Between Terms and Words for European Language IR and Between Words and Bigrams for Chinese IR

Jian-Yun Nie

Département d'Informatique et Recherche opérationnelle,
Université de Montréal
C.P. 6128, succursale Centre-ville
Montreal, Quebec, H3C 3J7 Canada

nie@iro.umontreal.ca

Jean-Pierre Chevallet,
Marie-France Bruandet

Laboratoire CLIPS, IMAG
BP. 53X, 38041 Grenoble cedex, France

Jean-Pierre.Chevallet@imag.fr
Marie-France.Bruandet@imag.fr

Université de Montréal, together with the MRIM research group of the CLIPS laboratory in IMAG institute, participated in the Cross-Language Retrieval track in TREC6. Université de Montréal also participated in the Chinese track. In this paper, we describe our approaches used in our experiments. In the cross-language retrieval track, we compared word-based retrieval and term-based retrieval. In the Chinese track, the approaches using bigrams and words are compared.

1. Introduction

The principal goal of our participation in TREC6 is to compare the following two pairs of approaches:

For French and Cross-language retrieval:

- the classical approach based on words;
- our approach using terms from a terminological base, and automatically built terms.

For Chinese retrieval:

- the approach using bigrams;
- the approach using word segmentation.

This report describes our approaches and the experimental results.

2. French and Cross-Language Retrieval using terms

Classical IR systems operate on a word-basis. That is, both documents and queries are represented by a set of weighted words (keywords). This approach has been criticized in a number of studies on the two following points:

1. The content of a document (or a query) cannot be represented precisely by a set of isolated words.

2. Different keywords are assumed to be independent. In reality, they are not.

To solve the first problem, it is often suggested that compound terms, instead of single words, should be used. For example, when the compound term "expert system" is used to represent a document content, it is more precise than using "expert" and "system" separately. Many studies have been concerned with the problem of extracting compound terms from texts. Many of them are based on statistics of word co-occurrences and syntactic analysis of phrases.

However, whether such an approach can bring significant improvements to the system's performance is still an issue [3, 10].

To solve the second problem, it is generally suggested that one must incorporate in the query evaluation some form of inference using relationships between terms. The following approaches have been proposed to establish relationships between different terms:

- by considering term co-occurrences in the document collection [9], or
- by considering user relevance feedback [4, 5].

Terminological bases may provide solutions to both problems. A terminological base may contain a large number of compound terms set up by experts. They can be used to index documents and queries. In addition, in some bases, relationships are established between different terms. These relationships may be used to retrieve related documents.

However, a manually established terminological base may not cover correctly all the terms in a particular application domain. An automatic term building mechanism may be useful to complement a manual base. In our experiments, we tested the use of a manual terminological base both alone and in combination with an automatically built base, using the French document collection and French queries. Later on we also used the manual base for English to French retrieval.

2.1. Using manual terminological base

We used the "Banque de Terminologie du Québec" (BTQ) as our source of terms. The BTQ has been developed by the "Office de la Langue Française du Québec" (Office of French Language of Quebec). It contains over 500 000 files classified in about 160 different domains. Each file in the BTQ contains (among others) a term, its domains, related terms, and definition. All these elements are in both French and English. Below is part of the file 1000012 (file id.) in the BTQ:

```
1000012    11 (domain)
       1615 (pyrotechny)
1000012    13 (English term)
       delay electric blasting cap
1000012    17 (English definition)
       An electric blasting cap with a delay element between the priming
       and detonating composition to permit firing of explosive charges in
       sequence with but one application of the electric current.
1000012    41 (French term)
       détonateur électrique à retard
1000012    45 (French definition)
       Ces détonateurs permettent d'échelonner plusieurs explosions dans
       le temps en n'employant qu'un seul circuit électrique et une seule
       charge électrique.
1000012    48 (French synonyms)
       amorce électrique à retard
       détonateur électrique à retardement
```

The BTQ has a good coverage for a number of specialized domains (especially some

scientific domains). Our intention is to use the BTQ to do the following processes:

- Using the terms in the BTQ to extract terms in both documents and queries in order to create a additional representation more precise than words;

- Using the synonyms (and other related terms) to extend the user's queries;

- In the cross-language retrieval, English terms recognized in a query are translated into their corresponding French terms which are then used to retrieve French documents.

The extraction of the BTQ terms from texts may be seen as illustrated in Figure 1.

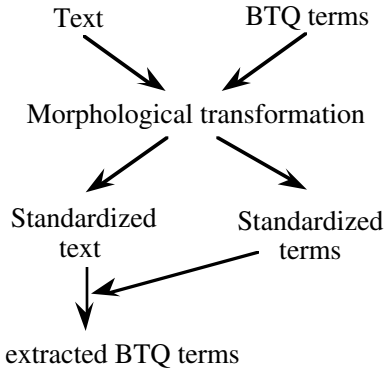


Figure 1. Process flow

The first step aims to transform the terms in the BTQ into a standard form with a stemming process. A French text is submitted to the same transformation. The stemming process removes 123 frequently used French word endings. For example:

- issement (élargissement)
- amment (indépendamment)
- able (traitable)
- ance (plaisance)

After the transformation, the text and the terms in the BTQ may be compared so that all the terms occurring in the text may be recognized and extracted. Below is an example to illustrate the result of the extraction.

Initial text

```
Description et schéma d'une cellule d'écoulement avec une
électrode de carbone vitreux recouverte d'un film de Hg,
utilisée pour cette méthode d'analyse. Etude de l'influence
de la concentration en Hg2+ et de l'oxygène dissous sur la
hauteur du signal lors du dosage de Pb et Cd. Observation
d'une augmentation de la sensibilité avec le système en
continu par rapport à un système en discontinu
```

Extracted terms:

```
Description, schéma, cellule, écoulement, électrode,
électrode de carbone, carbone, vitreux, film, Hg, méthode,
analyse, concentration, oxygène, oxygène dissous, dissous,
hauteur, signal, dosage, Pb, Cd, Observation, augmentation,
sensibilité, système, continu, rapport, système, discontinu
```

2.2. Automatic term building

This task was done by the IOTA system. The IOTA system was built in the middle of 1980s. It contains a component that builds a term base automatically from corpus [1]. This process is based on a syntactic analysis of French sentences, as well as the frequency of word co-occurrences in the corpus.

In order to build a term base, the following processes are performed: First, the texts in the corpus are analyzed so that a (sometimes more than one) syntactical category is attached to each word. This requires us to solve many ambiguities during the analysis. To do this, a precedence matrix is used which tells if a given category can follow or precede another category in French. In this way, many impossible sequences of syntactic categories may be removed. After the tagging, all the word groups that fit one of the pre-determined syntactic structures are identified as potential terms. Finally, a statistical analysis is used to determine a degree of potentiality for each group. Only highly potential groups are retained as terms and are put into the automatic term base. Below are some examples of the terms recognized by this process (together with their frequency in the corpus):

2394 conseil_fédéral	1026 grand_conseil	647 année_dernier
2025 premier_foi	1015 police_cantonale	601 croix_rouge
2017 million_de_franc	991 conseil_de_administration	596 annoncer_jeudi
1754 new_york	984 droit_de_homme	588 ronald_reagan
1672 premier_ministre	861 office_fédéral	578 également_été
1455 conseil_national	840 pouvoir_être	570 annoncer_mardi
1435 nations_unies	819 département_fédéral	568 anné_précédente
1385 affaire_étrangère	815 dernier_année	560 perez_de_cuellar
1326 chiffre_de_affaire	770 parti_communist	553 annoncer_mercredi
1243 million_de_dollar	725 source_officie	550 cour_de_conférence
1213 grande_bretagne	724 déclarer_m	548 nord_ouest
1164 affaires_étrangère	701 territoire_occuper	540 suisse_romand
1153 milliard_de_dollar	691 semaine_dernier	529 annoncer_vendredi
1093 chef_de_etat	678 premier_ministr	519 source_proche
1086 plus_grand	671 proche_orient	509 dire_m
1063 week_end	652 parti_socialiste	507 protection_de_environnement

Table 1. Samples of the terms built automatically

In comparison with the manual terminological base, the automatic base is less accurate. That is, many established items are not real terms. For example: 1086 plus_grand (larger), 724 déclarer_m (declare_me), 578 également_été (also been). These terms have been retained because of word ambiguities. For example, "plus" and "grand" may also be a noun in French, although they are respectively adverb and adjective in this example. Despite the noise, the automatic term base has a good coverage of the corpus. Thus, it is a good complement to a manual base. We used this base to recognize and extract terms from texts just in the same way as shown in Figure 1.

2.3. Retrieval results and discussion

In our experiments, we used a modified version of SMART system [2]. Prior to TREC6 experiments, we had a French corpus - OFIL - with a set of evaluated queries. This corpus is set up in the Amaryllis project [8]. It contains a set of articles published in the French journal "Le

Monde". We used this corpus to determine some variables for French IR as follows.

We first compared different stemming approaches: transforming plural to singular, and removing several sets of word endings. Finally, we chose the set of 123 word endings which gave the best results for OFIL.

We compared approaches using different combinations of terms and words: terms only, terms with words in a single vector and terms and words in two different vectors. The second approach gave the best results. Using terms only is not as good as the classical approach using words only, and separating terms from words is not as good as grouping them together. This shows that the terms alone do not have a good coverage of document contents. They have to be supplemented by single words.

We used *ltc* term weighting of SMART in all our experiments. Three sets of results have been submitted. Run2 (CLIPS2) is the result with the classical approach using (stemmed) words only. Run1 (CLIPS1) uses both the BTQ terms and words in a single vector. Run3 (CLIPS3) uses the BTQ terms, automatically built terms and words. Table 2 gives a comparison of these approaches for French to French retrieval.

Run	Run2	Run1	Run3
Total number of documents over all queries			
Retrieved:	21000	21000	21000
Relevant:	1239	1239	1239
Rel_ret:	1009	989	983
Interpolated Recall - Precision			
at 0.00	0.6859	0.6932	0.7113
at 0.10	0.5505	0.5949	0.6016
at 0.20	0.4650	0.4887	0.5028
at 0.30	0.4078	0.4072	0.4381
at 0.40	0.3644	0.3481	0.3823
at 0.50	0.3245	0.3109	0.3427
at 0.60	0.2886	0.2741	0.3061
at 0.70	0.2453	0.2331	0.2418
at 0.80	0.2173	0.1930	0.1943
at 0.90	0.1369	0.1307	0.1332
at 1.00	0.0288	0.0350	0.0401
Average precision (non-interpolated) over all rel docs			
	0.3171	0.3204	0.3404
		(1.04%)	(7.35%)
Precision:Precision:			
At 5 docs:	0.5048	0.5238	0.5714
At 10 docs:	0.4619	0.4714	0.5429
At 15 docs:	0.3841	0.4159	0.4698
At 20 docs:	0.3619	0.3857	0.4214
At 30 docs:	0.3222	0.3524	0.3540
At 100 docs:	0.2138	0.2229	0.2162
At 200 docs:	0.1493	0.1505	0.1507
At 500 docs:	0.0822	0.0783	0.0781
At 1000 docs:	0.0480	0.0471	0.0468
R-Precision (precision after R (= num_rel for a query) docs			
Exact:	0.3562	0.3439	0.3625

Table 2. Evaluation of French-French retrievals

We observe that using the terms of the BTQ together with words does not lead to a significant improvement in effectiveness. However, when the automatically built terms are also considered, an improvement of 7.35% is obtained. This suggests that, in our case, the automatically built terms have more impact on IR effectiveness than the manually established terms in the BTQ. Part of the reasons for this may be the following:

- The BTQ is not a general terminological base (as our test corpus is). Its aim is to suggest standard French terms to designate concepts in different specialized domains, as well as their correspondence with English terms. Many words in everyday language are not covered. For example, the word "mariage" (marriage) is not in the BTQ. This word designates the key concept in Query 2. So using the BTQ, we were not able to recognize a good part of important concepts in the queries and in the documents. On the other hand, some secondary concepts have been recognized as terms, for example "taux" (rate) in Query 2. When the recognized BTQ terms are combined with the words occurring in the original query (or document), the representation obtained often derive from the original content. So the retrieved documents do not correspond to the original query.

- The automatically built terms are closer to the test corpus than the manual base in the sense that all the built terms actually occur in the corpus, and frequently used terms are usually recognized. Although much noise is obtained, there is much less silence than when the manual base is used. This is why we did make some improvement in Run3.

In comparison with other experiments on French to French retrieval, our approaches performed reasonably well. Below is a comparison with the medium performance for our three runs:

Run	≥ medium	< medium
CLIPS2	13	8
CLIPS1	12	9
CLIPS3	14	7

Table 3. Comparison with the medium performance

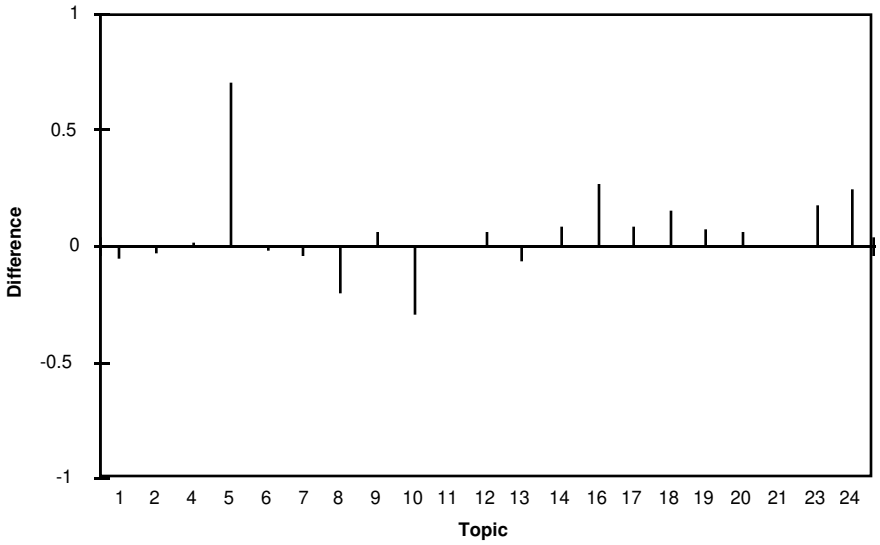
Figure 2 shows more details of the comparison for each query.

2.4. Cross-language retrieval

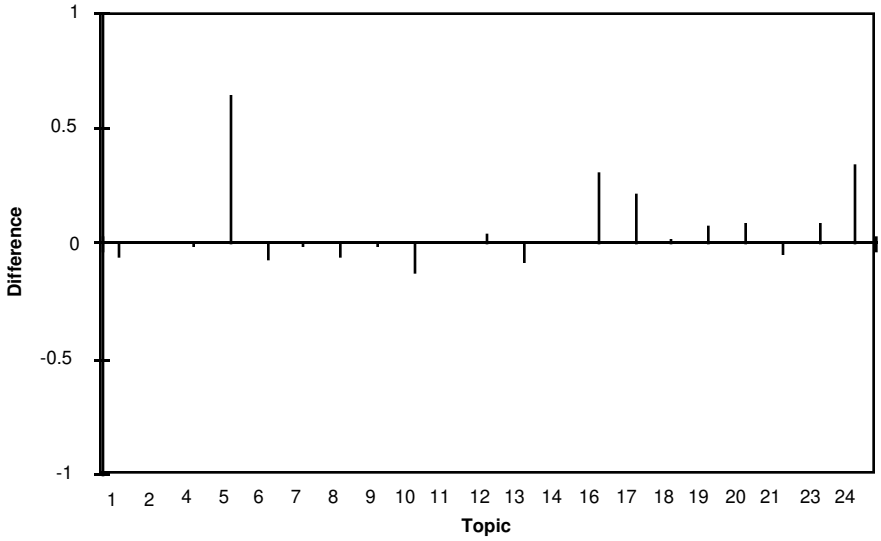
At the time of official submission of retrieval results, our implementation for cross-language retrieval was not finished. So no result was submitted. The implementation was completed later on. We tested the effectiveness of the BTQ for English to French retrieval (i.e. using English queries to retrieve French documents). As a point of comparison, we also used LOGOS™ translation system to translate the English queries into French, and then using the classical IR approach to retrieve documents. We obtained an average precision of 20.83% by this approach.

For cross-language retrieval using the BTQ, documents in French are indexed using both words and the BTQ French terms. English terms are extracted from the queries in a similar way to term extraction from documents. They are translated into their French equivalents by the BTQ, which are then used to retrieve French documents.

Run2:



Run1:



Run2

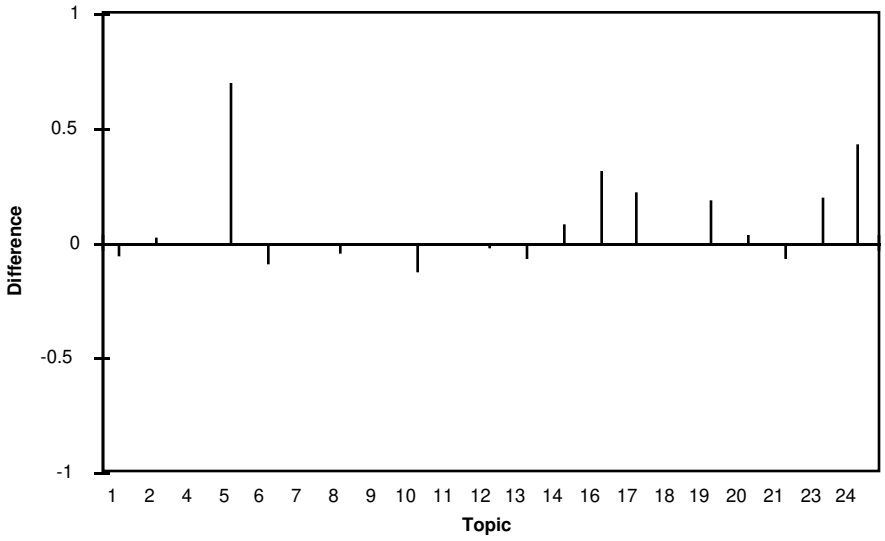


Figure 2. Comparison with the medium performance

Although this approach sounds valid in principle, we did not obtain many important concepts from the queries. We mentioned the problem with the word "mariage" in French for Query 2. This same problem also occurs for the English part of the BTQ. In addition, due to the language barrier, we could not combine the French BTQ terms with the English words in the original queries in order to diminish the silence in the representation (although the Cornell group made a surprising success in using English words to retrieve French documents, see in the same proceedings). Due to these facts, our use of the BTQ for cross-language retrieval led to a very low average precision of 7.98%. Our conclusion from this experiment is that the BTQ is not appropriate for general domain cross-language retrieval. It is still questionable whether it helps in cross-language retrieval in specialized domains.

3. Chinese IR

The difference between Chinese IR and IR for European languages lies in the fact that words are not separated in Chinese sentences. For example, the phrase "information retrieval system" is written as 信息检索系统. It is important to separate a sentence into smaller segments. Two types of segments may be used: N-grams or words.

3.1. Using N-grams

An N-gram is a subsequent string of N Chinese characters. For example, the string 信息检索系统 (information retrieval system) may be segmented into the following unigrams (N=1) and bigrams (N=2):

N=1: 信 息 检 索 系 统

N=2: 信息 息检 检索 索系 系统

In the case of bigrams, it is possible to consider the combinations of the first and the last characters with sentence boundaries or punctuation particles, that is _信 and 统_. However, punctuation particles and sentence boundaries are meaningless for IR purposes. We believe that such bigrams will have little impact on retrieval effectiveness. So we do not consider them.

It is possible to use longer N-grams for Chinese IR. However, it has been shown that bigrams are a good choice for Chinese IR [6].

3.2. Using words

This approach requires one to segment a Chinese sentence into words. This is not a trivial task because of the enormous amount of ambiguity. A sentence may often be segmented into several different sequences of legitimate words. For example, the sentence 现在本所有究生活动 (there is currently an activity for graduate students in our institute) contains the following legitimate words:

现 (now), 现在 (now),
在 (at),
本 (originally), 本所 (our institute),
所 (institute), 所有 (all, belong to),
有 (have),

研究 (research), 研究生 (graduate students),
生 (give birth), 生活 (life),
活 (live), 活动 (activity),
动 (move).

There are as many as 30 possible combinations of legitimate words. Only the following one is correct: 现在 本所 有 研究生 活动 (now / our institute / have / graduate student / activity). The key problem is to choose the correct segmentation among all the possible solutions.

There are two basic segmentation approaches for Chinese: the approach based on a dictionary, and the approach based on statistics (see [7] for discussions).

In the dictionary-based approach, one first finds all the legitimate words included in a sentence, then the longest-matching algorithm is applied to choose the sequence of words which covers the sentence with the longest words (or with the fewest words). A dictionary-based segmentation is usually augmented by a set of heuristic rules to recognize special sequences such as quantity-classifier sequences (e.g. 一千个 - one thousand [units of]).

On the other hand, a statistical approach relies on statistical data to determine possible words and to select the best word sequence. Statistical data are usually obtained from a set of manually segmented training texts. According to the frequency of occurrences and co-occurrences, one may determine how probable a string (possibly within some context) may be a word.

From the point of view of performance, the previous experiments showed that the two approaches have comparable accuracy. In our experiments, we used the dictionary-based word segmentation because no training text from this corpus was available. In addition of a word dictionary of 87 600 entries, we also dealt with some special character sequences which may be considered as words in Chinese. These sequences include: nominal pre-determiner and affix structure. A set of rules are set up for their recognition. For example, ２０００年 (year 2000), 第一回 (first time) are recognized as nominal pre-determiners, and "小朋友" (little friend) and "大众化" (popularize) as having an internal affix structure.

One particular problem we dealt concerns numbers. Numbers may be written in different ways in Chinese texts. For example, "the year 2000" may be written in Arabic numbers which may be encoded in ASCII (2000年) or in Chinese codes (２０００年). It may also be written in Chinese numbers as 二零零零年, or as a mixture of Chinese and Arabic numbers: 二０００年. Some of the queries contain numbers (e.g. a date). It is important to normalize them so that the same number becomes identical in documents and in queries. In our segmentation, a normalization is performed.

3.3. Particularity of segmentation for IR: Long vs. short words

When a dictionary is used, the maximum-matching algorithm is usually applied. However, as there is no clear definition of words in Chinese, in many Chinese dictionary, there are a number of long words/phrases that are composed of shorter words. For example:

long words	**component words**
环境污染 (environment pollution):	环境 (environment), 污染 (pollution)
安全措施 (security measure):	安全 (security), 措施 (measure)
电脑网络 (computer network):	电脑 (computer), 网络 (network)

If a long word/phrase is encountered, the shorter words contained in it are hidden. If we consider the segmentation problem from a different standpoint than IR (e.g. Machine Translation), this is not problematic. However, it will lower the recall ratio in IR. For example, if a document talks about 环境污染 (environment pollution) and a query asks for "污染" (pollution), the document will not be retrieved. In order to avoid this problem, our segmentation process extracts all the possible compound words (composed of two characters or more) from a given character string. So for the sequence 环境污染, three words will be extracted: 环境污染, 环境 and 污染. In one of our official runs submitted, this approach is used: all the compound words included in a long word are also extracted.

This approach may be further extended by also extract all the single-character words. So, for the string 环境污染, we also have the following segments extracted: 信, 息, 系, 统. This approach has been tested after our official submissions.

3.4. Experiments

We submitted two runs: one using bigrams and another using words. The answers submitted are also retrieved with SMART system. Notice that once the documents and the queries have been cut down into segments, SMART system may be used for their indexing and retrieval, after some modifications. A stop-list of 1460 elements is set up. This list contains frequently used functional words as well as symbols. The functional words included are usually prepositions, adverbs, tense particles and so on.

The evaluation of the two approaches is shown in Table 4. We observe that there is no significant difference between the two runs with regard to effectiveness. For most queries, both approaches perform quite well. However, for a few queries (CH33 and CH34 in particular), the average precision is quite low (from 0.0609 to 0.1320). This is the same for the other groups, too. In particular, Query CH33 is a difficult one for automatic IR systems. It asks for documents about particular events of airplane hijacking between mainland China and Taiwan, while documents discussing airplane hijacking at a general level (e.g. measures to prevent hijacking) are not relevant. In our top ranked answers, most documents concern airplane hijacking at a general level. A few other documents describe hijacking events in other countries.

We also tested the impact of number normalization on IR. Query CH35 is the one which involves most nominal pre-determiners. It asks for documents about the event of a particular date. Without normalization, we obtained 0.4663 for average precision. With the normalization, we obtained 0.5271. This shows that the normalization is necessary when querying documents with numbers.

After the official submissions, we tested the extended word segmentation approach by also extracting single characters. Using this approach, we obtained a better average precision of 46.15%. Still, this performance is not significantly different from that using bigrams. The reason possibly lies in the Chinese language itself: in Chinese, single characters (ideographs) may constitute a reasonably good representation of a text. In order to confirm this fact, we indexed the documents and the queries only by single characters. We obtained a quite high performance of 41.09%. Similar result is also obtained in [6]. This result clearly shows that the good performances obtained for both bigram- and word-based approaches is due in major part to the meaningfulness of Chinese characters. Then a possible explanation to the almost identical performances for both bigram- and word-based approaches is as follows: Both

Approach:	Bigrams	Words
Total number of documents over all queries		
Retrieved:	26000	26000
Relevant:	2958	2958
Rel_ret:	2709	2668
Interpolated Recall - Precision Averages:		
at 0.00	0.8546	0.8581
at 0.10	0.6844	0.7134
at 0.20	0.6165	0.6582
at 0.30	0.5556	0.5953
at 0.40	0.5236	0.5345
at 0.50	0.4815	0.4939
at 0.60	0.4252	0.4371
at 0.70	0.3711	0.3609
at 0.80	0.3043	0.2608
at 0.90	0.1912	0.1541
at 1.00	0.0284	0.0169
Average precision (non-interpolated) over all rel docs		
	0.4467	0.4524 (1.28%)
Precision:		
At 5 docs:	0.6615	0.6769
At 10 docs:	0.7000	0.6423
At 15 docs:	0.6359	0.6333
At 20 docs:	0.5981	0.6173
At 30 docs:	0.5705	0.5782
At 100 docs:	0.4408	0.4662
At 200 docs:	0.3379	0.3402
At 500 docs:	0.1899	0.1857
At 1000 docs:	0.1042	0.1026
R-Precision (precision after R docs retrieved):		
Exact:	0.4655	0.4748

Table 4. Comparison of the retrieval performances

bigrams and words allow to enhance the document and query representation in comparison with single characters. However, the enhancement is more limited in both cases than we could expect (increasing the average precision by about 10%). So the comparable performances of bigrams and words are due to their limited capacity to improve the document and query representation by single characters.

In comparison with the results of other groups, ours are not quite good. Table 5 shows the comparison of our results with the medium performance.

Run	≥ medium	< medium
Bigrams	6	20
Words	6	20

Table 5. Comparison with the medium performance

The reason of these poor performances is that we did not use any further techniques in our runs to improve the effectiveness, as several other groups did. Our goal in these tests is to compare

Chinese IR using bigrams and words on the same basis. By using some techniques such as using the top ranked documents to do feedback retrieval as well as combining bigrams with words, the effectiveness may be increased.

3.5. Time and space

It is also important to compare the two approaches with respect to the time and space requirements. Table 6 shows the comparison.

	Bigrams	Words
NB. of tokens	1 446 354	205 056
Segmentation time	1 h 04 mn	4 h 56 mn
Indexing time	13 h 12 mn	1 h 19 mn
Total document processing time	14 h 16 mn	6 h 15 mn
Retrieval time	7 mn	5 mn
Space	1.2 Gb	0.5 Gb

Table 6. Time and space for IR using bigrams and words

This table shows that the total processing time for Chinese IR using bigrams are more than twice longer than that using words. So is the comparison on the space needed for their indexing. We can conclude that on the time and space criterion, IR using words has advantages over IR using bigrams.

4. Conclusions

We compared an IR approach based on terms with the classical word-based approach in French and cross-language IR. For French IR, it is shown that terms, together with words, may bring improvements to the system performance. We noticed that using a manually established terminological base leads to little improvement over the classical approach. However, using both the manual base and the automatic base brings a significant improvement. The little improvement by using the manual terminological base is due to its poor coverage of the test corpus. Many important concepts are not recognized. By incorporating an automatically built term base, the coverage is increased. So, we were able to achieve a better performance.

The manual terminological base is also used in the cross-language retrieval after the official submission The result is disappointing. It is much lower than those obtained by the other groups (which is about 50% of the monolingual IR). Again, the primary reason is the poor coverage of the terminological base used. In particular, in this case, the base is left alone to recognize the concepts from the queries (without the help of words). The poor coverage has complete impact on the global result of this test.

For Chinese IR, two approaches have been compared: one is based on bigrams and the other on words. From the point of view of effectiveness, there is no significant difference between the two approaches. However, if we also consider time and space, the approach based on words has some advantages. So, despite the comparable effectiveness, we still advocate the word-based approach if one has to chose between them. Moreover, it seems much easier to do cross-language

retrieval with words than with N-grams. We also believe that the word-based approach has not reached its limits. Improvements are still possible, for example, by integrating more heuristic rules to recognize more special sequences, or by incorporating a thesaurus. This is part of our future work.

Acknowledgment: This research is partly supported by a grant for France-Quebec research cooperation in linguistic engineering.

References

1. M.-F. Bruandet, Outline of a knowledge base model for an intelligent information retrieval system. *Information Processing and Management*, vol. 25, pp. 89-115 (1989).
2. C. Buckley, Implementation of the SMART information retrieval system. Cornell University, Technical report 85-686, (1985).
3. J. Fagin, Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic methods. in *Computer Science*: Cornell University, 1988.
4. V. Güntzer, G. Jüttner, S. G., and F. Sarre, Automatic thesaurus construction by machine learning from retrieval sessions. *Information Processing & Management*, vol. 25, pp. 265-273 (1989).
5. H. Kimoto and T. Iwaderie, Construction of a dynamic thesaurus and its use for associated information retrieval. *13th ACM-SIGIR Conference*, 227-240 (1990).
6. K. L. Kwok, Comparing representations in Chinese information retrieval. *ACM-SIGIR'97*, 34-41 (1997).
7. J.-Y. Nie and M. Brisebois, On Chinese text retrieval. *ACM-SIGIR'96*, Zürich, 225-233 (1996).
8. J.-Y. Nie, J.-P. Chevallet, and Y. Chiaramella, Vers la recherche d'information à base de termes. *1er Journéees Scientifiques et Techniques FRANCIL*, Avignon, 119-125 (1997).
9. C. J. v. Rijsbergen, A theoretical basis for the use of co-ocurrence data in information retrieval. *Journal of Documentation*, vol. 33, pp. 106-119 (1977).
10. K. Sparck-Jones, Notes and references on early automatic classification work. *SIGIR Forum*, vol. 25, pp. 10-17 (1991).