Improving Low-resource Named Entity Recognition with Graph Propagated Data Augmentation

Jiong Cai, Shen Huang, Yong Jiang, Zeqi Tan, Pengjun Xie, Kewei Tu

School of Information Science and Technology, ShanghaiTech University
Shanghai Engineering Research Center of Intelligent Vision and Imaging
Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
College of Computer Science and Technology, Zhejiang University
DAMO Academy, Alibaba Group

Abstract

Data augmentation is an effective solution to improve model performance and robustness for low-resource named entity recognition (NER). However, synthetic data often suffer from poor diversity, which leads to performance limitations. In this paper, we propose a novel Graph Propagated Data Augmentation (GPDA) framework for Named Entity Recognition (NER), leveraging graph propagation to build relationships between labeled data and unlabeled natural texts. By projecting the annotations from the labeled text to the unlabeled text, the unlabeled texts are partially labeled, which has more diversity rather than synthetic annotated data. To strengthen the propagation precision, a simple search engine built on Wikipedia is utilized to fetch related texts of labeled data and to propagate the entity labels to them in the light of the anchor links. Besides, we construct and perform experiments on a real-world low-resource dataset of the E-commerce domain, which will be publicly available to facilitate the low-resource NER research. Experimental results show that GPDA presents substantial improvements over previous data augmentation methods on multiple low-resource NER datasets.

1 Introduction

Data augmentation is an effective solution to improve model performance and robustness, and is especially useful when the labeled data is scarce. In computer vision and speech, simple hand-crafted manipulations (Zhong et al., 2020; Zhang et al., 2018) are widely used to generate synthetic data that preserve the original information. However, when applied to natural language processing (NLP), it is challenging to edit a sentence without changing its syntax or semantics.

There are two successful attempts of applying data augmentation on sentence-level NLP tasks. One is manipulating a few words in the original sentence, which can be based on synonym replacement (Zhang et al., 2015; Kobayashi, 2018; Wu et al., 2019; Wei and Zou, 2019), random insertion or deletion (Wei and Zou, 2019), random swap (Sahin and Steedman, 2018; Wei and Zou, 2019; Min et al., 2020). The other is generating the whole sentence with the help of back-translation (Yu et al., 2018; Dong et al., 2017; Iyyer et al., 2018), sequence to sequence models (Kurata et al., 2016; Hou et al., 2018) or pre-trained language models (Kumar et al., 2020). However, when applied to token-level tasks such as NER, these methods suffer heavily from token-label misalignment or erroneous label propagation.

To overcome the issue of token-label misalignment, Dai and Adel (2020) extend the replacement from token-level to entity-level with entities of the same class, which proves to be a simple but strong augmentation method for NER. Li et al. (2020) adopt a seq2seq model to conditionally generate contexts while leaving entities / aspect terms unchanged. Ding et al. (2020) exploit an auto-regressive language model to annotate entities while treating NER as a text tagging task. Zhou et al. (2022) utilize labeled sequence linearization to enable masked entity language model to explicitly condition on label information when predicting masked entity tokens. Still, these methods generate synthetic data, which inevitably introduces incoherence, semantic errors and lacking in diversity.

In this work, we investigate data augmentation with natural texts instead of synthetic ones. We are inspired by the fact that professional annotators usually understand the semantics of an entity through its rich context. However, in low-resource
NER, the semantic information of a specific entity is relatively limited due to fewer annotations. To this end, we propose to improve the NER models by mining richer contexts for the existing labeled entities. More particularly, we propose a Graph Propagation based Data Augmentation (GPDA) framework for NER, leveraging graph propagation to build relationships between labeled data and unlabeled natural texts. The unlabeled texts are accurately and partially labeled according to their connected labeled data, which has more diversity rather than synthetic hand-crafted annotations. Furthermore, not restricted to the existing annotated entities in the training data, we explore external entities from the unlabeled text by leveraging consistency-restricted self-training.

The contributions of GPDA can be concluded:

- We propose a data augmentation framework that utilizes graph propagation with natural texts for augmentation, which is rarely investigated in previous work (Section 2);
- We utilize a simple Wikipedia-based search engine to build the graph with two retrieval methods (Section 2.2);
- With consistency-restricted self-training, we further make the most efficient utilization of externally explored unlabeled text (Section 2.3);
- By conducting experiments on both public datasets and a real-world multilingual low-resource dataset, GPDA achieves substantial improvements over previous data augmentation methods (Section 3).

2 Method

Fig. 1 presents the workflow of our proposed data augmentation framework. First, we build a graph between labeled data nodes and unlabeled text nodes according to their textual similarity. Then, the entity annotations are propagated to obtain augmented data. Finally, the marginalized likelihood for conditional random field (CRF) (Tsuboi et al., 2008) is applied during the training phase as the augmented data are partially labeled. Moreover, we adopt the consistency-restricted self-training strategy to further improve the model performance.

2.1 NER with Pure Labeled Data

We take NER as a sequence labeling problem, which predicts a label sequence $y = \{y_1, \ldots, y_n\}$ at each position for the input tokens $x = \{x_1, \ldots, x_n\}$, where $\mathcal{Y}$ denotes the label set. The sequence labeling model feeds the input $x$ into a transformer-based encoder (such as BERT (Devlin et al., 2019)) which creates contextualized embeddings $r_i$ for each token. Then a linear-chain CRF layer that captures dependencies between neighboring labels is applied to predict the probability distribution:

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

Unified Training Objective Instead of directly minimizing the negative log-likelihood, we unify the training objectives in Section 2.1, 2.2 and 2.3. Specifically, we compute the marginal probability of each token $P_\theta(y_i|x)$ with the forward-backward algorithm.

$$\alpha(y_i) = \sum_{\{y_0 \ldots y_{i-1}\}} \prod_{k=1}^i \psi(y_{k-1}, y_k, r_k)$$
$$\beta(y_i) = \sum_{\{y_{i+1} \ldots y_n\}} \prod_{k=i+1}^n \psi(y_{k-1}, y_k, r_k)$$
$$P_\theta(y_i|x) \propto \alpha(y_i) \times \beta(y_i)$$

The marginal distributions can be computed efficiently. Given a partially annotated label sequence $y^* = \{*, \ldots, y_i, \ldots, *\}$ that * denotes the label that is not observed, we can obtain the probability.

$$Q_\theta(y^*|x) = \prod_{i=1}^n Q_\theta(y_i|x)$$
where \( Q_\theta(y_i|x) \) is defined as \( P_\theta(y_i|x) \) if \( y_i \) is observed, otherwise \( Q_\theta(y_i|x) = 1 \).

The final model parameters can be optimized by minimizing the following objective:

\[
\mathcal{L}(\theta) = -\log Q_\theta(y^*|x)
\]

For the pure labeled data \( D = \{(x^{(i)}, y^{(i)})\}_{i=1}^N \), we direct set \( y^* = y_i \) and obtain the loss function:

\[
\mathcal{L}(\theta) = -\sum_{(x^{(i)}, y^{(i)}) \in D} \log Q_\theta(y^* = y^{(i)}|x^{(i)})
\]

### 2.2 NER with Propagated Unlabeled Data

**Building Propagating Graph** Compared to labeled data, large-scale unlabeled natural texts can be acquired much more easily. We attempt to utilize these natural texts for augmentation by building a graph between the labeled data nodes and the unlabeled text nodes according to their textual similarity. Given a labeled sample \( (x^{(i)}, y^{(i)}) \), we retrieve its corresponding augmented sentences \( \{x^{(i,j)}\}_{j=1}^m \) via a search engine. For common NER datasets, the search engine is built on the Wikipedia corpus with one of the two methods we explore: sparse retrieval based on BM25\(^2\) or dense retrieval\(^3\) based on L2 similarity. The top related sentences will be treated connected to the original labeled sentence in the graph.

\(^2\)Sparse retrieval is implemented with Elastic Search
\(^3\)Dense retrieval is implemented with ColBERT

**Label Propagation** While building the graph, label propagation is conducted from labeled data \( (x^{(i)}, y^{(i)}) \) to unlabeled data \( \{x^{(i,j)}\}_{j=1}^m \) to generate partially annotated \( \{(x^{(i,j)}, y^{(i,j)})\}_{j=1}^m \). To strengthen the precision, propagation will not happen unless the anchor text in Wikipedia matches the labeled entity. By graph propagation, we obtain the augmented data \( D' = \{(x^{(i)}, y^{(i)})\}_{i=1}^M \) sharing the same entities but with more diverse contexts. Along with the original labeled data \( D \), we train the NER model following the same objective in Section 2.1:

\[
\mathcal{L}(\theta) = -\sum_{(x^{(i)}, y^{(i)}) \in D \cup D'} \log Q_\theta(y^* = y^{(i)}|x^{(i)})
\]

### 2.3 NER with Explored Entity Annotations

To make the most efficient utilization of the explored annotations in \( D' \), we adopt consistency-restricted self-training. A well-trained model from Section 2.2 will be utilized to re-annotate the partially labeled augmented data under consistency restriction. Particularly, an augmented sample \( (x^{(i)}, y^{(i)}) \) will be re-annotated to \( (x^{(i)}, y^{(i)}) \). Now we have \( \tilde{D} = \{(x^{(i)}, \tilde{y}^{(i)})\}_{i,j=1}^M \). Along with the original labeled data \( D \), we train a better NER model following the objective in Section 2.1:

\[
\mathcal{L}(\theta) = -\sum_{(x^{(i)}, \tilde{y}^{(i)}) \in D \cup \tilde{D}} \log Q_\theta(y^* = y^{(i)}|x^{(i)})
\]
3 Experiments

3.1 Dataset

We conduct experiments on the CrossNER (Liu et al., 2020) dataset of 5 genres (AI, Literature, Music, Politics, Science) and an anonymous multi-lingual E-commerce query NER dataset (Ecom) consisting of 3 languages (English, Spanish, French) Detailed statistics about these two datasets is provided in the Table 2.

For CrossNER, the search engine is manually built on the Wikipedia corpus. While for Ecom, an off-the-shelf E-commerce search engine is utilized to build the augmentation graph.

3.2 Results and Analysis

Low-resource NER Tasks As illustrated in Table 1, the proposed GPDA consistently achieves the best F1 scores across the five genres of CrossNER and gains an average improvement of 2.2% over the baseline BERT-CRF model. It also outperforms other data augmentation methods, demonstrating its effectiveness on multi-domain low-resource NER.

Furthermore, GPDA with Explored Entity Annotation (EEA) strategy achieves 1.1% higher F1 than GPDA without EEA, suggesting that it is also crucial to extend unique entities rather then only diversifying entity contexts in data augmentation.

It can be noticed that GPDA with dense retrieval performs worse than with sparse retrieval, which is not intuitive. This may be attributed to dense retrieval requires careful supervised training in the target domain, but our pre-trained matching model is not finetuned. We will leave this part for future work.

Real-world Low-resource NER Scenarios Table 3 shows the F1 results on three languages from the real-world Ecom dataset. The augmented data generated by GPDA improves model performances for multilingual NER. For specific domain datasets where high-quality knowledge or texts can be fetched easily, GPDA are indeed helpful.

Size of Gold Samples We study the impact of GDPA on different size of gold samples in Fig. 2. On the low-resource settings where 10%-25% gold samples are available, the improvement is striking which outperforms the baseline model by at most 37%.

4 Discussion

Retrieving relevant texts from databases has been widely used in NLP tasks. RaNER (Wang et al., 2021) retrieves context using a search system to enhance the token representation for NER tasks. To help entity disambiguation in domain-specific NER, Zhang et al. (2022) retrieves the domain-specific database to find the correlated sample. In order to leverage the extensive information about entities in Wikipedia and Wikidata, Wang et al. (2022) and Tan et al. (2023) construct databases and retrieve context to enhance model performance.

In this study, we propose the utilization of retrieval techniques for data augmentation in low-resource settings. Furthermore, while they perform retrieval on both the training and testing datasets, we only use the small seed training dataset for retrieval. It’s noteworthy that our approach can also be combined with theirs to further enhance the performance of NER in low-resource settings.

5 Conclusion

We present GPDA as a data augmentation framework for low-resource NER, which utilizes graph propagation with natural texts for augmentation. To make the most efficient utilization of the explored partially labeled text, we adopt consistency-restricted self-training. Experiment results show
Table 2: The statistics of the dataset used and generated in our experiments.

<table>
<thead>
<tr>
<th>Method</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>76.54</td>
<td>85.50</td>
<td>72.78</td>
<td>78.27</td>
</tr>
<tr>
<td>DAGA</td>
<td>77.71</td>
<td>86.51</td>
<td>81.32</td>
<td>81.65</td>
</tr>
<tr>
<td>NERDA</td>
<td>77.70</td>
<td>87.05</td>
<td>81.64</td>
<td>81.93</td>
</tr>
<tr>
<td>GPDA</td>
<td><strong>77.83</strong></td>
<td><strong>87.23</strong></td>
<td><strong>82.48</strong></td>
<td><strong>82.51</strong></td>
</tr>
</tbody>
</table>

Table 3: Results on the Ecom dataset.

Gold Training Data

... with the 2016 introduction of the voice editing and generation software [PRODUCT Adobe Voco], a prototype slated to be a part of the [PRODUCT Adobe Creative Suite] and [ORGANISATION DeepMind] [PRODUCT WaveNet], ...

Augmented Data

1) Adobe Voco is an unreleased ... prototype software by [ORG Adobe] that enables novel editing and generation of audio. Dubbed "[PRO Photoshop]-for-voice", it was first previewed at the [PRO Adobe MAX] event in November 2016.
2) With the 2016 introduction of Adobe Voco audio editing and generating software prototype slated to be part of the [no Adobe Creative Suite] and the similarly enabled DeepMind [no WaveNet], a [ALG deep neural network] based audio synthesis software ...
3) Adobe Device Central is a software program created and released by [ORG Adobe Systems] as a part of the [no Adobe Creative Suite] 3 (CS3) in March 2007.
4) [no Adobe Creative Suite], a design and development software suite by Adobe Systems.

Figure 3: An illustration of diversity of augmented data. The pink annotations are propagated via anchor matching while the yellow ones are labeled with EEA

that our proposed GPDA achieves substantial improvements over previous data augmentation methods on multiple low-resource NER datasets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61976139) and by Alibaba Group through Alibaba Innovative Research Program.

6 Limitations

There are some limitations in the use of GPDA.

• The label propagation procedure requires anchor matching in the light of annotation precision, which limits the unlabeled data source. However, Wikipedia is a open-domain easy-to-fetch corpus with anchor links, which can somehow mitigate the issue.

• Augmented Data generated by GPDA provide more diversity. But for some datasets, simple modifications (NERDA) on the original words performs better. We are investigating a hybrid approach to apply GPDA and NERDA in the same framework.

References


Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. Damo-nlp at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition.


Adams Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. ICLR.


