Improving Model Generalization: A Chinese Named Entity Recognition Case Study

Guanqing Liang, Cane Wing-Ki Leung
Wisers AI Lab, Wisers Information Limited
{quincyliang, caneleung}@wisers.com

Abstract

Generalization is an important ability that helps to ensure that a machine learning model can perform well on unseen data. In this paper, we study the effect of data bias on model generalization, using Chinese Named Entity Recognition (NER) as a case study. Specifically, we analyzed five benchmarking datasets for Chinese NER, and observed the following two types of data bias that may compromise model generalization ability. Firstly, the test sets of all the five datasets contain a significant proportion of entities that have been seen in the training sets. These test data are therefore not suitable for evaluating how well a model can handle unseen data. Secondly, all datasets are dominated by a few fat-head entities, i.e., entities appearing with particularly high frequency. As a result, a model might be able to produce high prediction accuracy simply by keyword memorization. To address these data biases, we first refine each test set by excluding seen entities from it, so as to better evaluate a model's generalization ability. Then, we propose a simple yet effective entity rebalancing method to make entities within the same category distributed equally, encouraging a model to leverage both name and context knowledge in the training process. Experimental results demonstrate that the proposed entity resampling method significantly improves a model's ability in detecting unseen entities, especially for company, organization and position categories.

1 Introduction

Named Entity Recognition (NER) is a fundamental building block for various downstream natural language processing tasks such as relation extraction (Bunescu and Mooney, 2005), event extraction (Ji and Grishman, 2008), information retrieval (Chen et al., 2015), question answering (Diefenbach et al., 2018), etc. Due to the ambiguous word boundaries and complex composition (Gui et al., 2019), Chinese NER task is more challenging compared with English NER.

Recently, by leveraging upon the pretrained language model (e.g, BERT (Devlin et al., 2018), etc.), we have witnessed superior performances on Chinese NER datasets, including: MSRA, Weibo, Ontonotes 4.0 and Resume (Li et al., 2020, 2019; Xuan et al., 2020). Despite the superior performance of the fine-tuned models, we argue that there are two types of data bias that can compromise the model generalization ability.

First, we observe that in widely used Chinese NER datasets, 50% to 70% entities in test data are seen in the training data. Such test data would therefore not be able to evaluate the true generalization ability of a model.

Second, the datasets are dominated by a few fat-head entities, i.e., entities appearing with particularly high frequency. For example, within the organization category of Cluener (Xu et al., 2020), fat-head entity 曼联 (Manchester United) appears 59 times, while 法兰克福队 (Eintracht Frankfurt) occurs only once. As a result, a model might be encouraged to memorize those fat-head entities rather than leveraging context knowledge during training process. The rationale is that given the same entity and diverse contexts, the easiest way for model convergence is to memorize the entity rather than extracting patterns from the diverse contexts.

To address these data biases, we first refine each test set by excluding seen entities from it, so as to better evaluate a model's generalization ability. Then, we propose a simple yet effective entity rebalancing method to make entities within the same category distributed equally, encouraging a model to leverage both name and context knowledge in the training process.

The contributions of this paper are as follows.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 992–997
August 1–6, 2021. ©2021 Association for Computational Linguistics

Dataset	Categories	Train	Dev	Test
MSRA	LOC, ORG, PER	41728	4636	4365
OntoNotes 4.0	GPE, LOC, ORG, PER	15724	4301	4346
Resume	CONT, EDU, LOC, NAME, ORG, PRO, RACE, TITLE	3821	463	477
Weibo	GPE.NAM, GPE.NOM, LOC.NAM, LOC.NOM, ORG.NAM, ORG.NOM, PER.NAM, PER.NOM	1350	270	270
Cluener	movie, organization, company, game, book, scene, name, government, address, position	10748	1343	1345

Table 1: Chinese NER datasets overview: entity categories and the sentence number in train/dev/test data.

- We analyze five benchmarking Chinese NER datasets and identify two types of data bias that can compromise model generalization ability.
- We refine each test set by excluding seen entities from it, which can measure real model generalization. Specifically, the competitive BERT+CRF model only achieves 33.33% and 65.10% F1 score on detecting unseen organization entities of Cluener and MSRA dataset respectively, which are far from satisfactory.
- We design a simple yet effective algorithm to rebalance the entity distribution. The experiments show that the proposed method significantly improves the model generalization. In particular, the F1 score has been improved by 12.61% and 37.14% on the organization category of Cluener and MSRA dataset respectively.

2 Dataset Observation

2.1 Dataset Overview

In this study, we analyze five benchmarking Chinese NER datasets, including: (1) MSRA (Levow, 2006), (2) Ontonotes 4.0 (Weischedel et al.), (3) Resume (Zhang and Yang, 2018), (4) Weibo (Peng and Dredze, 2015) and (5) Cluener (Xu et al., 2020). The statistics of these datasets are shown in Table 1.

2.2 Seen vs Unseen Entity

If an entity in dev/test data has been covered by the training data, we refer it as a seen entity. Otherwise, it is an unseen entity. To quantify the degree to which entities in the dev/test data have been seen in the training data, we define a measurement called entity coverage ratio. The entity coverage ratio of data D^{te} is denoted by $r(D^{te})$, which is calculated using the below equation.

$$r(D^{te}) = \frac{|Ent(D^{te}) \bigcap Ent(D^{train})|}{|Ent(D^{train})|} \quad (1)$$

where $Ent(.)$ denotes a function to obtain the list of annotated entities and D^{train} represents the training data. As Table 2 shows, the entity coverage ratios of the dev and test data in different benchmarking datasets are very high, ranging from 0.429 to 0.709.

Dataset	r(dev)	r(test)
MSRA	0.554	0.709
OntoNotes 4.0	0.505	0.514
Resume	0.540	0.544
Weibo	0.498	0.429
Cluener	0.615	-

Table 2: Entity coverage ratio of dev and test data in different Chinese NER datasets.

Observation 1 The test sets of Chinese NER datasets contain a significant proportion of seen entities.

2.3 Fat-head vs Long-tail Entity

Fat-head entity is defined as the entity appearing with particularly high frequency, while long-tail entity is defined as the entity with very few mentions. To identify the existence of fat-head entity, we use kurtosis (Balanda and MacGillivray, 1988), a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. Usually, high kurtosis (greater than 3) indicates the existence of outliers, i.e., fat-head entities.

Table 3 shows the kurtosis score of each category in different datasets. For example, the kurtosis score of PER category of training data in MSRA dataset is 984.1, which is very high. We find that 1% distinct entities with the highest frequency contribute 21% of the overall annotation.

Observation 2 Fat-head entities prevail in different categories of Chinese NER datasets.

We think this finding is also valid in other NER datasets, since the annotated corpus is usually collected within a certain time frame when some entities (e.g., celebrities, organizations) get much more exposure than others.

We hypothesize that the dominance of fat-head entities will cause the model to simply memorize

those high-frequency entities without fully leveraging context knowledge. The rationale is that given the same entity and diverse contexts, the easiest way for model convergence is to memorize the entity rather than extracting patterns from the diverse contexts.

Dataset	Training	Test
MSRA	ORG : 609.2 PER : 984.1 LOC : 1272.1	ORG : 66.3 LOC : 226.6 PER : 387.4
OntoNotes 4.0	LOC : 46.9 PER : 77.2 GPE : 108.9 ORG : 151.4	LOC : 36.9 PER : 79.0 GPE : 184.7 ORG : 337.5
Resume	LOC : 0.0 RACE : 4.2 EDU : 11.1 CONT : 11.8 PRO : 45.6 NAME : 59.6 TITLE : 239.3 ORG : 1723.1	CONT : 1.0 RACE : 1.0 LOC : 3.3 EDU : 4.5 PRO : 9.1 NAME : 35.6 TITLE : 53.5 ORG : 212.4
Weibo	GPE.NOM : 1.5 LOC.NAM : 6.4 LOC.NOM : 8.0 ORG.NOM : 9.9 GPE.NAM : 13.4 PER.NOM : 38.5 ORG.NAM : 101.1 PER.NAM : 188.4	GPE.NOM : 0.0 ORG.NOM : 1.7 GPE.NAM : 4.5 ORG.NAM : 5.0 LOC.NOM : 6.1 LOC.NAM : 6.6 PER.NOM : 27.7 PER.NAM : 48.0
Cluener	movie : 25.4 organization : 35.3 company : 81.4 game : 90.2 book : 97.5 scene : 117.4 name : 261.8 government : 308.2 address : 511.0 position : 570.0	-

Table 3: Kurtosis score of different categories in various Chinese NER datasets.

3 Method

To improve model's generalization ability in detecting unseen entities, we argue that the model should be trained to leverage both name and context knowledge (Nie et al., 2020; Lin et al., 2020). Thus, we propose a simple yet effective entity rebalancing algorithm. The main idea is to make the annotated entity equally distributed within the same category.

There are two major reasons why the proposed entity rebalancing algorithm works. First, the equal distribution will encourage the model to leverage both name knowledge and context knowledge, since there are no simple statistical cues (Niven and Kao, 2019) to exploit due to uneven distribution. Second, different entities within the same category should be interchangeable semantically in most cases, which avoids the train-test discrepancy.

The proposed algorithm works as follows. First, rebalance the annotated entity frequency in the training data. Let C_l denotes the original entity frequency counter of category l. For example, given $C_l = \{e_1 : 11, e_2 : 1, e_3 : 1\}$, which means entity e_1 is annotated 11 times, and both e_2 and e_3 are annotated once in the category l, which is very imbalanced. Then we turn C_l to the balanced entity frequency counter C_l^b, which is $C_l^b = \{e_1 : 5, e_2 : 4, e_3 : 4\}$. In C_l^b, the difference between the maximum and minimum entity frequency is 1 at most. Second, replace the fat-head entity with randomly sampled entity of the same category, once its accumulated occurrence surpasses the rebalanced frequency in C_l^b. Details are shown in **Algorithm 1**.

Algorithm 1: Entity replacement algorithm

```
foreach sentence in Dataset do
    foreach ent_text, ent_label in sentence do
        l = ent_label;
        if C_l^b[ent_text] > 0 then
            keep ent_text as it is;
            C_l^b[ent_text] -= 1;
        else
            sample ent_s from c_l^b if c_l^b[ent_s] > 0;
            replace ent_text with ent_s;
            C_l^b[ent_s] -= 1;
```

4 Experiments

4.1 Experiment Settings

According to observation 1, the test sets of Chinese NER datasets contain a significant proportion of seen entities, which fails to evaluate the true model generalization ability. In our study, the test sample will be excluded if it contains entities that are covered in training data. For Cluener (Xu et al., 2020), we split the original training set into 90% train and 10% dev, and use the development set for test, as the test set is not publicly available. For Resume (Zhang and Yang, 2018) and Weibo (Peng and Dredze, 2015) datasets, we report evaluation results on the selected categories, since there are zero or very few unseen entities on other categories.

We use the BIOES tagging scheme to label named entities, since previous studies have shown optimistic improvement with this scheme (Ratinov and Roth, 2009). We report span-level micro-averaged F1 score obtained from `seqeval` (Nakayama, 2018) toolkit using IOBES scheme.

We use BERT+CRF as the model architecture. In particular, we use bert-base-chinese pre-

trained model [1] (12-layer, 768-hidden, 12-heads) released by google (Devlin et al., 2018). The hyperparameters of the model are tuned on the development set using grid search method (details are reported in Appendix A.). As shown in Table 4, the adopted BERT+CRF model is competitive with the complicated state-of-the-art models.

Model	MSRA	OntoNotes	Resume	Weibo
Glyce-BERT (Meng et al., 2019)	95.54	81.63	96.54	67.60
BERT+FLAT (Li et al., 2020)	96.09	81.82	95.86	68.55
BERT+CRF (Ours)	95.57	**82.29**	95.71	**69.89**

Table 4: Comparison between BERT+CRF and the state-of-the-art models using the same train/dev/test splits as (Li et al., 2020, 2019; Xuan et al., 2020)

4.2 Results

Table 5 presents the comparisons between the proposed method and the baseline on five Chinese NER datasets. The baseline uses the original training data, while the proposed applies entity rebalancing algorithm on the original training data.

For Cluener, MSRA and OntoNotes datasets with over 10K training samples, our proposed method outperforms the baseline on different categories. One exception is on the address category of Cluener dataset when the proposed method performs worse than the baseline by -2.58%. We believe it is due to the fact that the address category contains both geopolitical entities and location entities, which are not interchangeable semantically.

For Weibo dataset, the proposed outperforms the baseline by 8.89% in PER.NAM category, but performs worse in PER.NOM category. Note that the PER.NOM category contains entities such as man, woman and friend, which are hard to generalize based on context knowledge. For Resume dataset, the proposed method does not work well. We think it is due to the structure of the resume corpus, which is the mere concatenation of name, education and organization, etc. Thus, there is very few context knowledge to leverage.

Overall, the proposed entity rebalancing method is able to improve model's generalization ability in detecting unseen entities. However, the proposed method only works for categories which meet certain conditions. First, the entities of the same category require to be interchangeable semantically. Second, the entities should be dependent of context knowledge.

Cluener			
Category	**Baseline**	**Proposed**	**F1 Improvement**
address	58.48	56.97	-2.58%
book	77.65	83.72	+7.82%
company	62.34	64.86	+4.04%
game	61.29	62.50	+1.97%
government	80.00	83.78	+4.72%
movie	71.91	75.61	+5.15%
name	74.38	75.81	+1.92%
organization	33.33	45.71	+37.14%
position	35.90	52.63	+46.60%
scene	74.56	78.31	+5.03%
MSRA			
Category	**Baseline**	**Proposed**	**F1 Improvement**
LOC	86.79	89.17	+2.74%
ORG	89.69	89.69	+0.00%
PER	95.85	96.35	+0.52%
OntoNotes 4.0			
Category	**Baseline**	**Proposed**	**F1 Improvement**
GPE	64.93	66.94	+3.10%
LOC	37.88	45.03	+18.88%
ORG	65.10	73.31	+12.61%
PER	96.45	96.32	-0.13%
Weibo			
Category	**Baseline**	**Proposed**	**F1 Improvement**
PER.NAM	69.09	75.23	+8.89%
PER.NOM	46.67	45.28	-2.98%
Resume			
Category	**Baseline**	**Proposed**	**F1 Improvement**
NAME	1.00	1.00	0%
ORG	90.62	87.88	-3.02%

Table 5: Evaluation results (F1 score) of the proposed entity resampling method and the baseline on unseen test data

5 Conclusion and Future Work

In this paper, we take Chinese NER as a case study, aiming to improve the model generalization by mitigating the data bias. We first refine each test set by excluding seen entities from it, so as to better evaluate a model's generalization ability. Then, we propose an entity rebalancing method to make entities within the same category distributed equally. Experimental results show that the proposed entity rebalancing method significantly improves a model's ability in detecting unseen entities.

As future work, we will first investigate the generalizability of this study to non-Chinese NER. Second, we will improve the entity replacement algorithm by leveraging language model so that the replaced entity is more semantically plausible.

[1] https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip.

References

K. Balanda and H. MacGillivray. 1988. Kurtosis: A critical review. *The American Statistician*, 42:111–119.

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2018. Core techniques of question answering systems over knowledge bases: A survey. *Knowl. Inf. Syst.*, 55(3):529–569.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuan-Jing Huang. 2019. A lexicon-based graph neural network for chinese ner. 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 1039–1049.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262, Columbus, Ohio. Association for Computational Linguistics.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. Association for Computational Linguistics.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842, Online. Association for Computational Linguistics.

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*.

Hongyu Lin, Yaojie Lu, Jialong Tang, Xianpei Han, Le Sun, Zhicheng Wei, and Nicholas Jing Yuan. 2020. A rigorous study on named entity recognition: Can fine-tuning pretrained model lead to the promised land? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7291–7300, Online. Association for Computational Linguistics.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. In *Advances in Neural Information Processing Systems*, volume 32, pages 2746–2757. Curran Associates, Inc.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, et al. Ontonotes release 4.0.

Liang Xu, Qianqian Dong, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. Cluener2020: Fine-grained name entity recognition for chinese. *arXiv preprint arXiv:2001.04351*.

Zhenyu Xuan, Rui Bao, Chuyu Ma, and Shengyi Jiang. 2020. Fgn: Fusion glyph network for chinese named entity recognition. *arXiv preprint arXiv:2001.05272*.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564, Melbourne, Australia. Association for Computational Linguistics.

Appendix A: Hyper-parameter Settings

Parameter	Value
learning_rate	[5e-5, 3e-5, 2e-5]
warmup_proportion	[0]
train_batch_size	[32]
seed	[2020, 1, 10]
crf_learning_rate	[1e-3, 1e-4]
model_name_or_path	["bert-base-chinese"]
max_seq_length	[128]
eval_batch_size	[16]
num_train_epochs	[10]
weight_decay	[0]
is_learning_rate_linearly_decrease	["yes"]

Table 6: The range of hyper-parameters grid-search for BERT+CRF model.