AdGLUE: 広告言語理解ベンチマーク

張 培楠 [1]　坂井 優介 [2]　三田 雅人 [1]　大内 啓樹 [2,3]　渡辺 太郎 [2]
[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] 理化学研究所
{zhang_peinan,mita_masato}@cyberagent.co.jp
{sakai.yusuke.sr9,hiroki.ouchi,taro}@is.naist.jp

概要

近年、インターネット広告における自然言語処理技術の応用が盛んに行われている。広告分野では、学術的によく使われるデータセットでは出現しない表現も多く、文法的に誤った非文が許容されることもある。このような広告特有の言語現象に対処する技術が必要とされるが、共通のタスクやデータセットがないため、分野全体の研究発展が妨げられている。そこで本研究では、広告特有の言語現象や性質に対する言語理解を促進するため、5 つのタスクから構成される広告分野特化の言語理解ベンチマーク AdGLUE を提案する。また、一般的な言語タスクやデータセットとの違いを分析するため、主要な既存手法でベースライン実験も行った。

1　はじめに

近年、検索連動型広告（図 1）を含むインターネット広告市場の著しい成長 [1] に伴い、テキスト分類モデルを使ったクリック率予測 [2] やニューラル生成モデルを使った広告文生成 [3, 4, 5] など、自然言語処理技術の多くが広告分野で応用され始めている [6]。広告分野では、短文でユーザに商品の特徴と魅力を伝え、購買行動などに繋げるよう訴求する必要がある。そのため、使用される言語表現は一般的な文章とは異なることが多い。例えば、機能語や部分的な内容語の省略や強調のために記号の使用、場合によっては文法的な誤りを含んだ非文が許容されることもある。例として、「マンションを売却したい方は、今すぐ○○で無料で査定しませんか？」という汎用的な表現は、広告特有の表現に言い換えると「マンション売却／今すぐ無料査定」になる。このような広告特有の表現は人間の理解を妨げるどころか、注意を引くことに繋がることが広告心理学分野で報告されている [7]。

以上のような一般的な言語資源では存在しない

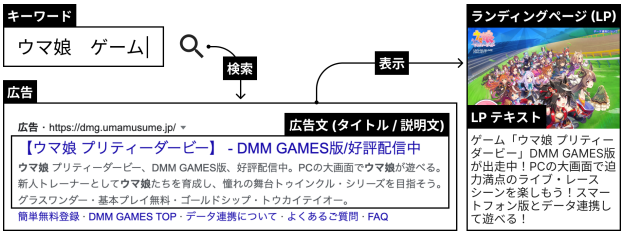

図 1　検索連動型広告に関連する要素と遷移の例

特徴を有しながらも、機械翻訳や自動要約などの分野と比較して、広告における自然言語処理の研究や分析は少ない。その一因として考えられるのは、共通するタスクやデータセット、そしてそれらで構成されるベンチマークの欠如である。共通ベンチマークがあることによって、手法の比較が容易になり、実験の手順が明確化するため、研究の活発化が期待できる。英語の一般的な言語理解を目指したベンチマークとして GLUE [8] や SuperGLUE [9] などが挙げられ、JGLUE [10] を始めとする多くの言語でも構築が進んでいる [11, 12, 13, 14]。また、特定の分野に特化したベンチマークも存在し、PLUE [15] はプライバシーポリシーのための言語理解ベンチマークであり、BLURB [16] は生物医学分野におけるベンチマークである。しかし広告分野には特化したベンチマークが現状存在せず、一貫したデータセットなども公開されていない。そのため、広告分野での研究の多くは独自のデータと指標で評価されており、横断的な比較が難しい状況になっている。また、広告データは配信プラットフォームを保有している一部の企業にしか所持していないこともあり、それ以外の企業や学術機関からの利用が困難である。

以上のことから、我々は広告分野に特化した言語理解ベンチマーク **AdGLUE** (**Ad**vertisin**G** **L**anguage **U**nderstanding **E**valuation) を提案する。このベンチマークでは、広告運用の手順を踏まえて、広告において重要視される容認性や一貫性、魅力度などの理

表 1　AdGLUE のタスクとデータセットの内訳

タスク	Train	Dev	Test	Total	タスク種類	入力	正解ラベル	評価指標
広告容認性	35,377	–	2,000	37,377	分類	広告文	2 値	Accuracy
広告一貫性	28,756	–	2,000	30,756	分類	広告文、LP	2 値	Accuracy
訴求表現認識	1,856	465	410	2,731	分類	広告文	多値	F1-Score (micro, macro)
広告類似性	4,980	623	629	6,232	回帰	広告文ペア	実数値	Peason/Spearman corr.
広告効果予測	138,358	–	965	139,323	回帰	広告文、キーワード	実数値	Peason/Spearman corr.

表 2　広告容認性、広告一貫性、訴求表現認識タスクのデータ例。LP テキストは広告一貫性タスクでのみ使用される。なお acceptable / not_acceptable と consistent / not_consistent はそれぞれ容認・非容認と一貫・非一貫を指している。

入力文	LP テキスト	容認性	一貫性	訴求表現
マンション売却／今すぐ無料査定	一戸建ての査定は○○におまかせ！	acceptable	not_consistent	無料, 利便性
エンジニアの転職／土日の求人多数	○○が運営する転職サイトで求人を見つけよう！	not_acceptable	consistent	品揃え

解に着眼して作られている。我々は一般的な言語理解と異なる広告特有の観点で 5 つのタスクを規定し、対応するデータセットを構築した。データは実際にインターネット上で掲載された広告であり、それらに対して人手によるアノテーションを行った。また、一般的な言語タスクとの違いを分析するため、提案ベンチマークに対して主要な既存手法でベースラインで実験・分析を行った。[1)]

2　AdGLUE

AdGLUE は、表 1 のとおり広告容認性、広告一貫性、訴求表現認識、広告類似性、そして広告効果予測の 5 つのタスクから構成されている。本研究で使用される広告データはすべて実際にインターネット上で掲載された検索連動型広告である。なお、広告効果予測以外の全データセット構築において、人手によるアノテーションが行われている。アノテーション作業者はいずれも日本語母語話者であり、ガイドラインの説明を受けた上で最大 2 回の訓練アノテーションを経験している。より詳細なデータセット情報は付録に示す。

2.1　広告容認性

広告容認性タスクは、表 2 に示すように、ある文を広告としてみた際に容認できる表現か、容認・非容認の 2 値ラベルを推定するタスクである。なお、推定結果は正解ラベルとの Accuracy で評価する。インターネット広告において、ほとんどの媒体は表示領域を確保するための文字長制限が設けられている[2)]ため、広告文は限られたスペースや少ない文字数で読者に印象付ける必要がある。そのため、広告文を制作する際には、人間が読み取れる範囲であえて文法を崩したり圧縮したりすることがある。一方で、過度な圧縮は読者の誤解を招いたり、意味理解を妨げる可能性もある。これは CoLA [17] など文法的に正しいかどうかを判断する一般的な言語容認性の考え方とは異なる。このように、読者の正確な意味理解の可否を広告容認性として定め、広告運用の現場においても人手で確認するステップが存在する。そこで我々は広告自動生成モデルによって生成された広告文に対して、広告運用と同じ基準で広告容認性のアノテーションを行った。アノテーション結果の一部を表 2 に示す。例えば 1 行目の「マンション売却／今すぐ無料査定」は述べている意図を問題なく読み取れるため「容認される」が、2 行目の「エンジニアの転職／土日の求人多数」は「土日の求人多数」の意味が不明なため「容認されない」結果となっている。

2.2　広告一貫性

広告一貫性タスクは、表 2 に示すように、広告文と広告文に紐づくランディングページ（LP）テキストが一貫しているか、一貫・非一貫の 2 値ラベルを推定するタスクである。なお推定結果は正解ラベルとの Accuracy で評価する。検索連動型広告には URL が紐づいており、図 1 のようにクリックすることでその URL 先の Web ページである LP に遷移する。LP には広告文で言及した商品の詳細情報が掲載されているが、その情報と広告文で述べられている内容に齟齬が生じる場合は顧客の信用を失うなど大きな損害が発生する。そのため、LP にある商品内容と広告文内での表現の一貫性は重要である。我々は LP テキストとして Web ページの HTML にあ

1)　作成されたデータセットは一般公開し、オンラインからアクセスできる評価プラットフォームを提供する予定である。
2)　例えば Google レスポンシブ広告ではタイトルに全角で最大 15 文字の制限を設けている。

表 3　広告類似性タスクのデータ例

例 1	**入力文 1**	すっぽん黒酢にセラミド贅沢配合
	入力文 2	すっぽん黒酢に贅沢セラミドまで
	類似度	5.00
例 2	**入力文 1**	ご予算に合わせて贈り物を検索
	入力文 2	お得な割引商品は最大 40 %OFF
	類似度	2.33

る meta タグの description 要素を使用し、紐付いている広告文と矛盾が存在するかどうかを人手でアノテーションした。アノテーション結果の一部を表 2 に示す。例えば 1 行目の広告文は「マンション売却／今すぐ無料査定」だが、対応する LP テキストは「一戸建ての査定は○○におまかせ！」で、「マンション」ではなく「一戸建て」について言及しているため「一貫していない」と判断できる。

2.3　訴求表現認識

訴求表現認識タスクは、表 2 に示すように、広告文がどのような訴求軸を含んでいるのか、該当する訴求ラベルをすべて推定するタスクである。なお推定結果は正解ラベルとの F 値で評価する。訴求とは、広告読者に対して、対象商品のメリットや魅力を訴えかけるような表現のことである。商品によって関心を持つ人が異なるため、適切に訴求できるかが広告効果に大きく影響する。このタスクでは、村上ら [18] の訴求表現データセットを使用しており、21 種類の訴求ラベルが 2,731 件の広告文に付与されている。[3] データ例として、表 2 の 1 行目の広告文「マンション売却／今すぐ無料査定」に対しては、「無料査定」から訴求ラベルの「無料」「利便性」が、2 行目の広告文「エンジニアの転職／土日の求人多数」に対しては、「求人多数」から訴求ラベルの「品揃え」が付与されている。本研究では村上らの研究の doc-base タスクにならって、span ではなく文全体を入力として考え、該当する全ての訴求ラベルを推定しその F 値を評価する。

2.4　広告類似性

広告類似性タスクは、表 3 に示すような広告文ペアがどれくらい広告的に類似しているのか、[1, 5] の範囲を持つ実数値を類似度として推定するタスクである。類似度は 1 に近づくほど低く 5 に近づくほど高く、推定結果は正解類似度とのピアソン相関係数およびスピアマン相関係数 [19] で評価する。

3)　先行研究では大分類と小分類からなる構造だが、本タスクでは小分類のみを使用する。

表 4　広告効果予測タスクのデータ例

業種	金融
キーワード	カードローン
タイトル 1	【No.1】カードローン比較サイト
タイトル 2	とにかく急ぎで借りたい方必見
タイトル 3	即日融資安心カードローン
説明文 1	周囲に内緒で借りるカードローン。24 時間スマホ完結カードローンランキング。免許証のみで申込 OK
説明文 2	初めての方も安心・コンビニ ATM で借入 OK なので便利・22 時迄の申込で最短即日融資が可能。
品質スコア	82.3

STS-B [20] などの一般的な文間類似度は意味の近さを評価するが、広告類似度では言及対象の商品の同一性（商品のカテゴリ、商品内容）と訴求の同一性（訴求軸、訴求内容）に関する評価を行う。例えば表 3 の例 1 の文ペアはどちらも同じ商品について言及しており、かつ訴求も同じであることから類似度は最も高い 5.00 である。一方で、例 2 では商品については同じカテゴリについて言及している可能性を持つが、訴求が異なるため類似度は 2.33 と低いことがわかる。より詳細なアノテーション方法については付録に示す。

2.5　広告効果予測

広告効果予測タスクは、表 4 に示すように、キーワードや広告文などの情報から広告そのものの品質スコアを推定し、その結果を正解スコアとのピアソン相関係数およびスピアマン相関係数で評価するタスクである。入力となる情報は大きく分けて 3 種類あり、「業種」「広告文」そして「キーワード」である。「業種」は株式会社サイバーエージェント内で定めた広告の業界種別であり、合計で 4 カテゴリ存在する。「広告文」は広告そのものを異なる表現や文長で述べたテキストであり、「タイトル」と「説明文」の 2 種類存在する。[4]「キーワード」は当該広告に関連する検索キーワードのことであり、同一広告文に対して複数存在することもある。表 4 の例では「カードローン」がキーワードとして設定されているが、同じ広告文で「カードローン 比較」といったキーワードの例も他に存在する。「品質スコア」は広告配信実績値をもとに算出した 1 から 100 の範囲を取る実数値で、高いほど高品質である。また、公開にあたって一部の情報を秘匿化加工している。秘匿化処理の詳細については付録で述べる。

4)　「タイトル」は全角で最大 15 文字で「説明文」は全角で最大 30 文字である。なお、「タイトル 3」と「説明文 2」は空であることも許容されている。

表 5　学習済み言語モデルによる AdGLUE の評価結果

モデル名	広告容認性	広告一貫性	訴求表現認識		広告類似性		広告効果予測	
	Acc.	Acc.	F1-micro	F1-macro	Pearson	Spearman	Pearson	Spearman
東北大 $\text{BERT}_{\text{BASE}}$	0.844	**0.889**	0.745	0.593	0.812	0.836	0.437	0.454
東北大 $\text{BERT}_{\text{BASE}}$ (char)	0.833	0.878	0.693	0.529	0.786	0.829	0.437	0.447
東北大 $\text{BERT}_{\text{LARGE}}$	0.849	0.888	0.763	0.641	0.828	0.845	**0.480**	**0.497**
早稲田大 $\text{RoBERTa}_{\text{BASE}}$	0.841	0.880	0.731	0.575	0.855	0.868	0.444	0.454
早稲田大 $\text{RoBERTa}_{\text{LARGE}}$	0.848	0.879	0.773	**0.699**	**0.898**	**0.892**	0.445	0.457
$\text{XLM-RoBERTa}_{\text{BASE}}$	0.850	0.888	0.742	0.602	0.818	0.853	0.425	0.439
$\text{XLM-RoBERTa}_{\text{LARGE}}$	**0.856**	0.886	**0.784**	0.646	0.856	0.865	0.453	0.457

3　AdGLUE を用いたモデル評価

本節では、自然言語処理における代表的な分類・回帰モデルを用いて、構築したベンチマークの評価実験を行う。実験結果を通して、提案ベンチマークの有用性や今後の課題について分析する。

3.1　実験設定

ベースラインモデルは日本語版 BERT [21]、RoBERTa [22]、多言語モデルとして XLM-RoBERTa [23] を用いて、以下のようにタスクの種類に応じて fine-tuning を行った。

- **広告容認性、広告一貫性** 特殊トークン `[SEP]` で広告文と LP を接続した文字列を入力として `[CLS]` トークンに対する分類問題を解く。
- **訴求表現認識** 村上ら [18] の Doc-Based Model と同じく広告文でマルチラベリング問題を解く。
- **広告類似性、広告効果予測** 広告類似性タスクでは文ペアを、広告効果予測タスクでは業種・キーワード・広告文を `[SEP]` で接続し、`[CLS]` トークンに対する回帰問題を解く。なお、広告効果予測タスクの秘匿化に使用された `[MASK_*]` トークンは語彙に追加して学習を行った。

検証用データが配布されていないデータセットは学習用データの 2 割を検証用データとした。ハイパーパラメータ等の実験の詳細は付録に示す。

3.2　結果と分析

表 5 に各言語モデルによる AdGLUE の実験結果を示す。モデルごとの性能評価を以下にまとめる。

- 広告一貫性タスクでは東北大 $\text{BERT}_{\text{BASE}}$ が最高値であったが、他との差は軽微であった。
- 全体的に RoBERTa、XLM-RoBERTa モデルが高い値である。要因として両モデルは大規模 Web クロールデータ CommonCrawl [24] を用いて学習している。CommonCrawl には砕けた表現の文章や広告文が多く含まれているため、広告文の特徴をうまく捉えることができ、結果的に Wikipedia のみで学習している BERT モデルよりも高い値となったと考えられる。
- 東北大 $\text{BERT}_{\text{BASE}}$ (char) は全体的に低く、広告文に対して文字単位の処理が不適切とわかる。
- 全体的に BASE モデルよりパラメータの大きい LARGE モデルのほうが値が高いことから、パラメータ数の増加は有効であることがわかる。
- 広告類似性タスクは AdGLUE 内では高いものの、日本語の文間類似度タスクである JSTS [10] と比べると低い。単純な比較はできないが、広告ドメインでの表現と一般的に使用される表現が異なることを示唆しており、より詳細な分析を必要である。
- 広告効果予測タスクでは Wikipedia のみで学習された東北大 $\text{BERT}_{\text{LARGE}}$ が最も高い値であり、CommonCrawl で学習された他モデルを上回っている。この原因まだ特定できておらず、これからの課題としたい。

また、今回実験に用いた学習済み言語モデル群は、学習データや単語分割などの条件が統一されていない。そのため、条件を統一した実験を行い、広告理解に最適な単語分割の粒度を明らかにすることが望ましいが、この点は今後の課題とする。

4　おわりに

本稿では広告言語理解ベンチマーク AdGLUE を提案し、タスク設計とデータセット構築について説明した。また、各タスクの性質や一般的な言語理解との違いを分析するため、構築ベンチマークで現在主要なアプローチでベースラインで実験した。実験結果から現状で調査不足な点や未網羅な課題点が発見され、それらを今後の改善につなげていきたい。

謝辞

本研究は株式会社サイバーエージェントと奈良先端科学技術大学院大学の共同研究により実施した。

参考文献

[1] 株式会社 CARTACOMMUNICATIONS, 株式会社 D2C, 株式会社電通, 株式会社電通デジタル. 2021 年 日本の広告費 インターネット広告媒体費 詳細分析, 2022. https://www.dentsu.co.jp/news/release/2022/0309-010503.html.

[2] Yanwu Yang and Zhai Panyu. Click-Through Rate Prediction in Online Advertising: A Literature Review. **Information Processing & Management**, Vol. 59, No. 2, p. 102853, 2022.

[3] 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. In **EMNLP 2022: Industry Track**, 2022.

[4] Haonan Li, Yameng Huang, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin, and Nan Duan. CULG: Commercial Universal Language Generation. In **NAACL-HLT 2022: Industry Track**, 2022.

[5] Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. An empirical study of generating texts for search engine advertising. In **NAACL-HLT 2021: Industry Papers**, pp. 255–262, 2021.

[6] 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.

[7] Taifeng Wang, Jiang Bian, Shusen Liu, Yuyu Zhang, and Tie-Yan Liu. Psychological advertising: Exploring user psychology for click prediction in sponsored search. In **KDD 2013**, p. 563–571, 2013.

[8] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In **ICLR 2019**, 2019.

[9] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. **Advances in neural information processing systems**, Vol. 32, , 2019.

[10] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese General Language Understanding Evaluation. In **LREC 2022**, 2022.

[11] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. KLUE: Korean Language Understanding Evaluation. In **NeurIPS 2021: Systems Datasets and Benchmarks Track**, 2021.

[12] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In **COLING 2020**, pp. 4762–4772, 2020.

[13] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised language model pre-training for French. In **LREC 2020**, pp. 2479–2490, 2020.

[14] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In **EMNLP 2020**, pp. 6008–6018, 2020.

[15] Jianfeng Chi, Wasi Uddin Ahmad, Yuan Tian, and Kai-Wei Chang. PLUE: Language Understanding Evaluation Benchmark for Privacy Policies in English. **arXiv preprint arXiv:2212.10011**, 2022.

[16] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. **ACM Transactions on Computing for Healthcare (HEALTH)**, Vol. 3, No. 1, pp. 1–23, 2021.

[17] Alex Warstadt and Samuel R. Bowman. Grammatical analysis of pretrained sentence encoders with acceptability judgments. **CoRR**, Vol. abs/1901.03438, , 2019.

[18] Soichiro Murakami, Peinan Zhang, Sho Hoshino, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. Aspect-based analysis of advertising appeals for search engine advertising. In **NAACL-HLT 2022: Industry Track**, pp. 69–78. ACL 2022, 2022.

[19] David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). **Pisani, R. Purves, 4th edn. WW Norton & Company, New York**, 2007.

[20] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In **SemEval-2017**, pp. 1–14, 2017.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT 2019**, pp. 4171–4186, 2019.

[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. **arXiv preprint arXiv:1907.11692**, 2019.

[23] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **ACL 2020**, pp. 8440–8451, 2020.

[24] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In **LREC 2020**, pp. 4003–4012, 2020.

[25] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, Vol. 76, No. 5, pp. 378—382, 1971.

[26] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. Transformers: State-of-the-art natural language processing. In **ACL 2020: System Demonstrations**, pp. 38–45, 2020.

表 6　実験に使用した学習済み言語モデルの一覧。CC は Common Crawl、Ja は日本語のデータを表している。

モデル名	PreTokenizer	Tokenize 単位	学習データセット	語彙数	パラメータ数
東北大 $\text{BERT}_{\text{BASE}}$	MeCab (IPADic+NEologd)	BPE	Wikipedia (Ja)	32K	111M
東北大 $\text{BERT}_{\text{BASE}}$(char)	MeCab (IPADic+NEologd)	文字	Wikipedia (Ja)	6K	90M
東北大 $\text{BERT}_{\text{LARGE}}$	MeCab (IPADic+NEologd)	BPE	Wikipedia (Ja)	32K	337M
早稲田大 $\text{RoBERTa}_{\text{BASE}}$	Juman++	Unigram LM	Wikipedia (Ja) + CC (Ja)	32K	110M
早稲田大 $\text{RoBERTa}_{\text{LARGE}}$	Juman++	Unigram LM	Wikipedia (Ja) + CC (Ja)	32K	336M
$\text{XLM-RoBERTa}_{\text{BASE}}$	–	Unigram LM	Multilingual CC	250K	278M
$\text{XLM-RoBERTa}_{\text{LARGE}}$	–	Unigram LM	Multilingual CC	250K	559M

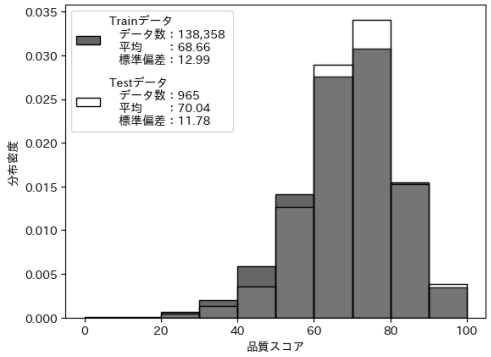

図 2　広告効果予測タスクの品質スコアの分布密度

表 7　広告容認性と広告一貫性のラベル分布

	広告容認性			広告一貫性		
ラベル	Train	Test	Total	Train	Test	Total
acceptable consistent	15,099	850	15,949	8,708	620	9,328
not_acceptable not_consistent	20,278	1,150	21,428	20,048	1,380	21,428
Overall	35,377	2,000	37,377	28,756	2,000	30,756

表 8　広告類似性のラベル分布。x は各値を表す。

分布レンジ	Train	Dev	Test	Total
$1 \leq x < 2$	527	66	67	660
$2 \leq x < 3$	845	105	108	1,058
$3 \leq x < 4$	2,739	343	344	3,426
$4 \leq x < 5$	790	99	100	989
$5 \leq x$	79	10	10	99
Overall	4,980	623	629	6,232

表 9　訴求表現認識のラベル付与数

ラベル付与数	Train	Dev	Test	Total
付与なし	337	94	84	515
1	769	198	165	1,132
2	485	98	100	683
3	182	44	47	273
4	69	26	10	105
5	10	5	3	18
6	4	0	1	5
合計	1,856	465	410	2,731

表 10　実験に使用したハイパーパラメータ。複数値があるものは最適値を学習データのみで探索し選択している。

ハイパーパラメータ名	値
Lerning Rate	2e-5, 5.5e-5, 2e-6
seed	0
Epoch	30
Early Stopping	3
Optimizer	Adam
Max Sequence Length	128

A　付録

アノテーションガイドライン　アノテーションは図 3 のように商材と訴求の 2 軸での一致度合いで判断される。村上ら [18] の訴求タイプの定義に基づいたアノテーションを依頼した。仕様書には各訴求タイプごとに説明文と例文を提示している。商材や訴求表現が一部存在しない広告文は、それ以降の要素が広告文に含まれていれば該当表現が存在したとして扱い、含まれていなければ存在しなかったとして扱う。広告文内の固有名詞などについて不明の場合は適宜 Web 検索を許可した。また、データセットは 3 人の非専門家のアノテータによって作成されており、2 回の訓練フェーズを経て本番データでアノテートし、アノテータ間の一致度合いは fleiss' kappa [25] で $\kappa = 0.51$ であった。

	商材		訴求	
選択肢	カテゴリの一致	商材の一致	訴求軸の一致	訴求内容の一致
1: 全く異なる	FALSE	FALSE	FALSE	FALSE
2: あまり似てない	TRUE	FALSE	FALSE	FALSE
3: ちょっと似てる	TRUE	TRUE	FALSE	FALSE
4: 似てる	TRUE	TRUE	TRUE	FALSE
5: ほとんど同じ	TRUE	TRUE	TRUE	TRUE

図 3　アノテーションの方法。左の要素から順に一致しているごとにスコアが 1 段階ずつ上がっていく。

広告効果予測のマスキング処理　直接的なクライアント名・サービス名のマスキングを行うことを条件に広告主からデータセットとして公開することを許可された。例文「[MASK_3] の仕事紹介サイト」ではクライアント名が [MASK_3] に置き換えられている。マスキング処理は 2 人のアノテータが手分けして行った。マスキング対象となったクライアント・サービス名は 24 個であり、それぞれ [MASK_0] から [MASK_23] に置換した。

各データセットのラベル分布　表 7 に広告容認性、広告一貫性のラベル分布、表 8 に広告類似性のラベル分布、表 9 に訴求表現認識のラベル付与数、図 2 に広告効果予測の品質スコアの分布密度を示す。

実験設定の詳細　各学習済み言語モデルの詳細を表 6、実験に用いたハイパーパラメータを表 10 に示す。表 10 に記載されていないパラメータは各言語モデルまたは Transformers の標準設定を使用している。実験は Hugging Face 社の Transformers [26] を使用した[5)]。

5)　実装は Transformers の AutoModelForSequenceClassification に修正を加えている。Transformers のバージョンは 4.25.1。