A Japanese Information Retrieval Method Using Syntactic and Statistical Information

Tsunenori Mine  Hiroki Fujitani  Makoto Amamiya
Graduate School of Information Science and Electrical Engineering, Kyushu University
6-1 Kasuga-kouen, Kasuga 816-8580, JAPAN
{mine,fujitani,amamiya}@al.is.kyushu-u.ac.jp

Abstract

This paper presents a Japanese information retrieval method using the dependency relationship between words and semantic and statistical information about them.

Our method gives a score to each document to be retrieved, based on a probabilistic model, giving an additional score based on the dependency relationship between words, and ranks the documents according to their score. In order to make use of semantic information, our method performs query expansion with a thesaurus, and zero pronoun resolution. Experimental results show that our method is more effective than a method using linguistic and statistical information independently.

1 Introduction

Information Retrieval is attractive as a crucial technology to find what users want. As the number of electronic documents, such as e-mail messages and web pages, increases, an information retrieval system providing high-speed and high-accuracy search performance has become desirable.

The fusion of a conventional information retrieval method based on statistical information about target documents with a method using linguistic knowledge is known as a promising approach to achieving high-accuracy search performance. However, most information retrieval systems utilize only statistical information about target documents, or at most shallow linguistic information such as cooccurrence information about noun phrases paired with verbs, and a thesaurus for expanding queries.

This paper presents the fusion of two information retrieval methods, namely, a Japanese information retrieval method using dependency relationships between words, and semantic information about them (which we call the DRB method for short) and Robertson’s approximation to the 2-Poisson Model (Robertson and Walker, 1994) based on a probabilistic model (the Robertson’s method for short). The DRB method treats a frame of a verb and nouns which have a dependency relationship between them as a key for judging whether or not a document is relevant to a query. Furthermore, it expands a query with a thesaurus, and performs zero pronoun resolution to create a frame when necessary. Since the targets of information retrieval are usually documents, it has been expected that the performance of information retrieval systems can be improved by using language processing techniques, and accordingly, we proposed the DRB method (Tateishi et al., 1999). This method proved that using the dependency relationship in information retrieval is highly effective in terms of precision, but on the other hand, is too exacting to retrieve an appropriate number of documents. It often does not even retrieve any document when quite a few relevant ones do exist. We therefore believe that combining the DRB method with a probabilistic model could lead to a more effective method. We call this method, combined method for short.

In what follows below, Section 2 describes the DRB method and the Robertson’s method, and Section 3 describes the combined method. Section 4 discusses experiments and their results.

2 The DRB Method and The Robertson’s Method

2.1 The DRB Method

The DRB method uses a frame that consists of a verb and nouns which have a dependency relationship as a key for judging whether or not a document is relevant to a query.
Basic Algorithm of the DRB Method

1. Transform a query into a frame structure of verb and nouns.
2. Transform every sentence in each document into a frame structure.
3. Compare the frame of the query with each frame of the document, and retrieve every sentence whose frame is recognized as identical to that of the query.

When a query is transformed into two or more frames, the sentence retrieved must match each frame of the query.

We define Frame match for judging whether two frames are identical or not as follows:

**Frame match:** If the verbs and nouns of two frames are respectively identical or synonymous, these frames are recognized as identical.

2.1.1 Special Frame

Although a frame usually consists of a verb and nouns, we also treat the phrase “noun A + attributive particle (‘no’) + noun B” in Japanese as a frame. We consider that two frames of this kind are identical if they meet at least one of the following conditions:

1. the A nouns and B nouns are identical or synonymous, and they are connected directly or indirectly with the attributive particle “no”.
2. the A nouns and B nouns are identical or synonymous, and they constitute a compound noun.
3. the A nouns and B nouns are identical or synonymous, and they are connected with an arbitrary verb.

If two or more nouns are connected with the particle ‘no’, the phrase is decomposed so as to make a set of “noun A ‘no’ noun B’ phrases. For example, the phrase “noun A ‘no’ noun B ‘no’ noun C” is decomposed into two phrases: “noun A ‘no’ noun B” and “noun B ‘no’ noun C”. They are then dealt with according to the method described above.

2.2 Query Expansion

In order to obtain synonyms for keywords in a query, we utilize the EDR conceptual dictionary of the EDR electronic dictionary (Yokoi, 1995) (EDR, 1996). It expresses relationships between concepts as a tree. Each word belongs to a vertex in a tree, and a word representing a concrete concept belongs closer to a leaf. When some words represent an identical concept with different written forms, such as Japanese Katakana characters, they belong to the same vertex. Taking account of usage in daily life, it is desirable to obtain not only the words which belong to the same vertex as the keyword, but also those which belong to, as synonyms, the neighboring vertices, because words with a tiny difference in meaning are treated as different concepts in the EDR conceptual dictionary. For this reason, the synonyms obtained for a keyword that belongs to a vertex v will be all the words of v, the descendants of v and the parent of v. The reason why the descendants, not just the children, are included is that most words we usually use belong to the vertices close to the leaves. When the synonyms are obtained by expanding the parent of v, some restrictions suggested in Ohta et al.’s report (Ota and Okumura, 1997) are applied. When a word represents multiple concepts, the synonyms are obtained for each concept in the way we mentioned above.

2.3 Zero Pronoun Resolution

In Japanese, once a word appears in a sentence, the word or pronoun can be left out in the following sentences. That phenomenon is known as zero pronoun. Considering the possibility that two frames compared by the DRB method could be identical if the elided word was restored, the following algorithm is used to compensate for the elision.

**2.3.1 Algorithm**

for a query “noun A + postpositional particle + verb B”, if noun A and verb B are not in the same frame of a document, execute the following steps.

**STEP1:** Stop this zero pronoun presumption analysis and conclude that noun A and verb B are not correlative, if noun A occurs in the sentence just after the sentence where verb B occurs.

**STEP2:** Conclude that noun A and verb B are correlative and that A has been elided, if A is either the topic, subject or object of the first sentence in the document.

**STEP3:** Conclude that noun A and verb B are correlative and that A has been elided, if A is either the topic, subject or object of one of M sentences that immediately precedes the sentence where B occurs. We set M to 4 based on experimental results (Tateishi et al., 1999).

It can readily be recognized whether A is the topic, subject or object by checking the postpositional particle which follows A. This is based on the Centering Theory (Kameyama, 1986) (Walker...
et al., 1994). In step 2, the first sentence in the document is checked in the first place following Nakaiwa et al.’s report (Nakaiwa and Ike-hara, 1993) that the first sentence usually summarizes the substance of the document, and therefore there is a great possibility that the word to be omitted later is present.

2.4 The Robertson’s Method

The Robertson’s method calculates a score for each target document, $Score^R$, with the following equation, provided that we adopt $\frac{Length_i}{\Delta} \simeq 1$ as an approximate value.

$$Score^R_{d_i} = \sum_{j=1}^{n} \frac{TF_{k_i}}{\Delta} \times \log \frac{N}{DF_{k_i}}$$

where $d_i$: the $i$th target document, $n$: the number of keywords in the query, $k_j$: the $j$th keyword in the query, $TF_{k_i}$: the number of times $k_j$ appears in $d_i$, $DF_{k_i}$: the number of documents where $k_j$ appears over all target documents, $N$: the total number of target documents, $Length_i$: the length of $d_i$, $\Delta$: the average length of the document over all target documents.

3 The fusion of the DRB method and the Robertson’s method

Combining the DRB method and the Robertson’s method, we calculate the score of document $d_i$ according to the following equation:

$$Score_{d_i} = \alpha \cdot Score^R_{d_i} + (1 - \alpha) \cdot Score^E_{d_i}$$

Where, $Score^R_{d_i}$ and $Score^E_{d_i}$ are the score of $d_i$ obtained by the Robertson’s method defined in Section 2.4 and that obtained by the DRB method, respectively. $\alpha$ is a weighting coefficient defined empirically.

$Score^E_{d_i}$ is defined as follows:

$$Score^E_{d_i} = \begin{cases} 
\sum_{f=1}^{F_N} \sum_{g=1}^{G_{FRM_f}} \frac{TF_{w_{(f,g)}}(d_i)}{\Delta} \times \log \frac{N}{DF_{w_{(f,g)}}} & (F_N \geq 1) \\
0 & (\text{otherwise})
\end{cases}$$

where $F_N$: the number of frames in a query, where they exist in $d_i$.

$FRM_f$: a frame that is in both the query and the target document

$G_{FRM_f}$: the number of keywords in $FRM_f$

$w_{(f,g)}$: $g$th keyword in $FRM_f$

$TF_{w_{(f,g)}}(d_i)$: the number of occurrences of $w_{(f,g)}$ in $d_i$

$DF_{w_{(f,g)}}$: the number of documents where $w_{(f,g)}$ appears over all target documents.

The other variables, $d_i$, $N$, $Length_i$, and $\Delta$, are the same as in the equation for the Robertson’s method defined in Section 2.4.

4 Experiments

4.1 Preliminary

Tools for Implementing the DRB Method

In order to implement the DRB method, we used QJP (Kameda, 1996) as a Japanese morphological and syntactic analyzer because QJP analyzes Japanese sentences very fast with heuristic rules. Since it unfortunately does not divide compound nouns, we used another Japanese morphological analyzer, Chasen (Matsumoto et al., 2001). After the compound noun is decomposed, it is transformed into a frame “noun A 'no' noun B” so that the dependency relationship between the nouns is clearly expressed.

The Test Set for Evaluation

In the experiments, we used BMIR-J2 (Kitani et al., 1998), which is a test collection for Japanese information retrieval systems. BMIR-J2 is based on the articles in the Mainichi Newspaper CD-ROM ’94 data collection (Mainichi-Newspaper, 1994), and contains 5080 articles or approximately 5M Bytes. BMIR-J2 provides 50 queries set$^1$. They are all divided into 5 categories, each of which requires one of the following functions (Kitani et al., 1998):

a) The basic function: every query consists of one word.

b) The numerical range function

c) The syntactic function

d) The semantic function

e) The world knowledge function

Category a) does not require dependency relationships between words, and the requirements of categories b) and c) are beyond the capacities of our current system. We consequently selected the queries of categories c) and d), of which there are 16. These 16 queries are shown in Table 1. Each of them has 6 to 50 relevant documents.

$^1$Although 10 additional queries are also provided, we did not use them because each of them has fewer than 4 relevant documents
Table 1: Queries requiring syntactic analysis and some knowledge of language for retrieval

<table>
<thead>
<tr>
<th>query number</th>
<th>Japanese query</th>
<th>the meaning in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>handotai seihin no seisan</td>
<td>the production of semiconductor products</td>
</tr>
<tr>
<td>(2)</td>
<td>denwa ryokin no nesage</td>
<td>a cut in telephone rates</td>
</tr>
<tr>
<td>(3)</td>
<td>seito ni taishuru kenkin</td>
<td>a donation to a political party</td>
</tr>
<tr>
<td>(4)</td>
<td>kokutoungun haken</td>
<td>dispatch of the UN forces</td>
</tr>
<tr>
<td>(5)</td>
<td>denki tsusin ni kansuru kisei kanwa</td>
<td>deregulation of electric communication</td>
</tr>
<tr>
<td>(6)</td>
<td>manshon no hanbai</td>
<td>sale of condominiums</td>
</tr>
<tr>
<td>(7)</td>
<td>chika no geraku</td>
<td>a fall in the value of land</td>
</tr>
<tr>
<td>(8)</td>
<td>kosokudoro no kensetsu</td>
<td>the construction of expressways</td>
</tr>
<tr>
<td>(9)</td>
<td>endaka ni yoru buka no teika</td>
<td>a fall in prices due to a strong yen</td>
</tr>
<tr>
<td>(10)</td>
<td>reka no higai</td>
<td>damage from a cool summer</td>
</tr>
<tr>
<td>(11)</td>
<td>meka no gen eki taisaku</td>
<td>a manufacture’s response to reduced profits</td>
</tr>
<tr>
<td>(12)</td>
<td>kabuka doko</td>
<td>a trend in stock prices</td>
</tr>
<tr>
<td>(13)</td>
<td>konpyutaseihin no sijodoko</td>
<td>a trend in the market for computer products</td>
</tr>
<tr>
<td>(14)</td>
<td>ginko no keiei keikaku</td>
<td>the management plan of a bank</td>
</tr>
<tr>
<td>(15)</td>
<td>yasuri wo okonai ryuisu gyosha</td>
<td>discount distributors</td>
</tr>
<tr>
<td>(16)</td>
<td>akaji kokusai no hakko</td>
<td>the issue of deficit-covering government bonds</td>
</tr>
</tbody>
</table>

4.2 Comparison of Information Retrieval Methods
The following four methods are compared:

**DRB method**: uses dependency relationships between words and semantic information about them when it retrieves documents on Frame match.

**Boolean AND matching**: retrieves every document which contains all keywords in a query.

**Robertson’s Method**: ranks documents according to the equation described in Section 2.4.

**Combined Method**: ranks documents according to the equation described in Section 3.

4.3 Evaluation Measures
As evaluation measures, Recall, Precision, Average Precision and Interpolated Recall-Precision are used.

**Recall, Precision**
Recall (REC for short) is the proportion of relevant material actually retrieved in answer to a search request and Precision (PRE for short) is the proportion of retrieved material that is actually relevant. Both are defined as follows:

\[
REC = \sum_{i \in Q} \frac{|A_i \cap B_i|}{|A_i|}, \quad PRE = \sum_{i \in Q} \frac{|A_i \cap B_i|}{|B_i|}
\]

- \(Q\): the set of queries
- \(A_i\): the set of documents relevant to the \(i\)th query
- \(B_i\): the set of documents retrieved for the \(i\)th query

\[
|\hat{A}| = \sum_{i \in Q} |A_i|, \quad |\hat{B}| = \sum_{i \in Q} |B_i|
\]

**Average Precision**
For a query, the average precision expresses the precision every time the relevant document is retrieved, and then takes their average. It is defined as follows:

\[
AveragePrecision = \sum_{j \in J_i} \frac{|A_i \cap B_{i,j}|}{|J_i|}
\]

- \(J_i\): the set of documents retrieved for the \(i\)th query
- \(B_{i,j}\): \(j\) documents retrieved for the \(i\)th query.

**Interpolated Recall-Precision**
The interpolated precision at a recall cutoff \(R\), denoted by \(P(R)\), is defined to be the maximum precision at all points \(\leq R\). \(P(R)\) over all queries is as follows:

\[
P_R = \frac{\sum_{i=1}^{\#Q} P_{R_i}}{\#Q}, \quad R = \{0.0, 0.1, 0.2, 0.3, \ldots, 1.0\}
\]

4.4 Experimental Results
Evaluation of Query Expansion and Zero Pronoun Resolution Methods
In order to evaluate the effectiveness of query expansion described in Section 2.2 and zero pronoun resolution described in Section 2.3, we compared the DRB method and simple Boolean AND matching method (AND matching for short). Table 2 shows the results with/without query expansion and with/without zero pronoun resolution. To deal with the retrieved results coordinately, we use the micro evaluation method. From the result that the precision rate of the DRB method
Table 2: Comparison between DRB method and Boolean AND matching method. QE:Query Expansion, ZPR:Zero Pronoun Resolution, REC:Recall, PRE:Precision

<table>
<thead>
<tr>
<th></th>
<th>No QE, No ZPR</th>
<th>No QE, ZPR</th>
<th>QE, ZPR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>REC</td>
<td>PRE</td>
<td>REC</td>
</tr>
<tr>
<td>DRB</td>
<td>13.4% (45/336)</td>
<td>78.9% (45/57)</td>
<td>16.4% (55/336)</td>
</tr>
<tr>
<td>AND</td>
<td>27.4% (92/336)</td>
<td>59.0% (92/156)</td>
<td>27.4% (92/356)</td>
</tr>
</tbody>
</table>

When applied with query expansion but without zero pronoun resolution, the DRB method slightly improves the recall rate (by 2.4%), although it also slightly reduces the precision rate (by 2.1%).

When applied without query expansion but with zero pronoun resolution, the DRB method improves both the recall and precision rate, although the precision rate improves only 0.8%.

When both query expansion and zero pronoun resolution are carried out, the precision rate of the DRB method is 39.3% higher than AND matching. The rate, however, is approximately 5.0% lower than that of the DRB method without query expansion or zero pronoun resolution. This is because the influence of query expansion for the wrong meaning of a keyword in a query is amplified by zero pronoun resolution.

From these results, we can see two things: 1) query expansion with the EDR conceptual dictionary improves the recall rate, but makes the precision rate worse, and 2) zero pronoun resolution improves both the recall rate and the precision rate. Furthermore, we deal with the F-measure, which is calculated by 2*PR*RE/(PR+RE) (PR:Precision, RE:Recall).

In the case of the DRB method, the F-measure is increased from 27.2%(No QE, ZPR) to 30.6%(QE, ZPR) because the constraint of the DRB method by Frame matching is effective and it makes the precision rate only 6% worse. On the other hand, the F-measure for AND matching decreases from 37.4%(No QE, ZPR) to 35.7%(QE, ZPR). From these results, we can say that both query expansion and zero pronoun resolution are effective constraints on the DRB method.

Average Precision of Each Query

Figure 1 depicts the average precision for each method applied to each query. All methods include query expansion as described in Section 2.2 and zero pronoun detection as described in Section 2.3.

4.5 Discussion

Although the DRB method achieves quite high average precision rates for some queries, it unfortunately returns no result for 5 queries. This is because the DRB method utilizes every frame in a query and retrieves only the documents including all frames in the query. This constraint is too severe as shown in Table 2. By constraining the DRB method, however, the combined method improves on the Robertson's method for queries (3), (5), (6), (7), (8) and (9) and (16), although for queries (4), (12), (14), (15), the opposite was true. This is because the current combined method makes use of any frame in a query even though the frame utilized does not always have strong relevance to the whole meaning of the query.

Averaged Interpolated Recall-Precision

Figure 2 shows that the combined method achieves the best precision rate at each recall rate compared to applying the Robertson’s method and the DRB method separately.
very poor (approximately 19%), and often returns no result. By combining the method with the Robertson’s method, the precision rates were improved compared to applying the DRB method and the Robertson’s method individually.

For some queries, the average precision rates of the combined method were inferior to those of the Robertson’s method. This is because irrelevant frames were yielded after expanding queries such as query (4), (14), (15), and irrelevant documents were given an excessive score by the frames. In the case of query (12), the DRB method retrieves 3 relevant documents and 1 irrelevant document. Since the 3 relevant documents are also ranked from the 1st to 3rd by the Robertson’s method, those documents do not contribute to increasing the precision rate. On the contrary, although the irrelevant document is given only a low score by the Robertson’s method, it decreases the precision rate because the additional score by the DRB method moves the document to a higher rank.

5 Conclusion

This paper discussed the combined method, which is a fusion of the DRB method and the Robertson’s method. The experimental results showed that the combined method was mostly superior to the Robertson’s method and to the DRB method. In terms of the average precision of each query, our method was, for some queries, inferior to the Robertson’s method because our method made use of any frame in a query to give a frame-match score to a document including the frame even when the frame was unfortunately not actually relevant to a query. We are accordingly investigating a method which handles only useful frames in a query to find documents relevant to it. We hope the results will be reported soon.

References


