Amélioration de la traduction automatique d'un corpus annoté

Marwa Hadj Salah[1, 2] Hervé Blanchon[1] Mounir Zrigui[2] Didier Schwab[1]
(1) LIG-GETALP, Univ. Grenoble Alpes, France
Prénom.Nom@imag.fr
http://getalp.imag.fr/WSD/
(2) LaTICE, Tunis, 1008, Tunisie
Prénom.Nom@fsm.rnu.tn

RÉSUMÉ

Dans cet article, nous présentons une méthode pour améliorer la traduction automatique d'un corpus annoté et porter ses annotations de l'anglais vers une langue cible. Il s'agit d'améliorer la méthode de (Nasiruddin *et al.*, 2015) qui donnait de nombreux segments non traduits, des duplications et des désordres. Nous proposons un processus de pré-traitement du SemCor anglais, pour qu'il soit adapté au système de traduction automatique statistique utilisé, ainsi qu'un processus de post-traitement pour la sortie. Nous montrons une augmentation de 2,9 points en terme de score F1 sur une tâche de désambiguïsation lexicale ce qui prouve l'efficacité de notre méthode.

ABSTRACT

Improvement of the automatic translation of an annotated corpus

In this article, we present a method to improve the automatic translation of an annotated corpus and transfer its annotations from English to any target language. The idea is to improve method of (Nasiruddin *et al.*, 2015) which leads to many untranslated segments, duplications and disorders. We propose a pre-treatment process for the English SemCor, to adapt it to the statistical machine translation system, as well as a post-treatment process for the output of SMT. We show an increase of 2,9 points in terms of F1 score on a Word Sense Disambiguation task which proves the effectiveness of our method.

MOTS-CLÉS : Portage d'annotations, Traduction Automatique Désambiguïsation lexicale.

KEYWORDS: Annotation transfert, Machine Translation, Word Sense Disambiguation.

1 Introduction

Le manque de corpus libres de droit annotés en sens pour la plupart des langues pose un problème crucial pour diverses applications et tâches du traitement automatique des langues. C'est évidemment le cas pour la désambiguïsation lexicale qui est la tâche qui consiste à déterminer quel est le sens le plus approprié pour chaque mot d'un texte dans un inventaire prédéfini.

Dans cet article, nous reprenons les travaux décrits dans (Nasiruddin *et al.*, 2015) qui portent sur la construction de corpus annotés par traduction automatique afin de créer rapidement un système de désambiguïsation lexicale supervisée. L'approche, illustrée avec le français et le bengali, présentait certaines limites liées en particulier à l'absence de normalisation des données.

Dans cet article, nous expliquons pourquoi et comment nous avons amélioré cette méthode en

l'enrichissant de traitements génériques, c'est-à-dire utilisables quelle que soit la langue cible sans adaptation particulière et de traitements facultatifs spécifiques c'est-à-dire à la fois optionnels et adaptés à la langue cible. Nous évaluons chacune des étapes et comparons notre méthode à celle de (Nasiruddin *et al.*, 2015) en l'appliquant au français et sur une tâche de désambiguïsation lexicale.

2 Contexte du travail

L'approche de (Nasiruddin *et al.*, 2015) consiste à traduire le SemCor et à porter ses annotations grâce à un système de traduction de l'anglais (une langue riche en corpus annotés) vers le français. Parmi les 714 759 mots du SemCor français traduit, nous avons remarqué la présence d'environ 5,84% mots non traduits ainsi que de nombreuses duplications. Nous avons repris leur code et leur version du *SemCor* en français, librement disponibles, et cherché à les améliorer.

2.1 SemCor

Le SemCor (Miller, 1995) est un sous-ensemble du corpus anglais Brown (Kucera & Francis, 1979) qui contient 234 000 mots annotés au niveau sémantique grâce au 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.

Le SemCor utilise le format de balisage structuré SGML (Standard Generalized Markup Language) normalisé et publié par l'ISO en 1986, et qui permet d'encoder le contenu d'un texte en attribuant des balises qui délimitent et indentifient chacun des éléments du texte. Par exemple dans le SemCor, la balise <s snum="id"> identifie une phrase, un mot est encodé entre <wf> et </wf> etc.

3 Méthode mise en œuvre

Dans cette section nous présentons la méthode mise en œuvre afin d'avoir une meilleure traduction possible du SemCor et porter ses annotations de l'anglais vers le français. La méthode de portage des annotations proposée par (Nasiruddin *et al.*, 2015) est très proche de celle de (Tiedemann *et al.*, 2014) dont nous avons eu connaissance plus tard. De plus, notre approche traite aussi le problème d'alignement en mots lié au transfert d'annotation. Comme nous le verrons dans la section 3.3 , elle est composée de trois étapes principales : le pré-traitement du SemCor anglais, la traduction et le portage de ses annotations vers le français, ainsi que le post-traitement appliqué sur les traductions produites. Chacune correspond à un script écrit en python. Les corpus et les scripts qui ont rendu possibles les travaux décrits dans cet article sont disponibles à l'adresse `https://github.com/getalp/WSD-TALN2016-Hadjsalahetal`.

3.1 Pré-traitement du SemCor

Parmi les données du SemCor nous trouvons des termes qui doivent être normalisés pour être exploités et adaptés aux données d'entraînement de notre système de TAS : des mots composés avec tiret bas, des mots non tokenisés, des mots commençant par une majuscule au début d'une phrase, ... Autant de mots non traités dans (Nasiruddin *et al.*, 2015).

Comme pré-traitement du SemCor, nous appliquons d'abord des traitements spécifiques pour le système de traduction puis, nous effectuons une normalisation adaptée au corpus SemCor anglais qui fait intervenir trois étapes :

— Segmenter les mots composés (effacer le tiret bas)
Federal_People's_Republic_of_Yugoslavia -> Federal People's Republic of Yugoslavia
— Appliquer la tokenisation Moses (ajouter des espaces entre mots et ponctuation :
People's rights. -> People 's rights .
— Mettre chaque mot du corpus dans une balise en suivant le format SGML du Semcor en lui affectant un id unique.

La Figure 1 présente un exemple de normalisation appliquée au mot composé "foster_homes" :

```
<wf cmd=done pos=NN lemma=foster_home wnsn=1 lexsn=1:14:00::>foster_homes</wf>
                                    ⇒
<wf id=15.1 cmd=done pos=NN lemma=foster_home wnsn=1 lexsn=1:14:00::>foster</wf>
<wf id=15.2 cmd=done pos=NN lemma=foster_home wnsn=1 lexsn=1:14:00::>homes</wf>
```

FIGURE 1 – Exemple de normalisation du mot composé "foster_homes"

3.2 Traduction et portage des annotations

Le système de traduction automatique statistique anglais-français que nous avons utilisé pour traduire le SemCor et porter ses annotations dans la langue cible, a été construit grâce à la boite à outils Moses (Hoang & Koehn, 2008) par (Besacier *et al.*, 2012), en exploitant l'ensemble 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).

La traduction d'un corpus annoté est rendue possible par *Moses* (Koehn *et al.*, 2007) qui permet d'obtenir l'alignement des mots cible-source. Comme (Nasiruddin *et al.*, 2015), nous exploitons ces données pour transférer les annotations d'un mot source à son correspondant dans le texte cible. La Figure 2 présente un exemple de traduction et de portage des annotations en français d'un mot composé pré-traité.

```
<wf cmd="done" pos="NN" lemma="foster_home" wnsn="1" lexsn="1:14:00::" id=15.1>des</wf>
<wf cmd="done" pos="NN" lemma="foster_home" wnsn="1" lexsn="1:14:00::" id=15.1>d'</wf>
<wf cmd="done" pos="NN" lemma="foster_home" wnsn="1" lexsn="1:14:00::" id=15.1>accueil</wf>
<wf cmd="done" pos="NN" lemma="foster_home" wnsn="1" lexsn="1:14:00::" id=15.2>foyers</wf>
```

FIGURE 2 – Traduction en français et portage des annotations du mot composé "foster_homes"

3.3 Post-traitement

L'exemple précédant montre clairement l'un des problèmes posé par l'étape de traduction. Il est nécessaire de passer par quelques étapes de post-traitement pour obtenir des résultats les plus pertinents possibles.

Ainsi, l'outil de portage d'annotations de (Nasiruddin *et al.*, 2015) produit parfois des traductions mal ordonnées (voir Figure 2) ou dupliquées (voir Figure 7) car il se base seulement sur les alignements en mots fournis par le décodeur (voir Figure 6) et en aucun cas sur la sortie de traduction. Ainsi, afin de résoudre ces problèmes, nous avons développé un outil permettant de compiler une chaîne de post-traitement sur la sortie de traduction, et qui enchaîne les trois étapes suivantes :

— Réordonnancement et suppression des mots ajoutés par Moses : afin de détecter les erreurs d'alignement en mot fournis par Moses, nous avons commencé par vérifier le positionnement de chaque mot annoté suivant la cible de traduction. Ainsi, si nous détectons un problème de désordre de mots, nous faisons le réordonnancement sinon, nous supprimons le mot ajouté. La Figure 3 présente un exemple de réordonnancement d'une traduction mal ordonnée (Figure 2).

Source : *[...] granted for child welfare services in foster homes .*

Cible : *[...] accordés pour les services de protection de l'enfance dans des foyers d'accueil .*

```
<wf cmd="done" pos="NN" lemma="foster_home" wnsn="1" lexsn="1:14:00::" id="15">des</wf>
<wf cmd="done" pos="NN" lemma="foster_home" wnsn="1" lexsn="1:14:00::" id="15">foyers</wf>
<wf cmd="done" pos="NN" lemma="foster_home" wnsn="1" lexsn="1:14:00::" id="15">d'</wf>
<wf cmd="done" pos="NN" lemma="foster_home" wnsn="1" lexsn="1:14:00::" id="15">accueil</wf>
```

FIGURE 3 – Réordonnancement des mots suivant la cible

— Concaténation : suite à la tokenisation des mots composés (ayant le même id) dans la chaine de pré-traitement du SemCor, nous avons concaténé ces derniers afin d'obtenir des id uniques. Dans la Figure 4, nous donnons un exemple précis de concaténation de mots ayant l'id=15 après le réordonnacement de la traduction en français du mot composé "foster_homes"

```
<wf cmd="done" pos="NN" lemma="foster_home" wnsn="1" lexsn="1:14:00::" id="15">des foyers d' accueil</wf>
```

FIGURE 4 – Concaténation des mots ayant le même id

— Segmentation : utiliser l'outil *Treetagger* (Schmid, 1995) qui nous permet d'avoir pour chaque mot, l'information grammaticale correspondante (POS), afin de mieux segmenter les suites de mots qui commencent ou se terminent par des conjonctions, déterminants, prépositions etc.

L'étape de post-traitement finale (Figure 5) appliquée sur la sortie de traduction du mot composé "foster_homes", est de filtrer *a posteriori* le résultat produit par l'outil Treetagger pour le déterminant "des" et de l'identifier suivant les annotations du SemCor anglais.

```
<wf cmd="ignore" pos="DT">des</wf>
<wf cmd="done" pos="NN" lemma="foster_home" wnsn="1" lexsn="1:14:00::" id="15">foyers d'accueil</wf>
```

FIGURE 5 – Exemple d'élimination du déterminant "des" à l'aide de Treetagger

Par ailleurs, comme il est indiqué précédemment, l'outil de portage d'annotation (Nasiruddin *et al.*, 2015) produit non seulement des sorties de TA mal ordonnées, mais aussi des traductions parfois dupliquées.
Si nous prenons l'exemple de la Figure 7, le mot composé anglais "city_council" qui n'existe pas dans le vocabulaire de notre système de traduction et qui a été (1) pré-traité, (2) traduit vers le français et (3) post-traité comme suit :

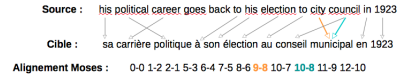

FIGURE 6 – Illustration du processus d'alignement

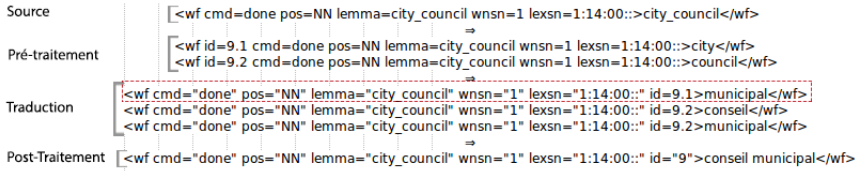

FIGURE 7 – Étapes de traduction et portage des annotations en français du mot composé "city_council"

La Figure 7 montre que notre outil de post-traitement a détecté que la séquence de mots "*municipal conseil municipal*" n'existe pas dans la sortie de traduction, et que le mot "*municipal*" a été dupliqué à cause des informations d'alignement fournies par Moses. Ainsi, il a cherché la séquence de mots exact dans la cible et a supprimé le mot dupliqué ayant l'identifiant *id=9.1* .

4 Évaluation

Nous effectuons une évaluation similaire à celle de (Nasiruddin *et al.*, 2015), c'est-à-dire en se basant sur une tâche de désambiguïsation lexicale et en reprenant le même protocole : les mêmes attributs (collocations locales, catégories grammaticales, mots du cotexte sur la même fenêtre de $-3, 3$) et le même classifieur, un classifieur bayésien naïf. En revanche, notre méthode d'évaluation est plus générique car elle est basée sur la forme de surface de mots et non pas sur les lemmes. En effet, il est plus simple de désambiguïser des textes bruts sans exiger qu'ils soient lemmatisés.

4.1 Corpus de Semeval 2013 tâche 12 : désambiguïsation lexicale multilingue

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). Pour l'évaluation, nous n'utilisons ici que la partie française.

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 $F_1 = \frac{2 \cdot P \cdot R}{P+R}$.

Puisque *SemCor* est annoté avec les sens de *Princeton WordNet* et que l'évaluation se fait avec des sens BabelNet, (Nasiruddin *et al.*, 2015) a réalisé une conversion en utilisant les alignements de BabelNet avec WordNet, conversion que nous réutilisons ici pour réaliser l'apprentissage sur nos corpus d'évaluation.

4.2 Résultats et Analyse

Nous avons réalisé un apprentissage sur le corpus SemCor traduit en français à chaque étape de post-traitement. Ainsi, nous pouvons analyser les performances de chacune.

- (Nasiruddin *et al.*, 2015) : le système de désambiguïsation est appris sur le corpus fourni dans la page compagnon de l'article ;
- WSD-INI : Le système est construit à partir du corpus issu de la traduction et du portage des annotations ; Contrairement au corpus précédant, une étape de pré-traitement (normalisation des données) a été réalisée ;
- WSD-REO : le système est construit à partir du corpus pour lequel a été effectué le réordonnancement ;
- WSD-Conc : le système est construit à partir du corpus pour lequel ont été effectués le réordonnancement et la concaténation ;
- WSD-Seg : le système est construit à partir du corpus pour lequel ont été effectués le réordonnancement, la concaténation ainsi que le post-traitement spécifique à la langue, l'étiquetage morpho-syntaxique.

	Précision	Rappel	Score F1
(Nasiruddin *et al.*, 2015)	49,6%	49,5%	49,55%
WSD-Ini	51,8%	51,6%	51,7%
WSD-Reo	52,2%	52,1%	52,15%
WSD-Conc	52,5%	52,4%	52,45%
WSD-Seg	52,2%	52,1%	52,15%

TABLE 1 – Performances des systèmes appris sur les différentes versions du corpus

La Table 1 présente les résultats des différents systèmes de désambiguïsation. Chacun est construit avec une version française du *SemCor*.

Rappelons, tout d'abord que la méthode d'évaluation de (Nasiruddin *et al.*, 2015) est basée sur les lemmes. Aussi pour pouvoir comparer nos résultats, nous avons repris l'expérience en utilisant ici notre méthode d'évaluation basée sur la forme de surface des mots qui est plus générique.

On peut remarquer tout d'abord, que le système WSD-Ini est meilleur que celui de (Nasiruddin *et al.*, 2015), avec une augmentation de 2,15 points en terme de score F1 ce qui signifie que notre méthode de pré-traitement du SemCor anglais est efficace. L'étape de réordonnancement et celle de concaténation permettent d'atteindre une meilleure performance à 52,45% (+2,9 points par rapport à la ligne de base). En revanche, l'étape spécifique à la langue, qui consiste à extraire certains mots outils dégrade légèrement les performances (-0,3 points). La différence de performance s'explique par un contexte d'analyse différent dû à la suppression de certains des mots outils dans la fenêtre d'analyse.

Une étude détaillée des raisons exactes n'est pas triviale vu le nombre important d'informations à vérifier manuellement. Il est vraisemblable qu'une étude d'attributs plus pertinents pour le français serait intéressante à mener, une fenêtre plus importante à gauche est, par exemple, une piste logique au vu de la perte de performance à la suite de la dernière étape.

5 Conclusion et perspectives

Dans cet article, nous avons présenté notre méthode pour améliorer la traduction d'un corpus annoté en sens vers une autre langue, méthode qui améliore fortement celle de (Nasiruddin *et al.*, 2015). Ensuite nous avons évalué nos résultats grâce à une méthode de désambiguïsation automatique supervisée. Dans l'avenir, nous envisageons d'utiliser cette méthode pour la création de corpus annotés pour différentes langues, en traduisant des corpus annotés et en portant leurs annotations, l'ensemble de ces corpus et les scripts permettant de les réaliser seront disponibles pour la communauté. Enfin, nous envisageons de mener une étude contrastive des attributs pertinents pour la désambiguïsation d'une langue (Français, Arabe), étude qu'il n'était pas possible de mener avant l'existence de ces corpus pour autant de langues.

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.

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.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R. *et al.* (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, p. 177–180 : Association for Computational Linguistics.

KUCERA H. & FRANCIS W. (1979). A standard corpus of present-day edited american english, for use with digital computers (revised and amplified from 1967 version).

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

NASIRUDDIN M., TCHECHMEDJIEV A., BLANCHON H. & SCHWAB D. (2015). Création rapide et efficace d'un système de désambiguïsation lexicale pour une langue peu dotée. In *TALN 2015 - 22ème Conférence sur le Traitement Automatique des Langues Naturelles,*, Caen, France.

SCHMID H. (1995). Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, **43**, 28.

TIEDEMANN J., AGIĆ Ž. & NIVRE J. (2014). Treebank translation for cross-lingual parser induction. In *Eighteenth Conference on Computational Natural Language Learning (CoNLL 2014)*.