*Natural Language Engineering* **11** (3): 247–261. © 2005 Cambridge University Press 
doi:10.1017/S1351324905003839 Printed in the United Kingdom

# *Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus*

L. BENTIVOGLI and E. PIANTA

*ITC-irst, Via Sommarive, 18 - 38050 Povo, Trento, Italy*
*e-mail*: bentivo,pianta@itc.it

(*Received May 1 2004; revised November 30 2004*)

**Abstract**

In this article we illustrate and evaluate an approach to create high quality linguistically annotated resources based on the exploitation of aligned parallel corpora. This approach is based on the assumption that if a text in one language has been annotated and its translation has not, annotations can be transferred from the source text to the target using word alignment as a bridge. The transfer approach has been tested and extensively applied for the creation of the MultiSemCor corpus, an English/Italian parallel corpus created on the basis of the English SemCor corpus. In MultiSemCor the texts are aligned at the word level and word sense annotated with a shared inventory of senses. A number of experiments have been carried out to evaluate the different steps involved in the methodology and the results suggest that the transfer approach is one promising solution to the resource bottleneck. First, it leads to the creation of a parallel corpus, which represents a crucial resource *per se*. Second, it allows for the exploitation of existing (mostly English) annotated resources to bootstrap the creation of annotated corpora in new (resource-poor) languages with greatly reduced human effort.

## 1 Introduction

In recent years, the importance of parallel corpora (i.e. texts accompanied by their translation in another language, also known as bi-texts) has become more and more evident within the Natural Language Processing (NLP) field, where these resources are used in many tasks such as multilingual lexical acquisition, machine and machine-aided translation, cross-language information retrieval, and for deriving multilingual and monolingual text processing tools.

In this paper we will focus on the exploitation of parallel corpora as a means for reducing the human effort needed for the creation of linguistically annotated corpora. Up until some years ago, linguistically annotated corpora were only produced through manual annotation, or by a manual check of automatically produced annotations. Unfortunately, manual annotation is a very difficult and time-consuming task, and constitutes a real bottleneck for the development of many data-driven approaches to NLP.

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

Two strategies have been developed to cope with the manual-annotation bottleneck. The first strategy aims at reducing as much as possible the need for annotated data. The second strategy aims at reducing the effort needed to produce manual-quality annotated data. As for the first strategy, in the machine learning community a number of learning strategies have been developed that try to get the best from small amounts of labeled data, e.g. weakly supervised learning techniques such as self- and co-training. However, even if the amount of labeled data needed by machine learning algorithms can be minimized, consistent amounts of manual-quality annotated data still have to be produced. This need is more acute for languages different from English, for which even the minimal amount of data necessary for bootstrapping machine learning algorithms is missing. Given this state of affairs, the second strategy (i.e. reducing the human effort needed to produce manual-quality labeled data) can be highly beneficial.

Parallel corpora present an opportunity for breaking the annotated resource bottleneck as they can be exploited in the creation of resources for new languages via projection of annotations available in another language. This article represents our contribution to the research in this field. We present a novel methodology to create a semantically annotated corpus by exploiting information contained in an already annotated corpus. Our approach is based on the assumption that, given a text and its translation into another language, the semantic information is mostly preserved during the translation process. Therefore, if the texts in one language have been semantically annotated and their translations have not, annotations can be transferred from the source language to the target using word alignment as a bridge.

The annotation transfer methodology exploits a situation where there is a large imbalance between the resources available for English and those available for other languages. Not only are manually annotated (or manually controlled) English corpora available, but given the availability of high quality linguistic processors working on English, automatically annotated English corpora can also be exploited with an acceptable degree of confidence.

The methodology has been applied for the creation of the MultiSemCor corpus, an English/Italian parallel corpus created on the basis of the English SemCor corpus (Landes, Leacock and Tengi 1998). In MultiSemCor the texts are aligned at the word level and word sense annotations are transferred from English to Italian. The final result of the project is an Italian corpus annotated with part of speech (PoS), lemma and word sense, but also an aligned parallel corpus lexically annotated with a shared inventory of word senses.

The MultiSemCor project started in 2002 with a pilot study on a pool of six SemCor texts. Given the promising results of that preliminary work (Bentivogli and Pianta 2002), the project has continued and currently, the methodology has been applied to 116 texts. This article reports on (i) the principles of our transfer methodology and the insights gained from its application; (ii) the thorough evaluation of the different steps involved in the transfer procedure.

In section 2 we summarize the related work which constitutes the background of our research. In section 3 we describe how the annotation transfer methodology has been applied in the creation of the MultiSemCor corpus. In section 4 we discuss

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

some problematic issues related to the annotation transfer methodology, which will be extensively tested and evaluated in section 5. In section 6 we report on the current composition of the MultiSemCor corpus, its main usages within the NLP field and our thoughts on future work.

## 2 Background

The idea of obtaining linguistic information about a text in one language by exploiting parallel or comparable texts in another language has been explored in the field of Word Sense Disambiguation (WSD) since the early 1990s. The key observation is that a polysemous word in the source language is likely to be translated into different words in the target language. This fact can be exploited in various ways.

In Brown, Della Pietra, Della Pietra and Mercer (1991) and Gale, Church and Yarowsky (1992) bilingual word-aligned parallel corpora are used to disambiguate words that are polysemous from a translation point of view. Dagan, Itai and Schwall (1991) and Dagan and Itai (1994) on the other hand use information from a bilingual dictionary and a monolingual target language corpus to handle the problem of target word selection in machine translation.

All of these early studies were committed to solving the problem of WSD in the context of machine translation. Thus they were only interested in those meaning distinctions that are reflected by the selection of different translation equivalents. Interestingly enough, some years later Resnik and Yarowsky (1997) proposed to restrict the sense distinctions used for NLP purposes to those distinctions that are typically lexicalized cross-linguistically. Along the same line, Ide, Erjavec and Tufis (2002) present a method to identify word meanings starting from a multilingual corpus composed of George Orwell's *1984* novel and aligned translations in six languages. The experiments carried out on 33 English nouns show that the sense distinctions obtained from this fully automatic approach are at least as reliable as those made by human annotators. A by-product of applying this method is that once a word in one language is word-sense tagged, the translation equivalents in the parallel texts are also automatically annotated.

Cross-language tagging is the goal of the work by Diab (2000) and Diab and Resnik (2002), who present a method for word sense tagging of both the source and target texts of parallel bilingual corpora with the WordNet sense inventory (Fellbaum 1998). The experiment has been carried out on the Brown corpus which has been automatically translated into French, German, and Spanish using commercially available machine translation packages. The use of commercially available machine-translation systems to obtain the parallel corpus to be annotated is the major distinctive feature of this approach. Note that contrary to the first studies, machine translation is an intermediate step to achieve WSD, and not vice versa. At the opposite, Magnini and Strapparava (2000) exploit parallel corpora made from existing manual translations to improve on the WSD task.

Parallel to the studies regarding the projection of semantic information, more recently the NLP community has also explored the possibility of exploiting translation

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

to project more syntax-oriented annotations, a kind of information traditionally viewed as language-specific. Yarowsky, Ngai and Wicentowski (2001) present a method for: (i) automatic annotation of English texts with information about PoS, NP chunking, named entities, and lemmatization; (ii) cross-language projection of annotations onto French, Chinese, Czech and Spanish translations; and (iii) induction of noise-robust taggers for target languages from the projected annotations.

A further step is made by Hwa, Resnik and Weinberg (2002), who address the task of acquiring a Chinese dependency treebank by bootstrapping from existing linguistic resources for English. The same methodology has been used by Cabezas, Dorr and Resnik (2001) (2001) to create a Spanish dependency treebank from English data. Finally, in Riloff, Schafer and Yarowsky (2002) a method is presented for rapidly creating Information Extraction systems for new languages by exploiting existing systems via cross-language projection.

The results of all the above mentioned studies show how previous major investments in English annotated corpora and tool development can be effectively leveraged across languages, allowing for the development of accurate resources and tools in other languages with reduced human effort.

## 3 The MultiSemCor project

The annotation transfer methodology proposed in this article has been tested and extensively applied within the MultiSemCor project. The project aims at building an English/Italian parallel corpus, aligned at the word level and annotated with PoS, lemma and word senses. The MultiSemCor corpus has been created on the basis of SemCor (Landes *et al.* 1998), an English corpus which is a subset of the Brown corpus and includes 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[1]. More in detail, the SemCor corpus is composed of 352 texts. In 186 texts, all open class words (nouns, verbs, adjectives, and adverbs) have been annotated with PoS, lemma and sense, whereas in the remaining 166 texts only verbs have been annotated with lemma and word sense. The "all-words" component of SemCor includes 359,732 tokens among which 192,639 are semantically annotated, whereas the "only-verbs" component includes 316,814 tokens among which 41,497 verb occurrences are semantically annotated. We are currently using the original release of SemCor (annotated with reference to WordNet 1.6 version), and working on the "all-words" component.

The procedure followed for creating MultiSemCor consists in: (i) obtaining Italian translations of the SemCor texts; (ii) automatically aligning Italian and English texts at the sentence and word levels; (iii) automatically transferring the word sense annotations from English to the aligned Italian words. The final result of the project

[1] WordNet (Fellbaum 1998) is an English lexical database, developed at Princeton University, in which nouns, verbs, adjectives, and adverbs are organized into sets of synonyms (synsets) and linked to each other by means of various lexical and semantic relationships. In recent years, within the NLP community WordNet has become the reference lexicon for almost all tasks involving word sense disambiguation (see, for instance, the Senseval competition).

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

is 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. More specifically, the sense inventory is taken from MultiWordNet (Pianta, Bentivogli and Girardi 2002), a multilingual lexical database in which the Italian component is strictly aligned with the English Princeton Wordnet.

### *3.1 Obtaining Italian translations*

The first problem to be solved in the creation of MultiSemCor was that Italian translations of SemCor texts did not exist. Our solution was to ask professional translators to translate the texts. Given the high cost of building semantically annotated corpora, we think that, when a corpus has already been annotated in one language, translating the corpus is a reasonable option, well worth taking into consideration. In fact, manually translating the annotated corpus and automatically transferring the annotations may be preferable to hand-labeling a new corpus from scratch. Not only are translators more easily available than linguistic annotators, but translations may be a more flexible and durable kind of annotation. A translation can be exploited in many different ways, whereas a manual annotation can turn out to be out of date or unusable for some new tasks; see Pianta and Bentivogli (2003) for a more detailed discussion about this topic. Moreover, 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. With respect to a situation in which the translation of a corpus is already available, a corpus translated on purpose presents the advantage that translations can be *controlled*, i.e. carried out following criteria aiming at maximizing alignment and annotation transfer.

In fact, MultiSemCor translators were asked to translate the English texts into Italian following various criteria aiming at enhancing the correspondence between source and target texts, namely:

- maintain the sentence segmentation of the original English texts;
- mark Italian multiword named entities with an underscore, following SemCor conventions (e.g. "Unione_Europea" as a translation of "European_Union");
- prefer the same dictionary used by the automatic word aligner (see the following section);
- choose the most "synonymous" translation equivalents and, more specifically, prefer those belonging to the same PoS.

The first three criteria are meant to facilitate the work of the word aligner, whereas the fourth is meant to maximize the correctness of the transfer of the information from English to Italian. It should be stressed that translators were also told that the controlled translation criteria should never be followed to the expense of good Italian prose.

### *3.2 Aligning the corpus at the word level with KNOWA*

Once the Italian translations of the SemCor texts were obtained, the second step in the creation of the MultiSemCor corpus was to align the texts at the word

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

level. For this, we resorted to 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. KNOWA also exploits a morphological analyzer and a multiword recognizer, for both Italian and English. The alignment algorithm expects the input bi-text to be sentence-aligned, but does not require any corpus for training. For a detailed description of this tool, see Pianta and Bentivogli (2004).

Some characteristics of the MultiSemCor scenario make the alignment task easier for KNOWA. First, multiword recognition is made easier by the fact that in SemCor all multiwords included in WordNet are explicitly marked. This implies that KNOWA does not need to recognize English multiwords, although it still needs to recognize them in Italian. Secondly, within MultiSemCor word alignment is done with the final aim of transferring lexical annotations from English to Italian. Since only content words have word sense annotations in SemCor, it is more important that KNOWA behaves correctly on content words, which are easier to align than functional words. Section 5 illustrates the evaluation of KNOWA in the MultiSemCor task.

### *3.3 Transferring annotations from English to Italian*

Once word alignment has been performed, the annotation transfer is a simple task: for each English-Italian word pair, (i) copy the sense annotation (if any) from SemCor to the Italian text; and, (ii) add lemma and PoS as selected during the alignment process.

The transfer of annotations from English to Italian is based on the assumption that translation keeps word meaning across languages. We will see in section 5 the extent to which this assumption holds.

## 4 Problematic issues

The MultiSemCor methodology raises a number of theoretical and practical issues. A first theoretical issue concerns the nature of translational language. Some studies show that, even in the case of very accurate translations, the language of translated texts has a number of peculiarities that set it apart from the language of original, non-translated texts (Baker 1993). As a consequence, MultiSemCor includes Italian texts which are not as fully representative of the general use of language as the original SemCor. 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. Thus, we can still maintain that MultiSemCor includes current, largely representative, annotated Italian texts, which are as useful as annotated original texts for tasks such as semantic concordancing and training of word sense disambiguation systems.

The second issue concerns the legitimacy of transferring word senses from one language to another. To what extent are the lexica of different languages comparable? A study on the comparability of English and Italian lexica has

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

shown that the vast majority of English words have an Italian cross-language synonym (Bentivogli and Pianta 2000). According to this study, only 7.8% of the English words correspond to lexical gaps in Italian. This figure suggests that transferring word meanings from English to Italian is reasonable, but also that there will be a relatively small number of cases in which the transfer will not be possible.

Besides these theoretical issues, the most crucial practical issue is represented by the quality of the Italian annotation resulting from the application of the methodology. As opposed to automatic word sense disambiguation tasks, the MultiSemCor project specifically aims at producing manual-quality annotated data. Therefore, there is a potential risk of degrading the quality of the Italian annotations through the various steps of the annotation transfer procedure. A number of factors must be taken into account.

- SemCor quality: annotation errors can be found in the original English texts.
- Word Alignment quality: the word aligner may align words incorrectly.
- Transfer quality: the transfer of the semantic annotations may not be applicable to certain translation pairs.

In the next Section we will describe and evaluate these quality issues in order to assess the extent to which they affect the cross-language annotation transfer methodology.

## 5 Evaluation of the annotation transfer methodology

To test and analyze the impact of the quality issues outlined in the previous section, an evaluation gold standard has been created.

The MultiSemCor evaluation gold standard is composed of four unseen English texts (`br-f43, br-g11, br-l10, br-j53`) taken randomly from the SemCor corpus. For each English text both a *free* and a *controlled* translation were made. Even though the MultiSemCor project only relies on *controlled* translations, we decided to create a gold standard also for the *free* translations in order to verify if the transfer methodology can be applied also to already existing parallel corpora, i.e. on translations carried out without following the MultiSemCor translation criteria. The resulting gold standard includes 8,877 English tokens, and 9,224 Italian tokens in controlled translations, i.e. Italian texts contain 3.9% more tokens. This difference is partly explained by grammatical characteristics specific to the Italian language, for instance in the usage of articles and clitics (cfr. the English sentence "as cells coalesced" translated into Italian as "quando *le* cellule *si* unirono"). Also, multiwords are represented as one token in the English section of MultiSemCor (e.g. *nucleic_acid*), whereas in the Italian section they are represented as more than one token (e.g. *acido nucleico*), with the exception of named entities (see section 3.1).

As a first step, necessary to evaluate the performance of the word alignment system, the eight pairs of texts in the gold standard were manually aligned, following guidelines that we defined, based on those of similar word alignment projects (Melamed 2001). Annotators were asked to align different kinds of units (simple

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

words, segments of more than one word, parts of words) and to mark different kinds of semantic correspondence between the aligned units, i.e. full correspondence (synonymic), non-synonymic correspondence, changes in lexical category and phrasal correspondence.

Inter-annotator agreement was measured with the Dice coefficient (Véronis and Langlais 2000) and can be considered satisfactory at 87% for *free* translations and 92% for *controlled* translations. As expected, controlled translations produced a better agreement rate between annotators.

The second step, necessary to evaluate the quality of the annotations automatically transferred to Italian, was to manually annotate the Italian texts. To this purpose, the four *controlled* Italian translations were manually semantically annotated taking into account the annotations of the English words. Each time an English annotation was appropriate for the Italian corresponding word, the annotator used it also for Italian. Otherwise, the annotator looked in WordNet for an alternative suitable annotation. Moreover, when the English annotations were not suitable for annotating the Italian words, the annotator explicitly distinguished between wrong English annotations and English annotations that could not be transferred to the Italian translation equivalents. The inter-annotator agreement for the semantic annotation task, calculated following the Dice coefficient method, amounts to 81.9%. By comparison,the annotator agreement calculated for the original SemCor annotation task was 78.6% (Fellbaum *et al.* 1998).

The second step of the annotation produced 4,313 Italian lexical annotations, compared to the original 4,101 English annotations (5.2% more annotations in Italian). The difference is explained by the fact that modal and auxiliary verbs (to have, to be, can, may, to have to, etc.) and partitives (some, any) where systematically left unannotated in the English text whereas they have been annotated for Italian.

### *5.1 SemCor quality*

The first factor which can affect the transfer methodology is the quality of the SemCor manual annotations. If we manage to correctly align an English and an Italian word, but the annotation of the original word is wrong, so will be the annotation automatically transferred to the Italian translation equivalent. Even if the SemCor corpus was manually annotated, a non-negligible percentage of the annotations turns out to be wrong (see Fellbaum, Grabowski and Landes (1998) for SemCor taggers' confidence ratings). As an example, the word *pocket* in the sentence "He put his hands on his pockets" was incorrectly tagged with the WordNet synset {`pouch, sac, sack, pocket -- an enclosed space`} instead of the correct one {`pocket -- a small pouch in a garment for carrying small articles`}.

As mentioned above, the English annotations which were considered wrong by our annotators were explicitly marked in the gold standard. They amount to 117, corresponding to the 2.8% of the total English annotations. Note that wrongly annotated English words only cause annotation errors in the Italian text if they are aligned.

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

Table 1. *Evaluation of KNOWA on full text*

| Translation | Precision (%) | Recall (%) | Coverage (%) |
|---|---|---|---|
| Free | 85.9 | 61.8 | 70.0 |
| Controlled | 89.2 | 69.4 | 76.1 |

Table 2. *Evaluation of KNOWA on sense tagged words only*

| Translation | PoS | Precision (%) | Recall (%) | Coverage (%) |
|---|---|---|---|---|
| Free | Nouns | 95.5 | 82.7 | 86.6 |
| | Verbs | 87.6 | 71.3 | 81.5 |
| | Adjectives | 95.9 | 66.5 | 69.3 |
| | Adverbs | 89.4 | 42.8 | 47.9 |
| | **Total** | **92.8** | **70.3** | **75.8** |
| Controlled | Nouns | 96.9 | 84.5 | 87.2 |
| | Verbs | 91.4 | 77.4 | 84.7 |
| | Adjectives | 96.0 | 72.3 | 75.3 |
| | Adverbs | 91.0 | 54.4 | 59.8 |
| | **Total** | **94.7** | **76.2** | **80.5** |

### *5.2 Word alignment quality*

As we have seen in section 3.2, the feasibility of the entire MultiSemCor project heavily depends on the availability of an English/Italian word aligner with very good performance in terms of recall and, more importantly, precision. In fact, if the aligner wrongly aligns two words then the word sense transferred from the source word will not be suitable for the target word.

The performance of KNOWA on MultiSemCor was compared to the gold standard alignments, and measured in terms of *alignment precision*, *recall* and *coverage*, as defined in Véronis and Langlais (2000): precision measures the proportion of test alignments that are also in the gold-standard; recall measures the proportion of gold-standard alignments that are present in the test; and coverage measures the proportion of English words for which the test contains alignments. See Och and Ney (2003) and Arhenberg, Merkel, Sagvall and Tiedemann (2000) for alternative evaluation metrics.

The evaluation results of KNOWA on the MultiSemCor task are shown in Tables 1 and 2. We report on both full text alignment and alignment of sense-tagged words only, for both *free* and *controlled* translations. For the alignment of sense-tagged words, results are broken down by PoS.

These results, which compare well with those reported in the literature (Arhenberg *et al.* 2000; Véronis 2000; Och and Ney 2003), show that, as expected, controlled translations allow for a better alignment. Moreover, we can see that the performance of the word aligner improves when function words are ignored.

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

### 5.3 Transfer quality

Even when both the original English annotations and the word alignment are correct, a number of cases remain for which the annotation transfer is not applicable. An annotation is not transferable from the source to the target language when the translation equivalent does not preserve the lexical meaning of the source word. In such cases, if the word alignment puts the two expressions in correspondence, the sense annotation transfer yields a wrong annotation in the target word.

The first main cause of incorrect transfer is represented by translation equivalents that are not cross-language synonyms of the source language words. For instance, in a sentence of the corpus, the English noun *meaning* was translated in Italian as *motivo* (reason, grounds) which is suitable in that specific context but is not a synonymic translation of the English word. A specific case of non-synonymous translation occurs when a translation equivalent does not belong to the same lexical category as the source word. For example, the English verb *to coexist* in the sentence "the possibility for man to coexist with animals" has been translated with the Italian noun *coesistenza* (coexistence) in "le possibilità di coesistenza tra gli uomini e gli animali". Sometimes, non-synonymous translations are due to errors in the Italian translation, as in *pull* translated as *spingere* (push). A more difficult case of non-synonymous translation is phrasal correspondence, occurring when a target phrase has globally the same meaning as the corresponding source phrase, but the single words of the phrase are not cross-language synonyms of their corresponding source words. For example, the expression *a dreamer sees* has been translated as *una persona sogna* (a person dreams). The Italian translation maintains the synonymy at the phrase level but the single component words do not.

The second, and somewhat controversial, potential cause of incorrect transfer occurs when the translation equivalent is indeed a cross-language synonym of the source expression, but is not a lexical unit. This usually happens with lexical gaps, i.e. when a language expresses a concept with a lexical unit whereas the other language expresses the same concept with a free combination of words, as for example the English word *successfully* which can only be translated with the Italian free combination of words *con successo* (with success). However, it can also be the result of a choice made by the translator who, for example, may decide to translate *empirically* as *in modo empirico* (in an empirical manner) instead of *empiricamente*. In these cases the problem arises because, in principle, if the target expression is not a lexical unit it cannot be annotated with one sense as a whole. On the contrary, each component of the free combination of words should be annotated with its respective sense.

One of the tasks of the annotators while building the MultiSemCor gold standard, was to mark translations pairs in which the English annotation could not be transferred to the Italian translation equivalent, for any of the above mentioned reasons. Non-transferable annotations amount to 692, i.e. 16.9% of the English annotations. Of these, 591 (85.4%) are due to translation equivalents which are lexical units but are not cross-language synonyms, while the remaining 101 (14.6%) are due to translation equivalents that are not lexical units.

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

Table 3. *Evaluation of the Italian annotation*

| | Precision (%) | Recall (%) | Coverage (%) |
|---|---|---|---|
| Italian *controlled* texts | 87.9 | 67.2 | 76.4 |

Table 4. *Source of incorrect transfer*

| | # | Error % | Annotation % |
|---|---|---|---|
| English annotation errors | 109 | 27.2 | 3.3 |
| Word alignment errors | 95 | 23.8 | 2.9 |
| Non-transferable annotations | 196 | 49.0 | 5.9 |
| **Total Incorrect transfer** | **400** | **100** | **12.1** |

### *5.4 Final quality of the Italian annotation*

In order to evaluate the final quality of the Italian annotation resulting from the application of the transfer methodology, the automatic procedures for word alignment and annotation transfer were run on the texts and evaluated against the gold standard.

Out of the 4,101 SemCor English annotations, the automatic procedure was able to transfer 3,297. Among these, 2,897 are correct and 400 are incorrect for the Italian words. Table 3 summarizes the results in terms of Precision, Recall and Coverage with respect to the Italian words to be annotated (4,313). The precision of the Italian annotation amounts to 87.9%, which can be considered acceptable.

Precision is crucial for assessing the quality of Italian annotation, and therefore we analyzed the 12.1% annotation errors, and classified them according to the different factors affecting the transfer methodology. Table 4 reports on the source of incorrect transfers.

**SemCor quality.** Comparing the number of annotation errors in the English source, as marked up during the creation of the gold standard (117, 2.8% of the original English annotations), with the number of errors in the Italian annotation due to SemCor wrong annotations (109, 3.3% of the transferred annotations), we can see that almost all of the source errors have been transferred, contributing in a consistent way to the overall Italian annotation error rate.

**Word alignment quality.** As already seen in Section 5.2, the precision of the word alignment system is quite high. As a matter of fact, the number of errors in the Italian annotation due to wrong alignments made by KNOWA (2.9%) does not affect the overall Italian annotation in an important way. Note that we included in this class only word alignment errors for word pairs that were correctly annotated for English, and whose annotation was transferable to the Italian translation equivalent. This explains why the reported percentage is lower than the word error rate reported for the word alignment task (5.3%).

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

**Transfer quality.** The last source of annotation errors is represented by words which have been aligned (correctly or not) but whose word sense annotation cannot be transferred. This may happen with (i) translation equivalents which are lexical units but are not cross-language synonyms at lexical level, and (ii) translation equivalents which are cross-language synonyms but are not lexical units. In practice, given the difficulty in deciding what is and what is not a lexical unit, we decided to accept the transfer of a word sense from an English lexical unit to an Italian free combination of words (see for instance *occhiali da sole* annotated with the sense of *sunglasses*). Therefore, only the lack of synonymy at lexical level has been considered an annotation error. The obtained results are encouraging, as only 196 of the 591 non-synonymous translations marked in the gold standard have been aligned by the word alignment system (33.2%). This is explained by the fact that KNOWA alignments rely on bilingual dictionaries where non-synonymous translations are quite rare.

A final remark about the error analysis concerns the distinction between errors due to wrong English annotations and to non-transferable word senses. As we are not native English speakers, it is sometimes very difficult for us to distinguish between the two. Thus, we preferred to be conservative in marking English annotations as errors, limiting these to very clear cases. As a result, the percentage of errors in the original English corpus may have been underestimated in our evaluation, and the percentage of non-transferable word senses overestimated.

The classification of errors presented in Table 4 shows that there is little room to improve on *precision*. If we analyze the three possible causes of Italian annotation errors, we see that the automatic transfer methodology cannot be improved on SemCor errors or non-synonymous translation equivalents. The only errors that we can improve on in principle are those caused by KNOWA. However, these errors make up only 2.9% of all annotations. On the other hand, *recall* could be improved through better alignment coverage, which currently stands at 80.5%. By breaking down the coverage by PoS, as shown in Table 2, one can see that coverage is particularly low for adjectives and adverbs. This seems due, at least in part, to the fact that KNOWA is not effective in recognizing multiwords belonging to these two categories. For this reason we plan to improve the multiword recognition component of KNOWA, by focusing on the Italian translation of English adverbial and adjectival multiwords contained in SemCor. Another strategy to improve the final quality of the MultiSemCor annotations, would be to manually correct annotation errors in the original SemCor. Such a revision exercise would be even more profitable if the MultiSemCor model was to be extended to other languages beyond Italian.

Summing up, the cross-language annotation transfer methodology produces an Italian corpus which is tagged with a final precision of 87.9%. After the application of the methodology, 23.6% of Italian words still need to be annotated (see the annotation coverage of 76.4% in Table 3). We think that, given the precision and coverage rates obtained from the evaluation, the corpus as it results from the automatic procedure can be profitably used. Even in a case where a manual revision is envisaged, we believe that hand-checking the automatically tagged corpus and

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

Table 5. *Italian annotation quality in free and controlled translations*

| | Precision (%) | Recall (%) | Coverage (%) |
|---|---|---|---|
| *Free* translation | 84.8 | 63.1 | 74.4 |
| *Controlled* translation | 87.7 | 70.8 | 80.7 |

manually annotating the remaining 23.6% would be cost-effective, compared to annotating the corpus from scratch.

### *5.5 Annotation quality in free Italian translations*

As mentioned above, we are interested not only in evaluating the annotation transfer methodology when applied to the MultiSemCor scenario, but also in understanding its effectiveness for already existing parallel corpora. To this purpose, the MultiSemCor gold standard has been extended by semantically annotating also the *free* translation of text `br-g11`. The results of the comparison of the Italian annotation quality for the *controlled* and *free* translations of `br-g11` are reported in Table 5. As expected, the quality of the annotation of the *controlled* translation is higher than that of the *free* translation. The gap between the two ranges from 2.9% for precision to 7.7% for recall.

## 6 Conclusion and future directions

We presented an approach to the creation of high quality semantically annotated resources based on the exploitation of aligned parallel corpora. This approach has been extensively applied in the creation of the MultiSemCor corpus. The various steps have been evaluated, yielding results that we rate as satisfactory.

At present MultiSemCor is composed of 116 English texts taken from the "all-words" component of SemCor (see section 3) along with their corresponding 116 Italian translations. The total amount of tokens is 258,499 for English and 267,607 for Italian. The texts are fully aligned at the word level and content words are annotated with PoS, lemma, and word sense. As regards English, we have 119,802 words semantically annotated (from SemCor). As for Italian, 92,820 words are annotated with word senses that have been automatically transferred from English. The first release of the corpus is distributed free for research purposes at the MultiSemCor web site `http://multisemcor.itc.it`.

MultiSemCor can be a useful resource for a variety of tasks, both as a monolingual semantically annotated corpus and as a parallel aligned corpus. As an example, we are already using it to automatically enrich the Italian component of MultiWordNet, the reference lexicon of MultiSemCor. As a matter of fact, out of the 92,820 Italian words automatically sense-tagged, 8,923 are not yet present in MultiWordNet and will be added to it. Moreover, the Italian component of MultiSemCor is being used as a gold standard for the evaluation of WSD systems for Italian (Gliozzo, Ranieri and Strapparava 2005). Besides NLP applications, MultiSemCor is also suitable for

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

consultation by humans through a Web interface (Ranieri, Pianta and Bentivogli 2004), which is available at the MultiSemCor web site.

As regards future research directions, we are planning a collaboration with other research institutes to extend the MultiSemCor methodology to other languages, e.g. Spanish and Romanian, for which a WordNet exists and can be aligned with MultiWordNet. Moreover, since the Brown Corpus, used to create SemCor, has been syntactically annotated within the English Penn Treebank (Marcus, Santorini and Marcinkiewicz 1993), the syntactic annotations of the SemCor texts are also available. We are exploring the feasibility of transferring syntactic annotations from the English to the Italian texts of MultiSemCor.

Another interesting research direction towards the full exploitation of parallel corpora in the creation of annotated resources is the projection of other types of linguistic annotation, for instance anaphoric reference and discourse-level information such as rhetorical relations.

## Acknowledgements

Sincere thanks to Pamela Forner, Massimiliano Bampi and Marcello Ranieri who contributed to the development of the MultiSemCor gold standard, to Christian Girardi for setting up and maintaining the software infrastructure, to Alberto Lavelli for proofreading and giving moral support, and to the MEANING and WEBFAQ projects for providing financial support.

## References

Ahrenberg, L., Merkel, M., Sagvall, H. and Tiedemann, A. J. (2000) Evaluation of word alignment systems. *Proceedings of LREC 2000*, Athens, Greece.

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

Bentivogli, L. and Pianta, E. (2000) Looking for lexical gaps. *Proceedings of the 9th EURALEX International Congress*, Stuttgart, Germany.

Bentivogli, L. and Pianta, E. (2002) Opportunistic semantic tagging. *Proceedings of LREC-2002*, Las Palmas, Canary Islands, Spain.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. and Mercer, R. L. (1991) Word-sense disambiguation using statistical methods. *Proceedings of ACL'91*, Berkeley, CA.

Cabezas, C., Dorr, B. and Resnik, P. (2001) Spanish language processing at University of Maryland: Building infrastructure for multilingual applications. *Proceedings of the 2nd International Workshop on Spanish Language Processing and Language Technologies*, Jaen, Spain.

Dagan, I. and Itai, A. (1994) Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics* **20**(4): 563–596.

Dagan, I., Itai, A. and Schwall, U. (1991) Two languages are more informative than one. *Proceedings of ACL'91*, Berkeley, CA, USA.

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

Diab, M. and Resnik, P. (2002) An unsupervised method for word sense tagging using parallel corpora. *Proceedings of ACL 2002*, Philadelphia, USA.

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press

Fellbaum, C. (ed.) (1998) *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.

Fellbaum, C., Grabowski, J. and Landes, S. (1998). Performance and confidence in a semantic annotation task. In Fellbaum, C. (ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Gale, W. A., Church, K. W. and Yarowsky, D. (1992) Using bilingual materials to develop word sense disambiguation methods. *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, Canada.

Gliozzo, R., Ranieri, M. and Strapparava, C. (2005) Crossing parallel corpora and multilingual lexical databases for WSD. *Proceedings of CICLing-2005*, Mexico City, Mexico.

Hwa, R., Resnik, P. and Weinberg, A. (2002) Breaking the resource bottleneck for multilingual parsing. *Proceedings of the LREC 2002 Workshop on "Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data"*, Las Palmas, Canary Islands, Spain.

Ide, N., Erjavec, T. and Tufis, D. (2002) Sense discrimination with parallel corpora. *Proceedings of the ACL 2002 Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"*, Philadelphia, USA.

Landes S., Leacock, C. and Tengi, R. I. (1998) Building semantic concordances. In: Fellbaum, C. (ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.
Resource Acquisition. *Proceedings of LREC-2002 Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*. Las Palmas, Canary Islands, Spain.

Magnini, B. and Strapparava, C. (2000) Experiments in word domain disambiguation for parallel texts. *Proceedings of the ACL 2000 SIGLEX Workshop on "Word Senses and Multi-linguality"*, Hong Kong.

Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A. (1993) Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* **19**(2): 313–330.

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

Och, F. J. and Ney, H. (2003) A systematic comparison of various statistical alignment models. *Computational Linguistics* **29**(1): 19–53.

Pianta, E., Bentivogli, L. and Girardi, C. (2002) MultiWordNet: developing an aligned multilingual database. *Proceedings of the 1st Global WordNet Conference*, Mysore, India.

Pianta, E. and Bentivogli, L. (2003) Translation as annotation. *Proceedings of the AI*IA 2003 Workshop on "Topics and Perspectives of Natural Language Processing in Italy"*, Pisa, Italy.

Pianta, E. and Bentivogli, L. (2004) Knowledge intensive word alignment with KNOWA. *Proceedings of Coling 2004*, Geneva, Switzerland.

Ranieri, M., Pianta, E. and Bentivogli, L. (2004). Browsing multilingual information with the MultiSemCor web interface. *Proceedings of the LREC-2004 Workshop on "The Amazing Utility of Parallel and Comparable Corpora"*, Lisbon, Portugal.

Resnik, P. and Yarowsky, D. (1997) A perspective on word sense disambiguation methods and their evaluation. *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, DC.

Resnik, P. and Smith, N. A. (2003) The Web as a parallel corpus. *Computational Linguistics* **29**(3): 349–380.

Riloff, E., Schafer, C. and Yarowsky, D. (2002) Inducing information extraction systems for new languages via cross-language projection. *Proceedings of Coling 2002*, Taipei, Taiwan.

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

Véronis, J. (ed.) (2000) *Parallel text processing*. Dordrecht: Kluwer Academic.

Yarowsky, D., Ngai, G. and Wicentowski, R. (2001) Inducing multilingual text analysis tools via robust projection across aligned corpora. *Proceedings of HLT 2001*, San Diego, CA.

https://doi.org/10.1017/S1351324905003839 Published online by Cambridge University Press