FaithCAMERA: 広告文生成タスクのための忠実性を担保した評価データセットの構築

加藤 明彦　三田 雅人　村上 聡一朗
本多 右京　星野 翔　張 培楠

株式会社 サイバーエージェント
{kato_akihiko, mita_masato, murakami_soichiro,
honda_ukyo, hoshino_sho, zhang_peinan}@cyberagent.co.jp

概要

広告文生成 (*ad text generation [ATG]*) において、望ましい広告文は、入力に忠実であると同時に、潜在顧客にアピールする重要な情報を含む、すなわち、情報性に優れている。既存の評価データである CAMERA [1] は、広告制作者が作成した参照文からなり、情報性の評価に適している。しかし、これらの参照文には **入力に忠実でない**情報が含まれていることが多く、ATG の研究を推進する上で顕著な障害となっている。そこで本研究では広告制作者と協力して CAMERA の参照文を修正し、忠実性を保証した新たな ATG の評価データセット (FaithCAMERA) [1)] を構築した。また、既存の忠実性向上手法が、忠実性を維持しながら情報性に優れた広告文を生成できるかどうかを評価した。

1　はじめに

オンライン広告の普及により、広告を大量に制作する需要が高まっているため、広告文の自動生成 (*ad text generation [ATG]*) に焦点を当てた研究への関心が高まっている [2, 3, 4, 5]。ここで我々は ATG を、ランディングページ (LP) などの入力文書と検索クエリなどのユーザー信号を入力として、訴求力のある広告文を出力するタスクとして定義する (表 1)。 ATG を NLP タスクとして初めて明示的に定義した Mita ら [1] によれば、広告文は入力文書に忠実で (忠実性[2)])、かつ、潜在顧客にアピールする重要な情報を含むべきである (情報性)。ATG 分野で初めて公開されたベンチマークデータセットである CAMERA [1] と、より多様な訴求を保証するその変種である CAMERA^3 [7] は、広告制作者によって作成された参照文で構成されているため、情報性の評価に適している。しかし、広告制作者は魅力的な広告文を作るために、外部知識や常識的知識に基づいて、入力では述べられていない情報を広告文に盛り込むことがある。そのため、これらの参照には、入力に現れないエンティティなどの**不忠実な**情報が含まれていることが多い (表 1-(c))[3)]。

表 1　検索連動型広告の例。各色で強調された部分は各々「忠実だが情報性に乏しい」、「事実だが忠実ではない」、「事実でなく忠実でもない」スパンを示す。なお、(d) に記載されている数量 (“30%”) は、ランディングページに記載されている数量 (“20%”) とは異なっている点に注意されたい。

検索クエリ	UV カット メガネ
ランディングページ	.. UV から目を守るメガネ.. .. メガネの UV カット率が 20% 向上.. .. All rights reserved...
広告文	(a) UV カットメガネで目を守る (b) メガネ | *All rights reserved.* (c) 有害な 紫外線から目を守る (d) メガネの UV カット率が *30%* 向上!

参照文の忠実性が保証されていない場合、生成された広告文が情報性と忠実性の両面で優れているかどうかを、参照ベースの単一の指標で評価することはできない。したがって、忠実性と情報性の双方を保証する ATG モデルの開発は、現状では困難である。忠実性が保証された評価データがないことは、実世界への適用性の高い ATG 研究を推進する上で顕著な障害となっている。この問題を解決するために、我々は社内の広告制作者と協力し、CAMERA の

1)　https://github.com/CyberAgentAILab/FaithCAMERA
2)　忠実性の定義は先行研究によって若干異なるが、本研究では [6] に従い、広告文が入力文書で意味的に支持されている場合、入力に忠実である、とする。
3)　Mita ら [1] は、出力側にのみ現れる n-gram を使って、この傾向を定量的に実証している。

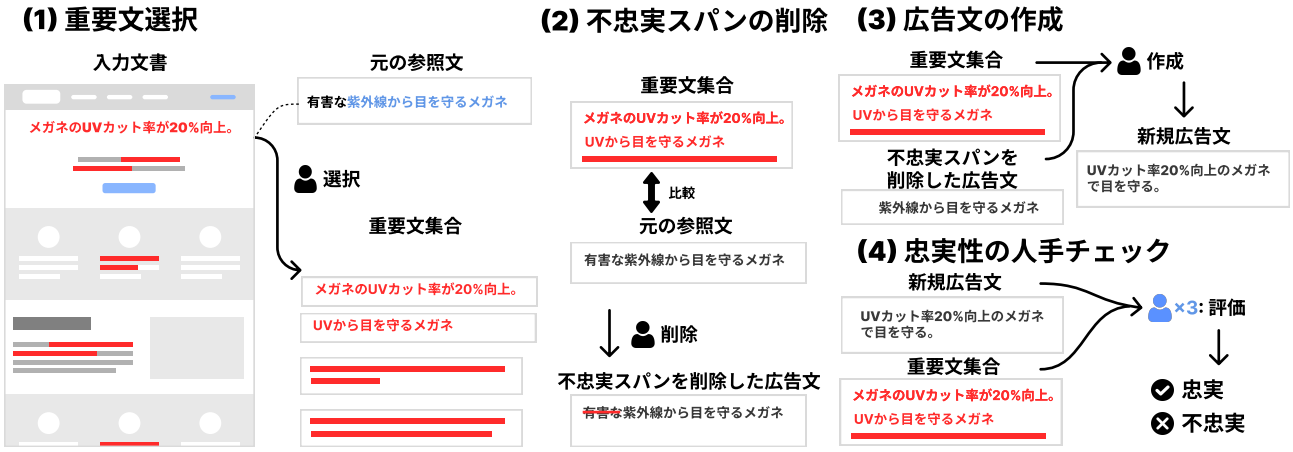

図 1　データセット構築の流れ

評価セットの参照文を修正することで、情報性を確保しつつ忠実性を保証した日本語の ATG 評価データセット (FaithCAMERA) を構築した [1)]。

我々は FaithCAMERA を用いて、既存の忠実性向上手法が、忠実性を維持しながら情報性に優れた広告文を生成できるか、評価した。実験の結果、 Loss Truncation [8] は、エンティティレベルと文レベルの双方で忠実性と情報性を向上させることが分かった。一方、入力に忠実でない出力を持つ学習事例のフィルタリング [9] は、エンティティレベルでは忠実性と情報性を向上させるが、文レベルでは双方を低下させることが分かった。

2　FaithCAMERA の構築

本研究では CAMERA の評価セットの参照文を修正することで、入力に忠実な参照文を持つ評価用データセットを構築することを目的とする。この目的を達成するためには、以下の 2 つのアプローチが考えられる: (a) CAMERA の参照文を忠実か否かで 2 値分類し、前者に該当する事例群のみを抽出する (フィルタリングアプローチ)、または (b) CAMERA の参照文を編集する (編集アプローチ)。本研究では、フィルタリングアプローチでは十分な量のテストデータが確保できないこと (§2.2)、またデータ分布が偏る可能性があることを懸念し、編集アプローチを採用する。実際、編集アプローチは、data-to-text [10] や対話 [11] など、他の自然言語生成タスクの関連研究でも採用されている。

2.1　データソース

CAMERA [1] の評価セットを元データとして使用する。各テスト事例は入力 x と参照広告文 y からなる。x は、(1) 検索クエリ、(2)LP の説明文、(3)OCR 処理された LP 本文で構成される。y は、配信済の 1 つの参照広告文と、広告制作経験を持つアノテーターによって作成された 3 つの追加の参照広告文で構成される。FaithCAMERA でも上記 (1)-(3) を入力とするが、修正対象の参照文としては、配信済広告文のみを利用する。これは、モデルの情報性や多様性に焦点を当てた評価を行いたい場合は CAMERA を、情報性を担保しつつモデルの忠実性に焦点を当てた評価を行いたい場合は FaithCAMERA を用いることを意図しているためである [4)]。

2.2　構築手順

本研究で実施したデータセット構築手順は 4 つのステップから構成される (図 1)。アノテーションは、社内の広告制作経験者が行った。以下のステップ (1)-(3) では、各事例に 1 名のアノテーターを割り当て、ステップ (4) では、参照文の忠実性を手動で確認するため、各事例に 3 名のアノテーターを割り当てた。以下、アノテーションの各ステップについて説明する。

(1) 重要文選択　入力文書から、 CAMERA の参照文を意味的に支持する一連の文を選択する (図 1-(1))。

(2) 不忠実なスパンの削除　重要文集合に含まれない情報 [5)] を持つスパン (以下、**不忠実スパン**) を元の参照文から削除する (図 1-(2))。不忠実スパンが存在しない場合 (872 事例中 199 事例) は、以下のス

4)　すなわち、 FaithCAMERA 構築の目的は「CAMERA を置き換え可能なデータセット」を作成することではなく、CAMERA を補完する評価用データセットを研究コミュニティに提供することである。

5)　表 1-(c) 中の「事実だが忠実ではない」スパンが該当し、*factual hallucination* [12] とも呼ばれる。

テップ (3) と (4) を省略し、元の参照文を入力に忠実な参照文として採用する。

(3) 広告文の作成　アノテーターは、不忠実スパンが削除された参照文に基づいて、重要文集合を参照しながら、入力に忠実な新しい参照文を作成する (図 1-(3))。アノテーターには以下 2 点の指示を行った: (a) 重要文集合に含まれている商品名や広告訴求を、可能であれば新たな参照文に盛り込む、(b) ステップ (1) で重要文が選択されなかった事例 (872 事例中 13 事例) においては、入力情報全体を重要文集合として扱う。

(4) 忠実性の人手チェック　ステップ (3) で作成された新しい参照文が入力に忠実になっているかどうかを、3 名のアノテーターが、重要文集合を参照しながら評価する (図 1-(4))。各事例で忠実性評価を担当するアノテータ 3 名は、ステップ (1)-(3) で当該事例の広告文を編集したアノテータ 1 名とは異なる。多数決によって、忠実ではないと評価された事例はステップ (3) に差し戻して、再編集を行った。これらの手順は、すべての事例が忠実であると評価されるまで繰り返し実施された。

CAMERA の参照文 872 事例の内、673 事例 (77.1%) が入力に忠実ではないことを考慮すると、FaithCAMERA は CAMERA と比較して忠実性の面で大幅に改善されていると言える。また、人間が生成した参照文には潜在顧客に訴える重要な情報が含まれているという前提に立つと、参照文を意味的に支持するという基準でステップ (1) で選択された重要文には、重要な情報が含まれていることが期待できる。したがってステップ (3) では、重要文集合に基づいて、忠実かつ情報性に優れた新しい参照文が作成される。

忠実性判定のアノテーションは、アノテーターによってどの程度異なるのか？　忠実性を確保するために修正を行った 673 事例のうち、ステップ (4) において、671 事例 (99.7 %) についてはアノテーター全員が、残り 2 事例においてはアノテーター 3 名中 2 名が、修正された参照広告文は入力に忠実であると判定した。

2.3　FaithCAMERA の抽出度

上記の構築手順を採用した結果、FaithCAMERA が過度に抽出的になり、忠実で情報性に優れた広告文を生成するタスクが FaithCAMERA において容易になり過ぎている可能性が懸念される。そこで我々は FaithCAMERA の抽出度を調査した。その結果、FaithCAMERA は CAMERA よりも抽出的ではあるが、過度に抽出的にはなっていないことがわかった。詳細は付録 A を参照されたい。この知見は、FaithCAMERA において忠実で情報性に優れた広告文を生成するには、単純な抽出的要約では不十分であることを示唆している。

3　FaithCAMERA を用いた忠実性向上手法の評価

本節では構築した FaithCAMERA を用いて、既存の忠実性向上手法 (§3.2) が、忠実性を維持しながら情報性に優れた広告文を生成できるか、調査する。生成広告文で **エンティティレベルの忠実性**が保証されていない場合、商業的なリスクが大きくなるため、忠実性の評価にあたっては、主にエンティティレベルの忠実性に着目する [6]。

3.1　ベースライン

ATG においても他の NLP タスクと同様に、大規模言語モデル (LLM) を利用するアプローチが標準的になりつつある [14] ため、事前学習済モデルに対して、CAMERA の学習セットを用いて指示チューニングを行ったモデルをベースラインとする。事前学習済モデルに関しては、(1) 日本語データを用いた継続事前学習を行なっていることから、日本語のテキスト生成タスクで高い性能が期待される、(2) 実験環境および実験コストの制約を考慮すると、オープンかつ 8B サイズ程度のモデルが望ましい、という理由から、Llama-3.1-Swallow-8B-Instruct-v0.2[7] を選定した。指示チューニングに用いたプロンプトを付録 B、その他の詳細を付録 C に示す。

3.2　忠実性向上手法

Data filtering (DF) [9]　入力に忠実でないエンティティが出力に含まれる学習事例を学習データから除去する手法である。具体的には、学習データから固有表現[8]、用語[9]、カタカナからなる語句[10]、

6)　実際、製品名 (例：*iPhone 15 Plus*) のような固有表現や、割引価格 (例：*50% OFF*) のような数値表現は、広告で効果的な訴求を行うために使用されている [13]。
7)　https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2
8)　GiNZA(https://megagonlabs.github.io/ginza/) を用いた。
9)　pytermextract(http://gensen.dl.itc.u-tokyo.ac.jp/pytermextract) を用いた。
10)　正規表現を用いた。

表 2　実験結果：最良の結果を**太字**で示す。シード値に応じて訓練データの提示順序が変わるため、抽出型要約（BM25）以外のモデルの各値は、異なるシードを用いた 3 回の実行の平均と標準偏差である。BM25 はニューラルネットワークベースの手法ではなく、ランダム性を持たないため、1 回の実行の結果を示す。

Models	prec_s	prec_t	ROUGE-L
Baseline	70.5 ± 0.9	24.6 ± 0.6	30.5 ± 0.4
Loss Truncation (LT)	75.8 ± 0.3	24.9 ± 0.2	**30.9** ± 0.2
Data Filtering (DF)	89.1 ± 1.9	27.2 ± 1.9	27.3 ± 0.9
DF+LT	89.0 ± 0.2	**27.9** ± 0.5	28.3 ± 0.7
BM25	**96.3**	9.5	11.4

時間表現[11]、数値表現[12]のいずれかに属するエンティティを抽出し、それらのエンティティが出力側のみに現れる学習事例を除外した。その結果、学習データは 12,395 事例から 5,878 事例に減少した。この学習データを用いて指示チューニングを行った。

Loss Truncation (LT) [8]　学習中の勾配計算において、ミニバッチ内の高い loss を持つ事例を動的に切り捨てる手法である。その他の詳細は付録 C に示す。

3.3　評価指標

エンティティレベルの忠実性と情報性は、precision-target(prec_t) [9] で評価する。これは参照ベースの評価指標であり、モデル出力中のエンティティの内、参照文にも含まれるものの割合である。忠実性のみに着目した評価では、precision-source(prec_s) [9] を用いる。これは参照によらない評価指標であり、モデル出力中のエンティティの内、入力文書にも出現するものの割合である。抽出対象のエンティティタイプと抽出方法は、[1] に従う。また、モデル出力と参照文の n-gram の重複に基づく評価指標である ROUGE-L [15] も用いる。FaithCAMERA は参照文の忠実性と情報性を保証しているため、ROUGE-L が高い広告文は、忠実性と情報性の両面で優れている。ROUGE-L を用いることで、エンティティレベルの忠実性と情報性だけでなく、エンティティの修飾句やエンティティ間の関係を、評価時に考慮することができる。

3.4　結果・考察

まず、 Loss Truncation は、prec_t で 0.3 ポイント、ROUGE-L で 0.4 ポイント、ベースラインを上回った。また、データフィルタリングほどではないものの、ベースラインと比較して prec_s を 5.3 ポイント改善した。これらの結果は、学習中に外れ値を勾配計算から動的に切り捨てることによって、忠実性と情報性が、エンティティレベルと文レベルの双方で向上したことを示唆している。

次に、データフィルタリングは、prec_t で 2.6 ポイント、prec_s で 18.6 ポイント、ベースラインを上回った (表 2)。一方、ROUGE-L はベースラインよりも 3.2 ポイント低い。これらの結果は、データフィルタリングによって、エンティティレベルでは忠実性と情報性が向上するが、文レベルでは双方が低下することを示唆している。これは、データフィルタリングによって学習データの量が約半分に減少したことが原因であると考えられる。

第 3 に、データフィルタリングと Loss Truncation を組み合わせることで (DF+LT)、データフィルタリング単体に比べて prec_s は微減したものの、prec_t は 0.7 ポイント、ROUGE-L は 1.0 ポイント向上した。DF+LT ではデータフィルタリングの影響が大きく、Loss Truncation に比べると ROUGE-L は 2.6 ポイント、低下した。

さらに我々は FaithCAMERA における抽出型要約の有効性を検証し、 FaithCAMERA に単純な抽出型要約手法を適用しても、忠実かつ情報性に優れた広告文を生成するには不十分であるという仮説 (§2.3) を裏付ける結果を得た (付録 D を参照)。

4　結論

我々は、忠実性と情報性を両立する広告文生成という目標がどの程度達成されたかを適切に評価するために、忠実性を保証した日本語の評価用データセット (FaithCAMERA) を構築した。実験の結果、以下の知見が得られた: (1) Loss Truncation [8] は、エンティティレベルと文レベルの双方で忠実性と情報性を向上させる、(2) データフィルタリング [9] は、忠実性と情報性をエンティティレベルで向上させる一方、文レベルでは双方を低下させる。今後の課題として、(1) ルールベースや LLM を用いたデータ合成 [16] によって、忠実な出力を持つ学習データを十分量確保する、(2) 忠実性や情報性に関する選好データを用いた LLM のアライメント [17] が挙げられる。

11）　ja_timex(https://github.com/yagays/ja-timex)
12）　pynormalizenumexp(https://pypi.org/project/pynormalizenumexp) を用いた。

参考文献

[1] Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. Striking gold in advertising: Standardization and exploration of ad text generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Annual Meeting of the Association for Computational Linguistics**, pp. 955–972, Bangkok, Thailand, August 2024.

[2] Atsushi Fujita, Katsuhiro Ikushima, Satoshi Sato, Ryo Kamite, Kouji Ishiyama, and Osamu Tamachi. Automatic generation of listing ads by reusing promotional texts. In **International Conference on Evolutionary Computation**, 2010.

[3] John 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 & Data Mining**, 2019.

[4] Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. An empirical study of generating texts for search engine advertising. In Young-bum Kim, Yunyao Li, and Owen Rambow, editors, **the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers**, pp. 255–262, Online, June 2021.

[5] 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 Wanxiang Che and Ekaterina Shutova, editors, **Empirical Methods in Natural Language Processing**, pp. 191–199, Abu Dhabi, UAE, December 2022.

[6] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Annual Meeting of the Association for Computational Linguistics**, pp. 1906–1919, Online, July 2020.

[7] Go Inoue, Akihiko Kato, Masato Mita, Ukyo Honda, and Peinan Zhang. CAMERA^3: An evaluation dataset for controllable ad text generation in Japanese. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **International Conference on Language Resources and Evaluation**, pp. 2702–2707, Torino, Italia, May 2024. ELRA and ICCL.

[8] Daniel Kang and Tatsunori B. Hashimoto. Improved natural language generation via loss truncation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **the Annual Meeting of the Association for Computational Linguistics**, pp. 718–731, Online, July 2020.

[9] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. Entity-level factual consistency of abstractive text summarization. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, **the European Chapter of the Association for Computational Linguistics**, pp. 2727–2733, Online, April 2021.

[10] Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. ToTTo: A controlled table-to-text generation dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Empirical Methods in Natural Language Processing**, pp. 1173–1186, Online, November 2020.

[11] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. FaithDial: A faithful benchmark for information-seeking dialogue. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 1473–1490, 2022.

[12] Meng Cao, Yue Dong, and Jackie Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Annual Meeting of the Association for Computational Linguistics**, pp. 3340–3354, Dublin, Ireland, May 2022.

[13] Soichiro Murakami, Peinan Zhang, Sho Hoshino, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. Aspect-based analysis of advertising appeals for search engine advertising. In Anastassia Loukina, Rashmi Gangadharaiah, and Bonan Min, editors, **North American Chapter of the Association for Computational Linguistics**, pp. 69–78, Hybrid: Seattle, Washington + Online, July 2022.

[14] Soichiro Murakami, Sho Hoshino, and Peinan Zhang. Natural language generation for advertising: A survey. **ArXiv**, Vol. abs/2306.12719, , 2023.

[15] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text summarization branches out**, pp. 74–81, Barcelona, Spain, July 2004.

[16] John Chung, Ece Kamar, and Saleema Amershi. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 575–593, Toronto, Canada, July 2023. Association for Computational Linguistics.

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in neural information processing systems**, Vol. 35, pp. 27730–27744, 2022.

[18] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, **North American Chapter of the Association for Computational Linguistics**, pp. 708–719, New Orleans, Louisiana, June 2018.

[19] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[20] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. **Found. Trends Inf. Retr.**, Vol. 3, No. 4, p. 333–389, apr 2009.

A FaithCAMERA の抽出度

§2 節で述べた構築手順により、 FaithCAMERA の広告文は忠実かつ情報性に優れることが保証されているが、その結果、過度に抽出的な問題設定になるのではないかという懸念がある。つまり、“ FaithCAMERA の広告文は完全に抽出的なのか、それとも、入力文書中のフレーズを並べ替えたり言い換えたりすることで、ある程度、抽象的になっているのか？” という疑問が生じる。もし前者が正しければ、単純な抽出型要約アプローチによって、入力に忠実で参照文と類似した広告文が生成されることになる。この疑問に答えるため、文書要約で一般的に使用される抽出度の指標である *coverage* と *density* [18] を使用して、FaithCAMERA の広告文がどの程度抽出的なのかを調査した。ここでは [18] にならい、入力文書と広告文で共通するトークン列を抽出フラグメントと呼ぶ。coverage は、広告文中の単語が抽出フラグメントの集合によってどの程度カバーされているかを表し、density は抽出フラグメントの平均長である。その結果、 FaithCAMERA は CAMERA よりも双方の指標で高い値を持つことが分かった (表 3)。一方で、 FaithCAMERA の広告文の平均長が 14.7 トークンで density の平均が 3.16 であることを考えると、 FaithCAMERA は過度に抽出的なものではないことが分かる。この主張を裏付ける他の証拠として、(1) FaithCAMERA の広告テキストにおける bi-gram の 38.4% は入力側には現れない、(2) FaithCAMERA には “12 時間”→“半日” といった多くの言い換えが含まれている、という点が挙げられる。

表 3 各コーパスの抽出度に関する統計量

	FaithCAMERA	CAMERA
Coverage	0.97 ± 0.10	0.81 ± 0.22
Density	3.16 ± 2.09	2.07 ± 2.15

B 指示チューニングに用いたプロンプト

以下は、タスクを説明する指示です。要求を適切に満たす応答を書きなさい
指示:
与えられた入力 (検索クエリ、ランディングページの説明文、ランディングページの本文) をもとに、ランディングページに記載されている商品・サービスの魅力をユーザーにアピールする広告文を全角 15 文字 (半角 30 文字) 以内で作成してください。ここで検索クエリとはユーザーが検索に利用する文字列であり、広告文は検索クエリと関連する内容であることが望ましいです。
入力:
検索クエリ: {keyword}
ランディングページの説明文: {lp_meta_description}
ランディングページの本文: {lp_sentences}
応答:

C 実装詳細

指示チューニングでは、開発セットの loss を基準とした EarlyStopping を採用し、patience は 1 とした。バッチサイズは 2 とした。学習の最大エポック数は基本 3 としたが、学習データが約半分であるデータフィルタリングについては 6 とした。また、メモリ使用量を削減するために QLoRA [19] を用いた。学習には A100 40GB 1 枚を用いた。最大入力長と最大出力長はそれぞれ 2048 トークン、30 トークン [13] とした。デコード手法としては greedy decoding を用いた。 Loss Truncation の実装にあたっては論文 [8] の公式実装 [14] を利用した。論文 [8] に従い、dropc は 0.4 とした。

D FaithCAMERA における抽出型要約の有効性

我々は FaithCAMERA における抽出型要約の有効性を BM25 [20] を用いて検証した。具体的には、キーワードをクエリとして、LP の説明文と LP 本文から、広告文の最大長を超えない範囲で、BM25 スコアの降順に文を選択して広告文を作成した。その結果、BM25 は、prec_s では他の手法を少なくとも 7.2 ポイント上回るが、ベースラインと比較すると、prec_t では 15.1 ポイント、ROUGE-L では 19.1 ポイント下回った (表 2)。この結果は、 FaithCAMERA に単純な抽出型要約手法を適用しても、忠実かつ情報性に優れた広告文を生成するには不十分であるという仮説 (§2.3) を裏付けるものである。

13) Google レスポンシブ検索広告のガイドライン (https://support.google.com/google-ads/answer/7684791?hl=ja) に従い、広告見出しの最大長を設定した。
14) https://github.com/ddkang/loss_dropper