

Construction (très) rapide de tables de traduction à partir de grands bi-textes

Aurélien Max, Li Gong, François Yvon

▶ **To cite this version:**

Aurélien Max, Li Gong, François Yvon. Construction (très) rapide de tables de traduction à partir de grands bi-textes. Conférence sur le Traitement Automatique des Langues Naturelles, Jan 2014, Marseille, France. hal-01908369

HAL Id: hal-01908369
https://hal.science/hal-01908369

Submitted on 31 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Construction (très) rapide de tables de traduction à partir de grands bi-textes

Li Gong[1,2] Aurélien Max[1,2] François Yvon[1]
(1) LIMSI-CNRS, Orsay, France (2) Univ. Paris Sud, Orsay, France
{prénom.nom}@limsi.fr

Résumé. Dans cet article de démonstration, nous introduisons un logiciel permettant de construire des tables de traduction de manière beaucoup plus rapide que ne le font les techniques à l'état de l'art. Cette accélération notable est obtenue par le biais d'un double échantillonnage : l'un permet la sélection d'un nombre limité de bi-phrases contenant les segments à traduire, l'autre réalise un alignement à la volée de ces bi-phrases pour extraire des exemples de traduction.

Mots-clés : traduction automatique statistique ; développement efficace ; temps de calcul.

Keywords: statistical machine translation ; efficient development ; computation time.

1 Introduction et motivations

De très grands bi-textes sont désormais disponibles pour l'apprentissage des systèmes de traduction automatique statistique. Cependant, la grande majorité des situations d'utilisation de ces systèmes n'imposera en pratique l'exploitation que d'une petite partie des données disponibles. Les approches standard, telles que celle du système à l'état de l'art `Moses` [1], reposent sur l'alignement au niveau des mots de l'intégralité des bi-textes, étape très couteuse en temps. Afin de réduire les temps de construction des tables de traduction, des travaux ont déjà proposé de recourir à un échantillonnage des exemples de traduction focalisé sur les seuls segments à traduire, permettant d'obtenir directement des tables de traduction très compactes dont la performance rivalise avec des tables apprises sur l'intégralité du bi-texte (Callison-Burch *et al.*, 2005). Toutefois, le gain en temps obtenu par cette approche ne correspond qu'à une réduction d'environ 15% du temps de traitement.

Nous présentons un logiciel permettant de considérablement modifier cette situation : non seulement les bi-phrases contenant des exemples de traduction utiles sont échantillonnées, mais l'alignement de ces bi-phrases est également construit à la demande à l'aide de la technique décrite dans (Gong *et al.*, 2013). Ce résultat peut par exemple être comparé à celui décrit dans (Zens *et al.*, 2012), où des tables de traduction sont filtrées *a posteriori*, soit en ajoutant à la procédure d'apprentissage standard un calcul (lui-même couteux) de détection de bi-segments redondants selon un critère d'entropie. Zens *et al.* (2012) décrivent par exemple, pour la paire de langue anglais-français, qu'un filtrage ne retenant que 8,1% des entrées d'une table de traduction ne mène pas à une perte supérieure à 1 point BLEU. Dans nos expériences ci-dessous, le même résultat est obtenu en n'effectuant que 6,9% du temps de calcul utilisé par un système de référence pour l'estimation des tables de traduction de développement et de test.

2 Description du logiciel

L'architecture de notre logiciel d'estimation de tables de traduction [2] est décrite sur la Figure 1. Pour l'ensemble des segments du texte à traduire, un échantillonnage aléatoire [3] (à taille constante) est effectué dans l'intégralité du bi-texte de manière efficace à l'aide d'un tableau de suffixes. Pour les bi-phrases trouvées et ne se trouvant pas déjà dans un cache d'alignements, un alignement est construit par une procédure d'alignement ciblé par échantillonnage proposée par Gong *et al.* (2013). Une fois ces phrases alignées, les comptes nécessaires au calcul d'un ensemble standard de traits pour des bi-segments sont extraits, et les tables de traduction sont finalement écrites sur disque pour être utilisé par un décodeur

1. `http://www.statmt.org/moses`
2. Nous regroupons sous le nom générique de "tables de traduction" les tables de bi-segments et tables de réordonnancement lexicalisées.
3. D'autres stratégies d'échantillonnage possibles ne seront pas décrites ici.

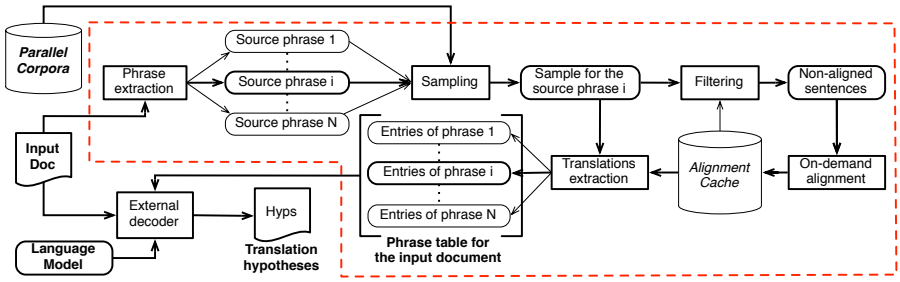


FIGURE 1 – Architecture de notre système (encadré rouge) permettant l'estimation rapide de tables de traduction.

fondé sur les segments. Le logiciel, intégralement développé en langage Java, permet une exécution *multi-thread*, et sera disponible en téléchargement libre à des fins de recherche.

3 Exemple de validation expérimentale

Systèmes		**Performance traduction**		**Performance temps**				
	α	BLEU	TER	TabSuff.	dev	test	tuning	total
`Moses`	-	34,12±0,10	48,59±0,22	-	1 212h		21h	1 233h
`otf`	1 000	33,35±0,02	49,62±0,05	2h	72h	81h	20h	175h
	500	33,05±0,10	50,03±0,11	2h	40h	44h	20h	106h
	250	32,81±0,03	49,87±0,07	2h	24h	25h	20h	71h
	100	32,52±0,08	50,49±0,15	2h	14h	14h	20h	50h
	50	32,73±0,06	49,46±0,06	2h	10h	11h	20h	43h

TABLE 1 – Résultats pour nos expériences de traduction anglais-français des documents Cochrane. La performance en temps est donnée en *user CPU time* tel que donné par la commande UNIX `time`.

Pour illustrer la performance de notre approche, nous avons exécuté les expériences suivantes. Nous avons utilisé un grand corpus de 16,6 millions de bi-phrases anglais-français de la tâche de traduction médicale de WMT'14 [4]. Nous avons choisi comme domaine d'évaluation des documents destinés à des spécialistes en médecine produit par la collaboration Cochrane [5], en utilisant 743 phrases pour le développement, et 1 800 pour l'évaluation. Nous avons suivi les procédures standard du logiciel `Moses`, utilisant notamment MERT pour l'optimisation des paramètres.

La Table 1 présente les résultats obtenus pour un système `Moses` standard, ainsi que pour différentes valeurs du paramètre d'échantillonnage (α) de calcul des scores d'association utilisé par notre procédure d'alignement (voir (Gong *et al.*, 2013)). On constate une perte en BLEU relativement au système `Moses` allant de 0,77 à 1,38 point BLEU, pour un gain en temps pour l'estimation des tables nécessaires de respectivement 87% à 98%.

Références

CALLISON-BURCH C., BANNARD C. & SCHROEDER J. (2005). Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of ACL*, Ann Arbor, USA.

GONG L., MAX A. & YVON F. (2013). Improving bilingual sub-sentential alignment by sampling-based transpotting. In *Proceedings of IWSLT*, Heidelberg, Germany.

ZENS R., STANTON D. & XU P. (2012). A Systematic Comparison of Phrase Table Pruning Techniques. In *Proceedings of EMNLP*, p. 972–983, Jeju Island, Korea.

4. `http://www.statmt.org/wmt14`
5. `http://www.cochrane.org`