Establishing sentential structure via realignments from small parallel corpora

George Tambouratzis
ILSP/Athena Res. Centre
6 Artemidos & Epidavrou,
Paradissos Amaroussiou,
Athens, GR-15125, Greece.
giorg_t@ilsp.gr

Vasiliki Pouli
ILSP/Athena Res. Centre
6 Artemidos & Epidavrou,
Paradissos Amaroussiou,
Athens, GR-15125, Greece.
vpouli@ilsp.gr

Abstract

The present article reports on efforts to improve the translation accuracy of a corpus–based hybrid MT system developed using the PRESEMT methodology. This methodology operates on a phrasal basis, where phrases are linguistically-motivated but are automatically determined via a dedicated module. Here, emphasis is placed on improving the structure of each translated sentence, by replacing the Example-Based MT approach originally used in PRESEMT with a sub-sentential approach. Results indicate that an improved accuracy can be achieved, as measured by objective metrics.

1   Introduction

In the present article, a corpus-based methodology is studied, which allows the creation of MT systems for a variety of languages using a common set of software modules. This methodology has been specifically designed to address the scarcity of parallel corpora needed to train for instance a Statistical Machine Translation system (Koehn, 2010), in particular for less widely-resourced languages. Thus the main source of information is a large collection of monolingual corpora in the target language (TL). This collection is supplemented by a small parallel corpus of no more than a few hundred sentences, which the methodology employs to extract information about the structural transfer from the source language (SL) to the target one. The aim in the present article is to investigate how the translation quality can be improved over the best results reported so far (Tambouratzis et al., 2014). Emphasis is placed on extracting the salient information from the small parallel corpus, to most accurately define the structure of the sentences being translated. The efficacy of this effort is verified by a set of experiments.

2   Summary of Translation Process

The PRESEMT methodology studied here is designed to address the very limited availability of parallel corpora (of a few hundred sentences at most) with large amounts of monolingual corpora, and achieve a competitive translation quality without explicit provision of linguistic knowledge. Instead, linguistic knowledge is extracted from the corpora available, via algorithmic means (Sofianopoulos et al., 2012). The parallel corpus comprises a number of aligned sentences (these are referred to as ACS – Aligned Corpus Sentences).

The PRESEMT methodology comprises two phases, which process the text to be translated on a sentence-by-sentence basis. The first phase determines the structure of the translation (Phase 1– Structure selection phase) using the parallel corpus. The second phase rearranges the sequence of tokens in each phrase and decides on the optimal translation of each token (Phase 2 – Translation equivalent selection).

PRESEMT adopts a phrase-based approach, where the phrases are syntactically motivated and the text-to-be-translated is processed on the basis of the phrases contained. These phrases are determined in a pre-processing phase, just before the beginning of the translation process, via the Phrase Model Generator (PMG) module. PMG is trained on the small parallel corpus to port the phrasing scheme from the target language (in which a chunker is available) towards the source language. Thus PMG is able to chunk arbitrary

input sentences into phrases, in the process eliminating the need for a suitable SL chunker.

As a result, in the first translation phase each input sentence (*InS*) is handled as a sequence of phrases. The reordering of these phrases in the TL translation is determined by comparing the *InS* structure to the SL-side structures of all sentences of the parallel corpus. To this end, a Dynamic Time Warping (DTW) algorithm, as discussed in Myers et al. (1980), is used. The DTW implementation chosen is that of Smith et al. (1981), with all comparisons being performed on a phrase-by-phrase basis, on the SL-side. When the best-matching SL-side sentence structure is determined, the structure of the *InS* translation is defined by the corresponding TL-side sentence. This process is summarized in Fig.1.
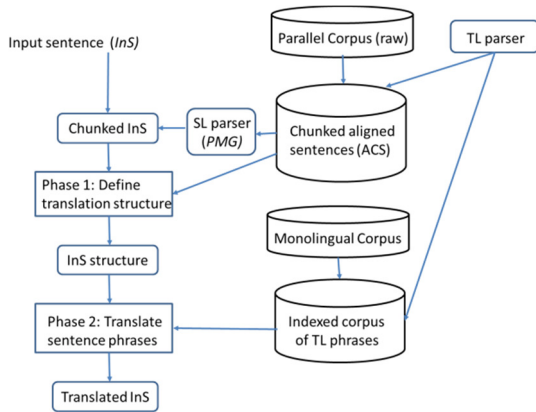

Figure 1: Schematic description of PRESEMT translation methodology.

In turn, Phase 2 samples the indexed monolingual TL corpus to determine the most likely token translations and sequence of tokens within the boundaries of each phrase, using the stable-marriage algorithm (Mairson, 1992). The PRESEMT translation methodology is of a hybrid nature, as Phase 1 is EBMT-inspired (Nagao, 1984 & Hutchins, 2005), while Phase 2 is conceptually much closer to SMT (Brown et al., 1988).

In the present article, emphasis is placed on improving specifically the first translation phase, aiming for an improvement in the resulting quality. The aim is to algorithmically establish realignment rules that cover sub-sentential segments. Conceptually, this possesses similarities to the preordering methods proposed for SMT systems (cf. Lerner et al., 2013 and Stymne, 2009). In (Lerner et al., 2013) preordering is aimed to pre-process the input text so as to render the sequence in SL closer to the sequence in

TL. This can simplify the translation process significantly, and result in improved translation scores. However, a dependency parser in the SL-side is assumed and preordering is performed prior to any processing of the input string to simplify the training of the SMT.

On the contrary, in the present article the aim is to determine sub-sentential re-orderings which are applied within the translation process. Furthermore, in PRESEMT the SL-side parsing scheme is induced via the TL-side shallow parser and thus is not sufficiently detailed to provide subject-object relationships or to determine dependencies as required by preordering algorithms. Finally, as the parallel corpus only numbers a few hundred sentences, the SL-side information is not sufficiently extensive to support the extraction of large numbers of rules as reported by e.g. Stymne (2009).

3 Porting the Sentence Structure from SL to TL

The structure selection phase serves to determine for each sentence to be translated the structure of the translation. For each phrase in the sentence, the following tuple is created:

$$phr_i = [phr\_type_i; pos(head_i); case(head_i)] \quad (1)$$

where phr_type_i indicates the phrase-type of the ith phrase, $pos(head_i)$ is the Part-of-Speech (PoS) tag of the phrase head, and $case(head_i)$ is the case of the phrase head. Then the kth sentence is expressed as an ordered sequence of j tuples:

$$struct(sent_k) = [phr_1; phr_2; ..; phr_j] \quad (2)$$

To determine the optimal structure of the translated sentence *InS*, the information existing in the small parallel corpus of *N* sentences is exploited. More specifically, this corpus contains a number of *N* aligned sentences ACS (Aligned Corpus Sentences), for which the SL and the TL sentence are direct equivalents of each other (denoted as *ACS_SL* and *ACS_TL* respectively). Then, the structure of the *InS* translation is the one of the cth sentence pair ACS, for which the following expression is maximized:

$$\max_{1 \le i \le N} \{simil(struct(InS), struct(ACS\_SL_i)\} \quad (3)$$

In (3), *simil* expresses the phrase-wise structural similarity between sentences *InS* and *ACS_SL* determined via a phrase-by-phrase comparison (as discussed in Sofianopoulos at el., 2012). The similarity of two phrases is calculated as the weighted sum of three constituent similarities, (a) the phrase type, (b) the phrase head PoS tag and (c) the grammatical case of the phrase head.

3.1 Analysing translation problems

An analysis of PRESEMT results has shown that a large proportion of translation errors are due to the first phase of the translation process. In such cases, the structure of the translation of an input sentence *InS* fails to be accurately determined, and the translation quality suffers accordingly. It should be noted that within the PRESEMT system, the diversity of sentences in the parallel corpus is limited. Due to the limited size of the parallel corpus, the number of archetypes supporting the transfer of structures from SL to TL is smaller than is desirable.

To indicate the effect of the limited coverage provided by the restricted number of parallel SL-TL sentence pairs, an example is provided in Figure 2, based on the Greek-to-English translation pair. In this example only the phrase types are quoted, without any additional differentiation (such as case or PoS of the phrase head). In this example, the original sentence in Greek is shown in (1) while in (2) and (3) the translation is shown in terms of lemmas and tokens, when using the standard Structure selection algorithm of PRESEMT. In (4) and (5), the translation using our proposed novel structure selection algorithm is shown, which has an improved structure.

(1) Initial Sentence	
PC2(O πατέρας της) VC6(προσπαθεί) ADVC8(μάταια) να PC11(τη) VC13(μεταπείσει) .	
Standard Structure Selection	
(2) Lemmas	**(3) Tokens**
PC2(her father) VC6(try) ADVC8(in vain) to PC11(her) VC13(dissuade).	Her father tries in vain to her dissuade.
Proposed Structure Selection	
(4) Lemmas	**(5) Tokens**
PC2(her father) VC6(try) ADVC8(in vain) to VC13(dissuade) PC11(her).	Her father tries in vain to dissuade her.

Figure 2: translation of input sentence (1) as generated by the standard PRESEMT structure selection [cf.(2) and (3)] & the new one [cf.(4) and (5)] in terms of lemmas and tokens respectively.

For this language pair, four phrase types exist (namely ADVC, PC, VC and ADJC), as determined by the Treetagger (Schmid, 1995) version for the English language. To simplify the analysis, it is assumed that all sentences have a fixed structure size k (that is, they comprise exactly k phrases each). Then, the number of possible combinations, N_{phr} depends on the number of phrase types *ptype* (not taking into account linguistic constraints that may render certain combinations ungrammatical):

$$N_{phr} = ptype^k \qquad (4)$$

In the case of only four phrase types and a sequence of ten phrases, the number of combinations as determined by eqn. (1) is 4^{10}, which is approximately equal to 10^6. However, the size of the parallel corpus in PRESEMT is typically constrained to only 200 sentences. Consequently, for the EBMT approach used by PRESEMT, the maximum number of possible structural transformations from SL to TL is at most 200. In reality there are bound to be identical entries within the structures of the aligned sentences, ACS, with more than one sentence pairs having the same structure in terms of phrase sequences in both SL and TL. For instance, in the default parallel corpus of 200 sentences used in the Greek-to-English PRESEMT system, the actual number of unique SL/TL phrase-based structures (defined as a sequence of phrase types) is approximately 100. Hence, the population of archetypes covers the pattern space much more sparsely than is ideal, and the likelihood of a representative exemplar existing in the small parallel corpus is very low.

On the basis of this observation, it is expected that for several input sentences *InS* a sub-optimal match will be established, as either no satisfactory match can be found or only a partial match will be determined, with conflicts occurring for one or more phrases. For instance, for a given 4-phrase input sentence with structure {PC ; VC ; ADVC; PC} the closest match may well be archetype {PC ; VC ; PC ; PC}, resulting in a mismatch at the third element. As a result the structure of the translation is defined by making arbitrary approximations, due to the constrained corpus size. If the proportion of mismatches is very high, it might be preferable to disregard the structural transformations indicated by the chosen template as they are probably inaccurate.

3.2 Replacing the classic structure selection

Based on the aforementioned discussion, the question becomes how to use more effectively the information inherent in the small parallel corpus of SL-TL sentences, to determine realignments when translating sentences from SL to TL. In the original Structure selection algorithm an EBMT-type algorithm (Nagao, 1984) is used where a single sentence from the parallel corpus defines the structure of the translation, implicitly assuming an appropriate coverage of the pattern space. An alternative approach is investigated in the present article, where from each sentence pair of the parallel corpus, knowledge is extracted about realignments of phrases when transferring a sentence from SL to TL. Thus sub-sentential templates (hereafter termed **realignment templates**) are created which describe the necessary reorderings of phrases for relatively short sequences in preference to longer templates that operate on the entire sentence (as is the case in the standard Structure selection). The aim then becomes to extract a representative set of such templates that is applicable to the large majority of sentences.
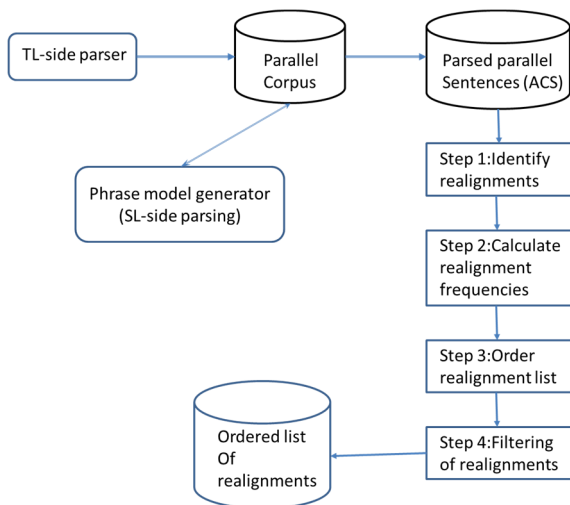

Figure 3: Sequence of steps to extract realignment templates from a parallel corpus.

The underlying assumption of this approach is that when translating from SL to TL, the structure should not be modified, in the absence of evidence that a realignment is required. Hence, the aim is to determine templates that model phrase realignments between the SL and TL sides, and which are then applied to the input sentence structure depending on certain criteria. An example of such phenomena includes the subject which in Greek may follow the corresponding verb chunk, though in English this order is reversed. What is needed is to determine realignments that consistently occur when transitioning from SL to TL and to estimate their corresponding likelihood. Regarding the linguistic resources available, this information may only be extracted from the small parallel corpus. The criteria for estimating the likelihood comprise:

-       the length of the realignment template in terms of phrases,
-       the frequency of occurrence of the template in the small parallel corpus.

A different realignment template is defined for each reordering of phrase sequences, from the SL and TL side sentences. To support direct comparisons with earlier results, each phrase is defined by the phrase type (e.g. verb phrase or noun phrase), the phrase head part-of-speech (PoS) tag and its case (if this exists for the given language). Of course, additional characteristics may also be chosen, depending on the specific language pairs studied, to attain a better performance.

The outline of the algorithm to create a set of realignment templates is depicted in Figure 3. Initially (in Step 1) every parallel sentence pair is scanned to find phrase realignments, and each realignment is recorded in a list. Then (in Step 2), identical realignments (where sequences of phrases in both SL and TL match exactly) are assimilated to record the frequency of occurrence of each realignment template. In Step 3, a heuristic is used to score each one of the templates and a new ordered list of templates is created. Based on the heuristic, a higher score indicates a higher likelihood of correct activation. Finally, a filtering Step 4 is used to eliminate templates that are considered unlikely to be correct, based on their frequency of occurrence in the parallel corpus (more details on the heuristic function and filtering process are provided in sub-section 3.4). The resulting list of templates is then used to define the structure of the input sentence *InS* when translated, by consecutively trying to apply each of the realignment templates, one at a time, starting with the highest-ranked one, as discussed in the next sub-section.

3.3 Application of realignment templates

When ordering the realignment templates, two distinct cases are defined, depending on whether context beyond the realignment template is taken into account. In the first case, the algorithm identifies the realignment template by finding only the sequence of phrases that are realigned, with-

out taking into account the identities of any neighbouring phrases (this being denoted as "Align-nC", where nC stands for No Context).

In the second case, the context of the realignment template to the left and the right is also considered. Thus, one additional phrase to the left and to the right is recorded within an extended realignment template. In this approach, three distinct variants are considered, depending on the degree of the context match. More specifically:

a. If the left and right contexts need to fully-match (i.e. the type of phrases need to coincide but the PoS tag and case of the phrase heads also need to agree), this is termed as **type-0** (and is denoted as "Align-C0", where C0 stands for Context-type-0). This type of match is the most restrictive as it requires matching of all characteristics but on the other hand it allows for more finely-detailed matching.

b. If the left and right context need to be matched only in terms of the type of phrases (but not the PoS tag and case of the phrase heads), this is termed as **type-1** (denoted as "Align-C1"). In contrast to context phrases, for the phrases within the realignment templates, matching extends to both the PoS tag and case of the phrase head. The alignment of type-1 is thus relaxed in comparison to that of type-0 in terms of context-defining phrases, allowing a potentially larger number of matches of the alignment template to the parallel corpus, as observed in Table 1.

c. If for both the context and realignment phrases, only the phrase-type is required to match, (i.e. not the head PoS tag or its case) then this is termed as **type-2**, (denoted as "Align-C2") and corresponds to the least restrictive match in terms of the context, giving the largest number of matches, as seen in Table 1. However, this relaxation in matching might allow for realignment cases where the PoS tags of the phrases and their neighbouring ones do not match, thus resulting in lower translation accuracy.

Table 1: increase of realignments detected for different variations with context in comparison to the no-context (Align-nC) case, for the standard Greek-to-English parallel corpus.

Realignment variations			
Align-nC	Align-C0	Align-C1	Align-C2
+0%	+75%	+150%	+825%

Two examples of realignment templates extracted from a small parallel corpus are depicted in Figure 4. For a specific realignment, the different realignment templates extracted with and without context are depicted in Figure 5.
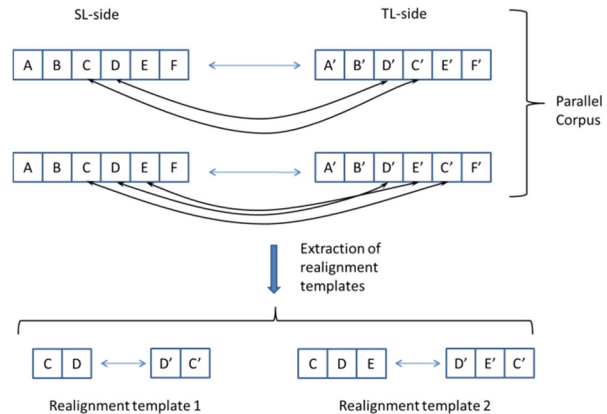

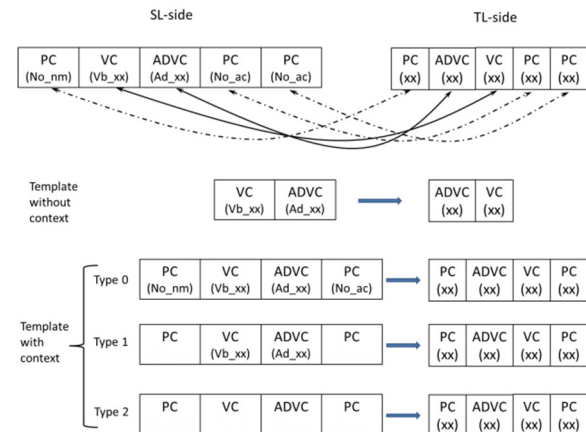Figure 4: Examples of realignment templates.

Figure 5: Types of realignment template without context and with context – the parallel SL/TL sentence is shown on top, followed by the different templates that can be extracted.

The optimal matching depends of course on the characteristics of the language pair being handled, as well as the amount of training data available. Thus more discriminative templates can be established, provided that the appropriate amount of training data is available. Else, it is likely that most templates will only be encountered once, and effectively a look-up table will be established for realignment templates found within the parallel corpus. In this case, no generalization by the system will be possible and the translation accuracy can be expected to suffer. Comparative performances of the aforementioned variants will be discussed in the experimental results' section.

3.4 Heuristic function for ranking realignment templates

The heuristic function has a key role in determining the system behavior, by defining the appropriate ranking of the templates. As the system attempts to iteratively match the sequence of phrases in the chunked input sentences with each realignment template, it first applies the highest-ranked templates and progressively moves to lower-ranked ones. A lower-ranked template is applied to a specific set of phrases provided that no higher-ranked template has been applied to any of these phrases. Thus, the ranking dictates the selection of one realignment template over another, and can affect the accuracy of the translation structure.

Based on a preliminary study, it was decided to rank higher realignment templates which occur more frequently within the given training set (parallel corpus). Also, the application of larger templates is preferred over smaller ones. The actual heuristic function chosen for translation simulations is expressed by equation (5):

$$score_i = freq_i + a_1 * len_i \qquad (5)$$

Where $score_i$ is the score of the *i-th* realignment template, $freq_i$ corresponds to the frequency of occurrence of the template in the training corpus and len_i is the length of the realignment template in terms of phrases. Parameter a_1 is used to weigh the two factors appropriately.

In addition, a number of constraints serve to eliminate cases where potentially spurious realignments may be chosen as valid ones. These constraints have been developed by studying initial translation results. For this description, a realignment between the SL and TL-sides is defined as $rseq_{SL-TL}$. On the other hand, $rseq_{SL}$ is used to denote only the part of the realignment in the SL-side of the parallel corpus.

Constraint 1: If a realignment involving a sequence of phrases $rseq_{SL-TL}$ is encountered very infrequently in comparison to the occurrences of the sequence in the SL-side of the parallel corpus, $rseq_{SL}$, then it is rejected. The aim of this constraint is to eliminate unlikely realignments, which are not applicable for the majority of SL-side patterns [1]. This is expressed by (6), where $freq\_thres$ is a user-defined threshold:

$$\frac{freq(rseq_{SL-TL})}{freq(rseq_{SL})} \geq freq\_thres \qquad (6)$$

Constraint 2: If a realignment $rseq_{SL-TL}$ occurs in the parallel corpus only very rarely, then it is removed from the list of applicable realignments. This is implemented by setting a minimum threshold value *min_freq* for a realignment template to be retained, allowing the reorderings that are rarely applied to specific phrase sequences to be filtered out.

Constraint 3 (hapax legomena): This constraint refines the elimination process of realignments dictated by Constraint 2. More specifically, it introduces an exception to Constraint 2, to prevent certain realignment templates from being filtered-out. If the filtering-out concerns a sequence $rseq_{SL-TL}$ that appears only once in the parallel corpus SL-side, Constraint 3 is activated to retain this rare realignment.

4 Experiments

4.1 Experimental Setup

In the present article, the Greek-to-English language pair is used for experimentation. To ensure compatibility with earlier results, the standard language resources of PRESEMT are used, including the basic parallel corpus of 200 sentences and the two test sets of 200 sentences each, denoted as testsetA and testsetB (all these resources have been retrieved from the www.presemt.eu website). Regarding the parameters related to the realignment templates, the value used for *freq_thres* is 0.50, while *min_freq* is set to 3. Finally, parameter a_1 of eqn (5) is set to 100 for the given experiments, indicating a strong preference to larger realignment templates. These parameter values have been chosen by performing trial simulations during the development phase.

Different PMG modules resulting in different phrase sizes have been studied to investigate alternative SL phrasing schemes applied on the sentences to be translated. This test is performed, to determine whether the proposed realignment method is robust. Comparative evaluation with a selection of PMG modules with different phrase sizes can indicate the effectiveness of realignment templates in this MT methodology. Experiments are performed by considering or not the context (cases: Align-nC, Align-C) or by varying

---

[1] In contrast to training, where both SL and TL information is available, during operation only the SL-side pattern is available and the TL-side one is unknown.

Thus an infrequent realignment cannot be relied upon to provide structure-defining information.

the type of match when Align-C is applied (cases: Align-C0, Align-C1, Align-C2). The best realignments have been compared to the baseline i.e. the case when the classic Structure selection algorithm is used (Tambouratzis et al., 2014).

Regarding the PMG modules, the first version, termed PMG-s gives the highest reported translation accuracy (Tambouratzis, 2014), splitting sentences into smaller phrases[2]. The alternative PMG (PMG-b) evaluated, favours larger phrases than PMG-s and results in smaller average sentence lengths expressed in terms of phrases. The average sentence sizes for each phrasing scheme in both testsets can be seen in Table 2, while the numbers of realignment templates applied to the input sets for testsets A and B are detailed in Table 3. The difference in realignments between the two testsets reflects the fact that TestsetA has smaller sentences of on average 15.3 words per sentence, while for TestsetB this is 22.6 words (the sentence size being increased by 48% in terms of words). Hence, the occurrence of realignments is higher for TestsetB.

Table 2: Average sentence sizes in terms of phrases for the two evaluation testsets when different PMG modules are applied.

	PMG-s	PMG-b
TestsetA	6.90	6.21
TestsetB	10.64	9.66

Table 3: Total number of realignments recorded per testset, for different realignment variations.

TestsetA	Realignment variations			
	Align-nC	Align-C0	Align-C1	Align-C2
PMG-s	4	7	10	37
PMG-b	11	11	14	38

TestsetB	Realignment variations			
	Align-nC	Align-C0	Align-C1	Align-C2
PMG-s	7	9	18	76
PMG-b	7	10	16	59

MT setups are evaluated regarding the translation quality, based on a selection of widely-used MT metrics: BLEU (Papineni et al., 2002), NIST (NIST, 2002), Meteor (Denkowski et al., 2011) and TER (Snover et al., 2006). For BLEU, NIST and Meteor, the score measures the translation accuracy and a higher score indicates a better translation. For TER the score counts the error rate and thus a lower score indicates a more successful translation. For reasons of uniformity, when comparing scores, an improvement in a metric is depicted as a positive change (for all metrics, including TER).

4.2 Experimental Results

Fig. 6 depicts the BLEU scores for different phrasing schemes when the different cases and variations of the realignments are applied (cases: Align-nC, Align-C0, Align-C1, Align-C2). As observed, the PMG-s variant achieves the highest score in general and especially when neighboring phrases are not taken into account (Align-nC BLEU score = 0.3626), thus not limiting the realignments to specific environments.
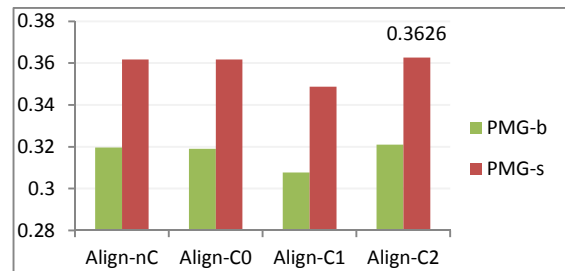

Figure 6: BLEU scores for the different realignment cases, for both PMGs and Testset A.

Figure 7 indicates how translation quality is improved when the best realignment case (Align-nC) is applied compared to the baseline, for different PMGs, using TestsetA. The use of realignments improves metric scores in both PMGs, indicating the improved robustness of the MT system towards this choice. The highest improvement of 1.63% observed for the BLEU score is obtained with PMG-s, which leads to sentences of larger length (with more phrases but of fewer words each).

When applying the best realignment variant (Align-nC) to TestsetB with the two phrasing schemes (i.e. PMG-s, PMG-b) a substantial improvement is achieved, reaching 1.12% for BLEU (cf. Figure 8). As before, PMG-s achieves the greatest improvement, showing that the realignment template algorithm benefits to a greater degree phrasing schemes that generate larger numbers of phrases per sentence.

---

[2] Based on the experiments reported in (Tambouratzis, 2014) both PMG-s and PMG-b correspond to criterion C_F, the differentiation being that PMG-b (PMG-s) allows (prevents) the formation of phrases containing multiple cases.
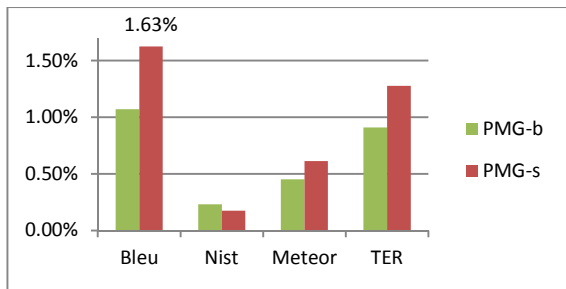
Figure 7: Improvement in scores for the Align-nC case compared to the baseline, for the two phrasing schemes when applied to TestsetA.

A further evaluation effort has involved examining how the proposed realignment template method compares to a zero-baseline, where the SL structure is retained without change in TL. In this case, the improvement amounts to 0.53% in terms of the BLEU score.
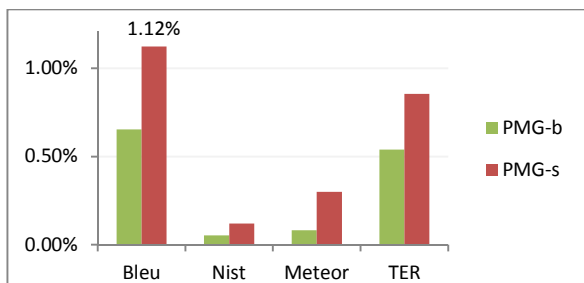

Figure 8: Improvement in scores for Align-nC case compared to the baseline, when the two phrasing schemes are applied to TestsetB.

To compare against another benchmark, TestSetA was translated with a MOSES-based SMT (trained with a parallel corpus of approx. 1.2 million sentences - the parallel corpus is 4 orders of magnitude larger than that used by PRESEMT) and resulted in BLEU and NIST scores of 0.3795 and 7.039 respectively. These MOSES scores are comparable to the scores achieved by PRESEMT with Align-nC (0.3626 and 7.086 for BLEU and NIST respectively).

4.3   Statistical Analysis of Results

To determine whether the results are statistically significant, paired-sample T-tests were applied at a sentence level. Comparing the use of realignment templates with and without context (Align-nC versus Align-C2), the scores for each of the 200 sentences were used to form two distinct populations for TestsetA. By comparing the two populations, for both PMG-b and PMG-s, a statistically significant difference is found at a confidence level of 95%, showing that Align-nC

gives a significantly better translation quality over Align-C2. On the other hand, the improvement of Align-nC compared to the baseline scenario is small, thus not resulting in statistically significant differences.

5   Conclusions and Future Work

The proposed method of applying realignments to sentence structure has been shown to provide a useful increase in translation accuracy over the best configurations established in earlier experiments. Still, a number of possible extensions of the work presented here have been identified. These focus primarily on how to extract a more comprehensive set of templates from the limited-size parallel corpus available. To achieve this, one method would be to integrate linguistic knowledge. For instance, by identifying grammatical categories (i.e. different PoS tags) which are equivalent, it is possible to extend knowledge to introduce new realignment templates based on known ones and thus cover more cases.

Also, it is possible to concatenate different realignment templates to larger groups, in order to make more accurate calculations of the statistics underlying each template. For instance, it may be assumed that whether the PoS tag of the phrase head is a noun or pronoun, the template remains the same and such cases can be grouped together. By extrapolating these new templates, an increase in the pattern space coverage can be expected, leading to an improved translation accuracy.

A point which is of interest is applicability to other language pairs. As is the case for the PRE-SEMT MT methodology as a whole, a key decision was not to design the methodology for one specific language pair. For instance, initial experimentation has shown that the application of realignment templates has correctly generated templates for the case of split verbs when German is the TL (here the Greek-to-German language pair). This is important, as split verbs have been identified as one of the key problems when translating into German. Of course, more experimentation is needed in terms of the generalisation abilities of such realignment templates to cover more cases than those encountered in the training set, and to efficiently model the shift of the second part of the verb to the end of the relevant sentence. Still, the ability of the proposed realignment template method to identify such occurrences is promising.

References

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, Robert L. Mercer, P. Roossin. 1988. A statistical approach to language translation. Proceedings of *COLING'88* Conference, Vol. 1, pp. 71–76.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems, Proc. of the 6th Workshop on Statistical Machine Translation, Edinburgh, Scotland, UK, July 30–31, pp. 85–91.

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

Harry Mairson. 1992. The Stable Marriage Problem. The Brandeis Review, Vol.12, No.1. Available at: http://www.cs.columbia.edu/~evs/intro/stable/writeup.html

Cory S. Myers, Lawrence R. Rabiner, and Aaron E. Rosenberg. 1980. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 6, pp. 623-635.

Philipp Koehn. 2010. Statistical machine translation. Cambridge University Press, Cambridge, UK [ISBN: 978-0-521-87415-1].

Uri Lerner, and Slav Petrov. 2013. Source-Side Classifier Preordering for Machine Translation. In Proceedings of EMNLP-2013 Conference, Seattle, USA, October 2013, pp.513-523.

Makoto Nagao. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. Artificial and human intelligence: edited review papers presented at the international NATO Symposium, October 1981, Lyons, France; A. Elithorn and R. Banerji (eds.), Amsterdam: North Holland, pp. 173-180.

NIST. 2002. Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrences Statistics (Report). Available at: http://www.itl.nist.gov/iad/mig/tests/mt/doc/ngram-study.pdf

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, U.S.A., pp. 311-318.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop.* Dublin, Ireland.

Temple F. Smith, 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. Proceedings of the 7th AMTA Conference, Cambridge, MA, USA, pp.223-231.

Sokratis Sofianopoulos, Marina Vassiliou, and George 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, 13 July, pp.1-10.

Sara Stymne. 2012. Clustered Word Classes for Preordering in Statistical Machine Translation Proceedings of the 13th EACL Conference, April 23-27, Avignon, France, pp.28-34.

George Tambouratzis. 2014. Conditional Random Fields versus template-matching in MT phrasing tasks involving sparse training data. Pattern Recognition Letters, Vol. 53, pp.44-52.

George Tambouratzis, Sokratis Sofianopoulos, and Marina Vassiliou. 2014. Expanding the Language model in a low-resource hybrid MT system. In Proceedings of the SSST-8 Workshop, held within EMNLP-2014, 25 October 2014, Doha, Qatar, pp. 57-66.