# Query Structuring with Two-stage Term Dependence in the Japanese Language

Koji Eguchi[1] and W. Bruce Croft[2]

[1] National Institute of Informatics, Tokyo 101-8430, Japan,
eguchi@nii.ac.jp
[2] University of Massachusetts, Amherst, MA 01003-9264, USA,
croft@cs.umass.edu

**Abstract.** We investigate the effectiveness of query structuring in the Japanese language by composing or decomposing compound words and phrases. Our method is based on a theoretical framework using Markov random fields. Our two-stage term dependence model captures both the global dependencies between query components explicitly delimited by separators in a query, and the local dependencies between constituents within a compound word when the compound word appears in a query component. We show that our model works well, particularly when using query structuring with compound words, through experiments using a 100-gigabyte web document collection mostly written in Japanese.

## 1 Introduction

Japanese text retrieval is required to handle several types of problems specific to the Japanese language, such as compound words and segmentation [1]. To treat these problems, word-based indexing is typically achieved by applying a morphological analyzer, and character-based indexing has also been investigated. Some researchers compared this kind of character-based indexing with word-based indexing, and found little difference between them in retrieval effectiveness (e.g., [1–3]). Some other researchers made use of supplemental phrase-based indexing in addition to word-based indexing for English (e.g., [4]) or for Japanese (e.g., [5]). However, we believe this kind of approaches is not appropriate for the languages, for instance Japanese, in which individual words are frequently composed into a long compound word and the formation of an endless variety of compound words is allowed.

Meanwhile, the structured query approach has been used to include more meaningful phrases in a proximity search query to improve retrieval effectiveness [6, 7]. A few researchers have investigated this approach to retrieval for Japanese newspaper articles [1, 3]; however, they emphasized formulating a query using n-grams and showed that this approach performed comparably in retrieval effectiveness with the word-based approach. We are not aware of any studies that have used structured queries to formulate queries reflecting Japanese compound words or phrases appropriately. Phrase-based queries are known to perform effectively, especially against large-scale and noisy text data such as typically appear

on the Web [8, 7]. Again, we have not seen any studies that used structured queries to effectively retrieve web documents written in Japanese.

In this paper, we use the structured query approach using word-based units to capture, in a query, compound words and more general phrases of the Japanese language. Our approach is based on a theoretical framework using Markov random fields [7]. We experiment using a large-scale web document collection mostly written in Japanese.

## 2 Retrieval Model and Term Dependence Model

'Indri' is a search engine platform that can handle large-scale document collections efficiently and effectively [9]. The retrieval model implemented in Indri combines the language modeling [10] and inference network [11] approaches to information retrieval. This model allows structured queries similar to those used in 'InQuery' [11] to be evaluated using language modeling estimates within the network. Because we focus on query formulation rather than retrieval models, we use Indri as a baseline platform for our experiments. We omit further details of Indri because of space limitations. See [9] for the details.

Metzler and Croft developed a general, formal framework for modeling term dependencies via Markov random fields (MRFs) [7], and showed that the model is very effective in a variety of retrieval situation using the Indri platform. MRFs are commonly used in statistical machine learning to model joint distributions succinctly. In [7], the joint distribution $P_\Lambda(Q, D)$ over queries $Q$ and documents $D$, parameterized by $\Lambda$, was modeled using MRFs, and for ranking purposes the posterior $P_\Lambda(D|Q)$ was derived by the following ranking function, assuming a graph $G$ that consists of a document node and query term nodes:

$$P_\Lambda(D|Q) \stackrel{rank}{=} \sum_{c \in C(G)} \lambda_c f(c) \tag{1}$$

where $Q = t_1 ... t_n$, $C(G)$ is the set of cliques in an MRF graph $G$, $f(c)$ is some real-valued feature function over clique values, and $\lambda_c$ is the weight given to that particular feature function.

*Full independence* ('fi'), *sequential dependence* ('sd'), and *full dependence* ('fd') are assumed as three variants of the MRF model. The full independence variant makes the assumption that query terms are independent of each other. The sequential dependence variant assumes dependence between neighboring query terms, while the full dependence variant assumes that all query terms are in some way dependent on each other. To express these assumptions, the following specific ranking function was derived:

$$P_\Lambda(D|Q) \stackrel{rank}{=} \sum_{c \in T} \lambda_T f_T(c) + \sum_{c \in O} \lambda_O f_O(c) + \sum_{c \in O \cup U} \lambda_U f_U(c) \tag{2}$$

where $T$ is defined as the set of 2-cliques involving a query term and a document $D$, $O$ is the set of cliques containing the document node and two or more

query terms that appear contiguously within the query, and $U$ is the set of cliques containing the document node and two or more query terms appearing noncontiguously within the query.

## 3 Query Structuring with Two-stage Term Dependence

In compound words that often appear for instance in Japanese, the dependencies of each constituent word are tighter than in more general phrases. Therefore, we consider that these term dependencies should be treated as global between query components that make up a whole query and as local within a compound word when the compound word appears in a query component. Metzler and Croft's term dependence model, which we summarized in the previous section, gives a theoretical framework for this study, but must be enhanced when we consider more complex dependencies as mentioned above. In this paper, we propose *two-stage term dependence model* that captures term dependencies both between query components in a query and between constituents within a compound word.

To achieve the model mentioned above, we extend the term dependence model given in Eq. (2), on the basis of Eq. (1), as follows:

$$P_\Lambda(D|Q) \stackrel{rank}{=} \sum_{c_q \in T(Q)} \lambda_T f_T(c_q) + \sum_{c_q \in O(Q)} \lambda_O f_O(c_q) + \sum_{c_q \in O(Q) \cup U(Q)} \lambda_U f_U(c_q) \tag{3}$$

where

$$\begin{aligned} f_T(c_q) &= f'_T\Big( \sum_{q_k \in c_q} \sum_{c_t \in T(q_k)} \mu_T g_T(c_t) \Big) \\ f_O(c_q) &= f'_O\Big( \sum_{q_k \in c_q} \sum_{c_t \in O(q_k)} \mu_O g_O(c_t) \Big) \\ f_U(c_q) &= f'_U\Big( \sum_{q_k \in c_q} \sum_{c_t \in O(q_k) \cup U(q_k)} \mu_U g_U(c_t) \Big) \quad . \end{aligned} \tag{4}$$

Here, $Q$ consists of query components $q_1 \cdots q_k \cdots q_m$, and each query component consists of individual terms $t_1 \cdots t_n$. $T(Q)$, $O(Q)$ and $U(Q)$ can be defined in the same manner as in Eq. (2) with the query components consisting of a whole query, while $T(q_k)$, $O(q_k)$ and $U(q_k)$ are defined with the individual terms consisting of a query component. The feature functions $f'_T$, $f'_O$ and $f'_U$ and another feature functions $g_T$, $g_O$ and $g_U$ can be given in the same manner as $f_T$, $f_O$ and $f_U$ that were defined in Section **2**, respectively. Hereafter, we assumed that the constraint $\lambda_T + \lambda_O + \lambda_U = 1$ was imposed independently of the query, and assumed $\mu_T = \mu_O = \mu_U = 1$ for simplicity. When $Q$ consists of two or more query components and each of which has one term, Eq. (3) is equivalent to Eq. (2). The model given by Eq. (2) can be referred to as the *single-stage term dependence model*. When $f'_T(x) = f'_O(x) = f'_U(x) = x$ for any $x$, Eq. (3) represents dependencies only between constituent terms within each query component, which can be referred to as the *local term dependence model*; otherwise, Eq. (3) expresses the *two-stage term dependence model*.

According to Eq. (3), we assumed the following instances, considering special features of the Japanese language [12].

**Two-stage term dependence models**

(1) *glsd+* expresses the dependencies on the basis of the sequential dependence both between query components and between constituent terms within a query component, assuming dependence between neighboring elements. The beliefs (scores) for the resulting feature terms/phrases for each of $f_T$, $f_O$ and $f_U$ are combined as in Eq. (3).

(2) *glfd+* expresses the dependencies between query components on the basis of the full dependence, assuming all the query components are in some way dependent on each other. It expresses the dependencies between constituent terms within a query component on the basis of the sequential dependence.

Here in $f_T$, $f_O$ and $f_U$, each compound word containing prefix/suffix word(s) is represented as an exact phrase and treated the same as the other words, on the basis of the empirical results reported in [12]. Let us take an example from the NTCIR-3 WEB topic set [13], which is written in Japanese. The title field of Topic 0015 was described as three query components, "オゾン層　オゾンホール　人体" (which mean 'ozone layer', 'ozone hole' and 'human body'). A morphological analyzer converted this to "オゾン" ('ozone' as a general noun) and "層" ('layer' as a suffix noun), "オゾン" ('ozone' as a general noun) and "ホール" ('hole' as a general noun), and "人体" ('human body' as a general noun). The following are Indri query expressions corresponding to Topic 0015, according to the glsd+ and glfd+ models, respectively:

#weight( $\lambda_T$ #combine( #1( オゾン 層 ) オゾン ホール 人体 )
$\lambda_O$ #combine( #1( オゾン 層 ) #od2( オゾン ホール ) 人体 )
$\lambda_U$ #combine( #uw$N_4$( #1( オゾン 層 ) オゾン ホール )
#uw$N_3$( オゾン ホール 人体 ) ) )

#weight( $\lambda_T$ #combine( #1( オゾン 層 ) オゾン ホール 人体 )
$\lambda_O$ #combine( #1( オゾン 層 ) #od2( オゾン ホール ) 人体 )
$\lambda_U$ #combine( #uw$N_4$( #1( オゾン 層 ) オゾン ホール )
#uw$N_3$( オゾン ホール 人体 )
#uw$N_3$( #1( オゾン 層 ) 人体 )
#uw$N_5$( #1( オゾン 層 ) オゾン ホール 人体 ) ) )

where #1$(\cdot)$ indicates exact phrase expressions; #od$M(\cdot)$ indicates phrase expressions in which the terms appear ordered, with at most $M-1$ terms between each; and #uw$N_\ell(\cdot)$ indicates phrase expressions in which the specified terms appear unordered within a window of $N_\ell$ terms. $N_\ell$ is given by $(N_1 \times \ell)$ when $\ell$ terms appear in the window.

**Local term dependence models**

(3) *lsd+* indicates the glsd+ model with $f'_T(x) = f'_O(x) = f'_U(x) = x$ for any $x$, ignoring the dependencies between query components.

(4) *lfd+* indicates the glfd+ model with $f'_T(x) = f'_O(x) = f'_U(x) = x$ for any $x$, ignoring the dependencies between query components.

The following is an example query expression of 'lsd+' on Topic 0015.

#weight( $\lambda_T$ #combine( #1(オゾン 層) オゾン ホール 人体 )
$\lambda_O$ #combine( #1(オゾン 層) #od2(オゾン ホール) 人体 )
$\lambda_U$ #combine( #1(オゾン 層) #uw$N_2$(オゾン ホール) 人体 ) )

# 4 Experiments

## 4.1 Data and Experimental Setup

For experiments, we used a 100-gigabyte web document collection 'NW100G-01', which was used for the NTCIR-3 Web Retrieval Task ('NTCIR-3 WEB') [13] and for the NTCIR-5 WEB Task ('NTCIR-5 WEB') [14]. We used the topics and relevance judgment data of the NTCIR-3 WEB for training the model parameters[3]. We used the data set that was used in the NTCIR-5 WEB for testing[4]. All the topics were written in Japanese. The title field of each topic gives 1–3 query components that were suggested by the topic creator.

We used the texts that were extracted from and bundled with the NW100G-01 document collection. In these texts, all the HTML tags, comments, and explicitly declared scripts were removed. We segmented each document into words using the morphological analyzer 'MeCab version 0.81'[5]. We did not use the part-of-speech (POS) tagging function of the morphological analyzer for the documents, because the POS tagging function requires more time. We used Indri to make an index of the web documents in the NW100G-01 using these segmented texts. We used several types of stopwords in the querying phase, on the basis of the empirical results reported in [12].

In the experiments described in the following sections, we only used the terms specified in the title field. We performed morphological analysis using the 'MeCab' tool described above to segment each of the query component terms delimited by commas, and to add POS tags. Here, the POS tags are used to specify prefix and suffix words that appear in a query because, in the query structuring process, we make a distinction between compound words containing prefix and suffix words and other compound words, as described in Section **3**.

## 4.2 Experiments for Training

Using the NTCIR-3 WEB test collection, we optimized each of the models defined in Section **3**, changing each weight of $\lambda_T$, $\lambda_O$ and $\lambda_U$ from 0 to 1 in steps of 0.1, and changing the window size $N$ for the unordered phrase feature as 2, 4, 8, 50 or $\infty$ times the number of words specified in the phrase

[3] For the training, we used the relevance judgment data based on the *page-unit document model* [13] included in the NTCIR-3 WEB test collection.

[4] We used the data set used for the *Query Term Expansion Subtask*. The topics were a subset of those created for the NTCIR-4 WEB [15]. The relevance judgments were additionally performed by extension of the relevance data of the NTCIR-4 WEB. The objectives of this paper are different from those of that task; however, the data set is suitable for our experiments.

[5] ⟨http://www.chasen.org/∼taku/software/mecab/src/mecab-0.81.tar.gz⟩.

**Table 1.** Optimization results using a training data set.

| | $\text{AvgPrec}_a$ | %increase | $\text{AvgPrec}_c$ | %increase |
|---|---|---|---|---|
| base | 0.1543 | 0.0000 | 0.1584 | 0.0000 |
| lsd+ | 0.1624 | 5.2319 | 0.1749 | 10.4111 |
| lfd+ | 0.1619 | 4.9120 | 0.1739 | 9.7744 |
| glsd+ | 0.1640 | 6.2731 | 0.1776 | 12.0740 |
| glfd+ | 0.1626 | 5.4140 | 0.1769 | 11.6788 |
| naive-sd | 0.1488 | -3.5551 | 0.1472 | -7.0743 |
| naive-fd | 0.1488 | -3.5427 | 0.1473 | -7.0496 |
| ntcir-3 | 0.1506 | -2.3774 | 0.1371 | -13.4680 |

expression. Additionally, we used $(\lambda_T, \lambda_O, \lambda_U) = (0.9, 0.05, 0.05)$ for each $N$ value above. Note that stopword removal was only applied to the term feature $f_T$, not to the ordered/unordered phrase features $f_O$ or $f_U$. The results of the optimization that maximized the mean average precision over all 47 topics ('$\text{AvgPrec}_a$') are shown in **Table 1**. This table includes the mean average precision over 23 topics that contain compound words in the title field as '$\text{AvgPrec}_c$'. '%increase' was calculated on the basis of 'base', the result of retrieval not using query structuring. After optimization, the 'glsd+' model worked best when $(\lambda_T, \lambda_O, \lambda_U, N) = (0.9, 0.05, 0.05, \infty)$, while the 'glfd+' model worked best when $(\lambda_T, \lambda_O, \lambda_U, N) = (0.9, 0.05, 0.05, 50)$.

For comparison, we naively applied the single-stage term dependence model using either the sequential dependence or the full dependence variants defined in Section **2** to each of the query components delimited by commas in the title field of a topic, and combined the beliefs (scores) about the resulting structure expressions. We show the results of these as 'naive-sd' and 'naive-fd', respectively, in **Table 1**. These results suggest that Metzler and Croft's single-stage term dependence model must be enhanced to handle the more complex dependencies that appear in queries in the Japanese language. For reference, we also show the best results from NTCIR-3 WEB participation [13] ('ntcir-3') at the bottom of **Table 1**. This shows that even our baseline system worked well.

### 4.3 Experiments for Testing

For testing, we used the models optimized in the previous subsection. We used the relevance judgment data, for evaluation, that were provided by the organizers of the NTCIR-5 WEB task. The results are shown in **Table 2**. In this table, '$\text{AvgPrec}_a$', '$\text{AvgPrec}_c$' and '$\text{AvgPrec}_o$' indicate the mean average precisions over all 35 topics, over the 22 topics that include compound words in the title field, and over the 13 topics that do not include the compound words, respectively. '%increase' was calculated on the basis of the result of retrieval not using query structuring ('base'). The results show that our two-stage term dependence models, especially the 'glfd+' model, gave 13% better performance than the baseline ('base'), which did not assume term dependence, and also better than the local term dependence models, 'lsd+' and 'lfd+'. The advantage of 'glfd+' over 'base',

**Table 2.** Test results of phrase-based query structuring.

| | $\text{AvgPrec}_a$ | %increase | $\text{AvgPrec}_c$ | %increase | $\text{AvgPrec}_o$ | %increase |
|---|---|---|---|---|---|---|
| base | 0.1405 | 0.0000 | 0.1141 | 0.0000 | 0.1852 | 0.0000 |
| lsd+ | 0.1521 | 8.2979 | 0.1326 | 16.2563 | 0.1852 | 0.0000 |
| lfd+ | 0.1521 | 8.2389 | 0.1325 | 16.1407 | 0.1852 | 0.0000 |
| glsd+ | 0.1503 | 6.9576 | 0.1313 | 15.1167 | 0.1823 | -1.5496 |
| glfd+ | 0.1588 * | 13.0204 | 0.1400 | 22.6950 | 0.1906 | 2.9330 |

'*' indicates statistical significant improvement over 'base', 'lsd+', 'lfd+' and 'glsd+' where $p < 0.05$ with two-sided Wilcoxon signed-rank test.

'lsd+', 'lfd+' and 'glsd+' was statistically significant in average precision over all the topics. The results of '$\text{AvgPrec}_c$' and '$\text{AvgPrec}_o$' imply that our models work more effectively for queries expressed in compound words.

## 5 Conclusions

In this paper, we proposed the two-stage term dependence model, which was based on a theoretical framework using Markov random fields. Our two-stage term dependence model captures both the global dependence between query components explicitly delimited by separators in a query, and the local dependence between constituents within a compound word when the compound word appears in a query component. We found that our two-stage term dependence model worked significantly better than the baseline that did not assume term dependence at all, and better than using models that only assumed either global dependence or local dependence in the query. Our model is based on proximity search, which is typically supported by Indri [9] or InQuery [11].

We believe that our work is the first attempt to explicitly capture both long-range and short-range term dependencies. The two-stage term dependence model should be reasonable for other languages, if compound words or phrases can be specified in a query. Application to natural language-based queries, employing an automatic phrase detection technique, is worth pursuing as future work.

## Acknowledgments

We thank Donald Metzler for valuable discussions and comments, and David Fisher for helpful technical assistance with Indri. This work was supported in part by the Overseas Research Scholars Program and the Grants-in-Aid for Scientific Research (#17680011 and #18650057) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, in part by the Telecommunications Advancement Foundation, Japan, and in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

## References


1. Fujii, H., Croft, W.B.: A comparison of indexing techniques for Japanese text retrieval. In: Proceedings of the 16th Annual International ACM SIGIR Conference, Pittsburgh, Pennsylvania, USA (1993) 237–246
2. Chen, A., Gey, F.C.: Experiments on cross-language and patent retrieval at NTCIR-3 Workshop. In: Proceedings of the 3rd NTCIR Workshop, Tokyo, Japan (2002)
3. Moulinier, I., Molina-Salgado, H., Jackson, P.: Thomson Legal and Regulatory at NTCIR-3: Japanese, Chinese and English retrieval experiments. In: Proceedings of the 3rd NTCIR Workshop, Tokyo, Japan (2002)
4. Mitra, M., Buckley, C., Singhal, A., Cardie, C.: An analysis of statistical and syntactic phrases. In: Proceedings of RIAO-97. (1997) 200–214
5. Fujita, S.: Notes on phrasal indexing: JSCB evaluation experiments at NTCIR ad hoc. In: Proceedings of the First NTCIR Workshop, Tokyo, Japan (1999) 101–108
6. Croft, W.B., Turtle, H.R., Lewis, D.D.: The use of phrases and structured queries in information retrieval. In: Proceedings of the 14th Annual International ACM SIGIR Conference, Chicago, Illinois, USA (1991) 32–45
7. Metzler, D., Croft, W.B.: A markov random field model for term dependencies. In: Proceedings of the 28th Annual International ACM SIGIR Conference, Salvador, Brazil (2005) 472–479
8. Mishne, G., de Rijke, M.: Boosting web retrieval through query operations. In: Proceedings of the 27th European Conference on Information Retrieval Research, Santiago de Compostela, Spain (2005) 502–516
9. Metzler, D., Croft, W.B.: Combining the language model and inference network approaches to retrieval. Information Processing and Management **40**(5) (2004) 735–750
10. Croft, W.B., Lafferty, J., eds.: Language Modeling for Information Retrieval. Kluwer Academic Publishers (2003)
11. Turtle, H.R., Croft, W.B.: Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems **9**(3) (1991) 187–222
12. Eguchi, K.: NTCIR-5 query expansion experiments using term dependence models. In: Proceedings of the 5th NTCIR Workshop, Tokyo, Japan (2005)
13. Eguchi, K., Oyama, K., Ishida, E., Kando, N., Kuriyama, K.: Overview of the Web Retrieval Task at the Third NTCIR Workshop. In: Proceedings of the 3rd NTCIR Workshop, Tokyo, Japan (2003)
14. Yoshioka, M.: Overview of the NTCIR-5 WEB Query Expansion Task. In: Proceedings of the 5th NTCIR Workshop, Tokyo, Japan (2005)
15. Eguchi, K., Oyama, K., Aizawa, A., Ishikawa, H.: Overview of the Informational Retrieval Task at NTCIR-4 WEB. In: Proceedings of the 4th NTCIR Workshop, Tokyo, Japan (2004)