TALN 2009, Senlis, 24–26 juin 2009

anymalign : un outil d'alignement sous-phrastique libre pour les êtres humains

Adrien Lardilleux Yves Lepage
GREYC, Université de Caen Basse-Normandie
prénom.nom@info.unicaen.fr

Résumé. Nous présentons *anymalign*, un aligneur sous-phrastique grand public. Ses résultats ont une qualité qui rivalise avec le meilleur outil du domaine, GIZA++. Il est rapide et simple d'utilisation, et permet de produire dictionnaires et autres tables de traduction en une seule commande. À notre connaissance, c'est le seul outil au monde permettant d'aligner un nombre quelconque de langues simultanément. Il s'agit donc du *premier aligneur sous-phrastique réellement multilingue*.

Abstract. We present *anymalign*, a sub-sentential aligner oriented towards end users. It produces results that are competitive with the best known tool in the domain, GIZA++. It is fast and easy to use, and allows dictionaries or translation tables to be produced in a single command. To our knowledge, it is the only tool in the world capable of aligning any number of languages simultaneously. It is therefore *the first truly multilingual sub-sentential aligner*.

Mots-clés : alignement sous-phrastique, multilinguisme, table de traduction.

Keywords: sub-sentential alignment, multilinguism, translation table.

1 Vue d'ensemble

L'alignement sous-phrastique consiste à extraire des traductions d'unités textuelles de grain inférieur à la phrase à partir de textes multilingues dont les phrases ont préalablement été mises en correspondance. Le plus connu des outils disponibles est sans conteste GIZA++ (Och & Ney, 2003). Il implémente les modèles probabilistes IBM. Bien que produisant des résultats d'une grande qualité, beaucoup de ses utilisateurs sont loin d'en saisir toutes les subtilités, car de tels modèles sont d'une grande complexité. Cette complexité se répercute sur l'interface d'utilisation de l'outil.

Pour répondre à toutes ces difficultés, nous mettons à disposition de tous un aligneur sous-phrastique qui se distingue de GIZA++ selon plusieurs points. Il est :

réellement multilingue : il permet d'aligner un nombre quelconque de langues simultanément. Contrairement aux modèles d'alignement traditionnels, la méthode sous-jacente n'est pas nécessairement bilingue. Elle repose sur un traitement *alingue*, ce qui la rend en pratique multilingue ;

simple d'utilisation : une *unique* commande suffit dans la plupart des cas. Il n'y a pas de paramètre superflu ;

rapide : la qualité des alignements produits ne dépend pas du temps d'exécution. C'est la couverture du texte d'origine par les alignements qui augmente en fonction du temps. L'uti-

lisateur peut à tout moment interrompre le traitement et récolter tous les alignements obtenus jusqu'alors. En pratique, quelques secondes suffisent à extraire les alignements les plus fréquents à partir de textes de plusieurs centaines de milliers de phrases ;

facilement parallélisable : plusieurs processus peuvent être lancés simultanément, sur plusieurs machines par exemple. Les sorties de chacun peuvent être fusionnées sans difficulté aucune ;

facilement intégrable : les formats d'entrée et de sortie sont de simples fichiers textes, aisément réutilisables dans tout environnement (par exemple un pipeline Unix ou une plateforme de traitement linguistique). D'autres formats sont disponibles, tels que des sorties directement compatibles avec le décodeur statistique *Moses*, HTML, ou encore le format professionnel de mémoire de traduction, TMX de LISA ;

complet : contrairement à un aligneur qui établit des liens entre les mots sources et cibles d'un couple de phrases traductions l'une de l'autre, *anymalign* produit directement des tables de traduction constituées d'alignements (suites de mots connexes ou non) et de leurs scores associés : actuellement fréquence, probabilités de traduction et poids lexicaux ;

portable : le programme est écrit dans le langage de programmation Python, disponible sur la plupart des systèmes d'exploitation (actuellement la version 2.4 suffit). Il ne nécessite aucun module externe : seule la bibliothèque standard est utilisée. Il est constitué d'un unique fichier exécutable : aucune installation n'est nécessaire.

2 En théorie

La méthode d'alignement originale a été décrite dans (Lardilleux & Lepage, 2008). Les étapes sont les suivantes :

1. transformer le corpus d'entrée multilingue en un corpus alingue, c'est-à-dire supprimer les limites entre langues après discrimination des formes surfaciques des mots selon leur langue d'origine. Sur le corpus trilingue français-anglais-arabe ci-dessous, cette discrimination a été faite par indiçage sur les langues (voir en particulier les points de fin de phrases) ;
2. extraire un sous-corpus par échantillonnage. En pratique, des sous-corpus de petite taille sont privilégiés car ils fournissent de meilleurs résultats (Lardilleux & Lepage, 2008), comme le sous-corpus suivant :

1	Un$_1$ café$_1$,$_1$ s'il$_1$ vous$_1$ plaît$_1$.$_1$ One$_2$ coffee$_2$,$_2$ please$_2$.$_2$ ٫$_3$ فضلك$_3$ من$_3$ ،$_3$ قهوة$_3$
2	Ce$_1$ café$_1$ est$_1$ excellent$_1$.$_1$ This$_2$ coffee$_2$ is$_2$ excellent$_2$.$_2$ ٫$_3$ ممتازة$_3$ قهوة$_3$ هذه$_3$
3	Un$_1$ thé$_1$ fort$_1$.$_1$ One$_2$ strong$_2$ tea$_2$.$_2$ ٫$_3$ ثقيل$_3$ شاي$_3$

3. regrouper les mots de ce sous-corpus en fonction des lignes sur lesquelles ils apparaissent : les mots d'un même groupe apparaissent strictement sur les mêmes lignes (l'ordre des mots importe peu) :

	Les mots :	apparaissent sur les lignes :
1	Un$_1$ One$_2$	1 3
2	café$_1$ coffee$_2$ قهوة$_3$	1 2
3	,$_1$ s'il$_1$ vous$_1$ plaît$_1$,$_2$ please$_2$ فضلك$_3$ من$_3$ ،$_3$	1
4	.$_1$.$_2$ ٫$_3$	1 2 3
5	Ce$_1$ est$_1$ excellent$_1$ This$_2$ is$_2$ excellent$_2$ ممتازة$_3$ هذه$_3$	2
6	thé$_1$ fort$_1$ strong$_2$ tea$_2$ ثقيل$_3$ شاي$_3$	3

4. pour chaque groupe, relire les lignes du sous-corpus où ce groupe apparaît, et en extraire deux alignements :
   (a) la séquence ordonnée constituée des mots du groupe ;
   (b) la séquence complémentaire de cette séquence ordonnée également.
   Par exemple, le groupe 2 permet d'extraire la séquence ordonnée « café$_1$ coffee$_2$ قهوة$_3$ » et son complémentaire « Un$_1$ _ ,$_1$ s'il$_1$ vous$_1$ plaît$_1$.$_1$ One$_2$ _ ,$_2$ please$_2$.$_2$ من$_3$ فضلك$_3$ ،$_3$ » de la ligne 1. Il permet aussi d'extraire « café$_1$ coffee$_2$ قهوة$_3$ » et son nouveau complémentaire « Ce$_1$ _ est$_1$ excellent$_1$.$_1$ This$_2$ _ is$_2$ excellent$_2$.$_2$ هذه$_3$ _ ممتازة$_3$.$_3$ » de la ligne 3. Chaque groupe est donc susceptible de produire deux alignements par ligne ;
5. répéter les étapes 2 à 4 tant que de nouveaux alignements sont produits, ou jusqu'à ce que l'utilisateur interrompe le processus. Le nombre de fois que chaque alignement a été obtenu permet de déduire les probabilités de traduction associées. Des poids lexicaux peuvent également être calculés à partir des cooccurrences de mots dans le corpus initial.

3 En pratique

Le programme implémentant la méthode décrite ci-dessus est téléchargeable à l'adresse :

`http://users.info.unicaen.fr/~alardill/anymalign/`

Il permet de traiter des quantités de données considérables tout en minimisant les ressources nécessaires. Étant donné un corpus multilingue disponible dans des fichiers séparés avec une phrase par ligne, la simple commande :

```
python anymalign.py langue1.txt langue2.txt [...]
```

suffit pour obtenir une table de traduction. La figure 1 donne un exemple de résultat en HTML.

No	Freq.	Translation probabilities	Lexical weightings	ja	en	ar
1	50759	0.98 0.74 0.75	0.95 0.97 0.98	か	?	؟
2	47610	0.83 0.71 0.72	0.64 1.00 1.00	。	.	.
3	8161	0.95 0.97 0.93	0.83 0.96 0.94	どこ	Where	أين
4	6446	0.92 0.10 0.10	0.41 1.00 1.00	です 。	.	.
5	5372	0.92 0.08 0.08	0.40 1.00 1.00	か 。	?	؟
6	3344	0.99 0.99 0.99	1.00 0.99 1.00	東京	Tokyo	طوكيو
7	3134	0.96 0.97 0.96	0.73 0.95 0.99	日本	Japan	اليابان

FIG. 1 – Exemple d'un début de sortie d'*anymalign* sur un corpus japonais-anglais-arabe. Les alignements sont triés par fréquence. Les meilleurs alignements sont mis en évidence par la couleur.

Références

LARDILLEUX A. & LEPAGE Y. (2008). A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. Dans *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008)*, p. 125–132, Waikiki, Hawai'i, États-Unis.

OCH F. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**, 19–51.