広告データセットに内在する幻覚の分析

加藤 明彦 [1]　大田 和寛 [1]　村上 聡一朗 [1]
三田 雅人 [1] 本多 右京 [1]　張 培楠 [1]

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

概要

Encoder–decoder 型の抽象型要約に基づく広告文生成モデルは, テスト時に入力情報と矛盾する, 事実でない広告文を生成することがある. その要因として, 入力に含意されない情報 (幻覚) を持つ広告文がデータセット中に含まれることが挙げられる. 広告文の事実性を向上させるためには, データセット中の幻覚を考慮したデータ編集やモデル学習上の工夫を行う必要がある. 高性能な幻覚検出器を構築するためには, 広告データセット中の幻覚の傾向を把握しておくことが望ましいが, 検索連動型広告データセット中の幻覚の分析はほとんど行われていない. このため本研究では, 検索連動型広告のデータセットについて幻覚の有無とタイプのアノテーションを付与し, 分析を行った.

1　はじめに

広告文は, 広告対象の商品やサービスについて述べた入力情報をもとに作成される. 広告文が満たすべき重要な性質として, 入力情報に矛盾していない, という点が挙げられる. 広告文が入力情報に含意される場合, 入力に忠実な広告文である (忠実性 [1, 2])(図 1). また, 入力情報に含意されない情報を含むが, それらが世界知識や常識的知識などの外部知識に基づいている場合, 事実である広告文である (事実性 [1, 3]).

広告文の自動生成モデルには, 事実性に優れた広告文生成が求められるが, 現在, 広範に用いられている encoder–decoder 型の抽象型要約モデル [5] で広告データセットを学習すると, テスト時に入力情報と矛盾する, 事実でない広告文が生成されることがある. このように, 入力に含意されない情報が出力側に現れる現象を幻覚 [1] と呼ぶ. モデルが幻覚を含む広告文を生成する 1 つの要因は, 入力に忠実ではないが事実である広告文 (事実である幻覚 (図 1) [4]) を広告データセットがしばしば含むことにある. その理由としては, 例えば以下が挙げられる: (1) ライターが外部知識に基づいて入力にない訴求表現を追加することがある [1) 2)], (2) 検索連動型広告をクリックした際に表示されるウェブページ, 即ちランディングページ (以下 LP) のスクリーンショットの OCR エラーのために, 入力の一部欠損やノイズ混入が発生し得る.

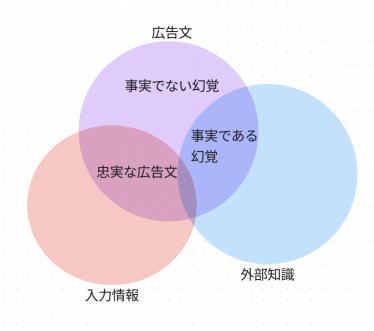

図 1　広告文の忠実性と事実性に関するベン図 ([4] 中の図を元に改変).

事実でない広告文の生成頻度を低減するためには, (A) 広告データセット中の幻覚を検出し, (B) 検出した幻覚を考慮したデータ編集やモデル学習上の工夫を行う必要がある. (A) の検出器を構築する上では, 広告データセット中にどのような幻覚がどの程度, 存在しているのかを把握しておくことが望ましい. 一般の文書要約タスク [1] やスローガン生成タ

1)　ここで訴求表現とは, 消費者の注意を惹き, 広告対象商品の購入を促すために用いられる表現を指す.

2)　上記 (1) の例を挙げると,「メガネの UV カット率が向上」と記載されているランディングページに対して,「UV カット率が向上! 長時間の PC 作業に」という広告文をライターが作成した場合,「UV カット機能を持つメガネは PC 作業に利用され得る」という外部知識を利用して商品の有用性に関する訴求表現を追加している, と解釈することができる.

表 1 入出力フォーマットと各フィールドの例. 表中の [NE] は, 固有表現の秘匿化処理を行なっていることを表す.

大分類	小分類	例
(1) 入力情報	(a) キーワードリスト	[NE] 市 介護 保険 料
	(b) ランディングページ (LP)	[NE] 保険相談キャンペーン実施中! 最新! 2022 年 12 月版..
	(c) LP の説明文	今おすすめの人気の介護保険ランキングを発表！...
(2) 広告文	見出しアセット	国内最大級の保険比較サイト
	説明文アセット	人気の介護保険をランキング形式で比較。 介護保険の相談...

スク [6] ではデータセットに含まれる幻覚の分析が行われているが, 検索連動型広告のデータセットに含まれる幻覚の分析は, ほとんど行われていないのが現状である. そこで本研究では (A) に向けた準備段階として, 日本語の検索連動型広告データセットに含まれる幻覚の分析に取り組む.

具体的には, 株式会社サイバーエージェントで扱っている日本語の広告データセットから抽出した 360 事例について, 幻覚の有無とタイプに関するアノテーションと分析を行った. その結果, 入力には出現しない広告文中のフレーズの内, 少なくとも 8 割は入力を単に要約しただけでは出現し得ないフレーズ, 即ち幻覚であることが明らかとなった. また, 幻覚の 70%以上が訴求表現に該当するという結果が得られた.

2 データ分析手法

本節ではデータ分析の手法について述べる. 詳細は各小節に譲るが, 主な狙いとしては, データセット全体の幻覚の傾向を把握するために, できるだけ多様な広告文をサンプルし (2.1), 各広告文から網羅的にフレーズを抽出して (2.2), 幻覚の有無とタイプをアノテーションした (2.3).

2.1 事例抽出

事例のフォーマット 本研究では表 1 の入出力フォーマットに従う 22,728 事例に対して, 後述のサンプリングを行った. 広告文生成における一般的な入出力フォーマットは表 1 に示すように, (1) 入力情報と (2) 広告文 からなる. (1) は (a) キーワードリスト, (b) LP, (c) LP の説明文, から構成される. 上記 (b) は LP のスクリーンショットに対する OCR によって取得し [3], (c) は LP の HTML 中の `meta` タグの要素 `description` から取得している. 一方, (2) は広告の見出しを構成するフレーズ (見出しアセット [4]; headline) と広告の説明文を構成するフレーズ (説明文アセット; description) から構成される. 見出しアセットは各入力に対して 3 ~ 15 個, 説明文アセットは 2 ~ 4 個存在する. 本研究では見出しアセット (半角 30 文字) を対象に分析を行う. これは説明文アセット (半角 90 文字) に比べて表示領域が狭いため, 略語や言い換えなどがより多く用いられ, 幻覚が発生し易いと考えられるためである.

事例のサンプリング 各クライアントからバランス良く事例をサンプルするために, 1 クライアントあたり 5 種の広告キャンペーンを, そして各広告キャンペーンについて 5 つの LP をサンプルした. また, 文長が短すぎる広告文を取り除くために, 各入力に紐づく見出しアセットの内, 7 文字以上のものだけを残した.

広告データセットには類似した広告文が多く含まれるため, ランダムサンプリングを行うと, 分析対象の事例集合の多様性が確保できないという問題がある. そこで本研究では, データセット全体としての幻覚の傾向を把握するために, できるだけ多様な広告文を分析対象とすることを狙い, 文ベクトルを用いたサンプリングを行った. 文ベクトル生成手法としては, 日本語を含む多言語モデルであり, 下流タスクで高い性能が報告されている LaBSE [5] [7] を採用し, 候補集合を 1 文ずつ拡張する方式を採用した. 具体的には, 最初の 1 文をランダムに選択し, その後は候補集合中のいずれの文との cosine 類似度も閾値 (0.5) 以下であるという条件を満たす 1 文を候補集合に加えるという手順とした. この結果得られた 360 種の広告文と, 対応する入力情報を対象に分析を行った.

3) Google cloud vision API(https://cloud.google.com/vision/docs/ocr) を利用した.

4) Responsive Search ads (RSA) においてはこの様に呼称される. 詳細は以下を参照されたい: https://support.google.com/google-ads/answer/7684791?hl=en&ref_topic=10284269

5) https://huggingface.co/sentence-transformers/LaBSE

表 2 幻覚有無アノテーションの例. 広告文側にのみ出現する 574 フレーズに対し,「この語句は入力情報 (キーワード, LP) の要約に含まれ得るか?」という質問に対する回答を選択する形式で幻覚有無のアノテーションを行った.

選択肢	理由	入力情報	広告文	#	%
はい	-	結婚式・披露宴	結婚式	99	17.2
いいえ	入力に類似語句は存在しない	-	-	430	74.9
いいえ	入力に類似語句は存在するが意味は異なる	約 30 分で視聴可能	即日	36	6.3
いいえ	入力に類似語句は存在するが数量表現違い	1,300 円	1,400 円	5	0.9
いいえ	入力に出現していない略称	フランチャイズ	FC	4	0.7

2.2 広告文からのフレーズ抽出

以下の手順により, 各事例中の広告文から網羅的にフレーズを抽出し, 1,173 フレーズを得た.

(1) Spacy[6] を介して提供されている GiNZA v5.1 [8] によって, 広告文に対して基礎解析を行った. これにより単語分割, 品詞タグ付け, 依存構造解析, 固有表現抽出 が行われる.

(2) 上記 (1) の基礎解析結果を利用して各文から名詞句以外のフレーズ (動詞句など) を抽出した. 具体的には主辞の品詞が NOUN, PROPN, PRON 以外である文節を抽出している.

(3) 上記 (1) の基礎解析結果を利用して各文から名詞句を抽出した. 具体的には (a) 固有表現, (b) 複合語, (c) 上記 (a)(b) に含まれない単一トークンの名詞, の和集合を名詞句とした. (a) については, GiNZA で抽出した固有表現を候補として, 人手で固有表現の範囲エラーを修正した[7]. (b) については, 最終トークンがフレーズに属する他のトークンの dependency head になっていて, 依存関係ラベルが compound または flat であるフレーズを抽出した. ただし上記手順で得られたフレーズが固有表現・数量表現を内包する場合には, 固有表現・数量表現とその前後でフレーズを分割した. 抽出したフレーズの具体例を付録 A の表 5 に示す.

2.3 アノテーション

2.2 節で抽出した 1,173 フレーズの内, 出力側にだけ出現する 574 フレーズに対し,「この語句は入力情報 (キーワード, LP) の要約に含まれ得るか?」という質問に対して, 表 2 に示す 5 つの選択肢から回答を選択する形式で幻覚有無のアノテーションを行った.

また, 2.2 節で抽出した各フレーズについて, 以下の 2 つの観点で分類を行った. ここでは各タイプのフレーズが幻覚になる割合を算出するために, 入出力の双方に出現するフレーズに対してもアノテーションを行っている.

(1) 広告訴求に関する分類

訴求フレーズ, 商材フレーズ, その他 の 3 つの選択肢から回答を選択する形式とした. ここで訴求フレーズは, 広告対象の商品やサービスの魅力や, 比較対象に対する優位性を述べたフレーズと定義し, [9] で規定されているいずれかの訴求タイプ ([9] の Table.1) に該当するかどうか, を基準としたアノテーションを行った.

(2) 一般的な観点での分類

固有表現, ジャンル特有の用語[8], 時間表現, 数量表現, その他 の 5 つの選択肢から回答を選択する形式とした. 固有表現, 時間表現, 数量表現の定義については関根の拡張固有表現階層 Ver 7.1.1[9] に従った.

アノテーションは株式会社サイバーエージェント内の広告アノテーション経験者 3 名により行った. アノテーションを実施する前に, 分析対象のデータとは異なる 50 事例についてパイロットアノテーションを実施し, 第 1 著者の想定回答とずれがある場合はフィードバックを行った.

上述した, 各フレーズの各分類タスク (幻覚の有無, 広告訴求, 一般的な観点) に対し, アノテータ 3 名によるアノテーションを得た. ラベルは多数決によって決定し, 3 名のアノテーションが全て異なる選択肢に分かれた場合には, 第一著者が 3 名のアノテーションを考慮して, どのアノテーションを採用するかを決定した. いずれの分類タスクにおいても,

6) https://spacy.io/
7) 商品名の一部だけが固有表現として抽出されているケースなどが該当する.
8) ここでジャンル特有の用語とは, そのフレーズによって, どの業種の広告文かが明確になるという性質を持つものと定義する ([例] ホールインワン).
9) https://sites.google.com/site/extendednamedentity711/

多数決でラベルが決定できなかったのは全体の 2% 以下であった.

最後にアノテータ間一致率について述べる. 幻覚有無と一般的な観点についての分類タスクについてはそれぞれ 0.78, 0.69 という高い Fleiss' kappa 値 [10] が得られた (Substantial agreement [11]). 広告訴求に関する分類タスクの一致率は前 2 者に比べれば低い値だがそれでも 0.33 (Fair agreement [11]) となった.

3 結果

幻覚有無アノテーションの結果を表 2 に示す. 表 2 の「はい」は, 当該のフレーズは入力情報 (キーワード, LP) の要約に含まれ得る, 即ち 幻覚ではない, というケースに対応する. 「はい」以外の 4 つの選択肢は幻覚に対応する. 574 フレーズ中, 82.8 % (475/574) が幻覚としてアノテーションされていることが分かる.

幻覚の内訳を見ると, 「いいえ (入力側に類似語句は存在しない)」が 430 / 475 と全体の 90.5%を占めるが, 「いいえ (入力側に類似語句が存在するが意味は異なる)」も 7.6% 存在する. 後者の例としては, LP 側の「カップル」が, 広告文側では「2 人」に言い換えられている事例などが挙げられる.

表 3 幻覚の内訳.

観点	タイプ				
訴求	訴求	商材	その他		
	334	8	133		
一般	用語	数量	時間	固有表現	その他
	49	36	11	10	369
主辞の品詞	名詞	動詞	副詞	形容詞	その他
	348	59	27	37	4

表 4 各フレーズが幻覚になる割合.

種別	幻覚率 [%]	幻覚フレーズ数	総数
訴求	42.1	334	794
商材	6.7	8	120
その他	51.4	133	259

幻覚を持つフレーズを各観点で分類したアノテーションの結果 (表 3) から, 幻覚フレーズの 70%以上が訴求としてアノテーションされていることが分かる ([例] 業界最大級, 全国). 商材に関する幻覚も 1.7%, 存在する. ジャンル特有の用語 ([例] くすみ, 最大減量) が少なくとも 約 10%, 数量表現も 約 7 % 含まれる. 品詞について見ると, 約 75%が名詞句だが, 用言 (動詞句, 副詞句, 形容詞句) も約 25% 存在する. また, 各フレーズを訴求に関して分類し, 種別ごとの幻覚率を算出した表 4 から, 広告文中の 商材フレーズ は 7% 程度しか幻覚にならないが, 訴求フレーズは 約 42%, その他のフレーズは 約 51% が幻覚になっていることが分かる.

4 議論

本分析の結果, 広告文側にのみ出現するフレーズの内, 少なくとも 8 割は入力を単に要約しただけでは出現し得ないフレーズ, 即ち幻覚であることが明らかになった. 1 節でも述べたように, 事実でない広告文の生成頻度を低減するためには, (A) 広告データセット中の幻覚を検出し, (B) 検出した幻覚を考慮したデータ編集やモデル学習上の工夫を行う必要がある.

上記 (A) に関する見通しを以下に述べる. (a) 数量表現, (b) 商材フレーズに関する幻覚は, 事実性の観点で広告品質に与える悪影響が大きいことから, また (c) 訴求フレーズ, (d) ジャンル特有の用語は高い出現頻度を持つことから, それぞれ検出が必要である. 従って, 入出力双方から上記 (a)-(d) の表現を検出し, 出力側にしか現れない表現を特定するアプローチが有効であると考えられる. これらの幻覚の検出は, 事前に各業種で辞書を作成し, ルールベースまたは distant supervision [12] で実現できると考えられる. 検出器の構築と性能評価は今後の課題とする.

一方, 上記 (B) には大別して (1) データ側での対処 [13, 14], (2) モデル側での対処 [15, 16, 17] という 2 つのアプローチが考えられる. (1) では, データセット中の広告文に幻覚が含まれる場合, 事例自体を廃棄するか [14], 広告文中の幻覚の削除や置換を行う必要がある [13]. 一方 (2) の例としては, 幻覚の低減に関する報酬関数を用いた強化学習 [15] や, 幻覚の度合いを制御コード (追加入力) として用いた制御可能な生成 [17] などが挙げられる.

5 おわりに

日本語の検索連動型広告のデータセットから抽出した事例に対して, 幻覚に関するアノテーションと分析を行った結果, 広告文側にのみ出現するフレーズは 8 割以上, 幻覚であることを確認した. また分析結果を踏まえ, 広告データセット中の幻覚への対処の見通しを述べた.

謝辞

本研究のアノテーションを実施して頂いた, 宇地原麻子氏, 宮城那南氏, 宮里竜士氏, 山城葵氏 に感謝します.

参考文献

[1] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics.

[2] Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5055–5070, Online, July 2020. Association for Computational Linguistics.

[3] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4812–4829, Online, June 2021. Association for Computational Linguistics.

[4] Meng Cao, Yue Dong, and Jackie Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3340–3354, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[5] Shaunak Mishra, Manisha Verma, Yichao Zhou, Kapil Thadani, and Wei Wang. Learning to create better ads: Generation and ranking approaches for ad creative refinement. In **Proceedings of the 29th ACM International Conference on Information & Knowledge Management**, pp. 2653–2660, 2020.

[6] Yiping Jin, Akshay Bhatia, Dittaya Wanvarie, and Phu T. V. Le. Generating coherent and diverse slogans with sequence-to-sequence transformer. **CoRR**, Vol. abs/2102.05924, , 2021.

[7] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] 松田寛. Ginza-universal dependencies による実用的日本語解析. 自然言語処理, Vol. 27, No. 3, pp. 695–701, 2020.

[9] Soichiro Murakami, Peinan Zhang, Sho Hoshino, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. Aspect-based analysis of advertising appeals for search engine advertising. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track**, pp. 69–78, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics.

[10] Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. The measurement of interrater agreement. **Statistical methods for rates and proportions**, Vol. 2, No. 212-236, pp. 22–23, 1981.

[11] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. **biometrics**, pp. 159–174, 1977.

[12] Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. Learning named entity tagger using domain-specific dictionary. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2054–2064, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[13] David Wan and Mohit Bansal. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1010–1028, Seattle, United States, July 2022. Association for Computational Linguistics.

[14] 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, Online, July 2020. Association for Computational Linguistics.

[15] Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. Plan-then-generate: Controlled data-to-text generation via planning. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 895–909, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[16] Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 1430–1441, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

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

A　広告文から抽出したフレーズ

2.2 節で広告文から抽出したフレーズの具体例を表 5 に示す.

表 5　抽出したフレーズの具体例.

カテゴリ	例
数量表現	0.95 %, 1,100 円, 15 階
時間表現	2022 年, 12 週間
ジャンル特有の用語	医療保険, HDD
固有表現	JR 桜木町駅, オメガ脂肪酸
その他	おすすめ, 安心, お任せください, まだまだ, わかりやすい