SAVING ENDANGERED LANGUAGES WITH A NOVEL THREE-WAY CYCLE CROSS-LINGUAL ZERO-SHOT SENTENCE ALIGNMENT

Eugene Hwang

Blue Core Labs, The Masters School, New York, USA

ABSTRACT

Sentence classification, including sentiment analysis, hate speech detection, tagging, and urgency detection is one of the most prospective and important subjects in the Natural Language processing field. With the advent of artificial neural networks, researchers usually take advantage of models favorable for processing natural languages including RNN, LSTM and BERT. However, these models require huge amount of language corpus data to attain satisfactory accuracy. Typically this is not a big deal for researchers who are using major languages including English and Chinese because there are a myriad of other researchers and data in the Internet. However, other languages like Korean have a problem of scarcity of corpus data, and there are even more unnoticed languages in the world. One could try transfer learning for those languages but using a model trained on English corpus without any modification can be sub-optimal for other languages. This paper presents the way to align cross-lingual sentence embedding in general embedding space using additional projection layer and bilignual parallel data, which means this layer can be reused for other sentence classification tasks without further fine-tuning. To validate power of the method, further experiment was done on one of endangered languages, Jeju language. To the best of my knowledge, it is the first attempt to apply zero-shot inference on not just minor, but endangered language so far.

KEYWORDS

Natural Language Processing, Large Language Models, Transfer Learning, Cross-lingual Zero-shot, Embedding Alignment, BERT, Endangered Languages, Low-resourced Languages

1. INTRODUCTION

Thanks to a deluge of information, automatic language classification systems becomes a more and more indispensable component in our daily life. Examples of language classification tasks are sentiment analysis [17, 26] question type classification [14] irony classification [28], politeness classification [5] and etc. Traditionally, models like MLP, RNN [25], LSTM [24] and BiLSTM were widely used for text classification tasks, due to their sequential nature that was suitable for data that could be broken down into a sequence, such as text. However, these models experienced major drawbacks such as 'gradient vanishing' . Furthermore, when processing large amounts of text data, these models were shown to be inefficient and computation-heavy due to their 'recurrent' nature, where the models have to go through an entire sequence of words to reach a certain part of the text.

Recently, transformers that combine self-attention and cross-attention mechanisms have proven to be much more effective than traditional models with a more sequential architecture. Transformers essentially showed that by increasing parameters and training data set size

David C. Wyld et al. (Eds): SIPR, NCWC, BIBC, SOFEA, DSML, ARIA, NLP, CSEN -2023
pp. 333-347, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131926

infinitely, we can also theoretically improve performance without limit. In summary, transformers started an era in which more resources equals ever-increasing performance, of course at the expense of computation power. Transformers paved the way for Large Language Models (LLMs). One such example is BERT [6], which utilizes the only encoder portion of the transformer, and was shown to be extremely effective for classification tasks that generally do not require a decoder. State-of-the-art (SOTA) LLMs like GPT-3 are also all transformer-based models trained with billions of parameters.

LLMs were also applied in multilingual situations, resulting in massively multilingual transformers (MMTs) including mBERT and more recently, XLM-Ra. These models are pretrained on a massive corpus of multilingual text to achieve SOTA performance on many multilingual tasks such as cross-lingual natural language inference (XNLI). Currently, due to a concerning gap in available text data for different languages, MMTs offer satisfactory performances to merely the top 20 languages including English, Spanish, Russian, Chinese, etc. In current MMT paradigm, the average number of languages supported (may or may not be with satisfactory performance) is around a 100. For example, Google Translate supports 133 languages, and ChatGPT, a recent SOTA chatbot MMT by OpenAI, supports 95 languages. This may seem like an impressive statistic, but considering there are over 7000 languages in the world, only ~2 percent of languages are supported by the current MMT paradigm. Population wise, this means more than 3 billion people are not able to access current SOTA multilingual services. Languages that are not supported by most models are classified as low-resources languages or LRLs. Most LRLs have sufficient native speakers in the real world but lack digital data. A vast majority of these languages are spoken in underdeveloped regions of the world without significant access to technology. Although they face a great deal of inaccessibility, most LRLs are not directly in danger of extinction and have a sufficient speaker base that can continue to accumulate data over time as more technological development takes place.

However, for endangered languages, the story is vastly different. Endangered languages face imminent extinction with a rapidly decreasing speaker base, a language going extinct every two weeks. Most endangered languages have less than 10,000 speakers which preclude the option of waiting for sufficient data to accumulate from user activity. For the speakers of these languages, NLP tools are almost never accessible because no one has ever conducted computational research for that certain language. It is important to realize that languages represent culture and history, and every time a language is lost, centuries of rich history and culture are lost. And as human society increasingly shifts online, more languages will die out even quicker due to a lack of data. This is why there need to be considerable efforts put into extending NLP technologies to not only LRLs but also endangered languages specifically, for the same reason we try to save endangered species.

However, this is no easy task. The fact that current paradigms require immense amounts of labeled data for training inhibits such technologies from being extended to endangered languages, because labeled data never springs up spontaneously, but requires 1) large amounts of available online text corpora 2) intensive human annotation. For endangered languages, there are far less text data than needed in order to reach performance on par with that of English (ex. BERT, which was trained on 3.3 billion words).

Traditional approaches to this problem of data scarcity has been to either obtain more data or use rule-based approaches. First, let us consider the means of obtaining more data for endangered languages. As per traditional data-collection approaches, there seem to be two main options: 1) annotate data for supervised learning for all tasks (such as NER, POS tagging, etc.) and domains (medical, legal, etc.) or 2) accumulate vast amounts of unannotated raw text data for unsupervised transformers. However, considering a lack of potential contributors, these methods

seem largely unviable. As per rule-based approaches, methods such as typological or morphological analysis can be a decent solution for low-level tasks such as lemmatization, dependency parsing and more. However, statistical models far outperformed rule-based models and thus rule-based models are no longer relevant in a vast majority of NLP tasks. Therefore, researchers studying LRLs increasingly rely on transfer learning techniques such as crosslingual zero-shot or few-shot to solve data scarcity.

There have been many attempts to implement such cross-lingual zero-shot training for many tasks [1, 3, 22]. However, previous research used task-specific word-level pairs like Hurtlex [2] for alignment, or pre-trained cross-lingual embeddings like MUSE [4] and LASER, which are black boxes that cannot undergo further fine-tuning. These approaches can be sub-optimal because there is no guarantee that making word embeddings aligned also makes sentence embeddings aligned either. For example, even if the word embeddings are aligned, there can be many morphological differences between two languages, such as English and Korean **as shown in figure 1**, which can prevent the model from accurately capturing the meaning of the sentence. Indeed, word embeddings proved to be effective for low-level semantic tasks including NER and dependency parsing which only require the understanding of individual words, but can be sub-optimal for sentence-length tasks related to sentence classification and prediction. One reason for this might be that although word embeddings are aligned, morphological and semantic differences between languages can prevent sentences from being aligned. This inspires us to rethink the primacy of cross-lingual word embeddings for lowresource sentence classification tasks. It may be better to utilize sentence embeddings that contain information at sentence length that can be directly applied to sentence classification tasks.

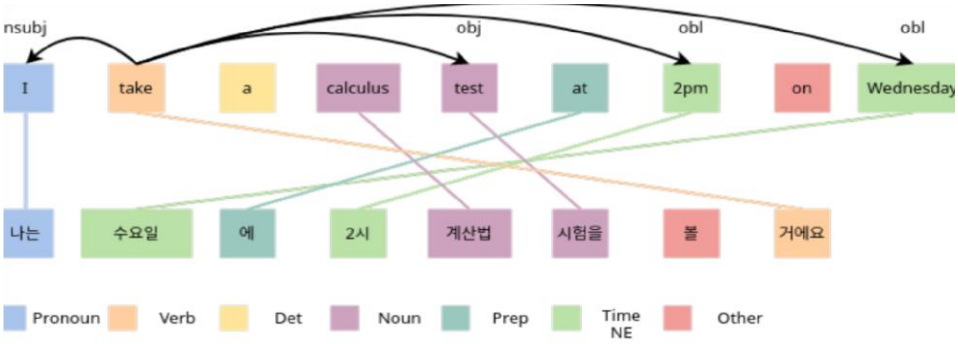

Figure 1. Morphological differences between languages cause problems for sentence classification tasks based on word embeddings.

Hence, we introduce a novel three-way cycle alignment technique targeted at not-only lowresource languages but primarily endangered languages, using sentence embeddings produced by pretrained cross-lingual language models (PXLMs). Although it may seem similar, our model aims to tackle a vastly different goal than most other cross-lingual zero-shot approaches. While other models tried to apply transfer learning to from a resource-rich language (such as English) to non-centered languages (Thai, Basque, etc.) that MMTs were still trained on, we are using a language that was never seen before by the PXLM, making it a much more challenging task.

The level of language speaker base can be ranked as follows, going from biggest to smallest: 1) resource rich langauges (English, Russian and many European Languages) 2) non-centered languages (Korean, Vietnamese, etc.) 3) vulnerable languages (usually less than 100,000 speakers) 4) definitely-endangered languages (less than 50,000 speakers) 5) severelyendangered languages (less than 10,000 speakers) 6) critically-endangered languages (less than 5,000 speakers) 7) Extinct. To make sure our methods are effective towards the most endangered

languages, we set the Jeju Language as our target language, which is a criticallyendangered language with a speaker base smaller than 5000 people. The Jeju language, despite being a Koreanic language, as vastly different semantic differences and cannot be understood by a native Korean speaker. However, it is true that Jeju language generally has similar syntactic structures as Korean. This reflects the current language distribution, as a vast majority of endangered languages are either variants (dialects) of a larger language or of similar origin thus evolved to have similar syntactic structures as a larger language. It is rare to find a complete language isolate. Although languages may be isolates, they can always be matched with a more resource-rich language with relatively similar syntactic structures.
In summary, the contribution of this paper is as follows:

1. I present a novel three-way cycle cross-lingual zero-shot sentence embedding alignment for effective cross-lingual zero-shot for endangered languages.
2. To the best of my knowledge, my work is the first zero-shot alignment attempt on an officially categorized endangered language, and therefore hopes to open the door to more research aimed at endangered languages.
3. The sentence embedding alignment method performs remarkably well on endangered langauges that PXLMs have not been trained on. The method also improves performances for non LRLs, showing that my approach can be effective even in non low-resource settings.

2. MOTIVATION

With many GPT[19]-variants sprouted, a myriad of global users have applied them to their daily tasks. This situation can be possible because GPT supports multilinguality, which means it was trained on multilingual data. This property seems impeccable at first glance. However, this multilinguality only applies to languages which receive enough attention to receive support, whereas critically endangered languages are out of the picture for researchers. This is because not only do those languages not have a large enough speaker base that can accumulate data over time, but also multilingual models can suffer from 'curse of multilinguality', which means that there can be a degradation of performance if language data that is too diverse is used during training. Curse of multilinguality prevents MMTs like XLM-R and GPTs from being able to accommodate for non-centered languages, let alone endangered languages further down the ladder. This makes me come to the conclusion that multilingual transformers alone cannot suffice in supporting endangered languages and motivates me to develop a further fine-tuning method for endangered language in extremely data-deficient settings. The inspiration of my embedding alignment process comes from CLIP's alignment process, a image-text based model that learns visual representations based on natural language supervision. Another motivation for this paper is the concerning lack of research on extending NLP to endangered languages, and to the best of our knowledge, there are none that actually experiments with an officially endangered language.

3. PRELIMINARIES

In this section, preliminaries to understand my methodology in detail will be discussed.

3.1. Word Embeddings

In natural language processing(NLP), a word embedding represents a real-valued vector-space representation of a word. Essentially, word embeddings are a way for the computer to understand the meaning of the natural languages that us humans use. Word embeddings are obtained through methods such as feature extraction and language modeling that maps each word onto a vector-space of varying dimensions as shown in **Figure 2**. Normally, how close each word is with

another represents the similarity in the meaning of the words. The most primitive form of word embedding is the frequency based embedding. Frequency based embeddings originated from the proposition that the more a word appears in a sentence, the more importance it will hold in that sentence. This form of word embeddings can be used for very simple tasks such as POS tagging and NER.

The earlier attempts to reduce the need for immense amounts of labeled data were to learn the general representations of words using unlabeled data, then integrating those representations for more task-specific frameworks such as GloVe [18] and Word2vec. Subsequently, sentence-level representation techniques like bidirectional transformers now generally outperforms word-level representations and BERT is one of many methods to extract such sentence embeddings.

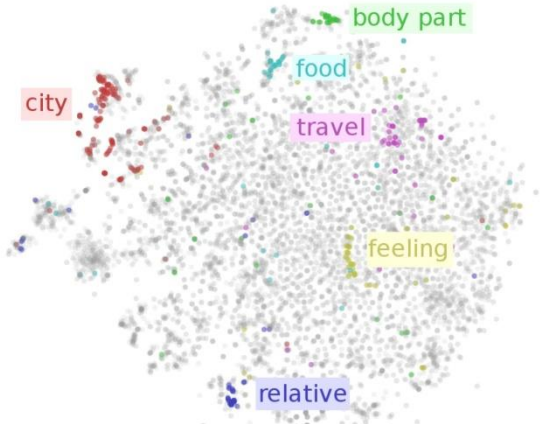

Figure 2. Word Embeddings

3.2. BERT(Bidirectional Encoder Representations from Transformers)

BERT [6] is one of the most advanced NLP models and has revolutionized the field of natural language processing. It processes the input text by encoding both the left and right context of each word in the sequence, allowing it to understand the context of each word and the relationships between the words in a sentence, which is not possible in word-embedding approach. This is different from traditional language models which only process text in a left-toright or right-to-left manner, thus limiting their ability to understand context. BERT can be finetuned on specific tasks such as question answering, sentiment analysis, and text classification, among others. The fine-tuning process is simple and fast, as the model is already pre-trained and only requires a small amount of training data for each new task. BERT is also transferable, meaning that it can be applied to different languages without the need for retraining from scratch.

3.3. Transfer learning and Zero-Shot training

Transfer learning is a technique where a model developed for one task is reused as the starting point for a model on a second task. This can save time and resources compared to training a model from scratch, especially when the sufficient target data cannot be obtained easily. The key idea is to transfer the knowledge gained while solving one problem to a different but related problem. For example, fine-tuning a language model like BERT to a specific task can be classified to transfer learning because its original task is predicting masked token so that context of a word in each sentence can be grasped by the model. Transfer learning can be applied in **Zero-shot training** too. Zero-shot training means training a model in a way that recognizes new classes of objects or perform new tasks without being explicitly trained on such examples. For instance, a multilingual transformer model which was trained to process English sentences can be

reused for other languages if the model is trained well so that it can generalize the context and structure of language. This technique will be mentioned again in the section of methodology later.

3.4. Embedding Alignment

Embedding alignment refers to the process of aligning two sets of embeddings, which are numerical representations of words or phrases in a high-dimensional space. This is typically done in order to achieve zero-shot or few-show training in some ways, such as measuring the cosine similarity and using it as a sort of labels instead of manually made annotation, so that similar embeddings are placed nearby, and disparate embeddings are placed very far. It has been globally used across many domains like computer vision, as well as natural language processing. For instance, CLIP[20] is one of the most successful application of embedding alignment approach to achieve open-vocabulary object detection, which means that unlike traditional object detection process, each image is paired with a raw sentence depicting the corresponding image, not a label. This feature facilitates gathering data from Internet because labor-intensive manual annotation process is not required. Likewise, for cross-lingual task in natural language process, instead of annotating labels to each target-language sentence, we can just use existing source-language sentence and label data, and align embeddings of targetlanguage sentence to the source-language one. This process can be done in word-level or sentence-level. The latter one was picked for the experiment in this paper.

4. RELATED WORKS

4.1. Monolingual Sentence Classification

Classifying monolingual data is one of the most long-researched topic in NLP field, and there are various approaches to deal with this task.

Kim. (2014) [11] was the first adoption of 1D Convolution operation for sentence classification task. Unlike previous usage of CNN only in computer vision tasks, he shows that convolution operation can also applied in word-embedding sequence matrix so that several words can considered for each operation at once. After this research, several other improved approaches include recurrent convolution network [12], attention-based convolution network [29], deeppyramid convolution network [10], etc. A series of researches shows that CNN can be a good choice for solving NLP tasks.

SVM(Support Vector Machine) is another way to solve the task without neural network. In this early days when neural network didn't get much attention, models like SVM played a central role doing such tasks. Joachims et al. (1998) [9], showed that SVM can categorize texts with relevant features.

LSTM(Long Short-Term Memory) [7, 21, 30] is another famous neural network model which was used with RNN(Recurrent Neural Network) for text classification before BERT was released. LSTM and RNN is a model very suitable for sequence data, but it has some weakness like vanishing gradients and also they are not parallelizable which can be a huge bottleneck in the era of GPU.

After Transformer network [27] was released, models like LSTM was replaced with Transformer-variants like BERT [6]. Thanks to effectiveness of attention mechanism and parallizablity, those Transformer-variants show state-of-the-art performances in NLP domain as

well as other domains like computer vision. BERT are further developed with many approaches and RoBERTa [15] and DistilBERT [23] are such examples.

4.2. Cross-lingual Sentence Classification

For researchers who use languages that can't get attention from others, It's not easy to glean enough data for training a model as opposed to major languages like English or Chinese.

Difficulty of gathering such data induces researchers to dive into contriving many methods to make it possible to improve performance of machine translation or zero-shot cross-lingual training without such data of minor languages [12, 8]. More recently, MUSE [4] was released, which is an universal multilingual unsupervised embedding, and has been a powerful tool for those researching zero-shot cross-lingual training [1]. However, I proposed that using such pretrained MUSE has a potential problem : using word embeddings rather than sentence embedding can be sub-optimal for solving specific tasks because hardly do word embeddings connote enough meanings related to specific tasks.

5. METHODOLOGY

5.1. Dataset

To validate our approach, three types of datasets are required, which are following.

1. Task-specifically annotated English sentence data
2. English-Korean translation corpus
3. Task-specifically annotated Korean sentence data

Using the first and second dataset, BERT language model and projection matrix are fine-tuned, and validate the model with the third dataset. Because the third one is only for validation, there's no need for much annotated data for such low-resource languages. Sentiment analysis is chosen as the task for classification, rotten-tomatoes movie reviews for the first dataset, and refer to AIHub and Korpora for the second and third dataset.

5.2. Task-Specific Fine Tuning

First, following a normal setting to fine-tune BERT, the model is trained with the first dataset. Each sentence is tokenized with BertTokenizer, and corresponding word embeddings are used to calculate attention score between themselves. Then, output embedding of CLS token from pooler layer is fed to feed forward network and the network finally outputs probability of each label. Then, the whole network will be updated comparing the output probability and groundtruth label. I decided to update the whole model without freezing BERT. The process is illustraed in **Figures 3, 4, 5**.

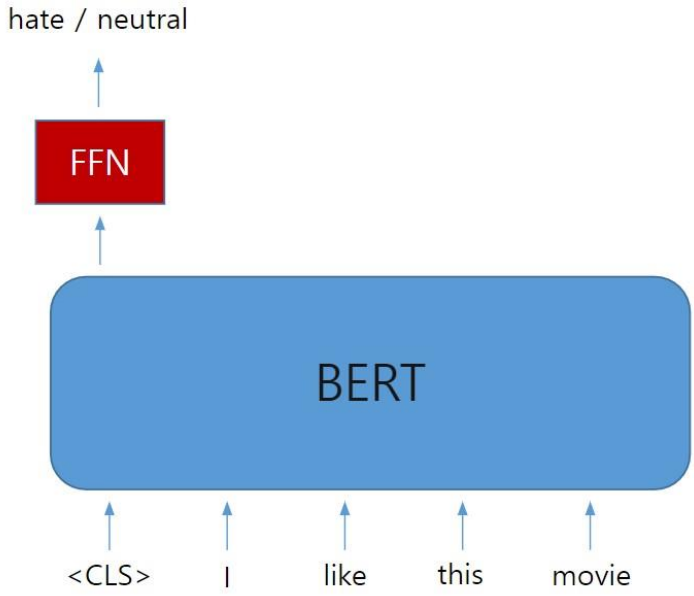

Figure 3. BERT model is firstly fine-tuned for any chosen task

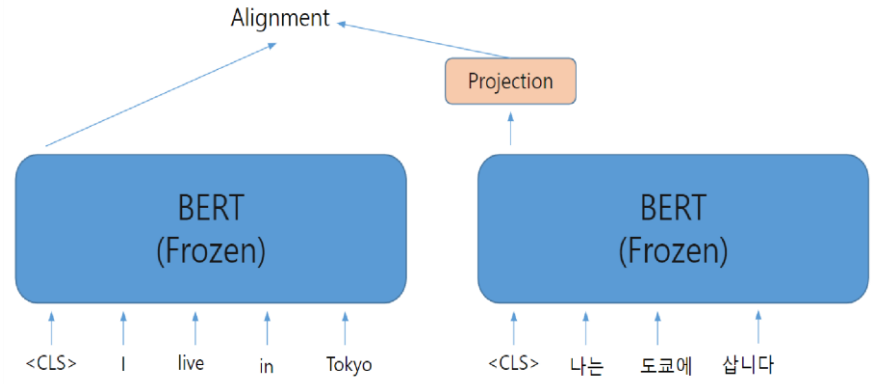

Figure 4. Sentence embeddings of two languages are aligned in the common space

5.3. Cross-Lingual Embedding Alignment

What we have to do is not update the language model entirely, but simply introduce a lightweight projection layer which maps target(minor) sentence embedding into source(major) sentence embedding in a way to make the embedding more suitable for our task-specific feed forward network so that the model is untouched during alignment fine-tuning.

Using parallel corpus of English and Korean, the matrix of projection layer is updated correspondingly. Because a pair of embeddings of two same meaning sentences are aligned to be near each other in the embedding space, it is expected that the model now can classify Korean sentences too if the pretrained FFN in phase 1 is used. You need to notice that the components that are updated are only weights of projection matrix, which means that we freeze all the weights in BERT because it is my objective to show that comparable performance can be obtained in zero-shot setting only with light-weight projection matrix without updating pretrained language model. Alignment process can be formulated as following. Let $E = \{e_1, e_2, ..., e_l\}$ be a set of word

embeddings consisting of a single English sentence, and $K = \{k_1, k_2, ..., k_{l2}\}$ is corresponding word embeddings consist of a Korean sentence. Then, the loss can be calculated as following.

$$\mathscr{L}_{align} = \mathscr{L}(\mathscr{B}(E), \mathscr{P}(\mathscr{B}(K)))$$

$\mathscr{B}$ means pooler output embeddings of BERT and $\mathscr{P}$ means projection layer. Mean-squared error is chosen as a loss function.

5.4. Inference

Because the model and FFN is trained to correctly classify English sentences in phase 1, and projection layer is also trained to output similar embeddings for same-meaning English sentence and Korean embeddings, now FFN can classify embeddings which is from Korean sentences if it was through pretrained projection layer that was trained in phase 2. It is meaningful that the result is without further-training the huge BERT model in phase 2; what was done is only updating light-weight matrix for alignment. This separates fine-tuning and alignment process, so performance can be improved if English corpus data is further augmented, and this doesn't require the alignment matrix be further updated correspondingly.

6. EXPERIMENT

6.1. Implementation Detail

For the implementation of multilingual BERT, XLM-RoBERTa, which is a multilingual version of RoBERTa, was chosen and implemented with help of Huggingface library. The model was

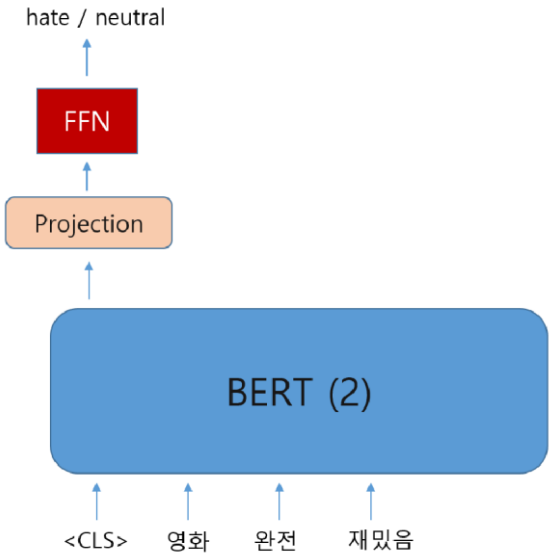

Figure 5. Now the model can classify Korean sentences without training dataset

chosen because it is meaningful to use well-trained multilingual model for showing that my methodology can further improve the performance of such pre-trained one. Tokenizer was imported from the same source. Hyperparameter settings are following : batch size to 16, embedding and hidden dimension size to 768 (which is a default size of Huggingface model), learning rate to 1e-4, and epoches to 5. As for optimizer, AdamW loss is selected for the first

phase, and MSE loss for the second phase. The number of sentences in each dataset is following : 751,549 English sentences for the first phase, 1 million general, daily-life sentence pairs for the second phase, 150,001 Korean sentences for the third phase. You need to notice that the third phase is only for inference, so actually there's no need to use such lots of sentences. This property would be desirable for users of languages which are minorer than Korean and have scarce data in Internet. The structure of feed forward layer for classification is following

$$\mathscr{D}(\mathscr{B}(\mathscr{M}(768,256)),0.5) \rightarrow \mathscr{D}(\mathscr{B}(\mathscr{M}(256,category)),0.5).$$

$\mathscr{M}(a,b)$ means a linear layer which maps dimension a to dimension b, $\mathscr{B}$ means batch normalization layer, and $D(\cdot,p)$ means a dropout layer with probability p.

6.2. Ablation Study

Table 1. All the combinations of each three component which are excluded(Projection layer) or not fine-tuned(Model and FFN layer) were tested. The penultimate accuracy is similar to the traditional approach

Model	FFN	Projection	Accuracy
X	X	X	0.575
O	X	X	0.519
X	O	X	0.424
X	X	O	0.558
X	O	O	0.475
O	X	O	0.431
O	O	X	0.866
O	O	O	**0.936**

To show the real effect of the method, several tests which are done with combinations with each of three component excluded or not fine-tuned, whose result can be seen in **Table 1**. The result shows that not only fine-tuning of BERT model and feed forward layer but also the projection method proposed has further efficacy to improve performance for performing bilingual zeroshot task. The effectiveness is evident in comparing between two last rows in table, which are without projection matrix (traditional zero-shot approach) and with it. According to [13], Korean is one of languages that show the most significant performance drops among other languages in cross-lingual zero-shot setting due to its huge semantic differences and also this tendency is displayed in the prior works. Even so, the result shows that our fine-grained approach can compensate for difference of those languages which have considerable difference from the source language like Korean. Also, the most prominent advantage of the method is that it is generalizable and can be applied for languages that have extreme suffer from resources in Internet. However, the number of training samples in training phase 2 (alignment process) was a million. Even though this number of data is not that extreme for languages that have moderatesized population like Korean, it can be quite burdensome and not realizable amount to collect for other languages even including ones that nearly face extinction. Therefore, another ablation experiment was done, where alignment process(training phase 2) was done with only a part of data, which can be seen in **Table 2**. The result clearly shows that the method can be applied with even more smaller data in that the performance was further improved even with 10,000 data in comparison with the model without projection matrix. Even though the test was done with a million data to achieve

highest accuracy as possible, using 100,000 data seems more adequate to balance between burdensomeness and performance.

Table 2. The accuracy improvement of enlarging the number of data from 100,000 to a million is quite marginal

Samples	Accuracy
1,000	0.796
10,000	0.889
100,000	0.930
1,000,000	**0.936**

6.3. Extending Zero-Shot to Few-Shot

To show that the method also has efficacy in a few-shot setting, further fine-tuning using a few target-language data was done in two models, which are a model without projection layer and the one with it, which can be seen in **Table 3**. The result clearly shows that projection layer can be helpful even in few-shot setting and performance can be further improved than zero-shot.

Table 3. The performance can be even further improved in a few-shot setting.

Few Shot	Projection	Accuracy
X	X	0.886
X	O	0.936
O	X	0.895
O	O	**0.941**

Table 4. Same experiment was done for validating possibility of applying the method to Jeju language

Model	FFN	Projection	Accuracy
X	X	X	0.665
O	X	X	0.276
X	O	X	0.334
X	X	O	0.644
X	O	O	0.678
O	X	O	0.717
O	O	X	0.789
O	O	O	**0.883**

6.4. Application to more Minor Language : Jeju Language

Even though Korean is a minor language compared to English or Chinese, it might be not minor enough compared to other endangered languages to validate the power of the method. Therefore, further experiment was done on more minor language, which is 'Jeju Language.' Jeju language is one of dialects of Korean, which is too disparate and abstruse for ordinary Korean speakers to comprehend. There has been an everlasting debate on whether or not Jeju language is a dialect of Korean or a distinct language, but it is chosen as second target language for two reasons. First,

only about ten thousands fluent speakers, mostly above 70 years of age are left and younger generation in Jeju island do not learn their dialect, which means that it is facing imminent extinction. Second, there is only one parallel dataset, JIT, created by few scientists, which obviously show its scarcity of data in Internet. To the best of my knowledge, it is the first zero-shot attempt on Jeju language, and I believe that the experiment can be applied similarly to other endangered languages. To construct data for inference, I manually picked about 10k sentences which clearly express sentiment in JIT. The result is shown in **Table 4**. Due to scarcity of data, not 1 million but 100k parallel data was used for training phase 2. Ablation study states that 100k parallel data is enough to get similar result compared to 1 million. Considering that experiment was not done on major language, the result is quite satisfactory. The interesting thing is, this fine result shows transitivity property of alignment process, meaning that even though I used parallel corpus of Korean-Jeju language, not English-Jeju language, and still the task-specific fine-tuned language model was trained on English, the result shows good result. This is possible because pre-trained English-Korean alignment layer is added for also the source language in alignment process in phase 2. This process is depicted in **Figure 6**. Transitivity can be very useful for minor languages in that even though researchers

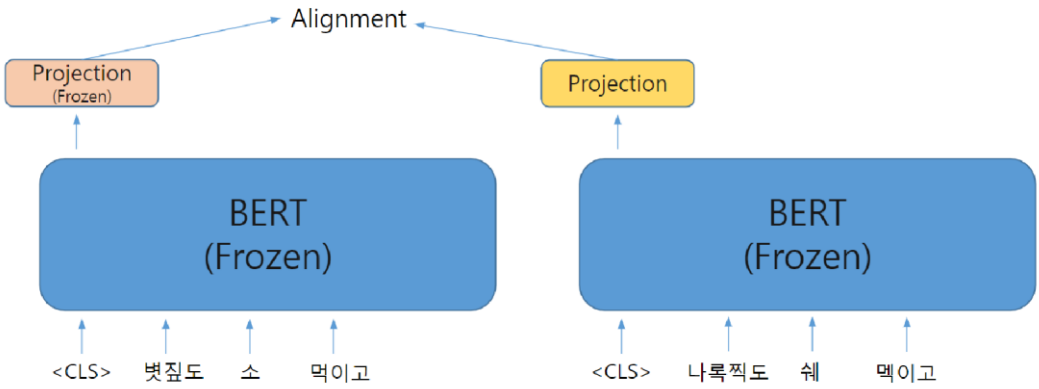

Figure 6. The alignment process has transitivity property

didn't find English-Target language corpus, they can utilize other corpus(for example, ChineseTarget language or Spanish-Target language, etc) because English-Chinese or English-Spanish parallel corpus is abundant in the Internet.

6.5. Application to Diverse Tasks

(this result shows generalizability) To further valid the method, one more experiment was done, where I chose NLI(Natural Language Inference) as a downstream task. Because NLI is more sophisticated task than simple sentiment analysis in that it needs to infer relation between two sentences rather than single one, it can be more challenging for the method. For training dataset in phase 1, I picked 100k sentence pairs from Multi NLI dataset for training, and 6k sentence pairs from Korean NLI dataset for inference due to limited hardware resources. Like previous experiments, ablation study of projection layer was done. The result is shown in **Table 5**. Even though it is zero-shot training, the model achieved decent performance. The noticeable thing is that the experiment was done using projection layer which was trained with a sentimentanalysis fine-tuned model before, not NLI dataset. This is quite interesting because it means that retraining of projection matrix is not necessary along diverse downstream tasks, but only once is sufficient. To validate this claim, one more experiment was done in **Table 6**. Each of row means that the projection layer was trained with a model that was not fine-tuned, fine-tuned with sentiment-analysis dataset, and NLI dataset.

Table 5. Ablation study of projection layer in NLI dataset

Model	FFN	Projection	Accuracy
X	X	X	0.144
O	X	X	0.353
X	O	X	0.421
X	X	O	0.353
X	O	O	0.503
O	X	O	0.353
O	O	X	0.631
O	O	O	**0.740**

Table 6. Comparison of projection layer that was trained on different dataset

Dataset	Accuracy
None	0.728
Sentiment	0.740
NLI	**0.747**

6.6.NLI in Jeju Language

Lastly, to validate the capability of the method in applying more complex zero-shot tasks, one more experiment evaluating the model in Jeju-NLI dataset, which was collated by labeling manually to each Jeju sentence. The number of sentences is around 1k, and the result in **Table 7**. As you can see in the table, the performance gap between using projection matrix and not using it is quite large. I suppose that it is partly because the language model has never seen Jeju language in its pre-training time. And this highlights the advantage of my method applying in minor languages in that the pre-trained model may have never seen that language before and my method can rectify the model's incapability of treating such languages.

Table 7. Ablation study of projection layer in Jeju NLI dataset

Model	FFN	Projection	Accuracy
X	X	X	0.089
O	X	X	0.080
X	O	X	0.339
X	X	O	0.089
X	O	O	0.061
O	X	O	0.419
O	O	X	0.476
O	O	O	**0.848**

7. CONCLUSION

Even though several embedding alignment techniques are proposed for cross-lingual zero-shot training, they all use pretrained cross-lignual embedding which can't be further fine-tuned, or update the whole model for alignment which can cause instability. In this paper, I proposed alignment methodology with only a light-weight projection matrix and the experiment shows that it is enough for alignment to use such simple layer. I hope that other refinement techniques (like using more sophisticated loss function) is done on my research so that performance can be boosted further.

REFERENCES

[1] A. Arango, J. Pérez, and B. Poblete. Cross-lingual hate speech detection based on multilingual domain-specific word embeddings. *arXiv preprint arXiv:2104.14728*, 2021.

[2] E. Bassignana, V. Basile, and V. Patti. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS, 2018.

[3] I. Bigoulaeva, V. Hangya, and A. Fraser. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, 2021.

[4] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. J´egou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[5] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*, 2013.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Z. Ding, R. Xia, J. Yu, X. Li, and J. Yang. Densely connected bidirectional lstm with applications to sentence classification. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 278–287. Springer, 2018.

[8] G. Dinu, A. Lazaridou, and M. Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014.

[9] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[10] R. Johnson and T. Zhang. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume1: Long Papers)*, pages 562–570, 2017.

[11] Y. Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014. URL http://arxiv.org/abs/1408.5882.

[12] S. Lai, L. Xu, K. Liu, and J. Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[13] A. Lauscher, V. Ravishankar, I. Vuli´c, and G. Glava˘s. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. arXiv preprint *arXiv:2005.00633*, 2020.

[14] X. Li and D. Roth. Learning question classifiers. In COLING 2002: *The 19th International Conference on Computational Linguistics*, 2002.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta:A robustly optimized bert pretraining approach. arXiv preprint *arXiv:1907.11692* , 2019.

[16] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation (2013). arXiv preprint *arXiv:1309.4168*, 2022.

[17] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. arXiv preprint *cs/0506075*, 2005.

[18] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[19] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[21] G. Rao, W. Huang, Z. Feng, and Q. Cong. Lstm with sentence representations for document- level sentiment classification. *Neurocomputing, 308:49–57*, 2018.

[22] M. S. Rasooli, N. Farra, A. Radeva, T. Yu, and K. McKeown. Cross-lingual sentiment transfer with limited resources. *Machine Translation, 32(1):143–165*, 2018.

[23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint *arXiv:1910.01108*, 2019.

[24] J. Schmidhuber, S. Hochreiter, et al. Long short-term memory. *Neural Comput, 9(8):1735–1780*, 1997.

[25] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing, 45(11):2673–2681*, 1997.

[26] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642*, 2013.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems, 30*, 2017.

[28] B. C. Wallace, L. Kertz, E. Charniak, et al. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 512–516*, 2014.

[29] Z. Zhao and Y. Wu. Attention-based convolutional neural networks for sentence classification. In *Interspeech, volume 8, pages 705–709*, 2016.

[30] C. Zhou, C. Sun, Z. Liu, and F. Lau. A c-lstm neural network for text classification. arXiv preprint *arXiv:1511.08630*, 2015.