Received: 24 August 2023 | Revised: 15 November 2023 | Accepted: 20 December 2023

DOI: 10.4218/etrij.2023-0321

SPECIAL ISSUE

ETRI Journal WILEY

# Named entity recognition using transfer learning and small human- and meta-pseudo-labeled datasets

**Kyoungman Bae** | **Joon-Ho Lim**

Language Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

**Correspondence**
Kyoungman Bae, Language Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.
Email: kyoungman.bae@etri.re.kr

**Funding information**
Institute for Information and Communications Technology Promotion, Grant/Award Numbers: 2013-2-00131, 2022-0-00369

**Abstract**

We introduce a high-performance named entity recognition (NER) model for written and spoken language. To overcome challenges related to labeled data scarcity and domain shifts, we use transfer learning to leverage our previously developed KorBERT as the base model. We also adopt a meta-pseudo-label method using a teacher/student framework with labeled and unlabeled data. Our model presents two modifications. First, the student model is updated with an average loss from both human- and pseudo-labeled data. Second, the influence of noisy pseudo-labeled data is mitigated by considering feedback scores and updating the teacher model only when below a threshold (0.0005). We achieve the target NER performance in the spoken language domain and improve that in the written language domain by proposing a straightforward rollback method that reverts to the best model based on scarce human-labeled data. Further improvement is achieved by adjusting the label vector weights in the named entity dictionary.

**KEYWORDS**
domain adaptation, KorBERT, meta pseudo-label, named entity recognition, transfer learning

## 1 | INTRODUCTION

Named entity recognition (NER) is a fundamental task in information extraction that focuses on locating and categorizing named entities from unstructured text into predefined classes, such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, and percentages [1]. NER can be formulated as a sequence labeling problem to assign an appropriate label to each word within a sentence [2]. In recent years, considerable research effort has been devoted to developing end-to-end neural-based sequence labeling models for NER [3–6]. In particular, neural network architectures based on pretrained language models have exhibited remarkable performance in single-domain NER [7–9]. Nevertheless, these models face challenges, such as dependency on large training datasets to prevent overfitting and considerable performance degradation under domain shifts [10]. Acquiring sufficient training data for new domains can be time-consuming and costly when constructing human-labeled data.

We aimed to train a high-performing NER model in both the written and spoken language domains. While written language had abundant human-labeled data (approximately 250 000 sentences), spoken language faced data scarcity with approximately 25 000 human-labeled samples. To address this problem, we adopted

transfer learning and used the Korean-specific KorBERT as the base model for NER in spoken language.

Transfer learning has emerged as a promising approach to handle data scarcity and domain shifts in various applications [11]. Its primary objective is to enable a target model to adapt swiftly to new domains using limited or automatically generated training data, thereby avoiding the need to retrain from scratch [12]. By employing domain adaptation, a subtype of transfer learning, we aimed to bridge the domain gap between written and spoken language, transferring pertinent knowledge from a well-resourced written language model to a spoken language model.

In this study, we adopted the meta-pseudo-label (MPL) method for domain adaptation in NER. The MPL method can enhance the classification performance on the ImageNet dataset and employs semi-supervised learning that leverages both labeled and unlabeled data within a teacher/student framework [13]. The student model learns from a minibatch of pseudo-labeled data annotated by the teacher model and human-labeled data, whereas the teacher model learns by applying a reward signal (feedback signal) that reflects the student model performance on a minibatch drawn from a labeled dataset.

In natural language processing, MPLs have been employed to complement human-labeled data. For example, a student model has been trained by evaluating the quality of pseudo-labeled data alongside human-labeled data [14]. In He et al. [15], a feedback score has been obtained to measure the improvement after updating a student model with labeled data. During the teacher model update, the feedback score has been used in the loss of pseudo-labeled data. Our proposed model differed from these approaches in two key aspects. First, we updated the student model using the average loss computed from both human- and pseudo-labeled data. Second, we mitigated the impact of noisy pseudo-labeled data by assessing the feedback score and updating the teacher model only when the score fell below a threshold set to 0.0005.

The proposed NER method achieved the target performance in the spoken language domain. However, because performance in the written language domain did not meet our expectations, we conducted additional research to enhance the model. Specifically, we adopted a simple yet effective rollback method that evaluates the performance at regular intervals and reverts to the best model based on scarce human-labeled data. In addition, we improved the model performance by adjusting the label vector weights in the named entity dictionary.

The contributions of this study are summarized as follows:

1. **Innovative use of MPL for NER with enhanced performance**. We innovatively apply the MPL method to NER, enhancing classification by leveraging both labeled and unlabeled data, and introduce a feedback mechanism to mitigate the impact of noisy pseudo-labeled data.
2. **Performance enhancement techniques in different language domains**. Our method substantially improves the NER performance in both spoken and written language domains by employing a rollback method with a named entity dictionary for performance evaluation and label vector weight adjustments. As a result, the F1-scores of our method considerably surpass those of baseline models.

The teacher model with the best performance was selected, with an F1-score of 93.5% in the written language domain on the evaluation set, surpassing that of the baseline model by 3.38%. Similarly, in the spoken language domain, the teacher model achieved an F1-score of 94.16% on the evaluation set, outperforming the baseline model by 1.89%.

The remainder of this paper is organized as follows. Related work is discussed in Section 2. The proposed method is detailed in Section 3. In Section 4, we report experimental results. Finally, conclusions and directions of future work are presented in Section 5.

## 2 | RELATED WORK

### 2.1 | Transfer learning

Transfer learning can be broadly categorized into inductive transfer learning, which transfers knowledge across different tasks when the source and target tasks differ, and transductive transfer learning, which leverages similar knowledge when the source and target tasks are the same [16]. Transductive transfer learning is further divided into domain adaptation and cross-lingual learning depending on variations in domains or languages.

Three main methods have been employed for domain adaptation: representation [17–22], data weighting and selection [23–26], and self-labeling [27,28–30]. Self-labeling, which belongs to the semi-supervised learning category, trains a model on labeled samples and subsequently uses this model to assign pseudo- or proxy labels to unlabeled samples. In subsequent iterations, these labels are used to refine the model. In this study, we adopted self-labeling for domain adaptation considering the same task and language but different domains.

## 2.2 | KorBERT

In natural language processing, transformer-based pretrained models have been widely used, such as bidirectional encoder representations from transformers (BERT) [31–33]. Although multilingual BERT demonstrates impressive performance across various natural language processing tasks owing to its pretraining on Wikipedia data for 104 languages, it may not fully capture specific linguistic characteristics of individual languages [34]. For example, Korean, an agglutinative language with morphologically rich properties, presents challenges when subjected to BERT standard tokenizers. As meaningful morpheme units may lose their meaning during tokenization, an approach tailored for Korean should be devised. Accordingly, we propose KorBERT (https://aiopen.etri.re.kr/bertModel, Table 1), a model specially trained for Korean using the morpheme unit byte-pair encoding method that preserves the essence and meaning of morpheme units during tokenization. KorBERT outperforms Google multilingual versions by 4.5% on average for several downstream tasks.

The proposed NER model used morpheme-analyzed sentences as inputs and performs tokenization at the morpheme level. We used KorBERT as the base model for training the NER model. Typically, BERT-based NER models employ conditional random fields (CRF) or bidirectional long short-term memory-CRF to determine the labels using the last hidden vector. We used KorBERT to obtain vector representations of deep features, followed by CRF as the downstream layer for sequence labeling and generating the NER results. By fine-tuning BERT on training data, the vector representation combined linguistic knowledge from the pretrained model with task knowledge from the NER training data. Additionally, CRF allowed to capture conditional transition probabilities between different tags, thus mitigating logic errors in entity tag sequences during prediction (e.g., an I tag following an O tag) [35]. To this end, we adopted KorBERT using the Hugging Face transformer package and then seamlessly combined KorBERT and CRF by importing the TorchCRF package (https://pytorch-crf.readthedocs.io/en/stable/) for PyTorch.

# 3 | NER BASED ON TRANSFER LEARNING

## 3.1 | Initialization of teacher and student models

We extensively fine-tuned a teacher model using abundant labeled data from the written language domain based on a base model (KorBERT + CRF) as follows:

$$\theta_{\mathrm{B}}^{\mathrm{WD}} = \underset{\theta}{\operatorname{argmin}} \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_m^{\mathrm{WD}}\left(\theta_{\mathrm{B}}^{m-1}\right),$$
$$\mathcal{L}_m^{\mathrm{WD}}\left(\theta_{\mathrm{B}}^{m-1}\right) = l\left(Y_m^{\mathrm{WD}}, f\left(X_m^{\mathrm{WD}}; \theta_{\mathrm{B}}^{m-1}\right)\right), \tag{1}$$

where $\theta_T$ and $\theta_S$ represent the parameters of the teacher and student models, respectively, $\theta_B$ represents the base model (KorBERT + CRF), $M$ is the number of batches, which corresponds to the number of training steps and model updates, and $\theta_{\mathrm{B}}^{m}$ is the parameter of the base model at step $m$. The expressions that distinguish the domains are indicated in superscripts. WD is the written language domain, SD is the spoken language domain, and WD → SD denotes the transition from the written to the spoken domain. For example, $\theta_{\mathrm{B}}^{WD}$ denotes the parameter of the base model trained on human-labeled data in the written language domain, $\left(X_m^{\mathrm{WD}}, Y_m^{\mathrm{WD}}\right)$ denotes the $m$th batch of sentences and their corresponding labels in the written language domain, $f\left(X_m^{\mathrm{WD}}; \theta_{\mathrm{B}}^{m-1}\right)$ denotes the label predictions of batch $X_m^{\mathrm{WD}}$ by the base model at step $(m-1)$, and $l\left(Y_m^{\mathrm{WD}}, f\left(X_m^{\mathrm{WD}}; \theta_{\mathrm{B}}^{m-1}\right)\right)$ denotes the cross-entropy loss between answer labels of the written language domain and predicted labels of the base model at step $(m-1)$, expressed as $\mathcal{L}_m^{\mathrm{WD}}\left(\theta_{\mathrm{B}}^{m-1}\right)$.

Subsequently, we performed additional fine-tuning of the teacher model, which was initially trained on human-labeled data from the written language domain and scarce human-labeled data from the spoken language domain, as expressed in (2). Fine-tuning in a transformer-based BERT resembles transfer learning.

$$\theta_{\mathrm{T}} = \theta_B^{\mathrm{WD}\rightarrow\mathrm{SD}} = \underset{\theta}{\operatorname{argmin}} \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_m^{\mathrm{SD}}\left(\theta_{\mathrm{B}}^{\mathrm{WD},m-1}\right),$$
$$\mathcal{L}_m^{\mathrm{SD}}\left(\theta_{\mathrm{B}}^{\mathrm{WD},m-1}\right) = l\left(Y_m^{\mathrm{SD}}, f\left(X_m^{\mathrm{SD}}; \theta_{\mathrm{B}}^{\mathrm{WD},m-1}\right)\right), \tag{2}$$

**TABLE 1** Characteristics of KorBERT.

| Tokenizer | Data | No. of vocabs | No. of parameters | Structure |
|---|---|---|---|---|
| Morpheme-level character-level (WodPiece) | 23 GB/4.7 billion morphemes | 30 349 | 110 million | 12 layers,768 hidden layers, 12 heads |

As shown in Figure 1, the teacher model was trained through simple transfer learning, while the student model was trained by combining written and spoken language labeled data as follows:

$$\theta_S = \underset{\theta}{\operatorname{argmin}} \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_m^{WD+SD}\left(\theta_B^{m-1}\right),$$
$$\mathcal{L}_m^{WD+SD}\left(\theta_B^{m-1}\right) = l\left(Y_m^{WD+SD}, f\left(X_m^{WD+SD}; \theta_B^{m-1}\right)\right), \quad (3)$$

where WD + SD indicates the combination of the written and spoken domains and $Y_m^{WD+SD}$ represents the labels merged from the two domains. Model training with the desired performance is difficult when a small training set is available, and collecting additional training data is time and labor intensive. Alternatively, we used transfer learning with an improved MPL method to increase the model performance using automatically generated pseudo-labeled data.

## 3.2 | Generation of pseudo-labeled data

After training the teacher model, we generated pseudo-labeled data in the spoken language domain, as shown in Figure 2. By employing the trained teacher model, we automatically labeled raw spoken language data, leveraging the teacher knowledge to annotate the dataset. This process supported training of the student model in the spoken language domain [36]. Notably, we generated pseudo-labeled data at learning step $t$ as follows:

$$\left(uX_m^{SD}, \widehat{uY}_{m,\theta_T^t}^{SD}\right) = f\left(uX_m^{SD}; \theta_T^t\right), \quad (4)$$

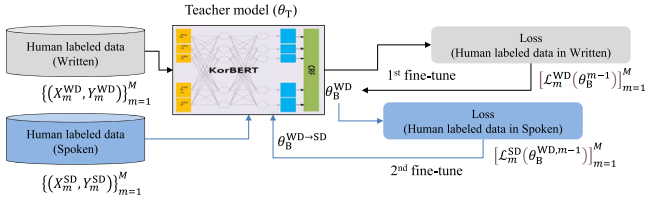


FIGURE 1 Diagram of initialization of teacher model.

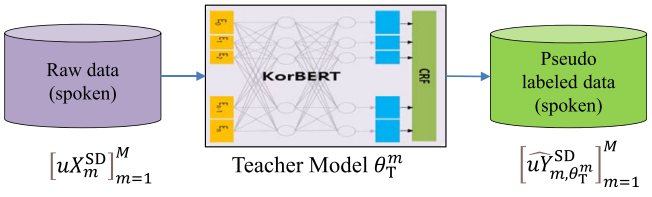


FIGURE 2 Diagram of generation of pseudo-labeled data.

where $uX_m^{SD}$ is the $m$th batch of unlabeled raw data in the spoken language domain and $\widehat{uY}_{m,\theta_T^t}^{SD}$ represents the predicted labels of the teacher model.

## 3.3 | Student model update

We trained a student model specifically tailored to the spoken language domain by using both the generated pseudo-labeled data and available human-labeled data. In [13], only pseudo-labeled data were used for student model training. Sufficiently large sets of raw data are required to generate and use pseudo-labeled data until the student model achieves the desired performance in fields like image processing. On the other hand, acquiring a large raw corpus with several named entities in the spoken language domain is challenging. Therefore, we improved the performance by incorporating small amounts of human-labeled data and pseudo-labeled corpora into training. To update the model, we calculated the loss from input data, which is equivalent to model learning. We computed the loss for both human- and pseudo-labeled data using the student model, as shown in (5) and Figure 3. Subsequently, we updated the student model by averaging the two losses.

$$\theta_S^1 = \underset{\theta}{\operatorname{argmin}}\, AVG\left(\mathcal{L}_1^{SD}\left(\theta_S^0\right), \hat{\mathcal{L}}_1^{SD}\left(\theta_S^0\right)\right),$$
$$\mathcal{L}_1^{SD}\left(\theta_S^0\right) = l\left(Y_1^{SD}, f\left(X_1^{SD}; \theta_S^0\right)\right), \quad (5)$$
$$\hat{\mathcal{L}}_1^{SD}\left(\theta_S^0\right) = l\left(\widehat{uY}_{1,\theta_T^0}^{SD}, f\left(uX_1^{SD}; \theta_S^0\right)\right).$$

To determine the feedback score of the student model, we used loss $\hat{\mathcal{L}}_1^{SD}\left(\theta_S^0\right)$ from the first pseudo-labeled data generated by the student model in the previous step ($\theta_S^0$). The performance was improved by using pseudo-labeled data for learning. By including human-labeled data in the model update, we mitigated noise in the automatically generated pseudo-labeled data.

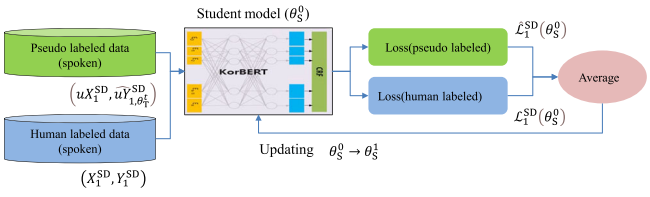


FIGURE 3 Diagram of updating student model with proposed loss averaging.

### 3.4 | Calculation of feedback score

The feedback score from the loss of the student model for labeled instances adjusted the teacher model to improve pseudo-labeled data [13], as shown in Figure 4, as follows:

$$H_{\mathrm{S}}^{1} = \hat{\mathcal{L}}_{1}^{\mathrm{SD}}\left(\theta_{\mathrm{S}}^{0}\right) * \mathcal{L}_{1}^{\mathrm{SD}}\left(\theta_{\mathrm{S}}^{1}\right) * lr_{\mathrm{S}}, \tag{6}$$

where $H_S^1$ is the feedback score of the first student model. The learning rate of the student model ($lr_S$) was used to calculate the feedback score.

### 3.5 | Teacher model update

After calculating the feedback score, the teacher model was updated as follows:

$$\begin{aligned} \theta_{\mathrm{T}}^{1} &= \underset{\theta}{\operatorname{argmin}}\, AVG\left(\mathcal{L}_{1}^{\mathrm{SD}}\left(\theta_{\mathrm{T}}^{0}\right), \hat{\mathcal{L}}_{1}^{\mathrm{SD}}\left(\theta_{\mathrm{T}}^{0}\right)\right), \\ \hat{\mathcal{L}}_{1}^{\mathrm{SD}}\left(\theta_{\mathrm{T}}^{0}\right) &= \hat{\mathcal{L}}_{1}^{\mathrm{SD}}\left(\theta_{\mathrm{T}}^{0}\right) * H_{\mathrm{S}}^{1} * lr_{\mathrm{S}}, \\ \mathcal{L}_{1}^{\mathrm{SD}}\left(\theta_{\mathrm{T}}^{0}\right) &= l\left(Y_{1}^{\mathrm{SD}}, f\left(X_{1}^{\mathrm{SD}}; \theta_{\mathrm{T}}^{0}\right)\right). \end{aligned} \tag{7}$$

The loss for the initial pseudo-labeled data generated by the teacher model, denoted as $\hat{\mathcal{L}}_1^{\mathrm{SD}}\left(\theta_{\mathrm{T}}^{0}\right)$, was determined based on the feedback score and learning rate of the student model. After the teacher model completed learning step $t$, the next iteration began at learning step $(t+1)$. This involved generating pseudo-labeled data using the updated teacher and student models. Thus, we developed a conditional update to the teacher model to enhance its performance, as shown in Figure 5. The teacher model was updated only when the student model

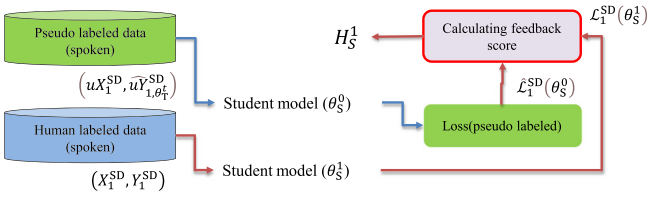


**FIGURE 4** Diagram for calculating feedback score.

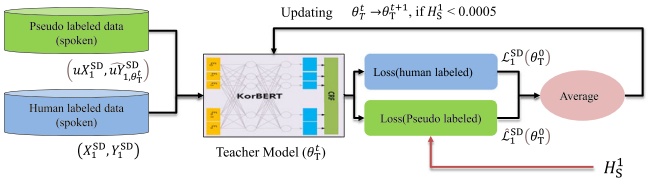


**FIGURE 5** Diagram of conditional teacher model update.

feedback score fell below a threshold of 0.0005. The threshold was determined experimentally to ensure that the distribution of new pseudo-labeled data reflected that of human-labeled data.

### 3.6 | Rollback-based training

Our primary aim was to enhance the NER performance in the spoken language domain under limited human-labeled data. Despite the improvements of the proposed method, we explored additional strategies to further increase the performance. Rollback learning employs the expected error reduction to eliminate outliers or relabel misclassified samples [37]. Accordingly, we devised a rollback method to maintain the best-performing model and evaluate its performance using scarce human-labeled data, as shown in Figure 6. We continuously assessed whether the teacher model exhibited performance improvement. If the performance was lower than that of the previous best model, the trained model was rolled back to the previous best model. Otherwise, if the current performance was high, the highest performance and model were updated to the current values. The evaluation was conducted over $N$ learning steps. By maintaining the teacher model with the best performance, we ensured the generation of a high-quality pseudo-corpus at each step. We used 10 steps in this study.

### 3.7 | Use of external named entity dictionary

Numerous studies have been aimed to enhance the NER performance using external resources. For instance, a method for retrieving and selecting semantically relevant texts through a search engine has been proposed using the original sentence as a query to find external contexts [38]. In this study, we explored the application of

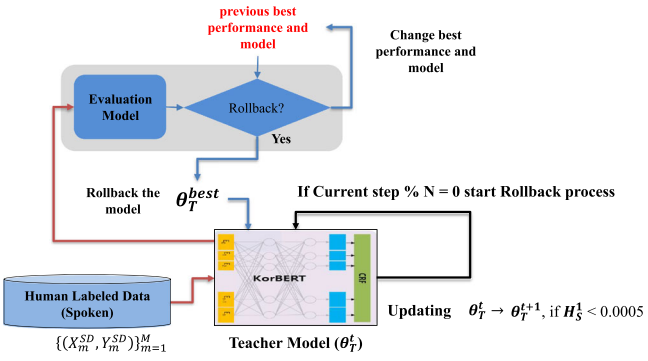


**FIGURE 6** Diagram of rollback-based training.

the named entity dictionary, which is a crucial resource in NER. For each token, we retrieved the corresponding label from the dictionary, and the matched labels were employed during learning or inference. Figure 7 illustrates the dictionary application during training.

KorBERT produced logits to compute the weights of all labels per token. The final label was determined using CRF based on these logits. The labels matched in the dictionary were incorporated into the logits as features with an additional weight value of $N = 9$. Owing to the ambiguity of the named entities, multiple labels could be matched. A specific named entity could be identified using this method during inference. For instance, legal names were recognized as named entities with label CV_LAW. In general, legal names have a negligible probability of belonging to other named entity labels. Consequently, unconditionally recognizing legal names as CV_LAW promoted the NER performance.

## 4 | EXPERIMENTS

We assessed the enhanced performance of the proposed method in comparison with a conventional transfer learning technique (BERT with fine-tuning). We aimed to achieve the target performance and conducted experiments on an established evaluation set, which has been widely used for testing NER in our projects.

### 4.1 | Datasets and experimental setup

The dataset employed in this study comprised 15 primary labels and 146 sublabels. Table 2 lists representative labels, such as person (PS), artifact (AF), and organization (OGG), along with concrete instances and selected sublabels. For instance, AF_CULTURAL_ASSET indicates cultural property, whereas AF_MUSICAL_INSTRUMENT designates a musical instrument. Likewise, OGG_ECONOMY indicates an enterprise, and OGG_EDUCATION indicates an educational institution. From an input sentence, named entities were recognized and aligned with the corresponding object name labels. Fifteen additional categories are listed in Table A1 in the Appendix.

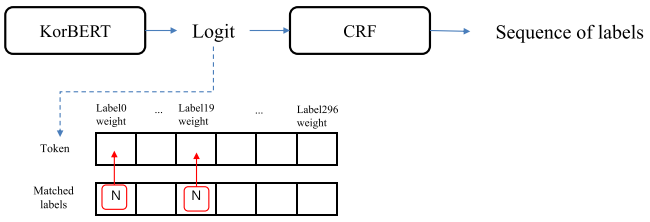


FIGURE 7 Diagram of using named entity dictionary for training and inference.

As outlined in Section 3.7, we employed a named entity dictionary comprising named entities and their corresponding labels. Table A2 in the Appendix illustrates the named-entity dictionary. The dictionary contained 3.5 million entries for learning and approximately 1000 entries for tuning. The labels for spoken and written languages were the same. The data used for training and evaluation are described in Table 3. For the training data, the original sentences were morphologically analyzed and tokenized. Each token was then constructed by tagging named entity tags using the BIO method. This is specified in Table A3 of the Appendix. In this study, we did not consider nested named entities.

KorBERT + CRF was used as the base model. The model parameter configurations are listed in Table 4. Moreover, the number of gradient accumulation steps was set to four. To facilitate learning and evaluation, we adapted the selection of source codes within the Hugging Face framework while retaining the default values of the parameters for those not listed in Table 4. We conducted experiments with either 146 sublabels or 15 representative labels.

### 4.2 | Performance of initial teacher and student models

As described in Section 3.1, we performed initial training of the teacher and student models. We evaluated the written and spoken language data against the initialized models. In this experiment, we used the spoken language

TABLE 2 Example of labels in our datasets.

| Named entity labels | Description |
|---|---|
| **Sublabels and examples** | |
| Person (PS) | Person's name |
| Son Heung-min (*PS_NAME*), Iron Man (*PS_NAME*), Zeus (*PS_NAME*), etc. | |
| Artifacts (AF) | Artifact |
| Eiffel Tower (*AF_CULTURAL_ASSET*), Oboe (*AF_MUSICAL_INSTRUMENT*), etc. | |
| Organization (OGG) | Organization name |
| Samsung Electronics (*OGG_ECONOMY*), MIT (*OGG_EDUCATION*), etc. | |

TABLE 3 Data statistics.

| Domain | No. of training samples | No. of test samples |
|---|---|---|
| Written | 255 624 | 2320 |
| Spoken | 25 177 | 2323 |

performance as the basis for improving the performance of the corresponding model considering 146 labels.

The performance gain achieved through basic transfer learning from the written language model surpassed that resulting from combining written and spoken language data for training. Nevertheless, we aimed to achieve a high performance within few epochs. As indicated in Table 5, although increasing the number of epochs led to incremental performance improvements, our target performance (92%) was not achieved.

## 4.3 | Performance of updated teacher and student models

Different criteria were used for updating the teacher and student models during simultaneous updates in a single training step. The student model considered pseudo- and human-labeled data. Averaging the losses from these two data sources outperformed the sequential update. Regarding pseudo-labeled data loss $Loss_{PLD}$ and human-labeled data loss $Loss_{HLD}$, this experiment reaffirmed that enhancements in the student model corresponded to improvements in the teacher model.

Typically, fewer labels lead to a lower processing complexity. Instead of 146 labels, we evaluated 15 labels in the conditional update experiment because the more labels would require a long computation time to reach the performance target. When the teacher model was conditionally updated, the performance improved for both 15 and 146 labels.

## 4.4 | Performance of proposed mothed

We conducted a comprehensive evaluation by collectively implementing all the proposed methods. For the final experiment, we employed the base model with the highest performance as both the teacher and student models while enabling swift learning. Our performance objectives per domain were distinct, achieving an F1-score of 93% for the written domain and 92% for the spoken domain. Given the anticipated challenges of improving the performance in the spoken language domain owing to labeled data scarcity, we conservatively set the target for spoken language NER. However, the difficulty in enhancing NER for spoken language was comparatively low. To achieve a comparable performance to that of the written language domain, we introduced two strategies: (1) rollback method to retain the best-performing model and (2) incorporation of named entity dictionary into learning and reasoning. Finally, we assessed the optimal performance achieved by the teacher and student models (Tables 6 and 7).

In Table 8, rollback indicates the utilization of the rollback method, whereas dic_train indicates the incorporation of a named entity dictionary during training and dic_tunning indicates the use of the named entity dictionary during inference, which fixes the recognition outcomes with labels from the dictionary. This dictionary application had a heightened significance in the written language domain. For instance, entity tenofovir alafenamide fumarate (TMM_DRUG) referred to a treatment for hepatitis B and acquired immunodeficiency syndrome developed by Gilead Sciences in the United States. Given its extended length and remote likelihood of being

**TABLE 6** Performance of training method for student model in spoken language domain (%).

| Training type | Teacher model | Student model |
|---|---|---|
| Base (no pseudo-labeled data) | 89.19 | 87.69 |
| $Loss_{PLD} \rightarrow Loss_{HLD}$ | 89.94 (+0.75) | 90.17 (+2.48) |
| Avg ($Loss_{PLD}$, $Loss_{HLD}$) | 90.21 (+1.02) | 90.48 (+2.79) |

**TABLE 4** Parameter settings of KorBERT + CRF.

| Max. length | Learning rate | No. of epochs | Warmup proportion | Batch size |
|---|---|---|---|---|
| 256 | $5.0e^{-5}$ | 2 or 5 | 0.1 | 8 |

**TABLE 5** Performance of initialized teacher and student models in spoken language domain (%).

| Model | Trained domain | No. of epochs | Spoken domain |
|---|---|---|---|
| – | WD | 2 | 79.07 |
| Student | WD + SD | 2 | 87.69 (+8.62) |
| | | 20 | 89.48 (+10.41) |
| Teacher | WD → SD | 2 | 89.19 (+10.12) |
| | | 20 | 89.50 (+10.43) |

TABLE 7 Performance of conditional updating teacher model in spoken language domain (%).

| No. of labels | Training type | Teacher | Student |
|---|---|---|---|
| 15 | Base | 92.27 | 90.78 |
| | Base + Avg. loss + conditional update | **93.14 (+0.87)** | **93.07 (+2.29)** |
| 146 | Base | 89.19 | 87.69 |
| | Base + Avg. loss | 90.21 (+1.02) | 90.48 (+2.79) |
| | Base + Avg. loss + conditional update | **90.11 (+0.92)** | **90.62 (+2.93)** |

TABLE 8 Performance of final method in spoken and written language domains (%).

| Pseudo-labeled data | Training type | Spoken | Written |
|---|---|---|---|
| Target performance | | 92.00 | 93.00 |
| Not used | WD + SD | 90.78 | 90.01 |
| | WD → SD (base) | 92.27 | 90.12 |
| Used | Base + Avg. loss + conditional update (#1) | 93.14 (+0.87) | 90.69 (+0.57) |
| | #1 + rollback + dict_train (#2) | 93.11 (+0.84) | 91.34 (+1.22) |
| | #2 + dict_tunning | **94.16 (+1.89)** | **93.5 (+3.38)** |

present in the training data, its accurate identification as an entity name was challenging. In such cases, a named entity dictionary must be incorporated. Consequently, in the written language domain, employing a named entity dictionary during learning can lead to a substantial performance enhancement.

## 5 | CONCLUSION

This study was aimed to develop a high-performance NER model adaptable to both written and spoken language domains. By leveraging transfer learning, we employed our specialized KorBERT model as the base to bridge the domain gap. The devised MPL method enabled domain adaptation by employing a teacher/student framework to enhance the quality of pseudo-labeled data. Our approach averaged the student model loss from human- and pseudo-labeled data while excluding noisy pseudo-labeled data under guidance of feedback scores. As a result, we achieved the target NER performance in the spoken language domain. Although our model fulfilled the requirements of the spoken language domain, further research was conducted to enhance the performance in the written language domain. We developed a rollback method and adjusted the label vector weights in the named entity dictionary to increase the performance. The results showed significant enhancements in the teacher model, achieving F1-scores of 93.5% and 94.16% in the written and spoken language domains, respectively. These results surpassed those of the baseline models by 1.89% and 3.38%, demonstrating the effectiveness of our approach. Overall, combining transfer learning, domain adaptation, and selective feedback-driven model updates notably enhanced NER across the written and spoken language domains.

## ORCID

*Kyoungman Bae* https://orcid.org/0000-0001-9007-4027

## REFERENCES

1. Named-entity recognition, [last accessed 10 August 2023], Available at: https://en.wikipedia.org/wiki/Named-entity_recognition
2. X Ma and E. Hovy, *End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF*, arXiv Preprint, 2016, DOI https://doi.org/10.48550/arXiv.1603.01354.
3. J. Li, A. Sun, J. Han, and C. Li, *A survey on deep learning for named entity recognition*, IEEE Trans. Knowl. Data Eng. **34** (2022), 50–70.
4. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, *Natural language processing (almost) from scratch*, J. Mach. Learn. Res. **12** (2011), 2493–2537.
5. Y. Lin, S. Yang, V. Stoyanov, and H. Ji, *A multi-lingual multi-task architecture for low-resource sequence labeling*, (Proc. 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia), 2018, pp. 799–809.
6. L. Liu, J. Shang, X. Ren, F. Xu, H. Gui, J. Peng, and J. Han, *Empower sequence labeling with task-aware neural language model*, (Proc. Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA), 2018, pp. 5253–5260.
7. W. Zhou and M. Chen, *Learning from noisy labels for entity-centric information extraction*, (Proc. 2021 Conf. Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic), 2021, pp. 5381–5392.

8. I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, *LUKE: Deep contextualized entity representations with entity-aware self-attention*, (Proc. 2020 Conf. Empirical Methods in Natural Language Processing, Online), 2020, pp. 6442–6454.
9. X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, *Dice loss for data-imbalanced NLP tasks*, (Proc. 58th Annual Meeting of the Association for Computational Linguistics, Online), 2020, pp. 465–476.
10. J. Li, S. Shang, and L. Shao, *MetaNER: Named entity recognition with meta-learning*, (Proc. Web Conference, Taipei, Taiwan), 2020, pp. 429–440.
11. S. Niu, Y. Liu, J. Wang, and H. Song, *A decade survey of transfer learning*, IEEE Trans. Artif. Intell. **1** (2020), 151–166.
12. X. Yang, Z. Song, I. King, and Z. Xu, *A survey on deep semi-supervised learning*, IEEE Trans. Knowl. Data Eng. **35** (2022), 8934–8954.
13. H. Pham, Z. Dai, Q. Xie, and Q. V. Le, *Meta pseudo labels*, (2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition, Online), 2021, pp. 11557–11568.
14. Y. Wang, S. Mukherjee, H. Chu, Y. Tu, M. Wu, J. Gao, and A. H. Awadallah, *Meta self-training for few-shot neural sequence labeling*, (Proc. 27th ACM SIGKDD Conf. Knowledge Discovery & Data Mining, Online), 2021, pp. 1737–1747.
15. K. He, R. Mao, T. Gong, C. Li, and E. Cambria, *Meta-based self-training and re-weighting for aspect-based sentiment analysis*, IEEE Trans. Affect. Comput. **14** (2022), no. 3, 1–13.
16. S. Ruder, Neural transfer learning for natural language processing, Ph.D. Dissertation, National Univ. of Ireland, 2019.
17. M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, (Proc. 34th International Conf. Machine Learning, Sydney, Australia), 2017, pp. 214–223.
18. A. Margolis, K. Livescu, and M. Ostendorf, *Domain adaptation with unlabeled data for dialog act tagging*, (Proc. 2010 Workshop on Domain Adaptation for Natural Language Processing, Uppsala, Sweden), 2010, pp. 45–52.
19. S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, *Domain adaptation via transfer component analysis*, IEEE Trans. Neural Netw. **22** (2010), 199–210.
20. X. Glorot, A. Bordes, and Y. Bengio, *Domain adaptation for large-scale sentiment classification: A deep learning approach*, (Proc. 28th International Conf. Machine Learning, Bellevue, WA, USA), 2011, pp. 513–520.
21. L. Qu, G. Ferraro, L. Zhou, W. Hou, and T. Baldwin, *Named entity recognition for novel types by transfer learning*, (Proc. 2016 Conf. Empirical Methods in Natural Language Processing, Austin, TX, USA), 2016, pp. 899–905.
22. Z. Wang, Y. Qu, L. Chen, J. Shen, W. Zhang, S. Zhang, Y. Gao, G. Gu, K. Chen, and Y. Yu, *Label-aware double transfer learning for cross-specialty medical named entity recognition*, (Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA), 2018, pp. 1–15.
23. B. Plank, A. Johannsen, and A. Søgaard, *Importance weighting and unsupervised domain adaptation of POS taggers: A negative result*, (Proc. 2014 Conf. Empirical Methods in Natural Language Processing, Doha, Qatar), 2014, pp. 968–973.
24. A. Søgaard and M. Haulrich, *Sentence-level instance-weighting for graph-based and transition-based dependency parsing*, (Proc. 12th International Conf. Parsing Technologies, Dublin, Ireland), 2011, pp. 43–47.
25. M. van der Wees, A. Bisazza, and C. Monz, *Dynamic data selection for neural machine translation*, (Proc. 2017 Conf. Empirical Methods in Natural Language Processing, Copenhagen, Denmark), 2017, pp. 1400–1410.
26. S. Ruder, P. Ghaffari, and J. G. Breslin, *Knowledge adaptation: Teaching to adapt*, arXiv Preprint, 2017, DOI https://doi.org/10.48550/arXiv.1702.02052
27. X. J. Zhu, Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, Univ. of Wisconsin-Madison, 2005.
28. D. McClosky, E. Charniak, and M. Johnson, *Effective self-training for parsing*, (Proc. Main Conf. Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, NY, USA), 2006, pp. 152–159.
29. O. Sandu, G. Carenini, G. Murray, and R. Ng, *Domain adaptation to summarize human conversations*, (Proc. 2010 Workshop on Domain Adaptation for Natural Language Processing, Uppsala, Sweden), 2010, pp. 16–22.
30. Y. He and D. Zhou, *Self-training from labeled features for sentiment analysis*, Inf. Process. Manag. **47** (2011), 606–616.
31. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, (31st Conf. Neural Information Processing Systems, Long Beach, CA, USA), 2017, pp. 5998–6008.
32. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, (Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019, Minneapolis, MN, USA), 2019, pp. 4171–4186.
33. C. Jia, Y. Shi, Q. Yang, and Y. Zhang, *Entity enhanced BERT pre-training for Chinese NER*, (Proc. 2020 Conf. Empirical Methods in Natural Language Processing, Online), 202, pp. 6384–6396.
34. S. Lee, H. Jang, Y. Baik, S. Park, H. Shin, *KR-BERT: A small-scale Korean-specific language model*, arXiv Preprint, 2020, DOI https://doi.org/10.48550/arXiv.2008.03979
35. Y. Gong, L. Mao, and C. Li, *Few-shot learning for named entity recognition based on BERT and two-level model fusion*, Data Intell. **3** (2021), no. 4, 568–577.
36. DH Lee, *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks*, (ICML 2013 Workshop on Challenges in Representation Learning, Atlanta, GA, USA), 2013, pp. 896–901.
37. M. U. Ahmed, Y. H. Kim, and P. K. Rhee, *EER-ASSL: combining rollback learning and deep learning for rapid adaptive object detection*, KSII Trans. Internet Inf. Syst. **14** (2020), 4776–4794.
38. X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu, *Improving named entity recognition by external context retrieving and cooperative learning*, (Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conf. Natural Language Processing, Online), 2021, pp. 1800–1812.

## AUTHOR BIOGRAPHIES

**Kyoungman Bae** received the BS, MS, and PhD degrees in computer engineering from the Department of Computer Engineering, Dong-A University, Busan, Republic of Korea, in 2004, 2006, and 2016, respectively. Since 2016, he has been working for the Language Intelligence Research Section at the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. His research interests include large language model, natural language processing, explainable artificial intelligence, and generative artificial intelligence.

**Joon-Ho Lim** received the BS and MS degrees in computer science from the Korean University, Seoul, Republic of Korea, in 2002 and 2005, respectively, and the PhD degree in computer engineering from Chungnam National University, Daejeon, Republic of Korea, in 2016. From 2005, he worked for the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. Since 2022, he has also been working as the Chief Technology Officer at Tutorus Labs, an educational artificial intelligence startup. His research interests include large language models, artificial intelligence alignment, and conversation-based tutoring artificial intelligence.

**How to cite this article:** K. Bae and J.-H. Lim, *Named entity recognition using transfer learning and small human- and meta-pseudo-labeled datasets*, ETRI Journal **46** (2024), 59–70, DOI 10.4218/etrij.2023-0321.

## APPENDIX A.

**TABLE A1** Fifteen representative labels from datasets used in this study.

| Named entity label | Description | Subcategories and examples |
|---|---|---|
| PERSON | Person's name | Son Heung-min, Iron Man, Zeus, etc. |
| STUDY_FIELD | Field of study | Social science (FD_SOCIAL_SCIENCE), engineering (FD_SCIENCE), medicine (FD_MEDICINE), etc. |
| THEORY | Theory, law, or principle | Special relativity (TR_SCIENCE), Heinrich's law (TR_SOCIAL_SCIENCE), etc. |
| ARTIFACTS | Artifact | Eiffel Tower (AF_CULTURAL_ASSET), oboe (AF_MUSICAL_INSTRUMENT), etc. |
| ORGANIZATION | Organization name | Samsung Electronics (OGG_ECONOMY), MIT (OGG_EDUCATION), etc. |
| LOCATION | Region/location | USA (LCP_COUNTRY), New York (LCP_CITY), etc. |
| CIVILIZATION | Civilization/culture | Indus culture (CV_NAME), soccer (CV_SPORTS), English (CV_LANGUAGE), etc. |
| DATE | Date | August (DT_MONTH), 23 years (DT_YEAR), etc. |
| TIME | Time | 12 hours (TI_HOUR), 30 seconds (TI_SECOND), etc. |
| QUANTITY | Quantity | 40 years old (QT_AGE), 55 m (QT_LENGTH), etc. |
| EVENT | Specific event/incident/accident | Opium War (EV_WAR_REVOLUTION), Seoul Olympics (EV_SPORTS), etc. |
| ANIMAL | Animal | Spider (AM_INSECT), salmon (AM_FISH), etc. |
| PLANT | Plant and derivatives | Apple (PT_FRUIT), cherry tree (PT_TREE), etc. |
| MATERIAL | Material | Aluminum (MT_METAL), ammonia (MT_CHEMICAL), etc. |
| TERM | Other entities | White (TM_COLOR), square (TM_SHAPE), COVID-19 (TMM_DISEASE), etc. |

**TABLE A2** Example of named entity dictionary.

| Named entity dictionary | |
|---|---|
| SonHeungmin = PS_NAME | 손흥민 = PS_NAME |
| SamsungElectronics = OGG_ECONOMY | 삼성전자 = OGG_ECONOMY |
| MIT = OGG_EDUCATION | MIT = OGG_EDUCATION |
| EiffelTower = AF_CULTURAL_ASSET | 에펠탑 = AF_CULTURAL_ASSET |
| Oboe = AF_MUSICAL_INSTRUMENT | 오보에 = AF_MUSICAL_INSTRUMENT |

**TABLE A3** Format of corpus used for training and testing.

| **경찰은 소래포구 어시장 화재 목격자 3명 진술 확보 (the police secured statements from three witnesses of the Sorae Port fish market fire)** | | |
|---|---|---|
| 경찰/NNG | B-OGG_POLITICS | Police |
| 은/JX | O | |
| 소래포구/NNP | B-EV_OTHERS | Sorae Port |
| 어/NNG | I-EV_OTHERS | Fish market |
| 시장/NNG | I-EV_OTHERS | |
| 화재/NNG | I-EV_OTHERS | Fire |
| 목격/NNG | B-CV_POSITION | Witnesses |
| 자/XSN | B-CV_POSITION | |
| 3/SN | B-QT_MAN_COUNT | Three |
| 명/NNB | I-QT_MAN_COUNT | |
| 진술/NNG | O | Secured statements |
| 확보/NNG | O | |
| ./SF | O | |

B-, beginning of named entity; CV_POSITION, position/position name; EV_OTHERS, another incident/incident name; I-, middle of named entity; JX, auxiliary; NNB, dependent noun; NNG, common noun; NNP, proper noun; OGG_POLITICS, government/administrative agency, public agency, political agency; QT_MAN_COUNT, number of people; SF, period, question mark, exclamation mark; SN, number; XSN, noun derived suffix.