Le système WoDiS - WOlf & DIStributions pour la substitution lexicale

Kata Gábor

Le système WoDiS - WOLF & DIStributions pour la substitution lexicale

Kata Gábor
Alpage, INRIA Paris–Rocquencourt & Université Paris 7
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
kata.gabor@inria.fr

Résumé. Le présent article décrit le système WoDiS pour la tâche de substitution lexicale SemDis-TALN 2014. L'algorithme mis en place exploite le WOLF (WordNet Libre du Français) pour générer des candidats de substitution ainsi que pour induire un regroupement des sens fondé sur la structure des synsets. Un espace vectoriel est ensuite créé pour caractériser les différents sens du mot cible à partir de données distributionnelles extraites d'un corpus. Lors de la désambiguïsation, cet espace est confronté au contexte par des méthodes empruntées au domaine de la classification thématique de documents. Pour surmonter le problème de l'insuffisance des données pour les sens peu fréquents, une expansion lexicale est appliquée au niveau des groupes de sens, qui permet de retrouver davantage de contextes caractéristiques et compenser le biais que présentent les vecteurs de mots induits de corpus. Le système a fini quatrième (sur neuf systèmes soumis) dans l'évaluation.

Abstract. In this paper we describe the WoDiS system, as entered in the SemDis-TALN2014 lexical substitution task. Substitution candidates are generated from the WOLF (WordNet Libre du Français) and are clustered according to the structure of the synsets containing them to reflect the different senses of the target word. These senses are represented in a vector space specific to the target word, based on distributional data extracted from a corpus. This vector space is then mapped to the context with simple topical similarity metrics used in document classification. To overcome the data sparseness problem while representing the less frequent senses, we apply a lexical expansion method which allows to extract a higher number of relevant contexts and to compensate for the bias present in corpus-based distributional vectors. Our system ranked fourth in the final evaluation.

Mots-clés : substitution lexicale, désambiguïsation de sens, sémantique distributionnelle, WordNet, WOLF.

Keywords: lexical substitution, word sense disambiguation, distributional semantics, WordNet, WOLF.

1 Introduction

La tâche de substitution lexicale consiste à proposer une ou plusieurs unités de substitution (mots simples ou composés) pour un mot dans une phrase, de manière à conserver le sens global de la phrase autant que possible. Il s'agit d'une adaptation au français de la tâche SemEval 2007 « Lexical Substitution » (McCarthy & Navigli, 2009). Le choix du substitut est libre : contrairement à la tâche de désambiguïsation de sens « standarde », aucun inventaire de sens ou de synonymes n'est proposé au préalable. L'évaluation s'étend ainsi aux ressources lexicales et/ou à la méthode de génération de candidats autant qu'à la désambiguïsation en contexte. L'évaluation des systèmes se fait en les confrontant aux réponses fournies par des annotateurs humains.

Les approches fréquemment utilisées consistent à extraire des candidats à partir d'un inventaire de sens ou de synonymes (typiquement un WordNet) et d'appliquer par la suite une méthode de désambiguïsation pour sélectionner le candidat qui s'adapte le mieux au contexte. C'est le cas notamment de la plupart des systèmes aboutis lors de la tâche de substitution lexicale pour SemEval 2007 (Hassan *et al.*, 2007; Martinez *et al.*, 2007). L'adéquation d'un répertoire de sens prédéfini a toutefois été contestée à plusieurs reprises (Véronis, 2003; Ide & Wilks, 2006) en raison de l'imprécision des distinctions de sens résultant d'une granularité trop fine, et de la rigidité des définitions due au caractère pré-établi de la ressource.
De nombreuses méthodes ont été proposées pour la désambiguïsation en contexte en s'appuyant sur des ressources externes d'information comme des définitions venant de dictionnaires (Lesk, 1986), des distances sémantiques extraites de WordNet (Aguirre & Rigau, 1996), de l'information sur la sélection sémantique (Carroll & McCarthy, 2000; Kohomban & Lee, 2005), où une combinaison de celles-ci (Stevenson & Wilks, 1999). La limitation de ces systèmes réside en leur

dépendance d'une (ou plusieurs) ressources externes avec une couverture finie. Ce problème peut être compensé par l'utilisation des résultats fournis par un moteur de recherche (Martinez *et al.*, 2007) ou un système de traduction automatique (Hassan *et al.*, 2007). D'un autre côté, des algorithmes d'apprentissage supervisé ont été appliqués avec succès à la tâche de désambiguïsation (Cabezas *et al.*, 2001; Lee *et al.*, 2004). Comme les modèles sont appris à partir de corpus annotés sémantiquement, ces systèmes doivent faire face au biais dans la distribution des sens : l'exactitude de la désambiguïsation baisse lorsqu'il y a un écart entre la distribution des sens dans les données d'apprentissage et les données de test (Aguirre & Martinez, 2000). L'exploration de dimensions sémantiques latentes, utilisées d'abord pour la classification de documents, commence récemment à gagner du terrain dans l'induction et la désambiguïsation non-supervisées de sens (Schütze, 1998; Véronis, 2004; Lin & Pantel, 2002; de Cruys & Apidianaki, 2011; de Cruys *et al.*, 2011).

L'algorithme de substitution WoDiS que nous proposons exploite le WOLF, une ressource de type WordNet pour le français, en tant que source primaire de candidats ; lorsque la couverture de celui-ci est insuffisante, il sera complété par une approche distributionnelle. Les candidats à la substitution fournis par le WOLF sont regroupés par sens, où un sens correspond à un ou plusieurs synsets imbriqués contenant le mot cible.
La phase de désambiguïsation est hybride : des connaissances venant de la structure du WordNet sont combinées avec un calcul de compatibilité contextuelle à partir du corpus FrWiki (de la Clergerie, 2010). Au coeur de la méthode est la notion « d'espace de désambiguïsation » spécifique à chaque mot cible. Il s'agit d'un espace vectoriel comprenant l'union des contextes spécifiques à chacun des sens du mot cible. Cet espace est construit à la volée à partir du corpus. Pour assurer une représentation équilibrée et ainsi minimiser le biais dans la distribution des sens, nous procédons à une expansion lexicale en consultants les synsets liés aux sens moins représentés.

Dans ce qui suit, nous allons présenter les ressources utilisées (2) et les détails de l'algorithme et de ses paramètres (3 et 4). La présentation est suivie de l'analyse des résultats (5). A la fin, nous tirons les conclusions et esquissons les perspectives (6).

2 Ressources utilisées

2.1 Le WOLF, inventaire de sens et de synonymes

Le WOLF (WordNet Libre du Français) est une ressource lexicale sémantique libre pour le français, de type WordNet. Cette ressource a été construite à partir du Princeton WordNet (PWN) (Fellbaum, 1998) et de diverses ressources multilingues (Sagot & Fiser, 2008). La méthode utilisée pour créer le WOLF est une méthode par *extension* (Vossen, 1999), suivant laquelle un ensemble de synsets (ensembles de synonymes) du PWN ont été traduits en français.

Dans la première version du WOLF, les traductions françaises des lexèmes monosémiques du PWN ont été notamment extraites à partir de Wikipédia et d'autres ressources wiki. Un corpus parallèle multilingue (JRC-Acquis (Steinberger *et al.*, 2006)) a permis de traiter également les lexèmes polysémiques de la manière suivante. Les informations fournies dans les lexiques obtenus à partir de l'alignement automatique du corpus multilingue en mots ont été combinées aux informations trouvées dans le PWN et dans les WordNets de plusieurs autres langues présentes dans le corpus (WordNets du roumain, du tchèque et du bulgare développés dans le cadre du projet BalkaNet). La désambiguïsation consistait à assigner un identifiant de sens (identifiant de synset) à chaque entrée polysémique du lexique. Pour chaque mot trouvé dans une entrée de ce type, à l'exception du lexème français, l'ensemble des identifiants des synsets auxquels il appartient ont été repérés dans le PWN (version 2.0) et les WordNets des autres langues (alignés sur le PWN 2.0). Ensuite, l'intersection des ensembles d'identifiants de synsets associés aux différents mots de chaque entrée était calculée. Si l'intersection était non vide, les synsets qu'elle contenait étaient attribués au lexème français de l'entrée.

Par la suite, plusieurs méthodes d'extension automatique ont été appliquées au WOLF dans le cadre du projet ANR EDyLex, notamment par induction et désambiguïsation de sens multilingues (Apidianaki & Sagot, 2012), ainsi que par la détection automatique de liens de dérivation (Gábor *et al.*, 2012). D'autres techniques ont été utilisées pour étendre le WOLF de façon massive (Sagot & Fišer, 2012; Hanoka & Sagot, 2012).

La version actuelle du WOLF, telle qu'elle a été utilisée pour la présente tâche, a bénéficié d'une validation manuelle partielle par deux annotateurs natifs. Des méthodes de filtrage automatique (Sagot & Fiser, 2012) ont également été utilisées, suivies d'efforts de validation manuelle des intrus, qui ont été retirés du WOLF. Au total, 4463 synsets ont été validés manuellement de façon partielle (pour certains lexèmes seulement) ou totale, pour un total de 7441 lexèmes

validés.

La ressource résultant de ces travaux contient 32 351 synsets non vides regroupant 38 001 lexèmes distincts et couvrant les quatre catégories principales (noms, verbes, adjectifs, adverbes). Pour comparaison, le WordNet français (Jacquin *et al.*, 2007) développé dans le cadre du projet EUROWORDNET (Vossen, 1999) ne contient que 22 121 synsets nominaux et verbaux. Le WordNet JAWS (Mouton & de Chalendar, 2010) couvre 26 807 lexèmes nominaux. Néanmoins, les synsets non vides du WOLF ne représentent qu'une partie des synsets de PWN, qui contient 115 24 synsets pour 145 627 lexèmes. Un des objectifs du présent travail est d'évaluer la couverture et la précision/fiabilité du WOLF dans le cadre d'une tâche sémantique.

2.2 Le corpus FrWiki

Ce corpus est utilisé par le système WoDiS pour compléter, selon la nécessité, la liste de candidats fournis par le WOLF en générant des candidats par similarité distributionnelle, ainsi que pour ordonner les candidats dans la phase de désambiguïsation en contexte. Le composant distributionnel de l'algorithme exploite les résultats d'analyse syntaxique produits par l'analyseur FRMG-TAG (de la Clergerie, 2010) sur le corpus FrWiki, constitué du Wikipedia français. Ce corpus contient 17.97M de phrases et 178.9M de mots. Il a été choisi pour son caractère encyclopédique, représentatif du domaine général. Plus spécifiquement, nous nous attendons à ce que tous les sens du mot cible y soient représentés, avec une distribution moins biaisée que dans le cas de corpus spécialisés tels que les corpus journalistiques où certains sens peuvent être complètement absents.

avocat_nc	modifieur	politique_adj	400
avocat_nc	modifieur	français_adj	330
homme_nc	et	avocat_nc	224
profession_nc	de	avocat_nc	138
cabinet_nc	de	avocat_nc	131
avocat_nc	modifieur	mûr_adj	1
graine_nc	de	avocat_nc	1
avocat_nc	attribut	arbre_nc	1

TABLE 1 – Exemples de dépendances

Le corpus a été parsé avec l'analyseur FRMG-TAG et les résultats d'analyse (de la Clergerie, 2010) sont fournis sous forme de dépendances (Tableau 1). Un triplet de dépendance tel qu'il est extrait du corpus contient une paire de mots et l'étiquette de la relation qui les relie (p.ex. sujet, objet, modifieur, complément de préposition). Les vecteurs de cooccurrence caractérisant la distribution des mots ont été calculés à partir de ces triplets, c'est-à-dire que l'espace sur lequel la distribution des mots est représentée se limite aux mots du contexte qui entrent dans une relation de dépendance directe avec celui-ci. Cependant, comme nous allons voir dans les sections 3.2 et 4.3, l'algorithme ne requiert pas d'analyse syntaxique et peut être appliqué à un espace vectoriel obtenu à partir d'une représentation « sac de mots ».

3 Génération de candidats de substitution

Notre méthode consiste à extraire des candidats-synonymes par groupes, correspondant aux différents sens du mot cible. Nous exploitons la structure du WOLF, identique à celle du Princeton WordNet. D'un côté, notre objectif est de retrouver tous les sens distincts liés au mot cible pour 1) générer des candidats pour chaque sens et 2) générer une représentation distributionnelle distinctive et caractéristique pour ces sens. De l'autre côté, nous souhaitons éliminer les synsets qui correspondent à des distinctions issues d'une granularité trop fine et qui seraient ainsi trop difficiles à désambiguïser.

3.1 Candidats extraits du WOLF

Dans le WOLF, ainsi que dans le Princeton WordNet, les mots sont regroupés dans des classes de synonymes appelées synsets. Pour obtenir des synonymes, nous avons besoin d'identifier les synsets qui contiennent le mot cible et d'extraire les autres mots présents dans le synset. Le problème de manque de synonymes, rapporté par rapport au PWN (Hassan

et al., 2007) utilisé par la majorité des participants de la tâche de substitution SemEval 2007, nous a également amenés à élargir la recherche aux hyperonymes. A défaut de synonymes dans un synset, nous avons donc extrait les hyperonymes directs. Par exemple, le mot *avocat* qui figure dans les données d'essai n'a pas de synonyme dans le sens « fruit » ; alors que les annotateurs ont recours à une paraphrase (*fruit de l'avocatier*), nous avons extrait l'hyperonyme « fruit ». Notons que selon les instructions SemEval 2007, l'utilisation d'hyperonymes est permise aux annotateurs : « *You may also put a substitute that is close in meaning, even though it doesn't preserve the meaning. In such cases, please aim for a word as close as possible to the meaning of the test word, and preferably one more general than the target word* [1]. »

WordNet est une ressource sémantique caractérisée par une granularité fine : certains synsets proches correspondent à des distinctions mineures et non pertinentes dans le cadre de la présente tâche. La construction du PWN et les autres WordNets suivant son modèle s'adaptent à la tradition lexicographique basée sur une énumération des sens, plus ou moins guidée par l'introspection. Cependant, il a été démontré (Véronis, 2003; Kuti *et al.*, 2010) que ces ressources sémantiques énumératives ne constituent pas un inventaire de sens fiable pour l'étiquetage en sens (dont la présente tâche est proche), d'où l'accord inter-annotateurs faible rapporté pour la discrimination de sens en contexte. Véronis (2003) explique ce problème par le manque d'informations distributionnelles dans les ressources actuellement utilisées, dont les WordNets. Nous sommes ainsi confrontés à des distinctions sémantiques non pertinentes du point de vue de la tâche. Bien que nous ne puissions pas induire la distance sémantique entre des noeuds du même niveau à partir de propriétés structurelles, nous pouvons toujours accéder au contenu lexical des synsets. C'est pour cette raison que nous avons décidé d'unifier les paires de synsets qui contenaient exactement les mêmes éléments lexicaux, ainsi que celles dont le plus petit constituait un sous-ensemble du plus grand. Désormais, les synsets résultant d'une unification seront gérés comme les synsets extraits tels quels ; pour les raisons mentionnées ci-dessus, nous n'accordons pas une présence supérieure aux candidats qui figurent dans plusieurs synsets.
Le tableau 2 montre la proportion des candidats obtenus pour les données d'évaluation après les unifications.

catégorie	# synsets par mot	# candidats par mot	# mots absents dans le WOLF
verbe	7.2	22.7	1
adjectif	1.9	5.5	3
nom	5.9	11.4	1

TABLE 2 – Candidats dans le WOLF

Outre le degré de polysémie du mot cible, les facteurs qui influencent la quantité de synsets et de candidats extraits incluent la granularité hérité du PWN et la couverture du WOLF pour les synsets en question.

3.2 Candidats extraits par similarité distributionnelle

Pour les mots cibles qui n'ont été trouvés dans aucun synset du WOLF, nous avons généré des candidats-synonymes par similarité distributionnelle, calculée selon leur représentation extraite du corpus FrWiki. Un espace vectoriel a été créé pour chaque mot absent du WOLF. Les vecteurs de co-occurrence ont été constitués en prenant les co-occurrences du mot cible avec les lemmes figurant dans son contexte, notamment ceux qui ont une relation de dépendance avec le mot cible. Toutes les relations sont considérées et le type de dépendance ne fait pas partie de la représentation. La méthode s'apprête ainsi à l'utilisation pour un espace « sac de mots » à défaut d'un corpus parsé.

Pour chaque mot candidat *c* et chaque élément du contexte *w*, les vecteurs ont ensuite été pondérés par le poids tf-idf adapté :

$$tf - idf_{c,w} = (tf_{c,w} \times idf_w) \quad (1)$$

où *tf* correspond à la fréquence de co-occurrence de *c* avec *w* observée sur l'ensemble des relations de dépendance, à l'échelle logarithmique [2] :

1. http ://www.informatics.susx.ac.uk/research/nlp/mccarthy/files/instructions.pdf
2. $tf_{c,w} = tf - idf_{c,w} = 0$ si $freq(c, w) = 0$

$$tf_{c,w} = \log freq(c, w) \tag{2}$$

et la mesure *idf* d'un élément de contexte *w* donne la spécificité de celui-ci sur la totalité des relations de dépendance *R* extraits du corpus [3] :

$$idf_w = \log \frac{|R|}{|r \in R : w \in r|} \tag{3}$$

La similarité entre le vecteur du mot cible *x* et ceux des candidats *y* a été calculée par la similarité cosine $sim_{cos}(x, y)$, en prenant en compte leurs co-occurrences pondérées avec les éléments du contexte *w* :

$$sim(x, y) = \frac{x \cdot y}{|x|\,|y|} = \frac{\sum_{w=1}^{n} x_w \times y_w}{\sqrt{\sum_{w=1}^{n} x_w^2} \times \sqrt{\sum_{w=1}^{n} y_w^2}} \tag{4}$$

Pour chaque mot cible, nous avons retenu les dix premiers candidats appartenant à la même catégorie grammaticale [4]. Comme notre approche distributionnelle ne permet pas d'induire un regroupement des sens du mot cible, nous ne savons pas avec quel sens les candidats distributionnels sont mis en correspondance. Par conséquent, nous traitons chaque candidat comme correspondant à un sens distinct. Nous générons ainsi dix classes, c'est-à-dire des pseudo-synsets, avec un candidat par classe.

4 Désambiguïsation en contexte

4.1 Caractérisation des sens - expansion lexicale

Bien que le corpus FrWiki soit une ressource encyclopédique, le tableau 1 indique un biais clair envers les sens dominants, ne fournissant que des exemples sporadiques pour les contextes caractéristiques aux sens moins fréquents. Ceci implique d'une part que nous ne disposons que d'un nombre limité de contextes pour les sens moins fréquents, dont l'apparition ou l'absence dans le corpus reste aléatoire. Une classification non-supervisée des contextes d'apparition du mot cible permet d'induire les différents sens de celui-ci, ainsi que d'associer des contextes spécifiques à chacun de ses sens ; cependant, le biais observé dans la distribution des contextes rend cette classification difficile à réaliser. D'autre part, nous disposons d'une classification de candidats basée sur la structure du WOLF et extraite lors de la génération de candidats (3.1). Nous nous sommes donc concentrés sur cette classification de sens pour identifier les contextes distinctifs associés. Sachant que dans le WOLF, qui est une ressource construite de manière automatique, les sens marginaux sont également moins bien représentés, nous avons eu recours à une expansion lexicale pour peupler davantage ces synsets.

L'objectif de cette expansion lexicale est de pouvoir caractériser chaque candidat-synset par un ensemble de contextes spécifiques et distinctifs : c'est-à-dire des contextes partagés entre les mots appartenant à ce synset ou y étant reliés par une relation sémantique forte. Pour chaque synset marginal ne contenant qu'un seul candidat de substitution, nous avons ainsi extrait les synsets reliés à celui-ci par une des relations suivantes : hyperonymie, « category_domain » ou « mero_part ». Les mots appartenant aux synsets reliés ont été rajoutés au contenu de candidat-synset en question. Ces synsets enrichis permettent de créer un espace vectoriel à partir des contextes distinctifs pour chaque sens de chaque mot cible (4.2). Il est cependant à noter que l'expansion lexicale ne change pas la liste des candidats à la substitution, qui reste limitée aux candidats générés comme décrit dans 3.1 et 3.2.

3. La mesure peut être adaptée à une représentation en sac de mots en remplaçant la spécificité des éléments de contexte sur les relations syntaxiques par leur spécificté sur l'ensemble des mots cibles.

4. Nous avons décidé de limiter le nombre des candidats distributionnels par rapport à la moyenne des candidats extraits du WOLF pour les données d'essai (14.5 par mot cible) : compte tenu du fait que le choix du candidat en fonction du contexte se fera également en s'appuyant sur des critères distributionnels, les candidats distributionnels erronés seront plus difficiles à exclure

4.2 Création de l'espace de désambiguïsation

La méthode conçue pour désambiguïser le mot cible en contexte repose sur l'idée de créer un « espace de désambiguïsation » propre à chaque mot cible, qui permet de calculer une valeur de compatibilité entre les candidats de substitution proposés et le contexte. Cet espace est constitué de l'union des contextes spécifiques aux différents sens du mot cible, sur lequel chaque candidat sera représenté en fonction de ses co-occurrences observées dans le corpus.

L'espace de désambiguïsation est construit de la manière suivante. Pour chaque synset *S* retenu pour le mot cible et enrichi, si besoin, selon ce qui est décrit dans 4.1, nous cherchons dans le corpus les contextes *w* (hors mots grammaticaux) qui lui sont spécifiques selon la formule suivante :

$$spec_{w,S} = \sum_{s \in S} tf - idf_{s,w} \tag{5}$$

Les contextes *w* seront donc ordonnés selon la somme de leurs valeurs *tf-idf* avec les mots *s* liés au synset [5]. Ces contextes peuvent être partagés entre les différents synsets du même mot cible, dans les cas des paires de synsets qui représentent des sens proches. Ceci ne représente pas un inconvénient, puisque notre but est d'ordonner directement les candidats-synonymes (qui peuvent éventuellement appartenir à plusieurs synsets), sans passer par l'étape de désambiguïser entre les synsets. Les expériences menées sur les données d'essai nous ont amenés à fixer en 200 la limite des contextes retenus par synset. L'espace de désambiguïsation du mot cible est créé en prenant l'union des contextes retenus pour chacun de ses sens ; la taille de l'espace est variable en fonction du nombre des sens entre lesquels nous devons désambiguïser.

Par la suite, chaque candidat (indépendamment de son synset d'origine) sera représenté sur cet espace, à la base de ses co-occurrences avec les éléments de contexte constituant l'espace. Trois représentations différentes ont été utilisées sur les données d'essai (tableau 3). La première représentation correspond simplement à la fréquence de co-occurrence du candidat *c* avec les éléments du contexte *w* ; la deuxième, à la fréquence relative ; la troisième est construite à partir de la deuxième, en normalisant les vecteurs par la moyenne et l'écart type pour atténuer le biais vers les candidats plus fréquents.

co-occurrences	$freq(c, w)$
fréquence relative	$\frac{freq(c,w)}{freq(c)}$
fréquence normalisée	$\frac{\frac{freq(c,w)}{freq(c)} - \mu}{\sqrt{\frac{\sigma^2}{N}}}$

TABLE 3 – Représentations des candidats de substitution sur l'espace de désambiguïsation

4.3 Classement des candidats selon le contexte

La phase de désambiguïsation consiste à confronter le contexte de phrase à la représentation vectorielle de chaque candidat. Pour ce faire, nous procédons d'abord à la lemmatisation de la phrase avec l'outil SxPipe (Sagot & Boullier, 2008). Nous créons ensuite un vecteur de phrase sur le même espace vectoriel que nous utilisons pour désambiguïser le mot cible. Le vecteur de phrase *p* est donné par la projection des mots *i* du vecteur de désambiguïsation par une simple fonction caractéristique :

$$p_i = \begin{cases} 1, & \text{si } i \text{ apparaît dans la phrase} \\ 0, & \text{ailleurs} \end{cases}$$

5. La valeur est toujours calculée par rapport au corpus entier.

Notons que les mots grammaticaux, absents des vecteurs de désambiguïsation, ne seront pas pris en compte lors de la désambiguïsation en contexte.
Finalement, pour associer une valeur aux candidats de substitution en fonction du contexte de phrase, nous prenons le produit scalaire du vecteur de désambiguïsation du candidat c avec le vecteur de phrase p :

$$compatibility(c, p) = c \cdot p = \sum_{i=1}^{n} c_i \times p_i \tag{6}$$

Autrement dit, la valeur de compatibilité du candidat est calculée à partir des mots de la phrase faisant partie de l'ensemble des contextes de désambiguïsation du mot cible, avec le poids qui leur est associé par le candidat. Les candidats seront ordonnés par la valeur de compatibilité, et les dix premiers seront retenus pour l'évaluation.

5 Résultats

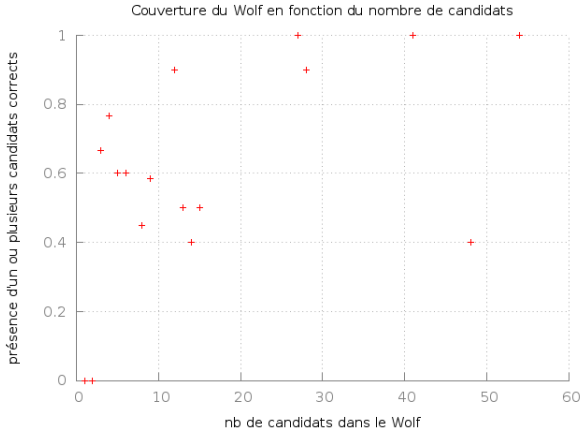
Les trois représentations du tableau 3 ont été appliquées aux données d'essai avec les résultats indiqués par le tableau 4. Nous avons retenu la fréquence relative normalisée, qui a produit les meilleurs résultats sur les données d'essai, pour l'évaluation.

données	type de vecteur	best	oot
test	co-occurrences	0.0402	0.2754
test	fréquence relative	0.0545	0.2600
test	fréquence normalisée	0.0601	0.2573
éval	fréquence normalisée	0.0626	0.2048

TABLE 4 – Résultats

Comme le système WoDiS utilise un nombre limité de candidats à la substitution, la mesure oot peut être interprétée en tant qu'indicateur de l'adéquation relative du WOLF comme source de candidats. Les résultats de l'évaluation suggèrent que le dictionnaire Dicosyn, qui sert de baseline, est plus adapté à la tâche : nous observons une valeur oot de 0.2048 pour WoDiS/WOLF contre une valeur de 0.3245 pour Dicosyn, qui s'est d'ailleurs montré meilleur que la plupart des systèmes en compétition en termes de mesure oot. La couverture du WOLF paraît donc encore limitée pour cette tâche. Il est intéressant de noter que le nombre des candidats trouvés dans le WOLF n'augmente pas automatiquement la probabilité d'y retrouver le bon candidat pour un contexte donné (figure 1). Ceci est certainement dû au fait que le WOLF a été rempli de manière automatique et non pas de manière exhaustive, à partir d'une liste de lemmes fréquents.

FIGURE 1 – Contextes couverts par le WOLF en fonction du nombre des candidats



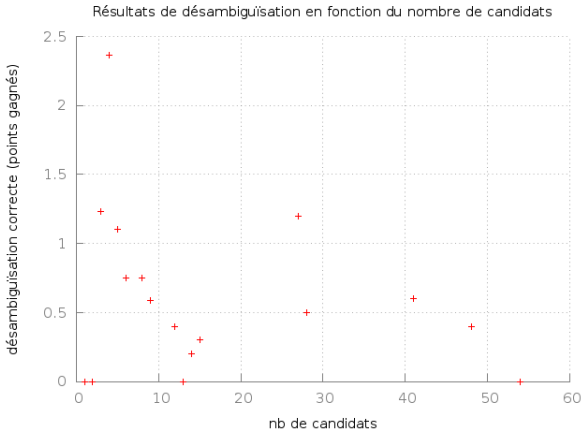
Si nous nous limitons à l'évaluation des candidats distributionnels (10 candidats par mot cible pour 5 mots : notamment

le verbe *taper*, le nom *montée* et les adjectifs *vaseux, hermétique* et *incorrect*), nous voyons que la mesure oot monte à 0.2461, nous produisons donc davantage de bons candidats à partir du corpus qu'à partir du WOLF. La qualité de ces candidats distributionnels reste cependant variable. Une des problématiques bien connues concernant la mise en relation de lemmes par la similarité de leurs contextes est que cette méthode ne permet pas de distinguer la synonymie des autres types de relations comme l'hyperonymie, l'antonymie ou une simple similarité thématique. Par exemple, les candidats distributionnels proposés pour le mot *incorrect* incluent *correct, exact, précis* et *approprié*. De l'autre côté, nous retrouvons les sens bien distincts du mot *taper* dans les candidats distributionnels *frapper, saisir, écrire* et *recevoir* (pour *se taper*).

Si nous nous concentrons sur la mesure best - plus informative que la mesure oot étant donné la quantité limitée de candidats - l'analyse des erreurs nous révèle que 51% des mauvaises substitutions (les cas où la meilleure proposition du système ne figure pas parmi les propositions des annotateurs) sont dues à l'absence d'un bon candidat, alors que dans 49% des cas, la désambiguïsation est erronée. Une évaluation effectuée uniquement sur les 180 phrases pour lesquelles nous avons pu extraire au moins un bon candidat du WOLF donne une mesure oot de 0.3342, soit proche du baseline Dicosyn, alors que la précision de désambiguïsation monte jusqu'à une valeur best de 0.1031, comparable au meilleur système.
Nous constatons également que le WOLF, malgré sa couverture limitée, s'apprête mieux à la tâche de désambiguïsation. Si nous comparons la performance en termes de la mesure best, nous observons une dégradation sur les mots pour lesquels nous n'avons que des candidats distributionnels (0.0520).

Notons également que l'algorithme du système WoDiS n'utilise pas de dimensions latentes : les valeurs de compatibilité sont estimées directement à partir de co-occurrences observées dans le corpus. Il peut arriver qu'aucun mot du contexte de phrase ne figure parmi les contextes de désambiguïsation retenus ; dans ce cas, chaque candidat aura une valeur de compatibilité de 0 et ils seront ordonnés de manière aléatoire. Il nous semblait donc judicieux de vérifier l'impact que peut avoir le manque d'information sur les résultats. Nous avons trouvé que le nombre total des décisions non informées lors du choix de candidat est de 29 (9.6%). Cependant, dans 5.6% des cas - soit la majorité des cas de manque d'information - aucun des candidats extraits n'est correct, ce qui explique l'impossibilité de la mise en relation avec le contexte de phrase.

FIGURE 2 – Bonnes substitutions selon le nombre de candidats



Comme nous pouvons remarquer sur le tableau 2, la quantité des candidats varie fortement en fonction de la catégorie du mot cible. Il est évident que la désambiguïsation devient plus difficile avec l'augmentation du nombre des candidats : la figure 2 illustre la dégradation des résultats en fonction du nombre des candidats, pour montrer une légère remontée pour les mots avec une très grande quantité de candidats, pour lesquels le problème de l'absence d'un bon candidat ne se présente plus. Sur l'ensemble des tâches de génération de candidats et de désambiguïsation, les meilleurs résultats sont obtenus pour les mots cible avec 3-6 candidats. Ceci peut expliquer que le résultat du système sur les adjectifs est significativement meilleur que sur les autres catégories.

6 Conclusion et perspectives

Nous avons présenté le système de substitution lexicale WoDiS. La tâche de substitution est accomplie en deux étapes. Les candidats à la substitution sont extraits à partir du WOLF ou, à défaut, à partir du corpus FrWiki par similarité distributionnelle. La méthode de désambiguïsation consiste à créer un espace vectoriel sur lequel chaque candidat sera représenté. La confrontation de cet espace aux mots du contexte nous permet d'ordonner les candidats selon leur compatibilité avec la phrase.
La méthode proposée s'appuie sur la structure du WOLF lors de la construction de l'espace de désambiguïsation. L'évaluation a permis de constater que la couverture de cette ressource est relativement limitée pour la tâche, puisque nous trouvons davantage de candidats corrects proposés par la méthode distributionnelle qu'en consultant le WOLF. Cependant, la structure du WOLF peut être exploitée pour obtenir davantage d'informations sur les différents sens du mot cible, et par conséquent, il permet d'aboutir à une meilleure désambiguïsation.
La méthode proposée est rapide et ne nécessite ni de données annotées, ni une analyse linguistique profonde. Bien que nous nous soyons servis des relations de dépendance extraites d'un corpus avec une analyse syntaxique, l'algorithme peut être également utilisé avec une représentation en sac de mots. Le problème de l'insuffisance des données est adressé par une expansion lexicale au niveau des groupes de candidats.

Les limitations connues du système WoDiS portent d'une part sur sa forte dépendance sur l'inventaire de synonymes utilisé, d'autre part sur le problème de l'insuffisance éventuelle des données contextuelles qui permettent d'ordonner les candidats. Par conséquent, les améliorations envisagées incluent l'utilisation d'une expansion lexicale pour les données du contexte. Une meilleure combinaison des candidats distributionnels avec les candidats proposés par le WOLF devrait également permettre d'augmenter la précision.

7 Remerciements

Je remercie Eric de la Clergerie d'avoir mis le corpus analysé à ma disposition, et Benoît Sagot pour son aide dans l'extraction des relations du WOLF et dans l'évaluation.

Références

AGUIRRE E. & MARTINEZ D. (2000). Exploring automatic word sense disambiguation with decision lists and the web. In *Procedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content*.

AGUIRRE E. & RIGAU G. (1996). Word sense disambiguation using conceptual density. In *Proceedings of COLING'96*, p. 16–22.

APIDIANAKI M. & SAGOT B. (2012). Applying cross-lingual wsd to wordnet development. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, p. 833–840 : European Language Resources Association (ELRA).

CABEZAS C., RESNIK P. & STEVENS J. (2001). Supervised sense tagging using support vector machines. In *In Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2*, p. 59–62.

CARROLL J. & MCCARTHY D. (2000). Word sense disambiguation using automatically acquired verbal preferences. *Computers and the Humanities*, **34**, 109–114.

DE CRUYS T. V. & APIDIANAKI M. (2011). Latent semantic word sense induction and disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* : The Association for Computer Linguistics.

DE CRUYS T. V., POIBEAU T. & KORHONEN A. (2011). Latent vector weighting for word meaning in context. In *Proceedings of the EMNLP 2011 Conference*, p. 1012–1022 : ACL.

DE LA CLERGERIE E. (2010). Convertir des dérivations TAG en dépendances. In *Actes de TALN'10 17e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2010)*, Montreal, Canada.

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

GÁBOR K., APIDIANAKI M., SAGOT B. & DE LA CLERGERIE E. (2012). Boosting the coverage of a semantic lexicon by automatically extracted event nominalizations. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, p. 1466–1473 : European Language Resources Association (ELRA).

HANOKA V. & SAGOT B. (2012). Wordnet creation and extension made simple : A multilingual lexicon-based approach using wiki resources. In *LREC 2012 : 8th international conference on Language Resources and Evaluation*, Istanbul, Turquie.

HASSAN S., CSOMAI A., BANEA C., SINHA R. & MIHALCEA R. (2007). Unt : Subfinder : Combining knowledge sources for automatic lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic : Association for Computational Linguistics.

IDE N. & WILKS Y. (2006). Making sense about sense. In *Word Sense Disambiguation : Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*, p. 47–74. Dordrecht, The Netherlands : Springer.

JACQUIN C., DESMONTILS E. & MONCEAUX L. (2007). French eurowordnet lexical database improvements. In *Proceedings of the CICLing Conference*, volume 4394 of *Lecture Notes in Computer Science*, p. 12–22 : Springer.

KOHOMBAN U. S. & LEE W. S. (2005). Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, ACL 2005*.

KUTI J., HÉJA E. & SASS B. (2010). Sense disambiguation - ambiguous sensation ? evaluating sense inventories for verbal wsd in hungarian. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)* : European Language Resources Association (ELRA).

LEE Y. K., NG H. T. & CHIA T. K. (2004). Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Senseval-3 : Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, p. 137–140.

LESK M. (1986). Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from a ice cream cone. In *Proceedings of SIGDOC-1986*.

LIN D. & PANTEL P. (2002). Concept discovery from text. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*.

MARTINEZ D., KIM S. N. & BALDWIN T. (2007). Melb-mkb : Lexical substitution system based on relatives in context. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic : Association for Computational Linguistics.

MCCARTHY D. & NAVIGLI R. (2009). The english lexical substitution task. *Language Resources and Evaluation*, **43**(2), 139–159.

MOUTON C. & DE CHALENDAR G. (2010). JAWS : Just another WordNet subset. In ATALA, Ed., *Actes de TALN 2010*, Montréal, Canada.

SAGOT B. & BOULLIER P. (2008). SxPipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts. *Traitement Automatique des Langues*, **49**(2), 155–188.

SAGOT B. & FISER D. (2008). Building a free french wordnet from multilingual resources. In *Ontolex 2008*, Marrakech, Maroc.

SAGOT B. & FIŠER D. (2012). Automatic Extension of WOLF. In *GWC2012 - 6th International Global Wordnet Conference*, Matsue, Japon.

SAGOT B. & FISER D. (2012). Cleaning noisy wordnets. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, p. 3468–3472 : European Language Resources Association (ELRA).

SCHÜTZE H. (1998). Automatic word sense discrimination. *Computational Linguistics*, **24**(1), 97–124.

STEINBERGER R., POULIQUEN B., WIDIGER A., IGNAT C., ERJAVEC T., TUFIS D. & VARGA D. (2006). The jrc-acquis : A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*.

STEVENSON M. & WILKS Y. (1999). Combining weak knowledge sources for sense disambiguation. In *Proceedings of the International Joint Conference for Artificial Intelligence (IJCAI-99)*.

VÉRONIS J. (2003). Sense tagging : does it make sense ? In *Corpus Linguistics by the Lune : a festschrift for Geoffrey Leech*. Frankfurt : Peter Lang.

VÉRONIS J. (2004). Hyperlex : lexical cartography for information retrieval. *Computer Speech & Language*, **18**(3), 223–252.

P. VOSSEN, Ed. (1999). *EuroWordNet : a multilingual database with lexical semantic networks for European languages*. Dordrecht : Kluwer.