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 adversial examples to evaluate a biomedical text mining task. Specifically, we evaluated the BlueBert model in the BioNER task.

Table 1: Adversarial Evaluation Sentence Examples

Original	Two mothers with heart valve prosthesis were treated with warfarin during pregnancy.
Swap Noise	Two mothers with herat vavle protshesis were terated with warafrin during pregannсy.
Keyboard Typo Noise	Two mothers with hea5t valce prosth3sis were trezted with warfsrin during pregnahcy.
Synonymy	Two mothers with heart valve prosthesis were treated with potassium warfarin during pregnancy.

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 2: Adversarial Evaluation Sentence Examples

Training Set	**BC5CDR Chemical**				**BC5CDR Disease**			
Test Set	**Orig**	**Keyb**	**Swap**	**Syno**	**Orig**	**Keyb**	**Swap**	**Syno**
Precision	.895	.734	.609	.730	.832	.543	.636	.337
Recall	.908	.683	.559	.748	.844	.278	.337	.390
F1-Score	.901	.708	.583	.739	.838	.368	.441	.362

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 3: Adversarial Training Results

Training Set	**BC5CDR Chemical + Keyboard**		**BC5CDR Chemical + Swap**		**BC5CDR Chemical + Synonymy**		**BC5CDR Disease + Keyboard**		**BC5CDR Disease + Swap**		**BC5CDR Disease + Synonymy**	
Test Set	**Orig**	**Keyb**	**Orig**	**Swap**	**Orig**	**Syno**	**Orig**	**Keyb**	**Orig**	**Swap**	**Orig**	**Syno**
Precision	.889	.850	.895	.684	.899	.872	.839	.723	.836	.773	.813	.788
Recall	.906	.792	.902	.630	.901	.908	.848	.712	.847	.746	.824	.841
F1-Score	.898	.820	.898	.656	.900	.890	.844	.717	.841	.759	.818	.814

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

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Carlos Aspillaga, Andrés Carvallo, and Vladimir Araujo. 2020. Stress test evaluation of transformer-based models in natural language understanding tasks. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1882–1894, Marseille, France, May. European Language Resources Association.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

O. Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D–270, January.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. 2019. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, March.

Nicolas Fiorini, Robert Leaman, David J Lipman, and Zhiyong Lu. 2018. How user intelligence is improving PubMed. *Nature Biotechnology*, 36(10):937–945, October.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information.

Lamy Jean-Baptiste, Venot Alain, and Duclos Catherine. 2015. Pymedtermino: an open-source generic api for advanced terminology services. *Studies in Health Technology and Informatics*, 210(Digital Healthcare Empowering Europeans):924–928.

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

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? a strong baseline for natural language attack on text classification and entailment.

J. Kringelum, S. K. Kjaerulff, S. Brunak, O. Lund, T. I. Oprea, and O. Taboureau. 2016. Chemprot-3.0: a global chemical biology diseases mapping. *Database*.

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*, 09.

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.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, June.

Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. 2018. Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 493–501. Springer International Publishing.

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 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Mengying Sun, Fengyi Tang, Jinfeng Yi, Fei Wang, and Jiayu Zhou. 2018. Identify susceptible locations in medical records via adversarial attacks on deep predictive models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, page 793–801, New York, NY, USA. Association for Computing Machinery.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *Proceedings of the 4th International Conference on Information Access Evaluation. Multilinguality, Multimodality, and Visualization - Volume 8138*, CLEF 2013, page 212–231, Berlin, Heidelberg. Springer-Verlag.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

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.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2015. Overview of the biocreative v chemical disease relation (cdr) task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, volume 14.