KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding

Jiyeon Ham*, **Yo Joong Choe***, **Kyubyong Park***, **Ilji Choi**, **Hyungjoon Soh**
Kakao Brain
{jiyeon.ham,yj.choe,kyubyong.park,ilji.choi,hj.soh}@kakaobrain.com

Abstract

Natural language inference (NLI) and semantic textual similarity (STS) are key tasks in natural language understanding (NLU). Although several benchmark datasets for those tasks have been released in English and a few other languages, there are no publicly available NLI or STS datasets in the Korean language. Motivated by this, we construct and release new datasets for Korean NLI and STS, dubbed KorNLI and KorSTS, respectively. Following previous approaches, we machine-translate existing English training sets and manually translate development and test sets into Korean. To accelerate research on Korean NLU, we also establish baselines on KorNLI and KorSTS. Our datasets are publicly available at https://github.com/kakaobrain/KorNLUDatasets.

1 Introduction

Natural language inference (NLI) and semantic textual similarity (STS) are considered as two of the central tasks in natural language understanding (NLU). They are not only featured in GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019), which are two popular benchmarks for NLU, but also known to be useful for supplementary training of pre-trained language models (Phang et al., 2018) as well as for building and evaluating fixed-size sentence embeddings (Reimers and Gurevych, 2019). Accordingly, several benchmark datasets have been released for both NLI (Bowman et al., 2015; Williams et al., 2018) and STS (Cer et al., 2017) in the English language.

When it comes to the Korean language, however, benchmark datasets for NLI and STS do not exist. Popular benchmark datasets for Korean NLU typically involve question answering[1][2] and sentiment analysis[3], but not NLI or STS. We believe that the lack of publicly available benchmark datasets for Korean NLI and STS has led to the lack of interest for building Korean NLU models suited for these key understanding tasks.

Motivated by this, we construct and release **KorNLI** and **KorSTS**, two new benchmark datasets for NLI and STS in the Korean language. Following previous work (Conneau et al., 2018), we construct our datasets by machine-translating existing English training sets and by translating English development and test sets via human translators. We then establish baselines for both KorNLI and KorSTS to facilitate research on Korean NLU.

2 Background

2.1 NLI and the {S,M,X}NLI Datasets

In an NLI task, a system receives a pair of sentences, a premise and a hypothesis, and classifies their relationship into one out of three categories: *entailment*, *contradiction*, and *neutral*.

There are several publicly available NLI datasets in English. Bowman et al. (2015) introduced the Stanford NLI (SNLI) dataset, which consists of 570K English sentence pairs based on image captions. Williams et al. (2018) introduced the Multi-Genre NLI (MNLI) dataset, which consists of 455K English sentence pairs from ten genres. Conneau et al. (2018) released the Cross-lingual NLI (XNLI) dataset by extending the development and test data of the MNLI corpus to 15 languages. Note that Korean is not one of the 15 languages in XNLI. There are also publicly available NLI datasets in a few other non-English languages (Fonseca et al., 2016; Real et al., 2019; Hayashibe, 2020), but none exists for Korean at the time of publication.

*Equal Contribution.
[1] https://korquad.github.io/ (Lim et al., 2019)
[2] http://www.aihub.or.kr/aidata/84
[3] https://github.com/e9t/nsmc

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 422–430

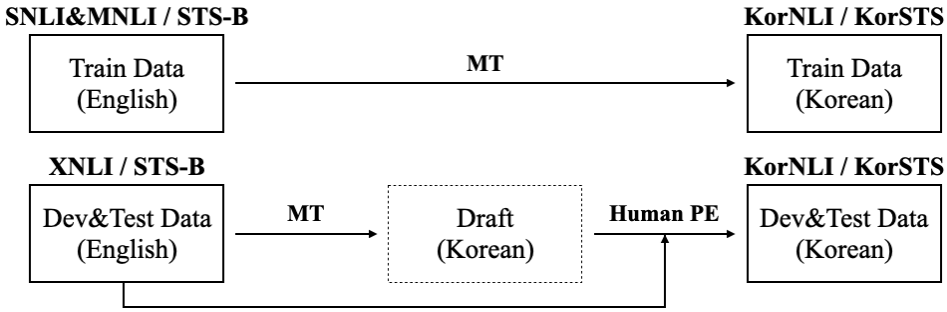

Figure 1: Data construction process. MT and PE indicate machine translation and post-editing, respectively. We translate original English data into Korean using an internal translation engine. For development and test data, the machine translation outputs are further post-edited by human experts.

2.2 STS and the STS-B Dataset

STS is a task that assesses the gradations of semantic similarity between two sentences. The similarity score ranges from 0 (completely dissimilar) to 5 (completely equivalent). It is commonly used to evaluate either how well a model grasps the closeness of two sentences in meaning, or how well a sentence embedding embodies the semantic representation of the sentence.

The STS-B dataset consists of 8,628 English sentence pairs selected from the STS tasks organized in the context of SemEval between 2012 and 2017 (Agirre et al., 2012, 2013, 2014, 2015, 2016). The domain of input sentences covers image captions, news headlines, and user forums. For details, we refer readers to Cer et al. (2017).

3 Data

3.1 Data Construction

We explain how we develop two new Korean language understanding datasets: KorNLI and KorSTS. The KorNLI dataset is derived from three different sources: SNLI, MNLI, and XNLI, while the KorSTS dataset stems from the STS-B dataset. The overall construction process, which is applied identically to the two new datasets, is illustrated in Figure 1.

First, we translate the training sets of the SNLI, MNLI, and STS-B datasets, as well as the development and test sets of the XNLI[4] and STS-B datasets, into Korean using an internal neural machine translation engine. Then, the translation results of the development and test sets are post-edited by professional translators in order to guarantee the quality of evaluation. This multi-stage translation strategy aims not only to expedite the translators' work, but also to help maintain the translation consistency between the training and evaluation datasets. It is worth noting that the post-editing procedure does not simply mean proofreading. Rather, it refers to human translation based on the prior machine translation results, which serve as first drafts.

3.1.1 Translation Quality

To ensure translation quality, we hired two professional translators with at least seven years of experience who specialize in academic papers/books as well as business contracts. The two translators each post-edited half of the dataset and cross-checked each other's translation afterward. This was further examined by one of the authors, who is fluent in both English and Korean.

We also note that the professional translators did not have to edit much during post-editing, suggesting that the machine-translated sentences were often good enough to begin with. We found that the BLEU scores between the machine-translated and post-edited sentences were 63.30 for KorNLI and 73.26 for KorSTS, and for approximately half the time (47% for KorNLI and 53% for KorSTS), the translators did not have to change the machine-translated sentence at all.

Finally, we note that translators did not see the English gold labels during post-editing, in order to expedite the post-editing process. See Section 5 for a discussion on the effect of translation on data quality.

3.2 KorNLI

Table 1 shows the statistics of the KorNLI dataset. There are 942,854 training examples translated automatically and 7,500 evaluation (development and test) examples translated manually. The premises

[4]Only English examples count.

KorNLI	Total	Train	Dev.	Test
Source	-	SNLI, MNLI	XNLI	XNLI
Translated by	-	Machine	Human	Human
# Examples	950,354	942,854	2,490	5,010
# Words (P)	13.6	13.6	13.0	13.1
# Words (H)	7.1	7.2	6.8	6.8

Table 1: Statistics of KorNLI dataset. The last two rows mean the average number of words in a Premise (P) and a Hypothesis (H), respectively.

Examples	Label
P: 너는 거기에 있을 필요 없어. "You don't have to stay there." **H**: 가도 돼. "You can leave."	E
P: 너는 거기에 있을 필요 없어. "You don't have to stay there." **H**: 넌 정확히 그 자리에 있어야 해! "You need to stay in this place exactly!"	C
P: 너는 거기에 있을 필요 없어. "You don't have to stay there." **H**: 네가 원하면 넌 집에 가도 돼. "You can go home if you like."	N

Table 2: Examples from KorNLI dataset. **P**: Premise, **H**: Hypothesis. E: Entailment, C: Contradiction, N: Neutral.

are almost twice as long as the hypotheses, as reported in Conneau et al. (2018). We present a few examples in Table 2.

3.3 KorSTS

As provided in Table 3, the KorSTS dataset comprises 5,749 training examples translated automatically and 2,879 evaluation examples translated manually. Examples are shown in Table 4.

4 Baselines

In this section, we provide baselines for the Korean NLI and STS tasks using our newly created benchmark datasets. Because both tasks receive a pair of sentences as an input, there are two different approaches depending on whether the model encodes the sentences jointly ("cross-encoding") or separately ("bi-encoding").[5]

4.1 Cross-encoding Approaches

As illustrated with BERT (Devlin et al., 2019) and many of its variants, the *de facto* standard approach for NLU tasks is to pre-train a large language model and fine-tune it on each task. In the cross-encoding approach, the pre-trained language model takes each sentence pair as a single input for fine-tuning. These cross-encoding models typically achieve the state-of-the-art performance over bi-encoding models, which encode each input sentence separately.

[5]These nomenclatures (cross-encoding and bi-encoding) are adopted from Humeau et al. (2020).

KorSTS	Total	Train	Dev.	Test
Source	-	STS-B	STS-B	STS-B
Translated by	-	Machine	Human	Human
# Examples	8,628	5,749	1,500	1,379
Avg. # Words	7.7	7.5	8.7	7.6

Table 3: Statistics of KorSTS dataset.

Examples	Score
A: 한 남자가 음식을 먹고 있다. "A man is eating food." **B**: 한 남자가 뭔가를 먹고 있다. "A man is eating something."	4.2
A: 한 여성이 고기를 요리하고 있다. "A woman is cooking meat." **B**: 한 남자가 말하고 있다. "A man is speaking."	0.0

Table 4: Examples from KorSTS dataset.

For both KorNLI and KorSTS, we consider two pre-trained language models. We first pre-train a Korean RoBERTa (Liu et al., 2019), both base and large versions, on a collection of internally collected Korean corpora (65GB). We construct a byte pair encoding (BPE) (Gage, 1994; Sennrich et al., 2016) dictionary of 32K tokens using SentencePiece (Kudo and Richardson, 2018). We train our models using `fairseq` (Ott et al., 2019) with 32 V100 GPUs for the base model (25 days) and 64 for the large model (20 days).

We also use XLM-R (Conneau and Lample, 2019), a publicly available cross-lingual language model that was pre-trained on 2.5TB of Common

Model	# Params.	†KorNLI	KorSTS
Fine-tuned on Korean training set			
Korean RoBERTa (base)	111M	82.75	83.00
Korean RoBERTa (large)	338M	**83.67**	**85.27**
XLM-R (base)	270M	80.56	77.78
XLM-R (large)	550M	83.41	84.68
Fine-tuned on English training set (Cross-lingual Transfer)			
XLM-R (base)	270M	75.17	-
XLM-R (large)	550M	80.30	-

Table 5: KorNLI and KorSTS test set scores for fine-tuned *cross-encoding* language models. KorNLI scores are accuracy (%) and KorSTS scores are 100 × Spearman correlation. †To ensure comparability with XNLI, we only use the MNLI portion of the KorNLI dataset.

Model	# Params.	KorSTS			
		Unsupervised		Supervised	
		-	*Trained on: KorNLI*	*Trained on: KorSTS*	*Trained on: KorNLI → KorSTS*
Korean fastText	-	47.96	-	-	-
M-USE$_{CNN}$ (base)	68.9M	-	†72.74	-	-
M-USE$_{CNN}$ (large)	85.2M	-	†76.32	-	-
Korean SRoBERTa (base)	111M	48.96	74.19	78.94	80.29
Korean SRoBERTa (large)	338M	51.35	75.46	**79.55**	80.49
SXLM-R (base)	270M	45.05	73.99	68.36	79.13
SXLM-R (large)	550M	39.92	**77.01**	77.71	**81.84**

Table 6: KorSTS test set scores (100 × Spearman correlation) of *bi-encoding* models. Note that the first two columns of results are unsupervised w.r.t. KorSTS, and the latter two are supervised w.r.t. KorSTS. †Trained on machine-translated SNLI only.

Crawl corpora in 100 languages including Korean (54GB). Note that the base and large architectures of XLM-R are identical to those of RoBERTa, except that the vocabulary size is significantly larger (250K), making the embedding and output layers that much larger.

In Table 5, we report the test set scores for cross-encoding models fine-tuned on KorNLI (accuracy) and KorSTS (Spearman correlation). For KorNLI, we additionally include results for XLM-R models fine-tuned on the original MNLI training set (also known as *cross-lingual transfer* in XNLI). To ensure comparability across settings, we only train on the MNLI portion when fine-tuning on KorNLI.

Overall, the Korean RoBERTa models outperform the XLM-R models, regardless of whether they are fine-tuned on Korean or English training sets. For each model, the larger variant outperforms the base one, consistent with previous findings. The large version of Korean RoBERTa performs the best for both KorNLI (83.67%) and KorSTS (85.27%) among all models tested. Among the XLM-R models for KorNLI, those fine-tuned on the Korean training set consistently outperform the cross-lingual transfer variants.

4.2 Bi-encoding Approaches

We also report the KorSTS scores of bi-encoding models. The bi-encoding approach bears practical importance in applications such as semantic search, where computing pairwise similarity among a large set of sentences is computationally expensive with cross-encoding.

Here, we first provide two baselines that do not use pre-trained language models: Korean fastText and the multilingual universal sentence encoder (M-USE). Korean fastText (Bojanowski et al., 2017) is a pre-trained word embedding model[6] trained on Korean text from Common Crawl. To produce sentence embeddings, we take the average of fastText word embeddings for each sentence. M-USE[7] (Yang et al., 2019), is a CNN-based sentence encoder model trained for NLI, question-answering, and translation ranking across 16 languages including Korean. For both Korean fastText and M-USE, we compute the cosine similarity between two input sentence embeddings to make an unsupervised STS prediction.

Pre-trained language models can also be used as bi-encoding models following the approach of SentenceBERT (Reimers and Gurevych, 2019), which involves fine-tuning a BERT-like model with a Siamese network structure on NLI and/or STS. We use the SentenceBERT approach for both Korean RoBERTa ("Korean SRoBERTa") and XLM-R ("SXLM-R"). We adopt the `MEAN` pooling strategy, i.e., computing the sentence vector as the mean of all contextualized word vectors.

In Table 6, we present the KorSTS test set scores (100 × Spearman correlation) for the bi-encoding models. We categorize each result based on whether the model was additionally trained on KorNLI and/or KorSTS. Note that models that are not fine-tuned at all or only fine-tuned to KorNLI can be considered as unsupervised w.r.t. KorSTS. Also note that M-USE is trained on a machine-translated version of SNLI, which is a subset of KorNLI, as part of its pre-training step.

[6] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.ko.300.bin.gz

[7] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3

First, given each model, we find that supplementary training on KorNLI consistently improves the KorSTS scores for both unsupervised and supervised settings, as was the case with English models (Conneau et al., 2017; Reimers and Gurevych, 2019). This shows that the KorNLI dataset can be an effective intermediate training source for bi-encoding approaches. When comparing the baseline models in each setting, we find that both M-USE and the SentenceBERT-based models trained on KorNLI achieve competitive unsupervised KorSTS scores. Both models significantly outperform the average of fastText embeddings model and the Korean SRoBERTa and SXLM-R models without fine-tuning. Among our baselines, large SXLM-R trained on KorNLI followed by KorSTS achieves the best score (81.84).

5 Effect of Translation on Data Quality

As noted in (Conneau et al., 2018), translation quality does not necessarily guarantee that the semantic relationships between sentences are preserved. We also translated each sentence independently and took the gold labels from the original English pair, so the resulting label might no longer be "gold," due to both incorrect translations and (in rarer cases) linguistic differences that make it difficult to translate specific concepts.

Fortunately, it was also pointed out in (Conneau et al., 2018) that annotators could recover the NLI labels at a similar accuracy in translated pairs (83% in French) as in original pairs (85% in English). In addition, our baseline experiments in Section 4.1 show that supplementary training on KorNLI improves KorSTS performance (+1% for RoBERTa and +4-11% for XLM-R), suggesting that the labels of KorNLI are still meaningful. Another quantitative evidence is that the performance of XLM-R fine-tuned on KorNLI (80.3% with cross-lingual transfer) is within a comparable range of the model's performance on other XNLI languages (80.1% on average).

Nevertheless, we could also find some (not many) examples the gold label becomes incorrect after translating input sentences to Korean. For example, there were cases in which the two input sentences for KorSTS were so similar (with 4+ similarity scores) that upon translation, the two inputs simply became identical. In another case, the English word *sir* appeared in the premise of an NLI example and was translated to 선생님, which is a correct word translation but is a gender-neutral noun, because there is no gender-specific counterpart to the word in Korean. As a result, when the hypothesis referencing the entity as *the man* got translated into 남자 (gender-specific), the English gold label (entailment) was no longer correct in the translated example. More systematically analyzing these errors is an interesting future work, although the amount of human efforts involved in this analysis would match that of labeling a new dataset.

6 Conclusion

We introduced KorNLI and KorSTS—new datasets for Korean natural language understanding. Using these datasets, we also established baselines for Korean NLI and STS with both cross-encoding and bi-encoding approaches. Looking forward, we hope that our datasets and baselines will facilitate future research on not only improving Korean NLU systems but also increasing language diversity in NLU research.

Acknowledgements

We thank Pulip Park for helping with hiring and contacting with the professional translators. We would also like to acknowledge Kakao Brain Cloud, which we used for our baseline experiments.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA.*

Stroudsburg (PA): ACL; 2016. p. 497-511. ACL (Association for Computational Linguistics).

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Yuta Hayashibe. 2020. Japanese realistic textual entailment corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6827–6834, Marseille, France. European Language Resources Association.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

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. Association for Computational Linguistics.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1. 0: Korean qa dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

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 preprint arXiv:1907.11692*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2019. Organizing the assin 2 shared task. In *ASSIN@ STIL*, pages 1–13.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual universal sentence encoder for semantic retrieval.

A Korean RoBERTa Pre-training

For the Korean RoBERTa baselines used in §4, we pre-train a RoBERTa (Liu et al., 2019) model on an internal Korean corpora of size 65GB, consisting of online news articles (56GB), encyclopedia (7GB), movie subtitles (∼1GB), and the Sejong corpus[8] (∼0.5GB). We use `fairseq` (Ott et al., 2019), which includes the official implementation for RoBERTa.

In Table 7, we list all hyperparameters we use for Korean RoBERTa pre-training. Note that, compared to the original RoBERTa (English), the model architectures are identical except for the token embedding layer, as we use different vocabularies (32K `sentencepiece` vocab instead of 50K byte-level BPE). After training, the base and large models achieve validation perplexities of 2.55 and 2.39 respectively, where the validation set is a random 5% subset of the entire corpora.

Hyperparameter	Large	Base
Total # of Parameters	338M	111M
Number of Layers	24	12
Hidden Size	1024	768
FFN Inner Hidden Size	4096	3072
Attention Heads	16	12
Attention Head Size	64	64
Dropout	0.1	0.1
Attention Dropout	0.1	0.1
Warmup Steps	30K	24K
Peak Learning Rate	2e-4	6e-4
Batch Size	2048	8192
Weight Decay	0.01	0.01
Scheduled # Updates	2M	500K
Performed # Updates*	502.3K	500K
Learning Rate Decay	Linear	Linear
Adam ϵ	1e-6	1e-6
Adam β_1	0.9	0.9
Adam β_2	0.98	0.98
Gradient Clipping	0.0	0.0

Table 7: Hyperparameters for Korean RoBERTa pre-training. *For the large model, we initially scheduled our learning rate to decay to zero at 2M steps. After 500K steps, however, we observed no significant improvement in the KorNLI and KorSTS fine-tuning performance.

B Fine-tuning with Cross-encoding Approaches

To fine-tune Korean RoBERTa and XLM-R models using the cross-encoding approach (§4.1), we follow the fine-tuning procedures of RoBERTa (Liu et al., 2019) on MNLI and STS-B, as described in RoBERTa's code release[9].

Hyperparameter	KorNLI	KorSTS
Batch Size	32	16
Learning Rate Schedule	Linear	Linear
Peak Learning Rate	1e-5	2e-5
# Warmup Steps	7318	214
Total # Updates	121979	3596

Table 8: Hyperparameters for Korean RoBERTa and XLM-R fine-tuning using the *cross-encoding* approach.

[8] https://ithub.korean.go.kr/user/guide/corpus/guide1.do

[9] https://github.com/pytorch/fairseq/blob/v0.9.0/examples/roberta/README.glue.md

The fine-tuning hyperparameters are summarized in Table 8. For each dataset and model size, we choose the hyperparameter configurations that are used in the corresponding English version of the dataset and model size (except for the XLM-R cross-lingual transfer using MNLI, where we also use the same hyperparameters as RoBERTa and XLM-R on KorNLI). We find that the hyperparameters used for English models and datasets give sufficiently good performances on the development set, so we do not perform an additional hyperparameter search. After training each model for 10 epochs, we choose the model checkpoint that achieve the highest score on the development set and evaluate it on the test set to obtain our final results in §4.1.

We also report the development set scores for the best checkpoint in Table 9. We observe that the XLM-R models fine-tuned on KorNLI and KorSTS achieve the highest scores on the development set, although the Korean RoBERTa models perform better on the test set (Table 5 in §4.1). Both models outperform the cross-lingual transfer models on the development set, as is the case on the test set.

Model	# Params.	†KorNLI	KorSTS
Fine-tuned on Korean training set			
Korean RoBERTa (base)	111M	81.97	84.97
Korean RoBERTa (large)	338M	83.17	87.82
XLM-R (base)	270M	79.20	83.02
XLM-R (large)	550M	**84.42**	**88.37**
Fine-tuned on English training set (Cross-lingual Transfer)			
XLM-R (base)	270M	74.34	-
XLM-R (large)	550M	81.45	-

Table 9: KorNLI and KorSTS **development** set scores for fine-tuned *cross-encoding* language models. KorNLI scores are accuracy (%) and KorSTS scores are 100 × Spearman correlation. †To ensure comparability with XNLI, we only use the MNLI portion of the KorNLI dataset.

C Fine-tuning with Bi-encoding Approaches

To fine-tune Korean RoBERTa and XLM-R models using the bi-encoding approach (§4.2), we train Korean Sentence RoBERTa ("Korean SRoBERTa") and Sentence XLM-R ("SXLM-R"), following the fine-tuning procedure of SentenceBERT (Reimers and Gurevych, 2019).

Unless described otherwise, we follow the experimental settings, including all hyperparameters, of SentenceBERT[10]. For each model size, we manually search among learning rates {2e-5, 1e-5} for training on KorNLI, {1e-5, 2e-6} for training on KorSTS, and {1e-5, 2e-6} for training on KorSTS after KorNLI. After training until convergence, we choose the learning rate that lead to the highest KorSTS score on the development set. These hyperparameters are shown in Table 10.

Model	KorNLI	KorSTS	KorSTS (after KorNLI)
Korean SRoBERTa (base)	2e-5	1e-5	1e-5
Korean SRoBERTa (large)	2e-5	1e-5	1e-5
SXLM-R (base)	2e-5	1e-5	1e-5
SXLM-R (large)	1e-5	2e-6	1e-5

Table 10: Learning rates for Korean SRoBERTa and SXLM-R fine-tuning using the *bi-encoding* approach.

We report the development set scores in Table 11. Korean SRoBERTa (large) achieves the best development set performance on both supervised settings, but SXLM-R (large) achieves the best performance for the *KorNLI* → *KorSTS* setting on test set.

[10] https://github.com/UKPLab/sentence-transformers

Model	# Params.	KorSTS			
		Unsupervised		Supervised	
		-	*Trained on: KorNLI*	*Trained on: KorSTS*	*Trained on: KorNLI → KorSTS*
Korean SRoBERTa (base)	111M	63.34	76.48	83.68	83.54
Korean SRoBERTa (large)	338M	60.15	77.95	**84.74**	**84.21**
SXLM-R (base)	270M	64.27	77.65	74.60	81.95
SXLM-R (large)	550M	55.00	**79.16**	82.66	84.13

Table 11: KorSTS **development** set scores (100 × Spearman correlation) of *bi-encoding* models. Note that the first two columns of results are unsupervised w.r.t. KorSTS, and the latter two are supervised w.r.t. KorSTS.