TALN 2005, Dourdan, 6–10 juin 2005

Paradocs : un système d'identification automatique de documents parallèles

Alexandre Patry et Philippe Langlais
Laboratoire de Recherche Appliquée en Linguistique Informatique
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
C.P. 6128, succursale Centre-ville
H3C 3J7, Montréal, Québec, Canada
{patryale,felipe}@iro.umontreal.ca

Mots-clefs : Corpus parallèles, apprentissage automatique, traduction automatique

Keywords: Parallel documents, machine learning, machine translation

Résumé Les corpus parallèles sont d'une importance capitale pour les applications multilingues de traitement automatique des langues. Malheureusement, leur rareté est le maillon faible de plusieurs applications d'intérêt. Extraire de tels corpus du Web est une solution viable, mais elle introduit une nouvelle problématique : il n'est pas toujours trivial d'identifier les documents parallèles parmi tous ceux qui ont été extraits. Dans cet article, nous nous intéressons à l'identification automatique des paires de documents parallèles contenues dans un corpus bilingue. Nous montrons que cette tâche peut être accomplie avec précision en utilisant un ensemble restreint d'invariants lexicaux. Nous évaluons également notre approche sur une tâche de traduction automatique et montrons qu'elle obtient des résultats supérieures à un système de référence faisant usage d'un lexique bilingue.

Abstract Parallel corpora are playing a crucial role in multilingual natural language processing. Unfortunately, the availability of such a resource is the bottleneck in most applications of interest. Mining the web for such a resource is viable solution that comes at a price : it is not always easy to identify parallel documents among the crawled material. In this study we address the problem of automatically identifying the pairs of texts that are translation of each other in a set of documents. We show that it is possible to automatically build particularly efficient content-based methods that make use of very little lexical knowledge. We also evaluate our approach toward a front-end translation task and demonstrate that our parallel text classifier yields better performances than another approach based on a rich lexicon.

1 Introduction

De nos jours, les corpus de *documents parallèles* (ensemble de documents exprimant le même contenu dans le même ordre) jouent un rôle crucial dans les applications multilingues de traitement automatique des langues (Véronis, 2000). Aligné au niveau des phrases, une tâche pouvant être accomplie avec fiabilité (Langlais *et al.*, 1998), un corpus parallèle s'avère très utile aux concordanciers bilingues (Macklovitch *et al.*, 2000) et est la pierre angulaire de la plupart des systèmes commerciaux de mémoire de traduction. Aligné au niveau des mots, une tâche maintenant bien maîtrisée (Brown *et al.*, 1993), un corpus parallèle peut servir à plusieurs applications telles que la traduction automatique, la désambiguïsation de mots ou l'extraction d'information translinguistique.

Malheureusement, il existe assez peu de *corpus parallèles* (ensemble de documents parallèles) riches et bien organisés comme le sont par exemple les *Hansards* canadiens (anglais/français), les débats parlementaires de Hongkong (anglais/chinois), les transcriptions des débats du parlement européen[1] (EUROPARL, disponibles en onze langues) ou encore les transcriptions des débats parlementaires du Nunavut (anglais/inuktitut)[2].

S'il existe également des ressources telles que la Bible qui sont traduites dans de nombreuses langues (mais pas nécessairement organisées en corpus parallèle), il n'en reste pas moins que la rareté des corpus parallèles demeure le goulot d'étranglement pour plusieurs applications d'intérêt. Plusieurs solutions ont été proposées pour palier leur absence. Il est par exemple possible d'extraire automatiquement des corpus parallèles à partir du Web (Ma & Liberman, 1999; Kraaij *et al.*, 2003; Resnik & Smith, 2003). Il est également possible de tirer profit de *corpus comparables* (corpus traitant du même sujet sans nécessairement être parallèles) (Munteanu *et al.*, 2004), voire même d'utiliser des corpus n'ayant aucune affinité (Rapp, 1999). D'autres misent à plus long terme sur des outils informatiques simplifiant la gestion des données parallèles (Hajlaoui & Boitet, 2004).

Dans cet article, nous nous intéressons à la détection des documents parallèles dans un corpus bilingue (par exemple extrait d'un site Web) à l'aide d'invariants lexicaux (par exemple données chiffrées, entités nommées, ponctuations). Cette idée était à la base d'un algorithme d'alignement bilingue de phrases décrit par Simard *et al.* (1993) ; nous montrons ici qu'elle s'applique à notre problème.

Nous décrivons en section 2 notre méthodologie et présentons les différentes métriques utilisées. Nous montrons en section 3 que notre approche permet d'identifier sans faute les paires parallèles d'une partie du corpus EUROPARL. Nous évaluons également notre approche à travers une tâche de traduction automatique et mesurons des performances supérieures à celles d'une approche faisant usage d'un lexique bilingue riche (section 4). Nous discutons en section 5 de travaux connexes et présentons en section 6 nos conclusions.

2 Méthodologie

Nous considérons dans cette étude que nous disposons de deux ensembles de documents: un ensemble $\mathcal{S}$ contenant les documents d'une langue source et un ensemble $\mathcal{T}$ contenant ceux

[1] `http://www.europarl.eu.int/home/default_fr.htm`
[2] `http://www.inuktitutcomputing.ca/NunavutHansards/`

d'une langue cible. Ces documents peuvent par exemple provenir du Web (Kraaij *et al.*, 2003) et leur langue peut avoir été identifiée automatiquement, comme ce sera le cas dans les expériences de la section 4.

Le problème que nous résolvons consiste à déterminer le sous-ensemble du produit Cartésien $\mathcal{S} \times \mathcal{T}$ qui contient les paires de documents parallèles. Nous ne faisons pas usage dans cette étude d'informations externes aux documents comme leur nom ou leurs balises structurelles, ce qui exclut l'usage d'heuristiques basées sur les noms de fichiers comme celles décrites dans (Resnik & Smith, 2003). Cette contrainte ne découle pas d'une pensée puriste, mais correspond à notre volonté d'évaluer objectivement différentes métriques n'utilisant que le contenu des documents. Ces caractéristiques externes pourraient cependant être incorporées facilement à notre approche.

L'identification des paires de documents parallèles est réalisée en deux étapes: le pointage de toutes les paires du produit Cartésien $\mathcal{S} \times \mathcal{T}$ et la classification de chacune d'elles comme parallèle ou non. Les différents pointages utilisés sont décrits dans la section 2.1 et l'algorithme de classification dans la section 2.2.

2.1 Métriques

Trois différentes familles de métriques sont utilisées pour mesurer le parallélisme de deux documents. La mesure de cosinus et la distance d'édition normalisée utilisent certaines des unités lexicales des documents: les séquences de chiffres (NOMBRE), certaines ponctuations (PUNCT) et les entités nommées (ENTITÉ). Les ponctuations que nous avons considérées sont les parenthèses, les crochets et les guillemets. De plus, nous avons considéré comme une entité nommée tout mot commençant par une lettre majuscule mais ne débutant pas une phrase. Ces types d'unités lexicales sont relativement indépendants des langues considérées. La troisième famille de métriques utilise la sortie d'un aligneur de textes au niveau des phrases pour juger du parallélisme de deux documents.

Mesure de cosinus (COS) Nous avons repris l'idée proposée par Nadeau et Foster (2004) et représenté un document par différents vecteurs où chaque dimension correspond à une unité lexicale et chaque coordonnée à la fréquence de cette unité dans le document. Dans nos expériences, chaque document est représenté par trois vecteurs: un pour les nombres, un pour les ponctuations et un pour les entités nommées. Un exemple d'une telle représentation est présenté en figure 1. La similarité entre deux documents est mesurée par la *mesure de cosinus* entre leur représentation vectorielle, mesure populaire en extraction d'information.

Distance d'édition normalisée (EDIT) La représentation vectorielle ne tient pas compte de l'ordre des unités lexicales dans le document, information qui peut être pertinente ici. Nous proposons de représenter un document par trois séquences d'unités lexicales (NOMBRE, PUNCT, ENTITÉ). Le parallélisme de deux documents peut ainsi être mesuré en comparant leurs séquences (voir la figure 1).

Pour mesurer la similarité de deux séquences, nous utilisons la distance d'édition (Levenshtein, 1966), qui compte le nombre minimal d'opérations nécessaires pour transformer la première séquence en la seconde (les opérations permises sont l'insertion, la suppression ou la substitution

d'une unité lexicale). Nous la normalisons ensuite par la longueur de la plus longue des deux séquences.

> Approximately **60**% very roughly, **60**% to **40**%, when the **60**% is paid by the tenant and **40**% is approximately paid by the Government subsidy.
>
> apiqqutiqaqqaujunga akunialuk, angiqqaugaluarakku $**60** milian kaivainnaqtuq kiinaujaqarvingmut, kisianittauq tusaqtitauvalliaqqaugama, takuvallialiqtugu $**39** milian **807** tausan ammalu taanna angiqtauguni taikkuali amiakkujut $**60** milianut tikillugu kisumut atuqtaugajaqpat ?

FIG. 1 – Si nous devions comparer les nombres dans les deux documents ci-haut (extraits anglais et inuktitut tirés des débats parlementaires du Nunavut), les représentations vectorielles utilisées pour la mesure de cosinus seraient $(0_{39}, 2_{40}, 3_{60}, 0_{807})$ et $(1_{39}, 0_{40}, 2_{60}, 1_{807})$. Alors que les représentations séquentielles pour mesurer la distance d'édition normalisée seraient $<60, 60, 40, 60, 40>$ et $<60, 39, 807, 60>$.

Scores d'alignements Une autre source d'information permettant de mesurer le parallélisme de deux documents est la sortie d'un aligneur de textes au niveau des phrases. Nous avons utilisé l'aligneur JAPA (Langlais *et al.*, 1998) qui produit une séquence d'alignements et un score global mesurant le *coût* de l'alignement produit. Les alignements qu'il produit sont de type m-n ($m, n \in \{0, 1, 2\}$) où m et n sont respectivement le nombre de phrases sources et le nombre de phrases cibles impliquées dans l'alignement.

Nous retenons cinq pointages: le ratio d'alignements *1-0* ou *0-1*, le ratio d'alignements *1-1*, le ratio d'alignements *1-2* ou *2-1*, le ratio d'alignements *2-2* et le score global d'alignement. Nous nommerons dorénavant les quatre ratios M-N et le score global COÛT. Intuitivement, le résultat de l'aligneur sur deux documents parallèles devrait contenir plusieurs alignements de type *1-1* et devrait être de faible coût.

2.2 Identification des paires parallèles

Chaque paire de documents est décrite par un ensemble de pointages. Une première approche pour identifier celles qui sont parallèles consiste à ajuster manuellement des seuils sur ces pointages, une tâche délicate ne se généralisant pas nécessairement bien. Nous avons plutôt utilisé AdaBoost (Y.Freund & Schapire, 1999), un algorithme d'apprentissage. Cet algorithme prend en entrée un ensemble de paires de documents, leurs pointages et leur *étiquette* (parallèle ou non) et produit à partir de cet ensemble d'entraînement une fonction classant une paire comme parallèle ou non à partir de ses pointages.

AdaBoost est un algorithme d'apprentissage itératif combinant plusieurs *classificateurs faibles* (classificateur juste plus d'une fois sur deux) en un classificateur plus robuste. À chaque itération, un classificateur faible est entraîné à reconnaître l'étiquette de toutes les paires de documents (à partir des pointages) en accordant plus d'importance à celles qui ont été moins bien étiquetées par les classificateurs faibles précédents. Les itérations se poursuivent jusqu'à ce qu'un classificateur faible ait un ratio d'erreur supérieur ou égal à 50% ou jusqu'à ce qu'un nombre maximal (fixé à l'avance) d'itérations ait été atteint. Le classificateur retourné par AdaBoost fait voter les différents classificateurs faibles afin de déterminer si une paire est parallèle

ou non.

Dans nos expériences, nos classificateurs faibles étaient des réseaux neuronaux (Bishop, 1996) à une couche cachée de cinq neurones et après quelques expériences informelles, nous avons décidé de borner le nombre d'itérations d'AdaBoost à 75. L'entraînement et les tests ont été réalisés à l'aide du logiciel PLEARN [3].

3 Expérience contrôlée

EUROPARL est un corpus parallèle tiré de la transcription des débats parlementaires européens s'étant tenus entre avril 1996 et septembre 2003 (Koehn, 2002). Les débats parlementaires européens sont traduits en onze langues, mais nous nous sommes concentrés sur les traductions anglaises et espagnoles. Notre corpus était composé de 487 textes anglais et de 487 textes espagnols ayant en moyenne environ 2800 phrases chacun.

3.1 Protocole d'évaluation

Parce que les paires de documents parallèles sont bien identifiées dans EUROPARL, les différentes configurations ont été comparées sur la base de leur *précision*, de leur *rappel* et de leur *f-mesure* (moyenne harmonique de la précision et du rappel). La précision (resp. rappel) est le ratio du nombre de paires vraiment parallèles que le classificateur a identifiées sur le nombre total de paires que le classificateur a identifiées (resp. sur le nombre total de paires parallèles dans le corpus). La précision indique la qualité de l'ensemble des paires trouvées et le rappel sa couverture.

Les différentes configurations ont été évaluées à l'aide d'une *validation croisée en cinq étapes*. Le produit cartésien $\mathcal{S} \times \mathcal{T}$ a été partitionné aléatoirement en cinq sous-ensembles de même taille. Ensuite, cinq expériences ont été lancées en testant chaque fois sur un sous-ensemble différent et en entraînant avec les paires ne faisant pas partie de ce sous-ensemble de test.

3.2 Système de référence (LEXIQUE)

Pour mettre en contexte les performances de nos différents classificateurs, un système de référence utilisant un *lexique bilingue* a été mis au point. Le lexique bilingue qui a été utilisé contient plus de 70 000 entrées et provient du projet PYTHOÑOL[4], qui vise à aider les locuteurs anglais à apprendre l'espagnol.

Un document est représenté par l'ensemble de ses *mots rares* (dans le cadre de ce projet, les mots rares sont ceux n'apparaissant qu'une seule fois dans le document) présents dans le lexique bilingue. Chaque document source est ensuite apparié avec le document cible partageant avec lui le plus grand nombre de mots rares.

[3] `http://plearn.sourceforge.net`
[4] `http://sourceforge.net/projects/pythonol/`

Configuration							Performances (%)		
COS	EDIT	NOMBRE	PUNCT	ENTITÉ	COÛT	M-N	précision	rappel	f-mesure
	√	√	√	√			100	100	100
√	√	√	√	√	√	√	99.8	99.8	99.8
	√	√					98.3	99.8	99.0
	√	√			√		96.6	99.8	98.1
					√		85.8	99.8	92.1
						√	65.6	99.4	77.1
					√	√	49.3	99.4	62.7
√		√	√	√			24.6	99.2	38.7
√		√					12.4	98.9	21.8

TAB. 1 – Précision, rappel et f-mesure de différentes configurations d'entraînement du classificateur. Notez que comme les f-mesures rapportées sont des moyennes sur les cinq étapes de la validation croisée, elles ne sont pas toujours cohérentes avec les mesures de précisions et de rappels.

3.3 Résultats

Nous avons entraîné des classificateurs sur plusieurs combinaisons des pointages décrits dans la section 2.1. Leurs performances sont présentées dans la Table 1. La meilleure de nos configurations et le système de référence ont tous deux obtenus des résultats parfaits.

Les meilleurs performances des métriques basées sur la distance d'édition semblent confirmer l'hypothèse selon laquelle l'ordre des unités lexicales est importante pour l'identification des documents parallèles. Il est à noter que le seul usage de la distance d'édition sur les nombres amène une f-mesure de 99%, ce qui suggère que les nombres sont de très bons indices de parallélisme pour ce genre de corpus. En effet, les débats parlementaires contiennent plusieurs nombres stables comme des dates, des numéros de lois ou encore les comptes de votes.

On observe également que les configurations utilisant les pointages d'alignements n'amènent pas de bons résultats. L'usage des ratios de types d'alignements donne en particulier une f-mesure moyenne inférieure d'au moins 20% aux meilleures configurations et ont été instables dans les différentes étapes de la validation croisée.

4 Tâche réelle

Nous avons montré dans la section précédente qu'il était possible d'identifier parfaitement les paires parallèles d'un corpus bilingue comme EUROPARL. Nous voulons maintenant mesurer si des performances satisfaisantes peuvent être obtenues dans un contexte d'utilisation plus représentatif. Nous avons pour cela aspiré le site Web de la *Pan American Health Organization*[5]. Bien qu'en principe simple, cette tâche s'est avérée particulièrement délicate (nombreux formats propriétaires, absence d'une nomenclature pour nommer et identifier les différentes ressources bilingues).

Le corpus résultant, PAHO, totalise 6878 documents dont 2523 ont été identifiés comme étant

[5] `http://www.paho.org`.

anglais (et 4355 comme espagnols) par SILC[6], l'outil que nous avons utilisé pour identifier la langue de chaque document. Au total, ce corpus totalise plus de 10 millions de paires potentielles. Chaque document contient en moyenne environ 180 phrases. Une inspection informelle du corpus a révélé que plusieurs de ces documents sont identiques ou très similaires et que certains sont bilingues.

4.1 Protocole d'évaluation

Pour cette expérience, nous avons mesuré l'impact de nos différents extracteurs de paires parallèles sur une tâche de traduction automatique (TA) de l'espagnol vers l'anglais. Deux raisons majeures ont mené à ce choix. Premièrement, l'identification de documents parallèles n'a d'intérêt que dans un cadre applicatif donné ; la traduction étant l'application bilingue par excellence. Deuxièmement, nous ne connaissons pas les documents parallèles du corpus PAHO ce qui complique les calculs de précision et rappel.

Le moteur de traduction que nous utilisons ici est un moteur probabiliste état de l'art (Koehn *et al.*, 2003). L'avantage d'un tel choix réside dans le fait que l'obtention d'un tel moteur est entièrement automatique une fois un corpus parallèle identifié.

Afin d'évaluer les traductions produites, nous avons téléchargé 520 nouvelles phrases du site de la *Pan American Health Organization* avec leur traduction. Pour les mêmes raisons d'automaticité, nous mesurons la qualité de nos traductions à l'aide de quatre métriques couramment utilisées en TA: deux taux d'erreurs au niveau des phrases (SER) et des mots (WER) et deux mesures de précision n-grammes (BLEU et NIST) calculées par le script `mteval`[7].

Les deux premières métriques varient entre 0 et 100 où 0 représente une traduction parfaite. SER (pour *Sentence-Error-Rate*) est le ratio des phrases produites par le moteur de TA qui sont différentes de la référence. WER (pour *Word-Error-Rate*) calcule la distance d'édition normalisée entre les mots de la traduction produite et ceux de la traduction de référence. BLEU et NIST comptent le nombre de séquences partagées entre la traduction automatique et la traduction de référence en donnant plus d'importance aux séquences plus longues. Le score BLEU varie entre 0 et 1 (où 1 est le score de la référence) alors que le score NIST n'est pas normalisé[8].

En plus de l'évaluation à l'aide d'un moteur de TA, la précision (voir la section 3.1) de chaque configuration a été calculée manuellement.

4.2 Résultats

Nous avons comparé les performances de notre moteur de TA lorsqu'il est entraîné sur quatre corpus parallèles différents. Le corpus COS-TOUS a été généré à l'aide de la mesure de cosinus sur les nombres, sur les ponctuations et sur les entités nommées (ligne 8 de la Table 1). Le corpus EDIT-TOUS a été obtenu à l'aide de la configuration ayant obtenue les meilleurs résultats sur EUROPARL, la distance d'édition normalisée sur les nombres, sur les ponctuations et sur les entités nommées (ligne 1 de la Table 1). Le corpus LEXIQUE a été produit à l'aide du système de référence (basé sur l'utilisation d'un lexique bilingue). Finalement, le corpus

[6] `http://rali.iro.umontreal.ca`.

[7] Disponible à l'adresse `http://www.nist.gov/speech/tests/mt/mt2001/resource`.

[8] Son calcul sur la référence produit dans notre cas une valeur de 13.11.

Corpus parallèle	N	SER	WER	NIST	BLEU	précision
COS-TOUS ∪ EDIT-TOUS	494	99.42	60.02	5.3125	0.2435	99.0
LEXIQUE	529	99.42	61.67	5.1989	0.2304	89.2
EDIT-TOUS	390	99.42	61.53	5.1342	0.2290	99.0
COS-TOUS	333	99.23	62.23	5.1629	0.2256	99.7

TAB. 2 – Performances de notre moteur de TA lorsque entraîné sur les corpus parallèles retournées par différentes configurations où N est le nombre de paires identifiées comme étant parallèles.

COS-TOUS ∪ EDIT-TOUS est l'union de COS-TOUS et de EDIT-TOUS. Une inspection de ces deux corpus nous a en effet révélé qu'ils ne partagent que 229 paires de documents.

Les classificateurs identifiant les paires parallèles ont été entraînés sur les paires du corpus EUROPARL. Pour chaque configuration, les paires partageant un document ont été rejetées (ce sont les paires les plus incertaines). Les performances de traduction des moteurs probabilistes correspondant sont présentées en Table 2.

Contrairement à nos expériences sur EUROPARL, la mesure de cosinus et la distance d'édition ont des performances comparables. Cela pourrait s'expliquer par la plus petite taille des documents de PAHO (rendant l'ordre des caractéristiques moins important) et par l'étape de suppression des paires partageant un document. Nous observons que les performances du moteur entraîné sur le corpus COS-TOUS ∪ EDIT-TOUS sont meilleures que celles du moteur entraîné sur le corpus LEXIQUE, et ce même si ce dernier contient plus de paires. Ce résultat est particulièrement intéressant puisqu'il montre qu'il n'est pas nécessaire de réunir un lexique bilingue. Un autre résultat encourageant est la forte précision de tous nos classificateurs (99% ou mieux).

5 Travaux connexes

Ce travail a été inspiré de celui de Nadeau et Foster (2004). Les auteurs ont proposé l'utilisation de la mesure de cosinus sur les nombres, les ponctuations, les entités nommées et le nombre de paragraphes pour détecter les paires de documents parallèles d'un corpus bilingue de communiqués. Ils ont montré qu'à l'aide d'un filtre sur les dates de publication, ils pouvaient identifier les documents parallèles du *Groupe Canada NewsWire*[9] avec une grande précision.

Nous avons étendu cette idée de trois façons. Premièrement nous avons validé l'utilisation d'unités lexicales invariantes sur des corpus de natures différentes. Deuxièmement, nous avons montré que l'ordre de ces caractéristiques est porteur d'information. Finalement, nous avons testé l'impact de cette approche sur une tâche concrète: la traduction automatique.

Notre travail, même si mené de façon indépendante, partage des points communs avec celui de Munteanu *et al.* (2004). Les auteurs ont montré qu'un moteur de traduction pouvait bénéficier d'un corpus parallèle extrait automatiquement de corpus comparables. L'approche qu'ils ont proposée est analogue à la nôtre: ils entraînent un classificateur (dans leur cas par une approche de *maximum entropie*) pour identifier les paires de phrases en relation de traduction (alors que nous travaillons au niveau du document). Ils font cependant l'hypothèse qu'un corpus parallèle

[9] http://www.newswire.ca

est disponible afin d'entraîner un modèle de traduction qu'ils utiliseront ensuite pour aligner les phrases au niveau des mots. Nous pensons que cette approche est complémentaire à la nôtre. Notre approche serait plus adaptée pour les corpus où nous savons à priori qu'ils contiennent plusieurs documents parallèles.

6 Conclusions et travaux futurs

Nous avons présenté une approche complètement automatique permettant d'identifier les paires de documents parallèles d'un corpus bilingue et ce, à l'aide d'un nombre restreint d'informations lexicales. Nous avons plus précisément étudié l'usage de certains invariants lexicaux comme les nombres, certaines ponctuations et les entités nommées. Nous avons montré que cette approche amenait des résultats comparables (voire supérieurs) à une approche de référence faisant usage d'un lexique bilingue riche.

L'un des avantages majeurs de notre approche est sa souplesse que nous devons à l'utilisation d'un algorithme d'apprentissage à la fois simple à mettre en place et efficace. Il est donc tout à fait possible d'étendre la liste des traits (pointages) que nous avons utilisés pour représenter nos documents. Ajouter comme trait le nombre d'entrées d'un lexique bilingue que partagent deux documents serait par exemple particulièrement aisé.

Nous travaillons actuellement sur l'amélioration de deux limitations du système proposé. Premièrement, nous avons considéré systématiquement dans cette étude toutes les paires du produit cartésien entre l'ensemble des documents sources et cibles. Cela impose des temps de traitement qui peuvent vite devenir prohibitifs. Certaines heuristiques conservatrices peuvent être appliquées pour limiter l'espace de recherche des paires de documents parallèles. Nous pouvons par exemple éliminer les paires de documents dont le rapport de longueur est anormalement grand ou faible (Kraaij *et al.*, 2003).

Deuxièmement, nous aimerions vérifier l'efficacité de l'approche si seulement une partie de chaque document est inspectée (par exemple les premières phrases). Cela diminuerait le temps de calcul des pointages et par le fait même accélérerait le processus au complet.

Références

BISHOP C. M. (1996). *Neural networks for pattern recognition*. Oxford University Press.

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

HAJLAOUI N. & BOITET C. (2004). PolyphraZ : a tool for the quantitative and subjective evaluation of parallel corpora. In *Proc. of the International Workshop on Spoken LanguageTranslation*, p. 123–129, Kyoto, Japan.

KOEHN P. (2002). Europarl: A multilingual corpus for evaluation of machine translation. Draft.

KOEHN P., OCH F. J. & MARCU D. (2003). Statistical phrase-based translation. In *Proceedings of the Second Conference on Human Language Technology Reasearch (HLT)*, p. 127–133, Edmonton, Alberta, Canada.

KRAAIJ W., NIE J.-Y. & SIMARD M. (2003). Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, **29**(3), 381–419.

LANGLAIS P., SIMARD M. & VERONIS J. (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 711–717, Montréal, Quebec, Canada.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, **6**, 707–710.

MA X. & LIBERMAN M. (1999). Bits: A method for bilingual text search over the web. In *Machine Translation Summit VII*, Kent Ridge Digital Labs, National University of Singapore.

MACKLOVITCH E., SIMARD M. & LANGLAIS P. (2000). Transsearch: A free translation memory on the world wide web. In *Second International Conference On Language Resources and Evaluation (LREC)*, volume 3, p. 1201–1208, Athens Greece.

MUNTEANU D. S., FRASER A. & MARCU D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL*, p. 265–272.

NADEAU D. & FOSTER G. (2004). Real-time identification of parallel texts from bilingual news feed. In *CLINE 2004*, p. 21–36: Computational Linguistics in the North East.

RAPP R. (1999). Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th conference on Association for Computational Linguistics*, p. 519–526: Association for Computational Linguistics.

RESNIK P. & SMITH N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, **29**, 349–380. Special Issue on the Web as a Corpus.

SIMARD M., FOSTER G. F. & ISABELLE P. (1993). Using cognates to align sentences in bilingual corpora. In *CASCON '93: Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research*, p. 1071–1082: IBM Press.

J. VÉRONIS, Ed. (2000). *Parallel Text Processing, Alignment and Use of Translation Corpora*. Kluwer Academic.

Y.FREUND & SCHAPIRE R. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, **14**(5), 771–780. Appearing in Japanese, translation by Naoki Abe.