Re-examination of Japanese Indexing: Fusion of Word-, Ngram- and Yomi-Based Indices

Nina Kummer (NII/Univ. Hildesheim), Christa Womser-Hacker (Univ. Hildesheim), Noriko Kando (NII)
nina@nii.ac.jp, womser@rz.uni-hildesheim.de, kando@nii.ac.jp

Abstract

The contribution of a yomi-based index to the retrieval effectiveness of a multiple-index system was analyzed. The results show that the yomi-based index, although insufficient as a standalone index, contributes to a slight increase in retrieval effectiveness, when combined with a bigram- and a word-based index. However, the weight attributed to the yomi-based index must not be too high.

Background

This work is part of a project with the aim of integrating Japanese language support into the MIMOR framework [8]. MIMOR stands for „Multiple Indexing for Dynamic Method-Object-Relation in Information Retrieval“ and adopts a multiple indexing and fusion approach in order to profit from the advantages of the best-performing technologies for an optimal retrieval result. Over time, the system learns the best combination of weights for the fusion of result lists.
The MIMOR model is originally inspired by the main outcomes of TREC[1], where it was found that many information retrieval systems perform similarly well in terms of recall and precision but do not lead to the same sets of documents. Multiple indexing and fusion approaches try to profit from these findings in order to gain access to a greater share of relevant documents through the integration of several approaches.

Fusion in Japanese IR

Similarly to the findings in TREC, the evaluations of the NTCIR Workshop series have not produced one clearly superior system, but rather comparably well performing systems using very different approaches. The two basic approaches are word-based indexing, which requires NLP techniques, and ngram indexing, which is completely language independent. Both strategies lead to similar results, however, their effectiveness varies case-by-case.
In order to take maximal 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 others “combination of evidence” or fusion approaches. These approaches merge the results lists obtained with more than one index type, usually coupling word-based and ngram-based indices. The results show that ranking documents on the basis of a multiple index search is a promising strategy in Japanese information retrieval [3][4].
Apart from the basic word-based and n-gram-based indices, we created a third, yomi[2]-based index and evaluated the effectiveness of their fusion.

Yomi-Based Indexing

Yomi-based indexing used to be employed in times 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 was abandoned since the introduction of double-byte character handling. Due to its fairly simple sound system, the Japanese language is very rich in

[1] Text Retrieval Conferences (http://trec.nist.gov/)
[2] Japanese for the “reading” or “pronounciation” of a word.

homophones. In written langue, the ideographic kanji characters or the use of different scripts normally help to keep these apart. A phonetical transcription of Japanese lacks this information and can therefore be very ambiguous at times. Whereas human readers may still be able to guess the meaning of ambiguous words from the context, an information retrieval system incurs many false drops, which can result in heavy losses in precision.
However, a yomi-based index might be valuable in combination with other index types, especially for the handling of orthographic varieties. The advantage of a pronounciation-based index is that it is not sensitive to orthographic variants, e.g. okurigana, kanji or kana variants, which are highly frequent in Japanese and represent a special challenge for Japanese IR [2].
The advantage of a yomi-based index lies in its independence from the orthography or written form of a word – also with regard to the form of the input provided by the user. Therefore, even if a purely yomi-based index would be questionable due to the high number of homophones in the Japanese language, it might prove effective in combination with other indexing approaches for the handling of orthographic variants.

Methods

In order to determine the influence of a yomi-based index on retrieval effectiveness, we carried out experiments with a triple index, word-based, ngram-based, and yomi-based. The results were merged with three different fusion strategies:

- Raw Score
- SumRSV: SumRSV = $\sum \alpha_I \cdot RSV_i$, where α_I may be used to represent the weight of an index
- Z-score (a normalized version of SumRSV, described in [5])

Raw Score can be considered as the most basic fusion strategy and was mainly used for first testing. SumRSV and Z-Score were successfully employed by [5] in NTCIR-4, where especially Z-Score yielded promising results.
The contribution of the individual indices to the final result set was controlled by weights.
The tests were carried out with the Mainichi Shimbun articles of 1998, which are part of the NTCIR-4 collection, and the corresponding NTCIR-4 topics.
For the ngram index, hiragana characters were discarded, katakana and roman character strings were left in their original form, and kanji character strings were divided in overlapping bigrams.
The morphological analysis for the word- and yomi-based indices was carried out with ChaSen[3]. Out-of-vocabulary words, i.e. words not recognized by ChaSen, were divided into bigrams. This can be called a hybrid approach [1].
For the yomi-based index, in the case of more than one suggested readings for one term, the readings were indexed as separate terms (e.g. ナマモノ and セイブツ for 生物). This leads to more single tokens in comparison to the word-based index (cf. Table 1). The low number of yomi types compared to the number of word types reflects the abundance of homophones in the Japanese language and already lets anticipate a certain loss in precision.

Table 1. Type-Token Ratio of the word-based and yomi-based indices

	tokens	**types**	**ratio**
word	21,426,876	94,124	0.00439
yomi	23,413,585	70,680	0.00302

Search requests were generated from the <TITLE> and <CONC> fields of the 46 NTCIR-4 topics with more than 5 relevant documents in the Mainichi 1998 collection. For each run, the average precision was calculated on the basis of the relaxed relevance judgements provided for the NTCIR-4 workshop.

[3] http://chasen.aist-nara.ac.jp/hiki/ChaSen/

Results

Figure 1 presents the average precision of individual runs with the bigram-, word-, and yomi-based indices.
Expectedly, the yomi-based index generally performs least well, in many cases even very poorly. In some individual cases, however, it outperforms the best-performing bigram strategy. We can observe a certain correlation between the results yielded by the yomi-based and the word-based index, which can be explained by the fact that they both depend on the quality of the output of ChaSen. In three cases (topics 12, 19, and 33), the yomi-based index slightly outperforms the word-based index.

Figure 1.

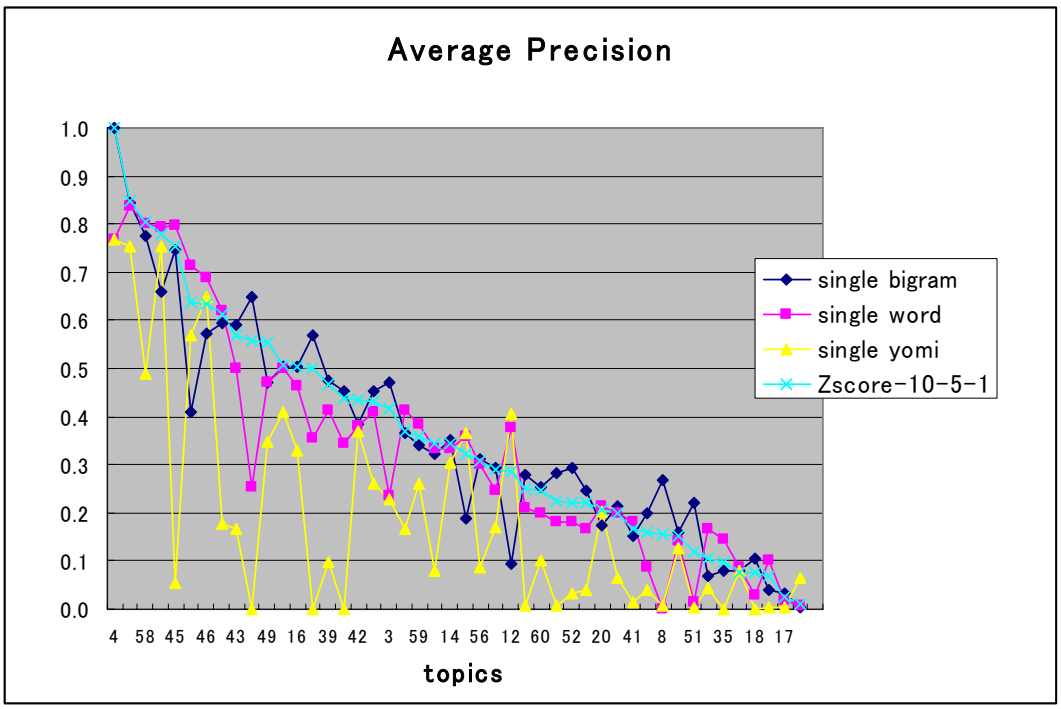


Table 2. Single-Run Results Compared to the best Fusion Run

Index	Avg. Precision
Bigram	0.35963047
Word	0.335239
Yomi	0.19794926
Score-10-5-1	0.36612925

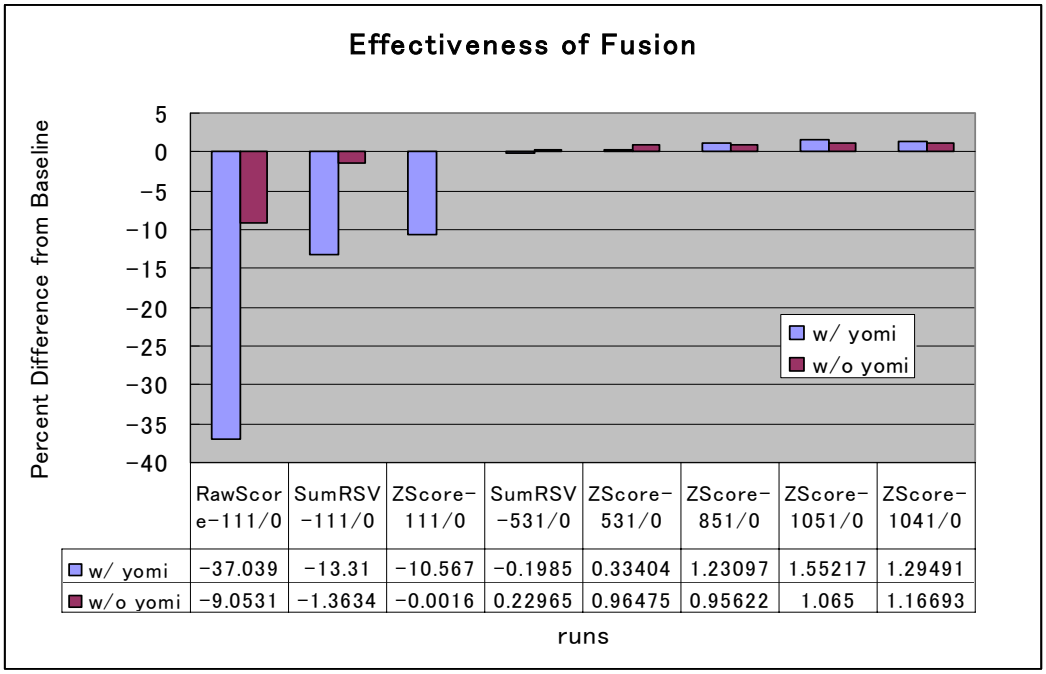
Figure 3 shows the results of different combinations of indices, fusion strategies and index weights in percent difference from the baseline (single bigram-based index). The run IDs are read as follows: fusion strategy – weight bigram, weight word, weight yomi.
The best-performing fusion strategy is Z-score with a high weight for the bigram index and a very low weight for the yomi-index (bigram: 10, word: 5, yomi: 1). However, the improvement gained through this combination is not statistically significant.
The nature of the influence of the yomi-based index on the retrieval result depends on merging strategy and weights. Only in an optimal combination does the additional yomi-based index lead to an improvement of the retrieval result.

Discussion

We could show that a yomi-based index may lead to a slight increase in retrieval effectiveness. If the weight is chosen too high, however, the yomi-based index clearly decreases the result. Given the

increase in storage cost and processing time caused by an additional index, a yomi-based index is not very effective for the collection and topics we examined in our analysis. The picture might be different with a less standardized collection and user-input queries, where orthographic varieties play a more important role.

Figure 3. Results of Fusion Runs



	RawScore−111/0	SumRSV−111/0	ZScore−111/0	SumRSV−531/0	ZScore−531/0	ZScore−851/0	ZScore−1051/0	ZScore−1041/0
w/ yomi	−37.039	−13.31	−10.567	−0.1985	0.33404	1.23097	1.55217	1.29491
w/o yomi	−9.0531	−1.3634	−0.0016	0.22965	0.96475	0.95622	1.065	1.16693

References

[1] Ken C. W. Chow, Robert W. P. Luk, K. F. Wong & K. L. Kwok (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.

[2] Jack Halpern (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.

[3] Gareth J. F. Jones, Tetsuya Sakai, Masahiro Kajiura & Kazuo Sumita (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.

[4] Tetsuya Sakai, Yasuyo Shibazaki, Masaru Suzuki, Masahiro Kajiura, Toshihiko Manabe & Kazuo Sumita (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.

[5] Jacques Savoy (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.

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