Corpus EN-Istex : un corpus d'articles scientifiques annoté manuellement en entités nommées

Enza Morale, Denis Maurel, Jeanne Villaneau, Jean-Yves Antoine

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Corpus EN-ISTEX : un corpus d'articles scientifiques annoté manuellement en entités nommées

Enza Morale[1], Denis Maurel[2],
Jeanne Villaneau[3], Jean-Yves Antoine[2]

(1) Inist, CNRS, Nancy, France

(2) Université de Tours, Lifat, Tours, France

(3) Ensibs, Irisa, Lorient, France

enza.morale@inist.fr, denis.maurel@univ-tours.fr,
Jean-Yves.Antoine@univ-tours.fr, jeanne.villaneau@univ-ubs.fr

RESUME

Nous présentons ici une nouvelle ressource libre : le corpus EN-ISTEX, un corpus de deux cents articles scientifiques annotés manuellement en entités nommées. Ces articles ont été extraits des deux éditeurs scientifiques les plus importants de la plateforme ISTEX. Tous les domaines sont concernés, même si les sciences dites *dures*, en particulier les sciences du vivant et de la santé, sont prépondérantes.

Parmi ceux-ci vingt articles ont été multi-annotés afin de vérifier l'adéquation du guide d'annotation et la fiabilité de l'annotation. L'accord inter annotateurs sur ces vingt textes s'élève à 91 %.

ABSTRACT

ISTEX-EN Corpus: a scientific paper corpus manually annotated in named entities
We present here a new free resource: the EN-ISTEX Corpus, a corpus of two hundred scientific papers manually annotated in named entities. These papers have been extracted from the two more representative scientific publishers of ISTEX platform. All fields are concerned, even if the so-called hard sciences, in particular the life sciences and health, are predominant.

Among these, twenty papers were multi-annotated in order to verify the adequacy of the annotation guide and the reliability of the annotation. The inter-annotator agreement on these twenty texts amounts to 91%.

MOTS-CLES : corpus annoté, entités nommées, ressource libre, articles scientifiques, accord inter annotateurs.

KEYWORDS: annotated corpus, named entities, free resource, scientific papers, inter annotator agreement.

---

1 Introduction

Le projet investissement d'avenir ISTEX[1] avait pour but la constitution d'une bibliothèque d'articles scientifiques disponibles en libre accès pour les acteurs de la recherche en France. Cette bibliothèque numérique comporte aujourd'hui « *23 millions de documents provenant de 30 corpus de littérature scientifique dans toutes les disciplines* »[2]. Pour améliorer la consultation de la plateforme ISTEX, des services à valeurs ajoutées ont été développés et sont disponibles via l'API d'ISTEX. Parmi ceux-ci, un service permet une interrogation de la base via les entités nommées (noms propres, dates et références). Ce service a bien sûr été testé en interne, mais l'idée est venue de constituer manuellement un corpus annoté en entités nommées avec un accord inter annotateurs, utilisable comme corpus d'apprentissage pour la création ou l'amélioration d'outils de détection d'entités nommées. Il s'agit d'une nouvelle ressource libre (sous licence ouverte Etalab[3]), disponible à l'URL https://corpus-gold.corpus.istex.fr/.

Ce corpus contient deux cents articles, issus des éditeurs Wiley et Elsevier, sélectionnés à partir des catégories scientifiques Science-Metrix de niveau 1, proportionnellement à l'ensemble constitué de ces deux éditeurs dans le fonds Istex. Parmi ceux-ci, vingt articles ont fait l'objet d'une annotation multiple avec calcul de l'accord inter annotateurs et choix collégial de la bonne annotation. Cet accord, calculé par le *alpha* de Krippendorff, est très bon, puisqu'il atteint 91 %. Le travail sur ces vingt articles a permis la formation des annotateurs qui ont ainsi acquis une capacité à annoter les textes de façon similaire. De ce fait, l'ensemble du corpus EN-ISTEX peut être considéré comme un *gold standard*. L'interprétation du *alpha* porte toujours à débat, mais pour un coefficient alpha de cet ordre, cette ressource peut être qualifiée d'une très bonne fiabilité.

2 Démonstration

Nous présenterons :

- la constitution du corpus à partir des catégories scientifiques Scopus ;
- la campagne d'annotation ;
- le guide d'annotation (téléchargeable sur le site) avec différents exemples, en particulier quelques points qui ont donné lieu à discussion ;
- L'accès au corpus annoté via le site de publication des corpus Istex : http://data.istex.fr/.

---

1 Le projet ISTEX (ANR-10-IDEX-0004-02) s'est déroulé d'avril 2012 à décembre 2018.

2 D'après le site https://www.istex.fr/, consulté le 26/11/2020.

3 https://www.etalab.gouv.fr/licence-ouverte-open-licence