# Linguistically Inspired Language Model Augmentation for MT

**George Tambouratzis, Vassiliki Pouli**,
Institute for Language and Speech Processing / Athena Research Centre,
6 Artemidos & Epidavrou, Paradissos Amaroussiou, Athens, GR-15125, Greece.
E-mail: giorg_t@ilsp.gr , vpouli@ilsp.gr

**Abstract**

The present article reports on efforts to improve the translation accuracy of a corpus–based Machine Translation (MT) system. In order to achieve that, an error analysis performed on past translation outputs has indicated the likelihood of improving the translation accuracy by augmenting the coverage of the Target-Language (TL) side language model. The method adopted for improving the language model is initially presented, based on the concatenation of consecutive phrases. The algorithmic steps are then described that form the process for augmenting the language model. The key idea is to only augment the language model to cover the most frequent cases of phrase sequences, as counted over a TL-side corpus, in order to maximize the cases covered by the new language model entries. Experiments presented in the article show that substantial improvements in translation accuracy are achieved via the proposed method, when integrating the grown language model to the corpus-based MT system.

**Keywords:** corpus-based MT, language model augmentation, extraction of knowledge from corpora.

## 1. Introduction

Currently a large proportion of MT approaches which are readily portable to new language pairs are based on the Statistical Machine Translation (SMT) paradigm. The main obstacle to the creation of an SMT system is the requirement for parallel corpora between the Source Language (SL) and the Target Language (TL), which are of a sufficient size to allow the extraction of meaningful translation models. As the amount of parallel corpora is limited for most language pairs, researchers of SMT systems are investigating the extraction of information from monolingual corpora, including creation of lexica (Koehn et al., 2002) and of topic-specific information (Su et al., 2012).

The present article studies a hybrid method for developing MT systems which has been developed as an alternative to SMT systems. This hybrid method is corpus-based and is intended to support the creation of MT systems using a common set of software modules, while having minimal requirements for specialized resources. This methodology is specifically designed to address the scarcity of parallel corpora from which an MT system is trained, using instead predominantly monolingual corpora. This allows the current methodology to be usable for language pairs for which parallel corpora are very limited. Based on this hybrid translation method and having as baseline the hybrid MT system developed in (Tambouratzis et al., 2013), the aim here is to investigate how the translation quality can be improved by augmenting the coverage of the Target Language Model (TLM).

## 2. The Translation Process

The hybrid translation methodology studied here is built on a two-stage core translation engine developed in the framework of the PRESEMT project (www.presemt.eu) which uses very small parallel corpora (ca. 200 sentences) supplemented by monolingual corpora. The final resource that is used is a bilingual lexicon between SL and TL which contains the translations of terms.

PRESEMT adopts a phrase-based approach, where the text-to-be-translated is processed on the basis of the syntactical phrases contained. Phrases are determined via a dedicated module (the Phrasing Model Generator - PMG) trained on the small parallel corpus. This dedicated module ports the phrasing scheme from the target language (a TL chunker is used) towards the source language.

In the first translation stage, each input sentence is handled as a sequence of phrases and the order of these phrases is determined by comparisons made at the SL side of the parallel corpus. Having located within the parallel corpus the SL sentence that best matches the input sentence, then the corresponding TL-side structure of that sentence is used as the structure for the translation. In this respect, the first PRESEMT stage strongly resembles Example-Based-Machine-Translation (EBMT).

In turn, the second stage samples the monolingual TL corpus to determine the most likely choice of word translations and sequence of tokens within the

boundaries of each phrase. To support this, a language model is created where the phrasal patterns (sequences of tokens) are recorded, together with the frequency-of-occurrence of each pattern. Thus, the second translation stage is more similar to Statistical Machine Translation (SMT) principles.

The PRESEMT Target Language Model (TLM) utilizes a phrase-based indexing scheme, in order to support fast and efficient translation of phrases. This scheme involves organizing phrases on (i) their type, (ii) their head and (iii) the head PoS tag and indexing by the phrase head. Phrase types are directly determined by the chunker chosen for the TL-side. As an example, for English as TL, the TreeTagger (Schmid, 1995) results in four main phrase types, namely ADVC, ADJC, PC and VC, as depicted in Table 1. When the tokens of a given chunk need to be ordered in the second translation stage, the phrases from the monolingual corpus (which form the TLM) are searched based on phrase head and type to determine the most appropriate sequence.

| Abbreviation | Type | Example |
|---|---|---|
| ADVC | Adverbial chunk | quickly |
| ADJC | Adjectival chunk | young |
| PC | Prepositional chunk | the city centre |
| VC | Verb chunk | were asked |

Table 1: Main phrase types and typical examples

An error analysis of the translation output has shown that some of the most severe output errors are attributable to the second PRESEMT translation stage, with sub-optimal intra-phrase ordering of tokens. Indicative examples are shown in Table 2, relating to translation errors for PC (entry 1) and VC (entry 2) types.

| EL Input | Actual trans. | Correct trans. |
|---|---|---|
| 1. Η οικονομική κατάσταση της πόλης | Of the economic city condition | The financial condition of the city |
| 2. Κλήθηκαν να εξετάσουν | to asked were consider | were asked to consider |

Table 2: Indicative examples of MT output errors

A study of these errors indicates that the second translation phase fails to locate any sufficiently close match for the SL-side input phrase among the sets of TL phrases extracted from the monolingual corpus.

## 3. Target Language Models

Errors such as those of Table 2 are attributable to a language model with limited coverage, where for a number of phrasal patterns no appropriate match for the SL input is found in a single phrase. The reasoning is that during chunking it is highly likely that a boundary between two phrases is not detected and two consecutive phrases are grouped into a larger one. This results in a sub-optimal word-reordering within the phrases.

To address this problem, an appropriate augmentation of the language model, by concatenating consecutive syntactic phrases, was chosen. This process is the equivalent of increasing the size *n* of n-gram LMs, as suggested by (Wang et al, 2014), in a model growing process. The n-phrase[1] growing approach is developed to define the expected likelihood of appearance of the most frequent n-phrases in a monolingual corpus, where $n>1$.

The possible phrase types are determined from the TL parser. In the case of English as TL, the set of phrase types is M = {VC, ADVC, PC, ADJC}. Following that, the TLM is enhanced by adding the appropriate composite sequences of n-phrases e.g. ADVC/VC (a VC appearing just after an ADVC) for only the phrase categories that appear in sequence with a very high frequency, so as to concentrate the effort on improving the coverage of the most likely-to-occur phrase sequences.

The first step is to examine which phrase types are eligible for augmentation depending on their high probability of occurrence. More specifically, for a total of $m=|M|$ types, the probability $P$ of a phrase $p$ of the $i^{th}$ type appearing in the corpus is expressed as follows:

$$P(p_i) = \frac{f_i}{\sum_{i=1}^{m} f_i} \quad (1)$$

In equation (1), $f_i$ is the frequency of occurrence of the $i^{th}$ type. The phrase types $p$ selected for augmentation are those belonging to set $S$ (where $S$ is a subset of $M$) for which the occurrence probability $P$ exceeds the value $threshold_A$ (eq. (2)). For English as TL, only types PC and VC exceed this threshold and are augmented.

$$S = \{p | P(p_i) > threshold_A\} \quad (2)$$

In the TL-side monolingual model, the probability of appearance of a composite n-phrase $p_1^n = < p, \dots, p_n >$ is estimated by computing the conditional probability of the phrase type $p_n$ (e.g. a phrase of type PC) appearing given the preceding phrase types $p_{1,\dots,} p_{n-1}$:

$$P(p_1^n) = P(p_n | p_{1\dots}p) = \frac{P(p_{1\dots}p_n)}{P(p_{1\dots}p_{n-1})} \quad (3)$$

[1] Where an n-phrase is defined as a sequence of n-consecutive phrases.

For the experiments presented in the present manuscript, the number of consecutive phrases $n$ is limited to 2. Among the composite phrases $p_1^n$ derived from the augmentation process, only those that adhere to the following conditions are considered valid and added in set $Q$ of composite phrases:

(i) their probability of appearance $P(p_1^n)$ (as expressed in eq. (4)) exceeds a specified $threshold_B$ (5a) and

(ii) the constituent phrases of the composite phrase belong to $M$ but do not correspond to more than one augmented types from $S$ (5b). As an example, let us consider the case where $S$={PC,VC}, i.e. S comprises phrases of type PC and VC. Then a PC phrase can be concatenated with another phrase of type PC, ADVC or ADJC. On the contrary, a PC phrase cannot be concatenated with a VC phrase since the phrases VC and PC are of two different types both of which belong to set $S$.

Condition (ii) prevents conflicts in the type of the resulting composite phrases due to the concatenation of phrases from incompatible linguistic types.

$$P(p_1^n) = P(p) * P(p_2|p_1) * \dots * P(p_n|p_1^{n-1}) \quad (4)$$

$$Q = \{p_1^n | P(p_1^n) > threshold_B, p_n \in S, \quad (5a)$$
$$(p_{n-1} = p_n) || (p_{n-1} \notin S)\} \quad (5b)$$

The composite phrase $p_1^n$ is then indexed based on the existing criteria of TL and added to the relevant indexed file maintaining the main phrase type belonging to $S$. For instance, a composite {VC/ADVC} is entered as an augmented VC.

## 4. Experiments

### 4.1. Experimental Setup

The proposed methodology is evaluated on the Greek-to-English (EL-EN) language pair. The MT system is trained with (i) a 200-sentence bilingual corpus (available at www.presemt.eu) and (ii) an extensive monolingual corpus in English of more than one billion tokens from which the indexed LM is extracted.

Three different PMG modules have been studied to investigate whether the proposed language model growing contributes to the robustness of the MT methodology. The first PMG is based on Conditional Random Fields (CRF) (Lafferty et al., 2001), which is the default choice for PRESEMT. An alternative PMG module, termed TEM-s, is based on TEmplate-Matching (TEM) principles and gives a higher translation accuracy in EL-EN (Tambouratzis, 2015). A further TEM variant is studied (TEM-b) that favours larger phrases than TEM-s.

Experiments employ two sets of 200 sentences each, denoted as Testset A and Testset B. Each testset results in different phrase sizes for each PMG, as noted in Table 3.

| Testset | CRF | TEM-s | TEM-b |
|---|---|---|---|
| A | 2.27 | 1.94 | 2.19 |
| B | 2.35 | 1.80 | 2.09 |

Table 3: Average phrase sizes (in words) for evaluation testsets with different PMG modules.

For the EL-EN language pair, the set $S$ of grown phrase types comprises PC and VC phrases. Regarding the indexed LM, a number of augmented versions have been compiled by augmenting the baseline model (denoted as V0) used in previous experiments. Versions V1 to V4 correspond to grown versions of V0, by incrementally adding composite phrases as shown in Fig. 1 (the number of new phrases introduced in the LM per step is noted, normalized over the number of baseline phrases). More specifically, V1 is created by adding composite phrases consisting of two consecutive PCs i.e. {PC/PC}. As an example, a sample of phrases from the basic indexed corpus (V0) corresponding to PCs with the lemma "condition" as their head is depicted in Table 4, together with a sample following augmentation (V1). The use of the grown model V1 leads to the improved translation of phrases containing two PCs, this being typified by example 1 of Table 2. Version V2 extends V1 to include a sequence of one ADVC phrase followed by a PC i.e. {ADVC/PC}. In V3, VC phrases are extended by composites comprising an ADVC followed by a VC. Finally, in V4 the VC class is grown by adding composite phrases of two successive VCs, solving errors such as example 2 of Table 2.

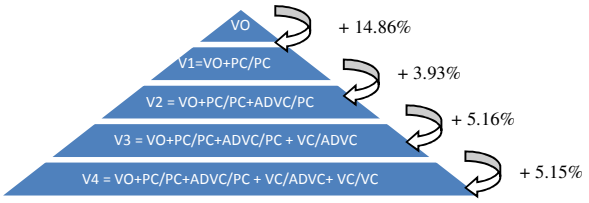


Figure 1: Incremental growing of LMs.

The translation quality is assessed using established MT metrics including BLEU (Papineni et al., 2002), NIST (NIST, 2002), Meteor (Denkowski et al., 2011) and TER (Snover et al., 2006). When comparing scores, an improvement in translation accuracy is depicted as a positive change.

| Phrase Type | Indexed corpus phrase example |
|---|---|
| PC (V0) | PC3(the condition)<br>‡PC0(of the economic living condition) |
| Composite PC (V1) | PC3(the condition)<br>PC0(of the economic living condition)<br>‡PC1(the financial condition) PC2(of the city)<br>PC12(the conditions) PC13(in France) |

Table 4: Samples from indexed corpus of lemma "condition" and type PC (i) from the baseline V0 and (ii) the augmented V1. The phrases used to translate example 1 of Table 2 are denoted by ‡

## 4.2. Experimental Results

For reasons of conciseness, most of the experiments reported here concern Testset A. Figure 2 depicts the translation quality variation for CRF, when the various augmented corpora are deployed, compared to the baseline V0 corpus. For all four metrics and all augmented LMs, improvements in terms of translation quality are reported compared to the baseline V0.

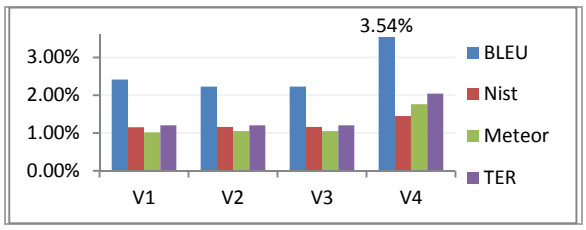


Figure 2: Improvement in metrics obtained with CRF, by using the grown LMs over V0.

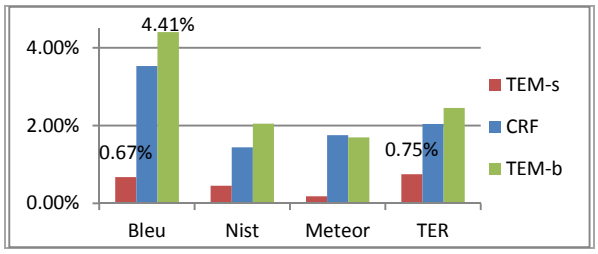


Figure 3: Improvement in metrics with grown language model V4, for three different PMGs with Testset A.

Figure 3 indicates how translation quality improves when LM V4 is used as compared to V0, for different PMGs. The use of the grown corpus improves metric scores in all three types of PMG modules used. The highest improvement is obtained with TEM-b, which creates longer phrases. Furthermore, the scores for all PMG models converge when using growing models, demonstrating the improved consistency of the MT system irrespective of the PMG choice. A similar situation is depicted for TestsetB in Figure 4.

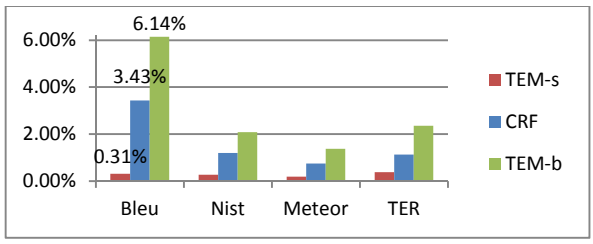


Figure 4: Improvement in metrics with grown language model V4, for various PMGs with Testset B.

To provide a reference system, a MOSES-based SMT system was chosen, created with a parallel corpus of 1.2 million sentences. Applied to the same testset (Testset A), the MOSES-based system achieved a slightly higher BLEU score, but our hybrid MT system showed increased performance in the other three metrics (i.e. NIST, Meteor, TER) (see Table 5). Since in this last experiment absolute values of metrics are compared (instead of improvements), it should be noted that for TER, a lower value (as achieved by the TEM-s hybrid system in comparison to MOSES) denotes better translation accuracy in contrast to the other metrics. Taking into account that the parallel resources of the SMT system are more than 3 orders of magnitude larger than those of the hybrid system, underlines the promising characteristics of our hybrid system.

| Number of sentences | 200 | Source | Web |
|---|---|---|---|
| Reference Translations | 1 | Lang. pair | EL-EN |
| MT System | Metrics | | |
| | BLEU | NIST | Meteor | TER |
| TEM-s | 0.3647 | **7.2176** | **0.4036** | **0.4487** |
| Moses | **0.3795** | 7.0390 | 0.3602 | 0.5711 |

Table 5: Comparison of metric scores between our hybrid MT system and Moses.

## 4.3. Statistical Analysis of Results

Paired t-tests were applied to determine whether the results are statistically significant, by comparing the scores for models V0 and V4. To this end, populations were created using the BLEU scores at a sentence level, for each experimental configuration. The aim was to compare the means of the populations and determine if they differ in a statistical sense. For testset A, at a confidence level of 95%, model V4 gives a significantly better translation quality over V0 for both TEM-b and CRF. For TestsetB, at a level of

99%, V4 improves significantly the translation over V0 for both TEM-b and CRF.

## 5. Conclusion

In this article, the use of grown LMs has been studied for improving the PRESEMT translation accuracy. Larger LMs usually perform better, thus efficient ways of constructing them is an important topic in SMT, and more generally in MT. According to the experimental results, the augmentation of a LM by concatenating appropriately consecutive phrases leads to improved translation accuracy, with reduced dependence on the choice of phrasing model. In many of the experimental configurations studied, statistically significant improvements have been shown.

Hence, the present article indicates how the augmentation of a language model may be algorithmically defined and implemented. This algorithm results in a measurable improvement in the translation quality, indicating the more effective extraction of information from the language resources available. It is believed that the proposed algorithm can be effective in other applications, beyond machine translation, thus allowing a more effective extraction of knowledge from language resources that have already been compiled.

## 6. Acknowledgements

The research reported in this article has been funded partly by a number of projects including the PRESEMT project (ICT-FP7-Call4/248307) and the POLYTROPON project (KRIPIS-GSRT, MIS: 448306).

The authors wish to acknowledge the assistance of Ms. M. Vassiliou of ILSP/Athena R.C. on the setting up of experiments and of Dr. S. Sofianopoulos of ILSP/Athena R.C., in integrating the new structure selection algorithm to the PRESEMT prototype and on providing the translation results for the MOSES-based SMT system.

## 7. Bibliographical References

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems, Proc. of the 6th SMT Workshop, Edinburgh, UK, July 30–31, pp. 85–91.

Koehn, P. and Knight, K. (2002). Learning a Translation Lexicon from Monolingual Corpora. Unsupervised Lexical Acquisition: Proc. of SIGLEX Workshop, Philadelphia, USA, July 2002, pp. 9-16.

Lafferty, J., McCallum, A. and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data, Proceedings of ICML Conference, June 28-July 1, Williamstown, USA, pp. 282-289.

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

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. Proc. of the 40th ACL Conference, Philadelphia, U.S.A., pp. 311-318.

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

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the 7th AMTA Conference, Cambridge, MA, USA, pp. 223-231.

Su, J., Wu, H., Wang, H., Chen, Y., Shi, X., Dong, H. and Liu, Q. (2012). Translation Model Adaptation for Statistical Machine Translation with Monolingual Topic Information. Proc. of ACL-2012, Jeju, Republic of Korea, 9-11 July, pp. 459-468.

Tambouratzis, G., Sofianopoulos, S. and Vassiliou, M. (2013). Language-independent hybrid MT with PRESEMT. Proc. of HYTRA-2013 Workshop, held within the ACL-2013 Conference, Sofia, Bulgaria, 8 August, pp. 123-130.

Tambouratzis, G. (2015). Conditional Random Fields versus template-matching in MT phrasing tasks involving sparse training data. Pattern Recognition Letters, Vol. 53, pp. 44-52.

Wang, R., Zhao, H., Lu, B.-L., Utiyama, M. and Sumita, E. (2014). Neural Network Based Bilingual Language Model Growing for Statistical Machine Translation. Proc. of the EMNLP-2014 Conference, Doha, Qatar, October 25-29, pp.189-195.