JOIV : Int. J. Inform. Visualization, 6(2-2): A New Frontier in Informatics - August 2022 587-592

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Verification of a Dataset for Korean Machine Reading Comprehension with Numerical Discrete Reasoning over Paragraphs

Gyeongmin Kim [a], Jaechoon Jo [b,*]

[a] *Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea*
[b] *Division of Computer Engineering, Hanshin University, Osan 18101, Republic of Korea*
*Corresponding author: *jaechoon@hs.ac.kr*

***Abstract*— Numerical reasoning in machine reading comprehension (MRC) has demonstrated significant performance improvements in the past few years. However, due to the process being restricted to specific languages, low-resource languages are not considered, and MRC studies on such languages are limited. In addition, the methods that rely on existing information extracted within the span of a paragraph have limitations in responding to questions requiring actual reasoning. To overcome these shortcomings, this study establishes a dataset for learning Korean Question and Answering (QA) models that not only answer within the span of passages but also perform numerical reasoning on passages and questions. Its efficacy was verified by training the model. We recruited eight annotators to tag the ground truth label, and they annotated datasets with 920, 115, and 115 passages in the train, dev, and test, respectively. A simple yet sophisticated automatic inter-annotation tool was created by effectively reducing the possibility of inaccuracy and error entailed by humans in the data construction process. This tool used common KoBERT and KoELECTRA. We defined four general conditions, and six conditions humans must inspect and fine-tune the pre-trained language models with numerically aware architecture. The KoELECTRA and NumNet+ with KoELECTRA were fine-tuned, and experiments in identical hyperparameter settings showed that compared with other models, the performance of NumNet+ with KoELECTRA was higher by more than 1.3 points. Our research contributes to the Korean MRC research and suggests potential and insight into MRC models capable of numerical reasoning.**

***Keywords*— Machine reading comprehension; numerical reasoning; language model; ELECTRA; low-resource language.**

Manuscript received 10 Oct. 2021; revised 23 Jan. 2022; accepted 18 Mar. 2022. Date of publication 31 Aug. 2022.

I. INTRODUCTION

Machine reading comprehension (MRC) trains the machine to read and answer questions with understanding the given text passages, which is a fundamental task in natural language processing (NLP). When the passage and question are given in MRC, the model should understand these attributes to find answers [1], [2]. With significant progress in reading comprehension tasks, pre-trained language models (PLMs), which are trained with the most popular dataset: Stanford Question Answering Dataset [3] and the Korean question answering dataset for machine comprehension of the Korean language [4], already outperform human performance [5], [6]. These two benchmark datasets are generally utilized as objective indicators to assess the model performance in reading comprehension. Recently, transformer [7] based models with pre-training and fine-tuning approach achieved dramatic performance in various downstream tasks, such as question answering [8], named entity recognition [9], text generation [10], and other NLP tasks [11]. Pre-trained language models (PLMs) such as Bidirectional Encoder Representations from Transformers (BERT) and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) are trained with large-scale text corpus with unsupervised objectives functions like masked language modeling, next sentence prediction, and replaced token detection (RTD). The PLMs can provide contextual features for better context-sensitive information within the passage and question. Multilingual PLMs such as mBERT for training models in Korean can be considered various forms of linguistic information. However, they have limitations in that Korean-specific data is not sufficient.

The conventional methods can extract spans from passages and questions and answer the questions being present in passages and questions. This means that there have been tremendous advances in the NLP field. However, more realistic and advanced answers that derive results through numerical reasoning by extracting significant numerical information in passages and questions, as in the reading comprehension benchmark dataset, Discrete Reasoning Over the content of Paragraphs (DROP) [12], are sought in real Question and Answering (QA), instead of simply asking for spans based on the passages and questions. For an optimal answer to a question such as "How many days after Game 5 will Game 6 happen?", a more realistic QA is required since it is necessary to infer with numerical expressions rather than simply understanding

For example, if we take a passage, a human can predict that this MRC requires more realistic QA as the correct answer to questions such as "How many days after Game 5 will Game 6 happen?"; the answer needs to not only be obtained simply through the evidence in the passages and question, but it has to be deduced through numerical reasoning.

> *"Victory in Game 5, and the 82% chance to win was secured. The* NC *Dinos beat the Doosan Bears in Game 5 of the 2020 Korean Series of the KBO League on the 23rd at the Seoul Gocheok Sky Dome ... The NC Dinos and Doosan announced Rucinski and Alcántara as their respective starting pitchers for Game 6 to be held on the 24th."*

This study follows NumNet [13] and NAQANet [14] structures directly motivated by DROP, constructed numerical reasoning data, and enabled numerical reasoning by implementing graphs into the model through symbolic reasoning beyond understanding the passages. In addition, other related works on this topic were proposed for DROP [15]–[17]. There is an English dataset, which selects necessary elements within the passage to answer a question fundamentally, that piles additional complexity to reading comprehension by tracking conversation status [18], demanding passage search [19], integrating external knowledge [20], and conducting multi-hop reasoning [21]. DROP differs in that it is a dataset built to perform numerical reasoning through the semantic understanding of the passage and a comprehensive understanding of the passage.

However, unlike certain languages in which data can be provided to models to enable them to learn and acquire abilities to understand complex contexts and numerical information, low-resource languages (LRLs) still lack data resources for learning [22]. Various groundbreaking studies [23] wherein DROP was translated to Korean to develop a Korean numerical reasoning language model and thus try to solve the three linguistic problems of the domain, spacing, and named entity have been carried out. However, these attempts suffered limitations in creating an optimal language model due to issues related to insufficient learning data volume and linguistic characteristics.

LRLs can be considered as languages relatively understudied in the NLP field. Transfer learning–based studies have been recently conducted for these LRLs, but further, more advanced research is required to provide more practical solutions. Therefore, in this study, we intend to establish a Korean dataset to create a QA model that can perform actual numerical reasoning in MRC and verify the effectiveness of the QA model. Thus, we aim to establish a dataset that can perform Korean numerical operations and prove its efficacy by learning with a model architecture that can perform numerical reasoning based on the pre-trained Korean language model.

The contributions are as follows:

- We collect numerally aware Korean MRC datasets from historical literature and Korean baseball news and recruit eight annotators to construct the dataset.
- We analyze the linguistic characteristics of Korean, which is a low-resource language, and describe the reasons for the necessity of the numerical Korean MRC dataset
- Finally, this corpus's effectiveness was constructed with Korean deep learning models.

The structure of this paper is as follows. Section 2 describes the data construction method and modeling. Section 3 details the experimental results and analysis of utilizing the language model based on the created data. Finally, in Section 4 we present our conclusions.

II. MATERIALS AND METHOD

This section explains the Korean numerical reasoning dataset's annotation process and fine-tune the Korean language model.

A. Korean Numerical Data Description and Annotation Procedure

TABLE I
THE OVERALL DATASET STATISTICS USED IN TRAINING. DEV & TEST WERE GENERATED MORE THAN TRAIN QA PAIRS ON AVERAGE.

	Train	Dev & Test
Number of Passages	920	230
Number of Questions	13471	3898
Number (72.01%)	9701	
Date (0.69%)	93	
Span (26.30%)	3677	
Avg. questions / passage	14.64	16.95

1) Numerical data Collection: In this study, we crawled the passages of the KBO News[1] data; this required numerical reasoning and the historical data of the Encyclopedia of Korean Culture[2]. Six hundred paragraphs for each domain in the overall data containing numbers in their passages were randomly extracted, and a total of 1,200 paragraphs were divided into Train, Dev, and Test sets in a ratio of 8:1:1. Table 1 shows the number of passages and questions, type of questions, and average on questions/passages the questions in the training dataset consist of numbers (72.01%), date (0.69%), and span (27.30%). Moreover, the questions were not classified for dev and test. The average questions/passage for the train was 14.64, and 16.95 for Dev and Test datasets.

2) Annotation Guidelines: We recruited eight annotators to annotate the data tagging task with guidelines and trained them periodically with feedback. Each annotator was equally allocated half of the baseball and historical dataset. The corpus was optimally composed in Korean based on DROP. It was required that the questions must be directly identifiable in the passage or deducible. If they are not, the span corresponding to the correct answer must be specifically identified. For non-text answer cases, each annotator must

[1] https://www.koreabaseball.com

[2] http://encykorea.aks.ac.kr

input the number if the answer is a numerical expression and to date if the answer is a date. First, all annotators were trained for a week, and each received 20 test samples to familiarize themselves with and perform the annotator tagging guideline. The annotator team discussed the work of each member on the annotation procedure during a weekly review. The eight annotators were divided into two groups, and all QA pairs were reviewed by four opposite cross-checking annotators and further reviewed by two expert supervisors. The set size annotated by annotators is equally divided by the category to which the raw text belongs. The discrepancy of the annotators on the annotated QA pairs was also fully discussed and matched within the execution period. These data-tagging procedures can enhance the reliability and value of the training data.

3) Automatic Inter-annotation Process: Data construction is generally a painstaking and costly process. We utilized automatic inter-annotation tools to build more efficient data in the process where annotators create questions by understanding the passage, especially when they are creating answers related to the span. Our automatic inter-annotation tool has the advantage of reducing cost by making the language models' performance similar to or exceed human performance and improving the quality of datasets that identify the intent of the annotators.

Fig. 1 shows that when the annotators create a dataset corresponding to the span type, the annotators must create a QA pair according to the given passage. Here, the language model evaluates whether the answer is already present in the given passage. The tool informs the annotator when the language model cannot provide an answer to the generated question and induces the question to be revised to the relevant passage if necessary. For example, an annotator wants to make the question "Who scored the first touchdown of the game?", which is related to the passage. Then, the question is fed into the inter-annotation tool to answer the question. In the feedback process, if the first answer is *Anna* from the BERT and the second answer is *Tom Brady* from the ELECTRA, the annotator should check whether the answer exists in the passage and then save it in the database.

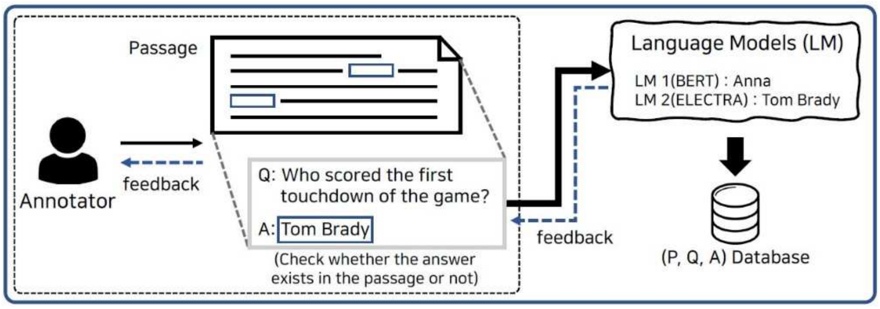

Fig. 1 Overall process of constructing automatic inter-annotation tool. Annotators reasonably determine whether the final P, Q, A save in database or not.

This is a simple yet sophisticated extra-automated tool to effectively reduce the possibility of inaccuracy and error entailed by humans and enables a high-quality dataset to be created by overcoming the existing shortcomings.

4) Differences from DROP dataset:

Domain: DROP is limited to history and American football topics and includes many English American figures and locations. Because researchers of other nationalities will be unfamiliar with these names, using DROP could lead to misunderstanding. For example, QB, which stands for the quarterback, refers to a position in American football. The term is alien to an individual who has no interest in and no knowledge about American football. Thus, in this study, training data used for learning were developed by excluding words that can lead to misunderstanding.

Linguistic characteristics: The Korean language is morphologically rich, which is challenging due to the intermediate characteristics positioned between isolating and inflectional language [23]. The word in the Korean language is composed of *eojeol* and grammatical morpheme. The first challenge is that the meaning of the *eojeol*, a white space unit, can vary depending on the grammatical morpheme behind the *eojeol*. For example, the pronoun '나 (I)' can be changed by postpending a postposition to it, such as '나의 (my)', '나랑 (with me)', or '나도 (me, too)'. Furthermore, the meaning of the verb '-이다 (be)' can be changed by adding suffixes to it, as in the case of '-오다 (come)' or '-가다 (go)'.

TABLE II
TRANSFORMABLE WORD TOKEN BY JAMO SUB-CHARACTER COMPOSITION. THE BOLD TEXT CHARACTER DESCRIBES EACH TRANSFORMED SUB-CHARACTER.

Original word	Text transformation
새(bird) ['ㅅ', 'ㅐ']	개(dog), ['ㄱ', 'ㅐ'] Initial consonant
	소(cow), ['ㅅ', 'ㅗ'] Vowel
	색(color), ['ㅅ', 'ㅐ', 'ㄱ'] Final consonant

The second challenge is that a *jamo* which is consisted of a consonant and vowel, can be represented differently depending on the meaning of the morpheme. In Table 2 shows the flexibility of *jamo* changing the meaning of the morpheme completely. In the first line, '새(bird)' composed of 'ㅅ' and 'ㅐ' is transformed in '개' which means 'dog' in Korean due to the variation of the initial consonant.

In the second line, the initial consonant 'ㅅ' remains fixed, and the vowel changes from 'ㅐ' to 'ㅗ', which changes the meaning to '소(cow)'. Finally, in the third line, a final consonant 'ㄱ' is added to the original word '새' and the meaning of the word is changed from '새(bird)' to '색(color)'. Therefore, the dataset was created by considering these factors during the process of answering the span type.

B. Constraints in Data Construction Process

It is extremely crucial to establish constraints in the data construction process. The following can be comprehensively checked in the source:

- Whether blank values exist in the answer type (number, span, date)
- Whether suitable information relevant to the number and span type was appropriately input into the answer
- Whether suitable numbers were input
- Whether there are any unnecessary blanks.

However, the following constraints must be manually inspected:

- All spans and dates must exist in the passages as text.
- Up to two decimal places are input for number types, and integers are input without decimal places.
- Reasoning with information other than the text is not considered as it is outside the span of data construction. An emphasis is placed on numerical reasoning.
- The generated answer types are classified into subtraction, comparison, selection, addition, count and sort, and other.
- Instead of simply having the correct answer to the passage and question, it is necessary to state in numbers whether the answer was generated through specific reasoning and whether the answer is correct.
- When the answer includes a unit, it must be included in the question to reduce the ambiguity of the question.

C. Numerical Reasoning Language Model

The Korean language model used in this study is KoELECTRA [24], [25], a language model pre-trained on big data. ELECTRA trains two neural networks, Generator G and Discriminator D, to perform RTD. Both networks are intrinsically composed of transformer encoders [7], and training occurs by receiving a sequence of tokens as input and mapping them as hidden sequences that reflect contextual information.

Our model follows the typical deep learning architecture flow of previous numerically aware graph neural network models, NumNet, which expresses passages and questions in the encoding module as graphs, and NumNet+, which is an enhanced version of NumNet for training numerical reasoning language model. It should be considered comparing information and conducting numerical reasoning, which is composed of embedding, encoding, attention between passage and question, and output layers.

To predict an answer probability in the passage, the model utilizes cross-attention embeddings to encode the question and passage with PLMs. Then directed graph with nodes corresponding to the numbers in the passage and question, and edges to encode numerical relationships in the numbers. The model should output the three types of answers in the prediction stage. These mean spans, count, and arithmetic expressions (AE). Especially, reasoning skills for AE can represent addition, subtraction, counting, and sorting from numbers. Compared with previous studies proposed for arithmetic word problems, this architecture with PLM has proved better natural language understanding ability.

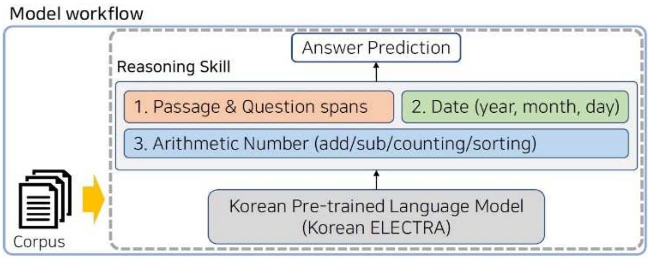

Fig 2. Workflow of the language model capable of Korean numerical reasoning. NumNet and NABERT structures were used to fine-tune into the Korean

Another model we consider is numerically aware BERT (NABERT), which uses a pre-defined numerical operator as a semantic tag to find the most optimal operator. As shown in Fig. 2, the pre-trained Korean language model learns reasoning skills that enable numerical reasoning to optimize the probability value of finding the most superlative answer to the input sentence. Subsequently, passages and questions are used to predict the best answer prediction.

III. Results and Discussion

A. Experimental Setup

The experimental environment is as follows. In this study, we constructed a dataset that is 1/6 the size of DROP and trained the PLMs to verify the effectiveness of the MRC model capable of numerical reasoning in Korean, which is a low-resource language. Table 3 shows the hyperparameter setting used in this study. An Adam optimizer with a learning rate 1e-5, 10 max epochs, 512 max sequence length, and 16 batch size was used in a GPU RTX 8000 * 2 environment. In the experiment process, we repetitively measure the performance of these models with 5 random seeds and record the average value of the measured results as the final performance of the models. We also leverage a 5-fold cross-validation method in the training process, which trains the models with various training split settings.

TABLE III
HYPERPARAMETER SETTING FOR FINETUNING PROCESS WITH KOELECTRA

Hyperparameters	**KoELECTRA**
Learning rate	1e-5
Optimizer	Adam
Max epochs	10
Max sequence length	512
Weight decay	0.1
Batch size	16

B. Evaluation Method

For evaluating the effectiveness of our proposed method, we use exact match (EM), which measures predictions that match any one of the ground truths answer exactly the same, and F1 score, which is a harmonic value of Precision P and Recall R, as a standard indicator to improve on weakness caused by using only accuracy for evaluation. The F1 score is calculated as follows:

$$P = \frac{|C|}{|E_p|}, R = \frac{|C|}{|E_r|}$$

$$F1\ score = \left(\frac{R^{-1} + P^{-1}}{2}\right)^{-1} = 2 \cdot \frac{P \cdot R}{P + R}$$

C. Experimental Results

Table 4 shows the experimental results of our models that trained the Korean numerical reasoning data. The following three primary model structures were used: the fundamental NumNet applied with Korean Glove-100d; NAKoELECTRA, which is KoELECTRA leveraged with the model structure of NAQANET [12], [14]; and NumNet+ exploited with KoELECTRA.

TABLE IV
EXPERIMENTAL RESULTS OF THE LANGUAGE MODEL TRAINED ON THE KOREAN NUMERICAL REASONING DATA

Models	**Train**		**Test**	**Valid**	
	EM	F1	EM	EM	F1
Numnet(Ko-Glove 100d)	0.293	0.301	0.14	0.14	0.14
NAKoELECTRA	0.331	0.339	0.1548	0.155	0.162
Numnet+ with KoELECTRA	0.947	0.975	0.176	0.184	0.189

The experimental results showed that NumNet using Ko-Glove-100d recorded the lowest performance in Train, Test, and Valid datasets, confirming that training of simple word in glove feature embeddings is not as proper as language models pre-trained with big data.

Low performances were recorded for NAKoELECTRA and NumNet+ with KoELECTRA despite finding optimal loss in the performance difference in Train, Test, and Valid datasets by applying early stopping. Various factors can cause this, such as insufficient model training, lack of training data, and data errors. In our study, the most numerically aware model is in the third row, and it is enough to prove to suggest its possibility, considering that it was fine-tuned with the 1/6 dataset of DROP. This study also aims to increase the possibility of researching Korean, an LRL, by constructing a numerically aware Korean language model as described above instead of enhancing model performance.

IV. CONCLUSION

In this study, we demonstrated to verify the effectiveness of training a Korean numerical reasoning model in the MRC task of NLP. PLMs trained with a large corpus showed a tremendous performance improvement in the NLP field, and we leveraged transformer-based pre-trained ELECTRA architecture to fine-tune our models. Since the linguistic characteristics of Korean, an agglutinative language, we analyze the challenges of the existing multilingual language models such as BERT and propose the necessity for the Korean language-specific PLMs to be leveraged.

One thousand two hundred passages and 17,369 QA pairs were created from the KBO, and history topics specialized for the Korean domain. Eight annotators who were recruited to create data built the training data based on the provided guidelines. During this process, an automatic inter-annotation tool was devised to reduce human inaccuracies and errors and create high-quality datasets effectively. KoELECTRA, pre-trained with big data used in this experiment, is a model that trained both the generator and discriminator neural networks by performing RTD and learned the model so that numerical reasoning is possible by utilizing the numerical reasoning skills of NumNet.

Our experimental results suggested the possibility of training the Korean language model to create an MRC model capable of reasoning. However, the significant performance variability in experiments is a hurdle that needs to be overcome in the future. Our subsequent research will entail utilizing the current dataset to design an improved model with performance variability issues, utilization of various language models, and hyperparameter tuning.

ACKNOWLEDGMENT

This work was supported by Hanshin University Research Grant.

REFERENCES

[1] C. A. Perfetti, N. Landi, & J. Oakhill (2005). The Acquisition of Reading Comprehension Skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Blackwell Publishing.

[2] N. K. Duke, and P. D. Pearson. "Effective practices for developing reading comprehension." *Journal of education* 189.1-2 (2009): 107-122.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

[4] S. Lim, M. Kim, & J. Lee. (2018). Korquad: Korean qa dataset for machine comprehension. In *Proceeding of the Conference of the Korea Information Science Society* (pp. 539-541).

[5] D. Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[6] R. Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.* 21.140 (2020): 1-67.

[7] A. Vaswani, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[8] G. Kim, et al. "AI Student: A Machine Reading Comprehension System for the Korean College Scholastic Ability Test." *Mathematics* 10.9 (2022): 1486.

[9] G. Kim, et al. "Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network." *International journal of machine learning and cybernetics* 11.10 (2020): 2341-2355.

[10] A. Radford, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

[11] G. Kim, et al. "Enhancing Korean Named Entity Recognition With Linguistic Tokenization Strategies." *IEEE Access* 9 (2021): 151814-151823.

[12] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

[13] Q. Ran, Y. Lin, P. Li, J. Zhou, and Z. Liu. 2019. NumNet: Machine Reading Comprehension with Numerical Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2474–2484, Hong Kong, China. Association for Computational Linguistics.

[14] A. W. Yu, et al. "Qanet: Combining local convolution with global self-attention for reading comprehension." *arXiv preprint arXiv:1804.09541* (2018).

[15] M. Hu, Y. Peng, Z. Huang, and D. Li. 2019. A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*,

pages 1596–1606, Hong Kong, China. Association for Computational Linguistics.

[16] K. Chen, et al. 2020. Question Directed Graph Attention Network for Numerical Reasoning over Text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6759–6768, Online. Association for Computational Linguistics.

[17] A. Saha, S. Joty, and S. C. Hoi, (2021). Weakly supervised neuro-symbolic module networks for numerical reasoning. *arXiv preprint arXiv:2101.11802*.

[18] S. Reddy, D. Chen, and C. D. Manning, (2019). Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, *7*, 249-266.

[19] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics. Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., & Van Durme, B. (2018). Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

[20] Z. Sheng, et al. (2018). Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

[21] J. Welbl, P. Stenetorp, and S. Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

[22] M. Alexandre, V. Carles, and E. Heetderks. "Low-resource languages: A review of past work and future challenges." *arXiv preprint arXiv:2006.07264* (2020).

[23] G. Kim, et al. "Reading Comprehension requiring Discrete Reasoning Over Paragraphs for Korean." *Annual Conference on Human and Language Technology*. Human and Language Technology, 2021.

[24] C. Kevin, et al. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555* (2020).

[25] J. Park, GitHub, 2020, "KoELECTRA: Pretrained ELECTRA Model for Korean", https://github.com/monologg/KoELECTRA