MEDICAL TEXTS CLASSIFICATION BASED ON KEYWORDS USING SEMANTIC INFORMATION

Roque López, Javier Tejada, Mikhail Alexandrov

Abstract: *This paper presents a method to classify medical texts based on keywords with the support of additional semantic information. The classification is performed in two phases. In the first phase, keyword sets are extracted for each type of disease presented in the training set. Keywords are ranked according to their semantic relatedness. In the second phase, medical texts are classified basing on the resulting keyword lists. The experimental results proved to be encouraging.*

Keywords: *Text Classification, Semantic Information, Natural Language Processing.*

ACM Classification Keywords: *I.2.7 Natural Language Processing.*

Introduction

Automatic text classification, also known as text categorization, is the task of assigning a text into a set of predefined classes or categories [Sebastiani, 2002]. During the last decades different methods of automatic document classification have been proposed. Text classification is commonly defined as a two-stage process. The first stage deals with learning a classification model on a set of pre-classified documents. At the second stage, the model is used to classify new documents. Most existing algorithms and methods are based on statistical data such as term frequency (TF), term frequency–inverse document frequency (TF-IDF), etc. The classification results based on this information can be enhanced by using some additional information.

Each document from the medical documents set includes data on clinical examinations, diagnoses, treatments, indications, medical monitoring, etc. Medical documents are short texts having lots of keywords in common (e.g.: patients, illness, treatment, etc.). In this context, relying on statistical findings alone does not help to distinguish properly between document categories.

In this work, an alternative solution is proposed, which aims to improve the classification of medical documents taking advantage of the semantic relatedness of keywords. The semantic relatedness data is obtained from the ontology of biomedical concepts UMLS (Unified Medical Language System). To evaluate the performance of the classifier, we used the OHSUMED corpus, a collection of medical documents, where each document is assigned a disease type.

The rest of the paper is organized as follows. In Section 2, the related work is outlined. Section 3 describes the proposed method of automatic medical texts categorization. The experiments and results are presented in Section 4. In Section 5, conclusions are drawn, and the future work items are identified.

Related Work

There is an extensive research on the algorithms of medical text classification. Naïve Bayes [Olszewski, 2003], Neural Networks [Farshchi, 2013], Rocchio Algorithm [Figuerola, 2001], etc., have been widely used in text classification. In most papers, statistical approaches are used [Elberrichi, 2012]. However, recently the interest has increased towards the use of semantic information for the improvement of clinical text classification. In this context, one of the earliest efforts is the work by [Wilcox, 2000]. This paper investigates the application of two knowledge resources (UMLS, a repository of biomedical vocabularies, and NLP, a medical language processor) to improve the classifier performance. The UMLS synonym set is used to enrich the representation of medical

records. [Perea, 2008] presents an automatic text categorization system, which uses the UMLS ontology to expand the set of terms in the training and test collections. The obtained results show that increasing the number of terms significantly improves the performance of categorization systems. [Elberrichi, 2012] proposes a method for clinical documents classification based on the information provided by the medical thesaurus MeSH (Medical Subject Headings). Instead of the standard bag-of-words approach, the document representation based on MeSH concepts is applied, which improves the classifier performance. In [Lakiotaki, 2013], a three-stage architecture is proposed: (1) data recovery and terms extraction, (2) representation and data modeling, and (3) documents classification. The main idea is to take advantage of the UMLS semantic network data. The semantic network provides a categorization of UMLS concepts.

The Proposed Approach

The method takes into account both the statistical data and the semantic relatedness between keywords in a medical document. It includes a training phase and a classification phase. In the first phase, keywords for each type of disease in the training set are extracted and ranked according to their semantic relatedness. In the second phase, we calculate the similarity between the medical text to be classified and the keywords of each disease type. Figure 1 shows the architecture of the proposed approach.

Training Phase

The main purpose of this phase is to automatically extract the most relevant keywords for each type of disease that is present in the training set (manually classified documents). This phase has three modules: preprocessing, keyword extraction and keyword ranking.

A. Preprocessing

This module filters out irrelevant passages from the medical documents that cannot contribute to the training process. It includes three steps: tokenization, stopwords removal and part-of-speech tagging.

- **Tokenization**: In this step, the text is divided into simple tokens such as words, numbers, punctuation marks, etc.
- **Stopwords Removal**: In this step, the most frequent words are removed (i.e. pronouns, prepositions, conjunctions, etc.), which do not convey any important semantics. The punctuation is also eliminated.
- **Part-of-Speech Tagging**: In this step, nouns, adjectives and verbs are selected, which carry most of the semantics [Liu, 2009]. For the experiments in this work, the tagger proposed in [Malecha, 2010] was used.

B. Keywords Extraction

The keywords of a document are the words and phrases that can precisely and compactly represent the content of the document [Jiang, 2009]. Some words, such as patient, infection, treatment, etc., appear frequently in all medical documents and do not provide important information about the class (disease) to which they belong. For this reason, in this module we use the method proposed by [Alvarez, 2009], where the weight of a keyword indicates its importance for a class and becomes discriminant for the other classes. Within this method, a word has more weight for a given class when it appears more times in this class and less in the others.

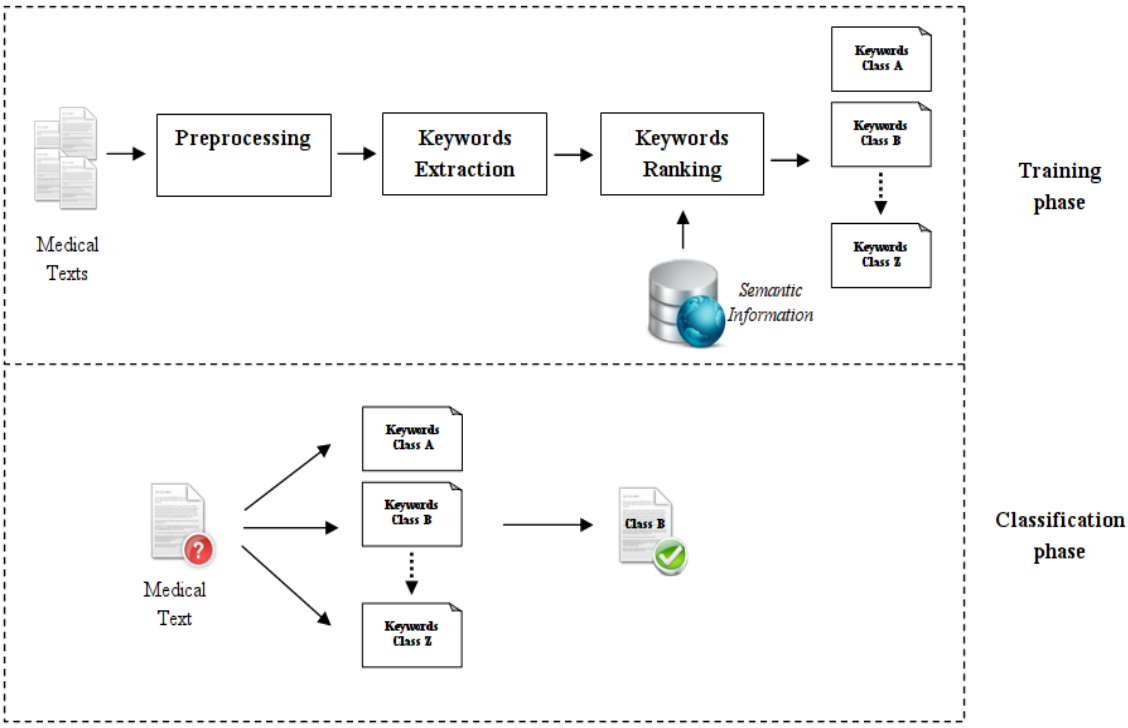

Figure 1: Architecture of the proposed approach

The weight of the word w_{iclass} for c given class is calculated as follows:

$$w_{i_{class}} = tf_i * \log\left(\frac{N_{classes}}{n_{i_{classes}}}\right) \quad (1)$$

where tf_i is the number of medical documents of the class c in which the word w_{iclass} appears. This value is normalized by the total number of documents in the class c; N_{clases} is the total number of classes; and $n_{iclases}$ is the number of classes that have medical documents containing the word w_{iclass}. Based on this statistical information, for each clinical document in the training set we extracted three words with the highest weights as keywords.

C. Keywords Ranking

At the following stage, the three keywords obtained for each medical document in the training set are ranked according to their semantic relatedness. The semantic relatedness considers the relations of all types between two concepts or terms in a taxonomy (i.e. hyponymic, meronymic and any kind of functional relations including *has-part*, *is-made-of*, *is-an-attribute-of*, etc.) [Strube, 2006]. If two concepts or terms tend to occur together more often than usual, their semantic relatedness level is deemed to be higher. For example, the words *endoscopic* and *epigastric* have more semantic relatedness than *endoscopic* and *brain*. We pretend to use this information to get the most similar keywords for each type of disease. To achieve this, we use the semantic relatedness provided by UMLS. UMLS is a widely used database of biomedical terminologies, it includes over 100 terminologies and contains more than 1.7 million active concepts [Liu, 2012].

To rank the keywords, we propose a modification of the PageRank algorithm [Page, 1998]. The PageRank algorithm is used by Google to determine the website importance level. This algorithm builds a graph with websites as nodes, and the input and output links as edges. The PageRank provides a numeric value that represents the relevance of a website on the Internet. In our case, this value represents the importance

of a keyword in the training set. Unlike the original PageRank algorithm, our proposed modification takes into account the weights between nodes, i.e. the semantic relatedness provided by UMLS. In this scenario, the importance of one keyword depends of the keywords that recommend it and the semantic relatedness shared between them. The modified PageRank algorithm is shown in Equation 2:

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{w_{ij}}{|Out(V_j)|} S(V_j) \quad (2)$$

where $S(V_i)$ is the PageRank value for the keyword V_i; d is a damping factor that can be set between 0 and 1; $S(V_j)$ are the PageRank values of each keyword that appears in the same medical document with V_i; $In(V_i)$ is the total number of the input links of keyword V_i; $Out(V_j)$ is the total number of the output links of keyword V_j. The weight of the edge that links keywords V_i and V_j is calculated as follows:

$$w_{ij} = tf_{ij} * UMLS_{V_i V_j} \quad (3)$$

where tf_{ij} is the number of occurrences of keywords V_i and V_j in the same medical document. $UMLS_{V_i V_j}$ is the weight assigned by the UMLS ontology, which corresponds to the semantic relatedness between these keywords.

Classification Phase

We calculated the similarity between a given medical document and keywords of each class (disease). The class with the highest similarity index is assigned to the medical document.

The main reason of calculating the similarity is to have an idea on how many features are shared between a given medical document and the selected keywords, and also on the level of importance of those features. In [Alvarez, 2009], it is denoted as *Heavy Intersection* to this way of comparing documents with classes and is defined as:

$$similarity(d,k) = \sum_{i \in d} w_{i_{doc}} * w_{i_{class}} \quad (4)$$

where d is the document that we want to classify; k is the set of keywords of the class K; w_{iclass} is the weight of keyword i in the class K; w_{idoc} represents the weight of the word i in the document (frequency of the word i).

Experiments

Dataset

For the experiments, we used the corpus OHSUMED [Hersh, 1994], which includes 50,216 medical documents written in English. Usually, the first 10,000 are used for training and the remaining 10,000 - for evaluation. This corpus contains medical documents describing 23 different cardiovascular diseases included in the MeSH vocabulary.

Results

The experiments are conducted to evaluate the utility of the semantic relatedness in clinical text classification. Additionally, we have made a comparison with Naïve Bayes and Rocchio algorithm.

We performed a comparative evaluation of the proposed method against a variation of the same, which does not use semantic information in keywords ranking. The two types of ranking are denominated Simple Ranking and

Semantic Ranking. The Simple Ranking uses the original PageRank algorithm, while the Semantic Ranking uses the modification proposed in this paper (Equation 2) and considers the semantic relatedness extracted from the UMLS ontology.

Table 1: Rankings comparison (5 classes)

	Simple Ranking			Semantic Ranking		
Class	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Cardiovascular (C14)	0.59	0.47	0.53	0.61	0.63	**0.62**
Digestive System (C06)	0.50	0.32	0.39	0.51	0.42	**0.46**
Immunology (C20)	0.63	0.41	0.50	0.64	0.51	**0.57**
Neoplasms (C04)	0.65	0.52	0.58	0.63	0.64	**0.64**
Pathology (C23)	0.45	0.67	**0.54**	0.51	0.55	0.53
Average	0.57	0.48	0.51	**0.58**	**0.55**	**0.56**

The performance evaluation of the proposed classifier is based on Accuracy, Precision, Recall and F-Measure. We have performed experiments using the 5 most frequent diseases in the OHSUMED corpus. Table 1 shows that the Semantic Ranking obtained the best results (0.58, 0.55 and 0.56 for Precision, Recall and F-Measure, respectively). Baseline for these 5 diseases is equal 0.34. With such a baseline these results demonstrate that semantics helps to improve the classifier performance. This improvement has been achieved due to the fact that in the proposed method the terms *gastric*, *esophageal* and *endoscopic* have stronger semantic relatedness and have more relevance to the class *Digestive System* than to the other disease types.

Classifying 23 types of diseases we achieved the accuracies 40.82%, 40.98% and 41.58% for the Rocchio algorithm, Naïve Bayes and the proposed method respectively. Here baseline were equal 16.90%, therefore all the methods showed good results. Some improvement of the results with the proposed method can be explained by more careful selection of keywords as it were described above. As explained in Section 3.1, the proposed method selects only the three most important keywords of each medical document in the training set, while the Rocchio algorithm and Naïve Bayes use all the tokens. Despite using fewer tokens, the proposed classifier ensures better results, which indicates that the tokens selected by the proposed method are representative for the disease classes.

Conclusion and Future Work

In this paper we present a method to classify medical documents, which improves the results of Naive Bayes and Rocchio algorithm. This method, in addition to considering statistical data, takes into account the semantic relatedness between keywords. The development of this classifier has been motivated by the specific features found in the medical texts. The most important points to highlight in this paper are: first, the proposed method ensures acceptable results in automatic classification of medical documents; second, the use of semantic information has proven to enhance the performance of the classifier.

The following is proposed as the future work: (1) use different weights for part-of-speech tags in the keywords ranking; (2) use different similarity measures in the classification phase; (3) run experiments on other datasets.

Bibliography

[Alvarez, 2009] Alvarez J. Clasificación automática de textos usando reducción de clases basada en prototipos. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, México, 2009.

[Elberrichi, 2012] Elberrichi Z., Amel B., Malika T. Medical Documents Classification Based on the Domain Ontology MeSH. arXiv preprint arXiv:1207.0446, 2012.

[Farshchi, 2013] Farshchi S., Yaghoobi M. Categorization of Medical Documents Using Hybrid Competitive Neural Network with String Vector, a Novel Approach. In Intelligence Computation and Evolutionary Computation, Volume 180 of Advances in Intelligent Systems and Computing, pp. 1045-1054, 2013.

[Figuerola, 2001] Figuerola C., Rodríguez G., Berrocal J. Automatic vs Manual categorisation of documents in Spanish. Volume 57, pp. 763-773, 2001.

[Hersh, 1994] Hersh W., Buckley C., Leone T., Hickam D. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 192-201, 1994.

[Jiang, 2009] Jiang X., Hu Y., Li H. A ranking approach to keyphrase extraction. In Proceedings of the 32nd International ACM SIGIR conference on Research and development in information retrieval, pp. 756-757, 2009.

[Lakiotaki, 2013] Lakiotaki K., Hliaoutakis A., Koutsos S., Petrakis E. Towards personalized medical document classification by leveraging UMLS semantic network. In Health Information Science, Volume 7798 of Lecture Notes in Computer Science, pp. 93-104, 2013.

[Liu, 2012] Liu Y., McInnes B., Pedersen T., Melton-Meaux G., Pakhomov S. Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pp. 363-372, 2012.

[Liu, 2009] Liu Z., Li P., Zheng Y., Sun M. Clustering to find exemplar terms for keyphrase extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1, pp. 257-266, 2009.

[Malecha, 2010] Malecha G., Smith I. Maximum Entropy Part-of-Speech Tagging in NLTK. Unpublished course-related report: http://www. people. fas. harvard. edu/gmalecha, 2010.

[Olszewski, 2003] Olszewski, R. Bayesian classification of triage diagnoses for the early detection of epidemics. In Proceedings of the FLAIRS Conference, pp. 412-416, 2003.

[Page, 1998] Page L., Brin S., Motwani R., Winograd T. The Pagerank Citation Ranking: Bringing Order to the Web. In Proceedings of the 7th International World Wide Web Conference, pp. 161-172, 1998.

[Perea, 2008] Perea J., Valdivia M., Ráez A., Díaz M. Categorización de textos biomédicos usando UMLS. Revista Procesamiento del Lenguaje Natural No 40, pp. 121-127, 2008.

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

[Strube, 2006] Strube M., Ponzetto S. Wikirelate! computing Semantic Relatedness using Wikipedia. In Proceedings of the 21st National Conference on Artificial intelligence – Volume 2, pp. 1419-1424, 2006.

[Wilcox, 2000] Wilcox A., Hripcsak G., Friedman C. Using Knowledge Sources to Improve Classification of. Medical Text Reports. In KDD-2000 Workshop on Text Mining, 2000.

Authors' Information

Roque López – *Master Student in Computer Science, Universidade de São Paulo, Avenida Trabalhador São-carlense, 400 – Centro, São Carlos, São Paulo, Brazil;*

e-mail: rlopezc27@gmail.com

Major Fields of Scientific Research: Natural Language Processing, Sentiment Analysis, Opinion Summarization

Javier Tejada – *Professor of Computer Science Department, San Pablo Catholic University; Campus Campiña Paisajista s/n Quinta Vivanco, Barrio de San Lázaro, Arequipa, Perú;*

e-mail: jtejada@itgrupo.net

Major Fields of Scientific Research: Natural Language Processing, Business Intelligence

Mikhail Alexandrov – *Professor, Academy of national economy and civil service under the President of Russia; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain; e-mail: malexandrov@mail.ru*

Major Fields of Scientific Research: Data Mining, Text Mining, Mathematical Modeling