Open Korean Corpora: A Practical Report

Won Ik Cho
Seoul National University
Seoul, Korea
wicho@hi.snu.ac.kr

Sangwhan Moon
Tokyo Institute of Technology
Odd Concepts Inc.
Tokyo, Japan
sangwhan@iki.fi

Youngsook Song
Kyung Hee University
Seoul, Korea
youngsoksong@khu.ac.kr

Abstract

Korean is often referred to as a low-resource language in the research community. While this claim is partially true, it is also because the availability of resources is inadequately advertised and curated. This work curates and reviews a list of Korean corpora, first describing institution-level resource development, then further iterate through a list of current open datasets for different types of tasks. We then propose a direction on how open-source dataset construction and releases should be done for less-resourced languages to promote research.

1 Introduction

The Korean language is less explored in terms of corpus and computational linguistics, but its prevalence is often underrated. It regards about 80 million language users and is recently adopted in multilingual research as it is bound to CJK (Chinese, Japanese, and Korean), also handling a distinguished writing system.

However, compared to the industrial need, the interest in Korean natural language processing (NLP) has not been developed much in international viewpoints, which recurrently hinders the related publication and further academic extension. Besides, in the recent NLP, where the benchmark practice is a trend, such systems lack at this point, deterring abroad and even native researchers who start Korean NLP from finding directions. Park et al. (2016) has shown a decent survey, but it seems that the techniques are mainly on the NLP pipeline. Also, albeit some curations on Korean NLP[1] and datasets[2], we considered that little more organization is required, and better if internationally available. Our attempts are expected to mitigate the challenges that the researchers who handle Korean from a multi- or cross-lingual viewpoint may face.

In this paper, we scrutinize the struggles of government, institutes, industry, and individuals to construct public Korean NLP resources. First, we state how the institutional organizations have tackled the issue by making up the accessible resources, and point out the limitation thereof regarding international availability and license, to finally introduce and curate the fully public datasets along with the proposed criteria. Through this, we want to find out the current state of Korean corpora across the NLP tasks and whether they are freely or conditionally available. Our survey is to be curated and updated in the public repository[3].

2 Accessible Resources

With the increase in popularity of machine learning-driven methods in NLP, constructing a novel dataset and releasing it to the public can be considered the cornerstone of advancing research of a given language. While we believe many useful datasets exist behind industry walls, this is not particularly useful for advancing open research. Fortunately, there are organizations that construct and distribute cleaned, pre-processed datasets which are occasionally accompanied by a task and the annotation. In the context of Korean, there are numerous efforts in this field driven by government-affiliated organizations.

2.1 Datasets from public institutions

National Institute of Korean Language (NIKL) is an institution that establishes the norm for Korean linguistics[4]. However, at the same time, it usually undergoes the massive dataset construction from the view of computational

[1] https://github.com/datanada/Awesome-Korean-NLP
[2] https://littlefoxdiary.tistory.com/42
[3] https://github.com/ko-nlp/Open-korean-corpora
[4] https://www.korean.go.kr/

Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), pages 85–93

linguistics, to apt to the new wave of language artificial intelligence (AI). Widely known ones include Korean word dictionaries[5] and Sejong Corpus (Kim, 2006). The dictionary contains fundamental and new lexicons that make up Korean (along with the content), and the Sejong Corpus is a large-scale labeled NLP pipeline corpus for the tasks such as constituency and dependency parsing, mainly provided in *.json* format. Besides, recently, labeled corpora of about 300 million word size is released[6], covering inter-sentence tasks such as similarity and entailment.

Electronics and Telecommunications Research Institute (ETRI) has been collecting, refining, and tagging language processing and speech learning data over a long period of time[7]. Aside from NIKL, which mainly focuses on classical NLP pipelines, ETRI has also built a database for semantic analysis and question answering (QA), which are the outcome of a project Exo-brain[8]. The project includes syntax-semantic ones such as part of speech (POS) tagging and semantic role labeling (SRL), simultaneously providing construction guidelines for the corpora.

AI HUB is a platform organized by National Information Society Agency (NIA) in which a large-scale dataset are integrated[9]. The datasets are built for various tasks at the government level, to promote the development of the AI industry. Provided resources are labeled or parallel corpora in real-life domains. Here, the domains are law, patent, common sense, open dialog, machine reading comprehension, and machine translation. Also, about 1,200 hours of speech corpus is provided to be used in spoken language modeling[10]. Recently, some new datasets have been distributed on wellness and emotional dialog, so that many people can have trials for social good and public AI. Also, open dictionary NIAdic[11] is freely available, provided by K-ICT Big Data Center.

[5]The search portal is provided in https://stdict.korean.go.kr/main/main.do while the full word and content list are available at https://github.com/korean-word-game/db

[6]https://corpus.korean.go.kr/

[7]https://www.etri.re.kr/intro.html

[8]http://exobrain.kr/pages/ko/result/outputs.jsp

[9]http://www.aihub.or.kr/

[10]https://www.aihub.or.kr/aidata/105

[11]https://kbig.kr/portal/kbig/knowledge/files/bigdata_report.page?bltnNo=10000000016451

2.2 Accessibility

The above datasets guarantee high quality, along with well-defined guidelines and the well-educated workers. However, their usage is often unfortunately confined to domestic researchers for procedural issues. Researchers abroad can indeed access the data, but they may face difficulty filling out and submitting the particular application form, instead of the barrier-free downloading system. Also, in most cases, modification and redistribution are restricted, making them uncompetitive in view of quality enhancement (Han et al., 2017).

Here, we want to introduce datasets that can be utilized as an alternative to the limitedly accessible Korean NLP resources. Instead of scrutinizing all available corpora, we are going to curate them under specific criteria.

3 Open Datasets

All the datasets to be introduced from now on are fully open access. This means that the dataset is downloadable with a single click or cloning, or at least one can acquire the dataset with simple signing. We set three checklists for the status of the corpus, namely *documentation*, *usage*, and *redistribution*. The first one is on how fine-grained the corpus description is.

- Does the corpus have any documentation on the usage? (**doc**)
- Does the corpus have a related article?[12] (**art**)
- Does the corpus have a internationally available publication? (**inter**)

Next, we check whether the dataset is commercially available, academic use only, or unknown **(com, acad, unk)**. For the last one, We also investigate if redistribution is available with or without modification, if neither, or unknown **(rd, rd/mod-x, no, unk)**. These attributes are noted along with each corpus title.

3.1 Parsing and tagging

KAIST Morpho-Syntactically Annotated Corpus [art, acad, no] applies morphological analysis to freely available KAIST raw corpus[13]. The scale is about 70M words and the domain includes novel, non-literature, article, etc.

[12]Article is here more a complete form of document than *doc* above, and some domestic publications are included here since they are not internationally available.

[13]http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus

KAIST Korean Tree-Tagging Corpus [inter, acad, no] Choi et al. (1994)[14] bases on independently collected 30K sentences that are annotated according to the tree tagging scheme for Korean.

UD Korean KAIST [inter, acad, no] Chun et al. (2018)[15] applies universal dependency (UD) parsing (McDonald et al., 2013) to the Korean Tree-Tagging Corpus (Choi et al., 1994).

PKT-UD [inter, acad, no] Chun et al. (2018); Oh et al. (2020)[16] applies UD parsing to the Penn Korean Treebank (Han et al., 2001)[17].

KMOU NER [art, acad, rd] is a named entity recognition (NER) dataset built by Korean Marine and Ocean University[18]. The named entities are tagged for about 24K utterances according to name, time, and number. The data source are Exo-brain (by ETRI) and their own data combined, while the redistribution is available only for the latter.

AIR×NAVER NER/SRL [doc, acad, no] adopted the NER[19] and SRL[20] data constructed by Changwon National University for the purpose of a public competition[21], and is annotated according to CoNLL format (Tjong Kim Sang and De Meulder, 2003). Corpus size is about 90K and 35K each.

3.2 Entailment and sentence similarity

Question Pair [doc, com, rd] consists of about 10,000 open domain sentence pairs[22], with the binary labels that are hand-annotated on whether the sentences are paraphrase or irrelevant.

KorNLI/KorSTS [inter, com, rd] Ham et al. (2020) is a natural language inference (NLI) and sentence textual similarity (STS) dataset for Korean[23]. For KorNLI, the train set was constructed by machine translating SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), and the valid and test set were constructed by human translation of XNLI (Conneau et al., 2018). Just as in the original dataset, the pairs are labelled with entailment, contradiction, or neutral. About 940K examples are provided for training, and 2,490 and 5,010 respectively for dev and test. For KorSTS, the scoring was done from 0 to 5 to elaborate rather than the binary label that determines paraphrase. Following the scheme of NLI, 5,749 training data were machine translated using the STS-B dataset (Cer et al., 2017) as a source, while 1,500 dev set and 1,379 test set pairs are human translated.

ParaKQC [inter, com, rd] Cho et al. (2020)[24] originally consists of 10,000 questions and commands, and each instance is labeled with 4 topics (mail, smart agent, scheduling, and weather) and 4 speech acts (*wh*-question, alternative question, prohibition, and requirement). The sentence set can be extended to about 540K sentence pairs that determine sentence similarity and paraphrase.

3.3 Sentence classification and QA

NSMC [doc, com, rd] is a review sentiment corpus[25] of size 200K, which consists of Naver movie comments automatically labeled according to the methodology of Maas et al. (2011). It adopts pos/neg binary labels, and it has been widely used as a benchmark for pretrained language models.

BEEP! [inter, com, rd] Moon et al. (2020) is a hand-labeled, crowd-sourced dataset of about 9.4K Naver entertainment news comments with hate speech and social bias[26]. Bias and hate attribute consists of 3 labels, namely gender/others/none and hate/offensive/none, respectively.

3i4K [inter, com, rd] Cho et al. (2018) aims an utterance-level speech act classification of the Korean language[27]. The volume reaches 61K, hand-labeled with 7 classes, namely fragment, statement, question, command, rhetorical question/command, and intonation-dependent utterances.

KorQuAD 1.0, 2.0 [inter, com, rd/mod-x] provides human-generated QA corpus and leaderboard for Korean[28]. KorQuAD 1.0 (Lim et al., 2019) benchmarks SQuAD 1.0 (Rajpurkar et al., 2016)

[14] http://semanticweb.kaist.ac.kr/home/index.php/Corpus4

[15] https://github.com/emorynlp/ud-korean

[16] Also available at UD-Korean repository, but currently previous version. PKT v2020 data will be uploaded.

[17] https://catalog.ldc.upenn.edu/LDC2006T09 LDC materials are not curated here.

[18] https://github.com/kmounlp/NER

[19] http://air.changwon.ac.kr/?page_id=10

[20] http://air.changwon.ac.kr/?page_id=14

[21] https://github.com/naver/nlp-challenge

[22] https://github.com/songys/Question_pair

[23] https://github.com/kakaobrain/KorNLUDatasets

[24] https://github.com/warnikchow/paraKQC

[25] https://github.com/e9t/nsmc

[26] https://github.com/kocohub/korean-hate-speech

[27] https://github.com/warnikchow/3i4k

[28] https://korquad.github.io/

and consists of total 70K questions. KorQuAD 2.0 of size 100K aims at machine reading comprehension for structured HTML natural questions, which was created referring to the scheme of Google Natural Questions (Kwiatkowski et al., 2019).

3.4 Parallel corpora

Sci-news-sum-kr [doc, acad, rd] contains about 50 Korean news summarizations generated by two Korean natives[29]. Since the size is not large, it is recommended to be used as a dev set.

sae4K [inter, com, rd] Cho et al. (2019b) contains the directive sentence summarization of the sentence level. It includes about 50K pairs of utterance and natural language query pair for questions and commands, where the data is partly based on 3i4K (Cho et al., 2018) and some are human-generate in concurrence with Cho et al. (2020).

Korean Parallel Corpora [inter, acad, rd/mod-x] Park et al. (2016) contains about 100K en-ko sentence pairs for machine translation (MT). The data mainly bases on news articles, and now also provides the data on North Korean[30].

KAIST Translation Evaluation Set [doc, acad, no] is an evaluation set of size about 3,000 for en-ko MT[31], augmented with index, original sentence, translation, related articles, and text source.

KAIST Chinese-Korean Multilingual Corpus [doc, acad, no] contains 60K short sentence pairs for zh-ko MT[32].

Transliteration Dataset [doc, com, rd] is not an official data repository[33], but en-ko transliteration is collected from public dictionaries such as NIKL or Wiktionary[34]. A total of about 35K en (word) - ko (pronunciation) pairs are included.

KAIST Transliteration Evaluation Set [doc, acad, no] is a word-pronunciation pair for phonotactics in en-ko[35], and consists of 7,186 words excerpted from the loanword dictionary[36].

3.5 Korean in multilingual corpora

Multilingual G2P Conversion [inter, com, rd] Gorman et al. (2020) is a shared task of SIGMORPHON 2020[37], which aims to transform grapheme sequence into a phoneme sequence. The dataset was created with WikiPron[38] (Lee et al., 2020), and has been built for 10 languages including Korean (3,600 pairs for train, and 450 for dev/test each).

PAWS-X [inter, com, rd] Yang et al. (2019) is a dataset that consists of 23,659 human translated PAWS evaluation pairs (Zhang et al., 2019) and about 300K machine-translated ones, for 6 languages including Korean[39]. Among them, Korean occupies about 5K train pairs, and 1,965 and 1,972 for dev/test each.

TyDi-QA [inter, com, rd] Clark et al. (2020) pursues typological diversity in QA, and provides a total of 200,000 question-answer pairs for 11 linguistically diverse languages, including Korean[40]. Among them, Korean occupies about 11K train pairs, and 1,698/1,722 for dev/test each.

XPersona [inter, com, rd] Lin et al. (2020) is a dataset for evaluating personalized chatbots[41]. It provides the dataset of Zhang et al. (2018) translated to 7 languages, including Korean, where Korean displays 299 dialogues with 4,684 utterances.

3.6 Speech corpora

Speech datasets are usually massive, that a downloading via a single click is not necessarily guaranteed. Thus, we listed some of them as open even if they require some application form.

KSS [doc, acad, rd] Park (2018) is a book corpus read by a female voice actress. 12K speech

[29] https://github.com/theeluwin/sci-news-sum-kr-50
[30] https://github.com/jungyeul/korean-parallel-corpora
[31] http://semanticweb.kaist.ac.kr/home/index.php/Evaluateset2
[32] http://semanticweb.kaist.ac.kr/home/index.php/Corpus9
[33] https://github.com/muik/transliteration
[34] https://en.wiktionary.org/wiki/Wiktionary:Main_Page
[35] http://semanticweb.kaist.ac.kr/home/index.php/Evaluateset3
[36] http://www-lib.tufs.ac.jp/opac/xc/openurl/search?rft.issn=0000200626
[37] https://sigmorphon.github.io/sharedtasks/2020/task1/
[38] https://github.com/kylebgorman/wikipron
[39] https://github.com/google-research-datasets/paws/tree/master/pawsx
[40] https://github.com/google-research-datasets/tydiqa
[41] https://github.com/HLTCHKUST/Xpersona

utterances and transcriptions are provided[42].

Zeroth [doc, com, rd] is an automatic speech recognition (ASR) dataset that contains approximately 50 hours of well-refined training data[43]. The speech corpus is provided free upon request and can be utilized for both research and commercial purposes.

ClovaCall [inter, acad, no] Ha et al. (2020) is an ASR dataset that consists of approximately 80 hours of telephone speech. The corpus is provided upon request, for only research purposes[44].

Pansori-TED×KR [inter, acad, rd/mod-x] Choi and Lee (2018) is an ASR dataset obtained by extracting the voices of Korean speakers from Pansori (Korean traditional song in colloquial style) and TED videos, with the transcription augmented[45]. The total reaches 3 hours, but it incorporates unique phonations that are not viable in other datasets.

ProSem [inter, com, rd] Cho et al. (2019a) is a spoken language understanding corpus for syntactic ambiguity resolution in Korean, classifying spoken utterances into 7 speech acts[46]. For about 7,100 utterances recorded by two speakers, namely a male and a female, the ground truth text and label are annotated along with the English translation.

4 Summary

In total, we surveyed 32 corpora, namely 18 Korean text corpora, 9 multilingual corpora, and 5 speech corpora. They are composed of 7 datasets on parsing and tagging, 7 datasets on entailment, paraphrasing, and summarization, 8 datasets on (spoken language) classification, QA, and dialog, 5 datasets on machine translation/transliteraion, and 5 datasets on speech (pre-)processing[47]. We provide the full specification in Table 1 in the Appendix A.

Documentation Ensuring that a curated list of resources is up-to-date is a challenge. In this regard, we aim to make our work open and canonical, as an online repository of curated resources for Korean. For the research community to have unconstrained access to all current open resources, while endorsing community contributions, the following criteria are crucial:

- The canonical, current version of this paper will be regularly published as a revision, e.g., on `arxiv.org`, based on a community-open version of this paper.
- The resources will also have a corresponding registry, following the same metadata protocol for usability in different types of research, as we used in this protocol.
- Each new contribution to the resource list will have a corresponding entry in the acknowledgments section.

We will make the registry machine parseable, so that other curated sites such as `nlpprogress.org`, can utilize the registry to automate updates. The project will be maintained as an open-source project, under a permissive license. A living document is a new territory for the field of academia, but we strongly believe that given the rapid progress of NLP research, this is an experiment worth attempting; and hope that a successful effort can inspire other languages to follow the same approach. Our approach is to be described in the public repository, guaranteeing the accessibility for domestic and abroad researchers. Also, a large portion of the data are expected to be more easily accessible via Koco[48] and Korpora[49], the recently constructed dataset wrappers for Korean NLP.

5 Conclusion

In this paper, we investigated the Korean NLP datasets constructed and released as public resources. Our curation suggests a variety of open corpora that are freely available. This information will not only be helpful for the Korean researchers who want to start NLP, but also for the abroad ones who are interested in Korean NLP. Nonetheless, we think that Korean open corpora are still less disclosed or not yet sufficient. It is notable that the Korean government is currently supplying substantial funds to build a database. To guide this well, appropriate management and documentation should be guaranteed, so that the construction is meaningful and the outcome is internationally available.

[42] `https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset`

[43] `https://github.com/goodatlas/zeroth`

[44] `https://github.com/clovaai/ClovaCall`

[45] `https://github.com/yc9701/pansori-tedxkr-corpus`

[46] `https://github.com/warnikchow/prosem`

[47] Note that these statistics do not incorporate the datasets provided by NIKL, ETRI, and AI HUB.

[48] `https://github.com/inmoonlight/koco`

[49] `https://github.com/ko-nlp/Korpora`

Acknowledgments

The authors are grateful for all the contributors of the open Korean corpora. Special thanks goes to Seungyoung Lim, Jiyeon Ham, Jiyoon Han, Hyunjoong Kim and Jihyung Moon for checking and proofreading. We also appreciate team Ko-NLP for accommodating the public repository of our project.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. pages 632–642.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. pages 1–14.

Won Ik Cho, Jeonghwa Cho, Jeemin Kang, and Nam Soo Kim. 2019a. Prosody-semantics interface in Seoul Korean: Corpus for a disambiguation of wh-intervention. In *Proceedings of the 19th International Congress of the Phonetic Sciences (ICPhS 2019)*, pages 3902–3906.

Won Ik Cho, Jong In Kim, Young Ki Moon, and Nam Soo Kim. 2020. 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*, pages 6819–6826.

Won Ik Cho, Hyeon Seung Lee, Ji Won Yoon, Seok Min Kim, and Nam Soo Kim. 2018. Speech intention understanding in a head-final language: A disambiguation utilizing intonation-dependency. *arXiv preprint arXiv:1811.04231*.

Won Ik Cho, Young Ki Moon, Sangwhan Moon, Seok Min Kim, and Nam Soo Kim. 2019b. Machines getting with the program: Understanding intent arguments of non-canonical directives. *arXiv preprint arXiv:1912.00342*.

Key-Sun Choi, Young S Han, Young G Han, and Oh W Kwon. 1994. KAIST tree bank project for Korean: Present and future development. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 7–14. Citeseer.

Yoona Choi and Bowon Lee. 2018. Pansori: ASR corpus generation from open online video contents. *arXiv preprint arXiv:1812.09798*.

Jayeol Chun, Na-Rae Han, Jena D Hwang, and Jinho D Choi. 2018. Building universal dependency treebanks in Korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. pages 2475–2485.

Kyle Gorman, Lucas FE Ashby, Aaron Goyzueta, Arya D McCarthy, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 40–50.

Jung-Woo Ha, Kihyun Nam, Jin Gu Kang, Sang-Woo Lee, Sohee Yang, Hyunhoon Jung, Eunmi Kim, Hyeji Kim, Soojin Kim, Hyun Ah Kim, et al. 2020. ClovaCall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers. *arXiv preprint arXiv:2004.09367*.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. *arXiv preprint arXiv:2004.03289*.

Chung-hye Han, Na-Rae Han, Eon-Suk Ko, Martha Palmer, and Heejong Yi. 2001. Penn Korean Treebank: Development and evaluation. In *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*, pages 69–78.

Gyeong-Eun Han, Seul-Ye Baek, and Jae-Soo Lim. 2017. Open sourced and collaborative method to fix errors of Sejong morphologically annotated corpora. In *Annual Conference on Human and Language Technology*, pages 228–232. Human and Language Technology.

Hansaem Kim. 2006. Korean national corpus in the 21st century Sejong project. In *Proceedings of the 13th NIJL International Symposium*, pages 49–54. National Institute for Japanese Language Tokyo.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron.

In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. KorQuAD 1.0: Korean QA dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020. XPersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. pages 25–31.

Tae Hwan Oh, Ji Yoon Han, Hyonsu Choe, Seokwon Park, Han He, Jinho D. Choi, Na-Rae Han, Jena D. Hwang, and Hansaem Kim. 2020. Analysis of the Penn Korean Universal Dependency treebank (PKT-UD): Manual revision to build robust parsing model in Korean. pages 122–131.

Jungyeul Park, Jeen-Pyo Hong, and Jeong-Won Cha. 2016. Korean language resources for everyone. In *Proceedings of the 30th Pacific Asia conference on language, information and computation: Oral Papers*, pages 49–58.

Kyubyong Park. 2018. KSS dataset: Korean single speaker speech dataset.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy 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.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. pages 1112–1122.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. pages 3687–3692.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. pages 1298–1308.

A Specification

The labels in Docu. denote the level of description.

- *doc*: If exists any document for the usage
- *art*: If exists any complete form of article
- *inter*: If exists a globally readable publication

Other attributes regarding license has the following order of usage and redistribution availability:

- *com* > *acad* > *unk*
- *rd* > *rd/mod-x* > *no* > *unk*

while no *unk* at this moment.

Dataset	Typical Usage	Provider	Docu.	License	Volume	Goal	Lang.
KAIST Morpho-Syntactically Annotated Corpus	Morphological analysis	Academia	art	acad/no	70M (w)	-	ko
KAIST Korean Tree-Tagging Corpus	Tree parsing	Academia	inter	acad/no	30K (s)	-	ko
UD Korean KAIST	Dependency parsing	Academia	inter	acad/rd	27K (s)	-	ko
PKT-UD	Dependency parsing	Academia	inter	acad/no	5K (s)	-	ko
KMOU NER	NER	Academia	art	acad/rd	24K (s)	-	ko
AIR×NAVER NER	NER	Competition	doc	acad/no	90K (s)	-	ko
AIR×NAVER SRL	SRL	Competition	doc	acad/no	35K (s)	-	ko
Question Pair	Paraphrase detection	Academia	doc	com/rd	10K (p)	-	ko
KorNLI	NLI	Industry	inter	com/rd	1,000K (p)	-	ko
KorSTS	STS	Industry	inter	com/rd	8,500 (p)	-	ko
ParaKQC	STS	Academia	inter	com/rd	540K (p)	-	ko
NSMC	Sentiment analysis	Academia	doc	com/rd	150K / 50K (s)	-	ko
BEEP!	Hate speech detection	Academia	inter	com/rd	8K / 500 / 1,000 (s)	-	ko
3i4K	Speech act classification	Academia	inter	com/rd	55K / 6K (s)	-	ko
KorQuAD 1.0	QA	Industry	inter	com/rd (mod-x)	60K / 5K / 4K (p)	-	ko
KorQuAD 2.0	QA	Industry	art	com/rd (mod-x)	80K / 10K / 10K (p)	-	ko
Sci-news-sum-kr	Summarization	Academia	doc	acad/rd	50 (p)	Eval	ko
sae4K	Summarization	Academia	inter	com/rd	50K (p)	-	ko
Korean Parallel Corpora	MT	Academia	inter	acad/rd (mod-x)	97K (p)	-	ko, en
KAIST Translation Evaluation Set	MT	Academia	doc	acad/no	3K (p)	Eval	ko, en

Dataset	Typical Usage	Provider	Docu.	License	Volume	Goal	Lang.
KAIST Chinese-Korean Multilingual Corpus	MT	Academia	doc	acad/no	60K (p)		ko, zh
Transliteration Dataset	Transliteration	Academia	doc	com/rd	35K (p)	-	ko, en
KAIST Transliteration Evaluation Set	Transliteration	Academia	doc	acad/no	7K (p)	Eval	ko, en
SIGMORPHON G2P	G2P conversion	Competition	inter	com/rd	3,600 / 450 / 450 (p)	-	ko, en, hy, bg, fr, ka, hi, hu, is, lt, el
PAWS-X	Paraphrase detection	Industry	inter	com/rd	5K / 2K / 2K (p)	-	ko, fr, es, de, zh, ja
TyDi-QA	QA	Industry	inter	com/rd	11K / 1,698 / 1,722 (p)	-	ko, en, ar, bn, fi, ja, id, sw, ru, te, th
XPersona	Dialog	Academia	inter	com/rd	299 (d) / 4,684 (s)	-	ko, en, it, fr, id, zh, ja
KSS	ASR	Academia	doc	acad/rd	12+ (h) / 13K (u) / 1 speaker	-	ko
Zeroth	ASR	Industry	doc	com/rd	51+ (h) / 27K (s) / 46K (u) / 181 speakers	-	ko
ClovaCall	ASR	Industry	inter	acad/no	80+ (h) / 60K (u) / 11K speakers	-	ko
Pansori-TED×KR	ASR	Academia	inter	acad/rd (mod-x)	3+ (h) / 3K (u) / 41 speakers	-	ko
ProSem	SLU	Academia	inter	com/rd	6+ (h) / 3,500 (s) / 7K (u) / 2 speakers	-	ko

Table 1: The specification on open Korean corpora. In *Provider*, *Academia* denotes universities and institutes, as well as the independent researchers who contribute to the community, while *Industry* means the companies or the research group thereof. *Competition* indicates the data used for the public competition, usually concerning both academia and industry. In *Volume*, (w) denotes words, (s) denotes sentences, (p) denotes pairs (either document or sentence pairs), (d) denotes dialogues, (h) denotes hours, and (u) denotes speech utterances. We note *Eval* only if the dataset is not for the training purpose.