Analysis of Zero-Shot Crosslingual Learning between English and Korean for Named Entity Recognition

♠Jongin Kim, ♠Nayoung Choi, ♢Seunghyun S. Lim, ♠Jungwhan Kim
♠Soojin Chung, ♠Hyunsoo Woo, ♣Min Song, ♢Jinho D. Choi
♠Department of Digital Analytics, Yonsei University, Seoul, South Korea
♣Department of Library and Information Science, Yonsei University, Seoul, South Korea
♢Department of Computer Science, Emory University, Atlanta, GA, USA
♠{jongin.kim, skdudenql11, jngwhnk, socheon, wkat}@yonsei.ac.kr
♣min.song@yonsei.ac.kr, ♢{slim72, jinho.choi}@emory.edu

Abstract

This paper presents a English-Korean parallel dataset that collects 381K news articles where 1,400 of them, comprising 10K sentences, are manually labeled for crosslingual named entity recognition (NER). The annotation guidelines for the two languages are developed in parallel, that yield the inter-annotator agreement scores of 91 and 88% for English and Korean respectively, indicating sublime quality annotation in our dataset. Three types of crosslingual learning approaches, direct model transfer, embedding projection, and annotation projection, are used to develop zero-shot Korean NER models. Our best model gives the F1-score of *51%* that is very encouraging, considering the extremely distinct natures of these two languages. This is pioneering work that explores zero-shot crosslingual learning between English and Korean and provides rich parallel annotation for a core NLP task such as named entity recognition.

1 Introduction

Crosslingual representation learning aims to derive embeddings for words (or sentences) from multiple languages that can be projected into a shared vector space (Conneau et al., 2018; Schuster et al., 2019b; Conneau and Lample, 2019). One important application of crosslingual embeddings has been found for transferring models trained on a high-resource language to a low-resource one (Lin et al., 2019; Schuster et al., 2019a; Artetxe and Schwenk, 2019). The latest multilingual transformer encoders such as BERT (Devlin et al., 2019) and XLM (Conneau et al., 2020) have made it possible to develop robust crosslingual models through zero-shot learning that requires no labeled training data on the target side (Jebbara and Cimiano, 2019; Chidambaram et al., 2019; Chi et al., 2020). However, these approaches tend not to work as well for languages whose words cannot be easily aligned.

Our team is motivated to create a rich crosslingual resource between English and Korean, which are largely different in nature as English is known to be rigid-order, morphologically-poor, and head-initial whereas Korean is flexible-order, morphologically-rich, and head-final (Choi et al., 1994; Han et al., 2002; Hong, 2009). Creation of a high quality parallel dataset to facilitate crosslingual research can reduce the gap between these two languages, and advance NLP techniques in both languages.

This paper provides a comprehensive analysis of crosslingual zero-shot learning in English and Korean. We first create a new dataset comprising a large number of parallel sentences and annotate them for named entity recognition (NER; Sec. 3). We then adapt the crosslingual approaches and build NER models in Korean through zero-shot learning (Sec. 4). All models are experimented on our dataset and thoroughly compared to evaluate the feasibility of this work (Sec. 5). Our results are promising although depicting few challenges in zero-shot learning for English and Korean (Sec. 6). The contributions of this work can be summarized as follows:

- To create a crosslingual dataset that enables to develop robust zero-shot NER models in Korean.
- To present a new data selection scheme that can notably improve zero-shot model performance.
- To provide a comparative analysis among several crosslingual approaches and establish the initial foundation of this research.

2 Related Work

For crosslingual representation alignment, Artetxe et al. (2016) and Smith et al. (2017) suggested orthogonality constraints on the embedding transformation that led to better quality translation. Aldarmaki and Diab (2019) derived a context-aware crosslingual mapping from a parallel corpus using word alignment. Schuster et al. (2019b) aligned

Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 224–237
November 11, 2021.

	Business	Lifestyle	Politics	Sci/Tech	Society	Sports	World	Regional	Others	Total
TRN	109,464	132,606	92,877	55,420	164,414	52,214	70,487	57,490	12,881	747,853
DEV	666	733	728	718	711	715	763	-	-	5,034
TST	676	709	687	680	686	650	722	-	-	4,810
Total	110,806	134,048	94,292	56,818	165,811	53,579	71,972	57,490	12,881	757,697

(a) The number of parallel sentences by each category.

	Business	Lifestyle	Politics	Sci/Tech	Society	Sports	World	Regional	Others	Total
TRN	65,901	63,131	41,233	27,924	76,450	34,662	29,765	34,510	6,197	379,773
DEV	100	100	100	100	100	100	100	-	-	700
TST	100	100	100	100	100	100	100	-	-	700
Total	66,101	63,331	41,433	28,124	76,650	34,862	29,965	34,510	6,197	381,173

(b) The number of parallel news articles by each category.

Table 1: The statistics of our English-Korean dataset. TRN/DEV/TST: the training/development/evaluation sets.

word embeddings from multilingual transformer encoders using context independent embedding anchors. Recent works based on multilingual pretrained language model aligns representations between languages in a unsupervised fashion. Devlin et al. (2019) proposed multilingual BERT that generates contextualized word embeddings for multiple languages in one vector space by simply sharing all languages' vocabulary. Conneau and Lample (2019) extends mBERT by introducing bilingual data and an extra pretraining task (Translation Language Modeling). Luo et al. (2021) adds a cross-attention module into the Transformer encoder to explicitly build the interdependence between langauges.

For cross-lingual NER, Ni et al. (2017) presented weakly supervised crosslingual models using annotation and representation projection. Huang et al. (2019) made an empirical analysis of how sequential order and multilingual embeddings are used in crosslingual NER. Artetxe and Schwenk (2019) presented multilingual transfer models that used few-shot learning adapting supervising BEA, ranking and retraining for massive transfer. Wu and Dredze (2019) and Wu et al. (2020) directly transfers the NER model trained on the source language to the target language using crosslingual representations from multilingual encoders (Direct model transfer).

3 English-Korean Crosslingual Dataset

3.1 Data Collection

AI Open Innovation Hub (AI Hub) is an integration platform operated by the Korea National Information Society Agency that provides data, software, and computing resources for AI research. It has released the Korean-English AI Training Text Corpus (KEAT)[1] containing 1.6M English-Korean parallel sentences from various sources such as news media, government website/journal, law & administration, conversation and etc. For the present study, 800K parallel sentences from the news portion of this corpus are extracted.

3.2 Data Preprocessing

Since KEAT is not organized into documents, each sentence is composed independently although it comes with the URL of its original source. Thus, we group all sentences into news articles based on the URLs. Although there exist news articles with single sentence after the grouping process, we still include them in the train set in order to make full use of the parallel sentences provided, which will be used to train the word alignment model and the transformation matrix in Section 5. As a result, 757,697 sentences are selected, that are composed into 381,173 news articles, to create our English-Korean crosslingual dataset.

The news articles can be categorized into 9 sections: *Business*, *Lifestyle*, *Science/Technology*, *Society*, *Sports*, *World*, *Regional*, and *Others*. Among those, 200 articles are randomly sampled from each of the first 7 categories for our annotation in Section 3.4 and they are split into 50/50 to create the development and test sets for our experiments in Section 5. Table 1 describes the statistics of our dataset. All sections are uniformly distributed in DEV and TST, enabling to conduct comparative studies among these sections.

3.3 Pseudo Annotation

English sentences are tokenized and tagged with named entities by the open-source NLP tool called

[1] http://www.aihub.or.kr/aidata/87

	CARDNIAL$_e$	DATE$_e$	EVENT$_e$	FAC$_e$	GPE$_e$	LANGUAGE$_e$	LAW$_e$	LOC$_e$	MONEY$_e$
TRN	207,618 (10.6)	281,936 (14.5)	29,548 (1.5)	39,579 (2.0)	315,014 (16.2)	3,050 (0.2)	11,821 (0.6)	27,865 (1.4)	26,751 (1.4)
DEV	1,314 (11.6)	1,534 (13.6)	139 (1.2)	146 (1.3)	1,845 (16.3)	21 (0.2)	32 (0.3)	151 (1.3)	123 (1.1)
TST	1,220 (11.1)	1,552 (14.1)	155 (1.4)	151 (1.4)	1,751 (15.9)	29 (0.3)	50 (0.5)	154 (1.4)	136 (1.2)
Total	210,152 (10.7)	285,022 (14.5)	29,842 (1.5)	39,876 (2.0)	318,610 (16.2)	3,100 (0.2)	11,903 (0.6)	28,170 (1.4)	27,010 (1.4)

	NORP$_e$	ORDINAL$_e$	ORG$_e$	PERCENT$_e$	PERSON$_e$	PRODUCT$_e$	QUANTITY$_e$	TIME$_e$	WOA$_e$
TRN	68,700 (3.5)	82,270 (4.2)	394,226 (20.2)	8,339 (0.4)	352,918 (18.1)	12,170 (0.6)	16,736 (0.9)	37,003 (1.9)	35,193 (1.8)
DEV	503 (4.4)	517 (4.6)	2,286 (20.2)	61 (0.5)	2,168 (19.2)	70 (0.6)	62 (0.6)	148 (1.3)	203 (1.8)
TST	513 (4.7)	443 (4.0)	2,264 (20.6)	52 (0.5)	2,049 (18.6)	91 (0.8)	47 (0.4)	156 (1.4)	205 (1.9)
Total	69,716 (3.5)	83,230 (4.2)	398,776 (20.2)	8,452 (0.4)	357,135 (18.1)	12,331 (0.6)	16,845 (0.9)	37,307 (1.9)	35,601 (1.8)

(a) The number of pseudo-annotated named entities in the English sentences. WOA: WORK_OF_ART.

	DAT$_k$	DUR$_k$	LOC$_k$	MNY$_k$	NOH$_k$	ORG$_k$	PER$_k$	PNT$_k$	POH$_k$	TIM$_k$
TRN	156,013 (9.0)	41,651 (2.4)	235,179 (13.6)	37,538 (2.2)	285,898 (16.5)	478,830 (27.6)	312,578 (18.0)	37,767 (2.2)	136,731 (7.9)	12,894 (0.7)
DEV	752 (7.4)	267 (2.6)	959 (9.5)	159 (1.6)	1,807 (17.9)	3,231 (32.0)	1,909 (18.9)	220 (2.2)	748 (7.4)	49 (0.5)
TST	741 (7.5)	257 (2.6)	947 (9.6)	174 (1.8)	1,626 (16.5)	3,142 (31.9)	1,818 (18.5)	249 (2.5)	850 (8.6)	40 (0.4)
Total	157,506 (9.0)	42,175 (2.4)	237,085 (13.5)	37,871 (2.2)	289,331 (16.5)	485,203 (27.6)	316,305 (18.0)	38,236 (2.2)	138,329 (7.9)	12,983 (0.7)

(b) The number of pseudo-annotated named entities in the Korean sentences.

Table 2: The statistics of pseudo-annotated named entities by each tag in our English-Korean dataset. The numbers in the parentheses indicate the percentages of the corresponding tags for each set.

	CARDNIAL	DATE	EVENT	FAC	GPE	LANGUAGE	LAW	LOC	MONEY
EN	1,235 (10.9)	1,496 (13.2)	190 (1.7)	160 (1.4)	1,755 (15.5)	25 (0.2)	39 (0.3)	150 (1.3)	156 (1.4)
KR	1,359 (12.3)	1,381 (12.5)	186 (1.7)	158 (1.4)	1,674 (15.1)	21 (0.2)	39 (0.4)	141 (1.3)	159 (1.4)
E ∩ K	1,124 (10.9)	1,324 (12.9)	173 (1.7)	154 (1.5)	1,566 (15.2)	21 (0.2)	34 (0.3)	135 (1.3)	156 (1.5)

	NORP	ORDINAL	ORG	PERCENT	PERSON	PRODUCT	QUANTITY	TIME	WOA
EN	625 (5.5)	461 (4.1)	2,217 (19.6)	191 (1.7)	2,118 (18.7)	157 (1.4)	65 (0.6)	146 (1.3)	147 (1.3)
KR	540 (4.9)	351 (3.2)	2,249 (20.3)	199 (1.8)	2,129 (19.2)	155 (1.4)	67 (0.6)	140 (1.3)	142 (1.3)
E ∩ K	529 (5.1)	329 (3.2)	2,042 (19.8)	189 (1.8)	2,048 (19.9)	145 (1.4)	65 (0.6)	132 (1.3)	131 (1.3)

(a) Statistics of the development set (DEV).

	CARDNIAL	DATE	EVENT	FAC	GPE	LANGUAGE	LAW	LOC	MONEY
EN	1,117 (10.1)	1,511 (13.7)	207 (1.9)	161 (1.5)	1,701 (15.4)	29 (0.3)	57 (0.5)	154 (1.4)	165 (1.5)
KR	1,253 (11.6)	1,406 (13.0)	205 (1.9)	159 (1.5)	1,635 (15.1)	26 (0.2)	52 (0.5)	147 (1.4)	172 (1.6)
E ∩ K	1,018 (10.2)	1,336 (13.4)	196 (2.0)	151 (1.5)	1,517 (15.2)	25 (0.3)	51 (0.5)	137 (1.4)	164 (1.6)

	NORP	ORDINAL	ORG	PERCENT	PERSON	PRODUCT	QUANTITY	TIME	WOA
EN	621 (5.6)	397 (3.6)	2,159 (19.6)	215 (2.0)	2,012 (18.2)	172 (1.6)	52 (0.5)	148 (1.3)	156 (1.4)
KR	509 (4.7)	287 (2.7)	2,196 (20.3)	223 (2.0)	2,017 (18.7)	176 (1.6)	51 (0.5)	133 (1.2)	155 (1.4)
E ∩ K	501 (5.0)	274 (2.7)	1,980 (19.8)	214 (2.1)	1,955 (19.5)	161 (1.6)	50 (0.5)	132 (1.3)	141 (1.4)

(b) Statistics of the evaluation set (TST).

Table 3: The statistics of manually annotated named entities on the parallel sentences in the DEV and TST sets. The numbers in the parentheses indicate the percentages of the corresponding tags for each set. EN/KR: # of entities in the English/Korean sentences respectively, E ∩ K: # of entities existing in both English and Korean sentences.

ELIT[2] using the Flair model trained on OntoNotes (Pradhan et al., 2013). Korean sentences are tagged by a CRF-based model adapting KoBERT (Korean BERT)[3] trained on the corpus distributed by Cheon and Kim (2018). Note that the named entity types pseudo-annotated on the Korean sentences don't match with those of the English sentences for now, which will be matched in Section 3.4 in the case of DEV and TST. In addition, Korean sentences are processed by the Mecab morphological analyzer[4] that produces more linguistically sounding tokens than SentencePiece (Kudo and Richardson, 2018) in KoBERT. All named entities from the CRF tagger are then remapped to the tokens produced by the Mecab analyzer using heuristics so they can better reflect the previous morphology work in Korean (Hong, 2009). Words in every parallel sentence pair, tokenized by the ELIT and Mecab analyzers, are aligned by GIZA++, that has been adapted by many prior crosslingual studies (Och and Ney, 2003).

Table 2 shows the statistics of pseudo-annotated named entities in our dataset. The detailed descriptions of these tags are provided in Appendix A.1. The overall statistics are comparable between English and Korean, 2.5 and 2.3 entities per sentence, respectively. GPE$_e$, the 3rd most frequent tag in English, is not supported by the Korean tagger but

[2] https://github.com/emorynlp/elit
[3] https://github.com/eagle705/pytorch-bert-crf-ner
[4] https://bitbucket.org/eunjeon/mecab-ko/

rather tagged as $\texttt{ORG}_k$ or $\texttt{LOC}_k$, explaining why the numbers of these two tags in Korean are much greater than those of $\texttt{ORG}_e$ and $\texttt{LOC}_e$, respectively.

3.4 Parallel Annotation

We conduct a team of graduate students majoring in Data Science to manually tag named entities on all parallel sentences in the DEV and TST sets by taking the following 3 steps:[5]

1. For English, the pseudo-annotated entities are revised by the OntoNotes named entity guidelines (BBN, 2014; Maekawa, 2018), and missing entities are annotated as necessary.

2. For Korean, the pseudo-annotated entities are revised to match the English tagset, and missing entities are annotated as necessary.

3. Let $E = \{e_1, \ldots, e_n\}$ and $K = \{k_1, \ldots, k_m\}$ be the lists of entities from Steps 1 and 2 for a English and Korean sentence pair, respectively. Every entity pair (e_i, k_j) is linked in our dataset if e_i is the translation of k_j.

Note that every article in DEV and TST consists of at least 5 sentences with at least 2 named entities. Table 3 shows the statistics of the gold annotation. Out of 22,367 and 21,892 named entities annotated in English and Korean sentences, 20,300 of them are linked across the languages (above 90%).

	English	Korean
Unlabeled	92.7	90.4
Labeled	90.9	88.3

Table 4: Cohen's kappa scores measured for the English and Korean annotation. Unlabeled: matches only entity spans, Labeled: matches both the spans and the labels.

To estimate the inter-annotator agreement, 10 news articles from each of the first 7 sections in Table 1 are randomly picked and double annotated; the rest of DEV and TST are single annotated and sample checked. Table 4 shows the Cohen's kappa scores measured for the English and Korean annotation. The high labeled matching scores of 90.9 and 88.3 are achieved for those two languages respectively, implying that the single annotation in this dataset is expected to be of high quality as well.

3.5 Analytics by Languages

A couple of challenges are found during the parallel annotation. First, subjects are obligatory in English for most sentence forms whereas Korean is a pro-drop language so that entities in the subject position can be missing in Korean but not in English, which explains the greater number of entities in English. Second, certain inflectional morphemes in Korean can be dropped without violating the grammar, that often makes the labeling ambiguous. For instance, the literal translation of "*Korean Church*" would be "한국(Korea)+의('s) 교회(Church)", although it is the standard practice to drop "의('s)" in this case such that it becomes "한국(Korea) 교회(Church)". Given this translation, the annotator can be easily confused to annotate "한국(Korea)" as a geopolitical entity (GPE) instead of a nationality (NORP), which may lead to annotation disagreement.

Additional analytics by news sections and entity types are described in Appendix A.6

4 Zero-shot Crosslingual Learning

4.1 Overview of Approaches

Three crosslingual learning approaches are adapted to develop zero-shot Korean models. One is direct model transfer method following Wu and Dredze (2019). We reproduce the previous work which fine-tunes mBERT on English NER dataset and transfers the trained model to a target language, in our case, Korean. We fine-tune on OntoNotes, whereas the previous work fine-tuned on CoNLL 2003 NER dataset. The other two approaches that will be experimented are embedding preojection and annotation projection following Ni et al. (2017), although some modules in the implementation are updated or added: the encoders used to derive the embeddings from sentences, the word alignment tool, the training data selection scheme heuristics. Figure 1 illustrates an overview of two crosslingual learning approaches adapted to develop zero-shot Korean NER models. One is embedding projection (R1) that takes a labeled English sentence (R2) and generates English embeddings, (R3) which are fed into an orthogonal mapping (R4) then transformed into Korean embeddings (Section 4.2). The other is annotation projection (A1) that aligns words across the two languages and pseudo-annotates the Korean sentence, (A2) which are fed into an encoder (A3) to generate Korean embeddings (Section 4.3). The Korean embeddings generated by individual approaches are fed into a trainer to build the Korean NER models. No manual annotation is added to the Korean data; thus they both are zero-shot learning.

[5]The full annotation guidelines are available at https://github.com/emorynlp/MRL-2021

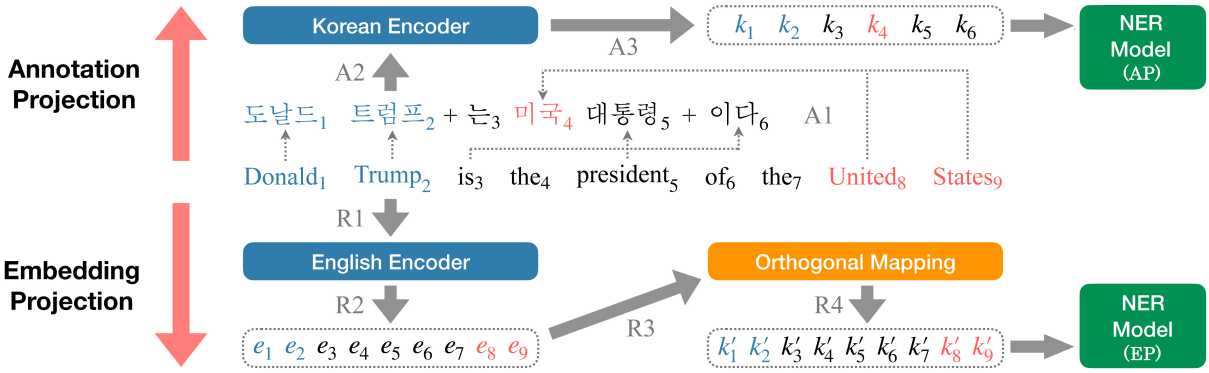

Figure 1: The overview of crosslingual methods, embedding projection (§4.2) and annotation projection (§4.3). The blue and red coded words and embeddings represent the PERSON and GPE entities, respectively. AP:Annotation Projection, EP:Embedding Projection

4.2 Embedding Projection

Let $X, Y \in \mathbb{R}^{n \times d}$ be parallel matrices between the source and target languages, where n is the number of parallel terms (words or sentences) in those two languages. Let $x_i, y_i \in \mathbb{R}^{1 \times d}$ be the i'th rows in X and Y, which are the embeddings of the i'th terms in the source and target languages respectively, that refer to the same content. Then, the transformation matrix $W \in \mathbb{R}^{d \times d}$ can be found by minimizing the distance between XW and Y as follows:

$$\underset{W}{\operatorname{argmin}} \|XW - Y\| \quad \text{s.t.} \quad W^T W = I$$

This optimization can be achieved by singular value decomposition as proposed by Artetxe et al. (2016), where $U, V \in \mathbb{R}^{d \times p}, \Sigma \in \mathbb{R}^{p \times p}$:

$$W = UV^T \quad \text{s.t.} \quad X^T Y = U\Sigma V^T$$

The transformation matrix W is used to convert any English embedding e_i into a Korean embedding k'_i in Figure 1 such that $e_i \cdot W = k'_i \approx k_j$ where k_j is the embedding from the Korean encoder that can be aligned with e_i. The NER model is trained on only English sentences represented by the transformed embeddings k'_* and a pseudo-label annotated with an existing English NER model. During decoding, the model takes Korean sentences represented by the encoded embeddings k_* and makes the predictions.

Given the latest contextualized encoders that generate different embeddings for the same word type by contexts (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019), the size of X and Y is as large as the number of all aligned words in the training data. It is worth saying that the transformed embedding space may be similar to the actual encoded space in the target language; however, the word order is still preserved as in the source language. Therefore, the model is limited to learn sequence information of the target language, which can be an issue for languages with very different word orderings.

4.3 Annotation Projection

Let $\mathbb{S} = \{S_1, \ldots, S_n\}$ and $\mathbb{T} = \{T_1, \ldots, T_n\}$ be lists of sentences in the source and target languages, and (S_i, T_i) be the i'th pair of parallel sentences in those two languages. Let $S_i = \{s_{i1}, \ldots, s_{in}\}$ and $T_i = \{t_{i1}, \ldots, t_{im}\}$ where s_i and t_i are the i'th word in S and T. Then, annotation projection can be performed as proposed by Ni et al. (2017):

1. Pseudo-label $S_{\forall i} \in \mathbb{S}$ using an existing model in the source language, in our case, ELIT (§3.2).

2. Pseudo-align words in every (S_i, T_i) using an existing tool, in our case, GIZA++ (§3.2). If a consecutive word span $S_i^{j,k} = \{s_{ij}, .., s_{ik}\}$ is pseudo-labeled as the entity type ℓ as well as pseudo-aligned with a span $T_i^{a,b} = \{t_{ia}, .., t_{ib}\}$, $T_i^{a,b}$ is also pseudo-labeled with ℓ.

The quality of pseudo annotation hugely depends on the performance of word alignment, which is generally not robust for the case of distant language pairs such as English and Korean. Thus, we propose a few constraints to filter out noisy annotation.

Entity Matching Let ψ be a boolean. If $\psi = \texttt{F}$, all parallel sentences in $(\mathbb{S}, \mathbb{T})$ are used for training. If $\psi = \texttt{T}$, (S_i, T_i) is selected for training only if all named entities in S_i are properly labeled in T_i by the above projection approach.

Relative Frequency Let e be an entity term such as "도널드 트럼프 (Donald Trump)" in Figure 1. Let L_e be a set of entity types pseudo-annotated for

all occurrences of e in the target language. Then, the relative frequency $P(\ell|e)$ for $\ell \in L_e$ and e can be measured as follows, where $\text{COUNT}(\ell, e)$ is the number of occurrences for e being labeled as ℓ:

$$P(\ell|e) = \frac{\text{COUNT}(\ell, e)}{\sum_{\ell' \in L_e} \text{COUNT}(\ell', e)}$$

Impurity Let F_e^ℓ be a set of unique terms in the source language that are pseudo-aligned with the term e labeled as ℓ in the target language such that $|F_e| \leq \text{COUNT}(l, e)$. Then, the impurity $M(\ell, e)$ is measured as follows where α is a smoothing factor:

$$M(\ell, e) = \frac{|F_e^\ell|}{\text{COUNT}(l, e) + \alpha}$$

The relative frequency $P(\ell|e)$ and the impurity $M(\ell, e)$ are used to assess pseudo-annotation reliability.

Ann. Reliability Let $E_i = \{(\ell_1, e_1), .., (\ell_q, e_q)\}$ be a list of all (entity term, label) pairs in the target sentence T_i. For each $T_i \in \mathbb{T}$, the following two scores, $f(T_i)$ and $g(T_i)$, are measured to estimate the reliability of the pseudo-annotation in T_i:

$$f(T_i) = \frac{\sum_{\forall(l,e) \in E_i} P(\ell|e)}{|E_i|}$$
$$g(T_i) = \frac{\sum_{\forall(l,e) \in E_i} M(\ell|e)}{|E_i|}$$

Given the annotation reliability metrics, our data selection scheme heuristic is as follows:

$$f(T_i) \geq \phi; \quad g(T_i) \leq \gamma; \quad |E_i| \geq \mu; \quad \psi = \texttt{T}|\texttt{F}$$

Only the target sentences satisfying all of the above constraints are used for training given the hyperparameters ψ, α, ϕ, γ, and μ.

5 Experiments

5.1 Direct Model Transfer

The experimental settings of direct model transfer approach are identical with Wu and Dredze (2019). We freeze the bottom n layers (including n) of mBERT, where layer 0 is the embedding layer. The cases of n are {-1, 0, 3, 6, 9}, where -1 denotes fine-tuning all layers in mBERT. For word-level classification, a simple linear classification layer with softmax is added on mBERT. The hyperparameters we experiment on are the combitations of batch size {16, 32}, learning rate {2e-5, 3e-5, 5e-5}, and number of max epochs {3, 4}.

5.2 Multilingual Encoders

For embedding projection and annotation projection, two types of transformer encoders, mBERT (Devlin et al. 2019) and XLM-RoBERTa (Conneau et al. 2020) are considered. Both mBERT and XLM-R are further pretrained on the training data (TRN in Table 1) by individually feeding 1.5M sentences in both English and Korean.[6]

5.3 Embedding Projection

Two types of transformation matrices are derived by the embedding projection method (Section 4.2). One is a word-level matrix and the other is a sentence-level matrix. To evaluate the zero-shot Korean NER model performance (Table 6) when different size of parallel sentences are available, we use different subsets of sentences of increasing sizes(0, 1K, 10K, 100K, 200K, 400K, 747K; 0 to total # of sentences in TRN). Size 0 means the embeddings from source language are not transformed when fed into the NER model for training.

Word embeddings from the last hidden layer of each transformer encoder are extracted. For every parallel sentence pair, let X_i and Y_i be lists of word embeddings of the i'th sentence extracted from the last layer in the source and target encoders, respectively. Only embeddings for words that find alignments are included in X_i and Y_i. If multiple words in the source language, s_i and s_j, are aligned to one word, t_k, in the target language (e.g., United States → 미국 in Figure 1), the embeddings of t_k are duplicated and added to Y_i and vice versa s.t. $|X_*| = |Y_*|$. Let x_{ij} and y_{ij} be the j'th embeddings in X_i and Y_i that are guaranteed to be the embeddings of aligned words; thus, X_i and Y_i are completely in parallel. For the word-level transformation matrix W^w, X_i and Y_i from all parallel sentences are appended together to create $\mathbb{X}_i^w$ and $\mathbb{Y}_i^w$ respectively such that $\mathbb{X}_i^w \cdot W^w \approx \mathbb{Y}_i^w$.

Sentence embeddings are simply created by averaging the word embeddings of parallel source and target sentences. Let X_i' and Y_i' be lists of word embeddings of the i'th sentence extracted from the last layer in the source and target encoders. Note that words in X_i' and Y_i' are not aligned, thus no duplications of word embedding unlike X_i and Y_i. For the sentence-level matrix W^s, the average em-

[6] There are 747.7K parallel sentence pairs in the training set (Table 1a); thus, the total number of individual sentences in both languages is 747.7K × 2 = 1.5M.

beddings of X'_i, Y'_i are appended to create $\mathbb{X}^s_i$ and $\mathbb{Y}^s_i$ such that $\mathbb{X}^s_i \cdot W^s \approx \mathbb{Y}^s_i$.

For each sentence in the source language, embeddings from last hidden layer are transformed by $W^{w|s}$ and fed into the NER model for training (Section 5.5).

5.4 Annotation Projection

The annotation projection is performed to generate the pseudo-annotated Korean dataset (Section 4.3). The following 5 hyperparameters are tuned to filter out noisy annotation for training, where ψ, α, and γ are newly introduced by our work:

- ψ: if True, keep only sentences whose entities are completely matching between the two languages.
- α: the smoothing factor to measure the impurity.
- ϕ: retain sentences whose annotation reliability scores by relative frequency $\geq$ this threshold.
- γ: retain sentences whose annotation reliability scores by impurity $\leq$ this threshold.
- μ: retain sentences that contain named entities whose quantities are $\geq$ this cutoff.

Once the pseudo-annotation is created, all Korean sentences are encoded by mBERT to generate Korean embeddings that are fed into the NER model.

5.5 NER Model

For embedding and annotation projections, a bidirectional LSTM-based NER tagger using a CRF decoder is adapted to build our NER models (Lample et al., 2016). Details of the hyperparameters are described in Appendix A.3

5.6 Results

Table 5 shows the best result of direct model transfer, mBERT fine-tuned on OntoNotes NER dataset and evaluated on our Korean TST set. All scores are reported in a form of mean (± standard deviation) after three developments. The best model is built under the setting when all layers including the embedding layer of mBERT are fine-tuned.

Precision	Recall	F1 score
46.92 (±0.85)	46.28 (±1.49)	46.59 (±1.02)

Table 5: The best result of the baseline model on TST

Table 6 shows the zero-shot results from the embedding projection models in Section 5.3. Both mBERT and XLM-R models showed a performance improvement over 2% with embedding transformation. F1 score improves 2.47% (36.67% to 39.14%) and 2.32% (39.36% to 41.68%) for mBERT and XLM-R, respectively. Both models showed the best performance with embedding transformation matrix made of 200k^w. The number of parallel sentences used for training transformation matrix has a considerable impact on the Zero-shot learning.

	Precision	Recall	F1 score
0	41.63 (±0.2)	32.78 (±1.2)	36.67 (±0.0)
1k^w	35.81 (±0.0)	23.60 (±0.1)	28.45 (±0.0)
10k^w	40.73 (±0.2)	36.43 (±0.6)	38.46 (±0.4)
100k^w	39.70 (±0.4)	36.02 (±0.0)	37.77 (±0.2)
200k^w	**42.17 (±0.0)**	36.52 (±0.2)	**39.14 (±0.1)**
400k^w	41.40 (±0.5)	**36.72 (±0.5)**	38.91 (±0.1)
747k^w	42.07 (±0.0)	36.31 (±0.6)	38.98 (±0.3)
1k^s	27.76 (±2.6)	06.32 (±0.9)	10.29 (±1.4)
10k^s	37.81 (±0.0)	22.20 (±1.3)	27.96 (±1.0)
100k^s	39.19 (±1.0)	29.54 (±0.6)	33.69 (±0.8)
200k^s	40.64 (±0.4)	29.74 (±1.1)	34.33 (±0.6)
400k^s	41.04 (±0.1)	30.20 (±0.6)	34.79 (±0.4)
747k^s	41.56 (±0.5)	30.27 (±0.9)	35.03 (±0.8)

(a) mBERT

	Precision	Recall	F1 score
0	48.04 (±0.5)	33.35 (±0.8)	39.36 (±0.5)
1k^w	42.20 (±0.7)	26.68 (±0.1)	32.69 (±0.2)
10k^w	47.15 (±0.2)	36.19 (±1.3)	40.95 (±1.5)
100k^w	47.87 (±0.1)	**36.57 (±0.5)**	41.46 (±0.3)
200k^w	48.58 (±0.2)	36.50 (±0.7)	**41.68 (±0.4)**
400k^w	47.55 (±0.6)	36.47 (±0.2)	41.28 (±0.1)
747k^w	47.30 (±0.6)	36.00 (±0.5)	40.88 (±0.5)
1k^s	41.02 (±1.9)	15.84 (±1.6)	22.76 (±1.3)
10k^s	46.49 (±1.3)	30.53 (±1.1)	36.84 (±1.0)
100k^s	47.64 (±0.3)	34.29 (±0.4)	39.88 (±0.3)
200k^s	47.47 (±1.0)	33.35 (±0.2)	39.17 (±0.4)
400k^s	48.19 (±0.4)	33.37 (±1.0)	39.43 (±0.8)
747k^s	**48.61 (±1.2)**	33.63 (±1.0)	39.75 (±1.0)

(b) XLM-R

Table 6: Zero-shot NER results on TST using the embedding projection models in §5.3. mBERT/XLM-R$^{w|s}$: the English embeddings from mBERT/XLM-R are transformed by $W^{w|s}$. 0 means not transformed.

Table 7 shows the results from the annotation projection models with various configurations. About 9% gain is shown by the best model using only the entity matching constraint ψ that effectively filters out 55% of the training data (row 3). A relative score of 50.25% is achieved by the model using only 21% of the training data, implying that a fair amount of noisy annotation is produced by the annotation projection approach.

The overall results show that the Annotation Projection approach achieves the best performance,

ψ	α	ϕ	γ	μ	T	Precison	Recall	F1 score
F	0	0	1	0	747K	**61.04 (±0.9)**	32.02 (±0.5)	42.00 (±0.5)
T	0	0	1	0	329K	57.04 (±0.6)	46.15 (±0.4)	**51.02 (±0.4)**
T	0	0.4	1	1	163K	53.87 (±0.2)	47.09 (±0.4)	50.25 (±0.2)
T	0	0.4	1	2	83K	50.43 (±0.8)	**49.49 (±0.3)**	49.95 (±0.4)
T	0.5	0.4	0.5	2	44K	49.92 (±0.8)	48.23 (±0.6)	49.05 (±0.3)
T	0	0.4	0.5	1	91K	49.88 (±1.0)	46.89 (±1.5)	48.31 (±0.4)
T	0.5	0.4	0.5	1	91K	50.95 (±0.4)	46.73 (±0.4)	48.74 (±0.1)
T	0	0.4	0.5	2	44K	47.52 (±0.8)	48.70 (±0.7)	48.10 (±0.2)
F	0	0.4	1	2	435K	57.78 (±1.8)	33.11 (±1.9)	42.06 (±0.5)

Table 7: Zero-shot NER results on TST using the annotation projection models in §5.4. T: number of parallel sentence pairs used for training.

implying that considering the word order of the target language is critical in cross-lingual learning, especially in the case of distant language pairs. We expect further improvement of the annotation projection approach when adapting a more accurate word alignment tool or a data selection scheme, which we will further investigate.

6 Analysis

6.1 Error Distribution

Given the results of the best models for embedding and annotation projection approaches (Section 5.6), a total of 105 parallel sentences (15 pairs per news section) are randomly selected for error analysis.[7] Table 8 shows the distributions of the 5 error types.

	NA	NE	WR	WL	WRL
EP	7.64%	35.03%	28.02%	14.01%	15.28%
AP	10.43%	28.83%	28.83%	23.31%	8.59%

Table 8: Distributions of 5 error types from the samples. EP: embedding projection, AP: annotation projection, NA: no annotation, NE: no extraction, WR: wrong range, WL: wrong label, WRL: wrong range and label.

6.2 Error Analysis

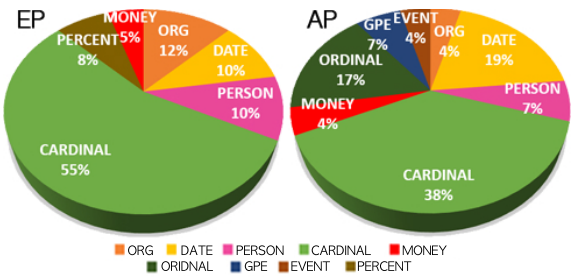

Figure 2: Comparison of entity type distribution of Wrong Range between EP and AP. Only entity types that have errors over 2 times are included in the chart.

Korean named entities describing cardinal are more prone to fall under Wrong Range than the other error types in both models. For example, the Korean entity "10개 (10 things)" comprises the quantity "10" and the metric "개 (things)" that is a generic measure word in Korean, whereas in English just write "10". The grammatical difference between English and Korean, where Korean uses measure words for quantifying the classes of objects while English does not in general, makes it difficult to accurately predict under the zero-shot learning setting.

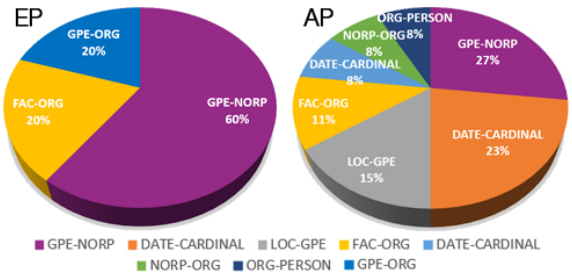

Figure 3: Comparison of entity type pair distribution of Wrong Label between EP and AP. Only entity type pairs that have errors over 2 times are included in the chart.

Wrong Label occurs frequently across all models when dealing with entities referring to nationality. As mentioned in Section 3.5, a single word in Korean can entail the meaning of both nationality and country. This overloaded word-sense characteristic makes entities that actually refer to nationality be mislabeled as GPE, which should have been labeled as NORP.

7 Conclusion

This paper presents a multilingual dataset that allows researchers to conduct crosslingual research between English and Korean. Our dataset contains high-quality annotation of named entities on parallel sentences from seven popular news sections. Given this dataset, Korean NER models are built by zero-shot learning using multilingual encoders. Our data selection scheme for annotation projection significantly improves the NER performance although it is still suboptimal. Our error analysis depicts unique characteristics in Korean that make it hard for zero-shot learning, challenges that we need to overcome in the future work.[8]

[7] Detailed descriptions of the five error types are provided in Appendix A.5.

[8] All our resources including dataset, source codes, and models are available at https://github.com/emorynlp/MRL-2021.

Acknowledgments

This work was partly supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)).

References

Hanan Aldarmaki and Mona Diab. 2019. Context-Aware Cross-Lingual Mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3906–3911, Minneapolis, Minnesota.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. In *Transactions of the Association for Computational Linguistics*, volume 7.

BBN. 2014. OntoNotes Named Entity Guidelines Version 14.0. Raytheon BBN Technologies.

Min-Ah Cheon and Jae-Hoon Kim. 2018. Definition of Korean Named-Entity Task. Technical report, Korea Maritime & Ocean University.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding Universal Grammatical Relations in Multilingual BERT. In *Proceedings of the 58rd Annual Meeting of the Association for Computational Linguistics*, ACL'20.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy.

Key-Sun Choi, Young S. Han, Young G. Han, and Oh W. Kwon. 1994. KAIST Tree Bank Project for Korean: Present and Future Development. In *In Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 7–14.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL'20.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Chung-Hye Han, Na-Rae Han, Eon-Suk Ko, Martha Palmer, and Heejong Yi. 2002. Penn Korean Treebank: Development and Evaluation. In *In Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*, PACLIC'02.

YoonPyo Hong. 2009. 21st Century Sejong Project Results and Tasks (21세기 세종 계획 사업 성과 및 과제). In *New Korean Life (새국어생활)*. National Institute of Korean Language.

Xiaolei Huang, Jonathan May, and Nanyun Peng. 2019. What Matters for Neural Cross-Lingual Named Entity Recognition: An Empirical Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6395–6401, Hong Kong, China.

Soufian Jebbara and Philipp Cimiano. 2019. Zero-Shot Cross-Lingual Opinion Target Extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2486–2495, Minneapolis, Minnesota.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition.

In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 1907.11692.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. VECO: Variable and flexible cross-lingual pretraining for language understanding and generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3980–3994, Online. Association for Computational Linguistics.

Emi Maekawa. 2018. Annotation guidelines for named entities version 1.1. Technical report, Search Results Web results National Institute of Information and Communications Technology.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards Robust Linguistic Analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019a. Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019b. Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F Karlsson, Biqing Huang, and Chin-Yew Lin. 2020. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9274–9281.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Named Entity Tagsets

A.1.1 English NER Tagset

There are 18 named entity tags annotated in the OntoNotes 5.0 as follows (Pradhan et al., 2013):[9]

- **CARDINAL**: Numerical terms not categorized in other categorizations. Numbers that indicate ages are included.
- **DATE**: Absolute or relative dates or periods. The period should last longer than 'TIME'. General expressions of dates are included too such as 'few months', 'that day', 'Next season' and 'First quarter'.
- **EVENT**: It means an official or widely known event, war, exhibition. Official events include ministerial meetings, general elections, presidential elections, exams (SAT), and prayers (U.S. national breakfast prayer). Social phenomena also include (Brexit) for widely known events.
- **FAC**: Objectives referring to facilities include buildings, airports, highways and bridge names.
- **GPE**: An object referring to a place or location, including the name of a country and the name of an administrative district, such as a city or state.
- **LANGUAGE**: Any named language.
- **LAW**: Named documents made into laws.
- **LOC**: Refers to the name of a place or location that does not belong to GPE. It also includes expressions covering the entire location of mountains, rivers, ocean names and Europe, Asia, etc.
- **MONEY**: Monetary values including units.
- **NORP**: It refers to nationality, religious groups and political groups (party).
- **ORDINAL**: All ordinal numbers such as first and second.
- **ORG**: It refers a community / group of people gathered together. For example, the name of the company, the name of the school, and the name of the sports team.
- **PERCENT**: Percentage expressions with % symbol or the word 'percent'.
- **PERSON**: Referring to a last name or full name of a particular person. It also includes nicknames for non-human creatures and characters in cartoons, dramas and movies.
- **PRODUCT**: Vehicles, Weapons, foods. IT services (including SNS) and medicine names are included.
- **QUANTITY**: Measurements as of weights or distances such as km, kg and etc.
- **TIME**: This tag indicates time expressions smaller than a day. This tag includes certain time indication, amount of time or any other expressions related to time. Even though an entity does not have numeral expressions but only words related to time (for instance, 'noon'), the words are tagged as 'Time'.
- **WORK_OF_ART**: Titles of books, songs, TV programs and art pieces. Title of games, awards, theories, records are included.

A.1.2 Korean NER Tagset

There are 10 tags annotated in the copus distributed by the Korea Maritime and Ocean University:[10]

- **DAT**: Absolute dates. Public holidays and day of the week is included.
- **DUR**: Duration of incidents. Academically clarified periods such as Cretaceous period are also included.
- **LOC**: The name of a country and the name of an administrative district, such as a city or state. Words representing certain locations such as tour spot and stadium is also included. When location word becomes compound nouns with other words, it is not included.
- **MNY**: Monetary values including units. Bitcoin is not included.
- **NOH**: Any numerical expressions such as measurements of heights, temperatures, weights. Ordinal numbers are included.
- **ORG**: A group consisting of 2 or more people. The name of the company, the name of the school, and the name of the sports team.

[9] https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf

[10] https://github.com/kmounlp/NER

- **PER**: Personal name including first and last name. Any name referring to living things and nicknames for non-human creatures and characters in cartoons, dramas and movies are included.
- **PNT**: Percentage expressions with % symbol or the word 'percent'.
- **POH**: Product name, medicine, game, event, meeting, movies, songs, drama series, TV channels, daily and weekly magazines, emails, phone numbers are included.
- **TIM**: This tag indicates time expressions smaller than a day. This tag includes certain time indication, amount of time or any other expressions related to time. Even though an entity does not have numeral expressions but only words related to time (for instance, 'noon'), the words are tagged as 'Time'.

A.2 English NER Performance

Table 9 shows the performance of the NER model in the ELIT toolkit on the English development and evaluation sets in our dataset.

Category	Precision	Recall	F1-Score
Business	81.01	80.39	80.70
Lifestyle	77.45	77.85	77.65
Politics	72.46	72.35	72.41
Society	74.19	73.52	73.86
Sci/Tech	80.76	81.20	80.98
Sports	80.07	79.76	79.92
World	81.90	82.02	81.96
Total	78.34	78.22	78.28

(a) Results on the development set.

Category	Precision	Recall	F1-Score
Business	77.44	77.72	77.58
Lifestyle	75.22	74.88	75.05
Politics	71.29	71.21	71.25
Society	71.24	70.48	70.86
Sci/Tech	77.48	77.42	77.45
Sports	80.86	79.91	80.38
World	82.25	82.73	82.49
Total	76.97	76.77	76.87

(b) Results on the evaluation set.

Table 9: Performance of the English NER model.

A.3 Experimental Settings

A.3.1 Task specific NER model for Embedding and Annotation Projections

The task specific NER model used : a 2-layer Bi-LSTM with a hidden size of 768 followed by a CRF layer. A dropout rate of 0.5 is applied on the input and the output of the Bi-LSTM. Adam with default parameters and a learning rate of 0.0001 are used for optimization. We trained the model for 10 epoch with a batch size of 32, and evaluate the model per a epoch.

A.4 Comparison of NER performances (Zero-shot VS Existing)

We compare our best performing Zero-shot Korean NER model with the existing Korean NER model[11] on TST.

A.4.1 Comparison Method

Since the number of named entity types that each model covers are different, named entity tags are mapped based on the definition of the tags. Our named entities are more fine-grained, which makes multiple tags (Zero-shot side) be mapped to one tag (Existing side). Named entity tags that cannot be mapped are discarded in both gold labels and predicted labels, thus not considered in the evaluation of the models.

A.4.2 Comparison Result

Our zero-shot model yields a better performance than the existing model although it may be difficult to directly compare the two models. In the case of the existing Korean model, the low performance may be caused by the different annotation scheme between the datasets. In the case of our zero-shot model, the improvement of the performance are seen due to the coarse-grained named entities after the mapping.

MODEL	Precision	Recall	F1-Score
Zero-shot	0.61	0.55	0.58
Existing	0.46	0.53	0.49

Table 10: NER performances of Zero-shot Korean model and Existing Korean model on TST

A.5 NER Error Types

Underlined words indicate entity boundaries, followed by their TAGs:

No Annotation occurs when the model extracts non-entity words as a named entity although they are not annotated in the gold data.

Gold: 신흥국 (Emerging Country)

Auto:: 신흥국 (Emerging Country)$_{\text{ORG}}$

[11] https://github.com/monologg/KoBERT-NER

`No Extraction` occurs when the model does not extract words as a named entity although they are annotated in the gold data.

Gold: 국회 (National Assembly)$_{\text{ORG}}$
Auto: 국회 (National Assembly)

`Wrong Range` occurs when the model extracts words as a named entity and only the range is wrong. This type of errors are words that are partially annotated with correct named entity tag.

Gold: 도널드 트럼프 (Donald Trump)$_{\text{PERSON}}$
Auto: 도널드 (Donald) 트럼프 (Trump)$_{\text{PERSON}}$

`Wrong Label` occurs when the model extracts words as a named entity with correct boundaries but only the tag type is wrong.

Gold: 하노이 (Hanoi)$_{\text{GPE}}$
Auto: 하노이 (Hanoi)$_{\text{PERSON}}$

`Wrong Range and Label` occurs when the model extract words as a named entity whose both range and tag type are wrong.

Gold: 60세 (60-years-old)$_{\text{CARDINAL}}$
Auto: 60$_{\text{DATE}}$ 세 (years-old)

A.6 Analytics by News Sections and Entity Types

Table 11a describes the proportions of entity types per news section on `DEV` and `TST` combined. For *Lifestyle*, *Politics*, and *Sports*, `PERSON` is the most frequent entity type since many topics in these sections are centered around famous people (e.g., someone's biography, politicians, sports players). `ORG` on the other hand is the most frequent entity type for *Business*, *Society*, and *Sci/Tech* although those entities in *Society* refer to social groups while they generally refer to industrial companies in the other sections. `GPE` is the most and the 2nd-most frequent entity types for *World* and *Politics* where these entities refer to countries or regions in *World* but they are often related to geographical relationships between political figures in *Politics*.

Table 11b describes the proportions of news sections per entity type. `CARDINAL`, `ORDINAL`, and `EVENT` appear the most in *Sports* that involves many game events and statistics. `DATE`, `ORG`, and `QUANTITY` show fairly even proportions in every section as they are elemental to a variety of topics. `GPE`, `LOC`, and `NORP` give high proportions to both *Politics* and *World*. `MONEY` and `PERCENT` appear the most in *Business* that often deals with monetary issues. `PERSON` show high proportions in *Sports* and *Politics* as discussed above. `FAC` takes good portions in *Lifestyle*, *Business*, and *World*, which often mention facilities that people encounter daily (e.g., airports, bridges). `TIME` appears the most in *Sports* and *Society* that are full of dynamic events and issues. `LANGUAGE` is mostly found in *Society* although the sample size is too small to generalize. `LAW`, `PRODUCT`, and `WORK_OF_ART` appear the most in *Politics*, *Sci/Tech*, and *Lifestyle*, that focus on legal issues, tech products, and entertainment (e.g., music, movies, shows), respectively.

	Business	Lifestyle	Politics	Society	Sci/Tech	Sports	World
CARDINAL	284 (10.0)	255 (11.3)	245 (6.4)	348 (13.2)	309 (12.2)	633 (15.4)	278 (6.7)
DATE	465 (16.4)	335 (14.8)	460 (12.0)	455 (17.2)	361 (14.3)	464 (11.3)	467 (11.2)
EVENT	18 (0.6)	51 (2.3)	58 (1.5)	18 (0.7)	35 (1.4)	142 (3.5)	75 (1.8)
FAC	61 (2.1)	71 (3.1)	28 (0.7)	41 (1.6)	26 (1.0)	34 (0.8)	60 (1.4)
GPE	489 (17.2)	230 (10.2)	857 (22.3)	239 (9.1)	235 (9.3)	305 (7.4)	1,101 (26.5)
LANGUAGE	4 (0.1)	4 (0.2)	-	43 (1.6)	-	1 (0.0)	2 (0.0)
LAW	14 (0.5)	2 (0.1)	27 (0.7)	19 (0.7)	13 (0.5)	8 (0.2)	13 (0.3)
LOC	26 (0.9)	56 (2.5)	81 (2.1)	16 (0.6)	28 (1.1)	17 (0.4)	80 (1.9)
MONEY	106 (3.7)	18 (0.8)	12 (0.3)	41 (1.6)	59 (2.3)	27 (0.7)	58 (1.4)
NORP	103 (3.6)	111 (4.9)	384 (10.0)	41 (1.6)	107 (4.2)	104 (2.5)	396 (9.5)
ORDINAL	98 (3.5)	81 (3.6)	95 (2.5)	110 (4.2)	87 (3.4)	300 (7.3)	87 (2.1)
ORG	728 (25.6)	258 (11.4)	609 (15.9)	613 (23.2)	736 (29.1)	822 (20.0)	610 (14.7)
PERCENT	129 (4.5)	8 (0.4)	29 (0.8)	79 (3.0)	55 (2.2)	21 (0.5)	85 (2.0)
PERSON	222 (7.8)	552 (24.4)	877 (22.8)	456 (17.3)	189 (7.5)	1,127 (27.5)	707 (17.0)
PRODUCT	42 (1.5)	28 (1.2)	29 (0.8)	29 (1.1)	163 (6.5)	3 (0.1)	35 (0.8)
QUANTITY	21 (0.7)	16 (0.7)	13 (0.3)	19 (0.7)	22 (0.9)	6 (0.1)	20 (0.5)
TIME	20 (0.7)	37 (1.6)	24 (0.6)	66 (2.5)	39 (1.5)	64 (1.6)	44 (1.1)
WORK_OF_ART	10 (0.4)	151 (6.7)	13 (0.3)	7 (0.3)	62 (2.5)	24 (0.6)	36 (0.9)
Total	2,840 (100.0)	2,264 (100.0)	3,841(100.0)	2,640 (100.0)	2,526 (100.0)	4,102 (100.0)	4,154 (100.0)

(a) Entities distribution by sections.

	Business	Lifestyle	Politics	Society	Sci/Tech	Sports	World	Total
CARDINAL	284 (12.1)	255 (10.8)	245 (10.4)	348 (14.8)	309 (13.1)	633 (26.9)	278 (11.8)	2,352 (100.0)
DATE	465 (15.5)	335 (11.1)	460 (15.3)	455 (15.1)	361 (12.0)	464 (15.4)	467 (15.5)	3,007 (100.0)
EVENT	18 (4.5)	51 (12.8)	58 (14.6)	18 (4.5)	35 (8.8)	142 (35.8)	75 (18.9)	397 (100.0)
FAC	61 (19.0)	71 (22.1)	28 (8.7)	41 (12.8)	26 (8.1)	34 (10.6)	60 (18.7)	321 (100.0)
GPE	489 (14.1)	230 (6.7)	857 (24.8)	239 (6.9)	235 (6.8)	305 (8.8)	1,101 (31.9)	3,456 (100.0)
LANGUAGE	4 (7.4)	4 (7.4)	-	43 (79.6)	-	1 (1.9)	2 (3.7)	54 (100.0)
LAW	14 (14.6)	2 (2.1)	27 (28.1)	19 (19.8)	13 (13.5)	8 (8.3)	13 (13.5)	96 (100.0)
LOC	26 (8.6)	56 (18.4)	81 (26.6)	16 (5.3)	28 (9.2)	17 (5.6)	80 (26.3)	304 (100.0)
MONEY	106 (33.0)	18 (5.6)	12 (3.7)	41 (12.8)	59 (18.4)	27 (8.4)	58 (18.1)	321 (100.0)
NORP	103 (8.3)	111 (8.9)	384 (30.8)	41 (3.3)	107 (8.6)	104 (8.4)	396 (31.8)	1,246 (100.0)
ORDINAL	98 (11.4)	81 (9.4)	95 (11.1)	110 (12.8)	87 (10.1)	300 (35.0)	87 (10.1)	858 (100.0)
ORG	728 (16.6)	258 (5.9)	609 (13.9)	613 (14.0)	736 (16.8)	822 (18.8)	610 (13.9)	4,376 (100.0)
PERCENT	129 (31.8)	8 (2.0)	29 (7.1)	79 (19.5)	55 (13.5)	21 (5.2)	85 (20.9)	406 (100.0)
PERSON	222 (5.4)	552 (13.4)	877 (21.2)	456 (11.0)	189 (4.6)	1,127 (27.3)	707 (17.1)	4,130 (100.0)
PRODUCT	42 (12.8)	28 (8.5)	29 (8.8)	29 (8.8)	163 (49.5)	3 (0.9)	35 (10.6)	329 (100.0)
QUANTITY	21 (17.9)	16 (13.7)	13 (11.1)	19 (16.2)	22 (18.8)	6 (5.1)	20 (17.1)	117 (100.0)
TIME	20 (6.8)	37 (12.6)	24 (8.2)	66 (22.4)	39 (13.3)	64 (21.8)	44 (15.0)	294 (100.0)
WORK_OF_ART	10 (3.3)	151 (49.8)	13 (4.3)	7 (2.3)	62 (20.5)	24 (7.9)	36 (11.9)	303 (100.0)

(b) Categories distribution by entities.

Table 11: Distribution comparisons between manually annotated entities and news sections in the English dataset (DEV and TST combined). Numbers in the parentheses indicate the percentages of the corresponding tags.