CLUE: A Chinese Language Understanding Evaluation Benchmark

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 team

CLUE@CLUEbenchmarks.com

Abstract

The advent of natural language understanding (NLU) benchmarks for English, such as GLUE and SuperGLUE allows new NLU models to be evaluated across a diverse set of tasks. These comprehensive benchmarks have facilitated a broad range of research and applications in natural language processing (NLP). The problem, however, is that most such benchmarks are limited to English, which has made it difficult to replicate many of the successes in English NLU for other languages. To help remedy this issue, we introduce the first large-scale Chinese Language Understanding Evaluation (CLUE) benchmark. CLUE is an open-ended, community-driven project that brings together 9 tasks spanning several well-established single-sentence/sentence-pair classification tasks, as well as machine reading comprehension, all on original Chinese text. To establish results on these tasks, we report scores using an exhaustive set of current state-of-the-art pre-trained Chinese models (9 in total). We also introduce a number of supplementary datasets and additional tools to help facilitate further progress on Chinese NLU. Our benchmark is released at https://www.CLUEbenchmarks.com

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. 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 **C**hinese **L**anguage **U**nderstanding **E**valuation (CLUE) benchmark that contains a collection of nine different natural language understanding tasks (two of which are created by us), including semantic similarity, natural language inference, short text classification, long text classification with large number of classes, and different types of machine reading comprehension tasks. 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. Also Chinese models are mainly

* Corresponding author. E-mail: lanzhenzhong@westlake.edu.cn

Proceedings of the 28th International Conference on Computational Linguistics, pages 4762–4772
Barcelona, Spain (Online), December 8-13, 2020

trained on different and relatively small corpora. Therefore, it is difficult to improve model performance and compare them across model architectures. This difficulty motivates us to construct and release a standard CLUE pre-training dataset: a corpus with over 214 GB 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 and 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 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 NLP 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 evaluation procedures are also used for sentence representations (Kiros et al., 2015). However, as different researchers use different evaluation pipelines on different datasets, results reported in the papers are not always fully comparable, especially in the case where the datasets are small, where 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 using a set of popular sentence embedding evaluation datasets. GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) further improve SentEval by providing benchmarks for natural language understanding tasks, ensuring that results from different models are consistent and comparable. 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 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., 2019) 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) nine 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 nine different tasks, to ensure that the benchmark tests 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.

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 many 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).

Although Chinese is not a low-resource language, it is still non-trivial to find and collect NLU tasks in Chinese, given a lack of diverse publicly available NLP datasets relative to English. Therefore apart from scrutinizing existing literature, we also sent out a call-for-tasks to the Chinese NLP community from which we received proposals or suggestions for several new datasets.[1] In addition, to help overcome the lack of publicly-available NLU-oriented sentence-/sentence-pair classification tasks for Chinese, we created two new tasks for our benchmark (CLUEWSC2020 and CSL, see section 4 for details). Based on the above standards, we gathered a total of nine tasks in the end, seven of them selected from our collected datasets plus two newly created by us. These tasks cover a broad range of text genres, linguistic phenomena and task-formats.

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, a total of 214 GB raw corpus with around 76 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.

4 Tasks

CLUE has nine 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 in the Appendix.

4.1 Single Sentence Tasks

TNEWS TouTiao Text Classification for News Titles[2] 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

[1] We only accepted some of them because other tasks were either not well-defined, or are normally not counted as NLU tasks (e.g., named-entity recognition).

[2] https://github.com/fatecbf/toutiao-text-classfication-dataset/

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.	academic (CNKI)
OCNLI	50k	3k	3k	natural language inference	acc.	5 genres
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; OCNLI has 3 classes, other classification tasks are binary classification.

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 dataset is an anaphora/coreference resolution task where the model is asked to decide whether a pronoun and a noun (phrase) in a sentence co-refer (binary classification), built following similar datasets in English (e.g., Levesque et al. (2012) and 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.

4.2 Sentence Pair Tasks

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

AFQMC The Ant Financial Question Matching Corpus[3] 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.

OCNLI Original Chinese Natural Language Inference (OCNLI, Hu et al. (2020)) is collected closely following procedures of MNLI (Williams et al., 2018). OCNLI is composed of 56k inference pairs from five genres: news, government, fiction, TV transcripts and Telephone transcripts, where the premises are collected from Chinese sources, and universities students in language majors are hired to write the hypotheses. The annotator agreement is on par with MNLI. We believe the non-translation nature of OCNLI makes it more suitable than XNLI (Conneau et al., 2018) as an NLU task specific for Chinese.

[3]https://dc.cloud.alipay.com/index#/topic/intro?id=3

4.3 Machine Reading Comprehension

CMRC 2018 CMRC 2018 (Cui et al., 2019) 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.

C^3 C^3 (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 (Raffel et al., 2020).

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, consisting of three different corpora: CLUECorpus2020-small, CLUECorpus2020, and CLUEOSCAR. Three models in this paper are pre-trained on the combined CLUE pre-training corpus (two ALBERT models and RoBERTa-large).

CLUECorpus2020-small It contains 14 GB of Chinese text, with the following genres:

- **News** This sub-corpus is crawled from the We Media (self-media) platform, with a total of 3 billion Chinese words from 2.5 million news articles of roughly 63K sources.
- **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 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 SophonPlus[4]. This subset has approximately 2.3 GB of raw texts with 0.8 billion Chinese words.

CLUECorpus2020 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.

[4]https://github.com/SophonPlus/ChineseNlpCorpus/

		Single Sentence			Sentence Pair			MRC		
Model	**Avg**	**TNEWS**	**IFLYTEK**	**CLUEWSC2020**	**AFQMC**	**CSL**	**OCNLI**	**CMRC**	**ChID**	**C^3**
BERT-base	69.20	56.58	60.29	63.45	73.70	80.36	72.20	69.72	82.04	64.50
BERT-wwm-ext-base	70.27	56.84	59.43	62.41	74.07	80.63	74.42	73.23	82.90	68.50
ALBERT-tiny	56.01	53.35	48.71	63.38	69.92	74.56	65.12	53.68	43.53	31.86
ALBERT-xxlarge	72.49	59.46	62.89	61.54	75.60	83.63	77.70	75.15	83.15	73.28
ERNIE-base	69.72	58.33	58.96	63.44	73.83	79.10	74.11	73.32	82.28	64.10
XLNet-mid	68.58	56.24	57.85	61.04	70.50	81.26	72.63	66.51	83.47	67.68
RoBERTa-large	71.01	57.86	62.55	62.44	74.02	81.36	76.82	76.11	84.50	63.44
RoBERTa-wwm-ext-base	71.17	56.94	60.31	72.07	74.04	81.00	74.72	73.89	83.62	63.90
RoBERTa-wwm-ext-large	74.90	58.61	62.98	81.38	76.55	82.13	78.20	76.58	85.37	72.32
Human	**85.09**	**71.00**	**66.00**	**98.00**	**81.0**	**84.0**	**90.30**	**92.40**	**87.10**	**96.00**

Table 2: Performance of baseline models on CLUE benchmark. *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.

CLUEOSCAR[5] 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

Baselines Our baseline models are built on different pre-trained transformers (Vaswani et al., 2017), on which an additional output layer is added for fine-tune on CLUE tasks. 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.

All the models are implemented in both TensorFlow (Abadi et al., 2016) and PyTorch (Paszke et al., 2019).

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

- BERT-base, we use the base model (12 layer, hidden size 768) published by (Devlin et al., 2019), which was pre-trained the on Chinese Wikipedia dump of about 0.4 billion tokens.
- 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., 2020).
- ALBERT-tiny/xxlarge, ALBERT (Lan et al., 2019) is a recent language representation model. We use: 1) a tiny version[6] with only 4 layers and a hidden size of 312, and 2) an xxlarge version[7] with 12 layers and a hidden size of 4096. Both are trained 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-mid[8], a model with 24 layers and a hidden size of 768, with sentencepiece tokenzier and other techniques from Yang et al. (2019).
- RoBERTa-large uses a 24 layer RoBERTa (Liu et al., 2019) with a hidden size of 1024, trained with the CLUE pre-training corpus.

[5]https://dumps.wikimedia.org/zhwiki/latest/
[6]https://github.com/brightmart/albert_zh
[7]https://github.com/google-research/albert
[8]https://github.com/ymcui/Chinese-PreTrained-XLNet

		TNEWS	AFQMC	CSL	IFLYTEK	CLUEWSC2020
Trained annotation	annotator 1	57.0	83.0	93.0	54.0	94.0
	annotator 2	66.0	81.0	80.0	80.0	97.0
	annotator 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.

- RoBERTa-wwm-ext-base (Cui et al., 2020) 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., 2020) has a network structure of RoBERTa-large and training procedure of RoBERTa-wwm-ext-base.

We believe these models are representative of most of the current transformer architectures. In particular, ALBERT-xxlarge and RoBERTa-wwm-ext-large are the largest models in Chinese at the time of writing, and are expected to give us an estimate of the upperbound of model performance. We include ALBERT-tiny to examine empirically how big the performance reduction is when switched to a much smaller model, which presents another estimate for scenarios with limited computing resources. A summary of the hyper-parameters of these models can be found in Table 6 in the Appendix.

Fine-tuning We fine-tune the pre-trained models separately for each task. Hyper-parameters are chosen based on the performance of each model on the development set. We also use early stopping to select the best checkpoint. Each model is fine-tuned three times and we choose the model with the best performance on the development set to report test results.

6.1 Human Performance

OCNLI, CMRC 2018, ChID and C^3 have provided human performance (Hu et al., 2020; Sun et al., 2019b; Cui et al., 2019; 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 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 Table 3 and the last row of Table 2. As we can see, most of the tasks are relatively easy for humans with a score in the 80s and 90s, except for TNEWS and IFLYTEK, both of which have many classes, potentially making it harder for humans. We will discuss human performance in light of the models' performance in the next section.

6.2 Benchmark Results

We report the results of our baseline models on the CLUE benchmark in Table 2.

Analysis of Model Performance The first thing we notice is that the results are better when: 1) the model is larger, or 2) the model is trained with more pre-training data, or 3) whole word masking is used. Specifically, RoBERTa-wwm-ext-large and ALBERT-xxlarge are the two best performing models, showing advantages over other models particularly for machine reading tasks such as C^3.

Next, we want to highlight the results from ALBERT-tiny, which has only about 1/20 of the parameters in BERT-base model. Our results suggest that for single-sentence or sentence-pair tasks, the performance drop compared with BERT-base can range from almost 0 (for CLUEWSC2020) to roughly 12 percentage points (IFLYTEK). However, for tasks involving more global understanding, small models have more serious limitations, as illustrated by ALBERT-tiny's low accuracy in all three machine reading tasks, with a performance drop of up to 40 percentage compared with BERT-base (ChID).

Finally, XLNet-mid, a model based on a common unsupervised tokenizer in English called SentencePiece (Kudo and Richardson, 2018), performs poorly in token level Chinese tasks like span-extraction based MRC (CMRC 2018). This highlights the need for our Chinese-specific benchmark which provides empirical results as to whether successful techniques in English can be readily applied or transferred to a very different language such as Chinese, where no word boundaries are present in running texts.

Analysis of Tasks It seems that what is easy for human may not be so for machine. For instance, 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 (about 17 percentage points). 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 roughly 11 points lower than human performance on average. This leaves much room for further improvement of models and methods, which we hope will drive the Chinese NLP community forward.

7 Diagnostic Dataset for CLUE

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).

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 OCNLI, and then tested on 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.

	#	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	47.9	58.3	47.9
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	C	C	C	60.0	60.0	54.0
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	N	N	C	44.0	58.0	48.0
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	36.0	56.0	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	E	E	E	54.2	62.5	62.5
Lexical semantics	100	小红很难过。 Xiaohong is sad.	小红很难看。 Xiaohong is ugly.	N	E	N	E	62.0	70.0	64.0
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	N	N	N	41.7	43.3	43.3
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	E	C	C	62.8	64.1	60.3
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	N	61.1	74.1	59.3
Total								53.5	61.5	54.7

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.

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 BERT also has a hard time dealing with comparatives. 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 two out of three models have decided to ignore. One possible explanation is that these 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 9 natural language understanding tasks and a linguistically motivated diagnostic dataset, along with an online leaderboard for model evaluation. In addition, we release a large clean crawled raw text corpus that can be directly used for pre-training Chinese models. 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. An analysis is conducted on the diagnostic dataset created by Chinese linguists, which illustrates the limited ability of state-of-the-art models to handle some Chinese linguistic phenomena.

In contrast to the English benchmarks such as GLUE and SuperGLUE, where model performance is already at human performance, we can see that Chinese NLU still has considerable room for improvement (i.e., models are $\sim$10% below our estimates of human performance), meaning that we expect that our benchmark will facilitate building better models in the short-term. Once models have reached human performance, however, we believe that extending our benchmark to newer tasks, or newer forms of evaluation (e.g., taking into account performance as a function of model size as in (Li et al., 2020)), could be a step forward. In this sense, we view CLUE, which is an entirely community-driven project, to be open-ended in that our current set of tasks serve as a first step in more comprehensively evaluating Chinese NLU.

9 Acknowledgement

The authors would like to thank everyone who has contributed their datasets to CLUE. We are also grateful to the annotators and engineers who have spent much of their time and effort helping with the creation of the CLUE benchmark. Special thanks to the following companies and organizations: OneConnect Financial Technology Co., Ltd, OpenBayes Co., Ltd, AI-Indeed.com, Alibaba Cloud Computing, Joint Laboratory of HIT and iFLYTEK Research (HFL). Research supported with Cloud TPUs from Google's TensorFlow Research Cloud (TFRC).

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*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. 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, November. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Findings of EMNLP*. 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, June. Association for Computational Linguistics.

Hai Hu, Kyle Richardson, Xu Liang, Li Lu, Sandra Kübler, and Larry Moss. 2020. OCNLI: Original Chinese natural language inference. In *Findings of Empirical Methods for Natural Language Processing (Findings of EMNLP)*.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

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. In *International Conference on Learning Representations*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Junyi Li, Hai Hu, Xuanwei Zhang, Minglei Li, Lu Li, and Liang Xu. 2020. Light pre-trained Chinese language model for nlp tasks. In Xiaodan Zhu, Min Zhang, Yu Hong, and Ruifang He, editors, *Natural Language Processing and Chinese Computing*, pages 567–578. Springer International Publishing.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. Lcqmc: 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, July. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Machine Learning Research,*, pages 1–67.

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, November. 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, 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, November. Association for Computational Linguistics.

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. *Neural Information Processing Systems*, pages 3266–3280.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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.