ISSN 1346-5597

NII National Institute of Informatics

NII Technical Report

Handling Orthographic Varieties in Japanese Information Retrieval: Fusion of Word-, N-gram-, and Yomi-Based Indices across Different Document Collections

Nina Kummer, Christa Womser-Hacker, Noriko Kando

NII-2005-011E
July 2005

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

Handling Orthographic Varieties in Japanese Information Retrieval: Fusion of Word-, N-gram-, and Yomi-Based Indices across Different Document Collections

Nina Kummer[1,2], Christa Womser-Hacker[1], Noriko Kando[2]

[1] Universität Hildesheim, Germany
[2] National Institute of Informatics, Tokyo, Japan

nina@nii.ac.jp, womser@uni-hildesheim.de, kando@nii.ac.jp

Abstract. Orthographic varieties are common in the Japanese language, and represent a serious problem for Japanese information retrieval (IR), as IR systems run the risk of missing documents that contain variant forms of the search term. We propose two different strategies for handling orthographic varieties: pronunciation or yomi-based indexing and "Fuzzy Querying", comparing katakana terms based on edit distance. Both strategies were integrated into our multiple index and fusion system, and tested using two different test collections, newspaper articles (Mainichi Shimbun '98) and scientific abstracts (NTCIR-1), to compare their performance across text genres. The fusion of the results obtained with a bi-gram-based, a word-based, and the additional yomi-based index was found to improve precision significantly for the NTCIR-1 collection, but only slightly for the Mainichi Shimbun '98 collection. Adding Fuzzy Querying as a fourth system and merging the results led to a further, but not significant, improvement in precision.

1 Introduction

This work is part of a project that aims to integrate Japanese language support into the Multiple Indexing for Dynamic Method-Object-Relation in Information Retrieval (MIMOR) framework [1, 2]. MIMOR adopts a multiple indexing and fusion approach to take advantage of the best-performing technologies to achieve optimum retrieval results. To address the challenges presented by automatic handling of orthographic varieties, we experimented with the fusion of a yomi- or pronunciation-based index in addition to word- and bi-gram-based indices. Further, we applied "Fuzzy Querying" for katakana terms as a strategy for coping with katakana variants. Fuzzy Querying matches terms similar to a specified term. The similarity between terms in the index and a specified target term is determined using the Levenshtein distance algorithm, normalized by the term length.

Building on previous experiments that evaluated the effectiveness of a yomi- or pronunciation-based index for Japanese IR within a multiple index and fusion system [3], we carried out comparative experiments using different document genres in order to determine the performance of our approach across text domains. The test corpora used were newspaper articles and scientific abstracts. Because a yomi-based index should help to compensate for orthographic variety within a text corpus, we expected the tested methods to be more effective using the collection of scientific abstracts, which is far more heterogeneous than the standardized newspaper corpus. The basic retrieval engine of our system is Lucene (http://lucene.apache.org).

2 Background

2.1 Orthographic Variety in the Japanese Language

A serious problem in Japanese IR, although rarely discussed outside the major difficulty of correct word segmentation and the resulting challenges for indexing, is the high degree of orthographic variety in the Japanese language. Owing to the combined usage of four different scripts within one writing system (kanji, hiragana, katakana, and Roman characters), many words can be written in a variety of ways. Traditional IR systems, which compare terms according to their written representation, run the risk of missing documents that contain variant forms of search terms. In this paper, we will provide an overview of the most frequent types of orthographic variation and discuss their impact on information retrieval. A comprehensive overview of the types of orthographic variety can be found in Halpern [4].

Cross-script variants. Although each of the four scripts used in Japanese has its own, well-defined function, cross-script variation is frequent, and often unpredictable. The same word may be written in hiragana, katakana, rōmaji or kanji, or even in a mixture of two scripts. Sometimes, an unusual script is chosen for certain words for stylistic reasons, to catch the reader's attention or to add an emotional component. Table 1 shows the most frequent cross-script variation patterns.

Table 1. Examples of cross-script variants (*see* Refs [4] and [5]).

Type	Variants	English
Kanji *vs*. hiragana	大勢　おおぜい	many; crowd; large number of people
Kanji *vs*. katakana	硫黄　イオウ	sulfur
Kanji *vs*. hiragana *vs*. katakana	猫　ねこ　ネコ	cat
Katakana *vs*. rōmaji	キログラム　kg	kg
Katakana *vs*. hybrid	ワイシャツ　Y シャツ	shirt (trans. = white shirt); business shirt
Kanji *vs*. katakana *vs*. hybrid	皮膚　ヒフ　皮フ	skin
Kanji *vs*. hybrid	彗星　すい星	comet
Hiragana *vs*. katakana	ぴかぴか　ピカピカ	glitter; sparkle

Okurigana variants. Okurigana are the hiragana used for grammatical endings. Since the Japanese adopted Chinese characters to frame their own written language, it is not always clear how much of a Chinese character is considered to represent the word stem, and what is considered to be an ending to be written in hiragana. In many cases, the actual ending and also a part of the stem is written in kana. Variants occur in the number of syllables expressed in hiragana. An extreme example is the verb "to express". Besides its standard form, 書き表す, it can also be written as 書き表わす, 書表わす, or 書表す.

Hiragana variants. Although hiragana orthography is generally regular, there are some irregularities, mainly owing to evolution in the orthographic rules over time.

Kanji variants. Although the Japanese writing system underwent major reforms in 1946 and 1981, and the character forms have now been standardized, there are still a significant number of variants in common use. Frequently used and complex Chinese characters have been simplified over time, sometimes in several ways. Their traditional forms also continue to exist, especially in proper nouns and classical works.

Phonetic substitutes. There are a large number of orthographic variants in Japanese, based on the principle of "phonetic substitution". In some cases, two characters are interchangeable in certain compounds. The original character and its phonetic replacement share the same reading, and often are similar in meaning.

Katakana variants. In recent years, there has been a sharp increase in the use of katakana, the script employed for writing loanwords, especially in technical terminology. As a modern and living language, Japanese is constantly evolving, producing new expressions for new concepts. These are very often adaptations of foreign, mostly English, terms. Unfortunately, the resulting katakana orthography is often irregular. Since katakana syllabary is used to transcribe the phonetic structure of foreign words, the orthography often depends on the interpretation of the correct pronunciation, which may vary from person to person [6].

In transcribing European words into Japanese, the nearest Japanese sound is chosen as a substitute for any sound not available in Japanese. Compared to European languages, the Japanese sound system is simpler, with only a small number of different sounds; thus, the Japanese rendition of foreign words is often very approximate [6]. Table 2 lists some examples of katakana variations.

Table 2. Katakana variants (*see* Ref [5]).

English	Reading	Standard	Variants
Computer	Konpyuuta or konpyuutaa	コンピュータ	コンピューター
Online	Onrain	オンライン	オン・ライン
Eye shadow	Aishadoo	アイシャドー	アイシャドウ
Maid	Meedo	メード	メイド
Diesel	Diizeru or jiizeru	ディゼル	ジーゼル チーゼル
Jerusalem	Erusaremu	エルサレム	イェルサレム

One possible solution for the automatic handling of orthographic variety would be a comprehensive dictionary of all variant forms along with an algorithm that performs a simple table-lookup and normalization of all variant forms to a base form. This solution, a lexicon-based disambiguation, is also suggested by Halpern [4, 5]. However, such a dictionary is costly to compile, and requires constant maintenance, as the language is evolving quickly. Therefore, we argue that a more flexible strategy for automatic handling of orthographic varieties is needed.

From the information retrieval point of view, we can classify orthographic varieties in Japanese into two groups:

1. Variants originating from a different written representation of the same phoneme (cross-script variants, okurigana variants, hiragana variants, kanji variants, and phonetic substitutes).
2. Variants originating from a different interpretation of the sound structure to be represented (katakana variants).

Variants in the first group share the same pronunciation. This fact can be exploited for information retrieval, if the terms are matched using their pronunciation instead of their written representation. Variants of the second group need a different treatment. When foreign words are transcribed into katakana, the nearest Japanese sound substitutes for any sound not available in Japanese. There are only a limited number of ambiguous cases, where there is more than one transcription of a foreign sound (*e.g.*, キ /ki/ and ク /ku/ for the rendering of the English sound /ik/ as in "cake"). Consequently, katakana variants still share most syllables, and only differ in minor aspects (*i.e.*, one or two characters). We supposed that Fuzzy Querying, a technique that matches terms based on their editing distance, may be an effective means of retrieving documents that contain katakana variants of a search term.

2.2 Yomi-Based Indexing

Yomi-, or pronunciation-based, indexing is not a new strategy for use in Japanese IR. In contrast, it is a rather old technique, which used to be employed before the introduction of double-byte processing on computers, when information processing systems used the katakana syllabary to represent Japanese text phonetically. The yomi-based index has been abandoned since the introduction of double-byte character handling, as the Japanese language is very rich in homophones, which are kept apart through the use of the ideographic kanji characters in written language. A phonetic transcription of Japanese lacks this information, and can therefore be very ambiguous at times. Whereas human readers may be able to infer the meaning of any ambiguous words from the textual context, an information retrieval system is unable to process data in this fashion, which can result in a large decrease in precision.

However, a yomi-based index may be valuable in combination with other index types, especially for the handling of orthographic varieties. The advantage of a pronunciation-based index is that it is independent from the orthography or written form of a word and therefore insensitive to orthographic variants (*e.g.*, okurigana, kanji, or kana variants). Therefore, even if the effectiveness of a purely yomi-based index is questionable owing to the large number of homophones in the Japanese language, it may prove effective in combination with other indexing approaches for the handling of orthographic variants.

2.3 Fuzzy Querying for Katakana Terms

We used the open source IR library Lucene as our basic retrieval engine, which is written in Java and licensed by Apache Software (http://lucene.apache.org/). Lucene offers a query option denoted as "FuzzyQuery", which matches terms similar to a specified term. If a query term is defined as "fuzzy", the similarity between terms in the index and a specified target term is determined using the Levenshtein distance algorithm. The edit distance affects the scoring, such that terms with lower edit distances are scored higher. Equation 1 shows how the FuzzyQuery distance is calculated [7]. The variable "targetlen" refers to the length of the target term.

$$sim_{fuzzy} = 1 - \frac{distance}{\min(textlen, targetlen)} \qquad (1)$$

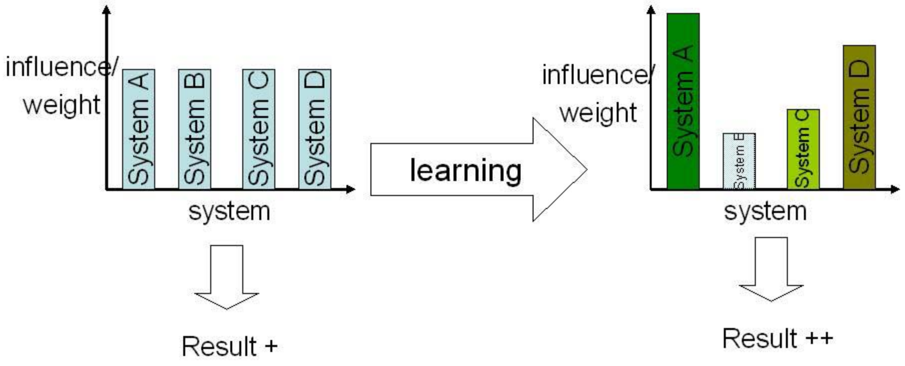
Since the calculation of fuzzy matches takes some time, the FuzzyQuery option needs to be employed with care. We decided to use it only for katakana terms to determine if it could prove helpful for the handling of katakana variants. We implemented Fuzzy Querying with a word-based index, and added it to our fusion approach as a fourth system besides the basic n-gram-, word-, and yomi-based systems.

2.4 The MIMOR Framework

MIMOR adopts a multiple indexing and fusion approach, combining the best-performing retrieval strategies to achieve an optimum retrieval result [1, 2]. The MIMOR model was originally inspired by the main outcomes of the Text Retrieval Conferences (TREC) (http://trec.nist.gov/), where it was found that many information retrieval systems perform similarly well in terms of their recall and precision, but do not lead to the retrieval of the same sets of documents. Multiple indexing and fusion approaches try to profit from these findings to gain access to a greater share of relevant documents through the integration of several approaches.

From a computational point of view, MIMOR is designed to act as a linear combination of the results of different retrieval systems or approaches [8]. The contribution of each system is controlled by a weight for that system. Relevance feedback is used to gradually optimize the fusion parameters, *e.g*. the weights of the individual indices for the fusion of the result lists. Figure 1 illustrates this learning process.

Figure 1. Fusion and learning in MIMOR. Systems A to D correspond to the three indexing strategies (bi-gram-based, word-based, and yomi-based), along with Fuzzy Querying.



2.5 Fusion in Japanese IR

Similar to the findings of TREC, the evaluations of the NTCIR Workshop series have not produced a clearly superior system, but rather, show systems performing equally well using very different indexing approaches. The two basic approaches used were word-based indexing, which requires Natural Language Processing (NLP) techniques, and n-gram indexing, which is language independent. Both strategies led to similar results, but their effectiveness varied case-by-case [9, 10]. To take maximum advantage of the strengths of the individual approaches, while at the same time minimizing their disadvantages, a number of enhanced approaches have been suggested. Among these are the "combination of evidence", or fusion approaches. These approaches merge the result lists obtained using more than one index type, usually by coupling word-based and n-gram-based indices. The results show that ranking documents based on a multiple index search is a promising strategy in Japanese information retrieval [11, 12, 13].

Jones *et al.* [11] used a combination of evidence techniques combining word-based and character-based indexing, and found that the use of a combination showed marginally better results. Sakai *et al.* [12] employed character-based and morpheme-based matching to avoid matching problems caused by nonexplicit word boundaries. Vines and Wilkinson [13] tried several different indexing strategies (*e.g.*, character-based, word-based, and bi-gram-based with unsegmented English strings), and subsequently combined the two best approaches: words without English, and bi-grams without English. The document score was calculated using the simple formula of $\text{sim}_{\text{new}} = 0.5 \cdot \text{sim}_1 + 0.5 \cdot \text{sim}_2$. This combination of evidence approach produced a further improvement on average of 1.2 percentage points.

3 Methods

3.1 Test Collections

Our experiments were carried out using two different test collections to compare the effectiveness of our approach across text genres. We chose part of the NTCIR-4 collection, the Mainichi Shimbun articles from 1998, as an example of a rather standardized collection. Major news companies, including Mainichi Shimbun, have strict usage guidelines concerning vocabulary and orthography for contributors. The second test corpus used was the NTCIR-1 collection of scientific abstracts. About half the documents were from the fields of electronic engineering and computer sciences, where new concepts are created frequently and rapidly, and the terminology is mostly borrowed from English. The frequency of katakana terms was therefore much higher than in the newspaper test collection (14.5% in the NTCIR-1 collection versus 10.0% in the Mainichi '98 collection). Additionally, we expected more variation in vocabulary and orthography in this collection because of the heterogeneity of the authorship.

Search requests were generated from all fields of the topic descriptions, *e.g.* title, description, narrative, and concept. For the experiments using the newspaper genre, we used the 46 NTCIR-4 topics with more than five relevant documents in the Mainichi '98 collection. For the experiments using the NTCIR-1 collection, we used the official test Topics 31 to 83. The calculation of the average precision for each run was based on the relaxed relevance judgments provided by the NTCIR-1 and NTCIR-4 workshops, respectively.

3.2 Indices

Three different indices were created: a bi-gram-based index, a word-based index, and a yomi-based index. For the bi-gram-based index, the hiragana characters were discarded, the katakana and roman character strings were left in their original forms, and the kanji character strings were divided into overlapping bi-grams. The morphological analysis for the word- and yomi-based indices was carried out using the Japanese morphological analyzer ChaSen (http://chasen.aist-nara.ac.jp/hiki/ChaSen/). Out-of-vocabulary words, *i.e.* words not recognized by ChaSen, were divided into bi-grams. This can be called a hybrid approach [14, 15]. For the yomi-based index, in the case of more than one suggested reading for a term, the readings were indexed as separate terms (*e.g.*, ナマモノ and セイブツ for 生物). This led to more single tokens compared to the word-based index (*see* Table 3). The low number of yomi types compared to the number of word types reflects the abundance of homophones in the Japanese language, and hints at a possible loss in precision.

Table 3. Number of index terms.

	Mainichi ’98	NTCIR-1
Word	184,657	113,542
Yomi	149,454	94,928
N-gram	513,118	910,893

Table 4. Index sizes.

	Mainichi ‘98 (146 MB)	NTCIR-1 (311 MB)
Word	356 MB	628 MB
Yomi	390 MB	706 MB
N-gram	355 MB	649 MB

A stoplist for each individual index was created determining the 100 most frequent index terms and we decided heuristically which of those terms should be discarded. In the case of the scientific abstracts collection, we decided to discard terms such as 研究 (research), 方法 (method), 実験 (experiment), 検討 (investigation, study), 結果 (result), and 目的 (purpose), which act as structure words, and are to be found in practically every scientific document. Similarly, we discarded terms such as 記事 (article) and 問題 (problem) for queries within the news domain. The yomi stoplist contained some equivalents of typical stop terms that were also to be found in the word-based stop list, such as モノ (thing), as well as the numerals 0 (レイ、ゼロ) to 9 (キュウ), and a number of individual syllables.

3.3 Fusion Strategy

To determine the influence of the yomi-based index on the retrieval effectiveness, we carried out experiments using a triple index: word-based, bi-gram-based, and yomi-based, and added Fuzzy Querying as a fourth system. After initial test runs to determine the performance of the individual systems, we adapted their weights manually to obtain an optimum fusion result. The fusion strategy we adopted was Z-Score, which was successfully employed by Savoy [16] in the NTCIR-4 data, and yielded the best results in our earlier study [3].

Z-score fusion allows for a normalized linear combination of the search results. The contribution of the individual systems is controlled using a weight represented by the parameter *α* (*see* Equation 2).

$$Z-ScoreRSV_k = \alpha \cdot \left[\frac{RSV_k - Mean^i}{Stdev^i} + \delta^i \right] \tag{2}$$

$$\delta^i = \frac{Mean^i - Stdev^i}{Stdev^i}$$

Key: RSV stands for “Retrieval Status Value”, i.e., the score assigned to a retrieved document

4 Results

4.1 Performance of the Individual Systems

In a first step, we evaluated the performance of our four individual systems for both document collections. Table 5 shows the mean average precision (MAP) per system.

Table 5. The MAP obtained using the individual systems.

	Mainichi ’98	NTCIR-1
Yomi-based index	0.3707	0.2776
Word-based index	0.3634	0.2775
N-gram-based index	0.3819	0.3072
Fuzzyword querying	0.3572	0.2392

The marked difference in retrieval performance across the two collections is probably owing to the fact that our system had originally been designed to handle newspaper articles. However, we can observe the same order of performance for the four individual systems. The best-performing system in both cases was the bi-gram-based approach.

Surprisingly, the yomi-based approach slightly outperformed the word-based approach for both collections. We were able to improve the performance considerably when comparing these results with the data we obtained in our former analysis of yomi-based indexing using the Mainichi ’98 collection [3]. This is owing to two adaptations:

1. an improved handling of numbers, as ChaSen was set to concatenate numerals; and
2. stopword filtering of the most frequent terms.

Fuzzy Querying clearly performs the worst. However, Figures 2 and 3 show that there are a number of cases where Fuzzy Querying outperformed the other approaches. An analysis of Topic 54 (marked with a circle in Figure 3), for example, revealed that one of the katakana query terms was ファイバ (faiba = fiber). However, the index contained only its variant ファイバー (faibaa). The variant was contained in 407 abstracts, 96 titles, and 203 keyword fields, while not a single document contained the original search term.

Figures 4 and 5 show the recall/precision curves of the individual systems. Interestingly, the yomi-based indexing performed best for high recall values using the Mainichi ’98 collection. For the NTCIR-1 collection, the n-gram-index system performed best for all recall values.

Figure 2. Average precision per topic using the Mainichi ’98 collection.

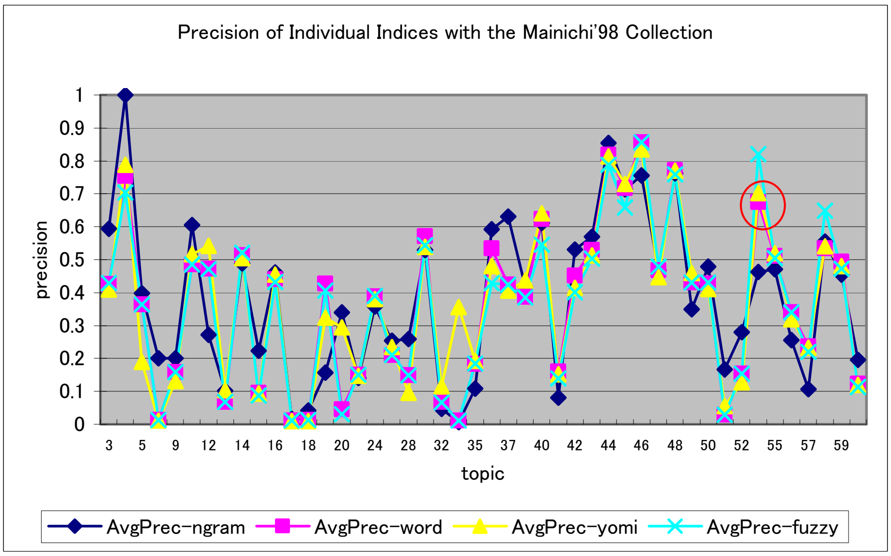


Figure 3. Average precision per topic using the NTCIR-1 collection.

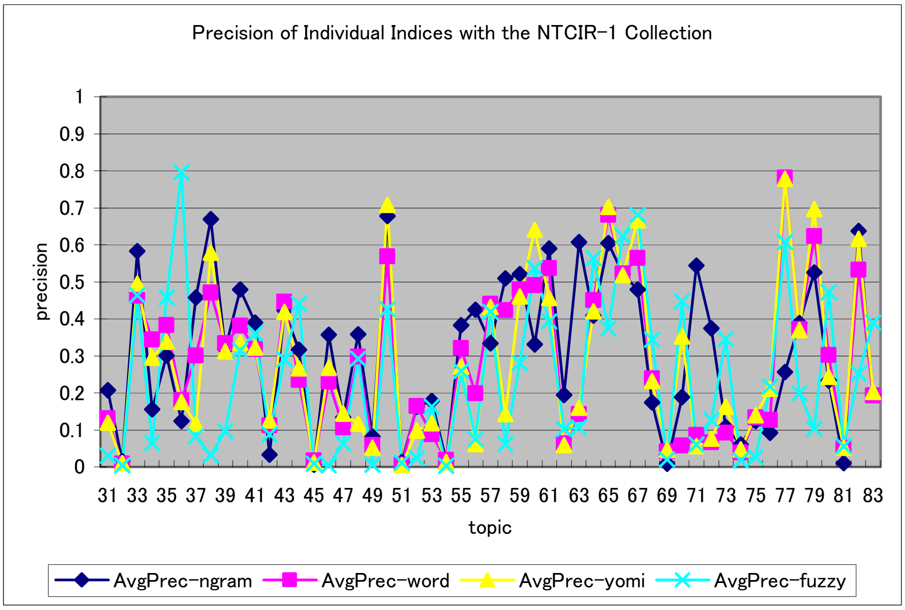


Figure 4. Recall/precision curves of the individual systems (Mainichi '98 collection)

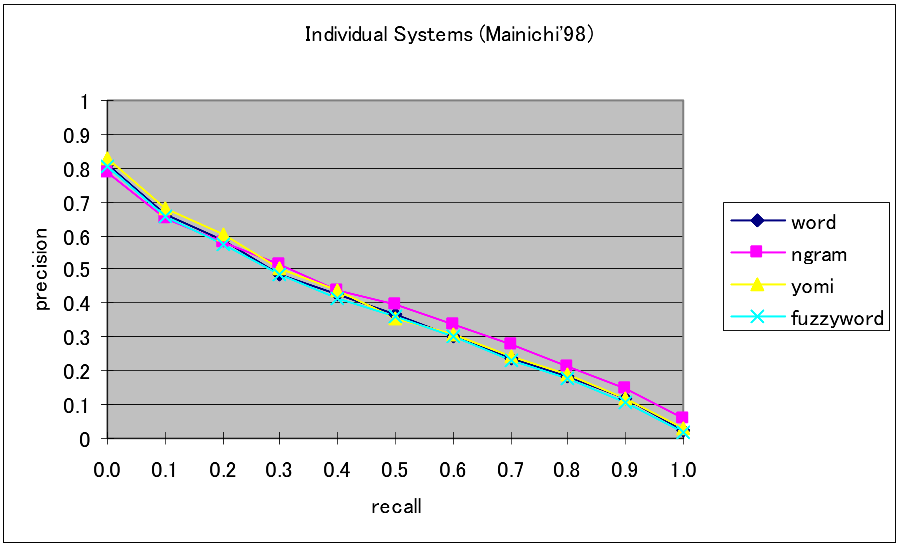
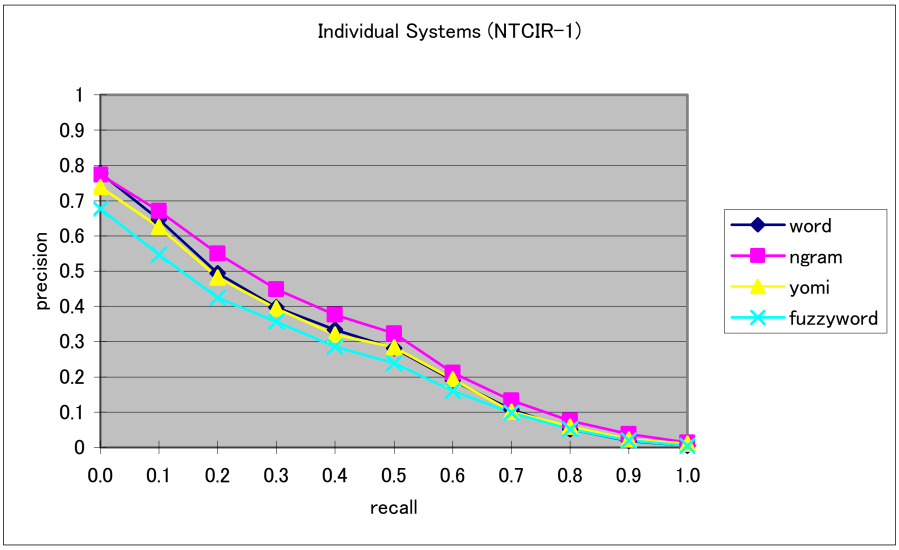


Figure 5. Recall/precision curves of the individual systems (NTCIR-1 collection).



4.2 Performance of Fusion Runs

For our fusion experiments, we first carried out a test run, assigning the same basic weight of unity to each of the indices. We then manually tuned the weights, using the heuristic that indices that performed better in the single runs should be assigned a higher weight. It turned out however, that the best performance for both test collections was yielded by the non-tuned runs. Tables 6 and 7 show the percentage improvement reached in the fusion runs. As a baseline for the comparisons, we chose to use the results of the best single run, which was the bi-gram-based system, with an average precision of 0.3072 using the NTCIR-1 collection, and of 0.3819 using the Mainichi ’98 collection. Comparing the fusion runs with the best single run helps to see how much increase in retrieval performance can be achieved with fusion at the cost of having several indices and taking longer time for query processing.

Table 6. The MAP of the fusion runs (NTCIR-1 collection).

Weight				Avg. Precision	% improvement from the single n-gram-based run
N-gram	Word	Yomi	Fuzzy word		
1	1	0	0	0.3094	0.71
1	1	1	0	0.3276*	6.64
3	1	1	0	0.3193*	3.93
2	1	1	0	0.3230*	5.14
1	1	1	1	0.3278*	6.69
3	3	3	2	0.3278*	6.68

Key: * = statistically significant compared to the performance of the single bi-gram-based system (T-Test, confidence level = 95%).

Table 7. The MAP of the fusion runs (Mainichi ’98 collection).

Weight				Avg. Precision	% improvement from the single n-gram-based run
N-gram	Word	Yomi	Fuzzy word		
1	1	0	0	0.3798	–0.56
1	1	1	0	0.3947	3.35
3	1	2	0	0.3910	2.38
2	1	1	0	0.3899	2.11
1	1	1	1	0.3953	3.50
3	3	3	2	0.3952	3.50

The results show that the additional yomi-based index leads to a significant improvement in retrieval effectiveness over the single bi-gram-based index using the NTCIR-1 collection of scientific abstracts. It also led to an increase in the MAP using the Mainichi ’98 collection. However, this increase was not significant. The precision can be improved further by adding Fuzzy Querying as a fourth system.

5 Discussion

We tested two strategies for the handling of orthographic varieties in Japanese: yomi-based indexing, and Fuzzy Querying. Integrating a yomi-based index system and merging the results obtained with a bi-gram-based, a word-based, and the yomi-based system led to a significant improvement in precision for the NTCIR-1 collection of scientific abstracts, and to a slight improvement in precision for the Mainichi Shimbun '98 collection. Fuzzy Querying was not effective as a single approach. However, it led to a minor improvement of precision within our fusion system.

So far, we have made use of an inbuilt function of the Lucene search engine for Fuzzy Querying. There is, however, room for an improvement in the algorithm. At the moment, all letter pairs are considered as being equally different in the calculation of the editing distance. However, as can be seen from Table 2, there are only a limited number of error patterns that can lead to differences in spelling. Confusion only occurs when there is more than one transcription of a foreign sound. Consequently, there are pairs of letters that are more likely to be confounded (*e.g.*, キ (/ki/) and ク (/ku/) for the rendering of the English sound /ik/ as in "cake"), and letters that are more likely to be added or omitted (such as the maguro "ー", or certain vowels for the indication of a long sound).

An improved Fuzzy Query algorithm should take these characteristics into account. This could be effected by a simple table listing the similarity between two syllables, and only the limited number of typical error patterns would need to be defined. This approach is comparable to the transliteration approach adopted for English/katakana transliteration [17, 18].

6 Acknowledgments

The authors would like to thank the DAAD (German Academic Exchange Service) for financing the research stay which made this study possible.

7 References

[1] Hackl, R.; Kölle, R.; Mandl, T. & Womser-Hacker, Ch. (2002): Domain Specific Retrieval Experiments at the University of Hildesheim with the MIMOR System. In: Proceedings of the CLEF 2002 Workshop. Berlin: Springer [LNCS 2406], pp. 343-348.

[2] Womser-Hacker, Ch. (1996): Das MIMOR-Modell. Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval, Habilitationsschrift, Universität Regensburg.

[3] Kummer, N.; Womser-Hacker, Ch. & Kando, N. (2005): Re-Examination of Japanese Indexing: Fusion of Word-, N-gram- and Yomi-Based Indices. In: Proceedings of the 11th Annual Meeting of The Association for Natural Language Processing, March 14–18, 2005, University of Kagawa, Kagawa Prefecture, Japan.

[4] Halpern, J. (2002): Lexicon-Based Orthographic Disambiguation in CJK Intelligent Information Retrieval. In: Proceedings of the 19th Conference on Computational Linguistics, COLING-2002, August 24–September 1, 2002, Taipei, Taiwan.

[5] Halpern, J. (2000): The Challenges of Intelligent Japanese Searching. Working paper (www.cjk.org/cjk/joa/joapaper.htm), The CJK Dictionary Institute, Saitama, Japan, revised 2003.

[6] Taylor, I. & Taylor, M. M. (1995): Writing and Literacy in Chinese, Korean and Japanese (Studies in Written Language and Literacy), Amsterdam / Philadelphia: John Benjamins Publishing Co.

[7] Gospodnetić, O. & Hatcher, E. (2004): Lucene in Action. Manning, Canada.

[8] Mandl, Th. & Womser-Hacker, Ch. (2001): Probability Based Clustering for Document and User Properties. In: Ojala, T. (ed.): Infotech Oulo International Workshop on Information Retrieval (IR 2001), Oulo, Finland. September 19–21 2001, pp. 100–107.

[9] Yoshioka, M.; Kuriyama, K. & Kando, N. (2002): Analysis of the Usage of Japanese Segmented Texts in NTCIR Workshop 2. In: Proceedings of the Second NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and Text Summarization, National Institute of Informatics, Tokyo, Japan, pp. 291–296.

[10] Ozawa, T.; Yamamoto, M.; Umemura, K. & Church, K. W. (1999): Japanese Word Segmentation Using Similarity Measure for IR. In: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, August 30–September 1, 1999, Tokyo, Japan, pp. 89–96.

[11] Jones, G.J.F.; Sakai, T.; Kajiura, M. & Sumita, K. (1998): Experiments in Japanese Text Retrieval and Routing Using the NEAT System. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp 197–205.

[12] Sakai, T.; Shibazaki, Y.; Suzuki, M.; Kajiura, M.; Manabe, T. & Sumita, K. (1999): Cross-Language Information Retrieval for NTCIR at Toshiba. In: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, August 30–September 1, 1999, Tokyo, Japan, pp. 137–144.

[13] Vines, P. & Wilkinson, R. (1999): Experiments with Japanese Text Retrieval Using mg. In: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, August 30–September 1, 1999, Tokyo, Japan, pp. 97–100.

[14] Chow, K.C.W.; Luk, R.W.P.; Wong, K.-F. & Kwok, K.-L. (2000): Hybrid Term Indexing for Different IR Models. In: Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages. Hong Kong, China, pp. 49–54.

[15] Luk, R.W.P.; Wong, K.-F. & Kwok, K.-L. (2001): Hybrid Term Indexing: An Evaluation. In: Proceedings of the Second NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and Text Summarization, National Institute of Informatics, Tokyo, Japan, pp. 130–136.

[16] Savoy, J. (2004): Report on CLIR Task for the NTCIR-4 Evaluation Campaign. In: Proceedings of the Fourth NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, pp.178-185.

[17] Fuji, A. & Ishikawa, T. (2001): Japanese/English Cross-language Information Retrieval: Exploration of Query Translation and Transliteration. Computers and Humanities 35, Kluwer Academic Publishers, Netherlands, pp. 389–420.

[18] Qu, Y.; Grefenstette, G. & Evans, D.A. (2003): Automatic Transliteration for Japanese-to-English Text Retrieval. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03), July 28–Aug. 1, 2003, Toronto, Canada, pp. 353–360.