

PAPER

Creating Chinese-English Comparable Corpora

Degen HUANG^{†a)}, Shanshan WANG^{†b)}, Nonmembers, and Fuji REN^{††}, Member

SUMMARY Comparable Corpora are valuable resources for many NLP applications, and extensive research has been done on information mining based on comparable corpora in recent years. While there are not enough large-scale available public comparable corpora at present, this paper presents a bi-directional CLIR-based method for creating comparable corpora from two independent news collections in different languages. The original Chinese document collections and English documents collections are crawled from XinhuaNet respectively and formatted in a consistent manner. For each document from the two collections, the best query keywords are extracted to represent the essential content of the document, and then the keywords are translated into the language of the other collection. The translated queries are run against the collection in the same language to pick up the candidate documents in the other language and candidates are aligned based on their publication dates and the similarity scores. Results show that our approach significantly outperforms previous approaches to the construction of Chinese-English comparable corpora.

key words: *Comparable Corpora, cross language information retrieval, keyword extraction, document alignment*

1. Introduction

Multilingual corpora, either parallel or comparable, are invaluable resources for cross-language information retrieval (CLIR) and many other linguistic studies, such as Discourse Analysis and Pragmatics, Terminology Extraction and Knowledge Engineering. Parallel corpora are usually obtained from official multilingual documents such as United Nations articles and EU documents or news services [1]. Such collections usually have a quite limited domain covering a limited number of languages apart from the cost and the time consuming work of creating parallel corpora, and the lack of sufficient parallel data for various languages and domains is currently one of the major obstacles to further linguistic research. On the other hand, comparable corpora are generally obtained from news articles [2], [3], available research corpora such as CLEF collections [4] or by crawling the Web [5], [6]. Such corpora can be collected easily by downloading electronic copies of newspapers, journals, articles, etc., from the Web. Compared with parallel corpora, comparable corpora are richer, more up-to-date and more accessible. Furthermore, as the EAGLES (<http://www.ilc.pi.cnr.it/EAGLES96/browse.html>) report

said, the possibilities of a comparable corpus are to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus. However, the currently available comparable corpora are either built for special purposes or not big enough for translation purposes. Because of the shortcomings, the creation and use of comparable corpora become a feasible and popular choice. Obviously, it's much easier to find text sets with similar topics than to find collections which are translations of each other. Considerable research on the construction of comparable corpora has achieved satisfactory results over the years.

Braschler and Scauble [7] employed content descriptors, publication dates, subject codes, common proper nouns, numbers and a small dictionary to align German and Italian news stories. Resnik [8] mined comparable corpora on the assumption that the pages presented in numerous languages on the Web are comparable to each other which share a similar structure such as headers, paragraphs and hyper links. Tao and Zhai [9] 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 or similar time periods. Talvensaari et al. [4] introduced a CLIR-based approach to align two document collections with different languages, Swedish and English. They extracted the query keywords from documents of Swedish collection using the relative average term frequency formula (RATF), and translated them into English. The English queries were run against the English collection and documents were aligned according to the similarity and publication dates. Their method is most akin to Braschler's method [7]. Rahimi and Shakery [10] used the similarity of the document topics and their publication dates to align the documents in their collections to build a Persian-English comparable corpus. Huang et al. [11] explored a CLIR-based approach to constructing Chinese-English comparable corpora, and the system contains three modules: keyword extraction, keyword translation, and retrieval and filtering.

Among the methods mentioned above, Talvensaari's [4], Rahimi's [10] and Huang's [11] systems share the same architecture. And we follow the general procedure proposed by them to create our comparable corpus. Nevertheless, there are many important differences between their methods and ours. The biggest difference is that they only choose one language (language I) as the source

Manuscript received November 1, 2012.

Manuscript revised March 21, 2013.

[†]The authors are with Dalian University of Technology, Dalian City, China.

^{††}The author is with the University of Tokushima, Tokushima-shi, 770-8501 Japan.

a) E-mail: huangdg@dlut.edu.cn

b) E-mail: xiaolanshan@mail.dlut.edu.cn

DOI: 10.1587/transinf.E96.D.1853

language and another one (language II) as the target language to construct their comparable corpora, while we also choose language II as the source language, with language I its target language. Therefore, we extract the keywords of one language as well as accomplish the same treatment on the other language. Another difference is the preprocessing of the source documents. Talvensaari et al. [4] utilize the TWOL lemmatizer program to lemmatize inflected source document words and to decompose compound words, and Rahimi and Shakery don't make any preprocessing on their source collection. The preprocessing of Huang's method includes segmentation and POS tags for Chinese using the method based on the marginal probabilities [12]. However, besides the process on Chinese texts, in this paper, the English texts are tokenized by POS tags by postagger[†] and stemmed by Porter stemmer^{††}. Furthermore, they make one-to-one aligned document pairs, while the number of target documents correlated with a source document is depended on the similarity scores of the target candidate documents and the source document. There are also some major differences in the approach of keyword extraction. The details will be discussed in the remainder of this paper.

The rest of this paper is organized as follows. Section 2 describes the main idea of comparable corpora construction. Experiments and results are reported and discussed in Sect. 3. And conclusion is made in Sect. 4.

2. Comparable Corpora Construction

Generally, documents with similar contents which are published on the same or similar dates most probably share the same event or topic. And the contents of the text can be described by the meaningful, appropriate, significant and corresponding words, namely keywords. In other words, we can extract the keywords to represent the core content of the document. So to create our comparable corpora, we first select the keywords for each Chinese document and English document. Then, the Chinese keywords are translated into English queries; simultaneously, the English keywords are translated into Chinese queries. After that, the translated English queries are run against the English document collections and the translated Chinese queries are run against the Chinese document collections to pick up the candidate documents in another language. Finally, to improve the quality of aligned document pairs in different languages, the document pairs which mismatch the given date window or whose similarity scores are below the given threshold will be abandoned.

2.1 Chinese Keywords Extraction

Lots of studies on keywords extraction have been made in recent years, and many approaches have been proposed including supervised machine learning algorithms and unsupervised methods. Native Bayes [13], decision trees [14] and SVM [15] are representative supervised methods which achieve satisfying accuracy and have excellent stability.

But, it's difficult for a large-enough corpus in which each document is annotated with keywords in advance to train a good classifier. TFIDF measure [16], [17] is a very popular method, which extracts keywords that frequently appear in a document, but don't appear frequently in the remainder of the corpus. Matsuo and Ishizuka [18] developed a word co-occurrence method to extract keywords from a single document, and achieved high performance comparable to TFIDF. Yang et al. [16] combined TFIDF and word co-occurrence features together to solve this problem. Wan and Xiao [19] proposed a method for key-phrase (units consisting of more than one word) extraction. However, it simply combined the adjacent candidate words to a multi-word phrase. Huang et al. [11] introduced a method based on multi-word expressions (MWE) extraction in combination with single word ranking method to extract keywords, but their approach focuses more on the construction of phrasal candidates. From the above approaches we can find that extracting keywords using TFIDF is the basic method and combination strategies are better than the individual methods.

We, therefore, propose a method to extract the keywords from the documents which combines key-phrases and key-words (units containing one word only) together. The selection of key-words is essential as well as key-phrases construction. The steps of keyword extraction are described as follows in detail. The initial Chinese and English document collections are crawled from XinHuaNet respectively. Thenceforward, the html pages are standardized uniformly after the discriminating noise is removed such as ads and pictures.

Step 1: preprocessing

The normalized texts are segmented and labeled with POS tags with the system proposed by Luo and Huang [12]. However, it's inevitable that there are numerous segment fragments in the processed results. Especially, some new words or other out-of-vocabulary words are more likely to be segmented to single Chinese characters. And incorrect segmented results will generate some inappropriate keywords. In order to solve the problem, we extract the strings consisting of single characters, and revise the string to be one word which appears more than *freq* times in the text and meets one of the following conditions:

- A. The string is composed of single Chinese characters without stopword, such as “克莉丝特丝”, “巴米安”.
- B. The string is an abbreviation which consists of capital letters and single Chinese characters or Arabic numerals, e.g. “G 八”, “G 20”.
- C. The string is an abbreviation in the form of “EP -3”, “F -14”.

Here, *freq* is an integer more than 1 and its value is determined by the experiments. Table 1 shows the changes before and after the string is modified.

Step 2: extracting the candidate key-phrases

Silva et al. [20] used LocalMaxs algorithm together

[†]<http://nlp.stanford.edu/software/tagger.shtml>

^{††}<http://tartarus.org/~martin/PorterStemmer/>

Table 1 Changes before and after modification.

	Before modified	After modified
String	摧毁 巴米安 佛像	摧毁 巴米安 佛像
POS tags	摧毁/v 巴/jb 米/n 安/Ag 佛像/n	摧毁/v 巴米安/mwe 佛像/n
String	出席 冲绳 G 八 会议	出席 冲绳 G 八 会议
POS tags	出席/v 冲绳/ns G/nx 八/j 会议/n	出席/v 冲绳/ns G八/mwe 会议/n
String	EP-3 高科技 侦察机	EP-3 高科技 侦察机
POS tags	EP/nx -3/m 高/a 科技/n 侦察/v 机/Ng	EP-3/mwe 高/a 科技/n 侦察/v 机/Ng

with a relevance measure calculation method (Symmetric Conditional Probability. SCP) for the extraction of multi-word lexical units and acquired a satisfactory result. In this paper, we adopt the method to select the candidate key-phrases. Beyond that, news web pages usually contain some symmetric labels, like “<” and “>”, “[” and “]”, which always indicate important meaningful units, such as “[勇往 号]”. For this reason, the strings which contain more than one word are extracted to be candidate key-phrases between the symmetric labels. Then all the key-phrases containing stopword(s) or those with an illegal grammar structure, such as the phrase “人士 参与” which consists of a noun and a verb, are removed. Consequently, the collection named C_{kp} of candidate key-phrases is obtained. Finally, the candidates are ranked in descending sort according to the value calculated by Formula 1 which refers to the method adopted to sort MWEs by Huang et al. [11]. But we increase the importance of the frequency of the candidate and take the location of the candidate in the document into account. In order to avoid redundancy of data, we remove the redundant candidates with lower rank, but if the lower one appears two more times than the higher one, the higher is removed. For example, the scores of “伦敦 金融 峰会” and “金融 峰会” are 0.4 and 0.3 respectively in one document, and “伦敦 金融 峰会” appears 2 times, if “金融 峰会” appears more than 4 times, “伦敦 金融 峰会” would be removed from the candidate key-phrase set, else “金融 峰会” would be removed. Intuitively, the word which appears in the title is more important than the word which only appears in the main body.

$$Weight(kp) = Title(kp) + f_{kp} + \log(\sqrt{len} + f_{kp}) \quad (1)$$

$$Title(kp) = \begin{cases} 1, & \text{when } kp \text{ in } Title \\ 0, & \text{else} \end{cases} \quad (2a)$$

$$(2b)$$

Here, kp refers to key-phrase; len is the number of words that the key-phrase contains; f_{kp} indicates the frequency of the kp in the document; $Title(kp)$ is the location weight of the kp , if the kp is in the title $Title(kp)$ equals 1, otherwise $Title(kp)$ equals 0; $Title$ is the title of the document.

Step 3: extracting the candidate key-words

According to the analysis of CSSCI[†] keywords, nouns account for a large proportion of all keywords, and verbs are in the second place. We can also get much meaningful information from the named entities (NEs) in the news reports besides nouns. Hence, the part-of-speeches (POSS) of candidate key-words are limited to NEs, nouns, English strings, mwes (modified words mentioned in Step 1) and

verbs. Meanwhile, we set a score to each POS, the score of NE is 5, the score of noun is 4, the score of English string is 3, the rest may be deduced by analogy according to this POSSs sequence. Similar to the extraction of the candidate key-phrases, the word in the stopword list will not be taken into account.

We use the TFIDF method integrated with other features to improve the performance to measure the importance of the candidate key-words of the documents.

As mentioned in the previous chapter, the location of a word is an important factor worth considering. And we calculate the location weight of the candidate key-word using Formula 3. The value of α is investigated in the following experiments.

$$k_{isTitle} = \begin{cases} \alpha, & word \subseteq Title \\ 1.0, & word \not\subseteq Title \end{cases} \quad (3a)$$

$$(3b)$$

In addition to the words themselves which appear in the title, the words that have a high correlation with them are worthy of consideration too. χ^2 test was adopted to measure the co-relationship of the candidate and the title by [16]. If probability distribution of co-occurrence between the candidate and the title words is biased to the unconditional distribution of occurrence of title words, the candidate is likely to be more important. Obviously, the document is composed of sentences, and a sentence consists of words. Therefore, if two words appear in one sentence, we consider that they co-occur once. The statistical value of χ^2 is defined as:

$$\chi^2(kw) = \sum_{t \in Title} \frac{(freq(kw, t) - n_{kw}p_t)^2}{n_{kw}p_t} \quad (4)$$

Where kw means key-word; t is a word in the title; $freq(kw, t)$ is the co-occurrence frequency of kw and t ; n_{kw} is the sum of co-occurrence frequencies of kw and each t in the title; p_t is the probability of t .

Generally speaking, in Chinese, the longer a word is, the more information it carries. And the candidate is set a length weight calculated by Formula 5. Where $word_{Len}$ is the length of the candidate; max_{Len} is the biggest length of all candidates in the document.

$$k_{Len} = word_{Len} / max_{Len} \quad (5)$$

Finally, for each candidate, the weight is measured by Formula 6. And the candidates will be ranked in a descending sort by their weights.

[†]http://www.cssci.com.cn

$$Weight(word) = TFIDF * (k_{isTitle} + k_{POS} + k_{Len} + k_{\chi^2}) \quad (6)$$

$$k_{POS} = Score_{POS} / max_{POS} \quad (7)$$

$$k_{\chi^2} = \chi^2(word) / max_{\chi^2} \quad (8)$$

Here, k_{POS} , k_{χ^2} are normalized by the largest value of POS score and χ^2 value of the candidate respectively; $Score_{POS}$ is the POS score of the candidate; $\chi^2(word)$ is the χ^2 value of the candidate.

With this, the set of ordering candidate key-words is obtained. The set is named C_{kw} .

Step 4: selecting the keywords

Since the process of the extraction of key-phrases is independent upon that of the extraction of key-words, there are many repeat strings, such as “金融峰会” and “金融”, in the two sets, C_{kp} and C_{kw} . For this reason, one of the repeat pairs should be removed. Here, the longer one is remained.

In the end, we choose the top-4 key-phrases in C_{kp} and top-6 key-words in C_{kw} as the keywords of the document.

2.2 English Keywords Extraction

Theoretically, different from Chinese, there are natural separators between English words. Moreover, English words have the morphological changes. Nonetheless, English news articles share some similar features with the segmented Chinese documents, such as TFIDF, location, POS and co-occurrence features described in Step 3 of 2.1. Hence, Formula 9 is applied to calculate the weight of the English word.

$$Weight(word) = TFIDF * (k_{isTitle} + k_{POS} + k_{\chi^2}) \quad (9)$$

The whole processes of the extraction of English article keywords are:

1. Preprocessing: POS tagging the English texts and stemming the English words.
2. Selecting the candidate words: Words appearing only once or in the stopword list are removed. Meanwhile, calculating the location weight $k_{isTitle}$, POS score k_{POS} and the co-occurrence feature k_{χ^2} .
3. Calculating the weights of all candidate words and ranking them in a descending order.
4. Outputting the top-10 candidates as the keywords extracted from the English document.

2.3 Translation and Alignment

After getting the source queries, we need to translate them into the target language. We use an English-Chinese dictionary composed of *LDC_CE_DICT2.0*[†] and *CEDICT*^{††}. For the Chinese words not in the dictionary, we translate them using the search engine-based method [11] and translate the English words missing from the dictionary using Google translator^{†††}.

The translated queries are run against the target collection, and the documents are aligned by the publication

dates and the values of the similarity. We define the date window 1, that is, only the target news reports published on the same day or one day around the publication date of the source news are taken into consideration. As Huang et al. [11] proved, the keyword similarity between document pairs (KSD) performs better than the similarity calculated by Indri^{††††}. KSD integrates two factors: the number of translated keywords appeared in the target document (NTK) and frequency information score (FIS) as Formula 10 describes.

$$FIS = \sum_{i=1}^{ktsLen} \frac{BM25(x_i, d_s) * IDF(x_i) * BM25(y_i, d_t) * IDF(y_i)}{norm(Dif(x_i, y_i))} \quad (10)$$

$$norm(score) = \begin{cases} 1, & score < 1 \\ \sqrt{score}, & score \geq 1 \end{cases} \quad (11a)$$

$$(11b)$$

Where, d_s is the source document; d_t is the target document; k_s is keywords set of d_s ; k_{ts} is the set of translated keywords; $ktsLen$ is the size of k_{ts} ; x_i is the i_{th} element in k_s ; y_i is the translation equivalent of x_i ; $BM25(w, d)$ is the normalized frequency of word w in document d . $Dif(x, y)$ is the absolute difference between $BM25(x, d_s)$ and $BM25(y, d_t)$. Before the calculation of the KSD value, the features NTK and FIS would be normalized by Formula (11a) and (11b), then $norm(NTK)$ and $norm(FIS)$ are obtained. And the KSD value is got by the following formula:

$$KSD = norm(NTK) + norm(FIS) \quad (12)$$

We use Formula 12 to measure the similarity between the source document and the target document. The top-N target documents returned by Indri and ranked by the similarity calculated by Formula 12 are considered as the candidate documents. Finally, we filter the document pairs whose KSD value is below the threshold which we name *sim*. In other words, the document pair would be removed from all candidate document pairs when it's KSD value is less than *sim*. The values of *top-N* and *sim* will be explored through the following experiments.

3. Experiments and Results

3.1 Experiments of Keywords Extraction

In this section, we first introduce the evaluation criteria of the experimental results and the source of our test dataset. Based on the set, our system is tested. The experimental results are reported and analyzed finally.

3.1.1 Experiment Setup

For evaluating the results, we compare the keywords extracted by our system with the manually labeled keywords.

[†]<http://www ldc.upenn.edu/>

^{††}<http://www.mandarintools.com/cedict.html/>

^{†††}<http://translate.google.com/>

^{††††}<http://www.lemurproject.org/>

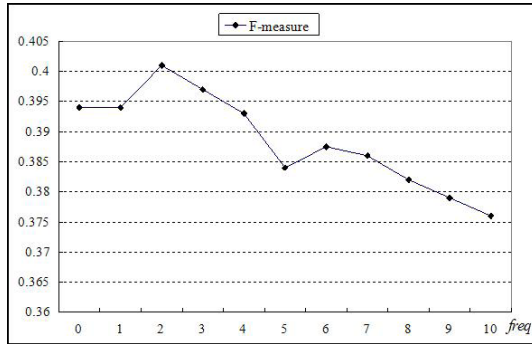


Fig. 1 F-measure varies with the value of $freq$.

The metrics of Precision (P), Recall (R) and F-measure (F) are used to evaluate the performance of our system.

$$P = \frac{N_{correct}}{N_{system}} * 100\% \quad (13)$$

$$R = \frac{N_{correct}}{N_{manual}} * 100\% \quad (14)$$

$$F = \frac{2 * P * R}{P + R} \quad (15)$$

Here, $N_{correct}$ is the number of correct keywords extracted matching with the manual results; N_{system} is the count of all extracted keywords; N_{manual} is the number of all manual keywords.

To evaluate performance, we test our system against a collection including 100 documents selected randomly from the corpus of NTCIR-5 since they are also news reports and are demonstrated by several subject words. Before that, we choose another 100 documents randomly from NTCIR-5 for estimating the optimal parameters. The two document collections are named testing set and training set respectively. For each document, the keywords which represent the essential contents have been manually assigned.

3.1.2 Results and Discussion

As the previous section mentioned, the parameters, $freq$ and α , exert influences on the results of keywords extraction, so it is worth estimating the values of the parameters to produce the best effect. Here, the experiments are worked on the training set mentioned on the last section. Figure 1 shows the results of keyword extraction under different values of $freq$ respectively. As can be seen from Fig. 1, when $freq$ equals 2, the system performs best. So we set the variable $freq$ to 2 and the next step, we will test the effect of α in the results of the experiments, and we can see the F-measure varies with the value of α from Fig. 2. From the curve of F-measure in Fig. 2, we can conclude that when α is equal to 1.2 or 1.3, F-measure reaches the maximum. Here, we set α to 1.3.

We also validate the impacts of different features on the performance of our system. Figure 3 shows the different results of different methods. Where, base means the basic TFIDF method, not considering any other features; T means

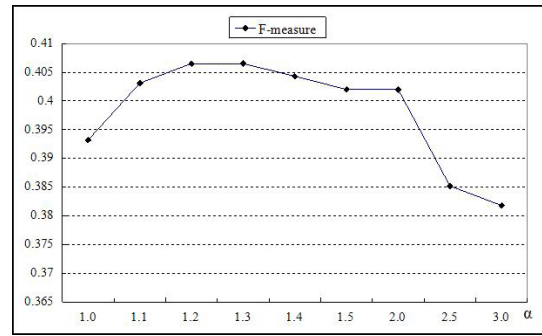


Fig. 2 F-measure varies with the value of α .

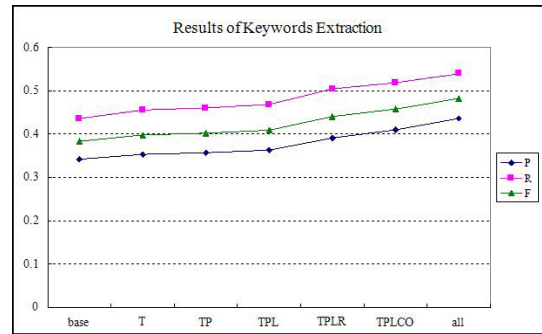


Fig. 3 Results of keywords extraction vary with different methods.

Table 2 Results of keywords extraction.

	P	R	F
only-word	39.01%	42.19%	40.54%
only-phrase	36.89%	19.35%	25.38%
word+phrase	47.24%	52.91%	49.92%
Ref. [11]	40.93%	47.09%	43.79%

the method which combines with the feature location; TP method adds the POS feature on the basis of T; TPL combines with the length feature besides the location feature and the POS feature; different from the TPL, TPLR removes the repeat candidates; TPLCO indicates that we take the co-occurrence feature into account besides the feature TPL; all means all features are merged together.

Finally, we do the experiments on the testing set under three separate conditions, and we reproduce the method produced by Huang et al. [11]. Method only-word only extracts words as keywords while phrases are extracted in method only-phrase. Method word+phrase is the method presented in this paper which makes a proper combination of key-phrases and key-words. Method Ref.[11] is the approach introduced in [11]. Table 2 displays the results of the four keyword extraction methods respectively. From Table 2, we can see that compared with the method simply extracting key-words or key-phrases, the method of combining key-words and key-phrases have remarkably improved P, R and F-measure. In addition, our method performs better than [11], the F-measure increases by 6.13 percentage point from 43.79% to 49.92%.

Table 3 The composition of original document collections.

Year	Chinese documents	English documents
2006	28672	25328
2007	25036	25247
2008	14021	24292
2009	7476	10887
total	75205	85754

3.2 Experiments of Alignment

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

3.2.1 Experiment Setup

To test the performance of the system proposed in this paper, 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 web pages are as follows:

1. Many websites, such as portal website, news agency, government websites and so on, provide massive news articles. At the same time, a large proportion of the articles can be crawled politely, so it's relatively easy to collect the news document collections.
2. The news pages concern multiple domains like politics, economy and sports, so the corpus made up of news pages can avoid the limitations of domain-specific corpora.

All the news web pages are processed uniformly to create the source and target document collections. The core contents of each web page are extracted and several tags describing the document ID, language, source, publication date and title are added. Table 3 shows the basic information of document collections.

3.2.2 Results and Discussion

The quality of comparable corpora highly depends on the alignment level between the source document and target document pairs. In our experiments, we assess the alignments quality using a five-level relevance scale proposed by [7]. The five levels of relevance to assessing the alignments are defined as follows:

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

Level 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.

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

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

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

In order to evaluate the results conveniently, two standards are established: (a) the number of high quality aligned document pairs created, which is the count of document pairs in Level 1 and Level 2 (Count); (b) the quality of the whole aligned document pairs, in other words, the percentage of alignment pairs in Level 1 and Level 2 (Per.) among all alignments.

For evaluating the performance of our system, the test sets are constructed. We randomly extract 500 Chinese documents in 2009 as the Chinese test dataset. After that, we extract the keywords of each document and translate the keywords into English. At last, the English queries are run against the English collections to retrieve a ranked list of related documents. Here, for each Chinese document, the candidate English document with the highest similarity score is selected to be the aligned English document. Then we get the English document set which is comparable with the Chinese document set as the sample set. And the English test collection contains 464 documents due to the inexistence of aligned English documents of some Chinese documents or the phenomenon which two or more Chinese documents align to one English document.

Experiments with changed parameters are conducted based on the sample sets. The quality of each alignment pair is manually assessed with labor force using the five-level relevance. First, we let Chinese as the source language and English as the target language (and we name the method CE) for making one-to-one aligned document pairs. Second, we make English as the source language and Chinese as the target language (and we call this method EC) to extract one-to-one document pairs. Table 4 and Table 5 show the results of CE method and EC method with discrete values of *sim* respectively (total means the total number of document pairs).

As illustrated in the two tables, the curves of CE method and EC method share the same trend with the change of the value of *sim*. The larger the *sim* score is, the bigger the percentage of high quality aligned document pairs is. Meanwhile, the count of high quality document pairs is reduced. Taking the two criteria into account, we determine the value of *sim* ranging from 7.3 to 11.3. After that, we design the following experiments to obtain the value of parameters top-N and *sim*. Table 6 shows the results of experiments vary with different *sim* and top-N. When *sim* is equal to 8.3, top-N is equal to 3, our system performs best. Therefore, we set *sim* to 8.3, top-N to 3.

We then test the performance of our system with different keywords extraction methods. Table 7 displays the details of comparable corpora construction results under method only-word, method only-phrase and method word+phrase discussed in Sect. 3.1.2. As can be seen in Table 7, the percentage of high quality alignments mined by method only-word reaches up to 93.75%, however, the number of document pairs is too small, which is only 16. Method only-phrase mines 251 document pairs, and the percentage of high quality alignments is 74.5%. It's lower than

Table 4 Results of CE method with different values of *sim*.

<i>sim</i>	11.3	10.3	9.3	8.3	7.3	6.3	5.3	4.3	3.3	2.3
Per.	100%	98.15%	84.54%	83.75%	71.25%	59.15%	48.40%	41.15%	40.08%	40%
Count	31	53	82	134	171	194	197	200	200	200
total	31	54	97	160	240	328	407	486	499	500

Table 5 Results of EC method with different values of *sim*.

<i>sim</i>	11.3	10.3	9.3	8.3	7.3	6.3	5.3	4.3	3.3	2.3
Per.	100%	96.43%	91.43%	85.71%	80.57%	70.64%	56.78%	46.23%	41.93%	40.74%
Count	16	27	64	102	141	166	180	184	187	187
total	16	28	70	119	175	235	317	389	446	459

Table 6 Count/Per. of high quality pairs of CE method varies different *sim* and top-N.

top-N	11.3	10.3	9.3	8.3	7.3
1	31/100%	54/98.15%	97/84.54%	160/83.75%	240/71.25%
2	33/100%	60/98.33%	108/84.26%	187/84.49%	298/70.13%
3	33/100%	60/98.33%	108/84.26%	189/85%	308/69.16%
4	33/100%	60/98.33%	108/84.26%	190/84.21%	312/68.59%
5	33/100%	60/98.33%	108/84.26%	190/84.21%	313/68.37%

Table 7 Results of comparable corpora construction.

Level	only-word		only-phrase		word+phrase	
	Count	Per.	Count	Per.	Count	Per.
level1	10	62.5%	82	32.67%	78	41.27%
level2	5	31.25%	105	41.83%	82	43.39%
level3	0	0%	35	13.94%	13	7%
level4	1	6.25%	19	7.57%	13	7%
level5	0	0%	10	3.98%	3	2%
total	16	100%	251	100%	189	100%
level1&2	15	93.75%	187	74.5%	160	85%

Table 8 Results of comparable corpora construction.

Level	CE		EC		CE+EC	
	Count	Per.	Count	Per.	Count	Per.
level1	78	41.27%	60	36.36%	90	36%
level2	82	43.39%	78	37.27%	114	45.6%
level3	13	7%	19	11.52%	27	10.8%
level4	13	7%	7	4.24%	15	6%
level5	3	2%	1	0.61%	4	1.6%
total	189	100%	165	100%	250	100%
level1&2	160	85%	138	83.64%	204	81.6%

Table 9 Results of comparable corpora construction.

Level	CE+EC		Huang [11]	
	Count	Per.	Count	Per.
level1	14	32.56%	10	20%
level2	15	34.88%	14	28%
level3	9	20.93%	10	22%
level4	4	9.3%	5	10%
level5	1	2.33%	11	22%
total	43	100%	50	100%
level1&2	29	67.44%	24	48%

both method only-word and method word+phrase since the number of the candidate phrases extracted in the most documents is less than 10, and the accuracy of translation of phrases is lower than that of translation of words. Considering the standard (a) and (b), we can draw the conclusion that method word+phrase gains as many document pairs as possible in the premise of guaranteeing alignment quality. Therefore, we choose method word+phrase to construct our comparable corpora.

Next, we test the three different approaches of the construction of comparable corpora on the sample datasets obtained in this chapter with the parameters *sim* and *top-N* set to be 8.3 and 3 respectively. Table 8 shows the results of comparable corpora construction under the three separate conditions respectively. And method CE+EC performs best.

Among the surveyed related work, Huang et al. [11] created Chinese-English comparable corpora based on

CLIR techniques and its framework of construction was similar to ours. However, the two systems are different in some aspects as mentioned in Sect. 1. So we implement the method of [11] and test the system on the same testing set to make a further comparison with our work. For judging whether our method is really effective or not, we randomly selected another 100 Chinese documents from the original document collections as the testing set. We align one document to each of the target document through the two systems respectively. Table 9 shows the distribution results gained by [11] and the results acquired by our system. The total number of aligned document pairs extracted by our system is less than that of [11], while the count of highly relevant document pairs of our system is larger than that of [11]. The performance of our system reaches 67.44%, which is 19.44 percentage point higher.

All the experimental results and analysis mentioned

above indicate that our approach is effective to create high quality comparable corpora. Up to now, both the source and target documents published in 2006–2009 years are used to build comparable corpora through our proposed system. It includes 70744 aligned document pairs.

4. Conclusion

In this paper, we explored a bi-directional CLIR-based method to create Chinese-English comparable corpora. First, we crawled Chinese and English news reports from XinHuaNet using open-source crawler as the original document collections. Then we extracted the core contents of each document through discriminating noise information and processed the documents into unity format. After that, we proposed our method of keywords extraction and the results demonstrated that our method achieved a satisfactory effect. We extracted the keywords from each Chinese document and translated them into the language of another collection (the target collection), and we made the same treatment for English document. Finally, the translated queries were run against the target collection to retrieve a ranked list of related documents. The documents were aligned based on their publication dates and the KSD similarity scores. The results show that our approach is applicable and effective in the construction of Chinese-English comparable corpora. In the future, we will optimize every module, including the extraction of topic words, the translation of the topic words and the alignment of documents, in the comparable corpora construction to improve the performance of the whole system. We will try to introduce Named Entity recognition techniques for both Chinese and English to improve the performance of keywords extraction. Furthermore, it will be well worth considering to extract translation knowledge, such as mining term associations, from the comparable corpora, which can serve as a feature for improving the mapping relationship between two documents in different languages in turn.

Acknowledgment

This work has been supported by the National Natural Science Foundation of China (No. 61173100, 61173101, 61272375).

References

- [1] C.C. Yang and K.W. Li, "Building parallel corpora by automatic title alignment using length-based and text-based approaches," *Inf. Process. Manage.*, vol.40, no.6, pp.939–955, 2004.
- [2] D.S. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Computational Linguistics*, vol.31, no.4, pp.477–504, 2005.
- [3] R. Steinberger, B. Pouliquen, and C. Ignat, "Navigating multilingual news collections using automatically extracted information," *J. Computing and Information Technology*, vol.13, no.4, pp.257–264, 2005.
- [4] T. Talvensaari, J. Laurikkala, K. Jarvelin, and M. Juhola, "Creating and exploiting a comparable corpus in cross-language information retrieval," *ACM Trans. Inf. Syst.*, vol.25, no.1, article 4, 2007.
- [5] T. Talvensaari, A. Pirkola, K. Jarvelin, M. Juhola, and J. Laurikkala, "Focused web crawling in the acquisition of comparable corpora," *Information Retrieval*, vol.11, no.5, pp.427–445, 2008.
- [6] S. Sharoff, "Creating general-purpose corpora using automated search engine queries," *WaCky! Working papers on the Web as Corpus*. Gedit., 2006.
- [7] M. Braschler and P. Scauble, "Multilingual information retrieval based on document alignment techniques," *Proc. 2nd European Conference on Research and Advanced Technology for Digital Libraries*, pp.183–197, 1998.
- [8] P. Resnik, "Mining the web for bilingual text," *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, pp.527–534, 1999.
- [9] T. Tao and C.X. Zhai, "Mining comparable bilingual text corpora for cross-language information integration," *Proc. 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp.691–696, 2005.
- [10] Z. Rahimi and A. Shakeri, "Topic based creation of a persian-English comparable corpus," *Proc. 7th Asia Conference on Information Retrieval Technology, LNCS 7097*, pp.458–469, 2011.
- [11] D.G. Huang, L. Zhao, L.S. Li, and H.T. Yu, "Mining large-scale comparable corpora from Chinese-English news collections," *Proc. 23rd International Conference on Computational Linguistics*, pp.472–480, 2010.
- [12] Y.Y. Luo and D.G. Huang, "Chinese word segmentation based on the marginal probabilities generated by CRFs," *J. Chinese Information Processing*, vol.23, no.5, pp.3–8, 2009.
- [13] O. Medelyan and I.H. Witten, "Thesaurus-based index term extraction for agricultural documents," *Proc. 6th Agricultural Ontology Service (AOS) Workshop at EFITA/WCCA 2005*, Vila Real, Portugal, 2005.
- [14] P. Turney, "Learning to extract key phrases from text," *National Research Council Technical Report ERB-1057*, 1999.
- [15] B. Hong and D. Zhen, "An extended keyword extraction method," *Physics Procedia*, vol.24, Part B, pp.1120–1127, 2012.
- [16] Y. Yang, L. He, and M. Qiu, "Exploration and improvement in keyword extraction for news based on TFIDF," *Proc. 3rd International Conference on Machine Learning and Computing (ICMLC)*, pp.26–28, Singapore, 2011.
- [17] J.Z. Li, Q.N. Fan, and K. Zhang, "Keyword extraction based on TF/IDF for Chinese news document," *Wuhan University J. Natural Sciences*, vol.12, no.5, pp.917–921, 2007.
- [18] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *Int. J. Artificial Intelligence Tools*, vol.13, no.1, pp.157–169, 2004.
- [19] X.J. Wan and J.G. Xiao, "CollabRank: Towards a collaborative approach to single-document keyphrase extraction," *Proc. 22nd International Conference on Computational Linguistics*, pp.969–976, 2008.
- [20] J.F. Silva, G. Días, S. Guilloire, and J.G.P. Lopes, "Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units," *Proc. 9th Portuguese Conference on Artificial Intelligence*, pp.113–132, 1999.



Degen Huang was born in 1965. He is a professor in the Dalian University of Technology. His main research interests include natural language processing, machine learning and machine translation. He is now working at the School of Computer Science and Technology, Dalian University of Technology. He is now a senior member of CCF, and an associate editor of Int. J. Advanced Intelligence.



Shanshan Wang is a graduate student in the Dalian University of Technology. Her main research interests include keyword extraction and comparable corpora construction.



Fuji Ren Ph. D. Professor. His main research interests include Natural Language Processing, Knowledge Engineering, Sentience Computer, Machine Translation, Machine-Aided English Writing, Automatic Abstracting, Dialogue machine translation, Information Retrieval.