Exploring Cross-Lingual Sense Mapping in a Multilingual Parallel Corpus

Francis Bond
Linguistics and Multilingual Studies,
Nanyang Technological University, Singapore
bond@ieee.org

Giulia Bonansinga
Filologia, Letteratura e Linguistica,
Università di Pisa, Italy
giuliauni@gmail.com

Abstract

English. Cross-lingual approaches can make sense annotation of existing parallel corpora inexpensive, thus giving new means to improve any supervised Word Sense Disambiguation system. We compare two such approaches that can be applied to any multilingual parallel corpus, as long as large inter-linked sense inventories exist for all the languages involved.

Italiano. *La disponibilità di corpora annotati a livello semantico è cruciale nei modelli di apprendimento supervisionato per Word Sense Disambiguation. Qualsiasi corpus parallelo multilingue può essere disambiguato -almeno parzialmente- sfruttando le similarità e le differenze tra le lingue incluse, facendo ricorso a reti semantiche quali WordNet.*

1 Introduction

Cross-lingual Word Sense Disambiguation (CL-WSD) aims to automatically disambiguate a text in one language by exploiting its differences with other language(s) in a parallel corpus. Since the introduction of a dedicated task in SemEval-2013 (Lefever and Hoste, 2013), work on CL-WSD has increased, but parallel corpora have been used to this purpose for a long time; see for instance Brown et al. (1991), Gale et al. (1992), Ide et al. (2002), Ng et al. (2003) and, more recently, Chan and Ng (2005) and Khapra et al. (2011). Diab and Resnik (2002) exploit the semantic information inferred by translation correspondences in parallel corpora as a clue for WSD; Gliozzo et al. (2005) represent the milestone behind one of the approaches here evaluated, i.e. sense disambiguation exploiting the polysemic differential between two languages. As Bentivogli and Pianta (2005) pointed out, Word Sense Disambiguation (WSD) is so challenging mainly because most approaches require large amounts of high-quality sense-annotated data. Ten years later, the **knowledge acquisition bottleneck** still needs to be addressed for most languages.

Given an ambiguous word in a parallel corpus, having access to the *semantic space* (here intended as all the senses associated to its lemma) of each of its aligned translations allows one to exploit similarities and differences in the languages involved and, consequently, to make more educated guesses of the intended meaning. This simple, yet powerful, intuition can be decisive, if not in disambiguating all words, at least in reducing ambiguity and thus the human effort in annotating a whole text from scratch.

We explore two approaches of annotating a multilingual parallel corpus in English, Italian and Romanian built upon SemCor (SC) (Landes et al., 1998). We describe it in Section 2 along with a brief outline of the first approach, **sense projection (SP)**, which was pioneered by Bentivogli and Pianta (2005). In Section 3 we list the requirements and the necessary preprocessing steps common to both approaches. In Section 4 we present the second approach, **multilingual sense intersection (SI)**. Section 5 discusses the results achieved on the multilingual corpus with each method. We conclude in Section 6 anticipating future work.

2 SemCor, a corpus made multilingual by sense projection

Developed at Princeton University, SC (Landes et al., 1998) is a sense-annotated subset of the Brown Corpus of Standard American English (Kučera and Francis, 1967). SemCor includes 352 texts, each around 2,000 words long; in 186 texts all content words are annotated, while in the remaining 166 only verbs are.

MultiSemCor: Bentivogli and Pianta (2005) built an English-Italian parallel corpus by manually translating 116 texts from SC all-words component into Italian. Using the word alignment as a bridge, the Italian component was automatically sense-annotated by projection of the annotations available in English. Assuming that translations preserve the meaning of a text, if a sense-annotated source text is aligned to its translation(s), then the annotations can be transferred, as long an inter-linked sense inventory is used by all languages. In this study, a multilingual WordNet with reference to WordNet 1.6 (WN 1.6), MultiWordNet[1] (MWN) (Pianta et al., 2002), was used.

Following Bentivogli and Pianta (2005), we replicated SP on MultiSemCor (MSC) after converting all sense annotations to WordNet 3.0 (WN 3.0), as described in Section **??**.

MultiSemCor+: Lupu et al. (2005) developed the Romanian SemCor (RSC) to build MultiSemCor+, which extended MSC with aligned Romanian translations. The MSC+ originally presented consists of 34 translations aligned to English (Lupu et al., 2005). Since then, the English-Romanian parallel corpus based on SC has grown, currently consisting of 81 texts (82 in the version released) (Ion, 2007) annotated following WN 3.0. Of these, 50 have Italian translations in MSC.

In conclusion, SP can bootstrap the creation of sense-annotated parallel corpora by exploiting existing resources in well-represented languages, with word alignment and connected sense inventories as the only requirements.

3 Preprocessing and requirements

Mapping to WN 3.0: As a preprocessing step, we mapped all annotations in MSC to WN 3.0. This is convenient in itself, as the corpus will be redistributed with reference to a widely used sense inventory, as comparison with related work will be easier. The English component is annotated with *sense keys*, stable across different WN versions, so the conversion was straightforward. On the sense keys alone, 95% of the WN 1.6 synsets can be correctly mapped to WN 3.0.[2] The Italian texts use an offset-based encoding that is not consistent across WN versions; fortunately, there are freely available mappings[3] inferred by exploiting both graph and non-structural information (Daudé et al., 2000; Daudé et al., 2001).

Sense inventories: Table 1 shows the coverage of WNs for our target languages. The Open Multilingual WordNet (OMW)[4] 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).

Another valid option for the multilingual sense inventory would be BabelNet, created from the automatic integration of WN 3.0, OMW, Wikipedia and many other resources (Navigli and Ponzetto, 2012), with an estimated accuracy of 91% for the WN-Wikipedia mapping (Navigli et al., 2013). However, we chose to use OMW since we wanted to test our hypothesis on resources that were purposely built to be mapped to one another.

The Romanian WordNet (RW) was created within the BalkaNet project (Stamou et al., 2002). The current version has 59,348 synsets in its latest release (Barbu Mititelu et al., 2014). The synsets were mapped to WN 3.0 with precision of 95% (Tufiş et al., 2013).

	Synsets	Senses
English	117,659	206,978
Italian	34,728	69,824
Romanian	59,348	85,238

Table 1: Coverage of the WNs used.

Aligning RSC to MSC: RSC is not word-aligned to any component of the parallel corpus, so it fails in meeting a necessary requirement to perform sense mapping. However, as the sentence alignment is available, we attempted to align all Romanian sense-annotated words to their English and Italian counterparts. For each aligned sentence pair, we first align all candidate pairs sharing the same sense annotation. If any words are left unaligned after this step, the remaining alignments are inferred by taking into account PoS information and synset similarity scores. Suppose the first step alone has aligned all Romanian content words but one, and that the corresponding English sentence has three content words left that are candidates for the alignment. Then, the aligner computes the most likely match by looking for

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

[2] According to the HyperDic project: http://www.hyperdic.net/en/doc/mapping

[3] http://www.talp.upc.edu/index.php/technology/tools/45-textual-processing-tools/98-wordnet-mappings/

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

PoS correspondence and for higher proximity in the WN network, by looking at a combination of the *path similarity score* and the *shortest path distance*. This latter alignment strategy (the only possible source of errors) achieved 97% precision on a small sample (12%) of the alignments found.

4 Multilingual Sense Intersection

Unlike SP, SI does not require any of the texts in a parallel corpus to be sense-annotated, so it can be applied to a wider range of existing resources. Its logical foundation is in that a polysemous word in a language is likely to be translated in different words in other languages, so the comparison with the semantic space of each translation should help select the sense actually intended. Consider, for instance, the problem of disambiguating the English word *administration* in Example 1.

(1) EN *The jury praised the administration and operation of the Atlanta Police Department.*

IT *Il jury ha elogiato l'amministrazione e l'operato del Dipartimento di Polizia di Atlanta.*

RO *Juriul a lăudat administrarea şi conducerea Secţ iei de poliţie din Atlanta.*

Given the alignments, we can retrieve the set of synsets associated with the lemmas in the Italian and Romanian translations. Figure 1 shows how the intersection helps detecting the correct sense, which is the only one shared by all the lemmas.

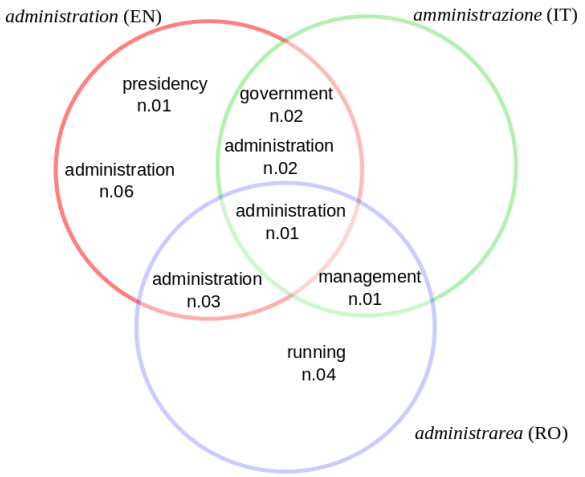

Figure 1: Disambiguation via SI

Most often, however, such a comparison will only partially reduce the ambiguity, especially as such a fine-grained sense inventory as WN is used. Yet, other approaches (employment of human annotators, or recourse to baselines) can be applied in a second phase to solve the disambiguation task, once it has been simplified.

The algorithm disambiguates one side of our multilingual parallel corpus at a time, having as target all texts aligned with at least one other component.[5] Table 2 displays the basic statistics of each corpus and, for the sake of clarity, the number of words to be annotated (target words) before the migration to WN 3.0, as the changes in the WN structure do not set ideal conditions for a meaningful comparison with previous work.

We use sense frequency statistics (SFS) whenever the target word is not fully disambiguated. These were calculated over all texts in the corpus **except** the one being annotated.

	#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	48,364

Table 2: Statistics for each text in the multilingual parallel corpus.

%	EN	IT	RO
Disambiguated	27.15	30.92	36.67
MFS-Subset	34.39	26.51	12.89
MFS-Overlap	13.59	26.69	50.45
No alignment	24.14	12.08	-
No match	0.67	0.65	-
No synset found	0.05	3.14	-

Table 3: Distribution of SI outcomes.

Algorithm: Given an ambiguous target word, each of its aligned translations in the parallel sentences contributes to the disambiguation process by bringing in all its 'set of senses' retrieved from the inter-linked sense inventory.

Intersection is then performed over each non-empty set retrieved. If the *overlap* only consists of one sense, then the target word is `Disambiguated` (see Table 3). If the overlap contains more than one sense, then it is further intersected with the set of most frequent senses available for the target lemma. If resorting to MFS statistics leads to an overlap containing one sense, the word is disambiguated (`MFS-Subset`); if the overlap still results in more than one sense, the

[5] With the exception of the English corpus, which we have considered made of the 116 texts included in MSC.

Method	English		Italian		Romanian	
	Precision	Coverage	Precision	Coverage	Precision	Coverage
MFS (baseline)	**0.761**	**0.998**	0.599	0.999	0.531	1
SP	-	-	**0.971**	**0.927**	**0.903**	**1**
SP (Bentivogli & Pianta)	-	-	0.879	0.764	-	-
3-way SI	0.750	0.778	0.653	0.915	0.590	1

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

most frequent one among the ones left is selected (`MFS-Overlap`). In the rare case in which no other language contributes to disambiguate, we assign the current target lemma its MFS. Disambiguation also fails when no match, synset or alignment is found. See also Table 3 for the distribution of all of the possible scenarios that may emerge.

5 Evaluation and discussion

Table 4 shows the precision and coverage scores achieved with the approaches here analyzed, along with the Most Frequent Sense (MFS) baseline. We report the original results for SP (Bentivogli and Pianta, 2005) and ours after the mapping to WN 3.0; we evaluate on different figures (see Table 2) as a part of the original annotations was lost in the mapping process. We performed SP also on the current release of RSC for completeness.

Coverage is overall reasonably high for all languages with SI and very high with the baseline. On the other hand, the precision achieved resorting to SFS is significantly lower for Italian, which makes more valuable the not very high score obtained by SI. Average ambiguity reduction is 54% (EN), 53% (IT) and 55% (RO).

Although SI and MFS perform comparably, we remind that SFS were computed on the same corpus, which is also not extremely large. Thus, we would expect MFS to compare at least slightly worse in more general cases (unfortunately, external statistics are hard to come by). This would make SI a valid and inexpensive cross-lingual disambiguation approach. We also performed 2-way intersection for each corpus pair. We find a slight decrease in precision (of 0.01 to 0.03) compared to the three-way intersection, depending on the corpus. While further restricting the semantic space does help in reducing ambiguity, the improvement is not striking. According to our error analysis, this is corpus-dependent, as the manually assigned correct senses against which we evaluate are very specific. Instead, as the WNs vary largely in coverage, senses found by intersection, though actually shared in all languages, are close, but not quite the same, to the very specific ones selected by the human annotator. In conclusion, coarse-grained evaluation would give a higher score, and in general the senses found by intersection would be just good enough in most cases. Also, as Italian and Romanian are quite similar, we would expect more differences if we added a language from a different language family.

6 Conclusions

To our knowledge, this is the first attempt to disambiguate a parallel corpus by using multilingual SI. The more languages are considered, the more ambiguity should be reduced and the better SI is expected to perform. In future work, we plan to include the Japanese SemCor (Bond et al., 2012) to test our hypothesis that translations from a different language family will discriminate further. We also plan to use a different parallel corpus built on open translations of *The Adventure of the Speckled Band* by Sir Arthur Conan Doyle. We will also try to calculate SFS from untagged text, following McCarthy and Carroll (2003).

Furthermore, we are investigating alternative ways to solve the ambiguity left whenever SI does not lead to a single synset; for instance, we plan to apply some implementation of Lesk (Lesk, 1986) on the subset found by SI. Finally, we aim to port to WN 3.0 the sense clustering carried out by Navigli (2006) to perform a coarse-grained evaluation, which would ignore minor sense distinctions. An initial comparison with Babelfly (Moro et al., 2014) would certainly be enlightening as well.

All data and scripts derived by our work will be made available, except for those derived from RSC, as its license currently forbids it.

Acknowledgments

This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2010-5094-7.

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.

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, 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.

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

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. 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 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.

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.

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.

Els Lefever and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. *Proc. of SemEval*, pages 158–166.

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. ACM.

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.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

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, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation.

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.

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