Knowledge discovery from patient forums: gaining novel medical insights from patient experiences

Dirkson, A.R.

6

VULNERABILITIES OF BERT FOR NAMED ENTITY RECOGNITION

Edited from: **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2021). Breaking BERT: Understanding its Vulnerabilities for Named Entity Recognition through Adversarial Attack. ArXiv. `https://arxiv.org/abs/2109.11308`

Both generic and domain-specific BERT models, like BioBERT and SciBERT, are widely used for natural language processing tasks including Named Entity Recognition (NER). In this chapter we investigate the vulnerability of BERT models to variation in the input data for NER using adversarial attack. Adversarial attack is the crafting of changes to the input data to deliberately try to fool the model.

We found that under these conditions BERT models are vulnerable to words in the local context of the entity being replaced with synonyms rarely seen during training. This type of variation resulted in 20.2 to 45.0% of entities being predicted completely wrong and another 29.3 to 53.3% of entities being predicted wrong partially. Often a single synonym replacement was sufficient to fool the model. The domain-specific BERT model trained from scratch (SciBERT) was more vulnerable than the original BERT model or the domain-specific model that retains the BERT vocabulary (BioBERT). We also found that BERT models are particularly vulnerable to entities that occur infrequently; BERT models could be fooled to predict 89.5% to 99.4% of entities wrongly when entities were replaced with more rare entities of the same type.

Our results chart the vulnerabilities of BERT models for NER and emphasize the importance of further research into uncovering and reducing these weaknesses.

6.1. INTRODUCTION

Self-attentive neural models, such as BERT [84], attain a high performance on a wide range of natural language processing (NLP) tasks. Despite their excellent performance, the robustness of BERT-based models is contested: Various studies [139, 143, 180, 290, 347] recently showed that BERT is vulnerable to adversarial attacks. Adversarial attacks are deliberate attempts to fool the model into giving the incorrect output by providing it with carefully crafted input samples, also called adversarial examples.

At present, the work on adversarial attack of Named Entity Recognition (NER) models is limited to a single study: Araujo et al. [11] attack biomedical BERT models by simulating spelling errors and replacing entities with their synonyms. They find that both attacks drastically reduce performance of these domain-specific BERT models on medical NER tasks.

Here, we aim to systematically test the robustness of BERT models for NER under severe stress conditions in order to investigate *which* variation in entities and entity contexts BERT models are most vulnerable to. This will, in turn, further our understanding of what these models do and do not learn. To do so, we propose two adversarial attack methods: replacing words in the context of entities with synonyms, and replacing the entities themselves with others of the same type. In contrast to previous work, the methods we propose are adaptive and specifically target BERT's weaknesses: We create adversarial examples by making the changes to the input that either manage to fool the model or bring it closest to making a mistake (i.e., lower the prediction score for the correct output) instead of randomly introducing noise or variation.

We address the following research questions:

1. How vulnerable are BERT models to adversarial attack on general and domain-specific NER?

2. To what extent is the vulnerability impacted by domain-specific training?

3. To which types of variation are BERT models for NER the most vulnerable?

Designing methods for direct adversarial attack of NER models poses additional challenges compared to the attack of text classification models as labels are predicted per word, sentences can contain multiple entities, and entities can contain multiple words. To ensure that labels remain accurate in our adversarial sentences, we constrain synonym replacements to non-entity words when altering the context of the entity (i.e., an *entity context attack*) and substitute entities only by entities of the same type when attacking the entity itself (i.e., an *entity attack*). In line with previous work [143, 180], we include a minimal semantic similarity threshold based on the Universal Sentence Encoder [57] to safeguard semantic consistency. Nonetheless, we acknowledge that for entity replacement adversarial examples may not be semantically consistent (e.g., if "Japan" is replaced with "China" in the sentence "Tokyo is the capital of Japan"). Although factually incorrect, the resulting sentences can be considered utility-preserving i.e., they retain their usefulness as valid input to the model [293], because BERT models should be able to identify that the final word in the sentence is a country even if it is not the correct country. In real-world data, sentences are not necessarily factually correct.

We assume a black-box setting, which means that the adversarial method has no knowledge of the data, parameters or model architecture [8]. This allows our methods to also be used for other neural architectures. Although we use English data, our methods are largely language-independent. Only an appropriate language model for synonym selection would be required.

The contributions of this chapter are twofold: We adapt existing adversarial attack methods to sequence labeling tasks and evaluate the vulnerability of general and domain-specific BERT models for NER. We make our code available for follow up research.[1]

6.2. RELATED WORK

In prior work, token-level black-box methods for adversarial attack have mainly been developed for classification and textual entailment [7]. Substituting tokens with their synonyms is the most popular choice for perturbing at the token level. Synonyms are often found using nearest neighbors in a word embeddings model. One major challenge when selecting synonyms based on word embeddings is that antonyms will also be close in the embedding space. To solve this issue, recent studies [139, 143, 181] require a minimal semantic similarity between the generated and original sentence. Additionally, some methods [8, 143] use word embeddings with additional synonymy constraints [211]. We will employ both techniques.

Approaches also differ in how they select the word that is perturbed: while some select words randomly, it is more common to use the importance of the word for the output [7]. The importance is often operationalized as the difference in output before and after removing the word. We follow this approach in our method.

Most adversarial attack methods were developed for attacking recurrent neural models. However, there has been a growing interest in attacking self-attentive models in the last year [11, 18, 139, 143, 180, 209, 290, 347]. Nonetheless, the only study that has attacked BERT models for NER is the study by Araujo et al. [11]. They perform two types of character-level (i.e., swapping letters and replacing letters with adjacent keys on the keyboard) and one type of token-level perturbation (i.e., replacing entities with their synonyms). The authors find that biomedical BERT models perform far worse on NER tasks when spelling mistakes are included or synonyms of entities are used.

Our work differs from Araujo et al. [11] in three ways. First, our adversarial examples are generated based on the importance of words for the correct output instead of through random changes. Thereby, we are able to test the robustness of BERT under the most severe stress conditions, while Araujo et al. [11] evaluate the scenario where the input data is noisy due to spelling mistakes and the use of synonyms. Second, we analyze the impact of replacing entities with others of the same type (e.g., 'France' with 'Britain') and replacing words in the context of entities (see Table 6.1 for an example) instead of replacing entities with their synonyms. Third, we will test our method on the original BERT model as well as biomedical BERT models and on both generic and biomedical NER.

[1] Our code (BSD-3 Clause license), URLs to the benchmark data and the annotation guideline are available at: `https://github.com/AnneDirkson/breakingBERT`

The	Republic	of	China	**bought**	flowers
<O>	<B-LOC>	<I-LOC>	<I-LOC>	<O>	<O>
The	Republic	of	China	**purchased**	flowers
<O>	<O>	<O>	<B-LOC>	<O>	<O>

Table 6.1: Example of a partial success. The **bold** word has been changed to attack the entity 'Republic of China'.

6.3. METHODS

In this section, we describe two methods for generating adversarial examples designed to fool NER models, namely through (1) synonym replacements in the entity context (*entity context attack*) and (2) entity replacement (*entity attack*). These are described in Sections 6.3.1 and 6.3.1, respectively.

6.3.1. AIM OF THE ATTACKS

We aim to generate adversarial examples in which a target entity is no longer recognized correctly. This can be either because it has become a false negative or it has been assigned a different entity type. The attack is considered a success when the correct label has been changed, unless it has changed from the I-tag to the B-tag of the same entity type under the IOB schema. An example of this can be seen in Table 6.1: Here, the start of the entity is mislabeled, but the last part of the entity is still recognized. We consider this a partial success.

METRICS FOR EVALUATION

The success of the attack and thus the vulnerability of the model is evaluated by the percentage of entities that were originally correctly labeled but are mislabeled after attack. For *entity context attacks* entities can also be partially mislabeled i.e., only some words in the entity are mislabeled. This is captured by the partial success rate: the percentage of entities for which not the whole entity but at least half of the entity is mislabeled. For context attacks we also include a metric ('Words perturbed') to measure how much the sentence needed to be changed before the attack was successful: the average percentage of words that were perturbed out of the total amount of out-of-mention words in the sentences. This metric functions as a proxy for how difficult it is to fool the model [143].

ENTITY CONTEXT ATTACK

To investigate the impact of the context on the correct labeling of the entity, we adapt the method of Jin et al. [143], which was designed for text classification, to sequence labeling tasks. For each entity in the sentence, a separate adversarial example is created, as models may rely on different contextual words for different entities.

Step 1: Choosing the word to perturb We use the importance ranking function shown in Equation 6.1 to rank words based on their importance for assigning the correct label to the entity. The importance (I_w) of a word w for a token in the entity is calculated as the change in the predictions (logits[2]) of the *correct* label before and after deleting the word

[2]Here logits refers to the vector of raw (non-normalized) predictions that the BERT model generates

Domain	Dataset	Entity types	Dev.	Train	Test	Eval. subset* (# Entities)
General	CoNLL-2003	Person, Location, Organization, Miscellaneous	3,466	14,987	3,684	500 (1,343)
General	W-NUT 2017	Person, Location, Corporation, Product, Creative-work, Group	1,008	1,000	1,287	500 (787)
Biomedical	BC5CDR	Disease, Chemical	4,580	4,559	4,796	500 (1,221)
Biomedical	NCBI disease	Disease	922	5,432	939	487 (897)

Table 6.2: Size of the data sets (number of sentences). *This subset is used for automatic evaluation

from the sentence [143]. If the deletion of the word leads to an *incorrect* label for the entity token, the importance of the word is increased by adding the raw prediction score (logits) attributed to the incorrect label.

If the entity consists of multiple words, we rank words based on their summed importance for correctly labeling each of the individual words in the entity. Besides stop words, we also exclude other entities from being perturbed. We adapt the function so that for any word with an I-tag, both the I and B label of the entity type (e.g., B-PER and I-PER) are considered correct.

Given a sentence of n words $X = w_1, w_2, ..., w_n$, the importance (I_w) of a word w for a token in the entity is formally defined as:

$$\begin{aligned} I_w = & F_Y(X) - F_Y(X_{-w}) \\ & \text{if} F(X_{-w}) = Y \vee (F(X) = Y_I \wedge F(X_{-w}) = Y_B) \\ & F_Y(X) - F_Y(X_{-w}) + F_{\_\tilde{Y}}(X_{-w}) - F_{\overline{\tilde{Y}}}(X) \\ & \text{if} F(X_{-w}) \neq Y \end{aligned} \tag{6.1}$$

where F_Y is the prediction score for the correct label, $F_{\tilde{Y}}$ is the prediction score of the predicted label, F is the predicted label, Y is the correct label, Y_I is the I-tag version of the correct label, Y_B is the B-tag version of the correct label and X_{-w} is the sentence X after deleting the word w.

Step 2: Gathering synonyms For each word, we select synonyms from the Paragram-SL999 word vectors [211] with a similarity to the original word above the threshold δ. Mrkšić et al. [211] injected antonymy and synonymy constraints into the vector space representation to specifically gear the embeddings space towards synonymy. These embeddings achieved state-of-the-art performance on SimLex-999 [134] and were also used by Jin et al. [143] and Alzantot et al. [8]. We chose 0.5 as the minimal similarity threshold δ for synonym selection in contrast to the threshold of 0 used by previous work to better guarantee semantic similarity. Regardless of δ, a maximum of 50 synonyms are selected. Examples of word pairs with a δ above 0.5 are 'bought' and 'obtained'; and

'cat' and 'puss'. Below this threshold but within the first 50 synonyms fall 'bought' and 'forfeited'; and 'cat' and 'dustpan'.

Step 3: Filtering synonyms To preserve syntax, synonyms must have the same POS tag as the original word. If the data did not include POS tags, we added POS tags using NLTK. We filter the generated sentences for a sufficiently high semantic similarity to the original sentence. Semantic similarity is calculated with the Universal Sentence Encoder (USE) [57]. We exclude synonyms that result in sentences falling below the similarity threshold ϵ.

Step 4: Selecting the final synonym After filtering, we check whether any of the synonyms can change the entity label(s) fully. If there are multiple options, we select the one that leads to the highest sentence similarity (ϵ) to the original sentence. If there are none, we select the synonym which can reduce the (summed) prediction scores of the correct label(s) the most. If no synonyms are left after filtering or none manage to reduce the prediction scores, we do not replace the original word.

For multi-word entities, it is possible that a synonym changes some, but not all, labels. From the synonyms that change the most labels, we select the one that leads to the largest reduction in the (summed) prediction scores for the unchanged labels (i.e., the labels that are still predicted correctly by the model) (see Equation 6.1). Which labels are still correct can differ per synonym.

Finalizing the adversarial examples For each word in this ranking, we go through step 2-4 until either the label(s) of the entity have been changed fully or there are no words left to perturb. Once the attack is partially successful, only the predictions of the not yet incorrectly labeled words in the entity are considered for subsequent iterations.

ENTITY ATTACK

To explore to what extent the models rely on the words of the entity itself, we replace the entity with one of the same type, e.g., we change 'Japan' to another location. If a sentence contains multiple entities, an adversarial sentence is generated for each entity. The replacement entity is selected from a list of all entities in the data that are of the same type. We randomly select 50 candidate replacements from the entity list. We exclude candidates that result in a sentence that is too semantically dissimilar from the original (i.e., falling below the semantic similarity threshold ϵ). For the remaining candidate entities, we check if the predicted label is incorrect. If so, we select the successful attack replacement with the highest semantic similarity at the sentence level. If not, the attack was unsuccessful.

6.4. EXPERIMENTS

6.4.1. DATA

We use two general-domain English NER data sets for evaluating our method: the CoNLL-2003 data [298] and the W-NUT 2017 data [83]. The goal of the latter was to investigate recognition of unusual, previously-unseen entities in the context of online discussions.

	CoNLL-2003	W-NUT 2017	NCBI-disease	BC5CDR
	BERT	BERT	BERT	BERT
Success rate (%)	36.3 ± 0.612	42.2 ± 0.677	20.2 ± 0.443	38.8 ± 0.862
Of which:				
– Missed entity (%)	47.4 ± 2.9	61.3 ± 5.1	100	90.4 ± 4.2
– Entity type error (%)	52.6 ± 2.9	38.7 ± 5.1	0	9.6 ± 4.2
Partial success rate (%)	51.0 ± 0.465	51.6 ± 1.6	29.3 ± 0.841	45.9 ± 1.1
Median semantic similarity	0.928 ± 0.009	0.926 ± 0.017	0.920 ± 0.040	0.946 ± 0.002
Words perturbed (%)	15.6 ± 0.306	13.2 ± 1.2	12.4 ± 1.0	12.3 ± 0.04

Table 6.3: Automatic evaluation results for context attacks on BERT models. Results are the mean of the three models

	NCBI-disease		BC5CDR	
	BioBERT	SciBERT	BioBERT	SciBERT
Success rate (%)	20.9 ± 0.762	26.4 ± 0.875	37.9 ± 0.388	45.0 ± 0.665
Of which:				
– Missed entity (%)	100	100	86.5 ± 2.3	87.1 ± 2.3
– Entity type error (%)	0	0	13.5 ± 2.3	12.9 ± 2.3
Partial success rate (%)	30.1 ± 1.032	39.0 ± 0.954	44.8 ± 0.331	53.3 ± 0.821
Median semantic similarity	0.921 ± 0.031	0.936 ± 0.030	0.921 ± 0.003	0.936 ± 0.008
Words perturbed (%)	9.1 ± 2.5	8.7 ± 2.9	9.8 ± 0.3	8.5 ± 0.8

Table 6.4: Automatic evaluation results for context attacks on biomedical BERT models. Results are the mean of the three models

Additionally, we use two English data sets from the biomedical domain: BC5CDR [179] and the NCBI disease corpus [90]. Both data sets have been used to evaluate domain-specific BERT models for NER in the biomedical domain [28, 174, 232]. See Table 6.2 for more details on the data sets.

6.4.2. TARGET MODELS

We fine-tune three BERT models (base-cased) for each data set with different initialization seeds (1, 2 & 4) using the Huggingface implementation [339]. We set the learning rate at 5×10^{-5} and optimized the number of epochs (3 or 4) as recommended in Devlin et al. [84] for NER. We select the number of epochs based on the first BERT model (seed=1). We find that for all data sets except W-NUT 2017, 4 epochs is optimal.

For the biomedical data sets, we additionally fine-tune two domain-specific BERT models, BioBERT (base-cased) [174] and SciBERT (scivocab-cased) [28]. Each model is trained in three-fold (seeds are 1, 2 & 4).

6.4.3. EVALUATION OF ADVERSARIAL ATTACKS

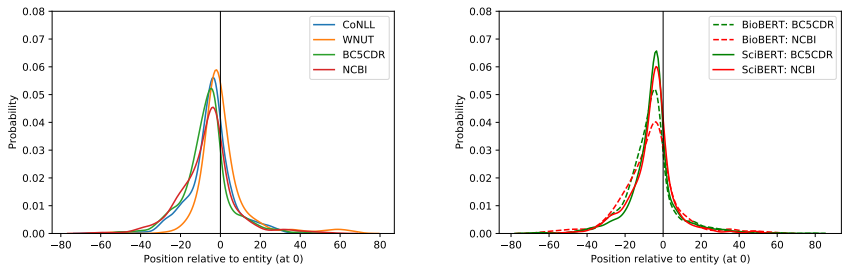
Automatic evaluation We randomly select 500 eligible sentences from each test set. Table 6.2 shows the number of entities in each subset. We considered sentences to be eligible if they contain at least one entity and one verb. For the NCBI-disease data, only 487 sentences fulfill these criteria.

We use models trained on the original training and development data to perform NER on the selected subset of the test data. We then generate one adversarial example for each entity in the sentence that was initially predicted correctly. We evaluate to what degree models are fooled only for entities that were predicted correctly in the original sentences. We set the semantic similarity threshold at $\epsilon = 0.8$ following Li et al. [181]. Experiments are run on a GPU machine (NVIDIA Tesla K80). An experiment of three runs (one model on one data set) on one GPU will take roughly 20-24hrs. The models have 110 M parameters.

Human evaluation To evaluate the quality of our adversarial examples from the CoNLL-2003 and BC5CDR data, 100 original sentences and 100 adversarial sentences from each type of attack are scored for grammaticality by human judges. Grammaticality is evaluated on a five-point scale following the reading comprehension benchmark DUC2006 [74]. Our annotators are four volunteering PhD students from our lab who have a background in linguistics[3]: two for each data set with 20% overlap. We choose to present annotators with different original sentences than the ones on which the adversarial sentences they evaluate are based to prevent bias.

6.5. RESULTS

6.5.1. ENTITY CONTEXT ATTACK



(a) BERT

(b) BioBERT and SciBERT

Figure 6.1: Distance of successful synonym replacements relative to the entity (at 0)

With an adversarial context attack, BERT models can be fooled into predicting entities partially or fully wrong (Partial + full success rate) for 87.3% and 93.8% of entities for CoNLL and W-NUT respectively. Moreover, for over 75% of the cases the BERT models were fooled by a single change.

[3]We opted for linguists as they are more acquainted with assessing grammaticality than biomedical domain experts

SciBERT appears more vulnerable than BERT, both to completely being fooled (+6.2 and +6.2% point) and being fooled partially (+9.7 and +7.4 % point) by context attacks. The domain-specific models were also often fooled by only one word being replaced with its synonym; BioBERT was fooled by a single change 65 and 75% of the time whereas SciBERT was fooled by a single change 68 and 76% of the time.

We analyzed the sentence statistics for successful and failed attacks. Specifically for BERT models, we see that the following cases are more vulnerable to attacks: longer sentences; sentences with more words that could be replaced by synonyms; and shorter entities. Manual analysis of successful attacks reveals that BERT models are vulnerable when common words are replaced by rare synonyms (e.g., replacing 'healthy' by 'salubrious').

Figure 6.1a shows where in the sentence changes have occurred in order to fool BERT. BERT models seem most vulnerable to changes in the local context of entities: only 1-2 words left or right of the entity. Manual analysis revealed that these words are often verbs. Although less influential, long distance context does appear to be used for predicting entities in some cases. We manually inspected sentences with long distance changes (>20 words). Lists stood out as a prime example of a sentence type for which long distance context is important (e.g., "The ministry said the group consisted of 13 nuns, seven Italians, and six Zaireans, and four priests, two from Belgium, one from Spain and one from Zambia.").

For the BioBERT model, the distribution is strikingly similar to that of the original BERT model (see Figure 6.1b). This is likely due to either the vocabulary or the training data[4] that these models share. SciBERT models which share neither training data nor vocabulary with the original BERT model are even more vulnerable to changes in the local context of the entity (see Figure 6.1b).

6.5.2. EVALUATING THE NECESSITY OF IMPORTANCE RANKING

To investigate the effect of adding the word importance ranking to the entity context attack, we perform an ablation study on the CoNLL-2003 test set. As can be seen in Table 6.5, removing the word importance ranking leads to a stark drop in both the average full success of adversarial attacks (from 37.3% to 9.5%) and the average partial success rate (from 52.8% to 20.1%). The number of words that need to be perturbed also drops, by 6.9% point, meaning that attacks require fewer synonym replacements on average to be successful. Thus, it appears that the word importance ranking is crucial to the success of the adversarial attack algorithm.

6.5.3. ENTITY ATTACK

The main results of adversarial *entity* attack on BERT models are presented in Table 6.6. BERT models appear highly vulnerable to adversarial attacks on the entities themselves despite the high similarity between adversarial and original sentences. On average, BERT models are fooled for 97.5% of entities that were initially predicted correctly on the CoNLL data and 89.2% on W-NUT data. BERT models appear even more vulnerable to entity attacks on domain-specific data with success rates above 99%.

[4] BioBERT includes all the original BERT training data as well as additional domain-specific data

	Importance ranking	
	Yes	No
Success rate (%)	37.3 ± 0.515	9.5 ± 4.3
Partial success rate (%)	52.8 ± 0.356	20.1 ± 9.2
Semantic similarity	0.922 ± 0.006	0.983 ± 0.006
Words perturbed (%)	13.8 ± 0.238	6.9 ± 2.2

Table 6.5: Comparison of context attacks with and without importance ranking on CoNLL-2003 data

	CoNLL-2003	W-NUT 2017	NCBI-disease	BC5CDR
	BERT	BERT	BERT	BERT
Success rate (%)	97.5 ± 0.037	89.5 ± 0.886	99.2 ± 0.114	99.4 ± 0.073
Of which:				
– Missed entity(%)	21.3 ± 12.3	71.4 ± 1.9	100	86.1 ± 0.5
– Entity type error(%)	78.8 ± 12.3	28.6 ± 1.9	0	13.9 ± 0.5
Median semantic similarity	0.959 ± 0.001	0.928 ± 0.003	0.952 ± 0.001	0.962 ± 0.000

Table 6.6: Automatic evaluation results for entity attacks on BERT models. Results are the mean of three models.

Table 6.7 shows that domain-specific BERT models do not resolve this issue. They are also highly vulnerable with over 99% of all initially correctly predicted entities now predicted incorrectly. The high success rates of entity attacks both on general domain and domain-specific data suggest that BERT models, similar to traditional models, are unable to predict entities correctly based solely on the context of the entity. Replacing the entity word itself with another of the same entity type, with the context unchanged, can easily fool the model. This suggests a strong dependency on the entities that the model has seen previously, making these models vulnerable to new or emergent entities.

This is corroborated by an analysis of which entities were chosen in successful attacks. For all BERT models and all data sets, except for the CoNLL data, these entities are significantly less frequent in training and development data than the original entities according to Wilcoxon signed rank tests ($p < 0.001$).

A possible explanation for why BERT models for CoNLL are the exception is that there is a stronger match between the pretraining data and the data at hand than for the other data sets. This may make the model less vulnerable to infrequent entities, despite not being less vulnerable to entity replacement overall. Manual inspection further revealed that BERT models appear to be sensitive to the capitalization of entities (e.g., BERT models trained on CoNLL were fooled by transforming 'New York' to 'NEW YORK').

6.5.4. RESULTS OF HUMAN EVALUATION

On CoNLL-2003, the annotators have a fair inter-annotator agreement (weighted κ = 0.353). On BC5CDR, the inter-annotator agreement is slight (weighted κ = 0.177). Investigation of the annotations reveals that this is most likely because biomedical sentences are more difficult to assess for laymen. Because of the limited agreement, we report grammaticality assessments per annotator.

	NCBI-disease		BC5CDR	
	BioBERT	SciBERT	BioBERT	SciBERT
Success rate (%)	99.2 ± 0.259	99.4 ± 0.054	99.4 ± 0.070	99.3 ± 0.089
Of which:				
– Missed entity(%)	100	100	94.0 ± 0.7	91.7 ± 1.5
– Entity type error(%)	0	0	6.0 ± 0.7	8.3 ± 1.5
Median semantic similarity	0.955 ± 0.001	0.953 ± 0.002	0.961 ± 0.001	0.962 ± 0.001

Table 6.7: Automatic evaluation results for entity attacks on biomedical BERT models. Results are the mean of three models.

	CoNLL		BC5CDR	
Annotator	1	2	3	4
Original	3.51	4.34	4.43	4.78
After context attack	3.05*	3.68**	3.86**	4.35**
After entity attack	3.30	4.37	3.85	4.67

Table 6.8: Mean grammaticality of the original and adversarial sentences. $^{*}p < 0.05$ $^{**}p < 0.01$ compared to the original sentences according to Mann-Whitney U tests.

Table 6.8 shows that although entity attacks do not significantly alter the grammaticality of the sentences, attacks on the context of the entity do. Although this reduction is consistent across data sets, the mean grammaticality of the adversarial sentences remains above 3 (acceptable) and the mean absolute reduction is less than a full point.

6.6. DISCUSSION AND LIMITATIONS

We manually analyzed the generated adversarial examples and found that our adversarial examples are susceptible to word sense ambiguity. For example, the top 50 synonyms for 'surfed' in 'surfed the Internet' includes both correct synonyms like 'googled' and incorrect ones like 'paddled'. There are also some cases where adversarial examples suffer from foreign words in the Paragram-SL999 word vectors [211]. Occasionally synonyms are not English words (e.g., 'number' to 'nombre'), or synonym choice is influenced by words that occur in multiple languages e.g., 'vie' in 'to vie for top UN post' is replaced with 'existence' which is a synonym of the French 'vie' (i.e., life).

Furthermore, our adversarial examples are susceptible to grammatical errors. Grammatically poor adversarial sentences often suffer from changes from verbs to nouns or vice versa that are not caught by the POS-filter (e.g., 'open' to 'openness' and 'influence' to 'implication'). These cases may be particularly difficult as 'open' and 'influence' can be both a verb and an adjective or noun. Another common error is singular-plural inconsistencies (e.g., 'one dossiers'). To mitigate these issues, future work could focus on removing non-English words from the embedding space, and altering how the POS-tag of the synonym is determined.

We find that semantic consistency can be an issue with broad entity types like location when attacking the entity itself. For example, in one case the country "U.S." is replaced by the village "Tavildara" (in Tajikistan). For more specific entity types like Disease, Chemical

or Person we do not encounter inconsistencies with subtypes of an entity category. On the contrary, often replacements are semantically close to the original. For instance, the anti-epileptic drug "clonazepam" was replaced by the anti-epileptic drug "lorazepam" and "Washington" in "Washington administration" was replaced by "Clinton".

Moreover, there are some caveats to keep in mind when interpreting weaknesses based on successful attacks. The architecture of self-attentive models means that the attention weight of a word is context-dependent. Thus, if changing that word fools the model, this might only be true in that context. Additionally, if multiple words were changed for a successful attack, their interaction may contribute to the success and it cannot simply be interpreted as caused by this combination of words.

6.7. CONCLUSIONS

We studied the vulnerability of BERT models in NER tasks under a black-box setting. Our experiments show that BERT models can be fooled by changes in single context words being replaced by their synonyms. They are even more vulnerable to entities being replaced by less frequent entities of the same type.

Our analysis of BERT's vulnerabilities can inform fruitful directions for future research. Firstly, our results reveal that rare or emergent entities remain a problem for both generic and domain-specific NER models. Consequently, we recommend further research into zero or few-shot learning. Moreover, the masking of entities during fine-tuning may be an interesting avenue for research. Secondly, BERT models also appear vulnerable to words it has not seen or rarely seen during training in the entity context. To combat this vulnerability, the use of adversarial examples designed specifically to include more infrequently used words could be explored. Another possible avenue for research could be alternative pre-training schemes for BERT such as curriculum learning [99]. Thirdly, we find that SciBERT is more vulnerable to changes in the entity context than BioBERT or BERT. This may be due to the domain-specific biomedical vocabulary that SciBERT employs, which could make it more vulnerable to out-of-entity words being replaced by more common English terms. This trade-off between robustness and domain-specificity of BERT models may be another worthwhile research direction.

We consider our work to be a step towards understanding to what extent BERT models for NER are vulnerable to token-level changes and to which changes they are most vulnerable. We hope others will build on our work to further our insight into self-attentive models and to mitigate these vulnerabilities.