A Hierarchical Approach to Encoding Medical Concepts for Clinical Notes

Yitao Zhang
School of Information Technologies
The University of Sydney
NSW 2006, Australia
yitao@it.usyd.edu.au

Abstract

This paper proposes a hierarchical text categorization (TC) approach to encoding free-text clinical notes with ICD-9-CM codes. Preliminary experimental result on the 2007 Computational Medicine Challenge data shows a hierarchical TC system has achieved a micro-averaged F_1 value of 86.6, which is comparable to the performance of state-of-the-art flat classification systems.

1 Introduction

The task of assigning meaningful categories to free text has attracted researchers in the Natural Language Processing (NLP) and Information Retrieval (IR) field for more than 10 years. However, it has only recently emerged as a hot topic in the clinical domain where categories to be assigned are organized in taxonomies which cover common medical concepts and link them together in hierarchies. This paper evaluates the effectiveness of adopting a hierarchical text categorization approach to the 2007 Computational Medicine Challenge which aims to assign appropriate ICD-9-CM codes to free text radiology reports. (Pestian et al., 2007)

The ICD-9-CM [1], which stands for International Classification of Diseases, 9th Revision, Clinical Modification, is an international standard which is used for classifying common medical concepts, such as diseases, symptoms and signs, by hospitals, insurance companies, and other health organizations. The 2007 Computational Medicine Challenge was set in a billing scenario in which hospitals claim reimbursement from health insurance companies based on the ICD-9-CM codes assigned to each patient case. The competition has successfully attracted 44 submissions with a mean micro-averaged F_1 performance of 76.70. (Pestian et al., 2007)

To the best of our knowledge, the systems reported were all adopting a flat classification approach in which a dedicated classifier has been built for every targeted ICD-9-CM code. Each classifier makes a binary decision of True or False according to whether or not a clinical note should be assigned with the targeted ICD-9-CM code. An incoming clinical note has to be tested against all the classifiers before a final coding decision can be made. The response time of a flat approach therefore grows linearly with the number of categories in the taxonomy. Moreover, low-frequency ICD-9-CM codes suffer the data imbalance problem in which positive training instances are overwhelmed by negative ones.

A hierarchical system takes into account relationships among categories. Classifiers are assigned to both leaf and internal nodes of a taxonomy and training instances are distributed among these nodes. When a test instance comes in, a coding decision is made by generating all possible paths (start from the root node of the taxonomy) where classifiers along path return favorable decisions. In other words, a node is visited only if the classifier assigned to its parent returns a True decision. This strategy significantly reduces the average number of classifiers to be used in the test stage when the taxonomy is very large. (Liu et al., 2005; Yang et al., 2003)

[1] see http://www.cdc.gov/nchs/icd9.htm

Proceedings of the ACL-08: HLT Student Research Workshop (Companion Volume), pages 67–72,
Columbus, June 2008. ©2008 Association for Computational Linguistics

2 Related Works

Most top systems in the 2007 Computational Medicine Challenge have benefited from incorporating domain knowledge of free-text clinical notes, such as negation, synonymy, and hypernymy, either as hand-crafted rules in a symbolic approach, or as carefully engineered features in a machine-learning component. (Goldstein et al., 2007; Farkas and Szarvas, 2007; Crammer et al., 2007; Aronson et al., 2007; Patrick et al., 2007)

Aronson et al. (2007) used a variant of National Library of Medicine Medical Text Indexer (MTI) which was originally developed for discovering Medical Subject Headings (MeSH) [2] terms for indexing biomedical citations and articles. The output of MTI was converted into ICD-9-CM codes by applying different approaches of mapping discovered Unified Medical Language System (UMLS) [3] concepts into ICD-9-CM codes, such as using synonym and built-in mapping relations in UMLS Metathesaurus. This approach can easily adapt to any subdomain of the UMLS Metathesaurus since it only requires very little examples for tuning purposes. However, MTI performed slightly behind an SVM system with only bag-of-words features, which suggests the difficulty of optimizing a general purpose system without any statistical learning on the targeted corpus. By stacking MTI, SVM, KNN and a simple pattern matching system together, a final F_1 score of 85 was reported on the official test set.

Farkas and Szarvas (2007) automatically translate definitions of the ICD-9-CM into rules of a symbolic system. Decision tree was then used to model the disagreement between the prediction of the system and the gold-standard annotation of the training data set. This has improved the performance of the system to a F_1 value of 89. Goldstein et al. (2007) also reported that a rule-based system enhanced by negation, synonymy, and uncertainty information, has outperformed machine learning models which only use n-gram features. The rules were manually tuned for every ICD-9-CM code found in the challenge training data set and therefore suffer the scaling up problem.

Total radiology records	1,954
Total tokens	51,940
Total ICD-9-Codes	45
Total code instances	2,423

Table 1: Statistics of the data set

On the other hand, researchers tried to encode domain knowledge into machine learning systems by developing more sophisticated feature types. Patrick et al. (2007) developed a variety of new feature types to model human coder's expertise, such as negation and code overlaps. Different combination of feature types were tested for each individual ICD-9-CM code and the best combination was used in the final system. Crammer et al. (2007) also used a rich feature set in their MIRA system which is an online learning algorithm.

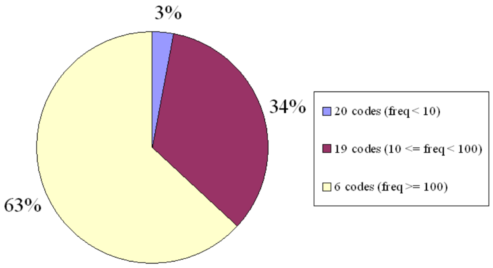

Figure 1: Distribution of ICD-9-CM codes in the challenge data set.

3 The Corpus

The corpus used in this study is the official data set of the 2007 Computational Medicine Challenge. The challenge corpus consists of 1,954 radiology reports from the Cincinnati Children's Hospital Medical Center and was divided into a training set with 978 records, and a test set with 976 records. The statistics of the corpus is shown in Table 1.

Each radiology record in the corpus has two sections: 'Clinical History' which is provided by an ordering physician before a radiological procedure, and 'Impression' which is reported by a radiologist after the procedure. A typical radiology report is shown below:

[2] http://www.nlm.nih.gov/mesh/

[3] http://www.nlm.nih.gov/research/umls/

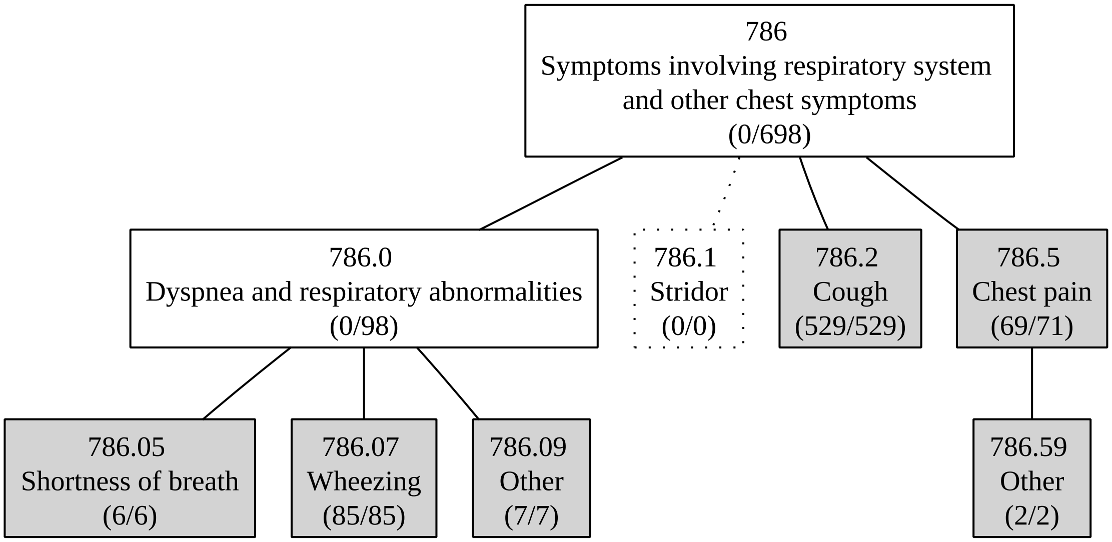

Figure 2: A part of the ICD-9-CM taxonomy: the tree covers symptoms involving respiratory system and other chest symptoms. There are two figures shown in each node: the first figure is the number of positive instances assigned to the current node, and the next figure shows the number of all the instances in its subtree.

> *Clinical history*
>
> Persistent cough, no fever.
>
> *Impression*
>
> Retained secretions vs atelectasis in the right lower lobe. No infiltrates to support pneumonia

Three different institutions were invited to assign ICD-9-CM codes to the corpus. The majority code with at least two votes from the three annotators was considered as the gold-standard code for the record. Moreover, a clinical record can be assigned with multiple ICD-9-CM codes at a time.

The general guideline of assigning ICD-9-CM codes includes two important rules:

- If there is a definite diagnosis in text, the diagnosis should be coded and all symptom and sign codes should be ignored.

- If the diagnosis is undecided, or there is no diagnosis found, the symptoms and signs should be coded rather than the uncertain diagnosis.

According to the guideline, the above radiology record should be assigned with only a 'Cough' code because 'Atelectasis' and 'Pneumonia' are not certain, and 'Fever' has been negated.

There are 45 ICD-9-CM codes found in the corpus and their distribution is imbalanced. Figure 1 shows a pie chart of three types of the ICD-9-CM codes found in the corpus and their accumulated category frequencies. The 20 low-frequency (less than 10 occurrences) codes account for only 3% of the total code occurrence in the challenge data set. There are 19 codes with a frequency between 10 and 100 and altogether they account for 34% total code occurrence. Finally, the most frequent six codes account for over 60% of total code instances.

4 Hierarchical Text Categorization Framework

In a hierarchical text categorization system, categories are linked together and classifiers are assigned to each node in the taxonomy. In the training stage, instances are distributed to their corresponding nodes. For instance, Figure 2 shows a populated subtree of ICD-9-CM code '786' which covers concepts involving respiratory system and other chest symptoms. Nodes in grey box such as 786.2 and 786.5 are among 45 gold-standard codes found in the challenge data set. Nodes in white box such as 786 and 786.0 are internal nodes which have nonempty subtrees. For instance, the numbers (0, 698) of '786' suggest that the node is assigned with zero instances for training while there are 698 positive instances assigned to nodes in its subtree. The node '786.1' is in dotted box because there is no instance assigned to it, nor any of its subtrees. In the experiment, all nodes (such as '786.1') with empty instance in its subtree were removed from the training and testing stage.

When training a classifier for a node A in the tree, all the instances in the subtree rooted in the parent of A become the only source of training instances. For instance, code '786.0' in Figure 2 uses all the 698 instances rooted in node '786' as the full training data set. The 98 instances rooted in node '786.0' itself are the positive instances while the remaining 600 instances in the tree as the negative ones. This hierarchical approach of distributing training instances can reduce the size of training data set for most classifiers and minimize the data imbalance problem for low-frequency codes in the taxonomy.

In the test stage, the system starts from the root of the ICD-9-CM taxonomy and evaluates an incoming clinical note against classifiers assigned to its children nodes. The system will then visit every child node which returns a positive classification result. The process repeats recursively until a possible path ends by reaching a node that returns a negative classification result. This strategy enables the sytem to assign multiple codes to a clinical note by visiting different paths in the ICD-9-CM taxonomy simultaneously.

5 Methods and Experiments

5.1 Experiment Settings

In this study, Support Vector Machines (SVM) was used for both flat and hierarchical text categorization. The LibSVM (Chang and Lin, 2001) package was used with a linear kernel.

5.1.1 Hierarchical TC

A tree of ICD-9-CM taxonomy was constructed by enquiring the UMLS Metathesaurus. During each iteration of 10-fold cross-validation experiment, the training instances were assigned to the ICD-9-CM tree and all nodes assigned with zero training instance in its subtree were removed from the tree. This ended with an ICD-9-CM tree with around 100 nodes for each training and test iteration.

Nodes in the tree were uniquely identified by their concept id (CUI) found in the UMLS Metathesaurus. However, two ICD-9-CM codes ('599.0' and 'V13.02') were found to share the same CUI in the UMLS Metathesaurus. As a result, 44 unique UMLS CUIs were used as the gold-standard codes in the experiment for the original 45 ICD-9-CM codes.

In the test stage, the hierarchical system returns the terminal nodes of the predicted path. Moreover, if the terminal ends in an internal code which is not one of the 44 gold-standard UMLS CUI found in the training corpus, the system should ignore the whole path.

5.1.2 Flat TC

In a flat text categorization setting, 44 classifiers were created for each UMLS Metathesaurus CUI found in the corpus. Each classifier makes a binary decision of 'Yes' or 'No' to a clinical record according to whether or not it should be assigned with the current code.

5.2 Preprocessing

The corpus was first submitted to the GENIA tagger (Tsuruoka et al., 2005) for part-of-speech tagging and shallow syntactic analysis. The result was used by the negation finding module and all the identified negated terms were removed from the corpus. The cleaned text was used by the MetaMap (Aronson, 2001) for identifying possible medical concepts in text. The MetaMap software is configured to return only concepts of ICD-9-CM and SNOMED CT which is another comprehensive medical ontology widely used for mapping concepts in free-text clinical notes.

5.3 Evaluation

The main evaluation metric used in the experiment is the micro-averaged F_1 which is defined as the harmonic mean between $Precision$ and $Recall$:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where

$$Precision = \frac{\sum_i TP(Code_i)}{\sum_i TP(Code_i) + \sum_i FP(Code_i)}$$

$$Recall = \frac{\sum_i TP(Code_i)}{\sum_i TP(Code_i) + \sum_i FN(Code_i)}$$

In the above equation, $TP(Code_i)$, $FP(Code_i)$, and $FN(Code_i)$ are the numbers of true positives, false positives, and false negatvies for the ith code. The micro-averaged F_1 considers every single coding decision equally important and is therefore dominant by the performance on frequent codes in data. Moreover, a hierarchical micro-averaged $F_1^{(hierarchical)}$ is also introduced by adding all ancestors of the current gold-standard code into calculation. The $F_1^{(hierarchical)}$ value helps to evaluate how accurate a system predicts in terms of the gold-standard path in the ICD-9-CM tree.

5.4 Features

The feature set is descibed in Table 2.

- Bag-of-words

  Both unigram (F1) and bigram (F2) were used.

- Negation and Bag-of-concepts

  An algorithm similar to NegEx (Chapman et al., 2001) was used to find negations in text. A small set of 35 negation keywords, such as ‘no’, ‘without’, and ‘no further’, was compiled to trigger the finding of the negated phrases in text based on the shallow syntactic analysis returned by GENIA tagger. After removing negated phrases in text, MetaMap was used to find medical concepts in text as new features in a bag-of-concepts manner (F3 and F4).

Different combination of feature types (F5, F6, and F7) were also used in the experiment. Information gain was used to rank the features and the feature cut-off threshold was set to 4, 000.

6 Result and Discussion

The 10-fold cross-validation technique was used in the experiments. The 1,954 radiology reports were randomly divided into ten folds. In each iteration of the experiment, one fold of data was used as the test set and the other nine folds as the training set.

The experimental results are shown in Table 2. The flat TC system has achieved higher F_1 scores than a hierarchical TC system in all experimental settings. However, paired t-test suggests the differences are not statistically significant at a ($p < 0.05$) level in most cases. This suggests the potential of adopting a hierarchical TC approach in the task. The effectiveness of the system is not sacrificed while the system now has the potential to scale up to much larger problems.

Similarly, the hierarchical TC system has better $F_1^{hierarchical}$ scores than the flat TC system while this difference is still not statistically significant at a ($p < 0.05$) level in most cases. This is partly due to the current strategy of not allowing unknown ICD-9-CM codes to be assigned in the system. As a result, many originally predicted internal nodes were removed in a hierarchical TC system.

Both the flat and hierarchical systems using bag-of-words feature set F1 have achieved a F_1 score above 0.85. Adding bigram features into F2 has shown minimum impact on the performance of both systems. Using a bag-of-concepts strategy in F3 and F4 has lowered the performance of the system. However, adding F3 and F4 into bag-of-words feature set has improved the performance of both systems. Finally, the best performance were reported on using feature set F5 which combines unigram and ICD-9-CM concepts returned by MetaMap software on the preprocessed text where negated terms were removed.

7 Conclusion and Future Work

Compared to a flat classification approach, a hierarchical framework is able to exploit relationships among categories to be assigned and easily adapts to much larger text categorization problems where real-time response is needed. This study has proposed a hierarchical text categorization approach to the task of encoding clinical notes with ICD-9-CM codes. The preliminary experiment shows that a hierarchical text categorization system has achieved a performance comparable to other state-of-the-art flat classification systems.

Future work includes developing more sophisticated features, such as synonym and phrase-level paraphrasing and entailment, to encode the knowl-

	Feature Description	Flat TC		Hierarchical TC	
		F_1	$F_1^{(hierarchical)}$	F_1	$F_1^{(hierarchical)}$
F1	Unigram	85.90 ± 2.00	89.50 ± 1.51	85.52 ± 1.30	90.49 ± 1.13
F2	Unigram, Bigram	85.99 ± 2.17	89.65 ± 1.70	85.27 ± 1.32	90.69 ± 1.20
F3	ICD-9-CM concepts on no negation text	81.96 ± 1.44	85.39 ± 1.47	81.45 ± 1.79	86.89 ± 1.65
F4	SNOMED CT concepts on no negation text	84.97 ± 1.55	89.00 ± 1.04	84.77 ± 1.04	89.82 ± 0.97
F5	F1 + F3	87.09 ± 1.70	90.26 ± 1.33	86.58 ± 1.30	91.08 ± 0.95
F6	F1 + F4	86.56 ± 1.69	89.99 ± 1.34	86.10 ± 1.80	90.70 ± 1.58
F7	F1 + F3 + F4	86.83 ± 1.34	90.23 ± 1.17	86.57 ± 1.28	91.06 ± 1.10

Table 2: 10-fold cross-validation experimental results

edge of human experts. How to manage a rich feature set in a hierarchical TC setting would be another big challenge. Moreover, this work did not use any thresholding tuning technique in the training stage. Therefore, a thorough study on the effectiveness of threshold tuning in the task is required.

Acknowledgments

I would like to thank Prof. Jon Patrick for his support and supervision of my research, and Mr. Yefeng Wang for providing his codes on negation finding. I also want to thank all the anonymous reviewers for their invaluable inputs to my research.

References

A.R. Aronson, O. Bodenreider, D. Demner-Fushman, K.W. Fung, V.K. Lee, J.G. Mork, A. Névéol, L. Peters, and W.J. Rogers. 2007. From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches. *Proceedings of the Workshop on BioNLP 2007*, pages 105–112.

A.R. Aronson. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap Program. *Proc AMIA Symp*, pages 17–21.

C. C. Chang and C. J. Lin, 2001. *LIBSVM: a Library for Support Vector Machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

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.

K. Crammer, M. Dredze, K. Ganchev, P.P. Talukdar, and S. Carroll. 2007. Automatic Code Assignment to Medical Text. *Proceedings of the Workshop on BioNLP 2007*, pages 129–136.

R. Farkas and G. Szarvas. 2007. Automatic Construction of Rule-based ICD-9-CM Coding Systems. *The Second International Symposium on Languages in Biology and Medicine*.

I. Goldstein, A. Arzumtsyan, and Ö. Uzuner. 2007. Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports. *AMIA Annu Symp Proc*.

T.Y. Liu, Y. Yang, H. Wan, H.J. Zeng, Z. Chen, and W.Y. Ma. 2005. Support Vector Machines Classification with a Very Large-scale Taxonomy. *SIGKDD Explorations, Special Issue on Text Mining and Natural Language Processing*, 7(1):36–43.

J. Patrick, Y. Zhang, and Y. Wang. 2007. Evaluating Feature Types for Encoding Clinical Notes. *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 218–225.

J.P. Pestian, C. Brew, P. Matykiewicz, DJ Hovermale, N. Johnson, K.B. Cohen, and W. Duch. 2007. A Shared Task Involving Multi-label Classification of Clinical Free Text. *Proceedings of the Workshop on BioNLP 2007*, pages 97–104.

Y. Tsuruoka, Y. Tateishi, J.D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pages 382–392.

Y. Yang, J. Zhang, and B. Kisiel. 2003. A Scalability Analysis of Classifiers in Text Categorization. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103.