TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Détection de zones parallèles à l'intérieur de bi-documents

Charlotte Lecluze[1, 2] Romain Brixtel[1]
(1) GREYC, Université de Caen Basse-Normandie, France
(2) Pertimm, Asnières-sur-Seine, France
prénom.nom@{unicaen.fr, pertimm.com}

Résumé. Dans cet article, nous présentons un des principaux axes de nos travaux en matière d'alignement multilingue et endogène de multidocuments. La question que nous abordons ici est celle du parallélisme entre les différents *volets* d'un bi-document. Loin de présupposer un parallélisme global entre ces volets, nous proposons une méthode endogène pour définir en contexte les zones qui maximisent le parallélisme : documents, séquences de paragraphes, paragraphes,... Cette étape repose sur un appariement de N-grammes de caractères répétés et constitue une étape préalable à un alignement lexical.

Abstract. In this paper, we present one of the main axes of our work in progress concerning multilingual and endogenous alignment of multidocuments. The question that we address here is the one of the parallelism between the different components of a bi-document. Far from presupposing a global parallelism between these components, we propose a endogenous method for defining in context the areas that maximize the parallelism: documents, sequences of paragraphs, paragraphs, ... This step is based on a matching of repeated character N-grams and is a preliminary step to a lexical alignment.

Mots-clés : détection et alignement de zones, appariement de N-grammes de caractères, collection de multidocuments.

Keywords: area detection and alignment, character N-grams matching, set of multidocuments.

1 Introduction

L'opération traduisante, réalisée par l'humain et visant à traduire un document donné dans une langue source dans une langue cible, donne lieu à de nombreuses modifications dans l'organisation interne des différents *volets* ou versions d'un multidocument tant au niveau micro qu'au niveau macro. Cette possibilité que l'ordre macro ne soit pas globalement maintenu d'un *volet* à un autre, par exemple lorsqu'un résumé présent au début d'un volet et à la fin d'un ou de plusieurs autres, ou encore en présence d'une liste triée par ordre alphabétique (cf. figure 1), constitue un des principaux obstacles aux méthodes d'alignement de documents.

Cet article présente un outil graphique de détection du parallélisme entre les volets, endogène et sans traitement préalable des multidocuments. Notre objectif est double :

- définir si la traduction est globalement littérale entre deux volets ;
- pour les cas où l'hypothèse de parallélisme ne s'avère pas globalement vérifiée, délimiter et aligner les zones entre les volets.

Plusieurs courants existent dans le domaine de l'alignement. Ils se distinguent notamment par le grain qu'ils proposent d'analyser : phrases, paragraphes, documents, ... Nous consacrerons donc la section 2 à un rapide tour d'horizon des principales méthodes d'alignement proposées à ce jour, avec un intérêt particulier pour l'alignement de documents. Dans la section 3, nous présenterons les points d'ancrage que nous utilisons dans le cadre de notre méthode d'alignement de zones à l'intérieur de bi-documents à la fois endogène et indépendante des langues. Enfin, dans la section 4, nous exposerons les outils graphiques de détection de zones que nous avons mis en place.

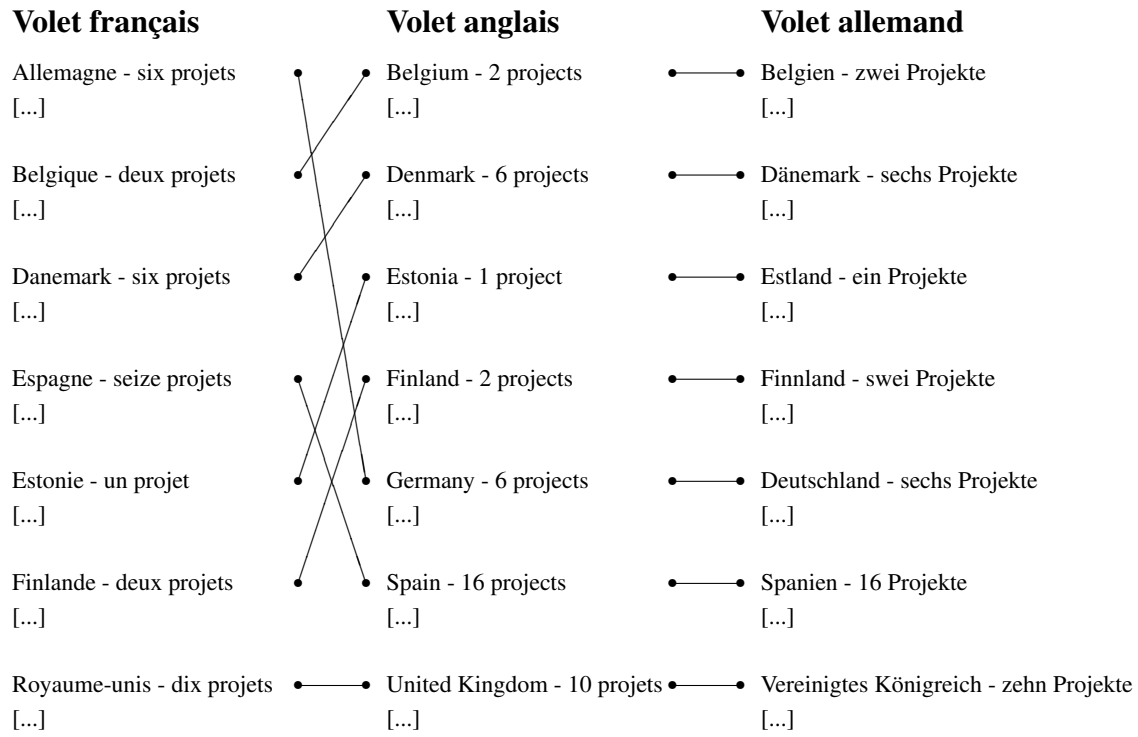

Figure 1: Maintien de l'ordre et inversions entre les différents volets d'un multidocument (communiqué de presse IP/05/1157 de l'Union Européenne) en anglais, français et allemand contenant des paragraphes triés par ordre alphabétique. Nous utilisons les [...] pour symboliser le contenu d'un paragraphe, dont nous ne conservons ici que le début soit le nom du pays dont il traite.

2 Contexte

Les méthodes d'alignement sous-phrastiques appliquées à des phrases alignées, si diverses soient-elles, trouvent toutes leur limite dans le fait qu'elles présupposent la disponibilité de corpus préalablement alignés en phrases (Hansard,...). De tels corpus sont cependant peu nombreux.
En revanche, l'accessibilité grandissante à des documents en différentes langues est avérée et laisse réellement envisager la pratique d'opérations de rétro-ingénierie massives et peu supervisées sur ces documents issus du travail du traducteur humain. Ces pratiques permettent d'extraire des informations linguistiques et des ressources lexicales pouvant être utiles tant aux traducteurs, qu'aux lexicographes, aux linguistes ou aux terminologues.
Plusieurs méthodes d'alignement sous-phrastique à partir de documents non préalablement alignés en phrases ont été proposées. (Simard *et al.*, 1993; Church, 1993; Church & Helfman, 1993; Dagan *et al.*, 1993), établissant un lien entre la similitude de graphie et la similitude de sens, proposent de s'appuyer sur les similitudes de chaînes de caractères. Néanmoins, si ces similitudes sont fréquentes entre les langues indo-européennes, elles s'avèrent plus rares et insuffisantes entre les langues indo-européennes et les langues asiatiques par exemple. (Fung & Church, 1994; Fung & Mckeown, 1994) quant-à-lui, à travers les systèmes K-vec et DK-vec, a proposé une méthode d'alignement de documents basée sur une similitude de répartition de mots. Les systèmes reposant sur la similitude de répartition de mots se heurtent à la nature flexionnelle de certaines langues, certaines mots pouvant recouvrir plusieurs formes selon leur fonction dans la phrase.

Para sur les outils de détection graphique Chang (Chang & Chen, 1997)

Nos recherches nous ont menés à nous intéresser aux N-grammes de caractères répétés propices à révéler des similitudes à la fois monolingues et multilingues susceptibles de nous aider dans la détection de zones maximisant le parallélisme entre les volets d'un bi-document. Grâce à eux, nous réalisons un alignement grossier des volets de bi-documents à partir d'une collection de bi-documents.

3 Appariemment multilingue de N-grammes de caractères répétés

Notre méthode consiste à obtenir de façon endogène et la plus indépendante des langues une série de points d'ancrage entre deux documents traductions. Le corpus que nous utilisons est constitué de X communiqués de presses de l'union européenne en X langues, disponibles sur le site Europa, le portail de l'Union Européenne. Notre travail se situe dans la lignée de ceux de Cromières (Cromieres, 2006), nous procédons à une recherche de N-grammes de caractères en contexte, indépendamment de leur taille. Après un découpage de l'ensemble des chaînes de notre corpus de documents entiers, pour lesquels nous supposons ne pas disposer d'alignement de phrases, notre critère d'extraction des chaînes est la répétition. Précisons que nous ne nous intéressons qu'aux chaînes répétées de longueur maximale, i.e. pour une chaîne de caractères répétée donnée, nous filtrons toutes les chaînes incluses de même effectif. L'intérêt que nous percevons dans ce découpage est double : révéler des facteurs communs monolingues et mettre en évidence des correspondances multilingues. Nous estimons le lexique entre deux langues, grâce à un algorithme qui prend en compte les similitudes de répartitions entre les chaînes de caractères d'effectifs proches sur la collection de bi-documents dans ces langues.

La formule de distance que nous utilisons est proche de celle du cosinus. Elle consiste à faire, pour deux N-grammes de caractères de deux langues différentes et d'effectifs proches, la somme des différences d'effectifs par document et sur l'ensemble de la collection, divisé par l'effectif du N-gramme le plus fréquent des deux :

$$distance(s_1,s_2) \quad = \quad \frac{\sum_{doc} |effectif(s_1,doc) - effectif(s_2,doc)|}{max(effectif\_corpus(s_1), effectif\_corpus(s_2))}$$

Ici exemple d'un appariement de N-grammes de caractères grec/français : (à reprendre et compléter)

langue	graphie	effectif corpus	effectif par multi-document			
			doc_0	doc_1	[...]	doc_{670}
el	' αερολιμέν'	(23)	4	2	[...]	3
fr	'aéroports'	(21)	4	2	[...]	2

Table 1: Exemple

4 Matrice : Outil graphique de détection de zones

Qu'est-ce qu'une matrice de points? Quelles sont les conventions de constructions? Chaque axe de nos matrices, axe horizontal et axe vertical, correspond à un des deux volets d'un multidocument à diagnostiquer. Une matrice est constituée de points, il y autant de points sur une ligne d'un axe que de zones définies en paramètre (les zones peuvent se chevaucher). Ainsi, un point se situe à l'angle de deux zones, x et y. La couleur d'une pixel sur la matrice est fonction de la densité de liens, i.e. d'appariements, présents entre les zones de textes qu'elle représente : plus un point est noir, plus il y a de liens entre la zone x et la zone y.

$$score(s_1,s_2) \quad = \quad \frac{nb\_link(s_1,s_2)}{max\_link(s_1)}$$

Si deux documents sont traduits de façon linéaire alors une diagonale se dessine de l'angle supérieur gauche à l'angle inférieur droit de la matrice. Une diagonale cassée signifie au contraire l'existence d'inversion dans l'ordre de la traduction.

figure 2 : 2 matrices une avec 1 diag et une avec plusieurs droites, à partir de IP/05/1157 cf figure 1

La méthode de détection est guidée par le modèle, l'attente que nous formulons est une diagonale au milieu de la matrice. Étant entendu que nous partons du principe que nos documents, lorsqu'ils portent le même nom, sont bien en relation de traduction les matrices nous servent à répondre à 2 questions principales, essentielles pour la suite des opérations à mettre en place en vue d'un alignement lexical des bi-documents :
- Ces documents sont-ils traduits de façon linéaire? Autrement dit sont-ils géométriquement parallèles? Leurs appariement laissent-ils apparaître une diagonale au centre de la matrice ou au contraire plusieurs bouts de droites?
- Pour ce dernier cas, quelles sont-les zones de textes qui maximisent le parallélisme? Comment délimiter ces zones de la façon la plus précise pour limiter au maximum les erreurs de l'alignement lexical qui en découlera.

5 Matrice : Segmentation arbitraire en zones et calcul de leur similarité

Exemple de IP/05/1157 2 Méthodes : => présegmentation en para, ali de para et fusion de para pour faire des bi-zones (pas de chevauchement) => ali flou de zones de textes de densité similaire (avec chevauchement)

Références

CHANG J. S. & CHEN M. H. (1997). An alignment method for noisy parallel corpora based on image processing techniques. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, p. 297–304, Madrid, Spain: Association for Computational Linguistics.

CHURCH K. W. (1993). Char_align: a program for aligning parallel texts at the character level. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, p. 1–8, Stroudsburg, PA, USA: Association for Computational Linguistics. ACM ID: 981575.

CHURCH K. W. & HELFMAN J. I. (1993). Dotplot: A program for exploring Self-Similarity in millions of lines of text and code. *Journal of Computational and Graphical Statistics*, **2**(2), 153–174. ArticleType: research-article / Full publication date: Jun., 1993 / Copyright © 1993 American Statistical Association, Institute of Mathematical Statistics and Interface Foundation of America.

CROMIERES F. (2006). Sub-sentential alignment using substring co-occurrence counts. In *Proceedings of the 21st International Conference on computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, p. 13–18, Sydney, Australia: Association for Computational Linguistics.

DAGAN I., CHURCH K. W. & GALE W. A. (1993). Robust bilingual word alignment for machine aided translation. *IN PROCEEDINGS OF THE WORKSHOP ON VERY LARGE CORPORA*, **1**, 1—-8.

FUNG P. & CHURCH K. W. (1994). K-vec: a new approach for aligning parallel texts. In *Proceedings of the 15th conference on Computational linguistics - Volume 2*, p. 1096–1102, Kyoto, Japan: Association for Computational Linguistics.

FUNG P. & MCKEOWN K. (1994). Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. *IN PROCEEDINGS OF THE FIRST CONFERENCE OF THE ASSOCIATION FOR MACHINE TRANSLATION IN THE AMERICAS, 81–88*, p. 81—-88.

SIMARD M., FOSTER G. F. & ISABELLE P. (1993). Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing - Volume 2*, p. 1071–1082, Toronto, Ontario, Canada: IBM Press.