PersianSMT: A first attempt to English-Persian Statistical Machine Translation

Mohammad Taher Pilevar, Heshaam Faili

University of Tehran, Tehran, Iran.

Abstract

In this paper, an attempt to develop a phrase-based statistical machine translation between English and Persian languages (PersianSMT) is described. Creation of the largest English-Persian parallel corpus yet presented by the use of movie subtitles is a part of this work. Two major goals are followed here: the first one is to show the main problems observed in the output of the PersianSMT system and set a baseline for further experiments and the second one is to check whether movie subtitles can provide a good quality corpus for the development of a general purpose translator or not. In the end, translations made by the PersianSMT system equipped with different language models are evaluated on test sets of different domains and the results are compared to the Google statistical machine translator. According to the obtained BLEU scores, the proposed SMT system strongly outperforms the Google translator in translating both in-domain (movie subtitle) and out-of-domain sentences.

Keywords: Statistical Machine Translation, English-Persian bilingual corpus, Natural Language Processing

1. Introduction

Automatic translation of text written in a natural language into another one by the use of computers is referred to as Machine Translation. Automatic machine translation was one of the first natural language processing applications investigated in computer science and has attracted many researchers around the globe for over 50 years. During this period, many paradigms have been introduced and developed: Dictionary-based machine translation in which the main translation source is a bilingual dictionary which in fact defines the strength of the translation by the number of successful dictionary lookups, Example-based machine translation which makes use of a knowledge-base and a bilingual corpus to perform case-based reasoning to map the source to the target, Rule-based machine translation which makes use of a set of rules written from linguistic knowledge by human.

First ideas of Statistical machine translation was proposed by Warren Weaver in 1947, but was abandoned at the time due to computational complexity. It was later followed as a valuable research topic in computational linguistics (Brown et al., 1993). Statistical machine translation tries to learn the translation by examining the translations made by humans. Since its revival in 1993, the statistical approach to machine translation has seen an increasing interest among natural language processing research communities. Many researchers have reported training statistical systems to produce translations in weeks or days.

In 1999, a summer workshop at the Center for Language and Speech Processing at Johns-Hopkins University hosted the creation of the EGYPT toolkit, on which the widely used training tool Giza++ (Och and Ney, 2003) is based. Giza++ is an extension of the program Giza which was developed in the same workshop. Giza++ includes IBM Models 1-5 training and HMM word alignment methods. It is actually a tool for aligning words and sequences of words in a sentence aligned corpus.

In 2004, the Pharaoh phrase-based decoder (Koehn, 2004) became available and distributed in binary form. More recently, another workshop [1] released an open source toolkit, which includes a decoder named Moses (Koehn et al., 2007) and a set of software and scripts to build a complete SMT system. Moses is a statistical machine translation system which makes use of an efficient beam-search algorithm to perform the decoding in the statistical translation process. It is a phrase-based decoder which takes short text chunks as translation units. It is also a factored decoder for being able to take factored representations of a word such as part-of-speech, morphology and surface forms.

2. Persian Language

Persian (locally called Farsi) is an Indo-Iranian branch of the Indo-European languages which uses a modified Arabic script and is spoken in Iran, Afghanistan, Tajikistan, by minorities in some of the countries in the south of the Persian Gulf, and some other countries. In total, it is spoken by approximately 134 million people around the world as first or second language (Languages of the World, 2005). It is written from right to left with some letters joined as in Arabic.

Persian is a highly inflective language in which a great number of different word-forms are created by the attachment of affixes. Words are very heterogeneous in Persian and they (especially adverbs) are allowed to freely move along the sentence. Persian is a null-subject, or pro-drop language, so personal pronouns (e.g. *I*, *he*, *she*) are optional.

In Persian, there can be up to 4 written forms for a character depending on its position in a word. Many words in Persian are written in multiple forms by elimination or addition of spaces within words or by using various forms of characters. Persian texts are stored using different scripts in the computers. The mentioned issues cause difficulty in the processing of Persian text by computers, especially during the tokenization task which is one of the early steps of text preprocessing.

Persian has rich inflectional morphology: Nouns can be either singular or plural. They may also be accompanied by the *ezafe* (a suffix that connects the elements in a phrase), the indefinite marker, the enclitic particle (a suffix that links words to the following relative clause), the possessive clitic pronoun and the copula (Megerdoomian, 2000)

In Persian, verbs inflect for person, number and mood. A verb lemma can have 450 different word-forms in a Persian text, however some of these word-forms are rarely used in the nowadays context. This means that a great amount of parallel data is needed in the training phase. In Persian construction of present stems from infinitives are mostly irregular. Many compound verbs can be derived by combining nouns with light verbs and they are usually separated by some other words like adjectives and have long distance dependencies.

The sparseness problem caused by the morphological richness of Persian can be partly solved by splitting the word-forms into lemmas and grammatical categories. Moses provides some features for this task and that is why Moses was chosen as we hope that in our future experiments we can work on this problem.

[1] http://www.clsp.jhu.edu/ws2006/.

The personal pronouns may be omitted in Persian sentences while they should be present in English (except for imperatives). English uses auxiliary verbs for negation and interrogation, while some morphotactics are used in Persian for negating or interrogating. Unlike English there is no female/male distinction for Persian pronouns while the gender must be considered for applying “she” and “he” in English.

3. English-Persian machine translation history

There have been few English-Persian Machine Translation systems developed, most of which are purely rule-based.

Shiraz project (Amtrup et al., 2000) is a prototype system that translates Persian text into English. This system uses typed feature structures and an underlying unification-based formalism to describe Persian linguistic phenomena. The Shiraz system uses a hand-crafted bilingual Persian to English dictionary consisting of approximately 50.000 terms, a complete morphological analyzer and a syntactic parser. It is based on two main architectural foundations: The use of a chart throughout the system, which allows an integrated view on results created on all levels of linguistic description, and the use of a complex typed feature structure formalism, which unifies the view on the descriptions itself. The Shiraz machine translation system is mainly targeted at translating news material.

Faili et al. (2004; 2005) and Faili (2009) propose a rule-based English to Persian machine translation system based on a rich formalism named tree-adjoining grammar (TAG). Later, they introduce an enhancement of the system with trained decision trees as a word-sense disambiguation module and also get the benefit of a statistical parser to generate intermediate syntactical structure during transfer phase.

Saedi et al. (2009) presented an automatic bidirectional English/Persian text translator called PEnTrans. Their system combines rule-based method with semantic approaches to improve the result. They compared PEnT1 (English to Persian side of PEnTrans) to some of the commercially available translators in WSD task and reported superiority of their system.

Google’s research group has recently developed a web-based English-Persian statistical machine translation which has been made available since June 2009 [2]. Except the Google translate, the other three mentioned systems have not presented any BLEU scores on a standard data set and are also not available for test. Therefore we are unable to compare the results of PersianSMT to their results.

4. Architecture of the system

The goal of statistical machine translation (SMT) is to produce a target sentence *f* from a source sentence *e* that maximizes the posterior probability *p(f |e)*:

$$f^* = \arg\max_f p(f \mid e) \tag{1}$$

This probability can be represented as a product of the language model probability *p(f)* and the translation model probability *p(e|f)*:

$$p(f \mid e) = p(e \mid f)\, p(f) \tag{2}$$

[2] translate.google.com.

Overall architecture of the PersianSMT is shown in Fig. 1. The main resources for constructing a statistical machine translation are monolingual and parallel corpora which are used to train language and translation models respectively.

The translation process involves segmenting the source sentence into source phrases e_p; translating each source phrase into a target phrase f_p, and optionally reordering the target phrases to produce the target sentence f.

It is today common practice to use phrases as translation units (Koehn et al., 2003; Och and Ney, 2003) and a log-linear framework in order to introduce several models explaining the translation process:

$$e^* = \arg\max p(e \mid f) = \arg\max_{e} \{\exp(\sum_i \lambda_i h_i(e,f))\} \tag{3}$$

Where the feature functions h_i are the system models and the λ_i weights are typically optimized to maximize a scoring function on a development set (Och and Ney, 2002).

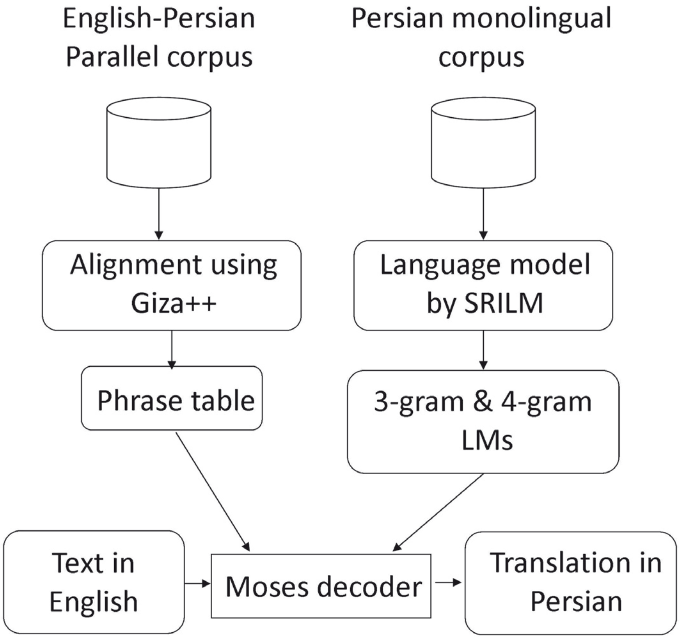

Figure 1: Architecture of the system

In our system fourteen feature functions were used, namely phrase and lexical translation probabilities in both directions, seven features for the lexicalized distortion model, a word and a phrase penalty and a target language model (LM).

The system is based on the Moses SMT toolkit (Koehn et al., 2007) and constructed as follows. The 3-gram and 4-gram language models are built using SRILM (Stolcke, 2002) toolkit from the two mentioned monolingual corpora which are of different domains. Giza++ is used to perform word alignments over the English-Persian parallel corpus. Phrases and lexical reordering are extracted using the default settings of the Moses SMT toolkit. Lexical reorderings which have substantially improved the performance of phrase-based systems, condition reordering probabilities on the words of each phrase pair. Moses is run and a 1000-best list is generated for each sentence. Using the cmert tool provided in the same toolkit, parameters of Moses are tuned on a small English-Persian parallel corpus.

4.1. English-Persian Parallel Corpus

Statistical machine translation is trained by using sentence-level aligned parallel corpora. The improvement of SMT is highly constrained by the lack of large parallel corpora which is the case for Persian language. Parallel corpora needs to be not only large in size but also balanced.

The job of the translation model, given a Persian sentence *f* and an English sentence *e*, is to estimate the probability that *f* generates *e* (*i.e. p(e|f)*). All statistical translation models are based on the idea of word alignment and are trained using a large parallel corpus. Obtaining this parallel corpus is one of the most important steps in the development of SMT, especially for low resource languages such as Persian. Publicly available bilingual data for English-Persian language pair are rarely available.

The intuitive idea would be to make parallel corpora out of literary translations, but this is less common for MT purposes as the translations are not usually literally and may involve many content omissions and therefore not suitable for word alignment. Translated books are not only unsuitable for the task but also protected by copyright. Literal translations such as Hansards are commonly used in the MT community as a resource to generate parallel corpora. For European languages, the Europarl corpus has become quite a standard one. Unfortunately, there is no similar resource available for Persian language.

To acquire a fairly large parallel corpus, we chose to mine movie subtitles which until recently has not been utilized by NLP tasks. There are various advantages in using movie subtitles such as (Itamar and Itai, 2008):

1. They grow daily in amount.
2. They are publicly available and can be downloaded freely from a variety of subtitle web sites.
3. The subtitle files contain timing information which can be exploited to significantly improve the quality of the alignment.
4. Translated subtitles are very similar to those in the original language – contrary to many other textual resources; the translator must adhere to the transcript and cannot skip, rewrite, or reorder paragraphs.

There are however disadvantages to using movie subtitles as a bilingual resource:

1. Movie subtitles contain daily conversations which are informal and therefore bias the system toward creating translations in spoken language.
2. After investigating the translated sentences in PersianSMT output, we observed that the length ratio of Persian to English sentences is about 0.7 (which is not the case while a human translator translates from English to Persian). This leads to a less understandable translation and a reduction in evaluation scores.
3. Punctuations are not usually included in movie subtitles, and therefore sentence limits are not available. Alignments are made between individual lines in subtitle files according to timing information. However these individual lines are sometimes neither complete sentences nor complete phrases.
4. In Persian, words are spoken in many ways, and therefore written in many different forms in a movie subtitle. Unifying these forms to avoid the scarcity is to be done manually and needs great effort.

Some of these problems can be solved by applying rule-based correction methods. Building aligned bilingual corpora from movie subtitles were first presented by Mangeot and Giguet (2005). They proposed a semi-automatic method which needs human operator to synchronize some of the subtitles.

Tiedemann created a multilingual parallel corpus of movie subtitles using roughly 23.000 pairs of aligned subtitles covering about 2.700 movies in 29 languages (Tiedemann, 2007).

He proposed an alignment approach based on time overlaps. Itamar and Itai (2008) proposed a methodology based on the Gale and Church's sentence alignment algorithm (1993) which benefits from timing information in order to obtain more accurate results.

Around 21.000 subtitle files were obtained from www.opensubtitles.org [3]. It contained subtitles of multiple versions of the same movie or even multiple copies of the same version created by different subtitle makers. For each movie, a subtitle pair was extracted by examining the file size and timing information of subtitle files. This information was used to confirm that both of the files contain the subtitles of the same version of the movie. Duplicates were then removed to make the resource unique and avoid redundancy. It resulted in about 1.200 subtitle pairs.

Each pair consists of two textual files (in srt format), containing subtitles of the same movie in both Persian and English languages. These files were aligned using a dynamic programming method proposed by Itamar and Itai (2008). Their method utilizes timing information in subtitle files, to obtain a good quality alignment. A parallel corpus of 150.000 sentences was obtained which has 4.100.000 tokens in Persian side and 4.400.000 tokens in English.

In addition to the subtitles, we have used some other resources to make the corpus larger and more domain-independent. We have made use of three English textbooks [4] in computer science whose translations were available in Persian. These translations were done by amateur translators in a literal and sentence by sentence manner which makes them suitable for alignment task. We have aligned these texts using a method based on the Gale and Church method (1993).

Another resource used for the creation of translation model was the documentation of KDE4 which is made available by the OPUS corpus (Tiedemann, 2004). OPUS is an attempt to collect translated texts from the web, to convert and align the entire collection, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and is also delivered as an open source package. It provides several gigabytes of multilingual corpora containing millions of tokens in many languages. For instance, the most recent release has been the EMEA corpus which includes 320 million tokens in 22 languages. Unfortunately, the only corpus which covers Persian language as a target language, is the KDE4 documentation which provided us with a parallel corpus of around 500.000 tokens in each side.

4.2. Translation Model

In order to estimate the translation model, an English-Persian parallel corpus was built using the three mentioned resources. Tab. 1 summarizes the size of used resources and their domain. The corpus size column in Tab. 1 suggests that the corpus extracted from movie subtitles constitutes a large portion of the overall parallel corpus used for the training of translation model in PersianSMT. This reduces the ability of the system as a general purpose translator and makes it more of a movie subtitle translator.

Available resources were combined, resulting in a 5-million-token parallel corpus. Giza++ was used to align these data. A phrase table containing more than 5 million phrases was resulted. The standard phrase table created by Giza++ includes computations of five phrase translation scores: phrase translation probability (in both directions), lexical weighting (in both directions)

[3] Thanks to Branislav Gerzo, the administrator of the site, for his support.

[4] Artificial Intelligence: Modern Approach (Stuart J. Russel), Build A Remote-Controlled Robot (David R. Shircliff) and Speech and Language Processing (Jurafsky and Martin).

and phrase penalty. This phrase table will be used by the Moses to generate the most probable translation for a source sentence during the decoding process.

	Corpus size (in tokens)	*Domain*
Movie subtitles	4.300.000	Movie
Textbooks	500.000	Computer Science
KDE4 documentation	150.000	Computer Science
Total	Around 5.000.000	

Table 1: Bilingual corpora used to train the translation model

4.3. Language Model

Unlike translation model, estimation of language model requires monolingual corpora which are much easier to obtain in large quantities. The following resources are available online which can be used to create the language model:

- Hamshahri is one of the most popular daily newspapers in Iran that has been publishing for more than 20 years. Hamshahri2 corpus (Darrudi et al., 2004) is a Persian test collection that consists of 1.4GB of news texts from this newspaper since 1996 to 2008. This corpus which is basically designed for the classification task, contains more than 318.000 news articles about variety of subjects.
- A 10-million-token monolingual corpus that was extracted from Persian movie subtitles.

Hamshahri2 text collection was used as monolingual corpus to create the language model. A part of this corpus containing news articles from the year 2005 to 2008 was chosen. This part whose size is 900 MB, consists of about 100 million tokens. In addition to that, the movie subtitle corpus whose domain is completely different to Hamshahri2's and has a closer domain to the parallel corpus used for translation model, has been used to create a language model. Tab. 2 summarizes the monolingual corpora used for creation of language model. SRILM toolkit (Stolcke, 2002) was used to create 3-gram and 4-gram language models of the mentioned resources.

	Corpus size (in tokens)	*Domain*
Hamshahri	100.000.000	News
Subtitle corpus	10.000.000	Movie

Table 2: Monolingual corpora used to train the language model

Tab. 3 shows the perplexities of 3-gram language models obtained from Hamshahri2 and subtitle corpora for the test data of different domains. The domain of EGIU test set is closer to that of Hamshahri2 corpus and the Subitle-test-set is in the same domain as subtitle language model. This is confirmed by the lower perplexity values obtained for these pairs. This concludes that, as expected the Hamshahri language model is more suitable for the translation of normal texts whereas subtitle language model suits the movie subtitle translation task.

	Language Model	
	Hamshahri	*Subtitle*
EGIU test set	**577**	2.903
Subtitle test set	21.635	**480**

Table 3: Perplexities of 3-gram language models for diff?erent domains

5. Experiments and Results

BLEU is one of the most popular metrics for automatic evaluation of machine translation quality. The translated output of a test set is compared with different manually translated references of the same set. The BLEU scores of translation by PersianSMT are summarized in Tabb. 4 and 5. Translation is performed on two test sets of different domains. The first one is a corpus of example sentences of the textbook "English Grammar in Use (EGIU)" which shows various grammatical aspects of English and the second one is taken from movie subtitles not included in the training process. Each of these test sets approximately contains 2.500 sentences of average ten words. In the tables, *Sub* stands for the corpus extracted from movie subtitles, *Ham* for the Hamshahri2 corpus and *Ham+sub* for the combination of both. Since there exists no big difference between *Ham+Sub* and *Ham* results, detailed results for *Ham+Sub* is not included in Tab. 4.

p_n is the n-gram precision and BP is the brevity penalty used when computing BLEU. N-gram precision in BLEU is computed as follows (Chin-Yew Lin and Hovy, 2003):

$$p_n = \frac{\sum_{C\in\{Candidates\}} \sum_{ngram\in C} Count_{clip}(ngram)}{\sum_{C\in\{Candidates\}} \sum_{ngram\in C} Count(ngram)} \tag{4}$$

Where *Count*(*ngram*) is the number of *n-gram*s in the candidate translation and $Count_{clip}(ngram)$ is the maximum number of *ngram*s co-occurring in a candidate translation and a reference translation. To prevent very short translations that try to maximize their precision scores, BLEU adds a brevity penalty (*BP*) to the formula:

$$BP = \begin{Bmatrix} 1 & if\,|c|>|r| \\ e^{(1-|r|/|c|)} & if\,\,|c|\le|r| \end{Bmatrix} \tag{5}$$

Where $|c|$ is the length of the candidate translation and $|r|$ is the length of the reference translation. The BLEU formula is then written as follows:

$$BLEU = BP \bullet \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{6}$$

In the output it was observed that there are several words which have not been translated (OOV). To overcome this problem, a dictionary database was obtained and word pairs were added to the bilingual corpus.

According to the obtained BLEU scores, the best configuration to translate the grammatical text of EGIU is the PersianSMT equipped with 4-gram Language Model taken from Hamshahri2 corpus. However, for translating movie subtitles, Language Model taken from movie subtitles improves system output by a great margin of 12 BLEU scores compared to the system with Hamshahri2 LM. These numbers also conclude that, with presence of large monolingual data, 4-gram LM provides a better estimation of the Persian language, but in the case of movie subtitles where fewer amounts of data is available, 3-gram outperforms the 4-gram LM. However subtitle corpus is about 10 times smaller than the Hamshahri2 corpus, it provides a better language model for the task of translation of subtitles.

Results of PersianSMT are compared to the baseline system and Google translate. As a baseline, we have chosen a very simple system which translates an English text by just replacing individual words with their first meaning in the dictionary. This is the architecture used by many English to Persian dictionary translators till recently. Google translator is tested on the same test data and results are included in Tab. 6. It is obvious that PersianSMT outperforms the Google translation by a great margin in translation of movie subtitles. In the task of translation of EGIU data, our system is again providing better results.

Having been trained on a parallel corpus mostly obtained from movie subtitles, PersianSMT is too biased towards the domain of movie conversations. This is shown in Tabb. 4 and 5 where BLEU scores obtained on the movie subtitle test corpus is more than twice of the EGIU test corpus.

	Language Model (3-gram)												
Test data	*Sub*						*Ham*						*Sub + Ham*
	BLEU	p_1	p_2	p_3	p_4	*BP*	*BLEU*	p_1	p_2	p_3	p_4	*BP*	*BLEU*
EGIU	7.76	52.5	15.8	5.4	1.9	0.812	11.22	61.5	23.8	9.8	3.9	0.73	11.22
Subtitle	24.20	65.3	47.5	37.5	30.0	0.56	11.69	50.8	28.4	18.6	13.6	0.475	13.56

Table 4: PersianSMT with different 3-gram LMs p_n is the n-gram precision and BP is the brevity penalty used when computing BLEU

Test data	*BLEU*	p_1	p_2	p_3	p_4	*BP*
EGIU	7.46	52.5	15.8	5.4	1.9	0.806
Subtitle	25.46	65.8	48.9	39.5	32.6	0.565

Table 5: PersianSMT with 4-gram Sub-LM $_n$ is the n-gram precision and BP is the brevity penalty used when computing BLEU

MT system	*Subtitle test corpus*						*EGIU test corpus*					
	BLEU	p_1	p_2	p_3	p_4	*BP*	*BLEU*	p_1	p_2	p_3	p_4	*BP*
PersianSMT	25.46	65.8	48.9	39.5	32.6	0.565	11.22	61.5	23.8	9.8	3.9	0.730
Google translator	0.96	15.2	2.4	0.5	0.1	0.900	5.35	37.5	10.1	2.8	0.8	1.00
Baseline		All less than 0.5				–		All less than 0.5				–

Table 6: Comparison of statistical English-Persian MT outputs

6. Discussion

The main problem observed in the translation is the failure of the system to translate the compound verbs. For example, the English verb “*watches*” is translated to a two-word verb *negah mikonad* in Persian. In translation from English to Persian, this is the case for most of the verbs. It is observed that in several places in the output of PersianSMT, a verb like *watches* is translated to *negah* only and the *mikonad* part is omitted. This is partly due to the deficiency in the Persian tokenization system used for preprocessing of the Persian corpus. We had expected the phrase-based system to overcome this problem but having a small training data, problems due to scarcity seem inevitable. This is due to high inflectional morphology phenomena of the Persian language. A verb like “*watch*” can take near to 450 forms depending on its subject, tense and mood in the sentence. Large number of inflected verb forms does not let the system learn to translate all the individual forms of a compound verb. It adds to the complexity to know that being a pro-drop language, Persian takes personal pronouns as an optional element in the sentence.

Another usual problem occurred in the translation is the failure of the system to place the elements of the sentence in the right order. We had hoped that using a phrase-based SMT system helps us to overcome the differences of word order between English and Persian languages, but apparently additional techniques are needed for that to be fulfilled. At the moment reordering of the target sentence is only taken care by distortion model of the SMT algorithm. This is not enough as many reordering are required to make the target sentences rightly ordered. One way to solve this problem is to re-rank the n-best output list and/or reorder the output sentences. Another solution to this problem is described in (Nießen and Ney, 2001). According to this method, prior to translation, the input sentence is reordered using morpho-syntactic information, so that the word order resembles better that of the target language.

Another examined disadvantage of the SMT output is that sentence lengths get shortened by a ratio of about 0.7 after translation. The ratio of sentence lengths in English to Persian is about 0.95 in the parallel corpus. Tab. 7 summarizes the average sentence ratio of English to translated Persian sentences. More shortening is observed while translating movie subtitles (about 0.6) compared to the 0.8 of the EGIU test data. This is due to the conversational nature of these test data.

The system has however performed a fairly good job in detecting the right word sense while translation. Another advantage of the system is the correct detection and translation of idioms which is very important in translation of movie subtitles.

	Language model used			
	3-gram LM			*4-gram LM*
Test data	*Sub*	*Ham*	*Sub + Ham*	*Sub*
EGIU	0.827	0.761	0.761	0.823
Subtitle	0.633	0.573	0.590	0.636

Table 7: Length ratio of English to Persian sentences (translated by PersianSMT)

7. Conclusions

This paper described a set of experiments, in which statistical machine translation was applied to the Persian language. The first objective of this work was to test, how well SMT translates from English into Persian, when trained on a small bilingual corpus.

There is no doubt that the parallel corpus used in our experiments is very small compared to the parallel corpora used nowadays for training SMTs for other languages such as Arabic and Chinese, but as far as we know, this is the largest English-Persian corpus yet presented. We hope to obtain better results by gathering more bilingual data and estimating a better language model of Persian language. Other methods such as local reordering using morpho-syntactic information can be used to improve the translation results.

References

Amtrup J.W., Rad H.M., Megerdoomian K. and Zajac R. (2000). Persian-English Machine Translation: An Overview of the Shiraz Project. In *Memoranda in Computer and Cognitive Science MCCS-00-319*, Colnputing Re-search Laboratory, New Mexico State University.

Brown P.F., Pietra S.A.D., Pietra V.J.D. and Mercer R.L. (1993). The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2).

Darrudi E., Hejazi M.R. and Oroumchian F. (2004). Assessment of a Modern Farsi Corpus. In *Proceedings of the 2nd Workshop on Information Technology & its Disciplines (WITID)*, ITRC, Kish Island, Iran.

Ethnologue (2005). *Languages of the World*, 15th ed.

Faili H. (2009). From Partial Toward Full Parsing. In *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, 14-16 Sep., Borovets, Bulgaria, pp. 71-77.

Faili H. and Ghassem-Sani G. (2004). An Application of Lexicalized Grammars in English-Persian Translation. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*, 24-27 Aug., Universidad Politecnica de Valencia, Valencia, Spain, pp. 596-600.

Faili H. and Ghassem-Sani G. (2005). Using a Decision Tree Approach for Ambiguity Resolution in Machine Translation. In *Proceedings of 10th Annual Int. CSI Computer Conference (CSICC'2005)*, 15-17 Feb., Tehran, Iran, Vol. II, pp. 252-256.

Gale W.A. and Church K. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19 (1): 177-184.

Itamar E. and Itai A. (2008). Using Movie Subtitles for Creating a Large-Scale Bilingual Corpora. In *6th Int. Conf. on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Koehn P. (2004). Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *AMTA 2004*, Washington, USA.

Koehn P., Hoang H., Birch A. and Callisono-Burch, C. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL, demonstration session*, Prague, Czech Republic, June.

Lin Chin-Yew and Hovy E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 71-78.

Megerdoomian K. (2000). Persian Computational Morphology: A Unification-Based Approach. NMSU, CRL, Memoranda in Computer and Cognitive Science Report (MCCS-00-320). http://citeseer.ist.psu.edu/megerdoomian00persian.html.

Mangeot M. and Giguet E. (2005). Multilingual aligned corpora from movie subtitles. Technical report, Condillac-LISTIC.

Nießen S. and Ney H. (2001). Morpho-syntactic analysis for reordering in statistical machine translation. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Galicia, Spain, pp. 1081-1085.

Och F.J. and Ney H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29 (1): 1951.

Saedi C., Shamsfard M., Motazedi Y. and Branch R. (2009). Automatic Translation between English and Persian Texts. *Journal Of The International Linguistic Association*, 23 : 299-330.

Stolcke A. (2002). Srilm – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, USA, vol. 2, pp. 901-904.

Tiedemann J. and Nygaard L. (2004). The OPUS corpus - parallel & free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 26-28.

Tiedemann J. (2007). Improved Sentence Alignment for Movie Subtitles". In *Int. Conf. on Recent Advances in Natural Language Processing (RANLP 2007)*, pp. 582-588, Borovets, Bulgaria.