Identification of “comment-on sentences” in online biomedical documents using support vector machines

In Cheol Kim*, Daniel X. Le, and George R. Thoma
Lister Hill National Center for Biomedical Communications
National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894

ABSTRACT

MEDLINE® is the premier bibliographic online database of the National Library of Medicine, containing approximately 14 million citations and abstracts from over 4,800 biomedical journals. This paper presents an automated method based on support vector machines to identify a “comment-on” list, which is a field in a MEDLINE citation denoting previously published articles commented on by a given article. For comparative study, we also introduce another method based on scoring functions that estimate the significance of each sentence in a given article. Preliminary experiments conducted on HTML-formatted online biomedical documents collected from 24 different journal titles show that the support vector machine with polynomial kernel function performs best in terms of recall and F-measure rates.

Keywords: “Comment-on” identification, Online biomedical documents, Support vector machine, Score function

1. INTRODUCTION

The Lister Hill National Center for Biomedical Communications (LHNCBC), a research and development division of the National Library of Medicine (NLM) has developed an automated system, the *Web-based Medical Article Records System* (WebMARS) that analyzes and extracts bibliographic information from online biomedical journal articles to create citations for MEDLINE [1][2], NLM’s premier bibliographic online database containing approximately 14 million citations and abstracts from over 4,800 biomedical journals. Since the online biomedical literature is continually growing, there is a strong motivation to develop automated systems like WebMARS to minimize human labor to provide bibliographic data in a timely fashion.

“Comment-on” is a field in a MEDLINE citation showing previously published articles commented on by a given article. In this paper, we call an article containing a “comment-on” list a “comment-in” article, while an article that is commented on is referred to as a “comment-on” article. Currently, the “comment-on” list is created manually, based on certain linguistic clues and contextual patterns. This is time-consuming and performance mainly depends on operators’ linguistic knowledge and understanding of scientific expressions and writing styles. In order to minimize this manual effort and to improve accuracy and processing speed, we propose an automated identification method based on support vector machines (SVMs). This proposed method will be incorporated into the WebMARS system as one of its major component modules.

Generally, authors of a “comment-in” article cite “comment-on” articles related to their research as primary external sources on which they may express complimentary or contradictory opinions. Thus the full bibliographical descriptions for these “comment-on” articles can usually be found in the reference section of a “comment-in” article. Note that in the scientific literature, all external sources including journal articles, books, or Web links listed in the reference section are generally cited at least once within sentences in the body of the paper. From this observation, our idea of identifying a “comment-on” list for a given article is to recognize the sentences that mention “comment-on” articles by analyzing bibliographic data in the reference section.

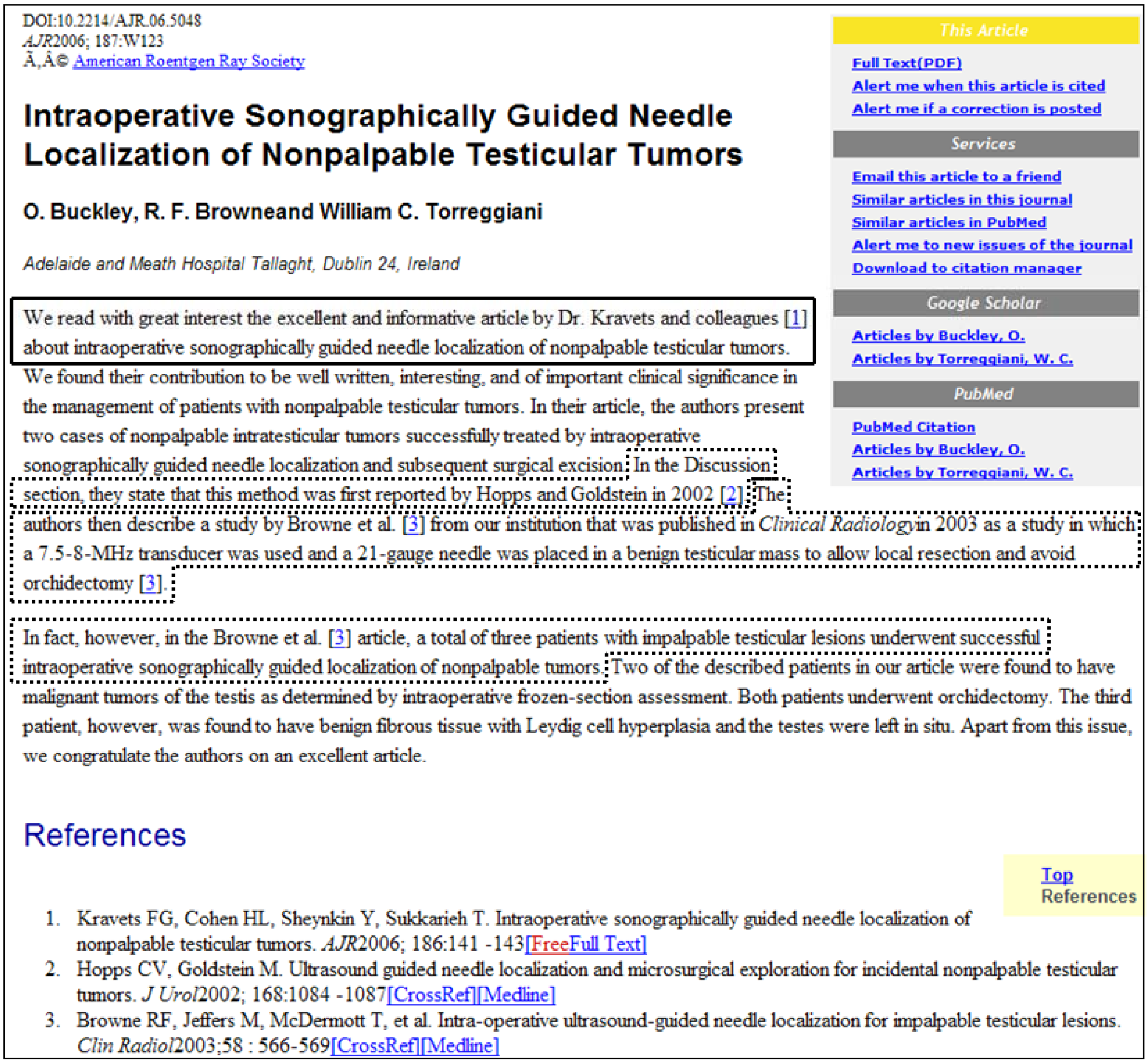
The SVM is a supervised learning algorithm for a two-class problem and it has been reported to show better generalization performance than other pattern recognition techniques such as neural networks. In our research, we implemented two types of SVMs: SVM with polynomial kernel function and SVM with radial basis function (RBF), and evaluated their recognition performance in terms of precision, recall, and F-measure rates. As an alternative way to identify “comment-on sentences”, we introduced another method based on the “significance of sentence” estimated by a specific scoring function and then compared its performance with that of SVMs.

The remainder of this paper is organized as follows. In Section 2, we provide a detailed description of the proposed method for identifying “comment-on sentences” including detection of “citation sentences”, extraction of input feature vectors, and learning and recognition process based on SVMs and score functions. Section 3 provides the results of recognition experiments on HTML-formatted online biomedical documents and error analysis. Final conclusions and issues related to future research are presented in Section 4.

2. AUTOMATED IDENTIFICATION OF “COMMENT-ON SENTENCE”

2.1 Extraction of “citation sentences” from HTML-formatted text

We define sentences located within the body text that cite external sources listed in the reference section as “citation sentences”. In the scientific literature, each “citation sentence” is usually associated with a tag (such as “(1)” or “[1]”) which we may call “in-text citation” or “parenthetical citation”. These tags point to the complete bibliographical description of the cited external source in the reference section. In addition, the sentence containing the “in-text citation” that specifically indicates the article commented on by the given article is defined as the “comment-on sentence”. A “comment-on sentence” is therefore a subset of “citation sentences”.



DOI:10.2214/AJR.06.5048
*AJR*2006; 187:W123
Ã‚Â© American Roentgen Ray Society

Intraoperative Sonographically Guided Needle Localization of Nonpalpable Testicular Tumors

O. Buckley, R. F. Browneand William C. Torreggiani

Adelaide and Meath Hospital Tallaght, Dublin 24, Ireland

This Article

Full Text(PDF)
Alert me when this article is cited
Alert me if a correction is posted

Services

Email this article to a friend
Similar articles in this journal
Similar articles in PubMed
Alert me to new issues of the journal
Download to citation manager

Google Scholar

Articles by Buckley, O.
Articles by Torreggiani, W. C.

PubMed

PubMed Citation
Articles by Buckley, O.
Articles by Torreggiani, W. C.

We read with great interest the excellent and informative article by Dr. Kravets and colleagues [1] about intraoperative sonographically guided needle localization of nonpalpable testicular tumors. We found their contribution to be well written, interesting, and of important clinical significance in the management of patients with nonpalpable testicular tumors. In their article, the authors present two cases of nonpalpable intratesticular tumors successfully treated by intraoperative sonographically guided needle localization and subsequent surgical excision. In the Discussion section, they state that this method was first reported by Hopps and Goldstein in 2002 [2]. The authors then describe a study by Browne et al. [3] from our institution that was published in *Clinical Radiology*in 2003 as a study in which a 7.5-8-MHz transducer was used and a 21-gauge needle was placed in a benign testicular mass to allow local resection and avoid orchidectomy [3].

In fact, however, in the Browne et al. [3] article, a total of three patients with impalpable testicular lesions underwent successful intraoperative sonographically guided localization of nonpalpable tumors. Two of the described patients in our article were found to have malignant tumors of the testis as determined by intraoperative frozen-section assessment. Both patients underwent orchidectomy. The third patient, however, was found to have benign fibrous tissue with Leydig cell hyperplasia and the testes were left in situ. Apart from this issue, we congratulate the authors on an excellent article.

References

Top
References

1. Kravets FG, Cohen HL, Sheynkin Y, Sukkarieh T. Intraoperative sonographically guided needle localization of nonpalpable testicular tumors. *AJR*2006; 186:141 -143[FreeFull Text]
2. Hopps CV, Goldstein M. Ultrasound guided needle localization and microsurgical exploration for incidental nonpalpable testicular tumors. *J Urol*2002; 168:1084 -1087[CrossRef][Medline]
3. Browne RF, Jeffers M, McDermott T, et al. Intra-operative ultrasound-guided needle localization for impalpable testicular lesions. *Clin Radiol*2003;58 : 566-569[CrossRef][Medline]

Fig. 1. An online biomedical article showing “citation sentences” (dotted line) and a “comment-on sentence” (solid line).

Figure 1 shows an example of an online biomedical article having “citation sentences” and a “comment-on sentence”. Our approach is first to detect “citation sentences” from the body text of a given article and then identify “comment-on sentences” from these “citation sentences” using SVMs. The list of “comment-on” articles is finally retrieved from the description of corresponding external source of the identified “comment-on sentence” in the reference section.

Very often, in HTML-formatted online articles, such an “in-text citation” is hyperlinked to the corresponding external source as shown in Fig. 2 (a). A hyperlink consists of both a source anchor and a destination anchor. The source anchor specified by “A” HTML element with “href” attribute appears at the “in-text citation” and points to the destination anchor. The destination anchor specified by “A” element with “name” attribute can be found at the beginning of the external source’s description. The source anchor and its destination anchor have the same unique name. Therefore, by recognizing this anchor name, we can reliably detect its associated “citation sentence”.

However, HTML documents converted from the PDF format do not have such hyperlink information. In this case, therefore, we detect “citation sentences” by recognizing specific tags indicating “in-text citations” such as a pair of square or round brackets enclosing reference numbers, or author names and publication year (Fig. 2(b)), or by tracking a superscript HTML tag pair enclosing reference numbers (Fig. 2(c)).

Hyperlink (Source anchor)	Richard Horton, editor of <I>The Lancet,</I> recently described the<SUP> </SUP> underrepresentation of poorer countries both on the editorial<SUP> </SUP>boards and in the pages of medical journals as a systematic<SUP> </SUP>bias against the diseases of poverty.<SUP**><A HREF=“#R1-44”>** 1</A> </SUP>
Hyperlink (Destination anchor)	<OL COMPACT> **<A NAME=“R1-44”>** <!-- null --></A><LI VALUE=1> Horton R. Medical journals: evidence of bias against diseases of poverty<I>Lancet</I> 2003;361:712-3.<!-- HIGHWIRE ID=“170:1: 65:1” --><A HREF=“/cgi/external_ref?access_num=12620731 &link_type=MED”> [Medline]</A><!-- /HIGHWIRE --><A NAME=“R2-44”><!-- null --></A>

(a)

Text symbols of in-text citation	The combination of improved bone marrow transplantation techniques, now called HSC transplantation, supportive animal data **[2]** and coincidental observations (improvement in coexisting autoimmune disease after HSC transplantation for conventional indications, such as aplastic anaemia, leukaemia and cancer **[3]**) has allowed the concept to move forward to the clinic.

(b)

HTML tags (superscript)	<SPAN CLASS="ps1p14"><NOBR><SPAN CLASS="ft3">Evidence continues to accumulate that children in families where there</SPAN></NOBR></SPAN><SPAN CLASS="ps1p15"><NOBR> <SPAN CLASS= "ft3"> is sub- stance &#160;abuse, &#160; including &#160;alcohol, &#160;have &#160;a &#160;number of &#160; difficulties. <SPAN CLASS= "em1p5">**<SUP>1</SUP>**</SPAN> &#160;The rate &#160;</SPAN></NOBR> </SPAN><SPAN CLASS="ps1p16"><NOBR><SPAN CLASS="ft3">of externalizing and internalizing problem in children of substance abusers appears </SPAN></NOBR></SPAN><SPAN CLASS= "ps1p17"><SPAN CLASS="ft3">to be high.<SPAN CLASS="em1p5">**<SUP>3 </SUP>**</SPAN>

(c)

Fig. 2. Detection of “citation sentences” using (a) hyperlink information, (b) text symbols of “in-text citation”, and (c) superscript HTML tags.

2.2 Feature extraction

Once “citation sentences” are extracted, their input feature vectors are calculated for training and testing the SVMs. We define a 30-dimensional binary feature vector created by combining three types of basic features which are experimentally found to be effective to represent a “comment-on sentence”.

The first basic feature is based on matching the score of cue phrases that are usually embedded in a “comment-on sentence” to indicate which article is being commented upon. This feature is defined as follows;

$$M_c(s_i) = \frac{w_c(s_i)}{W_c + w_g(s_i)} \tag{1}$$

Here, W_c denotes the number of words in a cue phrase *c*. $w_c(s_i)$ and $w_g(s_i)$ represent the number of matched words between the cue phrase *c* and the sentence s_i, and the number of other words existing between matched words, respectively. In our study, we collected a total of 52 cue phrases, some of which are shown in Fig. 3, by analyzing hundreds of samples of “comment-on sentences”.

1) We ***read with interest the study by*** Trochet and colleagues (2004), published in the April 2004 issue of The American Journal of Human Genetics, that described …..

2) We ***would like to comment on*** the Schork and Greenwood (2004) article dealing with the inherent “bias” toward the null hypothesis in the context of nonparametric linkage analysis.

3) As an ophthalmologist, I ***have a question regarding*** the clinical history for the 16-year-old boy with progressive vision loss described by Shaun Morris and associates[1.]

4) Ross Baker and associates, ***in their study of*** adverse events (AEs) in Canadian hospitals,[3] used a method that has been used in other countries to characterize this problem, but we feel …..

5) ***In a recent editorial in*** the THI Journal,[1] Dr. Ferguson wrote that he had learned two new words: “anabasis” and “apotheosis.”

Fig. 3. Examples of cue phrase embedded in “comment-on sentences”.

The second basic feature is based on sentence position. In many cases, “comment-on sentences” are located at the beginning of the body text of an article. Thus such position information can also serve as a good feature to distinguish a “comment-on sentence” from other “citation sentences”. The position information of each sentence is expressed as

$$P(s_i) = 1 - \frac{BD(s_i)}{|D|} \tag{2}$$

Where $|D|$ is the total number of characters in the given document *D*, and $BD(s_i)$ is the number of characters located before the sentence s_i.

The last basic feature is the frequency of occurrence of author names of “comment-on” articles, based on our observation that author names of articles commented on are more frequently mentioned in the text. The frequency score of author names of external sources listed in the reference section is defined as follows;

$$TF(a_i) = \frac{tf(a_i, D)}{tf_{\max}(a, D)} \tag{3}$$

Where $tf(a_i, D)$ and $tf_{\max}(a, D)$ denote the number of occurrences of author a_i and the maximum number of occurrences of an author name in the given document *D*, respectively.

Each basic feature is quantized and normalized to a real value ranged from 0 to 1. The normalized real-valued features are then converted to a 10-bit binary vector for SVMs. These three 10-bit binary vectors are then concatenated to produce the 30-dimensional input feature vector for each “citation sentence”.

2.3 Recognizing “comment-on sentences” using an SVM approach

SVM was originally introduced as a supervised learning algorithm for solving a two-class problem, though it can be easily extended to handle multi-class problems [3][4]. Owing to its superior generalization performance, SVM has been widely used in many pattern recognition applications such as handwriting recognition [5], face detection [6], and text categorization [7]. Identification of “comment-on sentences” from “citation sentences” is also a typical two-class problem, the classes being “comment-on sentences” or general “citation sentences”.

The basic idea using SVM to solve a non-linear pattern recognition problem is to map a non-linear separable input space to a linear separable higher dimensional feature space, using a predefined non-linear kernel function, and to find the optimal hyperplane that maximizes the margins between the classes in that feature space, as shown in Fig. 4.

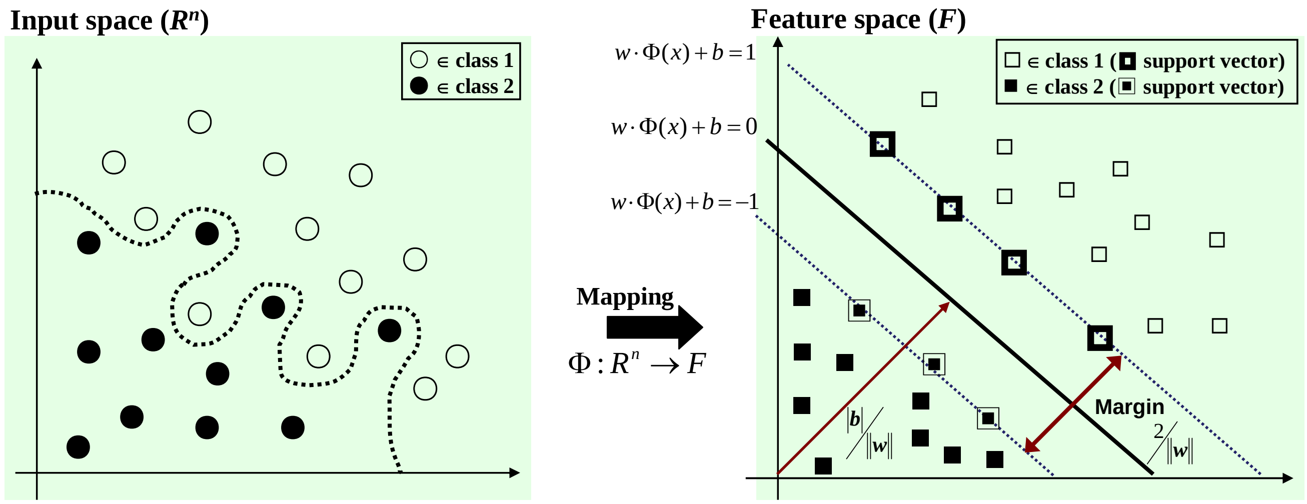

Input space (R^n)
○ ∈ class 1
● ∈ class 2
Feature space (F)
□ ∈ class 1 (▣ support vector)
■ ∈ class 2 (▣ support vector)
$w \cdot \Phi(x) + b = 1$
$w \cdot \Phi(x) + b = 0$
$w \cdot \Phi(x) + b = -1$
Mapping
$\Phi : R^n \rightarrow F$
$|b| / \|w\|$
Margin
$2 / \|w\|$

Fig. 4. SVM learning algorithm for nonlinear separable case.

Given that training samples $\{x_i, y_i\}$, $i = 1, \ldots, N, y_i \in \{-1,1\}, x_i \in R^n$ where y_i is the class label, the mapping Φ is implemented by a kernel function K such that $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$. Then the optimal hyperplane decision function $f(x)$ maximizing the margin and bounding the number of training errors is given by

$$\begin{aligned} f(x) &= \mathrm{sgn}(w \cdot \Phi(x) + b) \\ &= \mathrm{sgn}(\sum_{i=1}^{N} \lambda_i y_i \Phi(x_i) \Phi(x) + b) \\ &= \mathrm{sgn}(\sum_{i=1}^{N} \lambda_i y_i K(x_i, x) + b) \end{aligned} \quad (4)$$

where λ_i denotes a Lagrange multiplier. If λ_i is non-zero, the corresponding training sample x_i is called a support vector. We can see from (4) that such support vectors are crucial to determining the optimal hyperplane decision function. Other training vectors actually have no influence on the decision function due to their zero Lagrange multipliers. SVM learning is to find Lagrange multiplier λ_i by solving the following quadratic cost function:

$$\text{Maximize } W(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \quad (5)$$

$$\text{Subject to } \sum_{i=1}^{N} \lambda_i y_i = 0 \text{ and } 0 \le \lambda_i \le C, \quad i = 1, \ldots, N \quad (6)$$

Nonlinear decision function is finally realized by employing a specific nonlinear kernel function.

We implemented two types of SVMs: SVM with polynomial kernel function and SVM with RBF kernel function. These two kernel functions defined in equations (7) and (8) blow have been most commonly used in SVM-based pattern recognition applications. We evaluate their recognition performance using real online biomedical journal articles.

$$K(x, y) = (x \cdot y + 1)^p \quad (7)$$

$$K(x, y) = \exp(-\gamma \|x - y\|^2) \quad (8)$$

2.4 Recognizing "comment-on sentences" based on significance: an alternative approach

To compare with the SVM approach, we introduce an alternative method that recognizes "comment-on sentences" by estimating the "significance of sentence" using a predefined scoring function. The concept of measuring the significance of a given sentence using a score function is used conventionally in automatic text summarization [8][9].

In our research, we defined and implemented three types of scoring functions based on the three basic features mentioned previously, namely: the position of sentence, author name statistics, and linguistic and contextual information. In addition, the integration of the three individual scoring functions is introduced. At the learning stage, the threshold value of each score function is calculated using the training dataset (the same as that used in the SVM training). As a result, each score function shows the best performance in identifying "comment-on sentence" for the training dataset using that particular threshold. At the recognition stage, any "citation sentence" that has a significance value larger than the threshold calculated at the learning stage is labeled as a "comment-on sentence".

3. RECOGNITION EXPERIMENTS

3.1 Database

To build a dataset for the recognition experiments, we collected 1,236 "citation sentences" from 175 HTML-formatted online articles. These online articles appear in 24 different biomedical journal titles, and their publication types are Letter (62.3%), Review (5.1%), Editorial (4.0%), and Commentary (28.6%). We randomly selected 641 of these "citation sentences" to train the SVMs, and to calculate the threshold for the score functions. The remaining 595 "citation sentences" are used as the test set to evaluate the performance of SVMs and score functions.

3.2 Experimental results

We evaluated the performance of SVMs and score functions in terms of precision, recall, and F-measure rates that are defined as follows;

$$precision = TP/(TP + FP)$$

$$recall = TP/(TP + FN)$$

$$F - measure = 2 \times (precision \times recall)/(precision + recall)$$

Here, *TP*, *FP*, and *FN* denote true positive, false positive, and false negative, respectively. False (true) positive means that the target output is negative (positive) and the generated output is positive. False negative is the reverse of the above.

Our experiments show that SVM with polynomial kernel function yields the best performance in terms of recall (97.06%) and F-measure (97.06%) rates, as shown in Table 1. The SVM with RBF kernel function shows lower recall and F-measure rates than the score functions, though it yields a slightly better precision rate than SVM with polynomial function. Among the score function-based methods, the integrated method shows the best performance overall, as expected. Therefore we conclude that the SVM with polynomial kernel function is the most appropriate scheme for identifying "comment-on sentences".

Table 1. Precision, recall, and F-Measure rates of SVMs and score functions.

	Precision (%)	Recall (%)	F-Measure (%)
Score fn. 1 (sentence position)	65.35	64.71	65.02
Score fn. 2 (author name statistics)	83.19	97.06	89.59
Score fn. 3 (linguistic information)	85.71	52.94	65.45
Score fn. 4 (integrated method)	91.59	96.08	93.78
SVM with RBF	97.80	87.25	92.22
SVM with polynomial	97.06	97.06	97.06

An error analysis for the recognition results of the SVM with polynomial kernel function shows that many identification errors are caused by a very low matching score (see Eq. (1)) resulting from a lack of linguistic or contextual cues in a given "citation sentence". However, such errors may be reduced by gathering more cue phrases. In addition, a "comment-on sentence" can often be misrecognized as a general "citation sentence" when it is located at the middle or the end of the body text and the frequency of occurrence of its author name is significantly small relative to other author names. Conversely, a "citation sentence" located at the beginning of the body text can also be misrecognized as a "comment-on sentence". In order to minimize these problems, we recommend the addition of a rule-based approach based on cue phrases and other linguistic information in future work.

4. CONCLUSIONS

In this paper, we have presented an automated method based on SVMs for identifying "comment-on", a field in a MEDLINE citation denoting the list of previously published articles commented on by a given online biomedical article. Our strategy is first to detect all "citation sentences" from the body text of a given article using hyperlink information or specific tags indicating "in-text citations", and then to recognize the "comment-on sentences" that mention "comment-on" articles from these "citation sentences".

We have implemented and tested two types of SVMs, one with polynomial kernel function and the other with radial basis function. For comparative study, we have also introduced an alternative method that recognizes "comment-on sentences" by estimating the "significance of sentence" using four types of score functions. Through a series of experiments on HTML-formatted online articles collected from 24 different biomedical journal titles, we found that the SVM with polynomial kernel function performed best in terms of recall and F-measure rates.

Future work is planned to develop a rule-based method for compensating recognition errors made by SVMs, and to combine SVM-based and significance-based methods to further improve the recognition performance. Further study will also be considered to deal with scanned document images in which many OCR conversion errors are expected.

ACKNOWLEDGMENT

This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

REFERENCES

1. J. Kim, D. X. Le, and G. R. Thoma, "Automated labeling of bibliographic data extracted from biomedical online journals," *Proc. SPIE, Document Recognition and Retrieval*, 5010, 47-56, San Jose, Jan. (2003).
2. I. C. Kim, D. X. Le, and G. R. Thoma, "Automated cleanup processing for extracting bibliographic data from biomedical online journals," *Proc. 9th World Multiconference on Systemics, Cybernetics and Informatics*, 4, 401-405, Orlando, July (2005).
3. V. Vapnik, *The nature of statistical learning theory*, New York: Springer-Verlag, 1995.
4. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, 2(2), 1-43 (1998).
5. B. Zhao, Y. Liu, S. Xia, "Support vector machine and its application in handwritten numeral recognition," *Proc. 15th Int'l Conf. Pattern Recognition (ICPR)* 2, 2720-2723, Barcelona, Spain, Sept. (2000).
6. B. Heisele, T. Serre, M. Pontil, and T. Poggio, "Component-based face detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 1, 657-662, Hawaii, Dec. (2001).
7. T. Joachims, "Text categorization with support vector machine," *Proc. Euro. Con. Machine Learning (ECML)*, 137-142 (1998).
8. C. Nobata, S. Sekine, M. Murata, K. Uchimoto, M. Utiyama, and H. Isahara, "Sentence extraction system assembling multiple evidence," *Proc. 2nd NTCIR Workshop*, 319-324 (2001).
9. T. Kikuchi, S. Furui, and C. Hori, "Automatic speech summarization based on sentence extraction and compaction," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, I, 384-387, Hong Kong, April (2003).