Multi-Topic Aspects in Clinical Text Classification

Yutaka Sasaki[2] Brian Rea[2] Sophia Ananiadou[1,2]
[1]National Centre for Text Mining
[2]School of Computer Science
University of Manchester
MIB, 131 Princess Street, Manchester, M1 7DN, United Kingdom
{Yutaka.Sasaki,Brian.Rea,Sophia.Ananiadou}@manchester.ac.uk

Abstract

This paper investigates multi-topic aspects in automatic classification of clinical free text. In many practical situations, we need to deal with documents overlapping with multiple topics. Automatic assignment of multiple ICD-9-CM codes to clinical free text in medical records is a typical multi-topic text classification problem. In this paper, we facilitate two different views on multi-topics. The Closed Topic Assumption (CTA) *regards an absence of topics for a document as an explicit declaration that this document does not belong to those absent topics. In contrast, the* Open Topic Assumption (OTA) *considers the missing topics as neutral topics. This paper compares performances of various interpretations of a multi-topic Text Classification problem into a Machine Learning problem. Experimental results show that the characteristics of multi-topic assignments in the Medical NLP Challenge data is OTA-oriented.*

1 Introduction

Text Classification (or Categorization) has been investigated by many researchers over the past 20 years. Due to the drastic increase in online textual information, *e.g.*, email messages, online news, web pages, as well as a huge number of resources for scientific online abstracts such as MEDLINE, there is an ever-growing demand for Text Classification. It is an interesting question how to achieve high performance in the task of assigning multiple topics to documents in a targeted domain and how to make the most of the multi-topical features of the documents. The task to classify each document into multiple topics is called *multi-topic Text Classification*.

Automatically assigning multiple clinical codes to clinical free text is a typical multi-topic text classification problem. The Medical NLP Challenge 2007 [9] targeted the task of assigning ICD-9-CM codes [1] to radiology reports.

It is not straightforward to interpret a multi-topic text classification problem into a Machine Learning problem. Since a suitable Machine Learning approach depends on implicit and explicit characteristics of a target data set, we need to search for the best implementation among possible interpretations of the problem.

This paper carefully distinguishes the usage of the following terms in order to avoid a confusion between Text Classification problems and their implementation using Machine Learning algorithms.

Topic indicates categories that documents belong to. In this paper, the word *topic* is used only in the context of the Text Classification. The Text Classification problem is called single-topic Text Classification if every document belongs to a single topic. When documents belong to multiple topics, it is called multi-topic Text Classification.

Class is used in the context of Machine Learning in this paper. Trained classifiers classify data into predefined *classes*. The binary class classification problem has two classes, e.g., true or false, and the multi-class classification problem has multiple predefined classes.

Label is used in the context of Machine Learning in this paper. Classifiers assign *labels* (or *class labels*) to data. Single-label classifiers decide a single class label for each datum whereas multi-label classifiers decide multiple class labels for each datum.

2 Text Classification

Text Classification has been investigated by many researchers over the past 20 years. Traditionally, Text Classification has dealt with a single topic assigned to each doc-

[1] http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm

0-7695-3031-1/07 $25.00 © 2007 IEEE
DOI 10.1109/BIBM.2007.23

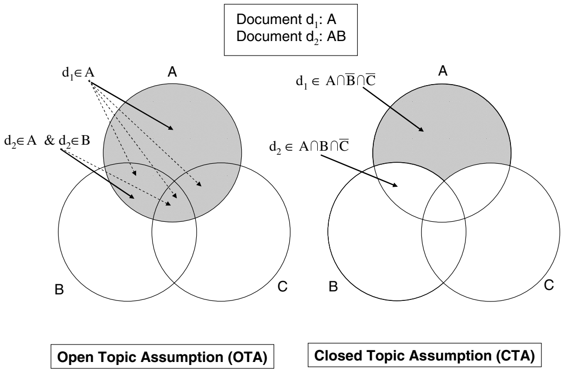
Document d_1: A
Document d_2: AB
$d_1 \in A$
A
$d_2 \in A$ & $d_2 \in B$
B
C
$d_1 \in A \cap \overline{B} \cap \overline{C}$
A
$d_2 \in A \cap B \cap \overline{C}$
B
C
Open Topic Assumption (OTA)
Closed Topic Assumption (CTA)

Figure 1. Venn Diagram of the CTA and OTA

ument. More formally, the problem can be defined as follows:

Definition 1 *(Single-Topic Text Classification)*
Given a set of m topics $\mathcal{T} = \{t_1, ..., t_m\}$, a set of documents $D = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_d\}$, and topic assignments $(\boldsymbol{x}_i, y_i)$ with $y_i \in \mathcal{T}$ for all $1 \leq i \leq d$, create a classifier C which correctly predicts the topic y_j for each document $x_j \in D^{test} \subset D$, based on topic assignments on $D^{train} \subset D$. Usually, the fairness condition $D^{train} \cap D^{test} = \emptyset$ is applied.

The single-topic Text Classification problem extends to the multi-topic Text Classification problem. The multi-topic Text Classification problem allows those cases where multiple topics are assigned to a document. More formally, the problem can be defined as follows:

Definition 2 *(Multi-Topic Text Classification) Given a set of m topics $\mathcal{T} = \{t_1, ..., t_m\}$, a set of documents $D = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_d\}$, and topic assignments $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ with $\boldsymbol{y}_i \subseteq \mathcal{T}$ for all $1 \leq i \leq d$, create a classifier C which correctly predicts the set of topics $\boldsymbol{y}_j$ for each document $\boldsymbol{x}_j \in D^{test} \subset D$, based on topic assignments of $D^{train} \subset D$. Usually, the fairness condition $D^{train} \cap D^{test} = \emptyset$ is applied.*

3 Machine Learning Algorithms

There are three types of Machine Learning algorithms in terms of the number of classes and labels.

Binary-Class (BC) Single-Label (SL) classification: A classifier decides a single label of a datum from two possible classes, e.g., Support Vector Machines (SVMs) [14].

Multi-Class (MC) Single-Label (SL) classification: A classifier decides a single class label from multiple possible classes, e.g., Maximum Entropy Models (MEMs) [1], Multi-Class SVMs[3], DAGSVM [10]. SVMs were extended to multi-class single-label SVMs (MC-SVMs) in [3].

Multi-Class (MC) Multi-Label (ML) classification: A classifier decides multiple class labels of a datum, e.g., Multi-Labelled MEM (MLME)[15]. Zhu et al. [15] extended MEMs to Multi-label MEMs.

4 From Multi-Topic Text Classification to Machine Learning

Implementing a multi-topic text classification problem with Machine Learning algorithms is not straightforward. We have to consider how to map multi-topics to classes of Machine Learning algorithms.

We define two concepts which can be facilitated to interpret multi-topic Text Classification problems into a Machine Learning problem. One is called the *Open Topic Assumption* (OTA) and the other is called *Closed Topic Assumption* (CTA).

Open Topic Assumption (OTA): multiple topics given to a document represent the topics that the text belongs to. The topics other than the given topics are neutral. This includes the case where only clear topics are given to a text and other topics are intentionally/unintentionally omitted from the topic assignments. For example, if there exist three topics $\mathcal{T} = \{A, B, C\}$ and documents d_1 and d_2 are given the topics $\{A\}$ and $\{A, B\}$, respectively, then these assignments are regarded as $d_1 \in A$ and $d_2 \in A$ & $d_2 \in B$. This means that d_1 would also be a member of the same class as d_2.

Closed Topic Assumption (CTA): multiple topics given to a text represent the topics that the text belongs to. At the same time, the topics other than the given topics are considered to be explicitly denied. This means that if there exist three topics $\mathcal{T} = \{A, B, C\}$ and a text d_1 is given the topic $\{A\}$, then this assignment is regarded as $d_1 \in A \cap \bar{B} \cap \bar{C}$, which means that d_1 does not belong to the same class of d_2 when d_2 has topics $\{A, B\}$ (*i.e.,* $A \cap B \cap \bar{C}$).

Figure 1 shows a Venn diagram of the CTA and OTA. These two assumptions can be considered as the end points on a spectrum of the nature of multi-topics. The characteristic of a specific data set could be placed somewhere on the spectrum.

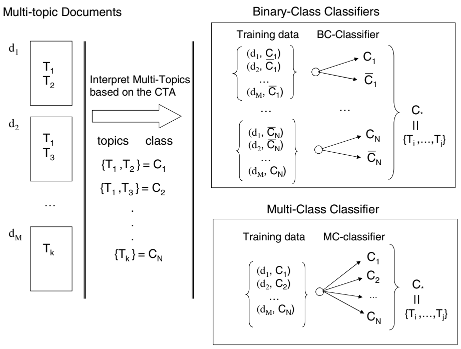

Figure 2. Implementing Multi-Topic Text Classification Problem

4.1 Interpretations

There are several possible ways to solve a multi-topic text classification problem with Machine Learning algorithms. Figure 2 shows another example on how to solve a multi-topic text classification problem using binary-class and multi-class machine learning algorithms based on the Closed Topic Assumption. Each set of topics attached to a document in the data set is used as the primitive unit of class labels. The class label C_* with the highest probability will be selected from classification results of N binary-class classifiers or a multi-class classifier.

BCSL-SVMs Based on the OTA, each SVM decides whether a given document belongs to a class that represents one of the topics. Multiple topics of the document are decided as a collection of topics that are determined by SVM classifiers.

BCSL-SVM/CTA Based on the CTA, each SVM decides whether a given document belongs to the specific class that represents a set of topics. The set of topics with the highest score is selected for the topics of the document.

MCSL-SVM/CTA Based on the CTA, each set of topics attached to a document in a data set is used as the primitive unit of a class. Each multi-class SVM decides the document's class which represents a set of topics.

MCSL-MEM/CTA The MEM is used in the usual multi-class single-label setting. Based on the CTA, each MEM decide the most probable class that represents a set of topics.

BCSL-MEM/CTA In order to assess the effect of the multi-class nature of MEM, MEMs are applied in a binary-class setting. Based on the CTA, each MEM determines whether a document belongs to a class which represent multiple topics. The class (*i.e.*, topics) having the highest probability is selected.

BCSL-MEMs MEMs are applied in a binary-class setting. Based on the OTA, each MEM decides whether or not a given document belongs to the single topic. The topics of the document are given in a collection of topics that are decided by MEMs.

5 Data set

We used the Medical NLP Challenge 2007 data whose number of texts are small and content focuses on the clinical field. The CMC Medical NLP Challenge 2007 Data Set [2] is the data set used in the shared task of the Medical NLP Challenge 2007. In total, the number of ICD-9-CM codes assigned to documents is 45. 1-3 topics are assigned to each document.

Because of confidentiality requirement, collecting sufficient amount of training data is inherently difficult in the clinical field. Computational Medicine Center anoynmized and provided a data set that consists of 978 radiology reports for training and 976 radiology reports for test. Every report has two sections: Clinical History and Impression.

Three annotators A, B, and C independently assigned ICD-9-CM codes to the data. Sometimes, different codes were assigned to the same text. Therefore, the annotated ICD-9-CM code is considered correct when two or more of three annotators assigned the code to the document.

6 Experiments

6.1 Evaluation Measures

Traditionally, following studies in Information Retrieval, the break-even point between precision and recall was used as an evaluation measure. However, the break-even point can be regarded as the optimal score within the test data whereas the parameters for the optimal point are unknown in advance in a practical setting.

Therefore, we adopt F_1 measures, a multi-topic accuracy, and a cost-sensitive accuracy measure as evaluation metrics in our work.

6.1.1 Multi-Topic Accuracy

In this paper, the accuracy of multi-topics is measured document by document. This means that if the output labels $\hat{\boldsymbol{y}}_i$ of a document are exactly the same as the correct labels $\boldsymbol{y}_i$, then the labeling of the document is judged to be correct.

[2] http://www.computationalmedicine.org/catalog/

Table 1. Results on the Medical NLP Challenge Data

Feature	Features used								
uni-gram	x	x	x	x					
bi-gram		x	x						
tri-gram			x						
c-value				x				x	
tf-idf					x	x	x	x	x
section						x			x
negation							x	x	x
Interpretation	Scores								
MCSL-MEM/CTA Micro-average F1	0.8268	0.8263	0.8275	0.8292	0.8336	0.8386	0.8390	0.8396	0.8433
Multi-Topic AC	0.7367	0.7377	0.7367	0.7367	0.7500	0.7572	0.7551	0.7561	0.7643
BCSL-MEM/CTA Micro-average F1	0.7651	0.8264	0.8246	0.7863	0.7376	0.7444	0.7454	0.7651	0.7561
Multi-Topic AC	0.6629	0.7408	0.7398	0.6844	0.6373	0.6475	0.6486	0.6670	0.6577
BCSL-MEM Micro-average F1	0.7200	0.8105	0.8164	0.7440	0.6603	0.6550	0.6698	0.7049	0.6632
Multi-Topic AC	0.5912	0.6301	0.6424	0.6290	0.5307	0.5277	0.5420	0.5809	0.5410
MCSL-SVM/CTA Micro-average F1	0.7727	0.7947	0.7953	0.7851	0.8039	0.8147	0.8035	0.8037	0.8147
Multi-Topic AC	0.6762	0.7079	0.7080	0.6854	0.7182	0.7295	0.7172	0.7172	0.7295
BCSL-SVM/CTA Micro-average F1	0.8158	0.8198	0.8208	0.8196	0.8344	0.8414	0.8322	0.8306	0.8417
Multi-Topic AC	0.7254	0.7326	0.7336	0.7285	0.7520	0.7623	0.7541	0.7520	0.7643
BCSL-SVM Micro-average F1	**0.8380**	**0.8454**	**0.8437**	**0.8452**	**0.8624**	**0.8584**	**0.8634**	**0.8672**	**0.8594**
Multi-Topic AC	**0.7396**	**0.7613**	**0.7602**	**0.7581**	**0.7859**	**0.7848**	**0.7859**	**0.7900**	**0.7848**

$$AC = \frac{1}{n}\sum_{i=1}^{n}\delta(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i),$$

where $\delta(\boldsymbol{x}, \boldsymbol{y})$ is:

$$\delta(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 1 & if\ \ \boldsymbol{x} = \boldsymbol{y} \\ 0 & otherwise \end{cases}$$

The above definition follows that of the accuracy in [15].

6.1.2 Cost-Sensitive Accuracy Measure [9]

The cost-sensitive accuracy measure that was used in the NLP Challenge 2007 is a generalized version of Jaccard's similarity metric, which was introduced in [2].

The under-coding (a false positive) leads to the loss of the amount of revenue that a hospital would have earned if it had assigned the code. The over-coding (a false negative) causes the penalty of three times of the revenue earned by the erroneous code. The cost-sensitive measure takes into account the penalties for over-coding and under-coding.

This economic aspect will be evaluated by the cost-sensitive accuracy measure.

Let Y_x be the set of correct labels for a test set and P_x be the set of labels predicted by a system. Define the set of false positives $F_x = P_x - Y_x$ and the set of false negatives $M_x = Y_x - P_x$. Note that $-$ denotes the set subtraction.

The score is defined as:

$$score(P_x) = \left(1 - \frac{\beta \mid M_x \mid + \gamma \mid F_x \mid}{\mid Y_x \cup P_x \mid}\right)^{\alpha}.$$

For the NLP Challenge 2007, $\alpha = 1$, $\beta = 0.33$, and $\gamma = 1.0$.

6.1.3 F_1 Measure

For the evaluation of multiple topics, we use the standard definition of F_1 for multi-topic Text Classification. Precision (P) and recall (R) are computed over the document-label pairs.

$$P = \frac{\#\ system's\ correct\ labeling}{\#\ system's\ labeling}$$

$$R = \frac{\#\ system's\ correct\ labeling}{\#\ correct\ labeling}$$

$$F_1 = \frac{2RP}{R + P}$$

Precision and recall are computed for each topic and then averaged in the following two ways: the micro average F_1 is a global average throughout the test data regardless of topics; the macro average F_1 is the average of F_1 scores for all topics.

Table 2. Scores of Top 10 Systems in Terms of the Cost Sensitive Measure and 3 Annotators in the Medical NLP Challenge 2007

Team Short Name	Cost Sensitive	Micro-average F1	Macro-average F1
Szeged	0.9180	0.8908	0.7691
University of Turku	0.9126	0.8769	0.7034
University at Albany	0.9091	0.8855	0.7291
PENN	0.9088	0.8760	0.7210
Annotator A	0.9056	0.8264	0.6124
MANCS	*0.9049*	*0.8594*	*0.6676*
otters	0.9010	0.8509	0.6816
LMCO-IS & S	0.9009	0.8719	0.7760
SULTRG	0.8998	0.8676	0.7322
Annotator B	0.8997	0.8963	0.8973
GMJ_JL	0.8975	0.8711	0.7334
ohsu_dmice	0.8938	0.8457	0.6542
Annotator C	0.8621	0.8454	0.8829

6.2 Feature Extraction

In preprocessing, the documents were all lower-cased and XML tags were removed from the documents. Stop words were removed using the SMART stoplist.

We tested the following features.

- uni-gram: word uni-grams in a document
- bi-gram: word bi-grams in a document
- tri-gram: word tri-grams in a document
- c-value: c-value terms [5]
- tf*idf: uni-gram, bi-gram, tri-gram with tf*idf values higher than a threshold.
- section: whether or not features are distinguished by section.
- negation: skip bi-grams where the first word is negation words, such as *no*, *nothing*, *not*.

Throughout our experiments, the feature values are frequencies of occurrence of the words in a document. The representation of feature vectors are the same regardless the machine learning algorithms. The tf*idf threshold is determined by 10-fold cross validation on the training data. The average of the micro-average F1 scores of the 10-fold cross validation was 0.86, which is consistent to the score for the test data.

The linear kernel is used for SVMs due to better performance on the data sets than other kernels.

6.3 Results

Tables 1 shows the comparative experimental results on the Medical NLP Challenge data. SVMs and MEMs showed different characteristics. In all feature settings, however, BCSL-SVM, traditional one-vs-rest approach using SVMs, achieved the best scores over all. This implies that SVMs successfully captured the nature of the training data and well generalized the training samples. Therefore, it suggests that the Medical NLP Challenge data should be OTA-oriented. In our past experience, MEMs tend more to overfit to the training data than SVMs when the data size is small. The best score F1=0.8672 was achieved when we used the c-value term, the tf*idf filtering, and the negation feature. The introduction of the negation feature improved the performance in most of the cases.

Table 2 shows scores of the top 10 systems among 44 participants to the Medical NLP Challenge. Our system MANCS ranked 5th in terms of the cost sensitive measure and 8th in terms of micro-average F1 score. The cost-sensitive score is almost the same as the best annotation of the three annotators. The system that participated in the Medical NLP Challenge employed tf*idf, negation, and section features.

7 Related Work

The origin of Text Classification goes back to the early '60s [11]. In the late '90s, Machine Learning techniques were successfully applied to Text Classification. Support Vector Machines were applied to Text Classification in [6, 4]. Maximum Entropy Models were also applied in [8].

Furthermore, multi-label classification of multi-topic text has been investigated in the last years. AdaBoost was enhanced to handle multi-labels in [12]. In this approach, the task of assigning multi-topics to a text is regarded as a ranking of labels for the text. Ranking-based evaluation was inspired by Information Retrieval. In a Text Classification problem, however, we need a set of labels for each document more clearly. McCallum [7] proposed to use the EM algorithm to train a mixture model of multi-labels. Parametric Mixture Models (PMM) were also proposed in [13].

Maximum Entropy Models were extended to multi-labeled MEMs (MLME) in [15].

These previous approaches to multi-labeled text classification are based on the OTA in our demarcation. A document with label $\{A\}$ is not explicitly isolated from a document with multi-labels $\{A, B\}$.

8 Conclusion and Remarks

This paper investigated multi-topic aspects in automatic classification of clinical free text. The Open and Closed Topic Assumption were proposed as the end points on a spectrum of the nature of multi-topics. We have used clinical records provided in the Medical NLP Challenge 2007 in which our classification system ranked 5th among 44 groups worldwide. This paper showed that for the small training data set, conventional interpretation of multi-topics to Support Vector Machines is the most suitable approach, which suggests that the multi-topic nature of the NLP Challenge data set should be OTA-oriented.

Additional experiments we have performed on newspaper articles showed that general text is more oriented towards the CTA than the clinical text tackled during the Medical NLP Challenge. Due to the page limitation, we omitted from this paper the experimental results on general text.

Acknowledgement

This research is partly supported by JISC ASSERT project and EC IST project FP6-028099 (BOOTStrep), whose Manchester team is hosted by the JISC/BBSRC/EPSRC sponsored National Centre for Text Mining.

References

[1] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra: A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71 (1996).

[2] M. Boutell, X. Shen, J. Luo, and C. Brown, Multi-label Semantic Scene Classification, *Technical Reports 813*, Department of Computer Science, Univerisity of Rochester, 2003.

[3] K. Crammer and Y. Singer, On the Algorithmic Implementation of Multi-class SVMs, *Journal of Machine Learning Research*, 2001.

[4] S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami, Inductive Learning Algorithms and Representations for Text Categorization, *Prof. CIKM '98*, pp.148–155, 1998.

[5] K. Frantzi, S. Ananiadou, The C-value / NC-value Domain Independent Method for Multi-word Term Extraction, *Journal of Natural Language Processing*, 6(3), 145–179, 1999.

[6] Thorsten Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *Proc. of 10th European Conference on Machine Learning (ECML-98)*, pp.137–142, 1998.

[7] A. McCallum, Multi-label Text Classification with a Mixture Model Trained by EM, *AAAI-99 Workshop on Text Learning*, 1999.

[8] K. Nigam, J. Lafferty, A. McCallum, Using Maximum Entropy for Text Classification, *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp.61–67, 1999.

[9] John P. Pestian, Christopher Brew, Pawel Matykiewicz, D.J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch, A Shared Task Involving Multi-label Classification of Clinical Free Text, *in Prof. of ACL-2007 Workshop on BioNLP*, 2007. (to appear)

[10] John C. Platt, Nello Cristianini, John Shawe-Taylor, Large Margin DAGs for Multiclass Classification, *in Proc. of NIPS-1999*, pp. 547–553, 1999.

[11] Fabrizio Sebastiani, Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, Vol. 34, No.1, pp.1–47, 2002.

[12] RE Schapire and Y Singer, BoosTexter: A Boosting-based System for Text Categorization *Machine Learning*, Springer, Vol. 39, pp.135–168, 2000.

[13] N. Ueda and K. Saito, Parametric Mixture Models for Multi-Labeled Text, *Advances in Neural Information Processing Systems 15*, MIT Press, pp.737–744, 2002.

[14] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[15] Shenghuo Zhu, Xiang Ji, Wei Xu, and Yihong Gong, Multi-labelled Classification Using Maximum Entropy Method, *Proc. the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-05)*, pp. 274–281, 2005.