UFSAC: Unification of Sense Annotated Corpora and Tools

Loïc Vial, Benjamin Lecouteux, Didier Schwab

UFSAC: Unification of Sense Annotated Corpora and Tools

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

Abstract

In word sense disambiguation, sense annotated corpora are often essential for evaluating a system and also valuable in order to reach a good efficiency. Always created for a specific purpose, there are today almost fifteen sense annotated English corpora, in various formats and using different versions of WordNet. The main hypothesis of this work is that it should be possible to build a disambiguation system by using any of these corpora during the training phase or during the testing phase regardless of their original purpose. In this article, we present UFSAC: a format of corpus that can be used for either training or testing a disambiguation system, and the process we followed for constructing this format. We give to the community the whole set of sense annotated English corpora that we know, in this unified format, when the copyright allows it, with sense keys converted to the last version of WordNet. We also provide the source code for building these corpora from their original data, and a complete Java API for manipulating corpora in this format.

Keywords: word sense disambiguation, sense annotated corpora, unified resource, tools

1. Introduction

Whether they are used for the evaluation or for the learning process of a Word Sense Disambiguation (WSD) system, the importance of sense annotated corpora in Natural Language Processing (NLP) is considerable. On one hand, the evaluation *in vivo*, i.e. the evaluation of a WSD system as part of a larger task, has never been really exploited. On the other hand, the evaluation *in vitro*, which uses directly sense annotated corpora by comparing the output of a system to manual annotations, is predominant. Moreover, WSD systems exploiting examples from sense annotated corpora are generally far better than those which do not (Navigli et al., 2007; Moro and Navigli, 2015).

At the time of its creation, WordNet (Miller, 1995) was undoubtedly the only lexical database freely available for English. Since the beginning of the 2000s, it has become the *de facto* standard for WSD in this language. Indeed, most of sense annotated corpora are either directly annotated with WordNet sense keys or they are annotated with a sense inventory linked to the senses of WordNet, such as BabelNet (Navigli and Ponzetto, 2010).

However, it is not trivial to use these corpora, because most of them differ in their format and on the version of WordNet they use. As a consequence, very few works in the literature of WSD are trained or evaluated on more than two annotated corpora.

Also, WSD systems are systematically evaluated on corpora that have been initially created for the purpose of evaluation, and never on corpora that have been created for another purpose, such as training or for sense distribution estimation, whereas there is no scientific reason for that.

This paper presents a work of unification of all existing English corpora annotated with any version of WordNet to our knowledge, in a unique format, easy to understand, and easy to work with in practice. We put on the same level the corpora originally created for the evaluation and those for the learning, so to facilitate the creation of robust WSD systems which could for example be evaluated in a way where all corpora except one are used for the learning, and the remaining one is used for the evaluation, then switch the corpora and do this for every existing corpus.

The language resource that we provide contains all English sense annotated corpora in UFSAC (Unified Format for Sense Annotated Corpora), the format that we propose, with sense annotations converted to the last version of WordNet (3.0), along with Java code to easily read, write and modify any corpus in this format, and scripts for converting a corpus from its original format to UFSAC.

Our work differs from the recent work of (Raganato et al., 2017) in several points. Their work is more about the evaluation of WSD systems, so they split their corpora set in a training and an evaluation set. We also propose five additional corpora in our resource among the most difficult to parse. And finally we provide a complete API for manipulating corpora in UFSAC, and conversion scripts allowing the full reconstruction of the corpora from the original data. In our resource, we provide a script for converting a corpus from our format to theirs, so existing WSD systems that rely on their format can be trained or evaluated on any of the corpus that we produced. We also provide a script for converting their format to ours in order to facilitate any collaborative work in the community.

2. Sense Annotated Corpora: rare and costly resources

Generally speaking, a corpus is a collection of documents which can be used as samples of text for a particular language (Habert et al., 1998). A corpus may contain several millions of words, which can be lemmatized and annotated with information concerning their part of speech for example. Among these corpora, we can find the *British National Corpus* (Burnard, 1998) (100 million words) and the *American National Corpus* (Ide and Macleod, 2001) (20 million words). The texts come from several sources such as newspapers, books, encyclopedias or from the Web.

A sense annotated corpus is a corpus in which some or all words are annotated with an identifier of sense that comes from a specific lexical database. For example, in the cor-

pus of the 7th task of the SemEval 2007 semantic evaluation campaign (Navigli et al., 2007), all words are annotated with sense identifiers from WordNet 2.1, whereas in the English corpus of the 13th task of SemEval 2015 (Moro and Navigli, 2015), all words are annotated with sense identifiers from WordNet 3.0, BabelNet 2.5 and Wikipedia pages.
There are at least three reasons to create a sense annotated corpus:

- Estimate the distribution of senses in the language. It is for this purpose that the SemCor (Miller et al., 1993) was annotated. And consequently, the senses in WordNet are, since version 1.7, sorted by this distribution of senses estimated on the SemCor.
- Build a Word Sense Disambiguation system which learns from examples contained in the annotated corpus. For instance, the OMSTI (Taghipour and Ng, 2015) was created for this purpose.
- Evaluate a WSD system by comparing its output to the annotations in the corpus, as it is the case for instance with corpora created as part of the evaluation campaigns SensEval-SemEval.

After their distribution, there is no scientific reason not to use indistinctly these corpora either for building a WSD system, for estimating the distribution of senses or for evaluating a WSD system. Indeed, the SemCor is used since a long time for the learning of WSD systems (Chan et al., 2007; Navigli et al., 2007) or more recently for the evaluation of different methods (Yuan et al., 2016). This last usage is still very rare, since it is one of the first experiment that we found in the literature, along with (Màrquez et al., 2002).
However, the format of the resources differs greatly depending on their original purpose. For the SemCor, a single file groups all the information, whereas in the case of the evaluation corpora, there are two files: one that contains the unannotated corpus, and the other that contains the sense annotations. In some corpora, like in the DSO and the OMSTI, there is one file for every lemma in the dictionary, and each file contains thousands of example sentences, where this lemma is the only word that is sense annotated.
Few data are manually sense annotated. The *Global WordNet Association* made a list of 26 corpora annotated with WordNet [1]. These corpora concern 17 languages, but only three of them reach 100,000 annotations. English, with more than 2 million words sense annotated ranks first, before Dutch with nearly 300,000 annotations and Bulgarian with 100,000 annotations. Thus, it is unsurprising that most of researches in WSD focus on English.

3. A single format for sense annotated corpora

The main purpose of this work is to help the construction and the evaluation of WSD systems, by giving to the community the set of all existing English sense annotated corpora to our knowledge, in the same format, using the same sense inventory, and tools to easily parse them, manipulate them, and convert corpora from their original format to our one.
Indeed, a large quantity of sense annotated data is vital for the construction of WSD systems. In evaluation campaigns, this often makes the difference. For example, looking at the data from the SemEval 2007 campaign (Navigli et al., 2007), which most of the recent systems were evaluated on, we observe that systems that did not use sense annotated data obtain a precision score up to $78-79\%$[2] (Schwab et al., 2013) (Chen et al., 2014) whereas those which use a lot of annotated data reach a score up to 82% (Chan et al., 2007) (Navigli, 2012) (Vial et al., 2016) and even 84% (Yuan et al., 2016).
Therefore, having all existing corpora in a unique format and using the same sense inventory offers several advantages: it allows to easily expand the quantity of data available for improving WSD systems, it allows to better estimate the distribution of senses in English, and finally, this format can help creating more robust WSD systems. Indeed, we still find a lot of works that focus on a single evaluation task (Vial et al., 2016; Chen et al., 2014), and in these cases, the analysis of the results concerning the robustness of the methods is limited. The unification of the format of sense annotated corpora could improve the evaluation process by facilitating a cross validation process for instance, where the system is evaluated sequentially on every corpus, with all others used for the training.

4. Provided resource

Our work consists in gathering all English corpora sense annotated with WordNet, and convert all of them to a unified format that is able to contain all the informations present in the original format. We created format conversion scripts for this purpose, as well as scripts for cleaning the corpora, and converting the sense annotation to the last version of WordNet (3.0). The resulting corpora are parts of the resource when the copyright allows it, along with the format conversion scripts, the cleaning scripts, and the sense conversion scripts. For the corpora that we cannot distribute because of the licence, anyone that possess them can still run our scripts to turn the original resource into our format. Finally, an API is provided for parsing, creating and manipulating corpora in our format.

4.1. Sense annotated corpora

Our resource contains the following corpora:

- The *SemCor* (Miller et al., 1993), a subset of the Brown Corpus (Francis and Kučera, 1964). Original annotations are done with WordNet 1.6.
- The *DSO (Defence Science Organisation)* (Ng and Lee, 1996), a non-free corpus, that is focused on 121 nouns and 70 verbs among the most frequently used and the most ambiguous words in English and have been annotated in various contexts with WordNet 1.5.
- The *WordNet Gloss Tag*, a corpus which consists of all definitions of WordNet [3] with every words sense

[1] http://globalwordnet.org/wordnet-annotated-corpora/

[2] This means that the system has chosen the same sense than the human annotators in 78 to 79% of cases

[3] http://wordnet.princeton.edu/glosstag.shtml

annotated since version 3.0.

- The *OMSTI (One Million Sense-Tagged Instances)* (Taghipour and Ng, 2015), a huge corpus of approximately one million words sense annotated with WordNet 3.0.
- The *MASC (Manually Annotated Sub-Corpus* (Ide et al., 2008), we used the version given in the article of (Yuan et al., 2016), annotated with the NOAD (New Oxford American Dictionary), but with corresponding WordNet 3.0 sense keys.
- The *Ontonotes 5.0* (Hovy et al., 2006), annotated with WordNet 3.0.
- The corpora of the English WSD evaluation campaigns SemEval-SensEval, and in particular the corpora of SensEval 2 (using WordNet 1.7), SensEval 3 (WordNet 1.7.1), SemEval 2007 (WordNet 2.1), SemEval 2013 (WordNet 3.0) and SemEval 2015 (WordNet 3.0).

Table 1 summarizes statistics concerning these corpora.
After the conversion of all these corpora into our format, we executed a cleaning step which consisted of trimming words, removing invisible characters, removing inconsistent annotations (i.e. when the part of speech annotation and the sense annotation differ), and finally, merging identical sentences which have annotations on different words. This last step is very important as it actually adds information to some corpora, especially the DSO and the OMSTI: because these corpora are constructed such that they contain lists of sentences with only one word that is sense annotated, surrounded by words not annotated, some sentences are present in different places across the corpus, but with different words that are sense annotated.
The merging phase identifies identical sentences with annotations on different words, and creates a single sentence containing all annotations. Thus, this steps adds a crucial information for some WSD systems. For instance, a similarity-based WSD system can now "learn" that two word senses are often located in the same sentence.
Finally, sense annotations have been converted, when necessary, from their original WordNet sense key to the last version of WordNet (3.0) thanks to conversion tables from (Daudé et al., 2000). However, because some senses have been dropped from the old versions of WordNet, some sense annotations have not been converted. In any case, the original sense annotations are always kept alongside the converted sense annotation.

4.2. UFSAC File format

Our approach for the unification of the different annotated corpora begins with a file format that is descriptive, easily understandable and readable by a human, and at the same time, efficient for a program to parse and create. Finally, it should be able to contain all the information contained in the original resources. These informations are represented with the following concepts:
– A Lexical Entity (LE) is something that contains a set of annotations.
– A Corpus is a LE which contains a set of documents.
– A Document is a LE hich contains a set of paragraphs.
– A Paragraph is a LE which contains a set of sentences.
– A Sentence is a LE which contains a set of words.
– A Word is a LE which has a special mandatory annotation "surface form", which is the value of the word.

In order to represent these concepts, UFSAC is based on a simple XML syntax with some conventions: lexical entities are represented by XML nodes (`corpus`, `document`, `paragraph`, `sentence` and `word`), and annotations are node attributes.

The annotations also follow a certain convention, we used the following to annotate words:
– The identifier (`id`) of a lexical entity, particularly useful for corpora originally created for the evaluation (e.g. "d001.s002.t003").
– The surface form (`surface_form`) of a word.
– The lemma (`lemma`) of a word.
– The part of speech (`pos`) of a word.
– The sense of a word, in a specific lexical database, for example WordNet 3.0 (`wn30_key`), WordNet 1.7.1 (`wn171_key`)...
The information of the sense is the one which is the most useful in our case, and it is specific to each lexical database, instead of having a unique "sense" annotation as we can find in most other formats. That way we allow multiples sense annotations from different lexical databases at the same time. For example, the DSO is originally annotated with senses from WordNet 1.5, and the conversion to WordNet 3.0 is sometimes impossible for some senses which were deleted between the two versions. This convention allows us to keep the original annotations, yet to have the annotations from the last version of WordNet, or any other lexical database (for instance BabelNet) at the same time.
The following is an example of the resulting UFSAC XML:

```
<corpus>
  <document id="d001" >
    <paragraph>
      <sentence >
        <word surface_form="A" pos="DT" />
        <word surface_form="precise"
              wn30_key="precise%3:00:00::" />
        <word surface_form="example"
              lemma="example" />
        <word surface_form="." />
      </sentence>
    </paragraph>
  </document>
</corpus>
```

Our format thus allows to integrate the whole corpus in a single file, and it is easily readable, especially comparing to most original formats (c.f. the end of section 2.).

4.3. API and tools

An easy-to-use Java API is also provided to read, write and modify efficiently corpora in our format. It allows two styles of programming: you can either load a full corpus in memory, perform all your calculations and save it entirely in a file; or you can sequentially scan, edit or print a corpus

Corpus	Sentences	Words		Annotated parts of speech			
		Total	Annotated	Nouns	Verbs	Adj.	Adv.
SemCor	37176	778587	229517	87581	89037	33751	19148
DSO	178119	5317184	176915	105925	70990	0	0
WordNet GlossTag	117659	1634691	496776	232319	62211	84233	19445
MASC	34217	596333	114950	49263	40325	25016	0
OMSTI	820557	35843024	920794	476944	253644	190206	0
Ontonotes	21938	435340	52263	9220	43042	0	0
SemEval 2007 task 07	245	5637	2261	1108	591	356	206
SemEval 2007 task 17	120	3395	455	159	296	0	0
SemEval 2013 task 12	306	8142	1644	1644	0	0	0
SemEval 2015 task 13	138	2638	1053	554	251	166	82
Senseval 2	238	5589	2301	1061	541	422	277
Senseval 3 task 1	300	5511	1957	886	723	336	12

Table 1: Statistics related to our set of annotated corpora, after the conversion and cleaning phase.

from a file, in a streaming manner. The latter is particularly useful when working with huge files which do not fit into memory. Finally, we offer a set of scripts that perform the conversion of a corpus from its original format to our one, and some pre-processing and analyses scripts.

4.3.1. Core API

The core API is a package containing the base classes for manipulating corpora. For simplicity, the class names match exactly what is described in section 4.2..

The class **`Annotation`** describes an annotation on a lexical entity. Concretely, it is a pair of Strings (name/value) and a pointer to the annotated lexical entity.

The class **`LexicalEntity`** describes something that has zero or more annotations, with public methods for accessing/modifying them.

The class **`Word`** inherits from LexicalEntity, has a special mandatory annotation `surface_form`, which is the value of the word, and a parent sentence.

The class **`Sentence`** inherits from LexicalEntity, contains a list of words and a parent paragraph.

The class **`Paragraph`** inherits from LexicalEntity, contains a list of sentences and a parent document.

The class **`Document`** inherits from LexicalEntity, contains a list of paragraphs and a parent corpus.

Finally, the class **`Corpus`** inherits from LexicalEntity and contains a list of documents.

These few classes, coupled with two functions `Corpus.saveToXML` and `Corpus.loadFromXML` allow to create, save, load and modify any corpus easily.

4.3.2. Streaming API

For some corpora particularly huge, like the OMSTI, we also provide a sub-package `streaming`, which allows to read, write or modify a corpus sequentially, without being fully loaded into memory. This is similar to the Java SAX library (Simple API for XML), events are fired when reading a word, sentence, paragraph, etc., and the user can choose to respond to this event or not.
In practice, we provide a set of classes which cover most use cases.

The class **`StreamingCorpusReader`** allows to respond to the events `readBeginCorpus`, `readBeginDocument`, `readWord`, etc.. This can be useful for printing every word that is sense annotated for example.

The class **`StreamingCorpusModifier`** allows to modify a corpus in-place. This is specially useful for pre-processing, for instance convert every word to lowercase.

The class **`StreamingCorpusWriter`** is used for creating a new corpus, with its methods `writeBeginSentence`, `writeWord` and so on.

4.3.3. Scripts

Finally, we provide a set of examples and useful scripts which use our format and our API, they will be more described in the final version of this article.
Important scripts are the conversion scripts, which allow to turn every corpus from its original format, to our XML format. This is specially valuable for non-free corpus like the DSO, that we cannot share directly in our format, but that people can still buy in their original format, and convert to our format.
Other useful scripts are pre-processing scripts, for example we provide a script which uses an external POS tagger and lemmatizer to annotate all words in a corpus.
To conclude, we offer also scripts that can compute the frequency of senses of each words in a set of corpora, scripts that evaluate a WSD system, and so on.

5. Conclusion

In this paper we advocate for a more uniform way of distributing sense annotated corpora, through a unique and uncomplicated file format. This unification can facilitate both the creation and the evaluation of word sense disambiguation systems. Indeed, sense annotated corpora are historically separated between those created for the purpose of training, and those created for the purpose of evaluation. In addition, the formats of these corpora are often very different from each other: different file hierarchy, different syntax, and different sense inventory are used.
The unification of all sense annotated corpora could hence allow to quickly expand a system which is trained on some

resources to new data without the effort of writing another parser. Also, a system could easily include to its training phase some corpora that were originally created for evaluation, and/or evaluate its performance on parts of corpora originally created for training. This would easily allow a much better coverage and a more fine-grained analysis of the WSD systems performance.

In our language resource, we gathered all existing English sense annotated corpora that we know, and we converted them in a simple and consistent XML file format that we named UFSAC. We also converted their sense annotations to the last version of WordNet (3.0). The corpora are only available when the licence authorizes it, but we also provide scripts that can easily convert a corpus from its original format to the one we propose. Thus, anyone who possess the corpora that we cannot distribute can still benefit from this work. In addition, we provide a complete Java API for reading, writing and modifying corpora in our unified format, along with example codes and tools for many applications such as pre-processing, sense distribution estimation, etc..

6. Extended Abstract

Ref (Schwab et al., 2015) LR ref (Speecon Consortium, 2014).

7. Bibliographical References

Burnard, L. (1998). *The British National Corpus*.

Chan, Y. S., Ng, H. T., and Zhong, Z. (2007). Nus-pt: exploiting parallel texts for word sense disambiguation in the english all-words tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 253–256. Association for Computational Linguistics.

Chen, X., Liu, Z., and Sun, M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, Doha, Qatar, October. Association for Computational Linguistics.

Daudé, J., Padró, L., and Rigau, G. (2000). Mapping wordnets using structural information. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 504–511, Stroudsburg, PA, USA. Association for Computational Linguistics.

Francis, W. N. and Kučera, H. (1964). A standard corpus of present-day edited american english, for use with digital computers (brown). Technical report, Brown University, Providence, Rhode Island.

Habert, B., Fabre, C., and Issac, F. (1998). *DE L'ECRIT AU NUMERIQUE. Constituer, normaliser et exploiter les corpus électroniques*. Number ISBN : 2-225-82953-5. ELSEVIER MASSON.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: The 90 In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ide, N. and Macleod, C. (2001). The american national corpus: A standardized resource of american english. In *Proceedings of Corpus Linguistics 2001*, volume 3.

Ide, N., Baker, C., Fellbaum, C., Fillmore, C., and Passonneau, R. (2008). Masc: the manually annotated sub-corpus of american english. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Moro, A. and 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)*, pages 288–297, Denver, Colorado, June. Association for Computational Linguistics.

Màrquez, L., Raya, J., Carroll, J., McCarthy, D., Agirre, E., Martínez, D., Strapparava, C., and Gliozzo, A. (2002). Experiment a : Several all-words wsd systems for english. Technical report, Meaning, Developing multilingual Web-scale Language Technologies.

Navigli, R. and 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*, pages 216–225. Association for Computational Linguistics.

Navigli, R., Litkowski, K. C., and Hargraves, O. (2007). Semeval-2007 task 07: Coarse-grained english all-words task. In *SemEval-2007*, pages 30–35, Prague, Czech Republic, June.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pages 115–129.

Ng, H. T. and 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, pages 40–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Raganato, A., Camacho-Collados, J., and 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*, pages 99–110, Valencia, Spain, April. Association for Computational Linguistics.

Schwab, D., Goulian, J., and Tchechmedjiev, A. (2013). Désambiguïsation lexicale de textes : efficacité qualitative et temporelle d'un algorithme à colonies de fourmis. *TAL*, 54(1):99–138.

Schwab, D., Tchechmedjiev, A., Goulian, J., and Sérasset, G., (2015). *Language Production, Cognition, and the Lexicon*, chapter Comparisons of Relatedness Measures Through a Word Sense Disambiguation Task, pages 221–243. Springer International Publishing, Cham.

Taghipour, K. and 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*, pages 338–344, Beijing, China, July. Association for Computational Linguistics.

Vial, L., Tchechmedjiev, A., and Schwab, D. (2016). Extension lexicale de définitions grâce à des corpus annotés en sens. In *Traitement Automatique des Langues Naturelles (TALN)*.

Yuan, D., Richardson, J., Doherty, R., Evans, C., and Altendorf, E. (2016). Semi-supervised word sense disambiguation with neural models. In *COLING 2016*.

8. Language Resource References

Speecon Consortium. (2014). *Dutch Speecon Database*. Speecon Project, distributed via ELRA, Speecon resources, 1.0, ISLRN 613-489-674-355-0.