A Study on Machine Translation-oriented Parallel Corpus Construction Techniques for Tibetan, Chinese and English

Cizhen Jiacuo, Sangjie Duanzhu, Zhoumao Xian

Qinghai Normal University

E-mail: 543819011@qq.com, sangjeedondrub@live.com, 513576745@qq.com

Abstract

The scarcity of parallel resources between Tibetan and other languages lays a great difficulty for application of current researches in the fields of neural networks and deep learning. The construction of a large-scale parallel Tibetan corpus for Chinese, English and other languages also serves as a great importance for Tibetan NLP in general. More importantly, machine translation between Tibetan and other languages also poses many challenges compared to some current mature machine translation systems like English and other languages. The availability of large-scale multi-lingual parallel language resources is essential to enable minority language machine translation services to better serve the "Belt and Road"'. In this work, through the research of the chapter-level, paragraph-level, sentence-level and word-level automatic acquisition techniques of Tibetan to other language texts, we proposed methods to acquire the knowledge needed for machine translation from the depth and breadth of knowledge mining. This first task in the work is to research on web-oriented automatic discriminant and extraction algorithms for acquiring the comparable corpus, at the same time, by maximizing local matching, to expand the size of the word alignment, phrase alignment library (block aligned library), in order to enrich the Tibetan related parallel language resources. The second is to study on individual paragraph representations based on the large-scale Chinese, Tibetan and English monolingual corpus. And by comparing the similarity of representations and optimizing the threshold to evaluate bilingual comparability both in horizontal and vertical directions. And third is to study the methods to improve the alignment of language resources using monolingual and trilingual word representations as well as the paragraph representations.

Keywords: Tibetan, Language resources, comparable corpus, alignment

1. Introduction

Availability of large-scale parallel corpus is the most indispensable resource for machine translation system, especially in Neural Machine Translation(NMT)(Bahdanau, Cho, & Bengio, 2014; Sutskever, Vinyals, & Le, 2014) Lacking large-scale language corpus poses a major practical problem for many language pairs (Artetxe, Labaka, Agirre, & Cho, 2017). Tibetan related studies in Natural Language Processing fields bloom in recent years, however, scarcity of parallel language resources is largely restraining the further researches. In this work, we present an automatic approach for constructing Tibetan-Chinese-English trilingual parallel corpus by exploiting sentence similarity which is computed via comparing the continuous representation of sentences(Le & Mikolov, 2014) extracted from comparable corpus we collected. Comparable corpus is a set of monolingual corpora which usually in same or similar topics in different languages. This kind of corpora paralleling on document level can be obtained from the web and other sources due to sentences in it are not necessarily translations of each other language pair, therefore building and using comparable corpora is often a more feasible option in multilingual information processing (Liu & Zhang, 2013).

Many Chinese governmental sites such as http://www.people.com.cn/ have multiple language locale settings including Chinese, Tibetan, English etc., and the content in those sites can serve as a good source to construct comparable corpus. Furthermore, we have large amount in-house parallel documents (shown in Table 1) in Tibetan, Chinese and English languages.

2. Related Works

Building comparable corpus attracted attention in the fields due to its flexibility and cost to construct relatively large parallel corpus which is usually hand-crafted with a huge amount of time and efforts. (Resnik, Philip, Smith, & Noah, 2003) presented an approach to mining parallel using STRAND software on Internet through supervised modeling based on structural features. (Talvensaari, Laurikkala, Juhola, & Keskustalo, 2007) presented a method to create comparable corpus by using relative term frequency (RAFT) value from collections in very different in origin. (Shang et al., 2017) proposed a framework, AutoPhrase, which extract high-quality phrases from the public knowledge base in an effective and automatic manner without the availability of POS tagger and rules designed by human experts. (Hashemi, Shakery, & Faili, 2010) align documents crawled from BBC news in different languages by comparing the similarity of document topics and corresponding publishing dates. In (Yasuda & Sumita, 2008) the authors reported an approach to translate original article and translate it into another language, then calculate the evaluation scores upon the translated and original articles. The evaluation is utilized to predict the similarity between two original Wikipedia articles.

In many recent researches, neural networks models are also proposed on the subject of extraction parallel sentences from comparable corpus, in (Chenhui Chu Raj Dabre, 2016) the author proposed a new method to first train a filter using

a seed parallel corpus and then use this filter to classify parallel sentence candidates.

3. Methods

3.1 Comparable Corpus

A *noisy parallel corpus* contains bilingual sentences that are not perfectly aligned or have poor quality translations. Nevertheless, most of its contents are bilingual translations of a specific document.

A *comparable corpus* is built from non-sentence-aligned and untranslated bilingual documents, but the documents are topic-aligned.

This article mainly uses the Internet to collect corpora, which will eventually be used in the construction of comparable Chinese-English-Tibetan corpora. The collected linguistic data should be of high research value while guaranteeing its stability and extensiveness. The obtained network text corpora are basically from Translation portal for Tibetan, Chinese, and English. In order to ensure the practicality of the trilingual data, we have collected a large amount of news corpora because of the high accuracy of news linguistic materials, clear logical logic, and certain representation. The corpora of this article are partly from Xinhua.net and people.com. Some of the nets come from their own data. The scale of the collected corpus is shown in Table 1.

Language	Xinhua	People.com.cn	In-house	Total
Chinese	3200	2500	5600	11300
English	2900	2450	3600	8950
Tibetan	2000	2400	5600	10000

Table 1: Comparable Corpus size in document counts.

In this paper, we use the similarity descending order to set the threshold value for selecting the comparable expectation. The experimental steps for putting the similarity greater than the threshold value into the comparable prediction library are as follows:

1. Filter the sample data
2. Segment the sample corpora, remove the stop words, and add the word bag model
3. Calculate the feature vectors of each word in the word bag and use these feature vectors to represent the document vector
4. Train the model and calculate the index with the calculated feature vector values
5. Mapping the trained sample corpora and candidate corpora to a two-dimensional space to calculate the Euclidean distance
6. Tag the text part-of-speech and filter the text not within the score range
7. Threshold Similarity Filtering, Compliant Documents Added in Comparable Languages, Deletes Not Threshold

Delete the corpus with accuracy less than 80%, and add the accuracy higher than 80% to the corpus3.2 3.2 Training Word Vector Words vector training for Tibetan, Chinese, and English trilingual words is performed using word2vec. Training samples are collected news corpus and own corpus.

3.2 Training word embedding

Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers. Conceptually it involves a mathematical embedding from a space with one dimension per word to a continuous vector space with much lower dimension. Methods to generate this mapping include neural networks (Bengio, Ducharme, Vincent, & Janvin, 2003),dimensionality reduction on the word co-occurrence matrix, and explicit representation in terms of the context in which words appear. Word and phrase embeddings, when used as the underlying input representation, have been shown to boost the performance in NLP tasks such as syntactic parsing and sentiment analysis.

Words vector training for Tibetan, Chinese, and English trilingual words is performed using word2vec. Training samples are collected news corpus and own corpus.

In this work, each word is export to a vector(Le & Mikolov, 2014) . The concatenation or sum of the vectors is then used as features for prediction of the next word in a sentence. More formally, given a sequence of training words $\{w_1, w_2, \dots, w_T\}$, the objective of the word vector model is to maximize the average log probability:

$$\frac{1}{T}\sum_{n=1}^{T} \log \mathrm{p}(w_t | w_{t-k}, \dots, w_{t+k})$$

The prediction task is typically done via a multiclass classifier, such as softmax. There, we have:

$$\mathrm{p}(w_t | w_{t-k}, \dots, w_{t+k}) = \frac{e_{y_{w_t}}}{\sum_i e^{y_i}}$$

3.3 Sentence Alignment Technique

Large corpora used as training sets for machine translation algorithms are usually extracted from large bodies of similar sources, such as databases of news articles written in the first and second languages describing similar events.

However, extracted fragments may be noisy, with extra elements inserted in each corpus. Extraction techniques can differentiate between bilingual elements represented in both corpora and monolingual elements represented in only one corpus in order to extract cleaner parallel fragments of bilingual elements. Comparable corpora are used to directly obtain knowledge for translation purposes. High-quality parallel data is difficult to obtain, however, especially for under-resourced languages.

Comparability is an important indicator of internal evaluation of comparable corpus. The current academic community does not clearly define its concept, but it is usually closely related to the concept of similarity. In most cases, the comparable degree of corpus can be considered as its

similar degree. We believe that comparability can be understood as the degree of similarity of the comparable corpus in terms of authorship, time, space, genre, source, domain, etc., or body and body information (grammar morphology, semantic content, pragmatic characteristics, etc.). This article mainly investigates the comparability of the comparable content of the Chinese-Tibetan-English news corpus, i.e. the similarity. Sentence similarity is calculated as follows: Suppose the sentence T_1 consists of n words, each word's word frequency weight is set to q_n, then there are:$T_1 = \{q_1, q_2, \dots, q_n\}$, T_1 is a multidimensional vector; Let sentence T_2 be composed of n words. The word frequency weight of each word is set to w_n. Then there are: $T_2 = \{w_1, w_2, \dots, w_n\}$, , T_2 is a multidimensional vector; then T_1 and T_2 similarity $\mathrm{Sim}(T_1, T_2)$ is calculated as:

$$\mathrm{Sim}(T_1, T_2) = \cos\alpha = \frac{\sum_{i=1}^{n}(q_i \times w_i)}{\sqrt{\sum_{i=1}^{n} {q_i}^2} \times \sqrt{\sum_{i=1}^{n} {w_i}^2}}$$

The corpus constructed can improve the use of corpus and its application scope by improving the alignment methods, corpus segmentation processing, similarity calculation, and the establishment of comparable relationships.

4. Experiments

In our experiments, we firstly train word embedding for Chinese, Tibetan and English with collected comparable corpus using fasttext[1]. And then two separate machine translation (MT) models were trained bi-directionally for two language pairs, namely Chinese↔English and Chinese↔Tibetan. For the former language pairs we used Google's NMT (Neural Machine Translation) system[2] due to WMT officially provides large amount training datasets. For the latter ones, given we have no access to enough Chinese-Tibetan parallel corpus, and as reported in (Chen, Liu, Cheng, & Li, 2017), training NMT models on low-resource language pairs usually perform poorly than SMT (Statistical Machine Translation) systems, we turn to Moses SMT toolkit [3]to train the model.

After training the MT models, the whole trilingual comparable corpus is processed with language dependent sentence segmenter. And then translate these sentences in to its corresponding languages.

We evaluate sentence similarity in two metrics: BLEU (Papineni, Roukos, Ward, & Zhu, 2002) score of original sentence and translated sentence (Yasuda & Sumita, 2008) and Cosine similarity (Luo, Zhan, Wang, & Yang, 2017) between the embeddings of the two sentences aforementioned.

The aligned sentences is pipelined into retraining process of both NMT and STM systems to gradually improve system's performance in a repeated iterative manner.

5. Conclusion

This paragraph collects and collates comparable corpus through the Internet, and gives a detailed account of the construction techniques and scale, uses word2vec to vectorize the collected corpora, and proposes a method for constructing a Chinese-English-Chinese-English comparative corpus. a corpus with a total of one million words, and evaluates the similarity calculation methods of sentences and the comparability of corpus.

6. Acknowledgements

This work is supported by grant (project ID: 2015-SF-520) from Provincial Science and Technology Department, Qinghai, PR China.

7. References

Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2017). UNSUPERVISED NEURAL MACHINE TRANSLATION. Retrieved from https://arxiv.org/pdf/1710.11041.pdf

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate, 1–15. https://doi.org/10.1146/annurev.neuro.26.041002.131047

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A Neural Probabilistic Language Model. The Journal of Machine Learning Research, 3, 1137–1155. https://doi.org/10.1162/153244303322533223

Chen, Y., Liu, Y., Cheng, Y., & Li, V. O. K. (2017). A Teacher-Student Framework for Zero-Resource Neural Machine Translation. Retrieved from http://arxiv.org/abs/1705.00753

Chenhui Chu Raj Dabre, S. K. (2016). Parallel Sentence Extraction from Comparable Corpora with Neural Network Features.

Hashemi, H. B., Shakery, A., & Faili, H. (2010). Creating a Persian-English Comparable Corpus. dx.doi.org.

Le, Q. V, & Mikolov, T. (2014). Distributed Representations of Sentences and Documents, 4, II-1188.

Liu, S., & Zhang, C. (2013). Termhood-based Comparability Metrics of Comparable Corpus in Special Domain. ArXiv E-Prints.

Luo, C., Zhan, J., Wang, L., & Yang, Q. (2017). Cosine Normalization: Using Cosine Similarity Instead of Dot Product in Neural Networks. Retrieved from http://arxiv.org/abs/1702.05870

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. Retrieved from http://www.aclweb.org/anthology/P02-1040.pdf

Resnik, Philip, Smith, & Noah, A. (2003). The Web as a parallel corpus. Computational Linguistics, 29(3),

[1] https://github.com/facebookresearch/fastText

[2] https://github.com/tensorflow/nmt

[3] http://www.statmt.org/moses/

349–380.

Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C. R., & Han, J. (2017). Automated Phrase Mining from Massive Text Corpora.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. Retrieved from http://arxiv.org/abs/1409.3215

Talvensaari, T., Laurikkala, J., Juhola, M., & Keskustalo, H. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. Acm Transactions on Information Systems, 25(1), 4.

Yasuda, K., & Sumita, E. (2008). Method for Building Sentence-Aligned Corpus from Wikipedia.