International Journal of Database Theory and Application
Vol.8, No.4 (2015), pp.273-280
http://dx.doi.org/10.14257/ijdta.2015.8.4.27

Tibetan-Chinese Bilingual Sentences Alignment Method based on Multiple Features

Lirong Qiu

School of Information Engineering, MinzuUniversity of China
Beijing, China
E-mail:qiu_lirong@126.com

Abstract

Sentence-level aligning bilingual parallel corpus is shown significant and indispensable status in machine translation, translation knowledge acquiring and bilingual lexicography research fields, which is the fundamental work for natural language processing. Given the great deal of work in sentence alignment and a variety of methods have developed for bilingual terminology extraction, those are unpractical for newly underway Tibetan information processing because those methods have to use a large number of manufactured sentences as training corpus while extracting inter-translatable word pairs. This paper proposes a multi-strategy Tibetan-Chinese sentence alignment method based on length of sentence, syntactic rules and bilingual dictionary. We test our approach on a bilingual corpus crawled from bilingual website and perform manual evaluation on bilingual sentences pairs extracted from Tibetan-Chinese corpora.

Keywords*: sentences alignment, comparable corpora, sentence pair, sentence bead*

1. Introduction

The enormous value of corpus resources for natural language processing research has received more and more recognition. Especially the sentence-level aligning bilingual parallel corpus resources have shown significant and indispensable status in machine translation, translation knowledge acquiring and bilingual lexicography research fields.

Sentence-level aligning bilingual parallel corpus is also the important basic resource for cross-language information retrieval, translation lexicography, bilingual term automatic withdrawal and multilingual comparison study, etc. There is an urgent need to do more research on Tibetan-Chinese bilingual sentence alignment technology.

The requiring of sentence-level aligning Chinese-Tibetan bilingual parallel corpus still exists serious shortcomings so far, plenty of works still need manual operation. The advantage of manual alignment is high aligning accuracy, meanwhile the disadvantage on slow speed is very clear. As manual alignment could not meet the needs of the rapid expansion of modern science and technology, it is of great significance to develop a feasible research method on automatic withdrawal of Tibetan-Chinese sentences.

The algorithm of bilingual sentences alignment has been quite mature in English-French English-German and Chinese-English currently [9]. Sentence alignment algorithm is generally divided into three categories: length-based algorithm, vocabulary-based algorithm and algorithm of comprehensive use of length and vocabulary.

Length-based algorithm is only appropriate for the text which has no or little noise, its dynamic planning frame is a good solution for solving sentence problems and widely used.

Vocabulary-based algorithm has nice robustness, but it costs a long time to acquire the vocabulary for its complex model. It is also unpractical for newly underway Tibetan

ISSN: 2005-4270 IJDTA

information processing because it has to use a large number of manufactured sentences as training corpus while extracting inter-translatable word pairs.

In practice, we have some sentences in Tibetan from the web, and most of the sentences are translated from the Chinese website. Our mission is to find the bilingual sentence pairs. We've found that, in most cases, if there is a sentence in Tibetan, there will be a sentence in Chinese. Conversely, given a Chinese sentence, the associated Tibetan sentence is often failed to found.

Under these conditions, this paper proposes a multi-strategy Tibetan-Chinese sentence alignment method based on length of sentence, syntactic rules and bilingual dictionary. Firstly, this method aims at the length of one sentence of Chinese, screens out all possible aligned Tibetan sentences making use of sentence-length-based algorithm. Secondly, judge with the syntactic rules, such as whether the Tibetan has obvious corresponding syntactic rules for Chinese interrogative sentence or exclamatory sentence. Thirdly, do accurate Tibetan-Chinese sentence alignment with bilingual dictionary.

The remainder of the paper is organized as follows. Section 2introduces the preliminaries of our work. Section 3 presents our work on multi-features Tibetan-Chinese sentence alignment approach. The empirical analysis and the results are presented in Section 4. In Section 5, we provide an overview of related work on sentence-level bilingual alignment, followed by the conclusions, discussions, and future work in Section6.

2. Related Work

Brown proposed word-number-length-based aligning method for sentence in 1991 [1], the basic idea was that the more the length of sentences approached, the more possible for them to become mutual translation sentence pairs. This method was applied in the large-scale bilingual corpus. In Brown's model, the length of sentence was decided by the number of words, the optimum sentence pairs were determined with the methods of using greatest-probability model and dynamic planning. Gale and Church published the improved algorithm in 1993 [2].The length of sentence was decided by the number of characters. It improved the efficiency of algorithm by dividing the bilingual corpus into smaller corpus segment.

The length-based algorithm is suitable for the circumstance when there is a strong correlation between the length of the text and the length of the translation. This method could be divided into word-number-length-based aligning method and character-number-length-based aligning method.

Length-based aligning method for sentence regards sentence alignment as the function of the sentence length. The advantage of which is it does not need extra dictionary information, while the disadvantage is the spread of mistake may be easily caused.

However, for Tibetan-Chinese sentence alignment, there are some issues of length-based aligning method. For example, the Tibetan-Chinese sentence pair is:

(1) ཕྱིང་མོ་ནི་སྐག་རྒྱུ་འཇིགས་རྒྱུ་མེད་པ་ཞིག་ཡིན་ཀྱང་།ཨོ་གང་ཟུར་དུ་དྲང་བྱེད་པ་གོ་མེད་ཅིག་ཡིན་པས།

尽管妹妹无恐无惧，但是不管怎么说，她还是一个不懂事的孩子。 (The sister is brave, but she is only a child.)

Also, there are some idioms in both the language. For example:

(2) ང་རྒྱལ་ལེང་ཏེགས་

骄傲的山岗上存不住知识之水. (Proud of the mountains cannot save water.)

(3) ཨ་མཆོག་རྒྱབ་ཀྱི་ཉགས་པ་

耳边风. (Taking advice like passing wind.)

In the three examples, the length of Tibetan sentence and the length of the Chinese translation share are no guarantee of logically associated. In most cases, The Tibetan

sentence is longer than related Chinese sentence, especially when there is a idiom or some named entities in the Chinese sentence.

The algorithm from Brown, Gale and Chen introduced the conception of anchor and divided the whole corpus into several smaller segments when aligning Hansard corpus [3]. It adopted the specific annotation from corpus to serve as anchor, and matched these anchors with dynamic planning algorithm. The text between these anchors could one-to-one correspond and the aligned text was formed after the matching. But method of adopting the specific annotation to serve as anchor is not available to general conditions.

By the means of counting the frequentness and location information of the words in the text, Fung adopted the high frequency inter-translatable words as candidate anchors in sentence aligning procedure [4]. He found out the real anchor by making use of dynamic planning algorithm to align the candidate anchors in bilingual text. This method had a large calculated amount for counting all the words. It also may cause the mistake of wrong anchor alignment for data sparseness.

Kay and Roscheisen proposed vocabulary-based aligning method for sentence in 1993 [5]. The basic idea was that assisting sentence alignment by making use of language feature or the matching for vocabulary pairs from bilingual dictionary.

For Chinese-English sentence alignment, there are some outstanding works, such as [7] and [12]. Wu et al. proposed Chinese-Tibetan sentence alignment method based on bilingual dictionary [10,15].

Given the great deal of similar work in bilingual sentence alignment and a variety of methods have developed for bilingual entity recognition, but less of them are focus on multi-features based Tibetan-Chinese sentence alignment.

Our work mainly focuses on length of sentence, syntactic rules and bilingual dictionary to improve the results of Tibetan-Chinese sentences alignment.

3. Multi-strategy Tibetan-Chinese Sentence Alignment Method

The realization of sentence alignment could basically divided into some procedures: 1) unite all the paragraphs from original text into one, 2) eliminate the disordered paragraph boundary, 3) apply sentence alignment method on the corpus, 4) evaluate the results.

The basic conception that referred in aligning method is introduced firstly.

Sentence pairs are formed by extracting a sentence from each of the bilingual texts and corresponding them together.

Anchor is a length of words that relatively easy to identify, and probably appears in the same position of each kind of language.

When aligning corpus, Brown proposed the conception of anchor sentence pairs, which he called the sentence pairs divided the whole text into aligned fragments.

Aligned sentence pairs is called the sentence bead.

According to above definitions, sentence bead has various kinds of formations, such as: (0:1), (1:0), (1:1), (1:2), (1:many), (2:1), (many:1).

Candidate anchor: the candidate anchors are the sentence pairs which possibly become the anchors. The (1:1)-type sentence beads are regarded as candidate anchor in this method.

The output result of sentence alignment is turning the bilingual text into the aligned fragment sequences which are the translation of each other. The statistics shows that there are plenty of (1:1) bilingual sentence pairs in the real texts. As research referred [7], after the alignment of complex font Chinese and English, the (1:1)-type sentence beads account for 89% of the whole paragraph. In this paper, we only consider the (1:1)-type sentence pairs.

3.1 Alignment Model based on Multiple Features

The analysis method for Tibetan syntactic analysis is the establishment of Tibetan dictionary and grammar rule base. The dictionary not only stores Tibetan vocabulary, but also the part of speech information, meanwhile the rule base stores Tibetan grammar rules.

The formalized description of sentence alignment research is: giving the bilingual corpus, find and formulize the greatest-probability alignment from all the possible ones.

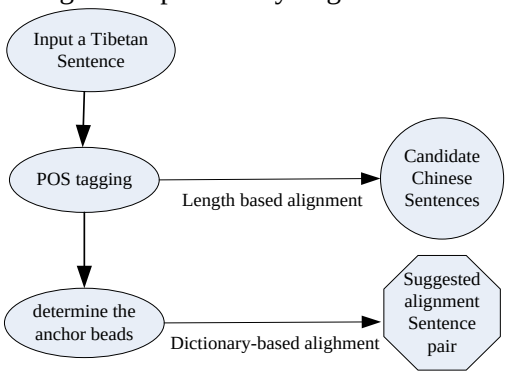

Figure 1. The Overview of the Alignment Method

There are three main characteristics as follows that occupied the core status of Tibetan sentences:

The word order of Tibetan sentences is "subject + object + verb" (SOV), this is a kind of predicate post position language. When Tibetan verb serves as predicate of the sentence, except the rich syntactic category information (person, tense, *etc*.), the semantic categories of verbs (verb attributes, possession, intention, *etc*.) are also included.

Words or phrases serves as subject, object, purpose and locations in Tibetan sentences, usually include grammatical markers.

Given a Tibetan sentence, calculate the number of the characters and select the longer Chinese candidate sentences firstly.

Secondly, we need determine the anchor beads based on similarity rules, which are manually created considering Tibetan and Chinese language characters.

We map the character on the right-hand side (Tibetan language) to one or more characters in the middle of the sentence in the middle of the sentence (Chinese language).

3.2 Sentence Alignment based on Dictionary

Get a bilingual key vocabulary list at first, make sure their corresponding relation by means of counting the vocabulary co-occurrence frequency from bilingual sentences. Taking the text of language *T* (Tibetan) and language *C* (Chinese) into account, the dictionary-based aligning method should be adopted.

Given the aligned bilingual corpus $D^{T\to C}$, the Tibetan sentence t_i consisting of *m* characters $\{x_{i,1}^T, x_{i,2}^T, \cdots, x_{i,m}^T\}$, the alignment method need to find the corresponding Chinese sentence c_j.

Score_of_match (t_i,c_j)=*number of matching words/maximum words*(t_i,c_j)

Our method works by first pairing each term extracted from the source language sentence *T* with related word extracted from the target language sentence *C*, by treating sentence alignment as a mapping problem.

Assume that *(X,Y)* is the corresponding key vocabulary list of language *T* and language *C*, *X* is the vocabulary of language *T* and *Y* is the vocabulary of language *C*. Assume that corresponding bilingual vocabulary list is: *M(X)→Y*. *X* is the vocabulary set

of the list included language *T*. *Y* is the vocabulary set of the list included language *C*. *M* represents mapping relation.

If the words $x_{i,m}^{T}$ in sentence t_i, and there is a word $y_{i,n}^{C}$according to the mapping relation $M(x_{i,m}^{T})\rightarrow y_{i,n}^{C}$.

For all the characters $\{x_{i,1}^{T}, x_{i,2}^{T}, \cdots, x_{i,m}^{T}\}$ in source sentence t_i, we use $d_i^{C} = \{y_{i,1}^{C}, y_{i,2}^{C}, \cdots, y_{i,n}^{C}\}$denotes the translations of $x_{i,m}^{T}$

We say a word $x_{i,m}^{T}$of t_i is translated by $y_{i,n}^{C}$ if either one of its translations in the dictionary *M* or it is a settled bead.

If the bilingual dictionary provides several translations for a source language word, we consider all of them but weight the different translations according to their frequency in the target language.

For example:

Tibetan:དེ་རིང་ཕྱག་རིང་འཛེན་past དོན་ཅི་ཡིན་ཞེས་འདྲི་past ཡ་ན་.

Chinese:今天这样叹气,到底是因为什么？

Word Alignment དེ་རིང་ཕྱག་རིང་འཛེན་དོན་ཅི་ཡིན་ཞེས་འདྲི་ཡ་ན་

今天力气长扯 PAST 事情什么（表询问）

The anchor mark is labeled

4. Experiments and Results

There are no open-sourced Tibetan sentence corpus, therefore, in our experiment we use the sentences from bilingual reading book as training data and the sentences extracted from bilingual website as test data.

We built a training corpus by gathering documents from school books and publications for the Tibetan-Chinese language pair. It contains 6135 Tibetan-Chinese sentence pairs. The average length of the Chinese sentences is 25.5 characters, and the average length of Tibetan sentences is comprised of 31.3 syllable points. The Tibetan-Chinese sentence corpora are collected from the text books and are manually aligned. Given the space constraint, detailed information about the development corpus is omitted here.

We pre-processed each document by performing monolingual term tagging using particular tools for Chinese and Tibetan respectively. For Chinese, we use the ICTCLAS from Chinese Academy of Science as POS-tagging tool [13] and Stanford CRF-based NER tagger as the monolingual component, which serves as a state-of-the-art monolingual baseline. For Tibetan POS-tagging, we use IEA-TWordSeg [14]. Next, we use grammar rules, in the form of sentence length and labeling words to identify candidate Chinese sentences. Then we apply a Tibetan-Chinese dictionary containing 150K entries with commonly used words.

We also built comparable corpora as test data by gathering bilingual documents from some particular government website, which always have Tibetan version and Chinese version. Since our aim is to a build a bilingual corpus, first, we manually chose Tibetan documents as seed document, which have an inter-language link to a Chinese document.

To test the performance of the alignment method, we evaluated it on a bilingual texts with almost 20K sentences. With these data and the collocations in section 4.1, we produced 10211 sentence pairs.

Experiment results indicate that our approach achieves 84.6% average precision and 72% recall respectively, which considerably outperform those method that only use dictionary or only use sentence length. If only lexical terms or length of sentences are adopted in the alignment process, the performance drops significantly. The accuracy of this alignment method is 95% for closed-corpora.

5. Conclusion and Future Work

Given the fact that we usually have large Chinese corpora (in some sense) and very limited bilingual corpora, especially minority languages, such as Tibetan, this paper proposes a method that tries to abstract exact Tibetan-Chinese sentences pairs from comparable corpora automatically.

A number of studies investigate the sentence alignment method, however, up to now, there have been few researches which directly address the problem of alignment of Tibetan-Chinese sentences pairs based on multi-features.

This paper proposes a multi-strategy Tibetan-Chinese sentence alignment method based on length of sentence, syntactic rules and bilingual dictionary. We test our approach on a bilingual corpus crawled from bilingual website and perform manual evaluation on bilingual sentences pairs extracted from Tibetan-Chinese corpora. The accuracy of this alignment method is 95% for closed -corpora.

The work of this paper is a part of our ongoing research work, which aims to provide a bilingual corpus with labeled for further bilingual named entity recognition, translation and other applications of Tibetan language.

Various experiments and applications have been conducting in our current research. Future work includes how to acquire and verify bilingual named entities from Tibetan and Chinese free text, how to obtain entity patterns automatically and how to acquire language features for entity recognition.

Acknowledgements

Our work is supported by the National nature science foundation of China(No. 61103161)and the Program for New Century Excellent Talents in University (NCET-12-0579).

References

[1] Brown, F. Peter, J. C. Lai and L. Robert, "Mercer. Aligning sentences in parallel corpora", Proceedings of the 29th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, **(1991)**.
[2] W. A. Gale and K. W. A, "Church program for aligning sentences in bilingual corpora. Computational linguistics", vol.19, no.1, **(1993)**, pp.75-102.
[3] Chen and F. Stanley, "Aligning sentences in bilingual corpora using lexical information", Proceedings of the 31st annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, **(1993)**.
[4] P. Fung, "A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora", Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup, **(1998)**.
[5] M. Kay and M. Röscheisen, "Text-translation alignment", Computational Linguistics - Special issue on using large corpora, vol.19, no.1, **(1993)**, pp.121-142.
[6] F. Smadja and K. McKeown, "Translating collocations for use in bilingual lexicons", Proceedings of the workshop on Human Language Technology, **(1994)**; Plainsboro, NJ.
[7] D. Wu, "Aligning a parallel English-Chinese corpus statistically with lexical criteria", Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, **(1994)**.
[8] T. Liu, M. Zhou, J. Gao, E. Xun and C. Huang, "PENS: a machine-aided english writing system for Chinese users", Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, **(2000)**; Hong Kong.
[9] Moore and C. Robert, "Fast and accurate sentence alignment of bilingual corpora", Springer Berlin Heidelberg, **(2002)**.
[10] X. Yu, J. Wu and W. Zhao, "Dictionary-based Chinese-Tibetan sentence alignment. 2010 International Conference on Intelligent Computing and Integrated Systems (ICISS)", **(2010)**; Guilin.
[11] A. Aker, M. Paramita and R. Gaizauskas, "Extracting bilingual terminologies from comparable corporaC", Annual Meeting of the Association for Computational Linguistics, **(2013)**.
[12] Y. Chen, C. Zong, and K.-y. Su, "A Joint Model to Identify and Align Bilingual Named Entities", Journal of Computational Linguistics, vol.39, no.2, **(2012)**, pp.229-266.

[13] ICTCLAS, **(2015)**, http://ictclas.nlpir.org/.
[14]. C. Long, Y. Lan and X. Zhao, "The analysis on mistaken segmentation of Tibetan words based on statistical method", 2014 International Conference on Asian Language Processing, **(2014)**.
[15] X. Yu, J. Wu and J. Hong, "Research and Application of Dictionary-based Chinese-Tibetan sentence alignment,M", Journal of Chinese Information Processing, vol.25, no.4, **(2011)**, pp.57-62.

Author

Lirong Qiu, she received her Ph.D. in Computer Sciences (2007) from Chinese Academy of Science. Now she is an associate professor of computer sciences at Information Engineering Department, Minzu University of China. Her current research interests include different aspects of natural language processing, artificial intelligence and distributed systems.

