Automatic Keyphrase Extraction from Medical Documents

Kamal Sarkar

Computer Science & Engineering Department,
Jadavpur University,
Kolkata – 700 032, India
jukamal2001@yahoo.com

Abstract. Keyphrases provide semantic metadata that summarizes the documents and enable the reader to quickly determine whether the given article is in the reader's fields of interest. This paper presents an automatic keyphrase extraction method based on the naive Bayesian learning that exploits a number of domain-specific features to boost up the keyphrase extraction performance in medical domain. The proposed method has been compared to a popular keyphrase extraction algorithm, called Kea.

Keywords: Domain specific keyphrase extraction, Medical documents, Text mining, Naïve Bayes.

1 Introduction

Medical Literature such as research articles, clinical trial reports, medical news available on the web are the important sources to help clinicians in patient care. The pervasion of huge amount of medical information through WWW has created a growing need for the development of techniques for discovering, accessing, and sharing knowledge from medical literature. The keyphrases help readers rapidly understand, organize, access, and share information of a document. Document keyphrases provide a concise summary of the document content. Medical research articles published in the journals generally come with several author assigned keyphrases. But, medical articles such as medical news, case reports, medical commentaries etc. may not have author assigned keyphrases. Sometimes, the number of author-assigned keyphrases available with the articles is too limited to represent the topical content of the articles. So, an automatic keyphrase extraction process is highly desirable.

A number of previous works has suggested that document keyphrases can be useful in a various applications such as retrieval engines [1], [2], [3], browsing interfaces [4], thesaurus construction [5], and document classification and clustering [6].

Turney [7] treats the problem of keyphrase extraction as supervised learning task. Turney's program is called Extractor. One form of this extractor is called GenEx, which is designed based on a set of parameterized heuristic rules that are fine-tuned using a genetic algorithm.

A keyphrase extraction program called Kea, developed by Frank et al. [8], uses Bayesian learning for keyphrase extraction task. In both Kea and Extractor, the candidate

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 273–278, 2009.

keyphrases are identified by splitting up the input text according to phrase boundaries (numbers, punctuation marks, dashes, and brackets etc.). Kea and Extractor both used supervised machine learning based approaches. Two important features such as distance of the phrase's first appearance into the document and TF*IDF (used in information retrieval setting), are considered during the development of Kea. Frank et al. [8] compares performance of the kea to Turney's work and shows that performance of Kea is comparable to GenEx proposed by Turney. Moreover, Frank et al. [8] claims that training Naïve Bayes learning technique is quicker than training GenEx that employs the special purpose genetic algorithm for training.

Compared to the previous works, our work differs in several ways: (1) we use POS tagger based noun phrase identification method, (2) our approach exploits a number of new domain-specific features and statistical features for keyphrase extraction and (3) a glossary database has been incorporated to discriminate between more-domain-specific and less-domain-specific phrases.

The paper is organized as follows. In section 2 we discuss how to build and use domain knowledge. In section 3, the proposed keyphrase extraction method has been discussed. We present the evaluation and the experimental results in section 4.

2 Domain Knowledge Preparation

One domain specific vocabulary is built up by using MeSH (Medical Subject Headings), which is NLM's (U.S. National Library of Medicine) controlled vocabulary thesaurus. All the MeSH terms are treated as domain specific key phrases. We prepare one table for medical keyphrases (MeSH terms), which is treated as a glossary database. We use this glossary database in our wok to determine domain specificity of a phrase.

In addition to this glossary database, we use another vocabulary of natural language terms to identify novel medical term. To decide whether a term is novel, we used two vocabularies: glossary database (medical vocabulary) and natural language vocabulary because absence of a word in the medical vocabulary is not a sufficient condition to consider it as a novel term. The vocabulary of natural language words has been constructed from a corpus of natural language texts (not related to the medical domain) downloaded from the site under the heading Yahoo news coverage. If a term is not available in these two vocabularies, we treat the term as novel term.

3 Proposed Keyphrase Extraction Method

The proposed keyphrase extraction method consists of three primary components: document preprocessing, noun phrase identification and keyphrase identification. The preprocessing task includes conversion from the pdf format to text format, formatting the document for removing non-textual content.

3.1 Noun Phrase Identification

We treat the noun phrases in the document as the candidate keyphrases [1]. To identify the noun phrases, documents should be tagged. We use GENIA[1] tagger for

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

tagging the input document. GENIA is part-of-speech tagger for biomedical text [9]. We use GENIA tagger 3.0.1. This version can assign tags to the terms in the medical text and separately mark terms which at the beginning and inside a noun phrase. By checking these markers of the noun phrases, we can easily identify noun phrases.

3.2 Keyphrase Extraction Using Naïve Bayes

To extract keyphrase from the medical documents, the Naïve Bayes classifier is trained on a set of medical documents and author assigned keyphrases available with those documents. Based on the features of the noun phrases discussed below, the classifier is trained to classify the candidate noun phrases as the keyphrases (positive examples) or not (negative example). Preparation of training and test instances, training procedure of the Naïve Bayes classifier are also discussed later in the subsequent subsections.

Features. To characterize the noun phrases in the medical documents we have designed and adopted a number of features discussed in full below:

Phrase Frequency and Positional Information. We adopt these two features used in [8] to our Keyphrase extraction task. Phrase frequency means number of times a phrase P occurs in a document. We also consider position of the first occurrence of the phrase in the document as a feature.

Frequency of Component Words in the Phrase. If a phrase frequency is not very high, but frequency of the individual words of the phrase is very high, we consider this phrase as an important one. The log-normalized value of the sum of the frequency of the component words is considered as a feature value.

Presence of the Phrase in the Glossary Database. We apply the following formula to compute the value of this feature.

G = 1 if the phrase is present in the glossary database
=1/n if the phrase partially matches with MeSH terms, where n is the number of partial matches.

If the number of partial matches increases, we assume that the phrase consists of more common words and the entire phrase is assumed to be less domain specific.

Acronyms and Novel Terms. A phrase gets a score based on the number of novel terms and acronyms it contains. In medical articles, authors frequently use acronyms for important complex medical terms, perhaps it helps them memorize the things better. Following two rules are used to detect novel term and acronym:

If the length of the term is greater than a threshold (5 characters) and the term is not found in any of two vocabularies (discussed in section 2), we consider the term as novel term.

If some letters (at least two letters) of a term are capital, we treat the term as an acronym (gene names, medical instruments etc.). For example, fMRI is a term, which is found in our test document, is recognized as an acronym by this rule.

Average Word Length. We also consider the average length of words in the phrase as a feature.

Keyphrase Identification. Training Naïve Bayesian learning algorithm for keyphrase extraction requires document noun phrases to be represented as feature vectors. Author assigned keyphrases are removed from the original document and stored in the different files with document identification number. For each candidate noun phrase in the given document we extract the feature values from the source document using the measures discussed above. If the noun phrase under consideration is found in list of author assigned keyphrases corresponding to the document, we label the phrase as "Positive" example and if it is not found we label the phrase as "negative" example. Thus the feature vector for each noun phrase looks like {<a_1 a_2 a_3 ….. a_n>, <label>} which becomes a training instance (example) for Naïve Bayesian learning algorithm where a_1, a_2 . . .a_n, indicate feature values for a noun phrase. After preparation of training data set, the Naïve Bayesian learning algorithm is trained on the training set to learn to classify candidate noun phrases as one of two categories: "Positive" (class 1) or "Negative" (class 0).

For our experiment, we use Weka (www.cs.waikato.ac.nz/ml/weka) machine learning tools. To build up the model based on Naïve Bayes learning algorithm, we used Weka's Simple CLI utility, which provides a simple command-line interface that allows direct execution of WEKA commands. Since all the features that we consider in this work are real numbers (feature values are continuous), we use Fayyad and Irani's [10] discretization scheme, which is based on the Minimum Description Length principle (MDL). The trained classifier is applied on the test document. During testing, we use –p option. With this option we can generate a probability estimate (posterior probability) for the class of each vector. This is required when the number of noun phrases classified as positive by the classifier is less than the desired number of the keyphrases. For a given document, if the user specifies that K keyphrases are desired, then we select the K vectors that have the highest estimated probability of being in class 1.

4 Evaluation and Experimental Results

There are two usual practices for evaluating the effectiveness of a keyphrase extraction system. One method is to use human judgment, asking human experts to give scores to the keyphrases generated by the system. Another method, less costly, is to measure how well the system-generated keyphrases match the author-assigned keyphrases. We prefer the second approach [7][8] [11] to evaluate the proposed keyphrase extraction system by computing its precision and recall using author-provided keyphrases for medical documents. In this experiment, precision is defined as the proportion of the extracted keyphrases that match the keyphrases assigned by a document's author(s). Recall is defined as the proportion of the keyphrases assigned by a document's author(s) that are extracted by the keyphrase extraction system.

To train and test our keyphrase extraction system, 75 journal articles have been downloaded from a number of online medical journals. The downloaded research articles are basically available as PDF files. All PDF files are converted to text files. Only the text content is considered, non-textual content is removed. Author assigned keywords are separated from the articles and stored in the different file with document identification number. Out of 75 medical research articles, 50 documents and the associated author provided keyphrases are randomly selected for training and the rest 25 documents are used for testing.

Kea [8] is now a publicly available keyphrase extraction system based on Naïve Bayes learning algorithm. Kea uses a limited number of features such as positional information and TF*IDF feature for keyphrase extraction. We download version 5.0 of Kea[2] and install it on our machine. Then it is trained with the same set of medical documents, which are used to train our system. After training Kea, a model is built based on Naïve Bayes. This pre-built model is used to extract keyphrases from the test set consisting of 25 documents.

We calculate the precision and recall for both systems when the number of extracted keyphrases is 5, 10 respectively. We also conduct the statistical significance test on the difference between precisions of the two systems, as well as their recalls, using a paired t test. From table 1, we can find that, in respect to precision and recall, the proposed system performs better than Kea. The results are significant at 95% confidence level.

Table 1. Precision and Recall for the proposed keyphrase extraction system and Kea in medical domain. P-values in the middle column indicates sigificance test on precision difference and P-values in the last column indicates sigificance test on recall difference.

Number of keyphrases	Average Precision ± SD		p-value	Average Recall ± SD		p-value
	Proposed System	Kea		Proposed System	Kea	
5	0.47± 0.20	0.28±0.21	<0.01	0.57±0.24	0.33±0.24	<0.01
10	0.28±.0.11	0.23±0.11	<0.01	0.66±0.24	0.54±0.22	<0.01

To interpret the results shown in table 1 we should mention some important points: some author-provided keyphrases may not occur in the document they are assigned to. According to Turney [7], about only 75% of author-provided keyphrases appear somewhere in the documents. This implies that the highest possible average recall for a system could only be 0.75, even when all the phrases are extracted from the documents. In our experiment, the average number of author-provided keyphrases for all the documents is only 4.33, so the precision would not be high even when the number of extracted keyphrases is large. For example, when the number of extracted keyphrases for each document is 10, the highest possible average precision is around 0.32475 (4.33 * 0.75/10 = 0.32475).

5 Conclusion

This paper discusses a keyphrase extraction method in medical domain. The proposed method uses Naïve Bayes learning algorithm that exploits a number of domain specific features and a number of statistical features for keyphrase extraction from medical documents. The experimental results also suggest that the proposed keyphrase

[2] http://www.nzdl.org/Kea/

extraction method is effective in medical domain and incorporation of domain specific features boosts up the system performance.

References

1. Wu, Y.B., Li, Q.: Document keyphrases as subject metadata: incorporating document key concepts in search results. Journal of Information Retrieval 11(3), 229–249 (2008)
2. Li, Q., Wu, Y.B., Bot, R., Chen, X.: Incorporating document keyphrases in search results. In: Proceedings of the tenth American conference on information systems, New York (2004)
3. Jones, S., Staveley, M.: Phrasier: A system for interactive document retrieval using keyphrases. In: Proceedings of SIGIR 1999, Berkeley, CA (1999)
4. Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., Frank, E.: Improving browsing in digital libraries with keyphrase indexes. Journal of Decision Support Systems 27(1-2), 81–104 (2003)
5. Kosovac, B., Vanier, D.J., Froese, T.M.: Use of keyphrase extraction software for creation of an AEC/FM thesaurus. Journal of Information Technology in Construction 5, 25–36 (2000)
6. Jonse, S., Mahoui, M.: Hierarchical document clustering using automatically extracted keyphrase. In: Proceedings of the third international Asian conference on digital libraries, Seoul, Korea, pp. 113–120 (2000)
7. Turney, P.D.: Learning algorithm for keyphrase extraction. Journal of Information Retrieval 2(4), 303–336 (2000)
8. Frank, E., Paynter, G., Witten, I.H., Gutwin, C., Nevill-Manning, C.: Domain-specific keyphrase extraction. In: Proceeding of the sixteenth international joint conference on artificial intelligence, San Mateo, CA (1999)
9. Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: Bozanis, P., Houstis, E.N. (eds.) PCI 2005. LNCS, vol. 3746, pp. 382–392. Springer, Heidelberg (2005)
10. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of 13th International Joint Conference on Artificial Intelligence, pp. 1022–1027. Morgan Kaufmann, San Francisco (1993)
11. Jones, S., Paynter, G.W.: Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications. Journal of American Society of Information Science and Technology 53(8), 653–677 (2000)