Un outil d'étiquetage rapide et un corpus libre en entités nommées du Français

Yoann Dupont[1]
(1) Laboratoire Lattice (CNRS, ENS, Université Sorbonne Nouvelle, PSL Research University, USPC)
1 rue Maurice Arnoux, 92120 Montrouge
yoa.dupont@gmail.com

RÉSUMÉ

Dans cet article, nous présentons un outil pour effectuer l'étiquetage rapide de textes bruts. Il peut charger des documents annotés depuis divers formats, notamment BRAT et GATE. Il se base sur des raccourcis claviers intuitifs et la diffusion d'annotation à l'échelle du document. Il permet d'entraîner des systèmes par apprentissage que l'on peut alors utiliser pour préannoter les textes.

ABSTRACT

A fast tagging tool and a free French named entity corpus

In this article we present a tool for fast tagging of raw texts. It handles multiple input and output formats, such as BRAT and GATE. For fast tagging, the tool relies on intuitive keyboard shortcut and document-wide annotation broadcasting. The tools allows to train machine learning systems that can be used to preannotate texts.

MOTS-CLÉS : étiquetage, entités nommées, corpus, annotation structurée, GUI.

KEYWORDS: tagging, named entities, corpus, structured tagging, GUI.

1 Introduction

Pour de nombreuses tâches de TAL, des textes annotées sont capitaux mais demeurent trop peu nombreux, ou ont une licence restrictive. Il existe déjà de nombreux outils pour annoter des textes bruts, parmi lesquels nous pouvons citer GATE (Cunningham *et al.*, 2013) ou BRAT (Stenetorp *et al.*, 2012). Ces outils ont cependant deux inconvénients principaux : le premier est d'être plutôt lents pour annoter et le second est qu'ils ne gèrent qu'un format, le leur. Pour cette raison, nous proposons ici un outil d'annotation rapide et capable de gérer des données de formats divers. L'outil que nous présentons est un module de SEM (Dupont, 2017).

2 L'outil et le corpus "preuve de concept"

L'outil que nous présentons a été conçu pour annoter rapidement dans le cadre de tâches comme l'étiquetage morphosyntaxique ou la reconnaissance d'entités nommées, mais peut se montrer utile pour toute tâche où des empans textuels doivent être annotés. Il est écrit en python en utilisant la librairie Tkinter (Shipman, 2013) Il permet d'annoter un corpus document par document. Afin

d'améliorer la vitesse d'annotation l'outil recourt à des raccourcis claviers déduits du jeu d'annotation qui doit être chargé (plusieurs jeux peuvent être gérés de manière indépendante). Si le jeu d'annotation contient un type "lieu", son raccourci par défaut sera "l". Nombre d'éléments peuvent se trouver répétés à de nombreuses reprises dans le texte. Par exemple, annoter toutes les occurrences d'une même personnes d'un roman peut s'avérer fastidieux et sujet à l'erreur s'il faut annoter les éléments un à un. Pour combler ce problème, si l'utilisateur souhaite annoter un élément textuel, il peut diffuser l'annotation à l'échelle du document. Cette opération n'est pas sans source d'erreur, il n'est pas impossible que certains "Rennes" annotés soient en fait "Inria de Rennes". Pour gérer ce cas, l'outil propose d'explorer l'historique des annotations, classées par date décroissante, effectuées par l'utilisateur afin de les réviser. Une autre source de vitesse est dans l'utilisation de données préannotatées et dans l'apprentissage de systèmes à partir des données annotées. L'outil propose actuellement d'entraîner des CRF (Lafferty *et al.*, 2001) à l'aide de Wapiti (Lavergne *et al.*, 2010) et prévoit d'intégrer l'entraînement de modèles neuronaux. À terme, il proposera d'entraîner des systèmes sur d'autres tâches, où les annotations ont une structure arborée.

Nous proposons, en preuve de concept, un corpus annoté en entités nommées. Le jeu d'annotation comprend les types suivants : les lieux ("Rennes", "la lune", "la Loire", etc.) ; les personnes ("Emmanuel Macron") ; les organisations (sans distinction entre les organisations et entreprises) ; les dates ("lundi 14 mai 2018", "mai 2018", mais pas "hier") ; les heures absolues ("midi", "14 heures", mais pas "le soir") ; les objets ("satellite James Webb", "la station spatiale internationale", etc.) ; les évènements ("CORIA-TALN", mais aussi les évènements climatiques comme "la tempête Egon"). Nous avons constitué le corpus en récupérant les articles de de la partie française de Wikinews pour l'année 2017. Le contenu de chaque document ne comprend que le titre et le corps de chaque article, nous avons ignoré les images avec légende ainsi que les sources citées. Nous avons également supprimé les documents qui énumèrent des résultats sportifs. Le corpus comprend à l'heure actuelle environ 7300 annotations structurées (par exemple "Tour de France" est un évènement, mais "France" est également annoté en tant que lieu), ce qui représente environ 7 heures de travail, pour une cadence moyenne d'environ 1000 annotations par heure (hors phases d'apprentissage et d'annotation automatique).

Références

CUNNINGHAM H., TABLAN V., ROBERTS A. & BONTCHEVA K. (2013). Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLoS computational biology*, **9**(2), e1002854.

DUPONT Y. (2017). Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique. In *Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*, p. 42–55.

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.

SHIPMAN J. W. (2013). Tkinter 8.4 reference : a gui for python.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). Brat : a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107 : Association for Computational Linguistics.