Mining Large-scale Comparable Corpora from Chinese-English News Collections

Degen Huang[1] **Lian Zhao**[2] **Lishuang Li**[3] **Haitao Yu**[4]
Department of Computer Science and Technology
Dalian University of Technology
[1]huangdg@dlut.edu.cn [3]lils@dlut.edu.cn
[2]zhaolian@mail.dlut.edu.cn [4]gengshenspirit@163.com

Abstract

In this paper, we explore a CLIR-based approach to construct large-scale Chinese-English comparable corpora, which is valuable for translation knowledge mining. The initial source and target document sets are crawled from news website and standardized uniformly. Keywords are extracted from the source document firstly, and then the extracted keywords are translated and combined as query words through certain criteria to retrieve against the index created using target document set. Meanwhile, the mapping correlations between source and target documents are developed according to the value of similarity calculated by the retrieval tool. Two methods are evaluated to filter the comparable document pairs so as to ensure the quality of the comparable corpora. Experimental results indicate that our approach is effective on the construction of Chinese-English comparable corpora.

1 Introduction

Parallel corpora are key resource for statistical machine translation, in which machine learning techniques are used to learn translation knowledge. Sufficient data is necessary for the data-driven approaches to estimate the model parameters reliably. However, as Munteanu (2006) stated, beyond a few resource-rich language pairs such as English-Chinese or English-French and a small number of contexts like parliamentary debates or legal texts, parallel corpora remain a scarce resource, despite the proposition of automated methods to collect parallel corpora from the Web. Researches on comparable corpora are motivated by the scarcity of parallel corpora. Compared with parallel corpora, comparable corpora are more abundant, up-to-date and accessible.

Comparable corpora are defined as pairs of monolingual corpora selected according to the same set of criteria, but in different languages or language varieties. When creating comparable corpora, the key process is to align the source document with relevant target documents. Early work by Braschler and Scäuble (1998) employed content descriptors and publication dates to align German and Italian news stories. Resnik (1999) mined comparable corpora on the assumption that the pages which are comparable of each other share a similar structure (headers, paragraphs, etc.) when text is presented in many languages in the Web. Tao and Zhai (2005) acquired comparable bilingual text corpora based on the observation that terms that are translations of each other or share the same topic tend to co-occur in the comparable corpora at the same/similar time periods. Recently, Talvensaari et al. (2007) introduced a CLIR-based approach to align two document collections with different languages. All the target documents were indexed with Lemur. Then appropriate keywords were extracted from the source language documents and translated into the target language as query words to retrieve similar target documents.

As we know, the problems may vary with the language of documents when using CLIR-based approach to construct comparable corpora, such as keyword extraction, out-of-vocabulary keyword translation and so on. This paper is a further endeavor to CLIR-based approach for com-

This work was supported by Microsoft Research Asia.

Coling 2008: Poster Volume, pages 472–480,
Beijing, August 2010

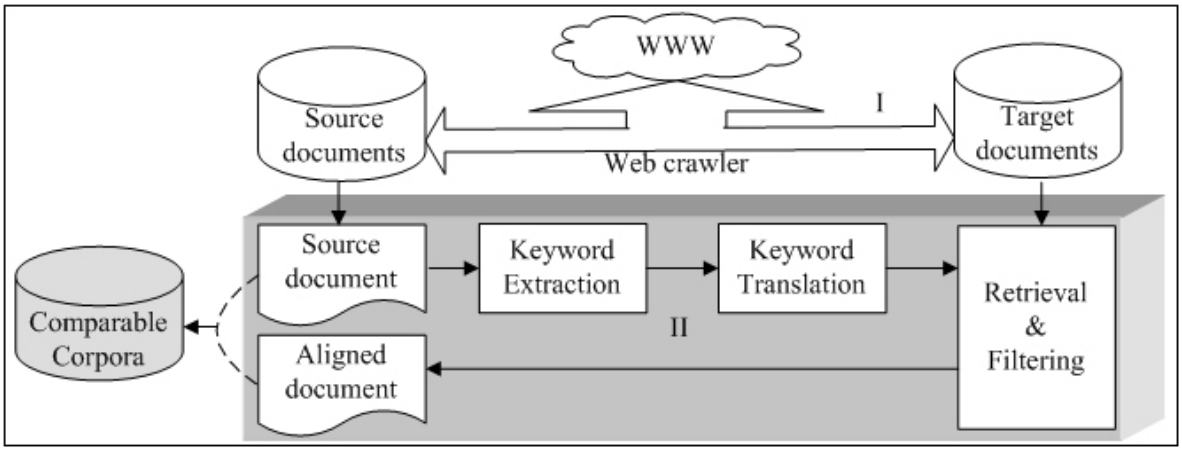

Figure 1. The general architecture of comparable corpora construction

parable corpora construction. We focus on the construction of Chinese-English comparable corpora, explore and address the issues during the construction. Experimental results show that our method is better through a rough comparison with Talvensaari et al. (2007) and it also outperforms our reconstruction of Tao and Zhai (2005) in respect to the quality of comparable corpora.

This paper is organized as follows. In section 2, the general architecture of our system is described, and each module is illuminated in detail. Section 3 reports and analyzes the experimental results followed by conclusions in section 4.

2 System Architecture

Figure 1 shows the general architecture of our comparable corpora construction system. It consists of two components: component I and component II. Component I is mainly composed by a web crawler, which is used to harvest source and target documents from selected web sites. We can get the final source and target document sets through content extraction and noise filtering. The core of the system is component II, which aligns a source document with target documents having comparable contents. It implements on the two document sets generated by component I. Component II is composed of three modules: keyword extraction, keyword translation, and retrieval & filtering. The methods for three modules are detailed respectively.

2.1 Keyword Extraction

A keyword is described as a meaningful and significant expression containing one or more words. Appropriate keywords briefly describe the theme of a document. In this paper, keywords are viewed as basic units of search indexes in order to retrieve closely related documents. Generally, phrases can capture the main idea of a document more effectively, inasmuch as they have more information than single words (an independent linguistic unit after word segmentation for Chinese).

Existing approaches for keyword extraction could be distinguished into two main categories: supervised or unsupervised methods. Supervised machine learning algorithms were widely used in keyword extraction such as Naïve Bayes (Frank et al., 1999; Witten et al., 1999), SVM (Zhang et al., 2006), CRF (Zhang et al., 2008), etc. These approaches had excellent stability. However, it was difficult for us to construct a big-enough golden annotated corpus to train a good classifier, especially for news web pages. Unsupervised methods hinged on evaluating various features to select keywords, such as word frequency (Luhn, 1957), word co-occurrence (Matsuo and Ishizuka, 2004), and TF*IDF (Li et al., 2007). The inherent problem in these methods was that most of their work came in the judgment whether a candidate was a keyword or not, but they had not paid sufficient attention to the identification of phrase candidates. Wan and Xiao (2008) proposed a method for keyphrase extraction from single document. However, it simply combined the adjacent candidate words to a multi-word phrase.

Based on the above observation, our approach for keyword extraction focuses more on the construction of phrasal candidates. It is mainly based on MWE (Multi-Word Expression) extraction together with relevant word ranking method.

MWE is a special lexical unit including compound terms, idioms and collocations, etc. The process of keyword extraction in this paper mainly depends on the following stages.

Stage 1: The generation of phrasal candidates

(1) The extraction of MWEs from the preprocessed document

Document preprocessing is a procedure of morphological analysis including segmentation and part of speech tagging for Chinese. The method based on the marginal probabilities detailed in (Luo and Huang, 2009) is adopted in this part.

We extract MWEs using LocalMaxs selection algorithm together with a relevance measure calculation method (FSCP) proposed by Silva et al. (1999). Suffix arrays and related structures in (Aires et al., 2008) are used to compute the FSCP value so as to raise efficiency. And the initial collection of MWEs named *G* for the document is generated after filtered by stopword list.

(2) The acquisition of new MWEs through the modification for segmentation

As a matter of fact, the results of segmentation for the document usually have some errors especially for out-of-vocabulary (OOV) words which are segmented to single Chinese characters in most cases. Inaccurate segmentation leads to some faults for keyword extraction. As stated in (Liu et al., 2007), OOV words can be identified by the method of MWEs extraction mentioned above. Therefore, we modify the segmentation like this: any MWE in *G* is merged to one word if it only consists of single Chinese characters and its frequency > *freq*. The changes before and after merging are shown in Table 1. Because the method of MWE extraction is based on statistical techniques, so low frequency of MWE will result in poor performance. But large value for *freq* means that very few MEWs can satisfy the frequency restriction. In our experiments, we set *freq*=2. The extraction process is called again to identify MWEs from the document with modified segmentation. Consequently, new collection of MWEs is acquired.

Additionally, some simple rules are defined according to language features to filter MWEs. In this paper, our method is tailored to extract keywords from news web pages which contain some special symmetric marks like "「, 」". The words in a specially marked area are usually important to the document. So we extract words within each paired marks and view them as a MWE on the condition that it contains two or more than two words. All of the MWEs are viewed as phrasal candidates and filtered by stopword list.

Stage 2: The generation of single words candidates

Our method also generates single word candidates with the account that both phrase and single word can be served as a keyword. The process of single word selection is independent of MWE extraction. The candidate words are restricted to nouns, verbs, strings (like WTO) and merged words as discussed in the previous stage. But the word will be removed if it only appears once in the document or is contained in the stopword list.

Stage 3: Keyword selection based on candidates ranking

As for MWE candidates, we calculate the weight for them using Formula 1 which refers to the formula used to sort NP phrases in (Bracewell et al., 2008). But the weight of *len* is reduced.

$$Weight(MWE) = \log(\sqrt{len} + f_{MWE} + \frac{1}{len} \times \sum_{i=1}^{len} tf(w_i)) \quad (1)$$

Where *len* is the length of MWE (in number of words); f_{MWE} is the frequency of the MWE within in the document; $tf(w_i)$ is the frequency of word w_i. The following rules are used to rank MWEs:

MWE	Segmentation before merging	Segmentation after merging	Pos before merging	Pos after merging
布 卡	鸟/ 人/ 布/ 卡/ 为/ 脚伤/ 所/ 苦/	鸟/ 人/ 布卡/ 为/ 脚伤/ 所/ 苦/	鸟/n 人/n 布/n 卡/n 为/vl 脚伤/n 所/us 苦/a	鸟/n 人/n 布卡/oov 为/vl 脚伤/n 所/us 苦/a
琼 丝	琼/ 丝/ 五金/ 梦/	琼丝/ 五金/ 梦/	琼/jb 丝/n 五金/b 梦/n	琼丝/oov 五金/b 梦/n

Table 1. Changes before and after merging

(a) more frequent MWEs are ranked higher; (b) MWEs with larger weight are ranked higher. In order to avoid redundancy, we remove the redundant MWEs with lower rank.

Single word candidates are ranked as follows: (a) the single word *w* with larger TF*IDF value is ranked higher; (b) the pos score for *w* in descending order is: named entity, merged words, nouns, strings, verbs. In the end, top-*a* MWEs and top-*b* single words are chosen to form the keyword set of the document.

Stage 4: Parameters evaluation and experimental results

The max number of keywords extracted from each document is limited to ten (a+b=10) and we run our approach on the dataset which include one hundred Chinese documents from the corpus of NTCIR-5 since they are also news articles. For evaluation of the results, the keywords extracted by our method are compared with the manually extracted keywords (at most ten keywords are assigned to each document). The F-measure is used as evaluation metric. It is defined like this: $F=(P+R)/2$; $P=num_{match}/num_{system}$; $R=num_{match}/num_{manual}$. Where num_{match} is the count of keywords extracted by our method matching with manually extracted keywords; num_{system} is the count of keywords extracted by our method; num_{manual} is the count of keywords assigned by human.

Figure 2 shows the performance curves for our extraction method. In this figure, *a* ranges from 0 to 10 while *b* is 10 to 0. It performs best when *a* = 4 and *b* = 6. So the two values are adopted in this paper.

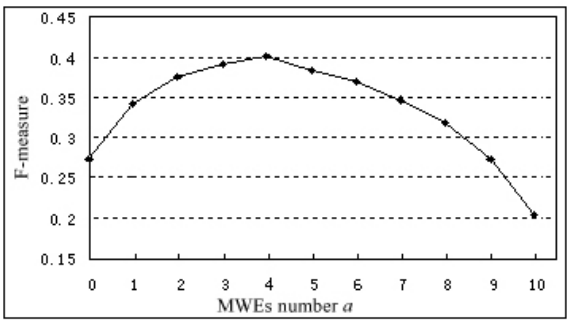

Figure 2. F-measure varies with the value of *a*

We test our approach on another dataset which also contains one hundred documents. In the experiments, the max number of keywords is set to ten. Table 2 shows the results of keyword extraction under three different conditions respectively. (A) Only extracts single words as keywords while just MWEs with (B). (C) The method presented in this paper which makes a proper combination of MWEs and single words.

	P	R	F
A (single words)	24.2%	28.5%	26.4%
B (MWEs)	18.1%	23.0%	20.6%
C (A+B)	34.2%	43.6%	38.9%

Table 2. Keyword extraction results

2.2 Keyword Translation

As for keyword translation, there are three main approaches: translation based on dictionary, parallel corpora and machine translation. Dictionary based approach is adopted in our system by taking the acquisition of translation resource into account.

Word Sense Disambiguation (WSD) and OOV problem are the main difficulties in CLIR (Cross Language Information Retrieval) task. A typical bilingual dictionary will provide a set of alternative translations for a given keyword, so how to choose the optimal translation is called Word Sense Disambiguation. Actually some keywords can not be found and translated due to the coverage limitation of a bilingual dictionary, which is called OOV problem.

In this paper, the keyword is given up if its size of translations gained from the bilingual dictionary is larger than two for the convenience of WSD. Additionally, both of the translations are treated as synonyms and equal weight is assigned to them when retrieval.

To address the OOV problem, researchers proposed methods using snippets returned by a search engine. For example, Wang et al. (2004) introduced a statistics-based approach called SCPCD to mine translations from the returned snippets. Different from (Wang et al., 2004), Zhou et al. (2007) used a pattern-based approach to analyze the mixed-languages snippets.

Leveraging on previous work, we analyze the co-occurrence mode of the OOV term and the corresponding translation in the returned snippets. Table 3 shows the typical co-occurrence modes collected during experiments, where the English words in bold are the corresponding translations of the underlined Chinese OOV terms. From Table 3, we can see the translations in number 1, 2 and 3 are included in the symmetric symbols, like bracket, quotation marks. However, the

Serial number	Segments extracted from the returned snippets
1	…原版英语论坛书名：廊桥遗梦《**The Bridges of Madison County**》作者：美国…
2	…英文影评：廊桥遗梦（**The Bridges of Madison County**）-52 影评网...
3	…具有布什特色的“牛仔外交”（**cowboy diplomacy**）反而被“现实主义”取代…
4	...用于数据挖掘的贝叶斯网络 **Bayesian Network** for Data Mining-作者：慕...
5	…以《布什牛仔外交终结》（The End of **Cowboy Diplomacy**）为题作封面故事…
6	…廊桥遗梦隐藏摘要. **The Bridges of Madison County**. Forrest Gump 阿甘正传...

Table 3. Chinese OOV and the corresponding translation in returned snippets

translations in number 4, 5, and 6 are embedded in the partial sentence while there are noise English words. In order to get the correct translation, the partial sentence needs to be segmented. By above analysis, we integrate the SCPCD method and the pattern-based method so as to extract more correct translations. The SCPCD method can be used to determine the boundaries for OOVs like number 4, 5, and 6; while pattern-based method makes use of the symmetric symbols like number 1, 2 and 3. Table 4 shows the experimental results for OOV translation methods. The average top-n inclusion rate is adopted as a metric. For a set of test OOV terms, its top-n inclusion rate is defined as the percentage of the OOVs whose translations can be found in the first n extracted translations.

	Pattern	SCPCD	Pattern + SCPCD
Top-1	40.0%	49.2%	68.1%
Top-3	41.5%	55.4%	70.2%

Table 4. The performance comparison of different OOV translation methods

The test dataset used is the Chinese topic terms in CLIR task of NTCIR-5. The search engine is Google. The bilingual dictionary used by us is LDC_CE_DICT 2.0. And we only adapt the pattern with symmetric symbols, which has the highest precision proposed by Cao et al. (2007).

2.3 Retrieval and Filtering

The process of retrieval is to construct the alignment relationship between source and target document pairs. It is a core module in our system since the quality of comparable corpora is greatly influenced by alignment level which depends on the relevance between document pairs. Our intention here is to retrieve high relevant target documents for the source documents. Open-source toolkit Indri is introduced to assist the retrieval process. Indri is a part of the Lemur project[1]. On the basis of Lemur, it combines inference networks with language modeling. And it's widely adopted by institution for scientific research since it is effective, flexible, usable and powerful. So it is employed by us to retrieve related documents. A query for each source document is formed by the translated keywords with Indri query language and then run against the target collection.

The essential of alignment is to compare the similarity between source and target document pairs. In order to reduce the workload of comparing, Pooling method is applied to assist the comparing process. We choose the top r documents returned by Indri retrieval system to build the related document pool. And g ($g<=r$) documents in the pool are selected to form the alignment document pairs together with the source document. In our experiments, we set r=10 and g=1.

In the process of alignment, three features are used to filter the alignment pairs for the sake of pruning the low relevant pairs. The first is publication date contained in documents. The second is similarity calculated by Indri between the query and the target document when retrieval. The last is KSD (Keyword similarity between document pairs) which is defined by our system. In this paper, we propose two methods to filter the alignment pairs by using various features.

(1) DSF filtering

This method depends on two features: date and similarity. At first, we give a priority to the target documents that have the closest date to the source document during the top-r documents searching. A date-window size d is defined to measure the date difference. We set d=1 in this paper. That is to say, the target documents with

[1] Lemur toolkit is developed by Carnegie Mellon University and University of Massachusetts. The open source code is available at http://www.lemurproject.org.

exactly the same date as the source document, and one day earlier or later are considered to be closest. Then, we select *g* documents with larger similarity from the related document pool. Finally, we rank all of the alignment pairs with the score of similarity and set a similarity threshold *s* to filter further. It should be noted that there are $n \cdot g$ alignment pairs, where *n* is the number of source documents having non-empty related document pool.

(2) DSKF filtering

This method utilizes all of the features: date, similarity and KSD. As for KSD, it integrates two factors. One is NTK, namely the number of translated keywords appeared in the target document, since the target document is more similar to the source document as increasing of NTK. The other is FIS, namely frequency information score. Inspired by paper (Tao and Zhai, 2005), we use the score of FIS to measure the correlations between the keywords in source document and translated keywords in target document which represent the matching for source and target document pair. We define d_s as the source document, d_t as the target document, *ks* as the set of keywords extracted from d_s, *kts* as the set of translated keywords. Formula 2 is used to compute the score of FIS:

$$Score_{FIS} = \sum_{i=1}^{ktsLen} (BM25(x_i, d_s) \cdot IDF(x_i) \cdot BM25(y_i, d_t) \cdot IDF(y_i) / norm(Dif(x_i, y_i))) \quad (2)$$

Where, *ktsLen* is the size of *kts*, y_i is an element in *kts*, x_i is the element in *ks* while y_i is the translation of x_i. Moreover, *BM25*(*w*, *d*) is the normalized frequency of word *w* in document *d*. It has been considered as one of the most effective matching functions for retrieval. *IDF* stands for Inverse Document Frequency which is also commonly used in information retrieval. *Dif*(*x, y*) is defined as the difference between *BM25*(*x*, d_s) and *BM25*(*y*, d_t). Formula 2 penalizes large difference due to the conditions like this: any keyword in source document appears many times while its translation appears rarely in target document. The process of its normalization is run by Formula 3 which makes the score less sensitive to the absolute value:

$$norm(score) = \begin{cases} 1, & score < 1 \\ \sqrt{score}, & else \end{cases} \quad (3)$$

Furthermore, the final KSD score is got by simply adding the normalized scores of NTK and FIS which are dealt with Formula 3. Actually, the two filtering methods differ principally in the last step. DSKF sorts all of the alignment pairs according to the KSD score while it is similarity in DSF. We also set a KSD threshold *k* for DSKF method to filter further. The values for *s* and *k* will be investigated in the following experiments.

3 Experiments

In this section, we first introduce how to acquire the source and target document sets. Then our system is tested on the two sets. The experimental results are reported and analyzed finally.

3.1 Experiment Setup

To test the effectiveness of the proposed system, large-scale of Chinese and English news web pages are crawled respectively from XinHuaNet and used as the document resource. The reasons for choosing news pages are:

(1) Many websites, like portal website, news agency, government and so on, provide large-scale news reports. At the same time, a large proportion of the reports can be crawled politely, so document acquisition is relatively easy.

(2) The news pages include various contents, such as politics, economy, sports, so the corpora made up of news pages can avoid the limitations of domain-specific corpora.

All the news pages are processed uniformly. The core content of each web page crawled is extracted and several tags describing the headline and publication date are added. Meanwhile, the original contents are kept with no change. Table 5 shows the basic information of document sets.

Year	Number of source documents	Number of target documents
2003	23747	3390
2004	25660	2943
2005	47333	11578
2006	28572	25320
2007	25036	25247
2008	14021	24292
2009	7476	10887
Total	171845	103657

Table 5. The composition of source document set and target document set

3.2 Results and Discussion

The quality of comparable corpora highly de-

pends on the alignment level between source and target document pairs. Braschler and Scäuble (1998) used five levels of relevance to assess the alignments as follows:

(1) Same story. The two documents deal with the same event.

(2) Related story. The two documents deal with the same event or topic from a slightly different viewpoint. Alternatively, the other document may concern the same event or topic, but the topic is only a part of a broader story or the article is comprised of multiple stories.

(3) Shared aspect. The documents deal with related events. They may share locations or persons.

(4) Common terminology. The events or topics are not directly related, but the documents share a considerable amount of terminology.

(5) Unrelated. The similarities between the documents are slight or nonexistent.

We randomly select 500 source documents published in 2009 as the test dataset. Experiments with different parameters are constructed based on this dataset. The quality of each alignment pair is manually assessed using the five-level relevance as discussed above. What should to be pointed out is that parameter s and k are not absolute values, but percentile rank level in our work. For instance, k = 10 means that we only choose the alignment pairs whose KSD score rank in top ten percent among all of the results.

Table 6 shows the results filtered by DSF method with different values of s ($s_1 < s_2 < s_3 < s_4$). Table 7 shows the results filtered by DSKF method with various values of k ($k_1 < k_2 < k_3 < k_4$). In order to evaluate the results conveniently, two standards are established: (a) the number of high relevant pairs created, which is the count of document pairs in Level 1 and 2; (b) the quality of the whole alignments, that is to say the percentage of alignment pairs with Level 1 and 2. Seen from Table 6 and 7, DSKF is better than DSF by considering the two standards. Compared with DSF, more high relevant pairs are left filtered by DSKF when they have the same total number of pairs. In other words, the DSKF method is more powerful to make high relevant pairs in higher rank so as to reduce alignment pairs which are rarely relevant. Therefore, DSKF is adopted in our system. Taking the first criterion into account, we give up the parameter k_1, k_2. Parameter k_4 is not the best considering the second criterion. Ultimately, k_3 is chosen as the final value for k. At this point, the number of alignment pairs in Level 1 and 2 is close to the maximum. Meanwhile, the percentage of high alignments reaches 68.5%.

Among the surveyed related work, Talvensaari et al. (2007) created Swedish-English comparable corpora based on CLIR techniques and its framework of construction is similar to ours. However, the two systems are different in the following aspects:

Level	s_1 = 10		s_2 = 30		s_3 =50		s_4 =70	
	Number	%	Number	%	Number	%	Number	%
Leve1 1	23	46.9%	54	36.5%	83	33.5%	96	27.7%
Level 2	18	36.7%	43	29.1%	62	25.0%	81	23.3%
Level 3	4	8.2%	21	14.2%	40	16.1%	57	16.4%
Level 4	4	8.2%	19	12.8%	41	16.5%	60	17.3%
Level 5	0	0.0%	11	7.4%	22	8.9%	53	15.3%
Total	49	100%	148	100%	248	100%	347	100%

Table 6. The distribution results filtered by DSF with different s parameters

Level	k_1 = 10		k_2 = 30		k_3 = 50		k_4 =70	
	Number	%	Number	%	Number	%	Number	%
Level 1	33	67.3%	78	52.7%	93	37.5%	98	28.2%
Level 2	15	30.6%	52	35.1%	77	31.0%	89	25.6%
Level 3	1	2.0%	9	6.1%	37	14.9%	62	17.9%
Level 4	0	0.0%	9	6.1%	34	13.7%	60	17.3%
Level 5	0	0.0%	0	0.0%	7	2.8%	38	11.0%
Total	49	100%	148	100%	248	100%	347	100%

Table 7. The distribution results filtered by DSKF with different k parameters

(1) The language is different. We focus on building comparable corpora of Chinese-English while they were Swedish-English.

(2) A series of sub problems are different due to language difference. As for keyword extraction, we propose a method to select both key phrases and single words, while they used RATF (Relative Average Term Frequency) method. For OOV problem, we combine the SCPCD method with the pattern-based method to extract OOV translations from snippets returned by a search engine. However, the classified s-gram matching technique was utilized by Talvensaari et al. (2007) to translate OOV words.

(3) Talvensaari et al. (2007) filtered their alignment pairs mainly depending on date and similarity, while we introduce new feature KSD to extend the original feature set.

Talvensaari et al. (2007) also randomly chose 500 source documents and assessed the quality of alignments using the same five-level relevance.

In addition to this, we implement the method of Tao and Zhai (2005) which is a purely statistical-based and language independent approach. The source and target documents published in 2009 are employed to test the method. The same sample as our system including 500 Chinese documents is chosen to make a further comparison with our work. We align each source document with one target document through the BM25Corr model in (Tao and Zhai, 2005). The alignment pairs are ranked according to mapping scores calculated by the BM25Corr model. And we select the top N (N = 248) alignment pairs for the benefit of comparison.

Table 8 shows the distribution results for the three systems. As illustrated in Table 8, we can roughly conclude that our approach creates more alignment pairs with the same number of source documents when compared with Talvensaari et al. (2007). Meanwhile, the percentage of high relevant document pairs is larger.

Likewise, our system outperforms BM25Corr in that it aligns more high relevant documents pairs when they use the same sample of test corpora and create the same total number of pairs. Obviously, the quality of comparable corpora gained by our system is better than BM25Corr.

All the experimental results and analysis mentioned above indicate that our method is effective to create alignment pairs. Up to now, both the source and target documents published in 2007-2009 years are used to build comparable corpora through our proposed system. It includes 23102 alignment pairs after filtered by DSKF.

Level	Talvensaari et al. (2007)		Our System (DSKF filtering)		BM25Corr (Top N = 248)	
	Number	%	Number	%	Number	%
Level 1	21	21.6%	93	37.5%	1	0.4%
Level 2	20	20.6%	77	31.0%	2	0.8%
Level 3	33	34.0%	37	14.9%	3	1.2%
Level 4	19	19.6%	34	13.7%	5	2.0%
Level 5	4	4.1%	7	2.8%	237	95.6%
Total	97	100%	248	100%	248	100%

Table 8. The distribution results for Talvensaari et al. (2007), Our System, and BM25Corr

4 Conclusions

In this paper, we propose a CLIR-based approach to create large-scale Chinese-English comparable corpora. Firstly, we harvest the original source and target document sets from news website using open-source crawler. Then the core content of each document is extracted through discriminating noise contents. Next, we delve into the approaches of problems such as keyword extraction and OOV translation followed by the process of retrieval to develop mapping correlations between source and target documents. Finally, three features as publication date, similarity score and KSD value are used to filter the aligned document pairs. Experimental results show that our approach is effective to mine Chinese-English document pairs with comparable contents. In the future, we will optimize the approach for every module in the construction of comparable corpora for the sake of improving the performance of the whole system. What's more, it will be worth consideration to mine mappings between terms which can be served as a feature for the process of developing mappings between document pairs in turn.

References

Aires, José, Gabriel Lopes, and Joaquim Ferreira Silva. 2008. Efficient Multi-word Expressions Extractor Using Suffix Arrays and Related Structures. In *Proceeding of the 2nd ACM workshop on Improving non english web searching*, pp. 1-8.

Bracewell, David B., Fuji Ren, and Shingo Kuroiwa. 2008. Mining News Sites to Create Special Domain News Collections. *International Journal of Computational Intelligence*, 4(1): 56-63.

Braschler, Martin, and Peter Scäuble. 1998. Multilingual Information Retrieval Based on Document Alignment Techniques. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, pp. 183-197.

Cao Guihong, Jianfeng Gao, and Jianyun Nie. A System to Mine Large-Scale Bilingual Dictionaries from Monolingual Web pages. 2007. In *Proceedings of Machine Translation Summit XI*, pp. 57-64.

Frank, Eibe, Gordon W. Paynter, and Ian H. Witten. 1999. Domain-Specific Keyphrase Extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp. 668-673.

Li Juanzi, Qi'na Fan, and Kuo Zhang. 2007. Keyword Extraction Based on tf/idf for Chinese News Document. *Wuhan University Journal of Natural Sciences*, 12(5): 917-921.

Liu Tao, Bingquan Liu, Xiaolong Wang, and Minghui Li. 2007. The Effectiveness Study of Local Maximum Feature for Chinese Unknown Word Identification. *Journal of Chinese Language and Computing*, 17(1): 15-26.

Luhn, Hans Peter. 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4): 309-317.

Luo Yanyan, and Degen Huang. 2009. Chinese Word Segmentation Based on the Marginal Probabilities Generated by CRFs. *Journal of Chinese Information Processing*, 23(5): 3-8.

Matsuo, Yutaka and Mitsuru Ishizuka. 2004. Keyword Extraction from a Single Document Using Word Co-occurrence Statistical Information. *International Journal on Artificial Intelligence Tools*, 13(1): 157-169.

Munteanu, Dragos Stefan. 2006. Exploiting Comparable Corpora. *Doctoral Thesis*. UMI Order No.3257825. University of Southern California.

Resnik, Philip. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 527-534.

Silva, Joaquim Ferreira, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. 1999. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence*, pp. 113-132.

Talvensaari, Tuomas, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola and Heikki Keskustalo. 2007. Creating and Exploiting a Comparable Corpus in Cross-Language Information Retrieval. *ACM Transactions on Information Systems*, 25(1):1-21.

Tao Tao, and Chengxiang Zhai. 2005. Mining Comparable Bilingual Text Corpora for Cross-Language Information Integration. In *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 691-696.

Wan Xiaojun, and Jianguo Xiao. 2008. CollabRank: Towards a Collaborative Approach to Single-Document Keyphrase Extraction. In *Proceeding of the 22nd International Conference on Computational Linguistics*, pp. 969-976.

Wang Jenq Haur, Jie Wen Teng, Pu Jen Cheng, Wen Hsiang Lu, and Lee Feng Chien. 2004. Translating Unknown Cross-Lingual Queries in Digital Libraries using a Web-based Approach. In *Proceedings of the 4th ACM/IEEE-CS joint Conference on Digital Libraries*, pp. 108-116.

Witten, Ian H., Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. KEA: Practical automatic keyphrase extraction. In *Proceedings of the 4th ACM Conference on Digital Libraries*, pp. 254-255.

Zhang Chengzhi, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, and Bo Wang. 2008. Automatic Keyword Extraction from Documents Using Conditional Random Fields. *Journal of Computational Information Systems*, 4(3): 1169-1180.

Zhang Kuo, Hui Xu, Jie Tang, and Juanzi Li. 2006. Keyword Extraction Using Support Vector Machines. In *Proceedings of the 7th International Conference on Web-Age Information Management*, pp. 85-96.

Zhou Dong, Mark Truran, Tim Brailsford, and Helen Ashman. 2007. NTCIR-6 Experiments Using Pattern Matched Translation Extraction. In *Proceedings of 6th NTCIR Workshop Meeting*, pp. 145-151.