N° d'ordre NNT : 2018LYSE2009

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512 Informatique et Mathématiques

Discipline : Informatique

Soutenue publiquement le 7 février 2018, par :

Edmundo-Pavel SORIANO-MORALES

Hypergraphes et fusion d'information pour l'enrichissement de la représentation de termes.

Applications à la reconnaissance d'entités nommées et à la désambiguïsation du sens des mots

Devant le jury composé de :

Sophie ROSSET, Directrice de Recherches, Université Paris 11, Présidente

Marc EL BEZE, Professeur des universités, Université d'Avignon, Rapporteur

Mathieu ROCHE, Chercheur HDR, CIRAD-La recherche agronomique pour le développement, Rapporteur

Farah ZITOUNE BENAMARA, Maîtresse de conférences, Université Toulouse 3

Sabine LOUDCHER, Professeure des universités, Université Lumière Lyon 2, Directrice de thèse

Julien AH-PINE, Maître de Conférences, Université Lumière Lyon 2, Co-Directeur de thèse

Contrat de diffusion

Hypergraphes et fusion d'information pour l'enrichissement de la représentation de termes. Applications à la reconnaissance d'entités nommées et à la désambiguïsation des sens des mots

1 Introduction

1.1 Contexte

Appréhender la sémantique portée par des documents textes joue un rôle essentiel dans l'évolution de l'intelligence artificielle. Compte tenu de la génération croissante de données textuelles, il existe, en effet, un besoin fort de systèmes capables d'extraire des informations pertinentes et porteur de sens à partir de grandes quantités de documents. Ces extractions aident à l'indexation, l'interprétation, l'exploitation, bref à l'analyse de l'information contenue dans des textes, et trouvent ainsi de nombreuses applications dans nos activités quotidiennes. De façon plus générale, rendre les ordinateurs capables de reproduire avec efficacité les aptitudes cognitives propres aux humains est l'objectif de la recherche en intelligence artificielle [Sugiyama 2016]. Issu de ce domaine multidisciplinaire, le traitement automatique du langage naturel (TALN) ou encore la linguistique computationnelle est la branche qui à rendre capables des machines de comprendre et de générer notre langage [Jurafsky 2009].

Les solutions aux tâches de TALN suivent généralement trois étapes pour être mises en oeuvre [Jurafsky 2009, Aggarwal 2012]. Premièrement, en pré/traitement, un corpus d'entrée est "normalisé" de sorte qu'il soit plus facile à traiter par la machine dans les étapes suivantes. Deuxièmement, dans la représentation des unités linguistiques (mots, groupes de mots, phrases, documents,...), de nombreuses caractéristiques pertinentes sont extraites du texte pré-traité. Troisièmement, une technique

d'apprentissage automatique est utilisée pour apprendre un modèle capable de résoudre de façon efficace la tâche sur les données d'apprentissage mais surtout sur de nouvelles données. La sortie de ce système est le modèle instancié qui révèle une certaine connaissance du langage permettant de traiter efficacement la tâche donnée.

1.2 Problématiques et contributions

Il y a plusieurs défis de recherche qui découlent des choix effectués dans chacune des étapes composant le flux d'un système TALN, décrit ci-dessus. Dans cette thèse, nous nous concentrons sur trois défis qui se posent à la fois dans les phases de représentation des caractéristiques et de découverte des connaissances. Ces défis sont : (1) la modélisation, l'extraction et le stockage de différents types d'éléments linguistiques à partir de textes bruts, (2) le traitement de la rareté inhérente aux données textuelles et leur combinaison pour obtenir de meilleures représentations ; (3) profiter des relations entre les mots et ensuite les exploiter afin de découvrir leur parenté latente et être capable de résoudre des tâches du TALN.

Afin de répondre à ces défis, nous proposons trois contributions :

— un modèle de réseau basé sur des hypergraphes pour contenir des données linguistiques hétérogènes.
— des méthodes de fusion permettant de combiner des représentations hétérogènes mais complémentaires, tout en atténuant le problème des données creuses.
— un algorithme basé sur le réseau du modèle proposé pour découvrir la relation sémantique entre mots liés.

Ces contributions sont évaluées et validées en utilisant deux taches sémantiques du TALN : la désambiguïsation lexicale (Word Sense Induction and Disambiguation ou WSDI/WSD) et la reconnaissance d'entités nommées (Named Entity Recognition ou NER). Nous choisissons ces tâches étant donné qu'elles sont centrales dans les systèmes de TALN les plus avancés.

2 Généralités

2.1 Hypothèse Distributionnelle

Le travail que nous présentons dans cette thèse repose principalement sur l'hypothèse distributionnelle (HD)[Harris 1954]. C'est aussi le cas pour la majorité des approches sémantiques actuelles dans le domaine du TALN. L'hypothèse est simple mais puissante : les mots qui apparaissent dans les mêmes contextes linguistiques partagent des significations similaires. Par conséquent, la signification d'un mot peut être déterminée par l'ensemble des contextes dans lesquels ce mot participe. Dans ce travail nous nous focalisons exclusivement sur deux contextes : la co-occurrence lexicale et la co-occurrence syntaxique des contextes. Nous définissons l'occurrence lexicale comme celle basée sur les mots qui co-occurrent avec un autre mot dans un voisinage prédéfini. Les co-occurrences syntaxiques sont basées sur une analyse (ou parsing) plus profonde du texte afin d'obtenir des relations (et le sens) existantes entre les mots qui co-occurrent.

2.2 Modèles de représentation

Le modèle d'espace vectoriel consiste à représenter des unités textuelles dans un espace multidimensionnel. Les unités textuelles ne sont pas forcément les mots. Nous pouvons décrire des caractéristiques co-occurrentes pour les documents, les phrases, les paragraphes ou d'autres types d'unités [Manning 1999]. Une matrice est utilisée comme la structure qui contient chaque objet et ses caractéristiques contextuelles. Les métriques de distance (ou mesures de similarité) sont utilisées sur les vecteurs de la matrice pour déterminer un niveau de dissimilarité ou de similarité entre ces objets.

D'autres types de modèles de représentation, basés sur des graphes, sont couramment utilisés dans la littérature. En effet, les modèles basés sur les réseaux ont été étudiés au cours des dernières années dans le champ du TALN [Mihalcea 2011]. Alors que nous pouvons représenter un graphe comme une matrice, et donc comme un modèle dans un espace vectoriel, les graphes sont un formalisme de représentation utile qui peuvent être appliqués à un large ensemble de caractéristiques linguistiques comme la relation entre les mots dans un texte ou entre les caractéristiques qui les décrivent. En effet, le langage étant un système complexe et dynamique, les réseaux fournissent un modèle adéquat pour représenter et étudier la structure et l'évolution des systèmes linguistiques [Choudhury 2009]. Dans cette thèse, nous basons notre

proposition de modèle linguistique sur un hypergraphe dont les hyperarêtes peuvent être de différents types.

2.3 Données creuses

Peu importe leur type, les réseaux d'information sont généralement transformés en matrices avant d'être traités par un calcul. Par conséquent, comme nous essayons de modéliser le langage, les matrices associées aux graphes ou aux réseaux d'informations sont souvent des matrices creuses. En effet, le manque de données ainsi que leur grande dimensionnalité sont des problèmes qui affectent les performances des approches de découverte de connaissances [Aggarwal 2012, Périnet 2015] appliquées aux données textuelles.

Une matrice de données éparse (ou creuse) a la plupart de ses entrées égales à zéro. Ainsi, dans notre contexte, la majorité des mots (lignes) dans le corpus sont décrits par très peu de contextes (colonnes) mais qui sont tout à la fois en grand nombre. Ceci est un problème important car pendant la phase de découverte de connaissances de tout système TALN, nous visons à former un modèle d'apprentissage qui finira par prédire, classifier, grouper les mots d'une manière ou d'une autre. Si les mots sont représentés par un nombre limité de contextes, les algorithmes d'apprentissage ne pourront pas généraliser correctement.

Une représentation textuelle explicite et distributionnelle est donc éparse. Il y a beaucoup de mots dans un corpus de documents et même si certains mots apparaissent dans plusieurs textes, tous les mots du corpus ne peuvent pas être présents dans un même texte. Cela devient un problème important avec les systèmes de TALN : les mots ne sont décrits que par un petit nombre de fonctionnalités. Dans ce qui suit, nous décrivons nos deux premières propositions qui traitent de l'utilisation d'informations hétérogènes pour représenter un terme et atténuer le manque de données qui accompagne ces types de représentations textuelles.

3 Modèle linguistique basé sur des hypergraphes et enrichi par la fusion

Les deux premières contributions de cette thèse sont contenues dans la proposition d'un modèle linguistique basé sur des hypergraphes et enrichi avec des techniques de

fusion. Ainsi, le modèle traite les deux premières questions de recherche exposées précédemment : l'exploitation de différents types de caractéristiques en utilisant un espace de représentation unique et qui, en même temps, permet de réduire l'eparsité des données caractéristique aux données textuelles.

Le modèle que nous présentons ici comporte trois caractéristiques importantes : premièrement, la possibilité de tirer profit des différents types d'informations textuelles par le biais de techniques de fusion visant à exploiter la complémentarité entre caractéristiques linguistiques distinctes. Deuxièmement, les relations de similarité sémantique se trouvant renforcées, nous pouvons exploiter la structure interne enrichie du réseau de similarité entre mots par le biais d'outils d'analyse de graphe. Troisièmement, notre modèle par l'utilisation de méthode de fusion, permet également de tenir compte du problème de l'éparsité de données en dépit d'une dimensionalité grandissante causée par l'hétérogénéité des descripteurs.

En tant que caractéristiques de co-occurrence de mots, nous sélectionnons à la fois des contextes lexicaux et syntaxiques, créant ainsi une ressource linguistique qui contient les deux types d'informations afin d'obtenir des vues complémentaires des relations entre les mots.

En effet, dans la littérature, nous constatons que peu d'approches traitent les attributs syntaxiques. Nous pensons que la recherche de similitudes sémantiques peut être améliorée en ajoutant des informations syntaxiques non seulement lors de l'utilisation de relations de dépendance, mais aussi en exploitant l'arbre de constituants de chaque mot. En outre, l'utilisation de données syntaxiques avec des co-occurrences sémantiques et/ou lexicales nous conduit à avoir un réseau d'informations hétérogènes enrichi, ce qui n'est pas le cas dans la plupart des approches existantes.

En prenant en compte les opportunités de recherche décrites ci-dessus, nous proposons une modélisation d'un réseau linguistique basée sur un hypergraphe hétérogène que nous enrichissons par l'utilsation de méthodes de fusion qui peuvent être vues telles des techniques d'enrichissement de liens sommets-sommets et/ou sommets-hyperarêtes.

Comme indiqué précédemment, notre proposition consiste en deux parties. La première, un modèle à base d'hypergraphes qui contient différents types de relations linguistiques extraites d'un corpus. Et la seconde, la combinaison de caractéristiques linguistiques afin de générer une représentation moins creuse et enrichie, à travers des techniques de fusion.

Notre modèle est basé sur l'utilisation d'un hypergraphe. La différence la plus importante avec les graphes classiques est de pouvoir relier plus de deux sommets en même temps, ce qui permet une meilleure caractérisation des interactions au sein d'un ensemble d'éléments individuels (dans notre cas, les mots) [Heintz 2014]. En effet, notre modélisation à base d'hypergraphes intègre initialement quatre types de relations entre les mots : la co-occurrence de phrase, les étiquettes morpho-syntaxiques, la structure syntaxique, et les relations de dépendance dans une structure linguistique unique. Ces relations ont été choisies car il est relativement facile de les obtenir pour les langues où il y a beaucoup de ressources linguistiques. Ainsi, ces caractéristiques peuvent être considérées comme des blocs de construction pour les modèles de TALN. Dans tous les cas, notre but est d'arriver à des annotations plus complexes (par exemple, des entités nommées) à partir des caractéristiques et des relations pertinentes.

Nous avons décidé de garder le contexte lexical au niveau de la phrase, de sorte qu'il puisse compléter l'information sémantique fournie par le contexte basé sur les relations de dépendance ainsi que le contexte syntaxique de la phrase. En bref, nous visons à couvrir trois niveaux de relation sémantique : un niveau proche avec les fonctions de dépendance, un niveau moyen avec une appartenance à une phrase nominale (avec la structure syntaxique), et un niveau plus long avec la coexistence lexicale de phrase, c-à-d, le voisinage des mots à niveau de toute la phrase. L'intuition est que lors de la résolution de tâches de TALN, avoir un accès direct à ces trois espaces sémantiques aidera à déterminer une relation de signification plus appropriée entre les mots.

La deuxième partie de notre méthode traite de la fusion des caractéristiques textuelles. À savoir, nous combinons les caractéristiques qui décrivent les termes en un seul espace de représentation unique. Ce nouvel espace vise à répondre à deux problèmes qui se posent lors de l'utilisation de données textuelles : comment utiliser efficacement des informations provenant de différents niveaux linguistiques (par exemple, lexical, syntaxique, sémantique) tout en allégeant l'eparsité typique dans les représentations textuelles.

Dans la littérature sur la fusion multimodale, nous pouvons discerner deux types principaux de techniques : la fusion précoce et la fusion tardive. Un troisième et un quatrième type de fusion, fusion croisée multimédias et fusion hybride sont également utilisés dans des tâches d'analyse multimédia. Nous utilisons ces techniques pour

enrichir le modèle.

Ces quatre types de fusion abordent naturellement la question du traitement de données hétérogènes car ils mélangent tous d'une manière ou d'une autre les colonnes de caractéristiques de chacune des deux représentations (dans le contexte d'une matrice de représentation). En ce qui concerne la réduction du taux des données creuses, l'intuition est qu'en combinant les matrices soit en les sommant, soit en les multipliant par éléments, la matrice résultante aura une structure plus dense. Par exemple, en additionnant deux matrices ayant la même forme, telles que deux matrices de similarité terme à terme, nous obtenons une matrice résultante qui contient les similarités des deux espaces caractéristiques. Dans le même esprit, en multipliant ces deux matrices, nous les combinons tout en obtenant une matrice de sortie plus dense. Néanmoins, le résultat de la somme et de la multiplication dépend évidemment de la nature des matrices utilisées.

Nous appliquons ces opérateurs de fusion à notre modèle basé sur les hypergraphes et nous obtenons un modèle unique qui réponde aux questions posées initialement.

Afin de matérialiser le modèle linguistique proposé, nous avons mis en place une procédure qui prend en compte un corpus et produit la ressource linguistique que nous venons d'introduire dans les paragraphes précédents. Nous avons basé notre processus sur l'encyclopédie en ligne Wikipédia [1] qui a été utilisée comme une source de données importante ainsi qu'un corpus de fond commun pour résoudre divers tâches du TALN et du text mining.

4 Applications à la reconnaissance d'entités nommées et à la désambiguïsation des sens des mots

Dans cette section, nous avons décidé de résoudre deux tâches de traitement du langage naturel en utilisant comme corpus, des données sous la forme de notre modèle . Nous abordons les tâches de reconnaissance d'entités nommées et d'induction et de désambiguïsation des sens des mots (WSI / WSD). Nous utilisons un réseau basé sur des hypergraphes enrichis, réseau construit sur des corpus de référence, pour valider l'utilité de nos propositions.

1. https://fr.wikipedia.org

4.1 Première application : reconnaissance d'entités nommées

Le but de la reconnaissance d'entités nommées est de découvrir automatiquement, dans un texte, des mentions appartenant à une catégorie sémantique bien définie. La tâche classique consiste à détecter, dans un texte, des entités de type Lieu (LOC), Organisation (ORG), Personne (PER), Divers (MISC) ou, si le terme n'est pas une entité nommée, lui attribuer une étiquette appropriée (en l'occurrence, l'étiquette O). La tâche est d'une grande importance pour les systèmes de TALN plus complexes comme par exemple, l'extraction de relations ou la fouille d'opinion [Nadeau 2007]. Généralement, la solution pour la reconnaissance d'entités nommées consiste à entraîner un algorithme d'apprentissage automatique supervisé avec de grandes quantités de texte annoté [Aggarwal 2012]. Comme pour d'autres tâches de TALN, la reconnaissance d'entités nommées nécessite des représentations (ou *features*) pour les mots afin de déterminer leur rôle dans une phrase. Nous proposons de construire ces représentations en nous basant sur notre modèle à base d'hypergraphes enrichi par fusion. Pour les expériences, nous avons choisi un algorithme d'apprentissage de type perceptron structuré en raison de ses performances et de son temps de d'entraînement réduit.

Nous expérimentons les quatre niveaux de fusion sur trois ensembles de données différents. Les matrices de représentation pour la reconnaissance d'entités nommées proviennent des caractéristiques contextuelles lexicales, des caractéristiques des dépendances syntaxiques ou d'autres caractéristiques dites standard pour la tâche en question. Notre objectif principal est de comparer l'efficacité des techniques de fusion primaires appliquées à la reconnaissance d'entités nommées. Aussi, nous déterminons empiriquement un opérateur de combinaison de fusion capable de tirer parti de la complémentarité des caractéristiques utilisées. Nous découvrons qu'en utilisant une combinaison de plusieurs opérations de fusion, une approche dite hybride, nous améliorons l'utilisation des caractéristiques individuellement et aussi sur l'utilisation de l'opérateur de fusion précoce trivial, qui consiste à la simple concaténation des caractéristiques dans un espace unique. Ceci indique que la matrice de caractéristiques unique, enrichie avec d'autres caractéristiques combinées, est suffisante pour améliorer les résultats desdites *baselines*. En général, dans nos expériences pour la reconnaissance d'entités nommées, nous voyons que les fonctionnalités enrichies ajoutées ne sont pas les plus importantes pour l'algorithme d'apprentissage lors de la prise d'une décision ; néanmoins, elles fournissent des informations supplémentaires nécessaires pour pousser le modèle vers la prédiction correcte, en enrichissant les caractéristiques

à travers la fusion croisée et tardive, en fournissant plus de descripteurs pour chaque mot et par conséquent en réduisant ainsi le niveau des données creuses dans les matrices de représentation.

Une fois que nous avons trouvé un ensemble d'opérations de fusion qui fonctionnent raisonnablement bien pour la reconnaissance d'entités nommées, nous expérimentons avec une autre tâche du TALN, la désambiguïsation lexicale, afin de confirmer l'intérêt d'utiliser des représentations enrichies par fusion.

4.2 Deuxième application : induction et désambiguïsation de mots

4.2.1 Représentations enrichies par fusion

Après avoir appris la meilleure configuration de fusion à partir de la tâche de reconnaissance d'entités nommées, nous avons testé dans les expériences qui suivent si les améliorations obtenues peuvent être transférées à une autre tâche de TALN, à savoir, l'induction et la désambiguïsation des sens des mots.

L'induction (Word Sense Induction, WSI) et la désambiguïsation (Word Sense Disambiguation, WSD) des sens des mots impliquent deux tâches étroitement liées. Le WSI vise à découvrir automatiquement l'ensemble des sens possibles d'un mot cible donné relativement à un corpus de textes contenant plusieurs occurrences dudit mot cible. D'un autre côté, le WSD prend un ensemble de sens possibles pour un mot et cherche à déterminer le sens le plus approprié pour chacune des instances du mot cible en fonction de son contexte. Le WSI est généralement traité comme une tâche d'apprentissage non supervisée, c-à-d qu'une méthode de *clustering* est appliquée aux mots apparaissant dans les contextes des instances d'un mot cible. Puis, les groupes trouvés sont interprétés comme les differents sens du mot cible. La tâche WSD est généralement résolue avec des approches basées sur des graphes de connaissances, ou plus récemment, avec des modèles supervisés qui nécessitent des données annotées. Il peut également être résolu raisonnablement bien en comparant les mots entourant chaque mot cible et les mots appartenant aux sens induits trouvés lors de l'étape WSI, comme nous le faisons dans cette section.

Nous utilisons un *clustering* spectral sur les matrices d'entrée afin de découvrir automatiquement les sens. En ce qui concerne la désambiguïsation de ces derniers, nous affectons trivialement les sens aux instances de mots cibles en fonction du nombre de mots communs dans chaque groupe et les mots du contexte du mot cible. En d'autres

termes, pour chaque instance de test d'un mot cible, nous sélectionnons le groupe (sens) avec le nombre maximum de mots partagés avec le contexte de l'instance actuelle.

De plus, afin de déterminer les systèmes les plus performants, éloignés des *baseline* triviales, nous proposons une mesure pour identifier les solutions les plus performantes. Selon cette métrique, nous trouvons que la combinaison des opérateurs de fusion est, à nouveau, la combinaison de caractéristiques la plus performante. En effet, en transférant des similarités lexicales de qualité dans la même matrice de caractéristiques, nous obtenons des relations plus utiles qu'en utilisant des données syntaxiques. Alors que les mêmes opérateurs qui ont surpassé les autre approches dans le cas de la reconnaissance d'entités nommées ne sont pas aussi adéquats dans cette expérience WSI/WSD, nous voyons que la plupart des techniques de combinaison améliorent les *baseline* à base de caractéristiques uniques et des opérations de fusion précoce.

4.2.2 Tirer parti de la structure du réseau linguistique

Nous tirons parti des relations qui existent au sein du réseau pour identifier les mots qui, avec leur voisinage, représentent un sens. Ainsi, nous proposons un algorithme basé sur le réseau pour résoudre l'induction des sens des mots. Notre méthode est inspirée des approches précédentes de [Véronis 2004] et de [Klapaftis 2007]. Dans Hyperlex, méthode basée sur les graphes [Véronis 2004], l'intuition principale est que les réseaux de co-occurrences ont des propriétés dites de "petit monde" et il est donc possible de détecter et d'isoler les nœuds importants connectés, appelés *hubs*. L'idée est que ces *hubs*, et leurs nœuds connectés, représentent eux-mêmes des sens. Nous utilisons également des espaces de représentation produits par des opérateurs de fusion.

En utilisant notre structure linguistique basée sur les hypergraphes, nous avons obtenu des résultats qui surpassaient ceux des méthodes similaires tout en étant plus flexibles en termes d'utilisation des paramètres. Encore une fois, nous avons constaté que l'utilisation d'opérateurs de fusion donne de meilleurs résultats par rapport à l'utilisation de caractéristiques uniques ou fusion précoce. Cependant, alors que nos systèmes de fusion donnent de meilleurs résultats que les *baseline*, les systèmes qui fonctionnent le mieux n'emploient pas de données hétérogènes. En effet, les meilleurs systèmes qui combinent les deux types de fonctionnalités possibles donnent de moins

bons résultats que les meilleurs espaces de fusion basée sur une information homogène.

5 Conclusion et travaux futurs

5.1 Conclusion

Les réseaux linguistiques sont des structures utiles pour comprendre la nature de notre langue. Dans la littérature, ils sont généralement utilisés pour comprendre la dynamique des mots et d'autres unités textuelles dans le langage, et pour résoudre des tâches pratiques de TALN. Néanmoins, quel que soit l'objectif, ils sont généralement basés sur l'hypothèse distributionnelle, c'est-à-dire que les mots seront trouvés dans des contextes similaires s'ils ont tendance à être liés sémantiquement. D'un autre côté, les représentations de données textuelles, décrites par des contextes dans un cadre distributionnel, sont rares par nature : la grande majorité des entrées dans une matrice de co-occurrence sont nulles. Pour traiter ces problématiques, nous avons proposé trois contributions dans cette thèse. Le premier et le second impliquent un réseau linguistique enrichi par des techniques de fusion, qui aboutit à des représentations de texte plus denses en combinant des espaces de caractéristiques hétérogènes. La seconde est une méthode basée sur la structure d'un graphe pour trouver des groupes de mots apparentés.

En ce qui concerne les techniques de fusion, nous les avons testées sur deux tâches : WSI/WSD et la reconnaissance d'entités nommées. En particulier, dans cette dernière, nous avons créé de nouvelles matrices de représentation qui ont montré une amélioration globale des performances. Afin d'obtenir ces améliorations, nous avons effectué un niveau élevé d'agrégation de fusion en combinant différents types de fusion. Concernant notre modèle basé sur la structure du réseau linguistique, nous l'avons testé sur la tâche WSI/WSD. En supposant l'hypothèse *scale-free*, nous avons trouvé des communautés de mots décrivant des sens en utilisant des contextes lexicaux au niveau des phrases avec des fréquences brutes pour pondérer les co-occurrences. Les opérateurs de fusion ont produit des espaces de représentation améliorés par l'utilisation de caractéristiques uniques, comme dans les expériences NER.

Enfin, l'hypergraphe proposé, à travers ses représentations de fusion, génère de grandes matrices qui doivent être correctement manipulées en machine afin d'améliorer le traitement du point de vue computationnel. Pour relever ce défi, nous utilisons

des approches telles que la parallélisation et les méthodes de calcul dites out-of-core [2].

5.2 Travaux futurs

En ce qui concerne les techniques de fusion, nous avons montré qu'il existait des approches hybrides permettant d'améliorer les résultats mais il serait nécessaire d'analyser plus finement quels types d'opérateurs de fusion et quels types de combinaisons hybrides seraient les plus efficaces afin de limiter l'espace des possibilités. Par ailleurs, la comparaison de ces méthodes avec d'autres approches de réduction de dimension bien établies serait intéressantes pour comprendre les compromis entre la réduction des performances et la réduction des dimensions, tout en se concentrant sur les corpus moins importants. En effet, les techniques récentes de représentations distributionnelles, ou de plongements de mots, ont un inconvénient majeur : elles ne fonctionnent pas très bien sur les corpus de petites tailles. Les méthodes de fusion de caractéristiques comme celle que nous avons utilisées pourraient apporter des réponses intéressantes dans ce cas.

D'une autre part, en ce qui concerne l'algorithme basé sur le réseau pour WSI/WSD, une analyse plus approfondie des erreurs permettrait un plus grand aperçu du comportement des noms et des verbes en fonction du contexte. Comprendre quelle est la différence syntaxique ou lexicale entre les contextes, qui induisent les bonnes ou mauvaises performances de chaque type de fonctionnalité, pourrait rendre le système plus flexible à d'autres domaines et d'autres tâches de TALN. En outre, l'hypergraphe pourrait être mieux exploité en utilisant des méthodes hypergraphes, principalement par l'analyse spectrale.

Références

[Aggarwal 2012] Charu C. Aggarwal and ChengXiang Zhai, editors. Mining text data. Springer, 2012. (Cited on pages 1, 4 and 8.)

[Choudhury 2009] Monojit Choudhury and Animesh Mukherjee. *The Structure and Dynamics of Linguistic Networks*. In Niloy Ganguly, Andreas Deutsch and Animesh Mukherjee, editors, Dynamics On and Of Complex Networks, Modeling

2. En bref, il s'agit d'algorithmes qui stockent dans la mémoire vive que les parties d'une matrice requises pour le calcul et qui gardent le reste sur le disque dur

and Simulation in Science, Engineering and Technology, pages 145–166. Birkhäuser Boston, 2009. (Cited on page 3.)

[Harris 1954] Zellig Harris. *Distributional structure*. vol. 10, no. 23, pages 146–162, 1954. (Cited on page 2.)

[Heintz 2014] Benjamin Heintz and Abhishek Chandra. *Beyond graphs : toward scalable hypergraph analysis systems*. vol. 41, no. 4, pages 94–97, 2014. (Cited on page 5.)

[Jurafsky 2009] Dan Jurafsky and James H. Martin. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition, 2nd edition. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, 2nd édition, 2009. (Cited on page 1.)

[Klapaftis 2007] Ioannis P. Klapaftis and Suresh Manandhar. *UOY : A Hypergraph Model for Word Sense Induction & Disambiguation*. In Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval '07, pages 414–417, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. (Cited on page 10.)

[Manning 1999] Christopher D Manning, Hinrich Schütze*et al.* Foundations of statistical natural language processing, volume 999. MIT Press, 1999. (Cited on page 3.)

[Mihalcea 2011] Rada F. Mihalcea and Dragomir R. Radev. Graph-based natural language processing and information retrieval. Cambridge University Press, 1st édition, 2011. (Cited on page 3.)

[Nadeau 2007] David Nadeau and Satoshi Sekine. *A survey of named entity recognition and classification*. Lingvisticae Investigationes, vol. 30, no. 1, pages 3–26, 2007. (Cited on page 8.)

[Périnet 2015] Amandine Périnet and Thierry Hamon. *Analyse distributionnelle appliquée aux textes de spécialité - Réduction de la dispersion des données par abstraction des contextes*. TAL, vol. 56, no. 2, 2015. (Cited on page 4.)

[Sugiyama 2016] Masashi Sugiyama. Introduction to statistical machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2016. (Cited on page 1.)

[Véronis 2004] Jean Véronis. *HyperLex : lexical cartography for information retrieval*. vol. 18, no. 3, pages 223 – 252, 2004. (Cited on page 10.)