DFKI: Multi-objective Optimization for the Joint Disambiguation of Entities and Nouns & Deep Verb Sense Disambiguation

Dirk Weissenborn
LT, DFKI
Alt-Moabit 91c
Berlin, Germany
dirk.weissenborn@dfki.de

Feiyu Xu
LT, DFKI
Alt-Moabit 91c
Berlin, Germany
feiyu@dfki.de

Hans Uszkoreit
LT, DFKI
Alt-Moabit 91c
Berlin, Germany
uszkoreit@dfki.de

Abstract

We introduce an approach to word sense disambiguation and entity linking that combines a set of complementary objectives in an extensible multi-objective formalism. During disambiguation the system performs continuous optimization to find optimal probability distributions over candidate senses. Verb senses are disambiguated using a separate neural network model. Our results on noun and verb sense disambiguation as well as entity linking outperform all other submissions on the SemEval 2015 Task 13 for English.

1 Introduction

The task of assigning the correct meaning to a given word or entity mention in a document is called word sense disambiguation (WSD) (Navigli, 2009) or entity linking (EL) (Bunescu and Pasca, 2006), respectively. Successful disambiguation requires not only an understanding of the topic or domain a document is dealing with (global), but also an analysis of how an individual word is used within its local context. E.g., the meanings of the word "newspaper" as the company or the physical product, often cannot be distinguished by the topic, but by recognizing which type of meaning fits best into the local context of its occurrence. On the other hand, for an ambiguous entity mention such as "Michael Jordan" it is important to recognize the topic of the wider context to distinguish, e.g., between the basketball player and the machine learning expert.

The combination of the two most commonly used reference knowledge bases for WSD and EL, e.g., WordNet (Fellbaum, 1998) and Wikipedia, by BabelNet (Navigli and Ponzetto, 2012) has enabled a new line of research towards the joint disambiguation of words and named entities. *Babelfy* (Moro et al., 2014) has shown the potential of combining these two tasks in a purely knowledge-driven approach that jointly finds connections between potential word senses in the global context. On the other hand, typical supervised methods (Zhong and Ng, 2010) trained on sense-annotated corpora are usually quite successful in dealing with individual words in a local context. Hoffart et al. (2011) recognize the importance of combining both local context and global context for robust disambiguation. However, their approach is limited to EL, where optimization is performed in a discrete setting.

We present a system that combines disambiguation objectives for both global and local contexts into a single multi-objective function. In contrast to prior work we model the problem in a continuous setting based on probability distributions over candidate meanings. Our approach exploits lexical and encyclopedic knowledge, local context information and statistics of the mapping from text to candidate meanings. Furthermore, we introduce a deep learning approach to verb sense disambiguation based on semantic role labeling.

2 Approach

The SemEval-2015 task 13 (Moro and Navigli, 2015) requires a system to jointly detect and disambiguate word and entity mentions given a reference knowledge base. The provided input to the system are tokenized, lemmatized and POS-tagged doc-

uments; the output are sense-annotated mentions.

Our system employs BabelNet 1.1.1 as reference knowledge base (KB). BabelNet is a multilingual semantic graph of concepts and named entities that are represented by synonym sets, called *Babel synsets*.

2.1 Mention Extraction & Entity Detection

We define a mention to be a sequence of tokens in a given document for which there exists at least one candidate meaning in the KB. The system considers all content words (nouns, verbs, adjectives, adverbs) as mentions including also multi-token words of up to 5 tokens that contain at least one noun. In addition, we apply a pre-trained stacked linear-chain CRF (Lafferty et al., 2001) using the FACTORIE toolkit of version 1.1 (McCallum et al., 2009) to identify named entity (NE) mentions. In our approach, we distinguish NEs from common nouns and treat them as two different classes because there are many common nouns also referring to NEs making disambiguation unnecessarily complicated.

2.2 Candidate Search

After potential mentions are extracted the system tries to identify their candidate meanings, i.e., the appropriate synsets. Mentions without such candidates are discarded. The mapping of candidate mentions to synsets is based on similarities of their surface strings or lemmas. If the surface string or lemma of a mention matches the lemma of a synonym in a synset that has the same part of speech, the synset will be considered a candidate meaning. We allow partial matches for BabelNet synonyms derived from Wikipedia titles or redirections. A partial match allows the surface string of a mention to differ by up to two tokens from the Wikipedia title (excluding everything in parentheses) if the partial string was used at least once as an anchor for the corresponding Wikipedia page. For example, for the Wikipedia title Armstrong_School_District_(Pennsylvania), the following surface strings would be considered matches: "Armstrong School District (Pennsylvania)", "Armstrong School District", "Armstrong", but not "School", since "School" was never used as an anchor. If there is no match we try the same procedure applied to the lowercased text or lemma.

Because of the distinction between nouns and named entities we treat NE as a separate POS tag. Candidate synsets for NEs are Babel synsets considered NEs in BabelNet, and additionally Babel synsets of all Wikipedia senses that are not considered NEs. Similarly, candidate synsets for nouns are noun synsets that are not considered NEs in addition to all synsets of WordNet senses in BabelNet. We add synsets of Wikipedia senses and WordNet senses, respectively, because the distinction of NEs and simple concepts is not always clear in BabelNet. For example the synset for "UN" (United Nations) is considered a concept whereas it could also be considered a NE. Finally, if there is no candidate for a potential noun mention we try to find NE candidates for it and vice versa.

2.3 Disambiguation of Nouns and Named Entities

We formulate the disambiguation problem in a continuous setting by using probability distributions over candidates. This has several advantages over a discrete setting. First, we can exploit well established continuous optimization algorithms, such as conjugate gradient or LBFGS, which guarantee to converge to a local optimum. Second, by optimizing upon probability distributions we are optimizing the actually desired result in contrast to densest subgraph algorithms where such probabilities need to be calculated artificially afterwards, e.g., Moro et al. (2014). Third, discrete optimization usually works on a single candidate per iteration whereas in a continuous setting, probabilities are adjusted for each candidate, which is computationally advantageous for highly ambiguous documents.

Given a set of objectives $\mathfrak{O}$ the overall objective function $\mathbf{O}$ is defined as the sum of all normalized objectives $O \in \mathfrak{O}$ given a set of mentions M:

$$\mathbf{O}(M) = \sum_{O \in \mathfrak{O}} \frac{O(M)}{O_{max}(M) - O_{min}(M)}. \quad (1)$$

We normalize each objective using the difference of their maximum and minimum value for the given document. For disambiguation we optimize the multi-objective function using Conjugate Gradient (Hestenes and Stiefel, 1952) with up to 1000 iterations per document.

Coherence Jointly disambiguating all mentions within a document has been shown to have a large impact on disambiguation quality. We adopt the idea of semantic signatures and the idea of maximizing the semantic agreement among selected candidate senses from Moro et al. (2014). We define the continuous objective function based on probability distributions $p_m(c)$ over the candidate set C_m of each mention $m \in M$ in a document as follows:

$$O_{\text{coh}}(M) = \sum_{\substack{m \in M \\ c \in C_m}} \sum_{\substack{m' \in M \\ m' \neq m \\ c' \in C_{m'}}} s(m, c, m', c')$$

$$s(m, c, m', c') = p_m(c) \cdot p_{m'}(c') \cdot \mathbb{1}((c, c') \in S)$$

$$p_m(c) = \frac{e^{\lambda_{m,c}}}{\sum_{c' \in C_m} e^{\lambda_{m,c'}}} \quad , \tag{2}$$

where S denotes the semantic interpretation graph, $\mathbb{1}$ the indicator function and $p_m(c)$ is a softmax function. The only free, optimizable parameters are the softmax weights $\lambda_{m,c}$. This objective can be interpreted as finding the densest subgraph of the semantic interpretation graph where each node is weighted by its probability and therefore each edge is weighted by the product of its adjacent vertex probabilities.

Type Classification One of the biggest problems of supervised approaches to WSD is the size and synset coverage of training corpora such as SemCor (Miller et al., 1993). One way to circumvent this problem is to use a coarser set of semantic classes that groups synsets together. Previous studies on using semantic classes for disambiguation showed promising results (Izquierdo-Beviá et al., 2006). WordNet provides a mapping, called lexnames, of synsets into 45 types based on the syntactic categories of synsets and their logical groupings[1].

A multi-class logistic (softmax) regression model was trained that calculates a probability distribution $q_m(t)$ over lexnames t given a potential WordNet mention m. The features used as input to the model are the following: embedding of the mention's text, sum of embeddings of all sentence words, embedding of the dependency parse parent, collocations of surrounding words (Zhong and Ng, 2010), surrounding POS tags and possible lexnames. We used pre-trained embeddings from Mikolov et al. (2013).

Type classification is included in the overall objective in the following form:

$$O_{\text{typ}}(M) = \sum_{\substack{m \in M \\ c \in C_m}} q_m(t_c) \cdot p_m(c) \tag{3}$$

Priors Another advantage of working with probability distributions over candidates is the easy integration of prior information. E.g., the word "Paris" without further context has a strong prior on its meaning as a city instead of a person. Our approach utilizes prior information in form of frequency statistics over candidate synsets for a mention's surface string. These priors are derived from annotation frequencies provided by WordNet for Babelsynsets containing the respective WordNet sense and from occurrence frequencies in Wikipedia extracted by DBpedia Spotlight(Daiber et al., 2013) for synsets containing only Wikipedia senses. Laplace-smoothing is applied to all prior frequencies. This prior is used to initialize the probability distribution over candidate synsets. Note that the priors are used "naturally", i.e., as actual priors and not during context based optimization itself.

Furthermore, because candidate priors for NE mentions can be very high we add an additional L2-regularization objective for NE mentions with $\lambda = 0.001$, which we found to work best on development data. Finally, named entities were filtered out if they were included in another NE, had no connection in the semantic interpretation graph with another candidate sense of the input document or were overlapping with another NE but were connected worse.

2.4 Disambiguation of Verbs

The disambiguation of verbs requires an approach that focuses more on the local context and especially the usage of a verb within a sentence. Therefore, we train a neural network based on semantic role labeling (SRL) and sentence words. Figure 1 illustrates an example network. The input is composed of the word embeddings (Turian et al., 2010) for each feature (word itself, its lemma, SRLs and bag of sentence words). All individual input embeddings are

[1] http://wordnet.princeton.edu/man/lexnames.5WN.html

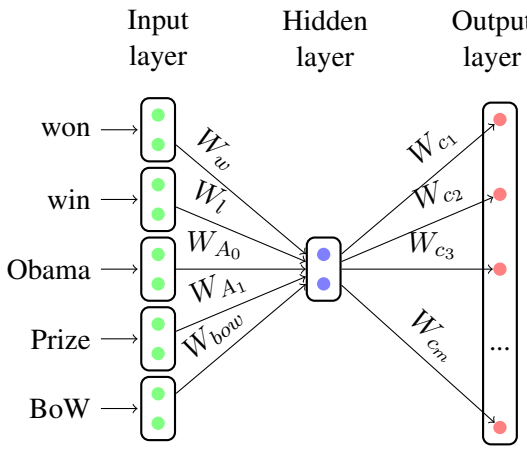

Figure 1: Disambiguation neural network for "won" in the sentence *"Obama won the Nobel Prize."*

50-dimensional and connected to a 100-dimensional hidden layer. The output layer consists of all candidate synsets of the verb. The individual output weights W_c are candidate specific. To ensure better generalization and to deal with the sparseness of training corpora, W_c is defined as the following sum:

$$W_c = W_{s(c)} + \sum_{s_p \in P_{s(c)}} W_{s_p} + \sum_{s_e \in E_{s(c)}} W_{s_e}, \quad (4)$$

where $s(c)$ is the respective synset of c, P_s is the set of all *hypernyms* of s (transitive closure) and E_s are the synsets *entailed* by s. We used ClearNLP[2](Choi, 2012) for extracting SRLs.

3 Results

The results of our system are shown in Table 1. Our approaches to the disambiguation of English nouns, named entities and verbs generally outperformed all other submissions across different domains as well as the strong baseline provided by the most-frequent-sense (MFS). This demonstrates the system's capability to adapt to different domains. However, results on the *math and computer* domain also reveal that performance strongly depends on the document topic. The results for this domain are worse compared to the other domains for almost all participating systems, which may indicate that existing resources do not cover this domain as well as the others. Another potential explanation is that enforcing only pairwise coherence does not take the hidden topics *computer* and *maths* into account that connect all concepts in the specific document. This might be an interesting point for further research.

[2]http://clearnlp.wikispaces.com

	bio	math	gen	all
MFS	75.3	43.6	69.2	66.7
best other	76.5	**51.4**	63.7	64.8
DFKI	**79.1**	44.9	**73.4**	**70.3**

(a) Nouns

	bio	math	gen	all
MFS	98.9	57.1	77.4	85.7
best other	98.9	**74.3**	89.7	87.0
DFKI	**100.0**	57.1	**90.3**	**88.9**

(b) Named Entities

	bio	math	gen	all
MFS	52.5	55.7	61.4	55.1
best other	53.8	**60.6**	**70.6**	57.1
DFKI	**58.3**	52.3	66.7	**57.7**

(c) Verb

Table 1: F1 scores of our system, the best other system and an MFS baseline on the disambiguation of English nouns, named entities and verbs for all domains of the SemEval 2015 task 13. *bio- biomedical; math- math & computer; gen- general*

4 Conclusion

We have presented a robust approach for disambiguating nouns and named entities as well as a neural network for verb sense disambiguation that we used in the SemEval 2015 task 13. Our system achieved an overall F1 score of 70.3 for nouns, 88.9 for NEs and 57.7 for verbs across different domains, outperforming all other submissions for these categories of English. The disambiguation of nouns and named entities performs especially well compared to other systems and can still be extended through the introduction of additional, complementary objectives. Disambiguating verbs remains a very challenging task and the promising results of our model still leave much room for improvement.

Acknowledgment

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) through the projects Deependance (01IW11003), ALL SIDES (01IW14002) and

BBDC (01IS14013E) and by Google through a Focused Research Award granted in July 2013.

References

[Bunescu and Pasca2006] Razvan C Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16.

[Choi2012] Jinho D Choi. 2012. Optimization of natural language processing components for robustness and scalability.

[Daiber et al.2013] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction.

[Fellbaum1998] Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

[Hestenes and Stiefel1952] Magnus Rudolph Hestenes and Eduard Stiefel. 1952. *Methods of conjugate gradients for solving linear systems*, volume 49. National Bureau of Standards Washington, DC.

[Hoffart et al.2011] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

[Izquierdo-Beviá et al.2006] Rubén Izquierdo-Beviá, Lorenza Moreno-Monteagudo, Borja Navarro, and Armando Suárez. 2006. Spanish all-words semantic class disambiguation using cast3lb corpus. In *MICAI 2006: Advances in Artificial Intelligence*, pages 879–888. Springer.

[Lafferty et al.2001] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[McCallum et al.2009] Andrew McCallum, Karl Schultz, and Sameer Singh. 2009. FACTORIE: Probabilistic programming via imperatively defined factor graphs. In *Neural Information Processing Systems (NIPS)*.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

[Miller et al.1993] George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

[Moro and Navigli2015] Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proc. of SemEval-2015*.

[Moro et al.2014] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2.

[Navigli and Ponzetto2012] Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

[Navigli2009] Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

[Turian et al.2010] Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

[Zhong and Ng2010] Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83. Association for Computational Linguistics.