Deciding Indexing Strings with Statistical Analysis

Yoshiyuki TAKEDA Kyoji UMEMURA
Toyohashi University of Technology,
1-1 Tempaku, Toyohashi, Aichi 441-8580, Japan
take@ss.ics.tut.ac.jp, umemura@tutics.tut.ac.jp

Eiko YAMAMOTO
Communications Research Laboratory
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
eiko@crl.go.jp

Abstract

Deciding indexing string is important for Information Retrieval. Ideally, the strings should be the words that represent the documents or query. Although each single word may be the first candidate of indexing strings for English corpus, it may not ideal due to the existence of compound nouns, which are often good indexing strings, and which depends on genre of corpus. The situation is even worse in Japanese or Chinese where the words are not separated by spaces. In this paper, we proposed a method to decide indexing strings based on statistical analysis. The novel features of our method are to make the most of the statistical measure called adaptation and not using language dependent resources such as dictionaries and stop words list. We have evaluated our method using Japanese test collection, and we have found that our method actually improves the precision of information retrieval systems.

1 Introduction

Classical information retrieval system selects documents by finding the matching of indexing strings between a query and documents. Therefore, the choice of indexing strings greatly affects the precision of information retrieval systems. Some systems have a list of good keywords to select the indexing strings, and this is the common case for Japanese in which the words are not separated by spaces and list of words are usually necessary to extract words. Forming predefined list of index strings causes several problems.

The first problem is the dependency on genre. The meaning and importance of a certain string may change according to the situations. For example, the term ”security” is more important in some collections (newspapers) than others (security alerts). Ideal index strings should express information needs of user and should express important concept in corpus [2]. In other words, importance of index string depends on the corpus.

The second problem is the difficulty of listing all good indexing strings in any language. Practically speaking, dictionaries can hardly contain every word. They tend to miss some words such as proper nouns, acronyms and technical terms. This situation is serious because these words are important for information retrieval. The situation is more serious for the Japanese retrieval system that uses morphological analysis system such as ChaSen[9], which uses list of pre defined words. These systems may completely fail to capture the words that are not contained in the list.

The third problem is boundary of indexing strings. Although English looks not having this problem, it is a problem because 85% of technical term is a compound noun [11]. This situation becomes worse in Japanese and Chinese where no spaces exist between words.

This paper proposes a method that extracts index strings by statistic analysis. This report is the enhanced material of the previous work [18] with additional evaluation. This method does not use dictionaries. Therefore it does not suffer from the three problems above. The main feature of our approach is to make the most of adaptation [1] based on recent work on Adaptive Language Model [3]. In the previous work, this method is proposed as dictionary-free extraction of keywords [14], which are natural keyword to documents. We have evaluated this system for information retrieval application. We confirm the effectiveness of the proposal method for Japanese and

Chinese test collection, in which the deciding indexing strings are much harder than in English.

2 The Feature of Adaptation

In English, it is reported that the value of adaptation is different with the contents word and function word [1]. We have found that this is also the case for in Japanese and Chinese [17]. We have also found that adaptation contains the information of boundaries of chunks of words [17]. We will briefly describe our findings here as the basis of our proposal.

2.1 Adaptation in English

Church[1] shows that words have a tendency to appear in one document repeatedly, and he also reported that this tendency varies for the kind of words: the function words have low adaptation and the content words have high adaptation. Let $p(x|y)$ be the conditional probability, $e_k(w)$ be the number of k times appearance of term w and df_k be the number of documents that contain a term k or more times. Adaptation is defined as follows [1]:

$$\text{adaptation}(w) = p\left(e_1(w)|e_2(w)\right) \approx df_2/df_1$$

Adaptation is the conditional probability that shows word w appears repeatedly when that word is contained in a certain document. Adaptation is estimated by df_k of the whole corpus.

The content words and the function words can be distinguished clearly by df_2/df_1 because df_2/df_1 has following features in English, which is reported by Church [1]:

- df_2/df_1 of the contents word is larger than that of the function words.
- df_2/df_1 of the function words is larger that of strings under Poisson distribution.
- df_2/df_1 of the contents word tend to be constant regardless of df_1/N where N is the number of documents.

The value of df_1/N is well studied for term recognition [4] and frequently used to estimate whether a word is good keyword or not. Since df_2/df_1 is not correlated with df_1/N, and is some information to distinguish contents word from function words, we may expect better result to use both of them.

2.2 Adaptation of Arbitrary Strings and Keywords

Since words in English always have larger adaptation than that of Poison distribution, we are interested in the adaptation of valid words and arbitrary strings in Japanese and Chinese, because arbitrary strings will have closer distribution of Poisson.

We have used NTCIR1/2[5, 13] and Chinese Information Retrieval Benchmark version 0.10[13] which are test collections of information retrieval. NTCIR2JG and NTCIR2K are Japanese abstracts collections. CIRB010 is Chinese newspaper collection. We also have used Mainichi Shinbun which is Japanese newspaper [8]. MAINICHI91-97 is Japanese newspaper collection. These corpora contain the keywords that are selected by their author or their publishers. Since the keywords are a kind of indexing strings selected by authors, we try to reproduce the selection based on statistical analysis.

We have used df_2/df_1 as vertical axis and df_1/N as horizontal axis. Plots of the arbitrary string and the keywords are shown in the figure 1. The smoothed version is shown in the figure 2. In this figure, we have plotted the value of averaged df_2/df_1 that have the same df_1/N.

Figure 1 shows that df_2/df_1 of keywords distribute larger value than arbitrary strings. This implies df_2/df_1 has the information to extract keywords from arbitrary string. Figure 2 shows that df_2/df_1 of the keywords are constant regardless of df_1/N in all the collections (for example Japanese and Chinese, abstracts and newspapers). This implies that df_2/df_1 is independent from df_1/N, and independent information source.

2.3 Adaptation on the Boundary of Keywords

Figure 3 shows a curve of df_2/df_1 for strings starting from the head of proper nouns. The vertical axis is the value of adaptation and the horizontal axis the length of the string. We can observe that adaptation is stable within the boundary of the proper noun. If the length of the string exceeds the length of proper noun, the adaptation drops. Figure 4 shows the same curve of df_1/N. We cannot observe the stable region within the boundary and it is rather hard to predict the boundary with df_1/N, whereas df_2/df_1 has clear dropping point. Figure 5 shows the adaptation of strings that consist of keywords and additional character. We call this string surrounding strings. The horizontal axis and vertical axis are same as figure 1 and figure 3. This figure implies that we can distinguish keywords from strings in df_2/df_1 vs. df_1/N plot. This figure also implies that df_1/N alone does not have enough information to distinguish keywords from surrounding. In other words, we need df_2/df_1 for the extraction.

Since the keywords provided by authors or publishers are natural candidate for indexing strings, we have build the system to extract strings using both df_1/N and df_2/df_1. We need to note that df_1/N and df_2/df_1 are statistical functions and language independent con-

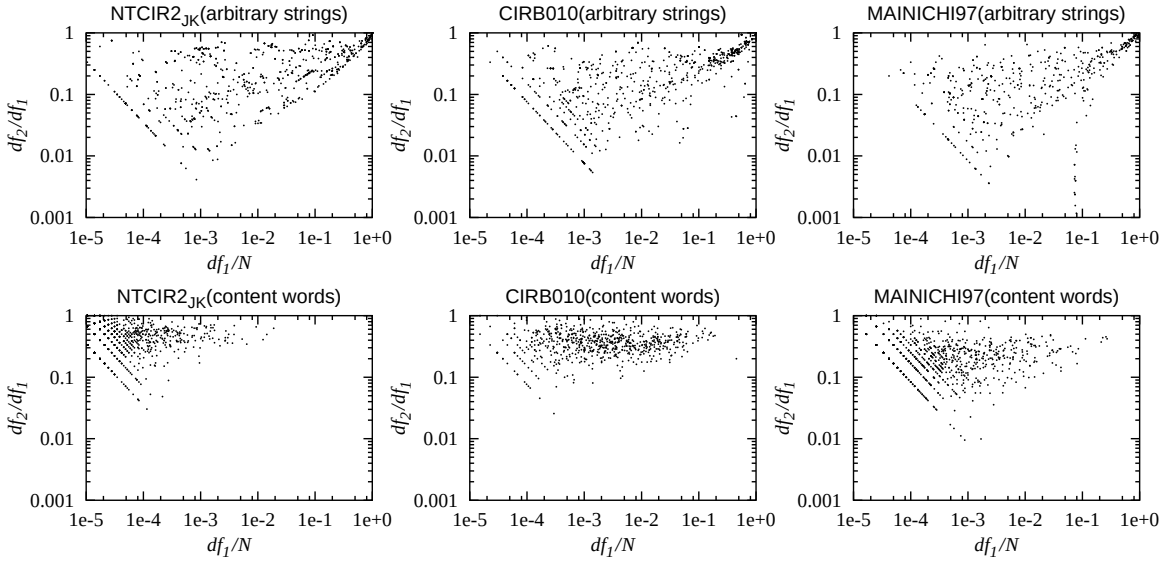

Figure 1. df_2/df_1 of the keywords are larger than df_2/df_1 of the arbitrary strings

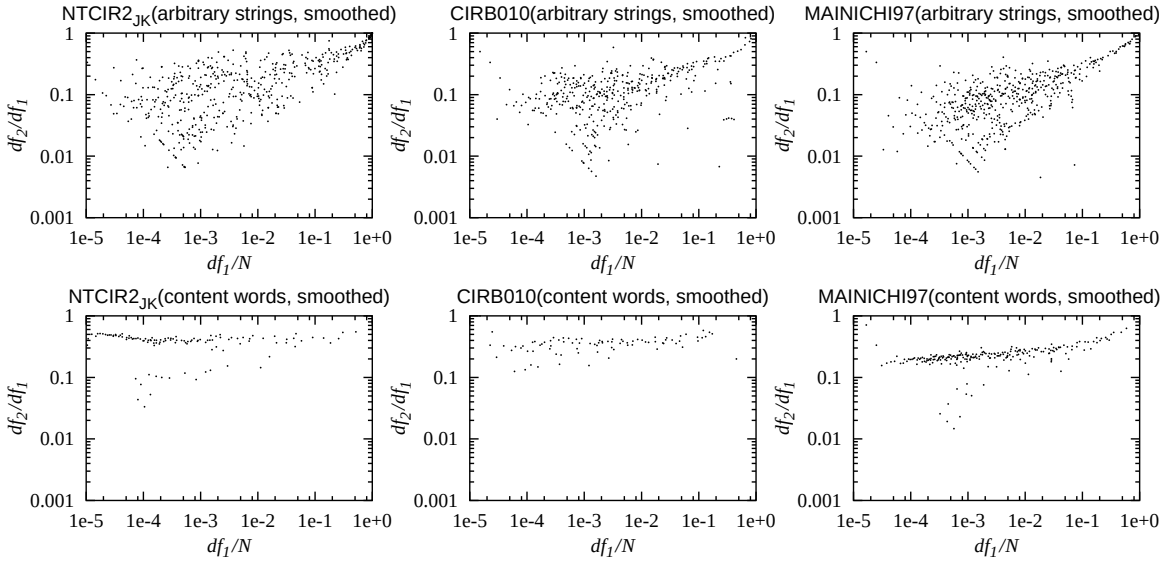

Figure 2. The averaged df_2/df_1 of the same df_1/N

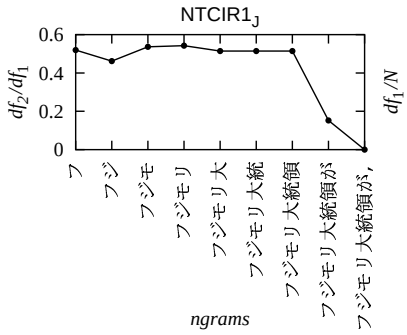

Figure 3. df_2/df_1 of various length of strings: we can predict the boundary of a proper noun

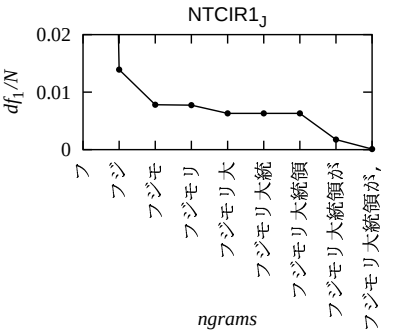

Figure 4. df_1/N of various length of strings: it is hard to tell where is the boundary of the proper noun

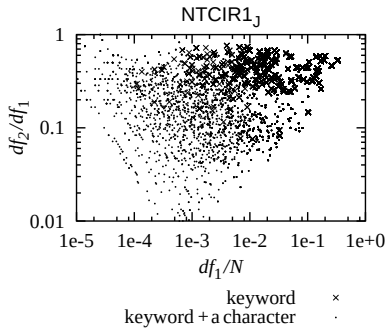

Figure 5. df_2/df_1 of keyword and keyword+a character: they are distributed in different region

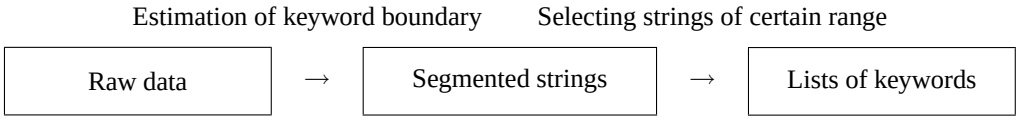

Figure 6. Indexing Strings Extraction Proceedure

cept, yet they have information to extract keywords from arbitrary string in Japanese and Chinese. We also need to note that df_2/df_1 (adaptation) may become large when the corpus tents to treat the string as keywords, even if the word generally is regarded as functional words.

3 Indexing String Extraction Procedure

We have developed a procedure to extract indexing strings observing the figures of previous section. We have utilized the facts that adaptation within the boundary is larger than surrounding strings, and the value within the boundary may be regarded as constant. We also utilized the fact that we can distinguish the keywords and surrounding strings in adaptation and frequency plain.

Indexing string extraction procedure is divided into two steps as figure 6. First, we estimate the boundary of keywords, and then we select keywords.

We have defined an empirical score obtained from results of figure 1 and figure 2. We regard this indexing string likelihood, assuming that keywords will have high indexing string likelihood. Please note that the strings with extremely high frequency also have large adaptation. They sometimes have larger adaptation than keywords even if the strings are not good indexing strings. Therefore, we adjusted the value of likelihood with the same level of good keywords. We also regard this score of the string with low frequency as negative infinity assuming that the value of adaptation is not reliable at all.

$$\mathrm{score}(w) = \begin{cases} \log \frac{df_2(w)}{df_1(w)} & \text{for } df_2(w) \geq 3, \frac{df_1(w)}{N} \leq 0.5 \\ \log 0.5 & \text{for } df_2(w) \geq 3, \frac{df_1(w)}{N} > 0.5 \\ -\infty & \text{for } df_2(w) < 3 \end{cases}$$

Using this likelihood, we have calculated the segmentation that will give the highest sum of these likelihood values of segmented strings. This is formulated as below:

$$words = \mathrm{argmax}_W \left(\sum_{W=\{w_1, w_2, \cdots, w_n\}} \mathrm{score}(w_i) \right)$$

where W is a segmentation of given string, and w_i is the i-th segmented string of W.

Next, for each string in the best segmentation: $words$, we calculate df_2/df_1 and df_1/N. We have observed that strings of one character exist as the result of segmentation problem for function words. Therefore, we add a heuristics that we will ignore the strings of one character. Observing the distribution of the keywords, as is shown in figure 1 and figure 2, we formulate the following condition.

$$conditions = \left\{ \begin{array}{l} 0.1 < df_2(w_i)/df_1(w_i), \\ 0.00005 < df_1(w_i)/N < 0.1, \\ \mathrm{length}(w_i) > 1 \end{array} \right\} \quad (1)$$

where $\mathrm{length}(w_i)$ is length of string w_i.

Although the constants including in formula (1) depend on target corpus, we can decide the constants from the existence range by examining a sample set of good indexing strings. The output of the system is the set of the strings that satisfies this condition.

Ozawa[15] reports segmentation of the similar kind. Ozawa used idf as the score function, and reported this segmentation improves the information retrieval precision. Although segmentation by idf may not produce natural indexing string, proposed segmentation by our likelihood tends to produce more natural words than Ozawa's system.

It should be noted that this segmentation method does not uses any kind of dictionaries. All that need is the sample of good keywords and reasonably large corpus. Yet, we have observed that it can decide meaningful string.

Although indexing strings might not necessary be the meaningful string for information retrieval, theoretically indexing unit should be statistically independent each other, and the strings corresponding to word boundary will be more independent than the strings within words. Therefore we may say this approach fits the theoretical framework, and provide an actual way to provide independent string unit.

It reported that using bigram or n-gram as indexing strings works well for Chinese or Japanese [14, 21]. We can still use our system with these systems.

After extracting indexing strings in a query, we can ignore other strings in a query as the source of noise. Although this approach is not theoretical at all, this makes it possible to use indexing extraction system with existing information retrieval systems.

Table 1. Example of Indexing Strings of Proposal Method

No.	Query
0001	**自立移動ロボット** / について **autonomous mobile robot** / about
0002	**複合名詞** / の / **構造解析** / において、 / **シンボリック** / な手法と / **統計的** / な手法を / **組み合わせ** / た / **アプローチ** / を取る研究はないか。 **compound noun** / of / **structural analysis** / in field of / **symbolic** / a method of / **statistical** / a method of / **combination** / **approach** / is there any research of
0003	**機械** / **学習** / における / **サンプル** / **複雑性** / について論じている / **文献** **machine** / **learning** / about / **sample** / **complexity** / is discussing about / **publication**

4 Experiments

In the evaluation, we have extracted indexing strings from each query with several different methods including proposed one. Then, we have compared information retrieval precision replacing each query with the corresponding extracted indexing strings. We have used NTCIR1 collection [5, 13]. This collection consists of 83 queries and about 330,000 documents of Japanese technical abstracts. NTCIR1 collection contains two kind of relevance judgment called RIGID and RELAXED. We have obtained 11 points average precision with the standard tools called TRECEVAL for both judgments.

We compare proposal method: SIS with four other methods. The first method is HMN. A human subject who is native speaker of Japanese and knows technical fields including the structure of information system selects the important indexing string by hand. We believe that this is the best method we can do for indexing string selection, even if it is very expensive to perform information retrieval. We choose this method as optimal case to measure how much efforts remain. The second method is ALL. All the strings in each query are used as it is. We choose this method as base line to measure the improvements as total. The third method is DF1. We select strings whose bigrams have similar document frequency as the keywords. We do not use df2 at all to show the importance of df2. The forth method is CHA. This method uses the output of Chasen and selects the string using part of speech information provided by Chasen. We selected the segmented strings labeled as ”nouns” and ”unknown words”. We choose this method as a representative of the dictionary based systems.

We also need several weighting schema because the precision is affected by term weighting. We choose two kind of schema. This first schema is TFIDF. TFIDF is standard weighting schema described in textbook. TFIDF is a variation of $tf(d,w) \cdot idf(w)$. It is reported that TFIDF shows better result than $tf(d,w) \cdot idf(w)$. TFIDF is defined as follows.

$$tfidf(d,w) = (1 + \log tf(d,w)) \left(1 + \log \frac{N}{df(w)}\right)$$

where $tf(d,w)$ is frequency of word w in the document d, $df(w)$ is the frequency of documents containing term w.

The second schema is ETW. ETW stands for Empirical Term Weighting. ETW is a schema that we have been using in our experience with NTCIR. It is kind of statistical model, and shows reasonable performance for NTCIR collection. η is defined as follows.

$$\begin{aligned}\eta(d,w) &= \max(0, \min(idf(w), \\ &\qquad a_{tf(d,w)} + b_{tf(d,w)} idf(w))) \\ idf(w) &= \log \frac{N}{df(w)}\end{aligned}$$

The coefficients of a_i and b_i are empirically decided by the regression of log odds from the relevance judgment of training collection. See Reference [21] for the detail of ETW.

5 Results

First, we show the example of indexing strings that SIS outputs. Table 1 shows the output for some queries. The indexing strings are in bold letter. The next line is direct translation of each string. It is interesting that we can have some translation for the string that is produced by the system without language dependent knowledge.

Table 2 and table 3 show performances of information retrieval for all methods. First of all, we have found the SIS is the best system in our experiment. It should be noted that HMN does not work well, even if SIS learns indexing strings by the examples of human. Let us recall the procedure of SIS. SIS will select strings within a region of df_1/N vs. df_2/df_1 plain. It is not sure whether human being will select every string within this range. This suggests that human can select good indexing strings because SIS works, but the human may not select all the indexing strings that should

Table 2. Gross 11pt. average precisions (RIGID)

	HMN	SIS	ALL	DF1	CHA
TFIDF	0.2226	0.3042	0.2293	0.2633	0.1997
ETW	0.2408	0.3220	0.2821	0.2841	0.2408

Table 3. Gross 11pt. average precisions (RELAXED)

	HMN	SIS	ALL	DF1	CHA
TFIDF	0.2278	0.3153	0.2356	0.2726	0.2121
ETW	0.2404	0.3318	0.2908	0.2936	0.2488

Table 4. The number of WINs of SIS

	RIGID				RELAXED			
	CHA	DF1	ALL	HMN	CHA	DF1	ALL	HMN
TFIDF	64	51	66	31	64	53	65	34
ETW	56	50	53	24	55	48	51	27

be included because the precision of HMN is less than SIS.

The precision of DF1 is in between SIS and ALL. This suggests that selecting indexing strings by statistical measure effective because DF1 is more precise than ALL. This also suggests that the contribution of adaptation (df_2/df_1) is not negligible.

The worst system is CHA. We estimates that this is due to poor dictionary because the query is about special technical documents.

Next, we are interested in the behavior of individual query and statistical significance. Therefore, we have compared 11 points average precision of each query. Table 4 shows the number of WINs for each method, where WIN represents the situation that proposed method has higher 11-point average than another.

The interesting case is HMN. The number of WIN out of 83 is contradict the result of previous one. One possible explanation is that Human sometimes makes a fatal mistake even if Human usually makes slightly better choice.

The results of other systems are consistent with previous experiment. The proposed system (SIS) is the best system, and DF1 is the second best. We can perform hypothesis test by the number of WIN. The null hypothesis is that the probability that the target system shows better performance is less than 0.5. Assuming that each query is independent, the number of WIN obeys binomial distribution. The probability that WIN is more than m times satisfies following formula:

$$P(X \geq m) = \frac{1}{2} \int_0^{1/2} \sum_m^n {}_nC_m p^m (1-p)^{n-m} dp$$
$$\leq \sum_m^n {}_nC_m \left(\frac{1}{2}\right)^n$$

From results of table 2, the hypothesis is rejected at the rate of 3.92×10^{-2} or less. The performance differences difference with SIS and other systems are statistically significant.

6 Related Work

Our indexing string extraction procedure consists of word segmentation and indexing strings selection. In information retrieval, it is important to choose the word segmentation method, because it directly affects the retrieval performance. Many word segmentation methods based on statistical analysis have been proposed for Japanese so far. In previous works, there are many methods using a lexicon [12, 19], a segmented training data [6, 10, 20], or an unsegmented training data [7]. These methods use language information such as a part-of-speech and types of character. Our method dose not use such language information. There is another method using frequency information in training data [7] like our methods. While our method uses only document frequency, this method needs collection frequency and knowledge of a part of speech. The below methods are proposed for morphological analysis in Japanese. As opposed to these, we propose a word segmentation method for obtaining effective words in information retrieval. In Chinese information retrieval, there is a report for over-segmentation phenomenon; thats' accurate segmenters lead to reduce retrieval performance [16]. This suggests that a segmenter for information retrieval is not necessarily an accurate segmenter. Therefore, we conclude that our proposal segmenter is useful for information retrieval at least, though our obtained words are not necessarily correct words.

7 Conclusion

We have described a method to decide indexing strings by statistical analysis. The method calculates document frequency and adaptation for every string of queries, then decide which strings are likely to be keywords learning from the examples of keywords. As the result, we have found that we can obtain meaningful strings from the system without and language dependent resources.

We have measured the effectiveness of the indexing strings of our system. We found that the average performance of the system is better than that of other system including human selection, even if the system try to follow the human behavior. We have also verified that adaptation have significant contribution for improving performance in our system.

References

[1] Church, K.W., Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2, *Coling2000*, pp. 173–179, 2000.

[2] Cleverdon, C., Optimizing convenient online access to bibliographic databases, Information Services and Use, vol. 4, pp.37–47, 1984.

[3] Jelinek, F. Statistical Methods for Speech Recognition *MIT Press*, 1997.

[4] Kageura, K. and Umino, B., Methods of Automatic Term Recognition, *Terminology*, vol. 3, no. 2, pp. 259–289, 1996.

[5] Kando, N., Overview of the Japanese and English IR Tasks at the Second NTCIR Workshop (Draft), *Proceedings of the Second NTCIR Workshop Meeting*, pp.4-37–4-60, 2001.

[6] Kashioka, H., Black, E., and Eubank, S., Decision-Tree Morphological Analysis without a Dictionary for Japanese, *NLPRS97*, pp.541–544, 1997.

[7] Kubota-Ando,
R. and Lee, L., Mostly-Unsupervised Statistical Segmetation of Japanese: Applications to Kanji, *ACL, ANLP2000*, pp.241–248, 2000.

[8] Mainichi Shinbun 91, 92, 93, 94, 95, 96, 97.

[9] Matsumoto, Y. Kitauchi, A., Yamashita, T., Hirano, Y., Imaichi, O., and Imamura, T., Japanese Morphological analysis System ChaSen Manual, NAIST Technical Report NAIST-IS-TR97007, 1997,
http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

[10] Mori, S. and Nagao, M., Unknown Word Extraction from Corpora using N-gram Statistics, *Journal of the Information Processing Society of Japan*, vol. 39, no. 7, pp.2093–2100, 1998, (in Japanese).

[11] Nagao, M., The Iwanami Software Science Series 15: Natural Language Processing, *Iwanami-Shoten*, 1988, (in Japanese).

[12] Nagata, M., A part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context, *ACL99*, pp.277–284, 1999.

[13] NTCIR Project, http://research.nii.ac.jp/ntcir/

[14] Ogawa, Y. and Matsuda, T., Overlapping statistical word indexing: A new indexing method for Japanese text, *SIGIR97*, pp. 226–234, 1997.

[15] Ozawa, T., Yamamoto, M., Umemura, K., Church, K., Japanese word segmentation using similarity measure for IR. *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 89–96, 1999.

[16] Peng, F., Huang, X., Schuurmans, D., and Cercone, N., Investigating the Relationship between Word Segmentation Performance and Retrieval Performance in Chinese IR, *Coling2002*, vol. 2, pp.793–799, 2002.

[17] Takeda, Y. and Umemura, K, Document Frequency Analysis which Realizes Keyword Extraction, *Mathematical Linguistics*, vol. 23, No. 2, 2001, (in Japanese).

[18] Takeda , Y. and Umemura, K., Selecting Indexing Strings using Adaptation, *SIGIR2002*, pp. 427–428, 2002.

[19] Takeuchi, K. and Matsumoto, Y., HMM Parameter Learning for Japanese Morphological Analyzer, *Journal of the Information Processing Society of Japan*, vol. 38, No. 3, pp.500–509, 1997, (in Japanese).

[20] Uchimoto, K., Sekine, S., and Isahara, H., Morphological Analysis Based on A Maximum Entropy Model – An Approach to The Unknown Word Problem –, *Journal of The Association for Natural Language Processing*, vol. 8, No. 1, pp.127–141, 2001, (in Japanese).

[21] Umemura, K. and Church, K.W., Empirical Term Weighting and Expansion Frequency, *Workshop SIGDAT, EMLNP2000*, pp. 117–123, 2000.