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]

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 (Şahin 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

[*] The email of the authors are: Jiong Cai (caijiong@shanghaitech.edu.cn), Shen Huang (pangda@alibaba-inc.com), Yong Jiang (yongjiang.jy@alibaba-inc.com), Zeqi Tan (zqtan@zju.edu.cn), Penjun Xie (chengchen.xpj@alibaba-inc.com) and Kewei Tu (tukw@shanghaitech.edu.cn). Yong Jiang and Kewei Tu are the corresponding authors.

[1]Our code is publicly available at `https://github.com/modelscope/AdaSeq/tree/master/examples/GPDA`.

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 $\boldsymbol{y} = \{y_1, \cdots, y_n | y_i \in \mathbb{Y}\}$ at each position for the input tokens $\boldsymbol{x} = \{x_1, \cdots, x_n\}$, where $\mathbb{Y}$ denotes the label set. The sequence labeling model feeds the input $\boldsymbol{x}$ 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:

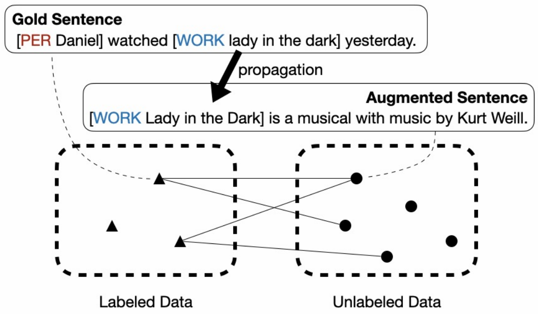

Figure 1: An example of GPDA. The graph is built on textual similarity with a Wikipedia-based search engine.

$$\psi(y_{i-1}, y_i, r_i) = \exp(\boldsymbol{W}_y^T r_i + \boldsymbol{b}_{y_{i-1}y_i})$$

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

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 $P_\theta(y_i|\boldsymbol{x})$ with the forward-backward algorithm.

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

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

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

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

$$Q_\theta(\boldsymbol{y}^*|\boldsymbol{x}) = \prod_{i=1}^{n} Q_\theta(y_i|\boldsymbol{x})$$

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**†

Table 1: Comparisons of different studies and our proposed GPDA on the CrossNER dataset. † means the result is significantly better than the compared baseline methods (with Student's t-test with $p < 0.05$).

where $Q_\theta(y_i|\boldsymbol{x})$ is defined as $P_\theta(y_i|\boldsymbol{x})$ if y_i is observed, otherwise $Q_\theta(y_i|\boldsymbol{x}) = 1$.

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

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

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

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

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 $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$, we retrieve its corresponding augmented sentences $\{\boldsymbol{x}'^{(i,j)}\}_{j=1}^m$ 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.

Label Propagation While building the graph, label propagation is conducted from labeled data $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$ to unlabeled data $\{\boldsymbol{x}'^{(i,j)}\}_{j=1}^m$ to generate partially annotated $\{(\boldsymbol{x}'^{(i,j)}, \boldsymbol{y}'^{(i,j)})\}_{j=1}^m$. 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 $D' = \{(\boldsymbol{x}'^{(j)}, \boldsymbol{y}'^{(j)})\}_{j=1}^M$ 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:

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

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 $(\boldsymbol{x}'^{(j)}, \boldsymbol{y}'^{(j)})$ will be re-annotated to $(\boldsymbol{x}'^{(j)}, \hat{\boldsymbol{y}}^{(j)})$. Now we have $\hat{D} = \{(\boldsymbol{x}'^{(j)}, \hat{\boldsymbol{y}}^{(j)})\}_{j=1}^M$. Along with the original labeled data D, we train a better NER model following the objective in Section 2.1:

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

[2]Sparse retrieval is implemented with Elastic Search
[3]Dense retrieval is implemented with ColBERT

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%.

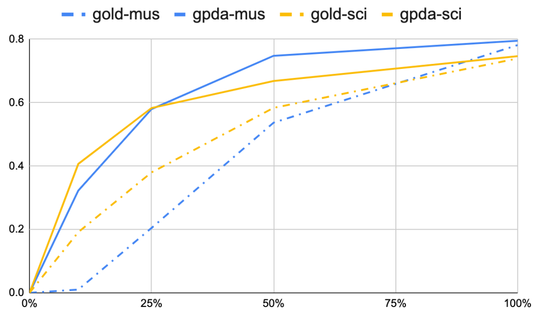

Figure 2: Results with different size of gold samples on Music and Science of CrossNER.

Case Study Taking a closer look at the augmented cases in Fig. 3, we notice that GPDA generates different contexts concerning the entity "Adobe Creative Suite". The augmented data generated by GPDA introduces more diversity to help reduce overfitting. Different from synthetic data, these generated data are all from natural texts so that there is no need to worry about the coherence in syntax or semantics.

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

		#Train / #Dev / #Test	#DAGA / Avg Ent	#NERDA / Avg Ent	#GPDA / Avg Ent	#GPDA+EEA / Avg Ent
CrossNER	Ai	100/350/431	866 / 1.60	6000 / 5.32	447 / 3.41	2428 / 7.79
	Lit	100/400/416	2814 / 2.21	6000 / 5.41	297 / 2.00	3967 / 7.71
	Mus	100/380/465	1102 / 1.93	6000 / 6.49	307 / 2.35	4273 / 10.40
	Pol	200/541/651	5274 / 2.70	12000 / 6.52	718 / 2.36	8463 / 8.95
	Sci	200/450/543	3896 / 2.79	12000 / 5.38	552 / 3.97	7494 / 9.53
Ecom	en	1000/1000/1000	4740 / 1.10	32000 / 0.96	30000 / 1.50	N/A
	es	1000/1000/1000	20362 / 1.07	32000 / 1.09	30000 / 1.41	N/A
	fr	1000/1000/1000	17340 / 1.08	32000 / 1.14	30000 / 1.24	N/A

Table 2: The statics of the dataset used and generated in our experiments.

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

Method	en	es	fr	Avg
Baseline	76.54	85.50	72.78	78.27
DAGA	77.11	86.51	81.32	81.65
NERDA	77.10	87.05	81.64	81.93
GPDA	**77.83**	**87.23**	**82.48**	**82.51**

Table 3: Results on the Ecom dataset.

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

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on 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.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jinpeng Hu, He Zhao, Dan Guo, Xiang Wan, and Tsung-Hui Chang. 2022. A label-aware autoregressive framework for cross-domain NER. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2222–2232, Seattle, United States. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Gakuto Kurata, Bing Xiang, and Bowen Zhou. 2016. Labeled data generation with encoder-decoder lstm for semantic slot filling. In *Interspeech*.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7056–7066, Online. Association for Computational Linguistics.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2020. Crossner: Evaluating cross-domain named entity recognition.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium. Association for Computational Linguistics.

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.

Minghao Tang, Peng Zhang, Yongquan He, Yongxiu Xu, Chengpeng Chao, and Hongbo Xu. 2022. DoSEA: A domain-specific entity-aware framework for cross-domain named entity recogition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2147–2156, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online. Association for Computational Linguistics.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, Wei Lu, and Yong Jiang. 2022. DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

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.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Xin Zhang, Yong Jiang, Xiaobin Wang, Xuming Hu, Yueheng Sun, Pengjun Xie, and Meishan Zhang. 2022. Domain-specific NER via retrieving correlated samples. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2398–2404, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Junhao Zheng, Haibin Chen, and Qianli Ma. 2022. Cross-domain named entity recognition via graph matching. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2670–2680, Dublin, Ireland. Association for Computational Linguistics.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. MELM: Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2251–2262, Dublin, Ireland. Association for Computational Linguistics.