The Chinese Causative-Passive Homonymy Disambiguation:
an Adversarial Dataset for NLI and a Probing Task

Shanshan Xu
L3S Research Center, Hannover, Germany/
Department of Informatics, Technical University of Munich, Germany
shanshan.xu@tum.de

Katja Markert
Institute of Computational Linguistics
Heidelberg University, Germany
markert@cl.uni-heidelberg.de
Motivation

Pretrained language models (PLMs) achieve fantastic performance in NLU tasks.
Do language models understand language?
Natural Language Inference (NLI) Task

Premise: Today is Friday
Hypothesis: Tomorrow is Saturday

Entailment

Premise: Today is Friday
Hypothesis: It is April

Non-entailment
NLI Dataset Examples

**SNLI** (Bowman et al. 2015)

Premise:
A soccer game with multiple males playing.

Hypothesis:
Some men are playing a sport.

**Entailment**

**OCNLI** (Hu et al. 2020)

Premise:

嗯，今天星期六我们这儿，嗯哼。
En, it’s Saturday today in our place, yeah.

Hypothesis:

昨天是星期天
It was Sunday yesterday.

**Contradiction/ Non-entailment**

A large annotated corpus for learning natural language inference (Bowman et al., EMNLP 2015)
OCNLI: Original Chinese Natural Language Inference (Hu et al. 2020, Findings of EMNLP 2020)
Adversarial NLI Datasets

• PLMs often exploit superficial patterns, and fail on examples with high lexical overlap
• **HANS**: carefully designed adversarial NLI dataset

(McCoy et al. 2019)

The doctor near the actor danced. → The actor danced.

Non-entailment

• However:
  • templates only for English
  • difficult to find the templates from scratch

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference (McCoy et al., ACL 2019)
Contributions

• We create the first Chinese adversarial NLI test set CANLI.
• Using the linguistic phenomenon Causative-Passive Homonymy (CPH).
• SOTA NLI system (RoBERTa finetuned on OCNLI) performs poorly on CANLI.
• We use word sense disambiguation as a probing task.
• The probe results demonstrate that RoBERTa’s performance on CANLI does not correspond to its internal representation of CPH.
The Causative-Passive Homonymy (CPH)

Canonical
1a. She gets them to do the cleaning. (causative: get + infinitive)
1b. Her wallet was stolen. (passive: be + past participle).

Causative - Passive Homonymy (CPH)
2a. She got them arrested (by the police). (Causative, get + past participle)
2b. She got her wallet stolen (by someone). (Passive, get + past participle)

- CPH also be observed in Korean, Chinese, Japanese, Manchu Tungusic languages, and others.
- There are no differences in the verbal constructions of CPH; it is the context that determines whether the verb should be read as causative or passive.
CPH in Chinese

(4)  a. 经济危机 让 公司 倒闭 了
    jingji-weiji rang gongsi dao bi le
    Economic-crisis CAUS company close-down PFV
    'The economic crisis caused the company to close down.'

b. 他 让 公司 开除 了
    ta rang gongsi kaichu le
    he PASS company fire PFV
    'He was fired by the company.'
Construction of CANLI

<table>
<thead>
<tr>
<th></th>
<th>Causative</th>
<th></th>
<th>Passive</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entailment</td>
<td>Non-entailment</td>
<td>Entailment</td>
<td>Non-entailment</td>
</tr>
<tr>
<td>Train</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Test</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Total</td>
<td>800</td>
<td>800</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Premises collection:
- CPH sentences marked by the CPH morpheme rang.
- Drawn from the genre of modern literature in the CCL online corpus
- Collected and annotated by a Chinese native speaker with a linguistics background

Hypothesis generation:
- Generated with templates.
- Proofread and edited by a native publishing house editor.
- The first author of this paper has double-checked the data after the editing process.
CANLI Template I

Template:

Premise: N1 rang N2 VP (passive)

Hypothesis 1: N1 VP N2

Hypothesis 2: N2 VP N1

Example:

Premise: Wo rang Baoqing chao xing le
"I was woken up by Baoqing"

Hypothesis 1: Wo chao xing le Baoqing.
“I woke Baoqing up.”

Hypothesis 1: Baoqing chao xing le Wo.
“Baoqing woke me up.”
CANLI Template II

Template:

Premise: N1 \textit{rang} N2 VP (causative)

Hypothesis 1: N1 VP

Hypothesis 2: N2 VP

Example:

Premise: Jing ji wei ji \textit{rang} gong si dao bi le.
"The economic crisis caused the company to close down."

Hypothesis 1: Jing ji wei ji dao bi le.
"The economic crisis closed down."

Hypothesis 1: Gong si dao bi le.
"The company closed down."
Experiments

• hfl/chinese-roberta-wwm-ext- large  (Cui et al. 2020)

• sequence classification head on top from the transformers library  (Wolf et al. 2020)

• Fine-tuned on OCNLI training set  (Xu et al. 2020)
Results

<table>
<thead>
<tr>
<th>Test data</th>
<th>OCNLI.val</th>
<th>CANLI.test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>accuracy</td>
<td>P</td>
</tr>
<tr>
<td>Fine-tuning data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OCNLI.train</td>
<td>87.4 (0.3)</td>
<td>81.5 (0.8)</td>
</tr>
<tr>
<td>OCNLI.train + CANLI.train</td>
<td>87.2 (0.2)</td>
<td>81.4 (0.4)</td>
</tr>
<tr>
<td>Human Performance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The OCNLI-fine-tuned model performed poorly on the CANLI.test.
- Fine-tuning with CANLI.train set indeed helps substantially when testing on CANLI.test.
- Has the model learned the linguistic feature of CPH after augmenting?
- To what extent can we find the CPH feature in the model’s internal representation?
The Representation of CPH

• There are no differences in the verbal constructions of CPH; it is the context that determines whether the verb should be read as causative or passive.
• Embeddings provided by Transformers depend on context.
• Hypothesis: The model has learned the CPH after fine-tuning with CANLI. -> CPH feature is captured in the model’s representation
Visualization with UMAP

- **green**: embeddings of *rang* from 200 passive sentences in the CANLI train set
- **purple**: embeddings of *rang* from 200 causative sentences in the CANLI train set
- **blue**: embeddings of *bei* (canonical passive marker) from 40 sentence drawn from CCL corpus.
- **red**: embeddings of *shi* (canonical causative marker) 40 sentence drawn from CCL corpus.

Embeddings pulled from the vanilla RoBERTA

Embeddings pulled from RoBERTA fine-tuned with OCNLI + CANLI
Quantitative Analysis

Causative/Passive disambiguation as a probing task

• Inspired by the method of Word Sense Disambiguation (WSD). (Coenen et al, 2019)
• Nearest-centroid classifier as probe.
• As the probe is not trained, selectivity is assured.
• Gold causative centroid: the centroid of 40 contextualized embeddings of *shi* (canonical causative marker)
• Gold passive centroid: the centroid of 40 contextualized embeddings of *bei* (canonical passive marker)
Probe Accuracies
Conclusions

• We present CANLI, the first adversarial NLI dataset for Chinese.
• The poor performance using RoBERTA fine-tuned on OCNLI demonstrates that CANLI is challenging for a state-of-the-art NLI system.
• WSD as probing task
• RoBERTa’s performance on CANLI does not correspond to its internal representation of CPH
• CANLI available @ https://huggingface.co/datasets/sxu/CANLI