Domain-Specific NER via Retrieving Correlated Samples

Xin Zhang[1], Yong Jiang, Xiaobin Wang, Xuming Hu[2], Yueheng Sun[3], Pengjun Xie, Meishan Zhang[4*]

[1]School of New Media and Communication, Tianjin University, China
[2]School of Software, Tsinghua University, China
[3]College of Intelligence and Computing, Tianjin University, China
[4]Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen), 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 calibrating 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 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[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.

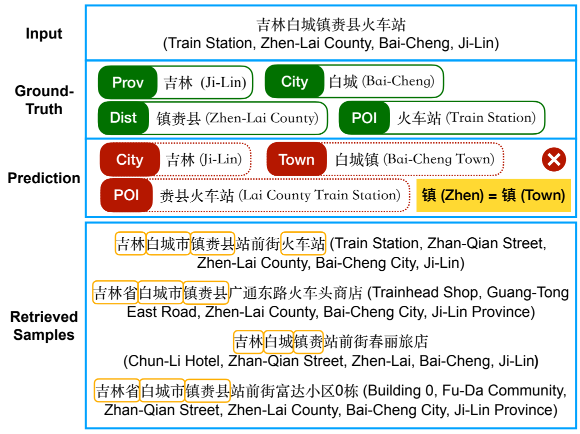

Figure 1: An address example with retrieved texts. The model incorrectly predicted “白城镇 (Baicheng Town)” and “赉县火车站 (Lai County Train Station)” because they match common patterns, i.e. “XX镇” (Xx Town) and “X县火车站” (X County Train Station). “吉林” is ambiguous, which is both a province and a city.

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

[*]Corresponding author.

[1]Because this pattern is indeed correct in most cases. This problem also exists in models with internal larger datasets.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2398–2404
October 12–17, 2022.

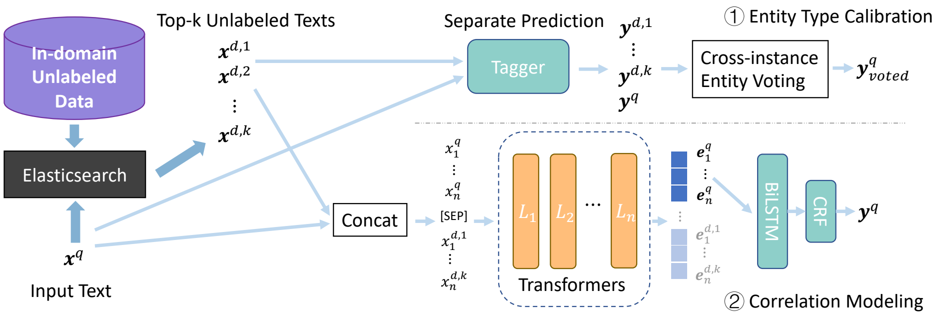

Figure 2: The overview of our suggested methods.

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 (Jilin)" and "白城 (Baicheng)" 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-

Method	Micro P	Micro R	Micro F1	Macro P	Macro R	Macro F1
Chinese Address						
Human	93.04	92.01	92.52	87.83	84.52	86.14
BC	85.56	83.90	84.72	82.20	76.48	79.24
NEZHA-BC	91.29	90.62	90.95	86.41	84.68	85.53
Entity-Voting†	91.67	91.00	91.34	86.70	84.83	85.76
Cross-Encoder†	**92.41**	**91.95**	**92.18**	**87.25**	**85.71**	**86.48**
Self-Training	91.57	91.02	91.29	86.65	85.37	86.01
Biaffine	91.35	90.25	90.80	86.32	84.59	85.45
Seq2set	89.43	87.69	88.55	83.89	80.12	81.96
Locate&Label	90.28	87.76	89.00	85.95	82.29	84.08
PIQN	90.27	87.83	89.03	86.04	80.28	83.06
E-commerce						
BC	65.31	62.54	63.90	58.88	50.38	54.30
NEZHA-BC	82.73	83.23	82.98	79.35	78.04	78.69
Entity-Voting†	82.83	83.33	83.08	79.56	78.19	78.87
Cross-Encoder†	**83.49**	83.74	**83.61**	**81.45**	**79.34**	**80.38**
Self-Training	81.51	**85.25**	83.34	78.89	79.29	79.09
Biaffine	81.91	84.06	82.97	80.14	79.05	79.59
Seq2set	82.77	81.65	82.21	81.39	76.44	78.84
Locate&Label	80.43	83.21	81.80	76.63	78.22	77.42
PIQN	83.43	82.54	82.98	81.23	75.60	78.31
BERT-CLS (2019)	77.06	80.65	78.81	-	-	-
MRC-NER (2020)	79.47	78.30	78.88	-	-	-
CoFEE-BERT (2020)	79.13	80.34	79.73	-	-	-
CoFEE-MRC (2020)	80.26	78.88	79.56	-	-	-

Table 1: Main results. † means statistically significant.

lation patterns among samples. Specifically, we concatenate the input text with retrieved samples by the separator (i.e., [SEP]), and then encode them by pretrained language models. Finally, only the contextual embeddings of the input text are fed into the NER tagger (here BiLSTM-CRF). With this simple 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 sentences[3] 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.

[3]We remove a few sentences that are particularly long and do not contain entities.

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-

Method		100%	50%	20%	10%	5%	3%
Chinese Address							
Micro	NEZHA-BC	90.95	90.04	88.79	87.57	86.56	84.69
	Cross-encoder	92.18 ↑1.23	91.56 ↑1.52	90.33 ↑1.54	89.13 ↑1.56	88.61 ↑2.05	86.82 ↑2.13
Macro	NEZHA-BC	85.53	84.22	82.14	78.08	75.99	72.78
	Cross-encoder	86.48 ↑0.95	85.36 ↑1.14	83.43 ↑1.29	79.66 ↑1.58	78.31 ↑2.32	75.41 ↑2.63
E-commerce							
Micro	NEZHA-BC	82.98	81.54	79.52	77.89	75.80	73.86
	Cross-encoder	83.61 ↑0.63	82.21 ↑0.67	80.29 ↑0.77	78.99 ↑1.10	77.18 ↑1.38	74.60 ↑0.74
Macro	NEZHA-BC	78.69	77.03	75.17	72.86	69.63	67.11
	Cross-encoder	80.21 ↑1.52	78.21 ↑1.18	76.47 ↑1.30	74.28 ↑1.42	71.47 ↑1.84	68.03 ↑0.92

Table 2: Test F1 scores at various low-resource settings.

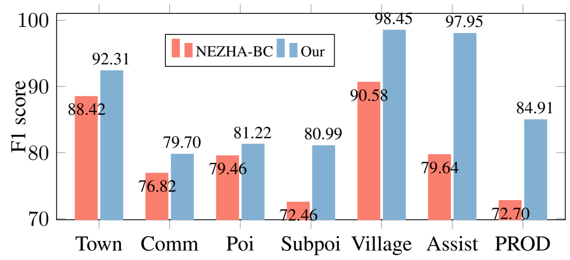

Figure 3: Test F1 of some hard entity types from main results, where their scores are less than the overall value.

nals for entity understanding (such as 镇赉县火车站 v.s 镇赉县站前街火车站). When compared with other recent state-of-the-art NER methods (Biaffine, Seq2set, Locate&Label, and PIQN), our approaches outperform them by a large margin. It is worth noting that our model outperforms the self-training (whose unlabeled corpus is in the same scale of samples we modeled), demonstrating that the correlation modeling is more effective. Then we plot detailed scores by categories whose F1 score is less than the overall F1 in Figure 3. All of these difficult categories are significantly improved, showing that the correlated samples are helpful.

Other Results. For the Chinese Address dataset, we also report the performance of human annotators without extra information provided. Notably, our approaches achieve comparable performance with humans, which empirically verifies that modeling text correlation with the retrieval perspective might have the possibility to simulate human expert annotations. For the E-commerce dataset, we also report other published results. Our NEZHA-BC is comparable with all the baseline implementations.

#Address	400M	40M	10M	4M	400k	100k
Micro F1	92.18	92.07	91.89	91.53	91.31	91.16

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

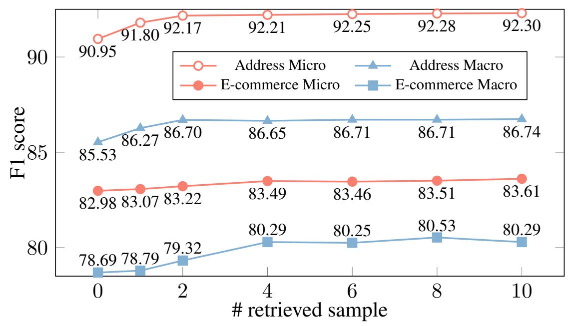

Figure 4: Test F1 scores of incooperating different correlated sample num by our cross-encoder.

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[4] 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,

[4]The 0 samples cross-encoder degrade to the baseline NEZHA-BC.

Method	NEZHA-BC	Entity-Voting	Cross-Encoder
Seconds	14.73	500+	38.41

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 droped 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 (Guu 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 and transformer-based correlation modeling, where the latter one is similar to Wang et al. (2021). Our recent work (Wang et al., 2022) also investigated retrieving knowledge from the Wikipedia, which can augment the context of NER inputs and shows significant improvements in SemEval-2022 Task 11 Multilingual NER.

This work could be further investigated with some more sophisticated techniques, such as example-based learning (Gao et al., 2021; Lee et al., 2022; Liu et al., 2022). Meanwhile, it also may help the NER task to extend to the low-resource and zero-shot scenarios (Meng et al., 2021; Zhang et al., 2021; Hu et al., 2021; Lu et al., 2022; Hu et al., 2020).

4 Conclusion

In this work, we investigated utilizing naturally correlated samples to improve current NER models on the Chinese address and e-commerce domain. We propose to retrieve correlated samples for the given text by the BM25 and elasticsearch engine. To explore the correlations in a light way, we suggest calibrating the predicted entity types 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

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proc. of the ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Youngwon Choi, Sungdong Lee, and Joong-Ho Won. 2021. Learning from nested data with ornstein auto-encoders. In *Proc. of the ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 1943–1952. PMLR.

Youngwon Choi and Joong-Ho Won. 2019. Ornstein auto-encoders. In *Proc. of the IJCAI*, pages 2172–2178. ijcai.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the NAACL*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruixue Ding, Pengjun Xie, Xiaoyan Zhang, Wei Lu, Linlin Li, and Luo Si. 2019. A neural multi-digraph model for Chinese NER with gazetteers. In *Proc.of the ACL*, pages 1462–1467, Florence, Italy. Association for Computational Linguistics.

Murat Dundar, Balaji Krishnapuram, Jinbo Bi, and R. Bharat Rao. 2007. Learning classifiers when the training data is not IID. In *Proc. of the IJCAI*, pages 756–761.

Elasticsearch. 2022. Elasticsearch: The official distributed search & analytics engine.

Zheng Fang, Yanan Cao, Tai Li, Ruipeng Jia, Fang Fang, Yanmin Shang, and Yuhai Lu. 2021. TEBNER: Domain specific named entity recognition with type expanded boundary-aware network. In *Proc. of the EMNLP*, pages 198–207, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proc. of the ACL-IJCNLP*, pages 3816–3830, Online. Association for Computational Linguistics.

Zhichao Geng, Hang Yan, Zhangyue Yin, Chenxin An, and Xipeng Qiu. 2022. TURNER: the uncertainty-based retrieval framework for chinese NER. *CoRR*, abs/2202.09022.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proc. of the ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. 2020. SelfORE: Self-supervised relational feature learning for open relation extraction. In *Proc. of the EMNLP*, pages 3673–3682, Online. Association for Computational Linguistics.

Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and Philip S. Yu. 2021. Gradient imitation reinforcement learning for low resource relation extraction. In *Proc. of the EMNLP*, pages 2737–2746, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alibaba Inc. 2022. Ccks 2021 chinese address element parsing dataset.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proc. of the EACL*, pages 874–880, Online. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proc. of the NAACL*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Dong-Ho Lee, Akshen Kadakia, Kangmin Tan, Mahak Agarwal, Xinyu Feng, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. 2022. Good examples make a faster learner: Simple demonstration-based learning for low-resource NER. In *Proc. of the ACL*, pages 2687–2700, Dublin, Ireland. Association for Computational Linguistics.

Hao Li, Wei Lu, Pengjun Xie, and Linlin Li. 2019. Neural Chinese address parsing. In *Proc. of the NAACL*, pages 3421–3431, Minneapolis, Minnesota. Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.*, 34(1):50–70.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proc. of the ACL*, pages 5849–5859, Online. Association for Computational Linguistics.

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In *Proc. of the ACL*, pages 1518–1533, Online. Association for Computational Linguistics.

Shuliang Liu, Xuming Hu, Chenwei Zhang, Shu'ang Li, Lijie Wen, and Philip Yu. 2022. HiURE: Hierarchical exemplar contrastive learning for unsupervised relation extraction. In *Proc. of the NAACL*, pages 5970–5980, Seattle, United States. Association for Computational Linguistics.

Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. 2019. Towards improving neural named entity recognition with gazetteers. In *Proc. of the ACL*, pages 5301–5307, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

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.

Ying Luo, Fengshun Xiao, and Hai Zhao. 2020. Hierarchical contextualized representation for named entity recognition. In *Proc. of the AAAI*, pages 8441–8448. AAAI Press.

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.

Xue Mengge, Bowen Yu, Zhenyu Zhang, Tingwen Liu, Yue Zhang, and Bin Wang. 2020. Coarse-to-Fine Pre-training for Named Entity Recognition. In *Proc. of the EMNLP*, pages 6345–6354, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. of the EMNLP-IJCNLP*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. of the SIGIR*, pages 232–241. ACM/Springer.

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.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and label: A two-stage identifier for nested named entity recognition. In *Proc. of the ACL*, pages 2782–2794, Online. Association for Computational Linguistics.

Yongliang Shen, Xiaobin Wang, Zeqi Tan, Guangwei Xu, Pengjun Xie, Fei Huang, Weiming Lu, and Yueting Zhuang. 2022. Parallel instance query network for named entity recognition. In *Proc. of the ACL*. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Zeqi Tan, Yongliang Shen, Shuai Zhang, Weiming Lu, and Yueting Zhuang. 2021. A sequence-to-set network for nested named entity recognition. In *Proc. of the IJCAI*, pages 3936–3942. ijcai.org.

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 *Proc. of the ACL-IJCNLP*, 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 *Proc. of the SemEval*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. NEZHA: neural contextualized representation for chinese language understanding. *CoRR*, abs/1909.00204.

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.

Xingcheng Yao, Yanan Zheng, Xiaocong Yang, and Zhilin Yang. 2022. NLP from scratch without large-scale pretraining: A simple and efficient framework. In *Proc. of the ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 25438–25451. PMLR.

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

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018a. Guiding neural machine translation with retrieved translation pieces. In *Proc. of the NAACL*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Qi Zhang, Jinlan 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.

Xin Zhang, Guangwei Xu, Yueheng Sun, Meishan Zhang, and Pengjun Xie. 2021. Crowdsourcing learning as domain adaptation: A case study on named entity recognition. In *Proc. of the ACL-IJCNLP*, pages 5558–5570, Online. Association for Computational Linguistics.