

Performance evaluation of various training data in English-Persian Statistical Machine Translation

Mahsa Mohaghegh *, Abdolhossein Sarrafzadeh **

* Masey University, ** Unitec

Abstract

Globalization and the continued increase in international travel and commerce have made automatic translation systems an attractive area of research and development. Even as technology opens up e-commerce opportunities, companies must overcome language barriers to reach new potential customers and business partners. With the advent of Web2.0 technologies, machine translation and tools like [Google Translate](#) have made the web more accessible. Machine translation is usually employed to translate text from one language into another. Statistical Machine Translation has 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 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.

Keywords: Natural Language Processing, SMT, Statistical machine Translation

1. Introduction

The web is a global community with rapidly growing international markets. A new Web 2.0 option is Machine translation. Machine Translation is the process of using computers for translation from one human language to another. Today many providers offer it for free. Recently, Google released its Translation Toolkit enabling the translation of HTML and DOC files via an easy to use web portal that offers a WYSIWYG translation environment integrated with its Machine Translation engine.

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 (Lopez, 2008). 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 (Schmidt, 2007).

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 in Persian requires special considerations (AleAhmad et al.).

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 (Amtrup et al., 2000). A Small English/Persian corpus has been built for information retrieval which was not found useful for SMT (Karimi et al., 2007).

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 is one of the most popular daily newspapers in Iran that has been publishing for more than 20 years. Hamshahri corpus is a Persian text 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 417,000 different words. Hamshahri corpus is used for information retrieval research (Darrudi et al., 2004).

Bijankhan corpus is a tagged corpus that is suitable for natural language processing research on the Persian (Farsi) language. This collection is gathered from daily news and common texts. In this collection all documents are categorized into different subjects (e.g. political, cultural and so on- totally 4,300 different 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 Farsi corpus. However, its size isn't good enough for extensive information retrieval research (Assi, 1997). 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 turns out to be the most difficult problem SMT faces.

Words and phrases in the source and target languages normally differ in where they are placed 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 not much previous work on Persian SMT. The first step we have taken 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).

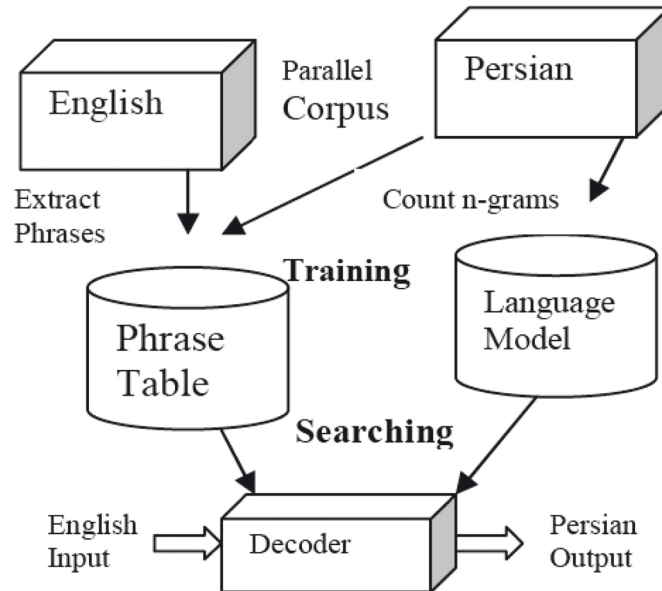


Figure 1: Schematic overview of an SMT system and its components

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 was used to train a 5-gram language model for experimentation purposes, as in (Stolcke, 2002).

3. Experiments and Results

3.1. Experiment setup

We used Moses (available online ¹) as the phrase-based statistical MT system development tool[3]. This included n-gram language models trained with the SRI language modelling toolkit, GIZA++ alignment tool, Moses decoder and the script for inducing phrase-based translation models from word-based ones. The automatic evaluation metric, used in the experiments is BLEUr1n4c ² (Stolcke, 2002).

3.2. Evaluation metrics

It is both expensive and time-consuming to evaluate the quality of machine translation and difficult to ensure that the process remains consistent when humans are used to perform this

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).

task. 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) developed at IBM's. The closer a machine translation is to a professional human translation, the better it is. This is the central idea behind the BLEU metric. The BLEU system gives a score between 0 and 1 depending on how close a machine translation output is to translations produced by a professional human translator.

NIST is another automatic evaluation metric. NIST has a score range between 0 and 100.

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. Corpus and training sets were aligned using the Microsoft bilingual sentence aligner developed by Moore (2002). 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. Evaluation results from these experiments are presented in Tab. 2. As expected BLEU scores improved as the size of the corpus increased. However, the small size of the corpus was a concern. The study will aim to grow the corpus size as a part of its future work.

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

Table 1: Size of test set and train set (language Model) En: English, Fa: Farsi

The first test was performed 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 an excerpt from the output of this first experiment is shown in Tab. 2.

Test No.	1	2	3	4	5
N-gram Precision	En/Fa	En/Fa	En/Fa	En/Fa	En/Fa
1-gPrec	0.059	0.055	0.089	0.016	0.099
2-gPrec	0.002	0.002	0.004	0.008	0.005
3-gPrec	0.001	0.001	0.002	0.004	0.002
4-gPrec	0.000	0.006	0.001	0.002	0.001
Prec Score	0.002	0.003	0.005	0.006	0.006
BLEUrln4c	0.002	0.003	0.005	0.006	0.0063

Table 2: Translation quality of SMT trained/tested on different corpora measured by BLEUrln4c En: English, FA: Farsi

In the second test 864 sentences were used for building a corpus but the Language Model was constructed with a Persian text collection comprising of 1.066 sentences. As shown below the results improved.

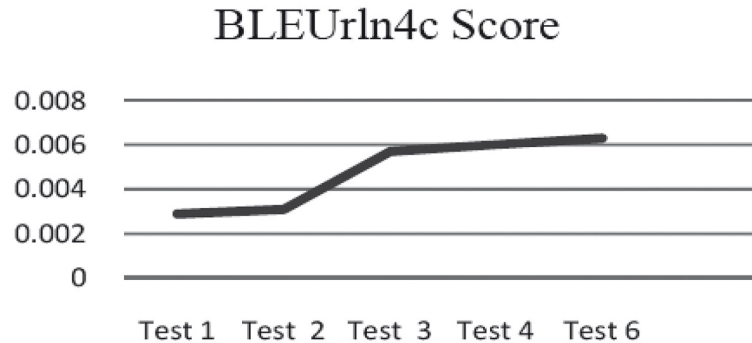


Figure 2 : Corpus Size vs. BLEU scores (EN-PE)

The same experiment was repeated with a larger number of sentences. Tests 3 and 4 were repeated for both languages but with a language model that was constructed using a collection of 864 and 7.005 Persian 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 2.343 sentence pairs were used. The increase in the BLEU score as the number of sentence pairs used for training increases is shown in Tab. 2 and Fig. 2. It must be noted that BLEU is only a tool to compare different machine translation systems. So an increase in BLEU scores may not necessarily mean an increase in the accuracy of translation.

In the initial part of the experiment corpuses of different sizes and language models of varying sizes were trialled with and the resulting translations were compared using BLEUrln4c measure.

The Second experimental part of this work consists of applying phrase-based statistical MT to a parallel English-Persian corpus is drawn from the BBC Persian News and from the United Nation, a web site which collects political commentary in multiple languages. The performance of the baseline English-Persian SMT system was evaluated by computing BLEU, IBM-BLEU-NIST (Li et al., 2009) scores from different automatic evaluation metrics against different sizes of the sentence aligned corpus and different sizes of the training set.

Tab. 3, 4 and 5 show the results obtained using corpuses of 817, 1.011, and 2.343 sentences respectively. The language model size was varied from 864 to 1.066 and finally to 7.005 sentences.

<i>Corpus size =817 sentences</i>			
Training set (LM)	864	1066	7005
BLEU	0.1061	0.0920	0.0805
NIST	1.8218	1.6838	1.6721
IBM-BLEU	0.0060	0.0060	0.0063

Table 3: Result obtained using corpus size=817

As evident from Tab. 3, with an increase in the language model size the quality of translation measured as shown in the table decreased. This may mean that proportionality of sizes of the language model and the corpus can help improve translation quality.

<i>Corpus size =1011 sentences</i>			
Training set (LM)	864	1066	7005
BLEU	0.0882	0.0986	0.0888
NIST	1.5338	1.5301	1.5512
IBM-BLEU	0.0050	0.0050	0.0051

Table 4: Result obtained using corpus size=1011

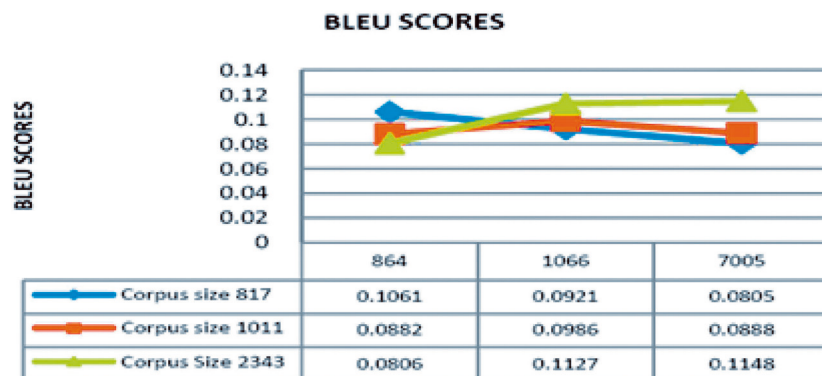
Moreover as shown in Tab. 4, using a corpus and language model of 1.011 and 1.066 in size respectively produces better results. This can clearly be noticed from graph in Fig. 3(b).

Finally, increasing the size of the corpus to 2.343 and language model constructed using 7.005 sentences produced the best translation results as shown in both Fig. 3(c) and Tab. 5. This shows that with an increase in the size of the corpus the quality of translation as measured improves provided that the size of the language model is proportional to the corpus size. Given the specifics of the Persian language, the experiments conducted to this study showed that in applying SMT to this language although the size of corpus affects the quality of the translation as measured using the BLEU metric, this is only true if the size of the corpus is proportional to that of the language model used. The literature refers to the fact that the size of the corpus although important does not have as great an effect as corpus and language model in the domain of translation (Ma and Way, 2009).

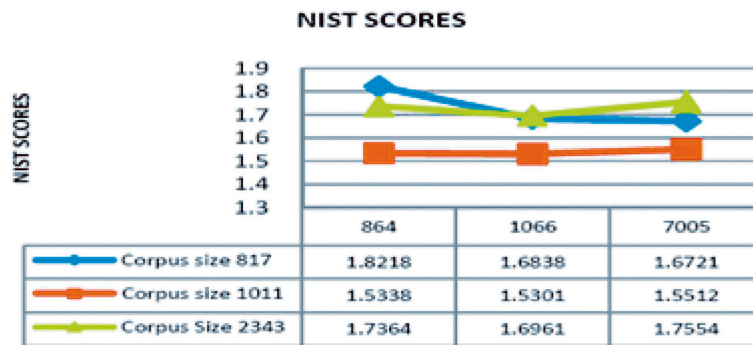
Because of the particular features of the Persian language including the script being written from right to left and the different character sets used in Persian and English and also the writing styles, there were problems like the large difference between the number of sentences in the original and 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.

<i>Corpus size =2343 sentences</i>			
Training set (LM)	864	1066	7005
BLEU	0.0806	0.1127	0.1148
NIST	1.7364	1.6961	1.7554
IBM-BLEU	0.0067	0.0069	0.0071

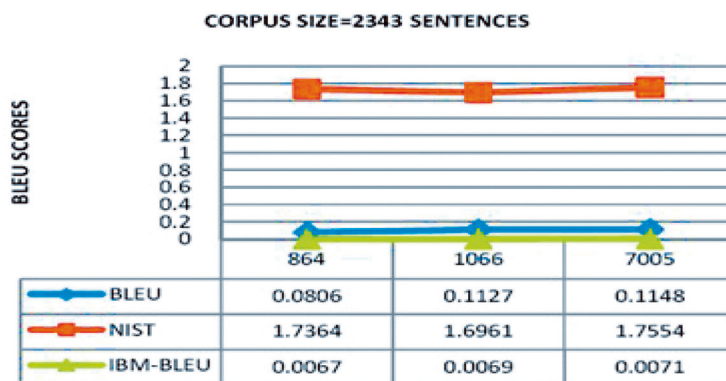
Table 5: Result obtained using corpus size=2.343



(a)



(b)



(c)

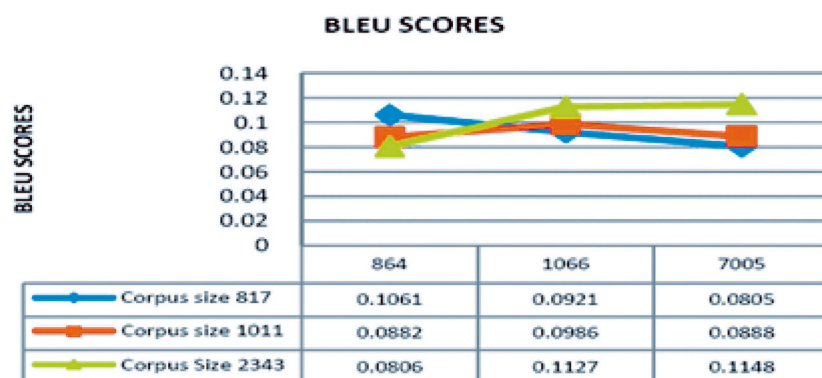
Figure 3: (a) Result obtained using corpus size=864 (b) result obtained using corpus size=1.011(C) result obtained using corpus size=2.343

4. Future work

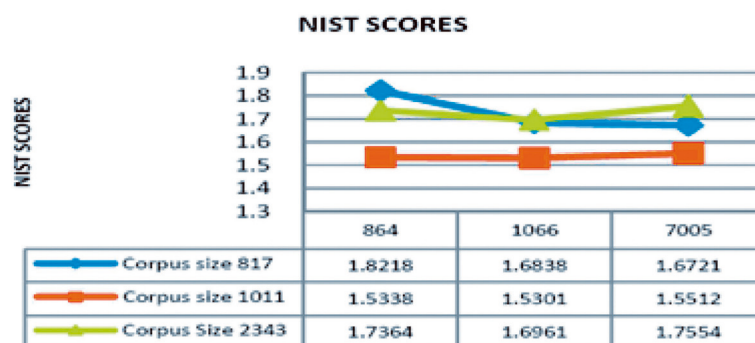
The accuracy should further increase if we categorize the corpus into different domains. At the moment our corpus includes different genres like news, short stories and poetry. Since Web 2.0 technology was used to develop corpora, translators now submit queries online in a wiki environment tied to the project's corpora and databases. There are several attempts to web-based tool for translators. For instance, Caitra is implemented in Ruby et al. (2008) on Rails as a web-based client-server architecture, using Ajax-style Web 2.0 technologies (Raymond, 2007) connected to a MySQL database-driven back-end (Koehn, 2009).

Timing the translation is the other major consideration. For web 1.0, it is conventional to perform all, or most, translation before launch and deliver many languages at once. As the number of languages increased this issues becomes less and less practical. Additionally, many Web 2.0 sites operate on user – supplied data, which by its nature won't be available until after launch.

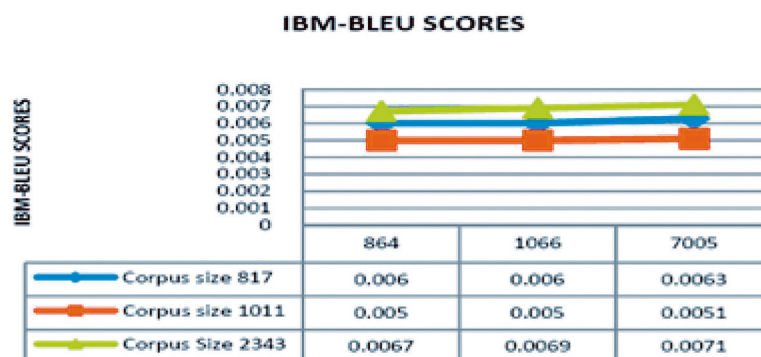
Incorporating linguistic inputs like part-of-speech tagging, parsing, morphological analysis, semantic model and a dictionary specific to the domain would make such a system more robust in terms of accuracy and is going to be explored in this project in the future. More research needs to be done in the 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.



(a)



(b)



(c)

Figure 4 : (a) Corpus Size vs. BLEU scores (b) Corpus Size vs. NIST scores (c) Corpus Size vs. IBM-BLEU scores (EN-PE)

5. Conclusion

This paper describes a set of experiments in which statistical machine translation 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 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 language model affect the translation. We already know from research reported in the literature that a corpus and language model in the domain of interest greatly affects translation results. In this study we showed that as the size of the corpus and language model increase, the BLEU score improves provided that the corpus and language model sizes are proportional. In fact the point where the best score is achieved is the point where those sizes are closest to one another. The size of the corpus is important however the sizes of the corpus and the language model need to be proportional. The finding that the corpus and language model being proportional affects translation is new and a contribution of this study of Persian SMT.

A number of problems occur when trying to align English and Persian sentences which require more investigation.

References

- AleAhmad A., Amiri, H., Oroumchian F. and Rahgozar Hamshahri M. (2008). A standard Persian text collection. *White Paper, Database Research Group*. University of Tehran.
- Amtrup J., Laboratory C.R. and University N.M.S. (2000). *Persian-English machine translation: An overview of the Shiraz project*. Computing Research Laboratory, New Mexico State University.
- Assi S. (1997). Farsi linguistic database (FLDB). *International Journal of Lexicography*, 10, 5.
- Darrudi E., Hejazi M. and Oroumchian F. (2004). Assessment of a modern farsi corpus. In *Proceedings of the 2nd Workshop on Information Technology & its Disciplines*.
- Karimi S., Scholer F. and Turpin A. (2007). Collapsed consonant and vowel models. In *New approaches for English-Persian transliteration and back-transliteration*, pp. 648-655.
- Koehn P. (2009). A Web-Based Interactive Computer Aided Translation Tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, 17-20.
- Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C. and Zens R. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL*, 2.
- Li Z., Callison-Burch C., Khudanpur S. and Thornton W. (2009). Decoding in Joshua. *The Prague Bulletin of Mathematical Linguistics*, 91: 47-56.
- Lopez A. (2008). Statistical machine translation. *ACM Computing Surveys*, 40, 3, Article 8.
- Ma Y. and Way A. (2009). Bilingually Motivated Domain-Adapted Word Segmentation for Statistical Machine Translation. *Association for Computational Linguistics*: 549-557.
- Moore R. (2002) Fast and accurate sentence alignment of bilingual corpora. *Lecture notes in computer science*: 135-144.
- Och F. and Ney H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30: 417-449.
- Raymond S. (2007). *Ajax on rails*. O'Reilly Media, Inc.
- Ruby S., Thomas D. and Hansson D. (2009). *Agile Web Development with Rails*. Pragmatic Bookshelf
- Schmidt A. (2007). Statistical Machine Translation Between New Language Pairs Using Multiple Intermediaries. Thesis.
- Stolcke A. (2002). SRILM-an extensible language modeling toolkit. In ISCA.