Master's Thesis of Arts

English to Korean Multilingual Transfer Learning with Sentence-BERT

Sentence-BERT를 활용한 영어-한국어 전이 학습

February 2022

Graduate School of Humanities
Seoul National University
Linguistics Major

Suyeon Bock

English to Korean Multilingual Transfer Learning with Sentence-BERT

Name of Examiner

Submitting a master's thesis of Arts

February 2022

Graduate School of Humanities
Seoul National University
Linguistics Major

Suyeon Bock

Confirming the master's thesis written by
Suyeon Bock
February 2022

Chair _____(Seal)
Vice Chair _____(Seal)
Examiner _____(Seal)

Abstract

Bock, Suyeon
Department of Linguistics
The Graduate School of Humanities
Seoul National University

This study focuses on constructing a Korean Sentence-BERT model in a novel method, using student-teacher knowledge distillation. The limitations of BERT have been well explored in previous publications. BERT has proven to be ineffective in deriving sentence-level embeddings and not applicable in practical situations where large amounts of sentence-level embeddings are required, such as document classification and clustering. Sentence-BERT was developed to alleviate these issues and create a model that can derive sentence embeddings in an efficient and accurate manner.

This study explores a transfer learning method in Sentence-BERT, which allows even low-resource language models to leverage the power of models trained in high-resource languages such as Korean. Using translated sentence pairs in the source and target languages, the student model learns to map the translated sentence to the same points in the vector space as the teacher model using a simple mean squared error loss method. In this experiment, an English model was used as the teacher model and a cross-linguistic model was used as the student model. To the

knowledge of this author, no Korean Sentence-BERT model has been trained using this novel method to the date of publication of this paper.

To conduct this knowledge distillation for Sentence-BERT, a large number of source and target language translated sentence pairs are needed. After collecting available datasets on the web, the data was augmented with crawled data from the web, which was then aligned using a novel method and then pre-processed for cleaning. This research evaluates the model trained on this data using the knowledge distillation method on sentence-level tasks and multilingual tasks. The model successfully performs well on all tasks, proving its wide applicability and cross-lingual abilities.

Keywords : Natural Language Processing, Language Modeling, BERT, Word Embeddings, Sentence Embeddings, Semantic Similarity

Student Number : 2020-25350

Table of Contents

List of Figures

List of Tables

1. Introduction

This publication presents a Korean Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) model that has been trained using a multilingual transfer learning method. This training method has the advantage of acquiring high levels of performance using a relatively small training dataset. This study evaluates the trained model on two existing Korean STS tasks and a novel cross-lingual task for evaluation.

Recent leaping advances in Natural Language Processing (NLP) tasks can largely be accredited to performance-enhancing developments in world embedding methods. From Word2Vec (Mikolov et. al., 2013) to BERT (Devlin et al., 2019), vector representations of natural languages have become more intricate, which in turn has led to advancements in traditional NLP tasks. However, most state-of-the-art research is currently being conducted in English. This is not only because many top-tier research facilities located in English-speaking nations, but also because textual data available to researchers skews heavily in favor of English-centric research. For this reason, English-trained language models lead in performance while those trained in other languages attempt to follow.

The performance of such complex NLP models is dependent on having large amounts of reliable and clean data for training. Although obtaining data of such magnitude is not a pressing problem for high-resource languages such as English and German, the same is not true for low-resource languages, for which data of a large scale is more difficult to come across. To alleviate the discrepancy in the size of training data, researchers have continually experimented with cross-lingual transfer learning methods. Transfer learning refers to a machine learning technique that distills knowledge gained in one model to give that knowledge to another

model that is lower in its performance. In the case of multilingual transfer learning in NLP, it ha become popular to take a high-performing English trained model and transfer its knowledge to that of a different language. The idea is that semantic and syntactic information acquired from training in English can be transferred to another model to improve its overall performance.

This paper will examine such transfer learning techniques in NLP from English to Korean using SBERT, a modification of the widely known BERT model. SBERT was specifically chosen for as the model for this experiment because it has shown to be efficient and accurate in deriving sentence embeddings, reaching state-of-the-art levels in English tasks, and because it is easily adaptable to many tasks that require sentence-level embeddings. The model is first pre-trained using parallel English and Korean data, part of which was crawled for this research. Then the model is fine-tuned on several Korean datasets for evaluation, including traditional STS tasks but also on a novel cross-lingual task.

The paper is structured in the following way. First, previous works, including BERT and SBERT will be discussed. Then, the methods used for data collection will be described. Next, the SBERT multilingual transfer learning method will be applied to train a new English and Korean SBERT model. Finally, the evaluation metrics of the final Korean model will be presented through several NLP tasks.

1.1. Purpose of Research

The purpose of this research is to explore the knowledge distillation method for Sentence-BERT. As a unique transfer learning method that can be utilized for Transformer models, it is an area of further research that can be explored in depth. By using this method, this research will train a new Sentence-BERT model that can function in both Korean and English.

Another purpose of this research is to explore parallel data collection and alignment methods. To conduct the previously mentioned knowledge distillation training method, large amounts of parallel sentences are required as training data. As most researchers know, the quality and quantity of training data are important factors that determine the success of any machine learning experiment. Thus, an effective and accurate method of collecting and aligning multilingual sentence pairs in the target and sources languages are explored in this paper.

Lastly, although multilingual models have many practical applications such as in machine translation or in document clustering for a search engine, it is difficult to find a multilingual task that will test the model's capabilities in an academic settings to prove its capabilities. A multilingual task is especially difficult to come across for a low-resource language such as Korean. This paper will propose one evaluation method for testing an English and Korean model's cross-lingual abilities.

2. Related Works

This section will introduce the model architectures of BERT and SBERT, including their setup, training methods, and advantages and disadvantages. Then, this section will discuss the currently existing pre-trained Korean SBERT models.

2.1 BERT

2.1.1 Setup

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a pre-trained language model based on the Transformer (Vaswani et al., 2017) encoder. The input of the model is given as a concatenation of two sentences that are separated and surrounded by special tokens. For instance, where x is the first sentence of length n and y is the second sentence also of length n: [CLS] x_1, ..., x_n [SEP] y_1, ..., y_n [EOS]. The first token of the sequence is the special classification token, [CLS]. In a classification task, this special token's final hidden state is utilized to predict a label.

To ensure the model can differentiate between the two segments, two measures are taken. First, the two segments are always separated by the special [SEP] token. Second, segment embeddings are added to the end of the token embeddings. The last token of the sequence is always the special end-of-sentence token, [EOS]. The length of the total sequence is limited by a value T, a parameter that can be adjusted. Unstructured data is given in this format as input to the model for pre-training. During fine-tuning, the input format may be modified according to the task's needs.

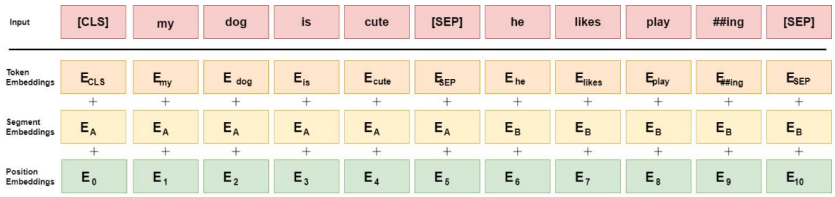


Figure 1. The BERT input representation adapted from Devlin et al. (2018). Instead of using regression, the positional embeddings allow the model to ingest sequential information of language.

2.1.2 Model Architecture

The model architecture is fundamentally built of a stack of Transformer encoder layers. The base model has 12 encoder layers, and the large model has 24 encoder layers. The more encoder layers a model has, the more intricate and accurate its performance. Each layer has multiple heads where self-attention is calculated by computing the key K, value V, and query Q vectors for every input token. The base model has 768 heads while the large model has 1024 heads. Through the attention mechanism, the head will obtain the attention weights between all tokens in the sequence as softmax normalized dot products of the query and key vectors, using the formula outlined in Equation 1. When the attention score is calculated from each head, they are combined and sent to a fully connected layer. Each of these fully connected layers is followed by a normalization layer.

$$\text{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

2.1.3 Pre-training Tasks

As stated previously, BERT is first pre-trained using unlabeled data. Two tasks are defined for the pre-training. The first is the Masked Language Modeling (MLM) task. 15% of the total data is masked with a [MASK] token, and the model is trained to predict the original token. Unlike traditional language models, this task allows BERT to learn bidirectional knowledge in the data. However, this task creates a discrepancy between the pre-training data and fine-tuning data, as [MASK] tokens do not occur in the latter. To narrow this gap, the task is designed so that not all 15% of the data is replaced with the [MASK] token. Instead, of the 15%, (1) 80% is masked with the [MASK] token, (2) 10% is replaced with a random token, and (3) 10% of the data is unchanged.

The second task used for pre-training BERT is Next Sentence Prediction (NSP). The NSP task is meant to model the relationship between two sentences, which is an important aspect of many NLP tasks such as question-answering (QA) and semantic textual similarity (STS). For this purpose, when the input data is prepared as two concatenated segments, segment y is the actual segment that follows segment x just half of the time. The model is trained to predict IsNext or NotNext. However, the effectiveness of this second task has been proven to be detrimental to the overall performance of the model by many researchers (Liu et al., 2019). For such reasons, NSP is often disregarded in current modifications of the BERT architecture.

2.1.4 Advantages and Disadvantages of BERT

The authors of the BERT publication evaluated the model across eleven NLP tasks and achieved state-of-the-art performances across the board. Although many of its original performance records have been wiped out by more recent advancements, BERT remains a cornerstone model in NLP research,

inspiring derivative models and ablation studies, both furthering model performance and our understanding of the mysteries of NLP. This section will examine the innovations of BERT that changed the landscape of language models.

First, it streamlines the language model architecture. Unlike many previous models, BERT has an elegantly simple structure: a stack of Transformer encoder layers. Not only is this simplicity good for simplicity's sake, but it also enables the model to make more efficient calculations, resulting in shorter computation times. This allows the model to train on larger amounts of data, ultimately learning to generate more accurate word embeddings.

Secondly, the idea of a pre-trained model was still relatively novel in the field of NLP at the time of BERT's inception, largely because it was difficult to fathom training a language model on unstructured data. The authors of BERT were able to design two pre-training tasks that allowed BERT to utilize the large inventory of unstructured, unlabeled text data available on the internet. Previous to BERT, only structured data that was difficult to come across and challenging to create was used to train models. Now, large sets of unstructured data from the web such as Twitter data or Reddit data can be used to train models for NLP with minimal preprocessing.

Despite swiping the leaderboard across multiple NLP tasks and constructing an efficient model architecture, BERT is not without its shortcomings. Because BERT uses a cross-encoder network, it is unsuitable for certain tasks that require sentence-level embeddings, especially when input sentences are long. Moreover, although the last hidden layer of the [CLS] token is usually used for sentence classification tasks, this method has later been shown to have inaccurate results (Reimers and Gurevych, 2019).

SBERT attempts to alleviate these issues by constructing an architecture that derives meaningful sentence embeddings. The details of SBERT are discussed in the next section.

2.2 Sentence-BERT

2.2.1 Setup

SBERT (Reimers and Gurevych, 2019) is a modification of the original BERT model that uses siamese and triplet networks to efficiently derive fixed-size vectors for input sentences. Instead of training from scratch, SBERT is fine-tuned from models that are already pre-trained, such as BERT or RoBERTa (Liu et al., 2019). This drastically decreases the needed training time of SBERT compared to that of previous neural sentence embedding models, allowing it to be efficiently applied to unsupervised tasks such as clustering and supervised tasks such as classification or STS.

2.2.2 Model Architecture

To extract a fixed-size sentence embedding vector for sentences of different lengths, SBERT adds a pooling operation to the output of BERT or the pre-trained model that has been fine-tuned. The authors of the paper experiment with three pooling strategies: (1) using the output of the [CLS] token (CLS), (2) computing the mean of all output vectors (MEAN), and (3) computing a max-over-time of the output vectors (MAX). They found that the MEAN pooling method consistently outperformed MAX and CLS method across classification and regression tasks and MEAN is the default pooling method for SBERT.

The specific model architecture depends on the objective function. The authors of SBERT experiment with three structures and objective functions.

1. **Classification Objective Function.** Sentence embeddings u and v are concatenated with their element-wise difference $|u-v|$. The resulting vector is multiplied

with trainable weight W and optimized with cross-entropy loss. Refer to Equation 2 below.

$$o = softmax(W_t(u, v, |u - v|)) \quad (2)$$

2. **Regression Objective Function.** The cosine similarity of sentence embedding u and v is calculated. Then the mean-squared-error loss is used as the objective function.

3. **Triplet Objective Function.** Whereas the previous two structures were siamese networks, the triplet objective function uses a triplet network. Three sentences are given as input: an anchor sentence a, a positive sentence p, and a negative sentence n. The objective is to train the network so that the distance between a and p is always smaller than that of a and n. The loss function is as Equation 3 below. Sentence embeddings for a, n, p is represented as s_x and the margin epsilon ensures that the positive sentence is at least epsilon closer to the anchor sentence than the negative sentence. In their experiments, epsilon is set as 1.

$$max\left(\left||s_a - s_p|\right| - \left||s_a - s_n|\right| + \epsilon, 0\right) \quad (3)$$

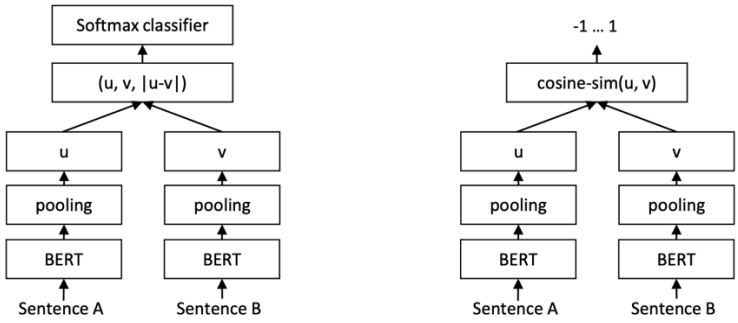


Figure 2. SBERT model architectures from Reimers and Gurevych, 2019. The classification objective function structure is on the left and the regression objective function structure is on the right.

2.2.3 Training and Evaluation

While BERT is trained on unlabeled data, SBERT requires quality data made up of accurately labeled sentence pairs. The original authors of SBERT fine-tune their model on a combination of the SNLI (Bowman et al., 2015) and the Multi-Genre NLI (Williams et al., 2018) datasets, containing 570,000 pairs and 430,000 pairs respectively.

After fine-tuning with the datasets above, the model is further trained and fine-tuned on several Semantic Textual Similarity (STS) tasks and the SentEval task (Conneau and Kiela, 2018) for evaluation. The details of all the evaluation tasks will not be discussed here, but it is worth noting that the SBERT-NLI-large was able to outperform state-of-the-art sentence encoder models by more than 4 points in the unsupervised STS benchmark (STSb) (Cer et al., 2017) evaluation task when both models were not trained on any STS data. This result delineates that even with a relatively small training dataset, SBERT was able to effectively capture semantic relationships between sentences.

2.2.4 Advantages and Disadvantages of SBERT

The obvious advantage of SBERT is that it can be efficiently fine-tuned by leveraging information retained by large pre-trained models such as BERT or RoBERTa. Thus, the amount of training data and training time needed for fine-tuning is drastically decreased compared to models trained from scratch. Another advantage is the computation speed of SBERT. In terms of computation speed, SBERT outperforms all other existing sentence embedding models on the GPU with smart batching by up to 55% (Reimers and Gurevych, 2019). This is advantageous because for tasks such as document classification or clustering, millions of sentences may need to be processed.

However, a disadvantage of SBERT is that to learn to derive accurate sentence embeddings, SBERT requires labeled sentence pairs that are of high quality. While English has the datasets, including the SNLI and STSb benchmarks, many other languages do not have the same luxury. To offset this imbalance, the authors of SBERT presented a knowledge distillation architecture that allows SBERT models of different languages to learn from the high-performing English model (Reimers and Gurevych, 2020). This architecture is used for training the English and Korean multilingual model presented in this paper. The specific details of the knowledge distillation architecture are discussed in Section 3.

2.3 Existing Korean SBERT Models

On the date of publishment, there are two existing Korean SBERT models currently publicly available. The first model, KR-SBERT[①], is developed by Computational Linguistics Lab at Seoul National University. KR-SBERT uses the KR-BERT-V40K[②] as its pre-trained model and is fine-tuned using the KLUE-NLI (Park, Moon, Kim, Cho et al., 2021) and KorSTS (Ham, Choe, Park et al., 2020) datasets. The KorSTS dataset is also augmented to obtain a larger training data set. The second model, Ko-SBERT[③], uses ETRI KorBERT[④] as its pre-trained model and is fine-tuned for SBERT using the KorNLU dataset, which includes KorNLI and KorSTS (Ham, Choe, Park et al., 2020). Both these models are trained using the SBERT architecture outlined in Section 2.2 and can be used to perform the same tasks as those outlined in the same section.

① https://github.com/snunlp/KR-SBERT
② https://github.com/snunlp/KR-BERT
③ https://github.com/BM-K/KoSentenceBERT
④ https://aiopen.etri.re.kr/service_dataset.php

3. Multilingual Transfer Learning for SBERT

In this section, the student−teacher architecture used for multilingual transfer learning between SBERT models, presented by Reimers and Gurevych (2020), will be reviewed. This architecture is the backbone to this experiment and will be used to create a multilingual English and Korean SBERT model from an English monolingual model.

3.1 Training Architecture

The core idea behind the SBERT knowledge distillation mechanism is that translated sentence pairs should be mapped to the same point in the vector space, as they should have the same semantic connotations.

The mechanism for achieving this is elegantly simple. A fine−tuned teacher model A that is trained in the source language s will train a student model $\hat{A}$ in the target language t using a set of n parallel translated sentence pairs of the source and target languages $((s_1, t_1)\ldots(s_n, t_n))$. The student model $\hat{A}$ will learn to map t_x to the same point in the vector space as s_x using a mean squared loss. The teacher model should obviously be a high−performing model that has been fine−tuned with reliable and accurate datasets. The objective function can be summarized mathematically as below in Equation 4 and a visualization of the architecture can be found in Figure 3.

$$\begin{aligned} \hat{A}(s_i) &\approx A(s_i) \\ \hat{A}(t_i) &\approx A(s_i) \end{aligned} \tag{4}$$

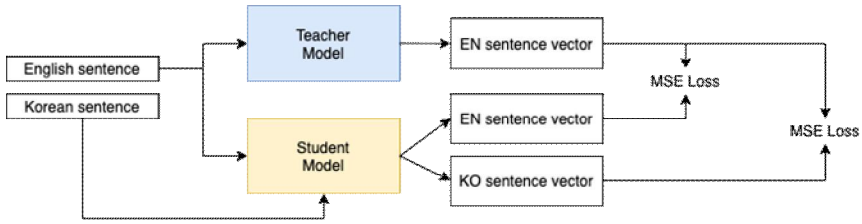


Figure 3. SBERT teacher–student knowledge distillation architecture adapted from Reimers and Gurevych, 2020.

The original authors of the paper call this transfer learning method "knowledge distillation" because the student model $\hat{A}$ distills the knowledge of the teacher model A. Like previously stated, this method will align the vector embeddings of the translated sentences into the same point. The vector space properties of the teacher model A are transferred to the student model. Although relatively simple in theory, the authors of the paper prove that this knowledge distillation method is effective across multiple language variations using several evaluation tasks, such as bitext retrieval and multilingual STS.

3.2 Advantages to SBERT Knowledge Distillation

There are several advantages to performing a multilingual knowledge distillation for SBERT instead of choosing to fine–tune a SBERT model using the siamese or triplet network training method. This section will explore some of the advantages of using this novel method to train a model.

As stated previously, it is difficult to obtain quality labeled sentence pair dataset for low–resource languages. Unlike data that is used for pre–training general language models such as BERT, the data used for fine–tuning SBERT must be carefully curated to obtain semantically meaningful sentence embeddings. By utilizing

the teacher model to fine-tune the student-model, the amount of necessary training data drastically decreases. Through this teacher-student knowledge distillation, you can fine-tune a model to high standards using a relatively small amount of data, which is advantageous for low-resource languages.

Secondly, this multilingual knowledge distillation method trains the student model to be one that can be used for cross-lingual tasks. The student model will acquire knowledge not only about the target language, but also the source language. Although this particular experiment only uses English and Korean sentence pairs and thus produces a bilingual model, theoretically, more than one target language can be used during fine-tuning, which would result in a truly multilingual model.

Lastly, unlike other previous multilingual models, this knowledge distillation approach is not trained with a specific task in mind. Previous research such as LASER (Artetxe and Schwenk, 2018) trains a LSTM model for translation. This knowledge distillation method using SBERT instead manipulates the properties of the vector space of the model's sentence embeddings. Through such manipulation, the resulting student model can identify not only exact pairs of translations, but also similar pairs that are not exact translations.

4. English and Korean Multilingual SBERT

This section will outline the training and evaluation details for the English and Korean Multilingual SBERT. Training procedures follow the teacher-student knowledge distillation architecture detailed in Section 3.

4.1 Setup

For this experiment, the teacher model is the English monolingual **bert-base-nli-stsb-mean-tokens**[⑤] SBERT and the student model is the base **XLM-RoBERTa** (XLM-R) (Conneau et al., 2020). The teacher model was chosen for its proven ability to derive accurate sentence embeddings in English across a variety of tasks including STSb and clustering. XLM-R is suitable as the student model in this experiment because it is a multilingual language model. It has been trained using 88 different languages, including English and Korean, which may be beneficial in learning the parallel data, since it already has acquired both languages. Furthermore, The XLM-R model uses the SentencePiece[⑥] tokenizer, which is not language-specific and applicable to non-Roman languages such as Korean.

4.2 Data

This section will discuss the datasets used in this experiment. A combination of existing datasets and augmented datasets were used for training.

⑤ https://huggingface.co/sentence-transformers/bert-base-nli-stsb-mean-tokens
⑥ https://github.com/google/sentencepiece

4.2.1 Existing Datasets

- **Bible.** The Bible is one of the most ubiquitous pieces of text in the world, having been translated into more than 700 languages. It is a valuable resource for multilingual machine learning training.
- **Conversations**[⑦]. A Kaggle dataset of English and Korean conversations extracted from an online dictionary. This is a relatively small dataset of around 4,500 sentences.
- **JHE**[⑧]**.** A Junior High English (JHE) evaluation dataset for Korean. This dataset has been built by pulling English and Korean reference sentences from middle school English textbooks used by Korean students and then aligning them using a self-training machine translation technique. Subjects vary from news articles, short stories, and letters, to even advertisements. This dataset is divided into train, dev, and test datasets.
- **En-Ko News Corpus**[⑨]**.** The authors of the JHE dataset also collected English and Korean news articles largely from Yahoo! Korea and CNN during 2010 and 2011. These articles were then aligned. This dataset contains a total of 96, 982 aligned sentences.
- **KAIST Parallel Corpus**[⑩]**.** The Semantic Web Research Center (SWRC) at KAIST, a national research university in Korea, released a dataset of 60,000 English and Korean sentence pairs in 2005. The dataset is comprised of largely short, colloquial sentences.
- **OPUS**[⑪]**.** The Open Parallel CorpUS (OPUS) is a growing collection of parallel datasets available in over 50

⑦ https://www.kaggle.com/rareloto/naver-dictionary-conversation-of-the-day
⑧ https://sites.google.com/site/koreanparalleldata/
⑨ https://sites.google.com/site/koreanparalleldata/
⑩ http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus
⑪ https://opus.nlpl.eu/

languages. It is an ongoing project that collects, aligns, and distributes publicly available resources to researchers. Through OPUS, three English and Korean parallel datasets were downloaded, including the Global Voices, TedTalks, and Tatoeba.

- **JW300.** Released in 2019, the JW300 Corpus (Agić, Vulić, 2019) is a parallel corpus of more than 300 languages with around 100,000 parallel sentences per language pair on average. The data is crawled from the website jw.org, which mostly publishes texts from magazines such as *Awake!* and *Watchtower*.
- **WikiMatrix.** The WikiMatrix (Schwenk et al., 2019) is a multilingual dataset crawled from Wikipedia articles in 85 languages. A distance-based mining approach and threshold optimization is used to align the multilingual datasets together. The recommended threshold for all language pairs is 1.04.

The above datasets result to a total of 927 MB of 5,264,321 English and Korean parallel sentences. The bulk of the training data stems from OPUS and the WikiMatrix. While these corpora are abundant in quantity, the quality of the sentence pairs cannot be guaranteed. When examining the data manually, there were frequent cases of sentence pairs that were not translations of each other.

4.2.2 Data Pre-Processing

To alleviate the issues of noise and incorrect alignment, the datasets listed above was pre-processed. Characters that were not essential punctuation, Korean, or English characters were removed. Content inside parentheses and brackets were removed because often in Korean texts, such content is comprised of Hanja characters that are not in the test data and thus disruptive to the evaluation performance of the model.

The most heavily processed dataset was the WikiMatrix. The authors of the dataset recommended a threshold of 1.04 for all languages. However, this threshold tended to include superficial matches (when the English sentence is included in both the English and Korean datasets, creating an exact match) but exclude actual translations. To alleviate this issue, the threshold was lowered to 0.999 and filtered the data aggressively. The occurrence of the same sentence in the same language in both the English and Korean datasets was filtered using a rule-based method. For examples of raw WikiMatrix data, refer to Table 1. Although this method resulted in a loss of ~100MB of data, the resulting dataset was much higher in quality.

#	Raw WikiMatrix Data
1	1.1766367946322376 "Why Spain were anything but boring". 2013 년 6 월 30 일에 확인함. “Why Spain were anything but boring”.
2	1.174842893439351 "How healthy is the air you breathe?". 2018 년 5 월 7 일에 확인함. “How healthy is the air you breathe?”
3	1.0200002206655117 A very pompous and proud person. 최고라는 자부심으로 충만한 사람이요.
4	1.0200016049267824 Early in its duration, Pongsona first affected Pohnpei as a tropical storm. 초기에 태풍 봉선화는 폰페이 섬에 열대 폭풍의 단계로 영향을 끼쳤다.

Table 1. Examples of raw data from the WikiMatrix. Data 1 & 2 are examples of the superifical matches that have high scores. Data 3 & 4 are examples of correct translations that have scores below the recommended threshold.

4.2.3 Data Augmentation

In addition to the aligned datasets listed in section 4.2.1, new training data was collected and aligned for data augmentation. The data was collected from English and Korean online dictionaries as well as English and Korean news sites.

The total data collected amounts to 135 MB and 592,858 sentences. The sentences scraped from English and Korean dictionaries have a 1:1 correspondence, meaning that each English sentence is paired with its Korean counterpart and no alignment work was necessary. In contrast, the data collected from the bilingual news sites did not have the same number of sentences for the English and Korean articles.

To align the data from the news sites, this study experimented with the following alignment methods. The data is divided by article, which is comprised of 20 to 30 English and Korean sentences each.

- LASER[12]. A library to calculate multilingual sentence embeddings, LASER provides an encoder that is trained using 93 languages, including English and Korean. LASER then uses FAISS[13] to efficiently calculate distances between sentence pairs, from which a similarity score for the pair is derived. This is the method that the authors of the WikiMatrix dataset aligned their data. The authors recommend setting a threshold for the score, to ensure the pairs were accurate translations. Because each article had a small number of sentences that needed to be aligned, the threshold was set to 0.65 for this alignment task.
- En-Ko SBERT. Using a multilingual model that was trained using the data outlined in section 4.2.1, the data collected from news sites was encoded into a vector space.

[12] https://github.com/facebookresearch/LASER
[13] https://github.com/facebookresearch/faiss

Then, the cosine distance between for all the sentence pairs in the article was calculated. Because each article has a relatively small number of sentences, calculating the distances was efficient and quick using SBERT. To avoid including inaccurate pairs in the final dataset, a filtering method was developed by calculating the distances between the first and second closest sentences. If the difference in the two distances exceeded 0.9 points, the first sentence was deemed the best translation. If it did not, the sentence was not included in the final dataset.

Despite the noise and error in alignment in the training data of SBERT, the performance of SBERT far outweighed that of LASER. For a comparison of the two alignment methods, refer to Table 2. As illustrated, the LASER alignment resulted in a steep loss of the number of pairs and a greater number of inaccurate pairs when examined manually. The data collected and aligned from these bilingual news sites was also pre-processed to remove extraneous symbols and noise and then used to augment the data in section 4.2.1.

Name	Alignment Method	Threshold	# of pairs above threshold	Examples
LASER	Euclidean distance using FAISS indexing	> 0.65	3,766	en: More than 20,000 people joined the event, and reviewers wrote that they had eaten their first theater popcorn in a long time. *ko: 일회용품 자제 차원에서 뚜껑 있는 다회용 식품용기를 가져오면 6000 원에 가득 채워주는 행사였다.
SBERT	Cosine distance using scipy	d1 – d2 > 0.9	13,171	en: More than 20,000 people joined the event, and reviewers wrote that they had eaten their first theater popcorn in a long time. ko: 전국 참여자가 2 만명에 이르렀고 "오랜만에 극장 팝콘 맛 봤다"는 후기가 줄이었다.

Table 2. A comparison of the LASER and SBERT alignment methods. As illustrated, the LASER alignment method produced a smaller amount of alignments, many of which were incorrect. The SBERT alignment method outperforms that of LASER in terms of both accuracy and efficiency.

4.3 Training

As stated in Section 4.1, teacher model is the English monolingual **bert-base-nli-stsb-mean-tokens**[14] SBERT and the student model is the base **XLM-R** (Conneau et al., 2020). Two variations of the student model were trained for this experiment. The first model was trained using the pre-processed training data that was collected online (as stated in 4.2.1) and the augmented data personally collected and aligned by me. The second model was trained solely using the unprocessed data that was collected online.

4.3.1 Training Details

Both models were trained under the same conditions. The parameters are as follows: learning rate 2e-5; eps 1e-6. The model trained with the clean and augmented data was trained for 10 epochs and will be referred to as **EN-KO-SBERT-CLEAN**. The model trained with the raw data only was trained for 5 epochs and will be referred to as **EN-KO-SBERT-RAW**.

During training, the models were evaluated on mean squared error (MSE) and two Korean semantic textual similarity (STS) tasks. The first of the STS tasks was KorSTS (Ham et al., 2020) and the second was the KLUE STS (Park et al., 2021). These STS evaluations were done in an unsupervised manner during model training. Evaluation was done using the test set of the KorSTS dataset and the dev set of the KLUE STS dataset. The results of the training are shown in Table 3 below.

[14] https://huggingface.co/sentence-transformers/bert-base-nli-stsb-mean-tokens

Model Name	MSE	KorSTS test	KLUE STS dev
EN-KO-SBERT-CLEAN	10.410109	0.8042	0.7288
EN-KO-SBERT-RAW	10.898891	0.8008	0.7062

Table 3. The MSE, KorSTS, KLUE STS results of the two models after unsupervised training. The model was evaluated on the KorSTS test set and the KLUE STS dev set.

The EN-KO-SBERT-CLEAN model outperformed the EN-KO-SBERT-RAW model in all three evaluation categories. However, the performance differences in the unsupervised KorSTS task was negligible. The largest performance difference in the unsupervised tasks between the two models came in the KLUE STS dev task, with the clean model outperforming the raw model by over 2 points.

4.4 Evaluation

The models were then evaluated with the same tasks, but in a supervised method, meaning the models were specifically trained for certain downstream tasks using training data before being evaluated on test datasets. The evaluation was conducted on the previously mentioned STS tasks and a novel translation matching task that was created to test the cross-lingual abilities of the model. Both the translation and STS tasks were chosen to evaluate SBERT because they are sentence pair tasks that require the model to gauge the semantic similarity of the sentence pairs.

Both the EN-KO-SBERT-CLEAN and EN-KO-SBERT-RAW models were evaluated on the same tasks. In addition to the models trained by me, two other models were evaluated for

comparison. The first was the XLM-R SBERT, which was used as a vanilla model. The second was the KR-SBERT mentioned in section 2.3. This model was evaluated to compare the models trained in this paper to a model that was purely pre-trained on Korean datasets.

4.4.1 Supervised KorSTS

The KorSTS dataset is constructed from the original English STS-B dataset (Agirre et al., 2012, 2013, 2014, 2015, 2016). The STS-B train data has been machine translated into Korean for train dataset while the dev and test datasets are machine translated then post-edited by humans into Korean. The sentence pairs are labeled with a gold label, a continuous score between 0 and 5, with 0 being the most semantically dissimilar and 5 being the most semantically similar. The results of the four models trained and evaluated on the KorSTS datasets are as below. The results are derived by computing the Spearman's rank correlation between the cosine-similarity of the sentence embeddings and the gold labels.

Model	KorSTS dev	KorSTS test
EN-KO-SBERT-CLEAN	**0.8501**	**0.8318**
EN-KO-SBERT-RAW	0.8496	0.8258
XLM-R	0.7834	0.7326

Table 4. The results of the four models for comparison on the KorSTS task. The EN-KO-SBERT-CLEAN model outperformed all other models, although it only outperformed the raw model by a small margin. The EN-KO-SBERT-CLEAN model outperformed XLM-R by almost 10 points.

As seen in Table 4, the EN-KO-SBERT-CLEAN model outperformed all other models, although it only outperformed EN-KO-SBERT-RAW by a small, insignificant margin. The clean model outperformed XLM-R by almost 10 points. These results are

indicative of the power of transfer learning through knowledge distillation for S-BERT.

However, there are limitations to relying on the KorSTS dataset results to evaluate the model. First, the results may be overconfident because the EN-KO-SBERT-CLEAN has learned from an English model that performed highly on the STSb task, from which the KorSTS dataset was constructed. Moreover, as discussed previously, the KorSTS relied heavily on machine translation to construct the training, dev, and test datasets. As a result, some sentence pairs are made of incomplete sentences or are nonsensical, superficial translations of the English dataset. For example, refer to the table below. The first example in Table 4 is one where sentence 1 is an incomplete sentence. Sentence 2 of the second example in Table 4 is an example of a nonsensical translation. For these two reasons, the results in Table 4 should be taken with a grain of salt and not as the gold standard for evaluation.

Label	Sentence 1	Sentence 2
4.40	파도를 타는 서퍼	서퍼가 파도를 타고 있다.
2.2	그것은 단지 세포의 무리일 뿐이다.	난 세포 다발이야.

Table 5. Examples taken from the KorSTS train dataset. These sentence pairs are examples of the limitations of the machine translation approach to constructing a dataset for an NLP task.

4.4.2 Supervised KLUE STS

To further evaluate the models on a high-quality Korean dataset that has not been seen by the teacher model in any shape or form, this study also evaluated the models on the KLUE STS dataset. The KLUE STS dataset is one that has been constructed solely from Korean resources, including colloquial reviews, formal news articles, and smart home utterances. However, because the authors of KLUE have not made the test set publicly available, the

models are trained on the KLUE STS train dataset but evaluated on the KLUE STS dev set and the KorSTS test dataset. The results are as shown in the table below.

Model	KLUE STS dev	KorSTS test
EN-KO-SBERT-CLEAN	**0.8865**	**0.8109**
EN-KO-SBERT-RAW	0.8843	0.7904
XLM-R	0.8682	0.7133

Table 6. The results from training on the KLUE STS dataset. Because the KLUE STS test dataset is not revealed to the public, the results here are of the KLUE STS dev set, and the model trained on the KLUE STS dataset evaluated on the KorSTS test dataset.

Like the KorSTS results, the EN-KO-SBERT-CLEAN model outperformed all other models. Surprisingly, the vanilla XLM-R model performed similarly well on the KLUE STS dev dataset.

4.4.3 En-Ko Translation Matching

The STS tasks tested the model in Korean tasks, but there were no existing NLP tasks that could evaluate the Korean and English cross-linguistic ability of a model. To evaluate the model across English and Korean, a novel task was created.

Using the alignment method outlined in section 4.2.3, new data from English and Korean news websites was collected and aligned into English and Korean sentence pairs. A total of 9,000 sentence pairs were collected. Each sentence pair was labeled as exact translation (1) or not a translation (0). Even the sentence pairs that were not translations of one another were taken from the same news article to mimic a slight similarity. To ensure the task was as accurate and informative as possible, the sentence pairs

were manually checked by native Korean speakers. Details of the dataset can be found in Table 7 below.

	Label 1	Label 0	Total
Number of sentences	4,399	4,601	9,000
Percentage	48%	52%	100%

Table 7. Details about the English and Korean cross-lingual translation matching task. There are 9,000 sentences in total.

The training for this task was conducted using the Cross Encoder class of the Sentence Transformers. The Cross Encoder is a wrapper around the original Transformers model where a sentence pair is passed to the model simultaneously. A classifier is trained on top of the model to predict the label. To visualize grasp how the Cross Encoder functions, refer to the figure below.

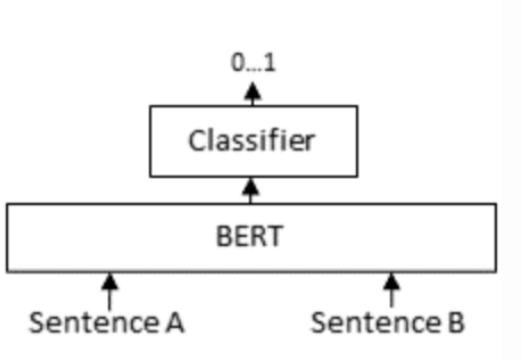


Figure 4. A visual representation of the Sentence Transformers' Cross Encoder class from Reimers and Gurevych, 2019.

The dataset of 9,000 sentence pairs was split into training and test datasets by a ratio of 9 to 1. Like the other tasks, several different models were trained under the same conditions to compare their performances. The training parameters were set as the

following: learning rate 1e-5; eps 1e-6. Each model was trained for 3 epochs. A softmax evaluator was used on the test results to evaluate model performance on classification accuracy. The results are shown in the table below.

Model Name	Test set evaluation (acc.)
EN-KO-SBERT-CLEAN	93.22
EN-KO-SBERT-RAW	93.00
XLM-R	**94.22**
KR-SBERT	72.44

Table 8. The results of the English and Korean translation alignment task. This is a cross-lingual task that requires knowledge of both languages.

As expected, the monolingual KR-SBERT model performed the worst, trailing behind the multilingual model by over 20 points. Because KR-SBERT was only trained on Korean data, it is expected that it will not perform well on this cross-lingual task, which requires knowledge of both English and Korean. Surprisingly, the base XLM-R model outperformed the EN-KO-SBERT model by a small margin, about one point. The reason for this performance is unclear, but it could be explained by the fact that the XLM-R model has retained more information about other languages that may have aided it during this simple simplification task, which is less reliant on semantic or syntactic information on sentences compared to the STS tasks. A further probe or examination of this difference in performances could also be a subject for further study.

4.4.4 Document Classification

This last evaluation task was conducted to further compare the Korean capabilities of the models trained in this research paper with the KR-SBERT model, which was trained only using Korean monolingual data.

The KorSTS and KLUE STS tasks were not sufficient for comparison with the KR-SBERT model because the KR-SBERT model was trained using the same STS datasets. Therefore, a document classification task was conducted. This evaluation task assesses the model's ability to classify Korean news articles into 9 labels. All the models were trained for 10 epochs. The results of the document classification evaluation task can be found in Table 9 below.

Model name	Evaluation (acc.)
EN-KO-SBERT-CLEAN	84.22
EN-KO-SBERT-RAW	83.78
KR-SBERT	85.89

Table 9. The results of the document classification task comparing the models trained in this paper and the monolingual KR-SBERT on a Korean monolingual task.

The clean model came within almost one point of the monolingual model. This test confirms that the model trained in this paper comes close to matching the performance of the monolingual Korean model, while also having multilingual capabilities.

4.5 Discussion

Importantly, the KorSTS task, KLUE STS task, and the cross-lingual translation classification task discussed in this section prove that the EN-KO-SBERT model is highly proficient not only on Korean tasks that require sentence-level embeddings but also on multilingual tasks, making it also widely applicable in various practical situations, such as machine translation or sentence alignment.

5. Model Analysis

In this section, the EN-KO-SBERT-CLEAN model will be analyzed by examining its attention patterns. Using a matplotlib, simple heatmaps were drawn using the attention patterns of the trained model. Although the model outputs a fixed-size embedding using SBERT, the base Transformer model was used for model analysis.

A few sentences from the training dataset were selected to be used as examples for deriving attention heatmaps. These examples were chosen for their simplicity and terseness. The short list of examples are shown in the table below.

No.	English	Korean
1	The teacher told Bobby to stop fooling around in class.	선생님은 바비에게 교실에서 그만 까불라고 하셨다.
2	The dog ran around and around trying to catch its tail.	개는 자기 꼬리를 잡으려고 빙빙 돌며 뛰었다.

Table 10. The examples chosen to analyze the attention of the trained EN-KO-SBERT model.

After analyzing all heatmaps of all the twelve layers of the Transformer model, it was found that the first layer included a variety of dispersion patterns of attention and was representative of all the other layers. Thus, only the heatmap of the first layers are shown in this section, but the attention heatmaps of all the twelve layers can be found in Appendix C. The first layer heatmaps of both the examples can be found in the figures below.

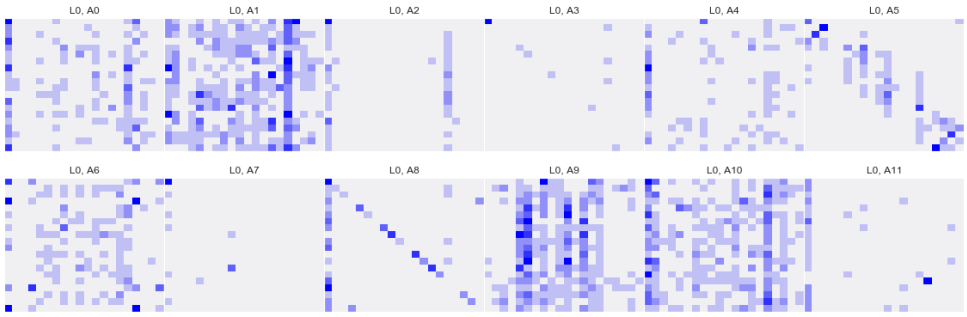


Figure 5. The heatmap of Example 1 - English sentence.

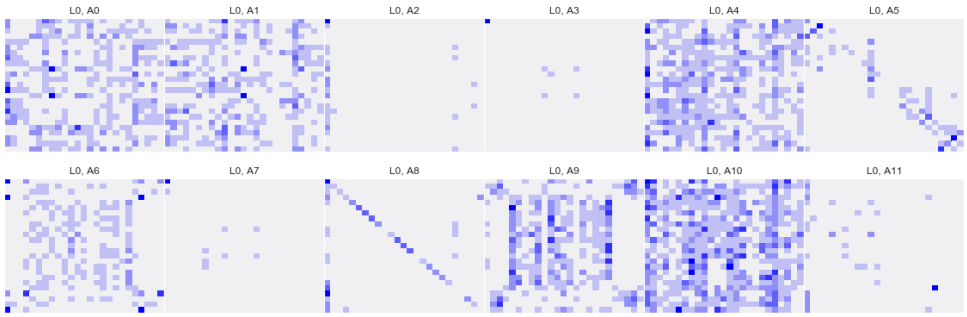


Figure 6. The heatmap of Example 1- Korean sentence

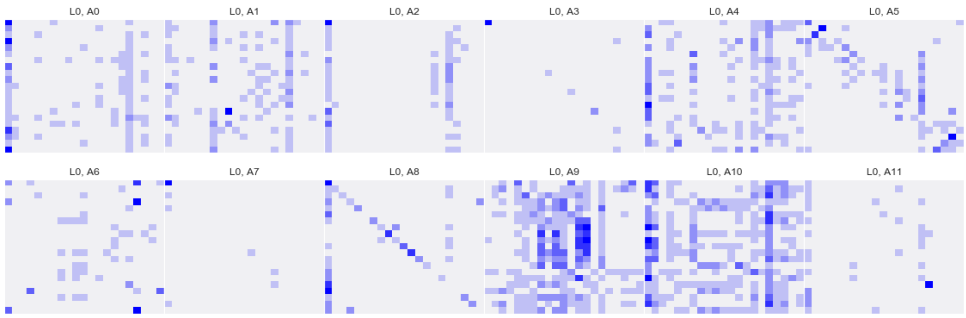


Figure 7. The heatmap of Example 2 - English sentence

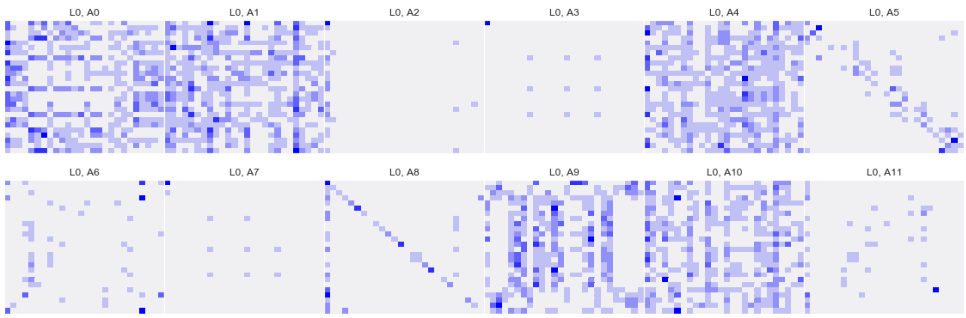


Figure 8. The heatmap of Example 2 - Korean sentence

Because XLM-R uses SentencePiece as its tokenizer, the Korean sentences are tokenized as below:

'▁[', 'C', 'LS', ']', '개', '는', '▁자기', '▁', '꼬', '리를', '▁잡', '으려', '고', '▁', '빙', '빙', '▁돌', '며', '▁', '뛰', '었다', '.', '[', 'S', 'EP', ']'

As seen in the example, the spaces are sometimes tokenized as separate tokens from the words. On the other hand, the English sentences are tokenized as below:

'▁[', 'C', 'LS', ']', 'The', '▁dog', '▁ran', '▁around', '▁and', '▁around', '▁trying', '▁to', '▁catch', '▁its', '▁', 'tail', '.', '[', 'S', 'EP', ']'

This difference in tokenization results in different lengths of the embeddings, which explains why the English heatmaps have a shorter axis than the Korean heatmaps. Also, the English heatmaps generally tend to have a more concentrated attention, as there are more darker spots in these maps than those of the Korean ones.

Despite these slight differences, it is clear that each attention head in the layer shows a similar attention pattern, signaling that the student-teacher knowledge distillation method successful aligned the attentions and embeddings of the Korean and English translations during training.

6. Conclusion

In this paper, the existing knowledge base of NLP models was first examined, starting with the highly influential BERT model. The advantages and disadvantages of BERT were discussed to show how it evolved into other models such as SBERT, which attempts to diminish the shortcomings of the BERT model and maximize its advantages. Next, there were discussions of the student-teacher knowledge distillation method proposed by the authors of SBERT, which allows models to be efficient by using transfer learning. This training method allows models to achieve high performances even in languages that are low in resources.

Then, in the main portion of this research, an experiment training an English and Korean multilingual SBERT model using this novel transfer learning method of knowledge distillation was conducted. In the process, parallel English and Korean data sentences was crawled and aligned, also using SBERT. To test this model's capabilities, two existing Korean STS tasks were used to evaluate this model, but a new classification task was also created to test the model's multilingual performance.

Further research on this topic could involve deeper comparisons with XLM-R, testing how the base XLM-R model was able to outperform the model trained in this paper. A detailed probe or experiments with other multilingual tasks would provide a more detailed look into the difference between the two models. Also, it would be interesting to compare the EN-KO-SBERT model with other models that were only trained in Korean. Because this research was limited in the scope of its application of models, further experiments involved with more practical applications of the model, such as document clustering and paraphrase mining, would further test the model's limits and capabilities.

Another possibility for further research could examine the quality of training parallel datasets and its effect on the overall training of the student model. In this experiment, although two

models were trained based on different qualities of data and their evaluations were compared, the amount of data augmented between the CLEAN and RAW models were not significant enough to yield significant differences in the evaluation process. If more quality parallel data could be collected and aligned in the gigabyte unit, there could be a larger difference in performance.

Despite leaving such opportunities for further research, this research is meaningful in that it is the first English and Korean multilingual SBERT model that has been effectively tested for its sentence-level embedding capabilities and its multilingual capabilities at the time of writing. Furthermore, through the process of training and testing, several aspects of SBERT were examined in more detailed and applied to Korean, such as its transfer learning training methods and cross-encoder classification evaluation process. Lastly, by collecting data for training and for testing, novel methods of parallel sentence alignment were examined and tested.

Bibliography

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 GonzalezAgirre, 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.

Željko Agić and Ivan Vulić. 2019. JW300: A widecoverage parallel corpus for lowresource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2018. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. arXiv preprint arXiv:1812.10464, abs/1812.10464.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In EMNLP. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iigo LopezGazpio, 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.

Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, 2018.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, GuillaumeWenzek, Francisco Guzman, Edouard Grave, Myle Ott, ´ Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Unsupervised Cross-Lingual Representation Learning at Scale. arXiv:1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In North American Association for Computational Linguistics (NAACL).

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. arxiv preprint arXiv:1907.11692.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In ICLR Workshop Papers.

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

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERTnetworks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzman. 2019. ′ Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. CoRR, abs/1907.05791.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017.

Adina Williams, Nikita Nangia, and Samuel Bowman. 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, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Appendix A. Collected Data Examples

Dataset	English	Korean
Bible	The angel swung his sickle on the earth, gathered its grapes and threw them into the great winepress of God's wrath.	그래서 그 천사가 낫을 땅에 휘둘러, 땅의 포도를 거두어서, 하나님의 진노의 큰 포도주를 만드는 술틀에다가 던졌습니다.
Conversations	No, I couldn't get his signature because there were too many people. I will definitey get his signature when I go to his next concert.	아니요, 사람이 너무 많아서 못 받았어요. 다음 콘서트에 가면 꼭 사인을 받을 거예요.
JHE	Searching through row upon row of Christmas trees, my husband and I picked one we liked.	여러 줄로 세워져 있는 크리스마스 트리 사이를 뒤지며 찾다가 남편과 나는 마음에 드는 나무 한 그루를 골랐다.
En-Ko News Corpus	It was unclear, however, whether the report referred to a plutonium- or uranium-based weapon.	그러나 그 보도가 언급한 것이 플루토늄에 의한 무기인지 우라늄에 의한 무기인지는 명확하지 않았다.
KAIST Parallel Corpus	The singer bowed out of the competition.	그 가수는 대회를 중도에 그만두었다.
OPUS	Social Media Week first took place in February 2009 in New York with over 2,500 people attending across forty	2009 년 2 월 뉴욕에서 처음 개최된 소셜 미디어 주간에서는 40 여가지 행사를 통해 2500 명 이상의 인원이

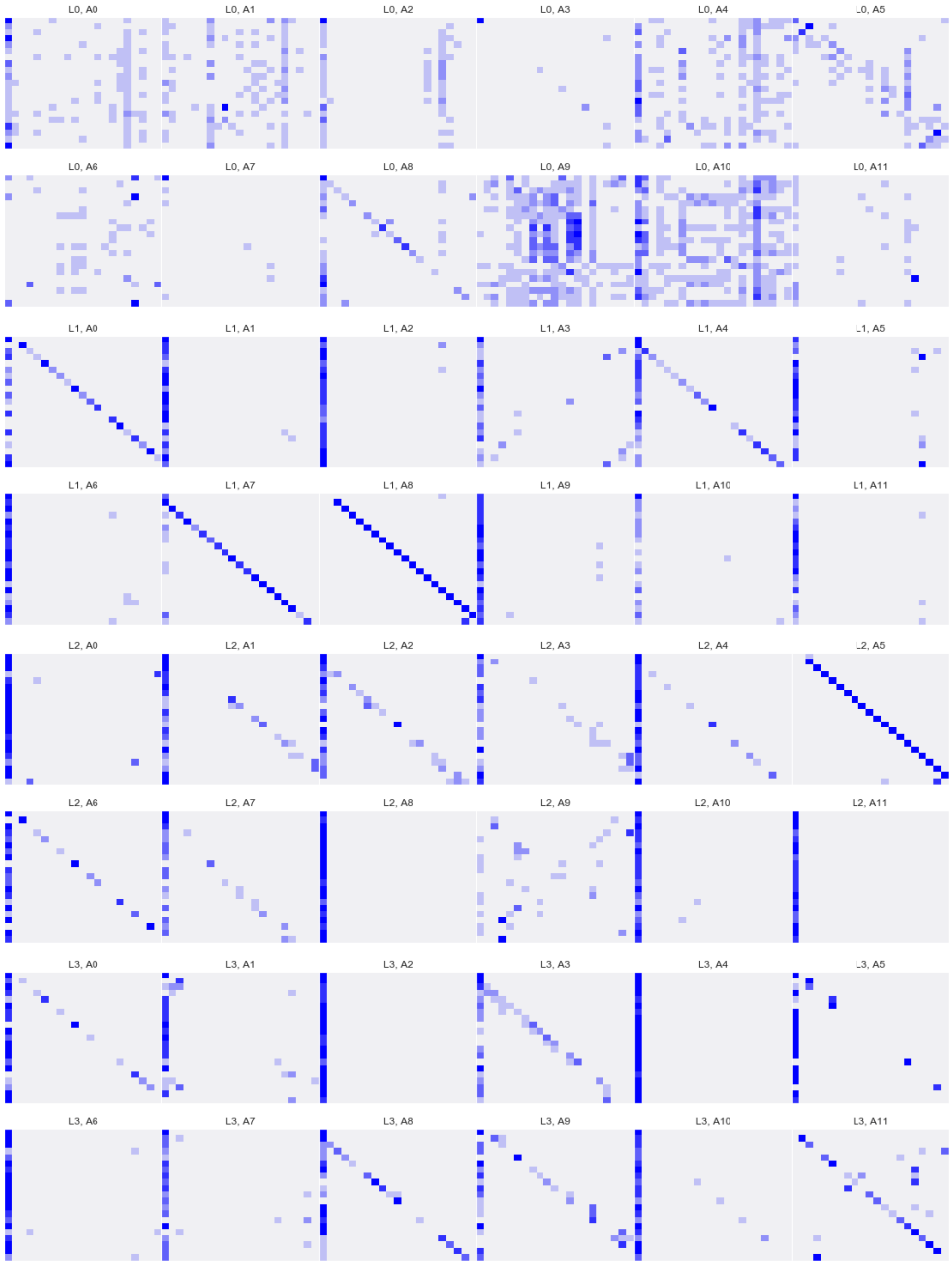
	different events .	참여했다 .
JW300	Examine the religious ‘ trees ’ with which you associate. Are they producing the kind of fruit God requires?	당신 이연합 하고있는종교 “나무 ” 를검사 해보라. 그 “나무 ” 는하나님 께서요구 하시는열매 를맺고있는가?
WikiMatrix	Depending on the cultural background, in some families this celebration is more important than Christmas.	문화적 배경에 따라 이탈리아의 일부 가족에서는 이 축하가 크리스 마스보다 중요하게 여겨진다.

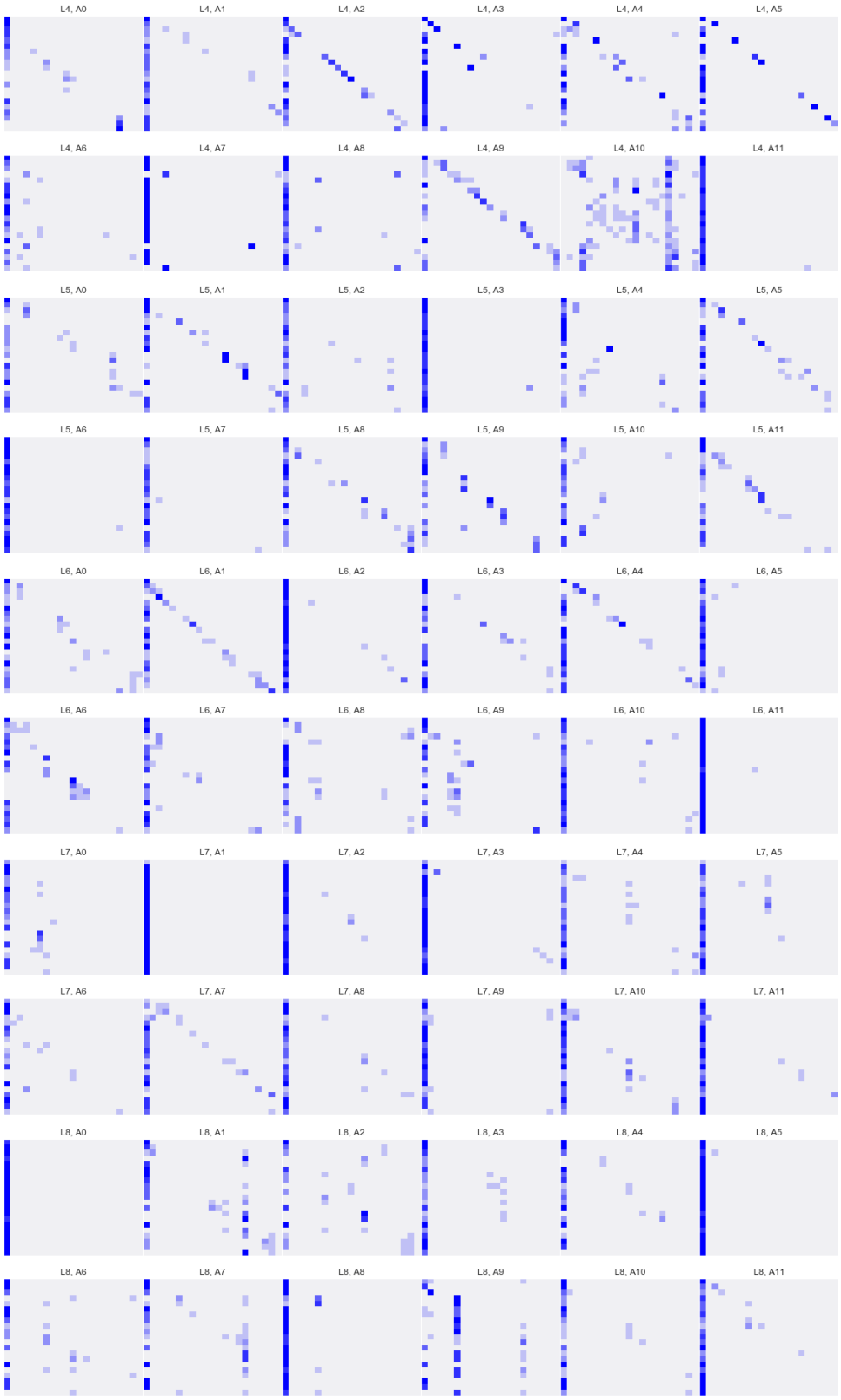
Appendix B. Crawled Data Examples

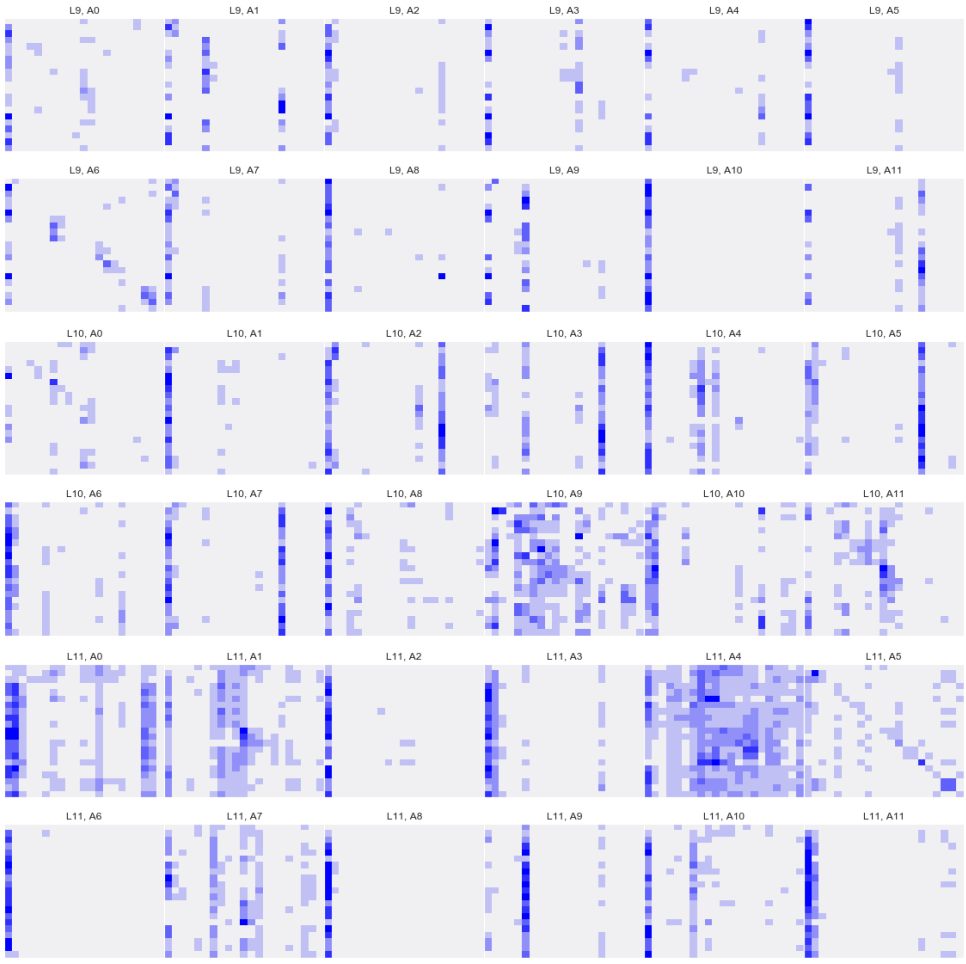
English	Korean
Some academics doubt dreamers can respond to specific instructions while asleep.	꿈을 꾸는 사람이 잠자는 동안 특정 지시에 반응할 수 있다는 주장에 대해 의문을 제기하는 학자들도 있습니다.
The White House has invited 1,000 guests who were at the frontline in the fight against coronavirus to an Independence Day event.	백악관은 코로나 최전선에서 사투를 벌인 1000 여 명을 독립기념일 행사에 초대한다.
A 30-something female defector who married to a Chinese man has a 12-year-old son.	30 대 탈북 여성은 중국인 남성과 결혼해 12 세 아들을 뒀고, 중국에서 상당히 많은 돈을 벌었던 것으로 알려졌다.
A feeling of dread settled over Robert when he heard noises in the empty house.	빈집에서 나는 소음을 들었을 때 로버트는 공포를 느꼈다.
The Supreme Court of South Korea has ruled that emergency measures issued by former president Park Chung-hee have been in violation of the constitution.	대법 민주주의 본질 침해 전원일치 판결유신정권 근거없이 발동 합헌 취지 판결 모두 폐기 박정희 전 대통령이 발령한 긴급조치는 위헌이라는 대법원 판결이 나왔다.
Several members of the Vietnam Lawyers Association came to Korea and announced a joint statement on the issue with the Seoul Bar Association (SBA), Wednesday.	베트남법률가협회 회원 여러 명이 한국에 와서 서울지방변호사회(서울변회)와 베트남전 민간인 학살 문제에 대한 공동 성명을 수요일 발표했다.

Appendix C. Heatmap of Layers 0–11 of Example 2

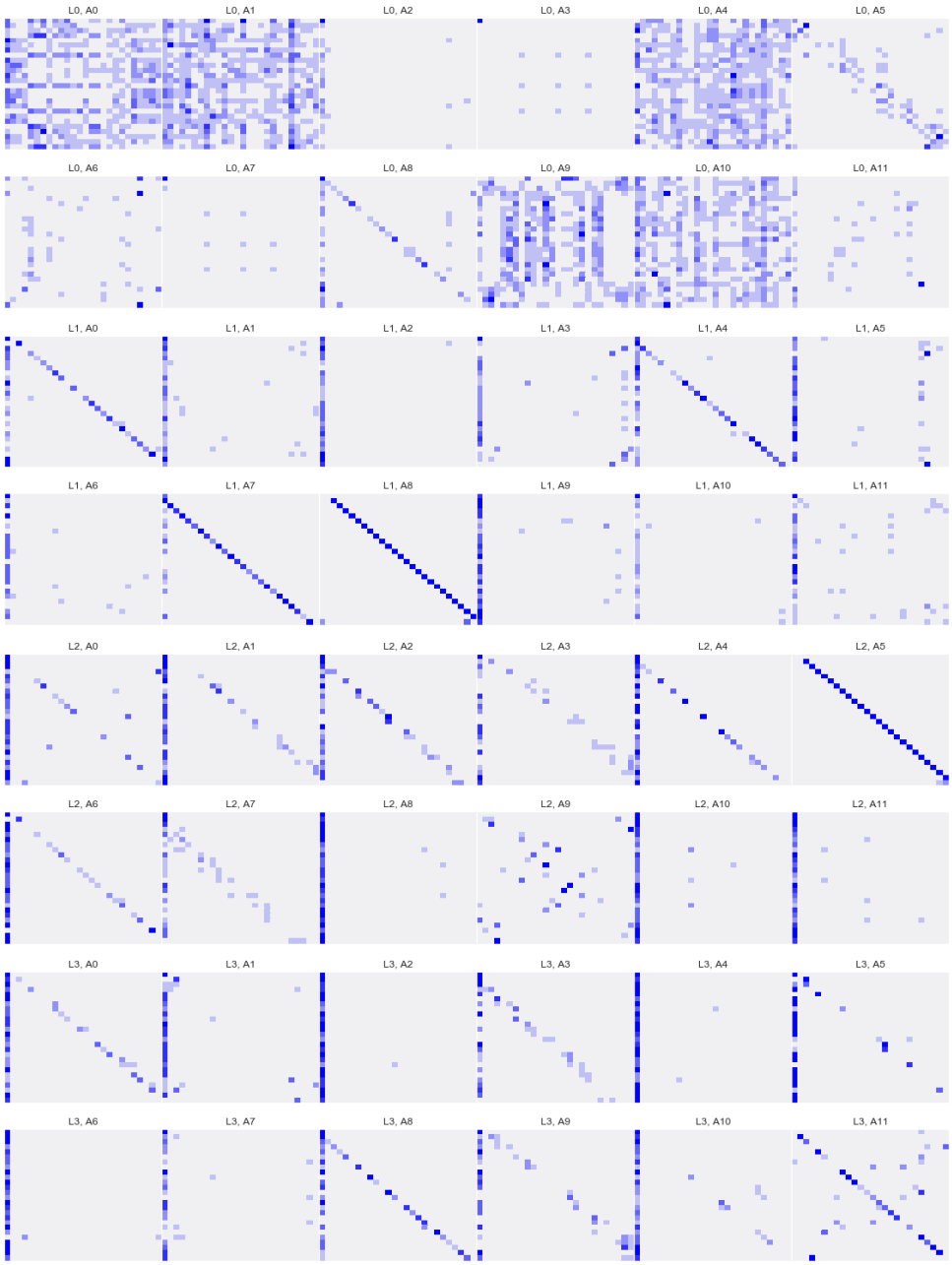
English Sentence

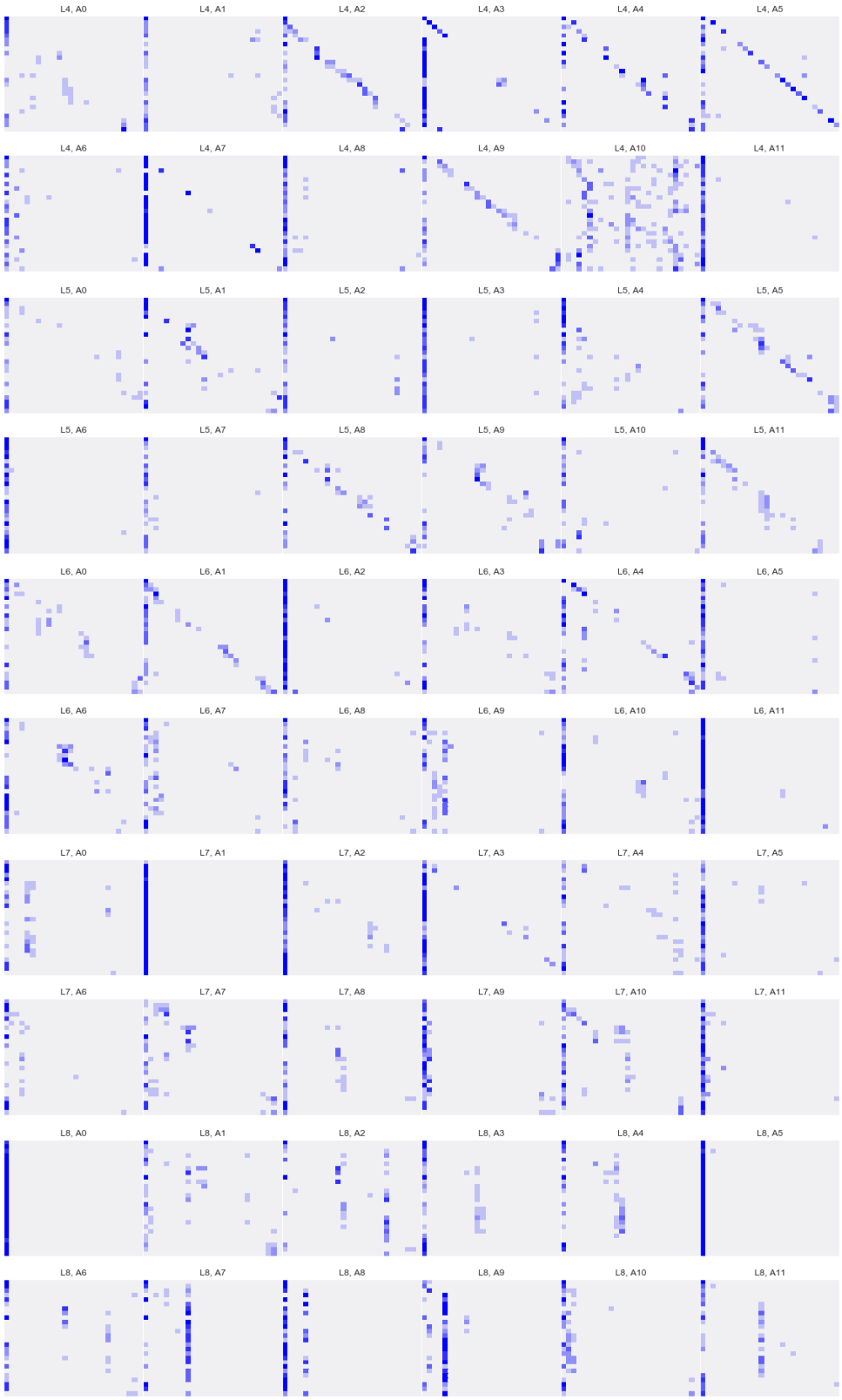


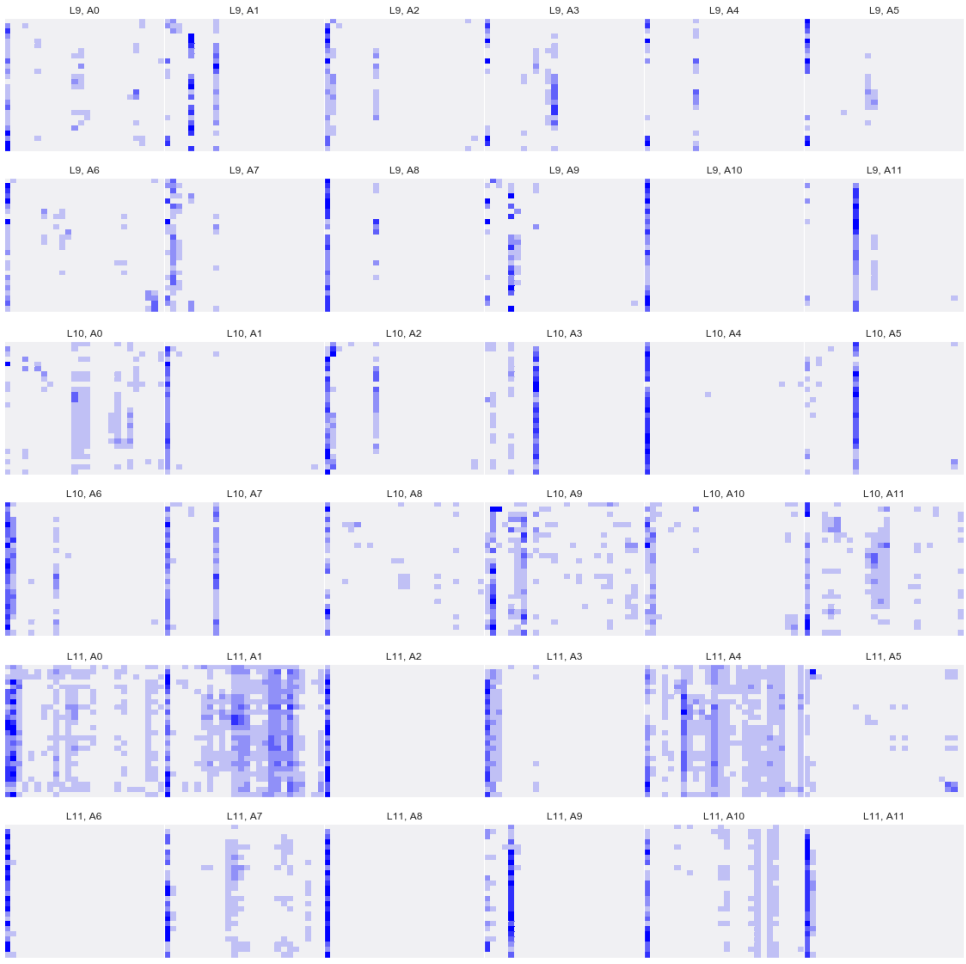

L4, A0
L4, A1
L4, A2
L4, A3
L4, A4
L4, A5
L4, A6
L4, A7
L4, A8
L4, A9
L4, A10
L4, A11
L5, A0
L5, A1
L5, A2
L5, A3
L5, A4
L5, A5
L5, A6
L5, A7
L5, A8
L5, A9
L5, A10
L5, A11
L6, A0
L6, A1
L6, A2
L6, A3
L6, A4
L6, A5
L6, A6
L6, A7
L6, A8
L6, A9
L6, A10
L6, A11
L7, A0
L7, A1
L7, A2
L7, A3
L7, A4
L7, A5
L7, A6
L7, A7
L7, A8
L7, A9
L7, A10
L7, A11
L8, A0
L8, A1
L8, A2
L8, A3
L8, A4
L8, A5
L8, A6
L8, A7
L8, A8
L8, A9
L8, A10
L8, A11


L9, A0
L9, A1
L9, A2
L9, A3
L9, A4
L9, A5
L9, A6
L9, A7
L9, A8
L9, A9
L9, A10
L9, A11
L10, A0
L10, A1
L10, A2
L10, A3
L10, A4
L10, A5
L10, A6
L10, A7
L10, A8
L10, A9
L10, A10
L10, A11
L11, A0
L11, A1
L11, A2
L11, A3
L11, A4
L11, A5
L11, A6
L11, A7
L11, A8
L11, A9
L11, A10
L11, A11

Korean Sentence




L4, A0
L4, A1
L4, A2
L4, A3
L4, A4
L4, A5
L4, A6
L4, A7
L4, A8
L4, A9
L4, A10
L4, A11
L5, A0
L5, A1
L5, A2
L5, A3
L5, A4
L5, A5
L5, A6
L5, A7
L5, A8
L5, A9
L5, A10
L5, A11
L6, A0
L6, A1
L6, A2
L6, A3
L6, A4
L6, A5
L6, A6
L6, A7
L6, A8
L6, A9
L6, A10
L6, A11
L7, A0
L7, A1
L7, A2
L7, A3
L7, A4
L7, A5
L7, A6
L7, A7
L7, A8
L7, A9
L7, A10
L7, A11
L8, A0
L8, A1
L8, A2
L8, A3
L8, A4
L8, A5
L8, A6
L8, A7
L8, A8
L8, A9
L8, A10
L8, A11


L9, A0
L9, A1
L9, A2
L9, A3
L9, A4
L9, A5
L9, A6
L9, A7
L9, A8
L9, A9
L9, A10
L9, A11
L10, A0
L10, A1
L10, A2
L10, A3
L10, A4
L10, A5
L10, A6
L10, A7
L10, A8
L10, A9
L10, A10
L10, A11
L11, A0
L11, A1
L11, A2
L11, A3
L11, A4
L11, A5
L11, A6
L11, A7
L11, A8
L11, A9
L11, A10
L11, A11

국문 초록

본 연구는 student-teacher knowledge distillation 전이학습 비법을 사용하여 영어-한국어 Sentence-BERT 모델을 학습한다. BERT의 한계는 이전 발표된 많은 연구에서 잘 탐구되었다. BERT는 문장 단위의 임베딩을 도출하는 데 효과적이지 않으며 문서 분류 및 클러스터링과 같이 대량의 문장 임베딩이 필요한 실용적인 상황에서는 적용할 수 없는 것으로 입증되었다. Sentence-BERT는 이러한 문제를 완화하고 효율적이고 정확한 방식으로 문장 임베딩을 도출할 수 있는 모델을 만들기 위해 개발되었다.

본 연구에서는 한국어와 같은 저자원 언어 모델도 영어와 같은 고자원 언어로 훈련된 모델과 같은 성능을 보일 수 있는 Sentence-BERT의 전이 학습 방법을 살펴본다. 의 모델은 Source 언어와 Target 언어에서 번역된 문장 쌍을 사용하여 Mean Squared Error Loss를 통해서 번역된 문장을 Teacher 모델과 동일한 벡터 공간에 매핑한다. 이 실험에서는 영어 모델을 Teacher 모델로, Cross-lingual 모델을 Student 모델로 사용한다. 저자가 알기로는, 이 논문의 출판일까지 이 새로운 방법을 사용하여 학습된 한국어 Sentence-BERT 모델은 없다.

Sentence-BERT에 대해 이러한 knowledge distillation을 수행하려면 많은 수의 소스 언어 및 대상 언어 번역 문장의 쌍이 필요하다. 웹에서 사용 가능한 데이터 세트를 수집한 후, 데이터는 웹에서 크롤링된 데이터로 증강되었고, 이 데이터를 새로운 방법을 사용하여 정렬한 이후에 전처리를 했다. 이 연구는 문장 단위 테스트 및 영어-한국어 테스트를 통해서 훈련된 모델을 평가했으며 모델의 광범위한 적용 가능성과 다국어 능력을 입증하였다.