TeamUFAL: WSD+EL as Document Retrieval*

Petr Fanta, Roman Sudarikov, Ondřej Bojar
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 11800 Praha 1, Czech Republic
p.fanta@seznam.cz, {sudarikov, bojar}@ufal.mff.cuni.cz

Abstract

This paper describes our system for SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. We have participated with our system in the sub-task which aims at monolingual all-words disambiguation and entity linking. Aside from system description, we pay closer attention to the evaluation of system outputs.

1 Introduction

Word sense disambiguation (WSD, i.e. picking the right sense for a given word from a fixed inventory) and entity linking (EL, i.e. identifying a particular named entity listed in a database given its mention in a text) are among the fashionable tasks in computational linguistics and natural language processing these days. WSD has been, after some debate, shown to help machine translation (Carpuat and Wu, 2007), other applications include knowledge discovery or machine reading in general (Etzioni et al., 2006; Schubert, 2006). WSD and EL are usually applied with large and rich context available (Navigli, 2009), but the arguably harder setting of short context has a wider range of applications, including text similarity measurements (Abdalgader and Skabar, 2011), Named Entities Extraction and Named Entities Disambiguation (Habib and Keulen, 2012) or handling data from social networks, such as attempts to translate tweets (Šubert and Bojar, 2014).

Our attempt at WSD and EL can be classified as unsupervised, corpus-based and our implementation relies on an information retrieval tool. We do not take longer context into account.

2 Task Description

As participants of SemEval-2015 Task 13 (Moro and Navigli, 2015), we were given only a very brief instructions, effectively just one example of a POS-tagged sentence:

The/X European/J/european Medicines/N/medicine Agency/N/agency (/X EMA/N/ema)/X is/V/be ,...

We were expected to provide such input with labels indicating that e.g. the words "European Medicines Agency" refer to the entity described in the English Wikipedia under the title European_Medicines_Agency ("`wiki:European_Medicines_Agency`"), the word "Medicines" refers to the BabelNet concept 00054128n etc. The repertoire of word sense and entities came from BabelNet 2.5 which included: Wikipedia page titles (2012/10 dump), WordNet 3.0 synsets, OmegaWiki senses (2013/09 dump) and Open Multilingual WordNet synsets (2013/08 dump). The output format accepted Wikipedia titles, BabelNet IDs and Wordnet sense keys.

The test set for SemEval-2015 task 13 was released for three languages: English, Italian and Spanish. We joined only the English task. All the data was gathered from 3 domains: biomedicine, mathematics and computers and general domain.

At the time of the shared task, neither a scoring

This research was supported by the grants FP7-ICT-2013-10-610516 (QTLeap). This research was partially supported by SVV project number 260 224. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

script, nor any development set with annotations was provided. It was also not very clear how the different allowed ID sources (Wikipedia, BabelNet and Wordnet) will be used concurrently.

The official scoring script and golden annotation was provided later, and we use it here to report the scores of our submission and a few variations of it.

3 Our System

Our system is unsupervised and relies on an information retrieval (IR) tool applied to a large collection of documents. We thus call it corpus-based.

Given an input sentence, we remove all stopwords (as defined by the IR tool) and punctuation, putting together even words which were originally not adjacent. For each span up to a given length in this abridged sentence, the system tries to find a document in the database. If found, this document implies the sense or entity ID for the given span.

The words in the span and (separately) other words in the sentence are used to construct the query for the IR engine. We construct multiple queries and merge their results in a candidate selection process, possibly returning no document at all.

Due to the different nature of our sources (see Section 3.1), we run sub-systems with different configurations for each of them. We return the union of responses from these sub-systems.

3.1 Data Sources

BabelNet alone is not a good resource for our approach, because it does not include textual data. We resort to the original sources of BabelNet and map them back to BabelNet. Our sources are thus the English Wikipedia, English Wiktionary and WordNet. Short of the original versions as used in BabelNet 2.5, we used Wiki dumps from November 2014 and WordNet 3.0, facing some ID mismatches.

3.2 Indexing

For each source, we create a full text index using Apache Lucene search engine which provides several ranking models. We experiment with models based on TF-IDF (Salton et al., 1975) and Okapi BM25 (Robertson et al., 1995), selecting the better one for each subsystem in our submission.

All indexes have a similar structure, they contain:

Score	Document ID (ie. Wikipedia Title)
8.201	Medical_condition → Disease
8.201	Medical_conditions → Disease
6.561	Frostbite_(medical_condition) → Frostbite

Figure 1: Query results (→ means redirection closure)

ID of the element in the given source (Wikipedia ID, Wiktionary ID or Wordnet sense key),

Title of Wikipedia or Wiktionary article or word from WordNet,

Body text of articles from Wiki sources (markup removed) or all textual data from Wordnet synset (including other words in the synset),

POS tag (only in WordNet index).

The Title and the Body field are stemmed by Porter stemmer implemented in Lucene.

3.3 Proposing Candidates

We use different sets of queries for each source. We query Wiktionary and Wordnet for single-word spans only, while Wikipedia seems suitable for both, single and multi-word spans.

The queries typically require all the words from the span to appear in the Title field of the document and the words from the context to appear in the Body field of the document. A number of slightly different queries, incl. queries that use n-grams of words or some boosting for some of the terms, is run in parallel, giving us multiple lists scored by the selected IR model (see Section 3.2). The results for a simple query *+TITLE:medical +TITLE:condition* for Wikipedia documents are shown in Figure 1.

3.4 Final Candidate Selection

Final candidates are picked from the results of the queries. Before this selection, the results for each span are grouped and scores for the same ID (coming from different lists or redirection) are summed.

For Wikipedia, we select the highest-scoring candidate and it is returned only if its score is greater than double the score of the second candidate. After this selection, the system checks if there are overlapping spans labeled with same ID and returns only the span with the best score.

For Wordnet and Wiktionary, we simply return the highest-scoring candidate for each span. Since Wiktionary IDs are not expected in the shared task,

	Official			Offic+penalty	Our Exact			Our Partial			Bag of IDs		
System	P	R	F1	P	P	R	F1	P	R	F1	P	R	F1
Submitted	40.4	36.5	38.3	30.4	25.9	48.2	33.7	26.6	49.4	34.6	24.0	50.5	32.5
Submitted-fix	41.2	37.3	39.1	30.7	25.7	49.6	33.9	26.3	50.8	34.7	23.4	52.0	32.3
DFKI	67.4	52.6	59.1	55.2	51.5	49.2	50.3	52.1	49.8	50.9	51.3	49.2	50.2
EBL-Hope	48.4	44.4	46.3	40.4	36.8	40.4	38.5	37.1	40.8	38.9	37.3	41.0	39.0
el92-Run1	69.9	21.4	32.8	62.6	59.9	20.4	30.5	61.2	20.9	31.1	62.2	21.2	31.7
el92-Run2	71.9	19.1	30.2	64.8	61.8	18.2	28.2	62.5	18.4	28.5	62.9	18.5	28.6
el92-Run3	**75.2**	18.5	29.6	**69.6**	**66.0**	17.5	27.7	**66.8**	17.7	28.0	**66.9**	17.8	28.1
LIMSI	68.7	63.1	65.8	57.3	55.4	60.9	58.0	55.6	61.2	58.3	55.6	61.2	58.2
SUDOKU-Run1	60.1	52.1	55.8	50.3	47.0	48.6	47.8	47.2	48.8	48.0	47.0	48.6	47.8
SUDOKU-Run2	62.9	60.4	61.6	53.0	49.2	56.1	52.4	49.7	56.6	52.9	49.3	56.2	52.5
SUDOKU-Run3	61.9	59.4	60.6	52.2	48.6	55.4	51.8	49.0	55.8	52.1	48.7	55.5	51.9
UNIBA-Run1	66.2	52.3	58.4	54.3	51.6	49.8	50.7	51.9	50.0	50.9	51.9	50.0	50.9
UNIBA-Run2	66.1	52.1	58.3	53.5	50.9	49.6	50.2	51.5	50.2	50.8	51.4	50.1	50.7
UNIBA-Run3	66.1	52.1	58.3	53.0	50.5	49.7	50.1	51.3	50.4	50.8	51.1	50.2	50.7
vua-background	67.5	51.4	58.4	56.3	52.1	47.5	49.7	52.3	47.8	50.0	52.3	47.7	49.9
WSD-games-Run1	57.4	48.8	52.8	47.9	44.1	45.0	44.6	44.3	45.2	44.7	44.3	45.2	44.7
WSD-games-Run2	58.8	50.0	54.0	49.0	45.3	46.2	45.7	45.5	46.4	45.9	45.5	46.4	45.9
WSD-games-Run3	53.5	45.4	49.1	44.6	40.7	41.5	41.1	41.0	41.8	41.4	41.0	41.8	41.4
MFS	67.9	**67.1**	**67.5**	67.9	65.2	**64.5**	**64.9**	65.5	**64.8**	**65.2**	65.2	**64.5**	**64.9**

Table 1: All submissions evaluated on all domains using various official and our scorings.

we map them to BabelNet IDs prior to picking the highest-scoring one. (Wiktionary IDs that cannot be mapped are discarded.)

4 Evaluation

Having thoroughly reviewed the official scoring script, we find some of its features unusual:

- The precision of a system is not penalized for spans, which don't occur in the golden set.
- The recall should consider only to what extent the expected answers are covered by the system's answers. The official scoring script reduces the recall score for any 'unexpected' answers.
- An exact match in span is needed to give any credit to the system answer.

We thus propose a slightly different evaluation procedure and apply it to all submitted systems.

4.1 Our proposed scoring

Our scoring is based on a credit for partially overlapping spans, similarly to Cornolti et al. (2013), who however disregard the overlap size. We call a 'label' $l = (l_1, l_2)$ the pair of a span (a range of words in the sentence; denoted l_1) and an ID attached to the span, l_2. For a label s in the system output and a label g in the golden annotation, we define their match as:

$$\text{match}(s, g) = \begin{cases} \frac{|g_1 \cap s_1|}{|g_1 \cup s_1|} & \text{if } |g_1 \cap s_1| > 0 \wedge g_2 = s_2 \\ 0 & \text{otherwise} \end{cases}$$

In other words overlapping spans labeled with the same ID get a credit proportional to the size of the overlap. We define precision and recall as follows:

$$\text{precision} = \frac{\sum_{s \in S, g \in G} \text{match}(s, g)}{|S|}$$

$$\text{recall} = \frac{\sum_{s \in S, g \in G} \text{match}(s, g)}{|G|}$$

where G and S are sets of labels from the gold standard and a system output, respectively. Our approach gives a partial credit for inexact, but overlapping, spans with correct identifiers.

Our precision and recall are only meaningful, if all IDs come from a single source. We pick BabelNet IDs for this purpose and map all system outputs as necessary. Note that the mapping from the Wikipedia IDs to the BabelNet IDs is ambiguous but not in more than 1 ‰ cases.

WSD-games and vua-background report only WordNet sense keys. We map them unambiguously to BabelNet IDs.

el92 produces lowercase Wikipedia titles so the ambiguous mapping to BabelNet IDs is slightly worse.

DFKI and our system produce both Wikipedia titles and WordNet IDs, we map both as above and union the results.

SUDOKU produces BabelNet IDs but some spans have no ID at all. We ignore these spans.

	Fix Wikt→BN mapping	Model for Wikipedia	Model for Wiktionary	Use context in Wikt. search	Precision	Recall	F1
Submitted	-	TF-IDF	BM25	no	40.4%	36.5%	36.5%
Submitted_fix	yes	TF-IDF	BM25	no	41.2%	37.3%	39.1%
Wiki_BM25	no	BM25	BM25	no	38.4%	35.0%	36.7%
Wikt+context_BM25	no	TF-IDF	BM25	yes	38.4%	35.3%	36.8%
Wikt+context_TF-IDF	no	TF-IDF	TF-IDF	yes	40.3%	37.0%	38.6%

Table 2: Our system outputs

4.2 Results

Table 1 reports systems' scores using these evaluation metrics:

Official Precision and recall as reported by the official scoring script.

Official+penalty A modified version of the official scoring script which treats spans in system output and no counterpart in the golden set in the same way as if the golden set assigned a different ID to the span.

Our Exact Our method (Section 4.1), but rounding the 'match' down to zero, so only exactly matching spans get the credit (of 1).

Our Partial Our method (Section 4.1).

Bag of IDs disregards spans altogether, checking just the match of the BabelNet IDs needed and produced. Precision is the fraction of correct (confirmed by the golden data) IDs among all labels produced by the system. Recall is the number of correct IDs divided by the number of labels in the gold set. This scoring gives an idea of how well the system guesses the "meaning" (bag of concepts) of the whole sentence.

The Table 1 documents that the official scoring heavily boosted our precision and hurt our recall. The performances of other systems are affected as well, but fortunately, the overall impression is similar across the scoring techniques.

4.3 Variants of our submission

As the official scores in the overview paper (Moro and Navigli, 2015) show, our system performed acceptably on Named Entities Recognition task, but it clearly failed on word senses disambiguation.

Table 2 reports the scores (official scoring) of a few variations of our approach. The first row is the submitted system, the second row is a correction which allows Wiktionary results to map to BabelNet senses of all parts of speech, not just nouns. The remaining rows use a different IR model or include sentence context in Wiktionary search but no improvement is obtained.

4.4 Recommendations for future evaluation

For future shared tasks, we recommend:

- Define precision and recall to better match the common meaning, e.g. as in our proposal.
- Preserve letter case in IDs to avoid ambiguity in Wikipedia to BabelNet mapping.
- Use only one repertoire of IDs in the gold set.

4.5 Future work

In future we want to evaluate other heuristics such as weighted words picking instead of first one, offered by search algorithms. Also we'll examine possibilities to enhance Wordnet and Wiktionary records to make search results more reliable. Another way of improvement is using Named Entities Recognition systems to define correct span boundaries and to achieve better results for Named Entities.

5 Conclusion

We described our system for SemEval Task 13 based on information retrieval. The system performs acceptably in Named Entity Linking (NEL) but fails in Word Sense Disambiguation. One of the reasons is that we used small information records for Wiktionary and especially for Wordnet and little or no sentence context in WSD queries, so the information retrieval algorithms performed poorly.

Additionally, we proposed different scoring techniques that, in our opinion, better reflect the performance of the systems. Fortunately, the overall ranking of systems ends up similar to the official scoring. We nevertheless recommend a few changes for future shared tasks.

References

Khaled Abdalgader and Andrew Skabar. Short-text similarity measurement using word sense disambiguation and synonym expansion. In *AI 2010: Advances in Artificial Intelligence*, pages 435–444. Springer, 2011.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *EMNLP-CoNLL*, volume 7, pages 61–72. Citeseer, 2007.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the International World Wide Web Conference (WWW) (Practice & Experience Track)*, 2013.

Oren Etzioni, Michele Banko, and Michael J Cafarella. Machine reading. In *AAAI*, volume 6, pages 1517–1519, 2006.

Mena B Habib and Maurice Keulen. Unsupervised improvement of named entity extraction in short informal context using disambiguation clues. 2012.

Andrea Moro and Roberto Navigli. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proc. of SemEval-2015*, 2015. In press (in this volume).

Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Lenhart Schubert. Turing's dream and the knowledge challenge. In *Proceedings of the national conference on artificial intelligence*, volume 21, page 1534. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Eduard Šubert and Ondřej Bojar. Twitter crowd translation – design and objectives. 2014.