L'alignement sous-phrastique multilingue pour les nuls

Adrien Lardilleux

MajecSTIC 2009
Avignon, France, du 16 au 18 novembre 2009

L'alignement sous-phrastique multilingue pour les nuls

Adrien Lardilleux

GREYC, Université de Caen Basse-Normandie, bd M[al] Juin, BP 5186, 14032 Caen CEDEX, France.

Contact : Adrien.Lardilleux@info.unicaen.fr

Résumé
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. Les méthodes les plus répandues actuellement, bien que produisant des résultats de grande qualité, sont complexes, supportent difficilement le passage à l'échelle, et ne peuvent traiter les langues que par couples. Elles mettent généralement l'accent sur les mots fréquents. Nous présentons une approche radicalement différente, tirant parti des mots rares. Elle permet l'alignement d'un nombre quelconque de langues simultanément et un passage à l'échelle naturel, tout en demeurant d'une grande simplicité.

Abstract
Sub-sentential alignment is the process by which multi-word unit translations are extracted from sentence-aligned multilingual parallel texts. While producing high quality results, common methods in the field are complex, may not scale up, and can only process pairs of languages. They usually focus on frequent words. We present a totally different approach that relies on rare words. It makes it possible to align any number of languages simultaneously and scales up naturally. The method is very simple.

Mots-clés : traitement automatique des langues, multilinguisme, traduction automatique, alignement sous-phrastique, hapax.

Keywords: natural language processing, multilinguism, machine translation, sub-sentential alignment, hapax.

---

1. Traduction automatique et alignement sous-phrastique

La traduction automatique est une des plus anciennes applications de l'informatique. Le principe est simple : étant donnée une phrase dans une *langue source*, la machine doit automatiquement en produire la traduction dans une *langue cible*. Des outils sont accessibles sur le web pour s'acquitter de cette tâche, citons entre autres *Systran* (http://www.systran.fr/) et *Google Traduction* (http://translate.google.com/). Ceux-ci témoignent de l'état de l'art : les traductions restent compréhensibles tant que les langues sont proches (par exemple français et italien), mais elles servent tout au plus à donner une vague idée du contenu d'un texte si les langues sont éloignées (par exemple chinois et français).
Les grands courants de la traduction automatique sont aujourd'hui au nombre de deux :

la traduction par règles est la plus ancienne. Elle a recours à des connaissances sur les langues, telles que dictionnaires, analyseurs morphologiques et syntaxiques, règles de transfert, etc. Le leader en la matière est *Systran*, qui présente le problème sur son site web comme suit :

> « La traduction automatique n'est pas un processus simple. Il ne s'agit pas de la simple traduction d'un mot par un autre, mais de la capacité à connaître tous les mots dans une phrase ou un contexte donné. Les langues naturelles se caractérisent par leur morphologie (composition des mots), leur syntaxe (structure de la phrase), leur sémantique (sens des mots) et présentent un grand nombre d'ambiguïtés. »

Du point de vue du Règlement, ces amendements sont parfaitement réglementaires.	↔	In accordance with the Rules of Procedure, they are perfectly permissible.
Par conséquent, nous passons au vote de la proposition de règlement.	↔	Therefore, we shall now proceed to the vote on the proposed regulation.
(Le Parlement approuve la proposition de la Commission)	↔	(Parliament approved the Commission' s proposal)

FIG. 1 – Extrait de la partie français-anglais du corpus parallèle Europarl [5], constitué de phrases en français et de leurs traductions en anglais. Le corpus d'origine contient plusieurs millions de phrases traduites en 11 langues (28 mots par phrase en moyenne).

Devant toute la richesse, les spécificités et autres irrégularités des langues, prévoir des règles pour tous les couples de langues est une tâche dantesque. De tels systèmes sont donc très coûteux en ressources et en temps humain ;

la traduction fondée sur les données est plus récente (20 ans d'histoire environ). Elle ne requiert aucune connaissance a priori sur les langues. À la place, elle a recours à des *textes parallèles* : un texte dans une langue source et sa traduction dans une langue cible, les correspondances étant établies au niveau de la phrase. La figure 1 en donne un exemple. Dans ce paradigme, des *exemples* de traduction suffisent, à partir desquels les connaissances nécessaires à la traduction d'une nouvelle phrase sont extraits. *Google Traduction* utilise cette approche (plus précisément une approche probabiliste).

La traduction fondée sur les données commence le plus souvent par une étape d'*alignement sous-phrastique*. Celle-ci consiste à extraire des traductions d'unités textuelles de grain inférieur à la phrase, telles que mots ou séquences de mots, à partir de textes parallèles. Des exemples de traduction de la figure 1, on attend par exemple en sortie des correspondances lexicales telles que *réglementaires* ↔ *permissible* ou encore *de la Commission* ↔ *the Commission's*. C'est sur cette tâche que se concentre cet article.

2. Problèmes inhérents aux approches actuelles

De nombreuses propositions ont été émises pour s'acquitter de la tâche d'alignement. Le problème a été abordé sous différents angles : heuristiques linguistiques (la traduction d'un nom a de grandes chances d'être un nom, des mots proches graphiquement sont probablement traductions les uns des autres) [9], similarités de distribution (les termes qui ont tendance à apparaître en même temps sont sûrement traductions l'un de l'autre) [4], ou encore purement statistiques [10]. Cette dernière approche est celle qui a produit le plus de littérature. Elle fut à l'origine de la traduction automatique statistique, avec les célèbres modèles IBM [1]. Souvent très complexe, elle est la plus répandue aujourd'hui, toujours largement fondée sur ces mêmes modèles.

Bien que radicalement différentes par essence, toutes ces approches ont un point commun : elles ne permettent de traiter les langues que par *couples*. En effet, la traduction automatique a été vue dès ses débuts comme un problème *bilingue*. Lorsque des alignement réellement multilingues (trois langues ou plus) sont nécessaires, deux options s'offrent alors :

1. on traite tous les couples de langues. Le nombre de couples à traiter est donc quadratique en le nombre total de langues [5, 11] ;
2. on a recours à une langue « pivot » (le plus souvent l'anglais). La perte de sens est alors presque inévitable.

Un autre problème inhérent à certains modèles est celui du passage à l'échelle : leur complexité est telle que la quantité de données pouvant être traitée est en pratique bornée. Bien que des textes parallèles de grande longueur mènent généralement à des résultats de meilleure qualité, leur traitement nécessite énormément de ressources informatiques, que ce soit en terme de mémoire, d'espace disque, ou plus particulièrement de temps de traitement. À l'inverse, lorsque peu de données sont disponibles, les résultats sont souvent médiocres, voire inexistants.

Avec l'avènement du web, de plus en plus de textes parallèles multilingues (dans plus de deux

Anglais :

parliament	approved	the	minutes	of	the	**previous**	**sittings**	.
412	96	1,573	45	592	1,573	**1**	**1**	362

Français :

le	procès-verbal	des	**séances**	**précédentes**	est	adopté	.
692	44	292	**1**	**1**	157	35	375

Allemand :

das	parlament	genehmigt	die	protokolle	der	**vorhergehenden**	**sitzungen**	.
556	427	46	460	2	642	**1**	**1**	802

FIG. 2 – Exemple d'alignement d'hapax. Une phrase en anglais issue d'un échantillon du corpus Europarl [5] ainsi que ses traductions en français et en allemand sont présentées (la casse a été supprimée). Les nombres correspondent à l'effectif des mots dans le corpus. Les séquences d'hapax (mots dont l'effectif vaut 1) sont traductions les unes des autres.

langues) sont accessibles en grandes quantités, et de plus en plus d'applications peuvent tirer avantage d'alignements réellement multilingues.
Nous proposons une méthode qui permet l'alignement d'un nombre quelconque de langues simultanément. Contrairement aux méthodes statistiques en vogue à l'heure actuelle, elle permet naturellement le passage à l'échelle. Plus exactement, le passage à l'échelle est au cœur de la méthode.

3. Jouons avec les fréquences

3.1. Se détacher des hautes fréquences…

La plupart des méthodes d'alignement sous-phrastique sont basées sur l'exploitation des mots de haute fréquence. Intuitivement, si un mot source et un mot cible apparaissent à plusieurs reprises dans les mêmes phrases, alors il y a de grandes chances pour qu'ils soient traductions l'un de l'autre — ou du moins, pour qu'ils interviennent dans ces traductions. Les mots peu fréquents sont ainsi souvent ignorés. Ils constituent pourtant la grande majorité du vocabulaire d'un texte [2], et méritent a fortiori d'être alignés. Une solution évidente est couramment employée : il suffit d'augmenter la quantité de données en entrée, de sorte que la fréquence de chacun de ces mots rares augmente également. Devenus mots fréquents, ils peuvent être alignés sans difficulté. Cependant, en ajoutant de nouvelles données, de nouveaux mots rares sont apparus… C'est un cercle vicieux.
Si, plutôt que de compter sur les mots fréquents, nous pouvions aligner les mots rares de façon sûre, nous n'aurions plus besoin d'ajouter perpétuellement des données. Au contraire, *supprimer* des données en entrée, de façon à ce que les mots fréquents deviennent rares, suffirait pour s'acquitter de la tâche d'alignement. Moins de données implique a priori moins de traitements, donc moins de temps de calcul.

3.2. … pour se concentrer sur les faibles fréquences…

Le cas le plus extrême des mots de faibles fréquences est celui des *hapax*,[1] mots qui n'apparaissent qu'une seule fois dans un texte ou corpus. Étant des mots rares, ils sont souvent rejetés. Pourtant, nous avons montré qu'ils pouvaient être alignés de façon sûre [6]. En effet, de la même façon que des mots fréquents apparaissant souvent simultanément ont de grandes chances d'être des traductions, si une phrase et sa traduction contiennent des hapax, alors il est hautement probable que ces *séquences d'hapax* soient également traductions les unes des autres. La figure 2 donne un exemple. Remarquons dès à présent que ce principe est valable quel que soit le nombre de langues.
Typiquement, le vocabulaire d'un texte est constitué de 50% d'hapax [2], quelle que soit sa nature. La conception d'une méthode d'alignement tirant parti de ces mots rares (mais très nombreux !) est donc justifiée.
Les hapax présentent un autre avantage : ils ne peuvent avoir qu'un seul et unique sens dans le texte où ils apparaissent. L'ambiguïté des mots est un problème récurrent en traitement des

[1] Du grec *hapax legomenon*, lit. « dit une seule fois. »

Corpus d'entrée :

	Anglais		Français		Arabe
1	One coffee , please .	↔	Un café , s'il vous plaît .	↔	قهوة ، من فضلك.
2	This coffee is excellent .	↔	Ce café n' est pas mauvais .	↔	هذه قهوة ممتازة.
3	One strong tea .	↔	Un thé fort .	↔	شاي ثقيل.

⇓

« Alignements parfaits » :

Les mots :					apparaissent aux lignes :
One	↔	Un	↔		1 3
coffee	↔	café	↔	قهوة	1 2
, please	↔	, s'il vous plaît	↔	، من فضلك	1
.	↔	.	↔	.	1 2 3
This is excellent	↔	Ce n' est pas mauvais	↔	هذه ممتازة	2
strong tea	↔	thé fort	↔	شاي ثقيل	3

FIG. 3 – Extraction « d'alignements parfaits » à partir d'un mini-corpus parallèle en anglais, français et arabe. Chaque ligne du corpus est un triplet de phrases traductions les unes des autres. Les séquences de mots qui apparaissent strictement sur les mêmes lignes sont traductions les unes des autres.

langues ; la traduction automatique et l'alignement sous-phrastique n'y dérogent pas. Dans notre cas, une certaine désambiguïsation est opérée implicitement dès lors que nous supprimons des données d'entrée, car des mots deviennent hapax à travers ce processus.

3.3. … ou les deux à la fois ?

À partir des remarques précédentes, nous pourrions imaginer une méthode d'alignement sous-phrastique qui consisterait à supprimer des phrases d'entrée de façon à ce qu'un mot particulier devienne hapax. Le traitement s'appliquant à la fois sur le texte source et le(s) texte(s) cible(s), les traductions de ce mot deviendraient également hapax, ce qui permettrait d'extraire ces traductions.

Quelques expériences (non détaillées ici) ont révélé que ce principe était prometteur. Il présente un gros désavantage cependant : il rend impossible l'alignement des mots de haute fréquence. L'exemple le plus cruel est celui du point, en tant que mot marqueur de fin de phrase. En supposant que celui-ci apparaisse dans *toutes* les phrases du corpus d'entrée, la seule configuration possible par laquelle il serait hapax serait de réduire ce corpus à une unique phrase. Mais tous les mots de cette phrase deviendraient alors hapax également, ce qui empêcherait de les aligner séparément. À la figure 2, les mots « séances » et « précédentes » ne pouvaient pas être dissociés parce qu'ils étaient tous deux hapax.

Ce problème n'en est en fait pas un. Il suffit de remarquer que les alignements d'hapax ne sont qu'un cas particulier de ce que nous appellerons par la suite des « alignements parfaits. » Il s'agit de séquences de mots qui apparaissent strictement dans les mêmes phrases, quelle que soit leur fréquence. La figure 3 en donne des exemples. En pratique, l'écrasante majorité de ces alignements sont des alignements d'hapax [6], mais les mots de haute fréquence ne sont pas négligés. Remarquons encore que ce principe est valable quel que soit le nombre de langues.

4. Description détaillée de la méthode

Cette section présente en détail une méthode d'alignement sous-phrastique complète, en se basant uniquement sur les remarques de la section précédente. Elle permet l'alignement d'un nombre quelconque de langues simultanément et devrait être accessible à tous. Une implémentation libre de cette méthode, dans le langage de programmation Python, est disponible à l'adresse suivante :

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

4.1. Plus fort que multilingue : *alingue*

Comme mentionné précédemment, un des avantages de la méthode est qu'elle permet l'alignement d'un nombre quelconque de langues simultanément. Une telle chose est possible car :

1 | One$_1$ coffee$_1$,$_1$ please$_1$.$_1$ Un$_2$ café$_2$,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$ ٣. ٣ فضلك ٣ من ٣ ،٣ قهوة ٣
2 | This$_1$ coffee$_1$ is$_1$ excellent$_1$.$_1$ Ce$_2$ café$_2$ n'$_2$ est$_2$ pas$_2$ mauvais$_2$.$_2$ ٣. ٣ ممتازة ٣ قهوة ٣ هذه ٣
3 | One$_1$ strong$_1$ tea$_1$.$_1$ Un$_2$ thé$_2$ fort$_2$.$_2$ ٣. ٣ ثقيل ٣ شاي ٣

FIG. 4 – Assimilation d'un corpus multilingue à un corpus monolingue. Il s'agit du corpus d'entrée de la figure 3, mais les mots ont été discriminés par indiçage sur les langues (1 pour l'anglais, 2 pour le français et 3 pour l'arabe). Les séparations entre les langues ont été supprimées.

– la méthode est endogène : les principes sur lesquels elle repose (extraire des « alignements parfaits ») ne nécessite aucune ressource extérieure. En particulier, aucune connaissance sur les langues n'est requise. Notons que cela n'est pas nouveau en soi, la plupart des méthodes d'alignement actuelles (principalement statistiques) étant également endogènes ;
– son champ d'application ne se limite pas aux couples de langues. La figure 3 présente des exemples en trois langues, et nous pourrions en ajouter davantage sans que le moindre changement ne soit nécessaire. Plus surprenant, le principe est toujours valable avec un corpus *monolingue*. En effet, le repérage de mots qui apparaissent strictement sur les mêmes lignes (en admettant qu'une ligne correspond à une phrase) peut sans problème être effectué au sein même d'une langue isolée. Ce qu'on obtient alors n'est rien d'autre qu'un cas particulier de *collocations*, mots qui ont fortement tendance à apparaître simultanément dans un texte monolingue. Procéder ainsi avec plusieurs langues simultanément revient donc en quelque sorte à extraire des « collocations multilingues. »

Par conséquent, le processus d'alignement peut être grandement simplifié en assimilant un corpus d'entrée multilingue à un corpus monolingue. Cette transformation peut être effectuée en discriminant toutes les formes de surface des mots en fonction de leur langue d'origine. On distingue ainsi les mots de même graphie mais issus de langues différentes. Les séparations entre langues n'ont plus de raison d'être ; elles sont purement et simplement supprimées et seront rétablies après le processus d'alignement, selon l'origine des mots. Un exemple de tel corpus est présenté à la figure 4.

Ce corpus étant une abstraction de plusieurs langues ne faisant intervenir aucune connaissance sur celles-ci, nous y faisons référence par la suite en tant que corpus *alingue*. Ce corpus est le point de départ de tous les traitements ultérieurs.

4.2. Échantillonnage des données d'entrée

Le cœur de la méthode consiste à supprimer des données d'entrée, dans le but de faire décroître toutes les fréquences des mots. Ce traitement fait émerger de nouveaux « alignements parfaits, » la plupart étant constitués d'hapax. Plus précisément, un certain nombre de sous-corpus, à partir desquels les alignements sont extraits, sont créés. Trois stratégies peuvent être envisagées :

1. construire tous les sous-corpus possibles du corpus alingue, et extraire les alignements de chacun d'eux. Cette approche n'est pas raisonnable lorsque de grandes quantités de données sont en jeu, car le nombre de sous-corpus est exponentiel en la taille du corpus alingue ;
2. construire un sous-corpus tel qu'une séquence de mots donnée correspond à un « alignement parfait » dans ce sous-corpus. Cela suppose de connaître à l'avance quelles séquences doivent être alignées, ce qui rend impossible la découverte de nouveaux alignements. Une solution serait de traiter toutes les sous-séquences de mots du corpus alingue, mais cette approche ne serait alors pas plus raisonnable que la précédente en terme de complexité. Plus important, sélectionner des phrases pour biaiser une distribution de mots particulière biaise en pratique *toutes* les distributions, ce qui altère la qualité des « alignements parfaits. » Des expériences préliminaires basées sur cette approche ont fourni de très mauvais résultats ;
3. échantillonner. Cette approche est non seulement la plus simple, mais elle est surtout la plus précise, parce qu'elle n'altère en aucune manière la distribution naturelle des mots du corpus alingue.

Nous choisissons l'approche par échantillonnage. Des sous-corpus seront donc créés à partir de phrases tirées aléatoirement dans le corpus alingue.

Du fait de son côté aléatoire, on peut s'attendre à ce que deux expériences identiques effectuées à partir du même corpus d'entrée produise des résultats différents. Ces différences sont minimes en pratique. La couverture du corpus d'entrée ne peut pas non plus être garantie, mais ce problème peut aisément être contourné en extrayant des alignements à partir de nombreux sous-corpus, avec des tailles de sous-corpus variées. Le traitement d'un grand nombre de sous-corpus n'est pas un problème en soi car en traiter un est très rapide. En outre, tous les sous-corpus étant indépendants, paralléliser l'ensemble est chose facile.

Optimiser l'échantillonnage

Nous proposons ici un biais possible par lequel l'échantillonnage doit assurer une certaine couverture du corpus alingue. Ce biais ne prend pas en compte le contenu des phrases ; il influence uniquement la taille des sous-corpus. Une fois déterminée la taille du sous-corpus suivant, les phrases seront piochées aléatoirement dans le corpus alingue selon une distribution uniforme.

Nous notons x le nombre de sous-corpus de taille k à traiter. x est défini comme suit : il doit garantir que la probabilité qu'aucune des phrases d'un sous-corpus de taille k (en nombre de phrases) ne soit jamais choisie soit inférieure à un certain seuil t. Ainsi, t est un indicateur de la couverture du corpus d'entrée : plus il est proche de 0, meilleure est la couverture.

Soit n la taille du corpus d'entrée alingue ($1 \leq k \leq n$) :
– la probabilité qu'une phrase donnée soit choisie dans un échantillon de taille k est k/n ;
– la probabilité que cette phrase ne soit pas choisie est $1 - k/n$;
– la probabilité qu'aucune des k phrases ne soit choisie est $(1 - k/n)^k$;
– la probabilité qu'aucune de ces k phrases ne soit jamais choisie est $(1 - k/n)^{kx}$.

Le nombre de sous-corpus de taille k à constituer par échantillonnage est ainsi défini par $(1 - k/n)^{kx} \leq t$, ce qui donne après résolution :

$$x \geq \frac{\log t}{k \log (1 - k/n)}$$

Le traitement de x sous-corpus aléatoires de taille k garantit ainsi la couverture voulue du corpus d'entrée.

Néanmoins, plutôt que de définir à l'avance un degré de couverture particulier (ce qui implique un nombre fixe de sous-corpus à traiter), nous déduisons du résultat précédent une distribution de probabilités pour tirer aléatoirement la taille du prochain sous-corpus à traiter :

$$p(k) = \frac{-1}{k \log (1 - k/n)} \qquad \text{(à normaliser)}$$

Le numérateur ($\log t$) a été remplacé par -1 parce que t est une constante : $t \leq 1 \Rightarrow \log t \leq 0$. Cette distribution privilégie grandement les sous-corpus de petite taille. Nous avons montré que de tels sous-corpus menaient à des résultats plus précis et plus nombreux, en plus d'être beaucoup plus rapides à traiter [7].

4.3. Extraction des alignments

L'étape suivante consiste à extraire tous les « alignements parfaits » de chacun des sous-corpus obtenus par échantillonnage. Le même principe que celui présenté à la figure 3 est appliqué, à la différence près que les phrases sont désormais alingues (comme celles de la figure 4). En outre, comme nous pouvons partir du principe que les « alignements parfaits » constituent de bonnes traductions, les parties restantes des phrases sur lesquels ils apparaissent ont de grandes chances d'être des traductions également [3].

En d'autres termes, chaque « alignement parfait » est susceptible de produire deux alignements par phrase où il apparaît :

1. la séquence de mots constituée de « l'alignement parfait » lui-même, en préservant l'ordre des mots de la phrase ;
2. le complémentaire de cette séquence sur la ligne (ses contextes), ordonné également.

Ce principe est illustré à la figure 5. Un alignement peut être obtenu plusieurs fois, à partir de différents sous-corpus ou différentes lignes. Le résultat est une liste d'alignements accompagnés du nombre de fois qu'ils ont été obtenus.

Corpus d'entrée : cf. figure 4

⇓

Extraction des « alignements parfaits » et de leurs contextes (étape 3.) :

Les mots :	apparaissent aux lignes :	d'où nous extrayons :
$\text{One}_1\ \text{Un}_2$	1	$\text{One}_1\ \text{Un}_2$ $\text{coffee}_1\ ,_1\ \text{please}_1\ ._1\ \text{café}_2\ ,_2\ \text{s'il}_2\ \text{vous}_2\ \text{plaît}_2\ ._2\quad {}_3.\ {}_3$ فضلك $_3$ من $_3$ ، $_3$ قهوة
	3	$\text{One}_1\ \text{Un}_2$ $\text{strong}_1\ \text{tea}_1\ ._1\ \text{thé}_2\ \text{fort}_2\ ._2\quad {}_3.\ {}_3$ ثقيل $_3$ شاي
$\text{coffee}_1\ \text{café}_2\quad {}_3$ قهوة	1	$\text{coffee}_1\ \text{café}_2\quad {}_3$ قهوة $\text{One}_1\ \_\ ,_1\ \text{please}_1\ ._1\ \text{Un}_2\ \_\ ,_2\ \text{s'il}_2\ \text{vous}_2\ \text{plaît}_2\ ._2\quad {}_3.\ {}_3$ فضلك $_3$ من $_3$ ،
	2	$\text{coffee}_1\ \text{café}_2\quad {}_3$ قهوة $\text{This}_1\ \_\ \text{is}_1\ \text{excellent}_1\ ._1\ \text{Ce}_2\ \_\ \text{n'}_2\ \text{est}_2\ \text{pas}_2\ \text{mauvais}_2\ ._2\quad {}_3.\ {}_3$ ممتازة $_3\ \_\ {}_3$ هذه
⋮	⋮	⋮

⇓

Collecte des alignements, décompte, et rétablissement des séparations entre langues (étape 5.) :

Anglais		Français		Arabe	Décompte
One	↔	Un	↔		2
coffee , please .	↔	café , s'il vous plaît .	↔	قهوة ، من فضلك.	1
strong tea .	↔	thé fort .	↔	شاي ثقيل.	1
coffee	↔	café	↔	قهوة	2
One _ , please .	↔	Un _ , s'il vous plaît .	↔	، من فضلك.	1
This _ is excellent .	↔	Ce _ n' est pas mauvais .	↔	هذه _ ممتازة.	1
		⋮			⋮

FIG. 5 – Extraction d'alignements multilingues à partir d'un corpus alingue. Les discontinuités au sein d'une langue sont mises en évidence par un tiret bas (_).

Dans le cas général, la méthode produit en sortie des séquences de mots discontinues. Elles peuvent être filtrées subséquemment selon des critères particuliers tels que la contiguïté des mots, le nombre de langues couvertes, ou encore le nombre de mots dans une langue donnée.

Résumé de la méthode :

1. transformer le corpus d'entrée multilingue en un corpus alingue ;
2. construire un sous-corpus par échantillonnage. La taille de ce sous-corpus est déterminée par une distribution de probabilités qui privilégie les petites tailles ;
3. extraire tous les « alignements parfaits » et leurs contextes de ce sous-corpus ;
4. répéter les étapes 2 et 3 jusqu'à ce que la couverture du corpus alingue ait atteint un certain seuil, ou jusqu'à ce que l'utilisateur demande explicitement l'arrêt du traitement ;
5. récolter tous les alignements obtenus à l'étape 3, compter le nombre de fois qu'ils ont été obtenus, et rétablir les séparations entre langues.

5. Qualité des alignements

Nous ne détaillerons pas ici le processus par lequel les alignements se voient octroyer des scores, tels que probabilités de traduction (calculées à partir des décomptes associés aux alignements) ou poids lexicaux (calculés à partir des occurrences des mots dans le corpus d'entrée, témoignent de la probabilité que les mots au sein d'un alignement soient traductions les uns des autres). Des évaluations sur diverses tâches de traduction automatique fondée sur les données ont montré que la qualité des alignements produits par notre méthode étaient comparables à ceux obtenus avec la référence du domaine, Giza++ [8]. Ces évaluations ont été menées dans le cas restreint de l'alignement bilingue, car à notre connaissance, aucun autre système ne permet à ce jour l'alignement de davantage de langues simultanément.

Notons que notre système se démarque des autres de par l'utilisation d'échantillonnage : dans ce paradigme, le temps n'influence pas la qualité des sorties ; c'est la couverture du texte d'origine par les alignements qui augmente en fonction du temps. Plus l'utilisateur laisse fonctionner le système, plus le nombre d'alignements en sortie est important. En pratique, la couverture des alignements produits par cette méthode est très rapidement supérieure à celle des alignements produits par Giza++.

6. Conclusion

Nous avons décrit une méthode d'alignement complète permettant l'alignement d'un nombre quelconque de langues simultanément à partir de textes parallèles. Contrairement aux méthodes actuelles, elle repose sur l'exploitation des termes de basse fréquence, ce qui la rend particulièrement flexible dès lors que de grandes quantités de données sont en jeu. L'approche présentée, basée sur l'échantillonnage des données d'entrée, permet de produire des alignements en un temps possiblement très court : l'utilisateur peut interrompre le traitement à tout moment. Les alignements produits ont une qualité qui rivalise avec la référence du domaine. La méthode présentée est de surcroît très simple. Nous envisageons une évolution de la méthode vers un traitement en caractères plutôt qu'en mots, afin de pouvoir l'appliquer directement aux langues dont la graphie ne sépare pas les mots par des espaces (par exemple le chinois ou le japonais) sans qu'aucune segmentation préalable ne soit requise.

Bibliographie

1. Peter Brown, Stephen Della Pietra, Vincent Della Pietra, et Robert Mercer. The mathematics of statistical machine translation : Parameter estimation. *Computational Linguistics*, 19(2) :263–311, 1993.
2. Bruno Cartoni. Constance et variabilité de l'incomplétude lexicale. Dans *Actes de TALN/RECITAL 2006*, pages 661–669, Louvain, Belgique, avril 2006.
3. Ilyas Cicekli. Similarities and differences. Dans *Proceedings of SCI2000*, pages 331–337, Orlando, États-Unis, juillet 2000.
4. Emmanuel Giguet et Pierre-Sylvain Luquet. Multilingual lexical database generation from parallel texts in 20 european languages with endogenous resources. Dans *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 271–278, Sydney, Australie, juillet 2006.
5. Philipp Koehn. Europarl : A parallel corpus for statistical machine translation. Dans *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, Phuket, Thaïlande, septembre 2005.
6. Adrien Lardilleux et Yves Lepage. The contribution of the notion of hapax legomena to word alignment. Dans *Proceedings of the 3rd Language and Technology Conference (LTC'07)*, pages 458–462, Poznań, Pologne, octobre 2007.
7. Adrien Lardilleux et Yves Lepage. 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)*, pages 125–132, Waikiki, États-Unis, octobre 2008.
8. Adrien Lardilleux et Yves Lepage. Sampling-based multilingual alignment. Dans *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgarie, septembre 2009.
9. I. Dan Melamed. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. Dans *Proceedings of the Third Workshop on Very Large Corpora*, pages 184–198, Boston, États-Unis, juin 1995.
10. Franz Och et Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29 :19–51, mars 2003.
11. Michel Simard. Text-translation alignment : Three languages are better than two. Dans *Proceedings of the Joint SIGDAT Conference of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, College Park, États-Unis, 1999.