Adversarial Evaluation of BERT for Biomedical Named Entity Recognition

Vladimir Araujo
Pontificia Universidad Católica de Chile
IMFD
vgaraujo@uc.cl

Andrés Carvallo
Pontificia Universidad Católica de Chile
IMFD
afcarvallo@uc.cl

Denis Parra
Pontificia Universidad Católica de Chile
IMFD
dparra@ing.puc.cl

Abstract

The success of pre-trained word embeddings of the BERT model has motivated its use in tasks in the biomedical domain. However, it is not clear if this model works correctly in real scenarios. In this work, we propose an adversarial evaluation scheme in a BioNER dataset, which consists of two types of attacks inspired by natural spelling errors and synonyms of medical terms. Our results indicate that under these adversarial settings, the performance of the models drops significantly. Despite the result, we show how the robustness of the models can be significantly improved by training them with adversarial examples.

1 Background

Biomedical Natural Language Processing (BioNLP) is the field concerned with developing tools and methodologies for processing biomedical textual information and generally applied to tasks such as Named Entity Recognition (NER), Sentence Similarity and Relation Extraction. In order to encourage the development of this area, public datasets and challenges have been shared with the community, such as BC5CDR (Wei et al., 2015), CLEF (Suominen et al., 2013), BioSSES (Soğancıoğlu et al., 2017), ChemProt (Kringelum et al., 2016) and i2b2 (Özlem Uzuner et al., 2011).

At the same time, general-purpose neural language models have recently shown significant progress with the introduction of ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018). These models have obtained remarkable results in several tasks. A natural choice has been to apply these models to BioNLP. As a result, several pre-trained models with medical corpus have been released, such as BioBERT (Lee et al., 2019), ClinicalBERT (Alsentzer et al., 2019), and BlueBERT (Peng et al., 2019).

Adversarial Examples have demonstrated the risk of using machine learning systems in real-world applications (Szegedy et al., 2014; Goodfellow et al., 2014). This evaluation strategy showed that slight disturbances in the input could cause severe failures in computer vision models. More recently, adversarial attacks have been applied to several NLP benchmarks (Jin et al., 2019; Aspillaga et al., 2020).

This type of evaluation has become relevant in the biomedical domain because an erroneous prediction could be very harmful to patients (Sun et al., 2018). Despite the existence of deployed systems in real-world clinical settings, researchers have shown that even the state of the art models in medical computer vision (Paschali et al., 2018; Finlayson et al., 2019; Ma et al., 2019) are vulnerable to adversarial attacks. However, perturbation methods developed for images cannot be directly applied to texts. Because of that, we proposed adversarial examples to evaluate a biomedical text mining task. Specifically, we evaluated the BlueBert model in the BioNER task.

Table 1: Adversarial Evaluation Sentence Examples

<table>
<thead>
<tr>
<th>Type</th>
<th>Original</th>
<th>Swap Noise</th>
<th>Keyboard Typo Noise</th>
<th>Synonymy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>Two mothers with heart valve prosthesis were treated with warfarin during pregnancy.</td>
<td>Two mothers with herat vavle protshesis were terated with warafarin during preganncy.</td>
<td>Two mothers with hea5t valce prosth3sis were trezted with warfsrin during pregnahcy.</td>
<td>Two mothers with heart valve prosthesis were treated with potassium warfarin during pregnancy.</td>
</tr>
</tbody>
</table>
2 Adversarial Evaluation

We propose a black-box attack methodology, which does not require the inner details of the model to generate adversarial examples (Ilyas et al., 2018). Specifically, we focus on making disturbances in the input data (edit adversaries) that could cause the models to fall into erroneous predictions (Table 1).

**Noise Adversaries** Motivated by the above and inspired by (Belinkov and Bisk, 2018), we constructed adversarial examples that try to emulate spelling errors committed by human beings. These edit adversaries consist of two types of alterations: (i) **Swap Noise**: For each word, one random pair of consecutive characters is swapped, (ii) **Keyboard Typo Noise**: For each word, one character is replaced by an adjacent character in traditional English keyboards.

**Synonymy Adversaries** These examples test if a model can understand synonymy relations. Replacing a medical term with an equivalent synonym is challenging. For that reason, we focus only on words of chemicals and diseases. We use PyMedTermino (Jean-Baptiste et al., 2015), which uses the biomedical vocabulary of UMLS (Bodenreider, 2004), to find the most similar or related words (synonyms) to the retrieved words. Finally, we replace the synonym found depending on whether it is a disease or chemical.

<table>
<thead>
<tr>
<th>Table 2: Adversarial Evaluation Sentence Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
</tr>
<tr>
<td>Test Set</td>
</tr>
<tr>
<td>Precision</td>
</tr>
<tr>
<td>Recall</td>
</tr>
<tr>
<td>F1-Score</td>
</tr>
</tbody>
</table>

3 Experiments

**Experimental Setup** We use the BC5CDR dataset (Wei et al., 2015) for the BioNER task, which consists of 1500 PubMed (Fiorini et al., 2018) articles with 4409 annotated chemicals and 5818 diseases. We evaluated the base version of the pre-trained BlueBERT model because it has been shown to perform better than its namesakes (Wada et al., 2020). We fine-tune the model with the original training set from each task for ten epochs, then evaluate them with the original test set and the adversarial sets.

**Results on Adversarial Evaluation** Table 2 shows the classification results of the BC5CDR task on the original test set and our adversarial examples. We see that the performance of BERT drops across all adversarial attacks. However, the task of recognizing the disease was the most affected. In the case of the chemical recognition task, the model shows a drop of approximately 20% of the F1 score. In contrast, the F1 score of the disease recognition task falls dramatically, below 50% of the original score.

**Adversarial Training Results** Training with adversarial examples is a methodology used in previous works (Belinkov and Bisk, 2018; Jia and Liang, 2017) to create robustness in neural language models. It ensures that the model is exposed to samples outside the training distribution and provides a form of regularization (Belinkov and Bisk, 2018). We first fine-tune the model with the original training set plus an adversarial version of the same set. Then we carry out the adversarial evaluation to measure how the models perform in the different test sets. Table 3 shows the results for NER of training with adversaries and testing with the original set compared with their respective adversaries. We see that training with adversarial examples significantly improves the robustness of the models to adversarial attacks, without significant impact on the original non-adversarial task.

<table>
<thead>
<tr>
<th>Table 3: Adversarial Training Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
</tr>
<tr>
<td>Test Set</td>
</tr>
<tr>
<td>Precision</td>
</tr>
<tr>
<td>Recall</td>
</tr>
<tr>
<td>F1-Score</td>
</tr>
</tbody>
</table>
4 Conclusions

In this paper, we investigated how the state-of-the-art model, BERT, is robust or brittle to simple adversarial attacks in a BioNER task. Our experimental results suggest the necessity of considering the robustness of the neural models for use in the biomedical field.

For future work, we plan to explore other tasks related to medicine. Also, investigate further why there is a different drop in performance between adversarial example types and datasets.

Acknowledgements

This work has been partially funded by Millennium Institute for Foundational Research on Data (IMFD).

References


Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. 2019. Understanding adversarial attacks on deep learning based medical image analysis systems.


Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2020. A pre-training technique to localize medical bert and enhance biobert.