CAMERA^3: An Evaluation Dataset for Controllable Ad Text Generation in Japanese

Go Inoue,[†*] Akihiko Kato,[‡] Masato Mita,[‡] Ukyo Honda,[‡] Peinan Zhang[‡]
[†]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[‡]CyberAgent, Tokyo, Japan
go.inoue@mbzuai.ac.ae
{kato_akihiko, mita_masato, honda_ukyo, zhang_peinan}@cyberagent.co.jp

Abstract

Ad text generation is the task of creating compelling text from an advertising asset that describes products or services, such as a landing page. In advertising, diversity plays an important role in enhancing the effectiveness of an ad text, mitigating a phenomenon called “ad fatigue,” where users become disengaged due to repetitive exposure to the same advertisement. Despite numerous efforts in ad text generation, the aspect of diversifying ad texts has received limited attention, particularly in non-English languages like Japanese. To address this, we present CAMERA^3, an evaluation dataset for controllable text generation in the advertising domain in Japanese. Our dataset includes 3,980 ad texts written by expert annotators, taking into account various aspects of ad appeals. We make CAMERA^3 publicly available, allowing researchers to examine the capabilities of recent NLG models in controllable text generation in a real-world scenario.

Keywords: ad text generation, controllable text generation, annotation, Japanese

1. Introduction

Ad text generation is a real-world application of natural language generation (NLG) in commercial contexts, where the goal is to create compelling and persuasive advertisements for specific products or services. Ad text generation has received limited attention in the research community, despite its practical implications and the challenges it presents to address limitations of current state-of-the-art NLG models (Murakami et al., 2023). This is partly due to the lack of publicly available datasets because of the proprietary nature of advertising, which is particularly pronounced in non-English languages.

One example that addresses these issues include the effort by Mita et al. (2023), where they released the CAMERA dataset, a publicly available corpus for ad text generation in Japanese, a language that exhibits interestingly contrasting linguistic differences from English (Bond and Baldwin, 2016). A dataset in such a language serves as a valuable resource to examine the cross-lingual ability of state-of-the-art large language models (LLMs). In this work, we build upon their effort by constructing an evaluation dataset for *controllable ad text generation* (Figure 1).

A key difference from previous work, including the efforts in English (Golobokov et al., 2022a; Chai et al., 2022), is that we annotate multiple ad texts conditioning on various *aspects of ad appeals (A^3)* (e.g., *“This book is [50% off*`DiscountPrice`*] for [the weekends only*`LimitedTime`*]”*) (Murakami et al., 2022). This is particularly important in diversifying generated text in the advertising domain, where a phenomenon called *ad fatigue* (Abrams and Vee, 2007) can significantly impact user engagement and the effectiveness of advertising campaigns. Ad fatigue occurs when a user sees the same advertisements repeatedly, leading to a negative perception of the advertised products.

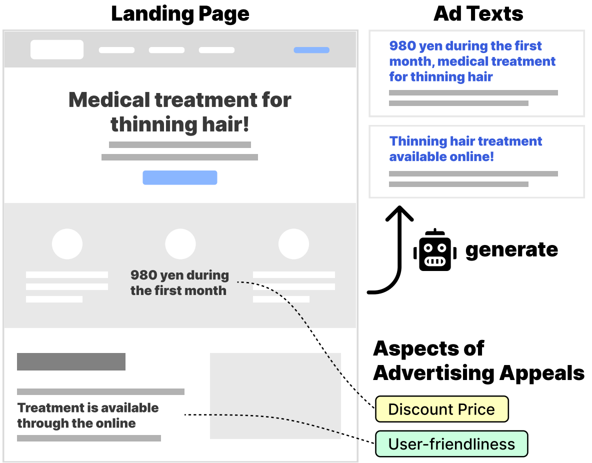
Landing Page
Ad Texts
Medical treatment for thinning hair!
980 yen during the first month
Treatment is available through the online
980 yen during the first month, medical treatment for thinning hair
Thinning hair treatment available online!
generate
Aspects of Advertising Appeals
Discount Price
User-friendliness

Figure 1: Controllable ad text generation.

To address these challenges, we introduce **CAMERA^3**, an evaluation dataset for controllable ad text generation in Japanese. We extend the CAMERA dataset (Mita et al., 2023) by further annotating LP images for A^3 and creating ad texts conditioned on these annotations. Our dataset comprises 3,980 expert-generated ad texts la-

[*]Work done during an internship at CyberAgent.

LREC-COLING 2024, pages 2702–2707

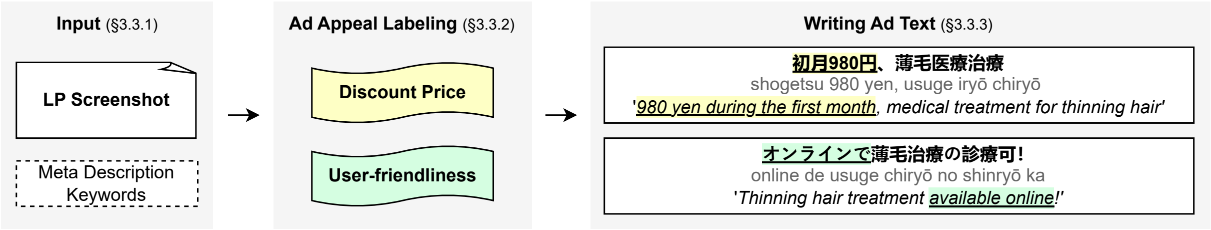
Input (§3.3.1)
LP Screenshot
Meta Description
Keywords
Ad Appeal Labeling (§3.3.2)
Discount Price
User-friendliness
Writing Ad Text (§3.3.3)
初月980円、薄毛医療治療
shogetsu 980 yen, usuge iryō chiryō
'980 yen during the first month, medical treatment for thinning hair'
オンラインで薄毛治療の診療可!
online de usuge chiryō no shinryō ka
'Thinning hair treatment available online!'

Figure 2: A schematic overview of the annotation process for CAMERA[3].

beled for various aspects of ad appeals. Each instance consists of three elements, an LP (Landing Page) screenshot, an ad appeal label identified within the LP, and a short ad text written conditioning on the annotated ad appeal. As an illustration, consider an LP promoting a medical hair treatment service includes a section highlighting the *"user-friendliness"* of their service. We annotate a short ad text that emphasizes this aspect (e.g., *"Thinning hair treatment available online!"*), as well as other aspects found in the LP. We make our dataset publicly available for research purposes[1].

Our contributions are as follows:

- We are the first to present an evaluation dataset for controllable text generation in the advertising domain in Japanese, which poses unique challenges due to its domain peculiarities and language characteristics.
- We release our dataset under the `CC BY-NC-SA 4.0` license, allowing researchers to examine approaches for controllable text generation in an underexplored real-world scenario.
- We report the first benchmark result of a simple baseline with an LLM, providing a foundation to uncover limitations of existing models.

2. Related Work

There is a limited number of publicly available resources in ad text generation, exceptions being UCL's open advertising dataset[2] and the CAMERA dataset (Mita et al., 2023). To the best of our knowledge, there is no publicly available dataset for controllable ad text generation. Previous studies on controllable ad text generation have addressed the control of rhetorical appeals or selling points of ad text (Chai et al., 2022; Golobokov et al., 2022b). Despite the effectiveness of their methods, the datasets they used are not publicly available, which hinders the development of research in this task. Jin et al. (2023) provided a dataset for syntactically diverse slogan generation, but it differs from our corpus in that it does not allow semantic control.

Outside of ad text generation, controllable text summarization is the most related task, considering the task similarity between ad text generation and text summarization (Murakami et al., 2023). Many datasets have been provided in this task (Fan et al., 2018; He et al., 2021; Zhong et al., 2021; Hayashi et al., 2021), but these were created by giving pseudo-attributes to existing datasets, not really written with the guidance to include attributes. Recently, Zhang et al. (2023) collected summaries by giving the annotators attributes to include in the summaries. Our work follows a similar approach and collected ad text with the guidance to include ad appeals.

3. Corpus Construction

3.1. Design Principle

The design philosophy of our corpus is inspired by the *prescriptive paradigm*, which discourages annotator subjectivity in the annotation process (Rottger et al., 2022). In this work, we discourage annotator subjectivity in writing ad text by focusing on a specific aspect of the LP to be annotated. More specifically, we explicitly ask annotators to write ad text taking into account a specific advertising appeal found in the LP, as described in detail in §3.3. This approach is in contrast to the *descriptive paradigm*, which encourages annotator subjectivity when creating datasets. Examples of such descriptive corpora include the CAMERA dataset (Mita et al., 2023), where they provide minimal instructions to the annotators to annotate LPs for ad text generation. Our corpus differs from theirs in the design principle: We aim to build a diversified corpus in ad text generation while incorporating the capability to control the specific aspect that introduces diversity to ad text.

3.2. Source Data

Our corpus is derived from a publicly available dataset for ad text generation in Japanese. We

[1] `https://github.com/CyberAgentAILab/camera3`

[2] `https://code.google.com/archive/p/open-advertising-dataset/`

use the test set of the CAMERA dataset (Mita et al., 2023), a collection of 869 LP images annotated for four ad text per instance. Each instance consists of an LP screenshot image, meta description, keywords, original ad text, and three annotated ad texts. In this work, we enrich their dataset with ad appeals found in the LP and additional ad texts that represent annotated ad appeals.

3.3. Annotation Task

Figure 2 shows a schematic example of an annotation instance. The annotation task is divided into two subtasks; ad appeal labeling (§3.3.2) and ad text creation (§3.3.3). Each annotation instance consists of an LP screenshot, extracted meta description, and keywords. We ask annotators to annotate each instance for aspects of ad appeals found in the LP, such as *"discount price"* and *"user-friendliness"*, followed by ad text that takes into account each specified aspect label. If no ad appeal is found in the LP, we label the instance as *"No Appeal"* without ad text annotation.

We used `Label Studio`,[3] an open-source annotation platform to annotate LPs for ad appeals and corresponding ad text. Annotation was performed by three annotators who are native Japanese speakers with experience in in-house ad production.[4]

3.3.1. Data Preparation

We manually verified LP screenshots in the test split of the CAMERA dataset and excluded instances with a screenshot error. We also excluded instances with a significantly lengthy LP image (above 30,000 pixels in height), yielding 819 LP images in total[5]. For efficiency purposes and to reduce the annotator's workload, we further split each LP screenshot into four segments horizontally. This also allow us to investigate which part of the original LP image contains the most relevant information for advertising. The resultant dataset contains 3,217 LP segments with 2,589 pixels in height on average, and 1,200 pixels in width.

3.3.2. Labeling Aspects of Ad Appeals

The first stage of the annotation is labeling aspects of ad appeals in the LP segment. The annotators first identify advertising appeals in the LP screenshot and annotate them with corresponding aspect labels. To avoid external influences, we ask annotators not to consult meta description and keywords at this stage. During the annotation session, annotators are encouraged to refer to the document that describes definitions of the labels along with the associated examples. We use an established label set for ad appeals based on Murakami et al. (2022) with an additional label for instances without any advertising appeal (*"No Appeal"*). Table 1 summarizes the label set we use in this work.

Ad Appeal Label	
1) No Appeal	12) Other features
2) Discount price	13) Limited time
3) Reward points	14) Limited target
4) Free	15) First-time limited
5) Special gift	16) Other limited offer
6) Other offer	17) Largest/no.1
7) Quality	18) Product lineup
8) Problem Solving	19) Trend
9) Speed	20) Other track record
10) User-friendliness	21) Story
11) Transportation	22) Other

Table 1: The aspect labels for advertising appeals, based on the scheme of Murakami et al. (2022).

3.3.3. Writing Ad Text

The second stage of the annotation is writing ad text *conditioned on* the annotated ad appeal label in the first stage. The annotators produce ad text guided by the ad appeal label assigned during the first stage (§3.3.2). In the absence of ad appeal, this phase will be disregarded.

A summary of the instructions to the annotators is as follows:

- Ad text must include expressions that represents the annotated ad appeal in the first stage.
- The length must be within 15 full-width characters (30 half-width characters)[6].
- Ad text should not be copy-pasted from the LP *as is* to ensure the diversity of the generated ad text[7].

During the second stage, annotators are allowed to consult meta description and keywords in case essential information, such as product or service name, is missing in the LP segment.

3.3.4. Quality Control

To ensure the annotation quality, annotators received training on the annotation process. The

[3] `https://labelstud.io/`

[4] Each instance takes approximately 90 seconds to annotate on average.

[5] The original dataset has 869 LP images.

[6] We follow the length limit set by Google's text ads guideline: `https://support.google.com/google-ads/answer/1704389`

[7] This is also to make sure that the task is not going to be a mere OCR task.

CAMERA^3 Statistics	
# LP Segments	3,217
# LP Segments w/ Ad Appeal	1,974
# LP Segments w/o Ad Appeal	1,243
# Ad Appeal & Ad Text	3,980

Table 2: The statistics of CAMERA^3. The corpus has 3,980 instances of the triplet (LP Segment, Ad Appeal, Ad Text).

annotators first received multiple sessions where the authors provided explanations of the annotation guidelines. All annotators went through two practice annotation rounds followed by a feedback session by the authors after each round. During the feedback session, the authors provided annotators with comments for each annotated instance, highlighting the difference from expected annotations, if any. Examples used for the practice sessions (77 instances in total) were carefully annotated by the authors.

To estimate the annotation quality, we compute Krippendorff's alpha (Krippendorff, 1980) on a sample of 100 instances annotated by the three annotators. Krippendorff's alpha is 0.33 for the labeling task, which is a fair agreement based on the guidelines in Landis and Koch (1977).

3.4. Corpus Statistics and Data Format

Table 2 shows statistics of our corpus. More than 60% of the LP image segments have at least one ad appeal, while the rest are labeled as *"No Appeal"*. On average, each LP segment has two ad texts associated with the corresponding ad appeal labels. In total, we obtain 3,980 ad text with ad appeal label annotations.

CAMERA^3 is released in json format, where each element has a file name of the LP segment, ad appeal label, and ad text, along with other supplemental information including OCR text, LP meta description, and keywords.

4. Ad Text Generation Baseline

4.1. Experimental Setup

To demonstrate the usefulness of CAMERA^3, we conduct baseline experiments with a state-of-the-art LLM. The purpose of this experiment is to provide results of a simple baseline as an initial comparison point for future efforts. To that end, we prompt a state-of-the-art LLM, GPT-3.5-Turbo[8] (Ouyang et al., 2022). We prompt it to generate an ad text given the specified ad appeal label, meta description, and an OCR-processed LP segment. We also include the definition of the ad appeal label in the prompt. To obtain textual representations from an LP screenshot, we use Google's Cloud Vision API[9].

	GPT-3.5-Turbo
Content	66.81%
Style	36.31%
Format	13.14%

Table 3: Evaluation of GPT-3.5-Turbo in terms of format, style, and content.

We prompt the LLM as follows: "*You are an advertising copywriter. Write only one advertisement containing {ad appeal label} ({description of the label}) from LP meta description and OCR text. Write it in 15 full-width characters (30 half-width characters) or less. LP meta description: {lp meta description} OCR text: {text}*"[10]

4.2. Evaluation

We evaluate controllable ad text generation models with the following three criteria, inspired by the evaluation criteria in style transfer (Madaan et al., 2020; Reid and Zhong, 2021).

- **Content**: BERTScore (F_1) (Zhang et al., 2020) to compute the content similarity of the generated ad text to the gold reference.
- **Style**: The percentage of generated ad texts classified to have the target ad appeal label. We use a BERT-based classifier by Murakami et al. (2022) to predict the ad appeal label.
- **Format**: The percentage of generated ad texts following the specified format, i.e., within 15 full-width characters.

4.3. Results

Table 3 shows the baseline result of GPT-3.5-Turbo on our dataset. Interestingly, the result shows a contrasting difference in performance between the content and the format specificity. We observe a significantly low format accuracy, even though the length of an ad text is a surface-level textual attribute. The low style accuracy suggests that generating an ad text with a specified ad appeal is a challenging task, calling for further developments in controllable ad text generation.

[8] We use the `2023-03-15-preview` version.

[9] https://cloud.google.com/vision/docs/ocr

[10] We use Japanese for instruction (Appendix A).

5. Conclusion and Future Work

We presented **CAMERA**[3], an evaluation dataset for controllable ad text generation in Japanese. Our dataset includes 3,980 ad texts written by expert annotators, taking into account various aspects of ad appeals within each LP. We also release our dataset under the `CC BY-NC-SA 4.0` license, to facilitate the exploration of approaches for controllable text generation in a real-world scenario. As an initial comparison point, we offer a simple baseline by prompting a state-of-the-art LLM. Our dataset serves as a testbed for evaluating the capabilities and limitations of state-of-the-art NLG models. In future work, we plan to explore diverse variants of LLMs and other model types, such as multi-modal models.

Limitations

One of the limitations in this work is the limited scope of baseline model exploration. Although we acknowledge the importance of exploring numerous LLMs that have shown remarkable generative abilities, the main focus of this paper is to provide the foundation for controllable ad text generation through the development of an evaluation dataset. As such, we leave extensive exploration of these models to future work. We also acknowledge the potential bias in annotation stemming from the demographic background of the annotators.

Acknowledgement

We thank the three anonymous reviewers for helpful suggestions. We also thank Kazuhiro Ota, Shota Sasaki, Sho Hoshino, Timothy Baldwin, and Nizar Habash for their valuable feedback.

6. Bibliographical References

Zoë Abrams and Erik Vee. 2007. Personalized ad delivery when ads fatigue: An approximation algorithm. In *Internet and Network Economics*, pages 535–540, Berlin, Heidelberg. Springer Berlin Heidelberg.

Francis Bond and Timothy Baldwin. 2016. Introduction to japanese computational linguistics. *Readings in Japanese Natural Language Processing, CSLI Publications, Stanford, USA*, pages 1–28.

Junyi Chai, Reid Pryzant, Victor Ye Dong, Konstantin Golobokov, Chenguang Zhu, and Yi Liu. 2022. Fast: Improving controllability for text generation with feedback aware self-training.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Konstantin Golobokov, Junyi Chai, Victor Ye Dong, Mandy Gu, Bingyu Chi, Jie Cao, Yulan Yan, and Yi Liu. 2022a. DeepGen: Diverse search ad generation and real-time customization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 191–199, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Junxian He, Wojciech Maciej Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2021. {CTRL}sum: Towards generic controllable text summarization.

Yiping Jin, Akshay Bhatia, Dittaya Wanvarie, and Phu TV Le. 2023. Towards improving coherence and diversity of slogan generation. *Natural Language Engineering*, 29(2):254–286.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Content Analysis: An Introduction to Its Methodology. Sage.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. 2023. Camera: A multimodal dataset and benchmark for ad text generation.

Soichiro Murakami, Sho Hoshino, and Peinan Zhang. 2023. Natural language generation for advertising: A survey.

Soichiro Murakami, Peinan Zhang, Sho Hoshino, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2022. 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*, pages 69–78, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Machel Reid and Victor Zhong. 2021. LEWIS: Levenshtein editing for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3932–3944, Online. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023. Macsum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A. Prompt Template

We use the following template to prompt an LLM in Japanese:

“あなたは広告ライターです。LPメタディスクリプションとOCRテキストから、{ad appeal label}({description of the label})を含む広告文を1つだけ書きなさい。全角15文字 (半角30文字) 以内で書きなさい。LPメタディスクリプション: {lp meta description} OCRテキスト: {text}”