Natural Language Engineering **12** (1): 91–108. © 2005 Cambridge University Press
doi:10.1017/S1351324905003931 First published online 6 December 2005 Printed in the United Kingdom

Linguistic knowledge in statistical phrase-based word alignment

A. DE GISPERT and J. B. MARIÑO

TALP Research Center, Universitat Politècnica de Catalunya (UPC),
Jordi Girona 1-3, Campus Nord D5, 08034 Barcelona, Spain
e-mail: {agispert,canton}@gps.tsc.upc.es

(*Received 4 May 2004; revised 10 May 2005*)

Abstract

In this paper, a novel phrase alignment strategy combining linguistic knowledge and cooccurrence measures extracted from bilingual corpora is presented. The algorithm is mainly divided into four steps, namely phrase selection and classification, phrase alignment, one-to-one word alignment and postprocessing. The first stage selects a linguistically-derived set of phrases that convey a unified meaning during translation and are therefore aligned together in parallel texts. These phrases include verb phrases, idiomatic expressions and date expressions. During the second stage, very high precision links between these selected phrases for both languages are produced. The third step performs a statistical word alignment using association measures and link probabilities with the remaining unaligned tokens, and finally the fourth stage takes final decisions on unaligned tokens based on linguistic knowledge. Experiments are reported for an English-Spanish parallel corpus, with a detailed description of the evaluation measure and manual reference used. Results show that phrase cooccurrence measures convey a complementary information to word cooccurrences and a stronger evidence of a correct alignment, successfully introducing linguistic knowledge in a statistical word alignment scheme. Precision, Recall and Alignment Error Rate (AER) results are presented, outperforming state-of-the-art alignment algorithms.

1 Introduction

The task of automatic word alignment focuses on detecting, given a parallel corpus, which tokens or sets of tokens from each language are connected together in a given translation context, revealing thus the relationship between these bilingual units. Among the many applications in natural language processing, such as bilingual dictionaries extraction or transfer rules learning, word alignment becomes particularly crucial in the context of statistical machine translation, where it represents an essential block in the learning process of current statistical translation models (single-word or phrase-based, conditional- or joined-probability based). Consequently, its correct generation shows a positive correlation with translation quality (Och and Ney 2003).

Actually, the relevance of word alignment has been corresponded by several previous works on the matter, including a shared task in the frame of the HLT-NAACL

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

2003 Workshop on Building and Using Parallel Texts (Mihalcea and Pedersen 2003). Several competing systems were presented and evaluated against a manual reference using Alignment Error Rate (AER), until now the prevailing evaluation measure. Two basic approaches stand out against the wide range of alternatives presented.

On the one hand, word alignment systems based on well-known IBM and HMM translation models (Brown, Della Pietra, Della Pietra and Mercer 1993; Vogel, Ney and Tillmann 1996)) are considered to be state-of-the-art. Nearly all current approaches to statistical translation depend on the results of these alignment processes to estimate their translation models, such as Och, Tillmann and Ney (1999), de Gispert and Mariño (2002) and Koehn, Och and Marcu (2003), among others. Typically, the implementation by Och (2003), which is freely-available in the GIZA++ package, is used. Despite not being entirely designed to tackle the alignment task (as it is regarded as a sort of hidden variable in a complete translation model), this solution provides quite satisfactory results even with small corpora. However, it suffers from two structural flaws that pose a severe limitation to its performance.

Due to the model definition of alignment as a function from positions in the target sentence to positions in the source sentence, the result is strictly asymmetric, generating one-to-many word alignments that do not account for many translation phenomena. Several kinds of symmetrization heuristics (all of them linguistically blind) have been proposed to deal with this effect, seeking the most accurate result to pass onto posterior phrase-based translation systems. Furthermore, the mathematical complexity of IBM models and their overload of parameters to estimate make it very hard to introduce linguistic information into this setting in a comfortable way, although some efforts have been done in Toutanova, Tolga Ilhan and Manning (2002).

From a considerably different standpoint, a word alignment can be produced using information on word cooccurrences and link probabilities, as introduced in Cherry and Lin (2003a). The relative simplicity of the approach, its flexibility to introduce more knowledge sources, its structural symmetry and its promising results (Mihalcea and Pedersen 2003) make it attractive despite its dependence on empirical data and tuning strategies. However, the most important disadvantage of the approach is the one-to-one constraint, producing high precision alignments with low recall, limiting thus its use in practical translation systems.

In this paper we propose an alignment strategy which is also based on bilingual cooccurrences, but aims at finding phrase-to-phrase alignments by using linguistic knowledge, and thereby, overcoming the one-to-one limitation. This knowledge is introduced by means of very simple rules made by non-linguists. A similar approach is followed in the framework of example-based machine translation in Brown (1999), in order to improve the coverage of the examples during translation. As it will be detailed in the following sections, in our case the rules provide a classification that leads to improved statistical word alignment performance.

The organization of the article is as follows. Section 2 describes this novel strategy in full detail, preceded by a brief introduction on word and phrase association

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

measures. Section 3 introduces the experimental setting used to evaluate the quality of the alignments produces, whereas sections 4 and 5 provide partial and complete alignment results, discussing the contribution of each part of the algorithm presented in the final result. Finally, section 6 concludes and outlines future research lines on the subject.

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 (sequences of consecutive words in one language). Typically, initial Giza-based alignments are generated and a symmetrization strategy follows to obtain the core alignment from which bilingual phrases are built. These units are then the input to statistical translation models 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 (Marcu and Wong 2002), we do so by producing very high-confidence links between phrases before proceeding onto word alignment, based on linguistic knowledge. 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 (Melamed 2001), 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 Gale and Church (1991), which is defined by the following equation:

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

where a is the number of times that two words (or phrases) cooccur in a document (the addition of the cooccurrence count for each sentence), b and c are the number of sentences where one word occurs and the other does not, and finally d accounts for the number of sentences neither one nor the other occur in the data set. All these counts are defined at document level, which means that they are the addition of the count for each sentence.

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

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

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

ϕ^2	please	
por	22.4	**0.9**
favor	1.2	

ϕ^2	maybe	
a	23.1	
lo	18.2	**8.0**
mejor	12.2	

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.

Even though 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 correct alignment, so long as we have a reasonable number of occurrences of the phrase.

The main problem is then the practical impossibility to compute all combinations for even relatively small corpora. 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.

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 measures, 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 Cherry and Lin (2003a). This stage takes advantage of the complexity reduction derived from the previous linking. Finally, in the fourth stage a postprocessing takes final decisions on unaligned words with certain linguistic criteria. These blocks are now described in detail.

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

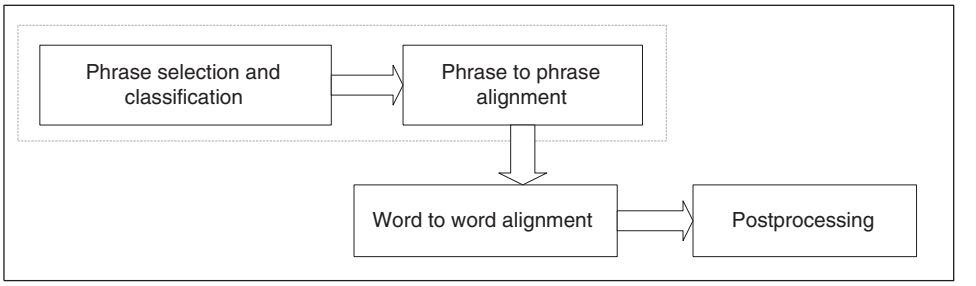


Fig. 1. Block diagram of the phrase-based word alignment strategy.

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 for the selection of phrases, we have followed a linguistically-guided strategy. Specifically, we have implemented three selection criteria using complementary knowledge.

Firstly, we detect **verb phrases** using deterministic automata that implement a few simple rules using word forms, POS-tags and word lemmas as input, and mapping the resulting phrase to the lemma of the head verb (see Figure 2 for some rules and examples of detected verb phrases). This way, the classification improves cooccurrence counts for verb phrases no matter how their full form is expressed, as long as they share the same base form of the head verb. Therefore, forms like

```
we have brought
will we bring
I would not bring
she probably brings
```

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 using heavily inflected verb forms (like languages belonging to the Romance family, as Spanish).

Our automata allow detection of non-sequential verb phrases, that is, phrases containing words that will not be part of the selected phrase, and therefore will not

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

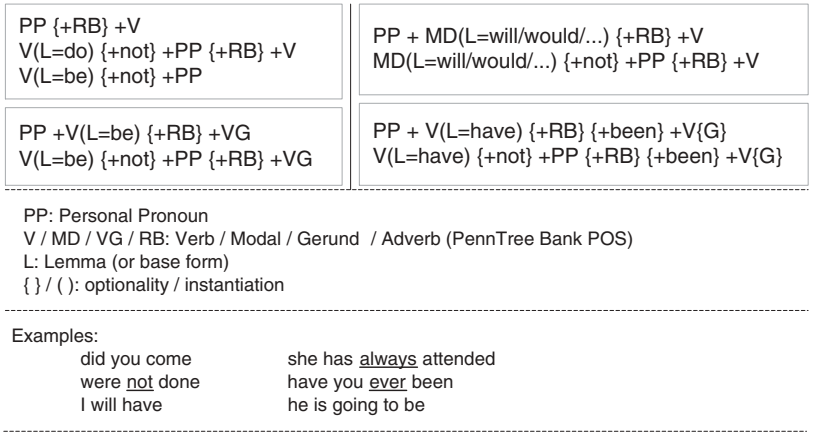


Fig. 2. Some verb phrase detection rules and detected forms in English.

be linked together (like the underlined words in the last two examples above). At the moment, we restrict this case to adverbs modifying English verb phrases (or negative forms), but other linguistic phenomena could be tackled similarly, such as separable phrasal verbs in English, provided a list of these verbs is available.

Since the rules detect only phrases with at least two tokens, all remaining words with POS-tag of a Verb are substituted by their base form to enforce the verb's cooccurrence evidence, after detection rules have been applied. This way, in the following example,

Stuart **will come** if Henry *comes* too
Stuart *vendrá* si Henry *viene* también

only the phrase 'will come' – in bold face – is detected (and classified to lemma 'come'), but words 'comes', 'vendrá' and 'viene' – in italics – are also substituted by their base forms 'come', 'venir' and 'venir', respectively, improving the cooccurrence measure between the pair 'come,venir' during the word-alignment stage. Of course, this is only possible if tagging gives a Verb tag for each of these tokens. Lemmatisation ambiguity did not affect the corpus used in our experiments.

Secondly, we have also implemented a selection based on **idiomatic expressions**. Specifically, we match the corpus against a list of 1497 and 50 usual idiomatic expressions that we have available from the *FreeLing* language analysis tool (Carreras, Chao, Padró and Padró 2004) for Spanish and English, respectively. These expressions (among which we find examples 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).

Finally, **date expressions** have also been detected. Although the corpus used already contained some tokens substituting 'dates' which included a month of the year (see section 3.2 for corpus description), several 'date' expressions (without a month of the year) were still found in the data. Therefore, we implemented a basic automaton

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

detecting expressions like 'on Monday the tenth', 'Sunday the fifth' or 'on the second' for English, and mapping the resulting phrase to a unified token, so that all date expressions contribute to the same cooccurrence measure with all the dates in the other language. Furthermore, we keep the *value* of the expression (meaning the day of the week and day of the month) for further use in the alignment phase discussed in the next section. Similarly for Spanish, expressions like 'el lunes día diez' or 'el domingo cinco' are detected.

Other possible linguistically-guided selection rules could include more regular expressions such as numbers or times of the day (that could also be classified) or even collocations, phrasal verbs or more complex structures. As this selection is language-dependent, every language will define its own adequate rules. If no linguistic knowledge is available, statistical procedures could also be used to obtain a set of possible phrases. For example, selecting the N most frequent bigrams, trigrams and Ngrams in general, or the ones having a very high bigram, trigram or Ngram probability (defining phrases of words that consistently appear together in the text).

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 is used, but not until all words or phrases are linked, but until a certain threshold is surpassed. Basically, this greedy strategy produces an alignment solution by iteratively choosing the link with best phrase-phrase or phrase-word cooccurrence measure as long as this is better than the threshold (Melamed 2001). Links are only selected if they link positions which have not been linked before.

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, in our case we have introduced all internal links between linked phrases, as this is a consistent alignment given the manual reference (see section 3.3 for details). In the case of 'date' expressions, we constrain their alignment to those having equal values (same day of the week and same day of the month).

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

After the initial phrase alignment, an one-to-one word alignment algorithm tries to link all unaligned tokens (including tokens inside unaligned detected phrases, which will be now considered separately). The details are presented in the following section.

2.2.3 Word alignment

As for the word alignment algorithm, we have implemented an iterative algorithm similar to the one presented in Cherry and Lin (2003a). 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 constraint (also called cohesion constraint (Cherry and Lin 2003b)) 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 contribute to the links 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 probability 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) and with the corpus used turned out to be optimal considering 8 words.

2.2.4 Postprocessing

As postprocessing, we have implemented three basic strategies based on linguistic criteria. First, we automatically align together those phrases selected during the first stage that have been left unaligned except for one of their tokens. This aligned token defines the alignment that is automatically generated for the other tokens of the phrase. The rationale behind this is to recover those detected phrases that did not have a good cooccurrence measure during the phrase alignment stage and were not linked, but got one token linked during posterior word alignment.

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

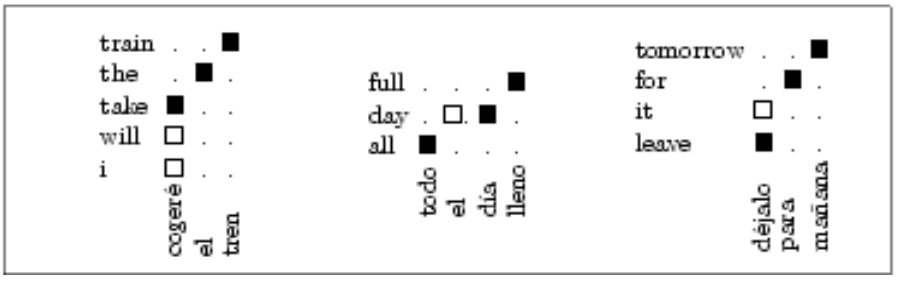


Fig. 3. Three linguistically-guided postprocessing strategies: completing selected phrases with only one token aligned (left), alignment of Spanish unaligned articles together with the noun they precede (center) and alignment of English personal pronouns following verbs linked to Spanish verbs with pronoun (right).

On the other hand, we also look for those Spanish articles preceding a noun which is linked to an English word, according to POS tags. For all these tokens, we automatically generate a link from the Spanish article to that English word, as these common Spanish articles are sometimes omitted in English. Finally, a similar approach is followed with English personal pronouns following a verb. If this verb is aligned to an Spanish verb form containing a pronoun, we conclude that the English pronoun must be aligned to the Spanish verb form as well. Figure 3 shows graphically one example of each of these postprocessing strategies (where black and white boxes represent links produced before and during postprocessing, respectively), whereas their impact is evaluated empirically in section 5.

The postprocessing stage could take other 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 could also feedback into the phrase selection/alignment blocks to reconsider previous decisions using global information of all sentences. Undoubtedly, this stage is strongly connected to the posterior translation model. Although alignment can be and must be evaluated separately, we are of the opinion that it is not completely independent from the translation model.

3 Experimental setting

In this section we present the English-Spanish parallel corpus used to evaluate the proposed phrase alignment strategy. Details on preprocessing performed and a discussion on the evaluation measure are provided.

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 appointment and meeting-planning domain. Table 2 shows the main statistics of the data used,

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

Table 2. *Statistics of the VerbMobil parallel corpus*

	Sentences	Words	Vocab.	Lmax	Lmean
English	30,054	228,328	3,276	66	7.6
Spanish		219,782	5,084	66	7.3

namely number of sentences, running words, size of vocabulary and maximum and mean sentence lengths for each language, respectively.

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 (Brants 2000), and base forms have been obtained using *wnmorph*, included in the WordNet package (Miller *et al.* 1991)
- Spanish data has been tagged using *FreeLing* package already mentioned. This software also generates a lemma or base form for each input word.
- Regular date and time expressions (which are numerous in the domain) have been substituted by a unified tag using a semi-automatic technique (de Gispert and Mariño 2003). Specifically, only dates containing a month of the year in both languages have been substituted, in contrast to date expressions dealt with in section 2.2.1.
- Finally, punctuation marks have been left out (as we expect to use the corpora for spoken language translation experiments).

3.3 Evaluation scheme

To assess the quality of the algorithm, we randomly selected from the corpus *two sets*: a validation set of 100 sentences (for tuning of parameters) and a test set of 400 sentences. For evaluation purposes, a manual gold standard alignment has been done in order to compute Alignment Error Rate (AER) as described in (Och and Ney 2000) and widely used in literature (including the aforementioned HLT-NAACL 2003 Shared task), together with appropiately redefined Recall and Precision measures. Mathematically, they can be expressed thus:

$$recall = \frac{|A \cap S|}{|S|}, \quad precision = \frac{|A \cap P|}{|A|}$$

$$AER = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

where A is the hypothesis alignment and S is the set of Sure (or unambiguous) links in the gold standard reference, and P includes the set of Possible (or ambiguous)

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

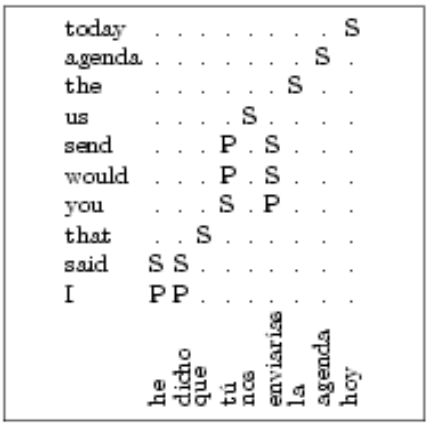


Fig. 4. Example of manual reference alignment (S for Sure links and P for Possible links).

and Sure links in the gold standard reference. If there is a possible link between two words in the manual reference, a link between these words is acceptable, but not compulsory (obtaining the same AER when generating or omitting it). On the contrary, if there is a sure link between two words in the reference, the link is compulsory (obtaining a worse AER when omitting it).

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 (Lambert and Castell 2004). 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, out of 100 links, 80 are Sure links and 20 are Possible links. Evaluations are performed without taking links to NULL into account.

Figure 4 shows an example of a manual reference alignment used in evaluation, where two alignments between verb phrases can be found ('I said - he dicho', 'you would send - tú enviarías'). As we are interested in aligning full verb phrases (with their pronoun, if any) as single meaning-bearing units during phrase alignment, the criterion used in reference has been to separate verb phrases into two parts, that is, personal pronoun (if any) and remaining tokens. These parts have been aligned with Sure links to their counterparts in the other language, whereas Possible links have been introduced between personal pronouns in one language and remaining tokens in the other (and vice versa). Introducing these latter links as Sure would favour our alignment result, since we produce all links between all tokens inside verb phrases. Therefore, the Possible links in the reference allow the generation of all internal links without falsely improving AER results with a misleading boost in Recall.

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

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

of the phrase selection and alignment stage previous to word alignment. Specifically, statistics on phrase selection and alignment are provided when dealing with verb phrases, idiomatic expressions and date expressions separately. Finally, complete alignment results are reported (phrase alignment + word alignment + postprocessing), comparing performance against state-of-the-art word alignment models.

4.1 Verb phrases

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

- 20,556 verb phrases in English (1,907 different), classified into 276 different verb lemmas,
- 2,007 verb phrases in Spanish (674 different), classified into 190 different verb lemmas.

In the case of English, 5% of detected phrases are 'gapped' in that they include adverbs modifying the verb, which will not be aligned together with the verb phrase (see section 2.2.1). Note that these figures include only phrases with more than one token (which are linked in the phrase alignment stage). For this reason, we get such a big difference between languages, given the usual omission of personal pronoun in Spanish by using inflected single-word verb forms. However, if we include also words with POS-tag of Verb that are substituted by their base form before word alignment (see section 2.2.1), we detect:

- 36,556 verb forms in English (2,584 different), classified into 482 different verb lemmas, and representing around 31% of all English tokens,
- 37,605 verb forms in Spanish (2,763 different), classified into 539 different verb lemmas, and representing around 20% of all Spanish tokens.

The results of the phrase alignment with only verb phrases for the development set are shown in the upper rows of Table 3, 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. The last two columns correspond to the percentage of verb phrases linked over the total number of phrases detected in the train set for English and Spanish, respectively. The margin of confidence at the 95% for Recall and Precision measure is 1.0%. as we do not compute cooccurrence counts for all words internal to the phrase, but just for the lemma of its head verb.

The selection strategy provides very encouraging results, as Precision is consistently nearly 99% for the three most-restrictive thresholds used, whereas Recall is just over 12% and nearly 87% of all selected blocks are being linked with a high precision. 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 single word cooccurrences. Given the significant descent in Precision without significant improvement in Recall

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

Table 3. *Separate phrase alignment results for development data set*

	Recall	Precision	LkEng	LkSpa
Verb phrases $\phi_v^2 < 5$	6.00	98.94	51%	40%
Verb phrases $\phi_v^2 < 10$	11.06	98.97	81%	77%
Verb phrases $\phi_v^2 < 15$	12.14	98.92	87%	86%
Verb phrases $\phi_v^2 < 20$	12.64	96.93	93%	91%
Idioms $\phi_v^2 < 5$	2.19	98.61	19%	47%
Idioms $\phi_i^2 < 10$	3.24	99.06	32%	77%
Idioms $\phi_i^2 < 15$	3.68	97.54	63%	84%
Idioms $\phi_v^2 < 20$	3.92	95.59	89%	90%
Date exprs. $\phi_i^2 < 5$*	4.11	98.82	88%	86%

setting the threshold to 20, a threshold of 15 has been selected as optimal for further experiments.

4.2 Idiomatic expressions

Regarding idiomatic expressions, the phrase selection stage detects the following cases by matching with the available lists presented in section 2.2.1:

- 525 idiomatic expressions in English (21 different), representing around 1% of all English tokens,
- 2,760 idiomatic expressions in Spanish (98 different), representing around 3% of all Spanish tokens.

The results when aligning only idiomatic expressions phrase-phrase and phrase-words links are shown in the middle rows of Table 3, again for different thresholds.

In this case, although the impact in Recall is much smaller than when considering verb phrases, two points are worth raising. First, we have again a nearly error-free alignment using a relatively small set of phrases (with thresholds up to 15). 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 phrases among the idiomatic expressions considered. Again, a threshold of 15 has been selected as optimal for further experiments.

4.3 Date expressions

Finally, regarding date expressions, which are all classified into one single class, the selection stage detects:

- 4,221 date expressions in English (336 different), representing around 5% of all English tokens,
- 4,303 date expressions in Spanish (410 different), representing around 5% of all Spanish tokens.

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

Table 4. *Comparison of final alignment results for test set*

	Recall	Precision	AER
giza++ eng2spa	76.70	92.90	15.78
giza++ spa2eng	78.52	94.02	14.14
giza++ union	**84.24**	90.64	12.52
giza++ intersection	70.98	**97.37**	17.78
one-to-one word align	72.78	95.85	17.15
full phrase align	80.75	96.37	**11.42**

The result when aligning only date expressions is shown in the lowest row of Table 3. This result does not change when increasing the threshold. As we only allow links between date expressions having the same 'value' (same day of the week and same day of the month), this result is constant once surpassed the cooccurrence measure between the tags 'date' of each language. Note that the 12% of expressions not linked correspond to not respecting the aforementioned constraint.

4.4 All phrases alignment

When aligning verb phrases, idiomatic expressions and date expressions before word alignment (using each optimal threshold), we achieve a **19.93%** of Recall with an outstanding **99.03%** Precision. This means that nearly 20% of the links are already done before proceeding to the statistical one-to-one word aligner. The complete alignment results are reported in the next section.

5 Complete alignment results and discussion

As baseline alignments for comparison, we have aligned our data using GIZA++ from English to Spanish and vice versa (performing five iterations of model IBM1 and HMM, and 3 iterations of models IBM3 and IBM4), and have evaluated two symmetrization strategies, namely the union and the intersection. Another symmetrization technique presented in Och and Ney (2000) has also been tested, without improvement over the union given this data and reference. The results are shown in the first four rows of Table 4. Note that the giza++ union rates the best among them, an expected result given the proportion of Sure and Possible links in the manual alignment (see section 3.3), which favours high-recall alignments. The margin of confidence at the 95% for Recall and Precision measures is 1.0%, and 0.8% for AER.

Complementarily, 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 8 to achieve high precision, always using cohesion constrain (initial NULL cost being set to 15). In contrast to giza++ intersection (the only baseline

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

Table 5. *Impact of postprocessing strategies in AER results*

	Recall	Precision	AER
phrase align, no postrocess.	78.71	97.20	12.35
postprocess. (3 strategies)	2.04	85.20	
postprocess. strategy 1	0.80	83.33	
postprocess. strategy 2	0.90	86.14	
postprocess. strategy 3	0.40	77.78	

alignment that is also one-to-one and thus subject to a fair comparison), we observe a reduction in Precision and an increase in Recall, leading to a non-significant AER reduction, i.e. comparable performance. However, both alignments skip around 30% of good links (far below the other alignments recall), which make them unpractical for posterior statistical translation modeling.

The results of our phrase-based word aligner are shown in the 'full phrase align' row (including phrase selection with each of the phrase optimal thresholds and postprocessing). As can be seen, it rates the best AER with 1.1 absolute points lower than the giza++ union (a relative 8.8% reduction). Our alignment achieves a very high Precision level (just one point lower than giza++ intersection), while providing a Recall increase of over 10 absolute points with respect to intersection, and of about 8 points with respect to the one-to-one aligner. The achieved Recall figure is still 3.5 points lower than that of the giza++ union (that nevertheless offers 6.3 points worse Precision results).

It is interesting to note the Precision improvement when comparing the word-to-word aligner with 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.

To analyse the impact of the postprocessing in this result, table 5 shows the AER without any postprocessing, which is not significantly lower than that of the giza++ union (with as high a Precision as the intersection). It is the addition of the three linguistically-guided postprocessing strategies presented in section 2.2.4 that provides a further boost to Recall for achieving the best AER.

To sum up, the presented alignment strategy improves state-of-the-art results while still making a relatively small use of linguistic knowledge. As its architecture is open to an easy introduction of more information, many other knowledge sources could be used. Furthermore, we performed the union between our results and the highest precision alignment, namely the giza++ intersection, obtaining a significant improvement in terms of Recall (over 3.5% absolute) and AER (1.3% absolute), as shown in Table 6. This result leads to two conclusions. On the one hand, the links produced by the algorithm proposed are complementary to those provided by the state-of-the-art statistical approaches, as linguistic knowledge plays a generalisation role where mere statistics are limited. And on the other hand, this means that there is

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

Table 6. *Results when combining two highest-precision alignments*

	Recall	Precision	AER
phr. align U giza++ intsct.	**84.36**	95.11	**10.10**

still room for improvement, which could be achieved by introducing more linguistic knowledge, as mentioned in the next section.

6 Conclusion and further research

In this paper, we have presented in detail a phrase alignment strategy combining linguistic knowledge and cooccurrence measures extracted from bilingual corpora. Following a different approach than state-of-the-art alignment algorithms, our proposal outperforms them in AER by introducing linguistic knowledge in an easy way. Specifically, simple detection rules for verb phrases, idiomatic expressions and date expressions in an English – Spanish task have been successfully introduced into the statistical framework, boosting recall at nearly no precision cost. Clearly, the cost of developing these basic rules is language-dependent, and though not relevant for the selected pair of languages, should be taken into account for more distant languages.

The algorithm proposed is the first step towards statistical machine translation with linguistic features, since it allows the system to learn the alignment with a certain knowledge that may allow a translation model to generalise the training data to all possible verb phrases, idiomatic expressions or date expressions during translation. In this sense, as current statistical machine translation decoders do not make use of such a knowledge, we plan to develop a linguistically-guided statistical decoder to evaluate the impact of these techniques in the translation task.

Future work also includes new and more linguistically-informed phrase selection schemes, such as searching for English phrasal verbs or compound nouns from available lists or using the output from named entities detectors and classificators during phrase selection and classification stages. Information from statistical chunkers and parsers could also be introduced to substitute current verb detection rules or include features based on chunks. Ontologies could also be included for dealing with synonymy and other semantic phenomena. However, the introduction of such data is only possible if these tools are available and with a similar performance level for both languages. 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 or ontologies must be provided.

Regarding the alignment algorithm, we plan to extend the use of base forms to nouns, adjectives and articles to tackle genre and number inflections and improve probability estimationsfreedom to reconsider phrase links, which are currently fixed (although their precision is very high). Finally, more postprocessing techniques will have to be tested, seeking a further boost in recall by using statistics of all aligned

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

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.

Acknowledgements

The authors want to thank Josep M. Crego, 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. This work has been partially funded by the Spanish Government under grant TIC2002-04447-C02 (ALIADO project), the Dep. of Universities, Research and Information Society (Generalitat de Catalunya) and the European Union under the integrated project TC-Star - Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738, http://www.tc-star.org)

References

Brants, T. (2000) TnT – a statistical part-of-speech tagger. *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Brown, P., Della Pietra, S., Della Pietra, V. and Mercer, R. (1993) The mathematics of statistical machine translation. *Computational Linguistics*, **19**(2): 263–311.

Brown, R. D. (1999). Adding linguistic knowledge to a lexical example-based translation system. *Proc. of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pp. 22–32.

Carreras, X., Chao, I., Padró, L. and Padró, M. (2004) Freeling: An open-source suite of language analyzers. *4th Int. Conf. on Language Resources and Evaluation, LREC'04*, pp. 239–242.

Cherry, C. and Lin, D. (2003a). A probability model to improve word alignment. *41st Annual Meeting of the Association for Computational Linguistics*, pp. 88–95.

Cherry, C. and Lin, D. (2003b). Word alignment with cohesion constraint. *Proc. of the Human Language Technology Conference, HLT-NAACL'2003*, pp. 49–51.

de Gispert, A. and Mariño, J. (2002) Using X-grams for speech-to-speech translation. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, pp. 1885–1888.

de Gispert, A. and Mariño, J. (2003) Experiments in word-ordering and morphological preprocessing for transducer-based statistical machine translation. *IEEE Automatic Speech Recognition and Understanding Workhsop, ASRU'03*, pp. 634–639.

Gale, W. and Church, K. (1991) Identifying word correspondences in parallel texts. *4th Speech and Natural Language Workshop*, pp. 152–157.

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

Lambert, P. and Castell, N. (2004) Alignment of parallel corpora exploiting asymmetrically aligned phrases. *Proc. of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*, pp. 26–29.

Marcu, D. and Wong, W. (2002) A phrase-based, joint probability model for statistical machine translation. *Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP'02*, pp. 133–139.

Melamed, D. (2001) *Empirical Methods for Exploiting Parallel Text*. MIT Press, Cambridge, MA.

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press

Mihalcea, R. and Pedersen, T. (2003) An evaluation exercise for word alignment. In: Mihalcea, R. and Pedersen, T., editors, *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pp. 1–10, Edmonton, Alberta, Canada.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. and Tengi, R. (1991) Five papers on wordnet. *Special Issue of International Journal of Lexicography*, **3**(4): 235–312.

Och, F. (2003) Giza++ software. http://www-i6.informatik.rwth-aachen.de/~och/ software/ giza++.html.

Och, F. and Ney, H. (2000) Improved statistical alignment models. *38th Annual Meeting of the Association for Computational Linguistics*, pp. 440–447.

Och, F. and Ney, H. (2003) A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1): 19–51.

Och, F., Tillmann, C. and Ney, H. (1999) 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.

Toutanova, K., Tolga Ilhan, H. and Manning, C. (2002) Extensions to hmm-based statistical word alignment models. *Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP'02*, pp. 87–94.

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

https://doi.org/10.1017/S1351324905003931 Published online by Cambridge University Press