参照例を使わないキャッチコピーの自動評価

新保 彰人　山田 寛章　徳永 健伸
東京工業大学 情報理工学院
{shimbo.a.aa@m, yamada@c, take@c}.titech.ac.jp

概要

広告文の一種であるキャッチコピーの人手によるオフライン評価は高コストである．キャッチコピーの自動生成研究の迅速化・効率化のためには自動評価器が必要となる．自動評価器の構築のために必要なデータセットが現存しないため，日本語としては初となる 23,641 件のキャッチコピーとその評価値から成るデータセットを構築した．このデータセットを利用して BERT と対照学習を用いた参照例を必要としない評価機を構築し，評価実験を行った結果，テストデータの評価値に対する相関係数が平均で 0.28 を超えた．対照学習を用いない学習との比較も行い，対照学習の有用性を確認した．

1　はじめに

近年，インターネット広告市場の拡大が著しい．総務省によると 2021 年には世界のデジタル広告の市場規模は 39 兆 396 億円となり前年比で 32.7%増であった．日本についてもインターネット広告の市場規模は 2 兆 7,052 億円となり，テレビ・新聞・雑誌・ラジオのマスコミ 4 媒体の合計を初めて上回った [1]．このような背景からインターネット広告における広告文生成の需要は高まっていると考えられる．

キャッチコピーは広告文の一種であり，人々の注目を引き商品の魅力を簡潔に伝えるための宣伝文句である．例えば株式会社日清食品の「すぐおいしい、すごくおいしい」[1)]や．株式会社大塚製薬の「それは小さな栄養士。」[2)]などが挙げられる．

キャッチコピーはその他の広告文と比較して評価が難しいことが知られている．村上ら [2] は「キャッチコピーは見る人の印象に残ることを主な目標として使用されるため．検索連動型広告やディスプレイ広告と異なり，広告効果の定量化が困難な性質を持つ.」という点を指摘している．

広告文の評価方法はオンライン評価とオフライン評価に大別され，オフライン評価はさらに人手評価と自動評価に大別される．オンライン評価は実際に広告を配信し．広告のクリック率や収益率を評価尺度として利用する．そのため，オンライン評価は広告文に対する評価を収集するのに時間がかかるデメリットがある．一方でオフライン評価は広告文を配信せずに行う．人手評価は人間が広告文に対して付けた評価を評価尺度として用いる．このため，オンライン評価と同様に実施に時間がかかるデメリットがある．一方で自動評価は人手を介さずに評価を行うため実施に時間がかからない利点がある．より質の良い広告文生成器を開発するためには効率良く学習と評価のサイクルを回す必要があり，そのために広告文の自動評価は不可欠である．

広告文生成の研究において，生成文の自動評価としては BLEU，ROUGE，METEOR，CIDEr が用いられることが多い [3]．しかし，これらの評価手法は参照文を必要とし，その参照文をあたかも「正解」の広告文として扱ってしまう問題がある．参照文を必要としない自動評価手法として，事前学習済みの汎用言語モデルのパープレキシティを評価尺度として用いるもの [4] があるが，汎用ドメインで学習されたモデルでは，文としての妥当性は評価できるが広告文としての質を評価できないと考えられる．

広告文生成の研究に関わる課題として村上ら [2] は「共通データセットの拡充」を挙げている．実際，パブリックに利用可能なキャッチコピーコーパスは我々が知る限りでは存在しない．本研究では，ウェブスクレイピングによってキャッチコピーコーパスを作成し，そのデータセットを用いて参照文を必要としないキャッチコピー評価器を構築する．

2　関連研究

花野ら [5] は RoBERTa を用いた俳句の評価器を提案している．ただし，花野らの研究では俳句の質の

1） https://www.chickenramen.jp/about/
2） https://www.otsuka.co.jp/cmt/cm/

評価というよりは，俳句とみなせる文とみなせない文の識別を行なっている．

丹羽ら [3] は宣伝会議が出版しているキャッチコピー集 SKAT に収録されているキャッチコピー 91,682 件を分析し，キャッチコピーの平均文字数が 18.91 文字であることやキャッチコピーに使われる文字種の割合を分析している．

黒木ら [6] は複数の指定語句を必ず含むリスティング広告文の生成手法を提案している．リスティング広告とは検索エンジンの検索結果画面に表示される広告のことである．黒木らの研究では生成文の評価は BLEU，ROUGE-1，ROUGE-2 の 3 つで行なっている．また，Chao らの研究 [4] では広告文生成器の評価を Pairwise-BLEU とパープレキシティで行なっている．

3 データセット

3.1 コピラ

キャッチコピーのデータセットを作成するためにコピラ [7] を利用した．コピラは東京コピーライターズクラブ（TCC）が提供しているコピー検索ツールである．コピラにはキャッチコピーだけでなく，テレビやラジオの CM のセリフ書き起こしなども収録されている．そのようなものを除外し，キャッチコピーとして適切なものだけを収集するため，掲載媒体が「新聞・ポスター・雑誌・パンフレット・その他」のいずれかに該当するものを収集対象とした．

コピラに収録されているコピーにはキャッチコピーの他に，広告本文に相当するボディコピーを含むとみなせるコピーも含まれ，平均文字数は 53.2 文字で最大文字数は 3,359 文字である．一方で，コピラの検索結果一覧に表示されるテキストの平均長は 21.8 文字であり丹羽らの分析による 18.91 文字と近い．このため，本研究ではコピラの検索結果一覧画面に表示されるテキストをキャッチコピーとみなして収集した．

収集したキャッチコピーは 24,331 件である．収集したキャッチコピーには改行文字やタブ文字が含まれる場合があり，それらはすべて半角スペースに置換した．スクレイピングに利用したコードは GitHub リポジトリ[3)]で公開予定である．

キャッチコピー評価器のためのデータセットには，キャッチコピー及びその評価が収録されている必要がある．キャッチコピーの評価は広告の対象・表現・世情をはじめ様々な観点を総合的に判断するものであるため，専門的知識を有する者がキャッチコピーに対して評価を付与することが望ましいが，新規に評価を行うことは時間的・金銭的な面で困難である．そこで，コピラ上で参照可能な受賞情報を活用する．コピラで収集したキャッチコピーには，テキストそのものに加えて受賞情報が付与されている．受賞情報の種類と件数は表 1 に示す．

表 1 キャッチコピーの受賞情報

受賞情報	件数
受賞なし	19,524
ファイナリスト	969
ノミネート	494
TCC 賞	684
部門賞	202
特別賞	119
TCC グランプリ	8
TCC 最高賞	52
TCC 広告賞	30
TCC クラブ賞	304
一般部門賞	81
審査委員長賞	73
会長賞	12
奨励賞	15
最高新人賞	162
新人賞	1,582
合計	24,311

本研究では受賞情報をキャッチコピーの質的尺度の一種として捉え，各キャッチコピーに対するスコアとして用いる．受賞情報とスコアを表 2 のように対応させ，順序を与える．スコアは高い方がキャッチコピーの質が高いことを意味する．コピラに収録されているキャッチコピーの多数を占める受賞なし作品を基準スコアの 0 と設定し，各賞の受賞作品にスコア 2 を割り当てた．なお，「ファイナリスト」と「ノミネート」は TCC 賞の最終選考まで残ったものの受賞はしなかった作品であることから，受賞作品と受賞なし作品の中間であるスコア 1 を割り当てた．また，新人賞は以前に TCC 賞を受賞したことがないコピーライターを対象として選考される賞であることから，新人賞以外の各賞との区別のためにスコア 1 を割り当てた．

3) https://github.com/ShimboAkito/copy-evaluator

表 2 に各スコア毎のキャッチコピー数を示す．表 1 と表 2 で件数が一致しないのは，表 2 ではキャッチコピーの重複を除いてカウントしているためである．

表 2　受賞情報とスコアの対応

受賞情報	スコア	件数
受賞なし	0	19,076
ファイナリスト ノミネート 新人賞	1	2,952
上記以外	2	1,613

4　実験設定

ベースラインモデルとしてランダム，パープレキシティ，SVR を用いた実験，さらに BERT を用いた回帰モデルと対照学習を利用したモデルでの実験を行う．データセットは各スコアのキャッチコピーの件数が等量になるように 1,613 件に揃えて実験する．データセットは学習データ 3,871 件，検証データ 484 件，テストデータ 484 件に分割した．

5　実験

5.1　Random

ランダムに 0,1,2 を出力するモデルをベースラインとして用意した．

5.2　Inverse Perplexity

2 つ目のベースラインとして rinna 株式会社が公開している日本語 GPT2[8] のパープレキシティの逆数 $PPL_{inv}(X)$ をキャッチコピーの評価として実験する．パープレキシティの逆数は次の式で計算する．

$$PPL_{inv}(X) = \exp\left(\frac{1}{t}\sum_{i}^{t}\log p(x_i|x_{0:i-1})\right)$$

ただし，X は入力文字列，t は入力テキストのトークン長，x_i は X の i 番目のトークン，$p(x_i|x_{0:i-1})$ はトークン列 $\{x_0, x_1, ...x_{i-1}\}$ の次に言語モデルがトークン x_i を出力する確率を表す．$PPL_{inv}(X)$ の計算に使用する言語モデルが十分に流暢であると仮定すると，$PPL_{inv}(X)$ は大きいほど X が流暢な文であることを意味する．流暢性のみによってキャッチコピーの評価が実施可能であるかどうかを検証するため，$PPL_{inv}(X)$ をキャッチコピーの評価のベースラインとして採用した．

5.3　SVR

独自に定義したキャッチコピーの特徴量を入力として SVR[9] で学習する．抽出した特徴量は，「数字を含む」，「3・5・7 のいずれかを含む」，「疑問文である」，「オノマトペを含む」，「断定表現である」，「体言止めである」，「ひらがなの割合」，「カタカナの割合」，「漢字の割合」，「文字数」の 10 種類である．それぞれの特徴量のデータタイプは 表 4 に示す．

抽出する特徴量の選定については弓削の書籍 [10] を参考にし，キャッチコピーの質に強い影響を及ぼすと思われるものを選んだ．「オノマトペを含む」の判定はウェブ上で公開されている擬音語・擬態語辞典 [11] に収録されているオノマトペを含むかどうかに基づいた．

カーネルは線形カーネルとしパラメータは $C = 0.1, \epsilon = 0.01$ とした．

5.4　BERT Regression

BERT Regression では，キャッチコピーを入力とし BERT の CLS トークンに対応する最終層に線形層を付け加えて得られたスカラーをスコアの予測値として学習する．正解のスコアを s とし，モデルの出力を p とする．損失関数は

$$L_{MSE} = (p - s)^2$$

で計算する．BERT の事前学習済みモデルは東北大学が公開しているモデル [12] を利用し，学習率は $1e-03$，バッチサイズは 64，エポック数は 100，tokenizer の max_length は 200 で実験を行なった．各エポックごとに検証データでモデルを評価して，その評価値が最も良かったエポックのモデルを使って学習後にテストデータでの評価をした．

5.5　BERT Contrastive

BERT Contrastive では，2 つのテキストを入力として対照学習を行う．モデル構造は BERT Regression と同じである．入力テキストを t_1, t_2 とし，それぞれに対するモデルの出力を p_1, p_2 とし，正解のスコアを s_1, s_2 とする．このとき $s_1 > s_2$ となるように入力データを作る．損失関数は

$$L_{MR} = \max(0, m - (p_1 - p_2))$$

で計算する．ただし m はマージンを表すハイパーパラメータである．ここで，正解スコア s_1, s_2 を使わずに損失関数を計算していることに留意されたい．

本実験では $m = 1$ とし，エポック数は 10 で行なった．その他のハイパーパラメータと利用した BERT の事前学習済みモデルは BERT Regression と同じである．

5.6 評価

モデルの評価はテストデータにおける，モデルの予測スコアと正解スコアの積率相関係数で行う．

6 結果と考察

実験結果を表 3 に示す．Random，BERT Regression，BERT Contrastive は 5 回ずつ実験を行い，他は 1 回のみ実験を行った．5 回行った実験には平均の相関係数を示し，標準偏差も示す．

表 3 実験結果

学習方法	相関係数	標準偏差
Random	0.009	0.077
Inverse Perplexity	−0.032	–
SVR	0.131	–
BERT Regression	0.124	0.039
BERT Contrastive	0.281	0.037

Inverse Perplexity は相関係数が −0.032 となった．これはキャッチコピーの評価として事前学習済み言語モデルのパープレキシティを用いることが不適であることを示唆している．この理由として，キャッチコピーとして成立している時点でそのテキストにはある程度の流暢性があり，それ以上の流暢性はキャッチコピーの質に影響を及ぼさないためだと考えられる．さらに，Inverse Perplexity より SVR の方が良い結果を示したことから，流暢性よりも「文字数」や「オノマトペを含む」などの特徴の方がキャッチコピーの質に大きな影響を及ぼすと考えられる．

BERT Regression はほとんど相関が無い結果となり，ニューラルモデルを用いない SVR よりも相関係数が劣る結果となった．一方，BERT Contrastive では平均相関係数が 0.28 を超え，弱い相関があるとみなせる水準に達した．BERT Contrastive の精度が BERT Regression よりも高かった要因として，BERT Contrastive は 2 つのテキスト間の相対的な評価を学習しているという点が挙げられる．本研究で扱っているキャッチコピーの評価というタスクにおいては，メタ評価指標として積率相関係数を利用しているため，評価器がスコアラベルの値自体を再現することよりもスコアラベルの相対的な差を再現することの方が本質的な意味を持つ．このため，スコアラベルの値をそのまま予測する BERT Regression よりも，相対的な差を学習する BERT Contrastive の方が高い精度を記録したと考えられる．

7 おわりに

本研究ではウェブスクレイピングによりキャッチコピーコーパスを作成した．さらに評価器の学習における対照学習の有用性を確認し，参照文を必要としないキャッチコピー評価器を提案した．今後のキャッチコピー生成の研究において，生成文や生成器の評価指標として広く利用されることを期待する．加えて本研究では事前学習済み言語モデルのパープレキシティをキャッチコピーの評価指標として利用することが不適であることを指摘した．なお実験に使用したソースコードは GitHub リポジトリ[4)]で公開予定である．

今後の課題としては次の 3 つの方向が考えられる．1 つ目は広告対象の商品情報を考慮したキャッチコピー評価器の開発である．キャッチコピーは文脈に依存したものが多く，人間がキャッチコピーの質を評価する際にも商品情報と照らし合わせて評価していると考えられる．従って商品情報を考慮したキャッチコピー評価器は，より高い精度を実現できると期待される．そのようなキャッチコピー評価器の開発には商品情報とキャッチコピーがセットになったコーパスの構築が必要となる．

2 つ目は修辞技法を考慮したキャッチコピー評価器の開発である．キャッチコピーには比喩，反復，対句のような修辞技法を用いたものが多く存在する．例えば，丹羽ら [3] の分析によると 17.9%のキャッチコピーが比喩を用いていて，10.6%は対句を用いている．修辞技法はキャッチコピーの質を左右する重要な要素であり，修辞技法を考慮することでより精度高い評価器を実装できると期待される．

3 つ目はキャッチコピー以外の広告文への応用である．広告文生成の研究はキャッチコピー生成以外にも広く行われているが，生成器の効率的な開発には自動評価手法が不可欠である．参照文を必要としない広告文評価器の開発は今後の広告文生成の研究を加速させることが期待される．

4) https://github.com/ShimboAkito/copy-evaluator

参考文献

[1] 総務省. 令和4年情報通信に関する現状報告の概要, (2022-12 閲覧). https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r04/html/nd233220.html.

[2] 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. 人工知能学会全国大会論文集, Vol. JSAI2022, pp. 1P5GS601–1P5GS601, 2022.

[3] 丹羽彩奈, 岡崎直観, 西口佳佑, 亀山千尋, 毛利真崇. キャッチコピーの自動生成に向けた分析. 言語処理学会 第25回年次大会発表論文集, pp. 558–561, 2019.

[4] Chao Zhang, Jingbo Zhou, Xiaoling Zang, Qing Xu, Liang Yin, Xiang He, Lin Liu, Haoyi Xiong, and Dejing Dou. Chase: Commonsense-enriched advertising on search engine with explicit knowledge. pp. 4352–4361, 10 2021.

[5] 花野愛里咲, 横山想一郎, 山下倫央, 川村秀憲ほか. マスク化言語モデル roberta を用いた俳句の評価. 第84回全国大会講演論文集, Vol. JSAI, No. 1, pp. 1061–1062, 2022.

[6] 黒木 開中田 和秀. 複数の指定語句を必ず含むリスティング広告の広告文自動生成. 言語処理学会 第28回年次大会発表論文集, pp. 1339–1343, 2022.

[7] 東京コピーライターズクラブ. コピラ, （2022-12 閲覧）. https://www.tcc.gr.jp/copira/.

[8] 趙天雨, 沢田慶. 日本語自然言語処理における事前学習モデルの公開. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 93, pp. 169–170, 2021.

[9] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In M.C. Mozer, M. Jordan, and T. Petsche, editors, **Advances in Neural Information Processing Systems**, Vol. 9. MIT Press, 1996.

[10] 弓削徹. 届く！刺さる！！売れる！！！ キャッチコピーの極意. 明日香出版社, 2019.

[11] 擬音語・擬態語 - 日本辞典, (2022-12 閲覧). http://nihonjiten.com/nihongo/giongo/.

[12] cl-tohoku/bert-base-japanese-whole-word-masking, (2022-12 閲覧). https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking.

A　参考情報

A.1　データタイプ

特徴量	データタイプ
数字を含む 3・5・7 のいずれかを含む 疑問文である オノマトペを含む 断定表現である 体言止めである	0 または 1
ひらがなの割合 カタカナの割合 漢字の割合	[0,1] の小数
文字数	整数

表 4　SVR に入力する特徴量のデータタイプ