Un système de segmentation du chinois basé sur des triplets

Yiping LI

LIC2M CEA Fontenay-aux-Roses – Université de Marne la Vallée
18 rue du Panorama BP 6, 92265 Fontenay aux Roses Cedex
li@zoe.cea.fr
Date de la thèse future : 2005

Mots-clefs – Keywords

Tokenisation, segmentation du chinois, ngrammes, approche statistique, maximum matching

Chinese segmentation, ngrams, statistical approach, maximum matching

Résumé – Abstract

Un des problèmes rencontrés lors de l'analyse de textes en chinois est qu'il n'existe pas de séparateur entre les mots dans cette langue. Le mot étant une unité linguistique fondamentale en traitement automatique de la langue, il est nécessaire d'identifier les mots dans un texte chinois afin que des analyses de plus haut niveau puissent être réalisées. Le but de cet article est de présenter un système d'identification des mots basé sur un algorithme utilisant des triplets de catégories grammaticales et des fréquences de mots. Ce système comprend deux dictionnaires : l'un dédié aux mots et à leurs fréquences, l'autre aux triplets des catégories correspondantes. Les tests qui ont été effectués révèlent que 98,5% des phrases sont découpées correctement. Certaines erreurs sont dues à la taille limitée du dictionnaire utilisé. Une réflexion sur la création de nouvelles catégories et des études proposant des règles grammaticales sont en cours de réalisation afin d'augmenter la performance du système.

One of the problems encountered by Chinese texts analysis is that there is no separator between the words in this language. As a fundamental linguistic unit in automatic treatment of the language, word is necessary to be identified in a Chinese text so that higher-level analyses can be carried out. The goal of this work is to develop a system, identifying words, based on an algorithm of triplets of grammatical categories and words frequencies. This system contains two dictionaries. One is dedicated to the words and their frequencies, the other, to the triplets of the corresponding categories. The tests carried out reveal that this system works very well, 98.5% of the sentences are segmented correctly. Thus, a reflection about the creation of new categories and the study proposing the grammatical rules are carrying out to improve the performance of the triplets.

1 Introduction

Le chinois s'écrit au moyen d'une écriture de type idéographique. Il y a ainsi des dizaines de milliers de caractères en chinois. Afin de faciliter la fluidité et la vitesse d'écriture, les caractères sont accolés, c'est-à-dire que l'écriture chinoise ne possède pas de délimiteur entre les mots (Hung, Tzeng, 1981). Ce phénomène, qui existe aussi pour d'autres langues asiatiques comme le japonais et le coréen, pose au Traitement Automatique des Langues (TAL) des problèmes spécifiques par rapport aux langues occidentales et celle de Moyen-Orient. Les mots étant des unités linguistiques fondamentales, il est nécessaire de les identifier dans un texte afin de pouvoir effectuer une analyse de plus haut niveau, comme par exemple l'analyse syntaxique ou la désambiguïsation sémantique.

Plusieurs types de phénomènes linguistiques sont à l'origine des ambiguïtés concernant le découpage en mots du chinois. Premièrement, presque tous les caractères peuvent constituer un mot en soi. Ils peuvent également se joindre à d'autres caractères pour former des mots. Deuxièmement, le chinois moderne utilise essentiellement des mots composés. Il est difficile de déterminer si un mot composé rare est un mot ou une expression. Troisièmement, les mêmes caractères sont généralement employés pour la construction des noms communs et des noms propres. La distinction entre les deux se révèle dès lors difficile. Quatrièmement, quelques structures morphologiques spéciales, comme la duplication et les constructions A_non_A, AA, A_un_A, et ABAB (Xia, 2000) doivent également être prises en compte. Par exemple au lieu de dire "regarder", le chinois utilise très souvent l'expression "regarder un regarder", etc. Deux types d'ambiguïtés sont très fréquentes : ambiguïté de croisement intérieur (si une chaîne de caractères ABC peut être découpée en A/BC ou AB/C, ABC est une chaîne ambiguë de croisement intérieur) et ambiguïté de combinaison (si PQ est un mot et P et Q peuvent aussi être des mots indépendants. PQ est une chaîne ambiguë de combinaison).

2 Les approches existantes

Dans le domaine de la segmentation du chinois, il existe trois types principaux d'algorithmes (Emerson, 2000) : les approches statistiques, les approches fondées sur les dictionnaires et les approches mixtes. Les approches statistiques sont basées sur la probabilité que deux ou plusieurs caractères apparaissent ensemble. Elles s'appuient généralement sur des modèles à base de HMM (modèles de Markov cachés). Ces approches présentent l'avantage d'être peu sensibles à l'effet des mots inconnus et des translittérations phonétiques. Toutefois, elles dépendent d'un modèle linguistique qui est lui-même contraint par la qualité et le volume des corpus d'apprentissage.

Les approches basées sur les dictionnaires peuvent être divisées en approches strictement basées sur les dictionnaires et approches combinant dictionnaire et connaissances linguistiques. L'idée générale consiste rassembler des caractères en mot en fonction de ceux présents dans un dictionnaire. Les ambiguïtés sont levées en utilisant une heuristique générale sur la façon de guider l'appariement : heuristique concernant la taille des mots, comme pour le *simple maximum matching* (Tsai, 1996) et le *complexe maximum matching* (Chen, Liu, 1992) ; ou heuristique relative au sens dans lequel l'appariement se fait, comme pour le *forward maximum matching* mis en œuvre dans Chinese Segment (Peterson, 2000) et le *backward maximum matching*. Les deux dernières sont parfois utilisées conjointement afin de résoudre des ambiguïtés de croisement intérieur. Comme aucun dictionnaire ne peut être

exhaustif, une désambiguïsation à l'aide de connaissances linguistiques peut être utilisée en complément. Ces connaissances, par exemple de nature grammaticale, permettent d'effectuer certains regroupements préférentiels à partir d'une segmentation de base considérant chaque caractère comme un mot. C'est le cas dans (Hockenmaier, Brew, 1998).

Les méthodes mixes combinent les deux types de méthodes présentés ci-dessus. Le *Chinese Morphological Analyzer* (Emerson, 2000) en est un exemple.

Au vu des avantages et des inconvénients de chaque méthode, nous avons choisi le troisième type de méthodes, en combinant un dictionnaire, des connaissances linguistiques et des statistiques. L'algorithme de *simple maximum matching* implique toujours des ambiguïtés de croisement intérieur et des traitements supplémentaires sont nécessaires pour améliorer la qualité du découpage. Pour la désambiguïsation, nous avons aussi besoin de connaissances grammaticales. Si des ambiguïtés persistent encore après ces traitements, l'utilisation de fréquences de mots constitue un très bon moyen pour trouver la meilleure solution.

3 Les travaux réalisés

Afin d'obtenir le dictionnaire et des informations grammaticales, nous avons utilisé le corpus "Chinese Treebank" de l'Université de Pennsylvanie. Ce corpus est constitué de 325 dépêches en chinois. Dans chacune de ces dépêches, les mots sont d'emblée balisés avec des étiquettes renfermant des informations sur les parties du discours, la syntaxe, les fonctions, et les catégories vides. Afin d'évaluer la performance des dictionnaires et la méthode de désambiguïsation à l'aide des triplets de catégories successives, nous avons utilisé 295 fichiers pour faire un apprentissage des règles de la segmentation ; les 30 fichiers restants ont servi à l'évaluation de cet apprentissage.

L'étude de la problématique de la segmentation du chinois et des algorithmes existants a révélé la nécessité de disposer d'un dictionnaire de bonne qualité. Nous avons créé deux dictionnaires en nous servant des informations disponibles du corpus : un dictionnaire de mots avec leurs catégories possibles et un dictionnaire de triplets de catégories successives en fonction du contexte. La fréquence de chaque mot figure également dans le dictionnaire de mots. Elle peut servir à lever certaines ambiguïtés. Le dictionnaire de triplets offre un contexte pour préciser et faciliter le choix.

Avec ces deux dictionnaires, nous pouvons entamer les principales étapes de la segmentation. Première étape : en travaillant au niveau de la ligne et en effectuant une comparaison par rapport aux entrées du dictionnaire, nous avons sélectionné tous les mots du texte à segmenter. Ceci nous permet d'obtenir un tableau de mots présents dans le dictionnaire avec leurs positions de début et de fin dans le texte. Il convient de remarquer que dans les méthodes classiques de segmentation du chinois, le traitement est effectué ligne par ligne. Comme il est possible de rencontrer un saut de ligne à l'intérieur d'un mot, nous traitons intégralement la partie entre deux ponctuations comme l'unité de traitement, sans quoi il n'est plus possible de trouver la bonne segmentation pour des mots séparés par un saut de ligne.

Deuxième étape : regrouper les mots en phrases. Le tableau de tous les mots possibles est semblable à une boîte de perles. Il faut enfiler ces perles sur une chaîne selon leurs positions. Après l'analyse de toutes les perles, une série de segmentations possibles est obtenue. Afin de faciliter les traitements ultérieurs, un numéro unique est donné à chaque segmentation.

Un dictionnaire ne peut pas contenir tous les mots. En effet, lorsque des mots inconnus sont présents dans une phrase, il est impossible d'enfiler les perles sur la chaîne, car il y a des "trous" entre les mots. La segmentation ne doit pas être interrompue pour cette raison. Tout comme nous le faisons pour le traitement des mots dans le dictionnaire, nous plaçons aussi les mots inconnus dans la boîte de perles. Pendant le processus de sélection des mots (étape deux), si aucun mot commençant par ce caractère n'est trouvé dans le dictionnaire, nous ajoutons ce caractère dans le tableau comme un mot appartenant à la catégorie "inconnu". Des caractères inconnus successifs peuvent constituer un mot intégral, ils sont considérés comme un mot à part entière avec la catégorie "inconnue". Cette technique peut aussi servir pour l'apprentissage de nouveaux mots.

Troisième étape : comparer des triplets. Nous effectuons une comparaison entre les triplets de catégories correspondant aux mots séparés en fonction de chaque possibilité de segmentation et les entrées du dictionnaire de triplets. En ce qui concerne le traitement des mots portant la catégorie "inconnu" qui ne figure pas dans le dictionnaire de triplets, nous avons fait le choix de leur affecter la catégorie "Nom Propre" puisqu'ils sont très souvent des noms propres. Un mot peut être associé à plusieurs catégories grammaticales. Lors de l'extraction des triplets possibles, nous traitons plusieurs catégories d'un même mot séparément en parcourant toutes les combinaisons possibles des triplets. Pour chaque possibilité de segmentation, si une combinaison de triplets correspond aux entrées du dictionnaire de triplets, la segmentation est pertinente. Si aucune combinaison de triplets n'est cohérente avec les entrées, cette segmentation est filtrée par l'analyse au moyen des trigrammes.

Les mots ayant plusieurs catégories augmentent considérablement le nombre de combinaisons des triplets et donc ralentissent la segmentation de manière importante. En moyenne, quand le nombre de caractères arrive à 35, le nombre de segmentations augmente exponentiellement jusqu'à 6000, et le nombre de combinaisons de triplets atteint presque 90 000 000. L'algorithme n'est plus utilisable, le temps de traitement devenant alors excessivement long. Il nous a donc failli effectuer un filtrage avant d'utiliser les triplets en choisissant la segmentation minimisant le nombre de mots découpés. Un test sur le corpus a révélé que 99 pour cent des résultats de cette stratégie sont corrects.

Quatrième étape: calculer et comparer la somme des fréquences de segmentation. L'utilisation de triplets ne produit pas nécessairement un découpage unique. A l'aide du dictionnaire des fréquences, nous avons sélectionné la segmentation ayant la fréquence la plus élevée.

4 Evaluation

Nous avons testé notre algorithme de segmentation et en avons analysé la performance. Comme nous l'avons signalé précédemment, nous avons conservé 30 fichiers pour l'évaluation. Nous avons extrait le texte brut comme le texte à découper et un texte où les mots sont balisés comme le résultat standard du découpage en supprimant des informations grammaticales supplémentaires.

Notre programme de segmentation a été lancé sur le texte brut. La F-mesure est utilisée pour évaluer le résultat (Peng, Huang, Schuurmans, Cercone, 2002). Le rappel est le nombre de mots découpés correctement divisé par le nombre total de mots dans notre résultat. La précision est le nombre de mots découpés correctement divisé par le nombre total de mots dans le résultat standard. Notre programme d'évaluation révèle que 9327 mots sont découpés

correctement parmi 9489 mots dans le texte découpé par notre algorithme. Il y a 9599 mots dans le résultat standard. Les valeurs du rappel, de la précision et de la F-mesure de notre algorithme sont respectivement de 97,2%, 98,9% et 97,7%.

Par rapport aux performances d'autres algorithmes existants, notre algorithme donne des résultats assez satisfaisants. Le rappel, la précision et la F-mesure de l'algorithme de MMSEG de Chih-Hao Tsai sur 1013 mots sont de 95,4%, 95,5% et 95.4% pour le *simple maximum matching*, et 98,1%, 98,4% et 98,3% pour le *complex maximum matching*. Le résultat de l'algorithme de Palmer (Palmer, 1997) sur le corpus XinHua a une F-mesure de 89,6%. Les F-mesures de l'algorithme *Error Driven Learning* (Hockenmaier, Brew, 1998) sont de 87,9%, 87,4% et 87,1%.

5 Discussion

Afin de mieux évaluer les performances de notre algorithme, nous l'avons comparé à deux systèmes actuellement disponibles (Peterson, Wu). Un découpage au moyen de ces deux algorithmes et du nôtre est réalisé sur le même corpus[1].

Texte à découper : 这是日本金融市场当前对中国银行的最高债券评级
Segmentation correct : 这•是•日本•金融•市场•当前•对•中国•银行•的•最高•债券•评级
Traduction : C'est l'évaluation des bons du Trésor la plus élevée du marché financier japonais pour la Banque de Chine.
Résultat d'Erik Peterson : 这•**是日**•**本金**•**融**•市场•当前•对•中国•银行•的•最高•债券•**评级**
Résultat de Zhibuiao Wu : 这•**是**•**日本**•**金融**•市场•当前•对•中国•银行•的•最高•债券•**评**•**级**
Notre résultat : 这•**是**•**日本**•**金融**•市场•当前•对•中国•银行•的•最高•债券•**评级**
La chaîne "是日本金融" est ambiguë. Sa segmentation peut être : 是•日本•金融, 是•日•本金•融, 是日•本金•融 ou bien 是•日•本•金融. La segmentation d'Erik est erronée dès le début. Il délimite directement 是日, le reste ne peut donc pas être correctement découpé. Par contre notre algorithme a trouvé le bon découpage, grâce au traitement intégral de la partie entre deux ponctuations, au lieu de sortir le mot rencontré le plus long immédiatement en comparant avec le dictionnaire de mots. Toutes les possibilités de découpage sont prises en compte. Il permet d'enlever un nombre important d'ambiguïtés.

Pendant l'évaluation, nous avons remarqué qu'un pourcentage important de mauvaises segmentations provient d'une faiblesse du dictionnaire. Nous avons testé de nouveau l'algorithme en rajoutant les mots manquants, et le résultat s'est nettement amélioré. Le rappel, la précision et la F-mesure sont de 97,17%, 98,85% et 97,73%.

6 Conclusion et perspectives

Dans notre algorithme de segmentation, à partir de toutes les segmentations possibles permises par le dictionnaire, basé sur une désambiguïsation par triplets de catégories et sur la fréquence des mots, nous obtenons un très bon résultat. La notion de l'unité de traitement a permis de prendre en compte des mots coupés par des sauts de ligne. Le traitement appliqué

[1] Le rappel, la précision, et la F-mesure de notre algorithme sur ce corpus sont de 83,8%, 90,0% et 86,8% ; ceux de l'algorithme d'Erik sont de 70,6%, 68,7%, 69,6% ; ceux de l'algorithme de Zhibiao sont de 68,3%, 62,6% et 65,3%. Nous voyons très clairement que notre algorithme fonctionne mieux.

aux noms inconnus a permis une segmentation complète de tous les textes. Néanmoins, la limite de volume des dictionnaires a affecté considérablement les résultats de la segmentation. De plus, les règles apprises à partir des sources de Chinese Treebank ne sont pas complètes.

Vu ces problèmes, nous sommes en train d'améliorer les deux dictionnaires. Nous pouvons intégrer des mots de dictionnaires existants dans notre dictionnaire de mots. Certains mots seront ajoutés manuellement, par exemple des unités de mesure, les chiffres, etc. Pour le dictionnaire des triplets, nous avons besoin de perfectionner notre connaissance de la grammaire chinoise. Avec ces connaissances approfondies, il sera possible d'écrire des règles de triplets. Par exemple, quelles catégories ne peuvent pas être consécutives, lesquelles peuvent l'être, lesquelles sont forcément placées les unes après les autres, etc. Cette analyse grammaticale des phrases permet aussi de mieux en comprendre le sens. Elle constitue également une excellente base afin de procéder ultérieurement à une analyse syntaxique.

Références

Tom Emerson (2000), Segmentation of Chinese Text, *Multilingual Computing & Technology*, Vol. 12 Issue 2, pp38.

Chih-Hao Tsai (1996), A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm.

Chen K. J., & Liu S. H. (1992), Word identification for Mandarin Chinese sentences. Proceedings, *Fifteenth International Conference on Computational Linguistics*, Nantes: COLING-92.

Hung D. L, & Tzeng O. J. L (1981), Orthographic variations and visual information processing, *Psychological Bulletin*, Vol.90, pp377-414.

David Palmer (1997), A Trainable Rule-Based Algorithm for Word Segmentation Proceedings, Acte de *the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid.

Fei Xia (2000), *The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0)*, Philadelphia, pp14-15.

Fuchun Peng, Xiangji Huang, Dale Schuurmans, Nick Cercone (2002), Investigating the relationship between word segmentation performance and retrieval performance in Chinese IR, Acte de *COLING 2002 the 19th International Conference on Computational Linguistics*, 793-799.

Julia Hockenmaier, Chris Brew, (1998), Error-driven segmentation of Chinese, Acte de *12th Pacific Conference on Language and Information*, 218-229.

Erik Peterson, (2000), http://www.mandarintools.com

Zhibiao Wu, (1999), http://www.ldc.upenn.edu/Projects/Chinese/segmenter/mansegment.perl http://www.ldc.upenn.edu/Projects/Chinese/segmenter/Mandarin.fre