IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 4, 2019, accepted May 29, 2019, date of publication June 4, 2019, date of current version June 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2920708

A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining

DONGHYEON KIM[1], JINHYUK LEE[1], CHAN HO SO[2], HWISANG JEON[2], MINBYUL JEONG[1], YONGHWA CHOI[1], WONJIN YOON[1], MUJEEN SUNG[1], AND JAEWOO KANG[1,2]

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, South Korea
[2]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul 02841, South Korea

Corresponding author: Jaewoo Kang (kangj@korea.ac.kr)

This work was supported in part by the National Research Foundation of Korea under Grant NRF-2017R1A2A1A17069645 and Grant NRF-2016M3A9A7916996, and in part by the National IT Industry Promotion Agency, Development Project of the Precision Medicine Hospital Information System (P-HIS), under Grant C1202-18-1001.

ABSTRACT The amount of biomedical literature is vast and growing quickly, and accurate text mining techniques could help researchers to efficiently extract useful information from the literature. However, existing named entity recognition models used by text mining tools such as tmTool and ezTag are not effective enough, and cannot accurately discover new entities. Also, the traditional text mining tools do not consider overlapping entities, which are frequently observed in multi-type named entity recognition results. We propose a neural biomedical named entity recognition and multi-type normalization tool called BERN. The BERN uses high-performance BioBERT named entity recognition models which recognize known entities and discover new entities. Also, probability-based decision rules are developed to identify the types of overlapping entities. Furthermore, various named entity normalization models are integrated into BERN for assigning a distinct identifier to each recognized entity. The BERN provides a Web service for tagging entities in PubMed articles or raw text. Researchers can use the BERN Web service for their text mining tasks, such as new named entity discovery, information retrieval, question answering, and relation extraction. The application programming interfaces and demonstrations of BERN are publicly available at https://bern.korea.ac.kr.

INDEX TERMS Biomedical text mining, decision rules, multi-type, named entity recognition, neural networks, normalization, Web service.

I. INTRODUCTION

There are over 29 million articles in PubMed as of May 2019, and the amount of biomedical literature has been growing rapidly in recent years. Fast and precise text mining tools can reduce the amount of effort and time it takes researchers to find and extract useful information from the vast amount of biomedical literature. Researchers have used named entity recognition (NER) and named entity normalization (NEN) models to develop effective biomedical text mining tools for information retrieval [1], question answering [2], relation extraction [3], and so on.

Existing Web-based text mining tools such as tmTool [4], ezTag [5], and PubTerm [6] have obtained excellent NER performance on various types of biomedical entities. However, they have a few limitations. First, the Web-based text mining tools use older NER models which obtain lower performance than recent NER models. Moreover, pre-trained NER models such as tmChem [7] and DNorm [8] used by Web-based text mining tools cannot effectively discover new entities. Second, the Web-based text mining tools do not consider the different types of entities that frequently overlap in NER results [9].[1] For instance, in "The androgen is synthesized from . . . ," NER models can tag "androgen" as both a gene/protein and a drug/chemical because an androgen is a natural or synthetic steroid hormone. The correct entity type should be determined based the context of the sentence; in

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

[1]In the GENIA corpus, the percentage of nested named entities among all entities is 18.1%.

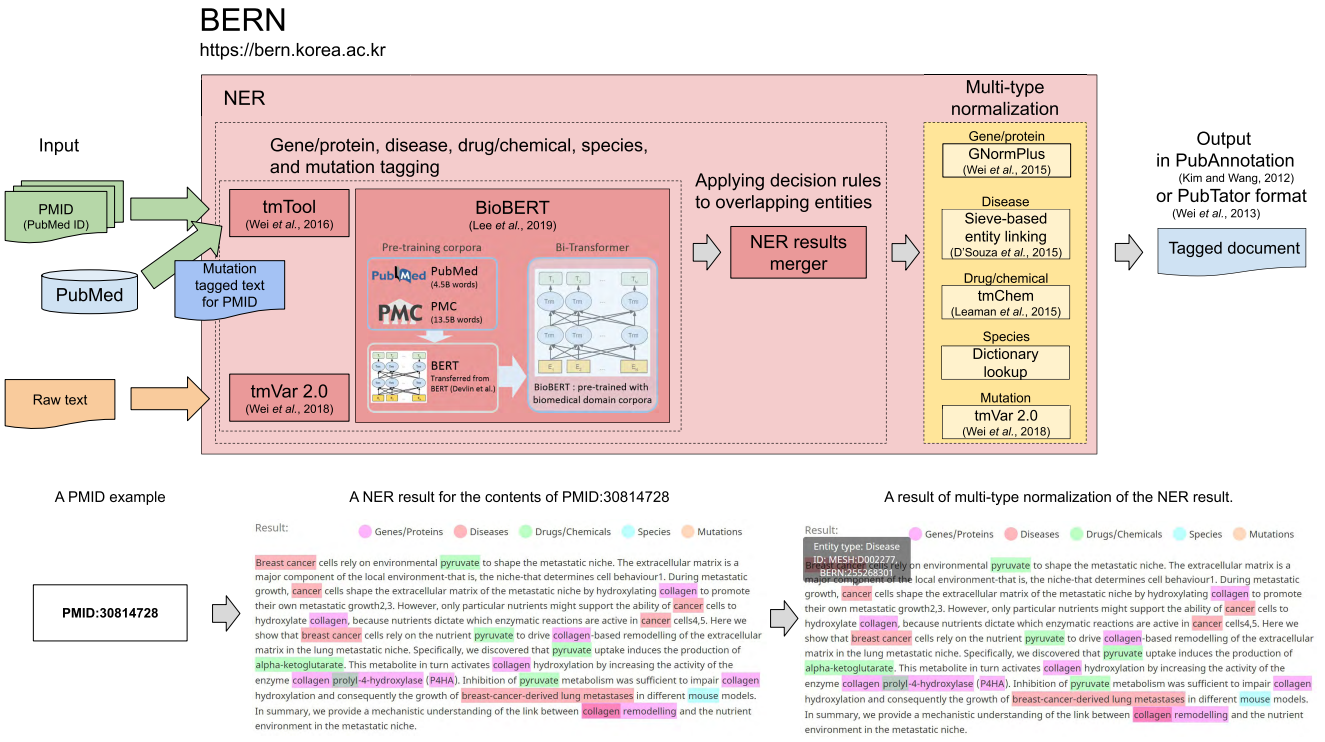

FIGURE 1. **Overview of the RESTful Web service of BERN.**

other words, if ''androgen'' refers to the synthetic hormone, NER models should tag ''androgen'' only as a drug/chemical. However, existing text mining tools have no pre-defined rules for such overlapping entities.

After NER, text mining tools need to normalize recognized entities since an entity can have multiple names (i.e., synonyms) and a name can be associated with multiple entities (i.e., polysemy). However, as normalization models vary greatly depending on the type of entity, it is difficult to build a text mining tool that performs normalization for multiple entity types. Generally, there are NER models for various types of biomedical entities in biomedical texts. However, it takes a considerable amount of time and effort to obtain computing resources and set up models for tagging entities.

We propose a neural **b**iomedical named **e**ntity **r**ecognition and multi-type **n**ormalization (BERN) tool that recognizes known entities and discovers new entities, and identifies the types of overlapping entities. The overview of the BERN Web service is shown in Fig. 1. When a PMID is inputted into BERN, it uses tmTool APIs to fetch texts annotated with mutations for the PMID. When raw text is inputted into BERN, it tags mutations in the text using tmVar 2.0 [10]. Next, BERN uses the BioBERT NER models of Lee *et al.* [11] to tag genes/proteins, diseases, drugs/chemicals, and species. After the multi-type NER, probability-based decision rules are applied to identify overlapping entities. Finally, normalization is performed for each entity type and the result is returned.

The BioBERT NER models obtained the highest F1-score in recognizing genes/proteins, diseases, and drugs/chemicals as shown in Table 1. Due to the lack of a high-quality public training set of mutations, BERN uses tmVar 2.0 as a pre-trained mutation NER model. Furthermore, BERN uses probability-based decision rules to determine whether to include all the overlapping entities or only the overlapping entities that are most likely to be entities predicted by the BioBERT NER models.

We also combined multiple named entity normalization models into one multi-type normalization model and integrated it into BERN to assign IDs to recognized entities. The multi-type normalization model uses a high-performance normalization model for each entity type, and uses a dictionary lookup such as SR4GN [23] for species. As a result, researchers can use the RESTful Web service of BERN to obtain NER and normalization results on PubMed articles or their raw text.

To the best of our knowledge, BERN is the first Web-based biomedical text mining tool that leverages neural network based NER models to recognize known entities and discover new entities. Our main contributions are as follows:

- BERN is a biomedical text mining tool that uses neural network based high-performance BioBERT NER models for recognizing known entities and discovering new entities.
- We developed probability-based decision rules for identifying the types of overlapping entities after conducting case studies.
- BERN uses the multi-type normalization model to assign a specific ID to each recognized entity.

TABLE 1. Performance comparison of NER models for genes/proteins, diseases, drugs/chemicals, species, and mutations at the entity mention level (The highest scores are in bold, and the second highest scores are underlined).

Text mining tools	Entity types	Pre-trained NER models	Test sets	Precision	Recall	F1-score
BERN		BioBERT [11]		**0.8516**	**0.8365**	**0.8440**
–		Sachan *et al.* [12]		0.8181	0.8157	0.8169
–		MTM-CW of Wang *et al.* [13]		0.8210	0.7942	0.8074
–	Gene/Protein	CollaboNet [16]	BC2GM [25]	0.8049	0.7899	0.7973
ezTag, tmTool, PubTerm		GNormPlus [14]		0.7840	0.7920	0.7880
–		Giorgi and Bader [15]		0.7862	0.7871	0.7866
–		LSTM-CRF (iii) of Habibi *et al.* [17]		0.7750	0.7813	0.7782
BERN		BioBERT [11]		0.8904	**0.8969**	**0.8936**
–		Sachan *et al.* [12]		0.8641	0.8831	0.8734
–		MTM-CW of Wang *et al.* [13]		0.8586	0.8642	0.8614
–		CollaboNet [16]		0.8548	0.8727	0.8636
–		Giorgi and Bader [15]		0.8262	0.8695	0.8472
–	Disease	LSTM-CRF (iii) of Habibi *et al.* [17]	NCBI disease [26]	0.8531	0.8358	0.8444
–		D3NER [18]		0.8503	0.8380	0.8441
–		Lou *et al.* [19]		**0.9072**	0.7489	0.8205
ezTag		[a]TaggerOne [20]		0.8510	0.8080	0.8290
ezTag		[b]TaggerOne [20]		0.8350	0.7960	0.8150
tmTool, PubTerm		DNorm [8]		0.8030	0.7630	0.7820
BERN		BioBERT [11]		0.9223	**0.9061**	**0.9141**
–		MTM-CW of Wang *et al.* [13]		0.9130	0.8753	0.8937
–		CollaboNet [16]		0.9078	0.8701	0.8885
–	Drug/Chemical	Giorgi and Bader [15]	BC4CHEMD [27]	0.8343	0.8883	0.8605
–		LSTM-CRF (iii) of Habibi *et al.* [17]		0.8783	0.8545	0.8662
–		Att-BiLSTM-CRF of Luo *et al.* [21]		**0.9229**	0.9001	0.9114
tmTool, PubTerm		tmChem (Model 2) [7]		0.8909	0.8575	0.8739
BERN		[c]BioBERT [11]		0.9384	0.8611	0.8981
–		Giorgi and Bader [15]		0.9280	0.9429	0.9354
–	Species	LSTM-CRF (iii) of Habibi *et al.* [17]	LINNAEUS [22]	0.9357	0.9324	0.9340
–		LINNAEUS [22]		**0.9710**	**0.9430**	**0.9570**
ezTag, tmTool, PubTerm		SR4GN [23]		0.8582	0.8528	0.8555
BERN, ezTag	Mutation	tmVar 2.0 [10]	tmVar2 [10]	0.9725	0.9040	0.9370
tmTool, PubTerm		tmVar [24]	MutationFinder [28]	0.9880	0.8962	0.9398

[a]Joint model of TaggerOne. [b]NER-only model of TaggerOne.
[c]For the LINNAEUS dataset, it is not clear whether training/dev/test sets of NER models are split in the same way.
For species, the BioBERT NER model uses the LINNAEUS dataset split of Pyysalo (https://github.com/spyysalo/linnaeus-corpus).

- BERN provides a Web service for tagging and normalizing entities in PubMed articles or raw text. The RESTful Web service of BERN is freely available at https://bern.korea.ac.kr.

II. RELATED WORK

A. NAMED ENTITY RECOGNITION FOR BIOMEDICAL TEXT MINING

For biomedical text, the word embeddings of Pyysalo *et al.* [29], which were trained on PubMed, PubMed Central Open Access (PMC OA) Subset, and English Wikipedia articles using word2vec [30], were widely used [13], [15]–[18]. Existing biomedical NER models [7], [10], [14], [23], [24] often use conditional random fields (CRFs) [31] or semi-Markov linear classifiers [20] with dictionaries to find entities in the biomedical literature. A CRF is flexible in terms of feature selection; however, it is computationally expensive.

In recent years, with the success of deep neural networks, Bi-LSTM with CRF [12], [13], [15]–[18] has been widely used to extract word sequence features. However, Bi-LSTM is limited to parallelization. The recently proposed Transformer [32], which is more parallelizable, obtained high-performance in natural language processing tasks without using recurrent networks or convolutional networks. The Transformer connects an encoder and a decoder through self-attention to be more parallelizable and to reduce its training time. Also, BERT (Bidirectional Encoder Representations from Transformers) [33], which can be used to understand deep contextual bidirectional language representations, was proposed. BERT pre-trains its weights on English Wikipedia and BooksCorpus, and then fine-tunes the pre-trained weights for each task.

B. RESOLVING OVERLAPPING ENTITIES

Since entities in biomedical text can overlap, it is necessary to decide which entities to select during or after NER. In previous studies, NER models were used to recognize entities in biomedical text even when the entities overlapped. Zhou [34] found patterns of nested entity names in the GENIA corpus, and proposed a pattern-based rule generation method to resolve the nested entity names. In recent years, Wang and Lu [35], and Katiyar and Cardie [36] proposed models for learning time-efficient hypergraph representations of overlapping entity mentions. Greenberg *et al.* [37] proposed a model which consists of Bi-LSTM and an expectation-maximization (EM) marginal CRF, and recognizes disjoint or partially overlapping sets of entity types.

However, their model does not have rules for determining the types of an entity mention if the mention belongs to different entity type spans.

C. NAMED ENTITY NORMALIZATION MODELS FOR BIOMEDICAL TEXT MINING

As mentioned in Section I, there are various types of entities which are referred to as different names in biomedical text. Thus, individual normalization models for each entity type have been proposed rather than an integrated normalization model. Lowercase conversion and abbreviation resolution are the most commonly used for normalizing biomedical entities. tmTool, PubTerm, ezTag, and BERN commonly use GNormPlus [14] for gene normalization, SR4GN [23] for species normalization (only dictionary lookup for BERN), and tmVar [10], [24] for mutation normalization. GNormPlus uses exact match and bag-of-words match to pair recognized names with concepts in Entrez Gene [38]. Also, GNormPlus applies Ab3P [39] to extract abbreviation pairs. SR4GN normalizes recognized species entities to the most specific concept if possible. Also, tmVar detects pairs of mutations and dbSNP RSIDs [40] using pattern matching and dictionary lookup.

On the other hand, text mining tools use different models for disease and chemical normalization. ezTag uses TaggerOne to normalize disease and chemical entities, and TaggerOne jointly performs NER and normalization using semi-Markov models. tmTool and PubTerm use DNorm [8], which is based on pairwise learning to rank (pLTR), to normalize disease entities. BERN uses the sieve-based entity linking approach of D'Souza and Ng [41] to normalize disease entities. Among the sieve based approaches, exact match, abbreviation expansion, and partial match were particularly effective. tmChem [7], which is used by tmTool, PubTerm and BERN, converts recognized chemical entity names and chemical entity names in the lexicon of tmChem to lowercase letters, and removes whitespace and punctuation. The lexicon is collected from MeSH [42] and ChEBI [43]. A chemical name in short form that can be recognized by Ab3P is assigned the same ID as the chemical name in long form.

III. METHODS

First, we describe the BioBERT NER models used for recognizing named entities in biomedical text, review cases of overlapping entities, and explain the decision rules developed for determining which entities to choose when they overlap. Next, we discuss the multi-type normalization model which normalizes the remaining entities.

A. BIOBERT FOR NAMED ENTITY RECOGNITION

BioBERT NER models used by BERN recognize known entities and discover new entities using WordPiece [44] embeddings. The word embeddings of Pyysalo *et al.* [29] suffer from the out-of-vocabulary problem. If a word in a text is not in the vocabulary of the embeddings, the embeddings cannot provide a rich representation for the word. On the other hand, the WordPiece embeddings are a way of dividing a word into several units (i.e., sub-word units) and expressing each unit. As a result, the WordPiece embeddings can be used to extract features of rare or unknown words, which is very helpful in discovering new entities.

BioBERT is initialized with the case-sensitive version of BERT-Base. BioBERT additionally pre-trains its weights on PubMed articles and PMC OA Subset articles, and fine-tunes the pre-trained weights for downstream tasks. The BioBERT NER models are fine-tuned as follows:

$$p(y_i = k|T_i) = \text{softmax}(T_i W^\top + b)_k, \quad k = 0, 1, \ldots, 6 \quad (1)$$

where p denotes the label probability, and $T_i \in \mathbb{R}^H$ denotes the final hidden representation for each token i. H is the hidden size, $W \in \mathbb{R}^{K \times H}$ is a classification layer, b is a bias, and K is 7. The classification loss L is calculated as follows:

$$L(\Theta) = -\frac{1}{N} \sum_{i=1}^{N} \log(p(y_i|T_i; \Theta)) \quad (2)$$

where Θ denotes trainable parameters, and N denotes sequence length.

BioBERT NER models compute the probabilities of the following seven tags: IOB2 tags (''I''nside, ''O''utside, ''B''egin) [45], ''X'' (a sub-token of WordPiece), ''[CLS]'' (the first token of every sequence for classification), ''[SEP]'' (a delimiter between sentences), and ''PAD'' (padding) of each word in a sentence. Note that the BioBERT NER models make predictions for ''I,'' ''O,'' and ''B'' tags but not for the ''X,'' ''[CLS],'' ''[SEP],'' or ''PAD'' tags. Words in a sentence are obtained using a tokenizer on a dataset with labels in CoNLL format [46] and then the sub-words of each word are obtained using the WordPiece tokenizer.

As a result, BERN can discover new entities using BioBERT NER models. As shown in Table 1, the BioBERT NER models used by BERN obtain the highest F1-scores on the test sets for all types except species. The BioBERT NER models of BERN outperform the NER models of tmTool, PubTerm, and ezTag on test sets of genes/proteins (BC2GM 5.6%), diseases (NCBI disease 6.46%), drugs/chemicals (BC4CHEMD 4.02%), and species (LINNAEUS 4.26%) in terms of F1-score. Also, the BioBERT NER models outperform the Bi-LSTM-CRF based NER models [13], [16]–[18], [21], the multi-task NER model [13], and the transfer learning NER models [12], [15] in recent years, on all the test sets except for the test set of species. Note that BioBERT NER models can recognize all types of entities if there is training data.

B. DECISION RULES FOR OVERLAPPING ENTITIES

1) CASE STUDIES

We performed comprehensive case studies on overlapping entities. First, we found that 26.2% of entities in 18.6 million PubMed articles[2] overlap. Among the entities in the

[2]We excluded PubMed articles that have titles but no abstracts.

TABLE 2. COMPLETE and PARTIAL overlap percentages of entity pairs (The numbers in each cell indicate the percentages of overlaps of the entity type pair in all COMPLETE (or all PARTIAL) overlaps. The numbers in parentheses are PARTIAL overlap percentages).

	Gene/ Protein	Disease	Drug/ Chemical	Species	Mutation
Gene/ Protein	–	1.76 (9.51)	**12.55** (**32.59**)	0.17 (12.83)	0.43 (1.49)
Disease	1.76 (9.51)	–	1.47 (5.57)	11.65 (8.93)	0.003 (0.012)
Drug/ Chemical	**12.55** (**32.59**)	1.47 (5.57)	–	0.03 (0.06)	0.06 (0.87)
Species	0.17 (12.83)	11.65 (8.93)	0.03 (0.06)	–	0.00001 (0.00163)
Mutation	0.43 (1.49)	0.003 (0.012)	0.06 (0.87)	0.00001 (0.00163)	–

articles, we also found 3.8 million cases where two or more entities overlap completely (COMPLETE overlap), and 9.6 million cases where they partially overlap (PARTIAL overlap). Table 2 shows the COMPLETE and PARTIAL overlap percentages of entities. Genes/proteins, diseases, and drugs/chemicals usually overlap. Genes/proteins and drugs/chemicals completely overlap the most (12.55%) of all entity types. Also, genes/proteins and drugs/chemicals have the largest number of PARTIAL overlaps (32.59%). Species usually overlap with genes/proteins (PARTIAL 12.83%) or diseases (COMPLETE 11.65%, PARTIAL 8.93%), and mutations generally overlap with genes/proteins (COMPLETE 0.43%, PARTIAL 1.49%). The proportion of species overlapping partially with genes/proteins (12.83%) is much greater than that of species overlapping completely with genes/proteins (0.17%) because there are many cases where genes/proteins are associated with Homo sapiens.

For a more detailed analysis, we calculate the PARTIAL overlap percentages of each entity pair, which are shown in Fig. 2. In our dataset, when genes/proteins and diseases partially overlap, a gene/protein mention is a part of a disease mention. And for each pair, a drug/chemical or a species mention is a part of a gene/protein or a disease mention. The sum of overlap percentage of overlap between the end of an entity and the beginning of another entity is quite low (< 1% except for ⟨drugs/chemicals and species⟩).

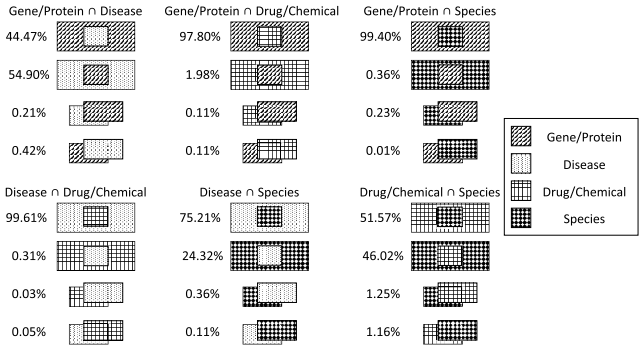

FIGURE 2. The percentages of partial overlapping entities of each entity pair.

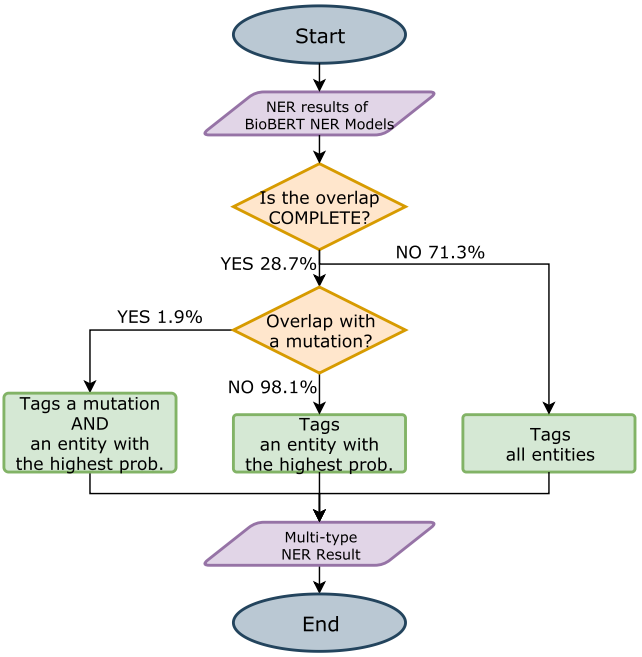

FIGURE 3. Decision rules for NER results.

2) DECISION RULES

After conducting the case studies above, we developed decision rules for overlapping entities in multi-type NER results. Unlike the model of Greenberg *et al.* [37], BERN uses decision rules for determining which entities to choose if overlapping entities are found. Fig. 3 shows the decision rules that BERN uses for the multi-type NER results. First, if the entities do not overlap completely (71.3% in our dataset), BERN tags all the entities. If entities overlap completely and there is a mutation among the entities (0.5% (28.7% × 1.9%)), BERN tags the mutation and entities that the mentions are most likely to be entities predicted by BioBERT NER models. Next, if entities completely overlap and all the entities are non-mutations (28.2% (28.7% × 98.1%)), only the entities with the highest probability of being an actual entity are tagged; the probability is calculated by BioBERT NER models. Because mutations in a text are typically in distinct formats, which makes it easier to more accurately recognize them, mutation NER models such as tmVar 2.0 used by BERN generally achieve much higher precision (over 97%) on mutations than on other entity types. In our evaluation, in most cases, if there is a mutation among the overlapping entities, all the remaining overlapping entities are wrong and the mutation is the correct answer.

We applied the decision rules to each test set, and the results are shown in Table 3. Since each test set in Table 3 has labels only for a particular entity type, there is no performance change if the NER model for the entity type labeled in the test set predicts that mentions are most likely to be entities of the entity type (e.g., In NCBI disease corpus, a NER model for diseases predicts that a mention is a disease.). If the NER models for the other entity types make stronger predictions on

TABLE 3. Examples of applying the decision rules of BERN to test sets.

Entity types	Test sets	Cases	Predictions	Examples
Gene/Protein	BC2GM	Preventing wrong answers	Drug/chemical	Results of these studies indicate that binding of **biotin** to the protein results in protection of regions of the central domain in the vicinity of the active site and the C-terminal domain from chemical cleavage.
Gene/Protein	BC2GM	Preventing correct answers	Drug/chemical	Similarly, **TAM-67** reverted the morphology of the AdoMetDC-antisense expressors.
Disease	NCBI disease	Preventing wrong answers	Gene/protein	A century after its recognition as a syndrome by Vaughan Pendred, the disease gene (**PDS**) was mapped to chromosome 7q22-q31.
Disease	NCBI disease	Preventing correct answers	Gene/protein	The APC gene was analysed in 190 unrelated **FAP** and 15 non-FAP colorectal cancer patients using denaturing gradient gel electrophoresis.
Drug/chemical	BC4CHEMD	Preventing wrong answers	Species	Understanding the interactions between metabolites isolated from **Achyrocline satureioides** in relation to its antibacterial activity.
Drug/chemical	BC4CHEMD	Preventing correct answers	Gene/protein	**Pinosylvin** Induces Cell Survival, Migration and Anti-Adhesiveness of Endothelial Cells via Nitric Oxide Production.

TABLE 4. The multi-type normalization model and dictionaries of BERN.

Entity types	Normalization models	Dictionaries	# of IDs	# of names	Avg. # of names per ID
Gene/Protein	GNormPlus	Entrez Gene [38]	139,375	248,581	1.8
Disease	Sieve-based entity linking [41]	MeSH [42], OMIM [53], SNOMED-CT [54], PolySearch2 [55]	32,954	172,650	5.2
Drug/Chemical	tmChem without Ab3P	MeSH [42], ChEBI [43], DrugBank [56], *US FDA approved drugs	518,223	2,571,570	5.0
Species	Dictionary lookup	NCBI Taxonomy	398,037	3,119,005	7.8
Mutation	tmVar 2.0	dbSNP [40], ClinVar [57]	208,474	302,498	1.5
Total			1,297,063	6,414,304	4.9

TABLE 5. The performance of the multi-type normalization model (i.e., the combined normalization models) of BERN. The authors of tmChem did not report the normalization performance of tmChem independently.

Entity types	Normalization models	Test sets	Precision %	Recall %	F1-score %	Accuracy %
Gene/Protein	GNormPlus	BC2 Gene Normalization, human species [25]	87.1	86.4	86.7	–
		BC3 Gene Normalization, multispecies [58]	–	–	50.1	–
Disease	Sieve-based entity linking	ShARe/CLEF eHealth Challenge corpus [59]	–	–	–	90.75
		NCBI disease	–	–	–	84.65
Mutation	tmVar 2.0	OSIRISv1.2 [60]	97.20	80.62	88.14	–
		Thomas [61]	89.94	88.24	89.08	–

which mentions are most likely to be entities on wrong mentions than predictions of the NER model for the entity type (e.g., In BC2GM, a NER model for drugs/chemicals predicts that a non-gene/protein mention is a drug/chemical.), they can avoid giving the wrong answer due to the decision rules, which helps reduce the false positive rate. Conversely, if the NER models for other entity types make stronger predictions on correct mentions than predictions of the NER model for the entity type (e.g., In BC4CHEMD, a NER model for diseases predicts that a drug/chemical mention is a disease.), the true positive rate is reduced. Therefore, the decision rules improve the precision and lower the recall.

In the GENIA corpus, only protein entities are labeled among the entity types of the NER models of BERN, making it difficult to compare the probabilities of other types of entities overlapping. Also, the BC5CDR corpus does not contain any labels for overlapping entities.

C. THE MULTI-TYPE NORMALIZATION MODEL

BERN uses the multi-type normalization model to more clearly distinguish entities. Table 4 shows the statistics of the normalization model used by BERN. We added the disease names in the PolySearch2 dictionary (76,001 names of 27,658 diseases) to the sieve-based entity linking dictionary (76,237 names of 11,915 diseases) to increase the number of normalizable entities. We also added the drug names in DrugBank [56] and US FDA to the tmChem dictionary. Due to the lack of normalization models for species, BERN normalizes species by dictionary lookup, as mentioned above. Using tmVar 2.0, we made a dictionary of mutations with normalized mutation names; a mutation with several names was given one normalized name or ID.

According to the statistics, drugs/chemicals have the highest number of unique IDs (40% of the total), and species have the most names per entity. If the normalization model fails to normalize a recognized entity, the model returns a Concept Unique Identifier-less (CUI-less) for the entity.

Table 5 shows the performance of the multi-type normalization model (i.e., integrated normalization models) of BERN. For genes/proteins, there are 75 kinds of species in the BC3 Gene Normalization (BC3GN) test set, but GNormPlus focuses on only 7 kinds of species. As a result, GNormPlus obtains a much lower F1-score of 36.6% on the multispecies test set (BC3GN) than F1-score on the human species test

TABLE 6. Normalization results of recognized entities in 19.4 million PubMed articles.

Entity types	# of recognized entities	# of normalized entities	# of articles with each entity type	Avg. # of recognized entities per article	Avg. # of normalized entities per article
Gene/Protein	73,655,197	39,299,648 (53.4%)	7,844,921	9.4	5.0
Disease	91,204,877	82,242,319 (90.2%)	12,461,907	7.3	6.6
Drug/Chemical	76,367,837	63,409,437 (83.0%)	9,568,465	8.0	6.6
Species	60,389,187	57,275,053 (94.8%)	13,580,610	4.4	4.2
Mutation	1,279,525	1,279,525 (100%)	310,210	4.1	4.1

TABLE 7. Runtime statistics for 10K PubMed articles (seconds per article, STD: standard deviation).

Models	Average ± STD
Getting a mutation tagged PubMed article using tmTool APIs	1.254 ± 0.103
Gene/protein, disease, drug/chemical, and species NER	0.411 ± 0.325
Multi-type normalization	0.022 ± 0.043
Total	1.688 ± 0.355

set (BC2GN). For mutations, tmVar 2.0 achieved F1-scores close to 90% on two corpora: OSIRISv1.2 and the Thomas corpus.

Also, we tested the multi-type normalization model on 19.4 million PubMed articles. The results are shown in Table 6. Because we constructed a gene dictionary of mostly Homo sapiens (i.e., human species), the percentage of normalized genes is low (53.4%). The result obtained by the sieve-based entity linking model had a high percentage (90.2%) of normalized diseases. Also, the percentage of normalized species was the highest (94.8%) and species were mentioned in most of the sample articles (70%). Mutations were mentioned the least in the articles (1.6%), but the percentage of normalized mutations was 100% since tmVar 2.0 for mutations always provides the normalized names of recognized entities. Although drugs/chemicals have the highest number of IDs and the second largest number of names in the dictionary as shown in Table 4, the percentage of normalized drugs/chemicals is 83.0%, which may be because Ab3P (abbreviation resolution), which is used by tmChem, was not applied.

IV. IMPLEMENTATION

The RESTful Web service of BERN was implemented using Python and Node.js. BERN run BioBERT NER models which are pre-trained with TensorFlow[3], on our server to recognize incoming biomedical text such as PubMed articles and raw text. The server specifications are as follows:

- Operating system: Ubuntu 18.04.2 LTS
- CPU: Intel Xeon E5-2687W v3
- RAM size: 128 gigabytes (GB)
- GPU: NVIDIA Titan X (Pascal) with 12 GB of memory
- Hard disk drive size: 2 terabytes

Four BioBERT NER models for genes/proteins, diseases, drugs/chemicals, and species, use 2.4 GB (4×0.6 GB) of GPU memory. We use 8 NVIDIA V100 GPUs for pre-training BioBERT, and we use a NVIDIA Titan X GPU for making predictions. And, we use the following training datasets to fine-tune each BioBERT NER model: BC2GM for genes, NCBI disease for diseases, BC4CHEMD for drugs/chemicals, and LINNAEUS for species. GNormPlus uses 8 to 16 GB, and tmVar 2.0 uses 4 to 8 GB of memory. And, the load time of the GNormPlus gene dictionary is about 5 seconds and the load time of the tmVar 2.0 part-of-speech tagger is about 1 second. To reduce their load time, we run GNormPlus and tmVar 2.0 processes in the background on the server.

Table 7 shows the runtime statistics of BERN. The statistics show that tmTool API calls and the NER models used by BERN have the longest time (98.6%) in each run. If there is no recognized entity, the multi-type normalization model is not used. In this experiment, only one article was assigned to each batch.

A. DEMONSTRATIONS

In the "Text" tab of BERN Web service, researchers can obtain NER+NEN results of submitted raw text in PubAnnotation JSON format, and see the visualized results under the text window. Also, as the BERN demonstration shows, entities are highlighted in their entity type color. When the mouse cursor is placed on an entity name, its entity type and entity ID are displayed in a tooltip.

Fig. 4 shows a BERN demonstration which uses the title and abstract of PMID:30429607 article. Also, in the "PMID" tab of BERN Web service, researchers can enter one or more PMIDs to obtain results in PubAnnotation JSON or PubTator format.

B. APPLICATION PROGRAMMING INTERFACES

BERN application programming interfaces (APIs) return NER+NEN results for PMIDs and a raw text. For PMIDs, the Uniform Resource Locator (URL) form used by BERN APIs is https://bern.korea.ac.kr/pubmed/<PMID(s)>[/pubtator]. In this URL form, the PMID parameter is required but the format parameter, "/pubtator", is optional. For the convenience of researchers, we made it possible to obtain NER+NEN results by simply including one or more PMIDs in the URL. For the "Single PMID" URL of Table 8, BERN returns an NER+NEN result for a PubMed article with the following PMID:29446767. In addition, researchers can enter multiple comma-separated PMIDs to obtain NER+NEN

[3]https://www.tensorflow.org

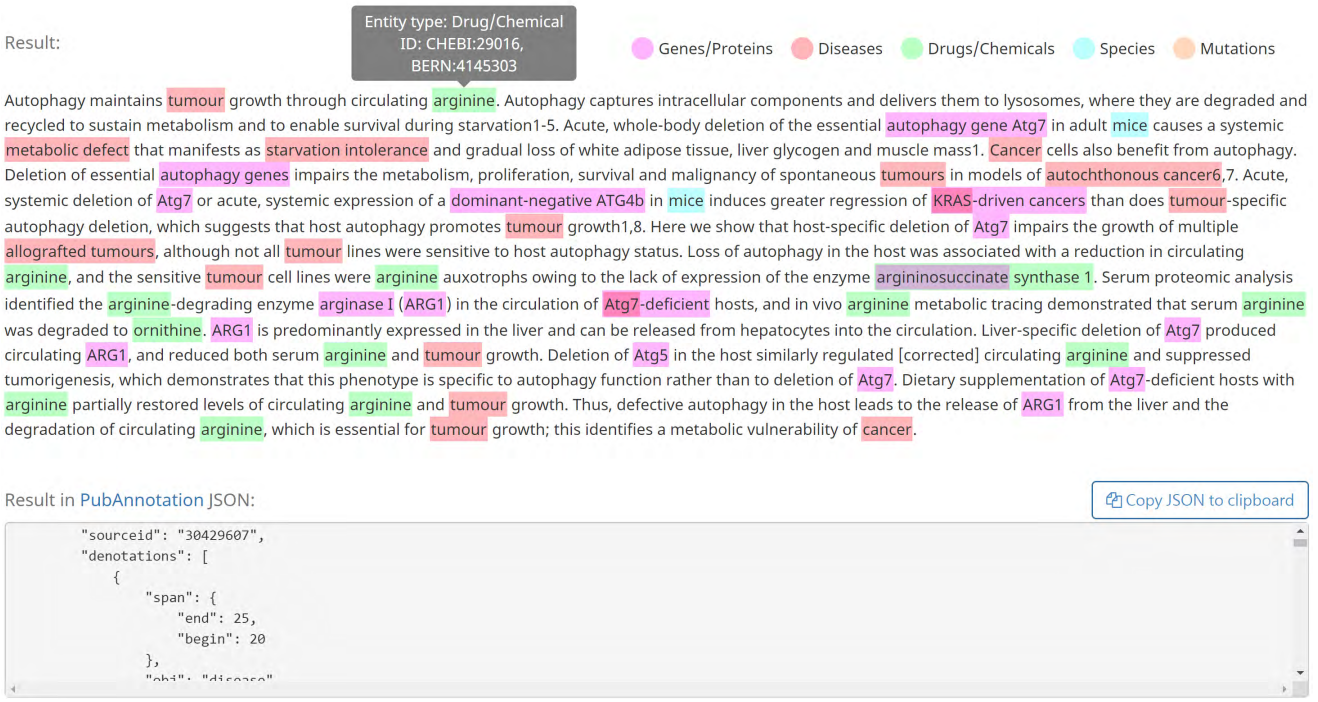

FIGURE 4. BERN demonstration of PMID:30429607 (best viewed in color mode).

TABLE 8. BERN APIs and URL examples (PMID: PubMed ID).

APIs	URL examples
Single PMID	https://bern.korea.ac.kr/pubmed/29446767
PMIDs	https://bern.korea.ac.kr/pubmed/29446767,25681199
In PubTator format	https://bern.korea.ac.kr/pubmed/29446767/pubtator
Raw texts	https://bern.korea.ac.kr/plain/

results for multiple PubMed articles at the same time.[4] For example, Table 8 shows the "PMIDs" URL for two PMIDs (PMID:29446767, PMID:25681199).

PubAnnotation JSON is the default format of the result of the APIs, which is also the format of the result of the demonstrations. Also, researchers can obtain results in PubTator format by adding "/pubtator" to the end of a URL. The "In PubTator format" URL of Table 8 is an example of the result in PubTator format. In the PubTator format result, the title is included in the first line, the abstract is included in the second line, and then the NER+NEN result ⟨PMID, start offset, end offset, entity name, entity type, entity ID⟩ of BERN is in each line. Furthermore, BERN uses a HTTPS POST method where researchers must include their raw texts and the "Raw texts" URL in Table 8 in their code. For ease of use, we provide a sample Python code for API calls at https://bern.korea.ac.kr.

V. DISCUSSION

A. USE CASES

There are many use cases where BERN can be used.

[4]Note that BERN allows up to 10 PMIDs at a time.

- Discovery of new named entities: As mentioned earlier, BioBERT NER models can be used to discover new entities from the latest biomedical literature. Table 9 shows new entity examples of each type discovered by the BioBERT NER models. Although the new entities are not in the dictionaries of BERN, the BioBERT NER models can discover the new entities. For instance, the chemical compound "pentandricine" is not included in the dictionaries of BERN, but the BioBERT NER models used by BERN accurately recognize the entity as a new chemical compound. Note that BioBERT NER models usually discovers new entities when sufficient contextual information (i.e., a complete sentence) is given (e.g., the chemical compound "osimertinib" without enough context may not be recognized by BERN).
- Information retrieval: BERN can serve as a fundamental NER+NEN model for various text mining tools. In the field of information retrieval, entity-based search engines such as LitVar [1] and BEST [62] can use BERN to find entities in queries and documents. BERN can greatly improve the performance of entity search engines in finding co-occurring entities in queries and documents.
- Question answering: BERN can recognize biomedical named entities in questions and passages in question answering tasks such as BioASQ Task B [63], [64], and help improve performance, especially on "what" and "which" questions by classifying whether a span in a passage is an entity or not.
- Relation extraction: BERN can generate rich datasets for downstream biomedical text mining tasks such as

TABLE 9. Discovered new entity examples of each entity type (discovered new entities (bold) and known entities (underlined)).

Entity types	Example sentences
Gene/Protein	... To decipher the synaptic substrate of hyperexcitability, we examined pan-neuronal Tsc1 knockout mouse and found a reduction in surface expression of a GABA receptor (GABAR) subunit but not **AMPA** receptor (**AMPAR**) subunit. ... (PMID:30683131)
Gene/Protein	... This metabolite in turn activates collagen hydroxylation by increasing the activity of the enzyme **collagen prolyl-4-hydroxylase** (P4HA). ... (PMID:30814728)
Disease	... A recent observational study in a large cohort of critically ill patients confirms the association between **hyperlactatemia** and mortality. ... (PMID:19691816)
Disease	... In contrast, high-performance liquid chromatography tandem mass spectrometry showed **hypercholanaemia** and high concentrations of biliverdin IXα in serum, urine, bile and milk. **Hyperbiliverdinaemia** disappeared after surgical correction of the cholestasis. ... (PMID:21278388)
Drug/Chemical	... MEK inhibition with trametinib synergized with **osimertinib** to block growth. Alternately, a pan-RAF inhibitor as a single agent blocked growth of all cell lines with **mutant EGFR** and **BRAF fusion**. ... (PMID:30831205)
Drug/Chemical	... A new limonoid, **pentandricine** (1), along with three known limonoids, **ceramicine B** (2), **6-de(acetyloxy)**-23-**oxochisocheton** (3), **6-de(acetyloxy)-23-oxo-7-O-deacetylchisocheton** (4), have been isolated from the stembark of Chisocheton pentandrus. ... (PMID:29368952)
Species	... Specific evidence of coastal contamination of the marine ecosystem with the zoonotic protozoan parasite, Toxoplasma gondii, and extensive infection of southern sea otters (**Enhydra lutris nereis**) along the California coast was documented by this study. ... (PMID:12076629)
Species	The complete mitochondrial genome sequence of the Chinese Serow, **Capricornis milneedwardsii** (Cetartiodactyla: Caprinae). ... (PMID:24438312)

relation extraction [65]. For instance, BERN can easily extract sentences with two or more recognized named entities from a biomedical corpus. Such sentences can be annotated, and the relationship of the recognized entities can be extracted from an existing database to generate a training dataset.

- A useful text mining tool: Using APIs, researchers can obtain NER+NEN results for texts from highly accessible Web services. Researchers can use commonly used entity IDs (e.g., HGNC IDs for genes, and MeSH IDs for diseases) [66] in the results of BERN more effectively for their text mining tasks.

B. ADVANTAGES AND LIMITATIONS OF HAVING A SEPARATE NER MODEL FOR EACH ENTITY TYPE

Using a separate NER model for each entity type has the following advantages. First, the best performing model can be used for each entity type. In BERN, we can substitute the BioBERT NER models with new state-of-the-art models. Second, adding a new NER model for a different entity type to existing NER models is relatively easy. On the other hand, a single NER model that recognizes multiple entity types may need to be trained again on the dataset due to the changes in the architecture of the model.

However, having a separate NER model for each entity type can lower the efficiency. Multithreading can be performed to reduce the processing time, but it requires a larger number of computing resources. Also, it is possible to improve NER performance by multi-task learning which trains a NER model to multiple tasks. A multi-task NER model can show higher performance than single-task models for various tasks with relatively few computing resources.

C. ADDITIONAL DEPENDENCIES OF BERN

As mentioned in the Section I, since BioBERT does not have a mutation NER model, BERN uses tmVar 2.0 and tmTool APIs for mutations. If we can obtain a high-quality training set and build a mutation NER model that achieves higher performance than tmVar 2.0, BERN would not have to use tmVar 2.0 and tmTool APIs.

D. OVERCOMING THE NETWORK DISTANCE BETWEEN A SERVER AND A CLIENT

The network delay tends to increase with the distance between a server and a client. In such cases, cloud computing can be used. Researchers can run a cloud machine in the region closest to the server to reduce the distance between the server and the client. For example, a researcher can launch Elastic Compute Cloud (EC2) instances or Lambda functions of Amazon Web Services in the same region or near the region of the server.

VI. CONCLUSION

Our proposed tool BERN recognizes known entities and discovers new entities using BioBERT NER models. The BioBERT models outperform NER models of existing Web-based text mining tools in terms of F1-score on genes/proteins, diseases, drugs/chemicals, and species. After reviewing a vast number of cases of overlapping entities, we developed and used the decision rules on identifying the entity types of overlapping entities which occur frequently in multi-type NER results. For assigning a specific ID to each recognized entity, multiple normalization models are combined and integrated into BERN. The RESTful Web service of BERN is freely available and can be used for various types of input. Researchers can use BERN for text mining tasks such as new named entity discovery, information retrieval, question answering, and relation extraction.

For future work, we plan to use a multi-task NER model for higher NER performance. Also, we will develop a novel entity type decision model that uses transfer learning to consider not only the entity types and probabilities of overlapping entities but also the deeper contextual meaning of a text.

ACKNOWLEDGMENT

We greatly appreciate Susan Kim for editing the manuscript.

REFERENCES

[1] A. Allot, Y. Peng, C.-H. Wei, K. Lee, L. Phan, and Z. Lu, "LitVar: A semantic search engine for linking genomic variant data in PubMed and PMC," *Nucleic Acids Res.*, vol. 46, pp. W530–W536, Jul. 2018.

[2] V. Sharma, N. Kulkarni, S. Pranavi, G. Bayomi, E. Nyberg, and T. Mitamura, "BioAMA: Towards an end to end biomedical question answering system," in *Proc. BioNLP Workshop*, Melbourne, VIC, Australia, 2018, pp. 109–117. [Online]. Available: http://aclweb.org/anthology/W18-2312

[3] B. Percha and R. B. Altman, "A global network of biomedical relationships derived from text," *Bioinformatics*, vol. 34, pp. 2614–2624, Aug. 2018.

[4] C. H. Wei, R. Leaman, and Z. Lu, "Beyond accuracy: Creating interoperable and scalable text-mining Web services," *Bioinformatics*, vol. 32, no. 12, pp. 1907–1910, Jun. 2016.

[5] D. Kwon, S. Kim, C. H. Wei, R. Leaman, and Z. Lu, "ezTag: Tagging biomedical concepts via interactive learning," *Nucleic Acids Res.*, vol. 46, pp. W523–W529, Jul. 2018.

[6] J. Garcia-Pelaez, D. Rodriguez, R. Medina-Molina, G. Garcia-Rivas, C. Jerjes-Sánchez, and V. Trevino, "PubTerm: A Web tool for organizing, annotating and curating genes, diseases, molecules and other concepts from PubMed records," *Database*, vol. 2019, p. 137, Jan. 2019.

[7] R. Leaman, C. H. Wei, and Z. Lu, "tmChem: A high performance approach for chemical named entity recognition and normalization," *J. Cheminformat.*, vol. 1, p. S3, Jan. 2015.

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

[9] M. G. Sohrab and M. Miwa, "Deep exhaustive model for nested named entity recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 2843–2849. [Online]. Available: https://www.aclweb.org/anthology/D18-1309

[10] C. H. Wei, L. Phan, J. Feltz, R. Maiti, T. Hefferon, and Z. Lu, "tmVar 2.0: Integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine," *Bioinformatics*, vol. 34, pp. 80–87, Jan. 2018.

[11] 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," Jan. 2019, *arXiv:1901.08746*. [Online]. Available: https://arxiv.org/abs/1901.08746

[12] D. S. Sachan, P. Xie, M. Sachan, and E. P. Xing, "Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition," in *Proc. Mach. Learn. Res.*, Palo Alto, CA, USA, vol. 85, Aug. 2018, pp. 383–402. [Online]. Available: http://proceedings.mlr.press/v85/sachan18a/sachan18a.pdf

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

[14] 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, Apr. 2015, Art. no. 918710.

[15] J. M. Giorgi and G. D. Bader, "Transfer learning for biomedical named entity recognition with neural networks," *Bioinformatics*, vol. 34, no. 23, pp. 4087–4094, Dec. 2018.

[16] W. Yoon, C. H. So, J. Lee, and J. Kang, "CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition," Sep. 2018, *arXiv:1809.07950*. [Online]. Available: https://arxiv.org/abs/1809.07950

[17] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, pp. i37–i48, Jul. 2017.

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

[19] Y. Lou, Y. Zhang, T. Qian, F. Li, S. Xiong, and D. Ji, "A transition-based joint model for disease named entity recognition and normalization," *Bioinformatics*, vol. 33, pp. 2363–2371, Aug. 2017.

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

[21] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition," *Bioinformatics*, vol. 34, pp. 1381–1388, Apr. 2018.

[22] M. Gerner, G. Nenadic, and C. M. Bergman, "LINNAEUS: A species name identification system for biomedical literature," *BMC Bioinf.*, vol. 11, p. 85, Feb. 2010.

[23] C. H. Wei, H. Y. Kao, and Z. Lu, "SR4GN: A species recognition software tool for gene normalization," *PLoS ONE*, vol. 7, p. e38460, Jun. 2012.

[24] C. H. Wei, B. R. Harris, H. Y. Kao, and Z. Lu, "tmVar: A text mining approach for extracting sequence variants in biomedical literature," *Bioinformatics*, vol. 29, pp. 1433–1439, Apr. 2013.

[25] 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.

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

[27] M. Krallinger *et al.*, "The CHEMDNER corpus of chemicals and drugs and its annotation principles," *J. Cheminf.*, vol. 7, p. S2, Jan. 2015.

[28] J. G. Caporaso, W. A. Baumgartner, D. A. Randolph, K. B. Cohen, and L. Hunter, "MutationFinder: A high-performance system for extracting point mutation mentions from text," *Bioinformatics*, vol. 23, pp. 1862–1865, Jul. 2007.

[29] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou, "Distributional semantics resources for biomedical text processing," in *Proc. 5th Int. Symp. Lang. Biol. Med.*, Tokyo, Japan, 2013, pp. 39–44. [Online]. Available: http://bio.nlplab.org/pdf/pyysalo13literature.pdf

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, Stateline, NV, USA, 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[31] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, Williamstown, MA, USA, 2001, pp. 282–289. [Online]. Available: https://dl.acm.org/citation.cfm?id=655813

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008. [Online]. Available: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

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

[34] G. D. Zhou, "Recognizing names in biomedical texts using mutual information independence model and SVM plus sigmoid," *Int. J. Med. Informat.*, vol. 75, pp. 456–467, Jun. 2006.

[35] B. Wang and W. Lu, "Neural segmental hypergraphs for overlapping mention recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 204–214. [Online]. Available: https://aclweb.org/anthology/D18-1019

[36] A. Katiyar and C. Cardie, "Nested named entity recognition revisited," in *Proc. NAACL-HLT*, New Orleans, LA, USA, 2018, pp. 861–871. [Online]. Available: https://www.aclweb.org/anthology/N18-1079

[37] N. Greenberg, T. Bansal, P. Verga, and A. McCallum, Brussels, Belgium. "Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2824–2829. [Online]. Available: https://www.aclweb.org/anthology/D18-1306

[38] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gCene: Gene-centered information at NCBI," *Nucleic Acids Res.*, vol. 39, pp. D52–D57, Nov. 2010.

[39] S. Sohn, D. C. Comeau, W. Kim, and W. J. Wilbur, "Abbreviation definition identification based on automatic precision estimates," *BMC Bioinf.*, vol. 9, p. 402, Sep. 2008.

[40] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: The NCBI database of genetic variation," *Nucleic acids Res.*, vol. 29, pp. 308–311, Jan. 2001.

[41] J. D'Souza and V. Ng, "Sieve-based entity linking for the biomedical domain," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, Beijing, China, 2015, pp. 297–302. [Online]. Available: http://www.aclweb.org/anthology/P15-2049

[42] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bull. Med. Library Assoc.*, vol. 88, pp. 265–266, Jul. 2000.

[43] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, "ChEBI: A database and ontology for chemical entities of biological interest," *Nucleic acids Res.*, vol. 36, pp. D344–D350, Oct. 2007.

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

[45] E. F. T. K. Sang and J. Veenstra, "Representing text chunks," in *Proc. 9th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Bergen, Norway, 1999, pp. 173–179. [Online]. Available: https://dl.acm.org/citation.cfm?id=977059

[46] S. Buchholz and E. Marsi, "CoNLL-X shared task on multilingual dependency parsing," in *Proc. 10th Conf. Comput. Natural Lang. Learn. (CoNLL-X)*, New York City, NY, USA, 2006, pp. 149–164. [Online]. Available: https://dl.acm.org/citation.cfm?id=1596305

[47] J. D. Kim and Y. Wang, "PubAnnotation: A persistent and sharable corpus and annotation repository," in *Proc. Workshop Biomed. Natural Lang. Process.*, Montréal, QC, Canada, 2012, pp. 202–205. [Online]. Available: http://www.aclweb.org/anthology/W12-2425

[48] C. H. Wei, H. Y. Kao, and Z. Lu, "PubTator: A Web-based text mining tool for assisting biocuration," *Nucleic Acids Res.*, vol. 41, pp. W518–W522, Jul. 2013.

[49] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[50] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[51] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 357–370, Dec. 2016.

[52] Y. Kim, Y. Jernite, D. Sontag, A. M. Rush, "Character-aware neural language models," in *Proc. 13th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016, pp. 2741–2749. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewFile/12489/12017

[53] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 33, pp. D514–D517, Jan. 2005.

[54] K. Donnelly, "SNOMED-CT: The advanced terminology and coding system for eHealth," *Stud. Health Technol. Informat.*, vol. 121, pp. 279–290, Jan. 2006.

[55] Y. Liu, Y. Liang, and D. Wishart, "PolySearch2: A significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more," *Nucleic Acids Res.*, vol. 43, pp. W535–W542, Jul. 2015.

[56] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, "DrugBank 4.0: Shedding new light on drug metabolism," *Nucleic Acids Res.*, vol. 42, pp. D1091–D1097, Nov. 2013.

[57] M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, and D. R. Maglott, "ClinVar: Public archive of interpretations of clinically relevant variants," *Nucleic Acids Res.*, vol. 44, pp. D862–D868, Jan. 2016.

[58] Z. Lu *et al.*, "The gene normalization task in BioCreative III," *BMC Bioinf.*, vol. 12, p. S2, Oct. 2011.

[59] S. Pradhan, N. Elhadad, B. R. South, D. Martinez, L. Christensen, A. Vogel, H. Suominen, W. W. Chapman, and G. Savova, "Task 1: ShARe/CLEF eHealth evaluation lab 2013," in *Proc. CLEF (Work. Notes)*, 2013, pp. 212–231.

[60] L. I. Furlong, H. Dach, M. Hofmann-Apitius, and F. Sanz, "OSIRISv1.2: A named entity recognition system for sequence variants of genes in biomedical literature," *BMC Bioinf.*, vol. 9, p. 84, Feb. 2008.

[61] P. E. Thomas, R. Klinger, L. I. Furlong, M. Hofmann-Apitius, and C. M. Friedrich, "Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers," *BMC Bioinf.*, vol. 12, p. S4, Jul. 2011.

[62] S. Lee, D. Kim, K. Lee, J. Choi, S. Kim, M. Jeon, S. Lim, D. Choi, S. Kim, A. C. Tan, and J. Kang, "BEST: Next-generation biomedical entity search tool for knowledge discovery from biomedical literature," *PLoS ONE*, vol. 11, p. e0164680, Oct. 2016.

[63] G. Tsatsaronis, M. Schroeder, G. Paliouras, Y. Almirantis, I. Androutsopoulos, E. Gaussier, P. Gallinari, T. Artieres, M. R. Alvers, M. Zschunke, and A.-C. N. Ngomo, "BioASQ: A challenge on large-scale biomedical semantic indexing and question answering," in *Proc. AAAI Inf. Retr. Knowl. Discovery Biomed. Text*, 2012, pp. 92–98. [Online]. Available: https://www.aaai.org/ocs/index.php/FSS/FSS12/paper/viewPaper/5600

[64] M. Wasim, W. Mahmood, M. N. Asim, and M. U. Khan, "Multi-label question classification for factoid and list type questions in biomedical question answering," *IEEE Access*, vol. 7, pp. 3882–3896, Dec. 2018.

[65] B. Xu, X. Shi, Z. Zhao, and W. Zheng, "Leveraging biomedical resources in bi-LSTM for drug-drug interaction extraction," *IEEE Access*, vol. 6, pp. 33432–33439, Jun. 2018.

[66] G. Wu, Y. He, and X. Hu, "Entity linking: An issue to extract corresponding entity with knowledge base," *IEEE Access*, vol. 6, pp. 6220–6231, Jan. 2018.

DONGHYEON KIM received the B.S. degree in computer science education from Korea University, Seoul, South Korea, in 2011, where he is currently pursuing the Ph.D. degree in computer science.

From 2014 to 2016, he was the CTO and a Co-Founder of Opinion8, Seoul, South Korea. From 2016 to 2018, he was a Data Scientist and a Software Engineer with Konolabs, Inc., Seoul. His current research interests include information retrieval, bioinformatics, machine learning, and recommender systems.

JINHYUK LEE received the B.E. degree in computer and communication engineering from Korea University, Seoul, South Korea, in 2016, where he is currently pursuing the Ph.D. degree. His current research interests include natural language processing, question answering systems, and biomedical text mining.

CHAN HO SO received the B.S. degree in computer science from Korea University, Seoul, South Korea, in 2018, where he is currently pursuing the M.S. degree in computer science. His current research interests include bioinformatics and named entity recognition.

HWISANG JEON received the B.S. degree in biotechnology from Korea University, Seoul, South Korea, in 2018, where he is currently pursuing the M.S. degree in computer science. His current research interests include bioinformatics, machine learning, and integrated multiomics.

MINBYUL JEONG received the B.S. degree in computer science from Korea University, Seoul, South Korea, in 2019, where he is currently pursuing the Ph.D. degree in computer science. His current research interests include natural language processing and bioinformatics.

YONGHWA CHOI received the B.S. degree in computer science from Korea University, Seoul, South Korea, in 2015, where he is currently pursuing the Ph.D. degree in computer science. His current research interests include natural language understanding, question answering, and bioinformatics.

WONJIN YOON received the B.S. degree in computer science from Korea University, Seoul, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in computer science. His current research interests include bioinformatics and named entity recognition.

MUJEEN SUNG received the B.S. degree in computer and communication engineering from Korea University, Seoul, South Korea, in 2015, where he is currently pursuing the Ph.D. degree in computer science. His current research interests include biomedical text mining, bioinformatics, and information retrieval.

From 2015 to 2018, he was a Software Engineer with General Electrics, Seoul, South Korea.

JAEWOO KANG received the B.S. degree in computer science from Korea University, Seoul, South Korea, in 1994, the M.S. degree in computer science from the University of Colorado at Boulder, CO, USA, in 1996, and the Ph.D. degree in computer science from the University of Wisconsin–Madison, WI, USA, in 2003.

From 1996 to 1997, he was a Technical Staff Member with AT&T Labs Research, Florham Park, NJ, USA. From 1997 to 1998, he was a Technical Staff Member with Savera Systems, Inc., Murray Hill, NJ, USA. From 2000 to 2001, he was the CTO and a Co-Founder of WISEngine, Inc., Santa Clara, CA, USA, and WISEngine, Inc., Seoul, South Korea. From 2003 to 2006, he was an Assistant Professor with the Department of Computer Science, North Carolina State University, Raleigh, NC, USA. Since 2006, he has been a Professor with the Department of Computer Science, Korea University, Seoul, South Korea. He also serves as the Department Head of the Interdisciplinary Graduate Program in Bioinformatics, Korea University, where he is jointly appointed as a Professor with the Department of Medicine, School of Medicine.