A resource-light phrase scheme for language-portable MT

George Tambouratzis
Inst. for Language & Speech Processing /
6 Artemidos & Epidavrou Str.,
Maroussi,151 25, Athens, Greece
giorg_t@ilsp.gr

Fotini Simistira
Inst. for Language & Speech Processing / 6 Artemidos & Epidavrou Str.,
Maroussi,151 25, Athens, Greece
fotini@ilsp.gr

Sokratis Sofianopoulos
Inst. for Language & Speech Processing /
6 Artemidos & Epidavrou Str.,
Maroussi,151 25, Athens, Greece
s_sofian@ilsp.gr

Nikos Tsimboukakis
Inst. for Language & Speech Processing / 6 Artemidos & Epidavrou Str.,
Maroussi,151 25, Athens, Greece
ntsimb@ilsp.gr

Marina Vassiliou
Inst. for Language & Speech Processing /
6 Artemidos & Epidavrou Str.,
Maroussi,151 25, Athens, Greece
mvas@ilsp.gr

Abstract

The present article introduces a phrase-alignment approach that involves the processing of a small bilingual corpus in order to extract suitable structural information. This is used in the PRESEMT project, whose aim is the quick development of phrase-based Machine Translation (MT) systems for new language pairs. A main bottleneck of such systems is the need to create compatible parsing schemes in the source and target languages. This bottleneck is overcome by combining two modules, the Phrase aligner module and the Phrasing model generator, both of them being based on pattern recognition principles.

1 Introduction - Summary of the PRESEMT MT approach

A large proportion of current Machine Translation (MT) systems translate sentences by operating at a sub-sentential level. However, this necessitates either (i) the development of matched segmentations that give similar outputs for the source and target languages (SL and TL) or (ii) the definition of a mapping between two given segmentations. Both these approaches constrain the MT system to language pairs for which suitable segmentation schemes exist, which either are or have been made compatible via additional processing. A typical example of such an approach is the METIS-II data-driven MT system (Markantonatou et al., 2006 and Carl et al., 2008), where the parsing tools are used for both the source and the target languages and are accordingly modified. This naturally limits the MT system portability to new languages, due to the need for developing and/or modifying the appropriate tools for generating the segmentation scheme.

The PRESEMT project investigates a novel paradigm, which circumvents this bottleneck, and supports the straightforward development of MT systems for new language pairs, using pattern recognition principles. Relying on the use of a large TL monolingual corpus and a small bilingual corpus, which typically comprises a few hundred sentences aligned at sentence level, PRESEMT is based on handling sub-sentential segments. It uses a parser only in one language and maps this 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), one

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 185–192
Leuven, Belgium, May 2011

can generate an appropriate phrasing model in the other language.

This approach supports the rapid creation of an MT system by using a bilingual parallel corpus to learn structural correspondences between source and target languages. This is achieved by grouping together elements (tokens) in the source and target languages, in order to create sub-sentential segments (phrases) which correspond to one another based on the structure of the parallel sentence. This approach exploits pattern-recognition-based clustering techniques.

The PRESEMT system entails a two-phase translation process. With respect to resources, the system draws on two different sources of linguistic content, these being a large TL monolingual corpus as well as a small parallel corpus. These are collected from the web, using as far as possible automated methods, to minimise the effort needed to create a new language pair.

1.1 General concept

The general architecture of the PRESEMT system is depicted in Figure 1. In the first phase of the translation process the effort focuses on defining the sentence structure in terms of sub-sentential segments (phrases), which do not necessarily coincide with syntactically-defined phrases, but are linguistically-motivated. To achieve this objective, only the small parallel corpus (containing a set numbering a few hundred sentences) is processed to detect the most similar sentence to the given SL one, following the assumption of a common underlying language structure. The corresponding structure of this sentence in the target language is considered the translation structure.

The defined structure is then handed over to the second phase, where micro-structural processing, involving disambiguation of multiple translations and establishment of word order within segments, takes place. This processing is based on the semantic-type and statistical information derived from the TL monolingual corpus.

So, the PRESEMT system is based on a learn-by-example approach encompassing pattern recognition principles. In this respect, it is closely related to the Example-Based Machine Translation (EBMT) family of MT systems.

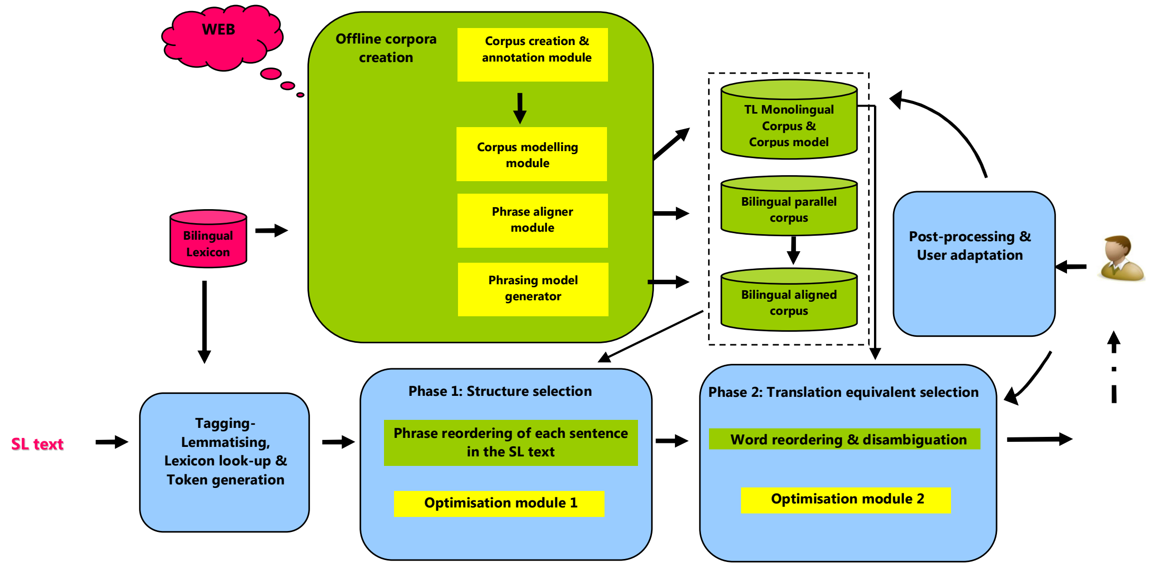
WEB
Offline corpora creation
Corpus creation & annotation module
Corpus modelling module
Phrase aligner module
Phrasing model generator
Bilingual Lexicon
TL Monolingual Corpus & Corpus model
Bilingual parallel corpus
Bilingual aligned corpus
Post-processing & User adaptation
SL text
Tagging-Lemmatising, Lexicon look-up & Token generation
Phase 1: Structure selection
Phrase reordering of each sentence in the SL text
Optimisation module 1
Phase 2: Translation equivalent selection
Word reordering & disambiguation
Optimisation module 2

Figure 1. Architecture of the PRESEMT system

2 Eliciting information from a bilingual corpus

The bilingual corpus processing and the extraction of the corresponding information involves two stages, the **Phrase aligner module** (**PAM**), which performs text alignment at word and phrase level within a language pair, and the **Phrasing model generator** (**PMG**), which elicits a phrasing model for a given language and applies it to input sentences. The present section provides a brief account of the two modules.

Phrase aligner module: It performs offline SL – TL word and phrase alignment within a bilingual corpus of parallel sentences. Intended to serve as a language-independent method for mapping corresponding terms within a language pair, it circumvents the problem of achieving compatibility between the outputs of different parsers for a language pair, by relying on a parser for one language and automatically obtaining an appropriate phrasing model for the other. PAM takes into account the parsing information in one language (in the present implementation the TL one) and makes use of a bilingual lexicon. The output is the bilingual corpus aligned at word, phrase and clause level, handed over to PMG for further processing.

Phrasing model generator: The Phrasing model generator has two modes of operation. First, it receives the PAM output and generates a probabilistic phrasing model for one side of a given language pair (in the present implementation the SL side). Second, this phrasing model is applied for segmenting SL text being input to the PRESEMT system for translation.

In the former operation mode, PMG works offline on aligned SL-TL sentence pairs in order to extract the phrasing model, while in the latter operation mode it utilises this model online. The PMG-processed SL text is then forwarded to the PRESEMT main translation engine.

3 Related work – Literature survey

3.1 Initial motivation for the Phrase aligner module

The Phrase aligner module was inspired by the work of Zakarian (2008) on generalised clustering methods. Zakarian's work focuses on the optimisation of automotive production lines, by representing the different parts to be assembled in a two-dimensional matrix. The iterative approach proposed continuously improves the quality of the solution, by defining increasingly larger non-overlapping groups on rectangular two-dimensional sub-matrices.

To draw an analogy with Zakarian's method, in the PRESEMT Phrase aligner the aim is to define appropriate segments in the sentence that can be largely translated independently, prior to being combined to provide the final translation (this is similar in concept to the recombination stage of EBMT systems).

However, problems have been encountered in mapping Zakarian's method to the Phrase aligner task, since in language the mapping of tokens is not necessarily one-to-one, but one SL token may be assigned to multiple TL tokens. In addition, it is difficult to combine multiple criteria for clustering sentence tokens into a single distance measurement for each element in the 2-D matrix.

3.2 Studies relevant to the Phrase aligner module

Several studies conceptually related to the Phrase aligner have been carried out in the field of linguistics in the past decade, to define the optimal alignment for bilingual corpora, to support the statistical MT, by defining word phrases. A similar process to the Phrase aligner, though based on different principles, has been proposed for parse trees by Yamada and Knight (2001).

Yarowski and Ngai (2001) proposed projecting linguistic annotations from a resource-rich language to a resource-sparse one, in the case of parallel corpora of sentences. These projections are aimed to support linguistic tasks in languages where the annotated material is sparse, using automatically word-aligned raw bilingual corpora to project annotations. Yarowsky and Ngai (2001) have proposed the creation of an NP-bracketer, which represents the first step towards the creation of a parser for resource-poor languages. Yarowsky and Ngai (2001) aimed at transferring shallow-processing tools such as noun phrase chunkers, being based on word-level alignment between the languages.

The motivation of Tillmann (2003) is to determine blocks of corresponding words in the two languages. A Viterbi-type approach is employed to determine high-precision alignments, which are then expanded to provide a higher recall by incorporating lower-precision alignments via dynamic-programming beam-search.

Och and Ney (2004) define consecutive sequences of words that do not necessarily correspond to linguistic phrases, using phrase-based dictionaries without linguistically-annotated corpora. A two-stage process is adopted, 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, in order to allow the coverage of additional linguistic phenomena. Hwa et al. (2005) propose creating a parser for a new language based on a set of parallel sentences coupled with a parser in a frequently-used language, transferring deeper syntactic structure and introducing fix-up rules to improve the chunking accuracy. Based on Hwa et al. (2005), Ganchev et al. (2009) use parallel texts to train parsers for resource-poor languages, while investigating language-specific constraints for disambiguating annotation choices.

More recently, Jiang et al. (2009) have proposed a strategy for automatically transferring knowledge from a source corpus treebank to a resource-scarce language, using a dynamic algorithm. Similarly, Smith et al. (2009) study cross-lingual parser projections and create a TL dependency parser by using bilingual text, a parser, and automatic word alignments. The starting point of the PRESEMT Phrase aligner is similar, since a two-phase approach is used, with word alignment of a parallel corpus being followed by the segmentation of the SL text. This segmentation, though, is limited to identifying the constituent phrases, without a detailed syntactic analysis.

An alternative approach frequently used in EBMT is the Marker Hypothesis, where specific words are used for signalling phrase boundaries in both the SL and TL (see for instance Gough and Way, 2004). This approach however entails the compilation of marker word lists per language; besides, in the PRESEMT approach the SL text segmentation is guided by the parsing scheme of the TL text.

3.3 Studies relevant to the Phrasing model generator

The second processing stage of the bilingual corpus generates an SL phrasing model based on the phrasing examples provided by the Phrase aligner. The task of extrapolating language models has been widely studied, including both speech recognition and machine translation. Initial efforts in extrapolating language models have been in the area of speech recognition, where most frequently the underlying language model is extracted via the Viterbi algorithm (Bahl et al., 1983). Theoretical aspects of the HMM-based modelling for speech recognition are presented by Rabiner (1989).

In computational linguistics, several methods for extrapolating the language model from a set of observations have been proposed. Klein and Manning (2002) employ Expectation Maximisation techniques for the induction of a grammar, comparing alternative methods by measuring precision, recall and their harmonic mean. Recently Conditional Random Fields (CRF) have been proposed for segmenting and labelling sequential data, based on the conditional approach (Lafferty et al., 2001). CRF is claimed to have a superior performance to both Hidden Markov Models and Maximum Entropy models while avoiding biasing solutions towards states with few successor states (Wallach, 2004).

4 Phrase aligner module (PAM)

4.1 Basic aspects

The Phrase aligner module processes a small bilingual corpus, for each sentence of which, in both languages, words are aligned to each other via the bilingual lexicon. In the present implementation, where the TL phrasing model is provided by a parser, the phrase alignment process relies on clustering all words in an SL sentence into phrases, on the condition that the given phrases in the two languages do not overlap.

For instance, following the phrase alignment, the parallel corpus could contain an SL sentence structure of the type *Phr.1:Phr.2:Phr.3:Phr.4* corresponding to a translation structure of the type *Phr.2':Phr.3':Phr.1':Phr.4'*. Hence, if the PRESEMT system is presented with an SL sentence structure *A:B:C:D*, for which the best-matching structure from the parallel corpus is the structure *Phr.1:Phr.2:Phr.3:Phr.4*, then the SL sentence should be reordered to *B':C':A':D'*, in accordance to the TL equivalent structure of the parallel corpus.

4.2 Design of the PAM algorithm

The Phrase aligner relies on the following resources: (1) Bilingual lexicon from SL to TL; (2) SL tagger-lemmatiser (the tagger may provide both basic PoS characterisation as well as detailed grammatical features such as case, number, person etc.); (3) TL tagger-lemmatiser and shallow parser; (4) TL clause boundary detection tool. So, the following information is available:

- Information on likely **word** and **lemma correspondences between source and target languages**, extracted from the bilingual lexicon. This is distinguished into:
  - one-to-one correspondence (an SL word translates into one TL word)
  - one-to-many correspondence (an SL word corresponds to a TL multi-word unit)
  - many-to-one correspondence (an SL multi-word unit corresponds to a TL single one)
- **Tag** correspondence between the source and the target languages, and in the case of languages with **rich morphology**, additional information such as case or number.
- **In-sentence distance** between words.
- **Decomposition** of the sentence in the target language in sub-sentential segments based on the output of the parser available.

Based on this set of inputs, PAM needs to decide on the optimal segmentation of the source sentence into phrases. Thus a multi-criterion-type comparison is involved, where the different inputs are accordingly prioritised and combined. Naturally, not all aforementioned inputs need be present for PAM to work, though use of all inputs results in a more accurate alignment.

4.3 Implementation of the PAM algorithm

Similarly to related approaches (cf. Och and Ney, 2004 and Ganchev et al., 2009), PAM operates in two steps, where (i) words in the SL sentence are aligned to those of the TL sentence and afterwards (ii) unaligned SL are grouped into phrases depending on agreement of grammatical features.

Step 1: Word aligner

The word aligner algorithm performs alignment of SL words to TL words via the bilingual lexicon. The algorithm allows the one-to-one alignment between SL words and TL ones, while rejecting any multiple alignments, unless the lexicon explicitly provides such information.

The idea underlying this approach is that for every word *k* in SL that is potentially aligned to more than one word in TL, the TL word chosen is the one (a) that has the minimum distance from the single-aligned TL word and (b) for which the corresponding single-aligned SL word has the minimum distance in tokens from word *k*.

In the following example pair of sentences, the tokens of the German article "*die*", "*das*", "*des*", "*der*" could all be aligned with the two instances of "*the*" in the English sentence:

German: "**Die** Europäische Union wurde gegründet, um **das** politische Ziel **des** Friedens zu erreichen, doch ihre Dynamik und ihren Erfolg stammen von ihrem Engagement in **der** Wirtschaft her."

English: "**The** European Union was created to achieve **the** political goal of peace, but its dynamism and success spring from its involvement in economics."

When an SL word remains unaligned, usually due to limited dictionary coverage, the algorithm transliterates it (in case of different SL and TL alphabets, e.g. Greek and English) and consequently attempts to match it to a highly similar word in the TL sentence. In this case, two words are considered similar when their letter-wise similarity, in terms of the longest common subsequence ratio exceeds a threshold.

At the end of Step 1, alignments using single-word information are resolved. SL words that remain unaligned are handled at the next step.

Step 2: Similarity of features

Operating on the output of Step 1, Step 2 handles the unaligned (hence not grouped into phrases) SL words and attempts to include them in phrases, by identifying those aligned SL words that are similar in terms of grammatical features, as these are reflected in the extended PoS-tags yielded by a tagger. Thus, for every unassigned SL word the algorithm calculates the similarity of its extended PoS tag with the extended PoS tags of all the already aligned SL words in the sentence. The extended PoS similarity for each word is then normalised by multiplication with a Gaussian function that takes as input the token-wise distance of words on the sentence. Normalisation allows PAM to cluster words that match to an acceptable extent in terms of tag but are also closely situated in the sentence. The variance of the Gaussian is tuneable to user requirements.

4.4 PAM experimental setup & results

The Phrase aligner module was tested on two language pairs, Greek-to-English and German-to-English. For each pair a bilingual corpus has been compiled manually from the web.

- **Greek → English corpus:** Extracted from a multilingual website[1], it comprises 200 sentences. The SL side of the corpus has been tagged and lemmatised by the ILSP FBT Tagger & Lemmatiser (Papageorgiou et al., 2000), while the TL side has been processed with the TreeTagger (Schmidt 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 (Schmidt and Laws, 2008), while the TL side has been processed with the TreeTagger, yielding tag, lemma and phrase annotations.

These bilingual corpora have been manually modified so that the SL and TL sides of the corpus are as "close" as possible to each other, 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 PAM performance on data devoid of errors.

Experimental results: In order to evaluate PAM, its results of the module were compared

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

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

with a gold-standard reference set that was manually created. This set comprises 50 sentences for the Greek-English corpus (EL-EN) and 30 sentences for the German-English (DE-EN) corpus. Thus, given the phrasing in the TL, for each SL word PAM determined the corresponding phrase label. The PAM output was then compared to the gold-standard segmentation. The degree of match is reported in Table 1.

	Baseline *(step1)*	**PAM v.1** *(steps 1&2)*	**PAM v.2** *(steps 1&2)*	**PAM v.3** *(steps 1&2)*
El-EN	78.0%	84.2%	90.0%	**92.0%**
De-EN	68.0%	77.3%	77.3%	**85.8%**

Table 1. PAM experimental results

Initially, only the information provided by the bilingual lexicon for the word alignment was used, to provide a baseline. Then, the information given by the bilingual lexicon (Step 1) was augmented with the similarity of PoS tags (Step 2), resulting in an error rate reduction of approximately 25%. An error analysis of the experimental results indicated that errors were due to incorrect handling of categories such as proper names, numbers and acronyms. To eliminate these errors, refinements were made to Step 1, addressing word transliteration, identification of highly similar tokens and verification of multi-aligned words. For the final PAM (v.3), the error rate was reduced by up to 50% over the original 2-step PAM (v.1).

5 Phrasing model generator (PMG)

5.1 Basic aspects & design

The PAM-generated phrasal segmentation of the SL side of the bilingual corpus is used to train a phrasing model. This in turn segments SL texts entered to the PRESEMT system for translation. The method for extracting the phrasing model is statistics-based, since substantial research has already been invested in creating similar models.

5.2 PMG Implementation

The PRESEMT system utilises the CRF model for phrasal segmentation in the SL. CRF's main purpose is to generate consistent phrases out of the set of words in any input sentence. Within the main translation process, the phrases created by PMG will be used to search for appropriate translations in the corpora. One main requirement for this module is language-independence, allowing the generation of a model for any language, provided that a suitable training set is available. To this end, in the current implementation only the PoS information per word is taken into account. This choice was made to reduce the amount of training data required, given that the number of possible tags is much lower than the number of potential words/lemmas.

5.3 PMG experimental setup & results

Datasets: In the final PRESEMT system, the PMG will use as input the PAM output and specifically the phrase-aligned SL side of the parallel corpus. To allow experimentation on this second phase before the finalisation of the PAM module, a golden corpus of aligned parallel sentences was manually segmented to give reference segmentations for evaluation purposes.

To examine the successful training of PMG, two independently-created sets of Greek sentences were used, the first one coming from the PRESEMT parallel bilingual corpus (**Set_A**), while the second one was obtained from news reports (**Set_B**). Both directions were used, i.e. PMG was trained with Set_A and then tested on Set_B, and vice versa. To give a representative result, the average of the accuracies for both directions is reported.

For the annotation of segments, a three-column format was used for training data, where each token occupies one line, the first column containing the actual token, the second one the PoS tag and the third column the segment information. The segment information is expressed as a single-letter label, ("**B**", for the first token in the given segment, "**I**" otherwise), followed by the type of segment. Only the type of phrase was included, without a case feature (e.g. the segmentation information would be of the form B-NP or I-NP). Tokens not belonging to a specific segment (for instance, punctuation marks) are identified by a "**0**" clause label.

Experimental results: For the implementation of the experiments presented in this section the Java programming language was chosen as it provides easy integration with web technologies and a wide variety of open source libraries.

Regarding the PMG algorithmic part, the **MALLET**[3] package was chosen as it is implemented in Java, which is also used for the PRESEMT prototype. For parameter passing and object injection the Spring[4] framework was employed, which simplified the execution of series of experiments.

[3] http://mallet.cs.umass.edu/

[4] http://www.springsource.org/

Different system setups were experimentally tested for the CRF model and the different options available within the MALLET toolkit were considered. Both the default CRF training method "*CRFTrainerByLabelLikelihood*" (hereafter denoted as "**std.**") and the alternative method "*CRFTrainerByL1LabelLikelihood*" (denoted as "**alt.**") were tested. Both the complete and reduced tagsets (denoted as "**std.**" and "**red.**" respectively) were considered for training.

Another set of parameters considered relates to the CRF structure. The CRF order parameter sets the time window, on the basis of which the model creates connections between the training sequence and the observed symbols. Additionally, a different feature counter functionality was added by creating a java class implementing the MALLET basic Pipe interface used for feature measurement. This custom implementation employs regular expressions (RegExp) to modify (by "find" and "replace" expressions) input sequences, conferring a wide range of capabilities.

The issue of combining different input parameters such as tags and lemmas has also been added. This functionality represents n-gram features by replacing each plain symbol at a given time with a more complex combination of "previous", "current", and "next seen" symbols based on configured time slots. For example, assuming the sequence "*X0 X1 X2*", the default feature approach would keep the sequence as is, while the modified variant would create the following set of feature triplets for timeslot [-1,0,+1]: *_ X0 X1; X0 X1 X2; X1 X2 _*. Timeslots that refer only to past observations have been studied, as well as alternative timeslots with spans of 3 and 2.

	Parameters			**Model order**		
Feature	**Tags**	**Method**	**Data size**	0	0-1	0-1-2
1-gram	std	std.	17	75.4	**80.4**	77.8
1-gram	red.	std.	17	82.4	**88.1**	84.4
1-gram	red.	std.	17	--	**88.3**	--
1-gram	red.	alt.	34	81.3	89.0	86.0
2-gram	std	std.	17	73.5	74.8	73.3
2-gram	red.	std.	17	85.5	86.7	84.5
2-gram	red.	std.	17	--	86.4	--
2-gram	red.	alt.	34	89.3	**90.0**	88.7

Table 2. PMG experimental accuracies (denoted in percentages)

Finally the effect of the training data size has been examined, keeping the test set intact. In the experiments performed, various set-ups were compared in order to perform a comprehensive evaluation of the CRF accuracy.

The experimental results are summarised in Table 2, where for each token, the PMG-generated phrasing information is compared to the manually-created gold-standard. The phrasing accuracy generated by CRF is adequate for the task at hand, peaking at 90%, for the best configuration, involving the larger training set of 34 sentences. Since a sizeable improvement in accuracy is obtained by increasing the training data size, even better performance may be achieved for a larger set, which is representative of the size of the bilingual corpus. The best results in most cases are achieved when adopting the model with 0-1 CRF order combined with n-gram features of zero value, while by further increasing the model order, and thus its complexity, no improvements are observed.

The reduction in terms of tag complexity also aids the segmentation accuracy, probably due to the fact that the training data is too sparse to support detailed tags. Still, the information of segment type and case of head is sufficient to provide results of a sufficient quality for the PRESEMT MT system. On the contrary, the choice of training algorithm does not affect the results.

6 Conclusions and Future Work

In this article, the phrase alignment approach of the PRESEMT project has been presented, which enables the PRESEMT MT system to be readily-extendable to new language pairs, requiring only widely available tools and resources.

The phrase aligner processes a bilingual corpus of parallel sentences and extracts phrasing models using only a TL parser, thus avoiding any incompatibility issues that arise when using both SL and TL parsers. The phrase aligner comprises two modules, the first one performing a cross-language segmentation of the parallel sentences, the latter extracting a phrasing model for SL.

First implementations of these two modules have been presented, together with experimental results. The resulting accuracies, compared to gold-standard data, indicate the methods' effectiveness in the given task. A number of extensions have been identified, these including the use of more extensive data, the further refinement of the algorithms and the integration of the two modules and the subsequent re-evaluation of the effectiveness of their combination. Of course, the main aim is the integration of the Phrase aligner module to the PRESEMT MT system in order to evaluate its actual effectiveness in MT tasks.

Acknowledgement

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 (PRESEMT project).

References

Bahl, Lalit R., Frederick Jelinek, and Robert L. Mercer. 1983. A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2): 179-190.

Carl, Michael, Maite Melero, Toni Badia, Vincent Vandeghinste, Peter Dirix, Ineke Schuurman, Stella Markantonatou, Sokratis Sofianopoulos, Marina Vassiliou, and Olga Yannoutsou. 2008. METIS-II: Low Resources Machine Translation, *Machine Translation,* Vol. 22, No.1-2, pp. 67-99.

Ganchev, Kuzman, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency Grammar Induction via Bitext Projection Constraints. *Proceedings of the 47th Annual Meeting of the ACL*, Singapore, 2-7 August 2009, 369–377.

Gough, Nano and Andy Way. 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.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas and Okan Kolak. 2005. Bootstrapping parsers via Syntactic Projections across Parallel Texts. *Natural Language Engineering*, 11: 311-325.

Jiang, Wenbin and Qun Liu. 2009. Automatic Adaptation of Annotation Standards for Dependency Parsing – Using Projected Treebank as Source Corpus. *Proceedings of the 11th International Conference on Parsing Technologies*, Paris, October 25-28.

Klein, Dan and Christopher D. Manning. 2002. A Generative Constituent-Context Model for Improved Grammar Induction. *Proceedings of the 40th ACL Meeting*, Philadelphia, U.S.A., July 2002, 128-135.

Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data, *Proceedings of ICML Conference*, June 28-July 1, Williamstown, USA 282-289.

Markantonatou Stella, Sokratis Sofianopoulos, Vassiliki Spilioti, George Tambouratzis, Marina Vassiliou and Olga Yannoutsou. 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.

Och, Franz Josef, and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4): 417-449.

Papageorgiou, Haris, Prokopis Prokopidis, Voula Giouli, and Stelios Piperidis. 2000. A Unified POS Tagging Architecture and its Application for Greek. *LREC-2000 Conference Proceedings*, Athens, Greece 1455-1462.

Rabiner, Lawrence R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257-286.

Schmid, Helmut, and Laws Florian. 2008. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained PoS Tagging. *Proceedings of COLING 2008*, Manchester, Great Britain 777-784.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees (ms.).

Simard, Michel, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with Non-Contiguous Phrases. *Proceedings of the Conferences on Human Language Technology and on Empirical Methods in Language Processing*, Vancouver, Canada 755-762.

Smith, David A. and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. *Proceedings of the 2009 EMNLP Conference*, Singapore 822-831.

Tillmann, Cristoph. 2003. A Projection Extension Algorithm for Statistical Machine Translation. *Proceedings of the EMNLP Conference* 1-8.

Wallach, Hanna M. 2004. Conditional Random Fields: An Introduction. CIS Technical Report, MS-CIS-04-21. 24 February 2004, University of Pennsylvania.

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

Yarowsky, David, and Grace. 2001. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. *Proceedings of NAACL-2001 Conference,* 200-207.

Zakarian, Armen. 2008. A New Non-binary Matrix Clustering Algorithm for Development of System Architectures. *IEEE Trans. on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, 38(1): 135-140.