A Joint Model for Medical Named Entity Recognition and Normalization

Ying Xiong[a], Yuanhang Huang[a], Qingcai Chen[a,b], Xiaolong Wang[a], Yuan Ni[c], and Buzhou Tang[a,b*]

[a] *Harbin Institute of Technology, Shenzhen, Xili university town, Shenzhen, China*
[b] *Pengcheng Labtorary, Xili Street, Shenzhen, China*
[c] *PingAn Health Technology Ltd, Shenzhen, China*

Abstract

Traditional pipeline models for medical named entity recognition and normalization (MER and MEN) suffer from error propagation. To tackle the error propagation problem, we propose a novel joint deep learning method for the 2020 IberLEF shared task on MER and MEN, where MER is regarded as a machine reading comprehension (MRC) problem and MEN as multiple sequence labeling problems corresponding to normalized hierarchical tumor codes. In the 2020 IberLEF shared task, our proposed joint model achieves an F1 score of 0.87 on MER and an F1 score of 0.825 on MEN, and significantly outperforms pipeline models for comparison.

Keywords[1]

Medica named entity recognition, medical entity normalization, joint deep learning, multiple sequence labeling

1. Introduction

In the past few years, researchers have taken great interest in clinical natural language processing (NLP) and have launched a number of shared tasks on clinical NLP in Spanish. In 2019, Martin Krallinger et al. [1] organized a shared task for Pharmacological Substances, Compounds and proteins and Named Entity Recognition (called PharmaCoNER). In the same year, Iberian Languages Evaluation Forum (IberLEF) launched a shared task on Medical Document Anonymization (MEDDOCAN) [2] and eHealth Knowledge Discovery [3]. In 2020, IberLEF first organizes CANcer TExt Mining Shared Task (CANTEMIST), a tumor morphology task, including named entity recognition and normalization of a critical type of medical concept related to cancer, named medical named entity recognition (MER) and medical named entity normalization (MEN) [4]. In this shared task, participants are required to recognize a kind of entity "MORFOLOGIA_NEOPLASIA". Pipeline methods are usually applied to MER and MEN, where MEN is a follow-up task of MER [5]. However, it is inevitable that the pipeline methods suffer from error propagation as reported by Xiong et al. [5]. To avoid alleviate error propagation problem, a few joint learning methods have been proposed to solve MER and MEN simultaneously. For example, Leamon et al. [6] proposed an ensembled method composed of two independent machine learning models to deal with chemical named entity recognition and normalization jointly. Lou et al. [7] proposed a transition-based model to recognize disease named entities and map them into normalized concepts. Zhao et al. [8] proposed a multi-task joint learning method to perform MER and MEN.

In this paper, we propose a novel joint deep learning method for MER and MEN and develop a system based this model for CANTEMIST in 2020. The model uses different neural network layers for MER and MEN, but the same word representation shared by MER and MEN. In this model, MER is regarded as a machine reading comprehension (MRC) problem inspired by Li et al. [9,10] and MEN as multiple sequence labeling problems corresponding to normalized hierarchical tumor codes as shown in Figure 1, where the first four digital chars denote tumor/cell type, the fifth digital char denotes behavior, the sixth digital char denotes differentiation, and the last char denotes whether there is a relevant modifier

[1] Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: xiongying@stu.hit.edu.cn; 18s051003@stu.hit.edu.cn; qingcai.chen@gmail.com; wangxl@insun.hit.edu.cn; niyuan442@pingan.com.cn; tangbuzhou@gmail.com (corresponding author);

not included in the terminology of this concept. In the 2020 IberLEF shared task, our proposed joint model achieves an F1 score of 0.87 on MER and an F1 score of 0.825 on MEN, and significantly outperforms pipeline models for comparison.

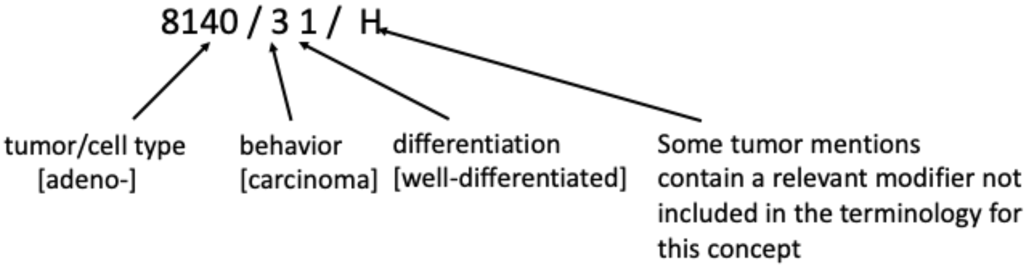
8140 / 3 1 / H
tumor/cell type [adeno-]
behavior [carcinoma]
differentiation [well-differentiated]
Some tumor mentions contain a relevant modifier not included in the terminology for this concept

Figure 1: Example of hierarchical architecture of tumor code

2. Material and method

As shown in Figure 2, we develop a joint deep learning model for the 2020 IberLEF shared task on MER and MEN. For the MER task, we adopt a machine reading comprehension model to detect ME spans. For MEN, we regard it as multiple sequence labeling tasks. Each part is presented in the following section in detail.

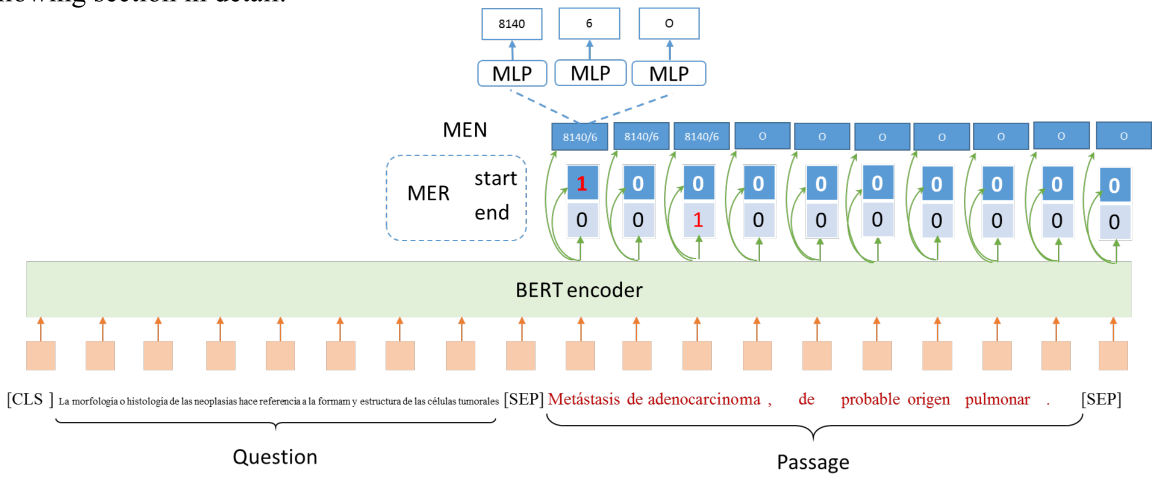
8140
6
O
MLP
MLP
MLP
MEN
8140/6
8140/6
8140/6
O
O
O
O
O
O
O
MER
start
end
1 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
BERT encoder
[CLS] La morfologia o histologia de las neoplasias hace referencia a la formam y estructura de las células tumorales [SEP] Metástasis de adenocarcinoma , de probable origen pulmonar . [SEP]
Question
Passage

Figure 2: Architecture of the proposed joint deep learning model for MER and MEN

2.1. Dataset

The 2020 IberLEF shared task organizers provide a corpus, called CANTEMIST, including a total of 6233 clinical notes in Spanish, 1301 out of which are manually annotated with "MORFOLOGIA_NEOPLASIA" entities mapped to 8410 normalized turor codes. The annoted corpus is further split into a training set of 501 notes, a development set (dev) of 250 notes, a supplementary development set (sdev) of 250 notes and a test set of 300 notes mixed with other 4932 notes as background (bg). A tumor code consists of six digital chars plus one relevant modifier denoted by "ABCD/EF/H" and can be divided into four parts of different meanings as follows: "ABCD"-tumor/cell type, 'E'-behavior, 'F'-differentiation and 'H'-relevant modifier not included in the terminology of this concept. Table 1 lists the statistics of the corpus in detail.

Table 1 Statistic of the CANTEMIST corpus

statistics	#training	#dev	#sdev	#test	#bg

document	501	250	250	300	4932

2.2. Medical named entity recognition

Different from most existing models that regard MER as a sequence labeling problem that needs to tag each token with entity boundary and type, in this paper, we regard MER as an MRC problem, whose task is to answer questions regarding different types of entities based on given passages. Following previous studies [9,10], we directly use the definition of each type of entity as the question regarding it. That is, the definition of MORFOLOGIA_NEOPLASIA, “La morfología o histología de las neoplasias hace referencia a la formam y estructura de las células tumorales”, is the question regarding MORFOLOGIA_NEOPLASIA, denoted by q. A sentence in any clinical record is regarded as a passage, denoted by p. The task of MRC is to determine the start and end position pairs of MORFOLOGIA_NEOPLASIA entities, given q and p. We define a start and end position pair as (s, e). In our model, BERT [11] is first used as backbone to represent the interactions between q and p, and outputs the representation of passage p, denoted by $H \in \mathbb{R}^{n \times d}$, where n is the length of p, and d is the representation dimension of each word. Then, multi-layer perception (MLP) is used to compute the possibilities of start position s and end position e as follows:

$$P_{s_i} = MLP(W_s H_i + b_s), \tag{1}$$

$$P_{e_i} = MLP(W_e H_i + b_e), \tag{2}$$

where $s_i = argmax(P_{si}) \in \{0,1\}$ represents whether the i-th position is the start position of an entity and $e_i = argmax(P_{e_i}) \in \{0,1\}$ represents whether the i-th position is the end position of an entity, W_s and W_e are parameter matrices, b_s and b_e are bias vectors.

During the training phase, we adopt the cross entropy loss to optimize the parameters of our MER model, which is defined as follows:

$$L_{start} = \sum_s CE(P_s, Y_s), \tag{3}$$

$$L_{end} = \sum_e CE(P_e, Y_e), \tag{4}$$

$$L_{mer} = L_{start} + L_{end}, \tag{5}$$

where CE is the cross entropy loss, P_s and P_e are the possibilities of predicted start position s and end position e, Y_s and Y_e are the possibilities of gold standard start position s and end position e.

2.3. Medical named entity normalization

The task of MEN is to map a medical named entity to a normalized code in a given vocabulary. In this paper, we convert MEN into a multiple sequence labeling problem, where each token is labeled with three normalized subcodes as shown in Figure 2, where the behavior code and differentiation code are combined together as we believe they are strongly related to each other. The subcodes of the i-th token in passage can be predicted by equations (6), (7) and (8) defined as follows:

$$c_i^{abcd} = MLP(W_{abcd} H_i + b_{abcd}), \tag{6}$$

$$c_i^{ef} = MLP\big(W_{ef} H_i + b_{ef}\big), \tag{7}$$

$$e_i^{hg} = MLP\big(W_{hg} H_i + b_{hg}\big), \tag{8}$$

$$c_i = c_i^{aabcd} / c_i^{ef} / c_i^{hg}, \tag{9}$$

where c_i^{abcd}, c_i^{ef} and c_i^{hg} are the three subcodes, W_{abcd}, W_{ef}, W_{hg} are parameter matrice and b_{abcd}, b_{ef}, and b_{hg} are bias vectors..

Similar MER, we adopt the cross entropy loss for model parameter optimization. The loss of MEN is defined as follows:

$$L_{men} = \sum_c CE\big(P_{c^{abcd}}, Y_{c^{abcd}}\big) + CE\big(P_{c^{ef}}, Y_{c^{ef}}\big) + CE\big(P_{c^{hg}}, Y_{c^{hg}}\big), \tag{10}$$

where P_{c^*} is the possibility of each predicted subcode c^* and Y_{c^*} is the possibility of each gold standard subcode c^*.

The total loss of our joint model is the weighted sum of the MER loss and MEN loss:

$$L_{total} = L_{mer} + \lambda L_{men}, \tag{11}$$

where λ is the loss weight.

2.4. Evaluation

The performances of all models on both MER and MEN task are evaluated by concept-level precision (P), recall (R) and F1-score (F1) under the exact-match criterion.

2.5. Experiments setup

We first investigate the effect of combination parameter λ of different values (0.5 vs 1.0) on our joint model and then compare the joint model with a pipeline method that uses the same method for MER and a generation model SGM [12] for MEN. The BERT model is initialized by "BERT-Base, Multilingual Cased" (https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip), and further pretrained on the CANTEMIST corpus. All other parameters are optimized on the supplementary development set.

3. Results

The performance of our joint model is listed in Table 2, where the highest P, R, and F1 scores of the model on MER and MEN are highligted in bold. The highest F1 score of MER is 0.87 when the loss weight λ is set as 0.5, and the highest F1 score of MEN is 0.825 when loss weight λ is 1.0. In total, our model achieves better performance when $\lambda = 1.0$.

Table 2
The overall performance of our model on MER and MEN

λ	Task	P	R	F1
0.5	MER	**0.871**	0.868	**0.87**
	MEN	0.82	0.808	0.814
1.0	MER	0.866	**0.87**	0.868
	MEN	**0.824**	**0.826**	**0.825**

3.1. Joint learning vs pipeline

As shown in Table 3, the joint learning method yields a higher F1 score than the pipeline method. This demonstrates the joint model benefits from the shared representation of word representation. For the MER task, the joint learning method outperforms the pipeline method by achieving a 0.868 F1 score. For the MEN task, the joint learning method can bring 4.3% F1 score improvements over the pipeline method. Due to the high imbalance of codes, there are ubiquitous code 8000/6. When removing 8000/6 mentions (denoted as No-Metastasis), the joint learning method has a slight change on P, R and F1 score, which indicates the robustness of the joint model.

Table 3
The performances of the pipeline method and joint learning method on MER and MEN

Task	method	P	R	F1	P- No-Metastasis	R- No-Metastasis	F- No-Metastasis
MER	pipeline	0.862	0.857	0.86	\	\	\
	joint	**0.866**	**0.87**	**0.868**	\	\	\
MEN	pipeline	0.794	0.791	0.792	0.799	0.765	0.782

	joint	**0.824**	**0.826**	**0.825**	**0.848**	**0.803**	**0.825**

4. Discussion

Though our joint learning method shows a great improvement over the pipeline method, there are still some errors on the MER task. 1) Long tail problem is the main obstacle. The number of the entity mentions containing over 10 words is about 60, but our model can only recognize 10 of them. 2) Nested entity mentions are difficult to recognize. Though our model tries to recognize the nested entities, it is difficult for our model to recognize them because of its rareness in the training set. 3) If a text has "A y (and) B", our model finds it difficult to judge whether A and B are both entities.

The joint learning method shows a great improvement in the MEN task, but there are some limits. The number of sequence labeling submodels has a huge impact on the results. When we further separate the behavior part and the differentiation part of tumor code, the MEN F1 score on the development set decreases from 0.794 to 0.708, indicating that the behavior and differentiation have a strong relationship. In the future, we plan to explore how to detect the relationships among different parts of tumor code automatically in the model.

5. Conclusion

In this study, we propose a joint learning method for medical named entity recognition and medical named entity normalization. We utilize a machine reading comprehension model to solve thee MER task and a multiple sequence labeling model to solve the MEN task. Experimental results show the effectiveness of our model.

6. Acknowledgements

This paper is supported in part by grants: National Natural Science Foundations of China (U1813215, 61876052 and 61573118), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), National Natural Science Foundations of Guangdong, China (2019A1515011158), Guangdong Province Covid-19 Pandemic Control Research Fund (2020KZDZX1222), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20180306172232154 and JCYJ20170307150528934) and Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052).

7. References

[1] Gonzalez-Agirre A, Marimon M, Intxaurrondo A, Rabal O, Villegas M, Krallinger M. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. Association for Computational Linguistics; 2019:1–10. doi:10.18653/v1/D19-5701

[2] Marimon M, Gonzalez-Agirre A, Intxaurrondo A, et al. Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In: *Proceedings of the Iberian Languages Evaluation Forum Co-Located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019.* ; 2019:618–638. http://ceur-ws.org/Vol-2421/MEDDOCAN_overview.pdf

[3] Piad-Morffis A, Gutiérrez Y, Consuegra-Ayala JP, et al. Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019. In: *Proceedings of the Iberian Languages Evaluation*

Forum Co-Located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019. ; 2019:1–16. http://ceur-ws.org/Vol-2421/eHealth-KD_overview.pdf

[4] Miranda-Escalada A, Farré E, Krallinger M. Named entity recognition, concept normalization and clinical coding: Overview of the CANTEMIST track for cancer text mining in Spanish, Corpus, Guidelines, Methods and Results. In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*. CEUR Workshop Proceedings. ; 2020.

[5] Xiong Y, Shen Y, Huang Y, et al. A Deep Learning-Based System for PharmaCoNER. In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. Association for Computational Linguistics; 2019:33–37. doi:10.18653/v1/D19-5706

[6] Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinform*. 2016;32(18):2839–2846. doi:10.1093/bioinformatics/btw343

[7] Lou Y, Zhang Y, Qian T, Li F, Xiong S, Ji D. A transition-based joint model for disease named entity recognition and normalization. *Bioinform*. 2017;33(15):2363–2371. doi:10.1093/bioinformatics/btx172

[8] Zhao S, Liu T, Zhao S, Wang F. A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.* ; 2019:817–824. doi:10.1609/aaai.v33i01.3301817

[9] Li X, Feng J, Meng Y, Han Q, Wu F, Li J. A Unified MRC Framework for Named Entity Recognition. *arXiv preprint arXiv:191011476*. Published online 2019.

[10] Levy O, Seo M, Choi E, Zettlemoyer L. Zero-Shot Relation Extraction via Reading Comprehension. In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017.* ; 2017:333–342. doi:10.18653/v1/K17-1034

[11] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* ; 2019:4171–4186.

[12] Yang P, Sun X, Li W, Ma S, Wu W, Wang H. SGM: Sequence Generation Model for Multi-label Classification. In: *Proceedings of the 27th International Conference on Computational Linguistics*. ; 2018:3915–3926.