TALN 2010, Montréal, 19–23 juillet 2010

JAWS : Just Another WordNet Subset

Claire Mouton[1, 2] Gaël de Chalendar[1]
(1) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, Fontenay aux Roses, F-92265, France;
(2) Exalead, 10 place de la Madeleine, 75008 Paris
claire.mouton@cea.fr, gael.de-chalendar@cea.fr

Résumé. WordNet, une des ressources lexicales les plus utilisées aujourd'hui a été constituée en anglais et les chercheurs travaillant sur d'autres langues souffrent du manque d'une telle ressource. Malgré les efforts fournis par la communauté française, les différents WordNets produits pour la langue française ne sont toujours pas aussi exhaustifs que le WordNet de Princeton. C'est pourquoi nous proposons une méthode novatrice dans la production de termes nominaux instanciant les différents synsets de WordNet en exploitant les propriétés syntaxiques distributionnelles du vocabulaire français. Nous comparons la ressource que nous obtenons avec WOLF et montrons que notre approche offre une couverture plus large.

Abstract. WordNet, one of the most used lexical resource until today has been made up for the English language and scientists working on other languages suffer from the lack of such a resource. Despite the efforts performed by the French community, the different WordNets produced for the French language are still not as exhaustive as the original Princeton WordNet. We propose a new approach in the way of producing nominal terms filling the synset slots. We use syntactical distributional properties of French vocabulary to determine which of the candidates given by a bilingual dictionary matches the best. We compare the resource we obtain with WOLF and show that our approach provides a much larger coverage.

Mots-clés : ressources lexicales françaises, WordNet, relations sémantiques, distributions syntaxiques.

Keywords: French lexical resources, WordNet, semantic relations, syntactical distributionality.

1 Introduction

La majorité des ressources lexicales ont d'abord été constituées pour l'anglais. Cependant, les différentes communautés non anglophones ont aussi besoin de telles ressources. Nous nous intéressons ici à la constitution d'une version française du réseau lexical WordNet de l'université de Princeton (Fellbaum, 1998). WordNet répertorie les mots du vocabulaire en fonction de leur sens et des relations sémantiques qui lient ces mots entre eux.

Il existe déjà plusieurs tentatives de constitution de WordNet pour le français telles que celles développées dans les travaux de (Vossen, 1998) ou (WOLF, (Sagot & Fišer, 2008)) mais aussi pour d'autres langues comme les travaux de (Barbu & Barbu Mititelu, 2005) par exemple. Le point le plus délicat de ces transferts réside dans la traduction des mots polysémiques, et c'est sur ce point particulier que nous souhaitons proposer une approche originale. L'idée principale est d'exploiter les propriétés des distributions des

contextes syntaxiques des noms dans un grand corpus afin de caractériser les relations sémantiques présentes dans WordNet. L'évaluation de ce type de travaux est difficile puisqu'il n'existe pas par définition de vérité terrain sur laquelle s'appuyer. L'évaluation de notre travail repose d'une part sur une comparaison avec une des précédentes tentatives de constitution d'un WordNet français (WOLF) et d'autre part sur une évaluation manuelle.

2 Travaux précédents

Parmi les méthodes proposées précédemment pour la constitution de nouvelles versions de WordNet, deux grandes tendances se dégagent : les approches par fusion parmi lesquelles se situe l'approche de (Kotis *et al.*, 2006) et les approches par extension comme celle proposée par (Sagot & Fišer, 2008). Les approches par fusion consistent à construire des ontologies indépendamment et de déterminer un mapping avec les WordNet existants *a posteriori*. L'avantage d'une telle approche est que l'on peut s'abstraire de la structure existante de WordNet. Au contraire, les approches par extension font l'hypothèse que la structure du WordNet anglais peut en première approximation être reprise dans la langue cible. Il s'agit alors de traduire les lexèmes de l'anglais vers la langue cible. Nous nous plaçons dans ce cadre.

Beaucoup de travaux utilisent un dictionnaire bilingue pour y sélectionner les traductions les plus pertinentes selon diverses heuristiques, c'est le cas de (Barbu & Barbu Mititelu, 2005). La difficulté majeure d'une telle traduction est le traitement des termes source polysémiques, qui sont associés à plusieurs synsets de WordNet [1]. En effet, les traductions données par un dictionnaire ne correspondent pas forcément à tous les synsets d'un même terme, il s'agit de déterminer la ou les traduction(s) adaptée(s) à chaque synset. Une approche originale de (Sagot & Fišer, 2008) utilise des corpus parallèles pour lesquels ils effectuent la désambiguïsation du corpus anglais à l'aide des synsets de WordNet et proposent les mots alignés de la langue cible comme nouveaux termes. Dans le présent article, nous proposons une méthode un peu différente, qui utilise un dictionnaire bilingue tout en caractérisant les relations sémantiques du réseau lexical par des propriétés syntaxiques distributionnelles.

3 Approche proposée

La structure du WordNet de Princeton (PWN) est tout d'abord reproduite pour la constitution du WordNet de langue cible. Après une phase d'extraction des candidats de traduction, chaque heuristique définie dans la suite de cette section est appliquée de façon itérative, de sorte que le WordNet cible se remplisse petit à petit et qu'à chaque itération, de nouvelles informations viennent rendre possible de nouvelles traductions. Nous ne traitons dans ce travail que des termes et syntagmes nominaux auxquels nous référerons dès lors plus simplement par l'emploi du mot *terme*.

La phase d'extraction consiste à traduire tous les termes associés à un seul synset par toutes les traductions proposées par notre dictionnaire bilingue [2]. Pour les autres termes et chacun de leurs synsets associés, nous conservons toutes les traductions comme termes cible candidats. La désambiguïsation consistera à

1. Rappelons ici que la structure de WordNet distingue les sens des mots par le regroupement en *synsets*. Un synset correspond à un ensemble de synonymes associé à une définition. Certains synsets sont reliés entre eux par des relations sémantiques.

2. Nous utilisons la concaténation du dictionnaire *SCI-FRAN-EuRADic* (`http://catalog.elra.info/product_info.php?products_id=666&language=fr`) et du Wiktionnaire français.

déterminer quel terme candidat (s'il existe) correspond au sens de chaque synset. La structure du PWN est ainsi conservée : l'appelation synset fait maintenant à la fois référence aux termes source et aux termes cible. Cette étape d'extraction sera notée *E* dans les résultats présentés plus loin. On parlera de synset instancié pour référer aux synsets auxquels on a assigné au moins un terme cible. Nous pouvons à présent définir des heuristiques de désambiguïsation qui exploiteront les relations sémantiques de PWN ainsi que des caractéristiques de distribution des termes cible dans les espaces sémantiques. Les espaces sémantiques que nous utilisons sont calculés à partir d'une analyse en dépendances syntaxiques sur un corpus français issu du Web. Les documents furent obtenus après avoir envoyé 600 000 mots d'un dictionnaire comme requêtes sur un moteur de recherche et téléchargé les 100 premiers résultats pour chaque requête. Ces espaces sont décrits plus en détails dans (Grefenstette, 2007) et (Mouton *et al.*, 2009).

La première heuristique, désignée par *S* dans les résultats, exploite une mesure de similarité sémantique dans les espaces sémantiques décrits ci-dessus. Nous utilisons une similarité cosinus caractérisée par l'information mutuelle spécifique (PMI). Cette mesure permet de trouver des relations proches de la synonymie (Turney, 2001). Soit un terme source d'un synset. S'il a plusieurs traductions candidates, et que le synset a déjà été instancié, alors la traduction choisie est celle la plus proche des termes cible instanciés. Par exemple, en français *saw* se traduit par *dicton* ou *scie*. Pour un des synsets de PWN associé à *saw*, les termes cible instanciés précédemment sont *adage*, *proverbe* et *sentence*. La proximité issue des espaces sémantiques indique alors que pour ce synset la meilleure traduction est *dicton*.

On se propose également d'exploiter les relations d'hyponymie et d'hyperonymie pour déterminer quel est le candidat de traduction le plus adapté. Un mot spécifique possédant des caractéristiques plus complètes que son hyperonyme, nous émettons la double hypothèse suivante : (1) les contextes syntaxiques d'un mot général apparaissent souvent comme contexte syntaxique de ses hyponymes (e.g. : *la vitesse du véhicule*, et *la vitesse du train, du bateau, du camion*) et (2) l'éventail des contextes syntaxiques d'un mot spécifique est plus grand que ceux de ses hyperonymes (e.g. : *la quille du bateau* mais pas *la quille du véhicule*). À partir de ces deux hypothèses, on déduit la caractérisation suivante : pour un synset *S* possédant au moins un synset hyponyme instancié h(S) et au moins un synset hyperonyme instancié H(S), on calcule pour chaque terme candidat c le score $\sigma(c)$ suivant :

$$\sigma(c) = \frac{1}{|h(S)|}. \sum_{\{T_{cible} \in h(S)\}} \frac{|ctx(T_{cible}) \cap ctx(c)|}{|ctx(c)|} + \frac{1}{|H(S)|}. \sum_{\{T_{cible} \in H(S)\}} \frac{|ctx(c) \cap ctx(T_{cible})|}{|ctx(T_{cible})|}$$

avec ctx(x) l'ensemble des termes cibles contextes de x. L'hypothèse (2) sert ici à limiter les diviseurs à $|ctx(c)|$ et $|ctx(T_{cible})|$ et non $|ctx(c) \cup ctx(T_{cible})|$. Le candidat de score le plus grand est validé. Nous utilisons cette heuristique distinctement sur les espaces sémantiques de complément du nom, sujet-verbe, et objet-verbe, en l'appelant respectivement *Hc*, *Hs* et *Ho*.

Les relations de méronymie ou d'holonymie (relation *est une partie de*) avec des synsets déjà instanciés peuvent également être exploitées pour déterminer le meilleur candidat cible. Notre hypothèse est qu'un concept compris dans un autre est fortement susceptible d'apparaître dans ses cooccurrents par la relation complément du nom : *la pédale du vélo*, *le toit de l'immeuble*. Pour d'autre langue que le français, la relation peut-être différente (i.e. *bicycle pedal*). Cette heuristique est discutable car certaines prépositions formant la relation de complément du nom ne réalisent que rarement la relation de méronymie dans le sens proposé. De plus, même si la préposition *de* correspond parfois à notre caractérisation, ce n'est pas toujours le cas (*tour du monde, coup de vent...*). L'ensemble des candidats étant restreint par les traductions des termes source, cette heuristique peut néanmoins permettre le choix du bon candidat. Le score d'un candidat est alors la moyenne des scores prenant en compte le nombre d'occurrences de la relation *complément du*

nom entre chaque méronyme (ou holonyme) et le candidat, divisé par le nombre d'occurences du candidat et du méronyme (ou holonyme) en position de complément du nom. Les candidats ayant les plus hauts scores sont conservés pour traduction. Cette heuristique sera notée *M* dans les résultats.

Nous appliquons une dernière heuristique (notée *F*) : la racine étymologique d'un mot pouvant être conservée d'une langue à l'autre, nous validons le meilleur candidat dont la distance de Levenshtein avec le mot source est en dessous d'un certain seuil.

A chaque itération de l'algorithme, on produit autant de nouvelles ressources (ensemble de traductions) que d'heuristiques. Puis, une évaluation automatique [3] est menée pour déterminer quelle heuristique fournit la meilleure ressource en précision. On élimine les autres et on réitère en utilisant la ressource conservée.

4 Évaluation

Afin de valider notre approche et la validité de nos hypothèses, nous nous comparons à la ressource WOLF qui présente l'intérêt d'avoir déjà été évaluée. La version de JAWS évaluée est donc construite à partir de la version du PWN 2.0 utilisée par WOLF. Nous mesurons d'une part la couverture obtenue par WOLF et JAWS en pourcentage du nombre de synsets polysémiques de PWN. D'autre part, nous classons les paires *Terme-Synset* P obtenues dans la ressource cible en trois catégories : dans la catégorie 1, P est présente dans WOLF. Dans la 2, P est absente dans WOLF mais il existe au moins une traduction du synset S et dans la catégorie 3, le synset S ayant produit P n'a pas de traduction dans WOLF. Les résultats sont présentés dans le tableau 1. Pour des raisons de place, seuls les résultats après extraction et ceux issus de la meilleure séquence d'heuristiques sont montrés ici. Nos résultats montrent que l'extraction pure produit moins de traductions que ce que l'on trouve dans WOLF (27 % contre 30 % des synsets de PWN). En revanche, chaque heuristique seule produit un plus grand nombre de traductions que WOLF. La séquence itérative E+FMHc produit 64 % du nombre de synsets polysémiques de PWN avec 13 % des paires présentes dans WOLF (précision des termes nominaux polysémiques de WOLF estimée à 77 % par leurs auteurs). Parmi les paires générées : 42 % sont produites par l'étape E, puis 47 % par F, 2 % par M, et 9 % par Hc.

	$Paires\ traduites$	Cat1. $P \in WOLF$	Cat2. $P \notin WOLF$	Cat3. $S \notin WOLF$
WOLF	30 %			
Extraction	27 %	8 %(31 %)	19 %(70 %)	73 %
E+FMHc	64 %	13 %(38 %)	21 %(62 %)	67 %

TABLE 1 – Pourcentage des paires nominales polysémiques traduites et répartition des paires sur 3 catégories. Entre parenthèses figure le cas où l'on considère uniquement les synsets appartenant à WOLF.

WOLF ne répertoriant pas exhaustivement l'ensemble des paires possibles pour un synset, nous procédons à l'analyse manuelle d'un extrait aléatoire des paires de catégorie 2 et 3. Nous proposons de classer les différences entre WOLF et JAWS selon le tableau 2. Le tableau 3 montre l'analyse manuelle (sur un échantillon de 40 paires) des paires absentes de WOLF mais présentes dans JAWS pour les synsets présents dans WOLF (61 % des paires concernant ces synsets). On constate que pour E+FMHc, 58 % de ces paires sont meilleures ou égales à celles de WOLF ($MP1 + D1 + D2 + D3$).

3. L'évaluation considère l'intersection avec WOLF comme vérité-terrain, cf. section suivante

Cat.2	Manque Partiel dans WOLF : au moins une traduction de S mais pas de traduction de L	MP1	Traduction JAWS correcte
		MP2	Traduction JAWS incorrecte
	Différence de traduction	D1	La traduction de WOLF est incorrecte et celle de JAWS est correcte
		D2	La traduction de WOLF est moins bonne
		D3	Les deux traductions sont correctes et équivalentes
		D4	La traduction de JAWS est moins bonne
		D5	La traduction de JAWS est incorrecte et celle de WOLF est correcte
	Non résolu	W	Aucune traduction n'est adaptée
Cat.3	Absent de WOLF : aucune traduction de S	MT1	Traduction JAWS correcte
		MT2	Traduction JAWS incorrecte

TABLE 2 – Différences par rapport à WOLF pour une paire P associée à un synset S issue d'un terme T

	MP1	MP2	D1	D2	D3	D4	D5	W	MP1+D1+D2+D3
Extraction	20	5	3	0	4	1	6	1	$68\ \% \pm 14$
E+FMHc	(16+4)	(2+9)	(1+0)	(0+2)	0	(0+2)	(1+3)	0	$58\ \% \pm 15$

TABLE 3 – Analyse des paires de catégorie 2 ($P \notin WOLF$) sur un échantillon de 40 paires. La dernière colonne est le pourcentage de précision pour cette catégorie.

Quand aux paires correspondant à la catégorie 3 (synsets absents de WOLF), leur analyse manuelle (sur un nouvel échantillon de 40 paires) montre qu'elles sont bonnes à 73 % pour E+FMHc (tableau 4). Ce dernier tableau indique aussi la micro précision estimée à l'aide de WOLF et des validations manuelles : $\sum_{i\in\{1,2,3\}} Précision(Cat(i)) * Pourcentage(paire \in Cat(i))$. On obtient un WordNet français couvrant deux fois plus de synsets nominaux polysémiques que WOLF pour une perte de précision de 6 points.

	MT1	MT2	$P_{estimée}$
Extraction	$83\ \% \pm 12$	$17\ \% \pm 12$	$80\ \% \pm 8$
E+FMHc	$73\ \% \pm 14$	$27\ \% \pm 14$	$71\ \% \pm 9$

TABLE 4 – Analyse des paires de catégorie 3 ($S \notin WOLF$). La dernière colonne est la probabilité estimée sur l'ensemble des catégories. (Ex : $0.13 * 77 + 0.21 * 58 + 0.67 * 73 = 71$)

5 Résultats et discussions

Après 3 itérations des heuristiques (soit F, M, Hc), nous obtenons la meilleure ressource avec une converture de 64 % du nombre de paires d'origine. La ressource obtenue contient un total de 26 807 termes nominaux uniques, et ceci avec une précision estimée à 71 % pour les termes nominaux polysémiques.

Un des inconvénients de la méthode proposée réside dans l'incapacité du système à ne choisir aucun candidat parmi les traductions proposées. Si le dictionnaire bilingue fournit un certain nombre de candidats mais ne fournit pas de traduction pour un des sens WordNet du terme source, la traduction choisie sera nécessairement fausse. Si le candidat le plus correct ne figure pas dans les entrées de l'espace sémantique (comme les noms propres dans notre cas), la traduction choisie sera nécessairement fausse. La méthode gagnerait donc à fixer quelques critères de non-choix de candidat.

L'heuristique fournissant les meilleurs résultats à la première itération est celle exploitant la distance de

Levenshtein. Ceci peut s'expliquer par le fait qu'un faible nombre de synsets sont instanciés avant la première itération, ceci rendant difficile l'exploitation des autres heuristiques. Par ailleurs, bien que toutes les heuristiques ne soient pas utilisées dans la séquence optimale, on remarque qu'elles produisent chacune des résultats intéressants. Nous souhaitons donc étudier dans des travaux ultérieurs le gain éventuel en précision apporté par une utilisation combinée (et non séquentielle) des différentes heuristiques.

6 Conclusion

Le WordNet français ainsi obtenu couvre deux fois plus de nominaux polysémiques que WOLF, avec une perte de précision estimée à 6 points. L'idéal serait maintenant de pouvoir combiner ces ressources.

La méthode peut être généralisée à d'autres langues, à condition que l'on dispose d'un dictionnaire bilingue riche, d'un analyseur syntaxique, et que la langue cible partage beaucoup de cognats avec la langue source (l'heuristique la plus efficace étant la distance de Levenshtein). Enfin, quelques modifications peuvent être nécessaires pour des langues dans lesquelles la structure de complément du nom ne s'emploierait pas de la même manière.

Ces heuristiques ne sont pas suffisamment robustes pour acquérir les mots et les relations qui les lient sans l'utilisation de la structure de WordNet. Nous projetons d'analyser de façon plus systématique les distributions syntaxiques caractérisant les relations sémantiques en utilisant le PWN et des espaces sémantiques constitués à partir de langue anglaise. S'il existe réellement une telle caractérisation, cette analyse mènera à une caractérisation distributionnelle plus fine et plus exploitable des relations sémantiques.

Références

BARBU E. & BARBU MITITELU V. (2005). Automatic building of Wordnets. In *Proc. of RANLP 2005*, p. 329–332.

C. FELLBAUM, Ed. (1998). *WordNet : An Electronic Lexical Database*. MIT Press.

GREFENSTETTE G. (2007). Conquering language : Using NLP on a massive scale to build high dimensional language models from the Web. In *Proc. of the 8th CICLing Conference*, p. 35–49, Mexico.

KOTIS K., VOUROS G. A. & STERGIOU K. (2006). Towards automatic merging of domain ontologies : The HCONE-merge approach. *Web Semantics : Science, Services and Agents on the World Wide Web*, **4**(1), 60–79.

MOUTON C., PITEL G., DE CHALENDAR G. & VILNAT A. (2009). Word Sense Induction from multiple semantic spaces. In *Proc. of RANLP 2009*, Borovets, Bulgarie.

SAGOT B. & FIŠER D. (2008). Construction d'un WordNet libre du français à partir de ressources multilingues. In *Actes de TALN 2008 (Traitement automatique des langues naturelles)*, Avignon : LIA.

TURNEY P. D. (2001). Mining the web for synonyms : PMI–IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, **2167**, 491–502.

VOSSEN P. (1998). EuroWordNet : A multilingual database with lexical semantic networks. *Computational Linguistics*, **24**(4), 628–630.