日本語の表記体系に着目した敵対的攻撃手法の提案

河野 竜士[1,a)]　小野 智司[1,b)]

概要：深層ニューラルネットワーク（Deep Neural Network：DNN）を使用したモデルは，入力に微小な修正を加えることで生成した敵対的事例で誤認識が生じることが知られており，自然言語処理においても同様の脆弱性が懸念されている．本手法では，日本語を扱う DNN における脆弱性の検証を目的として，単語単位では字種（ひらがな，カタカナ，漢字）の変換，フレーズ単位では語順の入れ替え，文単位では大規模言語モデルを用いた文置換の３つの手法を提案する．評価実験の結果，同義語置換を行う先行手法では発見できない日本語の表記体系特有の敵対的事例が存在することが確認できた．

A Proposal of An Adversarial Attack Method Focusing on the Japanese Language Grammar

Abstract: Models using Deep Neural Networks (DNNs) are known to be subject to misrecognition in adversarial cases generated by making small modifications to the input, and similar vulnerabilities are a concern in natural language processing. In this paper, we propose three methods to verify the vulnerability of DNNs for Japanese: word-level conversion of letter types (hiragana, katakana, and kanji), phrase-level word order replacement, and sentence-level sentence replacement using a large-scale language model. The results of evaluation experiments confirmed the existence of hostile cases peculiar to the Japanese grammar, which cannot be found by the previous method of synonym replacement.

1. はじめに

機械翻訳や文章生成人工知能（AI）などの自然言語を直接的に操作するサービスのほか，電子商取引サイトにおける商品推薦やメールや掲示板におけるコメントフィルタ機能などの自然言語を活用したサービスなど，機械学習を自然言語処理に利用した様々なサービスが提供されている．このようなサービスの急速な発展の背景には，機械技術の発展があり，特に近年の深層ニューラルネットワーク（Deep Neural Network: DNN）の急速な技術革新は，今後も自然言語の応用範囲を拡大し，多様なサービスの実現を可能にすると考えられる．

一方で DNN には脆弱性が存在することがいることが知られており，画像処理の分野ではパンダの画像に微小な摂動を付与することで CNN にテナガザルと誤推論することを明らかにしている [1]．このように，入力に微小な摂動を付与することで推論結果がもとの状態から変化するサンプルは敵対的事例（Adversarial Example: AE）と呼ばれており，このようなサンプルを意図的に生成する攻撃は敵対的攻撃と呼ぶ．そして，自然言語を扱う DNN においても同様に敵対的攻撃の影響を受け，入力文に微小な変化を付与することで推論結果が不正確となることが明らかにされている [1,2]．

上記の脆弱性は，データポイズニングなどの DNN の訓練段階における攻撃等と比較すると，実サービスに容易に悪用される可能性がある．例えば，ユーザによるコメントの記述を許可しているニュースサイトでは，深層学習ベースの大規模モデルを導入することにより「記事との関連性の低いコメント」や「過度な批判や誹謗中傷，不快な内容を含むコメント」を表示させない仕組みを取り入れている[*1]．上記の脆弱性を悪用して AE となるコメント文を入力すると，不適切なコメントが表示されるといった事態が生じてしまいサービスの安全性や信頼性の観点から懸念されている [3,4]．また，AE は，悪意を持った攻撃だけでなく，入力に対する意図しない変動などの最悪のケースとし

[1] 鹿児島大学大学院
Korimoto, Kagoshima, Kagoshima 8900065, Japan
[a)] k3933973@kadai.jp
[b)] ono@ibe.kagoshima-u.ac.jp

[*1] 大規模深層学習モデルによる Yahoo!ニュース「不適切コメント」対策
https://techblog.yahoo.co.jp/entry/2021041930133238/

ても存在の意味があり，頑健なシステムやサービスを構築する上で AE に対する頑健性を高めることは重要である．以上のような理由により，自然言語を扱う DNN に対して敵対的攻撃を行うことで AE を発見する様々な手法が提案されている．基本的な攻撃は，文字または単語単位での攻撃であり，文字単位の攻撃では文字の置換や挿入そして削除 [5–7]，単語単位ではもとの単語を同義語に置換を行うことで攻撃を行う [8–10]．また，その他の攻撃手法として文単位での攻撃や各言語の表記体系に特有の性質を利用した攻撃が存在する [11–13]．

本論文では，日本語に特化した摂動の付与を行う敵対的攻撃方式を提案する．自然言語 DNN を対象とした従来研究の多くは，特定の言語に特化しない汎用的な摂動の付与方法が検討されている．一方で日本語は，ひらがなや漢字など複数の表記体系を持つことや，フレーズの順序に対する文法的な制約が緩いなどの特徴を持つことから，他の言語にはない独自の脆弱性が存在すると考えられる．提案手法は，ひらがな，カタカナ，漢字の字種の間で変換を行う摂動を基本とし，さらに，語順の自由度の高さを活用したフレーズの入れ替えを行う摂動を提案する．特に，提案手法は，対象とする DNN の内部情報を利用しないブラックボックス条件下での攻撃を行う．これにより，近年のブラックボックス化がすすむ商用モデルの解析が可能となる．実験により，提案手法はブラックボックス条件下で日本語を扱う DNN で誤推論を引き起こす脆弱性を確認した．また主観評価では，提案手法の敵対的事例が文法的に破綻しておらず，入力文の内容を維持したものになっていることを確認した．

2. 関連研究

DNN は，微小な入力の変化によってその挙動が顕著に変わる脆弱性を有することが知られている．Szegedy らは，入力画像に微小な摂動を加えて生成した画像を画像分類用 DNN に入力した際，DNN が誤って分類する事象を発見し，DNN がこの種の摂動に対して頑健ではないことを示した [2]．このように，人間の識別には影響を及ぼさない程度であるが，DNN の誤分類を引き起こす摂動が付与された事例を敵対的事例（Adversarial Examples: AE）と命名した．Goodfellow はこの現象についてさらに議論を行い，AE を高速に生成する手法 Fast-Gradient Sign Method（FGSM）を提案した [1]．その後，FGSM を反復的に適用することでより摂動の少ない AE を生成する BIM [14] や，識別境界を線形的に近似しこれに垂直なベクトルを求めることでより原画像に類似する AE を発見する DeepFool [15] などが提案されている．

上記の敵対的攻撃手法はいずれも，対象となる DNN の内部情報を利用することから，ホワイトボックス（White-Box: WB）攻撃と呼ばれる．WB 攻撃は，対象となる DNN の損失関数の勾配などを活用することで高速に AE を生成できる点に特徴がある．一方で，WB 攻撃を用いて，DNN の内部情報を参照できない商用のシステムやサービスなどの脆弱性を外部から行うことは難しい．

このため，損失関数の勾配などの DNN の内部情報を参照しないブラックボックス（Black-Box: BB）敵対的攻撃手法も多く提案されている．Narodytska らは，原画像における正しいクラスの信頼度を最小化することで AE を生成する手法を提案した [16]．Su らは，1 画素のみに摂動を付与することで画像識別器の誤認識を生じさせることが可能であることを示した [17]．Brendel らは，正解クラスの信頼度を利用せず，識別器が出力する Top-1 ラベルのみを用いて AE を生成する境界攻撃を提案した．これらの BB 攻撃は，商用サービスやシステムの内部情報を利用しなくとも AE の発見が可能である点に特徴がある．

上記のように，敵対的攻撃の多くは画像を扱う DNN を対象としているが，自然言語を扱う DNN を対象とした研究も行われている．Jia らは，自然言語における AE の生成について議論を行い，画像と比較して，離散的であること，摂動が知覚されやすいこと，摂動により意味や文法が破綻しやすいことの 3 点により AE の生成が難しいことを示した [3]. Ebrahimi らや Liang らは，入力文において文字列を操作する，すなわち，文字の置換や挿入，削除を行うことで自然言語を扱う DNN における AE を生成できることを示した [5, 6]．Gao らは，上記を BB 条件下で行う手法を提案した [7].

上記のような文字単位の摂動に加えて，入力文内の単語を同義語に置換する手法が提案されている．Ren は，同義語がまとめられた概念辞書（WordNet）を使用して，同義語を生成する手法を提案した [8]．また，Alzantot らや Jin らは，単語の埋め込み空間の距離を計算することで，同義語を生成する手法を提案した [9, 10]．Jin らの手法では，入力文と生成した AE とを Universal Sentence Encoder（USE）[18] で埋め込み表現に変換し，コサイン類似度で比較することで 2 つの文章の類似度を測定し，類似度が一定の閾値を超える AE のみを採用することで意味的一貫性を高めた AE を生成できる手法となっている．上記のような単語単位の置換に加えて，機械翻訳を活用して類似文を生成することで摂動を付与する方法も提案されている [11].

なお，特定の言語に特化した摂動の付与方法についても研究が行われている．Zhang は，中国語の特徴である表意文字や文字構成などに着目し編と旁で構成された文字を 2 つの漢字に置換する摂動やピンインが同じ漢字に置換する摂動などを提案した [12]．Boucher らは，アラビア語など右から左に文字が配置される言語言語を表記するために使用される制御文字を摂動として使用する方式を提案した [13].

3. 提案する方式

3.1 基本アイデア

本研究では，日本語の特性に着目した摂動の付与を行うBB 敵対的攻撃方式を提案する．日本語の特性に起因するDNN の脆弱性を発見することで，システムやサービスの頑健性，安全性の向上に寄与する．また，ブラックボックス条件下での敵対的攻撃を可能とすることで，近年の商用大規模言語モデルなどを API 経由で第 3 者が利用してシステムやサービスを構築する際に，当該サービス等に特有の脆弱性を DNN の外部から発見することを可能にする．

提案手法は，日本語の特性を活用し，より知覚されにくい自然な AE を生成するため，単語単位，フレーズ単位，文単位の 3 種類の操作を組み合わせた AE の生成を行う．本手法は，ひらがな，カタカナ，漢字の字種間で単語を変換する操作を基本的に行う [19]．この操作は，日本語が複数の表記体系を使用するという特性を取り入れたものであり，同じ単語の異表記に書き換えを行うため，内容が変化する可能性が極めて低いという特徴を持つ．しかし，字種変換のみでは，入力文に対して付与可能な摂動の種類が限られてしまうため，本手法は上記の字種変換に加えて，フレーズ単位および文レベルの摂動の付与を行う．フレーズ単位の摂動は，入力文内のフレーズの順序を入れ替える操作とする．この操作は，日本語が語順の入れ替えを行っても，英語とは違い，文法や内容が破綻しないという日本語の表記体系における特徴を取り入れたものである．文単位の操作は，大規模言語モデルを活用して，入力文に類似する文に書き換えを行うことで摂動を付与する．

3.2 提案する方式の処理手順

提案手法のアルゴリズムの概要を図 3.2 に示す．本手法は，貪欲法にもとづいて入力文に対して徐々に摂動を付与し，AE を生成する．まず，入力文に対して，フレーズ単位および文単位の摂動を付与して類似文集合を作成し，このなかの文それぞれに対して，重要度スコアに基づく単語の選択，および，選択された単語への摂動の付与を誤認識が引き起こされるまで繰り返す．フレーズおよび文単位の摂動の生成，置換する表記の候補の抽出，重要度スコアでの単語のランク付け，解候補の生成，および評価の 5 つの手順について，以下にその詳細を示す．

(1) 文・フレーズ単位の摂動の生成: 図の 1，2 行目では，入力文 $\boldsymbol{X}$ に対して文単位，フレーズ単位の摂動を付与し，入力文に類似する文の集合 $\boldsymbol{L}$ を生成する．文単位の摂動は，大規模言語モデルでプロンプトを使用することで類似文を取得することで付与し，フレーズ単位の摂動は，構文解析器で取得したフレーズを入れ替えることで付与する．続いて，生成された類似文を攻撃対象モデルに入力し，正解以外のクラスの信頼度を取得する．上記の信頼度の降順で類似文をソートし，$\boldsymbol{L}$ とする．

Algorithm 1 提案手法における AE の生成アルゴリズム

Input: X: 入力文，Y: 入力文の正解ラベル，F: 攻撃対象モデル，$Sim(\cdot)$: 文の類似性を評価する関数
Output: X': 入力文に摂動を付与した AE

1: フレーズ単位，文単位の摂動を付与して類似文集合 $\boldsymbol{L}$ を生成
2: $\boldsymbol{L}$ をソート
3: **for** L_i in $\boldsymbol{L}$ **do**
4: L_i の単語のリスト $\boldsymbol{W_{L_i}} = \{w_1, w_2, ..., w_n\}$ を作成
5: **for** w_j in $\boldsymbol{W_{L_i}}$ **do**
6: 字種変換や同義語置換を適用し，置換候補集合 $\mathcal{C}_{ij}$ を作成
7: 重要度 $I_{W_{L_i}}$ を計算
8: **end for**
9: $\boldsymbol{W_{L_i}}$ を重要度の降順でソート
10: $\boldsymbol{W_{L_i}}$ 内のストップ語を除去
11: $X' \leftarrow L_i$
12: **for** w_j in W_{L_i} **do**
13: **for** c_k in $\mathcal{C}_{ij}$ **do**
14: $s_{ijk} \leftarrow$ replace (X', 注目する単語 w_j, c_k)
15: s_{ijk} を解候補リスト $\boldsymbol{S}_{ij}$ に追加
16: $\boldsymbol{S}_{ij}$ を $Sim(X, s_{ijk})$ の降順でソート
17: **end for**
18: **for** s_{ijk} in $\boldsymbol{S}_{ij}$ **do**
19: $Y_k \leftarrow s_{ijk}$ の分類結果 $F(s_{ijk})$
20: **if** $Y_k \neq Y$ **then**
21: $X' \leftarrow s_{ijk}$
22: **return** X'
23: **end if**
24: **end for**
25: $k_{min} \leftarrow \arg\min_k$ 正解クラスの信頼度 $P_Y(s_{ijk})$
26: $X' \leftarrow$ replace(X', w_j, c_k)
27: **end for**
28: **end for**
29: **return** None

図 1 提案手法のアルゴリズム

prompt = ”ユーザー: 入力文と殆ど同じ内容の文章を日本語で 5 個以上作成してください。
ただし、出力は python 用に文字列のリスト形式である [ ] に入れてを出力してください。
オリジナルの入力にない余計な前置きの文章は必要ありません。\ n\ n 入力文:” + [入力文] + ”\n\n 出力:”

図 2 入力文に類似した文を取得するために利用したプロンプト

(2) 単語の置換候補の抽出: 図の 4 行目から 6 行目では，形態素解析を行うことで入力文 X を構成する単語列 $X = \{w_1, w_2, \ldots, w_n\}$ を取得する．同時に，単語 w_i のひらがな，カタカナ，漢字の各表記を取得し，置換候補集合 $\mathcal{C}_{ij}$ に収納する．なお，先行手法である同義語辞書や埋め込み表現による同義語置換もオペレータとして併用する場合も，ここで置換語の候補として取得し，$\mathcal{C}_{w_i}$ に追加する．

(3) 重要度スコアにもとづく単語のランク付け: 前記の処理で得られた各単語にランク付けを行う（図 7-10 行目）．

すなわち，3.3 節で後述する重要度を用いて単語 w_i が信頼度に与える影響力 I_{L_i} を計算し，降順でソートした集合 W を生成する．

なお，ストップワードの除外もこの段階で行う．

(4) 解候補の生成: 入力文 X の w_i を摂動の候補である $\mathcal{C}_{w_i}$ のなかの置換候補の 1 つ c に置換し，AE である $X' = \{w_1, ..., w_{i-1}, c, w_{i+1}, ..., w_n\}$ を得る．攻撃対象モデル F を用いて X' を分類し，ラベル $F_{X'}$ と，それに対応した信頼度 P_k を計算する．

(5) 解候補の評価: 攻撃対象モデルの予測を変更できる候補を生成できた場合は，その候補 X' を AE として出力して探索を終了する．予測結果が入力文と同様である場合は，対象モデルの予測が変更されるまで，上記 (2) から (5) までの処理を繰り返す．

3.3 影響度に基づく単語のランク付け

単語 w_i の影響度 I_{w_i} は以下の式で算出する [10, 20]．

$$I_{w_i} = \begin{cases} P_Y(X) - P_Y(X_{\setminus w_i}) \\ \qquad \text{if } F(X) = F(X_{\setminus w_i}) \\ \\ (P_Y(X) - P_Y(X_{\setminus w_i})) + (P_{\overline{Y}}(X_{\setminus w_i}) - P_{\overline{Y}}(X)) \\ \qquad \text{if } F(X) \neq F(X_{\setminus w_i}) \end{cases} \tag{1}$$

ここで，$X_{\setminus w_i}$ は，入力文 X から任意の単語 w_i が削除された文を意味する．$F(\cdot)$ は分類器による分類結果を表し，$P_Y(X)$ は X を分類した際のクラス Y の信頼度を示す．これにより，DNN の勾配などの内部情報を参照することなく，少ない単語の置換回数で，誤分類を引き起こす AE を効率的に生成することが可能となる．

3.4 単語単位の摂動

提案手法では，基本的に，入力文を構成する単語に着目し，ひらがな，カタカナ，漢字の字種間で変換することにより摂動を付与する．字種変換の候補となる単語は形態素解析用の辞書 UniDic から取得する．なお，字種変換を適用する単語は，名詞，動詞，形容詞，形容動詞，副詞に限定する．

取得した解候補は，USE（Universal Sentence Encoder: USE）[18] で埋め込み表現に変換した入力文と解候補を cos 類似度で評価し，降順でソートしたものを字種変換の解候補として採用する．

3.5 フレーズ単位の摂動

3.4 節で述べた単語単位の摂動に加えて，提案方式ではフレーズ単位での摂動を付与し，単語単位の摂動のみでは AE を生成できない入力文に対処する．提案手法におけるフレーズ単位の摂動は，構文解析によって得られるフレーズの順序を入れ替える変化となる．すなわち，取得したフレーズ集合のうち述部に該当するフレーズを固定して，他のフレーズを並び替える全てのパターンを類似文の候補とする．得られた類似分の候補を攻撃対象のモデルに入力し，正解ラベル以外の信頼度について降順でソートし，正解ラベル以外の信頼度が高い類似文から AE の候補として利用する．

表 1 攻撃対象のモデルの精度

	KIT-BERT	Rakuten-BERT	TextAnalytics
Acc	60.4	95.2	91.4
Positive	70.4	95.8	94.8
Negative	50.4	94.6	88.0

3.6 文単位の摂動

3.5 節のフレーズ単位の摂動で得られる類似文の数は限られるため，提案手法では，大規模言語モデル（Large Language Model: LLM）を用いて入力文の類似文を生成することを文書全体に摂動を加える操作とみなす．提案手法では，入力文に類似しつつ入力文から大幅に変化しない文を得るため，図 2 に示すプロンプトを LLM に与えることで類似文を生成する．取得した類似文は，フレーズ単位の摂動で得られた類似文と同様に扱い，攻撃対象のモデルに入力し，正解ラベル以外の信頼度について降順でソートし，正解ラベル以外の信頼度が高い類似文から AE の候補として利用する．

4. 評価実験

4.1 実験設定

提案手法の有効性を検証するため，楽天市場のデータセット [21] を対象とした感情分析タスクを用い，商用モデルを含む 3 モデルに対して AE を生成することでそれらの脆弱性を分析した．楽天データセットに含まれる楽天グループのサービスや商品のレビュー文とレビュースコアが含まれる．本研究では，レビュースコアが 1 または 2 のレビュー文はネガティブな内容，4 または 5 のレビュー文はポジティブな内容として正解ラベルを割り当てることで，感情分析を二値分類問題として定式化してデータを作成した．

本実験では，以下の 3 つの言語モデルを攻撃対象モデルとして敵対的攻撃を行った．

- 北見工業大学が作成した感情分析モデル（KIT-BERT）[22]
- 楽天市場のデータを用いてファインチューニングを行ったモデル（Rakuten-BERT）
- Microsoft Azure の Azure Cognitive Services で提供される感情分析モデル（TextAnalytics）

KIT-BERT および Rakuten-BERT は，東北大学により訓練

表 2　従来手法と提案手法で採用する摂動付与方法

	従来手法 1	従来手法 2	提案手法 1	提案手法 2	提案手法 3	提案手法 4	提案手法 5	提案手法 6	提案手法 7
同義語置換（辞書）	○						○		
同義語置換（埋込表現）		○						○	○
字種変換			○	○	○	○	○	○	○
フレーズ置換				○		○			○
文置換					○	○			○

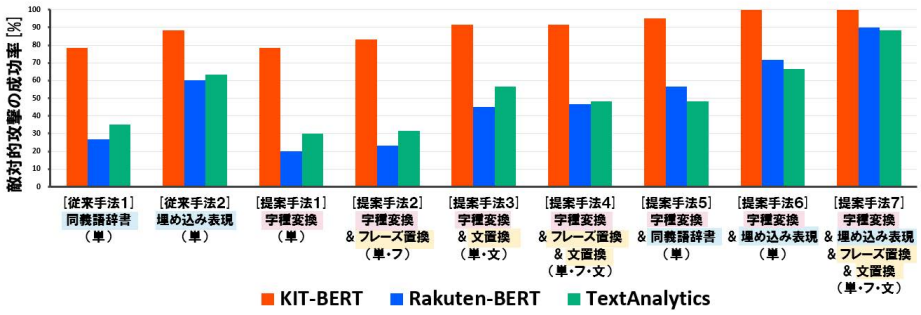

図 3　モデル別の攻撃成功率

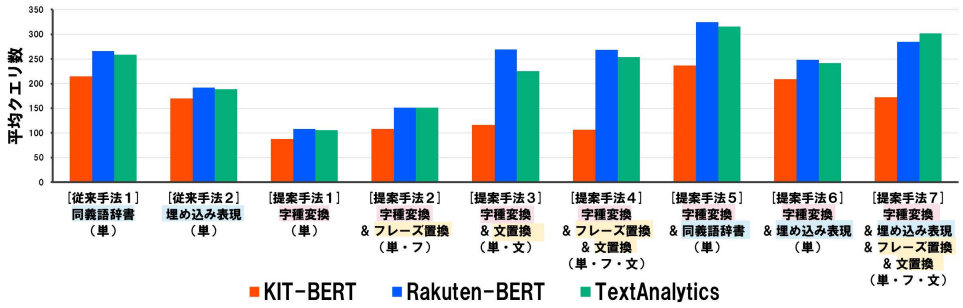

図 4　モデル別の平均クエリ数

された BERT モデル bert-base-japanese-sentiment をファインチューニングして作成しており，その際，KIT-BERT は皮肉や皮肉のツイートを用い，Rakuten-BERT は楽天市場のデータ 1 ヶ月分を用いている．

上記のモデルの感情分析の性能を，Rakuten-BERT の学習には使用していない 2018 年 1 月の楽天データセットを用いて評価した結果を表 1 に示す．ポジティブとネガティブの文章を 1:1 の割合で含む 1,000 サンプルにおける正解率を上段に，各ラベル（ポジティブ，ネガティブ）の分類性能に着目したときの結果を下段に示す．商用モデルである TextAnalytics やショッピングサイトのレビュー文における感情分析に特化させたモデルである Rakuten-BERT のスコアが高いことが確認できる．

提案手法の実装は，自然言語処理用敵対的攻撃フレームワーク TextAttack をベースに行った [23]．フレーズ単位の摂動を付与するための構文解析は，日本語自然言語処理ライブラリ GiNZA を用いて行った．文単位の摂動を生成する際の LLM として，OpenAI 社が提供する GPT-4 を使用した．

本実験では，同義語辞書を用いて単語の置換を行う敵対的攻撃手法 [8]（従来手法 1），および，単語の埋め込み空間の距離計算により同義語の置換を行う手法 [10]（従来手法 2）と提案手法の比較を行うこととした．従来手法 1 において，同義語のコサイン類似度による制約処理では USE [18] を用いた．また，従来手法 2 において，埋め込み表現による同義語の生成は東北大学乾研究室の word2vec モデルを[*2]を使用し，同義語の余弦類似度による制約処理は USE [18] を用いた．

提案手法は，採用する操作の組合せにより，表 2 に示す 7 パターンの手法を用意した．

4.2　実験結果

各手法を用いて 3 種類のモデル（KIT-BERT，Rakuten-BERT，TextAnalytics）に対して攻撃を行った結果を図 3 および図 4 に示す．検証用データとして，2018 年 1 月の楽天データセットのうち，3 種類の識別モデル（KIT-BERT，Rakuten-BERT，TextAnalytics）の全てが正しく識別を行った事例からランダムに 60 事例を抽出した．このとき，

[*2] 日本語 Wikipedia エンティティベクトル
https://github.com/singletongue/WikiEntVec/releases

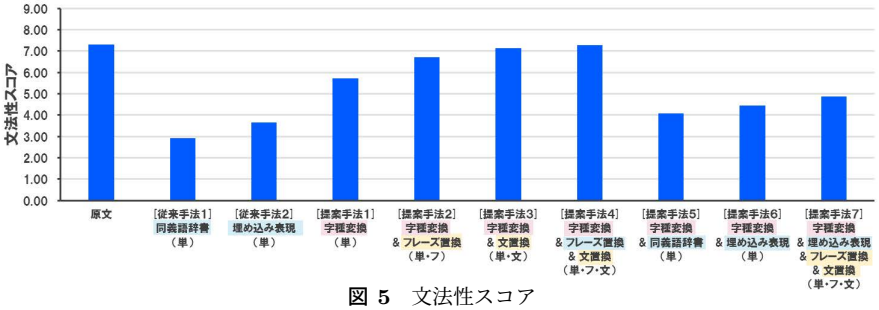

図 5 文法性スコア

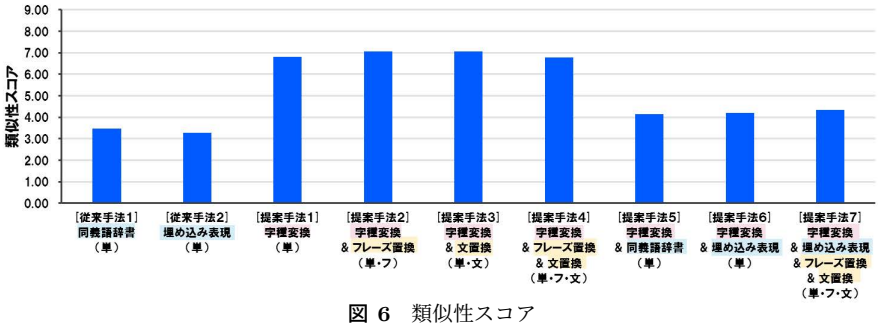

図 6 類似性スコア

文章の長さが偏らないように，3 種類の長さの事例（形態素数が 20 以下，21 以上 40 以下，41 以上 60 以下）を 20 個ずつ選択した．

図 3 のグラフは攻撃成功率を表しており，検証に利用した事例のうち攻撃に成功した割合を表す．図 3 より，字種変換のみを行う提案手法 1 で生成した AE であっても，同義語の置換を行う従来手法 1，2 ほど攻撃成功率は高くないものの，対象モデルの誤認識を引き起こせることがわかる．また，提案手法 2，3 および 4 の攻撃成功率が提案手法 1 よりも高いことから，字種変換にフレーズ置換や文置換を組み合わせることでより多くの事例で AE の作成に成功したことがわかる．なお，フレーズ置換と文置換を比較すると，文置換の方が効果が高いことも示された．

提案する字種変換に従来手法の同義語置換を組み合わせた提案手法 5 および 6 と，従来手法 1 および 2 の攻撃成功率を比較すると，提案手法 5 および 6 の方が高いことから，字種変換により，従来の同義語置換では行えない種類の摂動の付与が可能となったことがわかる．また，提案手法 7 の攻撃成功率は他の全ての手法よりも高く，4 種類の摂動付与方法を併用することで，商用サービスである TextAnalytics を含めてすべてのモデルで高い成功率で AE を発見できたことがわかる．なお，3 種類の攻撃対象モデルに着目すると，KIT-BERT の AE の発見は容易であったことに対して，他の 2 種類のモデルは KIT-BERT と比較してより敵対的攻撃に対して耐性を有していることがわかる．

また，図 4 のグラフは，攻撃に成功した際において，AE を生成するまでに必要とした対象モデルの呼び出し回数（クエリ数）の平均を示す．字種変換のみを行う提案手法 1 がもっとも平均クエリ数が少ないことが確認できるものの，攻撃成功率の低さとあわせて考えると，字種変換により加えることができる摂動の種類が限定されていることがわかる．また，組み合わせる操作の種類が増えることにより，攻撃に要するクエリ数が増加する傾向がみられる．例えば，提案手法 2，3 および 4 の結果から，字種変換にフレーズ置換や文置換を組み合わせることでクエリ数が増加する．また，字種変換と従来手法の同義語置換を組み合わせた提案手法 5 および 6 は，従来手法 1 および 2 と比較してクエリ数が増加した．なお，4 種類の操作を組み合わせた提案手法 7 は，攻撃成功率がもっとも高かったことに加えて，クエリ数も提案手法 5 と同程度に抑えられており，300 クエリ程度で AE を発見できたことがわかる．

続いて，生成された AE の品質を評価するために，20 名の被験者により，文法性および入力文との類似性に関する主観評価を行った．本評価では，多くの手法で共通して攻撃に成功した事例を選択するために KIT-BERT における AE を対象とすることとし，生成した 60 事例の AE のうち 5 事例を選択した．被験者はすべての手法で生成された AE について，9 段階で評価を行い，高い品質の AE ほど高い得点を割り当てることとした．20 名の被験者による主観評価を行った結果を図 5 および図 6 に示す．字種変換を用いる提案手法 1 と，同義語置換を用いる従来手法 1 および 2 とを比較すると，文法性および意味の類似性の双方において，字種変換の方が品質の高い AE を作成することがわかる[*3]

字種変換にフレーズ置換や文置換を組み合わせた場合は，文法性は改善されるものの類似性は同程度であった．提案手法 5，6 および 7 において従来手法の同義語変換を併用

[*3] 有意水準 5 %のマンホイットニーの U 検定により有意差があることを確認した..

表 3 生成された AE と主観評価結果の例

手法	文	予測ラベル	文法性	類似性
入力文	6 ヶ月のパピーに使用。 お散歩後の足の裏の汚れ落としに使用してます。使用後もしっとり。 足を洗うのを嫌う為、かなり重宝しそうです.	Pos		
従来手法 1	6 ヶ月のパピーに苦心。 お散歩後の足の裏の汚れ落としに要請してます。 使用後もしっとり。 足を洗うのを嫌う為、かなり重宝しそうです.	Neg	3.45	3.25
従来手法 2	6 ヶ月のパピーに使用。 お散歩後の足の裏の汚れ落としに使用してます。 使用後もツヤツヤ。 足を洗うのを嫌う結果的、極めて重宝しそうです.	Neg	5.35	5.90
提案手法 1	6 ヶ月のパピーにしよう。 お散歩後の足の裏の汚れ落としにシヨーしてます。 使用後もしっとり。 足を洗うのを嫌うタメ、かなり重宝しソウです.	Neg	4.70	6.10
提案手法 2	6 ヶ月のパピーに使用。使用後もしっとり。お散歩後の足の裏の汚れ落としに使用してます。足を洗うのを嫌うタメ、かなり重宝しソウです.	Neg	6.55	7.25
提案手法 3	6 ヶ月のコイヌに対してしよう。お散歩後の足のウラのヨゴレを落とすタメに使用しています。使用後も肌はシットリ。足を洗うのが嫌なので、かなりべんりソウです.	Neg	6.40	6.40
提案手法 4	6 ヶ月のパピーに使用。使用後もしっとり。お散歩後の足の裏の汚れ落としに使用してます。足を洗うのを嫌うタメ、かなり重宝しソウです.	Neg	7.45	7.85
提案手法 5	6 ヶ月のパピーに苦心。 お散歩後の足の裏の汚れ落としに使用してます。 使用後もしっとり。 足を洗うのを嫌う為、かなり重宝しソウです.	Neg	4.80	4.15
提案手法 6	6 ヶ月のパピーに使用。 お散歩後の足の裏の汚れ落としに使用してます。 使用後もしっとり。 足を洗うのを嫌う結果的、かなり重宝しソウです.	Neg	4.80	6.40
提案手法 7	使用後もしっとり。6 ヶ月のパピーに使用。お散歩後の足の裏の汚れ落としに使用してます。足を洗うのを嫌う結果的、かなり重宝しソウです.	Neg	6.15	6.25

することで文法性や類似性が低下することがわかる．以上のことから，提案手法の字種変換，フレーズ置換および文置換は，従来手法と比較して入力文に類似する自然な AE を作成できたことがわかる．

表 3 に，各手法で得られた AE と主観評価結果の例を示す．従来手法の同義語置換では，「使用」を「要請」に変更しているために文の意味が変化したり，「為」を「結果的」に変換しているために不自然な文となっている．提案手法の結果をみると，「しそう」が「しソウ」と変更されているなどやや不自然さを感じるが，ソーシャルネットワークサービス上でカジュアルに記述されるコメントなどにおいて，希に観察される表記のようにも見える．

5. 結論

本研究では，DNN の内部情報を用いないブラックボックス条件下で，日本語の特性に着目した摂動を付与することで敵対的攻撃を行う手法を提案した．本手法は，ひらがな，カタカナ，漢字の字種の間で変換を行う摂動と，フレーズの位置を入れ替える摂動，ならびに言語モデルを用いて文全体を類似文に置き換える変動，すなわち，単語，フレーズ，文全体の 3 種類の範囲の摂動を併用する方式である．感情分析タスクを用いた評価実験により，提案手法は商用サービスを含めた 3 種類の DNN における AE を生成でき，日本語の特性が活用される脆弱性が存在することを確認した．特に，同義語置換を行う従来方式とは異なる摂動であり，併用することで攻撃成功率を高めることができること，および，従来方式よりも文法の破綻を抑えつつ入力文との類似性も維持される AE を生成できることが確認された．

今後，摂動の候補を選択する際の優先度を見直すことで，生成される AE の品質の改善やクエリ数の削減を図る．

謝辞

本研究の一部は JSPS 科研費 JP22K12196 の助成による．また，国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」（https://rit.rakuten.com/data_release/）を利用した。

参考文献

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

[2] 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.

[3] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

[4] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*, 2017.

[5] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*, 2017.

[6] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. *arXiv preprint arXiv:1704.08006*, 2017.

[7] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 50–56. IEEE, 2018.

[8] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang

Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1085–1097, 2019.

[9] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018.

[10] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, pp. 8018–8025, 2020.

[11] Ang Li, Fangyuan Zhang, Shuangjiao Li, Tianhua Chen, Pan Su, and Hongtao Wang. Efficiently generating sentence-level textual adversarial examples with seq2seq stacked auto-encoder. *Expert Systems with Applications*, Vol. 213, p. 119170, 2023.

[12] Zihan Zhang, Mingxuan Liu, Chao Zhang, Yiming Zhang, Zhou Li, Qi Li, Haixin Duan, and Donghong Sun. Argot: Generating adversarial readable chinese texts. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 2533–2539, 2021.

[13] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. Bad characters: Imperceptible nlp attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1987–2004. IEEE, 2022.

[14] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

[15] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

[16] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *CVPR Workshops*, Vol. 2, p. 2, 2017.

[17] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, Vol. 23, No. 5, pp. 828–841, 2019.

[18] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[19] 河野竜士, 玉城大生, 小野智司. 日本語処理用深層学習器における脆弱性を検証する敵対的攻撃の基礎検討. 人工知能学会全国大会論文集 第 36 回 (2022), pp. 1P4GS601–1P4GS601. 一般社団法人 人工知能学会, 2022.

[20] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*, 2018.

[21] 楽天データセット - 国立情報学研究所. `https://www.nii.ac.jp/dsc/idr/rakuten/`.

[22] 団俊輔, プタシンスキミハウ, ジェプカラファウ, 桝井文人. 北見工業大学 テキスト情報処理研究室 electra base 皮肉検出モデル (daigo ver.). HuggingFace, 2022.

[23] John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*, 2020.