L'alignement au cœur du multilinguisme.

Jacques Guyot#1,#3, Gilles Falquet#2

#1 The Olanto Foundation,10 Chemin de Champ-Claude,1214 Vernier (Geneva),Switzerland

#2 University of Geneva, Centre Universitaire d'Informatique, Battelle Bat A, 7 Route de Drize,1227 Carouge, Switzerland

#1 jacques@olanto.org, #2 gilles.falquet@unige.ch, #3 www.olanto.org

Résumé. Nous montrons comment l'algorithme d'alignement au niveau phrase est implémenté dans les outils d'Olanto. L'algorithme est basé sur l'utilisation de lexiques pour calculer la similarité entre les phrases. Nous décrivons comment nous arrivons à une complexité O(n) pour le calcul de la similarité entre deux phrases. Nous décrivons aussi le processus pour produire les lexiques utilisés lors de l'alignement.

Abstract We present the implementation of the alignment algorithm in the Olanto tools. The algorithm is based on the computation of a sentence similarity measure. We describe how we reduced the time complexity of the similarity computation from $O(n^2)$ to O(n). We also describe the process to producing the lexicon used in the alignment.

Mots-clés : Alignement de phrases, complexité, lexique.

Keywords: sentence alignment, complexity, lexicon.

1 Introduction

Les outils d'Olanto sont utilisés dans par des entreprises et des organisations internationales dont les corpus sont volumineux (plusieurs centaines de milliers de documents) et les langues nombreuses et diverses (langues de la communauté européennes et onusiennes). La qualité des alignements et la performance sont des objectifs incontournables. Le temps d'alignement moyen de deux documents doit être bien inférieur à la seconde. L'utilisation des alignements produits est multiple, sous forme de TMX compatible avec les éditeurs avec MT (comme OmegaT[1]), de Bitexte (Format XML de l'ensemble des segments des deux documents similaire à ceux de Terminotix[2]), de corpus d'entrainement pour la traduction statistique automatique et de carte d'alignement pour le concordanciers myCAT. De plus, nous devons pouvoir aligner des corpus bilingues (traduction) et aligner des corpus comparables (même sujet, comme les fiches wikipédia).

[1] OmegaT: http://www.omegat.org/

[2] Terminotix: http://www.terminotix.com/

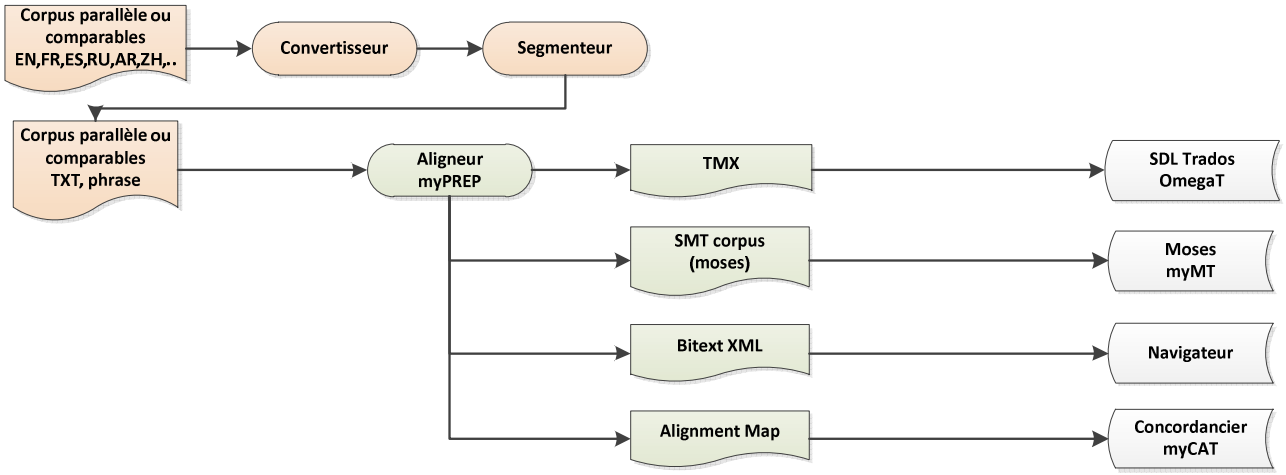

Figure 1 : Le module d'alignement est la source de différentes utilisations des corpus.

L'alignement à une longue histoire avec différentes méthodes reposant sur la longueur des phrases, les « cognate », l'utilisation de lexiques. On trouvera dans (SANTOS, 2011), une revue des différents outils et des méthodes déjà explorés. Lors de l'élaboration de l'aligneur, nous avons aussi choisi de ne pas inclure des ressources linguistiques complexes dépendant des langues (lemmatisation, …).

2 Algorithme

Nous nous plaçons dans le cas où les documents à aligner ont été convertis et segmentés en phrases. Les documents ont généralement un nombre de phrases différent. Nous sommes intéressés que par les phrases qui sont associés à une seule phrase. Si les documents étaient traduits phrase à phrase, l'alignement formerait une diagonale (figure 2) sur un graphe où les axes sont les numéros des phrases de chaque document. Les erreurs de conversion, de segmentation et une certaine liberté du traducteur font que la fonction d'alignement navigue autour de cette diagonale. Cette fonction peut être discontinue si des blocs de texte sont inversés dans le document (listes par ordre alphabétique). La tâche est donc de trouver cette fonction d'alignement.

2.1 Principe

Une évidence qui peut être expérimentée par chacun est qu'avec des connaissances minimum lexicales, il est possible de retrouver entre deux textes, les phrases qui sont des traductions mutuelles. Nous utilisons une approche mixte « géométrique et lexicale » qui est utilisée depuis de nombreuses années (MELAMED D.2001), (TIEDMANN 2011)

	Le	chat	noir	mange	une	souris	verte
The	1						
black			0.76				
cat		0.8					
eats				0.6			
a					1		
green							0.9
mouse						0.82	

lexique		
EN	FR	Score
black	noir	0.76
black	noires	1.00
cat	chat	0.8
cat	chats	0.9
…	…	…
…	…	…

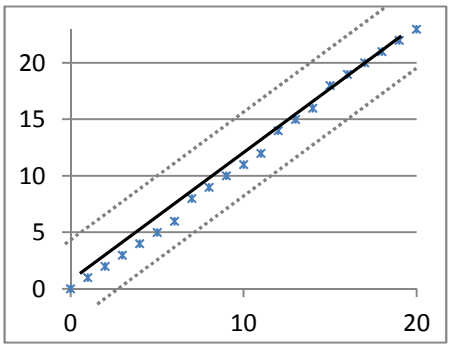

Figure 2 : Matrice de similiraté entre deux phrases, Structure du lexique, Alignement de deux documents avec la diagonale d'alignement et une fenêtre de recherche autour de la diagonale.

L'utilisation d'un lexique suffit donc à établir une mesure de similarité entre 2 phrases. La similarité sera donc la somme de tous les scores pour chaque entrée trouvée dans le lexique pondéré par la longueur des phrases.

Si **N** est le nombre de phrases à aligner, **w** la largeur de fenêtre d'exploration autour de la diagonale, et **n** la taille moyenne des phrases, on obtient Nwn^2 opérations de recherche dans le lexique. Sachant que la valeur moyenne de **n** est 40, le test de similarité coûte 1600 recherches.

Intuitivement on perçoit bien que l'on fait des tests inutiles. Nous allons montrer que l'on peut passer réduire cette complexité. La première observation est de constater que le calcul de similarité est indifférent à l'ordre des termes dans la phrase que le test avec la phrase « chat le mange noir souris une verte » dont les termes sont triés, obtient le même score. La deuxième observation est de voir que si l'on remplace la phrase source (anglaise) par toutes les traductions proposées par le lexique et que l'on trie le résultat, on obtient alors deux listes ordonnées à comparer.

Le	chat	noir	mange	une	souris	verte

Le (1)	chat(0.8)	chats(0.9)	noir(0.76)	noires(1)	mange(0.6)	une(1)	souris(0.82)	verte(0.9)

Figure 3 : la deuxième liste est obtenue par traduction des termes à travers le lexique

Ce problème est d'ordre O(n). Le tri et la traduction avec le lexique est utilisé une seule fois par phrase (à l'initialisation). On obtient donc bien Nwn opérations de comparaison.

L'apparition des termes supplémentaires dans la liste ne modifie pas le score. Soit ils n'ont pas de correspondance dans la liste cible et donc valent zéro, soit ils ont une correspondance dans la liste, mais avec ils auraient aussi contribué dans l'approche matricielle.

Cette réduction a largement facilité l'alignement pour les documents comparables. En effet, pour estimer Si deux phrases sont comparables entre deux documents, il faut tester deux à deux toutes les phrases de chaque document. La largeur w de la fenêtre d'exploration prend comme valeur le nombre de lignes du document cible. Nous avons pu expérimenter un gain en temps d'environ 100 sur l'alignement de 300'000 fiches Wikipédia.

Ceci ne constitue que le principe d'alignement. Nous utilisons aussi la longueur des phrases pour éliminer des associations trop dissemblables, les nombres et les entités nommées qui ne sont pas dans le lexique et aussi la « géométrie » des alignements déjà effectués pour restreindre l'espace des recherches.

2.2 Comment créer les lexiques ?

La méthode d'alignement est basée sur l'utilisation d'un lexique bilingue avec une valeur de confiance entre les deux termes. Dans l'introduction, nous avions énoncé une contrainte concernant un minimum de dépendance avec des ressources linguistiques. Ces lexiques sont difficiles à obtenir et généralement, ils ne sont pas pondérés. Par contre, il est assez facile d'obtenir un ensemble de phrases alignées soit en alignant manuellement des documents du corpus à aligner ou en utilisant un service de traduction en ligne. Notre expérience montre que quelques milliers de phrases sont suffisantes pour capter l'essence des deux langues. Dans une première version de l'aligneur, nous avons utilisé une mesure de corrélation (GUYOT,

2014) entre deux termes calculée sur la base d'un corpus de phrases déjà alignées entre deux langues bien que cette méthode soit satisfaisante, elle était lente. Une autre possibilité est d'utiliser Ghiza[3] le module de construction des modèles de traduction de Moses. Lors de l'élaboration du modèle, un alignement au niveau des termes est produit (KHOEN ,2010). Le résultat de cet alignement est compatible avec notre besoin.

La procédure est donc de créer un premier lexique d' « amorçage » avec le jeu initial de phrases alignées en utilisant Ghiza. Ensuite, on aligne le corpus. Ceci crée un nouvel ensemble plus conséquent et plus représentatif de phrases alignées. On crée alors le lexique définitif en utilisant à nouveau Ghiza. Le lexique produit est assez général pour être représentatif de la traduction des termes entre les deux langues à aligner.

2.3 Comparaison avec d'autres programmes d'alignement

Nous avons effectué une comparaison de la vitesse d'alignement avec deux autres programmes. Un produit commercial Terminotix et Hunalign. Terminotix spécifie la vitesse d'alignement sur son site de 500 documents de 650 segments (de 12000 caractères) alignés en une heure sur un pentium à 2.8Ghz. Nous avons corrigé ce temps par rapport au CPU qui est utilisé pour les autres tests. Pour Hunalign et myTerm, nous avons construit un corpus de 512 doc / 760 seg / 12550 char, pour avoir des tâches comparables. Les résultats de la figure 4 montre un net avantage pour myTerm dans cette tâche. Le rapport de 25 entre Terminotix et myTerm est bien celui que l'on a constaté en production chez les utilisateurs. Les tests ont été faits en mono-thread.

Aligneur	Corpus	processeur	durée alignement [s]	durée corrigée [s]	indice de performance
Terminotix[4]	500 doc / 650 seg / 12000 char	pentium @ 2.8GHz	3 600	1 256	1,0
Hunalign[5]	512 doc / 760 seg / 12550 char	i7-4710HQ @ 2.50GHz (portable)		760	1,7
myTerm	512 doc / 760 seg / 12550 char	i7-4710HQ @ 2.50GHz (portable)		50	25,1

Figure 4 : tableau comparatif des temps d'exécution

3 Conclusion

Bien que l'aligneur soit très performant, il est encore possible d'agir sur la largeur *w* de la fenêtre d'exploration qui est déterminée à partir de la taille du document et qui reste constante pour le reste du processus. Une approche dichotomique permettrait de la réduire en cherchant des points d'alignements intermédiaires.

[3] http://www.statmt.org/moses/
[4] http:// http://www.terminotix.com/
[5] http:// http://mokk.bme.hu/resources/hunalign/

Un autre axe d'intérêt est d'étendre l'alignement au niveau document. C'est-à-dire de rechercher dans un corpus de documents quels sont les documents qui sont susceptibles d'être des traductions mutuels. Ceci permettrait d'établir plus facilement les corpus bilingues.

Références

SANTOS A. (2011). A survey on parallel corpora alignment. In Proceedings of MI-Star, pages 117– 128.

GUYOT J, GHOULA N., FALQUET G. (2014) Terminology management revisited. Proc of ASLIB 2014, November 2014

KHOEN Philipp (2010) Statistical Machine Translation. Cambridge University Press, 2010

MELAMED D. (2001) Empirical Methods for Exploiting Parallel Texts. Cambridge, MA: MIT Press, 2001

TIEDMANN J. (2011) Bitext Alignement. Morgan & Claypool Publisher 2011