文脈指定可能な偽レストランレビューの自動生成とその対策法

代表研究者[1]　　孫　博　　早稲田大学　助手

1 はじめに※

ユーザーによって生成されたコンテンツ[6]は、ウェブ上の現代的なユーザーエクスペリエンスとして必要不可欠な部分である。 tripadvisor.com、yelp.com、Google Play などのサイトでは、ユーザーが作成したレビューを使用して、他のユーザーがお金と時間を費やす場所を選択するのに役立つ豊富な情報を提供する。ユーザーレビューは、サービスや製品の評価、および建設的な意見の提供に良く使用される。ユーザーのレビューとレーティングは、おすすめのサービスをランク付けするために利用されている。レーティングはビジネスを大きく左右することがある。yelp.com では、すでに 8 年前に、1 つ星のレーティングの上昇がビジネスの収益に 5〜9%の影響を及ぼすと推定している[7]。

ユーザー生成コンテンツの金銭的な影響のため、金銭的報酬と引き換えに労働者によって書かれた肯定的な評価を顧客に提供することを約束する、いわゆるクラウドターフエージェント[8]に頼っている企業もある。たとえば、Amazon コミュニティガイドラインに関しては、プロモーションに関連するコンテンツの購入は禁止されているが、作成されたコンテンツは違法とは見なされず、偽コンテンツの作成者も普通のユーザーと認識されている[9]。2015 年に、yelp.com のオンラインレビューの約 20%が偽造であると疑われていた[10]。

現在、yelp.com のようなユーザー作成のレビューサイトでは、フィルタと不正なレビュー検出技術が使用されている。これらの要因により、レビューサイトに提供されるレビューの要件が高まり、その結果、質の高いレビューのコストも増加している。コストが増加したため、既存研究はニューラルネットワークによって生成された偽のレビューの存在を仮定した。これらのニューラルネットワークベースの偽のレビューは、人間が書いた偽のレビューとは統計的に異なり、これらについて訓練された分類器によっては捉えられない[1]。

偽のレビューの検出は、個人のレベルでも、システム全体の検出ツール（つまり、規制）としても実行できる。個人レベルで偽のオンラインコンテンツを検出するには、重要な読書に関する知識とスキルが必要不可欠である。2017 年に、National Literacy Trust は、英国の若い人々は偽のニュースを本当のニュースと区別するためのスキルを持っていないと評価した[11]。たとえば、12〜15 歳の年齢層のオンラインニュースサイトを利用する子供たちの 20%は、ニュースサイトに関するすべての情報が正しいと考えている。

自動的に生成された偽のレビューはごく最近になって人間の読者をだますのに十分な自然さになってきた。Yao ら[1]ディープニューラルネットワーク（いわゆる 2 層 LSTM [2]）を使用することによって、偽のレビューを自動的に且つ大量に生成し、これらの偽のレビューは英語がネイティブである人を欺くために十分に本物に見えると結論付けた。Yao らは yelp.com [3]からの本当のレストランレビューを使って提案したモデルを訓練する。訓練されたモデルは、1 文字ずつレビューを生成するために使用される。すべての文字がランダムに生成されるため、特定のコンテキスト（意味のある側路情報）を簡単にターゲットにすることはできない。その結果、生成されたレビューの内容が特定の話題から外れる可能性は高い。たとえば、ラスベガスで和食レストランのレビューを生成する場合、レビュー生成プロセスにはボルチモアのイタリアンレストランへの参照が含まれる恐れがある。 [1]の著者たちは、食物関連の単語をより適切な単語（対象となるレストランから抽出）に置き換える後処理ステップ（カスタマイズ）を適用する。単語置換手法には欠点が存在する。周囲の単語とは無関係に、特定の単語を見逃したり、他の単語を置き換えたりする可能性があるため、知識のある読者に判別される可能性が高い。例として、[1]で説明されたカスタマイズ手法を日本のレストランのレビューに適用した場合、朝食用のスニペットガーリックノットから寿司用のガーリックノットに変換された。

そこで、本研究では、生成された各偽のレビューに対して文脈を定義することによって生成プロセスを改善するニューラルマシントランスレーション（ＮＭＴ）に基づく手法を提案する。本研究で用いたコンテキストは、レビューの評価、レストラン名、市区町村、都道府県名、食べ物のタグ（例：日本語、イタリア語）

[1]共同研究のため、孫博が代表として取り込んだ部分は“※”によってマークされている。

の平文のシーケンスである。提案のテクニックはトピックにとどまるレビューを生成できることを示した。提案モデルに使われている基本的なテクニックをいくつかの異なるモデルにインスタンス化することができる。 Amazon Mechanical Turk で実証実験を行ったところ、英語を母国語とするネイティブスピーカーは偽のレビューを認識するのが非常に苦手であることが判明された。 1 つのモデルでは、参加者のパフォーマンスはランダムに近く、各クラスにおける検知スコアは 47%である（テストの 1：6 の不均衡を考えるとランダムは 42%である)。経験豊富且つ高度な教育を受けた参加者によるユーザースタディを通して、この変種（以降、NMT-偽レビューと呼ぶ）を[1]の char-LSTM ベースの手法を使って生成された偽のレビューと比較した。結果としては、NMT-偽レビューはこれまでに知られている偽レビューのみを使用して訓練された自動分類器では検出できない新しい種類の偽レビューを生み出せることを証明した[1,4,5]。したがって、NMT-偽レビューは既存のオンラインレビューサイトでは検出されない可能性が高い。この課題を解決するために、NMT-偽レビューを高精度で検出する有効な分類器を開発した（F スコアは 97%である)。

2 システムモデル

2-1 攻撃モデル※

Wang ら[8]は、特定のプラットフォーム（例：Yelp）で特定のターゲット（例：レストラン）に対して偽のレビューを希望する顧客、顧客に偽のレビューサービスを提供するエージェント、偽のレビューを作成して投稿するようにエージェントによって調整された作業者の 3 つのエンティティから構成されるクラウドターフ攻撃のモデルを説明した。

自動クラウドターフ攻撃（ACA）は、生成モデルによって雇われた人間を置き換える。これには、非常に良い経済性、スケーラビリティ（人間の作業員の方はより高価でより遅い)、検知される可能性の削減（エージェントが偽のレビューを自動的に生成し、投稿される割合をより適切に制御できる）などいくつかの利点がある。エージェントはレビュープラットフォーム上でパブリックレビューにアクセスでき、それによって生成モデルをトレーニングできると仮定する。また、エージェントがレビュープラットフォーム上で多数のアカウントを作成するのは簡単であると想定しているため、アカウントベースの検出またはレート制限の手法は偽のレビューに対して効果がない。

生成モデルの品質は、攻撃において重要な役割を果たす。Yao ら[1]は、生成モデルを基礎として文字ベースの LSTM の使用を提案している。 LSTM は特定のターゲットに対するレビューを生成するようには調整されておらず[2]、ランダム方式の生成中に異なるコンテキストからの概念を混同する可能性がある。文脈上別々の単語を混在させることは、人間が偽のレビューを識別するために使用する重要な基準の 1 つである。これらは、偽のコンテンツの既知の指標に違反する可能性がある[19]。たとえば、レビューの内容が以前の予想や読者の持つ情報のニーズと一致しない場合がある。より適切なレビューを生成する、より有能な生成モデルであるニューラルマシントランスレーション（NMT）モデルを検討することで、攻撃モデルを改善することが期待できる。

2-2 生成モデル

（1）構造

偽のレビュー生成のための NMT モデルの使用を提案する。この方法はいくつかの利点を有する：1）文脈（キーワード）をレビューに関連付ける方法を学習する能力、2）速いトレーニング時間、および3）制作時間中の高度のカスタマイズ（レビューへの特定のウェイターや食品の名前などを導入する）

NMT モデルは、構造としてスタック型リカレントニューラルネットワーク（RNN）である。それらはエンコーダネットワークおよびデコーダネットワークを含み、これらは 1 つのシーケンスから別のシーケンスへの変換を生成するために共同で最適化されている。エンコーダは入力データを順番にロールオーバーして、文の 1 つの n 次元コンテキストベクトル表現を生成する。次に、デコーダは、埋め込みベクトルと、出力ワードを特定の入力ワードと関連付けるように教示されているアテンションモジュールに基づいて出力シーケンスを生成する。生成は通常、指定される EOS（文末）トークンに遭遇するまで続く。レビューの長さはさまざまな方法で制御できる。必要な長さに達するまで EOS トークンを生成する確率をゼロに設定するなどの方法がある。

NMT モデルはしばしばビームサーチ[15]を含み、それはいくつかの仮説を生成し、それらの中から最良の

ものを選択できる。本研究の目的では、欲張りビーム探索法を使う。出力の品質はすでに十分であり、並進フェーズの時間消費は使用される各ビームに対して直線的に増加することを確認したので、追加のビーム検索の使用を必要としない。

（２）データセット※

偽のレビューを生成するために Yelp Challenge データセット[3]を使用します。 データセット（2017 年 8 月）には、290 万件の 1〜5 つ星レストランのレビューが含まれている。機械によるレビュー攻撃の大規模な展開はまだ報告されていないため（2017 年 9 月）[20]、我々は、すべてのレビューを本物の人間が作成されるレビューとして取り扱う。 前処理として、処理不可能な（ASCII 以外の）文字と余分な空白を削除しておく。 句読点と単語の区別を考慮する。 訓練のために 15,000 のレビューを、テストのために 3,000 のレビューを配分し、残りはトレーニングに使用する。 ＮＭＴモデルは、原文と訳文の並列コーパス、すなわち（原文、訳文）ペアの大きな集合を必要とする。 データセットから（コンテキスト、レビュー）ペアを構築することによって、並列コーパスを構築する。 次に、入力コンテキストの作成方法について説明する。

（３）文脈

Yelp Challenge データセットには、レストランの名前、食べ物のタグ、都市、州など、レストランに関するメタデータが含まれている。各レストランのレビューでは、このメタデータを取得して NMT モデルの入力コンテキストとして使用する。対応するレストランレビューも同様に対象文として設定される。 この方法により、対訳コーパスに 290 万のペアが生成された。 以下の例 1 のパラレルトのレーニングコーパスの一例を示す。

Example 1.

5 Public House Las Vegas NV Gastropubs Restaurants > Excellent food and service . Pricey , but well worth it . I would recommend the bone marrow and sampler platter for appetizers

[rating name city state tags]の順序は一定に保たれている。 モデルを訓練することは、入力文中の単語の特定シーケンスを出力中の他の単語と関連付けるようにモデル化する。

（４）訓練用の設定※

本研究では、32GB の RAM、1 つの NVidia GeForce GTX 980 GPU、i7－4790k CPU（4.00GHz）を搭載した市販の PC で NMT モデルをトレーニングする。使用端末は、約 1,300〜1,500 のソーストークンと約 5,730〜5,830 の出力トークンを処理できる。1 エポックのトレーニングには平均 72 分掛かる。モデルは８つのエポック､すなわち一晩訓練される。このモデルによって生成された偽のレビューを NMT-偽レビューと呼ぶ。異なる評価のレビューを作成するために 1 つのモデルをトレーニングするだけで済む。トレーニング設定として Adam オプティマイザーを使用する[13]。学習率は推奨の 0.001 で設定する[15]。 ほとんどのパラメータはデフォルト値である。 特に、入力と出力の最大センテンス長はデフォルトで 50 トークンである。 本研究では NMT モデルとして広く使われているフレームワーク openNMT-py [15]を利用する。

2-3 偽レビュー生成の制御

図 1 の例 2 では、特定の文脈（例：すばらしい、食べ物、サービス、ビール、選択、例 1 の場合はハンバーガー）に対して与えられた一般的な単語が繰り返されている。 ジェネリックレビュー生成は、ジェネレータＬＭ、デコーダの確率（対数尤度［２］）を減少させることによって回避することができる。 ランダムに単語にペナルティを課すことによって文の生成を制限する。我々はいくつかの形式の付加されたランダム性を試みたが、ターゲット単語のランダムなサブセットに一定のペナルティを加えることが最も自然な文の流れをもたらすことを見出した。 確率変数は 1 または 0（オンまたはオフ）として選択されるため、これらのペナルティをベルヌーイペナルティと呼ぶ。

Example 2. Greedy NMT

Great food, great service, great *beer selection*. I had the *Gastropubs burger* and it was delicious. The *beer selection* was also great.

Example 3. NMT-Fake*

I love this restaurant. Great food, great service. It's *a little pricy* but worth it for the *quality* of the *beer* and atmosphere you can see in *Vegas*

図 1

Algorithm 1 Generation of NMT-Fake* reviews.

Data: Desired review context C_{input} (given as cleartext), NMT model
Result: Generated review *out* for input context C_{input}
set $b = 0.3$, $\lambda = -5$, $\alpha = \frac{2}{3}$, p_{typo}, p_{spell}
$\log p \leftarrow$ NMT.decode(NMT.encode(C_{input}))
out $\leftarrow$ [ ]
$i \leftarrow 0$
$\log p \leftarrow$ Augment($\log p$, b, λ, 1, [ ], 0) — random penalty
while $i = 0$ *or* o_i *not EOS* **do**
  $\log \tilde{p} \leftarrow$ Augment($\log p$, b, λ, α, o_i, i) — start & memory penalty
  $o_i \leftarrow$ NMT.beam($\log \tilde{p}$, out)
  out.append(o_i)
  $i \leftarrow i + 1$
end
return Obfuscate(out, p_{typo}, p_{spell})

一般的な文の構成要素を避けるために、デコーダのデフォルトの言語モデル p（・）を次のように拡張する。

$$\log p\tilde{}(t_k) = \log p(t_k|t_i, \ldots, t1) + \lambda q, \quad (3)$$

ここで、q∈RV は、値が 1 である確率 b と値が 0 である確率 1−bi を得るベルヌーイ分布ランダム値のベクトルで、λ<0 となる。パラメータ b は、ボキャブラリのどれだけを忘れるかを制御できる。 レビューに「忘れられた」単語を含める。 λqk は、ペナルティのない単語を使用した文形成を強調する。ランダム性は新しいレビューの生成の開始時にリセットされる。 言語モデルで Bernoulli のペナルティを使用すると、特定の割合の単語を「忘れる」ことができ、基本的にそれほど典型的でない文の作成を「強制する」ことができる。 第 2 節では、これら 2 つのパラメータ、ベルヌーイ確率 b と「忘れられた」単語 λ を含む対数尤度ペナルティの効果をテストする。

一般的な文章の開始を回避するために開始ペナルティを導入する（例：「最高の食事、最高のサービス」）。[18]に触発されて、我々は我々の言語モデルにランダムな開始ペナルティ λsi を追加し、それは生成された各トークンに対して単調に減少する。 生成される 5 語ごとに効果が 90%減少するので、α←0.66 に設定する。ベルヌーイの罰則は、文中の特定の単語を過度的に使用することを防止しない(例 2 の great のように)。単語の過剰な再利用を避けるために、各翻訳で以前に使用された単語に対するメモリペナルティを含めた。具体的には、欲張り検索によって生成された各単語にペナルティ λ を追加する。

NMT モデルにこれらのペナルティを適用した後、レビューを視覚的に分析した。 モデルは明らかに多様であるが、それらはインコヒーレントであった：ランダムなペナルティの導入は文の文法性を低下させた。 とりわけ、句読点の使用は不規則であり、代名詞は意味的に誤って使用されていた（例えば、「および」／「しかし」のようにこれらは置き換えられるかもしれない）。 レビューの信頼性を高めるために、文法ベースのルールをいくつか追加してみた。

英語には、文の自然な流れにとって重要な単語のクラスがいくつかある。一般的な代名詞（例えば、私、それら、私たちの）、接続詞（そして、したがって、のように）、句読点（例えば、/など）のリストを作成し、これらの単語に対して半分だけのメモリペナルティを適用する。この変更によりレビューの一貫性を高めることができた。 このステップと前のステップの疑似コードをアルゴリズム 2 に示す。文法ベースの規則と LM 拡張の組み合わせ効果は、図 1 の例 3 に示されている。

本研究の NMT モデルは文法ミスなしでレビューを生成することを目指す。これは実際の人間の作業者とは

異なり、文章は 2 つのタイプの言語の間違いを含む。1）人間の入力の間違いによって引き起こされたタイプミス、および 2）一般的な綴りの間違い。 オックスフォードの辞書[26]から一般的な英語のスペルミスのリストを削り取り、スペルミスをランダムに再導入するための 80 のルールを作成した。同様に、実際の英単語に小さな摂動を加えた誤字が強調されるように、誤字は weighted edit distance[27]に基づいてランダムに再導入される。これらの単語を見つけるための自動修正ツール[28]を使用している。 人間がそれらを書いたと思うように読者を混乱させることを目的としているので、我々はこれらの増強を難読化と呼ぶ。 簡潔にするために、擬似コードの説明は省略している。

Algorithm 2 Pseudocode for augmenting language model.

Data: Initial log LM $\log p$, Bernoulli probability b, soft-penalty λ, monotonic factor α, last generated token o_i, grammar rules set G
Result: Augmented log LM $\log \tilde{p}$

procedure AUGMENT($\log p$, b, λ, α, o_i, i)
  generate $P_{1:N} \leftarrow Bernoulli(b)$ — One value $\in \{0, 1\}$ per token
  $I \leftarrow P > 0$ — Select positive indices
  $\log \tilde{p} \leftarrow$ Discount($\log p$, I, $\lambda \cdot \alpha^i$,G) — start penalty
  $\log \tilde{p} \leftarrow$ Discount($\log \tilde{p}$, $[o_i]$, λ,G) — memory penalty
  return $\log \tilde{p}$
end procedure

procedure DISCOUNT($\log p$, I, λ, G)
  for $i \in I$ **do**
    if $o_i \in G$ **then**
      $\log p_i \leftarrow \log p_i + \lambda/2$
    else
      $\log p_i \leftarrow \log p_i + \lambda$
    end
  end
  return $\log p$
end procedure

2-4 実験；提案モデルにおけるパラメータ※

パラメータ b と λ は、偽レビューのさまざまな内容を制御する。 生成された偽のレビューの 6 つの異なる例を表 1 に示す。ここで、最大の違いは b の値が大きくなるにつれて発生する。明らかに、レストランのレビューはより極端になる。 これは、語彙の大部分が「忘れられている」ために起きている。b が 0.7 であるレビューには、もっとまれな単語の組み合わせが含まれている。例えば、「!!!!!」を句読点として認識されるとか、そして偶に文章の文法を破るとか（「経験は最高であった」）。低い b を持つレビューはより一般的である。それらは多くのレビューで発生する“素晴らしい場所、良いサービス”のような安全な単語の組み合わせが含まれる。パラメータ λ はもっと微妙である：それはランダムレビューの開始がある程度、そしてレビュー内のステートメント間の中断にある程度影響を与える。 どのような種類の NMT-偽レビューがネイティブスピーカーに説得力があるかを判断するために、Amazon Mechanical Turk（MTurk）調査を実施した。 次のセクションで、調査と結果について説明する。

表 1

(b, λ)	Example review for context
(0.3, −3)	I love this location! Great service, great food and the best drinks in Scottsdale. The staff is very friendly and always remembers u when we come in
(0.3, −5)	Love love the food here! I always go for lunch. They have a great menu and they make it fresh to order. Great place, good service and nice staff
(0.5, −4)	I love their chicken lettuce wraps and fried rice!! The service is good, they are always so polite. They have great happy hour specials and they have a lot of options.
(0.7, −3)	Great place to go with friends! They always make sure your dining experience was awesome.

(0.7, −5)	Still haven't ordered an entree before but today we tried them once.. both of us love this restaurant....
(0.9, −4)	AMAZING!!!!! Food was awesome with excellent service. Loved the lettuce wraps. Great drinks and wine! Can't wait to go back so soon!!

検知率のFスコアに関しては、レビューの平均はわずか56%であり、その中に偽のレビューは53%で、実際のレビューは59%である。結果としては、精度、再現率、およびFスコアがそれぞれ50%になっており、ランダム検出に非常に近づいている。具体的な結果は表2に示されている。カテゴリ全体での人の検出率はランダムに近いため、全体として、偽のレビュー生成は非常に成功している。

表2

Classification report				
Review Type	Precision	Recall	F-score	Support
Human	55%	63%	59%	994
NMT-Fake	57%	50%	53%	1006

異なる偽のレビューカテゴリの検出には多少のばらつきがあることを確認できた。 MTurk 調査の回答者は、真の陽性率が 40.4%であるのに対し、本物のレビューであるクラスの真の陰性率が 62.7%であるカテゴリ(b = 0.3、λ = 5) のレビューを認識するのが最も困難であった。精度はそれぞれ 16%と 86%であった。クラス平均Fスコアは 47.6%で、これはランダムに近い。通常の英語を母国語とする人は、実際のレビューと偽のレビューを区別することが非常に困難であるため、私たちの MTurk による研究では、NMT-偽のレビューがレビューシステムに大きな脅威をもたらすことが示唆されている。 MTurk の参加者はこれらのレビューを検出するのが最も困難であったため、このホワイトペーパーでは今後のユーザーテストにレビューカテゴリ (b = 0.3、λ = 5) を使用する。この研究では、このカテゴリを NMT-Fake *と呼ぶ。

3 評価

本研究は最初にそれらを以前に提案されたタイプの偽レビューと統計的に比較することによって提案されたモデルで生成される偽レビューを評価し、そして経験豊富な参加者によるユーザー調査を実施する。 前のタイプを検出し、分類性能を調査するために分類器をトレーニングすることによって、既存の偽のレビュータイプ[1,4,5]との統計的な違いを示す。

3-1 既存研究の再現※

Yao ら[1]は、偽のレビューに対する現在の最先端の生成モデルを提案した。 このモデルは、2 層文字ベースの LSTM モデルを使用して Yelp Challenge データセットについてトレーニングされている。 既存研究のLSTMモデル又はLSTMモデルによって生成された偽のレビューデータセットへのアクセスを[1]の著者たちに要求した。 残念ながら、既存研究の著者たちはこれらのどちらも共有してくれることができなかった。したがって、著者たちと電子メールのやりとりに基づいて、既存研究のモデルをできるだけ厳密に再現した。

本研究では同じグラフィックスカード (GeForce GTX) を使い、同じフレームワーク (Lua の torch-RNN) を使ってトレーニングした。 Yelp Challenge からレビューをダウンロードし、処理可能な ASCII 文字のみを含むようにデータを前処理し、レストラン以外のレビューを除外した。 モデルを約 72 時間訓練した。 [1] に言及されているカスタマイズ方法と電子メールによるやり取りを使用してレビューを後処理した。このモデルによって生成された偽のレビューを LSTM-偽レビューと呼ぶ。

3-2 既存偽レビューとの類似性※

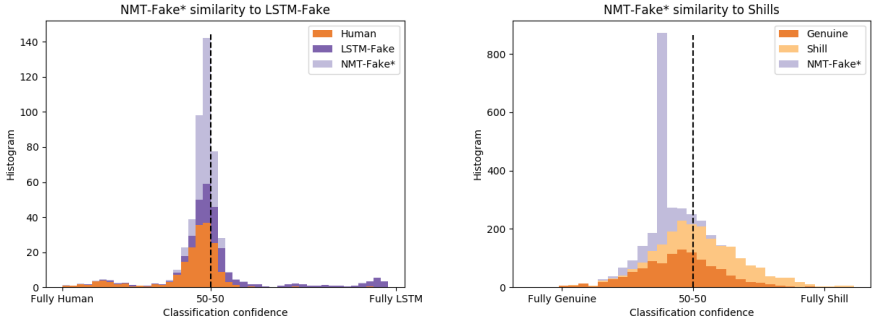
ここで、NMT-Fake *レビューが a) LSTM の偽レビューおよび b) 人為的な偽のレビューとどのような差分が存在するのかを究明する。これらのクラス間の統計的類似性を比較することによって実証実験を行う。

a (図 2a) には、Yelp Challenge データセットを使用する。 Yelp Challenge データセット (“ human”) からのランダムレビュー5,000件とLSTM-Fakeによって生成された偽のレビュー5,000件を使用して分類器をト

レーニングした。Yao ら[1]は、LSTM-偽レビューを識別するために文字の機能が必要不可欠であることを発見した。そのため、文字の特徴を使用する（最大 3 グラム）。
「b」（図 2b）の場合、「Yelp Shills」データセット（YelpZip [4]、YelpNYC [4]、YelpChi [5]の組み合わせ）である。このデータセットは、Yelp のフィルタリングメカニズムによって不正と識別されたエントリにラベルを付ける（ ″shill reviews″）。残りは人間のユーザーにより作成された本物のレビュー（「本物」）として扱われる。分類器を訓練するために、各カテゴリから 10 万件のレビューを使用する。生成された特徴には市販の心理測定ツール LIWC2015 [21]の特徴を採用する。
どちらの場合も、トレーニングには AdaBoost（200 本の浅い決定木を使用）を利用する。各分類子をテストするために、各ケースで両方のクラスから 1,000 件のレビューをランダムに選択し、保留されたテストセットを使用する。さらに、1,000 個の NMT-Fake *レビューをテストする。実験の結果を図 2a と図 2b に示す。50%の分類しきい値は破線でマークされている。
新しく生成されたレビューは以前の既知の種類の偽レビューと強い属性を共有していないことが確認された。どちらかといえば、本研究で提案されたモデルによる偽レビューは以前の偽のレビューより本物のレビューに似ていると言える。そのため、NMT-偽 *の偽のレビューは、オンラインのレビューサイトでは検出されない偽のレビューのカテゴリを表していると推測される。

3-3 ユーザースタディ

本研究では、機械で生成された偽のレビューを理解していると期待している、技術に詳しいユーザーに対して偽のレビューの有効性を評価する。著者たちは 20 人の参加者を持つユーザースタディを行った。全員がコンピュータサイエンス教育と少なくとも 1 つの大学の学位を持っていることを条件とした。各参加者はまずトレーニングセッションに参加し、そこでレビューにラベルを付ける（偽または本物）。その後、それらを集計して正しい回答と比較した。これらの参加者を経験豊富な参加者と呼ぶ。ユーザー調査中に個人データは収集されなかった。



(a) Human–LSTM reviews.　　(b) Genuine–Shill reviews.

図 2

各人にランダムに選択された 30 のレビュー（1 人あたり合計 60 のレビュー）のセットが 2 つずつ与えられ、レビューにはそれぞれ 10〜50 語が含まれている。各セットには Yelp からの 26（87%）の実際のレビューと 4（13%）の機械生成レビューが含まれており、その数は Yelp での疑わしいレビューの流行に基づいて決定したものである[4,5]。 1 つのセットは、2 つのモデルのうちの 1 つ（NMT（b = 0.3、λ = 5）または LSTM）からの機械生成レビューを含み、他のセットはランダムで他からレビューを選んだ。偽のレビューの数は、調査説明の際に各参加者に明らかにした。各参加者は、4 つのレビューに偽のマークを付けるよう求められた。
各レビューは実際のレストランを対象としていた。調査の前に、そのレストランの Yelp ページのスクリーンショットが各参加者に提示された。各参加者は、1 つの特定のランダムに選択されたレストランのレビューを評価した。
図 3 は、両方のタイプの検出されたレビューの分布を示している。比較のために仮定のランダム検出器も示した。 NMT-偽 *のレビューは、経験豊富な参加者にとっては検出がかなり困難である。平均して、検出率（再現率）は NMT-偽 *レビューで 20%となっているが、LSTM ベースのレビューでは 61%である。参加者は

30 のレビューの各セットで 4 つの偽物をラベル付けしたので、本研究において精度（そして F スコア）は再現性と同じである[2]。参加者間の検出分布は図 3 に示されている。差は 99%の信頼水準で統計的に有意である(Welch の t 検定)。NMT-Fake *レビューの検出率とランダム検出器を比較したところ、参加者の NMT-Fake *レビューの検出率は、95%の信頼度でのランダム予測と統計的に異ならないことが示唆された（Welch の t 検定)。

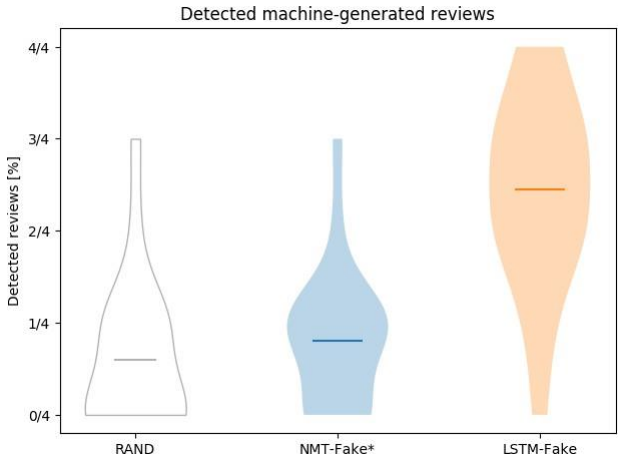


図 3

4 防御方法※

AdaBoost ベースの分類器を開発して、200 個の浅い決定木（深さ 2）からなる新しい偽のレビューを検出することができた。SpaCy トークン化に基づく単語レベルの機能を使用した[22]、そして POS タグと依存関係ツリータグの n-gram 表現による特徴を構築した。 NLTK から読みやすさの特徴も追加した[23]。

図 4 は、さまざまな種類の偽のレビューを検出したときの AdaBoost 分類器のクラス平均 F スコアを示している。分類器は、人間が検出するのが難しいレビューを検出するのに非常に有効である。たとえば、MTurk ユーザーが最も検出困難（b = 0.3、λ= 5）の偽のレビューは、97%という優れた F スコアで検出されることを確認した。分類のための最も重要な特徴は、読みやすさの特徴「Automated Readability Index」と同様に、偽のレビューで頻繁に出現する単語の数え上げ（句読点、代名詞、記事など）であった。したがって、NMT-Fake のレビューは人間にとって検出が困難であるが、適切なツールを使用して検出できることが証明された。

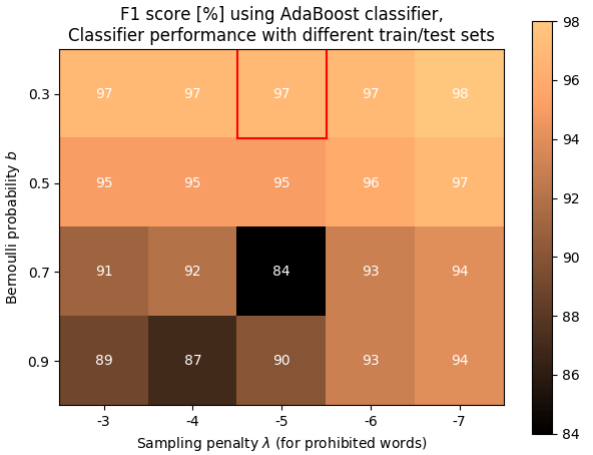


図 4

5 関連研究

Kumar and Shah [24]は、虚偽の情報調査を調査し、それらを分類している。自動的に生成された偽のレビューは、意見に基づく虚偽の情報の一種で、レビューの作成者が読者の意見や決定に影響を与える可能性がある。Yao ら[1]は、機械が生成した偽のレビューに関する研究を発表した。本研究と異なって、既存研究では生成前に特定の文脈を指定せずに、文字レベルの言語モデルを提案した。それに対して本研究では、レビューを生成する前に、既存の NMT ツールを利用してレストランの特定のコンテキストをエンコードする。本研究背景の根拠として、Everett ら[25]は、セキュリティ研究者が一般のインターネットユーザーと比較してマルコフ連鎖によって生成された Reddit コメントによってだまされる可能性が低いことを発見した。NMT モデル出力の多様化は[18]で研究されている。著者らは、最大の相互情報ベースの生成を強調するために、一般的に発生する文（n グラム）にペナルティを使用することを提案した。本研究では、チャットボットシステムにおける NMT モデルの使用について調査した。ランダムトークンに対するユニグラムのペナルティ(アルゴリズム 2）は、実装が簡単で、十分に多様な効果が出ることがわかった。

6 結論※

本研究では、経験豊富な技術に精通したユーザーでさえ欺くのに非常に効果的な偽のレビューを生成するために、ニューラル機械翻訳モデルを使用できることを検証した。これは事例証拠[11]をサポートしている。我々の技術は最先端技術より有効的である[1]。 人間のユーザーは偽のレビューを識別するのにあまり効果がないので、機械による偽のレビューの検出が必要であると結論が下せた。また、ある種類の偽のレビューを使用して訓練された検出器は、他の種類の偽のレビューを識別するのに有効ではないことも示した。 したがって、偽のレビューをロバストに検出することは今後の課題とする。

【参考文献】

1. Yao, Y., Viswanath, B., Cryan, J., Zheng, H., Zhao, B.Y.: Automated crowdturfing attacks and defenses in online review systems. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, ACM (2017)
2. Murphy, K.: Machine learning: a probabilistic approach. Massachusetts Institute of Technology (2012)
3. Yelp: Yelp Challenge Dataset (2013)
4. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: What yelp fake review filter might be doing? In: Seventh International AAAI Conference on Weblogs and Social Media (ICWSM). (2013)
5. Rayana, S., Akoglu, L.: Collective opinion spam detection: Bridging review net- works and metadata. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
6. O'Connor, P.: User-generated content and travel: A case study on Tripadvisor.com. Information and communication technologies in tourism 2008 (2008)
7. Luca, M.: Reviews, Reputation, and Revenue: The Case of Yelp. com. Harvard Business School (2010)
8. Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H., Zhao, B.Y.: Serf and turf: crowdturfing for fun and profit. In: Proceedings of the 21st international conference on World Wide Web (WWW), ACM (2012)
9. Rinta-Kahila, T., Soliman, W.: Understanding crowdturfing: The different ethical logics behind the clandestine industry of deception. In: ECIS 2017: Proceedings of the 25th European Conference on Information Systems. (2017)
10. Luca, M., Zervas, G.: Fake it till you make it: Reputation, competition, and yelp review fraud. Management Science (2016)

11. National Literacy Trust: Commission on fake news and the teaching of critical literacy skills in schools URL: https://literacytrust.org.uk/policy-and-campaigns/all-party-parliamentary-group-literacy/fakenews/.
12. Jurafsky, D., Martin, J.H.: Speech and language processing. Volume 3. Pearson London: (2014)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder–decoder for sta- tistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). (2014)
15. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.: Opennmt: Open-source toolkit for neural machine translation. Proceedings of ACL, System Demonstra- tions (2017)
16. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
17. Mei, H., Bansal, M., Walter, M.R.: Coherent dialogue with attention-based lan- guage models. In: AAAI. (2017) 3252–3258
18. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. In: Proceedings of NAACL-HLT. (2016)
19. Rubin, V.L., Liddy, E.D.: Assessing credibility of weblogs. In: AAAI Spring Sym- posium: Computational Approaches to Analyzing Weblogs. (2006)
20. news.com.au: The potential of AI generated 'crowdturfing' could undermine online reviews and dramatically erode public trust URL: http://www.news.com.au/technology/online/security/the-potential-of-ai-generated-crowdturfing-could-undermine-online-reviews-and-dramatically-erode-public-trust/news-story/e1c84ad909b586f8a08238d5f80b6982.
21. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015. Technical report (2015)
22. Honnibal, M., Johnson, M.: An improved non-monotonic transition system for dependency parsing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACM (2015)
23. Bird, S., Loper, E.: NLTK: the natural language toolkit. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Association for Com- putational Linguistics (2004)
24. Kumar, S., Shah, N.: False information on web and social media: A survey. arXiv preprint arXiv:1804.08559 (2018)
25. Everett, R.M., Nurse, J.R.C., Erola, A.: The anatomy of online deception: What makes automated text convincing? In: Proceedings of the 31st Annual ACM Sym- posium on Applied Computing. SAC '16, ACM (2016)
26. https://en.oxforddictionaries.com/spelling/common-misspellings
27. https://pypi.python.org/pypi/weighted-levenshtein/0.1
28. https://pypi.python.org/pypi/autocorrect/0.1.0

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Stay On-Topic: Generating Context-specific Fake Restaurant Reviews	Proceedings of the 23rd European Symposium on Research in Computer Security	2018