Developing Feature Types for Classifying Clinical Notes

Jon Patrick, Yitao Zhang and **Yefeng Wang**
School of Information Technologies
University of Sydney
NSW 2006, Australia
{jonpat, yitao, ywang1}@it.usyd.edu.au

Abstract

This paper proposes a machine learning approach to the task of assigning the international standard on classification of diseases ICD-9-CM codes to clinical records. By treating the task as a text categorisation problem, a classification system was built which explores a variety of features including negation, different strategies of measuring gloss overlaps between the content of clinical records and ICD-9-CM code descriptions together with expansion of the glosses from the ICD-9-CM hierarchy. The best classifier achieved an overall F_1 value of 88.2 on a data set of 978 free text clinical records, and was better than the performance of two out of three human annotators.

1 Introduction

Despite the rapid progress on text categorisation in the newswire domain, assigning meaningful labels to clinical notes has only recently emerged as a topic for computational linguists although health informatics researchers have been working on the problem for over 10 years. This paper describes constructing classifiers for the Computational Medicine Center's 2007 Medical Natural Language Processing Challenge which aims to assign ICD-9-CM codes to free text radiology reports. (Computational Medicine Center, 2007) It addresses the difficulties of medical text categorisation tasks by incorporating medical negations, term variations, and clues from hierarchy of medical ontologies as additional features.

2 The task of assigning ICD-9-CM codes

The corpus used in this study is a collection of radiology reports from the Cincinnati Children's Hospital Medical Center, Department of Radiology. (Computational Medicine Center, 2007) The data set is divided into a training set and a test set. The training set consists of 978 records and the test set consists of 976 records and 45 ICD-9-CM code. The task was considered as a multi-label text categorisation problem. For each code found in the corpus, we created a separate classifier which makes binary "Yes" or "No" decisions for the target code of a clinical record. Maximum Entropy Modeling (MaxEnt) (Berger et al., 1996) and Support Vector Machine (SVM) (Vapnik, 1995) were used to build the classifiers in our solution.

3 Features

A variety of features were developed to represent what we believed were the important determiners of the ICD-9-CM codes.

Bag-of-words (BOW) features: include only unigrams and bigrams in the text.

Negation features: were used in the classification system to capture the terms that are negated or uncertain, for example "pneumonia" vs "no evidence of pneumonia". We created a negation-finding system which uses an algorithm similar to (Chapman et al., 2001) to identify the negation phrase and the scope of negations.

Gloss matching feature: The ICD-9-CM provides detailed text definition for each code. This section explores different strategies for measuring gloss

BioNLP 2007: Biological, translational, and clinical language processing, pages 191–192,
Prague, June 2007.

Name	Description	P	R	F_1
S0	BOW baseline	83.9	78.4	81.1
S1	S0 + negation	88.5	78.2	83.0
S2	S1 + gloss matching	89.2	80.6	84.7
S3	feature engineering	89.7	86.0	87.8
S4	S3 + low-freq	89.7	86.9	88.2

Table 1: Experiment results for all ICD-9-CM codes

matchings between the content of a clinical record and the definition of an ICD-9-CM code.

Feature engineering: In experiments with a uniform set of feature types for all ICD-9-CM codes, we noticed that different codes tend to have a preference for different combinations of feature types. Therefore, different combinations of feature types for each individual code were used. The intuition is to explore different combination of feature types quickly instead of doing further feature selection procedures. The system trained on the best combination of feature types are reported as the final results for the target code.

Low frequency codes modeling: A rule-based system was also used to model low frequency ICD-9-CM codes which have only one occurrence in the corpus, or have achieved F_1 value of 0.0 by machine learning. The system assigns a low frequent code to a clinical record if the content of the record matches the words of the code definition.

4 Result

Table 1 shows the experiment results. Since the gold-standard annotation of the test dataset has not been released so far, the experiment was done on the 978 documents training dataset using 10-fold cross-validation. [1] The baseline system S0 was created using only BOW features. Adding negation features gives S1 an improvement of 1.9% on F_1 score. The gloss matching features gives a further increase of 1.7% on F_1 score.

In order to understand more about the ICD-9-CM code assignment task, this section evaluates the

[1]The official score of our system on the test dataset is $F_1 = 86.76$ which was ranked 7th among 44 systems. See http://www.computationalmedicine.org/challenge/res.php

Name	P	R	F_1	N
company1	78.3	89.8	83.7	1397
company2	82.6	95.2	88.5	1404
company3	90.4	75.0	82.0	1011
S4	89.7	86.9	88.2	1180

Table 2: Performances of Annotators

performance of the three annotators. Table 2 compares the performance of each annotator to the gold-standard codes. The item "N" in Table 2 stands for the total number of ICD-9-CM codes which an annotator has assigned to the whole corpus.

5 Conclusion

This paper presents an approach to the problem of assigning ICD-9-CM codes to free text medical records. We created a classification system which consists of multiple machine-learned classifiers on high-frequency codes, and a rule-based modeling module of low-frequency codes. By incorporating negations and a variety of gloss matching features, we successfully outperformed the baseline with only bag-of-words features by 7.1% on F_1 value. The best reported score is also considered as comparable to the performance of the best human annotator. We also consider the way our system selected the best combination of feature types for each individual ICD-9-CM code has a major contribution to the classification task of clinical records.

References

Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Computational Medicine Center. 2007. 2007 Medical Natural Language Processing Challenge. http://www.computationalmedicine.org/challenge/.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.