Opportunistic Semantic Tagging

Luisa Bentivogli, Emanuele Pianta

ITC-irst
Via Sommarive 18, 38050 Povo (Trento) -ITALY
{bentivo,pianta}@itc.it

Abstract

Building semantically annotated corpora from scratch is a time consuming activity requiring very specialized resources. In this paper we present a pilot study carried out to test a methodology that can be used to create a semantically annotated corpus by exploiting information contained in an already annotated corpus. The main hypothesis underlying the proposed methodology is that, given a text and its translation into another language, the translation preserves to a large extent the meaning of the source target. This means that if one of the two texts is already semantically tagged, and if we can align at the appropriate level the parallel texts, it should be possible to transfer the semantic annotation from the tagged text to its translation. More specifically, in our experiment we considered word level semantic annotation. The pilot study has been carried out on six texts taken from the SemCor corpus and their Italian translations. To test the methodology we implemented an annotation transfer system based on an English/Italian word aligner, developed at ITC-irst, which relies mostly on information contained in bilingual dictionaries.

1. Introduction

Semantically annotated corpora are useful for a variety of tasks in the fields of corpus and computational linguistics. For instance, corpora annotated with sentence level semantic representations are used in the field of interlingua based Machine Translation (Levin et al, 2000). A corpus annotated with word senses can be used to build conceptual concordancers (Fellbaum, 1998; Bentivogli et al., 2001a), or to train word sense disambiguation systems (Senseval 2).

Unfortunately, building semantically annotated corpora from scratch is a time consuming activity requiring very specialized resources. It is difficult to find official data about the time required to semantically annotate any of the existing corpora. However the personal experience of the authors and their colleagues at ITC-irst indicates that indeed manual semantic annotation is a highly time-consuming activity. For instance we estimate that to annotate 80,000 tokens of the SI-TAL Italian Treebank (Mana and Corazzari, 2002) with word-senses required more than one person-year. To annotate the dialogs of the NESPOLE! database with sentence level semantic representations (currently around 6,500 sentence units) required at least one person-year, not taking into consideration speech to text transcription. Even more problematic is the issue of the expertise required by the annotators. To annotate a corpus with semantic information requires specific skills and very specialized training which is not provided by current academic curricula. Thus if one plans to semantically annotate a corpus, he/she needs to provide not only time for the annotation itself but also a non-trivial amount of time and resources for training annotators. This state of affairs makes it clear that any strategy reducing the cost of producing manual annotated corpora would be highly beneficial to the field.

In this paper we present a pilot study carried out to test a methodology that can be used to create a semantically annotated corpus by exploiting information contained in an already annotated corpus. The main hypothesis underlying the proposed methodology is that given a text and its translation into another language (where the syntactic structures of the two texts are language-specific), the semantic information is mostly preserved during the translation process. This means that if one of the two texts is already semantically tagged, and if we can align the parallel texts at the appropriate level, it should be possible to transfer the semantic annotation from the tagged text to its translation. More specifically, we considered word level semantic annotation in our experiment.

As a matter of fact, often the texts of existing annotated corpora are not translated into other languages. Our assumption is that, even in this case, manually translating the annotated corpus and carrying out the cross-lingual annotation transfer may be preferable to hand-labeling a corpus from scratch. When compared to manual semantic tagging, translation is in fact less time consuming and requires human resources which are more easily available.

2. The MultiSemCor project

The MultiSemCor project aims at building an English-Italian parallel corpus based on SemCor (Fellbaum, 1998), a subset of the English Brown corpus containing almost 700,000 running words. In SemCor all the words are tagged by PoS, and more than 200,000 content words are also lemmatized and sense-tagged according to WordNet. To build MultiSemCor we intend to apply the following steps:

- Get the Italian *translations* of the SemCor texts
- *Align* Italian and English texts at sentence and word level
- *Transfer* the word sense annotations from English to the aligned Italian words

The final result of the project will be an Italian corpus annotated with PoS, lemma and word sense, but also an aligned parallel corpus lexically annotated with a shared inventory of word senses.

2.1. Problematic issues

The project raises a number of practical and theoretical issues that need to be settled in order to provide evidence for its feasibility. The first problem is that the Italian

translations of most SemCor texts are not available. To solve this problem we will ask professional translators to translate the texts. For reasons explained in the introduction, we think that, even when accurate manual translation is required, the transfer annotation methodology is preferable to manual semantic annotation from scratch. The cross-lingual annotation transfer has the further advantage of producing a parallel corpus aligned at the word level with a shared inventory of senses.

A second, more theoretical, issue concerns the legitimacy of transferring word senses from one language to another. To what extent are the lexica of different language comparable? A study carried out in the frame of the MultiWordNet project has shown that the vast majority of English words have an Italian cross-language synonym. According to this study, only 7.8% of the English words correspond to lexical gaps in Italian (Bentivogli and Pianta, 2000). This figure shows that transferring word meaning from English to Italian is a reasonable move, but shows also that there will be a relatively small number of cases in which the transfer will not be possible.

A third issue is related to the nature of translational language. Even if we assume that translations are very accurate, some studies show that the language of translated texts has a number of peculiarities that set it apart from the language of original, non translated texts. For instance, translated texts tend to be more explicit, less ambiguous, and grammatically and lexically more conventional than source texts (Baker, 1993). As a consequence MultiSemCor will include Italian texts which are not fully representative of the general use of language in the same way as the original SemCor is. However, we think that what is really important is that the translated texts are good written texts, even if they only partially use the potentialities of the current Italian language. Note that the use of a translated corpus would be much less plausible if the original texts were conversational speech transcriptions (as in the case of the NESPOLE! database). Preserving the features of spontaneous speech through the translation process is much more difficult than preserving most features of written, formal language. Thus, we can still maintain that MultiSemCor will include current, largely representative, annotated Italian texts, which will be as useful as annotated original texts for tasks such as semantic concordancing and training of word disambiguation systems.

A further issue related to the translation process can be formulated through the following question: to what extent does translation preserve lexical meaning? The issue at stake here is not the abstract comparability of the lexica of two languages, but what happens practically in the translation process. In other words, the translator can use, for a number of different reasons, translation equivalents that are not cross-language synonyms, even when such equivalents exist in the target language. In all these cases, transferring the word sense annotation from the source to the target language would not be legitimate. Assessing to what extent this happens is one of the aims of the pilot study that will be described in the next section.

Finally, the feasibility of the entire MultiSemCor project depends heavily on the availability of an English-Italian word aligner with very good performance in terms of recall and, more importantly, precision.

3. The MultiSemCor pilot study

To assess the feasibility of MultiSemCor, a pilot study has been carried out on a sample of 6 SemCor texts containing about 12,000 English running words. Four texts were used in the development stage of the annotation transfer system, while the other two were used as unseen test for evaluation purposes. Table 1 illustrates the composition of our pilot corpus.

Three texts in the development set were taken randomly from distinct components of the “Informative Prose” section of SemCor, while the fourth was taken randomly from the “Imaginative Prose” section. For the test set, one text was taken randomly from the “Informative” section and the other from the “Imaginative” section.

		Text type	Free Trans.	Controlled Trans.	Manual Alignment	Inter-coder Agreem.
Develop. Set	F-03	Informative	1	-	1 Free	-
	B-13	Informative	1	-	1 Free	-
	G-11	Informative	1	1	2 Free	87%
					2 Contr	92%
	P-12	Imaginative	1	1	1 Free + 1 Contr	-
Test Set	J-53	Informative	1	1	1 Free + 1 Contr	-
	L-10	Imaginative	1	1	1 Free + 1 Contr	-
	Total		6	4	12	

Table 1: Composition of the pilot corpus

In order to test to what extent the characteristics of the translation can affect the methodology, for 4 out of the 6 English texts we decided to experiment with two different Italian translations. In the first instance, a translator did a completely *free* translation of the English texts. Second, another translator was asked to do a *controlled* translation, using preferably the same dictionaries used by the word alignment system, and trying to maximize, whenever possible, the lexical correspondences between the source and target texts. The translator was also told that the controlled translation criteria should never be followed to detriment of good Italian prose.

Assuming that free translations are less suitable for word alignment, if we also obtain good results with them, it should be possible to apply the methodology to already existing parallel corpora.

As Table 1 shows, by translating the 6 texts of the sample corpus (2 in free modality and 4 in both modalities) we obtained a set of 10 English-Italian text pairs. All 10 pairs were manually aligned following a set of alignment guidelines (Bentivogli et al., 2001b) which have been defined taking into account the work done in similar word alignment projects (Merkel, 1999; Melamed, 2001).

Annotators were asked to align different kinds of units as illustrated in Table 2.

	English	Italian
Simple words	health	salute
Segments (multiwords)	rain dance	danza della pioggia
Segments (generic phrases)	open-mouthed	con la bocca spalancata
Parts of words	clasping *him*	afferrando*lo*

Table 2: Kinds of unit to be aligned

Also, the annotators had to mark different kinds of semantic correspondence between the aligned units, as illustrated in Table 3.

	English	Italian
Full (synonymic)	science	scienza
Non-synonymic	meaning	motivo (*reason, grounds*)
Trans-PoS non synonymic	dream previsions	sogni premonitori (*premonitory dreams*)
Fuzzy	the dreamer sees	una persona sogna (*a person dreams*)
Involving extra grammatical elements	my hands	(le) mie mani
	(he) wants	vuole
omissions	the (ocean of) mankind	il genere umano

Table 3: Kind of correspondences between aligned units

Table 4 shows an example of the Excel worksheet that was used by the annotators to manually align the sample corpus. To check inter-annotator agreement we asked two annotators to align the same text G-11 with both the free and the controlled translations (see Table 1). The agreement rate has been calculated with the following formula (Véronis and Langlais, 2000):

$$\text{Agree:} \quad 2 \frac{\text{N of common units}}{\text{N of units in the two texts}}$$

The agreement rate turned out to be 87% for free translations and 92% for controlled translations. These agreement measures can be considered satisfactory. As expected, controlled translations produced a better agreement rate between annotators.

4. Automatic word alignment and annotation transfer

Word alignment is a crucial step in the methodology proposed to build MultiSemCor. Within the pilot study we used KNOWA (KNOwledge-intensive Word Aligner), an English/Italian word aligner, developed at ITC-irst, which relies mostly on information contained in the Collins bilingual dictionary, available in electronic format. For each sentence pair KNOWA produces word aligments according to the following strategy:

- The *morphological analysis* produces a set of candidate lemmas for each English and Italian word.
- The candidate lemmas are ordered from the most to the least probable by means of a rule-based *PoS ordering* algorithm.
- A three phase incremental *alignment procedure* takes as input the two sentences annotated with sets of ordered candidate lemmas and outputs a set of pairwise word alignments.

The alignment procedure is crucially based on the relation of potential correspondence between English and Italian word tokens:

Given an English word token EW *and an Italian word token* IW, IW *is the* potential correspondent *of* EW *if one of the candidate lemmas of* EW *is the* translation equivalent *of one of the candidate lemmas of* EW, *according to a bilingual dictionary.*

The potential correspondence relation holds between word tokens, but is relative to a lemma pair. For instance we say that the word tokens *dreams* and *sogna* are potential correspondents relative to the lemma pair `<dream/verb, sognare/verb>`. Two word tokens can be potential correspondents relative to more than one lemma pair. For instance the word tokens *dream* and *sogno* are potential correspondents relative to the two lemmas pairs `<dream/verb, sognare/verb>` and `<dream/ noun, sogno/noun>`. In fact *dream* and *sogno* can be either first singular person of the verb *to dream* and *sognare*, or singular forms of the noun *dream* and *sogno* respectively.

The correspondence relation is called potential because it refers to tokens occurring in real texts rather than abstract word types. In real texts tokens that are potential correspondents may not in fact be translations of each other.

English words	*Eng. segments*	*Ita PoS*	*Italian words*	*Ita. segments*	*Word alignment*
1-There	1:mw	cli	1-Ci	1:mw	1-There
2-are	1:mw	v	2-sono	1:mw	2-are
3-certainly		p	3-,	p:nt	
4-large		adv	4-certamente		3-certainly
5-areas		p	5-,	p:nt	
6-of		adj	6-vasti		4-large
7-understanding	2	n	7-ambiti		5-areas
8-in		prep+art	8-della	2:gr:art=>la	6-of=>di
9-the		n	9-conoscenza	2	7-understanding
10-human	3:mw	prep+art	10-nelle		[8-in=>in,9-the=>le]
11-sciences	3:mw	n	11-scienze	3:mw	11-sciences
12-which		adj	12-umane	3:mw	10-human

Table 4: An example of the manual aligment

In the *first phase* of the alignment procedure the potential correspondence relation is exploited in the English to Italian direction:

For each English word *EW* in a certain position *P*:

1. Get the most probable candidate lemma of *EW*
2. Get the Italian word *IW* in the same position *P*
3. Check if *IW* is a potential *EW* relative to the current English candidate lemma
4. If yes, align *EW* and *IW* and record their lemmas
5. Otherwise consider the next probable candidate lemma of EW and go back to step 2
6. If no aligment is found, progressively extend the Italian word window and go back to step 1.

By extending the Italian word window we mean considering Italian words in position *P* ± *Delta*, where *P* is the position of the English word and *Delta* can vary from 1 to a *Max* value. The value of *Max* is adjustable (it was 5 in the experiment). Note that if the alignment is not found within the Italian word window, the English word is left unaligned. In the following table the box in the Italian column shows the maximal text window in which the potential correspondent of *dream* is searched.

…	…
9-the	9-l'
10-exact	10-esatta
11-pattern	11-riproduzione
12-of	12-di
13-a	13-un
14-previous	14-**sogno**
15-dream	**15**-precedente
16-we	16-abbiamo
17-have	17-un
18-an	18-caso
19-instance	19-di
20-of	20-deja_vu
21-deja_vu	21-,
…	…

Table 5: An example of a maximal text window

The search starts from *precedente* and ends after the first extension of the text window as *sogno* can be found in position *P-1*.

In the *second phase* of the alignment procedure the potential correspondence relation is exploited from Italian to English: For each Italian word which has not been aligned in the first phase, the same procedure is applied as above.

In the *third* and last phase, the algorithm tries to align the words which are still unaligned, resorting to the graphemic similarity of the Italian and English words. See (Yzaguirre, 2000) for a similar approach.

Note that given the way in which the alignment procedure works, each time an alignment is found it implies also selecting a PoS and a lemma for both English and Italian words. The selected PoS and lemma can be different from the ones that were considered most probable by the PoS ordering algorithm.

The KNOWA algorithm needs to be able to cope with at least two problematic aspects. The first are multiwords. To work properly, KNOWA needs to identify them in the source and target sentences, and needs knowledge about their translation equivalents. We have tried to exploit the information about multiwords contained in the Collins bilingual dictionary. However it is well known that dictionaries contain only a small part of multiwords actually used in language. Thus there is still wide room to improve KNOWA's capability to handle multiwords.

The second problematic aspect has to do with multiple potential correspondence relations. Given a source word in one language, more than one potential correspondent can be found within the maximal word window in the target language. This is particularly true if we pursue full text alignment. Whatever the number of potential correspondents, the alignment procedure selects the potential correspondent whose position is nearest to the position of the source word by first considering the most probable PoS of the source word. Unfortunately, the potential correspondent selected in this way is not always the right one. Thus multiple potential correspondents can be a source of alignment errors for KNOWA.

4.1. KNOWA for MultiSemCor

In the previous section the main characteristics of the KNOWA word aligner have been illustrated. When applied to the parallel texts of MultiSemCor, KNOWA needs some adaptations that in fact make its task easier. Unlike a generic pair of parallel texts, in the MultiSemCor case the words of the source text are already PoS tagged and lemmatized. More specifically, the SemCor texts have been first automatically tagged with the Brill PoS tagger, then all the content words have been manually checked and lemmatized. This fact makes the issue of multiple potential correspondents less relevant, as one of the two sources of lemma ambiguity is fixed.

Also the other problematic issue for KNOWA, i.e. multiwords, is made easier by the manner in which SemCor has been annotated. In fact, in SemCor multiwords included in WordNet have already been marked. See for instance *deja_vu* in Table 4. This implies that KNOWA does not need to recognize English multiwords.

Finally, there is another aspect that makes the alignment of MultiSemCor texts easier than the general case of full text alignment. Word alignment is done in MultiSemCor with the final aim of transferring lexical annotations from English to Italian. However, only content words have word sense annotations in SemCor. Thus it is more important that KNOWA behaves correctly on content words, which are admittedly easier to align than functional words. In fact in the pilot study we tried to also align functional words because we wanted to check how effective the described word alignment mechanism was in order to select the correct lemma and PoS of all the words of Italian texts. In other words we wanted to check if we can get an Italian corpus fully PoS tagged and lemmatized with acceptable precision, as a side product of the annotation transfer strategy.

4.2. Annotation transfer

Once the word alignment has been performed, the annotation transfer is a trivial task that can be described in the following way:

For each English-Italian word pair
1. Copy the sense annotation (if any) from SemCor to the Italian text.
2. Add lemma and PoS as selected during the alignment process. See Step 4 of the alignment procedure.

5. Evaluation

The performance of KNOWA applied to MultiSemCor has been evaluated comparing its output to the gold standard obtained by manually aligning the test set. The usual notions of Precision, Recall, and Coverage are defined as follows:

$$\text{Precision: } \frac{\text{N of correct alignments}}{\text{N of English words aligned}}$$

$$\text{Recall: } \frac{\text{N of correct alignments}}{\text{N of English words to be aligned}}$$

$$\text{Coverage: } \frac{\text{N of English words aligned}}{\text{N of English words to be aligned}}$$

The performance of KNOWA in a full-text alignment task is shown in the following table.

	Precision	Recall	Coverage
Free	0.72	0.46	0.63
Controlled	0.79	0.55	0.69

Table 6: KNOWA on Full-text

These results, which compare well with those reported in the literature (Ahrenberg et al., 2000; Véronis, 2000) show that, as expected, a controlled translation allows a better alignment.

However, as our purpose is the transfer of the semantic tagging from SemCor to the aligned Italian corpus, a more significant evaluation can be done by taking into account only English content words which have a semantic tag in SemCor. We can see that (ignoring function words) the performance of the word aligner improves.

	Precision	Recall	Coverage
Free	0.91	0.64	0.70
Controlled	*0.94*	*0.75*	*0.80*

Table 7: KNOWA on sense tagged words only

The second row in Table 7 shows the most interesting results in relation to the aims of the MultiSemCor project. The datum about Coverage indicates that after applying the automatic annotation transfer there still remain 20% of the content words that need to be manually annotated if we want to complete the annotation of content words.

5.1. Non synonymous translation equivalents

Note that the alignment process could put in correspondence words which are not cross-linguistic synonyms. If the translation is not synonymic and the two words are aligned, the transfer of the sense from English to Italian is not correct.

For example in a sentence of our gold standard the English word *meaning* was aligned with the Italian word *motivo* (reason, grounds) which is correct in that specific context but is not a synonymic translation of the English word. In our gold standard non-synonymous alignments have been marked. They amount to 3.1% of the total alignments.

The word aligner evaluation procedure takes this phenomenon into account, considering non-synonymous alignments as errors. However, unlike what can be expected from statistics-based word aligners, KNOWA makes very few errors of this kind. The reason is that it relies on bilingual dictionaries where non-synonymous translations are quite rare.

5.2. PoS tagging

As explained in Section 4 we expect that one of the side results of MultiSemCor be an Italian corpus fully lemmatized and PoS tagged. To check this hypothesis,

during the manual annotation of the gold standard annotators were also asked to specify the PoS of the Italian words. We checked the lemmatization and PoS tagging resulting from word alignment against the gold standard. The results show a precision of 91%. We think that this level of precision is not yet satisfactory, but we are confident that by improving the PoS ordering mechanism we will obtain better results.

6. Related work

In the literature many methods are proposed for the automatic semantic annotation of corpora. Most of them use monolingual material, while some others try to annotate texts in one language by using translations in other languages as a source for sense distinctions. However, we are not aware of works investigating the possibility of exploiting semantic annotation already available in one corpus to transfer it to an unannotated corpus.

The work of Diab (2000) is the most related to our work. Diab presents an unsupervised method for word sense tagging of both the source and target texts of a parallel corpus. Her method relies on translations as a source of word sense disambiguation. An unsupervised algorithm uses the parallel corpus to tag the English side and then project the results to the new language.

In principle the result of the two methodologies is the same: a parallel corpus aligned at word level with a shared inventory of senses (WordNet in both cases). However our work differs from Diab's in various aspects.

Firstly, to carry out our experiment we created a sample parallel corpus using translations done by professional translators. On the contrary, Diab uses Machine Translation systems to create the parallel corpora she worked with.

Secondly, our annotation transfer system crucially depends on the existence of an already annotated corpus while Diab's system automatically tags both corpora.

Thirdly, the knowledge-based word aligner we developed relies on bilingual dictionaries and works for English and Italian while Diab uses a statistical token aligner which is language independent. In fact, the experiment has been carried out on the Brown corpus automatically translated to French, German, and Spanish. However, she applied her system only to the tagging of nouns, whereas our system transfers the annotations of all parts of speech.

Diab's system does not need any prior linguistic knowledge apart from WordNet but accuracy rates are lower than ours. This can be explained by two facts: we rely on a manually annotated corpus and on knowledge intensive algorithms. The corpus resulting from our methodology is more reliable for tasks such as training of automatic word disambiguation systems.

7. Conclusions and future work

We have presented a methodology for cross-lingual semantic annotation transfer. This approach relies crucially on the performance of a word alignment algorithm which is still in an early stage of development, and can be further improved.

The results of our pilot study show that semantic annotation transfer is a promising approach for the development of semantically annotated parallel corpora.

We are planning to apply the annotation transfer methodology on a large scale to produce Italian-controlled translations for all the SemCor texts and have them word sense tagged. We are also studying how MultiSemCor can be used to enrich a multilingual lexical resource such as MultiWordNet (Pianta et al. 2002).

8. References

Ahrenberg, L., Merkel, M., Sagvall Hein, A. J. Tiedemann, 2000. Evaluation of word alignment systems. In *Proceedings of LREC 2000*, Athens, Greece.

Baker, M., 1993. Corpus linguistics and translation studies: implications and applications. In M. Baker, G. Francis, and E. Tognini-Bonelli (eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins.

Bentivogli, L., and E. Pianta, 2000. Looking for lexical gaps. In *Proceedings of the Ninth EURALEX International Congress*, Stuttgart, Germany.

Bentivogli, L., F. Pianesi, and E. Pianta, 2001a. PhiloNet: creating semantic concordances for the analysis of philosophical texts. *In Proceedings of ACH/ALLC 2001*, New York, USA.

Bentivogli, L., P. Forner, and E. Pianta, 2001b. *Manual Word Alignment Guidelines for the MultiSemCor Project*. ITC-irst Technical Report # 0112-31

Diab, M., 2000. An unsupervised method for multilingual word sense tagging using parallel corpora: A preliminary investigation. In *Proceedings of SIGLEX Workshop on Word Senses and Multi-linguality*, ACL-2000, Hong Kong.

Fellbaum, C. (ed.), 1998. *WordNet: An Electronic Lexical Database*. Cambridge (Mass.):The MIT Press.

Levin, L., D. Gates, A. Lavie, F. Pianesi, D. Wallace, T. Watanabe, M. Woszczyna, 2000. Evaluation of a Practical Interlingua for Task-Oriented Dialogues. In *Proceedings of the AMTA-SIG-IL Workshop On Interlinguas and Interlingual Approaches*, Seattle.

Mana, N. and O. Corazzari, 2002. The lexico-semantic annotation of an Italian Treebank. In *Proceedings of LREC 2002*, Las Palmas, Canary Islands, Spain.

Melamed, I. D., 2001. *Empirical Methods for Exploiting Parallel Texts*. Cambridge (Mass.):The MIT Press.

Merkel, M., 1999. *Annotation Style Guide for the PLUG Link Annotator*, Technical Report, Linkoping University.

Pianta, E., L. Bentivogli, and C. Girardi, 2002. MultiWordNet: developing an aligned multilingual database. *In Proceedings of the First Global WordNet Conference*, Mysore, India.

Senseval 2. http://www.sle.sharp.co.uk/senseval2/

Véronis, J., and P. Langlais, 2000. Evaluation of parallel text alignment systems. In *Parallel Text Processing*. Dordrecht: Kluwer Academic Publishers.

Véronis, J., (ed.), 2000. *Parallel Text Processing*. Dordrecht: Kluwer Academic Publishers.

Yzaguirre, L. de, M. Ribas, J. Vivaldi, M. T. Cabré, 2000. Some technical aspects about aligning near languages. In *Proceedings of LREC 2000*, Athens, Greece