Accurate phrase alignment in a bilingual corpus for EBMT systems

George Tambouratzis, Michalis Troullinos, Sokratis Sofianopoulos and Marina Vassiliou
Institute for Language and Speech Processing, Athena Research Center
6 Artemidos & Epidavrou Str., Paradissos Amaroussiou, 151 25, Athens, Greece
giorg_t@ilsp.gr; mtroullinos@ilsp.gr; s_sofian@ilsp.gr; mvas@ilsp.gr

Abstract

An ongoing trend in the creation of Machine Translation (MT) systems concerns the automatic extraction of information from large bilingual parallel corpora. As these corpora are expensive to create, the largest possible amount of information needs to be extracted in a consistent manner. The present article introduces a phrase alignment methodology for transferring structural information between languages using only a limited-size parallel corpus. This is used as a first processing stage to support a phrase-based MT system that can be readily ported to new language pairs. The essential language resources used in this MT system include a large monolingual corpus and a small parallel one. An analysis of different alignment cases is provided and the solutions chosen are described. In addition, the application of the system to different language pairs is reported and the results obtained are compared across language pairs to investigate the language-independent aspect of the proposed approach.

Keywords: phrase alignment, bilingual corpus, machine translation, EBMT systems,

1. Introduction

The current trend in MT systems is that of automatically extracting as much linguistic information as possible from corpora, either monolingual or bilingual ones. This applies to both Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT). The monolingual corpora are substantially easier to assemble, but cannot be used to create the translation models required by SMT systems, while bilingual corpora provide potentially more information and can be used to produce SMT translation models, but are more expensive to either collect from the web or create manually. The use, as far as possible, of monolingual rather than bilingual corpora can alleviate the need for expensive language resources. Hence, the motivation of the present article is to support the design of an MT system using as far as possible information extracted from monolingual corpora, while also maximising the utilisation of a small parallel corpus of a limited size (typically of a few hundred sentences).

The MT concept adopted here comprises a two-stage process. In the first stage, the structure of the sentence to be translated is transformed from the source language (SL) to the corresponding structure of the target language (TL), while in the second stage this structure is modified and enriched at a sub-sentential level to create the final translation. The entire process is data-driven and draws on linguistic information residing in two types of resources, namely (i) a limited-size bilingual corpus, the processing of which offers the essential information for transforming the SL sentence structure to the TL one, and (ii) a large monolingual corpus, compiled via web crawling, which is exploited in order to refine the translation at a sub-sentential level. This is in summary the concept of the PRESEMT project (www.presemt.eu), which aims to create an MT system that can be readily ported to new language pairs, using an EBMT-type approach.

The processing of this bilingual corpus to establish structural correspondences from the source to the target language is the main theme of the present article. In addition, it is useful to assess automatically the fidelity of translation from SL to TL for each sentence pair, so as to identify pairs where the match is not sufficiently accurate to provide information on the structure transformation from SL to TL.

In the remainder of this article, initially a survey of related research work is performed. This is followed by a description of the principles of the proposed approach. A detailed algorithmic description of the step-wise phrase alignment process is then provided. The required resources that have been assembled for experiments are subsequently presented, followed by the experimental results. This section includes an investigation of the approach accuracy when applied to different language pairs. In addition, an analysis of the source of errors (itemised in terms of the processing steps) is performed. Finally, potential extensions are investigated, such as the ability to assess the suitability of individual sentence pairs to serve as reference material for defining the TL structure, via the phrase aligner approach. This allows the creation of a more appropriate set of bilingual sentences.

2. Processing of bilingual corpora in MT

The majority of current MT systems, encompassing both statistical MT (SMT) and non-statistical MT systems, implement the translation of sentences by operating at sub-sentential level, for instance syntactic phrases, into which these sentence are split. In early SMT, the phrases were derived automatically based on sequences of tokens (Koehn, 2010). However, more recently, improvements are attained by introducing syntactically-valid phrasing. It has been found that the introduction of a parser in an SMT system enables the reordering of the SL side to better match the TL side of the corpus, thus conferring an improvement in translation quality (Collins et al., 2005).

A similar improvement in MT quality by introducing parsers has been identified in other MT paradigms such as EBMT systems, where sentences in SL are provided together with their reference translations in TL. However, in EBMT systems the definition of appropriate phrases necessitates either (i) the development of matched segmentations that give similar outputs for SL and TL or (ii) the definition of a mapping between SL and TL segmentation schemes. Both these approaches constrain the applicability of an MT system to language pairs for which the segmentation schemes are either directly compatible or are rendered compatible via additional processing (for example by generating transformation rules, mainly by trial-and-error, until a desired level of matching is achieved). A typical example of introducing phrasing in an EBMT approach is the METIS-II data-driven MT system (Markantonatou et al., 2006), where pre-existing parsing tools are employed for both the source and the target languages, but the tools' outputs are further processed to render them compatible. By definition, this heavily constrains the portability of an MT system to new language pairs, due to the need to ensure compatibility between the outputs of tools for different languages in advance.

An alternative solution, which is presented in this article, adopts a novel paradigm that circumvents this bottleneck of parsing scheme agreement, and thus can support the straightforward development of MT systems for new language pairs. This solution employs pattern recognition principles to create matching segmentations for the two languages, which then provide the basis for the transfer from the SL structure to the TL one. Relying on the use of a small bilingual corpus, which typically comprises a few hundred sentences aligned at sentence level, this approach is based on identifying sub-sentential segments in both SL and TL. Rather than trying to harmonise two already existing parsers, it uses a parser only in one language and maps this parsing information to the other language of a given language pair. In other words, given a parser (or more generally a phrasing model) in one of the two languages (either SL or TL), the aim is to generate an appropriate phrasing model for the other language. This is the main principle behind the PRESEMT approach (Tambouratzis et al., 2011). In the proposed implementation of the phrase alignment process, it is assumed that only a TL parser is available. The current work is based on the PAM approach proposed in Tambouratzis et al. (2011), though here the methodology has been extensively reworked to achieve a higher alignment accuracy coupled with enhanced language independence.

The process of defining SL-TL correspondences is achieved by grouping together SL elements (words) to sub-sentential segments (phrases) in accordance to the TL ones rendered by the parser. This approach exploits pattern recognition-based clustering techniques to extend these correspondences so that they cover the entire source language structure, dividing it into TL-based phrases.

3. Literature survey

A number of studies related to the phrase alignment approach proposed in this article have been carried out in the general field of linguistics, to determine the optimal alignment for bilingual corpora, by defining word phrases. A conceptually similar process to the one presented here has been proposed for parse trees by Yamada and Knight (2001), who assume a channel model. According to this model, during the machine translation process the segments (which are tree-based) are modified via three operations, namely reordering, insertion and translation. The information in this case is extracted via statistical methods.

Yarowski and Ngai (2001) propose projecting linguistic annotations from a resource-rich language to a resource-sparse one, in the case of parallel corpora of sentences. These projections are used to support the implementation of linguistic tasks in languages where the annotated material is sparse, via raw bilingual corpora which are automatically aligned. Yarowsky and Ngai (2001) have aimed at transferring shallow-processing tools such as noun phrase chunkers on the basis of word-level alignment between the languages.

The motivation of Tillmann (2003) is to determine blocks of corresponding words in the source and target languages that can then be used to perform statistical machine translation. This is achieved by a two-stage Viterbi-type approach which initially establishes high-precision alignments in terms of words that are in a second phase supplemented by incorporating lower-precision alignments to provide higher word coverage, thus generating blocks of words.

Och and Ney (2004) propose a data-driven approach that operates on corpora that are not linguistically-annotated to determine corresponding sequences of words. The definition of the sequences is performed via a two-stage process, where initially an alignment of words is performed and then aligned phrase pairs are extracted, employing a dynamic programming-type algorithm.

In contrast, Simard et al. (2005) propose a translation method using non-contiguous phrases, which is claimed to allow the coverage of additional linguistic phenomena in comparison to only allowing contiguous phrases. Ganchev et al. (2009) propose a methodology for inducing grammar knowledge for resource-poor languages. This methodology is based on bitexts between the resource-poor target language and a resource-rich language (such as English), where the resource-rich information is transferred to the resource-poor language. Ganchev et al. investigate the effect of introducing language-specific constraints for disambiguating annotation choices as compared to using only the bitext-based knowledge.

Melamed (1997) has studied the problem of correspondence of words in different languages with the aim of estimating a partial translation model that accounts for translational equivalence, only at a word level, based on word co-occurrences. Taskar et al. (2005) have proposed a discriminative method for defining word alignment models based on a selection of features of word pairs and compared this method to statistics-based models such as Giza++. Finally, DeNero et al. (2007) propose an alignment approach aimed to support syntactic machine translation, using HMM modelling.

An alternative approach for identifying corresponding words has been proposed for EBMT as the Marker Hypothesis. In this hypothesis, specific words are used for signalling phrase boundaries in both the SL and TL (see for instance Gough and Way, 2004). This approach however presupposes the compilation of marker word lists per language; besides, in the approach proposed in the present article, the SL text segmentation is guided by the TL text parsing scheme.

4. Extracting alignments from a bilingual corpus

The methodology proposed here, henceforth called Phrase aligner (PA), aims at extracting phrasal information via mutual alignment of the SL sentences and the TL ones of a parallel corpus. The Phrase aligner requires only one side of the parallel corpus to contain phrasing information that will be provided by an appropriate parser, while the other side only contains lemma and Part-of-Speech (PoS) tag information. By performing word alignments between the sentences of the parallel corpus and clustering all words into phrases based on the phrases found on the parsed side of the corpus, the Phrase aligner effectively extracts a phrasing scheme for the corpus side that has no phrasing information, on the condition that the given phrases in the two languages do not overlap. The extracted alignment information is then exploited to (a) create a phrasing model that can be applied for processing any input sentences for the parser-less language side and (b) create an SL-TL model for structural reordering during the machine translation process.

4.1 Design of the PA algorithm

The Phrase aligner needs three resources, namely an SL-TL bilingual lexicon, a tagger and lemmatiser for both the SL and TL sides of the corpus and a TL parser for yielding the appropriate phrasing scheme. Based on these resources, the following information is available:

- Likely SL-TL word correspondences, as furnished by the bilingual lexicon. These correspondences may be
  - one-to-one (a single SL word translates into exactly a single TL word)
  - one-to-many (a single SL word corresponds to a multi-word TL unit)
  - many-to-one (an SL multi-word unit corresponds to a TL single one)
- SL-to-TL tag correspondence; for languages with rich morphology, possibly additional morphological information.
- In-sentence distances between two words, measured in terms of the number of intervening tokens.
- Decomposition of the TL sentence in sub-sentential segments depending on the parser employed.

Based on this set of inputs, PA needs to decide on the optimal segmentation of the source sentence into phrases. A multi-criterion-type comparison must be performed, where the different inputs are accordingly prioritised and combined. Naturally, not all aforementioned inputs need to be present for the PA to generate results, though use of all inputs yields a more accurate alignment.

4.2 Implementation of the PA algorithm

Similarly to several of the aforementioned systems (cf. Och and Ney, 2004; Ganchev et al., 2009), PA employs a multi-stage process, according to which the establishment of word correspondences is performed in the first stage, and these correspondences are then extended in subsequent stages to eventually cover the entire sentence. More specifically, a three-stage process is implemented, where (i) SL-TL word correspondences are established based on the lexicon, (ii) alignments exploit the similarity of grammatical features and (iii) SL words aligned within the first two stages are used as the nuclei of phrases to which still unaligned SL words are assigned. Each of the three stages is described in detail below.

Stage 1: Alignment of single words

The word aligner algorithm performs alignment of SL words to TL ones based on the information of the bilingual lexicon. It is often the case that SL words have more than one candidate translations. So, let us assume that a given SL word ‘*A*’ has two candidate translations, ‘*B*’ and ‘*C*’, in the bilingual lexicon. If in the TL side of the sentence pair both ‘*B*’ and ‘*C*’ exist, then this multiple word alignment cannot be resolved without additional information, such as, for instance, the information residing in the neighbourhood of words ‘*A*’, ‘*B*’ and ‘*C*’ in the SL and TL sentences.

Figure 1 illustrates an example of such a case, where the SL side comprises four words, denoted ‘*SL1’* to ‘*SL4’*, and the TL side comprises four words denoted ‘*TL1’* to ‘*TL4’*. According to the lexicon, words ‘*SL1’* and ‘*SL4’* have each a single candidate translation (words ‘*TL3’* and ‘*TL4’* respectively); but the word ‘*SL2’* has two candidate translations in the TL sentence, namely ‘*TL1’* and ‘*TL2’*. Exploiting information on the environment of ‘*TL1’* and ‘*TL2’* to choose between the two candidate translations, a distance-based principle is used to determine the TL word (either ‘*TL3’* or ‘*TL4’*) to which an SL word within the vicinity of ‘*SL2’* is single-aligned and which has a minimum distance from one of the candidate words. In this example, the two distances corresponding to single-aligned words are *dis(SL2,TL1)* and *dis(SL2,TL2)*. Hence, the distance between the SL side and the TL side is expressed as the distance of the candidate translations (‘*TL1’* and ‘*TL2’)* from those TL words, to which other SL words, within a given neighbourhood to the SL word in question (‘*SL2’*), have already been single-aligned.

In the example of Figure 1, if a neighbourhood size of 1 is used, then only one neighbouring word, namely ‘*SL1’*, is single-aligned, to ‘*TL3’*. Since ‘*TL3’* is situated closer to ‘*TL2’* than to ‘*TL1’*, then ‘*TL2’* will be chosen as the most likely translation of ‘*SL2’*.

If a neighbourhood size of 2 is used, two neighbouring words are single-aligned, namely ‘*SL1’* and ‘*SL4’*, which translate into ‘*TL3’* and ‘*TL4’* respectively). In that case, the choice will be based on the smallest mean distance of the two candidate translations, ‘*TL1’* and ‘*TL2’* from the two reference points *‘TL3’* and *‘TL4’*. The computed distances are as follows:

dis(SL2,TL1)=[dis(TL3,TL1)+dis(TL4,TL1)]/2= [2 + 3]/2 = 2.5

dis(SL2,TL2)=[dis(TL3,TL2)+dis(TL4,TL2)]/2= [1 + 2]/2 = 1.5

Thus, based on the principle of smallest distance, word ‘SL2’ will again be chosen as the most likely translation of *‘TL2’*.

In the general case, for an assignment to be made, this cumulative distance must be below a given threshold, which is a system parameter, so as to avoid aligning words at a large distance to each other.

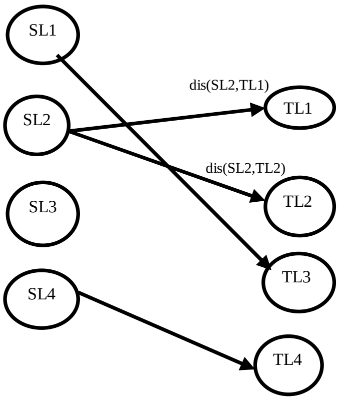

Figure 1: Example of multiple alignments

A similar process is followed in the case of multiple words from the SL side being translated to a TL single word. In this case, a mirror-application of the algorithm is performed, with words in the environment of the SL side being used to establish the minimum distance solution.

Naturally, within a sentence several multiple alignments may exist; their resolution is carried out in the first stage of PA so as to minimise the mean value of distances for all words being examined. In addition, a necessary property is that of independence to the order with which the sentence words are processed. To that end, all decisions aimed at resolving (some of) the multiple alignments are performed while ensuring that the collective distances for the entire sentence are minimised.

Given that (i) a single application of the algorithm will very likely not resolve all ambiguities within a sentence and (ii) the resolution of certain multiple alignments can facilitate the resolution of other pending ones, this algorithm is applied iteratively on a sentence basis, until there exist no further multiple alignments.

An example of a more complex situation is depicted in Figure 2 (distances between TL and SL elements are indicated on the relevant edges, while already aligned words are not shown in order to simplify the figure). There are two SL words, for each of which multiple possible alignments exist, and these alignments overlap. If it is attempted to resolve first the multiple alignment of ‘*SL1’*, the achievement of a global minimum cannot be guaranteed. On the contrary, by examining word ‘*SL2’*, it can be seen that ‘*TL3’* is at a smaller distance than ‘*TL2’*, and that this is the lowest global distance. By removing the possible association between ‘*SL2’* and ‘*TL2’* (as ‘*SL2’* has already been aligned to ‘*TL3’*), there remain two candidates for ‘*SL1’*, namely ‘*TL1’* and ‘*TL2’*. Thus, by examining in the second iteration their relative distances, it can be seen that ‘*TL2’* is a preferable alignment to ‘*TL1’*. Consequently, in a total of two iterations the entire sentence is disambiguated.

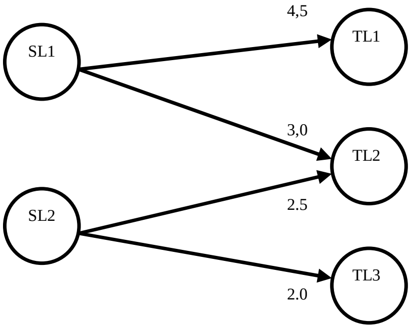

Figure 2: Example of resolvable multiple alignments needing more than one iterations to be resolved

A different situation is depicted in Figure 3. More specifically, though the number of words and of multiple alignments is exactly the same, the relevant distances differ. So, though the first iteration will again assign ‘*SL2’* to ‘*TL3’*, the second iteration cannot decide on a TL word to which word ‘*SL1’* should be assigned. This illustrates the effect of the relevant magnitude of distances on the disambiguation process. To avoid reaching sub-optimal solutions it has been decided not to force the resolution of such cases in stage 1, but re-examine candidate solutions at later stages.

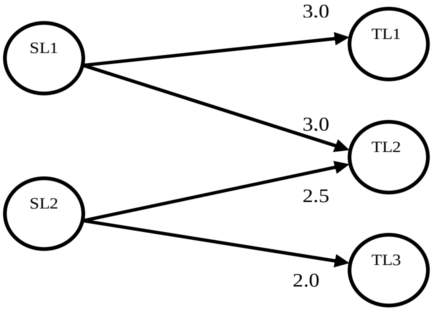

Figure 3: Example of non-fully resolvable multiple alignments needing more than one iterations

These examples illustrate the approaches that the PA employs in order to resolve as many as possible multiple alignments provided by the lexicon. The limited coverage of the lexicon is overcome through two language-independent mechanisms:

(i) Matching of numeric words, when their actual strings match. As this mechanism is almost certain to lead to the correct assignment, its application precedes accessing the lexicon.

(ii) Transliteration process to a common character set, when SL and TL differ in terms of alphabet (for instance English and Greek). A comparison between the transliterated words of the SL and TL sides is performed to map so far unmatched words, provided their transliterations have a similarity exceeding a given threshold. This operation is applied at the very end of stage 1, after all lexicon-extracted information has been used. This allows the similarity threshold to be set to a lower value without affecting the output of the lexicon-matching process.

At the end of Step 1, alignments using single-word information are resolved to the greatest extent possible. Any words that cannot be unambiguously aligned are forwarded to the next two stages for resolution.

Stage 2: Alignment based on feature similarity

Stage 2 processes the output of Stage 1, with the aim of increasing the percentage of words aligned between the SL and TL sentences. In this stage, the resolution of so far unassigned SL words is based on similarity of grammatical features (e.g. case, number etc.), to be found in the extended PoS tags. Hence, the extended tags of still unassigned SL words are matched to those of other SL words that have been unambiguously aligned in the previous stage. Amongst these matches, the one with the highest similarity is selected, since that indicates a high likelihood of association between the matched words. The tag similarity is normalised by multiplying with a Gaussian function that takes as its input the distance in terms of tokens of the two words on the sentence. Consequently, the tag similarity is reduced as the physical distance in the sentence increases. This normalisation allows the assignment of SL words to the same phrase, provided that they match to an acceptable extent in terms of grammatical features but are also relatively closely situated within the sentence. The variance of the Gaussian function is tuneable to the application requirements.

The aforementioned algorithm is effective only for inflected words such as verbs, nouns, adjectives, pronouns and yields good results in the case of morphologically rich languages. However, it can still be applied in morphologically poor languages without loss of generality, though naturally the number of words aligned by it will be limited.

The present phrase alignment stage is aimed to maximise the coverage and accuracy of word alignments. Hence, an additional effort involves aligning yet-to-be-assigned SL words to TL ones based on the inter-language tag correspondence. This type of information is of a statistical nature and is extracted in an unsupervised manner from the bilingual lexicon by studying macroscopically the average frequency with which any SL word of PoS type ‘X’ is translated to an also unaligned TL word of PoS type ‘Y’. Assuming that the majority (exceeding a chosen threshold) of words of PoS type ‘X’ do translate into words of PoS type ‘Y’, then an unaligned SL word of PoS type ‘X’ could be assigned to a TL word of PoS type ‘Y’ to improve the phrase aligner coverage. If there exist more than one TL words of PoS type ‘Y’, the most likely one can be determined by applying the neighbourhood-based principle, as described in Stage 1.

Stage 3: Alignment based on neighbourhood

Stage 3 operates on the output of Stage 2, with the aim of grouping the residual unaligned SL words to TL phrases. This is achieved via two methods. In the first method, grammatical feature similarity is taken into account, as introduced in stage 2, the difference being that at this third stage the principle of normalising over the distance applies to TL phrases instead of TL words. The second method forces an unaligned SL word to be assigned to the TL phrase to which the majority of its SL side immediate neighbours belong.

5. Experimental setup

Since the PA methodology is language-independent, the Phrase aligner module has been tested so far on three language pairs, Greek – English and German – English and English – German, all of which involve languages with a different word order (English has a fixed word order, Greek has a free word order, while German is a V2 language). In the present article, the experiments on the first two pairs are reported. For each pair a bilingual parallel corpus has been built from the web. For both the SL and TL sides the corpus has been processed using readily available language tools as detailed below.

The SL side of the corpus is then manually edited so that it would be “close” to the TL one, removing metaphors or elliptical constructions and smoothing out divergences between the two languages. Moreover, for the reported experiments, the corpus NLP annotations have been manually corrected, so as to focus on testing the PA performance on data devoid of errors. Future experiments will study the effect of the actual annotations (which will unavoidably contain errors) on the performance of the phrase aligner.

Greek - English corpus: Extracted from a multilingual website[1], this corpus comprises 200 sentences. The SL side of the corpus has been tagged and lemmatised by the FBT Tagger-Lemmatiser (Papageorgiou et al., 2000), while the TL side has been processed with the TreeTagger for English (Schmid 1994), yielding tag, lemma and phrase annotations.

German - English corpus: Also extracted from a multilingual website[2], it comprises 164 sentences. The SL side of the corpus has been tagged and lemmatised by the TreeTagger and the RFTagger (Schmid and Laws, 2008), while the TL side has been processed with the TreeTagger for English, generating tag, lemma and phrase annotations.

5.1 Experimental results

For assessing the segmentation accuracy obtained by the

[1] http://europa.eu/abc/history/index_en.htm
[2] http://europa.eu/abc/12lessons/index_en.htm

phrase aligner, its output was compared with a gold-standard reference set. This set included all SL sentences of the aforementioned corpora manually segmented into phrases in accordance to the TL side phrasal segmentation. In other words, the SL side was segmented in those phrases, which PA was expected to generate.

For the purposes of the experiment, two gold-standard sets have been created, of 50 sentences each, for the Greek – English corpus (EL-EN), and two sets, of 50 sentences each, for the German – English (DE-EN) corpus. The degree of match of the PA result to the gold-standard for both language pairs is reported in Table 1, where the best accuracies are denoted in boldface.

Different configurations have been examined, using different values for system parameters. Among the system parameters used, the configurations reported here vary in terms of only certain parameters to which the system is more sensitive, namely (i) the maximum distance for a single alignment to be made, (ii) the minimum required transliteration similarity, (iii) the minimum extended tag similarity threshold, and (iv) the minimum required number of lexicon entries of a given SL tag for which the most likely TL tag is defined in the latter part of Stage 2. The values of these parameters are listed in Table 2 for a number of experimental configurations.

Configuration	Accuracy			
	EL-EN Set1	EL-EN Set2	DE-EN Set1	DE-EN Set2
A	93.74	91.64	**88.50**	88.96
B	**94.51**	92.16	88.23	88.11
C	**94.51**	**93.09**	88.23	88.11
D	94.38	92.28	88.49	89.46
E	94.32	**93.09**	87.92	**90.09**

Table 1: PA experimental results for the EL-EN and DE-EN corpora with variant configurations

System Parameters	Configuration				
	A	B	C	D	E
Distance threshold	3.0	2.0	2.0	3.0	2.0
Translit. similarity threshold	0.5	1.0	1.0	0.5	0.5
Extended tag threshold	0.1	0.5	0.5	0.5	0.5
Threshold of lexicon entries per SL tag	100	0	100	10000	10000

Table 2: Configurations tested for the system parameters

A first observation is that the results across the two sets for each language pair are very similar, indicating that the PA behaviour can be expected to be consistent over a variety of texts. Furthermore, all tested configurations of parameter values (the configurations reported in Table 2 are the more effective ones out of the set examined) give rise to similar results, with a deviation of less than 2% in terms of accuracy.

Another observation concerns the actual accuracy of the phrase aligner. This averages over 94.5% in the case of the Greek – English language pair over a given set of sentences. Since certain sentences give very low alignment accuracies, the actual accuracy over the ‘better’ sentences is even higher. So, if the sentences to be aligned and then used in the translation process are filtered in advance to remove those with low alignment accuracy, the collective alignment can be substantially higher.

In the case of German – English, the peak accuracy is just over 90%. This is lower than the accuracy reported for the Greek – English pair but still represents a high accuracy. The reduced accuracy for German – English can be mainly attributed to the more complex alignments involved due to the very productive compounding mechanism of the German language, which increases the difficulty of identifying word-to-word alignments.

5.2 Studying the system performance

By analysing the system operation, it is possible to determine which stages are the more effective ones, and which may provide the basis for further improvement. The results summarised in Table 3 are yielded by the optimal configuration (configuration ‘C’) for the Greek – English corpus; those in Table 4 derive from the same configuration, when applied to the German – English corpus.

In both cases, the accuracy reported is calculated over the entire set of 100 sentences for which gold-standard phrases have been defined.

Greek – English			
	Erroneous alignments	Correct alignments	Accuracy
Stage 1	29	1198	97.6%
Stage 2	15	134	89.9%
Stage 3	61	324	84.2%
Total	**105**	**1656**	**94.0%**

Table 3: Accuracy of each stage of the alignment process for the EL-EN corpus

German – English			
	Erroneous alignments	Correct alignments	Accuracy
Stage 1	82	1601	95.1%
Stage 2	5	13	72.2%
Stage 3	191	325	63.0%
Total	**278**	**1939**	**87.5%**

Table 4: Accuracy of each stage of the alignment process for the DE-EN corpus

According to Tables 3 and 4, as the PA operation proceeds from stage 1 to stage 3, the alignment accuracy decreases in each subsequent stage. This is expected, as in each stage, less reliable information is employed to perform the alignments, in order to improve the coverage in terms of aligned words. However, for the given phrase alignment result to be useful in the MT process, it is essential to achieve a full coverage of the SL sentences and to this end all stages must be applied.

6. Comparison to Existing Methods

Comparative experiments have been performed in order to obtain a better perspective of the accuracy achieved by the Phrase aligner. GIZA++ was used as a baseline to perform alignments between the SL and TL phrases of the corresponding sentences in the bilingual corpora that PA has been developed on (even though it should be mentioned that GIZA++ is not primarily designed for such a task). The comparison results (cf. Table 5) are promising, as, for both Greek – English and German – English corpora, the accuracy attained by PA is substantially higher than that of GIZA++.

Comparison to Baseline	Corpus	GIZA++
Precision	EL-EN	**72.21%**
Recall		60.98%
Precision	DE-EN	**74.64%**
Recall		71.01%

Table 5: Giza-based experimental results

7. Evaluating sentence pairs' suitability

In the PRESEMT architecture, the limited-size parallel corpus determines the structure of the translation. As the creation of a parallel corpus is a labour-intensive process, it is essential to be able to determine the level of direct correspondence between the SL and TL sides. As described before, alignments are performed in three distinct stages, with each subsequent stage having a lower dependability than previous ones. Consequently, by measuring the percentage of words aligned after each stage for each sentence pair, an estimate of the sentence pair dependability is provided. This can then be used to filter out corpus sentence pairs with a low correspondence between SL and TL, this being reflected by the resolution of alignments for many sentence words in later stages (for instance stage 3). Of course, this estimate also depends on the coverage of the bilingual lexicon used, which can affect the accuracy of the given sentence pair alignments.

8. Further Extensions

In this article, a phrase alignment approach has been presented which generalises the phrasing scheme drawn from the parsed TL side of a bilingual corpus to the non-segmented SL side. This approach is used as a first processing stage to support a phrase-based MT system that is readily portable to new language pairs. A detailed analysis of alignment phenomena, coupled with the application of the system to different language pairs indicate the language independence of the proposed approach.

Within the next period, it is aimed to integrate this mechanism to the PRESEMT system in order to investigate the effectiveness of the approach.

Algorithm-specific improvements possibly entail the refinement of the distance definition, in order to take into account the phrase boundaries when identifying the limits of a word environment. Besides, it is planned to apply the algorithm to more language pairs, including Greek-to-German and English-to-German, with the aim of gaining further insight with respect to the characteristics of the proposed approach.

Up to date, the developed MT language pairs in PRESEMT have been based on the use of parallel corpora. In the following period, it is intended to employ SL-TL comparable corpora, with the aim of evaluating the PA performance on non-strictly parallel corpora and the consequent effect on the performance of the PRESEMT system. Provided the translation accuracy is of a sufficient level, this may allow the simpler development of new language pairs, potentially reducing the effort required for generating high-quality parallel corpora.

Upon completion, the phrase aligner will also be released as public software, available to be incorporated in other applications, with the expectation that it will be of interest and of benefit to the wider research community.

9. References

DeNero, J. and Klein, D. (2007). Tailoring Word Alignments to Syntactic Machine Translation. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, June 2007, pp. 17--24,.

Clause Restructuring for Statistical Machine Translation (2005) Collins, M., Koehn, P., Kucerova, I. (2005). Proceedings of the 43rd Annual Meeting of the Association for Computational Linguists (ACL-05), Ann Arbor, USA, June 2005, pp. 531--540.

Ganchev, K., Gillenwater, J., and Taskar, B. (2009). Dependency Grammar Induction via Bitext Projection Constraints. Proceedings of the 47th Annual Meeting of the ACL, Singapore, 2-7 August 2009, pp. 369--377.

Gough, N. and Way, A. (2004). Robust Large-Scale EBMT with Marker-Based Segmentation. Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04), pp. 95--104.

Koehn, P. (2010). Statistical Machine Translation. Cambridge University Press, Cambridge.

Markantonatou, S., Sofianopoulos, S., Spilioti, V., Tambouratzis, G., Vassiliou, M. and Yannoutsou, O. (2006). Using patterns for machine translation (MT). Proceedings of the 11th annual Conference of the European Association for Machine Translation. Oslo, Norway, pp. 239-246.

Melamed, D. (1997). A Word-to-Word Model of Translational Equivalence, Proceedings of the 35th Conference of the Association for Computational Linguistics, Madrid, Spain, pp. 490--497.

Och, F.J., and Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, 30(4), pp. 417--449.

Papageorgiou, H., Prokopidis, P., Giouli, V., and Piperidis, S. (2000). A Unified POS Tagging Architecture and its Application for Greek. LREC-2000 Conference Proceedings, Athens, Greece, pp. 1455--1462.

Schmid, H., and Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained PoS Tagging. Proceedings of COLING 2008, Manchester, Great Britain, pp. 777--784.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Simard, M., Cancedda, N., Cavestro, B., Dymetman, M., Gaussier, E., Goutte, C., Yamada, K., Langlais, P. and Mauser, A. (2005). Translating with Non-Contiguous Phrases. Proceedings of the Conferences on Human Language Technology and on Empirical Methods in Language Processing, Vancouver, Canada, pp. 755--762.

Tambouratzis, G., Simistira, F., Sofianopoulos, S., Tsimboukakis, N. and Vassiliou, M. (2011). A resource-light phrase scheme for language-portable MT, Proceedings of the 15th International Conference of the European Association for Machine Translation, (eds. M. L. Forcada, H. Depraetere and V. Vandeghinste) 30-31 May 2011, Leuven, Belgium, pp. 185--192.

Taskar, B., Lacoste-Julien, S. and Klein, D. (2005). A Discriminative Matching Approach to Word Alignment. Proceedings of the HLT/EMNLP Conference, Vancouver, October 2005, pp. 73--80.

Tillmann, C. (2003). A Projection Extension Algorithm for Statistical Machine Translation. Proceedings of the EMNLP Conference, pp. 1--8.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. Proceedings of the 39th Annual ACL Meeting, July 9-11, Toulouse, France, pp. 523--530.

Yarowsky, D. and Ngai, G. (2001). Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. Proceedings of NAACL-2001 Conference, pp. 200--207.

10. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 248307.