Multilingual Sense Intersection in a Parallel Corpus with Diverse Language Families

Giulia Bonansinga[1], **Francis Bond**[2]

[1]Filologia, Letteratura e Linguistica, Università di Pisa, Italy

[2]Linguistics and Multilingual Studies, Nanyang Technological University, Singapore

giuliauni@gmail.com, bond@ieee.org

Abstract

Supervised methods for Word Sense Disambiguation (WSD) benefit from high-quality sense-annotated resources, which are lacking for many languages less common than English. There are, however, several multilingual parallel corpora that can be inexpensively annotated with senses through cross-lingual methods. We test the effectiveness of such an approach by attempting to disambiguate English texts through their translations in Italian, Romanian and Japanese. Specifically, we try to find the appropriate word senses for the English words by comparison with all the word senses associated to their translations. The main advantage of this approach is in that it can be applied to any parallel corpus, as long as large, high-quality inter-linked sense inventories exist for all the languages considered.

1 Introduction

Cross-lingual Word Sense Disambiguation (CL-WSD) is an approach to Word Sense Disambiguation (WSD) that exploits the similarities and the differences across languages to disambiguate text in an automatic fashion. Using existing multilingual parallel corpora for this purpose is a natural choice, as shown by a long series of works in the literature; see for instance Brown and Mercer (1991), Gale et al. (1992), Ide et al. (2002), Ng et al. (2003), Chan and Ng (2005), and Khapra et al. (2011) more recently.

As Diab and Resnik (2002) showed, the translation correspondences in a parallel corpus provide valuable semantic information that can be exploited to perform WSD. For instance, Tufiş et al. (2004) used parallel corpora to validate the interlingual alignments in different WordNets (WNs). Specifically, they looked at the sense intersection between the lexical items found in all the reciprocal translations of a parallel corpus.

Gliozzo et al. (2005) showed how CL-WSD can help to sense-annotate a bilingual corpus by looking at the semantic differences in a language pair. Bentivogli and Pianta (2005), on the other hand, focused on how meaning is somehow preserved despite those differences, which allows us to transfer the semantic annotation of a text in a certain language to its translation in another language. The *sense projection* procedure that they used is simple yet powerful, but it can only be applied on corpora in which at least one parallel text is annotated with senses. Nevertheless, given the difficulty to come across sense-annotated data, any way to produce such data is of great benefit to WSD. The *knowledge acquisition bottleneck* is still a challenge to address for most languages.

Given the task of annotating an ambiguous word in a multilingual parallel corpus, some valuable information can be derived through the comparison of the *set of senses* of each of the word's translations. If fewer senses (or one only, in the optimal case) are retained across languages, then the cross-lingual information has helped reducing (or solving) the ambiguity.

In previous work (Bond and Bonansinga, 2015) we employed *sense intersection (SI)* to annotate a trilingual parallel corpus in English, Italian and Romanian built upon SemCor (SC) (Landes et al., 1998). We summarize the data used and our findings in Section 2.

In Section 3 we continue investigating in the same strand by introducing a further language, Japanese, to disambiguate English text. In Section 4 we show how an annotation task can benefit from coarser sense distinctions. In Section 5 we examine thoroughly how and how much each additional language helps the automatic sense disambiguation process. We conclude in Section 6.

2 Multilingual Sense Intersection

In Bond and Bonansinga (2015) we explored the cross-lingual approaches pioneered by Gliozzo et al. (2005) and Bentivogli and Pianta (2005) to annotate the SC corpus (Landes et al., 1998) and two corpora built upon it from its Italian and Romanian translations. This parallel corpus, though rather small (see Subsection 2.1), is ideal for the task as it is sense-annotated in all its translations, thus making the evaluation of alternative sense annotation methods straightforward. We briefly present the data used back then and introduce the last component of the corpus, the Japanese SemCor (Bond et al., 2012), which is included in the analysis presented in this paper.

2.1 Data

Developed at Princeton University, SC is a subset of the Brown Corpus of Standard American English (Kučera and Francis, 1967) enriched with sense annotations referring to the WN sense inventory (see Section 2.2).

Bentivogli and Pianta (2005) manually translated 116 SC texts and automatically aligned them to their English counterparts. Then the sense annotations of the English words were automatically transferred following the word alignment, thus leading to the creation of a sense-annotated English-Italian corpus, MultiSemCor (MSC).

With the purpose of providing a Romanian version of SC, Lupu et al. (2005) developed the Romanian SemCor (RSC) (Lupu et al., 2005; Ion, 2007), which shares 50 texts with MSC. Unfortunately, RSC is not word-aligned to any other component of the parallel corpus, which is a requirement to perform sense mapping with any of the mentioned procedures. Nevertheless, as the sentence alignment is available and as we are only interested in content words, we attempted a word alignment based upon the information already available. First, we aligned all the reciprocal translations in the same sentence pair having identical sense annotation. Then, we aligned the remaining content words, if any, using heuristics that exploit PoS information and path similarity in the WN ontology. Finally, we manually checked a sample of the alignment found in this fashion and we observed a precision of 97%; of course, errors can only be introduced in the second step, when the heuristics used to align the remaining unaligned content words come into play.

Bond et al. (2012) built a Japanese SemCor (JSC) matching the texts covered in MSC, after porting the sense annotations to WN 3.0 using the mappings provided by Daude et al. (2003). The sense annotation was carried out through sense projection by exploiting the word alignment, similarly to what Bentivogli and Pianta (2005) did for Italian.

JSC follows the Kyoto Annotation Format (KAF) (Bosma et al., 2009) and is released under the same license as SC.[1]

In Table 1 we remind the basic statistics of each corpus. For English and Italian we also specify the number of the target words after the migration to WordNet 3.0 (WN 3.0). In Table 2 we give a clearer picture of the alignments available in terms of the number of aligned sentences for each language pair.

	Texts	Tokens	Target words	After mapping
EN	116	258,499	119,802	118,750
IT	116	268,905	92,420	92,022
RO	82	175,603	48,634	=
JP	116	119,802	150,555	=

Table 1: Statistics for each component of the multilingual parallel corpus built from SemCor.

2.2 Sense Inventories

When MSC was released, MultiWordNet[2] (MWN) (Pianta et al., 2002), a multilingual WordNet aligned to Princeton WN 1.6, was used. As described in Bond and Bonansinga (2015), we ported all senses annotations in MSC to WN 3.0, so to make it possible a comparison between

[1] Both the Japanese WordNet and the Japanese SemCor are available at the following address: http://compling.hss.ntu.edu.sg/wnja/index.en.html

[2] http://multiwordnet.fbk.eu/

Language	Aligned sentences
EN-IT	12,842
EN-RO	4,974
EN-JP	12,781
IT-RO	4,974
IT-JP	12,781
RO-JP	4,913

Table 2: Number of aligned sentences for each language pair.

the different components of the parallel corpus. To this aim, we used automatically inferred mappings (Daudé et al., 2000; Daudé et al., 2001). However, the changes occurred between WN versions 1.6 and 3.0 led to the loss of 4,631 sense annotations (1,204 types, half of which are adjective satellites).

The Romanian WordNet (RW), created within the BalkaNet project (Stamou et al., 2002) and then consistently grown independently (Barbu Mititelu et al., 2014) was aligned to WN 3.0 with precision of 95% (Tufiş et al., 2013).

The Japanese WN (JWN) (Isahara et al., 2008; Bond et al., 2009), originally developed by the National Institute of Information and Communications Technology (NICT) and firstly released in 2009, is a large-scale semantic dictionary of Japanese and is available under the WordNet license.

	Synsets	Senses
English	117,659	206,978
Italian	34,728	69,824
Romanian	59,348	85,238
Japanese	57,184	158,069

Table 3: Coverage of the WNs used.

In Table 3 we give basic coverage statistics for the WNs of our target languages. The Open Multilingual WordNet (OMW)[3] is an open-source multilingual database that connects all open WNs linked to the English WN, including Italian (Pianta et al., 2002) among the 28 languages supported (Bond and Paik, 2012; Bond and Foster, 2013). A convenient interface to OMW is provided in the Python module NLTK[4] (Bird et al., 2009).

2.3 Findings

For the sake of completeness, in previous work we performed sense projection on the Italian and Romanian corpora using English as pivot, scoring a precision of over 90% in both cases. As for SI, we report the previous precision and coverage scores obtained through trilingual SI in Table 4, along with the Most Frequent Sense (MFS) baseline, that assigns each word its most frequent sense. In this step, sense frequency statistics (SFS) are therefore necessary, but unfortunately there are very few sense-annotated corpora from which we can derive such statistics. In the case of SC the issue is even more crucial, because in WN senses are ranked depending on their frequency in SC. So, whenever the first sense of a lemma follows a ranking order, we are using biased statistics.

Generally speaking, the coverage scores were quite good and higher with the baseline MFS. As for precision, the gap between SI and the baseline is smaller, probably due to the bias just mentioned. On the other hand, in languages other than English, the contribution of SFS is not as decisive and SI performs better than the baseline, and particularly so in the case of Italian.

3 Multilingual Sense Intersection with languages from different families

The theoretical justification behind Multilingual Sense Intersection (SI) is in that an ambiguous word will often be translated in different words in another language. As a consequence, the knowledge of all the senses associated to its translation can help detect the sense actually intended in the original text. More commonly, such a comparison will help reduce the ambiguity, but it will not identify one single, shared sense. On the other hand, a text whose ambiguity has been progressively reduced through automatic methods can be completely disambiguated by a human annotator at a lesser cost. Moreover, the more the languages available for comparison in the parallel corpus, the more likely is that SI actually manages to discern the correct sense in context.

Differently from our previous work, where we disambiguated all the texts that were aligned with at least one other language, in the following section we show results computed over 49 texts. Those constitute the subset of the corpus shared across all four components and for which we have alignments. As a result, we use an even smaller corpus through which, nevertheless, we can show more effectively the contribution of up to three languages.

Given an ambiguous word, all its translations provide their 'set of sense', as retrieved from the shared sense inventory. Then, intersection is performed over every non-empty set and successes when the final *overlap* contains only one sense, meaning that the target word has been disambiguated. Otherwise, the overlap is further intersected with the top most frequent senses available

[3] http://compling.hss.ntu.edu.sg/omw/summx.html

[4] http://www.nltk.org

Method	English		Italian		Romanian	
	Precision	Coverage	Precision	Coverage	Precision	Coverage
MFS (baseline)	**0.761**	**0.998**	0.599	**0.999**	0.531	**1**
3-way Intersection	0.750	0.778	**0.653**	0.915	**0.590**	1
Coarse-grained MFS	**0.850**	**0.998**	0.687	**0.999**	**0.794**	**1**
Coarse-grained SI	0.849	0.778	**0.761**	0.915	0.661	1

Table 4: Comparison of the results scored with SI and MFS baseline.

for the target lemma. We take note whether the sense selected was the most frequent one. As before, we resort to sense frequency statistics (SFS) whenever the target word is not yet disambiguated after SI. These frequencies were calculated over all texts in the corpus **except** the one being annotated.

4 Introducing coarse-grained senses

Sense inventories are a crucial part of this approach. Not only are a sufficient coverage and the alignment to the Princeton WN necessary: when it comes to deciding how to define close, very specific senses, a trade-off between the detail of the sense description and its actual usability in real contexts is highly desirable.

The fine granularity of WN senses can occasionally, depending on the application, be more of a practical disadvantage than a quality. In this analysis, for instance, error analysis suggested that the senses found through SI were often very close, but it may happen that they are discarded as wrong outputs just because one language has a WN more developed and granular than another. We should also bear in mind that the correct senses against which we evaluate were picked by trained human annotators in the first place, and human annotators tend to describe a word as precisely as possible.

Conscious of this limit, Navigli (2006) devised an automatic methodology to find a reasonable sense clustering for the senses in WN 2.1. Sense clustering can be of great help in tasks where minor sense distinctions can be ignored, allowing a coarse-grained evaluation.

They found 29,974 main clusters, some of which were manually validated by an expert lexicographer for the Semeval all-word task.

We mapped the senses in the clusters found to WN 3.0, losing 101 of them in the process (typically one-element clusters). When evaluating the results of SI, we performed a coarse-grained evaluation; in particular, whenever the sense found by SI was not correct, we checked whether it was part of a sense cluster and whether the correct sense was in it. If so, we considered the output of the algorithm correct.

Table 4 displays the difference in performance when coarse-grained evaluation is employed.

Method	English	
	Precision	Coverage
Coarse-grained MFS	0.851	**0.998**
Coarse-grained 4-SI	**0.854**	0.788

Table 5: Coarse-grained evaluation of the results scored with 4-way SI and MFS baseline, computed over the shared subset (49 texts).

5 Evaluation

In Table 4 we show the improvement in precision obtained thanks to coarse-grained evaluation with respect to the results in Bond and Bonansinga (2015). English and Italian show respectively a significant improvement of 0.1 and 0.11. In the case of Romanian, the improvement is not as big, but still meaningful (0.07). Of course, coarse-grained evaluation causes the MFS baseline to improve as well. In the case of English - which, again, is the component most subjected to the bias introduces by SFS - the difference between MFS and SI decreases a little, but MFS still performs better.

The case of Italian is unique, in that SI obtains better precision scores with both fine and coarse-grained senses. For Romanian, on the other hand, SI performs better until coarse-grained evaluation is employed, and the improvement achieved by MFS is striking.

In Table 5 we show our latest attempt to disambiguate English text by using the semantic information of its aligned translation in a parallel corpus. The languages that contribute to the dis-

ambiguation process are Italian, Romanian and Japanese, and all together they manage to beat MFS, if coarse-grained senses are considered.

6 Conclusions

For future work, it is important to analyze the progressive improvement that we can achieve by taking into account semantic information from one language at the time, so as to verify if it is true that it is the very diverse languages that contribute the most to the disambiguation process.

As for the sense inventories, it would be interesting to compare different lexical resources for Italian, that is MWN and ItalWordNet (ITW) (Roventini et al., 2002). ITW was born as the EuroWordNet Italian database, but even though compatible to a certain extent with EuroWordNet, it is released in XML format. ITW includes about 47.000 lemmas, 50.000 synsets and 130.000 semantic relations and is currently maintained by the Institute for Computational Linguistics (ILC) at the National Research Council (CNR). An updated version is freely available online. [5]

Finally, we could easily address, at least for English, the lack of unbiased sense frequency statistics by computing them over the WordNet Gloss Corpus, in which glosses are sense-annotated.[6] This corpus alone would provide sense frequencies for 157,300 lemma-pos pairs.

Acknowledgments

This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union (2010-5094-7) and the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

References

Verginica Barbu Mititelu, Stefan Daniel Dumitrescu, and Dan Tufiş, 2014. *Proceedings of the Seventh Global Wordnet Conference*, chapter News about the Romanian Wordnet, pages 268–275.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(03):247, September.

[5] http://datahub.io/dataset/iwn

[6] http://wordnet.princeton.edu/glosstag.shtml

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Francis Bond and Giulia Bonansinga. 2015. Exploring cross-lingual sense mapping in a multilingual parallel corpus. In *Second Italian Conference on Computational Linguistics CLiC-it 2015*. to appear.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362. Association for Computational Linguistics.

Francis Bond and Kyonghee Paik. 2012. A Survey of WordNets and their Licenses. In *GWC 2012*, pages 64–71.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 1–8. Association for Computational Linguistics.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese SemCor: A sense-tagged corpus of Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 56–63.

Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. KAF: a generic semantic annotation format. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon GL 2009*, Pisa, Italy.

Stephen A. Della Pietra Vincent J Della Pietra Brown, Peter F. and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the ACL*, Morristown, NJ.

Yee Seng Chan and Hwee Tou Ng. 2005. Scaling up word sense disambiguation via parallel texts. In *AAAI*, volume 5, pages 1037–1042.

Jordi Daudé, Lluís Padró, and German Rigau. 2000. Mapping wordnets using structural information. In *38th Annual Meeting of the Association for Computational Linguistics (ACL'2000).*, Hong Kong.

Jordi Daudé, Lluís Padró, and German Rigau. 2001. A complete WN1.5 to WN1.6 mapping. In *Proceedings of NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations". Pittsburg, PA*.

Jordi Daude, Luiss Padro, and German Rigau. 2003. Validation and tuning of wordnet mapping techniques. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 255–262, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Alfio Massimiliano Gliozzo, Marcello Ranieri, and Carlo Strapparava. 2005. Crossing parallel corpora and multilingual lexical databases for WSD. In *Computational Linguistics and Intelligent Text Processing*, pages 242–245. Springer.

Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, pages 61–66. Association for Computational Linguistics.

Radu Ion. 2007. *Metode de dezambiguizare semantica automata. Aplicat ii pentru limbile englezas i romana ("Word Sense Disambiguation methods applied to English and Romanian").* Ph.D. thesis, Research Institute for Artificial Intelligence (RACAI), Romanian Academy, Bucharest.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the japanese wordnet.

Mitesh M. Khapra, Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. 2011. Together we can: Bilingual bootstrapping for wsd. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 561–569, Stroudsburg, PA, USA. Association for Computational Linguistics.

Henry Kučera and W. Nelson Francis. 1967. Computational analysis of present-day American English.

Shari Landes, Claudia Leacock, and Randee I Tengi. 1998. Building semantic concordances. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 199–216. MIT Press, Cambridge, MA.

Monica Lupu, Diana Trandabat, and Maria Husarciuc. 2005. A Romanian SemCor aligned to the English and Italian MultiSemCor. In *1st ROMANCE FrameNet Workshop at EUROLAN*, pages 20–27.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112. Association for Computational Linguistics.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 455–462. Association for Computational Linguistics.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an Aligned Multilingual Database. In *In Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India.

Adriana Roventini, Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, Rita Marinelli, Bernardo Magnini, Manuela Speranza, and Antonio Zampolli. 2002. Italwordnet: a large semantic database for the automatic treatment of the italian language. In *First International WordNet Conference*.

Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit, and Maria Grigoriadou. 2002. Balkanet: A multilingual semantic network for the balkan languages. *Proceedings of the International Wordnet Conference, Mysore, India*, pages 21–25.

Dan Tufiş, Radu Ion, and Nancy Ide. 2004. Word sense disambiguation as a wordnets validation method in balkanet. In *Proceedings of the 4th LREC Conference*, pages 741–744.

Dan Tufiş, Verginica Barbu Mititelu, Dan Ştefănescu, and Radu Ion. 2013. The Romanian wordnet in a nutshell. *Language Resources and Evaluation*, 47(4):1305–1314, December.