An Approach for Automatic Multi-label Classification of Medical Sentences

Abeed Sarker[1], Diego Mollá-Aliod[1], and Cécile Paris[2]

[1] Centre for Language Technology
Department of Computing, Macquarie University
Sydney, NSW 2109, Australia
{abeed.sarker,diego.molla-aliod}@mq.edu.au,
URL: http://www.clt.mq.edu.au
[2] CSIRO – ICT Centre,
Locked Bag 17, North Ryde, Sydney, NSW 1670, Australia
cecile.paris@csiro.au
URL: http://www.csiro.au

Abstract. We present an approach for the automatic classification of sentences in medical abstracts to a set of predefined categories for use in Evidence Based Medicine. We use Support Vector Machines (SVM) to perform the classification task, and divide the multi-label classification problem into several binary classification problems. Through the use of a mixture of lexical, structural, sequential, and semantic information, our approach outperforms baselines and achieves comparable performance to state-of-the-art classification systems, obtaining micro-averaged f-scores greater than 84% for structured abstracts, and greater than 70% for unstructured abstracts. Our sytem also participated in the 2012 ALTA shared task, and performed significantly better than the benchmark system with an Area Under the ROC (AUC) score of 0.955. Our approach works particularly well for unstructured abstracts and infrequent classes.

Keywords: Evidence Based Medicine, Medical Text Classification.

1 Introduction

Evidence Based Medicine (EBM) practice requires practitioners to combine their clinical expertise with the best available literature-oriented medical evidence, when making decisions about patient care. Consequently, to satisfy EBM guidelines, practitioners have to search for available medical evidence, and synthesise evidence from multiple documents. Due to the plethora of medical documents available in specialised databases (e.g., PubMed[3] has over 22 million articles indexed), searching for appropriate evidence is a challenging task. A popular approach to facilitate the information retrieval process is to label medical text with a set of predefined medical categories, such as the *PICO* criterion [6]: ***P****opulation*,

[3] http://www.ncbi.nlm.nih.gov/pubmed

***I**ntervention*, ***C**omparison*, and ***O**utcome*. Due to the usefulness of such annotation, variants of the PICO criterion have been proposed (e.g., PIBOSO [3]).

In this paper, we address the problem of automatically labeling sentences in medical abstracts using a number of predefined categories. We apply supervised machine learning to solve the classification problem, and choose SVMs as the machine learning algorithm, primarily because of its proven effectiveness in text classification tasks. We show that our SVM-based approach outperforms baselines and performs comparably to benchmark systems for this task, with micro-averaged f-scores of approximately 84% and 71% for structured and unstructured abstracts, respectively. For the 2012 ALTA shared task[4], which was based on this problem, our system significantly outperformed the benchmark system with an AUC score of 0.955. Our approach works particularly well for unstructured abstracts and low-frequency classes.

2 Related Work

The PICO criterion was first proposed by Richardson et al. [6] as a framework for formulating clinical questions. Studies have shown that this framework improves the clarity of clinical problems and results in more precise search results. Primarily because of its usefulness, numerous variants of this framework have been proposed. Research has also focused on automatic methods for annotating medical text according to these criteria. The first approach to classify text into the PICO categories was proposed by Demner-Fushman and Lin [4], who use hand-crafted rules to identify the *Population*, *Intervention*, and *Comparison* elements in sentences. To detect *Outcome* sentences, the authors use a Naïve Bayes classifier with a range of features, including n-grams, and domain-specific information obtained using MetaMap [1]. The MetaMap tool is commonly used to map biomedical concepts to the UMLS[5] metathesaurus, which is a repository of over 1 million biomedical concepts, their semantic categories, and relations. The authors show that the classification of text into PICO elements improves information retrieval performance, and the authors also use this information for summarisation. Following on from this work, Chung [2] proposed the use of rhetorical roles for improving automatic classification accuracy.

A relatively recently proposed variant of PICO is the PIBOSO (Population, Intervention, Background, Outcome, Study Design, and Other) criterion [3]. The authors apply supervised machine learning on a manually annotated data set, and using features derived from context, semantic relations, structure and sequencing of the text, they obtain f-scores of 80.9% and 66.9% for structured abstracts and unstructured abstracts, respectively. Verbeke et al. [9] show the application of a statistical relational learning approach using *kLog* to perform classification on the same data set. The authors obtain classification f-scores of 84.3% and 67.1% for structured and unstructured abstracts, respectively.

[4] `http://alta.asn.au/events/sharedtask2012/`
[5] `http://www.nlm.nih.gov/research/umls/`

As for the machine learning algorithms for these classification problems, Conditional Random Fields (CRF) [7] are used by Chung [2] and Kim et al. [3] for the supervised classification task. The choice of CRFs for this task is influenced by the fact that CRFs are capable of modeling sequential effects and support the use of a large number of features. Another machine learning algorithm that is very popular for text classification, primarily because of its ability to handle very large feature spaces, is SVMs [8]. In our approach, we apply SVMs for the classification task and compare its performance against the benchmark systems proposed by Verbeke et al. [9] and Kim et al. [3].

3 Method

3.1 Data

We perform experiments on the publicly available NICTA-PIBOSO data set that was used for the 2012 ALTA shared task; and also by Kim et al. [3], and by Verbeke et al. [9]. The data set contains 1,000 hand-annotated abstracts; 384 of the abstracts are structured (section headings present), while the rest are unstructured (no section headings) abstracts. Each sentence of each abstract is annotated with one or more classes. Thus, for example, a sentence may be annotated as both *Background* and *Study Design*. A sentence that is not classified to be any of the five PIBOS elements, is annotated as *Other*. The task at hand, therefore, is a multi-class classification problem with a total of six possible classes. Specific details about this corpus can be found in [3].

We model this problem as a supervised classification problem and attempt to solve it using SVMs. We divide the multi-class classification problem to six binary classification tasks, and apply one-vs-all classification for each. Given the features associated with a sentence, each classification task attempts to determine if the sentence belongs to one of the six PIBOSO categories, and assigns a probability value to the sentence. If, for a category C, the probability assigned to a sentence by the classifier is greater than 0.5, C is considered to be a label for that sentence. If, for a sentence, the probability of $C = Other$ is greater than all the other probabilities, the sentence is labeled as *Other*.

We divide the corpus into two sets: training (800 abstracts) and evaluation (200 abstracts). We perform 10-fold cross validations on the training set and use it to optimise specific parameters of our SVMs. We use the 200 unseen abstracts to test the performance of our SVMs.

3.2 Features and Approach

One advantage of formulating the problem as a set of binary classification problems is that we can customise the features to each classification task. This means that if there are features that are particularly useful for identifying a specific class only, we can use those features for the classification task involving that class, and leave them out if they are not useful for the other classes. In our work, we

apply a class-specific feature set for the classification of *Outcome* sentences and show that the use of this feature improves performance.

Preprocessing. We run MetaMap on the abstract texts to identify the medical concepts and their categories. Each medical term is assigned a *Concept Unique Identifier* (CUI) that remains unique for different lexical representations of the same concept (e.g., high blood pressure and hypertension). Each concept also has one or more broad categories to which it belongs (e.g., disease or syndrome), called *Semantic Types* (ST). Therefore, each medical concept is assigned a CUI and one or more STs by MetaMap. MetaMap also contains the MedPost/SKR parts of speech (POS) tagger, which we use to annotate each word in the data set. We further preprocess the text by lowercasing all terms, removing stopwords and stemming the words using the Porter stemmer.

Lexical Features. We use word n-grams as our first feature set. We have experimented with n-grams of various lengths (up to $n = 4$). Our experiments on the training set showed that n-grams may be useful up to tri-grams (e.g., n= 1, 2, and 3). We also utilise the POS information of the text by adding another set of n-grams along with the POS tags for each token.

Structural Features. We use three structural features from our data set: sentence positions, sentence lengths, and section headings. The position of a sentence provides useful information about its category: a sentence towards the end of a document is much more likely to be an *Outcome* sentence than a *Background* sentence. We use both relative and absolute sentence positions as features. For structured abstracts, we use the section headings as a feature set. Section headings represent the themes of different parts of the abstracts and, therefore, provide key information about the contents of sentences belonging to the associated sections. For each sentence, we add a feature which specifies the heading of the section to which a sentence belongs (if any). Our third feature in this category is sentence length, in terms of number of words. Although previously unused for this task, our analysis show that lengths of sentences can provide useful information for specific classes.

Domain-specific Semantic Features. We incorporate domain-specific information by using the UMLS CUIs and STs as features. For each sentence, all the STs and CUIs associated with the terms in that sentence are used as features in a bag-of-words fashion (i.e., ordering of terms is not taken into account).

Sequential Features. Kim et al. [3] and Verbeke et al. [9] both show the importance of sequential information for this classification task. The former argue that sentence-level sequential information can be valuable for this task since sentences for a particular subtopic (e.g., *Background*) typically occur sequentially as a group, and do not tend to repeat in later context. As such, for a sentence, incorporating information from surrounding sentences is likely to be useful for classification. Guided by the experimental results presented in [3], we add, to each sentence, the preprocessed n-grams (n = 1, 2, and 3) of the previous sentences. We have experimented by including up to three previous sentences, but found that only two previous sentences are useful (the second previous sentence may only be very marginally useful).

Class-specific Features. We explore the use of class-specific features for the *Outcome* class only. We incorporate a set of cue phrases from [5] that are used to identify sentences presenting outcomes. In the *Outcome* classification task, the count of those specific cue phrases present in a sentence is added as a feature.

4 Results and Discussion

To execute the learning and classification tasks, we use the LibSVM[6] implementation for SVMs. We use an RBF kernel and optimise the c parameter using 10-fold cross validation over the training set. Table 1 presents the f-scores for each class for the 10-fold cross validations (CV) over the training set, and over the test set. F-scores for both structured and unstructured abstracts are shown, along with the class frequencies. Our training data set contains 304 structured abstracts and 496 unstructured abstracts, while the evaluation set contains 80 structured and 120 unstructured abstracts. The table enables us to compare our results with those obtained by Kim et al. [3], Verbeke et al. [9], and a simple Naïve Bayes baseline. It can be seen that the micro-averaged f-scores of our system are comparable to both the benchmark systems, and that our system tends to perform better for the less-frequent classes (e.g., *study design*). The f-scores for our 10-fold cross validations and evaluation sets values are comparable for frequent classes (e.g. *outcome*). However, for low-frequency classes, the f-scores vary more between our training and evaluation sets. This can be attributed to the small number of instances available, and also the low agreement among human annotators for these classes, as depicted in [3]. Importantly, our approach almost invariably performs better than both benchmark systems for unstructured abstracts, illustrating its low reliance on the discourse structure of documents. Notably, our system performs better despite using only 800 abstracts for the 10-fold cross validation task, compared to the benchmark systems' use of all 1000 abstracts. We also empirically verified that our system outperforms other similar baselines. A baseline with a good performance is a multi-class SVM classifier with the same features. It achieves an overall micro-averaged f-score of 0.752, compared to our system's micro-averaged f-score of 0.766). In the ALTA Shared Task 2012, where AUC scores were used for evaluation, our system obtained second position with a score of 0.955. This score was significantly better than the CRF-based benchmark system's score (0.804).

We performed an analysis of the features and found that, in general, lexical features are more useful for our classifier than sequential features, unlike the CRF-based benchmark system [3]. The use of sentence length as a feature largely improves classification accuracy for the *Other* class, but is not effective for some classes such as *Background*; sentence position related features are useful for identifying *Outcome* and *Background* classes, but not for others; and lexical and domain-specific features are comparably useful for all classes. The cue-phrase-count feature, when added to the other features, improves the classification f-score of the *Outcome* class by up to 1%, showing its usefulness.

[6] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

Class	10-fold CV		Eval. Set		Kim et al.		Verbeke et al.		Naïve Bayes		Class Freq.
	S	**U**	**S**	**U**	**S**	**U**	**S**	**U**	**S**	**U**	
Population	45.0	**59.9**	51.8	54.0	**56.3**	39.8	35.6	21.5	39.1	38.8	809
Intervention	29.8	**38.9**	27.2	24.7	20.3	12.9	26.1	16.1	**30.9**	30.6	687
Background	85.4	75.8	83.4	73.6	81.8	68.5	**86.2**	**76.9**	58.9	67.7	2,557
Outcome	91.1	**81.3**	90.8	80.3	92.3	72.9	**93.0**	77.7	77.6	70.0	4,523
Study	52.6	**60.3**	**59.6**	46.1	43.9	4.40	45.5	6.67	33.4	26.2	233
Other	87.7	**48.4**	**88.2**	40.6	70.0	24.3	88.0	24.4	80.7	35.0	3,396
Micro-average	84.1	**72.6**	**84.6**	71.6	80.9	66.9	84.3	67.1	65.5	56.8	

Table 1. Classification f-scores and their micro-averages for each of the 6 classes. Scores for structured (S) and unstructured (U) abstracts are shown separately for each class.

5 Conclusion and Future Work

We propose a 6-way binary classification approach using SVMs to solve the multi-label sentence classification problem. Our approach performs comparably to baselines and state-of-the-art classifiers for this task, and performs particularly well for unstuctured abstracts. Our experiments also show that the formulation of the problem as several binary classification tasks rather than a single multi-class problem is beneficial because of the possibility to use class-specific features. Our future work will focus on identifying useful class-specific features, for each of the six categories. We expect that the careful selection of features will significanly improve classification performance, particularly for the less frequent classes.

References

1. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: The metamap program. In: Proceedings of AMIA Symposium. pp. 17–21 (2001)
2. Chung, G.Y.: Sentence Retrieval for Abstracts of Randomized Controlled Trials. BMC Medical Informatics and Decision Making 9(10) (2009)
3. Kim, S.N.N., Martinez, D., Cavedon, L., Yencken, L.: Automatic classification of sentences to support Evidence Based Medicine. BMC bioinformatics 12(2) (2011)
4. Lin, J.J., Demner-Fushman, D.: Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics 33(1), 63–103 (2007)
5. Niu, Y.: Analysis of Semantic Classes: Toward Non-Factoid Question Answering. Ph.D. thesis, University of Toronto (2007)
6. Richardson, W.S., Wilson, M.C., Nishikawa, J., Hayward, R.S.: The well-built clinical question: a key to evidence-based decisions. ACP Journal Club 123(3) (1995)
7. Sutton, C., McCallum, A.: Introduction to Statistical Relational Learning, chap. Introduction to Conditional Random Fields for Relational Learning, pp. 93–128. MIT Press (2007)
8. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA (1995)
9. Verbeke, M., Van Asch, V., Morante, R., Frasconi, P., Daelemans, W., De Raedt, L.: A statistical relational learning approach to identifying evidence based medicine categories. In: Proceedings EMNLP-CoNLL. pp. 579–589 (2012)