Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation

Tristan Miller[1] *Chris Biemann*[1] *Torsten Zesch*[1,2] *Iryna Gurevych*[1,2]
(1) Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
(2) Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
http://www.ukp.tu-darmstadt.de/

ABSTRACT
We explore the contribution of distributional information for purely knowledge-based word sense disambiguation. Specifically, we use a distributional thesaurus, computed from a large parsed corpus, for lexical expansion of context and sense information. This bridges the lexical gap that is seen as the major obstacle for word overlap–based approaches. We apply this mechanism to two traditional knowledge-based methods and show that distributional information significantly improves disambiguation results across several data sets. This improvement exceeds the state of the art for disambiguation without sense frequency information—a situation which is especially encountered with new domains or languages for which no sense-annotated corpus is available.

TITLE AND ABSTRACT IN GERMAN

Über die Bestimmung lexikalischer Expansionen mittels distributioneller Ähnlichkeit und deren Einsatz in der wissensbasierten Lesartendisambiguierung

Wir untersuchen den Einfluss distributioneller Informationen auf die rein wissensbasierte Lesartendisambiguierung. Basierend auf einem distributionellen Thesaurus, den wir aus einem großen geparsten Korpus erzeugen, erweitern wir die Definition der Lesart und deren Kontext mit lexikalischen Expansionen. Dadurch schließen wir die 'lexikalische Lücke', die sich als Haupthindernis für Ansätze basierend auf Wortgemeinsamkeiten herausgestellt hat. Wir erweitern zwei klassische wissensbasierte Ansätze um lexikalische Expansionen und zeigen, dass dadurch die Qualität der Lesartendisambiguierung deutlich erhöht wird. Wir erzielen die bisher besten veröffentlichten Ergebnisse für Disambiguierung ohne Nutzung der Lesartenhäufigkeiten, was besonders für Domänen oder Sprachen relevant ist, für die keine Lesarten-annotierten Korpora zur Verfügung stehen.

KEYWORDS: word sense disambiguation, distributional thesaurus, lexical expansion.

KEYWORDS IN GERMAN: Lesartendisambiguierung, distributioneller Thesaurus, lexikalische Expansion.

1 Introduction

Word sense disambiguation (WSD)—the task of determining which sense a word carries in a particular context—is a longstanding core research problem in computational linguistics. Approaches to WSD can be classified according to what lexical resources are used: *knowledge-based* techniques rely only on machine-readable dictionaries (MRDs), lexical semantic resources (LSRs), and untagged corpora, whereas *supervised* approaches instead or additionally use manually annotated training examples. Though supervised systems generally perform better, their use is restricted to scenarios where a sufficient amount of hand-crafted training data is available. Estimates for the amount of time required to produce such training data are pessimistic (Mihalcea and Chklovski, 2003); this knowledge acquisition bottleneck is the principal motivation behind research into semi-supervised and knowledge-based WSD. The latter have the advantage that, unlike manually annotated corpora, MRDs and LSRs do exist for many languages and domains.

In the past, however, knowledge-based approaches have suffered from a variant of the lexical gap problem: when matching a sense description to a given context of a disambiguation target, it is often the case that the description and context do not have much vocabulary in common. We propose a new method to bridge this lexical gap which is based on statistics collected from a large, unannotated background corpus. Specifically, we enrich the textual information from the context and the MRD with lexical expansions produced by a distributional thesaurus.

We examine the contribution of these expansions to two popular knowledge-based algorithms, including one which tries to address the lexical gap through LSR-based augmentation of the sense description. We show that, especially in situations for which no sense frequency information is available, improvements from adding more knowledge and from adding lexical expansions add up, allowing us to improve over the state of the art for knowledge-based all-words disambiguation.

2 Background

MRD-based word sense disambiguation began with Lesk (1986), who proposed that two or more words in context could be simultaneously disambiguated by looking up their respective definitions in a dictionary and finding the maximum overlap between each combination of their senses. A popular variant is the "simplified" Lesk algorithm (Kilgarriff and Rosenzweig, 2000), which disambiguates one word at a time by comparing each of its definitions to the context in which the word is found. This variant avoids the combinatorial explosion of word sense combinations the original version suffers from when trying to disambiguate multiple words in a text.

Both the original and simplified versions of the Lesk algorithm suffer from low coverage due to the lexical gap problem: because the context and definitions are usually quite short, it is often the case that there are no overlapping content words at all. Various solutions to the problem have been proposed, with varying degrees of success. Lesk himself proposed increasing the size of the context window, though Vasilescu et al. (2004) found that performance was generally better for smaller contexts. Lesk also proposed augmenting the definitions with example sentences provided by some dictionaries; Kilgarriff and Rosenzweig (2000) found that including them (for simplified Lesk) led to significantly better performance than using the definitions alone. Banerjee and Pedersen (2002) observed that, where there exists a lexical resource like WordNet (Fellbaum, 1998) which also provides semantic relations between senses,

these can be used to augment definitions with those from related senses (such as hypernyms and hyponyms); their "extended" Lesk algorithm was found to be a great improvement over the original algorithm. Subsequent researchers (e.g., Ponzetto and Navigli (2010)) have combined the "simplified" and "extended" approaches into a "simplified extended" algorithm, in which augmented definitions are compared not with each other, but with the target word context.

Many successful approaches to automatic WSD in recent years rely on distributional information to model the "topicality" of the context and the sense definition.[1] They include using vector-space dimensionality reduction techniques like LSA (Gliozzo et al., 2005) or LDA (Cai et al., 2007; Li et al., 2010), additionally collected text material per sense as in topic signatures (Martínez et al., 2008), and clustering for word sense induction as features (Agirre et al., 2006; Biemann, 2012); the importance of bridging the lexical gap is reflected in all those recent advances, be it in knowledge-based or supervised WSD scenarios.

In this paper, we employ a source of semantic similarity whose application to automatic WSD has never before been explored: using a distributional thesaurus, or DT (Lin, 1998), we expand the lexical representations of the context and sense definition with additional terms. On this expanded representation, we are able to apply the well-known overlap-based methods to text similarity without any modification. Lexical expansion has already proven useful in semantic text similarity evaluations (Bär et al., 2012), which is a task related to matching sense definitions to contexts.

The intuition behind our approach is depicted in Figure 1: say we wish to disambiguate the word *interest* in the sentence, "The loan interest is paid monthly." The correct sense definition from our MRD ("a fixed charge for borrowing money") has no words in common with the context, and thus would not be selected by an overlap-based WSD algorithm. But with the addition of ten lexical expansions per content word (shown in smaller text), we increase the number of overlapping word pairs (shown in boldface) to seven.

Observe also that this expansion of linear text sequences into a two-dimensional representation makes conceptual associations (cf. the associative relations of de Saussure (1916)) explicit, allowing for purely symbolic matching instead of using a vector-space representation such as LSA. The main differences to vector-space approaches are the following: On the one hand, vector-space approaches usually use dimensionality reduction in order to handle sparsity, which results in a fixed number of topics/dimensions. While very salient collection-specific topics are handled well by this approach, rare topics are either conflated into a single rump topic, or distributed amongst the salient topics. Our DT-based expansion technique has no notion of dimensions since it works on the word level, and thus does not suffer from this kind of sampling error that is inevitable when representing a large vocabulary with a small fixed number of dimensions or topics. On the other hand, while vector-space models do a good job at ranking candidates according to their similarity,[2] they fail to efficiently generate a top-ranked list of possible expansions: due to its size, it is infeasible to rank the full vocabulary every time. Lexical expansion methods based on distributional similarity, however, generate a short list of highly similar candidates.

[1] Distributional information was also used in a much older, semi-automatic approach by Tugwell and Kilgarriff (2001). In their technique, "word sketches" consisting of common patterns of usage of a word were extracted from a large POS-tagged corpus and presented to a human operator for manual sense annotation. The pattern–sense associations were then used as input to bootstrapping WSD algorithm.

[2] See Rapp (2004) for an early success of vector-space models on a semantic task.

The	loan	interest	is	paid	**monthly**.
	mortgage			paying	annual
	loans			**pay**	weekly
	debt			pays	yearly
	financing			owed	quarterly
	mortgages			generated	**hefty**
	credit			invested	daily
	lease			spent	regular
	bond			collected	**additional**
	grant			raised	**substantial**
	funding			reimbursed	recent

interest :	a	fixed	charge	for	borrowing	money
		solved	charges		spending	dollars
		hefty	counts		borrow	cash
		resolved	charging		lending	funds
		monthly	cost		borrowed	billions
		additional	conviction		**debt**	monies
		existing	allegation		investment	millions
		reduced	**pay**		raising	trillions
		done	suspicion		inflows	**funding**
		current	count		investing	resources
		substantial	part		borrowings	donations

Figure 1: Example showing the intuition behind lexical expansion for matching a context (top) to a sense definition (bottom). The term to be disambiguated is underlined and the matching terms are in boldface.

The lexical expansions shown in Figure 1 were generated by the same DT used in our experiments. However, for the general case, we make no assumptions about the method that generates the lexical expansions, which could just as easily come from, say, translations via bridge languages, paraphrasing systems, or lexical substitution systems.

3 Experiments

Our experiments measure the contribution of various lexical expansion schemes to the simplified and simplified extended variants of the Lesk algorithm. We chose these algorithms because of their simplicity and transparency, making it easy for us to trace through their operation and see exactly how and where the lexical expansions help or hinder disambiguation. Furthermore, Lesk variants perform remarkably well despite their simplicity, making them popular choices as baselines and as starting points for developing more sophisticated WSD algorithms.

Our experiments with the simplified Lesk algorithm use only the definitions provided by WordNet; they are intended to model the case where we have a generic MRD which provides sense definitions, but no additional lexical semantic information such as example sentences or semantic relations. Such scenarios are typical of many languages and domains, where there is no WordNet-like resource and no manually sense-annotated corpus which could be used for supervised WSD or for a backoff to the most frequent sense. Accurate WSD systems that rely on the existence of an MRD only could pave the way to wider application of lexical disambiguation in NLP applications.

By contrast, the experiments with the simplified extended Lesk algorithm assume the existence of a WordNet-like resource with a taxonomic structure; the definition text for a sense is therefore constructed from the gloss, synonyms, and example sentences provided by WordNet, plus the

same information for all senses in a direct semantic relation. This setup specifically targets situations where such a resource serves as the sense inventory but no large sense-annotated corpus is available for supervised WSD (thus precluding use of the most frequent sense backoff). This is the case for many languages, where wordnets but not manually tagged corpora are available, and also for domain-specific WSD using the English WordNet. Whereas other approaches in this setting (Ponzetto and Navigli, 2010; Henrich et al., 2012) aim at improving WSD accuracy through the combination of several lexical resources, we restrict ourselves to WordNet and bridge the lexical gap with non-supervised, data-driven methods.

How one computes the overlap between two strings was left unspecified by Lesk; we therefore adopt the simple approach of removing occurrences of the target word, treating both strings as bags of case-insensitive word tokens, and taking the cardinality of their intersection. We do not preprocess the texts by lemmatization or stop word filtering, since the terms in the distributional thesaurus are likewise unprocessed (as in Figure 1), and because preliminary experiments showed that such preprocessing brought no benefit. We use the sentence containing the target word as the context. The sense with the highest overlap with the context is assigned a probability of 1; when $k \geq 2$ senses are tied for the highest overlap count, these senses are assigned a probability of $1/k$. All other senses are assigned a probability of 0. The probabilities are then used for scoring during evaluation (see §3.3).

3.1 Use of distributional information

We now describe the creation and the use of our distributional thesaurus. In the fashion of Lin (1998), we parsed a 10M sentence English news corpus from the Leipzig Corpora Collection[3] (Biemann et al., 2007) with the Stanford parser (de Marneffe et al., 2006) and used collapsed dependencies to extract features for words: each dependency triple (w_1, r, w_2) denoting a directed dependency of type r between words w_1 and w_2 results in a feature (r, w_2) characterizing w_1, and a feature (w_1, r) characterizing w_2. Words are thereby represented by the concatenation of the surface form and the POS as assigned by the parser. After counting the frequency of each feature for each word, we apply a significance measure (log-likelihood test (Dunning, 1993)), rank features per word according to their significance, and prune the data, keeping only the 300 most salient features per word. The similarity of two words is given by the number of their common features (which we will shortly illustrate with an example). The pruning operation greatly reduces run time at thesaurus construction, rendering memory reduction techniques like Goyal et al. (2012) unnecessary. Despite its simplicity and the basic count of feature overlap, we found this setting to be equal to or better than more complex weighting schemes in word similarity evaluations. Across all parts of speech, the DT contains five or more similar terms for a vocabulary of over 150 000 words.

To illustrate the DT, Table 1 shows the top three most similar words to the noun *paper*, together with the features which determine the similarities. Amongst their 300 most salient features as determined by the significance measure, *newspaper* and *paper* share 45, *book* and *paper* share 33, and *article* and *paper* share 28; these numbers constitute the terms' respective similarity scores.

The DT is used to expand the context and the sense definitions in the following way: For each content word (that is, adjectives, nouns, adverbs, and verbs) we retrieve the n most similar terms from the DT and add them to the textual representation. Since our overlap-based

[3]Available at `http://corpora.uni-leipzig.de/`; data for 229 languages and dialects is published.

term	score	shared features
newspaper|NN	45	told|VBD|-dobj column|NN|-prep|in local|JJ|amod editor|NN|-poss edition|NN|-prep|of editor|NN|-prep|of hometown|NN|nn industry|NN|-nn clips|NNS|-nn shredded|JJ|amod pick|VB|-dobj news|NNP|appos daily|JJ|amod writes|VBZ|-nsubj write|VB|-prep|for wrote|VBD|-prep|for wrote|VBD|-prep|in wrapped|VBN|-prep|in reading|VBG|-prep|in reading|VBG|-dobj read|VBD|-prep|in read|VBD|-dobj read|VBP|-prep|in read|VB|-dobj read|VB|-prep|in record|NN|prep|of article|NN|-prep|in reports|VBZ|-nsubj reported|VBD|-nsubj printed|VBN|amod printed|VBD|-nsubj printed|VBN|-prep|in published|VBN|-prep|in published|VBN|partmod published|VBD|-nsubj sunday|NNP|nn section|NN|-prep|of school|NN|nn saw|VBD|-prep|in ad|NN|-prep|in copy|NN|-prep|of page|NN|-prep|of pages|NNS|-prep|of morning|NN|nn story|NN|-prep|in
book|NN	33	recent|JJ|amod read|VB|-dobj read|VBD|-dobj reading|VBG|-dobj edition|NN|-prep|of printed|VBN|amod industry|NN|-nn described|VBN|-prep|in writing|VBG|-dobj wrote|VBD|-prep|in wrote|VBD|rcmod write|VB|-dobj written|VBN|rcmod written|VBN|-dobj wrote|VBD|-dobj pick|VB|-dobj photo|NN|nn co-author|NN|-prep|of co-authored|VBN|-dobj section|NN|-prep|of published|VBN|-dobj published|VBN|-nsubjpass published|VBD|-dobj published|VBN|partmod copy|NN|-prep|of buying|VBG|-dobj buy|VB|-dobj author|NN|-prep|of bag|NN|-nn bags|NNS|-nn page|NN|-prep|of pages|NNS|-prep|of titled|VBN|partmod
article|NN	28	authors|NNS|-prep|of original|JJ|amod notes|VBZ|-nsubj published|VBN|-dobj published|VBD|-dobj published|VBN|-nsubjpass published|VBN|partmod write|VB|-dobj wrote|VBD|rcmod wrote|VBD|-prep|in written|VBN|rcmod wrote|VBD|-dobj written|VBN|-dobj writing|VBG|-dobj reported|VBD|-nsubj describing|VBG|partmod described|VBN|-prep|in copy|NN|-prep|of said|VBD|-prep|in recent|JJ|amod read|VB|-dobj read|VB|-prep|in read|VBD|-dobj read|VBD|-prep|in reading|VBG|-dobj author|NN|-prep|of titled|VBN|partmod lancet|NNP|nn

Table 1: Illustration of a DT entry with features, showing the most similar terms to the noun *paper*.

approaches treat contexts and sense definitions as unordered bags of words, we do not need to take precautions with respect to the positions of words and expansions within the texts. The bags of words are filtered by removing occurrences of the disambiguation target. Then, we count the overlaps as usual between the expanded context and sense definitions. In our experiments we test $n = 10, 20, \ldots, 100$.

We had the intuition that the optimal number of expansions may depend on the part of speech of the word to be disambiguated, and perhaps also on the parts of speech of the words being expanded. Therefore, we parameterized our expansion procedure such that the part of speech of the target word determined the number of expansions, and also whether all words were expanded or only those of a certain part of speech.

3.2 Data sets

Data sets for WSD can generally be classified as *fine-grained* or *coarse-grained* according to the granularity of the sense inventory used for the annotations. Another common distinction is between the *all-words* task, in which the aim is to provide an annotation for every content word

in long running texts, and the *lexical sample* task, where several instances from the same small set of target words are annotated in (usually very short) contexts. We tested our systems on several coarse- and fine-grained data sets, and in both the all-words and lexical sample settings. However, most of our analysis will focus on the coarse-grained all-words scenario, as all-words provides a wider and more natural distribution of target words and senses, and because the fine sense distinctions of WordNet are considered a major obstacle to accurate WSD (Navigli, 2009). Additionally, as we discuss below, the fine-grained data sets available to us have various issues which render them unsuitable for comparisons with the state of the art.

Our coarse-grained data set is from the SemEval-2007 English all-words disambiguation task (Navigli et al., 2007). It consists of five non-fiction documents from various sources, where each of the 2269 content words (362 adjectives, 1108 nouns, 208 adverbs, and 591 verbs) has been annotated with clusters of WordNet 2.1 senses. For this data set only, we make a slight modification to our algorithm to account for this clustering: instead of choosing the WordNet sense with the highest overlap, we add up the overlap counts of each cluster's constituent senses, and then select the best cluster.

For our fine-grained experiments, we used the all-words and lexical sample tasks from Senseval-2 (Palmer et al., 2001; Kilgarriff, 2001) and Senseval-3 (Snyder and Palmer, 2004). With these data sets, however, several factors hinder direct comparison to previously published results. There are a number of errors in the gold standard annotations, and the methodology of the original task is different from what has subsequently become common. Specifically, not all of the target words have a corresponding entry in the sense inventory, and systems were originally expected to mark these "unassignable" senses as such. In the case of Senseval-2, the gold standard annotations were made using an unpublished (and now lost) version of WordNet. Subsequent researchers have adopted a variety of mutually incompatible methods for dealing with these issues. For our runs, we use Rada Mihalcea's WordNet 3.0 conversions of the corpora[4] and remove from consideration all "unassignable" target word instances. We do not fix the erroneous annotations, which means that even our baselines cannot achieve 100% coverage.

3.3 Baselines and measures

We use the evaluation metrics standard in word sense disambiguation research (Palmer et al., 2006; Navigli, 2009). Each disambiguation target receives a *score* equal to the probability the system assigned to the correct sense.[5] *Coverage* is the proportion of target word instances for which the system attempted a sense assignment, *precision* (P) is the sum of scores for the correct sense assignments divided by the number of target word instances for which the system made an attempt, and *recall* (R, also known as *accuracy*) is the sum of scores for the correct sense assignments divided by the number of target word instances. The *F-measure* is the harmonic mean of precision and recall: $F_1 = 2PR \div (P + R)$. Note that according to these definitions, $P \leq R$, and when coverage is 100%, $P = R = F_1$. In this paper we express all these measures as a percentage (i.e., in the range $[0, 100]$).

[4] `http://www.cse.unt.edu/~rada/downloads.html#sensevalsemcor`

[5] Where the probability is less than 1, this is mathematically equivalent to the average score which would have been obtained, over repeated runs, of choosing a sense at random to break any ties. It is effectively a backoff to a random sense baseline, ensuring 100% coverage even when there is no overlap.

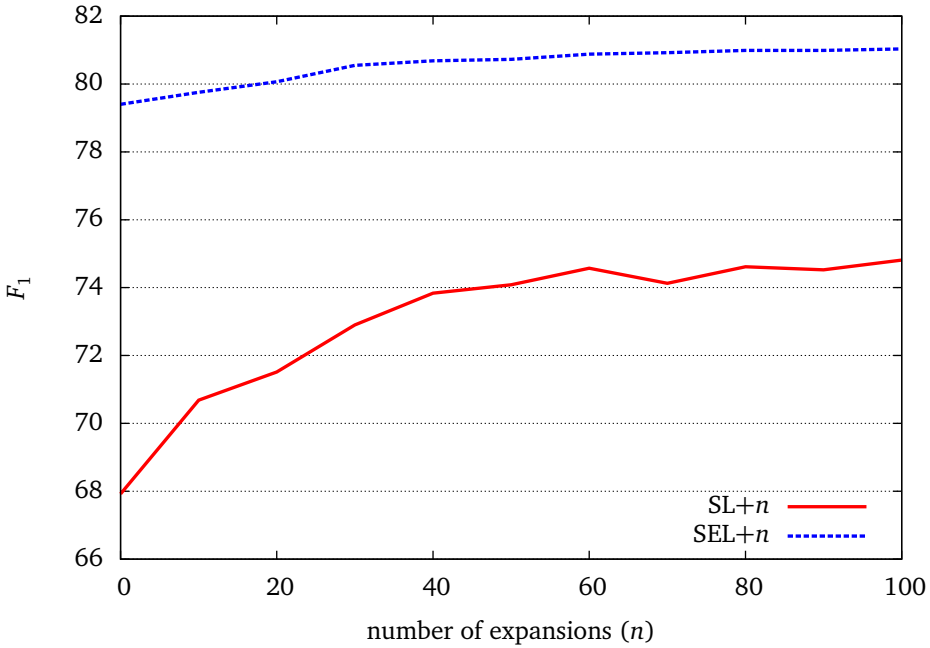

Figure 2: Results (F_1) on the SemEval-2007 corpus by number of lexical expansions

Our systems were compared against a computed random baseline which scores

$$P = R = F_1 = \frac{1}{|W|} \sum_{i=1}^{|W|} \frac{1}{|S(w_i)|},$$

where $W = \{w_1, w_2, \ldots\}$ is the set of target word instances in the corpus and $S(w_i)$ is the set of candidate senses for some target word w_i. This is equivalent to the score, averaged over repeated runs, of randomly choosing one of the candidate senses for each target word.

We also report accuracy of the most frequent sense (MFS) baseline, which always chooses the sense which occurs most frequently in SemCor (Mihalcea, 2008), a very large manually annotated corpus. Note that unlike our knowledge-based systems, MFS is a supervised baseline, and cannot actually be applied to the use cases for which our non-supervised systems are intended. Nonetheless, it is included here as it gives some idea of what accuracy could be achieved, at minimum, were one to go to the considerable expense of creating a manually tagged training corpus. Note that MFS is a notoriously difficult baseline to beat even for supervised systems.

3.4 Results

On the SemEval-2007 data set, the basic configuration of simplified Lesk (SL+0)—i.e., without any lexical expansions—achieves an overall F_1 of 67.92, which is already much better than the random baseline ($F_1 = 61.28$). When we tried adding a fixed number of lexical expansions to all content words, we observed that accuracy generally increased sublinearly with the number of

	part of speech				
system	**adj.**	**noun**	**adv.**	**verb**	**all**
MFS baseline	84.25	77.44	87.50	75.30	78.89
random baseline	68.54	61.96	69.15	52.81	61.28
SL+0	75.32	69.71	69.75	59.46	67.92
SL+100	82.18	76.31	78.85	66.07	74.81
SEL+0	87.19	81.52	74.87	**72.26**	79.40
SEL+100	**88.40**	**83.45**	80.29	72.25	**81.03**
TKB-UO	78.73	70.76	74.04	62.61	70.21
MII+ref	82.04	80.05	**82.21**	70.73	78.14
WN^{++}-DC	—	79.4	—	—	—

Table 2: Results (F_1) on the SemEval-2007 corpus by part of speech

expansions. The highest accuracy was obtained by using 100 expansions (the maximum number we tried); we denote this configuration SL+100. SL+100's F-measure of 74.81 represents a relative increase of more than 10% over SL+0. The simplified extended Lesk configuration also benefitted from lexical expansions, though the effect was less pronounced: the basic version without expansions (SEL+0) achieves $F_1 = 79.40$, and adding 100 lexical expansions (SEL+100) yields a relative performance increase of just over 2%, to $F_1 = 81.03$. As with simplified Lesk, accuracy increased sublinearly with the number of expansions. This effect is visualized in Figure 2, which plots the F-measure for the two algorithms according to the number of lexical expansions used.

Table 2 shows the F-measure of our two baselines (top), our algorithms (middle), and some state-of-the-art knowledge-based systems (bottom), broken down by target word part of speech. In each column the best result, excluding the supervised MFS baseline, is shown in boldface. TKB-UO (Anaya-Sánchez et al., 2007) was the best-performing knowledge-based system at the SemEval-2007 competition; it is a clustering-based system which uses WordNet 2.1 (but not its sense frequency information) as its sole source of knowledge. Among later knowledge-based systems using this data set, MII+ref (Li et al., 2010), a topic model approach, achieves the highest result we are aware of. This work maps sense descriptions and target word contexts to a topic distribution vector as sampled by LDA (Blei et al., 2003). WN^{++}-DC (Ponzetto and Navigli, 2010) uses an altogether different approach: it disambiguates nouns in a sentence by building a graph of candidate senses linked by semantic relations, and then for each target word selecting the sense with the highest vertex degree. When using semantic relations from WordNet alone the method achieves $F_1 = 74.5$, but when WordNet is enriched with additional semantic relations from an online encyclopedia performance increases to $F_1 = 79.4$. Note that, uniquely among the results in the table, WN^{++}-DC does not achieve full coverage ($P = 87.3$, $R = 72.7$).

POS-optimized results. We also tried using different expansion strategies for target words of different parts of speech: for each target word POS, we tried expanding only adjectives, only nouns, etc., and tried each of these scenarios for the same eleven values of n as previously. Because this procedure involved tuning on the test data, we do not include the results for comparison in Table 2. However, they are interesting as they give an upper bound on per-

system	Senseval-2 lexical sample	Senseval-2 all-words	Senseval-3 all-words
MFS baseline	41.56	65.36	65.63
random baseline	15.46	39.54	32.89
SL+0	17.10	39.02	35.41
SL+100	20.92	45.69	37.17
SEL+0	28.60	54.22	48.76
SEL+30	32.72	57.77	53.09

Table 3: Results (F_1) on the Senseval-2 and -3 corpora

formance for the case where the expansions-per-POS parameters are optimized on a set of manually annotated training examples—that is, a mildly supervised variant of our otherwise knowledge-based algorithms.

For simplified Lesk, we found that accuracy for nouns, adverbs, and verbs remained highest when all content words were given 100 expansions, but adjectives fared better when all content words were given only 60 expansions. With this configuration we achieve an overall F_1 of 74.94. The best simplified extended Lesk configuration achieves $F_1 = 81.27$ when for adjectives we apply 20 expansions to all content words; for nouns, 60 expansions to all content words; for adverbs, 80 expansions to all content words; and for verbs, 30 expansions to adverbs only. That verbs benefit from adverb expansions is not surprising, given that the latter often serve to modify the former. Why the optimal *number* of expansions should vary with the target word part of speech is not as clear. In any case, the extra performance gains from POS expansion optimization were quite small, not exceeding a quarter of a percentage point over the non-optimized versions.

Fine-grained results. As with the coarse-grained task, we found that using lexical expansions resulted in an improvement in accuracy in the fine-grained tasks. However, in this setting we did not observe the same continuously improving accuracy from using more and more expansions; in all but one case, adding expansions helped to a point, after which accuracy started to decrease. This effect was particularly noticeable with simplified extended Lesk, where peak accuracy was achieved with around 30 expansions. For simplified Lesk, the optimum was less stable across the corpora, ranging from 60 to 100 expansions. We believe that this is because the expanded terms provided by the DT reflect broad conceptual relations which, taken in aggregate, do not precisely map to the narrow sense distinctions of the sense inventory. This is not a problem when we stick to the first highly salient expansions provided by the DT, but beyond this the conceptual relations become too tenuous and fuzzy to facilitate disambiguation.

Table 3 shows the results of our systems and baselines on the Senseval-2 lexical sample and all-words tasks and the Senseval-3 all-words task. For simplified extended Lesk we show the results of using 30 expansions (SEL+30); as simplified Lesk had no consistent peak accuracy we stick with 100 expansions (SL+100). The results, while quite expectedly lower than the coarse-grained scores in absolute terms, nonetheless validate the utility of our approach in fine-grained tasks. Not only does the use of expansions significantly increase the accuracy, but in the case of the Senseval-2 corpora, the relative increase is much higher than that of the coarse-grained tasks. For SL+100, the relative improvements over the unexpanded algorithms

for the lexical sample and all-words data sets are 22.3% and 17.1%, respectively, and for SEL+100 they are 14.4% and 6.5%, respectively.

4 Discussion

In this section, we discuss our results and put them in the perspective of applicability of WSD systems. Our lexical expansion mechanism leads to a relative improvement of up to 22% in the fine-grained evaluation and 10% in the coarse-grained evaluation for the "simple" setup. This is achieved by merely adding lexical items to the representation of the sense description and context, and without changing the algorithm. Especially in situations where there exists a reasonably coarse-grained MRD for the language or domain, this is a major improvement over previous approaches on applications where one is not in the comfortable situation of having sense frequency information. In our opinion, this scenario has been neglected in the past, despite occurring in practice much more often than the case where one has access to a rich LSR, let alone sufficient training data for supervised disambiguation.

The expansions from distributional similarity are complementary to those coming from richer knowledge resources, as our results for fitting simplified extended Lesk with DT expansions show: even in the situation where a richer lexical resource allows for bridging the lexical gap via descriptions from related senses, we still see an additional relative improvement of 2% to 14% when comparing the F-measure of the SEL+n system against the SEL+0 baseline. Not only does this system outperform all previous approaches to coarse-grained WSD without MFS backoff, it is also able to outperform the MFS baseline itself, both generally and for certain parts of speech.

We emphasize that while the DT uses additional text data for computing the similarity scores used in the lexical expansion step, the overall system is purely knowledge-based because it is not trained on sense-labelled examples; the DT similarities are computed on the basis of an automatically parsed but otherwise unannotated corpus. This marks an important difference from the system described in Navigli and Velardi (2005) which, although it also uses collocations extracted from large corpora, avails itself of manual sense annotations wherever possible.

While the comparison of results to other methods on the same coarse-grained data sets suggests that lexical expansion using a distributional thesaurus leads to more precise disambiguation systems than word or topic vectors, our point is somewhat different: Realizing lexical expansions and thus explicitly generating associated terms to a textual representation opens up a new way of thinking about bridging lexical gaps and semantic matching of similar meaning. In light of the fact that distributional similarity (Lin, 1998) and overlap-based approaches to WSD (Lesk, 1986) have existed for a long time now, it is somewhat surprising that this avenue had not been explored earlier.

4.1 Error analysis

In order to better understand where and how our system is succeeding and failing, we now present an error analysis of the results, both in aggregate and for some individual cases. To begin, we computed a confusion matrix showing the percentage of the 2269 SemEval-2007 target word instances for which the SL+0 and SL+100 algorithms made a correct disambiguation, made an incorrect disambiguation, or failed to make an assignment at all without resorting to the random choice backoff (see Table 4). Table 5 shows the same confusion matrix for the SEL+0

		SL+100			
		unassigned	**incorrect**	**correct**	**total**
SL+0	**unassigned**	0.2	8.1	9.7	18.0
	incorrect	0.1	14.1	6.2	20.4
	correct	0.0	2.8	58.8	61.6
	total	0.4	25.0	74.7	100.0

Table 4: Confusion matrix for SL+0 and SL+100

		SEL+100			
		unassigned	**incorrect**	**correct**	**total**
SEL+0	**unassigned**	0.1	3.5	4.0	7.6
	incorrect	0.0	14.9	2.8	17.7
	correct	0.0	1.9	72.9	74.7
	total	0.1	20.3	79.6	100.0

Table 5: Confusion matrix for SEL+0 and SEL+100

and SEL+100 algorithms.[6] As can be seen, the pattern of contingencies is similar. Because of the sheer size of the expanded sense descriptions and contexts with this task, however, in the following analysis we stick to the simplified Lesk scenario.

As we hypothesized, using lexical expansions successfully bridges the lexical gap: whereas the basic simplified Lesk was able to make a sense assignment (be it correct or incorrect) in only 82.0% of cases, SL+100 could do so 99.6% of the time. SL+100 was able to correctly disambiguate over half of all the target words for which SL+0 failed to make any sense assignment. This contingency—some 9.7% of all instances—accounts for the majority of SL+100's improvement over SL+0. However, in 6.2% of cases SL+100's improvement resulted from successfully revising an incorrect answer of SL+0. We randomly selected ten of these cases and found that in all of them, all the overlaps for SL+0 were from a small number of non-content words (*the*, *of*, *in*, etc.), with the chosen sense achieving only one or two more overlaps than the runners-up; thus, the lexical gap is still at fault here. By contrast, the expanded sense definitions and contexts used by SL+100 for these cases always contained dozens of overlapping content words, and the overlap count for the chosen sense was markedly higher than for the runners-up.

What is also interesting to consider is the 0.2% of cases where both algorithms neglected to make a sense assignment, apparently signifying SL+100's failure to bridge the lexical gap. We manually examined all of these instances and found that for all but one, the systems failed to disambiguate the target words because the sentences containing them were extremely short, usually with no other content words apart from the target word. It is unlikely that any knowledge-based algorithm restricting itself to sentential context could succeed in such cases, and no reasonable number of lexical expansions is likely to help. Our choice to use sentential context was motivated by simplicity and expediency; a more refined WSD algorithm could, of course, using a sliding or dynamically sized context window and thereby avoid this problem.

[6]Totals in both tables may not add up exactly due to rounding.

The remaining case was a sentence of normal length where SL+0 found no overlapping content words between the definition and the context, but SL+100 produced a two-way tie between two of the clusters, one of which was the correct one.

It is also of interest to know why SL+0 was able to correctly disambiguate some words which SL+100 could not; these represent 2.8% of the instances. Again, we drew a random sample of these instances, and observed that in all of them, the only overlaps found by SL+0 were for non-content words; the fact that it happened to choose the correct sense cluster can therefore be chalked up to chance.

Though it has been relatively easy to identify the reasons behind SL+100's correct assignments, and behind its failures to make any assignment at all, it is not so easy to deduce the causes of its incorrect assignments. We observe that the system had disproportionate difficulties with verbs, which constitute 35% of the incorrect disambiguations but only 26% of all target words in the corpus. Particularly troublesome were verbs such as *be*, *go*, *have*, and *do*, which are often used as auxiliaries. On their own they contribute little or no semantic information to the sentence, and their dictionary definitions tend to explain their grammatical function, so there is little opportunity for meaningful lexical or conceptual overlaps. A related problem was observed for adverbs and adjectives: the problematic cases here were often generic terms of restriction, intensification, or contrast (e.g., *different*, *just*, *only*, *so*) which are used in a wide variety of semantic contexts and whose dictionary definitions focus on usage, or else constitute concise rephrasings using equally generic terms. Purely definition-based disambiguation approaches are unlikely to help in any of these cases; an accurate knowledge-based approach would probably need to be aware and make use of information beyond the lexical-semantic level, such as verb frames and semantic roles, or incorporate the grammatical structure around the target word for matching.

Conclusion and further work

We have proposed a new method for word sense disambiguation based on word overlap between sense descriptions and the target word context. Our method uses lexical expansions from a distributional thesaurus, which is computed over dependency-context similarities over a large background corpus. We found that applying our conceptually simple extension to two traditional knowledge-based methods successfully bridged the lexical gap, resulting in performance gains exceeding that of state-of-the-art knowledge-based systems that do not make use of sense frequency information, and approaching or even exceeding the MFS baseline. The concept of lexical expansion is a promising avenue to enrich classic, word-based NLP algorithms with additional lexical material. The intuitions of overlap-based approaches are thereby complemented by a method that makes associations explicit and bridges the lexical gaps for semantically similar contexts that are expressed in a different wording.

There are a number of ways how our method could be improved. First of all, since a DT is static and thus not dependent on the context, it generates spurious expansions, such as the similar terms for *charge* in Figure 1, which is obviously dominantly used in its "criminal indictment" sense in the background corpus. At best, these expansions, which implicitly capture the sense distribution in the background corpus, result in less overlap with the correct sense description—but they might well result in assigning incorrect senses. A straightforward improvement would alter the lexical expansion mechanism as to be sensitive to the context—something that is captured, for example, by LDA sampling (Blei et al., 2003). A further extension would be to have the number of lexical expansions depend on the DT similarity score (be it static or

contextualized) instead of the fixed number we used here.

In the future, we would like to examine the interplay of lexical expansion methods in WSD systems with richer knowledge resources (e.g., Navigli and Ponzetto (2010); Gurevych et al. (2012)) and apply our approach to other languages with fewer lexical resources. Also, it seems promising to apply lexical expansion techniques to text similarity, text segmentation, machine translation, and semantic indexing.

Acknowledgments

We thank Richard Steuer for computing and providing us access to the distributional thesaurus.

This work has been supported by the Hessian research excellence program *Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz (LOEWE)* as part of the research center *Digital Humanities*, and also by the Volkswagen Foundation as part of the Lichtenberg Professorship Program under grant № I/82806.

References

Agirre, E., Martínez, D., de Lacalle, O. L., and Soroa, A. (2006). Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing (TextGraphs-1)*, pages 89–96. (New York, NY, USA).

Anaya-Sánchez, H., Pons-Porrata, A., and Berlanga-Llavori, R. (2007). TKB-UO: Using sense clustering for WSD. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 322–325. (Prague, Czech Republic).

Banerjee, S. and Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012)*. (New Delhi, India).

Bär, D., Biemann, C., Gurevych, I., and Zesch, T. (2012). UKP: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval-2012)*, pages 435–440. (Montreal, Canada).

Biemann, C. (2012). Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*. DOI 10.1007/s10579-012-9180-5.

Biemann, C., Heyer, G., Quasthoff, U., and Richter, M. (2007). The Leipzig Corpora Collection: Monolingual corpora of standard size. In *Proceedings of the Corpus Linguistics Conference (CL2007)*. (Birmingham, United Kingdom).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cai, J. F., Lee, W. S., and Teh, Y. W. (2007). NUS-ML: Improving word sense disambiguation using topic features. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 249–252. (Prague, Czech Republic).

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluations (LREC 2006)*. (Genoa, Italy).

de Saussure, F. (1916). *Cours de linguistique générale*. Librairie Payot & Cie, Paris.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Gliozzo, A., Giuliano, C., and Strapparava, C. (2005). Domain kernels for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*, pages 403–410. (Ann Arbor, MI, USA).

Goyal, A., Daumé III, H., and Cormode, G. (2012). Sketch algorithms for estimating point queries in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP–CONLL 2012)*, pages 1093–1103. (Jeju Island, South Korea).

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). Uby – A large-scale unified lexical-semantic resource. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590. (Avignon, France).

Henrich, V., Hinrichs, E., and Vodolazova, T. (2012). WebCAGe – A Web-harvested corpus annotated with GermaNet senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 387–396. (Avignon, France).

Kilgarriff, A. (2001). English lexical sample task description. In *Proceedings of Senseval-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17–20. (Toulouse, France).

Kilgarriff, A. and Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities*, 34:15–48.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In *Proceedings of the Fifth Annual International Conference of Systems Documentation (SIGDOC '86)*, pages 24–26. (Toronto, Canada).

Li, L., Roth, B., and Sporleder, C. (2010). Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pages 1138–1147. (Uppsala, Sweden).

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL '98) and the 17th International Conference on Computational Linguistics (COLING 1998)*, volume 2, pages 768–774. (Montreal, Canada).

Martínez, D., López de Lacalle, O., and Agirre, E. (2008). On the use of automatically acquired examples for all-nouns word sense disambiguation. *Journal of Artificial Intelligence Research*, 33(1):79–107.

Mihalcea, R. (2008). Semcor 3.0. `http://lit.csci.unt.edu/~rada/downloads/semcor/semcor3.0.tar.gz`.

Mihalcea, R. and Chklovski, T. (2003). Open Mind Word Expert: Creating large annotated data collections with Web users' help. In *Proceedings of Fourth International Workshop on Linguistically Interpreted Corpora (LINC-03)*. (Budapest, Hungary).

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41:10:1–10:69.

Navigli, R., Litkowski, K. C., and Hargraves, O. (2007). SemEval-2007 Task 07: Coarse-grained English All-words Task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35. (Prague, Czech Republic).

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pages 216–225. (Uppsala, Sweden).

Navigli, R. and Velardi, P. (2005). Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1075–1086.

Palmer, M., Fellbaum, C., Cotton, S., Delfs, L., and Dang, H. T. (2001). English tasks: All-words and verb lexical sample. In *Proceedings of Senseval-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21–24. (Toulouse, France).

Palmer, M., Ng, H. T., and Dang, H. T. (2006). Evaluation of WSD systems. In Agirre, E. and Edmonds, P., editors, *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech, and Language Technology*. Springer.

Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pages 1522–1531. (Uppsala, Sweden).

Rapp, R. (2004). A freely available automatically generated thesaurus of related words. In *Proceedings of the Fourth International Conference on Language Resources and Evaluations (LREC 2004)*, pages 395–398. (Lisbon, Portugal).

Snyder, B. and Palmer, M. (2004). The English all-words task. In Mihalcea, R. and Edmonds, P., editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43. (Barcelona, Spain).

Tugwell, D. and Kilgarriff, A. (2001). WASP-Bench: A lexicographic tool supporting word sense disambiguation. In *Proceedings of Senseval-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 151–154. (Toulouse, France).

Vasilescu, F., Langlais, P., and Lapalme, G. (2004). Evaluating variants of the Lesk approach for disambiguating words. In *Proceedings of the Fourth International Conference on Language Resources and Evaluations (LREC 2004)*, pages 633–636. (Lisbon, Portugal).