabilité, les deux phases d'apprentissage et d'évaluation sont coûteuses en termes de temps et d'efforts à cause de la haute intervention de l'utilisateur lors de l'implémentation des outils d'analyse de textes arabes. En outre, notre approche souffre du problème de l'ambiguïté morphologique causée par l'absence des voyelles courtes dans les textes. Afin de résoudre ce problème, nous envisageons d'exploiter des textes partiellement ou complètement voyellés. Ainsi, nous pourrons profiter de certains livres voyellés dans le corpus du Hadith. En fait, si les principales entités logiques dans le document tels que les titres et les sous-titres sont voyellés, notre étape d'analyse linguistique s'améliore davantage.

Par ailleurs, la nature du corpus utilisé ainsi que sa structure sont deux facteurs pertinents dans l'évaluation de notre approche de désambiguïsation. Afin d'expliquer l'impact de la structure sur les performances de nos outils et de généraliser en conséquence nos résultats, il sera utile d'appliquer nos approches sur un ensemble de documents semi-structurés du Web. Dans ce cas, nous serons obligés de considérer une description plus détaillée de la structure, étant donné que les pages Web ne sont pas forcément hiérarchiques, contrairement à la structure arborescente des livres de Hadith. Notre solution préliminaire a été d'attribuer aux fragments particuliers des poids traduisant leur pertinence dans le document. Cependant, nous estimons que la structure des documents sera plus détaillée si nous exploitons une technique d'annotation automatique. En plus de la taille, le style et l'organisation spatiale, nous souhaitons exploiter des marqueurs rhétoriques qui pourront mieux définir les fragments. A ce niveau, il serait intéressant d'incorporer notre analyseur micro-logique dans notre outil d'analyse morphosyntaxique arabe. En fait, nous avons ignoré l'étape d'analyse morphologique dans le traitement lexical des entités nommées afin de se débarrasser des ambigüités. Néanmoins, nous envisageons d'extraire à la fois les entités nommées et les syntagmes nominaux arabes grâce à un mixage de deux types d'analyse en un seul outil.

De plus, nos tests doivent cibler non seulement tous les thèmes des livres du Hadith, mais aussi d'autres types de textes afin d'élargir nos expérimentations. En effet, l'exploitation des réseaux syntaxiques nous permet de découvrir d'autres types de relations sémantiques entre les termes ou les groupes de termes. Ces relations pourront être dépendantes ou non du domaine. En conséquence, les Hadiths seront mieux représentés grâce à cette analyse sémantique, ce qui nous encourage à développer des outils de raisonnement plus efficaces, particulièrement pour examiner d'autres critères dans l'évaluation de la fiabilité. Par exemple, les anomalies et l'excentricité seront identifiées suite à une confrontation des différentes versions du même Hadith.

Du côté de l'environnement de l'expérimentation, les mécanismes de visualisation et d'interaction dans notre plate-forme de fiabilité nécessitent des améliorations. En fait, le processus de recherche dans la plate-forme actuelle ne retourne que des Hadiths entiers, alors que l'utilisateur a besoin parfois des fragments des documents tels que les commentaires associés aux Hadiths ou les sous-chapitres. De plus, il sera intéressant de développer différentes stratégies d'adaptation, qui donnent à l'utilisateur la possibilité de constituer ses propres documents [Falquet et al., 2004] ou qui le guident lors de sa navigation [Iksal et Garlatti, 2002]. Mais, la considération des profils des utilisateurs est nécessaire dans toute recherche personnalisée. A ce stade, il sera possible de tenir compte de deux aspects indispensables dans le profil. Le premier tient compte de l'organisation sociale des utilisateurs en impliquant par exemple les orientations qui différencient le profil d'une communauté

de celui de ses membres [Ding et al., 2005]. Le deuxième tient compte de l'expertise de l'utilisateur lors de l'affichage de son résultat de recherche. Par exemple, les utilisateurs débutants ne s'intéressent ni aux longues chaînes de narrateurs ni à certains commentaires.

Récemment, nous avons proposé dans [Ben Khiroun et al., 2014b ; Ayed, 2017 ; Ayed et al., 2018] le prototype d'une plate-forme visant à transformer les livres de Hadiths en un standard de test utile pour la recherche d'information mono-, multi- et translinguistique. En fait, les textes de Hadiths ont été exploités et interprétés dans plusieurs régions au fil des siècles. Ainsi, ce corpus s'avère pertinent pour étudier l'évolution géographique et historique de la langue arabe. Nous avons commencé par collecter les diverses versions des livres de Hadiths sous différents formats. Notons que ces versions sont presque équivalentes du point de vue quantitatif et qualitatif. En outre, elles sont hétérogènes du point de vue crédibilité de leurs sources, couverture, taille, et même en terme de richesse en commentaires. Ainsi, nous avons profité des atouts de chacune des versions en exploitant et en combinant toutes les versions fiables disponibles. Ensuite, nous avons défini un ensemble de requêtes types ainsi que leurs documents pertinents inhérents en testant plusieurs modèles d'appariement. A l'heure actuelle, notre plate-forme traite cette dernière étape d'une manière collaborative et semi-automatique.

D'autre part, il existe aujourd'hui plusieurs autres domaines d'application qui souffrent du problème de la fiabilité de l'information. Ces domaines peuvent profiter de la méthodologie des sciences de Hadiths. Par exemple, le problème de confiance dans le Web socio-sémantique pourra profiter de cette méthodologie. En outre, certains travaux ont exploité cette méthodologie pour lutter contre les crimes électroniques [Yuso et al., 2010]. Il est aussi envisageable de réutiliser nos outils pour l'analyse et l'évaluation de la fiabilité dans les articles de journaux contenant des textes similaires aux chaînes de narrateurs des Hadiths.

Bibliographie

Agirre E. et Edmonds P. (2006). Word Sense Disambiguation. Algorithms and Applications (Text, Speech and Language Technology). Springer, Dordrecht, 364 p.

Agirre E. et Martinez D. (2000). Exploring automatic word sense disambiguation with decision lists and the Web. In: Proceedings of the Workshop COLING 2000, Luxembourg, pp. 11-19.

Agirre E., Arregi X. et Otegi A. (2010). Document expansion based on WordNet for robust IR. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 9-17.

Agirre E., Martínez D., de Lacalle O.L. et Soroa, A. (2006). Two Graph-based Algorithms for State-of-the-art WSD. In Proc. EMNLP, pp. 585–593.

Agirre, E., López de Lacalle, O. et Soroa, A. (2014). Random Walks for Knowledge-based Word Sense Disambiguation. Comput. Linguist., 40(1), pp. 57-84.

Al-Ansary S. (2005). Building a Computational Lexicon for Arabic : A corpus-based approach. In Alhawary, M. T. et Benmamoun, E., éditeurs : Current Issues in Linguistic Theory, volume 267, pp. 173–193.

Alkhatib M. (2010). Classification of al-Hadith al-shareef using data mining algorithm. In European Mediterranean & Middle Eastern Conference on Information Systems, Abu-Dhabi, UAE.

Alkuhlani S., Habash N. et Roth R., (2013). Automatic morphological enrichment of a morphologically underspecified Treebank. In: Proc. NAACL HLT, pp. 460–470.

Almasri M., Chevallet J-P. et Berrut, C. (2014). Exploiting Wikipedia Structure for Short Query Expansion in Cultural Heritage. In Proc. CORIA, pp. 287–302.

Al-Muhtaseb H.A., Mahmoud S.A. et Qahwahi R.S. (2009). A novel minimal script for arabic text recognition databases and benchmarks. In: International Journal of Circuits, Systems and Signal Processing, 3(3), pp. 145-153.

Audeh B. (2014). Reformulation sémantique des requêtes pour la recherche d'information ad hoc sur le Web. Thèse de doctorat, Ecole Nationale Supérieure des Mines de Saint-Etienne.

Audibert L. (2002). Etude des critères de désambiguïsation sémantique automatique : Présentation et premiers résultats sur les cooccurrences. Dans : Actes de TALN-RECITAL, pp. 415–424.

Audibert L. (2003). Outils d'exploration de corpus et désambiguïsation lexicale automatique. Thèse de Doctorat, Université d'Aix-Marseille I – Université de Provence, France.

Ayed R. (2017). Désambiguïsation Morphologique de Textes Arabes à Base de Classification Possibiliste pour la Recherche d'Information Socio-Sémantique. Thèse de Doctorat en Informatique, Ecole Nationale des Sciences de l'Informatique, Université de la Manouba, Tunisie, Décembre 2017.

Ayed R., Bounhas I., Elayeb B., Bellamine Ben Saoud N. et Evrard F. (2014a). Evaluation d'une approche possibiliste pour la désambiguïsation des textes arabes. Dans : Actes de TALN, pp. 316-327.

Ayed R., Bounhas I., Elayeb B., Bellamine Ben Saoud N. et Evrard F. (2014b). Improving Arabic Texts Morphological Disambiguation using Possibilistic Classifier. In: Proc. NLDB, pp. 138-147.

Ayed R., Bounhas I., Elayeb B., Evrard F. et Bellamine Ben Saoud N. (2012a). Arabic Morphological Analysis and Disambiguation Using a Possibilistic Classifier. In: Proc. ICIC, pp. 274–279.

Ayed R., Bounhas I., Elayeb B., Evrard F. et Bellamine Ben Saoud N. (2012b). A Possibilistic Approach for the Automatic Morphological Disambiguation of Arabic Texts. In: Proc. SNPD, pp. 187-194.

Ayed R., Elayeb B. et Bellamine Ben Saoud N. (2018). Possibilistic Morphological Disambiguation of Structured Hadiths Arabic Texts Using Semantic Knowledge. In: Proc. ICAART, pp. 565-572.

Ayed R., Chouigui A. et Elayeb B. (2018). A New Morphological Annotation Tool for Arabic Texts. In: Proc. AICCSA, Aqaba, Jordan, Oct 28 – Nov 01, 2018.

Banerjee S. et Pedersen T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In: Proc. CICLing, pp. 136–145.

Barathi M. et Valli S. (2010). Ontology Based Query Expansion Using Word Sense Disambiguation. In: International Journal of Computer Science and Information Security, 7(2), pp. 22-27.

Barque L. et Chaumartin F.R. (2008). La polysémie régulière dans WordNet. Dans : Actes de TALN, Avignon, France.

Ben Khiroun O. (2018). Recherche d'Information Monolingue & Translinguistique : de la Désambiguïsation vers l'Expansion Sémantique de Requêtes. Thèse de Doctorat en Informatique, Ecole Nationale des Sciences de l'Informatique, Université de la Manouba, Tunisie, Mars 2018.

Ben Khiroun O., Ayed R., Elayeb B., Bounhas I., Bellamine Ben Saoud N. et Evrard F. (2014b). Towards a New Standard Arabic Test Collection for Mono- and Cross-Language Information Retrieval. In: Proc. NLDB, pp. 168-171.

Ben Khiroun O., Elayeb B. et Bellamine Ben Saoud N. (2018). Towards a Query Translation Disambiguation Approach using Possibility Theory. In: Proc. ICAART, pp. 606-613.

Ben Khiroun O., Elayeb B., Bounhas I., Evrard F. et Bellamine Ben Saoud N. (2014a). Improving query expansion by automatic query disambiguation in intelligent information retrieval. In: Proc. ICAART, pp. 153-160.

Ben Khiroun O., Elayeb B., Bounhas I., Evrard F. et Bellamine Ben Saoud N. (2012). A possibilistic approach for automatic word sense disambiguation. In: Proc. ROCLING, pp. 261-275.

Ben Khiroun O., Elayeb B., Bounhas I., Evrard F. et Bellamine Ben Saoud N. (2011). A possibilistic approach for semantic query expansion. In: Proc. ITA'11, pp. 308-316.

Ben Romdhane W., Elayeb B. et Bellamine Ben Saoud N. (2017). A Discriminative Possibilistic Approach for Query Translation Disambiguation. In Proc. NLDB, pp. 366-379.

Ben Romdhane W., Elayeb B., Bounhas I., Evrard F. et Bellamine Ben Saoud N. (2013). A Possibilistic Query Translation Approach for Cross-Language Information Retrieval. In: Proc. ICIC, pp. 73-82.

Benferhat S., Dubois D., Garcia L. et Prade H. (1999). Possibilistic logic bases and possibilistic graphs. In: Proc. UAI, pp. 57–64.

Benferhat S., Dubois D., Garcia L. et Prade H. (2002). On the transformation between possibilistic logic bases and possibilistic causal networks. In: International Journal of Approximate Reasoning, 29(2), pp. 135-173.

Bhogal, J., Macfarlane, A. et Smith, P. (2007). A Review of Ontology Based Query Expansion. Information Processing and Management, 43(4), pp. 866-886.

Billerbeck B., Scholer F., Williams H.E. et Zobel, J. (2003). Query Expansion Using Associated Queries. In Proc. ACM CIKM, pp. 2-9.

Blansché A. (2006). Classification non-supervisée avec pondération d'attributs par des méthodes évolutionnaires. Thèse de Doctorat, Université Louis Pasteur, France.

Borlund, P. et Ingwersen, P. (1998). Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In: Proc. ACM SIGIR conference, pp. 24–28.

Boughanem M., Brini A. et Dubois D. (2009). Possibilistic networks for information retrieval. International Journal of Approximate Reasoning, 50(7), pp. 957-968.

Boughanem, M. et Brini, A. (2003). Introduction de la gradualité dans le jugement utilisateur. Dans : Actes d'EGC, vol. 17, pp. 343–348.

Boulaknadel S. (2006). Utilisation des syntagmes nominaux dans un système de recherche d'information en langue arabe. Dans : Actes de CORIA, pp. 341-346.

Boulaknadel S., Daille B. et Aboutajdine D. (2008). A multi-word term extraction program for arabic language. In Proc. LREC, pp. 1485-1488

Bounhas I. (2012). Construction et intégration d'ontologies pour la cartographie socio-sémantique de fonds documentaires arabes guidée par la fiabilité de l'information. Thèse de Doctorat en Informatique, Faculté des Sciences de Tunis, Université Tunis El Manar, Tunis, Tunisie.

Bounhas I. et Slimani Y. (2009a). A hybrid approach for Arabic multi-word term extraction. In Proc. NLPKE, pp. 1–8.

Bounhas I. et Slimani Y. (2009b). A social approach for semi-structured document modeling and analysis. In Proc. KMIS, Madeira, Portugal, pp. 95-102.

Bounhas I. et Slimani Y. (2011). Toward a methodology and a corpus for Arabic information sciences in the socio-semantic web. In: International Journal of Computing in Arabic, 3(3), pp. 67-80.

Bounhas I., Elayeb B., Evrard F. et Slimani Y. (2010). Toward a computer study of the reliability of Arabic stories. In: Journal of the American Society for Information Science and Technology, Wiley, 61(8), pp. 1686–1705.

Bounhas I., Elayeb B., Evrard F. et Slimani Y. (2011a). ArabOnto: Experimenting a new distributional approach for Building Arabic Ontological Resources. In: International Journal of Metadata, Semantics and Ontologies, Inderscience, 6(2), pp. 81-95.

Bounhas I., Elayeb B., Evrard F. et Slimani Y. (2011b). Organizing contextual knowledge for arabic text disambiguation and terminology extraction. In: Knowledge Organization Journal, Ergon Verlag of Würzburg, 38(6), pp. 473-490.

Bounhas I., Elayeb B., Evrard F. et Slimani Y. (2015a). Information reliability evaluation: from Arabic storytelling to computer sciences. In: ACM Journal on Computing and Cultural Heritage, 8(3), 14:1-14:33.

Bounhas M., Ghasemi M. H., Prade H., Serrurier M. et Mellouli K. (2014). Naïve possibilistic classifiers for imprecise or uncertain numerical data. In: Fuzzy Set and System, Elsevier, 239, pp. 137-156.

Bounhas M., Mellouli K., Prade H. et Serrurier M. (2013). Possibilistic Classifiers for numerical data. In: Soft Computing, 17(5), pp. 733-751.

Bounhas, I., Ayed R., Elayeb B., Evrard F. et Bellamine Ben Saoud N. (2015b). Experimenting a discriminative possibilistic classifier with reweighting model for Arabic morphological disambiguation. In: Computer Speech and Language, 33(2015), pp. 67-87.

Bounhas, I., Ayed R., Elayeb B., Evrard F. et Bellamine Ben Saoud N. (2015c). A hybrid possibilistic approach for Arabic full morphological disambiguation. In: Data & Knowledge Engineering, vol. 100, Part B, pp. 240-254.

Bourigault D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. Dans: Actes de TALN, pp. 75-84.

Bousmaha K.Z., Rahmouni M.K., Kouninef B. et Hadrich Belguith L. (2016). A Hybrid Approach for the Morpho-Lexical Disambiguation of Arabic. Journal of Information Processing Systems, 12(3), pp. 358–380.

Brin S. et Page, L. (1998). The Anatomy of a Large-scale Hypertextual Web Search Engine. In Proc. WWW, pp. 107–117.

Brini A. et Boughanem M. (2003). Relevance feedback: introduction of partial assessments for query expansion. In Proc. EUSFLAT, pp. 67–72.

Brini A., Boughanem M. et Dubois D. (2004). Towards a Possibilistic Approach for Information Retrieval. In: Proc. EUROFUSE Workshop, pp. 92-102.

Brown S.W., Dligach D. et Palmer M. (2011). Verbnet class assignment as a WSD task. In: Proc. IWCS, Stroudsburg, PA, USA, pp. 85–94.

Brun C., Jacquemin B. et Segond F. (2001). Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale. In : Traitement Automatique de la Langue, 42(3), pp. 667-691.

Cao G., Nie J.Y. et Bai J. (2005). Integrating word relationships into language models. In Proc. ACM SIGIR conference, pp. 298–305.

Carpineto C. et Romano G. (2012). A Survey of Automatic Query Expansion in Information Retrieval. In: ACM Computing Surveys, 44(1), pp. 1-50.

Carpuat M. et Wu D. (2007). Improving statistical machine translation using word sense disambiguation. In: Proc. EMNLP-CoNLL, pp. 61–72.

Chan Y.S., Ng H.T. et Chiang D. (2007). Word sense disambiguation improves statistical machine translation. In Proc. ACL, pp. 33–40.

Chen J., Hao Y. et Wang S. (2007). Improving Information Reliability in Mass Customization of Services: a Case Study from China's Catering Services. In: Proceedings of 6th Wuhan International Conference on E-Business, Wuhan, Hubei province, China.

Chevalier F., Huot S. et Fekete J. D. (2010). Visualisation de mesures agrégées pour l'estimation de la qualité des articles Wikipédia. Dans : Actes d'EGC'10), Hammamet, Tunisia, pp. 351-362.

Chifu A-G. et Ionescu R.T. (2012). Word sense disambiguation to improve precision for ambiguous queries. In: Central European Journal of Computer Science, 2(4), pp.398–411.

Chifu, A-G. et Mothe, J. (2014). Expansion sélective de requêtes par apprentissage. Dans: Actes de CORIA-CIFED, pp. 257-272.

Chouigui A., Ben Khiroun O. et Elayeb B. (2017). ANT Corpus: An Arabic News Text Collection for Textual Classification. In Proc. AICCSA, pp. 135-142.

Claveau V., Sébillot P. et De Beaulieu C. (2004). Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. Dans : Actes de TALN, Fès, Maroc.

Clough P. et Stevenson M. (2004). Cross-language information retrieval using Euro WordNet and word sense disambiguation. In: Proc. ECIR, UK, pp. 327–337.

Cohen J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), pp. 213-220.

Cohn T. (2003). Performance metrics for word sense disambiguation. In: Proceedings of the Australasian Language Technology Workshop, Melbourne, Australia, pp. 86-93.

Collins M. (2002). Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In Proc. ACL-EMNLP, pp. 1-8.

Crabtree D. (2009). Enhancing Web Search through Query Expansion. In: J. Wang (Ed.), Encyclopedia of Data Warehousing and Mining, Second Edition. IGI Global : Hershey, PA, USA, pp. 752-757.

Cui H., Wen J-R., Nie J-Y. et Ma W-Y. (2003). Query Expansion by Mining User Logs. IEEE Trans. on Knowledge and Data Engineering, 15(4), pp. 829-839.

Da Costa Pereira C. et Pasi G. (2007). Fuzzy Indices of Document Reliability. In: Applications of Fuzzy Sets Theory, LNCS 4578/2007, Springer Berlin Heidelberg, pp. 110-117.

Daoud D. (2009). Synchronized Morphological and Syntactic Disambiguation for Arabic. In: Advances in Computational Linguistics, 41, pp. 73-86.

Daoud D. et Daoud M. (2009). Arabic Disambiguation Using Dependency Grammar. In 16ème conférence Internationale de Traitement Automatique des Langues Naturelles, pp. 1-10.

Demsar J. (2006). Statistical comparisons of classifiers over multiple data sets. In: Journal of Machine Learning Research, 7, pp. 1-30.

Diab M. (2004). Word Sense Disambiguation within a Multilingual Framework. Ph.D. Thesis, University of Maryland, USA.

Diab M., Hacioglu K. et Jurafsky D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In: Proc. NAACL HLT, pp. 149-152.

Diab M.T. (2007). Improved Arabic Base Phrase Chunking with a New Enriched POS Tag Set. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages : Common Issues and Resources, pp. 89–96.

Diaz F., Mitra B. et Craswell N. (2016). Query Expansion with Locally-Trained Word Embeddings. In: Proc. ACL, Vol. 1, pp. 367-377.

Ding L., Li X. et Xing Y. (2005). Pushing scientific documents by discovering interest in information flow within e-science knowledge grid. In: Proc. GCC, pp. 498-510.

Dominich S. (2001). Mathematical Foundations of Information Retrieval. Kluwer Academic Publishers, Dordrecht, Boston, London, UK.

Dramé K., Mougin F. et Diallo, G. (2014). Query Expansion using External Resources for Improving Information Retrieval in the Biomedical Domain. In: Proc. CEUR, pp. 189-194.

Dubois D. et Prade H. (1987). Théorie des Possibilités : Application à la Représentation des Connaissances en Informatique. Paris, France: Edition Masson.

Dubois D. et Prade H. (1994). Possibility Theory: An Approach to computerized Processing of Uncertainty. Plenum Press, New York, USA.

Dubois D. et Prade H. (1998). Possibility Theory: Qualitative and Quantitative Aspects. In Handbook of defeasible reasoning and uncertainty management systems, Vol. 1, pp. 169–226.

Dubois D. et Prade H. (2000) An overview of ordinal and numerical approaches to causal diagnostic problem solving. In: Abductive Reasoning and Learning, Handbooks of Defeasible Reasoning and Uncertainty Management Systems. Klower Academic Publishers, pp. 231-280.

Dubois D. et Prade H. (2006) Représentations formelles de l'incertain et de l'imprécis. In: Concepts et méthodes pour l'aide à la décision - outils de modélisation, Paris, France, pp. 111-171.

Dubois D. et Prade H. (eds) (1988). Possibility Theory: An Approach to computerized Processing. Plenum Press, New York, USA.

Dubois D. et Prade H. (eds) (2009) Formal representations of uncertainty. In Decision-Making Process: Concepts and Methods. Wiley-ISTE, Hoboken, N.J. USA.

Duque A., Araujo L. et Martinez-Romo J. (2015). CO-graph : A new graph-based technique for cross-lingual word sense disambiguation. Natural Language Engineering, 21(5), pp.743-772.

Edmonds P. et Hirst G. (2002). Near-Synonymy and Lexical Choice. In: Computational Linguistics, 28(2), pp. 105–144.

Elayeb B. (2009). SARIPOD: Système multi-Agent de Recherche Intelligente POssibiliste de Documents Web. Thèse de Doctorat en Informatique, Institut National Polytechnique de Toulouse, France & Ecole Nationale des Sciences de l'Informatique, Tunisie.

Elayeb B. (2018). Arabic word sense disambiguation: A Review. In Artificial Intelligence Review, 51(xx), pp. 1-58. DOI:10.1007/s10462-018-9622-6 (to appear)

Elayeb B. et Bounhas I. (2016) Arabic Cross-Language Information Retrieval: A Review. In: ACM Transactions on Asian and Low-Resource Language Information Processing, 15(3), 18:1-18:44.

Elayeb B., Ben Romdhane W. et Bellamine Ben Saoud N. (2018). Towards a New Possibilistic Query Translation Tool for Cross-Language Information Retrieval. In Multimedia Tools and Applications, Springer, 77(2), pp. 2423-2465.

Elayeb B., Bounhas I., Ben Khiroun O. et Bellamine Ben Saoud N. (2015b). Combining Semantic Query Disambiguation and Expansion to Improve Intelligent Information Retrieval. In: Duval

B., van den Herik J., Loiseau S. et Filipe J. (eds.), ICAART2014 Revised Selected papers, LNAI 8946, pp. 280–295.

Elayeb B., Bounhas I., Ben Khiroun O., Evrard F. et Bellamine Ben Saoud N. (2011). Towards a Possibilistic Information Retrieval System Using Semantic Query Expansion. International Journal of Intelligent Information Technologies, 7(4), pp. 1-25.

Elayeb B., Bounhas I., Ben Khiroun O., Evrard F. et Bellamine Ben Saoud N. (2015a). A Comparative Study between Possibilistic and Probabilistic Approaches for Monolingual Word Sense Disambiguation. In Knowledge and Information Systems, 44(1), pp. 91-126.

Elayeb B., Evrard F., Zaghdoud M. et Ben Ahmed M. (2009). Towards An Intelligent Possibilistic Web Information Retrieval using Multiagent System. In: Interactive Technology and Smart Education: Special issue on New learning support systems, 6(1), pp. 40-59.

ElHadj Y., Al-Sughayeir I. et Al-Ansari A. (2009). Arabic Part-Of-Speech Tagging using the Sentence Structure. In: Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egypt, pp. 241-245.

Elshafei M., Al-Muhtaseb H.A. et Al-Ghamdi M. (2002). Techniques for high quality Arabic speech synthesis. In: Information Sciences, 140(3), pp. 255-267.

Ermakova, L. et Mothe, J. (2016). Query Expansion by Local Context Analysis. In: Proc. CORIA-CIFED, pp. 235-250.

Eugenio B.D. (2000). On the usage of Kappa to evaluate agreement on coding tasks. In: Proc. LREC, pp. 441-444.

Falquet G., Jiang C.L.M. et Ziswiler J.C. (2004). Intégration d'ontologies pour l'accès à une bibliothèque d'hyperlivres virtuels. In : Proc. AFRIF-AFIA, Toulouse, France.

Fang H. (2008). A re-examination of query expansion using lexical resources. In: Proc. ACL-HLT, pp. 139–147.

Fettweis G. et Meyr H. (1991). High-speed parallel Viterbi decoding: algorithm and VLSI-architecture. In: IEEE Communications Magazine, pp. 46- 55.

Forney G.D. (1973). The Viterbi algorithm. In: Proceedings of IEEE, 61, pp. 268-278.

Friedman N., Geiger D. et Goldszmidt M. (1997). Bayesian Network Classifiers. Machine Learning, 29(2-3), pp. 131–163.

Froehlich T. et Eisenberg M. (1994). Special topic issue on relevance research. Journal of the American Society for Information Science, 45(3), pp. 124–134.

Fu G., Jones C. B. et Abdelmoty A.I. (2005). Ontology-Based Spatial Query Expansion in Information Retrieval. In OTM Conferences and Workshops, pp. 1466-1482.

Fuhr N. (1992). Probabilistic models in information retrieval. The Computer Journal, 35(3), pp. 243–255.

Gan L. et Hong H. (2015). Improving query expansion for information retrieval using wikipedia. International Journal of Database Theory and Application,8(3), pp. 27-40.

Gaume B. (2004). Balades aléatoires dans les petits-mondes lexicaux. In : Information Interaction Intelligence, 4(2), pp. 39-96.

Gaume B. (2006). Cartographier la forme du sens dans les petits-mondes Lexicaux. In: Proceedings of the 8th International Conference on Textual Data and statistical Analysis. Presses Universitaires de Franche-Comté, Besançon, France, pp. 541-465.

Gaume B., Hathout N. et Muller P. (2004). Word sense disambiguation using a dictionary for sens similarity measure. In: Proc. ACL, pp. 1194-1200.

Ghazizadeh V., Zahedi M.H., Kahani M. et Minaei Bidgoli B. (2008). Fuzzy expertsystem in determining Hadith validity. In: Advances in Computer and Information Sciences and Engineering : Proceedings of the International Conference on Systems, Computing Sciences and Software Engineering, Bridgeport, USA, pp. 354-359.

Gonzalo J., Penas A. et Verdejo F. (1999). Lexical ambiguity and information retrieval revisited. In: Proc. the 1999 Joint SIGDAT-EMNLP, pp. 195–202.

Gonzalo J., Verdejo F., Chugur I. et Cigarrin J. (1998). Indexing with WordNet synsets can improve text retrieval. In: Proc. COLING-ACL, pp. 38–44.

Gruber T. (2006). Where the Social Web Meets the Semantic Web. Keynote Talk presented at the 5th International Semantic Web Conference, Athens, GA, USA.

Habash N. et Rambow O. (2005). Arabic Tokenization, Part-of-speech Tagging and Morphological Disambiguation in One Fell Swoop. In: Proc. ACL, pp. 573–580.

Habash N. et Rambow O. (2007). Arabic Diacritization Through Full Morphological Tagging. In: Proc. NAACL-HLT, pp. 53-56.

Habash N., Rambow O. et Roth R. (2009). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In: Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, The MEDAR Consortium, Cairo, Egypt.

Hadrich Belguith L. et Chaâben N. (2006). Analyse et désambiguïsation morphologiques de textes arabes non voyellés. Dans : Actes de TALN, pp. 493–501.

Hajic J. (2000). Morphological Tagging: Data vs. Dictionaries. In: Proc. NAACL, pp. 94-101.

Haouari B., Ben Amor N., Elouedi Z. et Mellouli K. (2009). Naïve possibilistic network classifiers. In : Fuzzy Sets and Systems, 160(22), pp. 3224-3238.

Harrag F., Alothaim A., Abanmy A., Alomaigan F. et Alsalehi S. (2013). Ontology Extraction Approach for Prophetic Narration (Hadith) using Association Rules. In: International Journal on Islamic Applications in Computer Science And Technology, 1(2), pp. 48-57.

Harrag F., Hamdi-Cherif A., Al-Salman A.M.S. et El-Qawasmeh E. (2009). Experiments in improvement of arabic information retrieval. In: Proc. CITALA, Rabat, Morocco.

Harter S. (1992). Psychological relevance and information science. Journal of the American Society for Information Science (JASIS), 43(9), pp. 602–615.

He B. et Ounis I. (2009). Studying query expansion effectiveness. In: Proc. ECIR, pp. 611-619.

Hersh, W.R., Bhupatiraju, R.T. et Price, S. (2003). Phrases, Boosting, and Query Expansion Using External Knowledge Resources for Genomic Information Retrieval. In : Proc. TREC, pp. 503-509.

Hoceini Y., Cheragui M.A. et Abbas M. (2011). Towards a New Approach for Disambiguation in NLP by Multiple Criterian Decision-Aid. In: The Prague Bulletin of Mathematical Linguistics, 95, pp. 19-32.

Ide N. et Véronis J. (1998). Word sense disambiguation: The state of the art. In: Computational Linguistics: Special Issue on Word Sense Disambiguation, 24, pp. 1-40.

Iksal S. et Garlatti S. (2002). Spécification déclarative pour des documents virtuels personnalisables. In : Actes du congrès Documents Virtuels Personnalisables (DVP), Brest, France, pp. 127-140.

Jansen B.J. et McNeese M.D. (2005). Evaluating the effectiveness of and patterns of interactions with automated searching assistance: Research Articles. In: Journal of American Society of Information Science and Technology, Wiley, 56(14), pp. 1480-1503.

Jenhani I., Ben Amor N. et Elouedi Z. (2008). Decision trees as possibilistic classifiers. In: International Journal of Approximate Reasoning, 48(3), pp. 784-807.

Jimeno-Yepes A.J., McInnes B.T. et Aronson, A.R. (2011). Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. In: BMC bioinformatics, 12(1), p. 223.

Jones R., Rey B., Madani O. et Greiner, W. (2006). Generating Query Substitutions. In Proceedings of the 15th International Conference on World Wide Web, pp. 387-396.

Jurafsky D. et Martin J.H. (2009). Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Pearson Prentice Hall, Upper Saddle River, New Jersey, USA.

Kekäläinen J. et Järvelin K. (2002). Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In Bruce, H., Fidel, R., P. Ingwersen, P. Vakkari, eds. Emerging Frameworks and Methods, Seattle, Colerado : Libraries Unlimited, pp. 253–270.

Keskes I., Beanamara F. et Hadrich Belguith L. (2013). Segmentation de textes arabes en unités discursives minimales. In : Actes de TALN, pp. 435-449.

Khapra M. M., Shah S., Kedia, P. et Bhattacharyya P. (2009). Projecting Parameters for Multilingual Word Sense Disambiguation. In: Proc. EMNLP, pp. 459–467.

Khemakhem A, Gargouri B. et Ben Hamadou A. (2013). Collaborative Enrichment of Electronic Dictionaries Standardized-LMF. In: Proc. NLDB, pp. 328-336.

Khoja Sh. (2001). APT: Arabic part-of-speech tagger. In: Proc. Student Workshop at NAACL, Carnegie Mellon University, Pennsylvania, USA.

Kilgarriff A. (1994). The myth of completeness and some problems with consistency (the role of frequency in deciding what goes in the dictionary). In: Proc. EURALEX, pp. 101–106.

Kim S.B., Seo, H.C. et Rim H.C. (2004). Information retrieval using word senses: root sense tagging approach. In: Proc. ACM SIGIR conference, pp. 258–265.

Knight S. et Burn J. (2005). Developing a Framework for Assessing Information Quality on the World Wide Web. In: Informing Science Journal, 8, pp. 59-73.

Koeling R, McCarthy D., et Carroll J. (2005). Domain-specific sense distributions and predominant sense acquisition. In: Proceedings HLT-EMNLAP, pp. 419–426.

Kohavi R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, Montréal Québec, Canada, pp. 1137-1143.

Koppula N., Rani B.P. et Rao K.S. (2017). Graph Based Word Sense Disambiguation. In Proceedings of the First International Conference on Computational Intelligence and Informatics, Advances in Intelligent Systems and Computing, Singapore, pp. 665-670.

Krause A. et Horvitz E. (2008). A utility-theoretic approach to privacy and personalization. In: Proceedings of the 23rd National Conference on Artificial Intelligence, pp. 1181–1188.

Krippendorff K. (1980). Content Analysis: an Introduction to its Methodology. Sage Publications, Beverly Hills, CA, USA.

Krovetz R. (1997). Homonymy and polysemy in information retrieval. In: Proc. European chapter of the ACL, pp. 72–79.

Krovetz R. et Croft W.B. (1992). Lexical ambiguity and information retrieval. In: ACM Transactions on Information Systems, 10(2), pp. 115–141.

Ksentini N., Tmar M. et Gargouri, F. (2016). The Impact of Term Statistical Relationships on Rocchio's Model Parameters for Pseudo Relevance Feedback. International Journal of Computer Information Systems and Industrial Management Applications, 8, pp. 135–144.

Kuzi S., Shtok A. et Kurland O. (2016). Query Expansion Using Word Embeddings. In Proc. ACM CIKM, pp. 1929-1932.

Lafourcade M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In: SNLP'07 : 7th international symposium on natural language processing, p. 7.

Lafourcade M. et Brun N.L. (2017). Extracting semantic relations via the combination of inferences, schemas and cooccurrences. In Proc. RANLP, Varna, Bulgaria, pp. 417-423.

Landis J.R. et Koch G.G. (1977). The measurement of observer agreement for categorical data. In: Biometrics, 33(1), pp. 159-174.

Latiri C., Haddad H. et Hamrouni T. (2012). Towards an effective automatic query expansion process using an association rule mining approach. In: Journal of Intelligent Information Systems, 39(1), pp. 209-247.

Lee H.M., Chen C.M., Chen J.M. et Jou Y.L. (2001). An efficient fuzzy classifier with feature selection based on fuzzy entropy. In: IEEE Transactions on Systems, Man, and Cybernetics, Part B, 31(3), pp. 426, 432.

Lefever E. et Hoste V. (2010). SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 15–20.

Lefever E. et Hoste V. (2013). SemEval-2013 Task 10: Cross-Lingual Word Sense Disambiguation. In: Proceedings of the 7th International Workshop on Semantic Evaluation, pp. 158–166.

Lioma C., He B., Plachouras V. et Ounis I. (2005). The University of Glasgow at CLEF 2004: French Monolingual Information Retrieval with Terrier, In : Proc. CLEF, pp. 253-259.

Liping J., Ng M.K. et Huang J.Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. In: IEEE Transaction on Knowledge and Data Engineering, 19 (8), pp. 1026–1041.

Liu S., Liu F., T. Yu C. et Meng W. (2004). An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In: Proc. ACM SIGIR conference, pp. 266–272.

Liu S., T. Yu C. et Meng W. (2005). Word sense disambiguation in queries. In: Proc. ACM CIKM, pp. 525–532.

Loupy C. (2000). Evaluation de l'apport de connaissances linguistiques en désambiguïsation sémantique et recherche documentaire. Thèse de Doctorat, Université d'Avignon, France.

Lucas S. (2002). The Arts of Hadith Compilation and Criticism. USA: University of Chicago, OCLC 62284281.

Lv C., Qiang R., Fan F. et Yang J. (2015). Knowledge-Based Query Expansion in Real-Time Microblog Search. In Information Retrieval Technology, pp. 43-55.

Lynch C.A. (2001). When Documents Deceive: Trust and Provenance as New Factors for Information Retrieval in a Tangled Web. In: Journal of the American Society for Information Science and Technology, Wiley, 52(1), pp. 12-17.

Maamouri M., Bies A. et Kulick S. (2009). Creating a Methodology for Large-Scale Correction of Treebank Annotation: The Case of the Arabic Treebank. In: the proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt, pp. 138–144.

Manning C.D., Raghavan P. et Schütze H. (2008). Introduction to Information Retrieval, New York, NY, USA: Cambridge University Press.

Mansour S., Sima'an K., et Winter Y. (2007). Smoothing a lexicon-based pos tagger for Arabic and Hebrew. In Proceedings of ACL07 Workshop on Computational Approaches to Semitic Languages, Prague, Czech, pp. 97-103.

Maron M. (1961). Automatic indexing: an experimental enquiry. Journal of the ACM, 24(8), pp. 404–417.

Mihalcea R. (2005). Unsupervised Large-vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling. In Proceedings HLTEMNLP, pp. 411–418.

Milne D. et Witten, I. H. (2008). Learning to Link with Wikipedia. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 509–518.

Mizzaro S. (1997). Relevance: the whole history. Journal of the American Society for Information Science (JASIS), 48(9), pp. 810–832.

Montgomery J., Si L., Callan J. et Evans D. A. (2004). Effect of Varying Number of Documents in Blind Feedback : Analysis of the 2003 NRRC RIA Workshop "Bf_numdocs" Experiment Suite. In: Proc. ACM SIGIR Conference, pp. 476–477.

Nasiruddin M. (2013). État de l'art de l'induction de sens : une voie vers la désambiguïsation lexicale pour les langues peu dotées. Dans: Actes de RECITAL, pp. 192–205.

Naumann F. et Rolker C. (2000). Assessment methods for information quality criteria. In: Proc. Information Quality (IQ), Cambridge, MA, USA, pp. 396-403.

Navigli R. (2009). Word sense disambiguation: A survey. In: ACM Computing Surveys, 41(2), pp.10:1–10:69.

Navigli R. et Lapata M. (2010). An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. IEEE Trans. Pattern Anal. Mach. Intell., 32(4), pp. 678-692.

Newman M. E. J. (2003). The structure and function of complex networks. In: SIAM Review, 45(2), 167-256.

Nguyen K-H. et Ock Ch-Y. (2013). Word sense disambiguation as a traveling salesman problem. In: Artificial Intelligence Review, 40(4), pp. 405-427.

Nie J-Y., Chevallet J-P. et Bruandet M-F. (1997). Between Terms and Words for European Language IR and Between Words and Bigrams for Chinese IR. In: TREC 1997, pp. 697-709.

Nilsson K., Hjelm H. et Oxhammar, H. (2006). SUiS–crosslanguage ontology-driven information retrieval in a restricted domain. In Proceedings of the 15th Nordic Conference of Computational Linguistics, pp.139-145.

Noorian Z. et Ulieru M. (2010). The state of the art in trust and reputation systems: A framework for comparison. In: Journal of Theoretical and Applied Electronic Commerce Research, 5(2), pp. 97-117.

Ogilvie P., Voorhees E. et Callan J. (2009). On the number of terms used in automatic query expansion. In: Information Retrieval, 12(6), pp. 666-679.

Othman E., Shaalan K. et Rafea A. (2004). Towards Resolving Ambiguity in Understanding Arabic Sentence. In: Proceedings of International Conference on Arabic Language Resources and Tools, NEMLAR, Egypt, pp. 118-122.

Pal D., Mitra M. et Datta K. (2014). Improving query expansion using WordNet. Journal of the Association for Information Science & Technology, 65(12), pp. 2469-2478.

Panchenko A., Ruppert E., Faralli S., Ponzetto S.P. et Biemann, C. (2017). Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation. In: Proc. European Chapter of ACL, pp. 86-98.

Parker M., Stofberg C. et De la Harpe R. (2006). Data quality: how the flow of data influences data quality in a small to medium medical practice. Paper presented at the Community informatics for developing countries: Understanding and organizing for a participatory future information society, Cape Town, South Africa.

Pasha A., Al-Badrashiny M., Diab M., El Kholy A., Eskander R., Habash N., Pooleery M., Rambow O. et Roth, R. (2014). MADAMIRA : A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In Proc. LREC, pp. 1094-1101.

Paskalis F.B.D. et Khodra M.L. (2011). Word sense disambiguation in information retrieval using query expansion. In: Proceedings of the IEEE-2011 International Conference on Electrical Engineering and Informatics. Bandung, Indonesia, pp. 1-6.

Picton A., Fabre C. et Bourigault D. (2008). Méthodes linguistiques pour l'expansion de requêtes. Une expérience basée sur l'utilisation du voisinage distributionnel. In : Revue française de linguistique appliquée, 13(1), pp. 83-95.

Pinto F.J. et Pérez-sanjulián C.F. (2008). Automatic query expansion and word sense disambiguation with long and short queries using WordNet under vector model. In : Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos, 2(2), pp.17–23.

Ploux S. et Victorri B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. In : Traitement automatique des langues, 39(1), pp.161-182.

Ponte J.M. et Croft W.B. (1998). A language modeling approach to information retrieval. In: Proc. ACM SIGIR conference, pp. 275–281.

Ponzetto S.P. et Navigli R. (2010). Knowledge-Rich Word Sense Disambiguation Rivaling Supervised Systems. In Proc. ACL, pp. 1522-1531.

Prade H. et Testemale C. (1987). Application of possibility and necessity measures to documentary information retrieval. In: Uncertainty in Knowledge-Based Systems, pp. 265-274.

Richardson M., Agrawal R., et Domingos P. (2003). Trust Management for the Semantic Web. In: Proceedings of International Semantic Web Conference, LNCS 2870, pp. 351-368.

Rijsbergen C.V. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. In Journal of Documentation, 33, pp. 106–119.

Rijsbergen C.V. (1979). Information Retrieval. Butterworth-Heinemann, Newton, MA, USA.

Robertson S. et Jones K.S. (1976). Relevance weighting of search terms. Journal of the American Society for Information Science (JASIS), 27(3), pp. 129–146.

Robertson S., Maron M.E. et Cooper, W.S. (1982). Probability of relevance : a unification of two competing models for information retrieval. Information Technology - Research and Development, 1, pp. 1-21.

Rocchio J. (1971). Relevance Feedback in Information Retrieval. In: The SMART Retrieval System. Prentice-Hall, Englewood Cliffs, New Jersey, USA, pp. 313–323.

Rodriguez H., Farwell D, Farreres J., Bertran M., Alkhalifa M. et Marti M.A. (2008). Arabic WordNet: Semi-automatic extensions using bayesian inference. In: Proc. LREC, pp. 1702-1706.

Romberg J. (2017). Comparing Relevance Feedback Techniques on German News Articles. In: Proceedings of Datenbanksysteme für Business, Technologie und Web, pp. 301–310.

Romney M. et Romney G.W. (2005). Security & reliability are provided by a Web-based classroom electronic document management process. In: Proceedings of the 6th International Conference on Information Technology Based Higher Education and Training, pp. T3A/1 - T3A/4.

Roth R., Rambow O., Habash N., Diab M. et Rudin C. (2008). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In: Proc. ACL, pp. 117-120.

Ruthven I. (2003). Re-examining the potential effectiveness of interactive query expansion. In: Proc. ACM SIGIR conference, pp. 213–220.

Salton G. (1971). The Smart retrieval system-experiments. Automatic Document Processing, Prentice Hall Inc.

Sanderson M. (1994). Word sense disambiguation and information retrieval. In: Proc. ACM SIGIR conference, pp. 142–151.

Sanderson M. (2000). Retrieving with good sense. In: Information Retrieval, 2(1), pp. 49–69.

Saracevic T. (1975). Relevance : A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 32(2), pp. 321–343.

Saracevic T. (1996). Relevance reconsidered. In Information science: Integration in perspectives, Proc. of the Conference on Conceptions of Library and Information Science, pp. 201–218.

Schamber L., Eisenberg M. et Nilan M. (1990). A re-examination of relevance : Toward a dynamic, situational definition. Information Processing and Management : an International Journal, 26(6), pp. 755–776.

Schamber L., Eisenberg M. et Nilan S.M. (1990). A re-examination of relevance toward a dynamic, situational definition. In: Information Processing and Management, 26(6), pp. 755-776.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, pp. 44-49.

Schütze H. et Pedersen J.O. (1995). Information retrieval based on word senses. In: Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, pp. 161–175.

Segond F. (2000). Framework and Results for French. In: Computers and the Humanities, 34(1), pp. 49-60.

Shaalan K. et Raza H. (2007). Person Name Entity Recognition for Arabic. In: Proceedings of the Workshop on Computational Approaches to Semitic Languages, Madison, USA, pp. 17-24.

Shaalan K. et Raza H. (2009). NERA: Named Entity Recognition for Arabic. In: Journal of the American Society for Information Science and Technology, 60(8), pp. 1652-1663.

Shafer G. (1976). A mathematical theory of evidence. Princeton, NJ: Princeton University Press.

Silberer C. et Ponzetto, S. P. (2010). UHD : Cross-Lingual Word Sense Disambiguation Using Multilingual Co-Occurrence Graphs. In Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 134–137.

Singhal A., Mitra M. et Buckley C. (1997). Learning routing queries in a query zone. In: Proc. ACM SIGIR Conference, pp. 25–32.

Sinha R. et Mihalcea, R. (2007). Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. In Proc. ICSC, pp. 363-369.

Smeaton A.F. (1997). Using NLP or NLP resources for information retrieval tasks. In: Natural Language Information Retrieval, Kluwer: Dordrecht, pp. 99-111.

Soto A, Olivas J.A. et Prieto M. E. (2008). Fuzzy Approach of Synonymy and Polysemy for Information Retrieval. In: Granular Computing: At the Junction of Rough Sets and Fuzzy Sets Studies in Fuzziness and Soft Computing, 224, pp. 179-198.

Stokoe C., Oakes M.P. et Tait J. (2003). Word sense disambiguation in information retrieval revisited. In: Proc. ACM SIGIR conference, pp. 159–166.

Stvilia B. (2008). A Workbench for Information Quality Evaluation. In: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries. New York, NY, USA, pp. 469-469.

Stvilia B., Gasser L., Twidale M.B. et Smith L.C. (2007). A Framework for Information Quality Assessment. In: Journal of the American Society for Information Science and Technology, 58(12), pp. 1720 – 1733.

Tlili-Guiassa Y. (2006). Hybrid Method for Tagging Arabic Text. In: Journal of Computer Science, 2(3), pp. 245-248.

Turtle H.R. et Croft, W.B. (1991). Evaluation of an inference network-based retrieval model. ACM Transaction on Information Systems, 9(3), pp. 7187–222.

van Rijsbergen C. J. (1986). A non-classical logic for information retrieval. Computer Journal, 29(6), pp. 481–485.

Vélez B., Weiss R., Sheldon M. et Gifford D. (1997). Fast and effective query refinement. In: Proc. ACM SIGIR conference, pp. 6-15.

Véronis J. (1998). A study of polysemy judgements and inter-annotator agreement. In: Program and advanced papers of the Senseval workshop. Herstmonceux Castle, England, pp. 2-4.

Véronis J. (2003). Sense tagging: does it make sense? In: Proceedings of the Corpus Linguistics 2001 conference. Peter Lang Frankfurt, Lancaster, UK.

Véronis J. et Ide N. (1990). Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In: Proc. COLING, Helsinki, Finland, pp. 389–394.

Véronis, J. (2004). HyperLex : lexical cartography for information retrieval. Computer Speech & Language, 18(3), pp. 223–252.

Vidhu Bhala R.V. et Abirami S. (2012). Trends in word sense disambiguation. In: Artificial Intelligence Review, pp. 1-13.

Viera A.J. et Garrett J. M. (2005). Understanding interobserver agreement: the kappa statistic. In: Family Medecine, 37(5), pp. 360-363.

Vignaux G. (2005). La recherche d'information: Panorama des questions et des recherches. Paris Nord: CNRS-MSH.

Voorhees E. M. (1994). Query expansion using lexical-semantic relations. In: Proc. ACM SIGIR conference, pp. 61-69.

Voorhees E.M. (1993). Using WordNet to disambiguate word senses for text retrieval. In: Proc. ACM SIGIR conference, pp. 171–180.

Watts D.J. et Strogatz S. H. (1998). Collective dynamics of 'small-world' networks. In: Nature, 393(6684), pp. 440-442.

Xu Y. et Chen Z. (2006). Relevance judgment: What do information users consider beyond topicality? In: Journal of the American Society for Information Science and Technology, 57(7), pp. 961 – 973.

Yarowsky D. (2000). Hierarchical decision list for word sense disambiguation. In: Computers and the humanities, 34(1-2), pp. 179–186.

Yarowsky D., Cucerzan S., Florian R., Schafer C. et Wicentowski R. (2001). The johns hopkins SENSEVAL2 system descriptions. In: Proceedings of The 2nd International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France, pp. 163–166.

Yin Z., Shokouhi M. et Craswell N. (2009). Query Expansion Using External Evidence. In: Advances in Information Retrieval, pp. 362-374.

Yue Y. (2012). A multi-classified method of support vector machine (SVM) based on entropy. In: Applied Mechanics and Materials, 241–244, pp. 1629–1632.

Yuret D. et Yatbaz M.A. (2010). The Noisy Channel Model for Unsupervised Word Sense Disambiguation. In: Computational Linguistics, 36(1), pp. 111-127.

Yuso Y., Ismail R. et Hassan Z. (2010). Adopting Hadith verification techniques in to digital evidence authentication. In: Journal of Computer Science, 6(5), pp. 484-489.

Zacklad M. (2007). Processus de documentarisation dans les Documents pour l'Action (DopA). In : Babel – edit, Le numérique: impact sur le cycle de vie du document, Montréal, Canada, pp. 1-28.

Zadeh L.A. (1965). Fuzzy sets. In: Information and control, 8, pp. 338-353.

Zadeh L.A. (1978). Fuzzy sets as a basis for a theory of possibility. In: Fuzzy Sets and Systems, 1(1), pp. 3–28.

Zhang J., Deng B. et Li X. (2009). Concept Based Query Expansion Using WordNet. In: Proceedings of the 2009 International e-Conference on Advanced Science and Technology, Washington, DC, USA, pp. 52-55.

Zhong Z. et Ng H.T. (2012). Word sense disambiguation improves information retrieval. In: Proc. ACL, pp. 273–282.

Zhou X. et Han H. (2005). Survey of word sense disambiguation approaches. In: Proc. FLAIRS, pp. 307-313.

Zingla M.A., Latiri C., Mulhem P., Berrut C. et Slimani Y. (2018). Hybrid query expansion model for text and microblog information retrieval. In: Information Retrieval Journal, 21(4), pp. 337-367.

Zitouni I., Sorensen J., Luo X. et Florian R. (2005). The Impact of Morphological Stemming on Arabic Mention Detection and Coreference Resolution. In: Proceedings of the ACL workshop on Computational Approaches to Semitic Languages, Madison, USA, pp. 63-70.

Zouaghi A., Merhbene L. et Zrigui M. (2012). Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation. In: Artificial Intelligence Review, 38(4), pp. 257-269.

Zribi C., Torjmen A. et Ben Ahmed M. (2006). An Efficient Multi-agent System Combining POS-Taggers for Arabic Texts. In: Proc. CICLing, pp. 121-131.