敵対的事例に対する日本語処理モデルの頑健性向上の試み

秋本 一樹 [1] 森本 文哉 [1] 小野 智司 [1]
[1] 鹿児島大学
{k3394902, k5801545}@kadai.jp ono@be.kagoshima-u.ac.jp

概要

深層ニューラルネットワーク（Deep Neural Network: DNN）は，微小な変更が加えられた入力データである敵対的事例（Adversarial Examples: AE）を誤認識してしまう脆弱性が存在する．日本語処理用の DNN も同様の脆弱性が存在し，このような脆弱性は，DNN を実世界の課題に応用する上で重大な障害となり得る．本研究では，AE に対する日本語処理モデルの頑健性向上を目的とし，日本語に特化した再攻撃による敵対的防御手法を提案する．実験により，提案手法が通常事例の分類精度を維持しつつ，AE の矯正を行えることを確認した．

1 はじめに

近年，DNN モデルは，自然言語処理分野においてカテゴリ分類や要約，機械翻訳，迷惑メールフィルタ，有害コメントの識別など様々なタスクで活用されている．一方で，画像処理分野において，Szegedy らは AE と呼ばれる特殊に設計された入力によって，DNN モデルが誤認識を引き起こす脆弱性の存在を明らかにした [1]．AE は，人間が知覚できないほど微小な摂動を DNN モデルが誤認識するよう設計し，入力に付与することで生成される．このような脆弱性は，実社会で DNN モデルを活用する際のリスクになり得る．例えば，自動車に搭載された道路標識識別モデルが，”止まれ”と書かれた標識を別の標識に誤認識した場合，自動車が止まらず事故につながる可能性が考えられる．

自然言語処理用の DNN モデルにおいても同様の脆弱性が確認されている．Liang らは，入力テキスト内の任意の文字を挿入，変更，削除する 3 種の摂動により AE を生成し，DNN ベースのテキスト分類器が誤認識を引き起こすことを示した [2]．このように AE を生成する手法を敵対的攻撃と呼び，DNN モデルの脆弱性解明を目的として研究されている．

DNN の持つ脆弱性は，DNN を実世界の課題に応用する上で重大な障害となり得る．このため，DNN モデルの AE に対する頑健性向上を目的とした敵対的防御手法の研究も広く行われている．

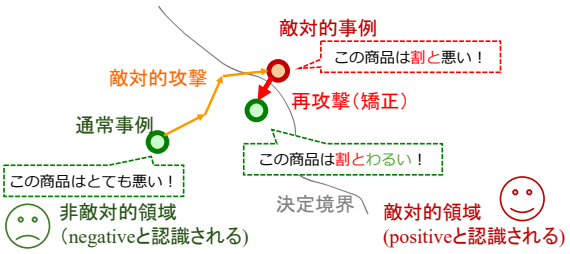

図 1 提案手法の概要

一般に，自然言語 DNN を対象とした敵対的攻撃や防御に関する研究は，特定の言語に特化しない汎用的な摂動の付与方法を検討している．これに対して河野らは，日本語に特化した敵対的攻撃手法を提案し，日本語処理モデルにおいて日本語の特性が活用される脆弱性が存在することを示した [3]．日本語処理用 DNN を対象とした敵対的防御の研究は十分に行われておらず，日本語に特化した攻撃に対して従来の防御手法 [4, 5] の性能が未知であるほか，日本語の特性を活かした防御方法も，著者らが調べた限りこれまでに研究されていない．

このため本研究では，AE に対する日本語処理モデルの頑健性向上を目的とし，日本語に特化した敵対的防御手法を提案する．AE に対して日本語特有の表記体系である字種を変換する再攻撃を行うことで，AE の分類ラベルの矯正を試みる．実験では，3 つの攻撃手法により生成した AE と通常事例を対象として，再攻撃を行い，通常事例の分類結果を維持しつつ，AE のラベルのみを矯正できることを示す．

2 関連研究

2.1 敵対的攻撃

敵対的攻撃は，画像分類モデルを対象とした研究が盛んに行われている．モデルの内部情報である勾配に基づいて AE を生成する FGSM[6] や，FGSM を

反復的に適用することでより摂動量の少ない AE を生成する BIM[7]，ヤコビアン行列に従って得られた敵対的顕著性マップを用いて，重要なピクセルを選択することで効率的に AE を生成する JSMA[8] など，これらの手法は敵対的事例研究の基盤となり，多くの後続研究に影響を与えている．

自然言語を対象とした敵対的攻撃のアプローチは，主に文字レベル，単語レベル，文レベルの 3 つに大別される [9]．文字レベルの攻撃では，与えられたテキスト内の文字を挿入，削除，置換，交換を行う [10, 11]．単語レベルの攻撃では，入力テキスト内の任意の単語を同義語などの別の単語に変更する．置換単語の選択には，WordNet を使用した辞書ベースの方法 [12]，学習済み単語埋込みベクトルを使用した方法 [13] や，言語モデル BERT にマスクしたトークンを予測させる生成ベースの方法 [14] などが提案されている．単語レベルは文字レベルの攻撃と比較して，意味の一貫性が保たれることで，人間が AE として認識することが困難になる．文レベルの攻撃では，新しい文の挿入や，元の文構造の変更といった手法がある [15, 16]．他のアプローチと比較して，AE の意味が維持され，多様性に富んでいる一方で，意味のないトークンが生成され可読性を低下させる場合がある．

日本語に特化した敵対的攻撃手法として，河野らは日本語の表記体系の特徴に着目し，字種変換による攻撃手法を提案した [3]．Jin らが提案した単語重要度スコア [13] に基づき，文中から重要度の高い単語を選択し字種変換することによって，原文に類似し文法性を維持した AE を生成できる．

2.2 敵対的防御

自然言語を対象とした敵対的防御のアプローチは，主に敵対的訓練，摂動制御，証明に基づく手法の 3 つに大別される [17]．

敵対的訓練は，敵対的攻撃によって生成された AE を訓練データに含めて訓練することで，モデルの頑健性向上を目的とする手法である．Wang らは，従来の研究において問題とされていた訓練段階における AE 生成の効率化に焦点を当て，新たな攻撃手法とともに効率的な敵対的訓練手法を提案し，モデルの頑健性を大幅に向上させた [18]．

摂動制御は，摂動が加えられた入力を識別し，AE の検出や AE のラベル矯正を目的とする手法である．Mozes らは，単語レベルの敵対的攻撃が，入力

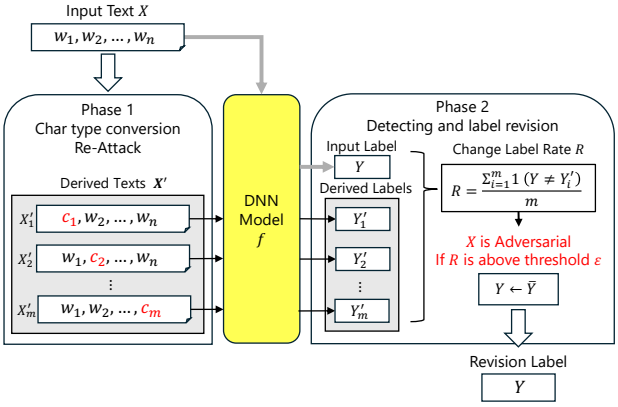

図 2　提案手法のアルゴリズム

単語を出現頻度の低い単語に置き換える傾向にあることを発見し，出現頻度がより高い単語に置換する FGWS[4] を提案した．また，Wang らは，単語を同義語とランダムに置換することで敵対的摂動を排除する RS&V[5] を提案した．入力テキスト内の単語を，同義語とランダムに置換することによって生成された複数のテキストのラベルを集計し，最も数の多いラベルを入力テキストのラベルとすることで，AE のラベルを矯正できることを示した．FGWS や RS&V はモデル構造の変更や追加訓練を必要としない単純な手法でありつつ，既存の研究より優れた防御性能を発揮した．

証明に基づく手法は，モデルの頑健性を理論的に保証した訓練方法により，敵対的空間を探索しないことでモデルの頑健性向上を目的とする手法である．証明に基づくアプローチは，DNN のような複雑なモデルや大規模データセットへの拡張が困難であることが指摘されている [17]．

2.3 先行研究

画像処理分野における敵対的防御として，Zhao らは AE の脆弱性に着目して検出を行う Attack as Defense（A^{2}D）を提案した [19]．AE の脆弱性とは，AE が特徴空間において決定境界付近に位置するために，再び攻撃を受けることで分類結果が容易に変わってしまう性質を示す．A^{2}D は，検出にのみ焦点をおいており，攻撃前の原画像の正しいクラスを得ることを考慮していない．

このため森本らは，検出された AE に対して再攻撃を行うことで，AE の分類結果を矯正できることを示した [20]．さらに，再攻撃時の摂動を微小に抑えることにより，AE と通常事例を区別することなく正しい分類結果に矯正できることを示した [21]．

3 提案手法

3.1 キーアイデア

本研究では，日本語特有の表記体系である字種に着目し，再攻撃による AE の検出及び矯正を行う敵対的防御手法を提案する．すなわち提案手法は，先行研究 [5] で提案された再攻撃による敵対的事例の検出・矯正方式において，日本語処理用 DNN 向けの攻撃方式 [3] を再攻撃に使用する手法である．

提案手法の概要を図 1 に示す．本手法は，先行研究 [19, 21, 5] と同様に，AE の脆弱性に着目する．優れた敵対的攻撃手法ほど摂動の量が最小化され，AE が決定境界付近に生成されることとなり，微小な再攻撃により通常事例と AE との識別やラベルの矯正が行えると考える．

提案手法は，再攻撃時の摂動として字種変換を用いる [3]．字種変換に基づく敵対的攻撃は，日本語に特有の脆弱性を発見できる一方で，摂動の量が限定的であり，攻撃成功率は同義語変換等と比較して高くない [3]．一方で，再攻撃を行う際は，通常事例への悪影響を抑えるために，AE を生成する際と比較して微小な摂動で攻撃を行うことが要求される．すなわち，字種変換にもとづく摂動の弱さは，再攻撃においては通常事例への悪影響を抑えた適切な摂動量となると期待できる．

3.2 提案手法のアルゴリズム

提案手法におけるアルゴリズムの全体像を図 2 に示す．アルゴリズムは再攻撃段階（Phase 1）と検出・矯正段階（Phase 2）により構成される．本手法は，入力された事例に対して再攻撃を行うことにより派生テキストを生成し，それらのラベル変化率によって AE の検出，矯正を試みる．

3.2.1 再攻撃

再攻撃段階では，RS&V[5] と同様，再攻撃により派生テキストを生成する．本手法と RS&V の異なる点として，置換対象となる単語と置換候補となる単語にランダム性がなく，摂動箇所が派生テキストにつき 1 単語であることが挙げられる．

再攻撃の手順について示す．まず，入力テキスト X を分かち書きした単語リスト $\boldsymbol{W} = \{w_1, w_2, \ldots, w_n\}$ を取得する．このとき，生成される派生テキストを選定し効率的な再攻撃を行うために，名詞，動詞，形容詞，形容動詞および副詞以外の単語は摂動を付与しないこととし，$\boldsymbol{W}$ から除外する．続いて，単語 w_i のひらがな，カタカナ，漢字の各表記を取得し，字種変換候補リスト $\boldsymbol{C_i}$ とする．$\boldsymbol{W}$ から i 番目の単語のみを $\boldsymbol{C_i}$ の各要素 C_{ij} で置き換えた派生テキストを生成する．この操作を $\boldsymbol{W}$ の各単語に対して行い，派生テキストの集合 $\boldsymbol{X'}$ を生成する．

表 1 字種変換による再攻撃の例

入力テキスト X	肌触りも良く、薄過ぎず、洗濯もできるので、良かったです。
派生テキスト X'_1	はださわりも良く、薄過ぎず、洗濯もできるので、良かったです。
派生テキスト X'_2	ハダザワリも良く、薄過ぎず、洗濯もできるので、良かったです。
派生テキスト X'_3	肌触りも良く、うす過ぎず、洗濯もできるので、良かったです。
派生テキスト X'_4	肌触りも良く、薄すぎず、洗濯もできるので、良かったです。
派生テキスト X'_5	肌触りも良く、薄過ぎず、せんたくもできるので、良かったです。
派生テキスト X'_6	肌触りも良く、薄過ぎず、センタクもできるので、良かったです。
派生テキスト X'_7	肌触りも良く、薄過ぎず、洗濯もできるので、よかったです。

再攻撃によって生成された派生テキストの例を表 1 に示す．表 1 より，入力テキスト X を元に，X から 1 単語のみ字種変換された派生テキストが複数生成されていることがわかる．

3.2.2 AE 検出及びラベル矯正

AE 検出段階では，再攻撃により生成した派生テキストのラベルを集計し，入力テキストのラベルと異なるラベルの割合，すなわちラベル変化率 R が閾値 ε 以上の入力を AE として検出する．ε は制御パラメータである．ε が低いほど，再攻撃に対して頑健な AE を検出できる反面，頑健性の低い通常事例に対して悪影響を及ぼすこととなる．

続いて，AE として認識された入力事例に対してラベル矯正を行う．二値分類問題では，AE として検出されたテキストのラベルを反転し，それを矯正ラベルとする．本稿における提案手法では二値分類タスクを想定しており，検出と矯正は同じ処理を意味する．多クラス分類問題では，入力事例と異なるラベルのうち，最も多いラベルを採用する．

4 実験

4.1 実験設定

提案手法の有効性を検証するため，通常事例と AE の双方に対して，本手法を適用した．本実験では，楽天市場のデータセット（みんなのレビュー・

口コミ情報）[22] の 2019 年 1 月分のデータ 3,000 事例（Positive:1,500, Negative:1,500）を，二値分類問題として扱うこととした．また，河野らが提案した 3 つの敵対的攻撃手法を用いて生成された AE と通常事例に対して防御実験を行った．すなわち，字種変換攻撃（CharType），同義語置換攻撃（WordNet），字種変換と同義語置換を組み合わせた攻撃（Combi）の 3 種類の攻撃手法を採用した．防御の対象となる識別モデルは，東北大学が公開している事前学習済みの BERT モデル[1)]を，楽天市場データ（2018 年 1 月分）を用いてファインチューニングを行ったモデルとした [3]．ファインチューニング後の BERT モデルに対して，各攻撃を行い生成された AE は付録の表 3 を参照されたい．

本手法の評価には，Clean Acc，Restored Acc，Adv Acc の 3 つの評価指標を用いた．Clean Acc は，通常事例 3,000 事例に対して再攻撃した後の分類精度であり，再攻撃がモデルの正常な分類精度を維持できているかを示す．Restored Acc は，敵対的攻撃に失敗した通常事例と生成された AE，攻撃前からモデルが誤分類していた事例（合計 3,000 事例）に対して再攻撃を行った後の分類結果であり，攻撃によって低下したモデルの分類精度が，再攻撃による防御を導入することでどの程度回復したかを示す [4]．Adv Acc は，AE のみの再攻撃後の分類結果であり，本手法の AE に対するラベル矯正率を示す．Adv Acc において，母数となる事例数が攻撃手法によって異なる点に注意されたい．本実験で生成された AE の事例数は，CharType が 592 事例，WordNet が 940 事例，Combi が 1,722 事例であった．

本手法は入力テキストのラベル変化率に基づいて AE を検出するため，本実験では，閾値を $\varepsilon \in \{0.1, 0.2, 0.3, 0.4\}$ と段階的に変更して防御性能を比較した．0.4 より閾値を大きく設定した際は，各評価指標において，閾値 0.4 までと同様の傾向，すなわち Clean Acc が閾値 0.4 の際と同様の 99.23%であり，Restored Acc と Adv Acc が減少する傾向が観察された程度であったため記載を省略した．

4.2 実験結果

実験結果を表 2 に示す[2)]．提案手法は，通常事例の分類結果を維持し，AE の分類結果の矯正に成功した．ε が小さい場合は AE の検出と矯正がより積極的に行われたことがわかる．ε を 0.1 とした場合は，どの攻撃手法でもモデルの分類精度が 90%以上に回復した．本手法は，字種変換を再攻撃として使用しており，同様に字種変換で生成された AE の矯正率が高かった．加えて，再攻撃とは異なる種類の摂動を用いている WordNet による同義語置換で生成された AE の結果に着目すると，防御を行わない場合は WordNet を用いた同義語置換攻撃によって正解率が 67.90%に低下していたが，$\varepsilon = 0.1$ と設定した提案手法を導入することで正解率が 94.43%に改善したことがわかる．Adv acc に着目すると，上記の場合に，AE のうち 84.26%を正しく矯正できたことが確認できる．さらに，字種変換と同義語置換を併用した攻撃手法である Combi は，防御を行わない場合の正解率を 41.83%まで低下させる攻撃であったが，提案手法の導入により，Combi に対するモデルの分類精度が 50%程度改善されたことがわかる．

表 2 再攻撃による防御性能 [%]

閾値 ε	Clean Acc	AE 生成時の攻撃手法					
		Restored Acc			Adv Acc		
		CharType	WordNet	Combi	CharType	WordNet	Combi
N/D	99.23	79.50	67.90	41.83	0.00	0.00	0.00
0.1	98.27	97.17	94.43	91.47	87.67	84.26	85.83
0.2	98.83	94.33	91.17	82.63	74.16	70.43	70.73
0.3	99.13	91.37	85.17	74.73	59.46	54.68	57.08
0.4	99.23	88.13	80.70	67.13	43.24	40.53	43.90

一方，ε を 0.4 まで高めた場合，提案手法を用いない場合と同程度に Clean Acc を維持できたものの，Adv Acc は 40%程度まで低下した．これは，ラベル変化率が小さい AE は，決定境界から離れていることが予想され，微小な摂動による再攻撃に対しても頑健であったために AE として検知されなかったためと考える．そのため，閾値を上げるごとに検出される AE が少なくなり，Adv Acc が低下したと考えられる．

5 おわりに

本研究では，AE に対する日本語処理モデルの頑健性向上を目的とし，日本語に特化した再攻撃による敵対的防御手法を提案した．提案手法は，AE の脆弱性に基づいており，実験結果から，通常事例への悪影響を抑えつつ，AE のラベルのみ矯正でき，日本語処理モデルの頑健性の向上が可能であることを示した．今後，再攻撃のアルゴリズムや摂動量を改善し，通常事例の分類精度を維持しつつ，モデルの回復性能の向上を図る．

1） https://huggingface.co/daigo/ bert-base-japanese-sentiment

2） N/D（Non-Defense）は再攻撃前のモデルの分類精度を表しており，N/D の場合の Restored Acc は各攻撃手法の攻撃性能を示している．

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] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. **CoRR**, Vol. abs/1312.6199, , 2013.

[2] 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.

[3] 河野竜士, 小野智司. Anjel: 日本語を扱う深層ニューラルネットワークを対象としたブラックボックス敵対的攻撃. 人工知能学会論文誌, Vol. 40, No. 2, 2025.（in press）.

[4] Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. Frequency-guided word substitutions for detecting textual adversarial examples. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 171–186. Association for Computational Linguistics, 2021.

[5] Xiaosen Wang, Xiong Yifeng, and Kun He. Detecting textual adversarial examples through randomized substitution and vote. In **Uncertainty in Artificial Intelligence**, pp. 2056–2065. PMLR, 2022.

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

[7] 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.

[8] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In **2016 IEEE European symposium on security and privacy (EuroS&P)**, pp. 372–387. IEEE, 2016.

[9] Shilin Qiu, Qihe Liu, Shijie Zhou, and Wen Huang. Adversarial attack and defense technologies in natural language processing: A survey. **Neurocomputing**, Vol. 492, pp. 278–307, 2022.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6174–6181. Association for Computational Linguistics, 2020.

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

[16] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. **arXiv preprint arXiv:1804.06059**, 2018.

[17] Shreya Goyal, Sumanth Doddapaneni, Mitesh M Khapra, and Balaraman Ravindran. A survey of adversarial defenses and robustness in nlp. **ACM Computing Surveys**, Vol. 55, No. 14s, pp. 1–39, 2023.

[18] Xiaosen Wang, Yichen Yang, Yihe Deng, and Kun He. Adversarial training with fast gradient projection method against synonym substitution based text attacks. **Proceedings of the AAAI conference on artificial intelligence**, Vol. 35, No. 16, pp. 13997–14005, 2021.

[19] Zhe Zhao, Guangke Chen, Jingyi Wang, Yiwei Yang, Fu Song, and Jun Sun. Attack as defense: characterizing adversarial examples using robustness. In **Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis**, ISSTA 2021, pp. 42–55. Association for Computing Machinery, 2021.

[20] 森本文哉, 赤垣敬吾, 小野智司. 敵対的事例の脆弱性を用いた分類結果矯正の試み. 人工知能学会全国大会論文集 第 37 回 (2023), pp. 2K5GS201–2K5GS201. 一般社団法人 人工知能学会, 2023.

[21] 森本文哉, 小野智司. 微小な再攻撃による敵対的事例の矯正に関する基礎検討. 第 86 回全国大会講演論文集, Vol. 2024, No. 1, pp. 517–518, 03 2024.

[22] 楽天グループ株式会社：楽天市場データ. 国立情報学研究所情報学研究データリポジトリ. 2020.

A　付録

A.1　敵対的攻撃により生成された AE

本実験で使用した 3 種類の敵対的攻撃手法によって生成された AE を表 3 に示す．赤字は摂動が付与されて変更された単語を示す．CharType では，「値段」をカタカナ表記の「ネダン」に字種変換することで分類結果が Negative から Positive に変化した．また，WordNet では，「少し」を「若干」に置換することで，文章全体の意味を変えずに誤認識を生じさせていることが確認できる．Combi は，字種変換と同義語置換を組み合わせた攻撃であり，例では，「円」を「エン」に，「ちゃんと」を「ぴったり」に変換することで，Negative な文章を Positive な文章と誤認識した．

表 3　生成された AE の例

攻撃手法	テキスト	分類結果 / 信頼度
攻撃前	実物を手にとってみて、実感、お値段が高いですね。	Negative / 99.9%
CharType	実物を手にとってみて、実感、おネダンが高いですね。	Positive / 99.0%
攻撃前	やはり訳アリだ、も少し良くならないかね。	Negative / 87.6%
WordNet	やはり訳アリだ、も若干良くならないかね。	Positive / 83.8%
攻撃前	2000 円足してちゃんとしたものを買うべきと猛省	Negative / 99.9%
Combi	2000 エン足してぴったりしたものを買うべきと猛省	Positive / 89.4%