Improving Biomedical Named Entity Recognition with Label Re-correction and Knowledge Distillation

Huiwei Zhou (zhouhuiwei@dlut.edu.cn)
Dalian University of Technology
Zhe Liu
Dalian University of Technology
Chengkun Lang
Dalian University of Technology
Yingyu Lin
Dalian University of Technology
Junjie Hou
Dalian University of Technology

Research Article

Keywords: Biomedical named entity recognition, Knowledge distillation, Label re-correction

Posted Date: December 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-125685/v1

Improving Biomedical Named Entity Recognition with Label Re-correction and Knowledge Distillation

Huiwei Zhou[1,*], Zhe liu[1], Chengkun Lang[1], Yingyu Lin[2], Junjie Hou[3]

[1]School of Computer Science and Technology, Dalian University of Technology, Address Chuangxinyuan Building, No.2 Linggong Road, Ganjingzi District, Dalian, Liaoning, 116024, China.

[2]School of foreign languages, Dalian University of Technology, Address Chuangxinyuan Building, No.2 Linggong Road, Ganjingzi District, Dalian, Liaoning, 116024, China.

[3]School of Business, Panjin Campus of Dalian University of Technology, No.2 Dagong Road, Liaodongwan New District, PanJin, Liaoning, 124221, China.

*To whom correspondence should be addressed.

Email: zhouhuiwei@dlut.edu.cn, njjnlz@mail.dlut.edu.cn, kunkun@mail.dlut.edu.cn, lyydut@sina.com, 470559283@mail.dlut.edu.cn

Improving Biomedical Named Entity Recognition with Label Re-correction and Knowledge Distillation

Abstract

Background: Biomedical named entities recognition is one of the most essential tasks in biomedical information extraction. Previous studies suffer from inadequate annotation datasets, especially the limited knowledge contained in them.

Methods: To remedy the above issue, we propose a novel Chemical and Disease Named Entity Recognition (CDNER) framework with label re-correction and knowledge distillation strategies, which could not only create large and high-quality datasets but also obtain a high-performance entity recognition model. Our framework is inspired by two points: 1) named entity recognition should be considered from the perspective of both coverage and accuracy; 2) trustable annotations should be yielded by iterative correction. Firstly, for coverage, we annotate chemical and disease entities in a large unlabeled dataset by PubTator to generate a weakly labeled dataset. For accuracy, we then filter it by utilizing multiple knowledge bases to generate another dataset. Next, the two datasets are revised by a label re-correction strategy to construct two high-quality datasets, which are used to train two CDNER models, respectively. Finally, we compress the knowledge in the two models into a single model with knowledge distillation.

Results: Experiments on the BioCreative V chemical-disease relation corpus show that knowledge from large datasets significantly improves CDNER performance, leading to new state-of-the-art results.

Conclusions: We propose a framework with label re-correction and knowledge distillation strategies. Comparison results show that the two perspectives of knowledge in the two re-corrected datasets respectively are complementary and both effective for biomedical named entity recognition.

Keywords: Biomedical named entity recognition, Knowledge distillation, Label re-correction

1 Introduction

Biomedical Named Entity Recognition (BioNER) is a fundamental step for downstream biomedical natural language processing tasks. BioNER is a great challenge due to the following reasons: various ways of naming biomedical entities, ambiguities caused by the frequent occurrences of abbreviations, and new entities constantly and rapidly reported in scientific publications [1]. To promote the performance of BioNER systems, many challenging BioNER tasks have been proposed, such as chemical and disease named entity recognition in the BioCreative V chemical-disease relation (CDR) task [2].

Recent Named Entity Recognition (NER) studies employ neural network models to generate quality features [3,4]. However, neural network models require very large datasets to train millions of parameters. It is too expensive and time-consuming to manually annotate large datasets.

This motivate some researchers to automatically create large-scale annotation datasets with semi-structured resources and semi-supervised methods [5,6]. They generate named entity annotations by leveraging the link structure of Wikipedia.

Inevitably, these methods generate many false labels during the annotation. Zhu et al. [7] design a neural correction model trained with a small human-annotated NER dataset to correct the false labels. They illustrate that correction process could greatly improve the quality of the annotation dataset. Nevertheless, noisy labels still exist and cannot be further reduced by their method.

Bagherinezhad et al. [8] propose an iterative process called Label Refinery to reduce noisy labels caused by crop-level augmentation, and

observe that labels improve iteratively even when the same architecture model is used to refine the dataset multiple times.

In biomedical domain, there is no large-scale semi-structured dataset like Wikipedia. Instead, many large-scale structured knowledge bases are constructed, such as CTDbase [9], MeSH [10] and RGD [11]. These repositories link PubMed identifiers (PMIDs) with entity identifiers (IDs), such as <PMID: 6893628, disease ID: D010264> from MeSH. How to make use of these resources for BioNER is more challenging, and becomes an urgent demand. Wei et al. [12] first collect mentions from structured knowledge bases, and then correlate them with the text mined span from a name entity recognition and link tool PubTator [13] for mention disambiguation.

In addition, it has been long observed that combining the predictions of multiple networks usually exceeds the performance of an individual network. Unfortunately, the space to store multiple networks and the time to execute them at prediction time prohibit their use, especially when the individual models are large neural nets. Recently, a promising ensemble method, knowledge distillation [14-16], is raised to overcome this problem. It could distill diverse knowledge from different trained (teacher) models into a single student model.

Considering coverage and accuracy of NER, we construct two datasets. As for coverage, we automatically annotate the spans of chemical and disease mentions in a large unlabeled dataset by PubTator to construct a weakly labeled dataset. And as for accuracy, multiple large-scale structured knowledge bases (i.e. CTDbase, MeSH, RGD) are utilized to filter out the mentions if their IDs are not contained in the current PMID. In this way, we construct two large-scale weakly labeled datasets.

Next, we propose a novel label re-correction strategy to reduce noises in the weakly labeled datasets iteratively, and obtain two high-quality complementary datasets. They are used to train two chemical and disease named entity recognition (CDNER) models, respectively.

Finally, to integrate diverse knowledge in the two models, we utilize them as teachers to teach a distilled student model with knowledge distillation.

In summary, we mainly make the following contributions:

- We construct two weakly labeled datasets considering from precision and recall respectively by utilizing multiple knowledge bases and PubTator.
- We propose a novel label re-correction strategy for iteratively reducing noises in the weakly labeled datasets, and obtain two high-quality datasets, each of which contains over 1.3 million chemical and disease mentions.
- We introduce knowledge distillation to compress the recognition models trained on the two datasets into a single recognition model. Experimental results show that our model yields state-of-the-art results on the CDR corpus.

2 Related work

Most existing approaches treat BioNER as a sequence tagging problem. Recently, various neural network architectures have been proposed for BioNER with word and character embeddings, among which bidirectional long short-term memory with conditional random field (BiLSTM-CRF) model exhibits promising results [4].

Besides word and character features, linguistic features and domain resource features [1,17,18] are also used to enrich the information of each token. These approaches heavily rely on quality and quantity of the labeled corpora. However, such BioNER resources of each entity type are scarce.

To address this problem, datasets of different types of entities are used to augment resources for knowledge transfer by multi-task learning [19-21]. However, combining several limited datasets of different tasks could hardly meet the needs of large-scale training parameters, and the relatedness among tasks usually limits NER performance.

A recent trend in transfer learning is to take advantage of unlimited amount of unlabeled data by unsupervised pre-training. BERT is designed to pre-train language representations with large-scale unlabeled datasets, which has been proved effective for improving many natural language processing tasks [22]. Lee et al. [23] pre-train BioBERT on general and biomedical domain corpora, and illustrate that it achieves better performance than BERT on BioNER tasks.

3 Methods

In this section, we introduce our neural network-based CDNER framework, as shown in Fig. 1. Firstly, two large weakly labeled datasets are constructed with Pubtator and knowledge bases (Section

3.1). Then we apply BiLSTM-CRF as the basic model (Section 3.2) and correct noisy labels iteratively with label re-correction strategy (Section 3.3). Finally, we utilize knowledge distillation to compress the knowledge in two teacher models trained on the two re-corrected datasets into a student model (Section 3.4).

Fig. 1 The framework of our CDNER with label re-correction and knowledge distillation

Fig. 2 Illustration of the dataset generation pipeline from the perspectives of coverage and accuracy.
The chemical and disease mentions are highlighted in yellow and green, respectively.

3.1 Weakly Labeled Dataset Construction

Inspired by Wei et al. [12], we use both the records in knowledge bases and the text mined span from PubTator for dataset generation. Two large-scale weakly labeled datasets are automatically constructed for coverage and accuracy, respectively. As shown in Fig. 2, the pipeline used to create two datasets is illustrated in the following steps:

Step 1: Download PubMed abstracts whose PMIDs are in CTDbase since these abstracts contain both chemical and disease entities. Note that the abstracts in CDR corpus are not included. CTDbase associates PMIDs with pairs of chemical-disease relations, such as <PMID: 6893628, Chemical ID: D003620, Disease ID: D015451>. However, none of the repositories provides the location of the mentions.

Step 2: Automatically recognize chemical and disease mentions with PubTator to obtain the first weakly labeled dataset. PubTator provides spans of mentions, which are automatically extracted by machine learning-based taggers. These taggers were previously evaluated and achieved 80-90% of F-score in recognition and normalization. From the perspective of coverage, we do not filter any mentions.

Step 3: Filter the spans whose entity IDs are not associated with the current PMID by using the repositories, i.e. CTDbase, MeSH, RGD. Obviously, many false positives exist in the first dataset. From the perspective of accuracy, only the spans that are paired with the repository records are remained. For example, the span of "spastic paraplegia" with entity ID D008223 and the span of "malignant lymphoma" with entity ID C536857 recognized by PubTator in Fig. 2 are filtered because they are not recorded in PMID 6893628 in repositories. Note that, CTDbase only associates PMIDs with pairs of chemical-disease relations. Therefore, MeSH and RGD are employed to complement records.

Table 1 Various statistics of the datasets.

Dataset		#Abstract	#Chemical	#Disease
Weakly labeled	CDWC	70026	706593	514964
	CDWA		503700	283293
	CDRC		770159	541235
	CDRA		781039	532198
Human annotated	CDR train	500	5203	4182
	CDR dev	500	5347	4244
	CDR test	500	5385	4424

#Abstract: the number of abstracts
#Chemical: the number of chemical mentions
#Disease: the number of disease mentions

In this way, two Chemical and Disease Weakly labeled datasets for Coverage and Accuracy are created called CDWC and CDWA with same abstracts but different annotations. We list the statistics of the two datasets in the first two rows in Table 1. We can see that only 71.29% chemical and 55.01% disease annotations are remained after filtering. We believe that many false positives are filtered.

3.2 BiLSTM-CRF model

We use BiLSTM-CRF model as our basic model, which has four layers as shown in Fig. 1. In the embedding layer, each token w_i in a sentence $\mathbf{w} =< w_1, w_2, ..., w_n >$ is represented as $\mathbf{X} =< \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n >$, where $\mathbf{x}_i$ is the concatenation of word embedding and character embedding learned by a character-level convolutional neural network.

$\mathbf{X}$ is then fed to a BiLSTM layer to obtain the hidden representation of each token by concatenating its left and right context representations.

The liner layer on the top of the BiLSTM is used to predict confidence scores $\mathbf{P} =< P_1, P_2, ..., P_n > \in R^{k \times n}$ for all tokens having each of the possible labels, where k is the number of distinct labels.

Finally, a CRF layer is applied to decode the best tag path in all possible tag paths. The score of $\mathbf{X}$ with a sequence of labels $\mathbf{y} =< y_1, y_2, ..., y_n >$ is defined as the sum of transition scores and confidence scores:

$$s(\mathbf{X},\mathbf{y}) = \sum_{i=1}^{n} (T_{y_{i-1},y_i} + P_{i,y_i}) \quad (1)$$

where $T_{i,j}$ represents the transition score from the *i*-th tag to the *j*-th tag.

The training loss of the basic model is defined by:

$$\mathcal{L}_{crf} = -\log \frac{e^{s(\mathbf{X},\mathbf{y})}}{\sum_{\mathbf{y}' \in \mathbf{Y_X}} e^{s(\mathbf{X},\mathbf{y}')}} \quad (2)$$

where $\mathbf{Y_X}$ are all possible tag paths.

3.3 Label Re-correction Strategy

Inevitably, many noisy annotations exist in CDWC and CDWA. In this paper, we propose a novel label re-correction strategy to reduce noises in the weakly labeled datasets by leveraging a small manually-annotated CDR training dataset. Here BiLSTM-CRF is used as our re-correction model.

There are two intuitions behind our label re-correction strategy: 1) the annotations in CDR training dataset can help us learn how to generate annotations in the large-scale dataset; 2) the iterative procedure to update labels can improve both the dataset and the trained re-correction model.

Given CDR training data ***T*** and a large-scale dataset ***L***, the process of label re-correct steps is defined as follows: we firstly train a new label re-correction model *C* on ***L***; then we transfer the model *C* to ***T***; finally, the re-correction model *C* is used to re-correct the label sequences in ***L***. We repeat such a re-correct procedure until the *F*-score on development set does not increase.

CDR corpus contains 1500 PubMed abstracts: 500 each for training, development and test set, as shown in Table 1.Following Luo et al. [1], the original training set and development set are merged as the training dataset D. We randomly select 10% of D as a validation dataset.

Two weakly labeled datasets CDWA and CDWC are Re-corrected to obtain two corresponding high-quality datasets called CDRA and CDRC, respectively. The number of annotations in the two Re-corrected datasets is listed in the third and fourth rows in Table 1. We can see that the chemical and disease annotations in CDRC are 9.00% and 5.10% more than those in CDWC, respectively. And the chemical and disease annotations in CDRA are 55.06% and 87.86% more than those in CDWA, respectively. We believe that label re-correction strategy could effectively correct the noisy labels, especially the false-negative entity labels.

3.4 Knowledge Distillation

Two datasets CDRC and CDRA aim to annotate chemical and disease entities from the perspectives of coverage and accuracy, respectively. We use them to train two models $T_j, j \in \{\mathrm{c},\mathrm{a}\}$, which are complementary.

We calculate the label similarity of each abstract predicted by two **R**e-corrected models over large-scale dataset, and the label similarity distribution is shown in Fig. 3. From this figure, we can see that though most of predicted labels are the same, there are still a lot of differences between the two models. It is natural to combine them to get a better model.

Fig. 3 Label similarity distribution over the large-scale dataset between the predictions of the two teacher models.
Each bar represents the number of the abstracts in the similarity interval.

We distill the knowledge from the two models (teacher) and transfer it to a new model (student). The structures of teachers and student could be identical or different. In this paper, the teachers $T_j, j \in \{\mathrm{c},\mathrm{a}\}$ and the student S are based on the same BiLSTM-CRF.

The label sequences (hard labels) $\mathbf{y}$ and the confidence scores $\mathbf{P}$ (used to calculate soft labels) predicted by the two teacher models are both used to teach the student model. Once the teacher models are trained, their parameters are frozen during the student model training.

With the hard labels, the student model is trained by minimizing the loss of $\mathcal{L}_{crf}^{T_j}$ in Eq. (2) based on $\mathbf{y}^{T_j}$ predicted by the teacher model T_j.

With the soft labels, the student model is trained by minimizing the similarity distance between soft labels of the reference teacher and the estimated student. The similarity metric can be formulated as:

$$\mathcal{L}_{sim}^{T_j} = \sum_{i=1}^{n} d(q_i^{T_j}, q_i^{S}) \quad (3)$$

where $q_i^{T_j}$ and q_i^{S} are soft labels generated by the teacher and the

student, respectively, d is referred to as a distance function. In this work, we investigate three distance metrics as follows:

- l_1 **distance**: We apply a softmax layer on confidence scores $P_i^{T_j}$ and P_i^S of each token to get the corresponding soft labels $q_i^{T_j}$ and q_i^S. l_1 distance is the absolute differences between the soft labels:

$$\mathcal{L}_{l_1\_sim}^{T_j} = \frac{1}{n}\sum_{i=1}^{n}|q_i^{T_j} - q_i^S| \tag{4}$$

- l_2 **distance**: Here, the soft labels are the same as those used in l_1 distance. l_2 distance is the straight-line distance in euclidean space between the soft labels:

$$\mathcal{L}_{l_2\_sim}^{T_j} = \frac{1}{n}\sum_{i=1}^{n}||q_i^{T_j} - q_i^S||_2^2 \tag{5}$$

- l_{KD} **distance**: Following Hinton et al. [14], we use a softmax layer to convert $P_i^{T_j}/t$ and P_i^S/t to soft labels $q_i^{T_j}$ and q_i^S, where t is the temperature. Then the l_{KD} distance is defined as the cross-entropy between the soft labels multiplied by t^2:

$$\mathcal{L}_{KD\_sim}^{T_j} = \frac{t^2}{n}\sum_{i=1}^{n}q_i^{T_j}\log q_i^S \tag{6}$$

The final objective loss for the distilled model is the sum of the hard label losses and the soft label losses:

$$\mathcal{L} = \sum_{j\in\{c,a\}}\mathcal{L}_{sim}^{T_j} + \mathcal{L}_{crf}^{T_j} \tag{7}$$

4 Experiment and Discussion

4.1 Experimental Settings

Dataset and Evaluation Metrics: We evaluate the proposed CDNER framework on CDR test dataset. The evaluation is reported by official evaluation toolkit, which adopts the standard Precision (P), Recall (R) and F-score (F) to measure the performance.

Implementation Details: Word2Vec [24] is employed to pretrain 100-dimension word embeddings on the PubMed articles provided by Wei et al. [13]. Other parameters are initialized randomly from uniform distributions. The dimension of character embeddings is 50. Mini-batch size is set to 32 and 4 for the model trained on the large-scale dataset and CDR dataset, respectively. RMSProp optimizer with learning rate 1e-3 is used to minimize the loss.

4.2 Effects of Label Re-correction

We first investigate the effects of the label re-correction strategy. Table 2 shows the results of the BiLSTM-CRF model trained on the two weakly labeled datasets, respectively. For CDWC and CDWA, the label re-correction process is repeated multiple times before convergence.

Table 2 Comparison of basic model results trained on CDWC and CDWA with different re-correction times.

Dataset	P(%)	R(%)	F(%)	Dataset	P(%)	R(%)	F(%)
CDWC	89.72	83.65	86.58	CDWA	94.75	67.27	78.68
CDWC^1	89.84	89.32	89.58	CDWA^1	90.16	88.94	89.55
CDWC^2	90.00	89.35	89.67	CDWA^2(CDRA)	91.03	88.31	**89.65**
CDWC^3(CDRC)	89.80	89.82	**89.81**				

All results are evaluated on the CDR test set. Each dataset is constructed by the re-correction model trained with the dataset right above it. The superscript represents the re-correction times. That is, CDWC^1 is the dataset constructed by the re-correction model trained on the CDWC. The first row datasets are the weakly labeled datasets without re-correction. What's more, CDWC^3 is CDRC, and CDWA^2 is CDRA.

From the table, we can observe that although the first re-correction process significantly improves the F-score, correcting only once is not enough. As the label re-correction process is further performed iteratively, the labels of the two datasets improve gradually, and thereby benefit the re-correction models.

In addition, we also find that the results on CDWC datasets keep a relatively high recall, while those on CDWA datasets have a relatively high precision, which is in line with our original motivation.

4.3 Effects of Knowledge Distillation

We further explore the effects of knowledge distillation, which are summarized in Table 3. The first row indicates the model simply trained on the combination of CDRC and CDRA datasets without knowledge distillation. Comparing this row with others, we find that the distilled knowledge can boost the F-score consistently.

The following four rows investigate the influences of different combinations of hard label losses and soft label losses for knowledge distillation. It is observed that using both soft labels and hard labels can achieve better performance than single hard labels, which is probably because soft labels contain some helpful information that may not be encoded in hard labels. Specially, using l_2 distance and hard labels achieves the best performance.

Table 3 Performance comparison of the distilled models trained with different combinations of losses.

$\mathcal{L}_{crf}$	$\mathcal{L}^T_{crf}$	$\mathcal{L}^T_{KD\_sim}$	$\mathcal{L}^T_{l_1\_sim}$	$\mathcal{L}^T_{l_2\_sim}$	Adv	F (%)
✓						89.99
	✓					90.13
	✓	✓				90.16
	✓		✓			90.13
	✓			✓		**90.35**
	✓			✓	✓	90.16

Adv: the short for adversarial learning

Besides, adversarial learning is commonly used in knowledge distillation. We also introduce adversarial learning into our model as Shen et al. [16] did (last row in Table 3). Unfortunately, it does not work. The possible reason is that there exists some potential conflict of information between the two teachers. It is difficult to force the student to generate similar outputs to the two teachers' at the same time.

4.4 Ablation Study

To better understand the function of key components of our framework, we conduct some ablation studies in Table 4.

Table 4 Ablation study results. w/o CDRA means that we train a single teacher without CDRA (i.e. only with CDRC)

Model	P (%)	R (%)	F (%)
Our best	90.71	**89.99**	**90.35**
w/o label re-correction	**91.34**	80.76	85.73
w/o CDRC	90.48	89.14	89.81
w/o CDRA	90.17	89.55	89.86

w/o label re-correction: we train the teachers on the two weakly labeled datasets CDWC and CDWA rather than CDRC and CDRA
w/o CDRC: we train a single teacher without CDRC (i.e. only with CDRA)
w/o CDRA: we train a single teacher without CDRA (i.e. only with CDRC)

Does label re-correction strategy really need to be applied to the weakly labeled datasets? See the second row, instead of using re-corrected datasets CDRC and CDRA, we use weakly labeled datasets CDWC and CDWA to train the teachers. The recall of the distilled student model drops significantly. This proves the effectiveness of label re-correction, especially for reducing false negatives in the weakly labeled datasets.

Are both the datasets for coverage and accuracy beneficial? See the last two rows, when we only use the dataset from one perspective, the performance of each student model drops but is still promising. This suggests that the datasets from two perspectives are complementary and both effective.

4.5 Main Results

We compare our distilled model with state-of-the-art methods on the BioCreative V CDR Task in Table 5. The BiLSTM-CRF model trained on the CDR training dataset is our baseline. These relevant models are divided into four groups. Except our model encoded with BioBERT and the method proposed by Lee et al. [23], all these methods are based on BiLSTM-CRF. To compare with other method in detail, the evaluation is performed on chemical type, disease type and both types.

Comparing group 1 and group 2, we find that rich features indeed improve the performance However, designing and extracting such features is laborious and time-consuming.

While comparing group 1 and group 3, we can see that multi-task learning could improve performance to a certain extent though data augmentation.

Table 5. Comparison with some state-of-the-art methods

	Methods	Chemical *F*(%)	Disease *F*(%)	Both *F*(%)
1	Habibi et al. [4]	91.05	83.49	87.63*
	Our Baseline (BiLSTM)	91.42	83.59	87.86
2	Luo et al. [1]	92.57	-	-
	Dang et al. [17]	93.14	84.68	89.30*
3	Wang et al. [20]	-	-	88.78
	Yoon et al. [21]	92.74	82.61	88.15*
4	Lee et al. [23]	93.47	87.15	90.60*
	Our model (BiLSTM)	94.17	85.69	90.35
	Our model (BERT)	**95.22**	**87.34**	**91.64**

1: models with word and character features
2: models with additional domain resource features and linguistic features
3: models with multi-task learning
4: models with large-scale unlabeled datasets
*: indicates that the results are calculated by us according to their reported results in chemical and disease

Our model and Lee et al. [23] leverage large-scale unlabeled datasets, significantly outperforming other methods. Lee et al. [23] pre-train BioBERT on the datasets with totally 21.3B words, and then fine-tune it on the CDR training data, while our model encoded with BiLSTM is trained on the datasets with only 14.8M words. The amount of their datasets and the parameter scale of their model are much larger than ours. Even though, our model with vector dimension 100 achieves a competitive performance of Lee et al. [23] with vector dimension 768 on both. This demonstrates the effectiveness of our label re-correction and knowledge distillation strategies. Our weakly labeled dataset is constructed specifically for chemical and disease entity recognition, which is more task-specific than directly using BioBERT. During the training process on the weakly labeled data, our word vector is fine-tuned at the same time, so the word vector could remain rich knowledge about chemical and disease entity recognition.

And when we use BioBERT as encoder to re-correct the weakly labeled datasets and train a student model, it outperforms Lee et al. [23], achieves a new state-of-art F-score of 91.64%.

4.6 Case Study

To better understand in which conditions the knowledge distillation helps, we give the annotations of the same input sentence predicted by the models before and after distillation in Fig. 4. To clearly explain why the student model out-performs the teacher models, we also output the label probabilities of the words “Coxon” and “scoline” in Table 6.

Fig. 4 Case study of knowledge distillation effectiveness. Yellow for chemical and green for disease

Table 6. Label probabilities of the words “Coxon” and scoline predicted by T_c, T_a and our model.

word	Model	B-Chemical (%)	I- Chemical (%)	B-Disease (%)	I-Disease (%)	O (%)
Coxon	T_c	89.66	0.81	0.33	0.07	9.14
	T_a	5.14	11.47	1.08	2.13	80.17
	Our model	23.38	11.69	3.17	6.46	55.31
scoline	T_c	97.39	0.27	0.80	0.01	1.54
	T_a	23.04	10.53	8.78	2.16	55.49
	Our model	90.70	3.86	2.67	0.47	2.30

For the word “Coxon”, see Fig. 4 and the first three rows in Table 6. Teacher model T_a correctly predicts it as “O” with the probability of 80.17%, while teacher model T_c incorrectly predicts it as “B-Chemical” with the probability of 89.66%. However, through the knowledge distillation, the student model selectively learns from the two teachers and balances their probability values. Finally, the probability of label “O” is 55.31%, which is larger than that of label “B-Chemical” with the probability 23.38%. This illustrates that student model can effectively distill the trustable knowledge from the teachers.

Similarly, for the word “scoline”, see the last three rows in Table 6. The label probabilities of the two teacher models are quite different. The student model effectively distills the knowledge from the two

models, finally assigning the probability of 90.70% to the right label "B-Chemical".

However, the student model also makes some errors. It identifies a false boundary for the disease entity "fasciculations" since there is no reliable knowledge from two teachers and previous learning.

5 Conclusion

In this paper, we address the problem of insufficient training corpus that BioNER suffers from. A novel label re-correction strategy is proposed to make full use of PubTator and knowledge bases to obtain two large-scale high-quality datasets for coverage and accuracy, respectively. Further, we introduce knowledge distillation to transfer knowledge from two teacher models into a distilled student model. Experiments on the BioCreative V CDR dataset show that label re-correction benefits recognition significantly and knowledge distillation further improves recognition. As a result, we achieve the new state-of-the-art results. In terms of further work, we would like to integrate semi-supervised learning and multi-task learning to construct large-scale datasets for broader knowledge transfer.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and code during the current study are available at https://github.com/ZheLiu1996/Label-Re-correction-and-Knowledge-Distillation.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by the National Natural Science Foundation of China [grant numbers 61772109] and the Humanities and Social Science Fund of Ministry of Education of China [grant numbers 17YJA740076].

Authors' contributions

HZ and ZL designed the study. ZL performed the experiments. All authors participated in manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank the editors and all anonymous reviewers for valuable suggestions and constructive comments

References

1. Luo L, Yang Z, Yang P, Zhang Y, Wang L, Lin H, Wang J. An attention-based BiLSTM-CRF approach to document-level chemidcal named entity recognition. Bioinformatics. 2017; 34(8):1381-1388.
2. Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Overview of the BioCreative V chemical disease relation (CDR) task. Proceedings of the fifth BioCreative challenge evaluation workshop. 2015; 14.
3. Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. ACL. 2016.
4. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017; 33(14):i37-i48.
5. Nothman J, Ringland N, Radford W, Murphy T,Curran JR. Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence. 2013; 194:151-175.
6. Ghaddar A, Langlais P. Winer: A wikipedia annotated corpus for named entity recognition. IJCNLP. 2017; 1(Long Papers):413-422.
7. Zhu M, Deng Z, Xiong W, Yu M, Zhang M, Wang WY. Towards Open-Domain Named Entity Recognition via Neural Correction Models. AAAI. 2020.
8. Bagherinezhad H, Horton M, Rastegari M, A, Farhadi. Label refinery: Improving imagenet classification through label progression. 2018. arXiv preprint arXiv:1805.02641.
9. Mattingly CJ, Colby GT, Forrest JN, Boyer JL. The Comparative Toxicogenomics Database (CTD). Environmental health perspectives. 2003; 111(6):793-795.
10. Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association. 2000; 88(3):265.
11. Nigam R, Laulederkind SJF, Hayman GT, Smith JR, Wang SJ, et al. Rat Genome Database: a unique resource for rat, human, and mouse quantitative trait locus data. Physiological genomics. 2013; 45(18):809-816.

12. Wei CH, Lee K, Leaman R, Lu Z. Biomedical Mention Disambiguation using a Deep Learning Approach. ACM. 2019; 307-313.
13. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic acids research. 2013; 41(W1):W518-W522.
14. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. NIPS. 2015.
15. Li Y, Yang J, Song Y, Cao L, Luo J, Li LJ. Learning from Noisy Labels with Distillation. ICCV. 2017; 1910-1918.
16. Shen Z, He Z, Xue X. Meal: Multi-model ensemble via adversarial learning. AAAI. 2019; 33:4886-4893.
17. Dang TH, Le HQ, Nguyen TM, Vu ST. D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. Bioinformatics. 2018; 34(20):3539-3546.
18. Wang J, Xu W, Fu X, Xu G, Wu Y. ASTRAL: Adversarial Trained LSTM-CNN for Named Entity Recognition. Knowledge-Based System. 2020; 197.
19. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics. 2016; 32(18):2839-2846.
20. 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. Bioinformatics. 2019; 35(10):1745-1752.
21. Yoon W, So CH, Lee J, Kang J. CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. BMC bioinformatics. 2019; 20(10):249.
22. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. NAACL-HLT. 2019.
23. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020; 1-7.
24. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. NIPS. 2013.

Figures

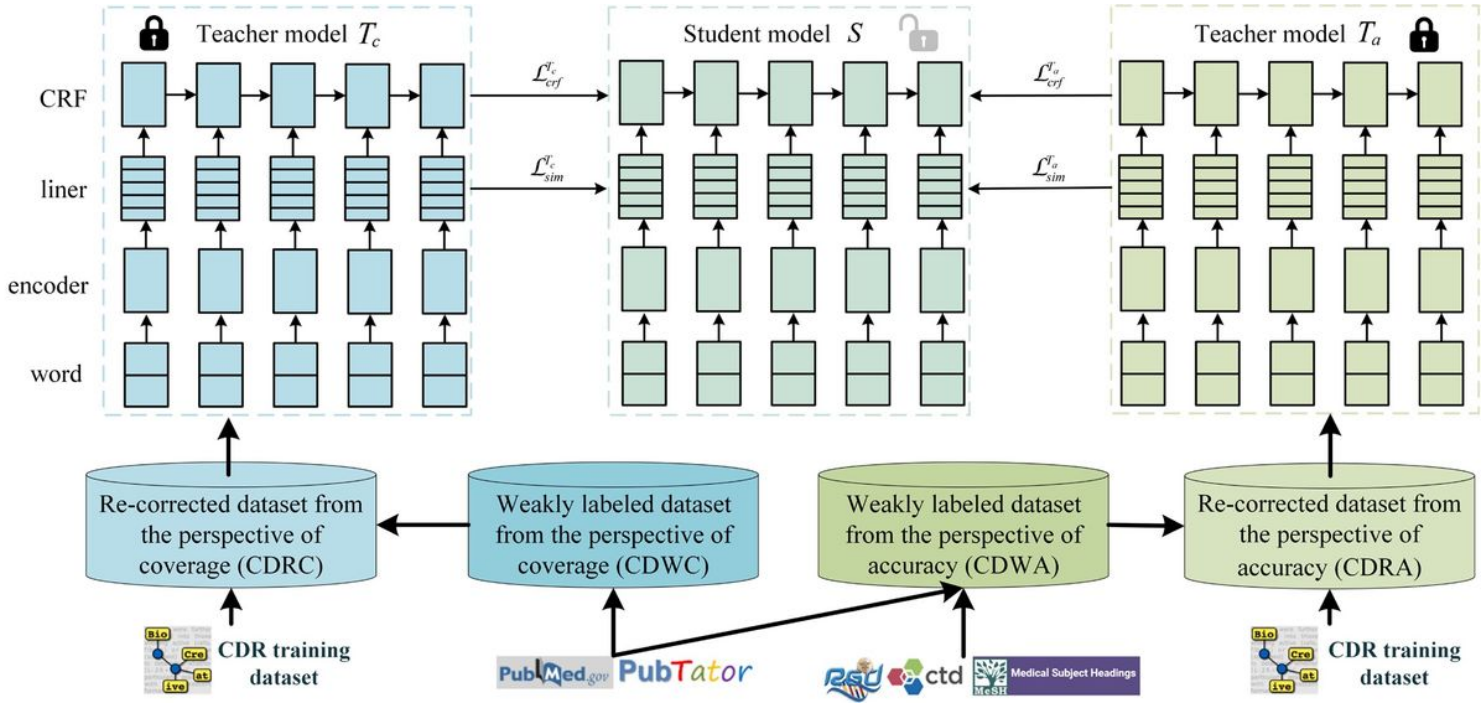

Figure 1

The framework of our CDNER with label re-correction and knowledge distillation

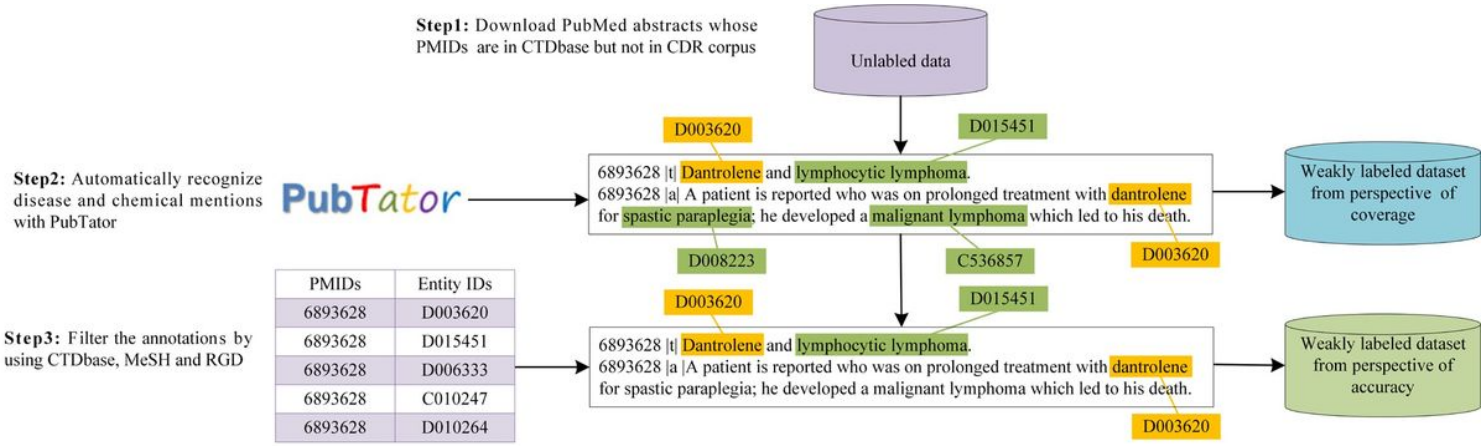

Figure 2

Illustration of the dataset generation pipeline from the perspectives of coverage and accuracy

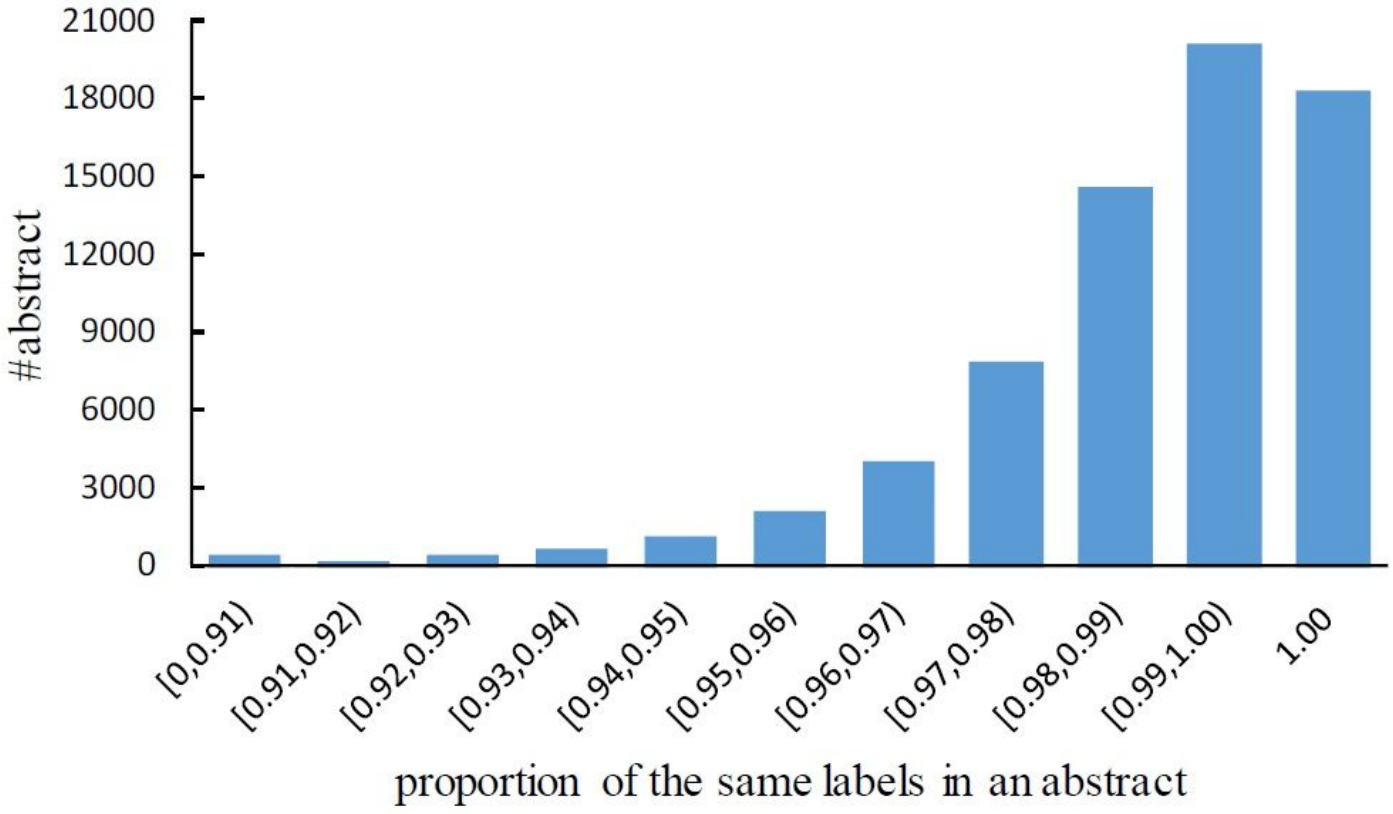

Figure 3

Label similarity distribution over the large-scale dataset between the predictions of the two teacher models

Model	Annotations
T_c	…report by Coxon, scoline pain occurs in African negroes… The abolition of muscle fasciculations...
T_a	…report by Coxon, scoline pain occurs in African negroes… The abolition of muscle fasciculations...
Our Model (BiLSTM)	…report by Coxon, scoline pain occurs in African negroes… The abolition of muscle fasciculations...
Gold	…report by Coxon, scoline pain occurs in African negroes… The abolition of muscle fasciculations...

Figure 4

Case study of knowledge distillation effectiveness