Yunnan-Deep at eHealth-KD Challenge 2021: Deep Learning Model for Entity Recognition in Spanish Documents

Zhengyi Guan[1][0000−0002−6265−7193] and Renyuan Liu[2][0000−0002−8355−2390]

[1] School of Information Science and Engineering Yunnan University, Yunnan, P.R. China
1941028528@qq.com

[2] School of Information Science and Engineering Yunnan University, Yunnan, P.R. China
bluewind159@gmail.com

Abstract. This paper describes our contribution of the Yunnan-Deep team for the shared task of IberLEF eHealth Knowledge Discovery Challenge 2021. This task includes two subtasks, subtask A and subtask B. we only finished the subtask A. The goal of subtask A is to identify all the entities per document and their types about the eHealth documents written in Spanish. This is a typical entity recognition problem for health text. We mainly used the classic BERT + BiLSTMs + CRF model architecture and replaced BERT with BETO as the pre-training model. BETO is a BERT model pre-training on Spanish text. And we add CNN before BiLSTM for further feature extraction. Our model architecture performed well in the training set and development set for identifying the types of entities and the F1-score was over 0.7. But in the end, the output of the test set was performed not so well because there may exist a few errors in the final calculation of the word span in the whole test. The accuracy of the final test set was 0.52036, and the F1-score only reached 0.3306. Our model is competitive among all the contestants of subtask A, ranking 6th in the final leaderboard.

Keywords: eHealth · Entity Recognition · Convolutional Neural Network· Natural Language Processing.

1 Introduction

This paper introduces our work in IberLEF eHealth Knowledge Discovery Challenge 2021 [13]. Name Entity Recognition (NER) is a subtask of Information Extraction. It mainly designs how to extract target entities from text and classify the entities to what they belong. As is shown in Fig. 1: there are four types

IberLEF 2021, September 2021, Málaga, Spain.

for entities. They are respectively "Concept", "Action", "Predicate" and "Reference". What we have to do is to identify these entities and determine the span of the entities in the entire text. There are four main types of solutions for NER: dictionary-based method, rule-based method, machine learning-based method and recent popular deep learning-based method. In the method based on deep learning, the classic BILSTM-CRF [6] model and Transformer [14] are applied to multiple sequence lists and tasks in the general field. Among them, the accuracy of part-of-speech tagging and NER have reached very good results. The deep learning algorithm eliminates the need to manually extract features and we can independently extract efficient discriminative features. In previous studies, Limsopatham et al. [9]used convolutional neural networks (CNN) to normalize phrase entities in user reviews and Tutubalina et al. [11]applied recurrent neural networks [8] (RNN) in practice. These are the applications of deep learning in NER. After reading the papers of Salvador Medina et al. [10] in the eHealth 2020 Challenge, we find that they tried to apply convolutional neural networks to NER and our model borrowed from this idea and fine-tuned our model.

Our system experiment tried three different model architectures. These three models use popular techniques in deep learning. They are BETO[3] [1], BiLSTM [5], CRF [7] and CNN. The first model is BETO + CRF, the second model adds BiLSTM after BETO and the third model proposed in this paper is the architecture adding CNN to the second model.

The remainder of the paper is organized as follows: Section 2 and 3 describe the different architectures used by the system and result. In Section 4, the discussion on the development set and the official results in the challenge are presented. Finally, Section 5 presents the conclusions of the paper along with some future work recommendations.

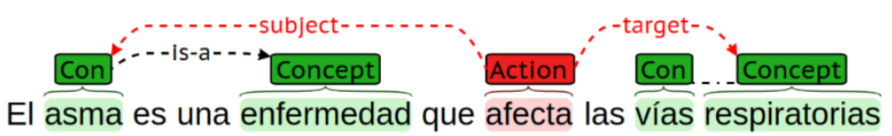

Fig. 1. Example of eHealth-KD annotations in the sentence "Asthma is a disease that affects the respiratory tract."

2 System Description

Because there are still some problems in the research of named entity recognition in the healthy field. Such as: the recognition accuracy of professional medical vocabulary is not high enough. The number of new medical entities is increasing

[3] https://github.com/dccuchile/BETO

rapidly, so a large number of unregistered words need to be recognized. Existing named entity recognition methods can't meet the requirements of cross-domain and multiple types of entity recognition. In response to the above problems, this paper uses the pre-training model BETO with stronger text feature representation capabilities as the feature representation layer. After reading the eHealth 2020 papers, we fine-tuned the above architecture. We added CNN to further extract features of text. We regard NER as a classification problem at the token level. We combined BETO with CNN, BiLSTM and CRF model to extract global and local features of the text. That means we input these word vectors into CNN, BiLSTM and CRF to calculate the corresponding scores and perform entity recognition and BIO labeling. The model architecture is shown in Fig. 2.

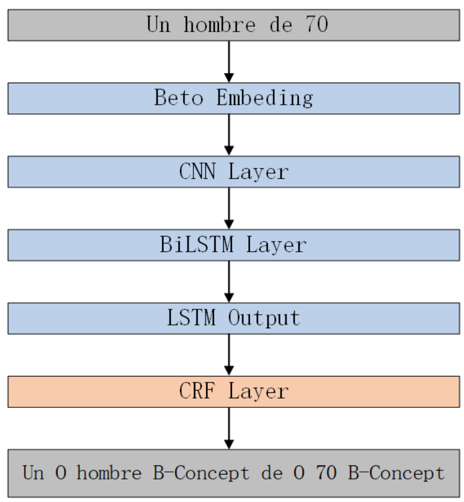

Fig. 2. Our model

2.1 Embedding Layer

Given the target sentence and the highlighted entities input as raw text, some preprocessing is done in order to derive useful structures from such text. Since both models make use of word-piece information, the input sentence must be tokenized first [12]. To obtain a representation of the corresponding inputs, model need to do the word embedding for each word. Since BERT [2] has been particularly popular in recent years, we plan to use other people's pre-training models as contextual embedding. But because BERT does not have a good effect in Spanish. We use BETO [1] to replace BERT. BETO is similar to BERT. They all have 12 hidden layers. BETO is a BERT model trained on a big Spanish corpus. BETO was trained with the whole word masking technique. Representing each word in the sentence as a vector, which includes word embedding and character embedding. The character embedding is initialized randomly. The word embedding is usually imported from a pre-training word embedding file. All embeddings will be fine-tuned during training.

2.2 CNN Layer

Because of the characteristics of convolution, CNN can't capture the long-distance dependence information in the text sequence. However, due to the window sliding mechanism in convolution operation, we can obtain obvious local features by controlling the size of convolution kernel. In the CNN module, we extracts local information between different distances by setting two convolutional layers with different kernel sizes. Then we performed data compression through maximum pooling operation while reducing data redundancy [16]. That means after BERT generated the word vector, we added two convolutional layers. Each layer is processed by Relu activation function and maximum pooling. Finally, the convolutional word vector is sent to the BiLSTM layer.

2.3 BiLSTM Layer

In sequence labelling tasks, the RNN model can dynamically capture sequence data information and store the information in memory, but it is easy to cause problems such as gradient disappearance. Compared with the RNN model, LSTM [15] adds a memory unit to the hidden layer, which solves the problems of gradient disappearance or gradient dispersion caused by long sequence information. Meanwhile, LSTM adds a threshold mechanism to selectively store and discard the required information. Therefore, it is widely used in named entity recognition tasks.

The traditional one-way LSTM model can't process contextual information at the same time, so Graves A et al. [4] used the basic memory unit of LSTM to construct the BiLSTM model, using a forward and a backward LSTM module, respectively, connected to the same output layer. Two different hidden layer representations are obtained by calculating in order (from left to right according to the sentence direction) and reverse order (from right to left according to the reverse of the sentence) for each sentence, and the final hidden layer representation is obtained through vector splicing. BiLSTM can better capture semantic dependencies, learn more comprehensive contextual and semantic co-occurrence information than LSTM. Therefore it can effectively use the context information of text sequences [3]. The output of the BiLSTM layer is the score of each label. These scores will be used as input to the CRF layer. Then, all the scores predicted by the BiLSTM layer are input to the CRF layer. In the CRF layer, the tag sequence with the highest prediction score is selected as the best answer.

2.4 CRF Layer

However, BiLSTM does not always get the correct prediction results. The CRF layer can learn the constraints of sentences. The CRF layer can add some constraints to ensure that the final prediction result is valid. These constraints can be automatically learned by the CRF layer during training data. And although the BiLSTM model can identify entity boundaries, it does not consider whether the relationship between the entity sequences is correct while the CRF model can obtain the global optimal tag sequence by considering the dependency relationship between adjacent tags, so it is often applied in tasks such as part-of-speech tagging and named entity recognition. CRF is a sequence labelling algorithm proposed on the basis of the EM model and the HMM model. Hence, it can solve the label bias problem by considering the global information of the label sequence and can better predict the label. The basic principle of CRF is to calculate the conditional probability distribution of the output random variable with a given random variable as input, usually using the Viterbi algorithm for decoding. The CRF model used in named entity recognition is to use the word sequence in the input sentence as the observation sequence, and the labelling process is to infer the most likely label sequence based on the known word sequence. Therefore, by combining CRF with BiLSTM neural network and reprocessing the output of BiLSTM, the output result of the BiLSTM is processed and revised again to obtain the best entity annotation [3].

3 Experimental Setup and Results

The official organizers provide training set, development set and test set. At the beginning of training, we first marked the official training set and development set with BIO(BIO is a commonly used mode for sequence labeling), as is shown in Fig 3. In the final prediction of the type of entity, we only predicted the first 50 Spanish sentences in subtask A. That means we did not include all the predicted sentences, which may affect the generalization ability of the model.

The deep learning framework used in this experiment is PyTorch [4]. All experiments are run on a GPU called Tesla P100-PCIE. The learning rate of all experiments is 5e-5. Because the experiments have made early stop judgments, the final epoch is 3 and the batch size of training and development is 4. Our experiment provides three runs for subtask A. They are:

- Run 1: BETO + CRF.
- Run 2: BETO + BiLSTM + CRF.
- Run 3: BETO + CNN + BiLSTM + CRF.

The results of our experiment are shown as Table 1. The metrics defined by the eHealth-KD challenge to evaluate the submitted experiments are those commonly used for some NLP tasks such as NER or text classification, namely precision, recall and F1-score.

[4] https://pytorch.org/

```
Se O
ha O
utilizado B-Action
contra O
COVID-19 B-Concept
debido O
a O
la O
similitud B-Concept
entre O
las O
dos B-Concept
enfermedades I-Concept
. O
```

Fig. 3. Example of BIO tagging scheme for entity recognition

Table 1. The result for our three runs on the devlepment set.

Task	Run	Precision	Recall	F1-score
	Run 1	**0.7327**	0.6786	0.7047
SubtasA	Run 2	0.7305	0.7114	0.7208
	Run 3	0.7314	**0.7184**	**0.7249**

4 Discussion

According to the results in Table 1, we can see that the pre-training model BETO + BiLSTM + CRF architecture has achieved good results for F1-score. The two layers of CNN that we added before BiLSTM are only slightly improved. But after adding CNN and BiLSTM, the accuracy of the experiment has slightly decreased. We should also notice that this may be related to the selection of experimental parameters. In our experiment, the parameters of our three Runs are the same. To some extent, we should try to fine-tune our models using different parameters. We should point out that since we only predicted 50 Spanish sentences, which affected our testing with the model. Our result of the precision, recall and the F1-score on the final leaderboard are 0.52036, 0.24599 and 0.33406. However, when calculating the span of words in the entire text, this obviously shows that our system still needs to be improved.

5 Conclusions

This work described the system proposed by Yunnan-Deep team at the IberLEF eHealth-KD 2021: eHealth Knowledge Discovery challenge. We just completed the research of subtask A. For subtask A, we mainly use BETO + CNN+ BiLSTM + CRF architecture. At the same time, we did two comparative experiments. Our model has a good effect on the entity annotation on the development set. Since we considered the generalization ability of the model, I believe that the recognition of entity types in the final test set should also have good result. This also shows that deep learning has a very good effect on NER problems. But the final result is not very perfect on the test set. That means our model should be perfected and We should use more different types of corpus to test our model.

In the future, We should work hard on data enhancement and take different types of language into consideration. In addtion, we can also try to integrate different pre-training models before the word vector is input to BiLSTM. What's more, we will study the performance of using more linguistic features such as Part-Of-Speech tags as an input in the neural network, as well as the use of ontologies related to the biomedical domain and other types of word embeddings. Finally, our team will compare the model with more advanced models and verify the effectiveness of the model on more data sets to improve generalization capabilities.

References

1. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
2. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805

3. Gao, W., Zheng, X., Zhao, S.: Named entity recognition method of chinese emr based on bert-bilstm-crf. Journal of Physics: Conference Series **1848**(1), 012083 (9pp) (2021)
4. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5–6), 602–610 (2005)
5. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5), 602–610 (2005). https://doi.org/https://doi.org/10.1016/j.neunet.2005.06.042, https://www.sciencedirect.com/science/article/pii/S0893608005001206, iJCNN 2005
6. Huang, Z., Wei, X., Kai, Y.: Bidirectional lstm-crf models for sequence tagging. Computer Science (2015)
7. Lafferty, J.D., Mccallum, A.K., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
8. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification (2015)
9. Limsopatham, N., Collier, N.: Adapting phrase-based machine translation to normalise medical terms in social media messages. Computer Science (2015)
10. Medina, S., Turmo, J.: TALP at ehealth-kd challenge 2020. In: Cumbreras, M.Á.G., Gonzalo, J., Cámara, E.M., Martínez-Unanue, R., Rosso, P., Zafra, S.M.J., Zambrano, J.A.O., Miranda, A., Zamorano, J.P., Gutiérrez, Y., Rosá, A., Montes-y-Gómez, M., Vega, M.G. (eds.) Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020. CEUR Workshop Proceedings, vol. 2664, pp. 85–93. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2664/eHealth-KD_paper1.pdf
11. Miftahutdinov, Z., Tutubalina, E.: Deep neural models for medical concept normalization in user-generated texts (2019)
12. Pérez, A.R., Caballero, E.Q., Alvarado, J.M., Linares, R.C., Consuegra-Ayala, J.P.: UH-MAJA-KD at ehealth-kd challenge 2020. In: Cumbreras, M.Á.G., Gonzalo, J., Cámara, E.M., Martínez-Unanue, R., Rosso, P., Zafra, S.M.J., Zambrano, J.A.O., Miranda, A., Zamorano, J.P., Gutiérrez, Y., Rosá, A., Montes-y-Gómez, M., Vega, M.G. (eds.) Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020. CEUR Workshop Proceedings, vol. 2664, pp. 125–135. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2664/eHealth-KD_paper5.pdf
13. Piad-Morffis, A., Gutiérrez, Y., Estevez-Velarde, S., Almeida-Cruz, Y., Muñoz, R., Montoyo, A.: Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021. Procesamiento del Lenguaje Natural **67**(0) (2021)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), http://arxiv.org/abs/1706.03762
15. Wang, J., Peng, B., Zhang, X.: Using a stacked residual LSTM model for sentiment intensity prediction. Neurocomputing **322**(DEC.17), 93–101 (2018)
16. Wang, T., Zhang, Y., Li, Y.: A parallel-attention model for tumor named entity recognition in spanish. In: Cumbreras, M.Á.G., Gonzalo, J., Cámara, E.M., Martínez-Unanue, R., Rosso, P., Zafra, S.M.J., Zambrano, J.A.O., Miranda, A., Zamorano, J.P., Gutiérrez, Y., Rosá, A., Montes-y-Gómez, M., Vega, M.G. (eds.) Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located

with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020. CEUR Workshop Proceedings, vol. 2664, pp. 438–446. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2664/cantemist_paper12.pdf