Evaluating the translation accuracy of a novel language-independent MT methodology

George TAMBOURATZIS, Sokratis SOFIANOPOULOS, Marina VASSILIOU
Institute for Language & Speech Processing, Athena R.C.
6 Artemidos & Epidavrou Str., Paradissos Amaroussiou, 151 25 Athens, Greece
giorg_t@ilsp.gr,s_sofian@ilsp.gr,mvas@ilsp.gr

ABSTRACT

The current paper evaluates the performance of the PRESEMT methodology, which facilitates the creation of machine translation (MT) systems for different language pairs. This methodology aims to develop a hybrid MT system that extracts translation information from large, predominantly monolingual corpora, using pattern recognition techniques. PRESEMT has been designed to have the lowest possible requirements on specialised resources and tools, given that for many languages (especially less widely used ones) only limited linguistic resources are available. In PRESEMT, the main translation process is divided into two phases, the first determining the overall structure of a target language (TL) sentence, and the second disambiguating between alternative translations for words or phrases and establishing local word order. This paper describes the latest version of the system and evaluates its translation accuracy, while also benchmarking the PRESEMT performance by comparing it with other established MT systems using objective measures.

KEYWORDS: hybrid machine translation; language-independent methodology; MT evaluation

1 Introduction

The Machine Translation (MT) task has been studied for a number of decades, but still remains to a large extent an issue unresolved, as the performance delivered by the best current systems still falls short of the required quality. Since the number of texts available over the World Wide Web is ever increasing, and these texts may be written in one of several hundred languages, the requirement for automatically performing translation of an acceptable quality remains a prime objective. A number of MT paradigms have been proposed, the main ones including Rule-Based MT (RBMT), Statistical MT (SMT) and Example-Based MT (EBMT). Furthermore, the requirement for covering an ever increasing combination of Source to Target language (SL to TL) combinations necessitates the development of language-independent methodologies.

Currently most MT approaches are based on the SMT paradigm (Koehn, 2010). SMT uses dedicated algorithms that do not employ language-specific rules and is thus portable to new language pairs, provided the necessary training data are available. The main SMT constraint is the need for SL-TL bilingual corpora of a sufficient size (of the order of a million parallel sentences) to allow the building of accurate translation models. Such corpora are hard to obtain, particularly when less widely-used languages are involved. Besides, the process of compiling and verifying such corpora is expensive in terms of both manpower and time.

In EBMT, translations are generated by analogy, where the system has available a set of known pairs of input sentence (in SL) and corresponding translation (in TL). Then, each new input sentence is broken down to non-overlapping phrases, which are translated using the translation examples as a reference. The translated sentence is finally composed by combining the translated phrases.

Another paradigm is hybrid MT, which combines ideas and techniques from more than one approaches, like for example EBMT and SMT techniques (cf. Groves & Way, 2005 and Phillips, 2011). Such approaches have been proposed for creating MT systems using more limited but easily obtainable resources. Even if these methods do not achieve accuracy as high as that of SMT systems, their ability to develop MT systems with limited resources is an advantage in the case of less-widely used languages. The PRESEMT system is based on such a methodology, as detailed below, its main characteristics being the use of only very small bilingual corpora and the employment of large monolingual corpora for extracting most of the necessary linguistic information.

A number of methods for the automatic inference of templates for the structural transfer from SL to TL have been proposed. For instance, Sanchez-Martinez et al. (2009) suggest using small parallel corpora only to extract transfer rules, assuming that a sufficient bilingual dictionary is already available. Carbonell et al. (2006) propose an MT method that requires no parallel corpora, but relies on a translation model utilising a full-form bilingual dictionary and a decoder using long-range context via large n-grams.

Another family of systems are METIS (Dologlou et al., 2003) and METIS-II (Markantonatou et al., 2009), both of which rely solely on extensive monolingual resources in order to generate translations employing pattern recognition-based algorithms. The METIS family represents the ancestor of PRESEMT, which has built upon the past experience, by (i) adding a small bilingual corpus to improve translation accuracy and (ii) using more advanced algorithms for pattern matching to provide a measurable increase in both speed and accuracy of the generated translations.

2 The principles of the PRESEMT system

In terms of resources, similarly to METIS-II, PRESEMT uses a bilingual dictionary providing SL – TL lexical correspondences and an extensive TL monolingual corpus collected automatically from the web. A small bilingual corpus containing parallel sentences is added in PRESEMT, in order to (a) reduce the number of possible translations that need to be evaluated by the system and (b) define examples of SL – TL structural modifications, thus improving the translation quality. The bilingual corpus need not cover a particular domain and only numbers a few hundred sentences (typically ~200) for determining structural equivalences between sentences in the source and target languages. Hence, in comparison to SMT systems, the size of the parallel corpus required is reduced by more than three orders of magnitude. Evidently, for a bilingual corpus of only a few hundred sentences, not all linguistic phenomena are likely to occur. However, it is expected that the most frequent ones will be covered and thus a sufficient coverage of the structure transformations from SL to TL can be achieved.

Both the bilingual and the monolingual corpora are annotated[1] with lemma and Part-of-Speech (PoS) information and, depending on the language, with additional morphological features (e.g. case, number, tense etc.). Furthermore, they are segmented into non-recursive syntactic phrases (e.g. noun phrase, verb phrase etc.). The next subsections describe the kind of information extracted.

2.1 Processing the bilingual corpus

The processing of the bilingual corpus involves the use of a pair of modules, namely the Phrase aligner module (PAM) and the Phrasing model generator (PMG). PAM operates on the bilingual corpus to achieve the establishment of matching phrasing schemes in the SL and TL sides. This is achieved by aligning the bilingual sentences initially at a word level and then porting these alignments at a phrase level. PAM aims at identifying how the SL structure is modified towards the TL one, allowing the deduction of a phrasing model for the source language. During initialisation, PAM takes as input a parsed text in the TL-side of the parallel corpus, via a chosen TL parser. It is assumed that in this corpus there is a high level of fidelity between the SL-side and TL-side, which extends to the phrasing schemes of the two languages. Then, PAM algorithmically segments the SL-side sentence into phrases in accordance to the TL side. To achieve that, PAM takes into account alignment information in the form of (a) lexicon-based correspondences, (b) alignment on the basis of grammatical feature similarity and PoS tag correspondence and (c) alignment information provided by already aligned neighbouring words in the SL and TL sides. Within this sequence, in each consecutive step additional SL words are aligned to TL words, the aim being for all words to be assigned to SL phrases that correspond to the TL phrasing, these phrases then being mapped to the TL phrases.

[1] For the annotation task readily available tools are employed, including statistical taggers and (to some extent) chunkers that provide shallow parsing. This alleviates the need for developing new linguistic tools.

The SL side of the aligned corpus is subsequently processed by PMG, with a two-fold purpose, namely to (i) deduce a phrasing model based on conditional random fields (CRF) (Lafferty et al., 2001) and (ii) employ this model for parsing any SL text submitted for translation. During the derivation of a phrasing model, the SL side of the aligned bilingual corpus is used to train a CRF model via a standard iterative process. During operation, this model is used to segment new sentences to be translated into their constituent phrases. Details on the algorithmic design and individual accuracy of PAM and PMG are provided in Tambouratzis et al. (2012a).

2.2 Extracting information from the monolingual corpus

The TL monolingual corpus is processed to create two distinct models, which are employed during the translation process. The first model is used solely for disambiguation purposes, when two or more translations are proposed for a word or set of words. In this account, different models have been studied, including a SOM-based model (Tsimboukakis et al., 2011 and Tambouratzis et al., 2012b) and an n-gram-based model. The second one is a phrase model that provides the micro-structural information on the translation output, to determine intra-phrasal word order. The model is stored in a file structure, where a separate file is created for phrases according to their (i) phrase type, (ii) phrase head and (iii) phrase head PoS tag. As most of the progress has involved developments in the second model, this is the one discussed in more detail in the remainder of the present section.

The number of files created as a result of this process is very large (of the order of millions of files), as for each combination of the three aforementioned criteria, a different file needs to be created; yet each of the files is of a small size and thus can be retrieved and loaded quickly. Due to the very large number of files, the actual data structure and implementation becomes very important. Currently PRESEMT uses a simple string representation to store the phrases in each file, ordered by their frequencies of occurrence. Initially the phrases were stored as serializable objects in hash tables, based on their order of appearance in the corpus. This redesigned model occupies substantially less disk space and provides faster retrieval. Also the ordered storage of phrases provides the algorithm a way to stop the search as soon as a relevant phrase has been retrieved. On the whole, the aforementioned revisions in the modelling of the phrases have led in a reduction in the translation time of approximately 40%, when averaged over a set of 200 sentences being translated (to avoid bias due to sentence-specific phenomena. Regarding the disk requirements, the use of the revised mapping has resulted in a substantial drop in the required space for storing the model (for a corpus of 80 Gbytes, the model size has been reduced by approximately 58%, from 22 Gbytes to 9.3 Gbytes).

2.3 Main translation engine

The translation process is split into two phases. Phase 1 (**Structure selection**) uses the bilingual corpus to determine, for a given input SL sentence, the appropriate TL structure in terms of the sequence of phrases and their order. The output of the Structure selection phase is the SL sentence with a TL structure, created by reordering the phrases according to the archetypes contained in the parallel corpus, and all words replaced by the TL lemmas and tag information as retrieved from the bilingual dictionary.

Phase 2 (**Translation equivalent selection**) uses the models extracted from the TL monolingual corpus as described in section 2 so as to specify the most likely word order within phrases, to handle functional words such as articles and prepositions and to resolve lexical ambiguities emerging from the possible translations provided by the bilingual dictionary. Finally, a token generator component generates tokens out of lemmas. Therefore, the first PRESEMT translation phase is closely related to EBMT, while the second phase is reliant upon information of a statistical nature (but extracted from monolingual corpora), resulting in a hybrid nature.

3 Phase 1: Structure selection

The task of Structure selection is to determine for each input sentence the type of TL phrases to which the SL ones translate and to order them in the TL sentence. To this end it consults the patterns of SL – TL structural modifications to be found in the parallel corpus, thus resembling EBMT (Hutchins, 2005).

Translation phase 1 receives as input an SL sentence (termed **ISS** – Input Source Sentence), bearing lexical translations from the dictionary, annotated with tag and lemma information and segmented into phrases by PMG. A dynamic programming algorithm is applied to determine for each ISS the most similar, in terms of phrase structure, SL sentence found in the bilingual corpus (termed **ACS** – Aligned Corpus Sentence)[2].

The similarity is determined on the basis of structural information such as phrase type, phrase head PoS tag, phrase functional head info and phrase head case. The phrases within ISS are reordered in accordance to the TL side of the chosen ACS by replicating the SL-TL phrase alignment mapping. The dynamic programming algorithm evaluates the similarity in the SL language. The most similar SL structure of the bilingual corpus, that determines the TL structure of the sentence to be translated, is thus selected purely on SL properties. The implemented method is based on the Smith-Waterman algorithm (Smith and Waterman, 1981), initially proposed for aligning DNA and RNA sequences. This algorithm is guaranteed to find the optimal local alignment between two input sequences.

The structural similarity between ISS and ACS is reflected on the similarity score, for the calculation of which a two-dimensional matrix is created with the ISS phrases along the top row and the ACS along the left side. As is standard practice in Dynamic Time Warping, movement across this matrix is from the top left corner towards the bottom right-hand side. The similarity for cell (i,j) is determined by examining the predecessor cells located directly to the left $(i, j-1)$, directly above $(i-1, j)$ and above-left $(i-1, j-1)$,, and is calculated iteratively as the maximum of the three similarities. The similarity of two phrases results by the weighted sum of the similarities of (a) the phrase type, (b) the phrase head PoS tag, (c) the phrase head case and (d) the functional phrase head PoS tag.

The similarity score ranges from *100* to *0*, these limits denoting respectively exact match and total dissimilarity between elements of ISS and ACS. In case of a zero similarity score,

[2] If the most similar ACS retrieved from the parallel corpus is very dissimilar, then ISS does not undergo any reordering. It is notable that in our experiments never did such an occasion appear, the similarity always reaching a high percentage (above 70%). The fact that comparisons involve sentences of the same language (SL) contributes to a high similarity score.

a penalty weight (-50) is employed, to further penalise the establishment of a mapping between dissimilar items.

After calculating the final similarity score between sentences, the comparison matrix indicates the optimal phrase alignment between the two SL sentences. By combining the SL sentence alignment from the algorithm with the alignment information between the ACS and the attached TL sentence, ISS phrases are reordered accordingly.

4 Phase 2: Translation equivalent selection

Following the completion of Phase 1, remaining translation issues include (i) establishing word order within phrases, (ii) handling functional words and (iii) resolving translation ambiguities. To establish the correct word order, the monolingual TL corpus is searched to determine the most similar phrase to each phrase in the SL sentence. The similarity measure takes into account the phrase type and the words contained in terms of lemma, PoS tag and morphological features. These factors enter the comparison with different weights, the relative magnitudes of which are the subject of an optimisation process.

The main issue at this stage is word reordering within each phrase. This entails that the words of a given phrase of the input sentence (denoted as **ISP –** Input Sentence Phrase) and the words of a retrieved TL phrase (denoted as **MCP** – Monolingual Corpus Phrase) are close to each other in terms of number of words and type.

When initiating Phase 2 of the translation process, the matching algorithm accesses the indexed TL phrase corpus (created as described in section 2) to retrieve similar phrases and select the most similar one through a comparison process, which is viewed as an assignment problem. This problem can be solved via both exact algorithms that guarantee the identification of the optimal solution and sub-optimal ones. Experiments when developing METIS-II have shown that the solution of the assignment problem is computationally-intensive. Consequently, to conform to the strict translation time constraints set for PRESEMT, the Gale-Shapley algorithm is used (Gale and Shapley, 1962 and Mairson, 1992), which solves the assignment problem in a reduced time. This process is possibly non-optimal but allows a substantial reduction in the computation time.

After the completion of this comparison process, the selected phrase from the monolingual corpus serves as a basis for resolving other issues such as the handling of functional words (e.g. insertion / deletion of articles). In this process, the TL information prevails over the SL entries, based on initial experiments performed, to provide a translation closer to the TL-provided information.

Translation equivalent selection receives as input the output of Structure selection, which contains sets of candidate translations for each SL lemma. One translation needs to be chosen from each set, thus disambiguating amongst the possible translations. The disambiguation process uses the semantic similarities between words as evidenced by the monolingual corpus. Different approaches are evaluated within PRESEMT for selecting the most appropriate translation, including Vector Space Modelling (Marsi et al., 2010) and Self-Organising Maps, following the work by Tsimboukakis et al. (2011).

Rather than employing these disambiguation processes, a simpler, corpus-based approach is proposed in the PRESEMT configuration discussed here, which relies on the extraction of statistical information with only limited pre-processing. This method reuses and enhances the indexed sets of the monolingual corpus phrases, by exploiting information on the frequency of occurrence of each TL phrase. When searching for the best matching TL phrase for each combination of lexical alternatives, the frequency of the TL phrase is taken into account. Notably, not all combinations are examined for lexical disambiguation; instead only the phrase mapped to the most frequent TL phrase is retained. A formula is used for selecting the most appropriate phrase based on both the similarity score and the frequency of the TL phrase. This formula ensures that even though one TL might achieve a higher comparison score than another, if its frequency is significantly lower, then the second phrase - which has a lower absolute score - will be selected, due to its substantially higher frequency of occurrence. This enables the algorithm to delete or add a word such as an article in the final translation of the phrase. This scoring mechanism can be easier to understand with a Greek to English translation example where the article in the Greek phrase needs to be removed from the English translation, for instance the Greek noun phrase '*Η Γαλλία*', which translates to "*France*" in English. When searching for relevant phrases in the TL model, the phrase "*the France*" scores *100* and appears in the corpus *34* times, while the phrase "*France*" scores *85* and appears *5,030* times. Using the aforementioned method and a threshold value of the score ratio being equal to 90% in this case, the ratio between the two scores is not high enough, so the selection will be based on the frequency of occurrence ratio between the two phrases, where the correct phrase (the one without the article) has a substantially higher number of occurrences in the corpus.

5 Example of the PRESEMT translation process

In this section a simple example is used to illustrate the translation process of the PRESEMT system in a step-wise manner. An SL sentence as the one in (1) is being input for translation:

(1) Εδραιώνονται σχέσεις καλής γειτονίας στις χώρες των Βαλκανίων
"Good neighbourhood relations are established in the Balkan countries"

Annotation at various levels [tagging & lemmatising; PMG-based segmentation to phrases (VC: verb chunk, PC: prepositional chunk); output of the lexicon look-up]

SL sentence annotated after being input for translation			
Phrase	**VC**	**PC**	**PC**
Word	εδραιώνονται	σχέσεις καλής γειτονίας	στις χώρες των Βαλκανίων
Lemma	εδραιώνω	σχέση, καλός, γειτονία	στου, χώρα, ο, Βαλκάνια
Tag	**vb**o3pl	**no**feplnm, **aj**fesgge, **no**fesgge	**as**feplac, **no**feplac, **at**neplge, **no**neplge

Lexicon	{consolidate; establish}	{relation; relationship}
		{nice; decent; good}
		{adjacency; neighbourhood}
			{on; at; to; into; in; upon}
			{country}
			{the}
			{Balkan}

1st translation phase: Search the bilingual corpus for the most similar SL sentence in structural terms, find the corresponding TL one and reorder the input SL sentence on the basis of TL; output an intermediate result (2).

Most similar SL sentence of the bilingual parallel corpus			
Phrase	**VC**	**PC**	**PC**
Word	σημειώνονται	διαμαρτυρίες φοιτητών	σε άλλες χώρες της ΕΕ
Lemma	σημειώνω	διαμαρτυρία, φοιτητής	σε, άλλος, χώρα, ο, ΕΕ
Tag	**vb**03pl	**no**feplnm, **no**maplge	**asppsp**, **pn**fe03plac, **no**feplac, **at**fesgge, **abbr**
Corresponding TL sentence of the bilingual parallel corpus			
Phrase	**VC**	**PC**	**PC**
Word	student protests	occur	in other EU countries
Lemma	student, protest	occur	in, other, EU, country
Tag	nn, nns	vv	in, jj, np, nns

(2) Output of the 1st translation phase (expressed as list of phrases and lemmas):

[$_{\text{PC}}${relation; relationship}; {nice; decent; good}; {adjacency; neighbourhood}]
[$_{\text{VC}}${consolidate; establish}]
[$_{\text{PC}}${on; at; to; into; in; upon}; {country}; {the}; {Balkan}]

2nd translation phase: Identify the correct word order within each phrase (3); disambiguate the translations (4); generate tokens out of lemmas (5); produce final translation (6).

(3) **Word reordering results:**

[$_{\text{PC}}${nice; decent; good}; {adjacency; neighbourhood}; {relation; relationship}]
[$_{\text{VC}}${consolidate; establish}]
[$_{\text{PC}}${on; at; to; into; in; upon}; {the}; {Balkan}; {country}]

(4) **Disambiguation results:**

[$_{\text{PC}}${good}; {neighbourhood}; {relation}]
[$_{\text{VC}}${establish}]
[$_{\text{PC}}${in}; {the}; {Balkan}; {country}]

(5) **Token generation:**

[$_{\text{PC}}${good}; { neighbourhood}; {relations }]
[$_{\text{VC}}${are established}]
[$_{\text{PC}}${in}; {the}; {Balkan}; {countries}]

(6) **Final translation:** Good neighbourhood relations are established in the Balkan countries

6 Experimental Results

The evaluation results reported here concern the Greek – English language pair[3] and are based on the development datasets used in PRESEMT for studying the system performance. For each SL, these datasets contain 1,000 sentences, collected via web-crawling. Sentence length ranges from 7 to 40 words.

[3] PRESEMT currently handles 8 language pairs: SL {Czech, English, German, Greek, Norwegian} – TL {English, German}.

From these datasets, 200 sentences were randomly chosen, and manually translated into each of the target languages. The correctness of these reference translations was checked independently by native speakers. For the current evaluation phase four automatic evaluation metrics have been employed, i.e. BLEU (Papineni et al., 2002), NIST (NIST 2002), Meteor (Denkowski and Lavie, 2011) and TER (Snover et al., 2006).

For the bilingual corpus, 200 sentences were used. The Greek-English dictionary contained a total of just over 40,000 entries. For PRESEMT, two versions are evaluated. The first one (PRESEMT-1) indicates the state of the system on April 2012 (i.e. after almost 28 months of development of the system, including the system specifications definition). PRESEMT-1 includes the basic configuration of the system as described in Sofianopoulos et al. (2012). PRESEMT-2 encompasses a number of improvements in comparison to PRESEMT-1, including refined algorithms for the two translation phases, an improved method of using the indexed monolingual corpus and later enhanced versions of the PAM/PMG modules (reflecting the current state in October 2012). Table 1 summarises indicative scores obtained together with scores achieved by four MT systems available online for the same set of data.

	BLEU	**NIST**	**Meteor**	**TER**
Google	0.5544	8.8051	0.4665	29.791
Systran	0.2930	6.4664	0.3830	49.721
WorldLingo	0.2659	5.9978	0.3666	50.627
Bing	0.4600	7.9409	0.4281	37.631
METIS-II	0.1222	3.1655	0.2698	82.8780
PRESEMT-1	0.1683	5.7389	0.3203	68.4670
PRESEMT-2	0.3011	6.6878	0.3733	54.5990

TABLE 1 – Comparison to other MT systems for the Greek-to-English language pair

In comparison to METIS-II[4], the latest version of PRESEMT offers a substantial improvement for all metrics, with for instance BLEU and NIST scores both being increased by more than 145%. This illustrates the improvements conferred by the new translation methodology as compared to the METIS-II family.

It is noteworthy that PRESEMT outperforms two of the other MT systems, Systran and WorldLingo, with scores increased by 2.7% and 13% respectively. As noted, PRESEMT is still under development and it is anticipated that more extensive experiments involving additional language pairs will provide improvements in the translation quality.

6.1 Detailed analysis of the evaluation results

In the present section, the aim is to visualise the evaluation results for the development set. In Figure 1 the BLEU results of the earlier PRESEMT prototype are indicated in a

[4] http://www.ilsp.gr/metis2/

scatter plot, as a function of the sentence size for the language pair Greek-to-English. It can be seen that, as the input sentence size increases in terms of words, the score shows a trend of reducing. Also, it can be noted that for most sentences, the BLUE score is less than 0.2, indicating a less than satisfactory translation.

The highest BLEU score of PRESEMT-1 is equal to 0.56 and is obtained for a relatively short sentence of 12 words, while for only a few medium to long sentences (of 15 words or more) is a BLEU score of 0.4 or more achieved. Finally, for sentences with length of 20 words or more the BLEU score rarely exceeds 0.2.

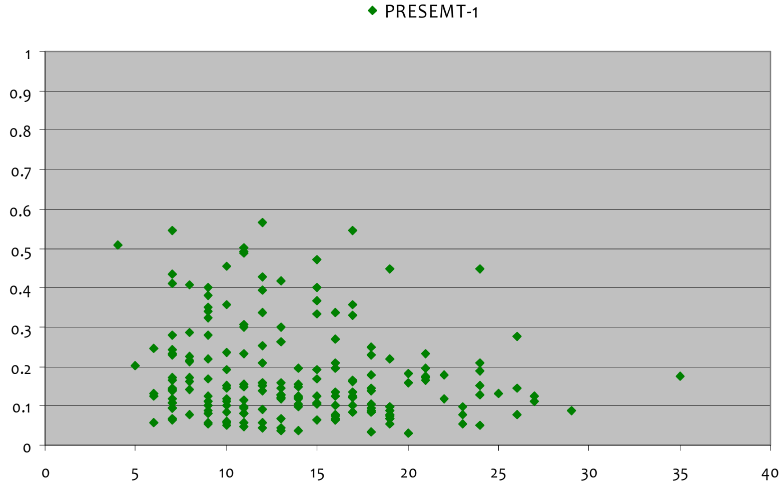

FIGURE 1 – Scatter plot of BLEU results for the EL-EN language pair (PRESEMT-1)

In Figure 2, the BLEU results for the PRESEMT-2 are indicated in a scatter plot, as a function of the sentence size for the language pair Greek-to-English. It is evident that the translation quality is improved, with BLEU scores exceeding 0.5 for a number of sentences. In addition, even for large input sentence sizes, relatively high BLEU scores are achieved (for instance, for the largest sentence of 35 words, a score of almost 0.6 is achieved). Furthermore, even for sentences of more than 25 words, the majority of translations approximate or exceed a score of 0.5, whilst when using PRESEMT-1 (cf. Figure 1) no sentences of this length manage a BLEU score exceeding 0.3.

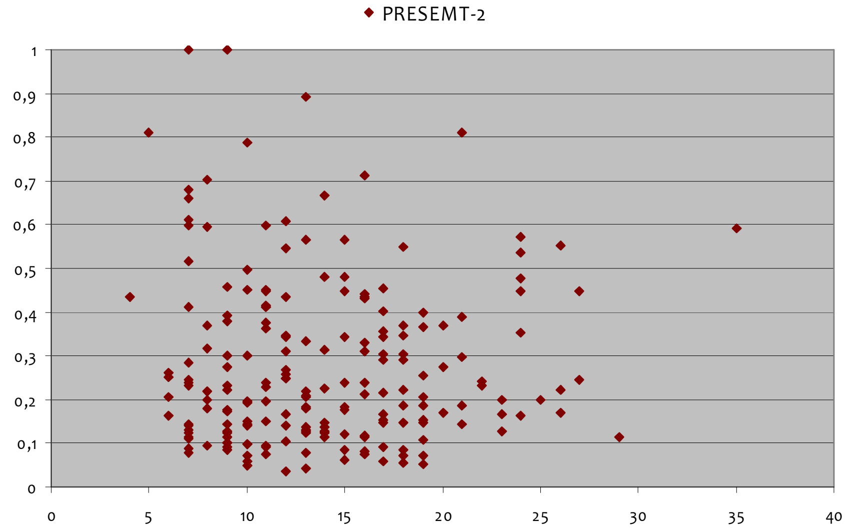

FIGURE 2 – Scatter plot of BLEU results for the EL-EN language pair (PRESEMT-2)

To perform a more systematic analysis, the different sentence sizes have been organised by defining bins, each of which spans 8 sentence sizes (i.e. the first bin concerns sentences of between 4 and 7 words, bin 2 comprises sentences between 8 and 10 words etc.). A boxplot diagram is used to indicate for each of the aforementioned bins the characteristics of BLEU scores, as shown in Figure 3 for PRESEMT-1 and in Figure 4 for PRESEMT-2.

By comparing the boxplots of the two PRESEMT versions for BLEU, it can be seen that boxplots for PRESEMT-1 occupy similar ranges of the score range to those of PRESEMT-2. However, the range for PRESEMT-1 is displaced towards lower values of BLEU in comparison to PRESEMT-2, while also a larger number of outliers exist for PRESEMT-1. Thus, most median values of PRESEMT-1 for different sentence sizes are placed at lower BLEU levels, below the 0.15 mark, with only a few outliers exceeding the limited range of the boxplots.

On the contrary, when turning to PRESEMT-2, the median values are higher, exceeding 0.200 in most cases and even reaching 0.400 in some of the cases. Besides, when comparing the median values, these are increased by 50% or more for most sentence sizes for PRESEMT-2 in comparison to PRESEMT-1. Also, for longer sentences (for instance bin6, which comprises sentences of 24 to 27 words), the improvement in BLEU score is substantial, increasing by a factor of approximately 2.5. This applies to the value corresponding to the 50% level (i.e. the median value) as well as to the levels of 25% and 75%.

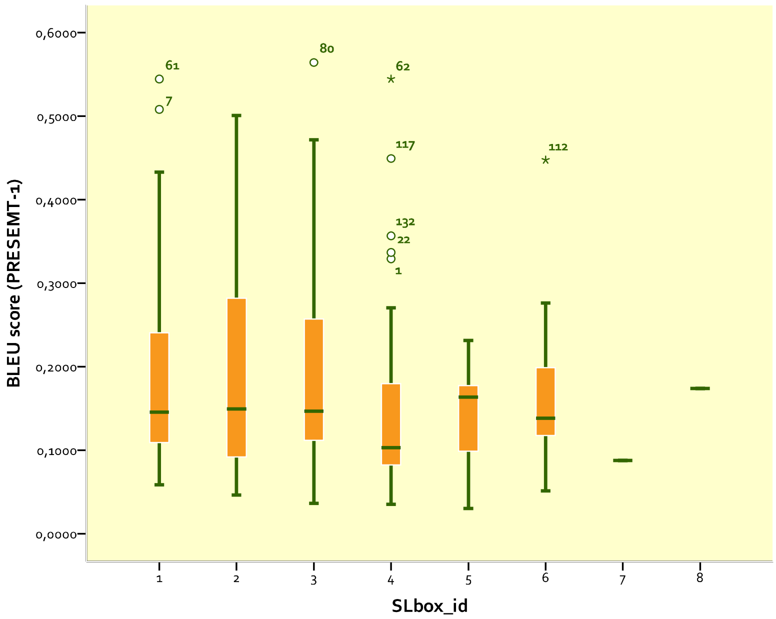

FIGURE 3 – Box plot of BLEU results for the EL-EN language pair (PRESEMT-1)

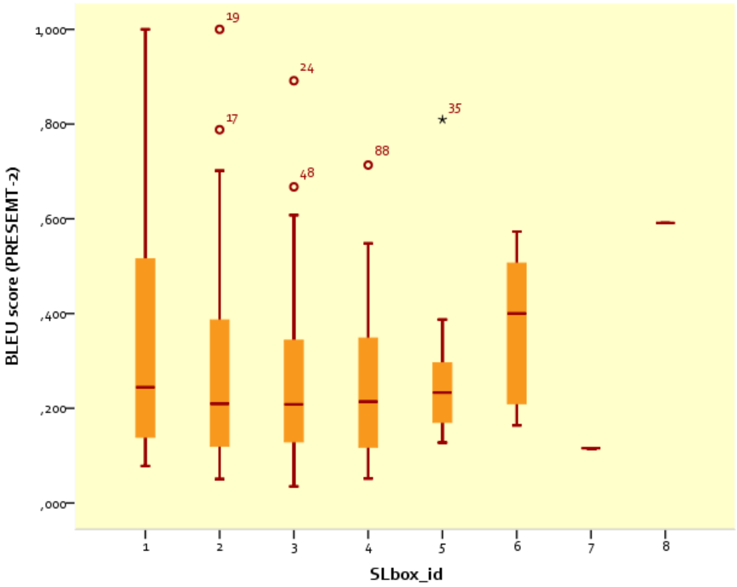

FIGURE 4 – Box plot of BLEU results for the EL-EN language pair (PRESEMT-2)

Furthermore, even though the y-axis scale is larger in Figure 4 than that of Figure 3 the population of solutions covers a much wider range and in several cases translations of a substantially higher quality are achieved. Though the variances are substantially higher for Figure 4 as compared to Figure 3, this is due to several sentences being translated much more accurately, thus reflecting a better translation performance. In addition, the boxplot outliers are fewer in the case of PRESEMT-2, while the variance does not appear to increase as the sentence size increases.

Finally, the BLEU score does not appear to reduce substantially as the sentence size increases, promising scalability of the PRESEMT system for more complex sentences (though this would need to be confirmed via more extensive experiments), with a dependable level of performance. This indicates that the algorithmic improvements integrated when transitioning from PRESEMT-1 to PRESEMT-2 result in a higher translation quality and also contribute to a more predictable performance.

Conclusions

In the present article the principles and the implementation of a novel language-independent methodology have been presented. The PRESEMT methodology draws on information residing in a large monolingual corpus and a small bilingual one for creating MT systems readily portable to new language pairs. Most of this information is extracted in an automated manner using pattern recognition techniques.

First experimental results and comparisons to established systems have been reported. These results are promising, especially taking into account the fact that several PRESEMT modules are still under development and the translation process is being refined, in particular with respect to the handling of internal phrasal structure. Initial studies of the PRESEMT translations have indicated that the handling of the bilingual corpus and the structure selection phase possess the greatest potential for further improvements. The outcome of these efforts will be reported in future articles.

References

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T. and Frey, J. (2006). Context-Based Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, pp. 19-28.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. *EMNLP 2011 Workshop on Statistical Machine Translation*, Edinburgh, Scotland, pp. 85-91.

Dologlou, I., Markantonatou, S., Tambouratzis, G., Yannoutsou, O., Fourla, A. and Ioannou, N. (2003). Using Monolingual Corpora for Statistical Machine Translation: The METIS System. In Proceedings of the EAMT- CLAW'03 Workshop, Dublin, Ireland, pp. 61-68.

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

Groves, D. and Way, A. (2005). Hybrid data-driven Models of Machine Translation. *Machine Translation*, Vol. 19, pp.301-323.

Hutchins, J. (2005). Example-Based Machine Translation: a Review and Commentary. *Machine Translation*, Vol. 19, pp.197-211.

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

Lafferty, J., McCallum, A. and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data. *28th International Conference on Machine Learning, ICML 2011*, Bellevue, Washington, USA, pp. 282-289.

Markantonatou, S., Sofianopoulos, S., Giannoutsou, O. and Vassiliou, M. (2009). Hybrid Machine Translation for Low- and Middle- Density Languages. *Language Engineering for Lesser-Studied Languages, S. Nirenburg (ed.),* IOS Press, pp. 243-274.

Marsi, E., Lynum, A., Bungum, L. and Gambäck, B. (2011). Word Translation Disambiguation without Parallel Texts. *International Workshop on Using Linguistic Information for Hybrid Machine Translation,* Barcelona, Spain, pp. 66-74.

Mairson, H. (1992). The Stable Marriage Problem. *The Brandeis Review*, 12:1. Available at: www.cs.columbia.edu/~evs/intro/stable/writeup.html

NIST 2002. Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrences Statistics.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Ma-chine Translation. *40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, USA, pp. 311-318.

Phillips, A. (2011). CUNEI: Open-source Machine Translation with Relevance-based models of each translation instance. *Machine Translation*, Vol. 25, pp. 161-177.

Sanchez-Martinez, F. and Forcada, M.L. (2009). Inferring Shallow-transfer Machine translation Rules from Small Parallel Corpora. *Journal of Artificial Intelligence Research*, Vol. 34, pp. 605-635.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, pp. 44-49.

Smith, T.F. and Waterman, M.S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, Vol. 147, pp. 195-197.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, pp. 223-231.

S. Sofianopoulos, M. Vassiliou & G. Tambouratzis (2012) Implementing a language-independent MT methodology. Proceedings of the First Workshop on Multilingual Modeling, held within the ACL-2012 Conference, Jeju, Republic of Korea, July 13, pp.1-10.

Tambouratzis, G., Troullinos, M., Sofianopoulos, S. and Vassiliou, M. (2012a). Accurate phrase alignment in a bilingual corpus for EBMT systems. Proceedings of the 5th BUCC

Workshop, held within the LREC2012 Conference, May 26, Istanbul, Turkey, pp. 104-111.

Tambouratzis, G., Tsatsanifos, G., Dologlou, I. and Tsimboukakis, N. (2012b). SOM-based corpus modeling for disambiguation purposes in MT. In Proceedings of the Hybrid Machine Translation Workshop, held within the TSD2012 Conference (MTW-2012), Brno, Czech Republic, September 3, pp. 29-36 (ISBN 978-80-263-0266-7).

Tsimboukakis, N. and Tambouratzis, G. (2011). Word map systems for content-based document classification. *IEEE Transactions on Systems, Man & Cybernetics – Part C*, Vol. 41(5), pp. 662-673.