Addressing polysemy in bilingual lexicon extraction from comparable corpora

Darja Fišer[1], Nikola Ljubešić[2], Ozren Kubelka[3]

[1]Department of Translation, Faculty of Arts, University of Ljubljana
[2]Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
[3]University of Applied Sciences Vern, Zagreb
darja.fiser@ff.uni-lj.si, nikola.ljubesic@ffzg.hr, ozren.kubelka@vern.hr

Abstract

This paper presents an approach to extract translation equivalents from comparable corpora for polysemous nouns. As opposed to the standard approaches that build a single context vector for all occurrences of a given headword, we first disambiguate the headword with third-party sense taggers and then build a separate context vector for each sense of the headword. Since state-of-the-art word sense disambiguation tools are still far from perfect, we also tried to improve the results by combining the sense assignments provided by two different sense taggers. Evaluation of the results shows that we outperform the baseline (0.473) in all the settings we experimented with, even when using only one sense tagger, and that the best-performing results are indeed obtained by taking into account the intersection of both sense taggers (0.720).

Keywords: bilingual lexicon extraction; comparable corpora; polysemy

1. Introduction

Automatic extraction of bilingual lexica is still a major bottleneck for many NLP applications for most languages and most domains. Using comparable corpora for finding translation equivalents has become increasingly popular in the past two decades. The main idea behind this approach is the assumption that a source word and its translation appear in similar contexts in their respective languages, so that in order to identify them their contexts are compared via a seed dictionary (Fung, 1998; Rapp, 1999).

Our earlier work shows that this approach gives good results for a specialized domain even though the seed dictionary is quite small (Fišer et al., 2011). What is more, if we are extracting translation pairs for closely related languages, we have shown that the same quality of the results can be achieved by exploiting the lexical overlap between the languages instead of using a seed dictionary (Ljubešić and Fišer, 2011).

However, all our attempts as well as most of the related work have so far neglected the issue of polysemy and considered the translation candidate to be correct if it was an appropriate translation for at least one possible sense of the source word. Manual evaluation of the results from our previous research has shown that in most cases the translation candidates of a polysemous word are all related to the most frequent sense of the polysemous source word because of the Zipfian distribution of senses for which the majority of data in the context model come from the most frequent sense. Thus, the goal of this paper is to refine the approach in order to be able to extract translations for the other senses of the polysemous words as well.

Distributional methods for word sense acquisition have been proposed before; Pantel and Lin (2002) for example produce overlapping clusters so that a polysemous word is assigned to multiple clusters, each of which represents one of its senses. In the bilingual setting, Kaji (2003) uses word clustering to extract sets of synonymous translation equivalents from comparable corpora.

However, the proposed method produces a hierarchy of clusters and it is far from trivial to terminate the merging of the senses automatically. More importantly, the method assumes that each translation equivalent represents only one target word, which is of course not always the case. And finally, the resulting clusters are very coarse-grained, which is not always useful, especially if the languages in question are very different and do not share the distribution of polysemy in lexicalizations of the concepts.

This is why we propose a somewhat different approach that relies on Princeton WordNet (PWN) as the sense inventory and uses third-party word-sense disambiguation algorithms to split the occurrences of a polysemous word into several groups, and build context vectors separately for each one of them. Then, vector features are translated into the target language with a seed dictionary and compared with all the vectors in the target language in order to find the most similar one, ideally the one that best captures that particular sense of the source word.

This paper is structured as follows: in the next section we present the resources we used in this research. In Section 3 we give a full account of the experimental setup for this research. In Section 4 we evaluate and discuss the results, and then conclude the paper with some concluding remarks and ideas for future work.

2. Resources and tools used

In this research we used two web corpora for Slovene and English in order to extract contextual information about polysemous words. Based on their contexts, each occurrence of the selected English polysemous words was automatically disambiguated with two different word-sense disambiguation tools. Two sense inventories with different levels of sense granularity were used to assign an appropriate sense to each occurrence of a polysemous word and a traditional bilingual English-Slovene dictionary was used

to translate contexts of English words into Slovene in order to find their most similar equivalents.

2.1. Corpora

Contextual information for English words was extracted from ukWaC (Baroni et al., 2009), a large corpus of English that was built by crawling the .uk Internet domain within the WaCky initiative. The corpus contains more than 2 billion tokens and is one of the largest freely available linguistic resources for English. The corpus contains basic linguistic annotation (part-of-speech tagging and lemmatization) and serves as a general-purpose corpus of English, comparable in terms of document heterogeneity to traditional balanced resources. The corpus is among the largest resources of its kind, and the only web-derived, freely available English resource with linguistic annotation.

For Slovene, its younger and smaller counterpart slWaC (Ljubešić and Erjavec, 2011) was used. It was built in parallel for Slovene and Croatian (hrWaC being the resulting corpus for that language, 1.2BW in size) using a modified WaCky pipeline that focuses on the limited amount of available web data. While trying to capture as much data as possible, the approach is still rigorous concerning the content it extracts from web pages through an updated boilerplate removal method. The corpus contains 380 million tokens and has been part-of-speech tagged and lemmatized. The corpus is freely available for research.

2.2. Lexical resources

The main lexical resource in this research is Princeton WordNet (Fellbaum, 1998), a large lexical database of English in which nouns, verbs, adjectives and adverbs (literals) are grouped into sets of synonyms (synsets), each expressing a distinct concept. Synsets are interlinked with semantic and lexical relations. The latest version of Princeton WordNet contains about 117.000 synsets and 155.000 unique literals.

However, because Princeton WordNet has been criticized for having too fine-grained sense distinctions that make it hard for NLP applications to use it as efficiently as desired, we also used the more coarse-grained Sense Inventory (Navigli, 2006) which contains automatically generated clusters of PWN senses that were obtained via a mapping to the Oxford Dictionary of English (ODE), a long-established dictionary which encodes coarse sense distinctions. The Sense Inventory has been successfully used in the Coarse-grained English all-words word-sense disambiguation task at the SemEval-2007 workshop.

For translating features of English context vectors in to Slovene, we used a traditional, medium-sized English-Slovene bilingual dictionary (Grad et al., 1999). It contains 41,405 content-word entries. Our previous experiments showed that using just the first (probably the most frequent) translation yields better results than using all translations with different weighting schemes.

2.3. WSD tools

Finally, two different freely available word-sense disambiguation tools were used: UKB (Agirre and Soroa, 2009) and WordNet::SenseRelate::AllWords (Pedersen and Kolhatkar, 2009). They both use Princeton WordNet (PWN) to assign a sense to each occurrence of the headword in the sentence. The UKB system uses the Personalized PageRank algorithm while the WordNet::SenseRelate::AllWords system maximizes the semantic relatedness of the headword and its context on WordNet::Similarity.

3. Experimental setup

In this section we present the preprocessing steps and the procedure for finding translation equivalents for different senses of polysemous words in comparable corpora. First, a sample of polysemous English words on which we test the proposed approach is selected, then the preprocessing steps, such as word-sense disambiguation of the sample words occurrences and their mapping to the coarse-grained Sense Inventory are explained and finally, the building and comparison of context vectors is described.

3.1. Lexical sample

In this pilot study, we test the proposed approach on finding Slovene translations for 8 English polysemous words that we divided into two groups according to their level of polysemy. At this stage we have limited ourselves to nouns only because they are treated best in Wordnet and because the best results of the state-of-the art word-sense-disambiguation tools are achieved for nouns. Nevertheless, the approach can be easily extended to other parts of speech in the future.

It is a known fact in the field of lexical semantics that some senses of polysemous words can be easily distinguished one from another, and are also used in very different ways, which is why they are relatively easy to tell apart by comparing their contexts. For example, the noun bat can refer to either the animal or the sports equipment. It is usually quite simple to say which one was meant by looking at the context in which the word has been used. Such senses are therefore considered here as examples of easy polysemy (see column A in Table 1). On the other hand, senses of polysemous words are sometimes much closer, either because they are related to each other or because a clear boundary between them cannot be drawn. For example, the word bass can, among others, mean the lowest adult male singing voice or an adult male singer with the lowest voice which are related concepts and are therefore used in similar contexts. This makes them much harder to decide which one exactly was meant in a particular instance by just looking at the surrounding words. We treat such words as examples of difficult polysemy (see column B in Table 1).

Easy words	Difficult words
bat	bass
nail	body
organ	game
present	object

Table 1: List of the English polysemous nouns included in the lexical sample

3.2. Pre-processing steps

In order to be able to carry out the experiment, we first had to extract contexts of the words from the sample from large corpora, which will serve as the basis for finding translation equivalents. All occurrences of the selected headwords were extracted from a random subset of 200 million tokens from the ukWaC corpus. For finding their translation candidates in Slovene we used a sample of the same size from slWaC. Both corpora had already been POS-tagged and lemmatized.

Next, we sense-tagged all the occurrences of the selected headwords with two different freely available word-sense disambiguation tools: UKB and WordNet::SenseRelate::AllWords that both assign to each occurrence of the analyzed polysemous word a sense from Princeton WordNet.

Because Princeton WordNet has often been criticized for having too fine-grained sense distinctions we have also mapped the results of the sense-tagging to the more coarse-grained Sense Inventory which contains automatically generated clusters of PWN.

3.3. Building and comparing vectors

Once the tagging was completed we built feature vectors for each sense of the polysemous headword. English vector features were translated into Slovene with the English-Slovene bilingual dictionary, and compared to all the context vectors for 8,760 Slovene nouns occurring in the Slovene web corpus more than 50 times.

For translating each context feature in the English vectors we used just the first translation from the seed dictionary since in our previous research this approach showed best results.

As a baseline we use our original method where no WSD on headword occurrences is performed, but for each headword only one vector is built based on all occurrences of the headword. The result of the baseline approach is a ranked list of candidate translations for each headword regardless of its different meanings.

In order to examine the impact of the quality of sense tagging on the final results, we built the vectors with 3 different settings:

1. using sense tags assigned by the UKB tool,
2. using sense tags assigned by the SenseRelate tool, and
3. using only those contexts and respective sense tags where the two tools give the same answer.

The third setting has two possible implications:

- it could achieve higher accuracy of sense tagging since the two WSD approaches are quite different, and
- peripheral senses, that are often the hardest ones to distinguish, could get filtered out since on them the two WSD tools could often not agree on.

Throughout the experiment we used the same settings for building context vectors:

- the minimum frequency of occurrences with the same sense tag was 50,
- the context window for building feature vectors was 3 content words to the left and right, not taking into account their position,
- TF-IDF was used as the feature weighting measure, and
- Dice coefficient was used as the similarity measure.

4. Evaluation of the results

In this section we report on the results obtained with the procedure explained above. The evaluation, which was completely manual due to the lack of appropriate gold standards, consists of two parts. Since the quality of the entire procedure heavily depends on the quality of the sense tagging, we first manually evaluated a sample of the word-sense disambiguation step. Then, we also evaluated the output of the translation equivalent extraction step, which was the main focus in our experiment.

4.1. Evaluation of sense tagging

In order to gain insight into the suitability of the proposed method which heavily depends on sense-tagging tools, we first manually evaluated 10 random occurrences of each automatically assigned sense by each sense tagger respectively and the quality of the same number of occurrences that were annotated with the same sense by both tools.

As Table 2 shows, the selected headwords had 33 senses in the coarse-grained Sense Inventory, 11 for easy words and 22 for the hard ones. Not all the senses were assigned by the sense taggers, at least not with the frequency required for building context vectors, which means that we were only able to look for translation candidates of some of the senses. The easy words are represented by 10 different senses and the difficult ones with 17 when disambiguated with just one of the tools, while the number of such senses goes down to 9 for the easy words and 10 for the difficult ones when only the intersection between the two tools is kept. This means that the total number of senses used for translation in the final phase of the experiment is 19 or 57% of the initial set.

WSD setting	Easy words	Difficult words	Total
Sense Inv.	11 (100%)	22 (100%)	33 (100%)
UKB	10 (90.1%)	17 (77.3%)	27 (81.2%)
SenseRelate	10 (90.1%)	17 (77.3%)	27 (81.2%)
Both	9 (81.2%)	10 (45.5%)	19 (57%)

Table 2: Word-sense distribution in the Sense Inventory and in the different WSD settings

Manual inspection of the results shows that senses which are eliminated in this way are mostly minor ones, such as *nail3* which is an obsolete unit of length in the easy category, or not well represented in the corpus we used for the experiment, such as *object2*, a computing term in the difficult category.

WSD setting	Easy words	Difficult words
UKB	0.547	0.476
SenseRelate	0.594	0.309
Both	0.731	0.716

Table 3: Accuracy of sense-tagging in the different WSD settings

	No. of senses	Max possible score	Obtained score	Accuracy
Baseline	33	74	35	0.473
UKB	27	81	45	0.555
SRel	26	78	46	0.590
UKB-19	19	57	39	0.684
SRel-19	19	57	33	0.579
Both	19	57	41	**0.720**

Table 4: Accuracy of translation equivalent extraction in the different WSD settings

As can be seen from Table 3, average accuracy for sense-tagging with UKB and SenseRealate is similar for easy words: 0.547 for UKB and 0.594 for SenseRelate but UKB outperforms its counterpart in the difficult category: 0.476 vs. 0.309. When the results of both taggers are combined, the improvement is much higher: 0.731 for the easy words and 0.716 for the difficult ones. These results show that using an intersection of both taggers is especially beneficial for highly polysemous words. The reason for such an improvement is the fact that, when using just the intersection of the two sense taggers, both of our previously stated assumptions about the increased accuracy and removal of peripheral senses are correct.

4.2. Evaluation of extracting translation equivalents

The goal of the second part of the evaluation was to evaluate the end results we obtained with the proposed approach for finding translation equivalents of polysemous words in comparable corpora. We took into account all three WSD settings. As a baseline we used our original technique that uses the same settings for building and comparing context vectors with the only exception that it builds a single vector for each headword, regardless of sense tags assigned to their specific occurrences. It is important to note that the baseline approach is overall evaluated in a somewhat favorable fashion. On the one hand the distinction between senses is not made explicitly and a correct translation of any sense of the headword is regarded as correct. In the new approach the distinction between senses is made, so the translation candidate has to be a correct translation for the specific sense. On the other hand, the baseline approach has to cope with all the 33 senses of the eight chosen words while in the new approach the number of senses diminishes as not enough tagged data for a specific sense is available.
The baseline approach suggests a correct first candidate in 21.2% of the cases. Taking into account 10 highest-ranking candidates, a correct translation is found in 36.3% of the cases, which is a lot worse than when UKB and SenseRelate are combined. In that case we are able to extract a correct first candidate in 36.8% of the cases while, when taking into account 10 highest-ranking candidates, a correct translation is found in as many as 78.9% of the cases.
It is also interesting that in case no appropriate translation is found among the candidates, the suggestions obtained with this method point to the correct sense of the word.
In an overall evaluation we assigned one of the 4 possible scores to extracted translation equivalents for each sense of the headwords in the pilot study:

- 0 if no correct translation has been found among the 10 highest scoring candidates and the sense of the source word cannot be determined from the extracted translation candidates,
- 1 if no correct translation has been found among the 10 candidates but the sense of the source word can be determined from the extracted translation candidates,
- 2 if a correct translation is found among the 10 highest scoring candidates, and
- 3 if the highest scoring candidate is a correct translation.

bass	bas (bass), boben (drum), zvok (sound), pevec (rooster), igralec (player)
body	človek (human), del (part), telo (body), primer (example), čas (moment)
organ	človek (human), del (part), primer (example), življenje (life), srce (heart)

Table 5: Examples of translation candidates obtained with the baseline method where no sense distinction is made

The overall accuracy of the four approaches is shown in Table 4. It is calculated as the quotient between the obtained and the maximum possible score. The maximum possible score is calculated by the number of senses for which the translations are sought (19*3=57 for the last three rows). In

bass1:	bas (bass), boben (drum), klavir (piano), zvok (sound), glasba (music)
bass2:	pesek (sand), jastog (lobster), zanka (trap), razpoka (crack), luža (puddle)
body1:	institucija (institution), interes (interest), služba (service), uslužbenec (employee), organizacija (organization)
body2:	telo (body), človek (human), del (part), življenje (life), oblika (form)
organ1:	orgle (organ), klavir (piano), kontrabas (contrabass), harfa (harp), flavta (flute)
organ2:	organ (organ), človek (human), srce (heart), telo (body), del (part)

Table 6: Examples of translation candidates obtained with the intersection of both taggers

case of the baseline method the maximum possible score is calculated in a different manner since it is not possible for translations of more senses to occur in the first position (8*3+(33-8)*2=74; since there are 8 headwords, for 8 senses it is possible to find the translation in the first position (score 3), for the remainder of the senses (33-8=25) the best-case-scenario is finding it among the ten first candidates (score 2)).

The results show that accuracy improves with our new methods (0.555 and 0.590 compared to 0.473). One could argue that the new methods have a simpler task with fewer senses to choose among since some are discarded because they do not pass the frequency threshold which is enforced for building context vectors. But one should be reminded that the baseline method does not differentiate between senses at all. Additionally, the senses not found in the corpora are obviously the minor ones. Using an intersection of both sense taggers (0.720) improves the results even further compared to the accuracy achieved on the same set of 19 remaining senses based on the output of each tagger respectively (0.684 for UKB and 0.579 for SenseRelate).

Tables 5 and 6 give some examples of Slovene translation candidates for the English polysemous headwords which we obtained with the baseline method where no sense distinction was made, and that we obtained with the intersection of both taggers (the English translations are given in brackets for better understanding of the examples):

5. Conclusions and future work

In this paper we proposed an approach to extract translations of polysemous words from comparable corpora, a problem which has so far been largely neglected by most of the related work. We use third-party sense taggers for determining the senses of source words and then build and compare separate vectors for each of the senses.

We ran the experiment in three different settings: first, translation equivalents were extracted for the senses assigned by the UKB tool, second, sense-assignment by the SenseRelate tool was taken into account, and finally, the results of both sense-taggers were combined and the context vectors were built only for those occurrences which were disambiguated in the same way by both taggers. All the settings outperform the baseline method, which builds a single context vector for all occurrences of a polysemous word in the corpus, regardless of what sense it is used in. The best-performing setting is the last one that only uses the intersection of both sense-taggers, suggesting that the quality of word-sense disambiguation is a crucial factor for building cleaner context vectors and a successful cross-lingual comparison.

The results of our pilot study are very promising since they significantly outperform the baseline method in all the settings. An important finding of the study is that even though word-sense disambiguation tools are still not very accurate, they can already be useful in stochastic approaches if sufficient data is available. Additionally, they can be combined in order to achieve better accuracy and filter out odd senses. In the future we wish to test the approach on a large scale, as well as try to extract translation equivalents for other parts of speech.

6. Acknowledgments

Research reported in this paper has been supported by the ACCURAT project within the EU 7th Framework Programme (FP7/2007-2013), grant agreement no. 248347, and by the Slovenian Research Agency, grant no. Z6-3668.

7. References

E. Agirre and A. Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of EACL 09*, pages 33–41.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

C. Fellbaum. 1998. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA.

D. Fišer, N. Ljubešić, Š. Vintar, and S. Pollak. 2011. Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *Proceedings of BUCC11*.

P. Fung. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In *AMTA98*, pages 1–17.

A. Grad, R. Škerlj, and N. Vitorovič. 1999. *English-Slovene Dictionary*. DZS.

H. Kaji. 2003. Word sense acquisition from bilingual comparable corpora. In *HLT-NAACL*.

N. Ljubešić and T. Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Proceedings of BSNLP 11*, pages 395–402.

N. Ljubešić and D. Fišer. 2011. Bootstrapping bilingual lexicons from comparable corpora for closely related languages. In *Proceedings of TSD 11*, pages 91–98.

R. Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of COLING-ACL 2006*, pages 105–112.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619.

T. Pedersen and V. Kolhatkar. 2009. Wordnet::senserelate::allwords - a broad coverage word sense tagger that maximimizes semantic relatedness. In *Proceedings of NAACL 09*, pages 17–20.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL 99*, pages 519–526.