Création rapide et efficace d'un système de désambiguïsation lexicale pour une langue peu dotée

Mohammad Nasiruddin, Andon Tchechmedjiev, Hervé Blanchon, Didier Schwab
LIG-GETALP
Univ. Grenoble Alpes
prenom.nom@imag.fr
http://getalp.imag.fr/WSD

Résumé. Nous présentons une méthode pour créer rapidement un système de désambiguïsation lexicale (DL) pour une langue L peu dotée pourvu que l'on dispose d'un système de traduction automatique statistique (TAS) d'une langue riche en corpus annotés en sens (ici l'anglais) vers L. Il est, en effet, plus facile de disposer des ressources nécessaires à la création d'un système de TAS que des ressources dédiées nécessaires à la création d'un système de DL pour la langue L. Notre méthode consiste à traduire automatiquement un corpus annoté en sens vers la langue L, puis de créer le système de désambiguïsation pour L par des méthodes supervisées classiques. Nous montrons la faisabilité de la méthode et sa généricité en traduisant le *SemCor*, un corpus en anglais annoté grâce au *Princeton WordNet*, de l'anglais vers le bangla et de l'anglais vers le français. Nous montrons la validité de l'approche en évaluant les résultats sur la tâche de désambiguïsation lexicale multilingue de Semeval 2013.

Abstract.

Rapid Construction of Supervised Word Sense Disambiguation System for Lesser-resourced Languages
We introduce a method to quickly build a Word Sense Disambiguation (WSD) system for a lesser-resourced language L, under the condition that a Statistical Machine Transation system (SMT) is available from a well resourced language where semantically annotated corpora are available (here, English) towards L. We argue that it is less difficult to obtain the resources mandatory for the development of an SMT system (parallel-corpora) than it is to create the resources necessary for a WSD system (semantically annotated corpora, lexical resources). In the present work, we propose to translate a semantically annotated corpus from English to L and then to create a WSD system for L following the classical supervised WSD paradigm. We demonstrate the feasibility and genericity of our proposed method by translating *SemCor* from English to Bangla and from English to French. *SemCor* is an English corpus annotated with *Princeton WordNet* sense tags. We show the feasibility of the approach using the Multilingual WSD task from Semeval 2013.

Mots-clés : clarification de texte, désambiguïsation lexicale, langues peu dotées, traduction automatique, portage d'annotations.

Keywords: clarification of texts, word sens disambiguation, under resourced languages, machine translation, annotation transfert.

1 Introduction

La clarification de texte est une tâche centrale pour le traitement automatique des langues. Elle peut, en effet, permettre d'améliorer de nombreuses applications comme l'extraction d'informations multilingues, le résumé automatique ou encore la traduction automatique. Il s'agit de lever, manuellement ou automatiquement, un certain nombre d'ambiguïtés : déterminer les différents acteurs qui interviennent dans l'énoncé, leurs rôles ou déterminer le sens des mots utilisés parmi un inventaire pré-défini (désambiguïsation lexicale). Par exemple, dans «*La souris mange le fromage.*», l'animal devrait être préféré au dispositif électronique de pointage et serait traduit en malais, par exemple, par ‹*tikus*› plutôt que par ‹*tetikus*›.

Les méthodes de désambiguïsation lexicale supervisée ont besoin de corpus annotés en sens de mot pour être entrainées. Malheureusement de tels corpus n'existent que dans très peu de langues, ce qui rend souvent impossible la désambiguï-

sation lexicale supervisée.

Dans cet article, nous présentons une méthode pour créer rapidement un système de désambiguïsation lexicale supervisée pour une langue L peu dotée. Cette méthode nécessite un système de traduction automatique statistique (TAS) d'une langue riche en corpus annotés en sens vers L. Il est, en effet, plus facile de disposer des ressources nécessaires à la création d'un système de TAS (des corpus alignés) que des ressources dédiées nécessaires à la création d'un système de désambiguïsation lexicale pour la langue L. Le système de DL pour la langue L sera alors construit en utilisant les traductions annotées dans la langue L produite.

Dans cet article, nous présentons la désambiguïsation lexicale en fonction des ressources qu'elle utilise et montrons que beaucoup de langues sont trop peu dotées pour permettre la construction d'un système de DL supervisée. Nous présentons notre méthode de construction de corpus annotés par traduction automatique et l'illustrons avec le français et le bengali. Enfin, nous évaluons la méthode sur le corpus de la tâche de désambiguïsation lexicale multilingue de Semeval 2013.

2 Désambiguïsation lexicale et langue

2.1 Processus général de la désambiguïsation lexicale

La mise en place d'un système de désambiguïsation lexicale se déroule en trois étapes :

1) constitution de ressources génériques : plusieurs ressources non dédiées à la désambiguïsation lexicale sont envisageables : dictionnaires, encyclopédies, corpus non annotés, corpus annotés, bases lexicales, ... Certaines sont construites automatiquement parfois en utilisant d'autres ressources. Cette étape est optionnelle et est souvent réalisée par des équipes spécialisées. Elle est celle qui demande le plus de supervision humaine.

2) constitution d'une ressource dédiée à la DL : utilisation d'une ou plusieurs ressources génériques pour associer une représentation informatique à chacun des sens d'un mot. Ces sens sont soit directement définis à partir de l'expertise humaine pour certaines ressources génériques comme les bases lexicales, soit induits à partir des contextes d'utilisation dans les textes (induction de sens). Structurellement, la ressource peut être, par exemple, un graphe, des chaines de caractères ou des représentations vectorielles ;

3) utilisation de la ressource dédiée pour désambiguïser des textes ; il s'agit de l'algorithme de désambiguïsation proprement dit. Plusieurs paramètres peuvent être définis pour le traitement. Certains sont communs à tous les algorithmes comme la taille du contexte considéré pour un mot cible (par exemple quelques mots avant ou après la cible, la phrase qui contient la cible, voire le texte) tandis que d'autres dépendent du type d'algorithme mis en œuvre (par exemple la limite à considérer pour la profondeur de la recherche dans un graphe ou encore les paramètres à considérer pour des algorithmes stochastiques).

Ainsi, selon ce point de vue, (Schwab *et al.*, 2013) utilisent WordNet comme ressource générique, une représentation sous forme de sacs de mots issus des définitions des sens et de leurs liens comme ressource dédiée, un algorithme à colonies de fourmis et une mesure de proximité entre les sacs de mots comme algorithme de désambiguïsation lexicale. Roberto Navigli et son équipe (Navigli & Ponzetto, 2012) utilisent *BabelNet* comme ressource générique, une représentation sous forme de graphe issu des sens et de leurs liens comme ressource dédiée, des algorithmes de graphes (*Pagerank*, *Degree*, ...) comme algorithmes de désambiguïsation.

2.2 Ressources pour la Désambiguïsation lexicale

En désambiguïsation lexicale, deux types de ressources sont importantes : des corpus manuellement annotés par des sens et des sources de connaissances. Les campagnes d'évaluation sur l'anglais ont globalement montré que plus un système utilise de corpus annotés, meilleurs sont les résultats. De même, meilleures sont les sources de connaissances en termes de quantité et de qualité, meilleurs sont les résultats. Dans le processus d'informatisation d'une langue, avant de pouvoir construire un corpus annoté manuellement par des sens, il faut disposer d'un inventaire de sens. Aucune autre langue que l'anglais ne bénéficie d'autant de corpus de textes manuellement annotés par des sens et de connaissances lexicales. La figure 1 permet d'illustrer de manière parlante l'état des ressources nécessaires pour la DL librement accessibles pour un certain nombre de langues. Il est donné pour que le lecteur puisse se faire une idée de la situation actuelle. Un recensement plus précis serait difficile à obtenir et il faut ainsi interpréter les positions des langues les unes par rapport aux autres plutôt

que de manière absolue sauf pour l'anglais que nous avons placé le plus en haut à droite. De fait si on peut considérer que la quantité de données annotées est un paramètre quantifiable (par exemple en nombre moyen d'occurrences par terme du lexique), la richesse des sources de connaissances disponibles est, elle, plus floue. C'est en particulier le cas entre deux langues différentes puisque la taille de leur vocabulaire est différente. Il faut noter également que certaines langues peuvent bénéficier de données provenant d'autres langues par des alignements (comme c'est le cas dans BabelNet, par exemple).

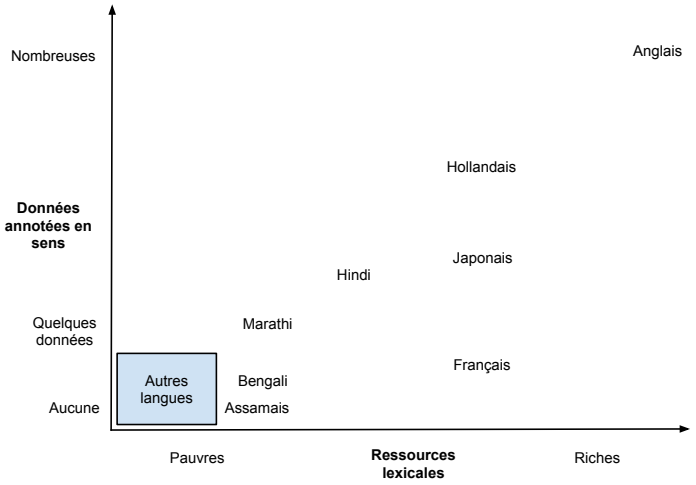

FIGURE 1: Données disponibles pour la désambiguïsation lexicale en fonction de la langue

2.2.1 Bases lexicales

WordNet Avant les années 1990, la désambiguïsation lexicale n'était pratiquement réalisée qu'à partir de dictionnaires électroniques. Le *Princeton WordNet* (Fellbaum, 1998), initié au milieu des années 1980, a permis la mise à disposition d'une ressource utilisable librement. Elle est devenue rapidement très populaire et a rapidement conduit à la disparition de l'usage des dictionnaires électroniques en désambiguïsation lexicale.

Le *Princeton WordNet* est organisé autour de la notion d'ensemble de synonymes (*synsets*) décrits par une partie du discours (nom, verbe, adjectif, adverbe), une définition et leurs liens (hyperonyme, hyponyme, antonyme, ...). Chaque sens d'un item lexical (entrée) correspond à un *synset*. La version courante du *Princeton WordNet*, la 3.0, comprend 155 287 items lexicaux pour un total de 117 659 *synsets*. Des versions pour d'autres langues existent mais, faute de moyens humains équivalents, leur qualité est encore inférieure à celle de l'anglais. Bien souvent, les mots de ces langues sont décrits grâce à des *synsets* du *Princeton WordNet*. La *Global WordNet Association* établit la liste des wordnets existants [1].

BabelNet BabelNet (Navigli & Ponzetto, 2012) est une ressource lexicale à grande échelle construite par alignement automatique des *synsets*, issus de *Princeton WordNet* et de pages Wikipedia correspondantes. BabelNet introduit la notion de *Babel Synset*, qui contient tout le contenu du *synset* correspondant dans le *Princeton WordNet*, ainsi qu'un ensemble de pages Wikipedia similaires. Cette correspondance entre *synsets* WordNet et pages Wikipédia se fait par un algorithme de désambiguïsation automatique. Les pages Wikipédia reliées par des hyperliens internes à Wikipédia ainsi que les articles associés dans les autres langues disponibles dans Wikipedia sont liés aux pages correspondantes. Pour toutes les pages dans les autres langues, si il n'y a pas de définition disponible ou extraite de la page, la définition anglaise *Princeton WordNet* ou un extrait venant de *SemCor* est traduit par *Google Translate* pour servir de définition.

BabelNet, dans sa dernière version, la 2.5.1, comprend 271 langues, 13 789 332 *Babel synsets*, 117 204 438 sens, 354 538 633 relations lexico-sémantiques et 40 328 194 définitions textuelles. Pour l'anglais il y a 11 812 887 entrées, 6 670 627 *Babel synsets*, 16 741 223 sens de mots et un degré de polysémie de 7,51. Pour le français, il y a 5 295 221

1. http://globalwordnet.org/wordnets-in-the-world/

entrées, 4 120 733 Babel synsets, 7 076 728 sens de mots et un degré de polysémie de 1,72. Pour le bengali il y a 188 511 entrées, 34 832 Babel synsets, 233 163 sens de mots et un degré de polysémie de 6,69.

2.2.2 Corpus annotés

Selon Benoit Habert, «*un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue*» (Habert *et al.*, 1998). Généralement, un corpus contient jusqu'à une douzaine de millions de mots et peut être lemmatisé et annoté avec des informations concernant les parties du discours. Parmi ces corpus, on trouve le *British National Corpus* (Burnard, 1998) (100 millions de mots) et le *American National Corpus* (Ide & Macleod, 2001) (20 millions de mots). Les textes proviennent de diverses sources comme des journaux, des livres, des encyclopédies ou du Web.

Exemples de corpus annotés En désambiguïsation lexicale, plusieurs corpus annotés en sens sont utilisés. On peut citer, par exemple :

1. La *Defense Science Organisation* (Ng & Lee, 1996) a produit un corpus non disponible librement. 192 800 mots ont été annoté avec des *synsets* du *Princeton WordNet*. L'annotation se concentre sur 121 noms (113 000 occurences) et 70 verbes (79 800 occurences) qui ont été choisies parmi les plus fréquents et les plus ambigus de l'anglais. Selon les auteurs, la couverture correspond à environ 20% des occurrences de noms et de verbes en anglais.

2. Le *SemCor* (Miller, 1995) est un sous-ensemble du Corpus de Brown (Francis & Kučera, 1964). Sur les 700 000 mots de ce dernier, environ 230 000 sont annotés avec des *synsets* du *Princeton WordNet*. L'annotation porte au total sur 352 textes. Pour 186 d'entre eux, 192 639 mots (soit l'ensemble des noms, verbes, adjectifs et adverbes) sont annotés. Sur les 166 autres, seulement 41 497 verbes sont annotés.

3. Les corpus issus des campagnes d'évaluation. Depuis 1998, il y a eu plusieurs campagnes (semeval-senseval) destinées à évaluer la désambiguïsation lexicale. La plupart ont concerné l'anglais mais également le japonais, l'espagnol, le chinois ou le français. La taille de ces corpus est de l'ordre d'une centaine de fois plus petite que celle des deux précédents corpus, soit quelques milliers de mots.

Difficultés liées à la construction d'un corpus annoté Il n'existe que peu de données manuellement annotés. La *Global WordNet Association* dresse la liste des 26 corpus annotés avec un wordnet [2]. Ces corpus concernent 17 langues. Seules trois d'entre elles (l'anglais, le hollandais et le bulgare) atteignent les 100 000 annotations. À notre connaissance, il n'existe pas de donnée annotée pour le bengali et très peu pour le français (environ 3600 mots annotés avec le dictionnaire Larousse pour la campagne Romanceval 1998 et 1656 mots annotés avec des sens de BabelNet pour la tâche 12 de la campagne SemEval 2013). Ces deux langues qui nous intéressent plus particulièrement sont donc peu dotées en ce domaine.

La construction d'un corpus manuellement annoté en sens est réputée comme une tâche très difficile par comparaison à d'autres tâches d'annotation. En effet, s'il n'y avait que 45 annotations possibles pour le *Penn Treebank* (Marcus *et al.*, 1993), un corpus annoté en parties du discours, il y en a autant que de synsets (117 000) pour une annotation en sens issus du *Princeton WordNet*. Ainsi, pour l'annotation du corpus de la *Defense Science Organisation*, alors que les conditions étaient plus favorables que celles des annotateurs du *SemCor* (uniquement 191 mots différents pour seulement 1 800 annotations possibles), le taux d'annotation était seulement de 150 à 250 mots par heure (1 homme-année pour les 192 800 occurrences de mots) tandis que les annotateurs du *Penn Treebank* réalisaient 6 000 annotations par heure.

Dans de telles conditions, on comprends mieux pourquoi assez peu de corpus annotés existent. Des recherches ont visé à faciliter cette annotation. Par exemple, (Vossen *et al.*, 2011) utilisent, pour le hollandais, un algorithme de désambiguïsation automatique dont les annotations les moins sûres sont vérifiées/modifiées par les annotateurs et (Mihalcea & Chklovski, 2003) utilisent des méthodes de *crowdsourcing* pour augmenter le nombre d'annotateurs.

Les corpus de grande taille annotés en sens sont pourtant le seul moyen de mettre en œuvre un processus de désambiguïsation lexicale supervisée.

2. `http://globalwordnet.org/wordnet-annotated-corpora/` consultée le 4 février 2015. Il existe d'autres corpus annotés avec des synsets de wordnets comme ces corpus de domaines annotés avec des *synsets* de l'*Hindi WordNet* `http://www.cfilt.iitb.ac.in/wsd/annotated_corpus/`

2.3 Désambiguïsation lexicale supervisée

Le principe, issu de l'apprentissage automatique, consiste à entraîner un classifieur pour chaque mot cible, afin de prédire le sens le plus vraisemblable en fonction de son contexte. Dans les termes utilisés à la section 2.1, la ressource générique est le corpus annoté en sens, la ressource dédiée est construite au moyen de classifieurs utilisés pour discriminer le contexte de chaque mot afin de déterminer son meilleur sens.

Ces approches (dites supervisées), sont très populaires pour l'anglais. Dans les campagnes d'évaluation Senseval-SemEval, elles ont obtenu, de loin, les meilleures performances sur l'anglais. La limitation principale est la nécessité de disposer de grands corpus annotés ce dont peu de langues disposent comme nous l'avons vu dans la partie précédente. Dans cet article, nous proposons une méthode fondée sur la traduction automatique statistique et le portage d'annotations d'un corpus comme nous le montrons dans la suite.

3 Construction d'un corpus annoté par traduction automatique

3.1 Principe général

Dans ce travail, nous proposons une méthode de traduction automatique et de transfert direct des annotations qui nous permet d'obtenir des corpus dans toutes les langues disposant d'un système de traduction avec comme source une langue possédant un corpus annoté en sens. Notre méthode a été mise en œuvre sur le *SemCor*, corpus en anglais annoté avec des sens issus du *Princeton WordNet* (voir section 2.2.2) à l'aide d'un système de traduction construit avec la boîte à outils *Moses*, de l'anglais vers le français et de l'anglais vers le bengali. Nous avons ainsi obtenu un corpus annoté en sens issus du *Princeton WordNet* pour le français et un autre pour le bengali.

3.2 Transfert d'annotations

À notre connaissance, le transfert d'annotations linguistiques a été utilisé à partir des années 1990 (Brown *et al.*, 1991). Dans cette approche, le principe consistait à exploiter des corpus parallèles (source, cible) annotés à la source et de construire un modèle d'alignements qui permet de transférer ces annotations vers la cible.

De tels transferts d'annotations ont été appliqués à une large gamme d'annotations : les parties du discours (Yarowsky & Ngai, 2001), les dépendances syntaxiques (Hwa *et al.*, 2005), les *chunks* (Yarowsky *et al.*, 2001), les rôles sémantiques (Padó & Lapata, 2009), *etc.* De plus récents travaux comme (van der Plas & Apidianaki, 2014) n'exploitent pas directement les données parallèles mais utilisent des techniques de désambiguïsation par traductions pour déterminer les alignements de mots et projeter les annotations. De leur côté, (Wang & Manning, 2014) utilisent un corpus parallèle et un système de désambiguïsation lexicale multilingue pour obtenir les traductions en contexte de chaque mot du corpus, ce qui leur permet d'ensuite transférer une annotation en rôle sémantique depuis l'anglais vers le français.

Dans le contexte plus spécifique de la désambiguïsation lexicale, (Diab & Resnik, 2002) utilisent un système de traduction automatique commercial pour traduire un corpus en anglais vers une langue cible, les mots anglais provenant de la source sont alors utilisés comme annotations sémantiques dans la cible. Le projet *MultiSemCor*, (Padó & Lapata, 2009) est très proche de l'approche que nous présentons ici puisqu'il s'agit de faire traduire en italien par des traducteurs professionnels une sous-partie du *SemCor*. Notre approche consiste à simultanément traduire le corpus et à porter les annotations de la source grâce à un système de TAS construit avec la boîte à outil *Moses*.

3.3 Traduction automatique statistique

La traduction automatique (TA) est le processus qui consiste à traduire un énoncé d'une langue naturelle (langue source) à une autre (langue cible). L'énoncé peut-être soit écrit ou oral mais nous ne nous intéresserons qu'à l'écrit dans cet article.

La traduction automatique statistique (TAS) est actuellement une approche très largement utilisée en TA. Schématiquement, un système est, dans un premier temps, entrainé sur des corpus parallèles langue source - langue cible. Il s'agit d'obtenir des informations statistiques permettant de calculer quelles sont les meilleures traductions pour un mot ou une

suite de mots (approche dite *phrase-based*). Dans un second temps, ces informations sont exploitées pour produire des traductions de la langue source à la langue cible.

Moses [3] (Hoang & Koehn, 2008) est une boîte à outils pour la traduction automatique statistique. Sa licence est libre (LGPL licence) ce qui facilite l'obtention d'un système statistique complet à l'état-de-l'art.

3.4 Mise en œuvre de notre approche

Notre approche consistant à traduire le *SemCor* et à porter ses annotations a pu être mise en œuvre grâce à un premier système de traduction de l'anglais vers le français et un second système depuis l'anglais vers le bengali que nous présentons maintenant.

3.4.1 Système anglais–français

Le système de traduction statistique anglais-français est celui mis au point par le Laboratoire d'Informatique de Grenoble pour participer à la campagne 2012 d'IWSLT (*International Workshop on Spoken Language Translation*) (Besacier *et al.*, 2012). Ce système a été construit avec des données alignées usuelles (*Europarl Parallel Corpus*, *United Nations Parallel Corpus*, ...). Il a été évalué avec la métrique BLEU — score de 24,85 — ainsi que par des juges humains — score de 11.

3.4.2 Système anglais–bengali

Notre système anglais-bengali n'a pas encore fait l'objet d'une publication, nous le décrivons donc de manière plus détaillée. Il est basé sur la boîte à outil *Moses* (version 2.1). Les données proviennent de différentes sources :

— Corpus parallèles : EMILLE corpora, OPUS corpus (KDE4, GNOME), INDIC, OpenSubtitles2013, Tanzil, Tatoeba, Bibel, Jehova Witness
— Corpus monolingues pour le bengali : l'ensemble des corpus parallèles cités ci-dessus, extraction du Wikipedia bengali du 28 décembre 2014 qui comporte 28393 articles.

Les ressources ont été pré-traitées comme suit : (1) filtrage des marqueurs HTML, (2) conversion de tous les caractères en UTF-8, (3) application de la normalisation forme D, (4) application de la normalisation des ponctuations (retrait des marques de ponctuation redondantes, conversion vers la version canonique des caractères, *etc.*), (5) tokenisation, (6) suppression des phrases de plus de 50 mots (7) suppression des phrases trop longues (ratio supérieur à 9).

Après ce pré-traitement, nous obtenons ces statistiques :

données parallèles				
	phrases	679 019		
		tokens	types	mots
	anglais	12 096 756	165 765	132 404
	bengali	11 288 581	229 634	14 497
données monolingues				
	phrases	1 107 776		
		tokens	types	mots
	bengali	21 273 100	317 265	18 632

Nous avons utilisé 85% de ces données pour l'entraînement, 10% pour l'optimisation du système et 5% pour le tester. Le score bleu de ce système est de 27,21.

3.4.3 Traduction du *SemCor* et portage des annotations

Pour traduire le *SemCor* (Miller *et al.*, 1993) (v. 3.0) et porter ses annotations vers la langue cible, nous utilisons nos systèmes de traduction automatique statistique. Un extrait du *SemCor* est présenté dans la figure 2. En même temps que le système traduit le texte, phrase par phrase, nous extrayons la correspondance mot à mot que nous fournit le décodeur *Moses* et nous l'utilisons pour transférer les annotations d'un mot source au mot correspondant dans le texte cible suivant

3. http ://www.statmt.org/moses/

```
<contextfile concordance=brown>
 <context filename=br-a01 paras=yes>
  <p pnum=1>
   <s snum=1>
    <wf cmd=ignore pos=DT>The</wf>
    <wf cmd=done rdf=group pos=NNP lemma=group wnsn=1 lexsn=1:03:00:: pn=group>
    Fulton_County_Grand_Jury</wf>
    <wf cmd=done pos=VB lemma=say wnsn=1 lexsn=2:32:00::>said</wf>
    <wf cmd=done pos=NN lemma=friday wnsn=1 lexsn=1:28:00::>Friday</wf>
    <wf cmd=ignore pos=DT>an</wf>
    <wf cmd=done pos=NN lemma=investigation wnsn=1 lexsn=1:09:00::>
    investigation</wf>
   <wf cmd=ignore pos=IN>of</wf>
   <wf cmd=done pos=NN lemma=atlanta wnsn=1 lexsn=1:15:00::>Atlanta</wf>
    [...]
    <punc>.</punc>
   </s>
  </p>
  [...]
 </context>
</contextfile>
```

FIGURE 2: *SemCor*(v. 3.0) dans le format SGML (Standard Generalized Markup Language).

l'algorithme 1. La traduction et le transfert des annotations a pris environ une semaine en utilisant un seul cœur sur un serveur 32-cœur Intel Xeon E5-2650, 2.0 GHz et 256 GB de mémoire physique.

4 Systèmes de désambiguïsation lexicale

À partir de ces corpus annotés en sens, nous pouvons utiliser des méthodes de désambiguïsation supervisée pour construire un ou plusieurs systèmes de désambiguïsation automatique. Dans cette section, nous présentons une méthode supervisée, un classifieur bayésien naïf et son évaluation sur le corpus de *SemEval 2013*.

4.1 Principe

Un corpus annoté contient un grand nombre de textes segmentés en phrases et en mots puis lemmatisés et annotés sémantiquement. Ainsi, pour chaque mot du corpus, nous pouvons obtenir, une liste de phrases dans lesquelles ce mot apparaît dans ses différents sens et en extraire des attributs du contexte prédicteurs du sens. Les prédicteurs habituels incluent les catégories grammaticales des mots du cotexte, les lemmes de ces mots ainsi que leur position dans la phrase.

Un classifieur est un algorithme nécessitant d'apprendre un modèle de prédiction afin d'affecter une série d'instance à un ensemble fini de classes. Pour cela, il faut fournir au classifieur une série d'instances annotées avec les classes auxquelles elle appartiennent afin de construire le modèle. Dans le cas de la désambiguïsation supervisée, les classes à prédire sont les identifiants des sens et les instances sont les attributs prédicateurs extraits.

4.2 Classifieur bayésien naïf

Pour un ensemble d'instances $I_x = (x_1, x_2, \ldots, x_n)$ annotées avec N classes $C_k, k \in 1..N$, un classifieur bayésien naïf estime la probabilité d'obtenir une classe C_k en fonction d'un certain ensemble d'attributs $x_1, \ldots, x_n$: $P(C_k \mid x_1, \ldots, x_n) = p(C_k)p(x_1 \mid C_k)p(x_2 \mid C_k, x_1) \ldots p(x_n \mid C_k, x_1, \ldots, x_n)$. L'algorithme fait la supposition de l'indépendance des attributs les uns par rapport aux autres, ce qui implique l'approximation suivante, avec $Z = p(x)$ un facteur de normalisation :

$$P(C_k \mid x_1, \ldots, x_n) \simeq \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

Algorithm 1: Processus de traduction du *SemCor* et portage de ses annotations

Entrées: SC : SemCor
Sorties: TSC : SemCor traduit annoté
`decoder` : chemin absolu vers le décodeur (dans notre cas, *Moses*);
`model` : chemin absolu vers le fichier du modèle entrainé (`moses.ini`);
`-config` : chemin absolu vers le fichier de configuration (option de *Moses*);
`-print-alignment-info` : sortie alignement mot à mot vers la sortie standard, séparée des traductions par ||| (option de Moses);
$initialization$;
`SemCorTranslation()`
- **pour tous les** *répertoires* $D \in SC$ **faire**
  - créer un nouveau répertoire TD;
  - **pour tous les** *fichiers* $F \in D$ **faire**
    - phrase $S \leftarrow \{\}$;
    - créer un nouveau fichier TF;
    - **pour tous les** *ligne* $L \in F$ **faire**
      - **pour tous les** *WF entrée* $E \in L$ **faire**
        - WF[cmd, rdf, $lemma$, pos, $lexsn$, $wnsn$, pn, ot, $word$, $punc$] $\leftarrow$ `ExtractWFInfoValue(`E`)`;
      - **fin**
      - $S \leftarrow S \cup \{WFInfo.word\}$;
    - **fin**
    - target-words $TWs \leftarrow$ `GetAlignedWords(`S`)`;
    - **pour** $i \in 1...|\,TWs\,|$ **faire**
      - $S[i].word \leftarrow TWs[i]$; $S[i].lemma \leftarrow TWs[i]$;
    - **fin**
    - target-sense $TS \leftarrow$ `SenseMapper` (*pos, lexsn, word*) ;
    - $S.lexsn \leftarrow TS$;
    - écrire S dans le fichier TF;
  - **fin**
- **fin**

`ExtractWFInfoValue(`E`)`
- WF Information $WFI \leftarrow \{\}$; WF Valeur $WFV \leftarrow \{\}$;
- **pour tous les** *item* $I \in E$ **faire**
  - split I by "=";
  - $WFI[] \leftarrow I[left\_item]$; $WFV[] \leftarrow I[right\_item]$;
- **fin**
- **retourner** $WFI[]$, $WFV[]$;

`GetAlignedWords(`S`)`
- alignments de mots $WAs \leftarrow \{\}$; target-words $TWs \leftarrow \{\}$;
- sortie de traduction$TO \leftarrow$ `MosesDecoder(`S`)` ;
- split TO by "|||";
- $WAs[] \leftarrow TO[right\_item]$;
- **pour tous les** *word alignment* $WA \in WAs$ **faire**
  - split WA by "$-$";
  - $TWs[] \leftarrow WA[right\_item]$;
- **fin**
- **retourner** $TWs[]$;

`MosesDecoder(`S`)`
- sortie de la traduction avec alignement de mots $TO \leftarrow -config\ model\ S\ -print\text{-}alignment\text{-}info$;
- **retourner** TO;

L'estimation du modèle est souvent faite par estimation de vraisemblance. Lorsque le modèle est construit, la décision (classification) est faite avec la règle du maximum *a posteriori* :

$$\hat{C_k} = \underset{k \in 1..N}{\arg\max}\, p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

Ce classifieur, simple à mettre en œuvre et extrêmement performant en termes de complexité d'apprentissage et de classification, est utilisé dans de nombreuses tâches, comme la détection de pourriels. Il s'utilise aussi couramment comme base de comparaison dans les travaux portant sur la classification supervisée. Nous l'avons utilisé ici car il offre de bonnes performances en désambiguïsation de l'anglais même avec peu de données d'entraînement (voir section 4.3) et permet de comprendre facilement d'où viennent les différences de performance dans la classification entre deux langues.

4.3 Systèmes implantés

Pour construire le système de désambiguïsation supervisée nous nous sommes basés sur les caractéristiques du système arrivé premier sur la tâche 7 de la campagne d'évaluation Semeval 2007, NUS-PT (Chan *et al.*, 2007). Suivant les travaux de (Lee & Ng, 2002) qui proposent une étude sur les meilleurs attributs à utiliser, NUS-PT en extrait trois types :

— *Les collocations locales* qui sont des séquences de mots autour du mot ciblé (inclus). Onze combinaisons du type $C_{i,j}$ sont utilisées, où i et j représentent les bornes de la collocation par rapport au mot ciblé : $C_{-1,-1}$; $C_{1,1}$; $C_{-2,-2}$; $C_{2,2}$; $C_{-2,-1}$; $C_{-1,1}$; $C_{1,2}$; $C_{-3,-1}$; $C_{-2,1}$; $C_{-1,2}$; $C_{1,3}$. Par exemple, dans la phrase «*Le chaton blanc est très heureux*», si le mot ciblé est ‹‹blanc››, alors $C_{-1,1}$ correspond à la collocation «*chaton blanc est*», et $C_{1,3}$ correspond à «*est très heureux*» ;
— *Les catégories grammaticales* avoisinantes qui correspondent aux catégories grammaticales de trois mots sur la gauche et de trois mots sur la droite ;
— *les mots du cotexte* correspondant à la même fenêtre de $-3, 3$.

Dans le système que nous avons implanté, nous avons réutilisé les mêmes attributs. Toutefois, pour cette expérience, qui consiste uniquement à examiner la faisabilité de l'approche, nous n'utilisons que le *SemCor* contrairement à NUS-PT qui utilise également le corpus de la DSO et des exemples extraits de corpus de tests parallèles.

Dans les cas où le classifieur ne retourne pas de réponse (pas assez d'exemples d'entraînement, mots non existants dans *SemCor*), nous utilisons la solution de repli classique, utilisée par exemple par NUS-PT, qui consiste à assigner le premier sens (le plus fréquent).

Trois systèmes ont été créés :

— un système permettant de désambiguïser de l'anglais avec des sens issus du *Princeton WordNet* ou des sens issus de *BabelNet* grâce aux alignements *synsets-Babel Synsets*. Il s'agit de notre système de référence construit directement à partir du *SemCor* ;
— un système permettant de désambiguïser du bengali avec des sens issus du *Princeton WordNet*. Ce système est construit à partir de la traduction vers le bengali du *SemCor*. Le bengali ne disposant pas de corpus annotés en sens à notre connaissance, une évaluation classique directe de ses performances (*in vitro*) n'est pas possible. Il fait l'objet, en revanche, d'étude de ses performances *in vivo*, c'est-à-dire utilisé dans une application, mais ce n'est pas l'objet de cet article. L'évaluation *in vitro* du troisième système présenté ici nous permettra d'avoir une idée des performances de cet annotateur du bengali ;
— un système permettant de désambiguïser du français avec des sens issus du *Princeton WordNet* ou des sens issus de *BabelNet* grâce aux alignements synsets-BabelSynsets. Ce système, construit à partir de la traduction vers le français du *SemCor*, est comparé au système de référence pour l'anglais sur la tâche de désambiguïsation multilingue (tâche 12) de Semeval 2013.

Ces trois systèmes sont mis à la disposition de la communauté et accessibles à l'adresse `http://getalp.imag.fr/static/wsd/TALN2015/`.

5 Évaluation de la méthode

Pour tester notre approche et vérifier qu'elle est générique, nous nous plaçons dans un contexte absolument similaire pour le système du bengali et les systèmes du français. La seule source de données utilisée dans les deux cas est donc une

traduction du *SemCor* (voir partie 3.4.3).

Nous allons chercher à estimer la perte de performance liée à l'utilisation de la traduction du corpus. Cela est rendu possible par l'utilisation du corpus de la tâche 12 (désambiguïsation lexicale multilingue) de Semeval 2013 qui est un corpus d'évaluation traduit en 5 langues (anglais, français, allemand, italien, espagnol) pour lequel nous utilisons les textes anglais et français.

5.1 Corpus de Semeval 2013 tâche 12 : désambiguïsation lexicale multilingue

Comme mentionné ci-dessus, le corpus de Semeval 2013 (tâche 12) comporte 5 langues dans lesquelles il a été traduit ainsi que les annotations sémantiques transférées puis validées manuellement. Il comprend 13 textes de différents domaines (politique, commentaire sportif, domaine général).

Puisque *SemCor* est annoté avec les sens de *Princeton WordNet* et que l'évaluation se fait avec des sens BabelNet, nous avons réalisé une conversion en utilisant les alignements de BabelNet avec WordNet [4]. Nous avons d'abord récupéré les offsets des synsets correspondant aux annotations de sens dans WordNet puis avons interrogé l'API java de BabelNet pour obtenir les identifiants BabelNet correspondants. Nous avons obtenu une correspondance de **96,10%** sur l'ensemble de *SemCor*. Nous avons ensuite réalisé l'apprentissage pour obtenir les systèmes présentés en 4.3 puis les avons appliqués sur notre corpus d'évaluation.

5.2 Mesures d'évaluation

La tâche de désambiguïsation lexicale multilingue de SemEval 2013 utilise les mesure classiques de précision P, de rappel R et de score F_1 qui correspond à la moyenne harmonique de P et R. La précision se définit comme $P = \frac{\text{annotés correctement}}{\text{total annotés}}$, le rappel comme $R = \frac{\text{annotés correctement}}{\text{total à annoter}}$ et le score F_1 comme $F_1 = \frac{2 \cdot P \cdot R}{P+R}$.

5.3 Résultats et analyse

Rappelons, tout d'abord, qu'il ne s'agissait pas ici d'obtenir des meilleurs scores possibles mais de montrer qu'il est possible de créer des systèmes de désambiguïsation lexicale pour des langues qui n'ont pas (ou trop peu) de données annotées en sens.

Système	Précision	Rappel	Score F1
SUP-EN	64,80%	64.70%	64,75%
MFS-EN	66,90%	66,60%	66,64%
SUP-FR	51,60%	51,50%	51,55%
MFS-FR	45,60%	45,10%	45,34%

TABLE 1: Les résultats (Précision, Rappel, score F1) de l'expérience. MFS-EN/FR sont respectivement les résultats de la référence avec sens le plus fréquent en anglais et en français. SUP-EN/SUP-FR sont les résultats pour le système supervisé en anglais (apprentissage directement sur le *SemCor*) et en français (apprentissage sur une traduction du *SemCor*.

Le tableau 1 présente les résultats de l'expérience. Les systèmes MFS-EN et MFS-FR sont les résultats obtenus en assignant systématiquement le sens le plus fréquent (*Most Frequent Sense* - heuristique du sens le plus fréquent) comme réponse. Nous pouvons déjà constater que cette référence classique en évaluation de la désambiguïsation lexicale, fournit des résultats pour le français bien inférieurs à ceux de l'anglais.

SUP-EN correspond à notre système de référence, pour l'anglais, construit directement à partir du *SemCor*. Il obtient un score inférieur d'environ 2% sur l'heuristique du sens le plus fréquent. Le système créé à partir de la traduction vers le français du *SemCor*, SUP-FR obtient $51,6\%$ soit une perte de $13,2\%$ si on compare à l'anglais mais il s'agit d'un score qui est supérieur de 6.21% de la référence du sens le plus fréquent.

À notre connaissance, il n'existe pas d'autre expérience utilisant un système supervisé sur ce corpus. Il convient donc de considérer avec précaution les comparaisons avec les systèmes ayant participé à la tâche car ils bénéficiaient de la

4. BabelNet est construit sur la base de Wordnet 3.1 alors que *SemCor* 3.0 est basé sur *Princeton WordNet* 3.0

richesse des informations de BabelNet. 3 équipes ont participé à la campagne. Notre système référence sur l'anglais aurait été entre la première ($-3,7\%$) et la seconde ($+4,4\%$) tandis que le système sur le français aurait terminé entre la seconde ($-2,25\%$) et la troisième ($+3,25\%$).

Nous voyons ici qu'avec un système supervisé non optimisé avec des données d'entrainement qui peuvent être facilement étoffées en ajoutant d'autres corpus, nous obtenons déjà des résultats qui valident notre approche.

6 Conclusion et perspectives

Dans cet article, nous avons montré qu'il existe des corpus annotés dans certaines langues (en anglais par exemple) alors qu'il en existe peu ou pas dans la plupart des langues. Ces corpus sont pourtant essentiels à la création de systèmes de désambiguïsation lexicale supervisée. Nous avons proposé une méthode qui consiste à traduire des corpus annotés et à porter ces annotations dans la langue pour laquelle on veut un système de désambiguïsation. Nous avons utilisé le même script sur un système *Moses* traduisant de l'anglais vers le bengali et de l'anglais vers le français pour créer un système de désambiguïsation du bengali et un système de désambiguïsation du français. Nous avons ainsi montré la faisabilité et la généricité de l'approche. Nous avons aussi montré qu'en utilisant un apprentissage supervisé naïf, sur assez peu de données, qui plus est traduites automatiquement, on obtient des performances suffisantes pour valider cette approche.

Dans l'avenir, nous envisageons d'utiliser plus de corpus annotés libres ou gratuits pour la recherche mais également d'acquérir et d'utiliser le DSO qui, lui, ne l'est pas. Nous souhaitons également comparer différents algorithmes supervisés voire les combiner par une fusion tardive, par exemple. Enfin une dernière piste d'amélioration consistera à essayer d'améliorer les attributs du contexte permettant l'apprentissage automatique en intégrant, par exemple, les attributs utilisés par d'autres systèmes.

Références

BESACIER L., LECOUTEUX B., AZOUZI M. & LUONG NGOC Q. (2012). The LIG English to French Machine Translation System for IWSLT 2012. In *In proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*, p. 102–108, Unknown.

BROWN P., PIETRA S. D., PIETRA V. D. & MERCER R. (1991). Word-sense disambiguation using statistical methods. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, p. 264–270 : Association for Computational Linguistics.

BURNARD L. (1998). *The British National Corpus*.

CHAN Y. S., NG H. T. & ZHONG Z. (2007). Nus-pt : exploiting parallel texts for word sense disambiguation in the english all-words tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, p. 253–256 : Association for Computational Linguistics.

DIAB M. & RESNIK P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 255–262, Stroudsburg, PA, USA : Association for Computational Linguistics.

FELLBAUM C. (1998). *WordNet*. Wiley Online Library.

FRANCIS W. N. & KUČERA H. (1964). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Rapport interne, Brown University, Providence, Rhode Island.

HABERT B., FABRE C. & ISSAC F. (1998). *DE L'ECRIT AU NUMERIQUE. Constituer, normaliser et exploiter les corpus électroniques*. Number ISBN : 2-225-82953-5. ELSEVIER MASSON.

HOANG H. & KOEHN P. (2008). Design of the moses decoder for statistical machine translation. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, p. 58–65 : Association for Computational Linguistics.

HWA R., RESNIK P., WEINBERG A., CABEZAS C. & KOLAK O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, **11**(3), 311–325.

IDE N. & MACLEOD C. (2001). The american national corpus : A standardized resource of american english. In *Proceedings of Corpus Linguistics 2001*, volume 3.

LEE Y. K. & NG H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, p. 41–48, Stroudsburg, PA, USA : Association for Computational Linguistics.

MARCUS M. P., MARCINKIEWICZ M. A. & SANTORINI B. (1993). Building a large annotated corpus of english : The penn treebank. *Computational linguistics*, **19**(2), 313–330.

MIHALCEA R. & CHKLOVSKI T. (2003). *Building sense tagged corpora with volunteer contributions over the Web*, p. 357–402. John Benjamin Publishing Compagny.

MILLER G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

MILLER G. A., LEACOCK C., TENGI R. & BUNKER R. T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, p. 303–308 : Association for Computational Linguistics.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

NG H. T. & LEE H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense : An exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, p. 40–47 : Association for Computational Linguistics.

PADÓ S. & LAPATA M. (2009). Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, **36**(1), 307–340.

SCHWAB D., GOULIAN J. & TCHECHMEDJIEV A. (2013). Worst-case complexity and empirical evaluation of artificial intelligence methods for unsupervised word sense disambiguation. *International Journal of Web Engineering and Technology*, **8**(2), 124–153.

VAN DER PLAS L. & APIDIANAKI M. (2014). Cross-lingual word sense disambiguation for predicate labelling of french. In *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, p. 46–55 : Association pour le Traitement Automatique des Langues.

VOSSEN P., GÖRÖG A., LAAN F., VAN GOMPEL M., IZQUIERDO-BEVIA R. & VAN DEN BOSCH A. (2011). Dutchsemco : building a semantically annotated corpus for dutch. In *Electronic lexicography in the 21st century : New Applications for New Users : Proceedings of eLex 2011, Bled, 10-12 November 2011*, p. 286–296.

WANG M. & MANNING D. C. (2014). Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1*, p. 55–66.

YAROWSKY D. & NGAI G. (2001). Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, p. 1–8, Stroudsburg, PA, USA : Association for Computational Linguistics.

YAROWSKY D., NGAI G. & WICENTOWSKI R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, p. 1–8 : Association for Computational Linguistics.