Un segmenteur-étiqueteur et un chunker pour le français

Isabelle Tellier[1,2], Yoann Dupont[1,2], Arnaud Courmet[2]
(1) LaTTiCe, université Paris 3 - Sorbonne Nouvelle
(2) LIFO, université d'Orléans
isabelle.tellier@univ-paris3.fr, yoann.dupont@etu.univ-orleans.fr,
arnaud.coumet@gmail.com

RÉSUMÉ

Nous proposons une démonstration de deux programmes : un segmenteur-étiqueteur POS pour le français et un programme de parenthésage en "chunks" de textes préalablement traités par le programme précédent. Tous deux ont été appris à partir du French Tree Bank.

ABSTRACT

A Segmenter-POS Labeller and a Chunker for French

We propose a demo of two softwares : a Segmenter-POS Labeller for French and a Chunker for texts treated by the first program. Both have been learned from the French Tree Bank.

MOTS-CLÉS : étiquetage POS, chunking, apprentissage automatique, French Tree Bank, CRF.

KEYWORDS: POS tagging, chunking, Machine Learning, French Tree Bank, CRF.

1 Introduction

Nous proposons de faire une démonstration de plusieurs programmes appris automatiquement à partir du French Treebank(Abeillé *et al.*, 2003) :
– un segmenteur combiné avec un étiqueteur morphosyntaxique (Constant *et al.*, 2011)
– un "chunker" ou analyseur syntaxique superficiel (Abney, 1991; Sha et Pereira, 2003; Tellier *et al.*, 2012)

Les deux programmes sont utilisables en séquence, le chunker s'appuyant pour fonctionner sur le résultat fourni par l'étiqueteur. Ils ont tous les deux été appris automatiquement par un CRF (Conditional Random Fields) (Lafferty *et al.*, 2001; Tellier et Tommasi, 2011). Ils sont libres et gratuits, disponibles en téléchargement mais ont surtout été testés sous Debian, Ubuntu et Mac (résultats non garantis sous Windows). Il faut pour les utiliser disposer des logiciels suivants :
– un interpréteur Python : http ://www.python.org/download/
– Wapiti, une implémentation des CRF linéaires : http ://wapiti.limsi.fr/
– Bazaar, un gestionnaire de versions donnant accès au serveur où ils sont stockés : http ://wiki.bazaar.canonical.com/

Nous décrivons brièvement ci-dessous les différentes options disponibles pour ces programmes et les résultats de leur évaluation.

2 Les programmes

Pour télécharger le segmenteur-étiqueteur, il faut saisir l'instruction suivante :
bzr branch lp : yoann-dupont/crftagger/stand-alone-tagger
Le processus ayant permis de l'apprendre est décrit dans (Constant *et al.*, 2011). Son originalité est de permettre plusieurs segmentations possibles :
– soit une segmentation "maximale" réalisée à l'aide de règles écrites manuellement
– soit une segmentation qui cherche à identifier les unités multimots du texte, en tenant compte de celles présentes dans le French Treebank ainsi que dans le Lefff (Sagot, 2010).

L'étiqueteur s'appuie sur la segmentation choisie et distingue 29 étiquettes : en validation croisée, il atteint une exactitude de 97,3% sans tenir compte des unités multimots, 95,2% avec elles.
Pour télécharger le "chunker", il faut saisir l'instruction suivante :
bzr branch lp : yoann-dupont/crftagger/chunker_models
Le processus ayant permis de l'apprendre est décrit dans (Tellier *et al.*, 2012). Il s'appuie sur les étiquettes fournies par le programme précédent et fonctionne suivant deux variantes possibles :
– une variante qui se concentre sur la seule identification de tous les "groupes nominaux simples" (i.e. non récursifs) NP. En supposant un étiquetage morphosyntaxique parfait et en réquérant la stricte égalité des frontières, ils sont identifiés en validation croisée avec une précision de 97,49%, un rappel de 97,40%, et donc une F-mesure de 97,45.
– une variante qui cherche à réaliser un parenthésage complet des phrases, en distinguant 6 types de chunks possibles. En supposant un étiquetage parfait, la "micro-average" (moyenne des F-mesures de chaque groupe pondérées par leur effectif) vaut 79,73, tandis que la "macro-average" (moyenne des F-mesure sans pondération) vaut 73,37.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer, Dordrecht.

Abney, S. (1991). Parsing by chunks. *In* Berwick, R., Abney, R. et Tenny, C., éditeurs : *Principle-based Parsing*. Kluwer Academic Publisher.

Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A. et Billot, S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. *In Actes de TALN'11*.

Lafferty, J., McCallum, A. et Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proceedings of ICML 2001*, pages 282–289.

Sagot, B. (2010). The lefff, a freely available, accurate and large-coverage lexicon for french. *In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

Sha, F. et Pereira, F. (2003). Shallow parsing with conditional random fields. *In Proceedings of HLT-NAACL 2003*, pages 213 – 220.

Tellier, I., Duchier, D., Eshkol, I., Courmet, A. et Martinet, M. (2012). Apprentissage automatique d'un chunker pour le français. *In Actes de TALN'12, papier court (poster)*.

Tellier, I. et Tommasi, M. (2011). Champs Markoviens Conditionnels pour l'extraction d'information. *In* Eric Gaussier et François Yvon, éditeurs : *Modèles probabilistes pour l'accès à l'information textuelle*. Hermès.