Database, 2022, 1–9
DOI: https://doi.org/10.1093/database/baac074
Original article

Full-text chemical identification with improved generalizability and tagging consistency

Hyunjae Kim[1], Mujeen Sung[1], Wonjin Yoon[1], Sungjoon Park[2] and Jaewoo Kang[1,3,*]

[1]Department of Computer Science and Engineering, Korea University, Seoul, South Korea
[2]Department of Medicine, University of California, San Diego, CA, USA
[3]AIGEN Sciences, Seoul, South Korea

*Corresponding author: Tel: +82-2-3290-4840; Email: kangj@korea.ac.kr

Citation details: Kim, H., Sung, M., Yoon, W. *et al.* Full-text chemical identification with improved generalizability and tagging consistency. *Database* (2022) Vol. 2022: article ID baac074; DOI: https://doi.org/10.1093/database/baac074

Abstract

Chemical identification involves finding chemical entities in text (i.e. named entity recognition) and assigning unique identifiers to the entities (i.e. named entity normalization). While current models are developed and evaluated based on article titles and abstracts, their effectiveness has not been thoroughly verified in full text. In this paper, we identify two limitations of models in tagging full-text articles: (1) low generalizability to unseen mentions and (2) tagging inconsistency. We use simple training and post-processing methods to address the limitations such as transfer learning and mention-wise majority voting. We also present a hybrid model for the normalization task that utilizes the high recall of a neural model while maintaining the high precision of a dictionary model. In the BioCreative VII NLM-Chem track challenge, our best model achieves 86.72 and 78.31 F1 scores in named entity recognition and normalization, significantly outperforming the median (83.73 and 77.49 F1 scores) and taking first place in named entity recognition. In a post-challenge evaluation, we re-implement our model and obtain 84.70 F1 score in the normalization task, outperforming the best score in the challenge by 3.34 F1 score.

Database URL: https://github.com/dmis-lab/bc7-chem-id

Introduction

Chemical entities include drugs, compounds, chemical formulas, identifiers, etc. (1). Extracting chemical entities from a vast amount of literature is an essential step in various downstream tasks such as relation extraction (2, 3) and literature search (4). Chemical identification is a task to support this scenario, which consists of (i) named entity recognition (NER) that locates chemical entities in the provided text and (ii) named entity normalization (NEN or normalization) that links the entities to unique identifiers pre-defined in biomedical knowledge bases.

Several datasets such as BC5CDR (5) are proposed to facilitate research on chemical identification. Most consist of the titles and abstracts of PubMed articles with manually annotated chemical mentions and the corresponding identifiers. Recent studies put a lot of effort into achieving high performance on these abstract-level datasets in the NER (6–8) and NEN tasks (9, 10). However, there are few studies on tagging a *full-text* corpus consisting of the main body of a paper as well as the title and abstract. Since detailed descriptions of background, methodology and findings are mostly included in the main body, automatically annotating the full text can be more informative than annotating only the abstract ('title and abstract' and 'abstract' are interchangeable).

In this work, we conduct a systematic study of full-text chemical identification. We analyze two limitations of current models in tagging full-text PubMed articles. First, models' generalizability to unseen entity mentions is limited (11), especially when entities appear in the main body. Our pilot experiment shows that the performance of an NER model on unseen mentions in the main body is lower than those in the abstract. Second, models make inconsistent predictions for the same entities within the same article due to the tagging inconsistency problem (6), which is worse in the main body than in the abstract.

We suggest using two methods to address the limitations. To improve the generalization ability of the model to unseen mentions, we use transfer learning, where the model is pre-trained on source data and then fine-tuned on the target data. This exposes models to more diverse chemical entities and contexts, improving entity coverage and generalizability. For the source data, we use existing chemical NER datasets (1, 5) and a synthetically generated dataset by synonym replacement (12).

To mitigate tagging inconsistency, we use a rule-based post-processing method called mention-wise majority voting. (While majority voting generally refers to an ensemble method in the context of machine learning, in this paper, it

Received 31 March 2022; Revised 11 July 2022; Accepted 22 August 2022

refers to our method for tagging consistency (Section Majority voting)). The method aggregates all inconsistent predictions for the same phrase in an article and changes all minority predictions to the majority, based on the assumption that (i) the majority is more accurate than the minority and (ii) the same words or phrases within the same article refer to the same concepts (e.g. entities). Our experiments show the method effectively improves NER performance despite its simplicity.

Additionally, we present a hybrid approach to improve recall while maintaining precision in the normalization task. Dictionary models usually achieve high precision but low recall due to the limited coverage of their dictionaries, whereas neural network models achieve higher recall but lower accuracy. We attempt to leverage the strengths of both while compensating for the weaknesses of each model. We first perform dictionary lookup and then use a neural model to further predict entities that do not match the dictionary. The hybrid model significantly improves recall, resulting in strong normalization performance.

We experiment with our methods using the NLM-Chem dataset (13), which consists of 150 full-text articles with chemical entity annotations. Based on the experiments, we submit our best models to the BioCreative VII NLM-Chem track challenge (14) and obtain 86.72 and 78.31 F1 scores in NER and NEN, significantly outperforming the median (83.73 and 77.49 F1 scores) and ranked first in NER. We found that our NEN models were underestimated in the official challenge evaluation due to implementation errors—after fixing the errors, we achieve 84.70 F1 score in NEN, surpassing the best score in the challenge by 3.34 F1 score. In sum, we make the following contributions:

- We identify the limitations of existing models in terms of full-text chemical identification: (i) low generalizability to unseen mentions and (ii) tagging inconsistency.
- We address the limitations using simple transfer learning and mention-wise majority voting methods. For normalization, we present a hybrid model combining dictionary and neural models to achieve higher recall while maintaining accuracy.
- Our system significantly outperforms the median in the official evaluation of the BioCreative VII NLM-Chem track challenge and even achieves the best score in NER (86.72 F1 score). In the post-challenge evaluation, our hybrid normalization model obtains the best score (84.70 F1 score).

Background

This section describes the background to understand the task and our methodology. We deal with full-text chemical identification as two separate tasks: NER and NEN. Specifically, we train NER and NEN models independently and combine them at the inference time. The NER model takes a sentence as input and is optimized to predict each token in the sentence. The NEN model uses the predictions of the NER model as input (i.e. predicted entity mentions) and assigns them to corresponding identifiers pre-defined in biomedical knowledge bases.

Named entity recognition

Let $\mathcal{D} = \{\mathrm{D}_1, ..., \mathrm{D}_N\}$ be a dataset, where N is the number of documents in the dataset and each document D_n ($n \in [1, N]$) consists of sentences. While entities are represented by character-level start and end indexes in the sentence (In a strict NER evaluation, models should predict the exact character-level indexes during inference.), they are usually treated as token-level labels in practice since most entity boundaries are per-token. In other words, we split the given sentence into L tokens and feed it into an NER model E^{NER} to predict the label of each token as follows:

$$\hat{y}_1, ..., \hat{y}_L = E^{\mathrm{NER}}([x_1, ..., x_L]), \tag{1}$$

where x_l and $\hat{y}_l$ ($l \in [1, L]$) are the l-th token and the predicted label. Following the `BIO` format (15), each label is assigned either `B` (Beginning) or `I` (Inside) or `O` (Outside). Finally, contiguous tokens with the first token corresponding to `B` and the others corresponding to `I` are considered a predicted entity $\hat{e}$.

Normalization

Once the corpus is annotated by the NER model, an NEN model then links the predicted mentions to pre-defined biomedical identifiers. Each $\hat{e}$ is fed into the NEN model E^{NEN}, and the model produces the identifier $\hat{c}$ as follows:

$$\hat{c} = E^{\mathrm{NEN}}(\hat{e}, \mathcal{V}), \tag{2}$$

where $\mathcal{V}$ is a dictionary consisting of identifier–mention pairs. The model E^{NEN} searches the most similar entity in the dictionary to the given mention $\hat{e}$ and outputs the identifier of the entity as the final prediction $\hat{c}$. While a dictionary model performs string matching between the input mention $\hat{e}$ and candidate entities in the dictionary $\mathcal{V}$, neural network models convert them into dense representations and compute the vector similarity between them (9, 10).

Full-text dataset

Previous datasets for chemical NER and NEN have a limitation in that they consist only of the title and abstract of the paper (1, 5, 16). Recently, Islamaj *et al.* (13) propose NLM-Chem, the first large-scale dataset with manually annotated chemical entity mentions and identifiers, consisting of 150 full-text PubMed articles. The data are designed to be rich in chemical entities that are difficult for models trained on other previous chemical NER datasets to identify.

NLM-Chem challenge

BioCreative VII introduces a new challenge, 'NLM-Chem Track: Full-text Chemical Identification and Indexing in PubMed articles' (14) (https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-2/). The challenge presents two tasks, but we focus on the Chemical Identification task in this work and leave the Chemical Indexing task for future work. In addition to 150 annotated articles in the original NLM-Chem data, the challenge provides additional 54 full-text articles manually annotated in a similar process when constructing NLM-Chem. Table 1 shows the statistics of the NLM-Chem and challenge data. We use the test set of NLM-Chem (50 articles) as the validation set. The additional 54 articles are used for the official evaluation.

Table 1. Statistics of the BioCreative VII NLM-Chem track challenge data. The test set is only used in the official challenge evaluation. **# Articles**: the number of articles. **# Sentences**: the number of sentences. **# Mentions**: the number of annotated entity mentions. **# CUIs**: the number of concept unique identifiers.

Type	# Articles	# Sentences	# Mentions	# CUIs
Train	100	23 560	26 566	29 089
Valid	50	11 183	11 772	12 211
Test	54	17 703	22 942	25 316

Table 2. Performance of Bio-LM-large (17) on the abstract and main body in the NLM-Chem validation set. **Prec.**, **Rec.** and **F1**: entity-level precision, recall and F1 score, respectively. Δ: performance difference. Note that we report only recall on `Mem`, `Syn` and `Con` since it is impossible to classify false positives into the splits, and precision for each split cannot be calculated (11).

Type	Prec.	Rec.	F1	`Mem`	`Syn`	`Con`
Full	86.5	88.7	87.6	92.6	77.8	86.7
Abstract	87.6	89.2	88.4	93.3	80.6	87.7
Main body	86.4	88.6	87.5	92.5	77.5	86.6
Δ	−1.2	−0.6	−0.9	−0.8	−3.1	−1.1

Preliminary study

We determine whether a current model is sufficient or limited in its ability to tag full-text articles. We focus on NER in this analysis because a strong NER model is a prerequisite for high normalization performance. We use the Bio-LM-large model (17) with a linear output layer as the NER model. We train the model on the full NLM-Chem training set and measure the performance on the abstract and the main body of the validation set separately. Table 2 shows that the performance on the main body is lower by 0.9 F1 score than that on the abstract, indicating that tagging full text is relatively challenging compared to tagging only abstract. We systematically analyze what factors make this difficulty in the following sections.

Generalization to unseen mentions

In the biomedical domain, it is of paramount importance to generalize unseen mentions that the model did not experience during training because synonyms and newly discovered biomedical concepts constantly emerge in the literature (11). Since the main body contains more diverse entities and complex context than the abstract, the generalizability issue (11, 18, 19) can be critical in the main body. Follow Kim *et al.* (11), we partition all mentions e in the NLM-Chem validation set into three splits as follows:

$$\begin{aligned} \texttt{Mem} &:= \{e : e \in \mathbb{E}_{\text{train}}, c \in \mathbb{C}_{\text{train}}\}, \\ \texttt{Syn} &:= \{e : e \notin \mathbb{E}_{\text{train}}, c \in \mathbb{C}_{\text{train}}\}, \\ \texttt{Con} &:= \{e : e \notin \mathbb{E}_{\text{train}}, c \notin \mathbb{C}_{\text{train}}\}, \end{aligned} \tag{3}$$

where $\mathbb{E}_{\text{train}}$ is the set of all entity mentions in the training set and $\mathbb{C}_{\text{train}}$ is the set of all Concept Unique Identifiers (CUIs) in the training set. Specifically, `Mem` consists of *memorizable* mentions that were seen during training. `Syn` consists of synonyms, where their surface forms are new/unseen but their CUIs are not. `Con` consists of new entities whose surface forms and CUIs are both unseen. Each data split corresponds to one of the recognition abilities that reliable NER models should possess: (i) memorization, (ii) synonym generalization and (iii) concept generalization. We focus on the last two abilities that are related to identifying unseen mentions.

Table 2 shows the performance on `Syn` and `Con` in the main body is consistently lower than that in the abstract, indicating that the model is limited in terms of generalizability to unseen mentions. Interestingly, the performance difference is very noticeable on `Syn` (3.1 F1 score). This may be because entities are often represented in different ways throughout the paper, especially in the main body.

Table 3. Label inconsistency and tagging inconsistency in the abstract and main body of the NLM-Chem validation set.

Type	Label inconsistency	Tagging inconsistency
Abstract	0.02	6.4
Main body	0.04	9.7

Tagging inconsistency

Since identical words or phrases within the same article often refer to the same concepts or entities, models should be consistent in predicting the same text spans. Unfortunately, current sentence-level models classify the same spans into different ones, which leads to the tagging inconsistency problem (6, 20). In this section, we measure how much tagging inconsistency occurs in the abstract and main body, respectively. Let W_n be all unique phrases (i.e. n-grams) within the n-th article, $g_n(p)$ be the total number of a phrase p within the n-th article and $h_n(p)$ be the total number of positive predictions for the phrase p within the n-th article. We consider model prediction for the phrase p to be *inconsistent* if a function ϕ_n returns 1, which is defined as follows:

$$\phi_n(p) = \begin{cases} 1 & \text{if} g_n(p) \neq h_n(p), \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Finally, we calculate tagging inconsistency in the dataset $\mathcal{D}$ as the average of inconsistent predictions for all unique phrases in the corpus as follows:

$$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{|W_n|}\sum_{p\in\mathcal{W}_n}\phi_n(p)\right). \tag{5}$$

Similarly, we define label inconsistency by assuming that $g_n(p)$ returns the total number of gold annotations for the phrase p within the n-th article.

Table 3 shows that label inconsistency is insignificant, supporting our assumption that phrases within the same surface forms are likely to be the same entity (or not entity). On the other hand, tagging inconsistency occurs frequently, and it is more pronounced in the main body than in the abstract, indicating that it needs to be addressed to obtain satisfactory performance in full-text chemical identification.

Method

From our analysis, we identified low generalizability to unseen mentions and tagging inconsistency as obstacles to tagging full-text articles. We use transfer learning and mention-wise

majority voting methods to address them. In normalization, we use a hybrid model to improve recall using a neural model while maintaining high precision of a dictionary model. See the paper (21) for a simpler system description.

Transfer learning

We pre-train a model on source data and then fine-tune it on the target data (i.e. NLM-Chem). Since pre-training with additional datasets exposes models to more diverse chemical entities and contexts, this can improve the generalizability. We use two popular chemical NER datasets CHEMDNER (1) and BC5CDR (5) as the source data. At the fine-tuning stage, we randomly initialize the output layer and only reuse the rest of the model parameters.

Data augmentation

Dai *et al.* (12) augment training data by replacing entity mentions with their synonyms. This allows the model to learn different representations of entities, which can help improve generalizability to morphological variations. Following the work, we generate the new synthetic data NLM-Chem(syn) by replacing the mentions in NLM-Chem with their synonyms, which are sampled from the Comparative Toxicogenomics Database. We use NLM-Chem(syn) as additional source data for transfer learning.

Majority voting

To alleviate tagging inconsistency, we use a majority voting method that aggregates model predictions in full text (Figure 1). First, we collect all inconsistent predictions in the same article, where the inconsistency is defined by Equation (4). We then compute the majority for model predictions and change all the minority predictions to the majority. Luo *et al.* (6) used a similar post-processing method to ours in their work, but the method only changes negative predictions to positives, which might be detrimental to precision. On the other hand, we also consider the direction from positives to negatives, which reduces false positives and improves precision. Since majority voting can be noisy if the target phrase does not frequently appear in the article, we apply the method only when the number of the phrase is greater than a threshold τ.

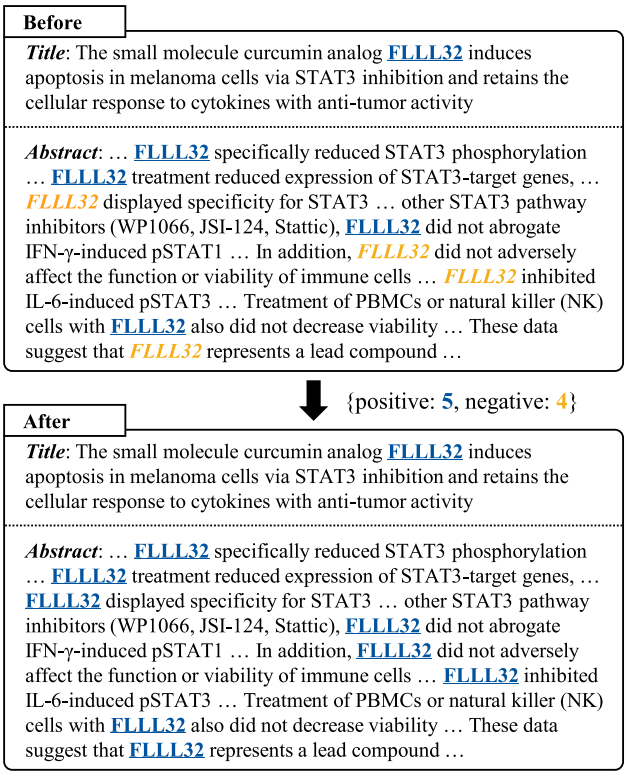


Figure 1. The tagging inconsistency problem and our majority voting method. We underline positive predictions and italicize negative predictions for the entity "FLLL32." Our method improves tagging consistency by changing the minority to the majority.

Hybrid model

Hybrid model consists of a dictionary model and a neural network model. The dictionary model first performs normalization based on string matching between the target mentions and the dictionary after applying several pre-processing rules such as lowercasing and removing punctuation. For mentions that are not normalized by the dictionary model, the neural model further performs the process. The neural model retrieves top-k similar entities to the given mention from the biomedical dictionary $\mathcal{V}$. To deal with the `CUI-LESS` class, which means that a given entity does not match any CUIs in the dictionary, we add a special embedding and classify mentions into the class if the embedding is included in top-k results.

Experiments

Evaluation

We evaluate our models in the BioCreative VII NLM-Chem track challenge. For NER, entity-level precision (Prec.), recall (Rec.) and F1 score (F1) are used. For normalization, unique CUI predictions and labels for each article are compared first and then precision, recall and macro-averaged F1 score are calculated based on the article-level true positives, false positives and false negatives (13, 14).

Implementation details

We select Bio-LM-large (17) as our NER model for its superiority compared to others (See Table 6). For NER, we search best checkpoints and hyperparameters of NER models, based on F1 score on the validation set at every training epoch. We further trained NER models on the validation set by 20 epochs for the final submission. The max length of input sequence is set to 512. We use the batch size of 24 and the learning rate of 1e−5. In synonym augmentation, NLM-Chem(syn) is 3x larger than the original data. For the majority voting method, we only use entities that are longer than 2 and appear more than 40 times in the same article (i.e. $\tau = 40$). For normalization, we use the 1 April 2021 version of the Comparative Toxicogenomics Database as our chemical dictionary. We further expand the dictionary using mentions annotated in NLM-Chem. For the neural model, we train BioSyn (9) with the SapBERT encoder (10) on NLM-Chem using the same hyperparameters as suggested by the authors. We search for the best neural NEN model checkpoints using F1 score in the validation set.

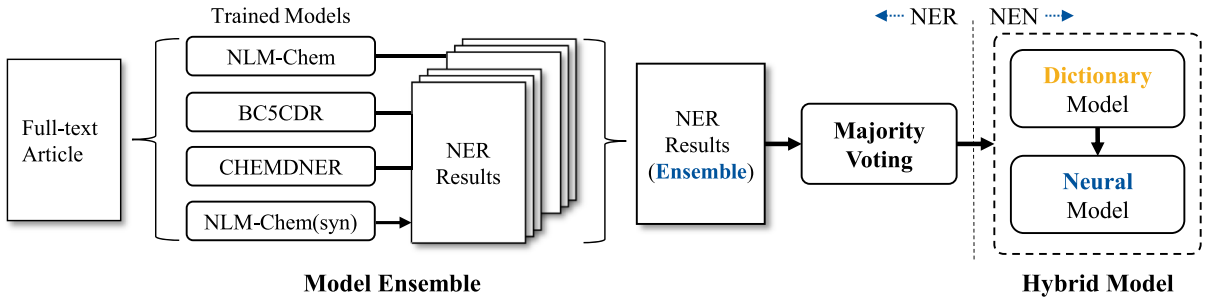


Figure 2. Overview of our final system for the BioCreative VII NLM-Chem track challenge.

Table 4. Top ten models and the scores in the official challenge evaluation.

	NER				NEN		
Team (Run)	Prec.	Rec.	F1	Team (Run)	Prec.	Rec.	F1
139 (3)[a]	87.59	85.87	86.72	110 (4)	86.21	77.02	81.36
139 (1)[a]	87.47	85.23	86.33	128 (2)	77.92	84.34	81.01
139 (2)[a]	87.75	84.47	86.07	110 (1)	85.82	76.41	80.84
128 (1)	85.44	86.58	86.00	128 (1)	78.33	83.39	80.78
143 (1)	85.35	86.08	85.71	121 (1)	78.74	82.81	80.72
128 (4)	84.57	86.17	85.36	121 (3)	78.76	82.72	80.69
128 (2)	86.43	84.03	85.21	110 (2)	82.21	78.98	80.56
121 (2)	84.61	85.83	85.21	128 (4)	77.55	83.18	80.27
121 (1)	86.16	84.15	85.15	121 (2)	77.48	83.15	80.21
121 (3)	85.80	84.09	84.94	121 (5)	78.21	82.26	80.19
Median	84.76	81.36	83.73	Median	71.20	77.60	77.49

[a]our models. The best scores in the table are underlined. See the challenge overview paper (14) for a full list of results.

Table 5. Performance of our NEN models on the test set. The best scores are underlined. Note that 'post-challenge' models were unofficially evaluated after the challenge was over, but on the same test set.

	Official			Post-challenge		
Run	Prec.	Rec.	F1	Prec.	Rec.	F1
1	72.12	84.71	77.91	85.39	83.27	84.32
2	72.56	85.05	78.31	85.80	83.64	84.70
3	71.20	84.99	77.49	85.42	83.49	84.44

Table 6. Differences between biomedical pre-trained language models. **Vocab.** and **Corpus**: the vocabulary and corpus type used in pre-training, respectively. The Bio-LM-large is the best in our experiment.

Model	Vocab.	Corpus	Size	F1
BioBERT (8)	Wiki+Books	Abstract	Base	84.8
PubMedBERT (23)	PubMed	Abstract	Base	87.2
PubMedBERT(full) (23)	PubMed	Full text	Base	87.4
Bio-LM-base (17)	PubMed	Full text	Base	87.0
Bio-LM-large (17)	PubMed	Full text	Large	87.6

Sub-token entities

The NLM-Chem data have many sub-token entities that are sub-strings of a token rather than the whole string. For example, the token 'Gly104Cys' has two sub-token entities 'Gly' and 'Cys'. In the official evaluation of the challenge, models should predict sub-token entities, not the whole tokens. We found that sub-token entities mostly appear within mutation names, and about 90% of sub-token entities can be processed with simple regular expressions. Based on this, we perform post-processing on sub-token entities, which greatly improves performance in the official evaluation.

Final submission

Ensemble methods theoretically reduce expected generalization errors by reducing the variance. To boost the performance in the challenge evaluation, we build majority voting ensemble models that combines predictions from different models trained on different datasets (See Table 7). For NEN, we use a single hybrid model. Figure 2 illustrates our final system for the challenge.

Results

Table 4 shows top ten submission results in NER and NEN, respectively. In NER, our top three systems significantly outperformed the median and other 88 submission results from 17 teams and ranked first, second and third, respectively. On the other hand, our systems did not make it into the top ten in NEN despite high performance in NER. After the challenge, we found some errors in our implementation of the normalization model, which significantly degraded the performance. Thus, we re-evaluate the NEN performance on the test set published after the challenge. As shown in Table 5, we achieved 84.70 F1 score after fixing the errors, which is higher than the best score in the challenge by 3.34 F1 score. From these results, we can conclude that the hybrid model is promising for future practical applications. Consistent with

Table 7. Ablation study for NER on the validation set. **Standard**: a single Bio-LM-large model. Performance differences between the standard and other models are shown in parentheses.

Model	Prec.	Rec.	F1
Single model			
Standard	86.5	88.7	87.6
+ BC5CDR	86.0 (−0.5)	89.4 (+0.7)	87.7 (+0.1)
+ CHEMDNER	86.5	89.5 (+0.8)	88.0 (+0.4)
+ NLM-Chem(syn)	86.7 (+0.2)	89.3 (+0.6)	88.0 (+0.4)
Ensemble model			
Fine-tune only	86.8 (+0.3)	89.2 (+0.5)	87.9 (+0.3)
Transfer only	87.2 (+0.7)	89.9 (+1.2)	88.5 (+0.9)
Both	87.2 (+0.7)	89.6 (+0.9)	88.4 (+0.8)
Ensemble model (with majority voting)			
Fine-tune only	87.3 (+0.8)	89.6 (+0.9)	88.4 (+0.8)
Transfer only	87.6 (+1.1)	90.1 (+1.5)	88.8 (+1.2)
Both	88.0 (+1.5)	89.8 (+1.1)	88.9 (+1.3)

our results, a concurrent work shows the hybrid approach improves the performance (22).

Analysis

Language model selection

We experiment with several pre-trained language models common in the biomedical domain to select the best sentence encoder in NER: BioBERT (8), PubMedBERT (23) and Bio-LM (17). As shown in Table 6, Bio-LM-large outperforms the other models. Although BioBERT usually performs well on many tasks and achieves similar performance with PubMedBERT and Bio-LM, it performed much worse on NLM-Chem. Differences in vocabulary may have a significant impact on chemical NER performance. Also, PubMedBERT-full performed better than PubMedBERT, indicating that pre-training on full-text articles may be effective for chemical NER at the full-text level. Bio-LM-large performed better than Bio-LM-base, showing that model size can affect performance.

Ablation study

Effect of transfer learning

Table 7 shows that transfer learning improved models' performance, especially recall. Although the synonym replacement method does not require the cost of manual annotations, it can be more effective than using existing hand-labeled datasets.

Effect of model ensemble

Table 7 shows that ensemble models outperform single models. Besides, we analyzed how the effect of ensembling varies according to the combinations of single models. We designed three ensemble models, 'Fine-tune only', 'Transfer only' and 'Both', which indicate the combination of models trained only with NLM-Chem, the combination of only transferred models and the combination of both types of models, respectively. As a result, we found that ensembling models trained on different sources can be effective.

Effect of majority voting

Table 7 shows that majority voting is simple but consistently improves the performance of ensemble models. Also, we see how the performance of the single Bio-LM-large model changes when changing the threshold of occurrence τ. Figure 3 shows the performance peaks at $\tau = 40$ and decreases, indicating that finding the optimal τ is important.

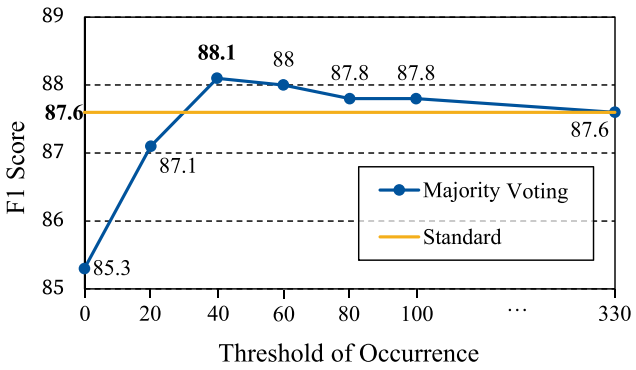


Figure 3. Performance of majority voting with different thresholds of occurrence τ on the validation set. **Standard**: a single Bio-LM-large model.

Table 8. Ablation study for NEN on the validation set. Gold NER annotations are used as input in this experiment.

Model	Prec.	Rec.	F1
Dictionary	94.4	83.8	88.8
Neural	83.9	88.4	86.1
Hybrid	91.6	87.2	89.3

Effect of hybrid model

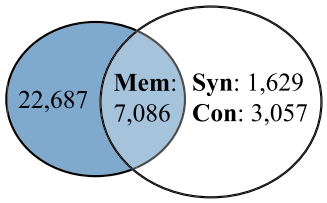
As shown in Table 8, the dictionary model works very well in normalization if we have a high-quality dictionary. However, the method has low recall due to the limited coverage of the dictionary. Our hybrid model significantly improved recall, resulting in a higher F1 score.

In-depth analysis

We pointed out the two limitations of existing models that hinder tagging full-text articles. We confirmed that transfer learning and majority voting improve the overall performance in Table 7 and Figure 3, but further analysis is needed to figure out the effect of the methods in depth.

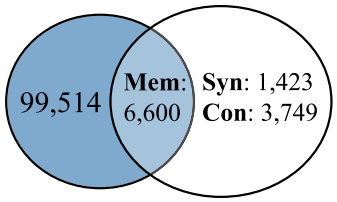
Q1. Does transfer learning actually improve generalization ability to unseen entities?

Table 7 shows that transfer learning improves model performance, especially recall. Furthermore, we see whether this performance improvement was achieved by simply increasing entity coverage during training or by improving true generalizability to unseen entities. A way of measuring generalizability is to split the dataset as in Equation (3) and compare performance of a model with/without applying transfer learning on `Syn` and `Con`, where the set of mentions $\mathbb{E}_{\text{train}}$ and the set of CUIs $\mathbb{C}_{\text{train}}$ include the NLM-Chem training set and a source dataset used in transfer learning scenarios. Figure 4 shows the number of mentions of `Mem`, `Syn` and `Con` of the validation set and the model performance, when source data are BC5CDR and NLM-Chem(syn). Regardless of source datasets, performance on `Syn` is improved, indicating that transfer learning can improve generalizability to synonyms. From these results,



Model	Syn	Con
Standard	72.9	86.9
Transfer	74.9	86.6

(a) NLM-Chem training + **BC5CDR**



Model	Syn	Con
Standard	76.0	86.7
Transfer	78.2	88.5

(b) NLM-Chem training + **NLM-Chem(syn)**

Figure 4. The number of mentions of `Mem`, `Syn` and `Con` and model performance on each split, when using BC5CDR and NLM-Chem(syn) as source data in transfer learning. The blue circles indicate the mentions in the validation set, and the others are the mentions in training sets (i.e. $\mathbb{E}_{\text{train}}$). **Standard** and **Transfer**: Bio-LM-large without/with applying transfer learning, respectively.

Table 9. Detailed analysis on majority voting using the validation set. **Standard** and **Majority**: Bio-LM-large without/with the majority voting method, respectively. Δ: performance difference.

Model	Prec.	Rec.	F1	`Mem`	`Syn`	`Con`
Abstract						
Standard	87.6	89.2	88.4	93.3	80.6	87.7
Majority	87.7	89.5	88.6	94.0	80.6	87.7
Δ	+0.1	+0.3	+0.2	+0.7	0.0	0.0
Main body						
Standard	86.4	88.6	87.5	92.5	77.5	86.6
Majority	86.9	89.1	88.0	93.5	77.5	86.4
Δ	+0.5	+0.5	+0.5	+1.0	0.0	−0.2

we confirm that the performance improvement is not simply due to increased entity coverage.

Q2. When is majority voting particularly effective?

The method is particularly effective when there are many mentions of the same entity in one article, and there is severe tagging inconsistency. For instance, the article with PMID 2 902 420 has 137 mentions of the entity 'FLLL32', and models predicted about 70% of the mentions as entities and the rest as not. In this case, the method corrected about 30% errors, which significantly improves performance. Also, Table 9 shows that majority voting is particularly effective in the main body, where the problem is much more severe than in the abstract.

Q3. Can majority voting improve generalization ability to unseen entities?

Since recognizing unseen mentions is more difficult than recognizing memorizable mentions, tagging inconsistency will occur more for unseen mentions. It will be interesting to see if majority voting can effectively mitigate tagging inconsistency for unseen mentions. As shown in Table 9, while the method significantly improved performance on `Mem`, it was not effective on `Syn` and `Con`. Since recall on unseen entities (i.e. entities in `Syn` and `Con`) is insufficient, the majority may be false negatives, and thus the method may not be as effective.

Error analysis (NER)

We analyze 100 error cases of our NER model using the test set.

Reoccurrence of the same errors

We found that the model repeated the same errors within the same article. For instance, 5% of the whole error cases occurred since the model failed to extract the entity mention 'pKAL' (the Korean plant *Artemisia annua* L.). Majority voting can be effective against these repeated errors if the majority predictions are correct and are greater than the minority repeated errors. However, since the model predicted all occurrences of 'pKAL' as negative, majority voting could not correct the repeated errors, which is a limitation of the method.

Abbreviations

Forty percentage of errors are due to abbreviations. It is challenging to deal with abbreviations as their names are ambiguous and have less information. The full names of abbreviations are often defined in the front parts of the paper such as the abstract or introduction; thus in further work, we can utilize these definitions to help identify abbreviations.

Other insights

The model sometimes made unexpected predictions including special characters, and these false positives accounted for 6% of all errors. For instance, the model predicted 'APO(' as an entity given the context 'The stability of APO(ANTR) nanodrugs was tested by storing them at 4 C for 30 days.', while the model correctly extracted 'APO' in most other contexts. Also, the model sometimes did not extract the entire entity 'Mg-PCL', rather it extracted 'Mg-' and 'PCL' separately. Many chemical entities are composed of complex combinations of alphabets and special characters, making it difficult for the model to distinguish exact boundaries.

The model appears to be sensitive to even small changes in entity forms. We found that the model successfully extracted the entity '11Cha1' but failed to extract other entities with similar forms, such as '11Cha2', '11Cha3', '11Cha10' and '11Cha11', even when they appeared in the same sentence 'Less hindered groups on ring A such as hydroxyl, methoxyl, and/or methoxymethoxyl (MOM) (e.g. 11Cha1, 11Cha2, and 11Cha3) increased the activity.' It seems that the model lacks the ability to understand sentence structure or context pattern. Such ability should be improved by developing better language models or incorporating syntactic information into the model.

Error analysis (NEN)

We manually analyze 300 error cases from the test set. The most common errors (71.3%) occurred due to limited coverage of the dictionary, and so the model incorrectly predicted entity as `CUI-LESS`. The second type of error, accounting for 14.3%, occurred when the model was misled by entities

with similar forms to a target entity. For instance, the target entity 'polyamide' and a synonym 'nylon' are not similar even though they are the same entity, so the model chose a more similar entity 'polymer'. Finally, some entity mentions with the same surface form can have different CUIs depending on the context, producing 14.3% errors. For instance, while 'DHA' in a test article refers to 'Docosahexaenoic Acid', in the dictionary, 'DHA' refers to 'Dihydroartemisinin', making a false prediction.

All types of errors we mentioned above can be addressed by using contextual information. Our model relies on surface forms of mentions to perform the task, which limits the NEN performance. Adopting recent models using contextual information (24, 25) to full-text chemical normalization would be interesting, and we leave this for future research.

Conclusion

In this paper, we studied chemical identification in full-text articles. We found that low generalizability to unseen entities and tagging inconsistency are problems and should be considered to effectively perform the task. We showed that the problems are addressable using transfer learning and mention-wise majority voting. Also, we showed that combining dictionary and neural models is effective for normalization. We demonstrated the effectiveness of all methods using the NLM-Chem dataset through ablation studies and achieved strong performance in the BioCreative VII NLM-Chem track challenge.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

We thank Rezarta Islamaj, Robert Leaman and Zhiyong Lu for organizing the NLM-Chem track and helping out during the challenge. Also, we thank the annotators of the NLM-Chem dataset and authors for their efforts and contributions.

Funding

Ministry of Science and Information and Communications Technology (ICT), Korea, under the ICT Creative Consilience program (Institute for Information and communications Technology Planning and Evaluation, IITP-2022-2020-0-01 819) supervised by the IITP; Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (HR20C0021); National Research Foundation of Korea (NRF-2014M3C9A3063541, NRF-2020R1A2C3010638), under project BK21 FOUR.

References

1. Krallinger,M., Rabal,O., Leitner,F. *et al*. (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminf.*, **7**, 1–17. 10.1186/1758-2946-7-S1-S1.
2. Zhang,Y., Zheng,W., Lin,H. *et al*. (2018) Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. *Bioinformatics*, **34**, 828–835. 10.1093/bioinformatics/btx659.
3. Lim,S. and Kang,J. (2018) Chemical–gene relation extraction using recursive neural network. *Database*, **2018**, 10.1093/database/bay060.
4. Lee,S., Kim,D., Lee,K. *et al*. (2016) BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS One*, **11**, e0164680. 10.1371/journal.pone.0164680.
5. Jiao,L., Sun,Y., Johnson,R.J. *et al*. (2016) BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, **2016**. 10.1093/database/baw068.
6. Luo,L., Yang,Z., Yang,P. *et al*. (2018) An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, **34**, 1381–1388. 10.1093/bioinformatics/btx761.
7. Yoon,W., So,C.H., Lee,J. *et al*. (2019) Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinf.*, **249**. 10.1186/s12859-019-2813-6.
8. Lee,J., Yoon,W., Kim,S. *et al*. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240. 10.1093/bioinformatics/btz682.
9. Sung,M., Jeon,H., Lee,J. *et al*. (2020) Biomedical entity representations with synonym marginalization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 3641–3650.
10. Liu,F., Shareghi,E., Meng,Z. *et al*. (2021) Self-alignment pretraining for biomedical entity representations. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
11. Kim,H. and Kang,J. (2022) How do your biomedical named entity recognition models generalize to novel entities? *IEEE Access*. 10.1109/ACCESS.2022.3157854.
12. Dai,X. and Adel,H. (2020) An analysis of simple data augmentation for named entity recognition. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain.
13. Islamaj,R., Leaman,R., Kim,S. *et al*. (2021) NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci. Data*, **8**, 1–12. 10.1038/s41597-021-00875-1.
14. Leaman,R., Islamaj,R. and Zhiyong,L. (2021) Overview of the NLM-Chem BioCreative VII track: Full-text chemical identification and indexing in PubMed articles. In: *Proceedings of the seventh BioCreative challenge evaluation workshop*.
15. Ramshaw,L.A. and Marcus,M.P. (1999) Text chunking using transformation-based learning. In: *Natural Language Processing Using Very Large corpora*. pp. 157–176.
16. Mohan,S. and Donghui,L. (2019) Medmentions: a large biomedical corpus annotated with umls concepts. *Automated Knowledge Base Construction (AKBC)*.
17. Lewis,P., Ott,M., Jingfei,D. *et al*. (2020) Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. pp. 146–157.
18. Augenstein,I., Derczynski,L. and Bontcheva,K. (2017) Generalisation in named entity recognition: A quantitative analysis. *Comput. Speech Lang.*, **44**, 61–83. 10.1016/j.csl.2017.01.012.
19. Lin,H., Yaojie,L., Tang,J. *et al*. (2020) A rigorous study on named entity recognition: Can fine-tuning pretrained model lead to the promised land? In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 7291–7300.
20. Gui,T., Jiacheng,Y., Zhang,Q. *et al*. (2021) Leveraging document-level label consistency for named entity recognition. In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. Yokohama, pp. 3976–3982.

21. Kim,H., Sung,M., Yoon,W. *et al*. (2021) Improving tagging consistency and entity coverage for chemical identification in full-text articles. In: *Proceedings of the seventh BioCreative challenge evaluation workshop*.
22. Sung,M., Jeong,M., Choi,Y. *et al*. (2022) BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, btac598. 10.1093/bioinformatics/btac598.
23. Yu,G., Tinn,R., Cheng,H. *et al*. (2020) Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, **3**, Article 2. 10.1145/3458754.
24. Angell,R., Monath,N., Mohan,S. *et al*. (2021) Clustering-based inference for biomedical entity linking. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 2598–2608.
25. Zhang,S., Cheng,H., Vashishth,S. *et al*. (2021) Knowledge-rich self-supervised entity linking. preprint, arXiv:2112.07887.