PC-Corpus: A Persian-Chinese Parallel Corpora

Mohammad Hadi Mohammadi, Qiao Pan*, Dehua Chen and Marjan Kamyab
School of Computer Science and Technology, Donghua University, Shanghai, China

*Corresponding author e-mail: panqiao@dhu.edu.cn

Abstract. With the intense use of neural networks in machine translation of NLP as the latest and most promising approach, neural machine translation (NMT), however, has a serious weakness in terms of corpus availability. This type of machine translation though has achieved a significantly remarkable results for the rich-resource language pairs but it has a challenging performance for the low-resource language pairs. Persian-Chinese is one of the low-resource language pairs that lacks a large scale freely available parallel dataset for training translation models. In this paper we illustrate the creation of bilingual Persian-Chinese corpus (PC-Corpus), which is the very first corpus for this language pair. Our work is significantly considerable for the future of machine translation on Persian-Chinese language pair.

1. Introduction
In Natural Language Processing (NLP) applications, a text corpus is a structured electronic data collection ready for analysis. Corpses are of different types: a corpus is called monolingual if all its texts are in only one language and it’s called multilingual if its texts are in multiple languages. Parallel corpus is a special class of multilingual corpus in which sentences in two languages are aligned side-by-side in the form of source-target pairs [1]. Each sentence in the source language has its aligned sentence (translation) in the target language. Parallel data are valuable resources for natural language processing applications, such as machine translation, and linguistic research. Finding the corresponding resource and documents between two languages is the first step to construct a parallel corpus, before more fine-grained paragraph and/or sentence alignments can be calculated.

Neural Machine Translation (NMT) as an end-to-end structure of machine translation is greatly relying on parallel bilingual dataset. However, for some language pairs such as Persian-Chinese, sentence-level aligned parallel resources are rarely available which make neural machine translation systems less ultimate translation system for such language pairs.

In this paper we describe the construction of Persian-Chinese bilingual corpus that contains above 100,000 sentence pairs. Our proposed algorithm for constructing PC-corpus is a length-based sentence alignment approach is originated from a sentence alignment algorithm described by Gale and Church [2]. The task of sentence alignment is difficult because sentences frequently do not have a one-to-one alignment fashion. Sometimes sentences align many-to-one, and often there can be many sentence deletions in one of the supposedly parallel corpora of a bilingual corpus. As far as our knowledge, this is the first Persian-Chinese parallel corpus in the history of Machine Translation.

2. Related Work
During the recent years, many researches and investigations have been made aiming to build up as lager parallel bilingual corpora as possible. Some works have used World Wide Web as the source for
building parallel corpus while others due to the lack of enough parallel data for many language pairs, they have proposed a pivot approach in which they used English as a pivot language to extract parallel resources.

Ren et al., 2018, proposed pivot approach for rare language translation by introducing a third rich-source language as a pivot [3]. Their method tries to leverage bilingual data \((X, Y)\) (a rich-resource pair) and \((Y, Z)\) (maybe small) to improve the translation performance of low-resource pair \((X, Z)\), during which translation models of \((X, Z)\) and \((Y, Z)\) can be improved jointly. Resnik et al., (2003), working on web pages, use STRAND to recognize parallel pairs [4]. This is a language-independent tool that automatically find bilingual text from the Web. To identify parallel pairs, their method specifies a set of pair-specific values. They achieved good performance according to an experiment on English-Chinese language pair. Koehn et al., (2005), used the proceedings of European Parliament to extract parallel texts for 11 languages [5]. Smith et al., (2010), proposed a technique to align texts in document level from Wikipedia data for three language pairs, Spanish-English, German-English, and Bulgarian-English [6]. They use Hidden Markov Model for word alignment purpose. Ansari et al., (2017), is another work that use English as the pivot language to create Persian-Italian language pair texts [7]. In their work, sentences from Persian and Italian are translated into English and compared with each other using a new similarity metric which is based on Normalized Google Distance (NGD). Linard et al., (2015), using English as the pivot language, proposed two approaches to bilingual lexicon extraction [8]. One approach is to translate the source language to a third rich-language (pivot) first and then translate back from the middle language to the target language. In their second approach they translate the source as well as the target languages into a pivot language and then extract bilingual vocabulary.

Table 1. Statistics of Chinese and Persian corpora (→ not related, -→ not available)

| Corpora               | Characters/Words/Tokens | Sentences |
|-----------------------|--------------------------|-----------|
|                       | Chinese                  | English   | Persian |
| MultiUN Corpus        | 629M                     | 220M      | ~        | 4M |
| UM-Corpus             | 56.8M                    | 33.7M     | ~        | 2.2M |
| CAS Chinese-English   | -                        | -         | ~        | 0.5M |
| English Corpus        | ~                        | 3.7M      | 3.7M     | 0.55M |
| TEP Corpus            | ~                        | -         | -        | 0.2M |
| IASBS English-Persian | ~                        | -         | -        | -    |

Chinese and Persian, despite being important languages of the world, no efforts have been made for this language pair due to rare parallel bilingual resources. However, good efforts have been made for Chinese-English, Chinese-German, and Chinese-Japanese etc. [9]. MultiUN corpus [10](2010) obtained from the official records of the United Nations (UN) consisting 4 million pair sentences, UM-Corpus [11](2014) consisting 2.2 million pair sentences, and Chinese-English Parallel Corpus [9] by Chinese Academy of Science consisting 0.5 million pair sentences are the largest parallel Chinese-English corpuses. For Persian language there are scant number of parallel corpus which are either inconsiderable in size or inaccessible for research purpose. Tehran English-Persian Parallel (TEP) Corpus [1](2011) consisting 55,462 Persian sentences and their related English translations and English-Persian Parallel Corpus [12] from Institute for Advanced Studies in Basic Sciences consisting 200,000 bilingual paired sentences are recent largest Persian corpuses. Table 1 shows statistics of Chinese-English and Persian-English parallel corpuses.

3. Corpus Construction

There are two major challenges for us for building Persian-Chinese bilingual corpus: collecting bilingual texts to be aligned, and building up the sentence-alignment algorithm for alignment purpose. We illustrate the construction process of the corpus is 4 main steps:
3.1. Corpus source selection
Bilingual text resources are quite limited for a specific language pair, regardless of what language pair is, comparing to monolingual text resources. Bilingual books are usually either rare or due to their content omission and copyright protection in literary translations such as translated books, they are not a suitable resource for parallel corpora construction purpose. Hence, collecting bilingual resources from the World Wide Web could be a good choice. Unfortunately for Persian-Chinese language pair there are almost either no web pages available providing the same contents in Chinese and its corresponding Persian translation or if available, their contents are not sufficient to construct a parallel corpora. However, we found out that TED.com is the only web source with relatively enough and comparably sufficient subtitles in Chinese and Persian languages to build a corpus. Therefore, the source of our corpus is Ted-Talk subtitles.

TED is a media organization that posts talks online for free distribution under the slogan "ideas worth spreading." TED Translators started as TED Open Translation Project in 2009 to provide talks’ translations and subtitles for non-English speakers in different languages. In 2015, over 70,000 sets of subtitles in 107 languages including Persian and Chinese had been completed by (an all-time total of) 38,173 volunteer translators.

3.2. Data collection
To collect parallel text from TED, we wrote a python code to crawl the HTML webpage of TedTalk. This program initially identifies if a specific talk has the interested Persian and its corresponding Chinese subtitles. When the talk meet the parallel subtitle requirement, the program start to parse the HTML page of the talk to get the subtitles in Persian and Chinese language. We downloaded and created a database of Persian-Chinese subtitles of 1950 talks. The format of subtitles are in SRT format. This bilingual database of subtitles includes topics from different domains namely technology, entertainment, design, scientific, cultural, and academic talks. All talks that have Chinese subtitles and corresponding Persian subtitles have been included in the database. Fig. 2 shows a sample of a subtitle with multiple separate displays and their related time frames. The number at the top of each separate display shows the count of that display in the whole talk. Each time frame is made up of three parts: the time in the left most part which indicate the start time of the display, an arrow which shows the time flow, and the time at the right most part which shows the end time for the display.
3.3. Data pre-processing and noise removal

Additional meaningless information in data is called noise. The database of subtitles were initially noisy with a large amount of useless whitespaces and contained duplicated sentences. Sentences in Fig. 3.1 and Fig. 3.2 show sentences repeated in almost each talk’s subtitle and sentences repeated multiple times in each single talk’s subtitle, respectively. We have coded a python program to check for these sentences in an interactive way to make sure we are not removing other useful sentences. Our next duty, before sentence alignment, is to eliminate the extra whitespaces in the gathered raw data. There are two principles regarding to the whitespaces in Persian and Chinese language: the first principle is that there should be no whitespaces or tab spaces between continual Chinese characters, while a sequence of tokens in Persian language can have just a mono-space in between. In the second principle, there should zero space between non-Chinese characters subsequent to a Chinese character. Elimination of such additional whitespaces are essential because these eliminations alters the different text formats into a mono-format and also can boost the quality of future functions that use the corpus. Punctuation marks that were misplaced in some files were taken care of by another pre-processing python code.

3.4. Sentence alignment

The text is aligned at sentence level. Sentence alignment of subtitles are similar to that of normal sentence alignment but with an extra advantage of carrying additional information such as time frames that in turn can help alignment process.

The algorithm that we used for alignment purpose in our work is originated from the algorithm explained by Gale and Church [2]. We made a slight modification in the algorithm to get the full benefits of available time-frames in the subtitles. Gale & Church’s method dynamically calculate a minimal cost alignment on the basis of normalized sentence length in characters satisfying some constraints.
Let \( S_h = (h^1 \cdots h^{|S_h|}) \) and \( S_f = (f^1 \cdots f^{|S_f|}) \) be the subtitles of Chinese and Persian languages respectively. Let \( h^1 \cdots h^i \) be the sentences of Chinese, \( f^1 \cdots f^j \) be the translations of those sentences in Persian and \( c(h, f) \) be the cost of aligning \( h \) with \( f \), according to this algorithm the minimum cost \( C(i, j) \) of aligning \( h^1 \cdots h^i \) with \( f^1 \cdots f^j \) is calculated using the following recurrence.

\[
C(i, j) = \min \begin{cases}
C(i - 1, j - 1) + c(h^i, f^j) \\
C(i - 1, j) + c(h^i, \emptyset) \\
C(i, j - 1) + c(\emptyset, f^j) \\
C(i - 2, j - 1) + c(h^{i-1} || h^i, f^j) \\
C(i - 1, j - 2) + c(h^i, f^{j-1} || f^j)
\end{cases}
\tag{1}
\]

Where \( \emptyset \) represent an empty string and \( x || y \) indicate the concatenation of \( x \) with \( y \). The cost function \( c(h, f) \) is defined on the basis of relatively normalized sentence length, namely \( \frac{\ell(h)}{\ell(S_h)} \) and \( \frac{\ell(f)}{\ell(S_f)} \) where \( \ell(S_h) \) and \( \ell(S_f) \) are the total lengths of the subtitle files of Chinese and Persian languages, respectively. Our modified algorithm allow substitution (1-1), insertion (0-1), deletion (1-0), expansion (1-2) and contraction (2-1) alignments. Since in subtitle translations many-to-many alignments are quite rare, our algorithm is designed to not allow (2-2) alignment for performance purpose.

In addition to the length of sentences in characters, our algorithm also take the advantage of timing information \([13]\). Timing information is of great value in subtitle alignment due to the fact that the subtitles matching each other must be viewed at the same time. However converting the absolute time values of frame numbers into real time values are not an easy task.

Let the duration \( \text{dur}(s) \) of the subtitle \( s \) be:

\[
\text{dur}(s) = \text{end}(s) - \text{begin}(s),
\tag{2}
\]

The cost function \( c(h, f) \) can be measured in absolute values as:

\[
c(h, f) = \lambda \left( \frac{\text{dur}(h)}{\text{dur}(S_h)} - \frac{\text{dur}(f)}{\text{dur}(S_f)} \right)^2 + (1 - \lambda) \left( \frac{\ell(h)}{\ell(S_h)} - \frac{\ell(f)}{\ell(S_f)} \right)^2
\tag{3}
\]

Where \( \text{dur}(S_h) \) and \( \text{dur}(S_f) \) are the overall duration of Chinese and Persian language subtitles, and \( 0 \leq \lambda \leq 1 \) is a language-dependent parameter indicating the relevant importance of the timing information and can be obtained using grid-search empirically. Using duration-length-based algorithm leaded us for more accurate alignment of subtitles. Fig. 4 is a portion of our Persian-Chinese aligned sentences.

**Figure 4.** A sample of pc-corpus
4. Statistics of PC-Corpus
The first release of CP-corpus is summarized in Table 2. The frequency of the sentences, based on their lengths in characters, are illustrated in Fig. 5 and 6 for Persian and Chinese languages respectively. As observed in the figures, the length of sentences in Persian, due to inflectional morphology, tends to be longer than that of Chinese translation.

| Language | Size in characters | Size in words | Num. of sentences |
|----------|--------------------|---------------|------------------|
| Persian  | 13189048           | 2680891       | 100,500          |
| Chinese  | 4288174            | 4288174       | 100,500          |

Figure 5. Frequency of Persian sentences based on length

Figure 6. Frequency of Chinese sentences based on length

5. Conclusion
Despite the fact that many researchers have carried out researches and investigations on the use of comparable corpora to generate initial training data for NLP, we still have a lack of bilingual corpus for different language pairs in machine translations.
In this paper, we seek to build a corpus that align texts in two languages, Persian and Chinese, in parallel fashion. We use Ted-Talk website as the source of our crawled raw data. We propose duration-length-based technique to develop Persian-Chinese bilingual corpus (PC-Corpus). We also provide the statistics of this corpus.

PC-corpus, as the first bilingual parallel corpus, can facilitate many researchers’ works in the area of Natural Language Processing such as Neural Machine Translation, etc. This corpus can be an important resource for cross-language information processing, and also important resource for Persian-Chinese bilingual lexicography, language research and teaching. We expect that this work can add up more determinations toward constructing wide-range parallel corpora for this language pair in future.

Acknowledgments
This work was supported by the Science and Technology Development Foundation of Shanghai (16JC1400802, 16JC1400803, 18511102703), the Special Fund of Shanghai Municipal Commission of Economy and Informatization (RX-RJJC-08-16-0483, 2017-RGZN-01004, XX-XXFZ-02-18-2666, XX-XXFZ-01-18-2604).

References
[1] Pilevar M.T., Faili H., Pilevar A.H., TEP: Tehran English-Persian Parallel Corpus. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2011. Lecture Notes in Computer Science, vol 6609. Springer, Berlin, Heidelberg.
[2] William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. Comput. Linguist. 19(1):75–102.
[3] Ren, Shuo & Chen, Wenhu & Liu, Shujie & Li, Mu & Zhou, Ming & Ma, Shuai. (2018). Triangular Architecture for Rare Language Translation.
[4] Resnik, P., & Smith, N. A. (2003). The web as a parallel corpus. Computational Linguistics, 29(3), 349-380.
[5] Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In MT summit (Vol. 5, pp. 79-86).
[6] Smith, J. R., Quirk, C., & Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 403-411). Association for Computational Linguistics.
[7] Ansari, E., Sadreddini, M. H., Sheikhalishahi, M., Wallace, R., & Alimardani, F. (2017). Using English as Pivot to Extract Persian-Italian Parallel Sentences from Non-Parallel Corpora. arXiv preprint arXiv:1701.08339.
[8] Linard, A., Daille, B., & Morin, E. (2015). Attempting to bypass alignment from comparable corpora via pivot language. ACL-IJCNLP 2015, 32.
[9] Sun, Le & Xue, Song & Qu, Waimin & Wang, Xiaofeng & Sun, Yufang. (2002). Constructing of a large-scale Chinese-English parallel corpus. 10.3115/1118759.1118770.
[10] Eisele, A., Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In Proceedings of the 7th International Conference on Language 1841 Resources and Evaluation (LREC’2010).
[11] Tian, Liang & Wong, Derek & Chao, Lidia & Quaresma, Paulo & Oliveira, Francisco & Li, Shuo & Wang, Yiming & Lu, Yi. (2014). UM-Corpus: a large English-Chinese parallel corpus for statistical machine translation. Proceedings of the 9th International Conference on Language Resources and Evaluation.
[12] Karimi, Akbar & Ansari, Ebrahim & Bigham, Bahram. (2017). Extracting an English-Persian Parallel Corpus from Comparable Corpora.
[13] Itamar, E., Itai, A.: Using movie subtitles for creating a large-scale bilingual corpora. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008). European Language Resources Association (ELRA), Marrakech (2008).