Évaluation et adaptation de plongements lexicaux au domaine à travers l'exploitation de connaissances syntaxiques et sémantiques

Alexandra Benamar

université PARIS-SACLAY

Évaluation et adaptation de plongements lexicaux au domaine à travers l'exploitation de connaissances syntaxiques et sémantiques

Domain evaluation and adaptation of word embeddings through the exploitation of syntaxic and semantic knowledge

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 d'accréditation, Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique. Référent : Faculté des sciences d'Orsay

Thèse préparée dans le **Laboratoire interdisciplinaire des sciences du numérique (Université Paris-Saclay, CNRS)**, sous la direction d' **Anne VILNAT**, Professeure, le co-encadrement de **Cyril GROUIN**, ingénieur de recherche

Thèse soutenue à Paris-Saclay, le 25 mai 2023, par

Alexandra BENAMAR

Composition du jury

Membres du jury avec voix délibérative

Fatiha SAÏS Professeure, Université Paris Saclay	Présidente
Farah BENAMARA ZITOUNE Maîtresse de conférences, Université Paul Sabatier	Rapportrice & Examinatrice
Benoît CRABBÉ Professeur, Université Paris Cité	Rapporteur & Examinateur
Delphine BERNHARD Maîtresse de conférences, Université de Strasbourg	Examinatrice

THESE DE DOCTORAT

NNT : 2023UPASG035

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Évaluation et adaptation de plongements lexicaux au domaine à travers l'exploitation de connaissances syntaxiques et sémantiques

Mots clés : plongements lexicaux, adaptation au domaine, réseaux de neurones profonds, traitement automatique des langues

Résumé : Les modèles de plongements lexicaux se sont imposés comme les modèles de représentation les plus populaires en TAL. Afin d'obtenir de bonnes performances, ils nécessitent d'être entraînés sur de grands corpus de données provenant principalement du domaine général et sont fréquemment affinés pour être appliqués à des données de spécialité. Cependant, l'affinage des données est une pratique coûteuse en termes de ressources et son efficacité est controversée.

Dans le cadre de cette thèse, nous évaluons l'utilisation de modèles de plongements lexicaux sur des corpus de spécialité et nous montrons que la proximité entre les vocabulaires des données d'entraînement et des données d'application joue un rôle majeur dans la représentation des termes hors-vocabulaire. Nous observons que cela est principalement dû à la tokenisation initiale des mots, et nous proposons une mesure pour calculer l'impact de la segmentation des mots sur leur représentation.

Pour résoudre ce problème, nous proposons deux méthodes permettant d'injecter des connaissances linguistiques aux représentations générées par les Transformer : une méthode intervient à l'échelle des données et l'autre à l'échelle du modèle. Notre recherche démontre que l'ajout de contexte syntaxique et sémantique peut améliorer l'application de modèles auto-supervisés à des domaines de spécialité, tant pour la représentation du vocabulaire que pour la résolution de tâches de TAL. Les méthodes proposées peuvent être utilisées pour n'importe quelle langue disposant d'informations linguistiques ou d'autres connaissances externes. Le code utilisé pour les expériences a été publié pour faciliter la reproductibilité et des mesures ont été prises pour limiter l'impact environnemental en réduisant le nombre d'expériences.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title : Domain evaluation and adaptation of word embeddings through the exploitation of syntaxic and semantic knowledge

Keywords : word embeddings, domain adaptation, deep neural networks, natural language processing

Abstract : Word embeddings have established themselves as the most popular representation in NLP. To achieve good performance, they require training on large data sets mainly from the general domain and are frequently finetuned for specialty data. However, finetuning is a resource-intensive practice and its effectiveness is controversial.

In this thesis, we evaluate the use of word embedding models on specialty corpora and show that proximity between the vocabularies of the training and application data plays a major role in the representation of out-of-vocabulary terms. We observe that this is mainly due to the initial tokenization of words and propose a measure to compute the impact of the tokenization of words on their representation. To solve this problem, we propose two methods for injecting linguistic knowledge into representations generated by Transformers : one at the data level and the other at the model level. Our research demonstrates that adding syntactic and semantic context can improve the application of self-supervised models to specialty domains, both for vocabulary representation and for NLP tasks.

The proposed methods can be used for any language with linguistic information or external knowledge available. The code used for the experiments has been published to facilitate reproducibility and measures have been taken to limit the environmental impact by reducing the number of experiments.

Remerciements

Je remercie toutes les personnes qui m'ont épaulée durant cette aventure.

Tout d'abord, je tiens à remercier les personnes avec qui j'ai le plus échangé durant cette thèse, à savoir mes encadrants. Merci à Anne et Cyril pour votre suivi durant ces trois dernières années, des débuts en visioconférence pendant le confinement aux dernières retouches du manuscrit. Je voudrais vous remercier pour votre patience et votre confiance durant ce projet de recherche. Votre expertise et votre expérience ont été précieuses pour la réussite de ma thèse.

Je tiens également à remercier Meryl, qui a su pousser une ébauche de sujet au sein d'EDF et ainsi obtenir un financement pour ce projet de thèse. Je te remercie de m'avoir permis de commencer et de terminer ma thèse dans des conditions sereines, et de m'avoir intégrée au sein de ton équipe de travail.

Je remercie ensuite l'ensemble des membres du jury, qui m'ont fait l'honneur de bien vouloir étudier avec attention mon travail. Nos échanges de grande qualité pendant ma soutenance ont été une expérience enrichissante dont j'ai pleinement bénéficié.

À mes collègues de travail. Je tiens à remercier chaleureusement Marie pour m'avoir accueillie au sein de son équipe pendant ma thèse, ainsi que Benoît pour m'avoir offert l'opportunité de travailler sur son projet à deux reprises. Merci à Philippe de m'avoir facilité la recherche de mon laboratoire et merci à Clément d'avoir partagé son vécu de doctorant avec moi. Je voudrais également adresser mes remerciements à Ghislain pour m'avoir prodigué des conseils judicieux et pour m'avoir partagé son expérience pendant trois ans, ce qui m'a été d'une grande aide tout au long de mon parcours.

À Sarah, je tiens à témoigner de ma profonde gratitude pour le précieux soutien que tu m'as apporté au cours de cette dernière année éprouvante. Ta présence m'a offert une échappatoire après de longues journées de travail, tandis que ton expertise académique m'a prodigué des conseils avisés pour la rédaction de mon manuscrit.

À ma famille, sans qui rien n'aurait été possible.

Je tiens tout d'abord à adresser mes remerciements les plus sincères à ma chère maman, dont l'instinct infaillible lui permet de deviner les épreuves à venir avant même qu'elles ne pointent le bout de leur nez. Ta patience et ta détermination ont été une source d'inspiration, et ton soutien inconditionnel a été pour moi un baume apaisant dans les moments de doute et de stress.

Enfin, j'adresse mes remerciements à ma sœur Sarah. Tu as été mon soutien indéfectible tout au long de cette période, et ta présence est devenue pour moi un refuge inconditionnel lorsque les choses ne tournaient pas rond. Tu as su me guider dans les moments de doute, me redonner confiance en moi et m'encourager à poursuivre mes efforts. Pour toi, un millier de fois.

Résumé

Les modèles de plongements lexicaux se sont imposés comme les modèles de représentation les plus populaires en TAL. Afin d'obtenir de bonnes performances, ils nécessitent d'être entraînés sur de grands corpus de données provenant principalement du domaine général et sont fréquemment affinés pour être appliqués à des données de spécialité. Cependant, l'affinage des données est une pratique coûteuse en termes de ressources et son efficacité est controversée.

Dans le cadre de cette thèse, nous évaluons l'utilisation de modèles de plongements lexicaux sur des corpus de spécialité et nous montrons que la proximité entre les vocabulaires des données d'entraînement et des données d'application joue un rôle majeur dans la représentation des termes hors-vocabulaire. Nous observons que cela est principalement dû à la tokenisation initiale des mots, et nous proposons une mesure pour calculer l'impact de la segmentation des mots sur leur représentation.

Pour résoudre ce problème, nous proposons deux méthodes permettant d'injecter des connaissances linguistiques aux représentations générées par les Transformer : une méthode intervient à l'échelle des données et l'autre à l'échelle du modèle. Notre recherche démontre que l'ajout de contexte syntaxique et sémantique peut améliorer l'application de modèles auto-supervisés à des domaines de spécialité, tant pour la représentation du vocabulaire que pour la résolution de tâches de TAL. Les méthodes proposées peuvent être utilisées pour n'importe quelle langue disposant d'informations linguistiques ou d'autres connaissances externes. Le code utilisé pour les expériences a été publié pour faciliter la reproductibilité et des mesures ont été prises pour limiter l'impact environnemental en réduisant le nombre d'expériences.

Mots-clés : plongements lexicaux, adaptation au domaine, réseaux de neurones profonds, traitement automatique des langues

Abstract

Word embeddings have established themselves as the most popular representation in NLP. To achieve good performance, they require training on large data sets mainly from the general domain and are frequently finetuned for specialty data. However, finetuning is a resource-intensive practice and its effectiveness is controversial.

In this thesis, we evaluate the use of word embedding models on specialty corpora and show that proximity between the vocabularies of the training and application data plays a major role in the representation of out-of-vocabulary terms. We observe that this is mainly due to the initial tokenization of words and propose a measure to compute the impact of the tokenization of words on their representation. To solve this problem, we propose two methods for injecting linguistic knowledge into representations generated by Transformers: one at the data level and the other at the model level. Our research demonstrates that adding syntactic and semantic context can improve the application of self-supervised models to specialty domains, both for vocabulary representation and for NLP tasks.

The proposed methods can be used for any language with linguistic information or external knowledge available. The code used for the experiments has been published to facilitate reproducibility and measures have been taken to limit the environmental impact by reducing the number of experiments.

Keywords: word embeddings, domain adaptation, deep neural networks, natural language processing

Table des Matières

Liste des Figures

Liste des abréviations

CNIL . . . . . Commission Nationale Informatique et Libertés, http://www.cnil.fr/

EDF . . . . . Électricité de France, entreprise française publique de production et de fourniture d'électricité en France et en Europe.

EN . . . . . Entité Nommée

IA . . . . . Intelligence Artificielle

LR . . . . . (*Learning Rate*) Le pas d'apprentissage est un hyperparamètre qui joue sur la rapidité de la descente de gradient lors de l'apprentissage d'un modèle.

OCR . . . . . L'oscérisation est le processus qui transforme automatiquement l'image à l'intérieur d'un fichier en un fichier texte distinct.

OOV . . . . . (*Out-Of-Vocabulary*) Les termes hors-vocabulaire désignent les termes qui n'apparaissent pas dans le vocabulaire d'un modèle.

POS . . . . . (*Part-Of-Speech*) Les étiquettes morphosyntaxiques permettent d'associer aux mots d'un texte les informations grammaticales correspondantes (e.g., le genre, le nombre, la partie d'un discours).

REN . . . . . Reconnaissance d'entités nommées.

RGPD . . . . . Le « Règlement Général sur la Protection des Données » encadre le traitement des données personnelles sur le territoire de l'Union européenne et s'inscrit dans la continuité de la Loi française Informatique et Libertés de 1978 et renforce le contrôle par les citoyens de l'utilisation qui peut être faite des données les concernant.

SMS . . . . . (*Short Message Service*) Le service de messagerie permet de transmettre de courts messages textuels. C'est l'un des services de la téléphonie mobile.

TAL . . . . . (*Natural Language Processing*) Le Traitement Automatique du Langage est une discipline qui a pour objectif de modéliser, grâce à des outils informatique, le langage écrit ou parlé.

Ce n'est pas bon de se combler dans les rêves, en oubliant de vivre.

— Harry Potter à l'école des sorciers

1 Introduction

Table des Matières

1.1 Contexte de la thèse

1.1.1 Contexte scientifique

Le traitement automatique des langues (TAL, parfois appelé traitement automatique du langage) est un domaine en constante évolution qui vise à améliorer l'interaction entre les machines et les humains grâce à l'analyse automatisée du langage naturel. Les modèles ont connu une progression rapide ces dernières années, notamment en raison de l'utilisation de techniques d'apprentissage automatique profond. Cependant, il reste encore des défis importants à relever pour améliorer les performances de ces modèles.

L'une des principales difficultés est de généraliser les modèles aux données de spécialité, c'est-à-dire des domaines spécifiques tels que la médecine, la finance, les sciences, etc. Les modèles entraînés sur des données générales peuvent ne pas être aussi précis pour des tâches spécifiques à un domaine. En outre, la disponibilité limitée de données de spécialité peut rendre difficile l'entraînement et l'affinage de modèles pour des domaines spécifiques.

Dans ce contexte, l'adaptation de domaine et l'ajout d'informations linguistiques sont des approches prometteuses pour améliorer les performances des modèles pour des domaines spécifiques. L'adaptation de domaine vise à utiliser des données de spécialité pour l'entraînement ou l'affinage des modèles, tandis que l'ajout d'informations linguistiques vise à utiliser des connaissances linguistiques supplémentaires pour la représentation des données ou pour la résolution de tâches. Les enjeux de l'adaptation de domaine sont les suivants :

1. Couverture de vocabulaire : les modèles formés sur des données générales peuvent avoir des difficultés à représenter des termes spécifiques à un domaine, tels que les termes techniques ou les abréviations. L'adaptation de domaine peut aider à améliorer la couverture de vocabulaire en utilisant des données spécifiques à un domaine pour l'entraînement.

2. Précision des tâches : les modèles entraînés sur des données générales peuvent ne pas être aussi précis pour des tâches spécifiques à un domaine, telles que la classification de documents scientifiques ou la détection de sentiments dans des messages de médias sociaux. L'adaptation de domaine peut aider à améliorer la précision en utilisant des données spécifiques à un domaine pour l'entraînement ou l'affinage.

3. Coût des ressources : l'entraînement et l'affinage de modèles sur des données de spécialité peuvent être coûteux en termes de temps et de calcul, surtout si les données sont rares ou difficiles à obtenir. Il est donc important de trouver un bon compromis entre les performances améliorées et les coûts des ressources.

4. Effet de transfert : l'adaptation de domaine peut améliorer les performances sur un domaine spécifique mais cela peut aussi entraîner une diminution des performances sur d'autres domaines. Il est donc important de considérer l'effet de transfert sur les performances globales des modèles.

La tâche d'adaptation au domaine se révèle ardue, car elle comporte une multiplicité de problématiques de recherche complexes, que nous examinerons avec diligence dans le cadre de notre recherche. Nous nous appliquons à explorer la

problématique de l'adaptation des modèles à un domaine spécifique, en évaluant les représentations existantes et en cherchant à les améliorer. Ainsi, nous soumettons des propositions pour adapter ces modèles au domaine en question, en ajoutant des informations linguistiques.

1.1.2 Applications pour EDF

Nos travaux s'inscrivent dans le cadre particulier d'une collaboration avec un partenaire industriel. Électricité de France (EDF) est une entreprise française et est le principal producteur et fournisseur d'électricité en France et en Europe. Elle a été créée en 1946 pour gérer la production, la transmission et la distribution d'électricité en France. Aujourd'hui, elle est responsable de la production d'électricité à partir de diverses sources d'énergie, notamment nucléaire, hydroélectrique, éolienne et solaire. Elle est également responsable de la distribution de l'électricité à travers le réseau national de distribution, ainsi que de la vente d'électricité aux particuliers, aux entreprises et aux collectivités.

En raison du nombre élevé de clients, elle cherche à disposer de solutions d'aide au traitement des messages de sa clientèle. Ainsi, Cameli@ [Dubuisson Duplessis et al., 2020] est un outil développé par EDF pour aider la Direction Commerce à surveiller et analyser les avis et les retours des clients. Cet outil utilise des techniques de traitement automatique des langues pour extraire des informations à partir de différents types de corpus textuels tels que les tweets, les réclamations, les courriels, les articles de blogs, etc. Il permet de classer ces informations dans des catégories liées aux métiers ou aux types de retours à l'aide d'algorithmes de classification supervisée. Cet outil permet à EDF de mieux comprendre les besoins et les préoccupations des clients et de les prendre en compte pour améliorer les services proposés. Il permet également d'identifier les tendances et les thèmes récurrents dans les commentaires des clients pour élaborer des stratégies d'amélioration efficaces.

L'utilisation des modèles de plongements lexicaux est une pratique commune dans l'entreprise. Pour cela, l'adaptation des modèles aux spécificités du domaine de l'énergie est essentielle. De manière générale, l'entreprise souhaite évaluer ces modèles sur ses données et les améliorer afin de mettre à jour les performances de classement qu'elle utilise quotidiennement pour ses analyses. Afin d'être en mesure d'identifier les problèmes de représentations spécifiques à leurs données, il est nécessaire d'effectuer des travaux de recherche et de mettre en place des modèles permettant de résoudre cette problématique. Bien que ces modèles aient été construits en utilisant notamment des données de la Direction Commerce, ils

pourront être appliqués dans d'autres services de l'entreprise.

1.2 Motivation et Objectifs

Les modèles de plongements lexicaux ont révolutionné la compréhension automatique du langage naturel. Cependant, ces modèles ont tendance à manquer de contexte linguistique spécifique à un domaine, ce qui peut entraîner des erreurs dans la compréhension de vocabulaire technique. L'ajout d'informations linguistiques aux modèles de traitement automatique des langues vise à améliorer les performances des modèles en utilisant des connaissances linguistiques supplémentaires pour la représentation des données ou pour la résolution de tâches.

Dans cette thèse, je cherche à évaluer et à améliorer les modèles de plongements lexicaux en ajoutant des informations linguistiques telles que les relations syntaxiques et sémantiques entre les mots. Les objectifs spécifiques de mon travail sont :

1. Analyser les limites de représentation des termes hors-vocabulaires et spécifiques à des domaines avec des modèles de plongements lexicaux.

2. Développer des méthodes pour intégrer des informations linguistiques dans les modèles de langage Transformer.

3. Évaluer les performances des modèles avec et sans informations linguistiques sur des tâches de compréhension automatique de phrase.

4. Explorer les limites de ces méthodes et identifier les situations où l'ajout d'informations linguistiques est le plus bénéfique.

1.3 Contributions

Le déroulement de cette thèse a permis d'aboutir aux contributions suivantes :

1. Nous menons une étude approfondie des stéréotypes de genre présents dans un corpus de pièces de théâtre datant du XVI^{e} au XX^{e} siècle. À notre connaissance, il n'existe pas de travaux similaires dans la littérature scientifique. Nous soulignons les différences remarquables entre les biais d'apprentissage contenus dans les modèles généraux et ceux spécifiques au théâtre. Nos recherches montrent l'importance d'aborder les biais d'apprentissage sociétaux à travers une approche transdisciplinaire (sociologie, littérature, etc.).

2. Nous examinons les problèmes liés à l'utilisation de modèles pré-entraînés sur des données spécifiques à un domaine. Nous soulignons les effets négatifs de l'apprentissage de modèles sur des données générales (Wikipédia, Web, etc.). Plus précisément, nous montrons comment la tokenisation préalable des termes spécifiques au domaine peut affecter leur représentation. Nous parvenons à démontrer que cet impact négatif se produit dans divers domaines spécialisés et pour de nombreux formats de rédaction différents.

3. Nous proposons deux solutions pour mesurer l'effet de la tokenisation sur la représentation des termes spécifiques au domaine. Nous expliquons l'utilité de chacune d'entre-elles et montrons, pour différents modèles de langue, qu'il est plus facile d'améliorer la représentation de nouveaux termes (termes spécifiques au domaine et erreurs orthographiques) que de modifier la représentation de termes existants.

4. Nous avons élaboré une méthode pour intégrer des informations morphosyntaxiques dans les données de domaine spécifique, afin de fournir un contexte supplémentaire aux modèles Transformer. Nous démontrons que cette méthode est plus performante que l'affinage pour projeter les termes hors-vocabulaire dans l'espace de représentation. De plus, cette méthode permet de réduire les effets négatifs de la tokenisation initiale de ces termes sur leur représentation, ce qui rend le contexte des termes environnants plus significatif.

5. Nous présentons une approche générique pour intégrer du contexte structurel et sémantique dans les modèles Transformer. Nous démontrons que cette méthode améliore considérablement les performances de ces modèles sur quatre tâches d'analyse des langues naturelles. Nous examinons en détail l'impact de l'ajout d'informations morphosyntaxiques pour chaque tâche ainsi que celui de la reconnaissance d'entités nommées.

6. Notre objectif global est de développer des solutions d'intelligence artificielle respectueuses de l'environnement. Pour y parvenir, nous avons tenu compte de l'impact écologique de nos activités expérimentales et nous avons cherché à réduire au minimum le nombre d'expériences nécessaires. Nous proposons également une méthode de représentation plus écologique que l'affinage. Nous travaillons donc en faveur d'une intelligence artificielle plus durable et responsable envers notre planète.

Dans l'ensemble, nos contributions visent à traiter des aspects complémentaires de l'évaluation et l'adaptation de modèles Transformer à des domaines de spécialité. Dans nos travaux, nous avons souhaité évaluer la reproductibilité de nos observations

et de nos expériences dans plusieurs contextes (c'est-à-dire dans différents domaines de spécialité) pour le français et pour l'anglais. Nous n'avons pas encore évalué l'applicabilité de nos méthodes à d'autres langues, cependant, nous croyons que les principes sous-jacents sont suffisamment généraux pour être appliqués à d'autres langues qui disposent d'outils d'annotation linguistique et d'informations spécifiques aux corpus. Nous espérons que ce manuscrit parvienne à établir des observations et des conclusions suffisamment robustes pour pouvoir être extrapolées à d'autres contextes. Pour faciliter la reproductibilité de nos travaux, nous rendons disponible en ligne tous les codes associés à nos méthodes proposées et à nos publications.

1.4 Structure du manuscrit

Les travaux de thèse présentés dans ce manuscrit se composent de quatre parties : la première recense les travaux de la littérature similaires à notre champ de recherche, la deuxième présente le matériel et les modèles nécessaires à nos expériences, la troisième détaille l'évaluation des performances des modèles de plongements lexicaux sur des corpus de spécialité et la quatrième présente les méthodes proposées pour ajouter des informations linguistiques à des modèles de langue.

Partie I. État de l'art. *Cette partie détaille l'état de l'art actuel de la littérature en apprentissage automatique dans le domaine du TAL.*

État de l'art. Le Chapitre 2 a pour objectif d'expliquer les concepts et de recenser les travaux liés à notre problématique de recherche. Plus précisément, nous introduirons ce chapitre en distinguant les approches de représentation symbolique (voir Section 2.3) et les approches plus récentes d'apprentissage par transfert (voir Section 2.4). Ensuite, nous détaillerons l'évolution de la représentation des termes hors-vocabulaires des modèles (voir Section 2.5). Nous nous intéresserons ensuite aux méthodes d'évaluation de la représentation des mots (voir Section 2.6), en présentant les avantages et les inconvénients des méthodes d'évaluation extrinsèques et intrinsèques. Enfin, nous montrerons un panorama de méthodes permettant d'ajouter du contexte aux modèles à l'aide de connaissances externes (voir Section 2.7).

Partie II. Matériel et Méthodes. *Cette partie vise à présenter les travaux préparatoires aux différentes expériences réalisées dans ce manuscrit.*

Présentation des corpus. Le Chapitre 3 décrit les tâches d'apprentissage automatique abordées dans le manuscrit et les jeux de données utilisés pour les

traiter. Les corpus sont divisés en deux catégories associées à deux objectifs de recherche : l'évaluation intrinsèque des modèles (analyse qualitative) et l'évaluation extrinsèque des modèles (analyse quantitative). Les corpus sont exploités en deux temps, en fonction du cadre de recherche dans lequel ils sont utilisés, soit l'évaluation intrinsèque des modèles (voir Section 3.2) qui vise à analyser les représentations générées du point de vue qualitatif, soit l'évaluation extrinsèque des modèles (voir Section 3.3) qui compare les performances des représentations sur diverses tâches (l'analyse de sentiments, la détection de paraphrase, l'implication textuelle et la reconnaissance d'entités nommées). Un jeu de données spécifique, le corpus de courriels EDF, privé et anonyme, sera discuté en détail en raison de ses caractéristiques spécifiques (voir Section 3.2.2).

Modèles de plongements lexicaux. Le Chapitre 4 a pour but de décrire les modèles de plongements lexicaux pré-entraînés pertinents pour notre recherche. Dans un premier temps, les corpus issus du domaine général nécessaires à l'entraînement de ces modèles seront présentés (voir Section 4.2). Dans un deuxième temps, nous détaillerons les modèles pré-entraînés appartenant aux familles Word2Vec, ELMo et Transformer (voir Section 4.3).

Partie III. Évaluations des modèles de plongements lexicaux. *Cette partie a pour but d'évaluer les modèles de plongements lexicaux utilisés dans la littérature, tels que les modèles de représentation statiques et contextuels, en mettant en évidence les biais de représentation liés aux données d'apprentissage de ces modèles.*

Les stéréotypes de genre. Le Chapitre 5 porte sur l'analyse des stéréotypes de genre présents dans les modèles de plongements lexicaux statiques de type Word2Vec. Ces modèles seront évalués dans le contexte particulier des pièces de théâtre, qui utilisent des stéréotypes existants à des fins comique ou tragique. La première étude vise à comparer les stéréotypes de genre contenus dans un modèle ayant appris sur les pièces de théâtre à ceux contenus dans des modèles ayant appris sur des corpus issus du domaine général. Ensuite, une étude approfondie sera menée afin de comparer les stéréotypes de genre sociétaux décrits dans la littérature et ceux retrouvés dans les pièces de théâtre. Enfin, une troisième analyse mettra en évidence la puissance de la représentation de concepts tels que les personnages types avec des modèles statiques.

La représentation des termes hors-vocabulaire. Le Chapitre 6 a pour but de montrer les difficultés liées à l'application de modèles pré-entraînés de l'architecture Transformer dans trois domaines différents : l'énergie, la

biologie et le droit. Nous verrons qu'il est difficile de représenter des termes qui ne sont pas dans le vocabulaire de ces domaines. Nos travaux montrent que la représentation de ces termes est principalement due à leur tokenisation préalable et nous proposerons deux méthodes pour évaluer la représentation de ces termes.

Partie IV. Les modèles Transformer appliqués à des données de spécialité. *Cette partie vise à proposer des méthodes d'incorporation de contexte linguistique à des modèles Transformer pour améliorer la représentation des termes hors-vocabulaire.*

Ajout de contexte linguistique. Le Chapitre 7 commence par aborder les problématiques actuelles de l'intelligence artificielle, en particulier en traitement automatique des langues, liées aux enjeux environnementaux des méthodes d'apprentissage automatique. Dans ce chapitre, nous proposons deux méthodes pour incorporer du contexte linguistique, syntaxique et/ou sémantique, dans des modèles Transformer pour améliorer la représentation des données de spécialité (voir Section 7.2) : une méthode qui vise à améliorer la représentation des termes hors-vocabulaire et une autre qui a pour but d'améliorer la performance des modèles sur diverses tâches.

Évaluation et Discussion. Le Chapitre 8 a pour objectif de présenter les résultats obtenus sur les expériences menées, que ce soit du point de vue intrinsèque ou extrinsèque. Tout d'abord, une évaluation intrinsèque est menée (voir Section 8.2) afin de comparer les méthodes proposées avec les modèles Transformer existants. Plus précisément, la représentation des mots sera évaluée à l'échelle globale (la structure du nuage de mots) et locale (les mouvements locaux sur des termes spécifiques). Ensuite, une évaluation extrinsèque sera menée sur quatre tâches (voir Section 8.3) pour comparer l'application des représentations dans différents contextes.

1.5 Publications

Certains des travaux que nous présentons dans ce manuscrit ont fait l'objet de soumissions et/ou de publications dans des conférences de TAL :

- A. Benamar. Segmentation de texte non-supervisée pour la détection de thématiques à l'aide de plongements lexicaux. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*

(RÉCITAL, 22e édition). Volume 3: Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 1–14. ATALA; AFCP, 2020

- A. Benamar, M. Bothua, C. Grouin, and A. Vilnat. Easy-to-use combination of POS and BERT model for domain-specific and misspelled terms. In *NL4IA Workshop Proceedings*, Milan, Italy, Nov. 2021

- A. Benamar, C. Grouin, M. Bothua, and A. Vilnat. Evaluating Tokenizers Impact on OOVs Representation with Transformers Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4193–4204, Marseille, France, June 2022b. European Language Resources Association

- A. Benamar, C. Grouin, M. Bothua, and A. Vilnat. Etude des stéréotypes genrés dans le théâtre français du XVIe au XIXe siècle à travers des plongements lexicaux. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 74–81, Avignon, France, 6 2022a. ATALA

- **[Soumission en cours]** A.Benamar, C. Grouin, and A. Vilnat. Adding Linguistic Features to Pre-trained Transformer Models. 2023

Partie I

Etat de l'art

Introduction de la partie

La première partie de ce travail a pour objectif de présenter les travaux qui nous permettent d'étudier les représentations de domaine associées au traitement du langage naturel et ses impacts dans les problématiques sociétales, problématique poursuivie dans ces travaux de thèse.

Dans un premier temps, nous passons en revue l'état de l'art du domaine en présentant l'évolution des méthodes utilisées. Plus précisément, nous présentons les différentes manières de représenter des documents et nous intéressons ensuite à deux sous-catégories d'apprentissage : l'adaptation de domaine et l'apprentissage séquentiel. Nous détaillons les méthodes d'apprentissage auto-supervisé (*self-supervised*) que nous traiterons dans ce manuscrit qui nécessitent un grand volume de données et qui sont à l'état de l'art pour de nombreuses tâches. Ces méthodes, bien qu'efficaces, ne permettent pas toujours d'obtenir de bonnes performances sur des données de spécialité. Pour pallier cette baisse de performance, différentes stratégies existent pour incorporer des données spécifiques à ces modèles. Enfin, nous présenterons les méthodes d'évaluation de ces représentations, et nous verrons que celles-ci consistent principalement à évaluer les modèles avec des méthodes supervisées des tâches et des jeux de données cibles.

Dans un deuxième temps, nous introduisons la question de l'éthique des modèles, en nous plaçant dans le contexte de la création de corpus. Cette mise en perspective implique d'évaluer deux dimensions des enjeux de société autour des problématiques de TAL : les biais contenus dans les modèles (par exemple, les biais de genre, d'âge ou d'ethnie) et l'impact environnemental de ces modèles.

$* * *$

Sachez le, vous n'avez rien à craindre si vous n'avez rien à cacher.

— Harry Potter et les reliques de la mort (Partie 1)

2

État de l'art

Table des Matières

2.1 Introduction

Ce chapitre vise à exposer les connaissances et les travaux de recherche existant en TAL, dans le domaine de la représentation de termes par les modèles. Ici, nous souhaitons contextualiser la recherche de manière approfondie afin de démontrer l'originalité et la pertinence de notre sujet de thèse. Nous examinerons les travaux précédents réalisés dans le domaine de notre étude, en détaillant les différents concepts et méthodes qui ont été utilisés par d'autres chercheurs pour aborder les mêmes questions ou problèmes. Nous effectuerons également une analyse critique des lacunes et des limites de recherches existantes, afin d'identifier les opportunités pour notre recherche.

2.2 Préambule

La première étape de ce chapitre consiste à définir quelques notions générales permettant d'introduire des concepts de TAL que nous utilisons dans notre recherche.

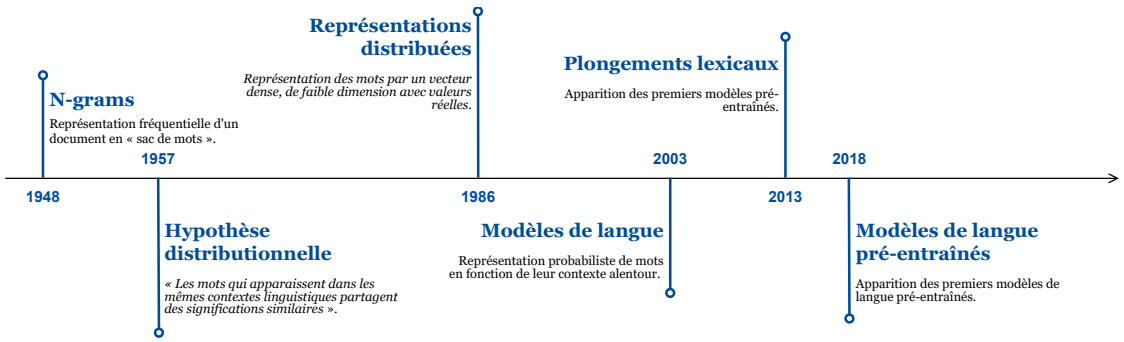


Figure 2.1: Frise chronologique des méthodes de représentation des mots

Beaucoup d'avancées ont été réalisées dans le domaine du TAL de 1948 à nos jours (voir Figure 2.1). Bien que de nombreux travaux consistent à créer des méthodes symboliques, nous présenterons seulement les méthodes plus récentes, plus pertinentes dans le cadre de notre recherche.

En 1948, la représentation en sac de mots apparaît (*n-grams*) et consiste à représenter un document en fonction de la fréquence des mots ou des cooccurrences de mots qui le compose. Cette famille de méthode pose deux principaux problèmes :

1. La fréquence des mots ou des cooccurrences de mots ne prend pas en compte le contexte des mots (l'entourage) et ne capture pas d'information sémantique ou syntaxique.

2. Bien que plusieurs méthodes de pré-traitement permettent de réduire la dimension des matrices (cf. Section 2.3.2), elles demeurent souvent trop grandes pour les algorithmes d'apprentissage automatique. Les matrices de données deviennent trop éparses et difficiles à traiter.

En 1954, l'hypothèse distributionnelle [Harris, 1954, Firth, 1957] démontre que des mots qui apparaissent dans des contextes proches partagent des significations similaires. Cette hypothèse a mené à des travaux sur les représentations distribuées, permettant de représenter des mots par des vecteurs de taille fixe. Les premiers travaux sur ces représentations émergent en 1986, soit trois décennies après l'émergence de l'hypothèse distributionnelle. Grâce à ces nouvelles représentations, il est désormais possible de traiter efficacement des matrices de données plus petites et plus représentatives de la sémantique des mots.

Puis, en 2013, on note l'arrivée de nouvelles représentations qui révolutionnent le traitement des langues : les plongements lexicaux [Mikolov et al., 2013]. Désormais, les mots d'une phrase ne sont plus représentés par des vecteurs projetés dans des dimensions indépendantes, mais par des vecteurs de taille fixe tous projetés dans un espace multi-dimensionnel. Ces modèles ne sont pas contextuels mais ils sont pré-entrainés. On parvient ainsi à transférer la connaissance apprise sur un grand jeu de données à un jeu de données plus petit. Cela permet d'améliorer la robustesse de l'apprentissage sémantique et syntaxique des mots.

Enfin, un dernier évènement majeur du domaine apparaît en 2018 : la construction de modèles de langue pré-entraînés [Peters et al., 2018a, Devlin et al., 2019]. Désormais, l'apprentissage d'informations sémantiques et syntaxiques (avec des marqueurs liés à la structure de la langue) sont apprises par un modèle destiné à être utilisé sur divers corpus et pour des tâches variées, en tenant compte d'informations contextuelles entre les mots. Ces modèles, désormais à l'état de l'art pour de nombreuses tâches, permettent de représenter les mots en fonction de leur contexte et d'adapter leur représentation à chaque occurrence. La polysémie des mots est

ainsi représentée grâce à des vecteurs sémantiques spécifiques.

Notations Nous présentons quelques notations que nous utiliserons dans cette sous-partie. Les variables sont présentées dans la Table 2.1. Un document $d_i \in D$ est composé de mots w_j appartenant à un vocabulaire V, où D est un ensemble de taille n_D et V est un ensemble de taille n_V. Le vocabulaire V est construit lors de la segmentation des documents d_i en *tokens* (ou jetons). Le vocabulaire est constitué de n_V *tokens* et chaque document est représenté comme une suite de *tokens* de taille L $(w_1, w_2, \ldots, w_L)$ avec $w_j \in [1, n_V]$.

Variable	Définition
D	Corpus
n_D	Taille du corpus D
d_i	Un document $d_i \in D$
V	Vocabulaire d'un document D
n_V	Taille du vocabulaire V
w_i	Un *token* $w_i \in \{1, w_L\}$

Table 2.1: Description des variables

2.3 Les méthodes statistiques

2.3.1 Représentation des documents

Matrice document-terme La représentation de documents la plus emblématique est la matrice document-terme (voir Figure 2.2). Cette représentation est également appelée sac de mots et ne conserve pas d'informations sur l'ordre des mots dans un document d_i, mais plutôt sur le contenu lexical de celui-ci (les mots qui le composent). Le document d_i est représenté sous la forme d'un vecteur v_D de taille n_V, où chaque entrée à l'index j correspond au nombre d'occurrences du mot w_j dans le document d_i. Ainsi, D est représenté sous forme d'une matrice X de taille $n_D \times n_V$, où l'entrée x_{ij} correspond au nombre d'occurrences du *token* w_j dans le document d_i. Les vecteurs de représentation peuvent ensuite subir diverses pondérations à des fins de normalisation de la matrice (par exemple, TF-IDF présenté dans le Paragraphe 2.3.1). Nous présentons quelques-unes de ces pondérations dans les paragraphes suivants. Cette représentation pose deux problèmes majeurs. Tout d'abord, chaque mot appartient à une dimension indépendante des autres parmi n_V. Cela signifie que la distance entre tous les termes w_j est la même, quelle que soit la

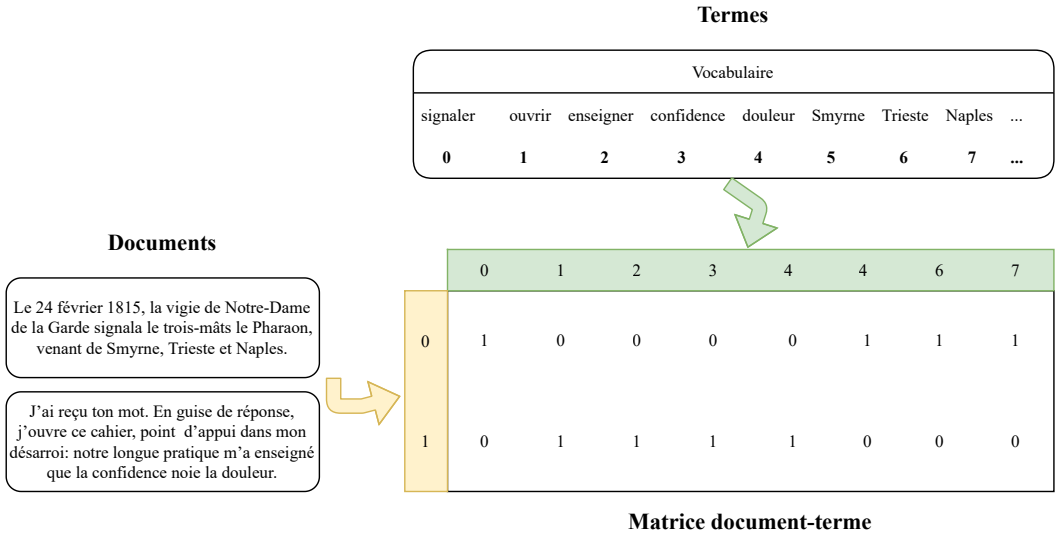


Figure 2.2: Matrice document-terme

proximité sémantique ou syntaxique entre ces mots. Ensuite, la dimension de X augmente de façon proportionnelle à n_V : plus il y a de mots dans le vocabulaire, plus la dimension de la matrice document-terme augmente et devient difficile à traiter. Richard Bellman définit le principe du fléau de la dimension ou de la malédiction de la dimension (*curse of dimensionality*) [Bellman and Kalaba, 1959] dans le cadre de ses travaux sur des problèmes d'optimisation dynamique. Cela signifie que lorsque le nombre de dimensions augmente, les données se retrouvent éparses et peuvent devenir dépourvues de sens. Tout cela a mené aux travaux visant à réduire la dimension des représentations créées.

N-grams Les *n-grams* de mots sont très utilisés en fouille de textes. Ils consistent à représenter un ensemble de mots qui co-occurrent dans une fenêtre de mots N (voir Figure 2.3). Lors du calcul des n-grams, une fenêtre allant de x_i à x_{i+N} est déplacée vers une fenêtre allant de x_{i+1} à x_{i+1+N}, jusqu'à atteindre $x_{i+1+N} = x_n$. Lorsque $n = 1$, on retrouve une configuration de matrice document-terme classique (cf. Paragraphe 2.3.1).

***Term Frequency - Inverse Document Frequency* (TF-IDF) [Jones, 2004]** Cette mesure très populaire a été créée par des linguistes et permet de représenter l'importance d'un mot par rapport à un document dans un corpus D. Elle est utile pour la recherche d'information, afin de distinguer les mots discriminants des mots non discriminants (ceux qui se retrouvent dans une majorité de documents). La

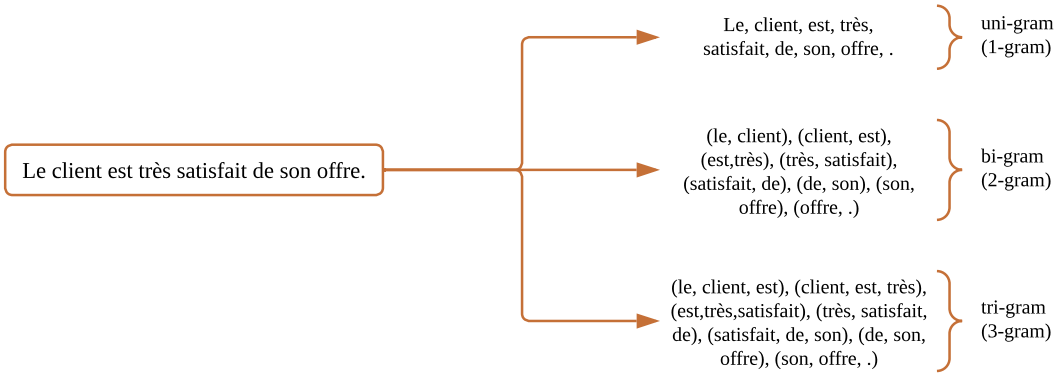


Figure 2.3: Exemple de n-grams obtenus pour $n = 1$ (*unigram*), $n = 2$ (*bigram*) et $n = 3$ (*trigram*)

pondération TF-IDF est caractérisée par le produit entre deux statistiques : la fréquence du terme dans la collection de documents (TF) et l'inverse de la fréquence dans le document (IDF) :

$$TF - IDF(w, d, D) = TF(w, d) \times IDF(w, D) \tag{2.1}$$

La fréquence d'un terme w_j dans un document d_i est normalisée et définie comme :

$$TF(w_j, d_i) = \frac{f_{w_j,d_i}}{\Sigma_{w' \in d} f_{w',d}} \tag{2.2}$$

L'inverse de la fréquence dans le document permet de calculer l'information apportée par un mot w_j dans D. Elle est définie par :

$$IDF(w_j, D) = \log \frac{n_D}{|d \in D \ : \ w_j \in d|} \tag{2.3}$$

où $|d \in D \ : \ w_j \in d|$ est le nombre de documents dans lequel w_j apparaît.

En d'autres termes, un poids $TF - IDF$ élevé signifie que la fréquence d'un terme est grande dans un document donné mais que la fréquence du terme est basse dans le reste du corpus. Cette métrique permet de filtrer plus simplement les mots outils (comme les conjonctions de coordination) ainsi que les mots qui apparaissent fréquemment dans plusieurs documents, et donc qui n'apportent pas beaucoup d'informations.

2.3.2 Méthodes de prétraitement

Comme nous l'avons vu précédemment, un des problèmes majeurs de ces représentations en n-grams est que la taille des matrices document-terme augmente linéairement avec la taille du vocabulaire, ce qui engendre des matrices de représentation très grande sur des corpus volumineux. Pour réduire ces matrices, différents prétraitements peuvent être appliqués. Ils permettent non seulement de pouvoir traiter plus facilement l'information contenue dans ces représentations, mais également de retirer du bruit[1] qui s'avère inutile. Dans cette sous-partie, nous évoquerons trois étapes de prétraitement que nous avons utilisées dans nos expériences.

La suppression de mots vides La suppression de mots vides (ou *stopwords*) est une étape importante dans beaucoup de processus de traitement automatique des langues. L'objectif est de réduire la taille de la matrice document-terme en supprimant des mots qui ne contiennent pas beaucoup de sens ou qui apparaissent avec une fréquence très élevée dans les corpus (comme les mots outils). Cette opération permet d'accélérer les traitements futurs effectués sur la matrice document-terme réduite.

La lemmatisation et la racinisation (*stemming*) La racinisation (ou désuffixation) [Lovins, 1968] est un processus qui consiste à extraire la racine d'un mot. C'est un processus très fréquent en anglais, mais il peut poser question dans des langues morphologiquement plus complexes comme le français, parce qu'il génère des mots qui n'existent pas. La lemmatisation consiste à représenter les mots sous leur forme canonique, c'est-à-dire à la forme que ce mot aurait par défaut dans un dictionnaire (par exemple, l'infinitif pour les verbes ou la forme au masculin singulier pour les noms et adjectifs). Contrairement à la racinisation, la lemmatisation permet de travailler sur de vrais mots et non des mots tronqués.

Désambiguïser par la morphosyntaxe Afin de désambiguïser certains termes dans un corpus, il convient parfois d'ajouter des informations de morphosyntaxe[2]

[1] Ici, le bruit est défini comme tout ensemble de mot qui dégrade les représentations de mots générées. Par exemple, la présence du symbole # dans des tweets peut être assimilé à du bruit dans certaines applications.

[2] *« La morphosyntaxe concerne l'ensemble des structures qui permettent de construire grammaticalement un énoncé. Elle porte aussi bien sur les formes des mots, flexions régulières et irrégulières, variantes irrégulières de certains noms et verbes, l'agencement des marques syntaxiques autour du nom (déterminants, etc.), du verbe (pronoms, etc.), de l'adjectif, de l'adverbe, et enfin de l'organisation des mots et groupes de mots dans un énoncé ou une phrase. »* Parisse [2009]

pour distinguer des homographes[3] en fonction de leur catégorie morphosyntaxique[4]. Guan et al. [2019] a démontré l'importance d'ajouter des informations morphosyntaxiques à des matrices TF-IDF pour la recherche d'information ; cette méthode a permis d'améliorer les résultats de 2 points de F-mesure. Yang [2017] insiste sur l'importance de ces caractéristiques pour des problématiques d'analyse de l'opinion publique.

2.3.3 Résumé

Les représentations de documents dites « classiques », construites à partir de méthodes statistiques ou lexicales, sont toujours d'actualité. Cependant, elles ont été largement remplacées par des méthodes de représentation permettant de construire des matrices de représentation moins grandes. De plus, les méthodes plus récentes projettent les mots dans des espaces multi-dimensionnels, visant à représenter une sémantique plus forte entre les termes en fonction de leur contexte environnant. Bien que les méthodes n-grams intègrent des informations de contexte dans un document (une suite de n mots apparaît x fois dans un document), elles ne sont pas considérées comme étant contextuelles, mais plutôt comme des méthodes fréquentielles (calcul de la fréquence d'un mot dans des documents). Dans la suite de cette revue littéraire, nous décrirons les méthodes de représentation évaluées dans ce manuscrit, apparues depuis les années 2010 : les plongements lexicaux.

2.4 Exploiter les similarités entre des domaines et des tâches

2.4.1 Apprentissage par transfert (*transfer learning)*

En apprentissage supervisé (*supervised learning*), lorsque l'on souhaite apprendre un modèle sur une tâche T^0 et un domaine D^0, on suppose que l'on dispose de données d'apprentissage pour cette tâche et ce domaine. L'hypothèse de cet apprentissage est que ce modèle sera performant sur un autre ensemble de données comportant les mêmes caractéristiques (appartenant au même domaine et pour la même tâche). En d'autres termes, si l'on souhaite utiliser un modèle sur des données d'un autre domaine D^1 ou sur une autre tâche T^1, un nouvel apprentissage devra être effectué sur ces données. Cette approche nécessite donc d'annoter les jeux de données pour chaque nouvelle tâche et chaque nouveau domaine (voir Figure 2.4).

[3] *« Se dit des mots qui ont même orthographe. « Son » (adjectif) et « son » (nom masculin) sont homographes et homophones »* - Dictionnaire le Robert

[4] Cette opération est généralement effectuée en accolant la catégorie morphosyntaxique aux mots à l'aide d'un séparateur (par exemple : Son_ADJ et son_NOM:MASC).

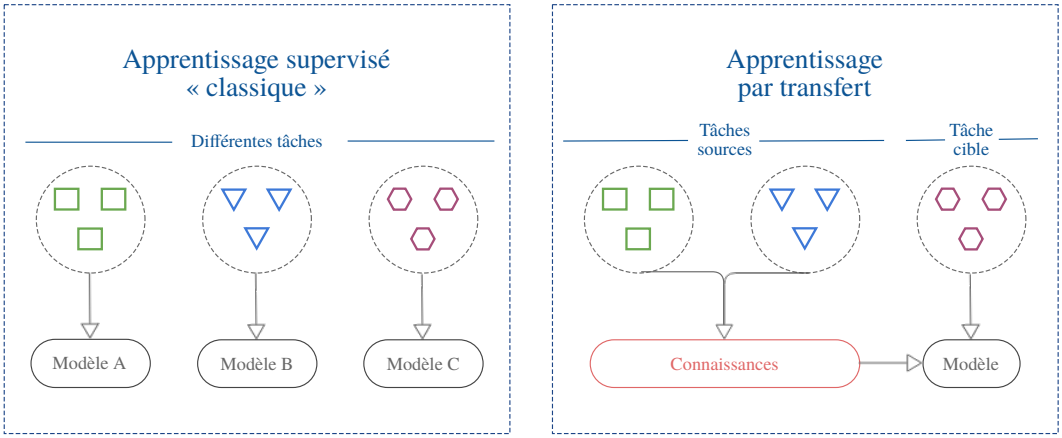


Figure 2.4: Processus en apprentissage supervisé « classique » et en apprentissage par transfert. Inspiré de Pan and Yang [2009]

L'apprentissage supervisé traditionnel s'effondre lorsque nous ne disposons pas de suffisamment de données annotées pour résoudre une tâche [Prakash et al., 2018, Douzas et al., 2022]. L'apprentissage par transfert (*transfer learning*) permet de faire face à ce scénario en exploitant des données sur une tâche ou un domaine connexe. Ainsi, l'objectif de cette approche est de stocker les connaissances acquises lors de la résolution d'une tâche source sur un domaine source afin de les transférer sur une tâche et un domaine cibles (voir Figure 2.4). Ces connaissances sont des caractéristiques intrinsèques au jeu de données d'apprentissage (de nature syntaxique et/ou sémantique) représentées sous plusieurs formes. Dans ce manuscrit, il s'agira de représentations vectorielles, ou plus généralement de représentations construites à partir de réseaux de neurones.

Taxonomie En apprentissage par transfert, plusieurs taxonomies ont été proposées afin de différencier les objectifs variés du transfert de connaissances [Pan and Yang, 2009, Mao, 2020, Liu et al., 2022]. Nous utilisons la taxonomie de Ruder et al. [2019] faisant foi en TAL et inspirée de celle proposée par Pan and Yang [2009], qui n'est pas spécifique à une tâche, et qui définit le principe d'adaptation de domaine crucial dans ce manuscrit. Dans cette taxonomie, l'auteur distingue l'apprentissage par transfert en deux grandes catégories qui aboutissent ensuite à quatre approches principales (voir Figure 2.5). D'une part, le transfert *transductif* représente les situations dans lesquelles les tâches source et cible sont similaires, mais où les domaines source et cible sont différents. Parmi les méthodes transductives, on distingue les méthodes d'adaptation de domaine (*domain adaptation*) – qui traitent

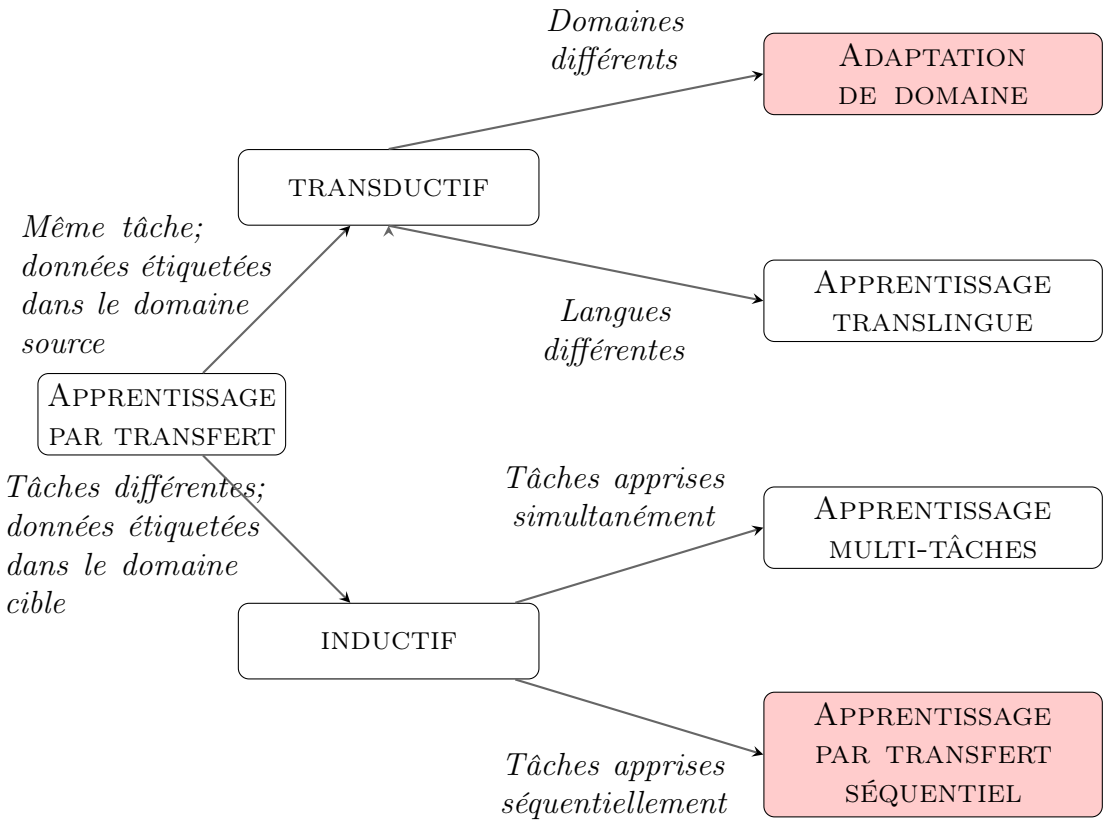


Figure 2.5: Taxonomie de l'apprentissage par transfert proposé par Ruder et al. [2019]. Les approches en rouge correspondent à celles auxquelles nous nous intéressons dans notre recherche

de domaines dans la même langue – des méthodes d'apprentissage translingue (*cross-lingual learning*) – qui définissent les domaines comme des instances de deux langues différentes. D'autre part, le transfert *inductif* se concentre sur les situations dans lesquelles les tâches source et cible sont différentes, sur des domaines source et cible similaires. Parmi les approches inductives, on distingue l'apprentissage multi-tâches (*multi-task learning*) lorsqu'au moins deux tâches sont apprises en parallèle et l'apprentissage par transfert séquentiel (*sequential transfer learning*), lorsque ces tâches sont apprises les unes après les autres.

Nous nous focalisons sur deux approches : l'apprentissage par transfert séquentiel (Section 2.4.2) et l'adaptation de domaine (Section 2.4.3). Nos travaux visent à traiter des modèles de langue qui s'inscrivent dans l'apprentissage auto-supervisé (Section 2.4.2.1) de tâches sur des données, ce qui fait partie intégrante de l'apprentissage séquentiel. De plus, nous nous intéressons aux spécificités des domaines dans des corpus (sous-domaines thématiques, genre littéraire, caractéristiques socio-démographiques, etc.), ce qui nous pousse à détailler les travaux

existants en adaptation de domaine.

2.4.2 Apprentissage par transfert séquentiel

L'apprentissage par transfert séquentiel résume les approches où les tâches source et cible sont différentes et où l'apprentissage est effectué en séquence. Concrètement, cela signifie que les modèles ne sont pas optimisés conjointement comme dans l'apprentissage multi-tâches, mais que chaque tâche est apprise indépendamment. L'objectif est d'améliorer les performances du modèle cible en transférant les informations du modèle source ; il s'agit du principe de transfert de modèles (*model transfer*) [Wang and Zheng, 2015]. De manière générale, l'apprentissage par transfert séquentiel est coûteux lors de l'apprentissage du modèle source, mais permet une adaptation rapide à une tâche cible. Dans le domaine du TAL, il existe des procédures de pré-entraînement purement non-supervisé comme l'allocation de Dirichlet Latente (*Latent Dirichlet Allocation*) [Blei et al., 2003] qui apprend automatiquement des thématiques (*topics*) décrivant les textes d'entrées. Toutefois, la plupart d'entre-elles sont auto-supervisées (cf. Section 2.4.2.1). En pratique, la tâche de pré-entraînement (ou tâche de reconstruction) est souvent une variante des modèles de langue ; soit le modèle source doit prédire le mot suivant une séquence donnée, soit le modèle source doit prédire les mots manquants dits masqués dans une séquence donnée. On peut supposer que l'entraînement de ces tâches encourage le modèle source à acquérir un certain niveau de connaissances linguistiques qui peut ensuite être transféré à des tâches cibles afin d'améliorer les performances. Ce transfert peut être réalisé par l'une des méthodes suivantes : soit en utilisant la sortie du modèle pré-entraîné comme caractéristiques fixes (extraction de caractéristiques), soit en adaptant le modèle pré-entraîné à la tâche cible par le biais d'un entraînement supplémentaire, le plus souvent par le biais d'une architecture plus large et spécifique à la tâche cible qui utilise le modèle comme un composant (affinage du modèle).

Dans ce manuscrit, nous nous focalisons sur l'apprentissage de représentations de *tokens*, et plus particulièrement aux plongements lexicaux statiques et contextuels. Nous expliquons, dans la suite de cette sous-partie, pourquoi l'entraînement et l'application de ces représentations peuvent être considérées comme un exemple de transfert séquentiel. Nous discuterons du lien qui existe entre les approches de transfert séquentiel et l'auto-supervision des modèles et nous verrons qu'il est difficile de distinguer les deux lorsqu'il s'agit d'approches qui visent à traiter les données et non le modèle pour transférer des informations entre deux tâches.

2.4.2.1 Apprentissage auto-supervisé (*self-supervised learning*)

En apprentissage automatique, il existe trois principales familles d'apprentissage : l'apprentissage non supervisé qui utilise uniquement des données non annotées, l'apprentissage supervisé qui utilise des données annotées et l'apprentissage semi-supervisé dans lequel il y a des données annotées et des données non annotées.

Une forme particulière d'apprentissage non supervisé suscite un intérêt croissant pour diverses tâches : l'apprentissage auto-supervisé . Historiquement, les termes « non-supervisé » et « auto-supervisé » étaient interchangeables dans la littérature, mais des travaux récents [Liu et al., 2021, Krishnan et al., 2022] ont préféré le terme d'apprentissage auto-supervisé pour sa spécificité. L'apprentissage auto-supervisé fait référence à toute approche d'apprentissage non-supervisé pouvant être facilement réduite en un problème supervisé en générant des étiquettes. L'apprentissage auto-supervisé nécessite toujours des étiquettes pour résoudre une tâche, mais il est non supervisé dans le sens où ces étiquettes sont dérivées des données elles-mêmes plutôt qu'annotées par des humains[5]. Ce qui caractérise cette approche est qu'au lieu d'entraîner des modèles à l'aide de données étiquetées à la main, l'étiquetage est fait automatiquement.

Les premiers travaux d'apprentissage auto-supervisé avec des réseaux de neurones profonds consistaient à entraîner des réseaux de neurones siamois [Becker and Hinton, 1992, Bromley et al., 1993], des auto-encodeurs empilés [Bengio et al., 2006] et des réseaux de croyance profonds (*deep belief networks*) [Hinton et al., 2006] sans étiquettes. Ces techniques ont vocation à entraîner les réseaux de neurones « une couche à la fois » afin de contourner les erreurs de minima locaux[6] lors de la descente de gradient [Bengio et al., 2007]. Une fois entraîné, le réseau de neurones est affiné (*fine-tuned*), c'est-à-dire que le réseau passe d'un objectif non-supervisé (apprentissage des poids) à un objectif supervisé (apprentissage d'une tâche cible). Cela peut mener à une amélioration des performances sur la tâche cible par rapport aux méthodes d'apprentissage supervisé directes. Depuis les années 2010, ces approches ont perdu de l'intérêt dans la communauté scientifique au profit des méthodes d'apprentissage « de bout en bout », où le réseau de neurones

[5] L'apprentissage auto-supervisé génère des signaux de supervision à partir des données elles-mêmes, en tirant souvent parti de la structure sous-jacente des données.

[6] En apprentissage statistique, l'algorithme de la descente du gradient est utilisé pour minimiser la fonction coût (une fonction convexe). Cet algorithme d'optimisation permet de trouver le minimum de n'importe quelle fonction convexe en convergeant progressivement vers celui-ci. Néanmoins, certaines méthodes se heurtent aux problèmes des minima locaux. En effet, rien ne garantit qu'il n'existe pas d'autre minimum local que celui obtenu, ni que sa valeur soit supérieure à celle du minimum trouvé.

est entraîné entièrement en une seule opération. Cette évolution a été causée (en partie) par la construction de nouvelles architectures [He et al., 2016], des opérations de normalisation [Ioffe and Szegedy, 2015] et de meilleures fonctions d'activation [Nair and Hinton, 2010] pour l'apprentissage de réseaux de neurones très profonds [Bachlechner et al., 2021] tout en évitant les minima locaux.

L'apprentissage auto-supervisé construit une tâche de pré-entraînement afin d'extraire de la connaissance à partir de données non annotées [Jing and Tian, 2020]. Ensuite, le modèle peut être adapté à la tâche cible avec des méthodes d'apprentissage par transfert.

2.4.2.2 Les plongements lexicaux

La représentation vectorielle de mots est une famille de techniques visant à représenter les mots sous forme de vecteurs dans un espace multidimensionnel. Elle vise à rapprocher des mots sémantiquement proches dans l'espace des dimensions. Ces mots sont placés en fonction de leur contexte : la cooccurrence dans l'ensemble du document (architectures en sac de mots) ou le contexte proche (approches séquentielles). Ces vecteurs de mots peuvent être construits par différentes approches ; les deux grandes familles de méthodes sont celles basées sur la factorisation de matrice et celles basées sur les réseaux de neurones. Dans ce manuscrit, nous nous intéresserons seulement à la deuxième.

Le plongement lexical (*word embedding*) est une méthode d'encodage des mots dans une phrase qui vise à représenter les mots d'un texte par des vecteurs de nombres réels, représentés dans un espace multi-dimensionnel fini. Cette méthode a révolutionné le domaine du TAL, principalement pour des tâches de résolution de relations sémantiques et syntaxiques entre des mots. L'avantage de cette représentation est que quelle que soit la taille du vocabulaire, tous les mots sont projetés dans un espace de taille prédéfinie. Autrement dit, la dimension des vecteurs de représentation ne dépend plus du nombre de mots contenus dans le vocabulaire. Cette méthode émerge grâce à la popularité grandissante des réseaux de neurones à cette période, et tout particulièrement à la famille des auto-encodeurs [Liou et al., 2014]. Le rôle d'un auto-encodeur est de réduire une représentation éparse (*sparse*) d'un jeu de données en une représentation compressée, tout en conservant les informations les plus importantes contenues dans les données. L'objectif est ensuite de reconstruire les données d'entrée en sortie du réseau de neurones avec un coût d'apprentissage (*loss*) minimal à partir de sa représentation compressée (voir Figure 2.6). Les données d'entrées sont alors soumises à un large volume d'informations afin que l'encodeur soit capable d'apprendre, par lui-même, la

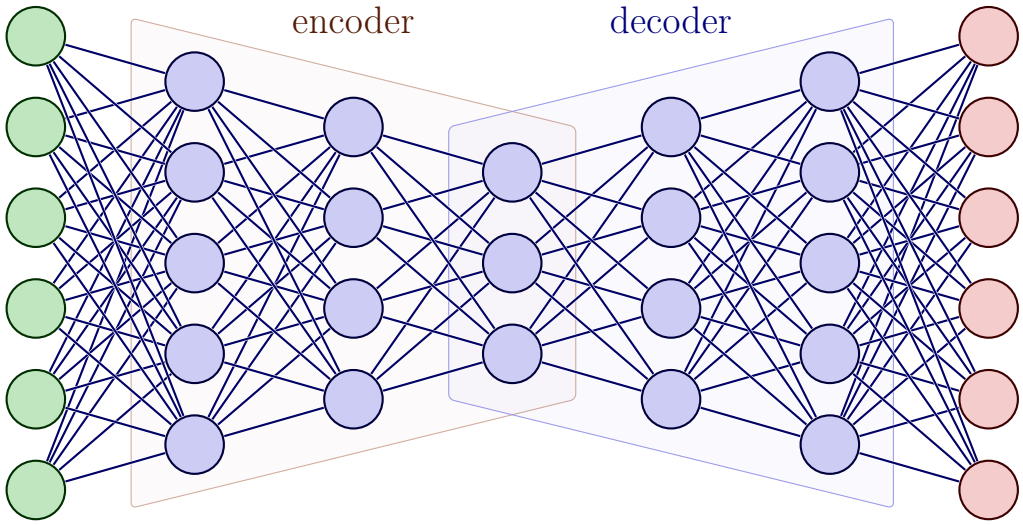


Figure 2.6: Schéma d'un auto-encodeur

représentation latente la plus efficace plutôt que de simplement la mémoriser.

Les plongements lexicaux sont des modèles différents des auto-encodeurs, mais intrinsèquement liés. Comme les auto-encodeurs, ces modèles apprennent des vecteurs dans un espace de données multi-dimensionnel. À partir d'un point qui existe dans un espace vectoriel (l'espace des données), le modèle le place dans un espace vectoriel différent (celui basé sur l'auto-encodeur) ; c'est un processus d'intégration d'espace vectoriel. Les points d'un espace vectoriel sont intégrés dans un espace vectoriel différent afin de représenter de l'information qui n'existait pas dans l'espace de données initial. L'utilisation de l'auto-encodeur est donc intéressante car elle permet de déplacer les mots dans des espaces vectoriels différents. Cependant, elle ne permet pas d'apprendre des informations sémantiques à partir des lettres qui composent un mot. Par exemple, un auto-encodeur rapprochera plus facilement « chat » et « chou » que « chat » et « oiseau », car ils partagent plus de caractères communs. Les plongements lexicaux sont les premiers modèles qui ont permis d'apprendre la sémantique et les relations entre les mots. La force de cette méthode est que le plongement lexical permet de construire des liens entre les relations de mots. Prenons par exemple les mots « Paris » et « France », qui sont reliés par un concept très spécifique qui est la notion de « Capitale ». Cette relation est la même que celle qui lie « Rome » et « Italie ». Cependant, cette relation est très différente de

celle qui existe entre les mots « théâtre » et « cinéma ». Dans l'idéal, les plongements lexicaux seront capables d'exprimer ces relations différemment. Par exemple, dans un espace de données, la distance entre les termes « Rome » et « Italie » devrait être la même (et dans la même direction) que celle qui sépare les termes « Paris » et « France », puisqu'ils sont reliés par la même information. Cependant, la direction entre les termes « théâtre » et « cinéma » devrait être différente. Ces modèles permettent donc, en théorie, de résoudre des analogies très fortes entre les relations [Drozd et al., 2016, Rhouma and Langlais, 2018], de sorte que la représentation des mots soit la plus significative possible[7]. Une fois que ces informations sont correctement retranscrites dans l'espace vectoriel final, les modèles d'apprentissage automatique parviennent à être bien plus performants sur diverses tâches (comme la traduction automatique ou le calcul de similarité). Il existe deux grandes familles de plongements lexicaux : les plongements lexicaux statiques [Mikolov et al., 2013, Pennington et al., 2014] qui attribuent une représentation unique à chaque mot et les plongements lexicaux contextuels [Melamud et al., 2016, Vaswani et al., 2017], qui tiennent compte de la polysémie des mots et attribue une représentation différente à chaque occurrence d'un même mot. Nous verrons que ces deux familles de plongements lexicaux peuvent bénéficier des méthodes d'apprentissage par transfert séquentiel.

Les plongements lexicaux statiques Ces méthodes attribuent une représentation unique à chaque mot, quel que soit son contexte. Il s'agit de méthodes telles que Word2Vec [Mikolov et al., 2013] qui génère des représentations aléatoires puis les adapte afin que les mots qui co-occurrent ensemble aient des vecteurs de mots similaires ; ou GloVe [Pennington et al., 2014] qui atteint un objectif similaire en décomposant essentiellement une matrice de statistiques de cooccurrence. Citons également fastText [Bojanowski et al., 2017] qui améliore Word2Vec en permettant aux mots hors-vocabulaire (*out-of-vocabulary* ou OOV) d'être représentés selon leur orthographe à l'aide de sous-mots (*sub-words*). Cependant, ces représentations statiques sont incapables de représenter correctement les mots polysémiques (qui varient en fonction du contexte). Pour résoudre ce problème, certaines méthodes visent à apprendre des représentations multiples pour chaque mot, soit en fonction d'un nombre prédéfini de sens [Chen et al., 2015], soit en apprenant automatiquement les différents sens d'un mot [Neelakantan et al., 2014]. Néanmoins, ces méthodes ont rapidement été mises de côté au profit de méthodes capables de prendre en compte le contexte pour produire un spectre continu de représentations. Bien

[7] Des travaux ont démontré qu'il existait, dans l'espace multi-dimensionnel des données, des analogies non linéaires plus difficiles à résoudre que l'analogie classique : *king:man :: woman:queen* [Drozd et al., 2016]

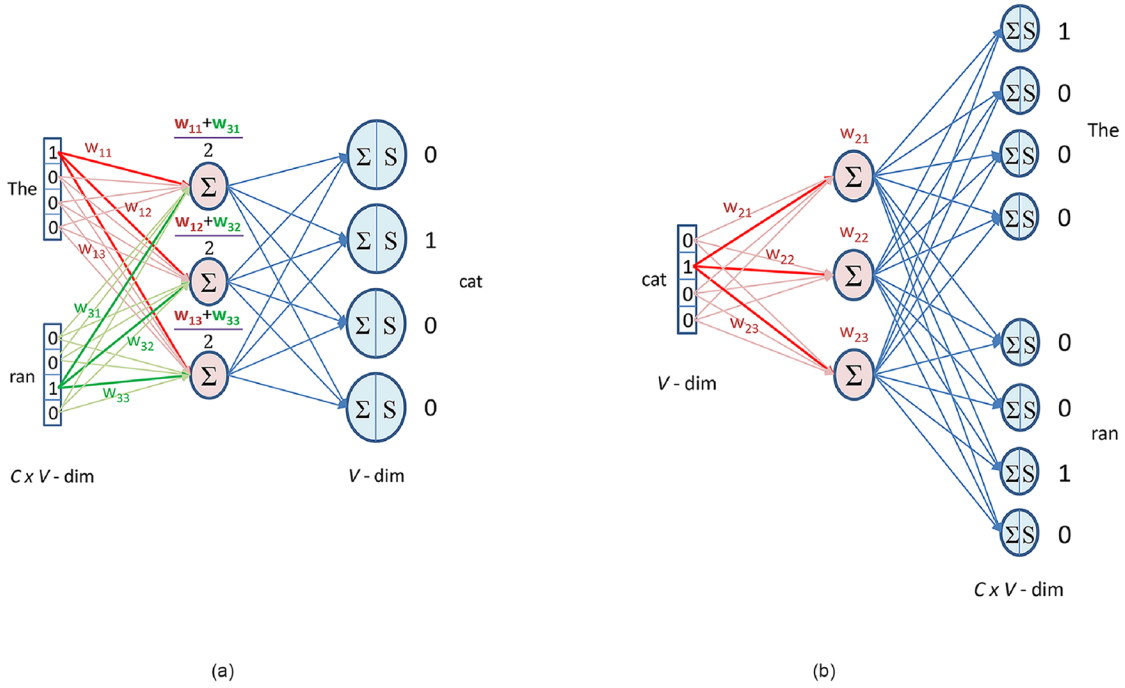


Figure 2.7: Architectures des modèles Word2Vec : (a) CBOW et (b) Skip-Gram [Khan and Chang, 2019]

que ces méthodes n'aient pas été proposées comme des approches contextuelles, elles pré-entraînent des réseaux de neurones récurrents capables de produire des représentations de mots dépendantes du contexte à chaque étape de temps. Ces représentations bénéficient de l'apprentissage par transfert, en entraînant un modèle sur un grand volume de données puis en l'affinant sur un corpus plus petit.

Les plongements contextuels Une représentation du contexte générique de mots a été proposée avec Context2Vec [Melamud et al., 2016] afin de générer des représentations de mots dépendantes du contexte. Leur modèle s'inspire de la version CBOW de Word2Vec (voir Figure 2.7), et remplace la représentation moyennée des mots dans une fenêtre fixe de mots alentours par un puissant réseau de neurones Bi-LSTM. Un large corpus de données [Ferraresi et al., 2008], composé de deux milliards de mots, a été utilisé pour entraîner le réseau de neurones intégrant du contexte à partir d'une phrase et de cibler des mots dans une dimension réduite. Une deuxième méthode a été proposée avec le modèle CoVe (*Contextualized word representations Vectors*) [McCann et al., 2017] qui est basé sur Context2Vec. Ce modèle utilise la traduction automatique pour construire CoVe au lieu de l'approche utilisée pour Word2Vec ou GloVe. Elle consiste à pré-entraîner un Bi-LSTM à deux couches pour effectuer une traduction séquence à séquence, en initialisant les vecteurs avec des représentations GloVe. À la suite de ces travaux utilisant des réseaux de neurones récurrents pour produire des représentations de mots dépendantes d'un contexte,

le modèle *Embeddings from Language Models* (ou ELMo) [Peters et al., 2018a] a émergé. L'entraînement d'ELMo a pour objectif de créer un modèle de langue en utilisant des réseaux de neurones récurrents. Le modèle a été pré-entraîné à l'aide d'un objectif de modélisation de la langue sur un grand corpus de domaine général. Ces représentations ont réussi à être bien transférées grâce à une extraction de caractéristiques afin de réduire les erreurs de prédiction de 6 à 20% sur six tâches[8]. Enfin, les modèles Transformers ont émergé [Vaswani et al., 2017]. Ils se composent de plusieurs couches formant l'encodeur et de plusieurs couches formant le décodeur. Ces modèles ont la spécificité d'utiliser des mécanismes d'attention qui encodent des informations de contexte de mots en utilisant la totalité des mots qui composent la phrase d'entrée. Ces mécanismes visent à appliquer plus ou moins d'attention aux mots alentours au mot pour lequel ils calculent une représentation. De nombreux modèles ont été construits à partir de cette architecture, comme des modèles d'encodage [Devlin et al., 2019, Liu et al., 2019c], des modèles de décodage [Radford et al., 2018] et des modèles d'encodage-décodage [Raffel et al., 2020]. Ces modèles sont à l'état de l'art pour la plupart des tâches, allant de l'analyse de sentiments [Mars, 2022] à la génération de texte [Brown et al., 2020].

La plupart des méthodes présentées ci-dessus effectuent un pré-entraînement auto-supervisé d'un modèle source sur des tâches liées aux modèles de langue. Notons cependant que des modèles pré-entraînés supervisés [Arora et al., 2022] et semi-supervisés [Celebi et al., 2012] ont également émergé. Même si ces approches ont permis des améliorations significatives des performances sur diverses tâches, l'apprentissage de modèles spécifiques à des tâches n'est pas souhaitable. De plus, des travaux ont démontré que les plongements lexicaux statiques, utilisés traditionnellement comme des caractéristiques figées, transfèrent mieux l'information lorsqu'un affinage est réalisé avec l'architecture d'apprentissage de la tâche cible [Vrbančič and Podgorelec, 2020]. Cela a mené à l'emploi généralisé de cette pratique en traitement automatique des langues. Néanmoins, l'affinage de gros modèles pré-entraînés n'est pas toujours pratique et requiert des données annotées et des capacités computationnelles fortes, ce qui constitue un impact environnemental non négligeable [Strubell et al., 2020].

2.4.3 Adaptation de domaine (*domain adaptation*)

Après avoir présenté le transfert séquentiel visant à apprendre des caractéristiques générales pouvant être utilisées pour diverses tâches, nous nous intéressons à l'adaptation de domaine (*domain adaptation*). L'adaptation de domaine consiste à

[8] Question-réponse, inférence (NLI), reconnaissance d'entités nommées (REN), résolution de coréférence, analyse de sentiment et annotation en rôles sémantiques (SRL)

apprendre des caractéristiques spécifiques à un domaine source qui sont pertinentes pour un domaine cible. Par exemple, cette pratique est bénéfique lorsque nous disposons de peu de données pour un domaine cible et de beaucoup de données issues du domaine général, ce que nous traiterons dans ce manuscrit. Dans ce manuscrit, nous ne discutons pas des approches qui visent à traiter différents domaines sources en même temps [Li and Zong, 2008, Yang and Eisenstein, 2015].

2.4.3.1 Définition et taxonomie

La définition d'un domaine est appréhendée de façon très inconsistante en TAL, sans consensus formel [Plank, 2016]. En fouille de textes, le domaine réfère à un ensemble de corpus cohérent (prédéterminé par un corpus général donné) [Plank and Moschitti, 2013]. Cela peut renvoyer à une notion de thématique, de style, de genre, ou de registre linguistique. La notion de domaine a évolué de manière significative ces dernières années, ce qui a mené à des objectifs de recherches variés. Les corpus *Penn Treebank WSJ* [Marcus et al., 1994] et *Brown* [Francis and Kucera, 1979] sont des exemples notables de ces évolutions, le corpus WSJ étant une référence du domaine canonique d'actualités. Ces dernières années, beaucoup de travaux se sont intéressés aux données dites *non canoniques*. Pour cause, la dichotomie entre les données dites canoniques – souvent représentées comme des données très structurées et bien écrites (comme l'actualité) – et les données non-canoniques a émergé avec l'intérêt croissant du traitement des données issues des réseaux sociaux. Le traitement de ces données représente un grand challenge en TAL au vu du bruitage spécifique du texte dans ce domaine [Sankaranarayanan et al., 2009], et des différences de performances notables sont observées (e.g. les données Twitter) [Peddinti and Chintalapoodi, 2011, Mejova and Srinivasan, 2012].

L'objectif principal qui consiste à déterminer l'impact des variations de langue sur les performances des modèles d'apprentissage a mené à des travaux sur l'impact des facteurs humains sur les données ; en d'autres termes, comment des caractéristiques démographiques impactent-elles les performances des modèles d'apprentissage ? [Hovy, 2015]; ou quels sont les effets des stratégies de collecte de données comme le *crowdsourcing*[9] sur la composition du corpus ? [Hsueh et al., 2009] ; ou enfin

[9]Le crowdsourcing a attiré l'attention de la communauté des chercheurs en facilitant l'obtention de données, en particulier de données étiquetées, plus rapidement et à moindre coût [Snow et al., 2008]. Cette nouvelle abondance de données étiquetées a été une aubaine pour l'apprentissage automatique et a conduit à une augmentation de l'utilisation du *crowdsourcing* dans les études. Cependant, il a été constaté que le *crowdsourcing* produit généralement des données plus bruitées que les pratiques traditionnelles d'annotation interne, ce qui a suscité un intérêt significatif pour le développement de mécanismes de contrôle de qualité efficaces afin d'améliorer la qualité des données. Cette pratique fait également débat du point de vue éthique (rétribution des contributeurs, propriété intellectuelle, main d'oeuvre à bas coûts, etc.).

comment des effets de fréquence dans des corpus influencent-ils les performances des modèles d'apprentissage ? [Ellis, 2002].

Dans ce manuscrit, nous choisissons une définition large du domaine, qui consiste à dire qu'un domaine renvoie à une distribution dans l'espace de caractéristiques. Par conséquent, l'adaptation de domaine peut être abordée comme l'utilisation de données sources associées à une distribution pour améliorer les performances d'un modèle sur des données cibles associées à une distribution différente sur une tâche cible (analyse de sentiments sur un domaine général > analyse de sentiments sur le domaine de l'énergie).

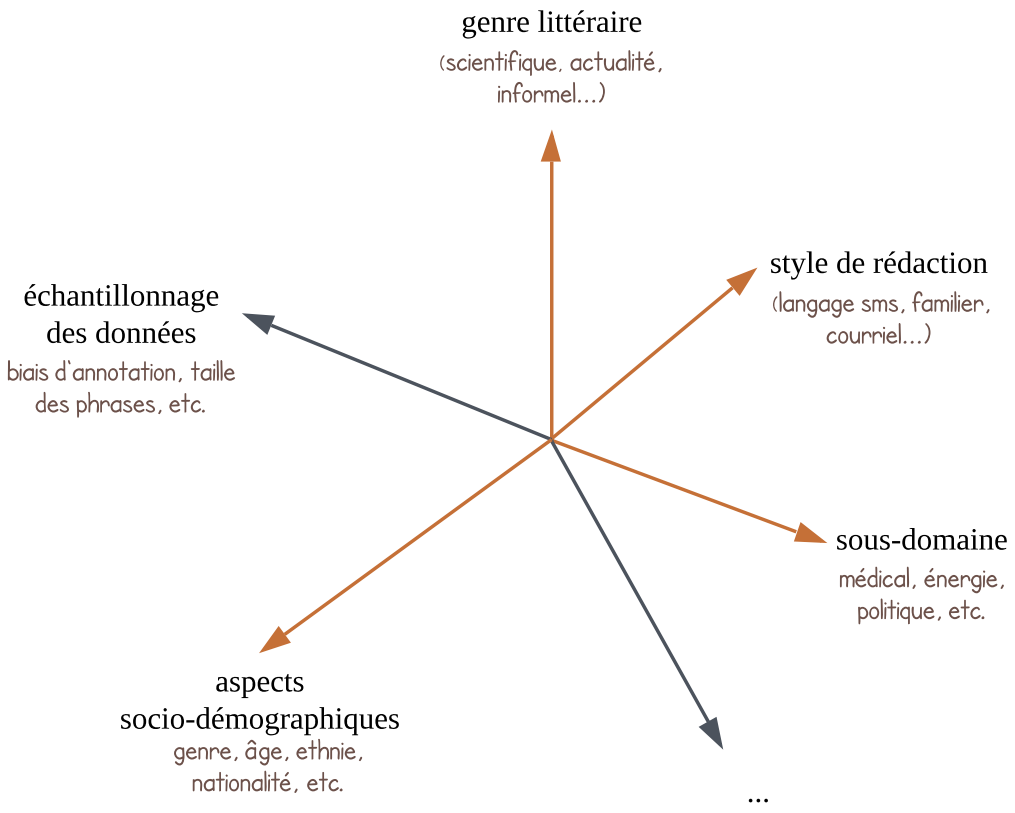


Figure 2.8: Espace de variété défini par Plank [2016]. Les axes colorés en orange correspondent aux dimensions étudiées dans ce manuscrit

Espace de variété En traitement automatique des langues, les données sont particulièrement hétérogènes sur beaucoup de dimensions (qui sont souvent inconnues). Plank [2016] formule une notion théorique d'un espace de variété (*variety space*), défini comme un espace dans lequel un corpus est vu comme un sous-espace de

l'espace de variété. Un corpus est constitué d'exemples tirés d'un espace sous-jacent inconnu à grande dimension, dans lequel les dimensions correspondent à des langages flous et à des aspects d'annotation. Ces caractéristiques latentes peuvent-être liées à plusieurs notions, comme le genre littéraire (par exemple, scientifique, actualité, informel), le sous-domaine (par exemple, médical, finances, politique, etc.) et des aspects socio-démographiques (par exemple, le genre, l'âge), parmi d'autres facteurs inconnus tels que le style de rédaction ou les impacts d'échantillonnage des données (par exemple, taille des phrases, biais d'annotation) (voir Figure 2.8). À la suite de cette définition, [Plank, 2016] suggère de ne plus parler de domaine mais de variété, afin de mieux cibler les différences linguistiques sous-jacentes au corpus et leurs implications. Inévitablement, tous les corpus sont biaisés vers un langage spécialisé et des aspects latents. Comme nous l'avons vu, il existe une multitude de dimensions à prendre en compte dans la composition et l'annotation des corpus, qui sont liées à la notion théorique d'espace des variétés. Elles remettent en question les véritables capacités de généralisation des modèles actuels. Il reste à étudier ce que recouvre la variété, comment les facteurs inconnus influencent les résultats, et à les prendre en compte dans la modélisation et l'évaluation.

Taxonomie Parmi les taxonomies formelles proposées dans la littérature, celle de Ramponi and Plank [2020] distingue plusieurs catégories d'approches d'adaptation de domaines (voir Figure 2.9) : les méthodes basées sur les modèles (*model-centric*), les méthodes basées sur les données (*data-centric*), et les méthodes hybrides. La première famille de méthode cible les approches qui agrandissent l'espace de caractéristiques, changent la fonction de coût, l'architecture ou les paramètres du modèle [Blitzer et al., 2006]. La deuxième famille de méthodes se focalise sur les données et impliquent soit un pseudo-étiquetage (*bootstrapping*) [Zhu and Goldberg, 2009] pour combler l'écart entre les domaines, soit une sélection des données, soit des méthodes de pré-entraînement [Han and Eisenstein, 2019]. Comme certaines approches utilisent des éléments des deux, il existe une troisième catégorie hybride. Nous présentons uniquement les méthodes d'adaptation de domaines basées sur les données, avec un intérêt tout particulier pour les méthodes de pré-entraînement, la préoccupation majeure de ce manuscrit.

2.4.3.2 Sélection de données et pseudo-étiquetage

Sélection de données Les approches d'adaptation de domaine basées sur les données utilisent de grands jeux de données pour construire des systèmes spécifiques à des domaines. Une de ces méthodes est la sélection de données, qui consiste à détecter des échantillons du jeu de données source qui seront pertinentes pour

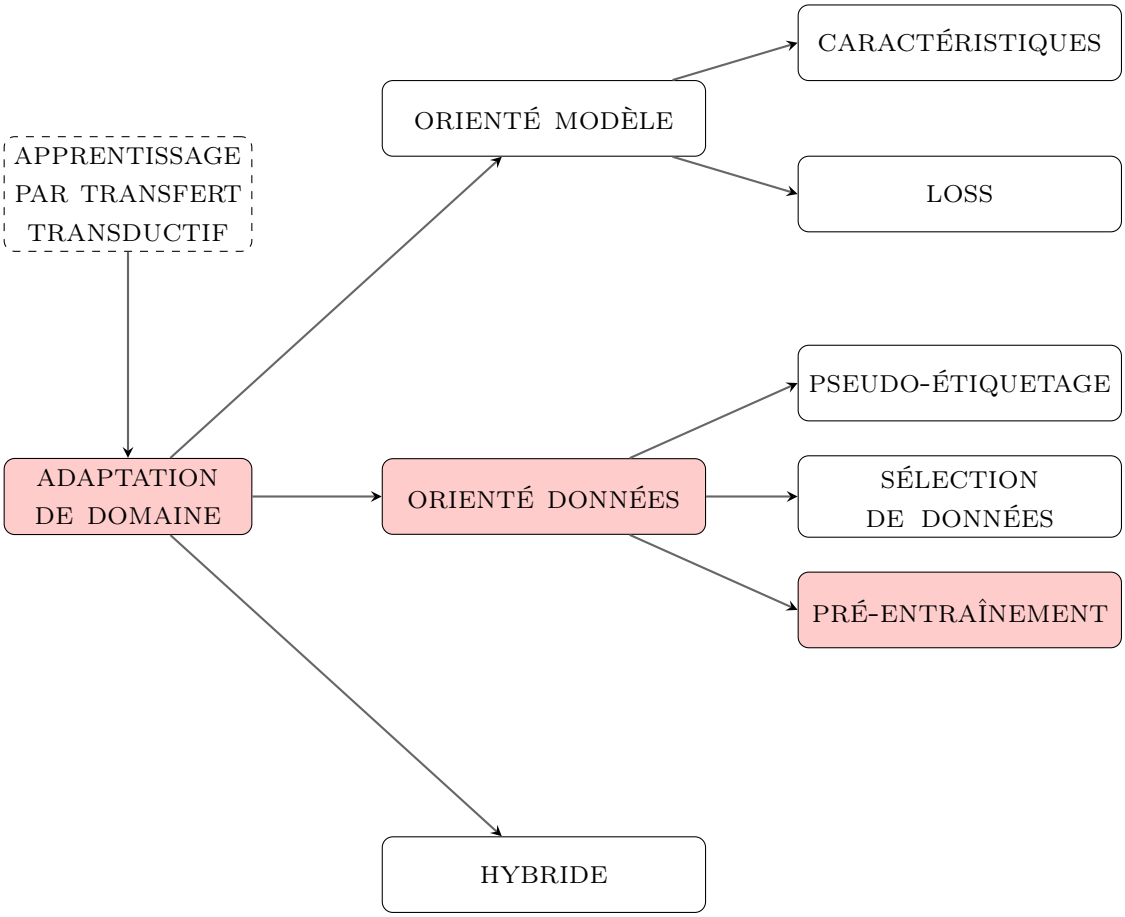


Figure 2.9: Taxonomie de l'adaptation de domaines proposée par Ramponi and Plank [2020]. En rouge, les méthodes sur lesquelles nous travaillons dans notre recherche

l'apprentissage d'un modèle sur le domaine cible [Liu et al., 2019a]. Ces méthodes ont été développées principalement pour la traduction automatique dans l'objectif de sélectionner un sous-ensemble de données parallèles qui correspondent le mieux à la distribution des textes au sein du domaine [Lü et al., 2007, van der Wees et al., 2017] ; ces méthodes ont également été proposées pour d'autres tâches comme l'analyse de sentiments [Xia et al., 2013] ou la détection de la parole (*speech detection*) [Xiong et al., 2020]. En pratique, une mesure de similarité est utilisée afin de sélectionner des exemples à partir de données sources en tant qu'étape de prétraitement avant la construction d'un modèle pour le domaine cible. Usuellement, cette métrique est soit la mesure de similarité de *Jensen-Shannon* [Cover and Thomas, 1991] qui calcule une distance entre les distributions source et cible, soit la mesure de perplexité, où un modèle entraîné sur des données d'un domaine source est utilisé pour calculer un score de perplexité sur les données cibles.

Pseudo-étiquetage Une autre approche basée sur les données est le pseudo-étiquetage qui appartient au domaine de l'apprentissage semi-supervisé. La méthode consiste à adapter des exemples du domaine source pour entraîner un modèle initial, puis à enrichir le modèle construit avec les exemples non annotés pour générer de nouvelles annotations, et ainsi enrichir l'ensemble de données source et réentraîner le modèle de nouveau. Cette boucle d'apprentissage s'arrête lorsque les performances du modèle ne s'améliorent plus. Une des méthodes les plus basiques de pseudo-étiquetage est l'auto-apprentissage (*self-training*) [Scudder, 1965, Yarowsky, 1995, Riloff, 1996], qui consiste à entraîner un seul modèle pour générer des étiquettes sur son propre entraînement futur. Néanmoins, cet apprentissage a tendance à amplifier les erreurs commises par le modèle, surtout lorsque les données annotées sont issues de domaines différents. Une des alternatives à l'auto-entraînement est le co-entraînement (*co-training*) [Blum and Mitchell, 1998, Chen et al., 2011], où deux modèles sont utilisés avec des représentations de caractéristiques indépendantes des données sources. Si ces représentations suffisent à entraîner de bons modèles, alors la génération d'étiquettes et l'entraînement sont dissociés par le fait d'avoir des étiquettes de bonne qualité. Cependant, trouver des représentations de caractéristiques à la fois suffisantes et indépendantes n'est pas toujours facile. Une alternative plus pratique à cette approche est basée sur le *tri-training* [Zhou and Li, 2005]. Dans cette configuration, trois modèles sont entraînés sur différentes versions des données sources[10]. Ensuite, chaque modèle est réentraîné sur des échantillons pour lesquels les deux autres modèles ont un taux de confiance élevé.

La sélection de données et le pseudo-étiquetage sont toujours explorés dans la littérature. Cependant, l'émergence des modèles de langue pré-entraînés a fait du pré-entraînement la méthode la plus utilisée pour l'adaptation de domaine à partir de données volumineuses de domaine général vers des données de spécialité.

2.4.3.3 Pré-entraînement de modèles

Les grands modèles pré-entraînés sont devenus omniprésents en TAL [Peters et al., 2018a, Devlin et al., 2019]. L'affinage d'un modèle basé sur une architecture Transformer avec une petite quantité de données étiquetées atteint souvent des performances à l'état de l'art sur diverses tâches (comme l'analyse de sentiments ou la reconnaissance d'entités nommées) et est devenu une norme *de facto*. Il s'agit de partir des poids du modèle pré-entraîné et d'entraîner une nouvelle couche spécifique à la tâche sur des données supervisées. Une question naturelle qui se pose est celle de l'universalité de ces grands modèles. Le plus gros est-il meilleur ? Et les domaines (ou les variétés) sont-ils toujours pertinents ? Nous revenons sur

[10] Généralement, ces modèles utilisent des techniques de *bootstrapping* pour l'entraînement

ces questions après avoir décrit les stratégies de pré-entraînement :

1. Pré-entraînement seul (par exemple, BERT multilingue ; BERT spécifique à une langue à partir de zéro) ;
2. Pré-entraînement adaptatif multi-phases (*multi-phase adaptive pretraining*) : cela comprend le pré-entraînement, suivi de phases secondaires de pré-entraînement sur des données non étiquetées ou sur des données étiquetées provenant de tâches auxiliaires intermédiaires à ressources supérieures :
   (a) Pré-entraînement multi-phases : deux phases ou plus de pré-entraînement secondaire, allant d'un pré-entraînement à large couverture à un pré-entraînement adapté au domaine ou à la tâche [Beltagy et al., 2019a, Han and Eisenstein, 2019, Lee et al., 2020]. Ils diffèrent par la source des données non étiquetées et des données cibles.
   (b) Pré-entraînement à des tâches auxiliaires : pré-entraînement, suivie de (éventuellement plusieurs étapes de) pré-entraînement à des tâches auxiliaires (par exemple, apprentissage supplémentaire sur des tâches intermédiaires à données étiquetées).

Le pré-entraînement seul (option 1) peut être considéré comme une adaptation simple, analogue au zéro pointé dans l'apprentissage inter-linguistique. L'idée principale est d'entraîner les encodeurs avec des objectifs auto-supervisés comme le modèle de langue (masqué) et des objectifs non supervisés connexes [Devlin et al., 2019, Beltagy et al., 2019b]. À la lumière d'un changement de domaine, un pré-entraînement adaptatif est bénéfique, dans lequel, pour une instanciation, les incorporations contextualisées sont adaptées au texte du domaine cible par la modélisation du langage masqué, comme cela a été introduit.

2.4.4 Résumé

Bien qu'il soit possible d'entraîner des modèles performants pour de nombreuses tâches avec les approches par transfert, certaines problématiques de la vie courante empêchent parfois son bon déroulement.

Tout d'abord, les méthodes d'affinage nécessitent un grand volume de données annotées. Cela implique de disposer d'annotateurs experts lorsqu'il s'agit de corpus spécifiques (par exemple, un corpus médical) et de langues ou de dialectes peu dotés (comme l'annotation du dialecte picard). Même lorsque ces annotateurs sont disponibles, cela reste coûteux en temps et en argent. De plus, les modèles d'affinage ont tendance à augmenter les erreurs d'annotation qui peuvent être présentes dans

le jeu de données d'entraînement pour une tâche cible. Sur des corpus où des annotations peuvent être difficiles à effectuer (comme l'annotation du niveau de paraphrases entre deux séquences avec des étiquettes comprises entre 1 et 5), le niveau d'erreurs sera lourdement impacté par les erreurs d'annotation.

Ensuite, des problématiques matérielles peuvent survenir lorsqu'il s'agit d'entraîner et d'affiner de gros modèles. Bien que des serveurs Cloud existent, cela nécessite une fois encore d'avoir les fonds suffisants. De plus, certaines de ces stratégies sont actuellement impossibles à mettre en œuvre aujourd'hui pour des industriels, étant donné que ces modèles sont principalement construits à des fins de recherche plutôt que d'industrialisation (ces modèles étant parfois trop lourds ou trop lents pour certaines tâches, comme celles en *streaming*). Puis, on se pose naturellement la question de l'éthique liée aux problèmes d'apprentissage. Ces modèles sont très coûteux en ressources énergétiques, et il semble important de se demander si l'apprentissage de gros modèles est toujours pertinent pour le cas d'usage que l'on souhaite traiter.

Enfin, il est aujourd'hui difficile de déterminer la durée de vie d'un modèle de transfert. Que se passe-t-il lorsqu'on travaille sur des données qui évoluent constamment ? À quelle fréquence doit-on réapprendre ces modèles de façon optimale ?

2.5 Les termes hors-vocabulaire : comment représenter des termes inconnus ?

2.5.1 De l'utilisation des mots aux sous-mots

Les documents sont traditionnellement segmentés en phrases et en mots pour des raisons linguistiques et/ou des contraintes techniques. Les unités macroscopiques (les phrases) sont souvent traitées indépendamment les unes des autres et sont elles-mêmes segmentées en unités microscopiques. La définition de ces unités a toujours été une question d'approximation et de compromis. D'une part, ces unités sont bien souvent annotées par des informations linguistiques (comme les parties du discours, la morphosyntaxe ou les dépendances syntaxiques), ce qui exige que la création de ces unités soit motivée linguistiquement. Par ailleurs, un grand nombre de phénomènes font qu'il est très difficile d'identifier et de définir de façon cohérente des unités linguistiques, désignées comme des formes de mots par le *Morphological Annotation Framework* (MAF) ISO standard [Clement et al., 2020]. Ces propriétés incluent les contractions (par exemple, à + le > aux), la dérivation (par exemple,

re + donner > redonner), la composition (tirer + bouchon > tire-bouchon), ainsi que des entités nommées variées ou encore des séquences de motifs spécifiques (par exemple, les URLs).

Par conséquent, les unités typographiques, généralement appelées *tokens* (ou jetons) ont été utilisées pour approximer ces unités linguistiques. Par exemple, la MAF définit un *token* comme une « séquence contiguë non vide de graphèmes ou de phonèmes dans un document ». Dans le cas des écritures qui utilisent l'espace comme séparateur (utilisé universellement avec l'écriture latine), les *tokens* sont largement utilisés et définis comme étant des séquences sans ponctuation ni espaces ou des marqueurs de ponctuation. Notons que certaines décisions arbitraires peuvent être prises concernant les marques de ponctuation, comme par exemple l'apostrophe en français – contenu dans certains mots (comme aujourd'hui) – ou encore le trait d'union. Grâce à cette définition formelle, une phrase peut être découpée de façon déterministe en unités atomiques afin d'obtenir une segmentation en *tokens*, une approximation acceptable des formes de mots. Comme détaillé par Clement et al. [2020], Sagot and Boullier [2008], il n'existe pas de correspondance directe entre les *tokens* et les formes de mots; une forme de mot peut être constituée de plusieurs *tokens* tandis que différentes formes de mots peuvent être représentées par le même *token*. Ce principe est décrit par les directives de l'*Universal Dependencies* (UD) [11] par les termes de « mots multi-*tokens* » et « *tokens* multi-mots » respectivement [More et al., 2018]. Ces deux phénomènes peuvent également interférer (par exemple, à l'image de + le > à l'image du).

Ces dernières années, la multiplication des approches à base de réseaux de neurones a entraîné une évolution dans la manière de découper les *tokens*. En se basant à la fois sur des exigences techniques (par exemple, les modèles Transformer qui nécessitent un vocabulaire de taille fixe) et des résultats scientifiques (par exemple, l'impact de cette segmentation sur les performances de traduction automatique [Sennrich et al., 2016]), la nécessité d'avoir des *tokens* qui approximent des formes de mots linguistiques s'est estompée. Aujourd'hui, la notion de *token* correspond toujours à sa définition MAF, mais elle ne correspond plus à la définition usuelle d'une unité typographique. La « tokenisation » désigne désormais la tâche de segmentation d'une phrase en des unités motivées par des raisons non typographiques et non linguistiques, qui sont souvent plus petites que les *tokens* et formes de mots classiques, et donc souvent appelés « sous-mots » ou « sous-unités ». Les unités typographiques sont désormais appelés pré-*tokens*, et l'ancienne technique de tokenisation est appelée aujourd'hui « pré-tokenisation ». La définition de ce

[11] UD Guidelines : universaldependencies.org

terme est motivée par le fait que les premières approches de la nouvelle tokenisation impliquaient fréquemment la segmentation de phrases en unités typographiques propres (ancienne tokenisation) avant de segmenter à nouveau les unités résultantes (anciens *tokens*) en sous-mots.

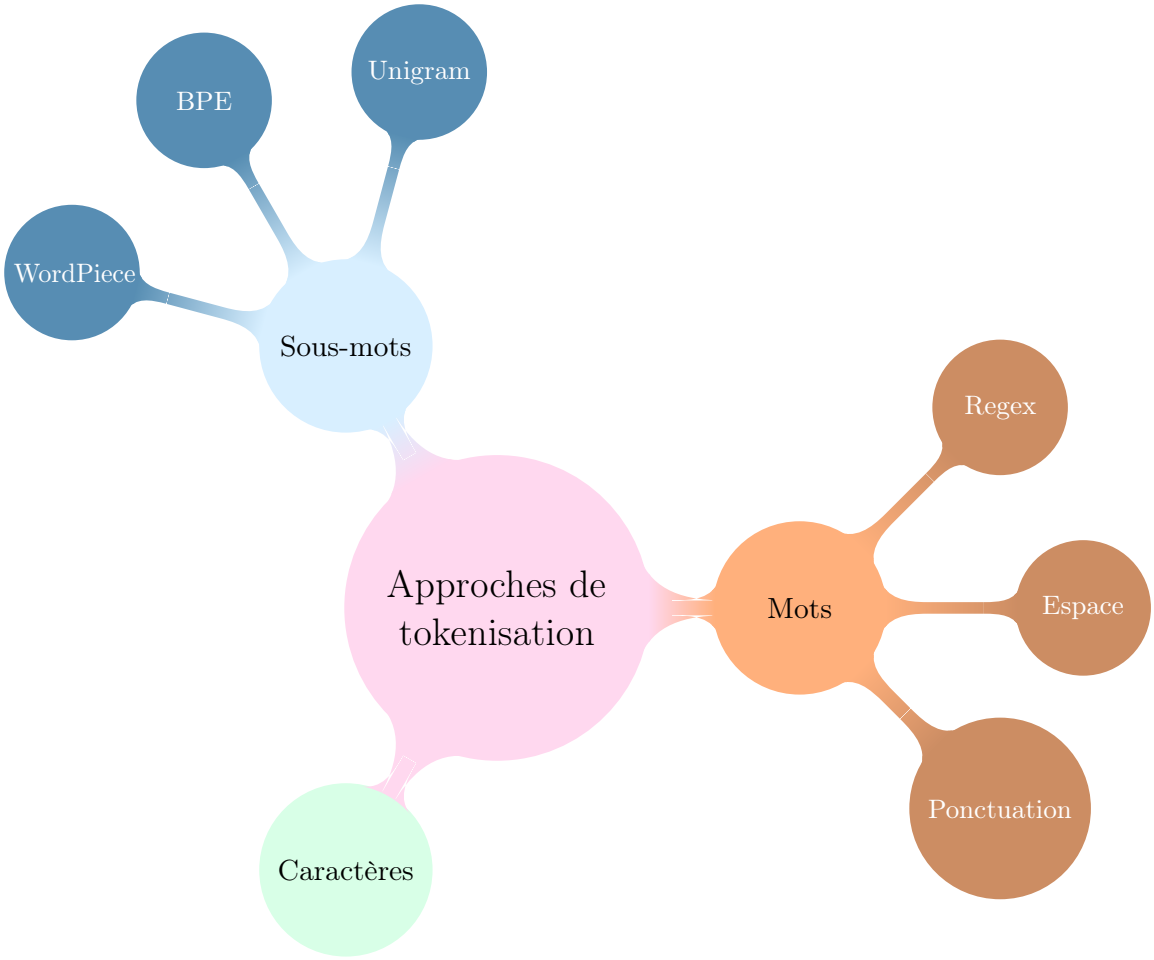


Figure 2.10: Taxonomie des méthodes de tokenisation

2.5.1.1 Algorithmes de tokenisation en sous-mots

Il existe trois principaux algorithmes de construction de vocabulaire de sous-mots, à savoir Byte Pair Encoding (BPE) [Gage, 1994, Sennrich et al., 2016], Unigram [Kudo and Richardson, 2018] et WordPiece dans la construction de vocabulaire de sous-mots.

BPE et WordPiece BPE et WordPiece initialisent le vocabulaire comme un ensemble de caractères, puis fusionnent itérativement une paire de mots dans le

vocabulaire et insèrent la paire fusionnée (c'est-à-dire un nouveau mot) dans le vocabulaire jusqu'à ce que la taille du vocabulaire atteigne une valeur prédéfinie (voir Algorithme 1). La différence entre eux réside dans la méthode de sélection de la paire de *tokens* à chaque itération. BPE remplace itérativement la paire de caractères consécutifs la plus fréquente dans un texte par un caractère unique qui n'apparaît pas dans le texte. En outre, il maintient une table de correspondance pour relier chaque paire remplacée à son caractère correspondant pour l'utilisation du décodage. Dans l'apprentissage du vocabulaire, chaque caractère est traité comme l'unité la plus fondamentale. Il convient de mentionner que l'espace ne peut être fusionné avec aucun autre caractère. Cela signifie que les caractères fusionnés dans toute itération doivent faire partie du même mot et que chaque élément du vocabulaire doit être une sous-chaîne d'un mot. WordPiece est identique à BPE, à l'exception de la sélection de la paire de caractères à chaque itération. WordPiece sélectionne celle qui maximise la vraisemblance d'un modèle de langage donné après la fusion de la paire. BPE et WordPiece construisent le vocabulaire de manière ascendante, en partant du vocabulaire au niveau des caractères et en enrichissant itérativement le vocabulaire par la fusion de deux *tokens* existants.

Algorithm 1: Byte-Pair Encoding [Gage, 1994, Sennrich et al., 2016]

Input : ensemble de chaînes de caractères D, taille du vocabulaire cible k
Output : Vocabulaire V
$V \leftarrow$ tous les caractères uniques $\in D$;
`/* fusion des tokens */`
while $|V| < k$ **do**
  $t_L, t_R \leftarrow$ Bigram le plus fréquent de D; $t_{NEW} \leftarrow t_L + t_R$;
  $V \leftarrow V + t_{NEW}$;
  Remplacer chaque occurence de t_L, t_R de D par t_{NEW};

Unigram À l'inverse, Unigram construit le vocabulaire de manière descendante, en partant d'un ensemble de mots/sous-mots et en enrichissant le vocabulaire en divisant les *tokens* existants au lieu de les fusionner. Plus précisément, il maintient initialement un ensemble de candidats de *tokens* (typiquement, tous les mots du corpus), puis divise itérativement les candidats par un modèle probabiliste et insère les candidats divisés dans la liste de candidats jusqu'à ce que la liste de candidats atteigne une certaine taille (voir Algorithme 2). Plusieurs techniques d'élagage sont également incorporées pour élaguer les candidats et améliorer la construction du vocabulaire. BPE et Unigram sont intégrés dans un outil renommé, à savoir la bibliothèque SentencePiece [Kudo and Richardson, 2018]. Mais la mise en œuvre de WordPiece, développée par Google, n'a pas été publiée en raison de problèmes

commerciaux et n'est pas disponible. Le vocabulaire des sous-mots a d'abord été déployé dans le domaine de la traduction automatique neuronale pour l'apprentissage du vocabulaire [Sennrich et al., 2016], puis a été introduit dans les modèles de langue pré-entraînés. BERT [Devlin et al., 2019] utilise WordPiece comme algorithme de construction de vocabulaire de base et ERNIE-Baidu, ERNIE-Tsinghua, NEZHA, ELECTRA utilisent simplement le vocabulaire de BERT publié par Google pour entraîner leurs modèles. ALBERT [Lan et al., 2019] et XLNET [Yang et al., 2019] utilisent respectivement Unigram et BPE en utilisant la bibliothèque SentencePiece.

Algorithm 2: Unigram LM [Kudo and Richardson, 2018]

Input : ensemble de chaînes de caractères D, taille du vocabulaire cible k
Output : Vocabulaire V et Likelyhood Θ
$V \leftarrow$ tous les caractères uniques $\in D$ et qui occurrent plus d'une fois;
`/* fusion des tokens */`
while $|V| > k$ **do**
  Ajuster l'unigram LM Θ à D;
  for $t \in V$ **do**
    `/* Θ' est le LM sans le token t */`
    $L_t \leftarrow p_\Theta(D) - p_{\Theta'}$; `// Estimation du coût du token`
  Supprimer $min(|V|k, [\alpha|V|])$ des tokens t qui maximise L_t dans V, où $\alpha \in [0, 1]$ est un hyperparamètre ;
Ajuster l'unigram final LM Θ à D ;

2.5.2 La tokenisation sur des domaines de spécialité

Les modèles les plus récents basés sur les réseaux neuronaux, tels que les réseaux de mémoire à long terme (LSTM) Peters et al. [2018a], la convolution attentive Yin and Schütze [2018] et les modèles Transformer Vaswani et al. [2017], sont entraînés et évalués sur des ensembles de données de domaine général. Les modèles Transformer pré-entraînés tels que BERT Devlin et al. [2018] ont prouvé leur efficacité en s'adaptant à de multiples tâches et domaines du langage naturel. Parmi les spécificités de cette architecture, sa capacité à s'adapter aux termes hors vocabulaire (*out-of-vocabulary* ou OOV) est un élément crucial de son succès.

Les modèles pré-entraînés reposent sur un algorithme Unigram ou sur l'algorithme Byte-Pair Encoding (BPE) présentés précédemment, qui divisent un mot en plusieurs sous-mots. L'idée principale de la division des mots en sous-mots est de réduire la taille du vocabulaire en calculant les sous-unités fréquentes dans les OOV. Malheureusement, ces algorithmes de tokenisation sont purement statistiques et entraînent une perte d'information sémantique lorsqu'il s'agit de termes spécifiques à un domaine. Bostrom and Durrett [2020] a montré que le BPE est sous-optimal

pour le pré-entraînement des modèles de langues, car il ne s'aligne pas bien sur la morphologie par rapport à Unigram pour l'anglais et le japonais. Cela suggère que les sous-unités ne contiennent pas d'informations sémantiques ou syntaxiques et que, par conséquent, les OOVs sont mal représentés dans l'espace de représentation.

En outre, les modèles doivent prendre en compte les textes bruités générés par les utilisateurs (par exemple, le contenu des médias sociaux ou des courriels). Les fautes de frappe (par exemple, l'insertion de caractères), les *smileys* et les abréviations sont les ajouts de bruit les plus courants. Selon Park et al. [2016], les OOV peuvent être classés en plusieurs catégories lorsqu'on travaille sur les médias sociaux (par exemple, les mots étrangers, les fautes d'orthographe, l'argot internet). On peut définir trois familles d'OOVs :

- nouveaux termes spécifiques au domaine (par exemple, "protozoon" et "eucaryote" en microbiologie) ; (voir Figure 2.12)
- les mots mal orthographiés contenant des fautes de frappe (par exemple, "infractus" au lieu de "infarctus" en médecine) ; (voir Figure 2.12)
- les homographes inter-domaines (c'est-à-dire des mots qui s'écrivent de la même façon mais qui ont des significations différentes) de mots existant dans la langue générale (par exemple, « bras », soit une partie anatomique dans la langue générale, soit une sous-partie d'une cohorte de patients dans les essais cliniques).

2.5.2.1 Problèmes de robustesse aux termes hors-vocabulaire

Les systèmes de pointe en traitement automatique des langues sont fragiles : de petites perturbations des textes, communément appelées « exemples adverses » (*adversarial examples*), peuvent entraîner des défaillances catastrophiques des modèles [Belinkov and Bisk, 2018, Ebrahimi et al., 2018, Alzantot et al., 2018, Ribeiro et al., 2018]. Par exemple, des fautes de frappe et des substitutions de mots soigneusement choisies ont trompé des systèmes de détection de discours haineux [Hosseini et al., 2017], de traduction automatique [Ebrahimi et al., 2018] et de filtrage de spams [Lee and Ng, 2005], entre autres.

Plusieurs travaux ont étudié la robustesse des modèles Transformer face à des données de spécialité contenant des OOVs. Par exemple, il a été démontré que ces modèles portent une attention déséquilibrée aux fautes de frappes, lorsque celles-ci étaient générées de façon automatique dans un corpus [Sun et al., 2020]. Ces travaux ont également prouvé que BERT n'était pas robuste face à ce bruit dans des tâches de réponse aux questions et d'analyse de sentiments [Sun et al., 2020, Bagla et al.,

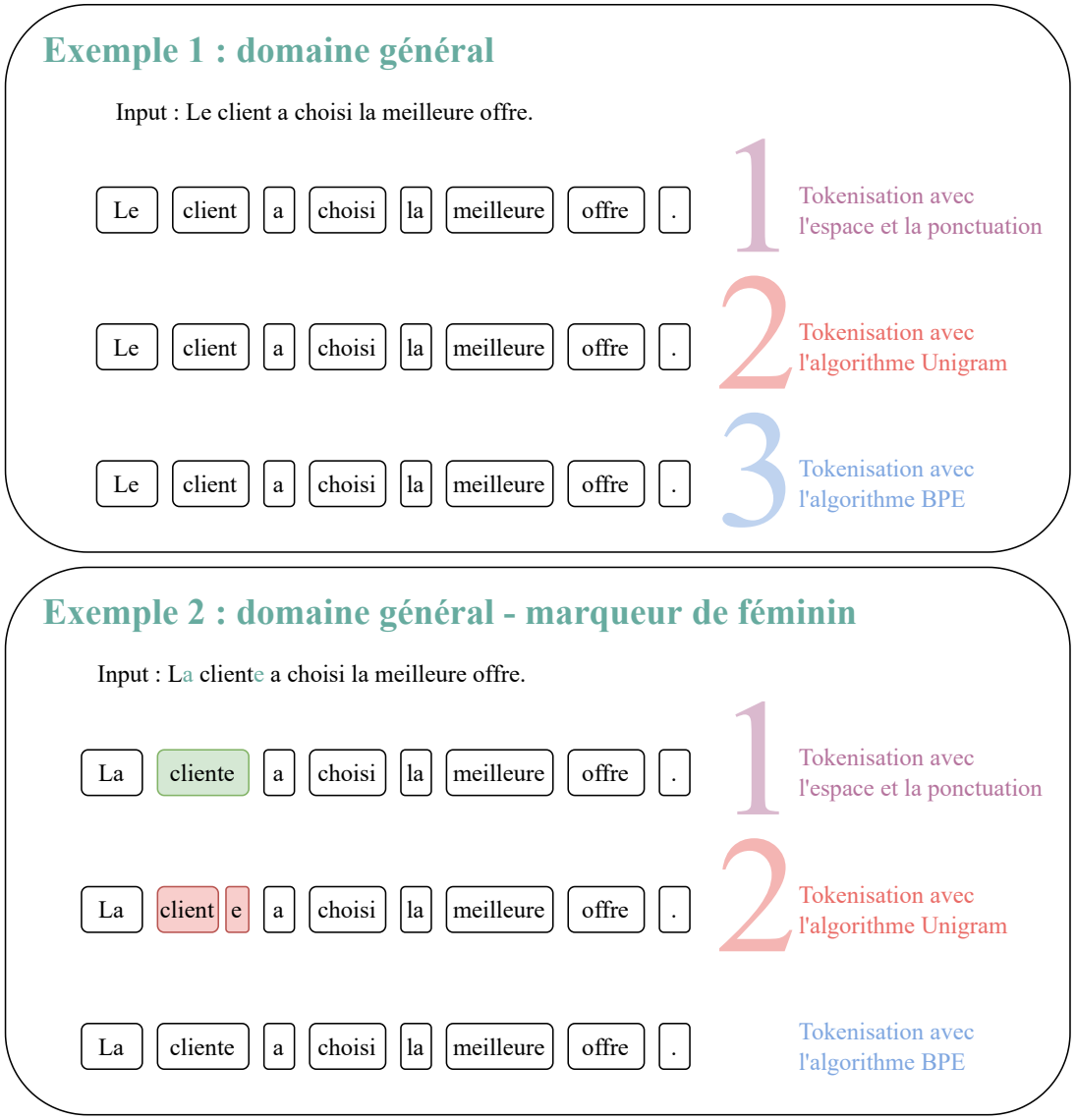


Figure 2.11: Exemple de tokenisation d'une phrase du domaine général dans 2 scénarios

2021]. D'autres tâches comme la similarité textuelle sont également impactées par des erreurs orthographiques [Sun et al., 2020, Bagla et al., 2021].

2.5.3 Résumé

Nous avons présenté différents algorithmes de tokénisation de mots utilisés pour construire un vocabulaire de tokens permettant représenter tous les mots du vocabulaire d'apprentissage. L'avantage de la représentation de mots en tokens est que les termes hors-vocabulaire peuvent être représentés, grâce à leur découpage en tokens connus du modèle. Nous avons présenté la problématique de la représentation de ces termes dans des domaines de spécialité, qui est que les termes ne sont pas

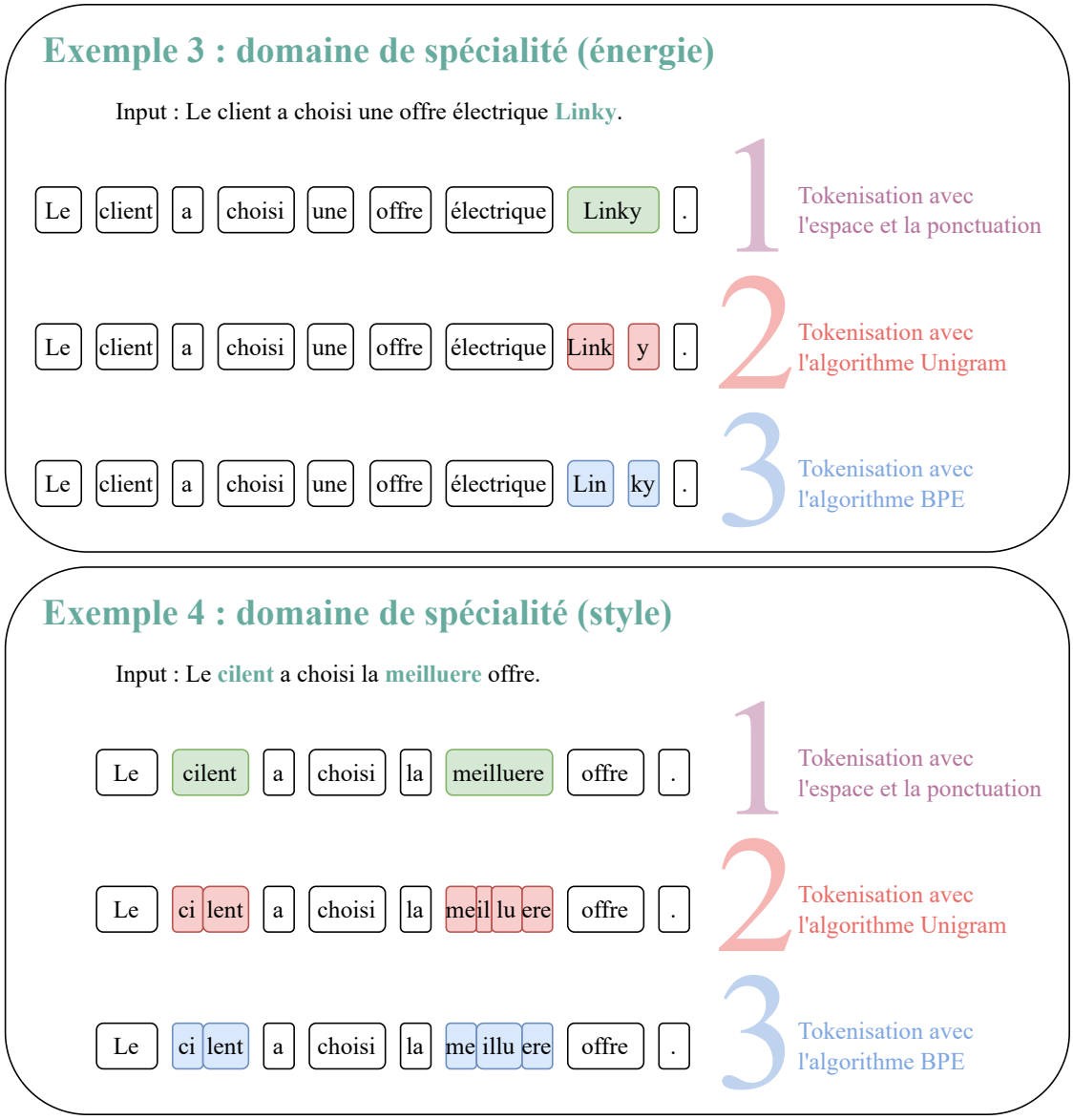


Figure 2.12: Exemple de tokenisation d'une phrase de domaines de spécialité dans 2 scénarios

toujours bien représentés par les modèles, même après l'affinage de ces modèles. Dans ce manuscrit, nous nous intéressons tout particulièrement à comprendre pourquoi ces termes sont mal représentés dans l'espace multi-dimensionnel et nous proposerons des méthodes permettant d'ajuster la représentation de ces termes. Enfin, nous comparerons les algorithmes BPE et Unigram, afin d'analyser l'impact de la tokenisation sur l'apprentissage de modèles pour des données de spécialités sur des tâches cibles.

2.6 Méthodes d'évaluation

Dans cette sous-partie, nous détaillons les principales méthodes d'évaluation des représentations qui existent actuellement. Dans un premier temps (cf. Section 2.6.1), nous distinguons les méthodes d'évaluation extrinsèques et intrinsèques qui existent pour mesurer les informations sémantiques et syntaxiques contenues dans des représentations. Nous verrons que les méthodes extrinsèques sont les plus utilisées mais ne sont pas efficaces pour évaluer les mots dans un grand espace de représentation. Les méthodes intrinsèques sont plus représentatives mais les mesures d'évaluation actuelles ne permettent pas toujours de trouver des corrélations interprétables. Dans un deuxième temps (cf. Section 2.6.2), nous présenterons quelques référentiels d'évaluation existants pour les méthodes extrinsèques. Nous discuterons de la variété de ces référentiels pour l'anglais et de leur adaptation beaucoup moins riche pour le français.

2.6.1 Évaluation des représentations de mots

Diverses méthodes d'évaluation (ou évaluateurs) ont été proposées pour tester la qualité des modèles de plongements lexicaux (contextuels ou non-contextuels). Comme présenté par Bakarov [2018], il existe deux catégories principales de méthodes d'évaluation : les méthodes intrinsèques et extrinsèques.

Evaluateurs extrinsèques Les évaluateurs extrinsèques ont pour objectif d'évaluer diverses représentations sur des tâches supervisées. Ils ont vocation à mesurer les différences de performances entre diverses représentations, sur certaines mesures cibles (par exemple, *accuracy*). Les tâches les plus utilisées pour ces évaluations sont par exemple le marquage morphosyntaxique (*POS tagging*) [Li et al., 2013], la reconnaissance d'entités nommées (*named-entity recognition*) [Xu et al., 2018], l'analyse de sentiments [Ravi and Ravi, 2015], et la traduction automatique [Bahdanau et al., 2015]. L'évaluation extrinsèque pose plusieurs problèmes. Tout d'abord, elle est coûteuse en termes de calcul, car elle nécessite la mise en place de référentiels d'évaluation contraignants (jeux de données volumineux, recherche d'hyperparamètres pour chaque méthode, etc.). De plus, elle nécessite des jeux de données annotés dans une langue, ce qui n'est pas toujours disponible. Enfin, la question de la résolution d'une tâche n'est pas toujours corrélée avec la connaissance sémantique ou syntaxique contenue dans les représentations.

Evaluateurs intrinsèques Les évaluateurs intrinsèques mesurent la qualité d'une représentation indépendamment de tâches spécifiques. Leur objectif est de mesurer

directement les relations syntaxiques ou sémantiques entre les mots. Les scores agrégés sont obtenus en testant les représentations dans des ensembles sélectionnés de termes de requête (*query terms*) et de mots cibles sémantiquement liés. Les évaluateurs intrinsèques peuvent être divisés en deux types : (1) l'évaluation absolue, où les représentations sont évaluées de manière individuelle et seulement leurs scores finaux sont comparés et (2) l'évaluation comparative, où des annotateurs sont interrogés sur leur préférence parmi différentes représentations [Schnabel et al., 2015]. Comme les évaluateurs intrinsèques comparatifs exigent des ressources supplémentaires pour les évaluations subjectives, ils ne sont pas aussi utilisés que les évaluateurs intrinsèques absolus.

Liens entre les évaluateurs Bien que la corrélation entre les évaluateurs intrinsèques et extrinsèques ait été étudiée auparavant [Chiu et al., 2016, Qiu et al., 2018, Bengio et al., 2000], ce sujet n'a jamais été traité de manière approfondie et sérieuse. Par exemple, la production de modèles en modifiant uniquement la taille de la fenêtre n'est pas fréquente dans les applications du monde réel, et la conclusion tirée dans [Chiu et al., 2016] pourrait être biaisée. Le travail de [Qiu et al., 2018] se concentre uniquement sur les caractères chinois avec des expériences limitées. Une étude complète réalisée par Wang et al. [2019b] a été menée pour comparer ces évaluateurs. Ce travail met en évidence la difficulté d'obtenir des corrélations fortes entre les évaluateurs intrinsèques et extrinsèques. De plus, des résultats très différents ont été obtenus sur toutes les tâches supervisées, indiquant que les évaluateurs extrinsèques ne permettent pas de déterminer quelle représentation encode le plus d'informations sémantiques ou syntaxiques. Pour finir, ils expriment des difficultés quant au choix de l'équilibre approprié qui réside entre les deux méthodologies de conception.

2.6.2 Outils d'évaluation

La plupart des travaux consiste à évaluer les plongements lexicaux sur des tâches, afin de déterminer si la sémantique de ces caractéristiques suffit à être performants. L'hypothèse sous-jacente à cette forme d'évaluation peut être formulée de la façon suivante : si les *tokens* sont bien représentés, alors ils sont représentés par des caractéristiques sémantiques et syntaxiques, et permettent donc de résoudre n'importe quelle tâche. Ainsi, un référentiel d'évaluation (*benchmark*) à grande échelle est nécessaire. Le GLUE (*General Language Understanding Evaluation*)[12] [Wang et al., 2018a] est une collection de 9 tâches de compréhension du langage naturel, y compris des tâches de classification de phrases, de classification de phrases

[12] Référentiel GLUE

parallèles, de similarité entre des textes et de réponse à des questions (*question answering*). Ce référentiel d'évaluation a été conçu pour évaluer la robustesse ainsi que la généralisation des modèles, sur des données de domaine général. Cependant, motivé par le fait que les progrès de ces dernières années ont considérablement réduit la marge de manœuvre du référentiel GLUE, un nouveau référentiel d'évaluation appelé SuperGLUE[13] [Wang et al., 2019a] a été présenté. Comparé à GLUE, SuperGLUE comporte des tâches plus difficiles et plus variées (par exemple, la résolution de coréférences). Plus récemment, la même équipe a proposé le nouveau référentiel multi-tâches Adversarial GLUE (AdvGLUE) [Wang et al., 2021] pour explorer et évaluer de manière quantitative et approfondie les vulnérabilités des modèles linguistiques modernes à grande échelle à différents types d'attaques adverses. En particulier, 14 méthodes d'attaques adverses textuelles sont appliquées sur les tâches GLUE pour construire AdvGLUE, qui est ensuite validé par des humains pour obtenir des annotations fiables. En français, cette tâche s'avère plus difficile. Le référentiel FLUE[14], inspiré de GLUE pour le français, contient seulement sept tâches : classement de textes, paraphrase, reconnaissance d'implication textuelles (ou inférence), analyse syntaxique, désambiguïsation lexicale des noms et des verbes (*word sense disambiguation*), et l'étiquetage en parties discours.

2.6.3 Résumé

Dans cette section, nous avons présenté deux formes d'évaluation distinctes de la représentation de termes par des modèles : l'évaluation intrinsèque et l'évaluation extrinsèque. Dans notre recherche, nous souhaitons évaluer les modèles que nous proposons à travers ces deux approches, afin d'évaluer les représentations de mots qualitativement et quantitativement. Nous avons également présenté divers référentiels d'annotation, et introduit GLUE et FLUE que nous utiliserons dans nos travaux. Ces référentiels ont l'avantage de positionner des méthodes par rapport à la littérature et de fournir un socle de reproductibilité commun à la recherche en TAL.

2.7 Ajout de contexte à l'aide de connaissances externes

2.7.1 Introduction

Les méthodes utilisant des plongements lexicaux sont actuellement les plus populaires pour représenter des mots en traitement automatique des langues.

[13] Référentiel SuperGLUE
[14] Référentiel FLUE

Cependant, ces plongements n'encodent pas toujours ce que l'on souhaite. Comme nous l'avons dit, ces plongements lexicaux sont construits d'après l'hypothèse distributionnelle formulée par Harris en 1954 [Harris, 1954]. Elle dit que dans un espace de représentation, les mots qui apparaissent dans des contextes similaires doivent être proches (en distance). Cela signifie que les synonymes et les antonymes d'un mot apparaissent dans le même nuage de mots voisins. Cela permet d'observer qu'avec ces méthodes, on n'encode pas réellement le sens des mots mais leur contexte. Par exemple, les mots « heureux » et « content » seront proches, tout comme les mots « heureux » et « malheureux ». Des travaux se sont intéressés à l'ajout de contexte externe (*external knowledge*) aux plongements lexicaux, soit en utilisant du contexte sur la sémantique entre des mots (par exemple, les synonymes ou les antonymes), soit en utilisant du contexte sémantique spécifique à un domaine (par exemple, un compteur Linky correspond à une offre spécifique). Dans un premier temps, nous détaillerons les approches à base d'ajout de connaissances et d'ontologies. Nous présenterons également les méthodes basées sur les graphes de connaissance. Les approches d'incorporation d'information structurées dans les plongements lexicaux sont les plus populaires et prometteuses actuellement. Dans un premier temps, nous verrons comment il est possible d'incorporer des informations structurées sous forme de graphes. Dans un deuxième temps, nous présenterons les approches qui visent à ajouter des contraintes durant l'apprentissage des plongements lexicaux. Ces méthodes reposent principalement sur l'ajout de contraintes pour forcer la similarité et la dissimilarité entre des mots d'un corpus. Dans un troisième temps, nous focaliserons cette revue sur l'ajout de contexte dans des modèles pré-entraînés.

2.7.2 Bases de connaissances, ontologies et graphes

Aujourd'hui, les bases de connaissances sont utilisées en traitement automatique des langues pour améliorer des modèles d'apprentissage en utilisant des connaissances externes. En pratique, il existe plusieurs modélisations de structures de connaissances : les lexiques, les thésaurus [Hudon, 1997, Bechhofer and Goble, 2001] les graphes de connaissance (*knowledge graphs*) [Schneider, 1973, Ehrlinger and Wöß, 2016] ou encore les ontologies. Nous détaillons brièvement les deux dernières dans cette sous-partie, qui sont celles qui se développent le plus rapidement en TAL.

Graphes de connaissances (*knowledge graphs (KG)*) Afin d'ajouter des données structurées dans des modèles d'apprentissage, il faut construire un graphe de connaissances par étiquette, dans lequel les étiquettes sont représentées par des entités. Par exemple, si l'on souhaite classer des médicaments, on utilisera des graphes de connaissances pour exprimer la sémantique qui existe entre des médicaments, comme la zone de traitement du point de vue anatomique (comme le

ventre ou la tête) ou encore le type d'administration (par exemple, par voie orale).[15] L'exploration des graphes de connaissances est principalement effectuée à partir de graphes existants et open source. Des bases de connaissances du domaine général sont fréquemment utilisées, comme DBPedia [Auer et al., 2007], Freebase [Bollacker et al., 2008]. Il existe également quelques bases de connaissances spécialisées, en particulier pour le domaine médical ou biomédical, comme UMLS [Bodenreider, 2004] pour le domaine biomédical ou encore AMiner [Tang, 2016] pour les sciences générales. Des bases de connaissances plus complexes ont ensuite vues le jour, principalement avec la création de ConceptNet [Speer et al., 2017]. Désormais, on ne cherche plus à construire des liens entre des entités et des relations (par exemple, l'entité Personne est liée à l'entité Ecole par une relation *aEtudieA*), mais on souhaite modéliser des évènements et leurs effets (par exemple, *nettoyer le sol si on fait tomber de la nourriture dessus*), des croyances et des désirs (par exemple, travailler dur pour obtenir un emploi) ou encore des propriétés sur des objets (par exemple, le doliprane est disponible en vente libre). Étant donné que certains graphes de connaissances sont souvent très larges et contiennent des informations non pertinentes pour un jeu de données cible, un graphe de connaissances spécifique à la tâche est souvent construit par l'extraction de connaissances et le traitement de données. Pour extraire les connaissances pertinentes, les données spécifiques à la tâche, telles que les noms de classe, sont mises en correspondance avec les entités du graphe de connaissances, soit en utilisant certaines associations existantes (par exemple, les classes d'ImageNet sont mises en correspondance avec des entités WordNet [Deng et al., 2009]) ou par des méthodes de mise en correspondance telles que l'alignement de chaînes de caractères. Avec le graphe de connaissances construit, les vecteurs sémantiques de classe peuvent alors être appris par une méthode d'intégration de graphe qui peut être soit des variantes du GNN comme le GCN relationnel de [Roy et al., 2020] et GCN d'attention de [Geng et al., 2021], ou des modèles d'intégration basés sur la traduction ou la factorisation, tels que TransE [Bordes et al., 2013] et DistMult [Yang et al., 2015][16].

Ontologies De manière générale, une ontologie est une conceptualisation d'un domaine [Gruber, 1995] qui prend souvent la forme d'une structure arborescente reliant différents concepts par un certain nombre de relations. Cette structure de données est très utilisée dans de nombreuses industries de systèmes experts. Par exemple, de nombreux travaux sont réalisés par l'industrie gazière et pétrolière ;

[15] La représentation de connaissances structurées nécessite une vraie expertise, étant donné qu'elle nécessite de construire des liens complexes entre des caractéristiques. En effet, il ne s'agit pas de représenter des relations « plates » sous forme de liste, mais bien de construire des schémas relationnels multi-dimensionnels.

[16] On invite le lecteur à consulter le *survey* de Wang et al. [2017a] pour plus d'informations sur les *Knowledge Embeddings*.

par exemple, pour des problématiques de gestion de réservoirs [Soma, 2008]. Cette structure de données est également très utile dans le domaine militaire, pour accéder à des informations [Halvorsen and Hansen, 2011], pour des procédures tactiques [Lacy et al., 2005], généralement pour aider la prise de décision et l'accès aux données. Aujourd'hui, l'ontologie WordNet [Miller, 1995] est la plus utilisée [Wang et al., 2018b, Zhu and Iglesias, 2016] car elle inclut des relations sémantiques entre des mots. Notons enfin que des systèmes existent pour gérer des bases de données en ligne, pour le commerce en ligne [Paolucci et al., 2003] ou encore pour des portails gouvernementaux [Haav, 2011].

Les ontologies sont donc très largement utilisées depuis des dizaines d'années, et permettent des résultats fiables. Cependant, ils nécessitent de disposer d'experts afin de créer les concepts et les relations. Le premier défi de cette famille de méthodes est le passage à l'échelle lorsqu'il s'agit de traiter des jeux de données très grands. Le deuxième défi est l'ouverture vers des méthodes non-supervisées. Pour que des systèmes fonctionnent, le peuplement automatique d'ontologies nécessite beaucoup de traitements ultérieurs par des experts.

2.7.3 Enrichir les modèles de langue

Dans cette sous-section, nous présentons les modèles qui utilisent des connaissances spécifiques au domaine dans des modèles de langue. Par exemple, SentiLARE [Ke et al., 2020] utilise la polarité des mots de sentiment de SentiWordNet [Baccianella et al., 2010]. Le modèle SKEP [Tian et al., 2020] incorpore la connaissance des sentiments à partir d'une formation auto-supervisée, y compris la détection des mots de sentiment, la polarité des mots et la paire aspect-sentiment. La connaissance du domaine médical [He et al., 2020] intègre l'ontologie biomédicale du système de langage médical unifié (UMLS) [Bodenreider, 2004] pour faciliter les tâches dans le domaine médical. K-BERT [Liu et al., 2020] exploite également les connaissances d'un concept médical pour une meilleure qualité de reconnaissance d'entités nommées (REN). Enfin, dans le domaine du commerce en ligne, E-BERT [Poerner et al., 2020] utilise le graphe d'association de produits (c'est-à-dire si deux produits sont substituables et complémentaires) [McAuley et al., 2015] construit à partir des statistiques d'achat des consommateurs. Il introduit des tâches supplémentaires pour reconstruire un produit en fonction de ses produits voisins dans le graphe d'association.

Bien que la plupart des approches existantes n'exploitent qu'une seule source de connaissances, il est intéressant de noter que certaines méthodes tentent d'incorporer

plus d'une source de connaissances. Par exemple, K-Adapter [Wang et al., 2020] incorpore des connaissances provenant de sources multiples en apprenant un adaptateur différent pour chaque source de connaissances. Il exploite à la fois la relation de dépendance en tant que connaissance linguistique et la connaissance relation/connaissance de Wikidata.

2.7.4 Résumé

Dans cette section, nous avons présenté différentes approches d'ajout de contexte dans des modèles à l'aide de connaissances externes. Nous avons distingué deux formes de contexte à ajouter dans des représentations : le contexte linguistique (syntaxique ou sémantique, à l'aide de caractéristiques liées à la langue) et le contexte spécifique au domaine (par exemple, à l'aide d'ontologies métiers). Dans notre manuscrit, nous nous intéressons tout particulièrement à l'ajout de contexte linguistique, syntaxique et sémantique, dans des modèles de langue afin d'améliorer la représentation de termes hors-vocabulaires dans des domaines de spécialité.

Conclusion de la partie

Dans cette partie, nous avons passé en revue un certain nombre de méthodes permettant de spécialiser des plongements lexicaux. Plus précisément, nous avons détaillé les différences majeures qui existaient entre les approches par transfert, avant de nous focaliser sur deux d'entre-elles, à savoir : les méthodes d'apprentissage par transfert séquentiel et les méthodes d'adaptation de domaines.

Dans la suite de ce manuscrit, nous travaillerons sur divers aspects de spécialité dans les corpus : les genres littéraires, des caractéristiques socio-démographiques (le genre), le style de rédaction et le sous-domaine de spécialité (comme le domaine de l'énergie).

Nous étudierons l'impact de l'apprentissage des modèles sur des données génériques lorsque l'on souhaite traiter des corpus comportant ces caractéristiques spécialisées. Nous proposerons également une mesure d'évaluation qualitative permettant d'évaluer la représentation des termes hors-vocabulaire dans un corpus.

Ensuite, nous observerons et analyserons les stéréotypes de genre présents dans ces corpus et nous verrons comment ils s'appliquent à de nouvelles données.

Enfin, nous étudierons l'impact environnemental de ces modèles de transfert, afin de construire des modèles moins gourmands et tout aussi robustes que ceux nécessitant une phase d'affinage.

Partie II

Matériel et Méthodes

Introduction de la partie

Cette partie a vocation à présenter les corpus, les tâches et les modèles de plongements lexicaux exploités dans nos travaux. Notre problématique de recherche est liée à l'apprentissage de modèles de plongements lexicaux sur des données issues du domaine général, et à leur application sur des données de spécialité. Pour évaluer et adapter ces modèles à des corpus spécifiques, nous devons quantifier l'impact des données d'apprentissage sur les modèles, que ce soit du point de vue de leur volume, leur richesse lexicale ou encore leur proximité avec les corpus de spécialité.

Pour cela, nous présenterons dans un premier temps les corpus sur lesquels nous avons travaillé dans cette thèse. Le détail des diverses propriétés associées à ces corpus, à l'exemple du domaine de spécialité, du style de rédaction et du format des documents, permettra de souligner la difficulté de traitement de certains corpus par rapport aux autres. Nous reviendrons plus en détail sur le corpus EDF-Courriels contenant des courriels de clients EDF. Ce corpus a un intérêt industriel mais possède également diverses propriétés peu exploitées de façon simultanées dans la littérature.

Dans un deuxième temps, nous dresserons la liste des modèles de représentation de mots par les corpus d'apprentissage utilisés puis par le détail des stratégies d'entraînement. Pour ce faire, nous préciserons la variété des paramètres et la taille des modèles entraînés. Ce chapitre vise à introduire les complexités d'apprentissage de différents modèles et à formuler les premières hypothèses de travail quant à l'application de ces modèles à des données de spécialité, au regard du type de données sur lesquels ces modèles ont appris.

* * *

La vérité, elle est toujours belle et terrible, c'est pourquoi il faut l'aborder avec beaucoup de précautions.

— Harry Potter à l'école des sorciers

3 Corpus

Table des Matières

3.1 Introduction

Dans ce chapitre, nous fournissons des connaissances de base sur les tâches qui seront abordées dans ce manuscrit, et plus particulièrement sur les jeux de données utilisés pour traiter ces tâches. Étant donné la variété des tâches et le grand nombre de jeux de données dont nous nous servons dans ce manuscrit, nous présentons les corpus en fonction des propriétés que nous étudierons dans nos travaux (comme le style de rédaction ou le domaine de spécialité). Tout d'abord, nous exploiterons des corpus sous deux angles d'approche.

Le premier est l'évaluation intrinsèque des modèles (voir Section 3.2), qui vise à analyser qualitativement les représentations générées par des modèles. Dans cette configuration, nous nous intéresserons tout particulièrement aux propriétés linguistiques des corpus cibles qui les différencient des corpus d'apprentissage des modèles ; par exemple, on tiendra compte de son domaine de spécialité (e.g., domaine de l'énergie, domaine médical, etc.), son format spécifique (e.g., courriels, articles, etc.) et du niveau de « propreté » de ces jeux de données. Nous détaillerons le jeu de données de courriels EDF[1] qui contient plusieurs spécificités (par exemple, des termes spécifiques au domaine, des abréviations ou le format de courriels) qui nécessitent d'être détaillées.

Ensuite, nous présenterons les corpus destinés à l'évaluation extrinsèque des modèles (voir Section 3.3), dans le but de comparer des approches de représentation différentes sur des tâches variées : l'analyse de sentiments, la détection de paraphrase, l'implication textuelle et la reconnaissance d'entités nommées.

3.2 Évaluation intrinsèque

3.2.1 Propriétés des corpus

Dans ce manuscrit, nous étudierons, dans différents contextes, la spécificité de certains corpus par leur domaine de spécialité ou leur format de rédaction (qui impactera également leur style de rédaction). Pour cela, nous avons sélectionné cinq corpus (voir Tableau 3.2) ayant des propriétés de style et de domaine différents (voir Figure 3.1). Ces corpus sont utilisés pour des évaluations intrinsèques (voir Section 2.6.1)

1. DEFT-Lois [Azé et al., 2006]: ce corpus a été collecté durant la campagne du DÉfi Fouille de Textes (DEFT) 2006, et se compose d'une collection d'articles de lois. Il contient donc à la fois des termes spécifiques au domaine juridique et un format de rédaction codifié.

2. Bio-Gallica : le jeu de données a été collecté à partir de la bibliothèque numérique GALLICA[2]. GALLICA contient un grand nombre de documents historiques tels que des livres ou des articles de presse et permet le téléchargement de documents océrisés. Nous avons extrait des documents rédigés en français pour le *« Journal de Microbiologie »*, car cette revue contient

[1] Ce corpus a été exploité dans le cadre du contexte CIFRE de cette thèse. Etant donné qu'il est sensible, il demeure privé et anonyme.

[2] https://gallica.bnf.fr/

de nombreux documents. Les documents sont donc très bien rédigés et contiennent deux types de termes hors-vocabulaire de spécialité : des termes mal orthographiés à la suite de l'océrisation et des termes du domaine de la biologie.

3. EASY [Paroubek et al., 2007]: le jeu de données est un extrait de la campagne EASY en ne retenant que la partie d'échanges de courriers électroniques entre des collègues d'une entreprise. Ils sont rédigés de manière formelle et informelle, en fonction du sujet de la conversation et des interlocuteurs. Ce jeu de données est pertinent pour cette étude, car il introduit des termes hors-vocabulaire spécifiques au format du courriel (e.g., « cordialement »). Cela permet d'effectuer l'analyse de ces termes dans des textes du domaine général.

4. EDF-Courriels : ce jeu de données, privé et anonyme, est détaillé dans la Section 3.2.2.

5. Théâtre : le corpus de pièces de théâtre que nous avons constitué[3] se compose de 817 pièces de théâtre dont 554 comédies et 263 tragédies. Elles ont été extraites au format XML et nettoyées « à la main » : suppression des entêtes, du contenu relatif au à l'œuvre plutôt qu'au contenu des œuvres (dates, liste des personnages, actes et scènes) et des noms dans les tours de paroles. Les documents ont ensuite été segmentés en phrases à l'aide d'expressions régulières. Le vocabulaire contient 79 183 mots pour une moyenne de 7900 mots/pièce. Le Tableau 3.1 présente la répartition des pièces par mouvement littéraire.

	Baroque 1580-1660	Classicisme 1660-1690	Lumières 1690-1789	Romantisme 1789-1860	Symbolisme 1850-1900
#Docs	144	147	404	77	45
#Mots	33 874	29 691	46 986	21 760	17 658
Moyenne #caractères/doc.	38 539	36120	35 282	34 269	37 908
Moyenne #mots/doc.	9 448	8 857	8 670	8 476	9 265

Table 3.1: Répartition des pièces de théâtre du corpus dans les mouvements littéraires

Dans nos travaux, nous nous intéressons à la représentation des mots avec des modèles de plongements lexicaux. Pour cela, nous évaluons les modèles sur plusieurs tâches et sur différents aspects (e.g., stéréotypes de genre). Nous nous intéressons tout particulièrement au comportement de ces modèles sur des jeux

[3] http://www.theatre-classique.fr/pages/programmes/PageEdition.php, licence CREATIVE COMMONS BY-NC-ND.

de données de spécialité. Nous présentons quelques propriétés pertinentes pour notre étude de façon détaillée (voir Tableau 3.2).

	Registre	**Format**	**Domaine**	**Langue**
EASY	Courant	Courriel	Général	Français
EDF-Courriels	Formel	Courriel	Energie	Français
Théâtre	Formel	Pièces	Général	Français
DEFT-Lois	Formel	Articles	Juridique	Français
Med-Gallica	Formel	Journal	Médical	Français

Table 3.2: Caractéristiques linguistiques des corpus

Plus précisément, nous sélectionnons les aspects pertinents pour notre étude parmi ceux proposés dans la nomenclature présentée dans le Chapitre 2, c'est-à-dire :

- Le registre de langue (ou niveau de langue) associé au corpus ; Selon le vocabulaire utilisé et la construction de la phrase, le registre peut être soutenu, courant ou familier.
- Le format de rédaction ; on distingue différents formats qui répondent à des règles de rédaction pouvant être très strictes et qui relèvent d'un champ lexical spécifique. Par exemple, le courrier électronique est un format qui nécessite l'utilisation d'un champ lexical spécifique et une organisation codifiée du document.
- Le domaine de spécialité ; ici, il s'agit du domaine général ou spécifique associé au corpus. Le domaine est représenté par un vocabulaire spécifique au domaine, qui n'existe pas dans la langue générale. Par exemple, les noms de virus appartiennent au domaine biomédical.
- La langue de rédaction ; dans nos travaux, nous nous intéressons uniquement à des corpus monolingues. Nos expériences seront principalement réalisées sur des corpus en français, mais des corpus en anglais seront également traités.
- La tâche à résoudre ; dans notre recherche, nous utilisons les jeux de données pour une application spécifique parmi quatre tâches principales, détaillées dans la Section 3.3.2.

3.2.2 Le corpus d'EDF

Le corpus de courriels EDF, privé et anonyme, appartient à l'entreprise Électricité de France et rassemble 99.993 courriers électroniques rédigés par des clients, reçus

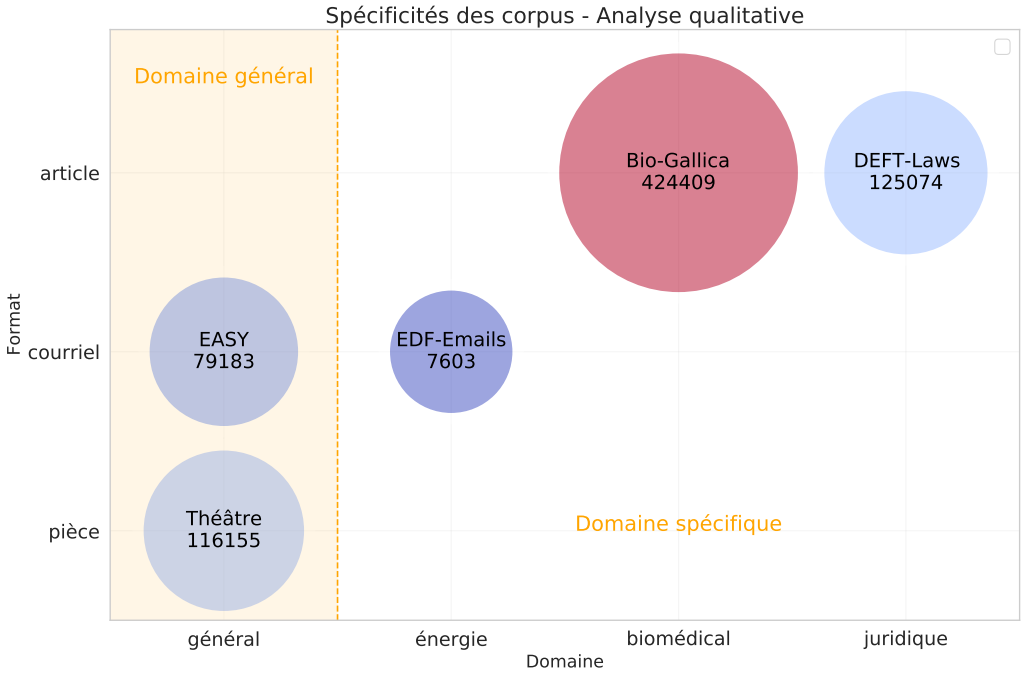


Figure 3.1: Spécificités des corpus d'analyse qualitative

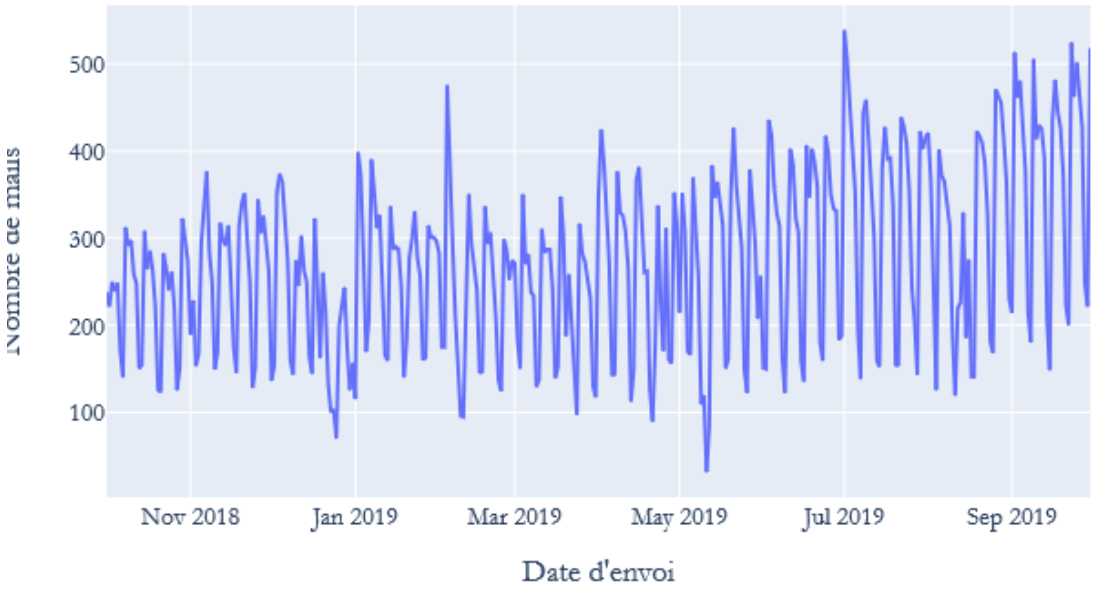


Figure 3.2: Répartition des courriels reçus chez EDF entre Octobre 2018 et Octobre 2019

entre octobre 2018 et octobre 2019. Afin d'être traités le plus efficacement possible, ces courriels sont classés automatiquement dans des catégories définies au préalable

par des experts métiers, puis traitées par l'outil Cameli@ [Dubuisson Duplessis et al., 2020]. Ce corpus contient 13 catégories et concerne des problématiques comme les contrats, les coupures d'électricité, les impayés ou encore l'accès à la plateforme numérique d'EDF. Ce corpus est multi-classes et 26% des courriels (soit 26.215 documents) sont annotées avec plus de deux classes. De plus, 241 courriels sont inexploitables, c'est-à-dire qu'ils n'appartiennent à aucune classe et ne peuvent pas être catégorisés.

Nous observons qu'en moyenne, le nombre de courriels envoyés à EDF reste constant toute l'année (voir Figure 3.2), avec une variance plus importante en 2019. Ce jeu de données est difficile à traiter, car il contient des courriels avec différents niveaux de formalité, mais également des erreurs orthographiques et syntaxiques. De plus, il contient un vocabulaire spécifique à l'énergie tel que des mots existants en français ou des mots n'appartenant pas au domaine spécifique. Le Tableau 3.3 contient plusieurs exemples de fautes d'orthographe, de langage SMS et de termes du domaine qui apparaissent dans le corpus.

Courriel
Bonjour je suis @firstname @name je envoye un message pour ve dire cset possible pour peyer la facture peu à peu pasque je pas bouceaoup l' argent .. S' il vous plait . Merci
Bonjour , Nous souhaitons être informés et bénéficiés de votre offre Mes jours Zen et mes jours Zen plus . Dans l' attente de votre retour par téléphone Cordialement @name
Bonjour Je voulais savoir comment cela se passe comme je vous ai fait parvenir un chèque énergie de @montant ??? ... Cordialement ☺☺

Table 3.3: Exemples de courriels dans le jeu de données d'EDF. bleu: étiquettes de désidentification; rouge: erreurs de syntaxe; violet: expressions spécifiques au domaine ; orange: émojis

Processus de désidentification. Les courriels ont été préalablement désidentifiés au sein d'EDF, conformément aux règlementations du RGPD. L'anonymisation a été réalisée avec des règles linguistiques combinées à l'apprentissage d'un réseau de neurones. Les entités nommées qui ont été désidentifiées sont décrites dans la Tableau 3.4. Lors de la désidentification, les entités nommées ne sont pas supprimées, mais elles sont remplacées par des étiquettes spécifiques à l'entité nommée. Dans ce manuscrit, nous n'avons pas quantifié l'impact de cette procédure sur les performances de nos modèles. Cependant, nous émettons l'hypothèse que les

évaluations menées dans nos études ne sont pas trop affectées par la désidentification, étant donné que nous n'utilisons pas les étiquettes.

Entité nommée	Étiquette
Nom	@name
Prénom	@firstname
Adresse	@address
Numéro de téléphone	@tel
Code postal	@postalcode
Adresse mail	@email
URL	@url
IBAN	@iban
RIB	@rib
Date	@date
Quantité	@amount
Numéro Compte Client	@numclientaccount
Numéro de facture	@numinvoice
Numéro de contrat	@numcontract
Numéro Compte Client	@numclientaccount
Numéro PDL[4]	@pdl

Table 3.4: Étiquettes de désidentification

3.2.2.1 Prétraitement du Corpus

Les formulaires. Les courriels de clients EDF peuvent être regroupés en deux catégories : ceux qui sont rédigés sous format libre et ceux qui contiennent des formulaires remplis par les clients (qui peuvent être suivis d'un autre courriel du client, un courriel automatique ou une réponse d'un conseiller EDF). Le jeu de données contient 18.412 courriels contenant des formulaires (voir Figure 3.3), ce qui représente 18% des courriels client. Les formulaires sont traités de la même manière que les courriels rédigés. Ces formulaires ont pour objectif de souscrire à une nouvelle offre d'électricité ou de gaz (pour les nouveaux clients EDF ou pour des clients qui souhaitent souscrire à une nouvelle offre) ou de modifier une souscription existante. Tous les formulaires contenus dans la collection de courriels contiennent au moins un champ non rempli, identifié avec l'étiquette « Non_Remplis ». Nous avons utilisé ce mot-clé pour supprimer tous les mails contenant des formulaires, afin de traiter uniquement les courriels rédigés librement par les clients.

[4]Le numéro PDL correspond à l'identifiant unique du compteur électrique, qui permet de faire le lien entre le compteur et le client.

Exemple de formulaire

Votre demande de souscription à l' offre en ligne Code INSEE : @address
GAZ : Bloc Offre Elec
ID offre : Non_remplis
Nom offre : Non_remplis
ID option : Non_remplis
Piscine chauffee : Pas de piscine
Véhicule électrique : non
Modèle véhicule électrique : Non_remplis
Lave vaisselle : non Sèche linge : non Climatisation : non
Plaque de cuisson : INDUCTION
Chauffage Alternatif : ELECTRICITE
Congélateur indépendant : 0
Type du compteur : EMC
Etat contractuel du PDL : QETGC

Figure 3.3: Exemple d'un courriel issu du corpus EDF-Courriels contenant un formulaire rempli par un client

Historique de conversation. Certains courriels du corpus possèdent un historique de discussion (voir Figure 3.4), soit parce que le client a eu plusieurs échanges avec des conseillers EDF, soit parce qu'il a envoyé plusieurs courriels d'affilée. Les historiques sont traités automatiquement par EDF, nous les conservons donc pour nos travaux.

Contenus automatiques. Dans la plupart des courriels qui sont sous format libre, la réponse des conseillers EDF est suivie d'un mail automatique concernant la régulation du RGPD ou la satisfaction des clients (voir Figure 3.4). Les courriels automatiques sont rédigés en anglais ou en français. Dans le cadre de nos recherches, nous considérons que ces ajouts automatiques bruitent le corpus, et nous décidons donc de les supprimer.

Détection de la langue. Les courriers électroniques envoyés par les clients EDF ne sont pas toujours rédigés en français. Ces courriels peuvent être rédigés dans une langue étrangère pour quatre principales raisons :

1. Le client rédige dans sa langue natale, différente du français.
2. Le client rédige en français et utilise des anglicismes
3. Le client rédige dans une langue étrangère et la réponse du conseiller est écrite en français.

Exemple de courriel

4 nov . 2018 à 15 : 09 , @firstname a écrit : vous eu mon message merci de me répondre . Cordialement @firstname
4 nov . 2018 à 13 : 33 , @firstname a écrit : i . Bonjour , Nous avons le plaisir de vous confirmer en date du 06 / 10 / 2018 l' utilisation de votre chèque énergie auprès du fournisseur ELECTRICITE DE FRANCE avec les références @numclientaccount et @bp de votre contrat . Le Ministère de la Transition écologique et solidaire Pour toutes informations complémentaires , veuillez contacter l' assistance utilisateur par téléphone ou par courriel en utilisant le formulaire disponible ici .
Madame , Monsieur Je viens a vous pour vous demander de me déduire le chèque que j' ai déclaré en ligne de 128e , sur ma facture de décembre car j' ai vu que vous pouvez le faire et qu' il fallait vous demander , je vous es remis en haut le reçu que sa été fais . en attendant votre réponse , Cordialement Conformément à la réglementation en matière de données personnelles , vous disposez d' un droit d' accès , de rectification , d' opposition , d' effacement et de portabilité de vos données que vous pouvez exercer par courrier électronique à l' adresse : @email . fr en justifiant de votre identité .

Figure 3.4: Exemple d'un courriel contenant un historique de conversation. Nous mettons en avant, en orange, les ajouts de contenus automatiques en fin de courriel. Les dates ont été remplacées ansi que les montants à des fins de désidentification des clients

4. Un conseiller EDF répond à un client et un message automatique en anglais apparaît en fin de mail (ou l'inverse).

Le dernier cas a été géré dans le paragraphe précédent, lorsque nous avons supprimé les contenus automatiques dans les courriels. Pour détecter la langue des courriels, nous utilisons un modèle FastText permettant d'identifier 176 langues. Ce modèle a été appris en utilisant des mots, des sous-mots et des n-grams de différentes tailles qui constituent les mots. Puis, une étape de sélection de caractéristiques est ajoutée pour permettre de supprimer celles qui n'ont pas beaucoup d'impact sur la décision du classifieur. Enfin, le modèle d'identification de la langue a pu être appris sur les 170 langues. Ce modèle a été testé sur trois jeux de données et a permis d'obtenir les résultats présentés dans le Tableau 3.5. Dans cette étude, nous avons utilisé le modèle *small* qui est plus petit mais qui nécessite moins d'espace mémoire. En utilisant le modèle, 98% des courriels sont identifiés comme étant rédigés en français (soit 98.822 courriels). La deuxième langue la plus fréquente est l'anglais, qui représente 878 courriels puis l'espagnol avec 92 courriels. Au total, 30 langues ont été identifiées par FastText. Dans la suite de cette étude, nous

traiterons uniquement les courriels rédigés en français.

		Wikipédia		TCL		EuroGov	
Model	Size	Acc.	Time	Acc.	Time	Acc.	Time
langid.py	2.5MB	91.3	11.4	93.1	10.8	98.7	16.1
fastText (Large)	126MB	93.1	0.9	95.1	1.1	98.9	2.7
fastText (Small)	917kB	92.7	1.5	94.6	1.6	98.9	4.9

Table 3.5: Performances de l'identification de la langue avec FastText

3.3 Évaluation extrinsèque

Dans cette section, nous présenterons les corpus utilisés pour l'évaluation extrinsèque (ou analyse quantitative) réalisée dans notre étude (voir Section 2.5.1.1). La présentation des corpus est réalisée en deux temps et débute par la présentation des tâches. Ensuite, nous présentons les corpus permettant de traiter ces tâches pour le français et pour l'anglais.

3.3.1 Tâches

Analyse de sentiment (*sentiment analysis* ou *opinion mining*). Cette tâche a pour objectif de classer la polarité des émotions (positives, négatives et neutres), les sentiments et émotions (colère, joie, tristesse, etc.) ou les intentions (e.g., intéressé ou non intéressé) dans des données textuelles à l'aide de techniques de fouille de textes. Cette tâche peut donc être découpée en trois types d'analyse. Voici quelques exemples d'analyse :

- La polarité : le système de notation d'un site de critiques de cinéma (*IMDB*, Allociné, etc.)
- Les émotions : retours clients sur le support d'un site internet.
- L'aspect ou les caractéristiques : une compagnie d'électricité peut chercher à savoir si les offres des concurrents font l'objet de critiques (positives ou négatives).

Dans cette étude, nous nous intéressons à la polarité car il s'agit de l'aspect pour lequel nous disposons de données annotées et qu'il s'agit d'une tâche qui peut aider la tâche de classification prévue sur les courriels EDF. Il s'agit d'une tâche plus simple que celle des émotions, car elle regroupe certaines émotions pour former une classe (e.g., la colère et la tristesse sont regroupées dans la classe de polarité négative).

Détection de paraphrases. Cette tâche consiste à calculer le niveau de similarité qui existe entre deux énoncés. Dans notre étude, nous nous intéressons à la détection de similarité entre une phrase A et une phrase B. Lejeune and Barbaresi [2020][5] définissent trois formes de similarité entre des textes, allant de la plus simple à la plus difficile à détecter :

1. La copie consiste à copier mot à mot tout ou une partie d'un texte dans un autre. Dans notre étude, il s'agit du cas où la phrase A est identique à la phrase B. *Exemple : Il fait beau aujourd'hui. <-> Aujourd'hui, il fait beau.*

2. La paraphrase consiste à reprendre une phrase pour la détailler ou l'expliciter. La phrase B conserve l'ordre des éléments évoqués dans la phrase A, en autorisant des variations de vocabulaire (exemple : l'utilisation de synonymes), ainsi que des opérations classiques de modification (ajout, suppression et substitution de mots). *Exemple : Il fait beau aujourd'hui <-> Alors qu'il pleuvait hier, aujourd'hui il fait beau.*

3. La reformulation autorise toutes les modifications textuelles d'une phrase, à condition que son sens soit conservé. Les mécanismes de reformulation sont plus lourds et plus élaborés que pour la paraphrase, et nécessite une réorganisation des concepts. *Exemple : Il fait beau aujourd'hui <-> Le temps est agréable aujourd'hui*

Inférence ou implication textuelle (*Textual entailment*). Cette tâche consiste à analyser la relation directionnelle entre une séquence source (S) et une séquence cible (C). Les séquences sont souvent des phrases, mais peuvent également être caractérisées par des énoncés courts. Dire que la séquence C peut être inférée de la séquence S signifie qu'un être humain lisant la séquence C peut raisonnablement conclure que la séquence S est vraie [Dagan et al., 2005]. En d'autres termes, le sens fourni par l'énoncé C peut être déduit du sens exprimé par l'énoncé S. Selon la nature des inférences mises en jeu, la relation d'implication textuelle englobe plusieurs formes de relations entre deux énoncés. Rossari [1990] distingue la reformulation paraphrastique – une relation d'équivalence entre les deux énoncés – et la relation non paraphrastique qui opère un changement de perspective énonciative. Dans cette étude, nous nous intéressons aux implications textuelles non paraphrastiques, étant donné que la tâche précédente est une tâche de reconnaissance de paraphrases. L'implication textuelle est encore plus difficile à résoudre que la détection de paraphrases, car elle nécessite une compréhension fine

[5]Il existe plusieurs définitions de la similarité entre des textes qui sont assez similaires. Nous choisissons cette définition car elle permet de distinguer finement trois niveaux de complexité pour la tâche de paraphrase.

des phrases ainsi que du raisonnement, voire des connaissances externes. *Exemple d'implication textuelle : Il fait beau aujourd'hui. Je n'ai pas besoin de mon parapluie.*

Reconnaissance d'entités nommées (*Named-entity recognition (NER)*). Cette tâche est une sous-tâche de l'extraction d'informations et vise à localiser et à classer les entités nommées dans un document. Ce classement est réalisé pour des catégories prédéfinies telles que des noms de personnes, des lieux, des entreprises / organisations, des codes médicaux, etc. La reconnaissance d'entités nommées (REN) a pour objectif, à partir d'un bloc de texte non annoté, de générer un bloc de texte annoté qui met en évidence les noms des entités. La difficulté de cette tâche réside dans la compréhension d'un contexte pour en déduire une classe de façon automatique. Un exemple d'annotation d'un texte avec spaCy (voir Figure 3.5) met en évidence les difficultés pour reconnaître la présence d'une entité (e.g., *Enel*), délimiter les frontières de l'entité (e.g., *China Energy Investment*) et catégoriser correctement l'entité (e.g., PER). Toutes ces difficultés sont accrues lorsqu'il s'agit d'un domaine de spécialité.

Figure 3.5: Exemple de reconnaissance d'entité nommées dans un texte avec spaCy (modèle *small* en français). Trois entités nommées : noms de personnes (PER), lieux (LOC) et organisations (ORG).

3.3.2 Jeux de données

Dans notre étude, nous avons choisi d'étudier des corpus en anglais et en français, afin de déterminer si la complexité de la syntaxe d'une langue pouvait affecter les résultats de nos méthodes (voir Chapitre 7).

Analyse de sentiments Pour l'analyse de sentiments, nous avons utilisé deux corpus en anglais (références corpus section données) et deux corpus en français (voir Tableau 3.6). Nous avons sélectionné des jeux de données avec des annotations binaire et multi-classe afin de comparer les résultats sur des classements difficiles.

Reconnaissance de paraphrases Pour la reconnaissance de paraphrases, nous avons utilisé deux corpus en anglais et trois corpus en français. Nous avons

sélectionné des jeux de données avec des annotations binaire et multi-classe afin de comparer les résultats sur des classements difficiles.

Implication textuelle Pour l'implication textuelle, nous avons utilisé un seul jeu de données en français et deux en anglais, qui sont fréquemment utilisés pour chacune de ces langues.

Reconnaissance d'entités nommées Enfin, deux jeux de données sont exploités pour l'anglais et le français, avec des sources équivalentes pour les deux langues (Wikipedia).

Corpus	Langue	Domaine	Tâche	#Labels	Train	Dev	Test
IMDB [Maas et al., 2011]	Anglais	Avis clients	Sentiment	2	40 000	-	10 000
SST-2 [Socher et al., 2013]	Anglais	Avis clients	Sentiment	2	6 920	872	1 821
CLS-FR [Le et al., 2020b]	Français	Avis clients	Sentiment	2	5 484	-	1 372
DEFT-18 [Paroubek et al., 2018]	Français	Tweets	Sentiment	4	22 537	-	5 635
PIT [Xu et al., 2015]	Anglais	Tweets	Paraphrase	2	11 530	4 142	838
Quora [Wang et al., 2017b]	Anglais	Forum	Paraphrase	2	323 432	80 858	234 579
DEFT-20 [Cardon et al., 2020]	Français	Médical	Paraphrase	6	360	120	120
OpusParcus-FR [Creutz, 2018]	Français	Sous-titres	Paraphrase	5	50 000	973	1 007
PAWS-X-FR [Le et al., 2020a]	Français	Général	Paraphrase	2	49 127	646	2 000
MNLI [Williams et al., 2018]	Anglais	Variés	NLI	3	392 702	9 815	
MNLI-Mis [Williams et al., 2018]	Anglais	Variés	NLI	3	392 702	9 832	
XNLI-FR [Le et al., 2020b]	Français	Variés	NLI	3	392 702	2 490	5 010
WikiNER-EN [Nothman et al., 2013]	Anglais	Wikipédia	NER	4	87 379	29 126	29 126
WikiNER-FR [Nothman et al., 2013]	Français	Wikipédia	NER	4	87 379	29 126	29 126

Table 3.6: Description des jeux de données

3.4 Synthèse

Dans ce chapitre, nous avons présenté les corpus qui nous permettront d'évaluer la performance de modèles dans la suite de ce manuscrit. Plus précisément, nous nous sommes intéressés aux caractéristiques des corpus que nous utiliserons dans la suite de ces travaux, à savoir le registre de langue, le format, le style de rédaction, le domaine de spécialité et la langue de rédaction (anglais ou français).

Insérer une citation

—

4

Modèles de plongements lexicaux

Table des Matières

4.1 Introduction

L'objectif de ce chapitre est de présenter les modèles de plongements lexicaux sur lesquels nous travaillerons. Nous détaillerons les modèles pré-entraînés utiles à notre recherche. Nous ne passerons en revue ni l'historique des modèles (abordé dans le Chapitre 2), ni l'architecture des modèles (présentée dans l'annexe A).

Dans un premier temps, nous détaillerons les corpus ayant servi à entraîner les modèles exploités dans notre recherche. Plus précisément, nous nous intéresserons aux propriétés de ces corpus (volumétrie, propreté, et richesse lexicale) afin de formuler des hypothèses de recherches quant à leur impact sur l'application des

modèles à des données de spécialité.

Dans un deuxième temps, nous détaillerons les modèles pré-entraînés évalués dans nos travaux, et appartenant à trois familles de modèles : Word2Vec, ELMo et Transformer. L'application de plusieurs modèles à des données de spécialité a pour objectif de comparer les biais d'apprentissage générés par plusieurs architectures neuronales.

4.2 Corpus d'entraînement

Dans cette partie, nous présentons les corpus ayant permis d'entraîner les modèles de plongements lexicaux pré-entraînés que nous exploiterons dans notre recherche. Ces corpus n'ont donc pas été utilisés dans nos expériences. Cependant, nous nous intéressons à la problématique de l'adaptation de ces modèles à des données de spécialité, ce qui nécessite de comparer différents corpus d'apprentissage afin d'évaluer l'impact de l'entraînement initial. Les données d'apprentissage des modèles exploités dans nos travaux sont issues de quatre sources de données du domaine général (voir Figure 4.1).

1. Les données Web. *Les données extraites à partir du Web sont faciles à récupérer et peuvent être exploitées en grande quantité. Ces deux avantages expliquent pourquoi tous les modèles que nous avons exploités utilisent des données issues du Web. Cependant, ces données ont l'inconvénient de nécessiter de nombreux prétraitements.*

Le corpus FrWac. FrWac[1] est un corpus de 1.6 milliard de mots construit à partir de données issues du Web, en limitant le *webcrawling* au domaine *.fr*. Des mots de moyenne fréquences du corpus du Monde Diplomatique et de listes du vocabulaire français basique ont été utilisés pour l'extraction de données. Le corpus a été étiqueté en morpho-syntaxe et lemmatisé avec TreeTagger[2]. Dans cette étude, nous n'utiliserons pas les étiquettes linguistiques.

Le corpus Oscar. Oscar [Martin et al., 2020] est un ensemble de corpus monolingues extraits à partir de Common Crawl. Les corpus ont été sélectionnés à l'aide d'un modèle de classification pour chaque langue suivant l'approche de Grave et al. [2018] basée sur FastText [Joulin et al., 2016]. Le classifieur a été préalablement pré-entraîné sur Wikipedia, Tatoeba et SETimes, et couvre

[1]Pour plus de détails sur FrWac : https://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf.

[2]Pour plus de détails sur TreeTagger : https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

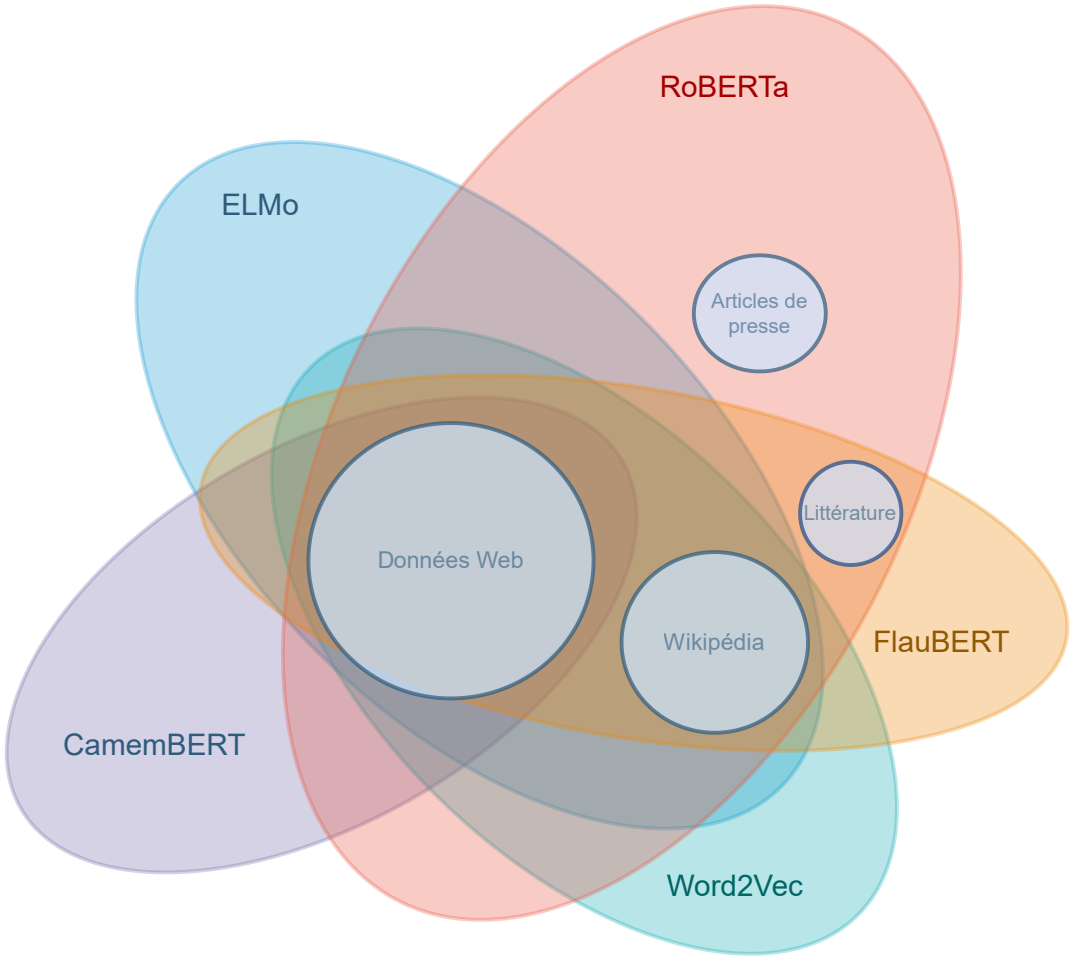


Figure 4.1: Comparaison des modèles de plongements lexicaux en fonction du type de données d'apprentissage ayant servi à les entraîner. La taille des sphères n'est pas à l'échelle, mais elle permet de représenter la taille des corpus de données en fonction de leur source. Par exemple, la récupération de données Web est la plus efficiente car elle permet de récupérer plus de données pour l'apprentissage des modèles.

176 langues. Dans ce manuscrit, nous utiliserons uniquement des modèles entraînés sur la version française d'Oscar.

Le corpus CCNet. CCNet [Wenzek et al., 2020] est un jeu de données extrait de Common Crawl mais avec un filtrage différent de celui d'Oscar. Il a été construit avec un modèle de langue utilisant Wikipédia, ce qui lui permet de filtrer le bruit (code, tableaux, etc.). CCNET contient donc des documents en moyenne plus longs qu'Oscar.

Le corpus OpenWebText. Le corpus OpenWebText[3] est une re-création

[3]Pour plus d'informations sur OpenWebText : https://github.com/jcpeterson/openwebtext.

OpenSource du jeu de données WebText utilisé pour entraîner GPT-2, après nettoyage de celui-ci.

Le corpus Stories. Le corpus Stories [Trinh and Le, 2018] a été construit dans l'objectif d'évaluer des questions à choix multiples posées par des tests de raisonnement de bon sens. Plus précisément, il a été construit pour les tâches de la désambiguïsation des pronoms [Levesque et al., 2012]. La collecte de données issues de CommonCrawl et divers prétraitements (méthodes de *ranking*) ont permis de générer ce corpus d'environ 1 million de documents.

2. Les données Wikipédia. *Les données issues de Wikipédia*[4] *sont moins diverses et moins volumineuses que les données Web. En effet, bien que la bibliothèque numérique comporte de nombreux articles, elle n'est pas aussi riche (en points de vue comme en thématiques) que l'ensemble du Web. Néanmoins, ces données possèdent l'avantage d'être structurées et ne nécessitent pas beaucoup de prétraitements spécifiques. De plus, les données Wikipédia sont des articles longs et plus propres, et permettent donc de créer beaucoup de contexte entre les mots du vocabulaire.*

Le corpus WikiDumps. Ce corpus[5] [Foundation] de 600 millions de mots a été construit à partir de *dumps* Wikipédia monolingues contenant uniquement des pages du domaine associé à une langue (par exemple, le domaine *.fr* pour le français). Dans ce manuscrit, nous nous intéressons uniquement aux corpus monolingues pour l'anglais et pour le français.

3. Les données de presse. *Les données issues d'articles de presse sont bien rédigées et ne nécessitent pas de prétraitements spécifiques. Ils ont l'avantage de discuter de sujets variés, allant de concepts généraux à des tendances spécifiques pouvant contenir du vocabulaire de spécialité. L'inconvénient de ces corpus est qu'ils représentent peu de données Open Source et sont donc limitant pour l'apprentissage de modèles de langue. De plus, ils ont tendance à être polarisés dans certaines revues, ce qui peut introduire des biais, des stéréotypes ou des préjugés très forts dans les modèles.*

Le corpus CCNews. Le jeu de données CCNews[6] contient 63 millions d'articles de presse en anglais extraits à partir du Web et récupérés entre septembre 2016 et février 2019.

4. Les données littéraires. *Les collections de corpus contenant des œuvres littéraires sont également très propres et nécessitent peu de préentraînement (sauf s'ils*

[4] https://www.wikipedia.org/

[5] Pour plus de détails sur WikiDumps : https://dumps.wikimedia.org/frwiki/.

[6] Pour plus d'informations sur CCNews https://commoncrawl.org/2016/10/news-dataset-available/

ont été récupérés avec des processus d'océrisation). L'avantage de ces corpus est qu'ils contiennent diverses visions de la réalité et permettent donc de construire du contexte riche entre les mots du vocabulaire. Cependant, ces corpus sont généralement trop peu volumineux pour être utilisés à eux seuls durant l'apprentissage de grands modèles de langue, et servent donc de complément aux corpus d'apprentissage plus volumineux.

Le corpus BookCorpus. Les livres sont une source riche d'informations à la fois fines (par exemple, l'apparence d'un personnage, d'un objet ou d'une scène) et sémantiques de haut niveau (comme ce que quelqu'un pense, ressent et comment ces états évoluent au cours de l'histoire). La création de BookCorpus [Zhu et al., 2015] a été motivée par l'alignement de livres sur leurs sorties au cinéma, afin de fournir des explications descriptives riches du contenu visuel (contenu cinématographique) qui vont sémantiquement bien au-delà des légendes disponibles dans les corpus classiques. Ce corpus contient 11 038 livres et se compose d'un vocabulaire de 1 316 420 mots.

4.3 Modèles de plongements lexicaux

Dans cette section, nous présentons tous les modèles pré-entraînés utilisés dans la suite de ce manuscrit.

4.3.1 Les modèles Word2Vec

Word2Vec [Mikolov et al., 2013] est un modèle de plongement lexical statique permettant de représenter des mots dans leur contexte d'utilisation (comme un document ou une phrase). Son utilisation repose sur l'apprentissage de représentations vectorielles de mots dans des corpus suffisamment grands. Nous exploiterons ce type de modèle pour représenter contextuellement les mots de notre corpus. De plus, nous comparerons ces vecteurs à ceux générés directement à partir de modèles pré-entraînés[7] sur Wikipédia et le Web proposés par Fauconnier [2015]. Nous nous limitons à l'entraînement Skip-Gram de Word2Vec, plus propice à la représentation de la sémantique que l'apprentissage CBOW, utile pour des calculs de similarité syntaxique. Afin d'effectuer une comparaison équitable des modèles, tous les modèles Word2Vec construisent des vecteurs de dimension 1 000.

4.3.2 Le modèle ELMo

ELMo [Peters et al., 2018a] est un modèle de langue pré-entraîné et supervisé permettant de générer des plongements lexicaux contextuels. Dans nos recherches,

[7] Pour plus de détails sur les modèles Word2Vec : `https://fauconnier.github.io/#data`.

nous avons utilisé un modèle pré-entraîné disponible pour le français [Che et al., 2018b, Fares et al., 2017], avec des plongements lexicaux de taille 512. Ce modèle a été entraîné sur un ensemble de données contenant 20 million de mots, sélectionnés aléatoirement à partir des corpus WikiDump et CommonCrawl[8].

4.3.3 Les modèles Transformer

Les trois modèles Transformer présentés dans cette section sont comparés dans le Tableau 4.1.

	RoBERTa-Base	**CamemBERT-Base**	**FlauBERT-Base**
Langue	Anglais	Français	Français
Données d'entraînement	160 GB	138 GB†	71 GB‡
Objectifs	MLM	MLM	MLM
Nombre de paramètres	125 M	110 M	138 M / 373 M
Tokeniseur	BPE 50K	SentencePiece 32K	BPE 50K
Masking dynamique	Sous-mots	Mots	Sous-mots

Table 4.1: Comparaison générale entre les modèles de langue réalisée par Le et al. [2020a]. †, ‡: 282 GB, 270 GB avant le prétraitement des données

Le modèle RoBERTa. RoBERTa est un modèle Transformer pré-entraîné sur un grand corpus de données en anglais de manière auto-supervisée. Cela signifie qu'il a été pré-entraîné sur les textes bruts uniquement, sans aucun étiquetage humain (c'est pourquoi il peut utiliser de nombreuses données publiques) avec un processus automatique pour générer des entrées et des étiquettes à partir de ces textes. Plus précisément, il a été pré-entraîné avec l'objectif de modélisation du langage masqué (MLM). À partir d'une phrase, le modèle masque aléatoirement 15% des mots de l'entrée, puis fait passer la phrase masquée entière par le modèle et doit prédire les mots masqués. Cette méthode est différente des réseaux neuronaux récurrents traditionnels (RNN) qui voient généralement les mots les uns après les autres, ou des modèles auto-régressifs comme GPT qui masquent les futurs tokens. Il permet au modèle d'apprendre une représentation bi-directionnelle de la phrase. De cette façon, le modèle apprend une représentation de la langue qui peut ensuite être utilisée pour extraire des caractéristiques utiles à des tâches en aval. Par exemple, à partir d'un corpus de phrases étiquetées, un classifieur standard peut être entraîné grâce aux représentations produites par le modèle BERT. Dans nos travaux, nous avons utilisé le modèle *Base* de RoBERTa, qui contient 12 couches cachées. RoBERTa a été pré-entraîné sur cinq corpus : BookCorpus, WikipediaDumps pour l'anglais

[8] https://commoncrawl.org

(à l'exception des listes, des tableaux et des en-têtes), CC-News, OpenWebText et Stories. Nous utiliserons RoBERTa pour nos expériences sur l'anglais.

Le modèle CamemBERT. CamemBERT [Martin et al., 2020] est le premier modèle de langue Transformer en français, issu de l'architecture RoBERTa. Dans notre manuscrit, nous utiliserons quatre modèles CamemBERT, détaillés dans le Tableau 4.2. L'avantage de ce modèle est qu'il en existe plusieurs variantes, que ce soit sur le nombre de couches, de paramètres mais aussi sur les corpus d'entraînement et leur taille. Par conséquent, c'est la famille de modèles que nous utiliserons lorsque nous souhaiterons effectuer une analyse approfondie de l'impact de l'entraînement des modèles sur la représentation des mots d'un corpus.

Modèle	**#Paramètres**	**#Couches**	**Corpus**
camembert-base	110M	12	OSCAR (138 GB)
camembert-large	335M	24	CCNet (135 GB)
camembert-base-wikipedia-4gb	110M	12	Wikipedia (4 GB)
camembert-base-oscar-4gb	110M	12	OSCAR (4 GB)
camembert-base-ccnet-4gb	110M	12	CCNet (4 GB)

Table 4.2: Présentation des modèles CamemBERT

Le modèle FlauBERT. Le modèle FlauBERT [Le et al., 2020a] est sorti quelques semaines après CamemBERT, avec le nouveau référenciel d'évaluation francophone pour le TAL appelé FLUE. Pour entraîner FlauBERT, les auteurs utilisent une configuration similaire à celle de CamemBERT. Dans l'ensemble, FlauBERT offre des performances très similaires à celles de CamemBERT alors qu'il est entraîné sur un volume plus restreint de données. Dans ce manuscrit, nous utilisons uniquement le modèle FlauBERT-Base ayant appris sur un réseau de neurones contenant 12 couches cachées. Les données d'apprentissage de FlauBERT consistent en 24 sous-corpus récoltés à partir de diverses sources, allant de textes bien rédigés (Wikipédia et des corpus littéraires)[9] à des corpus extraits à partir du Web (comme CommonCrawl). Ces modèles ont donc appris à partir de source plus variées que CamemBERT, ce qui pourrait avoir un impact positif sur la représentation de termes spécifiques au domaine.

[9] https://data.statmt.org/ngrams/deduped2017/

4.4 Synthèse

Dans ce chapitre, nous avons présenté les modèles de plongements lexicaux pré-entraînés que nous exploiterons dans la suite de ce manuscrit. Tout d'abord, nous avons énuméré les corpus d'apprentissage ayant permis d'entraîner ces modèles. Ces corpus sont issus du domaine général et correspondent à plusieurs styles de rédaction et types de contenus. Les données les plus faciles à extraire et les plus volumineuses sont les données Web, utilisées par tous les modèles. Cependant, ces données nécessitent des prétraitements qui peuvent être lourds et ne sont pas toujours correctement rédigées. Au contraire, des articles de presse ou des œuvres littéraires sont plus propres et nécessitent moins de prétraitements, mais sont plus rares. Dans ce manuscrit, nous nous intéressons à l'impact du choix du corpus d'apprentissage sur la représentation de données de spécialité. Nous pensons que plus les données sont propres et correctement structurées, et plus les modèles seront à même à capturer des relations entre des termes. Par conséquent, les modèles ayant uniquement appris sur des corpus Web bruités devraient générer plus de bruit. Néanmoins, nous émettons également l'hypothèse que l'apprentissage de modèles sur des données traitant de thématiques riches et nombreuses, comme les données issues du Web, pourrait faciliter l'apprentissage de vocabulaire de spécialité.

Ensuite, nous avons présenté les trois familles de modèles pré-entraînés pertinentes dans notre étude : les modèles Word2Vec, le modèle ELMo et les modèles Transformer (RoBERTa, CamemBERT et FlauBERT). Dans nos travaux, nous comparerons ces modèles sur leur capacité à représenter des termes hors-vocabulaire. Nous nous intéressons tout particulièrement à CamemBERT, qui a l'avantage d'avoir été entraîné sur des sources de données très différentes. Ces modèles seront comparés à FlauBERT, qui a appris sur une collection de corpus variée du point de vue du style de rédaction et du contenu. De plus, ces modèles utilisent des algorithmes de tokénisation différents, présentés dans le Chapitre 2.5.1.1. Plus récemment, le modèle FrAlbert [Cattan et al., 2021] a émergé comme une nouvelle alternative pour le traitement du français. Cependant, nous ne l'utiliserons pas dans nos expériences, mais des travaux complémentaires sur ce modèle seraient envisageables pour des recherches futures, afin d'étudier l'impact de l'apprentissage d'un modèle sur une architecture Transformer plus petite.

Conclusion et Discussion de la partie

Dans cette deuxième partie, nous avons introduit le matériel et les méthodes nécessaires aux expériences proposées dans notre manuscrit.

Corpus. Nous avons présenté des corpus adaptés à deux formes d'évaluation : l'évaluation intrinsèque (ou évaluation qualitative) et l'évaluation extrinsèque (ou évaluation quantitative) des modèles de langue. Dans le cadre de l'évaluation qualitative, l'objectif de nos travaux est d'analyser, de diverses manières, les spécificités liées au format des données, à leur style de rédaction et à leur domaine. Pour cela, nous avons présenté cinq corpus en langue française ayant des spécificités différentes. Nous traiterons trois domaines de spécialité : le domaine juridique, le domaine biomédical et le domaine de l'énergie. De plus, nous avons choisi de traiter une variété de formats différents, comme les courriers électroniques, les pièces de théâtre et des articles scientifique / juridique. Ensuite, nous avons présenté les corpus réservés à l'évaluation extrinsèque des modèles que nous réaliserons sur la langue française et anglaise. Ces jeux de données permettent de résoudre quatre tâches classiques de TAL : l'analyse de sentiments, la détection de paraphrases, l'implication textuelle et la reconnaissance d'entités nommées.

Modèles pré-entraînés. Nous avons détaillé trois familles de modèles permettant de générer des plongements lexicaux : Word2Vec (plongements statiques), ELMo (plongements contextuels) et les modèles Transformer (plongements contextuels). Nous avons présenté des modèles en français pour Word2Vec et ELMo et des modèles monolingues en français (CamemBERT et FlauBERT) et en anglais (RoBERTa) pour les Transformer. Dans nos travaux, nous utilisons donc des modèles de plongements lexicaux usuels, c'est-à-dire ayant appris sur des données du domaine général. Dans ce manuscrit, nous étudierons en détail l'impact de l'utilisation de ces modèles sur des corpus comportant des spécificités linguistiques (le domaine de spécialité et le style de rédaction).

* * *

Partie III

Évaluations de biais d'apprentissage dans des modèles de plongements lexicaux

Introduction de la partie

Cette partie vise à évaluer les biais d'apprentissage contenus dans des modèles de plongements lexicaux, et ce, en exploitant des modèles de représentation statiques et contextuels. Afin de mieux comprendre, dans le cadre de notre recherche, comment adapter des modèles de plongements lexicaux à des domaines spécifiques, nous devons tout d'abord qualifier les biais d'apprentissage liés au contenu des données d'apprentissage de ces modèles.

Pour ce faire, le premier chapitre visera à analyser les stéréotypes de genre induits dans des modèles de plongements lexicaux statiques. Bien que ces modèles ne soient pas les plus récents, leur interprétation est plus simple que les modèles contextuels. De plus, nos travaux démontrent des biais de représentation importants liés à l'apprentissage de ces modèles sur un corpus de pièces de théâtre ayant la particularité d'appuyer sur des stéréotypes existants dans une optique comique ou tragique. À notre connaissance, nos travaux sont les premiers à effectuer une étude trans-disciplinaire (en nous basant sur des travaux de littérature et de sociologie) des modèles sur les stéréotypes de genre au théâtre.

Le deuxième chapitre a pour objectif de présenter les difficultés liées à l'application de modèles Transformer pré-entraînés à trois domaines de spécialité : le domaine de l'énergie, le domaine de la biologie et le domaine juridique. Les données de spécialité sont souvent rares ou difficiles à obtenir, ce qui rend difficile l'entraînement de modèles de haute qualité. De plus, il arrive que les modèles formés sur des données de spécialité soient trop adaptés à ce domaine spécifique et ne pas fonctionner aussi bien dans d'autres domaines ou contextes. Nous verrons qu'il est difficile de représenter des termes hors-vocabulaire appartenant à un nouveau domaine, et qu'il s'agit de biais de représentation liés intrinsèquement avec les corpus d'apprentissage des modèles. Nos travaux démontrent que la représentation de ces termes est principalement due à leur tokenisation en amont et nous proposerons deux mesures permettant d'évaluer la représentation finale de ces termes.

* * *

Des temps sombres et difficiles nous attendent. Bientôt, nous devons tous faire face au choix entre ce qui est juste et ce qui est facile.

— Harry Potter et la coupe de feu

5 Les stéréotypes de genre

Table des Matières

5.1 Introduction

Les modèles de représentation les plus récents cherchent à capturer au mieux toutes les subtilités de la langue, ce qui implique de récupérer les stéréotypes qui y sont associés. Dans ce chapitre, nous souhaitons déterminer si les stéréotypes de genre attendus dans un corpus de pièces de théâtre sont capturés par les plongements lexicaux. Nous nous sommes intéressés aux modèles statiques, et plus particulièrement à Word2Vec pour l'analyse des stéréotypes de genre, de part leur

facilité d'utilisation, leur interprétabilité mais également en émettant l'hypothèse que, si des stéréotypes de genre sont capturés par des modèles de représentation statiques, des modèles contextuels plus complexes et plus fins permettront également de les modéliser. Enfin, ces modèles présentent l'avantage de pouvoir être ré-entraînés *from scratch* tout en conservant de bonnes performances et en consommant moins d'énergie, en moyenne, que les modèles contextuels.

Pour cette étude, nous avons constitué un jeu de données composé de pièces de théâtre françaises allant du XVI^{e} au XIX^{e} siècle. Nous avons choisi de travailler sur le genre théâtral car il tend à pousser à leur paroxysme certains traits de caractère représentatifs de hiérarchies sociales préexistantes. Nous présentons des expériences dans lesquelles nous parvenons à mettre en avant des stéréotypes de genre en relation avec les rôles et les émotions traditionnellement imputés aux femmes et aux hommes. De plus, nous mettons en avant une sémantique spécifique associée à des personnages féminins et masculins. Cette étude démontre l'intérêt de mettre en évidence des stéréotypes dans des corpus à l'aide de modèles de plongements lexicaux statiques. Au vu de la spécificité des genres du théâtre, nous étudions tout particulièrement les stéréotypes de genre existant dans deux genres : la comédie et la tragédie. Pour cela, nous nous intéressons d'abord à la sémantique des termes associés à « femme » et « homme » et notamment aux émotions qui les caractérisent. Puis, nous nous intéressons à quelques personnages types du théâtre (e.g., *valet, servante, maître*, etc.) afin d'étudier les spécificités de genre et de rôles captées par ces modèles.

5.2 Contexte et problématique

Les systèmes de décision automatique consistent à utiliser des outils permettant d'analyser des jeux de données à des fins de prédiction, classification, déclenchement d'action ou encore pour générer des scores. Dans ce contexte, les systèmes automatisés entraînent parfois des conséquences importantes sur des aspects cruciaux de la vie tels que l'éducation, la performance au travail, les opportunités bancaires ou encore la santé. Les retombées des décisions prises par les systèmes automatisés sont d'autant plus grandes pour les groupes désavantagés, lorsque les modèles penchent plus vers un groupe de personnes dans ces secteurs. Kendall and Gal [2017] a démontré qu'en traitement d'images, certains algorithmes échouent à détecter les visage de leurs utilisateurs Noirs, allant même jusqu'à en étiqueter certains comme des « gorilles » [Howard and Borenstein, 2019]. Dans le secteur de la santé, il a été découvert qu'un algorithme largement exploité dans les hôpitaux américains, conçu pour attribuer une sécurité sociale aux patients, était systématiquement discriminatoire à l'égard des patients Noirs [Obermeyer et al., 2019]. Les auteurs

ont remarqué que, à maladie égale, le système automatique était moins susceptible d'orienter les patients Noirs que les patients Blancs vers des programmes visant à améliorer les soins aux patients ayant des besoins médicaux complexes. De plus, les systèmes de prédiction de risques de récidive criminelle prédisent que des personnes issues de certaines ethnies ont plus de chances que d'autres de commettre un crime [Tolan et al., 2019]. Dans le domaine de la détection de la parole, il a été démontré que les systèmes de dictée vocale étaient plus performants pour les sujets masculins que pour les sujets féminins [Rodger and Pendharkar, 2004].

Dans notre recherche, nous souhaitons inclure ce sujet car il constitue une part importante de l'éthique de l'utilisation des modèles d'apprentissage automatique, mais aussi car il offre une problématique liée à aux biais d'apprentissage de ces modèles. En effet, le problème de ces systèmes automatiques est qu'ils sont entraînés sur des données rédigées par des humains, qui contiennent des stéréotypes et des préjugés existants dans la société. Les données représentent toujours une connaissance et une vision limitée du monde, avec ses biais. Ces travaux permettent de répondre à notre problématique de recherche liée à l'évaluation de problèmes de représentation suite à l'apprentissage de données massives brutes (au sens où ces données ne sont pas filtrées en amont). Afin d'analyser la présence ou non de stéréotypes dans cette étude, nous avons choisi de nous pencher sur les stéréotypes de genre, car ils peuvent être mis en évidence grâce à des méthodes binaires (un groupe représentatif des hommes et un autre des femmes). De plus, des méthodes plus complexes de pré-traitement sont nécessaires pour détecter qu'une phrase concerne une personne appartenant à un groupe spécifique (qu'il s'agisse du locuteur ou du sujet de la phrase). Enfin, nous souhaitons mener cette étude sur un corpus dont on sait qu'il contient ces biais, ce qui permet de vérifier si le modèle les capture. Nous émettons l'hypothèse que les plongements lexicaux, de par leur capacité à représenter la sémantique des termes d'un corpus, capturent les biais contenus dans les corpus et engendrent donc des risques lorsqu'ils sont implémentés dans des systèmes de prises de décision, s'ils ne contiennent pas de filtres en amont ou en aval de leur conception. Enfin, nous nous intéressons tout particulièrement à la subtilité de la représentation des stéréotypes dans les corpus, à travers l'analyse comparative des stéréotypes dans des modèles ayant appris sur différents corpus.

5.3 Définition : biais, stéréotype ou préjudice ?

5.3.1 Les biais en sciences humaines

L'étude des différents types de biais en sciences cognitives est menée depuis plus de cinq décennies [Kahneman and Tversky, 1973]. Depuis le tout début, les préjugés ont été considérés comme une stratégie humaine innée pour la prise de décision. Lorsqu'un biais cognitif est appliqué, nous présumons que la réalité se comporte en fonction de certains *a priori* cognitifs qui peuvent ne pas être vrais, mais avec lesquels nous pouvons former un jugement. Un préjugé peut être acquis par un processus d'induction incomplet (une vision limitée de tous les échantillons ou situations possibles) ou appris des autres (éducation ou observation). Dans tous les cas, un biais fournira un mode de pensée éloigné du raisonnement logique. Il existe beaucoup de biais cognitifs qui ont été identifiés et classés en différentes catégories comme les biais sociaux, comportementaux, les biais de mémoire, etc. Parmi eux, nous nous concentrons sur les biais sociaux, et tout particulièrement sur les stéréotypes. Alors que le biais cognitif peut être défini comme un cas dans lequel la cognition humaine[1] produit des représentations fiables qui sont systématiquement déformées par rapport à la réalité objective, le stéréotype peut être défini comme l'hypothèse de certaines caractéristiques appliquées à d'autres personnes sur la base de caractéristiques sociales (nationalité, genre, âge, ethnie, etc.). Par conséquent, les stéréotypes attribuent des caractéristiques à un individu parce qu'il appartient à un certain groupe social (par exemple, « les italiens parlent avec les mains »). Dans ce manuscrit, nous ne parlerons pas d'égalité de traitement mais d'équité de traitement (*fairness*), un concept largement traité en apprentissage automatique [Sahil and Julia, 2018].

5.3.2 Application en apprentissage automatique

L'équité est un concept intimement lié aux préjugés. Un système est considéré comme « équitable » lorsque ses résultats ne sont pas discriminatoires en fonction de certains attributs, comme le genre ou la nationalité. Dans l'évaluation de l'apprentissage automatique, la discrimination peut être estimée en examinant les matrices de confusion pour différents groupes. En d'autres termes, nous pouvons calculer des matrices de confusion et les taux qui leurs sont associés (taux de positifs, taux de vrais positifs, taux de faux positifs, etc.) pour chaque sous-ensemble d'échantillons (e.g., « femme »), et comparer les taux obtenus avec d'autres sous-ensembles (e.g., « homme ») pour la même caractéristique (e.g., le

[1] La cognition est le processus d'acquisition de la connaissance. (Source :TLFi https://www.cnrtl.fr/definition/cognition)

genre). Si ces taux sont loin d'être égaux, il s'agit d'une preuve potentielle d'un système de prédiction au comportement « injuste », c'est-à-dire, avec un biais clair sur la façon dont les décisions sont prises en fonction des valeurs de cette caractéristique. La parité démographique stipule que tous les groupes résultant des différentes valeurs d'une classe protégée (par exemple, le genre) devraient recevoir le même taux de résultats positifs.[2]

La question de l'analyse de biais dans des systèmes de TAL a émergé ces dernières années. Parmi ces études, nous distinguons deux grandes approches : l'évaluation des biais dans des espaces de plongements lexicaux [Bolukbasi et al., 2016, Caliskan et al., 2017, Gonen and Goldberg, 2019] ainsi que des travaux sur des biais relatifs à des tâches comme la traduction automatique [Wisniewski et al., 2021], l'analyse de sentiments [Thelwall, 2018], ou encore la résolution de coréférences [Alfaro, 2019]. Bien que ces travaux constituent un enjeu crucial dans l'analyse et la critique des systèmes de TAL et dans la mise en évidence des comportements négatifs que peuvent avoir ces systèmes, aucun consensus sur le concept de biais n'existe en apprentissage automatique, et les définitions sont parfois très éloignées de ce qui est présenté en sociologie. La définition du terme constitue donc un « parti pris » dès le début de l'étude : quels types de comportements sont préjudiciables, de quelle manière, à qui et pourquoi ? Par exemple, deux types de stéréotypes sont définis par Blodgett et al. [2020] :

1. Les stéréotypes qui propagent une généralisation négative sur des groupes sociaux particuliers (tels que stéréotypes de genre, stéréotypes ethniques, d'âge, etc.).

2. Les différences de performances de systèmes de TAL pour différents groupes sociaux.

Nous nous intéressons à la première catégorie de stéréotypes, et plus précisément aux stéréotypes de genre. Parce que le concept de « biais » est flou dans la littérature (parce que mal défini ou non défini) [Blodgett et al., 2020], nous utiliserons ici le terme « stéréotype » défini en sociologie comme « *une image préconçue, une représentation simplifiée d'un individu ou d'un groupe humain. Il repose sur une croyance partagée relative aux attributs physiques, moraux et/ou comportementaux, censés caractériser ce ou ces individus.* ». Ici, nous étudions les stéréotypes de genre dans un corpus de théâtre français du XVI^{e} au XIX^{e} siècle, qui présente l'avantage d'amplifier les stéréotypes existant dans le monde réel à des fins de critique sociale ou de comédie.

[2]Par exemple, si le système décide d'embaucher des personnes avec le même taux pour les femmes et les hommes, alors le système fait preuve de parité démographique.

5.4 Les stéréotypes au théâtre

Le théâtre se répartit en deux genres majoritaires : la tragédie et la comédie. Tandis que le premier, souvent illustré par Racine ou Corneille, est placé au sommet de la hiérarchie (avec son exigence de vraisemblance et ses convenances), la comédie française, représentée par Molière, est moins codifiée et valorisée. Durant le siècle des Lumières, le théâtre subit une révolution : « *la comédie, sous ses apparences de gaieté, de légèreté et de fantaisie, dénonce les injustices sociales, renverse les hiérarchies et tend vers la critique ou la satire* » [Marcandier, 2011]. Au contraire, la tragédie laisse place au drame bourgeois visant à « [...] *consacrer leurs écrits à des situations réalistes, proches du public bourgeois, à des intrigues reposant sur des conflits familiaux* ». Étant donné que les deux genres sont construits pour des publics différents, les stéréotypes présents dans ces œuvres le sont également. Gruffat [2012] étudie la représentation des héros classiques dans la tragédie et démontre que le héros tragique amoureux était le produit d'une époque (il parle ici d'un changement brutal des hommes, chez Racine comme chez Pradon, où le caractère guerrier laisse place à une faiblesse et à une transformation profonde des héros amoureux). Le théâtre constitue, comme toutes les œuvres littéraires et artistiques, un miroir de la société de l'époque à laquelle il renvoie. Ces écrits nécessitent d'être replacés dans leurs contextes culturels et sociaux afin de comprendre leurs impacts et d'analyser leur contenu. Selon la metteuse en scène Myriam Marzouki, la lourdeur des stéréotypes de genre est plus marquée dans le répertoire classique que dans le répertoire contemporain. Selon elle, « *les personnages de femmes dans le théâtre classique sont épouses, filles, mères ou servantes, leur rapport au monde étant toujours médiatisé par leur lien avec un homme.* » [3]. Dans cette étude, l'objectif est de déterminer dans quelle mesure des modèles de plongements lexicaux sont en mesure de mettre en avant les stéréotypes femmes/hommes présents dans les pièces de théâtre.

5.5 Corpus et Modèles

Pièces de théâtre. Ces travaux ont pour objectif de mettre en évidence la présence ou non de stéréotypes de genre dans des modèles de plongements lexicaux. Pour cela, nous élaborons un corpus de pièces de théâtre, contenant de nombreux stéréotypes associés au genre théâtral. Ce corpus a pour avantage de représenter un genre littéraire pour lequel les stéréotypes sont très marqués. Si les modèles ne parviennent pas à capturer des stéréotypes dans ce corpus, nous formulons

[3] Rapport d'information n° 704 (2012-2013) de Mme Brigitte Gonthier-Maurin, fait au nom de la délégation aux droits des femmes du Sénat. « La place des femmes dans l'art et la culture : le temps est venu de passer aux actes », 2013.

l'hypothèse qu'ils seraient difficiles à observer dans d'autres corpus. Le détail de la collecte et des statistiques concernant ce corpus est présenté dans la Section 3.2. Les huit-cent dix-sept pièces de théâtre contenues dans le corpus ont été publiées du XVI^{e} au XIX^{e} siècle.

Modèles de plongements lexicaux. Dans cette étude, nous souhaitons effectuer une analyse qualitative de la spécificité des stéréotypes de genre capturés par les modèles de plongements lexicaux statiques, si ceux-ci en capturent. Pour cela, nous nous servons des modèles Word2Vec [Mikolov et al., 2013]. L'avantage d'utiliser cette famille de modèles, par rapport à GloVe ou à FastText, est que plusieurs versions sont disponibles en ligne pour le français, et qu'ils sont faciles à ré-entraîner. Nous souhaitons comparer l'apprentissage de Word2Vec sur des données Wikipédia et sur des données Web avec l'apprentissage du modèle sur les corpus de pièces de théâtre. Nous émettons l'hypothèse que les stéréotypes seront plus marqués dans le théâtre que dans les autres corpus, étant donné qu'ils sont plus assumés et directs dans ce corpus. Des caractéristiques spécifiques au théâtre devraient émerger, comme la description de traits de caractères liés aux personnages, et nous observerons les caractéristiques spécifiques au genre. À l'inverse, nous émettons l'hypothèse que le corpus issu du Web devrait contenir des stéréotypes très actuels et qui ont à trait à la société (par exemple, les attendus sociaux des hommes et des femmes par rapport au travail ou à l'éducation). Enfin, le corpus de Wikipédia devrait être le plus nuancé et contenir moins de stéréotypes, étant donné qu'il traite de sujets plus factuels. Cette hypothèse est formulée en « prenant des pincettes » car nous avons formulé que les stéréotypes étaient inconscients, ils devraient donc être représentés dans le corpus. Nous exploiterons les modèles développés par Fauconnier [2015] et présentés dans la Section 4.3.1 de ce manuscrit.

5.6 Expériences

5.6.1 Stéréotypes de genre

Voisinage des termes « femme » et « homme » Afin de déterminer quels stéréotypes de genre sont spécifiques aux pièces de théâtre, nous comparons des modèles appris sur trois sources de données : deux sources généralistes (Web et Wikipedia) et les pièces de théâtre. Les plus proches voisins sont les termes ayant la similarité cosinus la plus grande avec le terme source. Le Tableau 5.1 présente une vision synthétique des cinq termes les plus similaires à « femme » et « homme ». Tout d'abord, les modèles renvoient des résultats plutôt similaires lorsque l'on s'intéresse au terme « femme », alors qu'ils sont bien plus variés pour le terme « homme ». Une

	Modèles pré-entraînés		**Modèles *from scratch***		
	Wikipédia	Web	C+T	Comédie	Tragédie
femme	fille, maîtresse, prostituée, fiancée, servante	fille, mari, jeune, prostituée, adolescente	mari, belle-sœur, vindicative, accoucha, veuve	mari, veuve, feue, épousé, vertueuse	mari, veuve, ressource, belle-mère, comtesse
homme	diplomate, femme, politicien, avocat, agriculteur	humaine, galop, humanité, nationalisme, sabre	garçon, hébété, désintéressé, aventurier, dissipateur	rustre, coureur, moine, joueur, bravoure	méchant, fou, franc, président, raisonnable

Table 5.1: Cinq plus proches voisins des mots « femme » et « homme » en utilisant des modèles Word2Vec. La distance cosinus est utilisée pour récupérer les mots les plus similaires. C : Comédie et T : Tragédie

faible richesse lexicale signifie que l'utilisation du terme « femme » sur le Web, sur Wikipédia et dans les pièces de théâtre est très similaire, et donc très stéréotypée. Nous montrons que les termes associés à « femme » sont principalement axés sur le rapport social qu'elles entretiennent par rapport aux hommes (e.g., *épousée, fille, maîtresse, veuve, fiancée*), sur leur âge et leur physique (e.g., *jeune, adolescente, beauté*) ou sur leur sexualité (*prostituée, maîtresse, vertueuse*). Les termes les plus similaires à « femme » renvoient principalement à leur rôle dans le contexte familial (e.g., *mère, tante, julie, épouse, mari*). Au contraire, nous observons une variété lexicale plus importante entre les modèles pour le terme « homme ». Sur Wikipédia, nous observons une prépondérance de termes qui renvoient aux métiers (e.g., *diplomate, politicien, avocat*). En revanche, deux champs lexicaux apparaissent sur le Web : le champ lexical relatif à la guerre (*galop, sabre*) et celui de « l'Homme » (*humanité, humaine*). Le deuxième est dû au passage en minuscule lors de l'apprentissage. Sur les pièces de théâtre, nous obtenons des termes typiques du genre (en comparaison avec les autres modèles). Par exemple, dans la comédie, différents personnages types sont mis en avant (tels que *coureur, joueur, rustre, bravoure*). Les personnages étant traités différemment en tragédie, le vocabulaire identifié est plutôt lié à des personnages odieux au statut social important. Nous remarquons ici que, contrairement au terme « femme », il n'y a pas de termes liés au champ lexical de la famille dans le voisinage du mot « homme ».

Homme	**Femme**	**Indéterminé**
Fierté, colère	Joie, soulagement, espoir, surprise, compassion, tristesse, honte, intérêt	Jalousie, dégoût, mépris, peur

Table 5.2: Tendance majoritaire attribuée aux émotions [Raymondie and Steiner, 2020]

Expression des émotions Dans cette partie, nous cherchons à déterminer quelles émotions sont les plus utilisées dans le contexte de « femme » et « homme ». Nous reprenons les émotions présentées par Raymondie and Steiner [2020] qui étudie le « genre des émotions » dans le contexte du travail. Le Tableau 5.2 présente la tendance d'attribution majoritaire observée pour quatorze émotions. La Figure 5.1 présente la proximité calculée entre les émotions et les termes « homme » et « femme ». Les émotions sont associées aux femmes et aux hommes très différemment en fonction du corpus. Nous notons un accord entre les trois modèles sur quatre émotions (surprise, peur, soulagement et honte), qui sont toutes attribuées aux femmes, avec une légère différence par rapport aux hommes, ce qui est corroboré avec l'étude d'origine. Dans les pièces de théâtre, trois émotions sont principalement associées aux hommes (fierté, colère et compassion). Bien que les deux premières soient attendues et soutiennent l'hypothèse que les stéréotypes sont renforcés dans le théâtre, la présence de la troisième est surprenante. Le terme « compassion » dans le corpus de théâtre est principalement utilisé dans trois contextes : un personnage parle d'un homme sans compassion[4], un personnage demande à un homme de faire preuve de compassion[5] ou un personnage (souvent féminin) éprouve de la compassion pour un homme[6]. Cette émotion n'est donc pas associée aux hommes parce qu'ils en font preuve mais plutôt parce qu'il s'agit d'une problématique récurrente dans ces œuvres. De façon générale, les stéréotypes de genre associés aux émotions dans le théâtre correspondent à ceux qui sont attendus. Nous notons qu'avec le modèle appris sur des données Web, les émotions « compassion » et « mépris » sont associées aux hommes, bien que contradictoires. Nous supposons que ces émotions ne sont pas uniquement employées en tant que qualificatif mais aussi comme requête (i.e., dans des contextes où on implorerait la compassion). De plus, étant donné que les hommes étaient souvent associés à « l'Homme » dans ce modèle, il se peut que ces adjectifs soient utilisés pour parler de l'humanité.

5.6.2 Voisinage de « personnages types »

Nous nous intéressons à la représentation de termes traditionnellement attribués à des personnages types au théâtre. L'objectif de cette étude est de déterminer si les modèles vectoriels parviennent à capturer les stéréotypes associés à ces rôles. Étant

[4] *« Il est toujours insensible et froid, l'homme qui se refuse à tes feux : son cœur qui s'isole se durcit, il n'est plus disposé à la compassion ni à la pitié. »* - « Les Tombeaux de Vérone », Louis-Sébastien Mercier (1782)

[5] *« Ami, qui que tu sois, si ton âme sensible à la compassion peut se rendre accessible, un jeune gentilhomme implore ton secours ; »* - « Clitandre », Corneille (1631)

[6] *« Petit bonhomme, veuille le ciel que vous ne vous trompiez pas, et que ce soit mon semblable que j'embrasse dans une créature pourtant si méconnaissable ! Vous me pénétrez de compassion pour vous. »* - « L'Île de la raison », Marivaux (1727)

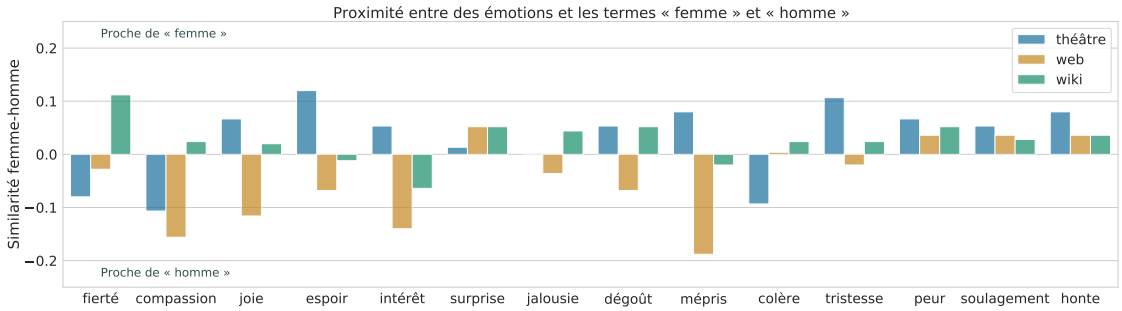


Figure 5.1: Similarité cosinus entre des émotions et les termes « femme » et « homme ». La similarité est calculée entre le vecteur d'une émotion $v_{\text{émo}}$ et les vecteurs de « femme » v_f et « homme » v_h avec la formule : $proximité = cos(v_{\text{émo}}, v_f) - cos(v_{\text{émo}}, v_h)$. Une proximité de 1 signifie que l'émotion est exclusivement associée à « femme » et une proximité de -1, qu'elle est associée à « homme »

Mot	Comédie	Tragédie
roi	prince, successeur, monarque, empereur, ambassadeurs	prince, nom, couronne, choix, droits
reine	souveraine, héritière, prisonnière, orgueilleuse, ingrate	faveur, honteux, princesse, hymen, jaloux
valet	coquin, vilain, fou, fripon, drôle	fripon, faquin, impertinent, domestique, arlequin
servante	demoiselle, andouille, soubrette, impertinente, gueuse	commère, minette, chonchon, chienne, hymen
maître	brave, scélérat, instruit, successeur, imposteur	valet, laquet, riche, province, méchant
enfant	bâtard, riche, jeune, mari, femme	veuve, gendre, comte, nièce, orphelin

Table 5.3: Cinq plus proches voisins (distance cosinus) de personnages types (modèles pré-appris et entraînés sur le corpus de théâtre) pour les genres comédie et tragédie

donné que les personnages types ont des rôles prépondérants et très caricaturaux, et qu'ils sont associés à des traits de caractères forts, nous nous attendons à retrouver de fortes différences dans la sémantique qui leur est associée avec les modèles vectoriels. De plus, les personnages au théâtre contiennent, à l'image de la vie réelle, des stéréotypes de genre assez forts qui orientent leurs caractéristiques. C'est pourquoi, nous nous attachons à distinguer les personnages types féminins et masculins dans cette étude. Cela nous permet d'étudier deux types de préjugés : ceux associés à des classes sociales (représentées par les statuts des personnages) et ceux associés au genre des personnages. Étant donné que les personnages ne sont pas traités de la même façon en comédie et en tragédie (nous nous attendons à des stéréotypes plus marqués en comédie), nous analysons le vocabulaire associé aux personnages

pour chaque genre. Zaragoza [2006] évoque plusieurs personnages types du théâtre. Pour cette étude, nous avons sélectionné des antagonistes pouvant être analysés par paire : roi et reine, valet et servante [Gunny, 1978], et maître (par opposition au valet) [Maija, 2012]. Nous avons ajouté le rôle de l'enfant afin de comparer le vocabulaire qui lui est attribué par rapport aux personnages adultes. Les résultats de l'étude sont présentés dans le Tableau 5.3.

Dans la comédie, le valet est un personnage clé, parfois qualifié de « maître du jeu » [Da, 2009]. La sémantique retrouvée dans les modèles est très spécifique de ce qui est attendu du personnage (*coquin, vilain, fou, fripon, drôle*). Dans la tragédie, Naugrette [2003] illustre le rôle de valet avec Ruy Blas : « *il contrevient à la loi, se révolte devant le sort qui lui est fait, et en meurt.* » Le modèle Word2Vec renvoie à la première définition du valet, similaire à la comédie (*fripon, faquin, impertinent*) allant jusqu'à relier le terme « arlequin » au rôle du valet. De même, la sémantique associée à la servante renvoie à son rôle emblématique de personne intrigante et commère dans la comédie alors qu'elle est plutôt renvoyée à son rôle marital dans la tragédie, tout comme la reine (le terme *hymen* apparaît pour ces deux rôles[7]). Nous observons que le vocabulaire du « roi » et de la « reine » demeurent spécifiques à leur genre. Bien que le roi apparaisse comme un dirigeant de haut rang, la reine est réduite aux aspects négatifs de sa personnalité (e.g., *orgueilleuse, ingrate*). Enfin, le maître est associé au statut de bourgeois riche et instruit avec des stéréotypes typiques de cette classe sociale (e.g., *méchant, imposteur*). La sémantique liée à l'enfant est très proche du celle de la femme, avec des termes du champ lexical de la famille utilisés dans le même contexte que les femmes[8].

5.7 Synthèse

Dans cette étude, nous avons analysé les stéréotypes de genre dans un corpus de pièces de théâtre en utilisant des modèles de plongements lexicaux statiques. Nous avons confirmé nos hypothèses de travail quant à la capture marquée des stéréotypes dans un corpus de pièces de théâtre que nous avons constitué pour l'expérience. De plus, nos travaux confirment notre deuxième hypothèse, et démontrent que les plongements lexicaux modélisent des stéréotypes fins et différents en fonction des données d'apprentissage. Nous observons ainsi des différences systématiques entre la sémantique utilisée pour qualifier les femmes et les hommes. Alors que les hommes sont perçus différemment sur Wikipédia, le Web et dans les pièces de théâtre, les

[7] En poésie comme en littérature, le terme hymen renvoie à l'union et au mariage. (Source :TLFi https://www.cnrtl.fr/definition/hymen)

[8] « *Vous seul nous arrachant à de nouvelles flammes nous avez fait laisser nos enfants et nos femmes.* » - « Iphigénie », Jean Racine (1794).

femmes sont qualifiées de façon très similaire dans ces corpus et souvent réduites au champ lexical de la famille et à leur sexualité. Nous relevons que les enfants, dans les pièces de théâtre, sont souvent mentionnés lorsqu'il s'agit de femmes et sont d'ailleurs très proches dans l'espace vectoriel. Au contraire, les hommes possèdent un champ lexical différent et plus varié (en fonction du genre de pièces). Nous relevons également que les stéréotypes liés aux émotions « genrées » peuvent être retrouvés facilement dans ce corpus. Enfin, nous avons observé que certains rôles de personnages-types au théâtre contiennent une sémantique attendue (les rôles de valet et servante sont très représentatifs de leurs attributs dans les comédies).

Nous avons mis en évidence des stéréotypes de genre dans les pièces de théâtre, ce qui montre que des modèles de plongements lexicaux statiques suffisent à révéler des stéréotypes et des biais d'apprentissage dans les modèles. Il serait intéressant de poursuivre ces travaux avec des modèles de représentation contextuels, afin de déterminer si l'amélioration globale des performances sur différentes tâches conduit au renforcement des stéréotypes dans les représentations. Divers travaux ont été conduits dans ce sens, comme la démonstration que le modèle GPT-3, créé pour de la génération automatique de textes, capturaient de nombreux stéréotypes de genre, d'ethnie et de religion. Par exemple, les termes « violent », « terrorisme » et « terroriste » (traduits de l'anglais à partir des exemples sources), étaient associés bien plus fortement à la religion musulmane qu'aux autres. Il est intéressant de noter que ChatGPT[9], un modèle de génération de textes plus récent, contient des filtres qui empêchent le système de générer des contenus discriminatoires. D'après les tests que nous avons réalisé sur la plateforme disponible en ligne, il semblerait que le filtre est déclenché par l'utilisation de certains mots-clés et qu'il peut être contourné grâce à diverses reformulations. Cependant, cela démontre que les travaux d'analyse de biais sociaux sont utiles et nécessaires à la création d'outils responsables.

Nous avons exploité des travaux de sciences cognitives et de littérature pour guider l'analyse des corpus. Nous pensons qu'il est important, sur des sujets qui touchent à des concepts de société, d'exploiter des connaissances théoriques d'autres domaines, afin de formuler des hypothèses de travail rigoureuses.

[9] https://chat.openai.com/

J'ai toujours été fier du talent que je possède pour tourner des phrases. Et les mots sont à mon avis, qui n'est pas si humble, notre plus inépuisable source de magie. Ils peuvent à la fois infliger des blessures et y porter remède.

— Harry Potter et les reliques de la mort

6

La représentation des termes hors-vocabulaire

Table des Matières

6.1 Introduction

Les modèles de langue permettent de représenter n'importe quel mot dans l'espace multidimensionnel, indépendamment du fait qu'ils aient été vus ou non durant l'apprentissage. Néanmoins, cela nécessite une contrainte forte fixée avant la phase d'apprentissage : un vocabulaire de tokens de taille fixe doit être appris. Cette procédure permet de représenter n'importe quel mot par des sous-tokens construits durant l'apprentissage du modèle (voir Section 2.5.1.1). Les modèles Transformer ont généralement un vocabulaire compris entre 30.000 et 50.000 tokens, fixé avant leur apprentissage. Par conséquent, la plupart des mots qui existent dans les données d'apprentissage n'apparaîtront pas dans le vocabulaire final du

modèle. En théorie, cela ne devrait pas poser de problèmes, car on considère qu'un mot est toujours découpé de la même façon pour un modèle donné, et donc représenté par les mêmes sous-tokens. Par exemple, le mot « cardiologie », inconnu d'un modèle de tokenisation, pourrait être segmenté en sous-mots de taille 1 à 10 vus durant l'apprentissage. Nous pourrions donc obtenir les tokens « car+dio+lo+gie », qui proviendraient de mots ou de sous-mots fréquents ayant été vus durant l'apprentissage du modèle (grâce à l'apprentissage de mots comme « car », « radio », « sociologique », etc.). Dans ce chapitre, nous verrons que cette hypothèse a ses limites et que le découpage de certains mots empêche une représentation pertinente du point de vue sémantique.

6.2 Définition : les termes hors-vocabulaire

Nous distinguons deux catégories de termes hors-vocabulaire dans cette étude : les termes hors-vocabulaire issus de la langue générale et les termes hors-vocabulaire spécifiques à un domaine. Dans un premier temps, nous étudions les termes de la langue générale afin d'analyser comment des formes classiques de la langue française sont segmentées par les modèles de tokenisation statistique (e.g., les formes conjuguées, les dérivations, etc.). Nous analyserons ensuite l'entourage de ces mots afin d'identifier l'impact que la tokenisation a dans leur représentation (voir Chapitre 2.5). Dans un deuxième temps, nous étudierons les termes hors-vocabulaire spécifiques à des corpus. Pour cela, nous distinguons trois types de termes hors-vocabulaire : les termes spécifiques à un domaine (e.g., domaine de l'énergie), les termes mal orthographiés (i.e., fautes d'orthographes involontaires et erreurs d'océrisation) et les termes spécifiques à un mode de rédaction (i.e., le format spécifique des courriels qui introduit des termes qui ne sont pas utilisés par ailleurs).

6.3 Jeux de données et Modèles

Jeux de données. Nous choisissons quatre corpus de spécialité contenant des termes hors-vocabulaire de nature différente parmi ceux présentés dans le chapitre 3 :

1. Bio-Gallica : le jeu de données a été collecté à partir de la bibliothèque numérique GALLICA. GALLICA contient un grand nombre de documents historiques tels que des livres ou des articles de presse et permet le téléchargement de documents océrisés. Nous avons extrait des documents rédigés en français pour le *« Journal de Microbiologie »*. Les documents sont donc très bien rédigés et contiennent deux types de termes hors-vocabulaire de spécialité : des termes mal orthographiés à la suite de l'océrisation (e.g.,

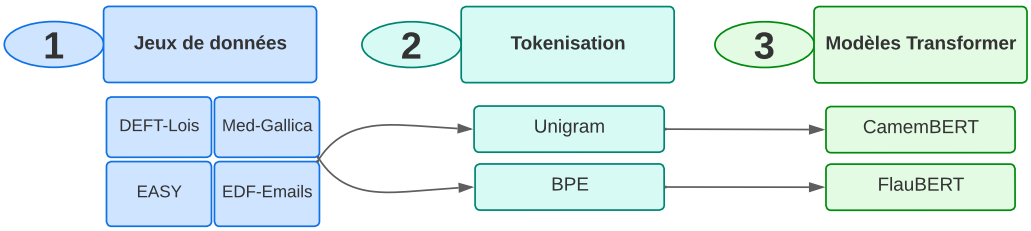


Figure 6.1: Expériences de comparaison qualitative entre les modèles

« dperme » au lieu de « derme » ou « typhïde » au lieu de « typhoïde ») et des termes du domaine de la biologie (e.g., « bacteriologique » au lieu de « bactériologique »). Avec ce jeu de données, nous pouvons étudier l'impact des erreurs d'océrisation sur la représentation des termes hors-vocabulaire.

2. DEFT-Lois : le jeu de données contient des articles de lois. Nous avons repéré de nombreux termes mal orthographiés. Ces erreurs sont dues à des problèmes de transcription du texte original ou à des erreurs d'extraction. Les fautes les plus fréquentes sont des erreurs orthographiques de type suppression d'accents.

3. EASY : le jeu de données est constitué d'échanges de courriers électroniques entre des collègues d'une entreprise. Ils sont rédigés de manière formelle et informelle, en fonction du sujet de la conversation et des interlocuteurs. Ce jeu de données est pertinent pour cette étude, car il introduit des termes hors-vocabulaire spécifiques au format du courriel (e.g., « cordialement ») ainsi que des usages (abbréviations). Cela permet d'effectuer l'analyse de ces termes dans des textes du domaine général.

4. EDF-Courriels : le jeu de données de courriers électroniques des clients de l'entreprise Électricité de France (EDF) est privé et anonyme. Il contient un mélange des termes hors-vocabulaire rencontrés dans les corpus précédents : des erreurs syntaxiques (e.g., ordre des mots, fautes de conjugaison), des abréviations (i.e., du langage général ou de spécialité) et des termes spécifiques au domaine de l'énergie (i.e., dans le contexte des courriels électroniques de clients, il s'agit principalement d'offres ou de tarifs EDF, ainsi que des termes relatifs aux relevés de compteurs électriques).

Modèles de langue Dans cette étude, nous utiliserons plusieurs modèles Transformer, présentés dans le Chapitre 4, suivant l'objectif des expériences réalisées. Pour l'évaluation intrinsèque des modèles de langue (voir Section 8.2), nous

souhaitons réaliser des expériences à deux niveaux de comparaison pour le français (voir Figure 6.1) :

1. l'algorithme de tokenisation : pour cela, nous utilisons FlauBERT qui utilise l'algorithme BPE et CamemBERT qui utilise l'algorithme Unigram.

2. les données d'entraînement des modèles de langue : nous comparons quatre modèles CamemBERT ayant été entraînés sur trois jeux de données différents (i.e., Wikipédia, OSCAR et CCNET) sur un petit volume de données ou un plus large.

6.4 La représentation des termes hors-vocabulaire

Termes hors-vocabulaire. La proportion des termes-hors vocabulaire dans chacun des corpus est comptabilisée dans la Figure 6.2. Pour cela, nous nous basons sur les résultats générés par le tokéniseur des modèles sur chaque corpus. Plus précisément, notre approche est fondée sur le calcul du nombre de tokens obtenus pour chaque mot du vocabulaire. Plus le nombre de tokens obtenu pour un mot est grand, plus la segmentation aura tendance à s'éloigner de la morphologie naturelle du mot telle qu'attendue en linguistique. En effet, étant donné que le découpage des mots est purement statistique, quel que soit l'algorithme utilisé (voir Section 2.5.1.1), un mot découpé de nombreuses fois perdra sa morphologie et la sémantique qui lui est associée. Nous observons que pour les jeux de données contenant des courriers électroniques (i.e., EASY et EDF-Courriels), il existe des différences significatives entre le vocabulaire d'entraînement Wikipédia et les vocabulaires issus du Web. Nous l'expliquons par le fait que le vocabulaire de Wikipédia est plus restreint que celui du Web. Cependant, cette différence n'apparaît pas sur les jeux de données DEFT et Bio-Gallica, qui semblent être aussi similaires aux vocabulaires de Wikipédia et du web. En revanche, les corpus de courriels sont bien plus proches des données du web que de Wikipédia. Nous concluons donc que les formules de politesse ou de structure spécifiques aux courriers électroniques sont plus présentes sur le web (e.g., forums) que sur des articles Wikipédia. En comparant les corpus, nous observons qu'EASY est le plus proche des jeux de données d'apprentissage des modèles, ce qui fait sens étant donné qu'il s'agit de textes appartenant au domaine général. De plus, le jeu de données Bio-Gallica est celui dont les mots sont les plus segmentés par les tokéniseurs. Cela nous semble logique puisque les mots spécifiques au domaine de la biologie ont tendance à être assez longs et composés de beaucoup de sous-mots (e.g., noms de molécules, pathologies, anatomie). Enfin, nous observons qu'il n'y a pas de différences significatives entre les modèles ayant appris sur le Web par rapport à nos jeux de données. Cela implique que l'étape

de prétraitement des données Web pour CCNet et OSCAR n'a pas d'impact sur nos jeux de données. De plus, cela signifie que l'augmentation de la taille du vocabulaire entre les modèles Base et Large n'a pas d'impact significatif sur la compréhension de ces corpus. Afin de mieux comprendre ce qui est contenu dans les termes hors-vocabulaire, nous analyserons quelques motifs de segmentation de termes qui apparaissent fréquemment dans le domaine général (voir Paragraphe 6.4) et d'autres, spécifiques au domaine de l'énergie (voir Paragraphe 6.4).

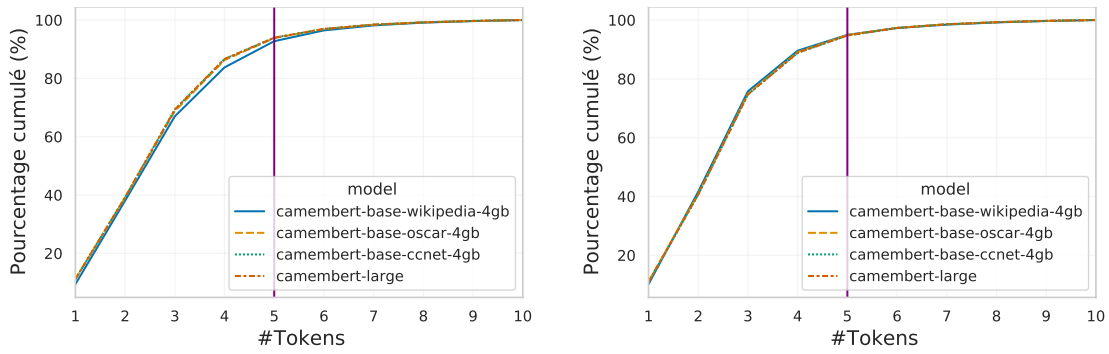
Les termes hors-vocabulaire du domaine général L'objectif de cette étude est de repérer des motifs de tokenisation des termes hors-vocabulaire de la langue générale puis de déterminer si cette tokenisation semble avoir un impact significatif sur la représentation de ces termes. Durant l'apprentissage d'un tokéniseur, la contrainte forte d'un vocabulaire de tokens fixe est nécessaire. Cela signifie que fréquemment, les mots comportant certaines flexions (i.e., conjugaison, déclinaison) sont découpés en conservant la racine reconnue du mot et sa flexion, parce que ces racines et flexions ont été identifiées dans d'autres mots du corpus d'apprentissage. Cela est bien souvent le cas avec les marqueurs de pluriels en -s ou les formes conjuguées. Nous distinguons trois types de mots construits dans cette étude (bien qu'il en existe d'autres en grammaire) :

1. Les mots dérivationnels qui sont créés à la suite de l'ajout d'un affixe (i.e., préfixe ou suffixe) ayant modifié le sens du mot. *Par exemple : le mot « réaménagement » est issu de « ménage », de l'ajout du préfixe ré- et du suffixe -ment.*

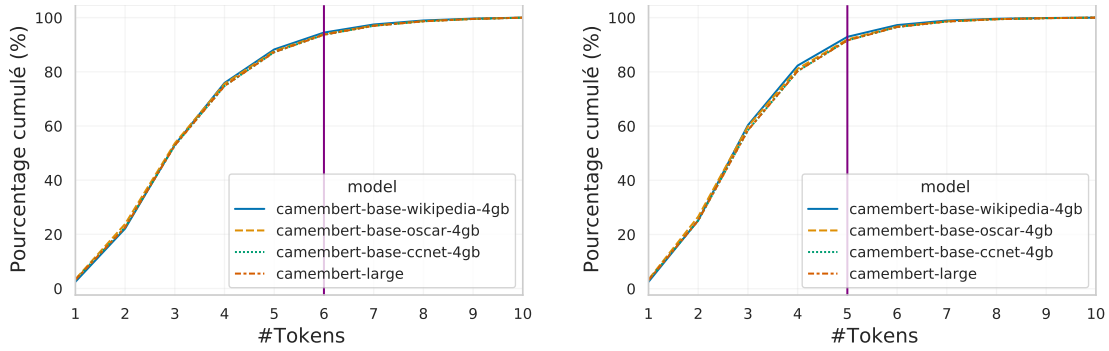
2. Les mots composés qui contiennent des mots autonomes qui existent d'eux-mêmes. *Par exemple : le mot « pomme-de-terre » est issu de « pomme », « de » et « terre ».*

3. Les mots contenant des désinences (i.e., conjugaison, dérivation, variations en genre et en nombre). *Par exemple : le verbe « vivrons » est issu du verbe « vivre » et de la désinence -ons.*

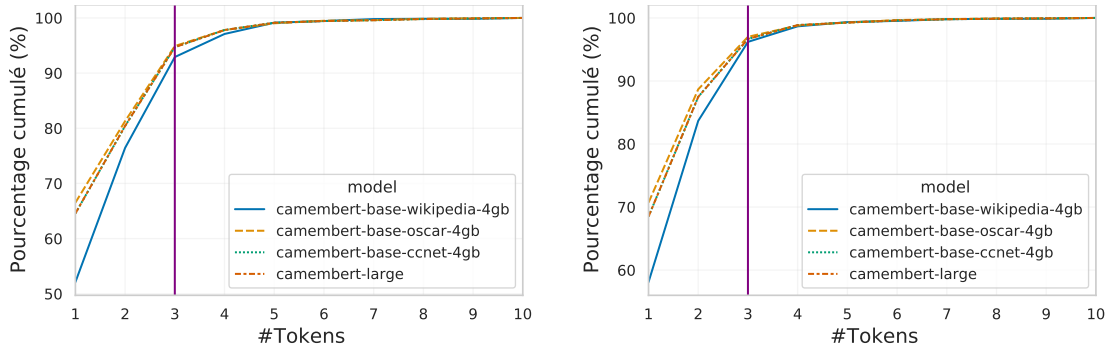
Nous présentons quelques exemples de tokenisation de certains de ces mots (voir Tableau 6.1), issus du corpus EDF-Courriels. La tokenisation a été réalisée avec le modèle CamemBERT ayant appris sur les données Wikipédia. Comme nous pouvons le voir, les mots sont parfois segmentés de la même manière que la segmentation morphologique du mot. Les mots présentés dans cet exemple ont une racine qui existe dans le vocabulaire du modèle, et ne devraient donc pas poser de problèmes durant la phase d'encodage. Ensuite, grâce à un calcul de similarité cosinus, nous pouvons récupérer les mots les plus similaires à ces termes dans le corpus EDF-Courriels (voir Tableau 6.2). Nous observons que les variantes conjuguées des verbes



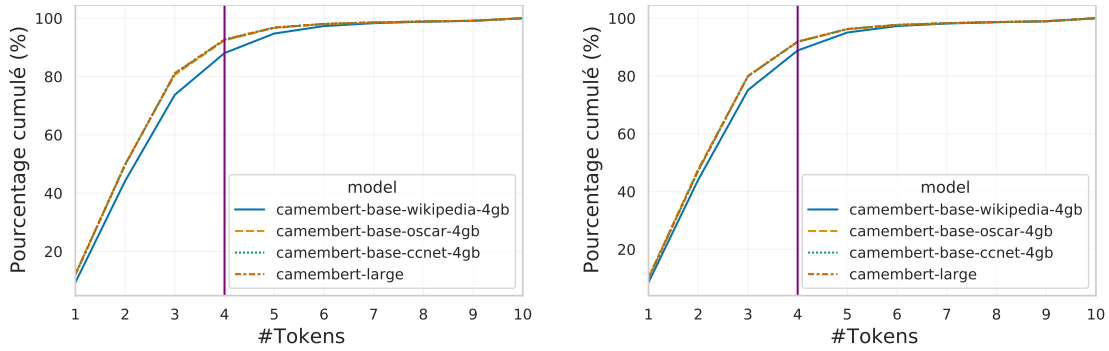
(a) DEFT



(b) Bio-GALLICA



(c) EASY



(d) EDF-Courriels

Figure 6.2: Pourcentage cumulé du nombre de sous-tokens obtenus pour chaque mot du vocabulaire. Les expériences à gauche sont menées sur les jeux de données bruts et les expériences à droite sur les jeux de données après lemmatisation. La droite verticale violette représente le seuil à partir duquel 90% du vocabulaire est traité

sont regroupées entre elles, ainsi que les variantes orthographiques d'un même mot. Cela n'est pas un problème en soi, mais il semblerait que les mots apparaissent regroupés en fonction de leur tokens plutôt qu'en fonction de leur contexte (on retrouve des formes nominales avec des formes verbales). Par exemple, le terme télétravail, découpé par le modèle en deux tokens (i.e., télé et travail) se retrouve à proximité, dans l'espace de représentation multidimensionnel, de tous les mots contenant le préfixe « télé », ce qui du point de vue sémantique, ne fait pas toujours sens. Dans nos travaux, nous souhaitons quantifier l'impact de la tokenisation sur la représentation des termes afin de valider ou non cette hypothèse.

Mot	**Morphologie**	**Flexion**	**Tokenisation**
bonjour	bon + jour	Composition	bon + jour
télétravail	télé + travail	Composition	télé + travail
hésitez	hésit- + ez	Désinence (conjug.)	hésite + z
contacter	contact- + er	Désinence (conjug.)	contact + er
consolider	con + -solid- + er	Désinence (conjug.)	cons + oli + der
e-mails	e + - + mail + s	Désinence (flexion)	e + - + mail + s
mensuellement	mensuelle- + ment	Dérivation	mensuelle + ment
cordialement	cordiale- + ment	Dérivation	cord + iale + ment

Table 6.1: Exemples de tokenisation des mots avec le modèle *camembert-base-wikipedia-4gb*. Conjug. : conjugaison

Les termes hors-vocabulaire de spécialité Pour cette analyse, nous nous concentrons sur le jeu de données EDF-Courriels, car il contient à la fois des termes spécifiques au domaine de l'énergie, au formalisme du courriel et à un registre de rédaction courant voire familier ce qui implique des fautes d'orthographe. Nous sélectionnons un nombre restreint de termes spécifiques à ce corpus (voir Tableau 6.4) et nous observons que certains termes sont découpés de façon à perdre toute sémantique liée à leur morphologie. Par exemple, le nom propre « Linky » est segmenté en sous-unités d'une à deux lettres par les modèles de tokenisation pré-entraînés. Dans l'ensemble, les mots appartiennent plus au vocabulaire des modèles ayant appris sur des données Web que sur des données Wikipédia. Si notre hypothèse selon laquelle l'étape de tokenisation a un rôle essentiel sur la représentation des mots s'avère correcte, les petits mots découpés en sous-unités incohérentes seront représentés de façon non optimale dans l'espace multidimensionnel. Nous nous intéressons, comme précédemment, au voisinage de ces termes pour tester cette hypothèse (voir Tableau 6.3). Tout d'abord, nous observons que les termes hors-vocabulaire segmentés par les tokéniseurs sont systématiquement rapprochés de termes ayant des sous-mots en commun avec eux. Cela fonctionne d'un point de vue

Mot	Voisins	Tokenisation
bonjour	bonjou	bon + jou
	bonour	bon + our
	bonjourmerci	bon + jour + mer + ci
	bonjours	bon + jour + s
	bonjjour	bon + j + jour
télétravail	téléopération	télé + opération
	téléfact	télé + fact
	télécharge	télé + charge
	téléopérateur	télé + opérateur
	télérelève	télé + rel + ève
hésitez	hésiterai	hésite + rai
	hésiterons	hésite + rons
	hésiterais	hésite + s
	hésites	hésite + s
	hésiterez	hésite + rez
contacter	contactera	contact + era
	contactez	contact + ez
	contactent	contact + ent
	contactons	contact + ons
	contacte	contact + e
consolider	structurer	structure + r
	résorber	ré + s + or + ber
	formaliser	form + alis + er
	complèter	complète + r
	matérialiser	mat + éri + alis + er
e-mails	emails	e + mail + s
	e-mail	e + - + mail
	courriers	courrier + s
	e-mailadresse	e + - + mail + adresse
	email	e + mail
mensuellement	mensuelles	mensuelle + s
	mensuelement	mensuel + ement
	mensuele	mensuel + e
	trimestrielle	trimestriel + le
	annuels	annuel + s
cordialement	cordialemement	cord + iale + me + ment
	cordialment	cord + ial + ment
	cordialemment	cord + iale + m + ment
	cordialemnt	cord + ial + ment
	cordiales	cord + iale + s

Table 6.2: Voisinage de termes issus du domaine général avec le modèle *camembert-base-wikipedia-4gb*. Les voisins correspondent aux termes les plus proches en terme de distance cosinus

Mot	Wikipedia	Oscar	CCNet
compteur	compteeur	disjoncteur	boîtier
	compteurs	ballon	chauffe
	compteru	boîtier	ballon
	compteir	chauffe	coffret
	compteurgent	coffret	transformateur
Linky	linki	linkie	linkys
	linki	linkis	linké
	linkin	linkdy	linked
	linké	llinky	linkie
	linker	linkdin	linkdy
remboursement	remboursements	retrait	retrait
	remboursment	débit	versement
	rembourssement	versement	réglement
	remboursés	sous	relevé
	remboursable	réglement	transfert
kwh	kwt	kwt	kva
	kws	kws	kwatt
	kwc	kwhr	kva
	kwhr	kwl	kwa
	kws	kwat	kv
gdf	grdf	gil	edfgdf
	edfgdf	gb	gp
	gdm	edfgdf	gil
	gd	gama	giv
	gsm	gdrf	grdf
edf	edl	eos	eos
	edt	edl	elect
	edge	edu	edl
	edmond	ej	ej
	eden	edfmoi	edfservice
ballon	véhicule	chauffe	chauffe
	connecteur	disjoncteur	boîtier
	four	compteur	réfrigérateur
	chauffe	boîtier	compteur
	tableau	lave	frigo
nucléaire	bouquet	club	réchauffement
	chauffage	reste	photovoltaïque
	théâtre	bois	gaz
	panneau	mien	nucléaires
	moteur	secteur	chauffage

Table 6.3: Voisinage de termes issus du domaine de spécialité avec les modèles *camembert-base-wikipedia-4gb*, *camembert-base-oscar-4gb* et *camembert-base-ccnet-4gb* sur EDF-Courriels

sémantique dans certains cas (e.g., les termes « gdf » et « grdf » sont proches, tout comme « kwh » et « kwat ») mais génère beaucoup d'incohérences dans d'autres (e.g., les acronymes « edf » et « gdf » sont également proches d'initiales apparaissant dans les signatures comme « ej » et « gd »). Dans ces cas de figure, nous observons que le découpage initial des tokens prévaut sur leur contexte dans la phrase pour la construction de leur vecteur de représentation. De plus, nous montrons avec l'analyse des voisins de « compteur » et « remboursement », qu'il existe des différences significatives entre la représentation d'un terme lorsqu'il appartient au vocabulaire d'entraînement des modèles ou non. Pour ces termes spécifiques, nous observons qu'ils sont rapprochés de leurs variantes orthographiques ou de leurs autres formes dérivationnelles (i.e., conjugaisons) lorsqu'ils sont hors-vocabulaire et à d'autres termes, parfois synonymes, le cas contraire. Ces résultats nous amènent à penser qu'il existe une corrélation très forte entre la tokenisation des termes hors-vocabulaire et leur représentation dans l'espace multidimensionnel. Nous souhaitons quantifier cet impact afin de déterminer si pour un terme hors-vocabulaire, sa représentation est due à ses tokens où au contexte dans la phrase.

Mot	**Modèle**	**Tokens**
compteur	*Wikipedia*	compte + ur
	Web	compteur
linky	*Wikipedia*	l + in + ky
	Web	l + ink + y
remboursement	*Wikipedia*	rem + bour + s + ement
	Web	remboursement
kwh[1]	*Wikipedia*	k + w + h
	Web	k + wh
GDF	*Wikipedia*	G + DF
	Web	G + DF
EDF	*Wikipedia*	E + DF
	Web	EDF
ballon	*Wikipedia*	ballon
	Web	ballon
nucléaire	*Wikipedia*	nucléaire
	Web	nucléaire

Table 6.4: Tokenisation des termes associés à un sens spécifique au domaine de l'énergie avec les modèles CamemBERT (c'est-à-dire le tokéniseur SentencePiece) sur les courriers électroniques EDF. Les modèles basés sur OSCAR, CCNet et le modèle *Large*, ayant généré les mêmes résultats, sont regroupés au sein de la catégorie *Web*

6.5 Mesures d'évaluation

Dans cette section, nous présentons des métriques d'évaluation de l'impact de la tokenisation sur la représentation des termes hors-vocabulaire. Dans la littérature, de telles mesures n'existent pas pour cette application. Nous proposons donc deux calculs de similarité permettant d'évaluer si la proximité entre deux termes est due à leur tokenisation ou à l'apprentissage de leur contexte. Nous utilisons pour cela les notations présentées dans le Tableau 6.5.

Notation	Description
X, Y	$X = x_1, \ldots, x_k$ et $Y = y_1, \ldots, y_l$ sont des chaînes de caractères de tailles respectives k and l, composées de symboles d'un alphabet de taille finie.
n_X, n_Y	nombre de n-grams contenus dans X et Y.
n_Z	nombre de n-grams communs entre X et Y
$t_M(X)$, $t_M(Y)$	fonction de tokenisation pour X et Y, respectivement, avec le modèle pré-entraîné M.
$n_{t_M(X)}$,$n_{t_M(Y)}$	nombre total de tokens obtenus après la tokenisation de X et Y, respectivement.
$n_{t_M(Z)}$	nombre total de tokens en commun entre X et Y.
$\|t_M(X)_i\|$,$\|t_M(Y)_i\|$	nombre total de caractères dans le $i^{\text{ème}}$ token de X et Y, respectivement.
$\|t_M(Z)_i\|$	nombre total de caractères dans le $i^{\text{ème}}$ token commun entre X et Y.

Table 6.5: Définitions et Notations

Le coefficient de Dice est une mesure très populaire pour calculer la similarité entre des mots ou des chaînes de caractères. Il consiste à calculer le ratio entre le nombre de n-grams de caractères partagés entre deux chaînes de caractères et le nombre total de n-grams de caractères dans les deux chaînes (i.e., l'intersection des n-grams de caractères dans les deux chaînes). Formellement, nous avons :

$$\text{Dice(X,Y)} = 2 \times \frac{n_Z}{n_X + n_Y} \tag{6.1}$$

Dans notre étude, nous souhaitons calculer le nombre de tokens partagés entre deux mots, après la phase de tokenisation. Pour cela, nous remplaçons les n-grams de

[1] 1 kWh ou kilowattheure correspond à 1000 wattheure (Wh). Cette unité sert à mesurer la consommation en énergie de chaque foyer. [SOURCE]

caractères par les tokens de mots générés durant la tokenisation. Mathématiquement, nous le formalisons comme ceci :

$$\text{Dice(X,Y)} = 2 \times \frac{n_{t_M(Z)}}{n_{t_M(X)} + n_{t_M(Y)}} \tag{6.2}$$

Le coefficient de Dice est très utile pour calculer le nombre de tokens partagés entre deux mots. Cependant, la génération de petits sous-tokens durant la tokenisation peut être sous-optimale, étant donné que certains tokens peuvent être trop petits pour contenir des informations sémantiques. Par exemple, si on mesure le coefficient de Dice entre les mots *chiens*, *chien* et *voitures*, tokenisés respectivement en « chien + s », « chien » et « voiture + s », nous obtenons des résultats très proches entre $Dice(chiens, chien) = 0.67$ and $Dice(chiens, voitures) = 0.5$, dus aux marqueurs de pluriel plutôt qu'à la proximité sémantique entre les mots. Nous proposons le coefficient *Dice for Sub-Units* (Dice-SU), une variante du coefficient de Dice, qui pénalise les petits tokens générés lors de la tokenisation. Cette mesure, comme celle de Dice, est comprise entre 0 et 1. Plus la valeur obtenue est élevée, plus les tokens partagés entre deux mots sont grands par rapport aux mots. Nous proposons la définition suivante :

$$\text{Dice-SU}(X, Y) = \frac{2 \times \sum_{i=0}^{n_{t_M(Z)}} |t_M(Z)_i|}{\sum_{i=0}^{n_{t_M(X)}} |t_M(X)_i| + \sum_{i=0}^{n_{t_M(Y)}} |t_M(Y)_i|} \tag{6.3}$$

D'après l'exemple précédent, nous obtenons $Dice - SU(chiens, chien) = 0.91$ et seulement $Dice - SU(chiens, voitures) = 0.2$. Cette mesure permet de mieux estimer l'information sémantique partagée entre deux mots.

Les coefficients de Dice et Dice-SU permettent de calculer la proximité entre les tokens de deux mots. Dans notre étude, nous souhaitons déterminer si la représentation d'un terme hors-vocabulaire dans l'espace multidimensionnel est principalement dû à son découpage initial (i.e., au fait qu'il s'agisse d'un terme hors-vocabulaire) ou à une sémantique déterminée par le modèle. Pour cela, notre algorithme consiste à calculer un coefficient de Dice-SU moyen entre un mot et ses k plus proches voisins (voir Algorithme 3). Cette méthode d'évaluation permet donc de déterminer si les termes hors-vocabulaire sont regroupés en fonction de leurs tokens grâce au coefficient de Dice. Appliquée à Dice-SU, la méthode d'évaluation permet de déterminer si les tokens couvrent une grande partie des mots d'origine (i.e., s'ils partagent une sémantique commune) ou non (e.g., s'ils contiennent seulement des marqueurs de flexions, dérivations, conjugaisons, ou des affixes en commun).

Algorithm 3: Unsupervised Evaluation of Tokenizer's Impact in the Embedding Space of OOVs

Input : A domain-specific word w; a dataset containing a vocabulary V of size N; a matrix of embeddings X of shape (N,length of embeddings); a Transformer model M; a similarity measure $sim(t_M(W_1), t_M(W_2))$ between the words W_1 and W_2; a number n of closest associates to evaluate.

Output : Averaged similarity s between d and its n closest associates

Step 1. Compute the n closest associates of w.;
$p \leftarrow$ position of w in V;
$c_p \leftarrow$ cosine similarity between X_p and X;
$V \leftarrow$ all unique characters $\in D$;
`/* Similarity between w and the rest of the vocabulary */`
$a_i \leftarrow \arg\max_{c_p}, i = 1, \ldots, n$ `/* Get the indices of the top n closest associates */`
$Y \leftarrow V[a_i]$;

Step 2. Tokenize w and its Y associates, and compute the similarity.;
$S_n = []$;
for y_i *in* Y **do**
  $S_i \leftarrow sim(t_M(w), t_M(y_i))$ `/* Similarity between the sub-units of w and` y_i `*/`
$s \leftarrow mean(S_n)$ `/* Averaged similarity between w and its top n associates */`

6.6 Synthèse

Dans ce chapitre, nous avons analysé les problèmes de représentation des termes hors-vocabulaire, issus du domaine général et de domaines spécifiques.

Définition de la problématique. Lors du découpage des mots en tokens avec des modèles de langue pré-entraînés, les termes hors-vocabulaire sont découpés en tokens qui appartiennent au vocabulaire de ces modèles. Nous avons démontré que le problème majeur de la représentation de ces mots est qu'elle dépend fortement des tokens qui les composent. Dans certains cas, cela ne pose pas de problème et peut permettre de répondre à des objectifs comme la correction orthographique lorsque des termes mal orthographiés sont regroupés parmi toutes leurs variantes. Nos travaux ont deux objectifs distincts : (1) réduire l'impact de la tokenisation sur les termes hors-vocabulaire de spécialité et (2) privilégier la représentation liée au contexte des mots dans l'entourage des termes hors-vocabulaire, afin de

mieux tenir compte du contexte de ces termes.

Termes hors-vocabulaire. Dans nos travaux, nous avons distingué les termes hors-vocabulaire du domaine général et de domaines de spécialité. Nous avons effectué une analyse des termes hors-vocabulaire du domaine général et nous avons présenté des motifs de tokenisation fréquents. Nous avons mis en évidence la différence de traitement de ces mots en fonction de leur type (composition, désinence et dérivation). Nous avons également mis en exergue les différences significatives de traitement de termes du domaine de l'énergie en fonction de leur appartenance ou non au vocabulaire des modèles. Ainsi, nous avons pu insister sur l'importance de notre problématique dans ce domaine de spécialité.

Mesures d'évaluation. Pour répondre à cette problématique, il n'existe pas de mesure d'évaluation permettant de calculer l'impact de la tokenisation des mots sur leur représentation dans l'espace multidimensionnel. Pour pallier ce manque, nous avons proposé deux métriques : (1) la mesure de Dice, populaire en TAL, appliquée non pas aux n-grams de caractères communs entre deux mots mais aux tokens communs ; (2) Dice-SU (*Dice for Sub-Units*) qui permet de pondérer la mesure de Dice en pénalisant les tokens communs de petite taille, qui ne contiennent pas d'information sémantique. Nous proposons un algorithme permettant, pour un terme hors-vocabulaire, de calculer ces coefficients à l'échelle de ses voisins et ainsi de déterminer si le voisinage d'un terme hors-vocabulaire est dû à sa décomposition en tokens ou à son contexte dans la phrase. Nous utiliserons ces métriques dans le Chapitre 8 sur les évaluations intrinsèques.

Conclusion et Discussion de la partie

Les modèles de langue peuvent présenter des biais d'apprentissage en raison de la nature des données d'entraînement sur lesquelles ils ont été entraînés. Ces biais peuvent prendre la forme de stéréotypes de genre ou d'une adaptation trop faible à un domaine spécifique, ce qui peut entraîner des prédictions inexactes ou discriminatoires lorsque le modèle est utilisé dans certains contextes. Il est donc important de prendre en compte ces biais lors de la conception et de l'évaluation des modèles de langue, afin de garantir qu'ils soient aussi justes et inclusifs que possible.

Stéréotypes de genre. Nous avons mis en exergue la prépondérance des stéréotypes de genre dans des modèles de plongements lexicaux statiques, utilisés pour représenter les mots dans un espace vectoriel. Ces stéréotypes peuvent être introduits dans les plongements lexicaux en raison de la nature des données d'entraînement sur lesquelles ils ont été appris. Ces données reflètent des préjugés sociaux et culturels. Il est donc important de prendre en compte ces stéréotypes lors de l'utilisation de plongements lexicaux pour diverses applications. Il est important de noter que l'analyse et l'évaluation de ces biais sont encore en cours de développement, et il est crucial de continuer à investiguer pour garantir l'équité dans les systèmes automatisés.

Termes hors-vocabulaire. Nous avons présenté la problématique de représentation des termes hors-vocabulaire par des modèles de langue Transformer. Nous avons réalisé cette étude sur trois domaines de spécialité : le domaine juridique, le domaine biomédical et le domaine de l'énergie. Nous avons proposé deux mesures permettant de calculer l'impact de la tokenisation sur la représentation du vocabulaire : Dice calculé à l'échelle des tokens et Dice-SU permettant de pondérer la similarité calculée en fonction de la taille des tokens partagée entre deux termes. Nous avons démontré des conséquences négatives de la tokenisation des termes hors-vocabulaire sur leur représentation avec les modèles Transformer. L'adaptation de modèles à des données de spécialité peut offrir de nombreux avantages en termes de précision et de pertinence des résultats. En utilisant des données spécifiques à un domaine, comme la médecine ou l'énergie, les modèles peuvent être entraînés pour comprendre les termes et les concepts spécifiques à ce domaine, afin d'améliorer les performances des tâches telles que la classification et la génération automatique de contenu. Cependant, il y a aussi des défis à relever lors de l'adaptation de modèles à des données de spécialité.

$* * *$

Partie IV

Les modèles Transformer appliqués à des données de spécialité

Introduction de la partie

Après avoir mis en exergue les problématiques de représentation des termes hors-vocabulaire et les biais d'apprentissage qui en découlent, nous proposons dans cette partie d'améliorer ces représentations grâce à l'ajout de contexte linguistique dans des modèles Transformer.

Tout d'abord, nous détaillons quelques problématiques actuelles de l'intelligence artificielle appliquée au TAL, notamment en matière d'enjeux environnementaux lors de l'utilisation de méthodes d'apprentissage automatique. Nous verrons que, bien que la tendance actuelle soit de construire des modèles de plus en plus gros, les contraintes environnementales nous mettent en garde sur l'importance d'une démarche écologique lors de nos recherches.

Nous présenterons deux modèles d'incorporation de contexte linguistique à des modèles de langue Transformer afin d'améliorer la représentation de données de spécialité. Les méthodes proposées sont faciles à implémenter, puisqu'elles utilisent des modèles pré-entraînés pour l'étiquetage des corpus.

Nous achèverons cette partie par l'évaluation intrinsèque et extrinsèque des performances obtenues par chacune des méthodes. Nous discutons alors des avantages et des inconvénients de ces méthodes du point de vue de la représentation et des performances obtenues sur plusieurs tâches.

* * *

L'impact environnemental des modèles de TAL

Ces dernières années, la fouille de texte a subi une transformation importante à la suite de l'arrivée des méthodes auto-supervisées. Comme ces méthodes passent à l'échelle et nécessitent plus de ressources de calcul et donc plus d'énergie, l'accent est de plus en plus mis sur l'efficacité et la durabilité des modèles [Strubell et al., 2019, Schwartz et al., 2020]. Par exemple, entraîner un seul modèle BERT (i.e., BERT-Base) [Devlin et al., 2019] sur GPU nécessite autant d'énergie[2] qu'un vol transaméricain [Strubell et al., 2019]. Si les modèles plus récents sont sans doute plus performants sur de nombreuses tâches, ils sont aussi plus grands d'un ordre de grandeur, ce qui soulève de nombreuses préoccupations environnementales [Bender et al., 2021]. Ce problème ne fait que s'aggraver avec le temps, puisque les besoins en calcul doublent en moyenne tous les mois [Sevilla et al., 2022]. Cela a été reconnu dans le domaine de l'intelligence artificielle, et tout particulièrement dans la communauté de TAL, car les modèles existants sont particulièrement longs et coûteux à entraîner. Un champ de recherche a récemment émergé à la suite de la publication d'un document de politique de recommandations[3] pour un TAL efficace, visant à réduire les émissions de gaz à effet de serre. Cette proposition s'inscrit dans l'objectif de construire une « IA verte » (*Green AI*) et un « TAL vert » (*Green NLP*) [Schwartz et al., 2020], qui se réfère à « la recherche qui vise à trouver de nouveaux résultats tout en prenant en compte le coût de calcul, encourageant une réduction des ressources dépensées »[4]. Bien que de nombreux efforts sont réalisés pour une prise de conscience collective de l'impact environnemental des modèles (i.e., l'entraînement et l'exécution de grands modèles de calculs), les modèles continuent de grossir.

Plusieurs outils ont été proposés pour mesurer l'empreinte carbone des modèles durant l'apprentissage [Lacoste et al., 2019, Anthony et al., 2020, Henderson et al., 2020], mais ils ne sont pas adoptés par une grande partie de la communauté de TAL. Dans l'idéal, l'impact environnemental devrait être pris en considération avant de décider quelles expériences doivent être menées. Stede and Patz [2021] et Rolnick et al. [2022] vont encore plus loin, en argumentant que l'impact positif doit être inhérent à la discussion des résultats. Malgré toute l'intelligence de ces méthodes, l'apprentissage de modèles profonds sur les GPU est une technique de force brute. Selon la fiche technique, chaque serveur DGX peut consommer jusqu'à 6,5 kilowatts, sans compter la puissance de refroidissement nécessaire dans le centre de données

[2] On parle d'énergie au sens d'émissions en dioxyde de carbone (CO_2).

[3] Politique de recommandations pour le TAL

[4] Citation d'origine : « AI research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent. »

utilisé. Depuis BERT, de nombreux modèles bien plus grands sont sortis (voir Figure 6.3), avec des apprentissages réalisés sur de nombreuses langues (e.g., les modèles CamemBERT [Martin et al., 2020] et FlauBERT [Le et al., 2020a] en français). Il est difficile d'ignorer le coût de ces apprentissages par rapport au gain de performance qui existe entre les modèles de langue existants.

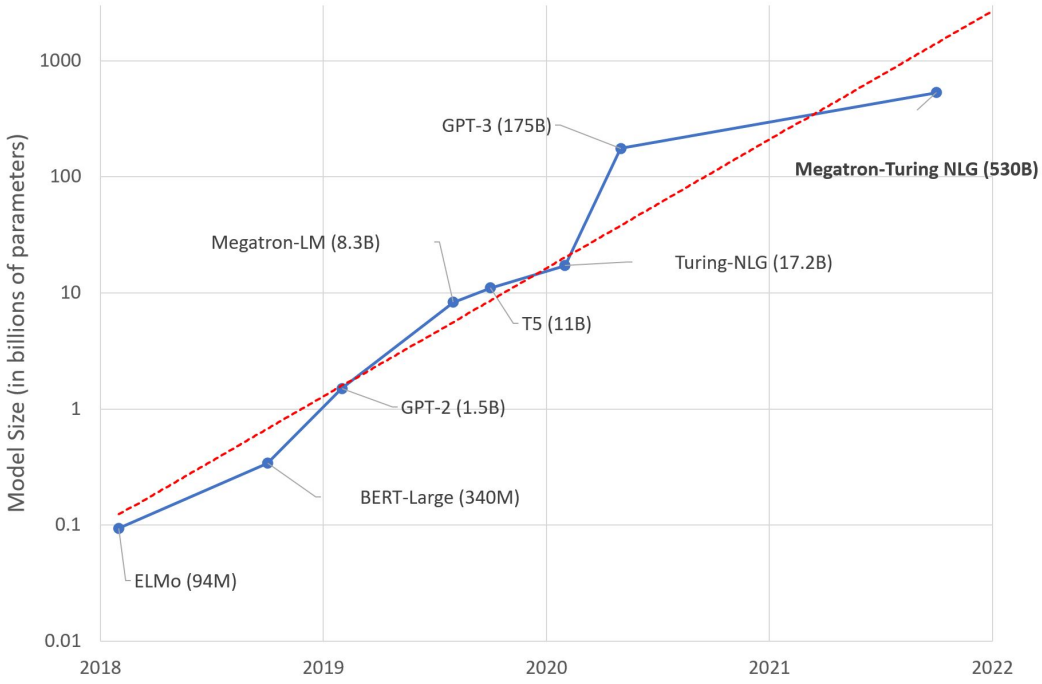


Figure 6.3: Taille de modèles pré-entraîné récents. Source : Hugging Face

Dans cette perspective, nous souhaitons produire des travaux écologiquement responsables, ce qui consiste à réfléchir en amont aux expériences menées et à limiter les expériences réalisées. De plus, nous regardons l'impact environnemental des processus d'auto-apprentissage en apprentissage automatique et nous proposerons une stratégie pour contourner l'affinage des modèles. Nos travaux s'intègrent dans le champ de recherche autour d'une intelligence artificielle plus durable et responsable.

L'avantage d'avoir grandi avec Fred et George, dit Ginny d'un air songeur, c'est qu'on finit par penser que tout est possible quand on a suffisamment de culot.

— Harry Potter et l'Ordre du Phénix

7 Ajout de contexte linguistique

Table des Matières

7.1 Introduction

Notre démarche s'inscrit dans la lignée des travaux visant à injecter des connaissances syntaxiques ou sémantiques à des modèles de langue [Sundararaman et al., 2020], dans le but d'enrichir ou de spécifier leur apprentissage. Dans nos travaux, nous traitons des données de spécialité contenant plusieurs niveaux de complexité : syntaxique (e.g., les erreurs orthographiques, l'absence de ponctuation, inversion de mots) et sémantique (i.e., vocabulaire spécifique). Comme nous l'avons vu précédemment (cf. Section 6), les modèles Transformer ne parviennent pas toujours à capturer du sens à partir des mots de spécialité. *A contrario*, ces mots ont tendance à être regroupés dans l'espace multidimensionnel avec d'autres mots ayant le même découpage en tokens lors de la tokenisation. Cette représentation

pose question, car la tokenisation des mots en tokens est une opération purement statistique et non morphologique. Afin d'améliorer la représentation de ces mots, nous émettons l'hypothèse que l'ajout de contexte syntaxique dans des modèles de langue Transformer permettrait d'enrichir la connaissance des termes hors-vocabulaire et ainsi de mieux les représenter dans l'espace multidimensionnel.

7.2 Présentation des méthodes

Afin de contourner les problèmes de représentation des mots liés à leur tokenisation, nous injectons des informations de contexte dans des modèles Transformer à l'aide d'outils *Open Source*. De ce fait, les méthodes que nous proposons ont l'intérêt d'être facilement utilisables, pour de nombreuses langues (à condition qu'elles bénéficient d'outils d'annotation en POS et en EN).

Les méthodes que nous proposons dans ce manuscrit se fondent sur l'hypothèse suivante : en ajoutant des informations contextuelles sur la structure des phrases (e.g., structures spécifiques comme sur Twitter ou dans des courriels) et le sens des termes hors-vocabulaire (e.g., des termes spécifiques à un domaine), nous pouvons améliorer la représentation du mots sur des domaines de spécialité. Nous proposons deux méthodes qui respectent des critères de recherche différents :

1. Transformer-POS : cette méthode a été construite avec un objectif d'IA responsable, afin d'améliorer des modèles Transformer sans aucun affinage des données. Elle est donc peu gourmande en ressources. Cette méthode consiste à modifier les données en entrée du réseau, sans changer son architecture.

2. Transformer-Embed : cette méthode a été construite de façon à améliorer les performances des modèles Transformer sur diverses tâches. Elle permet d'ajouter des caractéristiques linguistiques sans perturber le modèle pré-appris, en empilant de nouvelles couches à entraîner.

Ces deux méthodes sont faciles à implémenter et généralisables à n'importe quelle langue. Pour ces deux méthodes, les informations de structure ajoutées sont des étiquettes morphosyntaxiques, qui créent du contexte à deux niveaux : (1) l'ajout d'une information de « classe » des mots (i.e., la catégorie POS du mot) et (2) l'ajout d'une information liée à la structure de la phrase (e.g., un déterminant est souvent placé avant un verbe dans la langue française). Pour la deuxième méthode, nous incorporons également une information sémantique aux mots sous forme d'entités nommées. Ces propriétés contribuent à ajouter une sous-couche liée au sens des mots en plus de la structure ajoutée avec les POS. Nous détaillons Transformer-POS et Transformer-Embed dans cette section.

7.2.1 Prétraitement des données

Dans cette partie, nous présentons les briques de prétraitements nécessaires aux méthodes proposées (voir Figure 7.1). Ces méthodes sont différentes, mais contiennent des prétraitements communs, ce qui justifie une présentation commune aux deux méthodes.

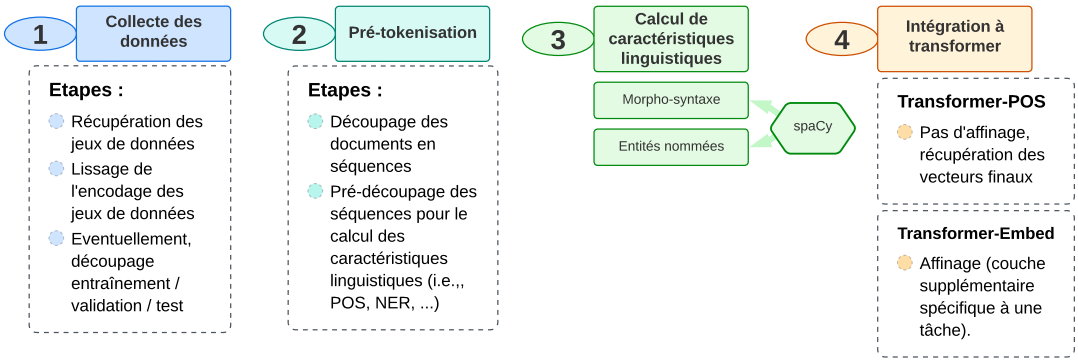


Figure 7.1: Étapes de prétraitements effectuées pour nos modèles

Pré-tokenisation des documents Les méthodes que nous proposons se fondent sur les modèles Transformer pour générer des plongements de mots. Comme pour les modèles Transformer originaux [Devlin et al., 2019, Radford et al., 2018, Martin et al., 2020], l'objectif est de ne « pas trop » prétraiter les corpus en amont, mais d'utiliser le texte brut pour la construction des plongements de mots. En dehors du nettoyage des corpus lié au lissage de l'encodage (i.e., certains corpus étaient encodés en ASCII et devaient être transformés en UTF-8), nous n'avons pas réalisé d'autre opération de prétraitement. La difficulté pour ces modèles réside surtout dans la segmentation des documents longs, en raison de la limite imposée de 512 tokens par document traité pour les modèles utilisés. Le découpage des séquences est différent en fonction de l'objectif à atteindre :

1. Évaluation intrinsèque du corpus (cf. Section 8.2) : cette tâche nécessite uniquement d'encoder les mots d'un corpus et de moyenner les représentations générées. Par conséquent, chaque document est découpé en phrases à l'aide des marqueurs de ponctuation présents dans les documents (i.e., point, point d'interrogation, point d'exclamation et points de suspension). Concernant les documents qui ne contiennent pas de ponctuation (e.g., certains courriels du corpus de clients EDF), un découpage statistique a été réalisé pour segmenter les documents en séquences ayant la taille maximale autorisée par le réseau.

2. Évaluation extrinsèque du corpus (cf. Section 8.3) : on traitera différemment les tâches à résoudre à l'échelle des documents et à l'échelle des mots.

   (a) Échelle du document (ou de la phrase) : les documents des corpus utilisés pour les tâches à l'échelle du document (i.e., analyse de sentiments, inférence, détection de paraphrases) sont souvent courts. En effet, il s'agit de tweets et de critiques de consommateurs en ligne. Pour faciliter les traitements, nous avons choisi de tronquer les documents à 512 tokens (la limite imposée par les modèles Transformer), en conservant seulement le début du document.

   (b) Échelle des mots : cette évaluation ne fonctionne que pour la tâche de reconnaissance d'entités nommées (REN). Cette tâche nécessite d'annoter les mots à l'échelle d'une phrase. Les corpus utilisés sont issus de Wikipédia, donc relativement propres, et l'annotation de ces corpus est effectuée à l'échelle de la phrase (pour chaque phrase, le découpage en mots est déjà réalisé avec leur annotation en EN). Nous avons donc traité les phrases telles quelles, en tronquant éventuellement celles qui étaient trop longues (plus de 512 tokens).

Étiquetage morphosyntaxique Ensuite, les séquences obtenues (après découpage des documents) doivent être annotés en morphosyntaxe pour les deux méthodes. Pour ce faire, plusieurs outils existent dans la littérature. Nous avons cependant choisi d'utiliser un outil gratuit et open-source, à des fins de reproductibilité de nos travaux. Nous avons sélectionné la bibliothèque spaCy[1] disponible avec Python pour deux raisons : (1) il s'agit d'une boîte à outil très complète qui obtient des résultats à l'état de l'art en terme d'annotation morphosyntaxique car elle est fréquemment mise à jour (par exemple, l'outil intègre désormais des vecteurs Transformers) et (2) elle est disponible avec plus de 66 langues (certaines langues sont néanmoins plus riches que d'autres). Cela signifie que nos travaux pourraient être appliqués à un grand nombre de problématiques à travers le monde ; nos expériences ont démontré que les différences de nombre et de classes de POS entre les langues ne sont pas un problème pour ces méthodes. Nous travaillons donc avec 15 classes d'étiquettes POS proposées par spaCy (voir Tableau 7.1). Il arrive qu'un problème d'encodage des espaces survienne où spaCy annotera les espaces avec la mention SPACE ; dans ce cas, l'étiquette sera supprimée *a posteriori*. Nous avons également fait le choix de rassembler les conjonctions de coordination et de subordination dans une classe de conjonctions afin de réduire le nombre de catégories morphosyntaxiques durant l'annotation. Enfin, nous avons rassemblé les auxiliaires et les verbes dans

[1] La bibliothèque spaCy est disponible à l'adresse suivante : https://spacy.io/

la catégorie de verbes pour limiter les erreurs d'annotations par spaCy des verbes être et avoir et ne pas introduire trop de bruit dans les modèles de langue. Pour ces raisons, nous ne conservons pas les informations détaillées (genre, nombre, temps et mode), bien que disponibles sur spaCy. De plus, ces ajouts pourraient engendrer des biais (par exemple, l'ajout de l'accord en genre pourrait renforcer des caractéristiques entre les femmes et les hommes). Dans nos travaux, nous utilisons seulement les classes grammaticales générales. Plus précisément, nous ne faisons pas de distinction entre les conjugaisons des verbes (e.g., entre l'imparfait et le présent), ce que peut faire spaCy. En revanche, nous distinguons les noms communs des noms propres dans ces prétraitements. Des travaux complémentaires restent à mener pour déterminer si ces ajouts sont pertinents pour certaines applications. Nous exploitons les modèles *fr_core_news_lg*[2] (un grand modèle ayant appris sur des articles de presse français) et *en_core_web_lg*[3] (un grand modèle ayant appris sur des extraits du Web en anglais), pour nos expériences respectives en français et an anglais. Ces modèles ont été choisi car ils obtenaient les meilleures performances d'étiquetage en entités nommées et en morpho-syntaxe sur les référentiels d'évaluation de spaCy, au moment de réaliser nos expériences.

POS	Description	Exemple
ADJ	Adjectif	grande, vieux, vert
ADP	Adposition	dans, vers, pendant
ADV	Adverbe	très, demain, cependant
CONJ	Conjonction	mais, où, à, dans
DET	Déterminant	une, le
INTJ	Interjection	psst, ouch, bravo
NOUN	Nom	chien, arbre, femme
NUM	Nombre	1, 2017, un, vingt-cing, IV
PART	Particule	's, not,
PRON	Pronom	je, tu, elle, on, nous, vous, ils
PROPN	Nom propre	Sarah, Houria
PUNCT	Ponctuation	.,()?
SYM	Symbole	&, @
VERB	Verbe	courir, mangeait, votera
X	Autre	mclkhdhdjsks

Table 7.1: Description des étiquettes morphosyntaxiques de spaCy

Étiquetage en entités nommées Pour la méthode Transformer-Embed, nous avons besoin d'annotations en entités nommées. Pour les mêmes raisons qu'avec

[2] https://spacy.io/models/fr#fr_core_news_lg
[3] https://spacy.io/models/en#en_core_web_lg

l'étiquetage en morphosyntaxe, nous avons utilisé spaCy pour nos travaux. Étant donné que nous travaillons en anglais et en français, nous utilisons les catégories d'entités nommées disponibles avec l'outil pour ces deux tâches, qui sont différentes pour les deux langues. Nous présentons les vingt classes d'entités nommées disponibles en anglais et les cinq disponibles en français (voir Tableau 7.2).

EN	**Description**	**Anglais**	**Français**
PERSON	Personne	✓	✓
MISC	Miscellanious		✓
GPE	Entité géopolitique	✓	✓
LOC	Localisation	✓	✓
ORG	Organisation	✓	✓
DATE	Date	✓	
MONEY	Valeurs monétaires	✓	
TIME	Indicateur de temps	✓	
PRODUCT	Objets, véhicules, ...	✓	
CARDINAL	Cardinal	✓	
ORDINAL	Ordinal	✓	
QUANTITY	Quantité	✓	
EVENT	Noms d'évènements	✓	
FAC	Équipements	✓	
LANGUAGE	Langues	✓	
LAW	Lois	✓	
NORP	Nationalités, ...	✓	
PERCENT	Pourcentage	✓	
WORK_OF_ART	Œuvre d'art	✓	
X	Autre	✓	

Table 7.2: Liste des étiquettes d'entités nommées de spaCy disponibles pour l'anglais avec le modèle *en_core_web_lg* et pour le français avec le modèle *fr_core_news_lg*

7.2.2 Spécificités des modèles

Transformer-POS Avec cette méthode, l'objectif est d'injecter les informations morphosyntaxiques dans des documents textuels. Pour cela, chaque mot est annoté avec sa classe morphosyntaxique, en utilisant le séparateur « _ ». Par exemple, la phrase « Ce manuscrit de thèse est super ! » serait annotée comme ceci : « Ce_DET manuscrit_NOM de_DET thèse_NOM est_VERB super_ADJ !_PUNCT ». L'architecture du modèle Transformer-POS est présentée sur la Figure 7.2. Cette méthode a été créée dans le but d'améliorer la représentation des termes dans l'espace multidimensionnel sans effectuer d'affinage. Elle s'inscrit dans une démarche d'IA verte et responsable, en utilisant les modèles en apprentissage

zero-shot. Elle a vocation a être utilisée pour des applications liées à la structure des vecteurs dans l'espace, comme des calculs de similarité ou de la résolution d'analogies. Elle sera analysée plus en détail dans la section sur l'évaluation intrinsèque des modèles (cf. Section 8.2).

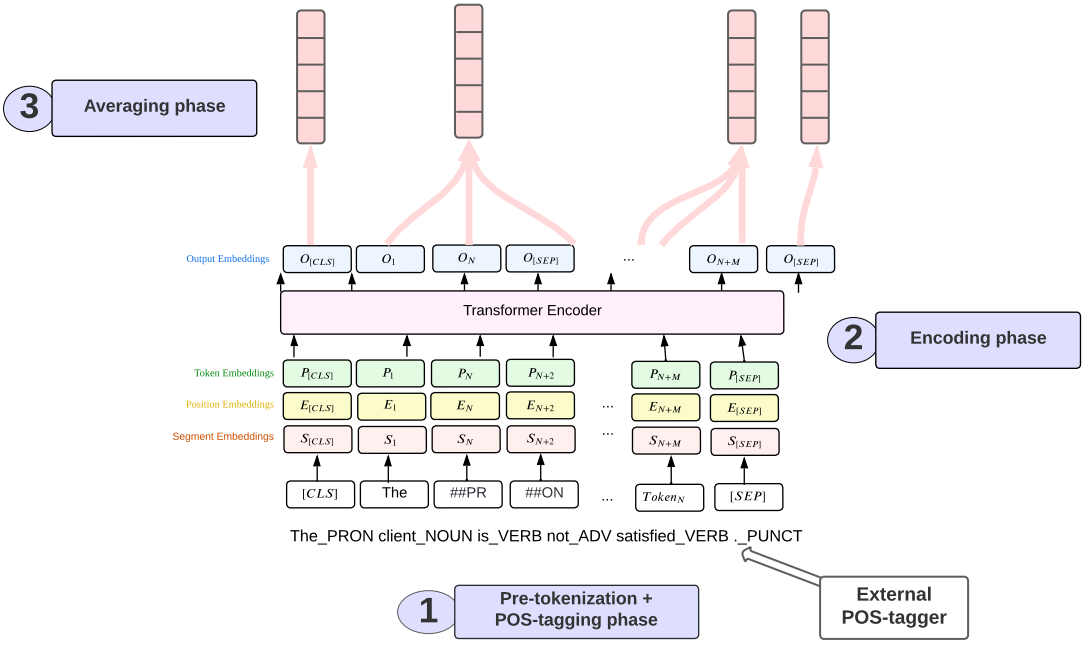


Figure 7.2: Présentation de la méthode Transformer-POS

Transformer-Embed Après la segmentation de la séquence source en tokens par le tokéniseur, un vecteur *one-hot* est calculé en associant chaque token à son étiquette POS. Nous utilisons un modèle pré-entraîné pour annoter les tokens en morphosyntaxe, ce qui confère à la méthode une certaine généricité. Ensuite, chaque mot est représenté par l'index de son étiquette POS. Nous constituons une liste de balises POS de taille m et calculons un vecteur *one-hot* en utilisant l'index des balises POS dans $\{0, m-1\}$ (appelé E_{POS}). Chaque token d'un mot est associé à l'étiquette POS du mot initial. Nous refaisons la même opération en utilisant le même modèle pré-entraîné pour les entités nommées afin de créer un encodage *one-hot* pour les entités (correspondant au vecteur E_{EN}). Les caractères spéciaux des modèles sont encodés, dans les vecteurs de POS et d'EN, avec l'étiquette « NA » qui les distinguent du reste des tokens. Enfin, nous combinons ces vecteurs avec les vecteurs en entrée des modèles pré-entraînés, les incorporations positionnelles. La représentation d'entrée est calculée en additionnant les matrices en entrée du modèle :

$$E = E_T + E_P + E_{POS} + E_{EN} + E_S \in R^{L \times D} \tag{7.1}$$

où la longueur d'entrée L est le nombre de séquences d'entrée (*input sequences*), D est la taille de la dimension des séquences d'entrée, E_T est la représentation du token pré-apprise par le tokéniseur, E_P est le vecteur contenant la position des tokens et E_S est le vecteur contenant l'information sur la position des phrases entre-elles. Plus de détails concernant les sous-couches des modèles Transformer sont fournis dans l'Annexe A. Il est probable qu'avec une méthode de découpage statistique, certaines entités nommées se retrouvent découpées en deux parties. Malheureusement, nous n'avons pas testé d'approches comme la fenêtre glissante pour contourner ce problème. Enfin, la sortie du token [CLS] est introduite dans une couche de classification *softmax* afin de prédire la classe du document. La Figure 7.3 présente l'architecture de la méthode Transformer-Embed. Cette méthode a été construite dans un objectif de résolution de tâches, et sera étudiée plus en détail dans la section sur l'évaluation extrinsèque des modèles (cf. Section 8.3).

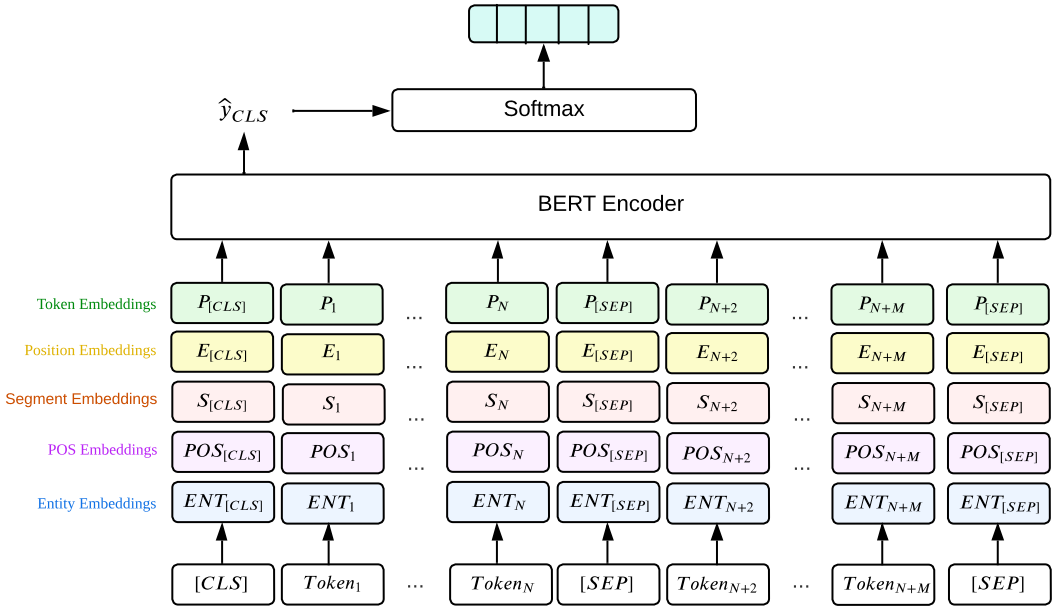


Figure 7.3: Présentation de la méthode Transformer-Embed

7.3 Évaluation des annotations

L'annotation en morpho-syntaxe et en entités nommées est effectuée automatiquement, avec des modèles entraînés spécifiquement sur ces tâches d'annotation. Cependant, afin de déterminer si les taux d'erreurs de nos approches sont influencés par les performances de ces annotations, nous appliquons une méthode d'évaluation simple pour chaque corpus utilisé dans le Chapitre 8.

Cette méthode consiste à effectuer un tirage aléatoire de documents, pour chaque corpus, et d'évaluer l'annotation de spaCy pour chaque token. Cette évaluation a été réalisée par une seule personne, mais les résultats pourraient varier si une étude plus rigoureuse des corpus était réalisée avec des principes d'annotation croisées. Les cinq cent premiers tokens obtenus pour chaque corpus ont été ré-annotés, mais une annotation plus longue aurait pu être menée pour cette évaluation.

Le taux d'erreur calculé pour chaque corpus est présenté dans le Tableau 7.3. Bien que le taux d'erreur en annotation morpho-syntaxique puisse donner une tendance claire d'erreur entre les corpus, les taux d'annotation en entités nommées sont à prendre avec plus de pincettes. Étant donné que ces entités nommées sont peu nombreuses dans le corpus par rapport aux autres termes, peu d'entités ont pu être annotées, et ce résultat dépend du nombre de ces entités annotées après tirage aléatoire. Cependant, sans grande surprise, nous observons que les corpus issus de Twitter (DEFT-18 et PIT) et les corpus de sous-titres sont ceux qui sont les plus difficiles à annoter automatiquement, de par leur style de rédaction spécifique. Les corpus les mieux annotés sont ceux issus de Wikipédia (WikiNER en anglais et en français) et les articles issus du domaine spécifique, mais propres (DEFT-Lois, Bio-Gallica). Enfin, le corpus EDF-Courriels comporte beaucoup d'erreurs d'annotation en POS (8,1%) et en EN (11,7%). Ces résultats visent à présenter un contexte d'analyse des résultats des modèles que nous proposons, qui pourraient être impactées par ces erreurs.

7.4 Synthèse

Dans ce chapitre, nous avons présenté les différentes méthodes d'incorporation de contexte linguistique à des modèles Transformer. L'hypothèse sous-jacente à nos travaux était que l'ajout de contexte, qu'il soit structurel ou sémantique, contribue à enrichir le contexte des termes hors-vocabulaire, et donc à mieux les représenter.

Dans un premier temps, nous avons établi la liste des prétraitements nécessaires à l'application de ces méthodes. Afin de bénéficier au mieux du pré-entraînement des modèles Transformer d'une part, et de faciliter la mise en place de nos modèles d'autre part, nous avons limité le prétraitement des données au minimum. Nous avons utilisé un outil *Open Source* et pré-entraîné pour étiqueter les corpus en morphosyntaxe et en entités nommées. La stratégie consistant à choisir des outils *Open Source* cherche à assurer la reproductibilité de nos travaux et à élargir l'application de ces méthodes à toute langue qui bénéficierait d'un outil similaire.

Corpus	POS (%)	EN (%)
DEFT-Lois	4.2	1.5
Bio-Gallica	3.8	1.1
EDF-Courriels	8.1	11.7
IMDB	9.2	2.5
SST-2	9.7	1.7
CLS-FR	10	1.6
DEFT-18	13.2	13.4
Quora	8.6	1.3
PIT	12.5	13.4
DEFT-20	6	0.9
PAWS-X	7	2.1
OpusParcus	11	12.2
MNLI	6.4	3.9
MIS	6.5	4.7
XNLI-FR	6.2	3.9
WikiNER-En	3.2	0.7
WikiNER-Fr	3.1	0.9

Table 7.3: Liste des étiquettes d'entités nommées de spaCy disponibles pour l'anglais avec le modèle *en_ core_ web_ lg* et pour le français avec le modèle *fr_ core_ news_ lg*

Dans un deuxième temps, nous avons présenté Transformer-POS, une méthode visant à modifier les données d'entraînement d'un modèle Transformer en intégrant des informations morphosyntaxiques. Cette méthode a été construite pour améliorer la représentation des termes hors-vocabulaire en leur ajoutant des propriétés linguistiques. Les avantages de cette méthode sont qu'elle ne nécessite pas d'affinage et qu'elle a été construite pour être plus écologique et performante qu'un modèle affiné sur des données de spécialité. Ce modèle n'a pas vocation à être appliqué à des tâches d'apprentissage automatique, mais plutôt à améliorer le nuage de représentation des mots dans des corpus de spécialité. Par conséquent, il pourrait être envisagé sur des tâches de similarité sémantique, de détection d'analogies et d'extraction d'informations.

Dans un troisième temps, nous avons proposé Transformer-Embed, une méthode consistant à injecter des informations structurelles et sémantiques avant l'encodage des données. Elle consiste à ajouter des vecteurs représentatifs de la morphosyntaxe et des entités nommées présentes dans un corpus. Cette méthode a vocation à être appliquée à des tâches, afin de déterminer si l'ajout de morphosyntaxe et/ou

l'ajout d'entités nommées est pertinent pour des tâches cibles. L'avantage de cette méthode, en dehors de sa facilité d'implémentation, est qu'elle peut être appliquée à n'importe quelle information linguistique que l'on souhaite intégrer aux modèles de langue. Nous pourrions imaginer l'ajout d'informations de dépendances dans une phrase, de chunks, de locutions adverbiales, etc. De plus, nous pouvons imaginer d'enrichir ces modèles avec des informations sémantiques telles que des ontologies spécifiques à des domaines.

Et pour finir, je dirais qu'il faut beaucoup de courage pour affronter ses ennemis mais qu'il en faut encore plus pour affronter ses amis.

— Harry Potter à l'école des sorciers

8

Évaluation et Discussion

Table des Matières

8.1 Introduction

Dans ce chapitre, nous présentons les résultats obtenus par les deux méthodes proposées (cf. Section 7.2) que nous avons testées et évaluées, avec une évalu-

ation intrinsèque d'une part, et une évaluation extrinsèque d'autre part. Nous discutons également des résultats obtenus et des performances obtenues avec ces deux approches, par comparaison avec des modèles Transformer classiques. Les mesures d'évaluation utilisées pour l'évaluation qualitative (ou intrinsèque) sont des mesures de similarité (Jaccard et les mesures de Dice et Dice-SU présentées dans la Section 6.5). Pour l'évaluation quantitative, nous présenterons trois mesures (exactitude, rappel et précision). Les données et les modèles utilisés pour ces évaluations ont été présentés respectivement dans les Chapitres 3 et 4.

8.2 Évaluation intrinsèque

8.2.1 Introduction

Dans cette étude, l'objectif est d'effectuer l'évaluation intrinsèque (ou analyse qualitative) de la méthode Transformer-POS, visant à améliorer la représentation des termes hors-vocabulaire dans des corpus de spécialité. L'évaluation intrinsèque a pour objectif d'observer et d'évaluer la représentation des mots dans l'espace avec nos modèles, en comparaison avec les Transformer usuels. Dans un premier temps, nous présenterons une mesure de similarité appelée Dice-SU qui quantifie l'impact des tokéniseurs sur la représentation des termes hors-vocabulaire. Dans un deuxième temps, nous visualiserons les modifications de l'espace multidimensionnel construite avec des modèles Transformer et avec nos méthodes. Dans un troisième temps, nous utiliserons principalement des méthodes de similarité (i.e., Dice-SU, cosinus et jaccard) pour calculer le voisinage local d'un mot dans l'espace multidimensionnel, et ainsi interpréter sa sémantique.

8.2.2 Détail des expériences

8.2.2.1 Corpus et Modèles

Nous utilisons les mêmes corpus et les mêmes modèles que ceux présentés dans l'étude de la représentation des plongements de mots, présentés dans la Section 6.3 du Chapitre 6.

8.2.2.2 Expériences comparatives

Dans cette section, nous évaluerons l'efficacité du modèle Transformer-POS pour représenter des termes hors-vocabulaire. Pour cela, nous devons comparer ce modèle à d'autres stratégies d'adaptation des modèles au domaine. Ici, nous avons choisi

l'affinage classique des modèles ainsi que l'utilisation d'un autre modèle appelé ELMo (voir la présentation du modèle en Section 4 et son architecture en Annexe A).

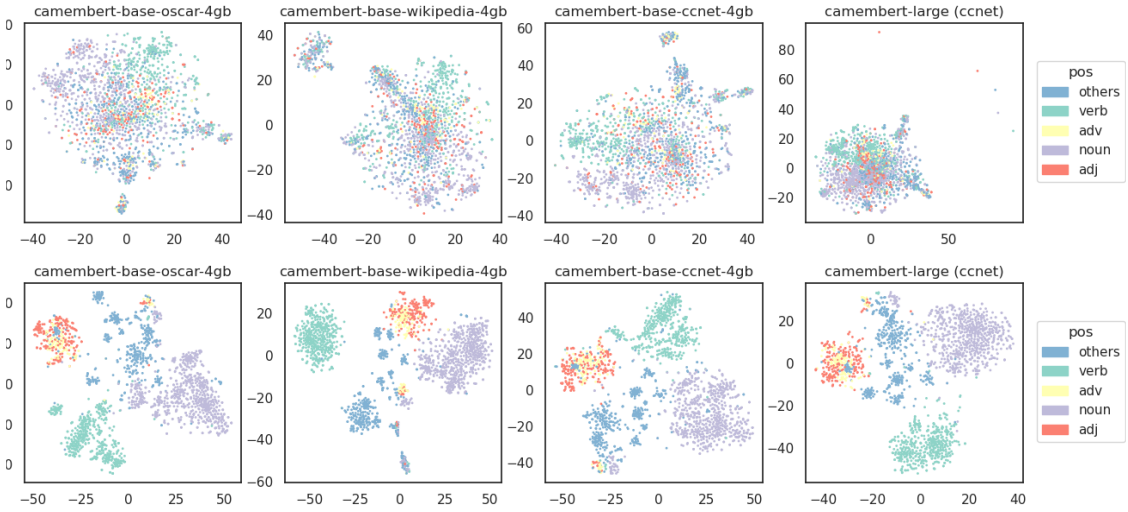
Affinage de CamemBERT/FlauBERT. Nous commençons par affiner les modèles de langue sur chaque ensemble de données en continuant à les entraîner avec de nouveaux textes. Nous effectuons une optimisation simple pour comparer les modèles sans utiliser des hyperparamètres trop complexes et spécifiques. Pour cela, nous choisissons d'affiner le tokéniseur (en amont du modèle) et les poids du modèle de langue pour chaque ensemble de données. Dans cette étude, il est crucial de ne pas figer les poids car nous ne voulons pas utiliser les modèles sur d'autres ensembles de données. Nous cherchons donc à améliorer les performances de chaque modèle sur son ensemble de données respectif.

Concaténation avec des représentations ELMo. Nous combinons un modèle contextuel pré-entraîné ELMo [Peters et al., 2018a] avec des Transformers pour ajouter plus de contexte dans les plongements de mots, comme suggéré par Polatbilek [2020]. Nous utilisons le modèle français pré-entraîné fourni par Che et al. [2018a], avec des enchâssements de dimension 512. Nous concaténons les représentations des modèles ELMo et Transformer (c'est-à-dire CamemBERT ou FlauBERT) pour chaque mot, obtenant ainsi une dimension de 1230.

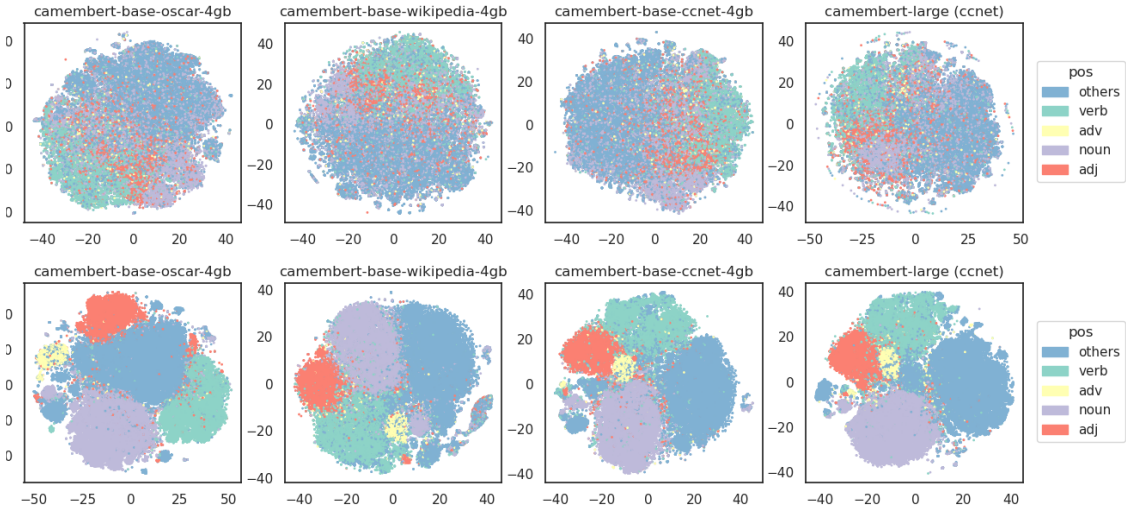
8.2.3 Différences de structure globale

8.2.3.1 Visualisation des plongements lexicaux

La Figure 8.1 présente l'impact significatif de l'application de Transformer-POS sur la distribution du vocabulaire en visualisant les représentations extraites des mots des ensembles de données EASY et EDF-Courriels. Les représentations sont observées grâce à une réduction de la dimension à deux composantes, avec l'algorithme t-SNE [van der Maaten and Hinton, 2008]. Comme prévu, nous démontrons que les vecteurs de mots de l'approche Transformer-POS sont plus séparables en ce qui concerne les catégories POS que ceux de CamemBERT. Cela indique que nous avons réussi à regrouper les mots syntaxiquement similaires en ajoutant des caractéristiques POS dans CamemBERT avant d'encoder les données. En effet, étant donné que le modèle CamemBERT n'ayant pas été entraîné pour la morpho-syntaxe, il est parfaitement normal que ce concept lui soit inconnu.



(a) EASY - CamemBERT (au dessus) et CamemBERT-POS (en dessous)



(b) EDF-Courriels - CamemBERT (au dessus) et CamemBERT-POS (en dessous)

Figure 8.1: Visualisation des termes avec CamemBERT après réduction de la dimension avec l'algorithme t-SNE

8.2.3.2 Évaluation des *clusters*

Pour valider nos observations, nous avons effectué un clustering k-means avec une distance euclidienne. Nous utilisons deux mesures d'évaluation pour évaluer objectivement les résultats du clustering : la pureté et l'information mutuelle normalisée (IMN).

La pureté La pureté est une mesure simple permettant d'évaluer la qualité des *clusters* générés grâce à un algorithme de *clustering*. Pour cela, chaque *cluster* est attribué à la classe la plus fréquente qui le compose, puis l'exactitude est calculée en comptabilisant le nombre de documents correctement classés et en divisant le

résultat par le nombre de classes. Formellement :

$$\mathrm{p}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j| \tag{8.1}$$

où $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ est l'ensemble des *clusters* et $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ est l'ensemble des classes réelles. Dans l'équation 8.1, ω_k est l'ensemble des documents contenus dans le *cluster* ω_k et c_j est l'ensemble des documents contenus dans la classe c_j. Un *clustering* éloigné des classes attendues correspond à une pureté proche de 0, alors qu'il correspond à une pureté proche de 1 lorsqu'il représente au mieux les classes réelles. Étant donné que nous ne cherchons pas à obtenir un seul cluster représentatif de chaque catégorie morphosyntaxique mais plusieurs clusters, la mesure de pureté est particulièrement intéressante dans cette étude. Seulement, une pureté de 1 est facile à obtenir lorsque le nombre de *clusters* est grand (en particulier, si chaque *cluster* est associé à chaque document), c'est pourquoi cette mesure n'est pas suffisante pour évaluer la qualité du *clustering* en tenant compte du nombre de *clusters*.

L'information mutuelle normalisée (IMN) La prise en compte du nombre de *clusters* dans l'évaluation est réalisée grâce à l'IMN :

$$\mathrm{IMN}(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2} \tag{8.2}$$

où I est l'information mutuelle :

$$I(\Omega; \mathbb{C}) = \sum_{k} \sum_{j} P(\omega_k \cap c_j) \log \frac{P(\omega_k \cap c_j)}{P(\omega_k) P(c_j)} \tag{8.3}$$

où $P(\omega_k)$, $P(c_j)$ et $P(\omega_k \cap c_j)$ sont les probabilités qu'un document appartiennent à un *cluster* ω_k, une classe c_j, et soient respectivement à l'intersection de ω_k et c_j.

et H correspond à l'entropie :

$$H(\Omega) = -\sum_{k} P(\omega_k) \log P(\omega_k) \tag{8.4}$$

La normalisation de l'information mutuelle, effectuée grâce au dénominateur $[H(\Omega) + H(\mathbb{C})]/2$ dans l'équation 8.2, permet de comparer les *clusterings* réalisés avec différents nombres de *clusters*, et est comprise entre 0 et 1.

8.2.3.3 Résultats obtenus

Nous répétons 10 fois le *clustering* k-means sur les EDF-Courriels, et à chaque implémentation nous générons aléatoirement les graines initiales. Nous sélectionnons le nombre de clusters avec la méthode du coude [Joshi and Nalwade, 2013]. Les résultats sont détaillés dans le Tableau 8.1 et mettent en évidence que la taille des données d'entraînement ne modifie pas la représentation syntaxique des termes. Il y a deux explications possibles à cela : 1) le petit ensemble de données contient des exemples représentatifs du plus grand ou 2) un petit ensemble de données est suffisant pour modéliser les propriétés syntaxiques des phrases, telles que calculées par CamemBERT.

		# clusters								
		13			*14*			*15*		
Modèle	**Mesure**	CBERT	CPOS	AFF.	CBERT	CPOS	AFF.	CBERT	CPOS	AFF.
oscar	IMN	.150	.481	.164	.151	**.498**	.165	.153	.496	.163
	Pureté	.518	.838	.584	.521	.853	.589	.521	**.862**	.589
wiki.	IMN	.128	**.484**	.164	.122	.470	.165	.124	.462	.157
	Pureté	.495	**.862**	.592	.490	.836	.601	.490	.838	.592
large	IMN	.130	**.519**	.168	.130	.515	.164	.131	.513	.165
	Pureté	.555	.869	.595	.559	.877	.584	.560	**.882**	.601

Table 8.1: Clustering K-means réalisé après réduction de la dimension avec l'algorithme t-SNE sur les courriels EDF. Nous comparons la qualité des résultats sur le regroupement des termes en fonction de leur catégorie POS. AFF.: Affinage du modèle

8.2.4 Analyse des voisinages locaux

Pour cette analyse, nous reprenons les termes étudiés précédemment (voir Tableau 6.4) spécifiques au domaine de l'énergie et au format du courriel. Nous avions observé (voir Tableau 6.3), pour les modèles CamemBERT, que la tokenisation des termes hors-vocabulaire semblait avoir un impact très fort sur leur représentation dans l'espace multidimensionnel. Ici, nous comparons ces résultats avec ceux obtenus avec la méthode Transformer-POS, qui consiste à ajouter du contexte morphosyntaxique dans les données (voir Tableau 8.2). Nous observons que la méthode Transformer-POS génère beaucoup moins de voisins ayant des tokens en commun avec le terme cible qu'avec les modèles usuels. Néanmoins, cela arrive dans certains cas, principalement avec le modèle ayant appris sur Wikipédia, lorsqu'il s'agit de traiter des acronymes hors-vocabulaire (e.g., « kwh » et « gdf »). Cependant, Transformer-POS parvient à trouver des termes similaires aux acronymes avec les modèles ayant appris sur le Web. Par exemple, nous remarquons que les entreprises d'électricité (distributrices ou productrices d'énergie) « enedis », « engie » et

Mot	Wikipedia	Oscar	CCNet
compteur	compteurs	comptage	relais
	disjoncteur	indicateur	contrat
	ballon	forfait	remplacement
	relevé	raccord	diagnostic
	compteeur	disjoncteur	boîtier
Linky	linki	ginko	linkie
	ld	zac	linkys
	li	cbe	linkdy
	link	log	linked
	log	installateur	linkee
remboursement	remboursements	retrait	retrait
	remboursment	débit	versement
	règlement	versement	réglement
	régularisation	sous	relevé
	télépaiement	réglement	transfert
kwh	kws	kwhat	occupant
	kw	reference	kws
	klwh	kwhs	energie
	kwt	pleines	élec
	kva	creuses	zac
gdf	grdf	grdf	grdf
	gdrf	pdl	gaz
	edfgdf	agde	engi
	gfr	engie	engie
	cgt	enedi	iban
edf	edt	electr	enedis
	enedi	edf&moi	engie
	edf&moi	enedi	enedi
	engie	enedis	gdf
	gaz	edfet	lcl
ballon	nid	chaudière	pompe
	chauffe	relais	chauffe
	relais	fusible	chaudière
	boîtier	disjoncteur	relais
	carton	boitier	dalle
nucléaire	matériel	photovoltaïque	monopole
	local	police	cancer
	secteur	ventilateur	lumière
	noir	cheminée	noir
	technique	pompe	pauvreté

Table 8.2: Voisinage de termes issus du domaine de spécialité avec les modèles *camembert-base-wikipedia-4gb*, *camembert-base-oscar-4gb* et *camembert-base-ccnet-4gb* sur EDF-Courriels. Les voisins sont classés par proximité décroissante

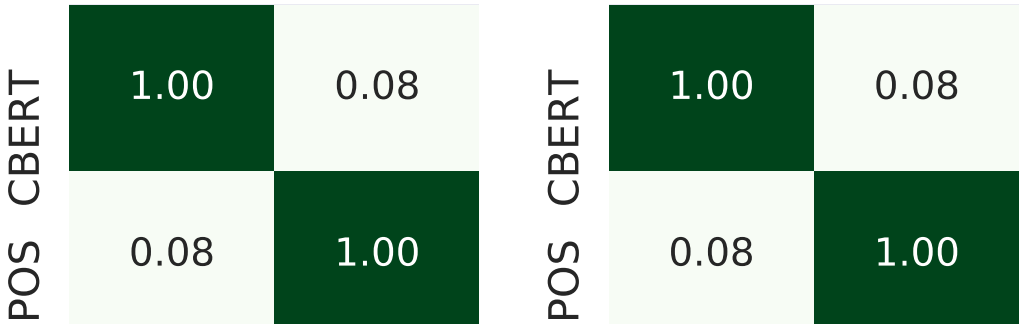
« gdf » apparaissent désormais dans le voisinage d'« edf » et de « gdf ». Nous observons également que de nouveaux termes apparaissent dans le voisinage des mots « compteur » et « remboursement » avec tous les modèles. Enfin, nous remarquons que, dans le voisinage du terme Linky, difficile à représenter pour les modèles, plusieurs termes hors-vocabulaire spécifiques au domaine de l'énergie et pertinents sont présents (CBE: Compteur Bleu Electronique, GINKO: Système d'information d'Enedis au service de Linky, ZAC: Zone d'aménagement Concerté). Dans l'ensemble, nous observons que le modèle Wikipédia est utile pour rapprocher des variantes orthographiques, alors que le modèle CCNet permet de rapprocher des synonymes ; le premier reste au niveau des formes de surface alors que l'autre se déplace au niveau sémantique grâce aux contextes.

Ces résultats valident notre hypothèse préliminaire selon laquelle l'ajout de contexte morphosyntaxique enrichit les connaissances des Transformer sur les termes hors-vocabulaire et améliore leur représentation dans l'espace multidimensionnel. Il semblerait que la méthode Transformer-POS que nous proposons le fait efficacement, car elle réduit l'impact significatif de l'étape de tokenisation sur la représentation des termes hors-vocabulaire. Nous allons tester cette hypothèse sur d'autres domaines dans la suite de notre étude, afin d'évaluer la robustesse de notre méthode dans différents contextes.

Pour quantifier les différences entre les voisins générés par CamemBERT et CamemBERT-POS, nous utilisons des mesures d'évaluation comparatives. L'indice de Jaccard [Jaccard, 1901] (ou coefficient de Jaccard) permet d'évaluer la similarité entre deux ensembles. Mathématiquement, l'indice de Jaccard s'écrit comme ceci :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{8.5}$$

Nous utilisons plutôt la distance de Jaccard, qui consiste à soustraire l'indice de Jaccard à 1. Elle mesure la dissimilarité entre deux ensembles. Pour calculer la distance, les cinquante premiers voisins obtenus par chaque méthode ont été utilisés et nous avons mesuré la dissimilarité entre les ensembles de voisins obtenus avec CamemBERT et CamemBERT-POS. Nous avons calculé la moyenne des similarités obtenues pour les cent mots les plus fréquents du corpus. Comme le montre la Figure 8.2, les deux modèles génèrent des voisins significativement différents avec une similarité Jaccard de 0,08 en moyenne, ce qui confirme que CamemBERT-POS change radicalement la représentation des mots.



(a) Oscar **(b)** Wikipedia

Figure 8.2: Similarité de Jaccard obtenue pour les cinquante plus proches voisins des 100 mots les plus fréquents

8.2.5 Termes hors-vocabulaire

Dans cette étude, nous évaluons l'impact de la tokenisation sur la représentation des termes hors-vocabulaire. En outre, les modèles doivent prendre en compte les textes bruités générés par les utilisateurs (par exemple, le contenu des médias sociaux ou des courriels). Les fautes de frappe (par exemple, l'insertion de caractères), les émojis et les abréviations sont les bruits les plus courants. Selon Park et al. [2016], les termes hors-vocabulaire peuvent être classés en plusieurs catégories lorsqu'on travaille sur les médias sociaux (par exemple, les mots étrangers, les fautes d'orthographe). En nous inspirant de leur typologie, nous allons évaluer la robustesse des modèles Transformer pour trois types de termes hors-vocabulaire :

- nouveaux termes spécifiques au domaine (par exemple, « enzyme » et « eucaryote » en microbiologie) ;
- les mots mal orthographiés contenant des fautes de frappe (par exemple, « infractus » au lieu de « infarctus » en médecine) ;
- les homographes inter-domaines (c'est-à-dire des mots qui s'écrivent de la même façon mais qui ont des significations différentes) de mots existant dans la langue générale (par exemple, « bras », soit une partie anatomique dans la langue générale, soit une sous-partie d'une cohorte de patients dans les essais cliniques).

8.2.5.1 Les termes hors-vocabulaires spécifiques à un domaine

Nous souhaitons évaluer la représentation de termes hors-vocabulaires spécifiques à un domaine avec des modèles Transformer. Cette expérience est menée sur deux

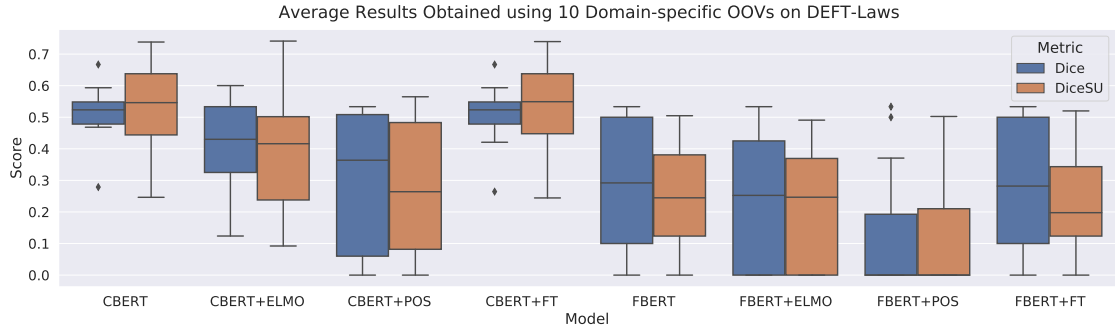
domaines : juridique et médical. Nous n'utilisons pas les courriers électroniques d'EASY et d'EDF dans cette section car cette catégorie de termes hors-vocabulaire n'est pas fréquente dans le corpus. Nous avons sélectionné 10 termes hors-vocabulaire fréquents et spécifiques au domaine juridique dans le corpus DEFT-Lois[1] et au domaine médical dans le corpus Bio-Gallica[2]. Pour chaque terme, nous avons calculé ses cinq plus proches voisins en utilisant la distance cosinus. Ensuite, nous avons tokenisé les termes hors-vocabulaires et leurs voisins en utilisant un modèle Transformer (CamemBERT ou FlauBERT). Nous avons mesuré et calculé la moyenne du coefficient de Dice et du coefficient de Dice-SU obtenu entre les termes hors-vocabulaires et leurs voisins. La procédure complète est détaillée dans l'Algorithme 3. Nous présentons les résultats obtenus sur les deux jeux de données sous forme de boîte à moustache (voir Figure 8.3). Un exemple détaillé des résultats obtenus pour le mot « discriminatoires » sur le corpus d'articles de lois DEFT-Lois est présenté dans le Tableau 8.3.

DEFT-Lois Nous observons que le *coefficient de Dice* est en moyenne plus élevé avec CamemBERT qu'avec FlauBERT, avec une différence de 20%. Par conséquent, la représentation des termes hors-vocabulaire est plus impactée par la tokenisation Unigram de CamemBERT qu'avec la tokenisation BPE de FlauBERT. De plus, l'ajout d'informations morphosyntaxiques réduit drastiquement l'impact de la tokenisation sur la représentation des termes hors-vocabulaires. FlauBERT-POS obtient une similarité moyenne nulle entre les termes hors-vocabulaire et leurs voisins, ce qui signifie que les tokens des voisins sont différents de ceux du terme source. Enfin, le *coefficient Dice-SU* est globalement inférieur à 50%, ce qui signifie que les termes hors-vocabulaire et leurs voisins partagent des tokens de petite taille, qui peuvent-être liés à une syntaxe commune (exemple : le marqueur du pluriel en -s ou du genre en -e). Par conséquent, nous supposons que les tokens partagées entre ces termes ne contiennent pas beaucoup d'information sémantique et donc que le rapprochement « naturel » des mots en fonction de leurs tokens communs réalisé par les modèles n'est probablement pas la plus souhaitable. Nous notons que l'affinage des modèles n'a pas modifié la représentation des termes hors-vocabulaire, ce qui est dû au fait que malgré le réapprentissage du tokéniseur n'a pas permis de modifier le vocabulaire de découpage de ces mots. Cependant, l'ajout d'informations contextuelles (avec ELMo ou POS) a réduit l'impact du tokéniseur, en particulier lors de l'ajout du contexte morphosyntaxique.

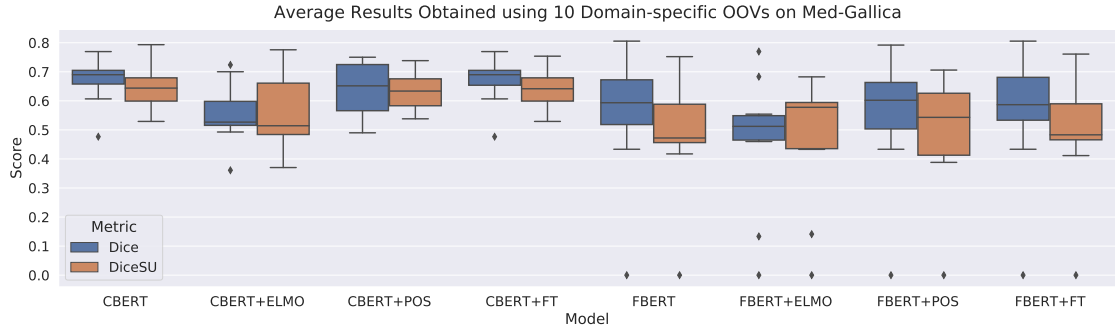
[1] *allegation, frauduleux, procès-verbal, délibéré, règle, apparence, discriminatoire, régularisé, cessionnaire, national, régularisé, enregistré, scellé et apposé*

[2] *incubation, bactériologique, épileptique, prophylactique, tuberculose, cautérisation, bacillophage, septicémie, hyperesthésie et anorexie*

Bio-Gallica Nous avons obtenu des tendances similaires sur Bio-Gallica, mais nous observons plus de tokens partagés entre les termes hors-vocabulaires et leurs voisins. Le *coefficient Dice-SU* est plus élevé, avec un score de similarité moyen compris entre 50% et 70%. Dans le domaine médical, la couverture entre les tokens des termes hors-vocabulaire et les tokens de leurs associés les plus proches est élevée, ce qui signifie que même si le tokéniseur a un impact significatif sur la représentation des termes hors-vocabulaire, les tokens sélectionnés sont probablement pleins d'informations sémantiques pertinentes.



(a) DEFT-Lois



(b) Bio-Gallica

Figure 8.3: Distribution des coefficients de Dice et Dice-SU sur 10 termes hors-vocabulaires spécifiques au domaine juridique (en haut) et au domaine médical (en bas). La moyenne des coefficients est rapporté entre chaque terme hors-vocabulaire et ses voisins (en utilisant la similarité cosinus). Les boîtes à moustache contiennent les scores obtenus par les modèles Transformers pour chaque coefficient

8.2.5.2 Erreurs orthographiques

Les termes mal orthographiés sont parfois difficiles à identifier avec des modèles utilisant des tokens de mots [Nayak et al., 2020, Sun et al., 2020]. La difficulté est double lorsque les mots mal orthographiés sont également des mots de spécialité. A l'échelle du mot, ces erreurs sont orthographiques ou grammaticales, et peuvent résulter de l'insertion, de la suppression ou de la substitution incorrecte d'un caractère ou de la transposition de deux caractères adjacents. A l'échelle de la phrase,

	Voisins	Tokens	Dice	Dice-SU
CamemBERT-Base discriminatoire + s	discriminatoire	discriminatoire	0,67	0,97
	discrimination	discrimination	0	0
	restrictives	restrictive + s	0,5	0,07
	injustifiées	injustifié + es	0	0
	inacceptables	inacceptable + s	0,5	0,07
+ ELMo	discriminatoire	discriminatoire	0,67	0,97
	souhaitables	souhaitable + s	0,5	0,07
	restrictives	restrictive + s	0,5	0,07
	exhaustives	exhaustive + s	0,5	0,07
	contraignants	contraignant + s	0,5	0,07
+ POS	discriminatoire	discriminatoire	0,67	0,97
	restrictives	restrictive + s	0,5	0,07
	injustifiés	injusti + fiés	0	0
	arbitraires	arbitraire + s	0,5	0,07
	unilatérales	unilatérale + s	0,5	0,07
FlauBERT-Base discriminatoires	discriminatoire	discriminatoire	0	0
	non-discriminatoires	non- + discriminatoires	0,67	0,91
	discriminations	discriminations	0	0
	non-discriminatoire	non- + discriminatoire	0	0
	discrimination	discrimination	0	0
+ ELMo	discriminatoire	discriminatoire	0	0
	non-discriminatoires	non- + discriminatoires	0,67	0,91
	restrictives	restrictives	0	0
	non-discriminatoire	non- + discriminatoire	0	0
	inefficaces	inefficaces	0	0
+ POS	discriminatoire	discriminatoire	0	0
	contradictoires	contradictoires	0	0
	contraignantes	contraignantes	0	0
	néfastes	néfastes	0	0
	monotone	monotone	0	0

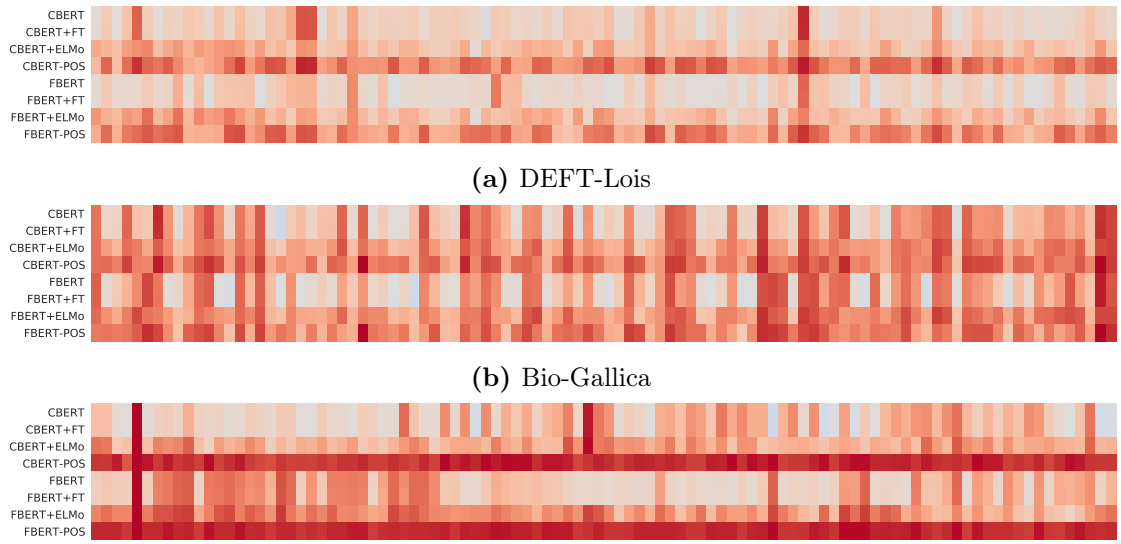
Table 8.3: DEFT-Lois - Cinq plus proches voisins du mot « discriminatoires » en terme de similarité cosinus. Les coefficients Dice et Dice-SU ont été moyennés sur l'ensemble des voisins et calculés entre le mot « discriminatoires » et ses cinq candidats

elles peuvent correspondre à des erreurs syntaxiques. Nous menons cette étude sur trois types spécifiques de fautes d'orthographe, un pour chaque ensemble de données :

- DEFT-Lois : nous analysons les fautes d'orthographe sur des mots qui existent dans les vocabulaires de CamemBERT et FlauBERT (par exemple, « xonditions » au lieu de « conditions »).
- Bio-Gallica : nous nous concentrons sur les fautes d'orthographe des mots spécifiques au domaine qui n'existent pas dans le vocabulaire des modèles (par exemple, "injoction" au lieu de "injection"). Les erreurs d'OCR sont à l'origine de toutes les fautes d'orthographe dans cet ensemble de données.
- EDF-Courriels : nous évaluons le traitement des fautes d'orthographe sur des mots spécifiques à la structure des emails et apparaissant à des positions similaires dans les textes (par exemple, "cordiallement" au lieu de "cordialement").

Protocole d'annotation. Nous avons effectué un tirage aléatoire de 100 mots mal orthographiés dans les corpus, afin de comparer le traitement de ces erreurs par les modèles. Plus précisément, nous avons généré un tirage aléatoire de mots par batchs de 50, puis nous avons annoté les mots à la main pour distinguer les mots bien orthographiés des mots mal orthographiés. Nous avons annoté des batchs jusqu'à ce que nous obtenions 100 mots pour chaque corpus. La liste de ces mots est détaillée dans l'Annexe B. Ensuite, nous avons associé, toujours à la main, les mots mal orthographiés à leur version correcte. Enfin, nous avons calculé la similarité cosinus entre les paires de mots $\text{word}_{\text{source}}, \text{word}_{\text{erreur}}$. Les résultats sont présentés à l'aide de cartes de chaleur (*heatmaps*), dans lesquelles nous représentons chaque case par la similarité entre le mot et sa variante orthographique. Nous présentons les résultats obtenus sur DEFT-Lois (voir Figure 8.4a), Bio-Gallica (voir Figure 8.4b) et EDF-Courriels (voir Figure 8.4c). La similarité moyenne obtenue pour les 100 termes hors-vocabulaire est également présentée dans le Tableau 8.4.

DEFT-Lois CamemBERT est légèrement plus performant que FlauBERT, avec une différence de 3% de la similarité en cosinus. Bien que cela puisse signifier que la tokenisation basée sur Unigram est légèrement plus efficace que BPE pour construire des tokens proches de la morphologie des mots mal orthographiés, la différence entre les résultats des deux méthodes reste trop mince pour valider l'hypothèse. L'affinage des modèles de langue sur les données n'a pas permis de modifier les résultats de façon significative sur les mots mal orthographiés sélectionnés. En définitive, l'affinage n'a pas permis de modifier la segmentation de ces termes, que ce soit avec BPE ou Unigram. Nous pensons que cela est dû à la taille des corpus, trop petite par rapport aux jeux de données d'entraînement des modèles.



(a) DEFT-Lois

(b) Bio-Gallica

(c) EDF-Courriels

Figure 8.4: Similarité cosinus obtenue sur 100 mots mal orthographiés et leurs 100 variantes bien écrites. Plus une case est rouge, plus la similarité entre la paire $word_{source}, word_{erreur}$ est élevée. CBERT et FBERT font référence respectivement aux versions Base des modèles CamemBERT et FlauBERT

	Droit	Médical	Energie
CamemBERT-Base	0,19	0,39	0,27
+ELMo	0,32	0,54	0,44
+POS	**0.63**	**0.66**	**0.92**
FlauBERT-Base	0,15	0,37	0,34
+ELMo	0.34	0.57	0.56
+POS	0.56	0.63	**0.93**

Table 8.4: Résultats moyens de la similarité en cosinus entre les 100 mots mal orthographiés sélectionnés au hasard et leurs associés corrects sur tous les ensembles de données

La concaténation des plongements lexicaux générés par les modèles Transformer et ELMo a permis d'améliorer les modèles, bien que ces mots n'appartiennent pas toujours au vocabulaire d'ELMo. Ainsi, le modèle ELMo semble rapprocher les mots mal orthographiés avec leur forme d'origine puisque la représentation concaténée augmente la similarité moyenne de 19% avec FlauBERT et de 13% avec CamemBERT. Quant à Transformer-POS, l'ajout de contexte morphosyntaxique améliore significativement les résultats d'un cran supplémentaire, avec 44% de similarité cosinus obtenue avec CamemBERT et 41% avec FlauBERT. En résumé, pour traiter plus efficacement les termes hors-vocabulaires, l'ajout de contexte linguistique comme la morphosyntaxe est efficace.

Bio-Gallica De même que sur le corpus issu du domaine juridique, les mots mal orthographiés du corpus – correspondant à des erreurs d'océrisation – sont difficiles à représenter pour les modèles Transformer. Le domaine biomédical est ici plus difficile à traiter que le domaine juridique, pour lequel les mots mal orthographiés étaient issus du domaine général. Non seulement les termes spécifiques au domaine doivent être compris, mais les erreurs doivent également être correctement interprétées (i.e., les variantes orthographiques d'un mot doivent être proches dans l'espace de représentation). Néanmoins, les modèles Transformer ont une meilleure compréhension sémantique des mots mal orthographiés tirés au hasard que ceux du domaine juridique. Les modèles de langue obtiennent 39% et 37% de similarité, respectivement, avec CamemBERT et FlauBERT. Bien que nous ne puissions pas comparer directement les résultats obtenus entre DEFT-Lois et Bio-Gallica, car le contexte des mots ainsi que le type de termes hors-vocabulaires sont différents, il est intéressant de noter les différences de représentation dans les deux domaines. Dans le domaine médical, les modèles Transformer obtiennent des performances médiocres, avec une moyenne de 38% de similarité entre les modèles, mais ont tout de même capturé une certaine proximité sémantique pour les mots mal orthographiés par rapport aux résultats obtenus dans le domaine juridique.

EDF-Courriels Dans ce corpus, nous analysons les fautes d'orthographe sur des termes spécifiques au format du courrier électronique. Cette étude se distingue des autres de par l'importance de la position des mots dans le courriel, et non plus le contexte de ces mots. Par exemple, l'entourage du mot « cordialement » est amené à évoluer d'un courriel à un autre, alors que sa position dans un courriel sera toujours attendue avant la signature. Nous observons que les performances des modèles ont des tendances similaires que sur les autres corpus, à l'exception que l'ajout de contexte morphosyntaxique produit une hausse de performance encore plus forte. En effet, 92% de similitude est obtenue avec CamemBERT-POS et 93% avec FlauBERT-POS, ce qui représente des écarts respectifs de performance de 60% et 59% par rapport aux modèles d'origine. L'ajout de contexte morphosyntaxique apparaît donc très efficace lorsque la position des termes hors-vocabulaire est récurrente dans un corpus.

Nayak et al. [2020] a émis l'hypothèse que les erreurs orthographiques en début de mot sont plus pénalisantes que les autres, car elles affectent plus la tokenisation du reste du mot, indépendamment de la langue traitée. Pour tester cette hypothèse, nous utilisons uniquement ce corpus car il contient des fautes d'orthographe réelles et non des fautes générées par des outils automatiques. Nous sélectionnons trois termes pour leur fréquence importante dans le corpus et leur taux élevé de variantes

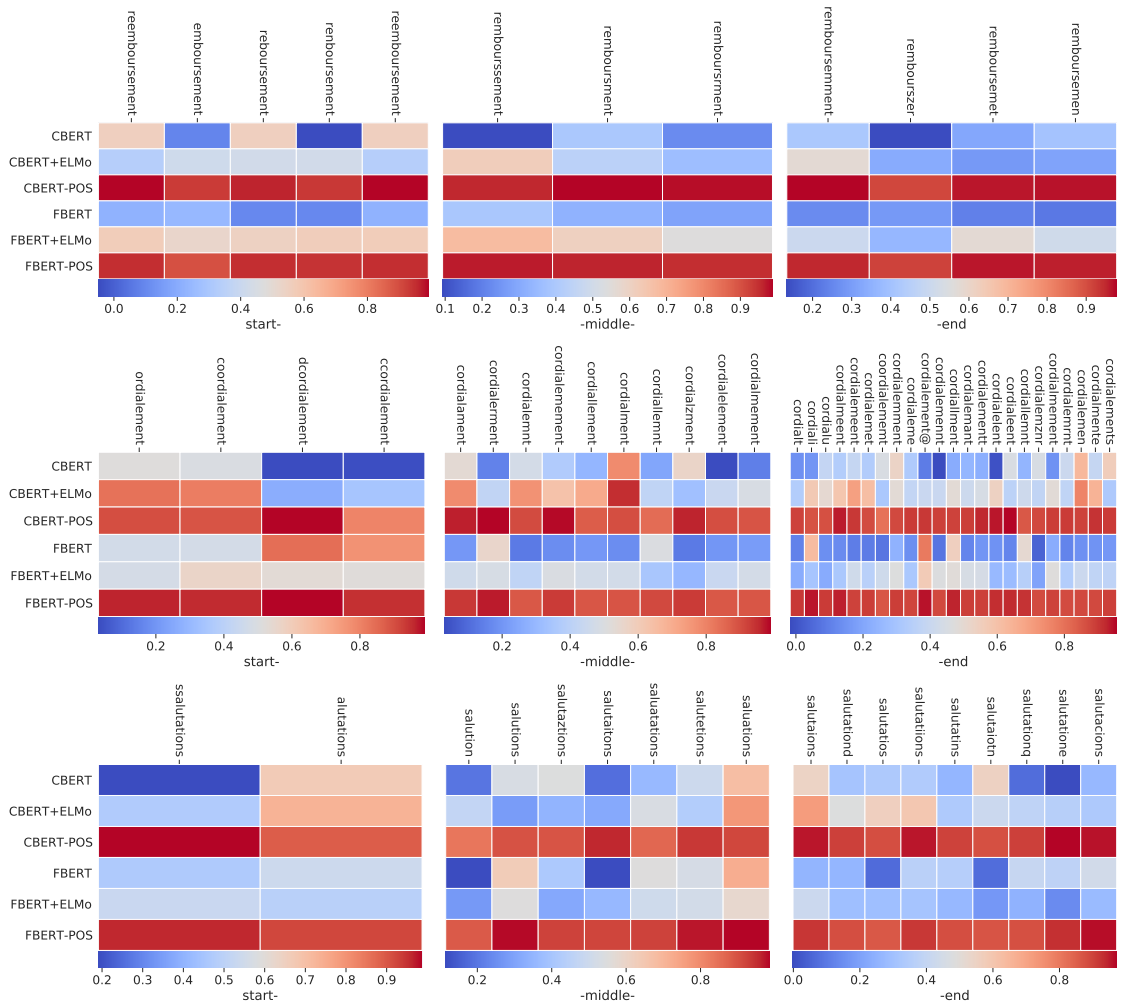


Figure 8.5: EDF-Courriels - Similarité cosinus calculée entre les mots "remboursement" (en haut), "cordialement" (au milieu) et "salutations" (en bas) et leurs variantes orthographiques. Nous distinguons les erreurs au début, au milieu et à la fin des mots (de gauche à droite)

orthographiques : « cordialement », « remboursement » et « salutations » (voir Figure 8.5). Nous choisissons des mots qui apparaissent fréquemment dans le corpus, car nous souhaitons les représenter avec des contextes variés. Nous distinguons les mots mal orthographiés en fonction de la position de l'erreur dans le mot (c'est-à-dire au début, au milieu et à la fin). Nous ne pouvons pas observer de différences significatives entre les erreurs d'orthographe pour ces mots. Cependant, nous précisons que les erreurs du même type ne sont pas équivalentes. C'est le cas du mot « cordialement » (tokénisé en _*cordialement* avec CamemBERT), qui a deux variantes orthographiques similaires : "dcordialement" et "ccordialement" (tokenisé en _*c,cord,iale,ment* avec CamemBERT), mais pour lequel "dcordialement" (découpé en _*d,cord,iale,ment* avec CamemBERT) obtient les meilleurs scores de similarité. C'est pourquoi nous concluons que des fautes d'orthographe tokenisées de

la même façon, peuvent générer des représentations très différentes en fonction des caractères qui sont modifiés (insertion, délétion et substitution). Cette conclusion est problématique, car la représentation ne peut pas être prédite à partir du découpage des mots en tokens.

8.2.5.3 Homographes inter-domaines (*cross-domain homographs*)

Avec cette troisième expérience, nous cherchons à déterminer si les termes spécifiques à un domaine sont traités contextuellement par les modèles Transformer. Plus précisément, nous souhaitons évaluer l'impact du domaine sur le traitement de ses homographes spécifiques. Pour ce faire, nous récupérons quatre termes ayant des significations différentes lorsqu'ils sont utilisés dans des contextes juridiques ou médicaux. La définition[3] de ces mots au sein de chaque domaine est donnée dans le Tableau 8.5.

Mot	Juridique	Médical
Preuve	Élément matériel (par exemple, document contractuel, certificat) qui démontre, indique, prouve la vérité ou la réalité d'une situation de fait ou de droit : preuve d'un crime.	Une preuve scientifique est une preuve utilisée pour soutenir ou réfuter une théorie ou une hypothèse en science.
Filiation	Relation juridique entre les parents et leurs enfants.	La continuité des différentes formes de vie, issues les unes des autres.
Observation	Le fait de se conformer à une règle, une loi, un règlement.	La démarche scientifique d'investigation consiste en l'examen attentif d'un fait, d'un processus, dans le but de mieux le connaître, de le comprendre, et en excluant toute action sur les phénomènes étudiés.
Isolement	Séparation d'un individu – ou d'un groupe d'individus – des autres membres de la société.	Technique de culture des bactéries et des virus permettant de les séparer dans un produit contaminé.

Table 8.5: Définition des homographes dans le contexte légal et dans le contexte médical

Afin d'évaluer si la sémantique des mots représentée par les modèles Transformer est spécifique au domaine (i.e., à son contexte d'utilisation), nous calculons pour

[3]La plupart des définitions proviennent du dictionnaire TLFi : https://www.cnrtl.fr/definition/

chaque mot, au sein de chaque domaine, ses dix plus proches voisins, en utilisant la similarité cosinus. Ensuite, l'objectif est de comparer pour chaque mot, son voisinage dans le domaine juridique et dans le domaine médical. Nous calculons la similarité de voisinage d'un mot entre deux domaines grâce à la formule suivante :

$$similarité = \frac{nombre\ de\ cooccurrences}{n} \tag{8.6}$$

où n est le nombre de voisins que nous utilisons dans les expériences. Ici, nous utilisons n=10. Nous motivons notre approche en affirmant que lorsqu'on calcule des vecteurs de représentation de mots dans un espace multidimensionnel, les voisins d'un mot sont associés à un champ lexical précis et symbolisent la compréhension d'une sémantique capturée par le modèle. Par conséquent, les synonymes sont censés être proches dans l'espace des vecteurs de représentation. Dans nos expériences, les homographes apparaissent dans des contextes différents (i.e., des textes de lois ou des articles de journaux médicaux) avec des vocabulaires différents (i.e., vocabulaires juridique et médical). Par conséquent, ils ne devraient pas partager beaucoup de voisins dans un espace de représentation sémantique.

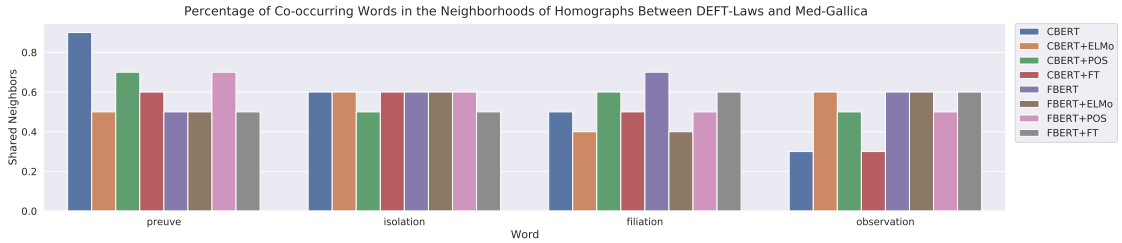


Figure 8.6: Pourcentage de termes co-occurants dans le voisinage des homographes entre le domaine juridique (DEFT-Lois) et le domaine médical (Bio-Gallica). Nous comparons les dix plus proches voisins pour chaque mot dans les deux jeux de données

La similarité obtenue pour les quatre mots sélectionnés avec les modèles Transformer est représentée sous forme d'histogramme (voir Figure 8.6) et la liste des voisins détaillée est fournie pour une meilleure compréhension des résultats (voir Tableau 8.6). Dans l'ensemble, les homographes sont similaires dans les deux domaines puisque nous obtenons au moins 50% de similarité entre leurs voisins pour la plupart des modèles. Par exemple, avec CamemBERT, la similarité entre les deux homographes du mot « preuve » est de 90%. En examinant les voisins du mot, nous remarquons qu'en dehors des mots spécifiques au domaine juridique (e.g., *témoignages, convaincre, contrepartie*), nous obtenons des termes non spécifiques dans le voisinage du mot pour les deux domaines. Nous pensons que cela est dû au fait que les données d'entraînement de CamemBERT et FlauBERT sont générales et que l'affinage des modèles n'a pas suffi à changer le voisinage proche de ces mots. Ce résultat est inattendu, étant donné que la méthode aurait dû favoriser la

Modèle	Voisins - Domaine juridique	Voisins - Domaine médical
Camembert-Base	preuves, confirmation, convaincre, prouve, prouvent, prouvé, particularité, démonstration, prouvant, certitude	preuves, preuveên, confirmation, convaincre, prouver, prouvent, prouvé, particularité, prouvant, démonstration
+ELMo	preuves, certitude, conclusion, confirmation, justification, promesse, différence, contrepartie, démonstration, possibilité	preuves, preuveên, certitude, conclusion, démonstration, vérité, chance, trace, justification, particularité
+POS	preuves, justification, mécanismes, résultant, préservation, arguments, démonstration, corrélation, témoignages, certitude	preuves, mécanismes, justification, enregistrement, arguments, démonstration, vestiges, corrélation, témoignages, renforce
+Affinage	preuves, confirmation, prouvent, prouvé, convaincre, certitude, démontrer, démonstration, prouver, prouvant	preuves, preuveên, convaincre, prouvé, prouvent, prouvant, confirmation, témoignent, particularité, prouve

Table 8.6: Exemples de voisins du mot « preuve » sur les jeux de données DEFT-Lois (à gauche) and Bio-Gallica (à droite) avec quatre modèles : CamemBERT, la concaténation de CamemBERT et d'ELMo, CamemBERT-POS and CamemBERT après affinage

capture d'une sémantique fine des mots de domaine par le modèle. Nous supposons que si l'affinage des modèles améliore la représentation de nouveaux mots, tels que des termes spécifiques ou des mots mal orthographiés, cette pratique n'est pas suffisante sur nos jeux de données pour modifier la représentation des mots du vocabulaire. Sur cette tâche, nous observons que ni les données d'entraînement des modèles, ni les tokéniseurs ne semblent impacter les résultats, étant donné que les modèles CamemBERT et FlauBERT génèrent des résultats similaires. Avec nos travaux, nous montrons qu'aucun modèle Transformer n'a continuellement distingué la sémantique des mots entre les domaines. Dans cette expérience, nous avons utilisé un référentiel classique d'affinage des modèles car notre objectif était d'évaluer les plongements dans une configuration fréquemment utilisée dans la littérature. Naturellement, ces résultats pourraient être améliorés par un affinage des hyperparamètres différent et plus complexe.

8.2.6 Consommation énergétique des modèles

Dans cette section, nous souhaitons quantifier la consommation énergétique des méthodes utilisées précédemment, et démontrer l'impact énergétique de l'utilisation

de BERT-POS par rapport à l'affinage. Les résultats obtenus pour l'évaluation intrinsèque ont été réalisés sur le serveur privé Quadro, détaillé dans l'Annexe C.

Nous calculons l'impact de l'utilisation de notre approche en termes de consommation électrique en kWh (voir Tableau 8.7). Nous observons que la méthode CamemBERT-POS, calculée sur l'architecture Base, consomme 10 fois plus d'énergie que la méthode usuelle de CamemBERT, ce qui est dû aux prétraitements réalisés avec spaCy. Bien que ces chiffres paraissent grands, la consommation énergétique globale est seulement d'un ordre de grandeur de 10^-2 en kWh, ce qui reste faible. En revanche, l'affinage est 1000 fois plus coûteux avec CamemBERT-POS, et génère, sur seulement trois jeux de données, une consommation énergétique finale de 6.28 kWh. En comparaison, cela équivaut à deux cycles de sèche-linge, 6 heures de chauffage en hiver contre 36 heures en été, ou encore 9 jours de travail sur un ordinateur portable.

	DEFT-Lois			**Bio-Gallica**			**EDF-Courriels**			
	PT	Aff.	Encod.	PT	Aff.	Encod.	PT	Aff.	Encod.	**Total**
CBERT-Base	0	0	$2x10^{-4}$	0	0	$6.72x10^{-4}$	0	0	$2x10^{-3}$	$2.9x10^{-3}$
+POS	1.2 x 10^{-5}	0	$2x10^{-4}$	0.02	0	$6.72x10^{-4}$	0.02	0	$2x10^{-3}$	$4.3x10^{-2}$
+Aff.	0	0.82	$2x10^{-4}$	0	2.29	$6.72x10^{-4}$	0	3.12	$2x10^{-3}$	**6.23**
Total (kWh)		0.821			2.312			3.146		**6.280**
Total (kg « équivalent CO2 »)		0.082			0.231			0.315		**0.628**

Table 8.7: Énergie consommée par l'entraînement et l'application des modèles en kWh. PT : Pré-traitements ; Aff. : Affinage du modèle (tokéniseur + poids) ; Encod. : Encodage des données avec le modèle

Le Potentiel de Réchauffement Global (PRG) est l'unité de mesure de l'effet d'un gaz à effet de serre sur le réchauffement global. Grâce au PRG, on peut exprimer l'impact de réchauffement global de chaque gaz à effet de serre (GES) à l'aide d'une unité commune : le kilo ou la tonne « équivalent CO2 ». En informatique, bien que le calcul des émissions est réalisé en fonction de la consommation d'électricité, la consommation généralement décrite est celle du dioxyde de carbone (CO_2). En France, près de 80% de l'électricité provient du nucléaire. Cela implique que le principal gaz à effet de serre émis est la vapeur d'eau (H_2O), loin devant le dioxyde de carbone. Cependant, la vapeur d'eau n'est pas prise en compte dans les bilans GES[4]. En France, un kWh électrique produit environ 0,1 kg de CO_2. Nous effectuons donc une règle de trois pour transformer nos résultats de consommation électrique en impact CO_2, et reportons les résultats dans le Tableau 8.7. La conversion des résultats permet d'ancrer ce travail dans des travaux plus globaux autour de la consommation énergétique, en dehors du cadre informatique.

[4]Pour plus d'informations sur les Bilans GES : https://bilans-ges.ademe.fr/

Afin de représenter au mieux les termes hors-vocabulaire, l'évaluation qualitative a démontré que l'utilisation de CamemBERT-POS était plus pertinente que l'affinage classique des modèles. Cette conclusion est renforcée par l'aspect écologique de l'utilisation de ces modèles, privilégiant une annotation automatique plutôt que l'affinage des modèles Transformer.

8.2.7 Synthèse

En résumé, nous avons évalué plusieurs méthodes pour aider à traiter la sémantique des termes hors-vocabulaire (c'est-à-dire en affinant les modèles de langue, en ajoutant des informations morphosyntaxiques avant l'encodage, et en concaténant la sortie avec une représentation externe). Nous démontrons que l'ajout du contexte morphosyntaxique améliore les représentations pour les trois catégories de termes hors-vocabulaire étudiées. Nous concluons que pour mieux traiter les termes hors-vocabulaire avec les modèles Transformer, l'ajout d'informations structurelles est plus efficace que l'ajout d'informations sémantiques dans les plongements (par exemple, avec un affinage).

Nous avons montré qu'il est plus facile d'améliorer la représentation des nouveaux termes hors-vocabulaire que celle des termes hors-vocabulaire qui existent dans le domaine général (homographes). De plus, nous avons démontré que l'ajout d'informations sur la structure des phrases (c'est-à-dire les balises POS) est bien plus efficace que l'apprentissage de nouveaux mots (affinage). Dans cette étude, nous n'avons pas pu démontrer l'efficacité de l'affinage des modèles Transformer pour améliorer la représentation des termes hors-vocabulaire. Nous analysons les spécificités de trois types de termes hors-vocabulaire (c'est-à-dire les termes spécifiques au domaine, les mots mal orthographiés et les homographes). Nous montrons que si la représentation des nouveaux termes hors-vocabulaire peut être améliorée à l'aide de diverses méthodes, modifier la représentation des mots existants (homographes) reste un défi. L'ajustement des modèles à l'aide d'hyperparamètres spécifiques, combiné à l'augmentation des données, pourrait aider à résoudre ce problème.

Maintenant que nous avons observé l'impact de nos méthodes sur la représentation des termes dans l'espace multidimensionnel, nous nous intéressons à l'évaluation extrinsèque des modèles (cf. Section 8.3) afin de quantifier l'apport de l'ajout d'informations structurelles et sémantiques sur des tâches d'apprentissage automatique.

8.3 Évaluation extrinsèque

Dans cette section, nous souhaitons quantifier l'impact de l'ajout d'informations morphosyntaxique et sémantique dans des modèles Transformer. L'étude conduite dans la Section 8.2 a démontré que l'ajout de morphosyntaxe modifie le nuage des mots et la position des mots dans l'espace euclidien.

Nous souhaitons poursuivre cette étude en quantifiant cet ajout sur différentes tâches. De plus, nous ajoutons également l'ajout d'informations sémantiques, grâce à des entités nommées. Cette expérience a également pour objectif de déterminer quelles tâches bénéficient de quels apports (morphologiques ou sémantiques). Pour cela, nous comparons les performances obtenues sur quatre tâches en anglais et en français : l'analyse de sentiments, la détection de paraphrases, l'inférence et la détection d'entité nommées.

8.3.1 Jeux de données

Dans cette étude, nous utilisons 14 corpus afin d'évaluer les modèles sur quatre tâches : l'analyse de sentiments, la détection de paraphrases, l'implication textuelle et la reconnaissance d'entités nommées et se composent d'avis clientèles, de tweets, de forums, de sous-titres, de pages Wikipédia, ou de documents du domaine médical sur l'anglais et le français. Les corpus sont décrits dans le Tableau 3.6 du Chapitre 3.

8.3.2 Détails des expériences

Sélection des hyperparamètres Nous entraînons les modèles Transformer indépendamment pour chaque jeu de données et pour chaque tâche. Nous utilisons l'optimiseur Adam [Kingma and Ba, 2015] avec un pas d'apprentissage (*learning rate*) fixe. Nous utilisons un optimiseur Adam distinct pour chaque couche que l'on ajoute avec le modèle Embed. Nous recherchons les hyperparamètres suivants : nombre d'*epochs*, taille des *batchs*, pas d'apprentissage et taille des séquences d'entrée. Le choix des hyperparamètres est présenté dans la Tableau 8.8. Dans cette étude, nous sommes limités en ressources GPUs. Par conséquent, nous ne pouvons pas utiliser des tailles de *batchs* plus grandes que 8 ou 16 sur nos jeux de données. Afin d'effectuer une comparaison juste, nous avons entraîné les modèles Transformer avec les mêmes paramètres que les modèles initiaux de RoBERTa et CamemBERT. Nous utilisons la troncation et le padding sur les vecteurs en entrée du modèle. Afin d'améliorer les résultats, nous appliquons deux stratégies de régularisation : la décroissance progressive des poids (*weight decay*) et le réchauffement du taux d'apprentissage (*learning rate warm-up*). Comme le modèle initial est pré-entraîné

et que les nouvelles couches d'entrée ne le sont pas, nous émettons l'hypothèse que le pas d'apprentissage optimal de ces couches devrait être plus grand que les autres (étant donné que l'apprentissage « part de plus loin »). Nous sommes contraints par notre infrastructure pour entraîner nos modèles dans cette étude, ce qui signifie que la taille maximale des *batchs* que nous pouvons utiliser est de 16. Cette limitation a un impact sur nos résultats, qui sont bien inférieurs à ceux rapportés dans la littérature pour les modèles RoBERTa et CamemBERT. En effet, les meilleurs résultats sont généralement obtenus avec des *batchs* de taille 256 pour la plupart des ensembles de données.

Optimisation des hyperparamètres En apprentissage automatique, le processus de recherche de la configuration des hyperparamètres a pour objectif de produire les meilleures performances possibles pour un modèle. Ce processus est généralement manuel et gourmand en ressources informatiques, en particulier lorsque les modèles sont volumineux et longs à entraîner. Nous pouvons distinguer trois méthodes principales d'échantillonnage : (A) l'échantillonnage par grille, (B) l'échantillonnage aléatoire et (C) l'échantillonnage bayésien. Dans cette étude, nous choisissons l'échantillonnage bayésien, basé sur l'algorithme d'optimsation bayésienne. Il choisit une configuration en fonction des performances du modèle obtenues avec la configuration précédente sur la mesure d'évaluation à améliorer (dans notre cas, l'exactitude de l'ensemble de validation). Cette optimisation est moins coûteuse que les autres, car elle ne parcourt pas tout l'espace hyperparamétrique, mais adapte ses recherches en fonction des résultats obtenus précédemment. Ce choix d'approche s'inscrit dans la lignée de l'IA responsable dans laquelle nous souhaitons nous intégrer. Nous présentons des exemples d'optimisation des hyperparamètres pour les tâches d'analyse de sentiments, de détection de paraphrases, d'implication textuelle et de reconnaissance d'entités nommées (REN), dans les Figures 8.7, 8.8, 8.9, et 8.10. La recherche a été effectuée avec l'objectif de maximiser la précision sur l'ensemble de validation, et nous visualisons les hyperparamètres de chaque expérience en fonction de la précision obtenue sur l'ensemble de test.

Hyperparamètres	**Ensemble de valeurs**
Epochs	{3,4,5,...,20}
Taille des *batchs*	{8,16}
Pas d'apprentissage	{1e-5,2e-5,5e-5,1e-6}
Pas d'apprentissage(couche POS)	{1e-5,2e-5,5e-5,1e-6}
Pas d'apprentissage (couche EN)	{1e-5,2e-5,5e-5,1e-6}

Table 8.8: Recherche des hyperparamètres

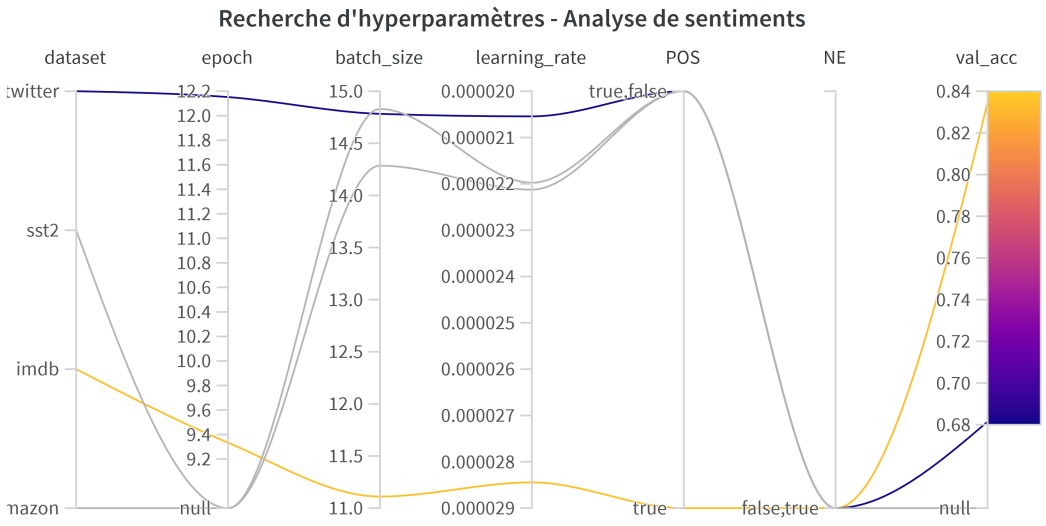


Figure 8.7: Optimisation des hyperparamètres pour la tâche d'analyse de sentiments

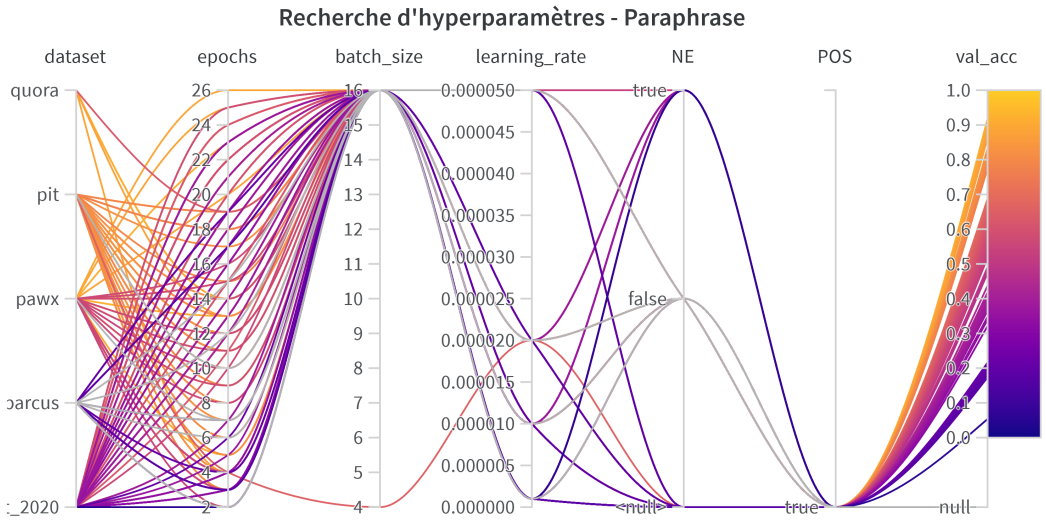


Figure 8.8: Optimisation des hyperparamètres pour la tâche de détection de paraphrases

8.3.3 Expériences

8.3.3.1 Expériences comparatives

Tout d'abord, nous comparons la méthode Transformer-Embed avec les modèles Transformer classiques. Cette comparaison vise, pour chaque tâche, à quantifier l'apport de l'ajout d'informations linguistiques à des modèles Transformer. Nous utilisons les modèles RoBERTa [Liu et al., 2019c] pour l'anglais et CamemBERT [Martin et al., 2020] pour le français. Ici, nous utiliserons les formats Base pour les deux

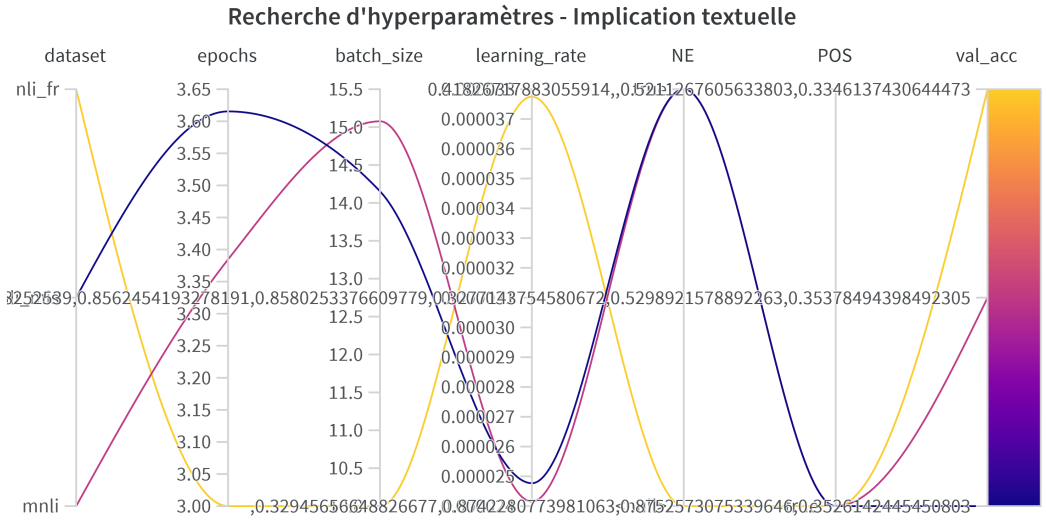


Figure 8.9: Optimisation des hyperparamètres pour la tâche d'implication textuelle

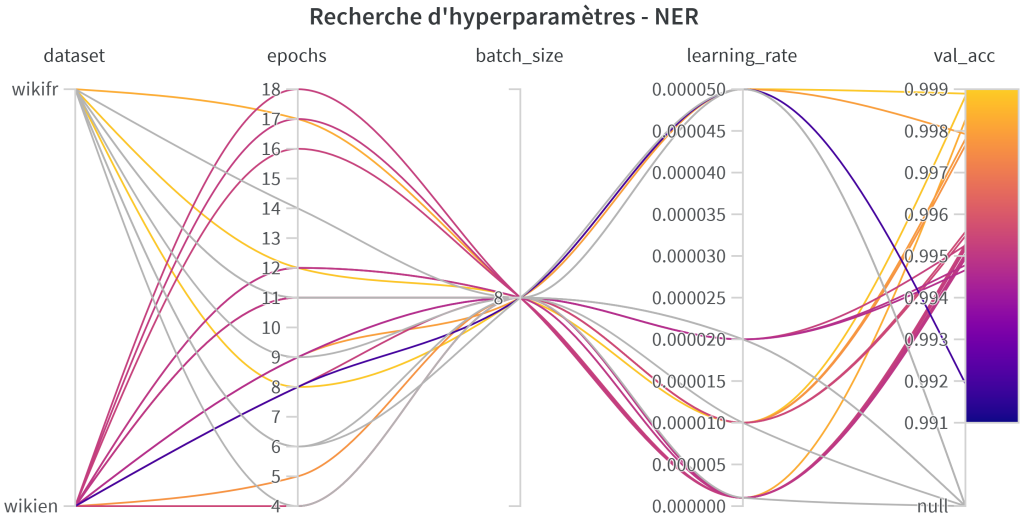


Figure 8.10: Optimisation des hyperparamètres pour la tâche de reconnaissance d'entités nommées

modèles (12 couches, 768 cachés et 12 têtes d'attention).

Ensuite, nous évaluons les performances obtenues par le modèle de Sundararaman et al. [2020], que nous appellerons Syntax-Infused dans la suite de cette étude (en référence au titre de l'article). Cette méthode est pertinente pour nous, car elle est la seule à s'approcher de ce que nous proposons. En effet, elle ajoute également des informations de morphosyntaxe mais elle supprime la couche positionnelle des

Transformer, en argumentant que les étiquettes POS sont un marqueur de position dans la phrase. Dans cet article, les auteurs proposent également d'ajouter la position des tokens grâce à un vecteur. Nous avons implémenté la méthode telle qu'elle était présentée dans l'article d'origine, mais nous utilisons l'architecture RoBERTa et non BERT. Cette méthode a été construite pour résoudre une tâche de traduction automatique et s'est avérée performante. Elle a également obtenu de bons résultats sur le benchmark GLUE. Cependant, nous pensons que la suppression de l'information de position engendre des erreurs par rapport au modèle Transformer d'origine, et que cela risque d'ajouter du bruit. De plus, nous pensons que l'utilisation des marqueurs de morphosyntaxe peuvent difficilement constituer un marqueur de position dans des corpus difficiles à annoter. Nous émettons alors l'hypothèse que notre méthode donnera lieu à l'amélioration des performances sur les quatre tâches ciblées.

8.3.3.2 Mesures d'évaluation

Dans cette étude, nous travaillons sur de nombreux jeux de données (voir Section 3), certains étant déséquilibrés. Nous choisissons trois mesures d'évaluation, fréquemment utilisées dans la littérature, pour toutes les expériences.

La première mesure d'évaluation calculée pour évaluer les modèles est l'exactitude (également appelée justesse ou *accuracy* en anglais). Cette mesure est simple, car elle représente le nombre de classes correctement prédites par le modèle. Concrètement, elle consiste à calculer le nombre de prédictions correctes (vrais positifs ou VP et vrais négatifs ou VN) par rapport au nombre total de prédictions (VP, VN, faux positifs ou FP et faux négatifs ou FN), c'est-à-dire au nombre de documents dans l'ensemble de test. La formule du calcul de l'exactitude est :

$$E = \frac{VP + VN}{VP + FN + VN + FP} \tag{8.7}$$

Dans cette étude, nous travaillons sur des corpus qui sont parfois déséquilibrés. Nous utilisons alors l'exactitude pondérée (*Balanced Accuracy*), définie comme ceci :

$$\text{Exactitude pondérée} = \frac{1}{2} \times \frac{VP}{VP + FN} + \frac{1}{2} \times \frac{VN}{VN + FP}$$

Pour chaque expérience, nous reportons les scores de précision et de rappel. La précision correspond au nombre de prédictions positifs correctement effectuées. En d'autres termes, cette mesure représente le nombre de vrais positifs bien prédit par

rapport à l'ensemble des positifs prédits. Mathématiquement, cela donne :

$$P = \frac{VP}{VP + FP} \tag{8.8}$$

Cette mesure est comprise entre 0 et 1, et plus elle est élevée, plus le modèle minimise le nombre de faux positif.

Le rappel est défini comme la proportion de résultats positifs réelle qui a été correctement identifiée par le modèle. Mathématiquement, elle est calculée comme ceci :

$$R = \frac{VP}{VP + FN} \tag{8.9}$$

En d'autres termes c'est le nombre de positifs bien prédit (Vrai Positif) divisé par l'ensemble des positifs (Vrai Positif + Faux Négatif). Comme pour la précision, le rappel est compris entre 0 et 1.

Étant donné que certains corpus sont déséquilibrés, nous choisissons d'utiliser la « macro-moyenne ». En somme, la macro-moyenne revient à calculer la moyenne des mesures d'évaluation pour chaque classe. En termes mathématiques, on obtient la formule suivante :

$$\text{Macro-mesure} = \frac{1}{n_{class}} \sum_{i=1}^{n_{class}} \text{mesure}_{class_i}$$

Respectivement, le macro-rappel et la macro-précision sont donc la moyenne des mesures de rappel et de précision de chaque classe. Cette mesure accorde autant d'importance à chacune des classes, quelle que soit la proportion d'individus qu'elles contiennent. Cette mesure est donc robuste au déséquilibre des classes, car elle ne néglige pas les classes sous-représentées.

8.3.3.3 Tests statistiques

Dans nos travaux, nous reportons la significativité des résultats grâce à un test de Student apparié (*paired t-test*) réalisé entre les classes réelles de l'ensemble de test et les classes prédites par un modèle sur ces données. Les travaux de Demšar [2006] concluent que ce test est le plus pertinent pour comparer des classifieurs lorsque les jeux de données sont grands (plus de 30 échantillons), et recommande le test des rangs signés de Wilcoxon [Wilcoxon, 1992] dans le cas contraire.

La valeur t de Student est donnée par la formule :

$$t = \frac{\mu}{\sigma/\sqrt{n}} \tag{8.10}$$

où μ et σ représentent la moyenne et l'écart-type de la différence des observations et n est la taille de l'échantillon.

Afin de déterminer si la différence entre les échantillons est significative, il faut lire dans la table de Student, la valeur critique correspondant à un risque α pour un degré de liberté (d.d.l). Ici, nous choisissons $\alpha = 0.05$, qui est le seuil communément utilisé en statistiques. Le degré de liberté est calculé avec $d.d.l = n-1$.

Les hypothèses du test de Student sont les suivantes :

- Hypothèse nulle (H0) : $\mu_1 = \mu_2$
- Hypothèse alternative (H1) : $\mu_1 - \mu_2 = 0$ ou $\mu_1 \neq \mu_2$

avec μ_1 la moyenne des résultats réels et μ_2 la moyenne des résultats observés. Si le test est rejeté, l'hypothèse alternative est validée, ce qui signifie que les résultats obtenus sont statistiquement différents pour un α donné. Sinon, cela signifie que le test n'a pas permis de montrer des différences entre les distributions.

Dans la section suivante, les résultats significativement différents sont indiqués avec des astérisques. Pour chaque jeu de données, le meilleur résultat est indiqué **en gras** lorsqu'il est significativement plus grand que la *baseline*, d'après un test de Student réalisé avec $\alpha = 0.05$.

8.3.3.4 Résultats

	IMDB			**SST-2**		
Modèle	Précision	Rappel	Exactitude	Précision	Rappel	Exactitude
RoBERTa-Base	94.47	94.46	94.46	93.55	93.47	93.47
+Syntax-Infused	90.21*	89.86*	89.91*	85.50*	85.50*	85.50*
+ Embed-POS	95.75	95.75	95.75	94.99	94.95	94.95
+ Embed-EN	95.21	95.21	95.21	95.36	95.28	95.28
+ Embed	**95.88***	**95.88***	**95.88***	95.26	95.23	95.23

Table 8.9: Performances obtenues sur les jeux de données anglais d'analyse de sentiments

Tâche 1 – Analyse de sentiments Pour cette tâche, nous effectuons une étape d'affinage en entraînant une couche de classification par-dessus les architectures de RoBERTa (voir Tableau 8.9) et CamemBERT (voir Tableau 8.10). Les résultats mettent en exergue l'importance d'ajouts des caractéristiques linguistiques pour l'analyse de sentiments Les résultats obtenus sur les jeux de données en anglais démontrent que l'ajout de caractéristiques d'entités nommées et de morphosyntaxe améliorent les performances du modèle RoBERTa-Base. Cependant, la comparaison avec le modèle de base est significative uniquement pour IMDB, pour lequel on remarque une différence d'exactitude de 1,4% pour le modèle complet, principalement due à l'ajout de morpho-syntaxe, qui a lui seul permet d'améliorer cette mesure de 1,2%. En analysant les résultats de plus près (voir Figure 8.11), la matrice de confusion met en évidence un gain de classement plus grand pour les émotions positives (avec 103 commentaires positifs correctement classés en plus par rapport au modèle de base) que pour les émotions négatives (avec 39 commentaires négatifs correctement classés en plus que le modèle de base).

En ce qui concerne les corpus français, on remarque que l'ajout de morpho-syntaxe a le même effet que pour l'anglais, avec des différences significatives notables entre le modèle d'origine et notre modèle, avec un gain de 1,1% d'exactitude sur CLS-FR et de 2,3% sur DEFT-18. Ce résultat met en évidence que la méthode Transformer-Embed semble peu sensible aux erreurs d'annotation en morpho-syntaxe, plus importante sur le corpus de tweets DEFT-18. L'analyse des matrices de confusion (voir Figures 8.13 et 8.14) démontrent que sur le corpus CLS-FR, l'ajout de morpho-syntaxe permet surtout d'améliorer le classement des commentaires négatifs, comme pour IMDB. Nous formulons l'hypothèse que les commentaires positifs contiennent donc une morpho-syntaxe moins spécifique que les commentaires négatifs. De plus, l'ajout de morpho-syntaxe a permis d'améliorer les résultats de classement des classes MIX (à la fois positifs et négatifs) et NEU (neutres) dans les tweets. Au vu des conclusions précédentes, cela signifierait que ces classes se distinguent par une morpho-syntaxe différente de celle contenue dans les tweets ayant une polarité forte, et donc que la morpho-syntaxe serait plus homogène dans ces classes que dans les positifs et négatifs. Nous souhaitons vérifier cette hypothèse en nous intéressant de plus près à la répartition de la morpho-syntaxe au sein des classes (voir Figure 8.15). Nous observons que la répartition est semblable entre les classes, et que les différences ont peu de chances d'affecter le classement final des documents. Nous supposons que l'ajout d'étiquettes morphosyntaxiques n'est pas utile pour l'analyse de sentiments grâce à leur fréquence d'utilisation (exemple : une ponctuation plus fréquente dans les tweets positifs que dans les tweets négatifs), mais à une utilisation différente, qui pourrait être due à l'ordre de l'utilisation de

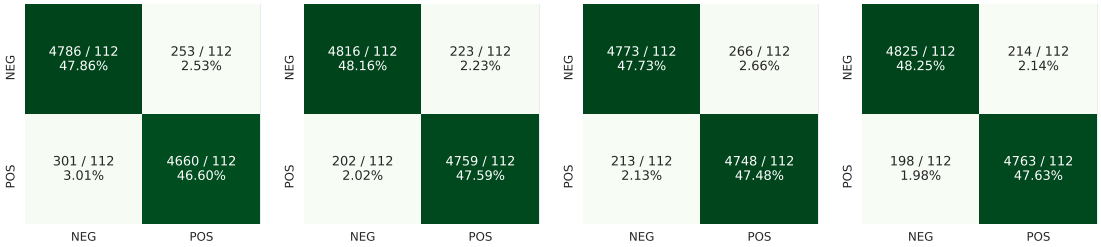
ces étiquettes (exemple : des interjections en début de tweets).

	CLS-FR			DEFT-18		
Modèle	Précision	Rappel	Exactitude	Précision	Rappel	Exactitude
CamemBERT-Base	93.84	93.78	93.78	68.50	66.24	75.29
+Syntax-Infused	86.11*	85.52*	85.56*	69.24*	69.22*	68.71*
+ Embed-POS	94.77*	94.73*	94.73*	**71.49***	**71.91***	**77.64***
+ Embed-EN	94.60	94.60	94.60	71.46*	72.13*	77.44*
+ Embed	**94.84***	**94.83***	**94.83***	71.00*	72.08*	76.24*

Table 8.10: Performances obtenues sur les jeux de données français d'analyse de sentiments

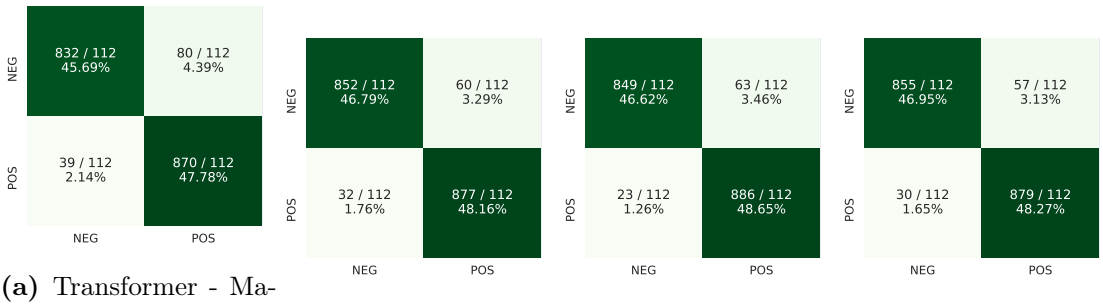
Enfin, l'ajout d'entités nommées est tout particulièrement pertinent pour DEFT-18, mais pas pour CLS-FR car les résultats sont meilleurs que le modèle de base mais ne sont pas significatifs. Nous remarquons que sur tous les jeux de données, l'ajout des entités nommées a le même effet sur les tweets que l'ajout de morpho-syntaxe, c'est-à-dire une amélioration de performance pour les émotions mélangées ou neutres. Nous observons que la répartition des entités nommées est similaire dans les différentes classes de tweets, sauf pour les tweets négatifs qui contiennent plus de lieux (environ 1% de plus que les autres classes) et les tweets neutres qui contiennent plus de MISC. Ces différences de fréquence d'entités nommées, comme pour les étiquettes morpho-syntaxiques, n'expliquent pas à elles seules les différences de classement des documents en sentiments. Afin d'approfondir cette étude, nous répétons l'expérience réalisée précédemment sur les jeux de données en anglais en conservant uniquement les cinq classes d'entités nommées présentes dans le modèle d'annotation en français (voir Tableau 8.11). Nous remarquons que les résultats sont nettement dégradés lorsqu'il y a moins d'entités dans le corpus, et donc que l'on construit des vecteurs plus épars. Avec ces expériences, nous avons démontré que la tâche d'analyse de sentiment bénéficie de connaissances grammaticales (i.e., morphosyntaxe) et sémantiques (i.e., entités nommées). De plus, nous remarquons que la méthode Syntax-Infused n'améliore pas les résultats des modèles d'origine. Nous formulons l'hypothèse que la suppression de la couche positionnelle du modèle perturbe le modèle pré-appris (les poids du modèle sont impactés par la modification de la distribution des données en entrée). Au contraire, nous obtenons de bons résultats avec Transformer-POS, ce qui signifie que la superposition des couches sans modifier l'existant est plus performant.

Tâche 2 - Détection de paraphrases Pour la tâche de détection de paraphrases, nous entraînons également une couche de classement par-dessus le modèle Transformer. En revanche, au lieu d'entraîner le réseau de neurones sur une seule phrase,



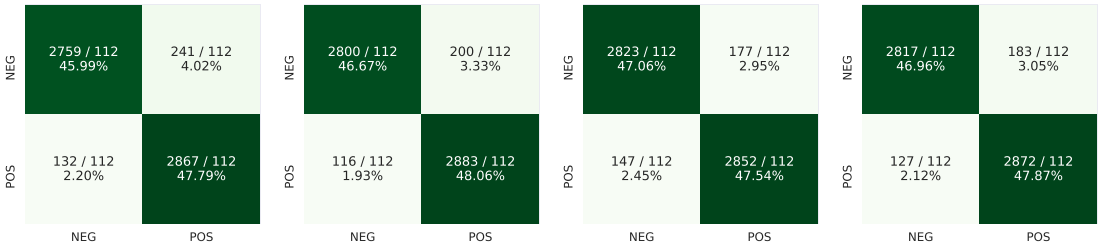
(a) Transformer **(b)** Embed-POS **(c)** Embed-EN **(d)** Embed

Figure 8.11: IMDB



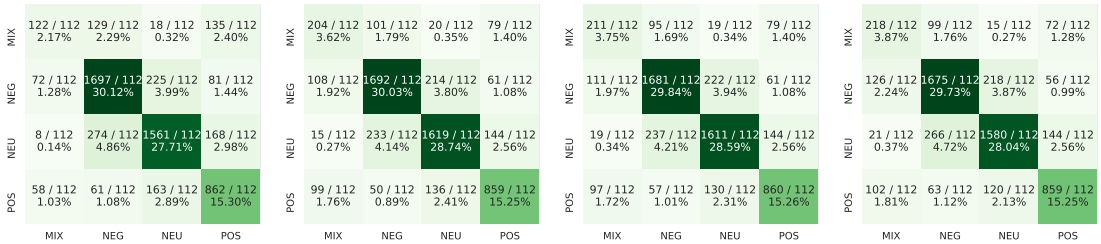
(a) Transformer - Matrices de confusion **(b)** Embed-POS **(c)** Embed-EN **(d)** Embed

Figure 8.12: SST-2 - Matrices de confusion



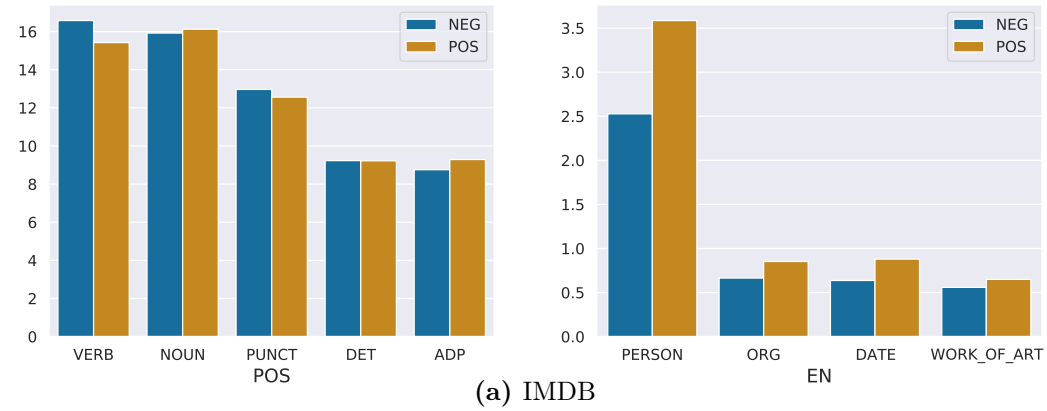
(a) Transformer **(b)** Embed-POS **(c)** Embed-EN **(d)** Embed

Figure 8.13: CLS-FR - Matrices de confusion

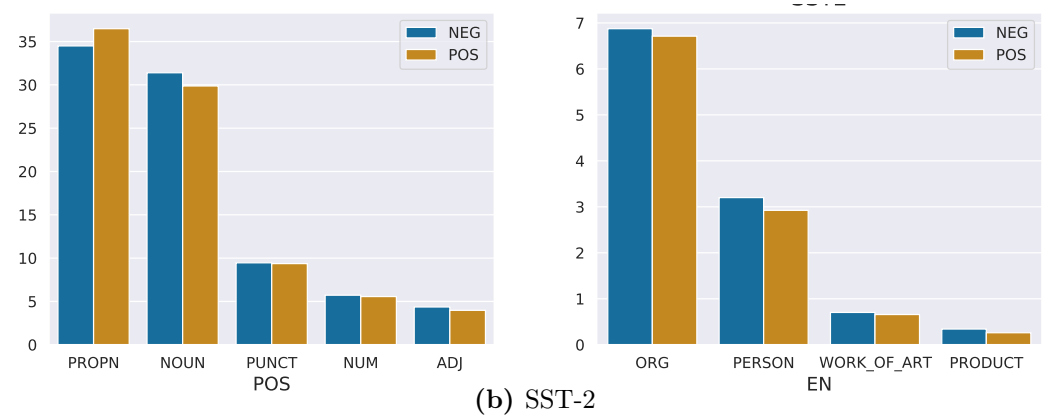


(a) Transformer **(b)** Embed-POS **(c)** Embed-EN **(d)** Embed

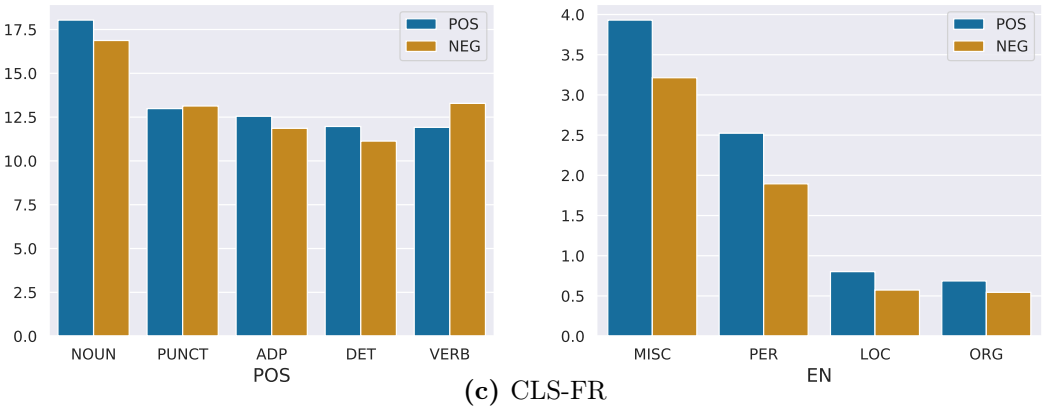
Figure 8.14: Twitter - Matrices de confusion



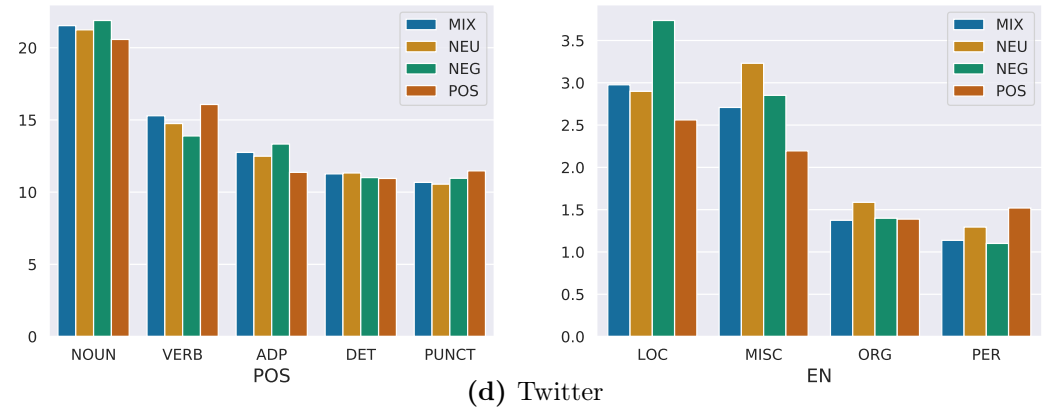
(a) IMDB



(b) SST-2



(c) CLS-FR



(d) Twitter

Figure 8.15: Comparaison de la proportion (en %) des étiquettes morpho-syntaxiques (à gauche) et des entités nommées (à droits) générées par spaCy pour chaque classe dans les corpus d'analyse de sentiments. Les cinq catégories les plus fréquentes sont présentées pour plus de lisibilité

	IMDB			SST-2		
Model	Précision	Rappel	Exactitude	Précision	Rappel	Exactitude
RoBERTa-Base	94.47	94.46	94.46	93.55	93.47	93.47
+ Embed (EN-Complet)	**95.88***	**95.88***	**95.88***	95.26	95.23	95.23
+ Embed (EN-Réduit)	95.40	95.40	95.40	95.04	95.00	95.00

Table 8.11: Performances obtenues sur les jeux de données anglais en étiquetant les corpus avec la liste complète d'entités nommées (EN-complet) et avec la liste réduite (EN-réduit) utilisée pour le français (PERSON, ORG, GPE, LOC, MISC)

nous utilisons des phrases parallèles, en les séparant par le token de séparation du modèle sélectionné. Par exemple, les phrases « elle est allée se coucher » et « elle est allée dormir » seraient concaténées pour obtenir l'entrée suivante : « <s> elle est allée se coucher </s> elle est allée dormir </s> » avec Nous récupérons une fois encore la représentation du premier token, utilisé traditionnellement pour le classement. Le Tableau 8.12 contient les résultats d'exactitude obtenus par les modèles sur les jeux de données de paraphrase.

Nous observons que l'ajout d'informations morphosyntaxiques améliore les performances des modèles Transformer sur tous les jeux de données et pour les deux langues. Cette augmentation est significative pour tous les jeux de données, sauf DEFT-20. L'amélioration significative des performances est très nette sur certains jeux de données, comme Quora (+2,5%), PIT (+7%) et OpusParcus-FR (+12,5%). Ce résultat est très étonnant, étant donné qu'il s'agit des trois jeux de données les plus difficiles à annoter par spaCy, et selon nos estimations, qui contiennent le plus d'erreurs (voir Tableau 7.2). Nous émettons l'hypothèse que ces jeux de données contiennent des structures phrases tellement spécifiques et différentes des jeux d'entraînement, qu'ils bénéficient de connaissances externes sur la structure des phrases, contrairement à des jeux de données issus du domaine général. L'ajout d'entités nommées dégrade les performances pour cette tâche, sauf pour PIT (+5,5%) et OpusParcus-FR (+8%). Comme précédemment, étant donné qu'il s'agit des corpus contenant des styles de rédaction très spécifiques, nous supposons que l'ajout de connaissances sémantiques permet d'améliorer les résultats sur ces corpus, difficiles à traiter pour les modèles Transformer ayant appris sur des données de spécialité.

Globalement, l'ajout d'informations linguistiques est pertinent pour cette tâche, tout particulièrement lorsqu'il s'agit de corpus comportant des styles de rédaction spécifiques. Les erreurs générées par spaCy durant l'étiquetage des mots ne sont pas trop pénalisantes à l'application de notre méthode sur les cinq jeux de données. Enfin, notons que nous n'observons pas de corrélation entre la difficulté de la tâche, représentée par le nombre de classes à prédire (le nombre de niveaux de paraphrases),

et l'amélioration des performances grâce à l'ajout de contexte linguistique. En effet, les corpus Quora, PIT et PAWS-X-FR contiennent deux classes (paraphrase et non paraphrase) et les performances de Transformer-Embed sur ces corpus varient.

Modèle	Quora	PIT	DEFT-20	PAWS-X-FR	OpusParcus-FR
RoBERTa/CamemBERT-Base	87.98	84.64	66.96	90.98	50.17
+Syntax-Infused	86.02*	80.50*	51.79*	56.71*	40.67*
+ Embed-POS	**90.35***	**91.72***	**67.86**	**91.88***	61.25*
+ Embed-EN	87.30	90.07*	66.85	91.01*	58.22*
+ Embed	88.89*	90.19*	66.86	91.19*	**62.92***

Table 8.12: Résultats d'exactitude obtenus sur la tâche de détection de paraphrases. R/C-Base : RoBERTa-Base pour les jeux de données

Tâche 3 - Inférence textuelle La tâche d'inférence textuelle ressemble à la détection de paraphrase, en ce qu'elle nécessite des corpus de phrases parallèles. Par conséquent, nous utilisons la même configuration pour l'entraînement de ces deux tâches, c'est-à-dire l'entrainement du classifieur sur la concaténation des phrases parallèles. Les résultats d'exactitude obtenus sur cette tâche sont présentés dans le Tableau8.13. Nous observons les mêmes résultats qu'en détection de paraphrases et en analyse de sentiments lors de l'ajout de contexte morphosyntaxique. Pour les trois jeux de données, nous observons des gains de performance très importants sur cette tâche, allant jusqu'à une amélioration de l'exactitude de 19% sur le corpus MNLI. La marge est nettement plus importante sur les jeux de données en anglais, avec une moyenne d'amélioration de 13% sur les deux corpus, que sur le jeu de données en français, où nous constatons un gain de 4% (ce qui représente environ 200 paires de phrases). Nous émettons l'hypothèse que ce résultat est dû à des implications textuelles plus redondantes en anglais et en français sur la structure des phrases.

Ensuite, l'ajout des informations sémantiques en entité nommées a permis une amélioration similaire que la morphosyntaxe. Ce résultat n'est pas surprenant, car nous avons remarqué plusieurs redondances d'entités nommées entre les phrases parallèles. Par exemple, la phrase « The best book review section in an American paper at the moment–the Los Angeles Times Book Review –is much closer to this variety model » contient deux étiquettes d'entités nommées pour spaCy : une zone géographique (NORP) « American » et une organization (ORG) « the Los Angeles Times Book Review". La phrase suivante est annotée comme étant le contraire de la première : « The Los Angeles times has never published book reviews », qui est annotée par spaCy avec une entité nommée « groupe » (GPE) pour l'entité « Los Angeles ». Bien qu'on remarque des incohérences d'annotation en entités nommées, le modèle est incapable de prédire le lien entre ces phrases avant l'ajout d'entités nommées. Cela signifie que pour cette tâche, l'apprentissage de relations

sémantiques entre les textes, grâce à l'ajout d'entités nommées, permet d'apprendre des motifs et des combinaisons d'entités nommées pertinents.

Enfin, pour cette tâche, la combinaison de contexte grammatical (avec la morphosyntaxe) et sémantique (avec les entités nommées) permet d'obtenir les meilleurs résultats.

Modèle	**MNLI**	**MNLI-Mis**	**XNLI-FR**
RoBERTa/CamemBERT-Base	72.93	73.04	81.63
Syntax-Infused	52.19	50.92	51.22
+ Embed-POS	91.15	86.94	85.27
+ Embed-EN	89.07	86.85	84.12
+ Embed	**94.27**	**87.89**	**85.65**

Table 8.13: Résultats d'exactitude obtenus pour la tâche d'implication textuelle

Model	**PER**	**ORG**	**LOC**	**MISC**
RoBERTa-Base	86.85	86.71	92.62	85.39
+Syntax-Infused	82.51*	68.05*	75.27*	67.23*
+ Embed-POS	88.15*	**89.48***	92.02	**86.52***
+ Embed-EN	94.28*	88.99*	**93.07***	78.22*
+ Embed	**93.17***	88.62*	92.89	80.53*

Table 8.14: Résultats de F-Mesure pour les tâches de REN sur le corpus WikiNER en anglais

Tâche 4 – Reconnaissance d'entités nommées La tâche de reconnaissance d'entités nommées (REN) est la seule qui évalue le modèle à l'échelle des mots et non à l'échelle des phrases. Pour cette tâche, une couche d'apprentissage est entraînée par-dessus le modèle Transformer et la représentation vectorielle de chaque *token* est utilisée pour les classer. Nous comparons les modèles en appliquant les modèles Transformer sur le jeu de données Wikipédia en anglais (voir Tableau 8.14) et en français (voir Tableau 8.15). Le jeu de données est étiqueté en cinq classes : les personnes (PER), les entreprises et organisations (ORG), les localisations (LOC), et d'autres entités (MISC). La dernière entité est la plus difficile à classer pour les modèles Transformer, ce qui peut s'expliquer par le fait que c'est la moins spécifique. Pour les données en français, toutes les catégories sont mieux classées avec notre méthode et grâce à l'ajout simultané d'entités nommées et de morphosyntaxe. Nous remarquons que l'ajout d'entités a été encore plus bénéfique que l'ajout de tags morphosyntaxique. Cela s'explique par le fait que les catégories d'entités

nommées proposées par spaCy en français correspondent parfaitement (à l'exception de la catégorie de groupes) à celle que l'on souhaite annoter. Les entités nommées ajoutées en amont sont donc tout à fait pertinentes. Pour le classement des *tokens* en anglais, nous observons la même chose pour les catégories de personnes et de localisations. Cependant, pour la catégorie des organisations, seules les étiquettes morphosyntaxiques sont plus efficaces que les étiquettes d'entités nommées. Nous supposons que cela est dû au fait que ce groupe ressemble à plusieurs catégories de spaCy en anglais comme la catégorie NORP. De plus, la catégorie MISC est la mieux catégorisée par Roberta-Base. Nous supposons que trop de catégories d'entités nommées et de morphosyntaxe sont représentées dans cette classe, ce qui a pour effet de perturber le classement de ces *tokens*. Dans l'ensemble, notre méthode a été la plus efficace pour classer les en entités nommées.

Model	**PER**	**ORG**	**LOC**	**MISC**
CamemBERT-Base	94.83	81.52	89.11	81.45
+Syntax-Infused	90.25*	75.12*	69.58*	62.23*
+ Embed-POS	95.99	85.26*	89.87	84.47*
+ Embed-EN	96.03*	85.56*	91.09*	86.14*
+ Embed	**96.53***	**87.38***	**92.09***	**87.12***

Table 8.15: Résultats de F-Mesure pour les tâches de REN sur le corpus WikiNER en français

8.3.4 Consommation énergétique des modèles

Dans cette section, nous souhaitons calculer l'impact énergétique de nos expériences, de façon similaire à ce qui a été réalisé pour les expériences intrinsèques (voir Section 8.2.6). Ces expériences ont toutes été réalisées sur le GPU Tesla (voir Annexe C.

Pour cela, nous calculons l'impact individuel de chaque modèle pour chaque tâche, en agrégeant les valeurs pour chaque corpus. Nous calculons cette fois-ci l'impact global de chaque modèle, sans distinguer les prétraitements de l'affinage des modèles, car nous avons vu précédemment que l'étiquetage automatique effectué par spaCy était dérisoire par rapport à l'impact énergétique de l'affinage. Dans l'ensemble, l'entraînement de tous les modèles (RoBERTa-Base, CamemBERT-Base, Syntax-Infused, Embed, Embed-POS et Embed-EN) ont eu le même impact énergétique durant la phase d'entraînement.

Pour ces expériences, nous avons consommé 427,4 kWh d'éléctricité, soit 42,74 kg d'équivalent CO_2 (voir Section 8.2.6 pour plus d'informations sur la conversion), ce qui représente 3 632,9 km parcourus avec une voiture Renault éléctrique Zoé, 17,8 kg de charbon brûlé et 0,59 plants d'arbres séquestrant le carbone pendant 10 ans.

8.3.5 Synthèse

Dans cette section, nous avons démontré l'intérêt d'ajouter des informations contextuelles syntaxiques et sémantiques dans des modèles de langue. Nous avons proposé deux méthodes d'ajout de connaissances linguistiques dans les modèles de langue avec Transformer-POS et Transformer-Embed. Transformer-POS consiste à ajouter des informations de manière peu coûteuse, en injectant des étiquettes morphosyntaxiques directement dans les données. Transformer-Embed permet quant à elle d'ajouter des informations en ajoutant des vecteurs, afin d'encoder la morphosyntaxe et les entités nommées dans un texte. Ces deux méthodes sont pertinentes pour améliorer la sémantique des termes hors-vocabulaires sur des corpus de domaine général et de spécialité. Nous avons enfin démontré l'efficacité de Transformer-Embed sur quatre tâches (analyse de sentiments, implication textuelle, détection de paraphrases et reconnaissance d'entités nommées) pour deux langues (anglais et français). La méthode Transformer-Embed, du fait de sa facilité de mise en œuvre, peut être utilisée avec d'autres caractéristiques linguistiques et ouvre les portes à d'autres ajouts structurels (e.g., dépendances syntaxiques, chunks) ou sémantiques (e.g., ontologies). Enfin, nous fournissons des renseignements concernant l'ajout de ces connaissances externes sous forme de vecteurs : ces informations ne doivent pas créer des vecteurs trop épars, au risque d'ajouter du bruit durant l'entraînement des modèles.

Conclusion et Discussion de la partie

Dans cette partie, nous avons analysé la représentation des termes hors-vocabulaire dans des corpus de spécialité.

Nous avons détaillé deux méthodes qui consistent à ajouter du contexte linguistique aux modèles Transformer : Transformer-POS et Transformer-Embed. Tandis que la première méthode a vocation à améliorer la représentation des termes hors-vocabulaire dans l'espace de représentation sans affinage (IA verte), la seconde vise à enrichir les modèles Transformer afin d'obtenir de meilleures performances dans diverses tâches. Les méthodes proposées utilisent des modèles open-source et sont parfaitement reproductibles.

Enfin, nous avons terminé cette partie par un chapitre portant sur l'évaluation intrinsèque et extrinsèque de ces méthodes. Nous avons alors discuté des avantages et des inconvénients de chacune d'entre-elles et démontré des gains de performances significatifs à la suite de l'ajout d'informations contextuelles. Nous avons également établi quels types de tâches pouvaient bénéficier de quelles informations et élargi le champ des possibles quant à l'ajout d'informations linguistiques à des modèles de langue.

$* * *$

Partie V

Conclusion et Discussion

On peut trouver le bonheur même dans les moments les plus sombres... Il suffit de se souvenir d'allumer la lumière.

— Harry Potter et le Prisonnier d'Azkaban

9 Conclusion

9.1 Problématique et Contributions

Dans le cadre de cette thèse, nous nous sommes intéressés à la problématique des biais d'apprentissage des modèles de plongements lexicaux, et plus précisément aux biais liés à l'apprentissage de grands volumes de données issues du domaine général pour traiter des données de domaine de spécialité. Nous avons mené des expériences pour évaluer les modèles de plongements lexicaux, enrichir les modèles et évaluer l'empreinte carbone. Nous avons travaillé sur deux langues et une diversité de tâches, dans un objectif final d'amélioration de la représentation des termes hors-vocabulaire.

Évaluation des modèles de plongements lexicaux. *Les trois premières contributions de ce manuscrit sont liées à l'évaluation de modèles de plongements lexicaux statiques et contextuels. Pour cela, nous avons évalué plusieurs aspects liés aux biais d'apprentissage de ces modèles sur de grands corpus issus du domaine général.*

La représentation du genre au théâtre. Nous avons démontré que les modèles de plongements lexicaux statiques capturent des stéréotypes existant dans des pièces de théâtre. Nous avons comparé des travaux de littérature et de sociologie à nos expériences et nous avons démontré l'intérêt interdisciplinaire de la recherche autour de la représentation des biais sociaux en TAL. Nous avons également mis en avant des stéréotypes du théâtre en comparant des modèles appris à partir de zéro et des modèles appris sur des corpus généraux (Wikipédia et Web). Grâce à des études antérieures de sociologie, de littérature

et de sciences cognitives, nous avons mis en avant des traits de caractères typiques des personnages de théâtre. Enfin, nous avons souligné l'importance de poursuivre la recherche sur l'évaluation des stéréotypes, des préjugés ou des biais sociaux dans les modèles de TAL. Bien que des solutions existent pour contrebalancer les biais sociaux, aucune n'a été réellement validée pour résoudre les problèmes de discrimination dans les applications de TAL.

Tokenisation des termes hors-vocabulaire. Dans le Chapitre 6, nous avons démontré que la représentation des termes hors-vocabulaire dans des corpus de spécialité était fortement impactée par leur tokenisation en amont. Nous avons confirmé ce résultat sur trois domaines de spécialité : le droit, la biologie et l'énergie. Nous avons démontré, pour ces trois corpus, que le corpus d'apprentissage de Wikipédia était plus éloigné que ceux du Web, et nous avons démontré l'intérêt de la proximité entre le corpus d'apprentissage et d'application pour la représentation de termes hors-vocabulaire.

Mesurer l'impact de la tokenisation. Dans la Section 6.5, nous proposons deux mesures permettant de quantifier l'impact de la tokenisation des mots sur leur représentation : Dice et Dice-SU. Ces mesures sont les premières à avoir été proposées dans la littérature pour cette application.

Adaptation des modèles Transformer au domaine. *Deux contributions de nos travaux visent à améliorer la représentation des termes dans des corpus de spécialité à travers l'incorporation d'informations linguistiques aux modèles de langue.*

Améliorer la représentation des termes hors-vocabulaire. Dans le Chapitre 7, nous proposons une méthode permettant d'améliorer la représentation des termes hors-vocabulaire pour trois domaines : le droit, l'énergie et la biologie. Nous démontrons que cette méthode est plus efficace que l'affinage sur deux types de termes : les variantes orthographiques (i.e., les termes mal orthographiés) et les termes spécifiques à des domaines. Nous démontrons qu'aucune méthode ne permet d'adapter les modèles efficacement pour adapter des représentations existantes à des domaines spécifiques lors du traitement d'homographes.

Enrichir les modèles pour des tâches de TAL. Dans le Chapitre 7, nous proposons une méthode d'enrichissement des modèles Transformer en utilisant des connaissances linguistiques, syntaxiques et sémantiques. Nous démontrons que cette méthode permet d'améliorer significativement les performances des modèles Transformer sur quatre tâches : l'analyse de sentiments, la détection de paraphrases, l'implication textuelle et la reconnaissance d'entités nommées. La méthode proposée est générique et peut-être appliquée à n'importe quelles connaissances extérieures, syntaxiques ou sémantiques.

Ajout de connaissances externes. Notre recherche a montré que l'évaluation et l'adaptation de modèles de TAL pour réduire les biais d'apprentissage sur des données de spécialité peuvent être améliorées par l'utilisation du contexte linguistique. Nous avons démontré que l'intégration d'informations contextuelles liées à la spécialité dans les modèles de TAL permet d'améliorer significativement les performances de ces derniers sur des tâches spécifiques à la spécialité. Ces résultats suggèrent que l'utilisation de contexte linguistique peut être un moyen efficace pour réduire les biais d'apprentissage dans les modèles de TAL et améliorer leur performance sur des données de spécialité.

Empreinte carbone des expériences. *Dans nos travaux, nous avons veillé à effectuer une recherche écologiquement responsable. Pour cela, nous avons appliqué trois principes tout au long de ce manuscrit :*

1. ***Limiter les expériences :*** *chaque expérience a été définie en amont afin de réaliser le minimum de calculs possibles en comparant un nombre restreint de modèles pour répondre à une problématique. Dans cette recherche, nous pensons que l'on se dirige vers une uniformisation des modèles Transformer, où l'important n'est plus de comparer différentes optimisations du modèle, mais plutôt d'essayer d'enrichir l'architecture existante.*

2. ***Calculer la consommation énergétique des modèles :*** *pour chaque expérience comparative effectuée dans le Chapitre 7, nous avons évalué l'impact énergétique des calculs réalisés en termes de consommation de CO_2 et de durée des calculs.*

3. ***Proposition d'une méthode peu coûteuse :*** *nous proposons BERT-POS, une méthode d'enrichissement des modèles visant à remplacer l'affinage pour des applications de représentation de termes hors-vocabulaire (e.g., calculs de similarité ou résolution d'analogies). Cette méthode est moins coûteuse en ressources que l'affinage.*

9.2 Perspectives

Nous prévoyons d'étudier plus en détail les possibilités d'ajout d'informations linguistiques à des modèles de TAL. Plus précisément, nous avons fait le choix d'implémenter des méthodes simples et généralisables sur nos corpus. L'utilisation des graphes de connaissances, présentés dans le Chapitre 2, permettrait de retranscrire plus finement les connexions entre les mots ainsi que les liens entre les mots, leur nature et leur fonction. Ces méthodes permettraient de structurer les

données avant de les transformer grâce à des modèles pré-entraînés tels que des Transformer. De plus, nous souhaiterions déterminer si l'utilisation de sous-couches d'informations linguistiques supplémentaire dans la méthode Transformer-Embed serait à envisager, avec les locutions adverbiales pour renforcer le contexte entre les mots qui les composent, ou encore avec des scores de valence de polarité pour améliorer la tâche de sentiments.

Dans notre recherche, nous avons souhaité réaliser des solutions écologiques (qui sont également économiques) pour répondre à notre problématique. Or, des modèles de distillation (plus petits que les modèles Transformer usuels) ont vu le jour, tels que DistillCamemBERT, et répondent à un usage plus écologique de cette architecture (mais pas à leur entraînement). Nous imaginons que les deux méthodes proposées seraient également utiles à ces méthodes, notamment pour combler la perte d'information engendrée par le processus de distillation.

L'arrivée récente de ChatGPT[1] et de BARD[2] marque un tournant dans la représentation des mots et des documents en traitement automatique des langues. Bien que nos travaux se soient principalement concentrés sur les modèles Transformer, ce manuscrit démontre l'importance de combiner des méthodes d'apprentissage avec des informations linguistiques. Il serait intéressant de déterminer comment ChatGPT modélise la langue, par comparaison avec les modèles Transformer, et s'il est intéressant ou non d'y incorporer d'autres caractéristiques.

[1] https://chat.openai.com/
[2] https://blog.google/technology/ai/bard-google-ai-search-updates/

La peur d'un nom ne fait qu'accroître la peur de la chose elle-même.

— Harry Potter et la Chambre des Secrets

Bibliography

F. Alfaro. Study and experimentation of gender bias in co-reference resolution. Master's thesis, Universitat Politècnica de Catalunya, 2019.

M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018.

L. F. W. Anthony, B. Kanding, and R. Selvan. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051*, 2020.

K. Arora, K. Shuster, S. Sukhbaatar, and J. Weston. Director: Generator-classifiers for supervised language modeling, 2022. URL https://arxiv.org/abs/2206.07694.

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

J. Azé, T. Heitz, A. Mela, A. Mezaour, P. Peinl, and M. Roche. Présentation de deft'06 (defi fouille de textes). *Proceedings of DEFT*, 6(1):3–12, 2006.

S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

T. Bachlechner, B. P. Majumder, H. Mao, G. Cottrell, and J. McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in Artificial Intelligence*, pages 1352–1361. PMLR, 2021.

K. Bagla, A. Kumar, S. Gupta, and A. Gupta. Noisy text data: Achilles' heel of popular transformer based NLP models. *CoRR*, abs/2110.03353, 2021. URL https://arxiv.org/abs/2110.03353.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, pages 1–15, 2015.

A. Bakarov. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*, 2018.

S. Bechhofer and C. Goble. Thesaurus construction through knowledge representation. *Data & knowledge engineering*, 37(1):25–45, 2001.

S. Becker and G. E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

Y. Belinkov and Y. Bisk. Synthetic and natural noise both break neural machine translation. *International Conference on Learning Representations (ICLR)*, 2018.

R. Bellman and R. Kalaba. A mathematical theory of adaptive control processes. *Proceedings of the National Academy of Sciences*, 45(8):1288–1290, 1959.

I. Beltagy, A. Cohan, and K. S. Lo. Pretrained contextualized embeddings for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China*, pages 3–7, 2019a.

I. Beltagy, K. Lo, and W. Ammar. Combining distant and direct supervision for neural relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1858–1867, Minneapolis, Minnesota, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/N19-1184. URL https://www.aclweb.org/anthology/N19-1184.

A. Benamar. Segmentation de texte non-supervisée pour la détection de thématiques à l'aide de plongements lexicaux. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3: Rencontre des Étudiants Chercheurs en Informatique pour le TAL*, pages 1–14. ATALA; AFCP, 2020.

A. Benamar, M. Bothua, C. Grouin, and A. Vilnat. Easy-to-use combination of POS and BERT model for domain-specific and misspelled terms. In *NL4IA Workshop Proceedings*, Milan, Italy, Nov. 2021.

A. Benamar, C. Grouin, M. Bothua, and A. Vilnat. Etude des stéréotypes genrés dans le théâtre français du XVIe au XIXe siècle à travers des plongements lexicaux. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 74–81, Avignon, France, 6 2022a. ATALA.

A. Benamar, C. Grouin, M. Bothua, and A. Vilnat. Evaluating Tokenizers Impact on OOVs Representation with Transformers Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4193–4204, Marseille, France, June 2022b. European Language Resources Association.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, 2006.

Y. Bengio, Y. LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5):1–41, 2007.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128, 2006.

S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.

T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

K. Bostrom and G. Durrett. Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*, 2020.

J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

R. Cardon, N. Grabar, C. Grouin, and T. Hamon. Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques (presentation of the DEFT 2020 challenge : open domain textual similarity and precise information extraction from clinical cases). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi*

Fouille de Textes, pages 1–13, Nancy, France, 6 2020. ATALA et AFCP. URL https://www.aclweb.org/anthology/2020.jeptalnrecital-deft.1.

O. Cattan, C. Servan, and S. Rosset. On the Usability of Transformers-based models for a French Question-Answering task. In *Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria, Sept. 2021. URL https://hal.science/hal-03336060.

A. Celebi, H. Sak, E. Dikici, M. Saraçlar, M. Lehr, E. Prud'hommeaux, P. Xu, N. Glenn, D. Karakos, S. Khudanpur, et al. Semi-supervised discriminative language modeling for turkish asr. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5025–5028. IEEE, 2012.

W. Che, Y. Liu, Y. Wang, B. Zheng, and T. Liu. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, Oct. 2018a. Association for Computational Linguistics. doi: 10.18653/v1/K18-2005. URL https://www.aclweb.org/anthology/K18-2005.

W. Che, Y. Liu, Y. Wang, B. Zheng, and T. Liu. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, October 2018b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K18-2005.

M. Chen, K. Q. Weinberger, and J. Blitzer. Co-training for domain adaptation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/93fb9d4b16aa750c7475b6d601c35c2c-Paper.pdf.

T. Chen, R. Xu, Y. He, and X. Wang. Improving distributed representation of word sense via WordNet gloss composition and context clustering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 15–20, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2003. URL https://www.aclweb.org/anthology/P15-2003.

B. Chiu, A. Korhonen, and S. Pyysalo. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP*, pages 1–6, 2016.

C. Clement, D. Drain, J. Timcheck, A. Svyatkovskiy, and N. Sundaresan. PyMT5: multi-mode translation of natural language and python code with transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9052–9065, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.728. URL https://www.aclweb.org/anthology/2020.emnlp-main.728.

T. M. Cover and J. A. Thomas. Information theory and the stock market. *Elements of Information Theory. Wiley Inc., New York*, pages 543–556, 1991.

M. Creutz. Opusparcus: Open Subtitles Paraphrase Corpus for Six Languages (version 1.0), 2018. URL http://urn.fi/urn:nbn:fi:lb-2018021221.

F. Da. En quoi le valet est-il maître du jeu au théâtre? *Publications Pimido*, 2009.

I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer, 2005.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

G. Douzas, M. Lechleitner, and F. Bacao. Improving the quality of predictive models in small data gsdot: A new algorithm for generating synthetic data. *Plos one*, 17(4):e0265626, 2022.

A. Drozd, A. Gladkova, and S. Matsuoka. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530, 2016.

G. Dubuisson Duplessis, E. Bartholme, S. Kerroua, M. Poulain, A. Roulier, and A.-L. Guénet. Désidentification de données texte produites dans un cadre de relation client (de-identification of customer relationship text data). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4 : Démonstrations et résumés d'articles internationaux*, pages 10–13, Nancy, France, 6 2020. ATALA et AFCP. URL https://www.aclweb.org/anthology/2020.jeptalnrecital-demos.3.

J. Ebrahimi, D. Lowd, and D. Dou. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, 2018.

L. Ehrlinger and W. Wöß. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2, 2016.

N. C. Ellis. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, 24(2):143–188, 2002.

M. Fares, A. Kutuzov, S. Oepen, and E. Velldal. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden, May 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-0237.

J.-P. Fauconnier. French word embeddings, 2015. URL http://fauconnier.github.io.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, 2008.

J. R. Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

W. Foundation. Wikimedia downloads. URL https://dumps.wikimedia.org.

W. N. Francis and H. Kucera. Brown corpus manual. *Letters to the Editor*, 5(2):7, 1979.

P. Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

Y. Geng, J. Chen, Z. Ye, Z. Yuan, W. Zhang, and H. Chen. Explainable zero-shot learning via attentive graph convolutional network and knowledge graphs. *Semantic Web*, 12(5):741–765, 2021. doi: 10.3233/SW-210435. URL https://doi.org/10.3233/SW-210435.

H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, 2019. doi: 10.18653/v1/n19-1061. URL https://doi.org/10.18653/v1/n19-1061.

E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*, 2018.

T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995.

S. Gruffat. La représentation du héros amoureux dans les tragédies classiques: pour une conception évolutive du moi? *Litteratures classiques*, (1):143–160, 2012.

X. Guan, Y. Li, and H. Gong. Improved tf-idf for we media article keywords extraction. In *Journal of Physics: Conference Series*, volume 1302, page 032003. IOP Publishing, 2019.

A. Gunny. Emelina, jean: Les valets et les servantes dans le théâtre comique en france de 1610 à 1700. *Zeitschrift für französische Sprache und Literatur*, 88: 265–266, 1978.

H.-M. Haav. A practical methodology for development of a network of e-government domain ontologies. In *Conference on e-Business, e-Services and e-Society*, pages 1–13. Springer, 2011.

J. Halvorsen and B. J. Hansen. Integrating military systems using semantic web technologies and lightweight agents. 2011.

X. Han and J. Eisenstein. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, 2019.

Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

B. He et al. Bert-mk: Integrating graph contextualized knowledge into pre-trained language models, 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

D. Hovy. Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 752–762, 2015.

A. Howard and J. Borenstein. Trust and bias in robots: These elements of artificial intelligence present ethical challenges, which scientists are trying to solve. *American Scientist*, 107(2):86–90, 2019.

P.-Y. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35, 2009.

M. Hudon. Multilingual thesaurus construction–integrating the views of different cultures in one gateway to knowledge and concepts. *Information services & use*, 17(2-3):111–123, 1997.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

P. Jaccard. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241 – 272, 1901.

L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

K. S. Jones. Idf term weighting and ir research lessons. *Journal of documentation*, 2004.

K. D. Joshi and P. Nalwade. Modified k-means for better initial cluster centres. *International Journal of Computer Science and Mobile Computing*, 2(7):219–223, 2013.

A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

D. Kahneman and A. Tversky. On the psychology of prediction. *Psychological review*, 80(4):237, 1973.

P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.567. URL https://www.aclweb.org/anthology/2020.emnlp-main.567.

A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

S. A. Khan and H.-T. Chang. Comparative analysis on facebook post interaction using dnn, elm and lstm. *PloS one*, 14(11):e0224452, 2019.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. 2015. URL http://arxiv.org/abs/1412.6980.

R. Krishnan, P. Rajpurkar, and E. J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, pages 1–7, 2022.

T. Kudo and J. Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In E. Blanco and W. Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-2012. URL https://doi.org/10.18653/v1/d18-2012.

A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres. Quantifying the carbon emissions of machine learning, 2019. URL https://arxiv.org/abs/1910.09700.

L. Lacy, G. Aviles, K. Fraser, W. Gerber, A. M. Mulvehill, and R. Gaskill. Experiences using owl in military applications. In *OWLED*, volume 188, 2005.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019. URL http://arxiv.org/abs/1909.11942.

H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020a. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.302.

H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab. FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français (FlauBERT : Unsupervised language

model pre-training for French). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 268–278, Nancy, France, 6 2020b. ATALA et AFCP. URL https://www.aclweb.org/anthology/2020.jeptalnrecital-taln.26.

H. Lee and A. Y. Ng. Spam deobfuscation using a hidden markov model. CEAS, Citeseer, 2005.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

G. Lejeune and A. Barbaresi. Bien choisir son outil d'extraction de contenu à partir du web (choosing the appropriate tool for web content extraction). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4 : Démonstrations et résumés d'articles internationaux*, pages 46–49, Nancy, France, 6 2020. ATALA et AFCP. URL https://www.aclweb.org/anthology/2020.jeptalnrecital-demos.12.

H. Levesque, E. Davis, and L. Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

S. Li and C. Zong. Multi-domain adaptation for sentiment classification: Using multiple classifier combining methods. In *2008 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–8. IEEE, 2008.

Z. Li, M. Zhang, W. Che, T. Liu, and W. Chen. Joint optimization for chinese pos tagging and dependency parsing. *IEEE/ACM transactions on audio, speech, and language processing*, 22(1):274–286, 2013.

C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou. Autoencoder for words. *Neurocomputing*, 139:84–96, 2014.

M. Liu, Y. Song, H. Zou, and T. Zhang. Reinforced training data selection for domain adaptation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1957–1968, 2019a.

S. Liu, L. Wang, V. Chaudhary, and H. Liu. Attention neural model for temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 134–139, Minneapolis, Minnesota, USA, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/W19-1917. URL https://www.aclweb.org/anthology/W19-1917.

W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908, 2020.

X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

X. Liu, C. Yoo, F. Xing, H. Oh, G. El Fakhri, J.-W. Kang, J. Woo, et al. Deep unsupervised domain adaptation: a review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019c.

J. B. Lovins. Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, 11(1-2):22–31, 1968.

Y. Lü, J. Huang, and Q. Liu. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350, 2007.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

N. Maija. La relation maître-valet dans le théâtre de comédie. *Publications Pimido*, 2012.

H. H. Mao. A survey on self-supervised pre-training for sequential transfer learning in neural networks. *arXiv preprint arXiv:2007.00800*, 2020.

C. Marcandier. Le théâtre, 2011.

M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.

M. Mars. From word embeddings to pre-trained language models: A state-of-the-art walkthrough. *Applied Sciences*, 12(17):8805, 2022.

L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.645. URL https://www.aclweb.org/anthology/2020.acl-main.645.

J. McAuley, R. Pandey, and J. Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2015.

B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30, 2017.

Y. Mejova and P. Srinivasan. Crossing media streams with sentiment: Domain adaptation in blogs, reviews and twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 234–241, 2012.

O. Melamud, J. Goldberger, and I. Dagan. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 51–61, 2016.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL http://arxiv.org/abs/1301.3781.

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

A. More, Ö. Çetinoğlu, Ç. Çöltekin, N. Habash, B. Sagot, D. Seddah, D. Taji, and R. Tsarfaty. Conll-ul: Universal morphological lattices for universal dependency parsing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

F. Naugrette. Le devenir des emplois comiques et tragiques dans le théâtre de hugo. *Littératures classiques*, 48(1):215–225, 2003.

A. Nayak, H. Timmapathini, K. Ponnalagu, and V. Gopalan Venkoparao. Domain adaptation challenges of BERT in tokenization and sub-word representations of out-of-vocabulary words. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 1–5, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.insights-1.1. URL https://www.aclweb.org/anthology/2020.insights-1.1.

A. Neelakantan, J. Shankar, A. Passos, and A. McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1113. URL https://www.aclweb.org/anthology/D14-1113.

J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194: 151–175, 2013.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

M. Paolucci, K. Sycara, T. Nishimura, and N. Srinivasan. Toward a semantic web e-commerce. In *Proc. of 6th Int. Conf. on Business Information Systems (BIS'2003)*, 2003.

C. Parisse. La morphosyntaxe : Qu'est ce qu'est ? - Application au cas de la langue française ? *Rééducation orthophonique*, 47(238):7–20, 2009. URL https://halshs.archives-ouvertes.fr/halshs-00495626.

S. Park, A. Fazly, A. Lee, B. Seibel, W. Zi, and P. Cook. Classifying out-of-vocabulary terms in a domain-specific social media corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2971–2975, 2016.

P. Paroubek, A. Vilnat, I. Robba, and C. Ayache. Les résultats de la campagne easy d'évaluation des analyseurs syntaxiques du français. In *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. Posters*, pages 243–252, 2007.

P. Paroubek, C. Grouin, P. Bellot, V. Claveau, I. Eshkol-Taravella, A. Fraisse, A. Jackiewicz, J. Karoui, L. Monceaux, and J.-M. Torres-Moreno. Deft2018: Recherche d'information et analyse de sentiments dans des tweets concernant les transports en île de france. In *DEFT 2018-14ème atelier Défi Fouille de Texte*, volume 2, pages 1–11, 2018.

V. M. K. Peddinti and P. Chintalapoodi. Domain adaptation in sentiment analysis of twitter. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL http://aclweb.org/anthology/N18-1202.

B. Plank. What to do about non-standard (or non-canonical) language in nlp. *Bochumer Linguistische Arbeitsberichte*, page 13, 2016.

B. Plank and A. Moschitti. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1498–1507, 2013.

N. Poerner, U. Waltinger, and H. Schütze. E-bert: Efficient-yet-effective entity embeddings for bert. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, 2020.

O. Polatbilek. *Enriching Contextual Word Embeddings with Character Information*. PhD thesis, Izmir Institute of Technology (Turkey), 2020.

C. Prakash, R. Kumar, and N. Mittal. Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges. *Artificial Intelligence Review*, 49(1):1–40, 2018.

Y. Qiu, H. Li, S. Li, Y. Jiang, R. Hu, and L. Yang. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In *Chinese computational linguistics and natural language processing based on naturally annotated big data*, pages 209–221. Springer, 2018.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. OpenAI, 2018.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

A. Ramponi and B. Plank. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, 2020.

K. Ravi and V. Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46, 2015.

R. A. Raymondie and D. D. Steiner. Stéréotypes de genre concernant l'expression des émotions: pensez subordonné–pensez femme? *Carrières, leadership et conflits*, page 203, 2020.

R. Rhouma and P. Langlais. Experiments in learning to solve formal analogical equations. In *International Conference on Case-Based Reasoning*, pages 612–626. Springer, 2018.

M. T. Ribeiro, S. Singh, and C. Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

E. Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the national conference on artificial intelligence*, pages 1044–1049, 1996.

J. A. Rodger and P. C. Pendharkar. A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies*, 60(5-6):529–544, 2004.

D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2): 1–96, 2022.

C. Rossari. Projet pour une typologie des opérations de reformulation. *Cahiers de linguistique française*, 11(345-359), 1990.

A. Roy, D. Ghosal, E. Cambria, N. Majumder, R. Mihalcea, and S. Poria. Improving zero shot learning baselines with commonsense knowledge. *CoRR*, 2020.

S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-5004. URL https://www.aclweb.org/anthology/N19-5004.

B. Sagot and P. Boullier. Sxpipe 2: architecture pour le traitement pré-syntaxique de corpus bruts. *Revue TAL*, 49(2):155–188, 2008.

V. Sahil and R. Julia. Fairness definitions explained. In *Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (FairWare'18)*, pages 1–7, 2018.

J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51, 2009.

T. Schnabel, I. Labutov, D. Mimno, and T. Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 298–307, 2015.

E. W. Schneider. Course modularization applied: The interface system and its implications for sequence control and data analysis. 1973.

R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

H. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos. Compute trends across three eras of machine learning. *arXiv preprint arXiv:2202.05924*, 2022.

R. Snow, B. O'connor, D. Jurafsky, and A. Y. Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263, 2008.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

R. Soma. *Applying semantic web technologies for information management in domains with semi-structured data*. University of Southern California, 2008.

R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

M. Stede and R. Patz. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18, 2021.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://www.aclweb.org/anthology/P19-1355.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13693–13696, 2020.

L. Sun, K. Hashimoto, W. Yin, A. Asai, J. Li, P. Yu, and C. Xiong. Adv-BERT: BERT is not robust on misspellings! generating nature adversarial samples on BERT. *arXiv preprint arXiv:2003.04985*, 2020.

D. Sundararaman, S. Si, V. Subramanian, G. Wang, D. Hazarika, and L. Carin. Methods for numeracy-preserving word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4742–4753, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.384. URL https://www.aclweb.org/anthology/2020.emnlp-main.384.

J. Tang. Aminer: Toward understanding big scholar data. In *Proceedings of the ninth ACM international conference on web search and data mining*, pages 467–467, 2016.

M. Thelwall. Gender bias in sentiment analysis. *Online Inf. Rev.*, 42(1):45–57, 2018. doi: 10.1108/OIR-05-2017-0139. URL https://doi.org/10.1108/OIR-05-2017-0139.

H. Tian, C. Gao, X. Xiao, H. Liu, B. He, H. Wu, H. Wang, and F. Wu. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, 2020.

S. Tolan, M. Miron, E. Gómez, and C. Castillo. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 83–92, 2019.

T. H. Trinh and Q. V. Le. A simple method for commonsense reasoning. arXiv, 2018. doi: 10.48550/ARXIV.1806.02847. URL https://arxiv.org/abs/1806.02847.

L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9: 2579–2605, 2008.

M. van der Wees, A. Bisazza, and C. Monz. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, 2017.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

G. Vrbančič and V. Podgorelec. Transfer learning with adaptive fine-tuning. *IEEE Access*, 8:196197–196211, 2020. doi: 10.1109/ACCESS.2020.3034343.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018a.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019a.

B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C. J. Kuo. Evaluating word embedding models: methods and experimental results. *APSIPA transactions on signal and information processing*, 8, 2019b.

B. Wang, C. Xu, S. Wang, Z. Gan, Y. Cheng, J. Gao, A. H. Awadallah, and B. Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021.

D. Wang and T. F. Zheng. Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237. IEEE, 2015.

Q. Wang, Z. Mao, B. Wang, and L. Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017a.

R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, G. Cao, D. Jiang, M. Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020.

X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018b.

Z. Wang, W. Hamza, and R. Florian. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150, 2017b.

G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and É. Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012, 2020.

F. Wilcoxon. *Individual comparisons by ranking methods*. Springer, 1992.

A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

G. Wisniewski, L. Zhou, N. Ballier, and F. Yvon. Biais de genre dans un système de traduction automatiqueneuronale: une étude préliminaire (gender bias in neural translation: a preliminary study). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1: conférence principale*, pages 11–25, 2021.

R. Xia, C. Zong, X. Hu, and E. Cambria. Feature ensemble plus sample selection: domain adaptation for sentiment classification. *IEEE Intelligent Systems*, 28(3): 10–18, 2013.

F. Xiong, J. Barker, Z. Yue, and H. Christensen. Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7424–7428. IEEE, 2020.

J. Xu, H. He, X. Sun, X. Ren, and S. Li. Cross-domain and semisupervised named entity recognition in chinese social media: A unified model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2142–2152, 2018.

W. Xu, C. Callison-Burch, and W. B. Dolan. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11, 2015.

B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. *ICLR*, 2015.

Y. Yang. Research and realization of internet public opinion analysis based on improved tf-idf algorithm. In *2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)*, pages 80–83. IEEE, 2017.

Y. Yang and J. Eisenstein. Unsupervised multi-domain adaptation with feature embeddings. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 672–682, 2015.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

W. Yin and H. Schütze. Attentive convolution: Equipping CNNs with RNN-style attention mechanisms. *Transactions of the Association for Computational Linguistics*, 6:687–702, 2018.

G. Zaragoza. *Le personnage de théâtre*. Armand Colin, 2006.

Z.-H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005.

G. Zhu and C. A. Iglesias. Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85, 2016.

X. Zhu and A. B. Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Annexes

A

Les réseaux de neurones profonds *Deep Neural Networks*

A.1 Préambule

Dans cette annexe, nous fournissons du contexte technique pour comprendre les principales architectures neuronales couramment utilisées dans le traitement du langage naturel, et plus particulièrement dans le contexte de ce manuscrit. Dans la Section A.2, nous examinons plus spécifiquement l'architecture de trois modèles qui jouent un rôle majeur dans le travail présenté dans ce manuscrit, à savoir : Word2Vec Mikolov et al. [2013], ELMo Peters et al. [2018b] et les modèles Transformer de type encodeur (BERT, CamemBERT, FlauBERT, etc.) Vaswani et al. [2017].

A.2 Les modèles de plongements lexicaux

A.2.1 *Embeddings from Language Models* (ELMO)

L'idée principale derrière la construction de plongements contextuels de mots est la création d'une représentation pour chaque mot, en fonction du contexte dans lequel il apparaît. Cette idée a été mise en œuvre à l'aide de réseaux de neurones récurrents (RNN) non supervisés, construits de façon similaire à Word2Vec. Plus précisément, les RNN sont construits pour prédire, à partir d'un mot dans une phrase, le mot suivant. Ces réseaux sont créés pour leur capacité à capturer et à maintenir des dépendances à long terme dans leurs états cachés. Afin de calculer la représentation du mot, il faut concaténer la couche cachée de la représentation Word2Vec avec la nouvelle représentation calculée. L'idée est qu'après avoir fourni le mot actuel, nous pouvons concaténer la représentation de la couche cachée avec une représentation statique afin de conserver l'information du contexte passé

et d'y intégrer le contexte actuel. Les représentations du modèle ELMo [Peters et al., 2018b] sont générées à partir d'un modèle Bi-LSTM pré-entraîné à deux couches pour produire le modèle de langue pré-entraîné, qui correspond à la brique contextuelle du modèle, qui construit des représentations de mots concaténées à des représentations statiques de caractères (CNN/RNN).

A.2.2 *Bidirectional Encoder Representations from Transformers* (BERT)

Les représentations bi-directionnelles de type encodeur issues de Transformer (BERT [Devlin et al., 2019]) constituent un autre modèle de langue neuronal, capable de générer des représentations de mots dépendantes du contexte. Néanmoins, contrairement à ELMo, qui est basé sur une représentation de caractères et deux couches récurrentes, les modèles Transformer comme BERT adoptent une approche différente qui consistent à un système de tokenisation comme WordPiece **?** ou SentencePiece Kudo and Richardson [2018] couplés à des couches d'attention Transformer.

Les systèmes de tokenisation sont basés sur un ensemble prédéfini de tokens appris sur le même corpus utilisé pour l'entraînement du réseau de neurones. Étant donné ce vocabulaire issu du domaine général, n'importe quel terme hors-vocabulaire est décomposé en tokens appartenant au vocabulaire, en minimisant le nombre de découpages possibles du mot. Par conséquent, contrairement à d'autres modèles de langue comme ELMo, les modèles BERT permettent d'entraîner une matrice de tokens grâce à WordPiece, de découper les termes hors-vocabulaire en tokens, et de représenter ces tokens en fonction du contexte.

BERT étant un modèle qui repose uniquement sur des couches basées sur l'attention, il est par défaut incapable d'encoder des informations sur la position d'un token dans une phrase comme le ferait une couche récurrente (i.e., pour une couche Transformer, une séquence dans un sens et la même séquence après avoir mélangé des mots produisent les mêmes représentations). Pour résoudre ce problème, les représentations statiques initiales de SentencePiece sont enrichies avec des incorporations de position et de segment (voir Figure **??**).

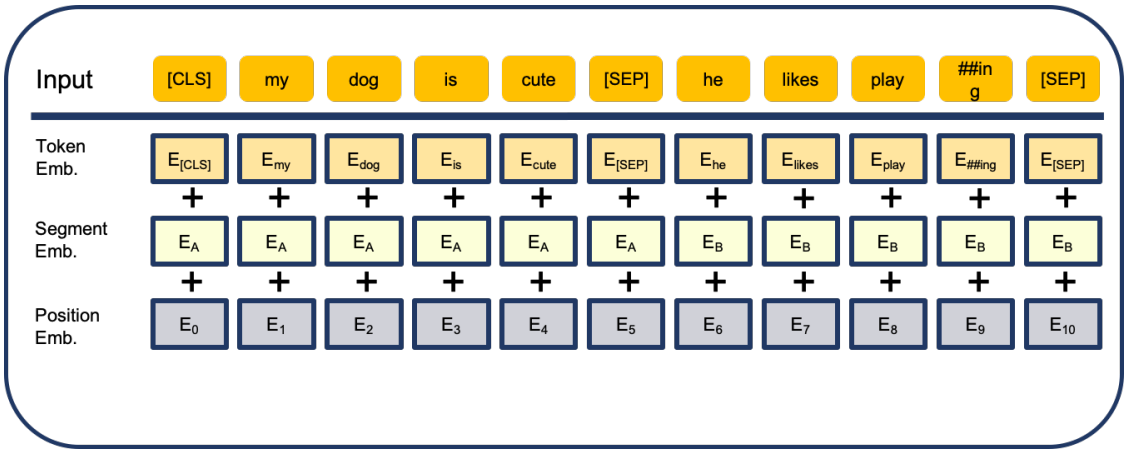


Figure A.1: Présentation de l'entrée de BERT. Les deux séquences d'entrées sont reliées en utilisant des tokens spéciaux [CLS] et [SEP], puis la représentation WordPiece de chaque token est construite et additionnée aux représentations de segment et de position. Source

Après la couche de représentation initiale, les représentations statiques de WordPiece traversent une série de couches Transformer, pour finalement aboutir à des représentations contextuelles de tokens. Cependant, contrairement à ELMo qui est utilisé dans une optique d'extraction de caractéristiques (i.e., en générant des représentations qui servent de caractéristiques fixes pour les couches en aval), BERT est utilisé comme un encodeur et est traditionnellement ajusté de bout en bout avec les têtes spécifiques à la tâche.

La dernière caractéristique spécifique à BERT est sa procédure d'entraînement. Alors que les modèles de langue plus traditionnels sont entraînés sur une tâche de prédiction des prochains mots d'une phrase, BERT tire parti de la nature bi-directionnelle du modèle qui utilise à la fois les contextes gauches et droits. En outre, ce modèle est entraîné sur une première tâche de *masking* de mots, visant à masquer 10% du corpus de préentraînement et à prédire correctement les mots masqués. Il est ensuite entraîné sur une deuxième tâche de prédiction d'une paire de phrases qui doivent être classées comme successives ou aléatoires. Cependant, cette tâche est largement considérée comme nuisible à la qualité globale des représentations générées par les modèles, et est ignorées dans des variantes ultérieures à BERT telles que RoBERTa [Liu et al., 2019b].

B

Erreurs orthographiques

Protocole d'annotation. Nous avons effectué un tirage aléatoire de 100 mots mal orthographiés dans les corpus, afin de comparer le traitement de ces erreurs par les modèles. Plus précisément, nous avons généré un tirage aléatoire de mots par batchs de 50, puis nous avons annoté les mots à la main pour distinguer les mots bien orthographiés des mots mal orthographiés. Nous avons annoté des batchs jusqu'à ce que nous obtenions 100 mots pour chaque corpus. Ensuite, nous avons associé, toujours à la main, les mots mal orthographiés à leur version correcte.

Corpus DEFT-Laws. Dans le corpus DEFT-Laws, nous avons effectué un tirage aléatoire sur les mots du domaine général contenant des erreurs orthographiques (voir Tableau B.1).

Corpus Bio-Gallica. Dans le corpus DEFT-Laws, nous avons effectué un tirage aléatoire sur les mots du domaine de spécialité (i.e., domaine de la biologie) contenant des erreurs orthographiques (voir Tableau B.2).

Corpus EDF-Emails. Dans le corpus EDF-Emails, nous avons effectué un tirage aléatoire sur les mots du domaine général, associés au registre des courriers électroniques et contenant des erreurs orthographiques (voir Tableau B.3).

Erreurs	Correct	Erreurs	Correct
condidat	candidat	privilege	privilège
xonditions	conditions	supéreurs	supérieurs
travails	travail	mofications	modifications
rspect	respect	acions	actions
investisssements	investissements	seines	saines
investissemets	investissements	evisagé	envisagé
producturs	producteurs	commisssion	commission
reponse	réponse	concertees	concertées
testosterone	testostérone	conformement	conformément
liberation	libération	limit	limite
annexei	annexe	annexeii	annexe
destinees	destinées	exportees	exportées
quantites	quantités	entreprosé	entreposé
etablit	établit	exprimee	exprimée
appliquees	appliquées	periodes	périodes
commencant	commençant	etablies	établies
elements	éléments	frontiere	frontière
dedouane	dédouané	releves	relevés
pésident	président	sanità	sanitaire
ministro	ministre	xpeuvent	peuvent
intérês	intérêt	objectis	objectif
aproprié	approprié	naure	nature
territioire	territoire	directrive	directive
cooperatieve	coopérative	communautare	communautaire
vieilesse	vieillesse	rénumération	rémunération
cordination	coordination	renvoyes	renvoyés
refuses	refusés	autorisee	autorisée
operation	opération	interet	intérêt
cetaces	cétacés	prets	prêts
delivrance	délivrance	accordees	accordées
entrepot	entrepôt	formalites	formalités
censee	censée	delivree	délivrée
presentee	présentée	enumeres	énumérés
etablir	établir	caractere	caractère
egard	égard	souspositions	sous-positions
categorie	catégorie	importee	importée
deuxieme	deuxième	competente	compétente
remplacee	remplacée	corresponje	corresponde
informee	informée	secretaire	secrétaire
executif	exécutif	debarquees	débarquées
capturees	capturées	precedent	précédent
numero	numéro	autorites	autorités
competentes	compétentes	proprietaire	propriétaires
affreteur	affréteur	recues	reçues
limitees	limitées	fixees	fixées
peche	pêche	infermiere	infirmière
dirigee	dirigée	depasser	dépasser
derniere	dernière	quantite	quantité
elevee	élevée	pechant	pêchant
decembre	décembre	especes	espèces

Table B.1: DEFT-Laws - Cent mots mal orthographiés et leurs équivalents corrects

Erreurs	**Correct**	**Erreurs**	**Correct**
hydrpphobe	hydrophobe	imporlance	importance
réfractairo	réfractaire	épilhélium	épithélium
pnèumonique	pneumonique	diphlhérie	diphtérie
sousrcutané	sous-cutané	alténuatjon	atténuation
obes	obèse	diphthéritiquhs	diphthéritiques
péiïtonéale	péritonéale	trachéotomise	trachéotomie
eongestion	congestion	caeutchpuc	caoutchouc
péiùtonéale	péritonéale	infiltratiou	infiltration
atëlectasique	atélectasique	supérieui'	supérieur
thsèe	thèse	lànceolatïts	lanceolatus
iujecté	injecté	injoction	injection
ouservations	observations	rotaiion	rotation
décoàgulation	décoagulation	prograiiime	programme
olynucléaires	polynucléaires	réatcion	réaction
synlhétique	synthétique	conlirmalion	confirmation
lémoin	témoin	iransparents	transparents
liypersensibiliser	hypersensibiliser	préparaion	préparation
propkiétés	propriétés	matleres	matières
iymphocytose	lymphocytose	septicemique	septicémique
hémorrhagiqu	hémorrhagique	desfinfection	désinfection
infractus	infarctus	ppinion	opinion
leuçqcytes	leucocytes	amibpïdes	amiboïdes
pjans	plans	migratipn	migration
leucopytes	leucocytes	cpjlules	cellules
dprme	derme	yaisseaux	vaisseaux
rpnferme	renferme	cellulps	cellules
bacifles	bacilles	zonp	zone
périphériqup	périphérique	inoffensiye	inoffensive
aotinobacilles	actinobacilles	inlermusculaire	intermusculaire
chimiôtaxisme	chimiotaxisme	vlrus	virus
yaccin	vaccin	baeilles	bacilles
vaccinogene	vaccinogène	ramoilissemenl	ramollissement
ampûule	ampoule	granulalations	granulations
folliculile	folliculite	éternuemenls	éternuements
chajeur	chaleur	éntérocoques	entérocoques
enkystemenl	enkystement	tracranienne	intracrânienne
bactériolytiquo	bactériolytique	vimmunité	immunité
auimaux	animaux	inloxicalion	intoxication
typhpïde	typhoïde	subslances	substances
antihematiques	antihématiques	jeucocytes	leucocytes
dégénéralives	dégénératives	interprét	interprété
streplococcus	streptococcus	sulfobàctéries	sulfobactéries
vacci	vaccin	àgglutinine	agglutine
anlienzyme	antienzyme	ântihématiques	antihématiques
urélrale	urétrale	aneslhésie	anesthésie
intoxicalion	intoxication	fintoxicalion	intoxication
symptomaliques	symptomatiques	anti-synrptomalique	anti-symptomatique
défervesçence	défervescence	hémotoxinc	hémotoxine
éfiologiques	étiologiques	stéréociiimie	stéréochimie
fhypoleucocytose	hypoleucocytose	byperleucocytose	hyperleucocytose

Table B.2: Bio-Gallica - Cent mots mal orthographiés et leurs équivalents corrects

Erreurs	Correct	Erreurs	Correct
ordialement	cordialament	coordialement	cordialament
dcordialement	cordialament	ccordialement	cordialament
cordialament	cordialament	cordialerment	cordialament
cordialemnt	cordialament	cordialemement	cordialament
cordiallement	cordialament	cordialment	cordialament
cordiallemnt	cordialament	cordialzment	cordialament
cordialelement	cordialament	cordialmement	cordialament
cordialmeent	cordialament	cordialemeent	cordialament
cordialemet	cordialament	coordialememt	cordialament
cordialemment	cordialament	cordialeme	cordialament
cordialement@	cordialament	cordialemennt	cordialament
cordiallment	cordialament	cordialemant	cordialament
cordialementt	cordialament	cordialelent	cordialament
cordialeent	cordialament	cordiallemnt	cordialament
cordialemznr	cordialament	cordialmement	cordialament
cordialemrnt	cordialament	cordialemen	cordialament
cordialmente	cordialament	cordialements	cordialament
reemboursement	remboursement	emboursement	remboursement
reboursement	remboursement	renboursement	remboursement
reemboursement	remboursement	rembourssement	remboursement
remboursment	remboursement	remboursrment	remboursement
remboursemment	remboursement	rembourszer	remboursement
remboursemet	remboursement	remboursemen	remboursement
bonjou	bonjour	bonjoiur	bonjour
bonjours	bonjour	bonjoir	bonjour
bonjoure	bonjour	bonjior	bonjour
bonjourje	bonjour	bonjout	bonjour
bonjiour	bonjour	lbonjour	bonjour
bonjiur	bonjour	bonjorur	bonjour
bonjor	bonjour	bonjoutr	bonjour
bonjr	bonjour	bonjpur	bonjour
bonjourr	bonjour	bonjourmme	bonjour
bonjjour	bonjour	bonjon	bonjour
ebonjour	bonjour	bonjeur	bonjour
bonjoursi	bonjour	bonj	bonjour
bonjean	bonjour	bonjourn	bonjour
bonjo	bonjour	bbonjour	bonjour
jbonjour	bonjour	nbonjour	bonjour
bonjouf	bonjour	bonjouir	bonjour
bonjous	bonjour	bonjpour	bonjour
salutions	salutations	salutaions	salutations
salutationd	salutations	salutaztions	salutations
saluations	salutations	salutatios	salutations
ssalutations	salutations	salutatiions	salutations
salutaitons	salutations	saluatations	salutations
salutatins	salutations	salutaiotn	salutations
salutetions	salutations	salution	salutations
salutationq	salutations	salutatione	salutations
salutacions	salutations	alutations	salutations
remerciemnts	remerciements	remercients	remerciements

Table B.3: EDF-Emails - Cent mots mal orthographiés et leurs équivalents corrects

C

Configuration des serveurs

C.1 Préambule

Dans ce chapitre, nous fournissons des détails sur les deux serveurs utilisés pour nos expériences. Pour chaque expérience, nous nous sommes assurés d'utiliser la même configuration, afin d'avoir des résultats comparables. En effet, en fonction des cartes graphiques utilisées pour une même expérience, des résultats différents sont obtenus. Les serveurs sont appelés Tesla et Quadro dans le manuscrit.

C.2 Présentation des serveurs

Serveur	**Référence CPU**	**RAM**	**Référence GPU**	**#GPU**
Quadro	Intel Xeon CPU E5-1650 v4	125.673 GiB	Quadro P6000	2
Tesla	Intel Xeon Gold 5120	32 GiB	NVIDIA Tesla V100	2

Table C.1: Résumé des informations sur les serveurs