Domain-Specific NER via Retrieving Correlated Samples

Xin Zhang\textsuperscript{1}, Yong Jiang, Xiaobin Wang, Xuming Hu\textsuperscript{2}, Yuecheng Sun\textsuperscript{3}, Pengjun Xie, Meishan Zhang\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1}School of New Media and Communication, Tianjin University, China
\textsuperscript{2}School of Software, Tsinghua University, China
\textsuperscript{3}College of Intelligence and Computing, Tianjin University, China

hsinz@tju.edu.cn, hxm19@mails.tsinghua.edu.cn
jiangyong.ml@gmail.com, czwangxiaobin@foxmail.com
yhs@tju.edu.cn, xpjandy@gmail.com, zhangmeishan@hit.edu.cn

Abstract

Successful Machine Learning based Named Entity Recognition models could fail on texts from some special domains, for instance, Chinese addresses and e-commerce titles, where requires adequate background knowledge. Such texts are also difficult for human annotators. In fact, we can obtain some potentially helpful information from correlated texts, which have some common entities, to help the text understanding. Then, one can easily reason out the correct answer by referencing correlated samples. In this paper, we suggest enhancing NER models with correlated samples. We draw correlated samples by the sparse BM25 retriever from large-scale in-domain unlabeled data. To explicitly simulate the human reasoning process, we perform a training-free entity type calibration by majority voting. To capture correlation features in the training stage, we suggest to model correlated samples by the transformer-based multi-instance cross-encoder. Empirical results on datasets of the above two domains show the efficacy of our methods.

1 Introduction

Named Entity Recognition (NER), which first locates entity positions and then labels their types sequentially, is a fundamental topic in both academia and industry (Li et al., 2022). Normal NER models consider the input samples to be independent of each other, learning the common intra-instance patterns and making predictions in a sequential way. This paradigm has shown surprising successes in decades, especially with the help of emerging deep learning (Shang et al., 2018; Zhang et al., 2018b; Liu et al., 2019; Luo et al., 2020; Lison et al., 2020; Fang et al., 2021; Meng et al., 2021).

However, learned models will fail at some hard cases, which would be inevitably encountered in

\begin{itemize}
\item Successful Machine Learning based Named Entity Recognition models could fail on texts from some special domains, for instance, Chinese addresses and e-commerce titles, where requires adequate background knowledge.
\item Such texts are also difficult for human annotators. In fact, we can obtain some potentially helpful information from correlated texts, which have some common entities, to help the text understanding.
\item Then, one can easily reason out the correct answer by referencing correlated samples.
\item In this paper, we suggest enhancing NER models with correlated samples. We draw correlated samples by the sparse BM25 retriever from large-scale in-domain unlabeled data.
\item To explicitly simulate the human reasoning process, we perform a training-free entity type calibration by majority voting.
\item To capture correlation features in the training stage, we suggest to model correlated samples by the transformer-based multi-instance cross-encoder. Empirical results on datasets of the above two domains show the efficacy of our methods.
\end{itemize}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example.png}
\caption{An address example with retrieved texts. The model incorrectly predicted “白城镇 (Baicheng Town)” and “赉县火车站 (Lai County Train Station)” because they match common patterns, i.e. “XX县” (Xx County) and “X县火车站” (X County Train Station). “吉林” is ambiguous, which is both a province and a city.}
\end{figure}

real scenarios (Li et al., 2019; Ding et al., 2019).

Figure 1 shows an example of the Chinese address domain. This kind of bad cases can not be easily solved by annotating more relevant training data\textsuperscript{1}. For human annotators, this case is ambiguous as well if no extra information is given, for instance, we can not distinguish the type of “吉林” (Jilin) without affixes “省” (Province) or “市” (City). This demonstrates that obtaining background knowledge and information is crucial to the text understanding.

Learning from correlated or nested data is mainly studied in Machine Learning and Computer Vision (Dundar et al., 2007; Choi and Won, 2019; Choi et al., 2021). Images in sub-groups naturally show a high degree of correlation on both features and labels, and come with nested structures (Dundar et al., 2007; Choi and Won, 2019), such as different

\textsuperscript{1}Because this pattern is indeed correct in most cases. This problem also exists in models with internal larger datasets.
regions of interest could be drawn from the same objects. In the address and e-commerce domain, texts are also highly correlated in nature. For example, two addresses may belong to the same city or refer to the same location, e-commerce product titles could come from the same brand, or they are just the same product. In Figure 1, with correlated texts, annotators can infer that the “吉林” and “白城” are short forms of “吉林省 (Jilin Province)” and “白城市 (Baicheng City)”. Hence, we argue that correlated samples could offer sufficient disambiguation information for NER models as well. Such kind of inductive bias is seldom considered in previous NLP studies.

In this work, we propose to enhance NER models by modeling and inferencing with the correlated samples. We first draw the correlated samples from in-domain large-scale unlabeled data by the retrieval engine (Elasticsearch, 2022).\(^2\) Then, we suggest two methods: (1) we perform an entity type calibrating by parallelly predicting the input text and all retrieved samples by the off-the-shelf NER model, and then aggregating the final labels by majority voting; (2) we propose to model the correlations by transformers via multi-instance cross-encoders to enhance the NER feature vectors.

To evaluate our methods, we conduct experiments on two open-access datasets (Inc., 2022; Ding et al., 2019) of the aforementioned two domains. We implement our methods based on a strong BiLSTM-CRF model with NEZHA (Wei et al., 2019) representation. Empirical results show that our methods outperform all baselines, and achieve promising results in the simulated low-resource setting. Finally, we present several analyses to understand our methods comprehensively.

\(^2\)Recently, Wang et al. (2021) and Geng et al. (2022) studied retrieving external contexts from Google or Baidu for standard NER datasets, which is quite different from our idea of modeling correlated samples for specific domains.

2 Approach

2.1 Unlabeled Data Retrieval

In the address (resp. e-commerce) domain, some texts naturally possess entity co-reference relations, for instance, they may belong to the same city (resp. brand) or represent the same location (resp. product). We call such texts, which usually have entities with the same semantic but different expressions, correlated samples. Since these texts are highly structured and of limited vocabulary, showing a high degree of lexical overlap. We could draw correlated samples for a given text by taking it as a query and retrieving the domain-specific database with text similarity measurements.

We implement an efficient BM25 (Robertson and Walker, 1994) retriever by an off-the-shelf retrieval engine (Elasticsearch, 2022). For a cleaned large-scale in-domain unlabeled corpus, we create the Elasticsearch index by the build-in standard analyzer. Then, we can retrieve top-K samples by BM25 scores of an input text in nearly real-time.

2.2 Entity Type Calibrating

As shown in Figure 1, correlated sample can help the entity disambiguation. If this kind of entity appears in correlated samples, human annotators can decide its type by referring to answers of correlated samples. For NER models, we suggest achieving this process by entity-level (or span-level) majority voting. Concretely, we first use a model (e.g., baseline) to extract entities of the input text and each correlated sample parallelly, and then re-assign labels of shared entities by majority voting.

2.3 Correlation Modeling

To further capture sample correlations in the training time, we suggest modeling correlated samples by the cross-encoder (Reimers and Gurevych, 2019), letting transformers learn complex corre-
<table>
<thead>
<tr>
<th>Method</th>
<th>Chinese Address</th>
<th>E-commerce</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Micro P R F1</td>
<td>Macro P R F1</td>
</tr>
<tr>
<td>Human</td>
<td>93.04 92.01 92.52</td>
<td>87.83 84.52 86.14</td>
</tr>
<tr>
<td>BC</td>
<td>85.56 83.90 84.72</td>
<td>82.20 76.48 79.24</td>
</tr>
<tr>
<td>NEZHA-BC</td>
<td>91.29 90.62 90.95</td>
<td>86.41 84.68 85.53</td>
</tr>
<tr>
<td>Entity-Voting†</td>
<td>91.67 91.00 91.34</td>
<td>86.70 84.83 85.76</td>
</tr>
<tr>
<td>Cross-Encoder†</td>
<td><strong>92.41</strong> 91.95 <strong>92.18</strong></td>
<td><strong>87.25</strong> 85.71 <strong>86.48</strong></td>
</tr>
<tr>
<td>Self-Training</td>
<td>91.57 91.02 91.29</td>
<td>86.65 85.37 86.01</td>
</tr>
<tr>
<td>Biaffine</td>
<td>91.35 90.25 90.80</td>
<td>86.32 84.59 85.45</td>
</tr>
<tr>
<td>Seq2set</td>
<td>89.43 87.69 88.55</td>
<td>83.89 80.12 81.96</td>
</tr>
<tr>
<td>Locate&amp;Label</td>
<td>90.28 87.76 89.00</td>
<td>85.95 82.29 84.08</td>
</tr>
<tr>
<td>PIQN</td>
<td>90.27 87.83 89.03</td>
<td>86.04 80.28 83.06</td>
</tr>
</tbody>
</table>

Table 1: Main results. † means statistically significant.

As shown in Table 1, our training-free calibrating strategy, NER models could benefit from the contrastive view between multiple correlated samples and understand the query instance better.

3 Experiments

3.1 Settings

Datasets. For the **Chinese Address** domain, we use the recently published dataset from CCKS competition (Inc., 2022). It is annotated by 21 classes of address elements and contains 8856, 1970, 4000 addresses for train, dev, and test sets. For the **E-commerce** domain, we use the dataset released by Ding et al. (2019). It is collected from e-commerce product titles and annotated by PROD (product) and BRAN (brand) types. It has 3983, 499, 498 sentences3 for train, dev, and test sets. For our **retrieval-based** methods, we process and index our internal in-domain unlabeled data with Elasticsearch, obtaining 400M and 600M samples for the address and e-commerce domain, respectively.

Evaluation. We employ entity-level exact precision, recall, and F1-measure and report both micro and macro aggregations. All experiments of the same setting are conducted by 8 different random seeds. We test the best model of the devset, and the average scores are reported. We regard a result as statistically significant when the p-value is below 0.05 by the paired t-test with baseline NEZHA-BC.

Implementation. We choose the BiLSTM-CRF (Lample et al., 2016) to achieve NER task, and use NEZHA-base (Wei et al., 2019) as the embedding module. The BiLSTM hidden size is set to 384 for each direction. We apply the dropout (Srivastava et al., 2014) with probabilities 0.5 and 0.2 to NEZHA embeddings for address and e-commerce, and 0.2 to BiLSTM features. We set the batch size to 32 and use the AdamW (Loshchilov and Hutter, 2017) optimizer with a constant lr 1e-3 and 1e-5 to update BiLSTM-CRF and NEZHA parameters.

For the entity type calibrating, we use the top 100 and 50 retrieved samples for address and e-commerce, respectively. For the correlating modeling, we limit the max sample number to 12 and the max sequence length to 256.

Baselines. We denote the BiLSTM-CRF with random character embedding (resp. NEZHA) by BC (resp. NEZHA-BC). We implement several state-of-the-art methods, i.e., **Biaffine** (Yu et al., 2020), **Seq2set** (Tan et al., 2021), **Locate&Label** (Shen et al., 2021), **PIQN** (Shen et al., 2022). We also implement **Self-Training** based on NEZHA-BC and the unlabeled data of the same size as our cross-encoder. We include e-commerce results from Mengge et al. (2020) for comparison.

3.2 Main Results

As shown in Table 1, our training-free calibrating method consistently outperforms our implemented baselines on both datasets, which verifies our intuition that modeling the correlation between samples is important in processing domain-specific texts. By leveraging the retrieved samples in the training stage (Cross-Encoder), our approaches gain a significant performance boost. This indicates that these retrieved samples not only provide extended entity information (such as 白城→白城市), but also supply sufficient disambiguate sig-
Table 2: Test F1 scores at various low-resource settings.

<table>
<thead>
<tr>
<th># retrieved sample</th>
<th>Micro F1</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>92.18</td>
<td>92.07</td>
</tr>
<tr>
<td>2</td>
<td>92.21</td>
<td>92.02</td>
</tr>
<tr>
<td>4</td>
<td>92.25</td>
<td>92.08</td>
</tr>
<tr>
<td>6</td>
<td>92.28</td>
<td>92.11</td>
</tr>
<tr>
<td>8</td>
<td>92.30</td>
<td>92.16</td>
</tr>
<tr>
<td>10</td>
<td>91.80</td>
<td>91.84</td>
</tr>
</tbody>
</table>

Table 3: Test F1 scores of our Cross-encoder in various sizes of unlabeled data for retrieval in address domain.

<table>
<thead>
<tr>
<th># Address</th>
<th>Micro F1</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>400M</td>
<td>90.95</td>
<td>88.79</td>
</tr>
<tr>
<td>40M</td>
<td>91.04</td>
<td>87.57</td>
</tr>
<tr>
<td>10M</td>
<td>86.56</td>
<td>84.69</td>
</tr>
<tr>
<td>4M</td>
<td>89.13</td>
<td>86.82</td>
</tr>
<tr>
<td>400k</td>
<td>91.31</td>
<td>91.16</td>
</tr>
<tr>
<td>100k</td>
<td>91.80</td>
<td>91.16</td>
</tr>
</tbody>
</table>

Figure 4: Test F1 scores of incorporating different correlated sample num by our cross-encoder.

3.3 Analysis

We conduct fine-grained analyses of cross-encoder.

Different Sizes of Labeled Data. Our idea essentially introduces extra in-domain data to the predictive models. Hence we can suppose that our methods will achieve larger improvements in the low-resource scenario. To verify this, we train the baseline and our cross-encoder in simulated smaller trainsets, which are sampled from the original trainset by different proportions. Table 2 demonstrates the test f1 scores of these two models in different settings. We can roughly say that the score difference increases as the sampling ratio decreases, which is in line with our intuition.

Different Sizes of Correlated Samples. In the above experiments, we limit the max sequence length of our cross-encoder to 256 for efficiency. Here we relax this constraint to investigate the influence of encoded sample num (from 0\(^2\) to top 10 retrieved texts) in cross-encoder on both two domains. As shown in Figure 4, the performance increment is significant at the lower sample number. And adding more relatively low-ranking samples is of limited gains.

Different Sizes of Unlabeled Data. All of the previous experiments are based on the same large-scale in-domain unlabeled data, which almost reach the billion-level (400M and 600M samples for address and e-commerce, respectively). We also sample several smaller unlabeled corpus (i.e., 40M, 4M, 400k, 100k).

\(^2\)The 0 samples cross-encoder degrade to the baseline NEZHA-BC.
<table>
<thead>
<tr>
<th>Method</th>
<th>NEZHA-BC</th>
<th>Entity-Voting</th>
<th>Cross-Encoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seconds</td>
<td>14.73</td>
<td>500+</td>
<td>38.41</td>
</tr>
</tbody>
</table>

Table 4: Running times of different methods on the address domain testset, which has 4,000 texts.

10M, 4M, 400k, 100k) and re-train our cross-encoder. As shown in Table 3, with the size of unlabeled data declines, the retrieved samples are less relevant, the improvements of our model are lower. Interestingly, this experiments also could reflect the effect of unlabeled data quality to the performance of our cross-encoder. The higher the quality of the data, the more correlated samples can be retrieved. The behavior of low-quality unlabeled data is similar to the small size data.

Running Speed of Different Methods. Another key concern of our methods is the running speed. The entity-voting needs parallelly decode dozens of texts, and the cross-encoder will significantly enlarge the text length. We measured the running time of several methods on the testset of the address domain dataset. As demonstrated in Table 4, the entity-voting is truely slower than other methods in an order of magnitude. But the cross-encoder just took about twice as long as the baseline NEZHA-BC. This is because the most time-consuming part is the CRF, where the concatenated samples are dropped before the CRF. So it can avoid the redundant decoding in the entity-voting, and has a higher running speed. Besides, the forward of pretrained language models are highly optimized.

3.4 Discussion

Retrieval-augmented models are showing state-of-the-art performance in many NLP tasks, such as Dialogue (Weston et al., 2018), Neural Machine Translation (Zhang et al., 2018a), Question Answering (Izacard and Grave, 2021), and Language Modeling (Gu et al., 2020; Yao et al., 2022; Borgeaud et al., 2022). Our work aims to model the internal correlation within sub-groups of samples. We first retrieve correlated sample groups for a given input by the off-the-shelf Elasticsearch engine. Then, we propose painlessly calibrating entity type by cross-instance entity voting. To further incorporate these correlated samples into model training, we use multi-instance cross-encoders to learn more complex correlations. Empirical results show that the painless entity type calibrating improved the performance to some extent, and modeling correlations by cross-encoders achieved the state-of-the-art performance. We hope this idea could benefit the similar scenario/domains of other tasks.

We will release our code and data at github.com/izhx/NER-unlabeled-data-retrieval to facilitate future research.

Acknowledgements

We thank all reviewers for their hard work. This research is supported by grants from the National Natural Science Foundation of China (No. 62176180).

Ethical Statement

All texts are anonymized.

References


Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. **Unified structure generation for universal information extraction.** In *Proc. of the ACL*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.


Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proc. of the EMNLP*, pages 10367–10378, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.


Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. **Learning named entity tagger using domain-specific dictionary.** In *Proc. of the EMNLP*, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.


Qi Zhang, Jilin Fu, Xiaoyu Liu, and Xuanjing Huang, 2018b. **Adaptive co-attention network for named entity recognition in tweets.** In *Proc. of the AAAI*, pages 5674–5681. AAAI Press.