Investigating Transfer Learning in Multilingual Pre-trained Language Models through Chinese Natural Language Inference

Hai Hu[†⁎] **He Zhou**[†] **Zuoyu Tian**[†] **Yiwen Zhang**[†] **Yina Ma**[‡]
Yanting Li[◁] **Yixin Nie**[◦] **Kyle Richardson**[⋄]
[†]Indiana University Bloomington [‡]Brigham Young University [⁎]Shanghai Jiaotong Univ.
[◁]Northwestern University [◦]UNC Chapel Hill [⋄]Allen Institute for AI
hu.hai@outlook.com; {hzh1,zuoytian,yiwezhan,yinama}@iu.edu;
yanting.li@northwestern.edu; yixin1@cs.unc.edu; kyler@allenai.org

Abstract

Multilingual transformers (XLM, mT5) have been shown to have remarkable transfer skills in zero-shot settings. Most transfer studies, however, rely on automatically translated resources (XNLI, XQuAD), making it hard to discern the particular linguistic knowledge that is being transferred, and the role of expert annotated monolingual datasets when developing task-specific models. We investigate the cross-lingual transfer abilities of XLM-R for Chinese and English natural language inference (NLI), with a focus on the recent large-scale Chinese dataset OCNLI. To better understand linguistic transfer, we created 4 categories of challenge and adversarial tasks (totaling 17 new datasets[1]) for Chinese that build on several well-known resources for English (e.g., *HANS*, NLI *stress-tests*). We find that cross-lingual models trained on English NLI do transfer well across our Chinese tasks (e.g., in 3/4 of our challenge categories, they perform as well/better than the best monolingual models, even on 3/5 uniquely Chinese linguistic phenomena such as *idioms, pro drop*). These results, however, come with important caveats: cross-lingual models often perform best when trained on a mixture of English and high-quality monolingual NLI data (OCNLI), and are often hindered by automatically translated resources (XNLI-zh). For many phenomena, all models continue to struggle, highlighting the need for our new diagnostics to help benchmark Chinese and cross-lingual models.

1 Introduction

Recent pre-trained multilingual transformer models, such as XLM(-R) (Conneau and Lample, 2019; Conneau et al., 2020), mT5 (Xue et al., 2020) and others (Liu et al., 2020; Lewis et al., 2020) have been shown to be successful in NLP tasks for several non-English languages (Khashabi et al., 2020; Choi et al., 2021), as well as in multilingual benchmarks (Devlin et al., 2019; Conneau et al., 2020; Xue et al., 2020; Artetxe et al., 2020). A particular appeal is that they can be used for *cross-lingual* and *zero-shot transfer*. That is, after pre-training on a raw, unaligned corpus consisting of text from many languages, models can be subsequently fine-tuned on a particular task in a resource-rich language (e.g., English) and directly applied to the same task in other languages without requiring any additional language-specific training.

Given this recent progress, a natural question arises: does it make sense to invest in large-scale task-specific dataset construction for low-resourced languages, or does cross-lingual transfer alone suffice for many languages and tasks? A closely related question is: how well do multilingual models transfer across specific linguistic and language-specific phenomena? While there has been much recent work on probing multilingual models (Wu and Dredze, 2019; Pires et al., 2019; Karthikeyan et al., 2019), *inter alia*, a particular limitation is that most studies rely on automatically translated resources such as XNLI (Conneau et al., 2018) and XQuAD (Artetxe et al., 2020), which makes it difficult to discern the particular linguistic knowledge that is being transferred and the role of large-scale, expert annotated monolingual datasets when building task- and language-specific models.

In this paper, we investigate the cross-lingual transfer abilities of XLM-R (Conneau et al., 2020) for Chinese natural language inference (NLI). Our focus on Chinese NLI is motivated by the recent release of the first large-scale, human-annotated Chinese NLI dataset OCNLI (**O***riginal* **C***hinese* **NLI**) (Hu et al., 2020)[2], which we use to directly in-

[1]All new datasets/code are released at https://github.com/huhailinguist/ChineseNLIProbing.

[2]To our knowledge, OCNLI is currently the largest non-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3770–3785
August 1–6, 2021. ©2021 Association for Computational Linguistics

	category	n
Chinese HANS	Lexical overlap	1,428
	Subsequence	513
stress tests	Distraction: 2 categories	8,000
	Antonym	3,000
	Synonym	2,000
	Spelling	11,676
	Numerical reasoning	8,613
diagnostics	CLUE (Xu et al., 2020)	514
	CLUE expansion (ours)	796
	World knowledge (ours)	38
	Classifier (ours)	139
	Chengyu/idioms (ours)	251
	Pro-drop (ours)	198
	Non-core arguments (ours)	186
semantic probing	Negation	1,002
	Boolean	1,002
	Quantifier	1,002
	Counting	1,002
	Conditional	1,002
	Comparative	1,002
sum		43,364

Table 1: Summary statistics of the four evaluation sets.

vestigate the role of high-quality task-specific data vs. English-based cross-lingual transfer. To better understand linguistic transfer, and help benchmark recent SOTA Chinese NLI models, we created 4 categories of challenge/adversarial tasks (totaling 17 new datasets) for Chinese that build on several well-established resources for English and the literature on model probing (see Poliak (2020)). Our new resources, which are summarized in Table 1, include: a new set of diagnostic tests in the style of the SuperGLUE (Wang et al., 2019) and CLUE (Xu et al., 2020) diagnostics; Chinese versions of the HANS dataset (McCoy et al., 2019) and NLI stress-tests (Naik et al., 2018), as well as a collection of the basic reasoning and logic *semantic probes* for Chinese based on Richardson et al. (2020).

Our results are largely positive: We find that cross-lingual models trained exclusively on English NLI do transfer relatively well across our new Chinese tasks (e.g., in 3/4 of the challenge categories shown in Table 1, they perform overall as well or better than the best monolingual Chinese models without additional specialized training on Chinese data, and have competitive performance on OCNLI). A particularly striking result is that such models even perform well on 3/5 uniquely Chinese linguistic phenomena such as *idioms, pro drop*, providing evidence that many language-specific phenomena do indeed transfer. These results, however, come with important caveats: on several phenomena we find that models continue to struggle and are far outpaced by conservative estimates of human performance (e.g., our best model on Chinese HANS remains ~19% behind human performance), highlighting the need for more language-specific diagnostics tests. Also, fine-tuning models on mixtures of English NLI data and high-quality monolingual data (OCNLI) consistently performs the best, whereas mixing with automatically translated datasets (XNLI-zh) can greatly hinder model performance. This last result shows that high-quality monolingual datasets still play an important role when building cross-lingual models, however, the particular type of monolingual dataset that is needed can vary and is best informed by targeted behavioral testing of the type we pursue here.

English NLI dataset that was annotated in the style of English MNLI without any translation.

2 Related Work

There has been a lot of work on trying to understand multilingual transformers (Wu and Dredze, 2019; Pires et al., 2019), which has focused on either examining the representation of different layers in the transformer architecture or the lexical overlap between languages. Karthikeyan et al. (2019) investigate the role of network depth and number of attention heads, as well as syntactic/word-order similarity on the cross-lingual transfer performance. In addition to studies cited at the outset, positive results of cross-lingual transfer across a wide range of languages are reported in Wu and Dredze; Nozza et al. (2020), with a focus on transfer across specific tasks such as POS tagging, NER; in contrast, we focus on different categories of linguistic transfer, which has received less attention, as well as the role of monolingual data for transfer in NLI.

Studies into the linguistic abilities and robustness of current NLI models have proliferated in recent years, partly owing to the discovery of systematic biases, or *annotation artifacts* (Gururangan et al., 2018; Poliak et al., 2018), in benchmark NLI datasets such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). This has been coupled with the development of new *adversarial tests* such as *HANS* (McCoy et al., 2019) and the NLI *stress-tests* (Naik et al., 2018), as well as several new linguistic *challenge datasets* (Glockner et al., 2018; Richardson et al., 2020; Geiger et al., 2020; Yanaka et al., 2019; Saha et al., 2020; Goodwin et al., 2020), *inter alia*, that focus on a wide range of linguistic and reasoning phenomena. All of this work focuses exclusively on English, whereas we focus on constructing analogous probing datasets tailored to Chinese to help advance research on Chinese NLI and cross-lingual transfer.

There has been a surge in the development of NLI resources for languages other than English. Such resources are often created in the following two ways: (1) from *scratch*, in the style of MNLI (Williams et al., 2018), where annotators are used to produce hypotheses and inference labels based on a provided set of premises, as pursued for Chinese OCNLI (Hu et al., 2020), or SciTail (Khot et al., 2018), where sentences are paired automatically and labeled by annotators (Amirkhani et al., 2020; Hayashibe, 2020). (2) Through *automatic* (Conneau et al., 2018; Budur et al., 2020; Real et al., 2020) or *manual* (Wijnholds and Moortgat, 2021) translation from existing English datasets. Studies on cross-lingual transfer for NLI have largely focused on XNLI (Conneau et al., 2018), which we show has limited utility for Chinese NLI transfer.

3 Dataset creation

In this section, we describe the details of the 4 types of challenge datasets we constructed for Chinese to study cross-lingual transfer (see details in Table 1). They fit into two general categories: **Adversarial datasets** (Section 3.1) built largely from patterns in OCNLI (Hu et al., 2020) and XNLI (Conneau et al., 2018) and **Probing/diagnostic datasets** (Section 3.2), which are built from scratch in a parallel fashion to existing datasets in English.

While we aim to mimic the annotation protocols pursued in the original English studies, we place the additional methodological constraint that each new dataset is vetted, either through human annotation using a disjoint set of Chinese linguists, or through internal mediation among local Chinese experts; details are provided below.

3.1 Adversarial dataset

Examples from the 7 adversarial tests we created are illustrated in Table 2.[3] Chinese HANS is built from patterns extracted in the large-scale Chinese NLI dataset OCNLI (Hu et al., 2020), whereas the **Distraction**, **Antonym**, **Synonym** and **Spelling** subsets are built from an equal mixture of OCNLI and XNLI-zh (Conneau et al., 2018) data; in the latter case, such a difference allows us to fairly compare the effect of training on expert-annotated (i.e., OCNLI) vs. automatically translated data (i.e., XNLI-zh) as detailed in Section 4.

Chinese HANS McCoy et al. (2019) discovered systematic biases/heuristics in the MNLI dataset, which they named "lexical/subsequence/constituent" overlap. "Lexical overlap" is defined to be the pairs where the vocabulary of the hypothesis is a subset of the vocabulary of the premise. For example, "*The boss is meeting the client.*" and "*The client is meeting the boss.*", which has an entailment relation. However, lexical overlap does not necessarily mean the premise will entail the hypothesis, e.g., "*The judge was paid by the actor.*" does not entail "*The actor was paid by the judge.*" (examples from McCoy et al. (2019)). Thus a model relying on the heuristic will fail catastrophically in the second case.

Inspired by the English HANS, we examine whether OCNLI also possesses such biases, as it has a similar annotation procedure as MNLI. We follow the design of the original HANS experiments, and adapt their scripts[4] to extract examples in OCNLI that satisfy the two heuristics. We find a heavy bias towards "entailment", where 79.5% of such examples are "entailment", similar to MNLI. To construct a Chinese HANS, we first look into syntactic structures of the examples having the two heuristics. Then we write 29 templates for the *lexical overlap* heuristic and 11 templates for *subsequence overlap*.[5] Using the templates and a vocabulary of 263 words, we generated 1,941 NLI pairs. See Table 2 for examples and Appendix A for details.

Distraction We add distractions to the premise or hypothesis, similar to the "length mismatch" and "word overlap" conditions in the NLI *stress tests* of Naik et al. (2018). The distractions are either tautologies ("true is not false") or a true statement from our world knowledge ("Finnland is not a permanent member of the UN security council"), which should not influence the inference label. We control whether the distraction contain a negation or not, and thus create four conditions: *premise-negation*, *premise-no-negation*, *hypothesis-negation*, and *hypothesis-no-negation*. See Table 2 for examples.

[3]A more detailed description of the data creation process can be found in Appendix A.

[4]https://github.com/tommccoy1/hans

[5]For details of the templates, see our Github repository.

	category	n	premise	hypothesis	label
Chinese HANS	Lexical overlap	1428	我们把银行职员留在电影院了。We left the bank clerk in the cinema.	银行职员把我们留在电影院了。The bank clerk left us in the cinema.	C
	Subsequence	513	谁说律师都是穿西装的。Who told you that all lawyers wear suits.	律师都是穿西装的。All lawyers wear suits.	C
stress tests	Distraction (add to premise)	4000	国有企业改革的思路和方针政策已经明确,而且刚做完手术出院的病人不应剧烈运动。The policy of the reform of state-owned enterprises is now clear, and patients who just had surgery shouldn't have intense exercise.	根本不存在国有企业。The state-owned enterprises don't exist.	C
	Distraction (add to hypothesis)	4000	这时李家院子挤满了参观的人。During this time, the Li family's backyard is full of people who came to visit.	这地方有个姓李的人家,而且真的不是假的。There is a Li family here, and true is not false.	E
	Antonym	3000	一些地方财政收支矛盾较大。The disagreement about local revenue is relatively big.	一些地方财政收支矛盾较小。The disagreement about local revenue is relatively small.	C
	Synonym	2000	海部组阁困难说明了什么。What can you tell from the difficulties from Kaifu's attempt to set up a cabinet?	海部组阁艰难说明了什么。What can you tell from the hardships from Kaifu's attempt to set up a cabinet?	E
	Spelling	2980	身上裹一件工厂发的棉大衣,手插在袖筒里。(Someone is) wrapped up in a big cotton coat the factory gave with hands in the sleeves	身上质少一件衣服。There's at least [typo] one coat on the body.	E
	Numerical reasoning	8613	小红每分钟打不到510个字。Xiaohong types fewer than 510 words per min.	小红每分钟打110个字。Xiaohong types 110 words per min.	N

Table 2: Example NLI pairs in Chinese HANS and stress tests with translations.

Antonym We replace a word in the premise with its antonym to form a contradiction. To ensure the quality of the resulting NLI pairs, we manually examine the initially generated data and decided to only replace nouns and adjectives, as they are more likely to produce real contradictions.

Synonym We replace a word in the premise with its synonym to form an entailment.

Spelling We replace one random character in the hypotheses with its homonym (character with the same *pinyin* pronunciation ignoring tones) as this is one of the most common types of misspellings in Chinese.

Numerical reasoning We create a probing set for numerical reasoning, following simple heuristics such as the following. When the premise is *Mary types x words per minute*, the entailed hypothesis can be: *Mary types less than y words per minute*, where $x < y$. A contradictory hypothesis: *Mary types y words per minute*, where $x > y$ or $x < y$. Then a neutral pair can be produced by reversing the premise and hypothesis of the above entailment pair. 4 heuristic rules (with 6 words for quantification) are used and the seed sentences are extracted from Ape210k (Zhao et al., 2020), a dataset of Chinese elementary-school math problems. The resulting data contains 8,613 NLI pairs.

For **quality control** and to compute human performance, we randomly sampled 50 examples from all subsets and asked 5 Chinese speakers to verify. Our goal is to mimic the human annotation protocol from Nangia and Bowman (2019), which gives us a *conservative* estimate of human performance given that our annotators received very little instructions. Their majority vote agrees with the gold label 90.0% of the time, which suggests that our data is of high quality and allows us to later compare against model performance.[6]

3.2 Probing/diagnostic datasets

While the Chinese HANS and stress tests are designed to adversarially test the models, we also create probing or diagnostic datasets which are aimed at examining the models' linguistic and reasoning abilities.

Hand-crafted diagnostics We expanded the diagnostic dataset from the Chinese NLU Benchmark (CLUE) (Xu et al., 2020) in the following two ways:

First, 6 Chinese linguists (PhD students) created diagnostics for 4 Chinese-specific linguistic phenomena. Here are two of the phenomena:[7] (1) *pro-drop*: subjects or objects in Chinese can be dropped when they can be recovered from the context (Li et al., 1981). Thus the model needs to figure out the subject/object from the context. (2) *four-character idioms* (i.e., 成语 *Chengyu*). They are a special type of Chinese idioms that has exactly four characters, usually with a figurative meaning different from the literary meaning, e.g., 打草惊蛇 *hit hay startle snake* (behaving carelessly and causing your enemy to become vigilant). We construct examples to test whether models understand the figurative meaning in the idioms. Specifically, we first create a premise $\mathcal{P}$ which includes the idiom, where there is enough context so that a human is highly likely to

[6]Specifically: 98.0% on Chinese HANS, 86.0% on the stress tests. For comparison, different subsets of the English stress tests receives 85% to 98% agreement (Naik et al., 2018).

[7]For the other two, please refer to Appendix A.

interpret the idiom figuratively. Then we create an entailed hypothesis that is based on the figurative (correct) interpretation, and a neutral/contradictory hypothesis that uses the literal (incorrect) meaning (see Table 11 in the Appendix for an example). For each $\mathcal{P}$ we write 3 hypothesis, one for each inference relation. We also added diagnostics involving world knowledge.

Second, we double the number of diagnostic pairs for all 9 existing linguistic phenomena in CLUE with pairs whose premises are selected from a large news corpus[8] and hypotheses are handwritten by our linguists, to accompany the 514 artificially created data in CLUE. The resulting new diagnostics is 4 times as large as the original one, with a total of 2,122 NLI pairs. For quality control, each pair is double-checked by local Chinese linguists not involved in this study and the controversial cases were discarded after a discussion among the 6 linguists. See Table 11 in Appendix A for examples.

Semantic fragments Following Richardson et al. (2020) and Salvatore et al. (2019), we design synthesized fragments to examine models' understanding ability of six types of linguistic and logic inference: **boolean**, **comparative**, **conditional**, **counting**, **negation** and **quantifier**, where each category has 2-4 templates. See example templates and NLI pairs in Table 3.

The data is generated using context-free grammar rules and a vocabulary of 80,000 person names (Chinese and transliterated), 8659 city names and expanded predicates and comparative relations in Richardson et al. (2020) to make the data more challenging. As a result, we generated 1,000 examples for each fragment. For quality control, each template was checked by 3 linguists/logicians; also 20 examples from each category were checked for correctness by local experts.

4 Experimental setup

Our main goal is to test whether cross-lingual transfer are robust against the adversarial and probing data we created when evaluated without additional training. Thus we need to compare the best Chinese monolingual models with the best multilingual models trained either on English NLI data alone, or on combinations of Chinese and English data.[9]

Chinese monolingual models We experimented with two current state-of-the-art transformer models: RoBERTa-large (Liu et al., 2019) and Electra-large-discriminator (Clark et al., 2019). We use the Chinese models released from (Cui et al., 2020)[10] implemented the Huggingface Transformer library (Wolf et al., 2020).

Multilingual model We use XLM-RoBERTa-large (Conneau et al., 2020). We choose XLM-R over mT5 (Xue et al., 2020) because XLM-R generally performs better than mT5 under the same model size (see original paper for details). Also, XLM-R as a RoBERTa model is most related architecturally to existing Chinese pre-trained models.

Fine-tuning data for Chinese models & XLM-R (1) XNLI: the full Chinese training set in the machine-translated XNLI dataset, with 390k examples (Conneau et al., 2018). (2) XNLI-small: 50k examples from XNLI, the same size as the training data of OCNLI. (3) OCNLI: Original Chinese NLI dataset (Hu et al., 2020). It is a Chinese NLI dataset collected from scratch, following the MNLI procedure, with 50k training examples. We use this to measure the effect of the quality of training data; that is, whether it is better to use small, high-quality training data (OCNLI), or large, low-quality MT data (XNLI). (4) OCNLI + XNLI: a combination of the two training sets, 440k examples.

Fine-tuning data for XLM-R To examine cross-lingual transfer, we finetune XLM-R on English NLI data alone and English + Chinese NLI data: (1) MNLI: 390k examples from MNLI.train (Williams et al., 2018). (2) English all NLI: we combine MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), FeverNLI (Thorne et al., 2018; Nie et al., 2019) with ANLI (Nie et al., 2020), a total of 1,313k examples. (3) OCNLI + English all NLI. (4) XNLI + English all NLI. These two are set to examine whether combining Chinese and English fine-tuning data is helpful.

[8] We use the BCC corpus (Xun et al., 2016): http://bcc.blcu.edu.cn/.

[9] We also run the same experiments for Chinese-to-English transfer, i.e., fine-tuning XLM-R with OCNLI and evaluate on the four English counterpart datasets. We find that transferring from OCNLI to English does not perform as well as monolingual English models, likely due to the small size of OCNLI. Detailed results are reported in Appendix C.

[10] We use `hfl/chinese-roberta-wwm-ext-large` from https://github.com/ymcui/Chinese-BERT-wwm and `hfl/chinese-electra-large-discriminator` from https://github.com/ymcui/Chinese-ELECTRA.

category	premise	hypothesis	label
Negation	库尔图尔只到过湛江市麻章区，丰隆格只到过大连市普兰店区…… $person_1$ only went to $location_1$; $person_2$ only went to $location_2$;	库尔图尔没到过大连市普兰店区。 $person_1$ has not been to $location_2$.	E
Boolean	何峥、管得宽、李国柱……只到过临汾市襄汾县。 $person_1$, $person_2$ ... have only been to $location_1$.	何峥没到过遵义市红花岗区。 $person_1$ has not been to $location_2$.	E
Quantifier	有人到过每一个地方，拥抱过每一个人。 Someone has been to every place and hugged every person.	王艳没拥抱过包一。$person_1$ hasn't hugged $person_2$.	N
Counting	韩声雄只拥抱过罗冬平、段秀芹……赵常。 $person_1$ only hugged $person_2$, $person_3$ ... $person_8$.	韩声雄拥抱过超过10个人。 $person_1$ hugged more than 10 people.	C
Conditional	……，穆肖贝夸到过赣州市定南县，如果穆肖贝夸没到过赣州市定南县，那么张本伟到过呼伦贝尔市阿荣旗。... $person_n$ has been to $location_n$. If $person_n$ hasn't been to $location_n$, then $person_m$ has been to $location_m$.	张本伟没到过呼伦贝尔市阿荣旗。$person_m$ hasn't been to $location_m$.	N
Comparative	龙银凤比武书瑾、卢耀辉……奈德哈特都小，龙银凤和亚厄纳尔普一样大。$person_1$ is younger than $person_2$, ..., $person_n$; $person_1$ is as old as $person_m$	亚厄纳尔普比梁培娟大。$person_m$ is older than $person_{n-2}$.	C

Table 3: Example NLI pairs for semantic/logic probing with translations. Each label for each category has 2 to 4 templates; we are only showing 1 template for 1 label. 1,000 examples are generated for each category.

Model	Fine-tuned on	Acc	Scenario
RoBERTa	zh MT: XNLI-small	67.44	monolingual
RoBERTa	zh MT: XNLI	70.29	monolingual
RoBERTa	zh ori: OCNLI	79.11	monolingual
RoBERTa	zh: OCNLI + XNLI	78.43	monolingual
XLM-R	zh MT: XNLI	72.55	monolingual
XLM-R	zh ori: OCNLI	79.24	monolingual
XLM-R	zh: OCNLI + XNLI	80.31	monolingual
XLM-R	en: MNLI	71.98	zero-shot
XLM-R	en: En-all-NLI	73.73	zero-shot
XLM-R	mix: OCNLI + En-all-NLI	**82.18**	mixed
XLM-R	mix: XNLI + En-all-NLI	74.12	mixed

Table 4: Results on OCNLI dev. "**Scenario**" indicates whether the model is fine-tuned on Chinese *only* data (**monolingual**), English data (**zero-shot**) or **mixed** English and Chinese data; results in gray show best performance for each scenario. Best overall result in **bold**. Same below.

We fine-tune the models on OCNLI-dev. Acknowledging that different training runs can produce very different checkpoints for behavioral testing (D'Amour et al., 2020), we run 5 models on different seeds and report the mean accuracy of the models with the best hyper-parameter setting (for details see Appendix B).

5 Results and discussion

5.1 Results on OCNLI_dev

Results on the dev set of OCNLI are presented in Table 4. For monolingual RoBERTa, we see a similar performance as reported in the OCNLI paper (Hu et al., 2020), with 79.11% accuracy. The monolingual Electra achieves a very close accuracy of 79.02% (not shown in the Table). As we see the same trend in the following experiments, we will therefore only report results on RoBERTa.

For XLM-R, fine-tuning on MNLI or En-all-NLI gives us reasonable results of around 72% to 74%, which is better than models fine-tuned on XNLI, indicating that fine-tuning on an English data (MNLI) alone can outperform monolingual models fine-tuned on the same data but machine-translated into Chinese (XNLI).[11] This is consistent with previous results on Korean (Choi et al., 2021) and Persian (Khashabi et al., 2020) for other NLU tasks.

What is also interesting is that combining OCNLI and En-all-NLI gives us a boost of 2% to 82.18% (a result that is comparable to the current published SOTA), showing the power of mixing high-quality English and Chinese training data.

5.2 Chinese HANS

Table 5 shows results of the Chinese HANS data tested on the aforementioned monolingual models and cross-lingual model.

Cross-lingual transfer achieves strong results. We first notice that when XLM-R is fine-tuned solely on the English data (En-all-NLI), the performance (~69%) is only slightly worse than the best monolingual model (~71%). This suggests that cross-lingual transfer from English to Chinese is quite successful for an adversarial dataset like HANS. Second, adding OCNLI to En-all-NLI in the training data gives a big boost of about 9%, and achieves the overall best result. This is about 12% higher than combining XNLI and the English data, demonstrating the advantage of the expert-annotated OCNLI over machine translated XNLI, even though the latter is about 8 times the size of the former. Despite these results, however, we note that all models continue to perform below human performance, suggesting more room for improvement.

Our results also suggest that examples involving the *sub-sequence* heuristics are more difficult than

[11] For these experiments we also tested with another Chinese machine-translated MNLI (CMNLI), translated by a different MT system, which was released by CLUE (https://github.com/CLUEbenchmark/CLUE), and obtained similar results.

Model	Fine-tuned on	Overall	Lexical Overlap	Sub-sequence	Entailment	Non-Entailment	Δ
RoBERTa	zh MT: XNLI-small	49.48	58.12	25.42	99.22	30.26	37.18
RoBERTa	zh MT: XNLI	60.80	68.99	38.01	99.74	45.76	24.53
RoBERTa	zh ori: OCNLI	71.72	75.39	61.48	99.67	60.91	18.20
RoBERTa	zh: OCNLI+XNLI	69.33	74.73	54.27	99.89	57.51	20.92
XLM-R	zh ori: OCNLI	61.82	65.83	50.68	99.89	47.11	32.13
XLM-R	zh MT: XNLI	57.74	66.47	33.45	99.96	41.42	31.13
XLM-R	zh: OCNLI+XNLI	70.31	74.25	59.34	100.00	58.84	21.47
XLM-R	en: En-all-NLI	69.56	77.62	47.13	100.00	57.80	15.93
XLM-R	en: MNLI	66.74	73.12	48.97	100.00	53.89	18.09
XLM-R	mix: OCNLI+En-all-NLI	**78.82**	**81.57**	**71.15**	**100.00**	**70.63**	11.55
XLM-R	mix: XNLI+En-all-NLI	66.90	76.25	40.90	99.93	41.89	32.23
Human		98.00					

Table 5: Accuracy on Chinese HANS. Δ = the difference of accuracy between OCNLI dev and Non-Entailment.

those targeting the *lexical overlap* heuristics for the transformers models we tested (see the "sub-sequence" and "lexical overlap" columns in Table 5). This is in line with the results reported in the English HANS paper (specifically Table 15 in McCoy et al. (2019) which also shows that the sub-sequence examples are more difficult for the English BERT model). Second, for the sub-sequence heuristics, results from monolingual model are 12% higher than those from XLM-R under the zero-shot transfer setting (61.48% versus 48.79% in "sub-sequence" column in Table 5). This stands in contrast with the lexical overlap heuristic, where the best monolingual model performs similarly to the best zero-shot cross-lingual transfer (75.39% versus 77.62%). This is one of the few cases where cross-lingual transfer under-performs the monolingual setting by a large margin, suggesting that in certain situations monolingual models may be preferred.

5.3 Stress tests

Table 6 presents the accuracies on all the stress tests. We first see that cross-lingual zero-shot transfer using all English NLI data performs even better than the best monolingual model (~74% vs. ~71%). This demonstrates the power of the cross-lingual transfer-learning. Adding OCNLI to all English NLI gives another increase of about 3 percentage points (to 77%), while adding XNLI hurts the performance, again showing the importance of having expert-annotated language-specific data.

Antonyms and Synonyms All models except those fine-tuned on OCNLI achieved almost perfect score on the synonym test. However, for antonyms, both monolingual and multilingual models fine-tuned with OCNLI perform better than XNLI. XLM-R fine-tuned with English NLI data only again outperforms the best of monolingual models (~80% vs. ~72%). Interestingly, adding XNLI to all English NLI data hurts the accuracy badly (a 14% drop), while adding OCNLI to the same English data improves the result slightly.

As antonyms are harder to learn (Glockner et al., 2018), we take our results to mean that either expert-annotated data for Chinese or a huge English NLI dataset is needed for a model to learn decent representations about antonyms, as indicated by the high performance of RoBERTa fine-tuned with OCNLI (71.81%), and XLM-R fine-tuned with En-all-NLI (80.36%), on antonyms. That is, using machine-translated XNLI will not work well for learning antonyms (~55% accuracy).

Distraction Results in Table 6 show that adding distractions to the hypotheses has a more negative negative impact on models' performance, compared with appending distractions to the premises. The difference is about 20% for all models (see "Distr H" columns and "Distr P" columns in Table 6), which has not been reported in previous studies, to the best of our knowledge. Including a negation in the hypothesis makes it even more challenging, as we see another one percent drop in the accuracy for all models. This is expected as previous literature has demonstrated the key role negation plays when hypotheses are produced by the annotators (Poliak et al., 2018).

Spelling This is another case where cross-lingual transfer with English data alone falls behind monolingual Chinese models (by about 4%). Also the best results are from fine-tuning XLM-R with OCNLI + XNLI, rather than a combination of English and Chinese data. Considering the data is created by swapping Chinese characters with others of the same pronunciation, we take it to suggest

Model	Fine-tuned on	Overall	Ant.	Syn.	Distr H	Distr H-n	Distr P	Distr P-n	Spell-ing	num.
RoBERTa	zh MT: XNLI-small	59.41	43.38	99.64	51.61	51.41	70.66	71.19	69.93	28.70
RoBERTa	zh MT: XNLI	66.22	52.28	99.79	54.83	53.8	74.55	74.57	72.22	53.53
RoBERTa	zh ori: OCNLI	64.49	71.81	73.66	52.95	51.8	73.43	73.86	71.79	54.16
RoBERTa	zh: OCNLI + XNLI	71.01	59.39	99.06	55.87	54.64	76.83	76.50	75.48	70.18
XLM-R	zh ori: OCNLI	69.08	71.29	88.63	55.93	55.05	76.84	77.00	71.42	65.51
XLM-R	zh MT: XNLI	66.87	55.53	**99.96**	56.11	55.29	77.69	77.9	74.37	46.81
XLM-R	zh: OCNLI + XNLI	71.49	61.85	99.45	58.15	57.92	79.16	79.28	**77.93**	61.88
XLM-R	en:MNLI	67.94	65.77	99.2	55.14	54.6	75.75	75.76	70.76	50.90
XLM-R	en: En-all-NLI	74.52	80.36	97.58	54.74	53.56	73.96	73.92	71.02	82.73
XLM-R	mix: OCNLI + En-all-NLI	**77.36**	**81.93**	95.09	**59.23**	**58.00**	**79.88**	**79.92**	74.53	**87.77**
XLM-R	mix: XNLI + En-all-NLI	73.57	66.15	99.68	57.02	55.51	78.38	78.53	75.15	80.33
Human		85.00	85.00	98.00	83.00	83.00	83.00	83.00	78.00	98.00

Table 6: Accuracy on the stress test. Distr H/P(-n): distraction in Hypothesis/Premise (with negation).

Model	Fine-tuned on	Overall	$\ast$ Classifier	$\ast$ Idioms	$\ast$ Non-core arguement	$\ast$ Pro-drop	$\ast$ Time of event	Anaphora	Argument structure	Common sense	Comparatives	Double negation	Lexical semantics	Monotonicity	Negation	World knowledge
RoBERTa	zh MT: XNLI-small	62.9	65.8	64.7	55.2	80.5	60.0	59.6	67.4	54.3	61.4	48.3	60.9	59.7	66.2	39.0
RoBERTa	zh MT: XNLI	67.7	67.6	66.2	59.4	82.3	65.1	69.9	72.0	56.8	70.4	64.2	67.5	61.7	72.9	52.1
RoBERTa	zh ori: OCNLI	67.8	62.0	68.0	59.4	80.7	77.5	70.3	70.0	56.0	66.6	64.2	68.4	61.7	72.4	57.9
RoBERTa	zh: OCNLI + XNLI	69.3	66.3	67.1	58.6	83.0	74.0	70.1	73.5	54.9	74.1	67.5	69.1	62.5	76.0	60.0
XLM-R	zh ori: OCNLI	68.0	57.6	70.1	58.0	79.6	76.3	67.4	70.3	55.3	69.8	75.8	71.1	62.5	71.1	62.1
XLM-R	zh MT: XNLI	60.9	61.2	62.3	50.4	71.9	59.7	60.3	63.3	51.7	65.2	54.9	61.0	53.5	66.9	58.3
XLM-R	zh: OCNLI + XNLI	71.5	70.4	71.6	57.5	84.6	77.8	74.5	74.7	55.3	75.5	76.7	72.8	62.7	76.3	65.3
XLM-R	en: MNLI	70.2	70.1	73.9	57.5	86.4	70.8	69.3	72.9	48.9	76.0	62.5	67.8	62.6	77.0	62.1
XLM-R	en: En-all-NLI	71.9	71.8	**74.3**	56.2	87.4	75.7	74.9	74.8	49.1	80.5	70.8	69.1	63.8	77.8	64.2
XLM-R	mix: OCNLI + En-all-NLI	**74.9**	**72.7**	**74.3**	60.1	**88.5**	**84.5**	**77.3**	**78.1**	56.6	**81.3**	**79.2**	**77.2**	65.6	**78.0**	67.9
XLM-R	mix: XNLI + En-all-NLI	71.4	70.2	58.5	**85.5**	71.3	75.2	75.5	55.1	**79.2**	70.0	69.1	62.4	**76.2**	72.1	**71.3**

Table 7: Accuracy on the expanded diagnostics. Uniquely Chinese linguistic features are prefixed with $\ast$.

that monolingual models are still better at picking up the misspellings or learning the connections between characters at the phonological level.

Numerical Reasoning Results in the last column of Table 6 suggest a similar pattern: using all English NLI data for cross-lingual transfer outperforms the best monolingual model. However, fine-tuning a monolingual model with the small OCNLI (50k examples, accuracy: 54%) achieves better accuracy than using a much larger MNLI (390k examples, accuracy: 51%) for cross-lingual transfer, although both are worse than XLM-R fine-tuned with all English NLI which has more than 1,000k examples (accuracy: 83%). This suggests that there are cases where a monolingual setting (RoBERTa with OCNLI) is competitive against zero-shot transfer with a large English dataset (XLM-R with MNLI). However, that competitiveness may disappear when the English dataset grows to an order of magnitude larger in size or becomes more diverse (En-all-NLI contains 4 different English NLI datasets).

5.4 Hand-written diagnostics

Results on the expanded diagnostics are presented in Table 7. We first see that XLM-R fine-tuned with only English performs very well, at 70.2% and 71.9%, even slightly higher than the best monolingual Chinese model (69.3%).

Most surprisingly, **in 3/5 categories with uniquely Chinese linguistic features, zero-shot transfer outperforms monolingual models**. Only in "non-core arguments" and "time of event" do we see higher performance of OCNLI as the fine-tuning data. What is particularly striking is that for "idioms (*Chengyu*)", XLM-R fine-tuned only on English data achieves the best result, suggesting that the cross-lingual transfer is capable of learning meaning representation beyond the surface lexical information, at least for many of the idioms we tested. The overall results (accuracy of 74.3%) indicate that cross-lingual transfer is very successful in most cases. Manual inspection of the results shows that for many NLI pairs with idioms, XLM-R correctly predicts the figurative interpretation of the idiom as entailment, and the literal interpretation as non-entailment, as described in section 3.2. Looking at OCNLI and XNLI, we observe that they perform similarly when fine-tuned on monolingual RoBERTa. However, when fine-tuned with XLM-R, OCNLI has a clear advantage (68.0% versus 60.9%), suggesting that OCNLI may produce more

model	finetune on	overall	boolean	comparative	conditional	counting	negation	quantifier
RoBERTa	zh MT: XNLI-small	46.57	32.81	34.41	61.48	81.82	33.27	35.63
RoBERTa	zh MT: XNLI	50.64	33.35	39.02	66.55	84.51	40.92	39.50
RoBERTa	zh ori: OCNLI	47.53	35.81	34.81	62.87	69.64	49.84	32.24
RoBERTa	zh: OCNLI + XNLI	51.13	38.16	37.98	66.19	75.73	53.31	35.43
XLM-R	zh ori: OCNLI	54.33	**54.19**	**49.02**	52.46	79.70	59.52	31.08
XLM-R	zh MT: XNLI	50.79	33.39	35.33	66.01	87.23	33.17	**49.60**
XLM-R	zh: OCNLI + XNLI	52.43	34.51	36.93	59.98	88.70	54.37	40.08
XLM-R	en: MNLI	49.09	33.27	37.98	66.25	89.70	34.69	32.65
XLM-R	en: En-all-NLI	55.37	33.43	39.70	**66.65**	92.34	64.11	35.99
XLM-R	mix: OCNLI + En-all-NLI	**57.95**	40.70	44.49	63.67	91.54	**74.47**	32.81
XLM-R	mix: XNLI + En-all-NLI	57.73	40.30	37.82	66.67	**93.19**	61.52	46.87

Table 8: Accuracy on the Chinese semantic probing datasets, designed following Richardson et al. (2020).

stable results than XNLI. Furthermore, when coupled with English data to be used with XLM-R, we see again that OCNLI + En-all-NLI results in an accuracy 3 percent higher than XNLI + En-all-NLI.

5.5 Semantic fragments

Results on the semantic probing datasets (shown in Table 8) are more mixed. First, the results are in general much worse than the other evaluation data, but overall, XLM-R fine-tuned with OCNLI and all English data still performs the best. The overall lower performance is likely due to the longer length of premises and hypotheses in the semantic probing datasets, compared with the other three evaluation sets. Second, zero-shot transfer is better or on par with monolingual Chinese RoBERTa in 4/6 semantic fragments (except Boolean and quantifier). Third, for Boolean and comparative, XLM-R fine-tuned with OCNLI has a much better result than all other monolingual models or XLM-R fine-tuned with mixed data. We also observe that all models have highest performance on the "counting" fragment. Note that none of the models have seen any data from the "counting" fragment during fine-tuning. That is, all the knowledge come from the pre-training and fine-tuning on general NLI datasets. The surprisingly good performance of XLM-R model (w/ En-all-NLI, 92.34%) suggests that it may have already acquired a mapping from counting the words/names to numbers, and this knowledge can be transferred cross-linguistically.

6 Conclusion and Future Work

In this paper, we examine the cross-lingual transfer ability of XLM-R in the context of Chinese NLI through four new sets of aversarial/probing tasks and a total of 17 new high quality and linguistically motivated challenge datasets. We find that multilingual transfer via fine-tuning solely on benchmark English data generally yields impressive performance. In 3/4 on our task categories, such *zero-shot transfer* outperforms our best monolingual models trained on benchmark Chinese NLI data, including 3/5 of our hand-crafted challenge tasks that test uniquely Chinese linguistic phenomena. These results suggest that multilingual models are indeed capable of considerable cross-lingual linguistic transfer and that zero-shot NLI may serve as a serious alternative to large-scale dataset development for new languages.

These results come with several important caveats. Model performance is still outperformed by conservative estimates of human performance and our best models still have considerable room for improvement; we hope that our new resources will be useful for continuing to benchmark progress on Chinese NLI. We also find that high-quality Chinese NLI data (e.g., OCNLI) *can* help improve results further, which suggests an important role for certain types of expertly annotated monolingual data in a training pipeline. In virtue of our study being limited to *behavioral testing*, the exact reason for *why* cross-lingual zero-shot transfer generally performs well, especially on some Chinese-specific phenomena, is an open question that requires further investigation. In particular, we believe that techniques that couple behavioral testing with *intervention* techniques (Geiger et al., 2020; Vig et al., 2020) and other analysis methods (Giulianelli et al., 2018; Belinkov and Glass, 2019) might provide insight, and that our new Chinese resources can play an important role in such future work.

Acknowledgments

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. He Zhou is sponsored by China Scholarship Council.

References

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*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of ACL*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

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*.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. Data and representation for turkish natural language inference. In *Proceedings of EMNLP*.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2020. N-LTP: A open-source neural Chinese language technology platform with pretrained models. *arXiv preprint arXiv:2009.11616*.

Hyunjin Choi, Judong Kim, Seongho Joe, Seungjai Min, and Youngjune Gwon. 2021. Analyzing zero-shot cross-lingual transfer in supervised nlp tasks. In *2020 25th International Conference on Pattern Recognition (ICPR)*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

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*.

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*.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

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*.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of BlackBoxNLP*.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of BlackboxNLP*, pages 240–248.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of ACL*.

Emily Goodwin, Koustuv Sinha, and Timothy O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of ACL*, pages 1958–1969.

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, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of EMNLP*.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. In *Proceedings of ICLR*.

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

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

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing. *Advances in Neural Information Processing Systems*, 33.

Charles N Li, Sandra A Thompson, and Sandra A Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Univ of California Press.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*.

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*.

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*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

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*.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific BERT models. *arXiv preprint arXiv:2003.02912*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*.

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

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*.

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

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. Conjnli: Natural language inference over conjunctive sentences. In *Proceedings of EMNLP*.

Felipe Salvatore, Marcelo Finger, and Roberto Hirata Jr. 2019. A logical-based corpus for cross-lingual evaluation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 22–30.

Tianqi Sun. 2009. *A Study on the License Pattern and Mechanism of Non-core Arguments in Mandarin Chinese*. Ph.D. thesis, Peking University.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of NAACL*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

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*.

Gijs Wijnholds and Michael Moortgat. 2021. SICKNL: A Dataset for Dutch Natural Language Inference. In *Proceedings of EACL*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP*.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP*.

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*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Endong Xun, Gaoqi Rao, Xiaoyue Xiao, and Jiaojiao Zang. 2016. The construction of the BCC corpus in the age of big data. *Corpur Linguistics (in Chinese)*, pages 93–109.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Help: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of *SEM*, pages 250–255.

Wei Zhao, Mingyue Shang, Yang Liu, Liang Wang, and Jingming Liu. 2020. Ape210k: A large-scale and template-rich dataset of math word problems. *arXiv preprint arXiv:2009.11506*.

Heuristic	entailment	contradiction	neutral
lexical overlap	944	155	109
subsequence	190	10	18

Table 9: Distribution of the two heuristics in OCNLI.

Heuristic	entailment	contradiction	neutral
lexical overlap	441	647	340
subsequence	100	193	220
Total	541	840	560

Table 10: Distributional statistics of our synthesized Chinese HANS dataset.

A Details for dataset creation

In this section, we list example NLI pairs and their translations. For examples of the Chinese HANS and stress tests, see Table 2. For the expanded diagnostics, see Table 11. For the semantic/logic probing dataset, see Table 3.

Chinese HANS Table 9 lists the number of examples in OCNLI for each inference label that satisfy the two heuristics we are examining. We observe that *entailment* examples take the majority for both heuristics. Therefore, we hypothesize that if the heuristics are learned, the *entailment* examples are likely to be correctly predicted while *non-entailment* (contradiction and neutral) examples are prone to receive wrong prediction.

To guarantee the generated sentences are syntactically and semantically sound, we add features for our vocabulary so that subject- predicate and verb-object constraints are satisfied, e.g., some verbs can only take animate subjects and objects. We then generate 50 premise-hypothesis pairs for each template described in our Github repository.[12] Excluding duplicated examples, our generated dataset has 1,941 pairs and the distribution of the three labels is shown in Table 10.

Antonym After looking at the quality of initially generated data, we decided to replace only the nouns and adjectives with their antonyms since such replacements are most likely result in grammatical and contradictory hypotheses.[13]

Synonym After inspecting the initially generated data, we decided to perform replacements only to verbs and adjectives. To ensure the quality of synonyms, we rank the synonyms from a commonly used synonym dictionary by their vector similarity to the original word, and pick the top ranking synonym.[14]

Distraction We created the distraction data similar to the stress test setting (Naik et al., 2018) but experimented with variations as to where "distractor statement"—either a tautology or a true statement—was added: the premise or the hypothesis. The distractor statement also varied w.r.t. whether it contains a negation:

- **Premise-no-negation**: A distractor statement is added to the end of the premise and it contains no negation.
- **Premise-negation**: A distractor statement containing a negation is added to the premise.
- **Hypothesis-no-negation**: A distractor statement is added to the end of the hypothesis.
- **Hypothesis-negation**: Same as the previous condition except that the distractor contains a negation.

Only two tautologies are used in Naik et al. (2018). In this paper, to thoroughly examine the influence of different true statements, we designed 50 tautology/statements varied in three factors: length, out-of-vocabulary, and negation word. There are 25 statements pairs in total (1 tautology and 24 true statements); each pair includes a true statement and its corresponding true statement with negation form. All the statements range from 5 to 16 characters. For the true statements in negation form, two common Chinese negation words 不 and 没 are used for negation. For the 24 true statement pairs, half of them contains at least one Out-of-Vocabulary word in OCNLI.

Experiments show that length, Out-of-Vocabulary words, and the choice of negator have little effects on the results.

Spelling We generate a set of data containing "spelling errors" by replacing one random character in the hypotheses with its homonym, which is defined as a character with the same *pinyin* pronunciation ignoring tones. We also limit the frequency

[12] https://github.com/huhailinguist/ChineseVariousNLI

[13] We use the LTP toolkit (https://github.com/HIT-SCIR/ltp) to annotate the POS tags and our antonym list is from https://github.com/liuhuanyong/ChineseAntiword.

[14] We use the synonym list from https://github.com/Keson96/SynoCN and the similarity score from the Python package Synonyms at https://github.com/chatopera/Synonyms.

of the homonym as within the range of 100 to 6000 so that the character is neither too rare nor too frequent.

Numerical reasoning We extracted sentences from Ape210k (Zhao et al., 2020), a large-scale math word problem dataset containing 210K Chinese elementary school-level problems[15]. We generate entailed, contradictory and neutral hypotheses for each premise, with the rules below:

1. **Entailment:** Randomly choose a number x and change it to y from the hypothesis. If the y > x, prefix it with one phrase that translate to "less than"; if y < x, prefix it with one phrase that translate to "more than".
   Premise: *Mary types 110 words per minute.* Hypothesis: *Mary types less than 510 words per minute.*
2. **Contradiction:** Perform either 1) randomly choose a number x from the hypothesis and change it; 2) randomly choose a number from the hypothesis and prefix it with one phrase that translate to "less than" or "more than".
   Premise: *Mary types 110 words per minute.* Hypothesis: *Mary types 710 words per minute.*
3. **Neutral:** Reverse the corresponding entailed premise-hypothesis pairs.
   Premise: *Mary types less than 510 words per minute.* Hypothesis: *Mary types 110 words per minute.*

The result contains 2,871 unique premise sentences and 8,613 NLI pairs.

Diagnostics The diagnstics for *classifiers* (or measure word) and *non-core arguments* are explained in detail below (see examples in Table 11).

1. *classifiers* (or measure word): in Chinese, when modified by a numeral, a noun must be preceeded by a category of words called classifier. They can be semantically vacuous but sometimes also carry semantic content: 一<u>匹</u>狼 *one <u>pi</u> wolf* (one wolf); 一<u>群</u>狼 *one <u>qun</u> wolf* (one <u>pack</u> of wolves). Our examples require the model to understand the semantic content of the classifiers.
2. *non-core arguments*: in Chinese syntax, sometimes a noun phrase at the argument position (e.g., object) is not serving as an object: 今天吃筷子，不吃叉子。*today eat chopsticks, not eat fork* (We eat **with** chopsticks today, not with fork). Sun (2009) shows that this structure is very productive in Chinese and we take example sentences from her dissertation.

Additionally, for the *pro-drop* examples, they are constructed such that the models return the correct inference relation only when they successfully identify what the dropped *pro* refers to. That is, our constructed premises involve several entities the dropped *pro* could potentially refer to, and the entailed hypothesis identifies the correct referent while the neutral/contradictory hypothesis does not (see Table 11 for an example).

B Hyperparameters for experiments

Table 12 presents the hyperparameters used for the models. The learning-rate search space for RoBERTa is: 1e-5, 2e-5, 3e-5, 4e-5 and 5e-5, for XLM-R: 5e-6, 7e-6, 9e-6, 2e-5 and 5e-5.

C Chinese-to-English transfer

We present Chinese-to-English transfer results in this section. As mentioned in the main text, for most of the cases, zero-shot transfer learning does not work well mostly likely due to the small size of OCNLI. However, for 3 out of the 4 datasets, XLM-R fine-tuned with the mix data outperforms the monolingual setting, suggesting that even OCNLI is only 1/20 of En-all-NLI, XLM-R can still acquire some useful information from OCNLI, in addition to what is present in En-all-NLI.

Specifically, (1) for English HANS, XLM-R fine-tuned with OCNLI is about 13 percentage points below the best English monolingual model, shown in Table 13. (2) For stress tests shown in Table 14, the gap is about 5 percent (XLM-R with OCNLI = 74%; RoBERTa with En-all-NLI = 79%). Interestingly, XLM-R with OCNLI performs the best for Negation and Word overlap. It even outperforms RoBERTa w/ MNLI on the Antonym, which seems to be consistent with the high performance of OCNLI-trained models on the Chinese Antonym in our constructed stress tests. (3) For semantic probing data, as shown in Table 15, XLM-R with OCNLI is 5 percent behind monolingual model fine-tuned with all English NLI, but performs better than the monolingual RoBERTa fine-tuned with

[15] We split all problems into individual sentences and filter out sentences without numbers. Then we remove sentences without any named entities ("PERSON", "LOCATION" and "ORGANIZATION") using the NER tool provided by LTP toolkit (Che et al., 2020).

category	n	premise	hypothesis	label
CLUE (Xu et al., 2020)	514	有些学生喜欢在公共澡堂里唱歌。 Some students like to sing in public showers.	有些女生喜欢在公共澡堂里唱歌。 Some female students like to sing in public showers.	N
CLUE expansion (ours)	800	雷克雅未克所有旅馆的床位加在一起才一千六百个。 There are only one thousand six hundred beds in all hotels in Reykjavik.	雷克雅未克有旅馆的床位超过一千个。 Some hotel in Reykjavik has over a thousand beds.	N
World Knowledge (ours)	37	上海在北京的南边。 Shanghai is to the south of Beijing.	北京在上海的南边。 Beijing is to the south of Shanghai.	C
Classifier (ours)	138	这些孩子吃了一个苹果。 These children ate an apple.	这些孩子吃了一筐苹果。 These children ate a basket of apples.	N
Chengyu/idioms (ours)	250	这帮人可狡猾得很啊，你一个电话打过去，打草惊蛇，后果不堪设想。 These people are so cunning! If you call them, it would alert them, and as we say in a Chinese idiom ”if you hit the grass, it would alert the snakes.” The consequences would be unimaginable.	你打电话过去会让这帮人察觉，造成不好的结果。 If you call them, it will alert them, and bring negative consequences.	E
		same as above	这些狡猾的人养了很多蛇。 These cunning people have raised a lot of snakes.	N
Pro-drop (ours)	197	见了很多学生，又给老师们开了两个小时会，校长和主任终于可以下班了。 After (*pro*) meeting many students and (*pro*) having two hours of meeting with the teachers, the principal and the director can finally get off work.	校长见了很多学生。 The principal met many students.	E
		same as above	老师们见了很多学生。 The teachers met many students.	N
Non-core arguments (ours)	185	平时范志毅都踢后卫的，今天却改当前锋了。 Zhiyi Fan usually plays full back in soccer, but today he switched to playing forward.	范志毅经常用腿踢对方的后卫。 Zhiyi Fan usually uses his legs to kick the other team's full back.	N

Table 11: Example NLI pairs in expanded diagnostics with translations.

Model	Training Data	LR
RoBERTa	zh MT: XNLI-small	3e-05
RoBERTa	zh MT: XNLI	2e-05
RoBERTa	zh ori: OCNLI	2e-05
RoBERTa	zh: OCNLI + XNLI	3e-05
XLM-R	zh ori: OCNLI	5e-06
XLM-R	zh MT: XNLI	7e-06
XLM-R	zh: OCNLI + XNLI	7e-06
XLM-R	en:MNLI	5e-06
XLM-R	en: En-all-NLI	7e-06
XLM-R	mix: OCNLI + En-all-NLI	7e-06
XLM-R	mix: XNLI + En-all-NLI	7e-06

Table 12: Hyper-parameters used for fine-tuning the models. All models are fine-tuned for 3 epochs with maximum length of 128.

MNLI (53.6% vs. 51.3%). This is quite surprising since the size of OCNLI is only 1/8 of MNLI. (4) For the English diagnostics as shown in Table 16 and Table 17, XLM-R with OCNLI is 7 percent behind RoBERTa fine-tuned with MNLI.

We leave it for future work to thoroughly examine transfer learning from a "low-resource" language such as Chinese to the high-resource one such as English.

Model	Fine-tuned on	Overall	Lexical_overlap	Subsequence	Constituent	Entailment	Non-entailment
RoBERTa	en: En-all-NLI	76.54	96.79	67.77	65.06	99.81	53.27
RoBERTa	en: MNLI	77.63	95.60	**68.08**	69.21	99.74	55.52
XLM-R	en: En-all-NLI	75.72	95.52	62.99	68.63	99.91	51.52
XLM-R	en: MNLI	74.80	92.92	65.24	66.23	98.83	50.76
XLM-R	zh ori: OCNLI	64.37	71.28	54.42	67.41	98.39	30.35
XLM-R	zh MT: XNLI	68.83	81.67	62.07	62.74	99.13	38.53
XLM-R	zh mix: OCNLI+XNLI	71.30	82.52	61.72	69.66	99.08	43.52
XLM-R	mix: OCNLI+En-all-NLI	**78.56**	**96.92**	64.91	**73.84**	**99.92**	**57.20**
XLM-R	mix: XNLI+En-all-NLI	74.65	93.93	60.97	69.04	99.96	49.34

Table 13: Results of English HANS (McCoy et al., 2019).

Model	Fine-tuned on	Overall	Antonym	Content word swap	Function word swap	Keyboard	Swap	Length mismatch	Negation	Numerical reasoning	Word overlap
RoBERTa	en: En-all-NLI	79.48	82.91	**86.22**	88.71	**87.8**	**87.48**	**88.28**	60.25	79.26	62.85
RoBERTa	en: MNLI	77.9	69.03	85.74	**88.75**	87.39	87.05	88.23	59.19	65.46	61.48
XLM-R	en: En-all-NLI	79.6	86.25	85.26	87.38	86.31	86.72	87.25	61.06	**82.84**	65.79
XLM-R	en: MNLI	77.6	74.65	85.09	87.33	86.08	86.42	86.96	60.95	54.66	65.13
XLM-R	zh ori: OCNLI	74.31	72.52	75.12	77.71	76.27	76.39	77.23	**72.86**	55.85	**72.79**
XLM-R	zh MT: XNLI	77.78	65.12	85.11	86.64	85.79	85.71	85.91	63.52	43.95	71.63
XLM-R	zh mix: OCNLI+XNLI	77.83	66.83	84.96	86.69	85.81	85.87	85.98	63.97	51.56	68.38
XLM-R	mix: OCNLI+En-all-NLI	**80.01**	**86.33**	85.22	87.40	86.26	86.77	87.23	62.52	81.79	67.54
XLM-R	mix: XNLI+En-all-NLI	79.38	85.27	85.35	87.20	86.28	86.74	87.22	60.29	80.50	66.19

Table 14: Results of English stress test (Naik et al., 2018).

Model	Fine-tuned on	Overall	Boolean	Comparative	Conditional	Counting	Monotonicity hard	Monotonicity simple	Negation	Quantifier
RoBERTa	en: MNLI	51.31	43.58	39.60	66.24	63.34	61.28	60.10	37.26	39.08
RoBERTa	en: En-all-NLI	58.72	60.18	40.28	66.30	66.22	59.60	58.98	64.46	**53.74**
XLM-R	en: MNLI	53.54	59.16	41.62	66.30	61.72	63.26	**62.82**	33.52	39.92
XLM-R	en: En-all-NLI	59.85	71.58	45.18	66.30	60.40	63.86	62.02	65.68	43.78
XLM-R	zh ori: OCNLI	53.61	66.02	**60.62**	41.10	58.00	47.86	49.88	51.88	53.50
XLM-R	zh MT: XNLI	52.29	43.24	39.00	66.22	65.66	58.08	62.74	34.12	49.24
XLM-R	zh mix: OCNLI+XNLI	54.68	54.64	38.84	66.28	**67.38**	58.18	61.38	41.88	48.82
XLM-R	mix: OCNLI+En All NLI	**60.20**	**71.20**	42.58	**66.30**	62.40	**64.72**	60.90	**68.58**	44.88
XLM-R	mix: XNLI+En-all-NLI	60.06	65.8	46.86	66.30	65.54	61.56	61.44	68.50	44.48

Table 15: Results of English semantic probing datasets (Richardson et al., 2020).

Model	Fine-tuned on	Overall	active passive	anaphora coreference	common sense	conditionals	anaphora coreference	coordination scope	core args	datives	disjunction	double negation	downward mono-tone	ellipsis implicits	existential	factivity	genitives partitives	intersectivity
RoBERTa	en: MNLI	66.87	**62.35**	**67.59**	**69.47**	62.50	78.00	63.50	69.62	85.00	39.47	**92.86**	**19.33**	65.29	65.00	62.06	**95.00**	60.43
RoBERTa	en: En-all-NLI	**68.03**	61.76	70.00	69.33	**63.75**	**82.50**	**68.00**	**75.77**	**85.00**	**41.58**	92.14	18.67	**67.65**	65.00	**62.35**	94.00	59.57
XLM-R	en: MNLI	63.03	61.76	62.76	59.73	55.62	76.00	61.50	61.54	85.00	26.84	91.43	16.00	64.12	69.00	51.47	90.00	60.00
XLM-R	en: En-all-NLI	64.57	61.76	65.17	61.47	60.00	76.00	66.00	65.77	85.00	33.16	89.29	14.00	62.94	**71.00**	58.53	90.00	**60.87**
XLM-R	zh ori: OCNLI	59.67	60.00	59.31	57.20	58.12	70.00	56.50	61.54	85.00	30.00	67.14	17.33	54.71	66.00	46.18	90.00	59.57
XLM-R	zh MT: XNLI	61.76	61.18	64.14	60.67	58.75	72.50	60.00	60.77	85.00	33.16	91.43	12.67	58.24	64.00	48.24	90.00	57.83
XLM-R	zh mix: OCNLI+XNLI	61.78	61.76	62.76	56.93	57.50	74.50	61.00	61.54	85.00	31.05	90.00	12.00	57.65	65.00	48.53	90.00	57.39
XLM-R	mix: OCNLI+En-all-NLI	64.51	61.76	63.45	61.60	58.75	76.00	66.00	67.31	85.00	35.26	90.71	15.33	60.59	68.00	60.59	91.00	60.87
XLM-R	mix: XNLI+En All LI	64.37	61.18	64.83	62.27	61.88	73.00	65.00	65.38	85.00	35.26	91.43	14.67	63.53	70.00	57.94	94.00	60.87

Table 16: Results of English Diagnostics from GLUE-Part I (Wang et al., 2018).

Model	Fine-tuned on	intervals numbers	lexical entailment	morphological negation	named entities	negation	nominalization	non-monotone	prepositional phrases	quantifiers	redundancy	relative clauses	restrictivity	symmetry collectivity	temporal	universal	upward monotone	world knowledge
RoBERTa	en: MNLI	54.74	66.71	89.23	57.22	66.59	82.14	56.00	86.18	78.46	79.23	63.75	55.38	67.86	56.25	**84.44**	76.47	48.51
RoBERTa	en: En-all-NLI	**63.16**	**71.57**	**89.23**	61.11	**69.02**	84.29	**57.33**	84.41	74.62	73.85	63.75	53.08	**70.00**	**69.38**	83.33	73.53	**49.55**
XLM-R	en: MNLI	45.79	65.57	84.62	61.11	61.95	82.14	52.00	85.88	**82.69**	78.46	62.50	52.31	57.14	51.25	80.00	74.12	44.03
XLM-R	en: En-all-NLI	45.79	69.71	84.62	61.67	64.15	**85.71**	48.67	84.71	79.23	69.23	63.12	46.15	59.29	60.00	77.78	75.29	45.82
XLM-R	zh ori: OCNLI	39.47	56.29	75.38	41.11	53.41	73.57	51.33	85.59	63.08	83.08	62.50	**70.77**	64.29	54.38	62.22	**78.82**	47.31
XLM-R	zh MT: XNLI	42.11	60.14	84.62	61.11	61.95	74.29	53.33	**86.76**	73.46	84.62	60.62	60.00	57.86	40.62	**84.44**	68.24	47.31
XLM-R	zh mix: OCNLI+XNLI	43.68	59.29	83.08	**63.33**	62.20	74.29	52.00	**86.76**	76.92	**85.38**	62.50	60.77	59.29	43.75	82.22	67.65	47.61
XLM-R	mix: OCNLI+En-all-NLI	45.26	69.86	85.38	62.22	65.12	85.71	50.67	85.00	74.62	69.23	**68.12**	47.69	60.71	56.25	77.78	75.29	45.67
XLM-R	mix: XNLI+En All LI	44.21	67.29	86.15	62.22	63.90	83.57	49.33	84.71	75.00	70.77	66.88	46.92	58.57	57.50	74.44	74.12	45.82

Table 17: Results of English Diagnostics from GLUE-Part II (Wang et al., 2018).