System Report of HW-TSC on the CAPITEL NER Evaluation

Lizhi Lei[a], Minghan Wang[a], Hao Yang[a], Shiliang Sun[b], Ying Qin[a] and Daimeng Wei[a]

[a]*Huawei Translation Service Center, Beijing, China*
[b]*East China Normal University, Shanghai, China*

Abstract

In this paper, we present participation in the Named Entity Recognition and Classification evaluation organized by CAPITEL at IberLEF 2020. We mainly aim to investigate the efficiency of data augmentation in transfer learning for NER. In addition to the official training set, we use Spacy to annotate the Spanish news monolingual corpus to create an augmented dataset. We perform experiments on two sets of data under 4 experimental settings, the experimental result show that pre-training on the augmentation set and then fine-tuning on the official set (i.e. cascadingly fine-tuning) could improve the performance compared to fine-tune on any of them separately or mixed.

Keywords

Named entity recognition, Pre-trained language models, Fine-tuning, Data augmentation

1. Introduction

The performance of the NER system has been greatly improved thanks to the introduction of pre-trained language models (PLM) as well as the pre-training and fine-tuning framework [1, 2, 3]. However, for some tasks with limited training data, it is still not be able to fully emerge the power of the pre-trained model. Therefore, researchers start to investigate better fine-tuning strategies.

In this paper, we mainly investigate the strategy of fine-tuning a PLM with data augmentation to improve the performance on the original task specific dataset. The official dataset is named as CAPITEL [4] which is composed by news articles in Spanish, and the augmented data is created by annotating WMT19 Spanish news corpus with Spacy. We will present details of our work and findings in following sections.

2. Model

Fine-tuning a pre-trained language model in downstream tasks has become a standard paradigm for many NLP tasks like sentiment classification, NER or POS-tagging, because models for these tasks can be easily designed as the combination of a unified encoder and a task specific classifier. [5]

Here we choose to use multilingual-BERT [1, 6] and distillBERT [7] as the encoder and a linear layer as the classifier. We use the hugging-face implementation released in the transformers library [5] and their pre-trained parameters. Experimental results show that the performance of multilingual-BERT is better than distillBERT therefore we only report results coming from the multilingual-BERT.

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
ORCID:

Table 1
The statistics of the dataset, where cap and aug represent for CAPITEL and augmented respectively.

	PER	LOC	ORG	OTH	sentence	token
train_cap	9,087	7,513	9,285	591,105	22,647	606,418
dev_cap	2,900	2,490	3,058	197,484	7,549	202,408
test_cap	2,996	2,348	3,143	194,730	7,549	199,773
train_aug	16,306	17,156	9,054	903,184	34,826	965,174
dev_aug	1,553	1,668	883	88,391	3,377	94,395

We use cross-entropy as the loss function. Note that BERT uses WordPiece tokenizer [8] to encode tokens, which may split a token into pieces of sub-tokens. In the hugging-face implementation, for token w which can be split into sub-tokens $w_{[0:c]}$, only the first sub-token w_0 is used to compute the loss, this doesn't affect un-split tokens and remains the distribution of each class unchanged.

3. Dataset

3.1. CAPITEL

As described by the organizers, the CAPITEL dataset is composed of Spanish news and has been annotated with Person (PER), Location (LOC), Organization (ORG), and Other (OTH). The corpus has been annotated with the BIOES format [9], where entities with a single token should be labeled as "S-ENT", otherwise should be labeled as "B/I/E-ENT" when there are multiple tokens, representing begin, inside and end of an entity. For example: Alex(S-PER) is(O) going(O) with(O) Marty(B-PER) A.(I-PER) Rick(E-PER) to(O) Los(B-LOC) Angeles(E-LOC).

3.2. Augmented

To create the augmented dataset, 38,000 sentences were sampled from the WMT news translation corpus. Then, we use Spacy to annotate the corpus and create the augmented dataset which contains 11,000 PER, 6,000 LOC, 4,000 ORG and 6,000 OTH. Note that the "MISC" annotation of Spacy is similar to the CAPITEL's OTH, so it can be directly converted.

Table 1 shows the detail of two datasets employed in experiments, where the cap and aug represents for the CAPITEL and the augmented set respectively. From the table we can see that the distribution of each entity type for CAPITEL and augmented is different, which means the augmented set should be used carefully to prevent from introducing such bias.

4. Experiment

Basically, our experiments are conducted under four settings:

- **Cap** Fine-tuning on all of the official training data.
- **Aug** Fine-tuning on all of the augmented training data.
- **Mix** Fine-tuning on the mixture of 10000 augmented data and all of the official data.
- **Cascade** Pre-training on the all of the augmented training data and fine-tuning on the official set.

Table 2
The table of the experimental results.

		Cap	Aug	Mix	Cascade
PER	**P**	93.90	79	93.44	94.83
	R	95.55	84	94.24	95.45
	F1	94.72	81	93.84	95.14
LOC	**P**	88.23	76	85.41	88.54
	R	88.84	77	88.84	89.64
	F1	88.53	77	87.09	89.08
ORG	**P**	83.02	68	82.17	84.49
	R	86.82	78	84.27	86.59
	F1	84.88	73	83.21	85.53
OTH	**P**	79.94	57	75.35	79.04
	R	76.53	59	76.36	79.00
	F1	78.20	58	75.85	79.02
Micro avg.	**P**	86.87	72	84.95	87.45
	R	87.99	75	86.83	88.52
	F1	87.43	74	85.88	87.99
Macro avg.	**P**	86.84	72	84.96	87.46
	R	87.99	75	86.83	88.52
	F1	87.39	73	85.88	87.99

The first experiment can be considered as a strong baseline, which shows that fine-tuning a PLM with in-domain data is already able to achieve an acceptable performance.

The second experiment shows that fine-tuning solely on the augmented set could have a negative impact on the performance. We consider that the distribution of each entity in the CAPITEL and augmented data is different which could be a reason. Another reason might comes from the noise of the annotation. Unlike some data augmentation method achieved by corrupting input features which doesn't introduce bias on the mapping of X to Y, Spacy annotation is not the golden truth, we tried to use the Spacy to tag the CAPITEL training set which achieves 51.18%, 72.03% 57.79%, 50.43% of F1 score for the PER, LOC, ORG and the Macro respectively. This means the augmented data could mislead the model to learn incorrect pattern, thus should be used carefully.

The third experiment aims to evaluate the performance of fine-tuning on the mixture of two set. Unfortunately, same as the second experiment, the noise of the augmented data still brings negative influence on the performance, although could be fixed to a certain extent by the correct pattern in the official data, which means that the mixture paradigm only applies to high-quality augmented data.

Despite the performance was unexpected for the second experiment, we decided to continue experiments fine-tuning the trained model with the CAPITEL training set. We can see that further fine-tuning on the unbiased clean data successfully fix the biased estimation of the noisy augmented data, at the same time, the performance improvement might comes from the more generalized knowledge learned from the correct annotation of the augmented data.

5. Conclusion

In this paper, we present our work in the CAPITEL NER evaluation and investigate the method of improving performance with transfer learning and data augmentation. We find that cascadingly fine-tuning a pre-trained model on the augmented set and official set could significantly improve the performance. Our submission is based on this strategy and achieves the third place.

References

[1] J. Devlin, M. Chang, K. Lee, K. Toutanova, 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423/.

[2] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 2145–2158. URL: https://www.aclweb.org/anthology/C18-1182.

[3] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, CoRR abs/1812.09449 (2018). URL: http://arxiv.org/abs/1812.09449. arXiv:1812.09449.

[4] J. Porta-Zamorano, L. Espinosa-Anke, Overview of CAPITEL Shared Tasks at IberLEF 2020: NERC and Universal Dependencies Parsing, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), 2020.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's Transformers: State-of-the-art natural language processing, CoRR abs/1910.03771 (2019). URL: http://arxiv.org/abs/1910.03771. arXiv:1910.03771.

[6] K. K, Z. Wang, S. Mayhew, D. Roth, Cross-lingual ability of multilingual BERT: an empirical study, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020. URL: https://openreview.net/forum?id=HJeT3yrtDr.

[7] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR abs/1910.01108 (2019). URL: http://arxiv.org/abs/1910.01108. arXiv:1910.01108.

[8] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's neural machine translation system: Bridging the gap between human and machine translation, CoRR abs/1609.08144 (2016). URL: http://arxiv.org/abs/1609.08144. arXiv:1609.08144.

[9] J. P. C. Chiu, E. Nichols, Named entity recognition with bidirectional LSTM-CNNs, Trans. Assoc. Comput. Linguistics 4 (2016) 357–370. URL: https://transacl.org/ojs/index.php/tacl/article/view/792.