Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007

Text Representations for Text Categorization: A Case Study in Biomedical Domain

Man LAN, Chew Lim TAN, Jian SU, Hwee Boon LOW

Abstract— **In vector space model (VSM), textual documents are represented as vectors in the term space. Therefore, there are two issues in this representation, i.e. (1) what should a term be and (2) how to weight a term. This paper examined ways to represent text from the above two aspects to improve the performance of text categorization. Different representations have been evaluated using SVM on three biomedical corpora. The controlled experiments showed that the straightforward usage of named entities as terms in VSM does not show performance improvements over the bag-of-words representation. On the other hand, the term weighting method slightly improved the performance. However, to further improve the performance of text categorization, more advanced techniques and more effective usages of natural language processing for text representations appear needed.**

I. Introduction

Text categorization is an important task for information management which automatically classifies unlabelled documents into a predefined set of categories. In the vector space model (VSM), the document d is usually represented as a vector in the term space, $d = (w_1, ..., w_k)$, where k is the size of the set of terms (*features*). The type of and the value of w_i are two key issues for text representation, that is, (1) what should a term be and (2) how to weight a term.

As the basic indexing unit, the term type can be at different levels, such as sub-word level (syllables), word level (single token), multi-word level (phrases, sentences), etc. Among them, the word level representation, i.e. the bag-of-words approach is the most widely-used one. In recent decade, researchers tried various alternative term types to represent text from semantic and syntactic aspects, such as phrases, word meaning, topics, semantic and syntactic relationships, etc. However, the experimental results found to date have not shown that these high level representations performed significantly better than the simple bag-of-words approach.

Different terms have different importance in a text and thus an important indicator w_i represents how much the ith term contributes to the semantics of document for text categorization. Our previous work in [1] classified the term weighting methods into *supervised term weighting methods* and *unsupervised term weighting methods*. The traditional methods borrowed from information retrieval field, such as $tf.idf$, $binary$, tf and their variants, belong to the unsupervised term weighting methods. In recent years, researchers have also introduced several new supervised term weighting methods ([1], [2], [3] and [4]).

With an overwhelming amount of textual information in biology and biomedicine, the needs for automatically extracting information, eg, protein protein interaction, from biomedical documents increase. For this purpose, first of all, we must classify whether the documents are relevant to protein protein interaction.

We thus investigate biomedical text classification through the two text representation issues mentioned above, which have not been explored so far. We explore the named entity representation in this study. Since we have earlier proposed a new term weighting method *tf.rf* in [1] and it has been confirmed to perform better than other methods through cross-classifier validation on Reuters corpus and 20 Newsgroup corpus, we further apply *tf.rf* to biomedical domain. Specifically, in this study, we adopted three biomedical corpora, i.e. Ohsumed corpus, BioCreAtIvE II corpus and 18 Journals corpus.

The rest of this paper is structured as follows. In Section II, we survey the existing text representations for text categorization from the term type and term weighting aspects. In Section III, we describe the methodology in this study including text representation methods adopted in this paper, learning algorithm, benchmark corpora and performance evaluation. In Section IV, we report results and discussion. The conclusions are in Section V.

II. Related Work on Text Representation

Even though text is already stored in machine readable form, such as HTML, PDF, DOC, Postscript and etc, it is generally not directly suitable for most learning algorithms. Before we apply one machine learning method to text categorization, the content of a textual document must be converted to a compact representation in order to be recognized and categorized by classifiers in a computer. Therefore, in VSM, there are two issues involved for text representation, i.e. term type and term weight.

A. Term Types

Generally, the indexing terms for representing documents can be at different levels, such as sub-word level (n-gram), word-level (single token), multi-word level (phrases, sentences), and semantic or syntactic level, etc.

Among them, the *bag-of-words* representation is so far the most common way to represent the content of text. The most

Man Lan is with the School of Computing, National University of Singapore, Singapore 117543 (phone: 65-65162784; email: lanman@comp.nus.edu.sg or lanman.sg@gmail.com). Chew Lim Tan is with the School of Computing, National University of Singapore, Singapore 117543 (phone: 65-65162900; email: tancl@comp.nus.edu.sg). Jian Su and Hwee Boon Low are with the Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 (email: sujian, hweeboon@i2r.a-star.edu.sg).

1-4244-1380-X/07/$25.00 ©2007 IEEE

advantage of this approach is simplicity as only the frequency of a word in a document is recorded, while all the structure and the ordering of the words is left out. However, due to this simplicity characteristic, it has often been criticized for its disregard of semantic relationships between words which are thought to be crucial to human understanding.

With the development of computational linguistic tools, large quantity of text can be analyzed efficiently with respect to their syntactic structure. Some researchers have used phrases and multiple words rather than individual words as indexing terms (see [6], [7], [8], [9], [10]). Moreover, to get rid of the problem of synonym in natural language, researchers also tried to use *word meanings* [11] or *term clustering* [12] to represent text. Furthermore, in order to capture the semantic relationships between words ignored by using the bag-of-words representation, in [9], the authors also included a hypernym-based representations by virtue of WordNet.

However, so far the experimental results showed that none of these new high level representations based on semantic and syntactic relationships performed significantly better than the bag-of-words representation. Even though the simple bag-of-words approach performs well in practice in most cases, it is still too early to draw a definite conclusion that the bag-of-words approach is better than the complicated representations. The current obstacle that hinders further research comes from the small annotated data corpora and the effective representations to preserve the information left out from the bag-of-words approach.

B. Term Weighting Method

No matter which term type we adopt to represent a document, each term in a document vector must be associated with a value (weight) which denotes their different importance in a text and contributions to text categorization task.

The traditional term weighting methods for text categorization are usually borrowed from information retrieval field and belong to the unsupervised term weighting methods, such as *tf.idf* [13], *binary*, *tf* and their different variants.

Recently, researchers proposed several supervised term weighting methods by using the prior information on the membership of training documents in predefined categories in the following ways. One approach is to weight terms by adopting feature selection metrics, such as χ^2, *information gain*, *gain ratio*, *odds ratio* and so on ([2] and [4]). However, these methods have not been shown to have a consistent superiority over the most widely-used $tf.idf$ method. Another approach is to weight terms in the interaction with a text classifier. Since the text classifier selects the positive documents from negative documents by assigning different scores to the documents, these scores are believed to be effective in weighting terms for text categorization, for example, in [14] terms are weighted using an iterative approach involving the kNN text classifier at each step. Most recently, the author in [3] introduced a new term weighting method called `ConfWeight` based on statistical confidence intervals. The experimental results showed that *ConfWeight* generally outperformed $tf.idf$ and *gain ratio* on three benchmark data collections.

III. Methodology

A. Text Representations in This Study

In this study, we examined two ways to represent biomedical text from term type and term weighting aspects as follows.

1) Named Entity: Recognizing named entities like gene, protein and virus, is quite important for biomedical information retrieval and information extraction. It is a challenging task because there is no standard naming conventions of named entities in the biomedical domain, being much more difficult than the one in the news domain. For example, many biomedical entity names are descriptive and have many words and numbers. One biomedical entity name may be with various spelling forms with capitalization or hyphen or even various irregular abbreviations. In [5], the authors presented a HMM-based named entity recognition system called PowerBioNE by exploring more evidential features to deal with various complex naming conventions in the biomedical domain.

In this study, we adopted named entity as term type to represent text based on the consideration that using named entities as features would capture some of the information left out from the bag-of-words representation. Due to lack of enough annotated training corpus, in this study, we only use PowerBioNE to extract protein names. Therefore, we only investigated the named entity-based representation in the BioCreAtIvE II corpus which are relevant to protein protein interaction documents.

2) Term Weighting Methods: Term weighting methods assign appropriate weights to terms to improve the performance of text categorization. We have earlier proposed a new effective supervised term weighting method *tf.rf* [1]. The *tf.rf* method has been confirmed to perform significantly better than other methods on two widely-used newswire benchmark corpora, i.e. Reuters corpus and 20 Newsgroups corpus, cross different learning methods. Therefore, in this new biomedical domain, we would like to see the results of *tf.rf* method. To examine the performance of different term weighting methods and make the comparison meaningful, we also included other widely-used term weighting methods as baseline, i.e. *tf.idf*, *binary* and *term frequency*.

B. Learning Algorithm

In this paper, linear SVM serves as benchmark learning algorithm due to its superior performance among these algorithms in previous studies (see [15], [16] and [17]). The SVM software we used is LIBSVM-2.8 [18].

C. Corpora

1) Ohsumed Data Corpus: The OHSUMED collection is a clinically-oriented MEDLINE subset from year 1987 to year 1991, consisting of 348, 566 references covering all references from 270 medical journals. The OHSUMED in year

1991 includes 74337 documents but only 50216 of which having abstracts. Joachims [16] selected the first 10000 documents for training and the second 10000 documents for testing from those with abstracts. In this study, we used the Ohsumed corpus adopted by Joachims as the comparison makes sense based on a benchmark data collection.

2) BioCreAtIvE II Corpus: The second BioCreAtIvE challenge evaluation in year 2006-2007 (*http://BioCreAtIvE.sourceforge.net/biocreative_2.html*) is a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain. One of the three tracks in BioCreAtIvE II is to identify protein protein interactions from biology literature. The study of protein protein interactions is one of the most pressing biological problems since characterizing protein interaction partners is crucial to understanding not only the functional role of individual proteins but also the organization of entire biological processes. Due to the rapid growth of the biomedical literature and the increasing number of newly discovered proteins, it is becoming difficult for the interaction database curators to keep up with the literature by manually detecting and curating protein protein interaction information. In practice, before detecting protein protein interaction descriptions in sentences, it is necessary to select those articles which contain relevant information relative to protein protein interactions. Thus the first sub-task of protein protein interaction track is concerned with the classification of whether a given article contains protein interaction information.

The training corpus of the first sub-task is a collection of PubMed article abstracts which contains a true positive set of 3536 documents (64.3%) which are relevant for protein protein interaction curation and a true negative set of 1959 documents (35.7%) which are not relevant for protein protein interaction curation from the two protein protein interaction databases, i.e. IntAct and MINT.

3) 18 Journals Corpus: From the digital library center in National University of Singapore, we chose 20 journals with the top largest impact factor in the subject categories of *Biochemistry and Molecular Biology*. Due to access limitation, two journals are not accessible[1]. The resulting data collection named 18 Journals Corpus consists of 7,903 documents from year 2004 to year 2005 of 18 journals in PubMed. Table I lists the statistical information of these 18 journals. After deleting the duplicates and removing the documents with blank abstract and/or blank MeSH keywords, the ultimate corpus has 5,417 documents and 933 MeSH keywords. Each of the 933 MeSH keywords is viewed as a category label and a document belongs to a category if it is indexed with at least one such keyword from these 933 keywords. Due to the large number of category labels (i.e. MeSH keywords), most categories contain only $1-2$ documents. Then we select 8421 document and category pairs from the top 132 categories each of which has at least 10 articles. For each category, the first half of documents are used to train the text classifier model and the last half of documents are used to test as unlabelled samples. After removing the 513 stop words, the whole vocabulary is 19018 words in all 132 categories.

[1]The two inaccessible journals are *Biochimica ET Biophysica ACTA-Reviews On Cancer* and *Reviews Of Physiology Biochemistry And Pharmacology*.

TABLE I
STATISTICAL INFORMATION OF THE 18 JOURNALS CORPUS

Abbreviated Journal Title	# of Articles
Annual review of biochemistry	78
Nature medicine	773
CELL	844
Molecular cell	688
Trends in biochemical sciences	229
PLoS biology	525
Annual review of biophysics & biomolecular structure	40
Nature structural & molecular biology	491
Current biology : CB	1396
The Plant cell	571
The EMBO journal	915
Genome research	501
Cytokine & growth factor reviews	98
Current opinion in structural biology	184
Progress in lipid research	39
Advances in microbial physiology	11
Current opinion in chemical biology	183
Cell death and differentiation	337

Generally, compared with the two previous corpora which involved domain experts in grouping the documents into certain categories, the documents in the 18 Journals corpus are grouped based on the indexed MeSH keywords. In some sense, the 18 Journals corpus is more difficult than the two previous corpora because the data are more "noisy". That is, the word/category correspondences are more "fuzzy" in the 18 Journals corpus. Consequently, the categorization for the 18 Journals corpus will be more difficult than those for the two previous corpora.

D. Performance Evaluation

Classification effectiveness is usually measured by using *precision* (P) and *recall* (R). *Precision* is the proportion of truly positive examples labelled positive by the system that were truly positive and *recall* is the proportion of truly positive examples that were labelled positive by the system. Usually, a classifier should thus be evaluated by means of a measure which combines *precision* and *recall*. The two most widely-used measures adopted by text categorization are F_1 function and *breakeven point*.

IV. RESULTS AND DISCUSSION

In this study, we conducted two series of experiments under different experimental circumstances. The purpose of the first series of experiments is to investigate the performance of four different term weighting methods, i.e. *binary*, *tf*, *tf.idf* and *tf.rf*, based on bag-of-words approach on these three corpora. The second series of experiments is to examine the performance of protein named entity-based representation for text categorization on the protein protein interaction corpus, i.e. BioCreAtIvE II corpus.

A. Performance of Different Term Weighting Methods for Text Categorization

1) Results on the Ohsumed Corpus: Figure 1 shows the micro-averaged breakeven results of the four term weighting methods on the Ohsumed corpus. Table II summarizes the

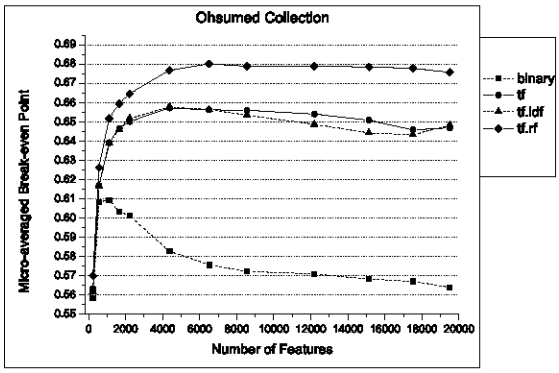

Fig. 1. Results of the four term weighting methods on the Ohsumed Data Collection.

best results of four different schemes on the Ohsumed corpus, where the best micro- and macro-F_1 scores are shown in bold font. Note that the micro-averaged *precision* and micro-averaged *recall* score in this table are almost equal to each other, thus the micro-averaged F_1 value actually coincides with the micro-averaged *breakeven point*.

TABLE II
THE BEST RESULTS OF FOUR TERM WEIGHTING METHODS ON THE OHSUMED CORPUS.

Scheme	micro-R	micro-P	micro-F1	macro-F1
$binary$	0.6091	0.6097	0.6094	0.5757
tf	0.6578	0.6566	0.6572	0.6335
$tf.idf$	0.6567	0.6588	0.6578	0.6407
$tf.rf$	**0.6810**	**0.6800**	**0.6805**	**0.6604**

It is clear to find that $tf.rf$ performs consistently and significantly better than other term weighting methods as the feature set size increases and it achieves the best performance in all experiments in terms of micro-averaged breakeven point, i.e. 0.6805. On the other hand, $binary$ performs consistently the worst among these four term weighting methods. The tf and $tf.idf$ method perform comparable to each other and better than $binary$ all along.

Since Joachims [16] conducted experiments on the same corpus by using $tf.idf$ and SVM in terms of micro-averaged breakeven point, it is easy to compare the two results. Based on the comparison between Joachims' results and our study, several observations are worth discussion.

- Our linear SVM gives a 68.05% micro-averaged breakeven point vs 60.7% for Joachims' linear SVM and 66.1% for his radial basis function with $\gamma = 0.8$. The observation that linear SVM outperforms other non-linear SVMs has already been supported by many researchers [15], [17] and our previous studies.
- The performances of $tf.idf$ in two experiments are almost identical, i.e. 65.78% for our linear SVM and 66.1% for his radial basis function SVM. This difference is not significant.
- Last but not least, $tf.rf$ performs significantly the best among these four methods in our experiment. Moreover, it outperforms $tf.idf$ in Joachims' experiments whether using linear SVM or non-linear SVMs.

Note that although our experiments use the same corpus and same evaluation measure as Joachims', there are minor differences in data preparation, such as stemming, stop words lists and feature selection measures. Joachims used *information gain* for feature selection, while we used χ^2 instead. The difference is not significant thus the comparison between the two experiments is reasonable.

2) Results on the BioCreAtIvE II Corpus: The result on the BioCreAtIvE II corpus is similar to that of Ohsumed corpus. Table III depicts the results of four term weighting methods on the BioCreAtIvE II training corpus using 5-folder cross validation. Again $tf.rf$ is the best term weighting method among all these methods and $binary$ is the worst method. The difference between tf and $tf.idf$ is not significant.

TABLE III
RESULTS OF THE FOUR TERM WEIGHTING SCHEMES ON THE BIOCREATIVE II CORPUS

Scheme	Micro-P	Micro-R	Micro-F1
$binary$	92.55 ± 0.94	91.99 ± 4.05	92.22 ± 1.77
tf	92.34 ± 1.23	94.26 ± 3.26	93.17 ± 1.43
$tf.idf$	92.19 ± 1.01	94.48 ± 3.69	93.28 ± 1.64
$tf.rf$	92.23 ± 1.24	95.11 ± 2.79	93.63 ± 1.23

Based on this result, we adopted the $tf.rf$ method on the test corpus in this BioCreAtIvE II competition. The recent released preliminary result shows that our method performs rather encouraging. We will report the detailed result later.

3) Results on the 18 Journals Corpus: Figure 2 shows the micro-averaged F_1 scores of the four methods on the top 10 categories. The result with respect to micro-averaged

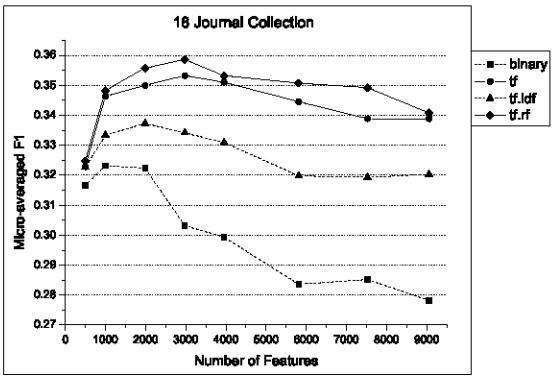

Fig. 2. F1 value of the four term weighting methods in the top 10 categories in 18 Journals corpus

breakeven point on the 18 Journals corpus is consistent with those on the two previous corpora. Again $tf.rf$ is the best term weighting method among these four methods. Again $binary$ is the worst method. However, tf performs better than $tf.idf$ and but still worse than $tf.rf$ with respect to micro-averaged breakeven point. These methods have shown consistent performance with respect to each other as the feature set size increases.

Moreover, we explored the performance of these term weighting methods on three different sizes of subsets of 18 Journals corpus as shown in Figure 3. Although the absolute

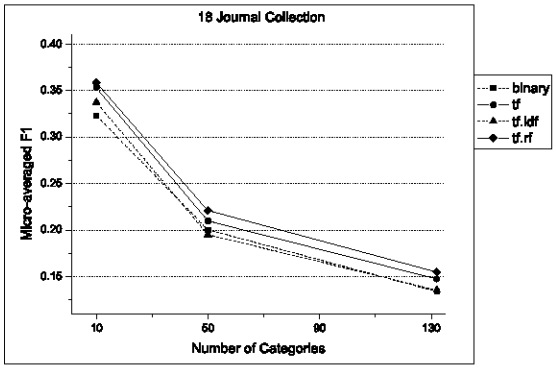

Fig. 3. Micro-averaged F_1 value of different number of categories in 18 Journals Data Collection

performance levels are not significant, their difference is, since this is somehow indicative of the relative "hardness" of these subsets, and allows us to compare these term weighting methods on different subsets. The fact that the subset consisting of the top 10 categories turns out to be the easiest subset is quite obvious, given that its categories are the ones with the highest number of positive samples. With the increase of "hardness" of subsets, the micro-averaged F_1 performance is decreasing. However, $tf.rf$ consistently has the best performance and tf also performs rather well. The performance of $binary$ and $tf.idf$ is the worst all along and the difference between them is not significant as the number of categories increases.

B. Performance of Named Entity-based Representation for Text Categorization

Based on the consideration that protein named entity-based representation may capture more information left out from the bag-of-words approach, we conducted experiment using alternative term type on the BioCreAtIvE II corpus.

The noticing phenomena of these extracted named entities are sparse and skewed distribution in the corpus. First, most of the named entities are in the positive category (76.7%) and only 23.3% are in the negative category. This is reasonable since the positive documents are relevant to protein protein interaction articles and thus they must contain more protein named entities than those in the negative category. Second, most of the named entities occur only once or few times in the corpus. For example, 25740 named entities (83.7%) occur only once in the corpus. 2529 named entities (8.2%) occur more than three times and only 380 named entities (1.2%) occur more than ten times in the whole corpus. This sparse distribution problem make the indexing of documents difficult since many documents will be represented as null vectors when the number of named entities used for indexing is quite small. Based on this consideration, we selected different number of named entities from positive and negative category. Specifically, the selected named entities occur at least ten times in the positive category and at least six times in the negative category.

Furthermore, we also combined named entity-based representation with the bag-of-words approach based on different term weighting methods. Table IV shows the results of these combined different representations, where NE denotes named entity and BOW means the bag-of-words approach.

TABLE IV
RESULTS OF DIFFERENT COMBINED REPRESENTS ON THE BIOCREATIVE II CORPUS.

Scheme	Micro-P	Micro-R	Micro-F1
NE(tf)	68.03 ± 0.81	92.98 ± 2.76	78.56 ± 1.28
NE+BOW($binary$)	91.51 ± 0.94	92.98 ± 4.05	92.20 ± 1.77
NE+BOW(tf)	91.90 ± 1.19	94.74 ± 2.76	93.27 ± 1.35
NE+BOW($tf.rf$)	91.97 ± 1.19	95.16 ± 2.76	93.52 ± 1.35

Based on the results from Table III and Table IV, we can find that named entity-based representation was the most disappointing. It only achieved 78.56% F_1 score. When combined with bag-of-words approach based on different term weighting methods, the named entity-based representation has not increased the performance of text categorization. The performance of the named entity-based representation also supports the conclusions of most past research that phrases do not add much classification power. On the other hand, the bag-of-words approach alone performs significantly better than named entity-based representation alone. This once again supports the superiority of bag-of-words approach. These findings are consistent with the previous work reviewed in Section II.

V. CONCLUDING REMARKS

This paper examined different representations for biomedical text categorization from the term type and term weighting aspects.

The term weighting methods result in the improvement in the classification performance. Specifically, our proposed $tf.rf$ method once again shows classification power in biomedical domain in this study and in newswire domain.

On the other hand, named entity-based representations have no improvement over the bag-of-words approach. This supports the general conclusion that simple NLP-based representations do not improve the performance of text classifier. As the researchers [19] believed that significant advances must be made before NLP techniques can be used to improve text classification.

We should point out that the observations above are made based on the controlled experiments. The accuracy of extracted named entities also have an effect on text classification. Incorporating named entities with higher accuracy, much more categories and more sophisticated usage might be able to improve the effectiveness. We still believe that it is worth to try more advanced NLP techniques and advanced ways of incorporating NLP output to further improve the performance of text categorization, for example, high performance coreference resolution to normalize the protein names through different variations, nominal or pronominal expressions could generate more occurrences of the same protein names to facilitate the further text classification.

ACKNOWLEDGMENT

The work is partially supported by a Specific Targeted Research Project(STREP) of the European Union's 6th Framework Programme within IST call 4, Bootstrapping Of Ontologies and Terminologies STrategic Project(BOOTStrep).

REFERENCES

[1] Lan, Man, Tan, ChewLim. and Low, HweeBoon. Proposing a New Term Weighting Scheme for Text Categorization. 2006. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI2006)*. Page 763-768.

[2] Debole, F., and Sebastiani, F. 2003. Supervised term weighting for automated text categorization. In *Proceedings of the 2003 ACM symposium on Applied computing*, 784–788. ACM Press.

[3] Pascal Soucy and Guy W. Mineau. Beyond tfidf weighting for text categorization in the vector space model. In *IJCAI*, pages 1130–1135, 2005.

[4] Zhi-Hong Deng, Shi-Wei Tang, Dong-Qing Yang, Ming Zhang, Li-Yu Li, and Kun Qing Xie. A comparative study on feature weight in text categorization. In *APWeb*, volume 3007, pages 588 – 597. Springer-Verlag Heidelberg, March 2004.

[5] Zhou Guo Dong and Su Jian. Exploring deep knowledge resources in biomedical name recognition. In *Proceedings of 2004 Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA'2004 shared task)* 99–102, Aug 28-29, 2004, Geneva, Switzerland.

[6] Norbert Fuhr, Stephan Hartmann, Gerhard Knorz, Gerhard Lustig, Michael Schwantner, and Konstadinos Tzeras. AIR/X – a rule-based multistage indexing system for large subject fields. In André Lichnerowicz, editor, *Proceedings of RIAO-91, 3rd International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 606–623, Barcelona, ES, 1991. Elsevier Science Publishers, Amsterdam, NL.

[7] Hinrich Schutze, David A. Hull, and Jan O. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Research and Development in Information Retrieval*, pages 229–237, 1995.

[8] Kostas Tzeras and Stephan Hartmann. Automatic indexing based on bayesian inference networks. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 22–35, New York, NY, USA, 1993. ACM Press.

[9] Sam Scott and Stan Matwin. Feature engineering for text classification. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, pages 379–388, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[10] Ron Papka and James Allan. Document classification using multiword features. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pages 124–131, New York, NY, USA, 1998. ACM Press.

[11] Athanasios Kehagias, Vassilios Petridis, Vassilis G. Kaburlasos, and Pavlina Fragkou. A comparison of word- and sense-based text categorization using several classification algorithms. *J. Intell. Inf. Syst.*, 21(3):227–247, 2003.

[12] David D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–50, New York, NY, USA, 1992. ACM Press.

[13] Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 24(5):513–523.

[14] Eui-Hong Han, George Karypis, and Vipin Kumar. Text categorization using weight adjusted k-nearest neighbor classification. In *PAKDD '01: Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 53–65, London, UK, 2001. Springer-Verlag.

[15] Dumais, S.; Platt, J.; Heckerman, D.; and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, 148–155. ACM Press.

[16] Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. In Nédellec, C., and Rouveirol, C., eds., *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, 137–142. Chemnitz, DE: Springer Verlag, Heidelberg, DE.

[17] Yang, Y., and Liu, X. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 42 – 49. ACM Press.

[18] Chang, C.-C., and Lin, C.-J. 2001. *LIBSVM*: a library for support vector machines. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[19] David Lewis and Spark Jones Karen. Natural language processing for information retrieval. In *Communications of the ACM* 39(1):92–101, 1996.