The Effect of Feature Representation on MEDLINE Document Classification

Meliha Yetisgen-Yildiz, M.S.,[1] Wanda Pratt, Ph.D.[1,2]

[1] The Information School, University of Washington, Seattle, USA.
[2] Biomedical and Health Informatics, School of Medicine, University of Washington, Seattle, USA.

Abstract

This work explores the effect of text representation techniques on the overall performance of medical text classification. To accomplish this goal, we developed a text classification system that supports the very basic word representation (bag-of-words) and the more complex medical phrase representation (bag-of-phrases). We also combined word and phrase representations (hybrid) for further analysis. Our system extracts medical phrases from text by incorporating a medical knowledge base and natural language processing techniques. We conducted experiments to evaluate the effects of different representations by measuring the change in classification performance with MEDLINE documents from the OHSUMED dataset. We measured classification performance with information retrieval metrics; precision (p), recall (r), and F1-score (F1). In our experiments, we achieved better classification performance with the hybrid approach (p=0.87, r=0.46, F1=0.60) compared to the bag-of-words approach (p=0.85, r=0.44, F1=0.58) and the bag-of-phrases approach (p=0.87, r=0.42, F1=0.57).

Introduction

With the rapid growth of online documents available, information organization and retrieval have become a great problem, and automated text classification has been promoted as one method to help with this problem. Automatic text classification is the process of labeling unstructured text with topic categories from a predefined set. The main approach to text classification is based on machine learning, where a general inductive process automatically builds a classifier by learning the characteristics of the categories from a set of pre-classified documents.

The problem of text classification within the medical domain is both an important and challenging one. MEDLINE, the primary source for medical literature contains over 13 millions online entries and over 2000 entries are added each day [1]. MEDLINE documents are manually categorized under 22,568 Medical Subject Headings (MeSH) category names by experts from the National Library of Medicine (NLM) [2]. Because the medical literature is contextually rich and grows rapidly, manual categorization is necessarily time consuming. As a solution to this problem, designing tools that automate the process of classifying MEDLINE documents has drawn great research interest.

The classification of MEDLINE documents is particularly challenging along at least two dimensions. First, each document is typically assigned to many categories. In our examination of the documents published in MEDLINE in 2003, each document averaged 12 descriptive MeSH terms. The minimum number of descriptive MeSH terms used was 1, and the maximum was 51. In contrast, for most other standard document classification collections, such as for Reuters or Yahoo!, most documents are classified into one or two categories. Second, the controlled vocabulary for MeSH is huge and detailed. MeSH contains 22,568 category names in hierarchy of depth eleven. In contrast, the Reuters-21578 collection, one of the standard collections of text classification research, contains only 672 category names [3], as is the case for many other document collections.

In this paper, we explore the effects of different text representation approaches on the classification performance of MEDLINE documents. The first step in text classification is to transform text data into a representation that is suitable for classification methods to use. Although text representation has a direct effect on the classification performance, questions remain about the optimal representation, particularly for domain-specific text classification, such as for MEDLINE. To make a comparison between different representation approaches, we developed a text classification system that supports the very basic word representation, the more complex medical phrase representation, and a hybrid representation that incorporated both basic words and medical phrases. To identify medical phrases our system incorporates natural language processing (NLP) techniques and medical domain knowledge. We conducted experiments to evaluate the effects of the text representation by measuring the change in classification performance with MEDLINE documents from the OHSUMED collection [4]. We used only document titles in our experiments and found interesting differences in performance among the text representations.

Related Work

One of the main efforts of the research in automated text classification area has been to adopt and enhance machine learning algorithms, such as decision trees, nearest neighbor, Rocchio, naïve-bayes, neural networks, and Support Vector Machines (SVM) [5]. Another focus has been to compare the performance

of existing classifiers [6,7]. From this body of research, the high dimensional nature of text data has been shown to be the main reason for the bad performance of many classifier methods. After investigating the classifier options listed above, we decided to use SVM in our experiments because its performance has been superior for high dimensional feature representations, such as those necessary for text [8].

Although text representation is another potential area that can affect the overall classification performance, it has not received as much attention as algorithm performance. The common approach taken by many researches is to represent the text with all the individual words that appear in documents, often referred to as the bag-of-words representation. In a number of studies [6,9], it has been reported that representations more sophisticated than bag-of-words do not yield better effectiveness. However, none of these studies focused exclusively on medical documents.

As an alternative to bag-of-words, Wilcox et.al, used medical domain knowledge and a natural language processor to extract medical concepts and to use those concepts as the features in classifying radiology reports [10]; they found that using domain knowledge increased the performance of their classification methods. Their classification problem was to classify radiology reports into six predefined clinical conditions. Even with a training set composed of 200 documents they achieved very good classification performance (>80%) which suggests the boundaries of their categories are very strictly defined. In MEDLINE, many MeSH categories are too broad and general, such as *Immunologic Diseases*, for a classifier to learn with a small set of training documents.

Mao and Chu developed a representation based on medical phrases [11]. They conducted retrieval experiments with OHSUMED dataset and reported that the phrase-based representation was superior to bag-of-words representation for vector space model. But they have not tested their representation for machine learning algorithms used for text classification.

System Architecture

We developed a text classification system to investigate the effects of different representation approaches on the classification performance. A high-level view of our prototype is illustrated in Figure 1. In the following sections, we describe, in detail, each of the major steps in the classification process.

Selecting Training and Test Titles

For a given set of titles, our system randomly picks 80% of the titles as the training titles and the rest as the test titles. We used OHSUMED test collection in our experiments. The section on experiments and results includes a brief description of OHSUMED and a detailed explanation of our experiment setup with this test collection.

Representing Text

Our system represents the training and the test titles with the following three representation approaches.

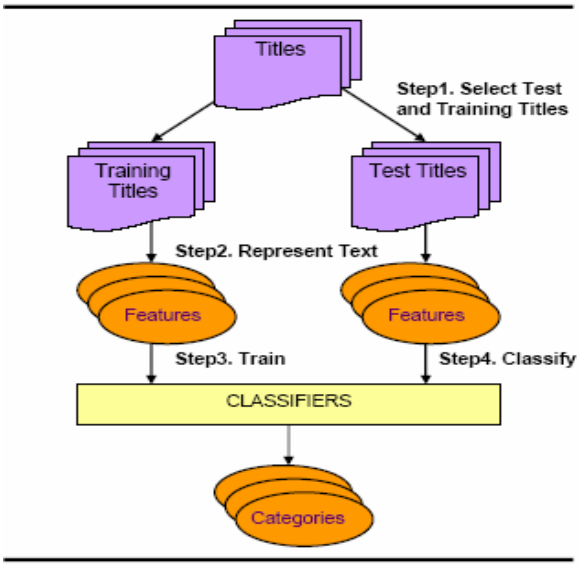

Figure 1. High-level view of text classification prototype

Bag-of-words Representation: Information retrieval research suggests that words work well as representation units for retrieving documents [9]. In the bag-of-words representation, each distinct word corresponds to a feature with a weight as its value that is correlated to the number of times the word occurs in the document. For each training and test title, our system first extracts all the words in the title, eliminates the stopwords by using the Princeton English Stopword List [12], and represents the title with the remaining words and their weights. We observed that words usually do not appear more than once in a title and used a binary weighting approach in our representation. The vector representation of a title has 1 as the weight of a word if the word appears in the title and 0 otherwise.

Bag-of-phrases Representation: The previous representation approach has two challenges. First, the percentage of long phrases, such as calcium channel blockers, in medical vocabulary is very high. This high prevalence of phrases represents a problem for text classification. For example, the meaning of *calcium channel blockers* is very different then that of *calcium* alone. Second, synonymy is a very common characteristic among the medical phrases. For example, medical researchers interchangeably use *heart failure* and *cardiac failure* even in the same docu-

ments. If not grouped explicitly, synonymous words or phrases are represented as different features in the feature vector, which leads to two major drawbacks. The first drawback is that the increase in the dimensionality of feature space is known to have a negative effect on the classification performance. The second drawback is that information is lost due to feature splits. Instead of having a stronger feature, the representation has multiple relatively weaker features that are synonyms of each other. For example, in the OHSUMED, *heart failure* appears in 982 titles and *cardiac failure* appears in 54 titles. We tried to handle these challenges with bag-of-phrases representation.

To identify biomedical phrases, our system uses a knowledge-based, natural-language-processing approach to process the document titles. A key part of our approach is to use a knowledge base to help to identify domain-specific terms. The biomedical domain already has a large publicly available knowledge base called the Unified Medical Language System (UMLS) [13]. In the latest version of UMLS (2005AA), there are over 1 million biomedical concepts as well as over 5 million concept names. NLM created this knowledge base by unifying hundreds of other medical knowledge bases and vocabularies to create an extensive resource that provides synonymy links as well as parent-child relationships among single or multi-word concepts.

To identify the medical phrases, we used MetaMap (Version 2.3.C), a tool created by NLM that maps from free text to biomedical concepts in the UMLS [14]. MetaMap uses the Xerox tagger to assign syntactic parts of speech and then uses the tags to identify phrases. It uses the UMLS to find the closest matching known concept to each identified phrase in the free text.

Our system sends each title from both training and test sets to MetaMap and stores all concepts that MetaMap identifies in a data field linked to the title in a database table. It also uses the UMLS to group synonymous concepts.

Hybrid Representation: Lewis has reported that although representations based on phrases have superior semantic qualities, they have inferior statistical qualities with respect to word representations [9]. To have both statistical and semantic advantages, we combined bag-of-words representation with bag-of-phrases representation for further analysis. For each title, we took the union of features from bag-of-words and bag-of-phrases representations and eliminated the duplicate ones. We also applied the same binary weighting approach that we used in the previous two representations.

Classifying Documents

After representing the titles with bag-of-words, bag-of-medical-phrases, and hybrid approaches, we trained our classification method with the training titles and tested its performance with the test titles. We picked Support Vector Machines (SVM) as our classification method due to its superior performance with text compared to other methods.

SVM's are based on the Structural Risk Minimization principle from computational learning theory [5]. They are linear classifiers that try to find a hyperplane that maximizes the margin between the hyperplane and the given positive and negative examples. For our text classification case, a medical document can be assigned to more than one MeSH category; thus, this problem can be viewed as a series of binary classification problems, one for each category, rather than as a multi-class classification problem.

In our system, we used the latest version of SVM-Light, a very commonly used implementation of SVM developed by Thorsten Joachims [8]. Our system generates a training and test set for each category by labeling all the training or test titles in the category as positive examples and the rest of training or test titles as negative examples. Then it trains the classifiers with the training sets, classifies each test set with the corresponding trained classifier, and measures the overall performance by calculating precision, recall, and F_1-Score metrics for the classification results.

Experiments and Results

We evaluated the effect of representation techniques on the performance of classification by measuring the change in the ability of the classifier to reproduce the manual MeSH assignments in the test set.

Dataset

We used the OHSUMED dataset for our experiments. The OHSUMED dataset, which was created for the TREC conference, has become an evaluation benchmark in automatic text classification research since 1994 [4]. OHSUMED is composed of 348,566 MEDLINE documents from 270 journals published between 1987-1991. The documents are classified under 14,321 MeSH categories.

Because OHSUMED includes only a small portion of MEDLINE, many categories have too few documents to train classifiers. For example, there are 753 categories with only 1 document in OHSUMED. To overcome this problem, many other researchers have limited their classification experiments to a small set of categories that have a minimum number of positive training titles. As an example, Ruiz et al. [15] and

Lewis et al. [7] used 49 categories related to heart diseases with at least 75 training documents.

Disease-related categories have been commonly used in classification experiments because many of them provide more training documents compared to other categories in OHSUMED. In our experiments, we followed a similar but more expansive approach. Rather than using a small subset of disease related categories, we identified all the MeSH categories grouped under the UMLS semantic type Disease or Syndrome. There were 1,928 MeSH categories in the retrieved set, and the number of distinct documents categorized under these disease categories was 179,796. Our system randomly picked 143,837 documents (80%) as the training titles and the rest 35,959 (20%) as the test titles. After creating the training and test sets, our system trained the classifiers for the categories with at least 75 positive training titles. There were 634 such disease categories, which is still considerably more categories than most other researchers have used.

Performance Metrics

We evaluated the category assignments of our binary classifier by using precision and recall performance metrics. To calculate these metrics, we first found the values of the following parameters for each classifier:

- TP: count of the documents correctly assigned to the MeSH category
- FP: count of the documents incorrectly assigned to the MeSH category
- FN: count of the documents incorrectly rejected from the MeSH category
- TN: count of the documents correctly rejected from the MeSH category

Then, we calculated precision and recall for each of the classifiers with the following formulas:

- $Precision_i = \frac{TP_i}{TP_i + FP_i}$

- $Recall_i = \frac{TP_i}{TP_i + FN_i}$

We used micro-averages of each of the metrics for summarizing the results in a more compact form. We computed the micro-averaged precision and recall with the following formulas:

- $Precision^{\mu} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)}$

- $Recall^{\mu} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)}$

where n is the number of classifiers.

We also calculated the F_1-Score that combines micro-averaged recall and precision under one value. F_1-Score is the harmonic mean of precision and recall.

- $F_1 = \frac{2 \times Precision^{\mu} \times Recall^{\mu}}{Precision^{\mu} + Recall^{\mu}}$

Results

Table 1 includes the micro-averaged precision and recall values, and F_1-Score values for each representation approaches.

Table 1. Micro-averaged precision, micro-averaged recall and F_1-Score

	Precision	**Recall**	**F_1-score**
Bag-of-words	0.85	0.44	0.58
Bag-of-phrases	0.87	0.42	0.57
Hybrid	0.87	0.46	0.60

As can be seen from Table 1, SVM generated 2% better precision values with bag-of-phrases and hybrid representations than with bag-of-words representation. In terms of recall, SVM generated the best recall value with hybrid representation. The recall with hybrid representation was 4% better than with bag-of-phrases representation and 2% better than with bag-of-words representation. In terms of F_1-Score, SVM generated the best value again with the hybrid approach. The confidence intervals for precision and recall are listed in Table 2.

Table 2. Confidence intervals for micro-averaged precision and recall (95%)

	Precision	**Recall**
Bag-of-words	0.8497 – 0.8583	0.4528 - 0.4616
Bag-of-phrases	0.8697 – 0.8781	0.4174 - 0.4260
Hybrid	0.8645 – 0.8725	0.4591 - 0.4678

We also performed 5-fold cross validation for 10 randomly selected categories. SVM performed the best with the hybrid approach (Table 3).

Table 3. 5-fold validation results for 10 randomly selected categories

	Precision	**Recall**	**F_1-score**
Bag-of-words	0.82	0.41	0.55
Bag-of-phrases	0.83	0.40	0.54
Hybrid	0.84	0.42	0.56

While investigating the precision and recall values calculated for each classifier, we noticed that some

classifiers were absolute rejectors. In other words, absolute rejectors cannot distinguish positive test documents from negative ones; they label all test documents as negatives. In addition, precision cannot be calculated for such rejector classifiers because both the nominator and the denominator values in the precision formula are equal to 0. The number of absolute rejectors with the bag-of-words representation is 65. This number falls to 45 with the bag-of-phrases representation, and 36 with the hybrid representation. The decrease in the number of absolute rejectors can be explained by the semantic richness of the phrases. With phrases, the classifiers' ability to identify the boundary between positive and negative training documents increases so that the trained classifiers can classify the test documents more precisely. We found that very general and broad categories, such as *nervous system diseases*, *vascular diseases*, and *infection*, often resulted in absolute rejectors. Although each of our classifiers was trained with at least 75 positive training examples, having any rejector classifiers indicated that there were not sufficient training examples in OHSUMED to capture the broad context of some of the very general categories.

Conclusions

In this paper, we presented the effects of using different representation approaches on the overall performance of medical document classification. Although we only used titles in the classification of MEDLINE documents, in general the results were surprisingly good. In the best case, our classifier was able to classify the documents with 87% precision and 46% recall. As part of our future plans, we will use abstracts to represent documents.

Although many researchers have reported that representations more sophisticated than bag-of-words do not yield better effectiveness, we demonstrated that the classification performance of the semantically rich hybrid approach outperformed the bag-of-words approach for MEDLINE. One of the reasons for this performance increase was that we used an NLP tool that was designed purely for medical phrase identification and a medical knowledge-base in our bag-of-phrases representation.

In this paper, we have reported the results from our classification experiments with the OHSUMED dataset. With absolute rejector classifiers, we have shown that the OHSUMED dataset does not include enough training data for many categories, especially for the more general and broad ones. We plan to continue to our work with larger and more representative datasets gathered from MEDLINE.

Because vast numbers of documents are now available in digital format, highly effective text classification systems have become critical. Our experimental results suggest that using semantically rich text representations is one of the potential ways to increase the overall classification performance and create such highly effective document classifiers.

Acknowledgements

This work was supported by the National Science Foundation under Grant IIS-0133973.

References

1. National Library of Medicine. *MEDLINE Fact Sheet*. Available at: http://www.nlm.nih.gov/pubs/factsheets/medline.html
2. National Library of Medicine. *Medical Subject Headings Fact Sheet.* Available at: http://www.nlm.nih.gov/pubs/factsheets/mesh.html
3. Reuters-21578 Test Collection. Available at: http://www.daviddlewis.com/resources/testcollections/reuters21578
4. OHSUMED Test Collection. Available at: http://trec.nist.gov/data/t9_filtering/
5. Sebastiani , F. *Machine Learning in Automated Text Categorization*. Computing Surveys. 2002. **31**(1): 1-47.
6. Dumais, S. T., Platt, J., Heckerman, D., and Sahami, M. *Inductive learning algorithms for text categorization*. In Proceedings of CIKM. 1998. Bethesda, MD.
7. Lewis, D. D., Schapire, R.E., Callan, J.P., and Papka, R. *Training Algorithms for Linear Text Classifiers*. In Proceedings of SIGIR. 1996.
8. Joachims, T. *Learning to Classify Text Using Support Vector Machines - Methods, Theory and Algorithms*. 2001. Kluwer Academic Publishers.
9. Lewis, D. D. *An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task*. In Proceedings of SIGIR. 1992.
10. Wilcox, A., Hripcsak, G., and Friedman, C. *Using Domain Knowledge Sources to Improve Classification of Text Medical Reports*. In Proceedings of ACM SIGKDD Workshop on Text Mining. 2000.
11. Mao, W., and Chu., W.W. *Free-text Medical Document Retrieval via Phrase-based Vector Space Model*. In Proceedings of AMIA. 2002.
12. English Stopword List. Available at: ftp://ftp.cs.cornell.edu/pub/smart/english.stop
13. National Library of Medicine. *Unified Medical Language System Fact Sheet*. Available at: http://www.nlm.nih.gov/pubs/factsheets/umls.html
14. National Library of Medicine. *Metamap*. Available at: http://mmtx.nlm.nih.gov/
15. Ruiz, M. E., and Srinivasan, P. *Hierarchical Neural Networks for Text Categorization*. In Proceedings of SIGIR. 1999.