Investigating Symptom Recognition in Colloquial Patient Narratives

Jennifer Xu and Tamara Babaian

Bentley University, Waltham, MA 02452, USA
{jxu, tbabaian}@bentley.edu

Abstract. This research-in-progress is motivated by the need for automated processing of healthcare-related texts. Specifically, we address the problem of symptom recognition from colloquial narratives of patients. Informal, colloquial texts are a challenge to the state-of-the-art named entity recognition (NER) approaches based on the BERT model due to the differences in style, structure, and terminology than formal texts. Using text data from COVID-19-related online patient forums and medical papers, we conduct experiments comparing performance of BERT on normalized and de-normalized variants of the data. These experiments will shed light on the impact of various text characteristics on model performance, based on which design principles for domain-specific model training for NER tasks can be developed.

Keywords: Named Entity Recognition, BERT, patient narratives, symptoms

1 Introduction

Named entity recognition (NER) seeks to automatically extract terms representing entities (e.g., persons, organizations, locations, etc.) from documents. NER has long been an important task in natural language processing (NLP) applications. Recently, the introduction of the BERT (Bidirectional Encoder Representations from Transformers) model [2] has brought impressive progress in open-domain NLP tasks, including NER. However, recognition of domain-specific entities beyond standard entity types remains a challenge. In the healthcare and medical domains, for instance, although it is relatively easy to identify standard entities from medical documents, such as person names, diseases, treatments, and medicines, it is more difficult to accurately recognize descriptions of symptoms. The terms used to describe a symptom (e.g., *loss of appetite*) may vary considerably (e.g., n*o appetite, not feel like eating, not hungry, lack interest in food*). The challenge is even greater in colloquial texts, such as patient narratives posted on social media platforms, which are often quite "noisy" due to informal style, grammatical errors, typos, and use of abbreviations.

As a part of a broader research project for automated processing of healthcare related documents, this research-in-progress study is motivated by the need to recognize important medical concepts and entities from text. Specifically, we focus on the recognition of symptoms from colloquial narratives of patients. We adopt the BERT-based

approach for this domain-specific NER task. From the design science perspective, we seek to develop a set of design artifacts, which will include not only effective, domain-specific NER methods for handling informal writings but also general design principles for such tasks. We start with identifying sources of impact on NER performance when dealing with colloquial writing, aiming to discover *what* affects performance and *how* we can improve it. Our research may help healthcare professionals, medical researchers, pharmaceutical manufacturers, and disease control organizations to mine valuable information and knowledge from text data for a variety of purposes, such as studying new epidemics or diseases, gathering patient reactions to treatments, documenting side effects of drugs, and monitoring spread of viruses (e.g., COVID-19).

In the following, we review the related work on open-domain BERT and NER in the healthcare and medical domain. We present our research design and preliminary results from symptom recognition in both patient narratives and medical literature. At the end, we lay out our plan for completing the research.

2 Related Work

Natural Language Processing (NLP) refers to processing and understanding human speech or text data. NLP tasks include speech recognition, machine translation, document categorization and summarization, information extraction, etc. NLP has been one of the most difficult challenges of artificial intelligence. Traditionally, NLP problems were usually tackled using computational linguistic approaches that relied on language models (e.g., *n*-gram) and statistical methods. The development of the **BERT** model [2] brought a breakthrough to the NLP research and practice, becoming the state-of-the-art technique for many NLP tasks and achieving significantly better performance than traditional computational linguistics-based approaches and recurrent neural network models. Built on the Transformer architecture [14], BERT consists of multiple self-attention layers and millions of weights. Trained on large text corpora of books and Wikipedia entries, the open-domain BERT has embedded into its weights a comprehensive amount of knowledge of language and a deep sense of language contexts. BERT can be fine-tuned to achieve specific downstream tasks (e.g., sentiment analysis and question answering) without having to use large datasets [2]. Moreover, researchers have fine-tuned the open-domain BERT for *domain-specific* documents such as legal contracts [3] and business documents (e.g., regulatory filing) [16].

In the healthcare and medical domain, there exit a large volume of text documents such as medical literature, clinical notes or reports, hospital discharge summaries, lab protocols, social media posts by patients and their caregivers, among many others. It has been shown that the open-domain BERT needs to be fine-tuned extensively to achieve satisfactory performance for processing medical documents [4, 8, 9]. This is mainly because medical documents often contain terms and expressions that are absent from general domain corpora, on which BERT was trained [4, 8, 9]. As a result, a few variants of BERT specifically for mining medical documents have been proposed [4, 8]. Most of these BERT versions are trained using formal medical documents, such as academic literature (e.g., Medline or PubMed Central) and clinical notes, and generally

outperform the open-domain BERT. Alternatively, CT-BERT is a model trained on COVID-19 related Twitter messages and has achieved a 10-30% improvement over BERT on sentiment analysis of tweets [11].

Named Entity Recognition (NER) is an information extraction task for identifying and extracting entities of interest, such as persons (e.g., John Doe), organizations (e.g., the National Science Foundation), locations (e.g., Boston), date and time (e.g., July 4th, Saturday morning) from text documents. Traditional NER methods include lexicon-based string matching and pattern recognition, rule-based heuristics, statistical models, and classification in machine learning. The performance of word-context-based statistical and machine learning approaches, such as Conditional Random Fields (CRF) and LSTM (Long Short-Term Memory), depends heavily on feature engineering, in which part-of-speech tags, the position of the word in a sentence, and other word-based features are constructed and used as input to train and build the classifier. A large amount of prior NER studies focus on constructing features for classification [13].

The research on NER has benefited tremendously from the advent of BERT. Without having to depend on extensive feature engineering, the open-domain BERT has outperformed traditional NER approaches ([2, 6]). BERT-based models have been applied to extracting *domain-specific* entities such as medicine and treatment names in medical documents. For instance, BioBERT [4] improves the F1 score of NER in biomedical texts. Incorporating knowledge about biomedical entities (e.g., chemicals and proteins) into BioBERT, Sun et al. [13] report over 90% F scores in several biomedical corpora.

Despite the progress, NER remains a challenge. Since nearly all the BERT-based NER models in the healthcare and medical domain have been trained using academic literature, performance of these models may drop significantly on documents containing a large amount of noise (e.g., misspellings, synonyms) [1] or written in informal styles. Researchers have sought to normalize informal descriptions by mapping them to formal medical terms (e.g. “head spinning,” versus “dizziness”) to improve performance [5, 7, 10].

3 Research Design

This study focuses on the recognition of symptoms from colloquial patient narratives on social media. We will examine how NER performance is affected by (a) normalizing the colloquial texts by replacing informal descriptions of symptoms with formal terminology, and (2) de-normalizing the formal texts from medical literature by substituting informal terms from colloquial texts. Through this comparative design, we hope to address two research questions: *(1) What text characteristics (e.g., writing style, terminology, domain) may impact the recognition of symptoms from colloquial texts? (2) How can we improve performance of NER on colloquial texts?*

Our design consists of two stages: (1) data collection and preprocessing, and (2) training and testing (see Fig. 1). The colloquial dataset (CLQ) is collected from posts from the public COVID-19 community forums at Patient.info and PatientsLikeMe websites, both of which are popular online patient communities. These sites were platforms

of early and open communication during the beginning of the pandemic, in which patients described their varying symptoms instead of focusing on a few common ones. Hence, the set of COVID-19 symptoms has been constantly expanding. For samples of formal writing, we retrieved papers from the COVID-19 Open Research Collection (CORD-19)[1] with over 500,000 academic articles.

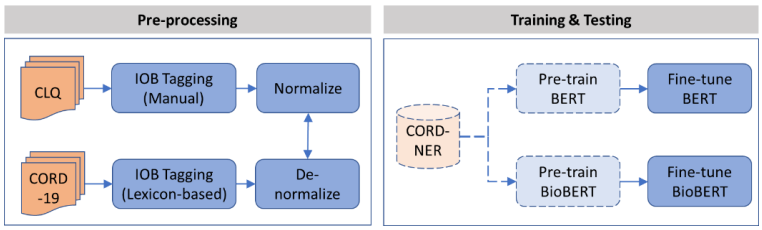

Fig. 1. Research design.

We used the IOB (Inside-Outside-Beginning) tagging scheme [12] to label the datasets. Each word in a sentence was tagged as either O – for non-entity, B-Sym, I-Sym, for the beginning and rest of the words in a symptom phrase, B-Dis or I-Dis for the beginning and rest of the words in a disease name. For example, words in "I can't smell anything!" were tagged as O, B-Sym, I-Sym, O, respectively. The CLQ dataset was labeled manually by a research assistant, reviewed and finalized by the authors. The CORD-19 dataset was labeled using a lexicon-based pattern recognition approach based on a list of 209 formal terms for COVID-19 related symptoms found in the literature.

The original, tagged CLQ dataset will be normalized by replacing informal descriptions of symptoms with corresponding formal terminology, such as "anosmia (B-Sym)" in place of "can't (B-Sym) smell (I-Sym)". The original, tagged CORD-19 dataset will be de-normalized by reversing this operation. Using the four datasets: (1) original CLQ, (2) CLQ-N: normalized CLQ, (3) original CORD-19, and (4) CORD-19-D: de-normalized CORD-19, we will identify the impacts of writing styles (academic vs. colloquial) and terminology (formal vs. informal) on model training and performance. To examine the impact of domain, we will fine-tune and compare the open-domain BERT model and the BioBERT model. An additional option is to first pre-train the two models using the CORD-NER corpus [15].

4 Preliminary Results

We extracted over 500 narratives (1,502 sentences) between February and July 2020 from the Patient.info and PatientsLikeMe websites (CLQ), and extracted 2,366 sentences containing symptom tags from the CORD-19 dataset (based on the CORD-19 collection of 534 papers, 98,186 sentences) between May 1 and May 5, 2020. To make the two datasets (**CLQ** and **CORD-19**) comparable in size, we have also created a **CORD-19-Short** dataset by selecting the first 1,500 sentences from the CORD-19 set. Table 1 presents the statistics of the above-mentioned datasets.

[1] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

We have fine-tuned the open-domain BERT-base model using the original CLQ and COVID-19 datasets, respectively. Table 1 presents the precision, recall, and F scores for each model. Apparently, the performance for the CLQ dataset is worse than that of CORD-19 and CORD-19-Short.

Table 1. Statistics of datasets and performance evaluation results.

	CLQ	**CORD-19**	**CORD-19-Short**
# Sentences	1,502	2,366	1,500
# Words/Tokens	26,288	107,073	68,328
# Distinct Symptom Terms	417	121	111
Precision	0.52	0.93	0.79
Recall	0.49	0.95	0.91
F Score	0.50	0.94	0.85

Table 2 reports the frequency counts of the top 5 symptoms in the three datasets, respectively. Not surprisingly, along with some overlap in the most frequent terms, the descriptions of symptoms vary significantly in the CLQ posts. For example, "tastebuds [sic] completely gone". "lack of taste", "odd taste in my mouth" that appear in CLQ, are referred to as "taste loss" or "loss of taste" in CORD-19 papers. Moreover, patients often used very vivid language to describe their symptoms, e.g., *very difficult time breathing, thinking is still slow, redness around the toes nails, strange mild burning sensation in the top of my chest*.

Table 2. Frequency counts of top 10 symptoms in each dataset.

	CLQ		**CORD-19**		**CORD-19-Short**	
	Term	**Freq.**	**Term**	**Freq.**	**Term**	**Freq.**
1	cough	41	fever	668	fever	437
2	fever	32	cough	418	cough	249
3	headache	17	anxiety	396	anxiety	176
4	sore throat	16	depression	218	discharge	132
5	fatigue	16	discharge	190	diarrhea	124

5 Research Plan

We have planned a series of experiments to investigate how the properties of colloquial text data affect NER performance using BERT. At this stage, we focus on symptom recognition given the difficulty of this task when performed on informal texts.

We have performed preliminary comparisons of BERT on two sets of data related to COVID-19: colloquial narratives of patients and formal academic papers. The next step of our study will be normalizing the CLQ set and de-normalizing the CORD-19 set. We will be able to compare between the following pairs of models/datasets:

- CLQ vs CLQ-N: same colloquial context/style, different terms
- CORD-19 vs CORD-19-D: same formal context/style, different terms
- CLQ vs CORD-19-D: different contexts/styles, same colloquial terms
- CORD-19 vs CLQ-N: different contexts/styles, same formal terms

By comparing performance of these pairs of models we hope to get insight into the importance of the structure of tagged entities versus the overall sentence context, which

will help us identify the ways for improving performance. Results of our work will inform the principles behind model training on colloquial texts and help develop methods and design principles for domain-specific NER.

References

1. Araujo, V., Carvallo, A., Parra, D.: Adversarial evaluation of BERT for biomedical named entity recognition. In: Proceedings of the The Fourth Widening Natural Language Processing Workshop, (2020).
2. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv: 1810.04805.
3. Elwany, E., Moore, D., Oberoi, G.: BERT goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding. ArXiv: 1911.00473.
4. Lee, J., et al.: BioBERT: A pre-trained biomedical language representation model for biomedical text mining Bioinformatics, 36, 1234-1240 (2019).
5. Li, L., Ma, R., Guo, Q., Xue, X., Qiu, X.: BERT-ATTACK: Adversarial attack against BERT using BERT. ArXiv:2004.09984.
6. Liang, C., et al.: BOND: BERT-assisted open-domain named entity recognition with distant supervision. KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1054-1064 (2020).
7. Limsopatham, N., Collier, N.: Normalising medical concepts in social media texts by learning semantic representation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany (2016).
8. Liu, N., Hu, Q., Xu, H., Xu, X., Chen, M.: Med-BERT: A Pre-training framework for medical records named entity recognition. IEEE Transactions on Industrial Informatics, (2021).
9. Liu, X., Hersch, G.L., Khalil, I., Devarakonda, M.: Clinical trial information extraction with BERT. In: IEEE 9th International Conference on Healthcare Informatics (ICHI), (2021).
10. Miftahutdinov, Z., Tutubalina, E.: End-to-end deep framework for disease named entity recognition using social media data. In: Proceedings of the 30th IEEE Jubilee Neumann Colloquium, Budapest, Hungary (2017).
11. Müller, M., Salathé, M., Kummervold, P.E.: COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. ArXiv:2005.07503.
12. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., Yarowsky, D. (eds.) Natural Language Processing Using Very Large Corpora, pp. 157-176. Springer, Dordrecht (1999).
13. Sun, C., et al.: Biomedical named entity recognition using BERT in the machine reading comprehension framework. Journal of Biomedical Informatics, 118, 103799 (2021).
14. Vaswani, A., et al.: Attention is all you need In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. 30, pp., (2017).
15. Wang, X., Song, X., Li, B., Guan, Y., Han, J.: Comprehensive named entity recognition on CORD-19 with distant or weak supervision. arXiv: 2003.12218.
16. Zhang, R., et al.: Rapid adaptation of BERT for information extraction on domain-specific business documents. ArXiv: 2002.01861.