ISCAS' Text Retrieval in NTCIR Workshop II

Zhang Yibo, Sun Le, Du Lin, Jin Youbing, Sun Yufang
Chinese Information Processing Center,
Institute of Software, Chinese Academy of Sciences,
P.O.Box 8718, Beijing, 100080, P.R. China
e-mail: zyb, lesun, ldu, ybjin, yfsun@sonata.iscas.ac.cn

Abstract

In NTCIR workshop II, we participated in both the Chinese and English-Chinese Information Retrieval tracks in an automatic way. In the Chinese Information Retrieval (CHIR) track, to improve the effectiveness of the word-based retrieval system, a new method, named PM-based (PM: Proximity and Mutual information) method, is designed and introduced into the system. Employing the proximity and mutual information of the term pairs, the effectiveness of the PM-based method is expected to outmatch that of the word-based method in processing dictionary-uncovered words. In the English-Chinese Information Retrieval (ECIR) track, we mainly focused our efforts on how to disambiguate the Chinese meanings of each English query word when it is translated with a given English-Chinese bilingual dictionary, and then used the word-based Chinese text retrieval system to carry out the ECIR task.

Keywords: bi-gram, PM-based, information retrieval, word segmentation

1. Introduction

It is well know that, unlike English text, Chinese text is written as a string of ideograms with no specific delimiters between Chinese characters. In the view of the fact, it is necessary to segment the Chinese text into words before a document is indexing into a database. In the state of the art of the Chinese information retrieval, there are three kinds of indexing method according to the indexing unit: n-gram based, word-based, and hybrid indexing method.

Among the n-gram based method, bi-gram based method is the fully investigated one. The results of [1][2] show that the bi-gram based method got better results than word-based method. We assume that the following three points result in the success of the bi-gram based method. Firstly, to some extent it reflects the frequency distribution of Chinese word. The single-character and two-character Chinese words in Chinese culture are so popular that their usage frequencies can reach up to 12.1% and 73.6% individually [3]. Secondly, those bi-grams that span across adjacent words can partially reflect the proximity of the words, which can help to find newborn or dictionary-uncovered words. Thirdly, Bi-gram based method behaves better than uni-gram based method. Despite of above advantages, the bi-gram method make use of only a mechanically segmentation method, and does not take into account the syntactic and semantic characteristics of Chinese words. Also, it cannot be combined with other related techniques such as nature language processing, with which the retrieval effectiveness can be further improved, nor be applied into other related field like cross-language information retrieval. As a result, the word-based method would take up the dominant position. However, there is still a tough problem to be solved, that is, the dictionary used in word segmentation cannot, also impossibly, cover all newborn, proper name and abbreviation words. It therefore leads to great effectiveness loss of word-based method.

In the NTCIR workshop II, we tried to put forward a new method named PM-based weighting and scoring method to solve the problem. Experiments on PM-based method were conducted to evaluate and compare with word-based and bi-gram based method. Although the experiments themselves did not produce a satisfactory result, it gave us a lot of hints to further our future work.

This paper is organized as follows. Section 2 details our PM-based method. Section 3 describes our experiments on CHIR task and gives analysis of the results. Section 4 concludes the paper.

2. PM-based Weighting and Scoring Methodology

In this section, we firstly detail our original PM-based weighting and scoring method, which introduces the computation of the local proximity weight and the global mutual information weight of

This research was supported by the National Science Fund of China for Distinguished Young Scholars under contact 69983009.

the word pairs, and then explains how to apply these weights into the similarity computation between query and document. Due to the time limit on the CHIR task, an easier method approximate to the original method is designed and described in the second part of the section.

2.1 Methodology Description

In practice, the dictionary cannot, also impossibly, cover all words, especially proper name, terminology and newborn words when the information retrieval system is used under many different circumstances. As a result, the dictionary-based segment method would wrongly segment those words, discomposing these words into some single-character Chinese words or many pieces of Chinese words (in this paper, we call them sub-string Chinese words) that may have no any relevance to the original Chinese word. For example, the title of the thirty-fourth topic in the CHIR track is "白鳳豆" (Bai-feng Bean). In this topic, "白鳳豆" is a brand of medicine, however, it would be segmented into "白/鳳/豆" three single-character Chinese words. If the retrieval system uses these sub-string words as the index units with no other measures to reflect the fact that they three are a Chinese word, it will definitely return user a lot of irrelevant documents and lead great loss on the retrieval effectiveness.

We also take "白鳳豆" as an example but from another point of view to investigate the problem. If the topic and content of a document are relevant to "白鳳豆", the sub-string Chinese words "白/鳳/豆" must be adjacent with each other or proximate, and according to [4], the bigger mutual information of those sub-string word in the document collection, the more possibility they would be a Chinese word or a phrase. We conclude that the word proximity in a given a document and word mutual information in the whole collection imply how possible these sub-string words be a word or a phrase. This heuristic information could be employed as the matching features to select the relevant documents from the collections. The following method description is mainly from [5].

More formally, supposing $c_1 c_2 \Lambda\ c_s$ is a Chinese string of sentence S, and $w_1 w_2 \Lambda\ w_r$ is the corresponding segmented word pieces. The word proximity coefficient $pc(w_i, w_j)$ is defined as:

$$pc(w_i, w_j) = \begin{cases} \dfrac{1}{j-i} & j > i \quad and \quad j - i < MaxPW \\ 0 & others \end{cases} \qquad (1)$$

where $MaxPW$ is the maximum windows size of word proximity coefficient. The proximity coefficient of word w_i and w_j is not symmetrical on this calculation, because of the proper nouns, subject terminology and most phrases are word (the new words may be wrongly segmented into sub-string word) order dependent. So the word proximity coefficient partly reflects the word modification relation of the phrases and the character order of the new words. The information, like the term frequencies in the traditional vector space model, could be used as the similarity parameters while matching the query and documents.

The local proximity weight of word pair x and y in document D_i, $lpc_i(x, y)$, is defined as the sum of the proximity coefficient of different occurrences:

$$lpc_i(x, y) = \sum_{D_i} pc(x, y) \qquad (2)$$

where $pc(x, y)$ is the proximity coefficient of word pair x and y in each sentence.

In the traditional vector space model [7], the weight of term T_j in document D_i is composed of two parts, local weight and the global weight. The global weight of term T_j in the document database is determined by the inversion of the document frequency idf. The lower the document frequency, the higher the global weight.

The idf weighting scheme could not be employed to describe global weight of proximity indexing. The low document frequency of word pair x and y indicates that the probability of word pair x and y being proper noun or phrase is also low. On the other hand, the mutual information could be employed to reflect the important relations of word pair x and y, so the global mutual information weight of word pair x and y, $gmi(x, y)$, is defined as:

$$gmi(x, y) = log \frac{p(x, y)}{p(x) \cdot p(y)} \qquad (3)$$

where, $p(x)$ and $p(y)$ are the occurring probabilities of word x and y in the proximity window respectively, and $p(x, y)$ is probability of word x and y occurring together in the proximity window. These probabilities could be estimated through the occurrence of word x and y in the document databases.

$$p(x) = \frac{\sum_{i=1}^{n} tf_{ix}}{\sum_{i=1}^{n} |D_i|} \times MaxPW \qquad (4)$$

$$p(y) = \frac{\sum_{i=1}^{n} tf_{iy}}{\sum_{i=1}^{n} |D_i|} \times MaxPW \qquad (5)$$

$$p(x, y) = \frac{\sum_{i=1}^{n} lpc_i(x, y)}{\sum_{i=1}^{n} |D_i|} \times MaxPW \qquad (6)$$

where tf_{ix} and tf_{iy} are the individual term frequency of word x and y in the maximum windows of document D_i, $lpc_i(x, y)$ is the local proximity weight of word pair x and y in document D_i. Replacing (3) with equation (4), (5) and (6), the global mutual information weight of word pair x and y is:

$$gmi(x, y) = log \frac{\sum_{i=1}^{n} lpc_i(x, y)}{\sum_{i=1}^{n} tf_{ix} \sum_{i=1}^{n} tf_{iy}} \times \frac{\sum_{i=1}^{n} |D_i|}{MaxPW} \quad (7)$$

So the PM weight of word pair x and y in document D_i, $w_i(x, y)$, is defined as:

$$w'_i(x, y) = lpc_i(x, y) \times gmi(x, y) \quad (8)$$

The weight $w_i(x, y)$ should be normalized according to the document length $|D_i|$ and maximum document length $Max(|D|)$, so the normalized PM weight of word pair x and y in document D_i, $PM_i(x, y)$ is:

$$PM_i(x, y) = \frac{Max(|D|)}{|D_i|} \times w'_i(x, y) \quad (9)$$

The similarity coefficient $sim(D_i, D_j)$ between document D_i and D_j is based on two parts, $sim_{TR}(D_i, D_j)$ for the traditional word subspace, and the other part, $sim_{PM}(D_i, D_j)$ for the word PM subspace, that is:

$$sim(D_i, D_j) = \alpha \cdot sim_{TR}(D_i, D_j) + \beta \cdot sim_{PM}(D_i, D_j) \quad (10)$$

where α and β are the adjusting coefficient, $sim_{TR}(D_i, D_j)$ and $sim_{PM}(D_i, D_j)$ could be calculated according to the cosine similarity coefficient of the sub-vectors in the traditional word subspace and word PM subspace respectively.

2.2 Approximate Implementation

Because of the time restriction of the CHIR task, we did not fully follow the theory described in section 2.1 but adopted analogous and relatively easier method to rank the documents. In this method, we only take word mutual information into account and do not investigate the word proximity and word order. The approximate PM weight of the word pairs is defined as:

$$PM(x, y) = max\left(\frac{p(x, y) - p(x) \cdot p(y)}{p(x) + p(y)}, 0\right) \quad (11)$$

In above equation, $p(x, y)$ cannot be calculated by formula (6) because $lpc_i(x, y)$ is not ready-made. It can be defined in the same way of $p(x)$ and $p(y)$ as:

$$p(x, y) = \frac{\sum_{i=1}^{n} tf_{ixy}}{\sum_{i=1}^{n} |D_i|} \times MaxPW \quad (12)$$

where tf_{ixy} is the term frequency of word x and y occurred in the same maximum windows. The similarity coefficient $sim_{PM}(Q_i, D_j)$ in formula (10) between query Q_i and document D_j can be calculated as:

$$sim_{PM}(Q_i, D_j) = \frac{\sum_{x, y \in Q_i, |x-y| \le MaxPW} PM(x, y)}{MaxPM_i} \quad (13)$$

where $|x - y|$ is the distance of word x and y, $MaxPM_i$ is defined as:

$$MaxPM_i = \max_{j=1 \to n}\left(\sum_{x, y \in Q_i} PM(x, y)\right) \quad (14)$$

Introducing equation (13) into (10), the similarity between query and document can be calculated and used to rank the documents.

3. Experiments and Analysis of Results

In the NTCIR Workshop II, we took part in the CHIR task, which includes CHIR track and ECIR track. The retrieval system used in the experiments is based on the conventional vector space model. In this section, our experiments on CHIR and ECIR are introduced individually and failure is analysed according to the experiment results.

3.1 CHIR Track

In the CHIR track, we carried out four groups of experiments according to the query types that are "LO" (any query uses the narrative of the topics), "SO" (any query uses no narrative of the query), "VS" (any query uses neither narrative nor question of the topics) and "TI" (any query uses the title of the topics only). Each group has three runs that make use of bi-gram based method, word-based method and PM-based method. The indexing structure of all runs is type of inverted file.

For the bi-gram based method, the text of queries and documents are segmented into two-character Chinese word and indexed into database without stop words. The similarity between query and document is calculated by the conventional cosine coefficient.

For the word-based method, the word segmentation is based on a dictionary coded in Big5 set which contains 138,000 words and has the usage frequency of each word. The forward and backward maximum matching algorithm is used to segment the text and find the word combinatorial ambiguities [6]. If there is combinational ambiguity, the product of the usage frequencies of the words is used to

determine the final combination of words. A stop word list of 287 elements is set up, which contains frequently used functional words as well as symbols. The weight of each word is calculated by the $tf \times idf$ [7]. The similarity between query and document is calculated by the cosine coefficient.

For the PM-based method, the same segmentation method and stop word list are used as in the word-based method. In the inverted file, the word position information is also stored to calculate the PM weight of word pairs. The maximum window size is set to 7. The similarity between query and document is calculated by traditional similarity and PM similarity. PM similarity is calculated by equation (13). The adjusting coefficients α and β in equation (10) are all set to 1.

The results of all runs by relaxed relevance judgment are showed in table 1 and figure 1. The results by rigid relevance judgment are showed in table 2 and figure 2.

Recall	LO			SO			TI			VS		
	bi-gram	Word	PM	bi-gram	Word	PM	bi-gram	Word	PM	bi-gram	Word	PM
0.0	0.9556	0.9432	0.7845	0.9513	0.9501	0.7818	0.7754	0.7561	0.6912	0.9900	0.9709	0.6703
0.1	0.8444	0.8439	0.6790	0.8357	0.8404	0.6704	0.6735	0.6548	0.6076	0.8887	0.8675	0.5800
0.2	0.7765	0.7852	0.6425	0.7688	0.7938	0.6326	0.6218	0.6125	0.5530	0.8091	0.8220	0.5079
0.3	0.7429	0.7425	0.6062	0.7370	0.7500	0.5931	0.5939	0.5677	0.5254	0.7607	0.7721	0.4434
0.4	0.7112	0.6998	0.5811	0.7032	0.6989	0.5758	0.5616	0.5131	0.4877	0.7305	0.7294	0.4200
0.5	0.6778	0.6566	0.5530	0.6778	0.6514	0.5540	0.5227	0.4700	0.4527	0.6919	0.6917	0.3840
0.6	0.6399	0.6106	0.5243	0.6404	0.6152	0.5194	0.4512	0.4102	0.3993	0.6501	0.6408	0.3677
0.7	0.5976	0.5485	0.4878	0.5938	0.5605	0.4824	0.4017	0.3540	0.3526	0.6088	0.5719	0.3485
0.8	0.5154	0.4781	0.4230	0.5122	0.4963	0.4238	0.3342	0.2925	0.2956	0.5279	0.5134	0.3254
0.9	0.4227	0.3746	0.3437	0.4152	0.3787	0.3454	0.2532	0.2118	0.2176	0.4273	0.4043	0.2804
1.0	0.2618	0.2355	0.2175	0.2585	0.2372	0.2203	0.1395	0.0978	0.1033	0.2623	0.2347	0.2308

Table 1. CHIR (Relaxed relevance) Interpolated Recall - Precision Averages

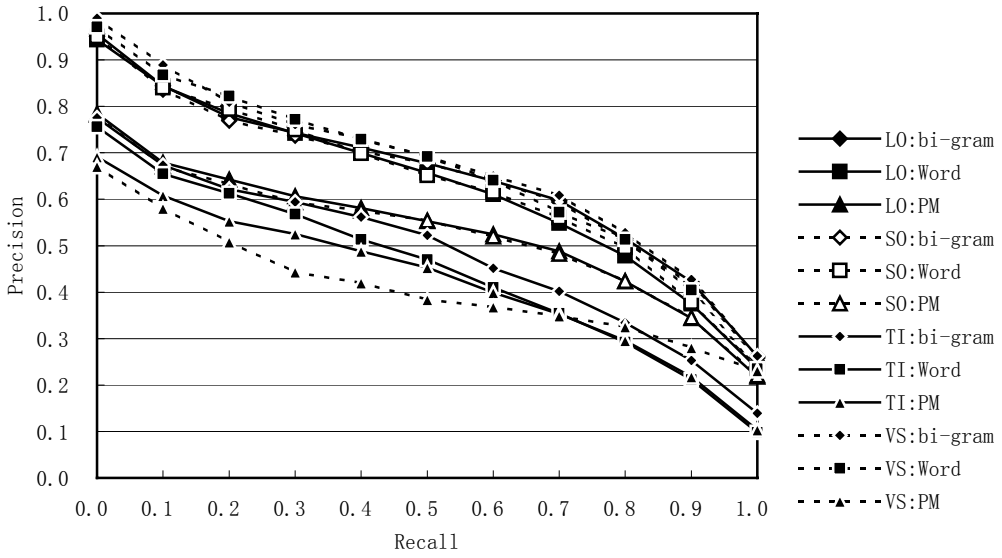

Figure 1. CHIR (Relaxed relevance) Runs

Recall	LO			SO			TI			VS		
	bi-gram	Word	PM	bi-gram	Word	PM	bi-gram	Word	PM	bi-gram	Word	PM
0.0	0.8451	0.8255	0.6323	0.8490	0.8564	0.6376	0.5968	0.5653	0.4903	0.8962	0.8743	0.8027
0.1	0.7630	0.7493	0.5596	0.7719	0.7804	0.5567	0.5372	0.4962	0.4459	0.8287	0.8142	0.6778
0.2	0.6769	0.6826	0.4968	0.6739	0.6953	0.4849	0.4715	0.4561	0.4037	0.7301	0.7521	0.6414
0.3	0.5974	0.6053	0.4356	0.5946	0.6105	0.4234	0.3933	0.3990	0.3517	0.6370	0.6449	0.5959
0.4	0.5733	0.5750	0.4135	0.5697	0.5853	0.4129	0.3719	0.3776	0.3227	0.6041	0.6205	0.5820
0.5	0.5276	0.5363	0.3867	0.5311	0.5367	0.3824	0.3443	0.3384	0.3033	0.5531	0.5733	0.5619
0.6	0.4893	0.4779	0.3681	0.4846	0.4765	0.3627	0.3246	0.3165	0.2851	0.5076	0.5030	0.5257
0.7	0.4518	0.4419	0.3446	0.4471	0.4471	0.3403	0.2981	0.2772	0.2604	0.4676	0.4584	0.4842
0.8	0.4058	0.4093	0.3189	0.4031	0.4104	0.3167	0.2526	0.2404	0.2394	0.4186	0.4255	0.4333
0.9	0.3388	0.3263	0.2707	0.3352	0.3387	0.2723	0.2091	0.1843	0.1801	0.3477	0.3532	0.3547
1.0	0.2940	0.2902	0.2399	0.2910	0.2945	0.2379	0.1651	0.1494	0.1414	0.2997	0.2943	0.2207

Table2. CHIR (Rigid Relevance) Interpolated Recall - Precision Averages

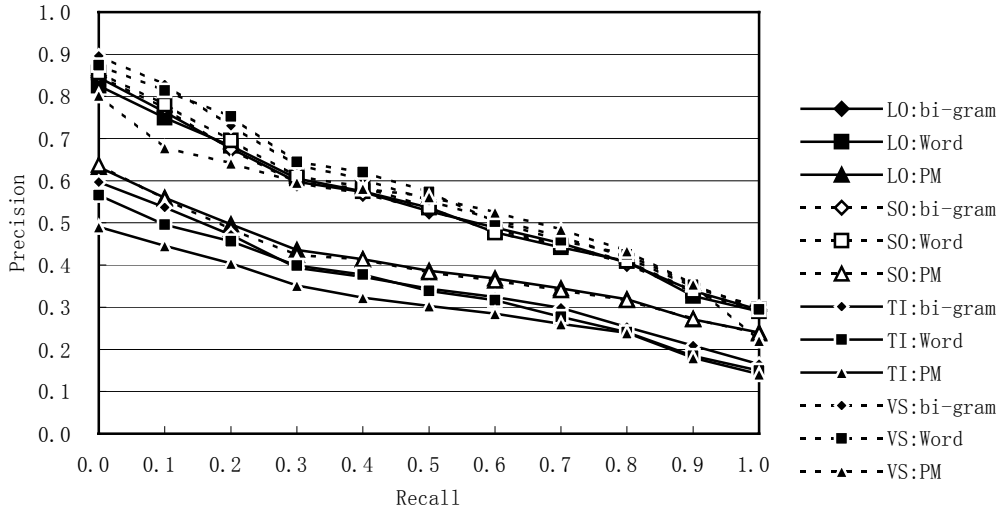

Figure 2. CHIR (Rigid relevance) Runs

RUN	1/R	2/R	3/R	1/X	2/X	3/X
LO	0.5184	0.5148	0.3746	0.6416	0.6200	0.5080
SO	0.5165	0.5205	0.3696	0.6347	0.6231	0.5029
VS	0.5482	0.5539	0.3844	0.6609	0.6496	0.5111
TI	0.3343	0.3219	0.2823	0.4681	0.4359	0.4047
AVG	0.4794	0.4778	0.3527	0.6013	0.5822	0.4817

Table 3. The average query precision of the three methods

Over all 12 runs, the average precisions of the three different methods on rigid and relaxed condition are shown in table 3. The column 1/R, 2/R and 3/R are rigid relevant precisions, corresponding to bi-gram based method, word-based method and PM based method respectively, while column 1/X, 2/X and 3/X are for relax precisions.

It has been shown that the average query precisions of the bi-gram and word indexing over four runs are almost equivalent. The performance of PM-based scoring is relatively low. Why the PM-based method does not function properly as expected?

In this evaluation, the PM similarity is simplified and depended only on the term mutual information. The proximity information is not taken part in the similarity calculation. Notice that the number of relevant documents in the collection is very small compared with the total number of documents in the collection. For an example, the relevance document of the forty-fourth topic by relaxed relevance judgment is only 8, while the document number of CHIR track is 132,173. The accuracy of the term mutual information is relatively low.

Secondly, the unimportant term-pairs with high PM weights may impair the retrieval precision. Since most terms in the field <concepts> of topics are single words, the two or more word terms maybe over weighted. For an example, the concept field of the thirty-first topic is “地震、斷層、活動斷層、觸口斷層、規模、能量、週期、板塊、地質、頻率、預測、災害性地震、台灣、嘉南、東部、西部、地震測報中心、中央氣象局” (“earthquake, fault, moving fault, active fault lines, scale, energy, cycle, tectonic plate, geology, frequency, forecast, disastrous earthquake, Taiwan, Chi-nan, east, west, Earthquake Forecast Center, the Central Weather Bureau”). Among these Chinese words, “地震、斷層、板塊、地質” (“earthquake, fault, moving fault, tectonic plate, geology”) are all a Chinese word and relatively important concepts as to the topic. However, they do not take part in the PM weight calculation, because they have no other words in the proximity window to consist of term pairs. On the other hand, “嘉南、中央氣象局” (“Chi-nan, the Central Weather Bureau”) are relatively unimportant concepts, but their PM weight is calculated and employed into the similarity calculation.

Thirdly, PM similarity between query and documents is based only on the dot product without any normalization on the document size.

Fourthly, the stop wordlist used in our experiments is only set up by the term frequency, and includes a lot of key query terms. The term database of the document collection does not contain any information of these terms. However, these terms are crucial to our PM-based method. For example, for the second topic “新三不政策與台獨” (The New 'Three Noes' and independence of Taiwan). The term database only includes two terms “政策” and “台獨”. The information of mutual information among the term “新”, “三”, “不” is lost.

At last, the adjusting coefficients are equivalent for word and the PM weights. The coefficients should be calculated according to the number of terms and term-pairs in the query so as to improve relevant scoring accuracy.

3.2 ECIR Track

For lack of the English and Chinese (Big5) bilingual or comparable corpus in hand, we cannot extract bilingual dictionary from corpora. The only English and Chinese (Big5) dictionary we found on web is about 170,000 entries. However, it is a pity that many translations are not consistent with those used in the Chinese test corpus. Such as, the English word ‘Disney’ is translated to Chinese word “狄斯尼”, not the Chinese word used in test corpus “迪士尼”; the word ‘stray dog’ is translated to Chinese word “喪家犬”, not the Chinese word “流浪狗”, etc. And there are also many English words you cannot find in this dictionary. What we focused on is to disambiguate the Chinese meanings of the each English query word when it is translated with a given English-Chinese bilingual dictionary by word frequency and co-occurrence in test corpus.

There are totally five result files we submitted, IOS-ECIR-VS-02-A, IOS-ECIR-VS-02-B, IOS-ECIR-VS-02-C, IOS-ECIR-VS-02-D and IOS-ECIR-VS-02-E. We firstly translated the English query words into Chinese words by directly looking up the bilingual dictionary in different translation methods, and then used the word-based Chinese retrieval system to find the relevant documents.

The method A use one English Chinese Machine Translation software, and its results are used to be contrasted. The method B chooses the first Chinese meaning of every English word. And it is the baseline of other methods. The method C chooses the Chinese meaning, whose frequency in whole corpora is maximal. In method D, we firstly filtered out the meaning of English words, whose frequency is below a threshold. And then chose two meanings, whose frequency are bigger. In method E, by co-occurrence frequency of different meanings of English words, to eliminate the translation ambiguity of method D. For example, in the term ‘Assembly Parade’, assembly can be translated into Chinese words “集會”, “會議”; parade can be translated into Chinese words “展覽”, “遊行”. By compare the co-occurrence frequency of Chinese words “集會” and “展覽”, “議會” and “展覽”, “集會” and“遊行”, “議會” and “遊行”, the biggest co-occurrence pair is selected.

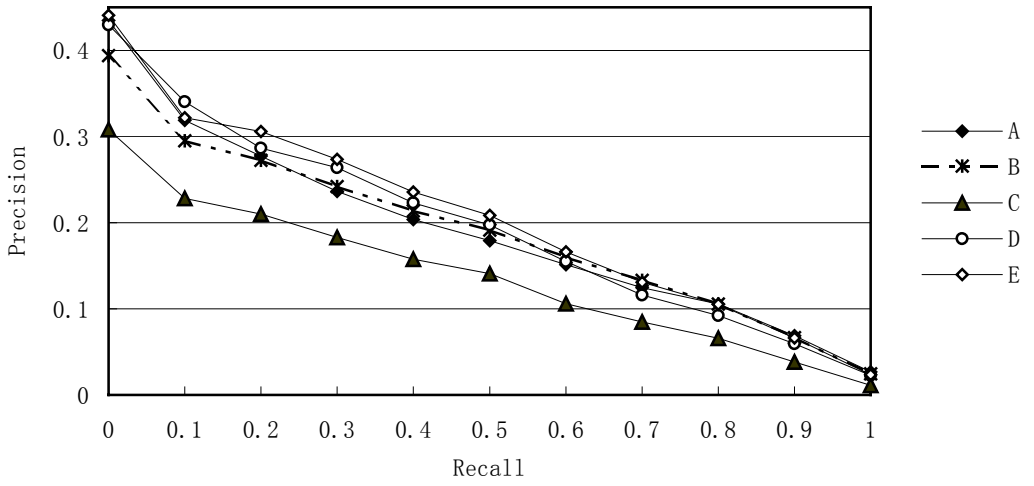

Figure 3. ECIR (Relaxed Relevance) Runs

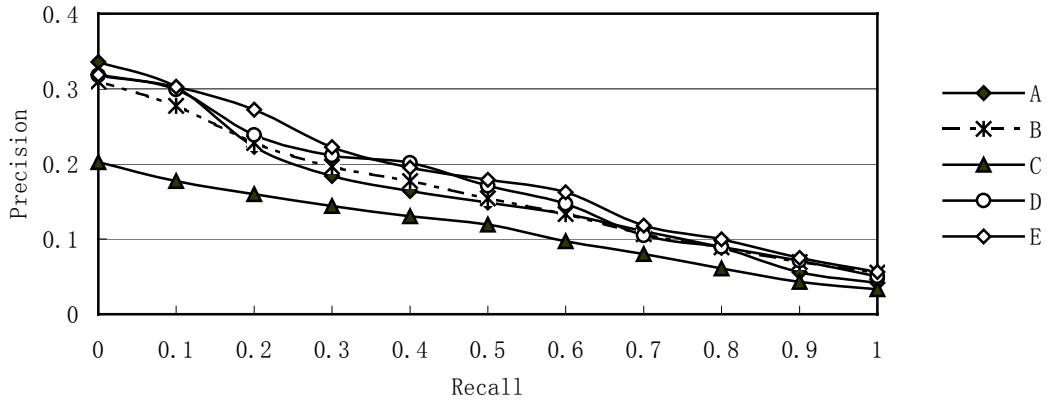

Figure 4. ECIR (Rigid Relevance) Runs

In figure 3 and figure 4, our five results of ECIR 11-point precision relaxed relevance and rigid relevance are given respectively. It is self-evident that the worst precision is method C and there is no big difference in result A and result B. When use the English-Chinese Machine Translation software to translate the very short query, due to short of enough context, the result is no better than the result of method B, which just select the first meaning from the dictionary. The result D and E are better than result B as we wish. However, the difference is too small to be mentioned. Because what we try to do is only eliminate the translation ambiguity, if there are no correct translation meanings in the dictionary, there would be no good results.

4. Conclusion

In the CHIR track, we designed the PM-based weighting and scoring method in order to solve the retrieval effectiveness loss, which is caused by the dictionary-uncovered words. The results show that there are several aspects in the PM-based method need to be improved, such as to normalize the PM similarity, to determine which weight is more important between word traditional weight and PM weight as to the similarity calculation, to further refine the PM weight calculation, etc.

In the ECIR track, because of unavailability of English and Chinese bilingual corpus, we only conducted experiments based on several query translation method. Results show that the co-occurrence information of word pairs may do some help in the disambiguity of the query translation, but not conclusive.

References

[1] Wilkinson R. Chinese document retrieval at TREC-6. *In Text Retrieval Conference (TREC-6)*, NIST, Gaithersburg, Maryland, 1997. 25-30.

[2] Leong M. K., Zhou H. Preliminary qualitative analysis of segmented vs bigram indexing in Chinese. *In Text Retrieval Conference (TREC-6)*, NIST, Gaithersburg, Maryland, 1997. 551-558.

[3] Language teaching lab of Beijing Language Institute. *Modern Chinese Frequency Dictionary*. Beijing Language Institute Press, 1985. (in Chinese)

[4] Church K. W., Hanks P. Word association norms, mutual information and lexicography. *Computational Linguistics*, 1990.16(1):22-29.

[5] Du, Lin and Sun, Yufang. A new indexing method based on word proximity for Chinese text retrieval. *Journal of Computer Science and Technology*, 2000. 15(3), pp.280-286.

[6] Liu Y., Tang Q., Shen X. *Modern Chinese Word Segmentation Specification and Methodology for Information Processing*. Tshinghua University Press, 1994. (in Chinese)

[7] Salton G. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.