Un étiqueteur en ligne du français

Yoann Dupont[1, 2] Clément Plancq[1]
(1) Laboratoire Lattice (CNRS, ENS, Université Sorbonne Nouvelle, PSL Research University, USPC)
1 rue Maurice Arnoux, 92120 Montrouge
(2) Expert System France, 207 rue de Bercy, 75012 Paris
yoa.dupont@gmail.com, clement.plancq@ens.fr

RÉSUMÉ
Nous proposons ici une interface en ligne pour étiqueter des textes en français selon trois niveaux d'analyse : la morphosyntaxe, le chunking et la reconnaissance des entités nommées. L'interface se veut simple et les étiquetages réutilisables, ces derniers pouvant être exportés en différents formats.

ABSTRACT
An online tagger for French

We propose here an online interface for tagging French texts according to three levels of analysis : morphosyntax, chunking and named entity recognition. The interface is simple and the taggings are reusable as they can be exported in different formats.

MOTS-CLÉS : Reconnaissance d'entités nommées, French Treebank, Apprentissage automatique, CRF, IHM, en ligne.

KEYWORDS: named entity recognition, French Treebank, machine learning, CRF, GUI, online.

1 Introduction

À l'heure où les interfaces en ligne se multiplient, de plus en plus d'outils de TAL deviennent accessibles facilement, s'ouvrant alors aux non spécialistes. Dans le cadre de cette démonstration, nous nous intéresserons à la reconnaissance d'entités nommées. Il existe pour cette tâche un certain nombre d'interfaces en ligne, parmi lesquelles on peut citer celle de Cognitive Computational Group de l'Université Illinois [1] et celle d'Explosion AI [2]). De telles interfaces pour le français sont au mieux rares et sont des ressources précieuses. Notre interface [3] est une surcouche à SEM (Tellier *et al.*, 2012; Dupont & Tellier, 2014), un programme également libre [4].

Afin d'effectuer les différentes tâches d'étiquetage, nous avons entraîné un CRF (Lafferty *et al.*, 2001) sur le French Treebank (FTB) (Abeillé *et al.*, 2003). Nous avons utilisé Wapiti (Lavergne *et al.*, 2010) comme implémentation des CRF. Le FTB annoté en entités nommées (Sagot *et al.*, 2012) distingue 7 types d'entités principaux : *Company* (les entreprises), *Location* (les lieux), *Organization* (les organisations à but non lucratif), *Person* (les personnes réelles), *Product* (les produits), *FictionCharacter* (les personnages fictifs) et finalement les *PointOfInterest* (les points

1. http://cogcomp.cs.illinois.edu/page/demo_view/NERextended
2. https://demos.explosion.ai/displacy-ent/
3. accessible à l'adresse suivante : http://apps.lattice.cnrs.fr/sem/
4. disponible à l'adresse suivante : https://github.com/YoannDupont/SEM

d'intérêt). Le découpage du corpus suit le protocole entraînement–développement–test défini par Crabbé & Candito (2008). Notre étiqueteur en entités nommées atteint une précision de 87.89, un rappel de 82.34 et une f-mesure de 85.02.

2 Interface

L'interface se veut la plus simple possible, afin d'être accessible au plus grand monde. Son fonctionnement se découpe en deux passes : 1. rentrer le texte à annoter 2. récupérer le résultat dans le format attendu. Plusieurs niveaux d'analyse sont offerts à l'utilisateur : étiquetage morphosyntaxique, chunking (Abney, 1991) et entités nommées.

Une visualisation du texte en HTML avec les différents niveaux d'annotation est disponible directement afin que l'utilisateur puisse évaluer les résultats. Chaque niveau d'annotation est visible séparément et chaque élément dans un niveau spécifique est surligné, les éléments d'un même type étant surlignés avec la même couleur. La structure initiale du texte est préservée au maximum afin d'en faciliter la lecture. L'un des intérêts de notre interface est d'offrir la possibilité de récupérer les données étiquetées, en différents formats. À notre connaissance, les données ne sont que très rarement téléchargeables ou seulement en un unique format. Ces formats d'export sont les suivants : texte linéaire, tabulaire, HTML pour visualiser les annotations ainsi que json pour améliorer la réutilisation sont disponibles. Il est prévu que d'autres formats de sortie soient ajoutés, notamment le XML-TEI.

La taille maximale du texte soumis est fixée à 50 000 mots graphiques, au-delà nous conseillons aux utilisateurs d'installer le programme sur leur ordinateur. Le texte de l'utilisateur ainsi que le produit de l'annotation sont stockés dans des fichiers dont l'existence est liée à un cookie de session. La fermeture de la session web déclenche la suppression des fichiers sur le serveur.

Références

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

Abney S. (1991). Parsing by chunks. In *Principle-Based Parsing*, p. 257–278 : Kluwer Academic Publishers.

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

Dupont Y. & Tellier I. (2014). Un reconnaisseur d'entités nommées du français. In *TALN 2014*, p. 40–41.

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

Lavergne T., Cappé O. & Yvon F. (2010). Practical very large scale CRFs. In *Proceedings of ACL'2010*, p. 504–513 : Association for Computational Linguistics.

Sagot B., Richard M. & Stern R. (2012). Annotation référentielle du corpus arboré de paris 7 en entités nommées. In *Traitement Automatique des Langues Naturelles (TALN)*, volume 2.

Tellier I., Dupont Y. & Courmet A. (2012). Un segmenteur-étiqueteur et un chunker pour le français. *JEP-TALN-RECITAL 2012*, p. 7–8.