Distantly Supervised Biomedical Named Entity Recognition with Dictionary Expansion

Xuan Wang[1*], Yu Zhang[1*], Qi Li[2], Xiang Ren[3], Jingbo Shang[1], Jiawei Han[1]
[1]Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA
[2]Department of Computer Science, Iowa State University, IA, USA
[3]Department of Computer Science, University of Southern California, CA, USA
{xwang174, yuz9, shang7, hanj}@illinois.edu, qli@iastate.edu, xiangren@usc.edu

***Abstract*—State-of-the-art biomedical named entity recognition (BioNER) systems apply supervised machine learning models (i.e., relying on human effort for training data annotation) which are not easy to be generalized to new entity types and datasets. We propose a distantly supervised approach, AUTOBIONER, that automatically recognizes biomedical entities from massive corpora with user-input dictionaries. AUTOBIONER does not need any human annotated data. It relies on incomplete entity dictionaries to provide seeds for each entity type and performs a novel entity set expansion step for corpus-level new entity recognition and dictionary completion. The expanded dictionaries are used as distant supervision to train a neural model for BioNER. Experimental results show that AUTOBIONER achieves the best performance among the methods that only use dictionaries with no additional human effort on BioNER benchmark datasets. It is also demonstrated that the dictionary expansion step plays an important role in the great performances.**

***Index Terms*—biomedical named entity recognition, distantly supervised learning, entity expansion**

I. INTRODUCTION

Biomedical named entity recognition (BioNER) is a fundamental step in the biomedical information extraction pipeline. It is a task that identifies text spans as candidate entities and classifies them into a set of semantic classes, such as genes, chemicals and diseases. BioNER facilitates many downstream tasks such as relation extraction [4], [17] and knowledge base completion [12], [27], [28], [33].

Although BioNER has been extensively studied in the community, most previous approaches apply fully supervised machine learning models [11], [14], [18]. This reliance on human effort for training data annotation leads to highly specialized systems that cannot be directly used on new entity types. Even for the same entity type, it has been proved that models trained on one dataset may not perform well on the other [10].

To alleviate human effort and make the system more generalizable, distant supervision has been applied to automatically generate labeled training data with existing knowledge bases (or dictionaries) [2], [9], [21], [24], [26]. There exists a rich resource of open knowledge bases for biomedical entities (e.g,. UMLS, MeSH and CTD), which is a unique advantage for developing distantly supervised BioNER models. Some BioNER systems, such as MetaMap [2] and QuickUMLS [26], use a shallow parser to generate candidate phrases and a simple dictionary-matching approach for entity recognition based on the UMLS database. A recent work, AutoNER [24], uses a neural model that leverages distant supervision from entity dictionaries. However, these existing studies can only use limited information from the user-input dictionaries, especially when the dictionaries are incomplete in real word applications.

In this paper, we propose a distantly supervised framework for automated recognition of biomedical named entities. Our AUTOBIONER framework does not need *any* human annotated data and relies on incomplete entity dictionaries. AUTOBIONER first exploits statistical signals from massive corpora for candidate entity generation and user-input dictionaries for training example annotation. Since the dictionaries are assumed to be incomplete, AUTOBIONER performs a novel automatic entity set expansion for corpus-level new entity recognition and dictionary completion. It treats matched entities as positive examples to infer the type of unmatched candidates using context information. The expanded dictionaries are then used as distant supervision to train a neural model for BioNER. Our approach recognizes not only common biomedical entities (e.g., chemicals and diseases) but also other entities specific to users' interest (e.g., pathways and biological processes) with only user-provided dictionaries.

We conduct experiments on two benchmark BioNER datasets, BC5CDR [16] and NCBI-Disease [7], and AUTOBIONER achieves the best performance among the methods that only use dictionaries with no additional human effort. On BC5CDR, our F1 score is already very close to that of the *fully* supervised benchmark [11]. Moreover, we prove that dictionary expansion plays an important role in our framework both quantitatively and qualitatively. We also conduct case studies on a large corpus of PubMed abstracts to show that AUTOBIONER performs well on new entity types (e.g., pathways and biological processes) that do not have any existing training data.

We summarize our major contributions as follows.

- A novel distantly supervised framework, AUTOBIONER, is proposed, to automatically recognize biomedical entities from massive corpora with distant supervision from user-input dictionaries.

*The first two authors contributed equally to this work and should be considered as joint first authors.

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

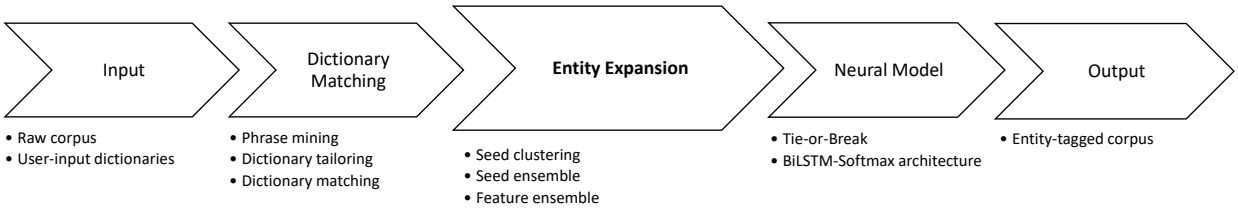

Fig. 1. The overall pipeline of AUTOBIONER. The entity expansion module is the unique part of AUTOBIONER.

- A novel entity set expansion method is developed that integrates automatic phrase mining with corpus-level new entity recognition and dictionary completion.
- Experiments on two BioNER benchmark datasets demonstrate that AUTOBIONER achieves the best performance using only dictionaries with no additional human effort.

II. RELATED WORK

BioNER has been popularly studied in the field of biomedical text mining. Most previous studies adopt a fully-supervised approach which leverages human annotated documents to train a sequence labeling model (e.g., conditional random fields (CRF)). Earlier BioNER systems require handcrafted features (e.g. capitalization, prefix and suffix) to be specifically designed for each entity type [1], [14], [18], [32], [35], [36]. This feature generation process takes the majority of time and cost in developing a BioNER system [15]. Recent NER studies consider neural network models (e.g., convolutional neural networks (CNN) and bidirectional long short-term memory networks (BiLSTM)) to automatically generate quality features [3], [11], [13], [19], [31], [34]. Semi-supervised methods have also been explored to further improve the accuracy in tasks like gene name recognition [29]. Unlike these existing approaches, our study focuses on the distantly supervised setting without any expert-curated training data.

Distant supervision, as a more recent trend, aims to reduce expensive human labor by utilizing entity information in dictionaries or knowledge bases. There are several studies on distantly supervised BioNER. Cotype [21] links mentions to knowledge base and then use linked entities to infer the type of unlinked ones using label propagation on heterogeneous graphs. SwellShark [9] leverages a generative model to unify and model noise across different supervision sources for named entity typing. However, it leaves the named entity span detection to a heuristic combination of dictionary matching and part-of-speech tag-based regular expressions, which requires extensive expert effort to cover many special cases. AutoNER [24] uses a neural model that leverage distant supervision from entity dictionaries. However, it can only use limited information from the user-input dictionaries, especially when the dictionaries are incomplete in real word applications. Our study mainly focuses on tackling this dictionary incompleteness problem using entity expansion.

III. FRAMEWORK

Our model AUTOBIONER is composed of three major components (Figure 1). It takes the raw corpus and the biomedical entity dictionaries as input. To make our system practical in real world applications, we do not assume the input dictionaries are complete. After dictionary matching, AUTOBIONER performs a novel entity set expansion for dictionary completion. The expanded dictionaries are used as distant supervision for a neural model for BioNER.

A. Phrase Mining and Dictionary Matching

Phrase Mining. We first obtain a large pool of candidate entities from the raw corpus for dictionary matching and completion. AUTOBIONER utilizes the state-of-the-art distantly supervised phrase mining method, AutoPhrase [23], and takes entries in the user-provided dictionaries as positive examples. After quality phrase generation, we type the entities with their corresponding types if they can be matched in the dictionary. Since the dictionaries are incomplete, unmatched candidate phrases may still be named entities. In our next step, we leverage the context similarity between matched phrases and unmatched ones to find more entities.

Dictionary Tailoring. Blindly using the full dictionary to match the phrases may introduce false-positive labels, as there exist many entities beyond the scope of the given corpus but their aliases can be matched. For example, when the dictionary has the term *Alzheimer's Disease* and its alias *AD*, many *AD*'s will be wrongly typed as DISEASE due to ambiguity. To tackle this problem, we perform a dictionary tailoring step [24] before dictionary matching. We tailor the original dictionary to a corpus-related subset by excluding entities whose canonical names (*Alzheimer's Disease* in the example) never appear in the given corpus. The intuition behind is that people will likely mention the canonical name of the entity at least once to avoid ambiguity. In our experiments, this is true for 88.12% and 95.07% of entity mentions on the BC5CDR and NCBI datasets, respectively. We expect the NER model trained on the tailored dictionary will have a higher precision compared to that trained on the original dictionary.

B. Entity Expansion

Figure 2 demonstrates our entity expansion step, where we automatically add the candidate phrases (e.g., *cisplatin*) to the correct entity type set (e.g., CHEMICAL type) for dictionary completion. A simple way to do this is to compare the semantic similarity between the candidate phrase and each entity type (described by the set of matched entities) and find the closest one. However, we face two major challenges for biomedical entity set expansion. First, the seed entities in each entity type set are diverse and sparse. For example, the

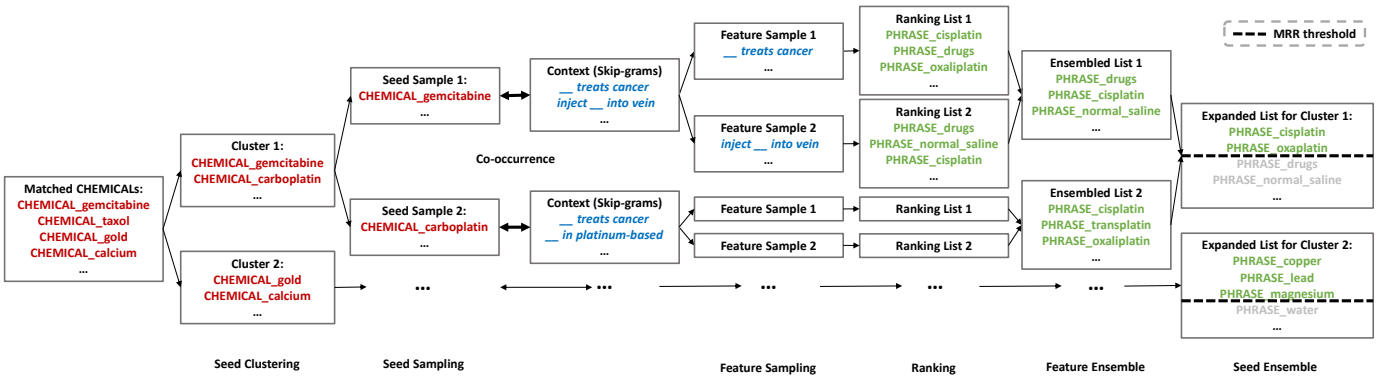

Fig. 2. The illustration of the entity expansion. Red tokens are seed entities for each type/cluster/sample. Blue tokens are skip-grams co-occurred with seed entities, serving as context features during expansion. Green tokens are expanded quality phrases. In the final ensemble step, phrases below the MRR threshold (in grey) will not be added into the entity sets.

CHEMICAL entity set contains highly distinct groups such as drugs (e.g., *gemcitabine*) and chemical elements (e.g., *gold*), which could be semantically distant from each other. Also, although there are abundant seed entities in each entity type set, each seed entity appears relatively infrequently in the corpus, providing sparse context information. Second, the candidate entities generated by phrase mining are noisy for entity set expansion. The candidate entities contain many noisy high-frequency entities such as "*p value*" and "*mm Hg*", which limits the performance of a simple semantic similarity comparison.

AUTOBIONER performs a novel entity set expansion with a careful design of the expansion mechanism to address the above challenges. We first group seed entities in each type set into clusters to reduce the seed diversity and sparsity. Then we perform both seed ensemble and feature ensemble to reduce the noises during set expansion. The candidate entities are ranked high for one entity type only if it satisfies both criteria: (1) it shares more context information with the seed entities of this type, and (2) it shares context information with more seed entities of this type. For example, in Figure 2, the expanded list for cluster 1 will contain "*cisplatin*" as a chemical with higher confidence than "*normal saline*", because it appears in the expansion results of more seed samples.

We give a formal definition of our entity expansion problem. For each entity type $t \in T$, we have an in-dictionary entity set $E_t = \{e_{t1}, ..., e_{tk}\}$. Besides, the remaining phrases not belonging to any E_t form a candidate set $P = \{p_1, ..., p_l\}$. In the following entity expansion step, we aim to "augment" each E_t by adding a set of candidates semantically close to $\{e_{t1}, ..., e_{tk}\}$. In other words, we are going to determine whether a candidate phrase has a desired type $t \in T$ or is Not-of-Interest (labeled as "NONE").

Semantic Closeness Scoring. Various ways have been proposed to model the semantic closeness, such as co-occurrence statistics and context features. We combine both signals during our set expansion.

Given a target phrase p, one of its skip-grams is "w_{-1} _ w_1" where w_{-1} and w_1 are two context words and p is replaced with a placeholder. For example, in Figure 2, a sentence "... *inject* CHEMICAL_*gemcitabine into vein* ..." provides one skip-gram of "*inject* _ *into vein*" for "CHEMICAL_*gemcitabine*". In our experiments, the maximum context window size is 4. One advantage of using skip-grams is that it imposes strong positional constraints.

Given a candidate phrase set P and a skip-gram (i.e., context feature) set C, we define the similarity between each pair of phrase p and context c using the TF-IDF transformation [22]:

$$f_{p,c} = \log(1 + X_{p,c})(\log|P| - \log \sum_{p' \in P} X_{p',c}),$$

where $X_{p,c}$ is the raw co-occurrence count between p and c. Empirically, [25] shows that such weight scaling outperforms some other alternatives such as point-wise mutual information (PMI) and BM25. Then the similarity between two phrases p_1 and p_2 under feature set C is defined as

$$sim(p_1, p_2|C) = \frac{\sum_{c \in C} \min(f_{p_1,c}, f_{p_2,c})}{\sum_{c \in C} \max(f_{p_1,c}, f_{p_2,c})}.$$

Given a seed entity set E and a skip-gram feature set C, each candidate phrase p can be scored as

$$score(p|E, C) = \frac{1}{|E|} \sum_{e \in E} sim(p, e|C).$$

Seed Clustering. Since each E_t is a highly diverse entity set, we cannot assume that one candidate phrase is close to *all* elements in E_t. For example, a chemical phrase as a candidate may be close to one subcategory of chemicals (e.g., drugs) but distant from others (e.g., chemical elements). We use k-Means to cluster each entity set into subcategories $E_{t1}, ..., E_{tH}$ according to their word2vec embeddings [20]. Then we conduct both *feature ensemble* and *seed ensemble* to select the best candidate entities for each seed cluster.

Seed Ensemble. For each cluster E_{th}, we sample N_E subsets $E_{th}^{(j)}$ $(i = 1, 2, ..., N_E)$. Each of the seed subsets contains M'_E $(M'_E < |E_{th}|)$ features.

Feature Ensemble. For a context feature set C, we first score each $c \in C$ based on its accumulated strength with entities in E (i.e., $\sum_{e \in E} f_{e,c}$). Then M_C skip-grams with the

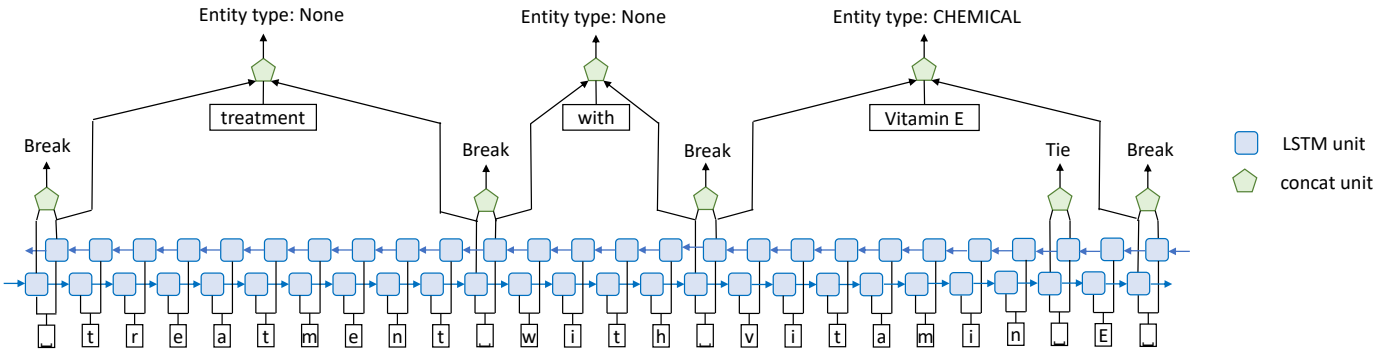

Fig. 3. The illustration of the neural architecture with the Tie or Break tagging scheme [24]. We first predict whether each connection is B (Break) or T (Tie), given it is not a connection within a quality phrase (Unknown). Then for a token span between two B's (e.g., "Vitamin E"), we predict its entity type as CHEMICAL.

highest scores will be selected, from which we sample N_C subsets C_i $(i = 1, 2, ..., N_C)$. Each of the subsets contains M'_C $(M'_C < M_C)$ features.

For each C_i and $E^{(j)}_{th}$, we obtain a ranking list of phrases according to their $score(\cdot|C_i, E^{(j)}_{th})$. Suppose the rank of p in terms of $score(\cdot|C_i, E^{(j)}_{th})$ is r_{pij}, the mean reciprocal rank of p is

$$MRR(p|E_{th}) = \frac{1}{N_C N_E} \sum_{i=1}^{N_C} \sum_{j=1}^{N_E} \frac{1}{r_{pij}}.$$

The phrases with MRR higher than a threshold MRR_{thrs} will be added into type E_t.

C. Distant Training

After dictionary matching and entity set expansion, we can obtain a training corpus (with an arbitrarily large size) with distant supervision. We proceed to the training process and adopt a "Tie-or-Break" tagging scheme other than the popular "BIOES" one, inspired by AutoNER [24].

In the "Tie-or-Break" scheme, instead of labeling each token, the connection between each two adjacent tokens has a label of 3 possible types: (1) T: "Tie", which means the two tokens belong to the same entity; (2) U: "Unknown", indicates that at least one of the two tokens belongs to an untyped quality phrase; and (3) B: "Break" for other cases. For example, in Figure 3, the connection between "*treatment*" and "*with*" is a Break since these two tokens do not belong to the same entity; whereas the connection between "*Vitamin*" and "*E*" is a Tie since these two tokens belong to the same CHEMICAL entity. Under this scheme, the NER model learning is divided into two steps: entity span detection and entity typing.

For entity span detection, a binary classifier is built to determine whether a connection has a label of B or T, while U will be skipped. To predict whether the connection y_i between tokens w_{i-1} and w_i is B, a BiLSTM layer is utilized to encode the character and word sequences. Then all the encoding vectors of y_i will be concatenated as one vector u_i and fed into a Sigmoid layer. I.e.,

$$p(y_i = \text{B}|u_i) = \text{Sigmoid}(w^T u_i).$$

Therefore, the loss function is computed as

$$\mathcal{L}_1 = \sum_{y_i \neq \text{U}} l(y_i, p(y_i = \text{B}|u_i)),$$

where $l(\cdot, \cdot)$ is the logistic loss.

After the entity boundary is determined, we can represent each token span with a new vector v_j and feed it into a Softmax layer to determine its entity type. I.e.,

$$p(t_j = t|v_j) = \frac{\exp(e_t^T v_j)}{\sum_{t' \in T} \exp(e_{t'}^T v_j)},$$

where t_j denotes the label of entity span j and e_t is the embedding vector of type t.

The loss function here is defined as

$$\mathcal{L}_2 = \sum_j H(\hat{p}(\cdot|v_i, T_j), p(\cdot|v_i)),$$

where $H(\cdot, \cdot)$ is the cross entropy function and $\hat{p}(\cdot|v_i, T_j)$ is the soft supervision distribution. T_j denotes the set of possible labels for t_j. Since in the distant training set, the type of a quality phrase is unknown (e.g., a phrase can be "CHEMICAL", "DISEASE" or "NONE"), the soft supervision distribution is adapted as

$$\hat{p}(t_j|v_i, T_j) = \frac{\mathbf{1}_{t_j \in T_j} \exp(e_{t_j}^T v_j)}{\sum_{t' \in T} \mathbf{1}_{t' \in T_j} \exp(e_{t'}^T v_j)}.$$

Note that the architecture in Figure 3 does not include a CRF layer. Therefore, the distant training step can be accelerated substantially in contrast to the popular LSTM-CRF models [13], [19], [31].

IV. EXPERIMENTS

We aim to answer three questions in our experiments. First, how does AUTOBIONER perform on benchmark datasets in comparison with state-of-the-art NER models (both fully supervised and distantly supervised)? Second, what is the contribution of entity expansion in our framework? Third, since AUTOBIONER does not require any human annotated training data, how does it perform on some new entity types (e.g., pathways and biological processes) with dictionaries?

TABLE I
NER PERFORMANCE COMPARISON. THE PERFORMANCES OF BILSTM-CRF ARE REPORTED IN [31]. THE PERFORMANCES OF OTHER ALGORITHMS, EXCEPT DICTIONARY-EXPANSION AND AUTOBIONER, ARE REPORTED IN [24]. SWELLSHARK HAS NO ANNOTATED DATA, BUT FOR ENTITY SPAN EXTRACTION, IT REQUIRES PRE-TRAINED POS TAGGERS AND EXTRA HUMAN EFFORTS OF DESIGNING POS TAG-BASED REGULAR EXPRESSIONS AND/OR HAND-TUNING FOR SPECIAL CASES.

Method	Human Effort	BC5CDR			NCBI-Disease		
		Prec	Rec	F1	Prec	Rec	F1
BiLSTM-CRF [11]	Gold Annotations	87.60	86.25	**86.92**	86.11	85.49	**85.80**
SwellShark [9]	Regex Design + Special Case Tuning	86.11	82.39	84.21	81.6	80.1	**80.8**
	Regex Design	84.98	83.49	**84.23**	64.7	69.7	67.1
Dictionary-Match	None	93.93	58.35	71.98	90.59	56.15	69.32
Dictionary-Expansion	None	92.09	63.85	75.42	90.17	56.88	69.76
Fuzzy-LSTM-CRF [24]	None	88.27	76.75	82.11	79.85	67.71	73.28
AutoNER [24]	None	88.96	81.00	84.79	79.42	71.98	75.52
AUTOBIONER	None	87.34	84.53	**85.91**	77.98	75.31	**77.58**

A. Experimental Setup

Datasets. Two benchmark datasets are used in our quantitative comparison.

- **BC5CDR** [16] was first released in the BioCreative V Chemical Disease Relation task. It has 1,500 articles containing 15,935 CHEMICAL and 12,852 DISEASE mentions. The whole corpus is divided into three parts, each of which has 500 articles, for training, development and testing. For distantly supervised methods, human annotations in training and development sets are not visible.
- **NCBI-Disease** [7] is a benchmark dataset for disease entity recognition and normalization. The corpus contains 793 abstracts with 6,881 DISEASE entities, and it is separated into training (593), development (100) and testing (100) subsets. Again, only raw text is provided in training and development sets.

On these two datasets, we use MeSH[1] and CTD[2] databases as the dictionaries to match Chemical and Disease entities.

Algorithms. We evaluate the performance of AUTOBIONER by comparing it with the following methods.

- **BiLSTM-CRF** [11] is the *fully* supervised benchmark. Its performances on BC5CDR and NCBI-Disease are listed to check whether AUTOBIONER can deliver competitive performance.
- **SwellShark** [9] is a benchmark distantly supervised method in the biomedical domain. It needs no human annotated data. However, it requires extra expert effort for entity span detection on building POS tagger, designing effective regular expressions, and hand-tuning for special cases.
- **Fuzzy-LSTM-CRF** [24] is adapted from BiLSTM-CRF [11], [13]. The char- and word-level BiLSTM architecture is retained but the CRF layer becomes Fuzzy-CRF so that it can support a "modified BIOES" scheme.
- **AutoNER** [24] is a recent state-of-the-art distantly supervised method. After dictionary matching, it trains a BiLSTM-Softmax architecture.
- **Dictionary-Match** is a simple distantly supervised baseline. We first generate quality phrases using AutoPhrase [23] and then type the entities if they can be matched from the dictionary.
- **Dictionary-Expansion** is an ablation of our AUTOBIONER framework. It only conducts the dictionary matching and entity expansion steps. We compare it with Dictionary-Match to show the improvement of our entity expansion step.

Parameters. In phrase mining, we generate quality phrases by setting the thresholds 0.5 and 0.9 for single-word and multi-word phrases, respectively. We have 552 Chemical entities and 449 Disease entities matched on BC5CDR and NCBI-Disease. In entity expansion, we divide each type of entities into 10 clusters. During seed ensemble and feature ensemble, we set M'_E, M_C and M'_C to be 10, 200 and 120, respectively. The MRR_{thrs} on these two datasets are 0.05. In distant training, we utilize a set of pretrained word embeddings[3]. The optimization method is gradient descent with momentum. The batch size and the momentum are set to be 10 and 0.9. The learning rate is set to 0.05. Dropout of a ratio 0.5 is applied. Gradient clipping of 5.0 is used for a better stability.

B. Quantitative Study

Comparison on Benchmark Datasets. Entity-level F1, precision and recall scores on the two benchmark datasets are reported in Table I. We first compare methods which require no extra human effort. In this setting, Dictionary-Match always has the highest precision. However, since dictionaries cannot cover every case in a corpus, its recall is rather low. Dictionary-Expansion improves the recall by 5.5% on BC5CDR, indicating the entity expansion step can accurately select quality phrases belonging to each type and improve the quality of distant training corpus. It also boosts the F1 by 3.4% on BC5CDR. Both Fuzzy-LSTM-CRF and AutoNER aim to use neural models for distant training. As we can see, the AutoNER architecture performs better on both datasets. By incorporating entity expansion into the neural model, AUTOBIONER consistently outperforms AutoNER by a clear margin (1.12% on BC5CDR and 2.06% on NCBI-Disease).

We would like to emphasize that Dictionary-Match, Dictionary-Expansion and AutoNER can all be viewed as

[1]https://www.nlm.nih.gov/mesh/

[2]http://ctdbase.org

[3]http://bio.nlplab.org/

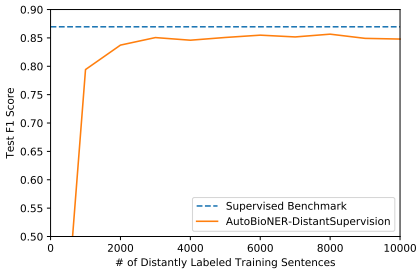

Fig. 4. Scale-up F1 scores for AUTOBIONER on BC5CDR. The x-axis is the number of unlabeled documents used for training.

ablation versions of the full AUTOBIONER framework. By outperforming the ablations, we demonstrate that every module in our method is necessary for improving the performance. In fact, at each step we tend to discover some new entities that are missed in the previous step. Since any predictor cannot be fully correct, the precision keeps decreasing. However, we managed to let the recall rise "faster" than the precision falls, which makes the F1 increase.

Although SwellShark utilizes much more expert effort, AUTOBIONER outperforms it on both corpora except when it has the "regular expression design" trick. AUTOBIONER still outperforms SwellShark on BC5CDR when the entity span matcher in SwellShark is carefully tuned by experts. Moreover, AUTOBIONER is competitive to the supervised benchmarks on BC5CDR, where the F1 score is only 1.01% away.

F1 vs. Size of Distant Training Corpus. Now we explore the change of test F1 scores when we have different sizes of distantly supervised corpus. On BC5CDR, we randomly sample sentences from the given training set (without gold labels) and then evaluate AUTOBIONER trained on the selected sentences. The curves can be found in Figure 4. The x-axis is the number of distantly supervised training sentences while the y-axis is the F1 score on the testing set. One can observe a significant growing trend of test F1 score in the beginning, but later the increasing rate slows down and fluctuates around some value slightly below the supervised benchmark.

C. Qualitative Study

Now we focus on *why* AUTOBIONER can outperform state-of-the-art distantly supervised methods. As we claimed, the unique module of AUTOBIONER is entity expansion. In Table I, we have quantitatively demonstrated the effectiveness of this step by showing that Dictionary-Expansion outperforms Dictionary-Match. We now lay out the expanded CHEMICAL and DISEASE entities.

As mentioned above, we divide each type of entities into 10 clusters and do entity expansion. Due to space limitation, we only show the expansion results of 5 (resp. 3) CHEMICAL (resp. DISEASE) clusters here (Tables II and III). For each cluster, we list 2 representative seed entities as well as 5 expanded entities with the highest MRR values. Entities in grey are judged as incorrect.

From Table II, we observe that the entity expansion step actually generates new entities with a high accuracy. 22 out of 25 (i.e., 88%) newly expanded chemicals are judged as correct. Although we only list 2 entities for each cluster, quite a few of them have corresponding similar entities in the expanded list. To name a few, in Cluster 1, *adenosine* (seed) and *oxprenolol* (expanded) are both drugs used in the treatment of arrhythmia; in Cluster 2, both *asenapine* (seed) and *quetiapine* (expanded) are used for bipolar disorders; in Cluster 3, *paracetamol* (seed) and *aspirin* (expanded) are for headaches. Similar observations can be found in Table III, where 13 out of 15 (i.e., 87%) newly expanded diseases are judged as correct.

Note that all the information we use in entity expansion are skip-grams and their co-occurrence statistics with the entities. One may think this is similar to directly using word embeddings. We have also tried this direction. We first adopt word2vec [20] to learn the representations of each entity. Given the seed entities, other candidates are ranked by the sum of distances away from the seeds. However, on the same clusters, such an embedding method has $\sim$72% accuracy for both CHEMICAL and DISEASE entities, which are significantly lower than those of our framework. This is because embedding similarities only consider semantics while ignores co-occurrence frequency. For entities that are not so frequent, their context in the corpus is limited, so the quality of their representations learned by word2vec may not be good. However, the embedding method gives vectors with good and bad qualities the same weight. In contrast, our framework cares both semantics and frequency. If the extracted entities co-occurs very often with some skip-grams, we have a good reason to trust the quality of its context information.

D. Case Study

Besides chemicals and diseases, many other entity types are also studied by biomedical researchers. For example, CTD [6] contains PATHWAY entities; the Gene Ontology[4] database [5] has Biological Processes, Molecular Functions and Cellular Components. To see how AUTOBIONER performs on these new entity types, we conduct a case study on a subset corpus of PubMed abstracts.

We randomly sample 248,064 tuples (two entities together with their relation) from the CTD database and extract all the PubMed abstracts associated with these tuples. This subset corpus with 302,736 sentences is much larger than the previous two benchmarks. We use MeSH and CTD to match GENE, CHEMICAL, DISEASE and PATHWAY entities, and we use Gene Ontology to match Biological Processes (BP), Molecular Functions (MF) and Cellular Components (CC).

Table IV shows the annotation results of AUTOBIONER and two baselines, Dictionary-Match and PubTator[5] [32]. PubTator is a well-established web-based tool for annotating PubMed papers. Note that it has a *fully-supervised* framework and incorporates massive training data.

In the first sentence, PubTator detects a long Disease entity "*lung tumor necrosis*", but it misses two synonymous

[4]http://www.geneontology.org/

[5]https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/

TABLE II

CHEMICAL ENTITY EXPANSION RESULTS OF AUTOBIONER ON BC5CDR. 5 OUT OF 10 CLUSTERS ARE SHOWN HERE. FOR EACH CLUSTER, WE LIST 2 REPRESENTATIVE SEEDS AS WELL AS 5 EXPANDED ENTITIES WITH THE HIGHEST MRR VALUES. ENTITIES IN GREY ARE JUDGED AS INCORRECT.

Seeds	{**adenosine, tamoxifen, ...**}	{**yohimbine, asenapine, ...**}	{**nicotine, paracetamol, ...**}	{**doxorubicin, haloperidol, ...**}	{**heparin, amiodarone, ...**}
Expanded Entities	thiabendazole	cefotetan	aspirin	argatroban	quetiapine
	oxprenolol	ketoconazole	hypokalemia	flunitrazepam	galantamine
	chinese herbs	quinpirole	cabergoline	melatonin	dl sotalol
	ketoconazole	NSAIDs	fenoldopam	erdosteine	OCs
	dihydrothienopyridine calcium	quetiapine	metformin	vincristin	nitroprusside

TABLE III

DISEASE ENTITY EXPANSION RESULTS OF AUTOBIONER ON BC5CDR. 3 OUT OF 10 CLUSTERS ARE SHOWN HERE. FOR EACH CLUSTER, WE LIST 2 REPRESENTATIVE SEEDS AS WELL AS 5 EXPANDED ENTITIES WITH THE HIGHEST MRR VALUES. ENTITIES IN GREY ARE JUDGED AS INCORRECT.

Seeds	{**myocardial ischemia, heart failure, ...**}	{**jaundice, hyperthermia, ...**}	{**headache, hyperalgesia, ...**}
Expanded Entities	bilateral optic neuropathy	antithymocyte globulin	axonal neuropathy
	NMS	mitochondrial abnormalities	bipolar mania
	haemopericardium and gastrointestinal haemorrhage	isotretinoin embryopathy	beta thalassemia
	posterior leukoencephalopathy	extrapyramidal signs	obese
	postoperative delirium	bevacizumab irinotecan	milk alkali syndrome

TABLE IV

CASE STUDY OF NER RESULTS ON A PUBMED SUBSET CORPUS. DETECTED ENTITIES ARE MARKED IN BOLD. IN CONTRAST TO DICTIONARY-MATCH, AUTOBIONER HAS A HIGHER RECALL. IN CONTRAST TO PUBTATOR [32], AUTOBIONER IS ABLE TO DETECT NEW TYPES OF ENTITIES.

	PMID: 22446026
PubTator	This is supported by findings that [**lung tumor necrosis**]$_{\text{DISEASE}}$ factor and lipocalin Lcn2 expression ...
Dictionary	This is supported by findings that lung tumor necrosis factor and [**lipocalin**]$_{\text{GENE}}$ Lcn2 expression ...
AUTOBIONER	This is supported by findings that [**lung tumor necrosis**]$_{\text{DISEASE}}$ factor and [**lipocalin**]$_{\text{GENE}}$ [**Lcn2**]$_{\text{GENE}}$ expression ...
	PMID: 22144121
PubTator	The most meaningful changes were observed amongst proteins involved in [**carbohydrate**]$_{\text{CHEMICAL}}$ metabolism, endoplasmic reticulum (ER) stress, [**calcium**]$_{\text{CHEMICAL}}$ homeostasis and apoptosis.
Dictionary	The most meaningful changes were observed amongst proteins involved in carbohydrate metabolism, [**endoplasmic reticulum**]$_{\text{CC}}$ (ER) stress, [**calcium**]$_{\text{CHEMICAL}}$ homeostasis and [**apoptosis**]$_{\text{BP}}$.
AUTOBIONER	The most meaningful changes were observed amongst proteins involved in [**carbohydrate**]$_{\text{CHEMICAL}}$ [**metabolism**]$_{\text{PATHWAY}}$, [**endoplasmic reticulum**]$_{\text{CC}}$ (ER) stress, [**calcium**]$_{\text{CHEMICAL}}$ homeostasis and [**apoptosis**]$_{\text{BP}}$.
	PMID: 22085608
PubTator	Additionally, the altered proteins were associated with the molecular functions of binding, catalytic activity, enzyme regulator activity and transporter activity, and involved in biological processes of apoptosis, developmental and immune system process, as well as response to stimulus.
Dictionary	Additionally, the altered proteins were associated with the molecular functions of [**binding**]$_{\text{MF}}$, [**catalytic activity**]$_{\text{MF}}$, [**enzyme regulator activity**]$_{\text{MF}}$ and [**transporter activity**]$_{\text{MF}}$, and involved in biological processes of [**apoptosis**]$_{\text{BP}}$, developmental and [**immune system process**]$_{\text{BP}}$, as well as [**response to stimulus**]$_{\text{BP}}$.
AUTOBIONER	Additionally, the altered proteins were associated with the molecular functions of [**binding**]$_{\text{MF}}$, [**catalytic activity**]$_{\text{MF}}$, [**enzyme regulator activity**]$_{\text{MF}}$ and [**transporter activity**]$_{\text{MF}}$, and involved in biological processes of [**apoptosis**]$_{\text{BP}}$, developmental and [**immune system process**]$_{\text{BP}}$, as well as [**response to stimulus**]$_{\text{BP}}$.

Gene entities "*lipocalin*" and "*Lcn2*". In contrast, Dictionary-Match finds "*lipocalin*" but ignores "*lung tumor necrosis*". AUTOBIONER is able to find all entities mentioned above. In the second sentence, the chemical "*carbohydrate*" is not in our dictionaries, but AUTOBIONER succeeds to recover it in the final result. For new entity types that PubTator does not cover, Dictionary-Match finds several instances for each of them, based on which AUTOBIONER can expand more (e.g., "*metabolism*" as a pathway in the second sentence). In the third sentence, there is a structure "*... with the molecular function of ...*" before "*binding*", "*enzyme regulatory activity*" and "*transporter activity*". AUTOBIONER manages to recognize all the three entities as MF. Similarly, there is a structure "*... and involved in biological processes of ...*" before "*apoptosis*", "*developmental and immune system process*" and "*response to stimulus*". AUTOBIONER also manages to recognize all the three entities as BP.

To summarize, in contrast to Dictionary-Match, AUTOBIONER can detect entities not included in dictionaries and achieve a high recall. In contrast to PubTator, AUTOBIONER is able to detect new types of entities accurately with the help of dictionaries.

V. CONCLUSIONS

We have proposed a distantly supervised approach, AUTOBIONER, to automatically recognize biomedical entities from massive corpora with user-input dictionaries. AUTOBIONER relies on incomplete entity dictionaries to provide seeds for each entity type and a novel entity set expansion method for corpus-level new entity recognition and dictionary completion. The expanded dictionaries are used as distant supervision to train a neural model for BioNER. AUTOBIONER achieves the

best performance among the methods that only use dictionaries with no additional human effort on BioNER benchmark datasets. Several cases are also demonstrated on new types of entities that do not have any existing training data, such as pathways and biological processes. Future work includes (1) extending the study to more entity types and fine-grained type levels (e.g., the UMLS database [26] has tens of entity types) and (2) exploring the potential of dictionary-based distant supervision in nested biomedical entity recognition [8], [30].

ACKNOWLEDGMENT

Research was sponsored in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreements No. W911NF-17-C-0099 and FA8750-19-2-1004, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, DTRA HDTRA11810026, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). Any opinions, findings, and conclusions or recommendations expressed in this document are those of the author(s) and should not be interpreted as the views of any U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] R. K. Ando. BioCreative II gene mention tagging system at IBM Watson. In *Proc. Second BioCreative Chall. Eval. Work.*, volume 23, pages 101–103, 2007.

[2] A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[3] J. P. C. Chiu and E. Nichols. Named Entity Recognition with Bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.*, 4:357–370, 2016.

[4] M. Cokol, I. Iossifov, C. Weinreb, and A. Rzhetsky. Emergent behavior of growing knowledge about molecular interactions. *Nature biotechnology*, 23(10):1243–1247, 2005.

[5] G. O. Consortium. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261, 2004.

[6] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly. The comparative toxicogenomics database: update 2017. *Nucleic acids research*, 45(D1):D972–D978, 2016.

[7] R. I. Doğan, R. Leaman, and Z. Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.

[8] J. R. Finkel and C. D. Manning. Nested named entity recognition. In *EMNLP'09*, pages 141–150. ACL, 2009.

[9] J. Fries, S. Wu, A. Ratner, and C. Ré. Swellshark: A generative model for biomedical named entity recognition without labeled data. *arXiv preprint arXiv:1704.06360*, 2017.

[10] J. M. Giorgi and G. D. Bader. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, btz504, 2019.

[11] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

[12] J. Huang, F. Gutierrez, D. Dou, J. A. Blake, K. Eilbeck, D. A. Natale, B. Smith, Y. Lin, X. Wang, Z. Liu, et al. A semantic approach for knowledge capture of microrna-target gene interactions. In *BIBM'15*, pages 975–982. IEEE, 2015.

[13] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *NAACL-HLT'16*, pages 260–270. ACL, 2016.

[14] R. Leaman and Z. Lu. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32(18):2839–2846, 2016.

[15] U. Leser and J. Hakenberg. What makes a gene name? Named entity recognition in the biomedical literature. *Briefings Bioinf.*, 6(4):357–369, 2005.

[16] J. Li, Y. Sun, R. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu. Annotating chemicals, diseases, and their interactions in biomedical literature. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, pages 173–182, 2015.

[17] Z. Li, Z. Yang, H. Lin, J. Wang, Y. Gui, Y. Zhang, and L. Wang. Cidextractor: A chemical-induced disease relation extraction system for biomedical literature. In *BIBM'16*, pages 994–1001. IEEE, 2016.

[18] Y. Lu, D. Ji, X. Yao, X. Wei, and X. Liang. CHEMDNER system with mixed conditional random fields and multi-scale word clustering. *J. Cheminf.*, 7(S1):S4, 2015.

[19] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL'16*, pages 1064–1074, 2016.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS'13*, pages 3111–3119. MIT Press, 2013.

[21] X. Ren, Z. Wu, W. He, M. Qu, C. R. Voss, H. Ji, T. F. Abdelzaher, and J. Han. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *WWW'17*, pages 1015–1024. IW3C, 2017.

[22] X. Rong, Z. Chen, Q. Mei, and E. Adar. Egoset: Exploiting word ego-networks and user-generated ontology for multifaceted set expansion. In *WSDM'16*, pages 645–654. ACM, 2016.

[23] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[24] J. Shang, L. Liu, X. Ren, X. Gu, T. Ren, and J. Han. Learning named entity tagger using domain-specific dictionary. In *EMNLP'18*, pages 2054–2064. ACL, 2018.

[25] J. Shen, Z. Wu, D. Lei, J. Shang, X. Ren, and J. Han. Setexpan: Corpus-based set expansion via context feature selection and rank ensemble. In *ECML-PKDD'17*, pages 288–304. Springer, 2017.

[26] L. Soldaini and N. Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, SIGIR*, 2016.

[27] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, and P. Bork. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, 45(D1):D362–D368, 2017.

[28] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn. Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic acids research*, 44(D1):D380–D384, 2015.

[29] A. Vlachos and C. Gasperin. Bootstrapping and evaluating named entity recognition in the biomedical domain. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 138–145, 2006.

[30] X. Wang, Y. Zhang, Q. Li, C. H. Wu, and J. Han. Penner: Pattern-enhanced nested named entity recognition in biomedical literature. In *BIBM'18*, pages 540–547. IEEE, 2018.

[31] 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*, 35(10):1745–1752, 2019.

[32] C.-H. Wei, H.-Y. Kao, and Z. Lu. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522, 2013.

[33] B. Xie, Q. Ding, H. Han, and D. Wu. mircancer: a microrna–cancer association database constructed by text mining on literature. *Bioinformatics*, 29(5):638–644, 2013.

[34] W. Yoon, C. H. So, J. Lee, and J. Kang. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC bioinformatics*, 20(10):249, 2019.

[35] G. Zhou and J. Su. Named entity recognition using an hmm-based chunk tagger. In *ACL'02*, pages 473–480. ACL, 2002.

[36] G. Zhou and J. Su. Exploring deep knowledge resources in biomedical name recognition. In *Proc. Int. Jt. Work. Nat. Lang. Process. Biomed. its Appl.*, pages 96–99, 2004.