Breaking BERT: Understanding its vulnerabilities for biomedical named entity recognition through adversarial attack

Dirkson, A.R.; Verberne, S.; Kraaij, W.

Breaking BERT: Understanding its Vulnerabilities for Biomedical Named Entity Recognition through Adversarial Attack

Anne Dirkson **Suzan Verberne** **Wessel Kraaij**
Leiden Institute of Advanced Computer Science, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, the Netherlands
{a.r.dirkson, s.verberne, w.kraaij}@liacs.leidenuniv.nl

Abstract

Biomedical named entity recognition (NER) is a key task in the extraction of information from biomedical literature and electronic health records. For this task, both generic and biomedical BERT models are widely used. Robustness of these models is vital for medical applications, such as automated medical decision making. In this paper we investigate the vulnerability of BERT models to variation in input data for NER through adversarial attack. Since adversarial attack methods for NER are sparse, we propose two black-box methods for NER based on existing methods for classification tasks. Experimental results show that the original as well as the biomedical BERT models are highly vulnerable to entity replacement: They can be fooled in 89.2 to 99.4% of the cases to mislabel previously correct entities. BERT models are also vulnerable to variation in the entity context with 20.2 to 45.0% of entities predicted completely wrong and another 29.3 to 53.3% of entities predicted wrong partially. Often a single change is sufficient to fool the model. BERT models seem most vulnerable to changes in the local context of entities. Of the biomedical BERT models, the vulnerability of BioBERT is comparable to the original BERT model whereas SciBERT is even more vulnerable. Our results chart the vulnerabilities of BERT models for biomedical NER and emphasize the importance of further research into uncovering and reducing these weaknesses.

1 Introduction

Biomedical named entity recognition (NER) is a key task in the extraction of information from biomedical literature and electronic health records. Both generic and biomedical BERT models are widely used for biomedical NER due to their excellent performance. However, the robustness of BERT-based models is contested: Various studies recently showed that BERT is vulnerable to adversarial attacks (Jin et al., 2020; Hsieh et al., 2019; Li et al., 2020; Zang et al., 2020; Sun et al., 2020). 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. Also in the biomedical domain it was recently shown that BERT models are also vulnerable to adversarial attack when used for biomedical text classification (Mondal, 2021) and named entity recognition (NER) (Araujo et al., 2020).

Yet, the robustness of BERT models for biomedical NER is vital for downstream medical applications, such as automated medical decision making (Hengstler et al., 2016). At present there is only a single study on the topic: Araujo et al. (2020) attack biomedical BERT models by simulating spelling errors and replacing entities with their synonyms. They find that both attacks drastically reduce performance on medical NER tasks.

Here, we aim to systematically test the robustness of BERT models for biomedical 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: using synonyms in the *context* of entities and replacing 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.

Using these methods, we compare generic and domain-specific BERT models for biomedical NER to understand to what extent domain-specific training impacts the vulnerability of these models. We also compare the robustness of generic BERT mod-

els on general and biomedical NER data to explore to what extent the vulnerability is specific to biomedical data.

We thus address the following research questions:

1. How vulnerable are BERT models to adversarial attack on general and biomedical 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 span multiple words. To ensure that labels remain accurate in our adversarial sentences, we substitute entities only by entities of the same type when attacking the entity itself (i.e. an *entity attack*) and constrain synonym replacements to non-entity words when altering the context of the entity (i.e. an *entity context attack*).

We assume a black-box setting, which means that the adversarial method has no knowledge of the data, parameters or model architecture (Alzantot et al., 2018). 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 paper are twofold: (1) We adapt existing adversarial attack methods to sequence labelling tasks and (2) we evaluate the vulnerability of general and biomedical BERT models for biomedical NER. We make our code available for follow up research.[1]

2 Related Work

Within the field of NLP, token-level black-box methods for adversarial attack have mainly been developed for classification and textual entailment (Alshemali and Kalita, 2019). Substituting tokens with their synonyms is the most popular choice for perturbing at the token level. Synonyms are often found using nearest neighbours 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 (Jin et al., 2020; Li et al., 2019; Hsieh et al., 2019) require a minimal semantic similarity between the generated and original sentence. Additionally, some methods (Alzantot et al., 2018; Jin et al., 2020) use word embeddings with additional synonymy constraints (Mrkšić et al., 2016). 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 (Alshemali and Kalita, 2019). 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 (Sun et al., 2020; Hsieh et al., 2019; Jin et al., 2020; Li et al., 2020; Zang et al., 2020; Araujo et al., 2020; Balasubramanian et al., 2020; Mondal, 2021). Nonetheless, the only study that has attacked BERT models for NER is the study by Araujo et al. (2020). 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. (2020) 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 on the most severe stress conditions, while Araujo et al. (2020) evaluate the scenario where the input data is noisy due to spelling mistakes and 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 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 is available at: https://github.com/AnneDirkson/breakingBERT

3 Methods

In this section, we describe two methods for generating adversarial examples designed to fool NER models, namely through (1) entity replacement (*entity attack*) and (2) synonym replacements in the entity context (*entity context attack*). These are described in Sections 3.1.2 and 3.1.3, respectively.

3.1 Aim of the attacks

We aim to generate adversarial examples in which the entire 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 1: Here, the start of the entity is mislabelled but the last part of the entity is still recognized. We consider this a partial success.

3.1.1 Metrics for evaluation of the attacks

The success of the attack and thus the vulnerability of the model is evaluated by the percentage of entities that were originally correctly labelled but are mislabelled after attack. For *entity context attacks* entities can also be partially mislabelled i.e. only some words in the entity are mislabelled. 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 mislabelled. For *entity 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 (Jin et al., 2020).

3.1.2 Entity Attack

To explore to what extent models rely on the words of the entity, 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 filter the generated sentences for a sufficiently high semantic similarity to the original sentence. Semantic similarity is calculated with the Universal Sentence Encoder (USE) (Cer et al., 2018). The USE encodes the sentences into 512 dimensional vectors and their cosine similarity (normalized between 0 and 1) approximates their semantic similarity. The resulting sentences need to be above the semantic similarity threshold ϵ. After filtering, we check if the predicted label of any of the replacements 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.

3.1.3 Entity Context Attack

To investigate the impact of the context on the correct labelling of the entity, we adapt the method of Jin et al. (2020), which was designed for text classification, to sequence labelling 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 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) of the *correct* label before and after deleting the word from the sentence (Jin et al., 2020). If the deletion of the word leads to an *incorrect* label for the entity token, the importance of the word is increased by also 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 labelling 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:

$$
\begin{aligned}
I_w = & F_Y(X) - F_Y(X_{-w}) \\
& \text{if} F(X_{-w}) = Y \vee (F(X) = Y_I \wedge F(X_{-w}) = Y_B) \\
& F_Y(X) - F_Y(X_{-w}) + F\_\bar{Y}(X_{-w}) - F_{\bar{Y}}(X) \\
& \text{if} F(X_{-w}) \neq Y
\end{aligned}
\tag{1}
$$

where F_Y is the prediction score for the correct label, $F_{\bar{Y}}$ is the prediction score of the predicted label, F is the predicted label, Y is the correct label,

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

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 (Mrkšić et al., 2016) with a similarity to the original word above the threshold δ. Mrkšić et al. (2016) 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 (Hill et al., 2015) and were also used by Jin et al. (2020) and Alzantot et al. (2018). We chose 0.5 as the minimal similarity threshold δ for synonym selection in contrast to the threshold of 0 used by previous work in order to better guarantee semantic similarity. Regardless of δ, a maximum of 50 synonyms are selected. Examples of word pairs above a δ of 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 also 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 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 labelled words in the entity are considered for subsequent iterations.

4 Experiments

4.1 Data

We use two general-domain NER data sets for evaluating our method: the English CoNLL-2003 data (Tjong Kim Sang and de Meulder, 2003) and the W-NUT 2017 data (Derczynski et al., 2017). The goal of the latter was to investigate recognition of unusual, previously-unseen entities in the context of online discussions. Additionally, we use two data sets from the biomedical domain: BC5CDR (Li et al., 2016) and the NCBI disease corpus (Doğan et al., 2014). Both data sets have been used to evaluate domain-specific BERT models for NER in the biomedical domain (Peng et al., 2019; Beltagy et al., 2019; Lee et al., 2019). See Table 2 for more details on the data sets.

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 (Wolf et al., 2019). We set the learning rate at 5×10^{-5} and optimized the number of epochs using the range (3,4) as described in Devlin et al. (2019) 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) (Lee et al., 2019) and SciBERT (scivocab-cased) (Beltagy et al., 2019). Each model is trained in three folds with three different initialization seeds (1, 2 & 4).

Domain	Dataset	Entity types	Dev.	Train	Test	Subset for Eval.* (# of 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 2: Size of the data sets (number of sentences). *This subset is used for automatic evaluation

4.3 Evaluation of the adversarial attacks

Automatic evaluation We randomly select 500 eligible sentences from each test set. Table 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. (2019).

Human evaluation In order 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 (Dang, 2006). Our annotators are four linguists[2]: 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.

5 Results

5.1 Entity Attack

The main results of adversarial entity attack on BERT models are presented in Table 3. 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%.

Table 4 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 suggests 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 with another of the same entity type 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 with the exception of the CoNLL data, these entities are significantly less frequent in training and development data than the original entities.

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').

[2]We opted for linguists as they would be more acquainted with assessing the grammaticality of a sentence than domain experts

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

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

5.2 Entity Context Attack

Attacks on the context of entities are less successful at completely fooling BERT models than entity attacks (see Table 5). Nonetheless, 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.

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 attack. Also the domain-specific models were 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 to rare synonyms replacing more common words (e.g. 'salubrious' instead of 'healthy').

Figure 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. For the BioBERT model, the distribution is strikingly similar to that of the original BERT model (see Figure 1b). This is likely due to either the vocabulary or the training data[3] 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 1b).

5.3 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 7, removing the word importance ranking leads to a stark drop in both the average full success of adversarial attack (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 need less changes on average to be successful. Thus, it appears that the word importance ranking is crucial to the success of the adversarial attack algorithm.

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

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

	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 7: Comparison of the effect of context attacks with and without importance ranking on CoNLL-2003 data

	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 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.

5.4 Results of human evaluation

On CoNLL-2003, the annotators have a fair inter-annotator agreement (weighted $\kappa = 0.353$). On BC5CDR, the inter-annotator agreement is slight (weighted $\kappa = 0.177$). 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.

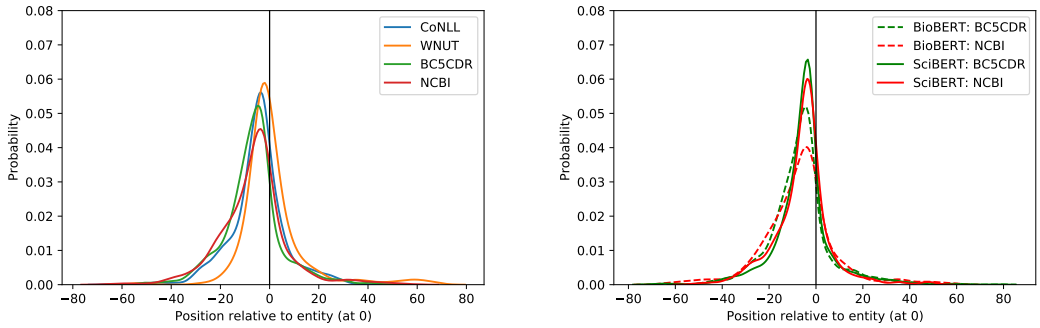
Table 8 show that although entity attacks do not significantly alter the grammaticality of the sentences, attacks on the context of the entity do. Although this reduction is consistent across data sets, the mean grammaticality of the adversarial sentences remains above 3 (acceptable) and the mean absolute reduction is less than a full point. Table 9 shows examples of adversarial sentences and their grammaticality judgments.

6 Discussion

6.1 Error analysis

We manually analysed the generated adversarial examples and found that our adversarial examples are susceptible to word sense ambiguity. For example, the top 50 synonyms for 'pound' in '10 mg/pound' contains both correct synonyms like 'lb' and incorrect ones like 'sterling'. Moreover, occasionally synonyms are not English words (e.g. 'number' to 'nombre', 'double' to 'doble' and 'life' to 'vie' or 'vida'). These issues are caused by foreign words influencing or being included in the Paragram-SL999 word vectors (Mrkšić et al., 2016).

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



(a) BERT

(b) BioBERT and SciBERT

Figure 1: Distance of successful synonym replacements relative to the entity (at 0)

Attack			Grammaticality
Context	Orig.	Major bleeding is of primary **concern** in patients receiving heparin therapy	
	Adv.	Major bleeding is of primary **preoccupation** in patients receiving heparin therapy	Good
	Orig.	Both drugs **influence** inositol metabolism	
	Adv.	Both drugs **implication** inositol metabolism	Poor
Entity	Orig.	Careful observation of patients receiving high-dose **MP** is recommended	
	Adv.	Careful observation of patients receiving high-dose **BEA** is recommended	Good
	Orig.	A man, 25 years old, was admitted to hospital with high fever, chill, vomiting, **jaundice**	
	Adv.	A man, 25 years old, was admitted to hospital with high fever, chill, vomiting, **arrhythmic**	Poor

Table 9: Examples of adversarial sentences from BC5CDR data. Sentences were selected from adversarial sentences (Adv.) with the lowest (=1) and highest (=5) grammaticality scores.

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.

Specifically for the biomedical adversarial examples we also see that synonyms are not always appropriate for the context (e.g. 'worst affected' is changed to 'shittiest affected'). Abbreviations are also particularly difficult in this regard. A domain-specific embeddings model may partly alleviate this problem.

6.2 Interpreting vulnerabilities

Analysis of which perturbations are able to fool the model might give valuable insight into its vulnerabilities and robustness to real-world data. However, 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 interactions may contribute to the success and it cannot simply be interpreted as caused by this combination of words.

7 Conclusions

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

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 biomedical NER models. Consequently, we recommend fur-

ther 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 (Elman, 1993). Thirdly, we find that SciBERT is more vulnerable to changes in the entity context than BioBERT or BERT. This may be due to the 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 first step towards understanding to what extent BERT models are vulnerable to token-level changes in biomedical NER tasks 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.

References

Basemah Alshemali and Jugal Kalita. 2019. Improving the reliability of deep neural networks in NLP: A review. *Knowledge-Based Systems*, 10520.

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

Vladimir Araujo, Andrés Carvallo, and Denis Parra. 2020. Adversarial evaluation of bert for biomedical named entity recognition. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 79–82.

Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi. 2020. What's in a Name? Are BERT Named Entity Representations just as Good for any other Name? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 205–214, Stroudsburg, PA, USA. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Hoa Trang Dang. 2006. Overview of DUC 2006. In *Proceedings of HLT-NAACL 2006*, pages 1–12.

Leon Derczynski, Eric Nichols, and Marieke van Erp. 2017. Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In *Proceedings ofthe 3rd Workshop on Noisy User-generated Text*, pages 140–147. Association for Computational Linguistics.

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

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Monika Hengstler, Ellen Enkel, and Selina Duelli. 2016. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, 105(C):105–120.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. 2019. On the Robustness of Self-Attentive Models. In *Proceedings of the 57th Annual Meeting ofthe Association for Computational Linguistics*, pages 1520–1529. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey A Tianyi Zhou, and Peter Szolovits. 2020. Is BERT Really Robust? A Strong

Baseline for Natural Language Attack on Text Classification and Entailment. In *AAAI*, page Accepted for publication.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Btz682.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:68.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. TEXTBUGGER: Generating Adversarial Text Against Real-world Applications. In *Network and Distributed Systems Security (NDSS) Symposium 2019*.

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

Ishani Mondal. 2021. BBAEG: Towards BERT-based biomedical adversarial example generation for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5378–5384, Online. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašíc, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting Word Vectors to Linguistic Constraints. In *Proceedings of NAACL-HLT 2016*, pages 142–148.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*.

Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. 2020. Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT. *arXiv*.

Erik Tjong Kim Sang and Fien de Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Transformers: State-of-the-art Natural Language Processing. *ArXiv*.

Yuan Zang, Bairu Hou, Fanchao Qi, Zhiyuan Liu, Xiaojun Meng, and Maosong Sun. 2020. Learning to attack: Towards textual adversarial attacking in real-world situations. *arXiv*.