Named Entity Recognition in COVID-19 tweets with Entity Knowledge Augmentation

Anonymous ACL submission

Abstract

The COVID-19 pandemic causes severe social and economic disruption around the world, raising various subjects that are discussed or argued over on social media. Identifying pandemic-related named entities as expressed on social media is fundamental and important for understanding the discussions on the pandemic. However, there is limited work on named entity recognition on this topic due to the following challenges: 1) annotated data is rare and insufficient to train a robust recognition model, and 2) named entity recognition in COVID-19 requires extensive knowledge of the pandemic. To address this, we propose a novel entity knowledge augmentation for named entity recognition systems in COVID-19 tweets. Experiments carried out on the COVID-19 tweets dataset show that our proposed entity knowledge augmentation improves NER performance, achieving an F1 score of 84.08.

1 Introduction

The COVID-19 pandemic has led to significant social and economic upheaval globally, sparking various topics of conversation and debate on social media. Identifying pandemic-related entities mentioned on social media is crucial for comprehending these discussions. Most existing named entity recognition datasets (Tjong Kim Sang and De Meulder, 2003; Pradhan et al., 2013; Strauss et al., 2016; Hou et al., 2020; Jiang et al., 2022) are not created with a focus on COVID-19 or public health research, making it difficult for epidemiologists to use them for analyzing COVID-19 topics.

From a public health research standpoint, Zhou et al. (2022a) released METS-CoV, a dataset of COVID-19 tweets annotated with seven types of entities, including four medical entity types (Disease, Drug, Symptom, and Vaccine) and three general entity types (Person, Location, and Organization). Given this dataset, a COVID-19 NER model can be designed and trained. However, the models on these benchmarks have limited performance due to the following challenges: 1) annotated data is rare and insufficient to train a robust recognition model, and 2) named entity recognition in COVID-19 requires extensive medical knowledge of the pandemic.

Recently, the widespread success of large language models in various text processing tasks has ushered in a new training paradigm. Recent models based on large language models demonstrate superiority in named entity recognition (Meoni et al., 2023; Sharma et al., 2023). We base our COVID-19 NER models on large language models, leveraging their superior text representation capabilities. We propose a **LLM**-based **E**ntity **K**nowledge **A**ugmentation (**LLM-EKA**) to enrich the COVID-19-related knowledge of the models. The proposed knowledge augmentation can be decoupled to enhance other domain-specific NER models.

The framework of LLM-EKA, as shown in Figure 1, consists of demonstration selection, entity augmentation, and instance augmentation. The demonstration selection aims to extract informative examples from the training data, and the extracted examples are used as demonstrations for NER model training. The entity augmentation is applied to obtain domain-specific entities via pretrained language models. The instance augmentation generates domain-specific training instances via prompts according to the selected demonstrations and augmented domain-specific entities.

The experiments carried out on the benchmark METS-CoV show that the NER models equipped with the proposed LLM-EKA outperform the baseline model by obtaining an F1 score of 84.10. The main contributions of this work are summarized as follows:

- We investigate named entity recognition in COVID-19 tweets from medical research per-

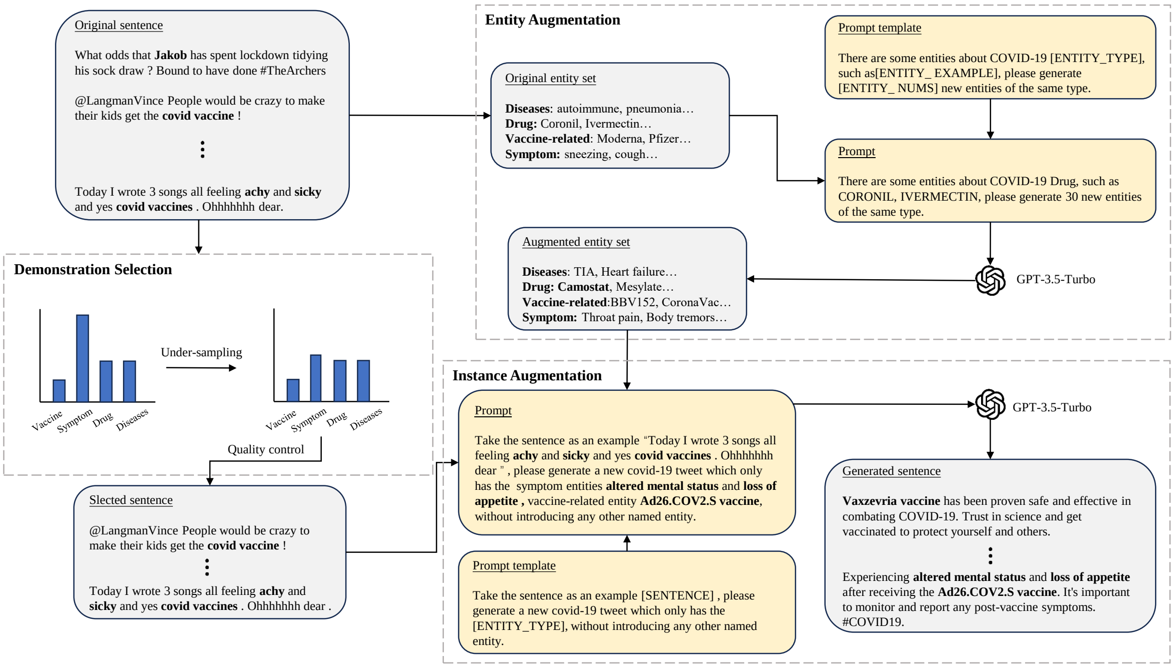

Figure 1: Framework of the LLM-based Entity Knowledge Augmentation

spectives that contribute to public health concerns.

- We propose a novel framework of entity knowledge augmentation for named entity recognition in COVID-19 tweets. The proposed method can be decoupled to enhance other named entity recognition models.
- Our final model, equipped with the proposed entity knowledge augmentation, achieves state-of-the-art results on the standard benchmark. The code is released at `https://anonymous`.

2 Related Work

Named Entity Recognition is widely investigated as a conventional task in NLP. Recent work has focused on specific domains, which are often limited to small-scale training data.

Domain Transfer Domain transfer aims to alleviate the data scarcity issue by transferring knowledge from source domains to target domains. This line of research on NER enhances the generalization of models by aligning the entity knowledge of source domains with the target domains (Daumé III, 2007; Kim et al., 2015; Lee et al., 2018; Lin and Lu, 2018; Wang et al., 2018; Yang et al., 2018; Jia et al., 2019; Wang et al., 2020; Liu et al., 2020). Moreover, some related works (Cui et al., 2021; Ma et al., 2022; Ding et al., 2022; Chen et al., 2023; Fang et al., 2023) have been done in few-shot settings. However, these methods are based on the assumption that the label spaces between the source domain and target domain are aligned.

Data Augmentation Data augmentation methods aim to directly expand the scale of training data to alleviate the data scarcity in low-resource domains. This line of research on NER focuses on rewriting and generating training instances (Dai and Adel, 2020; Ding et al., 2020; Zhou et al., 2022b). Moreover, Ye et al. (2024) leverage LLMs to generate a large quantity of diverse and high-quality new data for NER. However, these methods suffer from a lack of domain-specific knowledge and are not well-suited to NER in domains that require extensive knowledge.

3 Methodology

Named entity recognition is modeled as a sequence labeling task, where the input is a sequence of words, $X = [x_1, x_2, \ldots, x_n]$, and the output is a sequence of labels, $Y = [y_1, y_2, \ldots, y_n]$, where n is the length of the input sentence. We use the BIOES labeling system for NER.

3.1 Named Entity Recognition Model

We use a pre-trained language model to obtain the hidden representation, $H = [h_1, h_2, \ldots, h_n]$ for the input sentence, $X = [x_1, x_2, \ldots, x_n]$, where h_i is the hidden representation of i-th word x_i:

$$H = [h_1, h_2, \ldots, h_n] = \text{LLM}(x_1, x_2, \ldots, x_n). \tag{1}$$

The hidden representation h_i is transformed into the logits o_i using a linear layer:

$$o_i = Wh_i + b, \tag{2}$$

where W and b are tunable parameters. The label y_i is predicted by applying the $argmax$ function over the logits:

$$y_i = \arg\max_c o_{i,c} \tag{3}$$

where c indexes over the possible labels.

In training, we employ the cross-entropy loss function to measure the discrepancy between the predicted logits and the true labels:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid x_i) \tag{4}$$

where $p(y_i \mid x_i)$ is the probability assigned to the true label by the model, which is obtained via a softmax operation over the logits:

$$p(y_i = c \mid x_i) = \frac{\exp(o_{i,c})}{\sum_{k=1}^{C} \exp(o_{i,k})} \tag{5}$$

where C is the number of possible labels.

3.2 LLM-based Entity Knowledge Agumentation

The framework of LLM-EKA consists of demonstration selection, entity augmentation, and instance augmentation.

Demonstration Selection We select demonstration sentences from the training data, aiming to ensure that the demonstrations are domain-specific representatives containing COVID-19 entities, such as symptoms, drugs, vaccine-related entities, and diseases. We choose sentences that contain only domain-specific entities and filter out sentences that contain only general entity types (Person, Location, and Organization). Additionally, we rule out the sentences that have more than 6 entities by quality control, because these sentences are so complex that they can degrade the LLM generation. Table 1 shows the entity distribution which has the significant entity imbalance issue. We applied under-sampling methods to mitigate the bias towards majority class entities and achieve a more balanced data distribution.

Category	**Count**	**Ratio**
General Entity	3,108	56.32%
Person	1,387	25.13%
Location	608	11.02%
Organization	1,113	20.17%
Domain-specific Entity	2,411	43.68%
Vaccine-related	356	6.45%
Symptom	1,204	21.81%
Drug	428	7.76%
Disease	423	7.66%

Table 1: The entity distribution across categories in METS-CoV.

Entity Augmentation We augment the entities by leveraging the capabilities of large language models, aiming to enrich knowledge of COVID-19 and generate new domain-specific entities. We design targeted prompts based on original domain-related entities. The prompt template is “There are some entities about COVID-19 [ENTITY_TYPE] such as [ENTITY_EXAMPLE]. Please generate [ENTITY_NUMS] new entities of the same type." Given the prompt template, we generate prompts according to training data entities. Taking symptoms as an example, the generated prompt is: "There are some entities about COVID-19 symptoms such as coughing up blood, burning sensation in lungs, and head hurts. Please generate 30 new entities of the same type." We use the generated prompts to query GPT-3.5-Turbo and obtain the corresponding augmented entities.

Instance Augmentation We generate new COVID-19 tweets by querying GPT-3.5-Turbo using a prompt template “Take the sentence as an example [SENTENCE], please generate a new COVID-19 tweet which only has the [ENTITY], without introducing any other named entity." The prompts have demonstration slot, [SENTENCE], filled by sentences selected from the demonstration selection, aiming to guide LLM to generate tweets with high quality and consistency in structure and style. The entity slot, [ENTITY], are filled by domain-specific entities obtained from the entity augmentation. We merge the augmented data with original training data to train the NER model.

Models	Location	Organization	Person	Symptom	Vaccine-related	Disease	Drug	Average
CRF	76.37±0.62	54.64±2.08	64.43±1.59	74.05±0.56	84.85±0.82	73.61±0.44	77.34±1.60	71.58±0.54
WLSTM + CCNN + CRF	82.15±0.44	62.79±0.91	81.38±0.44	78.12±0.51	89.11±0.36	76.12±0.76	80.41±0.58	78.10±0.19
RoBERTa-lagrge	85.85±2.12	73.78±0.72	86.79±0.44	81.32±0.67	90.42±1.12	76.84±0.57	86.79±0.78	82.55±0.27
COVID-TWITTER-BERT	85.68±0.92	76.27±0.64	**91.29±0.42**	81.85±0.53	90.44±0.94	**77.48±0.81**	86.35±0.96	83.88±0.20
LLM-DA (Ye et al., 2024)	85.78±0.66	75.16±0.51	89.82±0.95	80.88±0.38	88.90±0.70	76.75±0.72	86.49±0.55	82.96±0.17
LLM-EKA-1000	**86.33±0.57**	76.98±0.27	90.62±0.58	**82.16±0.21**	90.46±1.05	77.46±0.88	**87.12±0.95**	**84.08±0.10**
LLM-EKA-all	85.30±0.50	**78.69±0.24**	90.88±0.40	82.08±0.10	**90.65±0.80**	77.37±1.11	86.16±0.48	84.05±0.05

Table 2: The results on test data in the standard benchmark. LLM-EKA-n is the COVID-TWITTER-BERT equipped with the proposed knowledge augmentation method, where n is the number of augmented training instances and all means that we use the all augmented instances. The bold are the best results.

4 Experiments

The experiments are conducted on the METS-CoV benchmark (Zhou et al., 2022a).

4.1 Settings

We set the temperature to 1 to fully leverage the diversity generation capabilities of the GPT-3.5-turbo model, leading to more varied and diverse outputs. We set the batch size to 8 and employ the AdamW optimizer with a learning rate of 3e-5. The models are trained for 100 epochs, and the best performing model on the validation set is selected for final testing. We use the micro F1 score as evaluation metrics. We use CRF and WSLTM + CCN + CRF to represent traditional NER models, and RoBERTa-large as a pre-trained language model. Additionally, COVID-TWITTER-BERT serves as our baseline due to its capability in representing COVID-19 tweets. COVID-TWITTER-BERT equipped with our entity knowledge augmentation is our final model.

4.2 Results

Table 2 shows the results on the test data across different models. Pre-trained language models significantly outperforms the traditional LSTM models with CRF. COVID-TWITTER-BERT outperforms RoBERTa because it has the ability to represent the COVID-19 tweets with help of the pre-training on tweets. Equipped with our proposed entity knowledge augmentation, the final model achieves the best results on the benchmarks.

4.3 Analysis

The Scale of Augmented Instances We experiment with different amounts of augmented samples, 1000 and 1831 samples (all), and merge them with the training data.Table 2 shows that using 1000 augmented samples, the model can achieves the marginal improvement with low variance. However, as the size of the augmented instance increase, the performance of models have no further improvements.

Comparison with Data Augmentations We reimplement LLM-DA method to obtain the same scale of augmented instance for comparison to our method. Table 2 shows that LLM-DA, with an average score of 82.96, underperforms compared to our final model. A possible reason is that LLM-DA is designed for few-shot entity recognition, and as the number of training instances increases, the additional context appears to introduce noise, adversely affecting performance. However, our method demonstrates marginal improvements across different sample sizes, indicating a more robust approach to handling augmented data.

Entity Type We perform a fine-grained analysis of the performance of different entity types across models. The models built on COVID-TWITTER-BERT has significant improvement in recognizing person entities. One reason for this is that person names in tweets are user IDs, which are difficult for models to recognize without pre-training on tweets. The final models achieve improvements in recognizing symptom and vaccine-related entities.

5 Conclusion

We present a novel LLM-based entity knowledge augmentation for named entity recognition in COVID-19 tweets for public health research. LLM-EKA leverages the sophisticated contextual reasoning capabilities and extensive knowledge base of LLMs to augment entity knowledge and improve recognition performance. Our proposed methods are general and can be decoupled to enhance other domain-specific NER models. The experimental results demonstrate that LLM-EKA is capable of addressing the challenges associated with scarce annotated data and the need for domain-specific expertise in COVID-19 NER tasks.

Limitations

The proposed LLM-based entity knowledge augmentation adopt the GPT-3.5-Turbo. So the performance of the final models are limited to the outcomes of the GPT-3.5-Turbo.

Ethnic Statement

The instances generated by querying GPT-3.5-Turbo is controlled by the prompts. Therefore, we foresee no ethical concerns in this work.

References

Yanru Chen, Yanan Zheng, and Zhilin Yang. 2023. Prompt-based metric learning for few-shot NER. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7199–7212, Toronto, Canada. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. OpenPrompt: An open-source framework for prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113, Dublin, Ireland. Association for Computational Linguistics.

Jinyuan Fang, Xiaobin Wang, Zaiqiao Meng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. MANNER: A variational memory-augmented model for cross domain few-shot named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4261–4276, Toronto, Canada. Association for Computational Linguistics.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393, Online. Association for Computational Linguistics.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the Tweebank corpus on named entity recognition and building NLP models for social media analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7199–7208, Marseille, France. European Language Resources Association.

Young-Bum Kim, Karl Stratos, Ruhi Sarikaya, and Minwoo Jeong. 2015. New transfer learning techniques for disparate label sets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 473–482, Beijing, China. Association for Computational Linguistics.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2012–2022, Brussels, Belgium. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020. Zero-resource cross-domain named entity recognition. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 1–6, Online. Association for Computational Linguistics.

Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. 2022. Decomposed meta-learning for few-shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1584–1596, Dublin, Ireland. Association for Computational Linguistics.

Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. Large language models as instructors: A study on multilingual clinical entity extraction. In *The*

22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 178–190, Toronto, Canada. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Saket Sharma, Aviral Joshi, Yiyun Zhao, Namrata Mukhija, Hanoz Bhathena, Prateek Singh, and Sashank Santhanam. 2023. When and how to paraphrase for named entity recognition? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7052–7087, Toronto, Canada. Association for Computational Linguistics.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Jing Wang, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2020. Multi-domain named entity recognition with genre-aware and agnostic inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8476–8488, Online. Association for Computational Linguistics.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1–15, New Orleans, Louisiana. Association for Computational Linguistics.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *arXiv preprint arXiv:2402.14568*.

Peilin Zhou, Zeqiang Wang, Dading Chong, Zhijiang Guo, Yining Hua, Zichang Su, Zhiyang Teng, Jiageng Wu, and Jie Yang. 2022a. Mets-cov: A dataset of medical entity and targeted sentiment on covid-19 related tweets. *Advances in Neural Information Processing Systems*, 35:21916–21932.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022b. MELM: Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2251–2262, Dublin, Ireland. Association for Computational Linguistics.