Comparing CRF and template-matching in phrasing tasks within a Hybrid MT system

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

Abstract

The present article focuses on improving the performance of a hybrid Machine Translation (MT) system, namely PRESEMT. The PRESEMT methodology is readily portable to new language pairs, and allows the creation of MT systems with minimal reliance on expensive resources. PRESEMT is phrase-based and uses a small parallel corpus from which to extract structural transformations from the source language (SL) to the target language (TL). On the other hand, the TL language model is extracted from large monolingual corpora. This article examines the task of maximising the amount of information extracted from a very limited parallel corpus. Hence, emphasis is placed on the module that learns to segment into phrases arbitrary input text in SL, by extrapolating information from a limited-size parsed TL text, alleviating the need for an SL parser. An established method based on Conditional Random Fields (CRF) is compared here to a much simpler template-matching algorithm to determine the most suitable approach for extracting an accurate model. Experimental results indicate that for a limited-size training set, template-matching generates a superior model leading to higher quality translations.

1 Introduction

Most current MT systems translate sentences by operating at a sub-sentential level on parallel corpora. However, this frequently necessitates parsers for both SL and TL, which either (i) develop matched segmentations that give similar outputs in terms of phrasing over the SL and TL or (ii) for which a mapping is externally defined between the two given segmentations. Both alternatives limit portability to new languages, due to the need for matching the appropriate tools. Another limitation involves the amount of parallel texts needed. Statistical MT (SMT) (Koehn, 2010) generates high quality translations provided that large parallel corpora (of millions of words) are available. However, this places a strict constraint on the volume of data required to create a functioning MT system. For this reason, a number of researchers involved in SMT have recently investigated the extraction of information from monolingual corpora, including lexical translation probabilities (Klementiev et al., 2012) and topic-specific information (Su et al., 2012).

A related direction in MT research concerns hybrid MT (HMT), where principles from multiple MT paradigms are combined, such as for instance SMT and RBMT (Rule-based MT). HMT aims to combine the paradigms' positive aspects to achieve higher translation accuracy. Wu (2009) has studied the trend of convergence of MT research towards hybrid systems. Quirk et al. (2007) have proposed an HMT system where statistical principles are combined with Example-Based MT (EBMT) to improve the performance of SMT.

The PRESEMT (www.presemt.eu) methodology (Tambouratzis et. al, 2013) supports rapid

Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 6–13,
Gothenburg, Sweden, April 27 2014.

development of hybrid MT systems for new language pairs. The hybrid nature of PRESEMT arises from the use of data-driven pattern recognition algorithms that combine EBMT techniques with statistical principles when modelling the target language. PRESEMT utilises a very small parallel corpus of a few hundred sentences, together with a large TL monolingual one to determine the translation. The MT process encompasses three stages:

Stage 1: this pre-processes the input sentence, by tagging and lemmatising tokens and grouping these tokens into phrases, preparing the actual translation.

Stage 2: this comprises the main translation engine, which in turn is divided into two phases:

Phase A: the establishment of the translation structure in terms of phrase order;

Phase B: the definition of word order and the resolution of lexical ambiguities at an intra-phrase level.

Stage 3: post-processing, where the appropriate tokens are generated from lemmas.

In terms of resources, PRESEMT requires:

(i) a bilingual lemma dictionary providing SL to TL lexical correspondences,

(ii) an extensive TL monolingual corpus, compiled via web crawling to generate a language model,

(iii) a very small bilingual corpus.

The bilingual corpus provides examples of the structural transformation from SL to TL. In comparison to SMT, the use of a small corpus reduces substantially the need for locating parallel corpora, whose procurement or development can be extremely expensive. Instead, a small parallel corpus can be assembled with limited recourse to costly human resources. The small size of the parallel corpus unavoidably places additional requirements on the processing accuracy in order to extract the necessary information. The main task studied here is to extract from a parallel corpus of 200 sentences appropriate structural information to describe the transformation from SL to TL. More specifically, a module needs to be trained to transfer a given TL phrasing scheme to SL, so that during translation the module segments arbitrary input text into phrases in a manner compatible to the TL phrasing scheme. The question then is which method succeeds in extracting from the parallel corpus the most accurate structural knowledge, to support an effective MT system.

For transferring a TL phrasing scheme into SL, PRESEMT relies on word and phrase alignment of the parallel corpus. This alignment allows the extrapolation of a model that segments the SL text. The SL–side segmentation is limited to phrase identification, rather than a detailed syntactic analysis.

The processing of a bilingual corpus and the elicitation of the corresponding SL-to-TL phrasing information involves two PRESEMT modules:

(i) The Phrase aligner module (PAM), which performs text alignment at word and phrase level within the parallel corpus. This language-independent method identifies corresponding terms within the SL and TL sides of each sentence, and aligns the words between the two languages, while at the same time creating phrases for the non-parsed side of the corpus (Sofianopoulos et al., 2012).

(ii) The Phrasing model generator (PMG), which elicits a phrasing model from this aligned parallel corpus. PMG is trained on the aligned parallel SL – TL sentences incorporating the PAM output to generate a phrasing model. This model is then employed to segment user-specified text during translation.

A number of studies relevant to this article involve the transfer of phrasing schemes from one language to another. These studies have focussed on extrapolating information from a resource-rich to a resource-poor language. Yarowski et al. (2001) have used automatically word-aligned raw bilingual corpora to project annotations. Och and Ney (2004) use a two-stage process via a dynamic programming-type algorithm for aligning SL and TL tokens. Simard et al. (2005) propose a more advanced approach allowing non-contiguous phrases, to cover additional linguistic phenomena. Hwa et al. (2005) have created a parser for a new language based on a set of parallel sentences together with a parser in a frequently-used language, by transferring deeper syntactic structure and introducing fix-up rules. Smith et al. (2009) create a TL dependency parser by using bilingual text, a parser, and automatically-derived word alignments.

2 Basic functionality & design of phrasing model generator

The default PMG implementation (Tambouratzis et al., 2011) adopts the CRF model (Lafferty at el., 2001, Wallach, 2004) to chunk each input

sentence into phrases. Earlier comparative experiments have established that CRF results in a higher accuracy of phrase detection than both probabilistic models (such as HMMs) and small parsers with manually-defined parsing rules. CRF has been used by several researchers for creating parsers (for instance Sha and Pereira, 2003, Tsuruoka et al., 2009).

Due to the expressiveness of the underlying mathematical model, CRF requires a large number of training patterns to extract an accurate model. Of course, the volume of training patterns is directly dependent on the size of the parallel corpus available. A more accurate CRF would require the use of a large parallel corpus, though this would compromise the portability to new language pairs. Even by moving from handling lemmas/tokens to part-of-speech tags when training the parser, to reduce the pattern space, it is hard to model accurately all possible phrase types via CRF (in particular for rarer PoS tags) via the small corpus. On the contrary, a lower complexity PMG model (hereafter termed **PMG-simple**) may well be better suited to this data. The work presented here is aimed at investigating whether a simpler PMG model can process more effectively this limited-size parallel corpus of circa 200 parallel sentences.

3 Detailed description of PMG-simple

3.1 PMG-simple Principles

PMG-simple follows a learn-by-example concept, where, based on the appearance of phrase patterns, the system learns phrases that match exactly patterns it has previously encountered. This approach is based on the widely-used template-matching algorithm (Duda et al., 2001), where the aim is to match part of the input sentence to a known phrase archetype. PMG-simple (i) does not generate an elaborate high-order statistical model for segmentation into phrases taking into account preceding and ensuing tag sequences, and (ii) cannot revise decisions so as to reach a global optimum. Instead, PMG-simple implements a greedy search algorithm (Black, 2005), using an ordered list of known phrases. Due to its simple design, it suffers a number of potential disadvantages in comparison to CRF-type approaches:

- PMG-simple only identifies exact matches to specific patterns it has previously seen (with some exceptions, as discussed below). On the contrary, more sophisticated approaches may extrapolate new knowledge. For example, let us assume that 'Aj', 'At' and 'No' represent PoS tags for adjectives, articles and nouns respectively, while 'Ac' indicates the accusative case. Then, if noun phrases (NP) [AjAc; AjAc; NoAc] and [AtAc; AjAc; NoAc] are seen in training, the unseen pattern [AtAc; AjAc; AjAc; NoAc] may be identified as a valid NP by CRF but not by PMG-simple.
- PMG-simple does not take into account the wider phrase environment in its decision.
- PMG-simple, as a greedy algorithm, does not back-track over earlier decisions and thus may settle to sub-optimal solutions.

Conversely, PMG-simple has the following advantages:

- As it relies on a simple learn-by-example process, all segmentation decisions are easily explainable, in contrast to CRF.
- The template-matching model is trained and operates much faster than CRF.
- Finally, modifications can be integrated to improve the base algorithm generalisation. These largely consist of incorporating linguistic knowledge to allow the template-matching approach to improve language coverage and thus address specific problems caused by the limited training data.

3.2 PMG-simple Steps

PMG-simple receives as input the SL-side sentences of a bilingual corpus, segmented into phrases. Processing consists of four main steps:

- Step 1-Accumulate & count: Each sentence of the bilingual corpus is scanned in turn, using the phrases of the SL-side as training patterns. More specifically, all SL-side occurring phrases are recorded in a phrase table together with their frequency-of-occurrence in the corpus.

- Step 2-Order: The table is ordered, based on an ordering criterion so that phrases with a higher likelihood of correct detection are placed nearer the top of the phrase table. As a consequence, matches are initially sought for these phrases.

- Step 3-Generalise: Recorded phrases are generalised, to increase the phrase table coverage. Thus, new valid templates are incorporated in the phrase table, which are missing from the limited-size training corpus. Currently, generali-

sation involves extending phrases for which all declinable words have the same case, to other cases. For instance, if NP [AtAc; AjAc; NoAc], with all tokens in accusative exists in the phrase table with a given score, NPs are also created for nominative, genitive and vocative cases ([AtNm; AjNm; NoNm] [AtGe; AjGe; NoGe] and [AtVo; AjVo; NoVo]), with the same score.

- Step 4-Remove: Phrases containing patterns which are grammatically incorrect are removed from the phrase table. As an example of this step, phrases involving mixed cases are removed in the present implementation.

Steps 3 and 4 allow the incorporation of language-specific knowledge to enhance the operation of PMG-simple. However, in the experiments reported in the present article, only limited knowledge has been introduced, to evaluate how effective this phrasing model is in a setup where the system is not provided with large amounts of linguistic knowledge. It is expected that by providing more language-specific knowledge, the phrasing accuracy can be further increased over the results reported here.

When PMG-simple is trained, it is likely that some phrase boundaries are erroneously identified in the training data. The likelihood of such an event is non-negligible as phrases are automatically transferred using the alignment algorithm from the TL-side to the SL-side. Errors may be attributed to limited lexicon coverage or only partial correspondence of SL-to-TL text. However, as a rule such errors can be expected to correspond mainly to infrequent phrases.

A mechanism for screening such errors has been introduced in PMG-simple. This is implemented as a threshold imposed on the number of occurrences of a phrase within the training corpus, normalised over the occurrences in the entire corpus of the phrase tag sequence. Thus, phrases identified very rarely in comparison to the occurrences of their respective tag sequence are penalised as unreliable. They are retained in the phrase table, but are demoted to much lower positions. This processing of the phrase table is performed after Step 4 and represents the optional final step (Step 5) of PMG-simple.

3.3 Ordering Criteria

The choice of template-ordering criterion dictates the order in which phrases are matched to the input text. Since PMG-simple performs no backtracking, the actual ordering affects the segmentation accuracy substantially. A variety of different criteria have been investigated for establishing the order of precedence with which phrases are searched for. Out of these, only a selection is presented here due to space restrictions, focussing on the most effective criteria. These are depicted in Table 1.

crit.1	If phrase_freq ≥ freq_thres : Crit1 = {[(1000*(phrase_freq/ tagseq_occur) + phrase_len*250] } If phrase_freq < freq_thres: Crit1 = {[phrase_len *10] }
crit.2	If phrase_freq ≥ freq_thres : Crit2 ={(phrase_freq[p_index]) + phrase_len*10000} If phrase_freq < freq_thres: Crit2 = {phrase_len *10 +floor(100*phrase_freq/ tagseq_occur)}
crit.3	If phrase_freq ≥ freq_thres : Crit3 = {phrase_freq + phrase_len*1000} If phrase_freq < freq_thres: Crit3 ={phrase_len + phrase_freq/tagseq_occur}
crit.4	If phrase_freq ≥ freq_thres : Crit4 = max {phrase_subfreq + phrase_len*100 } If phrase_freq < freq_thres: Crit4 = {phrase_len + phrase_subfreq/tagseq_occur}

Table 1: Definitions of phrase-ordering criteria.

Basically, the information according to which phrases may be ordered in the phrase table consists of two types, (i) the frequency of occurrence of a given phrase in the training corpus (denoted as *phrase_freq*) and (ii) the phrase length in terms of tokens (denoted as *phrase_len*). By combining these two sources of information, different criteria are determined. Parameter *tagseq_occur* corresponds to the number of occurrences of the phrase tag sequence within the training corpus. Finally *phrase_subfreq* is equal to the occurrences of a tag sequence as either an

entire phrase or as a sub-part of a larger phrase. This takes into account in the frequency calculations the instances of phrases which in turn are encapsulated within larger phrases, and is the main point of difference between criteria crit3 and crit4.

To summarise a series of earlier experiments involving different criteria, criteria using only one source of information prove to be less effective. Also, criteria using non-linear combinations of information types (i) and (ii) have been shown to be less effective and are not reported here. All criteria studied in the present article combine the two aforementioned types of information in a weighted sum, but using different multiplication factors to emphasise one information type over the other. The actual factors may of course be further optimised, as the values reported in Table 1 are chosen to differ in terms of order of magnitude.

All criteria reported here implement Step 5, by having a secondary formulation when the occurrences of a phrase fall below a threshold (parameter *freq_thres*). This results in assigning a lower priority to very infrequent phrases.

A mechanism has also been introduced for the proper handling of tokens with very infrequent part-of-speech (PoS) tags, which typically have a rate-of-appearance of less than 0.5% in the corpus. For such tags, the likelihood of appearing in the 200 parallel sentences is very low. Hence, in order to split them appropriately into phrases when they appear in input sentences, equivalence classes have been defined. A limited number of PoS equivalences are used, namely (i) abbreviations and foreign words are considered equivalent to nouns, (ii) numerals are considered equivalent to adjectives and (iii) pronouns are considered equivalent to nouns. This information is inserted in Step 3 of the phrase-ordering algorithm, allowing the generation of the appropriate phrases. Though the improvement in translation accuracy by introducing these PoS equivalences is not spectacular (no more than 0.005 BLEU points) this generalisation information allows the appropriate handling of unseen tag sequences during translation, leading to a more robust phrasing method.

It should be noted here that a non-greedy variant of PMG-simple has also been examined. This was expected to be more effective, since it extends the template matching approach to take into account a sentence-wide context. However, it has turned out that the complexity of the non-greedy approach is too high. By introducing backtracking, it becomes extremely expensive computationally to run this method for sentences larger than 12 tokens without a substantial pruning of the search space.

4 Experimental setup and results

4.1 Experiment Definition

To evaluate the proposed phrasing generator, the output of the entire translation chain up to the final translation result is studied. This allows the contribution of different PMG models to be quantified using objective metrics. For the purposes of the present article, the language pair Greek-to-English (denoted as EL→EN) is employed. Since the SL phrasing generated by PMG is based on the TL phrasing scheme, the phrase labels of the resulting SL phrases are inherited from the TL ones. In the experiments reported here (with English as TL), the TreeTagger parser (Schmid, 1994) is used. Thus the SL-side phrase types include PC, VC, ADVC and ADJC. As TreeTagger also allows for certain words (such as conjunctions) to remain outside phrases, it is possible that isolated words occur in SL too. For the purposes of modelling such occurrences, these words form single-token phrases, denoted as ISC (i.e. ISolated word Chunk).

Both the parallel corpus and the evaluation dataset employed here have been established in the PRESEMT project, and are available over the web (cf. www.presemt.eu/data). The parallel corpus has been retrieved from the web (from an EU website discussing the history of the Union), with an average size of 18 words per sentence, while the smallest sentence comprises 4 words and the largest 38 words. Only minimal editing was performed in the parallel corpus, to ensure parallelism between SL and TL. The evaluation set comprises 200 isolated sentences, each with a single reference translation (Sofianopoulos et al., 2012). These sentences have been drawn from the internet via web crawling, being required to have a length of between 7 and 40 tokens each.

4.2 Experimental Results for PMG-simple

Table 2 contains the translation accuracy results obtained with PMG-simple using the criteria of Table 1. In all experiments, the results concern the objective evaluation of the final translation, using four of the most widely used objective

evaluation metrics, namely BLEU, NIST, TER and METEOR (NIST, 2002, Papineni et al., 2002 & Snover et al., 2006). For TER a lower value indicates a more successful translation while for other metrics, a higher value corresponds to a better translation. Since other components of the MT implementation do not change, this set of metrics provides an accurate end-to-end measurement of the effect of the phrasing model on the translation process. As can be seen from Table 2, all four criteria result in translations of a comparable accuracy. For instance, the variation between the lowest and highest BLEU scores is approximately 1%, while for the other metrics this variation is even lower.

Criterion	BLEU	NIST	METEOR	TER
crit 1	0.3643	7.3153	0.4009	48.486
crit 2	0.3679	7.2991	0.4009	48.590
crit 3	0.3667	7.2937	0.4002	48.730
crit 4	0.3637	7.2730	0.3980	48.834

Table 2: Translation accuracy for EL→EN, using PMG-simple with various criteria.

cut-off freq.	BLEU	NIST	METEOR	TER
0	0.3637	7.2730	0.3980	48.834
1	0.3637	7.2730	0.3980	48.834
2	0.3732	7.3511	0.4017	48.138
3	0.3660	7.2911	0.4007	48.590

Table 3: Translation scores for EL→EN, using PMG-simple with criterion 4 and various cut-off frequencies.

A potential for optimisation concerns the cut-off frequency (*freq_thres*) below which a phrase is considered exceptionally infrequent and is handled differently. Indicative results are shown for the four metrics studied in Table 3. As can be seen, the best results are obtained with a cut-off frequency of 2, for the given parallel corpus. Of course, this value is to an extent dependent on the training set. However, based on detailed analyses of the experimental results, it has been found that phrases that represent hapax legomena (i.e. phrases which occur only once) are not reliable for chunking purposes. Here, there are two possible explanations: (i) either such phrases represent spurious chunkings resulting from errors in the automatic alignment or (ii) they represent very infrequent phrases which again should not bias the phrasing process disproportionately. In both cases, the activation of the cut-off frequency improves the translation accuracy.

4.3 Comparison of PMG-simple to CRF

Of course it is essential to examine how PMG-simple translation results compare to those obtained when PRESEMT is run with the standard CRF-based phrasing model. These results are shown in Table 4. As can be seen the optimal performance of PMG-simple leads to an improved translation accuracy over the best CRF-based approach, with a rise of more than 6.2% in the BLEU score. Similarly, the improvements obtained for NIST and Meteor by introducing PMG-simple in PRESEMT are 2.1% and 2.5%, respectively. Finally, in the case of TER, for which a lower score reflects a better translation, the score is reduced by circa 3.3%. Thus, based on the results quoted in Table 3, the performance of PMG-simple is superior to that of the CRF-based system for all four metrics reported. The higher performance of PMG-simple is in agreement to the observation that - as recently reported for other applications (Mao at al., 2013) - improvements over the performance of CRF and SVM are possible by appropriately weighing templates.

PMG	BLEU	NIST	METEOR	TER
PMG-simple (crit.4)	0.3732	7.3511	0.4017	48.138
CRF	0.3513	7.1966	0.3919	49.774

Table 4: Translation accuracy for EL→EN, using PMG-simple with crit.4 and using CRF.

To evaluate in more detail the results of Table 4, a preliminary statistical analysis was performed. More specifically, the scores in BLEU, NIST and TER for each of the 200 test sentences were collected. For each of these metrics, a paired T-test was performed comparing the measurements obtained with (i) PMG-simple using criterion crit.4 and (ii) CRF, over each sentence. It was found that the difference in means between the BLEU populations was indeed statistically significant at a 0.05 level. In the cases

of TER and NIST measurements, though, there was no statistically significant difference in the two populations.

5 Conclusions

PMG-simple has been proposed as a straightforward implementation to derive a phrasing model for SL text, based on template-matching. This operates on the same aligned corpus as the default CRF model, but is faster to train and has a more transparent operation. The results of PMG-simple have been compared to those of CRF, using the final PRESEMT translation output to gauge the phrasing effectiveness. The best results for PMG-simple are comfortably superior to those of CRF for all MT objective metrics used. This indicates that PMG-simple has a sufficiently high functionality. Though the modelling power of CRF is higher, the template-matching approach of PMG-simple is better harmonised to the amount of training data available. Thus PMG-simple appears to be the phrase generator of choice for PRESEMT.

One point that warrants further experimentation (currently under way) concerns the scaling-up effect of larger parallel corpora on the comparative performance of the models. Preliminary results with bilingual corpora of approximately 500 sentences have shown that the performance using PMG-simple remains superior to that with CRF, resulting in a difference of approx 0.02 for BLEU (equivalent to a 5%-6% improvement over the CRF baseline). In addition, PMG-simple has been shown to perform better than CRF when applied to the latest versions of PRESEMT, which are currently being tested and lie beyond the scope of this article.

Another topic of interest is to determine whether new improved criteria can be established. This is the subject of ongoing research.

In addition, an open question is whether the conclusions of this study are applicable to other thematic areas. In other words, could an approach such as PMG-simple be preferable to CRF in other applications involving relatively sparse data? It appears from the results summarised here that this could indeed be the case, though this remains the subject of future research.

Acknowledgements

The author wishes to acknowledge the invaluable help of Ms. Marina Vassiliou and Dr. Sokratis Sofianopoulos, both of ILSP/Athena R.C., in integrating PMG-simple within the PRESEMT prototype and performing a number of experiments.

The research leading to these results has received funding from the POLYTROPON project (KRIPIS-GSRT, MIS: 448306).

References

Paul E. Black. 2005. Dictionary of Algorithms and Data Structures. U.S. National Institute of Standards and Technology (NIST).

Richard O. Duda, Peter E. Hart and David G. Stork. 2001. *Pattern Classification (2nd edition)*. Wiley Interscience, New York, U.S.A.

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

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch and David Yarowsky. 2012. Towards Statistical Machine Translation without Parallel Corpora. *In Proceedings of EACL-2012 Conference*, Avignon, France, 23-25 April, pp. 130-140.

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

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

Qi Mao, and Ivor Wai-Hung Tsang. 2013. Efficient Multitemplate Learning for Structured Production. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 24, No. 2, pp. 248-261.

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

Franz Josef Och and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, Vol. 30, No. 4, pp. 417-449.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the *40th ACL Meeting*, Philadelphia, USA, pp. 311-318.

Chris Quirk and Arul Menezes. 2006. Dependency Treelet Translation: The convergence of statistical and example-based machine translation? *Machine Translation*, Vol. 20, pp. 43–65.

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.

Fei Sha and Fernando C. N. Pereira. 2003. Shallow Parsing with Conditional Random Fields. I*n Proceedings of HLT-NAACL Conference*, pp. 213-220.

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

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 Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, pp. 223-231.

Sokratis Sofianopoulos, Marina Vassiliou, and George Tambouratzis. 2012. Implementing a language-independent MT methodology. *In Proceedings of the First Workshop on Multilingual Modeling*, held within the ACL-2012 Conference, Jeju, Republic of Korea, 13 July, pp.1-10.

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong and Qun Liu. 2012. Translation Model Adaptation for Statistical Machine Translation with Monolingual Topic Information. *In Proceedings of the 50th ACL Meeting*, Jeju, Republic of Korea, pp. 459-468.

George Tambouratzis, Fotini Simistira, Sokratis Sofianopoulos, Nikolaos Tsimboukakis, and Marina Vassiliou. 2011. A resource-light phrase scheme for language-portable MT. *In Proceedings of the 15th EAMT Conference*, 30-31 May, Leuven, Belgium, pp. 185-192.

George Tambouratzis, Michalis Troullinos, Sokratis Sofianopoulos, and Marina Vassiliou. 2012. Accurate phrase alignment in a bilingual corpus for EBMT systems. *In Proceedings of the 5th BUCC Workshop*, held within the LREC-2012 Conference, May 26, Istanbul, Turkey, pp. 104-111.

George Tambouratzis, Sokratis Sofianopoulos, and Marina Vassiliou (2013) Language-independent hybrid MT with PRESEMT. *In Proceedings of HYTRA-2013 Workshop*, held within the ACL-2013 Conference, Sofia, Bulgaria, 8 August, pp. 123-130.

Yoshimasa Tsuruoka, Jun'ichi Tsujii and Sophia Ananiadou. 2009. Fast Full Parsing by Linear-Chain Conditional Random Fields. *In Proceedings of the 12th EACL Conference*, Athens, Greece, 30 March-3 April, pp. 790–798.

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

Dekai Wu. 2009. Toward machine translation with statistics and syntax and semantics. *In Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding*, 13-17 November, Merano, Italy, pp. 12-21.

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