KLUE: Korean Language Understanding Evaluation

Sungjoon Park[*1,4], **Jihyung Moon**[*1], **Sungdong Kim**[*2], **Won Ik Cho**[*8]
Jiyoon Han[†9], **Jangwon Park**, **Chisung Song**, **Junseong Kim**[6], **Youngsook Song**[11]
Taehwan Oh[†9], **Joohong Lee**[6], **Juhyun Oh**[†8], **Sungwon Lyu**[5], **Younghoon Jeong**[10]
Inkwon Lee[10], **Sangwoo Seo**[6], **Dongjun Lee**, **Hyunwoo Kim**[8], **Myeonghwa Lee**[2]
Seongbo Jang[6], **Seungwon Do**, **Sunkyoung Kim**[4], **Kyungtae Lim**[12], **Jongwon Lee**
Kyumin Park[4], **Jamin Shin**[7], **Seonghyun Kim**, **Lucy Park**[1]
Alice Oh[**4], **Jung-Woo Ha**[**2], **Kyunghyun Cho**[**3]

[1]Upstage, [2]NAVER AI Lab, [3]New York University, [4]KAIST,
[5]Kakao Enterprise, [6]Scatter Lab, [7]Riiid, [8]Seoul National University, [9]Yonsei University,
[10]Sogang University, [11]Kyung Hee University, [12]Hanbat National University

[*]`sungjoon.park@kaist.ac.kr`, [*]`jihyung.moon@upstage.ai`,
[*]`sungdong.kim@navercorp.com`, [*]`tsatsuki@snu.ac.kr`,
[**]`alice.oh@kaist.edu`, [**]`jungwoo.ha@navercorp.com`, [**]`kyunghyun.cho@nyu.edu`,

Abstract

We introduce Korean Language Understanding Evaluation (KLUE) benchmark. KLUE is a collection of eight Korean natural language understanding (NLU) tasks, including Topic Classification, Semantic Textual Similarity, Natural Language Inference, Named Entity Recognition, Relation Extraction, Dependency Parsing, Machine Reading Comprehension, and Dialogue State Tracking. We create all of the datasets from scratch in a principled way. We design the tasks to have diverse formats and each task to be built upon various source corpora that respect copyrights. Also, we propose suitable evaluation metrics and organize annotation protocols in a way to ensure quality. To prevent ethical risks in KLUE, we proactively remove examples reflecting social biases, containing toxic content or personally identifiable information (PII). Along with the benchmark datasets, we release pretrained language models (PLM) for Korean, KLUE-BERT and KLUE-RoBERTa, and find KLUE-RoBERTa$_{\text{LARGE}}$ outperforms other baselines including multilingual PLMs and existing open-source Korean PLMs. The fine-tuning recipes are publicly open for anyone to reproduce our baseline result. We believe our work will facilitate future research on cross-lingual as well as Korean language models and the creation of similar resources for other languages. KLUE is available at `https://klue-benchmark.com/`.

1 Introduction

A major factor behind the recent success of pretrained language models, such as BERT [30] and its variants [84, 22, 49] as well as GPT-3 [112] and its variants [113, 78, 9], has been the availability of well-designed benchmark suites for evaluating their effectiveness in natural language understanding (NLU). GLUE [135] and SuperGLUE [134] are representative examples of such suites and were

[*] Equal Contribution. A description of each author's contribution is available at the end of paper.

[**] Corresponding Authors.

[†] Work done at Upstage.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

designed to evaluate diverse aspects of NLU, including syntax, semantics and pragmatics. The research community has embraced GLUE and SuperGLUE and has made rapid progress in developing better model architectures as well as learning algorithms for NLU.

The success of GLUE and SuperGLUE has sparked interests in building standardized benchmark suites for other languages. Such efforts have been pursued along two directions. First, various groups in the world have independently created language-specific benchmark suites; a Chinese version of GLUE (CLUE [144]), a French version of GLUE (FLUE [74]), an Indonesian variant [139], an Indic version [57] and a Russian variant of SuperGLUE [127]. On the other hand, some have relied on both machine and human translation of existing benchmark suites for building multilingual versions of the benchmark suites that were created initially in English. These include XGLUE [80] and XTREME [54]. Although the latter approach scales much better than the former, the latter often fails to capture the sociocultural aspects of NLU and also introduces various artifacts arising from translation.

Hence, we build a new benchmark suite, Korean Language Understanding Evaluation (KLUE), for Korean which is the 13-th most used language in the world according to [34] but lacks a unified benchmark suite for NLU. Instead of starting from existing benchmark tasks or corpora, we build this benchmark suite from ground up by determining and collecting the base corpora, identifying a set of benchmark tasks, designing appropriate annotation protocols, and finally validating the collected annotations. This allows us to preemptively address and avoid properties that may have undesirable consequences, such as copyright infringement, annotation artifacts, social biases and privacy violations.

In summary, our contributions are:

- We build a new benchmark suite, Korean Language Understanding Evaluation (KLUE), consisting of eight NLU tasks constructed from scratch in a principled way
- We build and publicly release pretrained language models for Korean with introducing Korean-aware tokenization method.

2 KLUE Benchmark

2.1 Design Principles

We design KLUE with the following principles:

- *Covering diverse tasks and corpora*: To cover diverse aspects of language understanding, we choose eight tasks that address various task formats and domains of source corpora, including news, encyclopedia, user reviews, smart home queries and task-oriented dialogue. As Korean has distinct styles of formal and colloquial language, we explicitly include both.
- *Accessible to everyone without any restriction*: It is important for a benchmark suite to be available to everyone, such that the benchmark serves as a standard for evaluating and improving NLU systems. We thus ensure any corpora and resources in KLUE can be freely copied, redistributed, remixed and transformed for the purpose of benchmarking NLU systems.
- *Obtaining accurate and unambiguous annotations*: Ambiguity in benchmark tasks leads to less reliable evaluation, which often results in the discrepancy between the quality of an NLU system measured by the benchmark and its true quality. In order to minimize such discrepancy, we carefully design annotation guidelines of all tasks and improve them over multiple iterations, to assure accurate annotations.
- *Mitigating ethical issues in PLMs*: It has been repeatedly observed that large-scale language models often amplify the social biases in the training data [97]. We proactively remove examples from both unlabeled and labeled corpora that reflect social biases, contain toxic content or personally identifiable information (PII), both manually and automatically. Social biases are defined as overgeneralized judgment on certain individuals or groups based on social attributes (e.g., gender, ethnicity, religion). Toxic contents include insults, sexual harassment, and offensive expressions.

Table 1: Source corpora chosen for building KLUE. The top section, *News Headlines* is not protected by copyright act since they are not classified as a work due to their lack of creativity. The middle section is a collection of corpora under the permissive licenses. The bottom section, KED and Acrofan, is originally prohibited from creating derivative works, however, we release such condition by exclusive contract. For the column, *Volume*, we denote *Small* as corpus size under 1k, *Medium* as in between 1k and 50k, and *Large* as over 50k. All source corpora we collected, selection processes, and details on selected corpora are described in Appendix B.

Dataset	License	Domain	Style	Ethical Risks	Volume	Contemporary Korean
News Headlines	N/A	News (Headline)	Formal	Low	Large	o
Wikipedia	CC BY-SA 3.0	Wikipedia	Formal	Low	Large	o
Wikinews	CC BY 2.5	News	Formal	Low	Small	o
Wikitree	CC BY-SA 2.0	News	Formal	Medium	Large	o
Policy News	KOGL Type 1	News	Formal	Low	Medium	o
ParaKQC	CC BY-SA 4.0	Smart Home Utterances	Colloquial	Low	Medium	o
Airbnb Reviews	CC0 1.0	Review	Colloquial	Medium	Large	o
NAVER Sentiment Movie Corpus (NSMC)	CC0 1.0	Review	Colloquial	Medium	Large	o
Acrofan News	CC BY-SA 4.0 for KLUE-MRC by Contract	News	Formal	Low	Large	o
The Korea Economics Daily News	CC BY-SA 4.0 for KLUE-MRC by Contract	News	Formal	Low	Large	o

2.2 Source Corpora

We have actively sought corpora that are accessible, cover diverse domains and topics, and are written in modern Korean. This active search has resulted in ten sources from which we derive task-specific corpora in Table 1. These base corpora are released under CC BY(-SA) license or not considered as copyrighted work, permitting 1) derivative work, 2) redistribution, and 3) commercial use. Then we carefully preprocess them because the collected corpora came from various sources with varying levels of quality and curation. We remove noise, toxic or socially biased content, and PII, using predefined rules and machine learning models.

2.3 Considerations in Annotation

For all tasks in KLUE, we annotate a subset from the source corpora. We take into account three major considerations below:

- *Better reflection of linguistic characteristics of Korean*: Many existing Korean datasets were constructed as a part of multilingually aligned benchmarks, and they do not fully reflect linguistic characteristics of Korean such as agglutinative nature in named entity recognition (NER) [102], or tagset in part-of-speech (POS) tagging and dependency parsing (DP) [88, 46]. We write and revise annotation guidelines more appropriate for the linguistic properties of Korean.
- *Obtaining accurate annotations*: We provide crowdworkers or selected participants with a carefully designed annotation guideline and improve it over multiple iterations, in order to reduce the ambiguity in the annotation process as well as to mitigate the known artifact issues. In particular, we often filter out examples for which annotators cannot easily agree on.
- *Mitigating harmful social bias and removing PII*: To not incentivize socially biased NLU systems [7], we explicitly instruct both annotators and inspectors to manually mark and/or exclude examples that are unacceptable according to our principle of ethics. Our definitions of *bias* and *hate speech* follow Moon et al. [94]. We denote *bias* as an overgeneralized prejudice on certain groups or individuals based on the following traits: gender, race, background, nationality, ethnic group, political stance, skin color, religion, disability, age, appearance, (socio-)economic status, and occupations. In the case of *hate speech*, we include offensive, aggressive, insulting, or sarcastic contents. To deal with privacy risks, we identify a list of personally identifiable information (PII)

Table 2: Task Overview

Name	**Type**	**Format**	**Eval. Metric**	**# Class**	**{|Train|, |Dev|, |Test|}**	**Source**	**Style**
KLUE-TC (YNAT)	Topic Classification	Single Sentence Classification	Macro F1	7	45k, 9k, 9k	News (Headline)	Formal
KLUE-STS	Semantic Textual Similarity	Sentence Pair Regression	Pearson's r, F1	$[0, 5]$ 2	11k, 0.5k, 1k	News, Review, Query	Colloquial, Formal
KLUE-NLI	Natural Language Inference	Sentence Pair Classification	Accuracy	3	25k, 3k, 3k	News, Wikipedia, Review	Colloquial, Formal
KLUE-NER	Named Entity Recognition	Sequence Tagging	Entity-level Macro F1 Character-level Macro F1	6, 13	21k, 5k, 5k	News, Review	Colloquial, Formal
KLUE-RE	Relation Extraction	Single Sentence Classification (+2 Entity Spans)	Micro F1 (without *no_relation*), AUPRC	30	32k, 8k, 8k	Wikipedia, News	Formal
KLUE-DP	Dependency Parsing	Sequence Tagging (+ POS Tags)	Unlabeled Attachment Score, Labeled Attachment Score	# Words, 38	10k, 2k, 2.5k	News, Review	Colloquial, Formal
KLUE-MRC	Machine Reading Comprehension	Span Prediction	Exact Match, ROUGE-W (LCCS-based F1)	2	12k, 8k, 9k	Wikipedia, News	Formal
KLUE-DST (WoS)	Dialogue State Tracking	Slot-Value Prediction	Joint Goal Accuracy Slot Micro F1	(45)	8k, 1k, 1k	Task Oriented Dialogue	Colloquial

following KISA (Korea Internet and Security Agency) guideline,[1] whose information is related to a living individual based on personal information protection act of Korea.[2] We do not consider public figure's name as personal information.[3]

2.4 Tasks

We carefully choose the following eight tasks to cover diverse aspects of NLU in Korean while minimizing redundancy among the tasks. In Table 2, we illustrate important properties of the tasks, such as type, format, evaluation metrics, and annotated data characteristics.

Here, we list the tasks and describe why we include the task in KLUE, how we manage the construction process, and how/why we choose the evaluation metrics. For all tasks, we guide annotators to report examples that contain hate speech, biased expressions, or PII, to remove them from our benchmark[4].

KLUE-TC Topic classification (TC) is a single sentence classification task to predict the topic of a given text snippet. We include TC in our KLUE benchmark, as inferring the topic of a text is a key capability that should be possessed by a language understanding system. As a typical single sentence classification task, other NLU benchmarks such as CLUE [144] and IndicGLUE [57] also contain TNEWS and News Category Classification. For Korean, no dataset has been proposed for the task, which motivates us to construct the first Korean topic classification benchmark.

In KLUE-TC, given a news headline, a text classifier must predict a topic which is one of {politics, economy, society, culture, world, IT/science, sports}. We formulate TC as a single sentence classification task following previous works. The evaluation metric of KLUE-TC is a macro-F1 score to give the same importance to each class.

[1] `https://www.kisa.or.kr/public/laws/laws2_View.jsp?cPage=1&mode=view&p_No=282&b_No=282&d_No=3`

[2] `https://www.law.go.kr/LSW//lsInfoP.do?lsiSeq=213857&chrClsCd=010203&urlMode=engLsInfoR&viewCls=engLsInfoR#0000`

[3] See the precedent set by the Supreme Court in Korea: 대법원 2011. 9. 2. 선고 2008다42430 전원합의체 판결 available at `https://glaw.scourt.go.kr/wsjo/panre/sjo100.do?contId=2060159&q=2008%EB%8B%A442430`.

[4] All the workers were guaranteed to be paid minimum wage in Korea (about $7.5 per hour).

We collect news headlines from online articles distributed by Yonhap News Agency (YNA) and manually annotate the topics of the headlines, which is to address the gap between the headline and the predefined category. 13 selected workers labeled topics for 70,000 headlines. For each headline, 3 workers annotated topics. We filter invalid headlines and keep headlines whose topic is agreed by at least two annotators out of three, leaving 63,892 examples. For more information, see Appendix C.

KLUE-STS Semantic textual similarity (STS) is a regression task to measure the degree of semantic equivalence between two sentences. We include STS in our benchmark because it is essential to other NLP tasks such as machine translation, summarization, and question answering. Like STS [13] in GLUE [135], many NLU benchmarks include comparing semantic similarity of text snippets such as semantic similarity [144], paraphrase detection [135, 57], or word sense disambiguation [127, 74].

We formulate STS as a sentence pair regression task which predicts the semantic similarity of two input sentences as a real value. A model performance is measured by Pearson's correlation coefficient following the evaluation scheme of STS-b [13]. We additionally binarize the real numbers into two classes (paraphrased or not) with a threshold score, and use F1 score to evaluate the model.

We carefully sample and generate sentence pairs to cover all range of the similarities. For unlabeled corpora AIRBNB (colloquial review), POLICY (formal news), round-trip translation (RTT) is used to generate similar sentence pairs and greedy sentence matching (GSM) to sample less similar pairs. For labeled dataset, PARAKQC [18] (smart home utterances), we leverage the labeled intents of the commands to pair both similar and dissimilar sentences. 19 workers are employed and annotated the similarity between two sentences in integers from 0 (no meaning overlap) to 5 (meaning equivalence). We remove outlier annotations and then take average from the remaining labels for the final labels. The total number of KLUE-STS is 13,224 sentence pairs. Details are described in Appendix D.

KLUE-NLI The goal of natural language inference (NLI) is to train a model to infer the relationship between the *hypothesis* sentence and the *premise* sentence. Given a *premise*, an NLI model determines if *hypothesis* is true (entailment), false (contradiction), or undetermined (neutral). The task is also known as recognizing textual entailment (RTE) [27]. NLI datasets are also included in several NLU benchmarks such as GLUE [135] and superGLUE [134], and they are valuable as training data for other NLU tasks [24, 109, 117], which leads us to include NLI task in KLUE.

We formulate NLI as a classification task where an NLI model reads each pair of *premise* and *hypothesis* sentences and predicts whether the relationship is entailment, contradiction, or neutral. We use the classification accuracy to measure the model performance.

We construct KLUE-NLI by using a collection method similar to that of SNLI [8] and MNLI [140], while avoiding known annotation artifacts. We select premise sentences from multi-source corpora which allows to generate hypotheses. Then for each premise sentence, we ask one annotator to generate three hypothesis sentences that correspond to the three relationship classes, each. The writer is trained enough to aware of the annotation artifacts. To validate the generated pairs, we ask four additional annotators to label the relationship. The final output label is the majority of the five annotations - one original label and four additional ones. KLUE-NLI consists of 30,998 sentence pairs. In Appendix E, we compare ours against SNLI and MNLI dataset, and provide the details of the construction process and analysis.

KLUE-NER Named entity recognition (NER) is a sequence tagging task to detect the boundaries of named entities in unstructured text and identify the entity type. An entity can be a sequence of words that refers to a person, location, organization, time expression, quantity, or monetary value. Since NER is important for application fields like syntactic analysis, goal-oriented dialog systems, question and answering, and information extraction, many NLU benchmarks contain NER datasets [139, 57, 80, 54]. Despite the rise of necessity of NER datasets of various domains and styles, there are few existing Korean NER datasets to cover such needs.

In KLUE-NER, a model should detect the spans and classify the types of entities included in an input sentence. The six entity types used in KLUE-NER are person, location, organization, date, time, and quantity. We tag entity types with character-level BIO (Begin-Inside-Outside) scheme, as Korean word is mostly a combination of named entity and particle. To respect such characteristics, we evaluate a model performance using traditional entity-level F1 score and newly proposed character-level F1 score.

We choose WIKITREE (news articles) and NSMC (reviews) as the source corpora and sample sentences from them to mix the styles of written and spoken languages. Our six tag sets adapt Korean tag sets defined in Korean Telecommunications Technology Association (TTA) NER guidelines and colloquial tag sets defined MUC-7 [16]. 51 crowdworkers did the annotation first, and then two linguists validate the results. To correct erroneous annotations even after validation, six NLP researchers manually correct the annotation errors, resulting 31,008 sentences. Annotation scheme and statistics are attached in Appendix F.

KLUE-RE Relation extraction (RE) is a task to identify semantic relations between entity pairs. The relation is defined between an entity pair consisting of *subject entity* (e_{subj}) and *object entity* (e_{obj}). For example, in a sentence 'Kierkegaard was born to an affluent family in Copenhagen', the subject entity is 'Kierkegaard' and the object entity is 'Copenhagen'. The goal is then to pick an appropriate relationship between these two entities; '*place_of_birth*'. To ensure KLUE-RE captures this aspect of language understanding, we include a large-scale RE benchmark. Because there is no large-scale RE benchmark publicly available in Korean, we collect and annotate our own dataset.

We formulate RE as a single sentence classification task. A model picks one of predefined relation classes describing the relation between two entities within a given sentence. In other words, an RE model predicts an appropriate relation r of entity pair $(e_{\text{subj}}, e_{\text{obj}})$ in a sentence s, where e_{subj} is the subject entity and e_{obj} is the object entity. We refer to $(e_{\text{subj}}, r, e_{\text{obj}})$ as a relation triplet. The entities are marked as corresponding spans in each sentence s. There are 30 relation classes that consist of 18 person-related relations, 11 organization-related relations, and *no_relation*. Detailed explanation of these classes are presented in Table 13. We evaluate a model using micro F1 score, computed after excluding *no_relation*, and area under the precision-recall curve (AUPRC) including all 30 classes.

We draw sentences from WIKIPEDIA and news articles (WIKITREE and POLICY) to cover various sets of named entities and relational facts. Then we apply existing NER models to leave examples having at least two named entities, marking e_{subj} and e_{obj}, and sample sentences from those in two distinct ways. First is random sampling, which is similar to a real world scenario where the pair is highly likely to be irrelevant (*no_relation*). Second is distant supervision, which leverages Korean KB to have more chance to include relation-existing pairs. Our 30 relation classes are based on Text Analysis Conference Knowledge Base Population (TAC-KBP) [89]. We employ 163 qualified workers and assign three workers to each sentence to label the relation, taking majority-vote labels as gold labels. KLUE-RE consists of 48,001 annotated sentences. More details are in Appendix G.

KLUE-DP Dependency parsing (DP) aims to find the relations among words in a sentence. It is an important component in many NLP systems because it captures the syntactic structure of a sentence. We include DP in KLUE to evaluate the representational power of language models in terms of syntactic features.

Formally, a dependency parser predicts a graph structure of an input sentence based on the dependency grammar [29, 28]. In general, a parse tree consists of dependency arcs, connecting dependents to their heads, and the dependency labels attached to the arcs that represent the relations between dependents (DEPREL) and their heads (HEAD). Since each word in a sentence has a pair of dependency information (HEAD, DEPREL), we formulate DP as a word-level sequence tagging task. We evaluate a model's performance using unlabeled attachment score (UAS) and labeled attachment score (LAS).

For the source, we sample sentences from both WIKITREE (formal) and AIRBNB (informal). We annotate part-of-speech on the corpus in advance and use them when annotating the dependency relations. Both POS and DP are annotated and cross-validated by ten Korean PhD students who are majoring in linguistics. The final KLUE-DP consists of 14,500 sentences. The comprehensive process is demonstrated in Appendix H with dataset statistics.

KLUE-MRC Machine reading comprehension (MRC) is a task designed to evaluate a model's ability to read a given passage and then answer a question about the passage. Most existing MRC benchmarks are in English [21, 56, 60, 114, 115, 147, 152], and they are widely used in evaluating pre-trained language models for text comprehension. In Korean, however, an appropriate MRC benchmark is not available because existing Korean MRC datasets are less challenging, limited in access, or simply machine-translated from an English dataset [81, 1, 76]. We therefore include MRC in KLUE and create a new challenging Korean MRC benchmark. When building KLUE-MRC, we consider providing multiple question types, preventing reasoning shortcuts when answering to a multi-hop question, and using passage from several domains without any copyright violation.

We formulate MRC as a task of predicting an answer span of a question from a given text passage. A model input is a concatenated sequence of a question and a passage separated with a delimiter. A model output is start and end positions of a predicted answer span within a passage. If the question is unanswerable within the given passage, the model should predict the empty answer string. We evaluate a model with two metrics: 1) exact match (EM) and 2) character-level ROUGE-W. Note that character-level ROUGE-W is newly proposed character-level metric instead of character-level F1 score which have commonly used in other Korean MRC datasets. We find character-level F1 score could overestimate a model's performance as it gives score to any character overlap regardless of a sequential order.

We collect passages from Korean WIKIPEDIA and news articles provided by The Korea Economy Daily and ACROFAN. On each paragraph, annotators create questions 1) by paraphrasing a sentence in the passage, 2) requiring multi-sentence reasoning, and 3) that are unanswerable. Answers are annotated at the same time. To make question and answer pairs without having known artifacts, we prepare the guideline with specific do's and don'ts and train employed workers thoroughly. KLUE-MRC consists of 12,207 paraphrasing-based questions, 7,895 multi-sentence reasoning questions, and 9,211 unanswerable questions, for a total of 29,313 from 22,343 documents and 23,717 passages. We provide further information on each question type, statistics, and in-depth analyses in Appendix I.

KLUE-DST Dialogue State Tracking (DST) is about predicting *dialogue states* from a given task-oriented dialogue. Several recent papers have considered task-oriented dialogue (TOD) as an important problem of natural language understanding. For instance, DecaNLP [87] includes a DST, which is a key component of TOD, into one of their benchmark tasks, while DialoGLUE [90] releases the first task-oriented dialogue benchmark containing various sub-tasks including DST.

Specifically, DST is a task to predict slot (e.g. hotel type) and value (e.g. guest house, hotel, motel) pairs after each user utterance. The potential pairs are predefined by a task schema and knowledge base (KB), tied to the choice of a scenario. For evaluation, we use joint goal accuracy (JGA) and slot micro F1 score. JGA checks if all of the predicted slot-value pairs are exactly matched with the ground-truth for every turn, while the slot micro F1 computes F1 score for each slot-value pair independently. We also name this task as Wizard of Seoul (WoS).

We define a task schema and create a knowledge base, and then design an annotation system based on the schema. Then, we collect task-oriented dialogues with dialogue state annotations by following 'Self-dialog' scheme which requests a single worker to play both user and system roles [11]. Crowdworkers generate dialogues and the corresponding states by using the system. WoS contains overall 10,000 dialogues with 146,692 turns across 5 domains. Further information on the process is provided in Appendix J.

3 Experiments

In order to facilitate further research using KLUE, we provide strong baselines for all the benchmark tasks within it. As a part of this effort, we pretrain and release large-scale language models for Korean, which will reduce the burden of retraining these models from individual researchers. We also compare our models with existing multilingual pretrained language models and open-sourced Korean-specific models on the proposed KLUE benchmark.

3.1 Pretrained Language Models

Pretraining Corpora We collect publicly available data from diverse sources to cover a broad set of topics and many different styles. Having noticed quite a bit of PII and undesirable social biases in these large corpora, we pseudonymize PII while do not filter out socially biased contents nor hate speech for three reasons. First, manual inspection is infeasible. Second, it is a challenging problem on its own to automatically detect socially biased contents or hate space. Lastly, being blind to such harmful contents prevents the future use of a language model for detecting and correcting these harmful contents. Details of our choices are described in Appendix K.1.

Pretraining Korean Language Models We pretrain language models, namely KLUE-BERT and KLUE-RoBERTa, following the similar recipes of BERT [30] and RoBERTa [84], respectively. The models are trained on sequences of at most 512 tokens long with a static or dynamic masking strategy following the original training procedure. We use whole word masking (WWM) which masks all of the tokens that form a single word. We set the batch size to 256 for BERT and 2048 for RoBERTa and fix the learning rate to 10^{-4} for both. For the pretraining corpus, we gather publicly available Korean corpora of size approximately 62GB from diverse sources to cover a broad set of topics and many different styles. The most distinct part is tokenization. We use morpheme-based subword tokenization instead of BPE, considering the aggulutinative nature of Korean. The details of our pretraining are in Appendix K.

Comparison Models In addition to our own language models, we evaluate two existing multilingual language models and two Korean monolingual language models on our benchmark. For multilingual models, we employ mBERT [30] and XLM-R [26]. For the Korean models, we compare KR-BERT [77] and KoELECTRA [104]. You can find further information on each model in Appendix K.

3.2 Fine-tuning Language Models

Single Sentence Classification For **KLUE-TC**, we follow single sentence classification architecture in [30]. **KLUE-RE** on the other hand requires a special procedure to indicate entities within the input sentence. We use `<subj>`, `</subj>`, `<obj>`, and `</obj>` to mark the beginnings and the ends of subject and object entities, respectively, following Baldini Soares et al. [5].

Sentence Pair Classification and Regression For **KLUE-NLI**, a sentence pair classification task, we adopt the same approach to sentence pair classification framework suggested by Devlin et al. [30]. While for **KLUE-STS**, only the final layer is different, as **KLUE-STS** is a a regression task.

Multiple-Sentence Slot-Value Prediction **WoS** is a slot-value prediction task for a given dialogue context, where the prediction should be considered across multiple turns instead of a single utterance. We employ an encoder-decoder model following the architecture of TRADE [142], which consists of an utterance encoder, a state generator, and a slot gate classifier. In our implementation, we change the utterance encoder from GRU [17] to pretrained language model to get better representations. We also modify the slot gate classifier to predict additional two slot gate labels (*yes*, *no*), since WoS contains relatively more Boolean type slots than MultiWOZ [10]. We jointly minimize the cross-entropy loss of the state generator and slot gate classifier.

Sequence Tagging **KLUE-NER** is a subword-level tagging task and **KLUE-MRC** is a span prediction task, and in both tasks, each token is linearly mapped to a predefined label. We employ the same architecture provided in [30] and only use the given dataset for finetuning. We frame **KLUE-DP** as a sequence tagging problem. Our baseline architecture follows the model proposed in [39]. In our implementation, we use a pretrained language model to extract subword representations and concatenate the first and last subword token representations of each word, to form word vector representations, since the annotation is done at the word level. For the attention layers, we use biaffine attention [33] to predict the head, and bilinear attention [66] to predict the arc type for each word. Cross-entropy loss is minimized to tune all the parameters.

Results We present all results in Table 3 and summarize few observations. First, Korean monolingual models generally outperform multilingual models. Second, for sequence tagging tasks like KLUE-NER and KLUE-MRC, the tokenization matters. Although XLM-R_{LARGE} is on par with KLUE-$\text{BERT}_{\text{BASE}}$ on KLUE-MRC based on character-level score ROUGE, the performance gap is wider on entity-level score EM. Similar results are shown in KLUE-NER, and this demonstrates the effect of out tokenization method. Third, different models perform best on different tasks when controlled for their sizes; KLUE-BERT performs best for YNAT, KLUE-RoBERTa for KLUE-RE, KLUE-DP, KLUE-MRC and WoS, and $\text{KoELECTRA}_{\text{BASE}}$ for KLUE-STS, KLUE-NLI, and KLUE-NER. Lastly, as we increase the model size, KLUE-$\text{RoBERTa}_{\text{LARGE}}$ outperforms the other models in all tasks except for YNAT and KLUE-NER. More details of our experiments and further analyses are in Appendix L.

Table 3: Evaluation results of our pretrained LMs and other baselines on KLUE benchmark test set. The F1 refers to a macro-F1 score. The $\mathrm{F1}^E$ and $\mathrm{F1}^C$ of KLUE-NER indicates entity-level and character-level macro-F1 score, respectively. The $\mathrm{F1}^{mic}$ of KLUE-RE is micro-averaged F1 score ignoring the *no_relation*. The $\mathrm{F1}^S$ of WoS is an average of slot-value pair level micro-F1 scores. The R^P of KLUE-STS denotes Pearson correlation. **Bold** shows the best performance across the models, and underline indicates the best performance among BASE models.

	YNAT	**KLUE-STS**		**KLUE-NLI**	**KLUE-NER**		**KLUE-RE**		**KLUE-DP**		**KLUE-MRC**		**WoS**	
Model	F1	R^P	F1	ACC	$\mathrm{F1}^E$	$\mathrm{F1}^C$	$\mathrm{F1}^{mic}$	AUC	UAS	LAS	EM	ROUGE	JGA	$\mathrm{F1}^S$
mBERT$_{\textbf{BASE}}$	81.55	84.66	76.00	73.20	76.50	89.23	57.88	53.82	90.30	86.66	44.66	55.92	35.46	88.63
XLM-R$_{\textbf{BASE}}$	83.52	89.16	82.01	77.33	80.37	92.12	57.46	54.98	89.20	87.69	27.48	53.93	39.82	89.61
XLM-R$_{\textbf{LARGE}}$	**86.06**	92.97	85.86	85.93	82.27	**93.22**	58.39	61.15	92.71	**88.70**	35.99	66.77	41.20	89.80
KR-BERT$_{\textbf{BASE}}$	84.58	88.61	81.07	77.17	74.58	90.13	62.74	60.94	89.92	87.48	48.28	58.54	45.33	90.70
KoELECTRA$_{\textbf{BASE}}$	84.59	92.46	84.84	85.63	**86.11**	92.56	62.85	58.94	92.90	87.77	59.82	66.05	41.58	89.60
KLUE-BERT$_{\textbf{BASE}}$	85.73	90.85	82.84	81.63	83.97	91.39	66.44	66.17	89.96	88.05	62.32	68.51	46.64	91.61
KLUE-RoBERTa$_{\textbf{SMALL}}$	84.98	91.54	85.16	79.33	83.65	91.14	60.89	58.96	90.04	88.14	57.32	62.70	46.62	91.44
KLUE-RoBERTa$_{\textbf{BASE}}$	85.07	92.50	85.40	84.83	84.60	91.44	67.65	68.55	93.04	88.32	68.67	73.98	47.49	91.64
KLUE-RoBERTa$_{\textbf{LARGE}}$	85.69	**93.35**	**86.63**	**89.17**	85.00	91.86	**71.13**	**72.98**	**93.48**	88.36	**75.58**	**80.59**	**50.22**	**92.23**

4 Discussion

We develop KLUE with the aim of facilitating Korean NLP research, in response to the recent active development efforts of large Korean language models [63]. The entire NLP community has seen BERT [30] and its variants outperforming the previous NLU models for GLUE [135] and SuperGLUE [134], as well as the more recent GPT3 [9] with outstanding performance without fine-tuning (and with *in-context learning*) in natural language understanding and generation. Motivated by these models, many Korean researchers at various institutions rushed to pretrain large-scale Transformer-based Korean language models. Consequently, a number of nearly identical pretrained language models have been released to open-source communities. However, we could not systematically understand the behaviors and characteristics of these models because of the lack of well-designed general-purpose benchmarks like GLUE for Korean. KLUE will allow researchers to conduct controlled experiments to understand how and why various Korean LMs perform on certain tasks and thus obtain detailed insights into those models. Furthermore, since KLUE includes many representative NLU tasks that are also conducted in other languages, KLUE will function as a fundamental resource to NLP researchers who aim to conduct multilingual research with Korean and other languages.

From Scratch vs. Translation Translation has been the most straightforward approach for expanding English dataset to other languages. However, we insist to build KLUE from scratch to assure quality, considering its impact and role as the first Korean benchmark. To examine the quality difference quantitatively, we compare the correctness of sentence pairs and the corresponding labels of KLUE-NLI test set against that of KorNLI [45], a post-edited XNLI [25]. The results are shown in Table 4.

Table 4: Statistics of re-annotation results on randomly sampled 100 sentences from KorNLI and KLUE-NLI. Annotators are four native Korean undergraduates who are majoring in Korean linguistics and did not participate in the KLUE-NLI construction process. We compute the agreements of the additional labels to the gold labels for each dataset.

Statistics (n = 100)	**KorNLI**	**KLUE-NLI**
Unanimous Gold Label (4 Agree)	38.00%	**71.00%**
3 Agree with Gold Label	18.00%	24.00%
2 Agree with Gold Label	18.00%	3.00%
1 Agrees with Gold Label	16.00%	2.00%
0 Agrees with Gold Label	10.00%	0.00%
Individual Label = Gold Label	64.50%	**91.00%**
No Gold Label (No 3 Labels Match)	4.00%	**0.00%**
Majority Vote ≠ Gold Label	26.00%	**0.00%**

These numbers suggest that even human translation does not guarantee the quality. For KorNLI, annotators often report that they do not quite understand at least one of the two sentences or choose NEUTRAL because it is difficult to distinguish the semantic relationships of the sentences. On the other hand, for KLUE-NLI, there is no case where the annotators struggle to grasp the logical semantic relationship. Given that 83% of French XNLI recovers the original English consesus label [25], KorNLI seems to lost relatively more semantic meanings/relationships during the translation. This results imply that the difference in characteristics between the source and target language might have affected.

Overall Evaluation We do not average all scores gained from each task in KLUE. The performance of all tasks are measured by different evaluation metrics. This is because we carefully choose the metric for each task with considering its own characteristics. Their granularity differs by tasks, for example, KLUE-MRC and KLUE-NER employ character-level metrics because an entity can exist within a word in Korean whereas KLUE-STS and KLUE-NLI use sentence-level metrics. Furthermore, we use various metrics across tasks, such as F1 score, accuracy, AUPRC, UAS, LAS, ROUGE-W, joint goal accuracy, and Pearson's correlation. In this situation, simply computing the average of all tasks as in GLUE [135] results in misleading overall performance measure. The average will lose its interpretability as well as giving higher weights to a certain task in unintended ways. Accordingly, an alternative way to estimate a model's NLU capability is necessary. Recently, analyzing correctness of a model's prediction by using Item Response Theory (IRT) framework to estimate such capability is proposed [72], however, we find that it is not clear how it should be applied precisely in our benchmark. As of now, we thus decide to evaluate a model for each task separately without summarizing overall performance measure. This is our limitation, and we leave this problem for the future.

Ethical Consideration To secure and maximize the continued availability and usefulness of the benchmark, we include only source corpora for which we know we can release under a license that permits both redistribution and re-mix without any restriction on the use. Furthermore, we first automatically detect hate speech and gender-biased sentences using toxicity classifiers and remove those. Annotators are clearly instructed to mark any instance that exhibits social biases and/or is toxic. Finally, we manually examine these marked sentences to exclude them from the final dataset. For PII, we rely on manual inspection during annotation. We discard any sentences that was reported to contain PII after manual inspection. See Appendix M for our full statement of ethical considerations.

Broader Impact We distribute KLUE under CC BY-SA. The license allows everyone to freely copy and redistribute our benchmarks in any medium or format. In addition, one can improve our benchmark to build more challenging datasets after performance saturation. To function as a NLU *benchmark*, open access is a must. To set a good precedent for open access of data, we allow using our datasets for 1) any purpose, 2) derivative work, and 3) redistribution, as long as the existing copyrights in our benchmark datasets are respected. We also open our pretrained Korean language models and the implementation of pretraining and fine-tuning pipelines. This enhances reproducibility of our work, and allows anyone to fix and improve our data and models. We hope to contribute to the Korean NLP research community as well the wider NLP community.

5 Conclusion

We present KLUE, a suite of Korean NLU benchmarks that includes eight diverse tasks. KLUE is available for everyone, along with Korean language models trained to outperform multilingual models and other existing open-sourced Korean language models. We set high standards from the outset, as building the benchmark and training the models from scratch. We designed the benchmark datasets and trained the annotators rigorously to consider potential ethical issues including private information and hate speech. We documented in detail all of the benchmark construction and testing processes. We also discussed broader impacts and limitations of KLUE and our models. Despite the limitations, KLUE and the accompanying language models will facilitate future Korean NLP research by setting a valuable precedent describing how datasets and language models should be created and spread to a wider community.

Acknowledgments and Disclosure of Funding

Data annotation costs were provided by Upstage, NAVER CLOVA, Scatter Lab, SelectStar, Riiid!, DeepNatural and KAIST. The leaderboard is built and supported by Upstage. GPU cloud computing is provided by NAVER CLOVA NSML [64], Google TensorFlow Research Cloud (TFRC), and Kakao Enterprise BrainCloud. These three computing resources were used to pretrain and fine-tune the language models. News articles for the MRC datasets were provided by the Korea Economy Daily and Acrofan.

The authors thank Cheoneum Park for discussions about task selection and DP task, Jinhyuk Lee and Minjoon Seo for discussions on MRC task, Sujeong Kim and DongYeon Kim for considerable efforts to manage the annotation for MRC dataset, and Sangah Park for careful consideration of data construction in DP, NER, and RE. We thank Junyeop Lee, Geonhee Lee, Jiho Lee, Daehyun Nam, and Yongjin Cho for the leaderboard and the evaluation system.

This study is reviewed and approved by the KAIST Institutional Review Board (#KH2020-173).

References

[1] National Information Society Agency. MRC AI Dataset. https://aihub.or.kr/aidata/86, 2018.

[2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2045. URL https://www.aclweb.org/anthology/S15-2045.

[3] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1081. URL https://www.aclweb.org/anthology/S16-1081.

[4] Jens Allwood. An activity based approach to pragmatics. In Harry Bunt and William Black, editors, *Abduction, belief and context in dialogue: Studies in computational pragmatics*, chapter 2, pages 47–80. John Benjamins, Amsterdam, Netherlands, 2000.

[5] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1279. URL https://www.aclweb.org/anthology/P19-1279.

[6] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://www.aclweb.org/anthology/Q18-1041.

[7] Samuel R Bowman and George E Dahl. What will it take to fix benchmarking in natural language understanding? *arXiv preprint arXiv:2104.02145*, 2021.

[8] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://www.aclweb.org/anthology/D15-1075.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

[10] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL `https://www.aclweb.org/anthology/D18-1547`.

[11] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4516–4525, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1459. URL `https://www.aclweb.org/anthology/D19-1459`.

[12] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020.

[13] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL `https://www.aclweb.org/anthology/S17-2001`.

[14] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35, November 2017. ISSN 1931-0145. doi: 10.1145/3166054.3166058. URL `https://doi.org/10.1145/3166054.3166058`.

[15] Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. FairFil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=N6JECD-PI5w`.

[16] Nancy A. Chinchor. Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*, 1998. URL `https://www.aclweb.org/anthology/M98-1001`.

[17] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL `https://www.aclweb.org/anthology/D14-1179`.

[18] Won Ik Cho, Jong In Kim, Young Ki Moon, and Nam Soo Kim. 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, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://www.aclweb.org/anthology/2020.lrec-1.842`.

[19] Won Ik Cho, Sangwhan Moon, and Youngsook Song. Open Korean corpora: A practical report. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 85–93, Online, November 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.nlposs-1.12`.

[20] Jayeol Chun, Na-Rae Han, Jena D. Hwang, and Jinho D. Choi. Building Universal Dependency treebanks in Korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://www.aclweb.org/anthology/L18-1347`.

[21] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL `https://www.aclweb.org/anthology/N19-1300`.

[22] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/pdf?id=r1xMH1BtvB`.

[23] Korea Copyright Commission. Newspapers and copyright, 2009.

[24] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL `https://www.aclweb.org/anthology/D17-1070`.

[25] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL `https://www.aclweb.org/anthology/D18-1269`.

[26] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://www.aclweb.org/anthology/2020.acl-main.747`.

[27] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-33428-6.

[28] Marie-Catherine de Marneffe and Christopher D. Manning. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL `https://www.aclweb.org/anthology/W08-1301`.

[29] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2006/pdf/440_pdf.pdf`.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

[31] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf`.

[32] William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL `https://www.aclweb.org/anthology/I05-5002`.

[33] Timothy Dozat and Christopher D. Manning. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=Hk95PK9le`.

[34] David M. Eberhard and Charles D. Simons, Gary F. Fenning. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 24 edition, 2021. URL `http://www.ethnologue.com`.

[35] Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5526. URL `https://www.aclweb.org/anthology/W17-5526`.

[36] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5506. URL `https://www.aclweb.org/anthology/W17-5506`.

[37] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://www.aclweb.org/anthology/2020.lrec-1.53`.

[38] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL `https://www.aclweb.org/anthology/P19-1346`.

[39] Daniel Fernández-González and Carlos Gómez-Rodríguez. Left-to-right dependency parsing with pointer networks. 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 710–716, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1076. URL `https://www.aclweb.org/anthology/N19-1076`.

[40] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages

363–370, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219885. URL https://www.aclweb.org/anthology/P05-1045.

[41] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[42] Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2103. URL https://www.aclweb.org/anthology/P18-2103.

[43] Ralph Grishman and Beth Sundheim. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. URL https://www.aclweb.org/anthology/C96-1079.

[44] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://www.aclweb.org/anthology/N18-2017.

[45] Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.39. URL https://www.aclweb.org/anthology/2020.findings-emnlp.39.

[46] Ji Yoon Han, Tae Hwan Oh, Lee Jin, and Hansaem Kim. Annotation issues in Universal Dependencies for Korean and Japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 99–108, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.udw-1.12.

[47] Na-Rae Han, Shijong Ryu, Sook-Hee Chae, Seung-yun Yang, Seunghun Lee, and Martha Palmer. Korean treebank annotations version 2.0. *Linguistic Data Consortium (LDC), Philadelphia*, 2006.

[48] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1514. URL https://www.aclweb.org/anthology/D18-1514.

[49] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

[50] Matthew Henderson, Blaise Thomson, and Jason D. Williams. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A., June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4337. URL https://www.aclweb.org/anthology/W14-4337.

[51] Matthew Henderson, Blaise Thomson, and Jason D Williams. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329. IEEE, 2014.

[52] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/S10-1006`.

[53] Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.314. URL `https://www.aclweb.org/anthology/2020.findings-emnlp.314`.

[54] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR, 13–18 Jul 2020. URL `http://proceedings.mlr.press/v119/hu20b.html`.

[55] Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. A French corpus and annotation schema for named entity recognition and relation extraction of financial news. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2293–2299, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://www.aclweb.org/anthology/2020.lrec-1.279`.

[56] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL `https://www.aclweb.org/anthology/P17-1147`.

[57] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.445. URL `https://www.aclweb.org/anthology/2020.findings-emnlp.445`.

[58] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[59] J. F. Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41, January 1984. ISSN 1046-8188. doi: 10.1145/357417.357420. URL `https://doi.org/10.1145/357417.357420`.

[60] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1023. URL `https://www.aclweb.org/anthology/N18-1023`.

[61] Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, et al. ParsiNLU: a suite of language understanding challenges for persian. *arXiv preprint arXiv:2012.06154*, 2020.

[62] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in

multimodal memes. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf.

[63] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. What changes can large-scale language models bring? intensive study on billions-scale korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[64] Hanjoo Kim, Minkyu Kim, Dongjoo Seo, Jinwoong Kim, Heungseok Park, Soeun Park, Hyunwoo Jo, KyungHyun Kim, Youngil Yang, Youngkwan Kim, et al. Nsml: Meet the mlaas platform with a real-world case study. *arXiv preprint arXiv:1810.09957*, 2018.

[65] Youngmin Kim, Seungyoung Lim, Hyunjeong Lee, Soyoon Park, and Myungji Kim. KorQuAD 2.0: Korean QA dataset for web document machine comprehension. *Journal of KIISE*, 47:577–586, 2020. ISSN 2383-630X. URL https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=NART99691770&dbt=NART.

[66] Eliyahu Kiperwasser and Yoav Goldberg. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327, 2016. doi: 10.1162/tacl_a_00101. URL https://www.aclweb.org/anthology/Q16-1023.

[67] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075150. URL https://www.aclweb.org/anthology/P03-1054.

[68] Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.84. URL https://www.aclweb.org/anthology/2020.emnlp-main.84.

[69] K. Krippendorff. Computing Krippendorff's alpha-reliability. 2011.

[70] Taku Kudo. MeCab: Yet another part-of-speech and morphological analyzer, 2006. URL https://taku910.github.io/mecab/.

[71] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, March 2019. doi: 10.1162/tacl_a_00276. URL https://www.aclweb.org/anthology/Q19-1026.

[72] John P. Lalor and Hong Yu. Dynamic data selection for curriculum learning via ability estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 545–555, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.48. URL https://www.aclweb.org/anthology/2020.findings-emnlp.48.

[73] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

[74] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://www.aclweb.org/anthology/2020.lrec-1.302`.

[75] Junbum Lee. KcBERT: Korean comments BERT. In *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, pages 437–440, 2020.

[76] Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://www.aclweb.org/anthology/L18-1437`.

[77] Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. KR-BERT: A small-scale Korean-specific language model. *arXiv preprint arXiv:2008.03979*, 2020.

[78] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL `https://www.aclweb.org/anthology/2020.acl-main.703`.

[79] Shiyang Li, Semih Yavuz, Kazuma Hashimoto, Jia Li, Tong Niu, Nazneen Rajani, Xifeng Yan, Yingbo Zhou, and Caiming Xiong. CoCo: Controllable counterfactuals for evaluating dialogue state trackers. *arXiv preprint arXiv:2010.12850*, 2020.

[80] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.484. URL `https://www.aclweb.org/anthology/2020.emnlp-main.484`.

[81] Seungyoung Lim, Myungji Kim, and Jooyoul Lee. KorQuAD1.0: Korean QA dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*, 2019.

[82] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W04-1013`.

[83] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. On-the-fly controlled text generation with experts and anti-experts, 2021.

[84] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[85] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

[86] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL `https://www.aclweb.org/anthology/J93-2004`.

[87] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.

[88] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 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, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P13-2017`.

[89] Paul McNamee and Hoa Trang Dang. Overview of the TAC 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, volume 17, pages 111–113, 2009.

[90] Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. DialoGLUE: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*, 2020.

[91] Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.212. URL `https://www.aclweb.org/anthology/2020.acl-main.212`.

[92] Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1416. URL `https://www.aclweb.org/anthology/P19-1416`.

[93] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P09-1113`.

[94] Jihyung Moon, 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*, pages 25–31, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.socialnlp-1.4. URL `https://www.aclweb.org/anthology/2020.socialnlp-1.4`.

[95] Sangha Nam, Minho Lee, Donghwan Kim, Kijong Han, Kuntae Kim, Sooji Yoon, Eunkyung Kim, and Key-Sun Choi. Effective crowdsourcing of multiple tasks for comprehensive knowledge extraction. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 212–219, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://www.aclweb.org/anthology/2020.lrec-1.27`.

[96] Nikita Nangia and Samuel R. Bowman. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1449. URL `https://www.aclweb.org/anthology/P19-1449`.

[97] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL `https://www.aclweb.org/anthology/2020.emnlp-main.154`.

[98] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated MAchine reading COmprehension dataset. November 2016. URL `https://www.microsoft.com/en-us/research/publication/ms-marco-human-generated-machine-reading-comprehension-dataset/`.

[99] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL `https://www.aclweb.org/anthology/L16-1262`.

[100] National Institute of Korean Languages. NIKL CORPORA 2020 (v.1.0), 2020. URL `https://corpus.korean.go.kr`.

[101] Tae Hwan Oh, Ji Yoon Han, Hyonsu Choe, Seokwon Park, Han He, Jinho D. Choi, Na-Rae Han, Jena D. Hwang, and Hansaem Kim. Analysis of the Penn Korean Universal Dependency treebank (PKT-UD): Manual revision to build robust parsing model in Korean. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 122–131, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwpt-1.13. URL `https://www.aclweb.org/anthology/2020.iwpt-1.13`.

[102] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1178. URL `https://www.aclweb.org/anthology/P17-1178`.

[103] Eunjeong L. Park and Sungzoon Cho. KoNLPy: Korean natural language processing in Python. In *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, Chuncheon, Korea, October 2014.

[104] Jangwon Park. KoELECTRA: Pretrained ELECTRA model for Korean. `https://github.com/monologg/KoELECTRA`, 2020.

[105] Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China, December 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.aacl-main.17`.

[106] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://www.aclweb.org/anthology/N18-1202`.

[107] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. KILT: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2020.

[108] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf`.

[109] Jason Phang, Thibault Févry, and Samuel R Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018.

[110] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL `https://www.aclweb.org/anthology/S18-2023`.

[111] Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.67. URL `https://www.aclweb.org/anthology/2020.emnlp-main.67`.

[112] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[113] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

[114] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://www.aclweb.org/anthology/D16-1264`.

[115] Pranav Rajpurkar, Robin Jia, and Percy Liang. 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)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL `https://www.aclweb.org/anthology/P18-2124`.

[116] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696, Apr. 2020. doi: 10.1609/aaai.v34i05.6394. URL `https://ojs.aaai.org/index.php/AAAI/article/view/6394`.

[117] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL `https://www.aclweb.org/anthology/D19-1410`.

[118] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15939-8.

[119] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D11-1141`.

[120] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

[121] Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. *arXiv preprint arXiv:2012.15613*, 2020.

[122] Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1156. URL `https://www.aclweb.org/anthology/P18-1156`.

[123] Martin Schiersch, Veselina Mironova, Maximilian Schmitt, Philippe Thomas, Aleksandra Gabryszak, and Leonhard Hennig. A German corpus for fine-grained named entity recognition and relation extraction of traffic and industry events. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://www.aclweb.org/anthology/L18-1703`.

[124] Priyanka Sen and Amir Saffari. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.190. URL `https://www.aclweb.org/anthology/2020.emnlp-main.190`.

[125] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL `https://www.aclweb.org/anthology/P16-1162`.

[126] Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*, 2018.

[127] Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.381. URL `https://www.aclweb.org/anthology/2020.emnlp-main.381`.

[128] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D13-1170`.

[129] Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL `https://www.aclweb.org/anthology/W16-3919`.

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

[131] Erik F. Tjong Kim Sang and Fien De Meulder. 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*, pages 142–147, 2003. URL `https://www.aclweb.org/anthology/W03-0419`.

[132] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada,

August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL <https://www.aclweb.org/anthology/W17-2623>.

[133] Clara Vania, Ruijie Chen, and Samuel R. Bowman. Asking Crowdworkers to Write Entailment Examples: The Best of Bad options. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 672–686, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.aacl-main.68.

[134] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[135] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

[136] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, March 2019. doi: 10.1162/tacl_a_00290. URL https://www.aclweb.org/anthology/Q19-1040.

[137] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1042.

[138] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.494.

[139] Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.aacl-main.85.

[140] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://www.aclweb.org/anthology/N18-1101.

[141] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[142] Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1078. URL https://www.aclweb.org/anthology/P19-1078.

[143] Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. A discourse-level named entity recognition and relation extraction dataset for Chinese literature text. *arXiv preprint arXiv:1711.07010*, 2017.

[144] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.419. URL https://www.aclweb.org/anthology/2020.coling-main.419.

[145] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1237. URL https://www.aclweb.org/anthology/D15-1237.

[146] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1382. URL https://www.aclweb.org/anthology/D19-1382.

[147] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL <https://www.aclweb.org/anthology/D18-1259>.

[148] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

[149] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1074. URL https://www.aclweb.org/anthology/P19-1074.

[150] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://www.aclweb.org/anthology/Q14-1006.

[151] Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.444. URL https://www.aclweb.org/anthology/2020.acl-main.444.

[152] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*, 2018.

[153] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf`.

[154] Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1131. URL `https://www.aclweb.org/anthology/N19-1131`.

[155] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1004. URL `https://www.aclweb.org/anthology/D17-1004`.

[156] Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295, 2020. doi: 10.1162/tacl_a_00314. URL `https://www.aclweb.org/anthology/2020.tacl-1.19`.