SynWSD, Unsupervised WSD with Meaning Attraction Models

Radu ION

"Mihai Drăgănescu" Research Institute for Artificial Intelligence
Romanian Academy
radu@racai.ro

Table of Contents

1 BASIC INFORMATION

1.1 Tool name

"SynWSD" for Syntactic contex representation, Word Sense Disambiguation.

1.2 Overview and purpose of the tool

SynWSD is an unsupervised, knowledge-rich Word Sense Disambiguation algorithm that uses WordNet[1] to assign senses (or sense mappings such as SUMO/MILO classes[2] and/or IRST domains[3]) to content words (nouns, verbs, adjectives and adverbs) in a text.

1.3 A short description of the algorithm

SynWSD is based on the Constrained Lexical Attraction Models analysis (Ion and Barbu Mititelu, 2006) of the input sentence which is an undirected, unlabeled dependency analysis. Briefly, when training, each word in a linked pair of the dependency analysis is expanded to all the semantic labels (senses, SUMO/MILO categories or IRST domains) it can have in WordNet, and associative frequencies of these semantic labels are kept in a database. When a new word pair is given to the algorithm, each word in the pair receives the semantic label which has the highest association score within the trained database. Further details are given by Ion and Tufiș (2007).

2 TECHNICAL INFORMATION

2.1 Software dependencies and system requirements

SynWSD is written in Perl and it is at version 3.2. Thus, the only requirement is to have a recent version of Perl installed with package BerkeleyDB from http://www.cpan.org/.

2.2 Installation

No installation is required.

2.3 Execution instructions

In order to run SynWSD on a specific XML corpus (see the next section for details on the format and annotations) one has to run the command: '`synwsd.pl synwsd.conf`' (e.g. '`synwsd.pl synwsd_sample_en.conf`') from the directory containing the '`synwsd.pl`' script. The output of the disambiguation process is written in the file specified in the configuration file.

The configuration file has all the parameters that SynWSD uses. Among the most important ones are:

- Language to use: only English ('en') or Romanian ('ro'); set parameter LANG;

[1] http://wordnet.princeton.edu/
[2] http://sigmakee.cvs.sourceforge.net/viewvc/sigmakee/KBs/WordNetMappings/
[3] http://wndomains.fbk.eu/index.html

- The input file (for both training and disambiguation) which is the value of the parameter INFILE;
- The output file (for disambiguation) which is the value for parameter OUTFILE;
- The name of the associative measure when disambiguating: probability ('prob'), pointwise mutual information ('mi'), LogLikelihood ('ll' which is the best) and DICE ('dice'); set parameter TRAINMETH ('em' is not yet used);
- The type of the semantic label to use when/for training/disambiguating: WordNet sense ('ili'), SUMO/MILO category ('sumo') and IRST domain ('dom'). Please be careful to load the appropriate model for each of these labels from '`res/models/`'! Set the parameter SENSEINVENTORY;

The other parameters are documented in the configuration files that are included in the distribution. There are two such files: '`synwsd_sample_en.conf`' and '`synwsd_sample_ro.conf`'.

2.4 Input/Output data formats

The input file to SynWSD is an XML-encoded corpus that is POS-tagged, lemmatized, chunked and linked in both English and Romanian. The format of the corpus is exemplified by the test file located in the directory '`corpora/jrccorpus-seeeranet-en-ro-sgml-links-test.xml`'.

One can obtain the required annotations by using:

- **TTL** (POS tagging, lemmatization and chunking), which is another MetaShare4U deliverable: http://ws.racai.ro:9191/repository/browse/tokenizing-tagging-lemmatizing-and-chunking-free-running-texts/e8f4fe8ed58b11e1a3cb00226410db013c013c214f9a4de0a3bd54d88cef4ca3/. It is the user's responsibility to convert from TTL's output to the XML format required by SynWSD. Alternatively, the TTL web service (located at http://ws.racai.ro/ttlws.wsdl) has a function XCES which will perform the conversion on a sentence by sentence basis.
- **LexPar** (linking), also another MetaShare deliverable that is located at the following address: http://ws.racai.ro:9191/repository/browse/lexicalized-parsing/e053d184cb3f11e1a3cb00226410db0113b631d2cbfa4802b30bb61bb70c01e6/.

SynWSD will add the '`wns`' attribute for every word (`<w>` element) in the corpus which it has attempted to disambiguate. Sample output is given in the directory '`corpora/sample-output/`'.

2.5 Integration with external tools

Other than generating the input XML corpus (by using TTL and LexPar), no other tool is required/called by SynWSD.

3 CONTENT INFORMATION

3.1 A test input file

See the file '`corpora/jrccorpus-seeeranet-en-ro-sgml-links-test.xml`'.

3.2 The output file

See the directory '`corpora/sample-output/`'.

3.3 Running time

The test file was processed in approximately 1 minute and 54 seconds on an Intel(R) Core(TM) i7 CPU 980 @ 3.33GHz. The test set contains 200 sentences. Most of time is spent in accesing the Berkeley databases generating by the training phrase.

4 ADMINISTRATIVE INFORMATION

4.1 Contact

For further details and assistance, please contact the developer: Radu ION, radu@racai.ro.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Radu Ion and Dan Tufiș. *Meaning Affinity Models*. In Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval-2007, pages 282–287, Prague, Czech Republic, June 23–24 2007. ACL 2007.

Radu Ion and Verginica Barbu Mititelu. *Constrained Lexical Attraction Models*. In Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference, pages 297–302, Menlo Park, Calif., USA, 2006. AAAI Press.