A Survey on Awesome Korean NLP Datasets

Byunghyun Ban
Imagination Garden Inc.
Andong-si, Republic of Korea
bhban@kakao.com

2022 13th International Conference on Information and Communication Technology Convergence (ICTC) | 978-1-6654-9939-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICTC55196.2022.9952930

***Abstract*— English based datasets are commonly available from Kaggle, GitHub, or recently published papers. Although benchmark tests with English datasets are sufficient to show off the performances of new models and methods, still a researcher need to train and validate the models on Korean based datasets to produce a technology or product, suitable for Korean processing. This paper introduces 15 popular Korean based NLP datasets with summarized details such as volume, license, repositories, and other research results inspired by the datasets. Also, I provide high-resolution instructions with sample or statistics of datasets. Full-length version paper with summarization of each datasets with tables is available at arXiv (https://arxiv.org/pdf/2112.01624.pdf).**

Keywords*— *Artificial intelligence, Dataset, Korean, Natural Language Processing, NLP

I. Introduction

Natural language processing, NLP, with various neural network models has recently received a large amount of interest from both linguists and AI engineers. NLP has benefited from LSTM [1-3] and RNN models, due to their high performances on sequential data.

Although a CNN has hierarchical structures, YN Dauphin, et al. showed that application of CNN on natural languages outperformed traditional recurrent models [4]. Recent approaches on natural language processing utilize transformer architectures for language processing [5] – [8].

Noble approaches and brand-new applications inspired the researchers all around the world, to reproduce such high-performance models trained with various languages. For example, H Zhuang, et al. [9] and Y Cui, et al. [10] introduced Chinese based model, and M Al-Smadi, et al. [11] and S Romeo, et al. [12] showed models for Arabic natural language. Also, various Korean based NLP models have been introduced on international conferences [13] – [16] and Journals [17] – [18].

Researchers, who are working with non-English NLP technologies or multilanguage models, are highly interested in datasets in different languages. Even for researchers whose native language is a low-share language, the craving for datasets in their mother tongue langue would be much greater. The same goes for the industrial demands. NLP-related product holders are trying to research and develop models in official language of their country. Therefore, natural language processing researchers in Korea are very interested in Korean based NLP datasets. Not only academic institutions [19], [20], [25] and individual researchers [21], [22], but also large corporations [14], [23], [24] are actively producing Korean based natural language datasets.

II. Related Works

Simon Ruffieux, et al. provided summarized details of human pose estimation datasets to introduce various useful resources to researchers [49]. They wanted to reduce the time which other researchers spend for exploring useful datasets, and to provide useful tips for dataset construction.

Although there are several surveys on NLP technologies and algorithm themselves [26] – [28] or survey on Korean language processing technologies [29], an academic paper or comprehensive survey on Korean-based NLP dataset is not so many.

Cho, Won Ik, et al. established Open Korean Corpora [40] to provide assorted Korean NLP datasets. They provide the summarized detail of 29 different datasets from text parsing and tagging area, sentence similarity, sentence classification and QA, parallel corpora, multilingual corpora, and speech corpora. They summarized the information of each dataset into a short paragraph of one or two sentences. Their work is suitable for rapid and brief exploration among various datasets. But their work is insufficient to present a rationale for selecting a dataset suitable for the research purpose.

Yeongsook Song published a GitHub Repository named "AwesomeKorean_data" to provide information on various Korean datasets [30]. She introduced 31 datasets with academic references, and 9 datasets by individual researchers, and 5 government driven datasets. AwesomeKorean_data provides summarized detailed of datasets in a table which is easy to read quickly. It also provides hyperlink to the original webpage of each dataset. However, it categorized the license of datasets into too simple format (commercially available or not) so a researcher should figure out additional information on restrictions related to indication of same origin and impossibility of reprocessing and manipulations. Also, it does not provide published applications of datasets.

III. Survey

I selected 15 datasets which are famous among Korean research communities. Summarization over whole datasets is presented on Table 1. The table shows the name, category, domain and task, size, license, author, year, citation score and related research articles. This section also provides details of each dataset. For example, additional information such as repository URL, format, statistics of dataset or data sample, and application of the data.

A. *Chatbot_data*

Chatbot_data was published by Youngsook Song in 2018 [31]. The answers were composed of sentences that could comfort a person who has gone through a breakup with a lover. Answer sentences are labeled into 3 classes; 0 for daily life dialogue, 1 for negative (breakup), and 2 for positive (love).

This dataset has MIT license, which allows commercial use, modification, distribution and private use. However, the author is not responsible for any liability or warranty.

Hyeonseo Yun has published a paper on transformer-based AI model to detect unethical sentences from given text data, with this dataset [39].

B. *ClovaCall*

ClovaCall was provided by Jung-Woo Ha, et al. in 2020 [23]. It is a large scale Korean speech corpus dataset gathered from recorded phone calls, from more than 11,000 people under a given situation: restaurant reservation.

License-free ASR related datasets are old fashioned and most of them are established with English. Therefore, the authors constructed a license-free Korean phone call dataset with baseline codes.

Won Ik Cho, et al. included ClovaCall dataset in their work on Open Korean Corpora [40]. Jihwan Bang, et al. followed a variant version of LAS model proposed on ClovaCall paper [50].

As ClovaCall is recently released, a research article using ClovaCall for training or model validation could not be found yet. However, many recent works have mentioned ClovaCall as related recent works [51 – 54].

C. KorQuAD

KorQuAD dataset was published by Seungyoung Lim in 2019 [32]. This dataset is inspired by SQuAD [55], a reading comprehension-based question answering dataset driven from Wikipedia articles. It is the largest Korean question answering dataset. The authors drove data from Wikipedia.

KorQuAD is actively applied on Korean NLP conferences. Dongheon Lee, et al. applied KorQuAD for benchmark test of reading comprehension model in 2019 [56]. Minho Kim, et al. has developed a machine reading comprehension-based question and answering system and applied KorQuAD for performance test in 2020 [57]. Seohyung Jeon, et al. applied KorQuAD dataset on BERT-based multilingual model for machine reading comprehension task in 2020 [58].

D. Song-NER

Song_NER dataset was established by Hyewoong Park and Yeongsook Song in 2017 [33]. It consists of 4 categorical corpus; date (DT), location (LG), personal name (PS), time (TI). This dataset could be applied for named entity recognition model for chatbot system.

A chatbot model trained with Song_NER could perform information extraction from user's message. For example, a sentence "내일 저녁 7 시에 이태원 파스타집 5 명 예약해줘 (Please make a reservation for 5 people at the pasta restaurant in Itaewon tomorrow at 7pm)" could processed as below;

DT : tomorrow LC : Itaewon

PS : 5 people TI : 7pm.

Hyunjoong Kim, et al. has included this dataset on Korpora in 2019 [42].

E. KMOUNLP-NER

KMOUNLP-NER was openly provided at GitHub since 2016, and was officially published by Minah Cheon in 2021 [20]. Her paper proposed a criteria and method for Korean sentence tokenization. KMOUNLP-NER dataset was processed with the tokenization method on Korean sentences from various resources such as news articles.

Won Ik Cho, et al. cited KMOUNLP-NER on Open Korean Corpora in 2020 [40]. So-Yeop Yoo and Jeong Ok-Ran applied BERT and knowledge graph method for Korean contextual information extraction, trained with KMOUNLP-NER for named entity recognition in 2020 [43]. Yeon-SOO You and Hyuk-Ro Park also used corpuses from this dataset to establish syllable-based Korean NER model [59]. The number of citations of KMOUNLP-NER is underestimated because the author has opened dataset for 2 years without any citation request.

F. Sci-News-Sum-Kr-50

Sci-News-Sum-Kr-50 dataset was proposed by Jinsuk Seol and Sang-gu Lee in 2016 [34]. They collected 50 news articles from IT and Science section of news curation service. Each article is attached with human-written summarization. As copyright of each raw article belongs to the press, this dataset is not available for commercial use. Although the size of this dataset is small, the reliability of this dataset is quite high because the summarized sentences are written by human, without any augmentation algorithm. This dataset was enlisted on Open Korean Corpora [40].

G. SAE4K

SAE4K dataset was published by Won Ik Cho, et al. in 2019 [36]. They opened SAE4K_v1 dataset with 30,837 raw sentence pairs and SAE4K_v2 dataset with 50,837 augmented data. Each data pair has original sentence and extracted structured argument with sentence type label. The labels for sentence categorization and dataset statistics are described on Table XVI. This dataset was enlisted on Open Korean Corpora [40].

H. KLUE

KLUE is the first Korean-based natural language understanding evaluation dataset. KLUE dataset was created through collaboration among academia and industry, and published by Sungjoon Park, et al. in 2021.

KLUE consists of 8 different large datasets for various NLP tasks. KLUE-TC is a topic classification dataset with 63k data, driven from news headlines. KLUE-STS dataset stands for semantic textual similarity task with 12.5k samples, whose source are news, review and query. KLUE-NLI dataset has 31k samples for natural language inference task, from news, Wikipedia and review. KLUE-NER is a collection of 31k named entity recognition data constructed with news and review articles. KLUE-RE is a relation extraction dataset with 48k datasets driven from Wikipedia and news articles. KLUE-DP dataset consists of 14.5k news and review data for dependency parsing task. KLUE-MRC is a dataset for machine reading comprehension task, with 29k data from Wikipedia and news articles. And the last dataset, KLUE-DST, is a dialogue state tracking dataset with 10k data driven from task oriented dialogue.

As KLUE provides an amount benchmark data for various NLP tasks, it would provide a persuasive criteria for performance measurement on Korean language processing research area.

Myeongjun Jang, et al. has applied KLUE-NLI and KLUE-STS on natural language inference and semantic text similarity experiment for analysis on language understanding models in 2021 [60]. Boseop Kim, et al. has shown the effect and efficiency of large-scale language models by experiments on KLUE-TC and KLUE-STS datasets in 2021 [61].

I. KorNLU

KorNLU was established by Jiyeon Ham, et al. in 2020 [14]. It consists of two sub-datasets; KorNLI for natural language inference and KorSTS for semantic textual similarity.

KorNLI has 94k data for train, 2.5k data for development and 5k data for test, driven from SNLI [62], MNLI [63], and XNLI [64]. This dataset classifies the relationship between a pair of sentences into 3 classes: entailment, contradiction, neutral. KorSTS has 5.7k training data, 1.5k development data and 1.4 test data, and the source of KorSTS dataset is STS-B [65]. KorSTS provides a score between 0(dissimilar) and 5(equivalent) for each pair of sentences.

Myeongjun Jang applied KorNLI for experiments on language understanding models in 2021 [60]. Hyunjae Lee proposed a lite BERT model for Korean understanding with benchmark performance experiment on KorNLI and KorSTS [45]. Yongmin Yoo applied KorSTS for sentence similarity measurement in 2021 [66].

J. ParaKQC

ParaKQC is a dataset for parallel Korean questions and commands, proposed by Won Ik Cho, et al. in 2020 [36]. It contains 10k utterances that consist with 1,000 sets of 10 similar sentences, provided with an augmentation script which make up whole corpus of 545k utterances.

ParaKQC contributed as a resource of KLUE dataset [24]. Also it is enlisted on Open Korean Corpora [40]. Won Ik Cho, et al. applied the sentence classification scheme of ParaKQC for StyleKQC, a style-variant paraphrase corpus for Korean question and commands [67].

K. NSMC

NSMC dataset was constructed by Eunjeong Park in 2015, with sentences scrapped from Naver movie reviews [37]. Raw materials have user-written comments on the movie with score ratings between 1 (negative) to 10 (positive). The sentences were classified into 3 categories by rating scores; negative (1 ~ 4 score), neutral (5 ~ 8 score), positive (9 ~ 10 score). NSMC only contain the negative (label 0) and positive (label 1) sentences for semantic textual analysis task.

Although NSMC is a highly attractive dataset for Korean NLP, the author has not published related paper or provided official citation request text. Therefore, the exact citation number is not known; search result on Google Scholar indicates that at least 50 papers has applied NSMC for experiment or benchmark test. Yong-Jun Lee, et al. Used NSMC dataset for experiment on Korean-specific emotion annotation model in 2020 [46]. Sangwan Moon and Naoaki Okazaki applied NSMC to evaluate a just-in-time BERT model in 2020 [68]. Kyubong Park, et al. has analyzed the performance benchmark of various Korean NLP with NSMC dataset in 2020 [69]. NSMC also contributed to KLUE [24] and Open Korean Corpora [40].

L. Toxic Comment

Toxic Comment dataset is a collection of negative movie review sentences from NSMC datasets, with more specific emotional labels. The label of each sentence is a 5-dim vector, similar to one-hot expression. The parameters of label vector indicates toxic, obscene, threat, insult, and identity hate. Sentences has at least one '1' indicate for those categories.

M. KHateSpeech

KHateSpeech dataset was published by Jihyung Moon, et al. in 2020 [38]. They scrapped malicious replies from news articles from entertainment and celebrity sections, where the largest amount of negative comments are produced. This dataset provides 9,381 human-labeled data with 2,033,893 unlabeled raw comments. Each comment is annotated on two aspect; existence of social bias and hate speech. And additional binary label has attached to indicate whether a comment contains gender bias or not.

Chanhee Lee, et al. has applied KHateSpeech dataset to analyzed the data efficiency of cross-lingual post-training in pretrained language models in 2021 [70]. Seyoung Lee and Saerom Park used KHateSpeech dataset to proposed a hate speech classification algorithm using ordinal regression in 2021 [47]. Hyeonsang Lee, et al. proposed a CNN based model for toxic comments classification, trained with KHateSpeech dataset in 2020 [71].

N. 3i4K

3i4K was published by Won Ik Cho, et al. in 2018 [25]. They collected corpus data from released works of SNU-SLP(http://slp.snu.ac.kr/) and Natural Institute of Korean Language([1] https://krean.go.kr) to establish FCI (fragments, clear-cut cases, intonation-dependent utterances) dataset. They also manually created questions and commands for dataset.

3i4K dataset is enlisted on Open Korean Corpora [40]. Won Ik Cho, et al. has applied 3i4K to investigate the effectiveness of character-level embedding in Korean sentence classification in 2019 [48].

O. KAIST Corpus

KAIST Corpus dataset is a collection of various corpus and machine translation datasets, constructed since 1997 to 2005, published by KAIST Semantic Web Research Center [72]. IT has over 70 million Korean text corpus, position annotated corpus, tree-annotated corpus, Korean-Chinese parallel corpus, Korean-English parallel corpus.

KAIST corpus was established by Ki-sun Choi, and published in 2001.[72]. KAIST corpus has various sub-datasets described on Table XXXIII and TABLE XXXIV.

Qualified corpus has 7 Korean corpus datasets and 4 multilingual corpus datasets. Processed resources section provides 23 Korean corpus dataset and 1 Korean-English co-occurrence dataset, and an off-line Korean handwriting image dataset. KAIST corpus is also enlisted on Open Korean Corpora [40].

Considering the year of publication, the number of citations of KAIS corpus dataset is very low. So it's difficult to find a benchmark performance of recent algorithms tested with KAIST corpus. However, as the size of datasets are very large and the quality of each dataset is excellent, it still worth using for algorithm design and model training.

IV. Discussion

As the interests and demands of companies on Korean language processing dataset has dramatically increased, fund and investment seems to be increased too. Therefore, great dataset such as KLUE has been published and accepted to Neural IPS this year. As described on Table 1, most of the datasets popular today were published within last 3 years. This reflects the recent advance in popularity of Korean NLP dataset construction.

Since the datasets were recently published, I guess that many research projects using them are currently in progress, or have not yet published. This may explain the low citation scores of useful datasets. The number of citation is considered as a good indicator to evaluate a popularity of dataset. I agree that a dataset which are frequently cited is very useful to produce a reasonable benchmark result. However, many dataset papers are cited because of the novel method proposed by the authors, rather than the dataset itself.

A researcher should not underestimate or overestimate a dataset by the paper's citation score only. I suggest exploring various datasets to find one, best appropriate for your research purpose.

V. Conclusion

This paper provided a survey of Korean-based natural language processing datasets. The survey section proposed brief and comprehensive summarization and detailed information. It covered conversation log, question answering, named entity recognition, text summarization, natural language understanding, semantic sentence similarity, semantic textual analysis, semantic speech analysis and machine translation. The discussion outlined a brief guide to consider when choosing a dataset. In conclusion, I hope this paper to help researchers' to reduce time consuming works for dataset exploring and to choose appropriate dataset in Korean NLP area.

Acknowledgment

This paper was inspired while writing the book *"142 datasets for AI researchers."* with Life & Power Press Co., Ltd. I would like to express my gratitude to Jehoon Yoo for suggesting me collect various datasets to write such a book.

REFERENCES

[1] Hamid Palangi, Li Deng, Yelong Shen, Jian Feng Gao, Xiaodong He, Jianshu Chen, et al. “Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval.” in *IEEE/ACM Transactions on Audio, Speech, and language Processing* 24.4, 2016, pp.694-707, doi: 10.1109/TASLP.2016.2520371.

[2] Luukkonen, Petri, Markus Koskela, and Patrik Floréen. "LSTM-based predictions for proactive information retrieval." *arXiv preprint* arXiv:1606.06137, 2016. [Online] Available: https://arxiv.org/abs/1606.06137.

[3] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song and R. Ward "Semantic modelling with long-short-term memory for information retrieval." *arXiv preprint* arXiv:1412.6629(2014). [Online] Available: https://arxiv.org/abs/1412.6629

[4] Yann N. Dauphin, Angela Fan, Michael Auli and David Grangier. “Language modeling with gated convolutional networks.” In *International Conference on Machine Learning.* PMLR, 2017. [Online] Available: http://proceedings.mlr.press/v70/dauphin17a.

[5] Yang, X, He, X, Zhang, H, Ma, Y, Bian, J, Wu, Y. "Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models". *JMIR medical informatics* 2020. 8(11):e19735. [Online] Available: https://medinform.jmir.org/2020/11/e19735, doi: 10.2196/19735.

[6] Sun, L, Xia, C, Yin, W, Liang, T, Yu, P, He, L. "Mixup-Transformer: Dynamic Data Augmentation for NLP Tasks." *arXiv preprint* arXiv:2010.02394 2020. [Online] Available: https://arxiv.org/abs/2010.02394.

[7] Wolf, T, Debut, L, Sanh, V, Chaumond, J, Delangue, C, Moi, A, Cistac, P, Rault, T, Louf, R, Funtowicz, M, et al. "Huggingface's transformers: State-of-the-art natural language processing". *arXiv preprint* arXiv:1910.03771 2019. [Online] Available: https://arxiv.org/abs/1910.03771

[8] Gillioz, A, Casas, J, Mugellini, E, Abou Khaled, O. “Overview of the Transformer-based Models for NLP Tasks.” In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. 2020. pp. 179–183, doi: 10.15439/2020F20.

[9] Zhuang, H, Wang, C, Li, C, Wang, Q, Zhou, X. “Natural language processing service based on stroke-level convolutional networks for Chinese text classification.” In *2017 IEEE international conference on web services (ICWS)* 2017. pp. 404–411, doi: 10.1109/ICWS.2017.46.

[10] Cui, Y, Che, W, Liu, T, Qin, B, Wang, S, Hu, G. "Revisiting pre-trained models for chinese natural language processing". *arXiv preprint* arXiv:2004.13922 2020. [Online] Available: https://arxiv.org/abs/2004.13922.

[11] Al-Smadi, M, Al-Zboon, S, Jararweh, Y, Juola, P. "Transfer learning for Arabic named entity recognition with deep neural networks". *IEEE Access* 2020; 8:37736–37745. doi: 10.1109/ACCESS.2020.2973319.

[12] Romeo, S, Da San Martino, G, Belinkov, Y, Barron-Cedeno, A, Eldesouki, M, Darwish, K, Mubarak, H, Glass, J, Moschitti, A. "Language processing and learning models for community question answering in Arabic." *Information Processing & Management 56(2).* 2019. pp. 274–290. doi: 10.1016/j.ipm.2017.07.003.

[13] Yun, Hyungbin, Ghudae Sim, and Junhee Seok. "Stock prices prediction using the title of newspaper articles with korean natural language processing." in *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 2019. doi: 10.1109/ICAIIC.2019.8668996.

[14] Ham, J, Choe, Y, Park, K, Choi, I, Soh, H. “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.* 2020. pp. 422–430. doi: 10.18653/v1/2020.findings-emnlp.39.

[15] Stratos K. “A sub-character architecture for Korean language processing.” In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings. Association for Computational Linguistics (ACL).* 2017. p. 721-726. doi: 10.18653/v1/d17-1075.

[16] Hwang, MH, Shin, J, Seo, H, Im, JS, Cho, H. "KoRASA: Pipeline Optimization for Open-Source Korean Natural Language Understanding Framework Based on Deep Learning". *Mobile Information Systems 2021*; 2021. doi: 10.1155/2021/9987462.

[17] Jung, Haemin, and Wooju Kim. "Automated conversion from natural language query to SPARQL query." *Journal of Intelligent Information Systems.* 2020. pp. 1-20. doi: 10.1007/s10844-019-00589-2

[18] Kim, H, Lee, JK, Shin, J, Choi, J. "Visual language approach to representing KBimCode-based Korea building code sentences for automated rule checking". *Journal of Computational Design and Engineering* 6(2). 2019. pp.143–148. doi: 10.1016/j.jcde.2018.08.002.

[19] J.-H. Kim and G. C. Kim, “Guideline on Building a Korean Part-of-Speech Tagged Corpus: KAIST Corpus.” *Technical Report CS-TR-95-99*, Department of Computer Science, KAIST, 1995 (in Korean).

[20] Minah Cheon, "다중 생성 단위의 관계 점수를 이용한 학습 말뭉치 생성: 개체명 말뭉치를 중심으로", Ph.D. dissertion, department of Computer Science, Korea Maritime & Ocean University, Busan-si, Republic of Korea, 2021.

[21] Song, Youngsook, "Toxic Comment." 2018. [Online] Available: https://github.com/songys/Toxic_comment_data.

[22] Alan Kang and Jieun Kim, "Petition." 2017. [Online] Available: https://github.com/akngs/petitions.

[23] Ha, JW, Nam, K, Kang, J, Lee, SW, Yang, S, Jung, H, Kim, H, Kim, E, Kim, S, Kim, H, et al. “ClovaCall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers.” In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).* 2020. pp. 409. doi: 10.21437/Interspeech.2020-1136.

[24] Park, S, Moon, J, Kim, S, Cho, W, Han, J, Park, J, Song, C, Kim, J, Song, Y, Oh, T, et al. "KLUE: Korean Language Understanding Evaluation". *arXiv preprint* arXiv:2105.09680 2021. [Online] Avaiable: https://arxiv.org/abs/2105.09680.

[25] Cho, W, Lee, H, Yoon, J, Kim, S, Kim, N. "Speech intention understanding in a head-final language: A disambiguation utilizing intonation-dependency". *arXiv preprint* arXiv:1811.04231 2018. [Online] Avaiable: https://arxiv.org/abs/1811.04231.

[26] Zeng, Z, Shi, H, Wu, Y, Hong, Z. "Survey of natural language processing techniques in bioinformatics". in *Computational and mathematical methods in medicine 2015*. 2015. doi: 10.1155/2015/674296.

[27] Otter, D, Medina, J, Kalita, J. "A survey of the usages of deep learning for natural language processing". *IEEE Transactions on Neural Networks and Learning Systems.* 32(2). 2020. pp.604–624. doi: 10.1109/TNNLS.2020.2979670.

[28] Oshikawa, Ray, Jing Qian, and William Yang Wang. "A survey on natural language processing for fake news detection." *arXiv preprint* arXiv:1811.00770. 2018. [Online] Available: https://arxiv.org/abs/1811.00770.

[29] Manhui Han, Sungchan Park, Hanbit Lee, Jeongheum Yeon, Sang-goo Lee. "Natural language processing on Korean language: A survey." *Processing of the Korean Information Science Society Conference.* 2016. pp 681-683. [Online] Available: https://s-space.snu.ac.kr/handle/10371/95645.

[30] Song, Yeongsook. "Awesome Korean Data." [Online] Available: https://github.com/songys/AwesomeKorean_Data.
[31] Song, Youngsook, "Chatbot_data." 2018. [Online] Available: https://github.com/songys/Chatbot_data.
[32] Lim, Seungyoung, Myungji Kim, and Jooyoul Lee. "KorQuAD1. 0: Korean QA dataset for machine reading comprehension." *arXiv preprint* arXiv:1909.07005. 2019. [Online] Available: https://arxiv.org/abs/1909.07005.
[33] 박혜웅, 송영숙. "음절 기반의 CNN을 이용한 개체명 인식." in *한국어정보학회 학술대회*. Republic of Korea. 2017. pp. 330-332.
[34] 설진석, 이상구. "LexRanKR: LexRank 기반 한국어 다중 문서 요약." In *한국정보과학회 학술발표논문집*. Republic of Korea. 2016. pp. 458-460.
[35] Cho, W, Moon, Y, Moon, S, Kim, S, Kim, N. "Machines Getting with the Program: Understanding Intent Arguments of Non-Canonical Directives." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 2020. pp. 329–339. doi: 10.18653/v1/2020.findings-emnlp.31
[36] Cho, W, Kim, J, Moon, Y, Kim, N. "Discourse component to sentence (DC2S): An efficient human-aided construction of paraphrase and sentence similarity dataset." In *Proceedings of The 12th Language Resources and Evaluation Conference.* 2020. pp. 6819–6826. [Online] Available: https://aclanthology.org/2020.lrec-1.842.
[37] Eunjeong Park, "NSMC: Naver sentiment movie corpus v1.0." 2015. [Online] Available: https://github.com/e9t/nsmc.
[38] Moon, Jihyung, Won Ik Cho, and Junbum Lee. "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*. 2020. doi: 10.18653/v1/2020.socialnlp-1.4.
[39] Yun, Hyeonseo, and Sunyong Yoo. "Transformer-based Unethical Sentence Detection." *Journal of Digital Contents Society* 22.8. 2021. pp. 1289-1293. doi: 10.9728/dcs.2021.22.8.1289.
[40] Cho, Won Ik, Sangwhan Moon, and Youngsook Song. "Open Korean Corpora: A Practical Report." *arXiv preprint* arXiv:2012.15621 (2020). [Online] Available: https://arxiv.org/abs/2012.15621.
[41] Lee, Hanbum, Jahwan Koo, and Ung-Mo Kim. "A Study on Emotion Analysis on Sentence using BERT." In *Proceedings of the Korea Information Processing Society Conference*. Korea Information Processing Society, 2020. doi: 10.3745/PKIPS.y2020m11a.909
[42] Hyunjoong Kim, Gichang Lee, Tim Lee, Hank Kim, Won Ik Cho and Taewook Kim. "Korpora:Korean Corpora Archives." 2019. [Online] Available: https://github.com/ko-nlp/Korpora
[43] Yoo, S, Jeong, O. "Korean Contextual Information Extraction System using BERT and Knowledge Graph". *Journal of Internet Computing and Services* 21(3). 2020. pp.123–131. doi: 10.7472/jksii.2020.21.3.123.
[44] Kim, B, Kim, H, Lee, SW, Lee, G, Kwak, D, Jeon, D, Park, S, Kim, S, Kim, S, Seo, D, et al. "What changes can large-scale language models bring? intensive study on HyperClova: Billions-scale korean generative pretrained transformers". *arXiv preprint* arXiv:2109.04650 2021. [Online] Available: https://arxiv.org/abs/2109.04650.
[45] Lee, H, Yoon, J, Hwang, B, Joe, S, Min, S, Gwon, Y. "KoreALBERT: Pretraining a Lite BERT Model for Korean Language Understanding." In *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021. pp. 5551–5557. doi: 10.1109/ICPR48806.2021.9412023.
[46] Lee, Young-Jun, Chae-Gyun Lim, and Ho-Jin Choi. "Korean-Specific Emotion Annotation Procedure Using N-Gram-Based Distant Supervision and Korean-Specific-Feature-Based Distant Supervision." In *Proceedings of The 12th Language Resources and Evaluation Conference.* 2020. [Online] Available: https://aclanthology.org/2020.lrec-1.199.
[47] Lee, Seyoung, and Saerom Park. "Hate Speech Classification Using Ordinal Regression." In *Proceedings of the Korean Society of Computer Information Conference.* Korean Society of Computer Information. 2021. [Online] Available: http://www.koreascience.or.kr/journal/CPTSA9.page.
[48] Cho, Won Ik, Seok Min Kim, and Nam Soo Kim. "Investigating an Effective Character-level Embedding in Korean Sentence Classification." *In Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation.* Waseda Institute for the Study of Language and Information. 2019. [Online] Available: http://jaslli.org/files/proceedings/02_paclic33_postconf.pdf.
[49] Ruffieux, Simon, et al. "A survey of datasets for human gesture recognition." In *International conference on human-computer interaction*. Springer, Cham, 2014. doi: 10.1007/978-3-319-07230-2_33.
[50] Bang, Jihwan, et al. "Boosting Active Learning for Speech Recognition with Noisy Pseudo-labeled Samples." *arXiv preprint* arXiv:2006.11021 2020. [Online] Available: https://arxiv.org/abs/2006.11021.
[51] Cho, W, Kim, S, Cho, H, Kim, N. "Kosp2e: Korean Speech to English Translation Corpus". *arXiv preprint* arXiv:2107.02875 2021. [Online] Available: https://arxiv.org/abs/2107.02875.
[52] Bang, JU, Yun, S, Kim, SH, Choi, MY, Lee, MK, Kim, YJ, Kim, DH, Park, J, Lee, YJ, Kim, SH. "KsponSpeech: Korean spontaneous speech corpus for automatic speech recognition". *Applied Sciences*. 10(19). 2020; pp. 6936. doi: 10.3390/app10196936.
[53] Lee, SW, Jung, H, Ko, S, Kim, S, Kim, H, Doh, K, Park, H, Yeo, J, Ok, SH, Lee, J, et al. "Carecall: a call-based active monitoring dialog agent for managing covid-19 pandemic". *arXiv preprint* arXiv:2007.02642 2020. [Online] Available: https://arxiv.org/abs/2007.02642.
[54] Hwang, S, Kim, J. "Toward a Chatbot for Financial Sustainability". *Sustainability* 13(6). 2021. pp. 3173. doi: 10.3390/su13063173.
[55] Rajpurkar, P, Jia, R, Liang, P. "Know What You Don't Know: Unanswerable Questions for SQuAD." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* 2018. pp. 784–789. doi: 10.18653/v1/P18-2124.
[56] 이동헌, 박천음, 이창기, 박소윤, 임승영, 김명지, 이주열. "BERT 를 이용한 한국어 기계 독해". in *한국정보과학회 학술발표논문집*, Republic of Korea. 2019. pp.557–559.
[57] Kim, M, Cho, S, Park, D, Kwon, HC. "Machine Reading Comprehension-based Question and Answering System for Search and Analysis of Safety Standards". *Journal of Korea Multimedia Society.* 23(2). 2020. pp.351–360. doi: 10.9717/kmms.2020.23.2.351.
[58] You, Yeon-Soo, and Hyuk-Ro Park. "Syllable-based Korean named entity recognition using convolutional neural network." In *한국마린엔지니어링학회지* 44.1. 2020. pp. 68-74. doi: 10.5916/jamet.2020.44.1.68.
[59] 정서형, and 곽노준. "BERT 를 이용한 한국어 질의응답 데이터 셋에서의 기계 독해." In *대한전자공학회 학술대회*. Republic of Korea. 2020 pp.625-630.
[60] Jang, M, Kwon, D, Lukasiewicz, T. "Accurate, yet inconsistent? Consistency Analysis on Language Understanding Models". *arXiv preprint* arXiv:2108.06665 2021. [Online] Available: https://arxiv.org/abs/2108.06665.
[61] Kim, B, Kim, H, Lee, SW, Lee, G, Kwak, D, Jeon, D, Park, S,

Kim, S, Kim, S, Seo, D, et al. "What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers". *arXiv preprint* arXiv:2109.04650 2021. [Online] Available: https://arxiv.org/abs/2109.04650.

[62] Bowman, S, Angeli, G, Potts, C, Manning, CA large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*. 2015. pp. 632–642. doi: 10.18653/v1/d15-1075.

[63] Williams, A, Nangia, N, Bowman, S. "A broad-coverage challenge corpus for sentence understanding through inference." In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018* .2018. pp. 1112–1122. [Online] Available: https://arxiv.org/abs/1704.05426.

[64] Conneau, A, Rinott, R, Lample, G, Schwenk, H, Stoyanov, V, Williams, A, Bowman, S, "XNLI: Evaluating cross-lingual sentence representations." In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. 2020. pp. 2475–2485. [Online] Available: https://arxiv.org/abs/1809.05053.

[65] Cer, D, Diab, M, Agirre, E, Lopez-Gazpio, I, Specia, L. "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation." In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 2017. pp. 1–14. doi: 10.18653/v1/S17-2001.

[66] Yoo, Y, Heo, TS, Park, Y, Kim, K. "A novel hybrid methodology of measuring sentence similarity". *Symmetry* 13(8). 2021. pp.1442. doi: 10.3390/sym13081442.

[67] Cho, W, Moon, S, Kim, J, Kim, S, Kim, N. "StyleKQC: A Style-Variant Paraphrase Corpus for Korean Questions and Commands". *arXiv preprint* arXiv:2103.13439 2021. [Online] Available: https://arxiv.org/abs/2103.13439.

[68] Moon, S, Okazaki, N. "PatchBERT: Just-in-Time, Out-of-Vocabulary Patching." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 2020. pp. 7846–7852. doi: 10.18653/v1/2020.emnlp-main.631.

[69] Park, K, Lee, J, Jang, S, Jung, D. "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.* 2020. pp. 133–142. [Online] Available: https://aclanthology.org/2020.aacl-main.17.

[70] Lee, C, Yang, K, Whang, T, Park, C, Matteson, A, Lim, H. "Exploring the Data Efficiency of Cross-Lingual Post-Training in Pretrained Language Models". *Applied Sciences* 11(5). 2021. pp. 1974. doi: 10.3390/app11051974.

[71] 이현상, 이희준, and 오세환. "딥러닝 기술을 활용한 악성댓글 분류: Highway Network 기반 CNN 모델링 연구." in *한국경영학회 통합학술발표논문집*. Republic of Korea. 2020. pp. 343-351.

[72] Ki-sun Choi. "KAIST Language Resources 2001 edition." *Result of the core software project from Ministry of Science and Technology*, Republic of Korea. 2001.