Expanding Scope: Adapting English Adversarial Attacks to Chinese

Hanyu Liu, Chengyuan Cai, Yanjun Qi
Department of Computer Science
University of Virginia
Charlottesville, VA, USA
{hl2gn, cc4fy, yq2h}@virginia.edu

Abstract

Recent studies have revealed that NLP predictive models are vulnerable to adversarial attacks. Most existing studies focused on designing attacks to evaluate the robustness of NLP models in the English language alone. Literature has seen an increasing need for NLP solutions for other languages. We, therefore, ask one natural question: whether state-of-the-art (SOTA) attack methods generalize to other languages. This paper investigates how to adapt SOTA adversarial attack algorithms in English to the Chinese language. Our experiments show that attack methods previously applied to English NLP can generate high-quality adversarial examples in Chinese when combined with proper text segmentation and linguistic constraints. In addition, we demonstrate that the generated adversarial examples can achieve high fluency and sentiment consistency by focusing on the Chinese language's morphology and phonology, which in turn can be used to improve the adversarial robustness of Chinese NLP models.

1 Introduction

Adversarial examples are text inputs crafted to fool an NLP system, typically by making small perturbations to a seed input[1]. Recent literature has developed various adversarial attacks generating text adversarial examples to fool NLP predictive models [2]. These attack methods mainly focus on the English language alone, building upon components that use language-specific resources, such as English WordNet (Miller, 1995) or BERT models (Devlin et al., 2018a) pretrained on English corpus.

Literature has seen a growing need for NLP solutions in other languages; therefore, evaluating NLP solutions' robustness via adversarial examples is crucial. We ask an immediate question: "Can we extend the SOTA adversarial attacks in English to other languages by replacing those English-specific inner components with other languages' resources ?". For instance, we can attack a Chinese NLP model by replacing WordNet with HowNet (Dong et al., 2010). However, it is unclear if such a workflow is sufficient for generating high-quality adversarial examples, when a target language differs from English. In this work, we attempt to answer this question by adapting SOTA word substitution attacks designed for English to evaluate Chinese NLP models' adversarial robustness. Moreover, we introduce morphonym and homophone word-substitution attacks that are specific to the Chinese language; they function as a benchmark to the English adapted attack methods.

Our experiments on Chinese classification and entailment models show that both the English-adapted and Chinese-specific attack methods can effectively generate adversarial examples with good readability. The attack success rates of homophone-based and HowNet-derived methods are significantly better than the success rate of masked language model-based attacks or morphonym-derived attacks. We then combine the four attacks mentioned above into a composite attack that further increases the attack success rate to 96.00% in fooling Chinese classification models and 98.16% in attacking entailment models. In addition, we demonstrate that adversarially trained models significantly decrease attack success rate by up to 49.32%.

2 Method

Recent NLP literature includes a growing body of works on adversarial examples in NLP, mostly in English (more background details are in Section A). Most SOTA English adversarial attacks search for a

---

[1]Most existing work attempts to perturb an input using character-level (Ebrahimi et al., 2017a; Gao et al., 2018; Pruthi et al., 2019; Li et al., 2018) or word-level perturbations (Alzantot et al., 2018; Jin et al., 2019; Ren et al., 2019; Zang et al., 2020) to fool a target model's prediction in a specific way.

[2]We use "natural language adversarial example", "text adversarial example" and "adversarial attacks" interchangeably.

perturbation to change a given seed input $\mathbf{x}$ into an adversarial example $\mathbf{x}'$; $\mathbf{x}'$ fools a predictive NLP model and satisfies certain language constraints, like preserving the same semantical meaning as $\mathbf{x}$. Essentially each adversarial attack algorithm has four components: a goal function, a set of constraints, a suite of transformations, and a search algorithm (Morris et al., 2020b). The search algorithm attempts to find a sequence of transformations that results in a successful perturbation. The goal function can be like fooling a target model into predicting the wrong classification label.

Related literature: While most NLP adversarial attacks have focused on the English language, a few recent methods have been proposed for Chinese. Zhang et al. (2020) proposed a black-box attack that performs a glyph-level transformation on the Chinese characters. Related, Li et al. (2020a) and Zhang et al. (2022) added phonetic perturbations to improve the adversarial robustness of Chinese NLP models. All three attacks, however, are only applicable to the Chinese language. Another study (Wang et al., 2020) proposed a white-box attack against BERT models (Devlin et al., 2018b) that performs character-level swaps using gradient optimization. These character-level attacks extend poorly to other languages and tend to generate out-of-context partial substitutions that impact fluency. Later studies, such as Shao and Wang (2022) and Wang et al. (2022), included semantic-based word substitutions but did not consider the significance of constraints and adversarial training. We choose to generalize SOTA word synonym substitution attacks in English to the Chinese language (due to the prevalence of word substitutions) and our attacks consider a range of language constraints.

2.1 Determining Text Segmentation

The first step to crafting a new adversarial attack for the Chinese language is to select the level of transformation. Unlike English, which separates words with space, the Chinese language lacks native separators to determine different words in a sentence. A Chinese character may represent a word, while longer words may include multiple adjacent Chinese characters. To avoid out-of-context perturbations that replace partial components of a multi-character word, we use a Chinese segmentation tool provided by Jieba [3] to segment an input text into a list of words.

[3] https://github.com/fxsjy/jieba

2.2 General Overview of Proposed Attacks

The general perturbation strategy we propose is word synonym substitutions. Given an input text $\mathbf{x}$, we use the aforementioned segmentation tool to segment $\mathbf{x}$ into $[x_1, x_2, \ldots, x_n]$. Subsequent transformations (synonym substitution) are then getting applied to each eligible word [4]. This means we obtain perturbed text $\mathbf{x}'$ by replacing some x_i with its synonym x_i'. Our attack goal is to make the model mis-predict the $\mathbf{x}'$ (i.e. $\mathbf{F}(\mathbf{x}) \neq \mathbf{F}(\mathbf{x}')$), [5] which is also called an untargeted attack. If one substitution is not enough to change the prediction, we repeat the steps to swap another x_j to generate the perturbed text $\mathbf{x}'$. This process essentially solves the following objective:

$$
\begin{aligned}
\text{Find} \quad & \mathbf{x}' = \text{wordSubstitution}(\mathbf{x}) \\
\text{s.t.} \quad & \mathbf{F}(\mathbf{x}) \neq \mathbf{F}(\mathbf{x}') \\
& \mathbf{x}' \in \mathcal{X}, \ \mathbf{F}(\mathbf{x}) = y_{orig} \\
& \wedge C_i(\mathbf{x}, \mathbf{x}'; \epsilon_i), \ \forall i \in \{1, 2, ..., C\}
\end{aligned}
\tag{1}
$$

Here $C_1, ..., C_n$ denotes a set of language constraints including like semantic preserving and grammaticality constraints (Morris et al., 2020b). ϵ_i denotes the strength of the constraint C_i.

The critical component "wordSubstitution$(\mathbf{x})$" in Eq. (1) requires us to figure out what words in $\mathbf{x}$ to perturb first, and what words as next. Essentially this is a combinatorial search issue. Literature includes different search strategy (see Section B.1 for details). We adapt the greedy with word importance ranking based search algorithm here. Our attack chooses the order of words by estimating the "importance" of each x_i in $\mathbf{x}$. The importance of x_i is computed by replacing each x_i with an UNK token and then calculating the change in the model's confidence on the original label. Essentially we sort words x_i in $\mathbf{x}$ by the decreasing importance regarding the following $score$:

$$
\begin{aligned}
score(x_i) &= 1 - Prob(\mathbf{F}(\mathbf{x}'))_{y_{orig}} \\
\text{s.t.} \quad & \mathbf{x}' = \text{replace}(\mathbf{x}, x_i, \text{UNK})
\end{aligned}
\tag{2}
$$

This measures how much the target model's confidence decreases regarding the original label class

[4] In this paper, the phrase "Chinese characters" refers to one unit long token, and "Chinese words" refers to one or more Chinese characters in their semantically correct segmentation that may or may not be one unit long.

[5] Here $\mathbf{F} : \mathcal{X} \rightarrow \mathcal{Y}$ denotes a predictive Chinese NLP model taking Chinese text as input. $\mathcal{X}$ denotes the input space and $\mathcal{Y}$ is the output space.

y_{orig} when replacing x_i with "UNK" token. Then for each selected x_i, we find its best x_i' to swap with, from a candidate synonym set (Section 2.3)

2.3 Generating Synonyms for Words

Now for a selected word x_i in $\mathbf{x}$, we propose four different Chinese word transformation strategies to perturb a word x_i into x_i' through the following word transformations:

We design the first two transformations by adapting from English attack studies (Jin et al., 2019) and (Garg and Ramakrishnan, 2020).

- **Open HowNet.** Open HowNet (Qi et al., 2019) is a sememe-based lexical dataset that is consisted of a sememe set and the corresponding phrases annotated with different sememes. A sememe is defined as the minimum semantic unit in a language, and Open HowNet incorporates relations between sememes to construct a taxonomy for each sememes. The semantic similarity between two words can be calculated by comparing their annotated sememes. In our study, we use Open HowNet to generate synonyms by searching the top five words with the highest semantic similarity with an input Chinese word.
- **Masked Language Model.** We adapt the masked language model (MLM) method to generate perturbations based on the top-K predictions by a MLM. The XLM-RoBERTa model (Conneau et al., 2019) was used as the MLM in our study, as it is able to predict Chinese words consisting of multiple characters to preserve the fluency of the attacked sentence better, in comparison to other prevalent MLM (mac-bert, etc.) that predicts single characters alone.

The Chinese language, along with other Eastern Asian languages, differs from English, especially in phonology and morphology.[6] Using these in-

---

[6] Each Chinese character represents a monosyllabic word with unique combinations of pictographs, while English words consist of alphabetic letters. Though each Chinese character's morphology combination is unique, many characters with similar morphology structures can be substituted in an adversarial attack without impacting the readability of the attacked sentence. In addition, because there exist many homophones in modern Chinese, the same spoken syllable may map to one of many characters with different meanings. The phonology of Chinese characters is commonly transcribed into the Latin script using Pinyin. Typing the wrong character of a word in Pinyin despite having the same pronunciation is a common mistake in Chinese writing. Thus, replacing Chinese characters with the same pronunciation may serve as an additional attack method to test the adversarial robustness of NLP models while preserving the semantics for human readers.

tuitions, we design two special word transformations considering phomophones and morphonyms of Chinese language.

- **Homophone transformation.** Since the phonology of Chinese characters can be expressed by the romanization system Pinyin. To replace a Chinese character with its homophone, top-k words are randomly selected from a list of characters with the same Latin script.
- **Morphonym transformation.** Similarly, to replace a character with its morphonyms, top-k words are randomly selected from a list of characters that share partial pictographs with the target character, as it is a common mistake for Chinese writers to mistaken one pictograph with another.
- **Composite transformation.** We also design a composite transformation that consists of the four transformation methods listed above. For each target word, Open HowNet, Masked Language Model, Homophone, and Morphonym perturbations are separately generated to replace a candidate word from the input text. If none of the substitutions changes the target NLP model prediction, the attack then move on to replace the next important word in the input sentence.

In addition, for each perturbation, we want to ensure that the generated $\mathbf{x}'$ preserves the semantic consistency and textual fluency of $\mathbf{x}$. We use three constraints, namely (1) constraint to allow only non-stop word modification, (2) constraint to allow only no-repeat modification, and (3) multilingual universal sentence encoder (MUSE) similarity constraint that filter out undesirable replacements (Cer et al., 2018) [7]. These constraints can easily adapt to other languages. A detailed description of each constraint is in Section B.2. The pseudo-code of our proposed attacks is in Algorithm 1.

In summary, each word transformation strategy gets combined with the greedy word ranking algorithm (Section 2.2) plus the language constraints (see above), making a unique adversarial attack against Chinese NLP.

3 Results and Evaluation

Victim Models: We chose to perform attacks on two Chinese NLP models: one for sentiment classification and one for entailment. BERT and RoBERTa as selected as our victim models due to their reported robustness and SOTA performance.

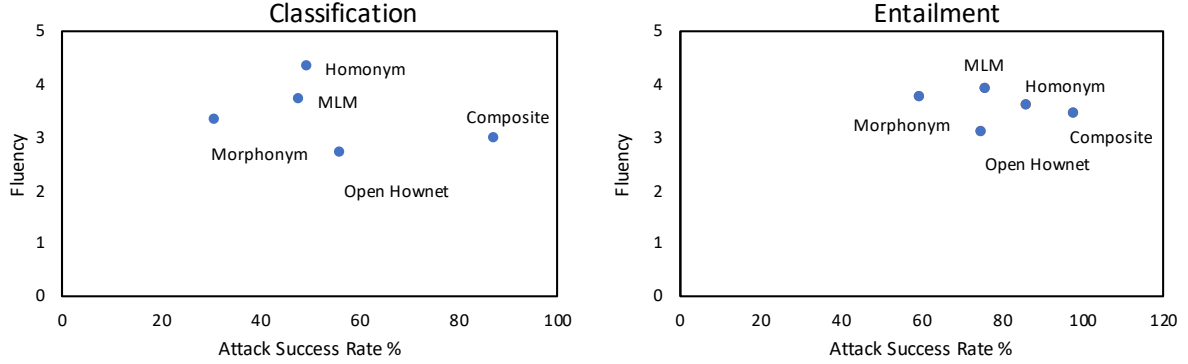[7] We require that the MUSE similarity is above 0.9.

Figure 1: The performance of composite attack method with STM-RM-MUSE constraint regarding the attack success rate and human-evaluated fluency on BERT classification model (left), and RoBERTa entailment model (right). For both classification and entailment tasks, composite transformation achieves the highest attack success rate without a significant trade-off in fluency, while morphonym transformation has the lowest attack success rate.

Details of the two models and its related two Chinese datasets are presented in Section C.2.

Metrics: For each attack method, we recorded the attack success rate and perturbation percentage, skipping samples that a target model fails to predict correctly before any perturbation.

Ablation: To measure how MUSE constraint impact the quality of Chinese adversarial examples, we add baseline attacks that use only the stop word constraint and repeat constraints for ablation study.

Figure 1 connects attack success rate and fluency in one figure. Figure 2 and Figure 3 show few Chinese adversarial examples generated by our attacks. More results can be found in Section C.3

Results on Attack Success: Figure 1, Table 1 and Table 2 present the quantitative results of our attacks. Figure 1 (left) is about our results on Chinese sentiment classification model. Among all non-composite-transformation based attacks, we can see that Open HowNet substitution achieves the highest success rate, while morphonym substitution has the lowest success rate. From Table 1, we can also see that having the MUSE constraint dramatically decreases the attack success rate and perturbation percentage for all attack methods, especially for Open HowNet and homonym substitutions based attacks. This makes sense as the MUSE constraint is designed to limit the amount of perturbation the attacks can do to improve the quality of generated adversarial example. In addition, when we compare the success rate and perturbation percentage of composite attack versus other individual attack methods, we see that composite attack achieves a 87.50% attack success rate without increasing the perturbation percentage. We can make similar conclusions from Figure 1 (right) and

Table 2.

Human Evaluation: For each of the attack method, we randomly sampled 30 adversarial examples produced from the same set of input texts for each attack (a total of five). We asked four volunteers to score the semantic consistency and fluency of the examples. Semantic consistency refers to how well the ground truth label of the adversarial example matches with the original label of the input, and fluency refers to the cohesiveness of the sentence. Both metrics are scored on a scale of 1 to 5, with a score of 5 being the most consistent or fluent.

Table 3 and Table 4 respectively summarize the human evaluations of adversarial examples generated by fooling classification and entailment models. For classification, Table 3 (plus Figure 1) shows that homonym substitution outperforms other attack methods, as its examples have both the highest consistency and fluency scores. On the other hand, Open HowNet substitution reports the lowest quality scores, indicating its generated adversarial examples either include out-of-context substitution or disrupt the cohesiveness semantics. Table 3, plus Figure 1(right) for entailment tasks, shows that homonym attack still achieves the highest consistency score, while MLM achieves the highest fluency score. Besides, we conjecture that the low consistency and fluency scores of the composite attack method may root to its inclusion of adversarial examples generated by Open HowNet.

Adversarial training and more result discussions: Furthermore, we conduct adversarial training (AT) (see details in Section C.1). Table 5 shows the positive results of AT that improve the robustness across all five proposed attacks over both models.

279

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Zhendong Dong, Qiang Dong, and Changling Hao. 2010. Hownet and its computation of meaning. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, page 53–56, USA. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017a. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017b. Hotflip: White-box adversarial examples for text classification. In *ACL*.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. 2019. On the robustness of self-attentive models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1520–1529, Florence, Italy. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? natural language attack on text classification and entailment. *ArXiv*, abs/1907.11932.

Jinfeng Li, Tianyu Du, Shouling Ji, Rong Zhang, Quan Lu, Min Yang, and Ting Wang. 2020a. Textshield: Robust text classification based on multimodal embedding and neural machine translation. In *USENIX Security Symposium*.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020b. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

John Morris, Eli Lifland, Jin Yong Yoo, and Yanjun Qi. 2020a. TextAttack: A framework for adversarial attacks in natural language processing. *ArXiv*, abs/2005.05909.

John X. Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020b. Reevaluating adversarial examples in natural language.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy (EuroS&P)*.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. *arXiv preprint arXiv:1905.11268*.

Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Qiang Dong, Maosong Sun, and Zhendong Dong. 2019. Openhownet: An open sememe-based lexical knowledge base. *arXiv preprint arXiv:1901.09957*.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Yuyao Shao and Liming Wang. 2022. Gpsattack: A unified glyphs, phonetics and semantics multi-modal attack

against chinese text classification models. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Zhouxing Shi and Minlie Huang. 2020. Robustness to modification with shared words in paraphrase identification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 164–171, Online. Association for Computational Linguistics.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Boxin Wang, Boyuan Pan, Xin Li, and Bo Li. 2020. Towards evaluating the robustness of chinese bert classifiers.

Boxin Wang, Chejian Xu, Xiangyu Liu, Yu Cheng, and Bo Li. 2022. SemAttack: Natural textual attacks via different semantic spaces. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 176–205, Seattle, United States. Association for Computational Linguistics.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics.

Xiang Zhang and Yann LeCun. 2017. Which encoding is the best for text classification in chinese, english, japanese and korean?

Zihan Zhang, Jinfeng Li, Ning Shi, Bo Yuan, Xiangyu Liu, Rong Zhang, Hui Xue, Donghong Sun, and Chao Zhang. 2022. RoChBert: Towards robust BERT fine-tuning for Chinese. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3502–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zihan Zhang, Mingxuan Liu, Chao Zhang, Yiming Zhang, Zhou Li, Qi Li, Haixin Duan, and Donghong Sun. 2020. Argot: Generating adversarial readable chinese texts. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2533–2539. International Joint Conferences on Artificial Intelligence Organization.

4 Limitations

We are optimistic that the algorithmic workflow presented in this paper can be generalized to other languages. When the victim models are in languages other than Chinese and English, however, we also acknowledge the uncertainty in achieving a high attack success rate while at the same time achieving fluency in generated examples. In addition, because of the variation in linguistic structures across different languages, further efforts are required to design language-specific transformation methods (such as the homophone and morphonym transformations for the Chinese language case in this paper).

5 Ethics Statement

In this study, we honor the ethical code in the ACL Code of Ethics.

A Background: NLP Adversarial Attacks

Adversarial examples are inputs crafted to fool a machine learning system, typically by making small perturbations to a seed input (Szegedy et al., 2013; Goodfellow et al., 2014; Papernot et al., 2016; Moosavi-Dezfooli et al., 2016). The study of natural language processing (NLP) in adversarial environments is an emerging topic as many online platforms provide NLP based information services, like toxic content detection, misinformation or fake news identification. These applications make NLP frameworks potential targets of adaptive adversaries.

Adversarial attacks aim to use a set of transformations $T_1...T_k$ to perturb a correctly predicted instance, $\mathbf{x} \in \mathcal{X}$, into an adversarial instance $\mathbf{x}'$. Attacks normally define a goal function $FoolGoal(\mathbf{F}, \mathbf{x}')$ that represents whether the goal of the attack has been met, for instance, indicating if the prediction $\mathbf{F}(\mathbf{x}')$ differs from $\mathbf{F}(\mathbf{x})$. Attacks in NLP normally needs another set of Boolean functions $C_1...C_n$ indicating whether the perturbation satisfies a certain set of language constraints.

Initial studies on NLP adversarial attacks performed character-level perturbations to create misspellings (Ebrahimi et al., 2017a; Gao et al., 2018; Pruthi et al., 2019; Li et al., 2018). Recent later works have explored various word substitutions methods to produce adversarial examples in natural language. Both Alzantot et al. (2018) and Jin et al. (2019) use counter-fitted word embeddings to find synonyms while Ren et al. (2019) and Zang et al. (2020) use lexical databases like WordNet (Miller, 1995) and HowNet (Dong et al., 2010). Lately, masked language models have been used to perform word substitutions to preserve fluency of the perturbed text better (Li et al., 2020b; Garg and Ramakrishnan, 2020; Shi and Huang, 2020).

Algorithm 1 Word Substitution Attack against Chinese NLP Models

---
1: **Input:** Input text $\mathbf{x}$
2: $\mathbf{x} = segment(\mathbf{x}) = [x_1, \ldots, x_n]$
3: $R =$ ranking $r_1, \ldots, r_n$ of words $x_1, \ldots x_n$
4: $\mathbf{x}^* = \mathbf{x}$
5: **for** $i = r_1, \ldots, r_n$ **do**
6:      $X_{\text{candidate}} = T_1(\mathbf{x}, i) \cup \cdots \cup T_k(\mathbf{x}, i)$
7:      $X'_{\text{candidate}} = \{x' | C_j(x', x) \quad \forall C_j \in \{C_1 \ldots C_n\}$
8:      **if** $X'_{\text{candidate}} \neq \emptyset$ **then**
9:          $\mathbf{x}^* = \arg\max_{\mathbf{x}' \in X'_{\text{candidate}}} score(\mathbf{x}')$
10:          **if** $F(\mathbf{x}^*) \neq F(\mathbf{x})$ **then**
11:              return $\mathbf{x}^*$
12:          **end if**
13:      **else**
14:          end search
15:      **end if**
16: **end for**

B More Method Details

B.1 Details on Search for Words to Perturb

Solving Eq. (1) needs the Chinese adversarial attacks to conduct a combinatorial search task and adapt search algorithms from the English adversarial attacks in this paper. The search algorithm aims to perturb a text input with language transformations such as synonym substitutions in order to fool a target NLP model while the perturbation adheres to linguistic constraints.

The potential search space is exponential by nature. Assuming $\mathbf{x}$ includes n words, and each word has S potential substitutions, the total number of possible perturbed inputs is then $(S+1)^n - 1$. The search space of all potential adversarial examples for a given $\mathbf{x}$ is far too large for an exhaustive search. This is why many heuristic search algorithms were proposed in the literature, including greedy method with word importance ranking (Gao et al., 2018; Jin et al., 2019; Ren et al., 2019), beam search (Ebrahimi et al., 2017b), and population based genetic algorithm (Alzantot et al., 2018). While heuristic search algorithms cannot guarantee an optimal solution, they can efficiently search for a valid adversarial example.

B.2 Details on Language Constraints

NLP adversarial attacks generate perturbations and use a set of constraints to filter out undesirable $\mathbf{x}'$ to ensure that perturbed $\mathbf{x}'$ preserves the semantics

and fluency of the original $\mathbf{x}$ (Morris et al., 2020a). Therefore, we propose to use three following constraints:

- **Stop word modification**: Replacing the coordinating conjunctions and pronouns within a sentence often changes the semantics of a target sentence. Therefore, words such as "but" and "I" cannot be perturbed.
- **Repeat modification**: This prevents replaced words to be modified again, as the targeted word may gradually diverge from its original meaning.
- **Multilingual Universal Sentence Encoder (MUSE)**: Using the multilingual sentence encoder, we encode both original x and x' and measure the cosine similarity between the two text. We require that the cosine similarity is above 0.9.

C More on Results and Setup

Attack Method	Constraints	Success Rate	% Perturbed
Open HowNet	STM-RM	81.20	32.54
	STM-RM-MUSE	56.49	27.99
MLM	STM-RM	53.67	40.33
	STM-RM-MUSE	48.03	28.71
Homonym	STM-RM	74.14	54.67
	STM-RM-MUSE	49.57	38.22
Morphonym	STM-RM	43.21	45.73
	STM-RM-MUSE	31.02	36.63
Composite	STM-RM	96.00	41.05
	STM-RM-MUSE	87.50	31.74

Table 1: Attack results of classification task performed on online-shopping review dataset. Attack success rate and amount of perturbations of each attack. "STM-RM" stands for stop word modification and repeat modification, and "STM-RM-MUSE" stands for stop word modification, repeat modification, and universal sentence encoder constraint.

Attack Method	Constraints	Success Rate	% Perturbed
Open HowNet	STM-RM	74.94	37.35
	STM-RM-MUSE	75.00	36.59
MLM	STM-RM	75.40	40.88
	STM-RM-MUSE	75.79	40.93
Homonym	STM-RM	86.21	49.63
	STM-RM-MUSE	86.21	49.58
Morphonym	STM-RM	60.23	46.47
	STM-RM-MUSE	59.81	46.60
Composite	STM-RM	98.16	39.06
	STM-RM-MUSE	98.02	38.74

Table 2: Attack results on Chinese entailment model using the Chinanews dataset. Attacks' setup same as Table 1.

Attack Method	Consistency(1-5)	Fluency(1-5)	Δ Fluency
Open HowNet	2.94	2.69	-2.31
MLM	3.41	3.70	-1.28
Homonym	4.44	4.31	-0.69
Morphonym	3.75	3.31	-1.69
Composite	3.13	2.94	-2.06

Table 3: Human evaluation of attacks on Online-shopping dataset. We report average consistency and fluency scores on examples generated from each attack method. STM-RM-MUSE constraints were used for all attack methods.

Attack Method	Consistency(1-5)	Fluency(1-5)	Δ Fluency
Open HowNet	3.10	3.07	-1.93
MLM	3.20	3.90	-0.87
Homonym	4.43	3.57	-1.33
Morphonym	3.90	3.73	-1.27
Composite	2.97	3.43	-2.53

Table 4: Human evaluation of attacks on Chinanews dataset. STM-RM-MUSE constraints were used for all attack methods.

C.1 Adversarial Training

Attack Method	Constraints	Pre Success Rate	AT Success Rate	Δ%
Open HowNet	STM-RM	81.20	65.05	-19.89
	STM-RM-MUSE	56.49	43.30	-23.35
MLM	STM-RM	53.67	36.10	-32.74
	STM-RM-MUSE	48.03	38.94	-18.93
Homonym	STM-RM	74.14	40.83	-44.93
	STM-RM-MUSE	49.57	35.86	-27.66
Morphonym	STM-RM	43.21	30.51	-29.39
	STM-RM-MUSE	31.02	15.72	-49.32
Composite	STM-RM	96.00	77.75	-19.01
	STM-RM-MUSE	87.50	60.34	-31.04

Table 5: Results of adversarial training performed on BERT model. "Pre Success Rate" stands for the success rate of composite attack on the pre-adversarial-trained model, and "AT Success Rate" stands for the success rate of composite attack on adversarial-trained model

To evaluate how adversarial examples generated by the attack methods could improve the adversarial robustness of a target model, we attacked the target BERT model (Chinese sentiment classification) with 1000 examples from the training set of an online shopping review dataset, and finetune the target model with the successfully attacked examples. The model was trained for 3 epochs with 1 initial clean epochs, learning rate of 5e-5, and effective batch size of 32 (8x4).

Table 5 shows the positive effect of adversarial training (AT) that improve the robustness of Chinese language models against all five of our proposed attacks. For instance, on the target BERT model, attack success rate decreased by up to 49.32% after being trained by adversarial examples generated by the Composite-MUSE attack method. Across all attacks, the AT-trained models result with a significant drop in attack success rate.

C.2 Victim Model and Dataset Setup

We chose to perform attacks on Chinese sentiment classification and entailment models, and chose BERT and RoBERTa as our victim models due to their reported robustness against perturbations when compared to other models such as LSTM and CNN (Hsieh et al., 2019).

The target BERT model for Chinese sentiment classification is from Huggingface [8], and the target

RoBERTa model for entailment was trained on the training set of the Chinanews dataset (Zhang and LeCun, 2017). The validation performance of the BERT sentiment classification model is 89.80% and is 89.71% for RoBERTa entailment model.

For both models, untargeted classification was set as the fooling goal function and the search method was greedy search with word-importance-ranking as aforementioned. 500 examples were attacked by using each attack method. Two related datasets are as follows.

- Dataset-1: An online shopping review dataset[9] for sentiment classification tasks was used to generate attacks against the BERT classification model, with 500 examples from the test set to check the model's adversarial robustness.
- Dataset-2: The Chinanews dataset was collected by the glyph project (Zhang and LeCun, 2017) and consists of the summary and first paragraphs of news articles from chinanews.com. Each set of summary and first paragraph was labeled with one of 7 topic classes, including mainland China politics, Hong Kong-Macau politics, Taiwan politics, International news, financial news, culture, entertainment, sports, and health. We randomly sampled 500 examples from the test set to attack against the entailment model.

C.3 Discussion of Results

After checking the generated adversarial examples, we realize that a leading cause behind inconsistent and unnatural adversarial examples for Open HowNet transformation is out-of-context substitutions, supported by it having the highest attack success rate yet the lowest consistency/fluency score. Most models are sufficiently robust to attacks with common synonyms, which means successful attacks are often accomplished by distant and unconventional synonym substitutions. On the other hand, cases of out-of-context word substitutions were observed less often in the other attack methods. This is reasonable as homonym and morphonym attack methods only perturb the presentation of the substituted words without changing its semantics to human, while a classification and entailment models fail to attend to the context. However, in rare cases, homonym transformations are also prone to out-of-context substitutions as some Chi-

---

[8] https://huggingface.co/Raychanan/bert-base-chinese-FineTuned-Binary-Best

[9] https://github.com/SophonPlus/ChineseNlpCorpus/tree/master/datasets/online_shopping_10_cats

nese characters have multiple pronunciations. In such scenarios, homonym attacks may result in a false successful attack due to failures to recognize the correct pronunciation and provide an appropriate substitution.

Furthermore, we also observe that perturbing certain characters results in almost guaranteed change in prediction, which was first reported by Wang et al. (2020). For instance, the Chinese character "bu" translates to "no" in English. As illustrated by the first example in Figure 2d, when "bu" is replaced by its morphonym or homonym, the prediction of the perturbed sentence often changes from negative to positive, as a strong negative cue was replaced by another character that the victim model not yet recognizes. Similarly, in the case of entailment models, when the name of a country/region is substituted with its morphonym or homonym, examples with region-specific labels (Hong kong-macau politic, Mainland china politics, etc.) were most often attacked successfully. The vulnerability of Chinese BERT and RoBERTa models against morphonym and homonym adversarial attacks indicates that there is still a large room for improvement in their adversarial robustness.

D Conclusion

In summary, we investigate how to adapt SOTA adversarial attack algorithms to the Chinese language. Our experiments show that the system of generating English adversarial examples can be sufficiently adapted to Chinese, given appropriate text segmentation, perturbation methods, and linguistic constraints. We also introduce two additional perturbation methods particular to the attributes of the Chinese language. Because most of the English/Chinese-specific components of the workflow can be substituted with other languages and resources, we are optimistic that the adaptation workflow presented in this paper can be generalized to other languages in building a language-agnostic attack algorithm in future research.

E Qualitative Examples

1. Adversarial Examples Generated by the OpenHownet Attack Method

Input Text, x: → Negative (95%)

屏幕色彩非常糟糕，不建议购买。真实评价 。
The color of the screen is very terrible, do not recommend purchase. True comment.

Perturbed Text, x': → Positive (85%)

屏幕色彩非常下乘，不建议购买。真实评价 。
The color of the screen is very mediocre, do not recommend purchase. True comment.

- -

Input Text, x: → Negative (83%)

包装很简陋，好失望的购物，一星是给快递小哥的，送货快，态度好。
The packaging was shabby, a disappointing shopping experience, the only star is for the delivery guy, he delivers fast with good attitude.

Perturbed Text, x': → Positive (56%)

包装很简陋，好万念俱灰的购物，一星是给快递小哥的，送货快，态度好。
The packaging was shabby, a depressing shopping experience, the only star is for the delivery guy, he delivers fast with good attitude.

(a) Adversarial examples of Open HowNet

2. Adversarial Examples Generated by the MLM Attack Method

Input Text, x: → Positive (86%)

手感冷冰冰的，除了小点好像没问题，蛮好的。
It feels cold when holding in hands, beside being a little small there is no problem, pretty good.

Perturbed Text, x': → Negative (51%)

手感冷冰冰的，除了小点的没啥，蛮好的。
It feels cold when holding in hands, beside being a little small there is nothing wrong, pretty good.

- -

Input Text, x: → Negative (86%)

没有明面的插座，需要把台灯下面的插座拔下来。没有Wi-Fi。为了赶早班飞机，凑合住住。
There is no accessible socket , need to unplug the socket under the lamp. There is no WiFi. Only staying to catch early flight in the morning.

Perturbed Text, x': → Positive (62%)

没有明面的插座，需要把台灯下面的插座拔下来。使用Wi-Fi。为了赶早点到达，凑合住住。
There is no accessible socket , need to unplug the socket under the lamp. There is no WiFi. Only staying to arrive earlier.

(b) Adversarial examples of MLM

3. Adversarial Examples Generated by the Homonym Attack Method

Input Text, x: → Negative (99%)

不好，有味非常刺鼻，而且还慢！
Not good, the scent is pungent, and the delivery is slow!

Perturbed Text, x': → Positive (89%)

部好，有味非常词鼻，而且还慢！
"Department good", the scent is "word nose", and the delivery is slow!

- -

Input Text, x: → Negative (95%)

容易死机通话声音小功能少。
The phone system is prone to crash, sound is weak and there aren't many functions.

Perturbed Text, x': → Positive (84%)

容易四机通话声音小功能少。
The phone system is prone to "four machine", sound is weak and there aren't many functions.

(c) Adversarial examples of Homonym

4. Adversarial Examples Generated by the Morphonym Attack Method

Input Text, x: → Positive (97%)

优点忒多了，不用多介绍了。
The product is amazing, there's no need to further introduction.

Perturbed Text, x': → Negative (61%)

优点忒多了，否用多价绍了。
The product is amazing, there's rejecting need to further "price introduction".

- -

Input Text, x: → Positive (54%)

看样子可以，没有苹果这么红，没吃不知道好吃不 。
The appearance is good, not as red as apples are, haven't tasted it to know whether it tastes good.

Perturbed Text, x': → Negative (72%)

看样子可以，没有苹果这么红，投吃不知道好吃不。
The appearance is good, not as red as apples are, "throw" tasted it to know whether it tastes good.

(d) Adversarial examples of Morphonym

Figure 2: Selected Adversarial Examples generated by proposed adversarial attacks on the online shopping review dataset (classification). Note for adversarial examples generated by the Homonym and Morphonym attack method (figure 2c and 2d), word substitutions are based on Chinese language characteristics instead of semantic meaning. For examples in figure 2c, substitutions were chosen from characters with similar sounds. For figure 2d, substitutions were from characters that look similar to human readers.

1. Adversarial Examples Generated by the OpenHownet Attack Method

Input Text, x: → Entertainment (67%)

日本动画导演芦田丰雄去世，曾执导《北斗神拳》。
Japanese anime director Toyoo Ashida passed away, who directed the show *Fist of the North Star*.

Perturbed Text, x': → Culture (60%)

日本动画领奏芦田丰雄去世，曾执导《北斗神拳》。
Japanese anime first chair (in band) Toyoo Ashida passed away, who directed the show *Fist of the North Star*.

Input Text, x: → Financial News (66%)

中国铁路总公司上半年负债3.4万亿元，亏损53亿。
Chinese Railway holds debts of 3.4 trillion in the first half of this year, with a gross loss of 5.3 billion.

Perturbed Text, x': → Mainland China Politics (51%)

中国铁路总公司上半年该欠3.4万亿元，亏损53亿。
Chinese Railway supposedly owes 3.4 trillion in the first half of this year, with a gross loss of 5.3 billion.

(a) Adversarial examples of Open HowNet

2. Adversarial Examples Generated by the MLM Attack Method

Input Text, x: → Sports (100%)

VOLVO中国公开赛群星荟萃，奖金提到2000万。
The Chinese Open, hosted by Volvo, was attended by various talents and public figures, with the prize raised to 20 million.

Perturbed Text, x': → Entertainment (92%)

VOLVO中国展群星荟萃，奖金提到2000万。
The Chinese Show, hosted by Volvo, was attended by various talents and public figures, with the prize raised to 20 million.

Input Text, x: → Entertainment (95%)

《私人订制》浮云场景被指仿蔡志松作品。
The "floating cloud" scene in the movie *Personal Tailor* was criticized for its similarity to Zhisong Cai's work.

Perturbed Text, x': → Culture (99%)

《私人订制》浮云画被指仿蔡志松作品。
The "floating cloud" drawing in the movie *Personal Tailor* was criticized for its similarity to Zhisong Cai's work.

(b) Adversarial examples of MLM

3. Adversarial Examples Generated by the Homonym Attack Method

Input Text, x: → Hong kong - macau politics (95%)

奥运冠军结束访澳行程，市民相送依依不舍。
Olympic champion ends trip in Australia, was welcomed by city residents.

Perturbed Text, x': → Sports (86%)

奥运冠军结束访熬行程，市民相送依依不舍。
Olympic champion ends trip in "persistence", was welcomed by city residents.

Input Text, x: → Culture (52%)

年要过完了，"大年精神"应该留下。
New Year is always over, but the "spirit of the new year" should remain.

Perturbed Text, x': → Sports (59%)

年要过完了，"大年精神"赢该留下。
New Year is always over, but the "spirit of the new year" win should remain.

(c) Adversarial examples of Homonym

4. Adversarial Examples Generated by the Morphonym Attack Method

Input Text, x: → Mainland china politics (91%)

人民日报：厉行节俭，人人有责。
China Daily: Promoting frugality with force is everyone's responsibility.

Perturbed Text, x': → Culture (88%)

人民日报：历行节捡，人人有责。
China Daily: Promoting "saving fetch" regularly is everyone's responsibility.

Input Text, x: → Financial news (84%)

韩国彩票连续两年超销售上限。
Korean lottery sales reached supply for the second straight year.

Perturbed Text, x': → Culture (77%)

韩国眯票连续两年超销售上限。
Korean "notice ticket" sales reached supply for the second straight year.

(d) Adversarial examples of Morphonym

Figure 3: Selected Adversarial Examples generated by proposed adversarial attacks on the Chinanews dataset (entailment task). As mentioned in the discussion section, substituting specific characters almost guarantees a change in the prediction result. As shown by the second example in figure 3c, the homonym substitution of the word "should" added the new semantic meaning of "winning", which is a strong cue for the Sports category.