

CLUE: A Chinese Language Understanding Evaluation Benchmark

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, Yiming Cui, Cong Yu, Qianqian Dong, Yin Tian, Dian Yu, Bo Shi, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaowei Hua Liu, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang and Zhenzhong Lan *

CLUE Team
CLUE@CLUEbenchmarks.com

Abstract

We introduce CLUE, a Chinese Language Understanding Evaluation benchmark. It contains eight different tasks, including single-sentence classification, sentence pair classification, and machine reading comprehension. We evaluate CLUE on a number of existing full-network pre-trained models for Chinese. We also include a small hand-crafted diagnostic test set designed to probe specific linguistic phenomena using different models, some of which are unique to Chinese. Along with CLUE, we release a large clean crawled raw text corpus that can be used for model pre-training. We release CLUE, baselines, pre-training dataset on Github¹

1 Introduction

Full-network pre-training methods such as BERT (Devlin et al., 2019) and their improved versions (Yang et al., 2019; Liu et al., 2019; Lan et al., 2019) have led to significant performance boosts across many natural language understanding (NLU) tasks. One key driving force behind such improvements and rapid iterations of models is the general use of evaluation benchmarks. These benchmarks use a single metric to evaluate the performance of models across a wide range of tasks. However, existing language evaluation benchmarks are mostly in English, e.g., GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019). To the best of our knowledge, there is no general language understanding evaluation benchmark for Chinese, whose speakers account for one-fourth of the world’s population, and one of the official languages of the United Nations. Also, Chinese is linguistically very different from English and other Indo-European languages, which necessitates an evaluation benchmark specifically designed for Chinese. Without such a benchmark,

it would be difficult for researchers in the field to check how good their Chinese language understanding models are.

To address this problem and facilitate studies in Chinese language, we introduce a comprehensive Chinese Language Understanding Evaluation (CLUE) benchmark that contains a collection of eight different natural language understanding tasks, including semantic similarity, natural language inference, short text classification, long text classification with large number of classes, and different types of machine reading comprehension. To better understand the challenges posed by these tasks, we evaluate them using several popular pre-trained language understanding models for Chinese. Overall, we find that these tasks display different levels of difficulty, manifest in different accuracies across models, as well as the comparison between human and machine performance.

The size and quality of unlabeled corpora play an essential role in language model pre-training (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Lan et al., 2019). There are already popular pre-training corpora such as Wikipedia and the Toronto Book Corpus (Zhu et al., 2015) in English. However, we are not aware of any large-scale open-source pre-training dataset in Chinese; and different Chinese models are trained on different and relatively small corpora. Therefore, it is difficult to compare performance across model architectures. This difficulty motivates us to construct and release a standard CLUE pre-training dataset: a corpus with over 214 GB of raw text and roughly 76 billion Chinese words. We also introduce a diagnostic dataset hand-crafted by linguists. Similar to GLUE, this dataset is designed to highlight linguistic common knowledge and logical operators that we expect models to handle well.

Overall, we present in this paper: (1) A Chinese natural language understanding benchmark

* Corresponding author. E-mail: lansysu@gmail.com

¹<https://github.com/CLUEbenchmark/CLUE>

that covers a variety of sentence classification and machine reading comprehension tasks, at different levels of difficulty, in different sizes and forms. (2) A large-scale raw corpus for general-purpose pre-training in Chinese so that the comparisons across different model architectures are as meaningful as possible. (3) A diagnostic evaluation dataset developed by linguists containing multiple linguistic phenomena, some of which are unique to Chinese. (4) A user-friendly toolkit, as well as an online leaderboard with an auto-evaluation system, supporting all our evaluation tasks and models, with which researchers can reproduce experimental results and compare the performance of different submitted models easily.

2 Related Work

It has been a common practice to evaluate language representations on different intrinsic and downstream linguistic tasks. For example, Mikolov et al. (2013) measure word embeddings through a semantic analogy task and a syntactic analogy task. Pennington et al. (2014) further expands the testing set to include other word similarity and named entity recognition tasks. Similar patterns have happened in sentence representations (Kiros et al., 2015). However, as different people use different evaluation pipelines on different datasets, results reported in different papers are not always fully comparable. Especially in the case where the datasets are small, a minor change in evaluation can lead to big differences in outcomes.

SentEval (Conneau and Kiela, 2018) addresses the above problem by introducing a standard evaluation pipeline on a set of popular sentence embedding evaluation datasets. GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) further improve SentEval along the line of getting consistent reported results for natural language understanding tasks. They introduce a set of more difficult datasets and a model-agnostic evaluation pipeline. Along with other reading comprehension tasks like SQuAD (Rajpurkar et al., 2016) and RACE (Lai et al., 2017), GLUE and SuperGLUE have become standard testing benchmarks for full-network pre-training methods such as BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2019).

We believe a similar problem exists in Chinese language understanding evaluation. Although more and more Chinese linguistic tasks (Liu et al., 2018; Cui et al., 2019b) have been proposed, there is still

a need for a standard evaluation pipeline and an evaluation benchmark with a set of diverse and difficult language understanding tasks.

3 CLUE Overview

CLUE consists of 1) eight language understanding tasks in Chinese, 2) a large-scale raw dataset for pre-training and a small hand-crafted diagnostic dataset for linguistic analysis, and 3) a ranking system, a leaderboard and a toolkit.

3.1 Task Selection

For this benchmark, we selected eight different tasks, to make sure this benchmark could test different aspects of pre-trained models. To ensure the quality and coverage of the language understanding tasks, we select tasks using the following criteria:

Diversity The tasks in CLUE should vary in terms of the task, the size of the text, the type of understanding required, the number of training examples, etc.

Well-defined and easy-to-process We select tasks that are well-defined, and we pre-process them for our users so that they can focus on modeling.

Moderate difficulty: challenging but solvable To be included in CLUE, a task should not be too simple or already solved so as to encourage researchers to design better models (e.g., multiple-choice machine reading comprehension task).

Representative and useful Our tasks should be representative of common language understanding tasks, easily applicable to real-world situations (e.g., classification task with lots of labels, or semantic similarity task).

Tailor to Chinese-specific characteristics Ideally, tasks should measure the ability of models to handle Chinese-specific linguistic phenomena (e.g., four-character idioms).

Following the above standards, we select eight tasks covering a wide range of tasks for Chinese. These include three single-sentence tasks, three sentence-pair tasks, and three machine reading comprehension tasks.

3.2 Large-scale Pre-Training Dataset

We collect data from the internet and preprocess them to make a large pre-training dataset for Chinese language processing researchers. In the end,

Corpus	Train	Dev	Test	Task	Metric	Source
Single-Sentence Tasks						
TNEWS	53.3k	10k	10k	short text classification	acc.	news title and keywords
IFLYTEK	12.1k	2.6k	2.6k	long text classification	acc.	app descriptions
CLUEWSC2020	1,244	304	290	coreference resolution	acc.	Chinese fiction books
Sentence Pair Tasks						
AFQMC	34.3k	4.3k	3.9k	semantic similarity	acc.	online customer service
CSL	20k	3k	3k	keyword recognition	acc.	CNKI
Machine Reading Comprehension Tasks						
CMRC 2018	10k	3.4k	4.9k	answer span extraction	EM.	Wikipedia
ChID	577k	23k	23k	multiple-choice, idiom	acc.	novel, essay, and news
C ³	11.9k	3.8k	3.9k	multiple-choice, free-form	acc.	mixed-genre

Table 1: Task descriptions and statistics. TNEWS has 15 classes; IFLYTEK has 119 classes; all other classification tasks are binary classification.

a total of 14 GB raw corpus with around 5 billion Chinese words are collected in our pre-training corpus (see Section 5 for details).

3.3 Diagnostic Dataset

In order to measure how well models are doing on specific language understanding phenomena, we handcraft a diagnostic dataset that contains nine linguistic and logic phenomena (details in Section 7).

3.4 Leaderboard

We also provide a leaderboard for users to submit their own results on CLUE. The evaluation system will give final scores for each task when users submit their predicted results. To encourage reproducibility, we mark the score of a model as “certified” if it is open-source, and we can reproduce the results.

3.5 Toolkit

To make it easier for using the CLUE benchmark, we also offer a toolkit named PyCLUE implemented in TensorFlow (Abadi et al., 2016). PyCLUE supports mainstream pre-training models and a wide range of target tasks. Different from existing pre-training model toolkits (Wolf et al., 2019; Zhao et al., 2019), PyCLUE is designed with a goal of quick model performance validations on the CLUE benchmark. We implement many pre-training model baselines on the CLUE benchmark and provide interfaces to support the evaluation of users’ custom models.

4 Tasks

CLUE has eight Chinese NLU tasks, covering single sentence classification, sentence pair classification, and machine reading comprehension. Descriptions of these tasks are shown in Table 1, and examples of these are shown in Table 5 Development set examples in the Appendix.

4.1 Single Sentence Tasks

TNEWS TouTiao Text Classification for News Titles² consists of Chinese news published by TouTiao before May 2018, with a total of 73,360 titles. Each title is labeled with one of 15 news categories (finance, technology, sports, etc.) and the task is to predict which category the title belongs to. To make the dataset more discriminative, we use cross-validation to filter out some of the easy examples (see Section D Dataset Filtering in the Appendix for details). We then randomly shuffle and split the whole dataset into a training set, development set and test set.

IFLYTEK IFLYTEK (IFLYTEK CO., 2019) contains 17,332 app descriptions. The task is to assign each description into one of 119 categories, such as food, car rental, education, etc. A data filtering technique similar to the one used for the TNEWS dataset has been applied.

CLUEWSC2020 The Chinese Winograd Schema Challenge dataset³ is an anaphora/coreference resolution task where the model is asked to decide whether a pronoun

²<https://github.com/fatecbf/toutiao-text-classification-dataset/>

³<https://github.com/CLUEbenchmark/LightLM>

and a noun (phrase) in a sentence co-refer (binary classification), built following similar datasets in English (e.g. Levesque et al., 2012; Wang et al., 2019). Sentences in the dataset are hand-picked from 36 contemporary literary works in Chinese. Their anaphora relations are then hand-annotated by linguists, amounting to 1,838 questions in total. Details of the dataset will be updated in <https://www.cluebenchmarks.com/>.

4.2 Sentence Pair Tasks

Tasks in this section ask a model to predict relations between sentence pairs, or abstract-keyword pairs.

AFQMC Ant Financial Question Matching Corpus⁴ comes from Ant Technology Exploration Conference (ATEC) Developer competition. It is a binary classification task that aims to predict whether two sentences are semantically similar.

CSL Chinese Scientific Literature dataset contains Chinese paper abstracts and their keywords from core journals of China, covering multiple fields of natural sciences and social sciences. We generate fake keywords through tf-idf and mix them with real keywords. Given an abstract and some keywords, the task is to tell whether the keywords are all original keywords of a paper. It mainly evaluates the ability of models to judge whether keywords can summarize the document.

4.3 Machine Reading Comprehension

CMRC 2018 CMRC 2018 (Cui et al., 2019b) is a span-extraction based dataset for Chinese machine reading comprehension. This dataset contains about 19,071 human-annotated questions from Wikipedia paragraphs. In CMRC 2018, all samples are composed of contexts, questions, and related answers. Furthermore, the answers are the text spans in contexts.

ChID ChID (Zheng et al., 2019) is a large-scale Chinese IDiom cloze test dataset, which contains about 498,611 passages with 623,377 blanks covered from news, novels, and essays. The candidate pool contains 3,848 Chinese idioms. For each blank in the passage, there are ten candidate idioms with one golden option, several similar idioms, and others are randomly chosen from the dictionary.

C³ C³ (Sun et al., 2019b) is the first free-form multiple-choice machine reading comprehension

dataset for Chinese. Given a document, either a dialogue or a more formally written mixed-genre text, and a free-form question that is not limited to a single question type (e.g., yes/no questions), the task is to select the correct answer option from all (2 to 4) options associated with the corresponding question. We employ all of the 19,577 general domain problems for 13,369 documents and follow the original data splitting. These problems are collected from language exams carefully designed by educational experts for evaluating the reading comprehension ability of language learners, similar to its English counterparts RACE (Lai et al., 2017) and DREAM (Sun et al., 2019a).

5 Pre-Training Dataset

Large-scale language data is the prerequisite for model pre-training. Corpora of various sizes have been compiled and utilized in English, e.g., the Wikipedia Corpus, the BooksCorpus (Zhu et al., 2015), and more recent C4 corpus (Colin Raffel, 2019).

For Chinese, however, existing public pre-training datasets are much smaller than the English datasets. For example, the Wikipedia dataset in Chinese only contains around 1.1 GB raw text. We thus collect a large-scale clean crawled Chinese corpus to fill this gap.

A total of 214 GB raw corpus with around 76 billion words are collected. It contains text from three parts, including CLUECorpus2020-small, CLUECorpus2020, and CLUEOSCAR.

CLUECorpus2020-small It contains 14 GB Chinese corpus, consisting of four sub-parts: News, WebText, Wikipedia and Comments. The details are as follows:

- **News** This sub-corpus is crawled from the We Media (self-media) platform, with a total of 2.5 million news articles from roughly 63K sources. It is around 8 GB raw corpus with 3 billion Chinese words.
- **WebText** With 4.1 million questions and answers, the WebText sub-corpus is crawled from Chinese Reddit-like websites such as Wukong QA, Zhihu, Sogou Wenwen, etc. Only answers with three or more upvotes are included to ensure the quality of the text.
- **Wikipedia** This sub-corpus is gathered from the Chinese contents on Wikipedia (Chinese

⁴<https://dc.cloud.alipay.com/index#/topic/intro?id=3>

		Single Sentence			Sentence Pair		MRC		
Model	Avg	TNEWS	IFLYTEK	CLUEWSC2020	AFQMC	CSL	CMRC 2018	ChID	C ³
BERT-base	68.83	56.58	60.29	63.45	73.70	80.36	69.72	82.04	64.50
BERT-wwm-ext-base	69.75	56.84	59.43	62.41	74.07	80.63	73.23	82.90	68.50
ALBERT-tiny	54.87	53.35	48.71	63.38	69.92	74.56	53.68	43.53	31.86
ALBERT-xxlarge	71.84	<u>59.46</u>	62.89	61.54	75.6	<u>83.63</u>	75.15	83.15	<u>73.28</u>
ERNIE-base	69.17	58.33	58.96	63.44	73.83	79.10	73.32	82.28	64.10
XLNet-mid	68.07	56.24	57.85	61.04	70.50	81.26	66.51	83.47	67.68
RoBERTa-large	70.29	57.86	62.55	62.44	74.02	81.36	76.11	84.50	63.44
RoBERTa-wwm-ext-base	70.72	56.94	60.31	72.07	74.04	81.00	73.89	83.62	63.90
RoBERTa-wwm-ext-large	<u>74.49</u>	58.61	<u>62.98</u>	<u>81.38</u>	<u>76.55</u>	82.13	<u>76.58</u>	<u>85.37</u>	72.32
Human	86.23	71.00	80.00	98.00	81.0	84.0	92.40	87.10	96.00

Table 2: Performance of Baseline Models on CLUE benchmark. For results of newly submitted models, including NEZHA-large(Huawei Noah’s Ark lab), ALBERT-xxlarge(Alibaba PAI) and UER(UER), check the leaderboard (<http://www.CLUEbenchmark.com>). Avg is the average of all tasks. **Bold** text denotes the best result in each column. Underline indicates the best result for the models. We report EM for CMRC 2018 and accuracy for all other tasks.

Wikipedia), containing around 1.1 GB raw texts with 0.4 billion Chinese words on a wide range of topics.

- **Comments** These comments are collected from E-commerce websites including Dianping.com and Amazon.com by SophonPlus⁵. This subset has approximately 2.3 GB of raw texts with 0.8 billion Chinese words.

CLUECorpus2020 (Xu et al., 2020) It contains 100 GB Chinese raw corpus, which is retrieved from Common Crawl. It is a well-defined dataset that can be used directly for pre-training without requiring additional pre-processing. CLUECorpus2020 contains around 29K separate files with each file following the pre-training format for the training set.

CLUEOSCAR⁶ OSCAR is a huge multilingual corpus obtained by language classification and filtering of the Common Crawl corpus. It contains 250 GB Chinese raw corpus. We do further filtering and finally get 100 GB Chinese corpus.

6 Experiments

6.1 Baselines

For baselines, we evaluate a variety of pre-trained models on the CLUE tasks. We implement models in both the TensorFlow library (Abadi et al., 2016) and PyTorch library (Paszke et al., 2019). Original code for these baselines will be made available at GitHub repository.

⁵<https://github.com/SophonPlus/ChineseNlpCorpus/>

⁶<https://dumps.wikimedia.org/zhwiki/latest/>

Architecture Our baseline architecture combines pre-trained models and fine-tunes the CLUE tasks with one additional output layer. For single-sentence tasks, we encode the sentence and then pass the pooled output to a classifier. For sentence-pair tasks, we encode sentence pairs with a separator and then pass the pooled output to a classifier. As for the extraction-style and multi-choice style for machine reading comprehension tasks, we use two fully connected layers after the pooled output to predict the start and end position of the answer for the former. For the latter, we encode multiple candidate-context pairs to a shared classifier and get corresponding scores.

Models We evaluate CLUE on the following public available pre-trained models:

- BERT-base, a BERT model with a 12 layer Transformer (Vaswani et al., 2017) and a hidden size of 768. It is trained on the Chinese Wikipedia dump with about 0.4 billion tokens and published by Devlin et al. (2019).
- BERT-wwm-ext-base, a model with the same configuration of BERT-base except it uses whole word masking and is trained on additional 5 billion tokens (Cui et al., 2019a).
- ALBERT-tiny, ALBERT (Lan et al., 2019) is the state-of-the-art language representation model, but at the time this paper was written, the author did not publish its Chinese versions. Due to the limitation of computational power, we train a tiny version of ALBERT with only 4 layers and a hidden size of 312 on the CLUE

pre-training corpus.

- ERNIE-base (Sun et al., 2019c) extends BERT-base with additional training data and leverages knowledge from Knowledge Graphs.
- XLNet-mid⁷, a model with 24 layers and a hidden size of 768, with sentencepiece tokenizer and other techniques from Yang et al. (2019).
- RoBERTa-large uses a 24 layer Transformer (Vaswani et al., 2017) with a hidden size of 1024, trained with the CLUE pre-training corpus and a sequence length of 256. It has a similar training procedure as that described in Liu et al. (2019)
- RoBERTa-wwm-ext-base (Cui et al., 2019a) uses a 12 layer Transformer (Vaswani et al., 2017) with a hidden size of 768, it uses whole word masking and is trained on the same dataset as BERT-base-wwm except following the training procedure of Liu et al. (2019).
- RoBERTa-wwm-ext-large (Cui et al., 2019a) has a network structure of RoBERTa-large and training procedure of RoBERTa-wwm-ext-base.

More details of these models can be found in Table 6 Parameters for pre-training in Appendix and the corresponding papers.

Fine-tuning We fine-tune the pre-trained models separately for each task. Hyperparameters are chosen based on the performance of each model on the development set. We also use early stopping to select the best checkpoint.

6.2 Human Performance

CMRC 2018, ChID and C³ have provided human performance (Sun et al., 2019b; Cui et al., 2019b; Zheng et al., 2019). For those tasks without human performance in CLUE, we ask human annotators to label 100 randomly chosen items from the test set and compute the annotators’ majority vote against the gold label.

We are first interested in how human annotators would perform if they have *only* seen the task instructions and very few pieces of labeled data, i.e. no training on the task. The results are shown

in the upper half of Table 3. We can see that the performance of our annotators is not very satisfactory when they have little or no training.

Next, we follow procedures in SuperGLUE (Wang et al., 2019) to train the annotators before asking them to work on the test data. Specifically, each annotator is first asked to annotate 30 to 50 pieces of data from the development set, and then compare their labels with the gold ones. They are then encouraged to discuss their mistakes and questions with other annotators until they are confident about the task. Then they annotate 100 pieces of test data, which is used to compute our final human performance, shown in the lower half of Table 3 and the last row of Table 2. We observe an average increase in the accuracy of 8.0% (AFQMC) to 19.5% (CSL). We will discuss human performance in light of the models’ performance in the next section.

6.3 Benchmark Results

Each model is fine-tuned three times and the test results come from the model that performs the best on development set. Results are shown in Table 2.

Models Performance Analysis The first thing we notice is that the larger pre-trained model, the more pre-training dataset, using whole word masking, the results become better. Specifically, RoBERTa-wwm-ext-large outperforms other models in all tasks, particularly for machine reading tasks such as C³.

Next, we want to highlight the results from ALBERT-tiny, which has fewer than 1/20 of the parameters in BERT-base model. Our results suggest that for single-sentence or sentence-pair tasks, the performance of small models are not very far behind much larger models. However, for tasks involving more global understanding, small model size negatively affects the results seriously, as illustrated by ALBERT-tiny’s low accuracy in all three machine reading tasks.

It should be noted that XLNet-mid is based on SentencePiece (Kudo and Richardson, 2018), which works as a quite common unsupervised text tokenizer in English, performs poorly in token level Chinese tasks like span-extraction based MRC (CMRC 2018). This shows the gap between the tokenizer used in English and Chinese since there are no natural blanks in Chinese texts.

Analysis of Tasks It seems that what is easy for human may not be so for machine. For instance, hu-

⁷<https://github.com/ymcui/Chinese-PreTrained-XLNet>

		TNEWS	AFQMC	CSL	IFLYTEK	CLUEWSC2020
Direct annotation	No.1	48.3	67.3	68.6	52.0	93.0
	No.2	54.7	78.7	69.0	48.0	95.0
	No.3	66.3	70.0	44.0	33.0	97.0
	avg	56.4	72.0	60.5	44.3	95.0
Trained annotation	No.1	57.0	83.0	93.0	54.0	94.0
	No.2	66.0	81.0	80.0	80.0	97.0
	No.3	73.0	76.0	67.0	50.0	95.0
	avg	65.3	80.0	80.0	61.3	95.3
	majority	71.0	81.0	84.0	66.0	98.0
	best model	58.61	76.5	82.13	62.98	81.38

Table 3: Two-stage human performance scores and the best accuracy of models comparison. “avg” denotes the mean score from the three annotators. “majority” shows the performance if we take the majority vote from the labels given by the annotators. **Bold** text denotes the best result among human and model performance.

humans are very accurate in multiple-choice reading comprehension (C^3), whereas machines struggle in it (ALBERT-tiny has a very low accuracy of about 32%, probably due to the small size of the model). The situation is similar for CLUEWSC2020, where the best score of models is far behind human performance. Note that in SuperGLUE, RoBERTa did very well on the English WSC (89% against 100% for humans). Whereas in our case, the performance of variants of RoBERTa is still much lower than the average human performance, though it is better than other models.

On the other hand, tasks such as CSL and ChID seem to be of equal difficulty for humans and machines, with accuracies in the 80’s for both. For humans, the keyword judgment task (CSL) is hard because the fake keywords all come from the abstract of the journal article, which has many technical terms: annotators are unlikely to perform well when working with unfamiliar jargon.

Surprisingly, the hardest dataset for both humans and machines is a single sentence task: TNEWS. One possible reason is that news titles can potentially fall under multiple categories (e.g., finance and technology) at the same time, while there is only one gold label in TNEWS.

The best result from machines remains far below human performance, with 11.6 points lower than human performance on average. This leaves room for further improvement. It indicates that solving CLUE is worthwhile as a driving force for the development of future models and methods.

7 Diagnostic Dataset for CLUE

7.1 Dataset Creation

In order to examine whether the trained models

can master linguistically important and meaningful phenomena, we follow GLUE (Wang et al., 2018) to provide a diagnostic dataset, setting up as a natural language inference task and predicting whether a hypothesis is *entailed* by, *contradicts* to or is *neutral* to a given premise. Crucially, we did not translate the English diagnostics into Chinese, as the items in their dataset may be specific to English language or American/Western culture. Instead, we have several Chinese linguists hand-crafting 514 sentence pairs in idiomatic Chinese from scratch. These pairs cover 9 linguistic phenomena and are manually labeled by the same group of linguists. We ensured that the labels are balanced (majority baseline is 35.1%). Examples are shown in Table 4. Some of the categories directly address the unique linguistic properties of Chinese. For instance, items in the “Time of event” category test models on their ability to handle aspect markers such as 着 (imperfective marker), 了 (perfective marker), 过 (experiential marker), which convey information about the time of event, whether it is happening now or has already happened in the past. We believe that for a model to make robust inferences, it needs to understand such unique Chinese phenomena, and also has other important linguistic abilities, such as handling anaphora resolution (Webster et al., 2018) and monotonicity reasoning (Yanaka et al., 2019; Richardson et al., 2020).

7.2 Evaluation and Error Analysis

We evaluate three representative models on the diagnostic dataset: BERT-base, XLNet-mid, RoBERTa-wwm-ext-large. Each model is fine-tuned on CMNLI training set, translated from English MNLI dataset into Chinese, then tested on

	#	Premise	Hypothesis	gold	Predictions			Accuracy		
					BE	RO	XL	BE	RO	XL
Anaphora	48	马丽和她的母亲李琴一起住在这里。 Ma Li and her mother Li Qin live here together.	马丽是李琴的母亲。 Ma Li is Li Qin's mother.	C	E	E	E	52.1	58.3	50.0
Argument structure	50	小白看见小红在打游戏。 Xiao Bai saw Xiao Hong playing video games.	小红在打太极拳。 Xiao Hong is doing Tai Chi.	C	N	N	C	58.0	76.0	50.0
Common sense	50	小明没有工作。 Xiaoming doesn't have a job.	小明没有住房。 Xiaoming doesn't have a place to live.	N	E	N	E	50.0	58.0	50.0
Comparative	50	这筐桔子比那筐多。 This basket has more oranges than that one.	这筐桔子比那筐多了不少。 This basket has much more oranges than that one.	N	E	E	E	50.0	54.0	46.0
Double negation	24	你别不把小病小痛当一回事。 Don't take minor illness as nothing.	你应该重视小病小痛。 You should pay attention to minor illness.	E	C	C	C	45.8	54.2	33.3
Lexical semantics	100	小红很难过。 Xiaohong is sad.	小红很难看。 Xiaohong is ugly.	N	E	E	E	62.0	66.0	57.0
Monotonicity	60	有些学生喜欢在公共澡堂里唱歌。 Some students like to sing in the shower room.	有些女生喜欢在公共澡堂里唱歌。 Some female students like to sing in the shower room.	N	C	N	N	40.0	45.0	46.7
Negation	78	女生宿舍，男生勿入。 Girls dormitory, no entering for boys.	女生宿舍只能女生进出。 Only girls can go in and out of the girls dormitory.	E	C	C	E	66.7	73.1	66.7
Time of event	54	记者去年采访企业家了。 The reporter interviewed the entrepreneur last year.	记者经常采访企业家。 The reporter interviews the entrepreneur very often.	N	N	N	E	48.1	55.6	48.1
total	514							54.3	61.3	52.1

Table 4: The CLUE diagnostics: Example test items in 9 linguistic categories, with their gold labels and model predictions, as well as model accuracy. E = entailment, N = neutral, C = contradiction. BE = BERT-base, RO = RoBERTa-wwm-ext-large, XL = XLNet-mid.

our diagnostic dataset. As illustrated in Table 4, the highest accuracy is only about 61%, which indicates that models have a hard time solving these linguistically challenging problems. We believe that both models and inference datasets suggest room for improvement.

A breakdown of results is presented in the last few columns of Table 4. Monotonicity is the hardest, similar to GLUE diagnostics (Wang et al., 2018). It seems that XLNet-mid also has a hard time dealing with double negation, highlighting the difficulty of monotonicity for some models. An interesting case is the example of lexical semantics in Table 4, where the two two-character words “sad” (难过 *hard-pass*) and “ugly” (难看 *hard-look*) in Chinese have the same first character (难 *hard*). Thus the premise and hypothesis only differ in the last character, which all three models have decided

to ignore. One possible explanation is that the current state-of-the-art models in Chinese are also using the simple lexical overlap heuristic, as illustrated in McCoy et al. (2019) for English.

8 Conclusions and Future Work

In this paper, we present a Chinese Language Understanding Evaluation (CLUE) benchmark, which consists of eight natural language understanding tasks and a linguistically motivated diagnostic dataset, along with an online leaderboard for model evaluation. To the best of our knowledge, CLUE is the first comprehensive language understanding benchmark developed for Chinese. We evaluate several latest language representation models on CLUE and analyze their results. In addition, we release a large clean crawled raw text corpus that can be directly used for pre-training Chinese mod-

els. An analysis is conducted on the diagnostic dataset created by our linguists, which illustrates the limited ability of state-of-the-art models to handle some Chinese linguistic phenomena.

Our results suggest that although current state-of-the-art models can achieve relatively high scores on many tasks, they still fall behind human performance in general. Also, small models such as ALBERT-tiny can have a close performance to the larger ones in simple tasks but may stumble on tasks requiring understanding of longer texts.

In the future, as more small models become publicly available, our leaderboard will reflect this trend. We will also include more results from state-of-the-art models such as ALBERT-xxlarge (Lan et al., 2019) and increase the diversity and difficulty of the tasks in CLUE.

References

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- Adam Roberts Katherine Lee Sharan Narang Michael Matena Yanqi Zhou Wei Li Peter J. Liu Colin Raffel, Noam Shazeer. 2019. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *arXiv e-prints*.
- Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019a. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.
- Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019b. [A span-extraction dataset for Chinese machine reading comprehension](#). 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 5886–5891, Hong Kong, China. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- LTD. IFLYTEK CO. 2019. Iflytek: a multiple categories chinese text classifier. *competition official website*, <http://challenge.xfyun.cn/2019/gamelist>.
- Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.
- Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv:1909.11942*.
- Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.
- Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. Lcqm: A large-scale chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1952–1962.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Kyle Richardson, Hai Hu, Lawrence S Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of AAAI*.
- Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019a. [Dream: A challenge data set and models for dialogue-based reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 7:217–231.
- Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019b. [Probing prior knowledge needed in challenging chinese machine reading comprehension](#). *CoRR*, cs.CL/1904.09679v2.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019c. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). *arXiv e-prints*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *ArXiv*, abs/2003.01355.
- Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can Neural Networks Understand Monotonicity Reasoning? In *ACL Workshop BlackboxNLP*.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv:1906.08237*.
- Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *arXiv preprint arXiv:1909.05658*.
- Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. Chid: A large-scale chinese idiom dataset for cloze test. *arXiv preprint arXiv:1906.01265*.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A DATASET SAMPLES

We have compiled examples of each data set for your reference in Table 5. Some of them are intercepted because the sentences are too long. For the complete data sets, you can refer to related papers. We will also release the download link of those datasets in the final version of the paper.

B ADDITIONAL PARAMETERS

B.1 Hyperparameters for pre-training

Although we did not train most of the models by ourselves, we list the hyperparameter for pre-training in Table 6 for reference purpose.

B.2 Hyperparameters for fine-tuning

Hyperparameters for fine-tuning in our experiments are listed in Table 7.

C ADDITIONAL BASELINE DETAILS

CSL In generating negative samples for CSL, we only replace one of the real keywords with a fake one. When fine-tuning on CSL task, we found that some of the larger models can only converge at very small learning rates, for example, $5e-6$.

IFLYTEK There are 126 categories in the original IFLYTEK dataset. However, some of them have few examples. We excluded those classes that have less than 10 examples so that we can apply the cross-validation filtering techniques as described in Section D. During the experiments, we also found when fine-tuning Albert-tiny requires a larger number of epochs to converge compare to other models. Also, sentences in IFLYTEK dataset are relatively long compared to other sentence classification tasks. However, most of the useful information is located at the beginning of the sentences. We, therefore, choose a max length of 128.

D Dataset Filtering

In order to increase the model differentiation and the difficulty of the dataset, we use four-fold cross-validation to filter iFLYTEK and TNEWS dataset. We divide the data sets in to four and use three of them to fine-tune ALBERT-tiny. After that, the fine-tuned model is used to select and filter those easy examples in the remaining set.

TNEWS	<p>sentence: 如果我的世界下架了，你会玩迷你世界吗？</p> <p>sentence (en): <i>If Minecraft is gone, will you play miniworld?</i></p> <p>label: 116(news_game)</p>
iFLYTEK	<p>sentence: 《钢铁英雄》是一款角色扮演类游戏。游戏拥有 带领他们逃出去。修复部分小错误，提升整体稳定性。</p> <p>sentence (en): <i>"Heroes of Steel" is a role-playing game. The game has all four heroes are imprisoned and you will lead them out. repair part small Errors to improve overall stability.</i></p> <p>label: 22(Strategy)</p>
SC2020	<p>text: 马克告诉皮特许多关于他自己的谎言，皮特也把这些谎言写进了他的书里。他应该多怀疑。</p> <p>text (en): <i>Mark told Pete many lies about himself, and Pete wrote them in his book. He should be more skeptical.</i></p> <p>label: false</p>
AEQMCW	<p>sentence1: 本月花呗还不上怎么办 sentence2: 花呗超时怎么办</p> <p>sentence1 (en): <i>What to do if Ant Credit Pay is not available yet this month</i> sentence2 (en): <i>How to deal with Ant Credit Pay overtime</i></p> <p>label: 0(different)</p>
CSL	<p>abst: 不同阶段电子数据的操作都会留下表现各异的轨迹.从操作系统、计算机应用系统 分析审计电子数据轨迹在计算机系统中表现形式,可以为审计人员提供有效的审计方法</p> <p>keyword: ["计算机审计", "数据轨迹", "日志文件"]</p> <p>abst (en): <i>The operation of electronic data in different stages will leave different traces. From operating system, computer application system provide effective audit methods for auditors by analyzing the expression of audit electronic data trace in computer system.</i></p> <p>keyword (en): ["computer audit", "data trace", "log file"]</p> <p>label: 0(false)</p>
CMRC 2018	<p>context: 萤火虫工作室是一家总部设在英国伦敦和康涅狄格州坎顿..... 目前，他们正在开发PC和Xbox360上的次时代游戏。</p> <p>question: 萤火虫工作室的总部设在哪里？ answer: 英国伦敦和康涅狄格州坎顿</p> <p>context (en): <i>Firefly Studios is a video game developer based in London, UK and Canton, Connecticut, with a quality department in Aberdeen, Scotland Currently, they are developing next-generation games on PC and Xbox 360.</i></p> <p>question (en): <i>Where is Firefly Studios headquartered?</i> answer (en): <i>London, UK and Canton, Connecticut</i></p>
ChID	<p>content: 中国青年报：篮协改革联赛切莫#idiom#.....</p> <p>candidates: ["急功近利", "画蛇添足", "本末倒置"(answer)]</p> <p>content (en): <i>China Youth Daily: Chinese Basketball Association should not #idiom# when reforming the league</i></p> <p>candidates (en): ["seeking instant benefit", "to overdo it", "take the branch for the root"(answer)]</p>
C ³	<p>document: 男：我们坐在第七排，应该能看清楚字幕吧？女：肯定可以，对了，我们得把手机设成振动。</p> <p>question: 他们最可能在哪儿？</p> <p>candidates: ["图书馆", "体育馆", "电影院"(answer), "火车站"]</p> <p>document (en): <i>Man: Our seats are in the seventh row. We should be able to see the subtitles clearly, right? Woman: Absolutely. By the way, we should set the phone to vibrate.</i></p> <p>question (en): <i>Where does the conversation most probably take place?</i></p> <p>candidates (en): ["In a library", "In a stadium", "In a cinema"(answer), "At a train station"]</p>

Table 5: Development set examples from the tasks in CLUE. **Bold** text represents part of the example format for each task. Chinese text is part of the model input, and the corresponding text in *italics* is the English version translated from that. Underlined text is specially marked in the input. Text in a monospaced font represents the expected model output.

	Masking	Type	Data Source	Training Tokens #	Device	Training Steps	Batch Size	Optimizer	Vocabulary	Init Ckpt
BERT-base	WordPiece	base	wiki	0.4B	TPU Pod v2	-	-	AdamW	21,128	Random Init
BERT-wwm-ext-base	WWM	base	wiki+ext	5.4B	TPU v3	1M	384	LAMB	~BERT	~BERT
ALBERT-tiny	WWM	tiny	CLUE corpus	5B	TPU Pod v3	500k	4k	LAMB	~BERT	Random Init
ERNIE-base	Knowledge Masking	base	wiki+ext	15B	NVidia v100	1M	8192	Adam	17964	Random Init
XLNet-mid	Sentence Piece	mid	wiki+ext	5.4B	TPU v3	2M	32	Adam	32000	Random Init
RoBERTa-large	WWM	large	CLUE corpus	5B	TPU Pod	100k	8k	AdamW	~BERT	Random Init
RoBERTa-wwm-ext-base	WWM	base	wiki+ext	5.4B	TPU v3	1M	384	AdamW	~BERT	~BERT
RoBERTa-wwm-ext-large	WWM	large	wiki+ext	5.4B	TPU Pod v3-32	2M	512	AdamW	~BERT	Random Init

Table 6: Parameters for pre-training. "BERT-base" is released by google (Devlin et al., 2019). "WWM" stands for whole word masking. "ext" presents for extended data, different models may use different extended data. "~BERT" means similar to Google's Chinese BERT.

	Model	Batch Size	Max Length	Epoch	Learning Rate
AFQMC	All*	16	128	3	2e-5
TNEWS	All*	16	128	3	2e-5
IFLYTEK	ALBERT-tiny	32	128	10	2e-5
	RoBERTa-large, RoBERTa-wwm-ext-large	24	128	3	2e-5
	All* except above	32	128	3	2e-5
CLUEWSC2020	ALBERT-tiny	8	128	50	1e-4
	All* except ALBERT-tiny	8	128	50	2e-5
CSL	RoBERTa-large	4	256	5	5e-6
	All* except above	4	256	5	1e-5
CMRC*	ALBERT-tiny	32	512	3	2e-4
	RoBERTa-wwm-ext-large	32	512	2	2.5e-5
	RoBERTa-large	32	256	2	3e-5
	XLNet-mid, RoBERTa-wwm-ext-base	32	512	2	3e-5
	All* except above	32	512	2	3e-5
CHID	All*	24	64	3	2e-5
C ³	All*	24	512	8	2e-5

Table 7: Parameters for fine-tuning. CMRC* presents for CMRC dataset in 2018. All* means ALBERT-tiny, BERT-base, BERT-wwm-ext-base, ERNIE-base, RoBERTa-large, XLNet-mid, RoBERTa-wwm-ext-base and RoBERTa-wwm-ext-large namely. It should be noted that RoBERTa-large is pre-trained with 256 sequence length, which is shorter than 512 length pre-trained for others. So we individually limit the length of RoBERTa-large to 256 for CMRC*, and use the striding text span to relieve this problem. However, this drawback of RoBERTa-large may decrease performances of some datasets whose length can not be effectively cut down, such as C³.