Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus

Alexandre Rademaker
(with Dan, Francis and others)

Jun, 2023

WordNet Glosses I

A synset gloss may contain a definition, one or more example sentences, or both. Glosses were introduced as redundancy to facilitate human understanding.

What is a butter?

A hyponym of solid food and dairy product.

A hypernym of lemon butter, drawn butter, stick, yak butter and beurre noisette.

But butter is "an edible emulsion of fat globules made by churning milk or cream"

WordNet Glosses II

The tennis problem, "a game played with **rackets** by two or four players who hit a **ball** back and forth over a **net** that divides the court"

WordNet Glosses III

Redundancy has its price! We usually pay missing consistency

WordNet 3.0 Pluto is "a small planet and.."

WordNet 3.1 Pluto is "a large asteroid ..."

but Tombaugh is in both releases "the astronomer who discovered the **planet** Pluto".

WordNet Glosses IV

Completeness is also relevant.

A Japanese oyster is "a large oyster native to Japan and introduced along the Pacific coast of the United States; a **candidate** for introduction in Chesapeake Bay"

A Algeripithecus minutus is "tiny (150 to 300 grams) extinct primate of 46 to 50 million years ago; fossils found in Algeria; considered by some authorities the leading **candidate** for the first anthropoid"

but what is candidate?

WordNet Glosses V

The machine learning approaches need data! We don't have many WordNet senses tagged corpora.

- SemCor 226,040 but with a lot of problems (Fellbaum talk!). 16% of the WordNet senses (here)
- OMSTI silver data, obtained from English-Chinese parallel corpus
- Senseval and SemEval tasks ($<$ 2K sentences)

See http://lcl.uniroma1.it/wsdeval/. How many Wordnet senses were annotated?

Processing the glosses I

The GlossTag project was started in Princeton, but not completed, 31% not annotated yet! **We call it GlossTag 2008**

Definitions and Examples are demarcated, tokenized and PoS tagged (only definitions). Tokens and globs (MWE).

Some spans were annotated with semantic classes: dates, time, number, currency, math expressions, etc.

Some spans are marked as auxiliary information (domain classification, verb arguments or contents that are secondary to the main sense of the synset (ignored to sense annotation).

Processing the glosses II

Data has been used by tools like UKB (graph-based word sense disambiguation library).

The eXtended WordNet from University of Texas at Dallas (website not available, based on WordNet 1.6 and 2.0). LF constructed from transformation rules applied to the syntactic analysis.

Standoff files from Priceton with logical forms from glosses. Generated by USC/ISI, California in 2006-2007. Also transformation rules, LFToolkit, applied to the output of Charniak syntactic parser.

Processing the glosses III

Our project started in 2019 (paper). The aim is to continue the annotation, fixing mistakes and adding extra layers to help on annotation. Thanks Fellbaum! **We call it GlossTag 2019**

We develop an annotation interface on top of Emacs sensetion.el

The annotation of verbs is hard, a syntactic/semantic parsing with an holistic interpretation of the sentence may help.

Processing the glosses IV

166,820	auto
664,175	ignore
334,533	man
449,967	un

We allow multiple senses whenever we can't distinguish the senses. We have 40% of the WordNet senses mentioned at least once.

MWE: $56,859$ tokens: $1,631,341$ sentences: $165,994$ (definitions: $117,658$ examples $48,336$).

Processing the glosses V

In 2023, we revised tokenization issues and the demarcation of definitions and examples. We also revised the quoted examples, moving the source of the text (author or references) to metadata.

We parsed the sentences with ERG and combine the sense annotation with the semantic representation.

We added PoS to examples. Hopefully more consistent semantic representation of texts. We call it **GlossTag 2022**.

English Resource Grammar I

After creating the profiles with 2000 sentences each, we processed them with the Ace parser in a cluster in 30 minutes. For each sentence, we asked for the top-best analysis of ERG.

From 165,976 sentences; only 5,282 (2%) were not parsed by ERG. Using some heuristics (e.g. 'get the votes of **X**'), 600 more sentences.

Preliminar evaluation gives us F1 80% for the first analysis be the expected one, future work aim to manually treebanking all sentences using FFTB tool.

See also here.

Examples I

- of a leaf shape; palmately cleft rather than lobed (here)
- raw milk that has soured and thickened (here)

Synset about verb with adjective example.

MWE: ADJ-NOUN, NOUN-NOUN, compounds, coordinations, verb phrases, light verbs, idioms, etc. Mismatch between ERG and WordNet.

Speeding up the annotations I

Manual word sense disambiguation (WSD) is an arduous task. One non-native speaker annotator doing it manually from the last 4 years (not full-time).

Many techniques for automatic WSD are being investigated: graph-based (or knowledge-based), supervised and unsupervised machine learning methods.

Automatic annotation would allow us to provide intermediary releases of the data (silver versions).

round trip... GlossTag 2008 was already used by UKB tool (graph-based WSD) and to training supervised WSD algorithms replacing the SemCor.

Speeding up the annotations II

We used UKB, data was transformed into UKB input for: (1) evaluation UKB performance; (2) complete annotation. From palmatifid

```
# text = of a leaf shape; palmately cleft rather than lobed
# id = 02173264-a
# type = def
1       wf      ignore  0:2     IN      of      of      _
2       wf      ignore  3:4     DT      a       a       _
3       glob|a  auto    _       _       _       leaf_shape%1    leaf_shape%1:25:00::
4       cf|a    un      5:9     NN      leaf    leaf%1|leaf%2   _
5       cf|a    un      10:15   NN      shape   shape%1|shape%2 _
6       wf      ignore  15:16   :       ;       _       _
7       wf      auto    17:26   RB      palmately       palmately%4     palmately%4:02:00::
8       wf      man     27:32   VBN     cleft   cleft%1|cleave%2|cleft%3        cleft%5:00:00:compound:00
9       wf      un      33:39   RB      rather  rather%4        _
10      wf      ignore  40:44   IN      than    than    _
11      wf      man     45:50   JJ      lobed   lob%2|lobed%3   lob%2:35:00::
```

```
ctx-02173264-a/a
leaf_shape#n#w4#1#1 palmately#r#w7#1#1 cleave#v#w8#1#1 rather#r#w9#1#1 lob#v#w11#1#1 02173264-a#a#fake1#2#1

ctx-02173264-a/b
leaf_shape#n#w4#1#1 palmately#r#w7#1#1 cleave#v#w8#1#1 rather#r#w9#1#1 lob#v#w11#1#1
```

We try with and without extra context. We compare results for annotated words with the annotations.

Speeding up the annotations III

Results

	Total	# (a)	# (b)	% (a)	% (b)
All	442782	413546	374648	93.39	84.61
Noun	329692	308245	287033	93.49	87.06
Adj	64298	60591	52008	94.23	80.89
Verb	41520	37832	29529	91.11	71.12
Adv	7272	6878	6078	94.58	83.58

For automatically complete annotation, words already annotated would increase UKB performance.

Final Thoughts I

The project is hosted in the https://github.com/own-pt/glosstag. We want to improvement the web interface http://openwordnet-pt.org to show annotations and links.

We need to make code available to the reproducibility of the experiments presented here.

We need to improve our annotation tool, fixed dependencies. Emacs? Web?

We need to finish the annotation and treebank the ERG analyses. Define a proper workflow to combine the two layers of annotation.

We plan to experiment with alternative WSD methods.

Final Thoughts II

This work is part of our effort in expanding and improving WordNet-like resources in an application-driven and domain-specific way.

Finally, we need to finish the migration to WordNet 3.1 before forking it from the Princeton official release (or further mapping to Open English Wordnet, http://en-word.net) for changes driven by the annotation.

How to release the data? Users? WSD tools? Wordnet itself?

How this data can help Dan's ERG lexicon expansion?

Thank You !