COMP 599 Project: Understanding Medical Records - NLP for Automatic Code Assignment using Structured and Contextual Information*

Pulkit Khandelwal - 260717316
pulkit.khandelwal@mail.mcgill.ca

Abstract

Natural Language Processing techniques have eased automatic categorization of medical reports in recent times. In this project, as part of COMP 599 course at McGill University, new methods have been proposed in this direction which exploit both structural and contextual knowledge of the text and that shows a good boost over existing methods.

1 Introduction

Electronic Health Record (EHR) Systems are generating enormous amounts of clinical data which serves as a rich source for research purposes. The modern hospital generates large volumes of data, which include discharge summaries, records of medicines administered, laboratory results, treatments provided and past medical histories. Automated extraction of textual information can be beneficial to predict patient outcomes, monitor drug administration, track diseases and improve medical procedures.

NLP techniques have been used to predict patient outcomes in the form of ICD codes (Ira Goldstein and Anna Arzumtsyan, 2007) assignment from discharge summaries. (Solt et al., 2009) did the first work in this direction which employed both rule based and machine learning approaches for automatic code assignment. (Saria et al., 2010) uses specialized cues such as uncertainty modifiers, negations to disambiguate cases where previous NLP approaches fail.

Previous work (Solt et al., 2009) has focused on Bag of Words model, language modeling techniques and hand crafted rules along with machine learning algorithms such as logistic regression. These methods fail to scale well and provide very little for any practical use. More recent work has used hierarchical SVMs to tackle the problem (Wei et al., 2016). With the boom of recent Deep Learning trend, one can achieve much better results given appropriate computational resources.

In this paper, the author tries to understand the previous work and arrives at a better representation of the data which has not been done before. The author leverages this information to propose deep learning models which explore more contextual and structural information in the data. Comparisons have been made with some baseline models to show practicality and novelty of the proposed models.

The rest of the paper is structured as follows: Section 2 describes the data and how it is preprocessed, Section 3 benchmarks some baseline models, Section 4 solves the class imbalance problem, Section 5 deals with newer models, Section 6 summarizes the results and finally Section 7 wraps up with a brief discussion.

2 Data and its preprocessing

MIMIC (Medical Information Mart for Intensive Care) III (Johnson et al., 2016) is used. It is a collection of radiology reports and discharge summaries of around 50,000 patients across 53,000 hospitals in the USA. All the reports have been assigned ICD 9

*The reader of this manuscript is encouraged to have a look at the supplementary material which follows at the end of the report.

codes. There are about 17,000 ICD 9 codes. These ICD 9 codes come from 18 main categories. For the purpose of this project all of the labels (code) for each entry have been converted to one of these 18 categories. Thus, there are 18 labels which is basically the type of disease associated with each patient such as *Mental Disorders* or *Congenital Anomalies* etc. The data contain over 2 million reports and hence has a very good record of each patient's history which forms a basis for longitudinal analysis. For this project a well represented portion of the data is used. 80:20 train and validation split is used.

The data is preprocessed using the standard pipeline: lemmatization according to POS tags, conversion to lower case, removal of stop words and punctuation.

There are some existing challenges in medical records such as class imbalance and label noise (Johnson and Thirman, 2015) as seen in Figure 3. Previous works have not tackled this issue which results in low performance in the classification task. Here, a number of techniques have been used to balance the dataset by under and over sampling methods such as Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), MCMC sampling (Li et al., 2016); the use of penalized algorithms and ensemble methods. Explanation for this is given in the following Sections.

3 Baseline Models

For classification of reports into its appropriate category using the baseline models, continuous bag of words (Damashek, 1995) were used as features. 24062 features were then extracted from the vocabulary. Unigrams were used to form the CBOW representation.

Performance of Logistic Regression, Naive Bayes, Support Vector Machine with RBF (Radial basis function) kernel, Linear and Quadratic Discriminant Analysis (LDA and QDA), a simple multilayer perceptron on the validation set are compared in Table 1. Ten hidden layers with 20 nodes each and optimizer *adam* is used to develop the MLP architecture. The choice made is obvious due to its performance record in the literature (LeCun et al., 2015).

Model	Acc	Pre	Recall	F1-score
Naive Bayes	43.00	41.00	43.00	42.00
Logistic Reg	50.86	49.00	51.00	50.00
SVM	45.63	34.00	46.00	33.00
LDA	16.09	36.00	16.00	20.00
QDA	11.00	9.00	6.00	5.00
MLP	51.90	50.00	52.00	51.00

Table 1: Comparison of Baseline Models on Validation Set. All values in percentage.

4 Solving the class imbalance problem

As mentioned the classes are not balanced. *Label 7* has around 19,000 data points in contrast to *Label 1* which has just 105 entries. Penalized versions of Logistic Regression and SVM were tested. The penalized versions adjust weights as inversely proportional to class frequencies. This gives more weight to under-represented classes. The second approach used was to use an ensemble of the baseline models to see if there is any boost in performance. Boosting and Bagging approaches are good for such issues.

Alternatively, we can try to go back to the data and start from there. Sampling the data to have a nearly equal representation of all the classes will improve performance. Here, SMOTE is used which increases the number of samples of the undersampled classes. *imbalanced-learn* (Lemaître et al., 2016) which is a *scikit* (Pedregosa et al., 2011) package is used. SMOTE's performance is observed for Logistic Regression and Naive Bayes as they showed better results than other models as seen in Section 3.

Earlier this month (Li et al., 2016) proposed a data sampling method based on Monte Carlo methods. Also, one can change the perspective of the problem in the sense that the underrepresented class can be considered as an outlier and hence outlier or anomaly detection can be used. Additionally the over represented class can be sub-categorized in some other classes.

The results for these experiments are tabulated in Table 2.

5 Deep Learning methods for contextual and structured knowledge

While CBOW are good feature descriptors they fail to capture the context of the words. Skip-grams

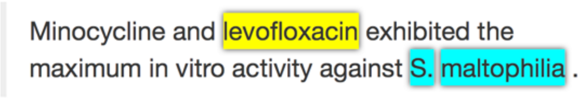
Minocycline and levofloxacin exhibited the maximum in vitro activity against S. maltophilia .

Figure 1: Contextual Information. *Source: ”Concept Linking for Text”*

can be used for this. It is now a common practice to use vector representations of words. Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) give adequate vectors for each word in the vocabulary of a corpus. These vectors can be used as feature descriptors for NLP tasks. Contextual information is stored in such vectors.

With the recent advances in Deep Learning, it is now also possible to find structure and relations among different regions in text documents. Convolutional Neural Networks (Krizhevsky et al., 2012), Recursive and Recurrent Neural Networks (Mikolov et al., 2010) have proven to be efficient at structured prediction and in dealing with longitudinal data. Deep learning Models scale very well with large amounts of data given appropriate computational resources.

Both Word2Vec and GloVe have been generated for the given data. Some examples: Similarity between *hospital* and *patient* is 90.43 %, between *blood* and *ventricle* is 88.30 % and between *scan* and *chest* is 97.48 %.

A ConvNet has been trained and validated. RNNs with three flavours namely: (Long Short-term Memory) LSTMs, (Gated Recurrent Units) GRUs and SimpleRNN have been implemented. *Adam* is the optimizer used with *relu* as the activation, *softmax* classifier for the final output layer and *categorical crossentropy* is used as the loss function. Each word vector has a dimension of 100. For these deep learning models, *GloVe* is used. The choice of hyperparameters is based on the results of state of the art hyperparamters used in the literature (LeCun et al., 2015). The representation of model architectures can be found in the supplementary material.

Performance of these models on structured classification task is tabulated in Table 3.

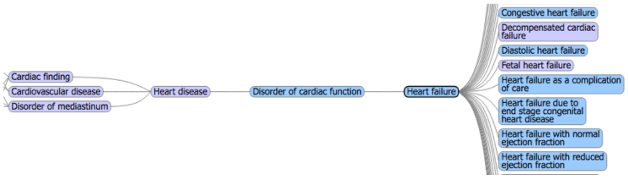
Cardiac finding
Cardiovascular disease
Disorder of mediastinum
Heart disease
Disorder of cardiac function
Heart failure
Congestive heart failure
Decompensated cardiac failure
Diastolic heart failure
Fetal heart failure
Heart failure as a complication of care
Heart failure due to end stage congenital heart disease
Heart failure with normal ejection fraction
Heart failure with reduced ejection fraction

Figure 2: Structural Information. *Source: ”Concept Linking for Text”*

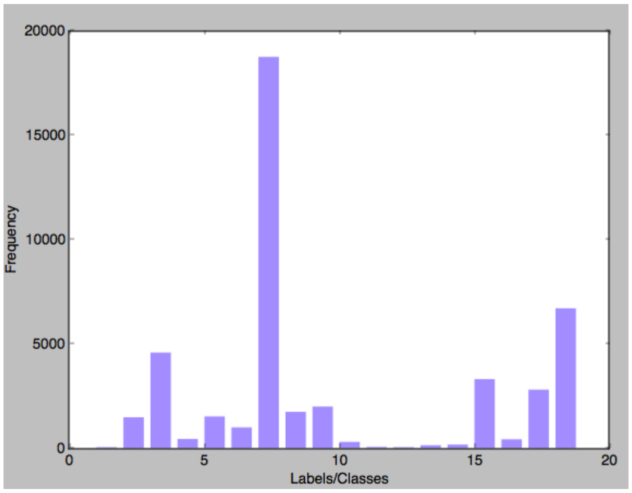

Figure 3: Label Distribution showing how imbalanced the classes are

Model	Acc	Pre	Recall	F1-score
Penalized Logistic Reg	2.30	1.00	3.00	1.00
Penalized SVM	1.90	0.00	2.00	0.00
Ensemble Method	51.90	50.00	52.00	51.00
Naive Bayes (SMOTE)	90.22	97.00	90.00	93.00
Logistic Reg (SMOTE)	98.78	100.00	99.00	99.00

Table 2: Comparison of Baseline Models after solving the class-imbalance problem on Validation Set. All values in percentage

Model	Train Loss	Val Loss	Train Acc	Val Acc
ConvNet	2.34	2.27	0.37	0.37
SimpleRNN	2.06	1.73	0.35	0.35
GRU	2.54	1.98	0.22	0.35
LSTM	1.97	2.05	0.39	0.39

Table 3: Comparison of performance of Deep Learning Models

6 Results

6.1 Baseline Models

The baseline models performed good enough with multi layer perceptron having the best results at 51.90 % accuracy. Logistic Regression and SVM followed next. QDA failed miserably.

6.2 Solving the class imbalance problem

The classes are now well represented in the dataset. This new distribution gave better results as expected but at the cost of training time. Niave Bayes and Logistic Regression achieved accuracies of 90.22 % and 98.78 % repectively. One can see a jump of about 40 % over the baseline models. Ensemble methods without SMOTE did not provide us with usable results. Penalized versions of SVM and Logistic Regression failed and performed even worse than the baseline models.

6.3 Deep Learning methods for contextual and structured knowledge

A vocabulary of around 40,000 word vectors was generated from the entire corpus of the medical reports. Word vectors and deeper architectures performed quite well but did not achieve better results than the models described in *Section 6.3*. The models converged after 10 epochs. ConvNet performed decently giving an accuracy of 37 % on the validation set. For the recurrent networks, SimpleRNN and GRUs were comparable at 35 %. LSTM performed the best at 39 %.

7 Discussion

A series of baseline models were implemented to classify medical reports into their appropriate categories. The class imbalance problem was solved after a careful examination of various methods. This was fruitful because the results were quite impressive as discussed above. Figures 1 and 2 shows how context plays an important role in seeing long term structures in the data. Given that the reports has medical records of the patients for a considerable amount of time, a longitudinal analysis is done by the extraction of cues. The cues here are structured data which was obtained through the use of recurrent neural networks. Word vector representations were found to be effective for these purposes.

8 Conclusion and Future Work

Thus, the classification of medical reports has been done in this project. The data is thoroughly understood and cleansed. A bunch of baseline models were implemented which were then improvised by solving the first issue being tackled in this project: class-imbalance. Newer and more advanced models were then employed in the second part of the project to discover more knowledge in the data.

There are several possible future works which can be proposed here. Alternate representation of words can be used such as tf-idf and Skip-thought vectors (Kiros et al., 2015). Outside knowledge can be incorporated to better model the data. More advanced and deeper architectures could be employed to boost performance.

The impact which NLP can make in the field of health care is unimaginable given the amount of resources at our disposal. Also, not much work has been done in this sector before and thus is a relatively new field for researchers to get their hands dirty with.

Acknowledgments

The author would like to thank the course instructor and the teaching assistant for putting up this course together. It was a pleasure studying NLP.

References

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Marc Damashek. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843.

MBA Ira Goldstein and MLS Anna Arzumtsyan. 2007. Three approaches to automatic assignment of icd-9-cm codes to radiology reports.

Justin Johnson and Daniel Thirman. 2015. Medical record understanding. *Stanford University*, 1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2016. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *CoRR*, abs/1609.06570.

Chengtao Li, Suvrit Sra, and Stefanie Jegelka. 2016. Markov chain sampling in discrete probabilistic models with constraints. In *Advances In Neural Information Processing Systems*, pages 4188–4196.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

Suchi Saria, Gayle McElvain, Anand K Rajani, Anna A Penn, and Daphne L Koller. 2010. Combining structured and free-text data for automatic coding of patient outcomes. In *AMIA Annu Symp Proc*, volume 2010, pages 712–716. Citeseer.

Illés Solt, Domonkos Tikk, Viktor Gál, and Zsolt T Kardkovács. 2009. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. *Journal of the American Medical Informatics Association*, 16(4):580–584.

Wei-Qi Wei, Pedro L Teixeira, Huan Mo, Robert M Cronin, Jeremy L Warner, and Joshua C Denny. 2016. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association*, 23(e1):e20–e27.

Supplementary Material

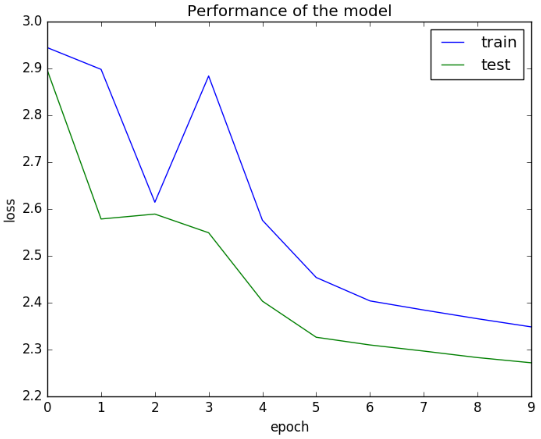

Figure 4: Loss vs Number of Epochs: ConvNet

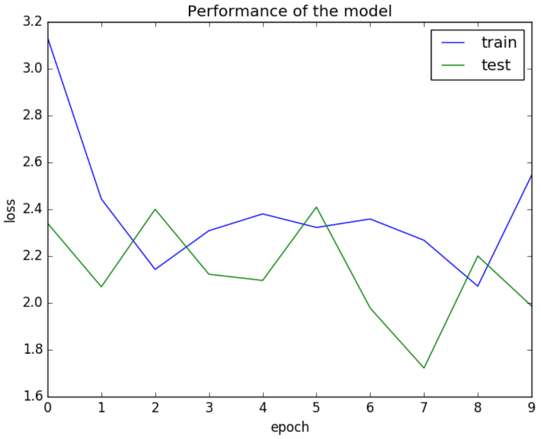

Figure 5: Loss vs Number of Epochs: SimpleRNN

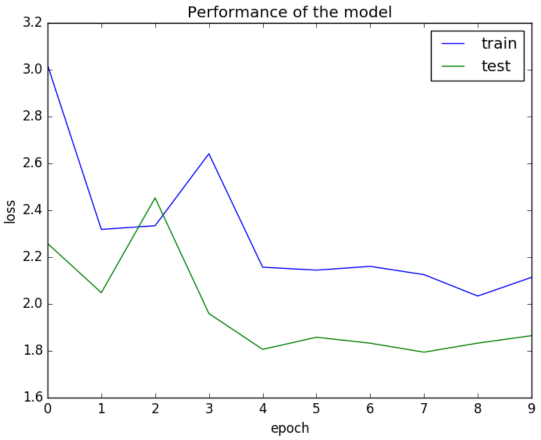

Figure 6: Loss vs Number of Epochs: GRU

Label	Description (diseases)
1	Infectious And Parasitic
2	Neoplasms
3	Endocrine, Nutritional And Metabolic
4	Blood And Blood-Forming Organs
5	Mental Disorders
6	Nervous System And Sense Organs
7	Circulatory System
8	Respiratory System
9	Digestive System
10	Genitourinary System
11	Pregnancy, Childbirth, And The Puerperiun
12	Skin And Subcutaneous Tissue
13	Musculoskeletal System And Connective Tiss
14	Congenital Anomalies
15	Perinatal Period
16	Ill-Defined Conditions
17	Injury And Poisoning
18	Contact With Health Services

Table 4: Labels

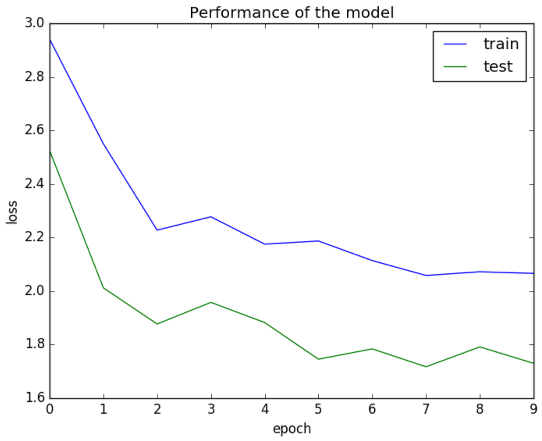

Figure 7: Loss vs Number of Epochs: LSTM

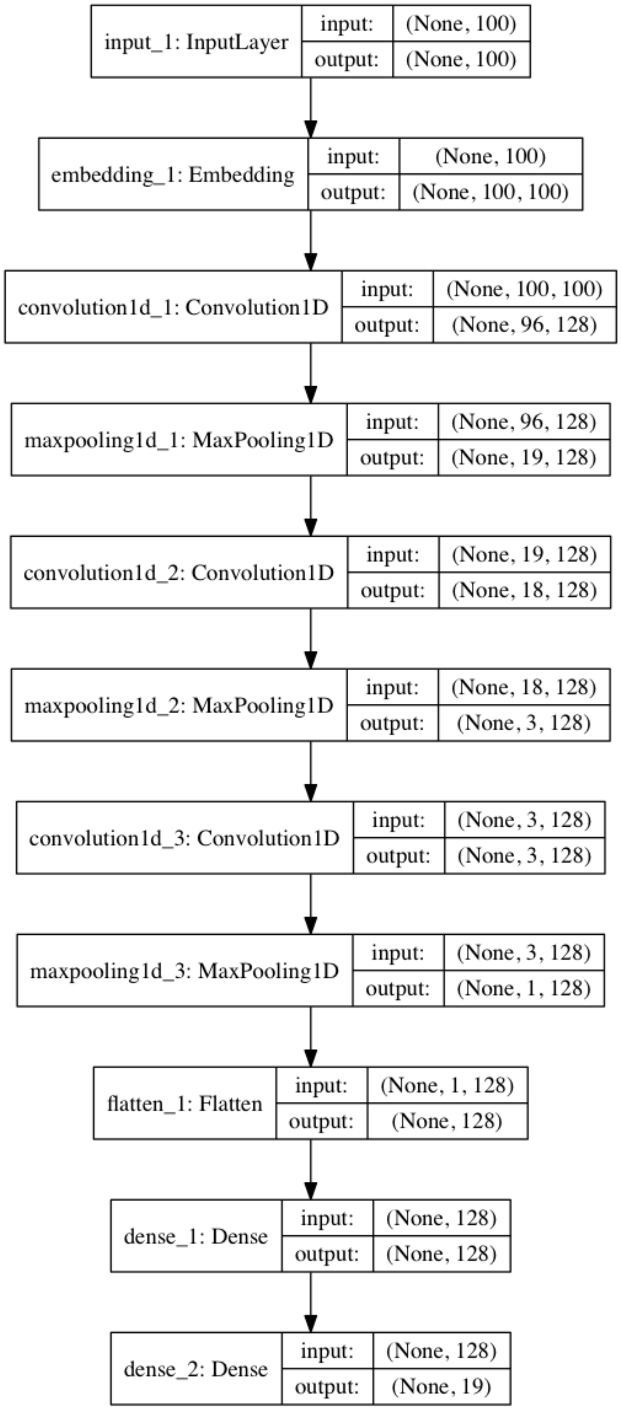

Figure 8: ConvNet Architecture

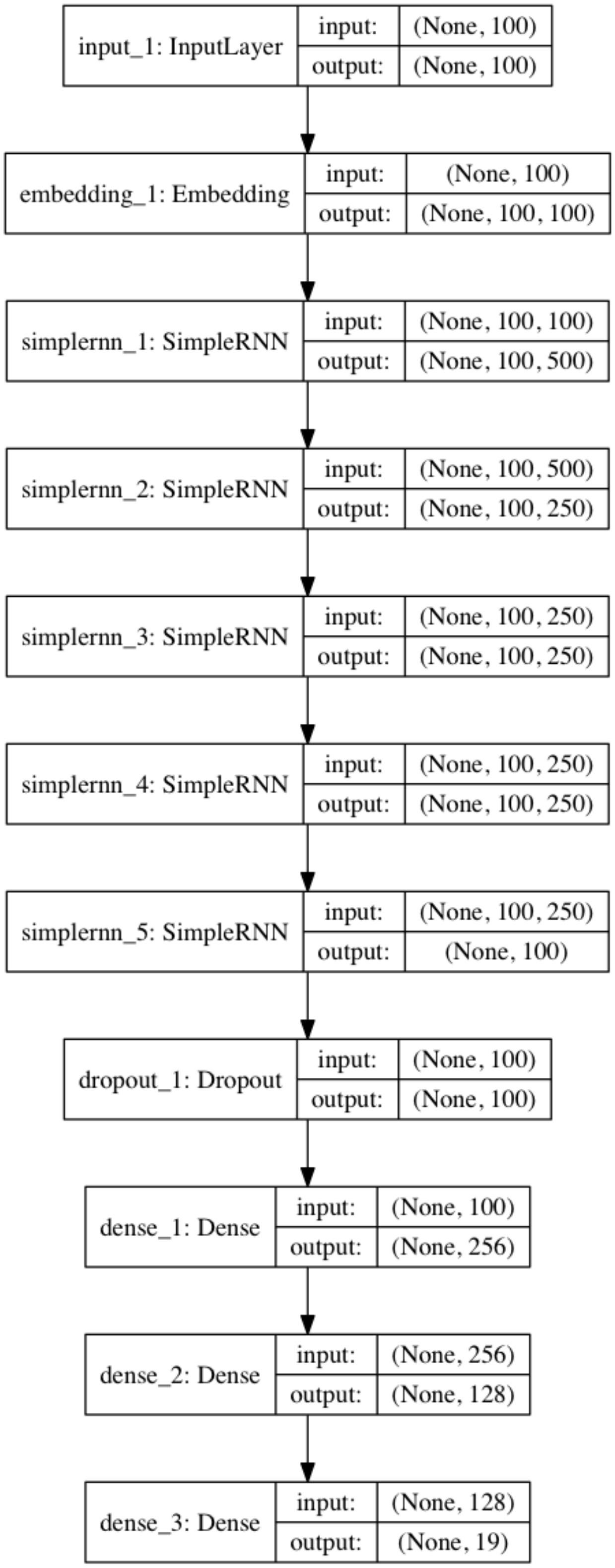

Figure 9: SimpleRNN Architecture

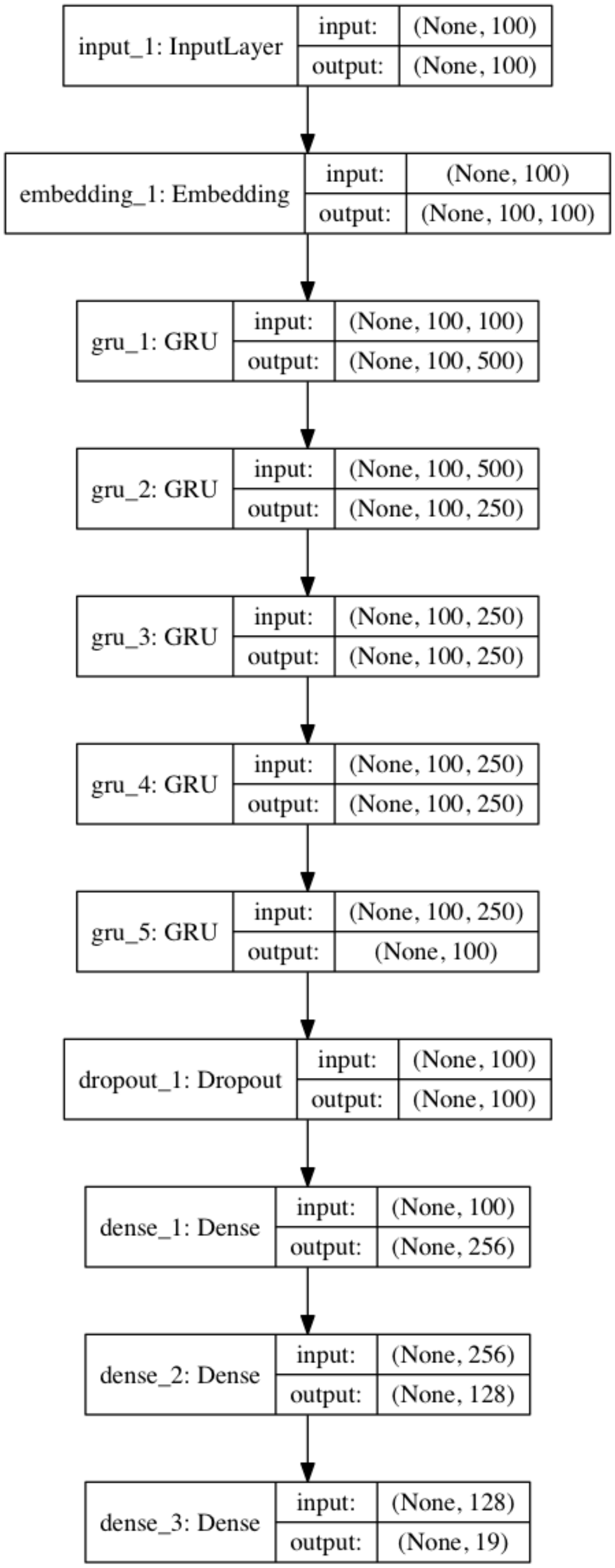

Figure 10: GRU Architecture

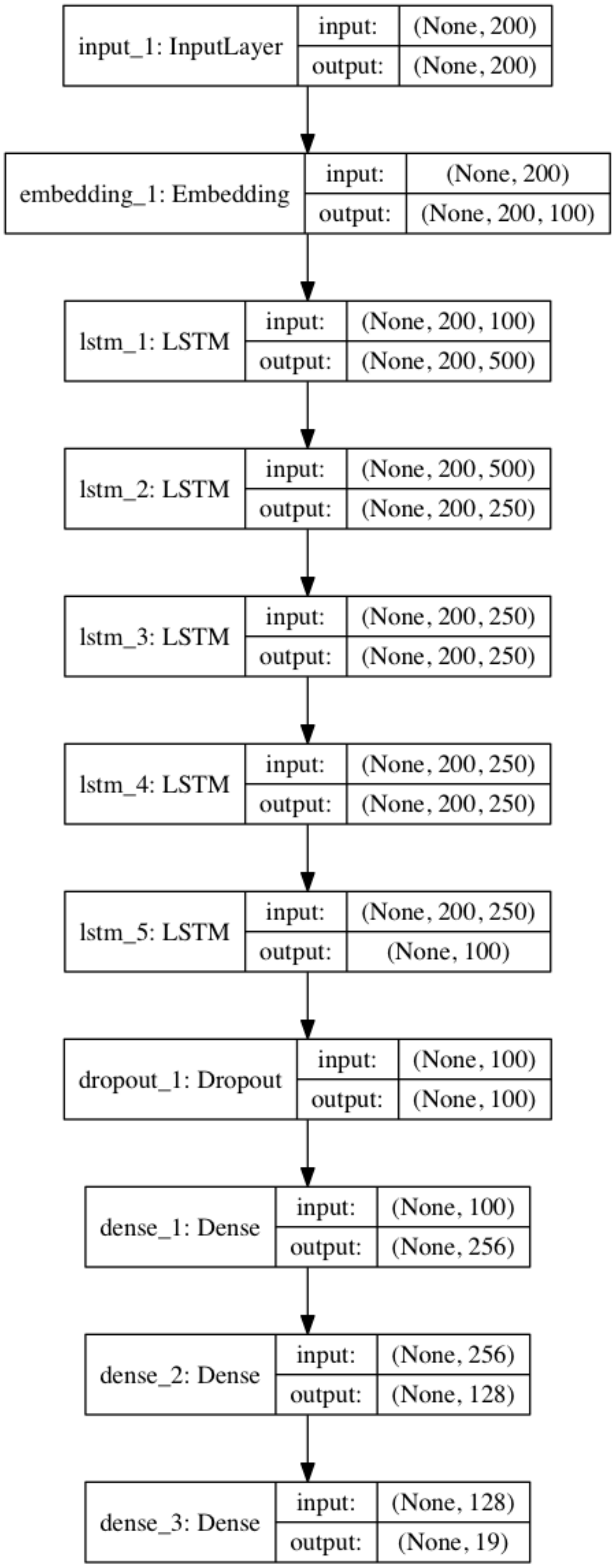

Figure 11: LSTM Architecture