Phrase-based alignment combining corpus cooccurrences and linguistic knowledge

Adrià de Gispert, José B. Mariño and Josep M. Crego

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
{agispert|canton|jmcrego}@gps.tsc.upc.es

Abstract

This paper introduces a phrase alignment strategy that seeks phrase and word links in two stages using cooccurrence measures and linguistic information. On a first stage, the algorithm finds high-precision links involving a linguistically-derived set of phrases, leaving word alignment to be performed in a second phase. Experiments have been carried out for an English-Spanish parallel corpus, and we show how phrase cooccurrence measures convey a complementary information to word cooccurrences, and a stronger evidence of a good alignment. Alignment Error Rate (AER) results are presented, being competitive with and even outperforming state-of-the-art alignment algorithms.

1. Introduction

Generally speaking, the automatic alignment task aims at revealing the relationship between bilingual units (be them words, subwords or phrases) in a given parallel corpus, detecting which words from each language are connected together in a given situation. This has many applications in natural language processing, such as bilingual dictionaries extraction or transfer rules learning. However, it is in the context of statistical machine translation where it becomes particularly crucial. As an essential block in the learning process of current statistical translation models (single-word or phrase-based, conditional- or joined-probability based), its correct production has a sound correlation with translation quality [1].

This relevance has been corresponded by many previous works on the matter, including a shared task in the frame of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts [2]. Several competing systems were presented and evaluated against a manual reference using AER, the most widely used evaluation measure. Among the wide range of approaches presented, two basic trends should be highlighted.

On the one hand, the one based on the freely-available GIZA++ software [3], implementing well-known IBM and HMM translation models [4, 5], is considered to be the state-of-the-art. Nearly all current approaches to statistical translation rely on the results of Giza-based alignment processes to learn their models (see [6, 7, 8] among others). Although this approach, which is not solely focused on the alignment task (as alignment is considered to be a sort of hidden variable in a complete translation model), provides quite satisfactory results even with small corpora, it suffers from two structural flaws that limit its performance. Due to the model definition of word alignment as a function from positions in target sentence to positions in source sentence, it is strictly asymmetric, generating one-to-many word alignments that do not account for many translation phenomena. This effect has been tackled by several kinds of symmetrization heuristics (all of them linguistically blind), in search of a strategy to provide posterior phrase-based translation systems with the most accurate possible source. Moreover, the complexity of IBM models and their overload of parameters to estimate turn it very hard to introduce linguistic information into this setting in a reasonable way (some efforts being done in [9]).

On the other side of the spectrum, we find the approach based on word cooccurrences and link probabilities, presented in [10]. Its relative simplicity, its flexibility to introduce more knowledge sources, its symmetry and its promising results [2] make it appealing despite its dependence on empirical data and tuning strategies. However, the most important disadvantage of this approach is the one-to-one constrain, producing high precision alignments with low recall, what can represent a severe limitation to its use in practical translation systems.

We present in this paper an alignment strategy that is also based on bilingual cooccurrences, but aims at finding phrase-to-phrase alignment by using linguistic knowledge and, thereby, overcoming the one-to-one limitation. In section 2 this new strategy is described in detail, whereas sections 3, 4 and 5 describe our experimental framework and provide and discuss partial and complete alignment results. Finally, section 6 concludes and out-

This work has been partially supported by the Spanish Government under grant TIC2002-04447-C02 (ALIADO project), the European Union under grant FP6-506738 (TC-STAR project) and the Dep. of Universities, Research and Information Society (Generalitat de Catalunya).

lines future research lines.

2. From word-based to phrase-based alignment

Recent research efforts in statistical translation have clearly focused on improving translation quality by training models not only based on single words, but also on phrases. Typically, initial Giza-based alignments are generated and a symmetrization strategy is followed to obtain the core alignment from which bilingual phrases are built. These in turn are fed to the statistical translation model for estimation. However, current symmetrization strategies lack linguistic knowledge to decide ambiguities, and the translation model faces the task of learning translation probabilities from a noisy source. To face this problem, we present an alignment strategy that generates directly a phrase alignment from corpus cooccurrence counts. In contrast to previous work [11], we do so by generating very high-confidence links between phrases before proceeding onto word alignment. This search for phrase links is limited to a small adequate set of possible phrases. In the following subsections our proposal is described in detail.

2.1. Word and phrases association measures

Association or cooccurrence measures extracted from parallel corpora give strong evidence of so-called translation equivalence [12], or simply alignment adequacy, between a pair of phrases or words. Among these measures we find Dice-score, ϕ^2 score and some others, offering a similar performance. In this paper, we have used ϕ^2 as presented in [13], which is defined by the following equation:

$$\phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \qquad (1)$$

where a is the number of times two words (or phrases) cooccur, b and c are the number of times one occurs and the other does not, and finally d accounts for the number of times neither one nor the other occur in the data set. In our implementation, we have defined the cooccurrence of a word (or phrase) appearing x times in a sentence and a word (or phrase) occurring y times in its translation as $min(x, y)$, for two reasons: on the one hand, the alternative option given by the product xy leads to confusing results when computing b and c, as these can be negative because the times a word cooccurs with another can overweight the total occurrences of the word. On the other hand, the word alignment algorithm used estimates link probabilities from existing one-to-one links (see section 2.2.3), being $min(x, y)$ the maximum number of links that can be established between the two words (in which case their probability is the highest). This way we preserve stochastic consistency.

Despite this score can be easily computed for each possible pair of words from both languages, computational problems arise when dealing with every bigram, trigram or, in general, phrase for each language. However, these scores can convey a useful complementary information in many cases, as in the examples of table 1, where the phrase-to-word score -in bold face- is comparatively much better than all word-word scores for all words involved in Spanish idioms 'por favor' and 'a lo mejor' (note that $-10log$ will be assumed when referring to ϕ^2). Furthermore, it is reasonable to expect that the longer the phrases considered, the stronger the evidence of a good alignment adequacy, so long as we have a reasonable number of occurrences of the phrase.

Table 1: *Examples of ϕ^2 between words and phrases.*

	please	
por	22.4	**0.9**
favor	1.2	

	maybe	
a	23.1	**8.0**
lo	18.2	
mejor	12.2	

The main problem is then the practical impossibility to compute all combinations for even relatively small corpora (see table 2 for the number of cooccurrence combinations considering all word, bigrams and trigrams in the relatively-small corpus used in section 3). To tackle this, we propose an algorithm that tries to extract as much useful information from these phrase cooccurrence measures by performing a selection of only a subset of all possible phrases. We have chosen to perform this selection using linguistic criteria, although statistics could also be used, as discussed in section 2.2.1. This way, we expect the algorithm to leverage linguistic knowledge and empirical evidence in a memory-efficient way.

Table 2: *Number of different cooccurrence entries considering unigrams, bigrams and trigrams for VerbMobil English-Spanish corpus.*

	# Coocs
unigrams	0.35M
+bigrams	4.36M
+trigrams	10.42M

2.2. A phrase alignment strategy

We propose a phrase alignment strategy in four stages, as shown in figure 1. Firstly, from all possible sets of words we select a small set of 'interesting' phrases for each language. This selection is linguistically guided and should produce a set of phrases containing words that play a unique semantic role. Secondly, a high-precision phrase alignment algorithm links these phrases together (with phrases or single words) using cooccurrence mea-

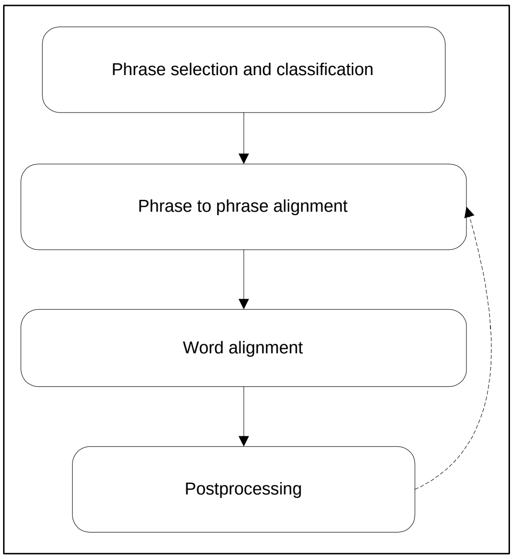

Phrase selection and classification
Phrase to phrase alignment
Word alignment
Postprocessing

Figure 1: *Phrase alignment stages.*

sures, and discarding uncertain links. After these phrase alignments have been produced, we run a word-to-word alignment algorithm based on link probabilities and shallow syntactic information in the fashion of [10]. This stage takes advantage of the complexity reduction derived from the previous linking. Finally, in the fourth stage a postprocessing of the resultant alignment is done, disambiguating cases such as unaligned words, and making decisions at a sentence level. These blocks are now described in detail.

2.2.1. *Phrases selection and classification*

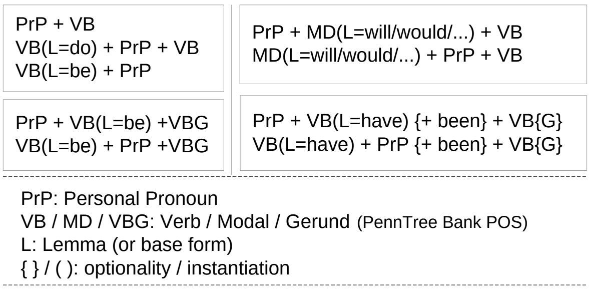
The objective of this stage is double. Given the exponential nature of the amount of different phrase cooccurrences shown above, which makes it impossible to work with cooccurrences between all combinations for each language, a first objective is the reduction of this huge space to those being 'interesting' from an alignment/translation point of view. Our criterion is one of so-called translational equivalence, so we define as interesting those phrases expressing a same concept or being semantically linked in one language, as it is reasonable to expect that these might be aligned to (or might translate into) a single word (or phrase) in another language. On the other hand, a semantic classification of these phrases should improve cooccurrence measures by adding different instantiations of the same concept to a same measure.

As about the selection of phrases, we have followed a linguistically-guided strategy. Specifically, we have implemented two selection criteria using complementary knowledge. Firstly, we detect **verb groups** using deterministic automata that implement a few simple rules, as shown in example 2. These rules take as input word forms, POS-tags and word lemmas (or base forms), and map the resulting phrase to the lemma of the head verb. This way, the classification improves coocccurrence counts for verb groups no matter how their full form is expressed, as long as they share the same base form of the head verb. This way, forms like

`we have brought` or `will we bring`

are considered equivalent and add a cooccurrence count for the base form ”bring”, increasing its evidence and reducing evidence for function words like ”have” and ”will” that act as modifiers and may therefore be expressed in many ways in the other language, as they do not convey a stand-alone meaning in the sentence. We expect this to produce a special gain in languages that have important declination in verb forms (like languages belonging to the Romance family, as Spanish).

After detection rules are applied, all remaining words with POS-tag of a Verb are substituted by their base form to enforce the verb's cooccurrence evidence. However, to avoid possible mistakes from the POS-tagger or lemmatizer and ensure we are dealing with a known verb, we filter out the resulting base form against two lists containing 25988 possible verb forms in English and 12668 possible verb lemmas in Spanish, available from the *maco+* and *relax* tagging package [14].



PrP + VB
VB(L=do) + PrP + VB
VB(L=be) + PrP

PrP + MD(L=will/would/...) + VB
MD(L=will/would/...) + PrP + VB

PrP + VB(L=be) +VBG
VB(L=be) + PrP +VBG

PrP + VB(L=have) {+ been} + VB{G}
VB(L=have) + PrP {+ been} + VB{G}

PrP: Personal Pronoun
VB / MD / VBG: Verb / Modal / Gerund (PennTree Bank POS)
L: Lemma (or base form)
{ } / ( ): optionality / instantiation

Figure 2: *Verb group detection rules used for English.*

We have also implemented a selection based on **idiomatic expressions**. Specifically, we match the corpus against a list of 1496 and 49 usual idiomatic expressions that we have available for Spanish and English, respectively (again from the *maco+* and *relax* package). These expressions (among which we find example like 'on the other hand' for English, and 'sin embargo' or 'a lo mejor' for Spanish) tend to convey a single meaning and we can expect them to be aligned *together* to one or more words in the other language. At the moment, we have not used any kind of dictionary, so these expressions are not classified according to their meaning(s). Other possible linguistically-guided selection rules could include regular expressions such as numbers, dates or times of the day (that could also be classified) or even collocations and phrasal verbs. As this selection is language-dependent, every language will define its own adequate rules.

If no linguistic knowledge is available, statistical procedures can also be used to obtain a set of possible phrases. For example, we can select the N most frequent bigrams, trigrams and Ngrams in general, or the ones having a very high bigram, trigram or Ngram probability (defining groups of words that consistently appear together in the text). We plan to investigate this in the short term, as commented in section 6.

At the moment, we have restricted our selection to phrases built by sequential groups of words. However, as word alignment is not always affected by this restriction, it should be eliminated, allowing phrases to contain words that are separated by other words (such as separable phrasal verbs in English or many separable verbs in German).

It is important to note that we do not expect this selection to be exhaustive, nor does it imply that the selected phrase will necessarily be linked *together* at the next stage (it is not a hard decision in terms of alignment). It is the phrase alignment stage that decides whether a phrase should be linked together, or whether the words should be left free to be linked word-to-word.

2.2.2. Phrases alignment

In this stage cooccurrence measures are computed for each selected phrase in one language against all selected phrases and single words in the other language. Then, a competitive linking strategy [12] is used, but not until all words or phrases are linked, but until a certain threshold is surpassed. Basically, we choose the link with best phrase-phrase or phrase-word cooccurrence measure as long as this is better than the threshold. This strategy relies on the fact that phrase cooccurrence measures are a stronger evidence of translational equivalence than word, and the threshold (which has to be empirically tuned) ensures that we generate only the highest-confidence links. This way, not all selected phrases will be linked, but only those having a high cooccurrence evidence in the data.

Once the linking of two phrases is decided, one can use several strategies to determine the internal links between words inside the phrases, if that is desired. For example, internal links can be solved using the general word alignment algorithm, but restricting the search inside the phrase positions. However, often selected phrases will contain function words that tend to depend on each language syntax and are not easily linked to the other language words. For this reason, we have decided to introduce all internal links between linked phrases.

2.2.3. Word alignment

As about the word alignment algorithm, we have implemented an iterative algorithm similar to the one presented in [10]. Basically, an initial alignment is generated using word cooccurrence measures, from which link probabilities are estimated. Then, a best first search is performed, following an heuristic function based on the global aligned sentence link probabilities. The search is further improved with a syntactic constrain (also called cohesion constrain [15]) and can introduce features on the links, such as a dependence on adjacent links. Our implementation allows certain positions to be prohibited, so that previous phrase alignment is fixed, although its links also compute in link probability estimation at each iteration.

Given the enormous space of possible word alignments to choose from, the heuristic function becomes the key to efficiency, so long as it is correctly defined. Basic parameters are:

- the initial **null probability**, or the prob. that a word links to null (no word), which is necessary to make fully- and partially-aligned solutions comparable

- and the **minimum score** to accept a link between two words (hereafter referred to as *mscore*)

These parameters must be set empirically for the optimal performance of the algorithm. We also found that restricting the search of possible links to a window in the other language not only made the algorithm much more efficient (turning it from exponential time to linear time with input sentence lengths), but also improved results by discarding the ambiguities generated by the repetition of frequent words (mostly function words). We define this window in the neighbourhood of the diagonal defined by the division between both sentence lengths. Of course, this window is completely dependent on the pair of languages considered (might even be eliminated for certain pairs), but in our case (English-Spanish) turned out to be optimal considering 8 words.

2.2.4. Postprocessing

The postprocessing stage should take the final alignment decisions using sentence-level information (ie. deciding whether unlinked words should be linked, looking for long-distance links, reconsidering the links for a word/phrase given all its links in all sentence pairs, etc). Ideally, it should also feedback into the phrase selection/alignment blocks to reconsider previous decisions using global information of all sentences. However, this stage is strongly connected to the posterior translation modeling. Although alignment can be and must be evaluated separately, we are of the opinion that it is not completely independent from the translation model. At the moment, we have not implemented any postprocessing technique in our system.

3. Experimental setting

3.1. Parallel corpus

To experiment with the alignment strategy described above, we have worked with an English-Spanish parallel corpus, namely the VerbMobil database, which has been translated to Spanish in the framework of the LC-Star project (IST-2001-32216). This data is the transcription of spontaneous dialogues in the appointments and meeting-planning domain. Table 3 shows the main statistics of the data used, namely number of sentences, words, vocabulary, percentage of singletons, and maximum and mean sentence lengths for each language, respectively.

Table 3: *Corpus used: 30054 sentences per language.*

VMobil	words	vocab.	singlets.	Lmax	Lmean
English	228328	3276	39 %	66	7.6
Spanish	219782	5084	43 %	66	7.3

3.2. Preprocessing

All preprocessing that has been carried out is described as follows:

- Normalization of contracted forms for English (ie. wouldn't = would not, we've = we have) and Spanish (del = de el)
- English data has been tagged using freely-available *TnT* tagger [16], and base forms have been obtained using *wnmorph*, included in the WordNet package [17].
- Spanish data has been tagged using *maco+* and *relax* package already mentioned. This software also generates a lemma or base form for each input word.
- Date and time expressions (which are numerous in the domain) have been substituted by a unified tag using a semi-automatic technique [18].
- Finally, punctuation marks have been left out (as we expect to use the corpora for spoken language translation experiments).

3.3. Evaluation scheme

For evaluation purposes, we have randomly selected from this data *two sets*: a validation set of 100 sentences (for tuning of parameters) and a test set of 400 sentences. These have been manually aligned following the criterion of Sure and Possible links, in order to compute Alignment Error Rate (AER) as described in [19] and widely used in literature (including the above-mentioned HLT-NAACL 2003 Shared task).

It has been shown that the percentage of Sure and Possible links in the gold standard reference has a strong influence in the final AER result, favouring high-precision alignments when Possible links outnumber Sure links, and favouring high-recall alignments otherwise [20]. Our criterion has been to produce Possible links only when they allow combinations which are considered equally correct, as a reference with too many Possible links suffers from a resolution loss, causing several different alignments to be equally rated. This way, we have 80% Sure links and 20% Possible links. Evaluations are performed without taking links to NULL into account.

4. Phrase alignment results

In this section, we evaluate separately the phrase selection, classification and alignment blocks described in sections 2.2.1 and 2.2.2. First we present results of the phrase alignment selecting only verb groups, to continue with results with only idiomatic expressions alignment. Finally, complete alignment results are presented, comparing performances for the isolated word-to-word alignment algorithm and the complete phrase-alignment strategy proposed against state-of-the-art alignments.

4.1. Verb groups

Verb groups detection rules include 14 rules for English language and just 6 for Spanish, which usually employs declined verb forms omitting thus personal pronouns and using thus a single word. Verb groups rules have detected a total of:

- 1156 verb groups in English, classified into 238 different verb lemmas
- 658 verb groups in Spanish, classified into 188 different verb lemmas

The classification of these phrases produces an efficient reduction of the cooccurrence table from 0.35M to 0.33M, we do not compute cooccurrence counts for all words internal to the phrase, but just for the lemma of its head verb. The results of the phrase alignment with these phrases are shown in the upper side of table 4, changing the value of the threshold to accept phrase links from more restrictive to less restrictive. A restriction that the linked pair cooccurs at least twice has also been used.

Surprisingly, this relatively simple selection strategy provides very encouraging results, as Precision is consistently higher than 98 % for the three thresholds used, whereas Recall achieves around 9%. We have to keep in mind that these phrase links will necessarily boost Recall with respect to the isolated word aligner, as it is a one-to-one algorithm, unable to produce these links. As about Precision, the high figures are due to the greater statistical evidence of phrase cooccurrence measures with respect to

Table 4: *Phrase alignment results considering verb groups only and idiomatic expressions only.*

	Recall	Precision
Verbs $\phi^2 < 8$	8.07	99.02
Verbs $\phi^2 < 10$	9.00	99.12
Verbs $\phi^2 < 15$	9.68	98.69
Idioms $\phi^2 < 5$	2.01	98.48
Idioms $\phi^2 < 10$	3.06	99.00
Idioms $\phi^2 < 15$	3.50	97.41

single word cooccurrences. Interestingly, no error is introduced from the first to the second case, whereas Recall is increased by 1.2 absolute points.

4.2. Idiomatic expressions

Regarding idiomatic expressions, a total of 20 English and 99 Spanish expressions have been detected in the parallel corpora, matching with the available lists presented in section 2.2.1. These have not been classified, leading to an increase in the cooccurrences table from 0.35M to 0.37M entries. The results when aligning only idiomatic expressions phrase-phrase and phrase-words links are shown in the lower side of table 4, again for different thresholds.

In this case, although Recall is much smaller than when considering verb groups, two points are worth raising. First, we have again a nearly error-free alignment using a relatively small set of phrases. And second, but not less important, that we expect this Recall to complement the previous experiment and further boost the global alignment Recall, as we find no verb groups among the idiomatic expressions considered. Results for the complete phrase alignment strategy follow in the next section.

5. Complete alignment results and discussion

As baseline alignments, we have aligned our data using GIZA++ from English to Spanish and vice versa (performing $1^5 H^5 3^3 4^3$ iterations), and have evaluated two symmetrization strategies, namely the union and the intersection. Their results are shown in the first four rows of table 5.

We have also used the word alignment algorithm presented in section 2.2.3 to align the data without any kind of previous phrase selection and alignment, thus producing the one-to-one alignment shown in the fifth row. For this result, we have run three iterations with a mscore = 30, and three iterations further restricting it to 6 to achieve high precision, always using cohesion constrain and adjacency features up to two positions (initial NULL cost being set to 15). In contrast to giza++ intersection (the only baseline alignment that is also one-to-one and thus subject to a fair comparison), we observe a slight reduction in Precision and a slight increase in Recall, leading to an also slight AER reduction. However, both alignments skip around 30 % of good links (far below the other alignments recall), which make them unpractical for posterior statistical translation modeling.

Table 5: *Comparison of final alignment results.*

	Recall	Precision	AER
giza++ eng2spa	76.99	93.15	15.51
giza++ spa2eng	78.75	94.19	13.94
giza++ union	**84.47**	90.85	**12.30**
giza++ intersection	71.27	**97.58**	17.52
one-to-one word aligner	72.56	96.69	16.96
phrase aligner $\phi^2 < 10$	76.31	**97.48**	**13.36**
phrase aligner $\phi^2 < 15$	76.88	**97.35**	**13.20**

On the contrary, the giza++ union alignment, which rates the best AER, suffers from a severe decline in Precision as compared to all other alternatives, compensated by a very important boost in Recall.

Remarkably, our phrase word aligner, whose results are shown in the last two rows (selecting verb and idiomatic phrases with the two less-restrictive ϕ^2 thresholds shown in table 4), preserves as high a Precision level as the giza++ intersection, while providing a Recall increase of over 5 absolute points with respect to intersection, and of about 4 points with respect to the one-to-one aligner. The achieved Recall figure is absolutely comparable with all other alignments (that nevertheless offer worse Precision results), except for the giza++ union, still to be beaten in Recall.

It is interesting to note the Precision improvement when comparing the word-to-word aligner to the phrase aligner proposed, which is due to two factors. On the one hand, the previous phrase alignment introduces links with a higher precision than that of the one-to-one aligner, and on the other hand, this previous linking results in a complexity reduction (less ambiguity) that simplifies the task of the one-to-one aligner, improving its performance. We find these results to be very promising: the alignment strategy achieves competitive results while still making a relatively small use of linguistic knowledge. However, this has already led to an important recall boost with respect to the word-to-word aligner, at no precision cost. As its architecture is open to an easy introduction of more information, many other knowledge sources could be used, as outlined in the following and concluding section.

6. Conclusion and further research

In this paper, we have presented a phrase alignment strategy that combines cooccurrence measures extracted from bilingual corpora and linguistic knowledge in detail. Results have been reported, that show a very positive tendency, with a state-of-the-art Precision figure and a very high Recall, which could be improved in the short term. Besides, we do not find any alignment beating the rest in both Precision and Recall measures. However, when it comes to further research, and bearing in mind that we are interested in statistical machine translation, our two main research priorities are to evaluate this alignment strategy with other parallel corpora, and to evaluate it in the framework of a translation experiment, in addition to AER results. We plan to do so using several translation models in the very short term.

Future work also includes new and more linguistically-informed phrase selection schemes, such as allowing phrases containing non-consecutive words (ie., negative verb groups or expressions like 'i will probably leave') or searching for English phrasal verbs from available lists. Categorisation of numbers, dates and time expressions during the phrase selection and classification stage should also be investigated. It is our view that the alignment strategy, which does not imply a hard alignment decision when selecting the phrases, could help resolve the inherent ambiguity when detecting these expressions from a monolingual point of view (ie., 'a las tres' might refer to a time of the day or simply a number). Connected with this, another interesting point to investigate is to allow phrases to add a cooccurrence count for each of their meanings, in case they have more than one, so that alignment plays a sense disambiguation role. In this case, adequate dictionaries must be provided. Information from statistical chunkers could also be introduced in the medium term.

Regarding the alignment algorithm, we plan to investigate ways of allowing a certain degree of freedom to reconsider phrase links, which are currently fixed (although their precision is very high). Finally, postprocessing techniques discussed in section 2.2.4 will have to be tested, seeking a further boost in recall by using statistics of all aligned sentences. For example, new cooccurrence measures could be computed for all phrases that can be built in the neighbourhood of unaligned words, to consider whether these should be linked or left alone.

7. Acknowledgements

The authors want to thank Patrik Lambert, Javier Pérez and Antonio Bonafonte (members of the TALP Research Center) for their valuable contributions to this paper. We would also like to thank Colin Cherry for his kind input on the cooccurrences-based word alignment algorithm.

8. References

[1] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, March 2003.

[2] R. Mihalcea and T. Pedersen, "An evaluation exercise for word alignment," in *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, R. Mihalcea and T. Pedersen, Eds. Edmonton, Alberta, Canada: Association for Computational Linguistics, May 31 2003, pp. 1–10.

[3] Giza++ software, "http://www-i6.informatik.rwth-aachen.de/˜och/software/giza++.html."

[4] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer, "The mathematics of statistical machine translation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[5] S. Vogel, H. Ney, and C. Tillmann, "Hmm-based word alignment in statistical translation," *Proc. of the Int. Conf. on Computational Linguistics, COLING'96*, pp. 836–841, August 1996.

[6] F. Och, C. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation," *Proc. of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 20–28, June 1999.

[7] A. de Gispert and J. Mariño, "Using X-grams for speech-to-speech translation," *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September 2002.

[8] P. Koehn, F. Och, and D. Marcu, "Statistical phrase-based translation," *Proc. of the Human Language Technology Conference 2003 (HLT-NAACL'03)*, May 2003.

[9] K. Toutanova, H. Tolga Ilhan, and C. Manning, "Extensions to hmm-based statistical word alignment models," *Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP'02*, July 2002.

[10] C. Cherry and D. Lin, "A probability model to improve word alignment," *41st Annual Meeting of the Association for Computational Linguistics*, July 2003.

[11] D. Marcu and W. Wong, "A phrase-based, joint probability model for statistical machine translation," *Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP'02*, July 2002.

[12] D. Melamed, *Empirical Methods for Exploiting Parallel Text*. Cambridge, MA: MIT Press, 2001.

[13] W. Gale and K. Church, “Identifying word correspondences in parallel texts,” *4th Speech and Natural Language Workshop*, pp. 152–157, 1991.

[14] J. Atserias, J. Carmona, I. Castellón, S. Cervell, M. Civit, L. Màrquez, M. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo, “Morphosyntactic analysis and parsing of unrestricted spanish text,” *1st Int. Conf. on Language Resources and Evaluation, LREC’98,* 1998.

[15] C. Cherry and D. Lin, “Word alignment with cohesion constraint,” *Proceedings of HLT/NAACL’03*, pp. 49–51, Edmonton 2003.

[16] T. Brants, “TnT – a statistical part-of-speech tagger,” in *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA, 2000. [Online]. Available: http://www.coli.uni-sb.de/˜thorsten/tnt

[17] WordNet: a lexical database for the English language, “http://www.cogsci.princeton.edu/˜wn.”

[18] A. de Gispert and J. Mariño, “Experiments in word-ordering and morphological preprocessing for transducer-based statistical machine translation,” *IEEE Automatic Speech Recognition and Understanding Workhsop, ASRU’03,* November 2003.

[19] F. Och and H. Ney, “Improved statistical alignment models,” *38th Annual Meeting of the Association for Computational Linguistics*, pp. 440–447, October 2000.

[20] P. Lambert and N. Castell, “Alignment of parallel corpora exploiting asymmetrically aligned phrases,” *4th Int. Conf. on Language Resources and Evaluation, LREC’04,* May 2004.