Contextualized Perturbation for Textual Adversarial Attack

Dianqi Li[♠] **Yizhe Zhang**[◇] **Hao Peng**[♠] **Liqun Chen**[♣]
Chris Brockett[◇] **Ming-Ting Sun**[♠] **Bill Dolan**[◇]
[♠]University of Washington [◇]Microsoft Research [♣]Duke University
`{dianqili, mts}@uw.edu, hapeng@cs.uw.edu, liqun.chen@duke.edu`
`{Yizhe.Zhang, Chris.Brockett, billdol}@microsoft.com`

Abstract

Adversarial examples expose the vulnerabilities of natural language processing (NLP) models, and can be used to evaluate and improve their robustness. Existing techniques of generating such examples are typically driven by local heuristic rules that are agnostic to the context, often resulting in unnatural and ungrammatical outputs. This paper presents CLARE, a **C**ontextua**L**ized **A**dversa**R**ial **E**xample generation model that produces fluent and grammatical outputs through a mask-then-infill procedure. CLARE builds on a pre-trained masked language model and modifies the inputs in a context-aware manner. We propose three contextualized perturbations, *Replace*, *Insert* and *Merge*, that allow for generating outputs of varied lengths. CLARE can flexibly combine these perturbations and apply them at any position in the inputs, and is thus able to attack the victim model more effectively with fewer edits. Extensive experiments and human evaluation demonstrate that CLARE outperforms the baselines in terms of attack success rate, textual similarity, fluency and grammaticality.

1 Introduction

Adversarial example generation for natural language processing (NLP) tasks aims to perturb input text to trigger errors in machine learning models, while keeping the output close to the original. Besides exposing system vulnerabilities and helping improve their robustness and security (Zhao et al., 2018; Wallace et al., 2019; Cheng et al., 2019; Jia et al., 2019, *inter alia*), adversarial examples are also used to analyze and interpret the models' decisions (Jia and Liang, 2017; Ribeiro et al., 2018).

Generating adversarial examples for NLP tasks can be challenging, in part due to the discrete nature of natural language text. Most recent efforts have explored heuristic rules, such as replacing tokens with their synonyms (Samanta and Mehta, 2017; Liang et al., 2019; Alzantot et al., 2018; Ren et al., 2019; Jin et al., 2020, *inter alia*). Despite some empirical success, rule-based methods are agnostic to context, limiting their ability to produce natural, fluent, and grammatical outputs (Wang et al., 2019b; Kurita et al., 2020, *inter alia*).

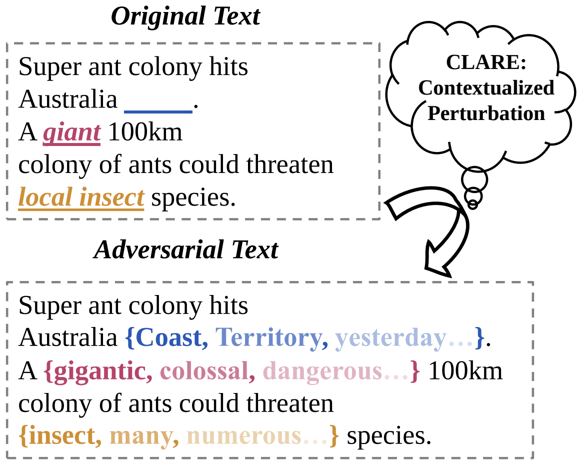

Original Text
Super ant colony hits Australia _____.
A *giant* 100km colony of ants could threaten *local insect* species.
CLARE: Contextualized Perturbation
Adversarial Text
Super ant colony hits Australia {Coast, Territory, yesterday…}.
A {gigantic, colossal, dangerous…} 100km colony of ants could threaten {insect, many, numerous…} species.

Figure 1: Illustration of CLARE. Through a mask-then-infill procedure, the model generates the adversarial text with three contextualized perturbations: ***Replace***, ***Insert*** and ***Merge***. A mask is indicated by "___". The degree of fade corresponds to the (decreasing) priority of the infill tokens.

This work presents CLARE, a **C**ontextua**L**ized **A**dversa**R**ial **E**xample generation model for text. CLARE perturbs the input with a mask-then-infill procedure: it first detects the vulnerabilities of a model and deploys masks to the inputs to indicate missing text, then plugs in an alternative using a pretrained masked language model (e.g., RoBERTa; Liu et al., 2019). CLARE features three contextualized perturbations: *Replace*, *Insert* and *Merge*, which respectively replace a token, insert a new one, and merge a bigram (Figure 1). As a result, it can generate outputs of varied lengths, in contrast to token replacement based methods that are limited to outputs of the same lengths as the inputs (Alzantot et al., 2018; Ren et al., 2019;

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5053–5069
June 6–11, 2021. ©2021 Association for Computational Linguistics

Jin et al., 2020). Further, CLARE searches over a wider range of attack strategies, and is thus able to attack the victim model more effectively with fewer edits. Building on a masked language model, CLARE maximally preserves textual similarity, fluency, and grammaticality of the outputs.

We evaluate CLARE on text classification, natural language inference, and sentence paraphrase tasks, by attacking finetuned BERT models (Devlin et al., 2019). Extensive experiments and human evaluation results show that CLARE outperforms baselines in terms of attack success rate, textual similarity, fluency, and grammaticality, and strikes a better balance between attack success rate and preserving input-output similarity. Our analysis further suggests that the CLARE can be used to improve the robustness of the downstream models, and improve their accuracy when the available training data is limited. We release our code and models at `https://github.com/cookielee77/CLARE`.

2 CLARE

At a high level, CLARE applies a sequence of contextualized perturbation actions to the input. Each can be seen as a *local* mask-then-infill procedure: it first applies a mask to the input around a given position, and then fills it in using a pretrained masked language model (§2.1). To produce the output, CLARE scores and descendingly ranks the actions, which are then iteratively applied to the input (§2.2). We begin with a brief background review and laying out of necessary notation.

Background. Adversarial example generation centers around a **victim** model f, which we assume is a text classifier. We focus on the black-box setting, allowing access to f's outputs but *not* its configurations such as parameters. Given an input sequence $\mathbf{x} = x_1 x_2 \ldots x_n$ and its label y (assume $f(\mathbf{x}) = y$), an **adversarial example** $\mathbf{x}'$ is supposed to modify $\mathbf{x}$ to trigger an error in the victim model: $f(\mathbf{x}') \neq f(\mathbf{x})$. At the same time, textual modifications should be minimal, such that $\mathbf{x}'$ is close to $\mathbf{x}$ and the human predictions on $\mathbf{x}'$ stay the same.[1]

This is achieved by requiring the similarity between $\mathbf{x}'$ and $\mathbf{x}$ to be larger than a threshold: $\operatorname{sim}(\mathbf{x}', \mathbf{x}) > \ell$. A common choice of $\operatorname{sim}(\cdot, \cdot)$ is to encode sentences using neural networks, and calculate their cosine similarity in the embedding space (Jin et al., 2020).

2.1 Masking and Contextualized Infilling

At a given position of the input sequence, CLARE can execute three perturbation actions: *Replace*, *Insert*, and *Merge*, which we introduce in this section. These apply masks at the given position with different strategies, and then fill in the missing text based on the unmasked context.

***Replace*:** A *Replace* action substitutes the token at a given position i with an alternative (e.g., changing "*fantastic*" to "*amazing*" in "The movie is *fantastic*."). It first replaces x_i with a mask, and then selects a token z from a candidate set $\mathcal{Z}$ to fill in:

$$\begin{aligned}\widetilde{\mathbf{x}} &= x_1 \ldots x_{i-1} \; [\textsc{Mask}] \; x_{i+1} \ldots x_n, \\ \operatorname{replace}(\mathbf{x}, i) &= x_1 \ldots x_{i-1} \; z \; x_{i+1} \ldots x_n.\end{aligned}$$

For clarity, we denote $\operatorname{replace}(\mathbf{x}, i)$ by $\widetilde{\mathbf{x}}_z$. To produce an adversarial example,

- z should fit into the unmasked context;
- $\widetilde{\mathbf{x}}_z$ should be similar to $\mathbf{x}$;
- $\widetilde{\mathbf{x}}_z$ should trigger an error in f.

These can be achieved by selecting a z such that

- z receives a high probability from a masked language model: $p_{\text{MLM}}(z \mid \widetilde{\mathbf{x}}) > k$;
- $\widetilde{\mathbf{x}}_z$ is similar to $\mathbf{x}$: $\operatorname{sim}(\mathbf{x}, \widetilde{\mathbf{x}}_z) > \ell$;
- f predicts low probability for the gold label given $\widetilde{\mathbf{x}}_z$, i.e., $p_f(y \mid \widetilde{\mathbf{x}}_z)$ is small.

p_{MLM} denotes a pretrained masked language model (e.g., RoBERTa; Liu et al., 2019). Using higher k, ℓ thresholds produces outputs that are more fluent and closer to the original. However, this can undermine the success rate of the attack. We choose k, ℓ to trade-off between these two aspects. [2]

The first two requirements can be met by the construction of the candidate set: $\mathcal{Z} =$

$$\left\{ z' \in \mathcal{V} \mid p_{\text{MLM}}(z' \mid \widetilde{\mathbf{x}}) > k, \operatorname{sim}(\mathbf{x}, \widetilde{\mathbf{x}}_{z'}) > \ell \right\}.$$

$\mathcal{V}$ is the vocabulary of the masked language model. To meet the third, we select from $\mathcal{Z}$ the token that, if filled in, will cause most "confusion" to f:

$$z = \arg\min_{z' \in \mathcal{Z}} p_f(y \mid \widetilde{\mathbf{x}}_{z'}). \tag{1}$$

[1] In computer vision applications, minor perturbations to continuous pixels can be barely perceptible to humans, thus it can be hard for one to distinguish $\mathbf{x}$ and $\mathbf{x}'$ (Goodfellow et al., 2015). It is not the case for text, however, since changes to the discrete tokens are more likely to be noticed by humans.

[2] k and ℓ are empirically set as 5×10^{-3} and 0.7, respectively. This also reduces the computation overhead: in our experiments $|\mathcal{Z}|$ is 42 on average, much smaller than the vocabulary size ($|\mathcal{V}| = 50,265$).

The *Insert* and *Merge* actions differ from *Replace* in terms of masking strategies. The alternative token z is selected analogously to that in a *Replace* action.

***Insert*:** This aims to add extra information to the input (e.g., changing "I recommend ..." to "I *highly* recommend ..."). It inserts a mask after x_i and then fills it. Slightly overloading the notations,

$$\begin{aligned}\widetilde{\mathbf{x}} &= x_1 \dots x_i \text{ [MASK] } x_{i+1} \dots x_n, \\ \text{insert}\,(\mathbf{x}, i) &= x_1 \dots x_i \; z \; x_{i+1} \dots x_n.\end{aligned}$$

This increases the sequence length by 1.

***Merge*:** This masks out a bigram $x_i x_{i+1}$ with *a single* mask and then fills it, reducing the sequence length by 1:

$$\begin{aligned}\widetilde{\mathbf{x}} &= x_1 \dots x_{i-1} \text{ [MASK] } x_{i+2} \dots x_n, \\ \text{merge}\,(\mathbf{x}, i) &= x_1 \dots x_{i-1} \; z \; x_{i+2} \dots x_n.\end{aligned}$$

z can be the same as one of the masked tokens (e.g., masking out "New York" and then filling in"York"). This can be seen as deleting a token from the input.

For *Insert* and *Merge*, z is chosen in the same manner as replace action. [3]

In sum, at each position i of an input sequence, CLARE first: (i) replaces x_i with a mask; (ii) or inserts a mask after x_i; (iii) or merges $x_i x_{i+1}$ into a mask. Then a set of candidate tokens is constructed with a masked language model and a textual similarity function; the token minimizing the gold label's probability is chosen as the alternative token. The combination of these three operations enables conversion between any two sequences.

CLARE first constructs the local actions for all positions in parallel, i.e., the actions at position i do not affect those at other positions. Then, to produce the adversarial example, CLARE gathers the local actions and selects an order to execute them.

2.2 Sequentially Applying the Perturbations

Given an input pair $(\mathbf{x}, y)$, let n denote the length of $\mathbf{x}$. CLARE chooses from $3n$ actions to produce the output: 3 actions for each position, assuming the candidate token sets are not empty. We aim to generate an adversarial example with minimum modifications to the input. To achieve this, we iteratively apply the actions, and first select those minimizing the probability of outputting the gold label y from f.

Algorithm 1 Adversarial Attack by CLARE

```
Input: Text-label pair $(\mathbf{x}, y)$; Victim model $f$
Output: An adversarial example
Initialization: $\mathbf{x}^{(0)} = \mathbf{x}$
$\mathcal{A} \leftarrow \varnothing$
for $1 \le i \le |\mathbf{x}|$ do
    $a \leftarrow$ highest-scoring action from $\{$
        $\text{replace}(\mathbf{x}, i), \text{insert}(\mathbf{x}, i), \text{merge}(\mathbf{x}, i)\}$
    $\mathcal{A} \leftarrow \mathcal{A} \bigcup \{a\}$
end for
for $1 \le t \le T$ do
    $a \leftarrow$ highest-scoring action from $\mathcal{A}$
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{a\}$
    $\mathbf{x}^{(t)} \leftarrow$ Apply $a$ on $\mathbf{x}^{(t-1)}$
    if $f(\mathbf{x}^{(t)}) \neq y$ then return $\mathbf{x}^{(t)}$
    end if
end for
return NONE
```

Each action is associated with a score, measuring how likely it can "confuse" f: denote by $a(\mathbf{x})$ the output of applying action a to $\mathbf{x}$. The score is then the negative probability of predicting the gold label from f, using $a(\mathbf{x})$ as the input:

$$s_{(\mathbf{x},y)}(a) = -p_f\big(y \mid a(\mathbf{x})\big).$$

Only one of the three actions can be applied at each position, and we select the one with the highest score. This constraint aims to avoid multiple modifications around the same position, e.g., merging "New York" into "Seattle" and then replacing it with "Boston".

Actions are iteratively applied to the input, until an adversarial example is found or a limit of actions T is reached. Each step selects the highest-scoring action from the remaining ones. Algorithm 1 summarizes the above procedure.[4]

Discussion. A key technique of CLARE is the local mask-then-infill perturbation. Compared with existing context-agnostic replacement approaches (Alzantot et al., 2018; Jin et al., 2020; Ren et al., 2019, *inter alia*), contextualized infilling produces more fluent and grammatical outputs. Generating adversarial examples with masked language models is also explored by concurrent work

[3] A perturbation will not be considered if its candidate token set is empty.

[4] *Insert* and *Merge* actions change the text length. When any of them is applied, we accordingly change the text indices of affected actions remaining in $\mathcal{A}$.

BERTAttack (Li et al., 2020) and BAE (Garg and Ramakrishnan, 2020).[5]

- BERTAttack only replaces tokens and thus can only produce outputs of the same lengths as the inputs. This is analogous with a CLARE model with the *Replace* action only. BAE entangles replacing and inserting tokens: it inserts *only* at positions neighboring a replaced token, limiting its attacking capability. Departing from both, CLARE uses three different perturbations (*Replace*, *Insert* and *Merge*), each allowing efficient attacking against *any* position of the input, and can produce outputs of varied lengths. As we will show in the experiments (§3.3), CLARE outperforms both these methods.
- When selecting the attack positions, neither BERTAttack or BAE takes into account the tokens to be infilled, whereas CLARE does. This results in better adversarial attack performance according to our ablation study (§4.1).
- CLARE demonstrates the advantage of using RoBERTa over BERT, which was used in the concurent works (§4.1).

3 Experiments

We evaluate CLARE on text classification, natural language inference, and sentence paraphrase tasks. We begin by describing the implementation details of CLARE and the baselines (§3.1). §3.2 introduces the experimental datasets and the evaluation metrics; the results are summarized in §3.3.

3.1 Setup

- We experiment with a distilled version of RoBERTa (RoBERTa$_{\text{distill}}$; Sanh et al., 2019) as the masked language model for contextualized infilling. We also compare to base sized RoBERTa (RoBERTa$_{\text{base}}$; Liu et al., 2019) and base sized BERT (BERT$_{\text{base}}$; Devlin et al., 2019) in the ablation study (§4.1).
- The similarity function builds on the universal sentence encoder (USE; Cer et al., 2018).
- The victim model is an MLP classifier on top of BERT$_{\text{base}}$. It takes as input the first token's contextualized representation. We finetune BERT when training the victim model.

Dataset	Avg. Length	# Classes	Train	Test	Acc
Yelp	130	2	560K	38K	95.9%
AG News	46	4	120K	7.6K	95.0%
MNLI[6]	23/11	3	392K	9.8K	84.3%
QNLI	11/31	2	105K	5.4K	91.4%

Table 1: Some statistics of datasets. The last column indicates the victim model's accuracy on the original test set *without* adversarial attack.

Baselines. We compare CLARE with recent state-of-the-art word-level black-box adversarial attack models, including:

- **TextFooler**: a state-of-the-art model by Jin et al. (2020). This replaces tokens with their synonyms derived from counter-fitting word embeddings (Mrkšić et al., 2016), and uses the same text similarity function as our work.
- **TextFooler+LM**: an improved variant of TextFooler we implemented based on Alzantot et al. (2018) and Cheng et al. (2019). This inherits token replacement from TextFooler, but uses an additional small sized GPT-2 language model (Radford et al., 2019) to filter out those candidate tokens that do not fit in the context with calculated perplexity.
- **BERTAttack**: a mask-then-infill approach by Li et al. (2020). It greedily replaces tokens with the predictions from BERT. BAE is not listed as it has a similar performance as BERTAttack (Garg and Ramakrishnan, 2020).

We use the open source implementation of the above baselines provided by the authors. More details are included in Appendix §A.1.

3.2 Datasets and Evaluation

Datasets. We evaluate CLARE with the following datasets:

- **Yelp Reviews** (Zhang et al., 2015): a binary sentiment classification dataset based on restaurant reviews.
- **AG News** (Zhang et al., 2015): a collection of news articles with four categories: *World*, *Sports*, *Business* and *Science & Technology*.
- **MNLI** (Williams et al., 2018): a natural language inference dataset. Each instance consists of a premise-hypothesis pair, and the model is supposed to determine the relation between them from a label set of *entailment*, *neutral*, and *contradiction*. It covers text from a variety of domains.

[5] Both Li et al. (2020) and Garg and Ramakrishnan (2020) are published concurrently to an initial report of this work.

[6] We only examine the performance on the matched set, since the mismatched set is easier to attack.

	Yelp (PPL = 51.5)					**AG News** (PPL = 62.8)				
Model	**A-rate**↑	**Mod**↓	**PPL**↓	**GErr**↓	**Sim**↑	**A-rate**↑	**Mod**↓	**PPL**↓	**GErr**↓	**Sim**↑
TextFooler	77.0	16.6	163.3	1.23	0.70	56.1	23.3	331.3	1.43	0.69
+ LM	34.0	17.4	90.0	1.21	0.73	23.1	21.9	144.6	1.07	0.74
BERTAttack	71.8	10.7	90.8	0.27	0.72	63.4	7.9	90.6	0.25	0.71
CLARE	**79.7**	**10.3**	**83.5**	**0.25**	**0.78**	**79.1**	**6.1**	**86.0**	**0.17**	**0.76**

	MNLI (PPL = 60.9)					**QNLI** (PPL = 46.0)				
Model	**A-rate**↑	**Mod**↓	**PPL**↓	**GErr**↓	**Sim**↑	**A-rate**↑	**Mod**↓	**PPL**↓	**GErr**↓	**Sim**↑
TextFooler	59.8	13.8	161.5	0.63	0.73	57.8	16.9	164.4	0.62	0.72
+ LM	32.3	12.4	91.9	0.50	0.77	29.2	17.3	85.0	0.42	0.75
BERTAttack	82.7	8.4	86.7	0.04	0.77	76.7	13.3	86.5	0.03	0.73
CLARE	**88.1**	**7.5**	**82.7**	**0.02**	**0.82**	**83.8**	**11.8**	**76.7**	**0.01**	**0.78**

Table 2: Adversarial example generation performance in attack success rate (A-rate), modification rate (Mod), perplexity (PPL), number of increased grammar errors (GErr), and textual similarity (Sim). The perplexity of the original inputs is indicated in parentheses for each dataset. Bold font indicates the best performance for each metric. All numbers are reported on 1000 test instances. ↑ (↓) represents that the higher (lower) the better.

- **QNLI** (Wang et al., 2019a): a binary classification dataset converted from the Stanford question answering dataset (Rajpurkar et al., 2016). The task is to determine whether the context contains the answer to a question. It is mainly based on English Wikipedia articles.

Table 1 summarizes some statistics of the datasets. In addition to the above four datasets, we experiment with DBpedia ontology dataset (Zhang et al., 2015), Stanford sentiment treebank (Socher et al., 2013), Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005), and Quora Question Pairs from the GLUE benchmark. The results on these datasets are summarized in Appendix A.2.

Following previous practice (Alzantot et al., 2018), we fine-tune CLARE on training data, and evaluate with 1,000 randomly sampled test instances of lengths $\leq$ 100. In the sentence-pair tasks (e.g., MNLI, QNLI), we attack the longer sentence excluding the tokens that appear in both.

Evaluation metrics. We follow previous works (Jin et al., 2020; Morris et al., 2020a), and evaluate the models with the following automatic metrics:

- **Attack success rate (A-rate)**: the percentage of adversarial examples that can successfully attack the victim model.
- **Modification rate (Mod)**: the percentage of modified tokens. Each *Replace* or *Insert* action accounts for one token modified; a *Merge* action is considered modifying one token if one of the two merged tokens is kept (e.g., merging bigram ab into a), and two otherwise (e.g., merging bigram ab into c).
- **Perplexity (PPL)**: a metric used to evaluate the *fluency* of adversaries (Kann et al., 2018; Zang et al., 2020). The perplexity is calculated using small sized GPT-2 with a 50K-sized vocabulary (Radford et al., 2019).
- **Grammar error (GErr)**: the absolute number of increased grammatical errors in the successful adversarial example, compared to the original text. Following (Zang et al., 2020; Morris et al., 2020b), we calculate this by the LanguageTool (Naber et al., 2003).[7]
- **Textual similarity (Sim)**: the cosine similarity between the input and its adversary. Following (Jin et al., 2020; Morris et al., 2020b), we calculate this using the universal sentence encoder (USE; Cer et al., 2018).

The last four metrics are averaged across those adversarial examples that successfully attack the victim model.

3.3 Results

Table 2 summarizes the results. Overall CLARE achieves the best performance on all metrics consistently across different datasets. Notably, CLARE outperforms BERTAttack, the strongest baseline, by a more than 5.4% attack success rate with *fewer* average modifications to the text. We attribute this to CLARE's flexible attack strategies obtained by combining three different perturbations at any position. Interestingly, using contextualized embeddings does *not* appear to guarantee better fluency:

[7] `https://www.languagetool.org/`

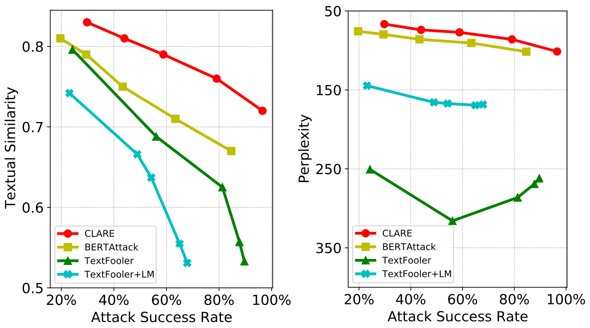


Figure 2: **Left**: Attack success rate and textual similarity trade-off curves (*both higher the better*). **Right**: Attack success rate (*higher the better*) and perplexity (*lower the better*) trade-off curve. The larger area under the two curves indicates the better trade-off between two metrics.

despite fewer modifications to the text, BERTAttack achieves similar perplexity to language-model-augmented TextFooler on three out of the four datasets, while CLARE consistently outperforms both. In terms of grammatical errors, contextualized models (CLARE and BERTAttack) are substantially better than the others, with CLARE performing the best. In terms of similarity, CLARE outperforms all baselines by more than 0.02, a larger gap than BERTAttack's improvements over TextFooler variants. We observe similar trends on other datasets in Appendix A.2.

Figure 2 compares trade-off curves between attack success rate and textual similarity. We tune the thresholds for constructing the candidate token sets, and plot textual similarity against the attack success rate. CLARE strikes the best balance, showing a clear advantage in success rate with least similarity drop. We observe similar trends for attack success rate and perplexity trade off.

Human evaluation. We further conduct human evaluation on the AG News dataset. We randomly sample 300 instances which both CLARE and TextFooler successfully attack. For each input, we pair the adversarial examples from the two models, and present them to crowd-sourced judges along with the original input and the gold label. We ask them which they prefer with a neutral option in terms of (1) having a meaning that is closer to the original input (similarity), and (2) being more fluent and grammatical (fluency and grammaticality). Additionally, we ask the judges to annotate adversarial examples, and compare their annotations against the gold labels (label consistency). We collect 5 responses for each pair on every evaluated aspect. Further details are in Appendix A.3.

Metric	CLARE	Neutral	TextFooler
Similarity	56.1$_{\pm 2.5}$	28.1	15.8$_{\pm 2.1}$
Fluency&Grammaticality	42.5$_{\pm 2.5}$	48.6	8.9$_{\pm 1.5}$
Label Consistency	68.0$_{\pm 2.4}$	-	70.1$_{\pm 2.5}$

Table 3: Human evaluation performance in percentage on the AG News dataset. ± indicates confidence intervals with a 95% confidence level.

As shown in Table 3, CLARE has a significant advantage over TextFooler: in terms of similarity 56% responses prefer CLARE, while 16% prefer TextFooler. The trend is similar for fluency & grammaticality (42% vs. 9%). This observation is consistent with results from automatic metrics. On label consistency, CLARE slightly underperforms TextFooler at 68% with a 95% condidence interval (CI) $(66\%, 70\%)$, versus 70% with a 95% CI $(68\%, 73\%)$. We attribute this to an inherent overlap of some categories in the AG News dataset, e.g., *Science & Technology* and *Business*, as evidenced by a 71% label consistency for original inputs.

Closing this section, Table 4 compares the adversarial examples generated by TextFooler and CLARE. More samples are listed in Appendix A.4.

4 Analysis

This section first conducts an ablation study (§4.1). We then explore CLARE's potential to be used to improve downstream models' robustness and accuracy in §4.3. In §4.2, we empirically observe that CLARE tends to attack noun and noun phrases.

4.1 Ablation Study

We ablate each component of CLARE to study its effectiveness. We evaluate on the 1,000 randomly selected AG news instances (§3.2). The results are summarized in Table 5.

We first investigate the performance of three perturbations when applied individually. Among three editing strategies, using INSERTONLY achieves the best performance, with REPLACEONLY coming a close second. MERGEONLY underperforms the other two, partly because the attacks are restricted to bigram noun phrases (§3.1). Combining all three perturbations, CLARE achieves the best performance with the least modifications.

[8] *Merge* perturbation can only merge noun phrases, extracted by the NLTK toolkit(`https://www.nltk.org/`). We find that this helps produce more grammatical outputs.

AG (Sci&Tech)	Sprint Corp. is in talks with Qualcomm Inc. about using a network the chipmaker is building to deliver live television to Sprint mobile phone customers.
TextFooler (Business)	Sprint *Corps*. is in talks with Qualcomm Inc. about *operated* a network the chipmaker is *consolidation* to *doing viva* television to Sprint mobile phone customers.
CLARE (Business)	Sprint Corp. is in talks with Qualcomm Inc. about using a network Qualcomm is building to deliver *cable* television to Sprint mobile phone customers.
MNLI (Neutral)	*Premise*: Let me try it. She began snapping her fingers and saying the word eagerly, but nothing happened. *Hypothesis*: She became frustrated when the spell didn't work.
TextFooler (Contradiction)	*Premise*: *Authorisation* me *attempting* it. She *triggered flapping* her *pinkies* and *said* the word eagerly, but nothing *arisen*. *Hypothesis*: She became frustrated when the spell didn't work.
CLARE (Contradiction)	*Premise*: Let me try it. She began snapping her fingers and saying the word eagerly, but nothing **unexpected** happened. *Hypothesis*: She became frustrated when the spell didn't work.

Table 4: Adversarial examples produced by different models. The gold label of the original is shown below the (bolded) dataset name. ***Replace***, **Insert** and **Merge** are highlighted in *italic red*, **bold blue** and sans serif yellow, respectively. (Best viewed in color).

To examine the efficiency of attacking order, we compare REPLACEONLY against BERTAttack. Notably, REPLACEONLY outperforms BERTAttack across the board. This is presumably because BERTAttack does not take into account the tokens to be infilled when selecting the attack positions.

We now turn to the two constraints imposed when constructing the candidate token set. Perhaps not surprisingly, ablating the textual similarity constraint (*w/o* sim $> l$) decreases textual similarity performance, but increases other aspects. Ablating the masked language model yields a better success rate, but much worse perplexity, grammaticality, and textual similarity.

Finally, we compare CLARE implemented with different masked language models. Table 6 summarizes the results. Overall, distilled RoBERTa achieves the fastest speed without losing performance. Since the victim model is based on BERT, we conjecture that it is less efficient to attack a model using its own information.

Module	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**
CLARE	79.1	**6.1**	**86.0**	0.17	0.76
MERGEONLY[8]	47.2	6.2	95.3	**0.08**	**0.79**
INSERTONLY	68.1	7.2	93.1	0.23	0.74
REPLACEONLY	66.7	7.7	85.6	0.10	0.72
BERTAttack	63.4	7.9	90.6	0.25	0.71
w/o sim $> \ell$	82.4	6.9	86.8	0.13	0.70
w/o $p_{\text{MLM}} > k$	**95.7**	6.8	162.8	0.71	0.61

Table 5: Ablation study results. "*w/o* sim $> \ell$" ablates the textual similarity constraint when constructing the candidate sets, while "*w/o* $p_{\text{MLM}} > k$" ablates the masked language model probability constraint.

MLM	**A-rate↑**	**Mod↓**	**PPL↓**	**Sim↑**	**Speed↑**
RoBERTa$_{\text{distill}}$	79.1	**6.1**	**86.0**	**0.76**	**0.14**
RoBERTa$_{\text{base}}$	**79.3**	6.3	88.9	0.75	0.07
BERT$_{\text{base}}$	78.4	8.3	95.2	0.71	0.06

Table 6: Results of CLARE implemented with different masked language models (MLM). **Speed** is measured by number of processed samples per second.

4.2 Perturbations by Part-of-speech Tags

In this section, we break down the adversarial attacks by part-of-speech (POS) tags in AG News dataset. We find that most of the adversarial attacks happen to nouns or noun phrases. Presumably, in many topic classification datasets, the prediction heavily relies on some characteristic noun words/phrases. As shown in Table 7, 64% of the *Replace* actions are applied to nouns. *Insert* actions tend to insert tokens into noun phrase bigram: two of the most frequent POS bigrams are noun phrases. In fact, around 48% of the *Insert* actions are applied to noun phrases. This also justifies our choice of only applying *Merge* to noun phrases.

4.3 Adversarial Training

This section explores CLARE's potential in improving downstream models' accuracy and robustness. Following Tsipras et al. (2018), we use CLARE to generate adversarial examples for AG news training instances, and include them as additional training data. We consider two settings: training with (1) full training data and full adversarial data and (2) 10% randomly-sampled training data and its adversarial data, to simulate the low-resource scenario. For both settings, we compare a BERT-based MLP classifier and a TextCNN (Kim, 2014) classifier without any pretrained embedding.

Whether adversarial examples, as data augmentation, can help achieve better test accuracy? As

Replace	*Insert*	*Merge*
NOUN: 64%	*(NOUN, NOUN): 12%*	*ADJ-NOUN: 31%*
ADJ: 17%	*(ADJ, NOUN): 10%*	*NOUN-NOUN: 22%*
VERB: 7%	*(NOUN, VERB): 9%*	*DT-NOUN: 12%*

Context: ... Amit Yoran, the government's *cybersecurity* chief, abruptly resigned yesterday after a year ...

Replace: cybersecurity ← *{security, surveillance, cryptography, intelligence, encryption ...}*

Insert: cybersecurity ___ chief ← *{technology, defense, intelligence, program, project ...}*

Merge: cybersecurity chief ← *{chief, consultant, administrator, scientist, secretary ...}*

Table 7: **Top**: Top-3 POS tags (or POS tag bigrams) and their percentages for each perturbation type. (a, b): insert a token between a and b. a-b: merge a and b into a token. **Bottom**: An AG news sample, where CLARE perturbs token "*cybersecurity*." TextFooler is unable to attack this token since it is out of its vocabularies.

Victim Model	Acc↑	A-rate↓	Mod↑
BERT (100% data)	95.0	79.1	6.1
+ 100% adversarial	-0.2	-18.0	+5.1
TextCNN (100% data)	91.2	92.7	5.0
+ 100% adversarial	-0.5	-10.3	+0.8
BERT (10% data)	92.5	96.1	5.4
+ 10% adversarial	+0.0	-12.3	+7.6
TextCNN (10% data)	83.6	99.0	5.6
+ 10% adversarial	+2.0	-3.5	+0.3

Table 8: Adversarial training results on AG news test set. "Acc" indicates accuracy.

shown in Table 8, when the full training data is available, adversarial training slightly *decreases* the test accuracy by 0.2% and 0.5% respectively. This aligns with previous observations (Jia et al., 2019). Interestingly, in the low-data scenario with adversarial training, the BERT-based classifier has no accuracy drop, and TextCNN achieves a 2.0% absolute improvement. This suggests that a model with less capacity can benefit more from silver data.

Does adversarial training help the models defend against adversarial attacks? To evaluate this, we use CLARE to attack classifiers trained with and without adversarial examples.[9] A higher success rate and fewer modifications indicate a victim classifier is more vulnerable to adversarial attacks. As shown in Table 8, in 3 out of the 4 cases, adversarial training helps to decrease the attack success rate by more than 10.3%, and to increase the number of modifications needed by more than 0.8. The only exception is the TextCNN model trained with 10% data. A possible reason can be that it is trained with little data and thus generalizes less well.

These results suggest that CLARE can be used to improve downstream models' robustness, with a negligible accuracy drop.

5 Related Work

Textual adversarial attack. An increasing amount of effort is being devoted to generating better textual adversarial examples with various attack models. Character-based models (Liang et al., 2019; Ebrahimi et al., 2018; Li et al., 2018; Gao et al., 2018, *inter alia*) use misspellings to attack the victim systems; however, these attacks can often be defended by a spell checker (Pruthi et al., 2019; Zhou et al., 2019b; Jones et al., 2020). Many sentence-level models (Iyyer et al., 2018; Wang et al., 2020; Zou et al., 2020, *inter alia*) have been developed to introduce more sophisticated token/phrase perturbations. These, however, generally have difficulty maintaining semantic similarity with original inputs (Zhang et al., 2020a). Recent word-level models explore synonym substitution rules to enhance semantic meaning preservation (Alzantot et al., 2018; Jin et al., 2020; Ren et al., 2019; Zhang et al., 2019; Zang et al., 2020, *inter alia*). Our work differs in that CLARE uses three contextualized perturbations that produces more fluent and grammatical outputs.

Text generation with BERT. Generation with masked language models has been widely studied in various natural language tasks, ranging from lexical substitution (Wu et al., 2019a; Zhou et al., 2019a; Qiang et al., 2020; Wu et al., 2019b, *inter alia*) to non-autoregressive generation (Gu et al., 2018; Lee et al., 2018; Ghazvininejad et al., 2019; Wang and Cho, 2019; Ma et al., 2019; Sun et al., 2019; Ren et al., 2020; Zhang et al., 2020b, *inter alia*).

6 Conclusion

We have presented CLARE, a contextualized adversarial example generation model for text. It uses contextualized knowledge from pretrained masked language models, and can generate adversarial examples that are natural, fluent and grammatical. With three contextualized perturbation patterns, *Replace*, *Insert* and *Merge* in our

[9] In preliminary experiments, we found that it is more difficult to use other models to attack a victim model trained with the adversarial examples generated by CLARE, than to use CLARE itself.

arsenal, CLARE can produce outputs of varied lengths and achieves a higher attack success rate than baselines and with fewer edits. Human evaluation shows significant advantages of CLARE in terms of textual similarity, fluency and grammaticality. We release our code and models at https://github.com/cookielee77/CLARE.

Acknowledgments

We would like to thank the reviewers for their constructive comments. We thank NVIDIA Corporation for the donation of the GPU used for this research. We also thank Tongshuang Wu, Guoyin Wang and Shuhuai Ren for their helpful discussions and feedback.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proc. of EMNLP*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proc. of ACL*.

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

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proc. of ACL*.

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

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

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proc. of EMNLP*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proc. of ICLR*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *Proc. of ICLR*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proc. of NAACL*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proc. of EMNLP*.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proc. of EMNLP*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? natural language attack on text classification and entailment. In *Proc. of AAAI*.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. In *Proc. of ACL*.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proc. of CNLP*, pages 313–323.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. of EMNLP*.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pre-trained models. In *Proc. of ACL*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proc. of EMNLP*.

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

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

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2019. Deep text classification can be fooled. In *Proc. of IJCAI*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019. Flowseq: Non-autoregressive conditional sequence generation with generative flow. In *Proc. of EMNLP*.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020a. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proc. of EMNLP: System Demonstrations*.

John X Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020b. Reevaluating adversarial examples in natural language. In *Proc. of EMNLP Findings*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proc. of NAACL*.

Daniel Naber et al. 2003. *A rule-based style and grammar checker*. Citeseer.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Proc. of NeurIPS*.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proc. of ACL*.

Jipeng Qiang, Yun Li, Yi Zhu, and Yunhao Yuan. 2020. A simple bert-based approach for lexical simplification. In *Proc. of AAAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proc. of ACL*.

Yi Ren, Jinglin Liu, Xu Tan, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. A study of non-autoregressive model for sequence generation. In *Proc. of ACL*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proc. of ACL*.

Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP*.

Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhihong Deng. 2019. Fast structured decoding for sequence models. In *Proc. of NeurIPS*.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2018. Robustness may be at odds with accuracy. In *Proc. of ICLR*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proc. of EMNLP*.

Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. of ICLR*.

Boxin Wang, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li. 2020. T3: Tree-autoencoder regularized adversarial text generation for targeted attack. In *Proc. of EMNLP*.

Xiaosen Wang, Hao Jin, and Kun He. 2019b. Natural language adversarial attacks and defenses in word level. *arXiv preprint arXiv:1909.06723*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of NAACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019a. Conditional bert contextual augmentation. In *Proc. of ICCS*.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019b. "mask and infill": Applying masked language model to sentiment transfer. In *Proc. of IJCAI*.

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 *Proc. of ACL*.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating fluent adversarial examples for natural languages. In *Proc. of ACL*.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020a. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proc. of NeurIPS*.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and Bill Dolan. 2020b. Pointer: Constrained text generation via insertion-based generative pre-training. In *Proc. of ACL*.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *Proc. of ICLR*.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019a. Bert-based lexical substitution. In *Proc. of ACL*.

Yichao Zhou, Jyun-Yu Jiang, Kai-Wei Chang, and Wei Wang. 2019b. Learning to discriminate perturbations for blocking adversarial attacks in text classification. In *Proc. of EMNLP*.

Wei Zou, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun Chen. 2020. A reinforced generation of adversarial samples for neural machine translation. In *Proc. of ACL*.

A Appendix

A.1 Additional Experiment Details

Model Implementation. All pretrained models and victim models based on RoBERTa and $\text{BERT}_{\text{base}}$ are implemented with Hugging Face transformers[10] (Wolf et al., 2019) based on PyTorch (Paszke et al., 2019). $\text{RoBERTa}_{\text{distill}}$, $\text{RoBERTa}_{\text{base}}$ and uncase $\text{BERT}_{\text{base}}$ models have 82M, 125M and 110M parameters, respectively. We use $\text{RoBERTa}_{\text{distill}}$ as our main backbone for fast inference purpose. TextFooler[11] and BERTAttack[12] are built with their open source implementation provided by the authors. In the implementation of TextFooler+LM, we use small sized GPT-2 language model (Radford et al., 2019) to further select those candidate tokens that have top 20% perplexity in the candidate token set. In the adversarial training (§4.3), the small TextCNN victim model (Kim, 2014) has 128 embedding size and 100 filters for $3, 4, 5$ window size with 0.5 dropout, resulting in 7M parameters.

During the implementation of *w/o* $p_{\text{MLM}} > k$ in the ablation study (§4.1), we randomly sample 200 tokens and then apply the similarity constraint to construct candidate set, as exhausting the vocabulary is computationally expensive.

Evaluation Metric. The similarity function sim builds on the universal sentence encoder (USE; Cer et al., 2018) to measure a *local* similarity at the perturbation position with window size 15 between the original input and its adversary. *All baselines* are equipped this sim when constructing the candidate vocabulary. The evaluation metric **Sim** uses USE to calculate a *global* similarity between two texts. These procedures are typically following Jin et al. (2020). We mostly rely on human evaluation (§3.3) to conclude the significant advantage of preserving textual similarity on CLARE compared with TextFooler.

Data Processing. When processing the data, we keep all punctuation in texts for both victim model training and attacking. This differs the preprocessing setting in TextFooler (Jin et al., 2020) as we empirically found that removing punctuation makes the victim model vulnerable. Since GLUE benchmark (Wang et al., 2019a) does not provide the label for test set, we instead use its dev set as the the test set for the included datasets (MNLI, QNLI, QQP, MRPC, SST-2) in the evaluation. For the sentence-pair tasks (e.g., MNLI, QNLI, QQP, MRPC), we attack the longer one excluding the tokens appearing in both sentences. This is because inference tasks usually require entailed data to have the same keywords, e.g., numbers, name entities, etc. All experiments are conducted on one Nvidia GTX 1080Ti GPU.

Dataset	Avg. Length	# Classes	Train	Test	Acc
SST-2	10	2	67K	0.9K	92.3%
DBpedia	55	14	560K	70K	99.3%
QQP	13/13	2	363K	40K	91.4%
MRPC	23/23	2	3.6K	1.7K	81.4%

Table 9: Some statistics of datasets. The last column indicates the victim model's accuracy on the original test set *without* adversarial attack.

A.2 Additional Results

We include the results of DBpedia ontology dataset (**DBpedia**; Zhang et al., 2015, Stanford sentiment treebank (**SST-2**; Socher et al., 2013), Microsoft Research Paraphrase Corpus (**MRPC**; Dolan and Brockett, 2005), and Quora Question Pairs (**QQP**) from the GLUE benchmark in this section. Table 9 summarizes come statistics of these datasets. The results of different models on these datasets are summarized Table 10. Compared with all baselines, CLARE achieves the best performance on attack success rate, perplexity, grammaticality, and similarity. It is consistent with our observation in §3.3.

A.3 Human Evaluation Details

For each human evaluation on **AG News** dataset, we randomly sampled 300 sentences from the test set combining the corresponding adversarial examples from CLARE and TextFooler (We only consider sentences can be attacked by both models). In order to make the task less abstract, we pair the adversarial examples by the two models, and present them to the participants along with the original input and its gold label. We ask them which one they prefer in terms of (1) having more similar a meaning to the original input (similarity), and (2) being more fluent and grammatical (fluency and grammaticality). We also provide them with a neutral option, when the participants consider the two

[10] https://github.com/huggingface/transformers

[11] https://github.com/jind11/TextFooler

[12] https://github.com/LinyangLee/BERT-Attack

	SST-2 (PPL = 99.5)					DBpedia (PPL = 37.3)				
Model	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**
TextFooler	89.8	14.9	227.7	0.53	0.69	56.2	24.9	182.5	1.88	0.68
+ LM	51.7	18.3	137.5	0.50	0.69	20.1	22.4	84.0	1.22	0.70
BERTAttack	87.8	8.1	142.9	0.03	0.67	60.7	9.1	57.8	0.20	0.69
CLARE	**97.8**	**7.5**	**137.4**	**0.01**	**0.75**	**65.8**	**7.0**	**53.3**	**-0.03**	**0.73**

	QQP (PPL = 56.2)					MRPC (PPL = 42.9)				
Model	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**
TextFooler	16.2	12.7	145.2	0.61	0.74	24.5	10.6	118.8	0.35	0.75
+ LM	7.8	12.9	78.8	0.21	0.77	12.9	9.5	71.0	0.29	0.79
BERTAttack	24.2	11.3	78.0	0.25	0.71	29.7	13.5	74.6	0.05	0.79
CLARE	**27.7**	**10.2**	**74.8**	**0.14**	**0.76**	**34.8**	**9.1**	**69.5**	**0.02**	**0.83**

Table 10: Adversarial example generation performance in attack success rate (A-rate), modification rate (Mod), perplexity (PPL), number of increased grammar errors (GErr), and text similarity (Sim). The perplexity of the original inputs is indicated in parentheses for each dataset. Bold indicates the best performance on each metric.

indistinguishable. Additionally, we ask the participants to annotate the adversarial examples, and compare their annotations against the gold labels (label consistency). Higher label consistency indicates the model is better at causing the victim model to make errors while preserving human predictions.

Each pair of system outputs was randomly presented to 5 crowd-sourced judges, who indicated their preference for similarity, fluency, and grammaticality using the form shown in Figure 3. The labelling task is illustrated in Figure 4. To minimize the impact of spamming, we employed the top-ranked 30% of U.S. workers provided by the crowd-sourcing service. Detailed task descriptions and examples were also provided to guide the judges. We calculate p-value based on 95% confidence intervals by using 10K paired bootstrap replications, implemented using the R Boot statistical package.

A.4 Qualitative Samples

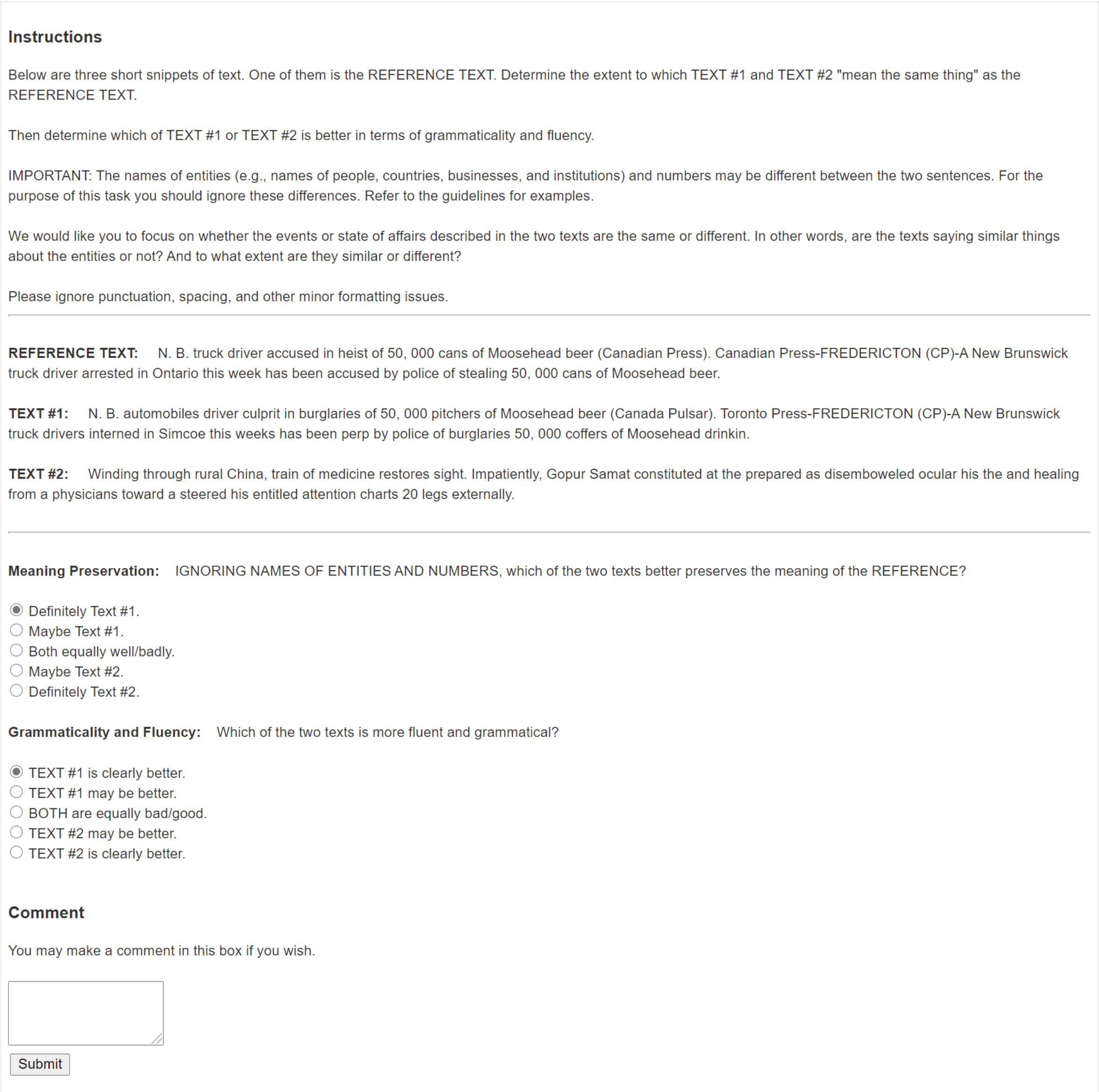
We include generated adversarial examples by CLARE and TextFooler on **AG News**, **DBpeida**, **Yelp**, **MNLI**, and **QNLI** datasets in Table 11 and Table 12.

AG (Business)	TECH BUZZ : Yahoo, Adobe team up for new Web services. Stepping up the battle of online search and services, Yahoo Inc. and Adobe Systems Inc. have joined forces to tap each other's customers and put Web search features into Adobe's popular Acrobat Reader software.
TextFooler (Sci&Tech)	TECH BUZZ : Yahoo, Adobe team up for *roman Cyberspace utilities*. Stepping up the battle of online *locating* and services, Yahoo Inc. and Adobe Systems Inc. have joined forces to tap each other's customers and put Web search features into Adobe's popular Acrobat Reader software.
CLARE (Sci&Tech)	TECH BUZZ : Yahoo, Adobe team up for new Web *Explorer*. Stepping up the battle of online search and services, Yahoo Inc. and Adobe Systems Inc. have joined forces to tap each other's customers and put Web search features into Adobe's popular Acrobat Reader software.
AG (Sport)	Padres Blank Dodgers 3 - 0. LOS ANGELES - Adam Eaton allowed five hits over seven innings for his career - high 10th victory, Brian Giles homered for the second straight game, and the San Diego Padres beat the Los Angeles Dodgers 3 - 0 Thursday night. The NL West - leading Dodgers' lead was cut to 2 1 / 2 games over San Francisco - their smallest since July 31 ...
TextFooler (World)	Dodger Blank *Yanks* 3 - 0. *Loos* ANGELES - *Adams Parades enabling* five hits over seven *slugging* for his career - high 10th *victoria*, Brian Giles homered for the second straight *matching*, and the *Tome José Dodger* beat the Los Angeles *Dodger* 3 - 0 Thursday *blackness*. The NL *Westerner* - *eminent Dodger*' lead was cut to 2 1 / 2 games over San *San* - their *tiny as janvier* 31 ...
CLARE (World)	Padres Blank Dodgers 3 - 0. Milwaukee **NEXT** - Adam Eaton allowed five hits over seven innings for his career - high 10th victory, Brian Giles homered for the second straight game, and the San Diego Padres beat the Los Angeles Dodgers 3 - 0 Thursday night. The NL West - leading Dodgers' lead was cut to 2 1 / 2 games over San Francisco - their smallest since July 31 ...
Yelp (Positive)	The food at this chain has always been consistently good. Our server in downtown (where we spent New Year's) was new, but that did not impact our service at all. She was prompt and attentive to our needs.
TextFooler (Negative)	The food at this chain has always been *necessarily ok*. Our server in downtown (where we spent New Year's) was new, but that did not impact our service at all. She was *early* and attentive to our needs.
CLARE (Negative)	The food at this chain has always been **looking** consistently good. Our server in downtown (where we spent New Year's) was new, but that did not *enhance* our service at all. She was prompt and attentive to our needs.
Yelp (Positive)	The pho broth is actually flavorful and doesn't just taste like hot water with beef and noodles. I usually do take out and the order comes out fast during dinner which should be expected with pho, it's not hard to soak noodles, slice beef and pour broth.
TextFooler (Negative)	The pho broth is actually flavorful and doesn't just *tasty* like *torrid waters* with *slaughter* and *salads*. I *repeatedly* do take out and the order *poses* out fast during dinner which should be expected with pho , it's not *strenuous* to soak noodles, *severing* beef and pour broth.
CLARE (Negative)	The pho broth is actually flavorful and doesn't just taste **bland** like hot water with beef and noodles. I usually do take out and the order comes out **awfully** fast during dinner which should be expected with pho, it's not hard to soak noodles, slice beef and pour broth.
MNLI (Neutral)	*Premise*: Thebes held onto power until the 12th Dynasty, when its first king, Amenemhet Iwho reigned between 1980 1951 b.c. established a capital near Memphis. *Hypothesis*: The capital near Memphis lasted only half a century before its inhabitants abandoned it for the next capital.
TextFooler (Contradiction)	*Premise*: Thebes *apprehended pour powers* until the 12th *Familial* , when its *earliest* king , Amenemhet Iwho reigned between 1980 1951 *c*.c. established a capital near Memphis . *Hypothesis*: The capital near Memphis lasted only half a century before its inhabitants abandoned it for the next capital.
CLARE (Contradiction)	*Premise*: Thebes held onto power until the 12th Dynasty, when its first king, Amenemhet Iwho reigned between 1980 1951 b.c. **thereafter** established a capital near Memphis. *Hypothesis*: The capital near Memphis lasted only half a century before its inhabitants abandoned it for the next capital.
MNLI (Entailment)	*Premise*: Hopefully, Wall Street will take voluntary steps to address these issues before it is forced to act. *Hypothesis*: Wall Street is facing issues, that need to be addressed.
TextFooler (Neutral)	*Premise*: Hopefully, Wall Street will take voluntary steps to *treatment* these issues before it is forced to act. *Hypothesis*: Wall Street is facing issues, that need to be addressed.
CLARE (Neutral)	*Premise*: Hopefully, Wall Street will take voluntary steps to *eliminate* these issues before it is forced to act. *Hypothesis*: Wall Street is facing issues, that need to be addressed.

Table 11: Adversarial examples produced by different models. The gold label of the original is shown below the (bolded) dataset name. ***Replace***, **Insert** and **Merge** are highlighted in *italic red*, **bold blue** and sans serif yellow, respectively. (Best viewed in color).

QNLI (Entailment)	*Premise*: Who overturned the Taft Vale judgement ? *Hypothesis*: One of the first acts of the new Liberal Government was to reverse the Taff Vale judgement.
TextFooler (Not-Entailment)	*Premise*: Who overturned the Taft Vale judgement ? *Hypothesis*: One of the first acts of the new Liberal Government was to *invest* the Taff Vale judgement.
CLARE (Not-Entailment)	*Premise*: Who overturned the Taft Vale judgement ? *Hypothesis*: One of the first acts of the new Liberal *Constitution* was to reverse the Taff Vale judgement.
QNLI (Entailment)	*Premise*: What are the software testers aware of ? *Hypothesis*: Black-box testing treats the software as a black box, examining functionality without any knowledge of internal implementation, without seeing the source code.
TextFooler (Not-Entailment)	*Premise*: What are the software testers aware of ? *Hypothesis*: Black-*boxes* testing *administers* the software as a black box, *investigating functions unless* any knowledge of internal *fulfil*, *unless* seeing the *wellspring* code.
CLARE (Not-Entailment)	*Premise*: What are the software testers aware of ? *Hypothesis*: Black-box testing treats the software as a black box, examining functionality without *awareness* of internal implementation, without seeing the source code.
DBpedia (Transportation)	Honda Crossroad. The Honda Crossroad refers to two specific types of SUVs made by Honda. One of them is a rebadged Land Rover Discovery Series I SUV while the other is a completely different vehicle introduced in 2008.
TextFooler (Album)	*Suzuki Junctions*. The *Suzuki* Crossroad refers to *three accurate typing* of *prius posed* by*Isuzu*. One of them is a rebadged Land Rover *Identify* Series I *LEXUS* while the other is a completely different vehicle introduced in 2008.
CLARE (Company)	Honda Crossroad. The Honda Crossroad refers to two specific *manufacturers* of SUVs made by Honda. One of them is a rebadged Land Rover Discovery Series I SUV while the other is a completely different vehicle introduced in 2008.
DBpedia (Company)	Yellow Rat Bastard. Yellow Rat Bastard is the flagship establishment in a chain of New York City retail clothing stores owned by Henry Ishay. It specializes in hip - hop-and alternative - style clothing and shoes.
TextFooler (Building)	*Yellowish Rats Schmuck* . *Yellowish Rats Dickwad* is the flagship *establishments* in a *chains* of New York City retail *uniforms* stores owned by *Henrik* Ishay . It *specialize* in hip - hop-and alternative - style *laundry* and *sneakers*.
CLARE (Building)	Yellow Rat Bastard. Yellow Rat Bastard **Mall** is the flagship establishment in a chain of New York City retail clothing stores owned by Henry Ishay. It specializes in hip - hop-and alternative - style clothing and shoes.
MRPC (Not Paraphrase)	*Premise*: The Americas market will decline 2.1 percent to $30.6 billion in 2003, and then grow 15.7 percent to $35.4 billion in 2004. *Hypothesis*: The US chip market is expected to decline 2.1 percent this year, then grow 15.7 percent in 2004.
TextFooler (Paraphrase)	*Premise*: The Americas market will decline 2.1 percent to $30.6 billion in 2003, and then grow 15.7 percent to $35.4 billion in 2004. *Hypothesis*: The US chip market is *prescribed* to decline 2.1 percent this year, then grow 15.7 percent in 2004.
CLARE (Paraphrase)	*Premise*: The Americas market will decline 2.1 percent to $30.6 billion in 2003, and then grow 15.7 percent to $35.4 billion in 2004. *Hypothesis*: The US chip market is expected to decline 2.1 percent this year, then grow 15.7 percent in 2004 **yr**.
MRPC (Paraphrase)	*Premise*: The Securities and Exchange Commission filed a civil fraud suit against the teen in Boston. *Hypothesis*: The Securities and Exchange Commission brought a related civil case on Thursday.
TextFooler (Not Paraphrase)	*Premise*: The Securities and Exchange Commission filed a civil fraud suit against the teen in Boston. *Hypothesis*: The Securities and Exchange Commission brought a *connect* civil case on *Yesterday*.
CLARE (Not Paraphrase)	*Premise*: The Securities and Exchange Commission filed a civil fraud suit against the teen in Boston. *Hypothesis*: The Securities and Exchange Commission brought a *Massachusetts* civil *lawsuit* on Thursday.

Table 12: Adversarial examples produced by different models. The gold label of the original is shown below the (bolded) dataset name. ***Replace***, **Insert** and **Merge** are highlighted in *italic red*, **bold blue** and sans serif yellow, respectively. (Best viewed in color).



Instructions

Below are three short snippets of text. One of them is the REFERENCE TEXT. Determine the extent to which TEXT #1 and TEXT #2 "mean the same thing" as the REFERENCE TEXT.

Then determine which of TEXT #1 or TEXT #2 is better in terms of grammaticality and fluency.

IMPORTANT: The names of entities (e.g., names of people, countries, businesses, and institutions) and numbers may be different between the two sentences. For the purpose of this task you should ignore these differences. Refer to the guidelines for examples.

We would like you to focus on whether the events or state of affairs described in the two texts are the same or different. In other words, are the texts saying similar things about the entities or not? And to what extent are they similar or different?

Please ignore punctuation, spacing, and other minor formatting issues.

REFERENCE TEXT: N. B. truck driver accused in heist of 50, 000 cans of Moosehead beer (Canadian Press). Canadian Press-FREDERICTON (CP)-A New Brunswick truck driver arrested in Ontario this week has been accused by police of stealing 50, 000 cans of Moosehead beer.

TEXT #1: N. B. automobiles driver culprit in burglaries of 50, 000 pitchers of Moosehead beer (Canada Pulsar). Toronto Press-FREDERICTON (CP)-A New Brunswick truck drivers interned in Simcoe this weeks has been perp by police of burglaries 50, 000 coffers of Moosehead drinkin.

TEXT #2: Winding through rural China, train of medicine restores sight. Impatiently, Gopur Samat constituted at the prepared as disemboweled ocular his the and healing from a physicians toward a steered his entitled attention charts 20 legs externally.

Meaning Preservation: IGNORING NAMES OF ENTITIES AND NUMBERS, which of the two texts better preserves the meaning of the REFERENCE?

● Definitely Text #1.
○ Maybe Text #1.
○ Both equally well/badly.
○ Maybe Text #2.
○ Definitely Text #2.

Grammaticality and Fluency: Which of the two texts is more fluent and grammatical?

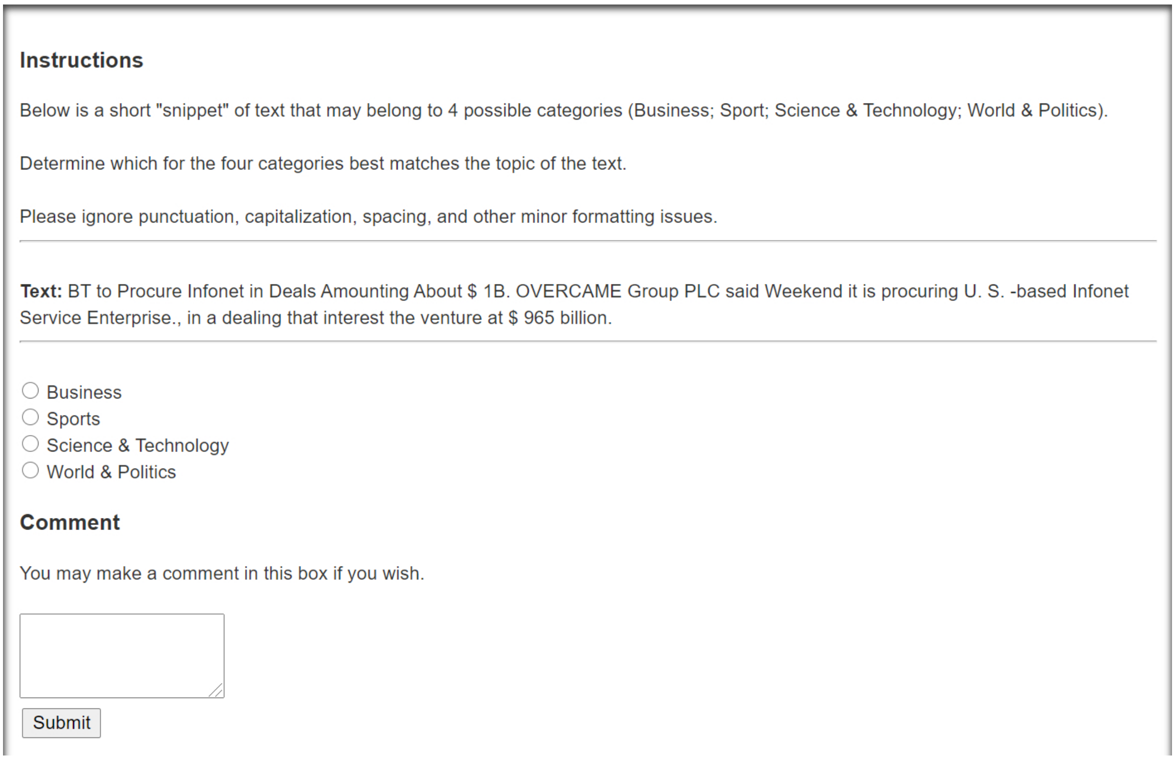
● TEXT #1 is clearly better.
○ TEXT #1 may be better.
○ BOTH are equally bad/good.
○ TEXT #2 may be better.
○ TEXT #2 is clearly better.

Comment

You may make a comment in this box if you wish.

Submit

Figure 3: Pair-wise comparison in terms of text similarity and fluency & grammaticality on human evaluation.



Instructions

Below is a short "snippet" of text that may belong to 4 possible categories (Business; Sport; Science & Technology; World & Politics).

Determine which for the four categories best matches the topic of the text.

Please ignore punctuation, capitalization, spacing, and other minor formatting issues.

Text: BT to Procure Infonet in Deals Amounting About $ 1B. OVERCAME Group PLC said Weekend it is procuring U. S. -based Infonet Service Enterprise., in a dealing that interest the venture at $ 965 billion.

○ Business
○ Sports
○ Science & Technology
○ World & Politics

Comment

You may make a comment in this box if you wish.

Figure 4: Label consistency task on human evaluation.