社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＩＣＳ－133　(5)
2003／9／14

A MEDLINE document search system using section information

Takahiro YAMASAKI†, Masashi SHIMBO†, and Yuji MATSUMOTO†

† Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
E-mail: †{takah-ya,shimbo,matsu}@is.aist-nara.ac.jp

Abstract We present an experimental text retrieval system to facilitate search in the MEDLINE database. A unique feature of the system is that users can search not only through the whole abstract text for his specified keywords, but also from limited sections in the abstracts. The sections reflect the structure of abstract texts, such as BACKGROUND and CONCLUSIONS. This feature makes it easier to narrow down search results when adding extra keywords does not work, and allows for ranking search results according to users' needs. The MEDLINE database contains a small portion of "structured" abstracts, in which sections are explicitly marked by the headings. They thus require no additional processing for inferring which section each sentence belongs to. They also provide training data for constructing classifiers that section the rest of the unstructured abstracts lacking explicit section heading, which form a majority of the MEDLINE corpus.

Key words MEDLINE database, structured abstracts, information retrieval, text classification.

文の役割を考慮したMEDLINE文書検索システム

山崎　貴宏†　　新保　　仁†　　松本　裕治†

† 奈良先端科学技術大学院大学情報科学研究科
〒630-0192 奈良県生駒市高山町 8916-5
E-mail: †{takah-ya,shimbo,matsu}@is.aist-nara.ac.jp

あらまし 本稿では, 試作したMEDLINEアブストラクト検索システムについて報告する. 本システムの特徴は, 指定した検索キーワードをアブストラクト文全体から検索できるのはもとより, 検索対象をアブストラクトの一部分に限定することが可能という点にある. 検索対象部分は, 研究背景, 実験方法, 結論 といった, アブストラクト中の文章構造上の役割によって指定する. この機能によって, 単なる検索語の追加では不可能な, 効率的な検索結果の絞り込みが可能になる. MEDLINEには, 一部 "構造化 (structured)" アブストラクトと呼ばれる, 各段落の先頭にその段落の役割 (BACKGROUND, CONCLUSIONS, 等) が明記されたものが含まれているものの, 大多数にはそのような役割ラベルは明記されていない. これら非構造化アブストラクトの各文に対して自動的に役割ラベルを付与するため, 構造化アブストラクト内の各文を用いて訓練したラベル判別器を構築した.

キーワード MEDLINEデータベース, 構造化アブストラクト, 情報検索, 文書分類

1. Introduction

With the rapid increase in the volume of scientific literature, demands are growing for systems with which the researcher can efficiently search relevant scientific documents with less effort. Online document retrieval services such as NLM PubMed [8] and NEC CiteSeer [5] are getting more and more popular, as they permit users to search in the large corpora of abstract text or full papers.

PubMed enables efficient retrieval of relevant papers with full-text search in the MEDLINE abstract [7]. It also provides a number of auxiliary ways to narrow down search results. For example, it is possible to restrict search to specific fields including titles and publication date. And most abstracted citations are given peer-reviewed annotation of topics and keywords chosen from controlled vocabulary, which can also be used to restrict search. All these facilities rely on information external to the content of the abstract text. In this report, by contrast, we explore the use of information that is inherent in the abstract text, to help efficient retrieval. Namely, we exploit the structure underlying the abstract texts.

Our search system allows search to be performed within restricted

portions of the texts, where 'portions' are determined in accordance with the structural role of constituent sentences in the text. We expect such a system to substantially reduce users' effort to narrow down an (overwhelming) amount of search results. Consider sentences in the abstracts were classified to their roles, say, Background and Objectives, Methods, Experimental Results, and Conclusions. The intention of users is closely related to some of these "sections," but not to the rest. For instance, if a clinician intends to find whether an effect of a chemical substance on a disease is known or not, she/he can direct the search engine to return the passages in which the names of the substance and the disease co-occur, but only in the sentences from the Results and the Conclusions sections. Such a restriction is not easily attainable by simply adding extra query terms. Furthermore, it is often not immediately evident to users what extra keywords are effective for narrowing down the results. Specifying sections to search in such a case should reduce users' effort to retrieve information they seek.

Labeling each sentence under human supervision is not a viable option due to the size of the MEDLINE corpus. We are hence forced to seek for a way to automate this process. In our previous work [11], [13], we have reported preliminary results concerning the use of text classification techniques for inferring the label of the sentences but with no particular application in mind. This paper reports the extension of the work with more focus on its application to the search system for MEDLINE.

The main topics of the paper are (1) how to construct training data used for training the sentence classifier without ascribing too much to human supervision, and (2) what classes, or sections, should be presented to users to which they can restrict search. This decision must be made on account of the trade-off between usability and accuracy of sentence classification. Another topic is (3) what types of features are effective for classification.

2. Statistics on the MEDLINE abstracts

A step in our construction of a MEDLINE search system, in particular, the classification of sentences into sections, relies on the "structured abstracts" found in the MEDLINE database. Since these abstracts have explicit sections, we use them as training data in constructing the sentence classifiers for the rest of the abstracts. However, we need to analyze the quality of the data, as it would affect the performance of the resulting classifiers. In the following, we describe the statistics on the structured abstracts as contained in MEDLINE, and their impact on the design of the system.

2.1 Structured abstracts

Since its proposal in 1987, a growing number of biological and medical journals have begun to adopt so-called "structured abstracts" [1]. These journals require authors to divide the abstract text into sections that reflect the structure of the text, such as BACKGROUND, OBJECTIVES, and CONCLUSIONS. The sectioning schemes are sometimes regulated by the journals, and sometimes left to the choice of the authors. As the sections in these structured abstracts are explicitly marked with a heading (usually written in all upper-case letters), this allows us to identify a heading as a category label for the sentences that follow. Unfortunately, the number of unstructured abstracts in the MEDLINE database far exceeds that of structured abstracts (Table 1). The frequencies of individual headings as well as sectioning schemes are shown in Tables 2 and 3, respectively.

2.2 Section headings = categories?

The fact that unstructured abstracts form a majority leads to the idea of automatically labeling each sentences when the abstracts are unstructured. This labeling process can be formulated as a text categorization task if we fix a set of sections (categories) into which each sentence should be classified. The problem remains what categories, or sections, must be presented to the users to specify the portion of the abstract texts to which search should be restricted. It is natural to choose the categories from the section headings occurring in the structured abstracts, as it will allow us to use those abstract texts to train the sentence classifiers. However, there are more than 6,000 distinct headings in MEDLINE 2002.

To maintain usability, the number of sections offered to the user must be kept as small as possible, but not too small to make the facility useless. But if we restrict the number of categories, then, how should a section in a structured abstract be treated if its heading does not match any of the categories presented to the users? If the selection were sensible, most sections translate into a selected category in a straightforward way, such as identifying "OBJECTIVES" and "PURPOSE" sections. But there are headings such as "BACKGROUND AND PURPOSES." If BACKGROUND and PURPOSES were two distinct categories presented to the user, we would have to determine which of these two classes each sentence in the section belongs to. Therefore, at least some of the sentences in the structured abstracts must go through the same labeling process we use for unstructured abstracts, namely, when sectioning does not

Table 1 Ratio of structured and unstructured abstracts in MEDLINE 2002.

	# of abstracts /	%
Structured	374,585 /	6.0%
Unstructured	5,912,271 /	94.0%
Total	11,299,108 /	100.0%

Table 2 Frequency of individual sections in the structured abstracts in MEDLINE 2002.

Sections	# of abstracts	# of sentences
CONCLUSION(S)	352,153	246,607
RESULTS	324,479	1,378,785
METHODS	209,910	540,415
BACKGROUND	120,877	264,589
OBJECTIVE	165,972	166,890
⋮	⋮	⋮
Total		2,597,286

Table 3 Frequency of sectioning schemes (# of abstracts). Percentages show the frequency relative to the total number of structured abstracts. Schemes marked with '*' and '†' are used for the experiment in Section 3.3.

Rank	# /	%	Section sequence
1	61,603 /	16.6%	BACKGROUND / METHOD(S) / RESULTS / CONCLUSION(S)
*2	54,997 /	14.7%	OBJECTIVE / METHOD(S) / RESULTS / CONCLUSION(S)
*3	25,008 /	6.6%	PURPOSE / METHOD(S) / RESULTS / CONCLUSION(S)
4	11,412 /	3.0%	PURPOSE / MATERIALS AND METHOD(S) / RESULTS / CONCLUSION(S)
†5	8,706 /	2.3%	BACKGROUND / OBJECTIVE / METHOD(S) / RESULTS / CONCLUSION(S)
6	8,321 /	2.2%	OBJECTIVE / STUDY DESIGN / RESULTS / CONCLUSION(S)
7	7,833 /	2.1%	BACKGROUND / METHOD(S) AND RESULTS / CONCLUSION(S)
*8	7,074 /	1.9%	AIM(S) / METHOD(S) / RESULTS / CONCLUSION(S)
9	6,095 /	1.6%	PURPOSE / PATIENTS AND METHOD(S) / RESULTS / CONCLUSION(S)
10	4,087 /	1.1%	BACKGROUND AND PURPOSE / METHOD(S) / RESULTS / CONCLUSION(S)
⋮	⋮	⋮	⋮
Total	374,585 /	100.0%	

coincide with the categories presented to the users.

Even when the section they belong to has a heading that seems straightforward to assign a class, there are cases in which we have to classify sentences in a structured abstract. The above mentioned OBJECTIVE (or PURPOSE) class is actually one such category that needs sentence-wise classification. Below, we will further analyze this case.

As Table 3 shows, the most frequent *sequences* of headings are (1) BACKGROUND, METHOD(S), RESULTS, and CONCLUSION(S), followed by (2) OBJECTIVE, METHOD(S), RESULTS, and CONCLUSION(S). Inspecting abstract texts that conform to formats (1) and (2), we found that in the BACKGROUND and OBJECTIVE sections, most of these texts actually contain the both the sentences describing the background, and those describing the research objectives.

We have verified this claim by computing Sibson's information radius (Jensen-Shannon divergence) [6] of the sentences in each sections. Information radius D_{JS} between two probability distributions $p(x)$ and $q(x)$ is defined as follows, using Kullback-Leibler divergence D_{KL}.

$$D_{JS}(p\|q) = \frac{1}{2}\left[D_{KL}\left(p\middle\|\frac{p+q}{2}\right)+D_{KL}\left(q\middle\|\frac{p+q}{2}\right)\right]$$
$$= \frac{1}{2}\left[\sum_x p(x)\log\frac{p(x)}{(p(x)+q(x))/2} + \sum_x q(x)\log\frac{q(x)}{(p(x)+q(x))/2}\right].$$

Hence, information radius is a measure of dissimilarity between distributions. It is symmetric in p and q, and is always well-defined, which are not always the case with D_{KL}.

Table 4 shows that the sentences under the BACKGROUND and OBJECTIVE sections have similar distributions of word bigrams as well as the combination of words and word bigrams. Also note the smaller divergence between these classes (bold faced figures), compared with those for the other class pairs. The implication is that these two headings are not reliable as separate category labels.

3. Classifier design

3.1 The number and the types of categories

In our previous work [11], [13], we used five categories, BACKGROUND, OBJECTIVE, METHOD(S), RESULTS, and CONCLUSION(S), based on the frequency of individual headings (Table 2). We believe this is still a reasonable choice considering the usability of the system and the ambiguity arising from limiting the number of classes. As we mentioned earlier, the BACKGROUND and the OBJECTIVE section headings are too unreliable to be taken as a category label for the sentences in the sections. But still, it is not acceptable to merge them as a single class. Since they are quite different in their structural roles, merging them would greatly affect the utility of the system.

3.2 Support Vector Machines and feature representation

Following our previous work, soft-margin Support Vector Machines (SVMs) [2], [12] were used as the classifier for each categories. We first construct SVM classifiers for each of the BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS classes, using the one-versus-rest configuration. Since SVM is a binary classifier while our task involves five classes, we combine the results of these classifiers as follows: the class i assigned to a given test example x is the one the one represented by the SVM whose value of $f_i(x)$ is the largest, where $f_i(x)$ is a decision function of SVM for the i-th class, i.e., the signed distance from the optimal hyperplane after the margin width is normalized to 1.

The basis of our feature representations is words and word bigrams. In the previous work [11], [13], we have used non-contiguous sequential word patterns as features. We use word bigrams here instead of sequential patterns due to practical reasons: the speed of the feature construction is prohibitive because of the

Table 4 Information radius between the sections.

(a) Word bigrams

Class	BACKGROUND	OBJECTIVE	METHODS	RESULTS	CONCLUSION(S)
BACKGROUND	0	**0.1809**	0.3064	0.3152	0.2023
OBJECTIVE	**0.1809**	0	0.2916	0.3256	0.2370
METHODS	0.3064	0.2916	0	0.2168	0.3201
RESULTS	0.3152	0.3256	0.2168	0	0.2703
CONCLUSIONS	0.2023	0.2370	0.3201	0.2703	0

(b) Word unigrams and bigrams

Class	BACKGROUND	OBJECTIVE	METHODS	RESULTS	CONCLUSION(S)
BACKGROUND	0	**0.1099**	0.2114	0.2171	0.1202
OBJECTIVE	**0.1099**	0	0.1965	0.2221	0.1465
METHODS	0.2114	0.1965	0	0.1397	0.2201
RESULTS	0.2171	0.2221	0.1397	0	0.1847
CONCLUSIONS	0.1202	0.1465	0.2201	0.1847	0

large number of documents to process.

3.3 Contextual information

Since we are interested in labeling a *series* of sentences, it is expected that incorporating contextual information into the feature set will improve classification performance. For example, it is unlikely that experimental results (RESULTS) are presented before the description of experimental design (METHODS). Thus, knowing that preceding sentences have been labeled as METHODS should condition the probability of the present sentence being classified as RESULTS section. And the sentences of the same class have high probability of appearing consecutively; after all, we would not expect the authors to interleave sentences describing experimental results (RESULTS) with those in CONCLUSIONS and OBJECTIVES classes.

Since it is not clear what kind of contextual information performs best, the following types of contextual representation were examined in an experiment (Section 4. 1).

(1) The class of the previous sentence.

(2) The classes of the previous two sentences.

(3) The class of the next sentence.

(4) The classes of the next two sentence.

(5) Relative location of the current sentence in the abstract text.

(6) The word features of the previous sentence.

(7) The word features of the next sentence.

(8) The word features of the previous and the next sentences.

(9) The class of the previous sentence and the length of the contiguous sentences having the same class.

4. Experiments

This section reports the results of preliminary experiments that we conducted to examine the performance of the classifiers used for labeling sentences.

4.1 Contextual information

In this experiment, structured abstracts from MEDLINE 2002 were used. The classes we considered (or, sections to which sentences are classified) are OBJECTIVE(S), METHOD(S), RESULT(S), and CONCLUSION(S). Note that this set does not coincide with the five classes we employed in the final system. According to Table 3, the section sequence consisting of these sections are only second after the sequence BACKGROUND / METHOD(S) / RESULT(S) / CONCLUSION(S). However, identifying the sentences with headings PURPOSE(S) and AIM(S) with those with OBJECTIVE(S) makes the corresponding sectioning scheme the most frequent.

Hence, we collected structured abstracts whose heading sequences matches the following patterns:

(1) OBJECTIVE(S) / METHOD(S) / RESULTS / CONCLUSION(S),

(2) PURPOSE(S) / METHOD(S) / RESULTS / CONCLUSION(S),

(3) AIM(S) / METHOD(S) / RESULTS / CONCLUSION(S).

We split each of these abstracts into sentences using UIUC Sentence Splitter [9], after removing all symbols and replacing every contiguous sequence of numbers with a single symbol '#'. After sentence splitting, we filtered out the abstracts that produced a sentence with less than three words, regarding it as a possible error in sentence splitting. This yielded a total of 82,936 abstracts.

To reduce the number of features, we only take into account word bigrams occurring in at least 0.05% of the sentences, which amounts to 9,078 bigrams. The number of (unigram) word features was 104,733.

We obtained 103,962 training examples (sentences) from 10,000 abstracts randomly sampled from the set of 82,936 structured abstracts described above, and 10,356 test examples (sentences) from 1,000 abstracts randomly sampled from the rest of the set.

The quadratic kernel is used with SVMs, and the optimal soft margin (or capacity) parameter C is sought for each of the SVMs using different context features. The results are listed in Table 5.

There were not much differences in the performance of contextual features as far as accuracy were measured on a per-sentence

Table 5 Performance of context features

	Features	Accuracy (%) sentence	abstract
(0)	No context features	83.6	25.0
(1)	The class of the previous sentence	88.9	48.9
(2)	The classes of the previous two sentences	89.9	50.6
(3)	The class of the next sentence	88.9	50.9
(4)	The classes of the next two sentences	89.3	51.2
(5)	Relative location of the current sentence	91.9	50.7
(6)	The word features of the previous sentence	87.3	37.5
(7)	The word features of the next sentence	88.1	39.0
(8)	The word features of the previous and the next sentences	89.7	46.4
(9)	The class of the previous sentence and the number sentences of the same class preceding the sentence	90.6	50.9

basis. All contextual features (1)–(9) obtained about 90% accuracy, which is an improvement of 4 to 8% over (0) when no context features were used. The performance on a per-abstract basis, in which a classification of an abstract is judged to be correct only if all the constituent sentences are correctly classified, was about 50% at maximum, which is 25% improvement. This maximum performance was obtained for features (3), (4), and (5).

4.2 Separating 'Objectives' from 'Background'

The analysis in the Section 2.2 suggest that it is unreliable to use the headings BACKGROUND and OBJECTIVE(S) as the label of the sentences in the sections, because the BACKGROUND section frequently contains sentences that should rather be classified as OBJECTIVES and vice versa. Yet, it is not acceptable to merge them as a single class, because they are quite different in their structural roles; doing so would severely impairs the utility of the system.

To resolve this situation, we construct an SVM classifier to distinguish between these classes again. To train this classifier, we use the sentences in the structured abstracts that contain both the BACKGROUND and the OBJECTIVES sections (such as in the scheme marked with a dagger in Table 3).

To assess the feasibility of this approach, we collected 11,898 abstracts that have both the BACKGROUND and the OBJECTIVE(S) headings. The texts in this collection were preprocessed in an identical manner as the previous subsection, and the number of sentences in the BACKGROUND and the OBJECTIVES sections from this collection was 34,761. The classification of individual sentences with SVMs exhibited an F1-score of 96.4 point (which factors into a precision of 95.6% and a recall of 97.2%), on average over 10-fold cross validation trials. The SVMs used quadratic kernels, and used the bag-of-words-and-word-bigrams features only. No context features were used.

5. The system

Using the feature set described in Section 3.2 as well as the context feature (5) of Section 3.3, we constructed five SVM classifiers for each of the five sections, BACKGROUND, OBJECTIVES, METHODS, RESULTS, and CONCLUSIONS. With these SVMs, we labeled the sentences in unstructured abstracts in MEDLINE 2003 whose publication year is 2001 and 2002. The same labeling process is applied to the sentences in structured abstracts as well, when their section headings do not match any of the five sections presented to users. We also classified each sentence in the BACKGROUND and OBJECTIVE (and equivalent) sections into one of the BACKGROUND and OBJECTIVE classes using the classifier of Section 4.2, when the structured abstract contains only one of them.

We implemented an experimental search system for these labeled data using eRuby on top of an Apache web server. The full-text retrieval engine Namazu was used as a back-end search engine. The screen shot for the service page is shown in Figure 1. The form on the page contains a field for entering query terms, a 'Go' button as well as radio buttons marked 'Any' and 'Select from' for choosing whether the keyword search should be performed on the whole abstract texts, or on limited sections. Plain keywords, phrases (specified by enclosing the phrase in braces), and boolean conjunction ('and'), disjunction ('or'), and negation ('not') are allowed for query field. If the user chooses 'Select from' button rather than 'Any,' the check boxes on its right are activated. These boxes corresponds to the five target sections, namely, 'Background,' 'Objectives,' 'Methods,' 'Results,' and 'Conclusions.'

Matching query terms found in the abstract text are highlighted in bold face letters, and the sections (either deduced from headings or from the content of the sentence with automatic classifier) are shown in different background colors.

6. Conclusions and future work

We have reported the first step towards construction of a search system for the MEDLINE database that allows the users to exploit the underlying structure of the abstract text. The implemented system, however, is only experimental, and surely needs more elaboration.

First of all, the adequacy of five sections presented to the user needs evaluation. In particular, OBJECTIVE and CONCLUSIONS are different as they each describes what has been sought and what is really achieved, respectively, but they are the same in the sense that they provides a summary of what the paper deals with. They are not about the details of experiments, and not about what is done elsewhere. Thus grouping them into one class might be sufficient for most users.

We plan to incorporate re-ranking procedure of label sequences based on the overall consistency of the sequences. By consistency,

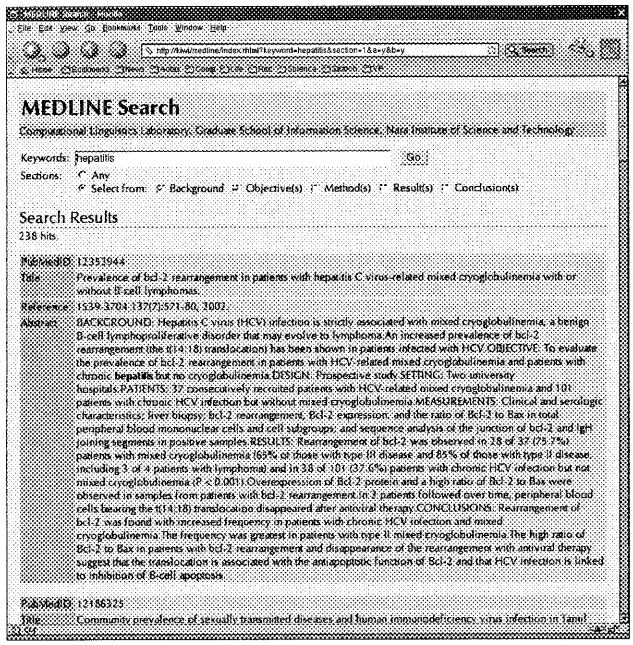

Fig. 1 A screen shot.

we mean the likelihood of sequences such as it is unlikely that conclusions appear in the beginning of the text, and the same section seldom occur twice in a text. The similar lines of research [4], [10] have been reported recently in ML and NLP communities, in which the sequence of classification results is optimized over all possible sequences. We also plan to incorporate features that reflects cohesion or coherence between sentences [3].

Acknowledgment

This research was supported in part by MEXT under Grant-in-Aid for Scientific Research on Priority Areas (B) no. 759. The second author is also supported in part by MEXT under Grant-in-Aid for Young Scientists (B) no. 15700098.

References

[1] Ad Hoc Working Group for Critical Appraisal of Medical Literature. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine*, 106(4):598–604, 1987.

[2] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[3] M. A. K. Halliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.

[4] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, pages 282–289. Morgan Kaufmann, 2001.

[5] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.

[6] L. Lee. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Comutational Linguistics (ACL-99)*, pages 25–32, 1999.

[7] MEDLINE. http://www.nlm.nih.gov/databases/databases_medline.html, 2002–2003. U.S. National Library of Medicine.

[8] PubMed. http://www.ncbi.nlm.nih.gov/PubMed/, 2003. U.S. National Library of Medicine.

[9] Sentence splitter software. http://l2r.cs.uiuc.edu/~cogcomp/cc-software.htm, 2001. University of Illinois at Urbana-Champaign.

[10] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of the Human Language Technology Conference North American Chapter of Association for Computational Linguistics (HLT-NAACL 2003)*, pages 213–220, Edmonton, Alberta, Canada, 2003. Association for Computational Linguistics.

[11] M. Shimbo, T. Yamasaki, and Y. Matsumoto. Automatic classification of sentences in the MEDLINE abstracts. In *Proceedings of the 6th Sanken (ISIR) International Symposium*, pages 135–138, Suita, Osaka, Japan, 2003.

[12] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[13] T. Yamasaki, M. Shimbo, and Y. Matsumoto. Automatic classification of sentences using sequential patterns. Technical Report of IEICE AI2002-83, The Institute of Electronics, Information and Communication Engineers, 2003. In Japanese.