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
LIG - GETALP, Univ. Grenoble Alpes, France

Corpus anglais annotés en sens

Les **corpus annotés en sens** sont une **ressource rare**, et pourtant **indispensable** pour l'apprentissage et/ou l'évaluation de systèmes de **désambiguïsation lexicale**.
En anglais, on compte moins d'une quinzaine de corpus annotés en sens, et leurs formats sont très différents les uns des autres.
Notre hypothèse est que l'**unification** des corpus dans un **format simple** à comprendre et **rapide** à utiliser permettra de faciliter la création de nouveaux systèmes de disambiguïsation, et l'évaluation des systèmes existant.

Statistiques sur les corpus

Ressource	Phrases	Mots		Parties du discours annotées				Version de WordNet
		Total	Annotés	Noms	Verbes	Adj.	Adv.	
SemCor [1]	37 176	778 587	229 517	87 581	89 037	33 751	19 148	1.6
DSO [2]	178 119	5 317 184	176 915	105 925	70 990	0	0	1.5
WordNet GlossTag [3]	117 659	1 634 691	496 776	232 319	62 211	84 233	19 445	3.0
MASC [4]	34 217	596 333	114 950	49 263	40 325	25 016	0	3.0
OMSTI [5]	820 557	3 5843 024	920 794	476 944	253 644	190 206	0	3.0
Ontonotes [6]	21 938	435 340	52 263	9 220	43 042	0	0	3.0
SemEval 2007 task 07 [7]	245	5 637	2 261	1 108	591	356	206	2.1
SemEval 2007 task 17 [8]	120	3 395	455	159	296	0	0	2.1
SemEval 2013 task 12 [9]	306	8 142	1 644	1 644	0	0	0	3.0
SemEval 2015 task 13 [10]	138	2 638	1 053	554	251	166	82	3.0
Senseval 2 [11]	238	5 589	2 301	1 061	541	422	277	1.7.1
Senseval 3 task 1 [12]	300	5 511	1 957	886	723	336	12	1.7.1

Ressources

Les corpus enregistrés dans le nouveau format sont accessibles, lorsque les droits le permettent, à l'adresse suivante :

Spécifications du nouveau format

Le **nouveau format** de corpus doit pouvoir contenir toutes les informations contenues dans les formats d'origine.

Ces informations sont compilées et généralisées aux concepts suivants :

- Le **corpus** est l'entité lexicale racine, qui contient un ensemble de document ;
- Le **document** est une entité lexicale qui contient un ensemble de paragraphe ;
- Le **paragraphe** est une entité lexicale qui contient un ensemble de phrases ;
- La **phrase** est une entité lexicale qui contient un ensemble de mots ;
- Le **mot** est l'entité lexicale feuille ;
- Enfin, chaque **entité lexicale** contient un ensemble d'**annotations**.

Pour représenter ces concepts, nous avons basé notre format sur une syntaxe **XML** simple, et respectant plusieurs conventions :

- Chaque **nœud** (balise XML) représente une **entité lexicale** ;
- Les **annotations** sont des **attributs** des nœuds.

Les **annotations** sont :

- La **forme de surface** (`surface_form`) d'un mot ;
- Le **lemme** (`lemma`) d'un mot ;
- La **partie du discours** (`pos`) d'un mot ;
- Le **sens** d'un mot, dans une base lexicale spécifique, par exemple WordNet 3.0 (`wn30_key`).

Exemple de format original

OMSTI : 2 fichiers **par mot**, plus de 20 000 mots

Extrait de "action.xml" :

```
<instance id="action.100476" docsrc="br-a01">
  <context>
    These
      <head>
        actions
      </head>
    should serve [...]
```

Extrait de "action.key" :

```
action action.100476 action%1:04:02::
```

Nouveau format

OMSTI : 1 seul fichier "omsti.xml"
Extrait :

```
<corpus>
  <document>
    <paragraph>
      <sentence>
        <word surface_form="These" pos="DT" />
        <word surface_form="actions" pos="NNS"
              wn30_key="action%1:04:02::" />
        <word surface_form="should" pos="MD" />
        <word surface_form="serve" lemma="serve"
                                   pos="VB" />
        [...]
```

Références

[1] George Miller *et al.* A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics.
[2] Hwee Tou Ng and Hian Beng Lee. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *ACL 1996*, pages 40–47, Stroudsburg, PA, USA, 1996.
[3] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
[4] Nancy Ide *et al.* Masc: the manually annotated sub-corpus of american english. In *LREC 2008)*, Marrakech, Morocco, may. http://www.lrec-conf.org/proceedings/lrec2008/.
[5] Kaveh Taghipour and Hwee Tou Ng. One million sense-tagged instances for word sense disambiguation and induction. In *CoNLL 2015*, pages 338–344.
[6] Eduard Hovy *et al.* Ontonotes: The 90% solution. In *NAACL 2006*, pages 57–60, Stroudsburg, PA, USA.
[7] Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. Semeval-2007 task 07: Coarse-grained english all-words task. In *SemEval-2007*, pages 30–35, Prague, Czech Republic, June 2007.
[8] Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. Semeval-2007 task 17: English lexical sample, srl and all words. In *Semeval 2007*, pages 87–92, Stroudsburg, PA, USA.
[9] Roberto Navigli, David Jurgens, and Daniele Vannella. Semeval-2013 task 12: Multilingual word sense disambiguation. In *SemEval@NAACL-HLT*, 2013.
[10] Andrea Moro and Roberto Navigli. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. *Proc. of SemEval-2015*, 2015.
[11] Philip Edmonds and Scott Cotton. Senseval-2: Overview. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, SENSEVAL 2001, pages 1–5.
[12] Kenneth C. Litkowski. Senseval-3 task: Word-sense disambiguation of wordnet glosses. In *SENSEVAL-3*.