Multi-Objective Optimization for the Joint Disambiguation of Nouns and Named Entities

Dirk Weissenborn, Leonhard Hennig, Feiyu Xu and Hans Uszkoreit
Language Technology Lab, DFKI
Alt-Moabit 91c
Berlin, Germany
{dirk.weissenborn, leonhard.hennig, feiyu, uszkoreit}@dfki.de

Abstract

In this paper, we present a novel approach to joint word sense disambiguation (WSD) and entity linking (EL) 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. The performance of our system on nominal WSD as well as EL improves state-of-the-art results on several corpora. These improvements demonstrate the importance of combining complementary objectives in a joint model for robust disambiguation.

1 Introduction

The task of automatically 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, but also a deep analysis of how an individual word is used within its local context. For example, the meanings of the word "newspaper", as in the company or the physical product, often cannot be distinguished by the global topic of the document it was mentioned in, but by recognizing which type of meaning fits best into the local context of its mention. On the other hand, for an ambiguous entity mention such as a person name, e.g., "Michael Jordan", it is important to recognize the domain or 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 employed reference knowledge bases for WSD and EL, WordNet (Fellbaum, 1998) and Wikipedia, in 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 on a global, document level. On the other hand, typical supervised methods (Zhong and Ng, 2010) trained on sense-annotated datasets are usually quite successful in dealing with individual words in their local context on a sentence level. Hoffart et al. (2011) recognize the importance of combining both local and global context for robust disambiguation. However, their approach is limited to EL and 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. The resulting system is flexible and easily extensible with complementary objectives. In contrast to prior work (Hoffart et al., 2011; Moro et al., 2014) we model the problem in a continuous setting based on probability distributions over candidate meanings instead of a binary treatment of candidate meanings during disambiguation. Our approach combines knowledge from various sources in one robust model. The system uses lexical and encyclopedic knowledge for the joint disambiguation of words and named entities, and exploits local context information of a mention to infer the type of its meaning. We integrate prior statistics from surface strings to candidate meanings in a "natural" way as starting probability distributions for each mention.

The contributions of our work are the following:

- a model for joint nominal WSD and EL that outperforms previous state-of-the-art systems on both tasks
- an extensible framework for multi-objective

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 596–605,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

disambiguation

- an extensive evaluation of the approach on multiple standard WSD and EL datasets
- the first work that employs continuous optimization techniques for disambiguation (to our knowledge)
- publicly available code, resources and models at `https://bitbucket.org/dfki-lt-re-group/mood`

2 Approach

Our system detects mentions in texts and disambiguates their meaning to one of the candidate senses extracted from a reference knowledge base. The integral parts of the system, namely *mention detection*, *candidate search* and *disambiguation* are described in detail in this section. The model requires a tokenized, lemmatized and POS-tagged document as input; the output are sense-annotated mentions.

2.1 Knowledge Source

We employ BabelNet 2.5.1 as our reference knowledge base (KB). BabelNet is a multilingual semantic graph of concepts and named entities that are represented by synonym sets, called *Babel synsets*. It is composed of lexical and encyclopedic resources, such as WordNet and Wikipedia. Babel synsets comprise several Babel senses, each of which corresponds to a sense in another knowledge base. For example the Babel synset of "Neil Armstrong" contains multiple senses including for example "armstrong#n#1" (WordNet), "Neil_Armstrong" (Wikipedia). All synsets are interlinked by conceptual-semantic and lexical relations from WordNet and semantic relations extracted from links between Wikipedia pages.

2.2 Mention Extraction & Entity Detection

We define a mention to be a sequence of tokens in a given document. The system extracts mentions for all content words (nouns, verbs, adjectives, adverbs) and multi-token units of up to 7 tokens that contain at least one noun. In addition, we apply a NER-tagger to identify named entity (NE) mentions. Our approach distinguishes NEs from common nouns because there are many common nouns also referring to NEs, making disambiguation unnecessarily complicated. For example, the word "moon" might refer to songs, films, video games, etc., but we should only consider these meanings if the occurrence suggests that it is used as a NE.

2.3 Candidate Search

After potential mentions are extracted, the system tries to identify their candidate meanings, i.e., the appropriate synsets. Mentions without any candidates are discarded. There are various resources one can exploit to map surface strings to candidate meanings. However, existing methods or resources especially for NEs are either missing many important mappings[1] or contain many noisy mappings[2]. Therefore, we created a candidate mapping strategy that tries to avoid noisy mappings while including all potentially correct candidates. Our approach employs several heuristics that aim to avoid noise. Their union yields an almost complete mapping that includes the correct candidate meaning for 97-100% of the examples in the test datasets. Candidate mentions are mapped to synsets based on similarity 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 as a candidate meaning. We allow partial matches for BabelNet synonyms derived from Wikipedia titles or redirections. However, partial matching is restricted to synsets that belong either to the semantic category "Place" or "Agent". We make use of the semantic category information provided by the DBpedia ontology[3]. A partial match allows the surface string of a mention to differ by up to 3 tokens from the Wikipedia title (excluding everything in parentheses) if the partial string occurred at least once as an anchor for the corresponding Wikipedia page. E.g., 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" or "District", since they were never used as an anchor. If there is no match we try the same procedure applied to the lowercase forms of the surface string or the lemma. For persons we allow matches to all partial names, e.g., only first name, first and middle name, last name, etc.

In addition to the aforementioned candidate extraction we also match surface strings to candidate entities mentioned on their respective disambigua-

[1] e.g., using only the synonyms of a synset

[2] e.g., partial matches for all synonyms of a synset

[3] `http://wiki.dbpedia.org/Ontology`

tion pages in Wikipedia[4]. For cases where adjectives should be disambiguated as nouns, e.g., "English" as a country to "England", we allow candidate mappings through the *pertainment* relation from WordNet. Finally, frequently annotated surface strings in Wikipedia are matched to their corresponding entities, where we stipulate "frequently" to mean that the surface string occurs at least 100 times as anchor in Wikipedia and the entity was either at least 100 times annotated by this surface string or it was annotated above average.

The distinction between nouns and NEs imposes certain restrictions on the set of potential candidates. Candidate synsets for nouns are noun synsets considered as "Concepts" in BabelNet (as opposed to "Named Entities") in addition to all synsets of WordNet senses. On the other hand, candidate synsets for NEs comprise all nominal Babel synsets. Thus, the range of candidate sets for NEs properly contains the one for nouns. We include all nominal synsets as potential candidates for NEs 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 nominal mention, we try to find NE candidates for it before discarding it.

2.4 Multi-Objective Disambiguation

We formulate the disambiguation as a continuous, multi-objective optimization problem. Individual objectives model different aspects of the disambiguation problem. Maximizing these objectives means assigning high probabilities to candidate senses that contribute most to the combined objective. After maximization, we select the candidate meaning with the highest probability as the disambiguated sense. Our model is illustrated in Figure 1.

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:

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

The continuous approach has several advantages over a discrete setting. First, we can ex-

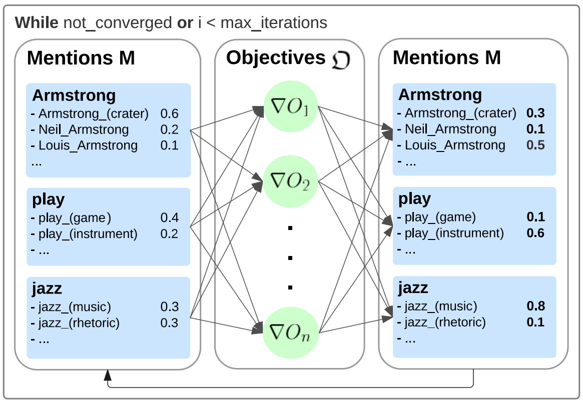

Figure 1: Illustration of our multi-objective approach to WSD & EL for the example sentence: *Armstrong plays jazz.* Mentions are disambiguated by iteratively updating probability distributions over their candidate senses with respect to the given objective gradients ∇O_i.

ploit well established continuous optimization algorithms, such as conjugate gradient or LBFGS. Second, by optimizing upon probability distributions we are optimizing the actually desired result, in contrast to densest sub-graph algorithms where normalized confidence scores are calculated 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.

We normalize each objective using the difference of its maximum and minimum value for a given document, which makes the weighting of the objectives different for each document. The maximum/minimum values can be calculated analytically or, if this is not possible, by running the optimization algorithm with only the given objective for an approximate estimate for the maximum and with its negated form for an approximate minimum. Normalization is important for optimization because it ensures that the individual gradients have similar norms on average for each objective. Without normalization, optimization is biased towards objectives with large gradients.

Given that one of the objectives can be applied to only a fraction of all mentions (e.g., only nominal mentions), we scale each objective by the fraction of mentions it is applied to.

Note that our formulation could easily be extended to using additional coefficients for each ob-

[4] provided by DBpedia at `http://wiki.dbpedia.org/Downloads2014`

jective. However, these hyper-parameters would have to be estimated on development data and therefore, this method could hurt generalization.

Prior Another advantage of working with probability distributions over candidates is the easy integration of prior information. For example, 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 and Wikipedia. We make use of occurrence frequencies extracted by DBpedia Spotlight (Daiber et al., 2013) for synsets containing Wikipedia senses in case of NE disambiguation. For nominal WSD, we employ frequency statistics from WordNet for synsets containing WordNet senses. Laplace-smoothing is applied to all prior frequencies. The priors serve as initialization for the probability distributions over candidate synsets. Note that we use priors "naturally", i.e., as actual priors for initialization only and not during disambiguation itself. They should not be applied during disambiguation because these priors can be very strong and are not domain independent. However, they provide a good initialization which is important for successful continuous optimization.

3 Disambiguation Objectives

3.1 Coherence Objective

Jointly disambiguating all mentions within a document has been shown to have a large impact on disambiguation quality, especially for named entities (Kulkarni et al., 2009). It requires a measurement of semantic relatedness between concepts that can for example be extracted from a semantic network like BabelNet. However, semantic networks usually suffer from data sparsity where important links between concepts might be missing. To deal with this issue, we adopt the idea of using *semantic signatures* from Moro et al. (2014). Following their approach, we create *semantic signatures* for concepts and named entities by running a random walk with restart (RWR) in the semantic network. We count the times a vertex is visited during RWR and define all frequently visited vertices to be the *semantic signature* (i.e., a set of highly related vertices) of the starting concept or named entity vertex.

Our coherence objective aims at maximizing the semantic relatedness among selected candidate senses based on their semantic signatures S_c. We define the continuous objective using probability distributions $p_m(c)$ over the candidate set C_m of each mention $m \in M$ in a document as follows:

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

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

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

where $\mathbb{1}$ denotes the indicator function and $p_m(c)$ is a softmax function. The only free, optimizable parameters are the softmax weights $\mathbf{w_m}$. This objective includes all mentions, i.e., $M_{O_{\text{coh}}} = M$. It can be interpreted as finding the densest subgraph where vertices correspond to mention-candidate pairs and edges to semantic signatures between candidate synsets. However, in contrast to a discrete setup, each vertex is now weighted by its probability and therefore each edge is weighted by the product of its adjacent vertex probabilities.

3.2 Type Objective

One of the biggest problems for supervised approaches to WSD is the limited size and synset coverage of available training datasets 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). For example, WordNet provides a mapping, called lexnames, of synsets into 45 types, which is based on the syntactic categories of synsets and their logical groupings[5]. In WordNet 13.5% of all nouns are ambiguous with an average ambiguity of 2.79 synsets per lemma. Given a noun and a type (lexname), the percentage of ambiguous nouns drops to 7.1% for which the average ambiguity drops to 2.33. This indicates that exploiting type classification for disambiguation can be very useful.

Similarly, for EL it is important to recognize the type of an entity mention in a local context.

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

For example, in the phrase “London beats Manchester” it is very likely that the two city names refer to sports clubs and not to the cities. We utilize an existing mapping from Wikipedia pages to types from the DBpedia ontology, restricting the set of target types to the following: “Activity”, “Organisation”, “Person”, “Event”, “Place” and “Misc” for the rest.

We train a multi-class logistic regression model for each set of types that calculates probability distributions $q_m(t)$ over WN- or DBpedia-types t given a noun- or a NE-mention m, respectively. The features used as input to the model are the following:

- word embedding of mention’s surface string
- sum of word embeddings of all sentence words excluding stopwords
- word embedding of the dependency parse parent
- collocations of surrounding words as in Zhong et al. (2010)
- POS tags with up to 3 tokens distance to m
- possible types of candidate synsets

We employed pre-trained word embeddings from Mikolov et al. (2013) instead of the words themselves to increase generalization.

Type classification is included as an objective in the model as defined in equation 3. It puts type specific weights derived from type classification on candidate synsets, enforcing candidates of fitting type to have higher probabilities. The objective is only applied to noun, NE and verb mentions, i.e., $M_{O_{\text{typ}}} = M_n \cup M_{NE} \cup M_v$.

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

3.3 Regularization Objective

Because candidate priors for NE mentions can be very high, we add an additional L2-regularization objective for NE mentions:

$$O_{L2}(M) = -\frac{\lambda}{2} \sum_{m \in M_{NE}} \|\mathbf{w_m}\|_2^2 \quad (4)$$

The regularization objective is integrated in the overall objective function as it is, i.e., it is not normalized.

Dataset	**|D|**	**|M|**	**KB**
SemEval-2015-13 (*Sem15*) (to be published)	4	757	BN
SemEval-2013-12 (*Sem13*)	13	1931	BN
SemEval-2013-12 (*Sem13*) (Navigli et al., 2013)	13	1644	WN
SemEval-2007-17 (*Sem07*) (Pradhan et al., 2007)	3	159	WN
Senseval 3 (*Sen3*) (Snyder and Palmer, 2004)	4	886	WN
AIDA-CoNLL-testb (*AIDA*) (Hoffart et al., 2011)	216	4530	Wiki
KORE50 (*KORE*) (Hoffart et al., 2012)	50	144	Wiki

Table 1: List of datasets used in experiments with information about their number of documents (D), annotated noun and/or NE mentions (M), and their respective target knowledge base (KB): *BN-BabelNet, WN-WordNet, Wiki-Wikipedia.*

4 Experiments

4.1 Datasets

We evaluated our approach on 7 different datasets, comprising 3 WSD datasets annotated with WordNet senses, 2 datasets annotated with Wikipedia articles for EL and 2 more recent datasets annotated with Babel synsets. Table 1 contains a list of all datasets.

Besides these test datasets we used SemCor (Miller et al., 1993) as training data for WSD and the training part of the AIDA CoNLL dataset for EL.

4.2 Setup

For the creation of semantic signatures we choose the same parameter set as defined by Moro et al. (2014). We run the random walk with a restart probability of 0.85 for a total of 1 million steps for each vertex in the semantic graph and keep vertices visited at least 100 times as semantic signatures.

The L2-regularization objective for named entities is employed with $\lambda = 0.001$, which we found to perform best on the training part of the AIDA-CoNLL dataset.

We trained the multi-class logistic regression model for WN-type classification on SemCor and for DBpedia-type classification on the training part of the AIDA-CoNLL dataset using LBFGS and L2-Regularization with $\lambda = 0.01$ until convergence.

Our system optimizes the combined multi-objective function using Conjugate Gradient

System	KB	Description
IMS (Zhong and Ng, 2010)	WN	supervised, SVM
KPCS (Hoffart et al., 2011)	Wiki	greedy densest-subgraph on combined mention-entity, entity-entity measures
KORE (Hoffart et al., 2012)	Wiki	extension of KPCS with keyphrase relatedness measure between entities
MW (Milne and Witten, 2008)	Wiki	Normalized Google Distance
Babelfy (Moro et al., 2014)	BN	greedy densest-subgraph on semantic signatures

Table 2: Systems used for comparison during evaluation.

(Hestenes and Stiefel, 1952) with up to a maximum of 1000 iterations per document.

We utilized existing implementations from FACTORIE version 1.1 (McCallum et al., 2009) for logistic regression, NER tagging and Conjugate Gradient optimization. For NER tagging we used a pre-trained stacked linear-chain CRF (Lafferty et al., 2001).

4.3 Systems

We compare our approach to state-of-the-art results on all datasets and a most frequent sense (MFS) baseline. The MFS baseline selects the candidate with the highest prior as described in section 2.4. Table 2 contains a list of all systems we compared against. We use *Babelfy* as our main baseline, because of its state-of-the-art performance on all datasets and because it also employed BabelNet as its sense inventory. Note that Babelfy achieved its results with different setups for WSD and EL, in contrast to our model, which uses the same setup for both tasks.

4.4 General Results

We report the performance of all systems in terms of F1-score. To ensure fairness we restricted the candidate sets of the target mentions in each dataset to candidates of their respective reference KB. Note that our candidate mapping strategy ensures for all datasets a $97\% - 100\%$ chance that the target synset is within a mention's candidate set.

This section presents results on the evaluation datasets divided by their respective target KBs: WordNet, Wikipedia and BabelNet.

System	Sens3	Sem07	Sem13
MFS	**72.6**	65.4	62.8
IMS	71.2	63.3	65.7
Babelfy	68.3	62.7	65.9
Our	68.8	**66.0**	**72.8**

Table 3: Results for nouns on WordNet annotated datasets.

System	AIDA	KORE
MFS	70.1	35.4
KPCS	82.2	55.6
KORE-LSH-G	81.8	64.6
MW	82.3	57.6
Babelfy	82.1	**71.5**
Our	**85.1**	67.4

Table 4: Results for NEs on Wikipedia annotated datasets.

WordNet Table 3 shows the results on three datasets for the disambiguation of nouns to WordNet. Our approach exhibits state-of-the-art results outperforming all other systems on two of the three datasets. The model performs slightly worse on the Senseval 3 dataset because of one document in particular where the F1 score is very low compared to the MFS baseline. On the other three documents, however, it performs as good or even better. In general, results from the literature are always worse than the MFS baseline on this dataset. A strong improvement can be seen on the SemEval 2013 Task 12 dataset (Sem13), which is also the largest dataset. Our system achieves an improvement of nearly 7% F1 over the best other system, which translates to an error reduction of roughly 20% given that every word mention gets annotated. Besides the results presented in Table 3, we also evaluated the system on the SemEval 2007 Task 7 dataset for coarse grained WSD, where it achieved 85.5% F1 compared to the best previously reported result of 85.5% F1 from Ponzetto et al. (2010) and Babelfy with 84.6%.

Wikipedia The performance on entity linking was evaluated against state-of-the-art systems on two different datasets. The results in Table 4 demonstrate that our model can compete with the best existing models, showing superior results especially on the large AIDA CoNLL[6] test dataset comprising 216 news texts, where we achieve an error reduction of about 16%, resulting in a new state-of-the-art of 85.1% F1. On the other hand, our system is slightly worse on the KORE dataset compared to Babelfy (6 errors more in total), which might be due to the strong priors and

[6] the largest, freely available dataset for EL.

System	Sem13	Sem15
MFS	66.7	71.1
Babelfy	69.2	–
Best other	–	64.8
Our	**71.5**	**75.4**

Table 5: Results for nouns and NEs on BabelNet annotated datasets.

System	Sem13	Sem15	AIDA
MFS	66.7	71.1	70.1
O_{typ}	68.1	73.8	78.0
$O_{coh} + O_{L2}$	68.1	69.6	82.7
$O_{coh} + O_{typ} + O_{L2}$	**71.5**	**75.4**	**85.1**

Table 6: Detailed results for nouns and NEs on BabelNet annotated datasets and AIDA CoNLL.

the small context. However, the dataset is rather small, containing only 50 sentences, and has been artificially tailored to the use of highly ambiguous entity mentions. For example, persons are most of the time only mentioned by their first names. It is an interesting dataset because it requires the system to employ a lot of background knowledge about mentioned entities.

BabelNet Table 5 shows the results on the 2 existing BabelNet annotated datasets. To our knowledge, our system shows the best performance on both datasets in the literature. An interesting observation is that the F1 score on SemEval 2013 with BabelNet as target KB is lower compared to WordNet as target KB. The reason is that ambiguity rises for nominal mentions by including concepts from Wikipedia that do not exist in WordNet. For example, the Wikipedia concept "formal language" becomes a candidate for the surface string "language".

4.5 Detailed Results

We also experimented with different objective combinations, namely "type only" (O_{typ}), "coherence only" ($O_{coh} + O_{L2}$) and "all" ($O_{coh} + O_{typ} + O_{L2}$), to evaluate the impact of the different objectives. Table 6 shows results of employing individual configurations compared to the MFS baseline.

Results for only using coherence or type exhibit varying performance on the datasets, but still consistently exceed the strong MFS baseline. Combining both objectives always yields better results compared to all other configurations. This finding is important because it proves that the objectives proposed in this work are indeed complementary, and thus demonstrates the significance of combining complementary approaches in one robust framework such as ours.

An additional observation was that DBpedia-type classification slightly overfitted on the AIDA CoNLL training part. When removing DBpedia-type classification from the type objective, results increased marginally on some datasets except for the AIDA CoNLL dataset, where results decreased by roughly 3% F1. The improvements of using DBpedia-type classification are mainly due to the fact that the classifier is able to correctly classify names of places in tables consisting of sports scores not to the "Place" type but to the "Organization" type. Note that the AIDA CoNLL dataset (train and test) contains many of those tables. This shows that including supervised objectives into the system helps when data is available for the domain.

4.6 Generalization

We evaluated the ability of our system to generalize to different domains based on the SemEval 2015 Task 13 dataset. It includes documents from the bio-medical, the math&computer and general domains. Our approach performs particularly well on the bio-medical domain with 86.3% F1 (MFS: 77.3%). Results on the math&computer domain (58.8% F1, MFS: 57.0%), however, reveal that performance still strongly depends on the document topic. This indicates that either the employed resources do not cover this domain as well as others, or that it is generally more difficult to disambiguate. Another potential explanation is that enforcing only pairwise coherence does not take the hidden concepts *computer* and *maths* into account, which connect all concepts, but are never actually mentioned. An interesting point for future research might be the introduction of an additional objective or the extension of the coherence objective to allow indirect connections between candidate meanings through shared topics or categories.

Besides these very specific findings, the model's ability to generalize is strongly supported by its good results across all datasets, covering a variety of different topics.

5 Related Work

WSD Approaches to WSD can be distinguished by the kind of resource exploited. The two main resources for WSD are sense annotated datasets and knowledge bases. Typical supervised approaches like IMS (Zhong and Ng, 2010) train classifiers that learn from existing, annotated examples. They suffer from the sparsity of sense annotated datasets that is due to the data acquisition bottleneck (Pilehvar and Navigli, 2014). There have been approaches to overcome this issue through the automatic generation of such resources based on bootstrapping (Pham et al., 2005), sentences containing unambiguous relatives of senses (Martinez et al., 2008) or exploiting Wikipedia (Shen et al., 2013). On the other hand, knowledge-based approaches achieve good performances rivaling state-of-the-art supervised systems (Ponzetto and Navigli, 2010) by using existing structured knowledge (Lesk, 1986; Agirre et al., 2014), or take advantage of the structure of a given semantic network through connectivity or centrality measures (Tsatsaronis et al., 2007; Navigli and Lapata, 2010). Such systems benefit from the availability of numerous KBs for a variety of domains. We believe that both knowledge-based approaches and supervised methods have unique, complementary abilities that need to be combined for sophisticated disambiguation.

EL Typical EL systems employ supervised machine learning algorithms to classify or rank candidate entities (Bunescu and Pasca, 2006; Milne and Witten, 2008; Zhang et al., 2010). Common features include popularity metrics based on Wikipedia's graph structure or on name mention frequency (Dredze et al., 2010; Han and Zhao, 2009), similarity metrics exploring Wikipedia's concept relations (Han and Zhao, 2009), and string similarity features. Mihalcea and Csomai (2007) disambiguate each mention independently given its sentence level context only. In contrast, Cucerzan (2007) and Kulkarni et al. (Kulkarni et al., 2009) recognize the interdependence between entities in a wider context. The most similar work to ours is that of Hoffart et al. (2011) which was the first that combined local and global context measures in one robust model. However, objectives and the disambiguation algorithm differ from our work. They represent the disambiguation task as a densest subgraph problem where the least connected entity is eliminated in each iteration. The discrete treatment of candidate entities can be problematic especially at the beginning of disambiguation where it is biased towards mentions with many candidates.

Babelfy (Moro et al., 2014) is a knowledge-based approach for joint WSD and EL that also uses a greedy densest subgraph algorithm for disambiguation. It employs a single coherence model based on semantic signatures similar to our coherence objective. The system's very good performance indicates that the semantic signatures provide a powerful resource for joint disambiguation. However, because we believe it is not sufficient to only enforce semantic agreement among nouns and entities, our approach includes an objective that also focuses on the local context of mentions, making it more robust.

6 Conclusions & Future Work

We have presented a novel approach for the joint disambiguation of nouns and named entities based on an extensible framework. Our system employs continuous optimization on a multi-objective function during disambiguation. The integration of complementary objectives into our formalism demonstrates that robust disambiguation can be achieved by considering both the local and the global context of a mention. Our model outperforms previous state-of-the-art systems for nominal WSD and for EL. It is the first system that achieves such results on various WSD and EL datasets using a single setup.

In future work, new objectives should be integrated into the framework and existing objectives could be enhanced. For example, it would be interesting to express semantic relatedness continuously rather than in a binary setting for the coherence objective. Additionally, using the entire model during training could ensure better compatibility between the different objectives. At the moment, the model itself is composed of different pre-trained models that are only combined during disambiguation.

Acknowledgment

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) through the projects ALL SIDES (01IW14002), BBDC (01IS14013E), and by the German Federal Ministry of Economics and Energy (BMWi) through the project SD4M (01MD15007B), and by Google through a Focused Research Award granted in July 2013.

References

[Agirre et al.2014] Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

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

[Cucerzan2007] Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716.

[Daiber et al.2013] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM.

[Dredze et al.2010] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proc. of the 23rd International Conference on Computational Linguistics*, pages 277–285. Association for Computational Linguistics.

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

[Han and Zhao2009] Xianpei Han and Jun Zhao. 2009. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proc. of the 18th ACM conference on Information and knowledge management*, pages 215–224. ACM.

[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 *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

[Hoffart et al.2012] Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. Kore: keyphrase overlap relatedness for entity disambiguation. In *Proc. of the 21st ACM international conference on Information and knowledge management*, pages 545–554. ACM.

[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.

[Kulkarni et al.2009] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466. ACM.

[Lafferty et al.2001] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Lesk1986] Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proc. of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

[Martinez et al.2008] David Martinez, Oier Lopez De Lacalle, and Eneko Agirre. 2008. On the use of automatically acquired examples for all-nouns word sense disambiguation. *J. Artif. Intell. Res.(JAIR)*, 33:79–107.

[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)*.

[Mihalcea and Csomai2007] Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proc. of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM.

[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 *Proc. of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

[Milne and Witten2008] David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proc. of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

[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 Lapata2010] Roberto Navigli and Mirella Lapata. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(4):678–692.

[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.

[Navigli et al.2013] Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (SEM)*, volume 2, pages 222–231.

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

[Pham et al.2005] Thanh Phong Pham, Hwee Tou Ng, and Wee Sun Lee. 2005. Word sense disambiguation with semi-supervised learning. In *Proc. of the national conference on artificial intelligence*, volume 20, page 1093. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[Pilehvar and Navigli2014] Mohammad Taher Pilehvar and Roberto Navigli. 2014. A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. *Computational Linguistics*, 40(4):837–881.

[Ponzetto and Navigli2010] Simone Paolo Ponzetto and Roberto Navigli. 2010. Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proc. of the 48th annual meeting of the association for computational linguistics*, pages 1522–1531. Association for Computational Linguistics.

[Pradhan et al.2007] Sameer S Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task 17: English lexical sample, srl and all words. In *Proc. of the 4th International Workshop on Semantic Evaluations*, pages 87–92. Association for Computational Linguistics.

[Shen et al.2013] Hui Shen, Razvan Bunescu, and Rada Mihalcea. 2013. Coarse to fine grained sense disambiguation in wikipedia. *Proc. of SEM*, pages 22–31.

[Snyder and Palmer2004] Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43.

[Tsatsaronis et al.2007] George Tsatsaronis, Michalis Vazirgiannis, and Ion Androutsopoulos. 2007. Word sense disambiguation with spreading activation networks generated from thesauri. In *IJCAI*, volume 7, pages 1725–1730.

[Zhang et al.2010] Wei Zhang, Jian Su, Chew Lim Tan, and Wen Ting Wang. 2010. Entity linking leveraging: automatically generated annotation. In *Proc. of the 23rd International Conference on Computational Linguistics*, pages 1290–1298. 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 *Proc. of the ACL 2010 System Demonstrations*, pages 78–83. Association for Computational Linguistics.