LREC 2022 Marseille

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

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Motivation

Leaderboard Version: **2.0**

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
+	1	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	2	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
	3	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
	4	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
+	5	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	6	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	7	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7

Screenshoot from super.gluebenchmark.com (retrieved on 06.05.2022)

- 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)

). $\boxed{\text{Non-entailment}}$

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

- 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	daobi	le
ecnomic-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

	Causative		Passive		
	Entailment	Non-entailment	Entailment	Non-entailment	Total
Train	200	200	200	200	800
Test	200	200	200	200	800
Total	800		800		

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

Non-entailment

Hypothesis 2: N2 VP N1

Entailment

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.”

Non-entailment

Hypothesis 1: Baoqing chao xing le Wo.

“Baoqing woke me up.”

Entailment

CANLI Template II

Template:

Premise: N1 *rang* N2 VP (causative)

Hypothesis 1: N1 VP

Non-entailment

Hypothesis 2: N2 VP

Entailment

Example:

Premise: Jing ji wei ji *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.”

Non-entailment

Hypothesis 1: Gong si dao bi le.

“The company closed down.”

Entailment

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)

Revisiting Pre-Trained Models for Chinese Natural Language Processing (Cui et al., Findings of EMNLP 2020)
Transformers: State-of-the-Art Natural Language Processing (Wolf et al., EMNLP 2020)
OCNLI: Original Chinese Natural Language Inference (Hu et al. 2020., Findings of EMNLP 2020)

Results

Test data	OCNLI.val				CANLI.test			
Fine-tuning data	accuracy	P	R	F1	accuracy	P	R	F1
OCNLI.train	87.4 (0.3)	81.5 (0.8)	78.8 (1.0)	80.1 (0.5)	48.1 (1.3)	48.9 (0.8)	88.2 (2.6)	62.9 (1.2)
OCNLI.train + CANLI.train	87.2 (0.2)	81.4 (0.4)	77.7 (0.2)	79.5 (0.3)	97.3 (0.6)	97.4 (0.6)	97.3 (0.9)	97.3 (0.7)
Human Performance					93.2 (2.4)	93.5 (7.2)	93.1 (5.10)	93.0 (2.5)

- 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

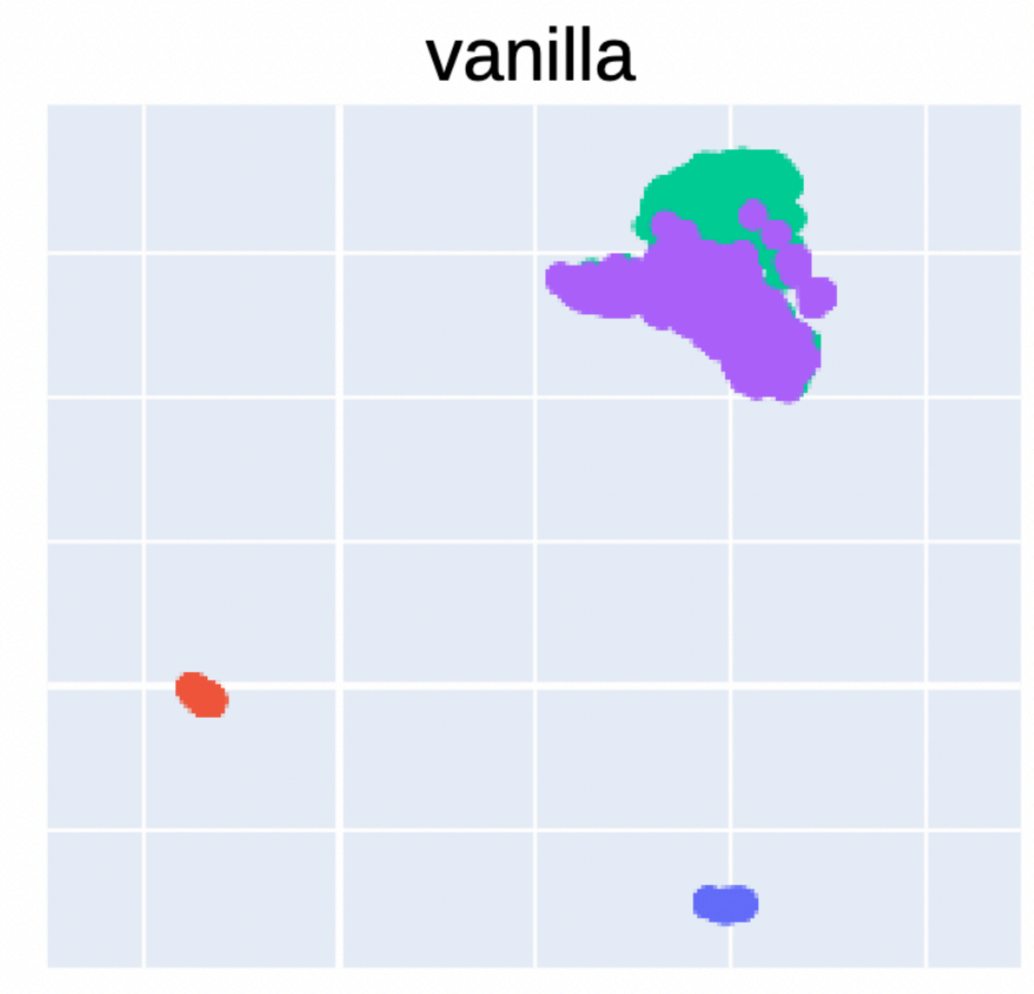
Visualization with UMAP

green: embeddings of *rang* from 200 passive sentences in the CANLI train set

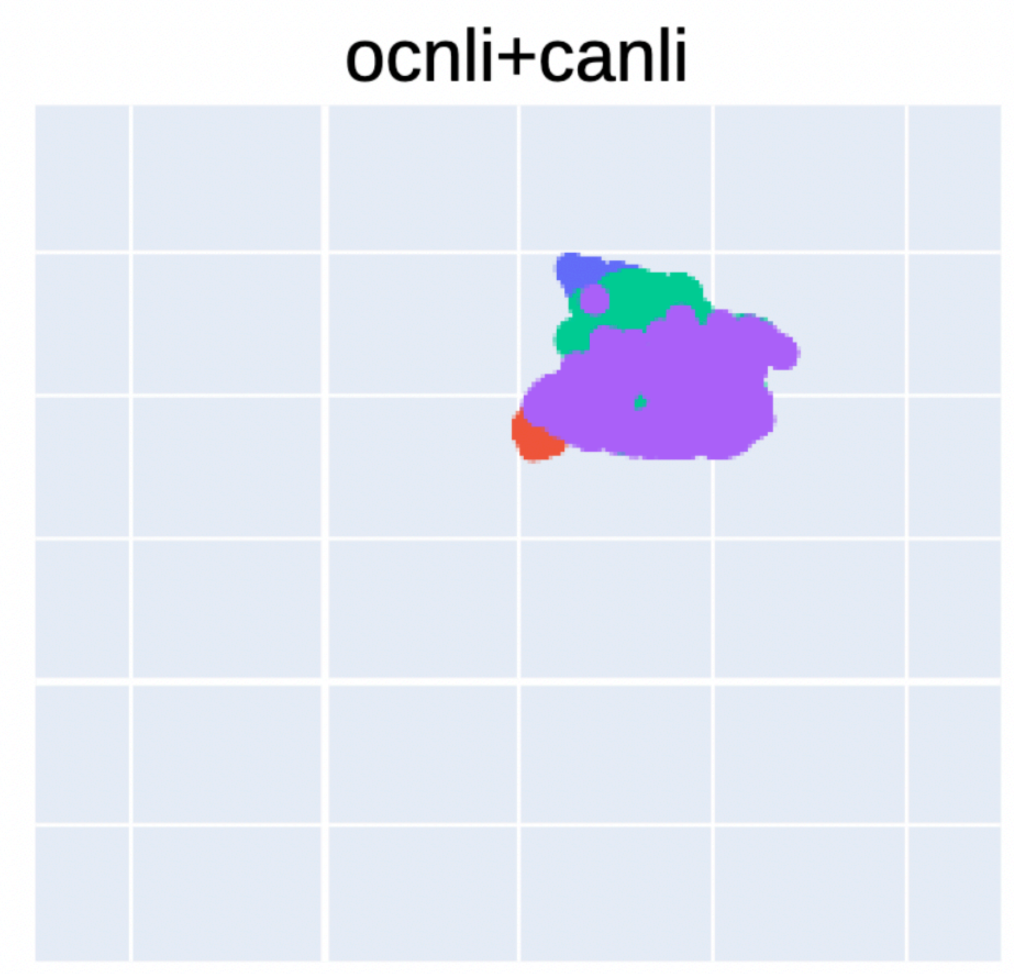
purple: embeddings of *rang* from 200 causative sentences in the CANLI train set

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

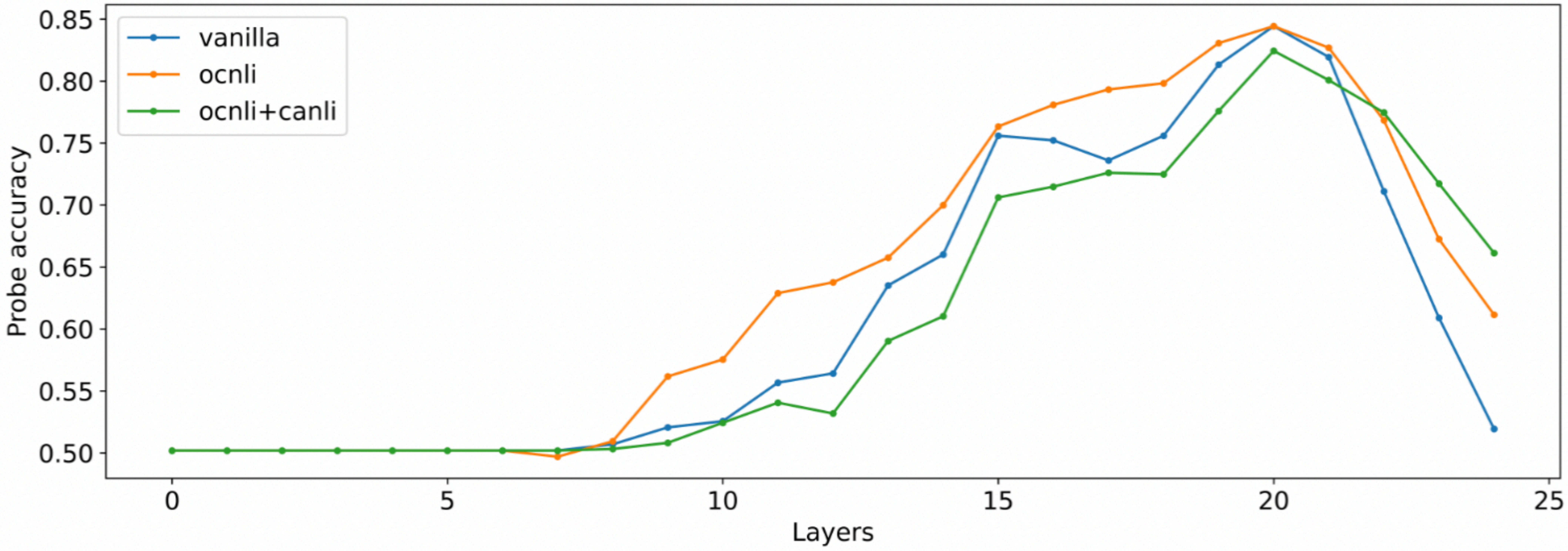
Quantitative Analysis

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)

Visualizing and Measuring the Geometry of BERT (Coenen et al., NeurIPS 2019)

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