A Deep Learning-Based System for PharmaCoNER

Ying Xiong[1], Yedan Shen[1], Yuanhang Huang[1], Shuai Chen[1], Buzhou Tang[1,2*] Xiaolong Wang[1], Qingcai Chen[1,2], Jun Yan[3], Yi Zhou[4*]
[1]Department of Computer Science, Harbin Institute of Technology, Shenzhen, China, 518055
[2]Peng Cheng Laboratory
[3]Yidu Cloud (Beijing) Technology Co., Ltd, Beijing
[4]Sun YAT-SEN UNIVERSITY
{xiongying0929, hyhang7, chenshuai726, tangbuzhou, qingcai.chen}@gmail.com
shenyedan@stu.hit.edu.cn, wangxl@insun.hit.edu.cn, Jun.YAN@Yiducloud.cn, zhouyi@sysu.edu.cn
*** Corresponding author**

Abstract

The Biological Text Mining Unit at BSC and CNIO organized the first shared task on chemical & drug mention recognition from Spanish medical texts called PharmaCoNER (Pharmacological Substances, Compounds and proteins and Named Entity Recognition track) in 2019. The shared task includes two tracks: one for NER offset and entity classification (track 1) and the other one for concept indexing (track 2). We developed a pipeline system based on deep learning methods for this shared task, specifically, a subsystem based on BERT (Bidirectional Encoder Representations from Transformers) for NER offset and entity classification and a subsystem based on Bpool (Bi-LSTM with max/mean pooling) for concept indexing. Evaluation conducted on the shared task data showed that our system achieves a micro-average F1-score of 0.9105 on track 1 and a micro-average F1-score of 0.8391 on track 2.

1 Introduction

Efficient access to mentions of clinical entities is very important for using clinical text. The way to extract clinical entities embedded in the text is natural language processing (NLP). In the last decades, clinical entity extraction has attracted plenty of attention of researchers, clinicians, and enterprises in the clinical domain. The development of technology for clinical entity extraction mainly benefits from related NLP challenges including tasks of biomedical entity recognition and normalization, such as the BioCreative (Critical Assessment of Information Extraction systems in Biology) challenges (e.g., the CHEMDNER (Chemical compound and drug name recognition) track (Leaman et al., 2013)), the i2b2 (the Center of Informatics for Integrating Biology and Bedside) challenges (Uzuner et al., 2011), SemEval (Semantic Evaluation) challenges (Elhadad et al., 2015) and the ShARe/CLEF eHealth Evaluation Lab shared tasks (Kelly et al., 2016). A large number of various kinds of methods have been proposed for biomedical entity recognition and normalization. Lots of machine learning methods such as conditional random fields (CRF) (Lafferty et al., 2001), structured support vector machines (SSVM) (Tsochantaridis et al., 2005) and bidirectional long-short-term memory with conditional random fields (BiLSTM-CRF) (Huang et al., 2015) have been applied for biomedical entity recognition, support vector machines (SVM) (Grouin et al., 2010) and ranking based on convolutional neural network (CNN) (Li et al., 2017) for clinical entity normalization. Although there have been a few promising results, most of them focus on the clinical text in English. Recently, clinical entity extraction for clinical text in other languages has also begun to receive much attention. For example, in 2016, NTCIR organized the first challenge about information extraction from clinical documents in Japanese (Morita et al., 2013). In 2017, CCKS organized the first challenge about information extraction from clinical records in Chinese (Hu et al., 2017).

To accelerate development of techniques of information extraction from clinical text in Spanish, Martin Krallinger et al. organized a shared task particular for chemical & drug mention recognition from Spanish medical texts called PharmaCoNER in 2019 (Gonzalez-Agirre, Aitor et al., 2019), which includes two tracks: track 1 for NER offset and entity classification and track 2 for concept indexing. The organizers provided an annotated corpus of 1000 clinical cases, 500 cases out of which were used as the

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 33–37
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

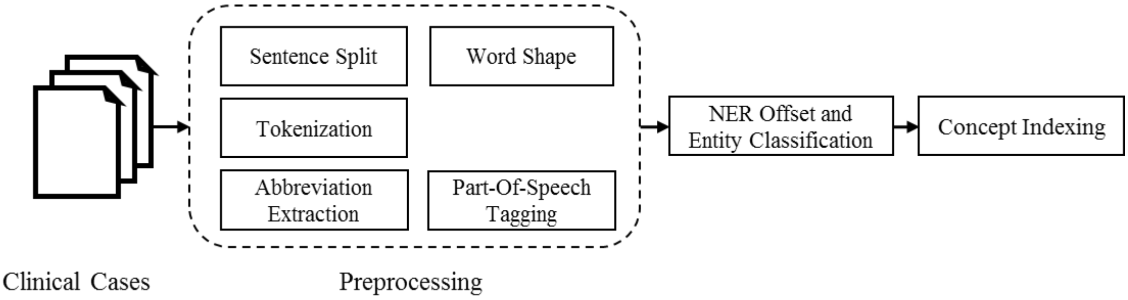

Figure 1: Overview architecture of our system for the PharmaCoNER task

training set, 250 cases as the development set and 250 cases as the test set. We participated in this shared task and developed a pipeline system based on two latest deep learning methods: BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) and Bpool (Bi-LSTM with max/mean pooling) (Conneau et al., 2017). The system developed on the training and development sets achieved a micro-average F1-score of 0.9105 on track 1 and a micro-average F1-score of 0.8391 on track 2 on the independent test set.

2 Material and Methods

As shown in Figure 1, We first developed a preprocessing module to split clinical cases into sentences, tokenized the sentences and extracted some features for each token, then a BERT-based subsystem for NER offset and entity classification, and finally a Bpool-based system for concept indexing. All of them were individually presented in the following sections in detail.

2.1 Dataset

The PharmaCoNER organizers asked medical experts to annotate a corpus of 1000 clinical cases with chemical & drug mentions for the shared task according to a pre-defined guideline. The corpus was divided into a training set, a development set and a test set. The test set was hidden in a background set of 3751 clinical cases when testing during the competition. The statistics of the corpus, including the number of documents, chemical & drug mentions in different types are listed in Table 1, where "UNK" denotes unknown. It should be noted that the chemical & drug mentions annotated with UNCLEAR were not considered during the competition.

2.2 Preprocessing

We split each clinical case into sentences using ';', '?', '!', '\n' or '.' which is not in numbers, and further split each sentence into tokens using the method proposed by Liu (Liu et al., 2015), which was specially designed for clinical text. We adopted Ab3P tools[1] to extract full names of abbreviations, and SPACCC_POS-TAGGER tool[2] for POS tagging and lemmatization. Besides, we used the same way as Liu (Liu et al., 2015) to get each word's word shape.

2.3 NER offset and entity classification

NER offset and entity is a typical NER problem usually recognized as a sequence labeling problem. In this study, we adopted "BIO" tagging schema to represent chemical & drug mentions, where 'B', 'I' and 'O' represent beginning, inside and outside of a chemical & drug mentions respectively, and developed a system based on BERT. First, character-level representation, POS tagging representation and word shape representation of each word were concatenated into the word representation of BERT, and then a CRF layer was appended to BERT for chemical & drug mentions recognition.

2.4 Concept Indexing

After chemical & drug mentions were recognized, we first constructed <mention, standard terminology> pairs as candidates for matching, and then built a Bpool-based matching model (Conneau et al., 2017) according to the candidates. Standard terminologies were selected into candidates in the following two ways:

[1] (https://github.com/ncbi-nlp/Ab3P)

[2] (https://github.com/PlanTL-SANIDAD/SPACCC_POS-TAGGER)

Statistic	#Training	#Development	#Test	#Background
DOCUMENT	500	250	250	3751
NORMALIZABLES	2304	1121	973	UNK
NO_NORMALIZABLES	24	16	10	UNK
PROTEINAS	1405	745	859	UNK
UNCLEAR	89	44	0	UNK

Table 1. Statistics of the PharmaCoNER Corpus.

1) Top *n* terminologies ranked by Levenshtein distance[3] with a given mention at char-level and at token-level.
2) Terminologies selected by 1) and the given mention's synonyms appearing in the standard terminology vocabulary.

After the terminology selection, a Bpool-based matching model at character-level was utilized to judge whether two mentions were matching or not.

2.5 Evaluation

The performance of our system was measured by micro-average precision (P), recall (R), and F1-score (F1), which were calculated by the official tool provided by the PharmaCoNER organizers[4].

2.6 Experiments Setup

In this study, for track1, we first optimized model on the development set and then fine-tuned the model on the training and development sets for 5 more epochs. For standard terminology selection, we optimized *n* from 10 to 50 with step 10, and finally set it to 40. For track2, we optimized the model on the training and development sets via 10-fold cross validation. The hyper-parameters and parameter estimation algorithm used for model training were listed in Table 2. The pre-trained BERT[5] was used as the initial neural language model and fine-tuned on all datasets provided by the shared task organizers. The embeddings of character, POS and word shape were randomly initialized from a uniform distribution. It is worth noting that in the BERT model, the update of the parameters included in the BERT used the learning rate of 2e-5, and the parameter update of other features used a learning rate of 0.003.

Hyper-parameter	Value
Dimension of character representation	BERT:30; Bpool:50
Dimension of POS representation	30
Dropout probability	0.1
Learning rate	BERT: 2e-5; Bpool: 1e-3
Training epochs	Bert:15; Bpool:20
Parameter estimation algorithm	BERT: adam with warmup; Bpool: adam

Table 2. Hyper-parameters and parameter estimation algorithm used for deep learning methods.

3 Results

The highest micro-average precisions, recalls and F1-scores of our system on the two tracks were listed in Table 3. Our system achieved a micro-average precision of 0.9123, recall of 0.9088 and F1-score of 0.9105 on track1, and a micro-average precision of 0.8284, recall of 0.8502 and F1-score of 0.8391 on track2. Among three types of chemical & drug mentions considered in the shared task, our system performed best on NORMALIZABLES and worst on NO_NORMALIZABLES for track1, which may be proportional to the number of mentions of each type.

Track	Type	P	R	F1
Track1	NORMALIZABLES	0.9426	0.9291	0.9358
	NO_NORMALIZABLES	1.0000	0.2000	0.3333
	PROTEINAS	0.8787	0.8941	0.8863
	Overall	0.9123	0.9088	0.9105
Track2	Overall	0.8284	0.8502	0.8391

Table 3. The highest results of our system for PharmaCoNER. (P: micro-average precision; R: micro-average recall; F1: micro-average F1 score)

[3] https://en.wikipedia.org/wiki/Levenshtein_distance

[4] https://github.com/PlanTL-SANIDAD/PharmaCoNER-CODALAB-Evaluation-Script

[5] https://storage.googleapis.com/bert_models/2018_11_03/multilingual_L-12_H-768_A-12.zip

3.1 Ablation Study

Table 4 provided additional ablation study results analyzing the contribution of individual features on track 1 and reporting the performance of each standard terminology selection method (STS) on track 2. We found that both character-level embedding, POS tagging representation, and word shape representation contributed towards our system on track 1. They brought 1.69%, 0.51%, and 0.63% improvements on F1-score, respectively. On track 2, when removing the extended synonyms, the F1 score declined from 0.8048 to 0.7932.

Track	model	P	R	F1
Track1	BERT	0.8989	0.9087	0.9037
	- Character-embedding	0.8981	0.8757	0.8868
	- POS tagging	0.8986	0.8986	0.8986
	- Word shape	0.8874	0.9076	0.8974
Track2	STS 1	0.7722	0.8153	0.7932
	STS 2	0.7826	0.8284	0.8048

Table 4. Ablation study of track 1 and track 2 on the development set. (P: micro-average precision; R: micro-average recall; F1: micro-average F1 score)

4 Discussion

For task 1, our analysis found that data processing had a great influence on the NER offset results. Separating alphabets and digitals in a word , for example, “PaO2” was split into ‘PaO’ and ‘2’ , caused some errors of entity boundary or entity type. Separating words by the hyphen ‘-‘ also caused some errors. For example, “4-methyilumbelliferyl α-D-galactosidasa” is totally identified as ‘PROTEINAS’, but in “daclizumab-tacrolimus-MMF-esteroide”, “daclizumab” is identified as “PROTEINAS”, “tacrolimus”, “MMF” and “esteroide” are identified as “NORMILIZED”. Our experiments on the development set showed that the effect of tokenization on micro-average F1 score on NER was about 2%.

There were mainly the following three types of errors caused by our system. (1) abbreviation recognition errors: it is difficult to identify abbreviations in a record correctly; (2) long entity: entities consisting of four or more tokens are hard to identify correctly, such as ‘anticuerpos antitransglutaminasa tisular IgA’. (3) drugs: model cannot recognize drugs such as ‘dasatinib’, ‘nilotinib’ and so on.

Since we experimented with a pipeline model, the mistakes of task 1 will be propagated to task 2 and there are about 8% errors caused by track1. In addition, about 10% errors are caused by the matching model. We summarized the modes of low recall rate by standard terminology selection methods when constructing <mention, standard terminology> pairs. The modes are: (1) about 40% entities are abbreviations, which is difficult to find the candidates from SNOMED-CT; (2) about 20% of entities have the same candidates in SNOMED-CT[6], which are not normalized entities in the shared task.

For further improvements, there may be two directions: (1) using joint learning methods for task 1 and task 2. (2) integrating knowledge graph into our system.

5 Conclusion

In this study, we developed a deep learning-based pipeline system for the PharmaCoNER shared task, a challenge specifically for clinical entity extraction from clinical text in Spanish.

Acknowledgments

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

References

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

[6] https://browser.ihtsdotools.org/?perspective=full&conceptId1=404684003&edition=MAIN/SNOMEDCT-ES&release=&languages=es,en

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. SemEval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310.

Gonzalez-Agirre, Aitor, Marimon, Montserrat, Marimon, Montserrat, Rabal, Obdulia, Villegas, Marta, and Krallinger, Martin. 2019. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In *Proceedings of the BioNLP Open Shared Tasks (BioNLP-OST)*, pages 1--X. Association for Computational Linguistics, November.

Cyril Grouin, Asma Ben Abacha, Delphine Bernhard, Bruno Cartoni, Louise Deleger, Brigitte Grau, Anne-Laure Ligozat, Anne-Lyse Minard, Sophie Rosset, and Pierre Zweigenbaum. 2010. CARAMBA: concept, assertion, and relation annotation using machine-learning based approaches. In

Jianglu Hu, Xue Shi, Zengjian Liu, Xiaolong Wang, Qingcai Chen, and Buzhou Tang. 2017. HITSZ CNER: A hybrid system for entity recognition from chinese clinical text. In *CEUR Workshop Proceedings*, volume 1976, pages 25–30.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, Joao Palotti, and Guido Zuccon. 2016. Overview of the CLEF eHealth evaluation lab 2016. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 255–266. Springer.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2013. NCBI at the BioCreative IV CHEMDNER Task: Recognizing chemical names in PubMed articles with tmChem. In *BioCreative Challenge Evaluation Workshop*, volume 2, page 34. Citeseer.

Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. 2017. CNN-based ranking for biomedical entity normalization. *BMC bioinformatics*, 18(11):385.

Zengjian Liu, Yangxin Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Haodi Li, Jingfeng Wang, Qiwen Deng, and Suisong Zhu. 2015. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of biomedical informatics*, 58:S47–S52.

Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. 2013. Overview of the NTCIR-10 MedNLP Task. In *NTCIR*. Citeseer.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(Sep):1453–1484.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.