Named Entity Recognition for Chinese biomedical patents

Yuting Hu
LIACS
Leiden University, the Netherlands
yukihuyt@gmail.com

Suzan Verberne
LIACS
Leiden University, the Netherlands
s.verberne@liacs.leidenuniv.nl

Abstract

There is a large body of work on Biomedical Entity Recognition (Bio-NER) for English, but there have only been a few attempts addressing NER for Chinese biomedical texts. Because of the growing amount of Chinese biomedical discoveries being patented, and lack of NER models for patent data, we train and evaluate NER models for the analysis of Chinese biomedical patent data, based on BERT. By doing so, we show the value and potential of this domain-specific NER task. For the evaluation of our methods we built our own Chinese biomedical patents NER dataset, and our optimized model achieved an F_1 score of 0.54±0.15. Further biomedical analysis indicates that our solution can help detecting meaningful biomedical entities and novel gene–gene interactions, with limited labeled data, training time and computing power.

1 Introduction

In the area of biomedical NLP, Named Entity Recognition (NER) is a widely discussed and studied topic. The aim of the task is to identify biomedical entities such as genes, proteins, cell types, and diseases in biomedical documents, to allow for knowledge discovery in this domain. Models for Biomedical NER (Bio-NER) offer the opportunity to mine information and knowledge and thereby foster biomedical and drug discovery research (Habibi et al., 2017). Several shared tasks addressing Bio-NER have been organized. These attempts and tasks resulted in benchmark datasets for solving English Bio-NER tasks, e.g. the GENIA corpus (Kim et al., 2003), JNLPBA (Kim et al., 2004), and BC2GM (Smith et al., 2008).

However, when turning our attention to Chinese bio-NER, only a few attempts have been made, and these attempts either had limitations in text resource types and amount (Gu et al., 2008) or did not predefine biomedical named entity categories (Wang et al., 2019). Another limitation is that these attempts mostly focus on clinical texts or biomedical scientific publications but not include other relevant text resources, in particular biomedical patents. Many biomedical discoveries are patented in China, not only because of the encouraging policy on patentability of genetic products, but also since the existence of the speedily progressed and cheaper gene sequencing services (Du, 2018).

Patent texts are highly technical with long sentences (Verberne et al., 2010). Two additional challenges of Chinese biomedical patents that we encountered are OCR errors and the heavy usage of code-mixing expressions, mixing English and Chinese in one entity. This is mainly because the protein and gene names are commonly written in English (or the English names are given after the Chinese ones), while the disease names and other contents are written in Chinese. For this reason, it is not possible to directly apply pre-trained NLP models to Chinese biomedical patents. Moreover, as mentioned before, because of the lack of related studies, we not only lack pre-trained models which were trained on Chinese biomedical text data, but also well-organized Chinese biomedical patents datasets.

The contributions of this paper are threefold. First, we release a hand-labeled dataset with 5,813 sentences and 2,267 unique named entities from 21 Chinese biomedical patents. Second, we obtain promising results for the extraction of genes, proteins, and diseases with BERT models using our labeled data in limited training time and with limited computing power. Third, we show that when we use our

Proceedings of the 28th International Conference on Computational Linguistics, pages 627–637
Barcelona, Spain (Online), December 8-13, 2020

NER model to extract entities from a large patent collection, we can potentially identify novel gene–gene interactions. We release our data and code for use by others.[1]

In the following parts of this paper, we discuss previous attempts to solve Chinese Bio-NER tasks and other related tasks in Section 2; our methods and implementation details are explained in Section 3; the results of all experiments, along with the post analysis results, are described and discussed in Section 4; in Section 5 we discuss challenges and limitations of our study, followed by conclusions in section 6.

2 Related Work

Given the abundance of work on English biomedical NER, we focus on Chinese NER in this section. A commonly used state-of-art model for general-domain Chinese NER is the Chinese Lattice LSTM (Zhang and Yang, 2018), which applied the Lattice-structured RNN framework and made some novel modifications to ensure that the original Lattice network can handle the segmentation-free Chinese NER task.

In the biomedical domain, the paper by Gu et al. (2008) is an early attempt to solve the Chinese Bio-NER task for scientific literature. The authors designed several feature groups from a small labeled dataset consisting of Chinese biomedical research abstracts, then applied a Conditional Random Fields (CRF) model. Although they showed their best model obtained a 0.68±0.05 F_1 score among 50 runs, since their dataset was small (481 sentences and 1062 entities in total), the train-test split may have influenced the results. Another issue is that they split their train-test set by randomly selecting sentences from their whole corpus, which might have led to overestimation of the quality of the model because entities from the same document can then end up in both the train and test set, which may cause overfitting.

More recent work by Wang and Wu (2019) addresses open concept extraction from 4,931 biomedical articles which contain 41,733 sentences and 97,373 entities in total. The NER part did not specify different entity categories. Their NER method is based on dictionary matching and rules; the main contribution of the paper is a classification model for relation extraction. The authors reported a 0.76 F_1 score for their non-category-specific open concept NER and finally reached a 0.52 F_1 score for their main relation extraction task.

Since the introduction of the BERT paradigm (Devlin et al., 2019) there have been attempts to solve Chinese NER in different domains with BERT models.

There are three recent studies that used BERT models for Chinese biomedical NER, all three for clinical data. Xue et al. (2019) address Chinese clinical NER and relation extraction using BERT. They finetune the model on data from coronary arteriography reports in Shanghai Shuguang Hospital for five entity types: negation, body part, degree, quantifier, and location. They obtain high precision and recall scores with their model for these entity types (>0.95), thereby beating the Bi-LSTM baseline. Since the entity types in this study are very different from the entities in our work, we cannot use this work as a relevant comparison.

The papers by Dai et al. (2019) and Li et al. (2020) both use data from the CCKS competitions, containing Chinese medical patient records. Six clinical entity types are defined in these data sets: diseases and diagnosis, image inspection, laboratory inspection, operation, drug, and anatomic site. Dai et al. (2019) compare multiple NER models and obtain the highest F-score (up to 75%) with a BERT-BiLSTM-CRF model. The method by (Li et al., 2020), based on an optimized BERT model, outperformed all other methods in both the CCKS2017 and CCKS2018 clinical NER competitions, with average F-scores above 90%.

In this paper, we address a new biomedical text type for Chinese NER: patents. We use a pre-trained Chinese BERT model, and compare different learning methods to finetune it to this highly specific domain and genre.

3 Methods

In this section we describe data collection and cleaning, data annotation, models and learning methods, evaluation, and post-analysis.

[1] Our git repository can be accessed at: `https://github.com/yukihuyt/Chinese_biomed_patents_NER`

3.1 Data Collection and Cleaning

We retrieved two sets of patents from Google Patents. The first set contains patents matching the query "人类AND基因" ("human AND gene"), from 1st January 2009 to 1st January 2019 with patent code starting with 'CN'. The second set contains patents matching the query "乳腺癌AND生物标记物" ("breast cancer AND biomarker"), from 1st December 2012 to 1st January 2019 with patent code starting with 'CN' as well. We call them HG and BC to refer to these two datasets in the remainder of the paper. We added a patent code filter based on the International Patent Classification (IPC) system[2] to further improve the relevance of the retrieved patents.[3] After this IPC code filtering, 2,659 patents and 53,007 patents were included in the BC and HG datasets, respectively.

As data cleaning we first removed blank lines and redundant whitespaces, then replaced whitespace inside English and code-mixing parts with the hyphen symbol. These steps were motivated by the typical Chinese language processing style, which is on the character level instead of the word level. The splitting of sentences was then implemented by splitting on newlines.

3.2 Data annotation

After we obtained our cleaned unlabeled patents data, we built a small labeled dataset to be used for training and evaluating the NER models. We randomly selected 21 patents from the total collection of our two unlabeled datasets. Two annotators labeled *gene*, *protein*, and *disease* entities using IOB tagging on the character level (for the Chinese, English and code-mixing parts).

The annotation guideline included the following instructions: First read the whole patent content to understand the meaning of the terms. Then read each sentence separately to understand the contents and contexts of the entities well. This is needed because the same term mention can refer to either the gene or its corresponding protein, depending on the context. No nested or overlapping named entities were allowed and the longest meaningful named entity should be annotated. In the case of a spelling or OCR error, if an entity can be recognized by the annotator even with those errors, we still annotated the entity without correcting it.

For each entity type, we set detailed rules on how to recognize and categorize them:

- The named entity should be annotated as *protein* in these cases: growth factor (e.g. cytokine and other signal proteins); most enzymes, enzyme families, or one category of special enzymes (e.g. DNA polymerase, acetyltransferase) except RNA enzymes (e.g. ribozyme); most antibiotics; protein family (e.g. histone, tubulin); protein expression of a specific gene; antibiotic drug conjugate (e.g. ADC); peptide(s) or amino acid(s); part of a protein structure.

- The named entity should be annotated as *gene* in these cases: a protein coding gene; primer; nucleotide(s), nucleotide analogs (e.g. adenine, 5-propanepyrimidine); ribozyme; gene probes, DNA microarrays and other gene products; gene family (e.g. DNA damage repair genes); expression vector: plasmid, specific ones constructed and named by the patent holder.

- The named entity should be annotated as *disease* in these cases: symptoms (e.g. headache, stomachache); disease names (e.g. B-cell lymphoma, breast cancer); a disease with some specific resistance; early or advanced period of disease; tumor or cancer molecular subclasses (e.g. 三阴型/triple negative breast cancer, 胃样癌/gastric-like).

Five of the 21 patents were labelled by two annotators for the purpose of measuring the inter-rater reliability.

Statistics of the resulting datasets are shown in Table 1. The annotation information of the labeled set is shown in Table 2. The row names indicate whether the value in table shows the total number of items or the unique count. The column names indicate each single category, while 'all' indicate all 3 types

[2] `https://www.wipo.int/classifications/ipc/en/`

[3] We kept patents with IPC code: A61, C07, C12N, C12Q, G01N, G16B, G16C, G16H in dataset BC; and patents with the following IPC codes: A61, C07, C12N, C12Q, C12Y, C12P, C12M, G01N, G16B, G16C, G16H in dataset HG; any patent with IPC code: A01, A21, A23, B09, C02, C09, C10, C11, C05, H01, H04, Y02 were discarded from both datasets

dataset name	n_docs	n_sents	n_chars	sents_per_patent	chars_per_sent
BC (Unlabled)	2,659	1.08M	161M	405	150
HG (Unlabled)	53,007	21.75M	2.84B	410	130
Gold Standard	21	5,813	0.78M	277	134

Table 1: Statistics of our data sets. BC is the "breast cancer AND biomarker" data set and HG is the "human AND gene" data set

<table>
<tr><th colspan="6">Textual information</th></tr>
<tr><th>source</th><th>n_docs</th><th>n_sents</th><th>n_chars</th><th>sents_per_patent</th><th>chars_per_sent</th></tr>
<tr><td>From BC</td><td>12</td><td>3361</td><td>479837</td><td>280</td><td>143</td></tr>
<tr><td>From HG</td><td>9</td><td>2099</td><td>296832</td><td>233</td><td>141</td></tr>
<tr><th colspan="6">Annotation statistics</th></tr>
<tr><th>source</th><th></th><th>gene</th><th>protein</th><th>disease</th><th>all</th></tr>
<tr><td rowspan="2">From BC</td><td>total</td><td>1203</td><td>2917</td><td>2534</td><td>6654</td></tr>
<tr><td>unique</td><td>199</td><td>695</td><td>698</td><td>1592</td></tr>
<tr><td rowspan="2">From HG</td><td>total</td><td>677</td><td>2096</td><td>202</td><td>2975</td></tr>
<tr><td>unique</td><td>271</td><td>338</td><td>54</td><td>663</td></tr>
<tr><td rowspan="2">Total</td><td>total</td><td>1888</td><td>5030</td><td>2739</td><td>9657</td></tr>
<tr><td>unique</td><td>482</td><td>1053</td><td>732</td><td>2267</td></tr>
</table>

Table 2: Annotation statistics and sources of our labeled data set

of annotations. Since the Gold Standard dataset was derived by randomly selecting patents from both BC and HG datasets, the composition of different patent source (from BC and HG datasets) in the Gold Standard dataset was shown in Table 2.

3.3 Models and Learning methods

Our methods are based on pre-trained BERT models (see Section 4.2). We designed and implemented three different learning methods to train our NER models using the labeled data (see also Figure 1):

- **Supervised original**: fine-tuning all weights (BERT model layers plus NER layer) using a relatively small learning rate ($5 * 10^{-5}$), with our labeled dataset;

- **LM mixed fine-tuning**: first tune the weights of the BERT language model layers with the unlabeled dataset; then repeat the supervised original learning step;

- **PartBERT+CRF fine-tuning**: fine-tune the weights of part of the BERT model (last 4 layers) plus an added CRF layer, trained with our labeled dataset.

We train our methods on existing infrastructures: the FlairNLP[4] package (Akbik et al., 2018), the BERT pre-trained Chinese model in PyTorch[5] and the huggingface BERT implementation[6]. We used the Bert-base-Chinese language model, which was trained on the whole Chinese Wikipedia (25M sentences) in both simplified and traditional Chinese.[7] With this pre-trained model, we can continue fine-tuning with our domain-specific data to make the language more suitable for our task.

For efficiency of the training process we create two smaller subsets out of our two large unlabeled datasets to fine-tune the BERT model: 'partBC' and 'partHG'. Both sets contain a train and test set to train the language models. Detailed information of these 2 datasets are shown in Table 3.

[4] https://github.com/flairNLP/flair

[5] https://github.com/Kyubyong/bert_ner

[6] https://github.com/huggingface/transformers

[7] https://github.com/google-research/bert/issues/155

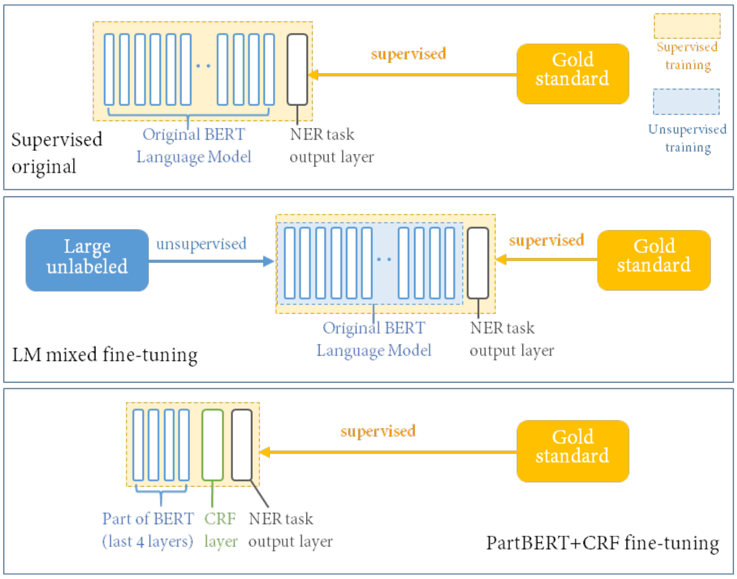


Figure 1: Overview of the three different learning methods that we compare.

dataset name	n_docs	n_sents	n_chars	sents_per_patent	chars_per_sent
partBC_train	100	41,016	6.13M	410	149
partBC_test	10	2,060	0.35M	206	172
partHG_train	10,000	4.18M	0.54B	418	130
partHG_test	100	36,097	5.15M	361	143

Table 3: Statistics of the data sets used for fine-tuning the BERT language model.

3.4 Evaluation

Data Quality For most annotation tasks, Cohen's Kappa is considered the standard metric to compute inter-rater agreement. However, it has been shown that for NER tasks, Kappa does not seem to be the best measure (Grouin et al., 2011). This is because Kappa needs the number of negative cases, which is unknown for named entities. For that reason we calculated the mutual F_1 score of the 5 patents that were annotated by both annotators. The mutual F_1 score is computed as follows. Let A be the set of entities labelled by annotator 1 and B the set of entities labelled by annotator 2. We can then define precision as $P = \frac{|A \cap B|}{|A|}$ and recall as $R = \frac{|A \cap B|}{|B|}$ (or vice versa) and F_1 as the harmonic mean of precision and recall: $F_1 = 2 \times \frac{P \times R}{P + R}$

Model Quality We make the train-test split on the document level, which means that one patent can only appear in either the train or the test set and cannot appear in both. Our labeled set only contains 21 Chinese biomedical patents. Although it has more then 5.8k sentences and 2.2k unique named entity appearance in total, it is still a relatively small dataset. We therefore used cross validation with the labeled data set for evaluation of our models. We make five different splits of the labeled dataset; in each split we use 18 patents as training set, 2 as test set and 1 as held-out development set in the case that we would ned to do hyperparameter tuning.

We evaluate the NER models using precision, recall and F_1 scores for each category of named entities and we report Micro and Macro averages over the named entity categories. Afterwards, the average and standard deviation among the 5 cross validation sets are calculated to monitor the stability of our implemented models and learning methods.

3.5 Post-analysis of the large corpus

Based on the evaluation results, we select the best learning setting and train a model on the full collection of our labeled dataset. The result is our final model. This model is then applied to large unlabeled datasets to extract biomedical entities from patent texts. We finally generated biomedical entity predictions from all 2,659 patents in the BC dataset and 10,100 patents from the HG dataset.

In the BERT output it is possible that there are some meaningless named entities or the ‘IOB’ tags are not generated in the correct order: a sequence of ‘I’ tags should always follow one ‘B’ tag and the type of ‘I’ tags cannot be mixed. The statistic results of the predictions generated on the whole BC dataset show that among all 2.26M continuous non-‘O’ sequences, 5% (0.12M) are single characters, 5% (0.11M) are sequences starting with tag ‘I’, 3% (58K) are sequences containing inconsistent category tags (for example, the first tag is ‘B-protein’ and the second comes an ‘I-gene’), and 4% (98k) sequences contain meaningless symbols in text regarding gene/protein/disease entities (such as ‘?’, ‘˜’ and the Chinese full stop mark ‘ 。’). Thus, before we start some biomedical analysis, some post-processing cleaning steps are needed. We apply the following post-processing steps:

1. If one non-‘O’ sequence only contains one single character, discard it since we assume a single character named entity of either gene, protein or disease category will be meaningless;[8]
2. Find any non-‘O’ sequence that starts with a B followed by only I-tags, and the sequence text should not contain invalid symbols;
3. After step 2, if the tags belongs to the same category (gene, protein or disease), store it; else, only store the first part with ‘BI...’ tags that do belong to the same category; discard other parts.

With these cleaned named entities, we first generate basic statistics of all predictions, then we build the gene–gene connection network based on co-occurrences. We assume that gene entities that frequently co-occur are semantically connected. We use a context window of two sentences to determine co-occurrence. The node weights and edge weights represent the number of patent documents which mention that node or edge. To get better visualization performance we set a threshold on edge weights; any edge with a weight smaller than the threshold will not be included in the final network that is visualized. The edge weight threshold will be defined in Section 4 when we describe and explain the post analysis results.

4 Experiments and Results

4.1 Data quality

4 For the inter-rater agreement, we obtained a 0.95 average F_1 for the five patents that were labelled by both annotators, and 0.98, 0.91 and 0.97 F_1 scores for gene, protein and disease type entities, respectively. This indicates that the annotations are sufficiently reliable.

4.2 Model quality

We ran our training experiments with the three learning methods. For the final NER layer, each model was trained for 40 epochs on the train set, then evaluated on the test sets as described in Section 3.4.

The results are in Table 4. Here we show the average F_1 score and standard deviation of each model for the five cross validation test sets. The row names ‘Supervised Original’, ‘BERT LM mixed’, and ‘PartBERT+CRF’ represent our three learning methods described in Section 3.3.

Table 4 shows that the BERT LM mixed model which was trained on the partHG dataset for only 1 epoch obtained the best results in all categories. It is suggested in the original BERT paper (Devlin et al., 2019) that more training epochs can lead to better performance. Our results do not confirm that. We do see a large standard deviation in the quality of the models for the different test folds, indicating that the model stability is relatively low. More labelled training data could be needed to prevent this.

Another point is that these categories clearly result in lower scores than the clinical categories addressed in related work (Section 2). However, we do see all models achieve higher performance on recognizing ‘disease’ entities than the other categories. This was probably mainly because most disease names in Chinese biomedical patents dataset were written in Chinese without any code-mixing, and the

[8] Almost all Chinese disease names end in a generic character, e.g. “病”(disease), “癌”(cancer), “炎”(inflammation) or words “感染”(infection), “结石”(stones), “综合症”(syndrome) and etc. This means that any Chinese disease name must contain more than one character. For example, “血友病”(hemophilia), “食道癌”(esophageal cancer), “涎腺导管结石”(salivary duct stones).

average F_1 score among 5 datasets	gene	protein	disease	macro avg	micro avg
Supervised Original	**0.31±0.21**	0.34±0.11	0.60±0.09	0.42±0.09	0.49±0.16
BERT LM mixed (partBC_1epoch)	0.21±0.18	0.33±0.23	0.67±0.16	0.40±0.06	0.51±0.15
BERT LM mixed (partBC_30epochs)	0.26±0.19	**0.36±0.24**	0.67±0.12	**0.43±0.06**	0.52±0.15
BERT LM mixed (partHG_1epoch)	0.27±0.21	0.33±0.22	**0.70±0.12**	**0.43±0.06**	**0.54±0.15**
PartBERT+CRF	0.21±0.08	0.25±0.20	0.67±0.12	0.38±0.03	0.47±0.11

Table 4: NER experiments results (averaged over categories, excluding the 'O' labels). The 'partBC' and 'partHG' indicates which unlabeled dataset the BERT language model was trained on, while the '1epoch' and '30epochs' denotes the number of epochs the language model was trained for. The best results of each category has been marked with boldface in the table.

	gene	protein	disease	all
total	410,523	933,106	548,871	1,892,500
unique	70,026	129,791	45,047	244,864
top10	HER2	单克隆抗体(Monoclonal antibodies)	乳腺癌(Breast cancer)	乳腺癌
	VEGFR2	半胱氨酸(Cysteine)	肺癌(Lung cancer)	肺癌
	EGFR	抗体片段(Antibody fragment)	前列腺癌(Prostate cancer)	前列腺癌
	VEGFA	EGFR	卵巢癌(Ovarian cancer)	单克隆抗体
	KRAS	贝伐单抗(Bevacizumab)	胰腺癌(Pancreatic cancer)	卵巢癌
	CDR3	双特异性抗体(Bispecific antibody)	胃癌(Gastric cancer)	胰腺癌
	c-MAF基因(c-MAF gene)	HER2	肝癌(Liver cancer)	胃癌
	PLGF	轻链可变区(Light chain variable region)	结肠癌(Colon cancer)	半胱氨酸
	CDR2	重链可变区(Heavy chain variable region)	膀胱癌(Bladder Cancer)	肝癌
	FGFR3	VEGF	白血病(leukemia)	结肠癌

Table 5: Statistics on the named entities extracted by our model from the large BC data set, with the top-10 most frequently occurring entities for each category.

original pre-trained BERT model we applied was a Chinese language model trained on the general domain Wikipedia dataset. It seems reasonable that our trained model can solve recognizing pure Chinese named entities better than code-mixing cases.

For example, our model annotated the string '任一项的含有T' ('any item containing T' in English) as a protein (false positive). The possible reason is that this string has a similar structure to a true protein entity, such as 'ro-i的单克隆抗体' ('The monoclonal antibodies of ro-i' in English), which contains English character(s) in the beginning or end, and the rest parts are all Chinese characters. Besides, some non-recognized entities (false negatives) show that our model seems to have difficulties dealing with long expressions, which is common in Chinese biomedical patent texts and even in protein and gene names. For example, our model failed to detect the protein entity from the string 'E6与E7癌蛋白', '针对plgR蛋白的抗体', and '长链脂肪酸辅酶A连接酶亚基' ('E6 and E7 oncoprotein', 'Antibodies against plgR protein', and 'Long-chain fatty acid coenzyme A ligase subunit' in English). The phenomenon is probably also because the model 'got confused' about the code-mixing expressions in Chinese biomedical patent texts. It is hard for the model to distinguish which code-mixing part indicates a protein or gene entity, and which part is just a common context, especially in a long sentence or expression.

4.3 Biomedical analysis of the large corpus

Most frequent entities in the data. We show the statistics of named entities extracted the large BC dataset in Table 5 (predictions for HG gave similar results but not shown here because of limited space).

The table indicates that, although we obtained most predictions for protein entities, the top 10 common protein mentions are actually the least meaningful compared to the other entity types. It is interesting that with these limited training data and far from perfect NER classification performance, the detected gene and disease entities are highly meaningful and the most common ones also seem very reasonable. Besides, we can notice that our model can successfully recognize code-mixing entities, such as the c-MAF基因(c-MAF gene) in top 10 common gene mentions shown in Table 5. This indicates that our model and solution at least partly solved the code-mixing problem of Chinese biomedical patents.

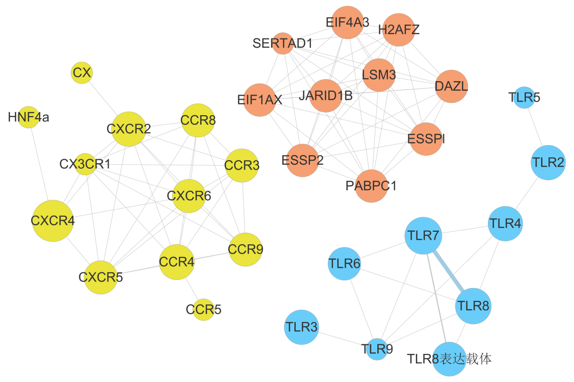
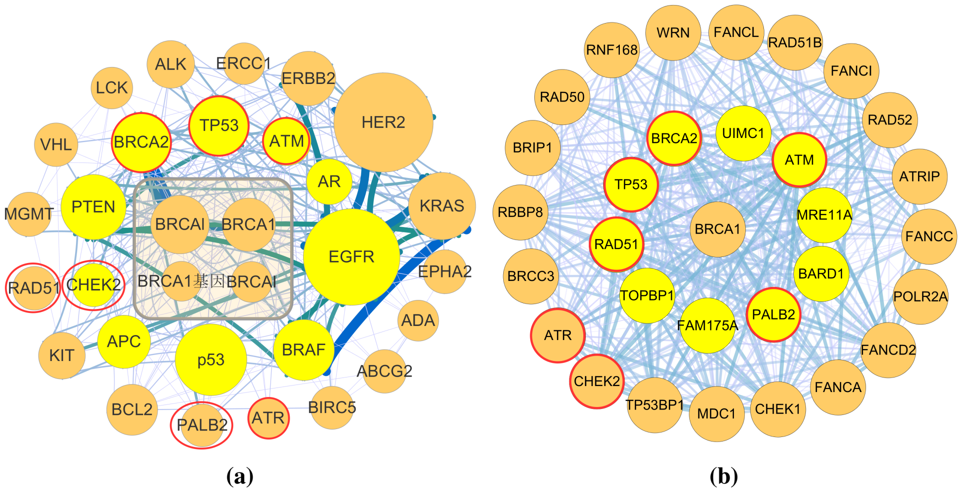


Figure 2: Part of the gene–gene connection network for the HG dataset. The colors indicate gene clusters.



(a)

(b)

Figure 3: Comparison of the BRCA1 gene network extracted from our data to the network based on an existing database. (a) is the BRCA1 gene network generated by predictions of our BC dataset, yellow colored nodes are connected with BRCA1 with top 10 edge weights; (b) is the BRCA1 gene network generated from the STRING database (Szklarczyk et al., 2016), yellow colored nodes are the top 10 common neighbours of BRCA1 in STRING. Nodes enclosed in red circles are the matching nodes between the two networks.

Relations between entities. We calculated and generated the gene–gene connection network from co-occurrences of gene entity predictions on both datasets. In Figure 2 part of the gene network for the HG dataset is shown, with an edge weight threshold of 2. This figure shows 3 different gene clusters found by our method, which are meaningful and reasonable in the sense that all genes in the same cluster indeed have real biological interactions or close relations (e.g. from the same gene family). There are also a lot other similar clusters in both whole version networks, which indicates that our predictions and co-occurrence network mining method can indeed find relatively reliable and meaningful connections among genes.

We also calculated a small cluster network only focusing on interactions with the specific gene ‘BRCA1’ in the BC dataset and its all first-degree neighbours. We compare the resulting network with the STRING database (Szklarczyk et al., 2016).[9] We calculated the ‘BRCA1’ gene cluster from BC dataset since ‘BRCA1’ is a breast cancer related biomarker and BC dataset is also a breast cancer related patents dataset. The visualization of our predicted BRCA1 network (edge weight threshold: 3), plus a BRCA1 network derived from STRING database with text mining edge sources (top 30 common first degree neighbours), are shown in Figure 3a and 3b.

As shown in the figure, the nodes connected with BRCA1 with top 10 edge weights in our predicted network match 4 genes (nodes) of the top 10 common neighbours of BRCA1 in the STRING network,

[9] https://string-db.org/

and 6 nodes in our predicted network match nodes in the STRING BRCA1 network. Our predictions contain novel gene nodes and edges compared to the relations included in STRING. Since we already only retrieved the edges with text mining sources from the STRING database, this indicates that the novel nodes and edges appeared in our predicted network are potentially new discoveries or some discoveries patented in China but not in Europe.[10] However, better performed NER classifiers and more network comparisons are still needed to prove these assumptions.

5 Discussion

If we compare our quantitative results to the results reported by Li et al. (2020) (see Section 2), it appears that NER in patents is a more difficult task than NER in clinical data. We have identified a number of directions for future work that can be done for further improvement:

First is that we may need more pre-processing steps focusing on the specialty of Chinese biomedical text. A vital problem of processing any Chinese text is that, if the source was PDF files or text converted by PDF files using optical character recognition (OCR), then, because of the large variety of Chinese characters and their complex shapes, the OCRed text will have relatively low quality compared with OCRed English text. In the general domain, there are several well developed tools that can detect and correct possible OCR errors automatically. But for domain-specific text, popular existing tools based on simple rules can not handle OCR errors well, especially if there is complex and difficult terminology use (D'hondt et al., 2016; Thompson et al., 2015). Facing this problem, it might be necessary to build an OCR correction tool adapted for this domain and genre specifically (Zhang et al., 2019).

Another challenge of our task is that, similar to other Chinese biomedical text, Chinese biomedical patents will have code-mixing or code-switching fragments, mainly because the protein and gene names are commonly written in English (or the English names been noted after the Chinese one), while the disease names and other contents will be written in Chinese. Moreover, even just inside each single named entity, it is possible that the code-mixed expression still appears. There have been some attempts trying to solve Chinese-English code-mixing problems in the general domain, and more commonly with speech data since this code-mixing usage seems more possible to happen in non-official situations, for example casual conversations or social media posts (Winata et al., 2018; Shen et al., 2011). Thus, it will be interesting to try some code-mixing language model from the general domain to check whether it can improve current biomedical NER performance.

In the final part of our study, we did some biological post analysis in order to mine some meaningful information from our generated predictions. There are other interesting topics possible in mining meaningful biological information from text mining resources, which we can try in the future. Our gene–gene connection network is based on co-occurrence; it would be interesting to follow-up with more advanced relation extraction methods to discover connections between entities.

6 Conclusions

In this study, we addressed Chinese biomedical NER for patents, which is a highly specific genre with English-Chinese code-mixing and a complex text writing style. We developed our own Chinese Biomedical patents dataset, including a humanly labeled dataset which contains 5,813 sentences and 2,267 unique named entities from 21 patent documents, and two large unlabeled datasets which contain 2,659 and 53,007 patent documents respectively.

After we implemented three different BERT models and learning methods, we trained and evaluated them on our evaluation sets. The results showed that the BERT LM mixed model, which was trained on a part of the human gene (HG) dataset containing 10,000 patents in the train set and 100 in the test set for only 1 epoch, obtained the optimal results over all entity categories. But there were only small differences between the models and training methods. The best model obtained a 0.54±0.15 micro average F_1 score among all entity types (among all evaluation sets). Although we consider these results to be reasonable,

[10] This is because STRING is part of ELIXIR Core Data Resources, which are a set of European data resources: https://elixir-europe.org/platforms/data/core-data-resources

one weakness is that the standard deviation between the different test folds is large. A larger training data set might be needed to make the model more stable.

We finally generated predictions with the trained best model (trained on the whole labeled dataset) on the large unlabeled datasets and then did some further biomedical analyses. These analyses indicate that our model can detect meaningful biomedical entities and find some novel gene–gene interactions, just with limited labeled data, training time and computing power.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using bert bilstm crf for chinese electronic health records. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Li Du. 2018. Patenting human genes: Chinese academic articles' portrayal of gene patents. *BMC medical ethics*, 19(1):29.

Eva D'hondt, Cyril Grouin, and Brigitte Grau. 2016. Low-resource OCR error detection and correction in French Clinical Texts. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 61–68.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th linguistic annotation workshop*, pages 92–100. Association for Computational Linguistics.

Baohua Gu, Fred Popowich, and Veronica Dahl. 2008. Recognizing biomedical named entities in Chinese research abstracts. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 114–125. Springer.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Citeseer.

Xiangyang Li, Huan Zhang, and Xiao-Hua Zhou. 2020. Chinese clinical named entity recognition with variant neural structures based on bert methods. *Journal of Biomedical Informatics*, page 103422.

Han-Ping Shen, Chung-Hsien Wu, Yan-Ting Yang, and Chun-Shan Hsu. 2011. CECOS: A chinese-english code-switching speech database. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 120–123. IEEE.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of BioCreative II gene mention recognition. *Genome biology*, 9(2):S2.

Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. 2016. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937.

Paul Thompson, John McNaught, and Sophia Ananiadou. 2015. Customised OCR correction for historical medical text. In *2015 Digital Heritage*, volume 1, pages 35–42. IEEE.

S Verberne, EKL D’hondt, NHJ Oostdijk, and CHA Koster. 2010. Quantifying the challenges in parsing patent claims. In *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval at ECIR 2010*, pages 14–21.

Xuwen Wang, Jiao Li, Yingjie Wu, and Junlian Li. 2019. BiLSTM-CRF based open concept relation extraction from Chinese biomedical texts. *Chinese Journal of Medical Library and Information Scinence*, 27(11):33–39.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Code-switching language modeling using syntax-aware multi-task learning. *arXiv preprint arXiv:1805.12070*.

Kui Xue, Yangming Zhou, Zhiyuan Ma, Tong Ruan, Huanhuan Zhang, and Ping He. 2019. Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text. *arXiv preprint arXiv:1908.07721*.

Yue Zhang and Jie Yang. 2018. Chinese NER Using Lattice LSTM. pages 1554–1564.

Congyue Zhang, Ziming YIN, Dayun SUN, and Wei DAI. 2019. Recognition technology of the laboratory sheet based on tesseract. *Beijing Biomedical Engineering*, (3):11.