Implementing a language-independent MT methodology

Sokratis Sofianopoulos
ILSP / Athena R.C.
Artemidos 6 & Epidavrou
Athens, Greece
s_sofian @ilsp.gr

Marina Vassiliou
ILSP / Athena R.C.
Artemidos 6 & Epidavrou
Athens, Greece
mvas@ilsp.gr

George Tambouratzis
ILSP / Athena R.C.
Artemidos 6 & Epidavrou
Athens, Greece
giorg_t@ilsp.gr

Abstract

The current paper presents a language-independent methodology, which facilitates the creation of machine translation (MT) systems for various language pairs. This methodology is implemented in the PRESEMT hybrid MT system. PRESEMT has 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 comprises two phases. The first one, **Structure selection**, determines the overall structure of a target language (TL) sentence, drawing on syntactic information from a small bilingual corpus. The second phase, **Translation equivalent selection**, relies on models extracted solely from monolingual corpora to implement translation disambiguation, determine intra-phrase word order and handle functional words. This paper proposes extracting information for disambiguation from the monolingual corpus. Experimental results indicate that such information substantially contributes in improving translation quality.

1 Introduction

Currently most language-independent MT approaches are based on the statistical machine translation (SMT) paradigm (Koehn, 2010). SMT has proved to be particularly amenable 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 (at least several hundreds of thousands of sentences) to allow the building of accurate translation models. Such corpora are hard to find, particularly for less widely used languages. Furthermore, SMT translation accuracy largely depends on the quality of the bilingual corpora as well as their relevance to the domain of text to be translated. For instance, parliament proceedings (among the most widely available corpora) may not suffice to train MT systems aimed towards technical manuals or news articles.

Example-Based Machine Translation (EBMT) is another MT paradigm, where a set of SL sentences are provided together with their TL reference translations. Translations are generated by analogy, where for an input sentence the most similar SL side from the sentence set is determined and the corresponding TL side sentence is used to generate the translation. Hybrid MT systems combining EBMT and SMT techniques have been proposed (cf. Groves & Way, 2005 and Phillips, 2011).

As an alternative to SMT, techniques for creating MT systems using more limited but easily obtainable resources have been proposed. Even if these methods do not achieve an accuracy as high as that of SMT, their ability to develop MT systems with very limited resources confers to them an important advantage. The present article focusses on the development of such a methodology.

Proceedings of the First Workshop on Multilingual Modeling, pages 1–10,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

2 MT systems utilising low-cost resources

A number of methods for the automatic inference of templates for the structural transfer from SL to TL have been proposed. Notably, Caseli et al. (2008) have proposed generating resources such as bilingual transfer rules and, more importantly, shallow transfer rules from parallel corpora. In a related set-up, Sanchez-Martinez et al. (2009) suggest using small parallel corpora only to extract transfer rules, assuming that a sufficient bilingual dictionary is already available. Sanchez-Martinez et al. (2009) report that the MT accuracy is substantially higher for related languages, the proposed method exceeding even SMT systems (for which the parallel corpora used, averaging approximately one million words each, are found to be too small to allow effective linguistic modelling). Both aforementioned approaches have been combined with the Apertium[1] MT system.

Other MT systems have been proposed to cater for the case of low resources. Habash (2003) has proposed the Matador system for translation from Spanish to English, as a typical example of Generation-Heavy Machine Translation (GHMT), where resource poverty in the source language is addressed by exploiting TL resources. Carbonell et al. (2006) propose an MT method that requires no parallel text, 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 using low-cost resources encompasses METIS (Dologlou et al., 2003) and METIS-II (Markantonatou et al., 2009; Carl et al., 2008). These rely solely on extensive monolingual corpora in order to translate SL texts. METIS and METIS-II employ pattern recognition-based algorithms to determine the translation.

3 The PRESEMT system in brief

The architecture of PRESEMT has been formulated on the basis of experience collected within METIS and METIS-II. However, PRESEMT has been substantially modified in order to provide a measurable increase in translation speed and accuracy.

More specifically, in terms of resources, PRESEMT uses a bilingual dictionary providing SL – TL lexical correspondences. It also uses, as does METIS-II, an extensive TL monolingual corpus, which is compiled automatically via web crawling; a small bilingual corpus is yet additionally employed, 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 the source and target languages. Hence, in comparison to SMT systems, the size of the parallel corpus required is reduced by at least three orders of magnitude.

Both the bilingual and the monolingual corpora are annotated[2] 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 section details the kind of information extracted.

3.1 Exploiting the corpora

The processing of the bilingual corpus involves the combined use of two modules, the Phrase aligner module (PAM) and the Phrasing model generator (PMG). Details on PAM and PMG are provided in Tambouratzis et al. (2011), though their operation is summarised here for reasons of completeness.

Initially, the bilingual corpus is aligned at word and phrase level by PAM. PAM aims at circumventing incompatibilities of different annotation tools, based on a learning-by-example principle. It identifies how the SL structure is modified towards the TL one, allowing the deduction of a phrasing model for the source language. To operate, PAM assumes the existence of a parser in TL, which provides chunking information. Based on lexical information combined with statistical data on PoS tag correspondences drawn from the bilingual lexicon, PAM transfers the parsing scheme from the TL side of the corpus (bearing lemma, tag and parsing

[1] www.apertium.org

[2] 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.

information[3]), to the SL side, which is only tagged and lemmatised. In other words, the SL side is segmented into phrases in accordance to the phrasal segmentation provided for the TL side. PAM follows a three-step process, involving (a) lexicon-based correspondences, (b) alignment based on similarity of grammatical features and PoS tag correspondence and (c) alignment guided by already aligned neighbouring words. In each consecutive step, additional SL words are assigned to phrases, but with a reduced accuracy, the aim being for all words to be assigned to phrases.

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.

The TL monolingual corpus serves as the basis for extracting two models, which are employed during the translation process. The first one is used solely for disambiguation purposes (cf. subsection 6.4). The second model provides the micro-structural information on the translation output to support word reordering. It derives from a phrase-based indexing of the TL monolingual corpus, which is performed offline during the pre-processing stage and is based on (i) phrase type, (ii) phrase head and (iii) phrase head PoS tag.

To implement a fast retrieval, the TL phrases are then organised in a hash map that allows the storage of multiple values for each key, using as a key the three aforementioned criteria. For each phrase the number of occurrences within the corpus is also retained. Each hash map is serialised and stored in a file with a unique name for immediate access by the search algorithm.

The number of files created as a result of this process is large, yet each of the files is of small size and thus can be loaded quickly. Furthermore, the existence of a given word in a phrase does not necessarily mean that this phrase will be grouped with other phrases containing the same word, since the model is based on the phrase head.

For the experiments reported here, the TL monolingual corpus is indexed based on the criteria listed above. However, a different indexing scheme may prove more effective, and thus experiments on the optimal indexing are continuing. For instance, the environment of the phrase may also be stored (i.e. the type of the previous and next phrases) and in this case the phrase organisation may be modified. These modifications may yield a decrease in computational load during translation, by reducing the number of phrase comparisons.

3.2 Main translation engine

The translation process is split into two phases, each of which makes use of only a single type of corpus. Phase 1 (**Structure selection**) uses the bilingual corpus to determine, for a given input SL sentence, the appropriate TL structure in terms of phrase type and order. The output of the Structure selection phase is the SL sentence with a TL structure, created by reordering the phrases according to 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 monolingual corpus 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 statistical information, resulting in a hybrid nature.

4 Example of the PRESEMT translation process

In this section the translation process of the PRESEMT system is illustrated via a simple example. Details on the algorithmic part are provided in the subsequent sections.

Input Sentence: 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; output of the lexicon look-up]

Input sentence annotation after being input for translation				
Phrase	**Word**	**Lemma**	**Tag**	**Lexicon**
VC[4]	εδραιώνονται	εδραιώνω	**vb**o3pl	{consolidate;

[3] For the experiments reported here, TreeTagger (Schmid, 1994) was used for the TL processing.

[4] VC: verb chunk, PC: prepositional chunk

Input sentence annotation after being input for translation				
Phrase	**Word**	**Lemma**	**Tag**	**Lexicon**
				establish}
PC	σχέσεις καλής γειτονίας	σχέση καλός γειτονία	**no**feplnm **aj**fesgge **no**fesgge	{relation; relationship} {nice; decent; good} {adjacency; neighbourhood}
PC	στις χώρες των Βαλκανίων	στου χώρα ο Βαλκάνια	**as**feplac **no**feplac **at**neplge **no**neplge	{on; at; to; into; in; upon} {country} {the} {Balkan}

1st translation phase: Establish the correct phrase order on the basis of TL. Search the bilingual corpus for the most similar SL sentence in structural terms, find the corresponding TL one and reorder the input sentence accordingly.

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

Output of 1st transl. phase (expressed as list of phrases and lemmas): [{relation; relationship}; {nice; decent; good}; {adjacency; neighbourhood} $_{PC}$] [{consolidate; establish} $_{VC}$] [{on; at; to; into; in; upon}; {country}; {the}; {Balkan} $_{PC}$]

2nd translation phase: Identify the correct word order within each phrase. Disambiguate the translations. Generate tokens out of lemmas

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

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

Token generation: [{good}; {neighbourhood}; {relations} $_{PC}$] [{are established} $_{VC}$] [{in}; {the}; {Balkan}; {countries} $_{PC}$]

Final Translation: [Good neighbourhood relations $_{PC}$] [are established $_{VC}$] [in the Balkan countries $_{PC}$]

5 Phase 1: Structure selection

The task of Structure selection is to determine 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 & lemma information and segmented into phrases by PMG. A dynamic programming algorithm then determines for each ISS the most similar, in terms of phrase structure, SL sentence found in the bilingual corpus (termed **ACS** – Aligned Corpus Sentence)[5].

The similarity is determined by taking into account structural information such as phrase type, phrase head PoS tag, phrase functional head info and phrase head case. The ISS phrases are then reordered in accordance to the TL side of the chosen ACS by replicating the SL-TL phrase alignment mapping. The data flow of the Structure selection is depicted in Figure 1.

The dynamic programming algorithm is essentially a monolingual similarity algorithm. 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 alignment of DNA and RNA sequences. This algorithm is guaranteed to find the optimal local alignment between two input sequences.

[5] 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) ensures a high similarity score.

5.1 Calculating structural similarity

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 along the top row and the ACS along the left side. A cell (*i,j*) represents the similarity of the sub-sequence of elements up to the mapping of the elements E_i of the ACS and E'_j of the ISS, where each element corresponds to a phrase. 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*), that contain values V1, V2 and V3 respectively, and is calculated iteratively as the maximum of the three numbers {max(V1, V2, V3)+ElementSimilarity(E_i, E'_j)}. The similarity of two phrases (PhrSim) is calculated as the weighted sum of four criteria, namely the similarities of (a) the phrase type (PhrTypSim), (b) the phrase head PoS tag (PhrHPosSim), (c) the phrase head case (PhrHCasSim) and (d) the functional phrase head PoS tag (PhrfHPosSim):

$$PhrSim(E_i,E'_j) = W_{phraseType}*PhrTypSim(E_i,E'_j) + W_{headPoS}*PhrHPosSim(E_i,E'_j) + W_{headCase}*PhrHCasSim(E_i,E'_j) + W_{fheadPoS}*PhrfHPosSim(E_i,E'_j)$$

For normalisation purposes, the sum of the four aforementioned weights (whose experimental values[6] are *0.4*, *0.1*, *0.1* and *0.4* respectively) is equal to *1*. The similarity score ranges from *100* to *0*, these limits denoting exact match and total dissimilarity between elements E_i and E'_j respectively. In case of a zero similarity score, a penalty weight (-50) is employed, to further penalise mapping of dissimilar items.

When the algorithm has reached the j^{th} element of the ISS, the similarity score between the two SL sentences is calculated as the value of the maximum j^{th} cell. The ACS that achieves the highest similarity score is the closest to the input SL sentence in terms of phrase structure.

After determining the similarity between sentences, as the final similarity score, 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 according to the TL side structure.

To illustrate this approach, an example is provided with Greek as SL and English as TL. Let us assume the ISS given in (1):

(1) Με τον όρο Μηχανική Μετάφραση αναφερόμαστε σε μια αυτοματοποιημένη διαδικασία (*"The term Machine Translation denotes an automated procedure"*)

The input sentence is segmented by PMG into the structure depicted in (2a); the structure elements being exemplified in (2b):

(2a) pc(as, no_ac) pc(-, no_ac) vp(-, vb) pc(as, no_ac)

(2b) <Phrase type> (<Phrase fhead PoS tag>, <Phrase head PoS tag>_<Phrase head case>)

An indicative ACS from the aligned corpus is given in (3):

(3) Οι ιστορικές ρίζες της Ευρωπαϊκής Ένωσης ανάγονται στο Δεύτερο Παγκόσμιο Πόλεμο. (*"The historical roots of the European Union lie in the Second World War"*)

The corresponding structural information for (3) is: pc(-,no_nm) pc(-,no_ge) vc(vb) pc(as,no_ac).

			Input source sentence (ISS)			
			pc (as, no_ac)	pc (-, no_ac)	vc (-, vb)	pc (-, no_ac)
		0	0	0	0	0
Aligned corpus sentence (ACS)	pc(-, no_nm)	0	60	80	-20	60
	pc(-, no_ge)	0	60	140	40	40
	vc(vb)	0	-50	10	240	140
	pc(as, no_ac)	0	100	30	-40	**340**

Table 1. Matrix defining phrase correspondence of sentences (1) and (3)

Then, the matrix of Table 1 is created to calculate the similarity scores between sentences (1) and (3) (cells forming the best aligned subsequence are highlighted). By choosing for each element the maximum similarity, the transformation cost is calculated (340 in this case). Based on this matrix, ISS is modified in accordance to the attached TL structure.

[6] An optimisation module has been designed as part of the PRESEMT system for defining the optimal values of these parameters (cf. subsection 5.3 for more details).

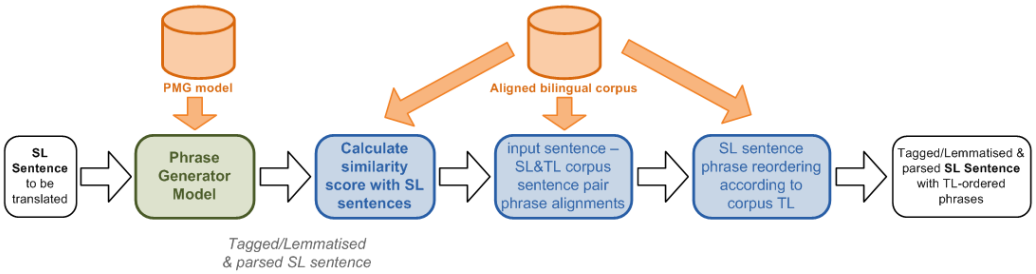

Figure 1. Data flow in Structure selection

6 Phase 2: Translation equivalent selection

Following Phase 1, the issues to be resolved in the second phase include (i) word ordering within phrases, (ii) handling of functional words and (iii) resolution of translation ambiguities.

6.1 Searching for phrasal equivalents

The monolingual TL corpus is searched to determine the most similar phrase to each phrase in the SL sentence, in order to establish the correct word order. The similarity measure takes into account the phrase type, and the words contained in the phrase in terms of lemma, PoS tag and morphological features. These factors enter the comparison with different weights, whose relative magnitudes are subject to an optimisation process.

The main issue at this stage is to reorder appropriately any items 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 (TL) Phrase), are close to each other in terms of number and type. The data flow of the Translation equivalent selection is depicted in Figure 2.

6.2 Establishing correct word order

When initiating Phase 2 of the translation process, the matching algorithm accesses the indexed TL phrase corpus 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 algorithms such as the Gale-Shapley (Gale and Shapley, 1962; Mairson, 1992) and Kuhn–Munkres ones (Kuhn, 1955; Munkres, 1957). The Kuhn-Munkres approach computes an exact solution of the assignment problem to determine the optimal matching between elements. Experiments with METIS-II have shown that the solution of the assignment problem is computationally-intensive.

On the contrary, the Gale-Shapley algorithm solves the assignment problem in a reduced time. In this approach, the two sides are termed suitors (in PRESEMT, the SL side) and reviewers (the TL side). The two groups have distinct roles, suitors proclaiming their order of preference of being assigned to a specific reviewer, via an ordered list. Each reviewer selects one of the suitors after evaluating them based on the ordered preference list, in subsequent steps revising its selection so that the resulting assignment is improved. This process is suitor-optimal but possibly non-optimal from the reviewers' viewpoint. As its complexity is substantially lower than that of Kuhn-Munkres, the Gale-Shapley algorithm is adopted in PRESEMT to limit the computation time.

For each SL phrase, it is necessary to establish the correct word order for all possible TL phrases that can be produced by combining the lexical equivalents of each word in the phrase.

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.

6.3 Optimising the selection process of phrasal equivalents

The search for the most similar phrase depends on a set of parameters. Within this set, different types of weights are included, such as weights governing the similarity of PoS tags, lemmas, phrase types and morphological features. The weights from both

translation phases are handled in a unified manner by the Optimisation module. Research in earlier MT systems has shown that the application of Genetic Algorithms (GAs) and multi-objective evolutionary algorithms such as SPEA2 (Improved Strength Pareto Evolutionary Algorithm) for the optimisation of parameters can considerably improve the translation quality (Sofianopoulos et al., 2010).

For the experiments presented in the next section, manually-defined preliminary weights are used for the parameters of both phases. To further improve the translation accuracy, an optimisation process is studied. This optimisation (which is beyond the scope of the present article) provides the prospect for a substantial improvement in the accuracy via the selection of appropriate parameter values.

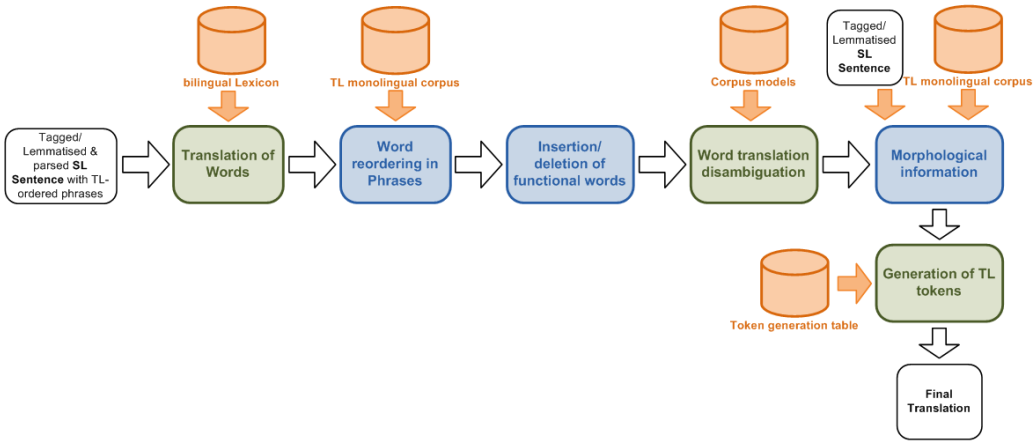

Figure 2. Data flow in Translation equivalent selection

6.4 Resolving translation ambiguities

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 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).

These disambiguation processes lie beyond the scope of the present publication. On the contrary, a simpler, corpus-based approach is proposed 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.

7 Experimental Results

The evaluation results reported here concern the Greek – English language[7] pair and were 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. 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.

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

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). Table 2 summarises indicative scores obtained.

Number of sentences	40	**Source**	web
Reference translations	1	**Language pair**	EL – EN
MT system	**Metrics**		
	BLEU	**NIST**	**Meteor**	**TER**
PRESEMT 1	0.1297	4.1568	0.2669	79.417
PRESEMT 2	0.2004	4.9995	0.3294	72.678
Metis-II	0.1222	3.1655	0.2698	82.878
Google[8]	0.5472	7.1360	0.4713	29.963
Systran[9]	0.3143	5.4615	0.3857	49.449
WordLingo[10]	0.2908	5.1853	0.3728	49.632

Table 2. Evaluation results

When using the base PRESEMT system with the phrase-frequency disambiguation component deactivated (denoted as PRESEMT 1), a BLEU score of 0.1297 and a Meteor score of 0.2669 are obtained. When the disambiguation component is activated (PRESEMT 2), these scores increase substantially, reaching a BLEU score of just over 0.20. The BLEU improvement over PRESEMT 1 is 0.07 points (representing a 50% improvement), while NIST is increased by 0.85 and Meteor by over 0.06. TER is reduced by 7 points, also marking an improvement.

To put these scores into perspective, a comparison is made to MT systems available on the Internet, both rule-based (SYSTRAN) and SMT ones (Google Translate). In addition, the results of METIS-II are quoted, to compare PRESEMT with a system based on monolingual corpora. As can be seen, web-based MT systems produce higher scores for all metrics, with Google Translate possessing the best values.

Yet these scores are, especially in the case of Systran and WordLingo, not far off the scores obtained for PRESEMT with disambiguation. In particular NIST scores are directly comparable whilst the Meteor ones are not substantially higher. It can be reasonably assumed that due to the language-independent methodology without direct provision of language-specific information, the scores obtained via PRESEMT will be lower. Still, it is expected that refined versions of the PRESEMT algorithm will allow the achievement of higher scores that render its performance directly comparable to that of Systran and WordLingo, for the given language pair. In comparison to METIS-II, PRESEMT offers a substantial improvement for all metrics, with for instance BLEU and NIST scores increased by over 50%. This illustrates the improvements conferred by the new translation methodology. 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.

[8] translate.google.com
[9] www.systranet.com
[10] www.worldlingo.com

8 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 using objective evaluation metrics and comparisons to established systems have also 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. These will be reported in future articles.

References

Michael Carl, 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: Background, Implementation, Results and Potentials. *Machine Translation, Vol. 22, No. 1-2*, pp. 67-99.

Jaime Carbonell, Steve Klein, David Miller, Michael Steinbaum, Tomer Grassiany, and Jochen Frey.

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.

Helena M. Caseli, Maria das Gracas V. Nunes, and Mikel L. Forcada (2008) Automatic Induction of Bilingual resources from aligned parallel corpora: Application to shallow-transfer machine translation. *Machine Translation, Vol. 20*, pp. 227-245.

Michael Denkowski and Alon Lavie. 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.

Ioannis Dologlou, Stella Markantonatou, George Tambouratzis, Olga Yannoutsou, Athanasia Fourla and Nikos Ioannou. 2003. Using Monolingual Corpora for Statistical Machine Translation: The METIS System. In Proceedings of the EAMT-CLAW'03 Workshop, Dublin, Ireland, pp. 61-68.

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

Declan Groves & Andy Way, 2005. Hybrid data-driven Models of Machine Translation. Machine Translation, Vol 19, pp.301-323.

Nizar Habash. 2003. Matador: A Large-Scale Spanish-English GHMT System. In *Proceedings of MT Summit IX*, New Orleans, LA, pp. 149-156.

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

Philip Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, Cambridge.

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

John Lafferty, Andrew McCallum and Fernando Pereira. 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.

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

Erwin Marsi, André Lynum, Lars Bungum, and Björn Gambäck. 2011. Word Translation Disambiguation without Parallel Texts. *International Workshop on Using Linguistic Information for Hybrid Machine Translation,* Barcelona, Spain, pp. 66-74.

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

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

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

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

Helmut Schmid. 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.

Temple F. Smith and Michael S. Waterman. 1981. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, *Vol. 147*, pp. 195-197.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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.

Sokratis Sofianopoulos, and George Tambouratzis. 2010. Multiobjective Optimisation of real-valued

Parameters of a Hybrid MT System using Genetic Algorithms. *Pattern Recognition Letters*, *Vol.* 31, pp.1672-1682.

George Tambouratzis, Fotini Simistira, Sokratis Sofianopoulos, Nikos Tsimboukakis, and Marina Vassiliou. 2011. A resource-light phrase scheme for language-portable MT. *15th International Conference of the European Association for Machine Translation*, Leuven, Belgium, pp. 185-192.

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