# Yunnan-1 at eHealth-KD Challenge 2021: Deep-Learning Methods for Entity Recognition in Medical Text

Maoqin Yang[1][0000−0001−6966−4436]

Yunnan University, Yunnan, P.R. China
maomaoq33@gmail.com

**Abstract.** The IberLEF eHealth-KD Challenge 2021, held at IberLEF 2021, proposes two subtasks to encourage the development of systems for automatically extracting knowledge from unstructured Spanish eHealth texts. I only participate in subtask A: Entity Recognition. This subtask aims to identify all the entities and their types for the given eHealth documents. This paper describes the system presented by team-Maoqin in the challenge. Several deep learning models are used in plain text documents, such as BERT-CRF, BiLSTM-CRF. Only the best result running in the local has been submitted. But the final result of my method is indeed very low, with a score of 0.173 (F1), ranking 9th on the leaderboard. Additional work needs to be done to improve the final result and complete the subtask B: relation extraction.

**Keywords:** eHealth · Entity Recognition · Deep-Learning Model.

## 1 Introduction

In recent years, the amount of medical documents produced by the scientific community has been increased. Those texts combine many corporates, it is necessary to extract useful knowledge with automatic methods. Therefore, various competitions have been held in the past years such as task 7 of SemEval 2018 [3] and the eHealth-KD challenge at IberLEF 2020 [9]. This paper presents the system description of team-Maoqin in the IberLEF eHealth-KD challenge subtask A at IberLEF 2021 [8]. The purpose of this task is to identify all entities of a given Spanish document and determining which of the four categories of "Action","Concept", "Predicate", and "Preference" these entities belong to.

The pre-trained and deep learning models have shown excellent performance in many NLP tasks such as text classification, reading comprehension and question answering [11]. So the system uses several deep learning methods. Advanced

*IberLEF 2021, September 2021, Málaga, Spain.*

methods for named entity recognition (NER) such as BERTbase [2], BiLSTM [4] have been the main components of the model. Both of these methods add a CRF [6] layer in that the CRF layer can add some constraints to the last predicted label to ensure that the predicted label is legal. In the training process of training data, these constraints can be automatically learned through the CRF layer.

The rest of this article is arranged as follows. Section 2 introduces the different architectures used by the system. In Section 3, the formal results obtained in the challenge are introduced. In Section 4, some ideas on improving the quality of each strategy and some unfinished experiments will be shared. Finally, Section 5 introduces the conclusions of this paper and some opinions for future work.

## 2 System Description

This part is used to comprehensively describe the model mentioned and how to process input and output data. Input handling, where I adjust the data format for use in the model. The description of the structure includes architecture and parameters set. After getting the prediction result, the result format should comply with official rules.

### 2.1 Input Handling

The official sentence example is shown in figure 1 and the corresponding format should be predicted as figure 2. The training and development corpora are both provided in Brat format which consists of an id increasing by row, the entity type, the next two numbers indicate the span of this entity and the word in last. If adjacent words belong to the same entity, write them on one line, the span should be separated by semicolons. The neural network only accepts token level so the documents will be converted to output the models can use with the provided txt and ann file. The input file is finally changed into $BIO$ format and input into the model. Among them, $B$ means beginning, $I$ means inside, and $O$ means non-entity. This scheme is the most popular in the NER task although it presents problems when the entity contains discontinuous tokens [7].

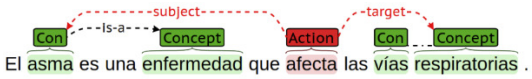


**Fig. 1.** The example of given annotations.

### 2.2 Model Structure

This section will introduce my models and show their structure. The model is trained and evaluated on the official training set. The training set and the

```
T1      Concept 3 7     asma
T2      Concept 15 25   enfermedad
T4      Action 30 36    afecta
T3      Concept 41 45;46 59     vías respiratorias
```

**Fig. 2.** The corresponding entities and spans.

development set each contains 100 pieces of data, and the test set has 50 pieces. The models are deep neural networks that receive the input tokens and jointly emit predictions for several different output variables. These predictions can be classified into tokens. All codes are executed on the GPU version of the colab platform.

**BERT-CRF Model** The BERT model has performed excellently in many areas in NLP. The tokens are fed into the BERT model to obtain their contextual embeddings. These embeddings are passed to a classification layer that emits logits with the prediction about each token being or not an entity of a certain type. Each entity type has $B$ and $I$ (8 in total) plus $O$, $PAD$, $SEP$, and $CLS$, so the input contains a total of 12 labels.

An early stopping mechanism (stop training when the accuracy drops within 2 consecutive epochs) is set up to prevent excessive invalid training in the experiment. The max sequence length is 200, the batch size is 4, the amount of hidden layer is 12, the number of training steps is 500, the number of attention heads is 12, the learning rate is 0.5, and the dropout probability is 0.1. The activation function used is ReLU. The best results on the validation set were obtained with an F1 score of 0.55 ( precision = 0.53, recall = 0.57).

**BiLSTM-CRF Model** By combining BiLSTM and CRF [5], the model can consider the correlation between the sequence before and after the sequence like CRF and have the feature extraction and fitting capabilities of LSTM. The whole model is presented in figure 3.

The input of the model is a word sequence, each word in the sentence is expressed as a vector, which should include word embedding and character embedding. And the output is the label predicted by the model for each word, which is a label sequence. The BiLSTM model is used to generate the emission matrix, that is, the probability that each word is marked as a certain label. The emission matrix of the BiLSTM model does not take the constraint relationship between labels into account. For example, in the $BIO$ system, $I$ can not appear after $O$. So the connection order of tags has to restrict, which will be generated by the CRF model. The CRF model is used to learn the constraint relationship between labels and generate a transition matrix, which can be understood as the probability of connecting another tag behind one tag. When the entire model predicts, it combines the transmission matrix and the transition matrix and uses the $Viterbi$ decoding algorithm to calculate the label sequence with the highest score. The whole process is as follows: the text data is input into the LSTM

network, and then into the upper CRF network, and finally, the annotation data is generated.

In this model, the learning rate is 0.001, the batch size is 5, the dropout is 0.5, the optimization algorithm used is Adam. The best results on the validation set are obtained with an F1 score of 0.56 ( precision = 0.53, recall = 0.61).

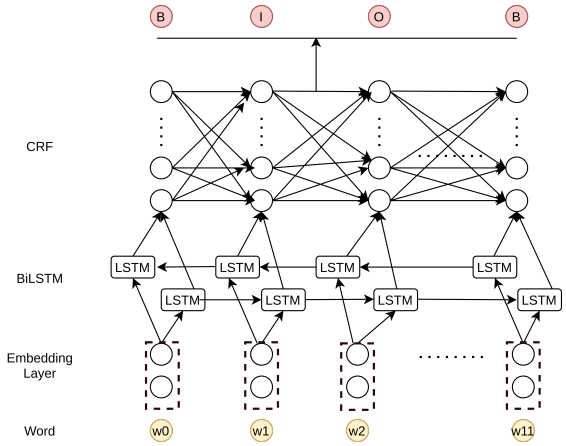


**Fig. 3.** The architecture of the BiLSTM-CRF model, where the $w_i$ represents the input token, and the last units mean the output.

### 2.3 Output Handling

The output of the neural network needs to be converted back to Brat's [10] format and the annotation schema proposed in the task. Please remind me that I do not use the official ann generator file to get the word span information. I only predict the entity type, because when the input text has span information, a little error occurred, so all the span produces from my own code. When getting the predicted file, then combine the span to get the submit file.

## 3 Results

The results for each method are shown in Table 1. I provide the results on the development data and the officially published results on the training data. I have tried to replace the BERT model with the ALBERT model, but the effect was not as good as BERT, so abandoned this idea. The BioBERT-CRF model was also had been tried, the effect was similar to BERT-CRF, so it was not been included in the selection.

The best performing version of both the BERT-CRF and the BiLSTM-CRF model is then run on the test set of the eHealthKD corpus. The best score obtained is from the BiLSTM-CRF model is 0.56 of F1. The BERT-CRF model, however, obtained an F1 score of 0.55 on the shared task. Even after downloading my own results, I manually checked the meaning of words with a browser and checked some entities' span. Most of the entity classifications are fine. I don't know why the score is so low.

**Table 1.** The first two lines represent the local performance score of the model and the last line is the final official score result. The results are run on the test set.

| Model | F1 | Precision | Recall |
|---|---|---|---|
| BERT-CRF | 0.55 | 0.53 | **0.57** |
| BiLSTM-CRF | **0.56** | 0.53 | 0.61 |
| Official Score | 0.17 | 0.27 | 0.12 |

All in all, this task is still a long way from being completely completed. Both the problem of the model and the problem of the task need to be solved in the future.

## 4 Discussion

Several models were trained and tested in the test set. Using BERT to embed words can benefit the performance of subtask A. In fact, only by using BERT representations and linear dense layers, competitive results can be obtained in the test set. Although the official gave us the code to extract the relationship, I still at the first step and failed to extract all the relationships. This is a very simple step, but this is the first time have done such a task, and this difficulty has not yet been effectively resolved. Later, due to time constraints, gave up subtask 2.

The final result only scores 0.17, the possible reasons are as follows:

. There is a problem with the evaluation method used locally, which makes the local score look good.
. The official evaluation result is for English plus Spanish, and I only completed the entity classification of Spanish, so the overall effect is much worse than that in the local area.
. The span straddling is realized by myself, and something may have gone wrong.
. The data number is less.

## 5 Conclusions

This paper describes the participation of team Maoqin in the eHealth-KD challenge at IberLEF 2021. For one method, the pre-trained BERT model has been

taken as the basic representation. For another method, BiLSTM has been preferred. The system achieved unsatisfied results in the challenge for official evaluation in subtask A. Due to the colab's time and resources limitation, some strategies were not achieved. And subtask B: Relation Extraction has not been completed.

Future work will be done to solve those remainders. More work needs to be done to improve the performance. Other methods such as transfer learning need to try. Still, further experimentation is required to understand the impact of the network's components and how to improve them, which I will explore in future work. Citing an external corpus may improve the results. The BETO [1] model, a BERT model pre-trained on Spanish text may improve the results.

## References


1. Canete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR **2020** (2020)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Gábor, K., Buscaldi, D., Schumann, A.K., QasemiZadeh, B., Zargayouna, H., Charnois, T.: Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 679–688 (2018)
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. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
6. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
7. López-Úbedaa, P., Perea-Ortegab, J.M., Díaz-Galianoa, M.C., Martín-Valdiviaa, M.T., Ureña-Lópeza, L.A.: Sinai at ehealth-kd challenge 2020: Combining word embeddings for named entity recognition in spanish medical records (2020)
8. 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)
9. Piad-Morffis, A., Gutiérrez, Y., Cañizares-Diaz, H., Estévez-Velarde, S., Muñoz, R., Montoyo, A., Almeida-Cruz, Y.: Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020. In: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing (2020)
10. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: a web-based tool for nlp-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107 (2012)
11. Su, D., Xu, Y., Winata, G.I., Xu, P., Kim, H., Liu, Z., Fung, P.: Generalizing question answering system with pre-trained language model fine-tuning. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering. pp. 203–211 (2019)