A Statistical Approach to English-Persian Machine Translation

Mahsa Mohaghegh
Massey University
School of Engineering and Advanced Technology
Auckland – New Zealand
+64 9 4140800 Extn 9831
m.mohaghegh@massey.ac.nz

ABSTRACT
Statistical Machine Translation has successfully been used for translation between many language pairs contributing to its popularity in recent years. It has however not been used for the English/Persian language pair. This paper presents the first such attempt and describes the problems faced in creating a corpus and building a base line system. Our experience with the construction of a parallel corpus during this ongoing study and the problems encountered especially with the process of alignment are discussed in this paper. The prototype constructed and its evaluation using the BiLingual Evaluation Understudy (BLEU) is briefly described and results are analyzed. In the final part of the paper, conclusions are drawn and work planned for the future is discussed.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Machine Translation, Language Model.*

Keywords
Statistical Machine Translation, Bi-lingual Corpora, Moses, Web 2.0

1. INTRODUCTION
Machine Translation is the process of using computers for translation from one human language to another. Machine translation was one of the first applications of natural language processing. Persian machine translation which is the focus of this paper is considered a challenge given the structure of the language and the fact that little work has been done in this area to date.

Many paradigms including rule-based, example-based, knowledge-based and statistical approaches to machine translation have been explored by researchers. The disadvantages of rule-based systems were soon to become clear. They were very expensive to build and maintain and difficult to adapt to other domains or languages. Statistical Machine Translation (SMT) seems to be the preferred approach of many industrial and academic research laboratories [1].

In recent years, so-called phrase-based machine translation approaches have become popular because they generally show better translation results. One major factor for this development is the growing availability of large monolingual and bilingual text corpora in recent years for some languages which do not include Persian. The advance of the Internet has produced many new resources for large text collections. The advantages of SMT compared to rule-based approaches lie in their adaptability to different domains and languages: once a functional system exists, all that has to be done in order to make it work with other language pairs or text domains is to train it on new data.

However, the effectiveness of SMT in translating between the language pair English and Persian needs to be explored further. This and the need for such a system has been the motivation for this study. This study is aimed at developing and evaluating the performance of an SMT system for use in translation of English and Persian texts in different domains.

2. Statistical Machine Translation

2.1 General
The goal of SMT is to produce a target sentence **e** from a source sentence **f** that maximizes the posterior probability. In other words, we want to find the string $\mathbf{e}'$ that maximises probability $P(e \mid f)$ [2].

By using Bayes Rule from equation (1):

$$P(e \mid f) = \frac{P(e)P(f \mid e)}{P(f)} \quad (1)$$

we are interested in the Persian sentence for which $P(e \mid f)$ is greatest. We therefore write:

$$e' = \arg\max_{e} P(e)P(f \mid e) \quad (2).$$

This paper was published in the proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC) 2010.

2.2 Statistical Machine Translation tools

There are a number of implementations of subtasks and algorithms in SMT and even software tools that can be used to set up a fully-featured state-of-the-art SMT system.

Moses [3] is a full-featured, open source SMT system developed at the University of Edinburgh, which allows one to train translation models using GIZA++ [4] for any given language pair for which a parallel corpus exists. This tool was used to build the baseline system discussed in this paper.

2.3 The Persian language

The Persian language (Farsi) is an Indo-European language and one of the dominant languages in the Middle East. Persian is spoken in several countries including Iran, Tajikistan and Afghanistan. It is very similar to Urdu which is spoken in Pakistan, parts of India and other parts of the world. Persian uses a script that is written from right to left. It has similarities with Arabic but has an extended alphabet and different words and/or pronunciations from Arabic.

During its long history, the language has been influenced by other languages such as Arabic, Turkish and even European languages such as English and French. Today's Persian contains many words from these languages and in some cases words from other languages still follow the grammar of their original language particularly in building plural, singular or different verb forms. Because of the special and different nature of the Persian language compared to other languages like English, the design of SMT systems for Persian requires special considerations [5].

2.4 Previous work

The only attempt at using the statistical approach to translate from Persian to English reported in the literature is the Shiraz project[6]. A Small English/Persian corpus has also been built for information retrieval [7] which was not found useful for SMT.

The Shiraz machine translation system is an MT prototype that translates Persian text into English. The project began in 1997 and the final version was delivered in 1999. Shiraz corpus is a 10 MB bilingual tagged corpus developed using on-line material for testing purposes in a project at New Mexico State University.

Hamshahri[8] is one of the most popular daily newspapers in Iran that has been publishing for more than 20 years. Hamshahri corpus is a Persian test collection that consists of 345 MB of news texts from this newspaper from 1996 to 2002 (corpus size with tags is 564 MB). This corpus contains more than 160,000 news articles on a variety of subjects (82 categories including politics, literature, art, economy, etc.). It includes nearly 417000 different words. Hamshahri corpus is used for information retrieval research.

Bijankhan corpus is a tagged corpus that is suitable for natural language processing research on the Persian (Farsi) language. This collection is gathered form daily news and common texts. In this collection all documents are categorized into different subjects (e.g. political, cultural and so on- totally 4300 subjects). The Bijankhan collection contains about 2.6 millions manually tagged words with a tag set of 40 POS tags.

FLDB is another Persian corpus comprising a selection of contemporary modern Persian literature, formal and informal spoken varieties of the language, and a series of dictionary entries and wordlists. It consists of about 3 million sentences. The comprehensiveness of FLDB presents it as a well-structured modern Persian corpus. However, its size isn't good enough for extensive information retrieval research [9].There has been very little work done in the area of SMT for Persian. The authors are however aware of the increasing interest in the topic.

2.5 Building a baseline SMT system

To build a good baseline system it is important to build a sentence aligned parallel corpus which is spell-checked and grammatically correct for both the source and target language. The alignment of words or phrases seems to be the most difficult problem SMT faces. Words and phrases in the source and target languages normally differ in where they are in a sentence. Words that appear on one language side may be dropped on the other. One English word may have as its counterpart a longer Persian phrase and vice versa.

The accuracy of statistical machine translation (SMT) relies heavily on the existence of large amounts of data which is commonly referred to as a parallel corpus. However, when a low or medium density language such as Persian comes to be one of the languages involved in a Parallel corpus, the case is much more difficult due to shortage of digitally stored materials and usable bilingual pages on the Web.

Building a parallel corpus for any domain is generally the most time consuming process as it depends on the availability of parallel texts. There has not been much work done in the construction of bilingual corpora involving Persian texts and there is thus not much previous work on Persian SMT. The first step we took was to develop the parallel corpus. This corpus is intended to be an open corpus in which more text can be added as they are collected. Sentences were aligned using Microsoft's bi-lingual sentence aligner developed by Moore (2002)[10].

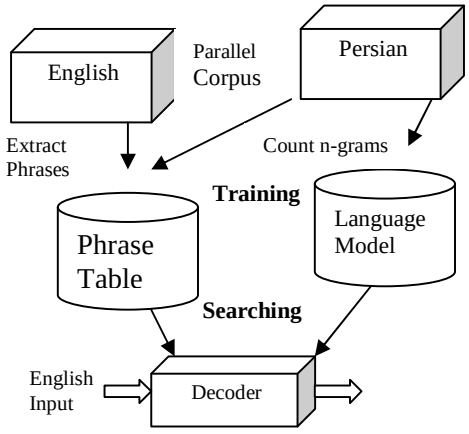

Figure 1: Schematic overview of an SMT system and its components

A language model (LM) is usually trained on large amounts of monolingual data in the target language to ensure the fluency of the language that the sentence is translated into. The SRILM toolkit developed by Stolcke (2002)[11] was used to train a 5-gram language model for experimentation purposes. An overview of an SMT system and its components is shown in Figure 1.

Test No.	1 En/ Fa	2 En/Fa	3 En/Fa	4 En/Fa	5 Fa/En	6 En/Fa	7 Fa/En
Test Sentences	730	864	1011	1011	1011	2343	2343
Train Sentences	864	1066	864	5514	7005	5514	7005

3. Experiments and Results

3.1 Experiment setup

We used Moses[1] [3]as the phrase-based SMT system. This included n-gram language models trained with the SRI language modeling toolkit [11], GIZA++ [4]word alignment software, the Moses decoder and the included script for inducing phrase-based translation models from word-based ones. The automatic evaluation metric, used in the experiments is BLEUr1n4c[2].

3.2 Evaluation metrics

It is expensive and time-consuming to use humans to evaluate the quality of machine translation and difficult to sustain any consistency in the process. Over the past several years, a number of automated means of measuring translation quality have been used. One of the most popular metrics is called BLEU (BiLingual Evaluation Understudy) which was developed by a team at IBM's Watson Research Lab. The BLEU system awards a score between 0 to 1 depending on how close a machine translation output is to that produced by a professional human translator.

3.3 Discussion and analysis of the results

A baseline system was built using Moses in this study. The system was trained and tested with an in-house corpus and repeated as the corpus size grew. The data available was split into a training and test set. Both sets were aligned using the Microsoft bi-lingual sentence aligner developed by Moore (2003). The test set was manually prepared. Blank lines and lines with a word in between were deleted. Alignment was also done manually with the aim of improving the results. Various experiments were conducted as we continued to increase the corpus size (see Table 1). In order to evaluate the Persian and English translation, two test sets were assigned to that purpose and LMs were built using 5514 sentences in English and 7005 sentences in Persian. Evaluation results from these experiments are presented in Table 2 As expected BLEU scores improved as the size of the corpus increased. However, because of the small size of the corpus the results obtained are not satisfactory when compared to those of other SMT systems for other languages.

Because of the particular features of the Persian language including the script being written from right to left and the different character sets used in Persian and English and also writing styles there were a lot of problems like the large difference between the number of sentences in the original and translated texts available and the differences in the types and symbols used for punctuation. These issues had to be resolved before any attempt at SMT could be made. Needless to stress on the fact that the better the alignment the better the results of the translation.

Table 1: Size of the test and train set (LM) - En: English, Fa: Persian

The first experiment was done on a corpus of 730 sentences in Persian and the same number for their translation in English. The training set used was 864 sentences. Results of translation were evaluated using the BLEUrln4c metric.

In the second experiment the same number of sentences was used for building a corpus but the Language Model was constructed with a Persian text collection comprising of 1066 sentences. As expected the results improved.

The same experiment was repeated with a larger number of sentences. Tests 4 and 5 were repeated for both languages but with a language model that was constructed using a collection of 50514 English sentences and 7004 Farsi sentences. The results were however very similar to previous round of testing. There was a small increase in the BLEU scores when a set of 2343 sentence pairs were used. The increase in the BLEU score as the number of sentence pairs used for training increases is shown in Table 2.

Table 2: Translation quality of SMT trained/tested on different corpora by BLEUr1n4c-EN: English, FA: Persian.

Test No.	1	2	3	4	5	6	7
N-gram Precision	En/Fa	En/Fa	En/Fa	En/Fa	Fa/En	En/Fa	Fa/En
1-gPrec	0.059	0.055	0.089	0.016	0.299	0.099	0.1287
2-gPrec	0.002	0.002	0.004	0.008	0.010	0.005	0.0050
3-gPrec	0.001	0.001	0.002	0.004	0.002	0.002	0.0025
4-gPrec	0.000	0.006	0.001	0.002	0.001	0.001	0.0013
PrecScore	0.002	0.003	0.005	0.006	0.009	0.006	0.0067
BLEUr1n4c	0.0029	0.0031	0.0057	0.0060	0.0062	0.0063	0.0067

It must be noted that BLEU is only a tool to compare different machine translation systems. So an increase in BLEU scores would not necessarily mean an increase in the accuracy of translation.

4. Future work

The accuracy should further increase if we categorize the corpus into different domains. At the moment our corpus includes different genres like news, short stories and poetry. Incorporating linguistic inputs like part-of-speech tagging, parsing, morphological analysis, semantic model and a dictionary specific to the domain would make such a system more robust in terms of accuracy and is going to be explored in this project in the future. More research needs to be done in the area of aligning of the text in the corpus. We intend to use a crawler with the aim of finding and using bilingual texts from the Web and work on this has already progressed.

5. Conclusion

This paper describes a set of experiments in which SMT was applied to the Persian language. The first part of this work was to test how well SMT translates from Persian to English when trained on the available corpora and to spot and try and resolve problems with the process and the output produced. The second

[1] http://www.statmt.org/moses

[2] the BLEU scores reported throughout this paper are for Case-sensitive BLEU. The number of references used is also reported (e.g., BLEUr1n4c: r1 means 1 reference, n4 means up to 4-gram are considered, c means case sensitive).

part of the study was to compare different sized parallel corpora for this language pair, and to find the extent to which increasing the size of the resulting SMT models affected the results. Both the size of the corpus and the collection used for building the LM affect the translation. The size of the corpus is however far more important. A number of problems occur when trying to align English and Persian sentences which require more investigation.

6. REFERENCES

[1] A. Schmidt, "Statistical Machine Translation Between New Language Pairs Using Multiple Intermediaries," 2007.

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

[3] 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.

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

[5] A. Aleahmad, P. Hakimian, F. Mahdikhani, and F. Oroumchian, "N-gram and local context analysis for persian text retrieval," 2007, pp. 1-4.

[6] J. Amtrup, C. R. Laboratory, and N. M. S. University, *Persian-English machine translation: An overview of the Shiraz project*: Computing Research Laboratory, New Mexico State University, 2000.

[7] S. Karimi, "Machine Transliteration of Proper Names between English and Persian," RMIT University, Melbourne, 2008.

[8] A. AleAhmad, H. Amiri, F. Oroumchian, and M. Rahgozar, "Hamshahri: A standard Persian text collection," *White Paper*.

[9] S. Assi, "Farsi linguistic database (FLDB)," *International Journal of Lexicography*, vol. 10, p. 5, 1997.

[10] R. Moore, "Fast and accurate sentence alignment of bilingual corpora," *Lecture notes in computer science*, pp. 135-144, 2002.

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