Learning Named Entity Tagger using Domain-Specific Dictionary

Jingbo Shang [†*] **Liyuan Liu** [†*] **Xiaotao Gu** [†] **Xiang Ren** [♯] **Teng Ren** [‡] **Jiawei Han** [†]
[†] University of Illinois at Urbana-Champaign, Urbana, IL, USA
[♯] University of Southern California, Los Angeles, CA, USA
[‡] CooTek Inc., Shanghai, China
[†]{shang7, ll2, xiaotao2, hanj}@illinois.edu [♯]xiangren@usc.edu [‡]teng.ren@cootek.cn

Abstract

Recent advances in deep neural models allow us to build reliable named entity recognition (NER) systems without handcrafting features. However, such methods require large amounts of manually-labeled training data. There have been efforts on replacing human annotations with distant supervision (in conjunction with external dictionaries), but the generated noisy labels pose significant challenges on learning effective neural models. Here we propose two neural models to suit noisy distant supervision from the dictionary. First, under the traditional sequence labeling framework, we propose a revised fuzzy CRF layer to handle tokens with multiple possible labels. After identifying the nature of noisy labels in distant supervision, we go beyond the traditional framework and propose a novel, more effective neural model AutoNER with a new `Tie or Break` scheme. In addition, we discuss how to refine distant supervision for better NER performance. Extensive experiments on three benchmark datasets demonstrate that AutoNER achieves the best performance when only using dictionaries with no additional human effort, and delivers competitive results with state-of-the-art supervised benchmarks.

1 Introduction

Recently, extensive efforts have been made on building reliable named entity recognition (NER) models without handcrafting features (Liu et al., 2018; Ma and Hovy, 2016; Lample et al., 2016). However, most existing methods require large amounts of manually annotated sentences for training supervised models (e.g., neural sequence models) (Liu et al., 2018; Ma and Hovy, 2016; Lample et al., 2016; Finkel et al., 2005). This is particularly challenging in specific domains, where domain-expert annotation is expensive and/or slow to obtain.

To alleviate human effort, distant supervision has been applied to automatically generate labeled data, and has gained successes in various natural language processing tasks, including phrase mining (Shang et al., 2018), entity recognition (Ren et al., 2015; Fries et al., 2017; He, 2017), aspect term extraction (Giannakopoulos et al., 2017), and relation extraction (Mintz et al., 2009). Meanwhile, open knowledge bases (or dictionaries) are becoming increasingly popular, such as WikiData and YAGO in the general domain, as well as MeSH and CTD in the biomedical domain. The existence of such dictionaries makes it possible to generate training data for NER at a large scale without additional human effort.

Existing distantly supervised NER models usually tackle the entity span detection problem by heuristic matching rules, such as POS tag-based regular expressions (Ren et al., 2015; Fries et al., 2017) and exact string matching (Giannakopoulos et al., 2017; He, 2017). In these models, every unmatched token will be tagged as non-entity. However, as most existing dictionaries have limited coverage on entities, simply ignoring unmatched tokens may introduce false-negative labels (e.g., "prostaglandin synthesis" in Fig. 1). Therefore, we propose to extract high-quality out-of-dictionary phrases from the corpus, and mark them as potential entities with a special "unknown" type. Moreover, every entity span in a sentence can be tagged with multiple types, since two entities of different types may share the same surface name in the dictionary. To address these challenges, we propose and compare two neural architectures with customized tagging schemes.

We start with adjusting models under the traditional sequence labeling framework. Typically, NER models are built upon conditional random

[*]Equal contribution.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2054–2064
Brussels, Belgium, October 31 - November 4, 2018.

fields (CRF) with the IOB or IOBES tagging scheme (Liu et al., 2018; Ma and Hovy, 2016; Lample et al., 2016; Ratinov and Roth, 2009; Finkel et al., 2005). However, such design cannot deal with multi-label tokens. Therefore, we customize the conventional CRF layer in LSTM-CRF (Lample et al., 2016) into a Fuzzy CRF layer, which allows each token to have multiple labels without sacrificing computing efficiency.

To adapt to imperfect labels generated by distant supervision, we go beyond the traditional sequence labeling framework and propose a new prediction model. Specifically, instead of predicting the label of each single token, we propose to predict whether two adjacent tokens are tied in the same entity mention or not (i.e., broken). The key motivation is that, even the boundaries of an entity mention are mismatched by distant supervision, most of its inner ties are not affected, and thus more robust to noise. Therefore, we design a new Tie or Break tagging scheme to better exploit the noisy distant supervision. Accordingly, we design a novel neural architecture that first forms all possible entity spans by detecting such ties, then identifies the entity type for each span. The new scheme and neural architecture form our new model, AutoNER, which proves to work better than the Fuzzy CRF model in our experiments.

We summarize our major contributions as

- We propose AutoNER, a novel neural model with the new Tie or Break scheme for the distantly supervised NER task.
- We revise the traditional NER model to the Fuzzy-LSTM-CRF model, which serves as a strong distantly supervised baseline.
- We explore to refine distant supervision for better NER performance, such as incorporating high-quality phrases to reduce false-negative labels, and conduct ablation experiments to verify the effectiveness.
- Experiments on three benchmark datasets demonstrate that AutoNER achieves the best performance when only using dictionaries with no additional human effort and is even competitive with the supervised benchmarks.

We release all code and data for future studies[1]. Related open tools can serve as the NER module of various domain-specific systems in a plug-in-and-play manner.

[1] https://github.com/shangjingbo1226/AutoNER

2 Overview

Our goal, in this paper, is to learn a named entity tagger using, and only using dictionaries. Each dictionary entry consists of 1) the surface names of the entity, including a canonical name and a list of synonyms; and 2) the entity type. Considering the limited coverage of dictionaries, we extend existing dictionaries by adding high-quality phrases as potential entities with unknown type. More details on refining distant supervision for better NER performance will be presented in Sec. 4.

Given a raw corpus and a dictionary, we first generate entity labels (including unknown labels) by exact string matching, where conflicted matches are resolved by maximizing the total number of matched tokens (Etzioni et al., 2005; Hanisch et al., 2005; Lin et al., 2012; He, 2017).

Based on the result of dictionary matching, each token falls into one of three categories: 1) it belongs to an entity mention with one or more known types; 2) it belongs to an entity mention with unknown type; and 3) It is marked as non-entity.

Accordingly, we design and explore two neural models, Fuzzy-LSTM-CRF with the modified IOBES scheme and AutoNER with the Tie or Break scheme, to learn named entity taggers based on such labels with unknown and multiple types. We will discuss the details in Sec. 3.

3 Neural Models

In this section, we introduce two prediction models for the distantly supervised NER task, one under the traditional sequence labeling framework and another with a new labeling scheme.

3.1 Fuzzy-LSTM-CRF with Modified IOBES

State-of-the-art named entity taggers follow the sequence labeling framework using IOB or IOBES scheme (Ratinov and Roth, 2009), thus requiring a conditional random field (CRF) layer to capture the dependency between labels. However, both the original scheme and the conventional CRF layer cannot handle multi-typed or unknown-typed tokens. Therefore, we propose the modified IOBES scheme and Fuzzy CRF layer accordingly, as illustrated in Figure 1.

Modified IOBES. We define the labels according to the three token categories. 1) For a token

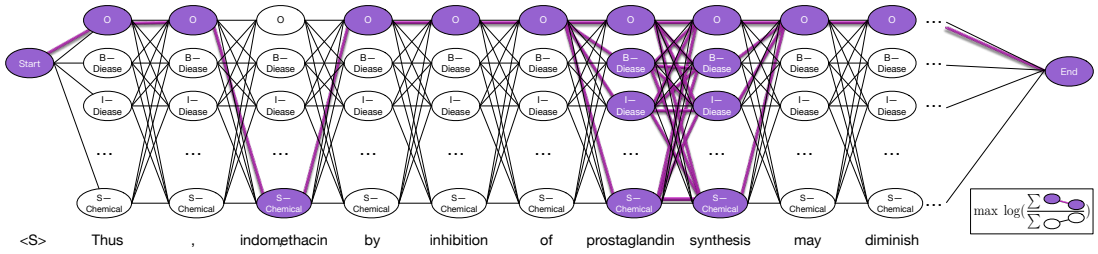


Figure 1: The illustration of the Fuzzy CRF layer with modified IOBES tagging scheme. The named entity types are {Chemical, Disease}. "indomethacin" is a matched Chemical entity and "prostaglandin synthesis" is an unknown-typed high-quality phrase. Paths from Start to End marked as purple form all possible label sequences given the distant supervision.

marked as one or more types, it is labeled with all these types and one of {I, B, E, S} according to its positions in the matched entity mention. 2) For a token with unknown type, all five {I, O, B, E, S} tags are possible. Meanwhile, all available types are assigned. For example, when there are only two available types (e.g., Chemical and Disease), it has nine (i.e., $4 \times 2 + 1$) possible labels in total. 3) For a token that is annotated as non-entity, it is labeled as O.

As demonstrated in Fig. 1, based on the dictionary matching results, "indomethacin" is a singleton Chemical entity and "prostaglandin synthesis" is an unknown-typed high-quality phrase. Therefore, "indomethacin" is labeled as S-Chemical, while both "prostaglandin" and "synthesis" are labeled as O, B-Disease, I-Disease, ..., and S-Chemical because the available entity types are {Chemical, Disease}. The non-entity tokens, such as "Thus" and "by", are labeled as O.

Fuzzy-LSTM-CRF. We revise the LSTM-CRF model (Lample et al., 2016) to the Fuzzy-LSTM-CRF model to support the modified IOBES labels.

Given a word sequence $(X_1, X_2, \ldots, X_n)$, it is first passed through a word-level BiLSTM (Hochreiter and Schmidhuber, 1997) (i.e., forward and backward LSTMs). After concatenating the representations from both directions, the model makes independent tagging decisions for each output label. In this step, the model estimates the score P_{i,y_j} for the word X_i being the label y_j.

We follow previous works (Liu et al., 2018; Ma and Hovy, 2016; Lample et al., 2016) to define the score of the predicted sequence, the score of the predicted sequence $(y_1, y_2, \ldots, y_n)$ is defined as:

$$s(X, y) = \sum_{i=0}^{n} \Phi_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \qquad (1)$$

where, $\Phi_{y_i,y_{i+1}}$ is the transition probability from a label y_i to its next label y_{i+1}. Φ is a $(k+2) \times (k+2)$ matrix, where k is the number of distinct labels. Two additional labels start and end are used (only used in the CRF layer) to represent the beginning and end of a sequence, respectively.

The conventional CRF layer maximizes the probability of the only valid label sequence. However, in the modified IOBES scheme, one sentence may have multiple valid label sequences, as shown in Fig. 1. Therefore, we extend the conventional CRF layer to a fuzzy CRF model. Instead, it maximizes the total probability of all possible label sequences by enumerating both the IOBES tags and all matched entity types. Mathematically, we define the optimization goal as Eq. 2.

$$p(y|X) = \frac{\sum_{\tilde{y} \in Y_{possible}} e^{s(X,\tilde{y})}}{\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}} \qquad (2)$$

where Y_X means all the possible label sequences for sequence X, and $Y_{possible}$ contains all the possible label sequences given the labels of modified IOBES scheme. Note that, when all labels and types are known and unique, the fuzzy CRF model is equivalent to the conventional CRF.

During the training process, we maximize the log-likelihood function of Eq. 2. For inference, we apply the Viterbi algorithm to maximize the score of Eq. 1 for each input sequence.

3.2 AutoNER with "Tie or Break"

Identifying the nature of the distant supervision, we go beyond the sequence labeling framework and propose a new tagging scheme, Tie or Break. It focuses on the ties between adjacent tokens, i.e., whether they are tied in the same entity mentions or broken into two parts. Accordingly, we design a novel neural model for this scheme.

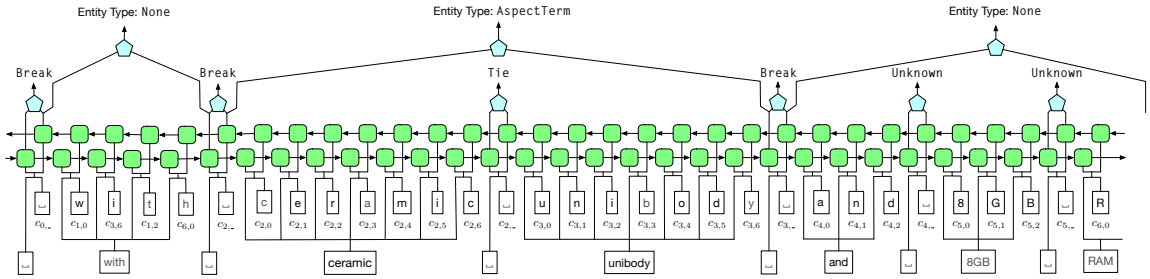


Figure 2: The illustration of AutoNER with `Tie or Break` tagging scheme. The named entity type is {`AspectTerm`}. "ceramic unibody" is a matched `AspectTerm` entity and "8GB RAM" is an unknown-typed high-quality phrase. `Unknown` labels will be skipped during the model training.

"Tie or Break" Tagging Scheme. Specifically, for every two adjacent tokens, the connection between them is labeled as (1) `Tie`, when the two tokens are matched to the same entity; (2) `Unknown`, if at least one of the tokens belongs to an unknown-typed high-quality phrase; (3) `Break`, otherwise.

An example can be found in Fig. 2. The distant supervision shows that "ceramic unibody" is a matched `AspectTerm` and "8GB RAM" is an unknown-typed high-quality phrase. Therefore, a `Tie` is labeled between "ceramic" and "unibody", while `Unknown` labels are put before "8GB", between "8GB" and "RAM", and after "RAM".

Tokens between every two consecutive `Break` form a token span. Each token span is associated with all its matched types, the same as for the modified `IOBES` scheme. For those token spans without any associated types, such as "with" in the example, we assign them the additional type `None`.

We believe this new scheme can better exploit the knowledge from dictionary according to the following two observations. First, even though the boundaries of an entity mention are mismatched by distant supervision, most of its inner ties are not affected. More interestingly, compared to multi-word entity mentions, matched unigram entity mentions are more likely to be false-positive labels. However, such false-positive labels will not introduce incorrect labels with the `Tie or Break` scheme, since either the unigram is a true entity mention or a false positive, it always brings two `Break` labels around.

AutoNER. In the `Tie or Break` scheme, entity spans and entity types are encoded into two folds. Therefore, we separate the entity span detection and entity type prediction into two steps.

For entity span detection, we build a binary classifier to distinguish `Break` from `Tie`, while `Unknown` positions will be skipped. Specifically, as shown in Fig. 2, for the prediction between i-th token and its previous token, we concatenate the output of the BiLSTM as a new feature vector, $\mathbf{u}_i$. $\mathbf{u}_i$ is then fed into a sigmoid layer, which estimates the probability that there is a `Break` as

$$p(y_i = \texttt{Break}|\mathbf{u}_i) = \sigma(\mathbf{w}^T\mathbf{u}_i)$$

where y_i is the label between the i-th and its previous tokens, σ is the sigmoid function, and $\mathbf{w}$ is the sigmoid layer's parameter. The entity span detection loss is then computed as follows.

$$\mathcal{L}_{\text{span}} = \sum_{i|y_i \neq \texttt{Unknown}} l\big(y_i, p(y_i = \texttt{Break}|\mathbf{u}_i)\big)$$

Here, $l(\cdot,\cdot)$ is the logistic loss. Note that those `Unknown` positions are skipped.

After obtaining candidate entity spans, we further identify their entity types, including the `None` type for non-entity spans. As shown in Fig. 2, the output of the BiLSTM will be re-aligned to form a new feature vector, which is referred as $\mathbf{v}_i$ for i-th span candidate. $\mathbf{v}_i$ will be further fed into a softmax layer, which estimates the entity type distribution as

$$p(t_j|\mathbf{v}_i) = \frac{e^{\mathbf{t}_j^T\mathbf{v}_i}}{\sum_{t_k \in L} e^{\mathbf{t}_k^T\mathbf{v}_i}}$$

where t_j is an entity type and L is the set of all entity types including `None`.

Since one span can be labeled as multiple types, we mark the possible set of types for i-th entity span candidate as L_i. Accordingly, we modify the cross-entropy loss as follows.

$$\mathcal{L}_{\text{type}} = H(\hat{p}(\cdot|\mathbf{v}_i, L_i), p(\cdot|\mathbf{v}_i))$$

Here, $H(p, q)$ is the cross entropy between p and q, and $\hat{p}(t_j|\mathbf{v}_i, L_i)$ is the soft supervision distribu-

tion. Specifically, it is defined as:

$$\hat{p}(t_j|\mathbf{v}_i, L_i) = \frac{\delta(t_j \in L_i) \cdot e^{\mathbf{t}_j^T \mathbf{v}_i}}{\sum_{t_k \in L} \delta(t_k \in L_i) \cdot e^{\mathbf{t}_k^T \mathbf{v}_i}}$$

where $\delta(t_j \in L_i)$ is the boolean function of checking whether the i-th span candidate is labeled as the type t_j in the distant supervision.

It's worth mentioning that AutoNER has no CRF layer and Viterbi decoding, thus being more efficient than Fuzzy-LSTM-CRF for inference.

3.3 Remarks on "Unknown" Entities

"Unknown" entity mentions are not the entities of other types, but the tokens that we are less confident about their boundaries and/or cannot identify their types based on the distant supervision. For example, in Figure 1, "prostaglandin synthesis" is an "unknown" token span. The distant supervision cannot decide whether it is a `Chemical`, a `Disease`, an entity of other types, two separate single-token entities, or (partially) not an entity. Therefore, in the FuzzyCRF model, we assign all possible labels for these tokens.

In our AutoNER model, these "unknown" positions have undefined boundary and type losses, because (1) they make the boundary labels unclear; and (2) they have no type labels. Therefore, they are skipped.

4 Distant Supervision Refinement

In this section, we present two techniques to refine the distant supervision for better named entity taggers. Ablation experiments in Sec. 5.4 verify their effectiveness empirically.

4.1 Corpus-Aware Dictionary Tailoring

In dictionary matching, blindly using the full dictionary 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 a non-related character name "Wednesday Addams"[2] and its alias "Wednesday", many Wednesday's will be wrongly marked as persons. In an ideal case, the dictionary should cover, and only cover entities occurring in the given corpus to ensure a high precision while retaining a reasonable coverage.

As an approximation, we tailor the original dictionary to a corpus-related subset by excluding entities whose canonical names never appear in the given corpus. The intuition behind is that to avoid ambiguities, people will likely mention the canonical name of the entity at least once. For example, in the biomedical domain, this is true for 88.12%, 95.07% of entity mentions on the BC5CDR and NCBI datasets respectively. We expect the NER model trained on such tailored dictionary will have a higher precision and a reasonable recall compared to that trained on the original dictionary.

4.2 Unknown-Typed High-Quality Phrases

Another issue of the distant supervision is about the false-negative labels. When a token span cannot be matched to any entity surface names in the dictionary, because of the limited coverage of dictionaries, it is still difficult to claim it as non-entity (i.e., negative labels) for sure. Specifically, some high-quality phrases out of the dictionary may also be potential entities.

We utilize the state-of-the-art distantly supervised phrase mining method, AutoPhrase (Shang et al., 2018), with the corpus and dictionary in the given domain as input. AutoPhrase only requires unlabeled text and a dictionary of high-quality phrases. We obtain quality multi-word and single-word phrases by posing thresholds (e.g., 0.5 and 0.9 respectively). In practice, one can find more unlabeled texts from the same domain (e.g., PubMed papers and Amazon laptop reviews) and use the same domain-specific dictionary for the NER task. In our experiments, for the biomedical domain, we use the titles and abstracts of 686,568 PubMed papers (about 4%) uniformly sampled from the whole PubTator database as the training corpus. For the laptop review domain, we use the Amazon laptop review dataset[3], which is designed for the aspect-based sentiment analysis (Wang et al., 2011).

We treat out-of-dictionary phrases as potential entities with "unknown" type and incorporate them as new dictionary entries. After this, only token spans that cannot be matched in this extended dictionary will be labeled as non-entity. Being aware of these high-quality phrases, we expect the trained NER tagger should be more accurate.

[2] https://en.wikipedia.org/wiki/Wednesday_Addams

[3] http://times.cs.uiuc.edu/~wang296/Data/

Table 1: Dataset Overview.

Dataset	BC5CDR	NCBI-Disease	LaptopReview
Domain	Biomedical	Biomedical	Technical Review
Entity Types	`Disease`, `Chemical`	`Disease`	`AspectTerm`
Dictionary	MeSH + CTD	MeSH + CTD	Computer Terms
Raw Sent. #	20,217	7,286	3,845

5 Experiments

We conduct experiments on three benchmark datasets to evaluate and compare our proposed Fuzzy-LSTM-CRF and AutoNER with many other methods. We further investigate the effectiveness of our proposed refinements for the distant supervision and the impact of the number of distantly supervised sentences.

5.1 Experimental Settings

Datasets are briefly summarized in Table 1. More details as as follows.

- **BC5CDR** is from the most recent BioCreative V Chemical and Disease Mention Recognition task. It has 1,500 articles containing 15,935 `Chemical` and 12,852 `Disease` mentions.
- **NCBI-Disease** focuses on Disease Name Recognition. It contains 793 abstracts and 6,881 `Disease` mentions.
- **LaptopReview** is from the SemEval 2014 Challenge, Task 4 Subtask 1 (Pontiki et al., 2014) focusing on laptop aspect term (e.g., "disk drive") Recognition. It consists of 3,845 review sentences and 3,012 `AspectTerm` mentions.

All datasets are publicly available. The first two datasets are already partitioned into three subsets: a training set, a development set, and a testing set. For the LaptopReview dataset, we follow (Giannakopoulos et al., 2017) and randomly select 20% from the training set as the development set. Only raw texts are provided as the input of distantly supervised models, while the gold training set is used for supervised models.

Domain-Specific Dictionary. For the biomedical datasets, the dictionary is a combination of both the MeSH database[4] and the CTD Chemical and Disease vocabularies[5]. The dictionary contains 322,882 `Chemical` and `Disease` entity surfaces. For the laptop review dataset, the dictionary has 13,457 computer terms crawled from a public website[6].

Metric. We use the micro-averaged F_1 score as the evaluation metric. Meanwhile, precision and recall are presented. The reported scores are the mean across five different runs.

Parameters and Model Training. Based on the analysis conducted in the development set, we conduct optimization with the stochastic gradient descent with momentum. We set the batch size and the momentum to 10 and 0.9. The learning rate is initially set to 0.05 and will be shrunk by 40% if there is no better development F_1 in the recent 5 rounds. Dropout of a ratio 0.5 is applied in our model. For a better stability, we use gradient clipping of 5.0. Furthermore, we employ the early stopping in the development set.

Pre-trained Word Embeddings. For the biomedical datasets, we use the pre-trained 200-dimension word vectors [7] from (Pyysalo et al., 2013), which are trained on the whole PubMed abstracts, all the full-text articles from PubMed Central (PMC), and English Wikipedia. For the laptop review dataset, we use the GloVe 100-dimension pre-trained word vectors[8] instead, which are trained on the Wikipedia and GigaWord.

5.2 Compared Methods

Dictionary Match is our proposed distant supervision generation method. Specifically, we apply it to the testing set directly to obtain entity mentions with exactly the same surface name as in the dictionary. The type is assigned through a majority voting. By comparing with it, we can check the improvements of neural models over the distant supervision itself.

SwellShark, in the biomedical domain, is arguably the best distantly supervised model, especially on the BC5CDR and NCBI-Disease datasets (Fries et al., 2017). 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.

Distant-LSTM-CRF achieved the best performance on the LaptopReview dataset without annotated training data using a distantly supervised

[4] https://www.nlm.nih.gov/mesh/download_mesh.html

[5] http://ctdbase.org/downloads/

[6] https://www.computerhope.com/jargon.htm

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

[8] https://nlp.stanford.edu/projects/glove/

Table 2: [Biomedical Domain] NER Performance Comparison. The supervised benchmarks on the BC5CDR and NCBI-Disease datasets are LM-LSTM-CRF and LSTM-CRF respectively (Wang et al., 2018). 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 other than Dictionary	BC5CDR			NCBI-Disease		
		Pre	Rec	F1	Pre	Rec	F1
Supervised Benchmark	Gold Annotations	88.84	85.16	**86.96**	86.11	85.49	**85.80**
SwellShark	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
Fuzzy-LSTM-CRF		88.27	76.75	82.11	79.85	67.71	73.28
AutoNER		88.96	81.00	**84.8**	79.42	71.98	**75.52**

Table 3: [Technical Review Domain] NER Performance Comparison. The supervised benchmark refers to the challenge winner.

Method	LaptopReview		
	Pre	Rec	F1
Supervised Benchmark	84.80	66.51	**74.55**
Distant-LSTM-CRF	74.03	31.59	53.93
Dictionary Match	90.68	44.65	59.84
Fuzzy-LSTM-CRF	85.08	47.09	60.63
AutoNER	72.27	59.79	**65.44**

LSTM-CRF model (Giannakopoulos et al., 2017).

Supervised benchmarks on each dataset are listed to check whether AutoNER can deliver competitive performance. On the BC5CDR and NCBI-Disease datasets, LM-LSTM-CRF (Liu et al., 2018) and LSTM-CRF (Lample et al., 2016) achieve the state-of-the-art F_1 scores without external resources, respectively (Wang et al., 2018). On the LaptopReview dataset, we present the scores of the Winner in the SemEval2014 Challenge Task 4 Subtask 1 (Pontiki et al., 2014).

5.3 NER Performance Comparison

We present F_1, precision, and recall scores on all datasets in Table 2 and Table 3. From both tables, one can find the AutoNER achieves the best performance when there is no extra human effort. Fuzzy-LSTM-CRF does have some improvements over the Dictionary Match, but it is always worse than AutoNER.

Even though SwellShark is designed for the biomedical domain and utilizes much more expert effort, AutoNER outperforms it in almost all cases. The only outlier happens on the NCBI-disease dataset when the entity span matcher in SwellShark is carefully tuned by experts for many special cases.

It is worth mentioning that AutoNER beats Distant-LSTM-CRF, which is the previous state-of-the-art distantly supervised model on the LaptopReview dataset.

Moreover, AutoNER's performance is competitive to the supervised benchmarks. For example, on the BC5CDR dataset, its F_1 score is only 2.16% away from the supervised benchmark.

5.4 Distant Supervision Explorations

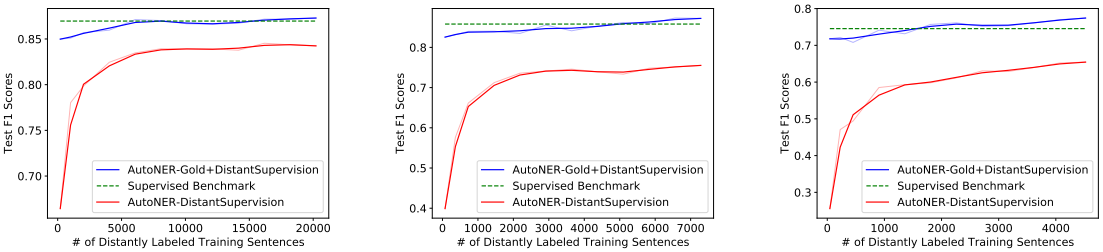
We investigate the effectiveness of the two techniques that we proposed in Sec. 4 via ablation experiments. As shown in Table 4, using the tailored dictionary always achieves better F_1 scores than using the original dictionary. By using the tailored dictionary, the precision of the AutoNER model will be higher, while the recall will be retained similarly. For example, on the NCBI-Disease dataset, it significantly boosts the precision from 53.14% to 77.30% with an acceptable recall loss from 63.54% to 58.54%. Moreover, incorporating unknown-typed high-quality phrases in the dictionary enhances every score of AutoNER models significantly, especially the recall. These results match our expectations well.

5.5 Test F_1 Scores vs. Size of Raw Corpus

Furthermore, we explore the change of test F_1 scores when we have different sizes of distantly supervised texts. We sample sentences uniformly random from the given raw corpus and then evaluate AutoNER models trained on the selected sentences. We also study what will happen when the gold training set is available. The curves can be found in Figure 3. The X-axis is the number of

Table 4: Ablation Experiments for Dictionary Refinement. The dictionary for the LaptopReview dataset contains no alias, so the corpus-aware dictionary tailoring is not applicable.

Method	BC5CDR			NCBI-Disease			LaptopReview		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
AutoNER w/ Original Dict	82.79	70.40	76.09	53.14	63.54	57.87	69.96	49.85	58.21
AutoNER w/ Tailored Dict	84.57	70.22	76.73	77.30	58.54	66.63	Not Applicable		
AutoNER w/ Tailored Dict & Phrases	88.96	81.00	**84.8**	79.42	71.98	**75.52**	72.27	59.79	**65.44**



(a) BC5CDR (b) NCBI (c) LaptopReview

Figure 3: AutoNER: Test F_1 score vs. the number of distantly supervised sentences.

distantly supervised training sentences while the Y-axis is the F_1 score on the testing set.

When using distant supervision only, one can observe a significant growing trend of test F_1 score in the beginning, but later the increasing rate slows down when there are more and more raw texts.

When the gold training set is available, the distant supervision is still helpful to AutoNER. In the beginning, AutoNER works worse than the supervised benchmarks. Later, with enough distantly supervised sentences, AutoNER outperforms the supervised benchmarks. We think there are two possible reasons: (1) The distant supervision puts emphasis on those matchable entity mentions; and (2) The gold annotation may miss some good but matchable entity mentions. These may guide the training of AutoNER to a more generalized model, and thus have a higher test F_1 score.

5.6 Comparison with Gold Supervision

To demonstrate the effectiveness of distant supervision, we try to compare our method with gold annotations provided by human experts.

Specifically, we conduct experiments on the BC5CDR dataset by sampling different amounts of annotated articles for model training. As shown in Figure 4, we found that our method outperforms the supervised method by a large margin when less training examples are available. For example, when there are only 50 annotated articles available, the test F1 score drops substantially to 74.29%. To achieve a similar test F1 score (e.g., 83.91%) as our AutoNER models (i.e., 84.8%), the supervised benchmark model requires at least 300 annotated articles. Such results indicate the effectiveness and usefulness of AutoNER on the scenario without sufficient human annotations.

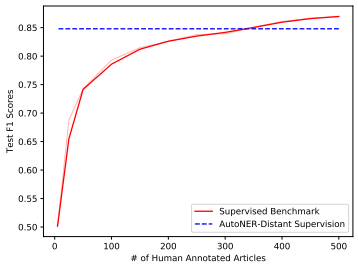


Figure 4: AutoNER: Test F_1 score vs. the number of human annotated articles.

Still, we observe that, when the supervised benchmark is trained with all annotations, it achieves the performance better than AutoNER. We conjugate that this is because AutoNER lacks more advanced techniques to handle distant supervision, and we leave further improvements of AutoNER to the future work.

6 Related Work

The task of supervised named entity recognition (NER) is typically embodied as a sequence labeling problem. Conditional random fields (CRF) models built upon human annotations and handcrafted features are the standard (Finkel et al., 2005; Settles, 2004; Leaman and Gonzalez, 2008). Recent advances in neural models have freed do-

main experts from handcrafting features for NER tasks. (Lample et al., 2016; Ma and Hovy, 2016; Liu et al., 2018). Such neural models are increasingly common in the domain-specific NER tasks (Sahu and Anand, 2016; Dernoncourt et al., 2017; Wang et al., 2018). Semi-supervised methods have been explored to further improve the accuracy by either augmenting labeled datasets with word embeddings or bootstrapping techniques in tasks like gene name recognition (Kuksa and Qi, 2010; Tang et al., 2014; Vlachos and Gasperin, 2006). Unlike these existing approaches, our study focuses on the distantly supervised setting without any expert-curated training data.

Distant supervision has attracted many attentions to alleviate human efforts. Originally, it was proposed to leverage knowledge bases to supervise relation extraction tasks (Craven et al., 1999; Mintz et al., 2009). AutoPhrase has demonstrated powers in extracting high-quality phrases from domain-specific corpora like scientific papers and business reviews (Shang et al., 2018) but it cannot categorize phrases into typed entities in a context-aware manner. We incorporate the high-quality phrases to enrich the domain-specific dictionary.

There are attempts on the distantly supervised NER task recently (Ren et al., 2015; Fries et al., 2017; He, 2017; Giannakopoulos et al., 2017). For example, SwellShark (Fries et al., 2017), specifically designed for biomedical NER, 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 require extensive expert effort to cover many special cases. Other methods (Ren et al., 2015; He, 2017) also utilize similar approaches to extract entity span candidates before entity typing. Distant-LSTM-CRF (Giannakopoulos et al., 2017) has been proposed for the distantly supervised aspect term extraction, which can be viewed as an entity recognition task of a single type for business reviews. As shown in our experiments, our models can outperform Distant-LSTM-CRF significantly on the laptop review dataset.

To the best of our knowledge, AutoNER is the most effective model that can learn NER models by using, and only using dictionaries without any additional human effort.

7 Conclusion and Future Work

In this paper, we explore how to learn an effective NER model by using, and only using dictionaries. We design two neural architectures, Fuzzy-LSTM-CRF model with a modified `IOBES` tagging scheme and AutoNER with a new `Tie or Break` scheme. In experiments on three benchmark datasets, AutoNER achieves the best F_1 scores without additional human efforts. Its performance is even competitive to the supervised benchmarks with full human annotation. In addition, we discuss how to refine the distant supervision for better NER performance, including incorporating high-quality phrases mined from the corpus as well as tailoring dictionary according to the given corpus, and demonstrate their effectiveness in ablation experiments.

In future, we plan to further investigate the power and potentials of the AutoNER model with `Tie or Break` scheme in different languages and domains. Also, the proposed framework can be further extended to other sequence labeling tasks, such as noun phrase chunking. Moreover, going beyond the classical NER setting in this paper, it is interesting to further explore distant supervised methods for the nested and multiple typed entity recognitions in the future.

Acknowledgments

We would like to thank Yu Zhang from University of Illinois at Urbana-Champaign for providing results of supervised benchmark methods on the BC5CDR and NCBI datasets.

Research was sponsored in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreement No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, DTRA HDTRA11810026, Google Ph.D. Fellowship 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

Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Jason Fries, Sen Wu, Alex Ratner, and Christopher Ré. 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data. *arXiv preprint arXiv:1704.06360*.

Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. 2017. Unsupervised aspect term extraction with b-lstm & crf using automatically labelled datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 180–188.

Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. 2005. Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(1):S14.

Wenqi He. 2017. Autoentity: automated entity detection from massive text corpora. *M.S. Thesis for Computer Science of University of Illinois at Urbana-Champaign*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Pavel P Kuksa and Yanjun Qi. 2010. Semi-supervised bio-named entity recognition with word-codebook learning. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 25–36. SIAM.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Robert Leaman and Graciela Gonzalez. 2008. Banner: an executable survey of advances in biomedical named entity recognition. In *Biocomputing 2008*, pages 652–663. World Scientific.

Thomas Lin, Oren Etzioni, et al. 2012. Entity linking at web scale. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 84–88. Association for Computational Linguistics.

Liyuan Liu, Jingbo Shang, Frank Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. *AAAI*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1064–1074.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, page 2735.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, pages 39–43.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*.

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, and Jiawei Han. 2015. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 995–1004. ACM.

Sunil Sahu and Ashish Anand. 2016. Recurrent neural network models for disease name recognition using domain invariant features. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2216–2225.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 104–107. Association for Computational Linguistics.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*.

Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. 2014. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, 2014.

Andreas Vlachos and Caroline Gasperin. 2006. 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. Association for Computational Linguistics.

Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 618–626. ACM.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2018. Cross-type biomedical named entity recognition with deep multi-task learning. *arXiv preprint arXiv:1801.09851*.