Chinese-Uyghur Sentence Alignment: An Approach Based on Anchor Sentences

Samat Mamitimin
Xinjiang University
Urumqi 830046, China
Communication University of China/
Beijing 100024, China
tilchin@hotmail.com

Min Hou
Communication University of China
Beijing 100024, China

houminxx@263.net

Abstract

This paper, which builds on previous studies on sentence alignment, introduces a sentence alignment method in which some sentences are used as "anchors" and a two step procedure is applied. In the first step, some lexical information such as proper names, technical terms, numbers and punctuation marks, location information and length information are used to generate anchor sentences that satisfy some conditions. In the second step, texts are divided into several segments by using the anchor sentences as boundaries, and then the sentences in each segment are aligned by using a length-based approach. By applying this segmentation technique, the method avoids complex computation and error spreading. Experimental results show that the precision of the method is 94.6% on the average for Chinese-Uyghur sentence alignment for multi-domain texts.

1 Introduction

Parallel corpora are very useful for both theory-oriented linguistic research and application-oriented cross-language information processing. For parallel corpora, the most important annotation is alignment, especially sentence alignment, which is a minimal and essential requirement for the annotation of a parallel corpus. Aligning Chinese-Uyghur parallel texts at the sentence level, however, is already very difficult because of the considerable differences in the syntactic structures and writing systems of the two languages.

A number of alignment techniques have been proposed for other language pairs, varying from statistical methods to lexical methods. There are basically three kinds of approaches on sentence alignment: the length-based approach (Gale and Church, 1991), the lexical approach (Kay and Röscheisen, 1993), and the combination of the two (Chen, 1993 and Wu, 1994).

The first approach is based on modeling the relationship between the lengths of sentences that are mutual translations. Similar algorithms based on this idea were developed independently by Brown, et al (1991) and Gale and Church (1993). However, their main targets are rigid translations that are almost literal translations. The method is applicable for structurally similar European languages (i.e. English-French or English-German).

One alternative alignment method is the lexicon based approach that uses lexical information to obtain higher accuracy. Kay and Röscheisen (1993) proposed a relaxation method to sentence alignment using the word correspondences acquired during the alignment process. Chen (1993) developed a method based on optimizing word translation probabilities which he showed gave better accuracy than the sentence-length based approach. Wu (1994) used a version of Gale and Church's method adapted to Chinese along with lexical cues in the form of a small corpus-specific bilingual lexicon to improve alignment accuracy in text regions containing multiple sentences of similar length. Melamed (1996) also developed a method based on word correspondences, for which he reported sentence-alignment accuracy slightly better than Gale and Church. The method does not capture enough word correspondences for structurally different languages such as Chinese and Uyghur, mainly for the following two reasons. One is the difference in the character types of the two languages. Chinese uses Chinese characters as its writing system while Uyghur uses alphabetic character. The other is the grammatical difference of the two languages. Chinese is an analytic language that has SVO word order. In contrast, Uyghur is

Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009, pages 38–45,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

a suffixing and agglutinative language that has SOV word order. Thus, it is impossible in general to apply the simple-feature based methods to Chinese-Uyghur sentence alignment.

This paper, on the basis of other sentence alignment methods, introduces an anchor sentence based sentence alignment method, in which some sentences are used as “anchors” and two steps are applied. In the first step, some lexical information such as proper names, technical terms, numbers and punctuation marks, location information and length information are used to generate anchor sentences that satisfy some conditions. In the second step, texts are divided into several segments by using anchor sentences as boundaries, and then the sentences in each segment are aligned by using a length-based approach.

2 The Chinese-Uyghur Parallel Corpus

Uyghur is a Turkic language spoken by Uyghur people in Xinjiang Uyghur Autonomous Region of China and adjoining areas, which has about 9 million speakers. As one of the official languages in Xinjiang, Uyghur is widely used in many fields such as education, communication, publication, etc. Bilingualism in Xinjiang requires translation from Chinese to Uyghur or in the opposite direction. Therefore, it is possible and essential to build a Chinese-Uyghur parallel corpus for teaching and research in translation, bilingual lexicography, linguistics, and other NLP applications. Consequently, we began to build a Chinese-Uyghur parallel corpus for linguistic research, translation studies, teaching and applications such as machine translation. The corpus is a sentence aligned general corpus of medium size.

So far, over 1 million characters of Chinese texts, in total 263 texts, and their corresponding Uyghur texts have been collected from several sources and included into the raw corpus after sampling. The corpus texts cover a variety of styles, such as fiction, scientific texts, government documents, law texts, daily conversation and other texts. Presently, the size of the corpus is smaller than we expected because it is not easy to obtain such digital text data which also needs to be processed before it can be included in the corpus. The main sources of text data are published books, news papers, magazines and some web pages. The proportions of the different genres in the corpus are shown in Figure 1.

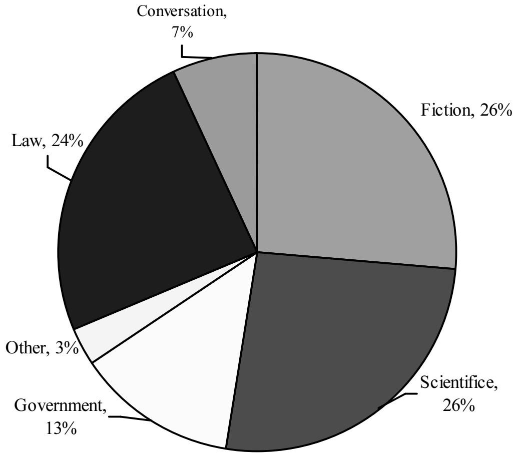

Figure 1. Genres and their percentages counted in tokens

3 System Overview

There is no previous work or approach specific to Chinese-Uyghur sentence alignment. So we firstly examined many papers related to the subject to find an appropriae method for Chinese-Uyghur sentence alignment. Most approaches share many common properties in the methods they use and suggest only small modifications to the earlier approaches. The length based method is suitable for aligning a very large bilingual corpus. Since it does not use any lexical information for the alignment task, it can be used between any pair of languages. However, in distant languages where characters differ, it is not so efficient. One alternative alignment method is the lexicon based approach that uses lexical information offering the potential for higher accuracy. However, it is not easy to capture enough word correspondences or cognates for Chinese and Uyghur. We may use bilingual dictionaries as an external resource to retrieve all possible word translations in such sentence alignment tasks. However, this is time-consuming and rather complex because word segmentation and lemmatization have to be done before the process of word matching can be started. Secondly, we tried some tentative methods to Chinese-Uyghur sentence alignment. According to the preliminary examination, it is generally not possible to apply the simple-feature based methods to Chinese-Uyghur sentence alignment.

Finally, we decided to apply a mixed approach to obtain better and more efficient results by combining the three criteria: length, lexical information and location information. Below are the detailed descriptions of this approach.

Our algorithm combines techniques adapted from previous work on sentence and word alignment. Our method is similar to Wu's (1994) in that it uses both sentence length and lexical information. But in our method, some lexical correspondences are used to find anchor sentences. Our method is similar to Simard's (1992) in that it uses cognates or anchors for sentence alignment. But in our method length information and anchors are used at different stages of sentence alignment. Our method is similar to Melamed's (1999) in that it uses a bitext mapping technique to locate anchor points, but it uses sentences as anchor points instead of words or characters. A segmentation technique that splits the text into several sections is also introduced to improve the length-based approach. As we can see from Figure 2, a two-step approach is applied to Chinese-Uyghur sentence alignment.

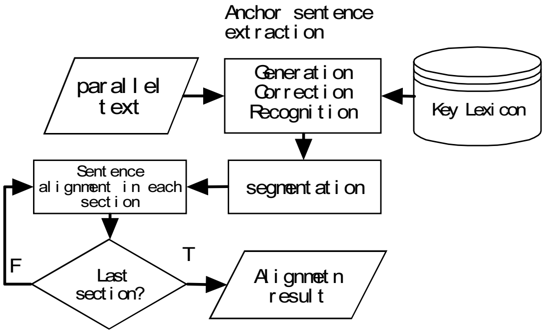
Anchor sentence extraction
parallel text
Generation Correction Recognition
Key Lexicon
Sentence alignment in each section
segmentation
F
T
Last section?
Alignmetn result

Figure 2: Flowchart of Chinese-Uyghur sentence alignment

In the first step, some (1:1) sentence pairs, called anchor sentences, are extracted by using lexical information, location and length information. A three-phase method is applied to anchor sentence extraction which will be explained in the following section.

In the second step, texts are divided into several segments by using these sentences as anchors, and then all sentences in each segment are aligned by using a length-based approach.

4 Anchor Sentence Extraction Algorithm

4.1 Anchor Sentence

Brown (1991) firstly introduced the concept of alignment anchors when he aligned the Hansard corpus. In our method, we also introduced this concept, which in our case are anchor sentences. In a parallel corpus, the anchor sentences are specific (1:1) sentence pairs that are strongly related and that satisfy some conditions. All such sentence pairs which were extracted from bilingual texts during the first step are seen as anchor sentences. These anchors divide the whole texts into short aligned segment. The goal of anchor sentence extraction is to divide the source text and the target text into one-to-one smaller segments. And using this segmentation, we attempt to improve the sentence alignments produced by the length based alignment. Sentence alignment tends to be better with shorter segments and, consequently, better sentence alignments are obtained.

For anchor sentence extraction, we applied a bitext mapping technique. A bitext map is a set of pairs (*x*, *y*), where *x* and *y* refer to precise locations in the first and second texts respectively, with the intention of denoting portions of the texts that correspond to one another (Simard, 1998). However, we used a bitext map of sentence pairs instead of words or characters to point out the correspondences between these anchor sentences (See Figure 3).

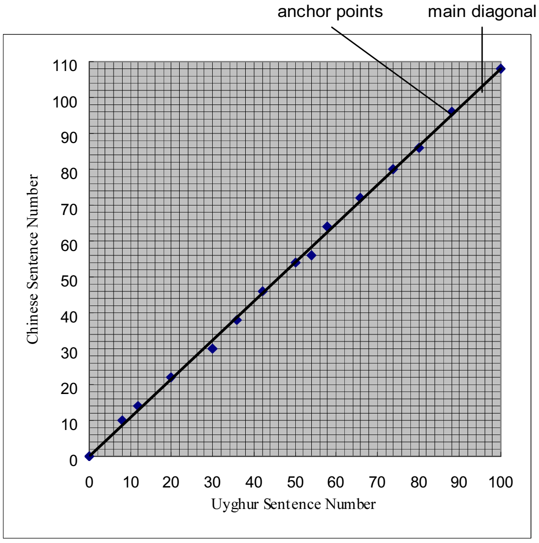

Figure 3. Bitext map of sentence alignments

The horizontal axis denotes the sentence number in the Uyghur text, and the vertical axis denotes the sentence number in the Chinese text. The anchor sentences, which are shown as anchor points in the bitext map, can be characterized by three properties:

Injectivity: no two anchor points in a bitext map can have the same x or y coordinates.

Linearity: anchor points tend to line up straight. In other words, all anchor points are to appear around a straight line.

Low variance of slope: The slope of the anchor points is rarely much different from the bitext slope.

4.2 Algorithm Description

In our anchor sentence extraction algorithm, a three-step process is applied to extract anchor sentences. In other words, the search for each anchor sentence pairs alternates between the following three steps: generation phase, correction phase and recognition phase.

- Generation phase

In the generation phase, the algorithm generates candidate anchor sentence points within a search rectangle. We define a search rectangle as follows: Rectangle(x, y, x+3, y+3) in which x=last anchor point(x) and y=last anchor point(y).

The first search rectangle is anchored at the origin of the bitext map where x=0, y=0. Subsequent search rectangles are anchored at the previously found points.

In this step, the search for an anchor sentence begins in a small search rectangle in the bitext map, whose diagonal is parallel to the main diagonal. If no candidate points are found, the search rectangle is proportionally expanded by the minimum possible amount, and the generation cycle is repeated. The rectangle keeps expanding until at least one acceptable point is found. Three kinds of information such as sentence length, location information and lexical information are used to generate anchor points. Sentence pairs that satisfy the following three conditions are added to the candidate anchor sentence array.

(1) Sentence length ratio

As was shown in the sentence alignment literature (Church, 1993), the sentence length ratio is also a very good indication of the alignment of a sentence pair.

In our method, for sentence pair P(c,u), if LenRatio(c,u)∈[MinLenRatio, MaxLenRatio], sentence pair P(c,u) would be candidate anchor sentences, in which LenRatio(c,u)= L_c/L_u (L_u is Uyghur sentence length, L_c is Chinese sentence length).

MinLenRatio and MaxLenRatio are calculated by using following formula:

$MaxLenRatio=C'+A/(L_c+B)$

$MinLenRatio= C'-A (L_c+B)$

$C'=(C+ Len(C)/Len(U))/2$

The constant C is the expected number of Chinese characters per Uyghur word. C′ is the weighted value when taking text size into account, the values of the constants are A=10， B=14.

(2) Matching score

If the matching score of a sentence pair is above the threshold (we set the threshold = 1.1), it is considered a candidate anchor sentence. By applying this condition, we reject some sentence pairs with a matching score smaller than the threshold. The matching score is calculated according to the matching degree of the key lexicon and punctuations as described in section 4.3.

(3) Maximum Angle Deviation (MAD)

According to the properties of the anchor sentences, the slope of the anchor points should not be much different from the bitext slope. So some sentence pairs are rejected by setting a maximum angle deviation. The angle of each anchor point's least-squares line is compared to the arc tangent of the bitext slope. The anchor point is rejected if the difference exceeds the maximum angle deviation threshold (MAD=3). The angle between the least-squares line and the bitext slope is calculated according to the following formula:

$$\theta = \arctan(\frac{|A-B|}{1+A*B})$$

In this formula, A is the slope of the least-squares line, B is the bitext map slope.

This filtering process generates anchor sentences with higher accuracy; however, it causes errors in some cases. So, we introduced another correction phase in order to reject some wrongly aligned sentence pairs.

- Correction phase

In this step, some candidate sentences that are no anchor sentences are eliminated according to characteristics of anchor sentences, namely the length ratios of corresponding segments.

First, the algorithm checks if there are any conflicts between anchor points. The injective property of anchor sentences implies that whenever two anchor points overlap in the x or y axis, but are not identical in the region of overlap, then one of the points must be wrong. To resolve such conflicts, we employed a lookup method to eliminate conflicting points.

Secondly, length ratios of corresponding segments divided by candidate anchor sentences are calculated according to a similar formula as used for the sentence length ratio in order to reject wrongly aligned anchor points.

If the length ratio of the segments LenRatio(c,u)∈[MinLenRatio, MaxLenRatio], the candidate anchor sentence must be an anchor sentence, otherwise it should be eliminated. MinLenRatio and MaxLenRatio are calculated by using the following formulae:

$\text{MaxLenRatio}=C'+A/(\ L_c+B)$

$\text{MinLenRatio}=\ C'-A\ (L_c+B)$

- Recognition phase

A number of candidate anchor sentences can be obtained in a certain search region during application of the above two steps. For anchor sentence alignment, accuracy is more important than recall rate. So it is essential to introduce a recognition step in order to achieve higher accuracy by eliminating some unlikely anchor sentences. In the recognition step, one best anchor sentence pair is selected from candidate anchor sentences according to two parameters: matching score and length similarity score. The anchor selection algorithm gives a score to each proposed sentence pair during the recognition phase, and finds the alignment with the largest sum of scores. A parameter estimation method is described in the following section.

4.3 Parameter Estimation

Matching score: As previous work suggests, lexical information is critical for sentence alignment, especially for finding anchor points. It is well-known that some proper names and technical terms have rigid translations in many languages; numbers and punctuations appear in the same or similar forms in both source text and translation text. In a parallel text, for instance, if a sentence contains a question mark, it is likely to be aligned to a sentence that also contains this mark, which can be a strong clue for sentence alignment. This is also true for Chinese-Uyghur translations.

However, in our method, lexical and non-lexical clues are not used to align all sentences, but to estimate matching scores and to find the best anchor sentences. We used multiple clues such as proper names, technical terms, punctuation marks and numbers.

In most cases, proper names, including person names, location names, organization names, and technical terms have unique translations that will be matched easily. But, the problem is that person names and technical terms are often unknown words. How to identify them is a difficult problem. In our case, we first collected some popular proper names and the most frequent technical terms into a small lexicon that we call the key lexicon. More than 2000 words are included in the key lexicon at present. Then, a very simple searching method is applied to match corresponding words.

In addition, punctuation marks, including other symbols (e.g. @#$%&), are the most obvious clues in Chinese and Uyghur translation. The correlation between Chinese and Uyghur punctuations is extremely high as depicted in Table1.

Punctuation	Chinese	Uyghur
full stop	。	.
question mark	?	؟
exclamation mark	!	!
comma	，	،
ideographic comma	、	،
semicolon	；	؛
colon	：	:
quotation mark	“” ‘’	“” «» ‘’
bracket	（） []	() []
Title mark	《》	«»

Table 1. Corresponding punctuation marks in Chinese and Uyghur

For punctuation and numbers, no external resources but some rules are applied to estimate the matching degree of these clues.

The matching scores are calculated according to the average number of matched clues. In other words, the more matched proper names, technical terms, punctuation and numbers, the higher the matching score.

Length Similarity: Length similarity is a score that reflects the similarity between the length ratio of the current sentence pair and the expected length ratio. The following formula will be applied to calculate the length similarity of a proposed sentence pair (A_iC, A_iU):

$\text{LenSimilar}(A_iC,\ A_iU\ )=|\text{Len}(A_iC)/\text{Len}(A_iU)-C|/C$

Hereby C is expected number of Chinese character per Uyghur words. We obtain C=2.01 experimentally. Len(A_iC) and Len(A_iU) are the sentence lengths of A_iC, A_iU, respectively.

However, the sentence length ratio is not stable when a Chinese sentence is shorter than 10 characters. So it is necessary to add a weighting factor WF:

$\text{LenSimilar}(A_iC,\ A_iU\ )=|\text{Len}(A_iC)/\text{Len}(A_iU)-C|/C\ {*}\text{WF}$

if Len(A_iC)<= StableLen, then

WF =a*Len(A_iC)/StableLen, else WF =1.

Hereby StableLen=10, a=0.5

The length similarity formula is also adjusted as follows:

LenSimilar(A_iC,A_iU)=|Len(A_iC)/Len(A_iU)-C′|/C′* WF

Hereby C′ is the value weighted by the whole text size.

5 Length Based Sentence Alignment

According to previous work by Gale and Church, length-based approaches are simple and can achieve good performance for different language pairs. Because of this simplicity, many later researchers integrated this method to their sentence alignment methods. We also applied the length-based approach to the second step of sentence alignment.

5.1 Measuring Length in Words and Characters

Different length measuring methods can be used in the length-based approach. Brown (1991) introduced the length-based algorithm based on the number of *words* in sentences, Gail and Church's algorithm is similar to the Brown's algorithm except that alignment is based on the number of *characters* in the sentences.

Uyghur is an alphabetic language while Chinese is a non-alphabetic language. Therefore, it is a difficult problem to select the best length measuring model. In general, a Chinese sentence does not have word boundary information; so one way to define Chinese sentence length is to count the number of characters in a sentence. Another way is to count how many words are in a sentence after word segmentation. For Uyghur sentences, we can similarly define the length in characters or in words.

In our case, we examined three possible length models described in the following Table 2:

L-1	Both Uyghur and Chinese sentences are measured in *characters*
L-2	Both Uyghur and Chinese sentences are measured in *words*[1]
L-3	An Uyghur sentence is measured in *words* and a Chinese sentence is measured in *characters*

Table 2. Three length models

The mean sentence length ratios, variances and correlation coefficients for each of the length models are calculated from hand aligned Chinese-Uyghur texts of 988 sentence pairs. Statistics of the three sentence length models are shown in Table 3.

	L-1	L-2	L-3
Mean	3.99	1.07	2.01
Var	0.71	0.23	0.21
Correl	0.976	0.953	0.977

Table 3. Statistics of different length measuring methods

In general, the smaller the variance, the better the sentence length model should be. From Table 3, we can see that the character based length ratio model has significantly larger variance (0.71) than the other two models (L-2:0.23, L-3: 0.21). This means L-1 is not as reliable as L-2 and L-3. Both L-2 and L-3 have similar variance, but L-3 is better than L-2 with regard to the correlation coefficient, which indicates that sentence lengths have higher correlation if the lengths of Chinese and Uyghur texts are measured in characters and words, respectively. A regression analysis of the three models also proved this result. So we applied the L-3 model to the length ratio examination and length based sentence alignment.

5.2 Preliminary Statistics for the Length-based Method

A length-based sentence alignment program is based on a very simple statistical model of sentence lengths. The model makes use of the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. A probabilistic score is assigned to each pair of proposed sentence pairs, based on the ratio of lengths of the two sentences and the variance of this ratio. This probabilistic score is used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences.

The parameters C and S^2 are used for likelihood estimation. C is the expected number of Chinese characters per Uyghur words. The parameters C and S^2 are determined empirically from a hand aligned parallel corpus of multi-domain texts. According to our statistical results, we obtained $C=2.01$ and $S^2 =3.24$.

Brown (1991) assume that every parallel corpus can be aligned in terms of a sequence of minimal alignment segments, which they call "beads", in which sentences align 1-to-1, 1-to-2, 2-to-1, 2-to-2, 1-to-0, or 0-to-1. The alignment model is a generative probabilistic model for

[1] Bbibst software is used for Chinese word segmentation.

predicting the lengths of the sentences composing sequences of such beads. The model assumes that each bead in the sequence is generated according to a fixed probability distribution over bead types. We also calculated the probability of different alignment types.

Type	Frequency	Percentage（%）
1:1	807	81.3
1:0 or 0:1	5	0.5
1:2 or 2:1	152	15.3
2:2	7	0.7
1:3 or 3:1	20	2.0
other	2	0.2
Total	993	100

Table 4. Proportion of alignment types

From the above statistical results, it is clear that the correlation between the length of a Chinese sentence in characters and the length of its Uyghur translation sentence in words is extremely high. This high correlation suggests that length might be a strong clue for sentence alignment.

In our cases, we applied the length-based approach suggested by Gale and Church after some parameters had been changed.

6 Experimental Results

In this section, we report the results of experiments on aligning sentences by using two methods.

6.1 Test Corpus

In our experiment, we selected ten texts as our testing corpus. The texts are varied in length and genres as summarized in Table 2. T1, T2 and T3 are fiction texts; T4 is a law text; T5 and T6 are official documents; T7 and T8 are scientific texts, T9 and T10 are news and other articles. The total size of the corpus is 72,000 tokens, about 1300 sentence pairs.

6.2 Results

Firstly, we aligned sentences by using two approaches: a length-based algorithm, and an anchor sentence based algorithm. Then we manually checked the alignment results for errors and calculated precision and recall scores. Experimental results show that our anchor sentence based approach yields higher accuracy than the purely length based approach. The precision of the method is 94.6% on the average for Chinese-Uyghur sentence alignment on multi-domain texts. This is 2% higher than that of a purely length based approach.

	length-based		anchor sentence based	
	Precision	recall	precision	recall
T1	89.9	89.3	94.2	93.6
T2	94.9	94.9	97.5	97.5
T3	83.1	84.5	86.4	87.9
T4	100	100	100	100
T5	100	100	100	100
T6	98.8	98.8	100	100
T7	98.5	98.9	98.5	98.9
T8	65	66.7	72.5	74.4
T9	89.1	86.0	96.4	93.0
T10	96.8	95.8	94.7	93.8
average	92.7	92.8	94.6	94.8

Table 5. Experimental results

As we can see from Table 5, the error rates of the two methods vary from text to text. We analyzed all errors during sentence alignment in order to find reasons and solutions. The following is an error analysis.

6.3 Error Analysis

Firstly, the style of a text affects the sentence alignment results. In law texts and official documents, precision is very high in comparison with the results in texts of other styles; even 100% accuracy has been achieved. The reason for this may be the language style of source texts and translated texts. The error rate is comparatively higher in fiction texts because of their free translation style.

Secondly, complex sentence beads that include deletion and insertion during translation affect the alignment accuracy. According to Table 7, complex alignment types that the current alignment algorithm did not take into consideration account for 2.2% of the errors in Chinese-Uyghur translations. So errors caused by these "unorthodox" translation patterns are unavoidable. There are many such errors in sample T8. By examination, we found that the number of sentences in the Chinese text (122 sentences) and corresponding Uyghur text (179 sentences) is so unbalanced that many complex alignment types are involved. This is a direct reason for the high error rate.

Finally, anchor sentences play an important role during alignment. However, we found that it leads to more mistakes once wrong anchor sentence are selected. For instance, in T9, just one wrong anchor sentence caused up to four errors during second-step sentence alignment. So it is crucial to align anchor sentences correctly.

7 Conclusions

We have developed a very effective sentence alignment method based on anchor sentences. In our method, firstly anchor sentences are extracted from bilingual texts according to key lexical information, location information and length information; secondly, whole texts are divided into small segments by using anchor sentence points; finally, sentences in each small segment are aligned by using a length-based approach. We have implemented the proposed method on the parallel Chinese-Uyghur corpus. Experimental results show that the precision rate of the method is 2% higher than that of a purely length-base approach. Differences and advantages of our anchor sentence based method are compared to other methods in Table 6.

Methods	Length based	Lexical based	Our method	Advantages
Length information	Yes	No	Yes	Quick
Lexical information	No	Yes	Yes	Higher accuracy
Language resource	No	Dictionary	Simple lexicon	Simple
Special character	No	No	Yes	Higher accuracy
Multi level	No	No	Yes	Avoids error spreading
For multi-domain	Good	Not good	Good	Applicable to different texts

Table 6. Differences of three alignment methods

References

Brown, Peter, J. Lai and R. Mercer. 1991. Aligning Sentences in Parallel Corpora. *in Proceedings of ACL-91*, 169-176.

Brown, P.F., Della Pietra, S. A., Della Pietra, V. J., Mercer, R.L. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics,* 19(2): 263–311.

Chen, S.F. 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. *In Proceedings of ACL-91.*

Chuang, Thomas and Kevin C. Yeh. 2005. Aligning Parallel Bilingual Corpora Statistically with Punctuation Criteria. *International Journal of Computational Linguistics and Chinese Language Processing ,* Vol. 10, No. 1.

Gale, William A. and Kenneth W.Church. 1991. A Program for Aligning Sentences in Bilingual Corpora . *Proceedings of ACL-91*, 177-184.

Fung, Pascale and Kenneth W. Church. 1994. K-vec: A new approach for aligning parallel texts. *In Proceedings of the 5th International Conference on Computational Linguistics*, 1096-1102, Kyoto, Japan.

Kay, M., Röscheisen, M. 1993. Text-Translation Alignment. *Computational Linguistics*, 19(1): 121-142.

Melamed, I. D., Bitext Maps and Alignment via Pattern Recognition, *Computational Linguistics,* 25(1), 107-130, March, 1999.

Melamed, I.D. 1996. A Geometric Approach to Mapping Bitext Correspondence. IRCS Technical Report, 96-22, University of Pennsylvania.

Melamed, I.D. 1997. A Portable Algorithm for Mapping Bitext Correspondence. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics,* Madrid, Spain, 305-312.

Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. *In Machine Translation: From Research to Real Users* (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, 135-244

Simard, M., Foster, G., and Isabelle, P. 1992. Using Cognates to Align Sentences in Bilingual Corpora. *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, Montreal, Canada.

Simard, M., Plamondon, P.1998. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Machine Translation,* 13(1), 59–80.

Weigang Li, Ting Liu, Zhen Wang and Sheng Li. 1994. Aligning Bilingual Corpora Using Sentences Location Information, *Proceedings of 3rd ACL SIGHAN Workshop,* 141-147.

Wu, D. 1994. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces,* New Mexico, 80-87.