ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 46-51
doi: 10.18178/wcse.2020.02.008

ICD-10 Auto-coding System Using Deep Learning

Ssu-Ming Wang [1], Feipei Lai [1,2+], Chang-Sung Sung [2] and Yang Chen [2]

[1] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan Univ., Taipei, Taiwan
[2] Department of Computer Science and Information Engineering, National Taiwan Univ., Taipei, Taiwan

Abstract. In this research, we aim to construct an automatic ICD-10 coding system. ICD-10 is a medical classification standard which is strongly related to scope of payment in health insurance. However, the work of ICD-10 coding is time-consuming and tedious to ICD coders. Therefore, we build an ICD-10 coding system based on NLP approach to reduce their workload. The result of f1-score in whole label prediction task is up to 0.67 and 0.58 in CM and PCS, respectively. In addition, recall@20 in whole label prediction task is up to 0.87 and 0.81 in CM and PCS, respectively. In the future, we will keep working on combining the current work with the rule-based coding system and applying the other brand new NLP techniques to improve our performance.

Keywords: Deep learning, Deep Neural Network, Natural Language Processing (NLP), ICD-10

1. Introduction

Auto-coding for ICD-10 (The International Statistical Classification of Diseases and Related Health Problems 10th Revision, ICD-10) based on free-text electronic health records (EHR) has drawn great attention in the field of clinical management system. The target of this research is to construct an automatic ICD-10 coding system via Natural Language Processing (NLP) technology.

The ICD-10 is a medical classification list released by World Health Organization (WHO) which defines the universe of diseases, disorders, injuries and other related health conditions and the classifying standard of diagnosis [1]. Since the first publication in 1893, ICD was widely used in fields such as health insurance.

After a statistical processing, the disease classification data can be applied to the clinical management system or be an evaluation factor for the health care quality. Also, since the bureau of national health Insurance, Taiwan started to use ICD code as a reference when evaluating the amount of premium subsidies in the diagnosis-related group prospective payment system, ICD codes have become one of the most important index for the hospital to apply for reimbursement and subsidy [2].

Currently, the reference material for ICD coding are mainly unstructured, i.e. free-text data. Different from structured data, unstructured data would not have fixed format and clear rules such as column length or type. The most common expression type of such data is free text included disease description, history, or diagnosis records of patients which are difficult to define the storage formality strictly. Hence, traditionally, the classification of ICD code was mainly relying on the person who has read a plenty of clinical language documents to handle the complicated procedure of ICD-10 classification. The wide and detail scope of this classification method and the need of referring to the classification rules and medical literature make the classifying task time-consuming and tedious, even the most professional staff in this field takes lots of efforts at finishing such classification. Therefore, an automatic ICD-10 coding system based on Deep Neural Network (DNN) model by applying supervised learning [3] illustrated in Figure 1 is proposed for providing assistance to ICD-10 coders.

+ Corresponding author. Tel.: + (8862) 3366-4888 #419; fax: +(8862) 2362-8167.
E-mail address: flai@ntu.edu.tw.

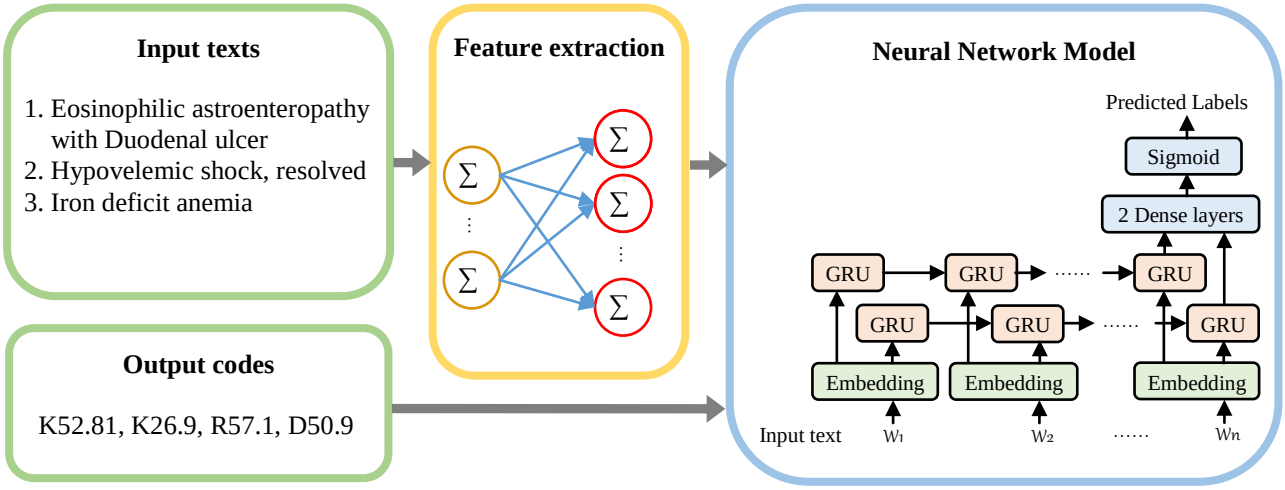

Fig. 1: Training pipeline of ICD-10 classification model.

2. Method

2.1. Data Description

Data including chief complaint, pathology report, physical examination, progress, history, transferring out of ICU diagnosis, and discharge diagnosis between January 2016 to July 2017 was acquired from patients in National Taiwan University Hospital (NTUH). The complete data are about 250,000 records in total and includes up to 14,602 ICD-10 labels which can be divided into 21 classes according to the classes of diseases such as nervous system, immune system, etc. The data will be split to 90% and 10% as training and validation sets.

2.2. Preprocessing

To ensure the uniqueness of the patient account identity and data record, duplicated and null records will be dropped and merged by account identity. Then, procedure including punctuation marks eliminating, case converting, stop words removal, and typos correction are applied for training more effective word embedding model and accelerating the training process. Word tokenization and sequence padding are based on Keras (version 2.2.4) tokenizer.

2.3. Feature Extraction

In this research, Word2Vec, a NLP technique, was applied to transform free-text words to numerical vectors by Continuous Bags of Words (CBOW) model [4]. In CBOW algorithm, Words of sentences were one-hot encoded to binary matrix and then trained with the fully-connected neural network. The previous output eventually pass a softmax classifier, giving the nearby word, i.e. the words within the giving window size, appearing probability for loss computing. Within the process, the hidden layer weight matrix can be extracted to build the word embedding matrix and the corresponding word dictionary.

2.4. Deep Neural Network model

The classification model constructed with four neural network layers, including Recurrent Neural Network (RNN) and Fully-Connected Neural Network (Dense), is shown in Table 1. The first layer is word embedding layer, which transforms the tokenized word list input into word vectors. The second layer is a bidirectional Gated Recurrent Unit (BiGRU) layer [5]. The gating mechanism of GRU solves the vanishing gradient problem that sometimes comes with the standard RNN. In comparison, GRU consumes less time for convergence than the Long Short-Term Memory (LSTM) [6]. The remaining two layers are Dense layers, where the final dense layers should output the vector with the dimension we expect to predict. In our case, there are 21 chapters of ICD-10 and 14602 labels in NTUH data records in total makes the final dense layer size set to 21-dimension and 14602 dimension separately. Each dimension indicates the probability of the code is associated with the diagnosis input.

Table 1: Hyperparameters of whole label classification model.

Hyperparameters	Size
Embedding layer	300
BiGRU layer	256
Dense layer 1	1024
Dense layer 2	14602
Dropout	0.4

2.5. Score Metric

F1-score is the harmonic mean of recall and precision, which are the number of correct positive results divided by the number of all positive results returned by the classifier and the number of correct positive results divided by the number of all relevant samples, hence, appropriate for evaluating the performance of a multi-label classification task. For the realistic application in ICD-10 auto-coding system, recall@20, which calculates the probability of correct answers in first 20 predicting result returned by classifier, is also applied for validating the model performance.

2.6. ICD-10 Predicting Interface

An ICD-10 auto-coding system prototype was built based on python3, ASP.NET Core 2.0 MVC, and Vue.js. The frontend discharge diagnosis input will call for the Web API and sent the case information to the backend. The well-trained DNN model will then be executed with python3 and return the predicting results, providing disease coders the top 20 related ICD-10-CM and ICD-10-PCS codes for auxiliary.

3. Result And Discussion

3.1. ICD-10 Chapter Classification

ICD-10 CM codes are composed of 3 to 7 characters and can be divided into 22 chapters as Table 2 21 classes and U00 to U99. Each of chapter has its title presenting the disease. For example, A00-B99 is about “Certain infectious and parasitic diseases”. Considering that U00-U99 blocks being about “Codes for special purposes” are not related to diseases, our model does not predict the ICD-10 codes between U00 to U99. Hence, in ICD-10 chapter classification task, the ICD-10 codes corresponding each diagnosis record were formatted to 3 characters by removing the last 0 to 4 characters for 21 categories training and predicting. The validation performance is shown in Figure 2.

As Figure 2 shows, discharge diagnosis can achieve the best performance on f1-score of 0.86 on the average of 21 chapters and achieve over 0.5 on f1-score in H60 to H95 and P00-P96 blocks, which only have 1,820 and 2,275 samples in our dataset. Hence, the discharge diagnosis was chosen as the ICD-10 auto – coding reference, i.e. the training data, in whole label classification task and ICD-10 auto-coding system.

3.2. ICD-10 CM Whole Label Classification

According to the ICD-10 chapter classification outcome, we use discharge diagnosis as training data in ICD-10 CM whole label classification task. In NTUH dataset, the complete ICD-10-CM codes, i.e. CM codes with 3 to 7 characters, corresponding to discharge diagnosis records are 14,602 labels. Comparing to the previous study on ICD-9 classification with 85,522 training data and f1-score 0.41 [7], with 300 as embedding dimension, our DNN classification model achieve 0.67 on f1-score and 0.86 on recall@20 as shown in Figure 3.

The model performance in each NTUH division is also tested in the whole label classification task. The results in Table 2 shows no significant difference on prediction results between the divisions while data amount over about 100 records, except of Department of Traumatology and Department of Dermatology. The diagnosis in these two departments partially depend on visual inspection, thus, leading to incomplete expression in text information.

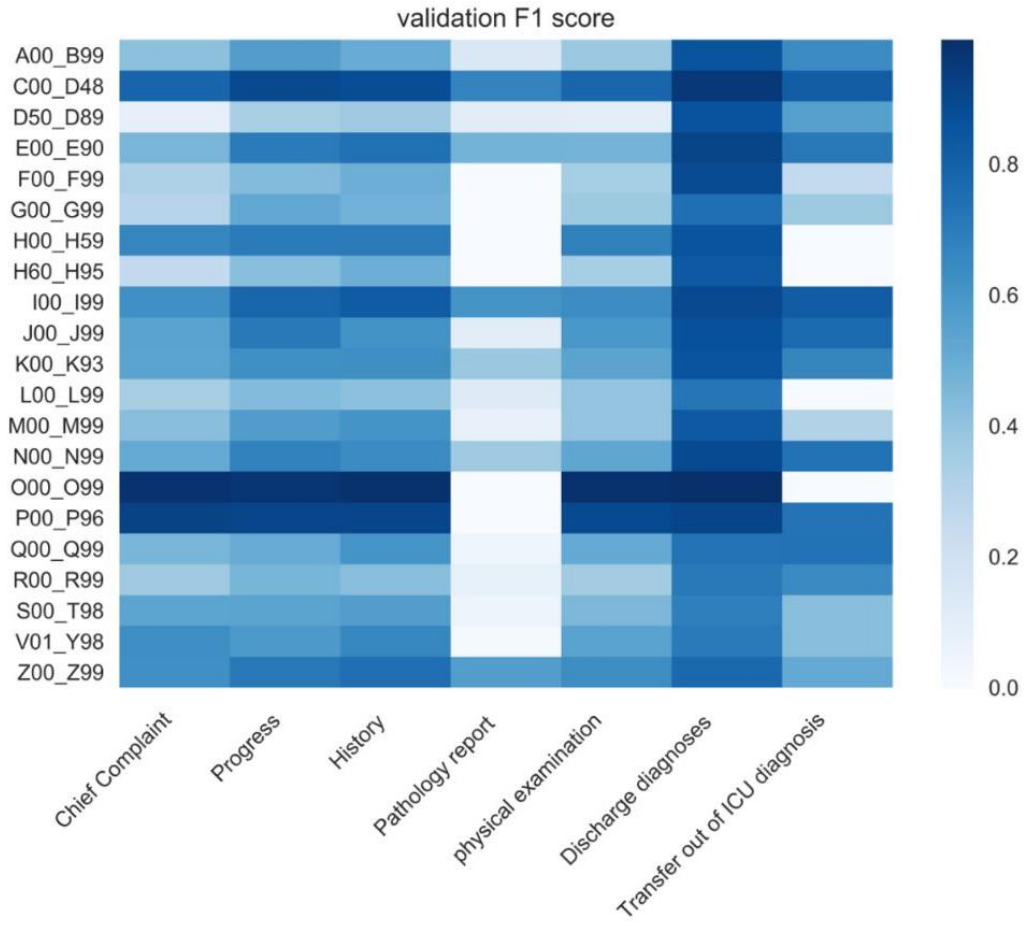

Fig. 2: Performance comparison with different input free-text data using F1 score in validation datasets.

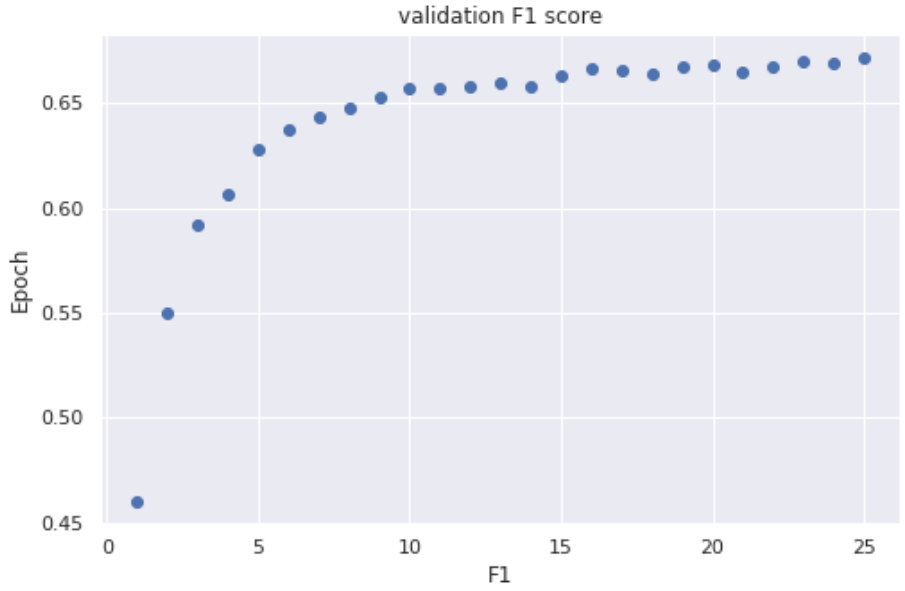

Fig. 3: F1-score performance on CM whole label prediction in validation dataset.

3.3. ICD-10 PCS Whole Label Classification

In ICD-10-PCS whole label classification task, the complete ICD-10-PCS code, i.e. PCS codes with 7 characters, corresponding to discharge diagnosis records are 9,513 labels. Progress, discharge diagnosis and physical examination records are applied for training DNN model. Result in Figure 4 implies that our model can achieve 0.58 on f1-score and 0.81 on recall@20 with progress as input data and word embedding size of 300 dimension.

Table 2: Validation data amount and the prediction for each departments on CM codes.

Department	F1-score	Data amount
Pediatrics	0.671	1,809
Orthopedics Surgery	0.68	1,554
Oncology	0.661	1,647
Obstetrics & Gynecology	0.67	2,136
Dentistry	0.668	258
Internal Medicine	0.669	5,470
Urology	0.67	1,532
Traumatology	0.678	293
Otolaryngology	0.667	929
Surgery	0.673	5,996
Neurology	0.663	243
Ophthalmology	0.674	631
Physical Medicine & Rehabilitation	0.659	104
Emergency Medicine	0.672	76
Family Medicine	0.681	256
Psychiatry	0.693	192
Dermatology	0.649	129
Geriatrics & Gerontology	0.654	6

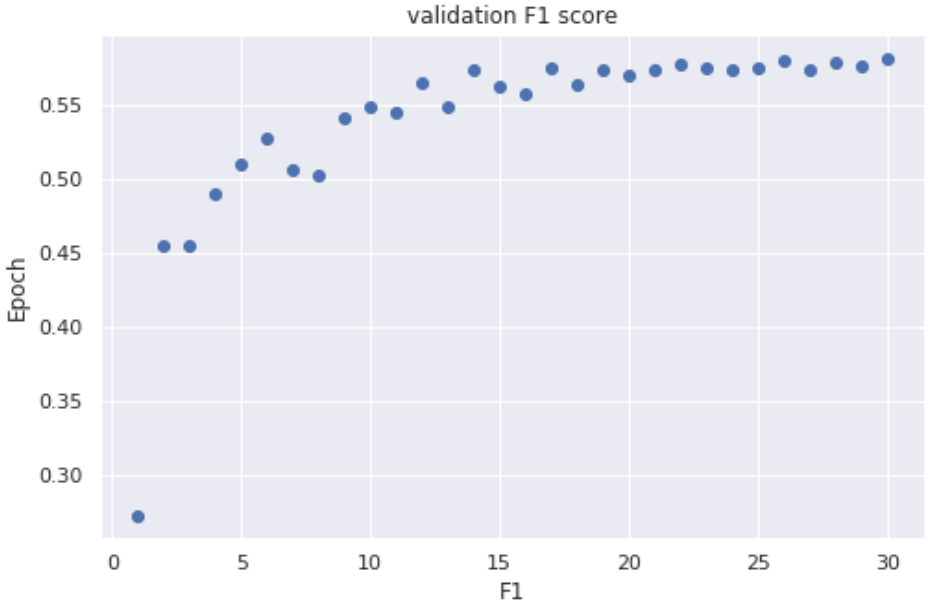

Fig. 4: F1-score performance on PCS whole label prediction in validation dataset.

3.4. ICD-10 Auto-coding System

The target of this research is to build an ICD-10 auto-coding system for assisting disease coders to elevate the work efficiency and coding accuracy. An ICD-10 auto predicting interface by taking discharge diagnosis as reference is published on http://nets.csie.ntu.edu.tw for accelerating coding efficiency. Architecture of the predicting system is shown in Figure 5. DNN model executed by python script will return the top 20 highest ICD-10-CM and ICD-10-PCS codes with recall@20 of 0.87 and 0.81 separately. The predicting process of each case takes less than 30 seconds which dramatically shorten the coding time of 30 mins per case in average.

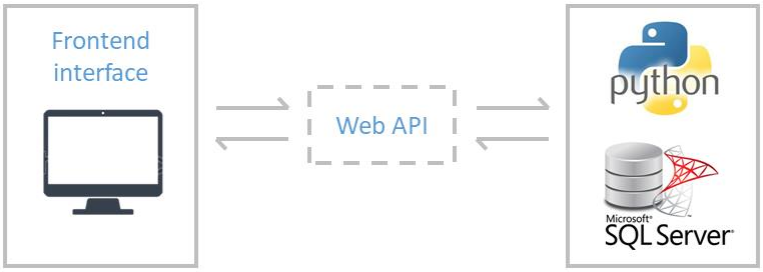

Fig. 5: ICD-10 auto-coding interface pipeline.

4. Conclusion

An ICD-10 classification model developed by NLP and deep learning model without any background knowledge from EHR data is realized in our research with f1-score 0.67 and 0.60 in CM and PCS, respectively. Also, the well-trained model is applied to the web service for assisting disease coders on ICD-10 coding work. In the future, we will keep working on applying BioBERT embedding approach and building a rule-based system to improve the ICD-10-CM classification task.

5. Acknowledgements

This study was supported by grants from the Ministry of Science and Technology, Taiwan (MOST 107-2634-F-002-015). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors have declared that no competing interests exist.

6. References

[1] WHO. 2017. ICD-10 Version: 2015. apps.who.int. (May 2017).

[2] Mills, Ronald E. “Estimating the impact of the transition to ICD-10 on Medicare inpatient hospital payments.” ICD-10 Coordination and Maintenance Committee presentation, March 15, 2015, Baltimore, MD.

[3] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” in Proceedings of the 23rd International Conference on Machine Learning (ICML 2006), 2006, pp. 161–168.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pages 3111–3119, 2013.

[5] KyungHyun Cho Junyoung Chung, Caglar Gulcehre and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555[cs] (Dec. 2014).

[6] Jurgen Schmidhuber Felix A. Gers. 2001. LSTM recurrent networks learn simple context free and context sensitive languages. IEEE Transactions on Neural Networks 12, 6 (2001), 1333–1340.

[7] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Liu PJ, et al. Scalable and accurate deep learning for electronic health records. 2018.