A Multi-Task Learning Framework for Extracting Bacteria Biotope Information

Qi Zhang[1], **Chao Liu**[2]*, **Ying Chi**[2], **Xuansong Xie**[2] **& Xiansheng Hua**[2]
[1]Zhejiang University, [2]Alibaba DAMO Academy
zhangqihit@zju.edu.com
{ maogong.lc, xinyi.cy, xiansheng.hxs}@alibaba-inc.com
xingtong.xss@taobao.com

Abstract

This paper presents a novel transfer multi-task learning method for Bacteria Biotope rel+ner task at BioNLP-OST 2019. To alleviate the data deficiency problem in domain-specific information extraction, we use BERT(Devlin et al., 2018) (Bidirectional Encoder Representations from Transformers) and pre-train it using mask language models and next sentence prediction (Devlin et al., 2018) on both general corpus and medical corpus like PubMed. In fine-tuning stage, we fine-tune the relation extraction layer and mention recognition layer designed by us on the top of BERT to extract mentions and relations simultaneously. The evaluation results show that our method achieves the best performance on all metrics (including slot error rate, precision and recall) in the Bacteria Biotope rel+ner subtask.

1 Introduction

Information extraction aims to recognize the entities and classify the relations between them in given unstructured text. It provides cornerstone for many downstream applications such as information extraction, knowledge base population, and question-answering. It is a challenging task partly because it requires elaborative human annotations (Riedel et al., 2010), which could be slow or expensive to get.

Bacteria Biotope (BB) task is an interesting information extraction task aiming at extracting knowledge about bacteria biotope from bioinfomatics literature related to microorganism. Rel+ner subtask focuses on extracting entity mentions of following types: Microorganism (MI), Habitat (HA), Phenotype (PH), Geographical (GE) and identification of the $Lives\_In$ relation between a Habitat/Geographical mention and a Microorganism mention as well as the $Exhibits$ relation between a Phenotype mention and a Microorganism mention. This task intends to extract structured triple of microorganism from unstructured biomedical text.

Some previous work has been done in handling such an information extraction problem, including some joint entity and relation extraction methodology and pipeline method which firstly do named entity recognition (NER) and then do relation extraction on the results of NER. (Zheng et al., 2017) proposes a novel tagging schema (NTS) that encodes relation type in the NER tag to recognize the named entity and extract the relation between them jointly. This methodology has a fatal flaw that it can not handle relation facts that share the same entity and this phenomenon is common in BB task. (Bekoulis et al., 2018) proposes a multi-head selection layer (MHS) to model the relation of each entity pair which is similar to our method. (Zeng et al., 2018) proposes a sequence to sequence model with copy mechanism (Copy RE). However, all above the previous work has been done on a large-scale general dataset. While the Bacteria Biotope rel+ner task only bases on a domain-specific and comparatively small dataset. Under this background, we adapt a recently widely used transfer learning framework, BERT(Devlin et al., 2018), and pre-train it on large-scale corpus using two novel unsupervised prediction tasks to mitigate the problem of insufficient data.

2 Model Architecture

The overall framework of the model is shown in Figure 1. Bottom parts of the model (including input representation, transformer encoder) are shared by both named entity recongnition task and relation extraction task.

*Corresponding author

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 105–109

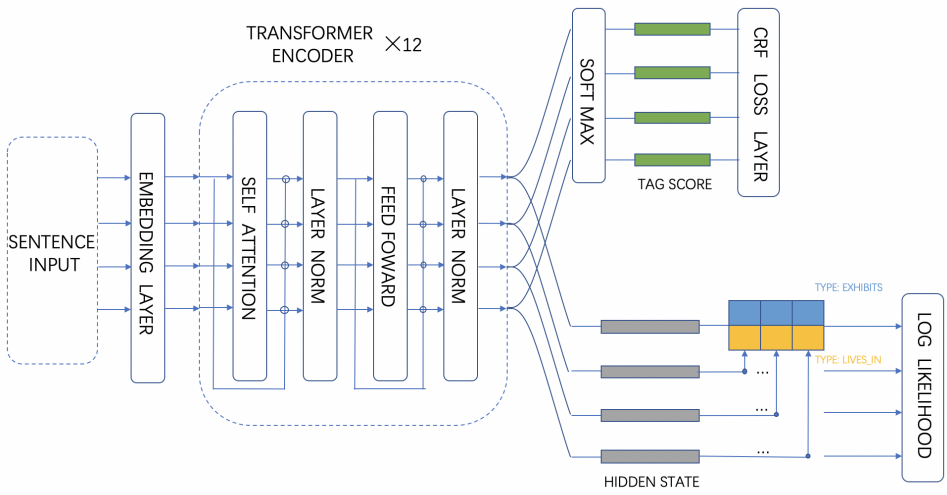

Figure 1: Diagram of Our Model

2.1 Input Feature and Representation

The input representation of each word w_i in sentence $S = \{w_1, w_2, ..., w_N\}$ consists of three parts: word vector, the embedding of features and positional encoding. A pre-trained word embedding using Skip-gram (Mikolov et al., 2013) model is used to map each word to a dense vector. The features we used are described in Table 1. Each feature is represented by a one-hot vector and pass a feature embedding layer. Positional encoding is added to make the model capture the relative and absolute position of each token (Vaswani et al., 2017). The three parts are concatenated and fed into transformer encoder.

2.2 Transformer Encoder

Transformer is widely used in various natural language processing task recently. We use transformer here to extract context features of each token. The encoder is composed of 12 layers. Each layer consists of a multi-head self attention sub-layer and point-wise fully connected feedfoward sub-layer, a residual connection is employed around each of the two sub-layers (Vaswani et al., 2017). The transformer is pre-trained using two novel unsupervised tasks including masked language model and next sentence predicting (Devlin et al., 2018) on the combination of BooksCorpus, English Wikipedia, PubMed and PubMed Central (PMC) corpus. The hyper-parameters we use to pre-train are exactly the same as the $BERT_{BASE}$ of (Devlin et al., 2018). In fine-tuning stage, the output of the transformer encoder H_i will be fed into both mention recognition layer and relation extraction layer.

2.3 Mention Recognition Layer

Commonly, in named entity recognition, annotated data is tagged using BIO tagging schema in which each token is assigned into one of following tag: B means beginning, I means inside and O means outside of an entity mention. However this tagging schema is insufficient since some entity mentions in BB task are disjoint concepts with overlapping words. Taking the phrase "serotypes A, B and C" as an example, this phrase contains three disjoint Microorganism mentions: "serotypes A", "serotypes B" and "serotypes C". To handle these special mentions, we apply an alternative tagging schema which introduce 'H' and 'D' flag, where 'H' indicates the overlapping tokens and 'D' indicates discontinuous tokens. Figure 2 shows an annotation example. The tagging label set of this new tagging schema can be written as $\{\{GE, HA, PH, MI\} \times \{H, D\}\} \times \{B, I\} \bigcup \{O\}$.

We feed the final state H_i of each token to the softmax classification layer over the tagging set. The conditional random field (CRF) layer takes the sequence of output score vector V_i from the softmax classification layer. The tag prediction of w_i in sentence s is denoted as y_i^S, and further the CRF score of the tag predictions $y^S = \{y_1^S, y_2^S, ..., y_N^S\}$ is defined as follows:

$$score_{y^S} = E_{y^S} + T_{y^S} \quad (1)$$

E represents emission score which can be defined

Feature Name	Description
Dot Flag Feature	Whether the word contains dot notations like “C. psittaci”.
Capitalization Feature	Whether the first letter of the word is capitalized.
POS Tagging Feature	The output for the tokenized sentence of the POS tagging tool.
Dependency Parsing Feature	The output for the tokenized sentence of the dependency parsing tool.

Table 1: Input Features of Our Model and Their Description

as:

$$E_{y^S} = \sum_{i=1}^{N} V_i \quad (2)$$

T represents transition score which can be defined as:

$$T_{y^S} = \sum_{i=1}^{N} TM_{y_{i-1}^S, y_i^S} \quad (3)$$

where TM_{y_{i-1},y_i} means the transition probability from tag y_{i-1} to y_i. The conditional probability $P(y|S)$ can be written as follows:

$$P(y^S|S) = \frac{e^{score_{y^S}}}{\sum_{y \in y^*} e^{score_y}} \quad (4)$$

where y^* is the collection of all possible tag predictions for sentence s.

Serotypes	A,	B	and	C
HB-MI	DB-MI	DB-MI	O	DB-MI

Figure 2: Examples of BIOHD Tagging

2.4 Relation Extraction Layer

As depicted in Figure 1, the sequence of final state H_i is also fed into the relation extraction layer. We observe that each Microorganism entity may have multiple relations with entities of other three types. Moreover, all types of relation must contain a Microorganism entity. Thus we take the Microorganism entity as the center of relation prediction task.

The Microorganism entity which ends with the word w_i will be calculated the following score with another entity end with the word w_j:

$$R_{i,j,r} = \sigma(W_r f(H_r * Vi + T_r * Vj + b_r)) \quad (5)$$

where H_r, T_r and b_r are parameter matrices associated with relation type r. The score $R_{i,j,r}$ represents probability that the Microorganism entity ends with words w_i has the relation r with another entity ends with words w_j. f is the activation function: relu. σ is used to normalize the probability.

2.5 Multi Task Training Objective

In training stage, we fine-tune the relation extraction layer and mention recognition layer simultaneously using a joint loss. The training loss defined by mention recognition layer can be written as:

$$L_{ner} = -logP(y^S|S) \quad (6)$$

Moreover, the loss function of the relation extraction layer can defined as

$$L_{rel} = \sum_i \sum_j -logR_{i,j,r} \quad (7)$$

The loss function of the whole system can be defined as

$$L = L_{ner} + L_{rel} \quad (8)$$

3 Experiment and Result

In this section, we briefly introduce the dataset, evaluation metrics and the external resources that we use. We present our performance on different relation type with different metrics provided by organizers and comparison with other jointly information extraction methodology mentioned in Section 1 on development data.

3.1 Dataset Description

Bacteria Biotope task includes two types of documents: PubMed references (titles and abstracts) related to microorganism, extracts from full-text articles related to microorganisms living in food products.

The statistics of the dataset is shown in Table 2. The training and development data released for this task contains 133 and 66 files respectively, with gold standard annotations. Test data contains 32 files which are used to evaluate participation. The number of entity mentions in different file is unbalanced, ranging from 0 to 85.

Table 2: The Statistics of The Dataset: Number of Files, Relations and Entities

	File	Entity	Relation
train	133	2266	1127
development	66	1271	608
test	32	Unknown	Unknown

Table 3: Performance for Each Relation Type

	SER	precision	recall
All-types	0.954	0.509	0.351
Exhibits	0.982	0.492	0.449
Lived_in-geo	1.318	0.316	0.273
Lived_in-habitat	0.927	0.530	0.311

3.2 External Resources

Here we introduce some external resources that we use in experiment. We use Google Word2vec tool to train word embeddings on corpora composed of PubMed, PubMed Central (PMC) corpus and English Wikipedia corpus. The LTP tool is used for sentence level dependency parsing and the NLTK tool is used for sentence tokenization and part of speech tagging.

3.3 Metric and Performance Comparison

Since the entity mentions which are potential arguments of each relation, are not given. In evaluation metrics (precision, recall), substitution errors are penalized. Morever, Slot Error Rate (SER) is taken as the main evaluation metric. Table 3 shows our results of different relation type.

We also evaluate some previous with famous jointly information extraction methodologies which are described in Section 1 on the BB 2019 development data for comparison:

NTS: Our implementation of (Zheng et al., 2017). Instead we use the tagging schema described in Section 2.3.

MHS: We use the code released by (Bekoulis et al., 2018) and train the model on the training data of BB rel+ner task.

Copy RE: Our implementation the sequence to sequence model using copy mechanism (Zeng et al., 2018). We train the model using the training data of BB rel+ner task.

Pipeline: The baseline method that we use includes two step separately: perform NER (Devlin et al., 2018) firstly, then perform relation extraction (Devlin et al., 2018) on the results of the NER task.

Table 4: Performance indicates statistically significant difference from our model, NTS, MHS, Copy RE and Pipeline.

	SER	precision	recall
Pipeline	1.472	0.231	0.294
NTS	1.456	0.261	0.288
MHS	1.183	0.381	0.302
Copy RE	1.128	0.376	0.291
Our model	0.947	0.493	0.339

As shown in Table 4, our model achieves improvements on BB dataset comparing with the other four models. Particularly, our model significantly outperforms the **Pipeline** baseline by -0.525 SER.

3.4 Factor Analysis

We propose several strategies to improve the performance including feature engineering and utilizing the transformer encoder. To investigate the influence of these two factors, we conduct ablation study and list results on Table 5 .

"No" prefix in Table 5 means that we train and evaluate our model without the corresponding feature. "No Transformer Encoder" indicates that we replace the transformer with bi-directional lstm.

Results show that each feature listed in Table 1 plays a key role. Our model suffers serious performance degradation without any one of the four input features.

Table 5: Ablation Study

Model	SER	P	R
Our Model	0.947	0.493	0.339
No Dot Flag Feature	0.961	0.485	0.313
No Capitalization Feature	0.956	0.489	0.324
No POS Tagging Feature	0.949	0.499	0.335
No Dependency Parsing Feature	0.951	0.487	0.333
No Transformer Encoder	0.998	0.470	0.321

4 Conclusions

In this paper, we describe our participation in Bacteria Biotope rel+ner subtask. We propose a transfer multi-task learning framework to overcome data deficiency and fine-tune a joint entity and relation extraction model using multi-task training objective. Though we achieve the best performance in this subtask, we have some future direc-

tions to improve this work furthermore: adapting adversarial training or posterior regularization to improve the performance of our system.

References

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, Jun Zhao, et al. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism.

Suncong Zheng, Wang Feng, Hongyun Bao, Yuexing Hao, Zhou Peng, and Xu Bo. 2017. Joint extraction of entities and relations based on a novel tagging scheme.