# Analysis of the Usage of Japanese Segmented Texts in NTCIR Workshop 2

Masaharu YOSHIOKA Kazuko KURIYAMA Noriko KANDO
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
{yoshioka,kuriyama,kando}@nii.ac.jp

## Abstract

*In this paper, we report on the usage of Japanese segmented texts and analyze the submitted search results to NTCIR Workshop 2, which used these texts. In these texts, each sentence is segmented into terms and term components (similar to phrases and words). However, the sizes of terms are inconsistent in the texts; e.g., some terms that should be decomposed into term components remain as terms. We analyze the effect of this inconsistency from the viewpoint of comparison between word-based indexing and phrasal indexing. Based on this analysis, we propose the desired specification of a morphological analyzer for Information Retrieval.*

**Keywords:** *NTCIR, Japanese Text Segmentation, Morphological Analyzer.*

## 1 Introduction

There are two main approaches in Information Retrieval (IR) for Japanese text retrieval. One is character-based indexing and the other is segmented text-based indexing. The former includes bigrams and n-grams, and the latter includes single words and phrasal indexings. In the last NTCIR workshop, several groups discussed the comparison and combination of these approaches [1, 7, 8].

However, as Chen *et al.* stated in [1], the comparisons between these results were not very appropriate because of the difficulties of segmenting Japanese text. In NTCIR workshop 2, we used a commercial Japanese morphological analyzer tool, which is used in several operational Japanese IR service systems, to generate Japanese segmented texts as baseline data to study this problem.

We used the Happiness/BASE3.5[1] system to generate Japanese segmented texts. This software uses a dictionary with approximately 130,000 entries and segments Japanese texts at two levels: ”hard segmentation” and ”soft segmentation”. The former indicates segmentation between terms, and the latter indicates segmentation between components of terms. From the viewpoint of indexing, usage of both forms of segmentation allows single-word indexing, while usage of hard segmentation only allows phrasal indexing. The user can use both segmentations in their IR system.

We received two submitted search results of runs that used these segmented texts. One was for the J-J task and the other was for the J-JE task from differnt group. In this paper, we discuss the relationships between the search effectiveness of the IR system and the consistency of the segmentation by the morphological analyzer, based on the submitted runs.

## 2 Effect of the Morphological Analyzer

### 2.1 Character-based Indexing and Segmented Text-based Indexing

Ozawa *et al.* [7] compared the bigram-based method with the word-based method from the viewpoint of term frequency. Their results showed that the word-based method was good at queries that included technical terms longer than a bigram. For example, “LFG” (that is an acronym of “Lexical Functional Grammer”) is an example of a good term for the word-based approach, because “LF” and “FG” have frequencies more than 10 times higher than “LFG.” On the contrary, the bigram-based method was good at queries that included terms as combinations of frequently used words. “文書画像理解” which means “text image understanding,” is an example, because “文書” (text), “画像” (image), and “理解” (understanding) are frequently used words. On the other hand, 90% of the occurrences of “像理,” which extends over “画像” and “理解” are related to “画像理解,” so “像理” is a good keyword for “画像理解.”

Based on this comparison, Ozawa *et al.* proposed an adaptive n-gram segmentation method that removes noisy n-grams, and changes the length of n-grams ac-

[1] Happiness is a registered trademark of Heiwa Information Center Co., Ltd

cording to the similarity between the query and the document.

Another issue related to Japanese segmented texts in IR is the comparison between single-word-based indexing and phrasal indexing. Fukushima *et al.* [3] used short units that corresponded to single words, and long units that corresponded to phrases, in their IR system. From their experiences, they concluded that short-unit indexing was better than long-unit indexing.

On the other hand, Fujita [2] compared single-word indexing and phrasal indexing, and discussed the characteristics of each indexing method. They concluded:

- *Especially in Japanese language, such noun compounds sometimes make domain specific terminology that is usually useful as a good discriminator of subject concept description. Given such terminological characteristics, indexers introduced precoordination of indexing terms, mainly adopting phrasal terms in order to preserve syntactic relations.*

- *Since in phrasal indexing miss-match in one word is propagated to phrase level, performance is more sensitive to noises than in single word indexing.*

Based on these ideas, Fujita proposed using a combination of single-word indexing and phrasal indexing with adequate term weighting.

### 2.2 Inconsistency in Japanese Morphological Analyzers

Because written Japanese language provides no explicit word boundary markers, most (perhaps all) Japanese morphological analyzers define single words based on dictionary entries.

This may cause problems when some phrasal terms are represented as single words. For example, consider the word "解析" (analysis). In the Japanese segmented texts for NTCIR workshop 2, we can find the single word "解析値" (analyzed value) and also the two-word phrase "解析　した　値" (analysis do value = analyzed value).[2] Because, in those terms, "解析" is used with a similar meaning, from consistency considerations, "解析値" should be segmented as "解析__値." Detailed discussion of consistent segmentation can be found in [5].

Before discussing the effects of inconsistencies in the real data, we should discuss the effect of inconsistencies by comparing the word-based approach with the character-based approach. We discuss the issues with the example of "開発" (development) in Table 1.

In the second case in Table 1, "再開発" should be segmented into "再__開発." However because in the dictionary entry "再開発" is stored as "再開発", no index entry is made for "開発" in word-based indexing. However, character-based indexing works well for that situation. On the other hand character-based indexing also makes an index entry for "公開__発表" in the third case in the table.

Consider using only the query term "開発" in an IR system. A character-based IR system finds all such terms, resulting in good recall with bad precision. In contrast, a word-based IR system finds some of the terms and rejects some, resulting in good precision with bad recall.

Therefore, in the analysis of *Interpolated Recall Precison Averages* for comparisons between character-based and word-based systems, if the system has inconsistent word segmentation, the precision in larger recall value decreases more than the character-based indexing result.

## 3 Analysis of the Submitted Data

### 3.1 General Analysis of the Morphological Analyzer

In this paper, we discuss what types of segmented texts work well in the word-based method, by analyzing query terms and search result for each topic. We use term frequency as the measure for analyzing results.

Initially, we analyzed the consistency of word segmentation by using query terms in the submitted query lists. First, we retrieved texts that included query terms from the original Japanese texts (non-segmented texts) by using character-based retrieval. Second, we checked the corresponding parts of the segmented Japanese texts and classified the results into three categories that were the same as those in Table 1. Percentages (approximate) of classification are as follows:

- Appropriate segmentation (query term is segmented as word): *Matched (M)* **75%**

- Inappropriate segmentation (word is composed with another element): *Overlapped (O)* **25%**

- Appropriate segmentation (query term is divided into different words): *Separated (S)* **0.1%**

Because the accuracy of segmentation in the commercial morphological analyzer from the viewpoint of grammatical analysis is higher than 95%, the mistakes in the text segmentation have less effect than the inconsistencies in the word boundary definitions.

We therefore focused on the effect of inconsistencies in analyzing the submitted papers. These inconsistencies included the identification of the small units and the lengths of the phrases that were represented as words because of the inconsistencies.

[2] "__" represents the word boundary that corresponds to soft segmentation and "　" represents the word boundary that corresponds to hard segmentation in our Japanese segmented texts.

**Table 1. Effect of segmentation in relation to relevance judgments**

| | Meaningful for relevance judgment | Character-based indexing | Word-based indexing |
|---|---|---|---|
| Appropriate segmentation *Match (M)*<br>e.g., 装置__開発<br>装置 (equipment) 開発 (development) | ○ | ○ | ○ |
| Inappropriate segmentation *Overlapped (O)*<br>e.g., 再開発<br>再 (re-) 開発 (development) | ○ | ○ | × |
| Appropriate segmentation *Separated (S)*<br>e.g., 公開__発表<br>公開 (public) 発表 (announcement) | × | ○ | × |

### 3.2 Analysis of the Submitted Papers

We received two submitted papers that used the segmented texts. One was for the J-J task and the other was for the J-JE task. Because the result of the J-JE task was biased by the translation between English and Japanese, we decided to use only the J-J task for the evaluation.

Comparison between the average of the average precisions for all systems and that for the submitted paper showed that some of its results were good, but some were not. In this analysis, we try to explain why, from the viewpoint of the effect of the consistency.

The worst example in this submitted paper was topic 0105 (41 relevant documents in level 2 (S+A+B)). From the topic information, a good keyword appeared to be "キノロン剤" (quinolone tablet). We checked how "キノロン剤" gathered from character-based retrieval data, was indexed in the word-based indexing format. Table 2 shows the result. The types in the table correspond to the categorization in Table 1. TF stands for Term Frequency and DF stands for Document Frequency.

**Table 2. Index entries for "キノロン剤" and their frequency**

| Type | TF(DF) | Index word |
|---|---|---|
| *M* | 61(23) | キノロン剤<br>quinolone tablet |
| *O* | 57(29) | ニューキノロン剤<br>new quinolone tablet |
| *O* | 3 (1) | フルオロキノロン剤<br>fluoro quinolone tablet |

From Table 2, we can see that "ニューキノロン剤" was not retrieved from the word index term "キノロン剤," even though "ニューキノロン剤" should be a good keyword because of the title of the topic "新規キノロン剤" (new quinolone tablet). In addition, on the subject of the consistency of segmentation, "キノロン剤" should be segmented as "キノロン__剤." In the text of the test collection, the frequency of "キノロン" (quinolone) was 561(208) and most of the relevant documents came from these documents.

On the other hand, there was no segmented word "キノロン" in the Japanese segmented texts for the topic. Searches could only find some of the "キノロン" documents that contained "キノロン剤" This may have a bad effect on recall value.

One of the good examples in this submitted result was topic 0104. This topic was "肺小細胞癌" (small-cell carcinoma of the lung) and was segmented as "肺__小細胞癌" (肺 (lung) 小細胞癌 (small-cell carcinoma)). For consistency of word segmentation, this should have been segmented as "肺__小__細胞__癌" (肺 (lung) 小 (small) 細胞 (cell) 癌 carcinoma). That would mean that the term "小細胞癌" would act as a phrasal index term, if we could assume consistent segmentation.

To evaluate the effectiveness of this phrasal index, we defined three types of segmented text index for the word "肺小細胞癌." A discussion of the effectiveness of the indexes follows. In each explanation, the number that follows the index word shows the frequency based on character-based indexing. In the second case, we also describe the frequency number for matching the term in the segmented texts.

1. "肺"(39, 683) "小" (321, 041) "細胞" (625, 207) "癌" (107, 453)
   Because all of the terms are high frequency terms, those words are not effective as keywords for retrieval. However, documents that include all of these terms (598 documents) include all of the relevant documents (41) in level 2 (S+A+B). This may have a positive effect on recall value.

2. "肺" (39, 683 (*Match* 25, 264)) "小細胞癌" (552 (*Match* 446))
   The "小細胞癌" index term has an appropriate frequency compared with the relevant documents and this explains why the system performed well. Documents that included all of these terms (220 documents) included most of

the relevant documents in level 2 (S+A+B) (39 from 41).

3. “肺小細胞癌” (321)
The “肺小細胞癌” index term has appropriate frequency compared with the relevant documents and this explains why the system performs well. However, documents that include this term (133 documents) miss 10 relevant documents in level 2 (S+A+B) from 41. In this case, the result should become worth by effects similar to those observed with topic 0104.

From the comparison between analyses 1 and 2, we can say that technical terms longer than bigrams work well. However, the comparison between analyses 2 and 3 shows that longer is not necessarily better.

The following topics are good examples compared with the average of average precisions for all of the system. We can identify an appropriate length for technical terms that is longer than bigrams and that has appropriate frequency according to the number of relevant documents. Most of those terms should be treated as phrases when considering consistency.

**0111** Number of relevant documents 231
Appropriate technical term: “ITS” (Abbreviation of Intelligent Transport Systems) (1010)

**0121** Number of relevant documents 202
Appropriate technical term: “vod” (Abbreviation of Video on Demand) (636)

**0122** Number of relevant documents 19
Appropriate technical term: “リテラシー” (literacy) (709)

**0123** Number of relevant documents 61
Appropriate technical term: “バイオフィルム” (biofilm) (135)

**0129** Number of relevant documents 57
Appropriate technical terms: “超対象性” (super-symmetry) (631)、宇宙項 (cosmological constant) (106)

**0130** Number of relevant documents 86
Appropriate technical term: “解析性” (Analyticity) (85)

**0141** Number of relevant documents 204
Appropriate technical term(s): “軌道法” (orbital methods) (1511)、非経験的 (not empirical: (ab initio in this context)) (534)

**0144** Number of relevant documents 94
Appropriate technical term: “パージング” (parsing) (68)

**Table 3. Indexes for “リテラシー” and their frequency**

| Type: | TF | Index word |
|---|---|---|
| *M* | 709 | リテラシー<br>literacy |
| *O* | 188 | コンピュータリテラシー<br>computer literacy |
| *O* | 31 | メディアリテラシー<br>media literacy |
| *O* | 19 | サイエンスリテラシー<br>science literacy |
| *O* | 48 | Other (ネットワークリテラシー<br>network literacy<br>コンピュータリテラシーテスト etc.)<br>computer literacy test |

For topic 0122, we found another element that made the result better. We checked how “リテラシー” was indexed from character-based retrieval data in the word-based indexing format (Table 3).

In this topic description, there is a restriction for the relevant document: “Such emerging concepts as information literacy, media literacy and computer literacy are not included.” Because the index for “リテラシー” did not include “メディアリテラシー” and “コンピュータリテラシー” by chance retrieved results were good at rejecting such documents.

The following topics are bad examples compared with the average of average precisions for all systems. Despite the bad results, most of those topics also included technical terms that were longer than bigrams.

**0115** Number of relevant documents 209
“ビデオストリーミング” (video streaming) (4) was a good keyword, but it had quite a low frequency. “ストリーミング” (streaming) (126) should be a good keyword, but the weight was low compared to the other keyword.

**0137** Number of relevant documents 15

All of the query terms have higher frequencies.

**0139** Number of relevant documents 225
“シックハウス” (Sick house) (11) was a good keyword, but it had quite a low frequency.

**0140** Number of relevant documents 217
“異性体” (isomer) (5075) should be a good keyword for “光学異性体” (enantiomer). However, because there are other varieties of “異性体,” such as “立体異性体” (stereo-isomer) and “構造異性体” (structure isomer), index term “異性体” gives the same effect for all oth these isomer.

**0143** Number of relevant documents 24
“障害者” (handicapped person) (5468) should be a good keyword for “視覚障害者” (sight-handicapped person; visually impaired person). However, because there are a variety of “障害者,” such as “聴覚障害者” (hearing-impaired person) and “身体障害者” (physically handicapped person), index “障害者” works well to select all of these handicapped persons.

**0148** Number of relevant documents 55
“プレストレストコンクリート橋” (prestressed concrete bridge) (11) was a good keyword, but it had quite a low frequency. “プレストレストコンクリート” (prestressed concrete) (538) should have been a good keyword, but the weight was low compared to the other keyword.

Based on these analyses, there are two types of bad effects in word segmentation. One derives from the inconsistency of word segmentation, i.e., long phrases are segmented as single words (0105, 0115, 0139, 0148). Despite the lower value of the average precision, *Interpolated Recall Precison Averages* at recall 0.00 equals 1 in these cases and that means the system works well for selecting good documents in higher rank. From this result, we assume that these inconsistencies may cause a reduction of the recall value.

The other bad effect is the inappropriate length of phrase; i.e., phrases are not long enough to identify specific topics (0140, 0143). For those cases, precision at five documents for topics 140 and 143 was much lower than the average. From this result, we assume these short phrases may retrieve irrelevant documents that include words that are similar to the index word.

For topic 0143, the identification of the specific concept for “視覚障害者” is important, because there is a restriction for the relevant document in this topic description: “Papers about handicapped persons who have no sight handicap do not satisfy the request.”

Compared with the results for 0122 and 0143, longer phrasal indexes may work well for the identification of a specific topic that excludes similar topics.

### 3.3 Discussion

From the analysis of the submitted paper, we can summarize the effect of the inconsistency of the morphological analyzer in this NTCIR workshop 2 text data.

- When the system segments texts by using longer phrases as words, the recall value may be reduced. In particular, when the system uses phrases that have omittable affixes and may thus by chance connect with a variety of affixes, there may be a remarkable reduction in recall value.
- When the system segments texts only by using small single words, most terms have high frequencies. This may cause good recall but reduction of precision.
- When the system uses domain-specific terminology as longer phrases, this may have good effects on the IR system. In addition, to identify specific terms and exclude similar ones, it is better to use longer phrases.

Based on this discussion, a morphological analyzer (or text segmentation system) for an IR system should have the ability to create segmented texts as combinations of consistent word-segmented texts and phrase identification. Phrase identification means the identification of technical terms that are combinations of segmented words.

We know it would be very difficult to implement such an ideal system, but the research result of the lexically motivated corpus [5, 4] will give some guidelines for implementing such an ideal system.

The Happiness morphological analyzer has the ability to handle these two levels (word level and phrase level), but it is not an easy task to make a good dictionary for all technical terms. ChaSen [6] has also started to deal with these two levels. We would therefore like to see these systems improved in the future.

When the ideal morphological analyzer (or text segmentation system) becomes available, IR systems should have a capability to select the types of index terms according to the topic type and the frequencies of the terms.

Based on the analysis of the J-J task result, we also analyzed the J-E collection result. The tendencies of the result were similar, but we could not find close relationships between them because other elements affected the retrieval results more.

## 4 Conclusion

In this paper, we discussed the relationships between the search effectiveness of an IR system and the consistency of the results from the morphological analyzer, based on the results of searches of segmented texts. We also proposed an ideal morphological analyzer (or text segmentation system) for IR based on the analysis result.

## Acknowledgment

We would like to thank the two groups who submitted results of using Japanese segmented texts.

## References

[1] A. Chen, F. C. Gey, K. Kishida, H. Jiang, and Q. Liang. Comparing multiple mthods for Japnese and Japanese-

English text retrieval. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 49–58, 1999.

[2] S. Fujita. Notes on phrasal indexing JSCB evaluation experiments at NTCIR AD HOC. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 49–58, 1999.

[3] T. Fukushima and S. Akamine. A character-based indexing and word-based ranking method for Japanese text retrieval. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 179–182, 1999.

[4] K. Kageura, M. Yoshioka, K. Takeuchi, and T. Koyama. Overview of the TMREC tasks. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, page 415, 1999.

[5] T. Koyama, M. Yoshioka, and K. Kageura. The construction of a lexically motivated corpus: The problem with defining lexical units. In *First International Conference on Language Resources and Evaluation*, pages 1015–1019, 1998.

[6] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. *Morphological Analysis System ChaSen version 2.2.1 Manual*. Nara Institute of Science and Technology, 2000.

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

[8] P. Vines and R. Wilkinson. Experiments with Japanese text retrieval using mg. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 97–100, 1999.