Knowledge Intensive Word Alignment with KNOWA

Emanuele PIANTA and Luisa BENTIVOGLI
ITC-irst
Via Sommarie, 18
38050 Povo - Trento
Italy
{pianta,bentivo}@itc.it

Abstract

In this paper we present KNOWA, an English/Italian word aligner, developed at ITC-irst, which relies mostly on information contained in bilingual dictionaries. The performances of KNOWA are compared with those of GIZA++, a state of the art statistics-based alignment algorithm. The two algorithms are evaluated on the EuroCor and MultiSemCor tasks, that is on two English/Italian publicly available parallel corpora. The results of the evaluation show that, given the nature and the size of the available English-Italian parallel corpora, a language-resource-based word aligner such as KNOWA can outperform a fully statistics-based algorithm such as GIZA++.

1 Introduction

Aligning a text and its translation (also known as bitext) at the word level is a basic Natural Language Processing task that has found various applications in recent years. Word level alignments can be used to build bilingual concordances for human browsing, to feed machine learning-based translation algorithms, or as a basis for sense disambiguation algorithms or for automatic projection of linguistic annotations from one language to another.

A number of word alignment algorithms have been presented in the literature, see for instance (Véronis, 2000) and (Melamed, 2001). Shared evaluation procedures have been established, although there are still open issues on some evaluation details (Ahrenberg et al. 2000).

Most of the known alignment algorithms are statistics-based and do not exploit external linguistic resources, or use them to a very limited extent. The main attractive of such algorithms is that they are language independent, and only require a parallel corpus of reasonable size to be trained.

However, word alignment can be used for different purposes and in different application scenarios; different kinds of alignment strategies produce different kinds of results (for instance in terms of precision/recall) which can be more or less suitable to the goal to be achieved. The requirement of having a parallel corpus of adequate size available for training the statistics-based algorithms may be difficult to meet, given that parallel corpora are a precious but often rare resource. For the most common languages, such as English, French, German, Chinese, etc., reference parallel corpora of adequate size are available, and indeed statistics-based algorithms are evaluated on such reference corpora. Unfortunately, if one needs to replicate in a different corpus the results obtained for the reference corpora, finding a parallel corpus of adequate size can be difficult even for the most common languages. Consider that one of the most appealing features of statistics-based algorithms is their ability to induce alignment models for bitexts belonging to very specific domains, an ability which seems to be out of reach for algorithms based on generic linguistic resources. However, for the statistics-based algorithms to achieve their objective, a parallel corpus for the specific domain needs to be available, a requirement that in some cases cannot be met easily.

For these reasons, we claim that in some cases algorithms based on external, linguistics resources, if available, can be a useful alternative to statistics-based algorithms. In the rest of this paper we will compare the results obtained by a *statistics-based* and a *linguistic resource-based* algorithm when applied to the EuroCor and MultiSemCor English/Italian corpora.

The statistics-based algorithm to be evaluated is described in (Och and Ney, 2003). For its evaluation we used an implementation by the authors themselves, called GIZA++, which is freely available to the scientific community (Och, 2003). The second algorithm to be evaluated is crucially based on a bilingual dictionary and a morphological analyzer. It is called KNOWA (*KN*owledge intensive *W*ord *A*ligner) and has been developed at ITC-irst by the authors of this paper. The results of the comparative evaluation show that, given specific application goals, and given the availability of Italian/English resources, KNOWA obtains results that are comparable or better than the results obtained with GIZA++.

Section 2 describes the basic KNOWA algorithm. Sections 3 and 4 illustrate two enhanced versions of the KNOWA algorithm. Section 5 reports an experiment in which both KNOWA and GIZA++ are first applied to the alignment of a reference parallel corpus, EuroCor, and then to the MultiSemCor corpus. Section 6 adds some conclusive remarks.

2 KNOWA – the basic algorithm

KNOWA is an English/Italian word aligner, 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. It does not require any corpus for training. However the input bitext must be sentence-aligned.

For each sentence pair, KNOWA produces word alignments 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 words:

Given an English word w^E *and an Italian word* w^I, w^I *is the* potential correspondent *of* w^E *if one of the candidate lemmas of* w^I *is the* translation equivalent *of one of the candidate lemmas of* w^E, *according to a bilingual dictionary.*

The potential correspondence relation holds between words, but is relative to a lemma pair. For instance we say that the words *dreams* and *sogna* are potential correspondents relative to the lemma pair <dream/verb, sognare/verb>. Two words can be potential correspondents relative to more than one lemma pair. For instance the words *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 in real texts, tokens that are potential correspondents may not in fact be translations of each other. Take for instance the following translation pair: “ll cane e il gatto”, “the dog and the cat”. The first occurrence of the Italian article “il” is a potential correspondent of both occurrences of the word “the” in the English sentence, but is the translation of only the first one.

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

For each English word w^E in a certain position p:

1. Get the most probable candidate lemma of w^E.
2. Get the Italian word w^I in the same position p.
3. Check if there is a candidate lemma of w^I which is a potential correspondent of w^E relative to the current English candidate lemma, on the basis of a bilingual lexicon.
4. If yes, align w^E and w^I and record their lemmas.
5. Otherwise consider the next probable candidate lemma of w^E 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 $p \pm Delta$, where p is the position of the English word and *Delta* can vary from 1 to a *MaxDelta* value. The value of *MaxDelta* is adjustable, but a number of experiments have shown that the best results are obtained when *MaxDelta*=14. Note that if the alignment is not found within the Italian word window, the English word is left unaligned. In Table 1 the box in the Italian column shows the maximal text window in which the potential correspondent of *dream* is searched (*MaxDelta*=5).

The search starts from *15-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 et al., 2000) for a similar approach.

Note that given the way in which the alignment procedure works, finding an alignment 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, due to the constraints added by the potential correspondence relation.

…	…
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 1: An example of a maximal text window

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 in a full text alignment task, that is trying to align also functional words. Articles and determiners can occur repeatedly in any sentence, and almost any Italian preposition can be the translation of any English preposition; this makes the task of aligning determiners and preposition on the basis of the potential correspondence relation and the absolute position in the sentence hard. 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. In the following section we describe an extension of the basic KNOWA algorithm that tries to cope with this limitation.

3 KNOWA – the pivot extension

In this section we illustrate a variation of the basic KNOWA algorithm, which tries to solve the problem of multiple potential correspondence relations. To illustrate the problem, let us consider the example in Table 2, where wrong alignments are marked with a cross.

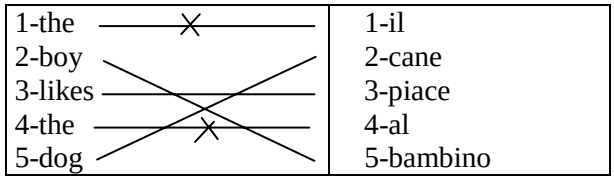

Table 2: Errors due to multiple potential correspondence relations

In the Italian translation the order of the English noun phrase is inverted. This is due to the fact that the Italian translation of "likes" follows a different verb subcategorization pattern. What is an object in English becomes a subject in Italian, causing a problem to the basic KNOWA algorithm. In fact, KNOWA correctly aligns 2-*boy* with 5-*bambino*, and 5-*dog* with 2-*cane*, even if the English and Italian nouns are not in the same position in the respective sentences, thanks to a search in the Italian word window. However, KNOWA would also align 1-*the* with 1-*il*, and 4-*the* with 4-*al*. Actually 1-*the* is a potential correspondent of both 1-*il*, and 4-*al* (the correct translation), but KNOWA chooses 1-*il* because its position is nearest to 1-*the*.

To solve these problems we need to use a different strategy. The solution is based on the observation that content words tend to be less involved in multiple potential correspondences than function words, and that function words tend to be attached to content words. Thus the basic idea amounts to trying first the alignment of content words, and only in a second phase trying the alignment of function words *relative to the position of content words to which they are attached*. Alignments between content words act as *pivots*, around which the alignment of function words is tried.

In the example above, first the algorithm finds the following correct alignments:

2-boy <> 5-bambino
3-likes <> 3-piace
5-dog <> 2-cane

Then, it takes the first alignment and tries to align the word before 2-*boy* and the word before 5-*bambino*, finding the correct alignment between 1-*the* and 4-*al*, and so on.

We do not expect that all content words are equally good pivots. To assess the goodness of nouns, verbs, adjectives, and adverbs as pivot words, we run various experiments, taking only the content words of a specific PoS and some combinations of them as pivot words. The results of these experiments show that nouns, taken alone as pivots, produce the best results in comparison with other PoS or combinations of PoS.

We also considered an alternative strategy for selecting pivots words. Instead of using the PoS as a predictor for the goodness of a word as pivot, which actually amounts to saying that words in a certain PoS can be aligned with a lower error rate than others, we selected as pivots the words for which the potential correspondence relation with their translation equivalents in the other language is *one-to-one*. Given a word w^E in the English sentence and a word w^I in its Italian translation, we select w^E as a pivot word if, and only if, w^I is the only potential correspondent of w^E, and w^E is the only potential correspondent of w^I. Of course, content words, and nouns in particular, tend to have such property much more frequently than words with other PoS. However, not all nouns have this characteristics. On the other hand certain function words, for instance conjunctions, may be involved in a one-to-one potential correspondence relation.

Table 3 shows a complete English sentence with its translation, taken from MultiSemCor. All the pivot words involved in one-to-one potential correspondence relations, according the Collins dictionary, are connected by a solid line. Note that the relation between *2-temperatures* and *2-clima* is indeed one-to-one, but is not recorded in the reference dictionary, so it is marked with a dotted line in the table.

Table 4 exemplifies instead typical cases of non-pivot words: *9-rovente* is the only potential translation of *1-sizzling*, but *9-rovente* can also translate *2-hot*, so neither *1-sizzling* nor *4-hot* are selected as pivot words.

The pivot extension of KNOWA has strong similarities with a strategy that is used by various statistics-based algorithms, aiming at selecting at first the translation correspondents that are most probably correct. Once these pivotal correspondences have been established, the remaining alignments are derived using the pivots as fixed points. Given that fact that these algorithms do not exploit bilingual dictionaries, the selection of the pivotal translation correspondent may be based on cognates, or specific frequency configurations. See among others (Simmard and Plamondon, 1998) and (Ribeiro et al., 2000).

The results obtained by applying the one-to-one potential correspondence as criterion for selecting pivot words are illustrated further on in Section 5.

1-Sizzling	1-Il
2-temperatures	2-clima
3-and	3-torrido
4-hot	4-e
5-summer	5-i
6-pavements	6-marciapiedi
7-are	7-dell’
8-anything	8-estate
9-but	9-rovente
10-kind	10-non
11-to	11-sono
12-the	12-niente
13-feet	13-di
	14-buono
	15-per
	16-i
	17-piedi

Table 3: *pivot words* involved in one-to-one potential correspondences

1-Sizzling	1-Il
2-temperatures	2-clima
3-and	3-torrido
4-hot	4-e
5-summer	5-i
6-pavements	6-marciapiedi
7-are	7-dell’
8-anything	8-estate
9-but	9-rovente
10-kind	10-non
…	…

Table 4: typical potential correspondences for *non-pivot words*

4 KNOWA - the breadth-first extension

The pivot extension to the basic KNOWA algorithm is based on two main hypotheses: first, certain words, which we call pivot words and which are mainly content words, are easier to align than others; second, the position of the other words, mainly function words, is anchored to the position of pivot words. This means for instance that if an article is near to a noun in Italian, we expect the English translation of the article to be near the English translation of the noun.

However if we look closer to the way the basic algorithm explores the search space of the potential correspondent in the word window, we will see that such strategy is inconsistent with the above two hypotheses. Suppose that we start from a pivot word w^E_1, in position p^E_1, as illustrated in Table 5, where pivot words are included in box. Then, we

try to align a non-pivot word w^E_2 occurring in position p^E_1+1. If the correspondent of w^E_1, that is w^I_1, occurs in position p^I_1, then we expect the correspondent of w^E_2, to occur in position p^I_1+1. Now, if w^I_2 turns out not to be the potential correspondent of w^E_2, possibly because w^E_2 has not been translated, KNOWA will extend the word window of w^I_2, and search the potential correspondents in position $p^I_1 \pm 2$, $p^I_1 \pm 3$, and so on, up to *MaxDelta*. We describe this by saying that the basic algorithm searches potential correspondents in the word window following a *depth-first* search strategy. Unfortunately, such strategy can cause alignment errors. Suppose that w^E_3 is another pivot word in position p^E_3, to be aligned with w^I_3 in position p^I_3, and that w^E_4 is a non-pivot word in position p^E_3+1, to be aligned with w^I_4, in position p^I_3+1. Suppose also that w^E_2 is a potential correspondent of w^I_4. Because of the depth-first search strategy, the basic KNOWA algorithm will align w^E_2 and w^I_4 wrongly. This kind of error can be avoided by adopting what can be called a *breadth-first* search strategy. In practice, for each pivot word we first search the potential correspondent in a word window of 0, that is in the expected initial position, then for each pivot word we search potential correspondents in a window of ±1, and so on up to the *MaxDelta*. The results of testing these strategy are reported in the following section.

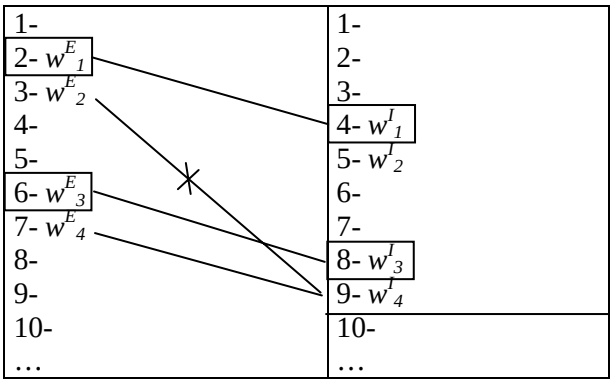

Table 5: Wrong alignment caused by the first-depth search strategy in the word window 1-9.

5 The experiments

We have run the experiments on two tasks, the EuroCor and the MultiSemCor alignment tasks. We call *EuroCor* a reduced and revised version of *EuroParl*, a multilingual corpus extracted from the proceedings of the European Parliament, see (Koehn, unpublished). EuroParl includes texts in 11 European languages, automatically aligned at the sentence level, whereas EuroCor includes only a part of the texts in EuroParl and only for English and Italian. On the other hand, MultiSemCor is a reference English/Italian corpus being developed at ITC-irst, including SemCor (part of the Brown Corpus) and its Italian translations. MultiSemCor has been created with the purpose of automatically transfer lexical semantic annotations from English to Italian (Bentivogli and Pianta, 2002).

For our experiments on EuroCor, we used as gold standard (and test set) a text that, following the EuroParl naming conventions, can be identified as ep-98-09-18. The revised version of this text includes 385 sentences, and has been manually aligned at the word level. Also sentence alignment has been manually revised.

For our experiments on MultiSemCor we used a gold standard composed of 6 files, manually aligned. Three of them have been exploited as development set and three as test set. In order to keep the test set as unseen as possible, the experiments whose main goal is tuning the algorithm by comparing various alignment strategies or parameters have been run on the development set. Once the best configuration has been obtained on the development set, we gave the results of running the algorithm with such configuration on the test set.

In our first experiment we run GIZA++ on both EuroCor and MultiSemCor. At first, we run GIZA++ on the entire English/Italian part of EuroParl, including around 694,000 sentences. The training of GIZA++ on this big corpus took around two weeks only for the English-to-Italian direction, on a high-level Sun Spark with 4 GB of memory. For this reason we decided to run the subsequent experiments on EuroCor, a reduced version of EuroParl, including around 21,000 sentences. EuroCor includes the following texts from EuroParl: ep-96-05-08, ep-97-04-07, ep-98-04-01, ep-90-11-04, ep-99-01-14, ep-99-10-05, ep-00-06-13, ep-00-09-04, ep-01-04-02, ep-01-04-03. the file in the gold standard, ep-98-09-18, should be added to these texts. These texts where chosen randomly, sampling them from as diverse periods of time as possible. Note that GIZA++ cannot be tested on a test set distinct from the training set. Thus we trained GIZA++ on the whole EuroCor corpus, including the file in the test set. Given the fact that we are simply using GIZA++ as a black box without having access to the internals of the alignment program, this seems acceptably safe from a methodological point of view. In all our experiments with GIZA++ we adopted a configuration of the system which is reported by the authors to produce optimal results, that is $1^5H^5 3^4 4^4 5^4$, where the number in the base refers to the IBM models 1, 3, 4, and 5, H refers to the HMM training, and the superscript figures refer to the number of iterations.

5.1 The EuroCor task

The first training of GIZA++ on EuroCor gave the following disappointing results on all-words alignment: 59.7% precision, 14.1% recall. After inspection of the corpus, we realized that the original files in EuroParl contain tokenization errors, and what counts more, a big number of sentence alignment errors. For this reason we produced a revised version of EuroCor, fixing these errors as extensively as possible.

A new run of GIZA++ on the revised EuroCor gave the following result: P:62.0%, R:34.7% on all word alignment; P:53.2%, R:38.3% on content words only. These results compare badly with those reported by (Och and Ney, 2003) on the Hansard alignment task. For this task, the authors report a precision of 79.6%, for a training on a corpus of 8,000 sentences. Explaining such a difference is not easy. A first explanation can be the fact the EuroCor task is inherently harder than the Hansard task. Whereas in the Hansard corpus the texts are direct translations of each other, in the EuroCor corpus it happens quite frequently that the English and Italian texts are both translation of a text in a third language. As a consequence, the texts are much more difficult to align. A better and more systematic revision of the sentence alignments could also improve the performance of GIZA++.

The basic version of KNOWA run on the EuroCor test file gives the results reported in Table 6. These results confirm the difficulty of the EuroCor task, but are quite encouraging for KNOWA, given that no special tuning was made to obtain them. It is interesting to note that whereas GIZA++ performs better on the all-word task than on the content-only-word task, KNOWA gets better results on the content-word-only task. Although it is true that aligning function words seems inherently more difficult than aligning content word, the worse result obtained by a statistics-based algorithm such as GIZA++ on the content-words-only task may be explained by the fact that data about content words are more sparse that data about function words.

		Precision	Recall
GIZA++ 22k	all	62.0	34.7
	content	53.2	38.3
KNOWA basic	all	63.4	41.6
	content	85.5	53.2

Table 6: GIZA++ and KNOWA-basic on the EuroCor task

5.2 The MultiSemCor task

The training of GIZA++ on MultiSemCor has been quite problematic, due to the small dimensions of MultiSemCor. In the current phase of the project, only 2,948 sentences are available. This is a small corpus which allows for only an approximate comparison with the experiment reported by Och and Ney (2003) on a set of 8,000 sentences from the Hansard corpus. Also, the authors report an improvement of around 7 points in precision, in passing from a corpus of 8,000 to 128,000 sentences. As the ultimate version of MultiSemCor is expected to include more than 20,000 sentences, we can expect a non negligible improvement in precision when GIZA++ will be applied to the final version of MultiSemCor.

To simulate at least partly the improvement that one can expect from an increase in the size of MultiSemCor, we trained GIZA++ on the union of the available MultiSemCor and EuroCor. The results of the training on MultiSemCor only, and on the union of MultiSemCor and EuroCor are reported in Table 7. Besides the row for the all-word task, the table contains also a SemCor row. This task concerns all the words that have been manually tagged in SemCor, and roughly corresponds to the content-word task. As the purpose of MultiSemCor is transferring lexical annotations from the English annotated words to the corresponding Italian words, it is particularly important that the alignment for the annotated words be correct. The results showed that GIZA++ works consistently better in the Italian-to-English direction, rather than vice versa, so we report the former direction. Only for the training on the union of the MultiSemCor and EuroCor data, we also report the results calculated by resorting to the symmetrization by intersection of the two alignments. Table 7 below shows that the MultiSemCor task is less difficult than the EuroCor Task; that GIZA++ consistently performs worse on content words; and finally that the increase in the dimensions of the training corpus produces a non marginal improvement in the precision, although not in the recall measure. Symmetrization produces a big improvement in precision but also an unacceptable worsening of the recall measure for GIZA++.

The two last rows in the table report the performances of the basic version of KNOWA in the same two tasks. These results show that given the available resources, KNOWA outperforms GIZA++ in all tasks. This is even clearer if we consider the extended versions of KNOWA, as reported in Table 8. Finally Table 9 reports the results of KNOWA on the test set.

	task	Prec.	Recall
GIZA++ 3k (MSC) It ->En	all	68.9	53.5
	semcor	*60.4*	*55.1*
GIZA++ 25k (MSC+EC) It ->En	all	73.4	55.2
	semcor	81.9	52.9
GIZA++ 25k (MSC+EC) intersec	all	95.2	38.8
	semcor	95.8	37.1
KNOWA basic	all	84.5	63.7
	semcor	92.0	73.4

Table 7: GIZA++ and KNOWA-basic on the MultiSemCor task (development set)

KNOWA version	task	Prec.	Recall
pivot (nouns) depth-first	all	86.8	65.3
	semcor	*92.5*	*73.6*
pivot (1-to-1) depth-first	all	88.1	66.5
	semcor	*92.8*	*74.4*
pivot (1-to-1) breadth-first	all	89.4	67.5
	semcor	***93.0***	***74.6***

Table 8: KNOWA-enhanced on constrained translation (development-set)

KNOWA version	task	Prec.	Recall
best *(on free tran.)*	all	82.1	56.9
	semcor	*89.1*	*66.5*
best *(constr. tran.)*	all	87.0	66.6
	semcor	***91.8***	***72.8***

Table 9: KNOWA-best on test set (free and constrained translation)

6 Conclusion

In this paper we compared the performances of two word aligners, one exclusively based on statistical principles, and the other intensively based on linguistic resources. Although statistics-based algorithms are very appealing, because they are language independent, and only need a parallel corpus of reasonable size to be trained, we have shown that, from a practical point of view, the lack of parallel corpora with the necessary characteristics can hamper the performances of the statistical algorithms. In these cases, an algorithm based on linguistic resources, if available, can outperform a statistics-based algorithm.

Also, knowledge-intensive word aligners may be more effective when word alignment is needed for special purposes such as annotation transfer from one language to another. This is the case for instance of the MultiSemCor project, in which, apart from a better performance in terms of precision and recall, a word aligner based on dictionaries, such as KNOWA, has the advantage that it will fail to align words that are not synonyms. The alignment of non-synonymous translation equivalents, which are hardly found in bi-lingual dictionaries, is usually a strength of corpus-based word aligners, but turns out to be a disadvantage in the MultiSemCor case, where the alignment of non synonyoums words causes the transfer of wrong word sense annotations from one language to the other.

References

Lars Ahrenberg, Magnus Merkel, Anna Sågvall Hein and Jörg Tiedemann. 2000. Evaluation of word alignment systems. In *Proceedings of LREC 2000*, Athens, Greece.

Luisa Bentivogli and Emanuele Pianta. 2002. Opportunistic Semantic Tagging. In *Proceedings of LREC-2002*, Las Palmas, Canary Islands, Spain (2002).

Philipp Koehn. Unpublished. *Europarl: A Multilingual Corpus for Evaluation of Machine Translation*, unpublised draft, available at http: //www.isi.edu/~koehn/publications/europarl.ps.

Dan I. Melamed. 2001. *Empirical Methods for Exploiting Parallel Texts*. The MIT Press, Cambridge, Massachussets.

Franz J. Och. 2003. GIZA++: Training of statistical translation models. Available at http://www.isi.edu/~och/GIZA++.html.

Franz. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51.

António Ribeiro, Gabriel Lopes and João Mexia. 2000. Using Confidence Bands for Parallel Texts Alignment. In *Proceedings of the 38th Conference of the Association for Computational Linguistics* (ACL 2000), Hong Kong, China, 2000 October 3–6. pp. 432–439.

Michel Simard and Pierre Plamondon. 1998. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. In *Machine Translation*, 13(1):59-80.

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

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