Chinese-Uyghur Parallel Corpus Construction and its Application

[1]Samat Mamitimin, [2]Umar Dawut
[1] Communication University of China，Beijing，100024；
[1、2] Humanities School of Xinjiang University，Urumqi，830046

Abstract

In this paper, our work towards the building the Chinese-Uyghur parallel corpus is presented. We elaborate on the design of the corpus, collection, annotation and mark-up of the parallel texts, and sentence aligning process to construct the corpus. In addition, we also introduce the corpus building tools for constructing and using the corpus easily for different purposes. Some preliminary study results and several prospective applications based on the Chinese-Uyghur parallel corpus are also proposed.

1 Introduction

It is well known that multilingual resources are very important to both theory-oriented linguistic researches and application-oriented cross-language information processing. Until now, many corpora have been built to different application. One of the most well-known and frequently used parallel corpora is Europarl (Koehn, 2002) which is a collection of material including 11 European languages taken from the proceedings of the European Parliament. Another parallel corpus is the JRC-Acquis Multilingual Parallel Corpus (Steinberger et al., 2006). It is the largest existing parallel corpus of today concerning both its size and the number of languages covered. The OPUS corpus (Tiedemann and Nygaard, 2004) and the English-Norwegian Parallel Corpus (Stig Johansson, 1994) are also very famous parallel language resource. Chinese-Uyghur parallel corpus also has very important application in cross-language information processing, Chinese-Uyghur bilingual

lexicography, Chinese-Uyghur comparative study, translation study and language teaching. But so far large-scale and balanced Chinese-Uyghur corpus is still unavailable yet, given the difficulties of collecting bilingual translated texts and the intensive labors required. Recently the project, “Construction and Application of Chinese-Uyghur Parallel Corpus”, has achieved some preliminary results in the field.

In this paper, our work towards the building the Chinese-Uyghur parallel corpus is presented. We elaborate on the design of the corpus, collection, annotation and mark-up of the parallel texts and sentence aligning process to construct the corpus. In addition, we also present our work toward building tools for constructing and using the corpus easily for different purposes. Some preliminary study results and several prospective applications based on the Chinese-Uyghur parallel corpus are also proposed. The remains of this paper will be organized as follows: in the second part, we introduce the aim, source data collection and construction workflow of the corpus; in the third part, corpus annotation process, including preprocessing, markup and sentence alignment, is presented in detail. In the forth part, we pointed out some preliminary results and potential value of the corpus.

2 The Chinese-Uyghur parallel corpus

The aim of the project, “Chinese-Uyghur Parallel Corpus Construction”, is to build a representative language resource for Chinese and Uyghur in order to be able to study the relations between these languages. More specifically, the goal is to build and annotate a sentence level aligned Chinese-Uyghur parallel corpus by using a set of tools. The parallel corpus is intended to be used in linguistic research, teaching and applications such as machine translation.

Before we present the corpus data, we give a short overview of the involved languages as they belong to different language types.

2.1 A Note on Chinese and Uyghur

Chinese belongs to the Han-Tibetan language family. It is the most commonly used language in China, and one of the most commonly used languages in the world. Modern Chinese is an analytic language, functions such as number in nouns or tense in verbs are expressed through syntax (word order and sentence structure) rather than morphology. One key feature is that all words in Chinese have only one grammatical form, as the language lacks declension, or any other inflection (there are minor exceptions). Chinese features subject verb object (SVO) word order similar to English. Uyghur is a Turkic language of Altaic family, spoken by the Uyghur people in Xinjiang Uyghur Autonomous Region of China. Uyghur is a suffixing and agglutinative language; in most of the cases, there is a one to one relationship between morpheme and function. The verbal system is rich and verbs have markers for tense, mood, aspect, and voice, as well as agreement markers in terms of the features person and number. Considering the syntactic characteristics, Uyghur is a left-branching type of language, where the dependents precede their head, for example adjective or genitive modifier precedes the modified head, and objects precede the verb. Uyghur is rather free in its word order which is based on the morphological structure. Uyghur has subject object verb (SVO) word order but other orders are possible depending on which element is put into the focus in the discourse. Modern Uyghur uses Arabic script as its writing system.

2.2 Corpus data collection

It is now a well recognized fact that a corpus is more than just a collection of electronic texts. Corpus data have to be selected with care with respect to the intended applications. Which means a corpus shall contain texts of different domain and different genres in reasonable proportions; the corpus thus can be a reasonable reflection of the language use. In this project we emphasize quality

with regard to content and translation. We focus on a collection of written texts to build a balanced corpus of the source and target language. However, when we decided to construct the Chinese-Uyghur corpus, we found it is not easy to construct a perfectly balanced Chinese-Uyghur corpus. That is because there are not so many electronic Chinese-Uyghur bilingual texts available. So, in the first step we decide to collect bilingual texts as many as we can, as long as the texts are of good quality. Bilingual texts in electronic format are collected from several resources such as books, newspapers, journals, and internet. Some texts that could not be obtained in electronic format were scanned, OCRed and reviewed. After one year efforts, there are totally 3 million words untagged Chinese-Uyghur parallel texts in hand.

In the second step, the texts have been normalized in their form (text-only), size and field in order to keep the balance in the corpus collection. Here, the balance means the weighting among the different sections in a general corpus. Obviously, it does not mean to have equal amounts of texts from different domains that are covered by the corpus. After sampling and normalization, the corpus texts cover a variety of domains, such as newspaper news, technical articles, government documents, law documents, daily conversation and fiction. (See table 1).

Genres	News	Technical articles	Government documents	Law documents	Daily conversation	Fiction	Total
Chinese character	120900	181350	145080	217620	157170	386880	1209000
Percent	10	15	12	18	13	32	100

Table 1 the genre and their percentage in Chinese-Uyghur parallel texts

So far, over 1 million characters of Chinese texts and their corresponding Uyghur texts have been collected and included into the raw corpus after sampling.

2.3 The workflow of the parallel corpus construction

To facilitate the construction of the parallel corpus, we also developed a systematic workflow based on our examination of the whole process of the corpus construction.

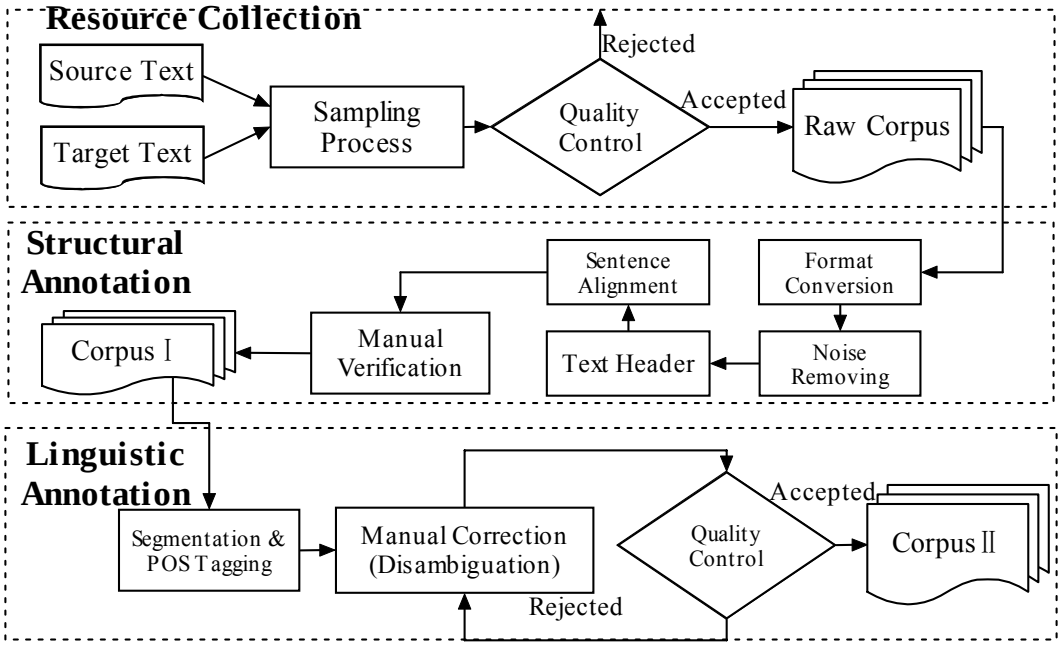

Figure1: The workflow of the parallel corpus construction

According to the workflow, any text must be firstly processed in the following steps after they are collected and before they included into the parallel corpus.

1) Preprocessing. This process includes format conversion and noise removing.

2) Textual attribute tagging. In this step, global textual attributes are tagged in the text.

3) Parallel Alignment. Parallel alignment at paragraph and sentence level is done in this step with the alignment tools.

4) Human verification of the alignment result. All the alignment results are verified and errors are corrected by human in this step.

5) Segmentation and POS tagging.

3. Annotation of the parallel corpus

The following steps below give an overview of the annotation procedure and involved tools.

3.1 Preprocessing

We start the annotation by cleaning up the original material that we received from publishers and Internet. This means that the various formats, for example rtf, doc, and pdf, are converted to plain text files. In some cases, we scanned and proofread the material and, where necessary, corrected it to ensure that the plain text file is complete and correct. In this step, all irrelevant links, tags are removed from the texts. Then, the texts are encoded according to international standards by using UTF-8 (Unicode) format and resaved using unique file name to indicate a pair of parallel files. Next, the corpus architecture and tools used to build the corpus is presented in more detail.

3.2 The markup of the parallel corpus

Parallel corpus could only be useful after it is annotated. In our study, to make the corpus application-independent and easier to exchange via the Internet, all parts of the corpus is clearly marked and annotated uniformly. For this reason, the international XML Corpus Encoding Standard (XCES) for the annotation format has been adopted and an XML-based framework, very similar to that of CHANG Baobao (2004)**,** has been designed (see Table 2 for detail).

		Tags	Attribute
Text		<TEXT>…</TEXT>	
Text Header	Text head	<TEXT_HEAD>…</TEXT_HEAD>	
	Chinese Title	<CH_TITLE>…</CH_TITLE>	
	Uyghur Title	<UY_TITLE>…</ UY _TITLE>	
	Chinese Subtitle	<CH_TITLE>…</CH_TITLE>	
	Uyghur Subtitle	<UY_TITLE>…</UY _TITLE>	
	Author	<AUTHOR>…</AUTHOR>	
	Translator	<TRANSLATOR>…</TRANSLAT OR>	
	Style	<STYLE>…</STYLE>	
	Field	<FIELD>…</FIELD>	
	Mode	<MODE>…</MODE>	
	Time	<TIME>…</TIME>	
	Source	<SOURCE>…</SOURCE >	
	Edition	<EDITION>…</SOURCE >	
	Translation direction	<DIRECTION>…</DIRECTION >	

Text Body	Text body	<TEXT_BODY>…</TEXT_BODY>	
	Paragraph	…	id
	Sentence	<s>…</s>	id
	Sentence alignment unit	<a>…	id, no
	Chinese Title	<CH_TITLE>…</CH_TITLE>	
	Uyghur Title	<UY_TITLE>…</ UY_TITLE>	
	Author	<AUTHOR>…</AUTHOR>	
	Translator	<TRANSLATOR>…</TRANSLATOR>	
	Time	<TIME>…</TIME>	
	Subtitle	<SUBTITLE>…</SUBTITLE>	
	Background information	<Background>…</Background>	
	Word	<w>…</w>	id, pos、lemma

Table 2: XML tags for Chinese-Uyghur Parallel Corpus

According to this framework, all the Chinese texts and Uyghur texts are encoded separately, each text, no matter what language it is, is composed of a text head and a text body. All the global textual attributes are put into the text head; the monolingual structural tags, linguistic information tags and the text itself are put within the text body. Alignments are indicated by an alignment attribute in the text body of the both languages. Structure of the each text is as following:

```
<?xml version=”1.0” encoding=”Unicode” ?>
<TEXT>
<TEXT_HEAD>
Text header
</TEXT_HEAD>
<TEXT_BODY>
Text body
</TEXT_BODY>
</TEXT>
```

Considering of the reliability of the corpus tools and possible use of the corpus, we decided to carry out the following three types of annotation:

1) Global textual attributes (text header). Global textual attributes are attributes applied to every full text in the corpus. They are features to specify the domain of the texts, whether a text is written or spoken, the author of a text, the translator of a text, the time when a text was authored and translated, the title of a text and so on. The global textual attributes will facilitate special research

based on the corpus, for example, language researchers might be interested only with texts belonging to a particular domain, and they can easily extract all texts belonging to that domain.

2) Monolingual textual structural annotation. Monolingual textual structural annotation deals with text unit of different levels. At present, boundaries of paragraph, sentence alignment units, and sentence have been annotated in the corpus.

3) Parallel alignment annotation. Parallel alignment annotation establishes the correspondence between the language units of the original texts and their translations. So far, the corpus is aligned only at the sentence level. Word alignment of the corpus seems still unpractical for the massive labor required and lacking of reliable tools.

4) Linguistic annotation. After the structural annotation, the Corpus is annotated linguistically for other purposes. Linguistic annotation actually covers any descriptive or analytic notations applied to raw language data. The added notations may include transcriptions of all sorts (from phonetic features to discourse structures), part-of-speech and sense tagging, syntactic analysis, "named entity" identification, co-reference annotation, and so on. However, linguistic annotation just includes word boundary detection and POS (part-of-speech) tagging in our study. Detection of word boundaries of Chinese texts, also known as word segmentation, is a very basic process in Chinese corpora building. For the Uyghur words, the detection of word boundaries is the word tokenization or word stemming that is a process for reducing inflected (or sometimes derived) words to their stem, base or root form. External morphological analyzers and part-of-speech taggers are used for the specific languages to the linguistic annotation.

3.3 Sentence alignment

For a parallel corpus, the most important annotation will be alignment, especially sentence alignment, which will be a minimal and essential requirement for a parallel corpus. Aligning

Chinese-Uyghur parallel texts, however, is already very difficult because of the great 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 & Church 1991), the lexical approach (key & Roscheisen 1993), and the combination of them (Chen 1993 and Wu 1994). Chen (1993) combines the length-based approach and lexicon-based approach together. His method is robust, fast enough to be practical and more accurate than previous methods.

In our project, the method we adopted is that of Chen (1993) and Simard (1992). Because the method considers both length similarity and cognateness as alignment criteria, the method is more robust and better able to deal with the text of very different language than pure length-based methods. However, in the case of Chinese-Uyghur alignment, where there are few cognates shared by the two languages, proper names, punctuations and numerals in both texts are taken as cognates or as anchors.

All sentences in parallel texts are aligned automatically by using newly developed tool, a Chinese-Uyghur sentence aligner, and then checked semi-automatically with the help of a sentence checker for alignment errors. Sentence alignment methods of our project will be presented soon in another paper.

3.4 The parallel corpus tool set

The corpus material is processed semi-automatically by using various tools making the annotation, alignment, and manual correction easy and straightforward for users with less computer skills. To facilitate the construction of the Chinese-English parallel corpus we have used a set of corpus tools developed by our team or others. So far we have the following tools in use: (1) The corpus builder:

It is an adopted text file editor like Ultra Edit, but has very strong application in corpus building such as text editing, xml encoding, and text indexing etc. (2) The Chinese-Uyghur sentence alignment program. The tool had been developed by us for the parallel corpus construction and for future application. (3) The Chinese segmentation and POS tagging program. It is software developed at Peking University. All Chinese texts are segmented and tagged with POS tags at the same time by the tool. (4) Uyghur morphological analyzer. The analyzer, developed by the researchers of Xinjiang University, has the function of word tokenization (stemming) and POS tagging. The four tools have been heavily used in the construction of the Chinese-Uyghur parallel corpus.

4 Preliminary result of the corpus application

At the moment of writing, only one year after its beginning, our project has not progressed far enough for us to carry out any major corpus-based studies; we can only present some very preliminary findings and some potential applications of the corpus at this point. These should, however, be of some interest in showing the sorts of analyses which can be carried out using a parallel corpus.

4.1 Some basic statistics

The statistics below are based on the texts translated from Chinese into Uyghur which amount to one million Chinese characters and corresponding over 500 thousand Uyghur words in the corpus. The texts are includes fiction and nonfiction texts such as news, government report, technical articles, law documents and daily conversation texts. Tables 2 summarizes the relationship between the number of paragraphs, S-units (orthographic sentences), words and mean sentence length in the original and translated texts of different genres.

	Field	Paragraph	Sentence	Character (word)	Mean Sentence Length (Chinese)	Mean Sentence Length (Uyghur)
1	Science	1.00	0.91	1.92	34.78	16.41
2	News	0.96	0.97	1.82	47.47	25.31
3	Law	1.01	0.99	1.85	40.55	21.77
4	Governm ent report	0.99	0.96	1.77	46.45	25.21
5	Fiction	0.96	0.93	1.98	24.86	11.63
6	Daily Conversa tion	1.00	1.00	2.15	10.63	4.95
Mean		0.99	0.95	1.91	29.03	14.44

Table 2: Paragraphs, sentences, words (characters), and sentence length in the Chinese and Uyghur parallel texts

As we can see, first three columns of the table are ratio between number of the paragraph, sentence, and words in Chinese texts and Uyghur texts respectively; it is not difficult to find that Uyghur translation of news and fiction uses fewer paragraphs than original Chinese text. There is an overall tendency for the Uyghur translated texts to contain fewer sentence, 100 sentences are translated as 95 sentences in Uyghur, than Chinese counterpart, it is especially true for technical articles. Another interesting result of the statistics is that translators use 1 Uyghur words to translate 1.91 Chinese characters on the average. However, the ratio changes slightly in different types of texts, for example, 2.15 Chinese characters to 1 Uyghur words in the daily conversation while 1.77 Chinese characters to 1 Uyghur words in government reports. This reflects that Uyghur translation of government report tend to be more literal than that of daily conversation. Mean sentence length is concerned, sentences in news and government documents of original or translated text are very long (47.47, 46.45 characters for Chinese and 25.31, 25.21 words for Uyghur respectively), which is 2-4 times of daily conversation and fiction (10.63, 24.86 characters for Chinese and 4.95, 10.63 words for Uyghur respectively).From this we can also find that longer sentence tend to have longer translation, and that shorter sentence tend to have shorter translation.

This proves that the correlation between the length of the sentence (or a paragraph) and the length of its translation was extremely high, which high correlation suggests that length might be a strong clue for sentence alignment.

4.2 future application of the parallel corpus

Generally speaking, parallel corpora are useful for all types of cross-lingual research. The value of a parallel corpus grows with its size and with the number of languages for which translations exist. The building of the Chinese-Uyghur parallel corpora is a great progress in corpus-based study of Uyghur language. At the current stage, however, we are chiefly focusing on developing of the corpus application tools, and we have not been able to carry out any large-scale investigations. The examples given in part should be sufficient to show the possibilities of using the Chinese-Uyghur parallel corpus.

(1) The parallel corpus offers specific uses and possibilities for contrastive and translation studies. it gives new insights into the languages that are not likely to be noticed in studies of monolingual corpus; it can be used for a range of comparative purposes and increase our knowledge of language-specific, typological and cultural differences, as well as of universal features; it illuminate differences between source texts and translations, and between native and non-native texts.

(2) The parallel corpus can be used for lexicography. Parallel corpus is necessary to clarify some terminological issues and acquisition of bilingual translation patterns; researchers can use it during bilingual dictionary compiling with the help of the concordance tools.

(3) The parallel corpus has application in building statistical machine translation and translation memory system. Useful data or knowledge could also be extracted from bilingual corpus based on statistical model providing translation examples for MT systems.

5. Conclusions and future works

We have just presented the construction process, including preprocessing, annotation and sentence alignment, of a Chinese-Uyghur parallel Corpus of about 2 million words. Some preliminary results and potential value of the corpus have also been introduced.

The parallel resource is relative rare at present, we would like to extend the material to other texts, both fiction and nonfiction, and to develop word alignment tools to improve the automatic word alignment in the near future. We hope that the corpus will provide ample material for text-based contrastive studies as well as for more specialized translation studies in the future.

6. Acknowledgments

The research work described in this paper is supported by the Social Science Foundation of China under grant number 07xyy019. We would like to thank the foundation and Prof. Hou Min for her guidance and great help during the study.

Bibliography

BAI Xiaojing, CHANG Baobao and ZHAN Weidong (2002), The construction of a large-scale of Chinese-English parallel Corpus. In proceedings of National Machine Translation Conference 2002. Electronic Industrial Publisher, Beijing. pp.124-131.

Beata B. Megyesi, Anna Sågvall Hein, and Eva Csato Johanson. 2006. Building a Swedish-Turkish Parallel Corpus. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). Genoa, Italy.

CHANG Baobao (2004). Chinese-English Parallel Corpus Construction and its Application. PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo

Chen S. F (1993). Aligning sentences in bilingual corpora using lexical information. In Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics, pages 9-16, Columbus, Ohio.

Dekai Wu(1994). Aligning a parallel English-Chinese corpus statistically with lexical criteria. In ACL-9$: 32nd Annual Meeting of the Assoc. for Computational Linguistics, pages 80-87, LasCruces, NM.

Dinh Dien and Hoang Kiem. Building an Annotated English-Vietnamese Parallel Corpus for Training Vietnamese-related NLPs.

Gale William & Kenneth Church (1993). A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19:1, pp. 75-102.

Graeme Kennedy. An Introduction to Corpus Linguistics. Foreign Language Teaching and Research Press, 2000.

Jörg Tiedemann and Lars Nygaard (2004). The OPUS corpus − parallel & free. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, May 26-28, 2004.

Koehn Philipp (2005). EuroParl: A Parallel Corpus for Statistical Machine Translation. Machine Translation Summit 2005. Phuket, Thailand.

Nerea Areta et al (2007). ZT Corpus: Annotation and tools for Basque. corpora. In Proceedings of Corpus Linguistics 2007. Birmingham, UK: University of Birmingham

Ralf Steinberger et al (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

Simard, M., Foster, G. and Isabelle, P. (1992). Using Cognates to Align Sentences in Bilingual Corpora', in Proceedings of the Fourth International Conference on Theoretical and Methodogical Issues in Machine translation (TMI92), (Montreal), 67-81

Stig Johansson and Cand.philol Jarle Ebeling (1994). The English-Norwegian Parallel Corpus: Introduction and Applications. In proceedings of International Conference on Cross-Language Studies and Contrastive Linguistics. 15 - 17 December 1994, Rydzyna, Poland