IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 3, 2021, accepted August 24, 2021, date of publication August 27, 2021, date of current version September 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108445

Re-Ranking System with BERT for Biomedical Concept Normalization

HYEJIN CHO[1], DONGHA CHOI[2], AND HYUNJU LEE[1,2]
[1]School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea
[2]AI Graduated School, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

Corresponding author: Hyunju Lee (hyunjulee@gist.ac.kr)

This work was supported in part by the Bio-Synergy Research Project of the Ministry of Science and ICT (MSIT) through the National Research Foundation of Korea (NRF) under Grant NRF-2016M3A9C4939665, and in part by NRF Grant through Korean Government (MSIT) under Grant 2021R1A2C2006268.

ABSTRACT In recent years, various neural network architectures have been successfully applied to natural language processing (NLP) tasks such as named entity normalization. Named entity normalization is a fundamental task for extracting information in free text, which aims to map entity mentions in a text to gold standard entities in a given domain-specific ontology; however, the normalization task in the biomedical domain is still challenging because of multiple synonyms, various acronyms, and numerous lexical variations. In this study, we regard the task of biomedical entity normalization as a ranking problem and propose an approach to rank normalized concepts. We additionally employ two factors that can notably affect the performance of normalization, such as task-specific pre-training (Task-PT) and calibration approach. Among five different biomedical benchmark corpora, our experimental results show that our proposed model achieved significant improvements over the previous methods and advanced the state-of-the-art performance for biomedical entity normalization, with up to 0.5% increase in accuracy and 1.2% increase in F-score.

INDEX TERMS Named entity normalization, natural language processing, text mining, text recognition.

I. INTRODUCTION

With the rapid development of computational technology, a large amount of literature has accumulated on various aspects regardless of domain. Based on a large amount of text data, many researchers consider constructing multiple knowledge bases (KB) of domain-specific ontologies. It is generally useful in many applications, from the general domain to specialized domains such as biomedicine, and beneficial for extracting key information related to entities of interest [1]. Because newly discovered biomedical evidence is written in natural language, accurate and efficient extraction of information from unstructured data has become important in natural language processing (NLP) [2], [3].

Named entities are meaningful terms or multi-word phrases and named entity recognition (NER) is an important task for identifying named entities and classifying the domain of pre-defined entities or entity types from informal texts [4]. After named entities in texts have been recognized, the next step is named entity normalization by mapping recognized mentions to suitable identifiers (IDs) in the pre-defined dictionary. The entity normalization task in the biomedical domain is necessary to resolve semantic ambiguity, as each biomedical entity may be written in numerous forms [5]. For example, although '*cancer*' and '*tumor*' are apparently different forms in the text, they can be normalized to '*neoplasms*' with the same concept ID (MeSH:D009369). On the other hand, '*AS*' can be expanded to various words after abbreviation resolution like '*Angelman Syndrome* (MeSH:D017204)' or '*Ammonium Sulfate* (MeSH:D000645).' Although many researchers consider the ambiguity resolution to avoid these difficulties, the normalization task in the biomedical domain is still challenging because of multiple synonyms, various acronyms, and numerous lexical variations [6].

The goal of this study is to improve the performance of the biomedical entity normalization by utilizing different scoring schemes between mentions and concept names. To achieve the goal, we generate a list of candidate concept names of the input biomedical mentions sorted by their similarity scores and then re-rank the retrieved candidate concept names by developing scoring systems. Additionally, we employ our architecture focused on the following

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

two points: a semi-supervised learning model using unlabeled task-specific data [7] and a calibrated classification model [8]. We evaluate our system using five biomedical corpora with four entity types and various assessment methods. Our experimental results show that the proposed method significantly outperforms the existing state-of-the-art (SOTA) models on biomedical corpora for the task of normalization.

The main contributions of our proposed study are as follows: (i) We demonstrate the effectiveness of word representations with pre-trained language models (LMs) rather than context-independent representation; (ii) We utilize pre-trained LMs with task-specific sentences in terms of the ranking tasks for biomedical normalization; (iii) We prove that our models employing the calibration method show significant improvements in normalization performance; and (iv) We show that a simple but effective strategy of implementing the incorporation of two different scoring systems is a key factor for performance improvement of our models.

II. RELATED WORKS

The biomedical entity normalization is a long-standing and important task in the biomedical NLP domain [9]–[11] and the goal of a biomedical normalization task is to map a mention in a document to a unique concept ID in a biomedical ontology [12]. Various challenges have been organized to solve the normalization problem, and many researchers have participated in these assessments of the NLP methods. As one of the representative challenges, the BioCreative workshops provided a set of biomedical tasks to encourage NLP research and related applications. Focusing specifically on the normalization track, the BioCreative I, II, and III workshops were designed to address a number of gene names [13]–[15], and the BioCreative V workshop aimed to normalize disease and chemical mentions from MEDLINE abstracts [16]. CLEF eHealth has been running an annual evaluation campaign in the medical and biomedical domain, and the Shared Annotated Resources (ShARe) project has created a disorder mention corpus from clinical texts. Therefore, the ShARe/CLEF eHealth 2013 Challenge offered the NER task for disorder mentions in clinical notes, along with the normalization task to map unique identifiers [17]. Furthermore, a part of the SemEval workshop was designed as a follow-up to the ShARe/CLEF eHealth 2013 Challenge. Using the ShARe/CLEF corpus, the SemEval-2014 Task 7 (Task B) [18] and the SemEval-2015 Task 14 (Task 1) [19] organized open challenges to recognize the span of a disorder mention in the clinical text and to normalize the disorder to a unique CUI in the SNOMED-CT subset of UMLS terminology. The Text Analysis Conference (TAC) is another series of workshops organized to assess a variety of NLP methods. To detect the adverse drug reactions (ADR) described in the structured product labels of drugs, the TAC 2017 challenge consisted of several intermediate tracks including the ADR extraction from drug labels and normalization through MedDRA terminology [20]. To provide various NLP tasks with annotated data in the clinical domain, the Informatics for Integrating Biology and the Bedside (i2b2) project has organized a series of shared tasks since 2006. In 2010, i2b2 with VA Salt Lake City Health Care System has run a medical NLP workshop for clinical records, called the 2010 i2b2/VA challenge, and they released a manually annotated corpus of patient reports [21]. The MCN (medical concept normalization) corpus is a subset of discharge summaries from the fourth i2b2/VA 2010 shared task [22] and this corpus is utilized as a shared-task dataset in the 2019 National NLP Clinical Challenges (n2c2)/Open Health NLP (OHNLP) track 3 [23].

The community-wide tasks have greatly promoted biomedical NLP research by building benchmark datasets and innovative methods. Through these challenges, many researchers have examined various techniques, such as dictionary-based, rule-based, machine learning-based, and deep learning-based methods.

The most common traditional normalization approaches are dictionary-based and rule-based methods, which use pattern matching based on dictionary lookup and heuristic matching rules, respectively. The sieve-based system [12] is a cascade architecture based on ten kinds of manual rules and Apache Lucene [24] is a Java-based text indexing and searching engine library by calculating the similarity between a document and a query. Although these approaches can be easily applied to broad areas such as disease, gene, and chemical name normalization tasks [25]–[29], they may often be inefficient and less accurate for words not in the dictionary or ungrammatical text with typos.

Although many normalization tools still tend to rely on the accuracy of well-constructed dictionaries or domain-specific rules, several studies have applied machine learning techniques to overcome the previous limitations. DNorm [10] proposed a pairwise learning-to-rank method to measure the similarities between entity mentions and candidate concepts. TaggerOne [30] is a machine learning-based system that jointly performs disease NER and normalization by utilizing semi-Markov models. Another machine learning technique for biomedical normalization is to utilize word representations in vector space. For instance, the Word2Vec-based method [6], convolutional neural network (CNN)-based ranking method [31], BNE [32] using a long short-term memory (LSTM), and NormCo [33] using a gated recurrent unit (GRU) network proposed entity representation architecture to calculate semantic similarities between biomedical mentions and candidate concepts.

Along with the success of deep learning, recent studies have focused on a paradigm shift in NLP from task-specific training methods to fine-tuning approaches based on general-purpose LMs. Following this trend, the most commonly used pre-trained model is bidirectional encoder representations from Transformers (BERT) [34] based on the transformer architecture [35]. BERT is a contextual language representation model that uses pre-trained deep bidirectional representations from the unlabeled text. Recently, BERT has been

adapted to the biomedical domain by further pre-training on additional corpora as follows:

BioBERT: BioBERT [36] is a domain-specific language representation model designed for biomedical text, and the model is initialized with the checkpoint of BERT, followed by training the BERT model on PubMed abstracts and PubMed Central full-text articles again. BioBERT achieves SOTA performance on various biomedical NLP tasks with task-specific fine-tuning while requiring only minimal architectural modifications.

SciBERT: Similar to BioBERT, SciBERT [37] is another BERT-based model following the same architecture as BERT. Although BERT was pre-trained using general-domain corpora, SciBERT was pre-trained from scratch using several scientific papers that consisted of the full text of computer science and biomedical domains. Furthermore, they constructed a new in-domain vocabulary on their scientific text corpora, called SciVocab.

PubMedBERT: PubMedBERT [38] is another pre-trained LM, following the same architecture as BERT. However, unlike the mixed-domain pre-training models, the weights of the PubMedBERT model were not initialized with those of BERT during pre-training. They constructed an in-domain vocabulary of the target biomedical domain and pre-trained from scratch on PubMed abstracts and additional data from PubMed Central full-text articles.

These fine-tuned versions of BERT-based models are often combined with various machine learning approaches to deliver good performance in biomedical normalization tasks. Ji *et al.* [39] applied an ensemble approach based on Lucene and a pair-wise BERT classifier, and Xu *et al.* [40] also proposed a hybrid system based on Lucene or a multi-class BERT classifier for the candidate generation, and a list-wise BERT classifier for ranking. BIOSYN [41] utilized entity representation from the BERT-based model and developed a synonym marginalization method with marginal maximum likelihood.

III. METHODOLOGY

In this section, we propose a method for entity normalization using the BERT-based model. First, we assume that an input mention m has its own concept ID c, and each c has at least one concept name n according to the dictionary. Our goal in this study is to assign a biomedical mention m to its unique concept ID c in the target dictionary. Formally, given a list of biomedical mentions $M = \{m_1, m_2, \ldots\}$ from a document and a set of concept IDs $C = \{c_1, c_2, \ldots\}$ and concept names $N = \{n_1, n_2, \ldots\}$ from the ontology, the goal of concept normalization is to map the i-th mention m_i to its correct concept c^* through a normalization function f:

$$c^* = f(m_i, N) = ID(\arg\max_{n \in N} P(n|m_i; \theta)) \tag{1}$$

where $ID(n)$ is a function that returns the unique ID of the concept name n, and θ denotes a parameter corresponding to our normalization model. As shown in Fig. 1, our system for biomedical entity normalization consists of three steps: candidate concept generation, candidate concept ranking, and entity disambiguation. A detailed description of these steps is provided in the following sections.

A. CANDIDATE CONCEPT GENERATION

Traditional word embeddings are context-independent representations such as Word2Vec [42] and GloVe [43], which constitute a single vector for each word regardless of the meaning and position of the word in the sentence.

With advances in contextualized representations, including ELMo [44] and BERT [34], the ability to share contextual information of words in sentences has further improved performance in various NLP tasks and demonstrated that relatively simple models using contextualized embeddings can outperform complex models using non-contextualized embeddings [45]. Therefore, we employed contextual representation models (i.e., BERT-based models) to extract feature embeddings for candidate concept generation.

Recent studies suggest further pre-training of a pre-trained LM with the in-domain data for task adaptation and show improved performance and effectiveness on downstream tasks from each target domain [7], [36], [37], [46]. To employ this strategy, we first collect corresponding texts from the same target task, which only makes use of the in-task text without any label as task-specific pre-training (Task-PT) data. Subsequently, we employ the original BERT-based models and continually execute an additional phase of pre-training with a masked language model (MLM) and next sentence prediction (NSP) approach on the Task-PT data.

The candidate concept generation step is retrieved to construct a list of candidate names $N_m \subseteq N$, which consists of possible k concept names in the ontology for the given mention $m \in M$. Based on the BERT-based model, we first extract embeddings e_m and e_n for mention m and each concept name $n \in N$, respectively. BERT uses WordPiece tokenization [47], which split an input word into pre-defined subword units to reflect rare words and morphological variation in linguistics. To retain linguistic information, we sum up embeddings of the subwords by the BERT encoder into one vector as desired embeddings of the input word. To retrieve relevant k candidate concept names for each mention m, we define the scoring function as a static BERT-score (score_{SB}) of each pair (m, n) as follows:

$$\text{Score}_{SB}(m, n) = sim(e_m, e_n) = \frac{e_m \cdot e_n}{\|e_m\| \, \|e_n\|} \in \mathbb{R} \tag{2}$$

where $sim(e_m, e_n)$ is calculated using the cosine similarity between two vectors e_m and e_n, which is a value $\in [0, 1]$.

B. CANDIDATE CONCEPT RANKING

In this section, we re-rank the list of candidate concepts by fine-tuning the BERT-based models, where we transform a binary classification task into a ranking task. Suppose there are k candidates in $N_m = \{n_1, \ldots, n_k\}$ for the input mention m. We can generate all mention-candidate name

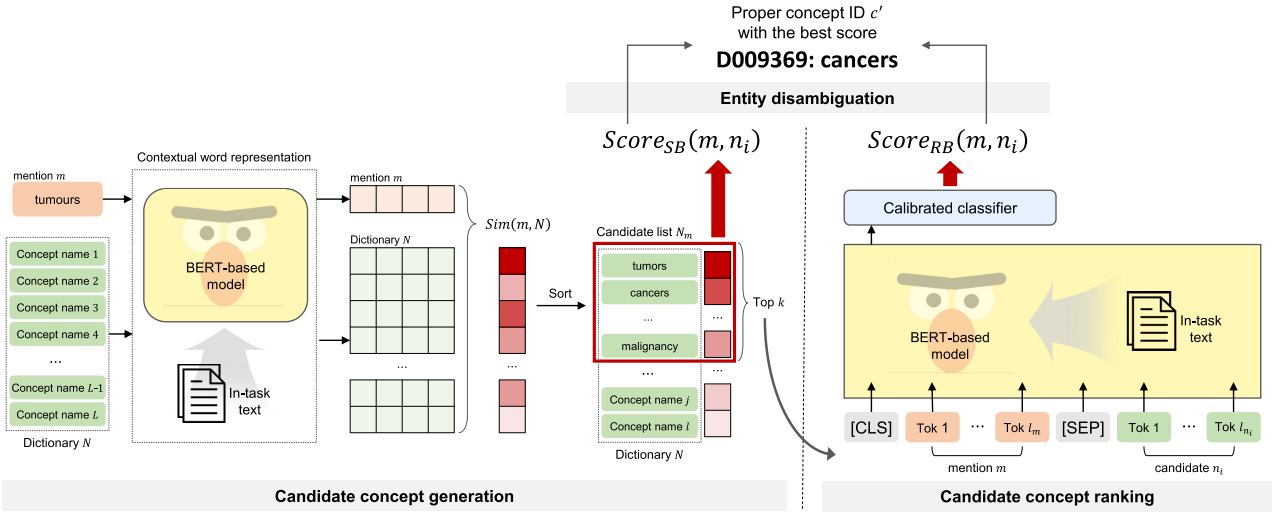

FIGURE 1. Overview of our normalization model. Candidate concept generation: An input mention and all concept names in a dictionary are represented by a task-specific pre-trained (Task-PT) LM, and we can generate a list of candidate names sorted by $score_{SB}$. Candidate concept ranking: We calculate the $score_{RB}$ of the list of candidates by utilizing the fine-tuned Task-PT LM along with the calibration method. Entity disambiguation: We re-rank the list of candidates using final scores and infer the proper concept ID with the best score.

pairs $\{(m, n_1), (m, n_2), \ldots, (m, n_k)\}$ from a list of the mention m and their candidate names N_m. We design the label of mention-candidate pairs as binary classes, which are represented as either 'correct' or 'incorrect.' If the i-th candidate concept name n_i is related to the mapping concept ID c for the target mention m, then it is labeled as '1', otherwise '0', which means that the candidate n_i is irrelevant with the mention m as a negative sample.

Concretely, for each pair of the mention m and the i-th candidate name n_i, we take an input sequence '[CLS] m [SEP] n_i' of the fine-tuning procedure, where '[CLS]' is a special token and '[SEP]' is a special separator token between m and n_i. To apply BERT for the classification task, we used the final layer of the special token '[CLS]' in the mention-candidate pair to compute the probability distribution of binary classes, and the output probability for each class was calculated using the softmax function.

However, these hard labels may adversely affect model generalization as the probabilistic models become overconfident about their predictions and overfit the training data with hard targets [48]. To solve this problem, we employ the confidence panelty as a regularization term for the loss function to alleviate the peaked distributions [8]. The conditional distribution $p_\theta(y|x)$ is described as:

$$H(p_\theta(y|x)) = -\sum_i p_\theta(y_i|x) \log(p_\theta(y_i|x)) \tag{3}$$

where p_θ is the probability of class y_i given an input sequence x and i indicates the index of each class. The confidence penalty loss (CPL) function ensures that the low entropy output distributions are penalized by adding the negative entropy H to the negative log-likelihood training objective as follows:

$$\mathcal{L}(\theta) = -\sum \log(p_\theta(y_i|x)) - \beta H(p_\theta(y|x)) \tag{4}$$

where β controls the strength of the confidence penalty. Thus, the incorporation of negative entropy into the original loss function enables overfitting and improves the generalization performance [49].

Similar to the previous method [39], we defined the probability of label '1' obtained from the softmax function as a ranking BERT-score ($score_{RB}$) of each mention-candidate pair (m, n) as follows:

$$\text{Score}_{RB}(m, n) = P(\text{label} = 1|m, n) \in \mathbb{R} \tag{5}$$

C. ENTITY DISAMBIGUATION

The final score is calculated for each candidate pair using the two aforementioned scores both $score_{SB}$ and $score_{RB}$ as follows:

$$\text{Score}(m, n) = \text{Score}_{SB}(m, n) + \text{Score}_{RB}(m, n) \tag{6}$$

Using the above equation, we re-calculated the scores of all the retrieved pairs and then re-ordered the list of candidates according to the final ranking score in decreasing order. Therefore, we could predict the proper concept ID c' with the highest score:

$$c' = ID(\arg\max_{n \in N_m} \text{Score}(m, n)) \tag{7}$$

IV. EXPERIMENTAL SETUP

A. DATASETS

We evaluated our normalization approach on the English biomedical benchmark corpora described in Table 1: the National Center for Biotechnology Information disease (NCBI) corpus [50], the BioCreative V Chemicals Disease Relationship (CDR) corpus [16], the BioCreative II Gene Normalization (GN) corpus [14], and the plant (Plant)

TABLE 1. Data statistics of five benchmark datasets for the biomedical entity normalization.

Corpus	Abstracts			Mentions		
	Train	Dev	Test	Train	Dev	Test
NCBI	592	100	100	5134	787	960
CDR-DIS	500	500	500	4182	4244	4424
CDR-CHEM	500	500	500	5203	5347	5385
GN	281	-	262	684	-	785
Plant	128	40	40	2647	709	629

corpus [6]. These corpora cover four major biological entity types: disease, chemical, gene, and plant. We briefly explain each corpus in the following sections.

NCBI for disease names: The NCBI corpus is the gold standard dataset of disease name recognition and normalization tasks, which consists of disease mentions mapped to their concept IDs in the MEDIC [51] vocabulary of the Comparative Toxicogenomics Database (CTD) project [52]. This corpus is available at http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/. In our experiments, we used the July 2012 version of MEDIC, which contains 11,915 MeSH and OMIM identifiers and 71,923 disease names with synonyms.

CDR for disease and chemical names: The CDR corpus is a dataset used for the BioCreative V challenge based on disease and chemical entity recognition and chemical-induced disease relation extraction tasks. It can be downloaded from http://www.biocreative.org/tasks/biocreative-v/track-3-cdr/. It should be noted that we denote a subset with disease entities and chemical entities as 'CDR-DIS' and 'CDR-CHEM', respectively. In this study, we used the MEDIC version published in June 2015 and the CTD chemical vocabulary published in July 2015 for the experiments of the CDR-DIS and the CDR-CHEM, respectively.

GN for human gene names: The GN corpus is the gold standard for the gene normalization task in the BioCreative II challenge to determine human genes or gene products mentioned in PubMed abstracts and to map them to the unique concept IDs. This corpus can be found at https://biocreative.bioinformatics.udel.edu/tasks/biocreative-iii/gn/. In this study, we used the EntrezGene [53] list, which contains 32,975 identifiers and 182,989 gene names and synonyms.

Plant for plant names: The plant corpus is a manually annotated abstract-based corpus for the plant normalization task with plant mentions and their unique concept IDs. The plant corpus is freely available for download (http://gcancer.org/plant/). Following a previous study [6], we also used the viridiplantae ontology from the NCBI taxonomy database [54].

B. IMPLEMENTATION DETAILS

1) PREPROCESSING

We performed several pre-processing strategies for each mention and each concept in the KB as follows: (i) combine mention information in the training set to increase the coverage of the ontology [12]; (ii) resolve abbreviations using the abbreviation resolution module [55]; (iii) lowercase all characters; (iv) remove all characters except for the lowercase alphanumeric for both mentions and concept names in the dictionary; and (v) split composite mentions into separate mentions using heuristic rules [12]. For example, we could separate 'breast and ovarian cancer' into 'breast cancer' and 'ovarian cancer', respectively.

TABLE 2. The hyperparameters of our proposed model.

Datasets	Training epoch	Weights of CPL
NCBI	10	0.2
CDR-DIS	5	0.2
CDR-CHEM	3	0.5
GN	3	0.1
Plant	5	0.1

2) TRAINING

Our experiments were conducted in a workstation with an Intel(R) Xeon(R) Gold 5120 CPU, 265 GB RAM, and three Tesla V-100-SXM2-32GB GPU. Our model was implemented in Python and based on the source codes of BERT, which were built using TensorFlow in the backend. To set hyperparameters, we performed a grid search on the training epochs and weights of CPL, and then selected the hyperparameters with the best performance in the development corpus.

To re-rank the list of candidate concepts, we performed fine-tuning by using the BERT-based models, where we transform a binary classification task into a ranking task. To fine-tune our models, we derived our training and development datasets from the candidate concept generation step. We heuristically selected $k = \{100, 10\}$ candidates for the input mention to generate mention-candidate name pairs, and then we employed the pairs to relation classification data as training and development, respectively. Note that the number of top candidates k was set to 20, as proposed by [41].

To set the training epoch and the weights of CPL, we performed a grid search over the epochs from 1 to 10 and weight values of CPL in {0.05, 0.1, 0.2, 0.3, 0.4, 0.5}, and then selected the hyperparameters with the best performance on the development dataset of each corpus. The hyperparameters of our proposed model are described in Table 2. We set the maximal sequence length to 64 and the batch size as 64, based on the recommendation options in BERT and other training settings including the optimizer are the same as those in the original BERT.

Although it takes approximately 8.5 hours to pre-train our models regardless of the length of task-specific data, each epoch, during fine-tuning, takes a different time ranging from 10 minutes to 1.3 hours, depending on the size of the training data.

3) EVALUATION

We applied two evaluation methods, namely accuracy and F-measure, for comparison of previous studies which used different evaluations.

TABLE 3. Comparison of the experimental results on five biomedical entity normalization corpora. The values in bold denote the best performance of each corpus and '–' denotes that results are not reported in the compared model.

Models		NCBI		CDR-DIS		CDR-CHEM	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Sieve-based Ranking [12]		84.7	-	84.1	-	90.7	-
TaggerOne [30]		87.7	-	88.9	-	94.1	-
CNN-based Ranking [31]		86.1	-	-	-	-	-
NormCo [33]		87.8	-	88.0	-	-	-
BNE [32]		87.7	-	90.6	-	95.8	-
BERT-based Ranking [39]		89.1	-	-	-	-	-
TripleNet [56]		90.0	-	-	-	-	-
BIOSYN [41]		91.1	93.9	93.2	96.0	**96.6**	**97.2**
Our models	Score_{SB}	89.3 (±0.64)	94.9 (±0.42)	92.4 (±0.22)	96.2 (±0.15)	95.4 (±0.09)	96.8 (±0.12)
	Score_{RB}	91.8 (±0.66)	95.8 (±0.34)	93.3 (±0.34)	96.5 (±0.35)	95.9 (±0.24)	97.1 (±0.14)
	Score	**92.1** (±0.48)	**95.8** (±0.21)	**93.7** (±0.14)	**96.8** (±0.26)	96.1 (±0.10)	**97.2** (±0.12)

Models		GN			Plant		
		Precision	Recall	F-score	Precision	Recall	F-score
GNAT [26]		**90.7**	82.4	86.4	-	-	-
GeNo [57]		87.8	85.0	86.4	-	-	-
GNormPlus [29]		87.1	86.4	86.7	-	-	-
Multi-stage system [58]		88.1	92.3	90.1	-	-	-
Word2vec-based Ranking [6]		-	-	-	59.0	82.2	68.7
Lucene [24]		-	-	-	82.2	93.0	87.3
Our models	Score_{SB}	85.5 (±0.60)	91.2 (±0.30)	88.3 (±0.44)	83.7 (±0.42)	94.2 (±0.14)	88.7 (±0.27)
	Score_{RB}	87.6 (±0.43)	93.8 (±0.35)	90.6 (±0.33)	**86.9** (±0.85)	95.0 (±0.26)	90.7 (±0.54)
	Score	88.7 (±0.45)	**94.1** (±0.27)	**91.3** (±0.32)	86.8 (±0.68)	**95.1** (±0.26)	**90.8** (±0.44)

Accuracy: We assessed our model using the accuracy of the top k predictions for the disease and chemical name normalization tasks. We define Accuracy@n (Acc@n) as the percentage of mentions in the corpus, which contain the correct concept ID within the top n retrieved candidates. We set $n = \{1, 5\}$ as Acc@1 and Acc@5, respectively.

F-score: We assessed our model in the gene and plant name normalization based on the following performance measures: precision (p), recall (r), and F-score (f). Precisely, the prediction is recognized in the following manner: (i) True positives (TP) if the identifiers match the answer; (ii) false positives (FP) if the identifiers do not match the answer; and (iii) false negatives (FN) if the gold standard identifiers do not match. The formulas are as follows:

$$p = \frac{TP}{TP + FP}, r = \frac{TP}{TP + FN}, f = \frac{2 * p * r}{p + r} \tag{8}$$

V. EXPERIMENTAL RESULTS

A. MAIN RESULTS

In Table 3, we compare the performance of our proposed model (Task-PT PubMedBERT with CPL) with previous normalization methods in terms of accuracy and F-score on a set of benchmark corpora. More precisely, our model with 'score_{SB}' and 'score_{RB}' shows that our model uses only the static BERT-scores and the ranking BERT-scores, respectively. Moreover, our model with 'score' represents our complete model with the default scoring method. It should be noted that we use the same list of candidates to test different scoring types of our models. Furthermore, we independently tested ten times for each corpus from scratch and calculated the mean accuracy and standard deviations for each evaluation type. We evaluate the effectiveness of our proposed scoring algorithm using experiments, and the results show that our approach is helpful in improving the performance of the biomedical normalization task.

First of all, we compare the mean accuracy of our models with that of the existing methods, based on the NCBI, CDR-DIS, and CDR-CHEM test datasets. As shown in the top section of Table 3, our proposed model achieves new SOTA performance on the NCBI and CDR-DIS. Especially, our experiments evaluated that the ensemble scoring algorithm showed significant improvement over the scoring methods using score_{SB} and score_{RB} separately. One example of such improvements is that '*deficiency of the second component of complement* (OMIM:217000)' was correctly predicted as '*deficiency of complement protein c2* (OMIM:217000)' in the final ensemble scoring algorithm, while '*deficiency of the fifth component of complement* (OMIM:609536)' incorrectly had the highest score in score_{SB}. As another example, '*autosomal recessive alport syndrome* (MeSH:C536587)' was correctly mapped into '*alport syndrome autosomal recessive* (MeSH:C536587)' in the final ensemble scoring algorithm, while '*alport syndrome* (MeSH:D009394)' incorrectly had the highest score in score_{RB}.

Compared with the current SOTA model on the CDR-CHEM, our model measured by Acc@1 shows slightly lower accuracy (approximately 0.5%); however, our system measured by Acc@5 obtained the same performance as the SOTA system. In the bottom section of Table 3, the F-score is used to compare the performance of our model with that of the existing methods based on the GN and plant test corpora. From the results, it can also be seen that our model consistently outperforms previous other models and achieves new SOTA performance in terms of F-score by 1.2% and

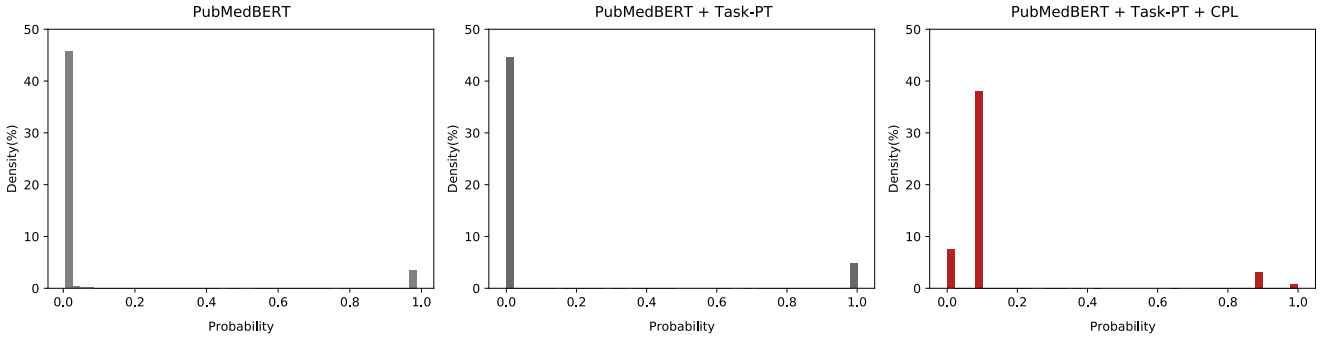

FIGURE 2. The distribution of output probabilities for our models using the NCBI test dataset. **Left (light gray): Vanilla PubMedBERT model. Middle (gray): Task-PT PubMedBERT with the in-task sentence. Right (red): Task-PT PubMedBERT with CPL.**

3.5% on the GN and Plant test datasets, respectively. Note that we use the BM25 similarity measure [59] provided by Lucene for gene and plant name retrievals, whereas the performance of other models was obtained from other studies.

B. TASK-PT EVALUATION

We illustrate the impact of Task-PT using several in-task data with five types of pre-trained LMs: BERT, BioBERT, SciBERT, PubMedBERT, and PubMedBERT-fulltext. It should be noted that we empirically set the hyperparameters of the training epoch as 3 and the value of CPL as 0.2 in this experiment. In Table 4, we describe our experiments in terms of three points: (i) Which model shows the best performance? (ii) How good is the performance of Task-PT with in-task data? (iii) Which type of in-task data is suitable for this approach?

With respect to Acc@1 among the vanilla BERT-based models on the NCBI development set, we observed that a significant benefit is achieved by using the PubMedBERT model, instead of the PubMedBERT-fulltext, as it contains more pre-training corpora. The surprisingly poor performance of the SciBERT model compared to other models is because SciBERT is an adaptation of BERT for biomedical and scientific domains, and computer science text is clearly out-domain from the perspective of biomedical applications.

As in-task data, training, development, and test sets of each corpus were used to represent sentences in the training set, in the combination of training and development sets, and the entire sets as 'TRAIN', 'TRAIN+DEV', and 'ALL', respectively. Although the performance of BioBERT with TRAIN is slightly lower than expected in terms of Acc@1, all results of Acc@5 show reasonable differences ranging from 0.2% to 2.1% when compared to the vanilla models. In the case of PubMedBERT, PubMedBERT with TRAIN+DEV and ALL appear to be approximately equivalent to Acc@1 and Acc@5. To compare models with the same highest scores, we denote the additional type of evaluation as Acc@3 ($n = 3$). Because the best accuracy is achieved in Acc@3, we demonstrated the effectiveness of Task-PT PubMedBERT with 'ALL' and utilized it to test other corpora for normalization tasks.

TABLE 4. Impact of Task-PT pre-trained LMs on in-task data. We compared several BERT-based models with or without Task-PT using different in-task data on the NCBI development dataset. The values in bold denote the best performance of each corpus.

Models		NCBI		
Pre-trained LM	In-task data	Acc@1	Acc@3	Acc@5
BERT	-	90.3	93.1	93.3
	TRAIN	90.9	93.6	93.7
	TRAIN+DEV	90.5	93.7	94.3
	ALL	90.4	93.5	94.1
BioBERT	-	90.3	92.7	93.2
	TRAIN	89.9	93.7	94.3
	TRAIN+DEV	90.8	94.3	94.7
	ALL	90.6	94.5	94.6
SciBERT	-	89.5	93.6	94.5
	TRAIN	90.3	94.2	94.7
	TRAIN+DEV	90.1	94.2	94.8
	ALL	90.5	94.2	94.8
PubMedBERT	-	91.6	94.1	94.1
	TRAIN	93.6	95.4	95.4
	TRAIN+DEV	**94.0**	96.1	**96.2**
	ALL	**94.0**	**96.2**	**96.2**
PubMedBERT -fulltext	-	91.1	94.1	94.2
	TRAIN	92.8	95.2	95.3
	TRAIN+DEV	92.8	94.9	94.9
	ALL	93.2	94.9	94.9

C. CALIBRATION EVALUATION

To evaluate the calibration performance, we compared each of our approaches with respect to the expected calibration error (ECE) and over-confidence error (OE) [60], [61]. Owing to the relatively small number of test mentions, we skipped the grouping interval bins, and then slightly modified the formula for calculating ECE and OE, which is described as follows:

$$\mathrm{ECE} = |\mathrm{acc}(M) - \mathrm{conf}(M)| \quad (9)$$

$$\mathrm{OE} = [\mathrm{conf}(M) \times \max\{\mathrm{conf}(M) - \mathrm{acc}(M), 0\}] \quad (10)$$

where M is a list of test mentions, $\mathrm{acc}(M)$ is the number of correct predictions for given M, and $\mathrm{conf}(M)$ is the summation of the winning softmax scores for M. When the ECE of a model is close to zero, the model is significantly well-calibrated because its accuracy and confidence

TABLE 5. Ablation study for the effect of calibration on the NCBI test dataset. The values in bold denote the best performance of each column.

Models	Confidence(%)	ECE	OE
PubMedBERT	97.22	0.0561	0.0545
PubMedBERT+Task-PT	99.20	0.0463	0.0459
PubMedBERT+Task-PT+CPL	98.60	**0.0341**	**0.0336**

are almost the same. In Table 5, it can be seen that the PubMedBERT model using CPL is better calibrated than vanilla PubMedBERT. We additionally visualized the distribution of the output labels for each model. As shown in Fig. 2, the confidence penalty promotes dispersed distributions, which may lead to better performance.

D. ERROR ANALYSIS

We performed an error analysis of the NCBI corpus by dividing the four major causes of false positive prediction produced by our model.

1) MISPREDICTIONS IN THE CANDIDATE CONCEPT GENERATION STEP

The majority of the errors (26.7%) were attributed to wrong candidate selections, where a list of candidate names for each mention did not include any concept name of a gold standard concept ID. If appropriate candidates are not extracted from the dictionary in the candidate concept generation step, it becomes difficult to improve the performance even after conducting additional post-processing.

2) MISPREDICTIONS BIASED TOWARDS Score_{SB}

The 21.4% of errors occurred when the score_{SB} of a candidate is significantly higher than other candidate names. Although we expect score_{SB} to reflect the linguistic regularity between the pairs of words, the slight spelling difference or overlap cases between the mention and candidate pairs can make prediction challenging. For example, although '*desmoid tumor*' is considered as 'MeSH:C535944' for a gold standard concept ID in the NCBI train set (PubMed ID:1351034), '*desmoid tumors*' in "No *desmoid tumors* were found in these kindreds. (PubMed ID:9585611)" is annotated with 'MeSH:D018222' in the NCBI test set.

As a special case, the score_{SB} is 1.0 when the input mention and the predicted name are the same. In this case, the concept ID with the identical name is placed in the top rank, because score_{SB} is too dominant in the final score to re-order the list of candidates. This could be due to an annotation error in the corpus construction or the same mention could have been interpreted differently depending on the context. These error cases would be equally represented in other studies using four corpora (i.e. NCBI, CDR-DIS, CDR-CHEM, and GN) except the Plant corpus. When we eliminated such annotation concerns from test sets, we can obtain the improved Acc@1 of up to 93.60% for NCBI, 94.79% for CDR-DIS, 97.34% for CDR-CHEM, and F-score of up to 92.4% for GN. (see Supplementary Material for more details.)

3) MISPREDICTIONS BIASED TOWARDS Score_{RB}

Contrary to the above-mentioned error, there are 21.4% errors when the score_{RB} of a candidate is significantly higher than the others, and it significantly contributes to semantically unrelated concept name. For example, although '*heart abnormalities* (MeSH:D006330)' is a gold standard concept name for the input '*cardiac defects* (MeSH:D006330)', '*cardiac abnormalities* (MeSH:D018376)' has a higher score_{RB} than '*heart abnormalities*', that is, 0.985 and 0.008, respectively.

4) MISPREDICTIONS IN THE ENTITY DISAMBIGUATION STEP

Although both score_{SB} and score_{RB} are not considered to be significant during the scoring and ranking step, when the sum of both scores is used as the total score, the final rank of candidates changes slightly. To solve this problem, instead of using a simple summation approach, an appropriate weight parameter can be used to balance the degree of importance between score_{SB} and score_{RB}.

VI. CONCLUSION

In this study, we applied and evaluated pre-trained language representation models for the biomedical entity normalization task as the re-ranking problem, which takes advantage of pre-trained LMs in modeling two different scoring strategies between entity mentions and candidate concepts. Among the five biomedical corpora, the results of our experiment showed that our model achieved SOTA performance for four biomedical corpora and obtained promising performance for the chemical entity normalization task. We found that PubMedBERT-based models outperformed other BERT-based models. Moreover, the performance can be further improved by additional Task-PT with in-task data, and we found that the calibration approach can significantly improve the performance of PubMedBERT-based models. In the future, our approach will be evaluated using more biomedical NLP tasks of various biological entities.

REFERENCES

[1] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings Bioinf.*, vol. 6, no. 1, pp. 57–71, Mar. 2005.
[2] Z. Zeng, H. Shi, Y. Wu, and Z. Hong, "Survey of natural language processing techniques in bioinformatics," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–10, Oct. 2015.
[3] D. Kim, J. Lee, C. H. So, H. Jeon, M. Jeong, Y. Choi, W. Yoon, M. Sung, and J. Kang, "A neural named entity recognition and multi-type normalization tool for biomedical text mining," *IEEE Access*, vol. 7, pp. 73729–73740, Jun. 2019.
[4] D. Campos, S. Matos, and J. L. Oliveira, "Biomedical named entity recognition: A survey of machine-learning tools," in *Theory and Applications for Advanced Text Mining*, Rijeka, Croatia: InTech, 2012, ch. 8, pp. 175–195.
[5] R. Leaman, R. Khare, and Z. Lu, "Challenges in clinical natural language processing for automated disorder normalization," *J. Biomed. Inform.*, vol. 57, pp. 28–37, Oct. 2015.
[6] H. Cho, W. Choi, and H. Lee, "A method for named entity normalization in biomedical articles: Application to diseases and plants," *BMC Bioinf.*, vol. 18, no. 1, p. 451, Oct. 2017.
[7] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," 2020, *arXiv:2004.10964*. [Online]. Available: http://arxiv.org/abs/2004.10964

[8] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," 2017, *arXiv:1701.06548*. [Online]. Available: http://arxiv.org/abs/1701.06548

[9] C.-C. Huang and Z. Lu, "Community challenges in biomedical text mining over 10 years: Success, failure and the future," *Briefings Bioinf.*, vol. 17, no. 1, pp. 132–144, Jan. 2016.

[10] R. Leaman, R. I. Doğan, and Z. Lu, "DNorm: Disease name normalization with pairwise learning to rank," *Bioinformatics*, vol. 29, pp. 2909–2917, Nov. 2013.

[11] R. Leaman, C.-H. Wei, A. Allot, and Z. Lu, "Ten tips for a text-mining-ready article: How to improve automated discoverability and interpretability," *PLOS Biol.*, vol. 18, no. 6, Jun. 2020, Art. no. e3000716.

[12] J. D'Souza and V. Ng, "Sieve-based entity linking for the biomedical domain," in *Proc. 53rd ACL 7th IJCNLP*, 2015, pp. 297–302.

[13] L. Hirschman, M. Colosimo, A. Morgan, and A. Yeh, "Overview of BioCreAtIvE task 1B: Normalized gene lists," *BMC Bioinf.*, vol. 6, no. 1, pp. 1–10, May 2005.

[14] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H.-H. Liu, R. Torres, M. Krauthammer, W. W. Lau, H. Liu, C.-N. Hsu, M. Schuemie, K. B. Cohen, and L. Hirschman, "Overview of BioCreative II gene normalization," *Genome Biol.*, vol. 9, no. 2, p. S3, 2008.

[15] Z. Lu and W. J. Wilbur, "Overview of BioCreative III gene normalization," in *Proc. BioCreative III Workshop*, 2010, pp. 24–45.

[16] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "BioCreative V CDR task corpus: A resource for chemical disease relation extraction," *Database*, vol. 2016, May 2016, Art. no. baw068.

[17] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. F. Jones, J. Leveling, L. Kelly, L. Goeuriot, D. Martinez, and G. Zuccon, "Overview of the ShARe/CLEF eHealth evaluation lab 2013," in *Proc. Int. Conf. CLEF Eur. Lang.*, 2013, pp. 212–231.

[18] S. Pradhan, N. Elhadad, W. Chapman, S. Manandhar, and G. Savova, "SemEval-2014 task 7: Analysis of clinical text," in *Proc. 8th SemEval*, 2014, pp. 1–9.

[19] N. Elhadad, S. Pradhan, S. Gorman, S. Manandhar, W. Chapman, and G. Savova, "SemEval-2015 task 14: Analysis of clinical text," in *Proc. 9th SemEval*, Jun. 2015, pp. 303–310.

[20] K. Roberts, D. Demner-Fushman, and J. M. Tonning, "Overview of the TAC 2017 adverse reaction extraction from drug labels track," in *Proc. TAC*, 2017, pp. 1–13.

[21] O. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *J. Amer. Med. Informat. Assoc.*, vol. 18, no. 5, pp. 552–556, Jun. 2011.

[22] Y.-F. Luo, W. Sun, and A. Rumshisky, "MCN: A comprehensive corpus for medical concept normalization," *J. Biomed. Informat.*, vol. 92, Apr. 2019, Art. no. 103132.

[23] S. Henry, Y. Wang, F. Shen, and O. Uzuner, "The 2019 national natural language processing (NLP) clinical challenges (n2c2)/open health NLP (OHNLP) shared task on clinical concept normalization for clinical records," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 10, pp. 1529–1537, Sep. 2020.

[24] The Apache Software Foundation. (2011). *Apache Lucene*. [Online]. Available: http://www.apache.org

[25] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: Historical perspective and recent advances," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 3, pp. 229–236, May 2010.

[26] J. Hakenberg, M. Gerner, M. Haeussler, I. Solt, C. Plake, M. Schroeder, G. Gonzalez, G. Nenadic, and C. M. Bergman, "The GNAT library for local and remote gene mention normalization," *Bioinformatics*, vol. 27, no. 19, pp. 2769–2771, Oct. 2011.

[27] T. Rocktäschel, M. Weidlich, and U. Leser, "ChemSpot: A hybrid system for chemical named entity recognition," *Bioinformatics*, vol. 28, no. 12, pp. 1633–1640, Jun. 2012.

[28] R. Leaman, C.-H. Wei, and Z. Lu, "tmChem: A high performance approach for chemical named entity recognition and normalization," *J. Cheminform.*, vol. 7, no. S1, p. S3, Jan. 2015.

[29] C.-H. Wei, H.-Y. Kao, and Z. Lu, "GNormPlus: An integrative approach for tagging genes, gene families, and protein domains," *BioMed Res. Int.*, vol. 2015, pp. 1–7, Aug. 2015.

[30] R. Leaman and Z. Lu, "TaggerOne: Joint named entity recognition and normalization with semi-Markov Models," *Bioinformatics*, vol. 32, pp. 2839–2846, Sep. 2016.

[31] H. Li, Q. Chen, B. Tang, X. Wang, H. Xu, B. Wang, and D. Huang, "CNN-based ranking for biomedical entity normalization," *BMC Bioinf.*, vol. 18, no. S11, pp. 79–86, Oct. 2017.

[32] M. C. Phan, A. Sun, and Y. Tay, "Robust representation learning of biomedical names," in *Proc. 57th ACL*, Jul. 2019, pp. 3275–3285.

[33] D. Wright, "NormCo: Deep disease normalization for biomedical knowledge base construction," Ph.D. dissertation, Dept. Comput. Sci., UC San Diego, San Diego, CA, USA, 2019.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*. [Online]. Available: http://arxiv.org/abs/1706.03762

[36] 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, Feb. 2020.

[37] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," 2019, *arXiv:1903.10676*. [Online]. Available: http://arxiv.org/abs/1903.10676

[38] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," 2020, *arXiv:2007.15779*. [Online]. Available: http://arxiv.org/abs/2007.15779

[39] Z. Ji, Q. Wei, and H. Xu, "Bert-based ranking for biomedical entity normalization," *AMIA Summits Transl. Sci. Proc.*, vol. 2020, pp. 269–277, May 2020.

[40] D. Xu, Z. Zhang, and S. Bethard, "A generate-and-rank framework with semantic type regularization for biomedical concept normalization," in *Proc. Conf. ACL*, Jul. 2020, pp. 8452–8464.

[41] M. Sung, H. Jeon, J. Lee, and J. Kang, "Biomedical entity representations with synonym marginalization," 2020, *arXiv:2005.00239*. [Online]. Available: http://arxiv.org/abs/2005.00239

[42] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, *arXiv:1310.4546*. [Online]. Available: http://arxiv.org/abs/1310.4546

[43] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. EMNLP*, Oct. 2014, pp. 1532–1543.

[44] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365

[45] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 11, pp. 1297–1304, Jul. 2019.

[46] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *Proc. China Nat. Conf. CCL*, Oct. 2019, pp. 194–206.

[47] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: http://arxiv.org/abs/1609.08144

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. 29th IEEE Conf. CVPR*, Jun. 2016, pp. 2818–2826.

[49] D. Choi and H. Lee, "Extracting chemical-protein interactions via calibrated deep neural network and self-training," in *Proc. Conf. EMNLP, Findings*, 2020, pp. 2086–2095.

[50] R. I. Doğan, R. Leaman, and Z. Lu, "NCBI disease corpus: A resource for disease name recognition and concept normalization," *J. Biomed. Inform.*, vol. 47, pp. 1–10, Feb. 2014.

[51] A. P. Davis, T. C. Wiegers, M. C. Rosenstein, and C. J. Mattingly, "MEDIC: A practical disease vocabulary used at the comparative toxicogenomics database," *Database*, vol. 2012, Mar. 2012, Art. no. bar065.

[52] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "Comparative toxicogenomics database (CTD): Update 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1138–D1143, Jan. 2021.

[53] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: Gene-centered information at NCBI," *Nucleic Acids Res.*, vol. 33, no. 1, pp. D54–D58, Jan. 2005.

[54] S. Federhen, "The NCBI taxonomy database," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D136–D143, Jan. 2012.

[55] S. Sohn, D. C. Comeau, W. Kim, and W. J. Wilbur, "Abbreviation definition identification based on automatic precision estimates," *BMC Bioinf.*, vol. 9, no. 1, p. 402, Sep. 2008.
[56] I. Mondal, S. Purkayastha, S. Sarkar, P. Goyal, J. Pillai, A. Bhattacharyya, and M. Gattu, "Medical entity linking using triplet network," in *Proc. 2nd Clin. NLP Workshop*, Jun. 2019, pp. 95–100.
[57] J. Wermter, K. Tomanek, and U. Hahn, "High-performance gene name normalization with GENO," *Bioinformatics*, vol. 25, no. 6, pp. 815–821, Feb. 2009.
[58] L. Li, S. Liu, L. Li, W. Fan, D. Huang, and H. Zhou, "A multistage gene normalization system integrating multiple effective methods," *PLoS ONE*, vol. 8, no. 12, Dec. 2013, Art. no. e81956.
[59] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proc. TREC*, 1995, pp. 109–126.
[60] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," 2017, *arXiv:1706.04599*. [Online]. Available: http://arxiv.org/abs/1706.04599
[61] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2019, pp. 13888–13899.

HYEJIN CHO received the B.S. degree in computer engineering from Korea Aerospace University, South Korea, in 2013, the M.S. degree from the School of Information and Communications, Gwangju Institute of Science and Technology (GIST), South Korea, in 2015, and the Ph.D. degree in electrical engineering and computer science from GIST, in 2021. She is currently a Postdoctoral Researcher with the School of Electrical Engineering and Computer Science, GIST. Her research interests include text mining, named entity recognition, and named entity normalization.

DONGHA CHOI received the B.S. degree from the School of Undergraduate Studies, Daegu Gyeongbuk Institute of Science and Technology, South Korea, in 2019, and the M.S. degree in electrical engineering and computer science from Gwangju Institute of Science and Technology (GIST), South Korea, in 2021, where he is currently pursuing the Ph.D. degree with the AI Graduated School. His research interests include text mining, relation extraction, and knowledge distillation.

HYUNJU LEE received the B.S. degree in computer science from Korea Institute of Science and Technology, South Korea, in 1997, the M.S. degree in computer engineering from Seoul National University, South Korea, in 1999, and the Ph.D. degree in computer science from the University of Southern California, USA, in 2006. From 2006 to 2007, she was a Postdoctoral Research Fellow with Brigham and Women's Hospital, Harvard Medical School. Since 2007, she has been with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea. Her research interests include machine learning, natural language processing, and bioinformatics.