Analyse automatique du chinois utilisant des ressources linguistiques

Zhen CAI[1]

(1) Laboratoire ELLIADD, Université de Franche-Comté,30 rue Mégevand, 25000 Besançon

zhencai1122@hotmail.com

RESUME

Au cours de ces dernières années, les recherches sur l'analyse automatique de la langue chinoise se multiplient et différentes approches sont proposées afin d'améliorer ces analyses. Nous proposons dans cet article de construire un analyseur automatique utilisant des ressources linguistiques en taille réelle. Les problèmes à résoudre concernent le codage et le lexique : en particulier, nous avons construit un dictionnaire électronique avec la plateforme linguistique NooJ. Nous citons quelques exemples d'applications, parmi lesquelles un outil de segmentation automatique du chinois, essentiel pour construire toutes les applications de TAL pour le chinois.

ABSTRACT

In recent years, research on automatic Chinese language processing has exploded and different approaches are proposed to improve parsers. In this article, we present an automatic parser that uses large-coverage linguistic resources. The two main problems to solve are managing the encoding system and the Chinese lexicon. We have constructed a large-coverage electronic dictionary, using the NooJ platform. We present various applications, among them an automatic Chinese tokenizer, crucial for many NLP applications.

MOTS-CLES : chinois, codage, segmentation, dictionnaire

KEYWORDS : Chinese, encoding, Tokenization, Dictionary

---

1 Le traitement automatique du chinois

Depuis les années 90, les chercheurs en TAL ont commencé à construire des logiciels capables d'analyser automatiquement des textes chinois. Plusieurs problèmes doivent être abordés :

— Les problèmes de codage des textes

— La segmentation du chinois, étape cruciale pour toute application de TAL.

— Le traitement des morphèmes et mots composés

— Le traitement de certains objets linguistiques spécifiques aux langues asiatiques, tels que les classificateurs et les subordinateurs.

2 Problème de codage des textes chinois

Pour que les caractères chinois puissent être traités et affichés correctement, il faut normaliser le codage de caractères. Nous utilisons la plateforme NooJ qui traite les textes en Unicode (UTF8).

3 Unité linguistique du chinois

La première étape de toute analyse d'un texte chinois est de reconnaître ses unités lexicales. Pour cela, nous avons construit un dictionnaire électronique qui contient environ 63 000 entrées. Nous présentons les critères que nous avons utilisés pour définir ce que sont ces unités lexicales, un problème crucial en particulier lorsque nous avons affaire à des morphèmes (mot ou suffixe ?) et à des mots composés (quand doit-on traiter en bloc une séquence de caractères). Noter que le problème de la segmentation est différent de celui pour les autres langues ; par exemple, en français, les mots composés se différencient des mots simples grâce à la présence de séparateurs comme l'espace ou le trait d'union, ce qui n'est pas le cas pour le chinois.

4 La segmentation du chinois

Nous proposons de comparer les outils de la segmentation du chinois existant (utilisant des techniques probabilistes ou statistiques) avec notre outil de segmentation du chinois (utilisant des dictionnaires et des grammaires). Nous commenterons les résultats et erreurs produits par notre outil.

Les deux exemples en utilisant notre dictionnaire électronique :

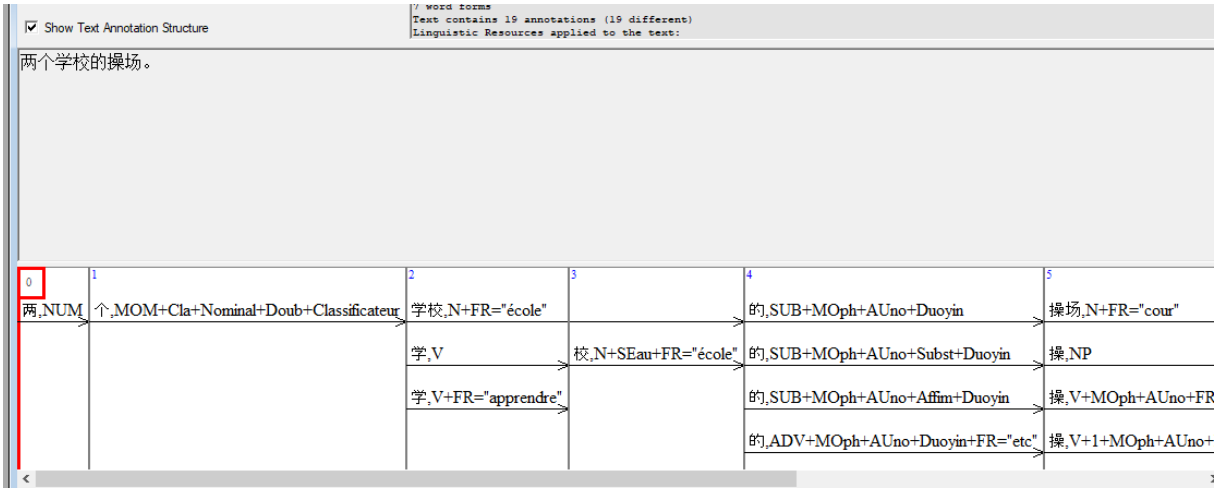

Figure 1

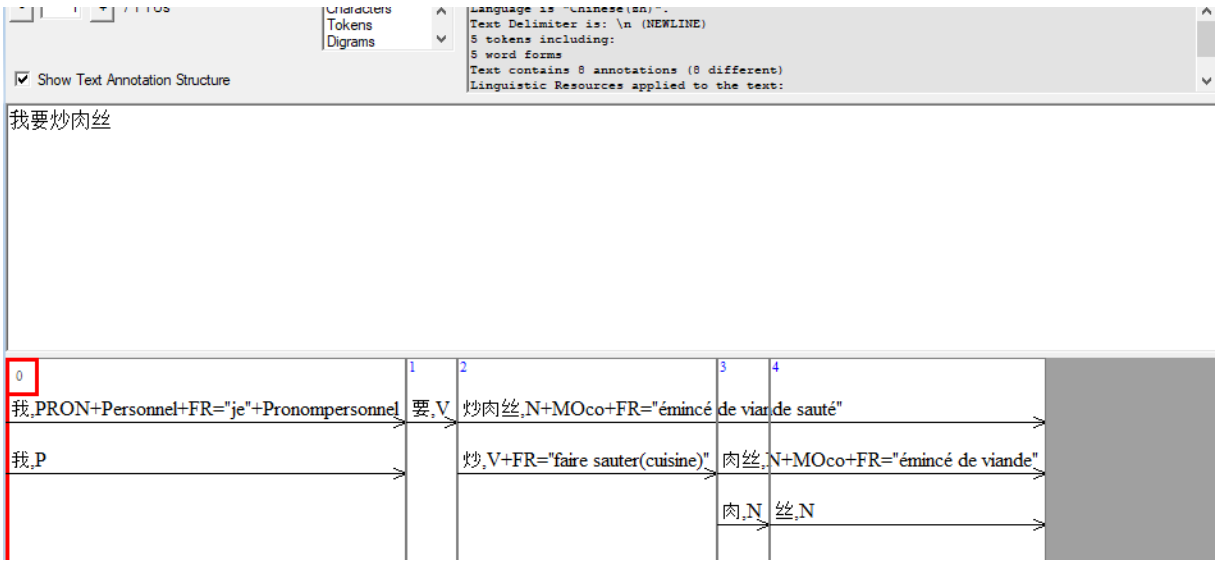

Figure 2

Nous pouvons enlever des résultats non-pertinents après quelques traitements :

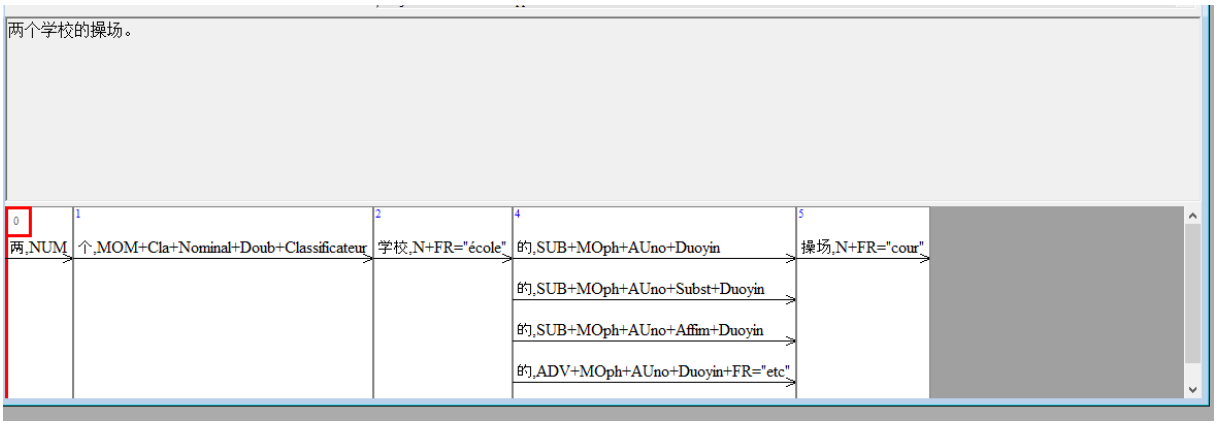

Figure 3

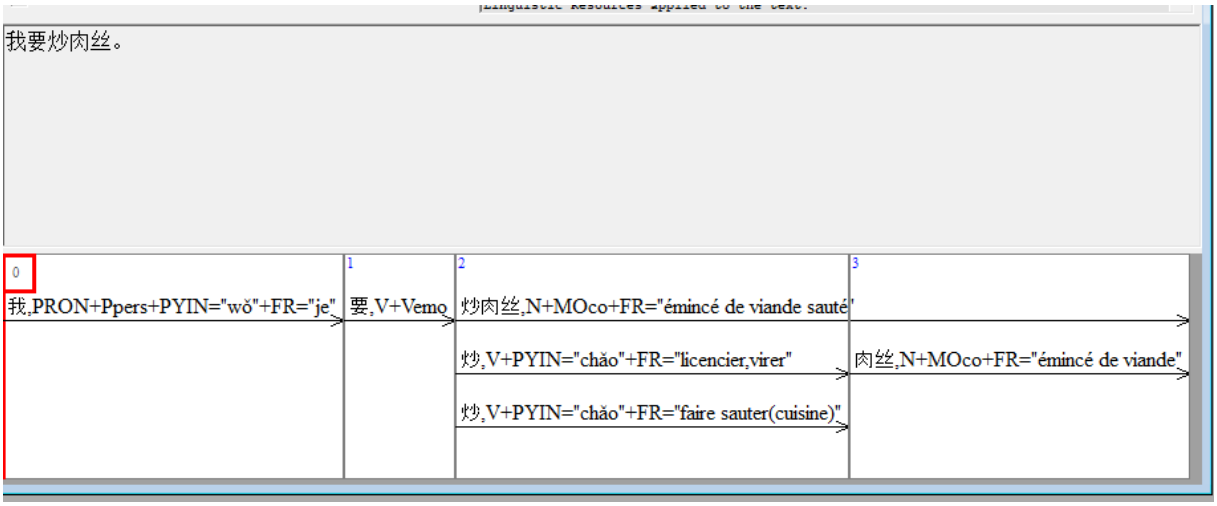

Figure 4

5 Conclusion

Après une présentation succinte de l'état de l'art dans le domaine de la segmentation automatique de textes chinois, nous présenterons un outil de segmentation utilisant des ressources linguistiques sous la forme d'un dictionnaire électronique de 63 000 entrées et de grammaires locales, morphologiques et syntaxiques. Nous commenterons enfin les résultats produits par notre outil.

Références

Feng, Z.冯志伟 (2004). 机器翻译研究 Jīqì Fānyì Yánjiū 'Etudes sur la traduction automatique' China Translation & Publishing Corporation.

Guo,R. 郭锐(2002). 现代汉语词类研究 Xiàndài Hànyǔ Cílèi Yánjiū 'Etudes sur les catégories en chinois moderne'. Commercial Press商务印书馆.

Paris, M. (1981). Problème de syntaxe et de sémantique en linguistique chinoise. INSTITUT DES HAUTES ETUDES CHINOISES.

Paris, M. (1996). La subordination en chinois standard : quelques contraintes d'agencement. Livre « Dépendance et intégration syntaxique » Edité par Claude Muller P233-P240

Paris, M. (2007). Un aperçu de la réduplication nominale et verbale en mandarin. Livre « La réduplication » par Alexis Michaud et Aliyah Morgenstern. Edition Ophrys. P63-P76.

Paris, M. (2013). Linguistique chinoise et linguistique générale. L'Harmattan.

Silberztein, M. (2015). La formalisation des langues l'approche de NooJ. Editions ISTE.

Yang-Drocourt, Z. (2007). Parlons chinois. L'Harmattan.