UA-KO at SemEval-2022 Task 11: Data Augmentation and Ensembles for Korean Named Entity Recognition

Hyunju Song, Steven Bethard
University of Arizona
{hyunjusong,bethard}@email.arizona.edu

Abstract

This paper presents the approaches and systems of the UA-KO team for the Korean portion of SemEval-2022 Task 11 on Multilingual Complex Named Entity Recognition. We fine-tuned Korean and multilingual BERT and RoBERTA models, conducted experiments on data augmentation, ensembles, and task-adaptive pretraining. Our final system ranked 8th out of 17 teams with an F1 score of 0.6749 F1.

1 Introduction

Named Entity Recognition (NER) is the task of recognizing and classifying named entities in unstructured text. NER is a critical component for NLP tasks such as question answering and relation extraction. Recent advances in neural NER have allowed state-of-the-art systems to perform well in recognizing persons, locations, and organizations in a variety of benchmark datasets. However, these systems still struggle to recognize complex named entities such as titles of creative works. SemEval Task 11 (Malmasi et al., 2022b) asks participants to build systems to identify both classic and complex named entities in multiple languages.

In this paper, we describe the UA-KO team's approach to deal with the challenge of recognizing complex named entities in Korean. We used both monolingual and multilingual transformer-based models. To improve the performance of the models, we conducted experiments on data augmentation, ensembles, and task-adaptive pretraining. Our final performance on the test set ranked 8th out of 17 teams with an F1 score of 0.675. Code to replicate our experiments is available at https://github.com/hyunssong/semeval2022-task11.

2 Related Work

NER in real-world settings is challenging. Identifying entities in short texts, such as search queries on the web, is not easy due to the lack of context. Identifying nontraditional named entities such as titles of creative works (movies, books, or TV shows) is challenging because similar phrases appear as non-named entities and the characteristics of creative work titles change over time (Ashwini and Choi, 2014). Approaches to such real-world named entity types include integrating English Wikipedia as a gazetteer within neural NER models (Meng et al., 2021), though simply relying on the robustness of large pre-trained models like XLM-R is more successful (Ushio and Camacho-Collados, 2021).

SemEval-2022 Task 11 asks participants to build systems that can perform well in real-world settings, providing a dataset with complex named entities in short sentences. The task is divided into three tracks: multi-lingual, monolingual, and code-mixed. We participated in the Korean portion of the monolingual track.

Korean is an agglutinative language, where each word (*eojeol*) is formed by combining a root morpheme with a bound morpheme (*josa*) or postposition (*eomi*) (Choi et al., 2017). More than 60 different forms can thus be created from each root morpheme. A model that is not aware of this morphological richness may end up significantly increasing the vocabulary size to represent all these different forms, at the costs of having only sparse data to learn each form and encountering a high rate of out-of-vocabulary (OOV) forms.

Recent studies have explored different representations to account for the agglutinative characteristics of Korean. Lee et al. (2020) explored both the syllable level and sub-character level representations of the text, achieving similar results to multilingual BERT with 1/10 of the training data. Kwon et al. (2017) proposed a deep learning based NER system that operates over syllables rather than words, resulting in a speedup by removing the need for morphological analysis. Kim et al. (2021) explored morpheme, syllable, and subcharacter rep-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1608 - 1612

Train	Dev	Test
15300	800	150K~500K

Table 1: The Korean language portion of the SemEval-2022 Task 11 data.

resentations and also found that syllables were the most effective representation for Korean NER.

3 Data

The SemEval-2022 Task 11 dataset follows the CoNLL format. Each token is classified into six entity types: person, location, group, corporation, product, or creative work. The size of the Korean language portion of the dataset is shown in table 1. More details about the dataset can be found in Malmasi et al. (2022a). Systems applied to this dataset are evaluated based on the macro-average F1 score.

Our data augmentation experiments (see section 4.2) utilized additional data. For locations, we used the GeoNames geographical database, which provides countries and place names in many different languages. We selected the 170 country names provided in the Korean language. For person names, we used the Encyclopedia of Korean Culture (of Korean Studies, 1989) , which is a Korean language encyclopedia. We selected their 18,506 names of Korean historical figures.

4 Methodology

We used both monolingual and multilingual pretrained models for the task. For the monolingual models, we used KoBERT[1], KR-BERT (Lee et al., 2020), Ko-ELECTRA[2], and KLUE-RoBERTa-large(Park et al., 2021) , which all have shown successful performance on Korean NLP tasks. For the multilingual model, we used XLM-RoBERTa-large (Conneau et al., 2019), which was provided as the baseline of the shared task.

We fine-tuned these models on the training set. Monolingual models were fine-tuned with a learning rate of 5e-5 for 5 epochs with a weight decay of 0.01. The multilingual model was trained with the same configuration but for 3 epochs. We also experimented with different methodologies to improve performance: task-adaptive pretraining, data augmentation, and ensembles. These approaches are described in the following sections.

[1] https://github.com/SKTBrain/KoBERT
[2] https://github.com/monologg/KoELECTRA

4.1 Task Adaptive Pretraining

Gururangan et al. (2020) showed that further pretraining models on the unlabeled task data, called task-adaptive pretraining, can improve model performance when the dataset is curated to capture language appropriate to the task. We follow the training environment as Gururangan et al. (2020). To avoid catastrophic forgetting, we additionally reduced the number of epochs following Zhao et al. (2021)'s approach of applying task-adaptive pretraining on small datasets. Specifically, we conduct task-adaptive pretraining with a batch size of 256 for 50 epochs. For the KoELECTRA model, we further train for 7k steps. XLM-R was not included in these experiments due to computing resource limitations.

4.2 Data Augmentation

Data augmentation is a strategy that enhances the amount of training data by modifying existing data or generating new synthetic data, usually by leveraging external resources. Dai and Adel (2020) proposed several data augmentation techniques for NER tasks, including generating sentences by replacing tokens or shuffling within sentence segments. Since the provided training dataset is small relative to the test set, we decided to conduct data augmentation to reduce the overfitting of the model on the training dataset. We specifically focused on person and location named entities as these entities had the highest percentage in the test set.

We generated new sentences by utilizing the external database described in section 3. For each of the sentences that contain location or person named entity, we generated k new sentences where $k = 3$ or 6. New sentences were generated by replacing the location named entity with Korean country names from GeoNames database or replacing the person named entity with historic person names from the Encyclopedia of Korean Culture. We then fine-tuned the language models on the original sentences together with these augmented sentences.

4.3 Ensembles

Ensemble methods are an effective way of combining multiple machine-learning models to make better predictions (Rokach, 2010). We created ensembles over the different monolingual and multilingual models using soft voting, which predicts the class label based on the argmax of the sums of

Model	Modification	F1
KoBERT	-	0.808
KoBERT	TAPT	**0.817**
KoBERT	Augment $k = 3$	0.816
KoBERT	Augment $k = 6$	0.806
KR-BERT	-	0.824
KR-BERT	TAPT	0.826
KR-BERT	Augment $k = 3$	**0.834**
KR-BERT	Augment $k = 6$	0.830
KoELECTRA	-	0.827
KoELECTRA	TAPT	0.767
KoELECTRA	Augment $k = 3$	**0.829**
KoELECTRA	Augment $k = 6$	0.827
KLUE	-	**0.850**
KLUE	TAPT	0.832
KLUE	Augment $k = 3$	0.846
KLUE	Augment $k = 6$	0.846
XLM	-	**0.831**
XLM	Augment $k = 3$	**0.831**
XLM	Augment $k = 6$	0.823
Ensemble(KLUE, XLM)	-	0.858
Ensemble(KLUE, XLM)	Augment $k = 3$	0.855
Ensemble(All Korean)	-	0.863
Ensemble(All Korean)	Augment $k = 3$	0.866
Ensemble(All Korean, XLM)	-	0.864
Ensemble(All Korean, XLM)	Augment $k = 3$	**0.868**

Table 2: Performance of different models on the Dev set. "All Korean" stands for all Korean monolingual models: KoBERT, KR-BERT, KoELECTRA, and KLUE. The best scoring system in each group is in bold.

the predicted probabilities of the various classifiers.

5 Results on Dev

Table 2 shows performance of task adaptive pretraining (TAPT), data augmentation (Augment $k = N$), and ensembles (Ensemble(...)) on the development set. Task adaptive pretraining yielded little benefit over the corresponding unadapted model, and sometimes dramatically worsened performance (e.g., the KoELECTRA model went from 0.827 to 0.767). Data augmentation either led to small gains over the non-augmented model or to roughly the same performance, and $k = 3$ was generally as good or better than $k = 6$. Ensembling led to consistent gains compared to the single models. The best overall model on the development set was an ensemble of KoBERT, KR-BERT, KoELECTRA, KLUE, and XLM combined with data augmentation where $k = 3$.

Given these results, for our official submissions on the test set, we applied data augmentation with $k = 3$, and we included several types of ensembles. We did not apply any task adaptive pre-training because it showed no benefits on the development data.

Model	F1
KLUE	0.651
KLUE, Augment $k = 3$	0.650
Ensemble(All Korean)	0.668
Ensemble(All Korean), Augment $k = 3$	0.626
Ensemble(All Korean, XLM)	**0.675**
Ensemble(All Korean, XLM), Augment $k = 3$	0.650

Table 3: Performance of different models on Test set. "All Korean" stands for all Korean monolingual models: KoBERT, KR-BERT, KoELECTRA, and KLUE. The best scoring system is in bold.

Class	Precision	Recall	F1
LOC	0.689	0.788	0.735
PER	0.776	0.748	0.761
PROD	0.706	0.665	0.685
GRP	0.678	0.595	0.634
CW	0.526	0.563	0.544
CORP	0.688	0.693	0.691

Table 4: Performance by class label for the best performing model, the ensemble of KoBERT, KR-BERT, KoELECTRA, KLUE, and XLM. Results for the other models look qualitatively similar.

6 Results on Test

Table 3 shows the performance of our submitted models on the test set. Unlike our results on the development set, data augmentation significantly reduced performance on the test set. For the runs without data augmentation, ensembles outperformed single models as in our development set results. The failure of data augmentation may imply a low overlap between the names we drew from GeoNames and the Encyclopedia of Korean Culture for data augmentation, and the named entities in the test data. That is, the coverage of these gazetteers may have been insufficient for the unique and nontraditional named entities of the test set, similar to problems mentioned in Meng et al. (2021).

Table 4 shows detailed performance of our best-performing model, the ensemble of KoBERT, KR-BERT, KoELECTRA, KLUE, and XLM. This model performs best in recognizing person names (0.761 F1), but has difficulty recognizing creative works (0.544 F1).

dataset	sentence	label
dev	(A1) 《**유희열의 스케치북**》/은/한국방송공사/음악/전문/텔레비전/ 프로그램이다. (A1) 《**Yu Hee-yeol's Sketchbook**》/is a/KBS/music/television/show.	CW
dev	(A2) 1993년/ 매크래컨은/ 해나/ 바베라/ 카툰스의/ 애니메이션/ 시리즈/ 《2 stupid dogs》/의 / 미술/ 감독으로/ 일했다. (A2) In 1993/ McCracken/ Hanna/ Babera/ Cartoons/ Animation/ series/ 《2 stupid dogs》 /'s/ art/ director/ worked.	CW
dev	(A3) 정도가/ 약한/ 경우는/ 완전히/치료하지는/ 않으며/ **일반 화장품**/ 간단히/ 감출/ 수/있다. (A3) Degree/weak/ in the case of/ completely/ cure/ don't/ **general cosmetics**/ simply/ cover/ is/ possible.	PROD
dev	(A4) 제 18회/ (1987년) / 영화/《 브라질 》/ **테리**/ 길리엄/ (A4) 18th/ (1987) / Film /《Brazil》/ **Terry**/ Gilliam	PER
test	(B1) 는/ 은여울역/ 카운티입니다. (B1) is/ Eunyeoul Station/ County.	- -
test	(B2) 에/ 로그인/ 텔레페 (B2) to/ login/ Telefe	- -

Table 5: Example sentences from the shared task data. Gold annotations are in bold. Model predictions are underlined.

6.1 Error Analysis

We analyze our best-performing system's predictions on the dev set to understand our system's strengths and weaknesses. By investigating the errors where the model is highly confident, we identified the following qualities from the errors on the dev set.

Annotation errors: We observed inconsistent labeling in named entities, especially in the creative work named entities. The development data contained many creative works enclosed in 《 and 》, such as example A1 in table 5. However the development data also contained creative works marked by the same punctuation that were not labeled as creative works by annotators, such as example A2. This led to a high error rate in predicting creative work named entities on the development data, so we were unsurprised to see similar low performance on the test data creative works. We speculate that the reason data augmentation was not helpful on the test data was these errors in annotation.

Token boundary issues: We see that often tokens are misclassified due to token boundary issues. For instance, in example A3 in table 5, the system found the product named entity but included an extra token before the start of the product name.

Foreign names: The model had difficulty recognizing transliterated named entities, such as names of foreign people or groups as in example sentence A4 of table 5. These names are different from traditional Korean words, likely leading to the system's difficulty in identifying them.

Grammatical errors: The sentences of the test set differ grammatically from the training and dev set. Table 5 shows some test set inputs that are incomplete (B1) or have grammatical errors (B2). B1 is missing the subject of the sentence, and B2 does not follow the subject-object-verb order of a Korean sentence structure. We found such grammatical problems to be frequent in the test set. As such grammatical problems were not frequently present in the development data, our models were not robust to them.

7 Conclusion

We have presented a description of our different approaches for identifying complex named entities in Korean language data. A monolingual model that considers characteristics of the Korean language performs well, and an ensemble of monolignual models and a multilingual model further improves performance. Though we also explored task-adaptive pretraining and data augmentation, task-adaptive pretraining did not help on the development data, and data augmentation helped on the development data but hurt on the test data. Our results suggest that while ensembles yield reliable gains for Korean named entity recognition, further research is needed to utilize external knowledge when dealing with complex named entity recognition.

References

Sandeep Ashwini and Jinho D Choi. 2014. Targetable named entity recognition in social media. *arXiv preprint arXiv:1408.0782*.

Sanghyuk Choi, Taeuk Kim, Jinseok Seol, and Sanggoo Lee. 2017. A syllable-based technique for word embeddings of Korean words. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 36–40, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Gyeongmin Kim, Junyoung Son, Jinsung Kim, Hyunhee Lee, and Heuiseok Lim. 2021. Enhancing korean named entity recognition with linguistic tokenization strategies. *IEEE Access*, 9:151814–151823.

Sunjae Kwon, Youngjoong Ko, and Jungyun Seo. 2017. A robust named-entity recognition system using syllable bigram embedding with eojeol prefix information. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2139–2142, New York, NY, USA. Association for Computing Machinery.

Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020. KR-BERT: A small-scale korean-specific language model. *arXiv preprint arXiv:2008.03979*.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512, Online. Association for Computational Linguistics.

Academy of Korean Studies. 1989. Encyclopedia of korean culture. *Acad Korean Stud*, 10:704.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Lior Rokach. 2010. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. *A Comparative Study of Using Pre-Trained Language Models for Toxic Comment Classification*, page 500–507. Association for Computing Machinery, New York, NY, USA.