Uniformisation de corpus anglais annotés en sens

Loïc Vial, Benjamin Lecouteux, Didier Schwab

Uniformisation de corpus anglais annotés en sens

Loïc Vial, Benjamin Lecouteux, Didier Schwab
GETALP – LIG – Univ. Grenoble Alpes
{loic.vial, benjamin.lecouteux, didier.schwab}
@univ-grenoble-alpes.fr

RÉSUMÉ

Pour la désambiguïsation lexicale en anglais, on compte aujourd'hui une quinzaine de corpus annotés en sens dans des formats souvent différents et provenant de différentes versions du *Princeton WordNet*. Nous présentons un format pour uniformiser ces corpus, et nous fournissons à la communauté l'ensemble des corpus annotés en anglais portés à notre connaissance avec des sens uniformisés du *Princeton WordNet* 3.0, lorsque les droits le permettent et le code source pour construire l'ensemble des corpus à partir des données originales.

ABSTRACT

Unification of sense annotated English corpora for word sense disambiguation

In word sense disambiguation, there are today about almost fifteen sense annotated English corpora, in various formats and using different versions of *Princeton WordNet*. We present a format that unifies these corpora, and we give to the community the whole set of corpora sense annotated in English that we know, with senses from *Princeton WordNet* 3.0, in this unified format, when the copyright allows it, and the source code for building these corpora from the original data.

MOTS-CLÉS : désambiguïsation lexicale, corpus annotés en sens, ressource uniformisée.

KEYWORDS: word sense disambiguation, sense annotated corpora, unified resource.

1 Introduction

Que ce soit pour l'évaluation ou l'apprentissage d'un système de désambiguïsation lexicale (DL), les corpus annotés en sens sont essentiels. En effet, les systèmes de DL exploitant les exemples issus de corpus annotés en sens sont généralement bien meilleurs que ceux qui n'en exploitent pas (Navigli *et al.*, 2007; Moro & Navigli, 2015).

En anglais, le *Princeton WordNet* (Miller, 1995) est aujourd'hui la base lexicale standard *de facto*. La plupart des corpus annotés en sens sont ainsi soit annotés directement grâce à WordNet soit annotés avec un inventaire de sens lié aux sens de WordNet comme BabelNet (Navigli & Ponzetto, 2010).

Il n'est toutefois pas aisé d'utiliser ces corpus car la plupart diffèrent grandement par leur format et par la version du *Princeton WordNet* utilisée. De plus, les systèmes sont systématiquement évalués sur les corpus destinés à l'origine à l'évaluation et jamais sur les corpus destinés à l'origine à un autre usage sans qu'il n'y ait de raison scientifique pour cela.

Nous présentons ainsi un travail d'unification de tous les corpus anglais annotés avec WordNet portés à notre connaissance, dans un format unique, simple à comprendre et rapide à utiliser en pratique. Nous mettons au même plan les corpus destinés à l'origine à l'évaluation et ceux destinés à l'apprentissage, pour faciliter la construction de systèmes de DL qui pourraient ainsi réaliser une

évaluation à plus large échelle en procédant, par exemple, à une validation croisée par rotation dans laquelle on utilise tour à tour chacun des corpus pour l'évaluation d'un système et l'ensemble des autres pour sa construction.

Nous avons aussi effectué la conversion de toutes les annotations en sens depuis leur version de WordNet d'origine à la dernière version (3.0) grâce à des tables de conversion dont la méthode de fabrication est issue de Daudé *et al.* (2000) [1].

Notre travail est proche de celui de Raganato *et al.* (2017) mais il diffère en deux points : premièrement, ils séparent les corpus en corpus d'évaluation et corpus d'apprentissage, et deuxièmement, ils n'intègrent que 7 corpus contre 12 pour nous.

Nous fournissons du code Java permettant de lire et écrire facilement ce format, ainsi que tous les corpus utilisés, à la fois dans leur format original ainsi que dans notre format. Le code qui nous a permis de faire la conversion est lui aussi fourni. Le tout est disponible à l'adresse suivante : `https://github.com/getalp/WSD-TALN2017-Corpus-Vialetal`

2 Corpus anglais annotés en sens

Notre ressource contient tous les corpus anglais annotés en sens *Princeton WordNet* à notre connaissance, c'est à dire :

- Le SemCor (Miller *et al.*, 1993), annoté originellement avec WordNet 1.6 ;
- Le DSO (Ng & Lee, 1996), annoté avec WordNet 1.5 ;
- Le corpus des définitions de WordNet [2] , annotées en sens depuis la version 3.0 ;
- L'OMSTI (Taghipour & Ng, 2015), annoté avec WordNet 3.0 ;
- Le MASC (Nancy Ide & Passonneau, 2008), annoté avec WordNet 3.0 ;
- L'Ontonotes (`https://catalog.ldc.upenn.edu/ldc2013t19`), annoté avec WordNet 3.0 ;
- Les 6 corpus des campagnes d'évaluation de DL pour l'anglais SemEval-SensEval.

3 Format de corpus unifié

Notre format de corpus est voulu pour être clair et facile à comprendre, tout en étant à la fois efficace à traiter, et contenant toutes les informations données par les différents corpus originaux.

Ainsi, nous avons opté pour un format descriptif XML, composé des 5 noeuds suivants : `corpus`, `document`, `paragraph`, `sentence` et `word`. À n'importe quel noeud, on peut y attacher un ou plusieurs attributs quelconques. Tous les attributs sont optionnels, sauf pour l'attribut `surface_form` d'un mot qui correspond à sa forme de surface. Les parties du discours sont annotés avec l'attribut `pos`, les lemmes avec `lemma`, les clés de sens pour une version spécifique de *Princeton WordNet* avec `wn{version}_key`.

L'extrait suivant est un exemple de XML résultant :

```
<corpus>
 <document id="d001" >
  <paragraph>
   <sentence id="d001.s001" >
    <word surface_form="exemple" lemma="exemple" wn30_key="example%1:09:00::" />
   </sentence>
  </paragraph>
 </document>
</corpus>
```

1. `http://www.talp.upc.edu/index.php/technology/tools/45-textual-processing-tools/98-wordnet-mappings/`
2. `http://wordnet.princeton.edu/glosstag.shtml`

Références

DAUDÉ J., PADRÓ L. & RIGAU G. (2000). Mapping wordnets using structural information. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, p. 504–511, Stroudsburg, PA, USA : Association for Computational Linguistics.

MILLER G. A. (1995). Wordnet : A lexical database. *ACM*, **Vol. 38**(No. 11), p. 1–41.

MILLER G. A., LEACOCK C., TENGI R. & BUNKER R. T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, p. 303–308, Stroudsburg, PA, USA : Association for Computational Linguistics.

MORO A. & NAVIGLI R. (2015). Semeval-2015 task 13 : Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 288–297, Denver, Colorado : Association for Computational Linguistics.

NANCY IDE, COLLIN BAKER C. F. C. F. & PASSONNEAU R. (2008). Masc : the manually annotated sub-corpus of american english. In B. M. J. M. J. O. S. P. D. T. NICOLETTA CALZOLARI (CONFERENCE CHAIR), KHALID CHOUKRI, Ed., *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA). http ://www.lrec-conf.org/proceedings/lrec2008/.

NAVIGLI R., LITKOWSKI K. C. & HARGRAVES O. (2007). Semeval-2007 task 07 : Coarse-grained english all-words task. In *SemEval-2007*, p. 30–35, Prague, Czech Republic.

NAVIGLI R. & PONZETTO S. P. (2010). Babelnet : Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, p. 216–225 : Association for Computational Linguistics.

NG H. T. & LEE H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense : an exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, p. 40–47, Stroudsburg, PA, USA : Association for Computational Linguistics.

RAGANATO A., CAMACHO-COLLADOS J. & NAVIGLI R. (2017). Word sense disambiguation : A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 99–110, Valencia, Spain : Association for Computational Linguistics.

TAGHIPOUR K. & NG H. T. (2015). One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, p. 338–344, Beijing, China : Association for Computational Linguistics.