SemAttack: Natural Textual Attacks via Different Semantic Spaces

*Boxin Wang[1], *Chejian Xu[1], Xiangyu Liu[2], Yu Cheng[3], Bo Li[1]
[1]University of Illinois at Urbana-Champaign
[2]Alibaba Group, [3]Microsoft Research
{boxinw2,chejian2,lbo}@illinois.edu
eason.lxy@alibaba-inc.com, yu.cheng@microsoft.com

Abstract

Recent studies show that pre-trained language models (LMs) are vulnerable to textual adversarial attacks. However, existing attack methods either suffer from low attack success rates or fail to search efficiently in the exponentially large perturbation space. We propose an efficient and effective framework SemAttack to generate natural adversarial text by constructing different semantic perturbation functions. In particular, SemAttack optimizes the generated perturbations constrained on generic semantic spaces, including typo space, knowledge space (*e.g.*, WordNet), contextualized semantic space (*e.g.*, the embedding space of BERT clusterings), or the combination of these spaces. Thus, the generated adversarial texts are more semantically close to the original inputs. Extensive experiments reveal that state-of-the-art (SOTA) large-scale LMs (*e.g.,* DeBERTa-v2) and defense strategies (*e.g.,* FreeLB) are still vulnerable to SemAttack. We further demonstrate that SemAttack is general and able to generate natural adversarial texts for different languages (*e.g.,* English and Chinese) with high attack success rates. Human evaluations also confirm that our generated adversarial texts are natural and barely affect human performance. Our code is publicly available at https://github.com/AI-secure/SemAttack.

1 Introduction

Deep neural networks have achieved remarkable success in many machine learning tasks. Particularly, BERT (Devlin et al., 2019) has inspired a suite of large-scale pre-trained language models (Yang et al., 2019; Zhang et al., 2019; Lan et al., 2019), which achieved new SOTA for many NLP tasks. In addition to BERT's dominant performance on English datasets, Tenney et al. (2019) points out that BERT is similarly effective on other languages such as Chinese, whose granularity of words is more complex, given the model's ability to disambiguate information from high-level representations (Ding et al., 2019).

Original Input: They need to hire experienced sales rep who are mature enough to handle questions and sales. **Adversarial Input:** They need to hire skilled sales rep who are mature enough to handle questions and sales. **Sentiment Prediction:** Most Negative → Most Positive
Original Input: 拿什么能吸引你：我们的海外学子? (**Translation:** What can attract you: our overseas students?) **Adversarial Input:** 拿甚么能吸引你：我们的海外学子? (**Translation:** What can attract you: our overseas students?) **Topic Prediction:** Education News → Entertainment News

Table 1: Adversarial texts generated against English and Chinese BERT classifiers by SemAttack on Yelp and THUCTC datasets. Replacing a word/character with an adversarial one misleads the correct prediction to a wrong class without fooling human.

Although effective for many NLP tasks, the robustness of these neural models is often challenged by carefully crafted adversarial examples. Specifically, attackers can add subtle human-imperceptible perturbation to the original input and induce dramatic changes in model output. Current adversarial text generation (Jia and Liang, 2017; Li et al., 2018; Alzantot et al., 2018) is mainly heuristic and only achieves low attack success rates for BERT-based models. Other work (Cheng et al., 2020; Ebrahimi et al., 2018) allows an input word to be substituted by any other word in the vocabulary, which fails to consider the semantic perturbation constraints and is prone to invalid adversarial examples. Recent work (Jin et al., 2020; Zang et al., 2020) relies on external knowledge to constrain the perturbation yet poorly handles large search space that grows exponentially with the input length, as it requires hundreds of queries to generate one adversarial example in practice.

Furthermore, most existing textual adversarial attacks are not generalizable to other languages, due to unique language-dependent characteristics

*Equal Contribution

Findings of the Association for Computational Linguistics: NAACL 2022, pages 176 - 205
July 10-15, 2022

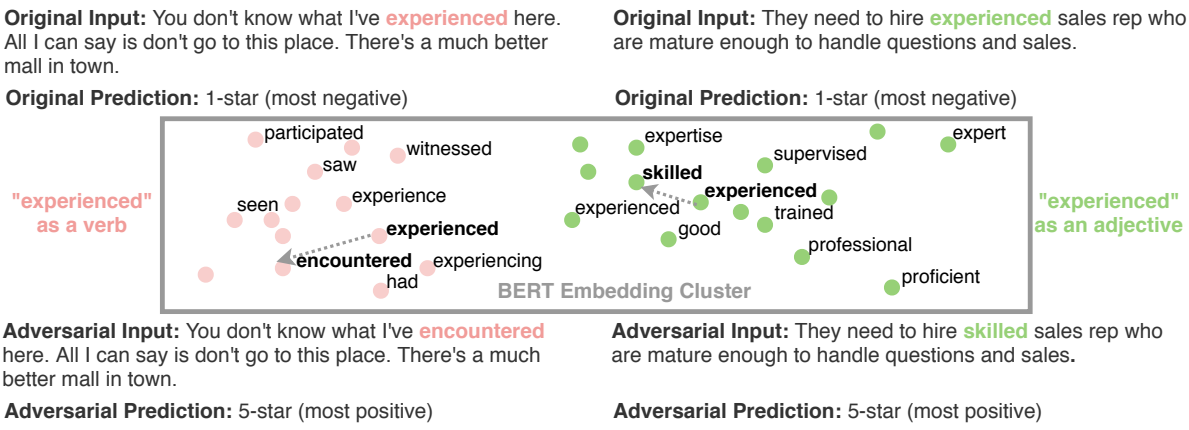


Figure 1: Adversarial texts against BERT sentiment classifier generated by SemAttack that formulates two different contextualized semantic perturbation spaces based on BERT embedding clusters (the embedding space is projected by PCA onto 2D space). The word "experienced" reveals different meanings (past tense of the verb "experience" or adjective form) in different contexts (clusters). Our contextualized semantic perturbation chooses "saw" or "encountered" as the perturbation for verb "experienced", while "skilled" or "trained" for the adjective form.

and the lack of universal linguistic resources. Moreover, character-level adversarial attacks designed in English context (Ebrahimi et al., 2017) are often ineffective for Chinese-character-level attacks, as the size of candidate characters increases by two orders of magnitude, resulting in surging computational costs especially for BERT-based models.

We tackle these limitations in textual adversarial attacks by proposing an effective and efficient framework SemAttack, which can be used to further evaluate the robustness of NLP models. We generalize existing word-level attacks and propose generic semantic perturbation functions, which optimize and constrain the perturbations within different semantic spaces, so that the generated adversarial texts retain their semantic meaning. We mainly consider three types of semantic spaces: (1) *Typo Space*, using typo words or characters that can fool the models but not human judges; (2) *Knowledge Space*, utilizing external linguistic knowledge base (*e.g.*, WordNet (Miller, 1995)) as valid perturbation candidates; and (3) *Contextualized Semantic Space*, exploiting the embedding space of BERT to generate a contextualized perturbation set semantically close to the original word (Figure 1). The contextualized semantic space does not require additional knowledge, and therefore can scale to other languages, especially low-resource languages where a large knowledge base is unavailable.

After the candidate semantic space is determined, SemAttack searches for the optimal perturbation combination. Instead of requiring thousands of queries to generate one adversarial example, optimal perturbations can be efficiently found in the embedding space by solving an optimization problem. We also control the magnitude of perturbation to be small as shown in Table 1. Extensive experiments on four datasets demonstrate that SOTA LMs and defense methods are still vulnerable to our adversarial attack, which are natural and barely affects human judgment. For example, the accuracy of BERT sentiment classifier drops from 70.6% to 2.4% by simply replacing fewer than 5% words with our method. Although these adversarial examples are generated in the whitebox setting, they can effectively transfer to two different blackbox attack settings while retaining higher than 90% attack success rate for BERT and other large-scale LMs such as DeBERTa-XXLarge.

Our contributions are summarized as follows: 1) We propose a unified and effective adversarial attack framework SemAttack by constructing semantic perturbation functions, which constraint perturbations within different semantic spaces and their combinations. 2) SemAttack generates contextualized perturbations that require no external knowledge and thus can easily adapt to different languages. 3) We conducted extensive experiments on different datasets and languages to show that adversarial texts generated by SemAttack are more semantically close to the benign inputs, and achieve much higher attack success rates than existing attack algorithms in different settings. 4) Comprehensive studies demonstrate that SOTA LMs and defenses are still vulnerable to SemAttack, and human evaluation verifies the naturalness and va-

lidity of our adversarial examples.

2 SemAttack

2.1 Problem Formulation

Given an input $\boldsymbol{x} = [x_0, x_1, ..., x_n]$, where x_i is the i-th input token, the classifier f maps the input to final logits $\boldsymbol{z} = f(\boldsymbol{x}) \in \mathbb{R}^C$, where C is the number of classes, and outputs a label $y = \arg\max f(\boldsymbol{x})$.

During attack, we evaluate the effectiveness of attack algorithms by calculating the targeted attack success rate (TSR):

$$\text{TSR} = \frac{1}{|D_{\text{adv}}|} \sum_{\boldsymbol{x'} \in D_{\text{adv}}} \mathbb{1}[\arg\max f(\boldsymbol{x'}) \equiv y^*] \quad (1)$$

and untargeted attack success rate (USR):

$$\text{USR} = \frac{1}{|D_{\text{adv}}|} \sum_{\boldsymbol{x'} \in D_{\text{adv}}} \mathbb{1}[\arg\max f(\boldsymbol{x'}) \neq y] \quad (2)$$

where the attack algorithm generates one adversarial sentence for each sample to form an adversarial dataset D_{adv} , y^* is the targeted false class, y is the ground truth label, and $\mathbb{1}(\cdot)$ is the indicator function.

2.2 Semantic Perturbation Functions

To control adversarial examples to be semantically close to the original input, we design a general form of semantic perturbation function $\mathcal{F}$, which takes one token x as input, and returns its candidate perturbation space $\mathcal{S} = \{x_0^*, x_1^*, ..., x_n^*\}$. We next discuss the types of perturbation function $\mathcal{F}$.

Typo-based Perturbation Function $\mathcal{F}_T$ constrain the search space $\mathcal{S}$ in the *typo space*, which uses typo words or characters to replace original tokens so that human can still understand the original meaning while models are fooled. In English, we follow the generation process introduced in TextBugger (Li et al., 2018) to generate typos.

In order to illustrate how our proposed method can be easily adapted to multilingual settings, we also generate typo-based semantic space for Chinese. Specifically, for each Chinese token x, we prepare a set of common Chinese characters $\mathcal{S}$ that look similar ("形近字") or have the same pronunciation ("音近字") as the original token x. We use the open-source similar Chinese character list that contains more than 9,000 common Chinese characters. To search for the Chinese characters with the same pronunciation (*i.e.,* pinyin), we first query the pronunciation of input x and then choose the characters returned based on the same pronunciation. If x is a heteronym that has multiple pronunciations, we only use one pronunciation to do the query. We also limit the size of Chinese characters of the same pronunciation to be less than 6 so that the search space is not too large. For the Chinese example shown in Table 1, we use "甚" to replace "什" as they share the same pronunciation and are a common typo that will not affect human understanding.

Knowledge-based Perturbation Function $\mathcal{F}_K$ considers the *knowledge space* to constrain the perturbation search space $\mathcal{S}$. Specifically, $\mathcal{F}_K$ utilizes existing knowledge base to build a candidate perturbation set. In our work, we use WordNet as an example to illustrate how our framework can integrate rule-based knowledge to enhance the quality of adversarial examples. WordNet is a large lexical dataset of more than 200 languages that groups words into sets of cognitive synonyms. With the manually labeled semantic relations among words, synonyms queried from WordNet (*i.e.*, synsets) share the same semantic meaning as the query word x. Therefore we choose these synonyms returned from WordNet to be the search space $\mathcal{S}$. We note that WordNet also contains hypernyms and hyponyms information, but including them into the search space may incur some unnatural replacement (*e.g.*, replacing "fifth" with "rank"). Therefore, we only consider synsets as the candidate search space $\mathcal{S}$. In addition, even for the same token (*e.g.*, "use") in WordNet, it may have different part-of-speech (POS) tags (*e.g.*, "use" as verb or as noun), and thus has different synonyms (*e.g.,* "exploitation" for noun "use" and "practice" as verb "use"), which may result in nonsensical replacement. In order not to include synonyms that have unusual part of speech, we counted the frequency of POS in the synset and only selected the words with the most frequent POS. Using the synonym set $\mathcal{S}$ after filtering, we are able to generate adversarial input texts that mislead models' prediction while barely affect on human understanding.

Contextualized Semantic Perturbation Function $\mathcal{F}_C$ is a novel perturbation function that explores the BERT embedding space and searches for contextualized perturbation to tackle the issue of most language tokens being polysemous. Previous work (Li et al., 2018; Jin et al., 2020) takes it as a standard practice to use the proximity in embedding space to query the semantic similarity. However, their embedding space is built on a non-contextualized word embedding from GLoVE (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013), thus failing to consider the polysemy when

generating the perturbation. We propose to explore the BERT embedding space, which is verified by (Hewitt and Manning, 2019; Coenen et al., 2019; Papadimitriou et al., 2021) that BERT embeddings can preserve syntactic and semantic information for word sense disambiguation better than GLoVE or Word2Vec. So the contextualized space from $\mathcal{F}_C$ is valid semantic perturbations. Similar to our parallel work (Li et al., 2020) of using BERT to generate adversarial perturbations, $\mathcal{F}_C$ also does not require external linguistic resources such as POS checker. Thus $\mathcal{F}_C$ can be adapted to other languages, as long as pre-trained BERT of such language models is available.

Specifically, we first choose a set of commonly used tokens $\mathcal{X}$. For each word $x \in \mathcal{X}$, we select at most 100 example sentences from Wikipedia that contain the word x so that these sentences represent different meanings of x in different contexts. We then feed these sentences into a pretrained BERT model to obtain the contextualized embeddings for each word x. Finally, the contextualized embeddings for all words in $\mathcal{X}$ formulate a large BERT embedding space. Figure 1 visualizes a BERT embedding space projected into 2D space by PCA.

To query the search space $\mathcal{S}$ for token x, we first calculate the BERT embedding of token x given its context sentence. Even for the same token, given different contexts and meanings, BERT will generate distinct representations in the high dimensional embedding space. For the example in Figure 1, the token "experienced" given different contexts have different latent representations and neighbors. Then we use k nearest neighbors (KNN) algorithm to choose the neighbors of the contextualized embedding of x as its perturbation search space $\mathcal{S}$. To ensure high quality of search space $\mathcal{S}$, we further filter $\mathcal{S}$ and only return the words that appear more frequently than a threshold ϵ among k nearest neighbors. In this way, we remove the noisy tokens that are rarely used and retain the high-quality neighbor tokens whose contextualized semantics are mostly close to the original token x.

Discussion. The final search space $\mathcal{S}$ can be the union of the search spaces mentioned above. This makes existing defense algorithm (Jones et al., 2020) difficult to apply, as they can only defend against typo-based perturbation but fails to detect other types of perturbation.

$\mathcal{F}$ is a generalization of most existing word-level textual adversarial attacks. Though $\mathcal{F}_T$ and $\mathcal{F}_K$ have been discussed in the previous literature (see §Related Work), we note that the goal of our paper is not to improve or propose better typo or knowledge perturbation, but to consider multiple semantic spaces at the same time to help generate natural high-quality adversarial examples.

2.3 Attack Algorithm `SemAttack`

The full pipeline is shown in Appendix Algorithm 1. Essentially, `SemAttack` searches for the optimal perturbations from different semantic spaces determined by semantic perturbation functions, which is efficiently solved as an optimization problem so that we only perturb as few tokens as possible while achieving the targeted attack.

Unlike generating adversarial examples in the continuous data domain, it is difficult to directly utilize the gradient to guide token substitution due to the discrete nature of text. Thus, we search perturbation in the embedding space and map the perturbed embedding back to tokens. Specifically, the one-hot representation of each discrete token $\boldsymbol{x_i} \in \mathbb{R}^{|V|}$ (V is the vocabulary set) is mapped into an embedding space of dimension d_c via the embedding matrix $\boldsymbol{M_e} \in \mathbb{R}^{d_c \times |V|}$

$$[\boldsymbol{e_1}; \boldsymbol{e_2}; ...; \boldsymbol{e_n}] = \boldsymbol{M_e}\Big[\boldsymbol{x_0}; \boldsymbol{x_1}; ...; \boldsymbol{x_n}\Big]. \quad (3)$$

We optimize perturbation $\boldsymbol{e^*}$ added to the original embedding $\boldsymbol{e}$ for m iterations. In each iteration, we freeze all the parameters of the classifier f and optimize variable $\boldsymbol{e^*}$ only. Following Carlini and Wagner (2016), we minimize the loss function as:

$$\mathcal{L}(\boldsymbol{e^*}) = ||\boldsymbol{e^*}||_p + c \cdot g(\boldsymbol{x'}), \quad (4)$$

where the first term controls the magnitude of perturbation, while $g(\cdot)$ is the attack objective function depending on the attack scenario. c weighs the attack goal against attack cost.

In *targeted attack* scenarios, we define $g(\cdot)$ as:

$$g(\boldsymbol{x'}) = \max[\max\{f(\boldsymbol{x'})_i : i \neq t\} - f(\boldsymbol{x'})_t, -\kappa],$$

where t is the targeted false class and $f(\boldsymbol{x'})_i$ is the i-th class logit. A larger κ encourages the classifier to output targeted false class with higher confidence.

In *untargeted attack* scenarios, $g(\cdot)$ becomes

$$g(\boldsymbol{x'}) = \max[f(\boldsymbol{x'})_t - \max\{f(\boldsymbol{x'})_i : i \neq t\}, -\kappa],$$

where t is the ground truth class.

After each iteration of gradient descent, we have an optimized perturbation $\boldsymbol{e^*}$ in the embedding space that tends to fool the classifier f with small perturbations. We choose the perturbed token $\boldsymbol{x'_i} \in \mathcal{S} = \mathcal{F}(x_i)$ that is from the semantic search space $\mathcal{S}$ returned by $\mathcal{F}(x_i)$ and semantically closest to

the perturbed embedding $\boldsymbol{e}'_i$.

$$\begin{aligned} \boldsymbol{e}'_i &= \boldsymbol{e}_i + \boldsymbol{e}^*_i, \\ \boldsymbol{x}'_i &= \underset{\boldsymbol{x}'_i \in \mathcal{S}}{\arg\min}(||\boldsymbol{e}'_i - \boldsymbol{M}_e \boldsymbol{x}'_i||_p). \end{aligned} \tag{5}$$

Finally, we obtain an optimal perturbation $\boldsymbol{e}^*$ after repeating the optimization step and token substitution step for m iterations. Under such settings and constraints, most tokens remain the same and very few are perturbed to their semantically close neighbors. Thus, the adversarial examples still look valid to humans but can fool the models.

3 Experimental Results

In this section, we conduct comprehensive experiments to evaluate our attack method in various settings. We **first** apply our attack method to two standard NLP models, BERT and SOTA Self-Attention LSTM. We evaluate on *two different types of NLP tasks*, sentiment analysis and natural language inference (NLI). **Secondly**, we investigate the effectiveness of `SemAttack` against SOTA large-scale language models and defense methods. **Thirdly**, we take Chinese as an example to measure `SemAttack`'s generalization ability across different languages. We evaluate BERT models finetuned on two Chinese datasets. **Finally**, we conduct extensive human evaluations on both English and Chinese datasets.

We find that: 1) `SemAttack` can achieve better attack success rates than existing textual adversarial attack methods with better language quality and adversarial transferability. 2) SOTA LMs and defense methods are still vulnerable to our `SemAttack`. 3) `SemAttack` is a general textual adversarial attack framework and can be easily adapted to other languages in addition to English with high attack success rates. 4) Adversarial examples generated by `SemAttack` are natural and barely affect human performance.

3.1 Whitebox and Blackbox Attack

Datasets For sentiment classification task, we choose the standard 5-class sentiment classification dataset, Yelp dataset. Note that unlike previous work (Li et al., 2020; Jin et al., 2020) that uses binary sentiment classification dataset, we focus on the standard 5-class Yelp dataset to further evaluate the **targeted attack capability** of `SemAttack`. For NLI task, we choose SNLI dataset. The detailed dataset descriptions are in Appendix §C.

Models We evaluate the robustness of *BERT* and *Self-Attention LSTM* (Lin et al., 2017). We present their test accuracy on the benign test sets in Table 2. More hyperparameter settings and training details are discussed in Appendix §B.

Attack Baselines We consider SOTA whitebox and blackbox attack baselines.

• **HotFlip** (Ebrahimi et al., 2017) is a whitebox attack method for generating adversarial examples on both character-level and word-level. In terms of preserving semantic meaning, we only use word-level attacks in our experiments, which uses gradient-based optimization method to flip words.

• **TextFooler** (Jin et al., 2020) is a blackbox attack method for generating adversarial text, which uses similarities between pre-calculated word embeddings to find synonyms for each word.

• **BERT-Attack** (Li et al., 2020) is a strong blackbox attack method using pre-trained masked language models such as BERT to replace words in input sentences, where pre-trained masked language models provide candidate words that have high semantic similarity between original texts.

These methods all perform untargeted attacks. We adapt them to both untargeted and targeted attack settings in our experiments.

Attack Goal In the sentiment analysis task, we consider the **targeted attack**, and choose the most opposite sentiment class as the targeted class, so sentences with original label lower than 2 (*negative*) are attacked to class 4 (*most positive*), and others are attacked to class 0 (*most negative*). In the NLI task, *Contradiction* and *neutral* will be attacked to *entailment* while *entailment* will be attacked to *contradiction*.

Adversarial Attack Evaluation We perform `SemAttack` on BERT and LSTM-based classifiers in both the whitebox and blackbox settings. The whitebox setting approximates the worst-case scenario, where attackers have the access to the model parameters and gradients; while the blackbox setting assumes that attackers can only access the model's output confidence.

For the **whitebox attack** shown in Table 2, `SemAttack` can outperform all the SOTA baselines and achieve the highest success rates in both untargeted and targeted settings for BERT and LSTM-based models with smaller or comparable perturbation rates. For example, untargeted `SemAttack` achieves 97.6% attack success rate

Model	Attack Method	% USR/TSR	% Perturbation
BERT (Acc: 0.706)	HotFlip	71.5/24.0	14.9/44.9
	SemAttack (+$\mathcal{F}_T$)	42.4/9.3	4.7/9.1
	SemAttack (+$\mathcal{F}_K$)	84.6/69.3	6.7/13.9
	SemAttack (+$\mathcal{F}_C$)	91.3/79.7	4.7/11.1
	SemAttack (+all)	**97.6/93.8**	4.3/10.2
Self-Attention LSTM (Acc: 0.705)	HotFlip	16.3/3.2	2.5/17.4
	SemAttack (+$\mathcal{F}_T$)	67.2/49.4	14.7/21.1
	SemAttack (+$\mathcal{F}_K$)	47.9/43.6	10.4/18.3
	SemAttack (+$\mathcal{F}_C$)	67.3/56.5	15.1/23.2
	SemAttack (+all)	**88.1/84.0**	19.2/29.2

(a) Yelp Dataset

Model	Attack Method	% USR/TSR	% Perturbation
BERT (Acc: 0.829)	HotFlip	83.3/44.9	27.0/30.3
	SemAttack (+$\mathcal{F}_T$)	21.2/10.2	13.1/16.5
	SemAttack (+$\mathcal{F}_K$)	53.8/23.2	14.8/22.3
	SemAttack (+$\mathcal{F}_C$)	90.2/69.7	15.3/26.9
	SemAttack (+all)	**92.6/72.6**	15.6/20.0
Self-Attention LSTM (Acc: 0.705)	HotFlip	32.3/17.8	11.6/13.4
	SemAttack (+$\mathcal{F}_T$)	53.8/33.4	23.9/29.1
	SemAttack (+$\mathcal{F}_K$)	40.7/23.2	21.4/22.2
	SemAttack (+$\mathcal{F}_C$)	76.5/63.8	30.9/36.3
	SemAttack (+all)	**86.2/68.5**	39.0/36.9

(b) SNLI Dataset

Table 2: Whitebox attack success rate for different attacks under targeted/untargeted attacks (TSR/USR) and corresponding word perturbation percentage against self-attention LSTM and BERT on Yelp and SNLI datasets.

Model	Attack Method	% USR/TSR	% Perturbation
DeBERTa (Large, Acc: 0.928)	TextFooler	83.2/57.1	22.5/21.3
	BERT-ATTACK	84.4/36.6	19.4/17.9
	SemAttack (+$\mathcal{F}_T$)	**88.1/58.3**	17.8/16.0
	SemAttack (+$\mathcal{F}_K$)	82.1/53.7	22.1/20.9
	SemAttack (+$\mathcal{F}_C$)	80.3/33.6	27.6/27.7
	SemAttack (+all)	83.0/41.2	21.4/20.5
DeBERTa (XXLarge-v2, Acc: 0.931)	TextFooler	86.4/57.1	22.1/20.3
	BERT-ATTACK	83.4/37.2	19.2/17.8
	SemAttack (+$\mathcal{F}_T$)	**90.5/65.5**	17.6/16.2
	SemAttack (+$\mathcal{F}_K$)	86.8/58.4	22.3/21.7
	SemAttack (+$\mathcal{F}_C$)	80.6/38.7	27.6/27.9
	SemAttack (+all)	82.7/42.9	21.2/20.2
FreeLB (Acc: 0.924)	TextFooler	63.0/31.5	22.1/22.0
	BERT-ATTACK	65.6/31.1	19.1/18.6
	SemAttack (+$\mathcal{F}_T$)	**71.4**/26.2	17.0/14.7
	SemAttack (+$\mathcal{F}_K$)	63.2/32.6	22.9/23.9
	SemAttack (+$\mathcal{F}_C$)	66.7/**32.7**	27.8/28.0
	SemAttack (+all)	64.3/32.2	20.9/20.5

Table 3: Zero-query blackbox attack success rate for different attacks under targeted/untargeted attacks (TSR/USR) and corresponding word perturbation percentage against large-scale LMs and defense methods on SNLI datasets.

for BERT models by perturbing 4% words on the Yelp dataset, when searching from the combination of the semantic spaces of $\mathcal{F}_T$, $\mathcal{F}_K$ and $\mathcal{F}_C$.

To adapt SemAttack to the blackbox attack setting, we distill the blackbox (teacher) model to train a whitebox (student) model, and *transfer* the adversarial examples from the whitebox student model to attack the blackbox model. More details can be found in Appendix §D.

For the **blackbox attack** shown in Appendix Table 8, the transferability-based SemAttack achieves higher attack success rates than SOTA blackbox attacks for self-attention LSTM. We also observe that BERT-ATTACK achieves a higher attack success rate on BERT than SemAttack. We think it is mainly because that BERT-ATTACK adopts an aggressively large candidate perturbation size (top-k=48), which may lead to large semantic changes (indicated by the worse human performance as shown in Table 5). For instance, we observe that some words are even changed to their antonym in BERT-ATTACK. On the contrary, the average size of search spaces for SemAttack (+all) is only 11.87, aiming to guarantee the naturalness and validity of the generated adversarial examples. We present more details of our semantic space in Appendix §D.3.

In addition, we observe that Self-Attention LSTM models are more robust than BERT in most settings. For example, we achieve the highest USR of 88.1% in whitebox attack on the Yelp dataset, which is 9.5% lower than BERT in the same setting. This suggests that self-attention mechanism can improve the robustness of vanilla WordLSTM by a large margin, as WordLSTM is known less robust than BERT (Jin et al., 2020).

3.2 Attack SOTA LMs and Defense Methods

In this section, we evaluate SemAttack and baseline attacks against various SOTA large-scale language models and defense methods.

Dataset and Attack Baselines Following §3.1, we evaluate SemAttack on SNLI dataset. We choose the same blackbox attack methods, TextFooler and BERT-Attack, as our baselines.

Models We consider the following models and defense methods following the Adversarial GLUE Benchmark (Wang et al., 2021). The selected large-scale models and defense methods not only represent SOTA performance on NLU tasks, but also achieve the highest robustness in the leaderboard.

- **DeBERTa (He et al., 2020)** improves BERT-based models by introducing disentangled attention mechanism and enhanced mask decoder, which is one of the best models in the GLUE leaderboard (Wang et al., 2018). In our experiment, we use DeBERTa (Large) and DeBERTa (XXLarge-v2).
- **FreeLB (Zhu et al., 2019)** is an adversarial training algorithm that defends adversarial attacks by adding perturbations to word embeddings and

minimizing the corresponding adversarial loss.

Attack Goal To demonstrate the model robustness in an approximately real-world scenario, we consider a **zero-query setting**, a more rigorous and common scenario that assumes the target models are not accessible during the attack phase. Since we can not access the target model, we perform a transferability-based backbox attack. Specifically, we attack the selected language models and defense methods using adversarial SNLI texts generated by `SemAttack` against BERT classifier in §3.1.

Adversarial Attack Evaluation We finetune the above models on the SNLI dataset and attack them using adversarial texts generated against BERT. The results are shown in Table 3.

For the **zero-query setting**, `SemAttack` always achieves the highest success rates. Specifically, among all the attack methods, `SemAttack` (+$\mathcal{F}_T$) always has the highest USR regardless of the model it is tested on. For example, on the largest model, DeBERTa (XXLarge-v2), we achieve 90.5% USR, which is 7.1% higher than BERT-ATTACK.

Furthermore, we find that *increasing the number of model parameters and expanding the model architecture have little effect on defense against adversarial attacks*. DeBERTa (XXLarge-v2), for example, is substantially larger than DeBERTa (Large), yet the attack success rates are similar. In some cases DeBERTa (XXLarge-v2) is even less robust than DeBERTa (Large). We also observe that introducing some defense strategies slightly improves the model's robustness. When we use the defense strategy of FreeLB, we can see that the robustness increases, but it is still not satisfactory to defend existing adversarial attacks.

3.3 Adapt `SemAttack` to Chinese

Datasets We evaluate our performance on the following two datasets in Chinese: 14-category news classification dataset THUNews (Sun et al., 2016) and 11-class Wechat Finance dataset. More details about these datasets are introduced in Appendix C.

Models We use BERT pre-trained on Chinese corpora and finetune on the two datasets separately. After finetuning, our BERT achieved 0.818 accuracy on THUNews dataset and 0.891 on Wechat Finance Dataset, as shown in Table 4

Attack Baselines Since both TextFooler and BERT-Attack adopt an aggressively large perturbation candidate space and thus require additional language resources (e.g., POS checker; stop words filtering) to ensure the proposed candidate words are valid, they cannot be adapted to Chinese due to the lack of corresponding language resources. Therefore, we adapt **HotFlip** for Chinese classification task, since it does not rely on any other linguistic resources. We also adapt it to transferability-based blackbox attack settings as well as the targeted attack setting for fair comparison.

Dataset	Setting	Attack Method	% USR/TSR	% Perturbation
THUNews (Acc: 0.818)	White-box Attack	HotFlip	81.4/40.4	21.7/27.9
		`SemAttack` (+$\mathcal{F}_T$)	96.6/81.7	20.1/34.7
		`SemAttack` (+$\mathcal{F}_K$)	15.6/3.6	16.1/17.4
		`SemAttack` (+$\mathcal{F}_C$)	95.0/78.3	17.4/29.4
		`SemAttack` (+all)	**99.0/92.1**	15.1/26.3
	Black-box Attack	HotFlip	44.3/10.0	15.4/10.8
		`SemAttack` (+$\mathcal{F}_T$)	52.3/34.0	19.7/35.3
		`SemAttack` (+$\mathcal{F}_K$)	8.4/1.3	12.7/13.1
		`SemAttack` (+$\mathcal{F}_C$)	55.9/37.0	17.6/28.6
		`SemAttack` (+all)	**58.6/48.2**	16.4/25.8
Wechat (Acc: 0.891)	White-box Attack	HotFlip	95.2/0.0	11.4/-
		`SemAttack` (+$\mathcal{F}_T$)	86.0/88.3	7.2/12.4
		`SemAttack` (+$\mathcal{F}_K$)	32.8/24.5	5.2/7.6
		`SemAttack` (+$\mathcal{F}_C$)	96.8/96.4	5.8/9.4
		`SemAttack` (+all)	**98.7/98.0**	4.6/8.7
	Black-box Attack	HotFlip	21.7/0.0	8.9/-
		`SemAttack` (+$\mathcal{F}_T$)	49.4/35.8	7.3/17.4
		`SemAttack` (+$\mathcal{F}_K$)	19.5/11.7	4.0/7.7
		`SemAttack` (+$\mathcal{F}_C$)	51.8/**42.4**	5.3/12.2
		`SemAttack` (+all)	**54.5**/36.7	4.0/11.7

Table 4: Whitebox and blackbox attack success rate for different attacks under targeted/untargeted attacks (TSR/USR) and corresponding word perturbation percentage against Chinese BERT on THUNews and Wechat Finance datasets.

Attack Goal In this paper, we choose the targeted attack class as "technology news" for THUNews dataset and "Bank" for Wechat dataset (when the ground truth label is the targeted class, we switch the target to another random class). This strategy achieves the highest targeted attack success rate as shown in Appendix F.7.

Adversarial Attack Evaluation In the **whitebox attack** scenario in Table 4, `SemAttack` is able to make the model mistakenly classify nearly all sentences with only a small number of characters being manipulated in both targeted and untargeted settings. The untargeted attack achieves 99% success rate by substituting merely two tokens on average on the THUNews dataset. On Wechat Finance dataset, it achieves 98.7% attack success rate by perturbing 4.6% tokens on average in the input sequences. In the targeted attack scenario, we always make BERT output as our expected false class on both datasets, resulting in a huge performance drop on BERT models. We achieve 92.1% and 98.0% on THUNews dataset and Wechat Finance dataset, respectively.

Dataset	Attack Method	% Perturbation	PPL	BertScore	Human Ratings
Yelp (English)	HotFlip	14.9	57.1	0.79	3.337 ± 1.650
	TextFooler	13.5	43.7	0.78	3.361 ± 1.326
	BERT-ATTACK	**4.2**	**31.4**	**0.92**	3.513 ± 1.280
	SemAttack (+all)	4.3	34.4	0.91	$\mathbf{3.524 \pm 1.584}$
THUNews (Chinese)	HotFlip	21.7	488.3	0.60	3.770 ± 1.061
	SemAttack (+all)	**15.1**	**317.4**	**0.76**	$\mathbf{3.846 \pm 0.906}$

Table 5: Language quality evaluation for the generated adversarial texts in both Chinese and English.

Dataset		Human	BERT
Yelp (English)	clean	0.9562 ± 0.0006	0.706
	adversarial	0.9390 ± 0.0010	0.000
THUNews (Chinese)	clean	0.9400 ± 0.0014	0.818
	adversarial	0.9369 ± 0.0015	0.000

Table 6: Human performance compared to BERT classifiers on the original and adversarial datasets.

We also present the **blackbox attack** results in Table 4. We can see that SemAttack (+all) achieves the highest success rates in most cases, which suggests that our semantic perturbation spaces have high adversarial transferability. Note that we do not present the targeted attack on Wechat Finance dataset for HotFlip since all attack attempts failed.

Ablation Studies We conduct a series of ablation studies such as exploration of BERT embedding space, attack strategies, ℓ_p norm selection for Eq.(4), hyper-parameter selection, and attack efficiency comparison, etc. in Appendix F.

3.4 Adversarial Text Quality Evaluation

To confirm that our generated adversarial texts are valid and natural to humans, we conduct both *automatic evaluation* and *human evaluation* on both English and Chinese NLP tasks, considering language quality and utility preservation. More evaluation details can be found in Appendix G.

Language Quality Evaluation We sample 100 original sentences from the test set for both Chinese and English such that all of them can be successfully attacked by SemAttack and our baselines. For automatic evaluation, we consider the average perturbation rate, perplexity (PPL) (based on GPT-2), and BertScore as metrics to indicate the language quality. For human evaluation, we present every generated adversarial sentence to 5 human annotators, ask them to rate the language quality from 1 to 5, and calculate the average ratings. We present the evaluation results in Table 5.

We can see that SemAttack has the best human ratings across different baselines for both Chinese and English. In terms of automatic evaluation metrics, we observe that SemAttack is quite close to the SOTA BERT-ATTACK. We think the reason why SemAttack is slightly weaker than BERT-ATTACK in terms of PPL and BertScore is that SemAttack also considers typos and knowledge-based perturbations. Such perturbations usually look good to humans, but may greatly impact the scores calculated by pretrained language models such as GPT-2 and BERT.

Utility Preservation Evaluation To evaluate human performance on our generated adversarial data, we randomly sample 50 clean sentences and 50 adversarial sentences generated by the targeted SemAttack (+all) for both the English Yelp and the Chinese THUNews dataset. For each sentence, we present the annotators with two labels: a ground truth label and a targeted wrong label (e.g., the most opposite sentiment), and request annotators to choose the correct one. Both clean text and adversarial text are randomly shuffled.

The detailed evaluation results with standard deviation are shown in Table 6. We find that our adversarial text barely impacts human perception, as the human performance on adversarial Yelp data is 93.9%, only 2% lower than the clean data. Human performance on the adversarial Chinese THUNews is 93.7%, which is very close to the performance of 94.0% on the clean dataset.

4 Related Work

Our proposed semantic perturbation functions generalize the existing textual adversarial attacks.

For typo-based perturbation function $\mathcal{F}_T$, existing work (Li et al., 2018; Ebrahimi et al., 2017) applies character-level perturbation to carefully crafted typo words (*e.g.*, from "foolish" to "fo0lish"), thus making the model ignore or misunderstand the original statistical cues.

Knowledge-based perturbation function $\mathcal{F}_K$ uses knowledge base to constrain the search space. For example, Zang et al. (2020) uses sememe-based knowledge base from HowNet (Dong et al., 2010) to construct a search space for word substitution.

Different from our contextualized semantic perturbation function $\mathcal{F}_C$, other work (Jin et al., 2020; Li et al., 2018) uses a non-contextualized word embedding from GLoVe (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013) to build synonym candidates, by querying the cosine similarity or eu-

clidean distance between the original and candidate word and selecting the closet ones as the replacements. However, some antonyms also have high cosine similarity in the Word2Vec space. Thus, additional hand-crafted filtering rules are needed to ensure that the meaning is not changed.

Other work (Garg and Ramakrishnan, 2020; Li et al., 2020, 2021) also leverages pre-trained models to generate contextualized perturbations by masked language modeling, which is a parallel work to `SemAttack`, where we explore the BERT embedding clusters to generate high-quality adversarial examples.

In terms of optimization, unlike the *heuristic-based* previous work that uses greedy (Jin et al., 2020) or genetic algorithms (Zang et al., 2020) which search for the optimal perturbations, or *gradient-based* methods (Wang et al., 2020; Guo et al., 2021) which search for perturbation on a tree-autoencoder with only syntactic constraints or a distribution of adversarial examples, we use an *optimization-based* method to efficiently and effectively search for the optimal adversarial perturbation in the semantic preserving spaces to ensure the validity and naturalness of perturbed sentences.

5 Conclusion

In this paper, we propose a novel semantic adversarial attack framework `SemAttack` to probe the robustness of LMs. Comprehensive experiments show that `SemAttack` is able to generate natural adversarial texts in different languages and achieve higher attack success rates than existing textual attacks. We also demonstrate that existing SOTA LMs and defense methods are still vulnerable to `SemAttack`. We expect our study to shed light on future research on evaluating and enhancing the robustness of LMs for different languages.

Acknowledgments

We gratefully thank the anonymous reviewers and meta-reviewers for their constructive feedback. This work is partially supported by the NSF grant No.1910100, NSF CNS 20-46726 CAR, and Sloan Fellowship.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Nicholas Carlini and David A. Wagner. 2016. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57.

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3601–3608. AAAI Press.

Andy Coenen, Emily Reif, A. Yuan, Been Kim, A. Pearce, F. Viégas, and M. Wattenberg. 2019. Visualizing and measuring the geometry of bert. In *NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Ruixue Ding, Pengjun Xie, Xiaoyan Zhang, Wei Lu, Linlin Li, and Luo Si. 2019. A neural multi-digraph model for Chinese NER with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1462–1467, Florence, Italy. Association for Computational Linguistics.

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. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 31–36. Association for Computational Linguistics.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *EMNLP*.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

John Hewitt and Christopher D. Manning. 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. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2752–2765. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Zhen-Zhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and William B Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069.

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

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

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

Isabel Papadimitriou, Ethan A Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual bert. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

M Sun, J Li, Z Guo, Z Yu, Y Zheng, X Si, and Z Liu. 2016. Thuctc: an efficient chinese text classifier. *GitHub Repository, https://github.com/thunlp/THUCTC (2016, accessed 17 May 2017). Google Scholar*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *ACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Boxin Wang, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li. 2020. T3: Tree-autoencoder constrained adversarial text generation for targeted attack. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6134–6150, Online. Association for Computational Linguistics.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *ArXiv*, abs/1906.08237.

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, ACL 2020, Online, July 5-10, 2020*, pages 6066–6080. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.

A Broader Impact

In this paper, we propose an effective and novel adversarial attack framework `SemAttack` to probe the robustness of state-of-the-art NLP models. Our experiments show that even pre-trained large-scale language models for different languages are not robust under `SemAttack`. We will open-source our code to shed light on future research to evaluate and enhance the robustness of NLP models. Considering attackers may leverage our code to perform adversarial attacks to NLP models, we suggest using adversarial training as an effective approach to improving adversarial robustness, and our proposed framework has provided an efficient way to generate these adversarial training data.

B Model Settings

Whitebox Classifier For English dataset, we use BERT and self-attention LSTM as the classifiers. BERT is a transformer (Vaswani et al., 2017) based model, which is unsupervised pretrained on large corpora. We use the 12-layer BERT-base model with 768 hidden units, 12 self-attention heads, and 110M parameters. For self-attention LSTM, we set the self-attention LSTM to 10 attention hops internally, and use a 300-dim BiLSTM and a 512-units fully-connected layer before the output layer.

We fine-tune BERT on Yelp dataset with a batch size of 64, learning rate of $2e-5$ and early stopping. We train the Self-attention LSTM-based model on 500K review training set for 29 epochs with stochastic gradient descent optimizer under the initial learning rate of 0.1. We run our experiments on i7-7820X CPU with 128GB memory on one RTX 2080Ti GPU.

For both Chinese datasets, we use BERT (Devlin et al., 2019) as the classifier. Chinese BERT is a transformer (Vaswani et al., 2017) based model, which is unsupervisedly pretrained on large Chinese corpora and is effective for downstream Chinese NLP tasks. We use the 12-layer BERT-base model with 768 hidden units, 12 self-attention heads and 110M parameters. We fine-tune BERT on each dataset independently with a batch size of 64, learning rate of 2e-5 and early stopping.

Blackbox Classifier The blackbox LSTM and BERT classifiers are trained/finetuned from scratch. The parameters of blackbox models are different from the whitebox ones.

C Dataset Details

• **Yelp Dataset** consists of 2.7M yelp reviews and each one has its corresponding star level to be predicted by our model. The target stars level is an integer number in the inclusive range of $[0, 4]$, which can be treated as 5 classes. We follow the process in Lin et al. (2017) to randomly select 500K review-star pairs as the training set, $2,000$ as the development set, and $2,000$ as the test set.

• **SNLI Dataset** (Bowman et al., 2015) consists of 570k human-written English sentence pairs and each pair contains one premise and one hypothesis. These pairs are manually labeled as entailment, contradiction, or neutral, which can be predicted by our model. We use 550k pairs as training set, 10k as the development set, and 10k as the test set. We follow the baseline setting (Li et al., 2020) and only allow perturbations on hypotheses (Table 2) or premises (Appendix Table 9 & 10).

• **THUNews (Sun et al., 2016)** is a public Chinese 14-category news classification dataset. It consists of more than 740k news articles from Sina News between 2005 and 2011. These articles are classified into 14 categories, such as education, technology, society and politics. To speed up the evaluation process, we use the news titles for classification. We evenly sample articles from all classes, and use $585,390$ articles as the training set, $250,682$ as the development set, and another $1,000$ as the testing set for the adversarial evaluation.

• **Wechat Finance Dataset** is a private dataset from the Wechat team, who collect $13,051$ subscription accounts in the finance domain. They use crowd-sourcing to classify the account into 11 sub-classes, such as insurance, banks and funds. Each account description has 94.18 Chinese characters on average. We split the dataset into training set ($10,000$ descriptions), validation set ($1,163$ descriptions) and test set ($1,888$ descriptions).

Dataset	avg length	LSTM Acc	BERT Acc
Yelp	135	0.705	0.706
SNLI	13(P)/7(H)	0.716	0.829

Table 7: Statistics of Yelp Dataset and SNLI Dataset together with benign accuracy of two models. In SNLI Dataset, we calculate the average length of premises (P) and hypotheses (H) separately.

Algorithm 1 `SemAttack`: Generating multilingual natural adversarial examples

Input: Input tokens $\boldsymbol{x} = [\boldsymbol{x_0}, \boldsymbol{x_1}, ..., \boldsymbol{x_n}]$, classifier $f : \boldsymbol{x} \rightarrow \boldsymbol{z}$ maps input to logits, attack objective function $g(\cdot)$, embedding matrix $\boldsymbol{M_e}$, constants c and κ, max iteration steps m, semantic perturbation function $\mathcal{F}$
Output: Adversarial text x'

Initialize perturbation $\boldsymbol{e}_0^* \leftarrow 0$
$\boldsymbol{e} \leftarrow \boldsymbol{M_e x}$
$\boldsymbol{e'} \leftarrow \boldsymbol{e} + \boldsymbol{e}_0^*$
$\boldsymbol{x'} \leftarrow \boldsymbol{x}$
for $k = 0, 1, ..., m-1$ **do**
  // Phase I: Optimize over the $\boldsymbol{e_k^*}$
  $\mathcal{L}(\boldsymbol{e_k^*}) \leftarrow ||\boldsymbol{e_k^*}||_p + c \cdot g(\boldsymbol{x'})$
  $\boldsymbol{e_{k+1}^*} \leftarrow \boldsymbol{e_k^*} - \alpha \nabla \mathcal{L}(\boldsymbol{e_k^*})$
  // Phase II: Token Substitution
  $\boldsymbol{e'} \leftarrow \boldsymbol{e} + \boldsymbol{e_{k+1}^*}$
  for $i = 1, 2, ..., n$ **do**
    $\mathcal{S} = \mathcal{F}(x_i)$ *// Get the perturbation search space*
    $\boldsymbol{x_i'} \leftarrow \arg\min_{\boldsymbol{x_i'} \in \mathcal{S}}(||\boldsymbol{e_i'} - \boldsymbol{M_e x_i'}||_p)$
  end for
end for
return $\boldsymbol{x'}$

D Experimental Setting

D.1 Attack Setup

`SemAttack` is a whitebox attack method which requires access to the model parameters and gradients. However, it can be easily adapted to blackbox settings. In our experiment, we consider the following two blackbox settings: a soft-label blackbox setting and a more rigorous zero-query blackbox setting. In soft-label blackbox setting, attackers can only query the classifier for output probabilities on a given input. We adapt our method to this setting by distillation. The output confidence of the blackbox (teacher) model is used to train a student model. Then we run whitebox attacks on the student model and attack the teacher model with adversarial instances provided by the student model. In zero-query blackbox setting, the target models (usually state-of-the-art large-scale language models enhanced with cutting-edge defense methods) are unavailable during the attacking phase, which is a common scenario in real-world applications and better demonstrates the algorithm's ability to generalize across models. We adapt `SemAttack` and baseline methods to this setting by performing a transferability-based backbox attack, in which we use adversarial texts created by BERT to attack the target models.

D.2 Embedding Space Construction

To construct the contextualized semantic perturbation function $\mathcal{F}_C$, we select $22,271$ English words commonly used as $\mathcal{X}$, which is also the vocabulary used by English BERT. For each word, We select at most 100 sentences that contain this specific word from wikidump. These contextualized embeddings form an embedding space of $2,181,622$ vectors in total. We choose $k = 700$ and $\epsilon = 8$, which means we only choose words that appear more than 8 times in the 700 nearest neighbors as the perturbation set $\mathcal{S}$. We apply similar strategies when constructing Chinese BERT embedding space, by choosing $5,178$ Chinese tokens appearing in the training data and up to 100 sentences from Chinese Wikipedia, which form an embedding space of $508,619$ vectors in total. When performing KNN, we choose $k = 700$ and $\epsilon = 5$. The query time of $\mathcal{F}_C$ is around 2.6s for English and 0.9s for Chinese. We provide more detailed settings in Appendix E.

D.3 Semantic Perturbation Functions

English We evaluate the following semantic perturbation functions for English corpus: typo-based perturbation function $\mathcal{F}_T$, knowledge-based perturbation function $\mathcal{F}_K$, and contextualized semantic perturbation function $\mathcal{F}_C$ based on BERT embedding clusters, together with the combination of $\mathcal{F}_T$, $\mathcal{F}_K$ and $\mathcal{F}_C$. The average sizes of search spaces obtained by $\mathcal{F}_T$, $\mathcal{F}_K$ and $\mathcal{F}_C$ are 5.03, 2.38 and 4.46, respectively.

Chinese We implement semantic perturbation functions for Chinese corpora as follows: (1) typo-based perturbation function $\mathcal{F}_T$, where typos are defined as Chinese characters with similar strokes or pronunciations, (2) knowledge-based perturbation function $\mathcal{F}_K$, where synonyms are obtained from Chinese WordNet, (3) contextualized semantic perturbation function $\mathcal{F}_C$ by Chinese BERT embedding clusters, and (4) the combination of these three functions.

Because in Chinese there are many characters with the same pronunciation, we limit the number of characters obtained by similar pronunciations to 5. The average sizes of perturbation search space collected by $\mathcal{F}_T$, $\mathcal{F}_K$ and $\mathcal{F}_C$ are 8.53, 0.27 and 17.06. $\mathcal{F}_K$ gives fewer candidate perturbations because Chinese WordNet has limited hand-crafted knowledge, while $\mathcal{F}_C$ gives more choices because it searches in BERT embedding space without human supervision.

D.4 Attack Hyper-parameter Settings

For English dataset, we set the max optimization steps m to 100 and use ℓ_2 norm in the loss function (equation 4) that is iteratively optimized via Adam (Kingma and Ba, 2014). Constants c and κ are set to $1e2$ and 1 in Yelp dataset, $1e4$ and 0 in SNLI dataset, which result in higher attack success rate and lower perturbation rate based on a series of ablation studies provided in Appendix Figure 5. We set our random seed to 1111 for reproducibility.

For Chinese dataset, we follow the experiment setting in English attacks for optimizing adversarial examples and training BERT models. Constants c and κ are set to 100 and 1 respectively to get the best performance. We set our random seed to 1111 for reproducibility. We experiment with different attack strategies in Appendix Table 11 to 13.

E SemAttack Implementation Details

E.1 Typo-based Perturbation Function Implementation

We use the similar Chinese character list[1] that contains more than 9,000 common Chinese characters. We use the existing Python library[2] to query the pronunciations for Chinese characters and another library[3] to search for the words that share the same pronunciations. Because in Chinese there are many characters with the same pronunciation, we limit the number of characters obtained by similar pronunciations to 5.

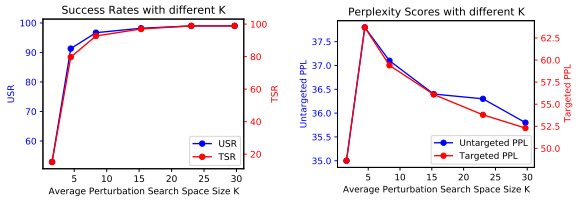
E.2 Knowledge-based Perturbation Function Implementation

In this paper, we use WordNet as an example to illustrate how our framework can integrate the rule-based knowledge to enhance the quality of our adversarial examples. For an input token x, we first query the synonym set s in the WordNet. For each meaning of the input word, the output synonym set s contains several synonyms that have this specific meaning. The output synonyms are given with their corresponding part-of-speech tags. In order not to include synonyms that have unusual part of speech, which may result in strange grammatical errors after replacement, we counted the frequency of each part of speech in set s and only selected the words with the highest frequency of part of speech. Using the synonym set after filtering, we are able to generate adversarial input texts that mislead models' prediction while having little effect on human understanding.

[1]Publicly available at https://github.com/zzboy/chinese/

[2]Publicly available at https://github.com/mozillazg/python-pinyin

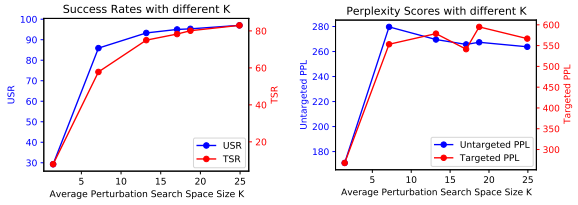
[3]Publicly available at https://github.com/letiantian/Pinyin2Hanzi



(a) Attack success rates with different perturbation search space size K.

(b) Perplexity scores with different perturbation search space size K.

Figure 2: English perturbation space size selection.



(a) Attack success rates with different perturbation search space size K.

(b) Perplexity scores with different perturbation search space size K.

Figure 3: Chinese perturbation space size selection.

F Ablation Studies

F.1 Perturbation space size selection

In Figure 2, 3, we present the attack success rates and perplexity scores of generated adversarial examples under different sizes of perturbation search space. We observe that in both languages, larger K lead to higher attack success rates. In English, PPL score decreases when K continues to increase, while in Chinese PPL score remains at a similar level.

F.2 Attack Efficiency

SemAttack is more efficient than existing baselines since it can substantially decrease the query time when performing attacks. SemAttack searches for the optimal perturbation e^* for a whole sentence in one query, instead of querying every word. Quantitatively, SemAttack is designed to query the model for less than 100 iterations, while BERT-ATTACK and TextFooler require hundreds of queries to generate one adversarial example on average.

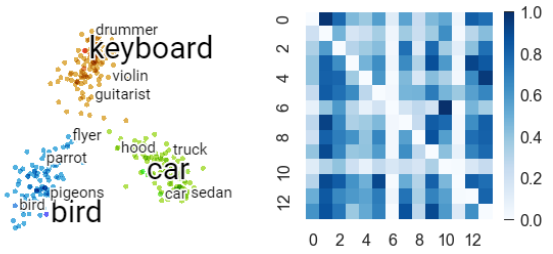
Model	Method	% USR/TSR	% Perturbation
BERT (Acc: 0.706)	TextFooler	84.7/48.6	13.5/32.2
	BERT-ATTACK	**95.4**/71.1	4.2/11.2
	SemAttack (+$\mathcal{F}_T$)	32.6/6.7	4.6/9.1
	SemAttack (+$\mathcal{F}_K$)	58.8/51.5	5.9/15.5
	SemAttack (+$\mathcal{F}_C$)	68.4/61.3	4.7/12.1
	SemAttack (+all)	67.5/**72.4**	4.0/11.7
Self-Attention LSTM (Acc: 0.705)	TextFooler	17.5/5.7	9.6/28.0
	BERT-ATTACK	65.0/24.7	2.2/3.7
	SemAttack (+$\mathcal{F}_T$)	51.2/25.0	18.3/22.4
	SemAttack (+$\mathcal{F}_K$)	39.2/24.0	15.0/19.2
	SemAttack (+$\mathcal{F}_C$)	57.7/33.7	23.4/26.7
	SemAttack (+all)	**74.1/67.0**	30.6/35.8

(a) Yelp Dataset

Model	Method	% USR/TSR	% Perturbation
BERT (Acc: 0.829)	TextFooler	73.2/30.8	22.3/24.7
	BERT-ATTACK	**88.9/61.8**	17.0/20.1
	SemAttack (+$\mathcal{F}_T$)	19.1/6.8	10.2/11.2
	SemAttack (+$\mathcal{F}_K$)	36.7/12.5	12.9/20.0
	SemAttack (+$\mathcal{F}_C$)	59.8/45.0	14.8/26.1
	SemAttack (+all)	63.9/40.5	15.2/17.1
Self-Attention LSTM (Acc: 0.705)	TextFooler	52.9/24.2	20.1/24.7
	BERT-ATTACK	62.8/36.9	17.9/18.7
	SemAttack (+$\mathcal{F}_T$)	49.9/33.3	26.4/32.9
	SemAttack (+$\mathcal{F}_K$)	40.3/22.5	22.1/25.6
	SemAttack (+$\mathcal{F}_C$)	68.9/56.9	33.0/39.5
	SemAttack (+all)	**75.4/57.0**	42.3/37.9

(b) SNLI Dataset

Table 8: Soft-label blackbox attack success rate for different attacks under targeted/untargeted attacks (TSR/USR) and corresponding word perturbation percentage against self-attention LSTM and BERT on Yelp and SNLI datasets.



(a) Visualization. (b) Confusion matrix.

Figure 4: Ablation studies. (a) shows the visualization of English words in BERT embedding clusters. (b) shows the TSR confusion matrix on THUNews dataset.

F.3 BERT Embedding Space

In Figure 4a, we visualize three clusters: "car", "bird" and "keyboard". Here "keyboard" is used as an instrument, not a peripheral device of PCs. As we can see, 'bird' has neighbors such as "pigeons", "parrot" and "flyer", which are not present in knowledge space. Word "keyboard" has neighbors such as "drummer", "violin" and "guitarist", which are contextualized based on the query context.

F.4 Additional Results on Attacking SNLI

We follow the setting of (Li et al., 2020) and perturb only hypotheses or premises for SNLI tasks. Attack results for perturbing hypotheses are shown in main paper Table 2. Attack results for perturbing premises only are shown in Table 9 and 10.

F.5 Ablation Studies on Attack Capability

In this section, we will evaluate the possible factors that will affect the attack success rate. Here, we set the candidate search space $\mathcal{S}$ to be the whole vocabulary V to eliminate variables introduced by the perturbation function.

Model	Method	% USR/TSR	% Perturbed
BERT (Acc: 0.829)	HotFlip	43.6/20.5	28.0/29.8
	SemAttack (+$\mathcal{F}_T$)	11.6/4.1	11.2/12.5
	SemAttack (+$\mathcal{F}_K$)	25.4/12.2	12.9/17.2
	SemAttack (+$\mathcal{F}_C$)	66.4/36.7	16.4/21.2
	SemAttack (+all)	**72.7/46.1**	17.5/21.6
Self-Attention LSTM (Acc: 0.716)	HotFlip	10.8/8.2	10.2/10.0
	SemAttack (+$\mathcal{F}_T$)	47.5/29.3	15.5/19.1
	SemAttack (+$\mathcal{F}_K$)	43.4/22.2	13.2/15.0
	SemAttack (+$\mathcal{F}_C$)	69.7/48.5	28.2/35.5
	SemAttack (+all)	**70.7/46.5**	29.5/36.6

Table 9: The whitebox attack success rate (in terms of "USR/TSR") and corresponding word perturbation percentage against LSTM and BERT on the SNLI dataset by only perturbing premises.

Model	Method	% USR/TSR	% Perturbed
BERT (Acc: 0.829)	TextFooler	61.3/31.1	15.0/17.0
	BERT-ATTACK	**60.2/34.8**	25.6/34.4
	SemAttack (+$\mathcal{F}_T$)	11.5/4.3	4.9/5.6
	SemAttack (+$\mathcal{F}_K$)	17.0/7.0	11.2/13.1
	SemAttack (+$\mathcal{F}_C$)	43.0/24.8	13.4/16.1
	SemAttack (+all)	47.0/30.2	14.6/17.5
Self-Attention LSTM (Acc: 0.716)	TextFooler	19.1/10.6	10.3/10.6
	BERT-ATTACK	42.9/31.5	19.4/23.0
	SemAttack (+$\mathcal{F}_T$)	29.4/22.7	23.1/27.6
	SemAttack (+$\mathcal{F}_K$)	23.2/15.8	20.7/23.0
	SemAttack (+$\mathcal{F}_C$)	55.9/46.3	43.5/45.7
	SemAttack (+all)	**59.0/49.7**	45.7/47.8

Table 10: The blackbox attack success rate (in terms of "USR/TSR") and corresponding word perturbation percentage against LSTM and BERT on the SNLI dataset by only perturbing premises.

F.6 Norm selection

In the main experiment, we use l_2 norm for our attack loss function (equation 7). However, because l_1 norm is known for good at feature selection and generating sparse features, we conduct the following experiments by setting l_p to l_1 and make an comparison with l_2 norm. The experimental results are shown in Table 11 and 12. We find the overall attack success rates decrease when switching to l_1 norm. However, given the same set of constants c and κ, we find the l_1 attack does change less words.

Dataset	Original		SemAttack (l_2 untargeted)			SemAttack (l_1 untargeted)			Baseline
	Acc	c/k	5/5	10/5	10/10	10/10	10/100	20/20	(untargeted)
THUCTC	0.818	target	-	-	-	-	-	-	-
		untarget	**1.000**	**1.000**	**1.000**	0.983	0.983	**0.995**	0.040
		#/chars	**1.583**	1.690	1.718	**1.577**	1.614	1.884	2.000

Table 11: Untargeted attack success rates on Chinese BERT-based classifier for THUCTC dataset. "target" and "untarget" calculate the targeted attack success rate (equation 1) and the untargeted attack success rate (equation 2). "#/chars" counts the number characters are modified in average.

Dataset	Original		SemAttack (l_1 targeted)			SemAttack (l_2 targeted)			Baseline
	Acc	c/k	10/10	10/20	30/30	5/5	10/5	10/10	(untargeted)
THUCTC	0.818	target	0.797	0.797	**0.898**	0.941	**0.945**	**0.945**	-
		untarget	0.828	0.828	**0.920**	0.953	**0.958**	**0.958**	0.040
		#/chars	2.000	**1.956**	3.280	**2.924**	3.186	3.045	2.000

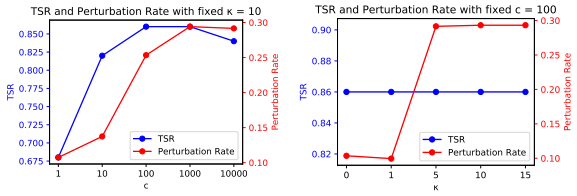
Table 12: Targeted attack success rates on Chinese BERT-based classifier for THUCTC dataset. "target" and "untarget" calculate the targeted attack success rate (equation 1) and the untargeted attack success rate (equation 2). "#/chars" counts the number charcters are modified in average.

F.7 Attack Strategy

As we have achieved 100% attack success rate in the untargeted attack scenario, we now focus on the targeted attack scenario and see which factor contributes to the targeted attack success rate. It is straightforward to think different targeted attack strategies will impact the targeted attack success rate, because maybe some classes look "farther" than semantic closer classes. So we tried two strategies on THUCTC dataset: 1) as used in the main paper, we set the targeted false class as "technology news". 2) we enumerate all the classes and set the targeted false class to be numerically the next class index. The targeted attack success rate is shown in Table 13. We do find choosing different attack strategies will impact the attack success rate.

F.8 Hyper Parameter Selection

We have two constants in our attack algorithm, c and κ, which control the attack success rates and the perturbation rates in our experiments. In order to find out the impact of these hyper parameters, we test with several combinations of different c and κ. We test on Yelp Dataset and we use BERT as our model. We show our results in Figure 5. As shown in Figure 5a, we first fix $\kappa = 10$ and test how TSR and perturbation rate will change according to different c. We find that under the same κ, c mainly controls the attack success rate at the cost of perturbation rate. In some certain range, a larger c encourages the algorithm to achieve our attack goal with the expense of more substitutions. And after exceeding a certain value, TSR will start to decrease while perturbation rate remains high. We then fix $c = 100$ and test different κ. We show our results in Figure 5b. We find that κ doesn't help to increase TSR and a smaller κ helps to limit the words changed without affecting TSR.



(a) Fixed κ and different c. (b) Fixed c and different κ.

Figure 5: Hyper parameter selection. In Figure 5a, we first fix $\kappa = 10$ and test different c to see how TSR and perturbation rate will change. we test $c = 1, 10, 10^2, 10^3, 10^4$ and find best $c = 100$ to obtain the highest TSR with less perturbations. A smaller or larger c will result in a low TSR or a high perturbation rate. In Figure 5b, after fixing $c = 100$, we test $\kappa = 0, 1, 5, 10, 15$. We find that κ has little influence on TSR while it can change perturbation rate dramatically. A smaller κ is able to effectively limit the number of words to be changed. In our experiment, we choose $\kappa = 0, 1$.

For hyper-parameter selection for Chinese datasets, we witness the same phenomenon in English attacks that increasing constant c can improve the attack success rate at the cost of more perturbed characters, while lowering constant κ limits the perturbation rate without affecting the attack success rate.

Dataset	Original Acc		SemAttack (targeted $c/\kappa = 10/10$) strategy 1	SemAttack (targeted $c/\kappa = 10/10$) strategy 2	Baseline (untargeted)
THUCTC	0.818	target	**0.945**	0.903	-
		untarget	**0.958**	0.913	0.040
		#/chars	3.045	4.543	2.000

Table 13: Attack success rates on Chinese BERT-based classifier for two datasets. "target" and "untarget" calculate the targeted attack success rate (equation 1) and the untargeted attack success rate (equation 2). "#/chars" counts the number characters are modified in average.

Transfer	Method	% TSR	% USR
Self-Attention LSTM → BERT	TextFooler	42.4	43.9
	BERT-ATTACK	8.1	33.5
	SemAttack (+$\mathcal{F}_T$)	44.4	32.5
	SemAttack (+$\mathcal{F}_K$)	57.7	62.0
	SemAttack (+$\mathcal{F}_C$)	**74.3**	**81.2**
	SemAttack (+all)	70.0	79.8
BERT → Self-Attention LSTM	TextFooler	30.8	31.9
	BERT-ATTACK	17.6	28.5
	SemAttack (+$\mathcal{F}_T$)	26.8	34.6
	SemAttack (+$\mathcal{F}_K$)	35.3	35.6
	SemAttack (+$\mathcal{F}_C$)	**35.5**	**36.0**
	SemAttack (+all)	30.9	31.0

Table 14: Targeted and untargeted attack success rate of transferability attack on Yelp Dataset, evaluating adversarial examples generated against Self-attention LSTMs on BERT, and vice versa.

F.9 Vulnerability Between Classes

In THUNews dataset, the article titles are classified into 14 categories. In order to find out the vulnerability of each class, we test the attack success rate of each source class and target class. The heatmap of results is provided in Figure 4b. We find that "technology news" and "entertainment news" as target classes have higher average success rates than other classes, while "lottery ticket" is the lowest. We also find that "constellation news" has the highest average success rate as source class, while "sports news" has the lowest, which means "constellation news" is vulnerable and easy to attack while "sports news" is much more robust.

F.10 Transferability Analysis

We evaluate the transferability of adversarial examples between different models by attacking a blackbox BERT classifier by using adversarial text generated from a whitebox LSTM, and vice versa.

The transferability-based attack results on Yelp Dataset are shown in Table 14. We find that the robustness of the two models is highly different from each other. When we feed adversarial texts generated from the LSTM model into the blackbox BERT model, attack success rate is higher than 70%. However, when we test the performance of the blackbox LSTM model on adversarial texts generated from the whitebox BERT, attack success rate is around 30%, which is much lower than previous experiment. These results show that Self-Attention LSTMs are more robust than BERT models, and the adversarial examples generated from a robust model has higher attack transferability than non-robust one. Therefore, we can attack blackbox BERT models using a strong Self-Attention LSTM trained by ourselves to generate adversarial texts with high success rates. We also observe that the USR of transferability-based attack is generally higher than that of targeted attack. Particularly, we achieve the highest success rate of 81.2% when attacking blackbox BERT with text generated by LSTM attacks under untargeted setting.

Furthermore, we find that the adversarial examples generated by the contextualized semantic perturbation functiuon $\mathcal{F}_C$ have the highest attack transferability, which suggests that our contextualized semantic perturbation is more generalizable than rule-based perturbation functions.

G Human Evaluation Details

Language Quality Evaluation Details We use Amazon Turk for English adversarial example quality annotations, and Alibaba Cloud for Chinese example quality annotations. Each sentence is annotated by 5 annotators. This evaluation only evaluates language quality and grammatical correctness, and thus does not require additional background or domain knowledge.

We present the annotation instructions on Amazon Turk below.

Please **rate the language quality** (from 1 to 5, in terms of coherence, fluency, and grammar correctness) of the presented sentence. 5 means the best language quality, and 1 means the lowest language quality.

- 5: The sentence looks totally correct. There

are no grammatical errors. I can fully understand the sentence.

- 4: The sentence looks somewhat correct. There are one or two grammatical errors or typos. But I can mostly understand the sentence.
- 3: The sentence looks OK to me. There are some grammatical errors or typos. I can partly understand the sentence.
- 2: The sentence looks bad to me. There are grammatical errors or typos everywhere. I can understand it a little.
- 1: The sentence totally does not make any sense. I cannot understand it.

Utility Preservation Evaluation Details We use the targeted `SemAttack` to generate the adversarial dataset with with $c/\kappa = 100/1$. In total, we collected annotations from 21 graduate students from US universities for English datasets and 26 annotators from native Chinese speakers for Chinese datasets. Both classification tasks do not require domain knowledge. The detailed human performance results are shown in Table 6.

H Perturbation Search Space Examples

H.1 English Perturbation Search Space $\mathcal{S}$ Examples

Table 15: English Perturbation Search Space $\mathcal{S}$ Examples Generated by `SemAttack` for BERT-based Classifier using $\mathcal{F}_T$, $\mathcal{F}_K$ and $\mathcal{F}_C$. In the first example, we list some words and corresponding candidate sets generated by these functions. We can see that words generated by $\mathcal{F}_C$ reflect the meaning of the current context. For example, when we say that a hotel is *good*, we may say it's *spacious*. When word *come* is followed by *back*, we may mean *return*. In the following two examples, we show that the same word may have different perturbation sets in different contexts. In the second example, by using *order*, the person means that he ordered food. Considering the context, $\mathcal{F}_C$ provides *eat*, *taste* in its candidate set. In the last example, *order* means the person orders a drink. As a consequence, we have *drink* as a verb with a similar meaning in its candidate set.

Input English Text: This was my fifth time traveling to vegas! I have stayed at hotels such as the Bellagio, Aria, Cosmopolitan, the venetian, and fortunately enough got a chance to **stay** at vdara. Considering the reviews I didn't expect vdara to be that-**good** of a hotel! Vdara was extremely **clean**, very modern, new, great customer service, **close** to the strip-connected to the bellagio. easy access to casinos and heart of the strip. Definitely **coming** back to vegas and booking a room at vdara.

$\mathcal{F}_T(stay)$ = stay
$\mathcal{F}_K(stay)$ = quell, last out, bide, persist, stay
$\mathcal{F}_C(stay)$ = staying, stay, vacationing, stays, relax, internship, enroll, stayed, visit, settle

$\mathcal{F}_T(good)$ = good, god
$\mathcal{F}_K(good)$ = estimable, adept, full, effective, dear, beneficial, dependable, good
$\mathcal{F}_C(good)$ = spacious, marvelous, marvel, wonderful, good

$\mathcal{F}_T(clean)$ = clean
$\mathcal{F}_K(clean)$ = blank, clean, uninfected
$\mathcal{F}_C(clean)$ = spacious, luxurious, lively, vibrant, cleanest, cozy, cleaned, renovated, clean

$\mathcal{F}_T(close)$ = close
$\mathcal{F}_K(close)$ = close, conclude, close up
$\mathcal{F}_C(close)$ = connected, near, close, nearer, closeness

$\mathcal{F}_T(coming)$ = coming
$\mathcal{F}_K(coming)$ = come, derive, issue forth, arrive, hail, total, occur, do, fall
$\mathcal{F}_C(coming)$ = returning, traveling, transferring, staying, relocating, visiting, talking, coming

Input English Text: Stopped by this place for lunch . **Ordered** the veggie slice and patty they put lettuce cheese and mayo in it and both the slice and patty were amazing. Definitely will be back for more.

$\mathcal{F}_T(Ordered)$ = ordered
$\mathcal{F}_K(Ordered)$ = rate, ordain, arrange, order, regulate
$\mathcal{F}_C(Ordered)$ = ate, tasted, ordered

Input English Text: Love this speakeasy bar. Last time I was at this location it was still the Panda bar. The place itself is super cozy and intimate. We went there to grab a drink before our Ali Wong show. Hubby **ordered** a Hendricks gin tonic (12$-happy hour price?) and I got a cocktail with Pimms (9$ before 9pm). The drinks were HUMONGOUS! So much so I couldnt finish mine and hubby was tipsy lol.

$\mathcal{F}_T(Ordered)$ = ordered
$\mathcal{F}_K(Ordered)$ = rate, ordain, arrange, order, regulate
$\mathcal{F}_C(Ordered)$ = ate, drank, ordered

H.2 Chinese Perturbation Search Space $\mathcal{S}$ Examples

Table 16: Chinese Perturbation Search Space $\mathcal{S}$ Examples Generated by `SemAttack` for BERT-based Classifier using $\mathcal{F}_T$, $\mathcal{F}_K$ and $\mathcal{F}_C$. Chinese characters are intrinsically polysemous, which requires candidate characters to be contextualized. We list four examples here. In the first two examples, we show two different meanings of character "美" in two different sentences. One referring to *the US* which has some other countries' names in its perturbation set, another meaning *poignant* which is used as an adjective. In the last two examples, we show "长", a well-known Chinese character that has multiple pronunciations and multiple meanings. We show that our two perturbation functions return different candidate sets. In the third example, "长" means a job title, while in the last example it means *growth*.

Input Chinese Text: 访谈：美国签证官解读学生签证获签要领 **Translation:** Interview: **U.S.** visa officer interprets the essentials of student visa $\mathcal{F}_T$(美) = 芥, 美, 界, 养, 镁, 每, 楣(mustard, nice, world, support, magnesium, each, lintel) $\mathcal{F}_K$(美) = 美(US) $\mathcal{F}_C$(美) = 美, 英, 香, 欧, 日, 澳, 俄, 荷, 德, 港, 华, 葡, 韩(US, Britain, Hong Kong, Europe, Japan, Australia, Russian, Netherlands, Germany, Hong Kong, China, Portugal, Korean)
Input Chinese Text: 陈嘉上《画皮》大换皮凄美爱情赢得眼泪(图) **Translation:** Chen Jia's "Painted Skin" changes skin, **poignant** love wins tears (photo) $\mathcal{F}_T$(美) = 芥, 美, 界, 养, 镁, 每, 楣(mustard, nice, world, support, magnesium, each, lintel) $\mathcal{F}_K$(美) = 美(poignant) $\mathcal{F}_C$(美) = 寞, 挚, 妙, 美, 腻, 酷, 烂, 凑, 坷, 凄, 惨, 悲, 慨(lonely, sincere, wonderful, nice, greasy, cool, rotten, make up, bumpy, sad, awful, sad, sad)
Input Chinese Text: 北京房协副秘书长陈志谈地产业诚信问题 **Translation:** Chen Zhi, Deputy Secretary-**General** of the Beijing Housing Association, talks about the integrity of the real estate industry $\mathcal{F}_T$(长) = 氏, 氐, 掌, 涨, 长(clan name, foundation, palm, rise, long) $\mathcal{F}_K$(长) = 长(general) $\mathcal{F}_C$(长) = 长, 授, 卿, 员, 师, 委, 厅, 秘, 副, 顾, 官, 董(general, professor, minister, member, teacher, committee, office, secretary, deputy, consultant, official, director)
Input Chinese Text: IMF大幅下调明年全球经济增长预期 **Translation:** IMF sharply lowered its forecast for global economic **growth** next year $\mathcal{F}_T$(长) = 氏, 氐, 掌, 涨, 长(clan name, foundation, palm, rise, long) $\mathcal{F}_K$(长) = 长, 加, 大(growth, increase, increase) $\mathcal{F}_C$(长) = 胀, 增, 膨, 速, 涨, 长, 加, 快(swell, increase, inflate, speed, rise, grow, plus, fast)

H.3 English Adversarial Examples

Table 17: English Adversarial Examples Generated by `SemAttack` for BERT-based Classifier using $\mathcal{F}_T$.

Input (red = Modified character, **bold** = original character.)
Original English Text: I went to AAA **for** their travel service. They could not help me at all with my trip to Belize. They have zilch information and resources. This is a prime destination of American tourists. I was disappointed. **Adversarial English Text:** I went to AAA tor their travel service. They could not help me at all with my trip to Belize. They have zilch information and resources. This is a prime destination of American tourists. I was disappointed. **Model Prediction:** 1-star (most negative) → 5-star (most positive)
Original English Text: I called numerous times and noted that they are going to deliver at a work address between 9 am to 5 pm. They attempted delivery three times after 5 pm. I **got** ups to pick up my parcel and got it delivered on time. **Adversarial English Text:** I called numerous times and noted that they are going to deliver at a work address between 9 am to 5 pm. They attempted delivery three times after 5 pm. I hot ups to pick up my parcel and got it delivered on time. **Model Prediction:** 1-star (most negative) → 5-star (most positive)
Original English Text: Mercedes does the **best** mani and pedi! You really have to go in at least once to see what I mean. **Adversarial English Text:** Mercedes does the bet mani and pedi! You really have to go in at least once to see what I mean. **Model Prediction:** 5-star (most positive) → 1-star (most negative)
Original English Text: I was charged $ 200 to **add** 6 lbs of Freon to my air conditioning. I went to amazon.com and 25 lbs cost $ 120 including shipping. That should be approx $ 29 **for** 6 lbs of Freon. So labor which was 20 min, transportation, and equipment up - keep for john, the service person who came, was $ 171. I feel that's somewhat unreasonable. Just fair warning for the next customer. Update: after listening to my complaint, the owner offered to refund my payment. That was quite reasonable of them. Therefore, I switch my review to 4 stars. **Adversarial English Text:** I was charged $ 200 to ad 6 lbs of Freon to my air conditioning. I went to amazon.com and 25 lbs cost $ 120 including shipping. That should be approx $ 29 fog 6 lbs of Freon. So labor which was 20 min, transportation, and equipment up - keep for john, the service person who came, was $ 171. I feel that's somewhat unreasonable. Just fair warning for the next customer. Update: after listening to my complaint, the owner offered to refund my payment. That was quite reasonable of them. Therefore, I switch my review to 4 stars. **Model Prediction:** 4-star (positive) → 1-star (most negative)
Original English Text: **Liked** how they were open late and also had happy hour specials after 10 pm. We really **liked** the bulgogi and korean prime kalbi. They were marinated very flavor-fully . the mushroom medley and sweet corn were also very good. Would definitely keep this place on my list of late night eats or when iia just craving korean barbecue. **Adversarial English Text:** Lied how they were open late and also had happy hour specials after 10 pm. We really lied the bulgogi and korean prime kalbi. They were marinated very flavor-fully . the mushroom medley and sweet corn were also very good. Would definitely keep this place on my list of late night eats or when iia just craving korean barbecue. **Model Prediction:** 4-star (positive) → 1-star (most negative)

Table 18: English Adversarial Examples Generated by SemAttack for BERT-based Classifier using $\mathcal{F}_K$.

Input (red = Modified character, **bold** = original character.)
Original English Text: Like the others below, I had a similar bad experience with this company. I also forgot to check here before I bought the living social deal. I am having some **issues** getting it refunded as well. Maid affordable was a no show, will not **call** back, and does not answer the phone or emails. Definitely take your business to someone else. **Adversarial English Text:** Like the others below, I had a similar bad experience with this company. I also forgot to check here before I bought the living social deal. I am having some topic getting it refunded as well. Maid affordable was a no show, will not shout back, and does not answer the phone or emails. Definitely take your business to someone else. **Model Prediction:** 1-star (most negative) → 5-star (most positive)
Original English Text: Just another reason why I will never bank with chase.... so now you can't deposit any amount of cash without **showing** your id..... so much for just running to the bank quick. **Adversarial English Text:** Just another reason why I will never bank with chase.... so now you can't deposit any amount of cash without usher your id..... so much for just running to the bank quick. **Model Prediction:** 1-star (most negative) → 5-star (most positive)
Original English Text: My 2017 camry got a check engine light and my car had a strong odor of gasoline after service closed, I asked the receptionist if there was anyway they could get me a rental and she said they were closed so she recommended me to come in bright and early at 7am on monday so they could look at my car so I told her I left for work at 6am cause I work in north scottsdale so I told her I didn't not want to drive my car to scottsdale and back because I was afraid my car would blow up or something from the strong odor of gasoline and she put me on hold to talk to a manager. When she came back on the phone she said her manager was going to get a hold of the rental manager to see if someone could come in tomorrow (today now) to get me a rental and I left my name and number and no one has **reached** out to me. It's great to know they don't care if their customer's car blows up on the freeway cause it's not a sale! Thanks avondale toyota you guys rock ! ! ! ! The dealership I work at teaches their receptionist to hand out rentals cause they know stuff like this happens, you guys might want to look into that ! **Adversarial English Text:** My 2017 camry got a check engine light and my car had a strong odor of gasoline after service closed, I asked the receptionist if there was anyway they could get me a rental and she said they were closed so she recommended me to come in bright and early at 7am on monday so they could look at my car so I told her I left for work at 6am cause I work in north scottsdale so I told her I didn't not want to drive my car to scottsdale and back because I was afraid my car would blow up or something from the strong odor of gasoline and she put me on hold to talk to a manager. When she came back on the phone she said her manager was going to get a hold of the rental manager to see if someone could come in tomorrow (today now) to get me a rental and I left my name and number and no one has achieve out to me. It's great to know they don't care if their customer's car blows up on the freeway cause it's not a sale! Thanks avondale toyota you guys rock ! ! ! ! The dealership I work at teaches their receptionist to hand out rentals cause they know stuff like this happens, you guys might want to look into that ! **Model Prediction:** 1-star (most negative) → 5-star (most positive)
Original English Text: I called numerous times and noted that they are going to deliver at a work address between 9 am to 5 pm. They attempted delivery three **times** after 5 pm. I got up to pick up my parcel and got it delivered on time . **Adversarial English Text:** I called numerous times and noted that they are going to deliver at a work address between 9 am to 5 pm. They attempted delivery three meter after 5 pm. I got up to pick up my parcel and got it delivered on time . **Model Prediction:** 1-star (most negative) → 5-star (most positive)

Table 19: English Adversarial Examples Generated by SemAttack for BERT-based Classifier using $\mathcal{F}_C$.

Input (red = Modified character, **bold** = original character.)
Original English Text: If you **think** Las Vegas is getting too white trash, don't go near here. This place is like a Steinbeck novel come to life. I kept expecting to see donkeys and chickens walking around. woo - pig - soooeeee this place is awful ! ! !   **Adversarial English Text:** If you senses Las Vegas is getting too white trash, don't go near here. This place is like a Steinbeck novel come to life. I kept expecting to see donkeys and chickens walking around. woo - pig - soooeeee this place is awful ! ! !   **Model Prediction:** 1-star (most negative) → 5-star (most positive)
Original English Text: My 2017 camry got a check engine light and my car had a strong odor of gasoline after service closed, I asked the receptionist if there was anyway they could get me a rental and she said they were closed so she recommended me to come in bright and early at 7am on monday so they could look at my car so I told her I left for work at 6am cause I work in north scottsdale so I told her I didn't not want to drive my car to scottsdale and back because I was **afraid** my car would blow up or something from the strong odor of gasoline and she put me on hold to talk to a manager. When she came back on the phone she said her manager was going to get a hold of the rental manager to see if someone could come in tomorrow (today now) to get me a rental and I left my name and number and no one has reached out to me. It's great to know they don't care if their customer's car blows up on the freeway cause it's not a sale ! Thanks avondale toyota you guys rock ! ! ! ! the dealership I work at teaches their receptionist to hand out rentals cause they know stuff like this happens, you guys might want to look into that !   **Adversarial English Text:** My 2017 camry got a check engine light and my car had a strong odor of gasoline after service closed, I asked the receptionist if there was anyway they could get me a rental and she said they were closed so she recommended me to come in bright and early at 7am on monday so they could look at my car so I told her I left for work at 6am cause I work in north scottsdale so I told her I didn't not want to drive my car to scottsdale and back because I was worry my car would blow up or something from the strong odor of gasoline and she put me on hold to talk to a manager. When she came back on the phone she said her manager was going to get a hold of the rental manager to see if someone could come in tomorrow (today now) to get me a rental and I left my name and number and no one has reached out to me. It's great to know they don't care if their customer's car blows up on the freeway cause it's not a sale ! Thanks avondale toyota you guys rock ! ! ! ! the dealership I work at teaches their receptionist to hand out rentals cause they know stuff like this happens, you guys might want to look into that !   **Model Prediction:** 1-star (most negative) → 5-star (most positive)
Original English Text: I have used this company twice. The first time they were great. We spent over 5,000 for installation of a new ac unit on a rental property. Since they did an **excellent** job, we had them do a redesign of ac system in our home to improve the cooling in our house. It was one of the most frustrating customer service experiences I've had with a contractor in the 25 years I have lived in phoenix. They didn't complete the job in the time frame they promised. They damaged the faux ceiling in the kitchen, they drilled holes and didn't repair them in the bedroom. They left marks on the ceiling in the **living** room, where they marked to cut a hole and then didn't. Which told me they installers were not skilled or professional. After waiting for 2 months for them to repair the mistake in the kitchen, we gave up and paid to have it repaired. I heard a lot of promises, no solution. I would never use this contractor again.   **Adversarial English Text:** I have used this company twice. The first time they were great. We spent over 5,000 for installation of a new ac unit on a rental property. Since they did an exemplary job, we had them do a redesign of ac system in our home to improve the cooling in our house. It was one of the most frustrating customer service experiences I've had with a contractor in the 25 years I have lived in phoenix. They didn't complete the job in the time frame they promised. They damaged the faux ceiling in the kitchen, they drilled holes and didn't repair them in the bedroom. They left marks on the ceiling in the attic room, where they marked to cut a hole and then didn't. Which told me they installers were not skilled or professional. After waiting for 2 months for them to repair the mistake in the kitchen, we gave up and paid to have it repaired. I heard a lot of promises, no solution. I would never use this contractor again.   **Model Prediction:** 1-star (most negative) → 5-star (most positive)
Original English Text: There's so many choices of food in Las vegas. Don't choose this place. It is no exaggeration that mcdonalds and arby's have better hash browns, eggs, and bacon. Missed items in the dishes we ordered. All around **disappointment** to the las vegas allure.   **Adversarial English Text:** There's so many choices of food in Las vegas. Don't choose this place. It is no exaggeration that mcdonalds and arby's have better hash browns, eggs, and bacon. Missed items in the dishes we ordered. All around sorrow to the las vegas allure.   **Model Prediction:** 1-star (most negative) → 5-star (most positive)

Table 20: English Adversarial Examples Generated by `SemAttack` for BERT-based Classifier using all perturbation functions.

Input (red = Modified character, **bold** = original character.)
Original English Text: I went to AAA for their travel service. They could not help me at all with my **trip** to Belize. They have zilch information and resources. This is a prime destination of American tourists. I was disappointed. **Adversarial English Text:** I went to AAA for their travel service. They could not help me at all with my voyage to Belize. They have zilch information and resources. This is a prime destination of American tourists. I was disappointed. **Model Prediction:** 1-star (most negative) → 5-star (most positive)
Original English Text: My wife and I have been to this location multiple times, and have only had 1 **bad** experience where the people in the check out area were a little brain dead that day. (they told us that the rug we purchased wasn't in stock, then it was, then wasn't, then was again...) Other than that, we are always **helped** right away, and checking out goes quickly. They also have free self serve Starbucks coffee which I always help myself to. **Adversarial English Text:** My wife and I have been to this location multiple times, and have only had 1 worst experience where the people in the check out area were a little brain dead that day. (they told us that the rug we purchased wasn't in stock, then it was, then wasn't, then was again...) Other than that, we are always served right away, and checking out goes quickly. they also have free self serve Starbucks coffee which I always help myself to. **Model Prediction:** 4-star (positive) → 1-star (most negative)
Original English Text: I love shopping at buffalo exchange but when it comes to selling I prefer selling to the phoenix location because the employees are a lot more genuine, there's less of a hipster pretentious vibe there, and I usually sell more there **too**. Not to mention the tempe location usually turns the music off at 8:30, which gives an unwanted feeling to their guests. I am giving two stars for the sake of finding things at all locations. Go phoenix location! **Adversarial English Text:** I love shopping at buffalo exchange but when it comes to selling I prefer selling to the phoenix location because the employees are a lot more genuine, there's less of a hipster pretentious vibe there, and I usually sell more there anyway. Not to mention the tempe location usually turns the music off at 8:30, which gives an unwanted feeling to their guests. I am giving two stars for the sake of finding things at all locations. Go phoenix location! **Model Prediction:** 2-star (negative) → 5-star (most positive)
Original English Text: There' s so many choices of food in Las Vegas. Don't choose this place. It is no exaggeration that mcdonalds and arby's have **better** hash browns, eggs, and bacon. Missed items in the dishes we ordered. All around disappointment to the Las Vegas allure. **Adversarial English Text:** There's so many choices of food in Las Vegas. Don't choose this place. It is no exaggeration that mcdonalds and arby's have delicious hash browns, eggs, and bacon. Missed items in the dishes we ordered. All around disappointment to the Las Vegas allure. **Model Prediction:** 1-star (most negative) → 5-star (most positive)
Original English Text: Not only is this place in my neighborhood, it is exactly what I'm looking for. I have pale skin, green eyes, and freckles yet I have been **cheated** out of having naturally red hair by mother nature!! Therefore I have been a fake redhead for at least a decade. You can imagine the cost and damage to my hair I have endured. Fringe has a new dye that is ammonia free! It's basically just a oil and water dying process! I've gone twice in a row and my hair has never been in such good condition. I'm paying the same amount for hair dying as my old salon except here I get a better cut and style and it's not frying my hair! Also Chanel (who dyes my hair) is a totally cool chic and always has interesting things to talk about! This is my new go to salon! **Adversarial English Text:** Not only is this place in my neighborhood, it is exactly what I'm looking for. I have pale skin, green eyes, and freckles yet I have been humiliated out of having naturally red hair by mother nature!! Therefore I have been a fake redhead for at least a decade. You can imagine the cost and damage to my hair I have endured. Fringe has a new dye that is ammonia free! It's basically just a oil and water dying process! I've gone twice in a row and my hair has never been in such good condition. I'm paying the same amount for hair dying as my old salon except here I get a better cut and style and it's not frying my hair! Also Chanel (who dyes my hair) is a totally cool chic and always has interesting things to talk about! This is my new go to salon! **Model Prediction:** 5-star (most positive) → 1-star (most negative)

Table 21: English Adversarial Examples Generated by `SemAttack` for BERT-based Classifier on SNLI Dataset using all perturbation functions.

Input (red = Modified character, **bold** = original character.)
Original Premise: Four boys are about to be **hit** by an approaching wave. **Adversarial Premise:** Four boys are about to be smashed by an approaching wave. **Hypothesis:** The wave missed the boys. **Model Prediction:** contradiction → entailment
Original Premise: A yellow race car **sliding** through a corner as spectators watch. **Adversarial Premise:** A yellow race car slipping through a corner as spectators watch. **Hypothesis:** A NASCAR is being watched. **Model Prediction:** neutral → entailment
Original Premise: A group of people on the bark, brightly lighten street, while one man with a gray hat holds a large colorful sign with **arrows**. **Adversarial Premise:** A group of people on the bark, brightly lighten street, while one man with a gray hat holds a large colorful sign with swords. **Hypothesis:** The people are walking down the street. **Model Prediction:** entailment → contradiction
Original Premise: A man takes a **drink** in the doorway of a home. **Adversarial Premise:** A man takes a drinking in the doorway of a home. **Hypothesis:** A man is looking out onto his front lawn from the doorway of his home. **Model Prediction:** neutral → contradiction
Original Premise: A dog **attacking** a man wearing protective gear. **Adversarial Premise:** A dog hurting a man wearing protective gear. **Hypothesis:** He was training a police dog. **Model Prediction:** neutral → entailment
Original Premise: A white man in a **red** shirt riding a bike. **Adversarial Premise:** A white man in a golden shirt riding a bike. **Hypothesis:** An old guy wears a shirt on a bike. **Model Prediction:** neutral → entailment
Original Premise: A child in a blue and white striped shirt crosses his arms and smiles while standing on red carpeted **stairs**. **Adversarial Premise:** A child in a blue and white striped shirt crosses his arms and smiles while standing on red carpeted terraces. **Hypothesis:** A child is smiling as he watches a clown. **Model Prediction:** neutral → contradiction
Original Premise: This man, with a red & white shirt has **water** bottles on this white truck. **Adversarial Premise:** This man, with a red & white shirt has beer bottles on this white truck. **Hypothesis:** The guy has bottles on the truck for me. **Model Prediction:** neutral → entailment
Original Premise: Three people are riding a carriage **pulled** by four horses. **Adversarial Premise:** Three people are riding a carriage hauled by four horses. **Hypothesis:** The oxen are pulling the carriage. **Model Prediction:** contradiction → entailment

H.4 Chinese Adversarial Examples

Table 22: Chinese Adversarial Examples Generated by `SemAttack` for BERT-based Classifier on THUNews Dataset using $\mathcal{F}_T$.

Input (red = Modified character, **bold**=original character.)
Original Chinese Text: 高露洁新品专效抗敏牙膏解决牙齿过敏 **Translation:** Gaolu**jie**'s new anti-hypersensitive toothpaste solves tooth hypersensitivity **Adversarial Chinese Text:** 高露吉新品专效抗敏牙膏解决牙齿过敏 **Translation:** Gaoluji's new anti-hypersensitive toothpaste solves tooth hypersensitivity **Model Prediction:** Fashion News （时尚新闻）→ Entertainment News （娱乐新闻）
Original Chinese Text: 组图：09巴黎高级定制秀最有看点8场次 **Translation:** Photos: 8 **highlights** of 09 Paris Haute Couture Show **Adversarial Chinese Text:** 组图：09巴黎高级定制秀最有看点8场炊 **Translation:** Photos: 8 cooking sessions of 09 Paris Haute Couture Show **Model Prediction:** Fashion News （时尚新闻）→ Entertainment News （娱乐新闻）
Original Chinese Text: 今秋男友新标准打造新时代型男 **Translation:** New standards for boyfriends in **this** autumn to create a new era of men **Adversarial Chinese Text:** 金秋男友新标准打造新时代型男 **Translation:** New standards for boyfriends in golden autumn to create a new era of men **Model Prediction:** Fashion News （时尚新闻）→ Entertainment News （娱乐新闻）
Original Chinese Text: 据称台联党可能下令赖幸媛辞去陆委会主委 **Translation: It is** said that the Taiwan Union Party may order Lai Xingyuan to resign as chairman of the MAC **Adversarial Chinese Text:** 剧称台联党可能下令赖幸媛辞去陆委会主委 **Translation:** The drama said that the Taiwan Union Party may order Lai Xingyuan to resign as chairman of the MAC **Model Prediction:** Politics news （时政新闻）→ Entertainment News （娱乐新闻）
Original Chinese Text: 猛犸象80%基因组破译完成史前巨兽有望复活 **Translation: Mammoth** 80% genome deciphered complete prehistoric behemoth is expected to be resurrected **Adversarial Chinese Text:** 孟犸象80%基因组破译完成史前巨兽有望复活 **Translation:** Mammoth 80% genome deciphered complete prehistoric behemoth is expected to be resurrected **Model Prediction:** Technology News （科技新闻）→ Entertainment News （娱乐新闻）

Table 23: Chinese Adversarial Examples Generated by `SemAttack` for BERT-based Classifier on THUNews Dataset using $\mathcal{F}_K$.

Input (red = Modified character, **bold**=original character.)
Original Chinese Text: 手袋进阶论：职场之路的秘密奠基石（组图） **Translation:** Handbag progression theory: the **secret** cornerstone of the road to the workplace (photo) **Adversarial Chinese Text:** 手袋进阶论：职场之路的机密奠基石（组图） **Translation:** Handbag progression theory: the confidential cornerstone of the road to the workplace (photo) **Model Prediction:** Fashion News （时尚新闻）→ Technology News （科技新闻）
Original Chinese Text: 中国银联发布十一黄金周用卡提示 **Translation:** China UnionPay releases card **tips** for Golden Week. **Adversarial Chinese Text:** 中国银联发布十一黄金周用卡提醒 **Translation:** China UnionPay releases card reminders for Golden Week. **Model Prediction:** Financial and economic news （财经新闻）→ Technology News （科技新闻）
Original Chinese Text: 买卖红木都是一项风险活 **Translation:** Buying and selling mahogany is a **risky** business. **Adversarial Chinese Text:** 买卖红木都是一项危险活 **Translation:** Buying and selling mahogany is a dangerous business. **Model Prediction:** Financial and economic news （财经新闻）→ Home News （家居新闻）
Original Chinese Text: 信用卡利润猛涨风险容忍度提高 **Translation:** Credit card profits soar with **increased** risk tolerance. **Adversarial Chinese Text:** 信用卡利润猛涨风险容忍度提升 **Translation:** Credit card profits soar with increased risk tolerance. **Model Prediction:** Financial and economic news （财经新闻）→ Stock News （股票新闻）
Original Chinese Text: 黎振伟：不同的城市有着各自的发展模式 **Translation:** Zhenwei Li: Different **cities** have their own development models. **Adversarial Chinese Text:** 黎振伟：不同的都市有着各自的发展模式 **Translation:** Zhenwei Li: Different cities have their own development models. **Model Prediction:** Real Estate News （房产新闻）→ Technology News （科技新闻）
Original Chinese Text: 韩国航空试验中心揭秘：战斗机被冰冻住测试 **Translation:** South Korea's aviation **experiment** center revealed: fighter jets were frozen in the test. **Adversarial Chinese Text:** 韩国航空检验中心揭秘：战斗机被冰冻住测试 **Translation:** South Korea's aviation test center revealed: fighter jets were frozen in the test. **Model Prediction:** Technology News （科技新闻）→ Current Affairs News （时政新闻）

Table 24: Chinese Adversarial Examples Generated by `SemAttack` for BERT-based Classifier on THUNews Dataset using $\mathcal{F}_C$.

Input (red = Modified character, **bold**=original character.)
Original Chinese Text: 高露洁新品专效抗敏牙膏解决牙齿过敏 **Translation:** Gaolu**jie**'s new anti-hypersensitive toothpaste solves tooth hypersensitivity **Adversarial Chinese Text:** 高露婕新品专效抗敏牙膏解决牙齿过敏 **Translation:** Gaolujie's new anti-hypersensitive toothpaste solves tooth hypersensitivity **Model Prediction:** Fashion News （时尚新闻）→ Entertainment News （娱乐新闻）
Original Chinese Text: 实录：张瑜阿穆隆王睿做客聊新片《八十一格》 **Translation: Record**: Zhang Yu, Amulon and Wang Rui as a guest to talk about the new film "Eighty-one Patterns" **Adversarial Chinese Text:** 实摄：张瑜阿穆隆王睿做客聊新片《八十一格》 **Translation:** Record: Zhang Yu, Amulon and Wang Rui as a guest to talk about the new film "Eighty-one Patterns" **Model Prediction:** Entertainment News （娱乐新闻）→ Technology News （科技新闻）
Original Chinese Text: 聚焦信用卡全额罚息：欠款44.6 元生千元利息 **Translation: Focus** on credit card full penalty interest: RMB 44.6 arrears generate interest of RMB 1,000 **Adversarial Chinese Text:** 聚盯信用卡全额罚息：欠款44.6 元生千元利息 **Translation:** Focus on credit card full penalty interest: RMB 44.6 arrears generate interest of RMB 1,000 **Model Prediction:** Financial and economic news （财经新闻）→ Technology News （科技新闻）
Original Chinese Text: 研究发现4000万年前鲸鱼长有4条腿（图） **Translation:** Research found that **whales** had 4 legs 40 million years ago (photo) **Adversarial Chinese Text:** 研究发现4000万年前鲤鱼长有4条腿（图） **Translation:** Research found that carp had 4 legs 40 million years ago (photo) **Model Prediction:** Technology News （科技新闻）→ Social News （社会新闻）
Original Chinese Text: 澳门博彩业后何鸿时代猜想 **Translation:** Post-Ho Hong Era **Conjecture** in Macau's Gaming Industry **Adversarial Chinese Text:** 澳门博彩业后何鸿时代预想 **Translation:** Post-Ho Hong Era Prediction in Macau's Gaming Industry **Model Prediction:** Stock news （股票新闻）→ Technology News （科技新闻）

Table 25: Chinese Adversarial Examples Generated by `SemAttack` for BERT-based Classifier on THUNews Dataset using all perturbation functions.

Input (red = Modified character, **bold**=original character.)
Original Chinese Text: 对话王辉灏：海归创业面临的困难（图） **Translation:** Dialogue with Wang **Huihao**: Difficulties faced by overseas returnees in starting a business (photo) **Adversarial Chinese Text:** 对话王辉耀：海归创业面临的困难（图） **Translation:** Dialogue with Wang Huiyao: Difficulties faced by overseas returnees in starting a business (photo) **Model Prediction:** Education News （教育新闻）→ Entertainment News （娱乐新闻）
Original Chinese Text: 拿什么能吸引你：我们的海外学子? **Translation:** **What** can attract you: our overseas students? **Adversarial Chinese Text:** 拿甚么能吸引你：我们的海外学子? **Translation:** What can attract you: our overseas students? **Model Prediction:** Education News （教育新闻）→ Entertainment News （娱乐新闻）
Original Chinese Text: 独家对话冯小刚：多个观众挺难少点观众挺容易 **Translation:** Exclusive dialogue with **Feng** Xiaogang: It's difficult for multiple audiences, and it's easy for less audiences **Adversarial Chinese Text:** 独家对话郜小刚：多个观众挺难少点观众挺容易 **Translation:** Exclusive dialogue with Gao Xiaogang: It's difficult for multiple audiences, and it's easy for less audiences **Model Prediction:** Entertainment News （娱乐新闻）→ Sports News （体育新闻）
Original Chinese Text: 高露洁新品专效抗敏牙膏解决牙齿过敏 **Translation:** Gaolujie's new anti-hypersensitive toothpaste solves **tooth** hypersensitivity **Adversarial Chinese Text:** 高露洁新品专效抗敏牙膏解决牙苔过敏 **Translation:** Gaolujie's new anti-hypersensitive toothpaste solves tooth coating hypersensitivity **Model Prediction:** Fashion News （时尚新闻）→ Entertainment News （娱乐新闻）
Original Chinese Text: 2010 艺术品秋拍上演六宗最 **Translation:** Six **most** of the 2010 art autumn auctions **Adversarial Chinese Text:** 2010 艺术品秋拍上演六综最 **Translation:** Six most comprehensive of the 2010 art autumn auctions **Model Prediction:** Financial and economic news （财经新闻）→ Entertainment News （娱乐新闻）
Original Chinese Text: 英属小岛发现罕见蓝色龙虾（组图） **Translation:** Rare blue **lobster** found on British island (photo) **Adversarial Chinese Text:** 英属小岛发现罕见蓝色龙鳖（组图） **Translation:** Rare blue turtle found on British island (photo) **Model Prediction:** Technology News （科技新闻）→ Social News （社会新闻）

Table 26: Chinese Adversarial Examples Generated by SemAttack for BERT-based Classifier on Wechat Finance Dataset using all perturbation functions.

Input (red = Modified character, **bold**=original character.)
Original Chinese Text: 翻倍网分享财富资产管理资讯知识技巧。关注信托、融资租赁、期货保险、私人银行等领域最新信息。 **Translation:** Fanbei.com shares wealth and asset management information knowledge and skills. Pay attention to the latest information in the fields of trust, financial leasing, **futures** insurance, and private banking. **Adversarial Chinese Text:** 翻倍网分享财富资产管理资讯知识技巧。关注信托、融资租赁、期祸保险、私人银行等领域最新信息。 **Translation:** Fanbei.com shares wealth and asset management information knowledge and skills. Pay attention to the latest information in the fields of trust, financial leasing, accident insurance, and private banking. **Model Prediction:** Comprehensive （综合）→ Bank（银行）
Original Chinese Text: 温泉邮政支局提供邮政服务、个性化邮票订制、快递小包上门取件、邮件查询。 **Translation:** The Post Office at Hot Spring **Branch** provides postal services, personalized stamp ordering, home delivery of small parcels, and mail inquiries. **Adversarial Chinese Text:** 温泉邮政驿局提供邮政服务、个性化邮票订制、快递小包上门取件、邮件查询。 **Translation:** The Hot Spring Post Office provides postal services, personalized stamp ordering, home delivery of small parcels, and mail inquiries. **Model Prediction:** Bank（银行）→ Insurance（保险）
Original Chinese Text: 中融华创（北京）基金有限公司（简称：中融华创）成立于2012 年3 月29 日。总部设立在首都北京，公司在国家发展改革委员会登记备案，由中国证券投资基金协会颁发金融牌照。 **Translation: Zhongrong** Huachuang (Beijing) Fund Co., Ltd. (abbreviated as Zhongrong Huachuang) was established on March 29, 2012. Headquartered in the capital, Beijing, the company is registered with the National Development and Reform Commission, and is a legal financial institution that is issued a financial license by the Securities Investment Fund Association of China. **Adversarial Chinese Text:** 申融华创（北京）基金有限公司（简称：中融华创）成立于2012 年3 月29 日。总部设立在首都北京，公司在国家发展改革委员会登记备案，由中国证券投资基金协会颁发金融牌照。 **Translation:** Shenrong Huachuang (Beijing) Fund Co., Ltd. (abbreviated as Zhongrong Huachuang) was established on March 29, 2012. Headquartered in the capital, Beijing, the company is registered with the National Development and Reform Commission, and is a legal financial institution that is issued a financial license by the Securities Investment Fund Association of China. **Model Prediction:** Fund （基金）→ Comprehensive （综合）
Original Chinese Text: 期货行业风起云涌，期市行情熟悉万变。交易帮玩转交易，携手众多期货高手，让交易更简单！ **Translation:** The **futures** industry is surging, and the futures market is familiar with ever-changing conditions. Trading helps fun trading, and join hands with many futures experts to make trading easier! **Adversarial Chinese Text:** 期券行业风起云涌，期市行情熟悉万变。交易帮玩转交易，携手众多期货高手，让交易更简单！ **Translation:** The futures bond industry is surging, and the futures market is familiar with ever-changing conditions. Trading helps fun trading, and join hands with many futures experts to make trading easier! **Model Prediction:** Futures （期货）→ Comprehensive （综合）
Original Chinese Text: 瑞倪资本专注于股权投资、证券投资及衍生品研究等领域，业务涵盖一、二级市场，包括天使投资以及对冲型、权益类与固定收益类证券投资。 **Translation:** Rui**ni** Capital focuses on equity investment, securities investment and derivatives research and other fields. Its business covers primary and secondary markets, including angel investment and hedging, equity and fixed income securities investment. **Adversarial Chinese Text:** 瑞券资本专注于股权投资、证投资及衍生品研究等领域，业务涵盖一、二级市场，包括天使投资以及对冲型、权益类与固定收益类证券投资。 **Translation:** Ruiquan Capital focuses on equity investment, securities investment and derivatives research and other fields. Its business covers primary and secondary markets, including angel investment and hedging, equity and fixed income securities investment. **Model Prediction:** Comprehensive （综合）→ Securities （证券）