Ricoh at CLEF 2004

Yuichi Kojima

Software R&D Group, RICOH CO., Ltd.
1-1-17 Koishikawa, Bunkyo-ku, Tokyo 112-0002, JAPAN
ykoji@src.ricoh.co.jp

Abstract. This paper describes the participation of RICOH in the monolingual and cross-lingual information retrieval tasks on German Indexing and Retrieval Testdatabase (GIRT) in the Cross-Language Evaluation Forum (CLEF) 2004. We used a morphological analyzer for word decompounding and parallel corpora for cross-lingual information retrieval. The performance of cross-lingual information retrieval was poor and that of monolingual information retrieval was not good. We need to check our modules and procedures.

1 Introduction

We are enhancing our information retrieval system for some languages [1, 2]. Our approach to the enhancement is to use same basic system and modify language depend modules. Our system showed reasonable performance for some European languages and the importance of word decompounding for the compound rich languages such as German in the participation in CLEF 2003 tasks [2].

This is our second participation in CLEF tasks. We used a commercial morphological analyzer for word decompounding and participated in GIRT tasks. Our focuses for this year were:

1. to confirm the performance of word decompounding
2. to find the problems in applying our approach to cross-lingual information retrieval

Section 2 of this paper outlines our system, section 3 describes the modifications made for the experiments, section 4 gives the results, and section 5 contains some conclusions.

2 Description of the System

The basic system is same as last year. Before describing our new modifications to European languages, we give an outline of the system as background information. It uses a document ranking method based on the probabilistic model [3] with query expansion using pseudo-relevance feedback [4] and was shown to be effective in TREC and NTCIR experiments.

In the following sections, we explain the processing flow of the system [5, 6].

2.1 Query term extraction

We used “title” and “description” fields of each topic. An input topic string is transformed into a sequence of stemmed tokens using a tokenizer and stemmer. Stop words are eliminated using a stopword dictionary. Two kinds of terms are extracted from stemmed tokens for initial retrieval: a “single term” is each stemmed token and a “phrasal term” consists of two adjacent tokens in a stemmed query string.

2.2 Initial retrieval

Each query term is assigned a weight w_t, and documents are ranked according to the score $s_{q,d}$ as follows:

$$w_t = \log\left(k_4^{'} \cdot \frac{N}{n_t} + 1\right) \tag{1}$$

$$s_{q,d} = \sum_{t \in q} \frac{f_{t,d}}{K + f_{t,d}} \cdot \frac{w_t}{k_4^{'} \cdot N + 1} \tag{2}$$

$$K = k_1\left(\left(1 - b\right) + b\frac{l_d}{l_{ave}}\right) \tag{3}$$

where N is the number of documents in the collection, n_t is the document frequency of the term t, $f_{t,d}$ is the in-document frequency of the term, l_d is the document length, l_{ave} is the average document length, and k'_4, k_1 and b are parameters.

Weights for phrasal terms are set lower than those for single terms.

2.3 Query expansion

As a result of the initial retrieval, the top 10 documents are assumed to be relevant (pseudo-relevance) to the query and selected as a "seed" for query expansion. Candidates for expansion terms are extracted from the seed documents in the same way as for the query term extraction mentioned above. Phrasal terms are not used for query expansion. The candidates are ranked on the Robertson's Selection Value [7], or RSV_t and the top ranked terms are selected as expansion terms. The weight is re-calculated as $w2_t$ using the Robertson/Sparck-Jones formula [8].

$$RSV_t = w2_t \cdot \left({r_t}/{R} - {n_t}/{N} \right) \tag{4}$$

$$w2_t = \alpha \cdot w_t + (1-\alpha) \cdot \log \frac{\dfrac{r_t + 0.5}{R - r_t + 0.5}}{\dfrac{n_t - r_t + 0.5}{N - n_t - R + r_t + 0.5}} \tag{5}$$

where R is the number of relevant documents, r_t is the number of relevant documents containing the term t and α is a parameter.

The weight of the initial query term is re-calculated using the same formula as above, but with a different α value and an additional adjustment to make the weight higher than the expansion terms.

2.4 Final retrieval

Using the initial query and expansion terms, the ranking module performs a second retrieval to produce the final result.

2.5 Cross-lingual retrieval

We performed English-to-German retrieval using well-known strategy and parallel corpora [9]. In the cross-lingual retrieval process, the English query is submitted against the English database and top-n documents are obtained. Their counterparts in the German database are exploited as seed documents to extract German query terms. The extraction can be performed using completely same mechanism for query expansion in pseudo-relevance feedback.

3 Experiments

There are five items in the system which needs adjustments depending on the language, 1) the tokenizer, 2) the stemmer, 3) the stopword dictionary, 4) the training data and 5) the parallel corpora.

We used mostly the same modules as last year and a commercial morphological analyzer which can tokenize a sentence, decompose a compound word, and stem a word.

Details of the items in the system are as follows:

3.1 Stemming and tokenizing

We had a selection of possible combinations of the stemmers and the tokenizers. The system can use Snowball stemmer [10] and simple tokenizer which were used for the last year's CLEF experiments, and the morphological analyzer which is imported into the system this year.

The possible combinations are limited by the behavior of the analyzer. It decomposes a compound word into its single words and stems each single word in the same procedure. So there is no selection of word decompounding without the stemming in the analyzer.

After some experiments, we selected the combination of 1) word decompounding and 2) a two step stemming which consists of the first stemming step of the decompounding and the second stemming step using snowball stemmer.

Table 1 shows the summary of our experiments.

Table 1. Summary of the experiments

Word decompounding	stemming	Average precision [*]
No	German Snowball stemmer	0.3149
Yes	Stemmer A[**]	0.2944
Yes	Stemmer A + German Snowball stemmer	0.3470

[*] Average precision using GIRT German monolingual task of CLEF 2003 after training

[**] German Stemmer in the analyzer

3.2 Stopword dictionary

This year, we used stopword dictionaries at Snowball site.

3.3 Parallel corpora

We prepared additional two document databases using English and German GIRT corpus. First database was made from the English corpus by extracting each tagged entity (TITLE, AUTHOR and ABSTRACT) as a document and used for making lists of seed documents. Second database was made from the German corpus by the same method and used for making German queries from the lists of seed documents.

Each document was tokenized and stemmed depend on its language using the method mentioned above.

Although we cannot expect all of corpus to be parallel in the real situation, we used all parallel corpora because we could not get average score even using them.

3.4 Training

We searched the parameters of the system by the hill-climbing method, using average precision values of search results with query expansion for the monolingual and cross-lingual retrieval task.

Table 2 shows the average precision values after training.

Table 2. Average precision values after training

Language	Average Prec.
DE -> DE	0.3470
EN -> DE	0.1370

4 Results

Table 3 shows the summary of our official results for CLEF 2004.

The result of monolingual task was worse than our expectation. According to our estimate using our experiments, the average precision value can be 0.1 point higher than the current value if attributes of queries are similar to last year. As for the result of cross-lingual task, although the value was bad, it was what we were expected.

Table 3. Official runs for CLEF 2004

Language	Run	Relevant	Rel. Ret.	Average Prec.	R-Precision
DE -> DE	rdedetde04	1663	922	0.2381	0.2759
EN -> DE	rendetde04	1663	684	0.1261	0.1678

5 Conclusions

We tested our new module for word decompounding and checked problems in applying our approach to cross-lingual retrieval. According to our experiments, word decompounding is effective. But the result of the official experiment was not good. We need check of our procedure to make data for submission. On the other hand, we could not achieve the average score in cross-lingual retrieval tasks (both our experiment and official experiment). We will check modules and the procedure for cross-lingual retrieval.

Further analysis with additional experiments will be shown in the CLEF workshop.

References

[1] Y. Kojima, H. Itoh, H. Mano, and Y. Ogawa. Ricoh at CLEF 2003. At http://clef.iei.pi.cnr.it:2002/2003/WN_web/26.pdf.
[2] Y. Kojima and H. Itoh. Ricoh in the NTCIR-4 CLIR Tasks. At http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/CLIR/NTCIR4WN-CLIR-KojimaY.pdf
[3] S. E. Robertson and S. Walker. On relevance weights with little relevance information. In Proceedings of the 20th Annual International ACM SIGIR Conference (SIGIR '97), pages 16-24, 1997.
[4] Y. Ogawa and H. Mano. RICOH at NTCIR-2. In Proceedings of the Second NTCIR Workshop Meeting, pages 121-123, 2001.
[5] H. Itoh, H. Mano and Y. Ogawa. RICOH at TREC-10. In the Tenth Text Retrieval Conference (TREC-2001), page 457-464, 2001.
[6] M. Toyoda, M. Kitsuregawa, H. Mano, H. Itoh and Y. Ogawa. University of Tokyo/RICOH at NTCIR-3 Web Retrieval Task. At http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-WEB-ToyodaM.pdf
[7] S. E. Robertson. On term selection for query expansion. Journal of Documentation, 46(4):359-364, 1990.
[8] S. E. Robertson and K. Spark-Jones. Relevance weighting of search terms. Journal of ASIS, 27:129-146, 1976
[9] H. Itoh. NTCIR-4 Patent Retrieval Experiments at RICOH. At http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/PATENT/NTCIR4WN-PATENT-ItohH.pdf
[10] Snowball web site. At http://snowball.tartarus.org/ visited 7th November 2002.