Fine-Grained Named Entity Recognition with Distant Supervision in COVID-19 Literature

Xuan Wang[1], Xiangchen Song[1], Bangzheng Li[1], Kang Zhou[2], Qi Li[2], Jiawei Han[1]
[1]Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA
[2]Department of Computer Science, Iowa State University, IA, USA
[1]{xwang174,xs22,bl17,hanj}@illinois.edu, [2]{kangzhou,qli}@iastate.edu,

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313126

***Abstract*—Biomedical named entity recognition (BioNER) is a fundamental step for mining COVID-19 literature. Existing BioNER datasets cover a few common coarse-grained entity types (e.g., genes, chemicals, and diseases), which cannot be used to recognize highly domain-specific entity types (e.g., animal models of diseases) or emerging ones (e.g., coronaviruses) for COVID-19 studies. We present CORD-NER, a fine-grained named entity recognized dataset of COVID-19 literature (up until May 19, 2020). CORD-NER contains over 12 million sentences annotated via distant supervision. Also included in CORD-NER are 2,000 manually-curated sentences as a test set for performance evaluation. CORD-NER covers 75 fine-grained entity types. In addition to the common biomedical entity types, it covers new entity types specifically related to COVID-19 studies, such as coronaviruses, viral proteins, evolution, and immune responses. The dictionaries of these fine-grained entity types are collected from existing knowledge bases and human-input seed sets. We further present DISTNER, a distantly supervised NER model that relies on a massive unlabeled corpus and a collection of dictionaries to annotate the COVID-19 corpus. DISTNER provides a benchmark performance on the CORD-NER test set for future research.**

***Index Terms*—fine-grained named entity recognition; distant supervision; COVID-19**

I. INTRODUCTION

COVID-19 is an infectious disease that was first identified in December 2019 and has since spread globally, resulting in the 2019–2020 coronavirus pandemic. Scholarly literature about COVID-19, SARS-CoV-2, and the coronavirus group has been pouring into the COVID-19 Open Research Dataset (CORD-19) [9] just in the past few months. It is critical to automatically extract the most relevant and accurate information from this large-scale and fasting growing COVID-19 literature corpus to facilitate COVID-19 studies.

Biomedical named entity recognition (BioNER) is a fundamental step for mining COVID-19 literature. Existing BioNER datasets (e.g., BC5CDR [10], JNLPBA [2], and BIONLP13CG [6]) cover a few common coarse-grained entity types (e.g., genes, chemicals, and diseases), which cannot be used to recognize highly domain-specific (e.g., animal models of diseases) or emerging entity types (e.g., coronaviruses) for COVID-19 studies.

We present CORD-NER, a fine-grained named entity recognized dataset of COVID-19 literature (up until May 19, 2020). CORD-NER contains over 12 million sentences annotated via distant supervision. Also included in CORD-NER are 2,000 manually-curated sentences as a test set for performance evaluation. CORD-NER covers 75 fine-grained entity types. In addition to the common biomedical entity types, it covers new entity types specifically related to COVID-19 studies, such as coronaviruses, viral proteins, evolution, and immune responses. These fine-grained entity types are highly related to research on COVID-19 related virus, spreading mechanisms, and potential vaccines. The dictionaries of these fine-grained entity types are collected from existing knowledge bases and human-input seed sets.

We further present DISTNER, a distantly supervised NER model that relies on the massive unlabeled corpus and dictionaries to annotate the COVID-19 corpus. DISTNER achieves high performance with dictionaries of different scales (from dozens to thousands of entities). It leverages a dictionary-guided representation learning model to expand the small dictionaries and further incorporates the newly-learned word embeddings into a NER neural model training. DISTNER automatically annotates the COVID-19 corpus with high quality and provides a benchmark performance on the CORD-NER test set for future research. Based on the DISTNER model, CORD-NER allows adding new documents as well as new entity types when needed by adding dozens of seeds as the input examples. CORD-NER can help the NLP community for downstream applications, such as relation extraction, knowledge graph construction, and information retrieval, in COVID-19 literature.

II. CORD-NER DATASET

In this section, we first introduce how we collected the input corpus and the fine-grained entity type dictionaries for CORD-NER. Then we introduce DISTNER, the distantly supervised NER model used to annotate the input corpus.

A. Corpus

The input corpus is generated from the CORD-19 dataset (up until May 19, 2020). We first combined the title and abstract of each paper in the meta-data file with their corresponding full-text from all the data sources (CZI, PMC, bioRxiv, and medRxiv) in CORD-19. This input corpus contains 12,698,615 sentences from 128,492 documents. Then we conducted automatic phrase mining and tokenization on the input corpus using AutoPhrase [7]. This tokenized corpus

is used for further NER annotations. We observed that incorporating the AutoPhrase tokenization results can improve the distantly supervised NER performance as it provides additional information for entity boundary detection.

B. Fine-Grained Entity Type Dictionaries

For each entity type to be annotated, we collect a dictionary containing a list of entities belonging to that type.

Existing Knowledge Bases. We use UMLS[1] knowledge base to collect the large-scale dictionaries. We collect the latest version of UMLS (the year 2020) that contains 127 fine-grained entity types. We further merged some fine-grained types into their more coarse-grained parent types according to the corpus counts and suggestions from domain experts. It results in 48 fine-grained types in UMLS used for our entity annotation. Each UMLS type includes thousands of entities as the input dictionary.

Human-Input Seed Sets. In addition to the types in UMLS, biomedical scientists and medical doctors are interested in some additional entity types specifically related to COVID-19 studies. These types are either new or too specific that have not been incorporated in the UMLS knowledge base. We included nine new types (coronaviruses, viral proteins, livestocks, wildlifes, evolution, physical science, substrates, materials, and immune responses) defined by the scientists and doctors. For each new type, the scientists and doctors provide 20 seed entities as the input dictionary.

C. Distantly Supervised NER Model

Based on the entity type dictionaries we collected in different scales (from dozens to thousands of entities), we propose DISTNER, a distantly supervised NER model that can automatically annotate the CORD-19 corpus.

Dictionary-Guided Representation Learning. The first step of DISTNER is dictionary-guide embedding learning. It takes the input dictionaries (Section II-B) as weak supervision and jointly embeds the entities, types and words into a shared space. The entities and types are from the input dictionaries (Section II-B). The words are from the input corpus (Section II-A). Note that the words here also include the phrases that we previously discovered during corpus tokenization.

To achieve the goal of making the words form discrete clusters around the types, we learn the joint embedding of entities, types and words by satisfying two criteria: *Coherence* and *Discriminativeness*. Coherence means that the entities should have embeddings that are close to their corresponding types' embeddings. Discriminativeness means that the embeddings of different types should be far apart from each other. Inspired by CatE [3], a category-guided embedding learning method, we first formulate a joint type and text generative process under the guidance of the input dictionary. Then we cast the learning of the generative process as a dictionary-guided embedding learning model.

[1]https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

The input to our dictionary-guided embedding learning model consists of two parts: (1) a set of dictionaries $\{D_t\}$, where each dictionary $\mathcal{D}_t = \{e_1, e_2, , \ldots, e_{|\mathcal{D}_t|}\}$ contains entities e for the type $t \in \mathcal{T}$, and (2) a text corpus containing sentences $s = [w_1, w_2, \ldots, w_{|s|}]$, where each sentence consists of words and entities that can be matched to the sentence. For ease of notation, we use w to denote both words and entities in the sentence.

We assume a joint type and text generative process in two steps: (1) each type t is generated conditioned on the semantics of the entities e in the dictionary $\mathcal{D}_t$; and (2) surrounding words and entities $\mathcal{C}(w_i, h)$ of a word/entity w_i in a sentence s are generated conditioned on the semantics of the center word/entity w_i, where $\mathcal{C}(w_i, h) = \{w_j : i-h \leq j \leq i+h, i \neq j\}$, h is the context window size. Putting the above two steps together, we have the following expression for the likelihood of the joint type and text generative process:

$$\mathcal{J} = \prod_t \prod_{e \in \mathcal{D}_t} p(t|e) \cdot \prod_s \prod_{w_i \in s} p(\mathcal{C}(w_i, h)|w_i).$$

The first part $\prod_t \prod_{e \in \mathcal{D}_t} p(t|e)$ of the likelihood J indicates the probability of observing all the types (e.g., "Coronavirus") given the entities (e.g., "SARS" and "MERS") in our input dictionaries. The second part $\prod_s \prod_{w_i \in s} p(\mathcal{C}(w_i, h)|w_i)$ of the likelihood J indicates the probability of observing the input corpus.

$$\begin{aligned} \mathcal{L} = &- \sum_t \sum_{e \in \mathcal{D}_t} log(p(t|e)) \quad (\mathcal{L}_{\text{type}}) \\ &- \sum_s \sum_{w_i \in s} \sum_{w_j \in \mathcal{C}(w_i, h)} log(p(w_j|w_i)). \end{aligned} \tag{1}$$

Then we formulate the optimization of the objective in Eq. (1) as an embedding learning problem. Similar to [4], we define the two conditional probabilities in Eq. (1) via log-linear models in the embedding space:

$$p(t|e) = \frac{\exp(\boldsymbol{t}^T \boldsymbol{e})}{\sum_{t' \in \mathcal{T}} \exp(\boldsymbol{t}'^T \boldsymbol{e})}, \tag{2}$$

$$p(w_j|w_i) = \frac{\exp(\boldsymbol{w}_j^T \boldsymbol{w}_i)}{\sum_{w'_j \in \mathcal{C}(w_i, h)} \exp(\boldsymbol{w}_j'^T \boldsymbol{w}_i)}, \tag{3}$$

where $\boldsymbol{t}$ is the embedding vector of the type t; $\boldsymbol{e}$ is the embedding vector of the entity e; and $\boldsymbol{w}$ is the embedding vector of the word or entity w.

Eqs. (2) and (3) can be directly plugged into Eq. (1) to train the joint type and text embeddings. To this end, we have enforced the first *Coherence* criterion in Eq. (2). Then we show how to satisfy the second *Discriminativeness* criterion. Let $\boldsymbol{p}_e = [p(t_1|e), ..., p(t_{|\mathcal{T}|}|e)]$ be the probability distribution of e over all types. To satisfy the second *Discriminativeness* criterion, if an entity e is known to belong to type t, $\boldsymbol{p}_e$ computed from Eq. (2) should become a one-hot vector $\boldsymbol{l}_e$ (i.e., the type of e) with $p(t|e) = 1$. To achieve this property, we minimize the KL divergence from each seed entity's distribution $p(t|e)$ to its corresponding discrete delta

distribution $\boldsymbol{l}_e$. Formally, given a dictionary of seed entities $\mathcal{D}_t$ for type t, the first term in Eq. (1) is implemented as:

$$\mathcal{L}_{\text{type}} = \sum_t \sum_{e \in \mathcal{D}_t} KL(\boldsymbol{l}_e \| \boldsymbol{p}_e) \quad (4)$$

From the embedding learning perspective, Eq. (4) is equivalent to a cross-entropy regularization loss, encouraging the type embeddings to become discriminative in the embedding space and are far apart from each other.

Finally, based on the newly-learned representations of the types and words, we expand each type's dictionary with the words that have high embedding cosine similarity (≥ 0.5) with its type embedding. Note that the words here also include the phrases that we previously discovered with AutoPhrase during corpus tokenization. We further incorporate the newly-learned word embeddings into the NER neural model training.

NER Neural Model. We adopt the AutoNER [8] neural model as the benchmark distant NER model on the CORD-19 corpus. The neural model learning is divided into two steps: entity span detection and entity typing.

For entity span detection, a binary classifier is built to determine whether a connection between two adjacent tokens should be labeled as *Break* or *Tie*. A BiLSTM layer is utilized to encode the character and word embeddings (learned from the Dictionary-Guided Representation Learning step) to predict whether the connection y_i between tokens w_{i-1} and w_i is *Break*. Then the output of the BiLSTM layer will be concatenated as one vector $\boldsymbol{u}_i$ and fed into a Sigmoid layer:

$$p(y_i = Break|\boldsymbol{u}_i) = \sigma(\boldsymbol{w}^T \boldsymbol{u}_i).$$

where y_i is the label between the i-th and its previous tokens, σ is the sigmoid function, and $\boldsymbol{w}$ is the sigmoid layer's parameter. The loss function of entity span detection:

$$\mathcal{L}_1 = \sum_{y_i = Break} l(y_i, p(y_i = Break|\boldsymbol{u}_i)),$$

where $l(\cdot, \cdot)$ is the logistic loss.

After the entity boundary is determined, each candidate entity span (tokens within two adjacent *Break*) is represented with a new vector $\boldsymbol{v}_j$ and fed into a Softmax layer to determine its entity type:

$$p(t_j = t|\boldsymbol{v}_j) = \frac{\exp(\boldsymbol{t}_j^T \boldsymbol{v}_j)}{\sum_{t' \in \mathcal{T}'} \exp(\boldsymbol{t}'^T \boldsymbol{v}_j)},$$

where t_j is the label of candidate entity span j and $\mathcal{T}' = \mathcal{T} \cup \{\text{None}\}$. The loss function of entity type prediction:

$$\mathcal{L}_2 = \sum_j H(\hat{p}(\cdot|\boldsymbol{v}_i, \mathcal{T}'), p(\cdot|\boldsymbol{v}_i)),$$

where $H(\cdot, \cdot)$ is the cross entropy function and $\hat{p}(\cdot|\boldsymbol{v}_i, \mathcal{T})$ is the supervision distribution.

III. Evaluation

Experimental Setup. Given the input corpus and the expanded dictionaries, we first conduct exact string matching [8] on a subset corpus of 3,000,000 sentences to generate a distantly labeled training corpus. Conflicted matches are resolved by maximizing the total number of matched tokens on each sentence. We split the distantly labeled training corpus into 9:1 for training and development. We randomly selected another 2,000 sentences from our input corpus and asked domain experts for manual annotation. We use this manually-annotated test set to compare the performance of different BioNER models on the CORD-19 corpus. We compare DistNER with AutoNER [8], the benchmark method for distantly supervised BioNER. We also compare DistNER with pre-trained supervised BioNER models, such as SciSpacy [5], a commercial supervised BioNER tool, and SciBERT [1], a benchmark method for supervised BioNER. We report the precision, recall, and F1 scores[2] of each method on our human-annotated test set.

Test Set Annotation. Three domain experts annotated each sentence. Due to a large number of fine-grained entity types, we only annotated 7 out of the 75 types in this test set for evaluation and resulted in 2,000 annotated sentences. The seven types include genes, chemicals, diseases, signs or symptoms, coronaviruses, evolution, and immune responses. Each pair of annotators reach a substantial agreement with a Fleiss's κ of 0.72.

Parameters. We used PyTorch for model implementations. For the baseline model AutoNER, we use 200-dimension word embeddings[3] trained on the entire Pubmed database of abstracts and full-text articles together with the Wikipedia corpus. For DistNER, we use the dictionary-guided word embeddings learned using our dictionary-guided representation learning model. The DistNER neural model parameter settings are the same as AutoNER. The character embedding dimension is 30, and the hidden state size for both the character-level BiLSTM and word-level BiLSTM is 300. The optimization method is gradient descent with momentum. The batch size and the momentum are set to be 10 and 0.9. The learning rate is set to 0.05. The dropout ratio is set to 0.5. For better stability, a gradient clipping of 5.0 is used.

Results. Table I shows the performance comparison of DistNER and AutoNER, the benchmark distantly supervised BioNER model. We use the original implementation of AutoNER[4] and trained the model on our distantly labeled training corpus. Then we evaluate the performance of DistNER and AutoNER on our test set. DistNER outperforms AutoNER by a large margin on the F1 scores. The performance gain is more significant when the input dictionary is small (e.g., dictionaries contain 20 seed entities used for types such as coronavirus, evolution, and immune response).

Table I also shows the performance breakdown of DistNER on the ablation models. DistNER $_{\text{w/o Emb}}$ uses Word2Vec embeddings and the expanded dictionary. DistNER $_{\text{w/o Exp}}$ uses our dictionary-guided word embeddings and

[2]https://github.com/chakki-works/seqeval
[3]http://bio.nlplab.org/
[4]https://github.com/shangjingbo1226/AutoNER

Supervision	UMLS (13K entities)			Seed Set (20 entities)			Seed Set (20 entities)			Seed Set (20 entities)		
Type	Sign or Symptom			Coronavirus			Evolution			Immune Response		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
AutoNER	63.18	72.54	67.54	47.70	24.21	32.12	43.11	75.78	54.96	13.22	21.70	16.43
DISTNER$_{\text{w/o Emb}}$	60.29	77.85	67.95	46.18	56.56	50.84	17.22	80.62	28.38	6.52	24.53	10.30
DISTNER$_{\text{w/o Exp}}$	**74.13**	75.42	74.77	**74.72**	75.33	**75.03**	**92.59**	78.74	85.11	**94.12**	90.57	92.31
DISTNER	72.07	**78.69**	**75.23**	72.81	**76.96**	74.83	91.87	**88.98**	**90.40**	92.66	**95.28**	**93.95**

TABLE I: Performance of DISTNER and AutoNER, the benchmark distantly supervised BioNER model, evaluated on our manually-annotated test set. We also show the performance of our ablation models.

Model	Supervision	Chemical			Disease or Syndrome			Gene or Genome		
		Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
SciSpacy	Human (BIONLP13CG)	55.46	36.74	44.20	54.55	3.75	7.02	21.66	**80.99**	34.18
SciSpacy	Human (BC5CDR)	**78.37**	57.35	66.23	**73.49**	61.25	66.81	-	-	-
SciBERT	Human (BC5CDR)	68.24	61.94	64.94	62.25	59.29	60.73	-	-	-
DISTNER	Dictionary (UMLS)	73.32	**65.9**	**69.41**	69.21	**70.55**	**69.87**	**57.54**	60.05	**58.77**

TABLE II: Performance of DISTNER and fully-supervised BioNER models on our manually-annotated test set.

the original input dictionary. We see that both the dictionary expansion and the dictionary-guided word embeddings help improve the DISTNER performance compared to AutoNER. The dictionary-guided word embeddings (DISTNER $_{\text{w/o Exp}}$) bring a more significant performance improvement compared to dictionary expansion (DISTNER $_{\text{w/o Emb}}$). The dictionary-guided word embeddings (DISTNER $_{\text{w/o Exp}}$) improve both the precision and recall significantly, while the expanded dictionary (DISTNER $_{\text{w/o Emb}}$) introduces an increase in recall but a decrease in precision compared to AutoNER.

Table II shows the performance comparison between DISTNER and the fully-supervised BioNER models, SciSpacy and SciBERT. For SciSpacy, we use its published pre-trained models[5] on both BIONLP13CG [2] and BC5CDR [10]. For SciBERT, since it does not release its pre-trained models, we use its SciBERT embeddings[6] and re-trained the model on BC5CDR. Then we conduct prediction and evaluation on our test set. DISTNER shows better performance on chemical and disease prediction compared to both SciSpacy and SciBERT due to a higher recall. DISTNER also shows better performance for gene prediction compared with SciSpacy trained on BIONLP13CG. We observe that SciSpacy tends to predict most coronaviruses as genes, leading to a very low precision.

IV. CONCLUSION

We present CORD-NER, a fine-grained named entity recognized dataset of COVID-19 literature (up until May 19, 2020). CORD-NER contains over 12 million sentences annotated via distant supervision. Also included in CORD-NER are 2,000 manually-curated sentences as a test set for performance evaluation. We further present DISTNER, a distantly supervised NER model that is used to annotate the COVID-19 corpus. DISTNER provides a benchmark performance on the CORD-NER test set for future research. CORD-NER can help other downstream NLP tasks for COVID-19 studies, such as relation extraction, knowledge graph construction, and information retrieval.

[5]https://allenai.github.io/scispacy/

[6]https://github.com/allenai/scibert

ACKNOWLEDGMENT

Research was sponsored in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and SocialSim Program No. W911NF-17-C-0099, National Science Foundation IIS-19-56151, IIS-17-41317, IIS 17-04532, and IIS 16-18481, and DTRA HDTRA11810026. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation hereon.

REFERENCES

[1] I. Beltagy, K. Lo, and A. Cohan. Scibert: Pretrained language model for scientific text. In *EMNLP*, 2019.

[2] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Citeseer, 2004.

[3] Y. Meng, J. Huang, G. Wang, Z. Wang, C. Zhang, Y. Zhang, and J. Han. Discriminative topic mining via category-name guided text embedding. In *Proceedings of The Web Conference 2020 (WWW20)*, 2020.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119. MIT Press, 2013.

[5] M. Neumann, D. King, I. Beltagy, and W. Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[6] S. Pyysalo, T. Ohta, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, J. Tsujii, and S. Ananiadou. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. *BMC bioinformatics*, 16(S10):S2, 2015.

[7] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837, 2018.

[8] J. Shang, L. Liu, X. Ren, X. Gu, T. Ren, and J. Han. Learning named entity tagger using domain-specific dictionary. In *EMNLP*. ACL, 2018.

[9] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al. Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*, 2020.

[10] 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*, volume 14, 2015.