An Overview of the Challenges and Progress in PeEn-SMT: First Large Scale Persian-English SMT System

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— This paper documents recent work carried out for PeEn-SMT, our Statistical Machine Translation system for translation between the English-Persian language pair. We give details of our previous SMT system, and present our current development of significantly larger corpora. We explain how recent tests using much larger corpora helped to evaluate problems in parallel corpus alignment, corpus content, and how matching the domains of PeEn-SMT's components affect translation outcome. We then focus on combining corpora and approaches to improve test data, showing details of experimental setup, together with a number of experiment results and comparisons between them. We show how one combination of corpora gave us a metric score outperforming Google Translate for the English-to-Persian translation. Finally, we outline areas of our intended future work, and how we plan to improve the performance of our system to achieve higher metric scores, and ultimately to provide accurate, reliable language translation.

Keywords-Statistical Machine translation- English-Persian, Language Model

I. INTRODUCTION

Machine Translation is one of the earliest areas of research in Natural Language Processing. Research work in this field dates as far back as the 1950's. Several different translation methods have been explored to date, the oldest and perhaps the simplest being rule-based translation, which is in reality transliteration, or translating each word in the source language with its equivalent counterpart in the target language. This method is very limited in the accuracy it can give. A method known as Statistical Machine Translation (SMT) seems to be the preferred approach of many industrial and academic research laboratories, due to its recent success [1]. Different evaluation metrics generally show SMT approaches to yield higher scores.

The SMT system itself is a phrase-based translation approach, and operates using a parallel or bilingual corpus – a huge database of corresponding sentences in two languages. The system is programmed to employ statistics and probability to learn by example which translation of a word or phrase is most likely to be correct. For more accurate translation results, it is generally necessary to have a large parallel corpus of aligned phrases and sentences from the source and target languages.

Our work is focussed on implementing a SMT for the Persian-English language pair. SMT has only been employed in several experimental translation attempts for this language pair, and is still largely undeveloped. This is due to several difficulties encountered with this particular language pair. Firstly, several characteristics of the Persian language cause issues with translation into English, and secondly, effective SMT systems generally rely on large amounts of parallel text to produce decent results, and there are no parallel corpora of appropriate size currently available for this language pair. These factors are prime reasons why there is a distinct shortage of research work aimed at SMT of this particular language pair.

This paper firstly gives a brief background to the Persian language, focusing on its differences to English, and how this affects translation between the two languages. Next, we give details of our PeEn-SMT system, how we developed and manipulated the data, and aligned our parallel corpora using a

hybrid sentence aligning method. We give a brief overview of previous tests with the earlier version of the system, and then show our latest experiments with a considerably larger corpus. We show how increasing the size of the bilingual corpus (training model), and using different sizes of monolingual data to build a language model affects the output of PeEn-SMT system. We focus on the aim for a general purpose translator, and whether or not the increase in corpora size will give accurate results. Next we show that with the PeEn-SMT system equipped with different language models and corpora sizes in different arrangements, different test results are presented. We explain that the improved result variations are due to two main factors: firstly, using an in-domain corpus even of smaller size than a mixed-domain corpus of larger scale; secondly, spending much focus on stringent alignment of the parallel corpus. We give an overview of the evaluation metrics used for our test results. Finally, we draw conclusions on our results, and detail our plan for future work.

II. PERSIAN LANGUAGE CHARACTERISTICS

Persian is an Indo-European language, spoken mostly in Iran, but also parts of Afghanistan, India, Tajikistan, the United Arab Emirates, and also in large communities in the United States. Persian is also known as Farsi, or Parsi. These terms are all interchangeable, and all refer to the one language. The written Persian language uses an extended Arabic alphabet, and is written from right to left. There are numerous different regional dialects of the language in Iran, however nearly all writing is in standard Persian.

There are several grammatical characteristics in written Persian which differ to English. There is no use of articles in Persian, as the context shows where these would be present. There is no capital or lowercase letters, and symbols and abbreviations are rarely used.

The subject in a Persian sentence is not always placed at the beginning of the sentence as a separate word. Instead, it is denoted by the ending of the verb in that sentence. Adverbs are usually found before verbs, but may also appear in other locations in the sentence. In the case of adjectives, these usually proceed after the nouns they modify, unlike English where they are usually found before the nouns.

Persian is a morphologically rich language, with many characteristics not shared by other languages [2]. This can present some complications when it is involved with translation into any language, not only English.

As soon as Persian is involved with statistical machine translation, a number of difficulties are encountered. Firstly, statistical machine translation of the Persian language is only recently being exploited. Probably the largest difficulty encountered in this task is the fact that there is very limited data available in the form of bilingual corpora.

The best language to pair with Persian for machine translation is English, since this language 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.

When compared to English however, Persian has many differing characteristics, some of which pose significantly difficult problems for the task of translation. Firstly, compared to English, the basic sentence structure is generally different in terms of syntax. In English, we most usually find sentence structure in its most basic form following the pattern of "subject – verb – object", whereas in Persian it is usually "subject – object – verb". Secondly, spoken Persian differs significantly from its written form, being heavily colloquial, to a much greater degree than in English. Thirdly, many Persian words are spelled in a number of different ways, yet all being correct. This in particular poses trouble for translation, since if one version of the spelling is not found in a bilingual corpus, such a word may be incorrectly translated, or remain as an OOV (out of vocabulary) word.

Any SMT system designed for this language pair needs to take these details into consideration, and specifics of the system developed to cater for these differences.

III. PEEN-SMT COMPOSITION

A. SMT System Architecture

The goal of a statistical machine translation system is to produce a target sentence ***e*** from a source sentence ***f***. It is common practice today to use phrases as translation units (Koehn et al., 2003; Och and Ney 2003) in the log-linear frame in order to introduce several models explaining the translation process.

The SMT paradigm relies on the probabilities of source and target words to find the best translation. The statistical translation process is given as:

$$\begin{aligned} \mathbf{e}^* &= \underset{\mathbf{e}}{\operatorname{argmax}} \Pr(\mathbf{e}|\mathbf{f}) \\ &= \underset{\mathbf{e}}{\operatorname{argmax}} \sum_{\mathcal{A}} \Pr(\mathbf{e}, \mathcal{A}|\mathbf{f}) \qquad (1) \\ &\approx \underset{\mathbf{e}}{\operatorname{argmax}} \max_{\mathcal{A}} \Pr(\mathbf{e}, \mathcal{A}|\mathbf{f}) \qquad (2) \end{aligned}$$

In the above equations, (***A***) denotes the correspondence between source and target words, and is called an alignment. The Pr(e, ***A*** |f) probability is modeled by combination of feature functions, according to maximum entropy framework [3]

$$\Pr(\mathbf{e}, \mathcal{A}|\mathbf{f}) \propto \exp \sum_i \lambda_i f_i(\mathbf{e}, \mathcal{A}|\mathbf{f}) \qquad (3)$$

The translation process involves segmenting the source sentence into source phrases *f*; translating each source phrase into a target phrase *e*, and reordering these target phrases to yield the target sentence **e***. In this case a phrase is defined as a group of words that are to be translated [4, 5] A phrase table provides several scores that quantize the relevance of translating *f* to *e*.

The PeEn-SMT system is based on the Moses SMT toolkit, by [6]. The decoder includes a log-linear model comprising a phrase-based translation model, language model, a lexicalized

distortion model, and word and phrase penalties. The weights of the log-linear interpolation were optimized by means of MERT[5]. In addition, a 5-gram LM with Kneser-Ney [7] smoothing and interpolation was built by means of the SRILM toolkit [8].

Our baseline English-Persian system was constructed as follows: first word alignments in both directions are calculated with the help of a hybrid sentence alignment method. This speeds up the process and improves the efficiency of GIZA++ [9], removing certain errors that can appear with rare words. In addition, all the experiments in the next section were performed using a corpus in lowercase and tokenized conditions. For the final testing, statistics are reported on the tokenized and lower-cased corpora.

B. Data Development

We used two news stories monolingual English corpora, originating from Europarl Corpus [10]. The Europarl corpus is extracted from the proceedings of the European Parliament in 11 languages: Romanic (French, Italian, Spanish, and And Portuguese), Germanic (English, Dutch, German, Danish, and Swedish), Greek and Finnish.

It is common to use huge bilingual corpora with SMT. Unfortunately for the Persian; there is a significant shortage of bilingual texts.

One English-Persian parallel text corpus we obtained consisted of almost 100,000 sentence pairs of 1.6 million words, and was mostly from bilingual news websites. There were a number of different domains covered in the corpus, but the majority of the text was in literature, politics, culture and science. Figure 1 shows the corpus divided into separate domains**.** To the best of our knowledge, the only freely available corpus for the English-Persian language pair is the TEP corpus, which is a collection of movie subtitles consisting of almost 3 million sentences- 7.8 million words. These two corpora were concatenated together to form News Subtitle Persian English Corpus (NSPEC) a single corpus of 3,100,000 sentences for use in one test, and will also be used in the future for further experiments.

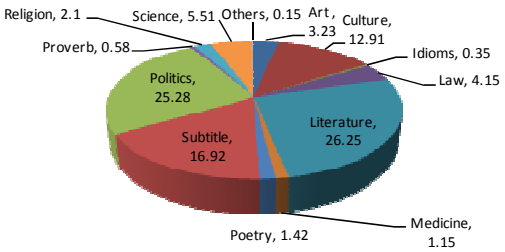

Figure 1. Domain percentages for NSPEC corpus

IV. EXPERIMENTS AND RESULTS

A. Experiments

To develop a translation model, an English-Persian parallel corpus was built as explained in Section B Data Development. We divided the parallel corpus into different sized groups for each test system. The details of the corpus size for each test are shown in Table I. Table II shows the size of each test's corpora after the text was tokenized, converted to lowercase, and stripped of blank lines and their correspondences in the corpora. This data was obtained after applying the hybrid sentence alignment method.

TABLE I. BILINGUAL CORPORA USED TO TRAIN THE TRANSLATION MODEL

Language Pair En-Pe	Data Genre	English Sentences	English words	Persian sentences	Persian Words
System1	Newswire	10874	227055	10095	238277
System2	Newswire	20121	353703	20615	364967
System3	Newswire	30593	465977	30993	482959
System 4	Newswire	40701	537336	41112	560276
System 5	Newswire	52922	785725	51313	836709
TEP	Subtitle	612086	3920549	612086	3810734
NSPEC	Subtitle-Newswire	678695	5596447	665678	5371799

TABLE II. BILINGUAL CORPORA AFTER TOKENIZED

Language Pair En-Pe	Data Genre	English Sentences	English Words	Persian sentences	Persian Words
System1	Newswire	9351	208961	9351	226759
System2	Newswire	18277	334440	18277	362326
System3	Newswire	27737	437871	27737	472679
System 4	Newswire	37560	506972	37560	548038
System 5	Newswire	46759	708801	46759	776154
TEP	Subtitles	612086	3920549	612086	3810734
NSPEC	Subtitle - Newswire	618039	5370426	618039	5137925

We constructed five different test systems increasing the size of the translation model by 10,000 sentences, up to the fifth test with 53,000 sentences. In addition to the news stories corpus, we had access to one another publicly available corpus, consisting of movie subtitles in Persian and English. This was shown to be in a completely different domain to our main corpus. So for most cases we preferred to run tests separately. Finally in NSPEC, we concatenated these two corpora, to ascertain the potential output with a combined corpus. We tested the subtitle corpus separately to see the effect of an out-of-domain corpus. In all cases, the test set consisted of a news article covering different domains and various grammatical aspects of each language.

Table III summarizes the monolingual corpora used for the construction of the language model. SRILM toolkit [8] was used to create up to 5-gram language models. We tested the baseline PeEn-SMT system against different sized aligned corpora and language models. Tables IV, V and VI show the results obtained using the Europal and News–Commentary language models respectively.

TABLE III. MONOLINGUAL CORPORA USED TO TRAIN THE LANGUAGE MODEL

Monolingual	Data Genre	Sentences	Words
Europarl	News	1658841	40624075
News-Commentary	News	18911860	44904370

B. Quality of Translation: Evaluation Metrics

One aspect of Machine Translation posing a challenge is automated evaluation metric. Most popular metrics yield scores primarily based on matching phrases in the translation produced to those in several reference translations. The metric scores differ in how they show reordering and synonyms.
BLEU is the most popular metric for both comparison of translation systems and tuning of translation models. The metrics we chose to use were BLEU, IBM-BLEU, METEOR, NIST, and TER.

C. Evaluation of the Results

Our first experiment was carried out with 10,000 sentences (System1) in the translation from Persian-English. For comparison we tested the SMT model on different models- Tables IV and V.
In each step, we increased the size of the parallel corpus by 10,000 sentences. The best result, as expected, was achieved when we trained System 5 (almost 53,000 parallel sentences) on the language model in the news commentary domain.

TABLE IV. AUTOMATIC EVALUATION METRICS OF PEEN-SMT SYSTEM

Scores	Language Model =Europarl v4					
System	BLEU_4	MULTI_BLEU	IBM-BLEU	NIST	METEOR	TER
System 1	0.1208	11.48	0.1175	2.5952	0.3841	0.7463
System 2	0.1277	16.09	0.1214	2.5592	0.4033	0.6376
System 3	0.2005	18.09	0.1936	3.5310	0.4410	0.6231
System 4	0.2415	16.23	0.2247	3.2908	0.43271	0.6449
System 5	0.2576	17.40	0.2542	3.1892	0.40149	0.6225
TEP	0.0414	3.33	0.0403	2.1196	0.2880	0.8623
NSPEC	0.1796	11.18	0.1771	3.1622	0.3950	0.6325

TABLE V. AUTOMATIC EVALUATION METRICS OF PEEN-SMT SYSTEM

Scores	Language Model =News-Commentary					
System	BLEU_4	MULTI_BLEU	IBM-BLEU	NIST	METEOR	TER
System 1	0.1318	13.35	0.1302	2.8344	0.3809	0.7535
System 2	0.2655	19.74	0.2611	3.2458	0.4470	0.6225
System 3	0.2910	20.14	0.2875	3.4425	0.4138	0.6952
System 4	0.3056	22.00	0.3018	3.7057	0.4414	0.6278
System 5	**0.3332**	**24.10**	**0.3258**	**3.8085**	**0.4685**	**0.5231**
TEP	0.0621	4.09	0.0435	2.2952	0.2978	0.8236
NSPEC	0.1975	15.50	0.1946	2.9907	0.3831	0.6429

In these experiments we used two different sized corpora, one with almost 1,700,000 sentences, and the other with almost 19,000,000 sentences. The size of the language model is important. For instance, in System 5, with the news commentary language model 18 times larger than Europarl, the BLEU score increased from 0.2576 to 0.3332. However, in TEP System, although the size of the corpus is dramatically larger than System 5, the BLEU score was not satisfactory at all. We determined that this was because the domains of the language model and bilingual corpus were completely different. In NSPEC system, which was a combination of movie subtitles and newswire domains, the score is also much lower than expected. Again it was determined that a combination of corpora of different domains would not necessarily lead to a better result. It was originally thought that the dramatic increase in the size of both models would yield a much higher metric score, since it gave the translation program more data to work with. However, these new tests proved that this was not *necessarily* always true, and corpus size alone was not synonymous with improved results.

V. CONCLUSIONS AND FUTURE WORK

In this paper we presented the development of our English/Persian system PeEn-SMT. We showed that increasing the amount of data alone cannot necessarily lead to better results. Instead, attention should be directed to the domain of the corpus, specifically we noted that keeping the domain of the corpus the same in a translation system is vital.

We wanted to evaluate how well PeEn-SMT would work in the Persian-to-English direction, as we had already performed tests and evaluation of translation in the opposite direction. In particular, we were interested to compare our results with Google's output scores. As shown, in our best translation system arrangement (System 5 – Table V), although the amount of data used in our tests was much smaller than the data Google has access to, we believe that the results from our system compare quite favorably (Tables V and VI).

TABLE VI. AUTOMATIC EVALUATION METRIC OF GOOGLE TRANSLATOR OUTPUT

Google (FA-EN)						
System	BLEU_4	MULTI_BLEU	IBM-BLEU	NIST	METEOR	TER
Google	**0.3453**	**22.56**	**0.3291**	**4.9075**	**0.5987**	**0.5072**

In the future we plan to develop a technique to find the most appropriate corpus and language model for PeEn-SMT system by detecting the domain of the input. We intend to perform tests using the matched-domain input, corpus and language models in an attempt to achieve even better translation.

REFERENCES

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

[2] K. Megerdoomian and N. M. S. U. C. R. Laboratory, *Persian Computational Morphology: A unification-based approach*: Computing Research Laboratory, New Mexico State University, 2000.

[3] A. Berger, V. Pietra, and S. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics,* vol. 22, pp. 39-71, 1996.

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

[5] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics,* vol. 29, pp. 19-51, 2003.

[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] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," 2002, pp. 181-184.

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

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

[10] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," 2005.