原文の書き換えによる広告文生成

村上 聡一朗[1] 菊田 洸[1] 張 培楠[1] 上垣外 英剛[2,4] 高村 大也[3,4] 奥村 学[4]
[1] 株式会社サイバーエージェント [2] 奈良先端科学技術大学院大学
[3] 産業技術総合研究所 [4] 東京工業大学
{murakami_soichiro,kikuta_ko,zhang_peinan}@cyberagent.co.jp
kamigaito.h@is.naist.jp takamura.hiroya@aist.go.jp oku@pi.titech.ac.jp

概要

本研究では商品ページ等から抽出した文（原文）を広告文として用いることを狙い，原文の体裁を整える書き換えモデルを提案する．同モデルでは，原文の長所である忠実性をいかに損なわず適切な広告文へ書き換えられるかが重要な課題となる．これに対し本研究では，書き換えモデルの学習データの忠実性を高めることで同モデルの忠実性を保証する方法を試みる．自動評価および人手評価の結果，書き換えモデルにより原文の忠実性を保ちつつ広告として適切な文へ書き換えられることを確認した．

1 はじめに

広告文は，消費者の興味を惹きつけ，商品・サービスの申込みや購入といった購買行動を促す重要な役割を担っている．その一方で，急拡大するオンライン広告の需要に伴い人手による広告制作は限界を迎えつつあり，広告制作の自動化が期待されている．近年ではニューラルベースの生成型手法が盛んに研究されている [1, 2, 3]．しかし，生成型手法には入力の商品説明ページ（Landing Page; LP）と一貫しない出力，すなわち忠実性の低い文を生成してしまう致命的な問題が指摘されており，自動生成技術を広告制作現場で運用する上で大きな障壁となっている [4]．そのため，正確な広告文の制作が求められる実運用では，誤りの可能性がほとんど無いテンプレート [5] や抽出型手法 [6] が重宝されている．

本研究では，LP から広告見出し文（以下，広告文）を生成する問題 [2, 3, 7] を例として，抽出型手法に焦点を当てる．図 1 のように，LP には購買行動を促すような有益な文が多く含まれる．しかし，抽出型手法により有益な抽出文（以下，原文）が得られたとしても広告文の文長制約[1)]や不要な記号を含む等の理由から広告として使用できないことが多々ある [8]．例えば，図 1 の原文 (22 文字) は不要な記号「■」を含み，文長制約も超過するため広告文として利用できない．したがって原文を適切に加工する必要がある．Fujita [8] らは依存木の刈り込みによる広告文生成手法を提案したが，刈り込みパターンを人手で構築する必要があるという課題があった．

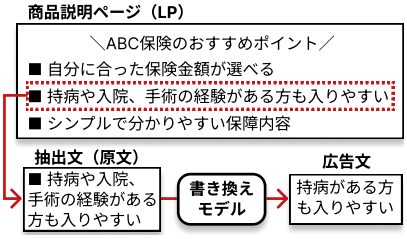

図 1 書き換えモデルの適用例

本研究では，原文の文長や文体等の体裁を整える処理を系列変換問題として考え，原文から広告文を生成する書き換えモデルを提案する．本研究では，後述する方法 (§2.1) によって構築した原文と広告文のペアデータを用いてモデルを学習する．ここで最も配慮すべきことは，**原文の忠実性を損なわずに広告として適切な文へ書き換えることができるか**であり，忠実性の保証は最重要課題である．そこで本研究では，忠実性の高い学習データが生成型要約の忠実性改善に寄与したという報告 [9] に基づき，書き換えモデルの学習データの忠実性を高めることで忠実性を保証できないかと考えた．具体的には原文と広告文の入出力ペアを構築する際に忠実性の低いペアを除去することで学習データの改善を狙う．加えて，フィルタリングによって学習事例数が減少し生成品質が劣化する問題に対して，本来除外される全ての事例を活用する学習方法を導入し，生成品質の改善に寄与するかを検証する．自動評価と人手評価の結果，提案モデルにより忠実性を保ちつつ広告として適切な文へ書き換えできることを確認した．

1） Google 広告では全角 15 文字の文字数制限を設けている．

2 提案手法

本研究では，LP から抽出された原文から適切な文長かつ文体を満たす広告文を生成する書き換えモデルを提案する．広告データは各 LP に対して複数の広告文が対応するデータである．N 文の原文を含む LP を $X = [x_1, x_2, \ldots, x_N]$，$M$ 文の広告文集合を $Y = \{y_1, y_2, \ldots, y_M\}$ と表記する．なお，原文 x_i，広告文 y_j は単語系列である．本研究では，モデルに事前学習済み Transformer [10] を用い，原文 x_i から広告文 y_j を生成する系列変換問題を後述の学習データにより追加学習する．以降では，書き換えモデルのデータセット構築および学習方法を説明する．

2.1 データセット構築

書き換えモデルの学習データとして，原文 x_i と広告文 y_j のペアを作成する．本研究では文書要約の先行研究 [11, 12] に倣い，各原文と広告文の ROUGE-1 (再現率) スコア [13] に基づいて文ペア (x_i, y_j) を作成する．以降では，文ペア (x_i, y_j) の同スコアを $\mathrm{Score}(x_i, y_j) \in [0, 1]$ と表記する．本研究では，原文 x_i の単語を多く含む広告文 y_j からなる文ペア，すなわちスコアが高い文ペアは忠実性が高いと仮定する．したがって，まず全ての原文 x_i と各広告文 y_j のスコアを算出し，スコアが高いペアを優先してデータセットへ加える．本研究では，より多くの文ペアを作成するために各広告文 y_j と全ての原文 X のスコアを算出したのち，スコアが上位 10 件の文ペアをデータセット $\mathcal{D}$ へ追加する．しかし，Y の中には原文 x_i と対応付けが難しい広告文 y_j も多く存在しており，結果的に忠実性の低いペアが $\mathcal{D}$ に混入する恐れがある．そのため本研究では，$\mathrm{Score}(x_i, y_j)$ が閾値 α 未満の文ペアを $\mathcal{D}$ から除去することで，より忠実性の高い学習データを構築する．本稿では，フィルタリング後の閾値 α 以上のペアからなる学習データを $\mathcal{D}_{\geq\alpha}$ と表記する．

2.2 学習方法

フィルタリングにより学習データの忠実性は期待できる一方で，学習事例数が減少することで生成文の品質が劣化することが懸念される．そこで本研究では，以下の各手法により学習事例数を減らすことなく生成文の忠実性を保証できるか検証する．

2 段階追加学習 (Two-stage Training)　$\mathcal{D}$ には閾値 α 未満の忠実性が低い文ペア（以下，ノイズ）も含まれるが，出力側は人手で書かれた品質の高い広告文であるため，全ペアデータ $\mathcal{D}$ を学習に活用したい．そこで，全学習事例 $\mathcal{D}$ とフィルタリング後の $\mathcal{D}_{\geq\alpha}$ に対する 2 段階追加学習を実施する．具体的には，まず事前学習済み Transformer に対して $\mathcal{D}$ を用いた 1 回目の追加学習を行い，その後，より忠実性の高い $\mathcal{D}_{\geq\alpha}$ を用いた 2 回目の追加学習を実施する．これにより，1 回目で広告文の生成方法を獲得し，2 回目で忠実性を改善することを期待する．

タグによる制御 (Control Codes)　$\mathcal{D}$ の各文ペア (x_i, y_j) の忠実性をスタイルと見なし，タグにより生成文の忠実性を制御する．具体的には先行研究 [14] に倣い，入力の先頭に各文ペアの忠実性を表すタグを連結する．各文ペアの忠実性を表す指標として $\mathrm{Score}(x_i, y_j)$ を用い，同スコアを 5 つの区間に等分割することで各区間に対応するタグ (<S0>, <S8> 等) を作成する．ここで，<S0> は $\mathrm{Score}(x_i, y_j) \in [0, 0.2]$ の文ペア，<S8> は $\mathrm{Score}(x_i, y_j) \in (0.8, 1]$ の文ペアに付与するタグを表す．これにより，学習時に全ての事例を活用できる．推論時には <S8> を用いることで忠実性が高い文が生成されることを期待する．

3 実験

書き換えモデルの効果検証のために，原文が事前に与えられる設定で同モデルの有用性を検証する．

3.1 実験設定

データセット　(1) 広告データの収集，(2) 原文-広告文ペア構築の工程により書き換えモデルのデータセットを構築した．(1) まず Web 上の広告配信データから全 31 社分の広告データ (以下，元データ) を 24,525 件用意し，15,000 件，4,919 件，4,606 件の学習, 開発, 評価データへ分割した．ここで LP と広告文集合のセット (X, Y) が 1 件であり，X と Y に含まれる平均文数はそれぞれ 212.2 文，12.5 文である．(2) 次に，元データから §2.1 の手法により原文-広告文のペア (x_i, y_j) を作成した．元データには同じ広告文も含まれるため重複削除を行った．その結果，書き換えモデルの学習，開発，評価データとして 39,430 件，534 件，1,824 件のペアが得られた．ここで，学習データはフィルタリング前のデータあり，開発，評価データには閾値 α が 0.7 以上のペアを採用した．構築したペアデータの例を付録 A に示す．

実装　書き換えモデルには事前学習済み mT5 [15] を用いる．その他の詳細は付録 B に記載する．

表 1 「日本語が不適切」と人手でラベル付けされた例. なお，[NE] は社名・サービス名等の秘匿化処理を表す.

、アニカ車&カーシェア，【１】個人事業主キャッシング
リボ払いまとめ [NE] 銀行，[NE][NE] 伊東市の宅配

自動評価 生成文の品質を測るために参照文との単語一致に基づく ROUGE を使用する．生成文の忠実性を測る指標として生成文中の単語のうち原文に含まれる割合，すなわち単語被覆率（以下，被覆率）を用いる．しかし，被覆率では原文と生成文で言い回しが異なる表現 (“50%OFF”, “半額”) の忠実性を考慮できないため人手評価も実施する．また, 生成文が広告文における文長制約を満たすかを評価する．本研究では，広告見出し文の入稿規定[2)]に従い，文長が全角 15 文字以内の生成事例の割合を報告する．さらに，生成文の自然さを測る指標として，パープレキシティ (PPL) を使用する．PPL の算出には評価データとは異なる 128,920 件の広告文で追加学習した GPT-2 [16] を用いた．また生成文が広告として適切な文であるかを評価するために，株式会社サイバーエージェント社内で過去に生成された広告文に対して広告品質チェックの専門家が「広告として日本語は適切か」を 2 値 (適切, 不適切) でラベル付けしたデータセットを用い，広告文としての適切さを評価するモデルを構築した．表 1 に人手で「不適切」とラベル付けられた事例を示す．これらの文は「不要な文字」「単語の繰り返し」「単語のつながりの不自然さ」等の理由により「日本語が不適切」とラベル付けされている．なお，本研究では適切さを評価するモデルを容認度判定モデルと呼称し，「適切」と予測された生成事例の割合を報告する．同モデルの評価データにおける F 値（マクロ平均）は 0.812 であった. その他の同モデルの詳細は付録 C に示す.

人手評価 忠実性（原文が生成広告文を含意するか），流暢性（内容が理解でき自然な文であるか）について 3 人の評価者による 2 段階評価（はい，いいえ）を実施した．評価データから無作為抽出した 200 件を評価対象とし，各事例について 2 人以上が「はい」と答えた割合を報告する．

3.2 実験結果

表 2 に評価対象のモデルおよび実験結果を示す．実験では，学習データの忠実度が生成品質に与える影響と §2.2 の学習方法を検証するために複数モデルを用意した．ここで，学習データ $\mathcal{D}$ で訓練したモデルを Baseline $_{\mathcal{D}}$，閾値 α のフィルタリングを適用したデータ ($\mathcal{D}_{\geq 0.5}$ 等) で訓練したモデルを Filter $_{\mathcal{D}_{\geq 0.5}}$ と表記する．なお，フィルタリング後の各データ ($\mathcal{D}_{\geq 0.5}$，$\mathcal{D}_{\geq 0.7}$，$\mathcal{D}_{\geq 0.9}$) の事例数は，11,022 件，3,291 件，1,688 件である．また，Baseline $_{\mathcal{D}}$ に対してフィルタリング後のデータ ($\mathcal{D}_{\geq 0.9}$ 等) で 2 回目の追加学習を実施したモデルを Two-stage $_{\mathcal{D}_{\geq 0.9}}$，タグで忠実性を制御するモデルを Control $_{\mathcal{D}}$ と表記する．

3.2.1 フィルタリングによる劣化を低減できるか

ROUGE スコアにより各モデルの生成品質を考察する．Filter $_{\mathcal{D}_{\geq 0.9}}$ 以外は高いスコアを得ることができた．Filter $_{\mathcal{D}_{\geq 0.9}}$ ではフィルタリングにより学習事例数が 1,688 件に削減されたことで十分な学習が出来ず生成品質が劣化したことが示唆される．これは PPL からも同モデルの生成品質が低いことが推察できる．また，Baseline $_{\mathcal{D}}$ に対して 2 回目の追加学習をするモデル (Two-stage $_{\mathcal{D}_{\geq 0.5}}$ 等) では，フィルタリング済みのデータだけを訓練に用いるモデル (Filter $_{\mathcal{D}_{\geq 0.5}}$ 等) と比べてスコア改善が見られた．この結果から 2 段階追加学習はフィルタリングによる生成品質の劣化を低減できていることが推察できる．同様に，タグで忠実性を制御する Control $_{\mathcal{D}}$ は Baseline $_{\mathcal{D}}$ から改善が確認できた．これはタグ制御により生成文の忠実性が改善したことで，参照文と類似した文が生成できたためと考えられる．

3.2.2 学習データの忠実度が生成品質に与える影響

生成文中の単語のうち原文に含まれる単語の割合を表す被覆率と人手評価に基づく各モデルの忠実性とを比較する．表 2 の結果から，より忠実度の高いデータ ($\mathcal{D}_{\geq 0.7}$，$\mathcal{D}_{\geq 0.9}$) で学習した Two-stage $_{\mathcal{D}_{\geq 0.7}}$，Two-stage $_{\mathcal{D}_{\geq 0.9}}$ は被覆率が高く，人手評価においても原文に対して忠実な文と判定される割合が高いことを確認した．表 4 に各モデルの生成例および忠実性の人手評価で「はい」と答えた人数を示す．表 4 に示すように，被覆率が高い Two-stage $_{\mathcal{D}_{\geq 0.7}}$，Two-stage $_{\mathcal{D}_{\geq 0.9}}$ は原文 (表題中の下線) から抽出的に生成することで忠実性を高めていることが分かった．しかし被覆率を高めるだけでは流暢性が損なわれる恐れがある．そこで流暢性を人手評価したところ，各モデルの流暢性はいずれも高いことを確認した (表 2)．なお，Baseline $_{\mathcal{D}}$ では事実に反する情報を含む不自然な生成文が多く，評価者にとっ

2) https://support.google.com/google-ads/answer/1704389

表 2 モデルの評価結果．ROUGE-1，ROUGE-2，ROUGE-L (R-1，R-2，R-L) は F 値を表す．各列の最高値を**太字**で表す．

モデル	自動評価							人手評価	
	R-1	R-2	R-L	PPL	容認度	文長	被覆率	忠実性	流暢性
BASELINE $\mathcal{D}$	0.434	0.350	0.428	**63.8**	0.807	0.721	0.494	0.290	0.800
FILTER $\mathcal{D}_{\geq 0.5}$	0.512	0.387	0.503	105.3	**0.834**	0.779	0.617	–	–
FILTER $\mathcal{D}_{\geq 0.7}$	0.692	0.576	0.672	295.3	0.675	0.828	0.963	–	–
FILTER $\mathcal{D}_{\geq 0.9}$	0.194	0.084	0.187	412.0	0.532	0.717	0.240	–	–
TWO-STAGE $\mathcal{D}_{\geq 0.5}$	0.593	0.478	0.582	78.6	0.702	0.884	0.768	0.645	0.940
TWO-STAGE $\mathcal{D}_{\geq 0.7}$	0.696	0.588	0.679	109.1	0.671	0.838	0.943	0.865	0.975
TWO-STAGE $\mathcal{D}_{\geq 0.9}$	**0.703**	**0.606**	**0.691**	144.9	0.648	0.855	**0.984**	**0.980**	**0.990**
CONTROL $\mathcal{D}$	0.566	0.461	0.556	81.4	0.752	0.872	0.679	–	–
原文	–	–	–	107.6	0.595	0.513	–	–	0.950
参照広告文	–	–	–	82.5	0.656	**0.992**	0.952	0.890	0.985

表 3 TWO-STAGE $\mathcal{D}_{\geq 0.9}$ による各原文の書き換え結果．なお，[NE] は社名・サービス名等の秘匿化処理を表す．

原文	生成文
| 奈良県奈良市の分譲マンション	奈良市分譲マンション
[NE] 生命の商品に詳しい「保険の専門家」のご紹介はこちら	[NE] 生命保険のご相談はこちら
業界最安水準の保険料!!「業界最安水準」について (	業界最安水準の保険料

表 4 原文「保証料 0 円! 一部繰上返済手数料 0 円! 返済シミュレーションはこちら」の書き換え結果

Model	生成文	忠実性
BASELINE $\mathcal{D}$	ローンの返済総額で比較	0
TWO-STAGE $\mathcal{D}_{\geq 0.5}$	初期費用 0 円/返済総額 0 円	0
TWO-STAGE $\mathcal{D}_{\geq 0.7}$	返済シミュレーションはこちら	3
TWO-STAGE $\mathcal{D}_{\geq 0.9}$	返済シミュレーションはこちら	3
参照広告文	一部繰上げ返済手数料が 0 円	2

て内容の理解が難しいことから流暢性が比較的低い結果となった．以上の結果より TWO-STAGE $\mathcal{D}_{\geq 0.7}$，TWO-STAGE $\mathcal{D}_{\geq 0.9}$ は忠実性と流暢性が優れていることが分かった．

3.2.3 原文を適切な広告文へ書き換えられるか

書き換えモデルによって原文から広告として日本語が適切な文を生成できるかを評価するために，原文と生成文が容認度判定モデルで「適切」と判定された割合を比較する．表 2 の結果から FILTER $\mathcal{D}_{\geq 0.9}$ 以外は，原文（0.595）よりも「適切」と判定された割合が高い (0.648~0.834) ことが確認できた．すなわち，書き換えモデルは広告として使用できない原文を適切な広告文へ書き換える手段として有用であることが示唆される．表 3 に TWO-STAGE $\mathcal{D}_{\geq 0.9}$ により原文を書き換えた結果を示す．これらの結果から TWO-STAGE $\mathcal{D}_{\geq 0.9}$ は，原文に含まれる単語を抽出しつつ，不要な文字を除去することで，原文から広告として適切な文を生成していることを確認できる．

また，参照広告文を容認度判定モデルで評価したところ，「適切」と判定された割合が比較的低い (0.656) ことが分かった．これは評価データ中に単語が列挙された広告文（例えば，「マンション 松本市」）が多く含まれており，こうした文が「不適切」と判定されたためと考えられる．

3.2.4 文長制約を満たすか

原文と生成文について文長制約を満たす割合を比較する．原文の中で文長制約を満たす割合は 0.595，生成文は 0.717~0.884 であり，書き換えモデルにより原文の多くを適切な文長へ圧縮できることを確認できた．一方，文長制約を超過する文も残存しており，文長制御 [17] の導入など改善の余地もある．

4 おわりに

本研究では，LP から抽出した文を広告として適切な文へ書き換える書き換えモデルを提案した．評価結果から書き換えモデルが原文を適切な広告文へ書き換える有用な手段であることを確認した．

一方で本研究には数多くの課題が残る．例えば，データ構築 (§2.1) における忠実性判定では原文と広告文の ROUGE を用いたが，同手法は表層に基づくため同義語に脆弱である．その対策として，表層ではなく 2 文の含意関係 (原文は広告文を含意するか) による判定が考えられる [18]．また，広告では忠実性に加えて魅力度 (消費者を惹きつける内容か) も重視される [2]．書き換えによる魅力度劣化を防ぐために，魅力度も考慮した手法などが求められる．

謝辞

本研究は，株式会社サイバーエージェントと東京工業大学の共同研究の成果をまとめたものです．

参考文献

[1] 村上聡一朗, 星野翔, 張培楠. 広告文自動生成に関する最近の研究動向. 2022 年度 人工知能学会全国大会, 2022.

[2] J. Weston Hughes, Keng-hao Chang, and Ruofei Zhang. Generating better search engine text advertisements with deep reinforcement learning. In **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, pp. 2269–2277, 2019.

[3] Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. An empirical study of generating texts for search engine advertising. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers**, pp. 255–262, 2021.

[4] Konstantin Golobokov, Junyi Chai, Victor Ye Dong, Mandy Gu, Bingyu Chi, Jie Cao, Yulan Yan, and Yi Liu. DeepGen: Diverse search ad generation and real-time customization. preprint, arXiv:2208.03438, 2022.

[5] Kevin Bartz, Cory Barr, and Adil Aijaz. Natural language generation for sponsored-search advertisements. In **Proceedings of the 9th ACM Conference on Electronic Commerce**, pp. 1–9, 2008.

[6] Stamatina Thomaidou, Konstantinos Leymonis, and Michalis Vazirgiannis. GrammAds: Keyword and ad creative generator for online advertising campaigns. In **Proceedings of the 1st International Conference on Digital Enterprise Design and Management**, pp. 33–44, 2013.

[7] 村上聡一朗, 星野翔, 張培楠, 上垣外英剛, 高村大也, 奥村学. Lp-to-text: マルチモーダル広告文生成. 言語処理学会第 28 回年次大会, 2022.

[8] Atsushi Fujita, Katsuhiro Ikushima, Satoshi Sato, Ryo Kamite, Ko Ishiyama, and Osamu Tamachi. Automatic generation of listing ads by reusing promotional texts. In **Proceedings of the 12th International Conference on Electronic Commerce**, pp. 179–188, 2010.

[9] Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. Improving truthfulness of headline generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1335–1346, 2020.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in Neural Information Processing Systems 30**, 2017.

[11] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In **Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence**, p. 3075–3081, 2017.

[12] Yen-Chun Chen and Mohit Bansal. Fast abstractive summarization with reinforce-selected sentence rewriting. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics**, pp. 675–686, 2018.

[13] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Proceedings of the ACL Workshop: Text Summarization Branches Out**, pp. 74–81, 2004.

[14] Katja Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 864–870, 2020.

[15] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 483–498, 2021.

[16] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[17] Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Controlling output length in neural encoder-decoders. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1328–1338. Association for Computational Linguistics, 2016.

[18] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 163–177, 2022.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **Proceedings of 3rd International Conference on Learning Representations**, 2015.

[20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, 2020.

[21] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, 2004.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. preprint, arXiv:1907.11692, 2019.

表 5　原文-広告文ペアの例．なお，[NE] は社名・サービス名等の秘匿化処理を表す．

	スコア	原文	広告文
(a)	1.000	最新!2022 年 4 月版 人気の女性保険ランキング【[NE]】	【2022 年女性保険ランキング】
(b)	1.000	定期おトク便 5 大特典付き!2 初回限定 50%off	定期初回限定 50%off
(c)	0.909	- 2022.04.05【静岡県民限定】「バイ・シズオカ ~ 今こそ! しずおか!!元気旅!!!~」再開のお知らせ	【伊豆】今こそしずおか元気旅
(d)	0.900	《厳選》今、注目の消費者金融を大紹介/web 完結・契約後すぐ口座へ振込・内緒で借入 ok	今、注目の消費者金融をご紹介
(e)	0.889	おかげさまで 2,200 万袋突破!※ 2021 年 10 月当社調べ	おかげさまで 2,100 万袋突破
(f)	0.800	|no.1 自動車保険一括見積もり|	自動車保険一括見積もりサイト
(g)	0.750	今おすすめの人気の生命 (死亡) 保険ランキングを発表!	死亡保険の人気 top3 を発表
(h)	0.400	高知県の結婚式場探しならこちら。	【ブライダル 高知】なら
(i)	0.375	保険を知る・学ぶ方はこちら	保険無料相談やお見積りはこちら

表 6　容認度判定モデルのデータセット内訳

	学習	開発	評価
適切	26,427	1,000	1,000
不適切	26,427	1,000	1,000
合計	52,854	2,000	2,000

表 7　容認度判定モデルの評価結果

	適合率	再現率	F 値
適切	0.766	0.902	0.829
不適切	0.881	0.725	0.795
マクロ平均	0.824	0.814	0.812

A　ペアデータの例

表 5 に作成した原文-広告文ペアデータの例を示す．表中のスコアは，§2.1 のデータセット構築における原文と広告文の ROUGE-1（再現率）を表す．

本研究では，原文中の単語を多く含む広告文は忠実性が高いと仮定する．例えば表 5 の原文-広告文ペア (a, b) では，広告文が原文中の単語を全て含むためスコアは 1.000 であり，広告文の内容も原文と一貫していることから忠実性が高いといえる．また，スコアが低い文ペア (h, i) の広告文では，「ブライダル」や「お見積り」と等の原文に含まれない内容が言及されており，忠実性が低いことが分かる．

一方，同スコアが高い文ペアであっても忠実性が高いとはいえない事例も存在する．例えば文ペア (c, e) はそれぞれ高いスコアではあるものの，広告文中の「伊豆」「2,100」といった固有名詞や数字が原文には含まれておらず，原文に対して広告文の内容が一貫しているとはいえない．広告文において固有名詞や数字等は消費者の購買行動に繋がる重要な情報であり，特に正確性に注意を払う必要がある．したがって，今後の課題として固有名詞や数字等の情報が原文と一貫しない事例をデータセットから除去するためのルールや手法の整備などが考えられる．

B　実装

モデルパラメータの最適化手法には Adam [19] を使用し，学習エポック数は 5，ミニバッチサイズは 10 とした．なお，10 ステップごとの累積勾配に基づいてパラメータを更新している．入力系列および出力系列の最大文長はそれぞれ 512，256 とした．生成時にはビーム探索を用い，ビーム幅は 5 とした．なお，実装には Transformers [20] を用いており，書き換えモデルとして用いた mT5[3]，PPL の算出に用いた GPT-2[4] のモデルは脚注に示す通りである．

データセットに対する前処理として，Unicode 正規化，アルファベットの小文字化を実施した．また，ROUGE や被覆率などの算出時には MeCab (IPA 辞書) [21] を用いて単語分割した．

C　容認度判定モデル

容認度判定モデルは，「入力として与えられた生成文は広告として日本語が適切な文であるか」を判定する二値分類モデルである．モデルには事前学習済み RoBERTa[5] [22] を使用し，株式会社サイバーエージェント社内で過去に生成された 56,854 件の広告文に対して広告品質チェックの専門家が「広告として日本語は適切か」を二値（適切，不適切）でラベル付けしたデータセットを用いて追加学習した．表 6 に学習，開発，評価データおよび各ラベルの事例数の内訳を示す．表 7 に容認度判定モデルの評価データに対する評価結果を示す．評価データに対する F 値（マクロ平均）は 0.812 であった．

3）https://huggingface.co/google/mt5-base
4）https://huggingface.co/rinna/japanese-gpt2-small
5）https://huggingface.co/rinna/japanese-roberta-base