A Hierarchical Phrase-Based Model for English-Persian Statistical Machine Translation

Mahsa Mohaghegh
Massey University
School of Engineering and advanced Technology
Auckland, New Zealand
m.mohaghegh@massey.ac.nz

Abdolhossein Sarrafzadeh
Unitec
Department of Computing
Auckland, New Zealand
hsarrafzadeh@unitec.ac.nz

Abstract— **In this paper we show that a hierarchical phrase-based translation system will outperform a classical (non-hierarchical) phrase-based system in the English-to-Persian translation direction, yet for the Persian-to-English direction, the classical phrase-based system is preferable. We seek to explain why this is so, and detail a series of translation experiments with our SMT system using various bilingual corpora each with both toolkits Moses (non-hierarchical) and Joshua (hierarchical).**

Keywords-component; statistical machine translation, natural language processing, hierarchical phrase-based models

I. Introduction

Most recent research in the area of statistical machine translation has been targeted at modelling translation based on phrases in both the source language, and matching them with their statistically-determined equivalents in the target language ("phrase-based" translation) – [1-4]; Many modern successful translation machines use this translation approach.

A significantly critical task in a phrase-based MT system is the determination of a translation model from a word-aligned parallel corpus. A phrase table containing the source language phrases, their target language equivalents and their associated probabilities, in most systems is extracted in a preprocessing stage before decoding a test set ([1, 5].

Moses toolkit [6]is an open source phrase-based toolkit, and uses such a preprocessing approach in their training scripts. Hierarchical phrase-based translation [7]expands on phrase-based translation by allowing phrases with gaps, modelled as synchronous context-free grammar (SCFG). The original hierarchical implementation trains its SCFG translation model in a pre-processing stage similar to standard phrase-based models. A subsample of occurrences of given source phrase are used to calculate translation probabilities. Phrase translation and their model parameters can be determined at run-time as the system accesses the target language corpus and word alignment data. A suffix array can also be used to obtain hierarchical phrases at run time [8].Joshua is another well-known open source machine translation toolkit [9] Using Joshua, sentences can be translated using an aligned parallel corpus without the need to extract an SCFG prior to decoding. This implementation enables any input sentence to be decoded, and data structures are not as large as full phrase tables, using less disk space. Due to this however, the decoder has a slower running time as phrase translations must take place while running.

We conducted experiments with hierarchical translation models using Joshua, with a range of corpora sizes, and compared the results with classical phrase-based models using Moses with the same corpora.

II. Difficulties with Persian in an SMT System

Statistical machine translation has proven itself to be successful for a number of language pairs. However, as soon as the Persian language is involved with any sort of machine translation, a number of difficulties are encountered. Of other common languages, English seems to be the best language to pair with Persian, since it is best supported by resources such as large corpora, language processing tools, and syntactic tree banks, not to mention it is the most widely used language online and in the electronic world in general. Persian is the complete opposite, with a significant shortage of digitally available text, both parallel and monolingual. Other language pairs make use of parallel corpora of many millions, even billions of sentences, giving any applied system a huge database to work from, and thus output much more accurate results.

The Persian-English pair poses several unique challenges. Persian is morphologically rich, with many characteristics not shared by other languages. Persian makes no use of articles ('a', 'an', 'the'), there is no distinction between capital and lowercase letters, and symbols and abbreviations are rarely used. Sentence structure is also different, Persian placing parts of speech such as nouns, subjects, adverbs and verbs in different locations in the sentence, and sometime even omitting them altogether. Some Persian words have many different (yet correct) versions of spelling, and it is not uncommon for translators to "invent" new words. This can result in an OOV (out-of-vocabulary) output. The difference between colloquial and formal Persian is also much greater than that of English. Any SMT system designed for this language pair needs to take all the characteristic differences between the languages into consideration, and construct specifics of the system to cater for these differences. Areas requiring special attention due to these language differences arise in the task of alignment.

III. JOSHUA TOOLKIT

Joshua is a general-purpose open source toolkit used for parsing-based machine translation, accomplishing the same purpose as Moses toolkit [6]does for regular phrase-based machine translation. The toolkit is written in Java and implements all the essential algorithms described in [7]: chart-parsing, n-gram language model integration, beam and cube pruning, and k-best extraction. The toolkit also implements suffix-array grammar extraction [8]and minimum error rate training [10]Additionally, parallel and distributed computing techniques are exploited to make it scalable [9].The toolkit was constructed to be user-friendly and readily extendable.

IV. DATA PREPARATION

In order to provide the best possible results, a statistical language model requires an extremely large amount of data, and this to be trained in order to obtain proper probabilities. For the purpose of this paper, we used IRNA as a monolingual corpus for training SMT translation from English to Persian. For the Persian to English translation direction we used the news commentary monolingual corpus. IRNA corpus, consisting of about 6 million sentences was derived from the Islamic Republic News Agency.

TABLE 1 – MONOLINGUAL CORPORA COMPOSITION

Monolingual	Data Genre	Sentences	Words
News-Commentary	News	18911860	44904370
IRNA	News	5852532	66331086

The test set consisted of 2K sentences with one human translation as a reference. This same test set was used in both directions of translation.

As far as we know, the only large, freely available parallel corpus available for the English-Persian language pair is the TEP corpus, developed on slang words with public domain, extracted from movie subtitles, and consisting of about 5.3M sentences of 7.8M words. This corpus, and another corpus privately obtained (MPEC) consisting of about 50K sentences, were concatenated together to form a single corpus of about 5.4M words (NSPEC) for use in one branch of tests.

Our tests used the MPEC corpus divided into sections of 20K, 30K, 40K, and 50K sentences, the NSPEC corpus, and also the TEP corpus in a separate test, every corpus used with both Moses and Joshua toolkits.

TABLE 2 – PARALLEL CORPA COMPOSITION

Language Pair En-Pe	**Data Domain**	**English**		**Persian**	
		Sentences	Words	Sentences	Words
20K	Newswire	20121	353703	20615	364967
30K	Newswire	30593	465977	30993	482959
40K	Newswire	40701	537336	41112	560276
50K	Newswire	52922	785725	51313	836709
NSPEC	Newswire -Subtitle	678695	5596447	665678	5371799
TEP	Subtitle	612086	3920549	612086	3810734

V. EXPERIMENT RESULTS AND EVALUATION

A. *Implementation*

Two systems are evaluated in this paper: Moses [6], and Joshua [9] – a reimplementation of Hiero. We perform translation in both directions – English – Persian and Persian - English.

In both systems, we use the default settings of Moses, i.e., we set the beam size to 200, the distortion limit to 6, we limit to 20 the number of target phrases that are loaded for each source phrase, and we use the same default eight features of Moses. In our previous work [11]we detail our specific work in the English-Persian language direction using only the Moses toolkit. Here, we also use Joshua (v1.3) with its default settings.

Our Joshua-based experiments used the Joshua implementation of the hierarchical phrase-based algorithms. Our maximum phrase length was set to 5, and maximum MERT iterations was set to 10, with the size of N-best list at 300. The language models used are 5-gram models.

As previously mentioned, the issue of word alignment in the parallel corpus in use is an area in need of much attention. Sentence-aligned parallel corpora are useful for the application of machine learning to machine translation, however unfortunately it is not usual for parallel corpora to originate in this form. Since there was a great shortage (comparatively) of bilingual text for Persian-English, great care needed to be taken to ensure that the text that was available was the best possible quality. Several different methods are able to perform alignment. Desirable characteristics of an efficient sentence alignment method include speed, accuracy and no need for prior knowledge of the corpus or the languages in the pair.

In our experiments using the Joshua toolkit, we used the Berkeley aligner, whereas with the Moses toolkit, we used the Microsoft bilingual aligner and later Giza ++[12]. All the corpora used in each test, in both the Moses and Joshua experiments were aligned on sentence level, and tokenized.

B. *Results*

In this section we discuss the results we achieved, and compare Moses and Joshua over our five systems that we detailed in chapter 4, Data Preparation. In the first stages of the test we apply Moses and Joshua for the Persian-English translation direction. We trained our machine on five different systems, each with a different corpus (Table 2). We also used news commentary for building a language model (Table 1). The language model in both systems was smooth, with a modified Kneser-Ney algorithm, and implemented in SLRIM [13]. We trained language models up to 5-grams. In our Joshua tests, we used N-best list of size 300. In the final evaluation, we report results using both BLEU and NIST evaluation scores.

We start by comparing the translations yielding the best configuration generated by both Joshua and Moses. As seen in (Tables 3 & 4), in system 50K we achieve the best score, where the BLEU score for Moses shows a better result in comparison to Joshua. The same trend is also observed in the NIST score for 50K. In (Tables 5 & 6) in 50K the NIST score for Moses is 4.4925 and for Joshua is 4.5269, and BLEU scores Moses at 0.3496 and Joshua at 0.3708. As you will observe, here Joshua achieves a better score in both BLEU and NIST when compared to Moses. One of the major differences between English and Persian is the word order. As previously mentioned, Persian as the target language possesses some features that negatively affect MT performance. It is rich in morphology, much more so than English, and there is greater noise in training data, and harder sparse-data problems due to vocabulary that combines words from various sources. Persian, being rich in morphology on the target side means that besides selecting a lexically correct Persian equivalent of an English word the SMT system must also correctly guess grammatical features. This means that significant reordering must take place during translation. Hierarchical phrase-based translation is based on synchronous context-free grammars (SCFG). Like classical phrase-based translation, pairs of corresponding source and target language phrases (sequences of tokens) are learnt from training data. The difference is that in hierarchical models, phrases may contain "gaps", and are represented by non-terminal symbols of the SCFG. If a source phrase contains a non-terminal, then the target phrase will also contain that non-terminal, and the decoder can replace the non-terminal by any source phrase and its translation respectively.

This follows the observation that hierarchical models have been shown to produce better translation results than classic phrase-based models [7].

As far as automatic evaluation is concerned, the best result report in this paper is 4.5269 NIST and 0.3708 BLEU using the Joshua based system trained on 50K corpus. Moses was not able to outperform these scores, despite its ability to learn factored models. The best Moses score is 4.4925 NIST and 0.3496 BLEU. Our Moses and Joshua systems are trained in identical conditions: both the translation and the language model are trained on the same monolingual corpus (IRNA) for the English-Persian direction, and news commentaries for the Persian-English direction. We wished to confirm that in Moses more data is more important, although in NSPEC and TEP corpora we didn't achieve a higher score compared to smaller size corpora, due to the differences in domain. We see that while the BLEU score indicates the superiority of the hierarchical model over the phrase based model in the English to Persian direction, we didn't achieve the same for the Persian to English direction.

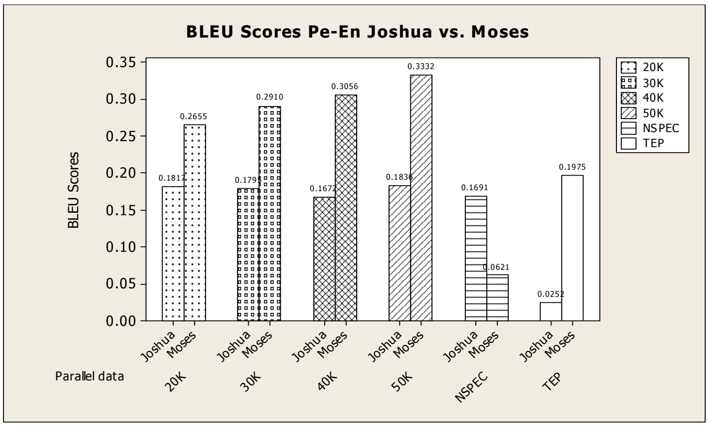

Figure 1. BLEU Scores Pe-En Joshua Vs. Moses

TABLE 2- BLEU SCORES Pe-En JOSHUA VS.MOSES

Parallel data	*Joshua*	*Moses*
20K	0.1817	0.2655
30K	0.1795	0.2910
40K	0.1672	0.3056
50K	**0.1836**	**0.3332**
NSPEC	0.1691	0.0621
TEP	0.0252	0.1975

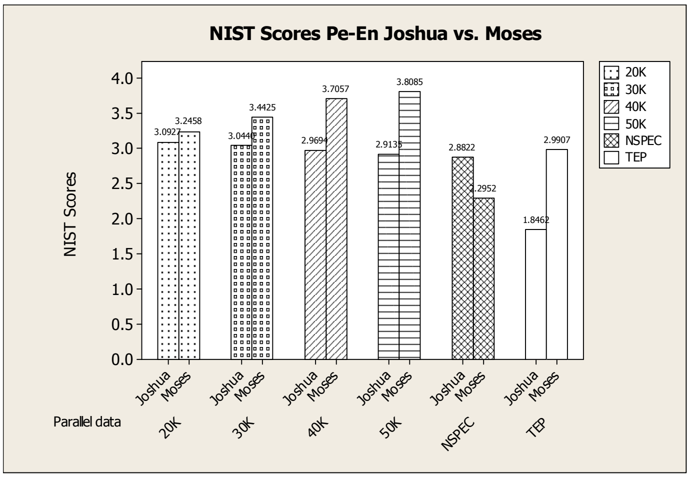

Figure 2. NIST Scores Pe-En Joshua Vs. Moses

TABLE 4- NIST SCORES Pe-En JOSHUA VS.MOSES

Parallel data	*Joshua*	*Moses*
20K	3.0927	3.2458
30K	3.0440	3.4425
40K	2.9694	3.7057
50K	**2.9135**	**3.8085**
NSPEC	2.8822	2.2952
TEP	1.8462	2.9907

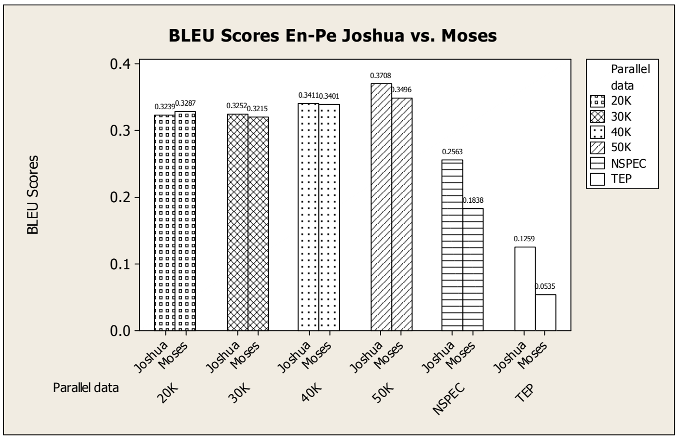

Figure 3. BLEU Scores En-Pe Joshua Vs. Moses

TABLE 5- BLEU SCORES En-Pe JOSHUA VS.MOSES

Parallel data	*Joshua*	*Moses*
20K	0.3239	0.3287
30K	0.3252	0.3215
40K	0.3411	0.3401
50000	**0.3708**	**0.3496**
NSPEC	0.2563	0.1838
TEP	0.1259	0.0535

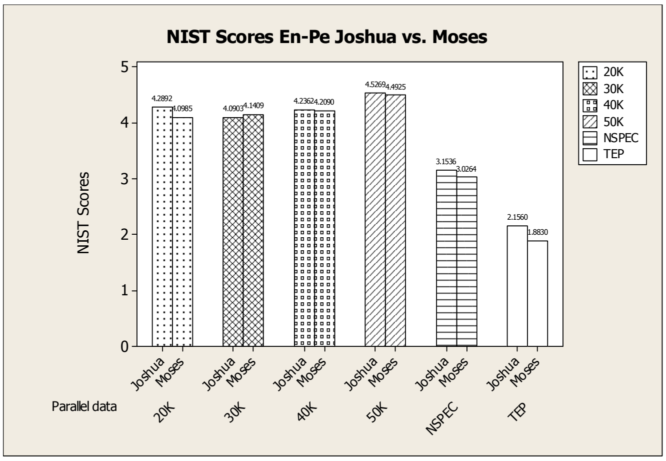

Figure 4. NIST Scores En-Pe Joshua Vs. Moses

TABLE 6- NIST SCORES En-Pe JOSHUA VS.MOSES

Parallel data	*Joshua*	*Moses*
20K	4.2892	4.0985
30K	4.0903	4.1409
40K	4.2362	4.2090
50000	**4.5269**	**4.4925**
NSPEC	3.1536	3.0264
TEP	2.1560	1.8830

VI. CONCLUSION

We showed the different behaviour of English/Persian language SMT towards a conventional phrase-based model and a hierarchical model. We observe several strange results. Adding more training data to the system for both translation directions either helps significantly, or (more often) brings down the BLEU score. Both BLEU and NIST scores improved when we trained with Joshua in the English-Persian direction, whereas Moses had a better performance in the Persian-English direction. In our future work we want to explore problems with existing data sets, the issue of morphology and its relation to output quality by combining those models together. Hierarchical decoder Joshua can capture word order even better than Moses. Its results tend to be always slightly better in the English to Persian direction, and as far as we know, our current result is the best that has been recorded for this language pair.

REFERENCES

[1] P. Koehn, F. Och, and D. Marcu, "Statistical phrase-based translation," 2003, pp. 48-54.

[2] D. Marcu and W. Wong, "A phrase-based, joint probability model for statistical machine translation," 2002.

[3] F. Och, C. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation," 1999, pp. 20–28.

[4] F. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics,* vol. 30, pp. 417-449, 2004.

[5] Y. Deng and W. Byrne, "MTTK: An alignment toolkit for statistical machine translation," 2006, p. 268.

[6] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, and R. Zens, "Moses: Open source toolkit for statistical machine translation," 2007, p. 2.

[7] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," 2005, pp. 263-270.

[8] A. Lopez, "Statistical machine translation," 2008.

[9] Z. Li, C. Callison-Burch, S. Khudanpur, and W. Thornton, "Decoding in Joshua," *The Prague Bulletin of Mathematical Linguistics,* vol. 91, pp. 47-56, 2009.

[10] F. Och, "Minimum error rate training in statistical machine translation," 2003, pp. 160-167.

[11] M. Mohaghegh, A. Sarrafzadeh, and T. Moir, "Improving Persian-English Statistical Machine Translation: Experiments in Domain Adaptation," 2011, p. 9.

[12] F. Och and H. Ney, "Improved statistical alignment models," 2000, pp. 440-447.

[13] A. Stolcke, "SRILM-an extensible language modeling toolkit," 2002.