A MULTI-TASK LEARNING FRAMEWORK FOR CHINESE MEDICAL PROCEDURE ENTITY NORMALIZATION

Xuhui Sui, Kehui Song, Baohang Zhou, Ying Zhang, Xiaojie Yuan*

College of Computer Science, TKLNDST, Nankai University, China

ABSTRACT

Medical entity normalization is a fundamental task in medical natural language processing and clinical applications. The task aims to map medical mentions to standard entities in a given knowledge base. In this paper, we focus on Chinese medical procedure entity normalization. This task brings an extra multi-implication challenge that a mention may link to multiple standard entities. To perform the task, we propose a novel deep neural multi-task learning framework to jointly model implication number prediction and entity normalization. Our model utilizes the multi-head attention mechanism to provide mutual benefits between the two tasks. Experimental results show that our method achieves comparable performance compared with the baseline methods.

***Index Terms*—** Named entity normalization, Chinese medical data, Text mining, Joint modeling framework

1. INTRODUCTION

Medical entity normalization (MEN) is the task of assigning mentions of medical terminology to corresponding entities in a knowledge base, such as the International Statistical Classification of Diseases and Related Health Problems 9th Revision (ICD-9). As a core medical information extraction task, MEN plays an important role in the medical language understanding pipeline, underlying a variety of downstream applications such as clinical research [1], diagnosis-related group [2] and medical Q&A system [3].

The major issues of medical entity normalization are the ambiguity [4] brought by noise in text such as misspelling and non-standard expression, and variation [5] which means that the same entity may need to be linked by different mentions. Many previous works [6, 7, 8, 9, 10] have focused on addressing these two issues and achieved outstanding performance. However, different from general medical entity normalization tasks [11, 12, 13, 14, 15], Chinese medical procedure entity normalization faces another issue of multi-implication. Multi-implication refers to the situation that a mention may link to multiple corresponding entities. Examples of Chinese medical procedure entity normalization are shown in Fig. 1. Mention 1 implicates two procedure concepts and needs to link to two different corresponding entities. In addition, the text-level features are not detailed enough to identify the exact number of corresponding entities for a given mention. For mention 2, which has a "+" delimiter, only needs to link to one entity. Thus, the uncertainty in the number of linking entities poses a significant challenge to Chinese medical procedure entity normalization.

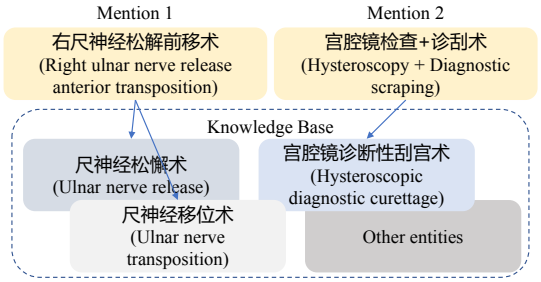

Fig. 1. Examples of multi-implication Chinese medical procedure entity normalization.

Some previous studies have focused on the task of Chinese medical procedure entity normalization. Yan et al. [16] is the first to perform this task and introduces a sequence generation model to generate all possible corresponding entities for the given mention. However, this method ignores the efficiency and may cause the problem of out of dictionary (OOD). Liang et al. [17] designs a framework to merge the results of recall and rank for this task. It solves the multi-implication issue on the candidate generation stage by using a simplistic way to jointly model implication number prediction and candidate generation, which can not ensure the essential mutual supports between these two tasks.

In this paper, we propose a novel deep neural multi-task learning model, which utilizes mutual benefits between implication number prediction and entity normalization to improve the performance of Chinese medical procedure entity normalization. Following prior works [16, 17], our model is designed as a two-stage framework: candidate generation and candidate ranking. The candidate generation stage only aims to decrease the scale of the candidates, while the candidate ranking stage aims to retrieve the corresponding entities from the candidates of the given mention. Implication number

*Corresponding author.

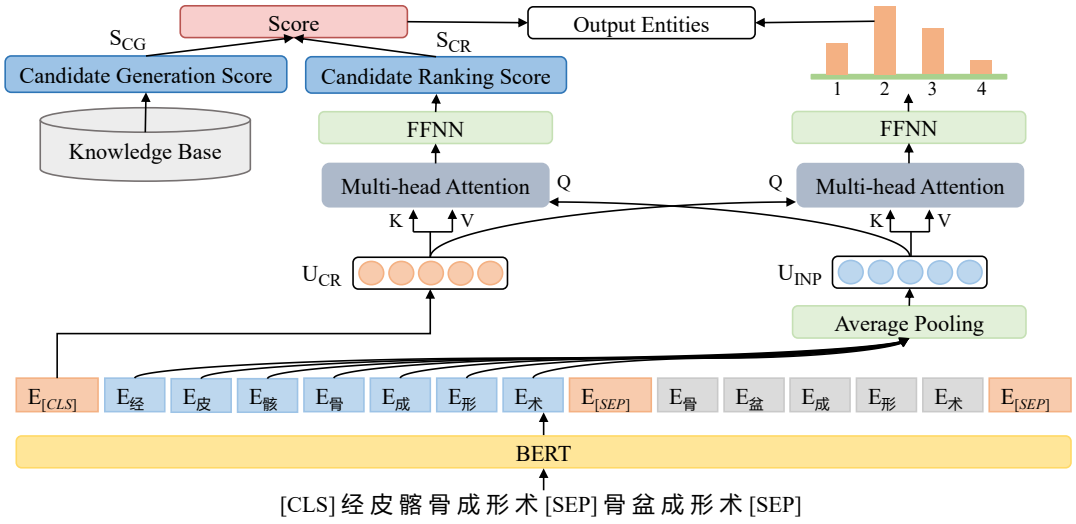

Fig. 2. The overall architecture of our multi-task learning model for Chinese medical procedure entity normalization.

prediction and candidate ranking are hierarchical tasks and their outputs potentially have mutual benefits for each other as well. Specifically, the output of implication number prediction, such as "2", is a clear signal for entity ranking stage to retrieve 2 corresponding entities, and reversely it could be inferred that the implication number is 2 if the scores of top-2 entities are extremely higher than others on candidate ranking stage. Therefore, it is reasonable to apply the implication number prediction task to candidate ranking stage due to their closer relationships, rather than the candidate generation stage. Unlike Liang et al. [17], we design a novel multi-task model for the implication number prediction and candidate ranking, which utilizes the multi-head attention mechanism to provide mutual supports between the two tasks. Experimental results demonstrate that our method performs significantly better than state-of-the-art approaches.

2. METHODOLOGY

We propose a novel multi-task learning model to perform Chinese medical procedure entity normalization. Fig. 2 shows the overall architecture of our model. Our model is designed as a two-step pipeline, which consists of the candidate generation and the multi-task candidate ranking.

2.1. Candidate Generation

We use deep metric learning to recall candidates, which encodes a given mention and all entities of the knowledge base into a dense space and compares the semantic similarity to choose the candidates of the given mention. We adopt triplet network architecture [18] for it and the used base model is BERT [19]. Suppose we have t entities in the knowledge base $E = e_i{}_{i=1}^{t}$ where e_i is a single entity. For a given mention m, both m and each entity e_i are encoded into vectors: $h_m = \text{BERT}(m)$, $h_{e_i} = \text{BERT}(e_i)$, which are the output of the last hidden layer corresponding to the position of the [CLS] token. We set the corresponding entities of given mention m as the positive samples while other entities are all possible negative samples. The objective of triplet network is to minimize the distance between the mention m and its corresponding entities while maximizing the distance between m and other entities. The triplet loss function is defined as follows:

$$L_{triplet} = \max(d_p - d_n + margin,\ 0) \tag{1}$$

where $d_p = ||h_m - h_{positive}||$, $d_n = ||h_m - h_{negative}||$ and $margin$ is a hyper-parameter.

To choose the negative samples from a large amount of entities, we use hard negative sampling. This strategy chooses top-k entities from possible negative samples, which are defined as the ones that are close to corresponding entities of m in terms of semantic distance. The candidates of m are selected by computing the semantic distance between h_m and each entity h_{e_i}.

2.2. Candidate Ranking

This stage aims to retrieve the corresponding entities of the given mention from a set of candidates. The correlation between implication number prediction and candidate ranking inspires us to utilize their potential mutual benefits.

Suppose that there are C candidates for the mention m and take the c-th candidate entity e_c as an example, our model concatenates m and e_c as a sequence pair together with special start and separator tokens ([CLS] m [SEP] e_c [SEP]) to form the input. By BERT, we obtain the vector representation U_{CR}, which is the output of the last hidden layer corresponding to the position of the [CLS] token. Each word embedding

of m could also be obtained by BERT, and an average pooling layer is used to produce the vector representation U_{INP}.

Next, we utilize the multi-head attention proposed in the transformer [20] to build potential mutual benefits for these two tasks. This mechanism can make the features fully interact with each other and is implemented as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}}V) \quad (2)$$

$$\bar{H}_i = \text{Attention}(U_{INP}\bar{W}_i^Q, U_{CR}\bar{W}_i^K, U_{CR}\bar{W}_i^V) \quad (3)$$

$$\hat{U}_{CR} = \text{Concat}(\bar{H}_1, ..., \bar{H}_n)\bar{W}_O \quad (4)$$

$$\tilde{H}_i = \text{Attention}(U_{CR}\tilde{W}_i^Q, U_{INP}\tilde{W}_i^K, U_{INP}\tilde{W}_i^V) \quad (5)$$

$$\hat{U}_{INP} = \text{Concat}(\tilde{H}_1, ..., \tilde{H}_n)\tilde{W}_O \quad (6)$$

In Equations 2-6, $\bar{W}_i^Q, \bar{W}_i^K, \bar{W}_i^V, \tilde{W}_i^Q, \tilde{W}_i^K, \tilde{W}_i^V \in \mathbb{R}^{d_{md} \times d_k}$, $\bar{W}_O, \tilde{W}_O \in \mathbb{R}^{d_{md} \times d_{md}}$ are parameters of the model where d_{md} is the size of hidden states of BERT, $d_k = d_{md}/d_{head}$ where d_{head} denotes the number of heads in the attention. As a result, we get the representation $\hat{U}_{CR}$ and $\hat{U}_{INP}$.

Finally, we put the encoding $\hat{U}_{CR}$ to a feed-forward neural network FFNN_1 and get the score of c-th candidate entity using a softmax function:

$$\hat{s}_{CR}^c = \text{FFNN}_1(\hat{U}_{CR}^c),\ s_{CR}^c = \frac{\exp(\hat{s}_{CR}^c)}{\sum_j^C \exp(\hat{s}_{CR}^j)} \quad (7)$$

We employ cross-entropy as the loss function of candidate ranking, which is shown as follows:

$$L_{CR} = y_{CR}^c \log s_{CR}^c + (1 - y_{CR}^c) \log(1 - s_{CR}^c) \quad (8)$$

where $y_{CR}^c \in \{0, 1\}$ and y_{CR}^c equaling to 1 means the c-th candidate entity is the gold entity otherwise it equals to 0.

Similar to the representation $\hat{U}_{CR}$, we put the representation $\hat{U}_{INP}$ to another feed-forward neural network FFNN_2 and softmax layer to produce the predicted implication number $\hat{y}_{INP}$. Cross-entropy loss is also applied for this task, which is shown as follows:

$$L_{INP} = y_{INP} \log \hat{y}_{INP} \quad (9)$$

where y_{INP} is the implication number label of the mention m. The final loss of the candidate ranking stage is calculated as follows:

$$L_{final} = L_{CR} + L_{INP} \quad (10)$$

2.3. Fusion Block

Followed Liang et al. [17], we also use a fusion block in inference time to merge the scores of candidate generation stage and candidate ranking stage. The candidate generation score is based on the semantic distance between the given mention m and its candidate entities, which is calculated as follows:

$$s_{CG}^c = 1 - \frac{d(m, e_c)}{\sum_j^C d(m, e_j)} \quad (11)$$

Dataset	Uni-implication	Multi-implication	Total
Train	3801	199	4000
Test	2851	149	3000

Table 1. Overall statistics of the dataset.

where $d(m, e_c)$ is the euclidean distance of the mention representation h_m and the c-th candidate entity representation h_{e_c}. We average the candidate generation score s_{CG}^c in Equation 11 and the candidate ranking score s_{CR}^c in Equation 7 as the final similarity score of m and the c-th candidate entity. Finally, based on the predicted implication number $\hat{y}_{INP}$, we select the top-$\hat{y}_{INP}$ candidate entities as the predicted corresponding entities of m.

3. EXPERIMENTS

3.1. Dataset and Metrics

The dataset used in this paper comes from the CHIP 2019 clinical entity normalization task. The corresponding entities of mentions are annotated from the ICD9-2017-PUMCH procedure codes(ICD9) knowledge base which contains 9867 standard entities. Table 1 shows the overall statistics of the training set and test set.

We evaluate the performance in terms of accuracy, which is the percentage of entity mentions that are correctly normalized. Considering the multi-implication issue, the normalization result of a given mention can only be correct when both the predicted implication number and predicted corresponding entities are exactly matched with their labels.

3.2. Implementation Details

The BERT we used is the Chinese $\text{BERT}_{\text{BASE}}$ model [19]. We set the maximum sequence length of input mentions and entities to be 40 and 25 respectively. Any string over the length is truncated. We use the Adam [21] to optimize both our candidate generation model and candidate ranking model.

For the candidate generation model, the number of negative samples for each mention is set to 20 and the hyperparameter $margin$ in Equation 1 is 1.0. We set the learning rate to 2e-5 and the batch size to 4. We choose the top-10 nearest entities for each mention as its candidates. For the candidate ranking model, the batch size is 32, and the learning rate is 2e-5 with a linear learning rate decay schedule. Our experimental code is available here [1].

3.3. Overall Performance

We compare our method with three types of baselines. The first type is the statistic methods, and we select two widely-

[1] https://github.com/suixuhui/MTCEN

Method	Uni -implication	Multi -implication	Total
Tf-idf	49.30	—	46.80
Edit-distance	50.80	—	48.30
BERT-based ranking	88.60	—	84.20
Yan et al. [16]	91.10	52.40	89.30
Liang et al. [17]	92.74	46.65	90.46
Ours	**93.13**	**53.02**	**91.14**

Table 2. Experimental results in terms of accuracy. Total represents the weighted average score of uni-implication and multi-implication data. All results are averaged 5 runs using different random seeds. The best scores are bold. The results indicate the improvement of our model over all baselines is statistically significant with $p < 0.05$ with t-test.

Method	Uni -implication	Multi -implication	Total
BERT-based ranking	100	—	96.30
Delimiter "+"	96.60	70.30	95.60
BERT-only	99.26	65.77	97.80
Yan et al. [16]	98.60	76.40	97.70
Liang et al. [17]	99.58	70.71	98.23
Ours	**99.61**	**76.51**	**98.46**

Table 3. The results of implication number prediction in terms of accuracy. Delimiter "+" denotes the method simply using "+" in the mentions as a delimiter to predict the implication number. BERT-only is the method that uses a BERT model only for implication number prediction. Total represents the weighted average score of uni-implication and multi-implication data.

used methods: Tf-idf and edit-distance. The second type is the BERT-based ranking model [8], which is one of the state-of-the-art methods for medical entity normalization. The last type is the methods for Chinese medical procedure entity normalization: Yan et al. [16] and Liang et al. [17]. Note that Liang et al. [17] utilize two unpublished glossaries collated by themselves in their methods. For a fair comparison, we re-implemented this method according to the experimental settings in their paper without these two glossaries.

Table 2 shows the results of baselines and our proposed model. It can be observed that our model achieves the best scores in both uni-implication and multi-implication data, and achieves the state-of-the-art for Chinese medical procedure entity normalization. Statistic methods only consider the text similarity and perform poorly in this task. The BERT-based ranking model losses competitiveness due to the neglect of the multi-implication issue. Yan et al. [16] introduces a sequence generation model to generate all possible corresponding entities for the given mention. This method needs to generate all words for each corresponding entity and the result is correct only when all words are completely matched, which is more complex and harder. Thus, this method performs worse than our model. Furthermore, our proposed method also performs better than Liang et al. [17], which uses a simplistic way to jointly model implication number prediction and candidate generation. This is consistent with our claim that the implication number prediction has a closer correlation to candidate ranking than candidate generation, and our proposed model can utilize the potential mutual benefits of implication number prediction and candidate ranking to improve the performance of Chinese medical procedure entity normalization.

3.4. Prediction of the Implication Number

We conduct further experiments to assess the ability of implication number prediction and the results are presented in Table 3. The BERT-based ranking model ignores the multi-implication issue and always predicts one corresponding entity for the given mention. Thus, it achieves 100% correctness on the uni-implication data and 0% correctness on the multi-implication data. Delimiter "+" only uses the text-feature to predict implication number and performs poorly. We can find that BERT-only performs worse than Liang et al. [17] and our method, which means that the implication number prediction task can get benefits from entity normalization task. It can also be observed that our proposed model performs better than Yan et al. [16] and Liang et al. [17] in both uni-implication and multi-implication data. This indicates that the implication number prediction task in our proposed model utilizes a more advanced and intelligent way to get benefits from the entity normalization task.

4. CONCLUSION

In this paper, we focus on the Chinese medical produce entity normalization task. It is a fundamental task in medical literature mining due to its close relevance to multiple applications in this area. In order to perform the task, we propose a novel neural multi-task learning framework to jointly model implication number prediction and entity normalization. Our proposed model utilizes the mutual benefits of the two tasks in a more advanced and intelligent way. Our method can convert hierarchical tasks, i.e., implication number prediction and Chinese medical produce entity normalization, into a parallel multi-task mode while maintaining mutual supports between tasks. The experimental results show that our model achieves state-of-the-art performance.

5. ACKNOWLEDGEMENTS

This research is supported by the Chinese Scientific and Technical Innovation Project 2030 (2018AAA0102100), NSFC-Xinjiang Joint Fund (No. U1903128), NSFC-General Technology Joint Fund for Basic Research (No. U1936206).

6. REFERENCES

[1] Robert Leaman, Ritu Khare, and Zhiyong Lu, "Challenges in clinical natural language processing for automated disorder normalization," *J. Biomed. Informatics*, vol. 57, pp. 28–37, 2015.

[2] Simon Hölzer, Ralf Schweiger, and Joachim Dudeck, "Application of information technology: Transparent ICD and DRG coding using information technology: Linking and associating information sources with the extensible markup language," *J. Am. Medical Informatics Assoc.*, vol. 10, no. 5, pp. 463–469, 2003.

[3] Kathy Lee, Sadid A. Hasan, Oladimeji Farri, Alok N. Choudhary, and Ankit Agrawal, "Medical concept normalization for online user-generated texts," in *ICHI*, 2017, pp. 462–469.

[4] Yizhou Zhang, Xiaojun Ma, and Guojie Song, "Chinese medical concept normalization by using text and comorbidity network embedding," in *ICDM*, 2018, pp. 777–786.

[5] Yi Luo, Guojie Song, Pengyu Li, and Zhongang Qi, "Multi-task medical concept normalization using multi-view convolutional neural network," in *AAAI*, 2018, pp. 5868–5875.

[6] Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang, "Cnn-based ranking for biomedical entity normalization," *BMC Bioinform.*, vol. 18, no. S-11, pp. 79–86, 2017.

[7] Pan Deng, Haipeng Chen, Mengyao Huang, Xiaowen Ruan, and Liang Xu, "An ensemble CNN method for biomedical entity normalization," in *EMNLP*, 2019, pp. 143–149.

[8] Zongcheng Ji, Qiang Wei, and Hua Xu, "Bert-based ranking for biomedical entity normalization," *AMIA*, vol. 2020, pp. 269, 2020.

[9] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang, "Biomedical entity representations with synonym marginalization," in *ACL*, 2020, pp. 3641–3650.

[10] Baohang Zhou, Xiangrui Cai, Ying Zhang, Wenya Guo, and Xiaojie Yuan, "MTAAL: multi-task adversarial active learning for medical named entity recognition and normalization," in *AAAI*, 2021, pp. 14586–14593.

[11] Sameer Pradhan, Noémie Elhadad, Brett R. South, David Martínez, Lee M. Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana K. Savova, "Task 1: Share/clef ehealth evaluation lab 2013," in *CLEF (Working Notes)*, 2013, vol. 1179.

[12] Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu, "NCBI disease corpus: A resource for disease name recognition and concept normalization," *J. Biomed. Informatics*, vol. 47, pp. 1–10, 2014.

[13] "Uniprot: a hub for protein information," *Nucleic Acids Res.*, vol. 43, no. Database-Issue, pp. 204–212, 2015.

[14] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu, "Biocreative V CDR task corpus: a resource for chemical disease relation extraction," *Database J. Biol. Databases Curation*, vol. 2016, 2016.

[15] Kirk Roberts, Dina Demner-Fushman, and Joseph M. Tonning, "Overview of the TAC 2017 adverse reaction extraction from drug labels track," in *TAC*, 2017.

[16] Jinghui Yan, Yining Wang, Lu Xiang, Yu Zhou, and Chengqing Zong, "A knowledge-driven generative model for multi-implication chinese medical procedure entity normalization," in *EMNLP*, 2020, pp. 1490–1499.

[17] Ming Liang, Kui Xue, and Tong Ruan, "A multi-perspective combined recall and rank framework for chinese procedure terminology normalization," *CoRR*, vol. abs/2101.09101, 2021.

[18] Elad Hoffer and Nir Ailon, "Deep metric learning using triplet network," in *SIMBAD*, 2015, vol. 9370, pp. 84–92.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.