UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation

Liang Tian[1], Derek F. Wong[1], Lidia S. Chao[1], Paulo Quaresma[2,3], Francisco Oliveira[1], Yi Lu[1], Shuo Li[1], Yiming Wang[1], Longyue Wang[1]

[1]NLP[2]CT Lab / Department of Computer and Information Science, University of Macau, Macau

[2]Department of Portuguese, University of Macau, Macau

[3]L[2]F@INESC-ID and Department of Computer Science, University of Évora, Portugal

E-mail: tianliang0123@gmail.com,{derekfw, lidiasc}@umac.mo, pq@di.uevora.pt, olifran@umac.mo, {takamachi660, leevis1987, wang2008499, vincentwang0229}@gmail.com

Abstract

Parallel corpus is a valuable resource for cross-language information retrieval and data-driven natural language processing systems, especially for Statistical Machine Translation (SMT). However, most existing parallel corpora to Chinese are subject to in-house use, while others are domain specific and limited in size. To a certain degree, this limits the SMT research. This paper describes the acquisition of a large scale and high quality parallel corpora for English and Chinese. The corpora constructed in this paper contain about 15 million English-Chinese (E-C) parallel sentences, and more than 2 million training data and 5,000 testing sentences are made publicly available. Different from previous work, the corpus is designed to embrace eight different domains. Some of them are further categorized into different topics. The corpus will be released to the research community, which is available at the NLP[2]CT[1] website.

Keywords: English-Chinese parallel corpus, statistical machine translation, different domains

1. Introduction

The essence of Statistical Machine Translation (SMT) is to use the knowledge mined from the existing corpora to translate new text. Both the monolingual corpora and parallel corpora are valuable resources for SMT task, which can be used to collect enough statistical evidences for SMT parameter estimation.

There are already many parallel corpora today. However, only a few parallel corpora for English-Chinese (E-C) are publicly available, usually with fees and licensing restrictions. For example, parallel corpus provided by the ELDA (Evaluations and Language resources Distribution Agency, 1995-2013), the LDC (Linguistic Data Consortium) (Ma, 2006) and the UCCCRL (University Centre for Computer Corpus Research on Language, 1994-2013), require subscription or fees. The Biblical text (Resnik et al., 1999) that contains about 33,000 sentences and the MultiUN corpus consisting of about 300 million words per language extracted from the United Nations website (Eisele et al., 2010) are domain specific. The parallel corpora provided by OPUS for E-C contains corpora from very different domains. However, the size of the corpora is usually not more than 20,000 sentences (Tiedamann, 2012).

In this paper, the constructed corpus is a balanced corpus, which contains texts of different domains and genres in a reasonable proportion. During the construction process, existing mature algorithms or methods are adopted to accelerate the building process, such as document alignment (Resnik and Smith, 2003; Patry and Langlais, 2005), sentence boundary detection (Koehn, 2005; Gillick, 2009; Wong and Chao, 2010) and statistical sentence alignment approach (Moore, 2002).

The main objectives are two-fold: (1) creation of a large, multi-domain, and free parallel English and Chinese corpus for the construction of SMT translation models; (2) serves as an important resource to the study of SMT domain adaptation. The built corpus contain more than 15 million parallel sentences, and around 2 million of them are released to the public.

The structure of the paper is as follows. In section 2, some related works on building English-Chinese parallel corpus are considered. Section 3 lays out the process of corpus construction and the statistics about the parallel corpus. In section 4, the translation performance based on the constructed parallel corpus and testing data will be provided, followed by conclusions to end the paper.

2. Related Works

There are a number of English and Chinese parallel corpora publicly available. Most of them can be found online. However, only a few can be downloaded and are suitable for machine translation development. Furthermore, most of them either focus on specific domains or the size of the corpora is usually less than 1 million sentences. The Pool of Bilingual Parallel Corpora of Chinese Classics (PBPCCC) (Sun et al., 2009a, 2009b), the Xiamen University Corpus (XUC) (Lu, 2005) and the parallel corpus collected by Hong Kong Institute of Education (HKIE) (Wang, 2005) are online accessible. The Lancaster's Babel Parallel Corpus contains about 33 thousand sentences (Resnik et al., 1999). Corpora with similar small size include the Chinese English News

[1] http://nlp2ct.cis.umac.mo/um-corpus/index.html

Magazine Parallel Text (CENMPT: LDC2005T10), Information Services Department of Hong Kong Special Administrative Region (HKSAR: LDC2000T46) and Hong Kong Parallel Text (HKPT: LDC2004T08) (Ma, 2006). The largest one is the MultiUN corpus that extracted from the official documents of the United Nations (UN), which is available for all the 6 official languages of the UN, including Chinese (Eisele et al., 2010). The statistic summary of these corpora is presented in Table 1.

Corpora	Characters/Tokens		Sentences
	English	Chinese	
PBPCCC	10M	15M	-
XUC	3.3M	5.4M	0.22M
HKIE	1.88M	3.15M	-
BabelCT	0.25M	0.29M	-
MultiUN	220M	629M	4M
CENMPT	9M	20M	0.36M
HKSAR	11.9M	18.15M	-
HKPT	59M	98M	2M

Table 1: Statistic information of different parallel corpora ("-" information is unavailable).

3. The UM-Corpus

The parallel corpus, named *UM-Corpus*, has been designed to be a multi-domain and balanced parallel corpus. Two issues were mainly considered before the construction: the *quality* and the *varieties* of the content. For quality concern, online sources which give high quality of parallel text between English and Chinese were extracted. This includes the sites from online news, online dictionary and translation, where the parallel texts are manually aligned in either sentence or document level. For varieties concern, eight different domains were embraced, including *News*, *Spoken*, *Laws*, *Thesis*, *Educational Materials*, *Science*, *Speech/Subtitles*, and *Microblog*.

3.1 The Sources

The quality of parallel corpus is one of the main concerns, the sources of the parallel texts, to some extent, is a very crucial step. All the sources of the collected websites that contain English and Chinese texts are carefully selected and manually verified. These include the sources from online journals (national and international), official websites, online language learning resources (e.g. online dictionary and translation portals), TED, and Microblogs. For example, in the *cuyoo* website, the translation of news is well aligned at sentence level. In reading, the user is allowed to read it in different ways, e.g. translation pairs organized side by side at sentence level, or by odd-even line patterns at paragraph level. The quality is very good, and this allows us an easy way to crawl and extract the parallel text in sentence level that fits to SMT.

3.2 The Crawling

The standard parallel corpora have been fully discussed in many papers or books (Koehn, 2005; Tiedemann, 2009; 2010; 2011; 2012), some well-designed algorithms and tools can be used in practice, such as the work of Patry and Langlais (2005) in document alignment (with a precision of 99%), the works of Koehn (2005) and Gillick (2009) in sentence boundary detection (error rates on test news data are less than 0.25%), and the work of Moore (2002) in sentence alignment (it achieves 99.34% in precision).

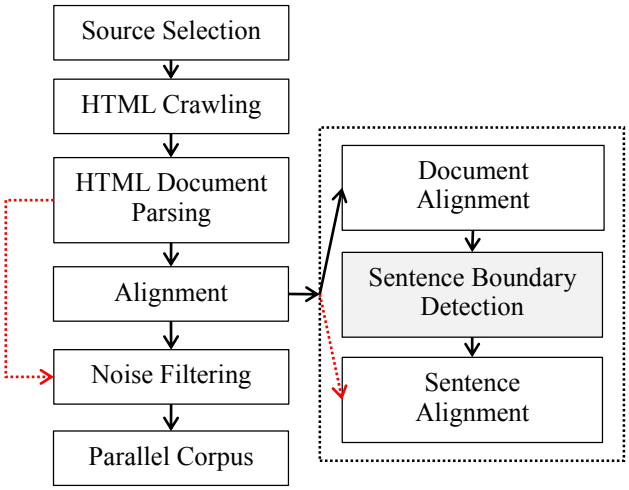

Figure 1: Construction process of UM-Corpus

The standard parallel corpus construction follows the process as illustrated in Figure 1. The overall construction process is divided into five major steps. The initial step is to identify the appropriate sources of the websites that contain the data and crawl the documents which are bilingual ready. In the second step of content extraction, the HTML files are parsed by discarding all the HTML tags profits from the function of NekoHTML[2] and XPath[3]. At the same time, the type of documents is analyzed for categorizing the text domain and topics in the subsequent stage. Documents are aligned in bilingual correspondence. Information together with the texts is stored in some unified formats. A key bridge between aligned documents and aligned sentences is the sentence boundary detection process. Different detection results will affect the alignment relation of sentence pairs, i.e. one to one or many to one alignment relationship. So far, the processing flow is automatically done. The final result is verified by human to get rid of the noisy texts, in particular the low quality translations. In total, there are 15,792,666 sentences prior to the removal of possible duplications.

3.2 Noise Removing

"Noise" is a general concept, which not only can refer to the extraneous information in the plain texts, but also can indicate the illegal data, such as the messy codes, mismatch

[2] http://nekohtml.sourceforge.net/

[3] http://www.w3schools.com/xpath/default.asp

	News	Spoken	Laws	Thesis	Education	Science	Subtitle	Microblog
Articles	173,994	-	64,630	26,853	-	-	-	-
Sentences	4,989,478	275,652	328,642	1,302,750	4,725,846	3,158,755	1,011,543	61,080
Percentage	31.59%	1.75%	2.08%	8.25%	29.92%	20.00%	6.41%	2.08%

Table 2: The constituent of UM-Corpus in terms of domains.

sentence alignments. In fact, the noise data exists throughout the standard collection steps.

During the document alignment, the HTML tags are the noises. Web pages often contain a lot of formatting information, indicating the color, font and style attribute of each piece of text, along with its position in the global layout of the page. This kind of noises can be removed with the help of the *Xpath* language.

The next task, before sentence segmentation, is to remove the extra spaces in the collected texts. One principle is that the continual Chinese characters should have no whitespaces or tab spaces, while continuous tokens of English or other similar language could only have a single space. Second, there should be no any space between the Chinese and non-Chinese text (i.e. Chinese characters followed by alphanumeric text). The removal of such extra whitespaces not only can transform the various text formats into a unified one, but also can improve the accuracy of subsequent tasks in processing and using the corpus.

The text is aligned at sentence level. However, there are still many mistakes that cannot be automatically identified. Manual alignment process was employed to improve the quality of the built corpora. 14 postgraduate students were arranged to do the job. Each student spends 54 days and about 324 hours on the task of English-Chinese language pairs. To accelerate the editing process, *SuperAlign*[4] was used in practice, which is a Windows application adapted from the *HunAlign* (Varga et al., 2005). As required in *Moses*[5], long sentences and unaligned sentences should be removed as they can cause problems with the training pipeline, and obviously sentences with very different sentence length as well as long sentences (more than 200 words) are discarded also in the target corpus.

3.3 The Analysis

Finally, more than 15 million sentences (15,764,200) are collected in the UM-Corpus after filtering, i.e. removal of duplicated sentences and those sentences with more than 200 words. The detailed statistics are summarized in Table 2 and Table 3. In Table 3, the tokens number (**Tokens**), average sentence length (**Avg. Len.**) and the vocabulary size (**Voc.**) of the constructed parallel corpus are listed, from which we see the parallel corpus is large enough for the development of machine translation. Figure 2 shows the distribution of different domain data and Figure 3 presents the length distribution of the corpus, where the Chinese text is counted by Chinese characters. From Table 2, it is easy to find that the News contains the most sentences (4,989,478), followed by the domain of Education (4,725,846), which profits from the large amount of resources in the web. The smallest portion is the parallel Microblog text, which is due to the rare available resources in the Internet. Long sentences ($Len_s \geq 200$ words) and unaligned sentences are removed as they may cause problems in the Moses training pipeline[6]. From Figure 3, we know that more than 85% of the corpus are sentences with length less than 50 words.

Lang.	Tokens	Avg. Len.	Voc.
English	381,921,583	23.90	1,690,792
Chinese	572,277,658	35.81	388,611

Table 3: Statistics of the 15 million UM-Corpus.

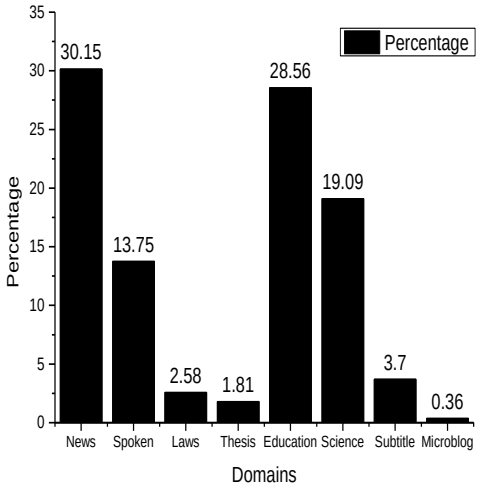

Figure 2: Distribution of domain data in UM-Corpus

3.4 Components of Corpus

The corpus is organized and categorized in eight different text domains. Texts from several topics and text genres are covered. We provide a brief information for each domain:

- **News**: News can cover a lot of topics, such as *Politics*, *Economy*, *Technology*, *Education*, *Agriculture*, and *Society*. The news collected mainly covers the articles published from year 2000. There are 173,994 articles and about 5 million sentences.

[4] http://sourceforge.net/projects/superalign/

[5] http://www.statmt.org/moses/?n=Moses.Baseline

[6] http://www.statmt.org/moses/?n=Moses.Baseline

Domains	Languages	Tokens	Average Length	Vocabularies	Sentences
News	English	8,646,174	19.2137	274,546	450,000
	Chinese	15,277,414	33.9498	47,902	
Spoken	English	1,836,670	8.3485	107,923	220,000
	Chinese	3,033,052	13.7866	9,011	
Laws	English	5,926,316	26.9378	66,330	220,000
	Chinese	8,783,941	39.9270	14,723	
Thesis	English	5,962,590	19.8753	378,679	300,000
	Chinese	10,514,430	35.0481	149,110	
Education	English	8,401,095	18.6691	293,595	450,000
	Chinese	13,749,570	30.5546	38,663	
Science	English	598,050	2.2150	115,968	270,000
	Chinese	1,527,849	5.6587	8,972	
Subtitles	English	2,299,742	7.6658	101,423	300,000
	Chinese	3,818,490	12.7283	13,854	
Microblog	English	72,144	14.4288	12,083	5,000
	Chinese	125,415	25.0828	3,525	
Total	English	33,742,781	13.2862	832,518	2,215,000
	Chinese	56,830,161	22.5039	209,729	

Table 4: Statistic summary of released 2.2 million UM-Corpus.

- **Spoken**: These texts mainly contain widely used spoken English in the English-speaking countries. Other types of text are from the video dialogue, such as the *Family Album U.S.A*. Without removing the duplicated sentences, there are about 0.3 million sentences in total.
- **Laws**: The law statements mainly come from the mainland China (1,561 documents), Hong Kong (150,000 documents), and Macau (186 documents). It has more than 0.3 million sentences.

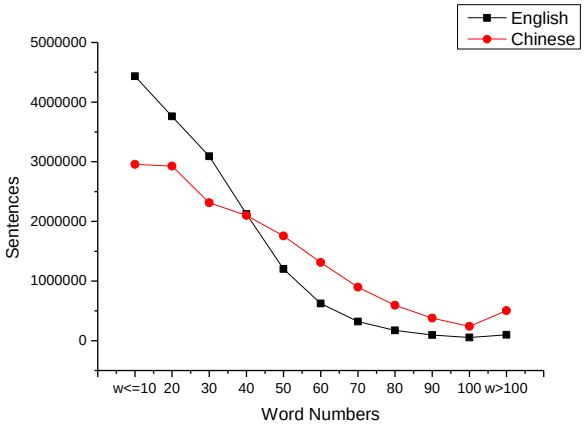

Figure 3: Sentence length distribution of UM-Corpus

- **Thesis**: This portion covers 15 journal topics in the research area, including electronics, traffic, agriculture, medicine, biology, aerospace, mathematic, economy and etc. However, it is not easy to distinguish the texts in terms of each topic because of its complex webpage structure. It has around 1.3 million sentences.
- **Education**: The texts in this domain are acquired from online teaching materials, such as language teaching resources, dictionaries, etc., which can be served as language education. Totally, about 5 million sentences are collected.
- **Science**: The texts mainly consist of parallel terminologies and sentences in science and technology areas. The total number is around 3.2 million.
- **Subtitles**: This collection of text covers the subtitles from TED talk and movie subtitles. They are included in this corpus to serve for dialog and spoken translation research. The data consists of around 1 million of sentences.
- **Microblog**: Compared to the edited genres that have played a central role in NLP research, microblog or twitter texts use a more informal register with nonstandard lexical items, abbreviations, and free orthographic variations. When confronted with such input, conventional text analysis tools often perform poorly. Therefore, twitter translation is very active recently in SMT community. To provide valuable training corpus in recent social network study in SMT, the parallel texts are also added in the UM-Corpus. There are more than 60,000 sentences.

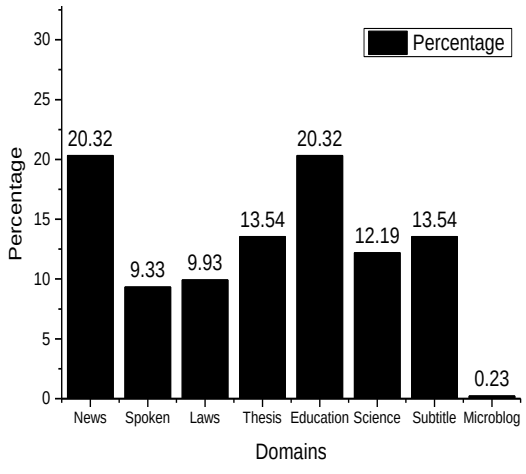

Figure 4: Distribution of domain data in released UM-Corpus

3.5 Released Corpus Analysis

Although more than 15 million sentences are collected in the UM-Corpus, only more than 2 million of sentences containing eight domains are released to the community for research purpose. The reason is that those sentences are carefully designed and manually proofread and edited by our group. The content is properly adjusted to embrace different domain data in proportion. This can be served as an important resource for the study of domain adaptation in statistical machine translation. The detailed statistics are summarized in Table 4. Figure 4 shows the distribution diagram of different domain data. From Table 4, it is easy to find that the *News* and the *Education* contain the most sentences (450,000). The smallest portion is the parallel *Microblog* text, which is due to the rare available resources in the Internet.

All the released parallel corpus is named by the format of *Bi-domain.txt*, for example, *Bi-News.txt* indicates the News parallel corpus.

3.6 Testing Data

Besides the training data, another set of sentences is also collected along with this parallel corpus for evaluation purpose. The designed testing data is fully considered for different domains/genres. There are 1,500 sentences for *news*, 500 sentences for *laws*, 500 sentences for *spoken*, 600 sentences for *thesis*, 700 sentences for *education*, 600 sentences for *science* and 600 for *subtitle*. The statistics summary for the 5,000 sentences is shown in Table 5.

Languages	Tokens	Average Length
English	107,709	21.5289
Chinese	144,018	28.7863

Table 5: Statistics for UM-Corpus testing data.

4. Copyright Problems

The use of crawling technique to extract texts from Internet for a research could raise legal and ethical questions. It is clear that storing whole texts and allowing retrieval on them would be an unacceptable violation of copyright. To avoid copyright restrictions as much as possible, some anonymizing operations on the corpus were conducted. That is, we replaced the proper names, such as person names, corporation names, product brands, etc., with certain placeholders during the manual edit process. In addition, sentence pairs that contain numbers or dates are replaced into random numbers.

5. SMT Experiment

The objective of this work is to provide a well design corpus that has large enough data for developing and evaluating a SMT. In setting up the experiments, the standard configurations (Koehn et al., 2007) were used, with MERT optimization (Och and Ney, 2003; Bertoldi et al., 2009) and pruning (Johnson et al., 2007). The phrases are extracted from the results generated from GIZA++ (Och and Ney, 2003). The word alignment was trained with ten iterations of IBM model 1 and model 4 (Brown et al., 1990; 1993) and six iterations of the HMM alignment model (Vogel et al., 1996). The language model is created by the external toolkits IRSTLM (Federico, 2008) from all the UM-Corpus with 5-gram model. 5,000 sentences UM-Testing was used for the development set and MT05 were used as the test data. During the training step, the Chinese texts were tokenized by ICTCLAS (Zhang, 2003). All the translation results were measured by BLEU (Papineni, 2002). The results are shown in Table 6.

Translation Directions	BLEU	
	UM-Testing	MT-05
English to Chinese	31.55	36.71
Chinese to English	28.67	29.48

Table 6: BLEUs for English-to-Chinese and Chinese-to-English MT.

6. Conclusions

In this paper, we introduce a high quality and large English-Chinese parallel corpus, UM-Corpus, designed for SMT research. The corpus consists of eight different domains and some of them are further categorized into different topics. This data is very suitable for the development of English-Chinese SMT and its study in domain adaption. Two million sentences are freely released to the community for research purpose. The corpus is licensed under the Creative Commons Non-Commercial 4.0 License[7].

7. Acknowledgements

The authors would like to thank all reviewers for the very careful reading and helpful suggestions. The authors are grateful to the Science and Technology Development Fund of Macau and the Research Committee of the University of Macau for the funding support for their research, under the Reference nos. MYRG076(Y1-L2)-FST13-WF and MYRG070(Y1-L2)-FST12-CS.

8. References

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, P. F., Pietra, V. J. Della, Pietra, S. A. D., Mercer, R. L. (1993). The mathematics of statistical machine translation. *Computational Linguistics*, 19(2): 263–311.

Bertoldi, N., Haddow, B., Fouet, J.B. (2009). Improved Minimum Error Rate Training in Moses. *Prague Bulletin of Mathematical Linguistics*, 91:7–16.

Eisele, A., Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In *Proceedings of the 7th International Conference on Language*

[7] http://creativecommons.org/licenses/by-nc-nd/4.0/

Resources and Evaluation (LREC'2010).

Federico, M., Bertoldi, N., Cettolo, M. (2008). IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association*.

Gillick, D. (2009). Sentence Boundary Detection and the Problem with the U.S. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 241–244.

Johnson, J. H., Martin, J., Foster, G., Kuhn, R. (2007). Improving Translation Quality by Discarding Most of the Phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 967–975.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions*, pages 177–180.

Lu, W. (2005). English-Chinese Corpus. Xiamen University. http://www.luweixmu.com/ec-corpus/query.asp (last accessed: March 21, 2014).

Ma, X., Cieri, C. (2006). Corpus support for machine translation at LDC. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 135–144.

Och, F. J., Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Patry, A., Langlais, P. (2005). Automatic identification of parallel documents with light or without linguistic resources. In *Canadian Conference on AI*, pages 311–318.

Resnik, P., Broman, M. B., Diab, M. (1999). The Bible as a Parallel Corpus: Annotating the Book of 2000 Tongues. *Computers and the Humanities*, 33: 129–153.

Resnik, P., Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Sun, H., Yang, J. (2009). Collected Chinese Documents Aligned with English versions at Sentential Level (CCDAE). December 25. http://corpus.usx.edu.cn/ (last accessed: 21 March, 2014).

Sun, H., Yang, J. (2009). Parallel Corpus of China's Legal Documents (PCCLD). August 12. http://corpus.usx.edu.cn/ (last accessed: 21 March, 2014).

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Recent Advances in Natural Language Processing (vol V)*, pages 237–248.

Tiedemann, J. (2010). Lingua-Align: An Experimental Toolbox for Automatic Tree-to-Tree Alignment. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

Tiedemann, J. (2011). *Bitext Alignment (Synthesis Lectures on Human Language Technologies)*. Morgan & Claypool Publishers.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596.

Vogel, S., Ney, H., Tillmann, C. (1996). HMM-Based Word Alignment in Statistical Translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, pages 836–841.

Wang, L. (2005). E-C Concord. Hong Kong Institute of Education. http://ec-concord.ied.edu.hk/paraconc/index.htm (last accessed: 21 March, 2014).

Wong, F., Chao. S. (2010). *i*Sentenizer: An incremental sentence boundary classifier. In *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 1–7.

Zhang, H.P., Yu, H.K., Xiong, D.Y., Liu, Q. (2003). HHMM-based Chinese Lexical Analyzer ICTCLAS. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, 17, pages 184–187.