A Simple but Effective Approach to Cross-domain Nested Named Entity Recognition

Kamil Agliullin
Innopolis University
Innopolis, Russia
k.agliullin@innopolis.university

Abstract

Until recently, researchers would only consider cross-domain flat NER. In this work, we propose an embarrassingly easy but effective approach to the double challenge of cross-domain nested NER. We use a RuBERT-base encoder and a Biaffine decoder with CNN block as a backbone nested NER model. The actual approach to cross-domain NER is simple: keep only the common categories between source and target domain datasets, train the model on a source domain and apply it to a target domain. The results show that proposed method has a drop in performance compared to the usual training approach, but, unlike latter, does not require any fine-tuning on target domain data.

Keywords: Named Entity Recognition, Cross-domain NER, Nested NER, RuBERT, Biaffine

1 Introduction

Recently, cross-domain NER started to become a topic of interest, since it has a number of practical applications, most importantly facilitating NER for limited-resource domains. However, domain generalization is a challenging task due to a number of reasons. (Jia and Zhang, 2020) define three main problems arising while solving cross-domain NER. First, instances of the same category could mean different things in source and target domain datasets (for more details, please refer to Section 5.2). Second, different categories have unequal similarities across domain. As an intuitive example, category "Country" could mean almost the same thing in news and biomedical domain, which cannot be surely said about the "Product" category. Last but not least, source and target dataset may contain varying number and types of categories.

To solve the problem of cross-domain NER, we propose a method that does not use any target domain data or external data for training. For the experiments, we use NEREL, the largest dataset in Russian annotated with named entities and relations (Loukachevitch et al., 2021), and Russian corpus of NEREL-BIO, an extension of NEREL dataset, which goes deeper into biomedical domain (Loukachevitch et al., 2023). Our main contributions can be summarized as the following:

- We explore the existing methods in nested and cross-domain NER.
- We implement a simple, but reasonably effective method for cross-domain nested NER that requires no fine-tuning on target domain data.

The code for preprocessing, training and experiments can be found at `https://github.com/kamilain1/Cross-Domain-Nested-NER`.

2 Related Work

Flat NER approaches, where a single label is predicted for each token, have difficulties with processing nested entities. To bypass that problem, various nested NER approaches were proposed. Three main types of nested NER methods can be highlighted: seq-to-seq, span-based

and MRC-based methods. In their work, (Loukachevitch et al., 2021) explore different approaches to nested NER, and present span-based approach of (Yu et al., 2020) and MRC-based approach of (Li et al., 2019) as two baselines with the highest performances. Similarly for NEREL-BIO (Loukachevitch et al., 2023), authors present MRC-based (Li et al., 2019) and second-best sequence learning approach (Shibuya and Hovy, 2020) as baselines. (Artemova et al., 2022) report the results of RuNNE-2022 shared task on NEREL, in which rule-based method outperformed neural network-based approaches.

Recently, span-based methods became a popular line of research within nested NER. Method of (Yu et al., 2020) was continuously improved by utilizing Triaffine mechanism (Yuan et al., 2021), boundary smoothing (Zhu and Li, 2022), and CNN (Yan et al., 2022). In this work, the approach of (Yan et al., 2022) was chosen as a backbone nested NER model due to its high performance, fast training and inference, and overall simplicity.

Most of the current approaches to flat and nested NER are limited to in-domain setting. Cross-domain NER aims to generalize the model to more than one domain at once. First, this task was approached as a supervised multi-task learning problem (Yang et al., 2017), (Wang et al., 2018), (Wang et al., 2019).

To solve the problem of probable data scarcity in target domain, several different works propose low-resource approaches to cross-domain NER through learning general representations of named entities (Liu et al., 2020), (Liu et al., 2021), and through data augmentation (Chen et al., 2021), (Wang et al., 2020), (Wang et al., 2021), (Yang et al., 2022).

Previous methods were designed and tested only for flat NER. The most recent update on cross-domain NER was made by (Ming et al., 2022). They were the first to consider few-shot nested NER setting by using BERT-multilingual encoder, Biaffine layer and contrastive optimization module. They use FewNERD as source dataset, and they use GENIA, GermEval, NEREL as target datasets. They achieve 33.71%, 39.56%, 44.47% F1 score in 1-shot setting, and 46.06%, 47.07%, 58.95% F1 in 5-shot setting on each respective target dataset.

In this paper, we make an intuitive assumption that strong in-domain generalization properties of the backbone model can help model achieve good cross-domain generalization *per se*. Therefore, we propose a method for cross-domain nested NER which, in contrast to the approach of (Ming et al., 2022), requires no fine-tuning on target domain data. The idea is embarrassingly simple yet effective: to keep only common entity types between datasets, train backbone model on source dataset and directly test it on target dataset. To the best of our knowledge, we are the first to test cross-domain nested NER on NEREL and NEREL-BIO as respective source and target datasets.

3 Method

3.1 Nested NER Method

To train a model, span-based method of (Yan et al., 2022) was followed:

- The input sentence of length n was encoded using the pre-trained language model.
- The encoded sentence X is then passed into two MLPs to obtain representations for the start and end of spans. Both are then sent into multi-head Biaffine decoder (Yan et al., 2022), (Yu et al., 2020), (Dozat and Manning, 2016), (Vaswani et al., 2017).

$$R = MHBiaffine(H_e, H_s) \tag{1}$$

The result is a score matrix of dimension $n \times n \times r$, where r is the feature size.

- CNN block is applied several times to model local relations between spans:

$$R' = Conv2d(R) \tag{2}$$

$$R'' = GeLU(LayerNorm(R' + R)) \tag{3}$$

- Predictions are obtained in the following way:

$$P = Sigmoid(Linear(R'' + R)) \tag{4}$$

  Resulting P has dimension $n \times n \times t$, where t is the number of named entity categories.
- Loss is calculated as binary cross entropy:

$$\mathcal{L} = -\sum_{0 \le i,j < n} y_{ij} \log(P_{ij}) + (1 - y_{ij}) \log(1 - P_{ij}) \tag{5}$$

- When inference, the score for span (i, j) is calculated as:

$$\hat{y}_{ij} = \frac{P_{ij} + P_{ji}}{2} \tag{6}$$

- Finally. decoding process is carried out. First, spans with all their scores less than 0.5 are discarded. Then, sorted spans are chosen in the descending score order. If the lower score span "clashes" (Yu et al., 2020) with higher score span, former is dropped.

3.2 Cross-domain Method

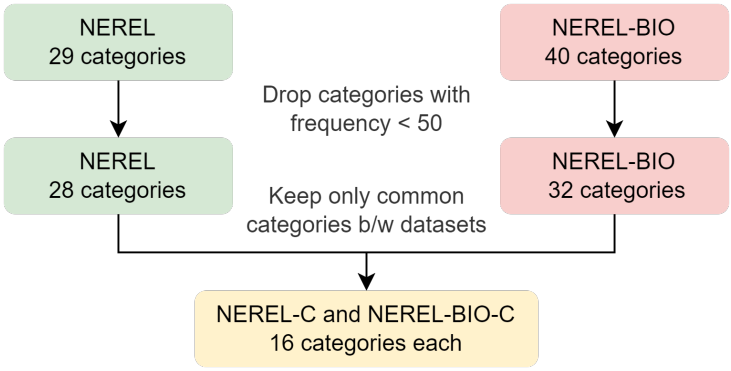

Figure 1: Preprocessing pipeline

To solve the task of cross-domain NER between NEREL and NEREL-BIO, a check was added to a preprocessing script, so that entity mention could be considered only if its type is common between NEREL and NEREL-BIO.

Fig. 1 shows the process of modifying NEREL and NEREL-BIO for using in cross-domain setting. In case of "DISO" entity type in NEREL-BIO, it was renamed to "DISEASE" to match the NEREL annotation scheme. Preprocessed datasets with common label types were named NEREL-C and NEREL-BIO-C.

4 Experiments and Results

4.1 Experiments Outline

Two types of experiments were conducted. First, CNN-NER was trained separately on NEREL and NEREL-BIO to obtain reference results. These results were compared with the baselines. Second, cross-domain method was tested: CNN-NER was trained only on NEREL-C and evaluated on both NEREL-C and NEREL-BIO-C. There is no comparison with the approach of (Ming et al., 2022), since they did not publish the code of their implementation.

4.2 Evaluation Metrics

Micro F1 score was used as a metric. In this work, strict evaluation was used: prediction is considered correct, if predicted span boundaries and label *exactly* match the ground truth.

4.3 Training Details

For CNN-NER backbone method, RuBERT-base (Kuratov and Arkhipov, 2019) was chosen as a pre-trained encoder since datasets are in Russian. All CNN-NER model instances were trained using AdamW optimizer with warmup-decay linear scheduler, training was done on Nvidia Tesla P100 GPU.

Table 1: Hyperparameters used

Hyperparams	NEREL	NEREL-BIO	NEREL-C
# Epoch	50	10	10
Learning rate	5e-6	7e-6	5e-6
Batch size	4	4	4
Hidden size h	200	400	200
Feature size r	100	200	100

Table 1 shows CNN-NER hyperparameters chosen for each dataset. For each dataset, 10 and 50 epochs setting was tested, and the best one was selected. For most of the hyperparameters, (Yan et al., 2022) were followed to select the appropriate values. In case of overfitting, the best epoch was chosen depending on dev F1 score.

4.4 Main Results

As mentioned in experimental setup, first experiment was done by training two models: one on NEREL and one on NEREL-BIO, and then comparing the results with the baselines. *RuBERT-base* is a default encoder for all methods which require it, and *Micro F1* score is a default score for all comparisons, if not specified otherwise.

Table 2: Performance on test sets for the usual training approach

Model	**F1**	**Precision**	**Recall**
NEREL			
MRC (Li et al., 2019)	79.64	78.70	80.24
Biaffine (Yu et al., 2020)	76.38	81.92	71.54
Rule-based (Artemova et al., 2022)	macro 81.1	-	-
CNN-NER (Yan et al., 2022)	**86.05**	**84.78**	**87.35**
NEREL-BIO			
MRC (Li et al., 2019)	76.94	66.83	59.90
Second-best (Shibuya and Hovy, 2020)	74.10	75.28	72.98
CNN-NER (Yan et al., 2022)	**78.23**	**77.03**	**79.47**

Table 3: Performance on test sets for the cross-domain approach

Model	**NEREL-C**			**NEREL-BIO-C**		
	F1	P	R	F1	P	R
Cross-domain CNN-NER	89.23	87.45	91.09	63.55	78.05	53.59

Table 2 compares the training results between baselines chosen by (Loukachevitch et al., 2021) and backbone method used in this work. For NEREL, approach with the highest performance from RuNNE shared task (Artemova et al., 2022) is also listed. It can be seen that for both NEREL and NEREL-BIO, CNN-NER (Yan et al., 2022) achieves better performance. Therefore, reasonable decision was made to further use this approach for cross-domain setting.

In the second experiment, CNN-NER was trained and evaluated on NEREL-C. After that, the same model instance was evaluated on NEREL-BIO-C without fine-tuning.

Table 3 shows the results of the cross-domain approach. It can be seen that there is a performance drop (-14.68% F1), when comparing the performance of cross-domain method on NEREL-BIO-C between the usual training approach on NEREL-BIO. As was mentioned before, there is no direct comparison with the method of (Ming et al., 2022). However, their results on NEREL as a target domain dataset were mentioned in the Related Work section.

5 Analysis

5.1 Datasets Examination

High percentage of matching tokens between source and target datasets can artificially improve cross-domain model performance. Therefore, these statistics should be examined.

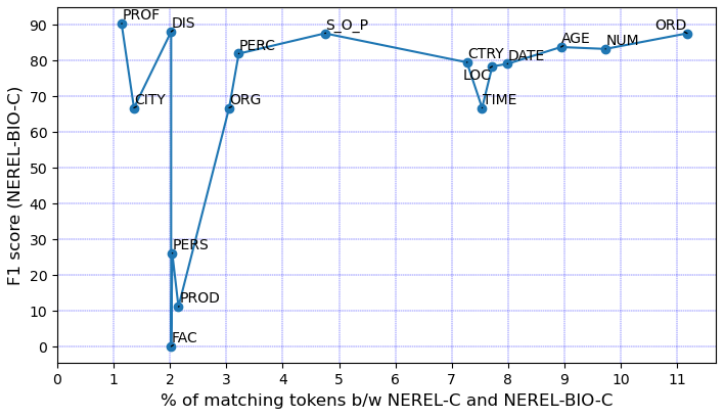

Figure 2: Comparison of matching tokens percentage and F1 score

Fig. 2 represents the information about coinciding tokens and corresponding performance. Each node represents a category. It can be seen that, apart from three outlying points, F1 score for entity types does not depend on percentage of matching tokens, since there is no discernible pattern which could prove the opposite. Therefore, quantity of similar tokens has none or very small contribution to model performance.

5.2 Performance Breakdown

Table 4 shows the comparison between ordinary training method on test set of NEREL-BIO and cross-domain method on test set NEREL-BIO-C. Additionally, for each category, corresponding count of occurrences in train sets of NEREL and NEREL-BIO is listed. We want to note that there can be slight inaccuracy in scores due to a low number of counts in test set for some categories.

For the cross-domain method, significant performance drop can be noticed on three categories, which were most affected by *semantic shift*. In common usage, *semantic shift* means evolutionary change of the word meaning over time, but in context of this work it could mean change of the word meaning over domain (Chen et al., 2018). For example, in source domain

Table 4: Performance breakdown on test sets for usual and cross-domain approaches

Cat.	NEREL-BIO			NEREL-BIO-C			Difference			Count (train set)	
	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	NEREL	NEREL-BIO
AGE	89.80	91.67	88.00	83.67	85.42	82.00	-6.13	-6.25	-6.00	794	389
CITY	66.67	55.56	83.33	66.67	66.67	66.67	0.00	11.11	-16.66	1517	83
CTRY	92.21	92.21	92.21	79.39	96.30	67.53	-12.82	4.09	-24.68	2950	182
DATE	78.43	91.95	68.38	78.97	79.31	78.63	0.54	-12.64	10.25	3068	967
DIS	93.09	93.52	92.66	87.96	94.71	82.11	-5.13	1.19	-10.55	322	9706
FAC	42.55	62.50	32.26	0.00	0.00	0.00	-42.55	-62.50	-32.26	491	151
LOC	21.05	80.00	12.12	78.26	75.00	81.82	57.21	-5.00	69.70	335	61
NUM	89.12	84.67	94.07	83.14	88.33	78.52	-5.98	3.66	-15.55	1162	3358
ORD	79.41	90.00	71.05	87.50	83.33	92.11	8.09	-6.67	21.06	637	873
ORG	82.05	87.67	77.11	66.67	86.54	54.22	-15.38	-1.13	-22.89	4785	353
PERC	92.50	92.50	92.50	81.82	75.00	90.00	-10.68	-17.50	-2.50	95	1498
PERS	90.98	88.55	93.55	26.03	86.36	15.32	-64.95	-2.19	-78.23	6551	5787
PROD	40.00	60.00	30.00	11.11	6.45	40.00	-28.89	-53.55	10.00	333	57
PROF	76.00	82.61	70.37	90.20	95.83	85.19	14.20	13.22	14.82	5973	153
S_O_P	70.00	63.64	77.78	87.50	100.00	77.78	17.50	36.36	0.00	420	66
TIME	71.43	62.50	83.33	66.67	55.56	83.33	-4.76	-6.94	0.00	218	107

dataset, NEREL (consequently, NEREL-C), "PERSON" usually represents proper nouns, e.g. "David Rockefeller", "Kevin Rudd". On the other hand, in target domain dataset, NEREL-BIO (hence, NEREL-BIO-C), "PERSON" is mainly a common noun, representing one person or group of people, e.g. "patient", "sick", etc. Moreover, "FACILITY" category denotes even more divergent concepts in both datasets. In NEREL, this label type mostly represents the names of city facilities, e.g. "house 76B", "St. Louis airport". In NEREL-BIO, the same category name refers to various objects and names within healthcare domain: breastfeeding tents, hospital room names, etc. Similar events, but to a different extent may be described for the other underperforming categories.

For "LOCATION", "PROFESSION", "STATE_OR_PROVINCE" categories, performance increase can be explained by the fact that NEREL contains much more training data for these categories than NEREL-BIO: 335 vs. 61, 5973 vs. 153, 420 vs. 66 instances respectively in NEREL and NEREL-BIO for each corresponding category. When cross-domain model was trained on NEREL-C, knowledge about such categories was transferred onto NEREL-BIO-C, while in usual training method, model saw data from NEREL-BIO only. This aspect may affect performance in some other categories.

Regarding the other categories, differences range considerably, in both negative and in positive directions. As was mentioned before, there is an overall -14.68 % micro F1 drop, which is arguably not a large reduction, considering that there was no fine-tuning on NEREL-BIO-C in cross-domain method.

6 Conclusion

This work explores a novel direction within Named Entity Recognition. A simple yet effective way was proposed to solve the double issue of cross-domain nested NER. Our method keeps only the common entity types between source and domain datasets, and utilizes generalization capabilities of the span-based nested NER approach. Experiments on NEREL and NEREL-BIO as source and target domain datasets show that our method achieves comparable performance to

the usual training approach without using target domain data for fine-tuning. There is no direct comparison with the existing approach to cross-domain nested NER by (Ming et al., 2022), since they did not publish the code of their implementation. We do not state that our research is complete, since there is a vast possibility for improvement. For example, experimenting with the set of categories to focus on some particular entity types, or adopting more sophisticated methods such as contrastive learning to solve the category discrepancy problem between source and target domain datasets can help achieve much better performance on this task.

References

Ekaterina Artemova, Maxim Zmeev, Natalia Loukachevitch, Igor Rozhkov, Tatiana Batura, Vladimir Ivanov, and Elena Tutubalina. 2022. Runne-2022 shared task: Recognizing nested named entities.

Baitong Chen, Ying Ding, and Feicheng Ma. 2018. Semantic word shifts in a scientific domain. *Scientometrics*, 117:211–226.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Data augmentation for cross-domain named entity recognition. *arXiv preprint arXiv:2109.01758*.

Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing.

Chen Jia and Yue Zhang. 2020. Multi-cell compositional lstm for ner domain adaptation. // *Proceedings of the 58th annual meeting of the association for computational linguistics*, P 5906–5917.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020. Zero-resource cross-domain named entity recognition. *arXiv preprint arXiv:2002.05923*.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. // *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, P 13452–13460.

Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Ilia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. 2021. Nerel: A russian dataset with nested named entities, relations and events. *arXiv preprint arXiv:2108.13112*.

Natalia Loukachevitch, Suresh Manandhar, Elina Baral, Igor Rozhkov, Pavel Braslavski, Vladimir Ivanov, Tatiana Batura, and Elena Tutubalina. 2023. Nerel-bio: a dataset of biomedical abstracts annotated with nested named entities. *Bioinformatics*, 39(4):btad161.

Hong Ming, Jiaoyun Yang, Lili Jiang, Yan Pan, and Ning An. 2022. Few-shot nested named entity recognition. *arXiv preprint arXiv:2212.00953*.

Takashi Shibuya and Eduard Hovy. 2020. Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. // I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. *arXiv preprint arXiv:1804.09021*.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2019. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2020. Adaptive self-training for few-shot neural sequence labeling. *arXiv preprint arXiv:2010.03680*.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2021. Meta self-training for few-shot neural sequence labeling. // *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, P 1737–1747.

Hang Yan, Yu Sun, Xiaonan Li, and Xipeng Qiu. 2022. An embarrassingly easy but strong baseline for nested named entity recognition.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.

Linyi Yang, Lifan Yuan, Leyang Cui, Wenyang Gao, and Yue Zhang. 2022. Factmix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. *arXiv preprint arXiv:2208.11464*.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 6470–6476, Online, July. Association for Computational Linguistics.

Zheng Yuan, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Fusing heterogeneous factors with triaffine mechanism for nested named entity recognition.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 7096–7108, Dublin, Ireland, May. Association for Computational Linguistics.