Building Korean NER models for a manually annotated corpus from clinical notes using cross-lingual transfer learning

Jianfu Li[1*], Jimyung Park[2*], Xinyue Hu[1], Jingqi Wang[3], Rae Woong Park[2, 4], Hua Xu[1]

[1]Shool of Biomedical Informatics, University of Texan Health Science Center at Houston, Texas, United States; [2]Dept. of Biomedical Sciences, Ajou University School of Medicine, Yeongtong-gu, Suwon, South Korea; [3]Melax Technologies Inc, Houston, Text, United States; [4]Dept. of Biomedical Informatics, Ajou University School of Medicine, Yeongtong-gu, Suwon, South Korea

*** These authors contributed equally to the work.**

Background

More and more clinical notes generated these days which contain very important clinical concepts like Problem, Tests, and Treatments. It is very important to develop efficient and accurate information extraction tools and methods to unlock structured information from this growing amount of raw unstructured text for use in clinical natural language processing research. Named entity recognition (NER) is one of the fundamental clinical NLP tasks and has long attracted much attention from researchers. In the medical domain, a lot of studies have been explored on NER in English clinical notes;[1] however, very limited NER research has been carried out on clinical notes written in Korean. The goal of this study was to systematically investigate cross-lingual transformer models for NER in Korean clinical notes.[2,3,4]

Methods

Three state-of-the-art pre-trained transformer-based language models were explored: BERT-KOR$_{BASE}$, BERT-MULTILINGUAL$_{BASE}$, and XLM-ROBERTA$_{BASE}$, where BERT-KOR$_{BASE}$ was a monolingual language model that was pre-trained on a 70GB Korean text dataset, BERT-MULTILINGUAL$_{BASE}$ was a multilingual BERT model that was pre-trained on 104 languages with the largest Wikipedias dataset, and XLM-ROBERTA$_{BASE}$ was a multilingual language model that was pre-trained on 100 languages with 2TB CommonCrawl dataset. An additional BI-LSTM-CRF model was trained and taken as baseline.

A total of 500 discharge summaries were randomly selected from a Korean tertiary hospital database, Ajou University School of Medicine (AUSOM), for this study. The clinical notes were manually reviewed and annotated with Problems, Tests, and Treatments entities using CLAMP (Clinical Language Annotation, Modeling, and Processing toolkit) according to a predefined guideline.[5] The annotated corpus was preprocessed with sentence boundary detection and tokenization and then transformed into the “BIO” format for training, where “B” represents the beginning of an entity, “I” represents tokens inside an entity, and “O” represents all other nonentity tokens.

The NER task was formulated as a sequence labeling task and performed by fine tuning the pre-trained transformer models using a linear classification layer to predict token tags using the training corpora. A 5-fold cross-validation (train/dev/test subsets with a ratio of 60%:20%:20%) was used to train and evaluate the performance of the NER models. The performance of all the NER models were evaluated using both the strict and relaxed micro precision, recall, and F1-score.

Results

Table 1 shows the strict and relaxed micro P/R/F1 scores of 4 NER models. It shows that the multilingual XLM-ROBERTA$_{BASE}$ model achieved the best performance (strict F1-score 0.821) and both XLM-ROBERTA$_{BASE}$ and BERT-MULTILINGUAL$_{BASE}$ models outperform the baseline BI-LSTM-CRF model, while BI-LSTM-CRF model outperforms BERT-KOR$_{BASE}$ model.

Table 1. The strict and relaxed overall performances (P/R/F1) on the test sets of the corpus. Numbers in the parentheses are results based on the relaxed matching criteria. The boldface represents the best performance on Precision, Recall and F1-score.

	P	R	F1
BERT-KOR$_{BASE}$	0.741(0.829)	0.779(0.874)	0.760(0.851)
BERT-MULTILINGUAL$_{BASE}$	0.787(0.861)	0.816(0.894)	0.801(0.877)
XLM-ROBERTA$_{BASE}$	**0.808(0.878)**	**0.835(0.908)**	**0.821(0.893)**
BI-LSTM-CRF	0.782(0.837)	0.792(0.870)	0.786(0.853)

Conclusion

In this study, we systematically explored the pre-trained language models for NER and built Korean NER models using clinical notes. Our results showed that the multilingual XLM-ROBERTA achieved the best performance on P/R/F1, which demonstrated the potential advantages of multilingual language models on cross-lingual task such as NER compared to monolingual language model. In the future analysis, we will utilize the cross-lingual corpus to improve the current NER models.

Acknowledgement

This research was partially supported by the National Institutes of Health under Award Numbers U01TR002062, R01AI130460, R01GM103859, and U24CA194215. This research was partially supported by a grant of the project for Infectious Disease Medical Safety, funded by the Ministry of Health, Republic of Korea (grant number: HG22C0024).

Conflict of Interest Statement

HX and The University of Texas Health Science Center at Houston have financial related interests at Melax Technologies Inc.

References/Citations

1. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011 Sep 1;18(5):552-6.
2. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
3. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116. 2019 Nov 5.
4. Kiyoung Kim. Pretrained language models for Korean. GitHub: https://github.com/kiyoungkim1/LMkor, 2020.