La place de la désambiguïsation lexicale dans la Traduction Automatique Statistique

Marianna Apidianaki
National Centre for Language Technology, Dublin City University
mapidianaki@computing.dcu.ie

Résumé. L'étape de la désambiguïsation lexicale est souvent esquivée dans les systèmes de Traduction Automatique Statistique (Statistical Machine Translation (SMT)) car considérée comme non nécessaire à la sélection de traductions correctes. Le débat autour de cette nécessité est actuellement assez vif. Dans cet article, nous présentons les principales positions sur le sujet. Nous analysons les avantages et les inconvénients de la conception actuelle de la désambiguïsation dans le cadre de la SMT, d'après laquelle les sens des mots correspondent à leurs traductions dans des corpus parallèles. Ensuite, nous présentons des arguments en faveur d'une analyse plus poussée des informations sémantiques induites à partir de corpus parallèles et nous expliquons comment les résultats d'une telle analyse pourraient être exploités pour une évaluation plus flexible et concluante de l'impact de la désambiguïsation dans la SMT.

Abstract. Word Sense Disambiguation (WSD) is often omitted in Statistical Machine Translation (SMT) systems, as it is considered unnecessary for lexical selection. The discussion on the need of WSD is currently very active. In this article we present the main positions on the subject. We analyze the advantages and weaknesses of the current conception of WSD in SMT, according to which the senses of ambiguous words correspond to their translations in a parallel corpus. Then we present some arguments towards a more thorough analysis of the semantic information induced from parallel corpora and we explain how the results of this analysis could be exploited for a more flexible and conclusive evaluation of the impact of WSD on SMT.

Mots-clés : Désambiguïsation lexicale, Traduction Automatique Statistique, sélection lexicale.

Keywords: Word Sense Disambiguation, Statistical Machine Translation, lexical selection.

1 Introduction

Les systèmes de SMT actuels effectuent une désambiguïsation lexicale (Word Sense Disambiguation (WSD)) implicite : la sélection lexicale est basée sur le contexte local des mots, sans considération d'informations linguistiques plus riches, comme celles exploitées pour la WSD.[1] Les architectures de SMT fondées sur les mots réalisent la WSD en combinant des probabilités a priori sur les candidats de sens (traductions) à partir de critères de fluidité modélisés par les probabilités du modèle de langue. Les architectures de SMT basées sur les segments intègrent

[1]Ces informations peuvent être syntaxiques ou collocationnelles, relatives à la position des mots, etc. (Chan *et al.*, 2007; Carpuat & Wu, 2007).

en plus des préférences de collocation lexicale. Néanmoins, les problèmes observés au niveau de la sélection lexicale ont réanimé l'intérêt autour des modèles de sélection basés sur la WSD[2], capables d'incorporer des traits contextuels de la langue source (LS) plus riches que ceux pris en compte par les systèmes de SMT "état de l'art". Dans cet article, nous expliquons comment la WSD est actuellement conçue dans la SMT et analysons les inconvénients posés par cette conception. Finalement, nous présentons des arguments en faveur d'une tâche de WSD basée sur une analyse de la sémantique lexicale.

2 Désambiguïsation dans les systèmes de SMT

2.1 Impact négatif de la WSD sur la qualité de traduction

Le travail de Carpuat et Wu (2005) met en cause le statut de la WSD en tant qu'étape bénéfique au sein des systèmes de SMT. Dans les travaux ultérieurs sur la question (Cabezas & Resnik, 2005; Chan *et al.*, 2007), les auteurs prennent le soin de se positionner par rapport à ces résultats en clarifiant les facteurs qui les ont provoqués, voire même en critiquant le cadre dans lequel les expériences ont été menées. Les principales raisons de cette détérioration sont les suivantes :

1. Le petit volume de données d'entraînement du système de WSD utilisé.
2. L'absence de lien entre, d'une part, les données d'entraînement du système de SMT et, d'autre part, les données d'entraînement du système de WSD et l'inventaire de sens utilisé.
3. L'intégration des prédictions du système de WSD à l'aide de contraintes dures et la non intégration du modèle de WSD et de ses prédictions dans le modèle de traduction. [3]
4. L'utilisation du score BLEU (Papineni *et al.*, 2002) pour l'évaluation.

2.2 Impact positif de la WSD sur la qualité de traduction

Des travaux plus récents sur le sujet démontrent l'effet positif de la WSD dans la SMT. Dans ces travaux, la WSD n'est pas restreinte à un nombre précis de sens issus d'inventaires définis a priori (Cabezas & Resnik, 2005; Chan *et al.*, 2007; Stroppa *et al.*, 2007; Carpuat & Wu, 2007). Au contraire, les sens des mots sont désignés par leurs équivalents de traduction (EQVs), repérés au sein du corpus d'entraînement du système de SMT. Ainsi, les systèmes de WSD et de SMT sont entraînés sur les mêmes corpus.

En outre, les prédictions de WSD sont intégrées dans les systèmes de SMT de manière dynamique : elles constituent des alternatives que le décodeur prend en compte en même temps que celles proposées par le modèle de traduction (Cabezas & Resnik, 2005). Le décodeur n'est donc pas contraint de sélectionner la traduction proposée par le système de WSD et la sélection finale est effectuée par le modèle de langue.

Il faut noter que les cibles de désambiguïsation dans ces travaux correspondent soit à des mots (Cabezas & Resnik, 2005; Carpuat *et al.*, 2006) soit à des segments (Stroppa *et al.*, 2007; Chan

[2]La première étude sur le sujet (Brown *et al.*, 1991) avait été encourageante mais limitée.

[3]Les prédictions sont exploitées soit en forçant le décodeur à effectuer un choix, soit en remplaçant les traductions choisies par le système de SMT lors d'une étape de post-traitement.

et al., 2007; Carpuat & Wu, 2007). L'unité de base de la sélection lexicale étant le segment et non le mot dans la plupart des systèmes de SMT actuels (Och & Ney, 2004; Koehn, 2004; Chiang, 2005), la désambiguïsation des segments est supposée adapter la WSD à la tâche à laquelle les systèmes de SMT sont confrontés, afin qu'ils puissent tirer profit de ses avantages (Carpuat & Wu, 2007). Dans ce cas aussi, les sens candidats des segments source sont leurs traductions repérées dans le corpus d'entraînement du système de SMT.

2.3 Assimilation des tâches de WSD et de sélection lexicale

2.3.1 Avantages

Lorsque les sens des mots ambigus sont désignés à l'aide de leurs EQVs, l'identification du sens d'une nouvelle instance coïncide avec la sélection de sa traduction. L'assimilation de ces deux tâches présente un certain nombre d'avantages :

1. Elle permet de dissocier les techniques de WSD d'inventaires de sens prédéfinis, qui peuvent contenir des sens non liés aux données d'entraînement du système de SMT (Resnik, 2007).
2. Elle met l'accent sur les distinctions sémantiques pertinentes pour la traduction entre les langues impliquées et permet d'éviter le repérage de distinctions de granularité trop fine.
3. Elle permet de tirer profit de la grande disponibilité de données étiquetées sous la forme de corpus parallèles alignés. L'étiquetage des mots source par leurs EQVs augmente fortement les données d'entraînement pour les algorithmes supervisés de WSD, dans la mesure où il y a davantage de corpus parallèles que de textes sémantiquement étiquetés.
4. Elle lie la WSD et son évaluation à la traduction, ce qui permet de résoudre le problème de la non-conformité, dans ce cadre, des méthodes de WSD monolingues ou visant d'autres applications.[4]

Cette manière de concevoir la WSD la rapproche des tâches multilingues de Senseval (Chklovski *et al.*, 2004) et de Semeval (Jin *et al.*, 2007), où les inventaires utilisés pour la WSD représentent des distinctions sémantiques effectuées dans d'autres langues.

2.3.2 Inconvénients théoriques

Malgré ses avantages, l'assimilation de la WSD et de la sélection lexicale présente également un ensemble d'inconvénients. Le souci de rendre l'inventaire de sens aussi relatif que possible au corpus d'entraînement du système de SMT devient tellement important que la pertinence théorique de l'approche adoptée n'est pas explorée. En effet, la considération de correspondances biunivoques (de type un-à-un) entre sens et EQVs paraît simpliste et peut facilement être mise en question. Chaque EQV est considéré comme indiquant un sens distinct et l'éventuelle similarité sémantique des EQVs n'est pas prise en compte. Ainsi, les sens sont simplement énumerés et considérés comme équivalents, et leurs relations ne sont pas décrites.

Cette approche néglige la complexité des relations entre sens et EQVs : un sens peut être traduit par des EQVs sémantiquement apparentés dans la langue cible (LC) ou, inversement, un EQV

[4]En raison des besoins divergents des applications concernant tant le degré de WSD, que le type et le niveau des distinctions sémantiques.

peut traduire des sens différents. En outre, il se peut que certains EQVs ne constituent pas de bons indices de sens, dans la mesure où ils peuvent réfléter des distinctions sémantiques propres à la LC ou présenter la même ambiguïté que le mot source.[5] Par conséquent, une analyse plus poussée de la sémantique des EQVs et de leurs éventuelles relations serait requise.

En ce qui concerne la désambiguïsation des segments, elle est considérée comme une véritable tâche de désambiguïsation. Néanmoins, les segments utilisés dans les systèmes de SMT "état de l'art" ne sont pas linguistiquement motivés (Koehn *et al.*, 2003). Il s'agit de séquences de mots (par ex. *house the*) qui sont considérées comme de bons segments parce que leurs mots sont alignés seulement entre eux et non à des mots extérieurs aux segments. Ces segments n'ont donc pas de statut linguistique défini, ce qui rend la définition de leurs sens discutable. La mention explicite à la WSD dans la SMT ne se justifie que par l'utilisation d'informations contextuelles étendues de la LS, permettant d'intégrer des probabilités contextuelles relatives aux traductions.

2.3.3 Inconvénients pratiques

Les sens induits par cette approche inter-langue d'induction de sens ont le même statut, ce qui a comme résultat leur traitement uniforme et l'application des mêmes contraintes pour la sélection lors de la WSD. En outre, lorsque la WSD coïncide avec la sélection lexicale, son évaluation obéit aux mêmes principes que l'évaluation de la TA : la sélection parmi les candidats de sens d'un mot pour une nouvelle instance est considérée comme correcte si le sens choisi correspond à la traduction de référence[6]. Etant donné que chaque EQV représente un sens distinct, la sélection d'un EQV différent de la référence est directement considérée comme fausse. Cette conception ne permet pas de prendre en compte l'importance des erreurs de WSD et de sélection lexicale.[7] Dans le cadre de la TA, Resnik et Yarowsky (1997) proposent de pénaliser uniquement les distinctions sémantiques lexicalisées différemment dans la LC. Nous estimons que pour une évaluation concluante il faudrait, en outre, considérer la distance entre les sens lexicalisés par des EQVs de traduction différents.

3 Vers une analyse de la sémantique lexicale

La conception actuelle de la WSD dans la SMT pose des problèmes aux niveaux de son intégration et de son évaluation. L'évaluation devient encore plus difficile en raison de l'incapacité des métriques d'évaluation de la TA à identifier des correspondances de sens et donner ainsi une image claire de l'impact de la WSD sur la traduction. BLEU (Papineni *et al.*, 2002) propose l'utilisation de références multiples, ce qui entraîne d'autres inconvénients (Callison-Burch *et al.*, 2006). METEOR traite la synonymie en considérant comme correctes des traductions sémantiquement similaires à la référence (trouvées dans le même synset de WordNet) (Lavie & Agarwal, 2007). Néanmoins, l'algorithme de détection de la synonymie met en correspondance les mots sans les désambiguïser. En outre, la méthode est fortement dépendante d'un inventaire

[5] Les ambiguïtés parallèles peuvent ne pas être résolues pendant la traduction, mais elles devraient être prises en considération dans des applications comme la recherche d'information multilingue.

[6] La traduction du mot dans le corpus d'évaluation (corpus de test).

[7] Les erreurs devraient être pénalisées en fonction de la granularité des distinctions sémantiques concernées : la classification erronée entre sens proches devrait être moins pénalisée que celle entre sens distants (Resnik & Yarowsky, 1997).

de sens prédéfini, besoin qui restreint son application aux langues disposant de telles ressources (Lavie & Agarwal, 2007).

Nous estimons qu'une analyse poussée des informations sémantiques extraites à partir de corpus parallèles rapprocherait davantage la WSD dans la SMT à une vraie tâche de WSD. Cette analyse pourrait être effectuée en appliquant une méthode d'apprentissage non supervisé sur le corpus d'entraînement du système de SMT. Une méthode de clustering sémantique permettant l'élaboration automatique d'inventaires de sens est proposée dans Apidianaki (2008). Cette méthode regroupe les EQVs des mots ambigus dans des clusters en fonction de leur similarité sémantique, qui correspond à la similarité distributionnelle des instances des mots ambigus qu'ils traduisent. Ainsi les EQVs sémantiquement proches et n'indiquant pas des sens distincts des mots ambigus sont distingués des EQVs qui traduisent des sens distincts, et qui constituent des indices pertinents de sens. Les sens des mots ambigus sont donc décrits, au sein de l'inventaire généré, par des *clusters de sens* regroupant leurs EQVs.

L'exploitation de cet inventaire de sens peut lier la WSD à la sélection lexicale, tout en la rendant sensible à la sémantique, et s'avère bénéfique au niveau de l'évaluation de la WSD, qui peut prendre en compte le phénomène de la variation lexicale dans la traduction (Apidianaki, 2009). Cet intérêt vers l'utilisation de clusters de sens acquis à partir de corpus est actuellement manifesté dans le domaine de l'évaluation de la WSD. Contrairement aux tâches de WSD multilingue des campagnes Senseval et Semeval07 – qui considéraient chaque EQV comme indice d'un sens distinct – dans la tâche correspondante proposée pour Semeval-2010 les sens seront décrits par des clusters d'EQVs sémantiquement proches, manuellement construits à partir d'un corpus parallèle[8].

Références

APIDIANAKI M. (2008). Translation-oriented Word Sense Induction based on Parallel Corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, p. 3269–3275, Marrakech, Morocco.

APIDIANAKI M. (2009). Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, p. 77–85, Athens, Greece.

BROWN P. F., COCKE J., PIETRA S. D., PIETRA V. J. D., JELINEK F., MERCER R. L. & ROOSSIN P. S. (1988). A statistical approach to language translation. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING)*, p. 71–76, Budapest, Hungary.

BROWN P. F., PIETRA S. A. D., PIETRA V. J. D. & MERCER R. L. (1991). Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the ACL*, p. 264–270, Berkeley, California.

CABEZAS C. & RESNIK P. (2005). *Using WSD Techniques for Lexical Selection in Statistical Machine Translation*. Rapport interne LAMP-TR-124,CS-TR-4736,UMIACS-TR-2005-42, University of Maryland, College Park.

CALLISON-BURCH C., OSBORNE M. & KOEHN P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL)*, p. 249–256, Trento, Italy.

[8] `http://webs.hogent.be/~elef464/lt3_SemEval.html`

CARPUAT M. & WU D. (2005). Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, p. 387–394, Ann Arbor, Michigan.

CARPUAT M. & WU D. (2007). How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In *Proceedings of the 11th International TMI Conference*, p. 43–52, Skovde, Sweden.

CARPUAT M., YU Y. X. & WU D. (2006). Towards Integrating Word Sense and Entity Disambiguation into Statistical Machine Translation. In *Proceedings of the International Workshop on Spoken Language Translation*, p. 37–44, Kyoto, Japan.

CHAN Y. S., NG H. T. & CHIANG D. (2007). Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL*, p. 263–270, Prague, Czech Republic.

CHIANG D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, p. 263–270, Ann Arbor, Michigan.

CHKLOVSKI T., MIHALCEA R., PEDERSEN T. & PURANDARE A. (2004). The SENSEVAL-3 multilingual English-Hindi lexical sample task. In *Proceedings of the 3rd International Workshop on Evaluating Word Sense Disambiguation Systems*, p. 5–8, Barcelona, Spain.

JIN P., WU Y. & YU S. (2007). SemEval-2007 Task 05 : Multilingual Chinese-English Lexical Sample. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, p. 19–23, Prague, Czech Republic.

KOEHN P. (2004). Pharaoh : A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA)*, p. 115–124, Washington, DC.

KOEHN P., OCH F. J. & MARCU D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology and North American ACL Conference (HLT/NAACL)*, p. 48–54, Edmonton, Canada.

LAVIE A. & AGARWAL A. (2007). METEOR : An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, p. 228–231, Prague, Czech Republic.

OCH F. J. & NEY H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, **30**(4), 417–449.

PAPINENI K., ROUKOS S., WARD T. & JING ZHU W. (2002). BLEU : A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, p. 311–318, Philadelphia, PA, USA.

RESNIK P. (2007). *WSD in NLP applications*, In E. AGIRRE & P. EDMONDS, Eds., *Word Sense Disambiguation : Algorithms and Applications*, p. 299–337. Springer.

RESNIK P. & YAROWSKY D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of SIGLEX Workshop "Tagging Text with Lexical Semantics : What, why and how ?"*, p. 79–86, Washington, D.C.

RESNIK P. & YAROWSKY D. (2000). Distinguishing systems and distinguishing senses : New evaluation methods for word sense disambiguation. *Natural Language Engineering*, **5**(3), 113–133.

STROPPA N., VAN DEN BOSCH A. & WAY A. (2007). Exploiting Source Similarity for SMT using Context-Informed Features. In *Proceedings of the 11th International TMI Conference*, p. 231–240, Skovde, Sweden.