PEN: Parallel English-Persian News Corpus

Mohammad Amin Farajian
Advanced Information and Communication Technology Research Center,
Sharif University of Technology, Tehran, Iran

Abstract - *Parallel corpora are the necessary resources in many multilingual natural language processing applications, including machine translation and cross-lingual information retrieval. Manual preparation of a large scale parallel corpus is a very time consuming and costly procedure. In this paper, the work towards building a sentence-level aligned English-Persian corpus in a semi-automated manner is presented. The design of the corpus, collection, and alignment process of the sentences is described. Two statistical similarity measures were used to find the similarities of sentence pairs. To verify the alignment process automatically, Google Translator was used. The corpus is based on news resources available online and consists of about 30,000 formal sentence pairs.*

Keywords: Parallel corpus, alignment, statistical machine translation.

1 Introduction

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent a language or language variety as a source of data for linguistic research [1]. A corpus may contain texts in a single language (monolingual corpus) or text data in multiple languages (multilingual corpus). Multilingual corpora that have been specially formatted for side-by-side comparison are called *aligned parallel corpora*.

A large and well-aligned parallel corpus plays a key role in the success of a statistical machine translation (SMT) system. Creation of such a corpus is the most critical and time consuming task in developing an SMT system for a given language pair, especially when limited resources are available to form a parallel corpus as it is the case for Persian language.

In this paper, the work towards building a sentence-level aligned English-Persian corpus in a semi-automated manner is presented. The corpus is open ended, which means that it can grow in size and/or more languages can be added to the corpus. The paper is organized as follows: Section 2 describes Persian language and the difficulties in automatic processing of this language. Existing parallel English-Persian corpora are also reviewed in Section 3. Section 4 discusses the use of news articles in parallel corpus construction. Finally, Section 5 describes the development process for the corpus presented in this paper.

2 Persian language

Persian is an Indo-European language, which is the official language of Iran, Afghanistan, Tajikistan, and Uzbekistan. The modern Persian, as written in Iran, is a right-to-left script, which looks like Arabic script but it has its own alphabet and grammatical rules. There are some difficulties in the Persian writing system that make it hard to be processed automatically [2], [3]. This is because of some characteristics of Persian language including the following:

- Character Encoding: In addition to the range of Unicode characters dedicated to Persian, some Arabic characters are sometimes used alternatively. For instance, the letters ک (kaf), and ي (ye) can be expressed by either Persian Unicode encoding (U+06A9 and U+064A) or Arabic Unicode encoding (U+0643 and U+06CC or U+0649). As another example, the Persian letter ه (he) with Unicode encoding (U+0647) is sometimes replaced by Arabic letter ة (Teh Marbuta) with Unicode encoding (U+0629). Such cases usually impose difficulties and errors on automatic Persian language processing systems.

- Tokenization: There are various scripts for writing Persian texts, differing in (a) the style of writing words, (b) using or eliminating spaces within or between words, or (c) using various forms of characters. This makes Persian texts hard to be processed automatically. Tokenization, which is one of the early steps of text processing, therefore becomes a complex and challenging part in Persian language processing [4]-[5].

- Word Order: In Persian, normal sentences are generally structured with a subject-object-verb format. However, sentences may have relatively free word order, referred to as scrambled order. This scrambling characteristic gives Persian a high degree of flexibility for versification and rhyming. This feature, however, makes Persian more difficult to be processed automatically [3], [6].

3 Existing English-Persian parallel corpora

Shiraz test corpus [7] is the first attempt reported on developing English-Persian corpora. This corpus consists of 3,000 Persian sentences collected from a Persian corpus of online material. It was manually translated into English at New Mexico State University to test Shiraz machine translation (MT) system. Some efforts in developing speech-to-speech English-Persian MTs for army force protection and urgent care medical interactions were supported by DARPA[1]. The corpora used in these works have been collected from available corpora in other languages (e.g., English-Iraqi) or in-domain resources such as Medical Phrasebooks and translated manually [8]-[11]. Qasemizadeh, *et al*, made some efforts on building a parallel multilingual corpus for Persian in MULTEXT-East framework [12]. They used the Orwell's 1984 as the main text to construct the corpus. The Persian side of the corpus comprised about 6,606 sentences with about 110,000 tokens. Mohaghegh, *et al*, developed an open corpus from movie subtitles consisting of about 10,000 sentence pairs [13]. Pilevar, *et al*, on the other hand, take advantage of movie subtitles to form the largest parallel English-Persian corpus to date, called TEP [6]. It consists of about 554,000 sentence pairs with about 3 million words in both Persian and English. They, however, admit that movie subtitles contain daily conversations, which are informal and therefore cannot be easily annotated in an automatic manner. This limits the usability of the corpus in Persian natural language processing (NLP) applications. As the last example, European Language Resources Association (ELRA) constructed a corpus (commercially available online) consisting 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 consists of several different domains, including art, culture, idioms, law, literature, medicine, poetry, politics, proverbs, religion, and science.

4 Using news in parallel corpus construction

There are many resources that can be used to construct parallel corpora (e.g., literary translations, movie subtitles, Wikipedia documents, news, etc.).

- **Literary translations**

Although literary translations have been used in parallel corpus construction, they are less common for machine translation purposes [12]. This is mainly because literary texts are hard to translate even by human translators because of cultural differences between source and target languages. Literary translations are hard to align at sentence level because they usually are translated conceptually. They not only require a thorough knowledge of the source and target languages, but they also need to be able to correctly translate the original feelings and to employ the most appropriate language means in the translation procedure.

- **Movie subtitles**

Another resource for constructing parallel corpora is movie subtitles. As addressed by Pilevar, *et al*, in [6], movie subtitles have various advantages such as large amounts of entries, public availability, similarity of the source and target sentences in length, etc. There is, however, a general disadvantage for using movie subtitles for corpus construction and that is the informality of the sentences used. The problem becomes worse for Persian language noting the fact that for mapping from formal words/phrases to their informal equivalents, there is no mapping table available, nor any tool has been developed for this purpose. Therefore the use of such resources in Persian corpus construction is rarely reported and prone to error.

- **News stories**

To acquire a fairly large parallel corpus that could provide the necessary training data for experiments on statistical machine translation, we chose to mine news stories. There are various advantages in using news stories, such as:

- Large amounts of news stories are written or translated in many languages including Persian and English.
- They are publicly available and can be downloaded freely from a wide variety of online news resources.
- All the news stories in Persian are formal, so it is more appropriate to be used in NLP tasks.
- There is an increasing demand by news agencies for immediate and more accurate machine translations to translate news documents.

According to the above characteristics of news stories, increasing interest has been attracted to develop corpora based on bilingual news stories. For example, Fry built an English-Japanese parallel corpus from RSS news feeds which publish Japanese news stories from English originals [14]. In this work, the links in the Japanese articles to their English equivalents were used to match corresponding document pairs. An English-Japanese comparable corpus was also developed by Utiyama, *et al*, [15]. Nadeau and Foster used Canada Newswire (CNW) news feeds to build a parallel corpus of English-French texts [16]. Huang, *et al*, reported an English-Chinese comparable corpus based on news stories [17]. To the best of our knowledge, Shiraz parallel corpus [2] and the comparable corpus introduced in [18] are the only attempts to build English-Persian parallel corpora from the

[1] Defense Advanced Research Projects Agency

news. Shiraz corpus was initially developed by collecting online news articles from Persian news web sites, extracting a set of sentences varied along syntactic and domain dimensions, and then translating into English by Persian native speakers, manually. Like any other comparable corpus, the corpus introduced in [18] needs more processing (both automatic and manual) to be suitable for statistical machine translation tasks.

In the next section, the procedure of building a new sentence-level parallel corpus for English-Persian machine translation is described.

5 PEN corpus development

Building a sentence-level aligned English-Persian parallel corpus in a semi-automated manner is presented, which can be used in linguistic researches such as statistical machine translation. The corpus development procedure is as follows:

1. A manually-aligned *control corpus* is built comprising 1,200 sentence pairs. This corpus is developed to determine the similarity measures for English-Persian sentences.

2. To build the *main corpus* (PEN), news stories in both Persian and English are downloaded from the website of a multilingual news agency. News story pairs (i.e., news stories in English paired with their translations in Persian) are aligned in document level automatically.

3. Main corpus documents are then aligned in sentence level using the similarity measures obtained in step 1.

4. Quality of sentence alignment procedure is evaluated using Google translator as a reference. In this step, sentence pairs with poor translation quality are first tagged and subsequently removed manually.

Corpus development procedure is illustrated in Fig. 1.

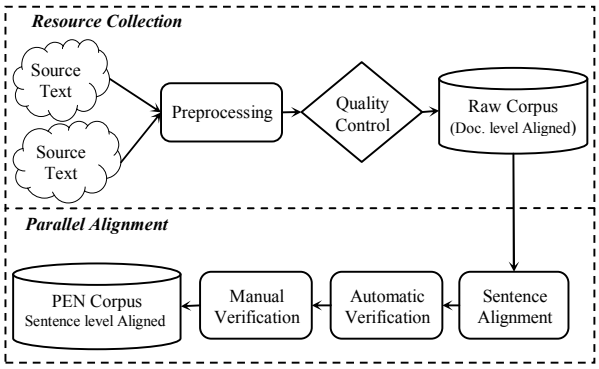


Figure 1 Development procedure for PEN

5.1 Resources

As described earlier, the raw material for constructing the corpus is collected from news stories available online on the website of a news agency. The stories (a total of about 5,700 documents) cover international news, brief news, articles, sports, interviews, etc.

5.2 Control corpus

Fortunately, most of the downloaded documents are already aligned in paragraph level, but in many cases the sentences in the paragraph pairs are not fully aligned. Misalignment of sentences could be classified in two categories:

1. *Sentence disorder*: All sentences in the source paragraph are exactly translated into the target language, but the order of sentences in the target paragraph is different from what appears in the source paragraph.

2. *Sentence mismatch*: (a) Target paragraph is an exact translation for the corresponding source paragraph, but it is expressed in a number of sentences different from what appears in the source paragraph. (b) When translating from one language into the other, some sentences are sometimes preferred to be removed or added.

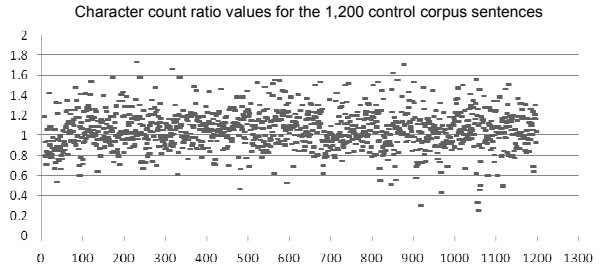
For automatic sentence alignment, some similarity measures are needed. Gale and Church used a simple statistical model of character lengths to align sentences in parallel corpora [19]. They used a normal distribution based on the lengths of sentences in both source and target languages in terms of the number of characters. Then, they specified the model by the mean value and standard deviation parameters of the distribution. In addition to the character-level measure introduced in [19], a similar statistical measure in word level was used for automatic sentence alignment in this work. These measures for the 1,200 control corpus sentence pairs are shown in Fig. 2, mean values and standard deviations of which were used for the alignment of the main corpus sentence pairs.

5.3 Preprocessing

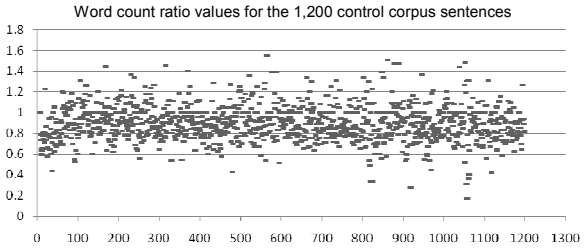
After document-level alignment, the main corpus needs to be prepared for sentence-level alignment. In this step, HTML tags of the aligned documents are removed. A normalization procedure is then performed to correct both character encodings and punctuation mark positions.

5.4 Sentence alignment

The document-pairs are classified into two categories according to their numbers of paragraphs: *Document pairs with different number of paragraphs* are discarded in this step and will be processed manually subsequently. *Document pairs with the same number of paragraphs*, on the other hand, are processed to extract the sentence pairs. Since the documents are selected from a news domain, they are full of



(a)



(b)

Figure 2 Distributions of (a) Character and (b) Word count ratios for the 1,200 sentence pairs of the control corpus

abbreviations[2] and also contain web addresses, both making the detection of sentence boundaries a rather complicated task. To overcome this problem, a sentence boundary detector (SBD) was developed for the extraction of parallel sentences from the corresponding paragraphs. It uses a small set of abbreviations in both Persian and English and some heuristics. This is to discriminate between the dots are used as 'full stops' from those used for other purposes, namely decimalpoint s and the ones used in web addresses (e.g., www.domain_name.com). Trying to process the paragraphs, the SBD first searches for non-full-stop dots and removes them keeping their locations in the text. When paragraph segmentation finishes, the non-full-stop dots are inserted in their original locations. Some sentence pairs may be missed in this step, due to the performance of the SBD module. So, improving the sentence boundary detector especially for Persian language is one of the future works.

Next step is the alignment of the extracted sentences using the two aforementioned similarity measures. In this step, only one-to-one correspondences in text units are used. In other words, a sentence in one language is matched with only one sentence in the other language. Application of this criterion may sometimes cause misalignments or even lead to not finding matching sentence pairs. Both automatic and manual verifications were done to avoid such problems.

[2] The abbreviations that are troublesome in this context are the ones that contain dots in between alphanumeric characters, e.g., U.S.A., U.N., E.U., Prof., and Mr.

5.5 Automatic verification

To improve the quality of the main corpus, Google translator was used to automatically verify the outcome of the sentence alignment procedure described in the preceding section. As illustrated in Fig. 3, automatic alignment verification is performed as follows: For each sentence pair, the English sentence is first translated into Persian using Google translator. Then, stop words of both the Persian sentence (from the corpus) and the output of Google translator are removed. The number of words identically appearing in both of the resulted Persian strings, referred to as the *matching factor*, is used to qualify how well the English-Persian sentence pairs are aligned. Aligned sentence pairs with a matching factor of more than 30% are considered as similar strings. Sentence pairs with the matching factor of less than 30% are tagged to be reviewed manually.

Table 1 shows some statistics of PEN corpus. Examples of the sentence pairs in PEN are presented in Table 2.

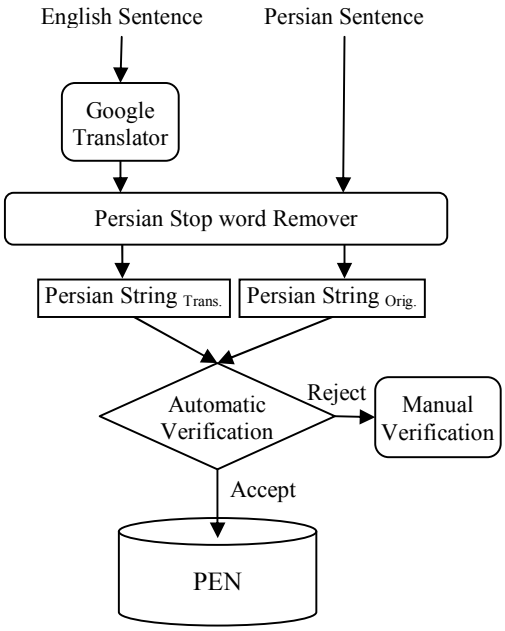


Figure 3. Automatic verification procedure

Table 1. Statistics of PEN

Parameters	English	Persian
Corpus size (in sentences)	30,479	30,479
Corpus size (in words)	567,078	716,089
Corpus size (in characters)	3,571,164	3,399,305
Average sentence length (in words)	18.6	23.5
Average sentence length (in characters)	117.2	111.5

Table 2. Some examples of sentence pairs in PEN

Unfortunately , were just not there yet. متاسفانه ما هنوز به این سطح نرسیده ایم .
In particular he noted that the conditions do not exist for developing major sports in the country. او به طور خاص اشاره کرد که در کشور شرایط رشد ورزش های اصلی مهیا نیست .
The tradition of celebrating the Muslim new year as navruz has been adopted by Uzbekistan , Kyrgyzstan , Azerbaijan , India , Iran , Pakistan and Turkey . برگزاری مراسم جشن سال نو مسلمانان با نام نوروز ، از سوی کشورهای ازبکستان ، قرقیزستان ، آذربایجان ، هند ، ایران ، پاکستان ، و ترکیه برگزیده شده است .
The reasons for delinquency vary . Some blame political and economic instability . Parents and teachers accuse each other of failing to disciple the children . Police say the children learn the behavior by watching adults . دلایل مختلفی برای ارتکاب جرم وجود دارد . برخی بی ثباتی سیاسی و اقتصادی را مقصر می دانند . والدین و آموزگاران یکدیگر را در آموزش ندادن انضباط به بچه ها متهم می کنند . پلیس می گوید که بچه ها رفتارها را با نگاه کردن به بزرگترها یاد می گیرند .

6 Conclusion

In this paper the work towards building a sentence-level aligned English-Persian corpus in a semi-automated manner was presented. The corpus was developed using about 30,000 sentence pairs extracted from about 57,000 news documents available online. The documents cover a wide variety of news domains including sports, politics, interviews, etc. Document-level aligned pairs were first preprocessed. Then, the documents were segmented into sentences and subsequently aligned using two similarity measures. Google translator was used to automatically verify the alignment of sentence pairs. In addition, some manual reviews were done to make PEN more accurate.

7 References

[1] Martin. Wynne. "Developing Linguistic Corpora: a Guide to Good Practice". Oxford: Oxbow Books, 2005.

[2] J. W. Amtrup, H. Mansouri Rad, K. Megerdoomian and R. Zajac. "Persian-English machine translation: An overview of the Shiraz project"; Technical report MCCS-00-319. New Mexico State University, Computing Research Lab, 2000.

[3] K. Megerdoomian. "Unification-Based Persian Morphology"; CICLing 2000, pp. 135-149, Feb 2000.

[4] S. Kiani. "Persian text tokenization and chunking"; 14th International CSI Computer Conference, Tehran, Iran, 2009.

[5] C. Saedi, M. Shamsfard, and Y. Motazedi, "Automatic Translation between English and Persian Texts"; Third International Workshop on Computational Approaches to Arabic-Script based languages (CAASL3), Ottawa, Canada, 2009.

[6] M. T. Pilevar, A. H. Pilevar, and H. Faili, "TEP: Tehran English-Persian Parallel Corpus"; CICLING 2011, Tokyo, Japan, 2011.

[7] R. Zajac, S. Helmreich and K.Megerdoomian, "Black-Box/Glass-Box Evaluation in Shiraz"; Workshop on Machine Translation Evaluation at LREC-2000, Athens, Greece, 2000.

[8] R. S. Belvin, W. May. S. Narayanan, P. Georgiou, S. Ganjavi, "Creation of a Doctor-Patient Dialogue Corpus Using Standardized Patients"; International Conference on Language Resources and Evaluation (LREC),2004 .

[9] E. Ettelaie, S. Gandhe, P. Georgiou, K. Knight, D. Marcu, S. Narayanan, D. Traum, R. Belvin, "Transonics: A Practical Speech-to-Speech Translator for English-Farsi Medical Dialogues"; International Committee on Computational Linguistics and the Association for Computational Linguistics, 2005.

[10] P. G. Georgiou, A. Sethy, J. Shin, S. Narayanan, "An English-Persian Automatic Speech Translator: Recent Developments in Domain Portability and User Modeling"; International Conference on Intelligent Systems and Computing (ISYC), 2006.

[11] N. Bach, M. Eck, P. Charoenpornsawat, T. Köhler, S. Stüker, T. Nguyen, R. Hsiao, A. Waibel, S.Vogel, T. Schultz, A. W Black, "The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System"; International Workshop on Spoken Language Translation(IWSLT-2007), Trento, Italy, 2007.

[12] Behrang Qasemizadeh, Saeed Rahimi, "The First Parallel Multilingual Corpus of Persian: Toward a Persian BLARK"; The second workshop on Computational Approaches to Arabic Script-based Languages (CAASL-2), California, USA, 2007.

[13] M. Mohaghegh, A. Sarrafzadeh, "Performance evaluation of various training data in English-Persian Statistical Machine translation"; 10th International Conference on the Statistical Analysis of Textual Data (JADT 2010), Rome, Italy, 2010.

[14] J. Fry, "Assembling a parallel corpus from RSS news feeds"; Workshop on Example-Based Machine Translation, MT Summit X, Phuket, Thailand, 2005.

[15] M. Utiyama and H. Isahara, "Reliable measures for aligning Japanese-English news articles and sentences";

Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 7-12, Sapporo, Japan, 2003.

[16] D. Nadeau, and G. Foster, “Real-time identification of parallel texts from bilingual newsfeed”; Computational Linguistic in the North-East (CLINE 2004), pp. 21-28, 2004.

[17] D. Huang, L. Zhao, L. Li, and H. Yu, “Mining Large-scale Comparable Corpora from Chinese-English News Collections”; the 23rd International Conference on Computational Linguistics (COLING’10), 2010.

[18] H. Baradaran Hashemi, A. Shakery, and H. Faili. “Creating a Persian-English Comparable Corpus”; Conference on Multilingual and Multimodal Information Access Evaluation (CLEF), pp.27-39, 2010.

[19] W. A Gale, and K. W. Chuch. “A program for aligninig sentences in bilingual corpora”; 2th Annual Meeting of the ACL, 1991.