A Survey on Existing Chinese-Japanese Bilingual Resources

Jing Sun and Yves Lepage

Graduate School of IPS, Waseda University
2-7 Hibikino, Wakamatsu, Kitakyushu 808-0135, Japan
{cecily.sun@akane, yves.lepage@aoni}.waseda.jp

Abstract

Most data-driven Natural Language Processing applications, such as Machine Translation or cross-lingual information retrieval require large parallel corpora. However, such resources only exist for limited language pairs and knowledge domains. Researchers have been putting great time and effort in building bilingual or multilingual parallel corpora mainly for language-pairs that involve English. Consequently, parallel corpora between languages other than English are scarce, even for well-documented languages like Chinese and Japanese. This paper reports on existing Chinese-Japanese bilingual corpora from the perspectives of their providers, sizes, annotation status and availability, and on state-of-the-art practices in constructing and evaluating bilingual corpora. This paper introduces also an initiative in building a publicly available, and on Chinese-Japanese bilingual resource automatically from the Web by applying bootstrapping loop and utilization of the linguistic features of both Chinese and Japanese languages.

Keywords: Survey, Bilingual Lexica, Parallel Corpora, Chinese, Japanese, LRL.

1. Introduction

Parallel corpora are collections of articles, paragraphs, sentences, or sub-sentential fragments in two or more languages. Parallel corpora are critical resources for most data-driven Natural Language Processing tasks, such as Machine Translation and Cross-lingual Information Retrieval.

Great time and effort have been spent in building bilingual or multilingual parallel corpora. Resources for language pairs that involve English are plentiful and many of them are freely accessible. Well known corpora include the European Parliament Proceedings Parallel Corpus (Koehn, 2005) which comprises of a collection of parallel text in 11 languages from the proceedings of the European Parliament, the Meteo parallel corpus (Langlais et al., 2005) which consists of naturally occurring weather forecasts written by humans in English and French, and the Multilingual UN Parallel Text (Graff, 1993) which consists of articles from the UN web site in all the six official languages of the UN.

However, parallel corpora for languages pairs that do not involve English are scarce, even for well-documented languages like Chinese and Japanese. Although some Chinese-English and Japanese-English corpora are downloadable from some resource providers, such as the Linguistic Data Consortium (LDC)[1], bilingual resources between Chinese and Japanese are rather rare. According to our survey, surprisingly, amongst currently existing Chinese-Japanese bilingual resources, the great majority is not publicly available or is limited in size or domain coverage.

Nowadays, various machine translation software and platforms are available that provide bi-directional translation service between Chinese and Japanese, yet the translation quality is far from the quality for language pairs that include English. So far as we know, most of these translation software and platforms use English as an intermediate language, i.e., they perform translation from Chinese into English first, and then translate from English into Japanese, or vice versa. By doing so, the error rate is inevitably increased. Such problems arise from the lack of Chinese-Japanese parallel corpora.

In Section 2, two specific problems for Chinese-Japanese parallel corpora will be described. In Section 3, a survey report on present existing Chinese-Japanese bilingual resources at both lexical and sentential levels will be introduced from the perspectives of their authorities, sizes, knowledge domains, annotation status, and availabilities for public. Besides, some state-of-the-art approaches in constructing Chinese-Japanese bilingual resources and the methods in evaluating such resources are also discussed. In Section 4, a collaborative initiative of automatic construction of a freely available Chinese-Japanese bilingual corpus from the Web will be described.

2. Specific Problems for Chinese-Japanese Parallel Corpora

Although we see the importance of bilingual corpora between Chinese and Japanese, we face specific difficulties in constructing such resources.

Standardization in segmentation: Word segmentation of written texts for Chinese, Japanese and Korean are usually the first step in most Natural Language Processing tasks like Machine Translation or Information Retrieval. Unlike European languages, there are no typographic boundaries, such as a space between words in written Chinese, Japanese or Korean. In recent years, word segmentation techniques have been discussed and improved greatly for each of these languages both theoretically and practically. However, each of them is limited for itself. It results in various problems. For instance, the different scales of the segmentation particles will directly affect the performance for phrasal extraction and sentence alignment. Choi (2009) discussed the principles for word

[1] LDC: http://www.ldc.upenn.edu/

segmentation that applied for Chinese, Japanese and Korean as well as the impact of word segmentation standards. In August 2011, an ISO standard for word segmentation for Chinese, Japanese and Korean was published.[2]

Copyright: In the past decades, great effort, human work and financial grants have been devoted in building large- scale Chinese-Japanese bilingual corpora. However, many of them cannot be released publicly with a long- time delay due to the matter of the copyright. In general, lengths of the standard copyright vary in different countries, ranging from 0 year[3] to 80 years plus the author's lifetime. For both China and Japan, the copyright length is 50 years after the author's death. This indicates that, most copyright-free written works were published at least 50 years ago. However, modern Japanese language has been changed significantly in 50 years time, thus Japanese text documents might not be usable for constructing parallel corpora which mainly work for Machine Translation and Information Retrieval tasks. In order to release the bilingual corpora publicly without copyright restrictions, the best way would be to construct such corpora from freely available resources.

3. Survey of Existing Chinese-Japanese Bilingual Resources

Bilingual corpora can be classified into spoken language corpora and written corpora. In recent years, construction of bilingual parallel corpora of spoken utterances has attracted attentions of many scholars and researchers. For instance, the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2007) contains edited spoken travel expressions and has over 588,000 utterances in English, Chinese and Japanese; the Spoken Language DataBase (SLDB) has over 16,000 utterances in hotel spoken language corpus. In this survey report, we mainly look into written bilingual corpora. From the content perspective, bilingual corpora can be classified into sentential and sub-sentential levels. For sub-sentential alignment, the mainstream is still at lexical and phrasal level, i.e. word-to-word alignment (Brown et al., 1993; Melamed, 1997; Och and Ney, 2003). Munteanu (2006) also proposed a method to extract parallel sub-sentential fragments from comparable bilingual corpora in Romanian and English. In the meanwhile, Cromieres (2006) suggested an algorithm to use the co-occurrence information on substrings and suffix arrays to tackle the problem of unbalanced segmentation for some Asian languages like Japanese and Chinese. In the following, we mainly discuss bilingual corpora at lexical and sentential levels.

3.1. Bilingual Lexicons

3.1.1. Approaches in Building Bilingual Lexicons

A bilingual lexicon is an essential part for multilingual processing tasks, such as Machine Translation (Och and Ney, 2003) or cross-lingual information retrieval (Grefenstette, 1998). Manual work in constructing a bilingual lexicon normally ensures good quality by translating from source into target languages by skilled human translators. Despite of that, manual work is known as tedious, costly and time-consuming. There are numerous approaches in building bilingual lexicon and extracting parallel sentences from parallel or comparable corpora. Hereunder, few selected methods in automatically building bilingual lexicons will be examined.

Using parallel corpora: Traditionally, one extracts bilingual lexicon from ready-made parallel corpora, using IBM Model as described in (Brown et al., 1993) and (Och and Ney, 2003). Such a method only works when there are adequate bilingual parallel corpora. Unfortunately, good-quality bilingual parallel corpora are scarce, especially for those language pairs which do not include English.

Using comparable corpora: Recently, great attention has been drawn to extract bilingual lexicons from comparable corpora, in which texts are not direct translation of each other (Fung, 2000; Otero, 2008). Yu and Tsujii (2009) proposed to extract bilingual dictionary from Wikipedia with both context heterogeneity and dependency heterogeneity. Li and Gaussier (2010) suggested to improve the quality of the comparable corpus from which the bilingual lexicon is to be extracted by improving the text comparability.

Using English as a Pivot Language: Tanaka and Umemura (1994) proposed to use an intermediate third language like English to build a Japanese-French corpus. By applying similar methods, Zhang et al. (2005) used the same approach together with the one time inverse consultation method to construct a Japanese-Chinese dictionary. Goh et al. (2005) used two freely available dictionaries (Japanese-English and Chinese-English) to construct a Chinese-Japanese bilingual lexicon using English as a pivot language. The main limitation for this method of using a third language as a pivot is the word ambiguities that arise during the joint product. Kaji et al. (2008) extracted word associations from monolingual corpora of both languages and effectively tackled this problem to some extent.

Using Chinese-Japanese Linguistics Features: Written Chinese and Japanese are closer to each other than to English and they share some common features in writing. The writing system of Chinese is Hanzi, while Japanese is composed with Hiragana, Katakana and Kanji. The Kanji characters were historically imported from China. Zhang et al. (2003) constructed a Kanji-Hanzi conversion table and calculated the similarity between kanji and Hanzi using Unicode and one time inverse consultation method. In the same time, Goh et al. (2005) suggested a direct conversion from Kanji to Hanzi based on the ideographs themselves. There is a certain amount of Chinese Hanzi which bear equivalent features with certain Japanese Kanji even though they share different pronunciations. Examples of some four-word idioms written in Simplified, Traditional Chinese and Kanji in Japanese are shown in Table 1. Differences in forms are in bold. All these words are equivalent in both form and semantic meaning. The challenge of such conversion is described in Table 2. In some cases, Chinese Hanzi and Japanese Kanji are of the same or similar form but bear completely different meanings. Such cases should not appear in a Chinese-Japanese lexicon.

[2] ISO 24614-2:2011

[3] Afghanistan, Laos and Marshall Islands have no copyright legislation.

Chinese		Japanese	Meaning
Simplified	Traditional		
半信半疑	半信半疑	半信半疑	‘doubtfully’
百**发**百中	百**發**百中	百**発**百中	‘in the bull’s eye’
恻隐之心	**惻隱**之心	**惻隠**之心	‘sympathy’
朝三暮四	朝三暮四	朝三暮四	‘capricious’
大**学**生	大**學**生	大**学**生	‘college student’
紧张	**緊張**	**緊張**	‘nervous’

Table 1: Examples of Hanzi-Kanji conversion that preserve meaning.

Japanese		Chinese		
Word	Meaning	Simp.	Trad.	Meaning
暗算	‘arithmetic’	暗算	暗算	‘secret plot’
処分	‘disposal’	**处**分	**處**分	‘punish’
床	‘floor’	床	床	‘bed’
手**紙**	‘letter’	手**纸**	手**紙**	‘toilet paper’
娘	‘daughter’	娘	娘	‘mother’
経理	‘cashier’	**经**理	**經**理	‘manager’

Table 2: Examples of Hanzi-Kanji conversion leading to different meanings (*false friends*).

3.1.2. Measuring the Quality of Bilingual Lexicons

The most common method in evaluating the quality of a bilingual lexicon consists in computing Recall and Precision scores. Researchers also use bilingual lexicons to extract parallel sentences and evaluate the parallel-ness of the test sets of the parallel corpus by applying SMT systems. Yu and Tsujii (2009) proposed a metric of Accuracy and MMR (Voorhees, 1999) to evaluate the extracted dictionary.

3.1.3. Current Existing Chinese-Japanese Bilingual Lexicon

Although many researchers and institutions are working on the automatic extraction of bilingual lexicon from parallel or comparable corpora, very few such resources have been released for Chinese and Japanese. Up to now, the most notable and rather easily available Chinese- Japanese bilingual lexicon is the EDR Electronic Dictionary (Japanese-Chinese Part) constructed by the NICT (National Institute of Information and Communication Technology) in Japan. Table 3 shows some features of the EDR Dictionary. And Table 4some Web-downloadable Chinese-Japanese/Japanese-Chinese Bilingual lexicons.

Provider	NICT
Released	2010
Publicly available	No
Annotation	Manually annotated
Size	240,000 entries

Table 3: Information and statistics of the EDR dictionary.

3.2. Parallel Corpora at Sentence Level

Similarly as for bilingual lexicons, parallel corpora of aligned sentences are a relatively scarce resource for Chinese-Japanese, although they are so important for many tasks, such as training and tuning statistical machine translation systems.

3.2.1. Approaches in Building Sentential-level Parallel Corpora

In this section approaches in extracting parallel sentences will be described according to their difficulties, i.e. the parallel-ness of the input bilingual documents.

Collect from parallel documents: This is the direct way to obtain parallel sentences, but it requires the high parallel-ness of the input bilingual parallel texts. Details are described in (Manning and Schütze, 1999; Wu and Fung, 2005).

Collect from comparable documents: Munteanu et al. (2006) used news articles published within the same 5-day windows. On the other hand, Otero and López (2010) converted the original Wikipedia into a codified corpus and managed to mine parallel sentences from it.

Collect from very-non-parallel documents: Researchers have been working toward the challenge of mining parallel sentences from very-non-parallel bilingual corpora. Fung and Cheung (2004) proposed to apply bootstrapping and IBM Model 4, and the EM lexical learning method in mining Chinese-English parallel sentences from the TDT corpus, a collection of news stories in different time ranges.

3.2.2. Measuring the Quality of Bilingual Parallel Corpora

Objective Evaluation by human is time consuming, costly and the standard varies from one to one. Precision, Recall, F-measure and AER metrics are often considered as measures for the quality of bilingual parallel corpora. Besides, it is also applicable to implement Statistical Machine Translation and evaluate the output results by comparing the TER and BLEU scores. Log-likelihood ratio statistics have also been proposed by Kumano et al. (2007) to assess the reliability of the alignment.

3.2.3. Current Existing Chinese-Japanese Bilingual Parallel Corpora

In Table 5, we list the most notable Chinese-Japanese bilingual parallel corpora to our best of our knowledge. The majority of these corpora are translated from source language into target language manually. None of them has been constructed automatically. Except from the smallest corpora, the JEC Basic Sentence Data, all the rest is not publicly or freely available. NICT Japanese- Chinese Parallel Corpus is expected to be released publicly in the future (Zhang et al., 2005) with no precision. Table 5 describes these parallel corpora in statistics. Abbreviations of the institutions are listed below.

NICT:National Institute of Information and Communications, Japan.

Kyoto U.: Kyoto University, Japan.

HIT: Harbin Institute of Technology, China.

BCJS: Beijing Center for Japanese Studies, affiliated to Beijing Foreign Studies University, China.

PKU: Institute of Computational Linguistics, Peking University, China.

Dictionary	Languages	Statistics	Source	URL
Omegawiki DB	zh-ja ja-zh	541 492	OmegaWikiProject	www.omegawiki.org/
Universal dict. DB	zh-ja/ja-zh	1,263	Dicts.info Project	www.dicts.info/
Wiktionary DB	zh-ja ja-zh	1,877 1,998	Wiktionary.org Project	witionary.org/
IPS dictionary	zh-ja/ja-zh	2,032	IPS, Waseda University	www.waseda.jp/ips/
Wikipedia interlink	zh-ja/ja-zh	7,173	Wikipedia database	www.wikipedia.org/

Table 4: Some free zh-ja/ja-zh bilingual lexicons as of September, 2011

Providers	**NICT**	**Kyoto U. & NICT**	**HIT**	**BCJS**	**PKU**
Corpus	NICT Japanese-Chinese Parallel Corpus	JEC Basic Sentence Data	Chinese-English-Japanese Trilingual Corpus for Olympics	Chinese-Japanese Bilingual Corpus	Chinese-Japanese Bilingual Corpus
Domains	News Mainichi News Paper	Tourism	Tourism, Beverage, Transportation, Sports and Business	Literature, Politics Reviews, Poem, Law etc.	Unkown
Annotation	Automatic annotation with human revision	Syntactic and case structure annotated	No	Automatic and manual annotation	Manually annotated
Nbr. Sentence Pairs	38,383	5,304	**52,227**	Unknown	2,000
Nbr. Characters	1,410,892 (Chinese)	197,597 (Chinese) 263,453 (Japanese)	750,809 (Chinese) 1,066,071 (Japanese)	**20,130,000** (Chinese and Japanese)	Unknown
Release (year)	2008	2011	2004	2003	2003
Public availability	No	**Yes**	No	No	No
Sentence length	Long	Short	Short	Long	Unknown

Table 5: Current Available Chinese-Japanese Parallel Corpora.

4. An Initiative in Building a Publicly Available Chinese-Japanese Bilingual Resource from the Web

We are currently working on a joint project with the NLPR-CASIA[4] to construct a large-scale, richly featured and expandable Chinese-Japanese bilingual resource from the Web by applying bootstrapping loop and utilization of the linguistic features of both Chinese and Japanese languages. Such resource shall comprise of both sentential and sub-sentential level alignments. Our ultimate goal is to make this resource publicly and easily available for all researchers, at no cost.

We built Chinese-Japanese bilingual Translation Tables using English as a pivot language from Chinese-English and Japanese-English lexica. Translation probabilities were estimated by using the number of occurrences in the lexica as initial values. The main limitation of using a pivot language is the word ambiguities that arise during the joint product. We converted the Chinese-English and Japanese-English dictionaries of XDXF[5] and Wiktionary. Manual check showed that accuracies reached 42.6% and 65.2% respectively. Known techniques mentioned in Section 3.1.1. will be applied to improve the accuracy.

During the past few months, over 98,000 Wikipedia pages in English, Chinese and Japanese have been crawled. Amongst them, there are 18,135 pairs of Chinese and Japanese documents which supposed to be comparable. Table 6 shows the statistics of this raw corpus in language pairs sharing the same topics. As we used interlanguage link words to locate page pairs, a limited number of identical pages has been crawled.

By using the previously converted dictionaries or the EDR dictionary as a reference, we tried to estimate similarities between sentences automatically in order to extract parallel sentences from this raw corpus. We performed some experiments by computing the word overlapping rate, still with inconclusive results, mainly because of inconsistencies in word segmentation across Chinese and Japanese. We thus consider computing rather a non-coverage rate at the character level. Moreover, we also try to perform the iteration to make such resource expandable.

lang	**ja-en-zh**	ja-en	**ja-zh**	en-zh	Total
doc pairs	16,176	21,084	1,959	2,087	41,306
doc nbr.	48,528	42,168	3,918	4,174	98,788

lang	sent nbr	uniq sent	characters	avg. length
zh	617,190	181,590	8,478,121	47
ja	857,313	262,299	14,862,887	57

Table 6: Statistics on ja & zh crawled docs

[4] NLPR-CASIA: National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, China

[5] XDXF - The Extensible (XML) Dictionary Exchange Format: http://xdxf.revdanica.com/down/

5. Conclusion

We have conducted a survey on existing Chinese- Japanese parallel resources in both sentential and sub-sentential levels and looked at the most notable ones. We also discussed some state-of-the-art approaches in extracting and building Chinese-Japanese bilingual lexicons and parallel corpora of aligned sentences as well as the methods to evaluate the quality of such parallel corpora. At the end, we introduced our research project in constructing a large scale and freely-available Chinese-Japanese bilingual resource from the Web.

Acknowledgments

This work was supported in part by Foreign Joint Project funds from Kitakyushu Foundation for the Advancement of Industry, Science and Technology (FAIS).

References

P. F. Brown, S.A.D. Pietra, V.J.D. Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translaiton: parameter estimation. *Computational Linguistics*, 19-2.

K. S. Choi, H. Isahara, K. Kanzaki, H. Kim, S.M. Pak, and M. Sun. 2009. Word segmentation standard in chinese, japanese and korean. In *Proc. of the 7th Workshop on Asian Language Resources, ACL-IJCNP*, pages 179–186, Suntec, Singapore.

F. Cromieres. 2006. Sub-sentential alignment using substring co-occurrence counts. In *Proc. of the COLING/ACL 2006 Student Research Workshop*, pages 13–18, Sydney.

P. Fung. 2000. *A Statistical View on Bilingual Lexicon Extraction from Parallel Corpora to Non-parallel Corpora*. Kluwer Academic Publishers.

P. Fung, P.and Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proc. of EMNLP*, pages 57–63.

C-L. Goh, M. Asahara, and Y. Matsumoto. 2005. Building a japanese-chinese dictionary using kanji/hanzi conversion. In LNAI 3651, editor, *R. Dale et al. (Eds.): IJCNLP*, pages 670–681.

D. Graff. 1993. The un multilingual text corpus. In *LDC Newsletter*, volume 1(3). Linguistic Data Consortium.

G. Grefenstette. 1998. The problem of cross-language information retrieval. In *Cross-language Information Retrieval*. Kluwer Academic Publishers.

H. Kaji, T. Shin'ichi, and E. Dashtseren. 2008. Automatic construction of japanese-dictionary via english. In *Proc. of the 6th Intl. Conference on language Resources and Evaluation*, pages 699–706.

P. Koehn. 2005. A parallel corpus for statistical machine translation. In *Proc. of MT Summit*, pages 79–86.

T. Kumano, H. Tanaka, and T. Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability smt model. In *Proc. of the 11th Intl. Conference on Theoretical and Methodological Issues in Machine Translation*, pages 95–103, Skövde, Sweden.

G. Langlais, S. Gandrabur, T. Leplus, and G. Lapalme. 2005. The long-term forecast for weather bulletin translation. In *Machine Translation*, volume 19(1), pages 83–112, March.

B. Li and E. Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proc. of the 23rd Intl. Conference on Computational Linguistics (Coling)*, pages 644–652.

C. Manning and H. Schütze. 1999. Foundations of statistical natural language processing. In *MIT Press, Cambridge, MA*.

I. D. Melamed. 1997. A portable algorithm for mapping bitext correspondence. In *35th Conference of the ACL)*, Madrid, Spain.

D. S. Munteanu. 2006. Extracting parallel sub- sentential fragments from non-parallel corpora. In *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 81–88, Sydney.

F. Och and H. Ney. 2003. A systematic comparison of v ariousstatisticalalignmentmodels. In *Computational Linguistics*, volume 29(1), pages 19–51.

P. G. Otero and I. G. López. 2010. Wikipedia as multilingual source of comparable corpora. In *Proc. of the 3rd Workshop on BUCC, LREC 2010*, pages 21–25, Malta.

P. G. Otero. 2008. Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. In *Proc. of LREC 2008 Workshop of BUCC*, pages 19–26.

T. Takezawa, G. Kikui, M. Mizushima, and E. Sumita. 2007. Multilingual spoken language corpus development for communication research. computational linguistics and chinese language processing. In *The Association for Computational Linguistics and Chinese Language Processing*, volume 12(3), pages 303–324.

K. Tanaka and Kyoji. Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proc. of the 15th Intl. Conference on Computational Linguistics*, pages 297–303.

E. M. Voorhees. 1999. Thetrec-8question answering track report. In *Proc. of the 8th Text Retrieval Conference*.

D. K. Wu and P. Fung. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *IJCNLP*, pages 257–268.

K. Yu and J. Tsujii. 2009. Bilingual dictionary extraction from wikipedia. In *Proc. of MT Summit XII*.

Y. Zhang, Q. Ma, and H. Isahara. 2003. Automatic acquisition of a japanese-chinese bilingual lexicon using english as an intermediary. In *Proc. of NLPKE*, pages 471–476.

Y. Zhang, K. Uchimoto, Q. Ma, and H. Isahara. 2005. Building an annotated japanese-chinese parallel corpus–a part of nict multilingual corpora. In *Proc. of MT Summit X*, pages 71–78.