Rapport de stage – *Juillet 2014*
Master 2 en Apprentissage et Traitement Automatique de la Langue

Acquisition de relations sémantiques dans Wikipédia à l'aide de patrons lexicaux

Agathe Mollé

Encadrant
Philippe Langlais

Tuteur
Solen Quiniou

Table des matières

Introduction

Contexte

Ce stage s'est déroulé à l'Université de Montréal, Québec, Canada. Plus particulièrement, il a eu lieu au sein du laboratoire RALI (Recherche Appliquée en Linguistique Informatique), encadré par le professeur Philippe Langlais. Ce stage aura duré 5 mois.

Remerciements

Je tiens à remercier l'équipe du RALI, qui m'a accueillie avec la réputée bienveillance du Québec. En particulier Philippe Langlais, qui malgré un emploi du temps chargé, a su trouver du temps pour m'encadrer et m'orienter, et ce toujours chaleureusement.

Je remercie également l'équipe du master ATAL ainsi que ses étudiants. Je suis en effet très fière d'avoir fait partie de la première promotion d'ATAL, et d'avoir par cette occasion découvert un domaine très intéressant : le traitement du langage.

Pour finir, merci à tous les québécois ou expatriés que j'ai pu rencontrer ici. Merci à mes colocataires et toutes les personnes qui m'ont permis de découvrir ce merveilleux pays.

Introduction à la problématique

Aujourd'hui, le domaine du Traitement Automatique du Langage (TAL), à savoir la manipulation de données textuelles à l'aide d'outils informatiques, occupe une part cruciale des enjeux de recherche. Nous faisons face à un besoin croissant de traiter à grande échelle, et de manière automatisée, de grandes quantités de textes. Nous vivons dans un monde avec un contenu textuel dynamique et changeant, et son traitement est facilité par le foisonnement de nombreuses sources. Afin de pleinement appréhender les textes, bon nombre de systèmes actuels et futurs de TAL nécessitent une connaissance externe au texte. En effet, un humain ne se contente pas de lire une suite de caractères afin de comprendre le sens d'une phrase, il puise également dans des connaissances qu'il a acquis au fil du temps. Pour pouvoir analyser correctement des données brutes, la machine a également besoin de posséder une base de connaissances similaires. A terme, les systèmes seront ainsi capables de lire, comprendre et raisonner.

Les bases de connaissances actuelles sont pour la plupart structurées sous forme de relations, par exemple la relation *réalisé-par* qui possède des instances telles que *réalisé-par(Kill Bill, Quentin Tarantino)* ou *réalisé-par(Le dictateur, Charles Chaplin)*. On peut d'ailleurs distinguer les relations purement sémantiques (par exemple la synonymie, l'antonymie ou la méronymie aussi appelée *partie/tout*) des autres plus spécifiques. L'enjeu est le suivant : comment alimenter ces bases de connaissances afin qu'elles soient le plus complet possible ? Ce travail titanesque ne peut être réalisé entièrement manuellement par des spécialistes du domaine. Cela serait bien trop coûteux en temps. C'est pourquoi des méthodes ont été mises en œuvre afin d'extraire de telles relations automatiquement. Certaines d'entre elles exploitent des corpus et le contexte des termes, à l'aide d'une analyse distributionnelle (Bourigault et al., 2004; Cimiano et al., 2003) ou bien de patrons lexico-syntaxiques (Hearst, 1992; Morin, 1999). D'autres exploitent des ressources existantes telles que des dictionnaires électroniques (Rigau, 1998) ou bien des bases de données lexicales (Harabagiu and Moldovan, 1998). D'autres enfin s'appuient sur la structure interne des termes (Daille, 2003).

Nous nous intéressons aux méthodes d'extraction à base de corpus, et plus particulièrement à base d'acquisition de patrons récurrents dans les contextes proches des termes. Alors que la plupart des travaux cherchent à identifier des

patrons lexico-syntaxiques, nous n'observons les données qu'au niveau lexical. Nous travaillons sur les textes issus de Wikipédia [1] français et nous intéressons aux relations sémantiques. Nous verrons par la suite que notre approche diffère des approches existantes car nous amorçons notre processus avec des données issues d'un jeu sérieux : JeuxDeMots [2] (Lafourcade and Joubert, 2008).

Un premier chapitre dressera un état de l'art de cette problématique. Nous détaillerons dans un second chapitre la méthode que nous proposons. Le chapitre 3 présentera le cadre expérimental, et le chapitre 4 les expérimentations réalisées.

1. http://fr.wikipedia.org/
2. http://www.jeuxdemots.org/jdm-accueil.php

Chapitre 1

État de l'art

1.1 Relations étudiées

Afin de pleinement analyser les textes, bon nombre de systèmes en Traitement du Langage ont besoin de connaissances externes. En effet, celles-ci leur permettront à terme d'être en mesure de raisonner, même à partir de textes bruts.

Ces connaissances sont principalement structurées sous forme de relations, le plus souvent binaires. Celles-ci permettent en effet de relier deux entités ou concepts par ce que l'on appelle un prédicat (par exemple, *est-capitale(Paris,France)* est une entité de relation).

Certaines relations, dites « sémantiques », sont particulièrement étudiées dans ce domaine. Elles ont en effet pour avantage d'être déjà beaucoup décrites dans des ressources existantes, telles que WordNet [1] (Miller, 1995).

On peut distinguer les relations sémantiques hiérarchisantes de celles qui ne le sont pas. L'hyperonymie et l'hyponymie, aussi désignées comme relation *est-un*, permettent de catégoriser les termes (*chat/animal*). La méronymie et l'holonymie permettent de relier une partie à son tout (*porte/maison*). Ces quatre relations sont hiérarchisantes puisqu'elles peuvent alimenter un treillis de termes. Par exemple : *chat* $\rightarrow$ *félin* $\rightarrow$ *animal* $\rightarrow$ *être vivant*. A l'inverse, les relations de synonymie et d'antonymie ne permettent pas d'établir de hiérarchie entre les termes.

On peut citer les travaux de Caraballo (1999), Morin (1999) ou encore Ravi-

1. http://wordnet.princeton.edu/

chandran and Hovy (2002) qui se sont intéressés à la relation d'hyperonymie. La relation de méronymie a également suscité l'intérêt de nombreux travaux tels que : Winston et al. (1987); Berland and Charniak (1999); Girju et al. (2003); Pennacchiotti and Pantel (2006).

Il est important de remarquer que ces relations ne sont pas exclusives, à savoir que le terme *chat* n'est pas le seul à être en relation d'hyperonymie avec le terme *animal*. Ceci est généralement différent pour les autres relations d'ordre plus général. En effet, des relations du type *date-de-naissance-de*, *réalisé-par* ou encore *est-capitale* auront des instances uniques. Ceci peut avoir un impact sur les méthodes d'extraction ou d'évaluation employées.

1.2 Ressources existantes

Des projets recensant une multitude d'instances de relations sémantiques ont déjà été implémentés. On peut citer parmi les plus connus *WordNet* (Miller, 1995), *MultiNet*[2] (Helbig, 2005) ou encore *UMLS*[3] (Lindberg et al., 1993). Ces bases de connaissances, bien que généreusement fournies, ne sont pas complètes, et font régulièrement l'objet de travaux afin de les alimenter. Il parait évident qu'une base de connaissances complète ne peut être décemment remplie à la main, d'où l'utilité des méthodes automatiques présentées ci-après.

1.3 Extraction automatique de connaissances

Construire des bases de relations manuellement est évidemment coûteux en temps, et nécessite l'intervention de spécialistes. C'est pourquoi l'on va chercher à réaliser cette tâche automatiquement, à partir de données textuelles ou autres ressources existantes, afin de réduire la quantité astronomique de travail manuel d'acquisition de connaissances. On peut distinguer trois grandes familles de méthodes étudiées jusqu'alors : celles qui nécessitent l'exploitation de corpus afin d'observer le contexte des termes, celles exploitant d'autres ressources, telles que

2. http://pi7.fernuni-hagen.de/forschung/multinet/multinet_en.html
3. http://www.nlm.nih.gov/research/umls/

les terminologies, dictionnaires ou autres ontologies et celles qui s'appuient sur la structure interne des termes.

1.3.1 Corpus et contexte des termes

Les corpus de textes contiennent quantité d'information qu'il peut être utile d'extraire. En effet, en observant le contexte de termes liés par une relation, on peut en déduire des méthodes d'acquisition de connaissances.

Propriétés distributionnelles des mots

Le contexte des termes peut être exploité en réalisant une analyse distributionnelle, afin d'identifier une relation de proximité entre termes. Par exemple, Bourigault et al. (2004) rapprochent les termes « *insuffisance rénale* » et « *détresse respiratoire* » car on les trouve dans des contextes syntaxiques identiques tels que : « *prise en charge d'une insuffisance rénale* » et « *prise en charge d'une détresse respiratoire* » ou encore « *admettre en réanimation chirurgicale pour insuffisance rénale* » et « *admettre en réanimation chirurgicale pour détresse respiratoire* ». Ces techniques d'association de mots (Hindle, 1990; Ruge and Schwarz, 1991; Agarwal, 1995; Grefenstette, 1994) permettent généralement d'établir des classes sémantiques à partir d'une faible quantité de connaissances *a priori*. Par exemple, Hindle rapproche les noms : *ship, plane, bus, jet, vessel, truck, car, helicopter, ferry, man* par similarité au nom *boat*. Ces approches ne permettent cependant pas d'extraire des relations typées à partir de données textuelles.

L'approche proposée par Cimiano et al. (2003) forme un treillis de noms partageant les mêmes contextes, afin d'établir une hiérarchie de concepts. Elle utilise pour cela l'Analyse Formelle de Concepts (FCA), une méthode proposée par Ganter and Wille (1997) pour analyser des données et les structurer en unités représentant des abstractions formelles de concepts. Alors que Cimiano et al. se limitent aux relations d'hyperonymie afin de hiérarchiser les termes, Bendaoud et al. (2010) vont s'intéresser à tout type de relation, par exemple l'instance *(objet_céleste, isObservedBy, téléscope)*.

D'autres travaux se sont intéressés plus particulièrement à l'extraction de synonymes (Grefenstette, 1994; Ferret, 2010). Les mesures distributionnelles peuvent également être réalisées sur des classes sémantiques, issues de thésaurus, par

exemple *WordNet* (Resnik, 1992).

Patrons

De nombreuses méthodes d'extraction à base de patrons se sont appuyées sur les premiers travaux réalisés par Hearst en 1992. Celui-ci a en effet mis en place un processus itératif (Fig.1.1) permettant d'extraire des patrons à partir de quelques couples dits « sources » (*seed*). Ces patrons permettent alors de découvrir de nouveaux couples, à leur tour ajoutés aux couples sources pour l'itération suivante.

1. Définir la relation d'intérêt, par exemple l'hyperonymie
2. Fournir un ensemble de couples qui respectent cette relation, par exemple « *England-country* ». Ces couples « sources » (*seed*) peuvent être définis manuellement ou bien extraits d'une base de connaissance existante.
3. Recueillir les phrases contenant les deux termes de ces couples sources
4. Chercher des environnements communs parmi ces phrases, ceux-ci seront alors considérés comme patrons pour notre relation d'intérêt
5. Les patrons émergeants permettent de trouver de nouvelles instances pour la relation d'intérêt. Il s'agit alors de recommencer le processus à l'étape 2.

Figure 1.1 – Processus d'extraction de patrons et de nouveaux couples, proposé par Hearst (1992).

Selon lui, les patrons sont particulièrement intéressants, puisqu'ils sont fréquents et se trouvent dans de nombreux genre de textes. Ils peuvent par ailleurs être détectés avec pas ou peu de connaissances pré-requises. Les patrons qu'il présente sont dits « lexico-syntaxiques » car ils nécessitent une analyse syntaxique préalable de la phrase. Il extrait par exemple le patron « *NP {, NP}* {,} or other NP* » pour la relation d'hyperonymie. Sa technique souffre cependant d'un manque d'automatisation, puisque la recherche de patrons (étape 4) est réalisée manuellement. Il a également essayé de la transposer pour les relations de

méronymie, ce sans succès.

D'autres travaux ont alors suivi ses traces, tels que Auger (1997); Bourigault and Aussenac-Gilles (2003); Malaisé et al. (2005); Pennacchiotti and Pantel (2006); Marshman and L'Homme (2006), etc.

Morin (1999) s'est en particulier intéressé à la relation d'hyperonymie pour le français. Il a en effet mis en place un système appelé *PROMÉTHÉE*, qui extrait des patrons lexico-syntaxiques à l'instar de Hearst. Sa méthode (Fig.1.2) est en revanche plus automatisée. En effet, l'extraction de l'environnement commun de l'étape 4 n'est plus réalisée manuellement. Cependant, un analyste doit superviser les résultats trouvés automatiquement, et ce à chaque itération.

1. Définir la relation d'intérêt, ici l'hyperonymie
2. Fournir un ensemble de couples qui respectent cette relation, par exemple (*glycérol,polyol*). Ces couples « sources » (*seed*) peuvent être définis manuellement ou bien extraits d'une base de connaissance existante.
3. Recueillir les phrases contenant les deux termes lemmatisés de ces couples sources. Par exemple, le couple (*glycérol,polyol*) sélectionne la phrase « *L'hydrolyse des substrats est activée par le glucose et les polyols tels que le sorbitol et le glycérol.* »
4. Chercher des environnements communs parmi ces phrases, sous la forme d'expressions lexico-syntaxiques. Des schémas candidats sont donc extraits, ici : « NP tel que LISTE »
5. Valider les schémas candidats les plus pertinents
6. Utiliser les nouveaux schémas pour extraire de nouveaux couples de termes candidats
7. Valider les couples de termes candidats les plus pertinents. Ces nouveaux couples sont ajoutés à la liste de couples sources initiale, et le processus est réitéré à partir de l'étape 3.

Figure 1.2 – Processus d'extraction de patrons et de nouveaux couples, proposé par Morin (1999).

Cette méthode travaille sur des textes bruts, mais ceux-ci doivent être préalablement analysés. En effet, *PROMÉTHÉE* effectue une première étape de présyntaxe (préédition, segmentation en phrases et segmentation en occurrences de

formes), puis un étiquetage morpho-syntaxique ainsi qu'une lemmatisation. Il extrait ensuite les sigles, puis détecte les syntagmes nominaux (SN) et enfin les successions de syntagmes nominaux (autrement dit, les listes). Ainsi, son analyse de surface réalisée sur la phrase « *L'hydrolyse des substrats est activée par le glucose et les polyols tels que le sorbitol et le glycérol.* » produit : [L'hydrolyse des substrats]$_{SN}$ est activée par [le glucose]$_{SN}$ et [les polyols]$_{SN}$ tels que [le sorbitol et le glycérol]$_{LISTE}$.

Afin d'extraire des schémas candidats, les expressions lexico-syntaxiques obtenues sont classifiées par l'intermédiaire d'une mesure de similarité. Puis une expression de chaque classe est choisie comme représentante privilégiée de celle-ci. Le tableau 1.1 liste les patrons extraits par *PROMÉTHÉE*.

	SN (LISTE) manuellement raffiné en :	
1ère itération	`{certain|quelque|plusieurs...} SN ( LISTE )`	(1)
	`{deux|trois|quatre...|2|3|4...} SN ( LISTE )`	(2)
	SN : LISTE manuellement raffiné en :	
	`{certain|quelque|plusieurs...} SN : LISTE`	(3)
2nde itération	`{deux|trois|quatre...|2|3|4...} SN : LISTE`	(4)
	`SN , particulièrement SN ,`	(5)
	`{de autre}? SN tel que LISTE`	(6)
	`SN {et|ou} de autre SN`	(7)
3ème itération	`{de autre}? SN comme LISTE`	(8)
	`SN tel LISTE`	(9)
	`SN et notamment SN`	(10)
4ème itération	`chez le SN , SN ,`	(11)

Table 1.1 – Patrons extraits par le système PROMÉTHÉE (Morin, 1999)

Morin, comme la plupart des travaux réalisés en extraction de patrons lexico-syntaxiques, a travaillé sur des corpus spécifiques. Ces corpus ne sont en général pas assez fournis. En effet, certains patrons y sont difficiles à extraire car ils apparaissent trop peu. Les corpus extraits d'internet pallient à ce problème. Ils ont pour intérêt d'augmenter le rappel des patrons et des couples extraits, mais la précision, elle, souffre du bruit fatalement présent dans ce type de corpus.

Ravichandran and Hovy (2002) ont montré que de tels corpus obtenaient de meilleurs résultats, à savoir une meilleure précision, pour des relations du type *date-de-naissance* que pour des relations sémantiques telles que l'hyperonymie.

Ils se sont en effet intéressés à l'extraction de connaissances pour un système de Question-Réponse. Leur méthode utilise un arbre de suffixes afin de trouver les patrons récurrents dans les contextes des termes (Fig.1.3).

1. Choisir un exemple pour un type de question. Par exemple, pour une date de naissance, la question est « *Mozart* » et la réponse est « *1756* »
2. Soumettre la question et la réponse à un moteur de recherche
3. Récupérer les 1 000 premiers documents retournés
4. Segmenter ces documents en phrases
5. Ne garder que les phrases qui contiennent la question et la réponse, les tokéniser
6. Construire un arbre de suffixes pour trouver toutes les sous-chaînes de toutes longueurs
7. Ne garder que les sous-chaînes qui contiennent et la question et la réponse
8. Remplacer « *Mozart* » et « *1756* » par <NAME> et <ANSWER>

Pour calculer la précision des patrons :

1. Soumettre seulement la question au moteur de recherche, ici « *Mozart* »
2. Récupérer les 1 000 premiers documents
3. Segmenter ces documents en phrases
4. Ne garder que les phrases contenant la question
5. Chercher tous les patrons obtenus précédemment. Compter ceux qui contiennent la bonne réponse dans le champ <ANSWER>, et ceux qui contiennent un autre terme
6. Ne garder que les patrons qui apparaissent suffisamment (plus de 5 fois)

Figure 1.3 – Processus d'extraction de patrons pour un système de Question-Réponse, proposé par Ravichandran and Hovy (2002).

Le score de confiance des patrons est calculé en se basant sur le fait que ce type de relations contient une réponse dite « correcte ». Ceci explique le fait que cette méthode soit moins performante sur des relations sémantiques telles que

l'hyperonymie.

Brin (1999) a également travaillé sur l'extraction de patrons et de relations à partir du Web. Son système, *DIPRE* (pour Dual Iterative Pattern Relation Extraction), cherche à extraire des paires de relations décrivant l'auteur d'un livre : *(author, title)*. Cette méthode suit encore une fois le même processus itératif (Fig. 1.4).

1. Fournir un petit échantillon de couples répondant à la relation étudiée, ici une liste de 5 livres et de leurs auteurs
2. Chercher toutes les occurrences de ces couples, et conserveur leur contexte : *(author, title, order, url, prefix, middle, suffix)*
3. Générer des patrons à partir de ces contextes
4. Chercher de nouveaux couples à l'aide des patrons extraits, puis réitérer à l'étape 2

Figure 1.4 – Processus d'extraction de patrons et de relations, proposé par Brin (1999).

Pour générer des patrons, *DIPRE* vérifie dans un premier temps que l'ordre des deux termes (*order*) et le contexte intérieur (*middle*) sont identiques. Pour tous les contextes qui répondent à ces critères, il conserve les plus longs préfixes et suffixes commun. Pour éviter que les patrons soient trop généraux et produisent par la suite des couples sans lien avec la relation étudiée, il établit un score de spécificité. Cette méthode semble très efficace pour la relation *(author,title)* puisque les résultats retournés sont pertinents.

Ce système a d'ailleurs été repris par Agichtein and Gravano (2000) avec *Snowball*, et adapté à la relation indiquant le lieu du siège d'une entreprise. En plus de la méthodologie de *DIPRE*, *Snowball* inclut les entités nommées. Celles-ci sont en effet utiles pour la relation qu'ils étudient, puisqu'ils sont en mesure de détecter les lieux (<LOCATION>) et les noms des entreprises (<ORGANIZATION>). *Snowball* établit également un score de confiance aux patrons en s'appuyant sur la sélectivité. Autrement dit, si un patron initialement fort génère un couple avec un bon score et une organisation identique à un couple déjà connu et également bien scoré, il vérifie si les lieux associés sont les mêmes. Si oui, le patron est considéré comme positif, si non, sa confiance est diminuée.

A l'instar du système proposé par Ravichandran and Hovy (2002), ces méthodes sont efficaces et ont pour particularité de s'intéresser aux relations exclusives. En effet, pour calculer un score de confiance, ces techniques doivent être en mesure de détecter si un couple généré est « faux », ce qui n'est pas possible avec les relations sémantiques.

A notre connaissance, la seule méthode faisant appel à des patrons uniquement lexicaux pour extraire des relations sémantiques, est l'approche proposée par Ortega-Mendoza et al. (2007). Ceux-ci travaillent en effet sur la relation d'hyponymie pour l'espagnol. Cette méthode reprend les principes précédents en partant d'un ensemble de couples sources afin d'extraire des patrons. Les scores de confiance attribués à ceux-ci sont obtenus à l'aide de l'Information Mutuelle (Kullback, 1959).

1.3.2 Terminologies, dictionnaires, ontologies

Nous détaillerons moins les méthodes alternatives d'extraction automatique de relations, puisqu'elles sont hors de notre champ de recherche. Il est toutefois intéressant de noter que des travaux ont été réalisés en s'appuyant uniquement sur des ressources existantes. On peut notamment citer les travaux de Rigau (1998) sur l'acquisition de connaissances à partir de dictionnaires électroniques (MRDs : Machine-Readable Dictionnaries). Ou encore ceux de Harabagiu and Moldovan (1998) qui infèrent des relations à partir de *WordNet*.

1.3.3 Exploitation de la structure interne des termes

Certains travaux se sont également appuyés sur la structure interne des termes, à savoir les informations morphologiques, syntaxiques voire sémantiques (Manser, 2012). On peut par exemple citer Zweigenbaum and Grabar (2000), Daille (2003), Claveau and L'Homme (2005), ...

1.4 Une alternative plus ludique

Les méthodes automatiques citées précédemment ont pour inconvénient d'être dépendantes des ressources utilisées. Lafourcade and Joubert (2008) ont ainsi pro-

posé une approche indépendante de tout corpus, et ne nécessitant pas le travail coûteux de spécialistes du domaine. Ils ont en effet mis en place ce que l'on appelle un « jeu sérieux », afin d'alimenter un réseau de relations typées et pondérées. L'approche ludique visant à faire contribuer des joueurs non-spécialistes a déjà été proposée dans d'autres domaines, tel que l'indexation d'images (von Ahn and Dabbish, 2004), la reconnaissance d'objets au sein d'images (von Ahn et al., 2006) ou encore la biologie (Cooper et al., 2010). En Traitement du Langage, cette approche à également été utilisée pour les paraphrases (Bouamor et al., 2009). Dans le domaine des réseaux lexicaux, un travail similaire a été réalisé sur le portugais par Mangeot and Ramisch (2012).

JeuxDeMots est indépendant, c'est-à-dire qu'il ne nécessite pas l'intervention d'un modérateur-expert humain. Afin d'éviter les erreurs intentionnelles ou non des utilisateurs, la méthode choisie est une validation des relations proposées par concordance des propositions entre paires de joueurs. Ainsi, pour une question posée, deux joueurs émettent des propositions en « double aveugle ». Leurs propositions sont ensuite comparées, et si elles concordent, elles alimentent le réseau lexical. Plus les joueurs ont de réponses communes, plus ils gagnent de points, leur intérêt est donc de donner des réponses « correctes ». Afin d'éviter de ne recueillir que les relations les plus évidentes, celles-ci perdent en valeur au fur et à mesure qu'elles sont jouées. Les utilisateurs vont ainsi balayer un spectre plus large de relations.

A ce jour, le réseau lexical obtenu avec *JeuxDeMots* compte plus de 280 000 nœuds (termes et concepts), et plus de 8 000 000 relations, allant de la simple relation non typée aux relations de synonymie, d'hyperonymie, de couleur ou même de conséquence.

Chapitre 2

Méthode proposée

Nous avons opté pour une approche à base de patrons lexicaux. Contrairement à la plupart des travaux faisant appel aux patrons, nous nous concentrons uniquement au niveau lexical. En effet, nous ne tenons compte d'aucune information syntaxique ou encore sémantique. Nous nous contentons d'observer le texte quasi-brut. Le seul traitement apporté à celui-ci est un découpage en phrases puis en unités lexicales.

Notre méthode se découpe en 3 phases principales. Dans un premier temps, nous recueillons des couples correspondant aux relations que nous étudions. Par exemple, nous récupérons le couple *(chat/animal)* pour la relation d'hyperonymie. Pour ce faire, nous utilisons une vaste ressource : JeuxDeMots. Ensuite, nous avons besoin d'extraire le contexte proche des couples récoltés à l'étape précédente, au sein d'une grande quantité de textes. Nous utilisons donc un corpus composé d'articles de Wikipédia (le chapitre suivant présente une description des ressources utilisées). De ces contextes, nous tirons des patrons lexicaux. Enfin, nous re-parcourons Wikipédia, munis de nos patrons, afin de trouver d'autres couples correspondant aux relations étudiées. Le schéma 2.1 présente la chaîne de traitement de notre système.

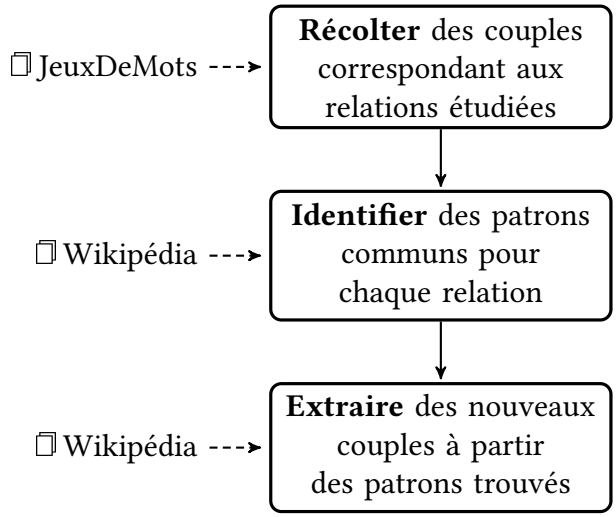

Figure 2.1 – Processus d'extraction de patrons et de couples

2.1 Extraction de couples à partir de JeuxDeMots

La plupart des travaux à base de patrons lexico-syntaxiques proposent une méthode itérative. Celle-ci ne nécessite que peu de couples sources à la première étape (également appelés *seed*). La suite du processus permet de récolter de nouveaux couples et donc de passer à l'itération suivante avec plus de couples sources. Il est ainsi nécessaire, à chaque étape, de valider les couples récoltés, afin de ne pas altérer la suite du processus.

Nous avons opté pour une approche sensiblement différente : utiliser un maximum de couples sources dès le début. Notre méthode est donc entièrement automatique et ne nécessite aucune intervention, à aucun moment du processus. Pour récolter ces couples sources, nous faisons appel à une ressource construite collaborativement : les données issues de JeuxDeMots. Comme expliqué précédemment, ces données comportent entre autres des relations sémantiques.

Nous avons choisi de nous intéresser aux relations suivantes : l'hyperonymie et l'hyponymie (relations de spécialisation), la méronymie et l'holonymie (relations partie-tout), la synonymie et enfin l'antonymie. Les études se limitent souvent à l'une de ces relations, mais puisque nous nous contentons d'observer le contexte lexical des couples, notre méthode n'a pas de raisons de se spécialiser dans l'une ou l'autre de ces relations. La première phase de notre méthode va donc consister à récupérer un nombre important de couples pour chacune de ces

relations.

2.2 Extraction des patrons lexicaux représentatifs d'une relation

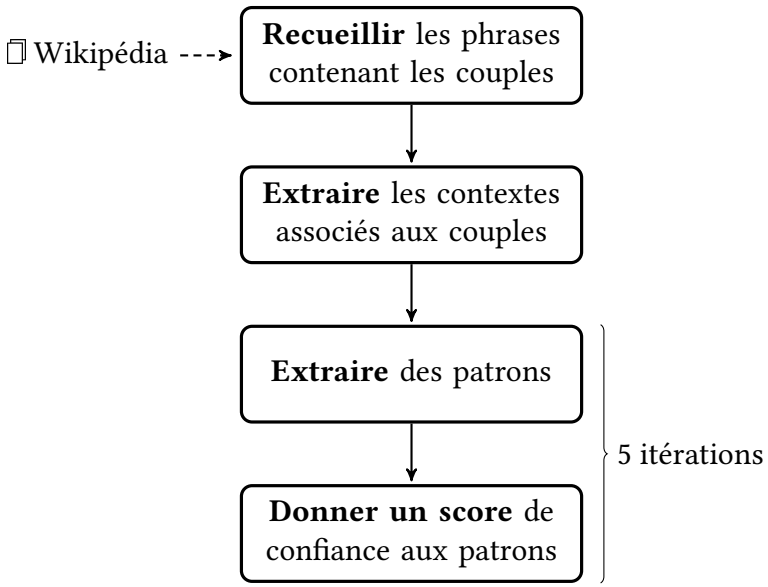

Figure 2.2 – Processus détaillé d'extraction de patrons

Cette seconde phase va permettre d'extraire les patrons lexicaux représentatifs de chacune des nos 6 relations étudiées. Pour ce faire, on va chercher dans un grand volume de textes les apparitions proches des deux éléments d'un couple d'une relation donnée. On extrait donc dans un premier temps les phrases contenant deux éléments d'un même couple. On va ensuite tâcher d'extraire des patrons récurrents dans le contexte proche de ces deux éléments. Notre fenêtre d'étude est la suivante : nous observons les 3 termes précédant et suivant les deux éléments de notre couple X/Y (voir Fig. 2.3).

Figure 2.3 – Fenêtre d'étude du contexte du couple hyperonyme `chat/animal`.

Ce que l'on considère comme « contexte » est donc constitué d'au maximum 12 termes. S'il y a par exemple moins de 3 termes à la gauche du X, seuls ceux-ci seront récupérés (cf exemple n°1 de la Fig. 2.4). De la même manière, s'il y a moins de 6 termes entre X et Y, ceux-ci seront d'abord considérés comme « à droite de X », puis s'il en reste, « à gauche de Y » (cf exemple n°7 de la Fig. 2.4). Ainsi, l'objectif est de capturer les termes proches du X et les termes proches du Y. Nous n'imposons pas de seuil de distance entre les termes X et Y. En effet, cette représentation du contexte a pour avantage d'être valide même si une proposition vient s'intercaler entre les deux termes.

1. le **chat** c' est l' **animal** qui se hérisse quand il fait le dos rond .
2. le **chat** est aussi un **animal** lié au diable , à plus forte raison s' il est noir .
3. le **chat** reste dans l' imaginaire collectif un **animal** inquiétant , rusé , magique .
4. le **chat** , commensal et **animal** de compagnie de l' homme , peut transmettre des maladies [...] .
5. ainsi le squelette d' un **chat** retrouvé auprès d' un tombeau humain indique qu' il s' agissait probablement d' un **animal** de compagnie .
6. bien qu' il ne s' agisse bien sûr pas forcément d' une canine de **chat** , mais d' un quelconque autre **animal** .
7. pendant longtemps le **chat** a été considéré comme un **animal** essentiellement solitaire et les groupes groupes comme [...] .
8. d'après des études , le **chat** est l' **animal** ayant la plus grande proportion de phases de sommeil paradoxal pendant lesquelles il rêve .
9. malgré la domestication , le **chat** reste un **animal** avec des instincts qu' il a besoin d' exprimer
10. le **chat** a en général tendance à affirmer son indépendance vis-à-vis des autres occupants d' un lieu , car c' est un **animal** avant tout territorial .
11. le **chat** est un petit **animal** félin .
12. le **chat** à pattes noires est un **animal** solitaire et nocturne , extrêmement discret .

Figure 2.4 – Exemples de contextes pour le couple hyperonyme *(chat/animal)*.

Il s'agit maintenant de détecter les redondances spécifiques aux relations étu-

diées au sein de ces contextes. Parmi les exemples de la figure 2.4, on peut constater que les contextes 2, 10, 11 et 12 sont en partie de la forme : « `X est un Y` ».

Nous optons pour accorder le plus de liberté possible aux patrons que nous cherchons à extraire. En effet, nous ne cherchons pas des mots à des positions fixes. Cela nous permet de capturer des patrons, même si des adjectifs ou adverbes sont venus se glisser entre les termes. Par exemple, nous considérons que les phrases « *le chat est un animal* » et « *le chat est un petit animal* » correspondent toutes deux au patron « `le X est un Y` ».

Nous cherchons à extraire des patrons contenant de 1 à 5 termes (excepté X et Y). Face à la complexité de recherche de sous-chaînes au sein de nos contextes, nous procédons par itérations. Dans un premier temps, nous extrayons des patrons de longueur 1, autrement dit, nous cherchons les mots les plus courants au sein de nos contextes. Pour passer à l'itération suivante, nous gardons les n premiers termes de l'étape précédente, et cherchons les patrons de longueur $i+1$ contenant tous les termes d'un patron de longueur i. Cela permet de limiter la recherche en mettant de côté les patrons plus rares. La figure 2.5 donne un exemple de ce processus.

…	X	…	…	Y	…
			Étape 1		
	X	,		Y	
	X	est		Y	
	X		un	Y	
	X			Y	de
le	X			Y	
			Étape 2		
	X	est un		Y	
le	X	est		Y	
	X	est		Y	de
			Étape 3		
le	X	est un		Y	
	X	est un		Y	de
un	X	est un		Y	

Figure 2.5 – Échantillons d'itérations pour l'hypéronymie.
En bleu, le patron qui sera « étendu » à l'itération suivante. En rouge, le patron de l'étape précédente.

Une fois que l'on récolte les patrons étendus (de longueur *i+1*), il s'agit alors de les scorer. Il sera ainsi possible de sélectionner à nouveau les *n* meilleurs patrons pour procéder à l'étape *i+2*.

Nous utilisons une fonction de score inspirée du TF-IDF. Celle-ci va en effet pondérer la fréquence des patrons observés dans le contexte des couples *seed* avec leur spécificité. En effet, nous souhaitons éviter que les patrons obtenus génèrent toutes sortes de couples qui n'ont rien à voir avec la relation d'intérêt. Nous réunissons donc tous les couples candidats qu'il est possible de trouver dans Wikipedia (voir section suivante), et observons leur contexte proche, puis nous comptons le nombre de fois où le patron apparaît. Ainsi, si un patron apparaît très fréquemment dans le corpus, il sera considéré comme peu spécifique, et son score diminuera.

$$score = freq\ patron \times log\left(\frac{\Pi}{\Pi_{patron}}\right) \tag{2.1}$$

avec :

Π le nombre de contextes associés aux couples candidats possibles

Π_{patron} le sous-ensemble de Π où le patron apparaît

2.3 Recherche de nouveaux couples dans Wikipédia

A ce stade, nous disposons de patrons scorés représentatifs de chaque relation. Cette troisième étape consiste à parcourir à nouveau le corpus Wikipédia afin de détecter des nouveaux couples susceptibles de partager une de nos relations étudiées.

Pour cela, nous recherchons dans un premier temps tous les couples de termes candidats. Pour éviter un calcul exponentiel, nous nous réduisons à la recherche de couples nom/nom, et ce au sein d'une phrase. Pour chacun de ces couples candidats, nous extrayons le contexte correspondant, comme défini auparavant. A partir de ce contexte, nous récupérons la liste de patrons potentiels, et vérifions si nous les avions trouvés à l'étape précédente. Ainsi, nous pouvons en déduire le score d'un couple candidat : il correspond à la somme des scores des patrons

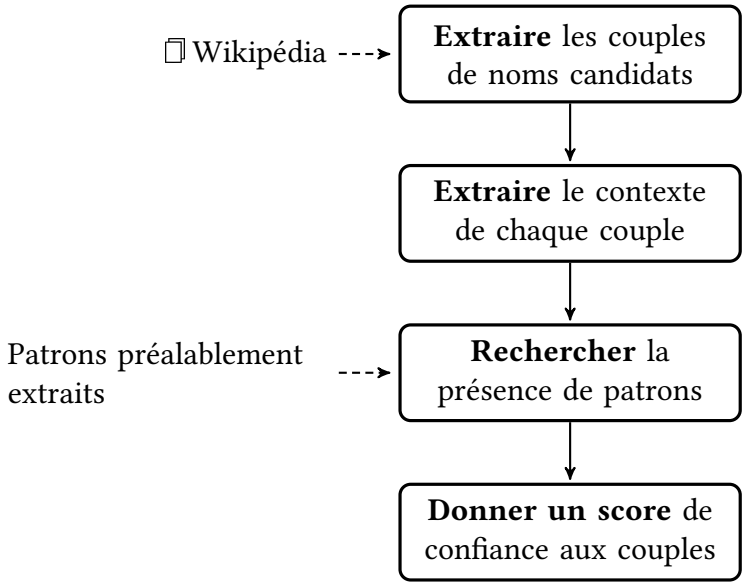

Figure 2.6 – Processus détaillé d'extraction de couples

générés par celui-ci.

Chapitre 3

Cadre expérimental

3.1 Description des données

3.1.1 JeuxDeMots

Pour recueillir les relations présentes dans JeuxDeMots, nous avons téléchargé les données datant du 9 février 2014. Le fichier se compose alors de 282 700 nœuds et de 8 004 458 relations différentes. Parmi ces données, nous ne conservons que les nœuds de type *n_term* et *n_concept*. En effet, les autres nœuds correspondent à des expressions ou des questions entières, des étiquettes grammaticales, etc. et donc ne nous intéressent pas. Au final, nous travaillons donc avec 275 660 nœuds :

- 274 658 de type 'terme'
- 1 002 de type 'concept'

Concernant les relations, nous ne gardons que celles sur lesquelles portent cette étude, à savoir :

- 154 753 hyperonymes
- 27 672 hyponymes
- 357 435 méronymes
- 32 823 holonymes
- 175 213 synonymes
- 16 174 antonymes

Soit plus de 750 000 relations.

Raffinement sémantique

Les nœuds sont parfois raffinés. Par exemple, le terme « *palmier* » compte deux raffinements sémantiques dans JeuxDeMots : selon si l'on parle de l'arbre ou du gâteau. Les données contiennent donc trois nœuds différents pour ce terme : « *palmier* », « *palmier(arbre)* » et « *palmier(gâteau)* ». Chacun possède son identifiant unique (eid).

On pourrait penser que le terme non raffiné contient toutes les relations des deux autres réunis, et que les deux raffinements servent à spécifier à quel type de « *palmier* » elles sont associées, mais ce n'est pas le cas. En effet, ces mots sont presque indépendants, et si un joueur a créé une association avec le terme « *palmier(arbre)* », celle-ci ne sera pas automatiquement associée au terme « *palmier* » non raffiné.

Pour ne pas supprimer trop de relations, nous gardons donc à la fois les relations des termes non raffinés, et celles de leurs raffinements sémantiques. Toutes ces relations seront associées à un seul terme, ici « *palmier* », car nous travaillons par la suite au niveau lexical, donc le sens précis des mots ne nous intéresse pas.

Seuil minimal des relations

Nous travaillons avec des données générées par des utilisateurs qui ne sont pas spécifiquement linguistes, bien au contraire. Ce sont donc forcément des données bruitées. En effet, un joueur a tout le loisir de proposer une relation erronée, que ce soit volontaire ou non. Il y a par exemple confusion pour certains joueurs entre les termes « *Dalila* » et « *Dalida* ». Pour éviter certaines erreurs, nous ne gardons pas les relations qui viennent juste d'apparaître, mais uniquement celles qui ont un poids minimum de 150. Les relations sont en effet pondérées, une relation entre en base de données avec le poids de 50, puis celui-ci est incrémenté de 10 à chaque fois qu'elle est rejouée. En ne conservant que les relations avec un poids supérieur à 150, nous nous assurons un minimum de leur fiabilité, puisqu'elles ont été jouées par au moins 11 paires de joueurs. Ce seuil est un paramètre qu'il est possible de modifier à tout moment. Plus il sera grand, plus les relations seront fiables, mais diminueront en nombre.

3.1.2 Wikipédia

Afin de faire des observations au niveau lexical, nous avons besoin d'un gros volume de textes. Nous avons donc opté pour Wikipédia. En effet, nous pouvons ainsi travailler sur de grandes quantités de textes, accessibles et non spécialisés. Ceci correspond donc au type de vocabulaire susceptible d'avoir été joué dans JeuxDeMots.

Le corpus que nous avons utilisé est tiré d'une version de 2008 de Wikipédia français, il a été mis à disposition par Franck Sajous [1]. Il est composé de 664 982 articles. A titre d'information, Wikipédia compte à ce jour plus de 1 500 000 articles français mais la version de 2008 était amplement suffisante pour nos recherches.

Ce corpus ne contient que les parties textuelles des articles. Les sommaires numérotés de début d'articles ont été supprimés, ainsi que les sections « voir aussi » (liens vers des références externes). Les parties « notes » ont été conservées.

Des expérimentations ont également été réalisées avec d'autres corpus mais ceux-ci étaient généralement trop spécifiques, on n'y retrouvait donc qu'une faible portion des termes et relations générées par JeuxDeMots. Il a donc été décidé de n'utiliser que le corpus WikipédiaFR.

3.2 Mesures d'évaluation

L'évaluation s'avère ardue pour ce type de tâche, car nous travaillons sur des données brutes, non annotées. En effet, dans l'idéal, il aurait fallu que notre corpus contienne des annotations pour chaque couple associé à chacune de nos relations étudiées. L'évaluation doit donc être manuelle. Il est également possible d'aller puiser des couples « référence » dans une ressource externe, mais rien ne garantit que ceux-ci soient suffisamment représentés dans le corpus.

Nous avons procédé de la manière suivante. Dans un premier temps, nous recherchons tous les couples initiaux de JeuxDeMots qui ont servi à extraire les patrons. La logique voudrait en effet que ces couples se retrouvent à terme dans les candidats générés par ces dits-patrons.

Nous avons ensuite réalisé une évaluation manuelle des couples extraits les

1. http://redac.univ-tlse2.fr/corpus/wikipedia.html

mieux notés. Ceci nous permet d'évaluer la précision des éléments retournés, mais non le rappel.

Chapitre 4

Expérimentations et résultats

4.1 Expérimentations

La première étape du travail consiste à extraire les phrases du corpus Wikipédia contenant les couples recueillis sur JeuxDeMots. Parmi les phrases extraites, nous avons remarqué que certains couples étaient bien trop majoritairement représentés. Par exemple, les couple hyperonymes *(un,nombre)* ou *(qui,mot)* permettaient de récolter une grande majorité des phrases, alors qu'ils ne sont absolument pas représentatifs de la relation en question. Pour le premier, toute phrase contenant le terme *nombre* est à peu près assurée de contenir aussi le déterminant *un*, sans pour autant qu'ils soient en quelconque relation d'hyperonymie. Cela peut aussi être le cas pour d'autres couples, et nous ne pouvons pas tous les vérifier, sinon la méthode perdrait sa qualification d'automatique. Cependant, nous avons malgré tout filtré les couples qui généraient « trop » de phrases, car ils auraient trompé le système aux itérations suivantes. Nous avons donc mis en place une liste noire d'une vingtaine de termes, comprenant notamment les termes se rapportant au nombre, au temps (mois, année, ...), à la couleur, ainsi que quelques adverbes. En définitive, le nombre de couples de JeuxDeMots extraits présent dans Wikipédia est présenté dans le tableau 4.1.

Les premières expérimentations ont été réalisées sur la relation d'hyperonymie. On peut cependant regretter de ne pas avoir été en mesure de mener nos investigations dans leur intégralité.

En ce qui concerne l'extraction de contextes contenant nos couples *seed*, nous

Relation	Seed	Seed trouvés dans Wiki	Contextes extraits
Hyperonymie	659	300	12 314
Hyponymie	555	243	19 574
Méronymie	769	565	50 347
Holonymie	467	377	56 648
Synonymie	505	277	12 484
Antonymie	313	204	21 170
TOTAL	3 268	1 966	172 537

Table 4.1 – Nombre de couples « source » pour chaque relation, ainsi que le nombre de phrases qui les contiennent

avons limité nos recherches aux phrases comprenant les deux termes du couple apparaissant dans le bon sens. Autrement dit, pour le couple *(chat,félin)*, nous ne sélectionnons pas une phrase du type : *Un félin tel que le chat […]*. Ce choix a uniquement été fait afin d'alléger le processus, et d'obtenir des premiers résultats, mais il serait également judicieux de s'intéresser aux patrons liant les deux termes en sens inverse.

Une fois nos contextes extraits, soit 12 314 pour la relation d'hyperonymie, nous nous sommes attelés à la tâche d'extraction de patrons. La première étape de cette extraction, à savoir celle qui génère des patrons de longueur 1 (en plus des deux termes), produit 22 590 patrons différents. Nous avons ensuite opté pour sélectionner les 500 patrons ayant le plus haut score, afin d'amorcer l'étape suivante. Celle-ci produit alors 10 244 patrons. La troisième étape en produit 14 948. Nous avons poursuivi l'expérience jusqu'aux patrons de longueur 5, même si ceux-ci semblent trop précis pour la tâche à laquelle nous nous intéressons. Nous estimons en effet que les patrons les plus fertiles seront plutôt de longueur 2 ou 3. En ce qui concerne la limite de 500 patrons choisie, cela a été fait de manière arbitraire afin de limiter la combinatoire. Il pourrait malgré tout être intéressant d'expérimenter la génération de patrons avec une limite plus importante, afin de ne pas éliminer de patrons potentiellement intéressants dès les premières étapes.

A ce stade, peu de tests ont pu être effectués, car le calcul du score de confiance d'un patron donné nécessite d'aller voir si celui-ci apparaît dans tous les contextes de tous les couples de noms possibles. En effet, ceci permet de mesurer si un patron est spécifique à une relation. Cependant, cela demande un temps de calcul

important. Nous avons essayé de « simplifier » notre méthode de calcul de score, en évitant cette combinatoire-ci, mais les résultats en souffraient significativement. En effet, pour calculer la spécificité d'un patron, nous nous contentions d'observer la somme des fréquences des termes composant le patron (les fréquences de chaque terme du corpus étant préalablement calculées, le calcul de score s'avérait bien plus rapide). Cette simplification de calcul de score est une tâche complexe, puisque la plupart des patrons sont constitués de termes-outils. Si l'on filtre trop les termes fréquents, on ne sortira pas de patrons valides, mais si on ne le fait pas assez, les patrons les mieux notés seront par exemple des patrons constitués uniquement de virgules ou de déterminants.

En ce qui concerne l'étape d'extraction de nouveaux couples, nous avons également affaire à des temps de calcul non négligeables. En effet, il faut encore une fois manipuler tous les couples de noms possibles de Wikipédia. A ce stade, peu de tests ont donc pu être effectués quant à la fonction de score utilisée pour noter les couples candidats.

Dans les premiers résultats que nous avons obtenu, nous avons réalisé une évaluation manuelle des 100 premiers couples d'hyperonymes générés par le système. Ces résultats sont calculés sur une portion des couples candidats (environ 1/8ème du corpus), soit 4 467 737 couples. Sur les 100 couples évalués, nous comptons 33 couples exacts, dont 21 n'étaient pas présents dans les relations sources de JeuxDeMots, par exemple le couple *(résistant, homme)*. Nous avons également fait l'observation suivante : 27 des couples obtenus sont ce que l'on appelles des co-hyponymes, à savoir des termes partageant la même relation d'hyperonymie. Par exemple, nous obtenons le couple *(bonda, prié)*, qui sont tous les deux des cépages de la Vallée d'Aoste.

Nous avons également réalisé l'évaluation consistant à mesurer le pourcentage de couples sources extraits sur la même portion de couples candidats. Il s'avère que l'on retrouve 23% des couples *seed*. Ce résultat n'est cependant pas significatif puisqu'il n'a pas été réalisé sur l'ensemble du corpus, certains couples ne sont donc pas ou peu représentés.

4.2 Discussion

A ce stade des expérimentations, il est évidemment trop tôt pour dresser des conclusions sur l'efficacité de cette méthode. On peut cependant réaliser quelques observations et en tirer des enseignements pour des travaux futurs.

Tout d'abord, il ne faut pas oublier qu'à plusieurs reprises lors de la chaîne de traitement, nous avons dû faire des choix plus ou moins arbitraires vis-à-vis de certains paramètres. Il serait bon de les faire varier afin de trouver une meilleure configuration pour nos extractions. Cela part du seuil imposé pour la récolte des couples de JeuxDeMots (à savoir 150), à la taille de la fenêtre d'étude, en passant par la limite de patrons choisie pour passer à l'itération suivante lors de l'extraction de patrons. De plus, nous aurions aimé avoir l'occasion d'expérimenter plusieurs fonctions de score pour l'extraction de couples candidats. En effet, nous nous contentons à l'heure actuelle de faire une somme des scores des patrons qu'un couple génère. Il faudrait donc introduire une certaine pondération afin de limiter l'impact des couples sur-représentés dans le corpus.

Les quelques résultats obtenus nous apprennent entre autres que les patrons d'hyperonymie permettent d'extraire des couples de co-hyponymes, ce que nous n'avions pas prévu initialement. Cela peut notamment s'expliquer par la présence de listes dans les phrases. Comme nous travaillons au niveau lexical, nous ne prenons pas en compte les énumérations, mais celles-ci peuvent contenir plusieurs co-hyponymes à proximité d'un hyperonyme.

L'évaluation réalisée sur une portion du corpus n'est pas assez représentative pour être analysée. Cependant, il est évident que si de tels résultats s'appliquaient au corpus dans sa totalité, le système en l'état ne serait pas assez performant. Nous restons cela dit optimistes, puisque ceux-ci ont toutes chances de s'améliorer lors d'expérimentations complètes (notamment le pourcentage de couples *seed* récupérés par le système : celui-ci devrait être bien supérieur sur le corpus en entier). Par ailleurs, de légères améliorations aux pondérations effectuées à chaque étape devraient permettre d'obtenir une meilleure précision, et d'éliminer les couples « négatifs », qui sont en réalité de simples couples de co-occurrence. Si Ortega-Mendoza et al. (2007) ont su obtenir des résultats satisfaisants sur l'espagnol avec une méthode itérative et des patrons lexicaux, il n'y a pas de raisons que cette méthode utilisant plus de couples *seed* ne fonctionne

pas.

Nous n'avons à l'heure actuelle trop peu d'éléments au sujet des autres relations. On peut néanmoins supposer que la relation de synonymie devrait obtenir de moins bons résultats que ses paires. En effet, il est rare de trouver deux synonymes au sein d'une même phrase. Ou alors ce sont des quasi-synonymes, et s'ils sont tous les deux présents, c'est sans doute pour mettre en avant ce qui les différencie justement.

Nous avons également constaté que le corpus Wikipédia utilisé n'était pas forcément idéal, et comprenait un bon nombre d'erreurs introduisant du bruit. Il aurait été fastidieux de reconstituer un corpus propre de texte brut, donc nous avons effectué les premiers tests sur cette version de 2008. Ce corpus contenait notamment beaucoup d'erreurs de termes accolés (sans doute les termes qui composaient des liens dans Wikipédia, et qui ont mal été transformés en texte brut), ou autres erreurs de segmentation. Ce point est assez facile à régler, puisqu'il existe sans doute d'autres corpus Wikipédia de textes bruts. Un co-stagiaire au RALI en a d'ailleurs constitué un ces derniers jours avec les 1 500 000 articles actuels de Wikipédia. Il serait ainsi intéressant de déployer notre système sur un tel corpus.

De la même manière, en filtrant les relations trop peu jouées sur JeuxDeMots, nous nous retrouvons au final avec un ensemble de moins d'un millier de couples *seed*, alors qu'à l'origine, JeuxDeMots propose plusieurs dizaines de milliers d'instances pour chaque relation. Ces instances sont beaucoup bruitées, on peut donc déplorer que le jeu ait été « trop peu joué ». En effet, cela aurait permis d'élargir notre ensemble d'amorces.

Pour finir, on remarque aussi que certaines des erreurs liées à l'utilisation de patrons uniquement lexicaux pourraient être évitées si l'on s'intéressait également au niveau syntaxique, comme par exemple le cas de la reconnaissance d'énumération. Cela étant, une analyse syntaxique introduirait aussi des erreurs, différentes, car les méthodes actuelles d'étiquetage morpho-syntaxique ou encore de reconnaissance de constituants syntaxiques ne sont pas forcément efficaces à 100%.

Conclusion

La méthode proposée ici permet donc d'extraire des patrons lexicaux significatifs des relations sémantiques. Son originalité réside dans le fait qu'elle se contente de travailler au niveau lexical, et que les couples fournis au début du processus sont tirés d'une ressource peu commune : JeuxDeMots. Elle a aussi pour avantage d'être indépendante : il n'y a à aucun moment intervention d'un analyste. On déplorera cependant le manque de résultats obtenus à ce stade : présenter une méthode sans pouvoir la vérifier correctement s'avère quelque peu frustrant. Les résultats sont cependant encourageants, et démontrent qu'il est effectivement possible d'extraire de nouveaux couples par simple observation du contexte des instances déjà connues.

Comme mentionné précédemment, il reste encore à mener de plus amples investigations pour valider le modèle proposé. Des expérimentations doivent être réalisées sur la variation des paramètres ainsi que sur la pondération du score attribué aux couples candidats.

Bibliographie

Agarwal, R. (1995). *Semantic Feature Extraction from Technical Texts with Limited Human Intervention.* PhD thesis, Mississippi State, MS, USA. UMI Order No. GAX95-35518.

Agichtein, E. and Gravano, L. (2000). Snowball : Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 85–94, New York, NY, USA. ACM.

Auger, A. (1997). *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles.* Université de Neuchâtel Faculté des lettres.

Bendaoud, R., Toussaint, Y., and Napoli, A. (2010). L'analyse Formelle de Concepts au service de la construction et l'enrichissement d'une ontologie. *Revue des Nouvelles Technologies de l'Information*, E-18 :133–164.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 57–64, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bouamor, H., Max, A., and Vilnat, A. (2009). Amener des utilisateurs à créer et évaluer des paraphrases par le jeu. In *Actes de TALN, session de demonstrations*, Senlis, France.

Bourigault, D. and Aussenac-Gilles, N. (2003). Construction d'ontologies à partir de textes . In *TALN 2003 : 10° conférence sur le Traitement Automatique des Langages Naturelles , Batz-sur-Mer (F), 11/06/2003-14/06/2003*, pages 27–47. Université de Nantes.

Bourigault, D., Aussenac-Gilles, N., and Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, pages 87–110.

Brin, S. (1999). Extracting patterns and relations from the world wide web. In *Selected Papers from the International Workshop on The World Wide Web and Databases*, WebDB '98, pages 172–183, London, UK, UK. Springer-Verlag.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 120–126, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cimiano, P., Staab, S., and Tane, J. (2003). Automatic acquisition of taxonomies from text : Fca meets nlp. In *Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia*, pages 10–17.

Claveau, V. and L'Homme, M.-C. (2005). Apprentissage par analogie pour la structuration de terminologie - utilisation comparée de ressources endogènes et exogènes. In *Terminologie et Intelligence Artificielle, TIA'05*, Rouen, France.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., and Players, F. (2010). Predicting protein structures with a multiplayer online game. *Nature*, 466(7307) :756–760.

Daille, B. (2003). Conceptual structuring through term variations. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment - Volume 18*, MWE '03, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ferret, O. (2010). Similarité sémantique et extraction de synonymes à partir de corpus. In *Actes de TALN 2010 Traitement Automatique des Langues Naturelles-TALN 2010*.

Ganter, B. and Wille, R. (1997). *Formal Concept Analysis : Mathematical Foundations*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edition.

Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.

Harabagiu, S. and Moldovan, D. (1998). Knowledge processing on an extended wordnet. *WordNet-An Electronic Lexical Database*, pages 379–405.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Helbig, H. (2005). *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*, ACL '90, pages 268–275, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

Lafourcade, M. and Joubert, A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, pages 657–666, France.

Lindberg, D., Humphreys, B., and McCray, A. (1993). The unified medical language system. *Methods of Information in Medicine*, 32(4) :281–291.

Malaisé, V., Zweigenbaum, P., and Bachimont, B. (2005). Mining defining contexts to help structuring differential ontologies. *Terminology*, 11(1) :21–53.

Mangeot, M. and Ramisch, C. (2012). A serious game for building a portuguese lexical-semantic network. In *Proceedings of the 3rd Workshop on the People's*

Web Meets NLP : Collaboratively Constructed Semantic Resources and Their Applications to NLP, pages 10–14, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manser, M. (2012). État de l'art sur l'acquisition de relations sémantiques entre termes : contextualisation des relations de synonymie. In *Actes des 14e Rencontres des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 163–175, Grenoble, France. Association pour le Traitement Automatique des Langues.

Marshman, E. and L'Homme, M.-C. (2006). Disambiguating lexical markers of cause and effect using actantial structures and actant classes. In Picht, H., editor, *Modern Approaches to Terminological Theories and Applications*, Linguistic Insights. Studies in Language and Communication. Vol. 36, pages 261–285, Pieterlen. Peter Lang AG.

Miller, G. A. (1995). Wordnet : A lexical database for english. *Commun. ACM*, 38(11) :39–41.

Morin, E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. PhD thesis, Nantes, Grenoble. Th. : informatique.

Ortega-Mendoza, R., Villaseñor-Pineda, L., and Montes-y Gómez, M. (2007). Using lexical patterns for extracting hyponyms from the web. In Gelbukh, A. and Kuri Morales, , editors, *MICAI 2007 : Advances in Artificial Intelligence*, volume 4827 of *Lecture Notes in Computer Science*, pages 904–911. Springer Berlin Heidelberg.

Pennacchiotti, M. and Pantel, P. (2006). Ontologizing semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 793–800, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ravichandran, D. and Hovy, E. (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 41–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Resnik, P. (1992). A class-based approach to lexical discovery. In *In Proc. of the 30th Ann. Meet. of the Assoc. for Computational Linguistics (ACL-92*, pages 327–329.

Rigau, G. (1998). *Automatic Acquisition of Lexical Knowledge from MRDs*. PhD thesis, Dept. LSI, Universitat PolitÚcnica de Catalunya.

Ruge, G. and Schwarz, C. (1991). Term associations and computational linguistics. *International Classification*, 18(1) :19–25.

von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA. ACM.

von Ahn, L., Liu, R., and Blum, M. (2006). Peekaboom : A game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 55–64, New York, NY, USA. ACM.

Winston, M. E., Chaffina, R., and Herrmann, D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11(4) :417–444.

Zweigenbaum, P. and Grabar, N. (2000). Liens morphologiques et structuration de terminologie.