Journal of Science and Technology Volume 48, Issue 4, 2010 pp. 127-132

USING MEDICAL OBJECTS FOR CLINICAL RECORDS CLASSIFICATION

ANH KHOA HONG, MINH CHAU NGUYEN, BAO QUOC HO

ABSTRACT

In this paper, medical objects are used as features to classify clinical records. Medical objects such as disease names, drug names, symptoms, examination indicators are extracted using an Unstructured Information Management Architecture (UIMA) based system. The extracted medical objects will be used against the "bag-of-words" as the features of the clinical record in some classification algorithms. The results show that the precision of the classification results using medical objects is better in all algorithms, suggesting that medical objects contribute a significant part to the semantic of a clinical record.

Keywords. information extraction, healthcare informatics

1. INTRODUCTION

Clinical records (or clinical notes) are records about real medical treatments of patients. They provide useful information which can be utilized in many different ways. However, clinical records are often written in free, natural language, with an informal fashion: they can contain incomplete sentences, inverted constructions, misspellings, etc[1]. Those attributes make the useful information contained in clinical reports difficult to be extracted by means of automatic tools, thus limiting the potential of applying analyzing algorithms to this domain's problems.

Our goal is to extract the medical objects appeared in clinical reports. Medical objects are terms in the clinical records which are related to the medical domain such as disease name, drug name, etc. These medical objects may contribute a significant part to the semantic of the clinical record in which they are contained.

Our system is based on the Apache implementation of the Unstructured Information Management Architecture (UIMA) [2], an architecture for building flexible and modular unstructured data processing system. To demonstrate the contribution of medical objects to the semantic of the clinical records, we have tested many types of features and many classification algorithms on the data set of the 2008 i2b2 obesity challenge [3], each algorithm was executed in two conditions: using the medical objects versus using words in the clinical record as the features for classification. The results of each algorithm in each condition will be compared to find out if using medical objects yields a better result.

2. RELATED WORK

In order to improve automated text classification, researchers have a tendency to concentrate on adopting and enhancing machine learning algorithms such as decision trees, Naive Bayes [4], or focus on comparing the performance of existing classifiers [5] to find out the most appropriate ones. Meanwhile, text representation, which is another potential area that can affect the overall classification performance, hasn't received much attention. The prevalent approach taken by many researches is to represent the text with all the individual words in documents, often known as the bag-of-words representation. Additionally, Dumais et al [5] stated that representations more sophisticated than bag-of-words don't provide better effectiveness. However, this study didn't focus exclusively on medical documents. Conversely, Jimmy Lin et al [6] assumed that bag-of-words is not enough for strength of evidence classification, especially in biomedical field, although their experimental results weren't indeed obvious. Furthermore, as an alternative to bag-of-words, Wilcox et al [7] proposed the use of medical domain knowledge and a natural language processor to extract medical concepts and use those concepts as the features in classifying radiology reports; they found that using domain knowledge increased the performance of their classification methods.

For medical object extraction, there are various approaches for this task. According to Leser et al [8], these approaches can be classified into three separate groups: dictionary based, rule based, and machine learning techniques. In dictionary based approaches, the text is matched against a fixed lexicon. Even though the precision is high, the recall is sometimes very low as new disease and drug names, for example, are presented. Rule based approaches are usually hand crafted by experts. The disadvantage of rule-based approaches is that the process is time consuming and such approaches have difficulties handling unseen name patterns. Various machine learning techniques have also been applied to solving this problem such as Support Vector Machine (SVM), Naive Bayes...

For objects' context analysis, several works have been done in the recent years, especially in negation handling. Aronow et al [9] applied syntactic processing techniques to identify noun phrases as well as determine negation scope. They also use machine learning algorithms to identify the negative patterns in the text. Another method developed by Averbuch et al [10] uses an information gain based selection algorithm to automatically learn negative patterns from the text. Chapman et al [11] proposed a regular expression based negation determination algorithm. Especially, it defines a wide range of negation phrases that either appears before or after a finding in medical domain.

3. OUR APPROACH

3.1. Introduction to UIMA

UIMA defines a common structure to store the analyzed data: the Common Analysis Structure (CAS). The CAS object contains the document to be analyzed and the current analysis result. CAS objects can be serialized and passed between components in the system. Upon receiving the CAS object, each component will read the document and the existing analysis results necessary for their task, then add their own results to the CAS object and send it to another component.

There are three main types of component in UIMA: Collection Readers, Analysis Engines and CAS Consumers. The Collection Readers will connect to the data source, read the collection

of documents to be analyzed, create a CAS object for each document, then send the CAS objects created to a pipeline of Analysis Engines. Each Analysis Engine usually performs one single task such as splitting sentences, tokenizing, etc. Finally, the CAS Consumers read the analysis results in the CAS object and save them to desired persistent sources such as database, XML files, indices, etc.

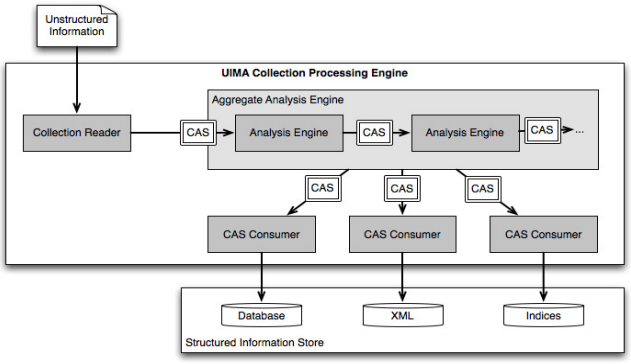

Figure 1. UIMA components and workflow

3.2. Our system model

Based on the UIMA, our system is composed by several UIMA components. A Collection Reader will read the clinical records presented in dataset then create a CAS object for each clinical record read. After that, the CAS object is passed through a pipeline of Analysis Engines; the Analysis Engines will perform natural language processing (NLP) tasks, medical object extraction tasks and then classification tasks. Finally, the CAS Consumers will convert the classification result to the specific XML format for comparison goal.

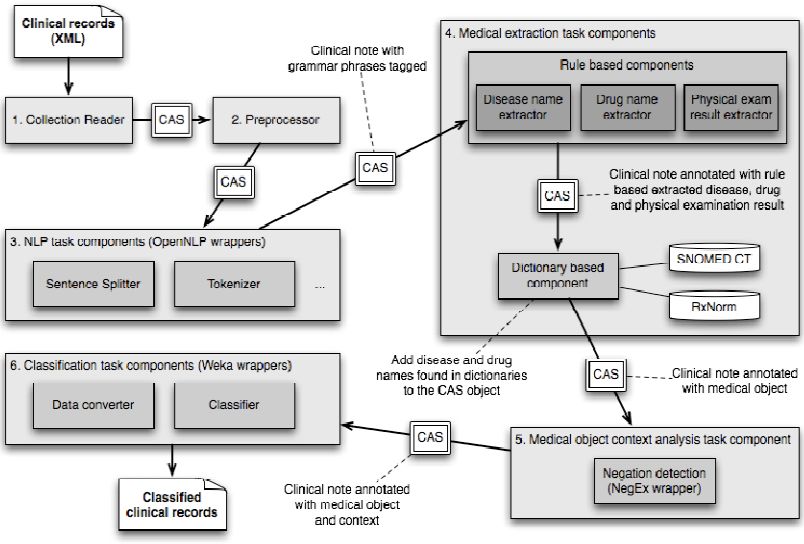

Figure 2. UIMA based classification system model

The NLP tasks is handled by the *UIMA wrapper components for OpenNLP* [12] developed by the Jena University Language & Information Engineering Lab. We used the Sentence Splitter, Tokenizer, Part-of-speech Tagger and Chunker components. After the CAS object passed through these components, we can determine phrases in the clinical record, such as noun phrase, verb phrase, etc. This creates the foundation for the next tasks, as medical objects are usually noun phrases and our extraction rules are mostly based on the phrases.

The medical object extraction task, the main task in our system, is carried out by using rule based method and dictionary based method. The *rule based medical object extraction components* use some simple rules that we have observed in our clinical record data. Our rules can detect disease names, symptoms, drug names and physical examination results appearing in the clinical note. For the dictionary based method, we utilized the *Dictionary Annotator* provided in the Apache UIMA project [13], using a subset of SNOMED CT [14] as the dictionary of diseases-symptoms and a subset of RxNorm [15] as the dictionary of drugs.

The classification tasks is handled by the *Mayo Weka/UIMA Integration (MAWUI) library* [16], using the medical objects extracted in the previous tasks as the features of the clinical record. We also use NegEx [11] to add the negation context to each feature before classifying the clinical record.

4. OUR EXPERIMENTS

In this section, we will present our results and observations. In order to evaluate our system, we conducted the experiment based on the dataset from the I2B2 obesity challenge 2008 [10]. This dataset was divided into two parts: 730 clinical records for training the system and 530 records for testing. In our experiment, besides using medical objects as the clinical records' features to classify the diseases, we also used the "bag of word" features enhanced by removing stop words to evaluate the effect of using medical objects. Three machine learning algorithms: SVM, Naive Bayes and J48 decision tree were applied to the classification.

4.1. Dataset

According to the I2b2 challenge 2008 [10], the dataset is the discharge summaries of patients obtained from different healthcare organizations. Each patient record which underwent a process of de-identification was assigned as present, absent, questionable or unmentioned for each disease in the disease set: Obesity, Diabetes mellitus - DM, Hypercholesterolemia, Hypertriglyceridemia, Hypertension - HTN, Atherosclerotic CV disease - CAD, Heart failure - CHF, Peripheral vascular disease - PVD, Venous insufficiency, Osteoarthritis - OA, Obstructive sleep apnea - OSA, Asthma, GERD, Gallstones, Depression and Gout.

4.2. Results

Here, we will compare the accuracy of disease classification predicated upon "medical objects" features and predicated upon "bag of word" features on three machine learning algorithms: Support Vector Machine (SVM), Naive Bayes, J48 decision tree.

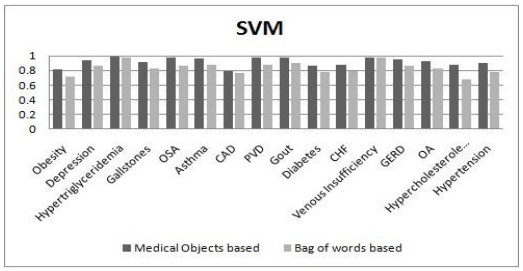

Figure 3. Accuracy of disease classification system based on medical objects versus bag of words on SVM

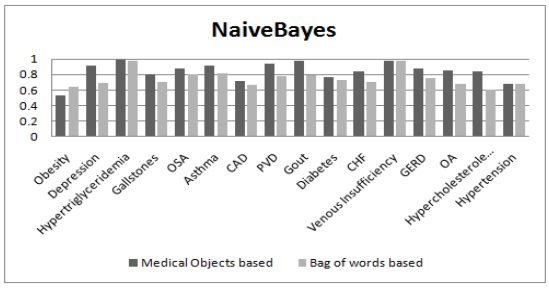

Figure 4. Accuracy of disease classification system based on medical objects versus bag of words on Naïve Bayes

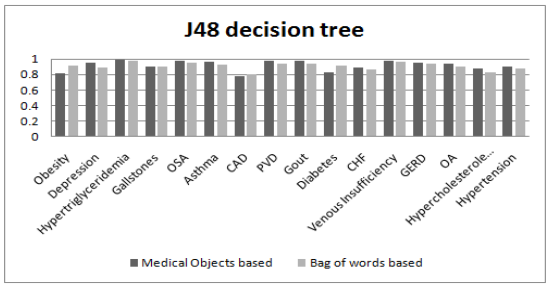

Figure 5. Accuracy of disease classification system based on medical objects versus bag of word on J48 decision tree

The graphs in Figs. 3, 4, 5 summarize the performance of the mentioned algorithms. In these three algorithms, medical objects based classification shows average accuracy higher than bag-of-words based classification. Especially in the SVM classification, using medical objects yields better results in all cases. Nevertheless, in some specific diseases, bag of word classification is predominant, especially in the J48 decision tree case. The results also show that in medical objects based classification, SVM is the best in three algorithms with an average accuracy of 92%; Naive Bayes performed the worst with an average accuracy of 84%. Meanwhile, in bag-of-words based classification cases, J48 is the best with an average accuracy of 88%; Naïve Bayes is the worst with 75%.

5. CONCLUSION

In this paper, we have explored various problems associated with free text processing of clinical reports and have also presented our approach to address some of those. The experimental results demonstrate that medical objects based classification does help in increasing the classification accuracy.

As a future work, we intend to explore other semantic aspects to make better predictions and find ways to utilize the annotated information for other problems in this domain such as retrieval, decision support, etc.

REFERENCES

1. Serguei V. P., Anni C., Christopher G. C. - Developing a corpus of clinical notes manually annotated for part-of-speech, International journal of medical informatics **75** (2006) 418-429.
2. Apache UIMA, http://incubator.apache.org/uima/
3. Informatics for Integrating Biology and the Bedside, https://www.i2b2.org/NLP/Main.php
4. Fabrizio S. - Machine learning in automated text categorization, ACM Comput. Surv. **34** (2002) 1-47.
5. Susan D., John P., David H., Mehran S. - Inductive learning algorithms and representations for text categorization, Proceedings of the seventh international conference on Information and knowledge management, ACM, Bethesda, Maryland, United States, 1998.
6. Jimmy L., Demner-Fushman D. - Bag of Words is not enough for Strength of Evidence Classification, AMIA Annu Symp. Proc., 2005, p. 1031.
7. Wilcox A., Hripcsak G., Friedman C. - Using Domain Knowledge Sources to Improve Classification of Text Medical Reports, Proceedings of ACM SIGKDD Workshop on Text Mining, 2000.
8. Leser U., Hakenberg J. - What makes a gene name? Named entity recognition in the biomedical literature, Brief Bioinform **6** (2005) 357-369.
9. Aronow D. B., Feng F., Croft W. B. - Ad Hoc Classification of Radiology Reports, Journal of the American Medical Informatics Association (1999) 393-411.
10. Averbuch M., Karson T., Ben-Ami B., Maimon O., Rokach L. - Context-Sensitive Medical Information Retrieval, Proc. of 11th World Congress on Medical Informatics. IOS Press, San Francisco, 2004.
11. Chapman W. W., Bridewell W., Hanbury P., Cooper G. F., Buchanan B. G. - A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inform. **34** (2001) 301-310.
12. NLP Toolsuite, http://www.julielab.de/Resources/Software/NLP_Tools.html
13. Apache UIMA - UIMA annotators, http://incubator.apache.org/uima/annotators.html
14. SNOMED Clinical Terms® (SNOMED CT®), http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
15. RxNorm, http://www.nlm.nih.gov/research/umls/rxnorm/
16. Text Analysis - MAWUI, http://informatics.mayo.edu/text/index.php?page=weka.

Address:

Faculty of Information Technology,

University of Science, VNU – HCMC.

Received June 16, 2010