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

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

ShanghaiTech University & Alibaba Group & Zhejiang University

Introduction

Named entity recognition (NER) aims to locate and classify named entities in unstructured text. For low-resource NER, the semantic information of a specific entity is relatively limited due to fewer annotations. Data augmentation methods are used to improve model performance and robustness in low-resource scenarios.

Problems of previous methods: Previous data augmentation methods in NER generate synthetic data which inevitably introduces incoherence, semantic errors and lacking in diversity.

Solutions: 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.

Graph Propagate Data Augmentation

Notation:

- $x = \{x_1, \cdots, x_n\}$: input tokens
- $y = \{y_1, \cdots, y_n | y_i \in \mathbb{Y}\}$: label sequence, $\mathbb{Y}$ is the label set

Building Propagation Graph:
Given a labeled sample $(x^{(i)}, y^{(i)})$, we retrieve its corresponding augmented sentences $\{x'^{(i,j)}\}_{j=1}^{m}$ via a search engine. The top related sentences will be treated as connected to the original labeled sentence in the graph.

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 or dense retrieval based on L2 similarity.

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 samples $\{(x'^{(i,j)}, y'^{(i,j)})\}_{j=1}^{m}$. To strengthen the precision, propagation will not happen unless the anchor text in Wikipedia matches the labeled entity.

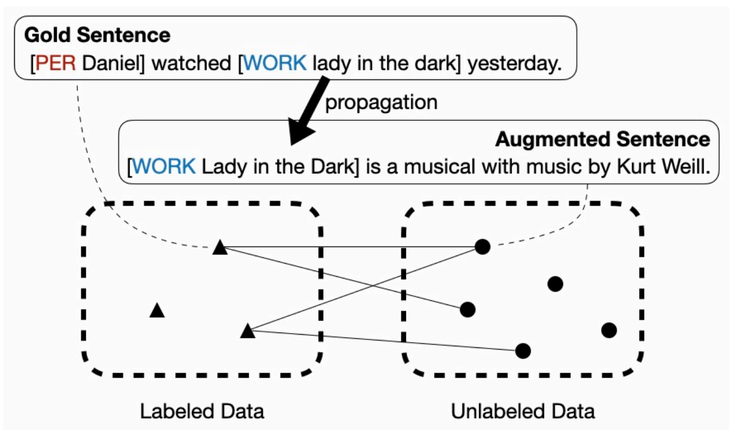

Figure 1. An example of GPDA. The graph is built on textual similarity with a Wikipedia-based search engine.

Training

We use a transformer-CRF sequence labeling model for the NER. Instead of directly minimizing the negative log-likelihood. We use a unified training objective with the labeled samples and generated partially annotated samples.

Unified Training Objective
We apply the forward-backward algorithm to compute the marginal probability of each token.

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

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

$$P_\theta(y_i|x) \propto \alpha(y_i) \times \beta(y_i)$$

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, $Q_\theta(y_i|x) = 1$ otherwise.
The model parameters can be optimized by minimizing the following objective:

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

where $D' = \{(x'^{(j)}, y'^{(j)})\}_{j=1}^{M}$ is the generated partially annotated samples and D is the original labeled data.

Explored Entity Annotations
To make the most efficient utilization of the explored annotations in D', we adopt consistency-restricted self-training. A well-trained NER model will be utilized to re-annotate the partially labeled augmented data under consistency restriction. Particularly, an augmented sample $(x'^{(j)}, y'^{(j)})$ will be re-annotated to $(x'^{(j)}, \hat{y}^{(j)})$. With the original labeled data D and re-annotated data $\hat{D} = \{(x'^{(j)}, \hat{y}^{(j)})\}_{j=1}^{M}$, we train a better NER model with the unified objective:

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

Experiments

Method	AI	Literature	Music	Politics	Science	Average
State-of-the-art Approaches						
Zheng et al. (2022)	63.28	70.76	76.83	73.25	70.07	70.84
Hu et al. (2022)	65.79	71.11	78.78	74.06	71.83	72.31
Tang et al. (2022)	66.03	68.59	73.1	71.69	75.52	70.99
Baseline w/o Data Augmentation						
BERT-CRF	65.06	71.39	78.18	74.46	73.95	72.61
Data Augmentation Approaches						
DAGA (Ding et al., 2020)	66.77	71.15	78.48	73.30	73.07	72.55
NERDA (Dai and Adel, 2020)	**70.20**	71.28	79.56	75.30	74.37	74.14
GPDA *(sparse retrieval w/o EEA)*	67.14	72.20	79.55	74.96	74.69	73.71
GPDA *(dense retrieval w/o EEA)*	67.76	72.11	77.54	74.86	73.07	73.07
GPDA *(sparse retrieval w/ EEA)*	70.05	**72.34**†	**80.16**†	**75.95**†	**75.55**†	**74.81**†

Figure 2. Comparisons of different studies and our proposed GPDA on the CrossNER dataset.

Case Study

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 [PRO Adobe Creative Suite] and the similarly enabled DeepMind [PRO 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 [PRO Adobe Creative Suite] 3 (CS3) in March 2007 .
4) [PRO Adobe Creative Suite], a design and development software suite by Adobe Systems.

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