Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5595–5600
Marseille, 20-25 June 2022

Developing Language Resources and NLP Tools for the North Korean Language

Arda Akdemir[†], **Yeojoo Jeon**[§], **Tetsuo Shibuya**[†]
[†]Institute of Medical Science, The University of Tokyo
[§] Graduate Schools for Law and Politics, The University of Tokyo
[†]{aakdemir, tshibuya}@hgc.jp
[§]jeon.yeojoo@gmail.com

Abstract

Since the division of Korea, the two Korean languages have diverged significantly over the last 70 years. However, due to the lack of linguistic source of the North Korean language, there is no DPRK-based language model. Consequently, scholars rely on the Korean language model by utilizing South Korean linguistic data. In this paper, we first present a large-scale dataset for the North Korean language. We use the dataset to train a BERT-based language model, DPRK-BERT. Second, we annotate a subset of this dataset for the sentiment analysis task. Finally, we compare the performance of different language models for masked language modeling and sentiment analysis tasks.

Keywords: North Korean, Language Model, Sentiment Analysis, Less-resourced languages, Language Resources

1. Introduction

It is now common sense in the Natural Language Processing (NLP) community that high-performing Machine Learning (ML) models require large amount of relevant datasets. However, such a large corpus is not available for many less-resourced languages. An extreme example is the Democratic People's Republic of Korea (DPRK) language. Due to the current circumstances, there are almost no digital resources available online for the DPRK Korean language. This resource scarcity makes it very difficult, if not impossible, to develop statistical models for the DPRK Korean language.

There are many publicly available labeled datasets for the Republic of Korea (ROK) language (Jang et al., 2013; Lim et al., 2019; Ham et al., 2020). Moreover, it is relatively easy to obtain large unlabeled corpora through web crawling, or Wikipedia dumps for the RoK Korean language. These resources have greatly fostered Korean NLP research. For example, there are various high-performing deep pre-trained language models (PLMs) for the ROK language such as KLUE-BERT (Park et al., 2021), and KR-BERT (Lee et al., 2020). These models are applied to various downstream tasks to achieve state-of-the-art results (Lee et al., 2020). Yet, the training data of these models do not include DPRK-specific text. Over the years, the Korean language used in the countries diverged significantly from each other. Table 1 highlights some of the important differences in the written Korean language between the two countries. [1] As a result of these differences, we can not directly apply the RoK PLMs to analyze the DPRK data.

The current NLP research on the DPRK language is highly limited in volume and scope because of a lack of NLP tools and resources. Most NLP research on the language tries to mitigate this issue either by relying on domain experts' manual readings of the DPRK articles or by only focusing on the English articles provided by the KCNA(Korean Central News Agency). The use of NLP methods in all these studies is limited to very conventional keyword-based approaches.

Increasing the available NLP tools and resources, and developing a high-quality DPRK language model is essential for better interpretations of the DPRK messages and prevent misperceptions of the DPRK's political agenda. Besides, there is an increasing amount of the DPRK's refugees worldwide and especially in the ROK (Green and Epstein, 2013). A genuinely inclusive AI requires developing tools for all communities regardless of its difficulty due to the scarcity of resource.

In this paper, we make the following contributions:

1. We present a large-scale corpus for the DPRK language.

2. We train a DPRK deep language model by fine-tuning the Korean KR-BERT model on a large unlabeled DPRK dataset, and show improvements over other PLMs.

3. We present our on-going effort annotating the DPRK corpus for the sentiment analysis task.

4. We train several PLMs and provide preliminary results on this sentiment analysis dataset.

2. A DPRK Language Corpus

There are no publicly available NLP datasets specifically for the DPRK language. In this section, we explain the dataset we compiled for the DPRK language.

[1] These differences are obtained fromKwon Jae Il (2015).

Description	DPRK	ROK	English
The middle ㅅ rule	나무잎	나뭇잎	leaf
Initial sound rule	로동	노동	labor
Different word usage	동무	동무	revolutionary comrade (DPRK), friend (the ROK)
DPRK-specific ideological terms	로동영웅	-	labor hero
White-space usage difference	갈바를_알수_없다	갈_바를_알_수_없다	don't know how to go

Table 1: Examples of distinctive differences between the two Korean Languages

2.1. Rodong News articles

The largest North Korean online data provider is the Korean Central News Agency (KCNA). KCNA is the only news agency in North Korea and provides content for both domestic and foreign audience (Shrivastava, 2007). KCNA releases daily news articles under Rodong Sinmun (Worker's Newspaper in English) toward its domestic audience. Every day, around 30 news articles from Rodong are made available online under the official website. [2] The online version continues from the beginning of 2018 and there are approximately 50 thousand articles including .

2.2. Web scraper for Rodong News

The official Rodong website supports searching. However, there are no mechanism to download the news articles in text format. Considering the large size of the database, this makes it practically impossible to obtain the whole dataset manually. For this reason, we developed a web-scraper for automatically parsing the official Rodong website to download all news articles in text format.

The web-scraper first downloads the raw html documents. This is followed by removing the markup and the boilerplate text. The Rodong website contains several different article types, each having a slightly different format. For example, these articles occasionally contain letters from/to leaders of other countries. Our web-scraper is designed to handle all different article types. We ignored the articles that only consist of a header and an image as the text portions are too short. The web-scraper stores each article in a system-friendly JSON format together with all its metadata. An example article inside the dataset looks as follows:

```
{"articles": [
      {"id": "2020-04-20-0002",
       "date": "2020-04-20",
       "year": 2020,
       "title": <Title redaccted>,
       "source": "Rodong",
       "type": "news_article",
       "data": <Text redacted>},
   ...]
  }
```

As online access to the news articles does not require any permission, the web-scraper tool can be used to obtain this dataset. [3] We recommend using the software, as the data keeps growing every day with the forthcoming news articles.

2.3. New Year Addresses

The second source we utilized is the Inaugural New Year Addresses of the Supreme Leader(s). Each year, the Supreme Leader of DPRK addresses its own citizens and these speeches contain valuable information about the political intentions and the agenda of Pyongyang (Park et al., 2015). The new year addresses are available for every year from 1946 to 2019 except 1957, when it was not released due to its domestic political issues (Yonhap News Agency, 2021). We used the annual addresses in text format [4]. More details about each dataset are given in Table 2.

We use these new year addresses in two main ways:

- As an out-of-domain test set for evaluating the performance of the trained DPRK language model.

- As an additional source for the sentiment analysis annotation.

for the construction of the sentiment analysis data that will be explained in Section 3. Next, we explain these addresses in more detail.

[2] http://www.rodong.rep.kp/ko/index.php

[3] Note that sharing it online as an open-source may have some legal issues. We will decide how to make this sharing process possible in consulting with the relevant parties.

[4] For data formation, we follow the criteria provided in Part et al. (Park et al., 2015) By referring to (pea, 1997), they coded 1) messages of congratulations, 2) speeches, 3) New Year's editorials, 4) joint editorials of the three major newspapers (Rodong Sinmun(로동신문), Joson Inmingun(조선인민군), and Chongnyon Jonwi(청년전위)) as new year addresses. Our dataset includes Kim Il-sung's New Year's Address (1946-1950, 1954-1956, 1957-1965, 1969, 1971-1994), the New Year's Editorial by the Rodong Sinmun during Kim Il-sung's regime (1966-1968, 1970), and New Year's Messages to the North Korean Army (1951-1953), New Year's Editorial during Kim Jong-il's regime (1995-2011), Kim Jong-un's New Year's Editorial (2012), and oral New Year's Address (2013-2019). (new, 1946 2019); The original texts of the North Korean New Year's addresses were quoted from (Il-Sung, 1979 1994) and (KCNA,). ; The DPRK has not been issuing its annual new year addresses since 2020 but substituting them with the Report of the Plenary Meeting of the Central Committee of the Workers party and a letter handwritten by Kim Jong-un. (Yonhap News Agency, 2021)

	New Year Speeches	Rodong Sinmun
Number of Documents	73	27,401
Number of Sentences	5,709	471,417
Date Span	1946-2019	2018-now

Table 2: Details about each DPRK dataset

3. Sentiment Analysis Dataset

States establish and change foreign policy predicated on their intentions. Therefore, it has been a critical task for policy-makers to understand the true intentions of other countries and predict the direction of their policy decisions correctly to establish pertinent foreign policy. Though understanding the intentions is a tricky job, deciphering the signals of a state can provide us with a good source for it.

Especially, in the case of autocracies, it's more difficult to understand the intentions due to the lack of information and secrecy of the decision-making process. However, the sentiment analysis on the leaders' speeches and propagandistic news reports can offer a good means for interpreting the intentions because the authoritarian countries make the speeches and news reports under the guideline of the regime that is affected by the leaders' thinking which is critical for its decision-making. Because we can assume that the sentiments in their official addresses and newspapers implicitly reveal the leaders' intentions on policy, we may be able to predict their next step. For example, when a country expresses negative sentiments more often than positive, the country is more likely to take negative policy afterward.

We aim to start a new attempt to understand DPRK's sentiments with the automated text-analysis method utilizing the DPRK language model. We expect this sentiment analysis will become the first step to enable this line of work.

3.1. Annotation Method

In this research, we assign labels only at the sentence level. We annotate sentences as either positive or negative. To annotate sentences, we refer to the standards were provided by Park et al. (2015). They categorized sentiments in DPRK's new years' addresses and mapped keywords into neutral and negative words [5]. After we annotate sentences once, we doubly check whether the annotations are consistent with the criteria of Park et al. (2015).

Positive class. We use the term positive to refer to any sentence that expresses or implies a positive sentiment, regardless of its subjectivity. An example of a positive sentence can be found in Table 3. The example contains the positive evaluation signaled by the adjective, "위대한(great)".

Negative class. Likewise, we use the term negative to refer to any sentence that expresses or implies a negative sentiment, regardless of its subjectivity. An example of a negative sentence can also be found in Table 3. The sentence contains the negative evaluation signaled by the ending of the sentence, "오산이다(is a misconception)".

3.2. Annotation Procedure

Annotation was conducted using Doccano tool [6] and we were able to check all the annotations' reviews simultaneously. An annotator who was trained to read DPRK texts annotated sentences and all the sentences were doubly annotated[7]. At this stage, to prevent wrong annotations, we only annotated sentences that explicitly contain negative or positive sentences. Ambiguous sentences were removed during the process. Table 4 gives the statistics about our on-going annotation process. We applied a random 60/40 splitting for constructing the training and test splits. [8]

4. Experiments and Results

4.1. Language Modeling

In this section, we present the experiments and the results for the masked language modeling task.

4.1.1. Methodology

We used **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT in short) (Devlin et al., 2019) as the deep language model. BERT is a bidirectional transformer encoder that is trained using the masked language model (MLM) and the next sentence prediction (NSP) tasks. BERT's architecture is identical to the Transformer-encoder architecture (Vaswani et al., 2017). BERT uses a WordPiece (Wu et al., 2016) tokenizer to represent an input sequence as a list of subword tokens.

Following the previous work, we used the masked language modeling (MLM) for pre-training the BERT model on the unlabeled Rodong News articles dataset. In the MLM task, the model is given a sentence where certain words are replaced with a `[MASK]` token and the model tries to predict the correct words from its vocabulary $\mathcal{V}$ of size $\mathcal{M}$.

Deep learning models such as BERT, require large amount of data in order obtain meaningful representations as they have billions of parameters to tune. The

[5] Because they performed the sentiment analysis mainly about the US and the ROK, they didn't capture positive sentiments of DPRK. However, because our research doesn't limit the scope of topics to the adversary-related ones, we can provide positive sentences.

[6] https://doccano.github.io/doccano/

[7] The annotation process has been mainly done by the Korean first author alone, since it requires the quality of linguistic efforts rather than the quantity of human power, and the good understanding of both South and North Korean languages.

[8] After the initial submission, we have extended this dataset to 600 sentences and balanced the negative and the positive sentences. We share this updated version as a language resource in our final submission.

Class	Example
positive	김정일동지는 우리 조국과 민족의 위대한 수호자이시다 (Comrade Kim Jong Il is a great guardian of our homeland and people.)
negative	원래 계급적원쑤들에게서 그 어떤 관용이나 인간성을 기대하는것자체가 오산이다. (Expecting any tolerance or humanity from class enemies has always been a misconception by itself.)

Table 3: Example sentences for each sentiment class.

	Total Number	Positive	Negative
Train	302	209	93
Test	202	76	126
Total	504	285	219

Table 4: Details about the sentiment analysis annotations.

Rodong News article dataset does not have enough coverage to learn a language model from scratch. For this reason, instead of training the BERT model from scratch, we initialized the model weights using the state-of-the-art South Korean LM, KR-BERT (Lee et al., 2020). In a way, we continue the pre-training of the KR-BERT model, on the DPRK-specific data.

4.1.2. Results

In this section we report the results obtained for the masked language modeling task. First, we report results on the Rodong test set. Following that, we evaluate the proposed methods on the New Year Addresses. In the following text, DPRK-BERT refers to the DPRK LM we trained.

We compare our approach with three other deep learning models:

1. KR-BERT (Lee et al., 2020): It is the state-of-the-art LM for the South Korean Language.

2. KR-BERT-MEDIUM: KR-BERT's extended version which is trained using, in addition to the original dataset, legal texts crawled from the National Law Information Center and a Korean comment dataset. KR-BERT-MEDIUM is shown to outperform KR-BERT on the sentiment analysis task. [9]

3. mBERT (Devlin et al., 2019): The multilingual BERT model trained on Wikipedia articles of 102 languages. Korean is also included in the training set however it is much less represented compared to the English and several other more resourced languages.

One might argue that evaluating different language models with varying vocabulary sizes is not a fair comparison. However, during our manual analysis, we observed that the average number of non-zero probability scores is independent of the vocabulary size and almost the same for all language models. In other words, all models consider only a small number of possible candidates for the masked tokens. Besides, except for the mBERT model, the vocabulary sizes for all three other models are also similar.

We used *perplexity* and *MLM accuracy* metrics to evaluate the performance of each model. *MLM accuracy* is the rate of the correctly predicted masked tokens, the higher the better. For *perplexity*, a lower score denotes a better language model. We repeated evaluation on each dataset three times as evaluation involves randomness (tokens are masked randomly).

First we present the result on the Rodong test set. We see that the multilingual BERT model performs significantly worse than all other PLMs (3.410 average log perplexity and 37.730% average MLM accuracy). This supports the previous research that illustrates the weaknesses of multilingual LMs in low-resource settings (Pires et al., 2019; Virtanen et al., 2019; Lee et al., 2020). However it must be noted that the vocabulary size of mBERT tokenizer is significantly larger than other LMs. A larger tokenizer vocabulary makes it more difficult to find the correct masked token, so mBERT has a disadvantage over other models. More importantly, we see that the DPRK-BERT model significantly outperforms all other approaches, 1.702 improvement on perplexity and 28.8% improvement on accuracy over the closest model (KR-BERT). DPRK-BERT achieves an average of 82.37% MLM accuracy.

Next we discuss to results on the New Year test set. Interestingly, we see that the performance of all compared models increase when evaluated on New Year Speeches. For example, the average MLM accuracy of mBERT increases by 7.55% (from 37.73% to 45.28%). This strongly suggests that *the use of Korean language in New Year speeches is closer to RoK Korean compared to Rodong news articles*. Conversely, *Rodong news articles have a much more tailored and unique use of the Korean language*. Similar to the Rodong results, we observe that the DPRK-BERT model significantly outperforms all other approaches. 4.49 improvement on perplexity and 16.72% improvement on accuracy over the closest model (KR-BERT). We see a performance drop for DPRK-BERT when evaluated on the New Year Speeches, 5.72% drop in average MLM accuracy and 0.74 increase in average perplexity. This is expected, as the DPRK-BERT is trained over the training split (80% of all dataset) of the Rodong news articles. Even though the training and validation splits are disjoint, the validation split naturally has a more similar data distribution to the training than New Year

[9] https://github.com/snunlp/KR-BERT-MEDIUM

	Rodong		New Year	
Model	Perplexity	Accuracy	Perplexity	Accuracy
mBERT	3.41	37.734	2.853	45.281
KR-BERT-MEDIUM	3.241	43.129	2.922	46.32
KR-BERT	2.505	53.57	2.009	59.933
DPRK-BERT	0.803	82.367	1.088	76.652

Table 5: MLM Results for all LMs on both datasets.

	F1	Acc
mBERT	90.7	88.6
KR-BERT	92.5	90.6
DPRK-BERT	92.4	90.6

Table 6: Sentiment analysis results.

speeches.

4.2. Sentiment Analysis

Next, we report the preliminary results for sentiment analysis on the dataset we annotated. After removing the ambiguous cases, we obtained a total of 504 sentences with sentiment annotations. We applied a 60/40 random split, which gave 302 and 202 training and test sentences, respectively.

As the sentiment analyzer model, we extended the vanilla BERT architecture with a sentence classification head whose weights are initialized randomly before fine-tuning. BERT weights are initialized using the pre-trained LMs explained above. During fine-tuning, we did not freeze the BERT weights.

Table 6 gives the results on the test split. Overall, we see that all three models achieved a significantly high performance on the test set. All models achieved a performance around 90% accuracy. As expected, we see that the multi-lingual BERT model performed slightly worse than the mono-lingual models. Surprisingly, we see that on this test set, DPRK-BERT model did not outperform the KR-BERT model in terms of accuracy or F1 score. Both models achieved almost identical performance.

5. Conclusion

In this paper, we presented a large-scale DPRK corpus by using Rodong News articles and New Year Addresses of the North Korean leaders. Following that, we presented our on-going attempt of building a sentiment analysis dataset using this corpus. Additionally, we provide several NLP tools for the DPRK Korean language such as a web-scraping tool for the Rodong News website, a BERT-based language model for the DPRK language, and a sentiment analyzer. Finally, we compare the proposed DPRK-BERT model with other Korean PLMs and show that pre-training on North Korean text brings significant performance improvements, especially in terms of masked language modeling accuracy.

Automated retrieval of sentiment from DPRK texts is an important task with several implications for social scientists and policy makers. In this paper, we presented our on-going annotation process and preliminary results for the sentiment analysis task where the performance of all language models was similar. We suspect that this results from our initial annotation process, which had a data imbalance and several duplicate sentences. We resolved these issues and increased the labeled sentences to 600. Besides, we only considered positive and negative sentiment classes in this preliminary step. A critical future research direction is to annotate a large-scale sentiment analysis with more fine-grained sentiment classes. We will expand the sentiment types to capture unique characteristics of the language used in the DPRK texts and annotate a much larger dataset that will enable better generalization on unseen DPRK sentences.

Currently, researchers and policy makers have to manually review new coming DPRK-related content to understand the policy changes of the DPRK government. Another important direction is to develop a web-based sentiment analyzer that will continuously process new data from DPRK sources. Such a tool will significantly help reduce the manual work of researchers for analyzing DPRK texts. After the annotation of the large-scale sentiment analysis dataset, we will implement a web-based sentiment analyzer that is trained on this dataset to process new coming content everyday.

6. Acknowledgements

This work is partially supported by JSPS KAKENHI Grant Number 20H05967, 20K21827, and 21H05052. We express our gratitude towards Dong-Joon Jo and Byoung-Lo Philo Kim for their valuable information and suggestions.

7. Bibliographical References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Green, C. K. and Epstein, S. J. (2013). Now on my way to meet who? south korean television, north korean refugees, and the dilemmas of representation. *Japan Focus: The Asia-Pacific Journal*, 11(41/2).

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 422–430.

Il-Sung, K. (1979–1994). *Kim Il-Sung Chojakchip(Collected Works of Kim Il-Sung)*. Pyongyang: Chosun Rodong-dang chulpansa(publishing company).

Jang, H., Kim, M., and Shin, H. (2013). Kosac: A full-fledged korean sentiment analysis corpus. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 366–373.

KCNA.). Rodong sinmun.

Kwon Jae Il. (2015). Unifying the vocabulary of the ROK and DPRK. *New Korean life*, 25(4):107–124.

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

Lim, S., Kim, M., and Lee, J. (2019). Korquad1. 0: Korean qa dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

(1946–2019). the original texts of the dprk's new year addresses.

Park, J. H., Park, E., and Jo, D.-J. (2015). Automated text analysis of north korean new year addresses, 1946–2015. *Korean Political Sci. Rev*, 49:27–61.

Park, S., Kim, S., Moon, J., Cho, W. I., Cho, K., Han, J., Park, J., Song, C., Kim, J., Song, Y., et al. (2021). Klue: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*. Advances in Neural Information Processing Systems.

(1997). Unification Problem Research. chapter Appendix: A Recent DPRK's Joint New Year's Editorial and Analysis, pages 299–349. The Institute for Peace Affairs.

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.

Shrivastava, K. (2007). *News agencies from pigeon to internet*. Sterling Publishers Pvt. Ltd.

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.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yonhap News Agency. (2021). Kim Jong-un, skip the New Year's address this year again… Reveal a handwritten New Year's card only (comprehensive report 2).