Un reconnaisseur d'entités nommées du Français

Yoann DUPONT[1] Isabelle TELLIER[1]
(1) Université Paris 3 Sorbonne Nouvelle, 13, rue de Santeuil - 75231 Paris Cedex 05
yoann.dupont@etud.sorbonne-nouvelle.fr, isabelle.tellier@univ-paris3.fr

Résumé. Nous proposons une démonstration d'un reconnaisseur d'entités nommées du Français appris automatiquement sur le French TreeBank annoté en entités nommées.

Abstract. We propose to demonstrate a french named entity recognizer trained on the French TreeBank enriched with named entity annotations.

Mots-clés : REN, POS, apprentissage automatique, French Treebank, extraction d'information, CRF.

Keywords: NER, POS, machine learning, French Treebank, information extraction, CRF.

1 Introduction

La reconnaissance d'entités nommées (REN) est une tâche importante du TAL, pour laquelle il existe de nombreuses tâches telles que le MUC-7 (Appelt *et al.*, 1995), CoNLL (Tjong Kim Sang & Erik F. & De Meulder, 2003) ou encore NLPBA (KIM *et al.*, 2004) pour les entités biomédicales. Cependant, peu de corpus en Français sont disponibles pour cette tâche, rendant difficile la création d'outils appris automatiquement tels que (Favre & Béchet, 2005). Un autre reconnaisseur d'entités nommées du Français étant CasEN (Maurel *et al.*, 2011), qui a recours à des cascades de transducteurs. Parmi les corpus français les plus connus sont le corpus de la campagne d'évaluation ESTER (Gravier *et al.*, 2004), celui de la campagne d'évaluation ETAPE (Gravier *et al.*, 2012) ainsi que le French Treebank (Abeillé *et al.*, 2003) annoté en entités nommées (Sagot *et al.*, 2012). Ce dernier nous a permis d'entraîner un CRF (Lafferty *et al.*, 2001; Tellier & Tommasi, 2011) pour obtenir un reconnaisseur d'entités nommées du Français qui est l'objet de cette démonstration.

SEM (pour segmenteur étiqueteur markovien) est gratuit sous licence GNU 3 et librement disponible [1], il ne fonctionne que sur les systèmes Linux et Mac, il ne fonctionne pas sous Windows à l'heure actuelle. Pour fonctionner il nécessite :

– Un interpréteur python 2.5 ou supérieur. (`http://www.python.org/download/`)
– Wapiti version 1.5.0, une implémentation des CRF linéaires (`http://wapiti.limsi.fr/`)
– Pour bénéficier du versionnement du programme, le gestionneur de versions Bazaar est requis (`http://wiki.bazaar.canonical.com/`).

2 Fonctionnement du programme

Pour télécharger la branche du programme, il faut entrer la commande suivante dans un terminal [2] :

```
bzr branch lp:~yoann-dupont/crftagger/stand-alone-tagger
```

Le programme peut enchaîner divers traitements les uns à la suite des autres, comme effectuer un étiquetage POS, intégrer des dictionnaires, puis effectuer une passe de reconnaissance d'entités nommées. Une part importante dans l'apprentissage automatique supervisé étant l'ajout d'informations, en particulier pour la REN où les traits morphologiques et contextuels

1. la page web du programme : `http://www.lattice.cnrs.fr/sites/itellier/SEM.html`
2. Il est possible de télécharger une révision sans versionnement depuis : `https://code.launchpad.net/~yoann-dupont/crftagger/stand-alone-tagger`

sont parmi les plus pertinents, un langage (syntaxe XML) permet de définir des informations à extraire afin d'enrichir le corpus telles que :
– des observations locales (ex : le mot courant commence-t-il par une majuscule ? Est-il en début de phrase ?).
– des conjonctions/disjonctions d'observations locales (ex : le mot courant commence-t-il par une majuscule sans être en début de phrase ?).
– ajouter des dictionnaires de mots et de séquences de mots.
L'étiqueteur POS a été appris sur le French Treebank et validé selon un processus de validation croisée. Il reconnaît les étiquettes définies dans (Crabbé & Candito, 2008) avec une F-mesure de 97,3% en supposant les unités multi-mots déjà segmentées (c'est-à-dire regroupées en un seul token), et une de 95,2% lorsqu'elles ne sont pas déjà segmentées.

Le programme intègre plusieurs ressources externes, dont le LeFFF (Sagot, 2010) et divers dictionnaires constitués à partir de wikipedia français.

Le FTB annoté en entité nommées dispose de 7 types généraux que nous cherchons à reconnaître : Company (entreprise), FictionCharacter (personnage de fiction), Location (lieu), Organization (association ou organisation à but non lucratif par exemple), POI (Point Of Interest), Person (Personne) et Product (Produit).

L'évaluation du reconnaisseur d'entités nommées s'est faite selon un processus de validation croisée sur cinq plis en partant d'une annotation POS parfaite et en considérant l'égalité stricte sur les séquences des entités nommées (égalité du type et des frontières). En micro-average, la précision est de 86.38, le rappel de 80.30 pour une f-mesure de 83.23. En macro-average, les précision, rappel et f-mesure sont respectivement de 77.38, 53.01 et 62.92. La différence de qualité entre la micro- et la macro-average est due aux classes FictionCharacter et POI, dont la faible représentation dans le corpus rend leur identification particulièrement difficile par des outils statistiques.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In A. ABEILLÉ, Ed., *Treebanks*. Dordrecht : Kluwer.

APPELT D. E., HOBBS J. R., BEAR J., ISRAEL D., KAMEYAMA M., MARTIN D., MYERS K. & TYSON M. (1995). Sri international fastus system : Muc-6 test results and analysis. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, p. 237–248, Stroudsburg, PA, USA : Association for Computational Linguistics.

CRABBÉ B. & CANDITO M. H. (2008). Expériences d'analyse syntaxique statistique du français. In *Actes de TALN'08*.

FAVRE B. & BÉCHET F. (2005). Robust named entity extraction from large spoken archives. In *Proc. of the Emperical Methods in Natural Language Processing*.

GRAVIER G., ADDA G., PAULSSON N., CARR'E M., GIRAUDEL A. & GALIBERT O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *LREC*.

GRAVIER G., BONASTRE J., GALLIANO S., GEOFFROIS E., TAIT K. M. & CHOUKRI K. (2004). Ester, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques. In *Journées d'Etude sur la Parole*.

KIM J.-D., OHTA T., TSURUOKA Y., TATEISI Y. & COLLIER N. (2004). An itroduction to the bio-entity recognition task at jnlpba. In *Proceedings of Natural Language Processing in Biomedical Applications (NLPBA 2004)*.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, p. 282–289.

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL I. & NOUVEL D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. In *Actes de TALN'11*.

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)*.

SAGOT B., RICHARD M. & STERN R. (2012). Annotation référentielle du corpus arboré de paris 7 en entités nommées. In *Actes de TALN'12, papier court (poster)*.

TELLIER I. & TOMMASI M. (2011). *Eric GAUSSIER et François YVON, Champs Markoviens Conditionnels pour l'extraction d'information*, chapter Modèles probabilistes pour l'accès à l'information textuelle. Hermès.

TJONG KIM SANG & ERIK F. & DE MEULDER F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL'03, p. 142–147, Stroudsburg, PA, USA : Association for Computational Linguistics.