Mining Comparable Bilingual Text Corpora for Cross-Language Information Integration

Tao Tao
Department of Computer Science
University of Illinois at Urbana Champaign

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana Champaign

ABSTRACT

Integrating information in multiple natural languages is a challenging task that often requires manually created linguistic resources such as a bilingual dictionary or examples of direct translations of text. In this paper, we propose a general cross-lingual text mining method that does not rely on any of these resources, but can exploit comparable bilingual text corpora to discover mappings between words and documents in different languages. Comparable text corpora are collections of text documents in different languages that are about similar topics; such text corpora are often naturally available (e.g., news articles in different languages published in the same time period). The main idea of our method is to exploit frequency correlations of words in different languages in the comparable corpora and discover mappings between words in different languages. Such mappings can then be used to further discover mappings between documents in different languages, achieving cross-lingual information integration. Evaluation of the proposed method on a 120MB Chinese-English comparable news collection shows that the proposed method is effective for mapping words and documents in English and Chinese. Since our method only relies on naturally available comparable corpora, it is generally applicable to any language pairs as long as we have comparable corpora.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Text Mining

General Terms: Algorithms

Keywords: Cross-lingual text mining, comparable corpora, frequency correlation, document alignment

1. INTRODUCTION

As more information becomes available online, we have also seen more and more information in different natural languages such as English, Spanish, and Chinese. The web today consists of documents in many different languages. For a user who is interested in finding information from all the documents in different languages, it would be very useful if we could integrate related information in multiple languages [1].

Currently, cross-lingual information integration is often achieved through performing cross-lingual information retrieval(CLIR) [17], which allows a user to retrieve documents in language A with a query in language B. Most CLIR techniques rely on manually created linguistic resources such as a bilingual dictionary or examples of direct translations of words and documents [16]. Such resources may not always be available, especially for minority languages; in such a case, how to perform cross-lingual information integration would be a significant challenge. In this paper, we propose a cross-lingual text mining method that can exploit comparable bilingual text corpora to perform cross-lingual information integration without requiring any additional linguistic resources.

Comparable text corpora are collections of text documents in different languages that are about the same or similar topics. For example, news articles published in the same time period tend to report the same important international events in various topics such as politics, business, science and sports. Such data are naturally available to us, so it would be very interesting to study how to exploit them to perform cross-lingual information integration. Even when we have manually created clean bilingual resources such as a bilingual dictionary, it may still be desirable to exploit such comparable corpora for two reasons: (1) New words and phrases are constantly introduced and it would be hard to keep updating a dictionary to include them all. Approaches such as what we propose in this paper can potentially be used to acquire translation knowledge about such new words/phrases from comparable corpora and help lexicographers update dictionaries. (2) Since comparable corpora are additional resources, we may expect to achieve better performance by combining the exploitation of comparable corpora with that of a bilingual dictionary.

We frame the problem of cross-lingual information integration as one involving mapping or linking words and documents in different languages. While comparable corpora have been studied extensively in the existing literature (e.g.,[6, 10, 15, 5, 2, 8, 13]), almost all existing work assumes some kind of bilingual dictionary or translation examples to start with. We study how to map words and documents from comparable bilingual corpora *without* requiring any additional linguistic resources such as a bilingual dictionary.

Our basic idea is to exploit the fact that the frequency distributions of topically related words in different languages

KDD'05, August 21–24, 2005, Chicago, Illinois, USA.
Copyright 2005 ACM 1-59593-135-X/05/0008 ...$5.00.

are often correlated due to the correlated coverage of the same events. For example, the earthquake and sea surge disaster that happened recently in Asia has been covered in the news articles in many different languages. We can thus expect to see a recent peak of words such as "earthquake", "India", and "Indonesia" in news articles published in multiple languages. In general, we can expect that topically related words in different languages tend to co-occur together over time. Thus if we have available comparable news articles over a sufficiently long time period, it is intuitively possible to exploit such correlations to learn the associations of words in different languages.

The general idea of exploiting frequency correlations to acquire word translations from comparable corpora has already been explored in several previous studies (e.g., [6, 10, 15]). However, none of them has adopted a direct comparison of frequency distributions of candidate words as we do; rather they tend to compute the associations between the words in the same language and then compare association patterns in two different languages. Our idea and the overall approach appear to be more similar to the method used in [7], but there the task is aligning sentences in parallel corpora.

With the word mappings, we can then try to match documents in two different languages based on how well the words in each document are correlated. We propose four different methods for computing cross-lingual document similarity, including a baseline expected correlation method, an IDF-weighted correlation method, a TF-IDF method, and a translation model method. These methods allow us to perform cross-lingual document retrieval as well as linking the most strongly correlated documents together.

We evaluated our methods on a 120MB Chinese-English news report corpus for both word association mining and document association mining. The results show that our method can discover meaningful word mappings and can generate meaningful document alignments for information integration. The top ranked word pairs show various kinds of interesting associations between words in the two different languages. The document alignment results have a high precision of 0.8 at a cutoff of 100, meaning that 80% of the document pairs among the top 100 matching results are correctly matched.

The rest of the paper is organized as follows. In Section 2, we discuss our data set. In Section 3 and Section 4, we present our methods for word association mining and document association mining respectively. The experiment results are reported in Section 5, and we summarize our work in Section 6.

2. DATA SET

The comparable corpora we experiment with are 6 months' of news articles of Xinghua English and Chinese newswires dated from June 8, 2001 through November 7, 2001. There are altogether 43488 documents in Chinese and 34751 documents in English. The average document length is 204.6 words. In this data set, there are many articles in English that have some comparable Chinese articles.

An example of comparable news articles is given in Figure 1 the same international swimming championship in English and Chinese, respectively. While these two articles are from the same newswire source, and they cover the same event, they are *not* translations of each other. However, some words in the two articles are clearly translations of each other. For example, the Chinese translations of "swimming", "World Swimming Championships", and "Men's 400M Freestyle Heats" all occur in the Chinese document. (They are underlined.)

World Swimming Championships Schedule ...
9th FINA world swimming championships here on Sunday ...
9:00, Men's 50M Freestyle Heats
Women's 100M Breaststroke Heats
Men's 400M Freestyle Heats
Women's 400M Individual Medley Heats
Men's 100M Backstroke Heats

游泳世锦赛明日赛事热点

...
明天是本届世锦赛开赛的第七天,将产生女子400米个人混合泳、男子400米自由泳、男子4X100米自由泳接力、男子三米板双人跳水和女子跳台双人跳水5枚金牌,同时进行男子50米自由泳等项目的预赛。

Figure 1: A fragment of an English article (top) and a comparable Chinese fragment (bottom) about an international swimming championship

3. MINING CROSS-LINGUAL WORD ASSOCIATIONS

In this section, we present our method for mining cross-lingual word associations. Our main idea is based on the observation that words that are translations of each other or about the same topic, tend to co-occur in the comparable corpora at the same/similar time periods. Thus if we have some large comparable corpora available, it is intuitively possible to exploit such correlations to learn the associations of words in different languages.

To see if our intuition can be supported empirically, in Figure 2, we compare the frequency distribution (over time) of *Megawati* with that of its Chinese tranlation "梅加瓦蒂 " (left) and with that of another randomly chosen Chinese name ("Arafat") (right). We see that the distributions of the English "Megawati" and its Chinese translation indeed look very similar with a high correlation of 0.855, while the distributions of "Megawati" and the Chinese translation of "Arafat" are quite different with a correlation of only 0.0324. These plots show that we can indeed expect to find semantically relevant mappings between words in different languages by exploiting frequency correlations. We can thus represent each word with a frequency vector and score each candidate pair of words (in different languages) by the similarity of the two frequency vectors.

Formally, suppose we have available comparable corpora $\mathcal{C} = \{(s_1, t_1), ..., (s_n, t_n)\}$, where s_i and t_i are a set of documents associated with an anchor point of "i" in language A and language B, respectively. Let x (or y) be a source (target) word in language A and language B, respectively. We use $c(x, s_i)$ (or $c(y, t_i)$) to denote the counts of x (or y) in s_i (or t_i). The raw frequency vectors for x and y are thus $(c(x, s_1), ..., c(x, s_n))$ and $(c(y, t_1), ..., c(y, t_n))$, respectively.

In order to make the frequency vectors more comparable across different languages, it is desirable to normalize a raw frequency vector so that it becomes a frequency distribution

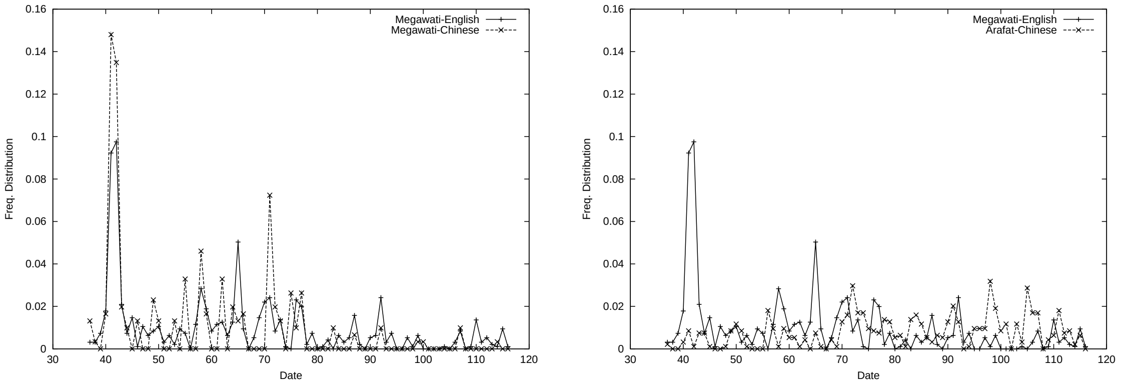

Figure 2: *Megawati* vs. its Chinese translatioin (left) and a random Chinese name (*Arafat*) (right).

over all the time points. That is, we divide all the counts by the sum of all the counts over the entire time period. Such a normalized frequency distribution would allow us to focus on the *relative* frequency on different days, which is presumably more comparable across different languages than the original non-normalized counts.

Let $\vec{x} = (x_1, ..., x_n)$ and $\vec{y} = (y_1, ..., y_n)$ be the normalized frequency vectors for x and y, respectively, where

$$x_i = \frac{c(x, s_i)}{\sum_{j=1}^{n} c(x, s_j)} \quad y_i = \frac{c(y, t_i)}{\sum_{j=1}^{n} c(y, t_j)}$$

In order to compute the similarity between $\vec{x}$ and $\vec{y}$ (or word x and word y), we use the Pearson's correlation coefficient, which is a commonly used statistic measure defined as

$$r(x, y) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{N}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2)(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2)}}$$

Using this correlation similarity measure, we can score every word in language A against every word in language B to obtain a matrix of correlations. Such word-level mappings can support a user to retrieve words in language A with a word in language B, providing some limited support of navigation across languages. Moreover, the mappings can also be used to discover matched documents between the two languages, which we discuss below.

4. MINING CROSS-LINGUAL DOCUMENT ASSOCIATIONS

We can score how well a document d_1 in language A matches a document d_2 in language B by computing a similarity score $s(d_1, d_2)$ based on how strongly correlated the words in d_1 and those in d_2 are. Technically, many different methods are possible. A natural baseline method is to compute the expected correlation between any word in d_1 and any word in d_2, i.e.,

$$\begin{aligned} s(d_1, d_2) &= \sum_{x \in d_1, y \in d_2} r(x, y) p(x|d_1) p(y|d_2) \\ &= \sum_{x \in d_1, y \in d_2} r(x, y) \frac{c(x, d_1)}{|d_1|} \frac{c(y, d_2)}{|d_2|} \end{aligned}$$

$|d_1|$ and $|d_2|$ are the lengths of d_1 and d_2, respectively. We call this method *Expected Correlation* (**ExpCorr**). Clearly, **ExpCorr** assigns a weight to every matching word pair based on the corresponding correlation and penalizes long documents due to the naturally high chances of matching, which are both reasonable.

One deficiency of **ExpCorr**, however, is that it does not distinguish a common word (e.g., "sport") from a more discriminative word (e.g., "swimming"). Intuitively, matching a common word is a weaker evidence for content matching between the two documents than matching a more discriminative word. A commonly used heuristic in information retrieval is to assign an Inverse Document Frequency (IDF) weight to each word, which penalizes popular (thus non-informative) words [14]. In our case, we associate an IDF weight for a matching pair (x, y), which is defined as

$$IDF(x, y) = IDF(x) IDF(y),$$

where $IDF(w) = \log \frac{n+1}{df(w)}$, and $df(w)$ is the number of documents in a language that contains word w, often called the document frequency of a word.

Adding IDF weighting to **ExpCorr**, we obtain the following *IDF-weighted Correlation* method (**IDFCorr**):

$$s(d_1, d_2) = \sum_{x \in d_1, y \in d_2} IDF(x, y) r(x, y) \frac{c(x, d_1)}{|d_1|} \frac{c(y, d_2)}{|d_2|}$$

In both **ExpCorr** and **IDFCorr**, the similarity score grows linearly to the count of a word in the document. However, intuitively, having one extra match after matching a word 100 times does not add so much extra evidence as matching the word the first time. We thus would like to have the similarity score to grow sub-linearly according to the count of a matching word. Again, in information retrieval, many formulas have been proposed to heuristically normalize the count of words to achieve this effect. A popular and effective method is the BM25 term frequency normalization method [11, 12]. According to this formula, the normalized count of word w in document d is given by

$$BM25(w, d) = \frac{k_1 c(w, d)}{c(w, d) + k_1(1 - b + b\frac{|d|}{AvgDocLen})}$$

where k_1 and b are parameters and $AvgDocLen$ is the average document length. In our experiments, we set $k_1 = 1.2$ and $b = 0.75$, which are the recommended default settings.

The BM25 weighting formula provides an alternative way of normalizing the count of a word, so $BM25(x, d_1)$ and $BM25(y, d_2)$ can effectively play the same role as $p(x|d_1)$

and $p(y|d_2)$, possibly with more reasonably normalization of the counts. Thus we use $BM25(x, d_1)$ and $BM25(y, d_2)$ to replace $p(x|d_1)$ and $p(y|d_2)$, respectively, in the **IDFCorr** approach to obtain the following formula, which we refer to as *BM25 Correlation* (**BM25Corr**).

$$s(d_1, d_2) = \sum_{x \in d_1, y \in d_2} IDF(x, y) r(x, y) BM25(x, d_1) BM25(y, d_2)$$

Finally, motivated by the language modeling approach to information retrieval [9, 17], we can also use the correlation between words to estimate a word translation model $t(x|y)$ [3] as $t(x|y) = \frac{r(x,y)}{\sum_{x'} r(x',y)}$. With this translation model, we can define the similarity between d_1 and d_2 as the likelihood of "generating" d_1 with a model based on d_2. That is,

$$\begin{aligned} s(d_1, d_2) &= \sum_{x \in d_1} c(x, d_1) \log p(x|d_2) \\ &= \sum_{x \in d_1} c(x, d_1)[(1 - \lambda) \sum_{y} p(y|d_2) t(x|y) + \lambda p(x|\mathcal{C})] \end{aligned}$$

where λ is a smoothing parameter to introduce a background language model $p(x|\mathcal{C})$ for modeling the noise (common words) in d_1 and $p(y|d_2)$ is the relative frequency of word y in d_2, i.e., $p(y|d_2) = \frac{c(y,d_2)}{|d_2}$ [18, 19]. $p(x|\mathcal{C})$ can be estimated as $p(x|\mathcal{C}) = \frac{\sum_{i=1}^{n} c(x,s_i)}{\sum_{x'} \sum_{i=1}^{n} c(x',s_i)}$. We refer this one as **CorrTrans**.

5. EXPERIMENTS

We use the data set described in Section 2 to evaluate the proposed word and document mapping methods. The documents published on the same day are aligned together; there are altogether 148 days. In order to support efficient computation of word correlations, we use the Lemur toolkit to index all the documents in the comparable corpora.

5.1 Mining word associations

Ideally, we can compute the correlation between every word in English and every word in Chinese. However, this involves a huge number of combinations. Since not all words are interesting to match (e.g., common functional words are not interesting for the purpose of information integration), we use the following heuristics to significantly reduce the space.

1. We first compute the entropy [4] of a word in each language using the formula $H(w) = -\sum_t p(t|w) \log p(t|w)$, where $p(t|w)$ is the normalized frequency of word w at time point t (i.e., a day).

2. We then filter out the high entropy words. These are usually frequent words as they tend to occur in everyday's news articles. The highest entropy is log(148), which is about 5, and we used a cutoff of 4.8.

3. We further filter out those words with an overall low frequency with a cutoff of 10. These words are rare, so their correlations may not be reliable.

English x	Chinese y	r(x,y)	English x	Chinese y	r(x,y)
26	26	0.929	october	10	0.875
31	31	0.920	25	25	0.875
23	23	0.918	27	27	0.873
22	22	0.917	19	19	0.873
28	28	0.916	2008	奥	0.870
16	16	0.915	apec	太	0.865
21	21	0.913	swimming	泳	0.862
d	吧	0.907	17	17	0.859
august	8	0.907	july	7	0.857
september	9	0.902	terror	怖	0.855
30	30	0.895	12	12	0.855
24	24	0.895	apec	APEC	0.852
afghan	汗	0.886	terror	恐	0.850
18	18	0.883	20	20	0.848
afghanistan	汗	0.882	terrorism	怖	0.846
14	14	0.880	games	运	0.841
2008	2008	0.879	taliban	汗	0.839
29	29	0.876	terrorism	恐	0.838
june	6	0.876	m	吧	0.836

Table 1: 38 highest correlated word pairs

In Table 1, we show the top 38 pairs with the highest correlations in the whole corpora along with their correlations, most of which are extremely high quality matchings. Many numbers are correctly aligned based on the dates in the news articles in the two different languages. There are also some direct translations of months, such as "august" vs. 8. They are clearly learned from the month information in the news articles; in Chinese news articles, August is written with the number 8. The character matching "afghan" and "afghanistan" is one of the three characters in the Chinese translation of "Afghanistan" (i.e. 阿富汗). The character matching "swimming" is also precisely its Chinese translation, and the top two characters returned for "terror" are the exact translation of this word in Chinese (i.e. 恐怖). Another interesting example is the matching of "APEC" and "apec". Interestingly, from this list, we can also infer that the two major common themes in this corpora appear to be sports and terrorism since the best matching words seem to fall into these two categories. This is an additional benefit of our word association mining algorithm.

Chinese x	corr: $r(swimming, x)$
泳 (swimming)	0.862
锦 (championship)	0.611
冈 (place name)	0.586
秒 (seconds)	0.567
牌 (medal)	0.544
跳 (diving)	0.532
半 (semi)	0.527
绩 (score)	0.512
男 (men)	0.497
赛 (championship)	0.487

Table 2: The 10 Chinese characters most correlated with "swimming" in English

In Table 2, we show the top 10 Chinese characters associated with the English word "swimming" along with their meanings in English. We see that almost all the top 10 characters are all closely related to the swimming activities. Naturally, as we go down the ranking list, the quality of matching gradually decreases.

5.2 Mining document associations

From the obtained word correlations, we select the pairs with a correlation larger than 0.6, and use these relatively more reliable correlations to match documents between the two languages. To evaluate the results quantitatively, we select a sample of representative topics from English by performing word clustering using the simple mixture model presented in [20]. We generated 30 clusters in this way.

We then take the top 5 English documents from the three randomly chosen clusters, to generate 15 seed English documents. For each English document, we use the baseline method **ExpCorr** to score all the Chinese documents of the same day, the previous day, and the day after, and retrieve top 20 Chinese documents for each English document. We read these documents and judge whether they are about the same topic/theme as the English seed document. The pairs judged as covering the same topic are assumed to be correct mappings. The seed English documents and the retrieved Chinese documents are combined together to form a working set of documents in both languages, which contains 15 English documents and 239 Chinese documents . We then use all the four methods to compute the matching scores of all the seed English documents and all the Chinese documents in the working set, and take the 20 top-ranked Chinese documents for each English seed document for evaluation. This way, we can compare the performances of these four methods with a controlled sample of the top-ranked pairs from the whole corpora.

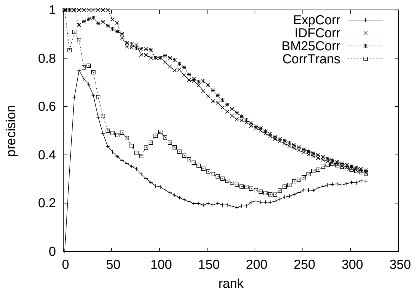

Figure 3: Alignment results on the working set.

Rank	Precision at Rank			
	ExpCorr	IDFCorr	BM25Corr	CorrTrans
3	0	1	1	0.67
5	0.2	1	1	0.8
10	0.6	1	1	0.9
30	0.67	1	0.97	0.77
50	0.42	0.98	0.92	0.48
100	0.27	0.8	0.81	0.5

Table 3: Precision at ranks on the working set.

In Figure 3 and Table 3, we show the precisions at different ranks on the working set for the four methods. We see that the baseline method **ExpCorr** performs the worst, while **BM25Corr** performs the best. A comparison between **ExpCorr** and **IDFCorr** shows that the incorporation of IDF weighting significantly improves the performance. A comparison between **IDFCorr** and **BM25Corr** indicates that the sublinear normalization of TF using BM25 helps further improve the front end precision. The translation model method **CorrTrans** performs better than the baseline but substantially worse than both **IDFCorr** and **BM25Corr**. We note that both **IDFCorr** and **BM25Corr** show roughly a monotonically decreasing curve, suggesting that the measures capture the semantic correlation between documents and a high score generally indicates a more accurate matching. However, **ExpCorr** and **CorrTrans** both show slightly improved precision at the tail, suggesting that the measures are not quite accurate. Indeed, for **ExpCorr**, quite a few top ranked pairs are non-relevant. The precisions of both **IDFCorr** and **BM25Corr** are as high as 0.8 even at a cutoff of 100, meaning that among the top 100 pairs, 80% of them are correct matchings.

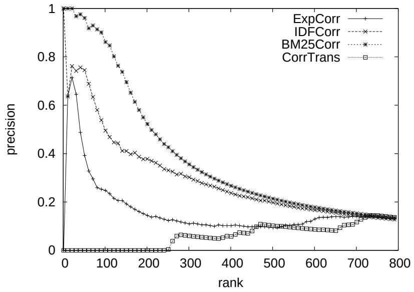

Figure 4: Alignment results on the augmented set.

Rank	Precision at Rank			
	ExpCorr	IDFCorr	BM25Corr	CorrTrans
3	0	0.67	1	0
5	0.2	0.8	1	0
10	0.6	0.7	1	0
30	0.67	0.73	0.97	0
50	0.4	0.74	0.96	0
100	0.25	0.49	0.86	0

Table 4: Precisions at ranks on the augmented set

In order to understand how much bias the working set might introduce, for each English seed document, we further rank *all* the Chinese documents from the same day as the English seed document, the day before, and the day after. This time, we take the top 50 Chinese documents for each English seed document and pool them together for evaluation. The results are shown in Figure 4 and Table 4. Note that because we have not judged all these Chinese documents, we assumed any matching with an unjudged document to be incorrect. This means that the performance we see actually represents a lower bound; the real performance can only be better.

Comparing Table 3 and Table 4 indicates that the baseline **ExpCorr** performs similarly, indicating that the additional 30 Chinese documents retrieved mostly have not made to the top pairs. The slight decrease in the precision at rank 50 and rank 100 suggests that there may be a couple unjudged Chinese documents showing up in the top 100 list. Note that, for the baseline method, these 30 Chinese documents are unjudged, thus can only decrease performance. The **BM25Corr** method also performs similarly; actually its performance on the larger set is even slightly better at rank 50 and 100. This suggests that the additional 30 Chinese documents retrieved may actually contain some correct matchings. Note that, in this case, the additional 30 Chinese documents may contain judged correct matchings, which are those documents that are among the top 20 documents returned using the baseline method, but failed to make to the top 20 documents by the **BM25Corr** method. Thus giving the **BM25Corr** method an opportunity to retrieve more results has helped it to improve the performance slightly.

Both **IDFCorr** and **CorrTrans** perform worse on the larger set, indicating that they are not very robust. Indeed, the precision of **CorrTrans** is all zeros for all the ranks. This suggests that the method cannot normalize the scores for different English seed document well; as a result, some incorrect results in the additional 30 Chinese documents may have turned out to dominate the top pairs.

Comparing the four methods on ranking the augmented working set, we see that both **IDFCorr** and **BM25Corr** again perform better than the other two methods, and **BM25Corr** is clearly the best method among the four.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose and explore a completely unsupervised cross-lingual text mining method that can exploit comparable bilingual corpora to perform cross-lingual information integration. Our basic idea is to exploit the frequency correlations of words about the same topic to first mine word associations and then mine document associations. These associations can be used to integrate multilingual text information and support cross-lingual information retrieval and navigation, which has been becoming more and more important due to the rapid growth of multilingual documents available on the Web. Evaluation of the proposed method on a 120MB Chinese-English comparable news collection shows that the proposed method is effective for mapping words and documents in English and Chinese.

The most important contribution of our work is that we have demonstrated the feasibility of mining word and document associations from comparable corpora without relying on any additional (manually created) linguistic resources. To the best of our knowledge, all previous attempts on cross-lingual information integration rely on some manually crafted linguistics resources such as a bilingual dictionary or translation examples. Since our approach does not depend on such resources, it is more general and robust than the existing methods.

Although we have shown promising results with our methods, our methods can be further improved in several ways. First, we could use the document matching results to induce new alignments for the whole corpora, which can then be used to improve our computation of word correlations. The new results of word correlations can be fed back to help generate improved document alignment. This way, we have an iterative algorithm for mining both word associations and document associations. Second, we have treated the whole document as an information unit. To improve information integration accuracy, it may be beneficial to alignment document segments. For example, we can use a sliding window to search for the best matching segments when matching two documents. Finally, it would be very interesting to explore how to design a mixture model that can mine word associations and document associations simultaneously.

7. ACKNOWLEDGEMENTS

We thank Richard Sproat and Dan Roth for helpful discussions. This work is supported in part by DOI under the contract number NBCHC040176.

8. REFERENCES

[1] J. Allan et al. Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval. *SIGIR Forum*, 37(1):31–47, 2003.

[2] L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *Research and Development in Information Retrieval*, pages 64–71, 1998.

[3] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229, 1999.

[4] T. M. Cover and J. Thomas. *Elements of Information Theory.* Wiley, 1991.

[5] M. Franz, J. S. McCarley, and S. Roukos. Ad hoc and multilingual information retrieval at IBM. In *Text REtrieval Conference*, pages 104–115, 1998.

[6] P. Fung. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of ACL 1995*, pages 236–243, 1995.

[7] M. Kay and M. Roscheisen. Text translation alignment. *Computational Linguistics*, 19(1):75–102, 1993.

[8] H. Masuichi, R. Flournoy, S. Kaufmann, and S. Peters. A bootstrapping method for extracting bilingual text pairs. In *Proc. 18th COLINC*, 2000.

[9] J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR'98*, pages 275–281, 1998.

[10] R. Rapp. Identifying word translations in non-parallel texts. In *Proceedings of ACL 1995*, pages 320–322, 1995.

[11] S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR'94*, pages 232–241, 1994.

[12] S. E. Robertson, S. Walker, S. Jones, M. M.Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 109–126, 1995.

[13] F. Sadat, M. Yoshikawa, and S. Uemura. Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval. http://acl.ldc.upenn.edu/P/P03/P03-2025.pdf.

[14] G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[15] K. Tanaka and H. Iwasaki. Extraction of lexical translation from non-aligned corpora. In *Proceedings of COLING 1996*, 1996.

[16] J. Veronis. Parallel text processing: Alignment and use of translation corpora. In *Kluwer Academic Publishers.*, 2000.

[17] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of ACM SIGIR 2001*, 2001.

[18] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'01*, pages 334–342, Sept 2001.

[19] C. Zhai and J. Lafferty. Two-stage language models for information retrieval. In *Proceedings of SIGIR'02*, pages 49–56, Aug 2002.

[20] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of KDD 2004*, 2004.