Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4316–4323
Marseille, 20-25 June 2022

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

Shanshan Xu[1,2], Katja Markert[3]
[1]L3S Research Center, Germany
[2]Department of Informatics, Technical University of Munich, Germany
[3]Institute of Computational Linguistics, Heidelberg University, Germany
shanshan.xu@tum.de, markert@cl.uni-heidelberg.de

Abstract

The disambiguation of causative-passive homonymy (CPH) is potentially tricky for machines, as the causative and the passive are not distinguished by the sentences' syntactic structure. By transforming CPH disambiguation to a challenging natural language inference (NLI) task, we present the first Chinese Adversarial NLI challenge set (CANLI). We show that the pretrained transformer model RoBERTa, fine-tuned on an existing large-scale Chinese NLI benchmark dataset, performs poorly on CANLI. We also employ Word Sense Disambiguation as a probing task to investigate to what extent the CPH feature is captured in the model's internal representation. We find that the model's performance on CANLI does not correspond to its internal representation of CPH, which is the crucial linguistic ability central to the CANLI dataset. CANLI is available on Hugging Face Datasets (Lhoest et al., 2021) at `https://huggingface.co/datasets/sxu/CANLI`
Keywords: natural language inference, causative-passive homonymy, Chinese, adversarial dataset

1. Introduction

Pretrained Language Models (PLMs) have recently achieved significant progress in natural language understanding tasks. Some models even claim to have surpassed human performance (Devlin et al., 2019; Wang et al., 2019a). However, recent research (Bender and Koller, 2020) questions whether these models really understand the meaning of natural language.

NLI is a canonical natural language understanding task, where a system must determine the relationship, such as entailment or contradiction, between a premise and a hypothesis. NLI performance is tested on large-scale datasets, of which several exist for English such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). Some non-English NLI datasets exist but are mostly generated either automatically or translated from existing English NLI datasets (Ham et al., 2020; Conneau et al., 2018). OCNLI (Hu et al., 2020) is a large-scale NLI dataset for Chinese that has *not* been generated by translation.

State-of-the-Art PLMs achieve high performance for NLI tasks, reporting up to 90% accuracy of English RoBERTa on MNLI (Wang et al., 2019b; He et al., 2020) and around 78% of Chinese RoBERTa on OCNLI (Hu et al., 2020). However, some researchers (Bender and Koller, 2020) have recently challenged whether these benchmarks test natural language understanding in full and pointed out that the models might achieve their excellent performance by relying on spurious statistical patterns in the data instead of by understanding meaning. This has been shown by feeding NLI systems with adversarial data sets where they might fail. These adversarial datasets are mostly limited to English,

We are to the best of our knowledge the first to construct an adversarial NLI dataset for Chinese[1]. Our hypothesis is that certain linguistic phenomena can be exploited for adversarial NLI testing. Linguists have studied homonymy and other *implicit* phenomena for decades (Panman, 1982; Lakoff, 1993; Sag, 1976). These phenomena often involve common sense reasoning and context information; in other words, abilities beyond spurious statistical patterns. For instance, when the same morpheme can mark both the causative voice and the passive voice, we call it a causative-passive homonym (CPH). There are no differences in the verbal constructions; it is the context that determines whether the verb should be read as causative or passive. The CPH exists in several languages such as Chinese, Korean, Turkish (and to some extent also in English and French) (Knott, 1995). In this paper, we use the CPH to create a linguistically motivated Chinese Adversarial NLI challenge set (CANLI). We hope that this can accelerate progress in adversarial NLI datasets across language borders. Furthermore, most previous research on homonymy/polysemy disambiguation focused on lexical semantics (Wiedemann et al., 2019; Garí Soler and Apidianaki, 2021). Instead, CPH is a morpho-syntactic phenomenon resulting from the historical grammaticalization process (Yap and Iwasaki, 2007). In this paper, we present a simple yet interpretable method to investigate to what extent the CPH feature is represented in the PLMs, using the example of RoBERTa.
The main contributions of this paper are as follows:

- We introduce the use of the linguistic phenomenon CPH to create adversarial NLI data sets. As a case study, we present CANLI, the first Chinese

[1]We use Standard Mandarin in this paper. Throughout this paper, when we refer to *Chinese*, we mean Mandarin.

Examples	Template	Voice of P
P: Jingji-weiji rang gongsi daobi le. ('The economic crisis caused the company to close down.') $\not\rightarrow$ H: Jingji-weiji daobi le. ('The economic crisis closed down.')	P: N1 *rang* N2 VP $\not\rightarrow$ H: N1 VP	Causative
P: Jingji-weiji rang gongsi daobi le. ('The economic crisis caused the company to close down.'). $\rightarrow$ H: Gongsi daobi le. ('The company closed down.')	P: N1 *rang* N2 VP $\rightarrow$ H: N2 VP	Causative
P: Ta rang gongsi kaichu le. ('He was fired by the company') $\rightarrow$. H: Gongsi kaichu le ta. ('The company fired him.')	P: N1 *rang* N2 VP $\rightarrow$ H: N2 VP N1	Passive
P: Ta rang gongsi kaichu le. ('He was fired by the company') $\not\rightarrow$ H: Ta kaichu le gongsi. ('He fired the company')	P: N1 *rang* N2 VP $\not\rightarrow$ H: N1 VP N2	Passive

Table 1: Templates for CANLI. The abbreviations we use are the following. P : Premise, H: Hypothesis, N1: the first Noun Phrase, N2: the second Noun Phrase, VP: Verb Phrase, $\rightarrow$: Entailment, and $\not\rightarrow$: Non-Entailment.

adversarial NLI dataset. CANLI is available on Hugging Face Datasets (Lhoest et al., 2021).[2]

- We test the large pretrained transformer model RoBERTa (Liu et al., 2019) on CANLI. We show that it performs poorly when fine-tuned on OCNLI and needs specific fine-tuning on CANLI to improve.
- We use word sense disambiguation as a probing task and find that RoBERTa's performance on CANLI does not correspond to its internal representation of CPH, which is exactly the linguistic ability required for CANLI.

2. Linguistic Background of CPH

In many languages, passive voice and causative voice are canonically marked by different morphemes, as the following English examples show:

(1) a. She gets her husband to do the cleaning.
(causative : get + infinitive)

b. Her wallet was stolen.
(passive: be + past participle).

However, a causative-passive homonymy has been observed in some languages, where one single morpheme can mark these two voices. The verb *get* in English serves as a case in point:

(2) a. She got them arrested (by the police).
(Causative: get + past participle)

b. She got her wallet stolen (by someone).
(Passive: get + past participle)

When the same morpheme can convey either of the two voices, we call it a causative-passive homonym. The disambiguation of this homonymy is easy for human readers. However, it is difficult for a machine, because there is no difference in the formal structure of the sentences. In order to distinguish the passive reading from the causative one, one has to apply either context information or common sense reasoning. Such phenomena can also be observed in Korean, Chinese, Japanese, Manchu-Tungusic languages, and others (Yap and Iwasaki, 2003; Robbeets, 2007). For instance, in Chinese, the canonical causative is marked by the morpheme *shi* (3a) and the canonical passive by the morpheme *bei* (3b); meanwhile *rang* can convey either causative in (4a) or the passive in (4b).

(3) a. 经济危机 使 公司 倒闭 了
jingji-weiji **shi** gongsi daobi le
ecnomic-crisis **CAUS** company close-down PFV
'The economic crisis caused the company to close down.'
b. 他 被 公司 开除 了
ta **bei** gongsi kaichu le
he **PASS** company fire PFV
'He was fired by the company.'

(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.'

Examples (4a) and (4b) are examples of full homonymy: there are no differences in the verbal constructions; it is the context that determines whether the voice marker *rang* should be read as causative or passive. Since there is no difference in the formal structure, this homonymy's disambiguation is complex for a machine. However, human readers can in most cases easily differentiate this causative/passive ambiguity.

3. Data Construction

CANLI consists of ordered pairs of sentences: one CPH sentence, including *rang*, selected from the Chinese online corpus CCL (Zhan et al., 2019) as premise and one template-generated sentence as a hypothesis. Each pair is labeled with one of the two labels *Entailment* or *Non-entailment*, following the two-label approach in HANS (McCoy et al., 2019).

Premises Selection. Our premise sentences consist of 400 causative sentences and 400 passive sentences. The first author of this paper (a Chinese native speaker with a linguistics background) collected and annotated the naturally-occurring CPH premises marked by the

[2]https://huggingface.co/datasets/sxu/CANLI

CPH morpheme *rang*. The sentences are drawn from the genre of modern literature in the CCL online corpus (Zhan et al., 2019). While preserving the original meaning and structure, the author slightly modified some of the collected sentences in order to make them easier to read and understand.

Hypothesis Generation. We use the templates in Table 1 to generate hypotheses. For instance, if a premise sentence 'N1 *rang* N2 VP' has a passive reading, then the premise entails 'N2 VP N1' but not 'N1 VP N2'. For each of the premises we therefore generate two hypotheses, leading to overall 1600 sentence pairs, 800 entailed and 800 not-entailed. To ensure the naturalness and quality of the template-generated hypotheses, a native publishing house editor has proofread and edited the collection. The first author of this paper has double-checked the data after the editing process.

4. NLI Experiments

Datasets. We use two datasets.

First, we use OCNLI (Hu et al., 2020) which contains 56,000 human-labeled premise-hypothesis pairs for Chinese, 50K for training and 3K for development/validation and testing each. It is one of the rare natural language inference datasets for Chinese. We believe that the fact that the hypotheses are generated by native speakers of Chinese makes it more suitable than, for example, the Chinese part of XNLI (Conneau et al., 2018), which has been generated by translating MNLI (Williams et al., 2018) to Chinese. In addition, XNLI has only (professionally) translated the development and testing part to English and is therefore much smaller than OCNLI. In addition, OCNLI is also part of CLUE (Xu et al., 2020). Although large and diverse, OCNLI has not focused on any specific linguistic phenomenon. OCNLI is annotated with three labels (*entailment*, *neutral* and *contradiction*) and its distribution is (almost) balanced across the three labels, in both train and validation set. To match our labeling, we merged *neutral* and *contradiction* as *non-entailment* in the OCNLI dataset, creating a dataset with 2/3 non-entailment, and 1/3 entailment. OCNLI comes with a predetermined training/validation/test split. As the labels of the test set are not publicly available we train on the training set `OCNLI.train` and evaluate on its validation set `OCNLI.val`.

Second, we use CANLI, which was annotated with two labels *entailment* and *non-entailment*. To generate a training/test split, we split the 800 premises into 50:50 for training and testing. As each premise comes with two hypotheses (one entailed and one not entailed), this leads to 800 pairs for training and 800 pairs for testing. Both training and test sets are balanced with regards to the classification variable (entilament vs. non-entailment) and with regards to causative/passive.

Upper Bound: Human Performance. To measure human performance on CANLI, we asked 5 Chinese native speakers to label a sample of the CANLI test set. First, we provided them with instructions and ten training examples. Then they were given the answers to the training examples. Finally, we gave each annotator a random sample of 100 examples from the CANLI test set for labeling.[3] We compared their labels against the gold labels in the CANLI test set to calculate accuracy. Human accuracy on CANLI is 93.2% on average. This is slightly better than the human accuracy on OCNLI and MNLI, as our task is somewhat easier because we distinguish between two instead of three labels.

Language Model and Experimental Setup. We choose to use the Chinese version of RoBERTa-large (Liu et al., 2019) as it is reported to achieve the best performance on OCNLI (Hu et al., 2020). We fine-tuned the model on the OCNLI training set (with the merged two labels). We used hfl/chinese-roberta-wwm-ext-large (Cui et al., 2020) with a sequence classification/regression head on top from the transformers library (Wolf et al., 2020). The model is fine-tuned with 3 epochs, a learning rate of 3e-5, following the hyperparameters used in OCNLI. We adjusted the batch size to 16 because of the limited memory size of our GPU.

As evaluation measures we use accuracy as well as precision, recall and balanced F1 for the entailment class.

Results. The OCNLI-fine-tuned model performs well on the OCNLI validation set with an accuracy of 87.4% and an F-measure of 80.1%. This result is higher than the 78.8% accuracy reported for RoBERTa in the original OCNLI paper (Hu et al., 2020) as we have merged the 3-label-problem into a 2-label problem.

Despite the high score on the OCNLI validation set, the OCNLI-fine-tuned model performed poorly when tested on the CANLI test set (Table 2). The model assigned the label *entailment* in the vast majority of cases, which leads to an accuracy of 48.1 and a precision for 48.9 for *entailment* cases, only, but obviously with a high recall of 88.2. As it is possible that OCNLI's poor performance on CANLI is due to the fact that the label distribution of `OCNLI.train` is unbalanced after label merging whereas the one in `CANLI.test` is balanced, we also ran an experiment where we fine-tuned on a balanced subset of `OCNLI.train`, which we call `OCNLI.train.bal`. We construct `OCNLI.train.bal` by using all entailment instances of OCNLI and a random sample of the same size of all other, non-entailment instances. Performance when fine-tuning on `OCNLI.train.bal` drops slightly to 86.2% accuracy when testing on `OCNLI.val` as expected. Performance on CANLI remains low which shows that label distribution mismatches are not the main reason for the low performance on CANLI when fine-tuning on OCNLI.

Augmenting OCNLI with CANLI. McCoy et al. (2019) show that PLMs perform significantly better when the training set is augmented with examples with

[3] As each annotator annotated potentially different samples, we could not compute agreement.

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.bal	86.2 (0.4)	75.1 (0.8)	85.1 (0.2)	79.8 (0.4)	47.8 (1.4)	48.8 (0.8)	91.8 (3.9)	63.7 (1.6)
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)

Table 2: Test performance on CANLI and OCNLI for RoBERTa when fine-tuned on the different training sets. We report mean accuracy across five fine-tuning runs with the standard deviation as well as balanced F1, precision and recall for the entailment class.

similar syntactic patterns to the test set. Although our problem is not characterized by syntactic variation, we hypothesize that it is still possible that the model just needs to see enough *rang* examples. We therefore add `CANLI.train` to `OCNLI.train`. Fine-tuning with this augmented training set indeed helps substantially when testing on CANLI without lowering performance on the OCNLI validation set (see line 3 in Table 2). This raises the questions: has the model learned the morpho-syntactic feature of CPH after augmenting? To what extent can we find the CPH feature in the model's internal representation? In the next section, we explore the model's representation of CPH both qualitatively and quantitatively.

5. The Representation of CPH

Contextualised Embedding of *Rang*. Contextualized Word Embeddings (CWE) provided by Transformers, such as RoBERTa, depend on the context. Previous studies (Wiedemann et al., 2019; Giulianelli et al., 2020) show that CWEs are able to disambiguate polysemous words. The idea is based on the distributional hypothesis (Harris, 1954; Firth, 1957): 'if the same word regularly occurs in different, distinct contexts, we may assume polysemy of a word's meaning.' as cited in Wiedemann et al. (2019). In this paper, we use CWEs of *rang* to disambiguate causative and passive. There are no differences in the verbal constructions of CPH; it is the context that determines whether *rang* should be read as causative or passive (Section 2). Moreover, *rang* in causative sentences can be replaced by the canonical causative marker *shi* (sentence 3a and 4a); and in passive sentences, *rang* can be replaced by the canonical passive marker *bei* (sentence 3b and 4b). Therefore, we expect that the CWEs of causative *rang* are closer to that of *shi*; and the CWEs of passive *rang* are closer to that of *bei*.

Exploratory Visualization. To explore the model's internal representation of CPH, we visualized the CWEs of the voice markers (*bei*, *shi*, *rang*). Apart from the 800 CPH premises in CANLI, we also collected 40 causative sentences marked by *shi* and 40 passive sentences marked by *bei* from the CCL corpus. We sent these sentences to the RoBERTa models as input, and for each sentence we retrieved the CWE for the voice markers from the last hidden layer. Then we visualize these CWEs using UMAP (McInnes et al., 2018) with the default configuration. Figure1 demonstrates that though there are clear clusters of different markers, there is no clear relationship regarding the causative and passive in the CWEs pulled from the vanilla RoBERTa. Meanwhile, when we look at the representation when fine-tuned with both `OCNLI.train` and `CANLI.train`, we can see that *shi* become more similar to causative *rang*; and *bei* become more similar to passive *rang*. The apparent detail in the visualization suggests that the fine-tuning with augmented `CANLI.train` helps improve the model's internal representation of CPH. However, dimension reduction tools (such as UMAP) for data visualization can be misleading; these algorithms can produce considerably different outputs on different hyperparameters (Wang et al., 2021).

Word Sense Disambiguation as a Probing Task. To quantitatively investigate the models' internal representations of CPH, we test whether a simple probe can perform well at a CPH disambiguation task. Inspired by the Word Sense Disambiguation (WSD) method used in Coenen et al. (2019), we use a nearest-neighbor classifier where each neighbor is the centroid of the CWEs of the canonical voice marker *bei* (passive) / *shi* (causative). Namely, given a CWE of *rang*, if its nearest neighbor is the centroid of *bei*, the probe then classifies it as passive, and vice versa. Note that this probe is unsupervised as it only computes similarity between the canonical voice marker(s) and *rang*.

We used the same sentences as in the visualization task and retrieved the CWEs for the voice markers from every hidden layer, and then calculated the probing accuracy. Figure 2 shows that, at the last layer, RoBERTa fine-tuned with `CANLI.train` achieved a better performance than the vanilla RoBERTa and RoBERTa when only fine-tuned on OCNLI. However, all three models obtained their best performance at the 20th layer and then degraded by the final layers. It suggests that RoBERTa captures the morpho-syntactic CPH feature in the late-middle layers and the feature gets diluted in the higher layers. These results are consistent with previous studies: PLMs embed "surface features in lower layers, syntactic features in middle layers and semantic features in higher layers (Jawahar et al., 2019)". In addition, the last layers of the models change most during fine-tuning (Kovaleva et al., 2019). However, fine-tuning is supposed to teach the model to rely more on

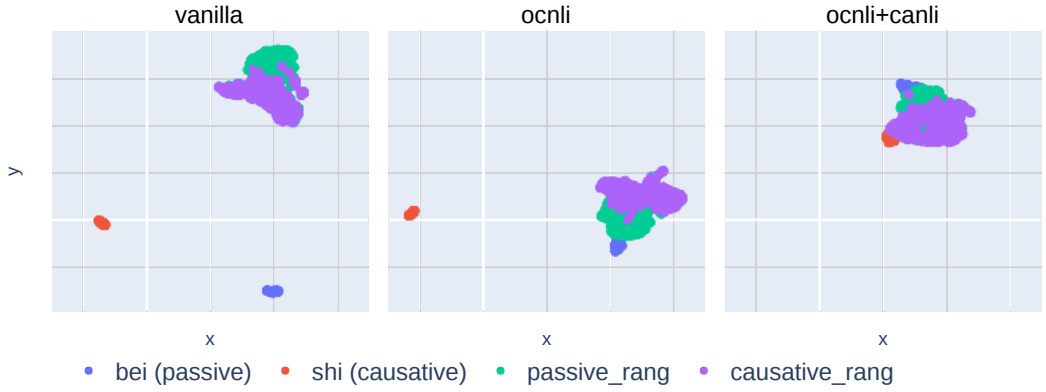

Figure 1: CWEs for the voice markers *rang, shi, bei*, visualized with UMAP. Frome left to right, CWEs pulled from: 1) Vanilla RoBERTa, 2) RoBERTa fine-tuned with OCNLI.train, 3) RoBERTa fine-tuned with OCNLI.train and CANLI.train

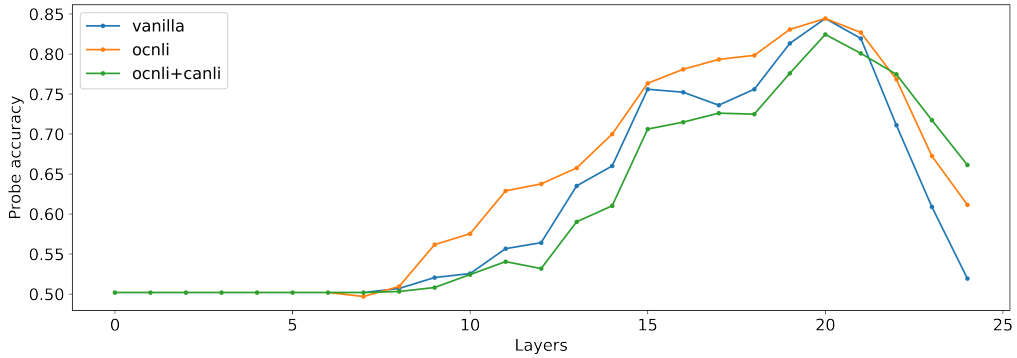

Figure 2: Probe accuracies on RoBERTa fine-tuned on the different training sets.

the representations useful for the task (Rogers et al., 2020). Therefore it is surprising that, even when fine-tuned on OCNLI+CANLI, the CPH probe accuracy still decreases in the final layer, albeit less than the other 2 models (see Figure 2). This result poses the question, how can we 'teach' the NLI systems reasoning? We leave analysis in this direction to future work.

6. Related Work

NLI adversarial datasets. Up to now adversarial datasets are mostly limited to English. To the best of our knowledge we present the first adversarial NLI dataset for Chinese.

The methods used for adversarial data collection for English datasets may not be appropriate for all languages. For example, HAMLET (Nie et al., 2020) requires an existing NLI corpus, which is not available for many non-English languages at present. Another example is HANS (McCoy et al., 2019), which generates the adversarial data set using syntactic heuristics templates. A similar approach has been used to generate Japanese adversarial data (JaNLI) (Yanaka and Mineshima, 2021). The template-based approach we use for generating hypotheses is a variant of the template-based approach in HANS (McCoy et al., 2019) and JaNLI (Yanaka and Mineshima, 2021). However, they use shallow syntactic variants of the premise sentence to generate hypotheses whereas we focus on a specific linguistic phenomenon — the causative-passive homonymy — to construct an adversarial dataset for Chinese. CPH generates two sentences that have the same syntactic pattern but differ in meaning.

Another potential approach would be to translate adversarial datasets into target languages as has been done for (non-adversarial) NLI datasets such as XNLI (Conneau et al., 2018). However, even if translation quality was guaranteed, issues such as unnatural 'translationese' (Hu et al., 2018) and cultural differences (Sechrest et al., 1972) still exist.

Probing Tasks. The success of PLMs in NLU tasks has excited interest in their internal representations. In an attempt to see what kind of knowledge they capture, researchers have employed probing strategies. Probes are classifiers used to predict linguistic properties from

a model's representation. Most work focused on the syntactic features captured in PLMs (Hewitt and Manning, 2019). Lately, semantic probing tasks that rely more on contextual information have also been investigated (Tenney et al., 2019). Existing probing tasks are mostly designed for English language only. Recently some researchers have extended the work to the multilingual setting (Şahin et al., 2020). To our knowledge, our work is the first Chinese probing task which focuses on a morpho-syntactic linguistic phenomenon and requires contextual understanding.

Some probing works achieved high accuracy on various linguistics tasks (Belinkov et al., 2017). However, it does not necessarily mean that the representations encode linguistic structure. Studies show that probes' accuracy can be similar when probing for genuine linguistic labels and probing for random synthetic tasks (Voita and Titov, 2020). Hewitt and Liang (2019) devise control tasks to measure the probes' selectivity. A good probe should be highly selective; namely, it should perform well on targeted linguistic tasks but poorly on control tasks. Another approach explores the geometry of internal representations for different syntactic and semantic information (Coenen et al., 2019). As the probe is not trained, selectivity is assured. Our probe is also not trained as it simply computes nearest-neighbour similarity.

7. Conclusion

In summary, we present the first Chinese adversarial NLI dataset CANLI based on the linguistic phenomenon of the causative-passive homonymy (CPH). Results using RoBERTa fine-tuned on OCNLI show that CANLI is challenging for a state-of-the-art NLI system such as RoBERTa. We also used word sense disambiguation as a probing task, and demonstrated that RoBERTa's performance on CANLI does not correspond to its internal representation of CPH. This paper is an initial exploration and leaves many open questions. Future work will focus on scaling the size of CANLI. Besides, CPH is a multilanguage phenomenon, so we hope our work will help accelerate progress on datasets for non-English languages other than Chinese.

8. Acknowledgements

This work was supported by German Federal Ministry of Education and Research (BMBF) under grant agreement No. 01IS19063A. We would like to thank Yinjun Wang for his diligent proofreading of the dataset; and our native speakers Yang Li, Tingxian Wu, Qinghua Chen, Jiaying Ma and Yong Xu for their great efforts. We also thank 3 anonymous reviewers for their insightful comments.

9. Bibliographical References

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July. Association for Computational Linguistics.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F. B., and Wattenberg, M. (2019). Visualizing and measuring the geometry of bert. In *NeurIPS*.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November. Association for Computational Linguistics.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online, November. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Garí Soler, A. and Apidianaki, M. (2021). Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Giulianelli, M., Del Tredici, M., and Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, July. Association for Computational Linguistics.

Ham, J., Choe, Y. J., Park, K., Choi, I., and Soh, H. (2020). KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online, November. Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

He, P., Liu, X., Gao, J., and Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hewitt, J. and Liang, P. (2019). Designing and interpreting probes with control tasks. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hu, H., Li, W., and Kübler, S. (2018). Detecting syntactic features of translated Chinese. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 20–28, New Orleans, June. Association for Computational Linguistics.

Hu, H., Richardson, K., Xu, L., Li, L., Kübler, S., and Moss, L. (2020). OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online, November. Association for Computational Linguistics.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July. Association for Computational Linguistics.

Knott, J. (1995). The causative-passive correlation. *Subject, voice, and ergativity. London*.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China, November. Association for Computational Linguistics.

Lakoff, G. (1993). The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and Thought*. Cmbridge University press.

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., and Wolf, T. (2021). Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

McCoy, T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July. Association for Computational Linguistics.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July. Association for Computational Linguistics.

Panman, O. (1982). Homonymy and polysemy. *Lingua*, 58(1-2):105–136.

Robbeets, M. (2007). The causative-passive in the transeurasian languages. *Turkic Languages*, 11:235–278.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sag, I. A. (1976). A note on verb phrase deletion. *Linguistic Inquiry*, 7(4):664–671.

Sechrest, L., Fay, T. L., and Zaidi, S. H. (1972). Problems of translation in cross-cultural research. *Journal of cross-cultural psychology*, 3(1):41–56.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S. R., Das, D., and Pavlick, E. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Voita, E. and Titov, I. (2020). Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online, November. Association for Computational Linguistics.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019a). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Wang, W., Bi, B., Yan, M., Wu, C., Bao, Z., Xia, J., Peng, L., and Si, L. (2019b). StructBERT: Incorporating language structures into pre-training

for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Wang, Y., Huang, H., Rudin, C., and Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *Journal of Machine Learning Research*, 22(201):1–73.

Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. (2019). Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. *ArXiv*, abs/1909.10430.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., Xu, Y., Sun, K., Yu, D., Yu, C., Tian, Y., Dong, Q., Liu, W., Shi, B., Cui, Y., Li, J., Zeng, J., Wang, R., Xie, W., Li, Y., Patterson, Y., Tian, Z., Zhang, Y., Zhou, H., Liu, S., Zhao, Z., Zhao, Q., Yue, C., Zhang, X., Yang, Z., Richardson, K., and Lan, Z. (2020). CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Yanaka, H. and Mineshima, K. (2021). Assessing the generalization capacity of pre-trained language models through Japanese adversarial natural language inference. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 337–349, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Yap, F. H. and Iwasaki, S. (2003). From causatives to passives: a passage in some east and southeast asian languages. *Cognitive linguistics and non-Indo-European languages*, 18:419.

Yap, F. H. and Iwasaki, S. (2007). The emergence of 'give'passives in east and southeast asian languages. *SEALS VIII*, page 193.

Şahin, G. G., Vania, C., Kuznetsov, I., and Gurevych, I. (2020). LINSPECTOR: Multilingual Probing Tasks for Word Representations. *Computational Linguistics*, 46(2):335–385, 06.

10. Language Resource References

Zhan, W., Guo, R., Chang, B., Chen, Y., and Chen, L. (2019). The building of the ccl corpus: its design and implementation. *Corp. Ling*, 6:77–86.