Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2841–2849
Marseille, 20-25 June 2022

Korean Language Modeling via Syntactic Guide

Hyeondey Kim[1], Seonhoon Kim[2], Inho Kang[2], Nojun Kwak[3], and Pascale Fung[1]
[1]The Hong Kong University of Science and Technology
[2]Naver Search
[3]Seoul National University
hdkimaa@connect.ust.hk, seonhoon.kim@navercorp.com, once.ihkang@navercorp.com,
nojunk@snu.ac.kr, pascale@ece.ust.hk

Abstract

While pre-trained language models play a vital role in modern language processing tasks, but not every language can benefit from them. Most existing research on pre-trained language models focuses primarily on widely-used languages such as English, Chinese, and Indo-European languages. Additionally, such schemes usually require extensive computational resources alongside a large amount of data, which is infeasible for less-widely used languages. We aim to address this research niche by building a language model that understands the linguistic phenomena in the target language which can be trained with low-resources. In this paper, we discuss Korean language modeling, specifically methods for language representation and pre-training methods. With our Korean-specific language representation, we are able to build more powerful models for Korean understanding, even with fewer resources. The paper proposes chunk-wise reconstruction of the Korean language based on a widely used transformer architecture and bidirectional language representation. We also introduce morphological features such as Part-of-Speech (PoS) into the language understanding by leveraging such information during the pre-training. Our experiment results prove that the proposed methods improve the model performance of the investigated Korean language understanding tasks.

Keywords: Neural language representation models, Semi-supervised, weakly-supervised and unsupervised learning, Part-of-Speech Tagging

1. Introduction

Recent progress in machine learning have enabled neural language models to move beyond traditional natural language processing tasks such as sentiment analysis and pos-tagging. Modern language processing systems are now equipped to handle complex tasks such as question answering (Rajpurkar et al., 2016), dialogue systems (Sun et al., 2019) and fact-checking (Thorne et al., 2018) that all require sophisticated language understanding capabilities.

The pre-trained language model (Devlin et al., 2018; Lewis et al., 2020) made significant breakthroughs in natural language processing. In most natural language processing tasks, contextual language representations trained from massive unsupervised learning with enormous plain texts achieve state-of-the-art performance. However, most of the computational linguistics research is focused on English. In order to build a language model for less commonly studied languages like Korean, it is necessary to focus on the target language's linguistics characteristics. Unfortunately, the Korean language has very different linguistic structures from the other languages; Korean is classified as a language isolate. As a result, language modeling is extremely challenging in Korean.

The concept of a language model can be explained as an algorithm that assigns probability values to words or sentences. Language models are typically trained by predicting the next token based on given context (Roark et al., 2007). However, the technique cannot be applied to languages with SOV order like Korean and Japanese. In a language with such structure, most vital information like verb is placed at the end of the sequence. What makes Korean language modelling even more difficult is that Korean is often order-free. Therefore, it is impossible to predict the next token in many cases. It creates a need to train the Korean language model with a new approach that can be helpful to understand its specific linguistic structure.

Although there are existing works on a language model for multiple languages such as Multilingual BERT, researches on Korean language modeling are extremely rare and limited. Various language versions of existing language models are available and show impressive performances. However, the multilingual version of BERT shows less performance compared to the English version (Pires et al., 2019), and most of the researches on the pre-trained language models are mainly focusing on English. Most of the recent works on language modeling such as BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), BART (Lewis et al., 2020), and ELECTRA (Clark et al., 2020) are trained for English. Therefore, we need to propose a new language model for the Korean language.

There is limited available data for the Korean language. The text contents on the web provide sufficient training corpora in English language modeling. Generally, knowledge plentiful corpus such as Wikipedia articles

are widely used for pre-training language model (Devlin et al., 2018), but the distribution of the number of articles in Wikipedia[1] by languages is very imbalanced. Thus, gathering sufficient corpus from the web content for less-studied languages is impossible or extremely difficult. Despite the low volume of data for less-studied languages, considering that significantly large numbers of people have a language other than English as their first language, designing a language model for such a minor language is necessary. Furthermore, the Korean language occupies less than 1% of web content. It only contains 75,184 articles on Wikipedia (English contains 2,567,509 articles). Therefore, we should focus on practical training for the Korean language model with smaller model size and less training data instead of leveraging tons of data and computational power.

Besides, typical language modeling with predicting the next tokens such as N-gram (Roark et al., 2007) is not applicable for order-free languages such as Korean and Japanese. Changing sequence order derives the changing of syntactic meaning in most Indo-European languages and Chinese languages. However, in an agglutinative language such as Korean and Japanese, not the sequential position of the word but its postposition primarily determines the syntactic meaning (Ablimit et al., 2010). Hence, clause or phrase level order shuffling does not influence the meaning of the entire sentence in many cases. Therefore, we need to build a language model for agglutinative languages with new approaches. Mainly focusing on a less studied agglutinative language, *Korean*, we enhance the language model to learn more about the grammar structure and features of the Korean language. Based on the masked language model (Taylor, 1953), we tag the PoS of the corpus and train the model to predict the part-of-speech of each token (NA and KIM, 2018). Also, we permute each sentence at a phrase and clause level to predict the original order and masked token simultaneously.

We conduct various experiments in several settings. The results show that our proposed method outperforms the baseline model in every downstream task. Furthermore, it proves that our approach guides the model to learn more generalized and robust features with low resources. Our contributions are summarized as follows:

- We propose a novel pre-training method, syntactic injection, to enhance the grammar understanding skill of the language model. Our proposed method improves performance on every Korean NLP task.
- We present chunk-wise reconstruction for pre-training Korean language modeling. Our approach shows effectiveness and robustness on some Korean NLP tasks that include scrambled sequence recognition.

[1] https://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics

2. Related Work

Out of vocabulary (OOV) is one of the main problems in modeling an agglutinative language. In Korean, too many combinations exist by combining different postpositions, such as Josa and Eomi. We introduce several works for the Korean language model.

A syllable-level language model (Yu et al., 2017) is proposed for the Korean language to solve the OOV problem. However, due to the agglutination of the Korean language, too many possible combinations exist for each verb and the nouns.

KR-BERT (Lee et al., 2020) is a BERT-based Korean language model. By considering the language-specific properties of the Korean language, the proposed KR-BERT model shows better performance than multilingual-BERT (Pires et al., 2019). Also, KR-BERT proposes sub-character level tokenization and Bidirectional BPE tokenization to enhance the understanding of Korean grammar. As a result, even with a smaller dataset and smaller model size, KR-BERT shows better or equal performance than BERT's multilingual version or other Korean-specific models.

Tokenization strategies on Korean language modeling are crucial to the performance of the language model. According to the investigation, results on the various tokenizers (Park et al., 2020) include a CV (consonant and vowel), Syllable, Morpheme, Subword, Morpheme-aware subword, and Word level, although CV tokenizer (character-level) and Syllable level tokenizer have the lowest OOV rate, however, Morpheme-aware sub-word tokenizer shows the best performance on most of the Korean NLU tasks. On the other hand, the word-level tokenizer shows the worst performance due to the OOV issue. This work indicates that linguistic awareness is a significant key to improving language model performance.

To sum up, most of the works are focused on the agglutinative of the Korean language and propose the tokenization methods on Korean language modeling. Various results show that separating postpositions from the words improves the effectiveness of the tokenizer and improves the final language representations. However, none of the works has focused on Korean as an order-free language. Moreover, linguistics phenomenon such as scrambling is not considered in Korean language modeling.

3. Methodology

Mainly focusing on a less studied agglutinative language, *Korean*, we enhance the language model to learn more about the grammar structure and features of the Korean language. Based on the masked language model (Taylor, 1953), we annotate the corpus with PoS tags and train the model to predict the part-of-speech of each token (Na, 2015) (NA and KIM, 2018). Also, we permute each sentence in a phrase and clause level to predict the original order and masked token.

Approach	Input Sequence
Original Sequence	언어모델 개발은 중요하다. *It is important to build language model.*
Baseline MLM	언어모델 [MASK]은 중요하다. *It is important to [MASK] language model.*
Chunk Reconstruction	언어모델 중요하다. [MASK]은 *language model important to [MASK] It is.*

Table 1: Input sequences and labels of each pre-training task. Italic sentences are the English translation of Korean sentences.

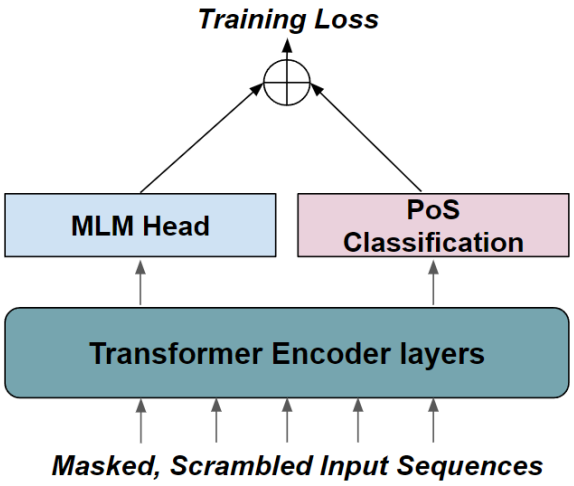

Figure 1: Overall framework of the proposed model. The loss value of the model is the combined value from the masked language model head and the PoS classifier.

PoS Tags	Meaning of Tags
JOSA	Postposition or particles
EOMI	Ending of Verb
SUFFIX	Suffix
CJK	Chinese Characters
VERB	Verb
MOD	Determiners
NOUN	Noun
NUMBER	Arabic Numbers (0-9)
ALPABET	Alpabets (A-Z and a-z)
PRONOUN	Pronoun
PREFIX	Prefix
NUMSUFFIX	Suffix of number
NUMNOUN	Noun of number and numerals
MIXED	Mixed Part-of-Speech
NBN_N	Dependent noun
PAD	Tag for PAD tokens
REST	Punctuation and etc

Table 2: Types of Part-of-Speech in the tokenizer

(1) 컴퓨터는 [computer-nun] 언어를 [eone-lul] 이해해 [ihae-hae]
computer-TOP language-ACC understand-DEC.INF
'Computer understands language.'

3.1. Masked Language Model

Masked language model, as known as cloze task, predicts *masked* tokens. We replace 15% of tokens to **[MASK]** token. Unlike BERT (Devlin et al., 2018), we do not modify or replace the masked tokens with the original or random token. Let $\hat{m}$ be the predictions, and we leverage the cross-entropy loss function. Hence, for each masked token, let m be the original token. Then the loss value $Loss_{mlm}$ for the masked language model is

$$\mathcal{L}_{mlm} = -\sum m \log \hat{m} \quad (1)$$

3.2. Syntactic Injection

Syntactic understanding is the most critical key for Korean language understanding to facilitate understanding of the syntactical structure and enhance the model's capacity for syntactic processing. We leverage an off-the-shelf PoS tagging module from KoNLPy (Park and Cho, 2014). Among the various PoS-tagging module KoNLPy provides, we select Twitter PoS-tagger. Twitter Korean Text[2] is an open-source Korean tokenizer written in Scala. The total types of PoS-tags are described in Table 2. Given the example sentence.

(2) 한국어를 [Hankukeo-lul] 처리하는 [cheori-hanun] 예시입니다[yesi-ipnida].
Korean-TOP process-ACC example-DEC.INF
'This is an example of processing Korean'

The output of the PoS tagging tokenizer is:

(3) 한국어 Noun, 를 Josa, 처리 Noun, 하다 Verb, 예시 Noun, 이다 Eomi.

We classify all tokens in corpus with part of speech (PoS) tag. Exclude PAD tag for padding tokens, the total amount of tags are 17. Table 2 describes the list of Part-of-Speech to classify. We implement a PoS classifier on the top of transformer encoders. Let $\mathcal{L}_{PoS}$ be the loss value, $\hat{p}$ be the predictions for the token, and p be the true PoS tags of the input sequence, the objective function of PoS tagging is

$$\mathcal{L}_{PoS} = -\sum p \log \hat{p} \quad (2)$$

[2] https://github.com/twitter/twitter-korean-text

Hyper parameter	Value
Epoch	5
Batch size	32
Learning rate	5e-5

Table 3: Hyper parameters for fine-tuning our models on test datasets

3.3. Scrambled Chunk-wise Reconstruction

Based on the given PoS information, we split the given sequences into chunks. Definition of Korean phrase is equal to the part of the sentence that is parsed by the postpositions (Josa and Eomi). By permuting chunks, some sequences are scrambled with no change of semantic meaning, and the semantic meaning of some sentences is damaged. We redefine the pre-training task by restructuring the scrambled and shuffled chunks.
Agglutinative of Koren language makes Korean hard to be trained by next-token prediction task. Therefore, we train our language model via masked language model (Devlin et al., 2018) (Cloze task (Taylor, 1953)). Also, based on the order-free character of the Korean language, we train our language model via permutation language model (Yang et al., 2019; Lewis et al., 2020) and Scrambling-based language model.
Given an example sentence:

(4) 선수가 [seonsu-ga] 쏜 [sso-n]
화살이[hwasal-i] 과녁의 [gwanyeog-ui]
한가운데를 [hangaunde-leul]
맞추었다 [majchu-eoss-da]
player-NOM shoot-MOD.PST
arrow-NOM target-GEN
center-ACC hit-PST-DEC
'The arrow that the player shoot has hit the center of the target'

We replace the 15% of input sequence to *[MASK]* tokens.

(5) 선수가 [seonsu-ga] [MASK]
화살이[hwasal-i] 과녁의 [gwanyeog-ui]
한가운데를 [hangaunde-leul]
맞추었다 [majchu-eoss-da]
player-NOM [MASK]
arrow-NOM target-GEN
center-ACC hit-PST-DEC
'The arrow that the player [MASK] has hit the center of the target'

For the typical permutation language model, we permute tokens randomly.

(6) 한가운데를 [hangaunde-leul] 선수가 [seonsu-ga]
화살이[hwasal-i] 과녁의 [gwanyeog-ui]
맞추었다 [majchu-eoss-da] [MASK]
center-ACC player-NOM
arrow-NOM target-GEN
hit-PST-DEC [MASK]
'The arrow that the player [MASK] has hit the center of the target'

However, for the chunk-wise reconstruction, we shuffle the sequences in chunk (clause) level.

(7) 과녁의 [gwanyeog-ui] 한가운데를 [hangaunde-leul]
선수가 [seonsu-ga] [MASK]
화살이[hwasal-i] 맞추었다 [majchu-eoss-da]
target-GEN center-ACC
player-NOM [MASK]
arrow-NOM hit-PST-DEC
'The arrow that the player [MASK] has hit the center of the target'

To process the scrambled chunk-wise reconstruction token by token. Let t_i be original tokens at i-th position, $\hat{t_i}$ be the prediction at i-th position, the objective function is:

$$\mathcal{L}_{chunk} = -t_i \log \hat{t_i} \tag{3}$$

3.4. Model

Merging all of the aforementioned methods, we train our model with a masked language model, syntactic injection (PTP), and scrambled chunk-wise reconstruction (SCR). Based on BERT (Devlin et al., 2018) model, we implement transformer (Vaswani et al., 2017) encoders with several layers. On the top of the encoder layers, we connect two linear layers, one for the masked language model head and the other for the PoS tagging classifier. The final loss value $loss_{final}$ is the sum of losses mentioned above. However, we perform a masked language model and scrambled chunk-wise reconstruction simultaneously. Therefore, the objective function of the entire model is:

$$\mathcal{L}_{total} = \mathcal{L}_{chunk} + \mathcal{L}_{PoS} \tag{4}$$

Given example 1 as the input sequence, we describe the different inputs of our models in Table 1. We make noise to the given sentence not only permutate the sentence in the chunk level but also mask 15% of tokens of the sentence. Therefore, the $\mathcal{L}_{mlm}$ and $\mathcal{L}_{chunk}$ play an identical role in the pre-training stage. Figure 1 illustrates the structure of our model.

4. Experiments

We train our model with 5e-4 of learning rate and 512 of batch size with 128 of max sequence length. Based on the BERT model, we have 6-layers encoders and 768 for the hidden size of each layer. For both pre-training and fine-tuning, we set 42 as the random seed.

4.1. Training Data

For the Training data, to attain general knowledge and generalize the feature, we collect corpus from Korean Wikipedia[3] and Namu-wiki[4], which are open to the public. The Korean Wikipedia is generally written in relatively formal language and contains academic knowledge. On the other hand, the Namu-wiki corpus

[3] https://ko.wikipedia.org
[4] https://namu.wiki

is generally written in informal languages and mainly contains non-academic information. The total amount of the corpora is 6GB.

4.2. Test Data

To verify trained language model, we test our models on various Korean tasks. We fine-tune our models to NSMC[5], Q-Pairs[6], KoreanNLI (Ham et al., 2020), QS, KoreanNER[7], and Korean hate-speech detection (Hate-speech) (Moon et al., 2020). Hyper parameters for fine-tuning are mentioned in Table 3.

4.2.1. NSMC

NSMC[8] is a movie review classification dataset in the Korean language. Reviews were scraped from Naver Movies. It contains 200K reviews for the training set. All reviews are shorter than 140 Korean characters. Each sentiment class are sampled equally. Thus random guess yields 50% accuracy. Reviews with 9-10 ratings are labelled as positive, reviews with 1-4 ratings are labelled as negative. All of the reviews are generated by humans. Therefore, most of the reviews contain noise such as 'ㅋ' or 'ㅇ'. Also, spacing is not strictly obeyed. For instance, see example 8, the example shows one of the sentences in the NSMC dataset.

(8) 교도소이야기구먼..솔직히재미는없닼ㅋㅋㅋ..
평점조정..
It is a jail story. Frankly speaking, it is not funny at all. Adjusting the ratings.

The user's comment contains a number of grammar errors. The sentence has spacing errors, misuse of '.', and sub-word noise 'ㅋ'. Also, it contains a part of a local accent, '구먼', which is not included in formal Korean. To conclude, improving robustness against such noisy, ungrammatical, nonstandard sentences is the key to solving this task.

4.2.2. Q-PAIR

Question pairs[9] dataset is a sentence classification dataset. The task requires predicting whether two given sentences describe the same meaning or not. Some of the datasets is generated by scrambling. Thus it is a good example to explain the robustness against the understanding of scrambling.

4.2.3. KorNLI

KorNLI (Ham et al., 2020) is a natural language inference dataset in Korean. KorNLI requires to predict the relationship of given two sentences into three labels(Entailment, Contradiction, Neutral). ham-etal-2020-kornli generated the training-set with machine translation, development-set and evaluation-set are generated by human translation. Since the training dataset is generated by machine translation, the development-set and test-set are generated by human translation. Adapting different domains is the key to improving performance.

4.2.4. QS

QS is a paraphrase identification dataset using questions from the community-based question answering system. Naver QA[10] is the largest question answering community in Korea. As Quora[11] and Stackoverflow[12], there are many duplicated questions being asked, even correct and the well-written answer already exists.

4.2.5. Korean Hate-speech Detection

Korean hate-speech detection dataset is a hate speech classification task that contains 3 labels for bias and 3 labels for hate speech. For the bias, it has 'gender', 'others' and 'none' labels. For the hate speech, it contains hate, offensive and none labels.

4.3. Result

Table 4 shows the result on downstream tasks compared to the baseline model and our models. The baseline model is only trained with a BERT-based masked-language model (MLM) without next sentence prediction (NSP). +PTP model represents the model that combines the PoS classifier on the top of the baseline model. +PTP model is trained to predict masked tokens and PoS-tag of each token. +PTP+SCR model refers to the model that combines all pre-training approaches, masked-language model, PoS-tagging, and reconstruction of scrambled chunks. For QPAIR and Hate-speech tasks, the +PTP+SCR model achieves the best result among the baseline and +PTP models. In the rest of the tasks, such as KorNLI, NSMC, KorNER, and QS, the +PTP model outperforms the other models. In every task, the +PTP model outperforms the baseline model.

4.4. Experiment on Sequence to Sequence Model

We also conduct additional experiment on BART (Lewis et al., 2020) based encoder-decoder transformer model to verify the effectiveness of chunk-wise reconstruction. Detailed parameters and comparison of our model and Korean version of BART model (KoBart) is demonstrated at table 5

4.4.1. Pre-training for Encoder-Decoder Model

To verify the effectiveness of our proposed approaches in a low resource setting, we only take 100Mb, 700K sentences of Korean corpus from Korean Wikipedia [13]. We train our models on three different nosing methods, masked language model (MLM), permutation language model with MLM, and chunk-wise reconstruction with

[5] https://github.com/e9t/nsmc
[6] https://github.com/songys/Question_pair
[7] http://air.changwon.ac.kr/?page_id=10
[8] https://github.com/e9t/nsmc
[9] https://github.com/songys
[10] https://kin.naver.com/
[11] https://www.quora.com/
[12] https://stackoverflow.com/
[13] https://ko.wikipedia.org/

Model / Task	KorNLI	NSMC	QPAIR	KorNER	QS	Hate-speech (F1)
Basline	71.0%	86.12%	84.17%	80.9%	73.0%	63.0%
+PTP	**72.2%**	**88.02%**	86.28%	**81.75%**	**78.3%**	63.5%
+PTP+SCR	64.1%	86.14%	**87.2%**	76.23%	74.5%	**63.7%**

Table 4: Comparison of the baseline model, the proposed model with syntactic injection (+PTP), and PoS tagging prediction plus scrambled chunk-wise reconstruction (+PTP+SCR). The results are accuracy values except for Hate-speech that measured by F1 score. We bold the best results of each task.

Model/Parameter	KoBART	Our Model
Training Data	40Gb	100Mb
Number of Parameter	124M	16M
Encoder Layer	6	2
Decoder Layer	6	2
Hidden Size of Model	768	256

Table 5: Comparison of KoBART (Lewis et al., 2020) and our encoder-decoder model

MLM. We train the 3 models with the same training steps.

4.4.2. Test Data for Encoder-Decoder Model

We evaluate our models on NSMC and Q-Pairs datasets. Also, we add another downstream dataset to evaluate the generation ability of our models. Dacon Korean document generation and summarization AI competition dataset[14] is a newspaper article summarization dataset. It is an extractive summarization task built by selecting the top 3 significant sentences from the article. We use the same hyperparameters mentioned in Table 3.

From the test result explained in table 6, we can see that our proposed approach chunk-wise reconstruction with masked language model improves baseline in all of the test datasets, including sentence classification, sentence pair classification, and generation task.

5. Analysis

In this section, we provide results and explanations of experiments over different approaches to the proposed pre-training procedure.

5.1. Effect of Syntactic Injection

Experiment results show that our syntactic injection method improves the performance of all of the test datasets. The outcome represents that understanding the syntactical structure of language is critical to comprehend its semantic meaning. Especially in the QS task, our model outperforms the baseline by an absolute 5.3%. This result indicates that our scheme provides better understanding of sequence syntactic structure.

[14] https://dacon.io/competitions/official/235671

5.2. Effect of Scrambled Chunk-wise Reconstruction

In some tasks such as KorNLI and KorNER, +PTP+SCR model performs worse than the baseline model. One possible explanation is that chunk-wise reconstruction requires more training steps to converge than other approaches. During the pre-training procedures, +PTP+SCR model fails to converge. This result indicates that the permutation-based language model cannot guarantee performance improvement with limited computation costs. In fact, we train all of the models with the same steps.

On the other hand, Chunk-wise reconstruction may not generalize the feature for every task, especially token-level and inference tasks. Another possible explanation is the model structure. Typically, permutation language models are trained with auto-regressive sequence to sequence mechanism on the top of the encoder layers (Yang et al., 2019; Lewis et al., 2020). However, we implement our model based on the BERT (Devlin et al., 2018). Our experiment result with the sequence to sequence model shows scrambled chunk-wise reconstruction improves performance generally compared to the baseline MLM model and permuation language model. We notice that the permutation language model improves the performance in Korean language tasks, while it performs less in ones in English(Lewis et al., 2020). Thus, permuting sequence order is helpful to understand an order-free language like Korean.

However, a few pieces of evidence imply that the +PTP+SCR model has more robust features than the +PTP model. First, the +PTP+SCR model outperforms other models on the QPAIRs dataset that contains generated data by scrambling. For order-free languages such as Korean, it is easy to generate adversarial examples via scrambling (Zhang et al., 2019). This result shows that both proposed approaches improve robustness on the scrambling issue. In the sentence classification task for example, even though the Hate-speech classification is obviously more challenging than NSMC, the +PTP+SCR model slightly outperform other models. It indicates SCR is one of the promising ways to build a robust feature for order-free languages.

5.2.1. Scrambled sentence classification

We test our models, which is fine-tuned on the Q-Pairs dataset to classify scrambled sentence. The baseline and syntactic injection models (+PTP) fail to recognize

Model / Task	NSMC	QPAIR	Dacon (F1)	Dacon (Precision)	Dacon (Recall)
Basline	84.2%	79.65%	0.281	0.289	0.283
Permutation	84.6%	89.50%	0.301	0.313	0.305
SCR	85.6%	92.42%	0.328	0.343	0.336
KoBart	90.2%	94.34%	0.415	0.440	0.415

Table 6: Comparison of the baseline MLM model, permutation language model, and the proposed scrambled chunk reconstruction model. The results are accuracy values except for Dacon summarization dataset.

the scrambled sentence in many cases. See example 9 and example 10.

(9) 삶과 죽음의 순환은 계속된다
life-AND death-GEN circulation-TOP continue-PRS-DEC
'The circle of life and death continues'

(10) 계속된다 순환은 죽음과 삶의
continue-PRS-DEC circulation-TOP death-AND life-GEN
'The circle of life and death continues'

Both examples are equal sentences, but only the word-order of the example 10 is scrambled. Only the +PTP+SCR model predicts the label accurately. Nevertheless, this case indicates that scrambled chunk reconstruction provides meaningful pre-training tasks to attain generalized features for the Korean language.

5.2.2. Focusing on the end of the sentence, not the beginning

In the Korean language, the speaker's real intentions are normally revealed at the end of the sentence, not at the beginning in many cases (Grice, 1975) (Kim, 2012). From the table 7, we can see that +PTP+SCR model predicted difficult sentences correctly. Normally, our model fails to predict only ambiguous sentences or those sentences that do not provide enough information to classify the ratings by the context. Moreover, the +PTP+SCR model successfully classifies very confusing examples. See example 11, which is an example from the table 7.

(11) 이게 9점인가. 우리나라에서만 유독 호평받는 영화. -_- 네이버평점에 낚였다... 참고로 이거 외국사이트에서 평점보면 완전 낮다... 이래서 네이버평점 믿기가 싫다...
Is this worth 9 rate? The movie is particularly well-received only in Korea. -_- I was tricked by Naver's rating... FYI, the rating of this movie is totally low on foreign websites... This is the reason why I do not trust Naver's rating...

The correct translation of the sentence *"이게 9점인가. 우리나라에서만 유독 호평받는 영화"* is *"Is this worth 9 rate? The movie is particularly well-received only in Korea."*. However, if we translate the sentence directly, the translation can be *"This is the 9 rate. The movie is highly welcomed, especially in Korea."*. Although the beginning of the review may imply that the movie is good, the end of the sentence explains that the movie's reputation is over-rated in Korea. The reviewer discloses his real intention in the last sentence *"이래서 네이버평점 믿기가 싫다..."("This is the reason why I do not trust Naver's rating...")*. Like the review, many Korean articles and comments define its intention at the end of the context.
The example 12 clearly proves the robustness of our model.

(12) 설 특집으로 봐서 다행입니다. 영화관에서 봤다면 폭풍후회했을듯
It is good to watch this movie as the lunar new year special, if I'd seen it in the cinema, I must have regretted it.

The reviewer expresses that the movie is too bad that it is not worth paying for the movie's ticket. Hence, it is good to watch it without payment. However, both MLM and +PTP models predict the review as a positive review, only because it says *"It is good to watch this movie as the lunar new year special."* However, the true intention is demonstrated at the end of the review. "if I'd seen it in the cinema, I must have regretted it."

5.2.3. Most of the wrong predictions are acceptable

From the table 7, we notice that most of the wrong cases are acceptable. For instance, both of reviews " 내 마음의 평화도 좀 지켜주지 (Why don't you protect the peace of my mind.)", and "양심수라는올바른정의는누가내리는건가?? (Who decides he's a prisoner of conscience?)" are not relevant to the movie review. Those reviews only describe some scenes of the movie, not the user's sentiment. Even humans cannot distinguish the sentiment of the reviews from the context. We find that many wrong answers occur when the review is too short or impossible to classify the sentiment from the context. The NSMC dataset labelled the reviews not based on the context but the user's ratings. So, many reviews and ratings do not match.

(13) 재미있게 봤는데 마지막이 너무 엉성하네요 뭘 시사하는지 모르겠네요
'I enjoyed it, but the last scene was too sloppy. I don't understand what the message is.'

Sentence	MLM	+PTP	+PTP+SCR	Label
요사이 이 재미있는 영화는 없고 쓰잘대기없는 영화만 마너.. *Recently, there is no such funny movie, only useless movies*	0	0	1	1
참 이런명작이 없네. 누구나 겪었을법한 이야기에 세상더러운것까지. *What a masterpiece.* *A story anyone woud have experienced, and dirty world*	0	0	1	1
평점이 너무 걸레네.. 사무엘잭슨 OO연기만 봐도 지루하진 않은데 ㅋㅋ *The rating is too low...* *Even if you just watching the Samuel Jackson's OO acting,* *it is not boring at all tho lmao*	0	0	1	1
이게 9점인가. 우리나라에서만 유독 호평받는 영화. -_- 네이버평점에 낚였다... 참고로 이거 외국사이트에서 평점보면 완전 낮다... 이래서 네이버평점 믿기가 싫다... *Is this worth 9 rate?* *The movie is particularly well-received only in Korea.* *-_- I was tricked by Naver's rating...* *FYI, it's totally low on foreign websites...* *This is the reason why I do not trust Naver's rating...*	1	1	0	0
설 특집으로 봐서 다행입니다. 영화관에서 봤다면 폭풍후회했을듯 *It is good to watch this movie as the lunar new year special,* *if I'd seen it in the cinema, I must have regretted it.*	1	1	0	0
내 마음의 평화도 좀 지켜주지 *Why don't you protect the peace of my mind.*	1	1	0	1
양심수라는 올바른 정의는 누가 내리는 건가?? *Who decides he's a prisoner of conscience?*	0	0	1	0
재미있게 봤는데 마지막이 너무 엉성하네요. 뭘 시사하는지 모르겠네요 *I enjoyed it, but the last scene was too sloppy.* *I don't understand what the message is.*	1	1	0	1

Table 7: Prediction result on NSMC dataset. Italic sentences are the English translation of Korean sentences.

The review of the example 13 says the movie is good. The reviewer enjoyed it. However, at the end of the review, the user says the last scene was too bad and cannot understand what the movie implies. Based on the context, most of the humans will label the review as a negative review because the review sounds like *'The beginning scene was good, but the last scene ruined everything. So, it is a bad movie.*' However, the user rated the review with a high score. In most cases, the +PTP+SCR model fails to predict the sentiment of such reviews when the context and label are not matched. This analysis proves that our model has more generalized, robust, and reasonable representation and features than the baseline models.

6. Conclusion

In this paper, we present a novel pre-training strategy for the language model that specializes in the Korean language by leveraging its linguistical characteristic. Moreover, we conduct experiments on various datasets with multiple conditions. The experiments display the effectiveness of our proposed approaches. Also, our result proves that linguistical specialized pre-training methods can build better language model with lower resources. Our work provides meaningful results to the computational linguistic community of the Korean language and other language communities. We highly believe that our work can be utilized in similar agglutinative languages such as Japanese, Mongolian, Turkish, etc. For future work, we are interested in building unique transformer architecture for Korean and other agglutinative languages. Also, it will be meaningful to investigate how different PoS taggers affect the performance of our syntactic injection model.

7. Acknowledgement

This research is supported by Naver corporation as a part of internship program. The authors would like to appreciate Hongjoon Choi, Won Ik Cho and Kyung-Seo Ki for helpful discussion and support.

8. References

Ablimit, M., Neubig, G., Mimura, M., Mori, S., Kawahara, T., and Hamdulla, A. (2010). Uyghur

morpheme-based language models and asr. In *IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS*, pages 581–584.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Grice, H. P. (1975). Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Ham, J., Choe, Y. J., Park, K., Choi, I., and Soh, H. (2020). KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online, November. Association for Computational Linguistics.

Kim, Y. K. (2012). Inter-relationship between sentence structure (word order) and cultural structure : A case study in korean and english. *Journal of British American Studies*, 27.

Lee, S., Jang, H., Baik, Y., Park, S., and Shin, H. (2020). Kr-bert: A small-scale korean-specific language model. *ArXiv*, abs/2008.03979.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Moon, J., Cho, W. I., and Lee, J. (2020). BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online, July. Association for Computational Linguistics.

NA, S.-H. and KIM, Y.-K. (2018). Phrase-based statistical model for korean morpheme segmentation and pos tagging. *IEICE Transactions on Information and Systems*, E101.D(2):512–522.

Na, S.-H. (2015). Conditional random fields for korean morpheme segmentation and pos tagging. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 14(3), June.

Park, E. L. and Cho, S. (2014). Konlpy: Korean natural language processing in python.

Park, K., Lee, J., Jang, S., and Jung, D. (2020). An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China, December. Association for Computational Linguistics.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Roark, B., Saraclar, M., and Collins, M. (2007). Discriminative n-gram language modeling. *Computer Speech & Language*, 21(2):373–392.

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., and Cardie, C. (2019). Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yu, S., Kulkarni, N., Lee, H., and Kim, J. (2017). Syllable-level neural language model for agglutinative language. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 92–96, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zhang, Y., Baldridge, J., and He, L. (2019). Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.