KUNLP System for NTCIR-3 English-Korean Cross-Language Information Retrieval

Hee-Cheol Seo, Sang-Bum Kim, Baeg-Il Kim, Hae-Chang Rim and Sang-Zoo Lee
Dept. of Computer Science and Engineering, Korea University
1, 5-ka, Anam-dong Seongbuk-Gu, Seoul, 136-701, Korea
{hcseo, sbkim, cedar, rim, zoo}@nlp.korea.ac.kr

Abstract

This paper describes KUNLP system for the English-Korean cross-language information retrieval track in NTCIR-3 workshop and some experiments after the workshop. Query translation method based on the bilingual dictionary and the document language corpus was used. To automatically transliterate some proper nouns such as Korean person names, Korean place names, and Korean company names, we have constructed the bilingual biographical dictionary, and collected the corresponding translations of Korean place names and Korean company names. We submitted a monolingual run and three cross-language runs, which used only a description field of each topic as a query. Cross-language runs were classified as to whether query expansion was used and whether manual transliteration was applied. Comparisons between cross-language runs show that query expansion is useful in the English-Korean cross-language information retrieval and transliteration also improves the system performance.

And additional experiments after NTCIR-3 workshop show that the Korean query which consists of the best translation equivalents for English query terms is more effective than that consisting of two or more translation equivalents. In addition, including English acronyms and initial words in the Korean query is helpful to retrieve Korean documents.

Keywords: *English-Korean cross-language information retrieval, query translation, query expansion, query transliteration.*

1 Introduction

For the task of the English-Korean cross-language information retrieval(CLIR), which retrieves Korean documents by using English queries, we try to translate English queries into Korean by looking up a bilingual dictionary and then to retrieve Korean documents using the translated queries.

Although there might be many issues related to English-Korean CLIR, we focused on a query expansion and a proper noun transliteration. It is well known that query expansion improves the CLIR system performance [1][2][6]. However, there had not been studied about the effect of the query expansion in English-Korean CLIR as far as we know. Therefore, we have investigated whether query expansion was also effective in English-Korean CLIR. In developing an English-Korean CLIR system, proper nouns cause a problem for the query translation because they are usually not in a bilingual dictionary and can't be translated into Korean. To solve the problem, we try to devise a method of the automatic construction of a bilingual dictionary which includes Korean person names, Korean company names, Korean place names and transliterated English equivalents.

The paper is organized as follows: Section 2 describes the architecture of KUNLP system, which consists of query term extraction, query expansion, query translation, and document retrieval. We present official result of NTCIR-3 workshop and analysis in section 3, and additional experiments after NTCIR-3 workshop in section 4. Finally, section 5 presents our conclusion.

2 Architecture of KUNLP System

KUNLP system retrieves Korean documents relevant to English queries by performing following steps: query term extraction step, query expansion step, query translation step, and document retrieval step. Figure 1 shows the architecture of KUNLP system.

At the query term extraction step, English collocations are recognized by looking up English dictionary which has the information of English collocations as well as words, and the stop words are removed from the English query. The Okapi relevance feedback method[5] was adopted to expand the English queries. At the query translation step, we collect every possible Korean equivalent of English terms according to the English-Korean dictionary, and then select feasi-

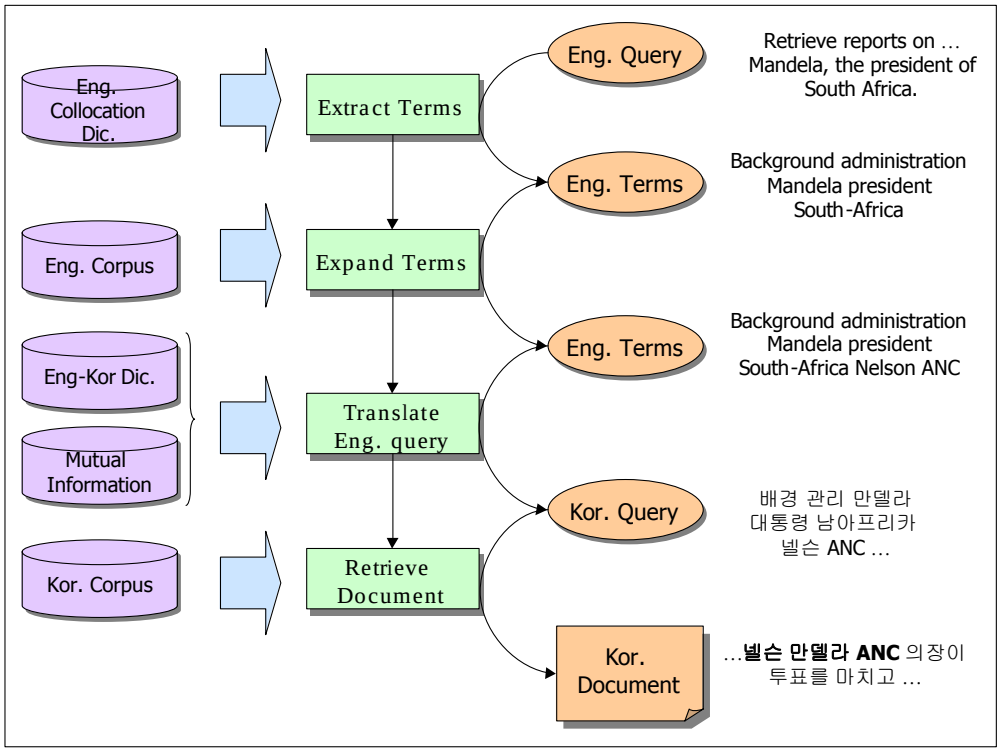

Figure 1. Architecture of KUNLP System

ble equivalents based on the co-occurrence information between Korean words. Some proper nouns that are not in the English-Korean dictionary are transliterated. Finally, we search the documents relevant to a translated Korean query as well as an original English query using the Okapi model[5]. Each step is described in detail in the following subsections.

2.1 Query Term Extraction

In this paper, query terms mean words or collocations which seem to be useful to search documents. Query terms are extracted by performing following processes: query word tokenization, collocation recognition, and stop words removal. The remaining terms in the query are regarded as query terms.

The tokenizer separates special symbols from words and filters out the special symbols. English collocations were recognized by looking up an English-Korean dictionary which has English collocation information as well as the word information. And finally, we removed the words belonging to a list of stop words, which consists of article, pronoun, proposition and so on. The list of stop words also includes the words irrelevant to the user information need, such as *retrieve, document* and *report*, which we manually picked out from English queries.

Table 1 shows the term extraction process with an example. The words in a collocation are linked with hyphen(-), such as "*South-Africa*".

2.2 Query Expansion

We expand queries before the query translation process. We use the Okapi relevance feedback method to expand English queries. The English corpus used for query expansion is Financial Times corpus extracted from TREC corpus , which is similar to target Korean corpus, KED010, for the set of economic newspapers.

To expand a query, we first retrieve English documents relevant to an English query based on the Okapi BM25. Every term that occurs in more than one document among the top 10 documents is sorted according to its relevance weight, which is calculated as follows[5]:

$$RW(t) = r_t log\frac{N}{n_t} - log\binom{R}{r_t} - logV \qquad (1)$$

where $RW(t)$ represents the relevance weight of the term *t* related to the relevance feedback, *R* the total number of the relevant documents(the value of *R* is 10, since we used the top 10 documents), r_t the number of the relevant documents in which term *t* occurs, *N* the

Table 1. Example of Query Term Extraction

query (NTCIR Topic 001)	*Retrieve reports on background and administration of Mandela, the president of South Africa.*
tokens	*Retrieve reports on background and administration of Mandela the president of South Africa*
collocations	*South-Africa*
stopwords	*Retrieve reports on and of the*
query terms	*background administration Mandela president South-Africa*

size of the collection, n_t the number of the documents which contain the term *t*, and *V* the size of the vocabulary.

And then the terms that belongs to the top 10 ranked terms and whose relevance weights are larger than 0 are added to the original query. For example, the English query is expanded to include terms "*Nelson, ANC, Congress ...*" at the NTCIR topic 001.

2.3 Query Translation

Translation equivalents are collected by referring to two English-Korean dictionaries: a biographical dictionary and a general dictionary. The English-Korean biographical dictionary includes Korean person names in English and their Korean transliterations. English-Korean general dictionary includes Korean translations of general English words.

At English-Korean CLIR, English queries may have Korean person names (e.g. *Kim Il Sung, Kim Jung Il*), and hence the English-Korean biographical dictionary is required. The biographical dictionary is used to transliterate the Korean person names written in English into Korean.

The English-Korean biographical dictionary was automatically constructed using English-Korean syllable mapping table (i.e. "*Kim*" - "김", "*Il*" - "일", "*Sung*"-"성"). Korean person name consists of the first name and the last name (i.e. "*Kim Jung Il*" consists of the first name "*Jung Il*" and the last name "*Kim*"). In Korean, while the number of the last names is about 100, the number of the first names is very large. Fortunately, the first name usually consists of one or two syllables and every possible pair of syllables constitute the first names. And then, we constructed the English-Korean biographical dictionary by connecting the first name and the last name. The method of constructing the bilingual biographical dictionary can be summarized as follows:

$$\begin{aligned} BioDic &= S_{ln} \cup S_{fn} \cup (S_{ln} \times S_{fn}) \\ S_{fn} &= S_{syllable} \cup (S_{syllable} \times S_{syllable}) \\ S_{ln} &= \{\textit{Lee-}\text{이}, \textit{Kim-}\text{김}, \textit{Park-}\text{박}, ...\} \\ S_{syllable} &= \{\textit{Il-}\text{일}, \textit{Sung-}\text{성}, ...\} \end{aligned} \quad (2)$$

where *BioDic* represents the English-Korean bilingual biographical dictionary, S_{ln} the set of Korean last names, S_{fn} the set of Korean first names, and $S_{syllable}$ the set of Korean syllables. Table 2 shows some examples of English-Korean biographical dictionary entries. The name of LN type consists of only the last name, the name of FN type only the first name and the name of LN+FN type the last name and the first name.

Table 2. Example of English-Korean biographical dictionary entries

English	Korean	type
Il Sung	일성	FN
Jung Il	정일	FN
Kim	김	LN
Kim Il Sung	김일성	LN+FN
Kim Jung Il	김정일	LN+FN

Translation equivalents of terms that are not Korean person name are collected by looking up an general dictionary. The information of Korean place names(i.e. *Pyongyang, Seoul*) and Korean company names(i.e. *Hyundai, Samsung*) are required in case where Korean documents are retrieved using the English query. Hence, we added the information of Korean place names and Korean company names to the general dictionary.

The English terms that are not in the two bilingual dictionaries are transliterated manually. For example, "*Mandela*" in the NTCIR topic 001 was transliterated as "만델라" manually.

After collecting the translation equivalents of the query terms, the plausible translation equivalents are selected by the co-occurrence information. We adopted a weighted mutual information(wMI) as a co-occurrence information. Because MI prefers the word pairs that the frequency of each word is very low, we used a weighted MI instead of MI. For each Korean word, we calculate a weighted MI with other words at a sentence in Korean text collection. The target Korean corpus, KED010 corpus, is used as the Korean

text collection. Equation of a weighted MI is followed:

$$wMI(w_1, w_2) = log(f(w_1, w_2)) \times MI(w_1, w_2) \quad (3)$$

$$MI(w_1, w_2) = log \frac{Pr(w_1, w_2)}{Pr(w_1)Pr(w_2)} \quad (4)$$

where $wMI(w_1, w_2)$ represents weighted MI between a term w_1 and a term w_2, $f(w_1, w_2)$ the frequency that w_1 and w_2 co-occurred in the same sentences, $Pr(w_1, w_2)$ the probability that w_1 and w_2 co-occurred in the same sentences, and $Pr(w_1)$ the probability of w_1.

Two translation equivalents are selected for each query term according to the cohesion of the translation equivalents, which is calculated as follows:

$$C(te_{ij}) = \sum_{k=1, k \neq i}^{n} \sum_{l=1}^{m_k} wMI(te_{ij}, te_{kl}) \quad (5)$$

where $C(te_{ij})$ represents the cohesion of the *j*-th translation equivalent of *i*-th query term te_{ij}, *n* the number of query terms, and m_k the number of translation equivalents of the *k*-th query term.

2.4 Document Retrieval

After translating the English query into Korean, the documents relevant to the query are searched. English query as well as the translated Korean query are used since Korean documents sometimes include English terms, such as "*WTO*", "*GATT*", and "*APEC*".

For Korean terms, just noun words were sent to the retrieval system, which are extracted by Korean POS tagger[3](hereafter HanTag). And for English terms, all English terms except stop words were used.

Korean document retrieval system uses the Okapi BM25(2,0, 0.0, inf, 0.8) formula including the length normalization controlled by *b*=0.8.

3 Official Results and anlysis

We have submitted one monolingual run and three cross-lingual runs using only description field:

- KUNLP-K-K-D-01(hereafter KKD01): a monolingual run without query expansion.
- KUNLP-E-K-D-01(hereafter EKD01): a cross-language run without query expansion.
- KUNLP-E-K-D-02(hereafter EKD02): a cross-language run with automatic English query expansion.
- KUNLP-E-K-D-03(hereafter EKD03): a cross-language run with automatic English query expansion, and manual transliteration of the terms that are not in two bilingual dictionaries.

Table 3 presents the official results of four runs. Figure 2 shows recall-precision curves and Figure 3 shows the average precision per query.

Table 3. official results

Run	relax		rigid	
	avg.P.	R-prec.	avg.P.	R-prec.
EKD01	0.1422	0.1933	0.1306	0.1547
EKD02	0.1795	0.2191	0.1691	0.1967
EKD03	0.1987	0.2374	0.1750	0.2102
KKD01	0.2693	0.3234	0.2292	0.2693

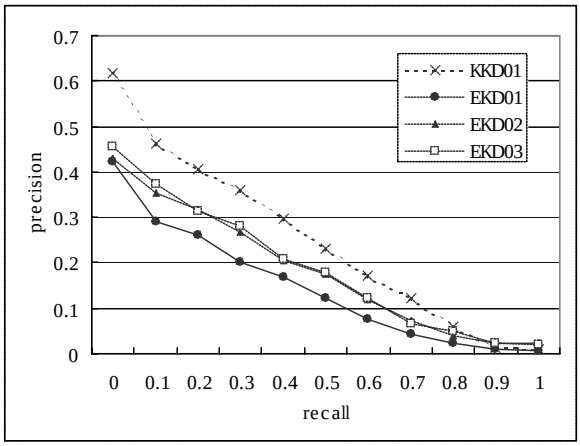

Figure 2. Recall-Precision Curve (Rigid Assessment)

3.1 Monolingual vs. cross-language

The results show that the monolingual run is better than the cross-language runs. This is due to the limited coverage of the bilingual dictionary for query translation and the query translation error. NTCIR topic 025 is about *Go* game in Asia, and hence one of the most important terms in the query is term "*Go*". But our dictionary doesn't have any information of "*Go*" in terms of game meaning, and the suitable equivalent can't be selected. As a result, our system can't retrieve documents related to the topic 025.

However, there are some queries that CLIR runs do better than the monolingual run. Some words in a translated query may express the user need better than those of the query for monolingual IR system. For example, NTCIR topic 006 is about "*Liner Caught Fire in South Korea*", and the Korean query for monolingual IR system includes a Korean word "*정기선*"(it means "*liner*") which doesn't occur in the relevant documents. The translated query for topic 006 had Korean words "*배*" and "*선박*"(both mean "*ship, vessel*"), which do occur in the relevant documents. These

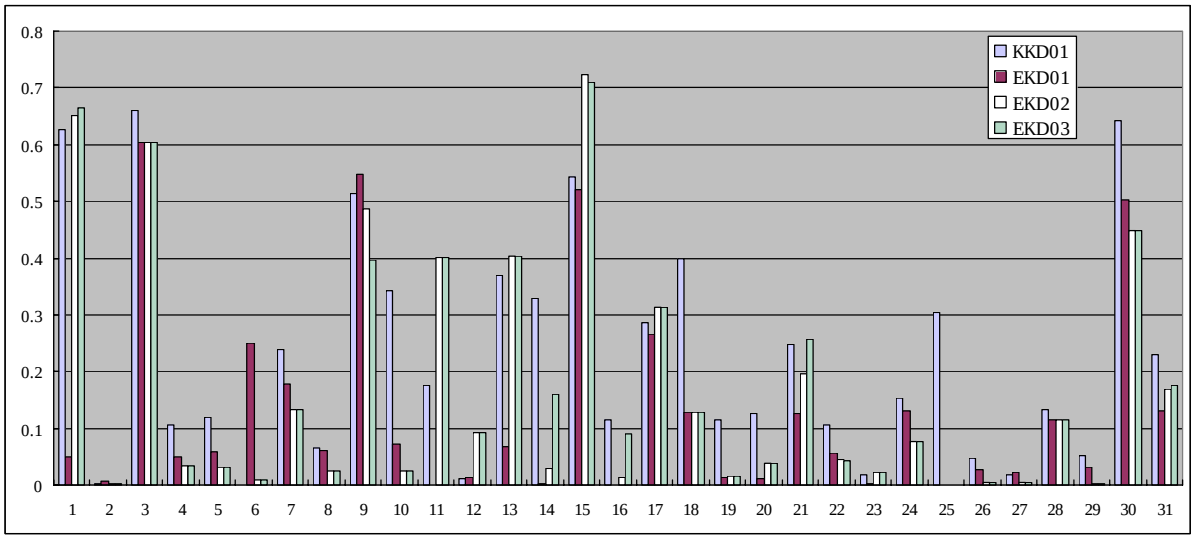

Figure 3. Average precision per query (Rigid Assessment)

words contributed to retrieving documents relevant to the topic 006.

3.2 Impact of Query Expansion

The different precision between EKD01 and EKD02 represented in Table **??** shows that the query expansion is useful for the English-Korean cross-language retrieval. That is to say, the words that are not in the original query but are added to the original query contribute to the system performance. For example, the expansion term "*Nelson*" is closely related with "*Mandela*" in the NTCIR topic 001, and helps retrieving documents relevant to the topic. Moreover, while "*Mandela*" is not in the bilingual dictionary, "*Nelson*" is in and hence plays an important role in retrieving relevant documents.

However, the expansion term which is not relevant to a user need may have a negative effect on the retrieval system. For NTCIR topic 006 about "*Liner Caught Fire in South Korea*", for example, the expansion term "*Caribbean*" dropped the system performance.

3.3 Impact of Transliteration

The different precision between EKD03 and EKD02 represented in Table 3 shows that transliteration in the English-Korean cross-language retrieval contributes to system performance. The proper nouns usually play an important role in IR. For example, NTCIR topic 001 is about "*Mandela, the president of South Africa*" and the most important term in the topic is the proper noun "*Mandela*". So, the CLIR system which cannot transliterate "*Mandela*" in the topic cannot be expected to perform well.

3.4 Impact of English query

We send the original query in English as well as the translated query in Korean to an information retrieval system because some Korean documents sometimes include English acronyms and English initial words, such as "*APEC*", "*WTO*" and "*ANC*". For example, "*ANC*[1]" in the NTCIR topic 001 and "*APEC*" in the NTCIR topic 030 affected a retrieval performance. At this point, using English query with Korean query seems to improve the system performance. But some Korean documents with English sentences and English query terms in query are ranked relatively high. For the NTCIR topic 001, for instance, the document that has many English sentences but is not relevant to the topic is ranked high for EKD01 run.

4 Post-Submission Experiments

We made some additional experiments after NTCIR-3 workshop. In this section, the post-submission experiments are described. A description field of each topic is used as a query in the experiments.

4.1 IR system based on NE2000

Our IR system had utilized HanTag to extract Korean noun words [2] at NTCIR-3 workshop. However, tagging all words by HanTag causes the IR system to be slow. On the other hand, extracting only noun words by Korean noun extractor(NE2000) leads to being faster IR system than using HanTag. According to

[1] term "*ANC* was included by query expansion.

[2] We consider only noun words among Korean words as index terms.

[4], the extraction speed of NE2000 is almost 17 times faster than that of HanTag.

Therefore, we evaluated two IR systems, one of which is based on the HanTag and the other of which is based on NE2000. Table 4 shows that the performance of two systems is similar in monolingual information retrieval.

Table 4. Performance of NE2000 and HanTag in monolingual IR

	relax		rigid	
	avg.P.	R-prec.	avg.P.	R-prec.
NE2000	0.2634	0.3139	0.2415	0.2656
HanTag	0.2693	0.3234	0.2292	0.2693

NE2000 is adopted to extract Korean noun words in the following experiments because IR system based on NE2000 is faster than that of HanTag and the performance of two systems is similar.

4.2 Number of Translation Equivalents

We have selected two translation equivalents for each English query term, where korean words with a similar meaning are included in Korean query. But, there is the problem that korean words irrelevant to a query might be selected during the query translation.

Hence, the performance of system need to be evaluated according to the number of translation equivalents. Table 5 shows the results.

Table 5. System Performance according to Number of Translation Equivalents

	relax		rigid	
	avg.P.	R-prec.	avg.P.	R-prec.
best one	**0.1717**	**0.2166**	**0.1464**	**0.1743**
best two	0.1350	0.1740	0.1221	0.1576
all	0.0703	0.1151	0.0751	0.1028

In Table 5, ”best one” denotes selecting only one translation equivalent per English query term, ”best two” two translation equivalents, and ”all” all possible translation equivalents. We can observe that choosing the best one translation equivalent for each English query term is desirable.

4.3 English Words in Korean Document

Some Korean documents include English words, and hence English words may be effective on searching relevant Korean documents. So, we tried to examine whether english words are helpful to retrieve Korean documents.

The effect of English words on English-Korean CLIR is provided in Table 6. In Table 6, KQ means Korean query consisted of translation equivalents, EQ English query, and EW English acronyms and English initial words. While English query with Korean query(KQ+EQ) drops performance, English acronyms and initial words with Korean query(KQ+EW) contributes to performance improvement.

Table 6. Effect of English Words

	relax		rigid	
	avg.P.	R-prec.	avg.P.	R-prec.
KQ	0.1717	0.2166	0.1464	0.1743
KQ+EQ	0.1516	0.1922	0.1304	0.1581
KQ+EW	**0.1807**	**0.2221**	**0.1564**	**0.1743**

4.4 Query Expansion

The overall results of query expansion are shown in Table 7. English query is expanded by adding W top ranked words from R top ranked documents at Financial Times Corpus in TREC data. Table 7 presents that query expansion always improves the performance of the system and the best performance is achieved at R=10 and W=40.

Table 7. Effect of Query Expansion

	relax		rigid	
	avg.P.	R-prec.	avg.P.	R-prec.
W=0	0.1807	0.2221	0.1564	0.1743
R=10,W=10	0.1950	0.2343	0.1700	0.1953
R=10,W=40	**0.2126**	**0.2554**	**0.2005**	**0.2134**
R=20,W=10	0.1962	0.2514	0.1712	0.1964
R=20,W=40	0.1904	0.2311	0.1687	0.1921

5 Conclusions

We have developed a English-Korean CLIR system for the NTCIR 3 workshop, based on the query translation approach using two bilingual dictionaries. When translating an English query into a Korean query, a bilingual biographical dictionary is used to handle proper names. Furthermore, Korean place names and Korean company names are also added to the general dictionary.

When we compare the results of three cross-language runs, we realize that query expansion and

transliteration in the English-Korean CLIR can improve the system performance.

And additional experiments after NTCIR-3 workshop show that choosing the best translation equivalent for each query term and utilizing English acronyms and initial words with Korean query improve the system performance.

References

[1] L. Ballesteros and W. B. Croft. Comparing representations in chinese information retrieval. In *Proceedings of the 20th ACM SIGIR International Conference on Research and Development in Information Retrieval*, 1997.

[2] J. Gao, J.-Y. Nie, J. Zhang, E. Xun, Y. Su, M. Zhou, and C. Huang. Trec-9 clir experiments at msrcn. In *Proceedings of the Nineth Text REtrieval Conference(TREC-9), NIST special publication, 500-249*, 2000.

[3] J. D. Kim, H. S. Rim, and H. C. Rim. Twoply HMM: A part-of-speech tagging model based on morpheme-unit considering the characteristics of korean. *Journal of Korean Information Science Society*, 24(12(B)):1502–1512, 1987.

[4] D.-G. Lee, S.-Z. Lee, and H.-C. Rim. An efficient method for korean noun extraction using noun occurrence characteristics. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, 2001.

[5] S. E. Robertson and S. Walker. Okapi/keenbow at TREC-8. In *Proceedings of The Eighth Text REtrieval Conference(TREC-8), NIST special publication, 500-246*, pages 151–162, Gaithersburg, November 1999.

[6] J. Xu and R. Weischedel. Trec-9 cross-lingual retrieval at bbn. In *Proceedings of the Nineth Text REtrieval Conference(TREC-9), NIST special publication, 500-249*, 2000.