TALN 2009, Senlis, 24-26 juin 2009

Un chunker multilingue endogène

démonstration

Jacques Vergne

GREYC – Université de Caen, campus 2 - BP 5186, F-14032 CAEN cedex
Jacques.Vergne@info.unicaen.fr

Résumé

Le chunking consiste à segmenter un texte en chunks, segments sous-phrastiques qu'Abney a défini approximativement comme des groupes accentuels. Traditionnellement, le chunking utilise des ressources monolingues, le plus souvent exhaustives, quelquefois partielles : des mots grammaticaux et des ponctuations, qui marquent souvent des débuts et fins de chunk. Mais cette méthode, si l'on veut l'étendre à de nombreuses langues, nécessite de multiplier les ressources monolingues. Nous présentons une nouvelle méthode : le chunking endogène, qui n'utilise aucune ressource hormis le texte analysé lui-même. Cette méthode prolonge les travaux de Zipf : la minimisation de l'effort de communication conduit les locuteurs à raccourcir les mots fréquents. On peut alors caractériser un chunk comme étant la période des fonctions périodiques correllées longueur et effectif des mots sur l'axe syntagmatique. Cette méthode originale présente l'avantage de s'appliquer à un grand nombre de langues d'écriture alphabétique, avec le même algorithme, sans aucune ressource.

Abstract

Chunking is segmenting a text into chunks, sub-sentential segments, that Abney approximately defined as stress groups. Chunking usually uses monolingual resources, most often exhaustive, sometimes partial : function words and punctuations, which often mark beginnings and ends of chunks. But, to extend this method to other languages, monolingual resources have to be multiplied. We present a new method : endogenous chunking, which uses no other resource than the text to be parsed itself. The idea of this method comes from Zipf : to make the least communication effort, speakers are driven to shorten frequent words. A chunk then can be characterised as the period of the periodic correlated functions length and frequency of words on the syntagmatic axis. This original method takes its advantage to be applied to a great number of languages of alphabetic script, with the same algorithm, without any resource.

Mots-clés : chunking, multilingue, endogène, longueur des mots, effectif des mots
Keywords: chunking, multilingual, endogenous, word length, word frequency

1 Le concept de chunk selon Abney et selon Déjean

Le concept de chunk a été proposé par Steve Abney (1991). Il est fondé sur des propriétés de la parole : Abney l'a défini approximativement comme un groupe accentuel. La parole étant contrainte par l'appareil vocal, invariant selon les langues, le chunk est alors un concept générique portant sur les langues, un concept du langage. Hervé Déjean (1998) en a proposé un modèle de structure : des débuts et des fins de chunk (mots ou morphèmes) entourant un noyau (Déjean 1998, page 117); notre méthode utilise ce modèle.

2 Principes du chunking endogène

La méthode que nous proposons est fondée sur les propriétés des fonctions longueur et effectif des mots sur l'axe syntagmatique. Ces deux fonctions sont correllées : périodiques, synchrones, en opposition de phase, et leur période est le chunk. Sur une période, la fonction longueur est monotone croissante, et la fonction effectif est monotone décroissante. Ces concepts prolongent les travaux de Zipf : la minimisation de l'effort de communication conduit le locuteur à raccourcir les mots fréquents (Zipf 1949). La métrique de longueur utilisée par Zipf est le nombre de caractères de la graphie; la métrique de notre méthode est le nombre de syllabes, ou plus exactement le nombre de noyaux vocaliques, calculable sur la graphie; cette métrique est enracinée dans l'origine orale du concept de chunk. L'effectif des mots est mesuré dans le texte à segmenter.

L'algorithme exploite une hypothèse et deux propriétés contextuelles. Voici l'hypothèse : dans un même texte, à 1 graphie correspond 1 mot; donc tous les contextes d'une même graphie informent sur le même mot dans un même texte. La première propriété contextuelle est la suivante : le "virgulot" est défini comme segment délimité par 2 ponctuations; il a été proposé par Hervé Déjean ("entre-ponctuations") et renommé "virgulot" par Nadine Lucas. Propriété : le chunk est un constituant du virgulot; donc un mot attesté en début de virgulot est un début de chunk, et un mot attesté en fin de virgulot est une fin de chunk. La deuxième propriété exploite la propriété du chunk (à l'intérieur d'un virgulot) d'être la période des fonctions périodiques longueur et effectif des mots, et leur caractéristique de croissance ou décroissance monotones.

Du fait de l'hypothèse de biunivocité graphie - mot, tous les contextes d'une même graphie informent sur le même mot dans un même texte; donc début et fin de chunk sont des attributs des mots du dictionnaire du texte et non pas des attributs des mots occurrences du texte. Pour chaque propriété contextuelle, l'algorithme consiste en un parcours du texte pour relever les contextes attestés, et renseigner les mots du dictionnaire, puis sort les résultats en informant les mots occurrences à partir des mots du dictionnaire. L'ordre d'application des deux propriétés est indifférent car elles sont indépendantes.

3 Quelques fragments segmentés en chunks

Le corpus de validation de la méthode est composé de communiqués de presse (environ 1000 mots chacun) du site de l'Union Européenne (http://europa.eu/rapid/), chaque communiqué étant rédigé en 1 à 22 langues. Les fragments suivants sont extraits du communiqué IP/05/1018 de 2005 :

[Die Laichgründe] [der Aale] befinden] [sich [im Sargassosee] [im mittleren Westatlantik] .

[Eels spawn] [in [the Sargasso Sea [in [the western central Atlantic] Ocean] .

[Las [anguilas] desovan] [en [el Mar [de [los Sargazos] , [en [las aguas] centro-occidentales] [del Océano Atlántico] .

[La zone [de frai] [de l'anguille] [se situe [en mer] [des Sargasses] , [dans [la partie centre-ouest] [de l'océan Atlantique] .

[Le anguille] [si riproducono] [nel mar [dei Sargassi] , [nell'Atlantico centro-occidentale] .

4 Conclusion

En caractérisant le chunk de manière purement numérique, comme période des fonctions longueur et effectif des mots sur l'axe syntagmatique, cette méthode originale consiste en des calculs sur le texte à segmenter; elle a l'avantage de s'appliquer à un grand nombre de langues, avec le même algorithme, sans aucune ressource monolingue : des langues d'écriture alphabétique, avec une pratique du mot écrit qui sépare les mots grammaticaux des mots lexicaux, et compatibles avec un modèle de structure de chunk où les mots grammaticaux sont généralement antéposés aux mots lexicaux; la méthode est prometteuse pour les 22 langues de la Communauté Européenne, hormis le finnois[1].

Étant indépendante des spécificités de chaque langue, cette méthode n'est pas "multilangue", ni "multi-monolangue", mais comme elle exploite des propriétés génériques des langues, c'est-à-dire des propriétés du langage, en tant qu'abstraction des langues, on pourrait simplement l'appeler méthode "linguistique".

Références

STEVEN ABNEY (1991). Parsing By Chunks, in *Principle-Based Parsing*, 257-278, Kluwer Academic Publishers.

HERVE DEJEAN (1998). Concepts et algorithmes pour la découverte des structures formelles des langues. *Thèse de doctorat de l'université de Caen.*

JACQUES VERGNE (2001). Analyse syntaxique automatique de langues : du combinatoire au calculatoire (communication invitée), Actes de *TALN 2001*, 15-29.

GEORGE K. ZIPF (1949) *Human Behavior and the Principle of Least-Effort*. Addison-Wesley.

[1] Voir des résultats sur : http://www.info.unicaen.fr/~jvergne/chunking_multilingue_endogene/