The 6th International Conference on Interdisciplinary research on Computer science, Psychology, and Education (ICICPE' 2022)
December 26-28, 2022. Pattaya, Thailand.

An Analysis of Korean Named Entity Recognition System using MLM-based Language Transfer Learning

Junyoung Son, Gyungmin Kim, Jinsung Kim, Aram So*

Department of Computer Science and Engineering, Korea University
South Korea
{s0ny, totoro4007, jin62304, aram}@korea.ac.kr

Abstract. With the advance of pre-trained language models, Named Entity Recognition (NER) systems have shown high performance. However, because there is not enough labeled data for most languages, including Korean, to learn the specific patterns for entities, it has been a challenging problem. In this paper, we analyze what language is the best for transferring from their trained knowledge to Korean in a low-resource setting. Specifically, we utilize an MLM-based language transfer learning approach, where target language is treated as a low-resource, sharing common prompt label words between languages. Our experiments on ten different languages demonstrate the effectiveness of our approach, outperforming fine-tuned model in our language transfer setting. In addition, we observe that syntactically similar languages, such as Persian which is an agglutinative language can be used to augment Korean.

Keywords: Named Entity Recognition, NER, Language Transfer, Prompt, MLM

1 Introduction

Named entity recognition (NER) is a task that seeks to locate and classify named entities in a text into predefined types, such as a person, location, etc. With the advance of pre-trained language models (PLMs), NER systems have shown high performance [1,4]. However, most of languages usually face data scarce problems, unlike high-resource languages such as English [2].

In this paper, we perform language transfer from ten languages in the MultiCoNER [3] to Korean in a low-resource scenario. Then, we analyze the correlation between the languages and Korean. Specifically, we utilize an MLM-based language transfer learning, where target language is treated as a low-resource, sharing common prompt label words between languages.

* Corresponding author

2 Problem Statement

We hypothesize the source languages are high-resource languages, while target languages are low-resource languages. Given training datasets of the source and target languages, D_S and D_t, we assume only K examples for each entity class in D_t. After we train model using D_S, we transfer the model using Ds. Then, the model is evaluated with an unseen test set D_t^{test}.

3 Approach

We employ MLM-based language transfer learning to analyze the correlation between languages. Specifically, we add virtual label words treated as entity classes to the model's vocabulary. Then, we train these virtual prompt words and the model using a source language and transfer it to a target language. For example, a prompt word '*[PER]*' is trained in the model's vocabulary to replace the superficial label '*person*.' Note that this prompt words can be utilized as language-independent features in out approach.

4 Experiments

4.1 Experimental settings

To get advantage of multilingual, we use MultiCoNER [3], a large multilingual dataset for NER that covers 11 languages, including Bangla (BN), Hindi (HI), German (DE), Chinese (ZH), Korean (KO), Turkish (TR), Dutch (NL), Russian (RU), Farsi (FA), English (EN), and Spanish (ES).

4.2 Experimental results

Table 1. Experimental results on the language transfer.

K	BN→ KO	DE→ KO	EN→ KO	ES→ KO	FA→ KO	HI→ KO	NL→ KO	RU→ KO	TR→ KO	ZH→ KO
K=0	26.68	23.88	26.43	23.29	**29.78**	28.24	27.48	19.07	27.98	27.64
K=5	29.35	28.58	29.13	29.68	**31.08**	29.49	30.12	29.44	28.68	29.85
K=10	33.19	33.63	35.23	33.69	33.73	**34.24**	33.98	33.91	32.98	32.87
K=20	37.73	36.59	38.26	37.30	37.64	38.00	37.85	37.45	36.98	**38.50**
K=50	42.52	42.71	42.55	43.06	41.23	43.15	41.79	42.36	42.46	**44.13**

The experimental results of language transfer on a low-resource setting is shown in Table 1. First, we observe that FA→KO shows the best performance when data is extremely scarce. The first thing to notice is that Farsi is an agglutinative language that is similar to Korean. When the amount of data gets larger, ZH→KO rapidly improves, especially *K=20*. We attribute this to common feature of Chinese characters between Korean and Chinese.

Table 2. Experimental results on a full-shot language transfer setting.

Method	BN→ KO	DE→ KO	EN→ KO	ES→ KO	FA→ KO	HI→ KO	NL→ KO	RU→ KO	TR→ KO	ZH→ KO
Fine-tuning	27.70	24.66	24.72	28.22	27.20	**29.79**	23.45	22.81	22.45	25.65
MLM-based	55.72	55.72	55.49	55.92	55.50	55.71	55.72	55.20	55.58	**56.41**

Table 2 shows the performance in a full-shot language transfer setting. We observe that MLM-based language transfer using prompt words also gets the advantage in a full-shot language transfer setting, outperforming fine-tuning model's performance.

5 Conclusion

In this paper, we study the correlation between languages and Korean by using MLM-based language transfer learning with entity prompt words. Our experiments on ten different languages demonstrate the effectiveness of our approach, outperforming fine-tuned model in our language transfer setting. In addition, we observe that syntactically similar languages, such as Persian which is an agglutinative language can be used to augment Korean. We hope that our study promotes low-resource language transfer learning in NER task.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques), the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

References

1. KIM, Gyeongmin, et al. Enhancing Korean Named Entity Recognition With Linguistic Tokenization Strategies. IEEE Access, 2021, 9: 151814-151823.
2. BARI, M. Saiful; JOTY, Shafiq; JWALAPURAM, Prathyusha. Zero-resource cross-lingual named entity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020. p. 7415-7423.
3. MALMASI, Shervin, et al. Multiconer: a large-scale multilingual dataset for complex named entity recognition. arXiv preprint arXiv:2208.14536, 2022.
4. KIM, Jin-Sung, et al. An Effective Segmentation Scheme for Korean Sentence Classification tasks. In: Annual Conference on Human and Language Technology. Human and Language Technology, 2021. p. 173-177.