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 Ricoh's participation is monolingual and bi-lingual information retrieval tasks done on the German Indexing and Retrieval Testdatabase (GIRT) at the Cross-Language Evaluation Forum (CLEF) 2004. We used a commercial morphological analyzer to decompound words and parallel corpora to retrieve bi-lingual information. While monolingual information retrieval was improved by using the analyzer, bi-lingual information retrieval still has room for improvement.

1 Introduction

We are enhancing our system of retrieving information in some languages [1, 2]. Our approach is to use the same basic system and modify language dependent modules. Our system performed reasonably with some European languages and revealed the importance of decompounding words in compound-rich languages such as German in the CLEF 2003 tasks [2].

This is the second time we have participated in CLEF tasks. We used a commercial morphological analyzer to decompound words and also participated in GIRT tasks. Our focus this year was:

1. To evaluate the effectiveness of word decompounding
2. To discover problems in applying our approach to bi-lingual information retrieval

Section 2 of this paper outlines our system, Section 3 describes the modifications we made to the experiments, Section 4 presents the results, and Section 5 is the conclusion.

2 System Description

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

We will now explain the processing flow for the system [5, 6].

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 292 – 297, 2005.

2.1 Query Term Extraction

We used "title" and "description" fields for 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 the 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 weight w_t, and documents are ranked according to score $s_{q,d}$ as follows:

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

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

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

where N is the number of documents in the collection, n_t is the document frequency of the term t, and $f_{t,d}$ is the in-document frequency of the term. Here, l_d is the document length, l_{ave} is the average document length, and k_4', k_1, and b are parameters.

The 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 were assumed to be relevant (pseudo-relevance) to the query and selected as a "seed" for query expansion. Candidates for expansion terms were extracted from the seed documents in the same way as for the query term extraction previously explained. Phrasal terms were not used for query expansion. The candidates were ranked on Robertson's Selection Value [7], or RSV_t, and the top ranked terms were selected as expansion terms. The weight was recalculated as $w2_t$ using the Robertson/Sparck-Jones formula [8].

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

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

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

The weight of the initial query term was 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 does a second retrieval to produce the final results.

2.5 Bi-lingual Retrieval

We did English-to-German retrieval using a well known strategy based on English-German parallel corpora [9]. The bi-lingual retrieval process involved the following: 1) an English query was used for retrieval from the English database, 2) top-n documents were used to extract German query terms, 3) German query terms were extracted from counterparts of documents in the German database using the same mechanism for query expansion as in pseudo-relevance feedback regarding the counterparts as seed documents, and 4) the terms were used for retrieval from the German database.

3 Experiments

There were five items in the system that needed adjustment depending on the language, 1) the tokenizer, 2) the stemmer, 3) the stopword dictionary, 4) the training data, and 5) the parallel corpora.

We mainly used the same modules as last year and a commercial morphological analyzer that could tokenize a sentence, decompose a compound word, and stem a word.

Details on the items in the system are given in the following.

3.1 Stemming and Tokenizing

We had a selection of possible combinations of stemmers and tokenizers. The system could utilize the Snowball stemmer [10] and simple tokenizer that we used for last year's CLEF experiments. The system could also utilize the morphological analyzer that we imported into the system this year.

The possible combinations were limited by the behavior of the analyzer. It decomposed a compound word into single words and stemmed each single word with the same procedure. In other words, word decompounding was not selected without stemming in the analyzer.

After various experiments, we selected a combination of 1) word decompounding and 2) two-step stemming, which consisted on the first stemming step for decompounding and the second stemming using the Snowball stemmer.

Table 1 lists the results of the preliminary experiments in CLEF 2003 tasks.

Table 1. Results of preliminary 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 for CLEF 2003 after training
** German stemmer in the analyzer

3.2 Stopword Dictionary

This year, we used stopword dictionaries at the Snowball site.

3.3 Parallel Corpora

We prepared two additional document databases using the English and German GIRT corpus. We first prepared a database from the English corpus by extracting each tagged entity (TITLE, AUTHOR and ABSTRACT) as a document and used these for making lists of seed documents. We prepared the second database from German corpus with the same procedure used for making the German query terms from the lists of seed documents.

Each document was tokenized and stemmed depending on its language with the above mentioned methods.

We used all, half and a quarter of the parallel corpora to evaluate the performance.

3.4 Training

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

Table 2 lists the average precision values after training.

Table 2. Average precision values after training

Language	Average precision	Years for documents used to prepare German query terms
DE->DE	0.3470	-
EN->DE	0.2644	1990-2000 (45-Mbyte English documents)
EN->DE	0.2449	1997-2000 (28 Mbytes)
EN->DE	0.1819	1999-2000 (16 Mbytes)

4 Results

Table 3 lists the summary of our results for CLEF 2004.

Our submitted results, rdedetde04 and rendetde04, had bugs during processing, so we prepared unofficial1 and unofficial4 instead of these. We also achieved results with other settings to observe the behavior of the system. The unofficial3 setting was the same as last year's. The unofficial5 and unofficial6 settings were to check what influence the document data capacity had.

The results for the monolingual task were improved with decompounding. Comparing unofficial1, unofficial2 and unofficial3, decompounding contributed to an improvement of about 17%. The results for the bi-lingual task were worse than those for training. The performance decreased by about 25% for bi-lingual retrieval while it only decreased by 2% for monolingual retrieval. The decreased performance from full-document to half-document size was smaller than that from half-document to quarter-document size. The former was 4% and the latter was 25%.

Table 3. Results for CLEF 2004

Language	Run-id	Relevant	Rel.Ret.	Average Prec.	R-Precision
DE->DE	Unofficial1	1663	1082	0.3393	0.3711
DE->DE	Unofficial2	1663	1072	0.2890	0.3203
DE->DE	Unofficial3	1663	1068	0.2828	0.3211
EN->DE	Unofficial4	1663	1030	0.1972	0.2392
EN->DE	Unofficial5	1663	961	0.1893	0.2198
EN->DE	Unofficial6	1663	917	0.1419	0.1827
DE->DE	Rdedetde04	1663	922	0.2381	0.2759
EN->DE	Rendetde04	1663	684	0.1261	0.1678

Unofficial1: Results using commercial morphological analyzer and Snowball stemmer
Unofficial2: Results using commercial morphological analyzer and Snowball stemmer without decompounding
Unofficial3: Results using Snowball stemmer and simple tokenizer
Unofficial4: Results using documents in 1990-2000 and unofficial1 setting
Unofficial5: Results using documents in 1997-2000 and unofficial1 setting
Unofficial6: Results using documents in 1999-2000 and unofficial1 setting
Rdedetde04: Results using commercial morphological analyzer and Snowball stemmer
Rendetde04: Results using documents in 1990-2000 and unofficial1 setting

5 Conclusion

We tested our new module for decompounding words and investigated problems we encountered in applying our approach to bi-lingual retrieval. The word decompounding that we used effectively improved performance by 17% according to our experiment. However, the results for bi-lingual information retrieval showed decreased performance from training to the experiment by about 25%, meaning there is room to improvement. The decreased performance from full to quarter documents indicates we require a reasonable document data capacity.

We intend to improve bi-lingual information retrieval and enhance target bi-lingual sets in future work.

References

1. Kojima, Y., Itoh, H., Mano, H., Ogawa, Y.: Ricoh at CLEF 2003. In: C. Peters, J. Gonzalo, M. Braschler, M. Kluck (eds.): Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. Lecture Notes in Computer Science, 3237: Berlin/Heidelberg/New York: Springer (2004) 367-372 online at http://www. springerlink.com/openurl.asp?genre=article&issn=0302-9743&volume=3237&spage=367
2. Kojima, Y., Itoh, H.: Ricoh in the NTCIR-4 CLIR Tasks. At http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/CLIR/NTCIR4WN-CLIR-KojimaY.pdf
3. Robertson, S. E., Walker, S.: On relevance weights with little relevance information. In: Proceedings of the 20th Annual International ACM SIGIR Conference (SIGIR '97), 16-24, 1997.
4. Ogawa, Y., Mano, H.: RICOH at NTCIR-2. In Proceedings of the Second NTCIR Workshop Meeting, pp. 121-123, 2001.
5. Itoh, H., Mano, H., Ogawa, Y.: RICOH at TREC-10. In: The Tenth Text Retrieval Conference (TREC-2001), pp.457-464, 2001.
6. Toyoda, M., Kitsuregawa, M., Mano, H., Itoh, H., Ogawa, Y.: University of Tokyo / RICOH at NTCIR-3 Web Retrieval Task. At http://research.nii.ac.jp/ntcir/workshop/Online Proceeding3/NTCIR3/NTCIR3-WEB-ToyodaM.pdf
7. Robertson, S. E.: On term selection for query expansion. Journal of Documentation, 46 (4): 359-364, 1990
8. Robertson, S. E., Sparck-Jones, K.: Relevance weighting of search terms. Journal of ASIS, 27: 129-146, 1976.
9. Itoh, H.: 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.