Crossing Parallel Corpora and Multilingual Lexical Databases for WSD

Alfio Massimiliano Gliozzo, Marcello Ranieri, and Carlo Strapparava

ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica, I-38050 Trento, ITALY
{gliozzo,ranieri,strappa}@itc.it

1 Introduction

Word Sense Disambiguation (WSD) is the task of selecting the correct sense of a word in a context from a sense repository. Typically, WSD is approached as a supervised classification task to get state-of-the-art performance (e.g. [1]), and thus a large amount of sense-tagged examples for each sense of the word is needed, according to the word-expert approach. This requirement makes the supervised approach unfeasible for "all-words" tasks, consisting on disambiguating all the words in texts. This problem has been called the Knowledge Acquisition Bottleneck and many solutions have been proposed for it (see for example [2]) .

In this paper we propose the use of aligned corpora and multilingual lexical databases to automatically acquire sense tagged data, exploiting the polisemic differential between two (or more) languages.

Even though the underlying idea of the approach proposed in this paper is not totally original in the WSD literature (see for example [3, 4]) our basic contribution is to show how far we can go in using parallel corpora to collect sense tagged data, by reporting both a quantitative and a qualitative evaluation. It will be shown that having an "ideal" aligned wordnet (i.e. a lexical resource such that all the sense distinctions in one language are reflected in the other), our simple strategy allows to disambiguate 51% of the English/Italian aligned pairs of words with 100% precision, while with the available resources this figures decreases to 67% precision for a subset of 40% words. In the rest of the paper we will evaluate this technique by exploiting two resources recently developed at ITC-irst: MultiWordNet and MultiSemCor.

2 MultiWordNet and MultiSemCor

MultiWordNet (http://multiwordnet.itc.it) is a multilingual computational lexicon, conceived to be strictly aligned with the Princeton WordNet. In our experiment we used the English and the Italian components. The last version of the Italian WordNet contains around 58,000 Italian word senses and 41,500 lemmas organized into 32,700 synsets aligned whenever possible with WordNet English synsets.

The MultiSemCor [5] (http://multisemcor.itc.it) corpus originates from the Princeton SemCor corpus. SemCor texts were taken from the Brown Corpus, and

were semantically annotated according with the synsets of WordNet. MultiSemCor has been built starting from a subset of the SemCor texts. 116 English texts were translated into Italian by professional translators. Then, the original texts and their translations were automatically aligned at the word level. Finally the annotations were transferred from each text to its alignment, creating a bilingual parallel corpus endowed with semantic annotation (about 116,000 semantically annotated English tokens, about 90,000 semantically annotated Italian tokens, being MultiWordNet the shared repository of senses).

3 A Bilingual WSD Algorithm

In this section we describe an unsupervised WSD technique that uses aligned corpora and multilingual lexical databases to automatically acquire sense tagged data, exploiting the polisemic differential between two languages. The basic assumption is that if two texts are one the translation of the other, they should refer to the same facts, and then words contained in them should refer to the same concepts. An aligned multilingual lexical resource (e.g. MultiWordNet) allows us to automatically disambiguate aligned words in both languages by simply intersecting their senses. If the intersection contains only one sense, then the words in both languages will be fully disambiguated, while if the cardinality of the intersection is higher, the words still remain ambiguous. In any case the number of possible senses is often sensibly reduced. For instance if the English word *soccer* is aligned with the Italian word *calcio*, the correct sense is "a football game" and not, for example, "a white metallic chemical element" (one of the four senses of the Italian word).

More formally let $S = \{c_1, c_2, \ldots, c_n\}$ be the set of aligned pairs of English/Italian lemmas such that $c_i = (l_i^E, l_i^I)$, $s(l)$ a function returning the set of senses corresponding to the lemma l, and $I(c_i) = s(l_i^E) \bigcap s(l_i^I)$ the intersection of the synsets corresponding to the two lemmas in the two languages. The function $WSD_{strict}(c_i)$, defined by equation 1, fully disambiguate the word pair if the intersection is a singleton.

$$WSD_{strict}(c_i) = \begin{cases} I(c_i) & : \texttt{if } |I(c_i)| = 1 \\ \emptyset & : \texttt{otherwise} \end{cases} \quad (1)$$

Equation 2 returns the set of all the possible senses.

$$WSD_{soft}(c_i) = I(c_i) \quad (2)$$

4 Evaluation and Discussion

We compared the results with a random baseline, being our method completely unsupervised. We also try to define an upper bound, assuming that all the senses annotated in the corpus are actually in the Italian WordNet. We evaluated our WSD method on the following two subsets of the original aligned pairs of lemmas in MultiSemCor. Let $G(c_i)$ be the gold standard function returning the correct sense annotated in MultiSemCor for c_i.

Evaluation	Language	Precision	Coverage	F1	#Valid
Ideal	both	1	0.51	0.68	39983
All	both	0.67	0.40	0.38	71421
Ideal-polysemous	English	1	0.39	0.56	32277
All-polysemous	English	0.56	0.37	0.30	61712
All-polysemous (random baseline)	English	0.22	1	0.22	61712
Ideal-polysemous	Italian	1	0.32	0.48	28890
All-polysemous	Italian	0.61	0.31	0.29	49206
All-polysemous (random baseline)	Italian	0.17	1	0.17	49206

Table 1. Multilingual WSD evaluation on word pairs and on polysemous words

Ideal: Only couples such that the gold standard annotation is a possible sense for the lemmas in both languages $S_C = \{c_i | G(c_i) \in s(l_i^E) \text{ and } G(c_i) \in s(l_i^I)\}$.
All: Only couples such that both lemmas are contained in MultiWordNet $S_A = \{c_i | s(l_i^E) \neq \emptyset \text{ and } s(l_i^I) \neq \emptyset\}$.

We distinguish among results for word pairs, polysemous terms in English and polysemous terms in Italian (see Table 1). As expected our WSD method is perfect (i.e. precision 100%) in the **Ideal** evaluation dataset, in which the sense in the Gold Standard is also a possible sense for the Italian lemma. Unfortunately this is not a realistic case, because the Italian resource does not still have the coverage of the English one (i.e. in the Italian part of MultiSemCor a word could be annotated with a sense not reachable from the lemma in Italian MultiWordNet). Thus in the **All** dataset, the precision of the algorithm drops to 0.67 for word pairs. We also evaluated the precision and coverage of the WSD algorithm only by considering polysemous words in English and Italian, and the results where encouraging (i.e. precision is the important feature in the case of acquisition of sense-tagged examples). Table 2 displays the polysemy reduction using the formula 2.

WSD	Pol-ENG	Pol-ITA	Pol-RES	Precision	Coverage	#Valid
Ideal	5.62	3.35	1.98	1	1	39983
All	6.72	3.28	1.54	0.56	1	71421

Table 2. Polysemy reduction using soft multilingual disambiguation

We showed that in the "ideal" case the methodology allows to disambiguate with 100% precision, while with available lexical resources the precision drastically drops to about 60%. A qualitative analysis (see Table 3) of 100 errors (randomly selected) showed that in about 77% of the errors are caused by *not covered* senses (i.e. senses in Italian that should be included in the resource even though they are not actually represented in MultiWordNet).

Causes of errors	# of cases
Senses not covered by the Italian WordNet	77
Alignment errors	7
Inter-lingual differences	16

Table 3. Qualitative analysis of 100 errors (randomly selected)

5 Conclusions and Future Works

In this paper an unsupervised WSD methodology has been presented. This methodology can be applied to parallel corpora allowing to fully disambiguate about the 50% of words, without requiring any external knowledge. Obviously the same approach can be applied also to aligned corpora composed by texts written in more than two languages. Intuitively the probability to obtain a smaller intersection among senses of a translated word in three (or more) languages is higher than the one for only two languages. For the future we plan to automatically acquire the most frequent *not covered* senses by exploiting MultiSemCor in order to improve the WSD performances in "real" parallel corpora, and to apply it extensively to disambiguate large scale parallel corpora (e.g. EuroParl [6]), in order to automatically acquire sense tagged data to train a supervised disambiguation system to be used in an "all-words" task.

Acknowledgments

We would like to thank Emanuele Pianta for useful discussions. This work was partially supported by the *Meaning* European Project (IST-200134460).

References

1. Strapparava, C., Gliozzo, A., Giuliano, C.: Pattern abstraction and term similarity for word sense disambiguation: Irst at senseval-3. In: Proc. of SENSEVAL-3, Barcelona, Spain (2004)
2. Mihalcea, R., Moldovan, D.: An automatic method for generating sense tagged corpora. In: Proc. of AAAI 99, Orlando, FL (1999)
3. Magnini, B., Strapparava, C.: Experiments in word domain disambiguation for parallel texts. In: Proc. of "Word Senses and Multi-Linguality", Hong Kong, Workshop held in conjunction of ACL2000 (2000)
4. Diab, M., Resnik, P.: An unsupervised method for word sense tagging using parallel texts. In: Proc. of ACL 02, Philadelphia (2002)
5. Bentivogli, L., Pianta, E.: Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus. Journal of Natural Language Engineering (NLE), Special Issue on Parallel Texts. (To appear)
6. Koehn, P.: EuroParl: A multilingual corpus for evaluation of machine translation. (Unpublished (http://people.csail.mit.edu/˜ koehn/publications/europarl.ps))