Automated Verb Sense Labelling Based on Linked Lexical Resources

Kostadin Cholakov[1] **Judith Eckle-Kohler**[2,3] **Iryna Gurevych**[2,3]
[1] Humboldt-Universität zu Berlin, kostadin.cholakov@anglistik.hu-berlin.de
[2] Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Dept. of Computer Science, Technische Universität Darmstadt
[3] Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
http://www.ukp.tu-darmstadt.de

Abstract

We present a novel approach for creating sense annotated corpora automatically. Our approach employs shallow syntactico-semantic patterns derived from linked lexical resources to automatically identify instances of word senses in text corpora. We evaluate our labelling method intrinsically on SemCor and extrinsically by using automatically labelled corpus text to train a classifier for verb sense disambiguation. Testing this classifier on verbs from the English MASC corpus and on verbs from the Senseval-3 all-words disambiguation task shows that it matches the performance of a classifier which has been trained on manually annotated data.

1 Introduction

Sense annotated corpora are important resources in NLP as they can be used as training data (e.g., for word sense disambiguation (WSD) or semantic role labelling) or as sources for the acquisition of lexical information (e.g., selectional preference information). Typically, a particular sense inventory from a lexical resource is used to annotate some or all words with word senses from this sense inventory. For instance, various sense-annotated corpora based on WordNet (WN; (Fellbaum, 1998)) exist, such as the data from the Senseval competitions,[1] or the SemCor corpus.[2] Such corpora are usually created manually which is expensive and time consuming. Furthermore, the corpora are often domain specific (e.g. newspaper texts) which makes statistical systems trained on them strongly biased.

We present a novel approach for creating sense annotated corpora automatically. Our approach employs shallow syntactico-semantic patterns derived from linked lexical resources (LLRs) to automatically identify instances of word senses in text corpora. We significantly extend previous work on this task by making two important contributions: (i) we employ a large-scale LLR for automatically creating sense annotated data and (ii) we perform meaningful intrinsic and application-based evaluations of our method on large sense annotated datasets.

LLRs are the result of integrating several lexical-semantic resources by linking them at the word sense level. Examples of large LLRs are the multilingual BabelNet (Navigli and Ponzetto, 2012), an integration of wordnets and Wikipedia[3], or UBY, (Gurevych et al., 2012), the resource we employ in our work here. UBY is an integration of multiple resources, such as wordnets, Wikipedia, Wiktionary (WKT)[4], FrameNet (FN; (Baker et al., 1998)) and VerbNet (VN; (Kipper et al., 2008)) for English and German.

A distinguishing feature of LLRs is the enriched sense representation for word senses that are interlinked since different resources provide different, often complementary information. Annotating corpora with such enriched sense representations turns them into versatile training data for statistical systems.

Our first contribution (i) also addresses a considerable gap in recent research regarding automated sense labelling of verbs. Most previous work is done on nouns. However, verbs pose a bigger challenge due to their high polysemy and the fact that, unlike nouns, syntax is of crucial importance because it often reflects particular aspects of verb meaning. That is why, here we focus on verbs and present results and evaluations for this previously neglected part-of-speech (POS). Our method, however, can be applied to other parts-of

[1] http://www.senseval.org
[2] http://www.cse.unt.edu/~rada/downloads.html#semcor
[3] http://www.wikipedia.org
[4] http://www.wiktionary.org

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 68–77,
Gothenburg, Sweden, April 26-30 2014.

speech as well.

Regarding (ii), we are the first to perform meaningful intrinsic and extrinsic evaluations of automatically labelled data on a larger scale. The intrinsic evaluation measures the performance of our method on the manually annotated SemCor corpus. The extrinsic evaluation compares the performance of a classifier for verb sense disambiguation (VSD) which has been trained (a) on automatically sense labelled data and (b) on manually annotated data. Both settings achieve very similar results which means that competitive VSD can be performed without the need of costly manually created training data. This could be beneficial in languages (e.g., German, Spanish) for which elaborate lexical-semantic resources exist but large, high-quality sense annotated corpora are unavailable. Moreover, we experiment with various linkings between lexical resources in order to investigate how different resource combinations affect the performance of automated sense labelling. We show that combining all available resources might not be the best option.

The remainder of the paper is organised as follows. Section 2 presents our method. Section 3 describes the data used in the experiments. Section 4 presents the results of the evaluations. Section 5 analyses in detail the differences between our method and previous work. Section 6 concludes the paper.

2 Automated Labelling of Verb Senses

This section describes our novel approach for automated sense labelling of verbs in a corpus, which exploits the added value of LLRs.

2.1 Approach

Our approach to automatically label corpus instances of verb senses with sense identifiers from an LLR is based on a *pattern-based representation of verb senses*. Such patterns constitute a *common format* for the representation of verb senses available in LLRs and verb instances found in corpora. The common format we developed resembles a syntactico-semantic clause pattern which we call a sense pattern (SP). Based on a comparison of the derived SPs by means of a similarity metric, verb instances in a corpus can automatically be labelled with sense identifiers from an LLR.

SPs can be derived from corpus instances and from information given in LLRs, in particular, sense examples and more abstract predicate argument structure information.

2.2 Step 1: Creation of SPs from LLRs

For the creation of SPs, we employ the large-scale LLR UBY which combines 10 lexical resources for English and German to make use of the enriched verb sense representations provided by the sense links between various resources available in UBY. Although our method can work with any LLR, we choose UBY because the various resources are represented in a standardised format (Eckle-Kohler et al., 2012) and sense links between them can uniformly and conveniently be accessed via the freely available UBY-API.[5]

Since we evaluate our method on data annotated with WN senses, we create SPs for *enriched* WN senses (see example given in Table 1). We enrich WN senses by aggregating lexical information that can be accessed through links given in UBY to corresponding verb senses in other resources.

In this setting, enrichment means that we make use of sense examples from WN, from FN via the WN–FN linking, and from WKT via the WN–WKT linking. In addition, we use abstract predicate-argument structure information from VN via the WN–VN linking (see Table 1).[6]

For phrasal verb senses (e.g., *write up*) and other verbal multiword expressions (e.g., *know what's going on*) listed in WN, UBY rarely provides links to other resources. Therefore, we induced sense links by following the one sense per collocation assumption.[7] Based on this assumption, we linked each sense of a verbal multiword verb lemma in WN with each sense of the same multiword lemma in FN and WKT.

From sense examples, we derive two different kinds of SPs. Based on a fragment of a sense example given by a window w around the target verb lemma we create: (i) lemma SPs (LSPs) consisting only of lemmas (including the target verb) and (ii) abstract SPs (ASPs) consisting of the target verb lemma and items from a fixed, linguistically motivated vocabulary. This is based on the intuition that LSPs are important to identify relatively fixed

[5] `http://code.google.com/p/uby/`

[6] Although VN is linked to sense examples given in the PropBank corpus, the rationale behind using just abstract predicate-argument structure information was to explore, which effect this type of information has on the performance of an automated labelling algorithm.

[7] It assumes that nearby words provide strong and consistent clues to the sense of a target word, see Yarowsky (1995).

	WN sense tell%2:32:00:: (*let something be known*)	Corresponding sense patterns (SPs)
WN	*Tell them that you will be late*	LSP – **tell** them that you will be ASP – **tell** PP that PP be JJ
WN–FN	*But an insider told TODAY : ' There was no animosity.'*	LSP – but an insider **tell** Today : ' there be ASP – person **tell** location be feeling
WN–WKT	*Please tell me the time.*	LSP – Please **tell** me the time ASP – **tell** PP event
WN–VN	Agent[$+animate\| + organization$] V Recipient[$+animate\| + organization$] about Topic[$+communication$]	ASP – PP **tell** group about communication

Table 1: Examples of SPs derived from an enriched WN sense in UBY. PP, JJ, and VV are POS tags from the Penn Treebank tagset, standing for personal pronoun, adjective and full verb.

verbal multiword expressions in a corpus, whereas ASPs are necessary to identify productively used verb senses that are constrained in their use only by their syntactic behaviour and particular semantic properties, such as selectional preferences on their arguments.

The fixed vocabulary used for the creation of ASPs consists of (i) the target verb lemma, (ii) selected POS tags from the Penn Treebank Tagset (Marcus et al., 1993), (iii) a list of particular function words that play an important role in fine-grained subcategorisation frames of verbs (Eckle-Kohler and Gurevych, 2012) and (iv) semantic categories of nouns given by WN semantic fields. We selected POS tags that play an important role in syntactic realisations of verbs, e.g. POS tags for personal pronouns which are potential verb arguments. In our experiments, we tried different sets of function words and POS tags. For instance, we found that some function words (e.g., reflexive pronouns) and some POS tags (e.g., those for past participles and comparative adjectives) introduced too much noise in the data and therefore we did not select them for the final vocabulary.[8]

In order to create SPs from sense examples, we apply POS tagging and lemmatisation using the TreeTagger (Schmid, 1994) and named entity tagging using the Stanford Named Entity Recogniser (Klein et al., 2003). The named entity tags attached by the Named Entity Recogniser are mapped to WN semantic fields.

For the generation of ASPs from sense examples, we used a window size of $w = 7$, while the generation of LSPs has been performed with $w = 5$ in order to put a focus on the closely neighbouring lexemes in multiword verb lemmas. The window size was set empirically using the English Lexical Sample task of the Senseval-2 dataset as a development set. The same set was also used for the development of the linguistically motivated vocabulary for ASPs.[9]

From the abstract predicate-argument structure information given in VN, we derived only ASPs. For this, we employed the subcategorisation frames, as well as the semantic role and selectional preference information from VN, and created ASPs based on manually created mappings between these information types and the controlled vocabulary used for ASPs.

2.3 Step 2: Automated Labelling

For the automated labelling of verbs in a corpus, we first derive SPs from each corpus sentence containing a target verb. SPs are derived from corpus sentences by applying the same procedure as described in Step 1 for the creation of SPs from sense examples, the window size used is $w = 7$.

To compare two SPs, we propose a similarity metric based on Dice's coefficient which calculates the sum of the weighted number of their common bi-grams, tri-grams, and four-grams. Formally, the similarity score $sim_w \in [0..1]$ of two SPs p_1, p_2 is defined as:

$$(1) \qquad sim_w(p_1, p_2) = \frac{\sum_{n=2}^{4} |G_n(p_1) \cap G_n(p_2)| \cdot n}{norm_w}$$

where $w >= 1$ is the size of the window around the target verb, $G_n(p_i), i \in \{1, 2\}$ is the set of n-

[8] The vocabulary used for the creation of ASPs is available at `http://www.ukp.tu-darmstadt.de/data/`.

[9] However, the Senseval-2 data are annotated with sense keys of the WN pre-release version 1.7 and therefore, we had to employ an automated mapping of WN 1.7 pre-release to WN 3.0 sense keys provided by Rada Mihalcea. Since this mapping turned out to be rather noisy, we did not use the Senseval-2 data in our evaluations.

Automated labelling of corpus instances

for each sentence s_i with verb v
  derive LSP_i and ASP_i

  forall $j = sizeOf$(UBY-LSP(v))
  compare LSP_i with LSP_j in UBY-LSP(v):
  $maxSim(LSP_i) = argmax_j score(LSP_i, LSP_j)$
  add sense($argmax_j$) to MostSimilarSenses(LSP_i)

  forall $k = sizeOf$(UBY-ASP(v))
  compare ASP_i with ASP_k in UBY-ASP(v):
  $maxSim(ASP_i) = argmax_k score(ASP_i, ASP_k)$
  add sense($argmax_k$) to MostSimilarSenses(ASP_i)

  if $maxSim_{i,j} >=$ threshold t and
    $maxSim_{i,j} >= maxSim_{i,k}$
    label(s_i) = random(MostSimilarSenses(LSP_i))
  else if $maxSim_{i,k} >=$ threshold t
    label(s_i) = random(MostSimilarSenses(ASP_i))
  end if
end for

Table 2: Algorithm for labelling corpus instances with WordNet senses.

grams occurring in SP p_i, and $norm_w$ is the normalisation factor defined by the sum of the maximum number of common bigrams, trigrams and fourgrams in the window w. Similarity metrics based on Dice's coefficient have often been used in Lesk-based WSD (Lesk, 1986) to calculate the overlap of two sets (e.g., Baldwin et al. (2010)). In our case, however, the elements of the two sets are bigrams, trigrams and fourgrams, while in Lesk-based algorithms typically sets of unigrams are compared, thus not accounting for word order.

Table 2 shows the algorithm used for automated labelling of corpus instances in pseudo-code. The algorithm assumes that for each verb v, the corresponding set of SPs derived from UBY sense examples (UBY-LSP(v) and UBY-ASP(v) in Table 2) has already been computed.

For each corpus sentence containing a target verb v, the corresponding SPs for verb v derived from UBY are scored by the similarity metric in (1). The SPs with the maximum score that is above a threshold t form the set of most similar senses. From this set, the algorithm picks one sense randomly as a label. How often this happens, depends on the value of t: the percentage of randomly selected senses ranges from about 33% for $t = 0.14$ to about 50% for $t = 0.04$.

3 Data

Web corpora. For the automated labelling of corpus data with WN senses, we use two very large web corpora: the English ukWaC corpus (Baroni et al., 2009) and the article pages extracted from the English Wikipedia using the Java-based Wikipedia API JWPL (Zesch et al., 2008). Further, for the evaluation of our method, we use three manually sense annotated data sets.

SemCor. We use the SemCor 3.0 corpus which is annotated with WN 3.0 senses.

MASC. MASC is a balanced subset of 500K words of written texts and transcribed speech drawn primarily from the Open American National Corpus (OANC).[10] The texts come from 19 different genres which allows us to test our method on real-life data from multiple sources. The corpus is annotated with various types of linguistic information, including WN 3.0 sense annotations for instances of selected words. Therefore, MASC is a *lexical sample* corpus.

We extracted instances of 16 MASC verbs (11,997 instances) which have been sense annotated. Most instances are annotated by multiple annotators and, to create a gold standard, we took the sense preferred by the majority of annotators and ignored instances where there were ties.

Senseval-3. In the test corpus of the Senseval-3 all-words disambiguation task sense annotations are provided for each content word in a chunk of the WSJ corpus (5,000 words of running text). The third annotated data set for our experiment is formed by extracting all verb instances from this test corpus. Note that the gold standard annotations in Senseval-3 were made using WN 1.7.1. In our experiments, we use Rada Mihalcea's conversion of the corpus to WN 3.0.[11] However, we found out that some verb instances were converted to sense labels that do not exist in WN 3.0. After removing those instances, there were 305 verbs with 592 instances left.

4 Experiments and Evaluation

Next, we present the intrinsic and the application-based evaluations of our method.

4.1 Intrinsic Evaluation

We intrinsically evaluate the performance of the automated labelling algorithm for the Senseval-3 verbs which occur in the SemCor corpus. Occurrences of these 152 verbs in SemCor are processed

[10] http://www.americannationalcorpus.org/

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

	WN–FN–WKT			WN–FN–WKT–VN		
t	Cov (Inst.)	Cov (Sense)	Acc	Cov (Inst.)	Cov (Sense)	Acc
0.04	0.55	0.27	0.32	0.48	0.25	0.35
0.07	0.15	0.17	0.36	0.13	0.15	0.42
0.1	0.11	0.14	0.35	0.10	0.13	0.42
0.14	0.02	0.07	0.41	0.02	0.05	0.47

Table 3: Performance of the automated labelling algorithm evaluated for occurrences of Senseval-3 verbs in SemCor.

by the labelling algorithm with a window size $w = 7$ and the automatically annotated WN 3.0 senses are compared with the gold senses available in SemCor 3.0.

Quantitative Evaluation. We calculated the accuracy as the percentage of correctly labelled instances and the instance coverage as the percentage of labelled instances. The sense coverage is calculated as the percentage of all predicted (not annotated) senses relative to all gold verb senses given in SemCor.

A random sense baseline yields 15% accuracy. Note that a MFS baseline based on WN would not be meaningful, because the WordNet MFS is based on the frequency distribution of annotated senses in SemCor.

Table 3 shows accuracy and coverage results of the automated labelling algorithm for different values of the threshold t and two combinations of sense links from UBY. Depending on the threshold t, 2% to 55% of the verb instances in SemCor can automatically be labelled, and the instance coverage goes largely in parallel to the coverage of predicted WN senses. Accuracy ranges between 32% and 47% and exceeds the random sense baseline by a large margin. Lowering the threshold increases the coverage of the labelling method, but it also leads to a decrease in accuracy of 9 percentage points (12 for the configuration with VN).

Adding more patterns from VN via the WN–VN alignment, leads to a decrease in both instance and sense coverage combined with an increase in accuracy. Since SemCor is a rather small corpus, the increase in instance coverage is not as clear as for large Web corpora such as the ukWaC corpus. Labelling a 1GB subset of the ukWaC corpus based on patterns derived from the WN–FN–WKT alignments resulted in 15MB of labelled data, whereas 25MB labelled data could be created from the same subset with the additional patterns from the WN–VN alignment.

Qualitative Analysis. In Table 4, we show examples of the highest ranking patterns and the corresponding labelled SemCor instances for senses that were correctly and falsely annotated. The examples in Table 4 show that the similarity metric assigns the highest values to instances where function words (e.g., *in, to, who*) or POS tags (e.g., PP, VV) from the ASP vocabulary occur in the immediate neighbourhood of the target verb. Since such functions words play an important role in the ASPs derived from VN, the VN ASPs possibly tend to dominate over the SPs derived from sense examples, which explains the observed decrease in coverage (see Table 3).

The falsely labelled instances turn out to be examples of WN senses where the gold sense is very similar to the automatically attached sense as evident from the synset definition given in the rightmost column.

4.2 Extrinsic Evaluation

We extrinsically evaluate our method for automated verb sense labelling by using it for learning a classifier for VSD in a train-test setting. We use features which have been widely used in supervised WSD systems, in particular features based on dependency parsing. While this might seem to be in contrast to our labelling algorithm which is based on shallow linguistic preprocessing, it is fully justified by the purpose of our extrinsic evaluation: The main purpose of the extrinsic evaluation is not to outperform state-of-the-art VSD systems, but to show that, when operating with reasonable features, a classifier trained on the data automatically labelled with our method performs equally well as when this classifier is trained on manually annotated data.

4.2.1 Features

The training and test data are parsed with the Stanford parser (Klein and Manning, 2003) which provides Stanford Dependencies output (De Marneffe et al., 2006) as well as phrase structure trees. We employ the Stanford Named Entity Recogniser to identify named entities. We then extract lexical, syntactic, and semantic features from the parse results for classification.

Lexical features include the lemmas and POS tags of the two words before and after the target verb. To extract **syntactic features** we select all dependency relations from the parser output in

SemCor instance	SP derived from SemCor	score	WN sense ID (gold sense in brackets)
Some of the New York Philharmonic musicians who **live** *in the suburbs spent yesterday morning digging themselves free from snow.*	of group person who **live** in location VVD time time VVG	0.29	live%2:42:08:: (live%2:42:08::)
These societies can **expect** *to face difficult times.*	group **expect** to VV JJ event	0.22	expect%2:31:01:: (expect%2:31:01::)
As autumn starts its annual sweep , few Americans and Canadians **realize** *how fortunate they are in having the world 's finest fall coloring.*	JJ attribute JJ person **realize** how JJ PP be in	0.22	realize%2:31:00:: – perceive (an idea or situation) mentally (realize%2:31:01:: – be fully aware or cognizant of)
Dan Morgan told himself he would **forget** *Ann Turner.*	person person VVD PP PP **forget** person location	0.16	forget%2:31:00:: – be unable to remember (forget%2:31:01:: – dismiss from the mind; stop remembering)

Table 4: Examples of SemCor instances with high similarity scores (upper half shows correctly labelled instances, lower half incorrectly labelled instances.

which the target verb is related to a noun, a pronoun, or a named entity. For each selected word, the lemma of the word (or the named entity tag in case of proper nouns) is combined with the type of the dependency relation which exists between it and the verb to form a separate feature. In a similar feature, the lemma of the selected word is replaced by its POS tag. The **semantic features** include all synsets found in WN for nominal arguments of the verb. Personal pronouns are mapped to 'person' and the three synsets found in WN 3.0 for this word are taken as features.

4.2.2 Train and Test Data

Using exactly the same method as intrinsically evaluated in section 4.1, we automatically labelled occurrences of the 16 MASC verbs and the 305 Senseval-3 verbs in both web corpora with WN senses. Only occurrences with similarity score above 0.1 are labelled – all other occurrences are discarded. We refer to the resulting data as *automatically labelled corpus* (ALC) and use it as training data for statistical VSD.

Instances of the test verbs found in SemCor are also used as training data in order to compare the performance of the classifier in a fully supervised setting.

MASC. There are 22 senses with instances in MASC which are not found in SemCor. For the ALC this number is 34. However, in the latter there are 27 senses, instances of which are unseen in MASC. 20 of those represent phrasal verbs which we attribute to the special treatment of such verbs in our method.

The classifier cannot correctly classify senses which are not seen in the training data. The coverage of the ALC is 88.05% and that of SemCor — 94.8%. The SemCor data can mainly cover more test instances of 3 verbs — *launch*, *rule*, and *transfer* — the WN senses of which lack sense examples or links to other senses in UBY. Unlike the hand-labelled SemCor data, our automated sense labelling method is limited to the information found in the LLR used. However, there are also 330 MASC instances covered by the ALC only. Those are mostly instances of phrasal verbs, such as *rip off* and *show up*. Note that the definition of coverage we use here makes its values the upper bounds for the performance of the classifier.

Senseval-3. We also generated training data automatically for the 305 Senseval verbs. However, only 152 of those verbs (442 instances) are found in SemCor. This means we cannot train the classifier for the remaining Senseval verbs. The coverage of the SemCor training data for the 152 verbs which can be classified is 96.15% and that of the ALC — 95.25%. For all 592 Senseval test instances, the coverage of the ALC is 90.38%.

4.2.3 Results and Analysis

We trained a separate logistic regression classifier for each test verb in the two datasets using the WEKA data mining software (Hall et al., 2009) with default parameters. The classifiers were trained with features extracted from (i) the SemCor hand-labelled data and (ii) the ALC.

MASC. The classifier achieves 50.23% accuracy when SemCor is used and 49% when the ALC is employed. The difference in the results is not statistically significant at $p < 0.05$. The MFS

baseline scores at 41.72%.

Senseval-3. The classifier achieves 43.24% with the ALC. We assigned the MFS to each of the 143 test verbs not found in SemCor since we cannot train the classifier for those. The achieved accuracy is 45.2%. We also measured accuracy in a setup where no MFS back-off strategy was employed for SemCor (152 test verbs with 442 instances). When trained on SemCor data, the classifier achieves 48.64% accuracy compared to 47.51% for the ALC. All differences in the results are not statistically significant at $p < 0.05$. Finally, the MFS baseline accuracy is significantly lower at 25.34% for all 305 test verbs.

For both test datasets, the overall performance of the classifier when trained on automatically labelled data is very close to the setting in which manually created training data is employed. We thus conclude that the quality of the data produced by our sense labelling method is sufficient and these data can be directly used for training a statistical VSD classifier. As a reference, the state-of-the-art supervised VSD system described in Chen and Palmer (2009) achieves 64.8% accuracy on the Senseval-2 fine-grained data. However, we cannot compare to this result due to the different sense inventory which the Senseval-2 data were annotated with.

4.2.4 Sense Links

In order to investigate the effect of LLRs, we performed experiments in which sense examples found in WN only were used. We also experimented with various combinations of the resources available in UBY to determine the contribution of each of those to our method. Table 5 shows the results. The setting which includes only WN has the worst performance, thus clearly showing the benefits of using LLRs. Next, the inclusion of WKT improves both coverage and accuracy. We conclude that WKT plays an important role in discovering additional verb senses. Finally, similarly to the results of the intrinsic evaluation, adding VN to the mix increases slightly the coverage but decreases accuracy.

	MASC		Senseval	
	Cov	Acc	Cov	Acc
WN	0.6573	0.3498	0.6372	0.3209
WN–FN	0.8562	0.4810	0.8812	0.4172
WN–FN–WKT	0.8805	0.4900	0.9038	0.4324
WN–FN–WKT–VN	0.8822	0.4688	0.9139	0.4054

Table 5: Performance of the various combinations of lexical resources.

5 Related Work and Discussion

Our work is related to previous research on (i) using a combination of lexical resources for knowledge-based WSD, (ii) using lexical resources for distant supervision, and (iii) the automated acquisition of sense-annotated data.

Knowledge-based WSD. While the combination of sense-annotated data and wordnets has been described for knowledge-based WSD before (e.g., Navigli and Velardi (2005; Agirre and Soroa (2009) who use graph algorithms), only recently Ponzetto and Navigli (2010) have investigated the impact of the combination of different *lexical resources* on the performance of WSD. They aligned WN senses with Wikipedia articles and employed two simple knowledge-based algorithms, i.e., a Lesk-based algorithm and a graph-based algorithm, to evaluate the resulting LLR for WSD. While their evaluation demonstrates that the use of an LLR boosts the performance of knowledege-based WSD, it is restricted to nouns only since Wikipedia provides very few verb senses. Moreover, lexical resources that are rich in lexical-syntactic information such as VN have not been involved.

Miller et al. (2012) employ a Lesk-based algorithm which makes use of a combination of WN and an automatically acquired distributional thesaurus. Lesk-based algorithms play a central role in knowledge-based WSD. Based on the overlap of the context of the target word and sense definitions in a given sense inventory, they assign the sense with the highest overlap as disambiguation result. We were kindly provided with the system described in Miller et al. (2012) and we were able to test its performance on our test sets. The system achieved only 33.86% and 30.16% accuracy for the MASC and the Senseval-3 verbs, respectively, which is far below the results we presented. This low performance is due to the fact that Lesk-based algorithms do not account for word order. Such information is important especially for verb senses, as the syntactic behaviour of a verb reflects aspects of its meaning.

Distant supervision. Distant supervision is a learning paradigm similar to semi-supervised learning. Unlike semi-supervised methods which typically employ a supervised classifier and a

small number of seed instances to do bootstrap learning (Yarowsky, 1995; Mihalcea, 2004; Fujita and Fujino, 2011), in distant supervision training data are created in a single run from scratch by aligning corpus instances with entries in a knowledge base. Distant supervision methods that have used LLRs as knowledge bases have been previously applied in relation extraction, e.g. Freebase (Mintz et al., 2009; Surdeanu et al., 2012) and BabelNet (Krause et al., 2012; Moro et al., 2013). However, as far as we are aware, we are the first to apply distant supervision to the task of verb sense disambiguation.

Acquisition of sense-annotated data. Most previous work on using lexical resources for automatically acquiring sense-annotated data either was mostly restricted to noun senses or, unlike us, did not present a meaningful evaluation. Leacock et al. (1998) describe the automated creation of training data for supervised WSD on the basis of WN as a lexical resource combined with corpus statistics, but they evaluate their approach just on one noun, verb, and adjective, and thus it is unclear whether their results can be generalized. Cuadros and Rigau (2008) used the approach of Leacock et al. (1998) to automatically build a large *KnowNet* from the Web, but they evaluated this resource only for WSD of nouns. However, the system based on KnowNet yields results below the SemCor-MFS baseline. Mihalcea and Moldovan (1999) use WordNet glosses to extract sense examples from the Web via a search engine and use this approach in a subsequent paper (Mihalcea, 2002) to generate a sense tagged corpus. For five randomly selected nouns, they performed a comparative evaluation of a WSD classifier trained on an automatically tagged corpus on the one hand, and on the manually annotated data from the Senseval-2 English lexical sample task on the other hand. The results obtained for these five nouns seem to be similar but the dataset used is too small to draw meaningful conclusions and moreover, it does not cover verbs. Mostow and Duan (2011) presented a system that extracts example contexts for nouns and apply these contexts in (Duan and Yates, 2010) for WSD by using them to label text and train a statistical classifier. An evaluation of this classifier yielded results similar to those obtained by a supervised WSD system.

Kübler and Zhekova (2009) extract example sentences from several English dictionaries and various types of corpora, including web corpora. They employ a Lesk-based algorithm to automatically annotate the target word instances in the extracted example sentences with WN senses and use them in one of their experiments as training data for a WSD classifier. However, the performance of the system decreased significantly achieving the lowest accuracy among all system configurations. The authors provide only the overall accuracy score, so we do not know how disambiguation of verbs was affected.

Summary. We consider the ability to establish a link between the rich knowledge available in LLRs and corpora of any kind to be the main advantage of our automated labelling method. However, to automatically label a suffcient amount of data for supervised learning, very large corpora are required. Our method can be extended to other POS (using sense examples and possibly other types of lexical information), as well as to other languages where (linked) lexical resources are available.

6 Conclusion

In this paper, we presented a novel method for creating sense labelled corpora automatically. We exploit LLRs and perform large-scale intrinsic and application-based evaluations. The results of those evaluations show that the quality of the sense labelled corpora created with our method matches that of manually annotated corpora.

In future research, we plan to use PropBank (Palmer et al., 2005) in order to extract sense examples for VN as well. This might improve the performance of lexical resource combinations which include VN. We will also apply our method to languages (e.g., German) for which lexical resources are available but no or little sense annotated corpora exist.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg- Professorship Program under grant No. I/82806 and by the German Research Foundation under grant No. GU 798/9-1. We would like to thank the anonymous reviewers for their valuable feedback.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 33–41, Athens, Greece.

C.F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90, Montreal, Canada.

Timothy Baldwin, Sunam Kim, Francis Bond, Sanae Fujita, David Martinez, and Takaaki Tanaka. 2010. A Reexamination of MRD-Based Word Sense Disambiguation. *ACM Transactions on Asian Language Information Processing (TALIP)*, 9(1):4:1–4:21.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Jinying Chen and Martha Palmer. 2009. Improving English verb sense disambiguation performance with linguistically motivated features and clear sense distinction boundaries. *Language Resources and Evaluation*, 43:181–208.

Montse Cuadros and German Rigau. 2008. Knownet: Building a large net of knowledge from the web. In *22nd International Conference on Computational Linguistics (COLING)*, pages 161–168, Manchester, UK.

M.C. De Marneffe, B. MacCartney, and C.D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 449–454, Genoa, Italy.

Weisi Duan and Alexander Yates. 2010. Extracting glosses to disambiguate word senses. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 627–635, Los Angeles, USA.

Judith Eckle-Kohler and Iryna Gurevych. 2012. Subcat-LMF: Fleshing out a standardized format for subcategorization frame interoperability. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 550–560, Avignon, France.

Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. 2012. UBY-LMF – A uniform format for standardizing heterogeneous lexical-semantic resources in ISO-LMF. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 275–282, Istanbul, Turkey.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Sanae Fujita and Akinori Fujino. 2011. Word sense disambiguation by combining labeled data expansion and semi-supervised learning method. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 676–685, Chiang Mai, Thailand.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY - a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42:21–40.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430, Sapporo, Japan. Association for Computational Linguistics.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 180–183, Edmonton, Canada.

Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. 2012. Large-scale learning of relation-extraction rules with distant supervision from the web. In *Proceedings of the 11th International Semantic Web Conference*, pages 263–278, Boston, Masachusetts, USA, 11. Springer.

Sandra Kübler and Desislava Zhekova. 2009. Semi-Supervised Learning for Word Sense Disambiguation: Quality vs. Quantity. In *Proceedings of the International Conference RANLP-2009*, pages 197–202, Borovets, Bulgaria.

Claudia Leacock, George A. Miller, and Martin Chodorow. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26, Toronto, Ontario, Canada.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.

Rada Mihalcea and Dan Moldovan. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of the American Association for Artificial Intelligence (AAAI 1999)*, Orlando, Florida, USA.

Rada Mihalcea. 2002. Bootstrapping large sense tagged corpora. In *Proceedings of the Third International Conference of Language Resources and Evaluation (LREC 2002)*, pages 1407–1411, Las Palmas, Canary Islands, Spain.

Rada Mihalcea. 2004. Co-training and self-training for word sense disambiguation. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, MA, USA.

Tristan Miller, Chris Biemann, Torsten Zesch, and Iryna Gurevych. 2012. Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1781–1796, Mumbai, India.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore.

Andrea Moro, Hong Li, Sebastian Krause, Feiyu Xu, Roberto Navigli, and Hans Uszkoreit. 2013. Semantic rule filtering for web-scale relation extraction. In *Proceedings of the 12th International Semantic Web Conference*, Sydney, Australia, 10. Springer.

Jack Mostow and Weisi Duan. 2011. Generating example contexts to illustrate a target word sense. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 105–110, Portland, Oregon, June. Association for Computational Linguistics.

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

Roberto Navigli and Paola Velardi. 2005. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1075–1086.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–105.

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

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, volume 8, pages 1646–1652, Marrakech, Morocco.