Learning cascaded latent variable models for biomedical text classification

Ming Liu **Gholamreza Haffari** **Wray Buntine**
Faculty of Information Technology, Monash University
ming.m.liu, gholamreza.haffari, wray.buntine @ monash.edu

Abstract

In this paper, we develop a weakly supervised version of logistic regression to help to improve biomedical text classification performance when there is limited annotated data. We learn cascaded latent variable models for the classification tasks. First, with a large number of unlabelled but limited amount of labelled biomedical text, we will bootstrap and semi-automate the annotation task with partially and weakly annotated data. Second, both coarse-grained (document) and fine-grained (sentence) levels of each individual biomedical report will be taken into consideration. Our experimental work shows this achieves higher classification results.

1 Introduction

In recent years, large amounts of biomedical text have become available with the development of electronic medical record (EMR) systems. The type of biomedical text ranges from reports of CT scans to doctoral notes and discharge summaries. Based on these biomedical text, there are medical tasks such as disease identification, diagnostic surveillance and evaluation and other clinical support services. Manual extraction and classification for these medical tasks from biomedical text is a time-consuming and often costly effort.

Biomedical text classification systems which consider both manual effort (e.g. annotation) and predictive performance are more appropriate in the medical context than those which only consider classification predictive performance. Early biomedical classification methods are rule-based (Tinoco et al., 2011; Matheny et al., 2012), which requires medical experts to develop logical rules to identify reports consistent with some diseases. The main advantages of such rule-based systems is that high precision can be achieved, but the weakness lies in the fact that the process is not easily transferable to similar tasks, because medical experts have to carefully develop specific types of rules and formulas for different kinds of diseases. In recent years, machine learning methods have been widely used in disease identification from biomedical text(Ehrentraut et al., 2012; Bejan et al., 2012; Martinez et al., 2015; Hassanpour and Langlotz, 2015), which also ask medical experts to do some annotation work for building training data. Unlabelled free biomedical text in hospitals and other clinical organizations is abundant but manual annotation is very expensive.

Exploiting fine-grained sentence level properties for coarse-grained document level classification has attracted large amounts of attention. Pang(Pang and Lee, 2004) first explored subjectivity extraction methods based on a minimum cut formulation, in which they performed subjectivity detection on individual sentences and implemented document level polarity classification by leveraging those extracted subjective sentences. McDonald(Täckström and McDonald, 2011) proposed a structured model for jointly classifying the sentiment of text at varying levels of granularity, they showed that this task can be reduced to sequential classification with constrained inference. Yessenalina(Yessenalina et al., 2010) described a joint two-level approach for document level sentiment classification that simultaneously extracts useful sentences, and Fang(Fang and Huang, 2012) extended it by incorporating aspect information to the structured model to aspect level sentiment analysis.

In this paper, we propose a cascaded latent variable model for biomedical text classification that combines logistic regression and EM, which is trained with a large number of unlabelled but limited amount of labelled biomedical text. Exper-

Ming Liu, Gholamreza Haffari and Wray Buntine. 2016. Learning cascaded latent variable models for biomedical text classification. In *Proceedings of Australasian Language Technology Association Workshop*, pages 139−143.

imental results show that the combined cascaded model is efficient in biomedical text classification tasks.

2 Methodology

In this section, we propose variants developed from a cascaded logistic regression model: the partially supervised model called as logistic regression with hard EM (LREM) and the weakly supervised model named as weak logistic regression with hard EM (WLREM). LREM is trained with part of the fully-annotated data and all of the partially-annotated data. WLREM is trained with the same part of the fully-annotated data and all of the weakly annotated data.

2.1 Preliminaries

Let d be a document consisting of n sentences, $\mathbf{X} = (X_i)_{i=1}^n$, with a document-sentence-sequence pair denoted $\mathbf{d} = (d, \mathbf{X})$. Let y^d denote the document level polarity and $\mathbf{Z} = (Z_i)_{i=1}^n$ be the sequence of sentence level polarity. In what follows, we assume that there are three types of training sets: a small set of fully labeled instances D_F which are annotated at both sentence and document levels, another small set of partially labeled instances D_P which are annotated only at the document level, and a large set of weakly annotated instances D_U (explained later). Besides, we assume that all Z_i take values in $\{POS(+1), NEG(-1), NEU(0)\}$ while y^d is in $\{POS(+1), NEG(-1)\}$.

The following three cascaded models are based on logistic regression, with the following standard parametrization

$$P_\theta\left(y^d|\mathbf{X}\right) = \sum_{\mathbf{Z}} P_\alpha(y^d|\mathbf{Z}) P_\beta(\mathbf{Z}|\mathbf{X}) \quad (1)$$

where $\theta = \{\alpha, \beta\}$, and α and β are the parameters of document and sentence level classifiers respectively.

2.2 The partially supervised model

The partially supervised model (LREM) is trained from the sets of fully labeled data D_F and partially labelled data D_P. Since the sentence polarity is unknown in D_P, a hard EM algorithm is used to iteratively estimate $\mathbf{Z}$ and maximize the cascaded goal function. Figure 1 outlines LREM. The parameters, α and β, of this model can be estimated by maximizing the joint conditional likelihood function

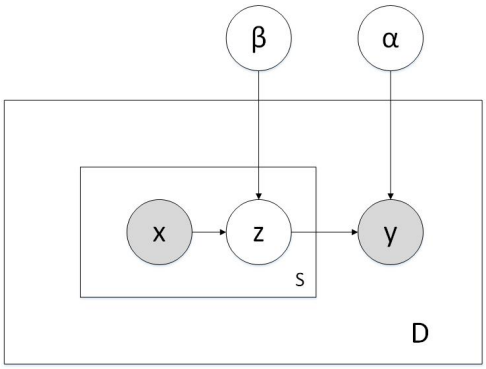

Figure 1: A partially supervised model.

$$\alpha, \beta = \arg\max_{\alpha,\beta} \Big(\sum_{d=1}^{N} \log P_\theta(y^d|\mathbf{X}) \Big) \quad (2)$$

where $N = |D_F \cup D_P|$.

2.3 The weakly supervised model

The weakly supervised model (WLREM) is trained from the sets of fully labeled data D_F and weakly labelled data D_U. In our case, the document polarity is unknown from D_U, while U represents the patient level diagnostic result in the treating hospital. Generally, if a patient is diagnosed with positive infection in the hospital, the reports of this patient are more likely to be positive. We get this estimated probability from a confusion matrix of D_F as shown in table 1. We

Table 1: Confusion matrix of fully-annotated dataset

D_F	y=POS	y=NEG
U=POS	167	68
U=NEG	41	82

notice that $P(U = POS|y = POS) = 0.803$, which is a trustful prior information for guessing y, thus we can extend the previous partially supervised model into a weakly one. Figure 2 shows WLREM. The parameters, α and β, of this model can be estimated by maximizing the joint conditional likelihood function

$$\begin{aligned} P_\theta(U^d|\mathbf{X}) &= \textstyle\sum_{y,\mathbf{Z}} P_\beta(\mathbf{Z}|\mathbf{X}) P_\alpha(y^d|\mathbf{Z}) P(U^d|y^d) \\ \alpha, \beta &= \arg\max_{\alpha,\beta} \Big(\textstyle\prod_{d=1}^{M} P_\theta(U^d|\mathbf{X}) \Big) \end{aligned} \quad (3)$$

where $M = |D_F \cup D_U|$.

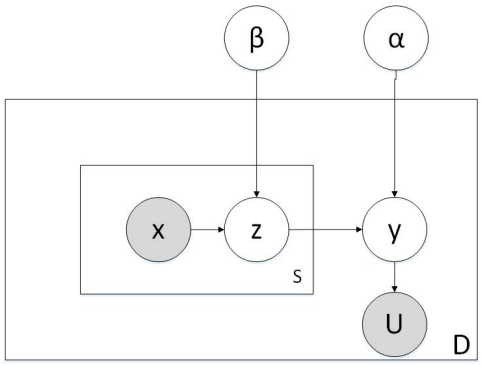

Figure 2: A weakly supervised model.

3 Combining partial and weak supervision

The partially and weakly supervised models both have their merits. The former requires document level annotation, while the latter can be used directly with available documents except for an initial guess of the document level polarity. In order to achieve the best predictive performance, we propose to combine the merits of these two models.

3.1 A combined cascaded latent variable model

Given in Algorithm 1, ComLREM is an integration of the above two models (LREM+WLREM), which can make full use of the partially and weakly annotated data.

Algorithm 1 ComLREM

$\alpha, \beta \leftarrow$ update for data D_F via Eqn (2)
$Z_i \leftarrow 0$ for all the sentences in $D_P \cup D_U$
$y \leftarrow 1$ for all the documents in D_U
while the convergence condition is meet **do**
  for every document $d \in D_P$ **do**
    $n_d \leftarrow$number of sentences of d
    for $k = 1$ to n_d in d **do**
      $Z_k^d = \arg\max_{Z_k^d} P_\theta\left(y^d|\mathbf{X}\right)$
        ▷ from Equation (1)
  for every document $d \in D_U$ **do**
    $n_d \leftarrow$number of sentences in d
    for $k = 1$ to n_d in d **do**
      $Z_k^d, Y^d = \arg\max_{Z_k^d, Y^d}$
        $P_\beta(\mathbf{Z}|\mathbf{X}) P_\alpha(y^d|\mathbf{Z}) P(U^d|y^d)$
  $\alpha, \beta \leftarrow$ update for all data via Eqns (2), (3)

Table 2: Feature representation

Feature level	Discription
Sentence-level	Uni-gram tokens + MetaMap concepts
Report-level	Pos sentence exists or not Neg sentence exists or not No. of pos sentences No. of neg sentences No. of other sentences Polarity of the first sentence Polarity of the last sentence Percentage of pos sentences Percentage of neg sentences Pos sentence exists in the beginning Pos sentence exists in the end Neg sentence exists in the beginning Neg sentence exists in the end

3.2 Feature representation

Two main types of features are explored: Bag and Structural. Bag features are applied to the sentence-level classification, while structural features are built on the results of sentence-level classification.

Dates, time and numbers are normalised into DATE, TIME, and NUM symbols. Reports are segmented into sentences using the JulieLab (Tomanek et al., 2007) automatic sentence segmentor. Stop words are terms and phrases which are regarded as not conveying any significant semantics to the sentences and reports, NLTK stop word list was chosen to do the filtering. The Genia Tagger (Tsuruoka et al., 2005) is used to do tokenization and lemmatization. The MetaMap concepts (Aronson, 2001) come from the mappings of biomedical knowledge representation. Table 1 illustrates the feature representation at the sentence and report levels.

4 Experiment

As shown in (Martinez et al., 2015), CT reports for fungal disease detection were collected from three hospitals. For each report, only the free text section were used, which contains the radiologist's understanding of the scan and the reason for the requested scan as written by clinicians. Every report was de-identified: any potentially identifying information such as name, address, age/birthday, gender were removed. Table 2 shows the number of distribution of reports over fully-annotated, partially-annotated and verified data sets.

Receiver operating characteristic (ROC) curve and Precision recall (PR) curve are used for the model evaluation. Area under ROC curve and

Table 3: Fully-annotated, partially-annotated and weakly annotated datasets

Datasets	D_F	D_P	D_U
Pos fungal	150	51	431
Neg fungal	208	53	816

PR curve is an estimated measure of the test accuracy.The results presented here are 5-fold cross validation outcomes on the fully-annotated data.

Fig. 3 and 4 show the ROC curves and PR curves of the four models: LR is the baseline algorithm, LREM is trained based on part of the fully-annotated data and all of the partially-annotated data, WLREM is trained based on part of the fully-annotated data and all of the unannotated data, and ComLREM is an integration of the above two models.

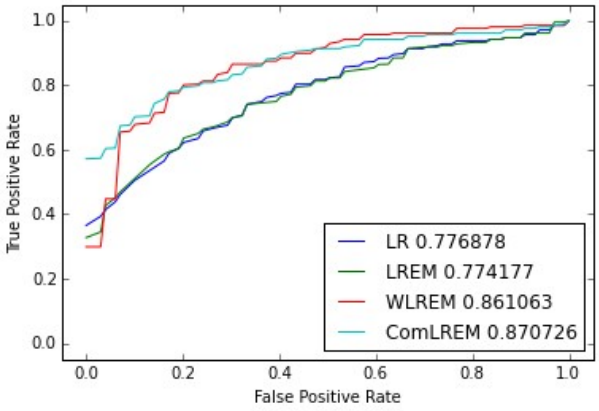

Figure 3: ROC curve of LR, LREM, WLREM and ComLREM.

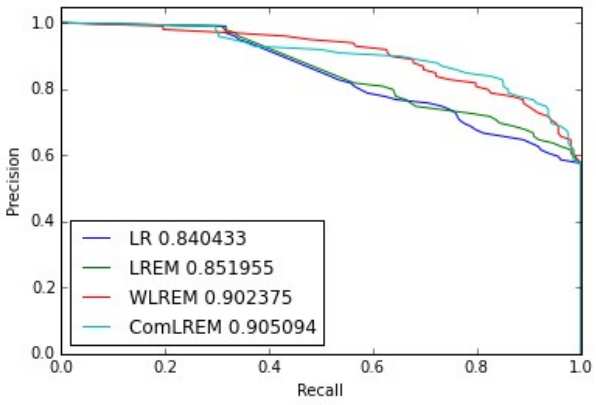

Figure 4: PR curve of LR, LREM, WLREM and ComLREM.

We can see from Fig. 3 that WLREM obtained higher ROC score than LR, the area under LREM and WLREM ROC curve is 0.774 and 0.861, which shows that the involvement of weakly annotated data contributes higher than that of partially annotated data to the improvement of classification performance. It is noticed WLREM achieved greater improvement than LREM, because the D_U contains big volume and trustful prior information. The highest ROC score (0.870) was achieved with a combination of the above two models, which is within our expectation. Fig. 4 shows the PR curves of the four models, there is a trade-off between precision and recall with recall as the most important metric. When the threshold is set to obtain a high recall (> 0.9), ComLREM obtained higher precision than other models. Overall, with true positive rate or recall as the first priority, the combined model ComLREM achieved the best classification performance.

We also compared our model with Martinez's system (Martinez et al., 2015), in which they applied conservative rules over sentence-classification output. Their sentence-level classifier used SVMs with Bag-of-words and Bag-of-concepts features. Since the conservative rules indicate that a report is labeled as positive if any sentence in it is labeled positive, the report-level prediction is not probabilistic and the PR curve can not be drawn accordingly. In order to make some comparison, we adjusted the threshold of our report-level logistic regression classifier to make our recall the same as theirs (0.930), and see whether the precision improves. Table 3 shows the compared results, we noticed that both WLREM and ComLREM outperforms the Conservative SVM approach, which indicates that the estimation we made from the unlabelled data is trustful and can be used to improve classification performance.

Table 4: Comparison of the experimental results

Models	**Recall**	**Precision**	**F score**
Conservative SVM	0.930	0.694	0.795
LR	0.930	0.646	0.762
LREM	0.930	0.656	0.769
WLREM	0.930	0.703	0.801
ComLREM	0.930	0.707	0.802

5 Conclusion

Learning classification models in a fully supervised manner is expensive in the biomedical do-

main. We therefore proposed a combined cascaded latent variable model, which effectively combines both partial and weak supervision for biomedical text classification. Sentence label is regarded as a latent variable in this model, and both fine-grained and coarse-grained features are considered in the learning process. In the future, we consider to develop active learning methods towards our cascaded latent variable model and further reduce manual annotation cost.

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Cosmin Adrian Bejan, Fei Xia, Lucy Vanderwende, Mark M Wurfel, and Meliha Yetisgen-Yildiz. 2012. Pneumonia identification using statistical feature selection. *Journal of the American Medical Informatics Association*, 19(5):817–823.

Claudia Ehrentraut, Hideyuki Tanushi, Hercules Dalianis, and Jörg Tiedemann. 2012. Detection of hospital acquired infections in sparse and noisy Swedish patient records. *A machine learning approach using Naïve Bayes, Support Vector Machines and C*, 4.

Lei Fang and Minlie Huang. 2012. Fine granular aspect analysis using latent structural models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 333–337. Association for Computational Linguistics.

Saeed Hassanpour and Curtis P Langlotz. 2015. Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine*.

David Martinez, Michelle R Ananda-Rajah, Hanna Suominen, Monica A Slavin, Karin A Thursky, and Lawrence Cavedon. 2015. Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. *Journal of biomedical informatics*, 53:251–260.

Michael E Matheny, Fern FitzHenry, Theodore Speroff, Jennifer K Green, Michelle L Griffith, Eduard E Vasilevskis, Elliot M Fielstein, Peter L Elkin, and Steven H Brown. 2012. Detection of infectious symptoms from va emergency department and primary care clinical documentation. *International journal of medical informatics*, 81(3):143–156.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Oscar Täckström and Ryan McDonald. 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 569–574. Association for Computational Linguistics.

Aldo Tinoco, R Scott Evans, Catherine J Staes, James F Lloyd, Jeffrey M Rothschild, and Peter J Haug. 2011. Comparison of computerized surveillance and manual chart review for adverse events. *Journal of the American Medical Informatics Association*, 18(4):491–497.

Katrin Tomanek, Joachim Wermter, Udo Hahn, et al. 2007. A reappraisal of sentence and token splitting for life sciences documents. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, page 524. IOS Press.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Junichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *Advances in informatics*, pages 382–392.

Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1046–1056. Association for Computational Linguistics.