OCNLI: Original Chinese Natural Language Inference

Hai Hu[†◇] **Kyle Richardson**[‡] **Liang Xu**[◇]
Lu Li[◇] **Sandra Kübler**[†] **Lawrence S. Moss**[†]
[†]Indiana University, Bloomington, IN, USA
[‡]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[◇]Chinese Language Understanding Evaluation (CLUE) benchmark
{huhai,skuebler,lmoss}@indiana.edu; kyler@allenai.org;
xuliang@i-i.ai; ll000@mails.ccnu.edu.cn; CLUE@CLUEbenchmarks.com

Abstract

Despite the tremendous recent progress on natural language inference (NLI), driven largely by large-scale investment in new datasets (e.g., SNLI, MNLI) and advances in modeling, most progress has been limited to English due to a lack of reliable datasets for most of the world's languages. In this paper, we present the first large-scale NLI dataset (consisting of ∼56,000 annotated sentence pairs)[1] for Chinese called the *Original Chinese Natural Language Inference* dataset (OCNLI). Unlike recent attempts at extending NLI to other languages, our dataset does not rely on any automatic translation or non-expert annotation. Instead, we elicit annotations from native speakers specializing in linguistics. We follow closely the annotation protocol used for MNLI, but create new strategies for eliciting diverse hypotheses. We establish several baseline results on our dataset using state-of-the-art pre-trained models for Chinese, and find even the best performing models to be far outpaced by human performance (∼12% absolute performance gap), making it a challenging new resource that we hope will help to accelerate progress in Chinese natural language understanding. To the best of our knowledge, this is the first human-elicited MNLI-style corpus for a non-English language.

1 Introduction

In the last few years, natural language understanding has made considerable progress, driven largely by the availability of large-scale datasets and advances in neural modeling (Peters et al., 2018; Devlin et al., 2019). At the center of this progress has been natural language inference (NLI), which focuses on the problem of deciding whether two statements are connected via an entailment or a contradiction. NLI profited immensely from new datasets such as the Stanford NLI (SNLI, Bowman et al. (2015)) and Multi-Genre NLI (MNLI, Williams et al. (2018)) datasets. However, as often the case, this progress has centered around the English language given that the most well-known datasets are limited to English. Efforts to build comparable datasets for other languages have largely focused on (automatically) translating existing English NLI datasets (Mehdad et al., 2011; Conneau et al., 2018). But this approach comes with its own issues (see section 2).

To overcome these shortcomings and contribute to ongoing progress in Chinese NLU, we present the first large-scale NLI dataset for Chinese called the *Original Chinese Natural Language Inference* dataset (OCNLI). Unlike previous approaches, we rely entirely on original Chinese sources and use native speakers of Chinese with special expertise in linguistics and language studies for creating hypotheses and for annotation. Our dataset contains ∼56,000 annotated premise-hypothesis pairs and follows a similar procedure of data collection to the English MNLI. Following MNLI, the premises in these sentence pairs are drawn from multiple genres (5 in total), including both written and spoken Chinese (see Table 1 for examples). To ensure annotation quality and consistency, we closely mimic MNLI's original annotation protocols for monitoring annotator performance. We find that our trained annotators have high agreement on label prediction (with ∼98% agreement based on a 3-vote consensus). To our knowledge, this dataset constitutes the first large-scale NLI dataset for Chinese that does not rely on automatic translation.

Additionally, we establish baseline results based on a standard set of NLI models (Chen et al., 2017) tailored to Chinese, as well as new pre-trained Chinese transformer models (Cui et al., 2019). We find that our strongest model, based on RoBERTa (Liu

[1]Our dataset and code are available at https://github.com/CLUEbenchmark/OCNLI.

Premise	Genre `Level`	**Majority label** All labels	Hypothesis
但是不光是中国，日本，整个东亚文化都有这个特点就是被权力影响很深 But not only China and Japan, the entire East Asian culture has this feature, that is it is deeply influenced by the power.	TV `medium`	**Entailment** E E E E E	有超过两个东亚国家有这个特点 More than two East Asian countries have this feature.
完善加工贸易政策体 (We need to) perfect our work and trade policies.	GOV `easy`	**Entailment** E E E E E	贸易政策体系还有不足之处 (Our) trade policies still need to be improved.
咖啡馆里面对面坐的年轻男女也是上一代的故事，她已是过来人了 Stories of young couples sitting face-to-face in a cafe is already something from the last generation. She has gone through all that.	LIT `medium`	**Contradiction** C C C N N	男人和女人是背对背坐着的 The man and the woman are sitting back-to-back.
今天，这一受人关注的会议终于在波恩举行 Today, this conference which has drawn much attention finally took place in Bonn.	NEWS `easy`	**Neutral** N N N N C	这一会议原定于昨天举行 This conferences was scheduled to be held yesterday.
嗯,今天星期六我们这儿,嗯哼. En, it's Saturday today in our place, yeah.	PHONE `hard`	**Contradiction** C C C C C	昨天是星期天 It was Sunday yesterday.

Table 1: Examples from the MULTICONSTRAINT elicitation of our Chinese NLI dataset, one from each of the five text genres. `easy`: 1st hypothesis the annotator wrote for that particular premise and label; `medium`: 2nd hypothesis; `hard`: 3rd hypothesis. **Bold** label shows the majority vote from the annotators.

et al., 2019), performs far behind expert human performance (∼78% vs. ∼90% accuracy on our test data). These results show that the dataset is challenging without using special filtering that has accompanied many recent NLI datasets (Le Bras et al., 2020).

Contributions of this paper: 1) We introduce a new, high quality dataset for NLI for Chinese, based on Chinese data sources and expert annotators; 2) We provide strong baseline models for the task, and establish the difficulty of our task through experiments with recent pre-trained transformers. 3) We also demonstrate the benefit of naturally annotated NLI data by comparing performance with large-scale automatically translated datasets.

2 Related Work

Natural language inference (NLI), or recognizing textual entailment (RTE), is a long-standing task in NLP. Since we cannot cover the whole field, we focus on existing datasets and current systems.

Data: To date, there exists numerous datasets for English, ranging from smaller/more linguistics oriented resources such as FraCaS (Cooper et al., 1996), to larger ones like the RTE challenges (Dagan et al., 2005) and SICK (Marelli et al., 2014). Perhaps the most influential are the two large-scale, human-elicited datasets: the Stanford Natural Language Inference Corpus (SNLI) (Bowman et al., 2015), whose premises are taken from image captions, and the Multi-Genre Natural Language Inference Corpus (MNLI) (Williams et al., 2018), whose premises are from texts in 10 different genres. Both are built by collecting premises from pre-defined text, then having annotators come up with possible hypotheses and inference labels, which is the procedure we also employ in our work.

These large corpora have been used as part of larger benchmark sets, e.g., GLUE (Wang et al., 2018), and have proven useful for problems beyond NLI, such as sentence representation and transfer learning (Conneau et al., 2017; Subramanian et al., 2018; Reimers and Gurevych, 2019), automated question-answering (Khot et al., 2018; Trivedi et al., 2019) and model probing (Warstadt et al., 2019; Richardson et al., 2020; Geiger et al., 2020; Jeretic et al., 2020).

The most recent English corpus Adversarial NLI (Nie et al., 2020) uses Human-And-Model-in-the-Loop Enabled Training (HAMLET) method for data collection. Their annotation method requires an existing NLI corpus to train the model during annotation, which is not possible for Chinese at the moment, as there exists no high-quality Chinese data.

In fact, there has been relatively little work on de-

Premise	Hypothesis
a. Louisa May Alcott和Nathaniel Hawthorne 住在Pinckney街道，而那个被Oliver Wendell Holmes称为“晴天街道 的Beacon Street街道住着有些喜欢自吹自擂的历史学家 William Prescott ***Eng.:*** *Louisa May Alcott and Nathaniel Hawthorne lived on Pinckney street, but on Beacon Street street, which is named “Sunny Street by Oliver Wendell Holmes, lived the bragging historian William Prescott. [sic]*	Hawthorne住在Main Street上 *Eng.: Hawthorne lived on Main Street.*
b. 运行 Slient，运行Deep，运行答案 *Eng.: run Slient, run Deep, run answer. [sic]*	悄悄的逃走 ***Eng.:*** *secretly escape.*

Table 2: Examples from crowd-translated XNLI development set (Conneau et al., 2018), showing problems of *translationese* (top) and poor translation quality (bottom).

veloping large-scale human-annotated resources for languages other than English. Some NLI datasets exist in other languages, e.g., Fonseca et al. (2016) and Real et al. (2020) for Portuguese, Hayashibe (2020) for Japanese, and Amirkhani et al. (2020) for Persian, but none of them have human elicited sentence pairs. Efforts have largely focused on automatic translation of existing English resources (Mehdad et al., 2011), sometimes coupled with smaller-scale hand annotation by native speakers (Negri et al., 2011; Agić and Schluter, 2017). This is also true for some of the datasets included in the recent Chinese NLU benchmark CLUE (Xu et al., 2020) and for XNLI (Conneau et al., 2018), a multilingual NLI dataset covering 15 languages including Chinese.

While automatically translated data have proven to be useful in many contexts, such as cross-lingual representation learning (Siddhant et al., 2020), there are well-known issues, especially when used in place of human annotated, quality controlled data. One issue concerns limitations in the quality of automatic translations, resulting in incorrect or unintelligible sentences (e.g., see Table 2b). But even if the translations are correct, they suffer from “translationese”, resulting in unnatural language, since lexical and syntactic choices are copied from the source language even though they are untypical for the target language (Koppel and Ordan, 2011; Hu et al., 2018; Hu and Kübler, 2020).

A related issue is that a translation approach also copies the cultural context of the source language, such as an overemphasis on Western themes or cultural situations. The latter two issues are shown in Table 2a, where many English names are directly carried over into the Chinese translation, along with aspects of English syntax, such as long relative clauses, which are common in English but dispreferred in Chinese (Lin, 2011).

Systems: As inference is closely related to logic, there has always been a line of research building logic-based or logic-and-machine-learning hybrid models for NLI/RTE problems (e.g. MacCartney, 2009; Abzianidze, 2015; Martínez-Gómez et al., 2017; Yanaka et al., 2018; Hu et al., 2020). However, in recent years, large datasets such as SNLI and MNLI have been almost exclusively approached by deep learning models. For examples, several transformer architectures achieve impressive results on MNLI, with current state-of-the-art T5 (Raffel et al., 2019) reaching 92.1/91.9% accuracy on the matched and mismatched sets.

Re-implementations of these transformer models for Chinese have led to similar successes on related tasks. For example, Cui et al. (2019) report that a large RoBERTa model (Liu et al., 2019), pre-trained with whole-word masking, achieves the highest accuracy (81.2%) among their transformer models on XNLI. In the CLUE benchmark (Xu et al., 2020), the same RoBERTa model also achieves the highest aggregated score from eight tasks. We will use this model to establish baselines on our new dataset.

Biases: The advances in dataset creation have led to an increased awareness of systematic biases in existing datasets (Gururangan et al., 2018), as measured through *partial-input baselines*, e.g., the *hypothesis-only* baselines explored in Poliak et al. (2018) where a model can achieve high accuracy by only looking at the hypothesis and ignoring the premise completely (see also Feng et al. (2019)). These biases have been mainly associated with the annotators (crowd workers in MNLI’s case) who use certain strategies to form hypotheses of a specific label, e.g., adding a negator for contradictions.

There have been several recent attempts to reduce such biases (Belinkov et al., 2019; Sakaguchi et al., 2020; Le Bras et al., 2020; Nie et al., 2020). There has also been a large body of work using probing datasets/tasks to stress-test NLI models trained on datasets such as SNLI and MNLI, in order to expose the weaknesses and biases in either the models or the data (Dasgupta et al., 2018; Naik et al., 2018; McCoy et al., 2019). For

this work, we closely monitor the hypothesis-only and other biases but leave systematic filtering/bias-reduction/stress-testing for future work. An interesting future challenge will involve seeing how such techniques, which focus exclusively on English, transfer to other languages such as Chinese.

3 Creating OCNLI

Here, we describe our data collection and annotation procedures. Following the standard definition of NLI (Dagan et al., 2006), our data consists of ordered pairs of sentences, one *premise* sentence and one *hypothesis* sentence, annotated with one of three labels: Entailment, Contradiction, or Neutral (see examples in Table 1).

Following the strategy that Williams et al. (2018) established for MNLI, we start by selecting a set of premises from a collection of multi-genre Chinese texts, see Section 3.1. We then elicit hypothesis annotations based on these premises using expert annotators (Section 3.2). We develop novel strategies to ensure that we elicit diverse hypotheses. We then describe our verification procedure in Section 3.3.

3.1 Selecting the Premises

Our premises are drawn from the following five text genres: government documents, news, literature, TV talk shows, and telephone conversations. The genres were chosen to ascertain varying degrees of formality, and they were collected from different primary Chinese sources. The government documents are taken from annual Chinese government work reports[2]. The news data are extracted from the news portion of the Lancaster Corpus of Mandarin Chinese (McEnery and Xiao, 2004). The data in the literature genre are from two contemporary Chinese novels[3], and the TV talk show data and telephone conversations are extracted from transcripts of the talk show *Behind the headlines with Wentao*[4] and the Chinese Callhome transcripts (Wheatley, 1996).

As for pre-processing, annotation symbols in the Callhome transcripts were removed and we limited our premise selection to sentences containing 8 to 50 characters.

[2] http://www.gov.cn/guowuyuan/baogao.htm, last visited 4/21/2020, same below.

[3] *Ground Covered with Chicken Features* by Liu Zhenyun, *Song of Everlasting Sorrow* by Wang Anyi.

[4] http://phtv.ifeng.com/listpage/677/1/list.shtml.

3.2 Hypothesis Generation

One issue with the existing data collection strategies in MNLI is that humans tend to use the simplest strategies to create the hypotheses, such as negating a sentence to create a contradiction. This makes the problem unrealistically easy. To create more realistic, and thus more challenging data, we propose a new hypothesis elicitation method called *multi-hypothesis* elicitation. We collect four sets of inference pairs and compare the proposed method with the MNLI annotation method, where a single annotator creates an entailed sentence, a neutral sentence and a contradictory sentence given a premise (Condition: SINGLE).

Multi-hypothesis elicitation In this newly proposed setting, we ask the writer to produce *three* sentences per label, resulting in three entailments, three neutrals and three contradictions for each premise (Condition: MULTI). I.e. we obtain a total of nine hypotheses if the writer is able to come up with that many inferences, which is indeed the case for most premises in our experiment. Our hypothesis is that by asking them to produce three sentences for each type of inference, we push them to think beyond the easiest case. We call the 1st, 2nd and 3rd hypothesis by an annotator per label `easy`, `medium` and `hard` respectively, with the assumption that they start with the easiest inferences and then move on to harder ones. First experiments show that MULTI is more challenging than SINGLE, and at the same time, inter-annotator agreement is slightly higher than for SINGLE (see section 3.3).

However, we also found that MULTI introduces more hypothesis-only bias. Especially in contradictions, negators such as 没有 ("no/not") stood out as cues, similar to what had been reported in SNLI and MNLI (Poliak et al., 2018; Gururangan et al., 2018; Pavlick and Kwiatkowski, 2019). Therefore we experiment with two additional strategies to control the bias, resulting in MULTIENCOURAGE (*encourage* the annotators to write more diverse hypothesis) and MULTICONSTRAINT (put *constraints* on what they can produce), which will be explained in detail below.

These four strategies result in four different subsets. Table 3 gives a summary of these subsets.

Instructions for hypothesis generation The basis of our instructions are very similar to those for MNLI, but we modified them for each setting:

Subsets	Instructions	# Pairs / Mean length of hypothesis *H* in characters			
		Total	`easy`	`medium`	`hard`
SINGLE	same as MNLI; one *H* per label	11,986 / 10.9	n.a.	n.a.	n.a.
MULTI	three *H*s per label	12,328 / 10.4	4,836 / 9.9	4,621 / 10.6	2,871 / 11.0
MULTIENCOURAGE	MULTI + encouraging annotators to use fewer negators and write more diverse hypotheses	16,584 / 12.2	6,263 / 11.5	6,092 / 12.5	4,229 / 12.7
MULTICONSTRAINT	MULTI + constraints on the negators used in contradictions	15,627 / 12.0	5,668 / 11.6	5,599 / 12.2	4,360 / 12.4
total		56,486 / 11.5			

Table 3: Information on the four subsets of data collected. Premises in all subsets are drawn from the same pool of text from five genres. `easy/medium/hard` refers to the 1st/2nd/3rd hypothesis written for the same premise and inference label. Number of pairs in the `hard` condition is smaller because not all premises and all labels have a third hypothesis. See section 3.2 for details of the subsets.

SINGLE We asked the writer to produce one hypothesis per label, same as MNLI[5].

MULTI Instructions are the same except that we ask for three hypotheses per label.

MULTIENCOURAGE We *encouraged* the writers to write high-quality hypotheses by telling them explicitly which types of data we are looking for, and promised a monetary bonus to those who met our criteria after we examined their hypotheses. Among our criteria are: 1) we are interested in *diverse* ways of making inferences, and 2) we are looking for contradictions that do *not* contain a negator.

MULTICONSTRAINT We put *constraints* on hypothesis generation by specifying that *only one out of the three contradictions can contain a negator*, and that we would randomly check the produced hypothesis, with violations of the constraint resulting in lower payment. We also provided extra examples in the instructions to demonstrate contradictions without negators. These examples are drawn from the hypotheses collected from prior data.

We are also aware of other potential biases or heuristics in human-elicited NLI data such as the lexical overlap heuristic (McCoy et al., 2019). Thus in all our instructions, we made explicit to the annotators that no hypothesis should overlap more than 70% with the premise. However, examining how prevalent such heuristics are in our data requires constructing new probing datasets for Chinese, which is beyond the scope of this paper.

[5]See Appendix A for the complete instructions.

Annotators We hired 145 undergraduate and graduate students from several top-tier Chinese universities to produce hypotheses. All of the annotators (*writers*) are native speakers of Chinese and are majoring in Chinese or other languages. They were paid roughly 0.3 RMB (0.042 USD) per P-H pair. No single annotator produced an excessive amount of data to avoid annotator-bias (for a discussion of this, see Geva et al. (2019)).

3.3 Data Verification

Following SNLI and MNLI, we perform data verification, where each premise-hypothesis pair is assigned a label by four independent annotators (*labelers*). Together with the original label assigned by the annotator, each pair has five labels. We then use the majority vote as the gold label. We selected a subset of the writers from the hypothesis generation experiment to be our labelers. For each subset, about 15% of the total data were randomly selected and relabeled. The labelers were paid 0.2 RMB (0.028 USD) for each pair.

Relabeling results Our results, shown in Table 4, are very close to the numbers reported for SNLI/MNLI, with labeler agreement even higher than SNLI/MNLI for SINGLE and MULTI.

Crucially, the three MULTI subsets, created using the three variants of the *multi-hypothesis* generation method, have similar agreement to MNLI, suggesting that producing nine hypotheses for a given premise is feasible. Furthermore, the agreement rates on the `medium` and `hard` portions of the subsets are only slightly lower than on the `easy` portion, with agreement rates of 3 labels at least

	SNLI†	MNLI†	XNLI†	OCNLI			
				SINGLE	MULTI	MULTIENC	MULTICON
# pairs in total	570,152	432,702	7,500	11,986	12,328	16,584	15,627
# pairs relabeled	56,941	40,000	7,500	1,919	1,994	3,000	3,000
% relabeled	10.0%	9.2%	100.0%	16.0%	16.2%	18.1%	19.2%
5 labels agree (unanimous)	58.3%	58.2%	na	62.1%	63.5%	57.2%	57.6%
4+ labels agree	na	na	na	82.2%	84.8%	82.0%	80.8%
3+ labels agree	**98.0%**	**98.2%**	**93.0%**	**98.6%**	**98.8%**	**98.7%**	**98.3%**
Individual label = gold label	89.0%	88.7%	na	88.1%	88.9%	87.0%	86.7%
Individual label = author's label	85.8%	85.2%	na	81.8%	82.3%	80.2%	79.7%
Gold label = author's label	91.2%	92.6%	na	89.8%	89.6%	89.6%	88.2%
Gold label ≠ author's label	6.8%	5.6%	na	8.8%	9.2%	9.0%	10.1%
No gold label (no 3 labels match)	2.0%	1.8%	na	1.4%	1.2%	1.3%	1.7%

Table 4: Results from labeling experiments for the four subsets. MULTIENC: MULTIENCOURAGE; MULTICON: MULTICONSTRAINT. † = numbers for SNLI, MNLI, XNLI are copied from the original papers (Bowman et al., 2015; Williams et al., 2018; Conneau et al., 2018). For XNLI, the numbers are for the English portion of the dataset, which is the only language that has been relabelled.

97.90% (see Table 10 in the Appendix), suggesting that our data in general is of high quality. Agreement is lower for MULTICONSTRAINT, showing that it may be difficult to produce many hypotheses under these constraints.

In a separate relabeling experiment, we examine the quality of human-translated examples from the XNLI dev set. The results show considerably lower agreement: The majority vote of our five annotators only agree with the XNLI gold-label 67% of the time, as compared to the lowest rate of 88.2% on MULTICONSTRAINT. Additionally, 11.6% of the XNLI dev examples in Chinese contain more than 10 Roman alphabets, which are extremely rare in original, every-day Chinese speech/text. These results suggest that XNLI is less suitable as validation set for Chinese NLI, and thus we excluded XNLI dev set in our evaluation. For further details, see Appendix C.

3.4 The Resulting Corpus

Overall, we have a corpus of more than 56,000 pairs of inference pairs in Chinese. We have randomized the total of 6,000 *relabeled* pairs from MULTIENCOURAGE and MULTICONSTRAINT and used them as the development and test sets, each consisting of 3,000 examples. All pairs from SINGLE and MULTI, plus the remaining 26,211 pairs from MULTIENCOURAGE and MULTICONSTRAINT are used for the training set, about 50,000 pairs[6]. This split ensures that all labels in the development and test sets have been verified, and the number of pairs in the `easy`, `medium` and `hard` portions are roughly the same in both sets. It is also closer to a realistic setting where contradictions without negation are much more likely. Pairs that do not receive a majority label in our relabeling experiment are marked with "-" as their label, and can thus be excluded if necessary.

4 Experimental Investigation of OCNLI

4.1 Experimental Setup

To demonstrate the difficulty of our dataset, we establish baselines using several widely-used NLI models tailored to Chinese[7]. This includes the baselines originally used in Williams et al. (2018) such as the continuous bag of words (CBOW) model, the biLSTM encoder model and an implementation of ESIM (Chen et al., 2017)[8]. In each case, we use Chinese character embeddings from Li et al. (2018) in place of the original GloVe embeddings.

We also experiment with state-of-the-art pre-trained transformers for Chinese (Cui et al., 2019)

[6]We note that given the constraints of having equal number of `easy`, `medium` and `hard` examples in dev/test sets, the resulting corpus ended up having high premise overlap between training and dev/test sets, in contrast to the original MNLI design. To ensure that such premise overlap does not bias the current models and inflate performance, we experimented with a smaller **non-overlap** train and test split, which was constructed by filtering parts of the training. This lead to comparable results, despite the non-overlap being much smaller in size, which we detail in Appendix G. Both the **overlap** and **non-overlap** splits will be released for public use, as well as part of the the public leaderboard at https://www.cluebenchmarks.com/nli.html.

[7]Additional details about all of our models and hyperparameters are included as supplementary material.

[8]We use a version of the implementations from https://github.com/NYU-MLL/multiNLI.

using the fine-tuning approach from Devlin et al. (2019). Specifically, we use the Chinese versions of BERT-base (Devlin et al., 2019) and RoBERTa-large (Liu et al., 2019) with whole-word masking (see details in Cui et al. (2019)). In both cases, we rely on the publicly-available TensorFlow implementation provided in the CLUE benchmark (Xu et al., 2020)[9]. Following Bowman et al. (2020), we also fine-tune *hypothesis-only* variants of our main models to measure annotation artifacts.

To measure human performance, we employed an additional set of 5 Chinese native speakers to annotate a sample (300 examples) of our OCNLI test set. This follows exactly the strategy used in Nangia and Bowman (2019) for measuring human performance in GLUE, and provides a *conservative* estimate of human performance in that annotators were provided with minimal amounts of task training (see Appendix E for details).

Datasets In addition to experimenting with OCNLI, we also compare the performance of our main models against models fine-tuned on the Chinese training data of XNLI (Conneau et al., 2018) (an automatically translated version of MNLI), as well as combinations of OCNLI and XNLI. The aim of these experiments is to evaluate the relative advantage of automatically translated data. We also compare both models against the CLUE diagnostic test from Xu et al. (2020), which is a set of 514 NLI problems that was annotated by an independent set of Chinese linguists.

To analyze the effect of our different hypothesis elicitation strategies, we look at model performance on different subsets of OCNLI. Due to the way in which the data is partitioned (all of SINGLE and MULTI are in the training set), it is difficult to fine-tune on OCNLI and test on all four subsets. We instead use an XNLI trained model, which is independent of any biases related to our annotation process, to probe the difficulty of our different subsets.

4.2 Baseline Results and Analysis

In this section, we describe our main results.

How Difficult is OCNLI? To investigate this, we train/fine-tune all five neural architectures on OCNLI training data and test on the OCNLI test set. The main results are shown in Table 5. All of the non-transformer models perform poorly while BERT and RoBERTa reach a ∼20 percentage-point advantage over the strongest of these models (ESIM). This shows the relative strength of pre-trained models on our task.

We find that while transformers strongly outperform other baseline models, our best model, based on RoBERTa, is still about 12 points below human performance on our test data (i.e., 90.3% versus 78.2%). This suggests that models have considerable room for improvement, and provides additional evidence of task difficulty. In comparison, these transformer models reach human-like performance in many of the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) tasks. For NLI specifically, the performance of the English RoBERTa on MNLI is 90.4%, and only about 2 percentage-points below the human score (Bowman et al., 2020; Nangia and Bowman, 2019). We see a similar trend for BERT, which is about 18 points behind human performance on OCNLI, but the difference is roughly 8 points for MNLI (Devlin et al., 2019). We also see much room for improvement on the CLUE diagnostic task, where our best model achieves only 61.3% (a slight improvement over the result reported in Xu et al. (2020)).

We also looked at how OCNLI fares on hypothesis-only tests, where all premises in train and test are replaced by the same non-word, thus forcing the system to make predictions on the hypothesis only. Table 7 shows the performance of these models on different portions of OCNLI. These results show that our elicitation gives rise to annotation artifacts in a way similar to most benchmark NLI datasets (e.g., OCNLI: $\sim 66\%$; MNLI $\sim 62\%$ and SNLI: $\sim 69\%$, as reported in Bowman et al. (2020) and Poliak et al. (2018), respectively). We specifically found that negative polarity items ("any", "ever"), negators and "only" are among the indicators for contradictions, whereas "at least" biases towards entailments. We see no negators for the MULTICONSTRAINT subset, which shows the effect of putting constraints on the hypotheses that the annotators can produce. Instead, "only" is correlated with contradictions. A more detailed list is shown in Figure 8, listing individual word and label pairs with high pairwise mutual information (PMI). PMI was also used by Bowman et al. (2020) for the English NLI datasets.

Given the large literature on adversarial filtering (Le Bras et al., 2020) and adversarial learning (Belinkov et al., 2019) for NLI, which have so far been

[9]See: https://github.com/CLUEbenchmark/CLUE

Maj.	CBOW	biLSTM	ESIM	BERT	RoBERTa
38.1	55.7 (0.5)	59.2 (0.5)	59.8 (0.4)	72.2 (0.7)	78.2 (0.7)

Table 5: Test performance on OCNLI for all baseline models. Majority label is *neutral*. We report the mean accuracy % across five training runs with random re-starts (the standard deviation is shown in parentheses).

Fine-tuning data / size		**OCNLI** / 50k		**XNLI** / 392k		**Combined** / 443k
Test data	size	BERT	RoBERTa	BERT	RoBERTa	RoBERTa
OCNLI human	300	90.3* (OCNLI.test)				
OCNLI.dev	3k	74.5 (0.3)	**78.8** (1.0)	66.8 (0.5)	70.5 (1.0)	76.4 (1.3)
OCNLI.test	3k	72.2 (0.7)	**78.2** (0.7)	66.7 (0.3)	70.4 (1.2)	75.6 (1.2)
CLUE diagnostics	0.5k	54.4 (0.9)	61.3 (1.3)	53.0 (0.9)	62.5 (2.9)	**63.7** (2.4)

Table 6: Accuracy on OCNLI, finetuned on OCNLI, XNLI and Combined (50k OCNLI combined with 392k XNLI). *: See Appendix E for details about the human baseline. As in Table 5, we report the mean accuracy % across five training runs with the standard deviation shown in parenthesis.

Test data	BERT	RoBERTa
OCNLI_dev	65.3	65.7
OCNLI_test	64.3	65.0
OCNLI_test_easy	63.5	64.0
OCNLI_test_medium	63.9	65.6
OCNLI_test_hard	65.5	65.5
MNLI	na.	62.0

Table 7: Hypothesis-only baselines for OCNLI (fine-tuned on OCNLI.train) and MNLI (retrieved from Bowman et al. (2020)).

limited to English and on much larger datasets that are easier to filter, we see extending these methods to our dataset and Chinese as an interesting challenge for future research.

Comparison with XNLI To ensure that our dataset is not easily solved by simply training on existing translations of MNLI, we show the performance of BERT and RoBERTa when trained on XNLI but tested on OCNLI. The results in Table 6 (column XNLI) show a much lower performance than when the systems are trained on OCNLI, even though XNLI contains 8 times more examples.[10] While these results are not altogether comparable, given that the OCNLI training data was generated from the same data sources and annotated by the same annotators (see Geva et al. (2019)), we still see these results as noteworthy given that XNLI is currently the largest available multi-genre NLI dataset for Chinese. The results are indicative of the limitations of current models trained solely on translated data. More strikingly, we find that when OCNLI and XNLI are combined for fine-tuning (column Combined in Table 6), this improves performance over the results using XNLI, but reaches lower accuracies than fine-tuning on the considerably smaller OCNLI (except for the diagnostics).

Figure 1 shows a learning curve comparing

[10]To ensure that this result is not unique to XNLI, we ran the same experiments using CMNLI, which is an alternative translation of MNLI used in CLUE, and found comparable results.

Word	Label	PMI	Counts
OCNLI			
任何 *any*	contradiction	1.02	439/472
从来 *ever*	contradiction	0.99	229/244
至少 *at least*	entailment	0.92	225/254
SINGLE			
任何 *any*	contradiction	0.89	87/90
没有 *no*	contradiction	0.83	582/750
无关 *not related*	contradiction	0.72	39/42
MULTI			
任何 *any*	contradiction	0.92	97/103
没有 *no*	contradiction	0.88	721/912
从来 *ever*	contradiction	0.75	42/46
MULTIENCOURAGE			
任何 *any*	contradiction	0.98	198/212
从来 *ever*	contradiction	0.96	131/137
至少 *at least*	entailment	0.82	81/91
MULTICONSTRAINT			
至少 *at least*	entailment	0.91	105/110
只有 *only*	contradiction	0.86	179/216
只 *only*	contradiction	0.77	207/280

Table 8: Top 3 `(word, label)` pairs according to PMI for different subsets of OCNLI.

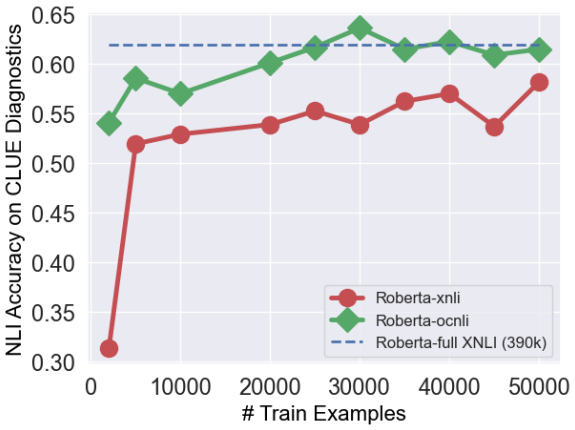

Figure 1: Ablation over the number of fine-tuning examples for RoBERTa fine-tuned on OCNLI vs. XNLI.

	SINGLE	MULTI	MULTIENC	MULTICON
BERT: fine-tune on XNLI				
dev_full	77.3	73.6	68.6	65.8
easy	na.	74.0	70.1	68.4
medium	na.	74.3	69.6	65.9
hard	na.	72.5	66.2	63.1
RoBERTa: fine-tune on XNLI				
dev_full	78.9	77.3	71.3	70.8
easy	na.	77.2	72.8	73.5
medium	na.	78.6	71.7	70.2
hard	na.	76.2	69.4	68.7

Table 9: Accuracy of XNLI-finetuned models, tested on relabelled parts of different OCNLI subsets.

model performance on the independent CLUE diagnostic test. Here we see that the OCLNI model reaches its highest performance at 30,000 examples while the XNLI model still shows improvements on 50,000 examples. Additionally, OCNLI reaches the same performance as the model finetuned on the full XNLI set, at around 25,000 examples. This provides additional evidence of the importance of having reliable human annotation for NLI data.

Understanding the OCNLI Subsets To better understand the effect of having three annotator hypotheses per premise, constituting three difficulty levels, and having four elicitation modes, we carried out a set of experiments with XNLI-finetuned models on the different subsets. We used XNLI to avoid imposing specific preferences on the models. Table 9 shows a consistent decrease in accuracy from SINGLE through MULTICONSTRAINT, and a mostly consistent decrease from easy to hard (exception: between easy and medium in MULTI). Both trends suggest that *multi-hypothesis* elicitation and improved instructions lead to more challenging elicited data.

5 Conclusion

In this paper, we presented the Original Chinese Natural Language Inference (OCNLI) corpus, the first large-scale, non-translated NLI dataset for Chinese. Our dataset is composed of 56,000 premise-hypothesis pairs, manually created by university students with a background in language studies, using premises from five genres and an enhanced protocol from the original MNLI annotation scheme. Results using BERT and RoBERTa show that our dataset is challenging for the current best pre-trained transformer models, the best of which is $\sim$ 12 percentage-points below human performance. We also demonstrate the relative advantage of using our human constructed dataset over machine translated NLI such as XNLI. To encourage more progress on Chinese NLU, we are making our dataset publicly available for the research community at https://github.com/CLUEbenchmark/OCNLI and will be hosting a leaderboard in the Chinese Natural Language Understanding (CLUE) (Xu et al., 2020) benchmark (https://www.cluebenchmarks.com/nli.html).

Given the wide impact that large-scale NLI datasets, such as SNLI and MNLI, have had on recent progress in NLU for English, we hope that our resource will likewise help accelerate progress on Chinese NLU. In addition to making more progress on Chinese NLI, future work will also focus on using our dataset for doing Chinese model probing (e.g., building on work such as Warstadt et al. (2019); Richardson et al. (2020); Jeretic et al. (2020)) and sentence representation learning (Reimers and Gurevych, 2019), as well as for investigating bias-reduction techniques (Clark et al., 2019; Belinkov et al., 2019; Le Bras et al., 2020) for languages other than English.

Acknowledgements

We thank all our annotators without whom this work wouldn't have been possible, and also Ruoze Huang, Jueyan Wu, Zhaohong Wu and Xiaojie Gong for their help in the annotation process. We are grateful for the suggestions from our anonymous reviewers and the CL colloquium at Indiana University. This work was supported by the CLUE benchmark and the Grant-in-Aid of Doctoral Research from Indiana University Graduate School.

Special thanks to the beaker team at AI2 for providing technical support for the beaker experiment platform. Computations on beaker.org were supported in part by credits from Google Cloud.

References

Lasha Abzianidze. 2015. A Tableau Prover for Natural Logic and Language. In *Proceedings of EMNLP*, pages 2492–2502.

Željko Agić and Natalie Schluter. 2017. Baselines and test data for cross-lingual inference. *Proceedings of LREC*.

Hossein Amirkhani, Mohammad Azari Jafari, Azadeh Amirak, Zohreh Pourjafari, Soroush Faridan Jahromi, and Zeinab Kouhkan. 2020. Farstail: A Persian Natural Language Inference Dataset. *arXiv preprint arXiv:2009.08820*.

Yonatan Belinkov, Adam Poliak, Stuart M Shieber, Benjamin Van Durme, and Alexander M Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. *Proceedings of ACL*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*.

Samuel R Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. Collecting entailment data for pretraining: New protocols and negative results. *arXiv preprint arXiv:2004.11997*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *Proceedings of ACL*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. *Proceedings of EMNLP*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of EMNLP*.

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 EMNLP*.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-Training with Whole Word Masking for Chinese BERT. *arXiv preprint arXiv:1906.08101*.

Ido Dagan, Oren Glickman, Alfio Gliozzo, Efrat Marmorshtein, and Carlo Strapparava. 2006. Direct word sense matching for lexical substitution. In *Proceedings of ACL*, pages 449–456.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognizing Textual Entailment*.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. In *Proceedings of CogSci*, pages 1596–1601.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading Failures of Partial-input Baselines. *Proceedings of ACL*.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference*, pages 13–15.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Modular Representation Underlies Systematic Generalization in Neural Natural Language Inference Models. *arXiv preprint arXiv:2004.14623*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *Proceedings of EMNLP-IJCNLP*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL*.

Yuta Hayashibe. 2020. Japanese realistic textual entailment corpus. In *Proceedings of LREC*.

Hai Hu, Qi Chen, Kyle Richardson, Atreyee Mukherjee, Lawrence S Moss, and Sandra Kuebler. 2020. MonaLog: a Lightweight System for Natural Language Inference Based on Monotonicity. In *Proceedings of SCiL*.

Hai Hu and Sandra Kübler. 2020. Investigating translated Chinese and its variants using machine learning. *Natural Language Engineering (Special Issue on NLP for Similar Languages, Varieties and Dialects)*, pages 1–34.

Hai Hu, Wen Li, and Sandra Kübler. 2018. Detecting syntactic features of translated Chinese. In *Proceedings of the 2nd Workshop on Stylistic Variation*, pages 20–28.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. *arXiv preprint arXiv:2004.03066*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTaiL: A Textual Entailment Dataset from Science Question Answering. In *AAAI*.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of ACL*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial Filters of Dataset Biases. *Proceedings of ICLR*.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of ACL*.

Chien-Jer Charles Lin. 2011. Chinese and English relative clauses: Processing constraints and typological consequences. In *Proceedings of the 23rd North American Conference on Chinese Linguistics (NACCL-23)*, pages 191–199.

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*.

Bill MacCartney. 2009. *Natural Language Inference*. Ph.D. thesis, Stanford University.

M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*.

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2017. On-demand injection of lexical knowledge for recognising textual entailment. In *Proceedings of EACL*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of ACL*.

Anthony McEnery and Zhonghua Xiao. 2004. The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In *LREC*, pages 1175–1178.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2011. Using Bilingual Parallel Corpora for Cross-lingual Textual Entailment. In *Proceedings of ACL*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress Test Evaluation for Natural Language Inference. In *Proceedings of COLING*.

Nikita Nangia and Samuel Bowman. 2019. Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark. In *Proceedings of ACL*.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and Conquer: Crowdsourcing the Creation of Cross-lingual Textual Entailment Corpora. In *Proceedings of EMNLP*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of ACL*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Proceedings of NAACL*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of *SEM*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv e-prints arXiv:1910.10683*.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. Organizing the ASSIN 2 shared task. In *Proceedings of the ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP*.

Kyle Richardson, Hai Hu, Lawrence S Moss, and Ashish Sabharwal. 2020. Probing Natural Language Inference Models through Semantic Fragments. In *Proceedings of AAAI*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WINOGRANDE: An adversarial Winograd schema challenge at scale. *Proceedings of AAAI*.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Arivazhagan, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. *Proceedings of AAAI*.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *Proceedings of ICLR*.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering Tasks. *Proceedings of NAACL*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*.

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 BlackboxNLP*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. *Proceedings of EMNLP*.

Barbara Wheatley. 1996. CALLHOME Mandarin Chinese Transcripts LDC96T16.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of NAACL*.

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. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of COLING*.

Hitomi Yanaka, Koji Mineshima, Pascual Martínez-Gómez, and Daisuke Bekki. 2018. Acquisition of Phrase Correspondences using Natural Deduction Proofs. In *Proceedings of NAACL*.

A Instructions for Hypothesis Generation

(*the instructions are originally in Chinese; translated to English for this paper*)

Welcome to our sentence writing experiment. Our aim is to collect data for making inferences in Chinese. In this experiment, you will see a sentence (A), which describes an event or a scenario, for example:

> Sentence A:
> **John won the first prize in his company's swimming competition last year.**

You task is to write three types of sentences based on the information in sentence A, as well as your common sense.

- Type 1: a sentence that is definitely true, based on the information in sentence A, e.g.,
  - John can swim
  - John won a prize last year
  - John's company held a swimming competition last year
- Type 2: a sentence that might be true (but might also be false), based on the information in sentence A, e.g.,
  - John's company held the swimming competition last March
  - Tom ranked second in last year's swimming competition
  - John can do the butterfly style
- Type 3: a sentence that cannot be true, based on the information in sentence A, e.g.,
  - John has not swum before
  - John did not get any prize from the company's swimming competition last year
  - John's company only hold table tennis competitions

You will see 50 sentence A. For each sentence A, you need to write three sentences, one for each type. In total you will write 150 sentences. If there is a problem with sentence A, please mark it as "x". Please refer to FAQ for more examples and further details of the task.

B Relabeling Results for Different Portions

In Table 10, we present labeler agreement for different portions of MULTI, MULTIENCOURAGE and MULTICONSTRAINT. We observe that the `medium` and `hard` portions in general have lower inter-annotator agreement, but still comparable to SNLI and MNLI. This suggests that writing three hypothese for each label is a feasible and reliable strategy.

Statistic	MULTI easy	MULTI medium	MULTI hard	MULTIENCOURAGE easy	MULTIENCOURAGE medium	MULTIENCOURAGE hard	MULTICONSTRAINT easy	MULTICONSTRAINT medium	MULTICONSTRAINT hard
# pairs relabelled	668	664	662	1,002	999	999	1,002	999	999
5 labels agree (unanimous)	66.5%	61.4%	62.5%	58.0%	56.5%	57.2%	60.8%	57.2%	54.9%
4+ labels agree	87.0%	82.1%	85.2%	82.2%	82.6%	81.2%	84.5%	78.6%	79.4%
3+ labels agree	**99.1**%	**99.1**%	**98.2**%	**98.5**%	**99.1**%	**98.4**%	**98.0**%	**99.0**%	**97.9**%
Indiv. label = gold label	90.1%	88.2%	88.5%	87.1%	87.3%	86.7%	87.9%	86.5%	85.6%
Indiv. label = author's label	84.5%	80.0%	82.4%	80.8%	80.8%	78.9%	82.2%	79.2%	77.6%
Gold label = author's label	91.5%	88.1%	89.3%	90.4%	91.4%	87.1%	90.1%	88.3%	86.1%
Gold label ≠ author's label	7.6%	11.0%	8.9%	8.1%	7.7%	11.3%	7.9%	10.7%	11.8%
No gold label	0.9%	0.9%	1.8%	1.5%	0.9%	1.6%	2.0%	1.0%	2.1%
%n_unrelated labels	0.2%	0.2%	0.4%	0.2%	0.6%	0.3%	0.1%	0.1%	0.4%

Table 10: Labeling results for different portions of MULTI, MULTIENCOURAGE and MULTICONSTRAINT.

C Relabeling Results for XNLI Development Set

For this experiment, we follow the same procedure as the relabeling experiment for OCNLI data. We randomly selected 200 examples from XNLI dev, and mixed them with 200 examples from our SINGLE (which has already been relabelled) for another group of annotators to label. The labelers for these 400 pairs were undergraduate students who did *not* participated in hypothesis generation so as to avoid biasing towards our data.

The labeling results for XNLI are presented in Table 11. Only 67% of the 200 pairs have the same label from our annotators and the label given in XNLI dev. 8.5% of the pairs are considered to be irrelevant by the majority of our annotators. As we mentioned in the introduction, there are other issues with XNLI such as the existence of many Roman alphabets (867 (11.56%) examples in XNLI dev have more than 10 Roman alphabets) which prevent us from using it as proper evaluation data for Chinese NLI.

Statistic	XNLI dev	SINGLE
# pairs relabelled (i.e., validated)	200	200
majority label = *original* label	**67.0**%	**84.0**%
5 labels agree (excl. "unrelated")	38.5%	57.5%
4+ labels agree (excl. "unrelated")	57.5%	83.5%
3+ labels agree (excl. "unrelated")	**86.0**%	**98.0**%
5 labels agree	41.0%	57.5%
4+labels agree	62.0%	83.5%
3+ labels agree	94.5%	98.0%
majority label = "unrelated"	8.5%	0%
# individual "unrelated" labels	125	11
# incomprehensible note	22	4

Table 11: Results for labeling a mixture of 200 pairs of XNLI dev Chinese and 200 pairs of SINGLE, by labelers who did not participated in the hypothesis generation experiment. Note the XNLI dev is translated by crowd translators (Conneau et al., 2018), not MT systems. The *original* label for XNLI dev Chinese comes with XNLI, which is the same for all 15 languages. The *original* label for SINGLE comes from our relabeling experiments.

D Model Details and Hyper-parameters

We experimented with the following models:

- Continuous bag-of-words (CBOW), where sentences are represented as the sum of its Chinese character embeddings, which are passed on to a 3-layer MLP.
- Bi-directional LSTM (biLSTM), where the sentences are represented as the average of the states of a bidirectional LSTM.
- Enhanced Sequential Inference Model (ESIM), which is MNLI's implementation of the ESIM model (Chen et al., 2017).
- BERT base for Chinese (BERT), which is a 12-layer transformer model with a hidden size of 768, pre-trained with 0.4 billion tokens of the Chinese Wikipedia dump (Devlin et al., 2019). We use the implementation from the CLUE benchmark (Xu et al., 2020)[11].
- RoBERTa large pre-trained with whole word masking (wwm) and extended (ext) data (RoBERTa), which is based on RoBERTa (Liu et al., 2019) and has 24 layers with a hidden size of 1024, pre-trained with 5.4 billion tokens, released in (Cui et al., 2019). We use the implementation from the CLUE benchmark.

For CBOW, biLSTM and ESIM, we use Chinese

[11] https://www.cluebenchmarks.com/

character embeddings from Li et al. (2018)[12], and modify the implementation from MNLI[13] to work with Chinese.

Our BERT and RoBERTa models are both fine-tuned with 3 epochs, a learning rate of 2e-5, and a batch size of 32. Our hyper-parameters deviate slightly from those used in CLUE and (Cui et al., 2019)[14], because we found them to be better when tuned against our dev sets (as opposed to XNLI or the machine translated CMNLI in CLUE).

E Determining Human Baselines

We follow procedures in Nangia and Bowman (2019) to obtain *conservative* human baselines on OCNLI. Specifically, we first prepared 20 training examples from OCNLI.train and instructions similar to those in the relabeling experiment. Then we asked 5 undergraduate students who did *not* participate in any part of our previous experiment to perform the labeling. They were first provided with the instructions as well as the 20 training examples, which they were asked to label after reading the instructions. Then they were given the answers and explanations of the training examples. Finally, they were given a random sample of 300 examples from the OCNLI test set for labeling. We computed the majority label from them, and compare that against the gold label in OCNLI.test to obtain the accuracy. For pairs with no majority label, we use the most frequent label from OCNLI.test (neutral), following Nangia and Bowman (2019). We have only 2 (0.7%) such cases. The results are presented in Table 12.

We performed the same experiment with 5 linguistics PhDs, who are already familiar with the NLI task from their research, and thus their results may be biased. We see a higher 5-label agreement and similar accuracy compared against the gold label of OCNLI.test. We use the score from undergraduate students as our human baseline as it is the "unbiased" score obtained using the same procedure as Nangia and Bowman (2019).

The human score of OCNLI is similar to that of MNLI (92.0%/92.8% for match and mismatch respectively).

F More Examples from OCNLI

We present more OCNLI pairs in Table 13.

[12] https://github.com/Embedding/Chinese-Word-Vectors
[13] https://github.com/NYU-MLL/multiNLI
[14] https://github.com/ymcui/Chinese-BERT-wwm/

annotator	undergrad	linguistics PhD
# pairs anno.	300	300
accuracy (agree w/ OCNLI.test)	**90.3**	**89.3**
5-label agree	55.3	60.6
4-label agree	82.0	83.3
3-label agree	99.3	99.0
no majority	0.7	1.0

Table 12: Human score for OCNLI

G Filtering training data

To mimic the MNLI setting where the training data and the evaluation data (dev/test) have no overlapping premises, we filtered out the pairs in the current training set whose premise can also be found in evaluation. This means the removal of about 20k pairs in OCNLI.train, and results in a new training set which we call OCNLI.train.small, while the development and test sets remain the same. We fine-tune the biLSTM, BERT and RoBERTa models on the new, filtered training data, and the results are presented in Table 14.

We observe that our models have a 1.5-2.5% drop in performance when trained with the filtered training data. Note that OCNLI.train.small is only 60% of OCNLI.train in size, so we consider this drop to be moderate and expected, and more likely the result of reduced training data, rather than the removal of overlapping premises.

We will release both training sets publicly and also in our public leaderboard (https://www.cluebenchmarks.com/nli.html). We note that similar strategies for controlling dataset size have been used for WINOGRANDE project (Sakaguchi et al., 2020) and their leaderboard (https://leaderboard.allenai.org/winogrande/submissions/public).

Premise	Genre Level	**Majority label** All labels	Hypothesis
是，你看他讲这个很有意思 Yes, look, what he talked about is very interesting.	TV hard	**Entailment** E E N E E	他讲的这个引起了我的关注 What he talked about has caught my attention.
要根治拖欠农民工工资问题，抓紧制定专门行政法规，确保付出辛劳和汗水的农民工按时拿到应有的报酬 (We need to) solve the problem of delaying wages for the migrant workers at its root and act promptly to lay out specific administrative regulations to ensure those hardworking migrant workers receive the wages they deserve in a timely manner.	GOV easy	**Neutral** N E N N N	专门行政法规是解决拖欠工资问题的根本途径 (Designing) specific administrative regulations is the most fundamental way of solving the issue of wage delays.
你要回来啦,就住在我这老房. If you are back, you can stay in my old house.	PHONE hard	**Contradiction** C C C C C	我没房 I don't have a house.
十月底回去,十一月份底回来. Going there at the end of October, be back at the end of November.	PHONE medium	**Contradiction** C C C C C	要在那边呆两个月才回来。 Will stay there for two months before coming back.
呃,对,我大概有,这. Er, yes, I may have (it), here.	PHONE hard	Neutral N N N N N	是别人想问我借这个东西 Someone else is trying to borrow this from me.
桥一顶一顶地从船上过去，好像进了一扇一扇的门 Bridge after bridge was passed above the boat, just like going through door after door.	LIT medium	**Entailment** E E E E E	有不止一座桥 There is more than one bridge.
此间舆论界普遍认为，这次自民党在众议院议席减少无疑，问题在于减少多少 It is generally believed by the media that the Liberal Democratic Party are going to lose their seats. The problem is how many.	NEWS medium	**Contradiction** C C C N N	舆论界普遍认为，这次自民党要被驱逐出众议院。 It is generally believed by the media that the Liberal Democratic Party will be ousted from the House of Representatives.

Table 13: More examples from OCNLI.

Train data / size	**OCNLI.train / 50k**			**OCNLI.train.small / 30k**		
	biLSTM	BERT	RoBERTa	biLSTM	BERT	RoBERTa
OCNLI.dev	60.5 (0.4)	74.5 (0.3)	78.8 (1.0)	58.7 (0.3)	72.6 (0.9)	77.4 (1.0)
OCNLI.test	59.2 (0.5)	72.2 (0.7)	78.2 (0.7)	57.0 (0.9)	70.3 (0.9)	76.4 (1.2)

Table 14: Comparison of models performance fine-tuned on OCNLI.train and OCNLI.train.small. As before, we report the mean accuracy % across five training runs with the standard deviation shown in parenthesis.