Two-perspective Biomedical Named Entity Recognition with Weakly Labeled Data Correction

Huiwei Zhou
School of Computer Science and Technology
Dalian University of Technology
Dalian, China
zhouhuiwei@dlut.edu.cn

Zhe Liu
School of Computer Science and Technology
Dalian University of Technology
Dalian, China
njjnlz@mail.dlut.edu.cn

Chengkun Lang
School of Computer Science and Technology
Dalian University of Technology
Dalian, China
kunkun@mail.dlut.edu.cn

Yibin Xu
School of Computer Science and Technology
Dalian University of Technology
Dalian, China
19xyb@ mail.dlut.edu.cn

Lei Du
School of Mathematical Sciences
Dalian University of Technology
Dalian, China
dulei@dlut.edu.cn

Abstract—**Biomedical Named Entity Recognition (BioNER) is one of the most essential tasks in biomedical information extraction. Previous studies suffer from inadequate annotation datasets, especially the limited knowledge inside. This paper proposes a two-perspective named entity recognition method with Weakly Labeled (WL) data correction. Firstly, from the perspective of coverage and accuracy, we utilize PubTator and multiple knowledge bases to construct two large-scale WL datasets, which are then revised by their corresponding label correction models respectively, obtaining two high-quality datasets. Finally, we compress the knowledge in the two datasets into a BioNER model with partial label integrating. Our approach achieves new state-of-the-art performances on three BioNER datasets.**

Keywords—biomedical named entity recognition, two-perspective named entity recognition, weakly labeled dataset construction, noise correction, partial label integrating

I. Introduction

Named Entity Recognition (NER) aims to identity the boundaries of entity mentions in texts and classify them into predefined categories. It is a foundation for further level of complex information extraction tasks. Domain-specific NER is a challenging problem, especially in Biomedical domain [1]. To promote the Biomedical Named Entity Recognition (BioNER) performance, many challenging tasks have been proposed, such as chemical NER in CHEMDNER [2], chemical and disease NER in CDR [3] and disease NER in NCBI Disease[4].

Recently, to improve NER performance, neural networks are used to automatically generate quality features [1], [5], [6]. However, neural network models have millions of parameters and require large datasets to reliably estimate the parameters. It is too expensive to keep manually annotated large datasets up-to-date.

To address this problem, datasets of different types of entities are used to augment resources for knowledge transfer by multi-task learning [7], [8]. However, the relatedness among tasks usually limits NER performance.

A recent trend is to leverage unlimited amount of unlabeled data. BioBERT has been pre-trained on large-scale unlabeled general and biomedical domain corpus, which has been proven effective in BioNER tasks [9].

Meanwhile, some researchers automatically create large-scale Weakly Labeled (WL) datasets with semi-structured resources [10]. They exploit the link structure of Wikipedia to generate named entity annotations. Inevitably, these WL datasets contain many false labels. Cao et al. [11] maximize the potential of WL data by assigning unlabeled words with all possible labels, achieving sequential optimum with partial-CRF [12].

Zhu et al. [13] leverage a human-annotated NER dataset to correct the false labels in a WL dataset. However, they need a parallel correction dataset.

In biomedical domain, there is neither large-scale semi-structured dataset nor parallel correction dataset. Instead, researchers construct multiple large-scale structured knowledge bases, such as CTDbase [14], MeSH [15] and RGD [16]. These repositories associate PubMed identifiers (PMIDs) with entity identifiers (IDs). How to make use of these resources for BioNER becomes an urgent demand. Wei et al. [17] collect mentions from multiple structured knowledge bases, and then correlate them with the text mined span from a named entity recognition and link tool PubTator [18] for mention disambiguation.

By utilizing multiple human-curated databases and PubTator, this paper constructs two datasets from the perspective of the coverage and accuracy. 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. As for accuracy, multiple large-scale structured knowledge bases are utilized to filter out the mentions whose IDs are not included in the current PMID. In this way, we construct two large-scale WL datasets. Then, two label correction models are trained on the two datasets with the guidance of a human-annotated dataset, obtaining two high-quality WL datasets. Finally, knowledge in the two complementary corrected datasets is compressed into a single BioNER model. For better reproduction, we openly release the entire project at https://github.com/ZheLiu1996/TBNER.

In summary, we mainly make the following contributions:

- We propose a novel label correction strategy to revise large-scale WL datasets with guide of a small human-annotated dataset, obtaining high-quality datasets.
- We integrate the knowledge in the two complementary corrected datasets by modeling two sets of possible labels derived from the same corpus.

978-1-7281-6215-7/20/$31.00 ©2020 IEEE

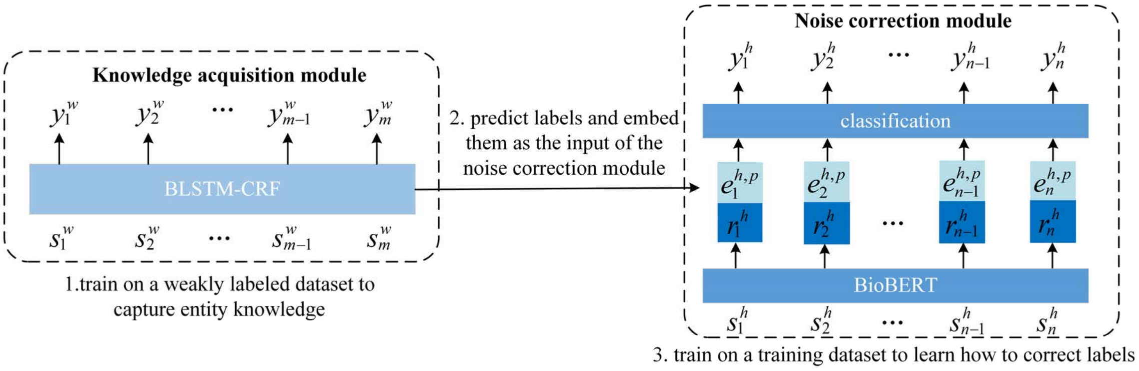
Knowledge acquisition module
y_1^w y_2^w … y_{m-1}^w y_m^w
BLSTM-CRF
s_1^w s_2^w … s_{m-1}^w s_m^w
1.train on a weakly labeled dataset to capture entity knowledge
2. predict labels and embed them as the input of the noise correction module
Noise correction module
y_1^h y_2^h … y_{n-1}^h y_n^h
classification
$e_1^{h,p}$ $e_2^{h,p}$ $e_{n-1}^{h,p}$ $e_n^{h,p}$
r_1^h r_2^h … r_{n-1}^h r_n^h
BioBERT
s_1^h s_2^h … s_{n-1}^h s_n^h
3. train on a training dataset to learn how to correct labels

Fig. 1. The architecture of the label correction model. When training the knowledge acquisition module, its input is the weakly labeled dataset. While training the noise correction module, the knowledge acquisition module predicts labels of the human annotated dataset, which are used as the input of the noise correction module. During correcting noises in the weakly labeled dataset, the knowledge acquisition module predicts labels of the weakly labeled dataset, which are used as the input of the noise correction module for label correction.

TABLE I. VARIOUS STATISTICS OF THE DATASETS.

Dataset	#Abstract	#Chemical	#Disease
CDWC	70026	706593	514964
CDWA		503700	283293
CDRC		770159	541235
CDRA		781039	532198
CCC		759093	-
CCA		752803	-
DCC		-	592105
DCA		-	499007

"#Abstract", "#Chemical" and "#Disease" mean the number of abstracts, chemical mentions and disease mentions, respectively.

- Experimental results show that our model yields state-of-the-art results on the CDR, CHEMDNER and NCBI disease corpus.

II. Methods

This section introduces our two-perspective named entity recognition method.

A. Weakly Labeled Dataset Construction

Inspired by Wei et al. [17], we automatically construct two large-scale WL datasets from the perspective of coverage and accuracy, respectively. The pipeline consists of the following three steps:

Step 1: Download PubMed abstracts whose PMIDs are in CTDbase since these abstracts contain both chemical and disease entities.

Step 2: Automatically recognize disease and chemical mentions by PubTator to obtain the weakly labeled dataset for coverage.

Step 3: From the perspective of accuracy, filter the mentions recognized in step 2, whose entity IDs are not associated with the current PMID by using the repositories (i.e. CTDbase, MeSH and RGD).

In this way, we create two **C**hemical and **D**isease **W**eakly labeled datasets for **C**overage and **A**ccuracy with the same abstracts but different annotations, which are called CDWC and CDWA, respectively. The statistics of annotations in two datasets are listed in the first two rows in Table I.

B. Label Correction

```
Algorithm 1 Knowledge acquisition algorithm
Input:   $\{S^h, Y^h\}$ : a human annotated dataset
         $\{S^w, Y^w\}$ : a weakly labeled dataset
Output: $M^k$ : the knowledge acquisition module
1. Learn a judge model $M^h$ from $\{S^h, Y^h\}$ based on BLSTM-CRF with randomly sample;
2. Use $M^h$ to predict each sentence $\mathbf{s}^w \in S^w$, obtaining its label sequence $\mathbf{y}^M$;
3. Calculate the entity prediction $F$-score of $\mathbf{y}^w \in Y^w$ based on the corresponding $\mathbf{y}^M$;
4. Rank all sentences $\mathbf{s}^w \in S^w$ according to their $F$-scores from low to high, obtaining the ranked $\{S^w, Y^w\}$;
5. Learn $M^k$ from the ranked $\{S^w, Y^w\}$.
```

Inevitably, two WL datasets CDWC and CDWA remain a large number of noisy labels. Therefore, we design a label correction model to correct false labels. Since two datasets are corrected with the same process, we take one dataset to illustrate. Hereafter, we denote data sources as superscripts of symbols and positions as subscripts of symbols.

Our label correction model contains two modules: the knowledge acquisition module and the noise correction module, as shown in Fig. 1. The knowledge acquisition module is first trained on a WL dataset with curriculum learning to capture its entity knowledge. The noise correction module is then trained on the human-annotated dataset to learn how to correct the noisy labels predicted by knowledge acquisition module.

Knowledge acquisition module:

BioBERT obtains state-of-the-art results on BioNER tasks by simply fine-tuning pre-training parameters. We believe that BioBERT remember many entity knowledge with large-scale trained parameters. However, there are many noisy labels in WL datasets. Therefore, we use BLSTM-CRF with fewer parameters to acquire entity knowledge with noisy labels instead of BioBERT.

We denote the **w**eakly labeled dataset as $\{S^w, Y^w\}$, and the **h**uman annotated dataset as $\{S^h, Y^h\}$, where S^w and S^h are token sequences, and Y^w and Y^h are their corresponding label sequences, respectively. Knowledge acquisition module M^k is trained on S^w with weakly labels Y^w to acquire the entity knowledge in the WL dataset. The knowledge acquisition algorithm is illustrated in Algorithm 1.

Bengio et al. [19] demonstrate that curriculum learning train models better than random. We use curriculum learning to assist knowledge acquisition. The higher the entity prediction F-score of the sentence is, the more correct its labels are. Because there are no gold labels for S^w, we use a judge model M^h with the BLSTM-CRF structure, to predict

```
Algorithm 2 Noise correction algorithm
Input:   $\{S^h, Y^h\}$ : a human annotated dataset
         $\{S^w, Y^w\}$ : a weakly labeled dataset
         $M^k$ : the trained knowledge acquisition module learned in
         Algorithm 1
Output: $M^c$ : the noise correction module
        $\{S^w, Y^c\}$ : the corrected weakly labeled dataset
1. Use $M^k$ to predict each sentence $\mathbf{s}^h \in S^h$, obtaining its noisy label
   sequence $\mathbf{y}^{h,p} = \{y_1^{h,p}, y_2^{h,p}, \ldots, y_n^{h,p}\}$ ;
2. Embed $\mathbf{y}^{h,p}$ to noisy label vector sequence $\mathbf{e}^{h,p} = \{e_1^{h,p}, e_2^{h,p}, \ldots, e_n^{h,p}\}$ ;
3. Input the same sentence $\mathbf{s}^h$ into a pre-trained BioBERT to obtain its
   token representation sequence $\mathbf{r}^h = \{r_1^h, r_2^h, \ldots, r_n^h\}$ ;
4. Concatenate the noisy label embeddings and the token
   representations in the corresponding positions as $c_i^h = e_i^{h,p} \oplus r_i^h$ ;
5. Fine-tune BioBERT, label embeddings $\mathbf{e}^{h,p}$ and classification layer
   to correct the noisy label sequence $\mathbf{y}^{h,p}$ to the gold label $\mathbf{y}^h \in Y^h$
   based on $\{S^h, Y^h\}$, obtaining the noise correction module $M^c$ .
6. Use $M^k$ to predict each sentence $\mathbf{s}^w \in S^w$, obtaining its noisy label
   sequence $\mathbf{y}^{w,p}$ ;
7. Use $M^c$ to correct each sentence $\mathbf{s}^w \in S^w$ with label sequence
   $\mathbf{y}^{w,p}$, obtaining its corrected label sequence $\mathbf{y}^c \in Y^c$ .
```

the labels as gold labels. All the sentences in a WL dataset are ranked in a wrong-to-right order.

Noise correction module:

After acquiring the entity label knowledge in a WL dataset, a noise correction module is followed to correct the noisy labels in the WL dataset as shown in the right part of Fig. 1. To accurately correct the noisy labels, we implement noise correction module with BioBERT. The noise correction algorithm is illustrated in Algorithm 2.

Instead of a parallel correction dataset, an arbitrary human-annotated dataset is used for training noise correction module M^c. The objective cross entropy loss function $L_{correction}$ is defined as follows:

$$L_{correction} = -\sum_{i=1}^{n} y_i^h \log p(y_i^h \mid s_i^h, y_i^{h,p}; \theta) \qquad (1)$$

where n is the token length of the sentence, $p(\cdot)$ denotes the classification probability, and θ is model parameters.

To construct the high-quality datasets, all sentences S^w in CDWC or CDWA are fed to the noise correction module to get their corrected label sequences Y^c. For **CD**R, **C**HEMDNER and NCBI-**D**isease, the **C**orrected labeled datasets CDCC and CDCA, CCC and CCA, DCC and DCA are constructed from the perspective of **C**overage and **A**ccuracy, respectively. The statistics of the corrected datasets are listed in Table I.

C. Partial label integrating

To compress the knowledge in the two complementary corrected datasets, we train a single BioNER model on both datasets. Note that the two datasets are derived from the same corpus, which means each training sentence have two set of labels. To effectively utilize the different labels in the two datasets and integrate their knowledge, we propose partial label integrating, as shown in Fig. 2. BioBERT is employed for entity recognition. Since the three dual datasets are compressed with the same process, we take CDCA and CDCC for illustration.

Partial-CRF (PCRF) [12] is trained by assuming the uncoupled words may refer to multiple labels. We utilize PCRF to maximize the probability of two possible sequential labels for a sentence. The labels from CDCC and CDCA are denoted as $Y^{c,c}$ and $Y^{c,a}$, respectively.

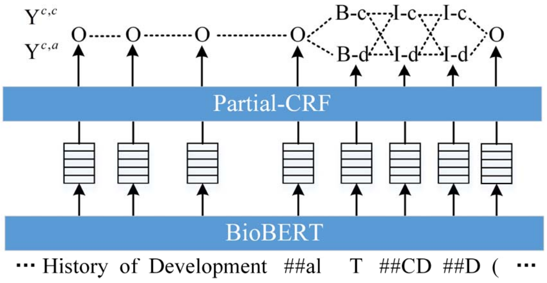

Fig. 2. The structure of our named entity recognition model with partial label integrating.

TABLE II. COMPARISON WITH SOME STATE-OF-THE-ART METHODS.

	Methods	CDR	CHEMDNER	NCBI disease
		F (%)	*F* (%)	*F* (%)
1	Habibi et al. [5]	87.63*	85.63	84.44
2	Luo et al. [1]	-	91.00	-
	Dang et al. [6]	89.30*	-	84.41
3	Wang et al. [7]	88.78	89.37	86.14
	Yoon et al. [8]	88.15*	88.85	86.36
4	Lee et al. [9]	90.60*	92.36	89.71
	Ours	**92.02**	**93.70**	**90.01**

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.

Specifically, given an input sentence and its two sets of labels $\{\mathbf{s}^w, \mathbf{y}^{c,c}, \mathbf{y}^{c,a}\}$, PCRF traverses two possible labels $\mathbf{y}^c$ for each token with different labels $\{s_i^w \mid y_i^c \in y_i^{c,c}, y_i^{c,a}\}$. The probability of possible two sets of labels $\mathbf{y}^c \in \{\mathbf{y}^{c,c}, \mathbf{y}^{c,a}\}$ (e.g. the dotted paths in Fig. 2) is computed as:

$$p(\mathbf{y}^c \mid \mathbf{s}^w) = \sum_{\mathbf{y}^c} \exp(score(\mathbf{s}^w, \mathbf{y}^c)) / \sum_{\mathbf{y}} \exp(score(\mathbf{s}^w, \mathbf{y})) \quad (3)$$

where $\mathbf{y}$ is all possible labeled sequences of the sentence $\mathbf{s}^w$, and the score function $score(\mathbf{s}, \mathbf{y})$ is the same as in CRF:

$$score(\mathbf{s}, \mathbf{y}) = \sum_{i=1}^{n} (T_{y_{i-1}, y_i} + P_{i, y_i}) \qquad (4)$$

where P_{i,y_i} is the score for labeling s_i with y_i, and T_{y_{i-1},y_i} is the transition score from label y_{i-1} to label y_i.

The loss function of the partial label integrating $L_{integrating}$ is to minimize the negative log-probability of two possible labeled sequences:

$$L_{integrating} = -\log p(\mathbf{y}^c \mid \mathbf{s}^w) \qquad (5)$$

III. EXPERIMENT AND DISCUSSION

A. Experimental Settings

Dataset and Evaluation Metrics: we conduct our experiments on the biomedical domain datasets CDR, CHEMDNER and NCBI disease.

Implementation Details: The dimension of character and noisy label embeddings is 50 and 20. RMSProp optimizer with learning rate 1e-3 and 5e-5 is used to minimize the loss of BLSTM and BioBERT, respectively.

B. Main Results

We compare our model with state-of-the-art methods on CDR, CHEMDNER and NCBI disease datasets. As shown in Table Ⅱ, we mainly divide these relevant models into four

TABLE III. ABLATION STUDY RESULTS ON CDR.

Model	P (%)	R (%)	F (%)
Ours	**92.02**	**92.02**	**92.02**
w/o correction	91.48	83.76	87.45
w/o CDCC (i.e. CDCA)	91.98	90.75	91.36
w/o CDCA (i.e. CDCC)	90.82	91.75	91.28
w/o curriculum learning	91.59	91.99	91.79

groups. Except the method proposed by Lee et al. [9] and our method, all the methods are based on BLSTM-CRF.

Habibi et al. [5] only use word and character features, while Luo et al. [1] and Dang et al. [6] further introduce additional features to enhance their models. From the results, we can see that rich features could improve the recognition performance, but designing such features is time-consuming and laborious.

Wang et al. [7] and Yoon et al. [8] employ multi-task learning to augment resources, which leads to some improvements.

Lee et al. [9] and our model take advantage of large-scale unlabeled datasets, which significantly outperforms other methods. Our model outperforms Lee et al. [9] on all the three datasets, which demonstrates corrected WL datasets are high-quality and partial label integrating is effective.

C. Ablation Studies

To better understand the function of key components, we conduct some ablation studies as shown in Table III.

Do we need to apply label correction strategy on the weakly labeled datasets? (w/o correction) Instead of using corrected datasets CDCC and CDCA, we use weakly labeled datasets CDWC and CDWA to train the partial label integrating model. Consequently, the recall drops significantly and the precision also drops. This demonstrates the label correction has high effectiveness.

Are both the datasets beneficial? (w/o CDCC and w/o CDCA) When we only use the dataset from one perspective, each performance drops but is still promising. In addition, CDCA achieves higher precision than CDCC, while CDCC achieves higher recall than CDCA. This suggests that datasets from two perspectives are both effective and complementary.

Do we need to adopt curriculum learning when training knowledge acquisition module? (w/o curriculum learning) Without curriculum learning, both the precision and recall drop, which proves the effectiveness of curriculum learning when training our knowledge acquisition module.

IV. CONCLUSION

In this paper, we address the problem of insufficient training corpus that BioNER suffers from. A novel label correction strategy is proposed to make full use of PubTator and knowledge bases to obtain large-scale high-quality datasets from the perspective of coverage and accuracy, respectively. Further, we introduce partial label integrating to compress the knowledge in the two datasets. Experiments show the effectiveness of our method.

In terms of further work, we would consider using multi-task learning to construct large-scale datasets for broader knowledge transfer.

ACKNOWLEDGMENT

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

REFERENCES

[1] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin and J. Wang, "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition," Bioinformatics, vol. 34, no. 8, pp. 1381–1388, 2018.

[2] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, and A. Valencia, "CHEMDNER: The drugs and chemical names extraction challenge," Journal of Cheminformatics, vol. 7, no.Suppl 1, pp. S1, 2015.

[3] C. H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wiegers, and Z. Lu, "Overview of the BioCreative V chemical disease relation (cdr) task," In Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, pp. 154–166, 2015.

[4] R. I. Doğan, R. Leaman, and Z. Lu, "NCBI disease corpus: a resource for disease name recognition and concept normalization," Journal of Biomedical Informatics, vol. 47 pp. 1–10, 2014.

[5] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," Bioinformatics, vol. 33, no. 14, pp. i37–i48, 2017.

[6] T. H. Dang, H.-Q. Le, T. M. Nguyen, and S. T. Vu, "D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information," Bioinformatics, vol. 34, no. 20, pp. 3539–3546, 2018.

[7] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, and J. Han, "Cross-type biomedical named entity recognition with deep multi-task learning," Bioinformatics, vol. 35, no. 10, pp. 1745–1752, 2019.

[8] W. Yoon, C. H. So, J. Lee, and J. Kang, "CollaboNet: collaboration of deep neural networks for biomedical named entity recognition," BMC Bioinformatics, vol. 20, no. Suppl 10, pp. 55–65, 2019.

[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.

[10] A. Ghaddar and P. Langlais, "WiNER: A Wikipedia annotated corpus for named entity recognition," In Proceedings of the Eighth International Joint Conference on Natural Language Processing, vol. 1, pp. 413–422, 2017.

[11] Y. Cao, Z. Hu, T. S. Chua, Z. Liu, and H. Ji, "Low-resource name tagging learned with weakly labeled data," In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, pp. 261–270, 2019.

[12] O. Täckström, D. Das, S. Petrov, R. McDonald, and J. Nivre, "Token and type constraints for cross-lingual part-of-speech tagging," Transactions of the Association for Computational Linguistics, vol. 1, pp. 1–12, 2013.

[13] M. Zhu, Z. Deng, W. Xiong, M. Yu, M. Zhang, and W. Y. Wang, "Towards open-domain named entity recognition via neural correction models," arXiv preprint arXiv:1909.06058, 2019.

[14] C. J. Mattingly, G. T. Colby, and John N. Forrest, and James L. Boyer, "The comparative toxicogenomics database (CTD)," Environmental Health Perspectives, vol. 111, no. 6: pp.793–795, 2003.

[15] C. E. Lipscomb, "Medical subject headings (MeSH)," Bull med libr assoc, vol. 88, no. 3, pp.265–266, 2000.

[16] R. Nigam, S. J. F. Laulederkind, G. T. Hayman, J. R. Smith, S. J. Wang, T. F. Lowry et al, "Rat genome database: a unique resource for rat, human, and mouse quantitative trait locus data," Physiological Genomics, vol. 45, no. 18, pp.809–816, 2013.

[17] C. H. Wei, K. Lee, R. Leaman, and Z. Lu, "Biomedical mention disambiguation using a deep learning approach," In 10th ACM International Conference, pp.307–313, 2019.

[18] C. H. Wei, H. Y. Kao, and Z. Lu, "PubTator: a web-based text mining tool for assisting biocuration," Nucleic Acids Research, vol. 41, no. W1, pp. W518–W522, 2013.

[19] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," In Proceedings of 26th Annual International Conference on Machine Learning, pp.41–48, 2009.