Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning

Xinyu Wang[◇‡], **Yong Jiang**[†*], **Nguyen Bach**[†], **Tao Wang**[†], **Zhongqiang Huang**[†], **Fei Huang**[†], **Kewei Tu**[◇*]
[◇]School of Information Science and Technology, ShanghaiTech University
[◇]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[◇]University of Chinese Academy of Sciences
[†]DAMO Academy, Alibaba Group
{wangxy1,tukw}@shanghaitech.edu.cn, yongjiang.jy@alibaba-inc.com
{nguyen.bach,leeo.wangt,z.huang,f.huang}@alibaba-inc.com

Abstract

Recent advances in Named Entity Recognition (NER) show that document-level contexts can significantly improve model performance. In many application scenarios, however, such contexts are not available. In this paper, we propose to find external contexts of a sentence by retrieving and selecting a set of semantically relevant texts through a search engine, with the original sentence as the query. We find empirically that the contextual representations computed on the retrieval-based input view, constructed through the concatenation of a sentence and its external contexts, can achieve significantly improved performance compared to the original input view based only on the sentence. Furthermore, we can improve the model performance of both input views by Cooperative Learning, a training method that encourages the two input views to produce similar contextual representations or output label distributions. Experiments show that our approach can achieve new state-of-the-art performance on 8 NER data sets across 5 domains.

1 Introduction

Pretrained contextual embeddings such as ELMo (Peters et al., 2018), Flair (Akbik et al., 2018) and BERT (Devlin et al., 2019) have significantly improved the accuracy of Named Entity Recognition (NER) models. Recent work (Devlin et al., 2019; Yu et al., 2020; Yamada et al., 2020) found that including document-level contexts of the target sentence in the input of contextual embeddings methods can further boost the accuracy of NER models. However, there are a lot of application scenarios in which document-level contexts are unavailable in practice. For example, there are sometimes no available contexts in users' search queries, tweets and short comments in various domains such as social media and E-commerce domains. When professional annotators annotate ambiguous named entities in such cases, they usually rely on domain knowledge for disambiguation. This kind of knowledge can often be found through a search engine. Moreover, when the annotators are not sure about a certain entity, they are usually encouraged to find related knowledge through a search engine (Wang et al., 2019). Therefore, we believe that NER models can benefit from such a process as well.

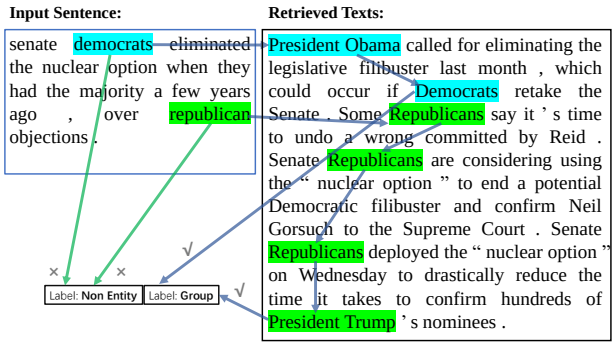

Figure 1: A motivating example from WNUT-17 dataset. The retrieved texts help the model to correctly predict the named entities of "democrats" and "republican".

In this paper, we propose to improve NER models by retrieving texts related to the input sentence by an off-the-shelf search engine. We re-rank the retrieved texts according to their semantic relevance to the input sentence and select several top-ranking texts as the external contexts. Consequently, we concatenate the input sentence and external contexts together as a new retrieval-based input view and feed it to the pretrained contextual embedding module, so that the resulting semantic representations of the input tokens can be improved. The token representations are then fed into a CRF layer

[*] Yong Jiang and Kewei Tu are the corresponding authors.
[‡]: This work was conducted when Xinyu Wang was interning at Alibaba DAMO Academy.

for named entity prediction. A motivating example is shown in Figure 1.

Moreover, we consider utilizing the new input view to improve model performance with the original input view that does not have external contexts. This can be useful in application scenarios when external contexts are unavailable or undesirable (e.g., in time-critical scenarios). To this end, we propose Cooperative Learning (CL) that encourages the two input views to produce similar predictions. We propose two approaches to CL which minimize either the L_2 distances between the token representations of the two input views or the Kullback–Leibler (KL) divergence between the prediction distributions of the two input views during training.

Our experiments show that including the retrieved external contexts can significantly improve the accuracy of NER models on 8 NER datasets from 5 domains. With CL, the accuracy of the NER models with both input views can be further improved. Our approaches outperform previous state-of-the-art approaches in each domain.

The contributions of this paper are:

1. We propose a simple and straight-forward way to improve the contextual representation of an input sentence through retrieving related texts using a search engine. We take the retrieved texts together with the input sentence as a new retrieval-based view.

2. We propose Cooperative Learning to jointly improve the accuracy of both input views in a unified model. We propose two approaches in CL based on the L_2 norm and KL divergence respectively. CL can utilize unlabeled data for further improvement.

3. We show the effectiveness of our approaches in several NER datasets across 5 domains and our approaches achieve state-of-the-art accuracy. By leveraging a large amount of unlabeled data, the performance can be further improved.

2 Framework

Given a sentence of n tokens $\boldsymbol{x} = \{x_1, \cdots, x_n\}$, the input sentence is fed into a search engine as a query. The search engine returns the top k relevant texts $\{\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_k\}$. Our framework feeds these texts into a re-ranking model. We concatenate l top-ranking texts output from the re-ranking model as the external contexts. The NER model is fed with either an input view with the input sentence (original input view) or a concatenation of the input sentence and external contexts (retrieval-based input view) as input. The model outputs the predictions of labels $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ at each position based on the CRF layer. To further improve the model, we use Cooperative Learning to train a unified model that is strong in both input views. With CL, the model is additionally constrained to be consistent in the internal representations or the output distributions of both input views. The architecture of our framework is shown in Figure 2.

2.1 Re-ranking

Given an input sentence as a search query, the search engine returns ranked relevant texts. However, the relevant texts may not be semantically similar to the search query. Since the NER task targets at semantically recognizing named entities, it is more helpful if the relevant texts are semantically similar to the input sentence. Therefore, we need to re-rank the the retrieved texts so that the most semantically relevant texts are chosen. We propose to apply BERTScore (Zhang et al., 2020) to score the relatedness of each retrieved text to the input sentence. BERTScore is a language generation metric that calculates a sum of cosine similarity between token representations of two sentences. Therefore, it is more likely that the search query and the retrieved texts have strong semantic relations when BERTScore is large. The token representations are generated from pretrained contextual embeddings such as BERT. Given the corresponding pre-normalized token representations $\{\mathbf{r}_1, \cdots, \mathbf{r}_n\}$ of the input sentence $\boldsymbol{x}$ and the pre-normalized token representations $\{\hat{\mathbf{r}}_1, \cdots, \hat{\mathbf{r}}_m\}$ of a certain retrieved text $\hat{\boldsymbol{x}}$ with m words, the Precision (P), Recall (R) of BERTScore measure the semantic similarities from one to another:

$$\mathrm{R} = \frac{1}{n}\sum_{x_i \in \boldsymbol{x}} \max_{\hat{x}_j \in \hat{\boldsymbol{x}}} \mathbf{r}_i^\top \hat{\mathbf{r}}_j; \quad \mathrm{P} = \frac{1}{m}\sum_{\hat{x}_j \in \hat{\boldsymbol{x}}} \max_{x_i \in \boldsymbol{x}} \mathbf{r}_i^\top \hat{\mathbf{r}}_j$$

We re-rank the retrieved texts by the F1 scores $\mathrm{F1}=2\frac{\mathrm{P}\cdot\mathrm{R}}{\mathrm{P}+\mathrm{R}}$ and concatenate l top-ranking texts $\{\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_l\}$ with F1 scores together as the external contexts:

$$\tilde{\boldsymbol{x}} = [sep\_token; \hat{\boldsymbol{x}}_1; \cdots; \hat{\boldsymbol{x}}_l]$$

where sep_token is a special token representing a separate of sentences in the transformer-based pretrained contextual embeddings (for example, "[SEP]" in BERT).

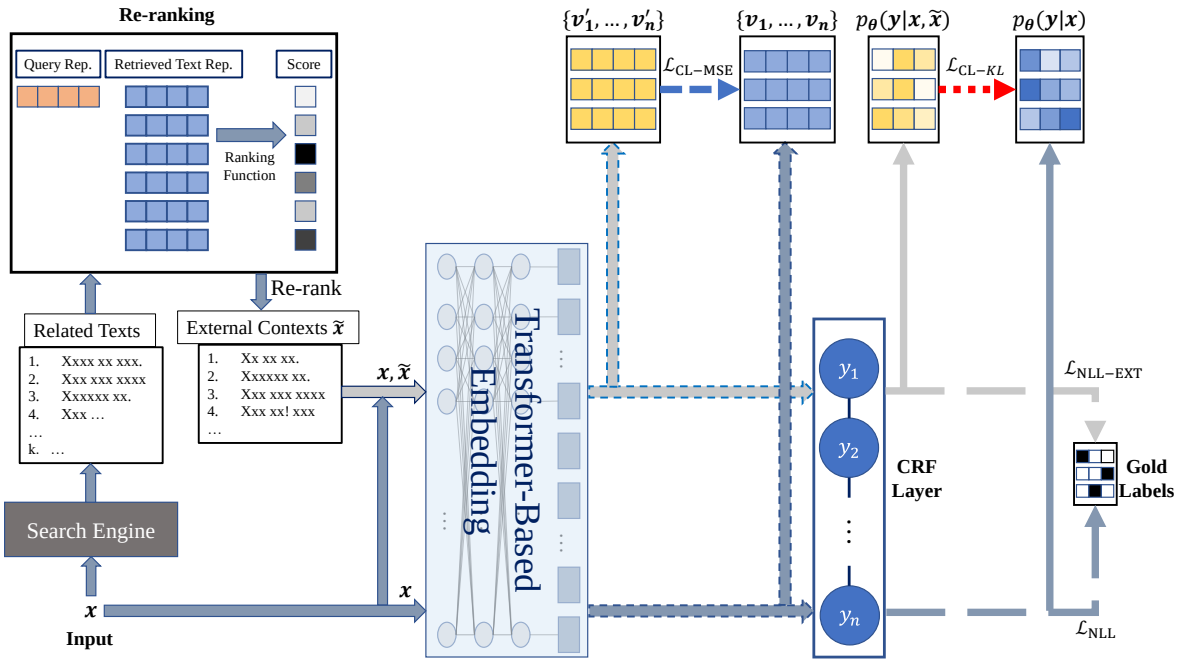

Figure 2: The architecture of our framework.

2.2 NER Model

We solve the NER task as a sequence labeling problem. We apply a neural model with a CRF layer, which is one of the most popular state-of-the-art approaches to the task (Lample et al., 2016; Ma and Hovy, 2016; Akbik et al., 2019). In the sequence labeling model, the input sentence $\boldsymbol{x}$ is fed into a transformer-based pretrained contextual embeddings model to get the token representations $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$ by $\boldsymbol{v}_i$=embed$_i(\boldsymbol{x})$. The token representations are fed into a CRF layer to get the conditional probability $p_\theta(\boldsymbol{y}|\boldsymbol{x})$:

$$\psi(y', y, \boldsymbol{v}_i) = \exp(\mathbf{W}_y^T \boldsymbol{v}_i + \mathbf{b}_{y',y}) \quad (1)$$

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{\prod_{i=1}^{n} \psi(y_{i-1}, y_i, \boldsymbol{v}_i)}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{x})} \prod_{i=1}^{n} \psi(y'_{i-1}, y'_i, \boldsymbol{v}_i)}$$

where ψ is the potential function and θ represents the model parameters. $\mathcal{Y}(\boldsymbol{x})$ denotes the set of all possible label sequences given $\boldsymbol{x}$. y_0 is defined to be a special start symbol. $\mathbf{W}^T \in \mathbb{R}^{t \times d}$ and $\mathbf{b} \in \mathbb{R}^{t \times t}$ are parameters computing emission and transition scores respectively. d is the hidden size of $\boldsymbol{v}$ and t is the size of the label set. During training, the negative log-likelihood loss for the input sequence with gold labels $\boldsymbol{y}^*$ is defined by:

$$\mathcal{L}_{\text{NLL}}(\theta) = -\log p_\theta(\boldsymbol{y}^*|\boldsymbol{x}) \quad (2)$$

In our approach, we concatenate the external contexts $\tilde{\boldsymbol{x}}$ at the end of the input sentence $\boldsymbol{x}$ to form the retrieval-based input view. The token representations are now given by:

$$\{\boldsymbol{v}'_1, \cdots, \boldsymbol{v}'_n, \cdots\} = \text{embed}([\boldsymbol{x}; \tilde{\boldsymbol{x}}])$$

The architecture of our NER model is shown in Figure 3. Now the conditional probability $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ becomes $p_\theta(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}})$. The loss function in Eq. 2 becomes:

$$\mathcal{L}_{\text{NLL-EXT}}(\theta) = -\log p_\theta(\boldsymbol{y}^*|\boldsymbol{x}, \tilde{\boldsymbol{x}}) \quad (3)$$

2.3 Cooperative Learning

In practice, there are two application scenarios for the NER model: 1) offline prediction, which requires high accuracy of the prediction but the prediction speed is less emphasized; 2) online serving, which requires a faster prediction speed. The retrieval-based input view meets the requirement of the first scenario for its strong token representations. However, it does not meet the requirement of the second scenario. The external contexts are usually significantly longer than the input sentence and a search engine may not meet the latency requirements. These two issues significantly slow down the prediction speed of the model. Therefore, it is essential to improve the accuracy of the original input views in a unified model to meet these two scenarios.

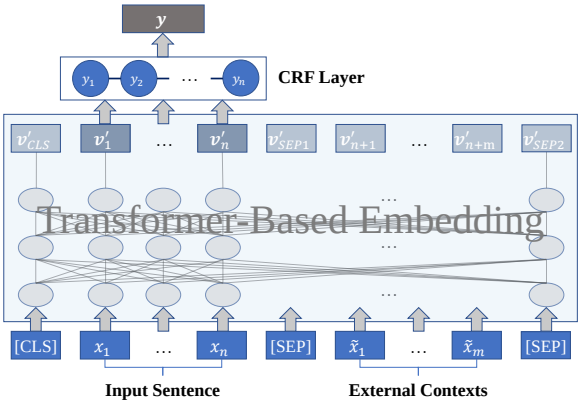

Figure 3: An illustration of our NER model architecture. "[CLS]" and "[SEP]" are an example of cls token and sep token in the embedding.

Cooperative Learning targets at using the retrieval-based input view to help improve the accuracy of the model when there are no external contexts available. CL adds constraints between the internal representations or the output distributions between two input views to enforce that the predictions of both views should be near. The objective function of CL is calculated by:

$$\mathcal{L}_{\text{CL}}(\theta) = D(h([\boldsymbol{x}; \tilde{\boldsymbol{x}}]), h([\boldsymbol{x}])) \quad (4)$$

where D is a distance function between a function h with different inputs. Because the representations or the distributions with retrieval-based input view are usually informative, we do not backpropagate the gradient through $h([\boldsymbol{x}; \tilde{\boldsymbol{x}}])$. We propose two approaches for CL.

Token Representations: Stronger token representations usually lead to better accuracy on the task. Therefore, CL constrains the token representations of two input views to be similar. This helps the model learn to predict the token representations with external contexts even if the contexts are not available. In this approach, D is the L_2 norm to represent the distances of the token representations:

$$\mathcal{L}_{\text{CL-}L_2}(\theta) = \sum_{i=1}^{n} ||\boldsymbol{v}'_i - \boldsymbol{v}_i||_2^2 \quad (5)$$

Label Distributions: Since CL enforces the label predictions of both input views to be similar, a straight-forward approach is constraining the label distributions predicted by the model to be similar with the two input views. In this approach, we use the KL divergence as the function D. Then objective function in Eq. 4 becomes the KL divergence between $p_\theta(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}})$ and $p_\theta(\boldsymbol{y}|\boldsymbol{x})$:

$$\mathcal{L}_{\text{CL-KL}}(\theta) = \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \text{KL}(p_\theta(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}}) || p_\theta(\boldsymbol{y}|\boldsymbol{x})) \quad (6)$$

With the CRF layer, the loss function is difficult to calculate because the output space of $p_\theta(\boldsymbol{y}|\bullet)$ is exponential in size. To alleviate this issue, we calculate the KL divergence between the marginal distributions $q_\theta(y_i|\boldsymbol{x}, \tilde{\boldsymbol{x}})$ and $q_\theta(y_i|\boldsymbol{x})$ at each position of the sentence to approximate Eq. 6. The marginal distributions can be obtained using the forward-backward algorithm:

$$\alpha(y_k) = \sum_{\{y_0, \ldots, y_{k-1}\}} \prod_{i=1}^{k} \psi(y_{i-1}, y_i, \boldsymbol{v}_i)$$

$$\beta(y_k) = \sum_{\{y_{k+1}, \ldots, y_n\}} \prod_{i=k+1}^{n} \psi(y_{i-1}, y_i, \boldsymbol{v}_i)$$

$$q_\theta(y_k|\boldsymbol{x}) \propto \alpha(y_k) \times \beta(y_k) \quad (7)$$

As mentioned earlier, we do not back-propagate the gradient through $p_\theta(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}})$. Therefore calculating the KL divergence is equivalent to calculating the cross-entropy loss between $q(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}})$ and $q(\boldsymbol{y}|\boldsymbol{x})$:

$$\mathcal{L}_{\text{CL-KL}}(\theta) = -\sum_{i=1}^{n} \sum_{y_i=1}^{t} q_\theta(y_i|\boldsymbol{x}, \tilde{\boldsymbol{x}}) \log q_\theta(y_i|\boldsymbol{x}) \quad (8)$$

Together with the negative log-likelihood losses in Eq. 2, 3, the total loss in training is a summation of label losses and a CL loss:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{NLL}}(\theta) + \mathcal{L}_{\text{NLL-EXT}}(\theta) + \mathcal{L}_{\text{CL}}(\theta) \quad (9)$$

where $\mathcal{L}_{\text{CL}}(\theta)$ can be one of the CL loss in Eq. 5, 8 or a summation of both of them.

3 Experiments

3.1 Settings

Datasets To show the effectiveness of our approach, we experiment on 8 NER datasets across 5 domains:

- **Social Media**: We use WNUT-16 (Strauss et al., 2016) and WNUT-17 (Derczynski et al., 2017) datasets collected from social media. We use the standard split for these datasets.
- **News**: We use CoNLL-03 English (Tjong Kim Sang and De Meulder, 2003) dataset and CoNLL++ (Wang et al., 2019) dataset. The

CoNLL-03 dataset is the most popular dataset for NER. CoNLL++ is a revision of the CoNLL-03 datasets. Wang et al. (2019) fixed annotation errors on the test set by professional annotators and improved the quality of the training data through their CrossWeigh approach. We use the standard dataset split for these datasets.

- **Biomedical**: We use BC5CDR (Li et al., 2016) and NCBI-disease (Doğan et al., 2014) datasets, which are two popular biomedical NER datasets. We merge the training and development data as training set following Nooralahzadeh et al. (2019).

- **Science and Technology**: We use CBS SciTech News dataset collected by Jia et al. (2019). The dataset only contains the test set with the same label set as the CoNLL-03 dataset. We use the dataset to evaluate the effectiveness of cross-domain transferability from the news domain.

- **E-commerce**: We collect and annotate an internal dataset from one anonymous E-commerce website. The dataset contains 26 named entity labels for goods in short texts. The dataset contains 38,959/5,000/5,000 sentences for training/development/test. We also collect 300,000 unlabeled sentences for semi-supervised training.

Retrieving and Ranking We use an internal E-commerce search engine for the E-commerce dataset. For the other datasets, we use Google Search as the search engine. Google Search is an off-the-shelf search engine and can simulate the offline search over various domains. We use summarized descriptions from the search results as the retrieved texts[1]. As Google Search limits the maximal length of searching queries to 32 words, we chunk a sentence into multiple sub-sentences based on punctuation if the sentence is longer than 30, feed each sub-sentence to the search engine, and retrieve up to 20 results. We filter the retrieved texts that contain any part of the datasets. Our re-ranking module selects top 6 sentences as the external contexts of the input sentence and chunk the external contexts if the total sub-token lengths of the input sentence and external contexts exceeds 510.

Model Configurations For the re-ranking module, we use Roberta-Large (Liu et al., 2019) for token representations which is the default configuration in the code[2] of BERTScore (Zhang et al., 2020). For token representations in the NER model, we use pretrained Bio-BERT (Lee et al., 2020) for datasets from the biomedical domain and use XLM-RoBERTa (Conneau et al., 2020) for datasets from other domains.

3.2 Results

We experiment on the following approaches:

- **LUKE** is a very recent state-of-the art model on CoNLL-03 NER dataset proposed by Yamada et al. (2020). We use the same parameter setting as Yamada et al. (2020) and use a single sentence as the input instead of taking document-level contexts in the dataset as in Yamada et al. (2020) for fair comparison.

- **W/O CONTEXT** represents training the NER model without external contexts (Eq. 2), which is the baseline of our approaches.

- **W/ CONTEXT** represents training the NER model with external contexts (Eq. 3).

- **CL-L_2** represents minimizing the L_2 distance between token representations. In evaluation, the NER model is evaluated by inputs without external contexts (**W/O CONTEXT**) and inputs with them (**W/ CONTEXT**) (also for **CL-KL**).

- **CL-KL** represents minimizing the KL divergence (Eq. 8) between CRF output distributions.

Besides, we also compare our approaches with previous state-of-the-art approaches over entity-level F1 scores[3]. We report the results averaged over 5 runs in our experiments. The results are listed in Table 1[4]. With the external contexts, our models with CL outperform previous state-of-the-art approaches on most of the datasets. Our approaches significantly outperform the baseline that is trained without external contexts with only one exception. Comparing with LUKE, our approaches and our

[1] If the descriptions are not available, we use the titles of the results instead.

[2] https://github.com/Tiiiger/bert_score

[3] We do not compare the results from previous work such as Yu et al. (2020); Luoma and Pyysalo (2020); Yamada et al. (2020) that utilizes the document-level contexts in CoNLL-03 NER here. We conduct a comparison with these approaches in Appendix.

[4] For the result of Bio-BERT (Lee et al., 2020) on NCBI-disease dataset, we report the results reported in official code (https://github.com/dmis-lab/biobert). The results (89.71 in NCBI-disease) reported in the paper used token-level F1 score instead of entity-level F1 score.

	Social Media		News		Biomedical		E-commerce
	WNUT-16	WNUT-17	CoNLL-03	CoNLL++	BC5CDR	NCBI	
Zhou et al. (2019)	55.43	42.83	-	-	-	-	-
Nguyen et al. (2020)	52.10	56.50	-	-	-	-	-
Nie et al. (2020)	55.01	50.36	-	-	-	-	-
Baevski et al. (2019)	-	-	93.50	-	-	-	-
Wang et al. (2019)	-	-	93.43	94.28	-	-	-
Li et al. (2020)	-	-	93.33	-	-	-	-
Nooralahzadeh et al. (2019)	-	-	-	-	89.93	-	-
Bio-Flair (2019)	-	-	-	-	89.42	88.85	-
Bio-BERT (2020)	-	-	-	-	-	87.70	-
Evaluation: **W/O CONTEXT**							
LUKE (2020)	54.04	55.22	92.42	93.99	89.18	87.62	77.64
W/O CONTEXT	56.04	57.86	93.03	94.20	90.52	88.65	81.47
CL-L_2	57.35†	58.68†	93.08	94.38†	90.70†	89.20†	82.43†
CL-KL	58.14†	59.33†	93.21†	94.55†	90.73†	**89.24**†	82.31†
Evaluation: **W/ CONTEXT**							
W/ CONTEXT	57.43†	60.20†	93.27†	94.56†	90.76†	89.01†	83.15†
CL-L_2	58.61†	60.26†	93.47†	94.62†	**90.99**†	89.22†	83.87†
CL-KL	**58.98**†	**60.45**†	**93.56**†	**94.81**†	90.93†	88.96†	**83.99**†

Table 1: A comparison among recent state-of-the-art models, the baseline and our approaches. † represents the model is significantly stronger than the baseline model (**W/O CONTEXT**) with $p < 0.05$ on Student's T test.

	Evaluation
	Science and Technology
Approach	**W/O CONTEXT** **W/ CONTEXT**
Jia et al. (2019)	73.59 -
W/O CONTEXT	75.87 75.74
W/ CONTEXT	75.72 75.94
CL-L_2	76.16 76.10
CL-KL	**76.37** **76.38**

Table 2: A comparison of different approaches transfer learning.

	Evaluation	
Approach	**W/O CONTEXT**	**W/ CONTEXT**
CL-L_2	82.43	83.87
CL-KL	82.31	83.99
CL–L_2+SEMI	**82.88**†	83.92
CL-KL+SEMI	82.58†	**84.10**

Table 3: A comparison between of CL approaches with and without semi-supervised learning. **SEMI** represents the approaches with semi-supervised learning. † represents the approach is significantly ($p < 0.05$) stronger than the approach without semi-supervised learning with the same input view.

baseline outperform LUKE in all the cases. The possible reason is that LUKE is pretrained only using long word sequences, which makes the model prone to fail to capture the information of entities based on short sentences[5]. For our approaches, with CL, the accuracy can be improved on both input views comparing with **W/O CONTEXT** and **W/ CONTEXT**, which shows adding constraints between two views during training can improve the accuracy of both views. For the two constraints in CL, we find that **CL-KL** is relatively stronger than **CL-L_2** in a majority of the cases.

3.3 Cross-Domain Transfer

For cross-domain transfer, we train the models on the CoNLL-03 datasets, evaluate the accuracy on the CBS SciTech News dataset, and compare the results with those in Jia et al. (2019). We evaluate our approaches with each input view and the results are shown in Table 2. Our approaches can improve the accuracy in cross-domain evaluation. The external contexts during evaluation can help to improve the accuracy of **W/ CONTEXT**. However, the gap between the two input views for the CL approaches is diminished. The observation shows that CL is able to improve the accuracy in cross-domain transfer for both views and eliminate the gap between the two views.

3.4 Semi-supervised Cooperative Learning

Cooperative learning can take advantage of large amounts of unlabeled text for further improvement. We jointly train on the labeled data and unlabeled data in training to form a semi-supervised training manner. During training, we alternate between minimizing the loss (Eq. 9) for labeled data and the CL loss for unlabeled data (Eq. 4). We conduct the experiment on the E-commerce dataset as an exam-

[5]We have confirmed with the authors of LUKE (Yamada et al., 2020) that the accuracy on the CoNLL-03 dataset is consistent with their experimental results.

	SE	FM	BS	BS+tf-idf
AVG.	59.95	59.54	**60.20**	59.71
BEST	61.79	60.89	**62.29**	60.96

Table 4: A comparison of different re-ranking approaches on WNUT-17. **SE**: Search engine. **FM**: Fuzzy match score. **BS**: BERTScore.

	WNUT-17
w/ Context (Ours)	**60.20**
w/o Context	57.86
w/ Context (Dataset)	57.21
w/ Context (Generated)	57.71
w/ Context (Random Retrieved)	57.53
w/ Context (Random Data)	47.69

Table 5: A comparison among different contexts types.

ple. Results in Table 3 show that the accuracy of both input views can be improved especially for the input without external contexts, which shows the effectiveness of CL in semi-supervised learning.

4 Analysis

We use the WNUT-17 dataset in the analysis.

4.1 Comparison of Re-ranking Approaches

Various re-ranking approaches may affect the token representations of the model. We compare our approach with three other re-ranking approaches. The first is the ranking from the search engine without any re-ranking approaches. The second is re-ranking through a fuzzy match score. The approach has been widely applied in a lot of previous work (Gu et al., 2018; Zhang et al., 2018; Hayati et al., 2018; Xu et al., 2020). The third is BERTScore with tf-idf importance weighting which makes rare words more indicative than common words in scoring. We train our models (**W/ CONTEXT**) with external contexts from these re-ranking approaches and report the averaged and best results on WNUT-17 in Table 4. Our results show that re-ranking with BERTScore performs the best, which shows the semantic relevance is helpful for the performance. However, for BERTScore with the tf-idf weighting, the accuracy of the model drops significantly (with $p < 0.05$). The possible reason might be that the tf-idf weighting gives high weights to irrelevant texts with rare words during re-ranking.

4.2 How the Context Quality Affects Accuracy

We analyze how the NER model will perform when the quality of external contexts varies. We train and evaluate the NER model in four conditions with various contexts. The first one takes each dataset split as a document and encodes each sentence with document-level contexts. In this case, we encode the document-level contexts following the approach of Yamada et al. (2020). The second one uses GPT-2 (Radford et al., 2019) to generate 6 relevant sentences as external contexts. The other two conditions randomly select from the retrieved texts or the dataset as external contexts. Results in Table 5 show that all these conditions result in inferior accuracy comparing with the model without any external context. However, our external contexts are more semantically relevant to the input sentence and helpful for prediction.

	Evaluation	
Approach	**W/O CONTEXT**	**W/ CONTEXT**
W/O CONTEXT	57.86	59.40
W/ CONTEXT	57.46	60.20
W/O CL	58.14	59.64
CL–L_2 + CL-KL	58.69	60.16
CL–L_2	58.68	60.26
CL-KL	**59.33**	**60.45**

Table 6: An ablation study of the training and prediction of models.

4.3 Ablation Study

To show the effectiveness of CL, we conduct three ablation studies for our approach. The first one is training the NER model based on one view and predict on the other. The second is jointly training both views without the CL loss term (removing $\mathcal{L}_{\text{CL}}(\theta)$ in Eq. 9). The final one is using both CL losses to train the model ($\mathcal{L}_{\text{CL}}(\theta) = \mathcal{L}_{\text{CL-}L_2}(\theta) + \mathcal{L}_{\text{CL-KL}}(\theta)$ in Eq. 9). Results in Table 6 show that the external context can help to improve the accuracy even when the NER model is trained without the contexts. However, when the model is trained with the external contexts, the accuracy of the model drops when predicting the inputs without external contexts. In joint training without CL, the accuracy of the model over inputs without contexts can be slightly improved but the accuracy over inputs with contexts drops, which shows the benefit of adding CL. For the model trained with both CL losses, we find no improvement over the models trained with a single CL loss.

5 Related Work

Named Entity Recognition Named Entity Recognition (Sundheim, 1995) has been studied

for decades. Most of the work takes NER as an sequence labeling problem and applies the linear-chain CRF (Lafferty et al., 2001) to achieve state-of-the-art accuracy (Ma and Hovy, 2016; Lample et al., 2016; Akbik et al., 2018, 2019). Recently, the improvement of accuracy mainly benefits from stronger token representations such as pretrained contextual embeddings such as BERT (Devlin et al., 2019), Flair (Akbik et al., 2018) and LUKE (Yamada et al., 2020). Very recent work (Yu et al., 2020; Yamada et al., 2020) utilizes the strength of pretrained contextual embeddings over long-range dependency and encodes the document-level contexts for token representations to achieve state-of-the-art accuracy on CoNLL 2002/2003 NER datasets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003).

Improving Models through Retrieval Retrieving related texts from a certain database (such as the training set) has been widely applied in tasks such as neural machine translation (Gu et al., 2018; Zhang et al., 2018; Xu et al., 2020), text generation (Weston et al., 2018; Kim et al., 2020), semantic parsing (Hashimoto et al., 2018; Guo et al., 2019). Most of the work uses the retrieved texts to guide the generation or refine the retrieved texts through the neural model, while we take the retrieved texts as the contexts of the input sentence to improve the semantic representations of the input tokens. For the re-ranking models, fuzzy match score (Gu et al., 2018; Zhang et al., 2018; Hayati et al., 2018; Xu et al., 2020), attention mechanisms (Cao et al., 2018; Cai et al., 2019), and dot products between sentence representations (Lewis et al., 2020; Xu et al., 2020) are usual scoring functions to re-rank the retrieved texts. Instead, we use BERTScore to re-rank the retrieved texts instead as BERTScore evaluates semantic correlations between the texts based on pretrained contextual embeddings.

Multi-View Learning Multi-View Learning is a technique applied to inputs that can be split into multiple subsets. Co-training (Blum and Mitchell, 1998) and co-regularization (Sindhwani and Niyogi, 2005) train a separate model for each view. These approaches are semi-supervised learning techniques that require two independent views of the data. The model with higher confidence is applied to construct additional labeled data by predicting on unlabeled data. Sun (2013) and Xu et al. (2013) have extensively studied various multi-view learning approaches. Recently, Clark et al. (2018) proposed Cross-View Training (CVT), which trains a unified model instead of multiple models and targets at minimizing the KL divergence between the probability distributions of the model and auxiliary prediction modules. Comparing with CVT, CL targets at improving the accuracy of two kinds of inputs rather than only one of them. We also propose to minimize the distance of token representations between different views in addition to KL-divergence. Besides, CL utilizes the external contexts and therefore we do not need to construct auxiliary prediction modules in the model. Moreover, CVT cannot be directly applied to our transformer-based embeddings. Finally, our decoding layer in the model uses the CRF layer instead of the simple Softmax layer as in CVT. The CRF layer is stronger but more difficult for KL-divergence computation.

Knowledge Distillation Knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015) transfers the knowledge of "teacher" models to smaller "student" models through minimizing the KL divergence of prediction probability distribution between the models. In speech recognition (Huang et al., 2018) and natural language processing (Wang et al., 2020), the marginal probability distribution of the linear-chain CRF layer has been applied to distill the knowledge between teacher models and student models. Comparing with these approaches, our approaches train a single unified model instead of transferring the knowledge between two models. We also show that the accuracy of both views can be improved with our approaches, unlike in knowledge distillation only the student model is updated and improved.

6 Conclusion

In this paper, we propose to improve the NER model's accuracy by retrieving related contexts from a search engine as external contexts of the inputs. To improve the robustness of the models when no external contexts are available, we propose Cooperative Learning. Cooperative Learning adds constraints between two input views over either the token representations or label distributions of both input views to be consistent. Empirical results show that our approach significantly outperforms the baseline models and previous state-of-the-art approaches on the datasets over 5 domains. We also show the effectiveness of Cooperative Learning in a semi-supervised training manner.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5360–5369, Hong Kong, China. Association for Computational Linguistics.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 535–541, New York, NY, USA. ACM.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1866–1875, Hong Kong, China. Association for Computational Linguistics.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2019. Coupling retrieval and meta-learning for context-dependent semantic parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 855–866, Florence, Italy. Association for Computational Linguistics.

Tatsunori B Hashimoto, Kelvin Guu, Yonatan Oren, and Percy Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10073–10083.

Shirley Anugrah Hayati, Raphael Olivier, Pravalika Avvaru, Pengcheng Yin, Anthony Tomasic, and Graham Neubig. 2018. Retrieval-based neural code generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 925–930, Brussels, Belgium. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Mingkun Huang, Yongbin You, Zhehuai Chen, Yanmin Qian, and Kai Yu. 2018. Knowledge distillation for sequence model. In *Proc. Interspeech 2018*, pages 3703–3707.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

Jihyeok Kim, Seungtaek Choi, Reinald Kim Amplayo, and Seung-won Hwang. 2020. Retrieval-augmented controllable review generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2284–2295, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jouni Luoma and Sampo Pyysalo. 2020. Exploring cross-sentence contexts for named entity recognition with BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 904–914, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1383–1391, Online. Association for Computational Linguistics.

Farhad Nooralahzadeh, Jan Tore Lønning, and Lilja Øvrelid. 2019. 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)*, pages 225–233, Hong Kong, China. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shreyas Sharma and Ron Daniel Jr. 2019. Bioflair: Pretrained pooled contextualized embeddings for biomedical sequence labeling tasks. *arXiv preprint arXiv:1908.05760*.

Vikas Sindhwani and Partha Niyogi. 2005. A co-regularized approach to semi-supervised learning with multiple views. In *Proceedings of the ICML Workshop on Learning with Multiple Views*. Citeseer.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.

Shiliang Sun. 2013. A survey of multi-view machine learning. *Neural computing and applications*, 23(7-8):2031–2038.

Beth M. Sundheim. 1995. Named entity task definition, version 2.1. In *Proceedings of the Sixth Message Understanding Conference*, pages 319–332.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, and Kewei Tu. 2020. Structure-level knowledge distillation for multilingual sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3317–3330, Online. Association for Computational Linguistics.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. CrossWeigh: Training named entity tagger from imperfect annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5154–5163, Hong Kong, China. Association for Computational Linguistics.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics.

Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Jitao Xu, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual adversarial neural transfer for low-resource named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3461–3471, Florence, Italy. Association for Computational Linguistics.

Approach	CoNLL-03
Yu et al. (2020)†	93.50
Yamada et al. (2020)	**94.30**
Luoma and Pyysalo (2020)†	93.74
W/ DOC CONTEXT	94.12
W/O CONTEXT	93.30
W/ CONTEXT	93.55
CL-L_2	93.68
CL-KL	93.85

Table 7: A comparison of retrieved contexts and document-level contexts. †: These approaches are trained on training and development sets.

A Detailed Settings

A.1 Annotations of E-commerce dataset

we manually labeled the NER labels for queries from a real-world E-commerce website. For each query, we asked one annotator to label the entities and ask another annotator to check the quality. After that, we randomly select 10% of the dataset and ask the third annotator to check the accuracy. As a result, the overall averaged query-level accuracy[6] is 95%.

A.2 Training

During training, we fine-tune the pretrained contextual embeddings by AdamW (Loshchilov and Hutter, 2018) optimizer with a batch size of 4. We use a learning rate of 5×10^{-6} to update the parameters in the pretrained contextual embeddings. For the CRF layer parameters, we use a learning rate of 0.05. We train the NER models for 10 epochs for the datasets in Social Media and Biomedical domains while we train the NER models for 5 epochs for other datasets for efficiency as these datasets have more training sentences.

B Retrieved Contexts Versus Document-level contexts on CoNLL-03

We conduct a comparison between our retrieved contexts and the document-level contexts on CoNLL-03 datasets. In Table 7, we report the best model on development set following Yamada et al. (2020). Comparing with previous state-of-the-art approaches with encoding document-level contexts, our approaches are competitive and even stronger than some of the previous approaches utilizing maximal document-level contexts. Comparing with our model trained on document-level contexts (**W/ DOC CONTEXT**), we find that there is still a gap between the document-level contexts and retrieved contexts but our CL approaches can reduce the gap between these two contexts.

[6] the accuracy of a query counts 1.0 if all the entities in the query are correctly recognized and 0.0 otherwise.