Using Parallel Corpora for Word Sense Disambiguation

Ahmad R. Shahid[1,2]
[1]Department of Computer Science
COMSATS Institute of IT
Islamabad, Pakistan
ahmadrshahid@comsats.edu.pk

Dimitar Kazakov[2]
[2]Department of Computer Science
University of York
York, United Kingdom
dimitar.kazakov@york.ac.uk

Abstract

This paper presents a method of lexical semantic disambiguation in multilingual corpora and describes the construction of an artificial word-aligned and lexically disambiguated gold-standard corpus from an existing multilingual resource. The suggested approach uses sets of aligned words and phrases across languages as unique semantic tags similar to WordNet synsets that can be used as a part of unsupervised natural language processing and information retrieval tasks. The approach goes beyond one-to-one word alignment, and uses an algorithm for the aggregation of results of pair-wise word alignment when the corpus contains several languages. When applied to the new corpus, this methodology has proven capable of reducing the ambiguity of a polysemous word by one third on average.

1 Introduction

This is a study of the specific potential that parallel corpora provide for word and phrase sense disambiguation (WPSD). We do not discuss any of the methods that can be applied to monolingual texts, as these can be considered complementary approaches that are not mutually exclusive, but, rather, can always be combined together. We focus instead on the specific contribution that the availability of multiple translations of the same text can make towards rejecting some of the alternative senses of the words and phrases in the corpus for any of the individual languages represented in it. We describe an approach in which the N translations in a parallel corpus are word-aligned, and the result used to group words and phrases that are translations of each other into N-tuples that can be seen as multilingual synsets akin to the sets of synonyms used in WordNet (Fellbaum, 1998). These synsets can then be used as semantic tags for word and phrase sense disambiguation. The approach was applied to a large, real-world parallel corpus, namely, Europarl (Koehn, 2005).

In this setting the full potential of the idea can be obscured by errors introduced by one pre-processing step, such as imperfect word alignment, or the lack of another, e.g., morpholexical analysis. We therefore use an existing multilingual lexical resource (Lefever, 2009) to develop a large, artificial parallel corpus containing semantically disambiguated polysemous words, and use it to calculate the maximum contribution that parallel corpora can make towards WPSD under ideal conditions, when all other processing steps are 100% accurate and therefore do not introduce any noise to the process. This result gauges the potential contribution of multiple translations to WPSD, providing its upper limit for the data studied.

The multilingual synsets produced in this framework represent a potentially valuable resource on its own, which could be used (as is, or after filtering out the errors) as a translation memory or as a lexicographer's resource. The unedited multilingual synsets from the experiments with Europarl have been made available online.[1] The Web interface includes search for words and phrases in the four languages used, and also displays all the contexts in which the word or phrase in question appears in the corpus.

[1]http://www.goodwithlanguages.com

Proceedings of Recent Advances in Natural Language Processing, pages 336–341,
Hissar, Bulgaria, 7-13 September 2013.

2 Background

Dagan *et al* (1991) first noted the usefulness of two corpora (one for each language) for lexical semantic disambiguation in the context of machine translation. Binary syntactic relations are identified in the source language and all of their possible translations are initially produced, and then gradually pruned based on the observed likelihood of these pairs of words in the target language corpus. It is noted that the target language word choice can indicate the sense of ambiguous words in the source language.

Gale *et al* (1992) used a parallel corpus to label ambiguous source words (along with their context) with the target language word translation. One could then learn from the labelled examples using the source context words as features to distinguish between the senses of unseen examples of the ambiguous word in the source language. All this of course assumed different target words were used for different senses of the source word.

More recently, parallel corpora have been used to create new linguistic resources, such as lexicons and WordNet-like resources (Fišer, 2007; Sagot, 2008; Shahid, 2009; Shahid, 2010; Lefever, 2009; Lefever, 2010a; Lefever, 2010b).

Fišer (2007) word aligned the translations of Orwell's *1984* (Dimitrova et al., 1998) in five languages: English, Czech, Romanian, Bulgarian and Slovene. She carried out pair-wise word alignment of nouns, verbs, adjectives, and adverbs using GIZA++ (Och, 2003). Only 1:1 alignments between words of the same part of speech were considered and alignments occurring only once were discarded. The bilingual word alignments (lexicons) thus generated were used to create a multilingual lexicon with 1500 entries. The multilingual lexicon was then compared against the existing WordNets: PWN (Fellbaum, 1998) for English; BalkaNet (Tufis, 2000) for Czech, Romanian and Bulgarian. If all the translations in a particular entry in the lexicon shared the synset ID, the same synset ID was assigned to the Slovene translation. Slovene words that shared the same synset ID were then grouped into synsets.

Sagot (2008) created WOLF, a freely available French WordNet. They used the *extend approach* (Vossen, 1998) whereby a subset of synsets was taken from the PWN and translated into the target language, preserving the structure of the PWN. 82% of the entries in the PWN are monosemous and only require a bilingual lexicon. For the polysemous words they pair-wise word aligned the subcorpus of the JRC-Acquis (Steinberger et al., 2006) in five languages, that is, English, Romanian, Czech, Bulgarian, and French. The bilingual lexicons thus created were used to create the multilingual lexicons. Translations in the multilingual lexicon were then compared against the corresponding WordNet in BalkaNet (Tufis, 2000). If all translations shared the same synset ID, the corresponding French translation was also assigned the same synset ID.

Shahid and Kazakov (2009; 2010) have used the notion of synsets in a multilingual context (cf. (Lavric, 2008)), defined as translation equivalences. They used the Europarl parallel corpus and word-aligned a subset of it for four languages, English, German, French and Greek, using an off-the-shelf tool (GIZA++ (Och, 2003)). English was used as the pivotal language. The resulting $1:1$ and $1:N$ mappings between words in each pair of languages were then grouped into 4-tuples of synonymous words, resp. phrases, using an in-house algorithm (Algorithm 1, (Shahid, 2010)). The resulting sets of translations are referred to as multilingual proto-synsets, to highlight the fact that they can be further improved, e.g., by merging those showing morphological variants of the same lexical entry. Similarly, one could consider merging multilingual proto-synsets if they only contained pairs of synonyms for each language.

It is also of relevance that Lefever and Hoste (2009; 2010a; 2010b) proposed an unsupervised multilingual Word Sense Disambiguation (WSD) task for polysemous English nouns. Rather than manually sense tagging individual occurrences of the nouns in the example sentences, they built a gold standard sense inventory using the Europarl parallel corpus in six languages: English, German, French, Spanish, Italian, and Dutch. The parallel corpus was word aligned using GIZA++. The word alignments were then manually verified by certified translators who were also asked to annotate 20 sentences per trial target word giving at most 3 suggested meanings at a time. These sense annotated sentences can also be treated as gold standard data.

3 Design

We have built here on Shahid and Kazakov's approach (Shahid, 2010) to use the multilingual proto-synsets they propose for word and phrase sense disambiguation, as described in the introduction.

The words were collated into phrases in the following way. Initially, each word in each language in the word-aligned parallel corpus is given a separate, unique identifier. Two data structures, an 'open' and a 'closed' list are created. Initially, all words are placed in the open list, and the closed list is empty. A simple recursive procedure, `fanout/1`, is used to extract all phrases. It takes a word from the open list, and gradually spread its ID to all words it is aligned with. Each processed word is transferred on to the closed list, which in the end, when the open list is empty, contains all words. All words that could be connected through one or more pair-wise alignments, now have the same ID. In other terms, all words forming a phrase and its translation into each language, are now indexed with the same ID. Each phrase and the corresponding translations form a multilingual proto-synset.

Algorithm 1 Multilingual Synset Construction

```
main(){
foreach Word in OpenList
  fanout(Word)
}

fanout(Word) {
  move Word from OpenList to ClosedList
  foreach W in OpenList that is aligned with Word
    W.ID=Word.ID
    fanout(W)
}
```

The process is deterministic and is not prone to introducing errors on its own. However, the errors introduced in the preceding steps are carried over to subsequent steps after phrase formation. In other words, the quality of proto-synsets is only as good as the quality of word alignment, not worse. Table 1 shows a larger sample of the results

3.1 Using Phrases in the Multilingual Synsets

In the word alignments generated by GIZA++ there are many words in a non-pivotal language that are aligned with N words in the pivotal language, or in other words they have $1 : N$ word mapping. Earlier research did not use this information to generate phrases from words (Fišer, 2007; Sagot, 2008). Our experiments with the parts of the Europarl corpus produced phrase alignments rather than $1 : 1$ word mapping in 28% of all cases. This is a substantial figure which shows that phrase alignment can have a substantial impact on the overall result. The quality of this alignment however cannot be tested without an appropriate annotated resource.

3.2 SemEval Parallel Corpora

We have have therefore set off to create a large, artificial parallel corpus where the semantics of selected key words (in their canonical lexical entry form) has been disambiguated. The result was to be used to evaluate the maximum contribution of multilingual synsets to the WSD process.

We made use of a resource which was part of the SemEval-2010 Task 3 on Cross-Lingual Word Sense Disambiguation (Lefever, 2009; Lefever, 2010a; Lefever, 2010b). This data is in six languages, namely, English, French, German, Dutch, Italian and Spanish.

Lefever and Hoste used the parallel corpus in all six languages to generate a gold standard data set and a sense inventory. They provided five target nouns to be disambiguated, namely, *bank*, *movement*, *occupation*, *passage*, and *plant* (Lefever, 2010b). They also provided a sense inventory for each of the target nouns.

The sense inventory defined meanings in which a target word could be used. It also contained combinations of words/phrases in all the six languages with semantics related to a particular meaning of the target word. For instance, the word *bank* had five different meanings: *Financial Institution*, *Supply/Stock*, *Sloping land beside water*, *Cisjordan*, and *group of similar objects (row/tiers)*. Further sub-meanings were also defined but for the purposes of this exercise we assumed them to be part of the main meanings.

The sense inventory was used by annotators to annotate 20 sentences per target word. They were asked to provide contextually relevant translations for each of the languages considered. The sentences were extracted from JRC-ACQUIS[2] and the British National Corpus (BNC)[3].

[2] http://langtech.jrc.it/JRC-Acquis.html
[3] http://www.natcorp.ox.ac.uk/

English	German	French	Greek
resumption of	wiederaufnahme	reprise de	επανάληψη της
session	sitzungsperiode	session	συνσδου
adjourned on friday	erkläre am freitag	interrompue vendredi	διακοπεί παρασκευή
thank you	vielen dank	merci	ευξαριστώ
shall do so gladly	will tun gerne	ferai volontiers	πράξω ευξαρίστως

Table 1: Sample multilingual synsets

There were 20 English sentences per each target word provided. Multiple translators were asked to translate the target words into 5 other languages, and a gold inventory of the possible translations of each word in each of its meanings was compiled. Annotators were asked to provide 3 or fewer relevant translations from the sense inventory. The proposed translations were stored with their frequency counts, of how many times a word/phrase from the sense inventory was used to translate a target word for a given language.

Given below is the list of possible translations of the word 'bank' to German for different senses with the frequency of its usage by a translator.

> bank.n.de 1 :: bank 4;bankengesellschaft 1;kreditinstitut 1;zentralbank 1;finanzinstitut 1;
> bank.n.de 2 :: bank 4;zentralbank 3;finanzinstitut 1;notenbank 1;kreditinstitut 1;nationalbank 1;
> bank.n.de 3 :: westjordanufer 3;westufer 2;westjordanland 2;westjordanien 2;westbank 2;west-bank 1;

This data can be the basis for a gold standard corpus: the translations of the words in question are perfectly aligned, and the words themselves are in their lexical entry form, that is, not needing any morphological analysis. Therefore, any experiments with this data will eliminate the errors introduced by GIZA++ and the lack of morphological analysis.

We used this data set to theoretically gauge the maximum by which the polysemy of an ambiguous word could be reduced by translations of a word across different languages. For the said purpose, we generated all possible multilingual synsets (combinations of possible translations) from the gold standard data and checked in the sense inventory to find all meanings to which this combination of translations across the six languages could possibly correspond. For instance, any combination of the words (bank:EN, westjordanien:GE...) could only correspond to the third and last meaning of the English word 'bank', that of a bank of a river.

On occasions, a combination of translations would correspond to more than one sense of the word. These combinations of translations (aka synsets) were weighted with the frequency with which its constituent words were proposed by the individual translators. We calculated polysemy (number of senses) for each word and synset, and the ratio by which such a synset would reduce the polysemy of the original English word.

Table 2 gives a summary of the results. It can be seen that polysemy is reduced by over 36% on average when translations of a word are used as sense tags. This is a significant result, which suggests that the previous negative results are due to other factors, some of which were already mentioned; however, the idea of using multilingual synsets for WSD is viable, and can be used when the other techniques needed reach a more mature stage of development.

3.3 Further Evaluation

For an evaluation of the synsets thus generated, we annotated the 5 target English words in the 20 trial sentences using the senses in the sense inventory. Two native speakers and one speaker with near-native proficiency were asked to annotate the target words. To generate consensus, only those senses were considered for evaluation where at least two annotators agreed. The annotated sentences were taken as gold standard (GS), against which the senses proposed by our synsets generated from the SemEval data were compared.

We used the Most Frequent Sense (MFS) as the first baseline for this comparison. Thus, among all the sense annotations for a target word the most fre-

Word	# of synsets	Before WSD	After WSD	Reduction [%]
bank	17,873	5	2.7	46%
movement	230,061	3	2.51	16%
occupation	81,706	4	3.39	15%
passage	95,363	7	3.71	47%
plant	91,830	3	1.67	44%
Total	516,833	4.4	2.796	36.45%

Table 2: Lexical ambiguity (polysemy) of English words before and after the use of multilingual synsets for disambiguation.

quent was taken and it was assumed that all the occurrences of the target word bore the same sense, referred to as 'GS-MFS.' We also took the top sense for a target word from PWN (Fellbaum, 1998), which orders them by frequency, and assumed that all the occurrences of the same target word bore the same meaning. It can be called as PWN-MFS. We compared the GS-MFS, PWN-MFS and senses proposed by our synsets for each occurrence of the target word against the GS. The results indicate that the accuracy of senses proposed by the multilingual synsets is 86%, 52% for PWN-MFS, and 59% for GS-MFS. This clearly shows the benefits of our approach.

4 Conclusion

We have demonstrated how a parallel corpus can be used for word (and phrase) sense disambiguation for each of its languages. The described approach also produces a new lexical resources as a side effect, which can be independently used for a variety of purposes. We demonstrated the viability and the upper limit of the potential of multilingual synsets for WSD on a novel data set specifically constructed for the purpose. There is a pleasing feeling about the fact that such an upper bound can be measured at all with rigor.

We have shown at the same time that the idea still has its limitations in practice due to the imperfections of other preprocessing techniques, such as word alignment, on which it is based.

5 Future Work

Rather than using existing resources to carry out morpholexical analysis in order to improve the results, we have considered the possibility of first learning such resources in the form of word paradigms from the parallel corpus. Once word paradigms are learned, they can be used for the above mentioned purpose of merging multilingual synsets, as the ambiguity such variant synsets indicate is spurious. We have chosen to frame these experiments as an unsupervised learning task, where the only resource available is the corpus. A comparison of the results to an existing gold standard and to another, monolingual unsupervised morphology learning approach have shown the clear potential of this approach, which will be the subject of a separate publication.

References

Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen. 1984. *Classification and Regression Trees*. Chapman and Hall/CRC, Florida, USA.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. *Two languages are more informative than one. ACL-91 Proceedings of the 29th annual meeting on Association for Computational Linguistics*, 114–133. Stroudsburg, PA, USA.

Ludmila Dimitrova, Nancy Ide, Vladimir Petkevic, Tomaz Erjavec, Heiki Jaan Kaalep, and Dan Tufis. 1998. *Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages*. Proceedings of the 17th International Conference on Computational Linguistics - Volume 1 (COLING 98). Stroudsburg, PA, USA.

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

Darja Fišer. 2007. *Leveraging parallel corpora and exisitng wordnets for automatic construction of the Slovene Wordnet*. Proceedings of L&TC 2007. Poznań, Poland.

William A. Gale, Kenneth W. Church and David Yarowsky. 1992. *Using bilingual materials to develop word sense disambiguation methods*. *TMI*. Montreal.

Philipp Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation* Proceedings of the MT summit 2005.

Eva Lavric, Gerhard Pisek, Andrew Skinner, and Wolfgang Stadler (Eds). 2008. *The Linguistics of Football*. Narr Francke Attempto Verlag.

Els Lefever and Véronique Hoste. June 2009. *SemEval-2010 Task 3: Cross-lingualWord Sense Disambiguation*. Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions. pp. 82-87. Boulder, Colorado.

Els Lefever and Véronique Hoste. 2010. *Construction of a Benchmark Data Set for Cross-Lingual Word Sense Disambiguation*. Proceedings of the Seventh International Conference on Language Resources and Evaluation. Malta.

Els Lefever and Véronique Hoste. July 2010. *SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation*. Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010. pp. 15-20. Uppsala, Sweden.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51.

Benoît Sagot and Darja Fišer. 2008. *Building a free French wordnet from multilingual resources*. Proceedings of OntoLex 2008. Marrakesh.

Ahmad Shahid and Dimitar Kazakov. 2009. Unsupervised Construction of a Multilingual WordNet from Parallel Corpora. Proc. of the RANLP Workshop on NLP methods and Corpora in Translation, Lexicography, and Language Learning. Borovets, Bulgaria.

Ahmad Shahid and Dimitar Kazakov. 2010. Retrieving Lexical Semantics from Multilingual Corpora. *Polibits* 5:25–28.

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis, and Dániel Varga. 24-26 May 2006. *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy.

Dan Tufis. 2000. Design and Development of a Multilingual Balkan WordNet. *Romanian Journal of Information Science and Technology Special Issue*, 7:1-2.

Piek Vossen. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

R Wagner and M Fischer. 1974. The String-to-String Correction Problem. *Journal of the Association for Computing Machinery*, 21(1):168-173.