Contextes multilingues alignés pour la désambiguïsation sémantique : une étude expérimentale

Boxing Chen, Meriam Haddara, Olivier Kraif (1)
Grégoire Moreau de Montcheuil, Marc El-Bèze (2)

(1) LIDILEM, Université Stendhal Grenoble 3
Meriam.Haddara@laposte.net, Olivier.Kraif@u-grenoble3.fr

(2) LIA – CNRS, Université d'Avignon et des Pays du Vaucluse
{Gregoire.Moreau-de-Montcheuil, Marc.El-Beze}@univ-avignon.fr

Mots-clés : Désambiguïsation sémantique, alignement multilingue, lexique sémantique
Keywords: Word sense disambiguation, multilingual aligning, semantic lexicon

Résumé Cet article s'intéresse à la désambiguïsation sémantique d'unités lexicales alignées à travers un corpus multilingue. Nous appliquons une méthode automatique non supervisée basée sur la comparaison de réseaux sémantiques, et nous dégageons un critère permettant de déterminer *a priori* si 2 unités alignées ont une chance de se désambiguïser mutuellement. Enfin, nous développons une méthode fondée sur un apprentissage à partir de contextes bilingues. En appliquant ce critère afin de déterminer pour quelles unités l'information traductionnelle doit être prise en compte, nous obtenons une amélioration des résultats.

Abstract This paper addresses the sense disambiguation of aligned words through a multilingual corpus. We apply an unsupervised disambiguation method using inter word net comparison. We study a criterion that allows to identify the cases for which disambiguation can take advantage of alignment. Finally, we implement a method based on a training stage using both monolingual and bilingual context, and we apply the previous criterion in order to select between monolingual or bilingual clues, showing some improvement of the results.

1 Introduction

Les corpus multilingues alignés présentent un intérêt particulier pour la désambiguïsation sémantique au niveau lexical. Dans la mesure où les distributions sémantiques des unités lexicales sont différentes d'une langue à l'autre, les unités alignées à travers des textes parallèles peuvent jouer un rôle réciproque de révélateur sémantique (Ide *et al.*, 2002). Dès les premiers développements des corpus multilingues alignés, il y a une quinzaine d'années, des méthodes de désambiguïsation multilingue ont été proposées (Gale *et al.* 1992, Dagan & Itai, 1994). Le principe est simple : les résultats obtenus grâce aux méthodes de désambiguïsation monolingue, qui se basent en général sur un examen du contexte proche du mot cible, devraient *a priori* être améliorés par l'apport d'information pertinente consistant en la prise en

compte des unités et contextes alignés. Fondé sur cette hypothèse de travail, le projet Carmel, financé par le réseau RIAM[1] vise au développement d'un corpus multilingue aligné concernant quatre langues européennes : l'anglais, l'espagnol, le français et l'italien. Outre son intérêt culturel (récits de voyage s'échelonnant de la fin du XIX^{e} au début du XX^{e} siècle), ce corpus quadrilingue doit nous permettre de coupler trois types de traitement : désambiguïsation sémantique, identification thématique et alignement. Nous décrivons dans un premier temps les résultats de quelques expériences de désambiguïsation non supervisée. Les sections 3 et 4 sont consacrées à la description d'une méthode avec apprentissage, basée sur un mélange d'indices monolingues et bilingues où nous montrons comment la triangulation des méthodes et des langues permet d'améliorer et de consolider les résultats.

2 Désambiguïsation multilingue non supervisée

Un des problèmes posés par les méthodes classiques de désambiguïsation sémantique est la nécessité de passer par une phase d'apprentissage, nécessitant le recours à des corpus étiquetés manuellement dont la production est longue et coûteuse. Récemment, des travaux originaux (Diab & Resnik, 2002, Tufis *et al.*, 2004) ont montré qu'il était possible d'utiliser des corpus alignés pour effectuer une désambiguïsation *non supervisée*. Comme Tufis *et al.* (2004), nous avons utilisé 2 réseaux sémantiques (les lexiques français et anglais livrés avec EuroWordNet), afin de comparer les unités par le biais d'index interlingue (ILI), qui permettent d'établir des équivalences de sens entre des unités. La désambiguïsation de 2 unités U_s et U_c est alors effectuée en cherchant le couple de sens (s_s,s_c) qui maximise une certaine mesure de similarité $Sim(s_s,s_c)$: $D(U_s,U_c) = \text{argmax}\_\{(s_s,s_c) \in S_s \times S_c\}\ Sim(s_s,s_c)$
$Sim(s_s,s_c)$ peut être calculée à partir du nombre de liens séparant chacun des ILI de leur plus proche parent commun dans la hiérarchie. Dans l'expérience décrite ci-après, afin de privilégier la précision (au détriment du rappel) et de s'approcher de la désambiguïsation manuelle effectuée précédemment, nous avons utilisé une définition maximaliste de la similarité, basée sur l'identité des ILI.

La première étape consiste à aligner automatiquement les textes. À partir des alignements phrastiques, nous avons procédé à l'extraction des correspondances lexicales, en utilisant une combinaison d'indices : fréquence des occurrences et des cooccurrences au sein des phrases alignées, positions dans les phrases, ressemblance graphique, identité des parties du discours. Ces indices, combinés de façon appropriée, nous ont permis d'obtenir une F-mesure voisine de 90% sur un corpus étiqueté (*Madame Bovary*, de Flaubert, et sa traduction en anglais) pour l'appariement des mots pleins (Kraif & Chen, 2004). Le corpus étudié est *The voyage of the Beagle* de Darwin, comportant environ 200 000 mots dans chaque langue, préalablement segmenté et étiqueté avec les parties du discours. Pour qu'un couple d'unités soit partiellement ou complètement désambiguïsé, il faut que les 2 unités alignées apparaissent chacune dans leurs réseaux respectifs. Seulement 21 133 couples ont satisfait cette condition.

Les couples totalement désambiguïsés représentent environ 4,3% de la totalité des mots, et 42% des couples pour lesquels les 2 réseaux n'étaient pas silencieux. Pour estimer la précision

[1] Le réseau RIAM, créé en 2001, dépend du Centre national de la cinématographie. Les partenaires du Projet sont l'association ACCE, la société SINEQUA et les laboratoires LIA (Avignon) et LIDILEM (Grenoble).

des résultats, nous avons soumis à l'évaluation manuelle d'un seul annotateur 100 couples totalement désambiguïsés, prélevés aléatoirement. Seul l'anglais a pour le moment été évalué. Ces résultats doivent être pris avec précaution du fait de l'incomplétude et du déséquilibre des réseaux (le réseau français contient 22 745 sens contre 91 600 pour Wordnet 1.5). Notons néanmoins que la précision est plutôt bonne pour les couples totalement désambiguïsés.

	Anglais	Français
Proportion moyenne de sens éliminés	63 %	46 %
Unités totalement désambiguïsées	34,6 % (7 316 / 21 133)	22,7 % (4 804 / 21 133)
Précision estimée	79 %	

Tableau 1 : Réduction des sens pour la désambiguïsation automatique

Nos observations manuelles indiquent que la désambiguïsation multilingue fonctionne très bien pour certains couples d'unités, pas du tout pour d'autres (quand les différentes acceptions sont trop similaires). Pour départager automatiquement les cas favorables des autres, on peut se tourner vers une troisième langue, qui fournira vraisemblablement des indices quant à cette similarité. Par exemple, pour (en-*scarcely*, fr-*presque*) si l'on considère les unités espagnoles alignées avec chaque mot du couple, indépendamment, dans notre corpus, on trouve :

en-*scarcely* -> es-*tampoco* es-*asegurar* es-*casi* es-*apenas* fr-*presque* -> es-*casi*

Un simple filtrage des fréquences nous permet d'éliminer les alignements erronés tel que es-*asegurar*. On constate alors que les 3 sens de *scarcely* indiqués par le dictionnaire, que l'on pourrait gloser par /presque pas/, /difficilement/, /à peine (sens temporel)/, se manifestent par des équivalents espagnols plus variés. La différence sémantique entre *presque* et *scarcely*, qui aboutit en fait à une désambiguïsation correcte de l'anglais, est donc rendue manifeste par cette projection dans une tierce langue. Pour prédire s'il est judicieux ou non d'employer, pour un couple donné, la désambiguïsation multilingue, nous proposons de recourir à un critère numérique, comme l'indice de DICE : $s = \frac{2 \cdot |ES(e) \cap ES(f)|}{|ES(e)| + |ES(f)|}$, où *ES(e)* et *ES(f)* représentent les ensembles d'équivalents espagnols dérivés de l'alignement pour les unités *e* et *f*. Calculé sur un corpus trilingue suffisamment important, *s* peut être un bon indicateur de la similarité sémantique de 2 unités : une valeur faible devrait indiquer de meilleures chances de désambiguïsation multilingue. Cette hypothèse semble confirmée par une certaine corrélation entre la similarité *s* obtenue par projection sur l'espagnol et la proportion de sens éliminés.

% sens éliminés	$0 \leq s < 0,25$	$0,25 \leq s < 0,5$	$0,5 \leq s < 0,75$	$0,75 \leq s \leq 1$
Anglais	75 %	65 %	62 %	60 %
Français	60 %	49 %	43 %	40 %

Tableau 2 : Corrélation entre *s* et la proportion des sens éliminés

3 Méthode de désambiguïsation supervisée

Puisque l'alignement des unités fournit, dans certains cas, une information exploitable pour la désambiguïsation, il paraît naturel d'intégrer cette information dans une méthode "classique", basée sur l'observation des contextes (de Loupy, 2000). Dans cette nouvelle tâche, nous disposons d'un ensemble d'exemples d'apprentissage, noté *Tr*, dont chaque élément représente un (ou plusieurs) sens pour un mot ambigu ; et nous recherchons pour un contexte particulier, noté C^0, les sens correspondants. La méthode employée est décrite en section 3.2.

3.1 Prétraitements

Chaque mot du contexte est lemmatisé, à l'exception du mot à désambiguïser, puis une série de réductions est effectuée : remplacement des nombres par CD, des mois par MONTH et des jours de la semaine par DAY, filtrage pour ne garder que les substantifs, adjectifs, verbes, adverbes, noms propres, prépositions et nombres. Parmi les différentes possibilités de contextes alignés, nous avons choisi de travailler sur la traduction *mot à mot* du contexte source : les termes alignés avec chacun des mots du contexte d'origine. Les contextes projetés (cf. Tableau 3) conservent l'ordre des mots du voisinage de la langue source. L'alignement vers une langue est total, même si quelques mots n'ont pas d'image.

En	strew:Verb	with:Prep	fine:Adj	sand:Noun	street:Noun	3
Pfr	poudrer:Verb	de:Prep	fin:Adj	sable:Noun	Rue:Noun	3

Tableau 3 - Contexte "projeté"

3.2 Algorithme des K plus proches voisins (KPPV)

L'algorithme des KPPV consiste à rechercher parmi les données d'apprentissage les *k* exemples qui ressemblent le plus au contexte C^0. Nous utilisons pour cela une mesure de similarité entre 2 contextes, *Simil(C,C')* qui permet de classer les différents exemples d'apprentissage et de ne retenir que les *k* plus proches (ensemble *KNN*) Chacun des exemples de l'ensemble des *KNN* vote pour le (ou les) sens qu'il caractérise, proportionnellement à sa similarité avec C^0, ce qui donne pour chaque sens *s*, le score : $Sc_{KPPV}(s) = p(s) + \sum_{C \in KNN \cap Tr(s)} Simil(C, C^0)$, où *p(s)* est la probabilité du sens *s* et *Tr(s)* est l'ensemble des exemples d'apprentissage pour le sens *s*. Pour comparer 2 contextes dans une même langue, nous avons décidé de comptabiliser le nombre de termes identiques placés à la même position (par rapport au mot ambigu) ou légèrement décalés (pour prendre en compte des phénomènes de décalage comme l'incise d'un adverbe). Si on note $C=(w_i)_{-g \le i \le +d}$ et $C'=(w'_i)_{-g' \le i \le +d'}$ deux contextes « centrés » autour d'un lemme ambigu (w_0 et w'_0 sont les 2 occurrences de ce lemme), la formule de calcul de similarité entre *C* et *C'* s'écrit :

$$Simil(C,C') = \left(\sum_{i=-G}^{+D} f_i \right)^2 \text{ avec } f_i = \begin{cases} 1 \text{ , si } w_i \equiv w'_i \\ ½ \text{, si } w_i \neq w'_i \wedge (w_i \equiv w'_{i-1} \vee w_i \equiv w'_{i+1}) \\ 0 \text{ , sinon} \end{cases} \text{ et } \begin{matrix} G = \min(g,g') \\ D = \min(d,d') \end{matrix} .$$

Enfin, pour comparer 2 contextes multilingues, nous avons choisi d'additionner les similarités monolingues (norme 1). Soit, si les contextes sont $C=\{C_l\}_{l \in Lang}$ et $C'=\{C'_l\}_{l \in Lang}$, la similarité multilingue est : $Simil_{Mu}(C,C') = \sum_{l \in Lang} Simil(C_l, C'_l)$.

4 Expériences et stratégie de décision

Dans cette série d'expériences, pour 15 lemmes anglais (6 verbes, 5 substantifs et 4 adjectifs), nous disposions de 2 886 contextes désambiguïsés manuellement extraits de 14 récits de voyage d'auteurs anglophones et francophones du XIX° siècle. Par un tirage au sort respectant la distribution des sens, nous en avons écarté environ 80% pour l'apprentissage, ce qui fait un

jeu de 659 tests. Dans l'apprentissage comme dans le test, pour plus de 95% des contextes nous disposons de l'alignement anglais-français.

Lemme	#App	#Test	en	%	en-Pfr	%	#Couple	%	Mel.Fin(0,7)	%	Opt.Fin	%
Total	2227	659	410	62,2	411	62,4	229	35	421	63,9	435	66,0

Tableau 4 - Résultats monolingues, bilingues et avec sélection fine

La première observation qu'il convient de faire, sur les expériences monolingues, est que le résultat global (62,2%) cache une disparité de comportement entre les 15 termes retenus. Il n'est pas étonnant, au regard de la littérature, de voir les noms (maximum atteint par *lady* à 87,9%) à un meilleur niveau de performance que les verbes (minimum atteint par *strike* à 38,3%). Le nombre de sens (21) de *strike* rapporté à une taille de corpus d'apprentissage assez faible (147 exemples) explique en partie le mauvais résultat obtenu par ce verbe.

Comme le montre la colonne *en-Pfr* du tableau 4, le recours systématique au français entraîne une toute petite amélioration des performances au niveau global (1 test). En fait, il n'est pas possible de faire état d'un apport quelconque à mettre au crédit de l'alignement, la différence n'étant pas suffisante pour pouvoir statuer quoi que ce soit. En outre, un examen détaillé des résultats montre que, pour certains mots, non seulement il n'y a pas de gain, mais il y a des pertes. C'est le cas de *carry, child*, *curious, nature, strike* et *use*. Ce comportement hétérogène ouvre la perspective d'une amélioration plus franche.

Il est naturel d'imaginer une sélection permettant de décider, en fonction de la traduction du mot à étiqueter, s'il convient ou non de recourir à l'alignement. Dans cette expérience nous avons fait une sélection fine, test par test, en fonction du coefficient de Dice du couple anglais-français. Cependant, cela n'est possible que dans environ un tiers des cas (colonne *#Couple* du Tableau 4). Dans les autres cas, nous utilisons la moyenne pondérée des coefficients de Dice pour les différents couples du mot. Dans l'idéal, un grain fin aurait pu permettre d'atteindre un résultat de 66%. Force est de constater que nous en sommes loin avec 63,9%. Ceci s'explique par le fait que la décision fine n'est prise qu'une fois sur trois. Cela constitue cependant une progression par rapport à l'usage systématique de l'alignement et laisse la porte ouverte à de nouvelles améliorations.

5 Conclusion et perspectives

Une étude manuelle a permis de dégager le potentiel important du recours aux unités alignées pour la désambiguïsation lexicale. Ainsi, environ 30% des noms ont pu être désambiguïsés totalement, sans que l'annotateur ne détecte de sens incorrect par rapport au contexte. Pour exploiter au mieux cette information par des méthodes automatisées, nous avons étudié 2 approches différentes. La première, basée sur la comparaison de lexiques sémantiques, a confirmé qu'il était possible d'obtenir une désambiguïsation complète avec une très bonne précision, mais pour un nombre faible d'unités (dépendant de la complétude des lexiques en question) ; la deuxième consistait à intégrer le contexte aligné comme si c'était un élément du contexte monolingue, et à appliquer une méthode de WSD supervisée basée sur les KPPV. De cette manière, nous avons observé que les contextes alignés jouaient un rôle parfois positif, parfois négatif, les traductions inopérantes semblant dans certains cas bruiter des indices

monolingues plus pertinents. Sur la base de nos observations préliminaires, nous avons dégagé un critère permettant d'estimer grossièrement la pertinence du contexte aligné : pour 2 unités, on peut comparer les équivalents proposés par l'alignement avec une tierce langue. Ainsi, 2 unités aux différents sens voisins, *a priori* peu intéressantes pour la désambiguïsation mutuelle, devraient avoir des équivalents similaires dans la langue tierce. Ce critère, appliqué au choix, selon les unités, entre WSD monolingue stricte et WSD mono et bilingue, a permis d'obtenir une légère amélioration des résultats. Nous pensons que ce critère de triangulation, s'il implique plus que 3 langues, et un plus grand volume de textes alignés peut permettre un gain important sur le plan de la précision, et ouvrir la voie à la constitution automatique de corpus étiqueté de bonne qualité, pouvant servir de corpus d'apprentissage peu coûteux, pour des méthodes monolingues sur des corpus non-alignés.

Remerciements

Nous remercions le réseau RIAM, qui finance le projet Carmel ainsi que nos partenaires ACCE et la société SINEQUA.

Références

DAGAN I., ITAI A. (1994), Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics*, 20(4):563-596.

DE LOUPY CL. (2000), Évaluation de l'Apport de Connaissances Linguistiques en Désambiguïsation Sémantique et Recherche Documentaire. Thèse de Doctorat, Université d'Avignon et des Pays du Vaucluse.

DIAB M., RESNIK P. (2002), An Unsupervised Method for Word Sense Tagging using Parallel Corpora, *Proc. of 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL-02), Philadelphia.

DIAB M. (2003), Word Sense Disambiguation within a Multilingual Framework. Ph.D. thesis,University of Maryland.

GALE W. A., CHURCH K. W., YAROWSKY D. (1992). Using Bilingual Materials to Develop Word Sense Disambiguation Methods. In *Proc.of the Fourth Inter-national Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101-112, Montreal.

IDE N., ERJAVEC T., TUFIS D. (2002). Sense Discrimination with Parallel Corpora. *Proceedings of ACL'02 Workshop on Word Sense Disambiguation*: *Recent Successes and Future Directions*, Philadelphia, 54-60.

KRAIF O., BOXING, CHEN (2004) Combining clues for lexical level aligning using the Null hypothesis approach, in *Proceedings of Coling 2004*, Geneva, August 2004, pp. 1261-1264.

TUFIS D., ION R., IDE N. (2004), Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING2004*, Geneva.