Developing an Open-domain English-Farsi Translation System Using AFEC: Amirkabir Bilingual Farsi-English Corpus

FattanehJabbari, SomayehBakhshaei, Seyed Mohammad Mohammadzadeh Ziabary, ShahramKhadivi
Human Language Technology Lab
Department of Computer Engineering and Information Technology
Amirkabir University of Technology
Tehran, Iran
fjabbari@ce.sharif.edu, {bakhshaei, mehran.m, khadivi}@aut.ac.ir

Abstract

The translation quality of Statistical Machine Translation (SMT) depends on the amount of input data especially for morphologically rich languages. Farsi (Persian) language is such a language which has few NLP resources. It also suffers from the non-standard written characters which causes a large variety in the written form of each character. Moreover, the structural difference between Farsi and English results in long range reorderings which cannot be modeled by common SMT reordering models. Here, we try to improve the existing English-Farsi SMT system focusing on these challenges first by expanding our bilingual limited-domain corpus to an open-domain one. Then, to alleviate the character variations, a new text normalization algorithm is offered. Finally, some hand-crafted rules are applied to reduce the structural differences. Using the new corpus, the experimental results showed 8.82% BLEU improvement by applying new normalization method and 9.1% BLEU when rules are used.

1 Introduction

Statistical Machine Translation (SMT), the most promising MT approaches, is producing acceptable translation for some languages, but not for all language pairs because of some challenges. For example, since it requires a big amount of training data, the translation quality is low for those languages with scarce resources. The problem is also more critical for morphologically rich languages. Farsi is an instance of such languages which has insufficient size of existing parallel corpora, in addition to its rich morphology. Although its morphology is not as rich as Arabic but is richer than most of the languages like English [1]. So, preparing a SMT system for the English-Farsi language pair, results in weak translation quality using small training data, as the previous researches on English-Farsi SMT systems. Considering the related problems in English-Farsi translation, we try to develop a more qualified system. To this end, we first generated a large open-domain parallel corpus, Amirkabir Bilingual Farsi-English Corpus (AFEC). The produced corpus can be considered as the best bilingual parallel English-Farsi corpus according to its size, quality, and domain generality in the news issues.

Furthermore, another difficulty rises when translating from/to Farsi texts, which is the existence of different written forms for each character in Farsi. To remove this character abusing, we offered a new algorithm for text pre and post processing called Essential for Statistical Machine Translation (E4SMT) which uses a high speed character-based algorithm for simultaneous normalization, tokenization and detection of special tokens (e.g. Numbers, Dates, Abbreviations, etc) by reviewing whole text in a single pass.

Finally, we try to handle another complication of English-Farsi language pair which is the effect of differences in the grammatical structures of English-Farsi language pair. For example, the part of speech order in a Farsi sentence is: Subject Object Verb (SOV), but it is SVO in English. This variation causes to long displacements which are hard to detect by many of the reordering models (since most of them consider the local short distortions). To moderate the differences in words order, we applied some hand-crafted rules which change the order of words in the source language to match the structure of the target side. For this task, we have extracted some manual rules making use of part of speech tags.

The previous considerable researches on English-Farsi languages are [2], [3], and [4] which are the first attempts for making a SMT system for English-Farsi language pair. These researches are developing Automatic Speech Recognition (ASR) systems and try to run speech to speech translation systems, with an essential SMT component as an inner core. They have used either a small corpus, or a limited-domain one. For instance, [2] is a speech to speech translation system in the medical care domain. Thus, our system outperforms the previous SMT systems for English-Farsi language pair since it uses a larger open-domain corpus. Recently, some new experiments are reported like [5] which offer how to build SMT system from limited resources. They have used normalization just on the English side according to the NIST standard table of normalization rules. Compared to this work we have offered a novel dynamic normalization algorithm for both English and Farsi sides. [6] uses a 130K lined corpus with 2.8M running words. This paper has improved the reordering model with a novel idea for Farsi-to-English SMT system. [7] offered a direct search for minimizing error rate for parameter optimization in Farsi-to-English SMT system, instead of MERT algorithm [8], using the corpus size of about 739K line. The corpus we collected in this research is more noticeable than the existing corpora in its size, domain generality, and the numbers of words it covers.

The remainder of this paper is as follows. We describe our corpus generation method in the second section. Then the data normalization scheme is explained in the third section. The forth part is about manual rules. Experiments are explained in section 5. Finally, section 6 concludes the paper.

2 Corpus Gathering

There are two main approaches to create a new corpus: 1) using automatic tools for document aligning, 2) by means of human translators. In this research, both of these methods are used. First, we crawled the web and extracted as much data as possible including parallel, comparable and monolingual texts. In addition to web pages, we used other resources like translated books, software manuals, subtitle-films, multilingual constitution of some countries, etc. Among the gathered data, a small volume was completely parallel, while the rest were the comparable documents.

Bilingual Corpus		Line Number	Singleton	Running Words	Lexicon
Central Asia	English	84807	27722	1971667	61565
	Farsi	84807	18735	2152752	41191
Ted	English	66534	10921	628963	24590
	Farsi	66534	14724	668450	29382
News	English	282227	61537	6993837	135365
	Farsi	282227	75225	7494634	135284
Verb-mobil	English	23145	1039	249356	2763
	Farsi	23145	2414	216577	5283
Misc	English	141602	54319	3343737	105713
	Farsi	141602	44634	3541859	82579

Table 1. Statistics of generated corpora

The qualified comparable data was selected and document aligned with aligner tools. We have used HunAlign [9] and Microsoft aligner [10]. Since these tools are not customized for Farsi language, many parts of the automatically aligned corpora were in such a bad condition that we ignored them. Thus, the produced data was not as much as we needed. We continued the work by translating some part of the documents by the help of human translators. The statistics of each created corpus are shown in Table 1.

In the following section, we will describe much about each of these prepared corpora and the existing ones.

2.1 The automatically aligned corpora

- **CentralAsia** - The first corpus named Central Asia is extracted from Central Asia news website: http://centralasiaonline.com. This website reports news in different languages such as Farsi, English, Urdu, Pashtu, but we have used only Farsi-English parts. It has 84K lines, with about 1.9M words in the English side and 2M words in the Farsi side. Its domain is news domain.

- **Ted** - Ted corpus is the subtitles of the ted website movies: http://www.ted.com/talks. Since the different subjects are presented in this website, the corpus is open-domain. The size of corpus is about 66K lines, with 620K words in the English side and 660K words in the Farsi side.

2.2 Human translated corpora

- **News** - This corpus is the monolingual documents downloaded from news websites such as CNN, BBC, etc. Its volume is about 280K lines with about 6.9M words in the English side and 7.4M words in the Farsi side.

- **Misc** - Misc corpus is a bunch of miscellaneous documents translated by human translators. It has general domain with size of 140K lines and 3.3M words in the English side and 3.5M words in the Farsi side.

- **Verbmobil** - Is a part of English side of Verbmobil project corpus [11] which includes some tourists' conversations about time scheduling and appointment settings and is translated by human translators. This dataset includes 23K lines in both sides, 249K and 216K words in Farsi and English sides respectively.

2.3 The existing corpora

We used some existing corpora in addition to the corpora that we made, which are:

- **Pen** - An existing corpus with about 30K lines. Its domain is news [12].

- **Elra** - An existing corpus with 50K lines which has the news domain [13].

- Another Farsi-English existing corpus is Tehran University Corpus [14]. This corpus is extracted from subtitle films. Its domain is general and sentences are transcriptions of spontaneous speech. The size of this corpus is 612K. The corpus is noisy, so we did not use it in or works.

- 20K transliterated names for further improvement was produced and added it to our integrated corpus [15].

2.4 The AFEC corpus

By integrating all generated and existing corpora, we produced our large corpus. The information of this new corpus is mentioned in Table 2. The lines number of this corpus is about 700M. This corpus covers 14.7G words of English sides and about 15.8G of Persian side.

Bilingual Corpus		Line Number	Singleton	Running Words	Lexicon
AFEC	English	700916	139041	14764413	267717
AFEC	Farsi	700916	133413	15807981	238571

Table 2. Statistics of AFEC corpus

3 DATA NORMALIZATION

Farsi has an important challenge in its written form. This dilemma originates from existence of different ASCII codes for each Farsi written character since there is not a standard format for Farsi written text. Moreover, some characters are misplaced by their Arabic format, because of their similar appearance, for example using "ئ" or "ي" instead of "ی" . We propose a text pre and post processing tool incorporated with an interactive text normalizer to remove this complication we called this tool E4SMT (Essential for Statistical Machine Translation).

The proposed tool is incorporated with a bunch of plugins where each one monitors the occurrence of a specific token. These specific tokens are something like numbers, dates, abbreviations, etc which must be treated different from other parts of the context or maybe does not need to be translated. Also, a built-in character normalizer module normalizes different character representations to be uniform. The innovative characteristic of the algorithm is the ability of processing, normalization and tagging the whole text in a single pass. By visiting each character, along with normalizing it, all of the plugin modules will process it and cache in case it is a valid character in the sequence. Whenever a plugin module detects new valid token, it will report it to be tagged. Plugins are controlled by a plugin manager and could be deactivated and/or prioritized to change tool behavior in case of similar tokens detection by different plugins.

E4SMT has been developed using C++ in a cross platform scheme thanks to Nokia Qt framework [16] and can be used as a standalone application, as a web service, and also can be integrated to other tools using its API. This tool has many features which are not used in the pre and post-processing parts but used in corpora generation and maintenance. Built-in modules and plugins are incorporated with external configuration files and tables which eases the use, maintenance and enhancement of the tool. Currently, the following built-in features and plugins are developed and activated:

- Character normalizer: This is a built-in feature which works in two interactive and non-interactive modes to convert each Unicode character to a uniform representation
- Built-in tokenizer and tagger: These will tokenize input text and tag specific tokens using plugins. Inline XML (IXML) is used for tagging. IXML tags will be removed in post-processing pass.
- URL plugin: This recognizes URL addresses in the text and tag them
- Email plugin: Similar to the URL plugin, this one recognizes e-mail patterns.
- Suffix plugin: Check for suffixes such as apostrophes by using the suffix tables and some manual rules to exclude them from tokenization process.
- Number plugin: This part recognizes and tags different number types in the text including general numbers, currencies, weights, etc.
- Abbreviation plugin: Recognizes and tags abbreviation words in the text using a dictionary and also some predefined rules. Abbreviations will be converted to their equivalent in post-processing of translated text.
- Transliteration plugin: This plugin will transliterate Name Entities recognized (NER) in input text.
- Virastyar Plugin: This one is a special plugin used for post-processing and correction of punctuations and dictation problems in the translated text.

One of the most important features of the E4SMT tool which caused high improvement in translation results is the normalization feature. At first, we had used a static mapping table to normalize characters both in Persian and English texts. But we found that there are many other unrecognized or multiform characters in texts (especially Farsi texts) downloaded from news agencies which need to be normalized. So, we developed an interactive normalizer which will ask for user decision on any new seen character. Valid decisions are:

- Keep it: the input character must be moved to output without any change
- Remove it: null will be passed as output
- Change it: another character will be replaced.

User decisions will be stored in normalization table and used next time the character is seen both in interactive and non-interactive use of the tool. Now, our normalization table has more than 600 entries covering whole AFEC corpora.

4 Grammatical Rules for English-Farsi Language Pair

As stated earlier, English and Farsi languages have different grammatical structures which results in low quality of translation. Some major challenges of this type, which also affect the translation quality, are discussed in this research. For example, Farsi usually follows SOV pattern in sentences, but this is SVO in English. Also, there may be multiple verbs in a Farsi sentence like English, but there is no clue to find out which verb belongs to which subject and object except the meaning of the sentence. “Ezafe” structure is another feature of Farsi language which makes it challenging in NLP tasks. Ezafe structure is composed of two or more related words within a phrase which are connected together by Ezafe vowel /e/ or /ye/. Ezafe structure includes:

- A noun before another noun,
- A noun before a possessor,
- A noun before adjectives,
- An adjective before another adjective,
- And combinations of above.

The Ezafe vowel is pronounced but it is not written in Farsi text, thus it raises ambiguities for NLP tasks. One way to reduce such problems is the reordering of words in the source language to simulate the word patterns in the target language. This can be done both by rule-based and data-driven methods where in this research we focus on rule-based reorderings. Regarding to the Farsi language structure compared to English, for English-to-Farsi SMT, two types of reorderings can be applied to the source sentences: Local reorderings which seems appropriate for Ezafe structure and global reorderings which is more suitable for verb reorderings. Global reorderings of verbs puts the verbs in source sentence to the end of the sentence to follow the Farsi structure. This requires the boundaries of clauses especially when there are multiple verbs in a sentence, but there are no obvious marks to determine these points in Farsi sentences. However, an application of hand-crafted rules to reorder the verbs of Farsi sentences in Farsi-to-English SMT is done in [6] by means of conjunctions and punctuations, but using such clues did not lead to notable improvements. Here, we extract some rules for local reorderings of Ezafe structures, which is very common in Farsi, using part of speech tags. These hand-crafted rules are described as follows:

Rule 1: In Farsi, the adjectives in Ezafe structure which describe a noun follow it, whereas in English this order is opposite, i.e. the adjectives precede the noun. For example:

English	a beautiful house and a kind landlord
Reordered English:	a house beautiful and landlord kind

The following rule can be applied to remove this mismatch:

JJ [JJ || CC JJ ||, JJ]* [NN ||NNS]
→ Rule (1)
[NN ||NNS] JJ [JJ || CC JJ ||, JJ]*

where JJ, CC, NN, and NNS are part of speech tags for adjectives, conjunctions, noun, and plural nouns respectively.

Rule 2: It is also useful to apply reordering when Ezafe occurs in the case of nouns modifying other nouns. In English such relations can be expressed in two ways: 1) using the preposition “of” like “the handle of the door”. This pattern matches the Farsi. 2) The order can be changed by removing “of” such as “the door handle”. This pattern conflicts Farsi Language. This can be lessened by applying this rule:

[NN || NNS]1 [[NN || NNS]2 … [NN || NNS]n
→ Rule (2)
[NN || NNS]n … [NN || NNS]2 [NN || NNS]1

Rule 3: Another incompatibility which occurs in Ezafe structure is the placement of pronoun after possessor. For example in English we say “your book”, but in Farsi it comes in reverse order “ کتاب شما” (ketab-e-shoma).

PRO [NN || NNS] → [NN || NNS] PRO Rule (3)

where PRO stands for pronoun.

Rule 4: Finally, the order between the noun and its possessor is changed in Farsi. For instance, we say “John’s book” in English, but “کتاب جان” (ketabe-e-jaan) in Farsi.

5 Experiments and results

To achieve a reasonable SMT system for English-Farsi, we focus on the bottlenecks of the Farsi language, i.e. limited data resource, text normalization, and grammatical structure of it. To overcome these problems, we gather a large corpus. The statistics of all corpora are shown in Table 2. Then to measure the quality of each of these corpora, we did an experiment. In the following experiments all of the conditions except the training corpora are the same. These conditions includes language model, tuning set, testing set and translation parameters. Table 3 shows the statistics of the test and tuning sets with four Farsi references and Table 4 demonstrates the quality of each corpus based on BLEU measure:

Test/Tune		Line Number	Singleton	Running Words	Lexicon
Test set	English	418	1945	10981	3144
	Farsi 1	418	1642	12208	2888
	Farsi 2	418	1555	13266	2913
	Farsi 3	418	1366	13021	2673
	Farsi 4	418	1529	12738	2827
Tune set	English	400	2052	10848	3204
	Farsi 1	400	1881	11759	3095
	Farsi 2	400	1825	13235	3136
	Farsi 3	400	1558	12911	2849
	Farsi 4	400	1716	12397	3003

Table 3. Statistics of multi reference test and tuning sets

Corpus	BLEU on Test Set	BLEU on Tuning Set
Central Asia	24.82	24.52
News	27.70	29.76
Misc	20.72	22.61
Ted	14.74	18.23
Verbmobil	4.62	5.68
Existing corpus (Pen, Elra)	7.66	8.34

Table 4. Translation quality on generated corpora (BLEU %)

It is obvious that the News corpus which is translated by human has the best quality.

After generating a big corpus by means of automatic aligners and human translators, we offered the first interactive text normalizer for English-Farsi language pair. This is the first text normalizer for this language pair, which can normalize the text interactively. To show the effectiveness of this tool, we performed three experiments using our big corpus, which is the concatenation of all gathered corpora, plus two existing corpora (Table 2), as the training set and the same corpora of Table 3 as test and tuning sets. In the first trial, an SMT system is created without doing any text normalization on training, testing or tuning sets. Afterward, we did another experiment in which these data sets were normalized statically, i.e. normalizing the text using only a fixed normalization table which consists of valid English-Farsi characters. The final experiment related to this part was to generate a SMT system using interactively normalized data sets. Table 5 indicates the efficiency of the proposed text normalizer on the translation system. Three experiments are done. First, we test the translation system without normalizing the texts. Then we use static text normalization. Finally, interactive normalization is used and the results are as below.

Text Normalization	BLEU on Test Set	BLEU on Tuning Set
None	26.73	28.65
Static approach	27.83	28.60
Interactive approach	29.09	31.04

Table 5. Efficiency of interactive text normalizer (BLEU %)

The experiments clarify that while the static normalization improves quality of the translation, the interactive normalization improves it much more efficiently.

Our final set of experiments is related to the hand-crafted rules which are applied in order to weaken the structural dissimilarities between Farsi and English languages. To this end, four rules, described in section 4, are applied on the source language (English) to make its structure similar to Farsi’s. To show the effectiveness of these rules, we perform four experiments. In the first experiment, the baseline system with monotone reordering is created without applying rules. Afterward, we apply the manual rules on the datasets and then create three more SMT systems with monotone, distance-based, and lexicalized reorderings. The results of these experiments are shown in Table 6.

Reordering	Manual Rule	BLEU on Test Set	BLEU on Tuning Set
Monotone	No	26.04	28.19
Monotone	Yes	27.50	30.03
Distance-based	Yes	27.90	30.72

Table 6. Effects of manual reordering (BLEU %)

As the results demonstrate, using manual reordering results in a better BLEU on test set compared to the baseline model with no manual rules and monotone reordering. Since the manual rules are local and we did not apply long range reordering rules, the combination of manual rules and distance-based reordering performs better than manual rules with monotone reordering. Because of the same reason, i.e. the manual rules do not completely cover the structural differences of English-Persian; it does not perform better than the system which uses lexicalized reordering (Table 5).

6 Conclusion And Future work

In this research we try to create and introduce the first open-domain bilingual English-Farsi corpus which is gathered according to the standard approaches. Then a new text tokenizer/normalizer tool is proposed to normalize, tokenize, and tag the English-Farsi corpus and it is especially designed to interactively normalize the Farsi side to remove the character anomalies in Farsi. Finally, some manual rules are offered to improve the translation quality by decreasing the structural differences of the English-Farsi language pair. Future works includes making use of some other aspects of the proposed normalizer, i.e. the detected tags for special words. Also, find some other effective rules to apply global reordering to English verbs and other useful kinds of distortions to match Farsi sentence patterns.

References

[1] R. Nilipour, “Task- Specific Agrammatism In A Farsi- English Bilingual Patient”. JOURNAL OF NEUROLINGUISTICS, NO.4, pages 243-253, 1989.

[2] S. Narayanan, S. Ananthakrishnan, R. Belvin, E. Ettaile, S. Ganjavi,. “Transonics: A Speech To Speech System For English-Persian.”In the Proceedings of ASRU. U.S., Virgin Islands, pages 670-675, 2003.

[3] N. Bach, M. Eck, P. Charoenpornsawat, T. Köhler, S. Stüker. “The CMU Transtac 2007 Eyes-Free And Hands-Free Two-Way Speech-To-Speech Translation System.”,In the Proceedings of IWSLT, Kyoto, Japan, 2007.

[4] E. Ettelaie, S. Gandhe, P. Georgiou, K. Knight, D. Marcu,S. Narayanan, D. Traum, R. Belvin. “Transonics: A Practical Speech-To-Speech Translator or English-Farsi Medical Dialogs.” International Committee on Computational Linguistics and the Association for Computational Linguistics, pages 89-92, 2005.

[5] A. Kathol, J. Zheng. “Strategies For Building A Farsi-English SMT System From Limited Resources.” In Interspeech ‘08, pages 2731–2734, 2008.

[6] E.Matusov, S.Köprü. “Improving Reordering In Statistical Machine Translation From Farsi.” in AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas. Denver, Colorado, USA, 2010.

[7] T. Chung, M. Galley. “Direct Error Rate Minimization For Statistical Machine Translation.” Association for Computational Linguistics, pages 468-479,2012.

[8] F. Och, “Minimum Error Rate Training In Statistical.” Association for Computational Linguistics. Sapporo, Japan, pages 160-167, 2003.

[9] http://mokk.bme.hu/resources/hunalign

[10] http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

[11] H. Ney, F. Och, S. Vogel. “Statistical Translation Of Spoken Dialogues In The Verbmobil System.” In Workshop on Multi-Lingual Speech Communication, pages 69-74, 2000.

[12] M.A. Farajian, “Pen: Parallel English-Persian News Corpus.” Proceedings of the 2011th World Congress in Computer Science, Computer Engineering and Applied Computing, 2011.

[13] http://www.elra.info/LRs-Announcements.html

[14] http://ece.ut.ac.ir/NLP/resources.htm

[15] S. Karimi, “Machine Transliteration Of Proper Names Between English And Persian”, PhD thesis, 2008.

[16] http://qt.nokia.com