TALN 2009, Senlis, 24–26 juin 2009

Données bilingues pour la TAS français-anglais : impact de la langue source et direction de traduction originales sur la qualité de la traduction

Sylwia Ozdowska
National Centre for Language Technology – Dublin City University
Glasnevin, Dublin 9, Ireland
sozdowska@computing.dcu.ie

Résumé. Dans cet article, nous prenons position par rapport à la question de la qualité des données bilingues destinées à la traduction automatique statistique en terme de langue source et direction de traduction originales à l'égard d'une tâche de traduction français-anglais. Nous montrons que l'entraînement sur un corpus contenant des textes qui ont été à l'origine traduits du français vers l'anglais améliore la qualité de la traduction. Inversement, l'entraînement sur un corpus contenant exclusivement des textes dont la langue source originale n'est ni le français ni l'anglais dégrade la traduction.

Abstract. In this paper, we argue about the quality of bilingual data for statistical machine translation in terms of the original source language and translation direction in the context of a French to English translation task. We show that data containing original French and English translated from French improves translation quality. Conversely, using data comprising exclusively French and English translated from several other languages results in a clear-cut decrease in translation quality.

Mots-clés : Traduction automatique statistique, corpus bilingue, direction de la traduction, langue source, langue cible.

Keywords: Statistical machine translation, bilingual corpus, translation direction, source language, target language.

1 Introduction

La traduction automatique statistique (TAS) s'appuie sur l'idée qu'il est possible d'apprendre à partir de corpus les données nécessaires à la traduction automatique sous forme d'un modèle de langue et d'un modèle de traduction. Le modèle de langue est estimé à partir d'un corpus monolingue dans la langue vers laquelle s'effectue la traduction. Le modèle de traduction est estimé à partir d'un corpus bilingue qui contient des textes en relation de traduction. Aucune importance ne semble être accordée à l'adéquation entre la direction de traduction représentée dans le corpus d'apprentissage et la direction de traduction dans laquelle le système est destiné à opérer. Dans cet article, nous nous positionnons par rapport la pertinence de cette question dans le cadre de l'entraînement d'un système de TAS traduisant du français vers l'anglais. Nous faisons le point sur la composition des corpus bilingues en terme de langue originale et de direction de traduction (section 2) avant de voir quel écho ces notions trouvent dans le contexte

de la recherche en TAS et dans celui des études sur la traduction (section 3). Ensuite, nous situons notre cadre expérimental (section 4) et présentons les résultats d'expériences (section 5) sur lesquels nous nous appuyons pour prendre position et conclure (section 6).

2 Corpus bilingue

Du point de vue des langues en présence dans le corpus, dans le cas le plus simple le corpus bilingue est composé de textes qui ont été originellement traduits de la langue A vers la langue B de sorte que la partie B du corpus est la traduction directe de la partie A. On dira dans ce cas que A est la langue source (LS) originale, B étant la langue cible (LC) par rapport à A. Dans des cas plus complexes, il peut arriver que certains textes aient été à l'origine traduits de la langue A vers la langue B et certains autres dans la direction opposée, à savoir de B vers A, et/ou que d'autres langues aient été impliquées dans le processus de traduction de sorte que tout ou partie des textes en langue A et tout ou partie des textes en langue B ont été traduits à partir d'une ou plusieurs autres langues. Dans ce cas, la partie du corpus correspondant à la langue A (resp. B) mélange LS originale A (resp. B), langue A (resp. B) cible de B (resp. A) et langue A (resp. B) cible d'une autre langue que B (resp. A). Les corpus multilingues de langues européennes comme Europarl (Koehn, 2005)[1] ou l'Acquis Communautaire (Steinberger *et al.*, 2006)[2], largement exploités dans le domaine de la TAS, relèvent de ce cas de figure. Ils rassemblent en effet des textes produits au sein d'institutions européennes ; ces textes peuvent avoir été formulés à l'origine dans n'importe laquelle des langues officielles avant d'être traduits dans toutes les autres langues.

Du point de vue de la TAS, étant donné un corpus bilingue contenant des textes en langue A et une version correspondante en langue B, l'entraînement d'un système visant à traduire de la langue A vers la langue B s'effectue en posant A comme LS et B comme LC indépendemment du statut de ces deux langues dans le processus de traduction original (LS originale et LC en relation de traduction directe, langues cibles d'une LS tierce en relation de traduction indirecte) et donc dans le corpus d'entraînement. Autremement dit, pour apprendre à traduire de A vers B, un système de traduction ne s'appuie pas nécessairement sur des textes effectivement traduits de A vers B. Cela revient à dire que la qualité de la traduction automatique d'une langue A vers une langue B ne serait pas corrélée avec le statut effectif des langues A et B dans le corpus d'apprentissage, *i.e.* la direction de traduction originale telle que reflétée dans le corpus. Cette non prise en compte de la composition des corpus bilingues en terme du statut des langues transparaît dans l'absence de réferences à la question dans le contexte de la recherche en TAS.

3 Travaux apparentés

Bien que la thématique de la langue originale *vs.* langue de traduction, et plus largement celle de la directionnalité et reversibilité du processus traduction, occupe une place importante dans les travaux sur la traduction, elle semble avoir été presque complètement négligée dans le domaine de la recherche en TAS. Dans ce dernier, les seuls travaux que l'on pourrait rattacher à la question posée dans cet article sont ceux portant sur la mise au point de systèmes de TAS par pivot, notamment dans le cas de langues pour lesquelles il n'existe que peu ou pas de données

[1] `http://www.iccs.inf.ed.ac.uk/\~pkoehn/publications/europarl/`
[2] `http://wt.jrc.it/lt/Acquis/`

bilingues (Wu & Wang, 2007). Le principe est de construire un modèle de traduction pour A et B à partir d'un modèle de traduction pour A et P et d'un modèle de traduction pour P et B. Bien qu'il n'en soit fait aucune mention, ce type de travaux semble suggérer que pour traduire entre deux langues, l'existence d'un lien de traduction direct (et orienté dans le même sens que celui visé par la traduction) entre ces langues dans les données d'apprentissage n'est pas essentiel.

Dans le contexte des recherches sur la traduction, W. Teubert (1996) souligne qu'étant donné un texte en langue A et ses traductions en langues B et C, chacune de ces dernières est susceptible de présenter plus de similitudes (syntaxiques, lexicales et sémantiques) avec la version A qu'il n'y en aurait entre elles. Globalement, il semble y avoir un consensus par rapport au fait que les textes traduits ne devraient pas être considérées comme des ressources bidirectionnelles et réversibles (Bowker, 2003), ce qui laisse à penser que le statut des langues dans les données bilingues pourrait avoir un impact sur la qualité de la TAS.

La question de l'adéquation des données d'apprentissage vis a vis de la tâche de traduction visée du point de vue du lien de traduction (direct ou indirect) entre les langues en présence et de son orientation semble donc légitime. À cet égard, l'hypothèse en faveur de laquelle nous argumentons et prenons position en nous appuyant sur des résultats d'expériences en TAS va dans le sens des propos de W. Teubert mentionés ci-dessus. Nous pensons en effet que le recours à des données bilingues dans lesquelles le statut traductionnel des langues A et B est en adéquation avec la direction de traduction testée constituent les conditions d'entraînement optimales pour ce qui est de la qualité de la TAS de A vers B telle que mesurée automatiquement à l'aide de métriques quantitatives standard. Dans ce qui suit, la qualité des données n'est pas envisagée *per se* mais à l'égard des résultats d'une tâche de traduction de A vers B évaluée automatiquement ; nous partons donc du principe que des données qui améliorent les scores de traduction sont de meilleure qualité pour la tâche de traduction envisagée.

Pour vérifier cette hyptohèse et justifier notre prise de position, nous mettons en place un cadre expérimental qui nous permet de comparer plusieurs configurations d'un même système de TAS état de l'art à base de segments. Ces configurations ne diffèrent qu'en terme de la LS originale en fonction de laquelle sont selectionnées les données d'apprentissage et de test. La comparaison se fait sur la base de la qualité de la traduction du français vers l'anglais obtenue avec les différentes configurations et mesurée automatiquement.

4 Données et outils de l'expérience

Les expériences que nous présentons s'appuient sur une version des parties française et anglaise du corpus Europarl. Pour chaque couple de phrases alignées, nous disposons de l'information indiquant la langue dans laquelle la phrase a été originellement produite. Ainsi, sur les 1 391 222 bi-phrases du corpus, 164 648 au minimum ont été à l'origine traduites du français vers l'anglais et 235 102 au minimum dans le sens inverse[3]. À partir de ce corpus, nous avons extrait quatre sous-corpus selon différents critères quant à la LS originale et entraîné les configurations de TAS français-anglais correspondantes.

config-1 Aucune condition sur la LS originale n'a été imposée, autrement dit la partie française du sous-corpus et la partie anglaise correspondante contiennent respectivement:
– du français traduit d'une langue autre que l'anglais, du français traduit directement de l'anglais ainsi que du français original ;

[3]La LS originale n'est pas connue ou spécifiée pour toutes les bi-phrases.

– de l'anglais traduit d'une langue autre que le français, de l'anglais traduit directement de l'anglais ainsi que de l'anglais original.

config-2 La LS originale n'est ni le français ni l'anglais, *i.e.* le corpus contient :
– du français traduit d'une langue autre que l'anglais ;
– de l'anglais traduit d'une langue autre que le français.

config-3 La LS originale est l'anglais, *i.e.* le corpus contient :
– du français traduit directement de l'anglais ;
– de l'anglais original.

config-4 La LS originale est le français, *i.e.* le corpus contient :
– du français original ;
– de l'anglais traduit directement un français.

Chaque configuration est entraînée sur 100 000 phrases et testée sur 500 phrases. Les deux ensembles sont disjoints. Le même critère de sélection en terme de LS originale vaut pour les données d'entraînement et de test. Chaque configuration dispose donc de données de test individuelles qui lui sont propres et qui sont différentes de celles des autres configurations ; nous les appelons test-500. Dans un second temps, nous évaluons également chaque configuration sur un même jeu de test résultant de l'union des tests individuels et contenant 2000 phrases. Nous parlerons dans ce cas de test-2000.

Toutes nos expériences de traduction du français à l'anglais sont effectuées avec des outils état de l'art disponibles publiquement. Les bi-phrases de chaque corpus d'entraînement sont alignées au niveau des mots suivant le modèle IBM 4 (Brown *et al.*, 1993; Och & Ney, 2003) implémenté dans GIZA++[4]. Le modèle de traduction est extrait sur la base de ces alignements avec le système Moses (Koehn *et al.*, 2003; Koehn *et al.*, 2007)[5]. Le modèle de langue 5-grammes pour l'anglais est obtenu grâce à l'outil SRLIM (Stolcke, 2002)[6]. Le décodage est effectué avec Moses. La qualité de la traduction est évaluée automatiquement avec trois métriques standard : BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002) et METEOR (Banerjee & Lavie, 2005).

5 Résultats de l'expérience

Comme dit précédemment, nous nous intéressons à l'impact de l'utilisation en TAS de corpus bilingues dont la partie source en langue A contient effectivement des textes produits à l'origine dans la langue A, *i.e.* il y dans ce cas adéquation entre la direction de traduction originale et la direction de traduction pour laquelle le système est entraîné, *vs.* des corpus bilingues dans lesquels cette condition d'adéquation n'est que partiellement ou pas vérifiée.

Le tableau 1 présente les résultats de nos expériences sur test-500 et test-2000 (la référence est unique ; les scores maximum sont en gras et les scores minimum en italique). Nous analysons tout d'abord les résultats obtenus par chaque configuration sur son test-500 respectif. Les tendances observées restent les mêmes quels que soient le système et la métrique d'évaluation considérés. En terme de BLEU par exemple, on observe une considérable augmentation absolue en qualité de traduction, soit 0.0956, lorsque l'on passe de config-2 à config-4. Config-2, qui est la moins performante, correspond à un entraînement sur le corpus dont les deux parties

[4] http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/
[5] http://www.statmt.org/moses/
[6] http://www.speech.sri.com/projects/srilm/

	test-500			test-2000		
configuration	BLEU	NIST	METEOR	BLEU	NIST	METEOR
config-1	0.2608	5.9771	0.5758	**0.2542**	6.4797	0.5646
config-2	*0.2008*	*5.1531*	*0.4867*	*0.2424*	*6.3211*	*0.5525*
config-3	0.2857	6.4717	0.6082	0.2520	**6.5385**	0.5558
config-4	**0.2964**	**6.5502**	**0.6162**	0.2500	6.4331	**0.5681**

TAB. 1 – Évaluation sur les tests individuels test-500 et le test unique test-2000

française et anglaise sont traduites à partir de plusieurs langues tierces. Config-4, qui est la plus performante, a été entraînée sur un corpus dans lequel l'anglais est une traduction directe du français. Ces résultats tendent à confirmer notre hypothèse, à savoir que des données dont la LS originale est le français sont optimales pour entraîner un système de TAS du français vers l'anglais. Par ailleurs, il semblerait que le recours à des données comprenant du français et de l'anglais en relation de traduction indirecte constituent les moins bonnes conditions d'entraînement. On peut supposer que le statut de langue traduite, qui de plus est à partir de différentes langues sources, confère au corpus des caractéristiques particulières, par exemple en termes de régularités et divergeances entre les deux langues en présence, qui rendent la généralisation sur les données plus ardue et affecte donc la qualité de l'apprentissage.

Si on regarde maintenant les résultats sur test-2000, on remarque premièrement que les différentes métriques donnent des résultats conflictuels. Le seul résultat cohérent d'une métrique à l'autre, ainsi qu'avec les évaluations individuelles, est celui concernant config-2 qui obtient les moins bonnes performances, à savoir 0.2424 BLEU. Pour ce qui est de config-4, elle se classe première uniquement en fonction de METEOR. Autrement dit, config-3 produit de meilleurs résultats que config-4 si on ne tient pas compte de METEOR et config-1 produit de meilleurs résultats que config-3 si on ne tient pas compte de NIST. Cependant, les meilleurs scores sur test-2000 restent inférieurs aux scores obtenus par config-4 sur son test individuel, par exemple 0.2542 *vs.* 0.2964 pour le meilleur BLEU.

6 Conclusion

Nous avons soulevé la question de la relation entre la qualité des données bilingues destinées à la TAS en terme de langue source et direction de la traduction originales et la qualité de la traduction automatique français-anglais issue de l'entraînement sur ces données. Les résultats montrent que le recours à des données où les parties française et anglaise ne sont pas en relation de traduction directe dégradent la qualité de la traduction. Le recours à des données produites originellement dans la même direction de traduction que celle effectuée par le systéme s'avère optimal dans le cas de l'évaluation sur des tests individuels sélectionnés sur les mêmes critères de langue source originale que les données d'entraînement. Les résultats sur le test unique ne sont cependant pas conclusifs.

Les résultats obtenus suggèrent que les systèmes de TAS pourraient bénéficier d'un l'entraînement sur des données qui respectent la même direction de traduction que celle dans laquelle ils sont censés opérer. Même si elles indiquent une piste intéressante, les différences quantitatives dans les scores ne sauraient être réellement interprétées comme un mieux au niveau de la qualité de la traduction que grâce à une véritable étude qualitative des résultats, incluant une corrélation

avec des évaluations humaines, pour identifier et expliquer les phénomènes en jeu.

Jusque-là négligée, la question de l'impact de la *qualité* des données d'entraînement sur les performances de la TAS devrait selon nous être étudiée au même titre que celui de leur *quantité*. Elle ouvre en effet des perspectives de recherche intéressantes, autant dans de la domaine de la TAS que dans celui des études sur la traduction. On peut notamment se demander pourquoi, d'un point de vue traductionnel, la relation d'équivalence n'est pas tout à fait réversible ou encore pourquoi les systèmes de TAS sont sensibles à cette dyssymétrie.

Références

BANERJEE S. & LAVIE A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-05)*, p. 65–72, Ann Arbor, MT.

BOWKER L. (2003). Investigate 'reversible' translation resources: are they equally useful in both translation directions ? In L. P. GONZÁLES, Ed., *Speaking in Tongues: Language across Contexts and Users*, p. 201–224.

BROWN P. F., PIETRA S. A. D., PIETRA V. J. D. & MERCER R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, **19**(2), 263–311.

DODDINGTON G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Human Language Technology: Notebook Proceedings*, p. 128–132, San Diego, CA.

KOEHN P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, p. 79–86, Phuket, Thaïlande.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., B. COWAN W. S., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A., & HERBST E. (2007). Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, p. 177–180, Prague, République Tchèque.

KOEHN P., OCH F. & MARCU D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL'03)*, p. 48–54, Edmonton, Canada.

OCH F. J. & NEY H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, **1**(29), 19–51.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphie, États-Unis.

STEINBERGER R., POULIQUEN B., WIDIGER A., IGNAT C., ERJAVEC T., TUFIŞ D. & VARGA D. (2006). The JRC-Acquis: A multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, p. 2142–2147, Genoa, Italie.

STOLCKE A. (2002). SRILM: an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, p. 901–904, Denver, CO.

TEUBERT W. (1996). Comparable or Parallel Corpora ? *International Journal of Lexicography*, **9**(3), 239–264.

WU H. & WANG H. (2007). Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, **21**(3), 165–181.