Approches en corpus pour la traduction: le cas MÉTÉO

Philippe Langlais, Thomas Leplus
Simona Gandrabur et Guy Lapalme
RALI
Université de Montréal
http://rali.iro.umontreal.ca/

Mots-clefs : Mémoire de traduction, traduction probabiliste, alignements multiples, réordonnancement à postériori

Keywords: Memory-based translation, statistical translation, multiple alignment, rescoring

Résumé La traduction automatique (TA) attire depuis plusieurs années l'intérêt d'un nombre grandissant de chercheurs. De nombreuses approches sont proposées et plusieurs campagnes d'évaluation rythment les avancées faites. La tâche de traduction à laquelle les participants de ces campagnes se prêtent consiste presque invariablement à traduire des articles journalistiques d'une langue étrangère vers l'anglais; tâche qui peut sembler artificielle. Dans cette étude, nous nous intéressons à savoir ce que différentes approches basées sur les corpus peuvent faire sur une tâche réelle. Nous avons reconstruit à cet effet l'un des plus grands succès de la TA: le système MÉTÉO. Nous montrons qu'une combinaison de mémoire de traduction et d'approches statistiques permet d'obtenir des résultats comparables à celles du système MÉTÉO, tout en offrant un cycle de développement plus court et de plus grandes possibilités d'ajustements.

Abstract Machine Translation (MT) is the focus of extensive scientific investigations driven by regular evaluation campaigns, but which are mostly oriented towards a somewhat artificial task: translating news articles into English. In this paper, we investigate how well current MT approaches deal with a real-world task. We have *rationally reconstructed* one of the only MT systems in daily production use: the METEO system. We show how a combination of a sentence-based memory approach, a phrase-based statistical engine and a neural-network rescorer can give results comparable to those of the current system while offering a faster development cycle and better customization possibilities.

1 Introduction

Depuis la reprise des campagnes d'évaluation NIST[1] la traduction automatique (TA) revêt un caractère de plus en plus compétitif. La tâche partagée à laquelle se prêtent les "compétiteurs" de ces campagnes d'évaluation consiste à traduire vers l'anglais des textes journalistiques. S'il est clair que cette tâche répond en partie à des préoccupations concrètes du pays organisateur, il n'est cependant pas immédiat d'imaginer des applications réelles de la technologie évaluée.

Des tâches de traduction plus spécifiques existent cependant. Lors du workshop IWSLT (Akiba *et al.*, 2004) dont l'objectif premier était de proposer un protocole d'évaluation adapté à la traduction de corpus oralisés, la tâche partagée consistait à traduire des phrases du corpus BTEC (Basic Travel Expression Corpus). Ce corpus regroupe des phrases susceptibles d'être utiles à un touriste à l'étranger. Une autre tâche de traduction plus ciblée et qui a fait l'objet de nombreuses études est la tâche Verbmobil (Wahlster, 2000) qui consiste à traduire des dialogues de tâches précises (comme la prise de rendez-vous) pour la paire de langue anglais/allemand.

Dans cette étude, nous nous intéressons à une tâche encore plus précise et dont l'applicabilité ne fait cette fois-ci aucun doute puisqu'elle est reconnue comme l'un des plus grand succès de la traduction automatique: la traduction de l'anglais vers le français de bulletins météorologiques émis par *Environnement Canada* (EC)[2]. Nous baptisons cette tâche MÉTÉO.

Récemment, Leplus et al. (2004) montraient qu'à l'aide d'une mémoire de traduction phrastique peuplée de bulletins météorologiques déjà traduits, il était possible d'obtenir des traductions de bonne qualité. Ils expliquaient leur succès par un fort taux de répétitivité des phrases que le système MÉTÉO traduit. Dans ce travail, nous étudions la pertinence de plusieurs approches basées sur les corpus à traduire les bulletins météorologiques.

2 Protocole

Nous avons utilisé dans ce travail le bitexte décrit dans (Leplus *et al.*, 2004). Nous avons repris le même découpage en trois partie de ce bitexte: TRAIN pour l'entraînement des systèmes, BLANC pour leur ajustement et TEST pour tester les différentes approches. Ce découpage avait été choisi de manière à ce que les textes soient d'une période disjointe et que la tranche de test soit d'une période postérieure à celle de l'entraînement; ceci afin de simuler autant que faire se peut les conditions réelles d'utilisation du système.

Pour évaluer nos différentes approches, nous utilisons des métriques automatiques qui bien que discutables n'en sont pas moins largement utilisées: deux taux d'erreurs — WER au niveau des mots et SER au niveau des phrases — que l'on cherchera à minimiser et deux mesures de couverture n-gramme — NIST et 100×BLEU — que l'on voudra maximiser, toutes les deux calculées par le script `mteval` (version 11a) disponible depuis le site de NIST.

[1] Consulter `http://www.nist.gov/speech/tests/mt/` pour plus d'information.

[2] Le bulletin en cours peut être consulté à l'adresse `http://meteo.ec.gc.ca/forecast/textforecast_f.html`

3 Mémoire de traduction phrastique

Nous avons mesuré que 83% des phrases du corpus BLANC sont présentes *verbatim* dans le corpus TRAIN. Cette couverture atteint 87% si nous introduisons quelques classes de mots comme les jours, les mois ou encore les numéros de téléphones. Nous avons donc commencé par reproduire l'approche mémoire de traduction phrastique proposée par Leplus et al. (2004).

Nous avons construit une mémoire en gardant de chaque phrase source de TRAIN, un maximum de 5 traductions. En pratique, 89% des phrases anglaises de TRAIN n'ont qu'une seule traduction, probablement en raison du fait que la plupart des phrases ont été produites automatiquement (nous reviendrons sur ce point dans la section 7).

Pour une nouvelle phrase à traduire, nous recherchons les phrases sources les plus proches (en terme de distance d'édition) dans la mémoire et trions les traductions associées selon un score dont le détail est décrit dans (Langlais *et al.*, 2005). Dans cette expérience, la première phrase cible retournée est la traduction retenue.

mémoire				Leplus et al.			
WER%	SER%	NIST	BLEU	WER%	SER%	NIST	BLEU
8.42	23.43	10.9571	87.68	9.18	23.56	10.8983	86.95

Table 1: Évaluation de l'approche mémoire phrastique sur le corpus TEST et comparaison avec l'approche Leplus et al. (2004).

Les scores sont très bons si on les compare avec ceux observés dans d'autres tâches de traduction. Nous référons le lecteur à l'étude de Zens et Ney (2004) pour des performances état de l'art sur trois tâches de traduction incluant Verbmobil. Nos performances sont également légèrement supérieures à celles mentionnées par (Leplus *et al.*, 2004). Il n'en reste cependant pas moins que le taux d'erreur au niveau des phrases (c'est-à-dire le pourcentage de traductions produites non identiques à la traduction de référence) n'est pas particulièrement bas.

4 Approche probabiliste

Nous avons testé dans un deuxième temps une approche état de l'art en traduction statistique (Koehn *et al.*, 2003). Elle s'appuie sur un modèle de la distribution conditionnelle d'une séquence de mots dans une langue étant donnée une séquence dans l'autre langue. Les détails de l'obtention des modèles probabilistes sous-jacents sont donnés dans (Langlais *et al.*, 2005). Nous avons fait usage du décodeur PHARAOH (Koehn, 2004) disponible gratuitement pour des fins de recherche.

Les performances du système probabiliste sont présentées en table 2. Une comparaison directe avec les résultats mesurés avec l'approche mémoire milite en faveur de la mémoire, surtout si l'on observe le taux d'erreur au niveau des phrases. Cependant, nous remarquons que la performance du traducteur probabiliste lorsque mesurée sur les phrases à traduire qui n'ont pas été vues *verbatim* dans le corpus TRAIN sont de loin supérieures à celles obtenues par l'approche mémoire. Nous reviendrons sur la complémentarité de ces deux approches en section 7.

WER%	SER%	NIST	BLEU
7.46	32.01	10.8725	84.03

Table 2: Évaluation de l'approche statistique sur TEST.

5 Approche consensuelle

Bangalore et al. (2002) ont montré qu'il était possible de combiner des traductions produites par différents moteurs de traduction afin de générer des traductions d'une qualité supérieure à celles produites par un seul des moteurs. L'idée sous-jacente à cette approche (*bootstrapping*) est l'alignement de plusieurs traductions candidates afin d'isoler des îlots de confiance capables de diriger la génération d'une traduction dite consensuelle. Nous retrouvons cette idée dans certains systèmes d'acquisition et de génération de paraphrases.

Nous avons reproduit cette approche et avons pour cela adapté à nos besoins le programme CLUSTALW (Thompson *et al.*, 1994) écrit pour aligner entre-elles plusieurs séquences de protéines. À partir d'un alignement multiple de traduction (dont le lecteur trouvera les détails dans (Langlais *et al.*, 2005)), nous pouvons construire un treillis qui permet de produire en sus des traductions alignées de nouvelles phrases que l'on espère plus robustes. Nous utilisons le package CARMEL (Knight & Al-Onaizan, 1999) pour trouver dans un treillis la traduction consensuelle; c'est-à-dire le chemin de plus faible coût dans le treillis.

Les résultats de cette approche sont présentés en table 3 pour les seules phrases de BLANC non rencontrées *verbatim* dans le corpus ayant servi à créer la mémoire. Nous observons que la traduction par consensus améliore la qualité (telle que mesurée) des traductions produites. Le taux d'erreur au niveau des phrases est en particulier réduit de 9 points (en absolu), ce qui constitue une amélioration notable.

mémoire				mémoire + consensus			
WER%	SER%	NIST	BLEU	WER%	SER%	NIST	BLEU
18.69	94.82	9.7853	66.56	18.97	85.53	9.9314	68.86

Table 3: Performance de l'approche consensuelle sur la sortie de la mémoire de traduction pour les 13 010 phrases de BLANC non rencontrées dans le corpus TRAIN.

6 Ré-ordonnancement par apprentissage neuronal

Dans notre cadre, le *rescoring* consiste à ré-ordonner une liste d'alternatives produites par un système (dit natif) avec l'espoir que des informations supplémentaires, ou différentes façons de les utiliser, permettent de produire un ordonnancement plus pertinent. Le *rescoring* a fait l'objet d'études récentes en traduction probabiliste (Blatz *et al.*, 2004).

Dans notre contexte, cela consiste à reclasser la liste des meilleures traductions générées par PHARAOH (Koehn, 2004) pour une phrase donnée. Chaque alternative de traduction t_j est représentée par un vecteur de traits v_j et est étiquetée comme correcte si elle est identique à la traduction de référence et incorrecte sinon. Nous avons utilisé le package TORCH (Collobert

et al., 2002) pour entraîner un réseau perceptron multi-couche à estimer $p(\oplus|v_j)$, la probabilité conditionnelle de la correctitude d'une alternative t_j.

Nous avons testé différentes configurations de la couche cachée du réseau et avons considéré de nombreux traits pour représenter nos alternatives, chacun encodant des caractéristiques particulières. Les plus utiles étaient a) le ratio des longueurs de la phrase source et de la traduction candidate, b) la probabilité *a posteriori* de l'alternative et c) les scores $p(t_j|s)$ calculés par les modèles IBM 1 et 2 (Brown *et al.*, 1993). De plus amples informations sur cette approche sont disponibles dans (Langlais *et al.*, 2005).

Nous présentons en table 4 les performances mesurées par l'étape de rescoring.

smt				smt + rescoring			
WER%	SER%	NIST	BLEU	WER%	SER%	NIST	BLEU
7.46	32.01	10.8725	84.03	5.73	25.03	10.9828	87.40

Table 4: Comparaison des performances du moteur probabiliste (smt) seul et des traductions produites par reclassement (smt + *rescoring*) sur TEST.

7 Discussion

La diversité des approches que nous avons implémentées nous donne la souplesse de pouvoir les combiner. Pour illustrer ce point, nous avons évalué une combinaison très simple où la mémoire seule est consultée lorsque la phrase à traduire est déjà dans la mémoire, et où le moteur de traduction probabiliste *rescoré* est consulté sinon. Les performances ainsi mesurées (voir la table 5) sont meilleures que celles de chaque approche prise isolément.

WER%	SER%	NIST	BLEU
4.85	20.80	11.3021	89.59

Table 5: Performance sur TEST de la combinaison de la mémoire et du moteur de traduction probabiliste reclassé.

Il est cependant approprié de s'interroger quant à la performance véritable d'un tel système. Il est en particulier intéressant de contraster ces résultats avec ceux mesurés par le *bureau de la traduction du Canada* (BTC) qui est en charge de produire les traductions des bulletins météorologiques produits par *Environnement Canada* (EC). Le BTC utilise en effet le système MÉTÉO pour traduire automatiquement les bulletins anglais, mais a la responsabilité de réviser tout ou partie des traductions ainsi produites.

(Macklovitch, 1985) décrit une évaluation du système MÉTÉO-II conduite par le BTC. L'auteur a sélectionné 1257 phrases françaises publiées sur une période de 24 heures par EC et a compté le nombre de fois où le système produisait exactement la même traduction que celle qui a été publiée. Les erreurs dues à des fautes flagrantes non imputables au système étaient cependant écartées (typos, erreur de transmission, etc.). Il rapporte que seulement 11% des phrases testées étaient différentes de celles publiées.

Ce protocole d'évaluation correspond grossièrement au nôtre lorsque nous mesurons un taux d'erreur au niveau des phrases. Les approches que nous avons implémentées ne montrent pas un tel niveau de performance. Cependant, une comparaison directe des deux protocoles n'est pas adéquate. Premièrement, nous évaluons nos approches sur un corpus bien plus grand (36 228 phrases). Deuxièmement, nous avons mesuré un bruit d'environ 7% dans notre référence. Troisièmement, une évaluation informelle d'un échantillon de 1000 traductions (choisies aléatoirement) différentes de celles de notre référence, nous a révélé que 77% d'entre-elles étaient des traductions correctes.

Références

AKIBA Y., FEDERICO M., KANDO N., NAKAIWA H., PAUL M. & TSUJII J. (2004). Overview of the IWSLT04 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, p. 1–12, Kyoto.

BANGALORE S., MURDOCK V. & RICCARDI G. (2002). Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. In *Proceedings of COLING*, p. 50–56, Taipei.

BLATZ, J., FITZGERALD, E., FOSTER, G., GANDRABUR, S., GOUTTE, C., KULESZA, A., SANCHIS, A., UEFFING & N. (2004). Confidence estimation for machine translation. In *Proceedings of COLING*, p. 315–321, Geneva.

BROWN P., PIETRA S. D., PIETRA V. D. & MERCER R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, **19**(2), 263–311.

COLLOBERT R., BENGIO S. & MARIÉTHOZ. J. (2002). *Torch: a modular machine learning software library*. Rapport interne IDIAP-RR 02-46, IDIAP.

KNIGHT K. & AL-ONAIZAN Y. (1999). *A Primer on Finite-State Software for Natural Language Processing*. http://www.isi.edu/licensed-sw/carmel/carmel-tutorial2.pdf.

KOEHN P. (2004). Pharaoh: a beam search decoder for phrase-based smt. In *Proceedings of AMTA*, p. 115–124, Washington.

KOEHN P., OCH F. & MARCU D. (2003). Statistical phrase-based translation. In *Proceedings of HLT*, p. 127–133, Edmonton.

LANGLAIS P., LEPLUS T., GANDRABUR S. & LAPALME G. (2005). From the real world to real words: The meteo case. In *in 10th Annual Conference of the European Association for Machine Translation*, Budapest, Hungary.

LEPLUS T., LANGLAIS P. & LAPALME G. (2004). Weather report translation using a translation memory. In *Proceedings of AMTA*, p. 154–163, Washington.

MACKLOVITCH E. (1985). *A Linguistic Performance Evaluation of METEO 2*. Rapport interne, Canadian Translation Bureau.

THOMPSON J., HIGGINS D. & GIBSON T. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**(22), 4673–4680.

WAHLSTER, Ed. (2000). *Verbmobil: Foundations of speech-to-speech translations*. Berlin, Germany: Springer Verlag.

ZENS R. & NEY H. (2004). Improvements in phrase-based statistical machine translation. In *Proceedings of HLT/NAACL*, p. 257–264, Boston.