PRESEMT: Pattern Recognition-based Statistically Enhanced MT

George Tambouratzis, Marina Vassiliou, Sokratis Sofianopoulos
Institute for Language and Speech Processing, Athena R.C.
6 Artemidos & Epidavrou Str., Paradissos Amaroussiou, 151 25, Athens, Greece.
{giorg_t; mvas ; s_sofian}@ilsp.gr

Abstract

This document contains a brief presentation of the PRESEMT project that aims in the development of a novel language-independent methodology for the creation of a flexible and adaptable MT system.

1. Introduction

The PRESEMT project constitutes a novel approach to the machine translation task. This approach is characterised by (a) introducing cross-disciplinary techniques, mainly borrowed from the machine learning and computational intelligence domains, in the MT paradigm and (b) using relatively inexpensive language resources. The aim is to develop a language-independent methodology for the creation of a flexible and adaptable MT system, the features of which ensure easy portability to new language pairs or adaptability to particular user requirements and to specialised domains with minimal effort. PRESEMT falls within the Corpus-based MT (CBMT) paradigm, using a small bilingual parallel corpus and a large TL monolingual corpus. Both these resources are collected as far as possible over the web, to simplify the development of resources for new language pairs.

The main aim of PRESEMT has been to alleviate the reliance on specialised resources. In comparison, Statistical MT requires large parallel corpora for the source and target languages. PRESEMT relaxes this requirement by using a small parallel corpus, augmented by a large TL monolingual corpus.

2. PRESEMT system structure

The PRESEMT system is distinguished into three stages, as shown in Figure 1:

1. Pre-processing stage: This is the stage where the essential resources for the MT system are compiled. It consists of four discrete modules: (a) the **Corpus creation & annotation module**, being responsible for the compilation of monolingual and bilingual corpora over the web and their annotation; (b) the **Phrase aligner module**, which processes a bilingual corpus to perform phrasal level alignment within a language pair; (c) the **Phrasing model generator** that elicits an SL phrasing model on the basis of the aforementioned alignment and employs it as a parsing tool during the translation process; (d) the **Corpus modelling module**, which creates semantics-based TL models used for disambiguation purposes during the translation process.

2. Main translation engine: The translation in PRESEMT is a top-down two-phase process, distinguished into the **Structure selection module**, where the constituent phrases of an SL sentence are reordered according to the TL, and the **Translation equivalent selection module** where translation disambiguation is resolved and word order within phrases is established. Closely integrated to the translation engine, but not part of the main translation process, is the Optimisation module, which is responsible for automatically improving the performance of the two translation phases by fine-tuning the values of the various system parameters.

3. Post-processing stage: The third stage is user-oriented and comprises (i) the Post-processing and (ii) the User Adaptation modules. The first module allows the user to modify the system-generated translations towards their requirements. The second module enables PRESEMT to adapt to this input so that it learns to generate translations closer to the users' requirements. The post-processing stage represents work in progress to be reported in future publications, the present article focussing on the actual strategy for generating the translation.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 65–68,
Avignon, France, April 23 - 27 2012.

3. Processing of the bilingual corpus

The bilingual corpus contains literal translations, to allow the extrapolation of mapping information from SL to TL, though this may affect the translation quality. The Phrase aligner module (PAM) performs offline SL – TL word and phrase alignment within this corpus. PAM serves as a language-independent method for mapping corresponding terms within a language pair, by circumventing the problem of achieving compatibility between the outputs of two different parsers, one for the SL and one for the TL. PAM relies on a single parser for the one language and generates an appropriate phrasing model for the other language in an automated manner.
The phrases are assumed to be flat and linguistically valid. As a parser, any available tool may be used (the TreeTagger (Schmid, 1994) is used in the present implementation for English). PAM processes a bilingual corpus of SL – TL sentence pairs, taking into account the parsing information in one language (in the current implementation the TL side) and making use of a bilingual lexicon and information on potential phrase heads; the output being the bilingual corpus aligned at word, phrase and clause level. Thus, at a phrasal level, the PAM output indicates how an SL structure is transformed into the TL. For instance, based on a sentence pair from the parallel corpus, the SL sentence with structure A-B-C-D is transformed into A'-C'-D'-B', where X is a phrase in SL and X' is a phrase in TL. Further PAM details are reported in Tambouratzis et al. (2011).

The PAM output in terms of SL phrases is then handed over to the Phrasing model generator (PMG), which is trained to determine the phrasal structure of an input sentence. PMG reads the SL phrasing as defined by PAM and generates an SL phrasing model using a probabilistic methodology. This phrasing model is then applied in segmenting any arbitrary SL text being input to the PRESEMT system for translation. PMG is based on the Conditional Random Fields model (Lafferty et al., 1999) which has been found to provide the highest accuracy. The SL text segmented into phrases by PMG is then input to the 1st translation phase. For a new language pair, the PAM-PMG chain is implemented without any manual correction of outputs.

4. Organising the monolingual corpus

The language models created by the Corpus modelling module can only serve translation disambiguation purposes; thus another form of interfacing with the monolingual corpus is essential for the word reordering task within each phrase. The size of the data accessed is very large. Typically, a monolingual corpus contains 3 billion words, 10^8 sentences and approximately 10^9 phrases. Since the models for the TL phrases need to be accessed in real-time to allow word reordering within each phrase, the module uses the phrase indexed representation of the monolingual corpus. This phrase index is created based on four criteria: (i) phrase type, (ii) phrase head lemma, (iii) phrase head PoS tag and (iv) number of tokens in the phrase.

Indexing is performed by extracting all phrases from the monolingual corpus, each of which is transformed to the java object instance used within the PRESEMT system. The phrases are then organised in a hash map that allows multiple values for each key, using as a key the 4 aforementioned criteria. Statistical information about the number of occurrences of each phrase in the corpus is also included. Finally, each map is serialised and stored in the appropriate file in the PRESEMT path, with each file being given a suitable name for easy retrieval. For example, for the English monolingual corpus, all verb phrases with head lemma "*read*" (verb) and PoS tag "*VV*" containing 2 tokens in total are stored in the file "*Corpora\EN\Phrases\VC\read_VV*". If any of these criteria has a different value, then a separate file is created (for instance for verb phrases with head "*read*" that contain 3 tokens).

5. Main translation engine

The PRESEMT translation process entails first the establishment of the sentence phrasal structure and then the resolution of the intra-phrasal arrangements, i.e. specifying the correct word order and deciding upon the appropriate candidate translation. Both phases involve searching for suitable matching patterns at two different levels of granularity, the first (coarse-grained) aiming at defining a TL-compatible ordering of phrases in the sentence and the second (fine-grained) determining the internal structure of phrases. While the first phase utilises the small bilingual corpus, the second phase makes use of the large monolingual corpus. To reduce the translation time required, both corpora are processed in advance and the processed resources are stored in such a form as be retrieved as rapidly as possible during translation.

5.1 Translation Phase 1: Structure selection module

Each SL sentence input for translation is tagged and lemmatised and then it is segmented into phrases by the Phrasing model generator on the basis of the SL phrasing model previously created. For establishing the correct phrase order according to the TL, the parallel corpus needs to be pre-processed using the Phrase aligner module to identify word and phrase alignments between the equivalent SL and TL sentences.

During structure selection, the SL sentence is aligned to each SL sentence of the parallel corpus, as processed by the PAM and assigned a similarity score using an algorithm from the dynamic programming paradigm. The similarity score is calculated by taking into account edit operations (replacement, insertion or removal) needed to be performed in the input sentence in order to transform it to the corpus SL sentence. Each of these operations has an associated cost, considered as a system parameter. The aligned corpus sentence that achieves the highest similarity score is the most similar one to the input source sentence. This comparison process relies on a set of similarity parameters (e.g. phrase type, phrase head etc.), the values of which are optimised by employing the optimisation module.

The implementation is based on the Smith-Waterman algorithm (Smith and Waterman, 1981), initially proposed for determining similar regions between two protein or DNA sequences. The algorithm is guaranteed to find the optimal local alignment between the two input sequences at clause level.

5.2 Translation Phase 2: Translation equivalent selection module

After establishing the order of phrases within each sentence, the second phase of the translation process is initiated, comprising two distinct tasks. The first task is to resolve the lexical ambiguity, by picking one lemma from each set of possible translations (as provided by a bilingual dictionary). In doing so, this module makes use of the semantic similarities between words which have been determined by the Corpus Modelling module through a co-occurrence analysis on the monolingual TL corpus. That way, the best combination of lemmas from the sets of candidate translations is determined for a given context.

In the second task, the most similar phrases to the TL structure phrases are retrieved from the monolingual corpus to provide local structural information such as word-reordering. A matching algorithm selects the most similar from the set of the retrieved TL phrases through a comparison process, which is viewed as an assignment problem, using the Gale-Shapley algorithm (Gale and Shapley, 1962).

6. Experiments & evaluation results

To date MT systems based on the PRESEMT methodology have been created for a total of 8 languages, indicating the flexibility of the proposed approach. Table 1 illustrates an indicative set of results obtained by running automatic evaluation metrics on test data translated by the 1st PRESEMT prototype for a selection of language pairs, due to space restrictions.

In the case of the language pair English-to-German, these results are contrasted to the ones obtained when translating the same test set with Moses (Koehn et al., 2007).It is observed that for the English-to-German language pair, PRESEMT achieved approximately 50% of the MOSES BLEU score and 80% of the MOSES with respect to the Meteor and TER scores. These are reasonably competitive results compared to an established system such as Moses. Furthermore, it should taken into consideration that (a) the PRESEMT results were obtained by the 1st system prototype, (b) PRESEMT is still under development and (c) only one reference translation was used per sentence.

Newer versions of the PRESEMT system, incorporating more advanced versions of the different modules are expected to result in substantially improved translation accuracies. In particular, the second translation phase will be further researched. In addition, experiments have indicated that the language modelling module can provide additional improvement in the performance. Finally, refinements in PAM and PMG may lead in increased translation accuracies.

7. Links

Find out more about the project on the **PRESEMT website**: www.presemt.eu. Also, the PRESEMT **prototype** may be tried at: presemt.cslab.ece.ntua.gr:8080/presemt_interface_test

Acknowledgments

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.

References

Gale D. and L. S. Shapley. 1962. College Admissions and the Stability of Marriage. American Mathematical Monthly, Vol. 69, pp. 9-14.

Koehn P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the ACL-2007 Demo and Poster Sessions.

Kuhn H. W. 1955. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, Vol. 2, pp.83-97.

Lafferty J., A. McCallum, F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data. Proceedings of ICML Conference, pp.282-289.

Munkres J. 1957. Algorithms for the assignment and transportation problems. Journal of the Society for Industrial and Applied Mathematics, Vol. 5, pp.32-38.

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

Smith T. F. and M. S. Waterman. 1981. Identification of Common Molecular Subsequences. Journal of Molecular Biology, 147: 195–197.

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

Table 1 – PRESEMT Evaluation results for different language pairs.

Language Pair		Sentence set		Metrics				
SL	TL	Number	Source	BLEU	NIST	Meteor	TER	
English	German	189	web	0.1052	3.8433	0.1939	83.233	
German	English	195	web	0.1305	4.5401	0.2058	74.804	
Greek	English	200	web	0.1011	4.5124	0.2442	79.750	
English	German	189	web	0.2108	5.6517	0.2497	68.190	**Moses**

Figure 1 – PRESEMT system architecture.