Topic Models for Word Sense Disambiguation and Token-based Idiom Detection

Linlin Li, Benjamin Roth, and Caroline Sporleder
Saarland University, Postfach 15 11 50
66041 Saarbrücken, Germany
{linlin, beroth, csporled}@coli.uni-saarland.de

Abstract

This paper presents a probabilistic model for sense disambiguation which chooses the best sense based on the conditional probability of sense paraphrases given a context. We use a topic model to decompose this conditional probability into two conditional probabilities with latent variables. We propose three different instantiations of the model for solving sense disambiguation problems with different degrees of resource availability. The proposed models are tested on three different tasks: coarse-grained word sense disambiguation, fine-grained word sense disambiguation, and detection of literal vs. nonliteral usages of potentially idiomatic expressions. In all three cases, we outperform state-of-the-art systems either quantitatively or statistically significantly.

1 Introduction

Word sense disambiguation (WSD) is the task of automatically determining the correct sense for a target word given the context in which it occurs. WSD is an important problem in NLP and an essential preprocessing step for many applications, including machine translation, question answering and information extraction. However, WSD is a difficult task, and despite the fact that it has been the focus of much research over the years, state-of-the-art systems are still often not good enough for real-world applications. One major factor that makes WSD difficult is a relative lack of manually annotated corpora, which hampers the performance of supervised systems.

To address this problem, there has been a significant amount of work on unsupervised WSD that does not require manually sense-disambiguated training data (see McCarthy (2009) for an overview). Recently, several researchers have experimented with topic models (Brody and Lapata, 2009; Boyd-Graber et al., 2007; Boyd-Graber and Blei, 2007; Cai et al., 2007) for sense disambiguation and induction. Topic models are generative probabilistic models of text corpora in which each document is modelled as a mixture over (latent) topics, which are in turn represented by a distribution over words.

Previous approaches using topic models for sense disambiguation either embed topic features in a supervised model (Cai et al., 2007) or rely heavily on the structure of hierarchical lexicons such as WordNet (Boyd-Graber et al., 2007). In this paper, we propose a novel framework which is fairly resource-poor in that it requires only 1) a large unlabelled corpus from which to estimate the topics distributions, and 2) paraphrases for the possible target senses. The paraphrases can be user-supplied or can be taken from existing resources.

We approach the sense disambiguation task by choosing the best sense based on the conditional probability of sense paraphrases given a context. We propose three models which are suitable for different situations: Model I requires knowledge of the prior distribution over senses and directly maximizes the conditional probability of a sense given the context; Model II maximizes this conditional probability by maximizing the cosine value of two topic-document vectors (one for the sense and one for the context). We apply these models to coarse- and fine-grained WSD and find that they outperform comparable systems for both tasks.

We also test our framework on the related task of idiom detection, which involves distinguishing literal and nonliteral usages of potentially ambiguous expressions such as *rock the boat*. For this task, we propose a third model. Model III calculates the probability of a sense given a context according to the component words of the sense paraphrase. Specifically, it chooses the sense type which maximizes the probability (given the context) of the paraphrase component word with the highest likelihood of occurring in that context. This model also outperforms state-of-the-art systems.

2 Related Work

There is a large body of work on WSD, covering supervised, unsupervised (word sense induction) and knowledge-based approaches (see McCarthy (2009) for an overview). While most supervised approaches treat the task as a classification task and use hand-labelled corpora as training data, most unsupervised systems automatically group word tokens into similar groups using clustering algorithms, and then assign labels to each sense cluster. Knowledge-based approaches exploit information contained in existing resources. They can be combined with supervised machine-learning models to assemble semi-supervised approaches.

Recently, a number of systems have been proposed that make use of topic models for sense disambiguation. Cai et al. (2007), for example, use LDA to capture global context. They compute topic models from a large unlabelled corpus and include them as features in a supervised system. Boyd-Graber and Blei (2007) propose an unsupervised approach that integrates McCarthy et al.'s (2004) method for finding predominant word senses into a topic modelling framework. In addition to generating a topic from the document's topic distribution and sampling a word from that topic, the enhanced model also generates a distributional neighbour for the chosen word and then assigns a sense based on the word, its neighbour and the topic. Boyd-Graber and Blei (2007) test their method on WSD and information retrieval tasks and find that it can lead to modest improvements over state-of-the-art results.

In another unsupervised system, Boyd-Graber et al. (2007) enhance the basic LDA algorithm by incorporating WordNet senses as an additional latent variable. Instead of generating words directly from a topic, each topic is associated with a random walk through the WordNet hierarchy which generates the observed word. Topics and synsets are then inferred together. While Boyd-Graber et al. (2007) show that this method can lead to improvements in accuracy, they also find that idiosyncracies in the hierarchical structure of WordNet can harm performance. This is a general problem for methods which use hierarchical lexicons to model semantic distance (Budanitsky and Hirst, 2006). In our approach, we circumvent this problem by exploiting paraphrase information for the target senses rather than relying on the structure of WordNet as a whole.

Topic models have also been applied to the related task of word sense induction. Brody and Lapata (2009) propose a method that integrates a number of different linguistic features into a single generative model.

Topic models have been previously considered for metaphor extraction and estimating the frequency of metaphors (Klebanov et al., 2009; Bethard et al., 2009). However, we have a different focus in this paper, which aims to distinguish literal and nonliteral usages of potential idiomatic expressions. A number of methods have been applied to this task. Katz and Giesbrecht (2006) devise a supervised method in which they compute the meaning vectors for the literal and nonliteral usages of a given expression in the trainning data. Birke and Sarkar (2006) use a clustering algorithm which compares test instances to two automatically constructed seed sets (one literal and one nonliteral), assigning the label of the closest set. An unsupervised method that computes cohesive links between the component words of the target expression and its context have been proposed (Sporleder and Li, 2009; Li and Sporleder, 2009). Their system predicts literal usages when strong links can be found.

3 The Sense Disambiguation Model

3.1 Topic Model

As pointed out by Hofmann (1999), the starting point of topic models is to decompose the conditional word-document probability distribution $p(w|d)$ into two different distributions: the word-topic distribution $p(w|z)$, and the topic-document distribution $p(z|d)$ (see Equation 1). This allows each semantic topic z to be represented as a multinominal distribution of words $p(w|z)$, and each document d to be represented as a multinominal distribution of semantic topics $p(z|d)$. The model introduces a conditional independence assumption that document d and word w are independent conditioned on the hidden variable, topic z.

$$p(w|d) = \sum_z p(z|d)p(w|z) \tag{1}$$

LDA is a Bayesian version of this framework with Dirichlet hyper-parameters (Blei et al., 2003).

The inference of the two distributions given an observed corpus can be done through Gibbs Sampling (Geman and Geman, 1987; Griffiths and Steyvers, 2004). For each turn of the sampling, each word in each document is assigned a semantic topic based on the current word-topic distribution and topic-document distribution. The resulting topic assignments are then used to re-estimate a new word-topic distribution and topic-document distribution for the next turn. This process repeats until convergence. To avoid statistical coincidence, the final estimation of the distributions is made by the average of all the turns after convergence.

3.2 The Sense Disambiguation Model

Assigning the correct sense s to a target word w occurring in a context c involves finding the sense which maximizes the conditional probability of senses given a context:

$$s = \arg\max_{s_i} p(s_i|c) \tag{2}$$

In our model, we represent a sense (s_i) as a collection of 'paraphrases' that capture (some aspect of) the meaning of the sense. These paraphrases can be taken from an existing resource such as WordNet (Miller, 1995) or supplied by the user (see Section 4).

This conditional probability is decomposed by incorporating a hidden variable, topic z, introduced by the topic model. We propose three variations of the basic model, depending on how much background information is available, i.e., knowledge of the prior sense distribution available and type of sense paraphrases used. In Model I and Model II, the sense paraphrases are obtained from WordNet, and both the context and the sense paraphrases are treated as documents, $c = dc$ and $s = ds$.

WordNet is a fairly rich resource which provides detailed information about word senses (glosses, example sentences, synsets, semantic relations between senses, etc.). Sometimes such detailed information may not be available, for instance for languages for which such a resource does not exist or for expressions that are not very well covered in WordNet, such as idioms. For those situations, we propose another model, Model III, in which contexts are treated as documents while sense paraphrases are treated as sequences of independent words.[1]

Model I directly maximizes the conditional probability of the sense given the context, where the sense is modeled as a 'paraphrase document' ds and the context as a 'context document' dc. The conditional probability of sense given context $p(ds|dc)$ can be rewritten as a joint probability divided by a normalization factor:

$$p(ds|dc) = \frac{p(ds, dc)}{p(dc)} \tag{3}$$

This joint probability can be rewritten as a generative process by introducing a hidden variable z. We make the conditional independence assumption that, conditioned on the topic z, a paraphrase document ds is generated independently of the specific context document dc:

$$p(ds, dc) = \sum_z p(ds)p(z|ds)p(dc|z) \tag{4}$$

We apply the same process to the conditional probability $p(dc|z)$. It can be rewritten as:

$$p(dc|z) = \frac{p(dc)p(z|dc)}{p(z)} \tag{5}$$

Now, the disambiguation model $p(ds|dc)$ can be rewritten as a prior $p(ds)$ times a topic function $f(z)$:

$$p(ds|dc) = p(ds) \sum_z \frac{p(z|dc)p(z|ds)}{p(z)} \tag{6}$$

As $p(z)$ is a uniform distribution according to the uniform Dirichlet priors assumption, Equation 6 can be rewritten as:

$$p(ds|dc) \propto p(ds) \sum_z p(z|dc)p(z|ds) \tag{7}$$

Model I:

$$\arg\max_{ds_i} p(ds_i) \sum_z p(z|dc)p(z|ds_i) \tag{8}$$

Model I has the disadvantage that it requires information about the prior distribution of senses

[1] The idea is that these key words capture the meaning of the idioms.

$p(ds)$, which is not always available. We use sense frequency information from WordNet to estimate the prior sense distribution, although it must be kept in mind that, depending on the genre of the texts, it is possible that the distribution of senses in the testing corpus may diverge greatly from the WordNet-based estimation. If there is no means for estimating the prior sense distribution of an experimental corpus, generally a uniform distribution must be assumed. However, this assumption does not hold, as the true distribution of word senses is often highly skewed (McCarthy, 2009).

To overcome this problem, we propose Model II, which indirectly maximizes the sense-context probability by maximizing the cosine value of two document vectors that encode the document-topic frequencies from sampling, $v(z|dc)$ and $v(z|ds)$. The document vectors are represented by topics, with each dimension representing the number of times that the tokens in this document are assigned to a certain topic.

Model II:

$$\arg\max_{ds_i} \cos(v(z|dc), v(z|ds_i)) \qquad (9)$$

If the prior distribution of senses is known, Model I is the best choice. However, Model II has to be chosen instead when this knowledge is not available. In our experiments, we test the performance of both models (see Section 5).

If the sense paraphrases are very short, it is difficult to reliably estimate $p(z|ds)$. In order to solve this problem, we treat the sense paraphrase ds as a 'query', a concept which is used in information retrieval. One model from information retrieval takes the conditional probability of the query given the document as a product of all the conditional probabilities of words in the query given the document. The assumption is that the query is generated by a collection of conditionally independent words (Song and Croft, 1999).

We make the same assumption here. However, instead of taking the product of all the conditional probabilities of words given the document, we take the maximum. There are two reasons for this: (i) taking the product may penalize longer paraphrases since the product of probabilities decreases as there are more words; (ii) we do not want to model the probability of generating specific paraphrases, but rather the probability of generating a sense, which might only be represented by one or two words in the paraphrases (e.g., the potentially idiomatic phrase 'rock the boat' can be paraphrased as 'break the norm' or 'cause trouble'. A similar topic distribution to that of the individual words 'norm' or 'trouble' would be strong supporting evidence of the corresponding idiomatic reading.). We propose **Model III**:

$$\arg\max_{qs_i} \max_{w_i \in qs} \sum_z p(w_i|z)p(z|dc) \qquad (10)$$

where qs is a collection of words contained in the sense paraphrases.

3.3 Inference

One possible inference approach is to combine the context documents and sense paraphrases into a corpus and run Gibbs sampling on this corpus. The problem with this approach is that the test set and sense paraphrase set are relatively small, and topic models running on a small corpus are less likely to capture rich semantic topics. One simple explanation for this is that a small corpus usually has a relatively small vocabulary, which is less representative of topics, i.e., $p(w|z)$ cannot be estimated reliably.

In order to overcome this problem, we infer the word-topic distribution from a very large corpus (Wikipedia dump, see Section 4). All the following inference experiments on the test corpus are based on the assumption that the word-topic distribution $p(w|z)$ is the same as the one estimated from the Wikipedia dump. Inference of topic-document distributions for context and sense paraphrases is done by fixing the word-topic distribution as a constant.

4 Experimental Setup

We evaluate our models on three different tasks: coarse-grained WSD, fine-grained WSD and literal vs. nonliteral sense detection. In this section we discuss our experimental set-up. We start by describing the three datasets for evaluation and another dataset for probability estimation. We also discuss how we choose sense paraphrases and instance contexts.

Data We use three datasets for evaluation. The coarse-grained task is evaluated on the Semeval-2007 Task-07 benchmark dataset released by Navigli et al. (2009). The dataset consists of 5377 words of running text from five different articles: the first three were obtained from the WSJ corpus, the fourth was the Wikipedia entry for *computer programming*, and the fifth was an excerpt of

Amy Steedman's *Knights of the Art*, biographies of Italian painters. The proportion of the non news text, the last two articles, constitutes 51.87% of the whole testing set. It consists of 1108 nouns, 591 verbs, 362 adjectives, and 208 adverbs. The data were annotated with coarse-grained senses which were obtained by clustering senses from the WordNet 2.1 sense inventory based on the procedure proposed by Navigli (2006).

To determine whether our model is also suitable for fine-grained WSD, we test on the data provided by Pradhan et al. (2009) for the Semeval-2007 Task-17 (English fine-grained all-words task). This dataset is a subset of the set from Task-07. It comprises the three WSJ articles from Navigli et al. (2009). A total of 465 lemmas were selected as instances from about 3500 words of text. There are 10 instances marked as 'U' (undecided sense tag). Of the remaining 455 instances, 159 are nouns and 296 are verbs. The sense inventory is from WordNet 2.1.

Finally, we test our model on the related sense disambiguation task of distinguishing literal and nonliteral usages of potentially ambiguous expressions such as *break the ice*. For this, we use the dataset from Sporleder and Li (2009) as a test set. This dataset consists of 3964 instances of 17 potential English idioms which were manually annotated as *literal* or *nonliteral*.

A Wikipedia dump[2] is used to estimate the multinomial word-topic distribution. This dataset, which consists of 320,000 articles,[3] is significantly larger than the BNC, which is the dataset used by Boyd-Graber et al. (2007). All markup from the Wikipedia dump was stripped off using the same filter as the ESA implementation (Sorg and Cimiano, 2008), and stopwords were filtered out using the Snowball (Porter, October 2001) stopword list. In addition, words with a Wikipedia document frequency of 1 were filtered out. The lemmatized version of the corpus consists of 299,825 lexical units.

The test sets were POS-tagged and lemmatized using RASP (Briscoe and Carroll, 2006). The inference processes are run on the lemmatized version of the corpus. For the Semeval-2007 Task 17 English all-words, the organizers do not supply the part-of-speech and lemma information of the target instances. In order to avoid the wrong predictions caused by tagging or lemmatization errors, we manually corrected any bad tags and lemmas for the target instances.[4]

Sense Paraphrases For word sense disambiguation tasks, the paraphrases of the sense keys are represented by information from WordNet 2.1. (Miller, 1995). To obtain the paraphrases, we use the **word forms**, **glosses** and **example sentences** of the *synset itself* and a set of selected *reference synsets* (i.e., synsets linked to the target synset by specific semantic relations, see Table 1). We excluded the 'hypernym reference synsets', since information common to all of the child synsets may confuse the disambiguation process.

For the literal vs. nonliteral sense detection task, we selected the paraphrases of the nonliteral meaning from several online idiom dictionaries. For the literal senses, we used 2-3 manually selected words with which we tried to capture (aspects of) the literal meaning of the expression.[5] For instance, the literal 'paraphrases' that we chose for 'break the ice' were *ice, water* and *snow*. The paraphrases are shorter for the idiom task than for the WSD task, because the meaning descriptions from the idiom dictionaries are shorter than what we get from WordNet. In the latter case, each sense can be represented by its synset as well as its reference synsets.

Instance Context We experimented with different context sizes for the disambiguation task. The five different context settings that we used for the WSD tasks are: collocations (1w), ±5-word window (5w), ±10-word window (10w), current sentence, and whole text. Because the idiom corpus also includes explicitly marked paragraph boundaries, we included 'paragraph' as a sixth type of context size for the idiom sense detection task.

5 Experiments

As mentioned above, we test our proposed sense disambiguation framework on three tasks. We start by describing the sampling experiments for

[2] We use the English snapshot of 2009-07-13

[3] All articles of fewer than 100 words were discarded.

[4] This was done by comparing the predicted sense keys and the gold standard sense keys. We only checked instances for which the POS-tags in the predicted sense keys are not consistent with those in the gold standard. This was the case for around 20 instances.

[5] Note that we use the word 'paraphrase' in a fairly wide sense in this paper. Sometimes it is not possible to obtain exact paraphrases. This applies especially to the task of distinguishing literal from nonliteral senses of multi-word expressions. In this case we take as paraphrases some key words which capture salient aspects of the meaning.

POS	Paraphrase reference synsets
N	hyponyms, instance hyponyms, member holonyms, substance holonyms, part holonyms, member meronyms, part meronyms, substance meronyms, attributes, topic members, region members, usage members, topics, regions, usages
V	Troponyms, entailments, outcomes, phrases, verb groups, topics, regions, usages, sentence frames
A	similar, pertainym, attributes, related, topics, regions, usages
R	pertainyms, topics, regions, usages

Table 1: Selected reference synsets from WordNet that were used for different parts-of-speech to obtain word sense paraphrase. N(noun), V(verb), A(adj), R(adv).

estimating the word-topic distribution from the Wikipedia dump. We used the package provided by Wang et al. (2009) with the suggested Dirichlet hyper-parameters [6]. In order to avoid statistical instability, the final result is averaged over the last 50 iterations. We did four rounds of sampling with 1000, 500, 250, and 125 topics respectively. The final word-topic distribution is a normalized concatenate of the four distributions estimated in each round. In average, the sampling program run on the Wikipedia dump consumed 20G memory, and each round took about one week on a single AMD Dual-Core 1000MHZ processor.

5.1 Coarse-Grained WSD

In this section we first describe the landscape of similar systems against which we compare our models, then present the results of the comparison. The systems that participated in the SemEval-2007 coarse-grained WSD task (Task-07) can be divided into three categories, depending on whether training data is needed and whether other types of background knowledge are required: What we call Type I includes all the systems that need annotated training data. All the participating systems that have the mark *TR* fall into this category (see Navigli et al. (2009) for the evaluation for all the participating systems). Type II consists of systems that do not need training data but require prior knowledge of the sense distribution (estimated sense frequency). All the participating systems that have the mark *MFS* belong to this category. Systems that need neither training data nor prior sense distribution knowledge are categorized as Type III.

We make this distinction based on two principles: (i) the cost of building a system; (ii) the portability of the established resource. Type III is the cheapest system to build, while Type I and Type II both need extra resources. Type II has an advantage over Type I since the prior knowledge of the sense distribution can be estimated from annotated corpora (e.g.: SemCor, Senseval). In contrast, training data in Type I may be system specific (e.g.: different input format, different annotation guidelines). McCarthy (2009) also addresses the issue of performance and cost by comparing supervised word sense disambiguation systems with unsupervised ones.

We exclude the system provided by one of the organizers (UoR-SSI) from our categorization. The reason is that although this system is claimed to be unsupervised, and it performs better than all the participating systems (including the supervised systems) in the SemEval-2007 shared task, it still needs to incorporate a lot of prior knowledge, specifically information about co-occurrences between different word senses, which was obtained from a number of resources (SSI+LKB) including: (i) SemCor (manually annotated); (ii) LDC-DSO (partly manually annotated); (iii) collocation dictionaries which are then disambiguated semi-automatically. Even though the system is not "trained", it needs a lot of information which is largely dependent on manually annotated data, so it does not fit neatly into the categories Type II or Type III either.

Table 2 lists the best participating systems of each type in the SemEval-2007 task (Type I: NUS-PT (Chan et al., 2007); Type II: UPV-WSD (Buscaldi and Rosso, 2007); Type III: TKB-UO (Anaya-Sánchez et al., 2007)). Our Model I belongs to Type II, and our Model II belongs to Type III.

Table 2 compares the performance of our models with the Semeval-2007 participating systems. We only compare the F-score, since all the compared systems have an attempted rate[7] of 1.0,

[6] They were set as: $\alpha = \frac{50}{\#topics}$ and $\beta = 0.01$.

[7] Attempted rate is defined as the total number of disambiguated output instances divided by the total number of input

which makes both the precision and recall rates the same as the F-score. We focus on comparisons between our models and the best SemEval-2007 participating systems within the same type. Model I is compared with UPV-WSD, and Model II is compared with TKB-UO. In addition, we also compare our system with the most frequent sense baseline which was not outperformed by any of the systems of Type II and Type III in the SemEval-2007 task.

Comparison on Type III is marked with $'$, while comparison on Type II is marked with $*$. We find that Model II performs statistically significantly better than the best participating system of the same type TKB-UO (p<<0.01, χ^2 test). When encoded with the prior knowledge of sense distribution, Model I outperforms by 1.36% the best Type II system UPV-WSD, although the difference is not statistically significant. Furthermore, Model I also quantitatively outperforms the most frequent sense baseline BL$_{mfs}$, which, as mentioned above, was not beat by any participating systems that do not use training data.

We also find that our model works best for nouns. The unsupervised Type III model Model II achieves better results than the most frequent sense baseline on nouns, but not on other parts-of-speech. This is in line with results obtained by previous systems (Griffiths et al., 2005; Boyd-Graber and Blei, 2008; Cai et al., 2007). While the performance on verbs can be increased to outperform the most frequent sense baseline by including the prior sense probability, the performance on adjectives and adverbs remains below the most frequent sense baseline. We think that there are three reasons for this: first, adjectives and adverbs have fewer reference synsets for paraphrases compared with nouns and verbs (see Table 1); second, adjectives and adverbs tend to convey less key semantic content in the document, so they are more difficult to capture by the topic model; and third, adjectives and adverbs are a small portion of the test set, so their performances are statistically unstable. For example, if 'already' appears 10 times out of 20 adverb instances, a system may get bad result on adverbs only because of its failure to disambiguate the word 'already'.

System	Noun	Verb	Adj	Adv	All
UoR-SSI	84.12	78.34	85.36	88.46	83.21
NUS-PT	82.31	78.51	85.64	89.42	82.50
UPV-WSD	79.33	72.76	84.53	81.52	**78.63***
TKB-UO	70.76	62.61	78.73	74.04	**70.21**$'$
MII–ref	78.16	70.39	79.56	81.25	76.64
MII+ref	80.05	70.73	82.04	82.21	**78.14**$'$
MI+ref	79.96	75.47	83.98	86.06	**79.99***
BL$_{mfs}$	77.44	75.30	84.25	87.50	**78.99***

Table 2: Model performance (F-score) on the coarse-grained dataset (context=sentence). Paraphrases with/without reference synsets (+ref/-ref).

Context	Ate.	Pre.	Rec.	F1
$\pm 1w$	91.67	75.05	68.80	71.79
$\pm 5w$	99.29	77.14	76.60	76.87
$\pm 10w$	100	77.92	77.92	77.92
text	100	76.86	76.86	76.86
sent.	100	**78.14**	**78.14**	**78.14**

Table 3: Model II performance on different context size. attempted rate (Ate.), precision (Pre.), recall (Rec.), F-score (F1).

Paraphrase analysis Table 2 also shows the effect of different ways of choosing sense paraphrases. MII+ref is the result of including the reference synsets, while MII-ref excludes the reference synsets. As can be seen from the table, including all reference synsets in sense paraphrases increases performance. Longer paraphrases contain more information, and they are statistically more stable for inference.

We find that nouns get the greatest performance boost from including reference synsets, as they have the largest number of different types of synsets. We also find the 'similar' reference synset for adjectives to be very useful. Performance on adjectives increases by 2.75% when including this reference synset.

Context analysis In order to study how the context influences the performance, we experiment with Model II on different context sizes (see Table 3). We find *sentence context* is the best size for this disambiguation task. Using a smaller context not only reduces the precision, but also reduces the recall rate, which is caused by the all-zero topic assignment by the topic model for documents only containing words that are not in the vocabulary. As a result, the model is unable to disambiguate. The context based on the whole text (article) does not perform well either, possibly because using the full text folds in too much noisy information.

instances.

System	F-score
RACAI	52.7 ±4.5
BL$_{mfs}$	55.91±4.5
MI+ref	**56.99**±4.5

Table 4: Model performance (F-score) for the fine-grained word sense disambiguation task.

5.2 Fine-grained WSD

We saw in the previous section that our framework performs well on coarse-grained WSD. Fine-grained WSD, however, is a more difficult task. To determine whether our framework is also able to detect subtler sense distinctions, we tested Model I on the English all-words subtask of SemEval-2007 Task-17 (see Table 4).

We find that Model I performs better than both the best unsupervised system, RACAI (Ion and Tufiş, 2007) and the most frequent sense baseline (BL$_{mfs}$), although these differences are not statistically significant due to the small size of the available test data (465).

5.3 Idiom Sense Disambiguation

In the previous section, we provided the results of applying our framework to coarse- and fine-grained word sense disambiguation tasks. For both tasks, our models outperform the state-of-the-art systems of the same type either quantitatively or statistically significantly. In this section, we apply Model III to another sense disambiguation task, namely distinguishing literal and nonliteral senses of ambiguous expressions.

WordNet has a relatively low coverage for idiomatic expressions. In order to represent nonliteral senses, we replace the paraphrases obtained automatically from WordNet by words selected manually from online idiom dictionaries (for the nonliteral sense) and by linguistic introspection (for the literal sense). We then compare the topic distributions of literal and nonliteral senses.

As the paraphrases obtained from the idiom dictionary are very short, we treat the paraphrase as a sequence of independent words instead of as a document and apply Model III (see Section 3). Table 5 shows the results of our proposed model compared with state-of-the-art systems. We find that the system significantly outperforms the majority baseline (p<<0.01, χ^2 test) and the cohesion-graph based approach proposed by Sporleder and Li (2009) (p<<0.01, χ^2 test). The system also outperforms the bootstrapping system by Li and Sporleder (2009), although not statistically significantly. This shows how a limited amount of human knowledge (e.g., paraphrases) can be added to an unsupervised system for a strong boost in performance (Model III compared with the cohesion-graph and the bootstrapping approaches).

System	Prec$_l$	Rec$_l$	F$_l$	Acc.
Base$_{maj}$	-	-	-	78.25
co-graph	50.04	69.72	58.26	78.38
boot.	71.86	66.36	69.00	87.03
Model III	67.05	81.07	73.40	87.24

Table 5: Performance on the literal or nonliteral sense disambiguation task on idioms. literal precision (Prec$_l$), literal recall (Rec$_l$), literal F-score (F$_l$), accuracy(Acc.).

For obvious reasons, this approach is sensitive to the quality of the paraphrases. The paraphrases chosen to characterise (aspects of) the meaning of a sense should be non-ambiguous between the literal or idiomatic meaning. For instance, ‘fire’ is not a good choice for a paraphrase of the literal reading of ‘play with fire’, since this word can be interpreted literally as ‘fire’ or metaphorically as ‘something dangerous’. The verb component word ‘play’ is a better literal paraphrase.

For the same reason, this approach works well for expressions where the literal and nonliteral readings are well separated (i.e., occur in different contexts), while the performance drops for expressions whose literal and idiomatic readings can appear in a similar context. We test the performance on individual idioms on the five most frequent idioms in our corpus[8] (see Table 6). We find that ‘drop the ball’ is a difficult case. The words ‘fault’, ‘mistake’, ‘fail’ or ‘miss’ can be used as the nonliteral paraphrases. However, it is also highly likely that these words are used to describe a scenario in a baseball game, in which ‘drop the ball’ is used literally. In contrast, the performance on ‘rock the boat’ is much better, since the nonliteral reading of the phrases ‘break the norm’ or ‘cause trouble’ are less likely to be linked with the literal reading ‘boat’. This may also be because ‘boat’ is not often used metaphorically in the corpus.

As the topic distribution of nouns and verbs exhibit different properties, topic comparisons across parts-of-speech do not make sense. We

[8] We tested only on the most frequent idioms in order to avoid statistically unreliable observations.

Idiom	Acc.
drop the ball	75.86
play with fire	91.17
break the ice	87.43
rock the boat	95.82
set in stone	89.39

Table 6: Performance on individual idioms.

make the topic distributions comparable by making sure each type of paraphrase contains the same sets of parts-of-speech. For instance, we do not permit combinations of literal paraphrases which only consist of nouns and nonliteral paraphrases which only consist of verbs.

6 Conclusion

We propose three models for sense disambiguation on words and multi-word expressions. The basic idea of these models is to compare the topic distribution of a target instance with the candidate sense paraphrases and choose the most probable one. While Model I and Model III model the problem in a probabilistic way, Model II uses a vector space model by comparing the cosine values of two topic vectors. Model II and Model III are completely unsupervised, while Model I needs the prior sense distribution. Model I and Model II treat the sense paraphrases as documents, while Model III treats the sense paraphrases as a collection of independent words.

We test the proposed models on three tasks. We apply Model I and Model II to the WSD tasks due to the availability of more paraphrase information. Model III is applied to the idiom detection task since the paraphrases from the idiom dictionary are smaller. We find that all models outperform comparable state-of-the-art systems either quantitatively or statistically significantly.

By testing our framework on three different sense disambiguation tasks, we show that the framework can be used flexibly in different application tasks. The system also points out a promising way of solving the granularity problem of word sense disambiguation, as new application tasks which need different sense granularities can utilize this framework when new paraphrases of sense clusters are available. In addition, this system can also be used in a larger context such as figurative language identification (*literal* or *figurative*) and sentiment detection (*positive* or *negative*).

Acknowledgments

This work was funded by the DFG within the Cluster of Excellence “Multimodal Computing and Interaction”.

References

H. Anaya-Sánchez, A. Pons-Porrata, R. Berlanga-Llavori. 2007. TKB-UO: using sense clustering for WSD. In *SemEval ’07: Proceedings of the 4th International Workshop on Semantic Evaluations*, 322–325.

S. Bethard, V. T. Lai, J. H. Martin. 2009. Topic model analysis of metaphor frequency for psycholinguistic stimuli. In *CALC ’09: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, 9–16, Morristown, NJ, USA. Association for Computational Linguistics.

J. Birke, A. Sarkar. 2006. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-06*.

D. M. Blei, A. Y. Ng, M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Reseach*, 3:993–1022.

J. Boyd-Graber, D. Blei. 2007. PUTOP: turning predominant senses into a topic model for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 277–281.

J. Boyd-Graber, D. Blei. 2008. Syntactic topic models. *Computational Linguistics*.

J. Boyd-Graber, D. Blei, X. Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 1024–1033.

T. Briscoe, J. Carroll. 2006. Evaluating the accuracy of an unlexicalized statistical parser on the PARC DepBank. In *Proceedings of the COLING/ACL on Main conference poster sessions*, 41–48.

S. Brody, M. Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 103–111.

A. Budanitsky, G. Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

D. Buscaldi, P. Rosso. 2007. UPV-WSD: Combining different WSD methods by means of Fuzzy Borda Voting. In *SemEval ’07: Proceedings of the 4th International Workshop on Semantic Evaluations*, 434–437.

J. Cai, W. S. Lee, Y. W. Teh. 2007. Improving word sense disambiguation using topic features. In *Proceedings of the 2007 Joint Conference on Empirical*

Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 1015–1023.

Y. S. Chan, H. T. Ng, Z. Zhong. 2007. NUS-PT: exploiting parallel texts for word sense disambiguation in the English all-words tasks. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, 253–256.

S. Geman, D. Geman. 1987. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in computer vision: issues, problems, principles, and paradigms*, 564–584. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

T. L. Griffiths, M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

T. L. Griffiths, M. Steyvers, D. M. Blei, J. B. Tenenbaum. 2005. Integrating topics and syntax. In *In Advances in Neural Information Processing Systems 17*, 537–544. MIT Press.

T. Hofmann. 1999. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50–57.

R. Ion, D. Tufiş. 2007. Racai: meaning affinity models. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, 282–287, Morristown, NJ, USA. Association for Computational Linguistics.

G. Katz, E. Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*.

B. B. Klebanov, E. Beigman, D. Diermeier. 2009. Discourse topics and metaphors. In *CALC '09: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

L. Li, C. Sporleder. 2009. Contextual idiom detection without labelled data. In *Proceedings of EMNLP-09*.

D. McCarthy, R. Koeling, J. Weeds, J. Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, 279–286.

D. McCarthy. 2009. Word sense disambiguation: An overview. *Language and Linguistics Compass*, 3(2):537–558.

G. A. Miller. 1995. WordNet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

R. Navigli, K. C. Litkowski, O. Hargraves. 2009. SemEval-2007 Task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluation (SemEval-2007)*.

R. Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 44th Annual Meeting of the Association for Computational Liguistics joint with the 21st International Conference on Computational Liguistics (COLING-ACL 2006)*.

M. Porter. October 2001. Snowball: A language for stemming algorithms. `http://snowball.tartarus.org/texts/introduction.html`.

S. S. Pradhan, E. Loper, D. Dligach, M. Palmer. 2009. SemEval-2007 Task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluation (SemEval-2007)*.

F. Song, W. B. Croft. 1999. A general language model for information retrieval (poster abstract). In *Research and Development in Information Retrieval*, 279–280.

P. Sorg, P. Cimiano. 2008. Cross-lingual information retrieval with explicit semantic analysis. In *In Working Notes for the CLEF 2008 Workshop*.

C. Sporleder, L. Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of EACL-09*.

Y. Wang, H. Bai, M. Stanton, W.-Y. Chen, E. Y. Chang. 2009. Plda: Parallel latent dirichlet allocation for large-scale applications. In *Proc. of 5th International Conference on Algorithmic Aspects in Information and Management*. Software available at `http://code.google.com/p/plda`.