Evaluating the Robustness of Biomedical Concept Normalization

Sinchani Chakraborty
IIT Kharagpur
sinchanichakraborty@gmail.com

Harsh Raj
Delhi Technological University
harsh777111raj@gmail.com

Srishti Gureja
University of Delhi
srishtigureja1110@gmail.com

Tanmay Jain
Delhi Technological University
tanmayj2020@gmail.com

Atif Hassan
IIT Kharagpur
atif.hit.hassan@gmail.com

Sayantan Basu
IIT Guwahati
sayantan18@iitg.ac.in

Abstract

Biomedical concept normalization involves linking entity mentions in text to standard concepts in knowledge bases. It aids in resolving challenges to standardising ambiguous, variable terms in text or handling missing links. Therefore, it is one of the essential tasks of text mining that helps in effective information access and finds its utility in biomedical decision-making. Pre-trained language models (e.g., BERT) achieve impressive performance on this task. It has been observed that such models are insensitive to word order permutations and vulnerable to adversarial attacks on tasks like Text Classification, Natural Language Inference. However, the effect of such attacks is unknown for the task of Normalization, especially in the biomedical domain. In this paper, we propose heuristic-based Input Transformations (word level modifications and word order variations) and Adversarial attacks to study the robustness of BERT-based normalization models across various datasets consisting of different biomedical entity types. We conduct experiments across three datasets: NCBI disease, BC5CDR Disease, and BC5CDR Chemical. We observe that for Input Transformations, pre-trained models often fail to detect invalid input. On the other hand, our proposed Adversarial attacks that add imperceptible perturbations result in affecting the ranking of a concept list for a given mention (or vice versa). We also generate natural adversarial examples that lead to performance degradation of $\sim$30% in the F1-score. Additionally, we explore existing mitigation strategies to help a model recognize invalid inputs.

1 Introduction

Biomedical concept normalization aims to map an entity mention occurring in free-form text to a unique concept in a knowledge base or an ontology [Xu and Bethard, 2021;Ji et al., 2020;D'Souza and Ng, 2015]. In the biomedical domain, concepts exhibit different surface forms, that include various morphological and orthographic variations [Zhou et al., 2004]. A concept can be linked to different mentions, e.g., the following entity mentions {*breast cancer, breast tumor, breast carcinoma*} are linked to the same concept *breast neoplasm* with a unique ID "D001943" in the MeSH ontology [Zieman and Bleich, 1997]. This makes it crucial to link mentions with their standard canonical forms present in an ontology or a knowledge base, rich with semantic structure.

Workshop on Transfer Learning for Natural Language Processing, 2022, New Orleans.

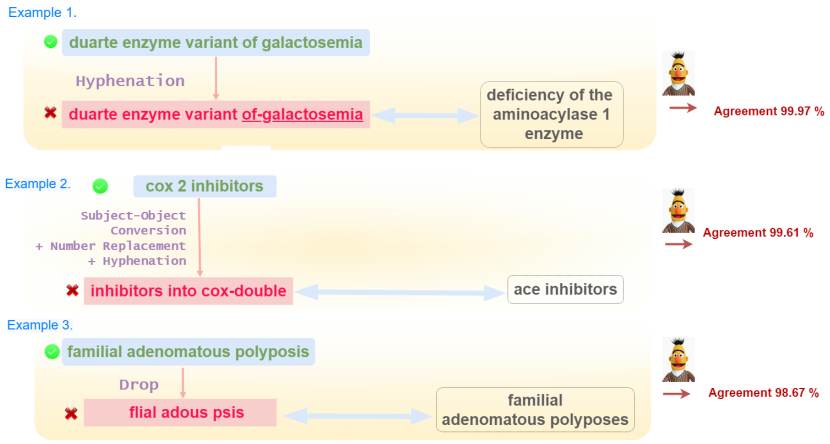

Figure 1: A Fine-tuned Clinical BERT model assigns high probability to invalid mention names that does not have correspondence to the candidate in an Ontology. Green : Original Concept/Mention; Red : Transformed Concept/Mention; Grey: Mention in a text/Concept in an Ontology.

Normalization is one of the fundamental tasks of data mining, that is utilized in automated medical decision making which is directly responsible for the well-being of patients [Matheson, 2019]. The robustness of such models is therefore crucial to maintain. Biomedical normalization models [Xu and Bethard, 2021;Ji et al., 2020;Li et al., 2019] achieve impressive results using pre-trained contextualized models like BERT [Devlin et al., 2018]. However, recent studies reveal BERT being vulnerable to adversarial attacks [Szegedy et al., 2013; Goodfellow et al., 2014] and insensitive to certain textual transformations that permute the word order [Gupta et al., 2021; Hessel and Schofield, 2021]. Figure 1 shows a BERT-based normalization model that retains the same predictions, for three different examples consisting of invalid transformed inputs as for a valid term, with extremely high agreement scores. This can lead to the linking of wrong invalid inputs to a concept in an ontology. In the biomedical domain, [Araujo et al., 2020] proposes adversarial attacks that introduces natural spelling errors and typos made by humans to analyze the robustness for Named Entity Recognition and Semantic Textual Similarity tasks. [Mondal, 2021] utilizes BERT to generate adversaries that constitute domain specific synonym replacement, spelling variations and number replacement for the task of biomedical Text Classification. To the best of our knowledge we are the first to study the effect of input transformations and adversarial attacks for the task of biomedical concept normalization.

In this paper, we study the responses of BERT-based normalization models to Input Transformations (word-level modifications and word-order variations). The word-level modifications are inspired from hand-crafted rules that were used to compose a multi-pass sieve in order to find a match between a candidate and mention [D'Souza and Ng, 2015]. While these kinds of transformations lead to changes detectable by a human, modifications that are adversarial in nature lead to imperceptible changes [Michel et al., 2019]. Adversarial attacks confuse the models and lead to performance degradation. In our work, we also focus on generating natural-looking adversarial examples to study the response of normalization models. Moreover, it has been observed for the task of image classification, adversarial ranking attacks can alter the ranks of a candidate by adding perturbation to a candidate list [Zhou et al., 2020]. We adopt this approach to study the effect of candidate ranking, on adding imperceptible perturbations, of existing normalization approaches that consist of a two-step framework: a candidate generator (produces potential candidates from an ontology) and a supervised candidate ranker (ranks potential candidates) [Xu and Bethard, 2021;Ji et al., 2020]. Finally, we also explore mitigation strategies with the aim of helping the model in identifying invalid inputs that result in the application Input transformations.

To demonstrate the efficacy of the proposed input transformations and adversarial attacks, we select ranking-based normalization models [Ji et al., 2020]. Based on this model we perform experiments on 4 types of BERT models : BERT [Devlin et al., 2018], BioBERT [Lee et al., 2019], ClinicalBERT [Alsentzer et al., 2019], PubMed BERT [Gu et al., 2021]. Additionally we consider a separate model that performs a triplet-search based ranking for normalization and trains the candidate generator space unlike the previous model [Xu and Bethard, 2021]. The evaluation of the task was performed on three

datasets: NCBI [Doğan et al., 2014], BC5CDR-Disease [Li et al., 2016], and BC5CDR-Chemical [Li et al., 2016].

Our experiments and observations reveal that most of the individual input transformations that involve word-level modifications and word-order variations, go undetected by a model. However, when transformations are chained or certain important words are affected we find the model starts detecting invalid input. For adversarial attacks, our experiments reveal that the best-performing models are brittle on imperceptible perturbations and natural-looking adversaries. Summarily, we make the following contributions in this paper[1].

1. We systematically study the effect of input transformations to understand the sensitivity of the model to word-level modifications and word order variations for biomedical concept normalization. We apply thirteen individual transformations for effective analysis.
2. We propose Adversarial Attacks that lead to a significant drop in model performance (86.0 to 58.05 in F1-score) on NCBI [Doğan et al., 2014] dataset, revealing the brittle nature of top-performing Normalization models.
3. Finally, we explore existing mitigation strategies to make models sensitive to invalid input transformations for the task of Normalization.

2 Task Formulation

Given a corpus H and an Ontology O, the aim of concept normalization is to find a mapping function f, that maps a mention m in a free-form text corpus H to a concept c which is a unique identifier in Ontology O, i.e., $c = f(m)$. A concept c is part of a set C consisting of all concepts in an Ontology.

In the following section 3 and 4 we describe each *Input transformations* and *Adversarial Attacks* in details, that we apply to analyze the robustness of biomedical concept normalization.

3 Input Transformations

For the given task consisting of an input $x \epsilon X$ and a function f that maps inputs to $y \epsilon Y$ which corresponds to a label space. An input transformation is a function $\sigma : X \rightarrow X$ that acts on input x to produce $x^{'}$ such that f($x^{'}$) is not defined. For each of the transformations defined below there is a high probability that one can map to the original candidate. In that case we remove it from the list. While this is a single instance of such a case to appear, the remaining combinations can serve as invalid mention entities that are formed through the following transformations.

Hyphenation A mention term undergoes hyphenation if it does not consist of hyphens whereas it undergoes dehyphenation if it consists of hyphens in the original term by considering consecutive pairs at a time. E.g., a mention "hereditary breast and ovarian cancer" will form the following variations: {"hereditary-breast and ovarian cancer", "hereditary breast-and ovarian cancer", "hereditary breast and-ovarian cancer", "hereditary breast and ovarian-cancer"}. This is performed only with multiword mentions.

Number Replacement We perform this transformation only on those mentions that consist of a pre-existing numeral while the other mentions devoid of any numeral retain their original form. The numbers are replaced as their equivalent form of the roman, cardinal, or multiplicative numeral. E.g., "c9 deficiency" has the following forms {"cnine deficiency", "cix deficiency"}.

Disorder Synonym/Mention Term Concatenation Given an entity mention *m* we generate samples of the form $(m + q)$ where q is a disorder synonym or a frequently appearing modifier for chemicals (depends on the domain usage). We collected a list of roughly fifteen disorder synonyms from [D'Souza and Ng, 2015] and concatenated this at the end of every mentions where these terms are absent. So for each mention there are fifteen different variants. Similarly, for chemicals we manually enlist fourteen terms and concatenate those at the end of every mention.

[1]Our code is available at https://github.com/deepwizai/robust-normalization

Table 1: Agreement scores on Transformations for three datasets on finetuned-BioBERT model

Transformations	NCBI disease	BC5CDR disease	BC5CDR chemical
Unperturbed	0.93	0.92	0.94
Hyphenation (H)	0.89	0.88	0.92
Subject Object Conversion (SOC)	0.90	0.89	0.93
Number Replacement (Num-R)	0.88	0.89	0.91
Disorder Synonym/Mention Term Concatenation (C)	0.90	0.89	0.91
Stemming (S)	0.87	0.87	0.92
H + SOC	0.89	0.90	0.94
H + SOC + Num-R	0.87	0.89	0.93
H + SOC + Num-R + C + S	0.86	0.87	0.92
Copy-Sort	0.91	0.90	0.14
Sort	0.90	0.89	0.92
Reverse	0.91	0.89	0.92
Shuffle	0.91	0.89	0.92
Drop	0.85	0.85	0.88
Repeat	0.84	0.84	0.88
Replace	0.83	0.82	0.89
Copyone	0.76	0.77	0.87

Stemming Using Porter Stemmer [Porter, 1980] from the NLTK library [Bird, 2006], the words of the mentions are stemmed and are then replaced with the corresponding occurrence. E.g., the mention of the term "chromosomal fragmentation during meiosis" changes to "chromosom fragment during meiosi". The stems are hard to interpret since they do not form actual words, thus altering the meaning of the mention terms.

Subject Object Conversion We generate examples for Subject-Object conversion using the steps detailed in [D'Souza and Ng, 2015]. There are four ways to generate samples for Subject-Object conversion. Given a mention m that contains a preposition: 1) Replacing with other prepositions; 2) Deleting the preposition and swapping surrounding token; 3) Shifting last token to the front and then inserting a preposition while shifting the other remaining tokens; 4.) Shifting first token to the end and adding a preposition as the second last token followed by shifting remaining tokens to the rest.

In the above transformations, the first four are word-level transformations that make changes in the word level or append additional terms to the neighboring words. The transformation- Subject Object Conversion- leads to change in word order. Following [Gupta et al., 2021] we perform two classes of additional transformations which are: 1.) Lexical-overlap based transformations retains the bag of word collection but it changes the word order in four different ways which are as follows: *Sort* - sorting input tokens; *Reverse* - reversing the token sequence; *Shuffle* - shuffling tokens randomly; *Copy Sort* - transforming a candidate c as a copy of the mention m with the words sorted alphabetically. The transformations Reverse and Copy Sort are part of Lexical overlap based transformations which exhibit random word order. 2.) Gradient level transformations consider change in the loss for i'th input token in a given mention m to measure the token importance. Based on this calculation four operations are defined: *Drop* - this drops the least important token in a mention; *Repeat* - least important token is repeated ; *Replace* - Least important token is replaced by random tokens; *Copy One* - The most important token is copied from the mention m and put as the only token in the candidate c . These transformations target to destroy the semantic meaning entirely conveyed in the phrase. In Figure1, we see Example 3 as the transformation Drop is being applied, where a pre-trained model is unable to catch the transformation and treating it as correct input.

Each of these transformations are performed individually as well as we perform chained evaluations. The second example in Figure 1 is based on chaining three transformations: *Subject-Object Conversion* followed by *Number Replacement* and finally *Hyphenation*. Since the task of normalization involves pair of texts as input constituting the mention level text and candidate concepts, performing transformations on any one of these should suffice in order to carry out robustness analysis.

4 Adversarial Attacks

We perform two types of Ranking Attacks for the task of normalization. Both of these attacks are adversarial in nature thus involving imperceptible changes to form an adversary.

Adversarial Ranking Attack (Adv-Rank) For a set of chosen candidate $X = \{c_1, c_2, ..., c_n\}$ with respect to a specific mention from the set M = $\{m_1, m_2, ..., m_n\}$ we perform two types of Adversarial Ranking Attacks: 1) Mention Attack (MA): Attack targeted to mention m and 2) Candidate Attack (CA) : Attack targeted to candidate c. We define MA+ and MA- as variants of Mention Attacks to raise or lower the rank of a candidate set C by perturbing a single mention m. Similarly, we define CA+ and CA- as variants of Candidate Attacks to raise or lower the rank of a single candidate c with respect to the mention set M. The ranks are altered by adding universal perturbation r. The final ranking order for a Deep Neural Network is defined by the sample positions in a common embedding space so adding an adversarial perturbation to it can lead to potential alteration in ranking. This is performed using a surrogate loss in the form of Triplet loss. For CA+ it is defined as:

$$L_{CA^+}(c, M; X) = \sum_{q \epsilon M} \sum_{x \epsilon X} [d(q, c) - d(q, x)]_+ \tag{1}$$

where X denotes set of all candidates; M denotes set of all mentions; c is the candidate whose rank is raised w.r.t mention q ϵ M. In the same way the Triplet loss is defined for CA-, MA+ and MA-. We refer the reader to consult the original work for a detailed overview on Adversarial Ranking [Zhou et al., 2020]. In our work we can utilize any one of the shifts (CA+ or CA-; MA+ or MA-) to make imperceptible changes to the ranking order.

Table 2: Adversarial Ranking Attack on BERT based ranking model. The "+" in MA indicates that the rank of the chosen candidate is raised. The changes in average rank using Cosine Distance Triplet Loss is reported with (%) omitted. Emb Shift: Embedding Shift.

Dataset	Emb Shift	Attack	Clinical BERT	Bio BERT
NCBI	1.3	MA+	$\sim$50 $\rightarrow$ 51.8	$\sim$50 $\rightarrow$ **44.2**
BC5CDR Chemical	1.3	MA+	$\sim$50 $\rightarrow$ **42.6**	$\sim$50 $\rightarrow$ 50.9

Least Similar Entity Concatenation (LSEC) We present Least Similar Entity Concatenation (LSEC) that modifies the candidate set by concatenating the most dissimilar entity belonging to a parent class same as the original candidate, corresponding to an ontology or a knowledge base. In the following steps we lay down the approach taken by LSEC attack. Given a mention m and a candidate c :

Step 1: Find the concept identifier of c that links it to an ontology.

Step 2: Access the immediate root which corresponds to the parent concept and find the set of existing siblings.

Step 3: Find pair-wise cosine similarity between the concept and the existing set of candidates.

Step 4: Select the most dissimilar entity and append it with the candidate c to form c'.

For example, given a mention-candidate pair - ("meningitis", "encephalomeningitis") - the identifier for the candidate c, "encephalomeningitis" is MESH:D008590 corresponding to MEDIC Ontology [Davis et al., 2012]. The immediate parent is tracked and the most dissimilar sibling based on cosine similarity, corresponding to the candidate under the same parent ID is determined, in this case, it is "Pseudorabies". Finally, the modified candidate c' is formed by appending both the terms as a composite mention: "encephalomeningitis and pseudorabies". The attacks consisting of entities belonging to the same class corresponding to an Ontology or a Knowledge Base are considered to be adversarial in nature [Lin et al., 2021]. Since both c and c' belong to the same parent class, hence we tag instances generated from LSEC attack as natural adversarial examples. Furthermore, the

adversarial examples obtained are natural since it consists of real and valid entities, devoid of any grammatical disfluency.

5 Experimental Setup

Datasets: We evaluate transformation on three different biomedical normalization datasets : 1) NCBI [Doğan et al., 2014] 2) BC5CDR-Disease [Li et al., 2016] and 3) BC5CDR-Chemical [Li et al., 2016].

Models: For evaluating transformations we perform experiments on BERT-based Ranking [Ji et al., 2020] . We use the following BERT models for reporting the results on BERT-based Ranking : BioBERT [Lee et al., 2019], Clinical-BERT [Alsentzer et al., 2019], PubMed BERT [Gu et al., 2021] and BERT [Devlin et al., 2018]. We perform fine-tuning on the training data using Adam Optimizer [Kingma and Ba, 2014]. For evaluating Ranking Attacks we only use BERT-base-uncased models. We set a threshold α of 0.75 for cosine similarity for LSEC attack. We provide more details in the Appendix.

6 Metrics

6.1 Input Transformations

We use Agreement and Confidence scores following [Gupta et al., 2021] for all the Transformations defined over the input.

Agreement: It is defined as the percentage of examples that retains the same prediction after applying transformations.

Confidence: The average probability scores of the predicted level gives the Confidence score.This depends on the number of classes N (in our case N=2).

A low confidence score and an agreement score close to random suggests that a model is reliable that is successfully able to detect invalid input.

6.2 Adversarial attacks

We perform two types of Adversarial attacks: 1) Adversarial Ranking (Adv-Rank) 2) Least Similar Entity Concatenation (LSEC). Below are the metrics that are used to evaluate these attacks.

Adversarial Ranking Attack (Adv-Rank) The effectiveness of the attack is calculated by the magnitude of change in normalized rank. Given a candidate c the normalized rank is given as follows:

$$R(m, c) = \frac{Rank_X(m, c)}{|X|} \times 100\% \tag{2}$$

where c ϵ X; $|X|$ is the length of full ranking list and $R(q, c)$ ϵ [0,1]

Least Similar Entity Concatenation (LSEC) We report the F1-score on the original test set and the F1-score after the attack . In addition, we also report the average confidence with which the predictions are performed.

7 Results and Analysis

This section discuss the results for Input Transformations and Adversarial Attacks. We also investigate through various analyses: 1) How do the model performance change for a varying percentage of invalid inputs? 2) Does invalid examples go undetected for a different approach? 3) Using existing strategies to make a model performing the normalization task, sensitive to invalid input transformations.

7.1 Main results

Input Transformations We apply various morphological, lexical and gradient-based transformations to the input belonging to three different datasets. Table 1 shows that the model exhibits high agreement scores. This is indicative of the fact that the model maintains its original predictions. *CopySort* and *CopyOne* are the transformations for which the model can identify invalid output, since those consists of a repetition of a single selected word lacking of other terms in a mention/candidate. The agreement scores are comparatively lower for *Gradient-based perturbations* where the important tokens are altered. A similar drop in scores is observed for *Chained Transformations*. On the contrary, for each individual transformation (except for stemming which leads to a meaning change in the input) the agreement scores are much higher. The model might pick up spurious correlations related to the common representation of biomedical entity names or phrases during training. Also, high agreement scores on lexical transformations confirm that the model is insensitive to word order.

Adversarial Ranking Attack (Adv-Rank) As shown in Table 2 for a designated embedding shift there is a considerable shift in the average rank for Mention Attack (MA+). The default value without the attack is 50%. The target of this attack is to make random changes in the ranked candidates after the model generates it. This is a black-box attack and significant shifts lead to 4% to 5% dip in model performance for BERT based ranking attacks.

Least Similar Entity Concatenation (LSEC) Table 3 shows the performance of the attack performed for Least Similar Entity Concatenation (LSEC). Compared to Adv-Rank this is a better attack that confuses the model and its performance is degraded.

Table 3: F1 scores before and after LSEC attack for two variants of BERT based ranking models on NCBI. Confidence : The average confidence with which the predictions are made after an attack is reported

Models	Before Attack	After Attack	Confidence
PubMed BERT	0.853	0.580	0.986
Bio BERT	0.859	0.582	0.990

7.2 Varying transformation of input (%)

Table 4 shows the results when 20%, 60% and 80% of the test data undergo transformations. We report the scores for transformations that include individual *Input transformations*, chained *Input transformations* and destructive transformations including *Lexical-overlap* and *Gradient-based perturbations*. The highlighted values of the accuracy scores denote instances where the model learns to detect the invalid inputs successfully, due to which there is a dip in the score. We observe a considerable change in the reported accuracy for every *Chained Transformations* and *Gradient-based transformations* this suggests that when there is a considerable transformation, like in the case of *Chained Transformations* (where the modifications leading to invalid changes in the input become more detectable), and for Gradient level transformations (where the most important token is affected), the model does a better job in classifying the invalid inputs.

Table 4: Accuracy Scores (%) for the BioBERT model undergoing transformations with different data coverage. **H:** Hyphenation; **SOC:** Subject Object Conversion; **NR:** Number Replacement; **C:** Disorder Synonym /Mention Term Concatenation; **S:** Stemming

Input (%)	Transformations											
	H	SOC	NR	C	S	H+ SOC	H + SOC+ NR	H + SOC+ NR+ C+S	Copy Sort	Shuffle	Drop	Copy One
20%	0.87	0.88	0.88	0.87	0.88	0.88	0.87	0.87	0.87	0.79	0.88	0.87
60%	0.88	0.88	0.87	0.88	0.87	0.88	0.85	0.84	0.87	0.81	0.87	0.80
80%	0.89	0.89	**0.87**	0.89	0.88	**0.84**	**0.82**	**0.81**	0.87	**0.77**	**0.86**	**0.77**

7.3 Are the transformations undetectable across different approaches?

To better understand the effect of the transformations, we conduct the experiments on a different model that performs the task of normalization which is Triplet-search ConNorm [Xu and Bethard, 2021]. Table 5 shows the results on the Input Transformations. Compared to the BERT-based ranking model in Table1 we find that this model shows a considerable accuracy drop for NCBI dataset on individual transformations apart from the lexical based transformation. For BC5CDR-Disease and BC5CDR Chemical the model performance is almost similar to BERT based ranking model. For detecting Gradient level perturbation, Triplet Search based Concept Normalization model does a better job.

Table 5: Accuracy Scores (%) of Triplet Search based Concept Normalization model across three datasets NCBI, BC5CDR disease, BC5CDR chemical. The "None" value is on untransformed training set. **H:** Hyphenation; **SOC:** Subject Object Conversion; **NR:** Number Replacement; **C:** Disorder Synonym /Mention Term Concatenation; **S:** Stemming; **Rev:** Reverse

	Transformations											
Dataset	None	H	SOC	NR	C	S	Sort	Rev	Shuffle	Drop	Repeat	Replace
NCBI	0.95	0.64	0.63	0.87	0.66	0.55	0.94	0.87	0.87	0.87	0.79	0.88
BC5CDR disease	0.94	0.92	0.94	0.94	0.94	0.79	0.93	0.93	0.92	0.48	0.46	0.46
BC5CDR chemical	0.99	0.97	0.97	0.93	0.75	0.95	0.97	0.98	0.98	0.24	0.25	0.21

7.4 Analysis of mitigation strategies for Input Transformations

We explore two strategies to enhance the ability of the model to recognize invalid input transformations.

Augmented Training : We follow a straightforward approach wherein we randomly select 50% of training data which then undergoes all possible transformations mentioned in Section 4. This set of invalid examples is augmented with the original training data, and the final version of the training set is obtained.

Entropic Regularization : This approach helps in training the model to be less certain on invalid inputs [Feng et al., 2018]. Given an original dataset D and the dataset constituting invalid examples D'. This method adds a loss term for the invalid examples weighted by the additions of a new hyperparameter.

Table 6 shows a comparative analysis between Augmented Training (Aug-Data) and Entropic Regularization (Ent Reg). The experiment is conducted for the NCBI dataset for the BERT-based Ranking model consisting of PubMed BERT and Bio-BERT. We observe that the Data Augmentation (Aug-Data) reports a marginally better performance when compared to Entropic Regularization.

Table 6: Comparative Analysis of Mitigation Strategies reported on NCBI for BERT-based ranking model. **Aug-Data**: Agreement on Augmented Training; **Ent Reg**: Agreement after Entropic Regularization

Model	Aug-Data	Ent Reg
PubMed-BERT	**0.828**	0.826
Bio-BERT	**0.769**	0.759

8 Conclusions and Future Work

In this paper, we study the effect of Input Transformations and Adversarial Attacks on the task of biomedical concept normalization. The Input transformations include word level modifications

and word order variations. A part of our proposed input transformations is motivated by surface form variations in the biomedical domain. Our work shows surprising results on invalid Input transformation, wherein BERT-based normalization models find it hard to identify such samples. The models are seen to be invariant to random word order permutations. In addition, we propose two types of Adversarial attacks, one of which form natural looking adversaries while the other affect the ranking of candidate/mention sets on adding imperceptible perturbations. These attacks lead to model performance degradation. We conduct this empirical study considering different BERT-based normalization models that operates on a ranking-based approach. We explore mitigation strategies and techniques to increase model sensitivity to input transformations. For future work, we plan to incorporate better mitigation strategies to increase sensitivity to invalid samples and defence mechanisms to reduce the brittleness of the model to adversarial attacks.

Acknowledgments

We thank the anonymous reviewers for their helpful insights and feedback. We would also like to thank past members of NLP group at IIT Kharagpur for providing us with valuable suggestions.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Vladimir Araujo, Andres Carvallo, Carlos Aspillaga, and Denis Parra. On adversarial examples for biomedical nlp tasks. *arXiv preprint arXiv:2004.11157*, 2020.

Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. Medic: a practical disease vocabulary used at the comparative toxicogenomics database. *Database*, 2012, 2012.

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

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.

Jennifer D'Souza and Vincent Ng. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 297–302, 2015.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. *arXiv preprint arXiv:1804.07781*, 2018.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. Bert & family eat word salad: Experiments with text understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12946–12954, 2021.

Jack Hessel and Alexandra Schofield. How effective is bert without word ordering? implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, 2021.

Zongcheng Ji, Qiang Wei, and Hua Xu. Bert-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings*, 2020:269, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: A pre-trained biomedical language representation model for biomedical text mining. bioinformatics, btz682, 2019.

Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, Hong Yu, et al. Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study. *JMIR medical informatics*, 7(3):e14830, 2019.

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. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. Rockner: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. *arXiv preprint arXiv:2109.05620*, 2021.

Rob Matheson. Automating artificial intelligence for medical decision-making, 2019.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. On evaluation of adversarial perturbations for sequence-to-sequence models. *arXiv preprint arXiv:1903.06620*, 2019.

Ishani Mondal. Bbaeg: Towards bert-based biomedical adversarial example generation for text classification. *arXiv preprint arXiv:2104.01782*, 2021.

Martin F Porter. An algorithm for suffix stripping. *Program*, 1980.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

Dongfang Xu and Steven Bethard. Triplet-trained vector space and sieve-based search improve biomedical concept normalization. Association for Computational Linguistics (ACL), 2021.

Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and ChewLim Tan. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190, 2004.

Mo Zhou, Zhenxing Niu, Le Wang, Qilin Zhang, and Gang Hua. Adversarial ranking attack and defense. In *European Conference on Computer Vision*, pages 781–799. Springer, 2020.

Yuri L Zieman and Howard L Bleich. Conceptual mapping of user's queries to medical subject headings. In *Proceedings of the AMIA Annual Fall Symposium*, page 519. American Medical Informatics Association, 1997.

A Appendix

A.1 Reproducibility

Model Used. We performed the attacks on two types of entity normalization methods, a ranking based normalization [Ji et al., 2020] and Triplet Search ConNorm [Xu and Bethard, 2021] which is trained via a triplet objective. For both methodologies we took a set of pretrained BERT models, namely - "bio-clinical-bert", "pubmed-bert" and "bio-bert-cased" to base our experiments. These are BERT like models but are trained on biomedical specific dataset. For such pretrained models, we used Huggingface's transformers repository [Wolf et al., 2020].

Dataset. We used 3 benchmark datasets in the biomedical domain for biomedical normalization, namely, NCBI [Doğan et al., 2014], BC5CDR-Disease [Li et al., 2016], BC5CDR-Chemical [Li et al., 2016]. The hyperparameters were kept the same irrespective of the data used.

The NCBI disease corpus contains 17324 manually annotated disorder mentions from 792 PubMed abstracts. The disorder mentions are mapped to 750 MEDIC lexicon [Davis et al., 2012]) concepts. We split the released training set in 5134 training mentions and 787 development mentions, and keep the 960 mentions from the original test set as evaluation. We use the 2012 version of MEDIC ontology which contains 11915 concepts and 71923 synonyms.

BC5CDR-Disease and BC5CDR-Chemical corpora were used in the BioCreative V chemical-induced disease (CID) relation extraction challenge [2]. BC5CDR-Disease and BC5CDR-Chemical contain 12850 disease mentions and 15935 chemical mentions, respectively. The annotated disease mentions are mapped to 1075 unique concepts out of 11915 concepts in the 2012 version of MEDIC ontology [Davis et al., 2012]. The chemical mentions are mapped to 1164 unique concepts out of 171203 concepts from the 2019 version of Comparative Toxicogenomics Database (CTD) chemical ontology [3]. We use similar train-dev-test splits as outlined in the BioCreative V challenge.

Hyperparameters

BERT-based Ranking for Entity Normalization: For the text encoder we used 3 pretrained models, namely, BioBERT [Lee et al., 2019], Clinical-BERT [Alsentzer et al., 2019], PubMed BERT [Gu et al., 2021] and BERT [Devlin et al., 2018]. For fine-tuning the sentence-pair classifier, same model hyperparameters were used as those saved in the pre-trained model, with the exception of the batch size, learning rate, and number of training epochs. In this study, we fixed the learning rate at $1e^{-5}$, weight decay to $1e^{-2}$, tuned the batch size to 16 and 32, tuned the number of training epochs from 1 to 10, and saved the model with the best performance.

Triplet Search ConNorm: For the text encoder we used 3 pretrained models, namely, BioBERT [Lee et al., 2019], Clinical-BERT [Alsentzer et al., 2019], PubMed BERT [Gu et al., 2021] and BERT [Devlin et al., 2018]. We used the Pytorch implementation of sentence-transformers [4] to train the Triplet Network for concept normalization [Xu and Bethard, 2021]. We used the following hyper-parameters during training of the triplet network: `max_sequence_length` $= 16$, `train_batch_size` $= 1024$, `epoch_size` $= 3$, `optimizer` $=$ Adam, `learning_rate` $= 3e^{-5}$, `warmup_steps` $= 0$.

Regularization Parameter in Entropic Regularization. The parameter λ in Entropic regularization provides a trade-off between the amount of regularization and original cross-entropy loss (refer to sub-section Entropic Regularization in section 7.4). It was found via multiple experiments with different values that large values of λ tend to decrease accuracy on the original validation set. We thus fix $\lambda = 0.1$ for all our experiments. Set of values tried: $\{0.01, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0, 5.0\}$.

Compute Infrastructure Most of our experiments (except most Input Transformations other than gradient-based transformations) required access to GPU accelerators. We primarily ran our experiments on 2 machines: NVIDIA Quadro P5000 (16GB VRAM) and NVIDIA Quadro RTX4000 (8GB VRAM). We used Paperspace [5] and Google Colab [6] platforms to run our experiments.

[2] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/

[3] URL: http://ctdbase.org/

[4] https://www.sbert.net/

[5] https://www.paperspace.com/

[6] https://colab.research.google.com/