Knowledge-light Asian Language Text Retrieval at the NTCIR-3 Workshop

Paul MCNAMEE
Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, Maryland 20723-6099, USA
mcnamee@jhuapl.edu

Abstract

To combat the inherent complexity of text retrieval in a large number of disparate languages, scalable techniques must be developed and refined. We have been studying how well language-neutral approaches to text processing and retrieval can perform. With that goal, we participated in the third NTCIR workshop and conducted experiments using knowledge-light approaches, ones that did not attempt to segment words or normalize word forms directly, but rather addressed these issues in other ways. Chiefly, we investigated term selection using a combination of character n-grams of various lengths.

We found that representing text using a combination of character n-grams of lengths one, two, and three is effective for monolingual retrieval of Chinese, Japanese, and Korean texts. We also performed bilingual retrieval using pre-translation expansion, machine translation, and n-gram indexing. We discovered an unforeseen difficulty in term weighting when attempting word-for-word query translation with multiple length n-gram indexing that limits the bilingual applications that would benefit from this tokenization technique.

Keywords: Text processing, Asian languages, multilingual information access, cross-language information retrieval, character n-grams

1 Introduction

Text retrieval research at the Johns Hopkins University Applied Physics Laboratory (JHU/APL) has emphasized retrieval approaches that are suitable for many languages. We have adopted this philosophy because customized techniques that benefit retrieval in one or two languages, even significantly, cannot typically be applied to other human languages without significant adaptation and software recoding. Governments and multinational enterprises create large repositories of electronic content in tens of languages. We contend that the amount of linguistic knowledge (and accompanying software complexity) utilized should be minimized to facilitate retrieval over such archives with a single software system. We believe such an approach has the greatest likelihood of scaling to multiple, disparate languages. However, it is clearly desirable to obtain the greatest possible retrieval performance using such techniques. It remains to be seen whether a performance gap exists between knowledge-based and knowledge-light systems and whether any differences in precision and efficiency are significant.

With these principles in mind we have developed the Hopkins Automated Information Retriever for Combing Unstructured Text, or HAIRCUT, a research retrieval engine. In agreement with this language-neutral posture the system has been developed in the Java programming language. Java's support for the Unicode standard enables queries and documents to be processed in arbitrary encodings without transforming source files to a particular encoding; the in-memory processing takes place using the Unicode character set. More significantly, in addition to internationalized software development, we rely on simplified methods for representing text and for combating problems like word segmentation, difficulties in morphological normalization, and phrase identification. In particular, over several years we have investigated the use of overlapping character n-grams for indexing and retrieval. Though long popular for use in Chinese and Japanese text processing, n-grams have been used far less often in European languages. We have found n-grams to be tremendously effective in Asian, European, and Semitic languages. We used them exclusively for our experiments with Asian languages at the NTCIR-3 evaluation.

Of the 22 groups that participated in the CLIR Task, only 4 groups submitted results in all four document languages: Chinese, English, Japanese, and Korean. This may be due to difficulties in employing resources or techniques that are specific to one or two languages. Alternatively it may simply reflect the different research interests of workshop participants.

This year APL submitted results in single-language retrieval (SLIR), bilingual retrieval (BLIR), and multilingual retrieval (MLIR) tasks. We had only few translation resources available and could only seriously consider translation where English was one of the language pairs; for the BLIR and MLIR Tasks, we submitted runs using English as the source language and translation from on-line Web services.

None of the members of our team had any ability to understand Chinese, Japanese, or Korean, thus our approach was truly based on minimal knowledge. The only prior exposure to Japanese text retrieval that we had was our participation in the NTCIR-2 workshop [7]. We had investigated Chinese retrieval at the TREC workshops [5], but until now have never worked with Korean text.

2 Background

In Asian languages it is important to consider the issues that arise from the nature of each language. For example, words are not typically delimited by a space, multiple alphabets can be used (e.g., in Japanese), and methods for translating phrases borrowed from other languages differ. These issues underlie the very basic problem of determining what elemental units should be used to represent text and how they should be identified. Approaches based on character n-grams appear most often in the literature; these are popular because they can be used without attempting word segmentation. Recently several studies have appeared that suggest alternative term representations may outperform a single length of character n-gram.

Ogawa and Matsuda have studied a variety of n-gram methods for indexing Japanese. Using the BMIR-J1 and BMIR-J2 collections they observed that bi-grams were the most effective; however, they experimented with several modifications that slightly improved performance. In one experiment they found that a combination of using 1-grams, 2-grams, and 3-grams as indexing terms was more effective than a single choice of bigrams [9]. In later work they found that ignoring certain n-grams (those containing hirigana) was beneficial [10].

At the NTCIR-1 workshop [3] several groups examined the role of segmentation and the merits of different approaches to tokenization. For example, Ozawa et al. found that an adaptive method of segmentation that produces n-grams of various lengths outperforms simple bigrams [11]. Their hypothesis was that bigrams are insufficient in technical language where word length increases. At the NTCIR-2 workshop McNamee found that when Japanese documents were indexed twice, once using 2-grams and once with 3-grams, better document rankings could be obtained by merging scores computed from the two separate indexes [7].

In a recent unpublished experiment, we have confirmed that a combination of 1-, 2-, and 3-grams conferred about a 10% advantage over the use of 2-grams alone using the TREC-9 Chinese data. It is possible that this effect is most significant for Chinese, where proper names tend to be three or four characters in length. We were sufficiently motivated by this work to decide upon a combination of 1-, 2-, and 3-grams for our submissions to the CLIR Task. Of course, it may certainly be that no single term representation works best across these three languages.

3 Overview

We submitted results for all three evaluations in the CLIR Task (SLIR, BLIR, and MLIR). Although we have previously shown that larger length n-grams are effective for indexing and retrieval in English *(e.g., n=6),* here we simply used unnormalized words [6]. At first glance, this appears a contradiction to our expressed philosophy of uniformly applying simplified methods (such as n-grams) in each language; however, this is not the case. Because we wanted to participate in the BLIR and MLIR tasks, we needed suitable translation resources. Due to monetary constraints and limits on available software development time, we were not able to develop quality translation resources for use at the workshop. With some reluctance we decided to rely solely on automated machine translation software to perform query translation from English into Chinese, Japanese, and Korean. We have recently confirmed Ballesteros and Croft's conclusion [1] that source-language pre-translation expansion is remarkably effective in improving bilingual retrieval performance and also found this when using only poor translation resources [4]. Thus we decided to index the English sub-collection using unstemmed words and to use this collection for query expansion prior to translation using MT software. The resulting translated terms were then converted into character n-grams and used for retrieval.

For monolingual retrieval we used words for English, and character n-grams in the other languages. During NTCIR-2 we evaluated the relative merits of 2-grams and 3-grams for Japanese retrieval. Since the dictionary size grows significantly for longer lengths of n, especially for a language with many symbols such as Japanese, few groups have seriously explored the use of 3-grams for a large collection; however, some believe that longer lengths may be important for domain-specific (e.g., technical) terminology. In our work at NTCIR-2 we discovered that 2-grams and 3-grams perform equivalently when used as indexing terms. This was somewhat surprising since experiments comparing the two in Chinese showed that 2-grams were best.

We intend to compare this approach to other types of n-grams, but we have not yet had the opportunity to complete this analysis.

3.1 Index Construction

The document collections for each language differed widely in size; the largest sub-collection included over half a million documents in Chinese, but only about 23,000 English articles were available. The encodings also varied. The respective Extended UNIX Codes were used for the Japanese and Korean collections, while Chinese documents were in Big-5.

For our official runs we created an index for each language using different term representations as described above. Information about each index can be seen in Table 1. In the Asian language collections no attempt was made to perform word segmentation. We did attempt to identify various forms of punctuation and sentence boundaries. In English words were lowercased, space-delimited tokens. In Japanese and Korean n-grams were formed of lengths 1, 2, and 3, over character sequences; however, they did not span sentence boundaries or punctuation. In Chinese, only n-grams of lengths 1 and 2 were used for performance reasons. All of the n-grams produced for a passage of text were added to the representation of each document. As can be seen from Table 1, the inverted files contained many postings lists.

	Docs	Type	Distinct Terms	Index (MB)
English	22,926	words	94,162	32
Chinese	513,854	1-2-grams	3,483,677	1391
Japanese	236,664	1-2-3-grams	9,223,696	1199
Korean	66,146	1-2-3-grams	1,538,748	232

Table 1. Statistics for the indexes.

3.2 Query Processing

We removed stop structure from queries using a list of about 1000 phrases such as "… would be relevant" or "relevant documents should…." These were mined from topic statements prepared for the TREC evaluations. For monolingual retrieval in Chinese, Japanese, and Korean, we used a translated version of this list. After this step, information requests were tokenized and handled in the same fashion as documents. Specifically, characters were lower-cased (if appropriate), punctuation was normalized to Latin-1 equivalents (e.g., full stops for each language were mapped to ASCII 46 (decimal)), and words or n-grams were used as indexing terms.

In our experiments a statistical language model was used to compute similarity scores. These models have recently received significant attention in the literature; we believe that they outperform traditional vector cosine-based measures, though no theoretical justification for this claim is available. Despite the fact that these models were developed with words or normalized word forms in mind, we have used them without adaptation with n-grams even though terms dependencies are probably greater when n-grams are used. Various papers describing the language modeling approach have appeared recently in the literature (*e.g.,* [2], [8], and [11]). Each query term was weighted by the query term frequency; the calculation performed was:

$$Sim(q,d) = \prod_{t=terms} \left(\alpha \cdot f(t,d) + (1-\alpha) \cdot mrdf(t)\right)^{f(t,q)}$$

Equation 1. Language model similarity score

where *f(t,d)* is the relative frequency of term *t* in document *d* and *mrdf(t)* denotes the mean relative document frequency of *t* (*i.e.,* the term document frequency for *t*, averaged over all documents). α is the probability that a term is generated by a model based on a single document instead of a model based on the language in general. We used *α=0.30* for all our official experiments.

For our monolingual runs and for pre-translation expansion (in English) of our bilingual runs, blind relevance feedback was performed. The top 20 documents were used and 60 terms for the expanded query are selected based on three factors, a term's initial query term frequency, the (α=3, β=2, γ=2) Rocchio score, and a metric that incorporates an IDF component. The top-scoring terms are then used as the revised query. The same parameters were used for feedback regardless of language.

4 Monolingual Experiments

For our monolingual experiments, we submitted three runs in each language using different parts of the topic statements. This was motivated by a desire to examine the effect of topic-length and to insure that our work was comparable with that of other track participants. The relative performance of our official runs compared to other official CLIR Track submissions is shown in Table 2.

	Topic Length	#runs	MAP	
			Median	APL
English	T	1	0.3087	0.3087
	D	13	0.2641	0.3551
	TDNC	9	0.4460	0.4460
Chinese	T	1	0.1928	-
	D	14	0.1908	0.2752
	TDNC	8	0.2653	0.3224
Japanese	T	2	0.2627	0.2877
	D	18	0.2550	0.2847
	TDNC	4	0.3429	0.3437
Korean	T	1	0.2463	0.2463
	D	9	0.1957	0.1907
	TDNC	4	0.3534	0.3113

Table 2. Summary of the performance of APL's official monolingual runs using the rigid relevance criteria[1]. Comparisons to median and top score are based on the set of official runs using the same language and topic fields.

Examining Table 2 we note that very few official runs were submitted using only the title portion of queries. Comparing with mean average precision we observe that our official monolingual results were at

[1] In this paper our analysis was based only on the relevance judgments produced using rigid criteria.

or above the median performance in Chinese, English, and Japanese retrieval. This result suggests that our choice of multiple-length n-grams as indexing terms was reasonable. However, our results in Korean retrieval were a bit worse, so it may be that a different representation is required. It would be premature to make this conclusion without further failure analysis for the Korean collection.

Figures 1 through 4 contain Precision-Recall graphs for the monolingual runs. Across the four languages, the best results are obtained when the longest topics are used (TDNC). Note the use of title-only queries appears better than description-only queries in Japanese and Korean; this might be due to how the topic statements were created.

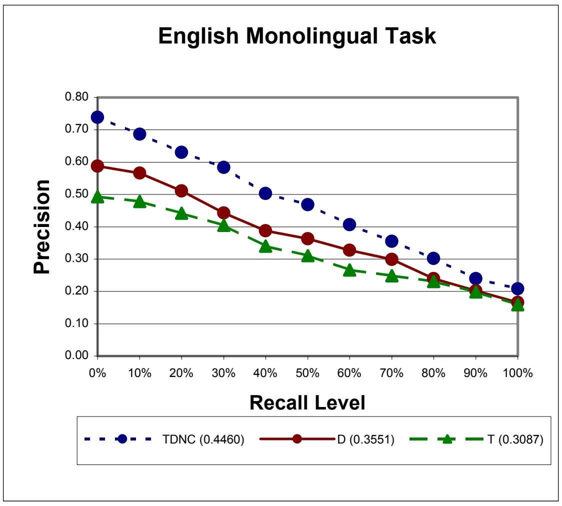

Figure 1. Comparison of English monolingual retrieval performance using different topic fields.

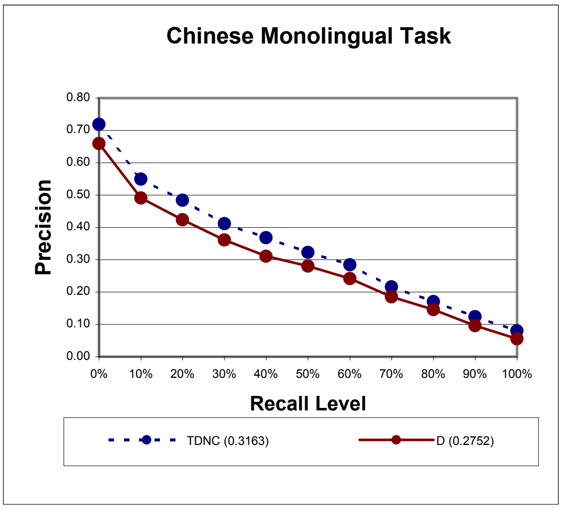

Figure 2. Comparison of Chinese monolingual retrieval performance using different topic fields.

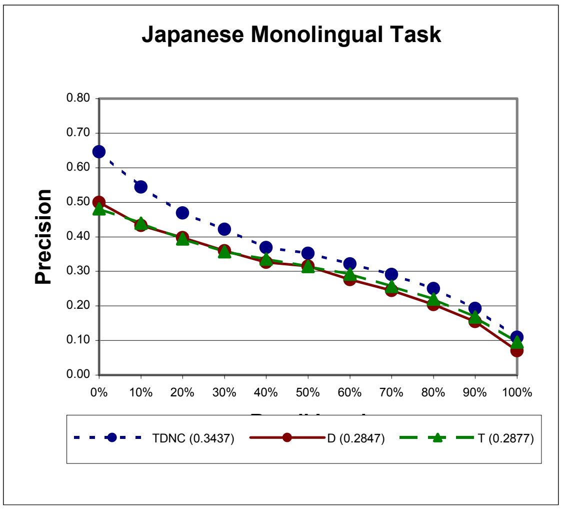

Figure 3. Comparison of Japanese monolingual retrieval performance using different topic fields.

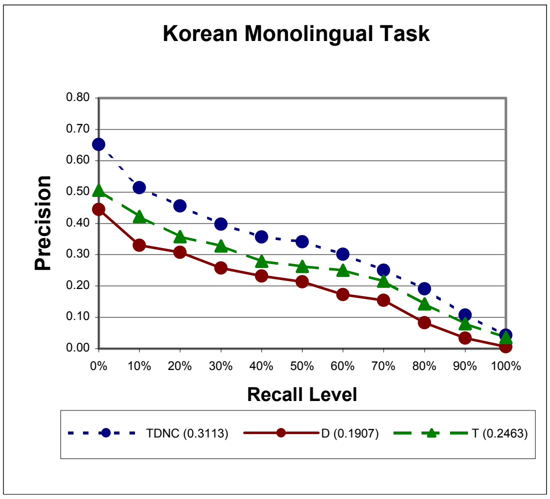

Figure 4. Comparison of Korean monolingual retrieval performance using different topic fields. Like the Japanese monolingual results, the title-only runs outperform the description-only ones; however, here the difference is pronounced.

5 Bilingual Results

We also submitted bilingual and multilingual results for the evaluation, but only English was used as a source language. We used pre-translation expansion using the CLIR Task English sub-collection as an expansion corpus. A set of 60 English words was then translated using the on-line Babelfish translator to process the queries; n-grams of multiple lengths were then produced from the translated terms (lengths 1-3 in Japanese and Korean and lengths 1-2 in Chinese). No post-translation query expansion was performed. We report the median mean average precision for each language/topic-field pair in Table 3 along with APL's official run for that condition.

	Topic Length	#runs	MAP	
			Median	APL
Chinese	T	1	0.0079	0.0079
	D	6	0.1086	0.0070
	TDNC	6	0.1803	0.0231
Japanese	T	1	0.0802	0.0802
	D	7	0.1899	0.0878
	TDNC	1	0.1338	0.1338
Korean	T	1	0.0484	0.0484
	D	4	0.1499	0.0234
	TDNC	1	0.0476	0.0476

Table 3. Official bilingual performance

These results seem rather bad; our bilingual runs exhibit much lower performance, relative to the median, than our monolingual runs. This could be due to the translation resource that we use, could be a result of our methods for pre-translation expansion and tokenization, or could be caused by a systematic error. These results were so low, we considered whether the poor performance was attributable to mistakes in character decoding, but the correct character sets for the documents and translated queries were used. Our complicated method for pre-translation expansion is where we first looked.

Since the English document collection is rather small, it is possible that an insufficient number of on-topic documents are available to capture the query semantics and produced a good set of terms over which to attempt translation. However, upon examining the queries, we do not think this is the case. The expanded terms produced for topic 001, "The Exhibition 'Art and Culture of the Han Dynasty'", are shown below in Figure 5. They seem very reasonable, and, furthermore, there were no relevant documents in the English collection for this topic.

dynasty	0.1054	osaka	0.0111	famous	0.0090
exhibition	0.0881	closed	0.0110	paris	0.0090
art	0.0840	mondays	0.0110	produced	0.0089
han	0.0786	th	0.0109	palace	0.0089
culture	0.0760	history	0.0103	1000	0.0089
museum	0.0230	ancient	0.0102	keihan	0.0088
paintings	0.0165	pieces	0.0102	5	0.0088
display	0.0156	chinese	0.0101	hours	0.0087
works	0.0141	artifacts	0.0101	nara	0.0086
exhibitions	0.0124	75	0.0100	100	0.0085
kyoto	0.0121	masterpieces	0.0097	painting	0.0085
admission	0.0120	temple	0.0096	dolls	0.0084
schedule	0.0116	m	0.0096	cho	0.0084
tofu	0.0114	7	0.0095	store	0.0084
festival	0.0114	treasures	0.0094	adults	0.0084
collection	0.0114	station	0.0094	titled	0.0084
9	0.0113	painters	0.0092	f	0.0083
call	0.0112	featured	0.0091	series	0.0082
information	0.0112	sogo	0.0090	period	0.0082
century	0.0112	hall	0.0090	buddhist	0.0082

Figure 5. Expanded query terms with associated query term weights for topic 001. Terms were identified prior to translation using documents retrieved from the source language (English) collection.

We also wondered whether the low performance is due our use of n-grams, even over untranslated terms. Numerous English words were not translatable, and these words were then processed as ordinary text; thus n-grams of lengths 1, 2, and 3 were produced over these (English) words. These n-grams retained the unbalanced weight of the original source-language query term. In hindsight, this does not seem like a promising approach, but our goal was to translate individual query words with MT software and retrieve documents that were indexed using n-grams. To ascertain whether this was the case, we selectively removed n-grams containing only Roman letters from queries; little difference was observed, so we do not belief this is the principal cause.

After the workshop, we examined the use of query-based machine translation, without using pre-translation expansion, both using character bi-grams as indexing terms and using multiple length n-grams. This enables comparison between two tokenization methods and also our method for pre-translation expansion. These results are reported in Table 4. Here we find a significant improvement when pre-translation is not used with the multiple n-gram lengths; since we do not know how multiple n-grams produced from a single translation should be weighted, we weighted each n-gram the same as the original query term from which it was derived – we belief that this difficulty in weighting the numerous n-grams. When no expansion is used, this is not a problem, as n-grams are just produced over the free-running text of the translated query. We also see that multiple n-gram lengths performed about the same as just 2-grams alone, though some small improvements were observed.

The revised performance numbers are still below the median for the bilingual task; we conjecture that a sizeable part of this is due to the particular translation source we based our experiments on.

	Topic Length	Mean Average Precision		
		Pre-trans. expansion	No expansion	
		mult. lengths	2-grams	mult. lengths
Chinese	T	0.0079	0.0371	0.0444
	D	0.0070	0.0333	0.0413
	TDNC	0.0231	0.0667	0.0739
Japanese	T	0.0802	0.0842	0.0768
	D	0.0878	0.0876	0.0787
	TDNC	0.1338	0.1807	0.1817
Korean	T	0.0484	0.1135	0.1068
	D	0.0234	0.0751	0.0803
	TDNC	0.0476	0.1326	0.1479

Table 4. Comparison of multiple methods for bilingual retrieval using machine translation.

6 Conclusions

Our monolingual runs performed well with respect to median results for the mandated description-only runs submitted by other track participants; thus we feel that a combination of n-grams can indeed be successful for retrieval in disparate Asian languages, something that has been suggested by several previous studies.

Our official bilingual runs exhibited very low performance, which we were able to investigate retrospectively. We found that indexing using multiple lengths of character n-grams does work as well as indexing of character bi-grams in Chinese, Japanese, and Korean when queries are free-running text. However, it appears that when the technique used in our official runs, word-by-word translation, is employed, it is less effective to use multiple n-gram lengths because appropriate term weights are difficult to determine. At present this would appear to restrict the application of multiple n-gram indexing to monolingual retrieval, or bilingual retrieval when fully translated queries are available.

7 Acknowledgements

We are grateful to the copyright holders who graciously provided the data used for the evaluation and to the workshop organizers and assessors. Without the careful work of the later, these experiments would not have been possible.

References

[1] L. Ballesteros and W. B. Croft, 'Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval.' *In the Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-97*), pp. 84-91, 1997.

[2] D. Hiemstra and A. de Vries, 'Relating the new language models of information retrieval to the traditional retrieval models.' CTIT Technical Report TR-CTIT-00-09, May 2000.

[3] N. Kando, K. Kuriyama, and T. Nozue, 'NACSIS Test Collection Workshop (NTCIR-1)'. In the *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR-99)*, August 1999.

[4] Paul McNamee and James Mayfield, Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In the *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR-2002)*, Tampere, Finland, pp. 159-166, August 2002.

[5] Paul McNamee, James Mayfield, and Christine Piatko, "The HAIRCUT System at TREC-9". In E. Voorhees and D. Harman (eds.), *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, NIST Special Publication 500-249, Gaithersburg, Maryland, pp. 273-279, October 2001.

[6] Paul McNamee, James Mayfield, and Christine Piatko, "A Language-Independent Approach to European Text Retrieval". In Carol Peters (ed.) *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF-2000 Workshop*, Lecture Notes in Computer Science 2069, Springer, Lisbon, Portugal, pp. 129-139, 2001.

[7] Paul McNamee, "Experiments in the Retrieval of Unsegmented Japanese Text at the NTCIR-2 Workshop". In the *Proceedings of the Second NTCIR Workshop*, Tokyo, Japan, pp. 5:157-162, March 2001.

[8] D. R. H. Miller, T. Leek, and R. M. Schwartz, 'A Hidden Markov Model Information Retrieval System.' In the *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR-99)*, pp. 214-221, August 1999.

[9] Y. Ogawa and T. Matsuda, 'Overlapping statistical word indexing: A new indexing method for Japanese text'. In the *Proceedings of the 20th International Conference on Research and Development in Information Retrieval (SIGIR-97*), pp. 226-234, July 1997.

[10] Y. Ogawa and T. Matsuda, 'Overlapping statistical segmentation for effective indexing of Japanese text'. In *Information Processing & Management*, 35(1), pp. 463-480, 1999.

[11] T. Ozawa, M. Yamamoto, K. Umemura, and K. W. Church, 'Japanese word segmentation using similarity measure for IR'. At the *First NTCIR Workshop on Research in Text Retrieval and Term Recognition (NTCIR-1)*, 1999.

[12] J. Ponte and W. B. Croft, 'A Language Modeling Approach to Information Retrieval.' In the *Proceedings of the 21st International Conference on Research and Development in Information Retrieval (SIGIR-98),* pp. 275-281, August 1998.