

RESEARCH

Open Access



# Improving deep learning method for biomedical named entity recognition by using entity definition information

Ying Xiong<sup>1,2</sup>, Shuai Chen<sup>1</sup>, Buzhou Tang<sup>1,2\*</sup> , Qingcai Chen<sup>1,2</sup>, Xiaolong Wang<sup>1</sup>, Jun Yan<sup>3</sup> and Yi Zhou<sup>4\*</sup>

From The 5th workshop on BioNLP Open Shared Tasks Hong Kong, China. 4 November 2019

\*Correspondence:

tangbuzhou@gmail.com;  
zhouyi@sysu.edu.cn

<sup>1</sup> Department of Computer Science, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, China

<sup>4</sup> Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou 510080, China  
Full list of author information is available at the end of the article

## Abstract

**Background:** Biomedical named entity recognition (NER) is a fundamental task of biomedical text mining that finds the boundaries of entity mentions in biomedical text and determines their entity type. To accelerate the development of biomedical NER techniques in Spanish, the PharmaCoNER organizers launched a competition to recognize pharmacological substances, compounds, and proteins. Biomedical NER is usually recognized as a sequence labeling task, and almost all state-of-the-art sequence labeling methods ignore the meaning of different entity types. In this paper, we investigate some methods to introduce the meaning of entity types in deep learning methods for biomedical NER and apply them to the PharmaCoNER 2019 challenge. The meaning of each entity type is represented by its definition information.

**Material and method:** We investigate how to use entity definition information in the following two methods: (1) Squad-style machine reading comprehension (MRC) methods that treat entity definition information as query and biomedical text as context and predict answer spans as entities. (2) Span-level one-pass (SOne) methods that predict entity spans of one type by one type and introduce entity type meaning, which is represented by entity definition information. All models are trained and tested on the PharmaCoNER 2019 corpus, and their performance is evaluated by strict micro-average precision, recall, and F1-score.

**Results:** Entity definition information brings improvements to both Squad-style MRC and SOne methods by about 0.003 in micro-averaged F1-score. The Squad-style MRC model using entity definition information as query achieves the best performance with a micro-averaged precision of 0.9225, a recall of 0.9050, and an F1-score of 0.9137, respectively. It outperforms the best model of the PharmaCoNER 2019 challenge by 0.0032 in F1-score. Compared with the state-of-the-art model without using manually-crafted features, our model obtains a 1% improvement in F1-score, which is significant. These results indicate that entity definition information is useful for deep learning methods on biomedical NER.



**Conclusion:** Our entity definition information enhanced models achieve the state-of-the-art micro-average F1 score of 0.9137, which implies that entity definition information has a positive impact on biomedical NER detection. In the future, we will explore more entity definition information from knowledge graph.

**Keywords:** Biomedical named entity recognition, Entity definition information, Machine reading comprehension, Span-level one-pass method

### Background

Biomedical named entity recognition (NER) is a fundamental task of biomedical text mining to identify biomedical entity mentions of different types in biomedical text. Most biomedical NER studies focus on the biomedical text in English. To accelerate the development of Spanish biomedical NER techniques, Martin Krallinger et al. organized a specific challenge for chemical & drug mention recognition in Spanish biomedical text, called PharmaCoNER, in 2019 [1]. Participants were required to recognize the entities in Spanish biomedical text, as shown in Fig. 1.

Biomedical NER is a typical sequence labeling problem, and lots of state-of-the-art methods have been proposed for this problem, such as BiLSTM-CRF [2]. Almost all these methods do not consider the meaning of different entity types, which may benefit biomedical NER. The meaning of each entity type can be represented by its definition. For example, the definition of PROTEINAS in the guideline of PharmaCoNER 2019 is: “Las menciones de proteínas y genes incluyen péptidos, hormonas peptídicas y anticuerpos.” (Protein and gene mentions include peptides, peptide hormones, and antibodies). In this paper, we explore how to encode entity definition information in two kinds of deep learning methods for NER. They are: (1) Squad-style MRC methods designed to find a continuous span of entity mentions in given text for each type. We use each type’s entity definition as a query instead of a naive query generated by simple rules in MRC methods. For convenience, we adopt MRC to represent Squad-style MRC in the following sections in this paper. (2) Span-level one-pass (SOne) methods that predict entity spans of one type by one type. We use entity definition information to represent each entity type’s meaning and introduce the entity type meaning into SOne. The definition information of each type includes the original definition of each type in the guideline and entity mentions in the text. We compare them in the SOne model.

In order to evaluate the performances of MRC and SOne, we conduct experiments on the PharmaCoNER 2019 corpus. Experiments show that the entity definition information brings improvements to both MRC and SOne methods. The improvement in micro-averaged F1-score is about 0.003. The MRC method using entity definition information

...

a) Nutrición Parenteral: la **NPT** se sustituye por 2 litros de **Isoplasmar G\*** con adición de una ampolla de 10 mEq de **potasio**/litro y el módulo de **vitaminas (Addamei®)** a días alternos con el módulo de **oligoelementos (Cernevit\*)**.

...

e) Medicación: para disminuir la secreción clorhidropéptica y el estímulo a la secreción de agua y **bicarbonato** pancreático se asocia **omeprazol** (40 mg/día) y para facilitar la digestibilidad de los alimentos naturales y su disponibilidad absorbiva se asocian 2 cápsulas de **enzimas pancreáticas (Kreón 10.000\*)** con la toma de alimentos.

**Fig. 1** Examples of the biomedical named entities in Spanish records. (NORMALIZABLE entities in green, PROTEINAS entities in blue, NO\_NORMALIZALLE entities in yellow and UNCLEAR entities in red. Notice that UNCLEAR entities are not included in the final evaluation.)

as query achieves the best performance with a micro-average precision of 0.9225, a recall of 0.9050, and an F1-score of 0.9137, respectively. It outperforms the best model of the PharmaCoNER 2019 challenge by 0.0032 in micro-averaged F1-score.

### Related work

The natural language processing (NLP) community has made a great contribution to the development of NER in the biomedical text through challenges, such as I2B2 (Informatics for Integrating Biology and the Bedside) [3, 4], BioCreative (Critical Assessment of Information Extraction systems in Biology) [5, 6], SemEval (Semantic Evaluation) [7, 8], CCKS (China Conference on Knowledge Graph and Semantic Computing) [9, 10] and IberLEF [11]. A large number of methods have been proposed for biomedical NER. Most of them can be classified into the following three categories: (1) Rule-based methods that extract named entities using specific rules design by experts. The earlier clinical NLP tools are rule-based systems relying on clinical dictionaries, such as MedLEE [12], KnowledgeMap [13] and MetaMap [14]. (2) Supervised machine learning methods with hand-crafted features Maximum Entropy (ME) [15, 16], Support Vector Machines (SVM) [17], CRF [18, 19], Hidden Markov Models (HMM) [20, 21] and Structural Support Vector Machines (SSVM) [22]. They usually treat NER as a sequence labeling task, which tags a sentence with a label sequence. The common features used in the supervised machine learning methods include orthographic information (e.g. capitalization, prefix, suffix and word-shape), syntactic information (e.g., POS tags), dictionary information, n-gram information, disclosure information (e.g. section information in EHRs) and some features generated from unsupervised learning methods [23]. (3) Deep learning methods that can learn features from large unlabeled data without costly feature engineering. Convolutional Neural network (CNN) [24], Recurrent Neural Network (RNN) [25] and Long Short Term Memory neural network (LSTM) [2] have been widely used for biomedical NER and show good performance. Besides the methods mentioned above, there are also some other attempts. For example, to tackle the low-resource problem in the biomedical domain, researchers introduce multi-task learning methods to learn more abundant information from other tasks, such as NER from other sources, chunking, and POS tagging [26–28], and deploy transfer learning methods to first learn knowledge from related sources and then finetune on target [29–33].

Nowadays, there is an upward trend in defining NLP tasks in the MRC framework. MRC models [34–36] extract answer spans from the context given a pre-defined question. Generally, SQuad-style MRC models can be formalized as predicting the start position and the end position of the answer. Li et al. [37] treat the entity-relation extraction task as a multi-turn question answering and propose a unified MRC framework to recognize entities and extract relationships. Li et al. [38] propose an MRC method to recognize both flat and nested entities.

### Material and methods

#### Datasets

In this study, all experiments are conducted on the PharmaCoNER 2019 corpus annotated by medicinal chemistry experts according to a pre-defined guideline. The corpus contains 1000 clinical records with 24,654 chemical & drug mentions. The corpus is

divided into a training set of 500 records, a development set of 250 records and a test set of 250 records, where the test set is hidden in a background set of 3751 records during the test stage of the competition. In experiments, we first split each record into sentences by sentence ending symbols, including ‘\n’, ‘.’, ‘;’, ‘?’, and ‘!’. About 95% of sentences are no longer than 230 tokens. The corpus statistics, including the number of records, sentences, and chemical & drug mentions of different types, are listed in Table 1. It should be noted that the UNCLEAR mentions are not considered during the competition.

### Task definition

Given a sequence  $X = \{x_1, x_2, \dots, x_n\}$  of length  $n$ , we need to assign a label sequence  $Y = \{y_1, y_2, \dots, y_n\}$  to  $X$ , where  $y_i$  is the possible label of token  $x_i$  ( $1 \leq i \leq n$ ) (e.g., PROTEINAS, NORMALIZABLES, NO\_NORMALIZABLES, UNCLEAR).

**MRC definition:** the sequence labeling problem can be redefined in the MRC framework as follows, For each label type  $y$ , its definition information is regarded as a query  $q^y = \{q_0, q_1, \dots, q_m\}$  of length  $m$ , a sentence  $X$  is regarded as the context of  $q^y$ , the span of an entity of type  $y$ , and  $x_{start:end}^y = \{x_{start}, x_{start+1}, \dots, x_{end-1}, x_{end}\}$ , is recognized as an answer. Then, the original sequence labeling problem can be represented by  $(q^y, X, x_{start:end}^y)$ . The goal of MRC is to find the spans of all entity mentions of all types, given all sentences.

**SOne definition:** SOne takes sequence  $X$  as inputs and predicts the spans of all entities of one type by one type using a multi-layer pointer network [39]. The number of network layers depends on the number of entity types. For each type of entity, we add entity definition information  $e$  to enhance SOne by concatenating it to all tokens.

### Query generation for MRC

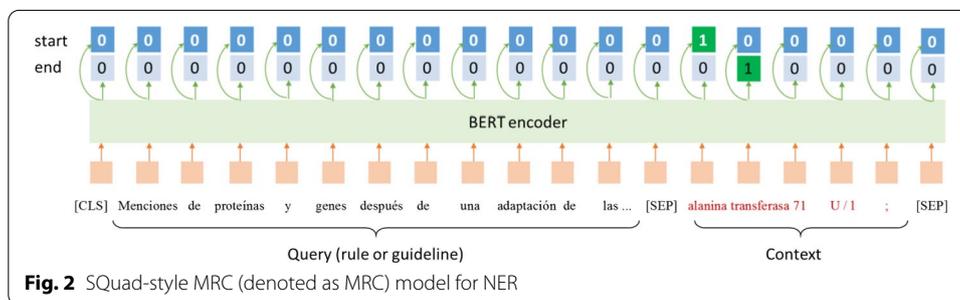
Query generation is critical for MRC, since queries usually contain some prior knowledge (e.g. entity type definition) about tasks. Li et al. [40] introduce various kinds of query generation methods, including keywords, Wikipedia, rule-based template filling, synonyms, keywords combined synonyms and annotation guideline notes, and compare them. The results show that annotation guideline is the best choice for query generation. Following Li et al. [40], we compare two kinds of query generation: annotation guideline and rule-based template filling. Table 2 shows our generated queries for each type of entity.

**Table 1** Statistics of the PharmaCoNER 2019 Corpus

Statistic	#Training	#Development	#Test	#Background
RECORDS	500	250	250	3751
SENTENCES	8776	4028	4260	\
NORMALIZABLES	2304	1121	973	\
NO_NORMALIZABLES	24	16	10	\
PROTEINAS	1405	745	859	\
UNCLEAR	89	44	0	\

**Table 2** Generated queries for each type of entity

Entity type	Query type	Generated query
PROTEINAS	Guideline	<b>“Las menciones de proteínas y genes incluyen péptidos, hormonas peptídicas y anticuerpos.”</b> (Protein and gene mentions include peptides, peptide hormones, and antibodies.)
	Rule-template	<b>¿Qué entidades PROTEINAS se mencionan en el texto?</b> (Which PROTEINAS entities are mentioned in the text?)
NORMALIZABLES	Guideline	<b>“Menciones de productos químicos que pueden normalizarse manualmente a un identificador de concepto único.”</b> (Chemical mentions that can be manually normalized to a unique concept identifier.)
	Rule-template	<b>¿Qué entidades NORMALIZABLES se mencionan en el texto?</b> (Which NORMALIZABLE entities are mentioned in the text?)
NO_NORMALIZABLES	Guideline	<b>“Menciones de productos químicos que no se pudieron normalizar manualmente a un identificador de concepto único.”</b> (Chemical mentions that could not be manually normalized to a unique concept identifier.)
	Rule-template	<b>¿Qué entidades No_NORMALIZABLES se mencionan en el texto?</b> (Which NON-NORMALIZABLE entities are mentioned in the text?)
UNCLEAR	Guideline	<b>“Casos de menciones generales de clase de sustancias de relevancia clínica y biomédica, incluidas ciertas formulaciones farmacéuticas, tratamientos generales, programas de quimioterapia, vacunas y un conjunto predefinido de sustancias generales ( por ejemplo: Estragón, Silimarina, Bromelaina, Melanina, Vaselina, Lanolina, Alcohol, Tabaco, Marihuana, cannabis, opio y gluten).”</b> (Cases of general mentions of class of substances of clinical and biomedical relevance, including certain pharmaceutical formulations, general treatments, chemotherapy programs, vaccines and a predefined set of general substances (for example: Tarragon, Silymarin, Bromelain, Melanin, Vaseline, Lanolin, Alcohol, Tobacco, Marijuana, cannabis, opium and gluten))
	Rule-template	<b>¿Qué entidades UNCLEAR se mencionan en el texto?</b> (Which UNCLEAR entities are mentioned in the text?)



**Model detail**

In this study, We utilize BERT (Bidirectional Encoder Representations from Transformers) [41] as our model backbone. Figure 2 shows the skeleton of the MRC model. Given query  $q^y$  and sentence  $X$ , we need to predict the span of every entity of type  $y$ , including a start position  $x_{start}^y$  and an end position  $x_{end}^y$ . The model first takes the following input and encodes it by BERT:

$$input_{MRC} = \{ [CLS], q^y, [SEP], X, [SEP] \}, \tag{1}$$

where [CLS] and [SEP] are special tokens of BERT, denoting whole sentence and sentence separator, respectively. Suppose that the last layer output of BERT is  $H \in \mathbb{R}^{s \times d}$ , where  $s$  is the total length of [CLS],  $q^y$ , [SEP],  $X$  and [SEP], and  $d$  is the dimension of the last layer output of BERT, the model then predicts the possibilities of start position and end position as follows:

$$P_{start} = \text{softmax}(H \cdot W_{start} + b_{start}) \in \mathbb{R}^{m \times 2}, \tag{2}$$

$$P_{end} = \text{softmax}(H \cdot W_{end} + b_{end}) \in \mathbb{R}^{m \times 2}, \tag{3}$$

where  $W_{start}$  and  $W_{end}$  are trainable parameters,  $b_{start}$  and  $b_{end}$  are biases.

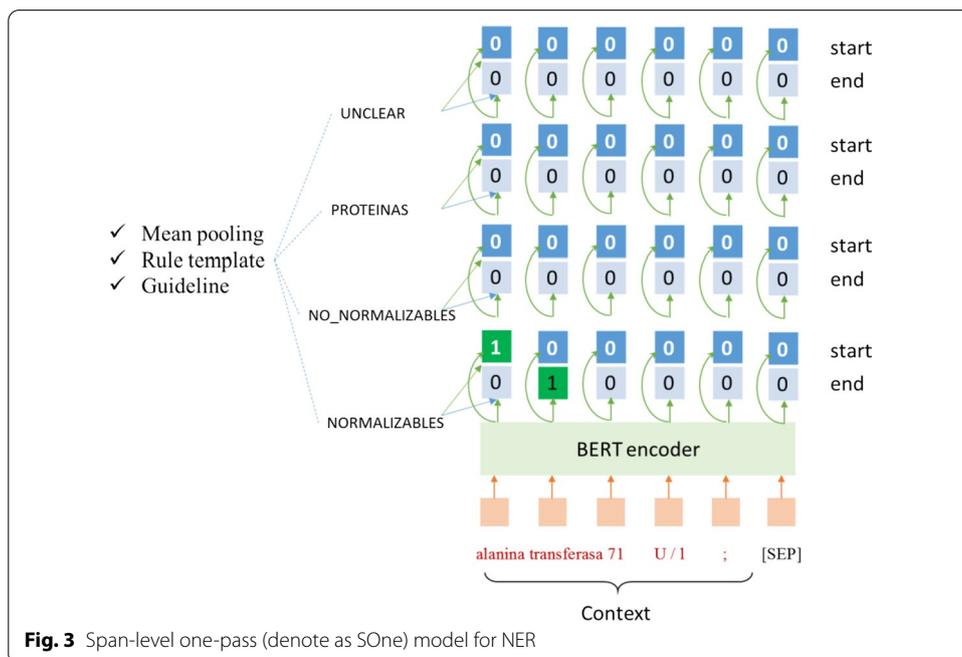
The predicted start index  $I_{start}$  and end index  $I_{end}$  are:

$$I_{start} = \{j | \text{argmax}(P_{start}^j) = 1, j = 1, 2, 3, \dots, m\} \tag{4}$$

$$I_{end} = \{k | \text{argmax}(P_{end}^k) = 1, k = 1, 2, 3, \dots, m\} \tag{5}$$

We use MRC\_rule and MRC\_guideline to denote MRC using rule-based template filling for query generation and MRC using annotation guideline as query, respectively.

Figure 3 shows the skeleton of the SOne model. In this model, we first use BERT to encode the input sentence  $X$  as  $Z \in \mathbb{R}^{n \times d}$  (i.e., the output of the BERT's last layer), and then concatenate the entity definition information representation  $e \in \mathbb{R}^{d_e}$  to all tokens, where  $d_e$  is the dimension of the entity definition information representation. Here, we consider three kinds of entity definition information: (1) entity mentions word embedding. each entity type definition information is represented by the mean pooling of word2vec embeddings



**Fig. 3** Span-level one-pass (denote as SOne) model for NER

of all tokens in all mentions of that type [42] (denoted as SOne\_w2v). (2) Rule-based query. We use BERT to encode each query generated by rules (denoted as SOne\_rule). (3) Annotation guideline encoded by BERT (denoted as SOne\_guideline). The entity definition information enhanced sentence representation is represented as follows:

$$\text{input}_{\text{SOne}} = [Z, E], \quad (6)$$

where  $E \in \mathbb{R}^{n \times d_e}$  is  $n$  copied  $e$ , and  $[]$  denotes the concatenation operation.

Finally, the SOne model makes the same prediction for start position and end position as the MRC model. The only difference is that SOne has four input-shared span predictors with the same structure and different parameters, while MRC has four separate span predictors. The overall objective function of MRC and SOne is:

$$L = L_{\text{start}} + L_{\text{end}}, \quad (7)$$

where  $L_{\text{start}}$  is the start position prediction loss and  $L_{\text{end}}$  is the end position prediction loss.

### Evaluation metrics

The performances of all models are measured by micro-averaged precision (P), recall (R), and F1-score (F1) under the “exact-match” criterion:

$$P = \frac{\#TP}{\#(TP + FP)}, \quad (8)$$

$$R = \frac{\#TP}{\#(TP + FN)}, \quad (9)$$

$$F1 = 2 \times \frac{P \times R}{P + R}, \quad (10)$$

where TP is true positive, FP is false positive, and FN is false negative.

These measures can be calculated by the evaluation tool [43] released by the official organization of the PharmaCoNER 2019 challenge.

### Experiment setting

Following Xiong’s work [44], we first train our models on the training set and development set, and then further finetune the model for 20 epochs. The max sentence lengths of the MRC model and SOne model are set as 250 and 230, respectively. The difference in the max length is due to the query in the MRC model. The learning rate of BERT is set as  $2e-5$ , the batch size of all models is set as 20. The dimension of entity definition information representation  $d_e$  is set as 300. Other parameters are set as the default. The code is available at [45].

## Results and discussion

### Performance evaluation

Table 3 presents the results of our proposed MRC and SOne model (lower part) and summarizes some reported results on the PharmaCoNER Corpus (upper part).

First, the micro-average precision, recall and F1-score of MRC\_rule and MRC\_guideline is 0.915, 0.9055, 0.9109 and 0.9225, 0.9050, 0.9137, respectively. Results show that both MRC\_rule and MRC\_guideline outperform the baseline model SOne by 0.44% and 0.72% in micro-averaged F1-score. The reason why MRC\_guideline performs better than MRC\_rule lies in the expertness of guideline definition. For SOne extended models, all kinds of entity definition information representation can bring improvements to the baseline model SOne. Compared with SOne, the micro-averaged F1-score of SOne\_rule increases to 0.912, SOne\_guideline increases to 0.9128, and SOne\_w2v increases to 0.9094. The overall micro-averaged F1-score improvements of extended SOne models range from 0.29 to 0.63%.

Second, MRC-guideline outperforms all existing systems on the PharmaCoNER corpus, creating new state-of-the-art results and pushing the micro-averaged F1-score of the benchmark to 0.9137, which amounts to 0.32% absolute improvement over the top-1 system of the PharmaCoNER 2019 challenge, developed by us that using lots of features, and 1% absolute improvement over our previous system without using features [44], which is a significant improvement. We perform a significance test by comparing the model without using any feature with our MRC model or SOne model, and the results show that the improvement is significant (t-test < 0.05) [46]. This implies that entity definition information has a positive impact on entity recognition.

Third, Table 4 shows the detailed results of each entity type of MRC\_guideline and SOne\_guideline. Both MRC\_guideline and SOne\_guideline perform best on NORMALIZABLES and worst on NO\_NORMALIZABLES. Though MRC\_guideline outperforms SOne\_guideline in terms of micro-averaged F1-score, it wrongly predicts all NO\_NORMALIZABLES type. The probable reason is that queries of NORMALIZABLES and NO\_NORMALIZABLES are too similar, which may confuse our models. Overall, MRC\_guideline outperforms better than SOne\_guideline on micro-averaged precision but worse on micro-averaged recall. Besides, we analyze all our proposed models and find that the SOne model can recognize the NO\_NORMALIZABLES entities, but the MRC model cannot. It may be because that concatenation of entity definition representation benefits to few samples.

**Table 3** Results on PharmaCoNER Corpus

Models	Features	Precision	Recall	F1-score
Xiong et al. [44]	Yes	0.9123	0.9088	0.9105
Stoeckel et al. [47]	No	0.9079	0.9030	0.9052
Sun et al. (2019) [48]	No	0.9046	0.8806	0.8924
Lange et al. [49]	Yes	0.8895	0.8827	0.8861
Hakala et al. [50]	No	0.8758	0.8719	0.8738
Lahuerta et al. [51]	No	0.9022	0.8366	0.8682
Sohrab et al. [52]	Yes	0.8688	0.8665	0.8676
MRC_rule		0.915	0.9055	0.9109
MRC_guideline		<b>0.9225</b>	0.9050	0.9137*
SOne (w/o entity definition)		0.9158	0.8974	0.9065
SOne_rule		0.9153	0.9088	0.912
SOne_guideline		0.9135	<b>0.9121</b>	0.9128
SOne_w2v		0.9167	0.9023	0.9094

The method with the highest F-score among all methods is highlighted in bold

\* Compared with the model without any feature, this is a significant improvement (t-test < 0.05)

**Table 4** Detailed results of each entity type of MRC\_guideline and SOne\_guideline

Entity type	Model	Precision	Recall	F1 score
NORMALIZABLES	MRC_guideline	<b>0.9428</b>	<b>0.9322</b>	<b>0.9375</b>
	SOne_guideline	0.937	<b>0.9322</b>	0.9346
NO_NORMALIZABLES	MRC_guideline	0.0	0.0	0.0
	SOne_guideline	<b>1.0</b>	<b>0.1</b>	<b>0.1818</b>
PROTEINAS	MRC_guideline	<b>0.8994</b>	0.8847	0.892
	SOne_guideline	0.8874	<b>0.8987</b>	<b>0.893</b>

The methods with the highest F-scores in each entity type are highlighted in bold

### Error analysis

Comparing with previous state-of-the-art models, our model can recognize more named entities due to the domain knowledge embedded in the entity definition information. For example, because of the introduction of the PROTEIN information, our model can recognize “timoglobulina (thymoglobulin)”, “protrombina (prothrombin)” and so on, which are ignored by previous state-of-the-art models. To visualize the effect of the added domain knowledge, we calculate the cosine similarity of some words based on their word2vec embeddings. For example, the similarity of “protrombina” and “proteínas” is more than 0.5 but has a lower similarity with “normalizar” or words in the question of the UNCLEAR type.

Though the MRC\_guideline model outperforms other models, there are also some errors, mainly of the following five kinds. (1) About 20% of errors are due to the predicted entities not included in the gold test set. Although these predicted entities are the ones that have appeared, such as “vimentina (vimentin)”, they are wrong because they are not officially annotated. (2) About 30% of errors are due to that the model omits some entities. (3) About 16% of the errors are because the model predicts the correct entity type, but the boundary is too long. For instance, the correct entity is “anticuerpos anticitoplasma (cytoplasmic antibodies)”, but the model predicts “anticuerpos anticitoplasma de neutrófilo (antineutrophil cytoplasmic antibodies)”, or the correct entity is “hormonas de crecimiento (growth hormones)”, but the model predicts “hormonas de crecimiento y antidiurética (growth hormones and antidiuretics)”. (4) About 20% of errors are because the model predicts the correct entity type, but the boundary is too short. For example, “tinción de auramina” is wrongly predicted as “auramina (auramine)”, “anticuerpos antimembrana basal glomerular (glomerular basement membrane antibodies)” is wrongly predicted as “nticuerpos antimembrana basal (basal membrane antibodies)”, and “(Ig)A-kappa” is wrongly predicted as “Ig”. (5) About 10% of the errors are caused by that the model predicts the wrong entity type, and 70% of them are because that “NO\_NORMALIZABLES” entity type is mistakenly predicted as “NORMALIZABLES”, such as “Viekirax”, “Tobradex” and “Harvoni”.

### Conclusion

This paper proposed two kinds of entity definition information enhanced model, MRC and SOne for biomedical NER. Compared with the previous models, our methods do not require features and achieve state-of-the-art performance with a micro-average

F1-score of 0.9137 on the PharmaCoNER Corpus. It indicates that the introduction of entity definition information is effective. In the future, we are planning to introduce more effective entity category definition information through domain knowledge graphs and to explore more valid methods to add the entity definition information, such as attention mechanism.

#### Abbreviations

NER: Named entity recognition; MRC: Machine reading comprehension; CRF: Conditional random fields; NLP: Natural language processing; I2B2: Informatics for Integrating Biology and the Bedside; BioCreative: Critical Assessment of Information Extraction systems in Biology; SemEval: Semantic Evaluation; CCKS: China Conference on Knowledge Graph and Semantic Computing; EHRs: Electronic health records; ME: Maximum Entropy; SSVM: Structural Support Vector Machines; HMM: Hidden Markov Models; SVM: Support Vector Machines; CNN: Convolutional Neural network; RNN: Recurrent Neural Network; LSTM: Long Short Term Memory neural network.

#### Acknowledgements

We thank all anonymous reviewers for suggesting the update of a draft of this manuscript.

#### About this supplement

This article has been published as part of BMC Bioinformatics Volume 22, Supplement 1 2021: Recent Progresses with BioNLP Open Shared Tasks - Part 2. The full contents of the supplement are available at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-22-supplement-1>.

#### Authors' contributions

The work presented here was carried out in collaboration between all authors. Y.X., S.C. and B.T. designed the methods and experiments. Y.X. and S.C. conducted the experiment. Y.X. analyzed the data and interpreted the results. Y.X. and B.T. wrote the paper. Q.C., X.W., Y.J., and Y.Z. provided detailed edits and critical suggestions. All authors have approved the final manuscript.

#### Funding

This paper is supported in part by grants: National Natural Science Foundations of China (U1813215, 61876052 and 61573118), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), National Natural Science Foundations of Guangdong, China (2019A1515011158), Guangdong Province Covid-19 Pandemic Control Research Fund (2020KZDZX1222), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20180306172232154 and JCYJ20170307150528934) and Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052).

#### Availability of data and materials

The dataset is available on <https://temu.bsc.es/pharmaconer/> [accessed on February 9, 2021].

#### Declarations

##### Ethics approval and consent to participate

This study was approved by the Institutional Review Board of Harbin Institute of Technology (Shenzhen).

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

##### Author details

<sup>1</sup>Department of Computer Science, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, China. <sup>2</sup>Peng Cheng Laboratory, Shenzhen, China. <sup>3</sup>Yidu Cloud (Beijing) Technology Co., Ltd, Beijing, China. <sup>4</sup>Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou 510080, China.

Received: 30 March 2021 Accepted: 4 June 2021

Published: 17 December 2021

#### References

1. Gonzalez-Agirre A, Marimon M, Intxaurreondo A, Rabal O, Villegas M, Krallinger M. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019. p. 1–10. doi:<https://doi.org/10.18653/v1/D19-5701>.
2. Lyu C, Chen B, Ren Y, Ji D. Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinform.* 2017;18:462.

3. Sun W, Rumshisky A, Uzuner Ö. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc.* 2013;20:806–13.
4. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task Track 1. *J Biomed Inform.* 2015;58:S11–9.
5. Smith L, Tanabe LK, nne Ando RJ, Kuo C-J, Chung I-F, Hsu C-N, et al. Overview of BioCreative II gene mention recognition. *Genome Biol.* 2008;9:S2.
6. Hirschman L, Yeh AS, Blaschke C, Valencia A. Overview of BioCreativeE: critical assessment of information extraction for biology. *BMC Bioinform.* 2005;6 S-1.
7. Elhadad N, Pradhan S, Gorman SL, Manandhar S, Chapman WW, Savova GK. SemEval-2015 Task 14: Analysis of Clinical Text. In: Cer DM, Jurgens D, Nakov P, Zesch T, editors. Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4–5, 2015. The Association for Computer Linguistics; 2015. p. 303–10.
8. Bethard S, Savova G, Chen W-T, Derczynski L, Pustejovsky J, Verhagen M. SemEval-2016 Task 12: Clinical TempEval. In: Bethard S, Cer DM, Carpuat M, Jurgens D, Nakov P, Zesch T, editors. Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16–17, 2016. The Association for Computer Linguistics; 2016.
9. Zhang J, Li J, Jiao Z, Yan J. Overview of CCKS 2018 Task 1: Named Entity Recognition in Chinese Electronic Medical Records. In: Zhu X, Qin B, Zhu X, Liu M, Qian L, editors. Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding—4th China Conference, CCKS 2019, Hangzhou, China, August 24–27, 2019, Revised Selected Papers. Springer; 2019. p. 158–164.
10. Hu J, Shi X, Liu Z, Wang X, Chen Q, Tang B. HITSZ\_CNER: a hybrid system for entity recognition from Chinese clinical text. In: CEUR workshop proceedings. 2017.
11. Marimon M, Gonzalez-Agirre A, Intxaurrenada A, Rodriguez H, Lopez Martin JA, Villegas M, et al. Automatic de-identification of medical texts in Spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). vol. TBA, p. TBA. CEUR Workshop Proceedings (CEUR-WS.org), Bilbao, Spain (Sep 2019), TBA. 2019.
12. Friedman C. Towards a comprehensive medical language processing system: methods and issues. In: Proceedings of the AMIA annual fall symposium. American Medical Informatics Association; 1997. p. 595.
13. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard 3rd A. The KnowledgeMap project: development of a concept-based medical school curriculum database. In: AMIA... Annual Symposium proceedings. AMIA Symposium. 2003. p. 195.
14. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc.* 2010;17:229–36.
15. Saha SK, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition. *J Biomed Inform.* 2009;42:905–11.
16. Lin Y-F, Tsai T-H, Chou W-C, Wu K-P, Sung T-Y, Hsu W-L. A Maximum Entropy approach to biomedical named entity recognition. In: Proceedings of the 4th International Conference on Data Mining in Bioinformatics. Berlin, Heidelberg: Springer-Verlag; 2004. p. 56–61.
17. Takeuchi K, Collier N. Bio-medical entity extraction using support vector machines. In: Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13. USA: Association for Computational Linguistics; 2003. p. 57–64. doi:<https://doi.org/10.3115/1118958.1118966>.
18. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *J Biomed Inform.* 2015;58:S47–52.
19. Settles B. Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets. In: Collier N, Ruch P, Nazarenko A, editors. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, NLPBA/BioNLP 2004, Geneva, Switzerland, August 28–29, 2004. 2004. <https://www.aclweb.org/anthology/W04-1221/>.
20. Shen D, Zhang J, Zhou G, Su J, Tan C-L. Effective adaptation of a hidden Markov model-based named entity recognizer for biomedical domain. In: Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13. USA: Association for Computational Linguistics; 2003. p. 49–56. doi:<https://doi.org/10.3115/1118958.1118965>.
21. Ponomareva N, Pla F, Molina A, Rosso P. Biomedical named entity recognition: a poor knowledge HMM-based approach. In: Kedad Z, Lammari N, Métais E, Meziane F, Rezgui Y, editors. Natural Language Processing and Information Systems, 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007, Paris, France, June 27–29, 2007, Proceedings. Springer; 2007. p. 382–387. doi:[https://doi.org/10.1007/978-3-540-73351-5\\_34](https://doi.org/10.1007/978-3-540-73351-5_34).
22. Tang B, Cao H, Wu Y, Jiang M, Xu H. Clinical Entity Recognition Using Structural Support Vector Machines with Rich Features. In: Proceedings of the ACM Sixth International Workshop on Data and Text Mining in Biomedical Informatics. New York, NY, USA: Association for Computing Machinery; 2012. p. 13–20. doi:<https://doi.org/10.1145/2390068.2390073>.
23. Tang B, Cao H, Wang X, Chen Q, Xu H. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international.* 2014;2014:240403.
24. Li H, Chen Q, Tang B, Wang X, Xu H, Wang B, et al. CNN-based ranking for biomedical entity normalization. *BMC Bioinform.* 2017;18:385.
25. Li L, Jin L, Jiang Z, Song D, Huang D. Biomedical named entity recognition based on extended recurrent neural networks. In: 2015 IEEE International Conference on bioinformatics and biomedicine (BIBM). IEEE; 2015. p. 649–652.
26. Wang X, Zhang Y, Ren X, Zhang Y, Zitnik M, Shang J, et al. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinform.* 2019;35:1745–52.
27. Crichton G, Pyysalo S, Chiu B, Korhonen A. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinform.* 2017;18:368.

28. Xiao S, Ouyang Y, Rong W, Yang J, Xiong Z. Similarity Based Auxiliary Classifier for Named Entity Recognition. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. p. 1140–1149.
29. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on Ten benchmarking datasets. arXiv preprint arXiv:190605474; 2019.
30. Giorgi JM, Bader GD. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*. 2018;34:4087–94.
31. Sachan DS, Xie P, Sachan M, Xing EP. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. arXiv preprint arXiv:171107908; 2017.
32. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36:1234–40.
33. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on Ten benchmarking datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019. p. 58–65.
34. Seo MJ, Kembhavi A, Farhadi A, Hajishirzi H. Bidirectional attention flow for machine comprehension. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. 2017. <https://openreview.net/forum?id=HJOUKP9ge>.
35. Cui Y, Chen Z, Wei S, Wang S, Liu T, Hu G. Attention-over-attention neural networks for reading comprehension. In: Barzilay R, Kan M-Y, editors. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. Association for Computational Linguistics; 2017. p. 593–602. doi:<https://doi.org/10.18653/v1/P17-1055>.
36. Wang W, Yang N, Wei F, Chang B, Zhou M. Gated self-matching networks for reading comprehension and question answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017. p. 189–198.
37. Li X, Yin F, Sun Z, Li X, Yuan A, Chai D, et al. Entity-relation extraction as multi-turn question answering. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 1340–1350. doi:<https://doi.org/10.18653/v1/P19-1129>.
38. Li X, Feng J, Meng Y, Han Q, Wu F, Li J. A Unified MRC Framework for Named Entity Recognition. In: Jurafsky D, Chai J, Schluter N, Tetreault JR, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020. Association for Computational Linguistics; 2020. p. 5849–5859. <https://www.aclweb.org/anthology/2020.acl-main.519/>.
39. Vinyals O, Fortunato M, Jaitly N. Pointer Networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada. 2015. p. 2692–2700. <http://papers.nips.cc/paper/5866-pointer-networks>.
40. Li X, Feng J, Meng Y, Han Q, Wu F, Li J. A Unified MRC Framework for Named Entity Recognition. arXiv preprint arXiv:191011476; 2019.
41. Devlin J, Chang M-W, Lee K, Toutanova K. 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). 2019. p. 4171–4186. <https://github.com/dccuchile/spanish-word-embeddings>. [accessed on February 9, 2021].
42. <https://github.com/PlanTL-SANIDAD/PharmaCoNER-CODALAB-Evaluation-Script>. [accessed on February 9, 2021].
43. Xiong Y, Shen Y, Huang Y, Chen S, Tang B, Wang X, et al. A Deep Learning-Based System for PharmaCoNER. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019. p. 33–37. doi:<https://doi.org/10.18653/v1/D19-5706>.
44. [https://github.com/xy-always/NER\\_Knowledge](https://github.com/xy-always/NER_Knowledge). [accessed on February 9, 2021].
45. <https://github.com/rtmdrr/testSignificanceNLP>. [accessed on February 9, 2021].
46. Stoeckel M, Hemati W, Mehler A. When Specialization Helps: Using Pooled Contextualized Embeddings to Detect Chemical and Biomedical Entities in Spanish. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019. p. 11–15. doi:<https://doi.org/10.18653/v1/D19-5702>.
47. Sun C, Yang Z. Transfer Learning in Biomedical Named Entity Recognition: An Evaluation of BERT in the PharmaCoNER task. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019. p. 100–104. doi:<https://doi.org/10.18653/v1/D19-5715>.
48. Lange L, Adel H, Strötgen J. NLNDE: Enhancing Neural Sequence Taggers with Attention and Noisy Channel for Robust Pharmacological Entity Detection. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019. p. 26–32. doi:<https://doi.org/10.18653/v1/D19-5705>.
49. Hakala K, Pyysalo S. Biomedical Named Entity Recognition with Multilingual BERT. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019. p. 56–61. doi:<https://doi.org/10.18653/v1/D19-5709>.
50. Lahuerta X, Goenaga I, Gojenola K, Atutxa Salazar A, Oronoz M. IxaMed at PharmaCoNER Challenge 2019. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019. p. 21–25. doi:<https://doi.org/10.18653/v1/D19-5704>.
51. Sohrab MG, Pham MT, Miwa M, Takamura H. A Neural Pipeline Approach for the PharmaCoNER Shared Task using Contextual Exhaustive Models. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019. p. 47–55. doi:<https://doi.org/10.18653/v1/D19-5708>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.