GRAFIX: Automated Rule-Based Post Editing System to Improve English-Persian SMT Output

Mahsa Mohaghegh[1] *Abdolhossein Sarrafzadeh*[2] *Mehdi Mohammadi*[3]
(1) Massey University, School of Engineering and Advanced Technology, Auckland, New Zealand
(2) Unitec, Department of Computing, Auckland, New Zealand
(3) SheikhBahaee University, Department of Computer Engineering, Isfahan, Iran
M.mohaghegh@massey.ac.nz, Hsarrafzadeh@unitec.ac.nz
mehdi.mka@gmail.com

ABSTRACT

This paper describes the latest developments in the PeEn-SMT system, specifically covering experiments with Grafix, an APE component developed for PeEn-SMT.

The success of well-designed SMT systems has made this approach one of the most popular MT approaches. However, MT output is often seriously grammatically incorrect. This is more prevalent in SMT since this approach is not language-specific. This system works with Persian, a morphologically rich language, so post-editing output is an important step in maintaining translation fluency.

Grafix performs a range of corrections on sentences, from lexical transformation to complex syntactical rearrangement. It analyzes the target sentence (the SMT output in Persian language) and attempts to correct it by applying a number of rules which enforce consistency with Persian grammar.

We show that the proposed system is able to improve the quality of the state-of-the-art English-Persian SMT systems, yielding promising results from both automatic and manual evaluation techniques.

KEYWORDS : Machine Translation, Post-editing of Machine Translation, Evaluation of Machine Translation

Proceedings of COLING 2012: Posters, pages 873–882,
COLING 2012, Mumbai, December 2012.

1 Introduction

Since most mistakes associated with machine translation are of a repetitive nature, the task of post-editing can be made automatic (Allen & Hogan, 2000). Furthermore, the process of automatic post-editing (APE) is very similar in nature to a machine translation process (Simard, Goutte, & Isabelle, 2007). Because of this, certain MT systems can be used to model the APE process.

The advantages and disadvantages of RBMT and SMT approaches may be summarised as follows: RBMT is strong in syntax, morphology, structural semantics, and lexical reliability, but demonstrates weakness in the areas of lexical semantics and lexical adaptivity. SMT, while being weak in the areas of syntax, morphology, and structural semantics, is superior to RBMT in areas of lexical semantics and adaptability, although the advantage of adaptability to other language pairs is only valuable when the system is to be used with a wider range of languages.

The Grafix APE system's main algorithm follows a Transfer-based approach. Transfer-based MT is among the most commonly used approaches for MT. This method involves capturing the meaning of a source sentence using intermediate representations, and from it generating a target output (Mohamed, 2000). The Grafix system developed by the authors attempts to correct some frequently occurring grammatical SMT system errors in English-to-Persian translations.

2 Related Work

Simard et al. (2007), Lagarda, Alabau, Casacuberta, Silva, and Diaz-de-Liano (2009) present APE systems that are added to commercial RBMT systems. Their APE components utilise a phrase-based SMT system using Moses as a decoder.

In his recent work, Pilevar (2011) demonstrates a statistical post-editing (SPE) module that is used to improve RBMT output for the English-Persian language pair in order to improve the translation of subtitles for movies. The results show that the SPE module can improve the performance of the RBMT system's output when used in a new domain. However, they found that the use of the SMT system alone yields a better result compared to the combination of RBMT + SPE. To our knowledge this is the only post-editing system reported for the English-Persian language pair, and it did not succeed in improving the output of the main system.

Marecek, Rosa, and Bojar (2011) report on experimental work in correcting the output of an English-Czech MT system by performing several rule-based grammatical corrections on sentences parsed to dependency trees. Their baseline SMT system relies on Moses, a phrase-based translation system. In their post-processing system, DEPFIX, they used a two-step translation that is a setup in which, the English source is first translated into simplified Czech, and then the simplified Czech is monotonically translated to fully inflected Czech. Both steps are simple phrase-based models. Rosa, Marecek, and Duˇsek (2012) enriched the rule set of DEPFIX and used a modified version of MST Parser. Their results show that both modifications led to better performance of DEPFIX 2012; however, they mention that since the effect of DEPFIX on the output in terms of BLEU score is not significant, the results are not as reliable as results obtained through manual evaluation.

3 Description of the System

Our approach to the system architecture differs from what is commonly used in most other systems in that the APE does not use an SMT system to automatically post-edit the output of an MT system, as described, for example, in Simard et al. (2007) and Lagarda et al. (2009).

In this study, we couple the PeEn-SMT system we previously developed (Mohaghegh, Sarrafzadeh, & Moir, 2011)with an RBMT-based APE. Since post-editing an MT system's output usually seeks to improve grammatical structure in order to render sentences and phrases with greater fluency, the advantage of RBMT's linguistic knowledge can be utilised well here.

3.1 The Underlying SMT System

Most recent research in the area of statistical machine translation has been targeted at modelling translation based on phrases in the source language and matching them with their statistically-determined equivalents in the target language ("phrase-based" translation) – (Koehn, Och, & Marcu, 2003; Marcu & Wong, 2002; Och & Ney, 2004; Och, Tillmann, & Ney, 1999). After conducting numerous experiments with Moses, we decided to experiment with some modifications of the Joshua 4.0 toolkit, to compare them and see if a better score could be achieved. To the best of our knowledge, this is the first time a hierarchical SMT system is being used for the Persian-English language pair. One motivation for this is the fact that since Persian is a morphologically rich language, word disordering is a common issue that we face. Hierarchical SMT takes syntax into account to some extent, with phrases being used to learn word reordering. This improvement is due to the word order differences between Persian and English, which are better handled with a hierarchical phrase based system than a standard phrase-based approach. Hierarchical phrase-based translation (Chiang, 2005) expands on phrase-based translation by allowing phrases with gaps, modelled as synchronous context-free grammars (SCFGs). Joshua is a well-known open source machine translation toolkit based on the hierarchical approach (Li, Callison-Burch, Khudanpur, & Thornton, 2009). In the latest version of Joshua (Version 4.0), the main changes include implementation of Thrax, which enables extended extraction of Hiero grammars, and a modified hypothesis exploration method (Ganitkevitch, Cao, Weese, Post, & Callison-Burch, 2012).

3.2 The Proposed APE Model

The proposed rule-based APE module consists of three levels of transformation.

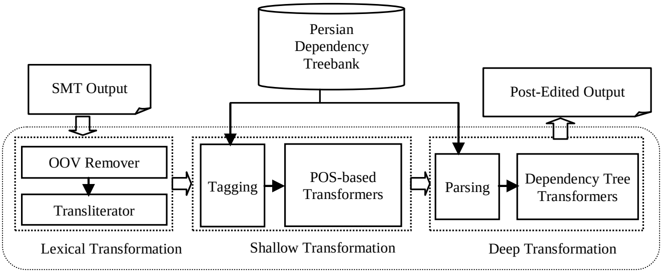


FIGURE 1 – High-Level diagram of the proposed Rule-based APE system

As shown in Figure 1, these three levels are lexical transformers, shallow transformers and deep transformers. First OOVRemover and Transliterator as lexical transformers are run using a bilingual dictionary, after which some shallow transformers are run based on POS tag patterns. Deep transformation at the third level is applied in which the rules exploit the tree dependecy structure of sentences.

Lexical Transformation: The first level benefits from the outcome of two components. OOV[1] remover is a simple substitute rule to replace an English word with the correct translation in Persian. However, there are instances like named entities where OOV remover could not find equivalent Persian translations for English words appearing as OOV in the output. In this case, a transliterator is used to replace English words by their equivalents in Persian scripts. The transliterator component uses a training data set containing over 4600 of the most frequently used Persian words and named entities written using English letters, and also the equivalent in Persian script.

Shallow Transformation: The second stage of the system involves a shallow transfer module. POS-tagging the input text is a pre-requisite process for both shallow and deep transformation levels. The MLE POS-tagger is used in this stage and trained with the Persian Dependency Treebank[2] data. Shallow transformers are developed, based on some POS patterns identified as wrong ones.

Deep Transformation: In the third level, the input is parsed by a dependency parser. Once the text is tagged, some preparation is performed to parse the input, based on the parsing input format (McDonald, Pereira, Ribarov, & Hajic, 2005). The Persian Dependency Treebank is also used in the parser training process.

We used MSTParser, which is an implementation of Dependency Parsing using, the Maximum Spanning Tree (Kübler, McDonald, & Nivre, 2009). The rules here are used for examination of the sentence's dependency tree in order to have some syntactical and grammatical constraints.

3.3 Training Data Source

In a sentence dependency tree, words and relations are graphed, with each word either modifying or being modified by another word, and the root in each tree being the only word which does not modify any other word. We have used Persian Dependency Treebank as our main source of training data for both tagging and data-driven parsing. It contains about 125,500 annotated sentences. The data format is based on CoNLL Shared Task on Dependency Parsing (Buchholz & Marsi, 2006). The sentences are manually annotated in the corpus, which contains about 12,500 sentences and 189,000 tokens

3.4 Pre-Processing and Tagging

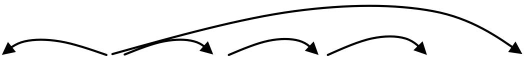
The pre-processing of input Persian sentences consists of tokenizing the sentences using our implemented tokenizer. We chose the Maximum Likelihood Estimation (MLE) approach as the POS-tagging component for our APE, due to its ability to be implemented easily and its consistency in yielding promising results for tagging the Persian language (Raja et al., 2007).

[1] Out Of Vocabulary

[2] http://dadegan.ir/en

3.5 Parsing

In dependency parsing, words are linked to their arguments by dependency representations (Hudson, 1984). These representations have been in use for many years. In Figure 2, the sentence, shown in sentence tree form, is a dependency tree. Each word depends on a “parent” word or a root symbol.



Label	PUNC	ROOT	OBJ	PREDEP	NPREMOD	SBJ
Token	.	می خوانم	را	نامه	یک	من
Pronunciation	.	/mi:khãnam/	/rã/	/nãmæ/	/yek/	/man/
POS	PUNC	V	POSTP	N	PRENUM	PR
English Equivalent	.	read		letter	a	I

FIGURE 2 – Dependency Parsing Example

3.6 Rule-based Transformers

The translation rules were gathered manually by investigating a broad range of incorrect translations. By considering the dependency parser output for these sentences, and determining frequent wrong patterns among them, we have defined the most common incorrect patterns under four rules in the shallow transformers, and six in the deep transformers. The following sections cover some of them regarding the transfer level.

3.6.1 Shallow Transformers

IncompleteDependentTransformer: In Persian, as in English, dependent clauses are usually connected by relative pronouns such as«که»(English *“that”*). The rule below identifies a lack of verb in a dependent sentence and corrects it by adding a verb. Currently, in most instances the verb «است» (English *“is”*) is suggested. In the notation below, * denotes any number of POS, and ^ denotes ‘except’.

***If** POS-sequence matches [* SUBR *^V PUNC] → modify as [* SUBR V(است) PUNC]*

IncompleteEndedPREMTransformer: Pre-modifiers (denoted by PREM) are a class of noun modifiers that precede nouns and are in complementary distribution with other members of the class. In the POS sequence in which a pre-modifier precedes a punctuation mark (PUNC) deemed as incorrect. Since there is no logical translation for given inputs with this pattern, these sequences were removed from the sentence altogether. The rule is described as:

***If** POS-sequence matches [$*_a$ N PREP PREM PUNC $*_b$] → modify as [$*_a$ $*_b$]*

3.6.2 Deep Transformers

NoSubjectSentenceTransformer: SMT output occasionally contains instances of sentences with a third person verb, no definite subject and an object labelled as OBJ in the parse tree and tagged as POSTP (postposition) in the POS sequence. Compared to known reference sentences, it was seen that what was parsed as the object in the sentence was actually the subject. The transformer is designed to revise the sentence by removing the postposition «را» which is the indicator of a direct object in the sentence. Removal of this postposition changes the sentence to one with a subject.

VerbArrangementTransformer: As a natural language, Persian has a preferred word order, with SOV (subject-object-verb) followed by SVO. One frequently violating case is sentences in which a main verb as Root does not occur immediately before the period punctuation. The matching procedure is as follows: For the verb of the sentence tagged as Root, reordering is performed by moving the root verb and its NVE dependants (in the case of compound verbs) to the end of the sentence, immediately before the period punctuation.

MissingVerbTransformer: In this transformer, any subject with a referred verb preceding the subject is identified as an incorrect linked subject to any verb, since the sentence does not follow the standard SOV structure. In this case, it can be assumed that the last word in the sentence can act as a candidate in order to find the non-verbal element in the verb Valency Lexicon (Rasooli, Moloodi, Kouhestani, & Minaei-Bidgoli, 2011). If such a verb is found, that verb will be suggested to fill the space of the missing verb. The tense of the verb is then modified to match that of the subject of the sentence.

MozafOfAlefEndedTokenTransformer: In Persian, there are certain nouns or pronouns following a head noun which signify relationships with the head noun, such as possession or name relation. Such nouns/pronouns are known as Ezafe dependents. Indication of such in the language is given as the vowel sound /e/, coming immediately after pronunciation of the head noun. If the head-word ends in «ا »/a/, then the character « ی» must be added to the end of that word. This character is a representation of the /e/ vowel that is written in such cases to ease the pronunciation. This transformer recognizes the Ezafe dependents which require a « ي» character between them and add it properly.

4 Experiments and Results

The SMT system evaluated in this paper is based on Joshua 4.0 with default setting. The parallel corpus used for the training set was based on the NPEC corpus tested by (Mohaghegh & Sarrafzadeh, 2012), but we built a modified version consisted of almost 85,000 sentence pairs in which we removed the subtitle addition. The language model was extracted from IRNA[3] website. The details of the components of the baseline system prior to alignment are shown in Table 1.

	English		Persian	
Training Set	Sentences	83042	Sentences	82496
	Words	1322470	Words	1399759
Tunings Set	Sentences	1578	Sentences	1578
	Words	40044	Words	41287
		Language Model	Sentences	5852532
			Words	66331086

TABLE 1 – Baseline System Components

[3] http://www.irna.ir/ENIndex.htm

4.1 Test Data Set

We used eight test sets based on text extracted from certain bilingual websites for our experiments, as shown in Table 2.

Testing Data Set #		1	2	3	4	5	6	7	8	Total
English	**Word**	163	218	371	362	101	354	555	259	2383
	Character	878	1381	1941	1922	589	1887	2902	1325	12825
Persian	**Word**	158	222	403	337	115	386	653	297	2571
	Character	551	955	1663	1230	430	1717	2551	1063	10160

TABLE 2 – Statistics of eight test set used in automatic and manual evaluation

Test sentences have been selected randomly covering different domains, regardless of whether or not they had potential to be covered by any post-editing rules. We performed translation in the English-Persian translation direction. The Persian side of the test sets was used as the translation reference when using scoring metrics to evaluate the output quality of both the baseline system and the final post-APE output.

4.2 Automatic Evaluation

The translation output before and after the APE is scored with BLEU, the results of which are shown in Table 3.

Input	1	2	3	4	5	6	7	8
Before APE	0.6523	0.2232	0.5914	0.1365	0.7925	0.2738	0.2945	0.4048
After APE	0.6770	0.2187	0.7388	0.1214	0.8716	0.2779	0.2951	0.4089
BLEU Difference	**0.0247**	**-0.0045**	**0.1474**	**-0.0151**	**0.0791**	**0.0041**	**0.0006**	**0.0041**

TABLE 3 - Scores of APE based on SMT Joshua version 4.0

The results generally show increases in BLEU metric, which is also shown in Figure 3. The greatest increase in BLEU score due to the APE was achieved in test set #3, with an increase of about 0.15 BLEU. However, in certain test sets the scoring metrics report a decrease in output quality, the worst BLUE score being at a difference of -0.0151.

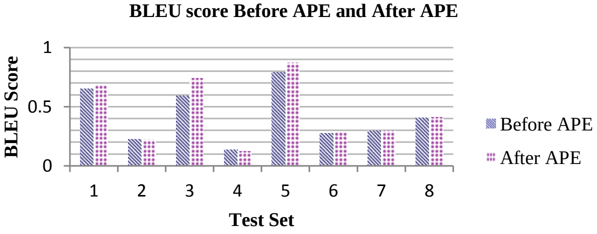


FIGURE 3 –Difference of BLEU score after applying APE on eight test sets

We propose that the weakened results are mainly due to the lack of training data for the Transliterator module in which some proper names and terms are scripted incorrectly in Persian.

Since we use the output of the SMT system, the quality of statistical translation (in terms of BLEU metric score) affects the APE module directly. Test set #4 yielded poor quality since the parallel corpus contained much less data in the religious genre. Furthermore, where there were some English words in the SMT output that OOVRemover was unable to correct, Transliterator generated a Persian script which completely changed the meaning of the original sentence.

4.3 Manual Evaluation

Marecek et al. (2011) show that grammatical correctness cannot simply be drawn from BLEU metrics alone. Because of this, we manually evaluated the proposed model. We used the same test sets as the automatic evaluation containing 153 sentences and the sentences were translated using SMT and post-edited by the proposed APE system. We assigned the APE output to two separate annotators, who were to rank the APE output based on the following criteria:

No Change: There is no difference to APE output and SMT output.
Improved: There are certain changes improving fluency.
Weakened: There are certain changes decreasing fluency.

The results of the manual evaluation are shown in Table 4.

Annotator/Rank	Improved	No Change	Weakened
Annotator 1	47	95	11
Annotator 2	43	99	11

TABLE 4 – Scores of two human evaluators for 153 test sentences

Both annotators completed the evaluation separately, but had very similar judgments of the APE system's output. The results show an improvement of the quality of the baseline SMT system output by 29.4% and that the rules developed in the APE system are not applicable to more than a half (63.4%) of the SMT output. On the other hand, human evaluation also shows that in some cases, the output is weakened after applying APE.

Both annotators' scores (Table 5) show a sentence quality improvement of 25% due to the APE.

I / II	Improved	No Change	Weakened
Improved	39	5	5
No Change	3	90	2
Weakened	3	4	4

TABLE 5 – Mutual score for both human evaluator I and evaluator II

Conclusion

We present an uncommon APE model for English-Persian statistical machine translation modeled on a rule-based approach in different levels of transformation. The automatic and manual evaluation results show encouraging improvement in quality of translation after post editing. While the improvement in some test sets is small, it still improves the SMT output up to 0.15 BLEU. Manual evaluation scores show that a rule-based APE system can yield even better results. From our results we can see at least a 25% improved output for a loss of at most 7%.

References

Allen, J., & Hogan, C. (2000). *Toward the Development of a Post editing Module for Raw Machine Translation Output: A Controlled Language Perspective*.

Buchholz, S., & Marsi, E. (2006). *CoNLL-X shared task on multilingual dependency parsing*.

Chiang, D. (2005). *A hierarchical phrase-based model for statistical machine translation*.

Ganitkevitch, J., Cao, Y., Weese, J., Post, M., & Callison-Burch, C. (2012). Joshua 4.0: Packing, PRO, and paraphrases.

Hudson, R. A. (1984). *Word grammar*: Blackwell Oxford.

Koehn, P., Och, F., & Marcu, D. (2003). *Statistical phrase-based translation*.

Kübler, S., McDonald, R., & Nivre, J. (2009). Dependency parsing. *Synthesis Lectures on Human Language Technologies, 1*(1), 1-127.

Lagarda, A. L., Alabau, V., Casacuberta, F., Silva, R., & Diaz-de-Liano, E. (2009). *Statistical post-editing of a rule-based machine translation system*.

Li, Z., Callison-Burch, C., Khudanpur, S., & Thornton, W. (2009). Decoding in Joshua. *The Prague Bulletin of Mathematical Linguistics, 91*, 47-56.

Marcu, D., & Wong, W. (2002). *A phrase-based, joint probability model for statistical machine translation*.

Marecek, D., Rosa, R., & Bojar, O. (2011). *Two-step translation with grammatical post-processing*.

McDonald, R., Pereira, F., Ribarov, K., & Hajic, J. (2005). *Non-projective dependency parsing using spanning tree algorithms*.

Mohaghegh, M., & Sarrafzadeh, A. (2012). *A hierarchical phrase-based model for English-Persian statistical machine translation*.

Mohaghegh, M., Sarrafzadeh, A., & Moir, T. (2011). *Improving Persian-English Statistical Machine Translation: Experiments in Domain Adaptation*.

Mohamed, A. A. E. M. (2000). Machine Translation of Noun Phrases from English to Arabic. *Faculty of Engineering, Cairo University, Giza*.

Och, F., & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics, 30*(4), 417-449.

Och, F., Tillmann, C., & Ney, H. (1999). *Improved alignment models for statistical machine translation*.

Pilevar, A. H. (2011). USING STATISTICAL POST-EDITING TO IMPROVE THE OUTPUT OF RULE-BASED MACHINE TRANSLATION SYSTEM. *Training, 330*, 330,000.

Raja, F., Amiri, H., Tasharofi, S., Sarmadi, M., Hojjat, H., & Oroumchian, F. (2007). Evaluation of part of speech tagging on Persian text. *University of Wollongong in Dubai-Papers*, 8.

Rasooli, M. S., Moloodi, A., Kouhestani, M., & Minaei-Bidgoli, B. (2011). *A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank*.

Rosa, R., Marecek, D., & Duˇsek, O. (2012). *DEPFIX: A System for Automatic Correction of Czech MT Outputs*.

Simard, M., Goutte, C., & Isabelle, P. (2007). Statistical phrase-based post-editing.