Les N-grammes de caractères comme moyen de comparaison à grande échelle de corpus multilingue

Charlotte Lecluze

Recherche d'une granularité optimale pour l'alignement multilingue : N-grammes de caractères ou N-grammes de mots ?

Charlotte Lecluze
GREYC, Université de Caen Basse-Normandie, France
charlotte.lecluze@unicaen.fr
Pertimm, Asnières-sur-Seine, France
charlotte.lecluze@pertimm.com

Résumé. Dans cet article, nous présentons un des principaux axes de nos travaux en matière d'alignement multilingue et endogène de textes et de segments de textes. Nous soulevons la question de la granularité optimale, N-grammes de caractères ou N-grammes de mots, pour mettre en évidence des correspondances sémantiques entre des documents traductions les uns des autres.

Abstract. In this paper, we present one of the main axes of our work in progress concerning multilingual and endogenous alignment of texts and text segments. We raise the question of the best granularity, characters N-grams or word N-grams, to bring out some semantic correspondences between documents translations of each others.

Mots-clés : Linguistique de corpus, Traitement Automatique des Langues (TAL), méthode fondée sur des N-grammes de caractères, alignement.

Keywords: Corpus linguistics, Natural Language Processing (NLP), character N-gram based method, alignment.

1 Introduction

L'accessibilité grandissante à des informations en différentes langues laisse envisager la pratique d'opérations de rétro-ingénierie massives et peu supervisées sur des 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 courants existent dans le domaine de l'alignement. Ils se distinguent notamment par le grain qu'ils proposent d'aligner : mots, chunks, propositions,... Nous consacrerons donc la section 2 à un rapide tour d'horizon des principales méthodes proposées à ce jour du point de vue du grain qu'elles proposent d'aligner. Dans la section 3, nous aborderons l'intérêt et les limites d'une segmentation en N-grammes de caractères. Enfin, dans la section 4, nous exposerons nos perspectives en matière d'implémentation et d'évaluation de ces principes dans un système d'alignement.

2 Contexte

Les méthodes d'alignement automatique proposées vont du tout statistique (Gale & Church, 1993), à des méthodes hybrides, alliant tant des indices de longueurs, de fréquences, que des indices lexicaux (Langlais, 1997). Historiquement, les recherches ont d'abord porté sur des méthodes d'alignement de phrases. Mais la quasi-résolution de ce problème, et surtout le constat que l'alignement de phrases est intimement lié à celui des mots, et plus généralement aux unités sous-phrastiques, quelles qu'elles soient, ont fait émerger rapidement des méthodes proposant d'aligner aux grains inférieurs à celui de la phrase : mots (Gale & Church, 1991), chunks (Zhou *et al.*, 2004), propositions (Nakamura-Delloye, 2007), ...
Les systèmes d'alignement et d'extraction d'information au sens large passent généralement par une segmentation en mots. La question du statut du mot se pose.
En TAL, le mot est généralement décrit comme un segment de discours compris entre deux espaces et/ou ponctuation. Or ce mot graphique, au travers des langues, recouvre des réalités très diverses d'un point de vue sémantique. En outre, certains systèmes d'écriture ne marquent pas les frontières du mot par des espaces, c'est le cas notamment en chinois.
Le concept de mot est donc complexe. Elle dépend en fait du point de vue adopté : lexical ou graphique. Ces deux points de vue ne sont pas toujours en correspondance.

langue	Mot polylexical	Nombres de mots graphiques
fr	transport en commun	3 mots graphiques
en	public transport	2 mots graphiques
hu	a tömegközlekedés	2 mots graphiques
fi	joukkoliikenne	1 mot graphique

Tableau 1: Illustration du décalage interlangue entre le niveau lexical et le niveau graphique du concept de mot, à partir de l'exemple de "transport en commun"

Cette question est d'autant plus complexe que l'on a à traiter des *mots polylexicaux* (ou complexes) à savoir "toute unité composée de deux mots simples ou mots dérivés préexistants [...] les mots polylexicaux (ou complexes) peuvent être soudés (et alors, du point de vue informatique, ils peuvent être assimilés à des mots simples) [...] ou comporter un séparateur"[1]. La forme graphique d'une unité lexicale composée tient de propriétés intralangues. Elle dépend des particularités morphologiques de flexions et de dérivations de chaque langue.
Au regard de ces caractéristiques morphologiques, le mot graphique n'apparaît pas suffisamment universel pour répondre au besoin de comparativité d'un système multilingue d'alignement et d'extraction d'information et qui plus est sans ressource. Ainsi, nous envisageons de réaliser en contexte un découpage en N-grammes de caractères[2] pour faire émerger des correspondances que ne révèle pas un découpage en mots.

3 Les N-grammes de caractères

La notion de N-grammes de caractères est déjà utilisée pour l'identification d'auteurs (Jardino, 2006), l'identification de la langue (Dunning, 1994), l'analyse de l'oral, la catégorisation de textes (Damashek, 1995), la classification numérique multilingue de documents (Biskri & Delisle, 2001) ou encore la recherche d'informations (Majumder *et al.*, 2002; Mcnamee & Mayfield, 2004). Cependant, à notre connaissance, il n'existe qu'une tentative (Cromières, 2006) pour

[1] G. Gross (2004) cité par F. Neveu, *Dictionnaire des sciences du langage*, 2004, Armand Collin
[2] Nous utilisons N de façon générique, sa valeur n'étant pas prédéfinie.

appliquer une telle méthode à l'alignement multilingue. Cromières réalise un alignement sous-phrastique par calcul de coefficients de corrélation entre des N-grammes de caractères. Il conseille particulièrement l'utilisation du grain caractère sur les langues asiatiques, où le mot n'est pas facile à définir. Pour les langues occidentales, Cromières a également appliqué son algorithme au grain caractère sur un petit corpus de bi-phrases tirées du corpus Europarl, à cause de limites de mémoire. Si, dans les applications de TAL évoquées ci-dessus, les n-grammes de caractères ont un nombre de caractères constants défini a priori, ce sont généralement des bi-grammes ou des tri-grammes de caractères (4-grammes ou 5-grammes dans le cas de Mcnamee & Mayfield, 2004), chez Cromières leur taille n'est pas prédéfinie.
Notre travail se situe dans la lignée de ceux de Cromières, 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.

3.1 Capacité des N-grammes de caractères à révéler des facteurs communs monolingues

Pour un document donné dans une langue, une segmentation en N-grammes de caractères met en évidence des facteurs communs que ne révèle pas un découpage en N-grammes de mots.

Langue	Mots	Chaînes de caractères
fr	transport, transports, transporter, transportation	transport-

Tableau 2: Mise en évidence de la chaîne de caractère commune à quatre mots formés par dérivation

3.2 Capacité des N-grammes de caractères à mettre en évidence des correspondances multilingues

Le problème de l'alignement multilingue est un problème de similarités et de différences de : sens, graphie et répartition. Les facteurs communs monolingues, d'ordre graphique, précédemment révélés, mettent en évidence des segments de textes sémantiquement proches. Celles-ci peuvent à leur tour servir à révéler des similarités multilingues de répartition. Entre deux langues, des formes différentes mais sémantiquement équivalentes ont des répartitions semblables entre deux documents traductions l'un de l'autre.

Entre deux documents traductions l'un de l'autre, l'écart entre les effectifs de ***N-grammes de caractères*** *sémantiquement équivalents est inférieur à l'écart entre les effectifs des* ***N-grammes de mots*** *graphiques sémantiquement équivalents. L'alignement des mots graphiques échoue d'autant plus que les langues comparées sont morphologiquement différentes.*

Ici, comme en témoigne la deuxième colonne du tableau 3, les écarts d'effectifs entre des mots alignés dans un échantillon sont déjà considérables. Or si l'on s'intéresse désormais aux répétitions de chaînes de caractères, on s'aperçoit qu'il existe dans chaque langue une sous-chaîne commune à l'ensemble des équivalents sémantiques de "transport".

Langue	Mots graphiques signifiant “transport” et (leur effectif)
fr	transports (3), transport (3)
es	transporte (5), transportes (1)
el	μεταφορών (3), μεταφορέας (1), μεταφορές (1), μεταφορέα (1)

Tableau 3: Liste des mots graphiques signifiant "transport" dans un échantillon de textes en fr, es et el, et leur effectif.

Langue	Chaînes de caractères répétées signifiant "transport"	Effectifs
fr	transport- (3+3)	6
es	transporte- (5+1)	6
el	μεταφορ- (3+1+1+1)	6

Tableau 4: Chaînes de caractères (d’au minimum 3 caractères) communes aux mots signifiant "transport" dans le même échantillon de textes en fr, es et el et leur effectif respectif.

Dans chacune des langues de cet échantillon, il existe une sous-chaîne commune aux mots lexicaux signifiant “transports”. Cette sous-chaîne commune apparaît donc comme un moyen de comparaison des langues susceptible de passer à l’échelle de plus gros volumes à moindre coût. Les écarts d’effectifs entre les mots partiellement ou intégralement équivalents se trouvent lissés.

3.3 Limites

Nous présentons dans cette dernière section, trois limites à la segmentation-alignement de N-grammes de caractères. Celles-ci trouvent une solution via la mise en place d’un traitement informatique spécifique et/ou adapté :

- Les mots lexicaux ou polylexicaux dont une ou plusieurs lettres changent, dans le cas de diphtongaison comme celle du verbe “contar” en espagnol, aux premières personnes du présent : “cuento”, “cuentas”, “cuenta” (i.e. skip-grams dans Mcnamee & Mayfield, 2004). Ici, sans autre traitement, l’alignement de N-grammes de caractères ne permettent pas de révéler davantage qu’un alignement basé sur des N-grammes de mots.
- Le risque de mettre en rapport des chaînes de caractères non liées au niveau du mot, entre “transport” et “transparence” par exemple.
- La surgénération de chaînes répétées “inintéressantes” dans le but de construction de ressources lexicales par une méthode d’alignement. Le fait de supposer que tout N-gramme de caractères d’une langue puisse être aligné avec n’importe quel N-gramme dans une autre langue nous permet de trouver beaucoup d’associations mais impose de fixer des règles pour parcourir ce très grand espace de recherche. Nous avons résolu ce problème en comparant les positions de N-grammes de fréquences similaires.

4 Perspectives et conclusion

Nous procédons actuellement à l’implémentation d’une méthode d’alignement multilingue de documents basée les N-grammes de caractères. Nous en préparons une évaluation par rapport

à une méthode en N-grammes de mots. Les premiers résultats témoignent que l'alignement de ces chaînes de caractères est non seulement réalisable, mais également ne requiert pas de ressource. Le découpage en N-grammes de caractères constitue une voie prometteuse en terme de comparaison et d'alignement des langues, et ce d'autant plus que les langues sont morphologiquement différentes.

Références

BISKRI I. & DELISLE S. (2001). Les n-grams de caractères pour l'extraction de connaissances dans des bases de données textuelles multilingues. In *Actes de la 8ème conférence annuelle sur le Traitement Automatique des Langues Naturelles*, Tours, France.

CROMIÈRES F. (2006). Sub-sentential alignment using substring Co-Occurrence counts. In *Sub-sentential Alignment Using Substring Co-Occurrence Counts*, Sydney, Australia.

DAMASHEK M. (1995). Gauging similarity with n-Grams: Language-Independent categorization of text. *Science*, **267**, 843–848.

DUNNING T. (1994). *Statistical Identification of Language*. Technical report MCCS 94-273, New Mexico State University.

GALE W. A. & CHURCH K. W. (1991). Identifying word correspondence in parallel texts. In *Proceedings of the workshop on Speech and Natural Language*, p. 152–157, Pacific Grove, California: Association for Computational Linguistics.

GALE W. A. & CHURCH K. W. (1993). A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, **19**(1), 75–102.

JARDINO M. (2006). Identification des auteurs de textes courts avec des n-grammes de caractères. In *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles*, Besançon, France.

LANGLAIS P. (1997). Alignement de corpus bilingues : intérêts, algorithmes et évaluations. *Bulletin de Linguistique Appliquée et Générale*, (numéro Hors Série), 245–254.

MAJUMDER P., MITRA M. & CHAUDHURI B. B. (2002). N-gram : a language independent approach to IR and NLP. In *Proceedings of the international Conference on Universal Knowledge and Language, 25-29 novembre*.

MCNAMEE P. & MAYFIELD J. (2004). Character N-Gram tokenization for european language text retrieval. *Information Retrieval*, **7**, 73–97. ACM ID: 961313.

NAKAMURA-DELLOYE Y. (2007). Méthodes d'alignement des propositions : un défi aux traductions croisées. In *Actes de la 14ème conférence annuelle sur le Traitement Automatique des Langues Naturelles, 12-15 juin*, Toulouse, France.

ZHOU Y., ZHONG C. & XU B. (2004). Bilingual chunk alignment in statistical machine translation. In *Proceedings of the 2004 IEEE international conference on systems, man & cybernetics, 10-13 october*, The Hague, Netherlands.