Automatic Extraction of Semantic Content from Medical Discharge Records

György Szarvas[1], Richárd Farkas[2], Szilárd Iván[2], András Kocsor[2], Róbert Busa-Fekete[2]

[1] University of Szeged, Department of Informatics, 6720 Szeged, Árpád tér 2., Hungary

[2] Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged H-6720 Szeged, Aradi vértanúk tere 1., Hungary

Abstract

Semi-structured medical texts like discharge summaries are rich sources of information that can exploit the research results of physicians with statistical analysis of similar cases. In this paper we introduce a system based on Machine Learning (ML) algorithms that successfully classifies discharge records according to the smoking status of the patient (we distinguish between current smoker, past smoker, smoker /where a decision between the former two classes cannot be made/, non-smoker and unknown /where the document contains no data on smoking status/ classes). Such systems are useful for examining the connection between certain social habits and diseases like cancer or asthma. We trained and tested our model on the shared task organized by the I2B2 (Informatics for Integrating Biology and the Bedside) research center [1], and despite the very low amount of labelled training data available, our system shows promising results in identifying the smoking habits of patients based on their medical discharge summaries.

INTRODUCTION

The classification of documents into different categories based on their content can really be regarded as an Information Extraction (IE) task where the aim is to derive some sort of semantic knowledge from the text. This problem arises in many real-life problems from spam filtering to the retrieval of relevant articles in huge databases like MedLine or the grouping of medical records according to the social habits/behaviour of the patients.

Processing of medical records

The main purpose of processing medical discharge records is to facilitate medical research carried out by physicians by providing them with statistically relevant data for analysis. An example of such an analysis might be a comparison of the runoff and effects of certain illnesses among patients with different social habits. The relevance drawn from the direct connection between social characteristics and diseases (like the link between smoking status and lung cancer or asthma) is of key importance in treatment and prevention issues. Such facts can be deduced automatically by applying statistical methods on large corpuses of medical records.

Related work

The identification of smoking habits based on discharge orders was studied earlier in the literature. [2, 3] reported an accuracy of 90% on the identification of smoker status. They constructed a classification model using about 8500 smoking-related sentences retrieved from discharge records and Support Vector Machine (SVM) as a classifier and word phrases of length 1-3 as features. Our approach differs from the one reported by them in the amount of data used (about 170 smoking-related sentences) and the variety of features employed (our system exploits syntactic information as well).

OUR APPROACH

Keyword-level classification

After some preliminary examinations of the structure of medical discharge records, we came to the conclusion that it was not whole discharge records that were relevant to the semantic information we aimed to extract, but rather short excerpts of the texts (or their absence) contained enough information to distinguish patients belonging to different smoker classes. As the classification of smaller text pieces with the same information content is always easier, we searched the documents for relevant parts or sentences that

word chunk	relative freq. (known/unknown)	document freq. (known)
tacrolimus	infinity	2
larynx	infinity	6
cigar	infinity	17
mgs.	infinity	4
smoke	59.5	108
tobacco	30	50
habit	18	—
father	12.5	—
AICD	12	—
palliation	12	—

Table 1: The frequencies of relevant words

appeared in documents that belonged to the four smoker classes (referred to as *known* texts later on) but were almost never seen in records that held no information on the patient's smoking status.
The most characteristic word chunks that distinguished *unknown* texts from others along with their relative known/unknown frequency and known-document frequency can be seen in Table 1. These word chunks that appear with the highest relative frequency (characteristic) and high known-document frequency (representative) really tell us that a document contains relevant information on the smoking status of the patients. The four most informative word chunks came to be {*cigar, smoke, tobacco, habit*}, which is an interesting but not surprising result. Since 'Habit :' is a heading of discharge records and the heading is usually filled with sentences containing one or more of the 3 other key words, we discarded this from our experiments and restricted our classification model to sentences containing {*cigar or smoke or tobacco*}

This way we built a keyword-level classifier, and since a document might contain more than one keyword, a joint decision had to be made to have a document-level classification. This was why we chose to test two different voting schemes. However, we did find that their efficiency was not significantly different.

Description of our classification model

The general structure of our document classification model can be seen in Figure 1, and the key steps of processing a discharge record are the following:

1. Preprocessing filters out documents belonging to the unknown class, and collects relevant sentences from known-class documents.

2. The feature extractor builds a feature vector for each keyword found in the text for an inductive learning task.

3. A classifier model assigns one of the known-class labels (current smoker, non-smoker, past smoker, smoker) to each instance generated from the same document.

4. A majority voting scheme makes the final decision on which class the document belongs to.

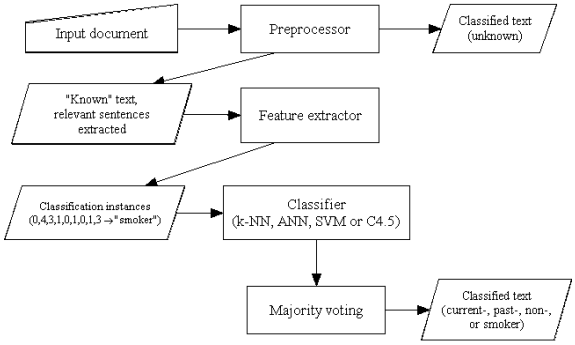

Figure 1: A schematic overview of our system

Features used

Our smoker status classifier system uses similar features to those employed by Zeng et. al. [2], considering phrases of length 1-3 words that we found characteristic to one or more of the smoker classes. In addition we also tried to incorporate deeper knowledge about the meaning of the sentence with several features by describing the part of speech information or some very basic properties about the syntactic structure. To get POS and syntactic information we used the publicly available Link Parser [4]. We should mention here that the sentences we extracted from the discharge records were out-domain texts for the parser and were often poorly formed sentences. These facts made the results of the parser somewhat poorer in quality than expected, but we think that since these sentences have very similar characteristics. Even the parse errors are similar in many cases and these features remain consistent and useful for the task.
The features we eventually opted for were the following:

1. We assigned 11 different values to the important 2-3 word long phrases for the class (or subset of classes) they indicated.
2. Which of the three keywords the sentence corresponded to.
3. Part of speech code of the keyword.
4. Whether the keyword was inside a Noun Phrase or Verb Phrase structure or not in the syntax tree of the sentence.
5. The lemma of the verb nearest to the keyword (in the syntax tree).
6. The part of speech code of the verb nearest to the keyword (in the syntax tree).
7. Whether the sentence contained a negative word (any of *no, none, never, negative, neither*) or not.
8. words seen in the training data several times (unigrams).

As regards the features described above, we collected 62 different attributes for each keyword in each sentence acquired from a document. The final decision on the patient's smoking status was made based on all the instances that originated from the same discharge summary, using a majority voting rule.

Learning methods

Nearest Neighbour Classifiers (k-NN) assign new instances to pre-defined classes by considering the known class labels to those training examples that are nearest to the new instance based on a distance measure. These methods are called k-NN classifiers where k denotes the number of training points in question to decide the class label of a new example. With our features, we can give an interesting interpretation to the labels assigned by a k-NN model: nearest neighbour classification is based on a kind of sentence similarity as our training instances characterise sentences. Since choosing the class label of the most similar sentence observed in the training data is a very simple and straightforward decision, we treated k-NN as a baseline in our experiments.

C4.5 decision tree is based on the well-known ID3 tree learning algorithm. It is able to learn pre-defined discrete classes from labelled examples. The result of the learning process is an axis-parallel decision tree. This means that during the training, the sample space is divided into subspaces by hyperplanes which are parallel to every axis but one. In this way, we get many n-dimensional rectangular regions that are labelled with class labels and organised in a hierarchical way, which can then be encoded into the tree. Splitting is done by axis-parallel hyper-planes, and hence learning is very fast. One great advantage of the method is its low time complexity.

Artificial Neural Networks (ANNs). Since it was realized that, under the right conditions, ANNs can model the class posteriors, neural nets have become evermore popular in the Natural Language Processing community. But describing here the mathematical background of ANN theory is beyond the scope of this article. Besides, we believe that ANNs are well known to those who are acquainted with pattern recognition. In our experiments we used the most common feed-forward multilayer perceptron network with the backpropagation learning rule.

Boosting (AdaBoost, AB) was introduced by Shapire as a way of improving the performance of a weak learning algorithm. The algorithm generates a set of classifiers (of the same type) by applying bootstrapping on the original training data set and it makes a decision based on their votes. The final decision is made using a weighted voting schema for each classifier that is many times more accurate than the original model. Here 10 iterations of Boosting were performed on the C4.5 model.

Support Vector Machines (SVM) is a kernel method that separates data points of different classes with the help of a hyperplane. The created separating hyperplane has a margin of maximal size with a proven optimal generalisation capacity. Another significant feature of margin maximisation is that the calculated result is independent of the distribution of the sample points. Perhaps the success and the popularity of this method can be attributed to this property.

We used the publicly available WEKA library for our experiments [5].

Feature selection

Solving a classification problem using a high-dimensional feature space often leads to overfitting on the training data. This means that, despite the seemingly low error-rates observed on the training data, the model cannot generalise well and performs poorly on unseen examples. In our experiments we had to handle the problem of having extremely low amounts of training data (about 200 instances) and numerous features collected for

each instance, hence we obtained a relatively high dimensional feature space.

A common solution to avoid overfitting on the training data is to reduce the dimensionality of the feature space using feature selection. The purpose of feature selection is to discard irrelevant attributes and keep those few that have the highest predictive power.

Chi-squared statistic (CSS): We used the well known chi-squared statistic to estimate the conditional dependence between individual features and the target attribute (that is, the class label). This statistical method computes the strength of dependency by comparing the joint distribution and the marginal distributions of the feature in question and the target variable. This way, the attributes could be ranked based on their individual relevance and this enabled us to discard insignificant features automatically.

CSS has some limitations though: e.g. as it compares attributes to the target attribute just one at a time. Thus it is possible that when a feature is not really informative on its own, but is useful when combined with other attributes, it might get a low rank by the chi-squared statistic.

Best subset selection (BSS): Another possibility is to rank subsets of features together, rather than measuring their individual association with the class values. This method has a very high computational time complexity as the number of possible subsets of features grows exponentially with the dimensionality of the initial feature space. Since we had an extremely low amount of training data available, this kind of subset evaluation became computationally feasible with classifiers that were fast to train, thus we decided to perform a best subset evaluation for the various attributes we used.

EXPERIMENTS AND RESULTS

Using the features described earlier, we constructed a learning model by assigning 62 different attributes for each keyword found in the discharge records. As we had only 200 training examples (originating from about 170 sentences extracted from 143 documents) available, it was quite apparent to us that dealing with such a high dimensional representation of the data could not be beneficial for classification accuracy.

We performed two different kinds of feature selection that helped us to reduce the dimensionality of our representation. Interestingly, both chi-squared attribute ranking and best subset selection (we applied a C4.5 decision tree classifier for evaluation) indicated that retaining 16 out of our 62 attributes was a good choice, but in the top ranked features they gave somewhat different results.

Both the CSS and BSS evaluations benefited from our deep knowledge features describing the syntactic and morphological properties of text, and important phrases of length 2-3 that indicated a single class value were also chosen by both evaluations. Best subset evaluation retained several features that described phrases indicating more than one class and several characteristic unigrams, while CSS underranked phrases that indicated 2 or more classes (indeed, these features can prove to be useful in combination with others and CSS is barely able to capture this evidence) and thus kept more unigram features, a few of which were hard to interpret.

The results of our feature evaluation clearly show that deep knowledge features that describe the syntactic properties of the text contribute greatly to the identification of a patient's smoking status. The features selected by one or both of the methods were the following:

Both: *lemma and POS of the verb nearest to keyword; negative word in the sentence; 2-3 word long phrases indicating 'current smoker', 'past smoker', 'non-smoker', 'current/past smoker' or 'smoker/non-smoker'; unigram in sentence: 'ago'*

BSS: *lemma of keyword; inside Noun Phrase; 2-3 word long phrases indicating 'smoker/current smoker' or 'smoker/past smoker'; unigram in sentence: 'use', 'drinks', 'quitting'*

CSS: *POS of keyword; unigram in sentence: 'years', 'does', 'smoke', 'per', 'smoker', 'approximately'*

As the features chosen by BSS were much easier to interpret, in our experiments we decided to use the 16 features that performed the best in the best subset selection.

We tested an ANN, SVM, AdaBoost+C4.5 decision tree learner, and a voting of ANN, SVM and C4.5 with performing a 5-fold evaluation on the training data. We chose a 5-fold cross-validation to get test sets of reasonable size (around 40 instances per fold) We used a k-NN classifier that implements a kind of sentence-similarity based classification as a baseline in our experiments.

To compare the classifiers with each other, we performed a 5-fold cross-validation 10 times, with randomized instances in the folds to eliminate any comparison's sensitivity to the low amount of data that might cause one method or another to perform better than the rest.

	ANN	SVM	AB+C4.5	VOTE
AVG F %	85.17	84.28	84.57	85.97
DEV %	1.64	1.96	1.45	1.34

Table 2: The average keyword-level accuracies and deviations

	5-fold		i2b2 eval	
	4-class	5-class	4-class	5-class
k-NN	76.92	90.95	–	–
SVM	77.62	91.21	–	–
AB-C4.5	81.11	92.46	63.41	85.58
ANN	81.11	92.46	63.41	85.58
VOTE	83.22	93.22	65.85	86.54

Table 3: The document accuracies of our models (weighted with the no. of documents in each class)

The average performance (keyword-level F measure) of the methods, along with their standard deviation can be seen in Table 2. As the reader can see, the voting model performed the best on average, and also it had the lowest deviation, so we can say it was the most reliable model. We evaluated each model at the document level later on. The results at the document-level were not so good, as at the keyword-level evaluation, instances originating from the same document often fell into different folds (and thus aided the proper classification of each other). In document-level evaluations all the instances from the same document appeared in the same fold (and thus were used as test instances at the same time, not helping each other). In Table 3 the document-level accuracies on the 4 known classes and for all 5 classes are given for all classifiers. Our system gave somewhat poorer results regarding the unweighted evaluation metrics of the i2b2 challenge (48.4% F-measure in 4-class and 58.7% F-measure in 5-class evaluation).

DISCUSSION

As our experiments show, the classification model we built can indeed identify the smoking status of patients, based on the analysis of their medical discharge records, with reasonable success. However the lack of training data is clearly visible from the significant deviation between the results among different random 5-fold cross-evaluations.

We extended the model introduced by Zeng et. al. [2] with several deep-knowledge features that describe the syntactic and morphological properties of the texts analysed. It is interesting to observe that our deep knowledge features are top ranked with different feature selection methods, thus here they proved to be extremely relevant in the classification of discharge records.

Taking into account the short number of training examples (e.g. we only had 9 samples for the 'smoker' class) the results we obtained look most promising. We think that with a decent amount of training samples the accuracy of the classification can be improved to give an F measure score of 90% or more.

SUMMARY AND CONCLUSIONS

In our studies we applied several inherently different Machine Learning algorithm for the semantic classification of structured documents based on their content. The advantage of these heterogenous classifiers is noticeable in a hybrid model that predicts the class label that seems to be the most certain in respect of the decisions of the individual models. In our paper we also introduced deep knowledge features that proved to be useful for the classification task.

The accuracy of our hybrid model achieved an F measure score over 80%, and this result is extremely promising considering the small amount of training examples used here.

Acknowledgements

Address for Correspondence

György Szarvas, University of Szeged, Department of Informatics, 6720 Szeged, Árpád tér 2., Hungary,

E-mail: szarvas@inf.u-szeged.hu

References

[1] I2B2. Informatics for integrating biology & the bedside. http://www.i2b2.org.

[2] Qing Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6(1):30, 2006.

[3] M Sordo and Q Zeng. On sample size and classification accuracy: A performance comparison. *Lecture Notes in Computer Science*, 3745:193–201, 2005.

[4] Dennis Grinberg, John Lafferty, and Daniel Sleator. A robust parsing algorithm for LINK grammars. Technical Report CMU-CS-TR-95-125, Pittsburgh, PA, 1995.

[5] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.