EMPIRICAL ANALYSIS OF UNLABELED ENTITY PROBLEM IN NAMED ENTITY RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

In many scenarios, named entity recognition (NER) models severely suffer from *unlabeled entity problem*, where the entities of a sentence may not be fully annotated. Through empirical studies performed on synthetic datasets, we find two causes of the performance degradation. One is the reduction of annotated entities and the other is treating unlabeled entities as negative instances. The first cause has less impact than the second one and can be mitigated by adopting pretraining language models. The second cause seriously misguides a model in training and greatly affects its performances. Based on the above observations, we propose a general approach that is capable of eliminating the misguidance brought by unlabeled entities. The core idea is using negative sampling to keep the probability of training with unlabeled entities at a very low level. Experiments on synthetic datasets and real-world datasets show that our model is robust to *unlabeled entity problem* and surpasses prior baselines. On well-annotated datasets, our model is competitive with state-of-the-art method[1].

1 INTRODUCTION

Named entity recognition (NER) is an important task in information extraction. Previous methods typically cast it into a sequence labeling problem by adopting IOB tagging scheme (Huang et al., 2015; Ma & Hovy, 2016; Akbik et al., 2018). A representative model is Bi-LSTM CRF (Lample et al., 2016). The great success achieved by these methods benefits from massive correctly labeled data. However, in some real scenarios, not all the entities in the training corpus are annotated. For example, in some NER tasks (Ling & Weld, 2012), the datasets contain too many entity types or a mention may be associated with multiple labels. Since manual annotation is too hard, some entities are neglected by human annotators. Situations in distantly supervised NER (Ren et al., 2015; Fries et al., 2017) are even more serious. To reduce handcraft annotation, distant supervision (Mintz et al., 2009) is applied to automatically produce labeled data. As a result, large amounts of entities in the corpus are missed due to the limited coverage of knowledge resources. We refer this to *unlabeled entity problem*, which largely degrades the performance of models.

There are several approaches used in prior works to alleviating this problem. Fuzzy CRF and AutoNER (Shang et al., 2018b) allow models to learn from the phrases that may be potential entities. However, since these phrases are obtained by using a distantly supervised phrase mining method (Shang et al., 2018a), many unlabeled entities in the training data may still not be recalled. In the context of only resorting to unlabeled corpora and an entity ontology, Peng et al. (2019) employ positive-unlabeled (PU) learning (Li & Liu, 2005) to unbiasedly and consistently estimate the task loss. In implementations, they build a distinct binary classifiers for each label. Nevertheless, the unlabeled entities still impact on the classifiers of the corresponding entity types and, importantly, the model can't disambiguate neighboring entities. Partial CRF (Tsuboi et al., 2008) is an extension of commonly used CRF (Lafferty et al., 2001) that supports learning from incomplete annotations. Yang et al. (2018); Nooralahzadeh et al. (2019) use it to circumvent the unlabeled entities. However, as fully annotated corpus is still required to get true negative instances, this approach is not applicable to the cases where no high-quality data is available.

In this work, our goal is to study what are the impacts of *unlabeled entity problem* on the models and how to effectively eliminate them. Initially, we construct some synthetic datasets and introduce

[1]The source code of this work is publicly available at https://XXX.

degradation rates. The datasets are constructed by randomly removing the annotated named entities in well-annotated datasets, e.g., CoNLL-2003 (Sang & De Meulder, 2003), with different probabilities. The degradation rates measure how severe an impact of *unlabeled entity problem* degrades the results of models. Extensive studies are investigated on synthetic datasets. We find two causes: the reduction of annotated entities and treating unlabeled entities as negative instances. The first cause is obvious but has far less impact than the second one. Besides, it can be mitigated well by using a pretraining language model, like BERT (Devlin et al., 2018), as the sentence encoder. The second cause seriously misleads the models in training and exerts a great negative impact on their performances. Even in less severe cases, it can sharply reduce the F1 score by about 20%. Based on the above observations, we propose a novel method that can almost completely eliminate the misguidance of unlabeled entities in training. The core idea is applying negative sampling to keep the probability of training with unlabeled entities at a very low level. In Section 5, we provide a proof to confirm its theoretical correctness.

We have conducted extensive experiments to verify the effectiveness of our approach. Studies on synthetic datasets and real-world datasets (e.g., EC) show that our model handles unlabeled entities well and significantly outperforms previous baselines. On well-annotated datasets (e.g., CoNLL-2003), our model is competitive with current state-of-the-art method.

2 PRELIMINARIES

In this section, we formally define the *unlabeled entity problem* and briefly describe a strong baseline, BERT Tagging (Devlin et al., 2018), used in empirical studies.

2.1 UNLABELED ENTITY PROBLEM

We denote an input sentence as $\mathbf{x} = [x_1, x_2, \cdots, x_n]$ and the annotated named entity set as $\mathbf{y} = \{y_1, y_2, \cdots, y_m\}$. n is the sentence length and m is the amount of entities. Each member y_k of set $\mathbf{y}$ is a tuple (i_k, j_k, l_k). (i_k, j_k) is the span of an entity which corresponds to the phrase $\mathbf{x}_{i_k,j_k} = [x_{i_k}, x_{i_k+1}, \cdots, x_{j_k}]$ and l_k is its label. The *unlabeled entity problem* is defined as, due to the limited coverage of machine annotator or the negligence of human annotator, some ground truth entities $\widetilde{\mathbf{y}}$ of the sentence $\mathbf{x}$ are not covered by annotated entity set $\mathbf{y}$.

2.2 BERT TAGGING

BERT Tagging adopts IOB tagging scheme, where each token x_i in a sentence $\mathbf{x}$ is labeled as B-tag if it's the beginning of an entity, I-tag if it's inside but not the first one within the entity, or O otherwise. Hence, its output is a n-length label sequence $\mathbf{z} = [z_1, z_2, \cdots, z_n]$.

Formally, BERT tagging firstly uses BERT to get the representation $\mathbf{h}_i$ for every token x_i:

$$[\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n] = \mathrm{BERT}(\mathbf{x}). \tag{1}$$

Then, the label distribution $\mathbf{q}_i$ is computed as $\mathrm{Softmax}(\mathbf{W}\mathbf{h}_i)$. In training, the loss is induced as $\sum_{1 \le i \le n} -\log \mathbf{q}_i[z_i]$. At test time, it obtains the label for each token x_i by $\arg\max \mathbf{q}_i$.

3 EMPIRICAL STUDIES

To understand the impacts of *unlabeled entity problem*, we conduct empirical studies over multiple synthetic datasets, different methods, and various metrics.

3.1 PREPARATIONS

Synthetic Datasets. To simulate poorly annotated datasets, we randomly remove the labeled entities of well-annotated datasets with different masking probabilities. The datasets are CoNLL-2003 (Sang & De Meulder, 2003) and OntoNotes 5.0 (Pradhan et al., 2013). Both datasets are usually used to test the performances of NER models. The probabilities p are respectively set as $0.0, 0.1, 0.2, \cdots, 0.9$. In this way, 2×10 synthetic datasets are constructed.

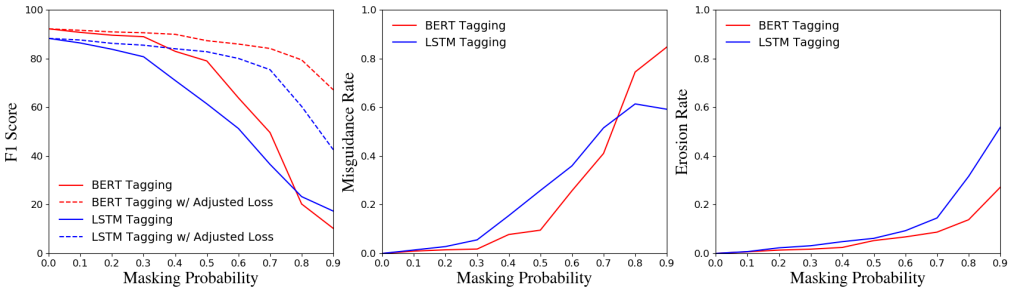

Figure 1: The empirical studies conducted on CoNLL-2003 dataset.

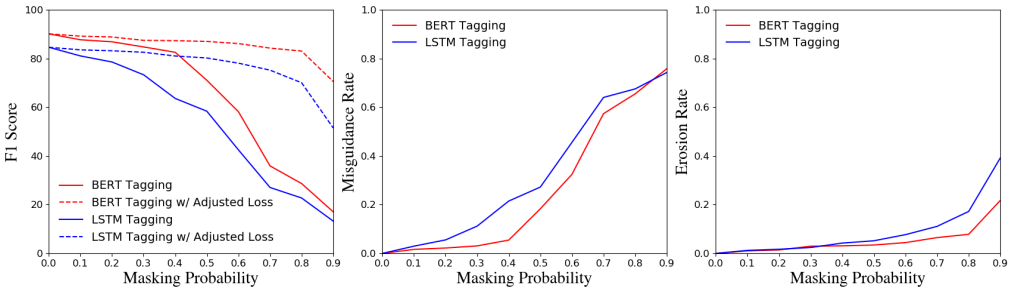

Figure 2: The empirical studies investigated on OntoNotes 5.0 dataset.

Methods. We adopt two models. One of them is BERT Tagging, which has long been regarded as a strong baseline. The other is LSTM Tagging that replaces the original encoder (i.e., BERT) of BERT Tagging with LSTM (Hochreiter & Schmidhuber, 1997). We use it to study the effect of using pretraining language model. To explore the negative impact brought by unlabeled entities in training, we present an adjusted training loss for above two models:

$$\Big(\sum_{1\leq i\leq n} -\log \mathbf{q}_i[z_i]\Big) - \Big(\sum_{(i',j',l')\in\widetilde{\mathbf{y}}} \sum_{i'\leq k\leq j'} -\log \mathbf{q}_k[z_k]\Big). \tag{2}$$

The idea here is to remove the incorrect loss incurred by unlabeled entities. Note that missed entity set $\widetilde{\mathbf{y}}$ is reachable in synthetic datasets but unknown in real-world datasets.

Metrics. Following prior works, the F1 scores of models are tested by using conlleval script[2]. We also design two degradation rates to measure the different impacts of *unlabeled entity problem*. One is erosion rate α_p and the other is misguidance rate β_p:

$$\alpha_p = \frac{f_0^a - f_p^a}{f_0^a}, \quad \beta_p = \frac{f_p^a - f_p}{f_p^a}. \tag{3}$$

For a synthetic dataset with the masking probability being p, f_p and f_p^a are the performances of a model and its adjusted version, respectively. Erosion rate α_p measures how severe the reduction of annotated entities degrades the F1 scores of a model. Misguidance rate β_p measures how serious unlabeled entities misguide the model during training.

3.2 Overall Analysis

The left parts of Figure 1 and Figure 2 show the results of empirical studies. From them, we can draw the following observations. Firstly, the significant downward trends of solid lines confirm the fact that NER models severely suffer from *unlabeled entity problem*. For example, by setting the masking probability as 0.4, the performance of LSTM Tagging decreases by 33.01% on CoNLL-2003 and 19.58% on OntoNotes 5.0. Secondly, in contrast, the dashed lines change very slowly,

[2]https://www.clips.uantwerpen.be/conll2000/chunking/conlleval.txt.

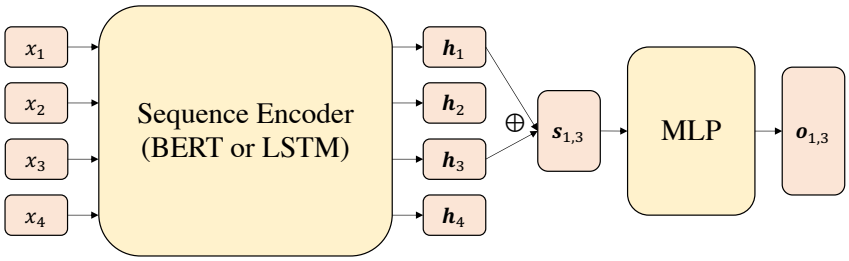

Figure 3: This demonstrates how our model scores possible entities.

indicating that the models with adjusted training loss (see Equation 2) are much less influenced by the issue. For instance, when masking probability is 0.7, adopting adjusted loss preserves the F1 scores of BERT Tagging by 41.04% on CoNLL-2003 and 57.38% on OntoNotes 5.0. Lastly, for high masking probabilities, even though the negative impact of unlabeled entities is eliminated by adjusting the training loss, the performance still declines to a certain extent. For example, when masking probability is set as 0.8, the F1 scores of adjusted LSTM Tagging decrease by 31.64% on CoNLL-2003 and 17.22% on OntoNotes 5.0.

Reduction of Annotated Entities. From the last observation, we can infer that a cause of performance degradation is the reduction of annotated entities. In the middle parts of Figure 1 and Figure 2, we plot the change of erosion rates α_p (see Equation 2) with respect to masking probabilities. We can see that its impact is not very serious when in low masking probabilities but can't be neglected when in high ones. Besides, using pre-training language models greatly mitigates the issue. As an example, when the probability is 0.8, on both CoNLL-2003 and OntoNotes 5.0, the erosion rates of adjusted BERT Tagging are only about half of those of adjusted LSTM Tagging.

Misguidance of Unlabeled Entities. From the last two observations, we can conclude that the primary cause is treating unlabeled entities as negative instances, which severely misleads the models during training. To better understand it, in the right parts of Figure 1 and Figure 2, we plot the change of misguidance rates β_p (see Equation 3) with masking probabilities. These rates are essentially the percentage decreases of F1 scores. From them, we can see that the impact of misguidance is very much serious even when in low masking probabilities.

4 METHODOLOGY

Motivated by Section 3.2, we present a model that is robust to the *unlabeled entity problem*.

4.1 SCORING MODEL WITH BERT

Based on the findings in Section 3.2, we use BERT as the default encoder to mitigate the reduction of annotated entities. Specifically, given a sentence $\mathbf{x}$, we firstly obtain the token representations $\mathbf{h}_i$ with Equation 1. Then, we get the representation for every phrase $\mathbf{x}_{i,j}$ as

$$\mathbf{s}_{i,j} = \mathbf{h}_i \oplus \mathbf{h}_j \oplus (\mathbf{h}_i - \mathbf{h}_j) \oplus (\mathbf{h}_i \odot \mathbf{h}_j), \tag{4}$$

where $\oplus$ is vector concatenation and $\odot$ is element-wise product.

Finally, multi-layer perceptron (MLP) computes the label distribution $\mathbf{o}_{i,j}$ for a span (i, j):

$$\mathbf{o}_{i,j} = \mathrm{Softmax}(\mathbf{U} \tanh(\mathbf{V}\mathbf{s}_{i,j})). \tag{5}$$

The term $\mathbf{o}_{i,j}[l]$ is the predicted score for an entity (i, j, l).

4.2 TRAINING VIA NEGATIVE SAMPLING

From Section 3.2, we know that regarding all the unlabeled spans as negative instances certainly degrades the performances of models, since some of them may be missed entities. Our solution to

this issue is negative sampling. Specifically, we randomly sample a small subset of unlabeled spans as the negative instances to induce the training loss.

Given the annotated entity set $\mathbf{y}$, we firstly get all the negative instance candidates as

$$\{(i, j, \mathrm{O}) \mid (i, j, l) \notin \mathbf{y}, 1 \leq i \leq j \leq n, l \in \mathcal{L}\}, \tag{6}$$

where $\mathcal{L}$ is the label space and O is the label for non-entity spans.

Then, we randomly sample a subset $\widehat{\mathbf{y}}$ from the whole candidate set. The size of sample set $\widehat{\mathbf{y}}$ is $\lceil \lambda * n \rceil, 0 < \lambda < 1$, where $\lceil\rceil$ is the ceiling function.

Ultimately, a span-level cross entropy loss used for training is incurred as

$$\Big(\sum_{(i,j,l) \in \mathbf{y}} -\log(\mathbf{o}_{i,j}[l]) \Big) + \Big(\sum_{(i',j',l') \in \widehat{\mathbf{y}}} -\log(\mathbf{o}_{i',j'}[l']) \Big). \tag{7}$$

In Section 5, we prove that, through negative sampling, the probability of training with unlabeled entities is kept at a very low level.

4.3 Inference

At test time, firstly, the label for every span (i, j) is obtained by $\arg\max_l \mathbf{o}_{i,j}[l]$. Then, we select the ones whose label l is not O as predicted entities. When the spans of inferred entities intersects, we preserve the one with the highest predicted score and discard the others.

5 Discussion

We prove that, through negative sampling, the probability of not treating a specific missed entity in a n-length sentence as the negative instance is larger than $1 - \frac{2}{n-5}$:

$$\begin{aligned} \prod_{0 \leq i < \lceil \lambda n \rceil} \Big(1 - \frac{1}{\frac{n(n-1)}{2} - m - i}\Big) &> \Big(1 - \frac{1}{\frac{n(n-1)}{2} - m - \lceil \lambda n \rceil}\Big)^{\lceil \lambda n \rceil} \\ &> \Big(1 - \frac{1}{\frac{n(n-1)}{2} - n - n}\Big)^{n} \geq \Big(1 - n * \frac{1}{\frac{n(n-1)}{2} - n - n}\Big) = 1 - \frac{2}{n-5} \end{aligned}. \tag{8}$$

Here $\frac{n(n-1)}{2} - m$ is the amount of negative instance candidates. Note that we use the facts $\lambda < 1$, $m \leq n$, and $(1 - z)^n \geq 1 - nz, 0 \leq z \leq 1$ during the proof.

6 Experiments

We have conducted extensive experiments on multiple datasets to verify the effectiveness of our method. Studies on synthetic datasets show that our model can almost eliminate the misguidance brought by unlabeled entities in training. On real-world datasets, our model has notably outperformed prior baselines and achieved the state-of-the-art performances. On well-annotated datasets, our model is competitive to current state-of-the-art method.

6.1 Settings

The setup of synthetic datasets and well-annotated datasets are the same as what we describe in Section 3.1. For real-world datasets, we use EC and NEWS, both of which are collected by (Yang et al., 2018). EC is in e-commerce domain, which has 5 entity types: Brand, Product, Model, Material, and Specification. The data contains 2400 sentences tagged by human annotators and are divided into three parts: 1200 for training, 400 for dev, and 800 for testing. Yang et al. (2018) also construct an entity dictionary of size 927 and apply distant supervision on a raw corpus to obtain additional 2500 sentences for training. NEWS is from MSRA dataset (Levow, 2006). Yang et al. (2018) only adopt the PERSON entity type. Training data of size 3000, dev data of size 3328, and testing data of size 3186 are all sampled from MSRA. They collect an entity dictionary of size 71664 and perform distant supervision on the rest data to obtain extra 3722 training cases by using

Masking Prob.	CoNLL-2003		OntoNotes 5.0	
	BERT Tagging	Our Model	BERT Tagging	Our Model
0.1	90.71	**91.37**	87.69	**89.20**
0.2	89.57	**91.25**	86.86	**89.15**
0.3	88.95	**90.53**	84.75	**88.73**
0.4	82.94	**89.73**	82.55	**88.20**
0.5	78.99	**89.22**	71.07	**88.17**
0.6	63.84	**87.65**	58.17	**87.53**

Table 1: The experiment results on two synthetic datasets.

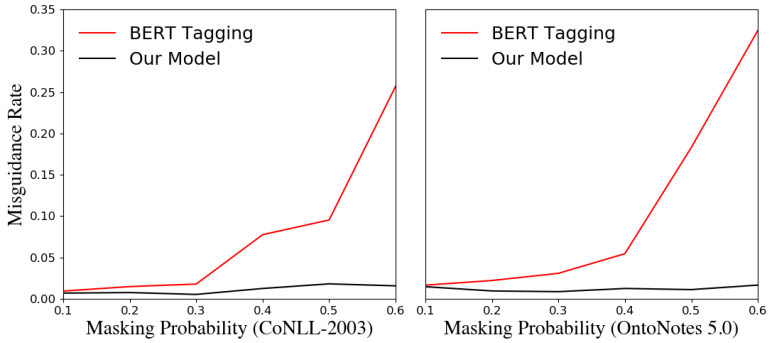

Figure 4: The misguidance rates of BERT Tagging and our model.

the dictionary. Both EC and NEWS contain a large amount of incompletely annotated sentences, and hence naturally suffer from the *unlabeled entity problem*.

We adopt the same neural network configurations for all the datasets. L2 regularization and dropout ratio are respectively set as 1×10^{-5} and 0.4 for reducing overfit. The dimension of scoring layers is 256. Ratio λ is set as 0.35. When the sentence encoder is LSTM, we set the hidden dimension as 512 and use cased, 300d GloVe (Pennington et al., 2014) to initialize the word embedding. We utilize Adam (Kingma & Ba, 2014) as the optimization algorithm and adopt the suggested hyper-parameters. At test time, we convert the predictions of our models into IOB format and use conlleval script to compute the F1 score. In all the experiments, the improvements of our models over the baselines are statistically significant with $p < 0.05$.

6.2 Results on Synthetic Datasets

In this section, our model is compared with BERT Tagging on the synthetic datasets of the masking probabilities being $0.1, 0.2, \cdots, 0.6$. From Table 1, we can get two conclusions. Firstly, our model significantly outperforms BERT Tagging, especially in high masking probabilities. For example, on CoNLL-2003, our F1 scores outnumber those of BERT Tagging by 1.88% when the probability is 0.2 and 27.16% when the probability is 0.6. Secondly, our model is very robust to the *unlabeled entity problem*. When increasing the masking probability from 0.1 to 0.5, the results of our model only decrease by 2.35% on CoNLL-2003 and 1.91% on OntoNotes 5.0.

Figure 4 demonstrates the misguidance rate comparisons between BERT Tagging and our models. The way to adjust our model is reformulating Equation (7) by defining the negative term via $\{(i, j, \mathrm{O}) \mid \forall l : (i, j, l) \notin \mathbf{y} \cup \widetilde{\mathbf{y}}\}$ rather than the negatively sampled $\hat{\mathbf{y}}$. The idea here is to avoid the unlabeled entities being sampled. From Figure 4, we can discover that, in all masking probabilities, the misguidance rates of our model are far smaller than those of BERT Tagging and are consistently lower than 2.50%. These indicate that, in training, our model indeed eliminates the misguidance brought by unlabeled entities to some extent.

Method	CoNLL-2003	OntoNotes 5.0
Flair Embedding (Akbik et al., 2018)	93.09	89.3
BERT-MRC (Li et al., 2019)	93.04	91.11
HCR w/ BERT (Luo et al., 2020)	93.37	90.30
BERT-Biaffine Model (Yu et al., 2020)	**93.5**	**91.3**
Our Model	93.42	90.59

Table 2: The experiment results on two well-annotated datasets.

Method		EC	NEWS
Yang et al. (2018)	String Matching via Ontology	44.02	47.75
	BiLSTM + CRF	54.59	69.09
	BiLSTM + CRF w/ RL	56.23	73.19
	BiLSTM + Partial CRF	60.08	78.38
	BiLSTM + Partial CRF w/ RL	61.45	79.22
Nooralahzadeh et al. (2019)	BiLSTM + Partial CRF w/ RL	63.56	80.04
This Work	Our Model	**66.17**	**85.39**
	Our Model w/o BERT, w/ BiLSTM	**64.68**	**82.11**

Table 3: The experiment results on two real-world datasets.

6.3 Results on Fully Annotated Datasets

We additionally apply our model with negative sampling on the well-annotated datasets where the issue of incomplete entity annotation is not serious. To train our model with negative sampling, λ is set to be 0.35 as in experiments of real-world datasets. As shown in Table 2, the F1 scores of our model are very close to current best results. Our model slightly underperforms BERT-Biaffine Model by only 0.09% on CoNLL-2003 and 0.78% on OntoNotes 5.0. Besides, our model surpasses many other strong baselines. On OntoNotes 5.0, our model outperforms HCR w/ BERT by 0.32% and Flair Embedding by 1.44%. On CoNLL-2003, the improvements of F1 scores are 0.41% over BERT-MRC and 0.35% over Flair Embedding. All these results indicate that our model is still very effective when applied to high-quality data.

6.4 Results on Real-world Datasets

For two real-world datasets, a large portion of training data is obtained via distant supervision. As stated in Yang et al. (2018), the F1 scores of string matching through an entity dictionary are notably declined in terms of the low recall scores, although its precision scores are higher than those of other methods. Therefore, *unlabeled entity problem* is serious in the datasets.

As shown in Table 3, the baselines come from two works (Yang et al., 2018; Nooralahzadeh et al., 2019). Yang et al. (2018) use Partial CRF to circumvent all possible unlabeled entities and leverage reinforcement learning (RL) to skip noisy annotation in an adaptive manner, and Nooralahzadeh et al. (2019) further improve their work by optimizing the policy of RL. A disadvantage of Partial CRF is that it can't be trained with the true negative instances in poorly labeled corpora. Our model has significantly outperformed prior baselines and obtained new state-of-the-art results. Compared with prior best model (Nooralahzadeh et al., 2019), we achieve the improvements of 3.94% on EC and 6.27% on NEWS. Compared with strong baseline, BiLSTM + Partial CRF, the increases of F1 scores are 9.20% and 8.21%. To make fair comparisons, we replace BERT with LSTM. Even so, we still outperform (Nooralahzadeh et al., 2019) by 1.76% on EC and 2.52% on NEWS. All these strongly confirm the effectiveness of our model.

6.5 Ratio λ in Negative Sampling

Intuitively, setting ratio λ (see Section 4.2) as too large values or too small values both are inappropriate. Large ratios amplify the contribution of negative entities, leading to underfitting. Small ratios reduce the amount of negative instances used for training, leading to overfitting. Figure 5 shows the experiments on some synthetic datasets with the ratio λ of our method being $0.1, 0.2, \cdots, 0.9$. From

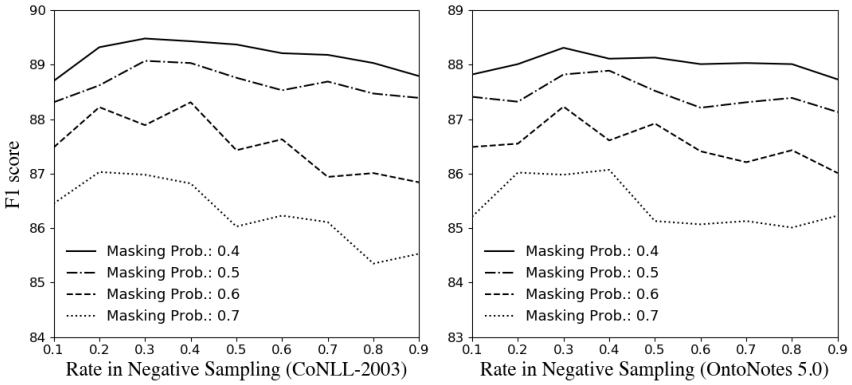

Figure 5: The results of our models with different ratio λ on synthetic datasets.

it, we can see that all the score curves are roughly arched, which verifies our intuition. Besides, we find that $0.3 < \lambda < 0.4$ performs well in all the cases.

7 Related Work

NER is a classical task in information extraction. Previous works generally regard it as a sequence labeling problem by using IOB tagging scheme (Huang et al., 2015; Lample et al., 2016; Akbik et al., 2018; Luo et al., 2020). Each word in the sentence is labeled as B-tag if it is the beginning of an entity, I-tag if it's inside but not the first one within the entity, or O otherwise. This approach is extensively studied in prior works. For example, Akbik et al. (2018) propose Flair Embedding that pretrains character embedding in large corpora and uses it rather than token representations to represent a sentence. Recently, there is a growing interest in span-based models (Li et al., 2019; Yu et al., 2020). They treat the spans, instead of single words, as the basic units for labeling. For example, Li et al. (2019) present BERT-MRC that regards NER as a MRC task, where named entities are extracted as retrieving answer spans.

In some practical applications (e.g., data annotation in fine-grained NER), NER models are faced with *unlabeled entity problem*, where the unlabeled entities seriously degrade the performances of models. Several approaches to this issue have been proposed. Fuzzy CRF and AutoNER (Shang et al., 2018b) allow learning from high-quality phrases. However, since these phrases are obtained through distant supervision, the unlabeled entities in the corpora may still be missed. PU learning (Li & Liu, 2005) unbiasedly and consistently estimates the training loss. Nevertheless, the unlabeled entities still impact on the classifiers of the corresponding entity types and, importantly, the model can't disambiguate neighboring entities. Partial CRF (Yang et al., 2018; Nooralahzadeh et al., 2019) supports learning from incomplete annotations. However, because fully annotated corpus is still in need to training models with true negative instances, this type of approach is not applicable to the situations where no high-quality data is available.

8 Conclusion

In this work, we study what are the impacts of unlabeled entities on NER models and how to effectively eliminate them. Through empirical studies performed on synthetic datasets, we find two causes: the reduction of annotated entities and treating unlabeled entities as negative instances. The first cause has less impact than the second one and can be mitigated by adopting pretraining language models. The second cause seriously misleads the models in training and greatly affects their performances. Based on the above observations, we propose a novel method that is capable of eliminating the misguidance of unlabeled entities in training. The core idea is applying negative sampling to keep the probability of training with unlabeled entities at a very low level. Experiments on synthetic datasets and real-world datasets show that our model handles unlabeled entities well and significantly outperforms previous baselines. On well-annotated datasets, our model is competitive with existing state-of-the-art approach.

REFERENCES

Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

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

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

Gina-Anne Levow. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pp. 108–117, 2006.

Xiao-Li Li and Bing Liu. Learning from positive and unlabeled examples with different data distributions. In *European conference on machine learning*, pp. 218–229. Springer, 2005.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*, 2019.

Xiao Ling and Daniel S Weld. Fine-grained entity recognition. In *AAAI*, volume 12, pp. 94–100, 2012.

Ying Luo, Fengshun Xiao, and Hai Zhao. Hierarchical contextualized representation for named entity recognition. In *AAAI*, pp. 8441–8448, 2020.

Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 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*, pp. 1003–1011, 2009.

Farhad Nooralahzadeh, Jan Tore Lønning, and Lilja Øvrelid. Reinforcement-based denoising of distantly supervised ner with partial annotation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 225–233, 2019.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. Distantly supervised named entity recognition using positive-unlabeled learning. *arXiv preprint arXiv:1906.01378*, 2019.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 143–152, 2013.

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, and Jiawei Han. 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*, pp. 995–1004, 2015.

Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30 (10):1825–1837, 2018a.

Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. Learning named entity tagger using domain-specific dictionary. *arXiv preprint arXiv:1809.03599*, 2018b.

Yuta Tsuboi, Hisashi Kashima, Shinsuke Mori, Hiroki Oda, and Yuji Matsumoto. Training conditional random fields using incomplete annotations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 897–904, 2008.

Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. Distantly supervised ner with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2159–2169, 2018.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150*, 2020.