Identifying "Comment-on" Citation Data in Online Biomedical Articles Using SVM-based Text Summarization Technique

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 - *Comment-on (CON), a MEDLINE® citation field, indicates previously published articles commented on by authors expressing possibly complimentary or contradictory opinions. This paper presents an automated method using a support vector machine (SVM)-based text summarization technique that identifies CON data by distinguishing CON sentences from "citation sentences" and analyzes their corresponding bibliographic data in the references. We compare the performance of two types of SVM, one with a linear kernel function and the other with a radial basis kernel function (RBF). Input feature vectors for these SVMs are created by combining five feature types: 1) word statistics, 2) frequency of occurrence of author names, 3) sentence positions, 4) similarity between titles, and 5) difference of publication years. Experiments conducted on a set of online biomedical articles show that the SVM with a RBF is more reliable in terms of precision, recall, and F-measure rates than the SVM with a linear kernel function for identifying CON.*

Keywords: "Comment-on" identification, online biomedical documents, support vector machine, MEDLINE

1 Introduction

MEDLINE is the premier bibliographic online database of the U.S. National Library of Medicine (NLM) containing more than 20 million citations from over 5,500 selected worldwide biomedical journals, and accessed through NLM's PubMed service. With rapid growth of biomedical literature, both the number of journals indexed and the number of citations produced by NLM increase dramatically; 130 journal titles are newly added each year on average and nearly 700,000 citations were added to MEDLINE in 2010. Bibliographic citation data describing the article consists of more than 50 fields such as author names, article title, affiliation, etc. Currently, the majority of data for these fields are provided electronically in XML format from journal publishers. However, publishers leave out data for several important fields: Databank accession numbers, Grant supports/numbers, Comment-on/Comment-in, and investigator names, almost certainly because including these fields would be highly labor-intensive and costly. As manual extraction and entry of bibliographic data missing from publisher-provided XML files would be equally burdensome for NLM, there is a strong motivation to develop automated systems to minimize human labor and to provide bibliographic data accurately and in a timely fashion.

The Lister Hill National Center for Biomedical Communications (LHNCBC), a research and development division of 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]. This paper presents one of the major components of WebMARS, an automated method for identifying and extracting "Comment-on" (CON) citation data. CON is a field in a MEDLINE citation listing previously published articles commented on by authors of a given paper in a complimentary, or sometimes contradictory, manner. We refer to the "Commented on" articles as CON articles, and the papers in which such opinions are expressed as "Comment-in" (CIN) articles.

Manually extracting the CON list from a given article is time-consuming and labor-intensive, and relies heavily on human operators' linguistic knowledge and their understanding of scientific expressions and writing styles. Generally, authors of a CIN article cite CON articles related to their research as primary external sources on which they may express complimentary or contradictory opinions. Thus the full bibliographic descriptions for these CON articles can usually be found in the reference section of a CIN article. Furthermore, all external sources (journal articles, books, or Web links) listed in the reference section of the CIN paper are generally mentioned at least once within sentences ("citation sentences") in the body text.

From this observation, our idea of identifying the CON list for a given article is to recognize the sentences ("CON sentences") that mention CON articles from the "citation sentences" in the body text using a support vector machine (SVM)-based text summarization technique and analyze the corresponding bibliographic data in the reference section. In our research, we implemented two types of SVMs: one with a linear kernel function and the other with a radial basis function (RBF), and compared their performance in terms of precision, recall, and F-measure rates. Five types of features were employed to create an input feature vector for these SVMs: 1) word statistics representing how differently a word is distributed in CON sentences and other "citation sentences", 2) frequency of occurrence of author names of external sources listed in the reference section of a given input article, 3) sentence positions within an article body text, 4) similarity of titles between a given input article and external sources, and 5) difference of publication years between an input article and each external source.

2 Related work

CIN articles are usually short papers such as commentaries, letters, editorials, or brief correspondences, written mainly for the purpose of supporting, refuting, or discussing other articles (CON). Accordingly, a sentence specifically referring to a CON article, called a "CON sentence", can be considered a key part of a CIN article because it is indicative of the article's subject and purpose. Detecting and extracting such key sentences within a document is a text summarization task. A summary can be loosely defined as text that conveys important information in the original text(s) and is a condensed representation (no longer than half) of the original text(s) [3]. Automated text summarization is the process of automatically constructing a summary for an input text. This summary can either be an "extract" created by merely reusing portions of the input text such as phrases, sentences, or paragraphs that are likely to be most important, or an "abstract" that is a newly generated text after an analysis of the original text.

Since creating such an abstract requires the high complexity of natural language processing techniques and knowledge engineering technology, most text summarization studies have focused on the extraction-based method. Our task of identifying CON sentences from the body text of CIN articles can also be considered as a typical extraction-based text summarization method. Text summarization has been addressed by a variety of methods and applied in different domains and genres of documents. Most early studies were based on surface-level features that do not require linguistic analysis, such as word frequency, paragraph or sentence position, and cue phrases to determine the most important concepts within a document [4][5][6].

There is another group of studies that builds an internal representation of the text by modeling text entities and their relationships to determine salient information. For example, Barzilay and Elhadad [7], and Silber and McCoy [8] employed lexical chains representing semantic relations between words to generate a summary of the original document. In addition, an approach exploiting the global structure of the text such as document format, rhetorical structure, etc. has also been reported [9][10].

All these aforementioned approaches can be implemented as either linguistic knowledge or machine-learning techniques. Linguistic knowledge-based methods that try to semantically analyze the structure of the text involve very sophisticated and expensive linguistic processing. Therefore, most methods employed in the recent literature are based on statistical theories and machine learning techniques; e.g., Naïve Bayes [11], decision tree [12], neural networks [13], hidden Markov models [14], and SVMs [15].

3 CON and CIN articles

CIN and CON articles are indicated in MEDLINE citation fields as "Comment in" and "Comment on" respectively, and linked together. As an example, Fig. 1(a) is the MEDLINE citation of an article (CIN) in which a "Commented on" article is cited. This CON information, shown enclosed in a dotted box, consists of abbreviated journal title, publication year, volume and issue number, and pagination. Conversely, as shown in the dotted box in Fig. 1(b), the MEDLINE citation for this CON article cites the CIN article in which it is mentioned. Thus readers may get to either citation from the other.

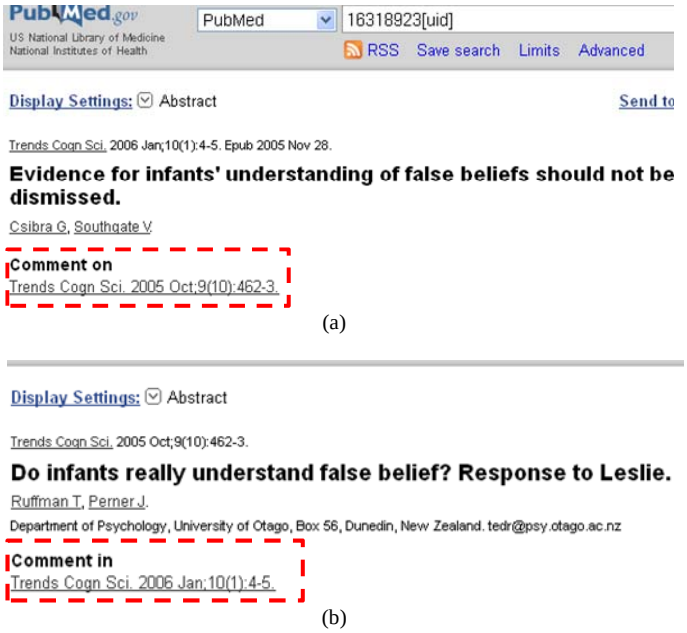


Fig. 1. (a) "Comment on" and (b) "Comment in" citations in MEDLINE

3.1 Issues

Currently, the CON list is created manually, based on certain linguistic clues and contextual patterns; operators are required to look for a particular sentence that contains a cue phrase expressing either complimentary or contradictory opinions on other articles from the body text of a given article. Typical examples of such cue phrases are listed in Table 1.

Table 1. Examples of cue phrases frequently found in CON sentences.

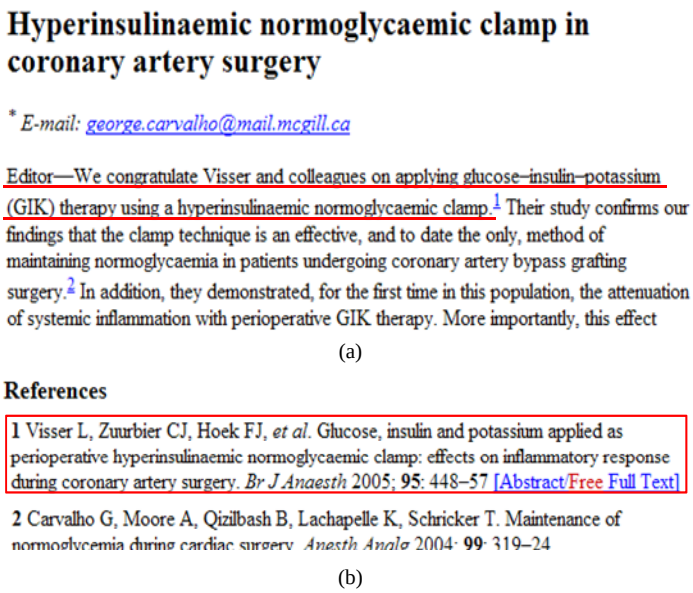
We congratulate [authors] for …
We question the interpretation by …
The {article | paper | letter | study | research} by …
We read with interest …
We would like to {reply | comment} to …
We agree with …

This manual method is highly labor-intensive and time-consuming. Furthermore, owing to a wide variety of linguistic expressions, and linguistic and contextual similarity between CON sentences and "citation sentences" generally citing other external sources, the overall performance relies heavily on human operators' experience, linguistic knowledge, and their understanding of scientific expressions and writing styles. In order to minimize manual efforts and to improve

accuracy and processing speed in detecting sentences that comment on other articles, we propose an automated method using a support vector machine (SVM)-based text summarization technique.

4 Method

As mentioned earlier, authors of a CIN article cite CON articles as primary external sources. Accordingly, full bibliographical descriptions for these CON articles can usually be found in the reference section of a CIN article. Note that in the scientific literature, all external sources listed in the reference section are generally mentioned at least once within sentences ("citation sentences") in the body text. The "citation sentence" that specifically indicates an article commented on by a given article is defined as the CON sentence. CON sentences are therefore a subset of "citation sentences". Based on this observation, our approach to identify a CON list for a given article is to first extract all "citation sentences" from its body text, and then to recognize the sentences among these that mention CON articles, using SVMs and to analyze the corresponding bibliographic data in the reference section. Figure 2(a) shows an example of a CON sentence (solid underline) and Fig. 2(b) shows its corresponding reference (solid box).



Hyperinsulinaemic normoglycaemic clamp in coronary artery surgery

* *E-mail:* george.carvalho@mail.mcgill.ca

Editor—We congratulate Visser and colleagues on applying glucose–insulin–potassium (GIK) therapy using a hyperinsulinaemic normoglycaemic clamp.[1] Their study confirms our findings that the clamp technique is an effective, and to date the only, method of maintaining normoglycaemia in patients undergoing coronary artery bypass grafting surgery.[2] In addition, they demonstrated, for the first time in this population, the attenuation of systemic inflammation with perioperative GIK therapy. More importantly, this effect

(a)

References

1 Visser L, Zuurbier CJ, Hoek FJ, *et al*. Glucose, insulin and potassium applied as perioperative hyperinsulinaemic normoglycaemic clamp: effects on inflammatory response during coronary artery surgery. *Br J Anaesth* 2005; **95**: 448–57 [Abstract/Free Full Text]

2 Carvalho G, Moore A, Qizilbash B, Lachapelle K, Schricker T. Maintenance of normoglycemia during cardiac surgery. *Anesth Analg* 2004; **99**: 319–24

(b)

Fig. 2. (a) A CON sentence and (b) its corresponding bibliographic description in the reference section

Our method consists of four main steps: 1) extraction of text zones of interest, 2) extraction of "citation sentences" and the corresponding bibliographical description of external sources, 3) creation of input feature vectors for SVMs, and 4) classification of CON sentences by SVMs.

Since our method takes advantage of clues from the article title, the body text, and the reference section in a given HTML-formatted online article, we need to segment the entire article into smaller logical zones, and detect such zones first. In our research, these text zones of interest are extracted using Zoning and Labeling modules detailed in [1] and [16]. Here, we focus on and provide details about the remaining three steps.

4.1 Extraction of "citation sentences" and the corresponding external source's description

In the scientific literature, each "citation sentence" is usually associated with a citation tag (such as "(1)" or "[1]") that points to the complete bibliographical description of the cited external source in the reference section. In addition, in HTML-formatted online articles, such a "citation sentence" is hyperlinked to its corresponding external source as shown in Table 2. A hyperlink consists of both a source anchor and a destination anchor. The source anchor specified by an "A" HTML element with a "*href*" attribute appears before or behind a citation tag in a "citation sentence" and points to the destination anchor. The destination anchor specified by an "A" element with a "*name*" attribute can be found at the beginning of the external source's description. The source anchor and its destination anchor should have the same unique name. Therefore, by recognizing this anchor name, we can reliably detect its associated "citation sentence" and its corresponding external source.

4.2 Feature extraction

In our research, five types of features were employed to build an input feature vector for SVM: 1) word statistics representing how differently a word is distributed in CON sentences and other "citation sentences", 2) frequency of occurrence of author names of external sources, 3) sentence positions within the body text, 4) similarity of titles between an input article and its external sources, and 5) difference of publication years between an input article and its external sources. These features were experimentally found to be effective to distinguish CON sentences from other "citation sentences". The first feature—word statistics—is based on a

Table 2. An example of a "citation sentence" and its hyperlinked external source.

Hyperlink (Source anchor)	<SPAN>Editor—We congratulate Visser and colleagues on applying glucose–insulin–potassium (GIK) therapy using a hyperinsulinaemic normoglycaemic clamp.</SPAN>^{**<A href="#B1">**1</A>}
Hyperlink (Destination anchor)	<P> **<A name=B1>** <!-- null --></A><STRONG>1</STRONG><SPAN> Visser L, Zuurbier CJ, Hoek FJ, </SPAN><I>et al</I><SPAN>. Glucose, insulin and potassium applied as perioperative hyperinsulinaemic normoglycaemic clamp: effects on inflammatory response during coronary artery surgery. </SPAN><I>Br J Anaesth</I><SPAN> 2005; </SPAN><B>95</B><SPAN>: 448–57 </SPAN>

bag of words, a vector of words. Using words as an input feature requires a very high dimensional feature space (21,314 dimensions in our case). Although SVM can manage (lead to a convergence) such a high dimensional feature space, many have suggested the need for word selection or dimension reduction to employ other conventional learning methods, to reduce the computational cost, to improve the generalization performance, and to avoid the over-fitting problem. A typical approach for word selection is to sort out words according to their importance. Many functions have been proposed to measure the importance of a word, including term frequency (TF), inverse document frequency (IDF), χ^2 statistics, and simplified χ^2 ($s\chi^2$) statistics [17]. The use of $s\chi^2$ has been reported as delivering the best performance since it removes redundancies, and emphasizes extremely rare features (words) and rare categories from χ^2 [18].

In our task, $s\chi^2$ of word t_k for CON sentences (class c_0) and other "citation sentences" (class c_1) can be defined as follows;

$$s\chi^2(t_k, c_i) = P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i) \quad i = 0, 1 \qquad (1)$$

where $P(t_k, c_i)$ denotes the probability that, for a random sentence *x*, word t_k occurs in *x*, *x* belongs to class c_i, and is estimated by counting its occurrences in the training set. The importance of word t_k is finally measured as follows;

$$s\chi^2_{max}(t_k) = max_i s\chi^2(t_k, c_i) \qquad i = 0, 1 \qquad (2)$$

Accordingly, the more differently a word is distributed in CON sentences and other "citation sentences" classes the higher its $s\chi^2_{max}(t_k)$. Our 21,314 words are sorted according to their $s\chi^2_{max}$ and a bag of words feature is created by selecting words having highest $s\chi^2_{max}$ scores. A series of experiments to investigate the influence of word reduction and to discover the number of words showing the best classification performance is also performed. These experiments are described in Section 5. A bag of words feature is converted to a binary vector; each vector component is assigned 1 if the corresponding word is found in a given sentence, or 0 otherwise.

The second feature is based on the sentence position. In many cases, CON 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 CON sentence from other "citation sentences". The position information of each sentence is expressed as

$$P(s_i) = 1 - \frac{BD(s_i)}{|D|} \qquad (3)$$

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.

Next, the frequency of occurrence of author names of external sources listed in the reference section is employed as another feature. Based on our observation, author names of CON articles are more frequently mentioned in the text of a CIN article. The frequency score of author names of external sources is defined as follows:

$$TF(s_i) = \frac{tf(s_i, D)}{tf_{max}(s, D)} \qquad (4)$$

where $tf(s_i, D)$ and $tf_{max}(s, D)$ denote the number of occurrences of author name of the external source associated with the "citation sentence" s_i and the maximum number of occurrences of an author name in the given document *D*, respectively.

The next input feature is based on similarity of titles between an input article and its external sources in the reference section. Basically, CIN and its CON articles have the same research topic because a CIN article is mainly for commenting about particular external sources (some CIN articles explicitly mention author names and/or titles of CON articles in their titles). Therefore, it is expected that their titles would be quite similar or have common keywords closely related to their research topic.

The similarity score between titles of CIN and external sources is simply measured using the ratio of the number of common words to the total number of words in the title of an external source excluding stop words, as shown below;

$$TS(t_{IN},\ s_i) = \frac{W_c(t_{IN},\ s_i)}{W(s_i)} \qquad (5)$$

Here, $W(s_i)$ denotes the number of words in the title of the external source associated with the "citation sentence" s_i in a given input article and $W_c(t_{IN}, s_i)$ is the number of words commonly found in the titles of input article and the external source for "citation sentence" s_i.

Each of the aforementioned three features, $P(s_i)$, $TF(s_i)$, and $TS(t_{IN},\ s_i)$, has a real value ranging from 0 to 1 and is converted to a 10-bit binary vector for SVM (*i*-th bit position corresponding to real values between *i*/10 and (*i*+1)/10). For $P(s_i)$ and $TF(s_i)$, one more bit component is additionally attached to represent if a given "citation sentence" is located at the first paragraph of the body text, and if this "citation sentence" includes the author name of its corresponding external source, respectively, thereby resulting in an 11-bit binary vector for each feature.

Lastly, the difference of publication years between an input article and its external source is also employed as an input feature. Based on our findings from the training dataset, authors of many CIN articles are found to be interested in and comment on recently-published articles. This feature would not be used alone because other recently-published general external sources are also found in the reference section. However, it is expected to be helpful for improving the accuracy of identifying CON sentences when combined with other features. The difference of publication years is represented using a 10-bit binary vector of which the index of each bit corresponds to the years of difference; bit 0 is set to 1 if the input article and its external source are published in the same year, and bit 9 is set to 1 if there is a difference of 9 or more years.

Finally, all these feature vectors are concatenated to build an input feature vector for the SVM-based training and classification tasks.

4.3 SVM classifiers

SVM [19] was originally introduced as a supervised learning algorithm based on the structural risk minimization principle for solving a two-class problem, though it can be easily extended to handle multi-class problems. Owing to its consistently superior performance compared to other existing methods, SVM has been widely used in many text categorization and summarization tasks. The basic idea of 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 kernel function, and to find the optimal hyperplane that maximizes the margins between the classes in that feature space.

As mentioned earlier, we employed two types of SVMs: one with a linear kernel function and the other with a RBF. These two kernel functions, defined in equations (6) and (7) below, respectively, have been commonly used in SVM-based pattern recognition applications. We implemented these SVMs using MYSVM (for linear kernel function) and LibSVM (for RBF), free software packages for non-commercial use [20][21], and evaluated their recognition performance using HTML-formatted online biomedical journal articles.

$$K(x_i, x_j) = (x_i^T \cdot x_j) \quad (6)$$

$$K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2) \quad (7)$$

5 Recognition experiments

5.1 Database

To build a dataset for our experiments to distinguish CON sentences from "citation sentences", we first collected a total of 5,848 HTML-formatted biomedical articles containing CON citations from a collection of indexed articles in MEDLINE. These online articles appeared in 414 different biomedical journal titles, and their publication types were Letter (49.0%), Review (2.1%), Editorial (25.4%), Commentary (14.5%), and others (9.0%). Full-length articles were not included in our dataset because CIN articles are generally letter-like short papers and MEDLINE does not typically acquire CON data from conventional full-length articles. We also developed an automated text categorization system to distinguish CIN articles from regular ones and to submit only articles classified as CIN to the proposed method of extracting CON citation data [22].

From these articles, 11,939 "citation sentences" were extracted; among them, 8,531 sentences (4,184 CON sentences + 4,347 other "citation sentences") were randomly selected to train the SVMs. The remaining 3,408 sentences (1,659 CON sentences + 1,749 other "citation sentences") were used to evaluate and compare the performance of the SVMs.

5.2 Experimental results

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

$$precision = TP/(TP + FP)$$
$$recall = TP/(TP + FN)$$
$$F - measure = \frac{2 \times (precision \times recall)}{(precision + recall)}$$

Here, *TP*, *FP*, and *FN* denote true positive, false positive, and false negative, respectively. False positive means that a "citation sentence" is misrecognized as a CON sentence. False negative is the reverse of the above.

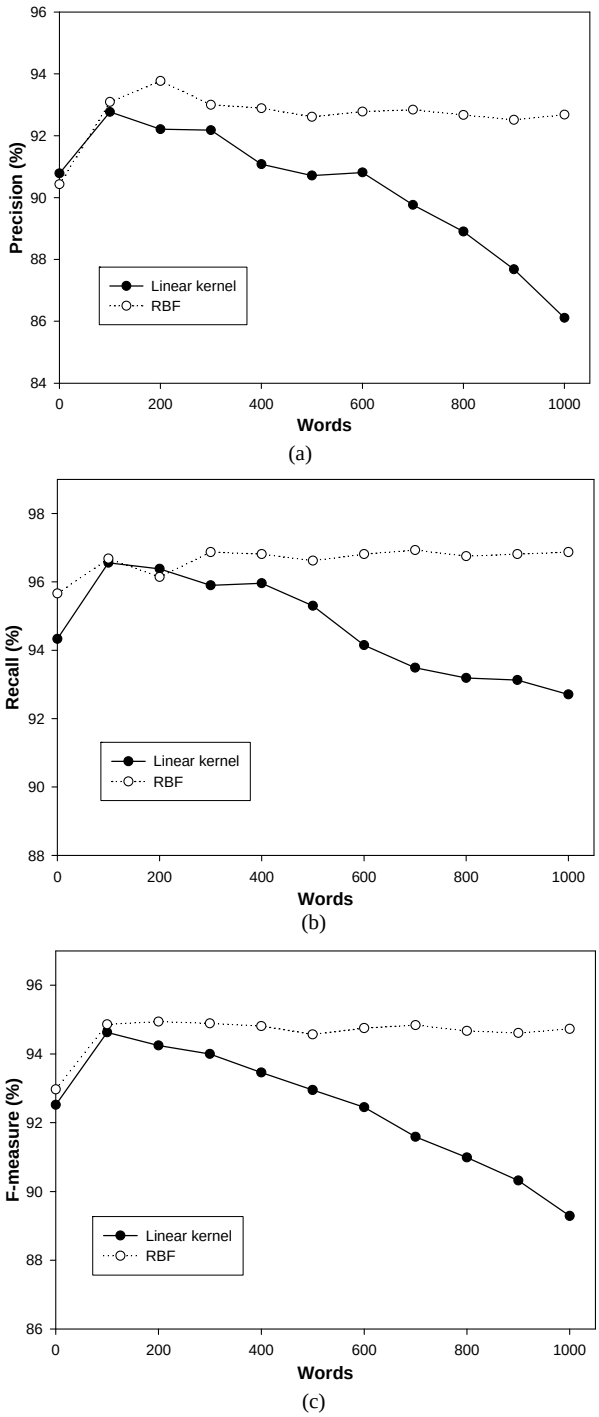


Fig. 3. (a) Precision, (b) recall, and (c) F-measure rates plotted against different word dictionary sizes.

Figure 3 shows precision, recall, and F-measure rates as functions of the size of the word dictionary in the bag of words feature ($s\chi^2_{max}$). We can see that the SVM with a RBF is slightly better than the SVM with a linear kernel, but both yield the best performance overall when the word size is 100, the performance in all three measures exceeding 93%. The SVM with a RBF shows consistent and reliable performance with respect to word selection in $s\chi^2_{max}$; its precision, recall and F-measure rates do not significantly vary with dictionary size. In contrast, the performance of SVM with a linear kernel function deteriorates as the word dictionary size increases. Therefore, we conclude that the SVM with a RBF is a more reliable classifier than the SVM with a linear kernel function for identifying CON sentences.

Table 3 shows examples of false-negative and false-positive errors. Two CON sentences shown in Table 3 (a): CON-I and CON-II, belong to the same input article and have a similar expression to comment other articles. However, unlike CON-I, CON-II is misrecognized as a general "citation sentence". It can be seen from its feature values that CON-II is located in the middle of the body text ($P(s_i)$ = 0.57), the frequency of occurrence of author name of the corresponding external source ($TF(s_i)$) is significantly small relative to other author names. Moreover, there are no words in common ($TS(t_{IN}, s_i)$) found between titles of input article and the external source corresponding to CON-II.

In contrast, the "citation sentence" (CIT-I) shown in Table 3 (b) is misrecognized as a CON sentence. CIT-I is found to contain several words frequently found in many CON sentences and located at the upper middle of the body text. Specially, it has a high similarity score ($TS(t_{IN}, s_i)$ = 0.83) between titles of CIN and its corresponding external source.

In order to minimize such problems, we recommend the addition of heuristic rules based on cue phrases and other linguistic information in future work.

Table 3. Error examples showing (a) false negative error and (b) false positive error

Input article	Growth attenuation: a diminutive solution to a daunting problem
CON-I	Now examine the proposed treatment: in this issue of the ARCHIVES, Gunther and Diekema (1) offer a medical solution to families who will likely face the harrowing choice of what to do when their child becomes too big to care for at home.
	Ref.: 1. Gunther DF, Diekema DS. *Attenuating growth in children with profound developmental disability: a new approach to an old dilemma*. Arch Pediatr Adolesc Med. 2006;160:1013-1017.
	$P(s_i)$ = 0.91 | $TF(s_i)$ = 1.00 | $TS(t_{IN}, s_i)$ = 0.10|
CON-II	Indeed, as Lee and Howell (2) point out in this issue of the ARCHIVES, estrogen has long been used to attenuate growth in girls destined to be taller than average.
	Ref.: 2. Lee JM, Howell JD. *Tall girls: the social shaping of a medical therapy*. Arch Pediatr Adolesc Med. 2006;160:1035-1039.
	$P(s_i)$ = 0.57 | $TF(s_i)$ = 0.29 | $TS(t_{IN}, s_i)$ = 0.00|

(a)

Input article	Ischemic hepatitis and collateral damage to the liver in severe viral respiratory tract infections
CON-III	Polakos and colleagues (1) investigated immunological causes of hepatic involvement by influenza virus respiratory tract infection manifesting itself in alanine and aspartate aminotransferase elevation and found evidence for a role of antigen-specific T cells in their pathogenesis.
	Ref.: 1. Polakos NK, Cornejo JC, Murray DA, Wright KO, Treanor JJ, Crispe IN, Topham DJ, Pierce RH: *Kupffer cell-dependent hepatitis occurs during influenza infection.* Am J Pathol 2006, 168:1169-1178
	$P(s_i)$ = 1.00 | $TF(s_i)$ = 1.00 | $TS(t_{IN}, s_i)$ = 0.17|
CIT-I	Adams and Hubscher (2) mention in their commentary on the work of Polakos and colleagues (1) our observational study, (3) in which we reported on the finding of elevated transaminase levels in 46% of children ventilated in the pediatric intensive care unit with severe respiratory syncytial virus bronchiolitis.
	Ref.: 2. Adams DH, Hubscher SG: *Systemic viral infections and collateral damage in the liver*. Am J Pathol 2006, 168:1057-1059
	$P(s_i)$ = 0.69 | $TF(s_i)$ = 0.5 | $TS(t_{IN}, s_i)$ = 0.83|

(b)

6 Conclusions

CON ("Comment-on") is a MEDLINE citation field showing previously published articles commented on by authors of a given article ("Comment-in" or CIN) as primary external sources on which they may express complimentary or contradictory opinions. Manually extracting the CON list from a given article is time-consuming and labor-intensive, and the overall performance relies heavily on human operators' experience, linguistic knowledge, and their understanding of scientific expressions and writing styles.

In this paper, we have presented a SVM-based text summarization method to automatically identify such CON data from online biomedical documents, thereby minimizing manual effort and improving accuracy and processing speed. Our main idea is to extract "citation sentences" using hyperlink information and then to recognize from the "citation sentences" CON sentences using SVMs. In our research, we have implemented two types of SVMs: one with a linear kernel function and the other with a radial basis function (RBF), and compared their performance in terms of precision, recall, and F-measure rates. Input feature vectors for these SVMs are created by combining five types of features: 1) word statistics representing how differently a word is distributed in CON sentences and other "citation sentences", 2) frequency of occurrence of author names of external sources listed in the reference section of a given input article, 3) sentence position within the body text, 4) similarity of titles between a given input article and external sources, and 5) difference of publication years between an input article and each external source.

Through a series of experiments on HTML-formatted online articles collected from 414 different biomedical journal titles, we can see that the SVM with a RBF and the SVM with a linear kernel both yield the best performance overall (over 93%) when the word size in the bag of words feature is 100. In addition, we found that the SVM with a RBF yields consistent and reliable performance in terms of precision, recall, and F-measure rates than the SVM with a linear kernel function with respect to word selection in the bag of words feature. Future work is planned to develop a rule-based method for compensating for recognition errors made by SVMs.

Acknowledgment

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

7 References

[1] J. Kim, D.X. Le, and G.R. Thoma, "Naïve bayes classifier for extracting bibliographic Information from biomedical online articles," *Proc. 5th Int'l Conf. Data Mining*, II, pp. 373-378, Las Vegas, 2008.

[2] I. Kim, D.X. Le, and G.R. Thoma, "Hybrid approach combining contextual and statistical information for identifying MEDLINE citation terms," *Proc. 15th SPIE Document Recognition and Retrieval*, 6815, 68150P (1-9), San Jose, 2008.

[3] D. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on text summarization," *Computational Linguistics*, 28(4), pp. 399-408, 2002.

[4] H.P. Luhn, "The automatic creation of literature abstracts", *IBM Journal of Research and Development*, 2(2), pp.159-165, 1958.

[5] P.B. Baxendale, "Man-made index for technical literature – An experiments", *IBM Journal of Research and Development*, 2(4), pp. 354-361, 1958.

[6] H.P. Edmundson, "New methods in automatic extracting", *Journal of the Association for Computing Machinery (JACM)*, 16(2), pp. 264-285, 1969.

[7] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization", *Proc. ACL'97/EACL'97 workshop on intelligent scalable text summarization*, pp. 10-17, Madrid, Spain, 1997.

[8] H.G. Silber and K.F. McCoy, "Efficient text summarization using lexical chains," *Proc. 5th int'l Conf. Intelligence User Interfaces*, pp. 252-255, New Orleans, 2000.

[9] W. Mann and S. Thompson, "Rhetorical structure theory: Toward a functional theory of text," *Text*, 8(3), pp. 243-281, 1988.

[10] D. Marcu, *The Rhetorical parsing, summarization, and generation of natural language texts*, PhD thesis, Dept. Computer Science, Univ. Toronto, Toronto, Canada, 1997.

[11] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," *Proc. 18th ACM-SIGIR Conf. Research and Development in Information Retrieval*, pp. 68-73, New York, 1995.

[12] E. Hovy and C.Y. Lin, "Automated text summarization in SUMMARIST," *In I. Mani, and M. Maybury editors, Advances in Automatic Text Summarization*, pp. 81-94, MIT press, 1999.

[13] K. Svore, L. Vanderwende, and C. Burges, "Enhancing single-document summarization by combining RankNet and third-party sources," *Proc. EMNLP-CoNLL*, pp. 448-457, 2007.

[14] J.M. Conroy and D.P. O'leary, "Text summarization via hidden Markov models," *Proc 24th ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 01)*, pp. 406-407, New York, 2001.

[15] T. Hirao, H. Isozaki, E. Maeda, and Y. Matsumoto: "Extracting important sentences with support vector machines", *Proc. 19th Int;l Conf. Computational Linguistics(COLING 2002)*, pp. 342-348, Taipei, Taiwan, 2002.

[16] Zou, J., Le, D. X. and Thoma, G. R. "Online medical journal article layout analysis," *Proc. 14th SPIE Document Recognition and Retrieval*, 6500, 65000V (1-12), San Jose, 2007.

[17] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, 34(1), pp. 1-47, 2002.

[18] L. Galavotti, F. Sebastiani, and M. Simi, "Experiments on the use of feature selection and negative evidence in to automated Text categorization," *ECDL 2000 LNCS 1923*, pp. 59-68, Springer, Heidelberg, 2000.

[19] V. Vapnik, *The nature of statistical learning theory*, New York: Springer-Verlag, 1995.

[20] S. Rüping, *mySVM-Manual*, Univ. Dortmund, 2000. [*http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM*]

[21] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," 2001. [*http://www.csie.ntu.edu.tw/~cjlin/libsvm*]

[22] I. Kim, D.X. Le, and G.R. Thoma, "Automated Identification of Biomedical Article Type Using Support Vector Machines," *Proc. 18th SPIE Document Recognition and Retrieval*, 7874, 787403, San Francisco, 2011.