Improved Language Modeling for English-Persian Statistical Machine Translation

Mahsa Mohaghegh
Massey University,
School of Engineering and Advanced Technology
m.mohaghegh@massey.ac.nz

Abdolhossein Sarrafzadeh
Unitec,
Department of computing
hsarrafzadeh@unitec.ac.nz

Tom Moir
Massey University,
School of Engineering and Advanced Technology
T.J.Moir@massey.ac.nz

Abstract

As interaction between speakers of different languages continues to increase, the ever-present problem of language barriers must be overcome. For the same reason, automatic language translation (Machine Translation) has become an attractive area of research and development. Statistical Machine Translation (SMT) has been used for translation between many language pairs, the results of which have shown considerable success. The focus of this research is on the English/Persian language pair. This paper investigates the development and evaluation of the performance of a statistical machine translation system by building a baseline system using subtitles from Persian films. We present an overview of previous related work in English/Persian machine translation, and examine the available corpora for this language pair. We finally show the results of the experiments of our system using an in-house corpus and compare the results we obtained when building a language model with different sized monolingual corpora. Different automatic evaluation metrics like BLEU, NIST and IBM-BLEU were used to evaluate the performance of the system on half of the corpus built. Finally, we look at future work by outlining ways of getting highly accurate translations as fast as possible.

1 Introduction

Over the 20th century, international interaction, travel and business relationships have increased enormously. With the entrance of the World Wide Web effectively connecting countries together over a giant network, this interaction reached a new peak. In the area of business and commerce, the vast majority of companies simply would not work without this global connection. However, with this vast global benefit comes a global problem: the language barrier. As the *international* connection barriers continually break down, the *language* barrier becomes a greater issue. The English language is now the world's lingua franca, and non-English speaking people are faced with the problem of communication, and limited access to resources in English.

Machine translation is the process of using computers for translation from one human language to another(Lopez, 2008). This is not a recent area of research and development. In fact, machine translation was one of the first applications of natural language processing, with research work dating back to the 1950s(Cancedda, Dymetman, Foster, & Goutte, 2009). However, due to the complexity and diversity of human language, automated translation is one of the hardest problems in computer science, and significantly successful results are uncommon.

There are a number of different approaches to machine translation. Statistical Machine Translation (SMT) however, seems to be the preferred approach of many industrial and academic research laboratories (Schmidt, 2007). The advantages of SMT compared to rule-based approaches lie in their adaptability to different domains and languages: once a functional

Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, pages 75–82,
COLING 2010, Beijing, August 2010.

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.

Research work on statistical machine translation systems began in the early 1990s. These systems, which are based on phrase-based approaches, operate using parallel corpora – huge databases of corresponding sentences in two languages, and employ statistics and probability to learn by example which translation of a word or phrase is most likely correct. The translation moves directly from source language to target language with no intermediate transfer step. In recent years, such phrase-based MT 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 a number of languages.

The focus of this paper is on statistical machine translation for the English/Persian language pair. The statistical approach has only been employed in several experimental translation attempts for this language pair, and is still largely undeveloped. This project is considered to be a challenge for several reasons. Firstly, the Persian language structure is very different in comparison to English; secondly, there has been little previous work done for this language pair; and thirdly, effective SMT systems rely on very large bilingual corpora, however these are not readily available for the English/Persian language pair.

1.1 The Persian Language

The Persian language, or Farsi as it is also known as, belongs to the Indo-European language family and is one of the more dominant languages in parts of the Middle East. It is in fact the most widely spoken language in the Iranian branch of the Indo-Iranian languages, being the official language of Iran (Persia) and also spoken in several countries including Iran, Tajikistan and Afghanistan. There also exist large groups and communities in Iraq, United Arab Emirates, People's Democratic Republic of Yemen, Bahrain, and Oman, not to mention communities in the USA.

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.

1.2 Related Work

Several MT systems have already been constructed for the English/Persian language pair.

One such system is the Shiraz project, (Amtrup, Laboratory, & University, 2000). The Shiraz MT system is an MT prototype that translates text one way from Persian to English. The project began in 1997 and the final version was delivered in 1999.

The Shiraz corpus is a 10 MB manually-constructed bilingually tagged Persian to English dictionary of about 50,000 words, developed using on-line material for testing purposes in a project at New Mexico State University. The system also comprises its own syntactic parser and morphological analyzer, and is focused on news stories material translation as its domain.

Another English/Persian system was developed by (Saedi, Motazadi, & Shamsfard, 2009). This system, called PEnTrans, is a bidirectional text translator, comprising two main modules (PEnT1, and PEnT2) which translate in opposite directions (PEnT1 from English to Persian; PEnT2 from Persian to English). PEnT1 employs a combination of both corpus based and extended dictionary approaches, and PEnT2 uses a combination of rule, knowledge and corpus based approaches. PEnTrans introduced a new WSD method with a hybrid measure which evaluates different word senses in a sentence and scores them according to their condition in the sentence, together with the placement of other words in that sentence.

ParsTranslator is a machine translation system built to translate English to Persian text. It was first released for public use in mid-1997, the latest update being PTran version in April 2004. The ParsTran input uses English text typed or from a file. The latest version is able to operate for over 1.5 million words and terminologies in English. It covers 33 fields of sciences, and is a growing translation service, with word banks being continually reviewed and updated, available at: http://www.ParsTranslator.Net/eng/index.htm. Another English to Persian MT system is the rule-based system developed by (Faili & Ghassem-Sani, 2005)This system was based on tree adjoining grammar (TAG), and later improved by implementing trained decision trees as a word sense disambiguation module.

Mohaghegh et al. (2009) presented the first such attempt to construct a parallel corpus from BBC news stories. This corpus is intended to be an open corpus in which more text may be added as they are collected. This corpus was used to construct a prototype for the first statistical machine translation system. The problems encountered, especially with the process of alignment are discussed in this research (Mohaghegh & Sarrafzadeh, 2009).

Most of these systems have largely used a rule based approach, and their BLEU scores on a standard data set have not been published. Nowadays however, most large companies employ the statistical translation approach, using exceedingly large amounts of bilingual data (aligned sentences in two languages). A good example of this is perhaps the most well-known Persian/English MT system: Google Translate recently released option for this language pair. Google's MT system is based on the statistical approach, and was made available online as a BETA version in June 2009.

The Transonics Spoken Dialogue Translator is also partially a statistically based machine translation system. The complete system itself operates using a speech to text converter, statistical language translation, and subsequent text to speech conversion. The actual translation unit operates in two modes: in-domain and out-of-domain. A classifier attempts to assign a concept to an utterance. If the object to be translated is within the translation domain, the system is capable of significantly accurate translations. Where the object is outside the translation domain, the SMT method is used. Transonics is a translation system for a specific domain (medical: doctor-to-patient interviews), and only deals with question/answer situations (Ettelaie, et al., 2005).

Another speech-to-speech English/Persian machine translation system is suggested by Xiang et al. They present an unsupervised training technique to alleviate the problem of the lack of bilingual training data by taking advantage of available source language data(Xiang, Deng, & Gao, 2008).

However, there was no large parallel text corpus available at the time of development for both of these systems. For its specific domain, the Transonics translation system relied on a dictionary approach for translation, using a speech corpus, rather than a parallel text corpus. Their Statistical Translation approach was merely used as a backup system.

2 Corpus Development for Persian

A corpus is defined as a large compilation of written text or audible speech transcript. Corpora, both monolingual and bilingual, have been used in various applications in computational linguistics and machine translation.

A parallel corpus is effectively two corpora in two different languages comprising sentences and phrases accurately translated and aligned together phrase to phrase. When used in machine translation systems, parallel corpora must be of a very large size – billions of sentences – to be effective. It is for this reason that the Persian language poses some difficulty. There is an acute shortage of digitally stored linguistic material, and few parallel online documents, making the construction of a parallel Persian corpus is extremely difficult.

There are a few parallel Persian corpora that do exist. These vary in size, and in the domains they cover. One such corpus is FLDB1, which is a linguistic corpus consisting of approximately 3 million words in ASCII format. This corpus was developed and released by (Assi, 1997) at the Institute for Humanities and Cultural Studies. This corpus version was updated in 2005, in 1256 character code page, and named PLDB2. This new updated version contains more than 56 million words, and was constructed with contemporary literary books, articles, magazines, newspapers, laws and regulations, transcriptions of news, reports, and telephone speeches for lexicography purposes.

Several corpora construction efforts have been made based on online Hamshahri newspaper archives. These include Ghayoomi (2004), with 6 months of Hamshahri archives to yield a corpus of 6.5 million words, and (Darrudi, Hejazi, & Oroumchian, 2004), with 4 years' worth of archives to yield a 37 million-word corpus.

The 'Peykareh' or 'Text Corpus' is a corpus of 38 million words developed by Bijankhan et al. available at: http://ece.ut.ac.ir/dbrg/bijankhan/ and comprises newspapers, books, magazines articles, technical books, together with transcription of dialogs, monologues, and speeches for language modeling purposes.

Shiraz corpus (Amtrup, et al., 2000)is a bilingual tagged corpus of about 3000 aligned Persian/English sentences also collected from the Hamshahri newspaper online archive and manually translated at New Mexico State University.

Another corpus, TEP (Tehran English-Persian corpus), available at: http://ece.ut.ac.ir/NLP/ resources.htm , consists of 21,000 subtitle files obtained from www.opensubtitles.org. Subtitle pairs of multiple versions of same movie were extracted, a total of about 1,200(Itamar & Itai, 2008) then aligned the files using their proposed dynamic programming method. This method operates by using the timing information contained in subtitle files so as to align the text accurately. The end product yielded a parallel corpus of approximately 150,000 sentences which has 4,100,000 tokens in Persian and 4,400,000 tokens in English.

Finally, European Language Resources Association (ELRA), available at: http://catalog.elra.info/product_info.php?products_id=1111, have constructed a corpus which consists of about 3,500,000 English and Persian words aligned at sentence level, to give approximately 100,000 sentences distributed over 50,021 entries. The corpus was originally constructed with SQL Server, but presented in access type file. The format for the files is Unicode. This corpus consists of several different domains, including art, culture, idioms, law, literature, medicine, poetry, politics, proverbs, religion, and science; it is available for sale online.

3 Statistical Machine Translation

3.1 General

Statistical machine translation (SMT) can be defined as the process of maximizing the probability of a sentence s in the source language matching a sentence t in the target language. In other words, *"given a sentence s in the source language, we seek the sentence t in the target language such that it maximizes P(t | s) which is called the conditional probability or the chance of t happening given s"* (Koehn, et al., 2007).

It is also referred to as the most likely translation. This can be more formally written as shown in equation (1).

$$arg\ max\ P(t \mid s) \qquad (1)$$

Using Bayes Rule from equation (2), we can write equation (1) for the most likely translation as shown in equation (3).

$$P\ (t \mid s) = P\ (t) * P(s \mid t) = P\ (\mathrm{s}) \qquad (2)$$

$$\mathrm{arg\ max\ P(t \mid s) = arg\ max\ P(t) * P(s \mid t)} \qquad (3)$$

Where (t) is the target sentence, and (s) is the source sentence. P (t) is the target language model and P(s | t) is the translation model. The argmax operation is the search, which is done by a so-called decoder which is a part of a statistical machine translation system.

3.2 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 (Koehn, et al., 2007) is an open-source statistical machine translation system which allows one to train translation models using GIZA++ (Och & Ney, 2004).for any given language pair for which a parallel corpus exists. This tool was used to build the baseline system discussed in this paper. MOSES uses a beam search algorithm where the translated output sentence is generated left to right in form of hypotheses. Beam-search is an efficient search algorithm which quickly finds the highest probability translation among the exponential number of choices.

The search begins with an initial state where no foreign input words are translated and no English output words have been generated. New states are created by extending the English output with a phrasal translation of that covers some of the foreign input words not yet translated.

The algorithm can be used for exhaustively searching through all possible translations when data gets very large. The search can be optimized by discarding hypotheses that cannot be part of the path to the best translation. Furthermore, by comparing states, one can define a beam of good hypotheses and prune out hypotheses that fall out of this beam (Dean & Ghemawat, 2008).

3.3 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 SMT relies heavily on the existence of large amounts of data which is commonly referred to as a parallel corpus. The first step 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).

The next step we plan to take involves the construction of a statistical prototype based on the largest available English/Persian parallel corpus extracted from the domain of movie subtitles. This domain was chosen because the maximum number of words that can be displayed as a subtitle on the screen is between 10- 12 which means both training and decoding will be a lot faster. Building a parallel corpus for any domain is generally the most time consuming process as it depends on the availability of parallel text. But the domain of subtitling makes it easier to get the source language in the form of scripts and the target language in the form of subtitles in many different languages.

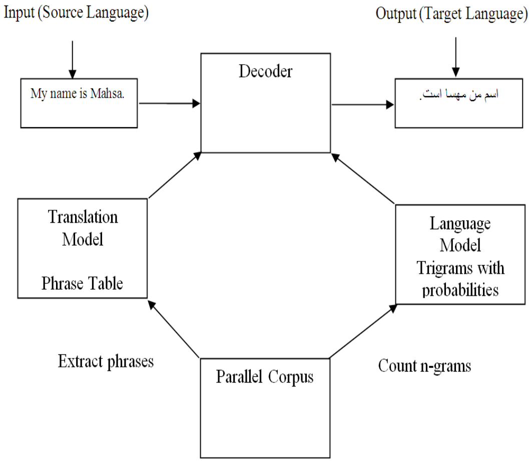


Figure1. A typical SMT System

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 getting translated into. Language modeling is not only used in machine translation but also used in many natural language processing applications such as speech recognition, part-of-speech tagging, parsing and information retrieval. A statistical language model assigns probabilities to a sequence of words and tries to capture the properties of a language.

The Language Model (LM) for this study was trained on the BBC Persian News corpus and also an in-house corpus from different genres. The SRILM toolkit developed was used to train a 5-gram LM for experimentation as in (Stolcke, 2002).

4 Experiments and Results

4.1 Experiment setup

We used Moses a phrase-based SMT development tool for constructing our machine translation system. This included n-gram language models trained with the SRI language modeling tool, GIZA++ alignment tool, Moses decoder and the script to induce phrase-based translation models from word-based ones.

4.2 Performance evaluation metrics

A lot of research has been done in the field of automatic machine translation evaluation. Human evaluations of machine translation are extensive but expensive. Human evaluations can take months to finish and involve human labor that cannot be reused which is the main idea behind the method of automatic machine translation evaluation that is quick, inexpensive, and language independent.

One of the most popular metrics is called BLEU (BiLingual Evaluation Understudy) developed at IBM. The closer a MT is to a professional human translation, the better it is. This is the central idea behind the BLEU metric. NIST is another automatic evaluation metric with the following primary differences compared to BLEU such as Text pre-processing, gentler length penalty, information-weighted N-gram counts and selective use of N-grams (Li, Callison-Burch, Khudanpur, & Thornton, 2009); (Li, Callison-Burch, Khudanpur, & Thornton, 2009).

4.3 Discussion and analysis of the results

In this study, Moses was used to establish a baseline system. This system was trained and tested on three in-house corpora, the first 817 sentences, the second 1011 sentences, and the third 2343 sentences. The data available was split into a training and test set. Microsoft's bilingual sentence aligner (Moore, 2002) was used to align the corpus and training sets. Aligning was also performed manually to aid in the improvement of the results. As the corpus size increased, we performed various experiments such as increasing the language model in each instance.

Test No.	EN/FA 1	EN/FA 2	EN/FA 3
Test Sentences	817	1011	2343
Training Sentences	864	1066	7005

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

Evaluation results from these experiments are presented in Tables 2, 3 and 4. As expected, BLEU scores improved as the size of the corpus increased. The BLEU scores themselves were significantly low; however this was expected due to the small size of the corpus. We plan to update and increase the corpus size in the near future, which will undoubtedly yield more satisfactory results.

LM=864	BLEU	NIST	IBM-BLEU
Corpus size 817	0.1061	1.8218	0.0060
Corpus size 1011	0.0882	1.5338	0.0050
Corpus size 2343	0.0806	1.7364	0.0067

Table 2. Result obtained using Language Model size=864

LM=1066	BLEU	NIST	IBM-BLEU
Corpus size 817	0.0920	1.6838	0.0060
Corpus size 1011	0.0986	1.5301	0.0050
Corpus size 2343	0.1127	1.6961	0.0069

Table 3. Result obtained using Language Model size=1066

LM= 7005	BLEU	NIST	IBM-BLEU
Corpus size 817	0.0805	1.6721	0.0063
Corpus size 1011	0.0888	1.5512	0.0051
Corpus size 2343	0.1148	1.7554	0.0071

Table 4. Result obtained using Language Model size=7005

The first test was performed on a corpus of 817 sentences in Persian and the same number for their aligned translation in English. In this instance, the training set used was 864 sentences. Results of this translation were evaluated using three evaluation metrics (BLEU, NIST, and IBM-BLEU) An excerpt from the output of this first experiment is shown in figure2 (a).

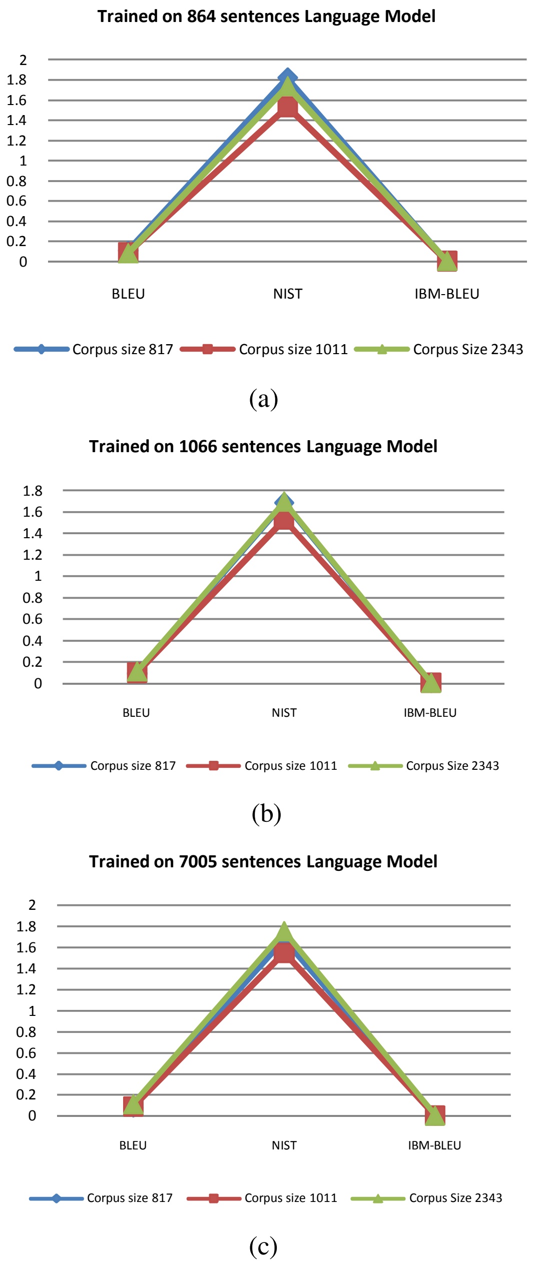
The second test comprised of a 1011 sentences corpus, with a 1066 sentence training set. As can be seen, the evaluation metric results improved.

The same experiment was repeated for a third time, this time with an even larger corpus of 2343 sentences, and a training set of 7005 sentences. The result can be seen in table 4. The results obtained in this test were close to those in the previous test, apart from a small increase in BLEU scores. It must be noted that BLEU is only a tool to compare different MT systems. So an increase in BLEU scores may not necessarily mean an increase in the accuracy of translation. 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 .

Tables 2, 3 and 4 show the results obtained using corpuses of 817, 1011, and 2343 sentences respectively. The language model size was varied from 864 to 1066 and finally to 7005 sentences.

Moreover as shown in table 3, using a corpus and language model of 1011 and 1066 in size respectively produces better results. This can clearly be noticed from graph in Figure 2(b).

Finally, increasing the size of the corpus to 2343 and language model constructed using 7005 sentences produced the best translation results as shown in both Figure 2(c) and Table 4. This data shows that an increased corpus size will yield an improved translation quality, but only as long as the size of the language model is proportional to the corpus size. 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 & Way, 2009). In the Persian language, some problems and difficulties arise due to natural language ambiguities, anaphora resolution, idioms and 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.



(a)

(b)

(c)

Figure 3. (a) Results obtained using training size=864 (b) Results obtained using training size=1066 (c) Results obtained using training size=7005

5 Future work

Despite the fact that compared to other language pairs, the available parallel corpora for the English/Persian language pair is significantly smaller, the future of statistical machine translation for this language pair looks promising. We have been able to procure several very large bilingual corpora, which we intend to combine with the open corpus we used in the original tests. With the use of a much larger bilingual corpus, we expect to produce a significantly higher evaluation metric score. Our planned immediate future work will consist of combining these corpora together, addressing the task of corpus alignment, and continuing the use of a web crawler to obtain further bilingual text.

6 Conclusion

This paper presented an overview of some of the work in the area of English/Persian MT systems that has been done to date, and showed a set of experiments in which our SMT system was applied to the Persian language using a relatively small corpus. The first part of this work was to test how well our system 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. According to the results we obtained, it was concluded that a corpus of much greater size would be required to produce satisfactory results. Our experience with the corpus of smaller size shows us that for a large corpus, there will be a significant amount of work required in aligning sentences.

References

Amtrup, J., Laboratory, C. R., & 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*(3), 5.

Cancedda, N., Dymetman, M., Foster, G., & Goutte, C. (2009). A Statistical Machine Translation Primer.

Darrudi, E., Hejazi, M., & Oroumchian, F. (2004). *Assessment of a modern farsi corpus*.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113.

Ettelaie, E., Gandhe, S., Georgiou, P., Knight, K., Marcu, D., Narayanan, S., et al. (2005). *Transonics: A practical speech-to-speech translator for English-Farsi medical dialogues*.

Faili, H., & Ghassem-Sani, G. (2005). *Using Decision Tree Approach For Ambiguity Resolution In Machine Translation*.

Itamar, E., & Itai, A. (2008). *Using Movie Subtitles for Creating a Large-Scale Bilingual Corpora*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). *Moses: Open source toolkit for statistical machine translation*.

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

Lopez, A. (2008). Statistical machine translation.

Ma, Y., & Way, A. (2009). *Bilingually Motivated Domain-Adapted Word Segmentation for Statistical Machine Translation*.

Mohaghegh, M., & Sarrafzadeh, A. (2009). *An analysis of the effect of training data variation in English-Persian statistical machine translation*.

Moore, R. (2002). Fast and accurate sentence alignment of bilingual corpora. *Lecture notes in computer science*, 135-144.

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

Saedi, C., Motazadi, Y., & Shamsfard, M. (2009). *Automatic translation between English and Persian texts*.

Schmidt, A. (2007). *Statistical Machine Translation Between New Language Pairs Using Multiple Intermediaries*.

Stolcke, A. (2002). *SRILM-an extensible language modeling toolkit*.

Xiang, B., Deng, Y., & Gao, Y. (2008). *Unsupervised training for farsi-english speech-to-speech translation*.