Stress Test Evaluation of Biomedical Word Embeddings

Vladimir Araujo[1,2], Andrés Carvallo[1,2], Carlos Aspillaga[1], Camilo Thorne[3], Denis Parra[1,2]
[1]Pontificia Universidad Católica de Chile
[2] Millennium Institute for Foundational Research on Data (IMFD)
[3]Elsevier
{vgaraujo,afcarvallo,cjaspill}@uc.cl
c.thorne.1@elsevier.com
dparra@ing.puc.cl

Abstract

The success of pretrained word embeddings has motivated their use in the biomedical domain, with contextualized embeddings yielding remarkable results in several biomedical NLP tasks. However, there is a lack of research on quantifying their behavior under severe "stress" scenarios. In this work, we systematically evaluate three language models with adversarial examples – automatically constructed tests that allow us to examine how robust the models are. We propose two types of stress scenarios focused on the biomedical named entity recognition (NER) task, one inspired by spelling errors and another based on the use of synonyms for medical terms. Our experiments with three benchmarks show that the performance of the original models decreases considerably, in addition to revealing their weaknesses and strengths. Finally, we show that adversarial training causes the models to improve their robustness and even to exceed the original performance in some cases.

1 Introduction

Biomedical NLP (BioNLP) is the field concerned with developing NLP tools and methods for the life sciences domain. Some applications of these techniques include e.g., discovery of gene-disease interactions (Pletscher-Frankild et al., 2015), development of new drugs (Tari et al., 2010), or automatic screening of biomedical documents (Carvallo et al., 2020). With the exponential growth of digital biomedical literature, the importance of BioNLP has become especially relevant as a tool to extract relevant knowledge for making decisions in clinical settings as well as in public health. In order to encourage the development of this area, public datasets and challenges have been shared with the community to solve these tasks, such as BioSSES (Soğancıoğlu et al., 2017), HOC (Hanahan and Weinberg, 2000), ChemProt (Kringelum et al., 2016) and BC5CDR (Li et al., 2016), among others. At the same time, neural language models have shown significant progress since the introduction of models such as W2V (Mikolov et al., 2013), and more recent models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). These models, trained over large corpora (MEDLINE and PubMed in the biomedical domain) have obtained remarkable results in most NLP tasks, including BioNLP benchmarks (Peng et al., 2019). However, they have not been systematically evaluated under severe stress conditions to test their robustness to specific linguistic phenomena. For this reason, the objective of this paper is to evaluate three well-known neural language models under stress conditions. As a case study, we evaluate NER benchmarks since it a key BioNLP information extraction task.

Our stress test evaluation is inspired by the work of Naik et al. (2018), which proposes the use of adversarial evaluation for natural language inference by adding distractions in sentences, and evaluating models on this test set. We propose an adversarial evaluation black-box methodology, which does not require access to the inner workings of the models in order to generate adversarial examples (Zhang et al., 2019). Specifically, we make perturbations to the input data, also known as edit adversaries, that could cause the models to fall into erroneous predictions. Additionally, we train the models with the proposed adversarial examples, which is a methodology used in previous works (Belinkov and Bisk, 2018; Jia and Liang, 2017) to strengthen the neural language models during the training process. We hope that our work will motivate the development and use of adversarial examples to evaluate models and obtain more robust biomedical embeddings.

2 Related Work

Adversarial Evaluation of NLP Models One way to test NLP models is by using adversarial tests, which consist of applying intentional distur-

Proceedings of the BioNLP 2021 workshop, pages 119–125

Original (O)	Linoleic acid autoxidation inhibitions on all fractions were higher than that on alpha-tocopherol.
Keyboard (K)	Linoleic avid autoxidatiob inh9bitions on all fractjons were higher than that on zlpha-toclpherol.
Swap (W)	Linoleic aicd autoxidtaion inhibtiions on all fractoins were higher than that on aplha-tocohperol.
Synonymy (S)	Linoleic acid autoxidation inhibitions on all fractions were higher than that on vitamin E.

Table 1: Examples of sentences of the stress tests.

bances to a gold standard, to test whether the attack leads the models into incorrect predictions. Previous works on adversarial attacks have demonstrated how dangerous it can be to use machine learning systems in real-world applications (Szegedy et al., 2014; Goodfellow et al., 2014). Indeed, it is known that even small amounts of noise can cause severe failures in neural computer vision models (Akhtar and Mian, 2018). However, such failures can be mitigated through adversarial training (Goodfellow et al., 2014). These properties have in turn motivated novel adversarial strategies designed for various NLP tasks (Zhang et al., 2019), as well as work on adversarial attacks focused on recurrent and transformer networks applied to *generic* NLP benchmarks (Aspillaga et al., 2020).

Evaluation of Biomedical Models Models used in BioNLP tasks elicit particular interest in this context because an erroneous prediction can potentially be very harmful in practice – e.g., put at risk the health of patients (Sun et al., 2018). Although adversarial attacks have been widely studied in tasks related to image analysis (Paschali et al., 2018; Finlayson et al., 2019; Ma et al., 2019), to the best of our knowledge, a gap still exists regarding BioNLP models and tasks (Araujo et al., 2020).

3 Methodology

We follow a black-box attack methodology (Zhang et al., 2019), which consists of making alterations in the input data to cause erroneous predictions in the models. The following subsections describe each of the adversarial sets, and their construction[1]. We show examples of the stress tests in Table 1.

Noise Adversaries These adversaries test the robustness of models to *spelling errors*. Inspired by (Belinkov and Bisk, 2018), we constructed adversarial examples that try to emulate spelling errors made by human beings. We used SpaCy models (Neumann et al., 2019) to retrieve the medical words of each corpus and add noise to them. We used two types of alterations: i) **Keyboard typo noise (K)** involves replacing a random character in each relevant word with an adjacent character on QWERTY English keyboards. This methodology could be adapted to keyboards with other designs or languages. ii) **Swap noise (W)** consists of selecting a random pair of consecutive characters in each relevant word and then swapping them.

Synonymy Adversaries (S) These adversaries test if a model can *understand synonymy relations*. Unlike the noise adversaries, this set focuses on modifying chemical and disease words (entities). We used PyMedTermino (Jean-Baptiste et al., 2015), which uses the vocabulary of UMLS (Bodenreider, 2004), to find the most similar or related term (synonym) to a certain word. If a synonym is retrieved, the original word is replaced; otherwise, it remains the same. In some cases, this method changes a simple entity (one word) to a composite one (multiple words), so the gold labels are also adjusted to avoid a mismatch in the dataset.

Task and Datasets Biomedical NER is the task that aims at detecting biomedical entities of interest such as proteins, cell types, chemicals, or diseases in biomedical documents. We conducted our evaluation on three biomedical NER benchmarks using the IOB2 tag format (Ramshaw and Marcus, 1999). The **BC5CDR** corpus (Li et al., 2016) is composed of mentions of chemicals and diseases found in 1,500 PubMed articles. The **BC4CHEMD** corpus (Krallinger et al., 2015) contains mentions of chemicals and drugs from 10,000 MEDLINE abstracts. The **NCBI-Disease** corpus (Doğan et al., 2014) consists of 793 PubMed abstracts annotated with disease mentions. Table 2 lists the datasets used in this work along with their most relevant statistics.

Embeddings and NER Models We evaluated both word (W2V) and contextualized embeddings. On the one hand, we assessed BioMedical W2V (Pyysalo et al., 2013) and ChemPatent W2V (Zhai et al., 2019). The ChemPatent embeddings were trained on a 1.1 billion word corpus of chemical patents from 7 patent offices, whereas all the other embeddings were trained on the PubMed corpus. On the other hand, we evaluated BioBERT v1.1 (Lee et al., 2019) and BlueBERT (P) (Peng et al., 2019), both in their base version for convenience.

[1]All stress tests available at https://github.com/ialab-puc/BioNLP-StressTest.

Train / Test	Entity	# of sentences (annotated)	# of tokens	% K	% W	% S
BC5CDR	Chemical	4560 (1609) / 4797 (1706)	122730 /129547	36.3 / 36.1	33.7 / 33.2	6.8 / 6.5
BC5CDR	Disease	4560 (1902) / 4797 (1955)	122730 /129547	36.3 / 36.1	33.7 / 33.2	10.6 / 9.9
BC4CHEMD	Chemical	30681 (16175) / 26363 (13935)	922609 / 792369	37.8 / 37.6	33.9 / 33.9	5.2 / 5.3
NCBI-Disease	Disease	5423 (2501) / 939 (401)	141092 / 25397	37.4 / 37.5	33.4 / 33.3	9.2 / 8.6

Table 2: Details of the datasets used. The last three columns present the percentage of tokens modified for each of the adversarial datasets. The slash separates the values belonging to the training and the test set.

Model	BC5CDR-Chemical				BC5CDR-Disease				BC4CHEMD				NCBI-Disease			
	O	**K**	**W**	**S**	**O**	**K**	**W**	**S**	**O**	**K**	**W**	**S**	**O**	**K**	**W**	**S**
BioBERT	.937 ±.004	.745 ±.006	.635 ±.008	.770 ±.011	.863 ±.004	.407 ±.008	.473 ±.010	.366 ±.007	.919 ±.004	.585 ±.005	.675 ±.007	.678 ±.009	.887 ±.004	.483 ±.007	.628 ±.011	.683 ±.006
BlueBERT	.901 ±.003	.583 ±.005	.708 ±.008	.739 ±.010	.838 ±.004	.368 ±.007	.441 ±.011	.362 ±.007	.820 ±.003	.472 ±.004	.570 ±.009	.607 ±.010	.773 ±.003	.332 ±.006	.438 ±.009	.615 ±.006
BERT	.887 ±.004	.563 ±.007	.684 ±.010	.738 ±.015	.816 ±.006	.356 ±.009	.431 ±.013	.336 ±.008	.808 ±.004	.443 ±.006	.509 ±.008	.598 ±.013	.771 ±.005	.305 ±.008	.433 ±.014	.583 ±.007
BioELMo	.923 ±.001	.838 ±.003	.726 ±.010	.757 ±.032	.845 ±.002	.656 ±.018	.482 ±.025	.408 ±.013	.915 ±.001	.770 ±.003	.634 ±.004	.668 ±.004	.869 ±.005	.711 ±.017	.543 ±.026	.677 ±.012
ChemPatent ELMo	.910 ±.001	.822 ±.004	.745 ±.005	.757 ±.016	.824 ±.001	.637 ±.013	.508 ±.013	.380 ±.017	.898 ±.001	.766 ±.003	.662 ±.005	.642 ±.005	.863 ±.004	.693 ±.018	.586 ±.020	.655 ±.009
ELMo	.879 ±.002	.702 ±.010	.637 ±.017	.720 ±.018	.800 ±.003	.461 ±.023	.373 ±.020	.378 ±.014	.866 ±.001	.612 ±.007	.507 ±.011	.611 ±.005	.848 ±.004	.575 ±.034	.495 ±.023	.643 ±.008
BioMedical W2V	.873 ±.004	.231 ±.012	.238 ±.021	.719 ±.016	.788 ±.008	.132 ±.009	.133 ±.011	.351 ±.015	.846 ±.005	.233 ±.008	.244 ±.013	.589 ±.012	.827 ±.005	.284 ±.014	.292 ±.019	.596 ±.021
ChemPatent W2V	.871 ±.003	.224 ±.011	.221 ±.012	.715 ±.015	.772 ±.007	.127 ±.005	.122 ±.009	.347 ±.016	.828 ±.007	.253 ±.009	.260 ±.010	.584 ±.012	.816 ±.007	.269 ±.021	.252 ±.019	.582 ±.013
W2V	.818 ±.004	.237 ±.013	.227 ±.013	.641 ±.017	.760 ±.003	.120 ±.008	.120 ±.009	.341 ±.013	.766 ±.007	.264 ±.011	.260 ±.012	.513 ±.008	.785 ±.005	.281 ±.022	.271 ±.019	.526 ±.009

Table 3: Stress test evaluation results in terms of terms F1-score for each model and dataset. We report means and standard deviations by training and evaluating ten times with different seeds.

BioBERT embeddings were trained on PubMed abstracts and full-text corpora consisting of 4.3 billion and 13.5 billion words each. BlueBERT was trained on 4 billion words from PubMed abstracts. We used the implementation provided by Peng et al. (2019) for NER with default hyperparameters.[2] Finally, we evaluate BioELMo (Jin et al., 2019) and ChemPatent ELMo (Zhai et al., 2019). As NER models we either (a) fine-tuned BERT as proposed by Peng et al. (2019) or (b) used AllenNLP's basic biLSTM-CRF implementation[3], with no hyperparameter tuning other than changing the initial embedding layer with one of the ELMo or W2V embeddings. For comparison purposes, we also include the "vanilla" version of the models mentioned above, which are pretrained with general corpora. We trained each model 10 times using different random seeds, for 15 epochs every time. We use CoNLL evaluation (Agirre and Soroa, 2007), reporting the F1 score for all datasets.

4 Experiments

In this section we report the results of our experiments. Note that all percentage drops or increases are expressed relative to the original score, not as percentage points.

Adversarial Evaluation Results Table 3 shows the evaluation results on the original (**O**) and adversarial test sets (**K**, **W**, and **S**). In general, the performance of models drops across all adversarial attacks. For BERT-based models, we observe that **K** attacks decrease performance by on average 43.1%, **W** by 34.3% and **S** by 30.8%. BioBERT has the smallest decrease in performance, 34.4%, followed by BlueBERT, with a 37.9% decrease. We hypothesize that BioBERT is more robust than BlueBERT since the former was trained on a larger and more varied corpus. Furthermore, when comparing the performance across all datasets, we see that **BC5CDR-Disease** is the most affected in all stress tests, with a 37.7% performance drop, and the least affected is **BC5CDR-Chemical**, with 16.1%.

The performance reduction of ELMo-based models is similar to those of BERT-based models. An exception is when subject to **W** and **S** noise, where they showed increased robustness with respect to BERT and W2V models (**W**: 55.3% better, **S**: 6.9% better). In almost all the tests, BioELMo performed better than ChemPatent ELMo, except under **W** noise, where ChemPatent ELMo performed con-

[2] https://github.com/ncbi-nlp/bluebert
[3] https://github.com/allenai/allennlp-models

Model	Training	BC5CDR-Chemical		BC5CDR-Disease		BC4CHEMD		NCBI-Disease	
BioBERT	O + K	.934 (O)	.888 (K)	.863 (O)	.755 (K)	.920 (O)	.874 (K)	.886 (O)	.820 (K)
	O + W	.931 (O)	.899 (W)	.865 (O)	.781 (W)	.922 (O)	.892 (W)	.872 (O)	.848 (W)
	O + S	.933 (O)	.910 (S)	.840 (O)	.819 (S)	.919 (O)	.923 (S)	.874 (O)	.875 (S)
BlueBERT	O + K	.898 (O)	.820 (K)	.844 (O)	.717 (K)	.819 (O)	.750 (K)	.789 (O)	.668 (K)
	O + W	.896 (O)	.656 (W)	.841 (O)	.759 (W)	.818 (O)	.785 (W)	.784 (O)	.729 (W)
	O + S	.900 (O)	.890 (S)	.818 (O)	.814 (S)	.820 (O)	.788 (S)	.773 (O)	.804 (S)
BioELMo	O + K	.923 (O)	.870 (K)	.833 (O)	.732 (K)	.912 (O)	.837 (K)	.864 (O)	.820 (K)
	O + W	.922 (O)	.825 (W)	.838 (O)	.654 (W)	.913 (O)	.820 (W)	.875 (O)	.777 (W)
	O + S	.919 (O)	.901 (S)	.826 (O)	.799 (S)	.912 (O)	.901 (S)	.871 (O)	.848 (S)
ChemPatent ELMo	O + K	.910 (O)	.859 (K)	.823 (O)	.713 (K)	.898 (O)	.828 (K)	.860 (O)	.793 (K)
	O + W	.907 (O)	.835 (W)	.813 (O)	.682 (W)	.899 (O)	.824 (W)	.863 (O)	.804 (W)
	O + S	.904 (O)	.895 (S)	.813 (O)	.757 (S)	.895 (O)	.874 (S)	.848 (O)	.819 (S)
BioMedical W2V	O + K	.888 (O)	.467 (K)	.773 (O)	.303 (K)	.832 (O)	.486 (K)	.820 (O)	.543 (K)
	O + W	.873 (O)	.598 (W)	.796 (O)	.482 (W)	.836 (O)	.609 (W)	.819 (O)	.639 (W)
	O + S	.867 (O)	.883 (S)	.781 (O)	.787 (S)	.837 (O)	.852 (S)	.836 (O)	.804 (S)
ChemPatent W2V	O + K	.867 (O)	.454 (K)	.768 (O)	.307 (K)	.817 (O)	.482 (K)	.822 (O)	.548 (K)
	O + W	.785 (O)	.619 (W)	.765 (O)	.477 (W)	.819 (O)	.626 (W)	.792 (O)	.663 (W)
	O + S	.868 (O)	.864 (S)	.738 (O)	.779 (S)	.818 (O)	.835 (S)	.797 (O)	.801 (S)

Table 4: Adversarial training results in terms of F1-score for each model and dataset. The training column shows the **O** set merged with **K**, **W**, or **S**. The test set is shown in parentheses for each scenario.

sistently better, by 5.1% on average. We hypothesize that these results are due to ELMo using a character-based input representation, which would allow handling of swap characters inside the words.

W2V-based models were the most brittle but showed similar patterns to the previous models. Adversaries examples produced performance drops ranging from 53.8% on **NCBI-Disease** to 74.1% on **BC5CDR-Disease**. In the case of **S** adversaries, W2V-based showed performance drops ranging from 17.8% on **BC5CDR-Chemical** to 55.3% on **BC5CDR-Disease**.

Regarding the "vanilla" models, we see that they are all the worst in the original dataset (**O**) compared to their biomedical counterparts. In the same way, they are more fragile to adversary attacks in the biomedical scenario. In average, BERT has a decrease in performance of 39.6%, ELMo of 34.4% and W2V of 59.6% across all datasets.

Even though the **BC5CDR** dataset covers both chemicals and diseases, the disease task is more affected by **S** adversaries. We believe this is due to the higher number of words affected by the attacks compared to the other benchmarks (Table 2). Another possible cause is the kind of synonyms used to replace the entities, which tend to be both superficially dissimilar and more extensive than their originals, e.g., *arrhythmia* is replaced by *heart conduction disorder*. By contrast, chemical synonyms often include terms derived from the original, e.g., *morphine* is changed to *morphine sulfate*.

Training on Adversarial Examples Additionally, we subjected the training sets to adversarial attacks, and evaluated the models both against the original test sets and their noisy counterparts. When training with **K** noise, we observed performance decreases by 21.2%, followed by **W**, 15.8%, and **S** with a slight decline of 0.8%, compared to 44.4%, 46.3% and 31.3% respectively in the Adversarial Evaluation setting. Besides, and interestingly, training with **S** improves performance in some cases, by up to 5.5% compared to the original **S** test set. We hypothesize that this is because the introduced adversarial samples work as a data augmentation mechanism. In terms of datasets, we see that **BC5CDR-Disease** is the most affected by adversaries, with an average 17.5% drop, and the least affected is **NCBI-Disease**, with an average 9.7% drop compared to the non-adversarial test set. When comparing the three architectures we see that BERT is affected by 6.3%, ELMo by 7.6% and W2V by 24.0% on average compared to the original test set. This result stands in line with findings on other NLP tasks, where BERT comes up first, followed by ELMo and W2V (Peng et al., 2019). This is because BERT uses recent methods and techniques like Transformer (Vaswani et al., 2017) and WordPiece tokenizer (Schuster and Nakajima, 2012) that allow it to learn better representations.

BioBERT Error Analysis This section seeks to understand how the most robust model – BioBERT – behaves under adversarial evaluation. To this end, we analyzed NER model confusions with respect to the original datasets, synonym (**S**), swap (**W**), and keyboard (**K**) perturbations on the BC5CDR chemical and disease dataset(s).

In the original dataset (Figure 1(a)), we see that

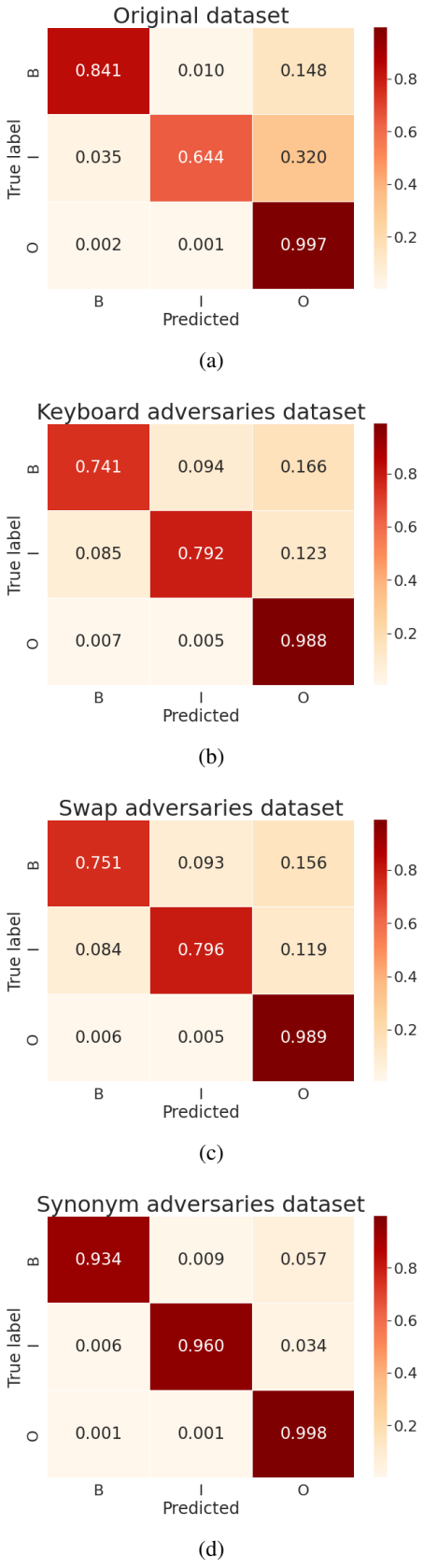

Figure 1: Normalized confusion matrices for test results with (a) original (**O**), (b) keyboard (**K**), (c) swap (**S**) and (d) synonym (**S**) BC5CDR-Disease and Chemical datasets on average.

most of the errors come from confusing I and O labels (32% of the cases). Under adversarial attacks, this type of error spreads to other IOB labels. For keyboard (**K**) errors (Figure 1(b)), the most frequent mistake is to confuse B with O, with 16.6% of these cases. The same goes for swap (**W**) perturbations (Figure 1(c)), where this error is repeated 15% of the time. When using synonyms (**S**) (Figure 1(d)), error rates become by contrast globally low compared to **K** and **W**. We believe that this happens because entities are converted into similar ones. For instance, "stomach neoplasm" gets transformed into "stomach tumor".

Lastly, regardless of the adversaries, there are confusions with numbers and special character sequences that the model classifies as I (i.e., lie inside an entity span) but whose ground truth label is O (i.e., lie outside an entity span).

5 Conclusions

In this work, we have investigated whether large scale biomedical word (W2V) and contextualized word embeddings (BERT and ELMo) are robust with respect to black-box adversarial attacks in the biomedical NER task. Our experimental results show different sensitivities of the models to misspellings and synonyms. Among the main findings, we show that BERT-based models are generally better prepared for adversarial attacks, but they are still fragile, leaving room for future improvement in the field. ELMo-based models show lower robustness in most cases but consistently outperformed BERT in some specific scenarios. W2V proves to be more brittle but shows similar patterns in terms of relative performance drops. We also demonstrate that by training with adversaries, we can considerably decrease the drop in performance and even improve the models' original performance when trained with synonyms, as they act as a form of regularization and augmentation of data.

Acknowledgements

We are grateful to the anonymous reviewers for their valuable feedback on earlier versions of this paper. This work was partially funded by ANID - Millennium Science Initiative Program - Code ICN17_002 and by ANID, FONDECYT grant 1191791, as well as supported by the TPU Research Cloud (TRC) program.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the fourth international workshop on semantic evaluations (semeval-2007)*, pages 7–12.

Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430.

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*, Seattle, USA. 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. 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.

Andres Carvallo, Denis Parra, Hans Lobel, and Alvaro Soto. 2020. Automatic document screening of medical literature using word and text embeddings in an active learning setting. *Scientometrics*, 125(3):3047–3084.

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.

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.

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

Douglas Hanahan and Robert A. Weinberg. 2000. The hallmarks of cancer. *Cell 100(1)*, pages 57–70.

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. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, Roger A Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, SV Ramanan, Senthil Nathan, Slavko Žitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A Akhondi, Jan A Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C Lee Giles, Hongfang Liu, Komandur Elayavilli Ravikumar, Andre Lamurias, Francisco M Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usié, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzabal, and Alfonso Valencia. 2015. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(S1).

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*, 36(4):1234–1240.

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:baw068.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *SciSpacy:Fast and Robust Models for Biomedical Natural Language Processing*.

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. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X Binder, and Lars Juhl Jensen. 2015. Diseases: Text mining and data integration of disease–gene associations. *Methods*, 74:83–89.

S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, pages 39–44.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

M. Schuster and K. Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

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 and Data Mining*, KDD '18, page 793–801, New York, NY, USA. Association for Computing Machinery.

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*.

Luis Tari, Saadat Anwar, Shanshan Liang, James Cai, and Chitta Baral. 2010. Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics*, 26(18):i547–i553.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zenan Zhai, Dat Quoc Nguyen, Saber Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory, and Karin Verspoor. 2019. Improving chemical named entity recognition in patents with contextualized word embeddings. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 328–338, Florence, Italy. Association for Computational Linguistics.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2019. Adversarial attacks on deep learning models in natural language processing: A survey.