2024

2022 Vol.1
2023 Vol.1
2023 Vol.2

White Paper Project

CONTENTS

2022 Vol.2

意味的類似度計算システムによるチャットボットFAQシステムの性能向上

栗原 健太郎
Kentaro, Kurihara
株式会社 AI Shift
ML/DS Engineer
kurihara_kentaro@cyberagent.co.jp

二宮 大空
Hirotaka, Ninomiya
株式会社 AI Shift
ML Engineer
ninomiya_hirotaka@cyberagent.co.jp

友松 祐太
Yuta, Tomomatsu
株式会社 AI Shift
DS Engineer
tomomatsu_yuta@cyberagent.co.jp

keywords: 意味的類似度 (STS) 計算システム, Dense Retriever, チャットボット FAQ

Summary

カスタマーサポートや社内ヘルプデスクなどにおける問い合わせ対応に、チャットボットが適用されつつある。AI Shift では、各種サービスにおけるユーザの質問に自動回答するシステムとして Dense Retriever を活用したチャットボット FAQ システムの構築を検討している。現状、AI Shift のボイスボット事業において用いている Dense Retriever の学習に、人手によるデータフィルタリングを適用した対話データを用いているが、顧客の多様さとデータ量の多さ故に非常に手間がかかる作業となっている。そのため、チャットボット FAQ システムの構築においても同様の課題が懸念される。本研究では、意味的類似度計算システムを用いた学習データのフィルタリング自動化を検討する。提案手法によるフィルタリングを適用したデータセットで FAQ システムを学習することで、フィルタリング未適用の FAQ システムの性能を上回ることから、フィルタリングが性能向上に効果的であることを示した。

1. はじめに

多くの企業や団体が提供するサービスにおけるカスタマーサポートなどにおいて、チャットボットが適用されつつある。チャットボットが提供する機能の一つに、各種サービスにおける「よくある質問」などと呼ばれる Frequently Asked Questions (FAQ) 検索を用いたユーザ質問への回答機能が存在する。FAQ 検索では、企業が保持する FAQ のデータベースに基づき、ユーザ質問に対して最もマッチする回答を得ることができる。

我々は現在構築を検討中のチャットボットの FAQ システムにおける検索手法として、Open-Domain QA で有効とされている Dense Retriever [Karpukhin 20] を採用する。Dense Retriever の学習には、自社のチャットボット事業で収集している <ユーザ質問, FAQ 質問> を対話ペアと見做した対話データを用いる [*1]。本対話データにおいて、ユーザ質問とユーザが選択した FAQ 質問の対話ペアを正例として学習に用いる。

しかし、プロダクトで収集される対話データには、ユーザが選ぶ FAQ 質問の内容が質問内容とマッチしていない品質の悪い対話ペアが含まれている。これらを正例と見做して学習することで、Dense Retriever の学習の際にノイズとなる恐れがある。品質の悪い正例の例を表 1 に示す。いずれもユーザ質問と FAQ 質問の内容が大きく異なるため品質の悪い正例と見做す。また、複数顧客のデータからランダムサンプリングした対話ペアに対して、筆者による品質の良い正例であるか否かについてのアノテーションを実施した結果は、品質の { 良い正例: 455 件, 悪い正例: 328 件 } となっており、品質の悪い正例が多く含まれていると言える。さらに、多様な顧客から多数の対話ペアデータを収集していることから、人手での品質の悪い正例の除去は大変手間がかかる作業となっている。人手フィルタリング以外の品質の悪い正例を除去する手

表 1 品質の悪い正例の例

ユーザ質問	FAQ 質問
アンケート	最新情報を教えてください
アカウントが作れない	ログインできない
あああああああ	定休日はいつですか？

*1 対話データの収集方法の詳細については [二宮 22] らの 4 章に原則従う。

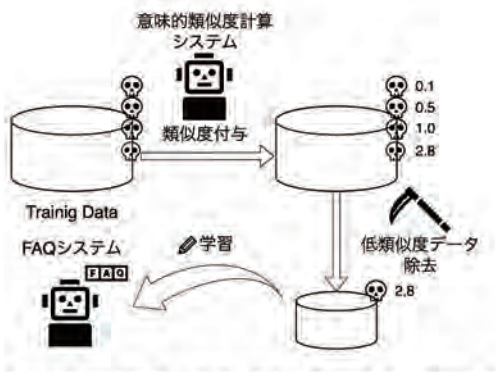


図 1　意味的類似度計算システムを用いた訓練データの自動フィルタリングのフロー

段として、FAQ 質問選択後のフィードバック結果 *2によるフィルタリングが可能だが、フィードバックに回答するユーザの少なさ故のデータの少なさが課題である。

本研究では、意味的類似度計算システムを用いた対話ペアの類似度付与による品質の悪い正例の除去の自動化を提案する。提案手法によるフィルタリングを適用したデータで学習した FAQ システムは、フィルタリング未適用のデータで学習した FAQ システムと比較して、より高い推論性能を達成した。実験結果は、提案手法を用いた学習データフィルタリングの自動化による人手の労力削減が実現可能であることを示している。

2. 関 連 研 究

Open-Domain QA タスクにおける文書検索では、表層情報による文書検索手法として TF-IDF や BM25 が用いられていた。昨今では密なベクトル表現に基づいた文書検索を行う Dense Retriever が採用されつつある。[加藤 21] らは日本語の Open-Domain QA データセット JAQKET を用いた DPR における Retriever の性能評価も実施しており、一定の性能の発揮を報告している。[Talukdar 21] らは学習に用いるデータのフィルタリングが言語理解モデルの性能向上に寄与すると報告しているが、その調査は主に SST などの分類タスクを対象としている。

本研究においては、２文間の類似度が低い対話ペアを品質の悪い対話ペアと見做し、回帰タスクである意味的類似度計算 (Sentence Textual Similarity: STS) タスクに帰着させることでデータのフィルタリングを実現する。STS タスクでは、正解類似度は 0（意味が完全に異なる）から 5（意味が等価）の実数値で定義されることが一般的である。STS タスクに関するデータセットとして、英語の sts-b や日本語の JSTS [Kurihara 22] が存在する。いずれのデータセットも、特定のドメインに依らない一般ドメインのデータセットとして構築されている。

*2　ユーザの FAQ 質問の選択後に、回答を閲覧することで質問内容が解決したか否かの 2 択（「はい」と「いいえ」）を問うフィードバック質問をユーザに提示している。

表 2　各種訓練データのサンプルサイズ

訓練データ	データ数
Raw Data	178,152
JSTS Data ($\mathcal{S} < 1.0$)	160,684
JSTS Data ($\mathcal{S} < 1.5$)	114,130
vanilla-BERT Data ($\mathcal{S} < 3.2$)	145,937
vanilla-BERT Data ($\mathcal{S} < 3.9$)	21,984

表 3　各種評価データのサンプルサイズ

評価データ	Known-Domain		Unknown-Domain	
	dev	test	dev	test
Normal-Data	22,270	22,279	684	684
Filtered-Data	6,763	6,777	251	251

3. 意味的類似度計算システムを用いた低品質対話ペアのフィルタリング

検討中のチャットボット FAQ システムにおける Dense Retriever の学習に用いる対話データには、品質の悪い正例が存在しており、モデルの学習の際にノイズとなる恐れがある。一方で、顧客の多様さとデータの多さ故に人手フィルタリングは非常に手間がかかる作業となっている。

そこで、品質の悪い正例の対話ペアの多くは、2 文間の内容が大きく異なることから意味的類似度が低いという仮定の元、図 1 に示す自動フィルタリングのフローを提案する。意味的類似度計算システムを用いて類似度 $\mathcal{S}$ を獲得し、$\mathcal{S}$ が一定以下である対話ペアを除去することで自動フィルタリングを実施する。その後フィルタリングしたデータを用いて FAQ システムを学習させる。意味的類似度計算システムの構築には、BERT [Devlin 19] を活用する。学習については、一般ドメインの STS データセットである JSTS で BERT の事前学習済みモデルを fine-tuning することによって意味的類似度計算システムを構築する。システムが算出する意味的類似度 $\mathcal{S}$ は、2 章における STS タスクの先行研究にならって、0 から 5 の間の実数値となるように構築する。

4. FAQ システムの評価実験

4·1 実 験 設 定

自社のチャットボット事業で収集した対話ペアを用いて、チャットボット FAQ システムにおける Dense Retriever の学習･評価を実施する。意味的類似度計算システムを用いたデータセットのフィルタリングの有効性検証のため、フィルタリングを適用した対話データと適用していないデータそれぞれで Dense Retriever を学習し、推論性能を評価する。評価には Macro Average Top {1, 3, 5} Accuracy を使用し、各顧客データ毎に Top {1, 3, 5} Accuracy を算出した後、全体顧客数で平均することで算出する。

i.　ベースライン手法

提案手法との比較のため、ベースラインとして BERT の事前学習済みモデルを fine-tuning せず (vanilla-BERT) に構築した意味的類似度計算システムによるフィルタリ

表 4　意味的類似度計算システム、および FAQ システムの fine-tuning 時のハイパーパラメータ

	意味的類似度計算システム	FAQ システム
pretrained-model	cl-tohoku/bert-base-japanese-v2	cl-tohoku/bert-base-japanese-wholo-word-masking
Batch Size	8	64
learning rate	5e-5	1e-5
epoch	4	10
max-seq-length	512	64

表 5　各評価手法による FAQ システムの性能評価結果 (3 つのスコアは左から順に Top{1, 3, 5} Accuracy を表す)

Normal-Data				
訓練データ	Known-Domain		Unknown-Domain	
	dev	test	dev	test
Raw Data	0.284 / 0.472 / 0.550	0.290 / 0.477 / 0.564	0.302 / 0.489 / 0.570	0.291 / 0.463 / 0.577
JSTS Data ($\mathcal{S} < 1.0$)	**0.296 / 0.500 / 0.584**	**0.305 / 0.504 / 0.589**	**0.365 / 0.574 / 0.638**	0.323 / **0.564 / 0.628**
JSTS Data ($\mathcal{S} < 1.5$)	0.278 / 0.468 / 0.566	0.285 / 0.474 / 0.566	0.352 / 0.534 / 0.586	**0.364** / 0.544 / 0.610
vanilla-BERT Data ($\mathcal{S} < 3.2$)	0.285 / 0.472 / 0.556	0.290 / 0.487 / 0.571	0.344 / 0.534 / 0.606	0.303 / 0.524 / 0.616
vanilla-BERT Data ($\mathcal{S} < 3.9$)	0.231 / 0.402 / 0.483	0.235 / 0.413 / 0.495	0.308 / 0.496 / 0.575	0.325 / 0.511 / 0.564
Feedback-Data				
訓練データ	Known-Domain		Unknown-Domain	
	dev	test	dev	test
Raw Data	0.296 / 0.511 / 0.589	0.310 / 0.505 / 0.602	0.232 / 0.369 / 0.531	0.246 / **0.432** / 0.477
JSTS Data ($\mathcal{S} < 1.0$)	0.328 / **0.544 / 0.624**	0.326 / **0.545 / 0.635**	0.281 / **0.546 / 0.592**	**0.303** / 0.430 / 0.492
JSTS Data ($\mathcal{S} < 1.5$)	**0.333** / 0.525 / 0.620	**0.327** / 0.516 / 0.595	**0.289** / 0.530 / 0.567	0.289 / 0.401 / 0.450
vanilla-BERT Data ($\mathcal{S} < 3.2$)	0.305 / 0.501 / 0.579	0.311 / 0.506 / 0.578	0.267 / 0.443 / 0.581	0.299 / 0.412 / 0.479
vanilla-BERT Data ($\mathcal{S} < 3.9$)	0.248 / 0.410 / 0.479	0.246 / 0.412 / 0.489	0.212 / 0.473 / 0.521	0.266 / 0.388 / **0.523**

図 2　ランダムサンプリングした対話ペアの類似度の分布を示す箱ひげ図（valid: 品質の良い正例データ, invalid: 品質の悪い正例データに対応している）

ングを適用した訓練データで FAQ システムを学習させる。ベースラインでは、対話ペアのそれぞれの文を独立にモデルに入力し、最終層から出力される [CLS] トークンの分散表現の cos 類似度を算出することによって対話ペアの類似度を獲得する。その際、STS タスクにおける 0 から 5 の実数値に対応させるため、獲得した cos 類似度を 5 倍したスコアを意味的類似度計算システムの出力とする。ベースライン、および提案手法それぞれにおける FAQ システムの訓練に用いたそれぞれの対話データのサンプルサイズを表 2 に示す。

ii.　評価データの設計

評価データについては、ドメインに依らない汎化性能を評価するために [二宮 22] らの先行研究に倣い、Dense Retriever の学習に用いられている既知ドメイン（Known-Domain）と、学習に用いていない未知ドメイン（Unknown-Domain）の 2 種類のデータを用意する。ただし、訓練データと同様にユーザ質問とユーザが選択した FAQ 質問の対話ペアを正例として収集した評価データ (Normal-Data) も、品質の悪い対話ペアデータを含んでいる。そのため、1 章で述べたフィードバック結果において「はい」が選ばれている対話ペアのみを抽出した対話データ (Feedback-Data) でも評価を行う。それぞれの評価データのサンプルサイズを表 3 に示す。

iii.　除去する対話ペアの類似度の閾値

本実験では、除去する対話ペアの意味的類似度 $\mathcal{S}$ の閾値による FAQ システムの性能差を比較するため、各フィルタリング手法において複数の閾値を設定する。効果的な閾値設定のため、1 章で述べた人手アノテーションした対話データについて、ベースライン, 提案手法それぞれで意味的類似度を獲得した。同データの類似度の分布を図 2 に示す。提案手法について、$\mathcal{S}$ の閾値を品質の良い正例 (valid データ) の類似度分布の第一四分位、および品質の悪い正例 (invalid データ) の類似度分布の第三四分位に相当する 1.5 程度に設定することで、valid データを 1/4 程除去しつつも、invalid データを 3/4 程除去することができる。以上より、提案手法における除去する対話ペアの類似度 1.5 を閾値の 1 つに設定する。また閾値設定の妥当性検証のため、invalid データを最低限除去することを目的として 1.0 も閾値に設定する。ベースライン手法については valid, invalid データの類似度分布の重なりが大きく、効果的な閾値を設定することが比較的困難である。本実験では、提案手法における閾値設定に倣い、valid データの第一四分位および invalid データの第三四分位に相当する 3.2, 3.9 をベースライン手法の閾値

に設定する。

iv. その他のハイパーパラメータ

意味的類似度計算システムの構築に用いる BERT、および FAQ システムの構築に用いる Dense Retriever の fine-tuning におけるハイパーパラメータを表 4 に示す。

4.2 結 果・考 察

Dense Retreiver の推論性能の評価結果を表 5 に示す。全般的に JSTS Data で学習した Dense Retriever ベースの FAQ システムの方が、Raw Data で学習した場合や、vanilla-BERT Data で学習した場合と比較して Accuradcy が高スコアとなっている。JSTS Data ($\mathcal{S} < 1.0$) と Raw Data それぞれで学習した結果を比較した場合、Normal-Data, Known-Domain, test において Top {1, 3, 5} Accuracy はそれぞれ 1.5%, 2.7%, 2.5%上回っており、Normal-Data, Unknown-Domain, test では 3.2%, 10.1%, 5.1%上回っている。一方で、Known-Domain における性能について、vanilla-BERT Data で学習した場合のスコアは、Raw Data で学習した場合と同等またはそれ以下のスコアという結果になっている。これは、JSTS による fine-tuning を行なっていない vanilla-BERT を用いた invalid データの効果的な除去は困難であること、及び JSTS を用いた fine-tuning による 2 文間の意味的類似度の学習の有用性を示す結果となっている。

Normal-Data の Known-Domain における評価において、JSTS Data ($\mathcal{S} < 1.5$) で学習したモデルのスコア (Top 1 Accuracy) が Raw Data で学習したモデルと比べ低い。これは表 2 より、Raw Data での学習したモデルのスコアを上回っている JSTS Data($\mathcal{S} < 1.0$) や vanilla-BERT Data ($\mathcal{S} < 3.2$) と比較して JSTS Data($\mathcal{S} < 1.5$) のデータ数が少なく、学習に寄与する品質の良い正例対話ペアが比較的多く除去された可能性が考えられる。

一方で、Feedback-Data の Known-Domain における評価では、JSTS Data($\mathcal{S} < 1.5$) で学習したモデルのスコア (Top 1 Accuracy) が Raw Data で学習したモデルと比べ高い。この結果は、Normal-Data に含まれていてモデルが正解できなかった invalid データが、Feedback-Data ではフィルタリングされていることに起因すると考えられる。

5. お わ り に

本論文では、Dense Retriever ベースの FAQ システムの学習データのフィルタリング手法として、JSTS で fine-tuning した BERT ベースで構築した意味的類似度計算システムの活用を提案する。実験結果は、品質の悪い対話ペアのフィルタリングに伴うチャットボット FAQ システムの性能向上、及び JSTS を用いた意味的類似度計算システムの学習の有効性を示す結果となった。加えて、本手法で大量のデータに対して一挙にフィルタリングを適用することで、人手でのフィルタリングの手間を大幅に削減することが可能である。今後は、T5 や BART, GPT などの生成モデルを用いて、FAQ に紐づいた回答文章と答えから質問文を生成することなどによって、データセットの拡張をすることを検討する。さらに、データセット拡張とフィルタリングを相互に実施することで FAQ システムの更なる性能向上を目指す。

◇ 参 考 文 献 ◇

[Devlin 19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota (2019), Association for Computational Linguistics

[Karpukhin 20] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t.: Dense Passage Retrieval for Open-Domain Question Answering, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online (2020), Association for Computational Linguistics

[Kurihara 22] Kurihara, K., Kawahara, D., and Shibata, T.: JGLUE: Japanese General Language Understanding Evaluation, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2957–2966, Marseille, France (2022), European Language Resources Association

[Talukdar 21] Talukdar, A., Dagar, M., Gupta, P., and Menon, V.: Training Dynamic based data filtering may not work for NLP datasets, in *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 296–302, Punta Cana, Dominican Republic (2021), Association for Computational Linguistics

[加藤 21] 加藤 拓真, 宮脇 峻平, 西田 京介, 鈴木 潤：オープンドメイン QA における DPR の有効性検証, 言語処理学会 第 26 回年次大会, pp. 1403 – 1407 (2021)

[二宮 22] 二宮 大空, 邊土名 朝飛, 杉山 雅和, 戸田 隆道, 友松 祐太：チャットボット事業における Dense Retriever を用いた Zero-shot FAQ 検索, 第 96 回言語・音声理解と対話処理研究会 (第 13 回対話システムシンポジウム) (2022)

著 者 紹 介

栗原 健太郎

2023 年に早稲田大学大学院基幹理工学研究科博士前期課程を修了予定。同年 4 月に株式会社サイバーエージェントに入社予定。大学院では自然言語処理の特にデータセット・ベンチマークに関する研究に従事。現在の関心は、DX 分野とゲーム開発への AI 適用、及びデータ分析系。

二宮 大空

2021 年奈良先端科学技術大学院大学先端科学技術研究科博士前期課程修了後、国立研究開発法人情報通信研究機構有期雇用研究員として質問応答に関する研究に従事。2022 年サイバーエージェントに入社。現在は株式会社 AIShift で FAQ 検索の開発に従事。

友松 祐太

2018 年明治大学理工学研究科博士前期課程修了後、サイバーエージェントに新卒入社。現在は株式会社 AI Shift の ML/DS チームのマネージャーを務め、Chatbot, Voicebot の研究開発に従事。

音楽サブスクに旧来の消費手段は置き換わるのか
様々な音楽消費についてのアンケート調査の分析

武内 慎
Makoto Takeuchi
メディア統括本部 Data Science Center
Data Mining Engineer
takeuchi_makoto@cyberagent.co.jp

森下 壮一郎
Soichiro Morishita
学際的情報科学センター
Researcher
morishita_soichiro@cyberagent.co.jp

keywords: 音楽，サブスクリプションサービス，探索的因子分析，ロジスティック回帰，消費者行動論

Summary

音楽には様々な役割がある．観賞されるばかりではなく，友人や知人とのコミュニケーションのツールにもなり得る．そして音楽の役割と消費手段との間には密接な関係がある．たとえばCD（コンパクトディスク）は，音源を入手して消費するための手段であると同時に，貸し借りなどによるコミュニケーションのための手段でもある．近年では，サブスクリプション型ストリーミングサービス（以下，サブスク）が音楽市場でシェアを拡大し，人々の音楽の聴き方を変化させている．しかし，この変化が音楽の役割にどう影響を与えるかについては，解明されていない．以上の背景の下，本研究ではアンケート調査データを分析し，人々が音楽に期待する役割と，サブスクを含む音楽消費手段との関係を明らかにした．そして，サブスクが旧来の手段の単なる代替ではなく，音楽市場における新しい消費価値を提供する上に，音楽による個人のアイデンティティー構築にも寄与する可能性を示す．

1. は じ め に

音楽は個人のパーソナリティに応じて様々な役割を持つ．音楽心理学や消費者行動論の観点では，楽曲や音楽ジャンルと性格特性との関係が研究されている [Lacher 94, Chamorro-Premuzic 07].

そして音楽の消費手段は，消費者が期待する役割に応じて選択される．CDなどの音盤の購入が主な音楽の消費手段だった頃は，購入前の試聴行動に関わる金銭的コストや情報探索コストが主要な研究テーマの一つであった [Chellappa 05, Dewan 12]．しかし，サブスクリプション型ストリーミングサービス（以下，サブスクと呼称）の登場により楽曲の探索が容易になり，音楽市場全体における消費者行動が変化している．

以下，サブスク利用の消費者行動への影響に着目した研究について述べる．Datta ら [Datta 18] は，有料/無料を含む音楽ストリーミングサービスの利用が，その後の音楽聴取傾向に与える影響を調査した．その結果，音楽ストリーミングサービスの利用は，個人差はあるものの，音楽消費量と多様性を有意に増加させたと報告している．これは，コスト面での制約があった一部の消費者が，音楽ストリーミングサービスを利用することで音楽消費の需要を満たすことができたと解釈できる．他の研究では，音楽ストリーミングサービスは既存の音源購入とカニバリゼーションが起こるが，一部の場合を除き音楽市場全体の収益に貢献することを示している．[Papies 11, Wlömert 16].

以上のように，サブスクの市場への影響についての研究はあるものの，音楽が果たす個人にとっての役割と，サブスクを含む音楽消費手段の選択との関係については，いまだ不明な点が多い．録音/配信技術によって音楽を聴く手段は常に変化しており，それらが，音楽が本来持っている様々な役割を果たすのに適しているのかは自明でない．したがって，この関係を確認することは，満たされていない市場ニーズを把握するというビジネスの観点からも，音楽文化の発展という社会的な観点からも重要である．また，サブスクビジネスの今後の成長可能性，つまり，サブスクが他の既存の手段に完全に取って代わるのか，それとも共存していくのかを議論するためにも、この関係の知見は有益である．

以上の背景の下，本研究では人々が音楽に期待する役割を明らかにし，それが音楽消費手段の選択にどのように関わっているか検証する．具体的には，音楽に対して人々が期待する役割と選択している消費手段についてア

本稿は [Takeuchi 22] を改稿したものである．

ンケート調査を実施する．そして探索的因子分析により音楽の役割を代表的に示す因子を抽出する．さらにロジスティック回帰分析により，それぞれの因子と音楽消費手段との関係を見出す.

2. 手　　法

2·1 音楽の役割

複数の先行研究により, 音楽の目的は様々なカテゴリに分類されている [Lonsdale 11, Manolika 21, Ikegami 21]. これらの研究の中で分類方法は一部異なるが，[Lonsdale 11] では音楽の使用目的を 6 つのカテゴリ (**ポジティブな気分管理，ネガティブな気分管理，暇つぶし，対人関係，自己アイデンティティ，他人を知る**) に分類した．この分類は一般的で頻繁に採用されているため，本研究でもこの分類を部分的に採用した．具体的には，6 つの目的のうち，**対人関係**と**自己アイデンティティ**の 2 つの役割に対する期待を，特に音楽消費手段の選択に影響を与えると想定し，構成要素として選択した．他の 4 つの目的は，音楽消費手段の選択というより，聴く音楽（曲やジャンルなど）の好みに影響すると想定されるため，考慮しなかった．さらに，本研究の調査を実施した日本の状況として，ポップミュージックなどの分野では，ファンダムに関する特異性が報告されている [Stevens 10, Masae 13]. また，日本の音楽市場において，アーティストファンによる音楽消費は無視できないことから，ファンが音楽に期待する役割としての構成要素も採用した.

2·2 アンケート調査

我々は，2020 年 6 月に，インターネットリサーチ会社日本在住のモニターを対象に，スクリーニング調査と本調査の 2 部構成で調査を実施した．スクリーニング調査は 10,552 名を対象とした．本調査はその中から，普段音楽を聴いており、かつ設問中に提示した定額制サービスの概要説明を理解したと回答した回答者を 453 名サンプリングした．この際，サブスク利用経験者のサンプル数を十分確保する目的で，サブスク利用経験者と、非利用経験者をほぼ 1 対 1 の割合でサンプリングした.

アンケートでは，音楽の使用目的，音楽消費手段，および年齢性別に関して質問した．音楽の使用目的に関しては，[Lonsdale 11] の**対人関係**と**自己アイデンティティ**のそれぞれに対応する，音楽に関する行動を 4 問ずつと，アーティストファンに関する行動を 2 問，計 10 問質問した．これらは 5 水準のリッカート尺度で回答を得た．音楽消費手段については，無料，購入，レンタル，ライブ，サブスクリプションの 5 つの分類に属する複数の具体的な手段の使用/不使用に関する設問を提示し，具体的な手段を 1 つ以上使用している場合は，その分類を使用していると解釈した．また，本調査が実施された 2020 年時点では，COVID-19 の感染拡大が始まっていたため，COVID-19 による影響を防ぐ目的で,「新型コロナウィルスの流行以前のことについてお答えください。」という文言を入れた．音楽の使用目的と設問の対応関係や，具体的な文言等については，[Takeuchi 22] を参照されたい.

2·3 分　　析

まず，音楽の使用目的に関する回答について探索的因子分析を行い，音楽に期待される役割の構成要素を得た．次に，ロジスティック回帰モデルを用いて，手段の選択に影響を与える要因の分析を行った．この際，5 つの音楽消費手段の利用有無を従属変数とし，音楽に期待される役割の構成要素と，性別年齢を説明変数とした.

3. 結　　果

3·1 探索的因子分析

音楽の使用目的に関する 10 問の回答に対して，因子分析を行った．まず，平行分析で因子数を求め，結果は 4 となった．この 4 因子を F1-F4 と呼ぶ．そして，斜め回転の一種であるオブリミン回転と最小残差法を用いて因子分析を行った．結果を表 1 に示す.

F1 は，因子負荷量が最も大きい 4 項目がすべて**対人関係**に関する項目であるため，この構成概念を**対人関係への期待**と呼ぶ．F2 は，因子負荷量が最も大きい 3 項目がすべて**自己アイデンティティ**に関する項目であるため，この構成概念を**自己アイデンティティ構築への期待**と呼ぶ. また，F2，F3 の両方で因子負荷量が高い**自分に合うアーティストや楽曲を積極的に探す**も**自己アイデンティティ**に関する項目である．F3 は，**自分に合うアーティストや楽曲を積極的に探す**に加えて，**自分が好きなアーティストの新曲を常にチェックして聴く**に対する因子負荷量が大きい．これは，好きなアーティストを軸に自分に合う音楽を探し，そのアーティストのファンとしてのアイデンティティを構築していくような構成概念であると解釈できる．従って，F3 を**ファンアイデンティティ構築への期待**と呼ぶ．F4 は，**アーティストを応援するために楽曲を共有したい**という項目において最大の因子負荷量を持つ. また，**まだ世の中に知られていない音楽を勧めたい**と**世の中で流行っている楽曲に興味を持って聴く**の項目における因子負荷量は，それぞれ無視できない絶対値の正と負の値を持つ．これは，マイナーなアーティストを応援することにより構築されるファンアイデンティティに関する構成概念と考えられるため，これを**アーティスト貢献に対する期待**と呼ぶ.

Cronbach の α 係数は、内部一貫性の信頼性を示す指標である．各因子について因子負荷量が最も大きい変数群の Cronbach の α 係数を計算すると，F1 は 0.82，F2 は 0.73，F3 は 0.72 であった．F4 は因子負荷量が最も大きい変数が 1 つしかなく計算不可となる．得られた 4 つの因子の因子得点と年齢との相関行列を表 2 に示す.

表 1　探索的因子分析で得られた因子負荷量．太字箇所は各変数毎の最大の因子負荷量を示す．

質問項目	F1: 対人関係への期待	F2: 自己アイデンティティ構築への期待	F3: ファンアイデンティティ構築への期待	F4: アーティスト貢献への期待
相手が好きそうな音楽を勧めたい	**0.868**	−0.003	−0.017	0.034
定番の音楽を勧めたい	**0.805**	−0.050	0.037	−0.026
良い音楽を見つけた時に勧めたい	**0.641**	0.080	0.110	0.085
世の中で流行っている楽曲に興味を持って聴く	**0.458**	0.220	0.140	−0.395
知らないアーティストや楽曲を積極的に探す	−0.066	**0.753**	0.155	0.024
音楽を受動的に探す	0.281	**0.531**	−0.151	−0.105
まだ世の中に知られていない音楽を勧めたい	0.304	**0.461**	−0.016	0.395
好きなアーティストの新曲を常にチェックして聴く	0.083	0.013	**0.726**	0.010
自分に合うアーティストや楽曲を積極的に探す	−0.025	0.458	**0.459**	−0.055
アーティストを応援するために楽曲を共有したい	0.383	0.083	0.286	**0.422**

表 2　因子得点と年齢の相関係数行列．
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

変数	F1	F2	F3	F4	年齢
F1	1	0.566***	0.271***	0.454***	−0.237***
F2		1	0.130**	0.592***	−0.207***
F3			1	0.138**	−0.026
F4				1	−0.271***
年齢					1

F1–F2，F2–F4 はそれぞれ 0.5 以上の比較的大きな相関係数を持つ．

3·2　ロジスティック回帰分析

探索的因子分析で得られた 4 つの構成概念が音楽消費手段の使用に及ぼす影響を調べるため，ロジスティック回帰分析を行った．説明変数は，4 つの構成概念の因子得点，性別，および年齢とした．性別は男性を 0，女性を 1 としたダミー変数，年齢は連続変数とした．従属変数は，各音楽消費手段の利用有無であり，5 つの手段それぞれ 1 つモデルを作成した．5 つのモデルの結果を表 3 に示す．ここで，R^2 は Nagelkerke の R^2 である．分析の結果，サブスク利用については，F2: **自己アイデンティティ構築への期待**のみ，統計的有意な寄与が示された．これは，聴き放題のサブスクが低コストでの楽曲探索手段を提供していることによると考えられ，[Datta 18] の結果を支持している．購入とライブ利用については，F3: **ファンアイデンティティ構築への期待**と F4: **アーティスト貢献への期待**がそれぞれ統計的有意に寄与することが示された．特にライブ利用については，F3，F4 ともに購入利用よりも寄与度が大きく，かつ女性であることも統計的有意に寄与している．レンタルと無料の利用については，5%水準で統計的有意な寄与を示す要因は確認できなかった．

4.　まとめと今後の展望

本研究では，音楽に期待する役割と音楽消費手段の選択との関係を明らかにした．探索的因子分析により，[Lonsdale 11] と一致する 2 つの構成概念，**対人関係への期待**と**自己アイデンティティ構築への期待**が特定された．また，アイデンティティの構築に関して，**自己アイデンティティ構築への期待**と**ファンアイデンティティ構築への期待**という 2 つの構成概念が得られた．このことは，[Stevens 10] で指摘されている日本のポップカルチャー特有のファンダムが影響していると考えられ，**ファンアイデンティティ構築への期待**は日本市場に特有の構成概念である可能性がある．さらに，ファンに関する別の構成概念として，**アーティスト貢献への期待**が得られた．これは，特にマイナーアーティストのファン心理が他のファン心理とは異なることを意味する．

ロジスティック回帰分析により，サブスク，購入，ライブの各消費手段の利用に影響を与える構成概念を得た．サブスクの利用には**自己アイデンティティ構築への期待**のみ有意に寄与し，他の手段の利用には寄与していないという結果になった．このことは，音楽市場に新たに登場したサブスクが，音楽による自己アイデンティティの構築を促進していることを示唆している．また，購入とライブという 2 手段の利用に対しては，2 つの構成概念 (**ファンアイデンティティの構築への期待**，**アーティスト貢献への期待**) が寄与している．特に，ライブ利用への寄与がどちらも大きく，アーティストファンにとってライブが特に重要であると言える．これらの結果から，音楽市場においてサブスクは，他の手段と役割を分担して新たな価値を提供していると言える．したがって，今後も，サブスクの市場シェアは旧来の購入やライブといった手段に完全に取って代わることはなく，その前に飽和状態になると考えられる．

一方で，**対人関係への期待**の構成概念は，5 つの手段のいずれに対しても有意な効果を示さなかった．このことから，現在の音楽市場における主要な手段が，音楽を通じた他者とのコミュニケーションを重視する音楽リスナーにとって魅力的でない可能性があり，彼らの需要を満たす別の手段が必要であることが示唆された．サブスクは音楽市場で徐々に成熟しつつあり，他と差別化を図

表 3　ロジスティック回帰分析で得られた回帰係数と Nagelkerke の R^2．*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

	モデル 1 サブスク	モデル 2 購入	モデル 3 ライブ	モデル 4 レンタル	モデル 5 無料
切片	0.088	−0.651*	−1.594***	−1.218***	0.140
F1	−0.014	−0.214	−0.084	0.084	0.234
F2	0.444**	−0.121	0.010	0.139	−0.149
F3	0.045	0.255*	0.473***	0.102	0.095
F4	0.023	0.764***	0.824***	0.028	0.089
性別	−0.174	0.084	0.730**	−0.217	0.083
年齢	−0.011	0.009	0.007	−0.000	0.004
R^2	0.082	0.126	0.225	0.020	0.022

る際のヒントとなるという意味で，この結果はビジネス観点でも有用である．これに関して，音楽ストリーミングサービス AWA のオンライン空間「LOUNGE」は，リスナー同士がリアルタイムで同じ音楽を聴きながらコミュニケーションできる機能であり，この差別化を体現する機能と言えるかもしれない．

最後に，限界と今後の展望に言及する．本研究は日本の音楽市場に焦点を当てたものであるが，国別の差異がある可能性がある．特に，ファンに関する構成要素は他の国では検出されないか，別の形で現れる可能性がある．音楽を聴く手段については，現代の音楽市場で主要な手段をカバーする 5 つの分類を採用したが，今後も手段の多様化は続くと思われる．特に，SNS での音楽視聴については，無料手段を細分化することで新たな知見が得られる可能性がある．COVID-19 の効果については，本研究の関心の範囲外であるが，音楽消費行動に対する長期的・短期的な効果は興味深いテーマである．

謝　　辞

本研究の遂行にあたり，筑波大学 佐野幸恵准教授から多くの助言をいただきました．また，株式会社 AI Shift の二宮大空さん，同僚の高野雅典さん，他の皆様から有益なコメントをいただきました．深く感謝致します．

◇ 参　考　文　献 ◇

[Chamorro-Premuzic 07]　Chamorro-Premuzic, T. and Furnham, A.: Personality and Music: Can Traits Explain How People Use Music in Everyday Life?, *British journal of psychology (London, England : 1953)*, Vol. 98, pp. 175–85 (2007)

[Chellappa 05]　Chellappa, R. K. and Shivendu, S.: Managing Piracy: Pricing and Sampling Strategies for Digital Experience Goods in Vertically Segmented Markets, *Information Systems Research*, Vol. 16, No. 4, pp. 400–417 (2005)

[Datta 18]　Datta, H., Knox, G., and Bronnenberg, B. J.: Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery, *Marketing Science*, Vol. 37, No. 1, pp. 5–21 (2018)

[Dewan 12]　Dewan, S. and Ramaprasad, J.: Research Note — Music Blogging, Online Sampling, and the Long Tail, *Information Systems Research*, Vol. 23, No. 3-part-2, pp. 1056–1067 (2012)

[Ikegami 21]　Ikegami, S., Sato, N., Hatoh, T., Ikoma, S., Miyazawa, S., Konishi, J., and Hoshino, E.: The psychological functions of and individual differences in music listening in Japanese people, *The Japanese Journal of Psychology*, Vol. advpub, p. 92.20005 (2021)

[Lacher 94]　Lacher, K. T. and Mizerski, R.: An Exploratory Study of the Responses and Relationships Involved in the Evaluation of, and in the Intention to Purchase New Rock Music, *Journal of Consumer Research*, Vol. 21, No. 2, pp. 366–380 (1994)

[Lonsdale 11]　Lonsdale, A. J. and North, A. C.: Why do we listen to music? A uses and gratifications analysis, *British Journal of Psychology*, Vol. 102, No. 1, pp. 108–134 (2011)

[Manolika 21]　Manolika, M., Baltzis, A., and Gardikiotis, A.: Individual Differences in Music Listener Motivations: The Neglected Values, *Empirical Studies of the Arts*, Vol. 39, No. 1, pp. 17–35 (2021)

[Papies 11]　Papies, D., Eggers, F., and Wlömert, N.: Music for free? How free ad-funded downloads affect consumer choice, *Journal of the Academy of Marketing Science*, Vol. 39, No. 5, pp. 777–794 (2011)

[Stevens 10]　Stevens, C. S.: You Are What You Buy: Postmodern Consumption and Fandom of Japanese Popular Culture, *Japanese Studies*, Vol. 30, No. 2, pp. 199–214 (2010)

[Takeuchi 22]　Takeuchi, M., Morishita, S., and Sano, Y.: Music Roles Affect the Selection of Consumption Means: A Questionnaire Survey of People's Expectations for Music and Exploratory Factor Analysis, *The Review of Socionetwork Strategies*, Vol. 16, No. 2, pp. 453–464 (2022)

[Wlömert 16]　Wlömert, N. and Papies, D.: On-demand streaming services and music industry revenues — Insights from Spotify's market entry, *International Journal of Research in Marketing*, Vol. 33, No. 2, pp. 314–327 (2016)

[Masae 13]　Masae, Y.: Popular Music and Female fans, *Journal of the Faculty of Global Communication*, No. 14, pp. 265–276 (2013)

著　者　紹　介

武内慎

2015 年 株式会社サイバーエージェントに中途入社．2024 年 筑波大学大学院 システム情報工学研究群 博士後期課程修了 (社会工学)．メディアサービスのデータ分析に従事．

森下　壮一郎

2005 年 埼玉大学大学院 理工学研究科 博士後期課程中退．2009 年 同大学 博士（工学）．東京大学，電気通信大学，理化学研究所を経て，2016 年より株式会社サイバーエージェント．メディアサービスのデータ分析や社会的受容性調査等に従事．

チャットボット事業における Dense Retrieverを用いたZero-shot FAQ検索

二宮 大空
Hirotaka Ninomiya
株式会社 AI Shift
Data Scientist
ninomiya_hirotaka@cyberagent.co.jp

邊土名 朝飛
Asahi Hentona
株式会社 AI Shift
Data Scientist
hentona_asahi@cyberagent.co.jp

杉山 雅和
Masakazu Sugiyama
株式会社 AI Shift
Data Scientist
sugiyama_masakazu@cyberagent.co.jp

戸田 隆道
Takamichi Toda
株式会社 AI Shift
Data Scientist
toda_takamichi@cyberagent.co.jp

友松 祐太
Yuta Tomomatsu
株式会社 AI Shift
Data Scientist
tomomatsu_yuta@cyberagent.co.jp

keywords: 情報検索, FAQ, 自然言語処理

Summary

チャットボットが提供する機能の一つに，よくある質問集（FAQ: Frequently Asked Questions）を用いてユーザの質問に回答するFAQ検索がある．我々はFAQ検索手法としてオープンドメイン質問応答で有効なDense Retrieverを，チャットボット事業の対話ログをもとに作成したデータで学習させた．ただし，この学習では対話ログが十分に存在する既存顧客では比較的高精度で検索できる一方で，新たに導入される新規顧客では十分な検索精度を保持することが難しい．そこで我々は，FAQ検索において効果的な負例選択方法を探索し，GPT-2を用いた訓練データ拡張の有効性の検証を行った．実験の結果，新規顧客を想定した評価において正解率が3.2%向上した．

1. はじめに

現在カスタマーサポートの分野においてチャットボットが注目を集めており，多くの企業においてユーザの課題解決を促すツールとして導入されている．そして，チャットボットが提供する機能の一つであるFAQ検索は現在盛んに研究が行われている [Sakata 19][Mass 20].

FAQ検索では，顧客のFAQとユーザの質問の間で表記揺れがあり，不明瞭な質問が与えられることが多いため，単語の表層形だけでなく発話全体の意味を考慮して検索する必要がある．そこで，我々は事前学習済み言語モデルを用いた検索モデルを構築する．特に，オープンドメイン質問応答や日本語クイズタスクJAQKET[鈴木 20]において有効性が確認された[加藤 20]Dense Passage Retrieval（DPR）[Karpukhin 20]の検索部であるDense Retrieverを利用する．ここで，Dense Retrieverは与えられた質問と検索対象である文書それぞれを個別にベクトルに変換するDual-Encoderである．DPRは検索を行うDense Retrieverと回答抽出を行うReaderから構成されるが，本稿で扱うFAQ検索は回答抽出を行わないため，Readerは利用しない．

Karpukhinらによると，Dense Retrieverの学習には正例と判別が困難な負例を用いることで検索精度が向上することが確認されている [Karpukhin 20]．ただし，オープンドメイン質問応答では検索対象となる文書数が1,000万件を超える大規模なものである一方で，我々が扱うFAQ検索では検索対象となるFAQ質問は顧客ごとに用意された平均数百件程度の小規模なFAQ DBである．そこで我々はFAQ検索においても同様の傾向があるかを調査するために，複数の負例選択方法を比較する．さらに，対話ログが存在せず訓練データが作成できないような新規顧客においても十分な検索精度をもつ検索モデルを得るために，GPT-2[Radford 19]を用いた訓練データの拡張を行う．

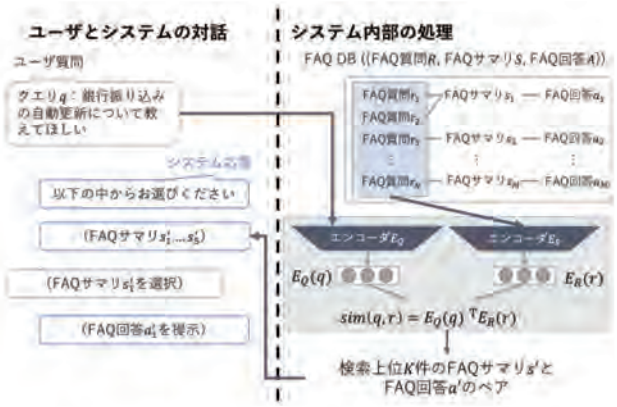


図 1　Dense Retriever を用いた FAQ 検索の概要図

2. Dense Retriever の学習

図 1 に Dense Retriever を用いた FAQ 検索の概要図を示す．FAQ 検索とは，事前に用意された FAQ のデータベース（FAQ DB）から，ユーザの質問に最もマッチする回答を選択するタスクと考えられる．例えば図 1 のように，「銀行振り込みの自動更新について教えて欲しい」といったユーザ質問が入力された場合，システムは事前に用意された FAQ DB を検索する．ここで FAQ DB は (1)FAQ 質問，(2)FAQ サマリ，(3)FAQ 回答を 1 組とする FAQ の集合である．FAQ サマリがユーザからよく問い合わせられる質問文を，FAQ 回答はそれに対する回答文を表し，FAQ 質問は FAQ サマリの複数の言い換え表現を表す．FAQ 質問を複数設定することにより検索性能が向上することが知られている．検索結果として最大 5 件の FAQ サマリを回答候補としてユーザに提示し，ユーザがいずれかを選択すると，対応する FAQ 回答が出力される．

学習時，Dense Retriever はユーザ質問 q に対する Encoder の出力ベクトルと FAQ サマリ s に対する Encoder の出力ベクトルの類似度を式 1 により算出する．

$$\mathrm{sim}(q,s) = E_Q(q)^T E_R(s) \tag{1}$$

ここで E_Q,E_R はそれぞれユーザ質問と FAQ サマリに対する Encoder を，^{T} は転置を表す．q_i をユーザ発話，s_i^+ を正例の FAQ サマリ，$s_{i,j}^-$ を負例の FAQ サマリ，n を負例の FAQ サマリの総数とすると，訓練データは $D = \{< q_i, s_i^+, s_{i,1}^-, ..., s_{i,n}^- >\}$ と表される．Dense Retriever は式 2 の損失関数が最小になるように学習される．

$$L\left(q_i, s_i^+, s_{i,1}^-, ..., s_{i,n}^-\right) = -\log\left(\frac{e^{\mathrm{sim}\left(q,s_i^+\right)}}{e^{\mathrm{sim}\left(q,s_i^+\right)} + \sum_{j=1}^{n} e^{\mathrm{sim}\left(q,s_{i,j}^-\right)}}\right) \tag{2}$$

ユーザ質問と正例の FAQ サマリが類似しているほど損失が小さくなり，負例の FAQ サマリとの類似度が高いほど，式 2 の損失は大きくなる．そのため，高精度の Dense Retriever を得るためには負例の選択方法が重要である．負例の選択に関する詳細は 3·1 節で述べる．

推論時はユーザの多様な言い回しに対応するために，FAQ サマリではなく FAQ 質問を用いて検索を行う．つまり，FAQ 質問を r とすると式 3 により類似度を計算し，最も類似度が高い FAQ 質問に対応する FAQ サマリを選択する．

$$\mathrm{sim}(q,r) = E_Q(q)^T E_R(r) \tag{3}$$

3. データセットの作成

自社のチャットボット事業 *1で収集した対話ログを用いてデータセットを作成する．まず，チャットボット事業における対話ログ収集時の FAQ 検索について述べる．最初に，入力されたユーザ質問に対して (1) 単語の表層形に基づく BM25[Robertson 09] による検索と，(2)BERT の出力ベクトルの類似度をもとに近傍探索ツール faiss*2 を用いた検索を行う．そして，(1) の検索結果上位 3 件と (2) の検索結果の重複しない上位 2 件の FAQ サマリの組をユーザに提示する．ユーザはそのいずれかを選択し，システムはそれに紐づく FAQ 回答を出力する．我々はユーザ質問と選択された FAQ サマリを正例として収集し，データセットを作成した．この FAQ サマリは対象の顧客が持つ FAQ DB 内の FAQ サマリのいずれかである．データセットは複数の顧客の FAQ DB と対話ログから作成されており，顧客ごとに事例数と FAQ 数が異なる．

実験では Known-Domain と Unknown-Domain という 2 つの実験設定を設け，8 対 2 の割合で 2 つの顧客群にランダムに分割した．Known-Domain の顧客の場合は訓練データが存在し，Unknown-Domain の顧客の場合は訓練データが存在しないようにした．これはそれぞれ，既に一定数の対話ログを収集できる既存顧客における評価と，新規顧客における評価を行うためである．顧客ごとのデータセットの作成時，Known-Domain の顧客では訓練データ，検証データ，評価データの件数が 8:1:1 の割合となるように，Unknown-Domain の顧客では検証データ，評価データの件数の割合が 1:1 の割合になるようにランダムに分割した．作成したデータセットの詳細を表 1 に示す．

表 1　データセットの統計

実験設定	顧客数	訓練	検証	評価
Known-Domain	20	147,401	18,426	18,434
Unknown-Domain	6	0	13,187	13,189

訓練，検証，評価はそれぞれ訓練データ，検証データ，評価データの件数を表す．

*1　https://www.ai-messenger.jp
*2　https://github.com/facebookresearch/faiss

3·1 負例の選択方法

Dense Retriever の学習時，FAQ 検索において効果的な負例の選択方法を探索するため，以下の 4 通りを比較した．

TargetTenant
ユーザ質問に対する顧客の FAQ サマリの集合．

PositiveCases
ユーザ質問に対する顧客の FAQ サマリの内，バッチ内に存在する正例の FAQ サマリの集合．

AllRandom
ユーザ質問に対する顧客の FAQ サマリの集合に対して，Known-Domain の顧客の FAQ サマリをランダムに複数件追加したもの．

AllBm25
ユーザ質問に対する顧客の FAQ サマリの集合に対して，Known-Domain の顧客の FAQ サマリから BM25 によるユーザ質問との類似度が高い順に複数件追加したもの．

TargetTenant は，検索対象の顧客の FAQ DB からのみ負例を選択したため，推論時と最も近い学習方法であると考えられる．次に，繰り返し正例として出現する FAQ を答えと適切に判別することが重要であると考えて PositiveCases を設定した．また，本研究では複数の顧客の FAQ DB が存在することから，対象外の顧客の FAQ DB を学習に用いた場合である AllRandom を設定した．そして，より困難な負例を作成するために対象外の顧客の FAQ DB から正解 FAQ サマリに類似する FAQ サマリを用いた AllBm25 を設定した．実験では計算コストの観点から AllRandom と AllBm25 は検索対象となる FAQ サマリ数が最大 320 件となるように FAQ サマリを追加した．

3·2 GPT-2 を用いた訓練データの拡張

Dense Retriever の訓練データ中には一度も選択されない FAQ サマリが存在し，それらは推論時に選択されにくい傾向がある．そこで，GPT-2 を用いて FAQ サマリからそれを正解とするユーザ質問を擬似的に生成することで，訓練データを拡張する．これにより全ての FAQ サマリに対して正解がある訓練データ事例を作成することができるため，頑健性の高いモデルとなることが期待できる．

GPT-2 は Known-Domain の顧客の訓練データを用いて，正例の FAQ サマリからユーザ質問を生成するように Fine-tuning を行った．その後，学習済みの GPT-2 に FAQ サマリを入力し，ユーザ質問を擬似的に生成することで，生成したユーザ発話と FAQ サマリの組を 1 事例として訓練データに追加した．ただし，生成したユーザ発話が正例の FAQ サマリと一致する場合は除外した．

実験では，(1)Known-Domain の顧客のデータを拡張する場合（*Known*），(2)Unknown-Domain の顧客のデータを拡張する場合（*Unknown*），(3) 全ての顧客のデータを拡張する場合（*All*）に分けて実験を行った．

4. 実　　験

Dense Retriever の検索性能の評価を Known-Domain と Unknown-Domain に分けて評価を行った．さらに，ベースライン手法として BM25 と，2 章で述べた学習を行わない場合の Dense Retriever との性能比較を行った．

4·1 実 験 設 定

Dense Retriever を構成する 2 つの Encoder の重みの初期値には，公開されている日本語事前学習済み BERT[*3] の重みを用いた．Fine-tuning 時の学習率は 1×10^{-5}（warmup rate=10%）とし，Epoch は 10，バッチサイズは 64，Dropout 率は 0.3，Optimizer は Adam を用いた．GPT-2 は公開されている日本語事前学習済みモデル [*4] の重みを初期値に利用しており，学習率は 1×10^{-5}，バッチサイズは 8，Dropout 率は 0.1，Optimizer は Adam を用いた．

評価時，顧客ごとに Top{1, 5, 10} Accuracy を算出し，それらを平均することで Macro Average Top{1, 5, 10} Accuracy を算出した．

また，ベースライン手法と負例選択方法で利用する BM25 では，顧客ごとの FAQ DB に含まれる FAQ 質問の集合を IDF 値の計算に利用し，名詞と動詞原型のみ用いて計算した．

4·2 実 験 結 果

表 2 の Macro Average Top1 Accuracy で各手法を比較する．

まず，GPT-2 を用いた訓練データ拡張を行わず 4 通りの負例選択を比較した結果，Known-Domain では TargetTenant が，Unknown-Domain では PositiveCases が検索精度が最も高い．さらに，AllRandom と AllBm25 は TargetTenant よりも検索性能が低下している．これは，FAQ 検索ではオープンドメイン質問応答と比べて検索対象の文書が小規模であるため，他の顧客の FAQ DB から負例を選択するよりも，推論時と同じ顧客の FAQ DB から負例を選択して，それらを適切に判別できるよう学習した方が良かったと考えられる．また，AllBM25 の結果より，BM25 を用いて正例との判別が困難な負例を選択することで Dense Retriever の性能が向上することが知られているが [Karpukhin 20]，我々の FAQ 検索に関する実験では同様の傾向は確認できなかった．

ベースラインとなる BM25 は Dense Retriever と同等の検索精度がある．これは，チャットボット事業ではユーザ質問が名詞のみの場合や短文の場合が存在するため，そのような事例に対しては単語の表層形に基づく検索が効果的であったと考えられる．

次に，これまでの実験で効果があった TargetTenant と PositiveCases の 2 つの負例選択において，GPT-2 を用

*3 https://huggingface.co/cl-tohoku/bert-base-japanese
*4 https://huggingface.co/rinna/japanese-gpt2-medium

表 2　Macro Average Top{1,5,10} Accuracy （%）

モデル	訓練データ拡張	負例選択	Known-Domain	Unknown-Domain
BM25	-	-	27.8/56.1/64.7	31.7/57.6/65.4
DR w/o training	-	-	11.1/29.1/38.3	17.0/34.1/43.9
DR	-	TargetTenant	**34.5**/60.6/70.3	29.8/57.1/67.8
DR	-	PositiveCases	33.3/59.3/69.8	31.5/57.0/68.4
DR	-	AllRandom	33.5/59.4/69.4	29.3/55.6/67.3
DR	-	AllBm25	33.1/61.1/71.3	29.3/55.6/66.4
DR	*Known*	TargetTenant	33.1/**62.1/72.4**	**32.0**/59.2/**70.2**
DR	*Unknown*	TargetTenant	33.4/59.6/72.0	31.7/58.8/69.1
DR	*All*	TargetTenant	31.9/61.0/70.1	31.2/58.0/69.6
DR	*Known*	PositiveCases	32.1/59.4/71.4	31.5/59.3/69.9
DR	*Unknown*	PositiveCases	31.4/59.5/71.3	30.7/**60.1**/69.5
DR	*All*	PositiveCases	33.3/59.4/70.8	30.1/58.7/69.9
BM25 + DR	*All*	TargetTenant	33.3/**62.8/74.2**	**36.6/66.4/75.7**

DR は Dense Retriever を表し，DR w/o training はチャットボット事業のデータを用いた学習を行っていない，BERT の重みの初期値を用いた Dense Retriever を表す．実験結果は，モデルと訓練データ拡張と負例選択の組み合わせに対して，Known-Domain と Unknown-Domain ごとに評価し，Macro Average TopK Accuracy を K={1, 5, 10} の順に”/”区切りで記す．

いた訓練データの拡張の有効性を検証した．その結果，負例選択が TargetTenant の場合，訓練データ拡張により Known-Domain で低下し，Unknown-Domain で増加した．つまり，訓練データが十分に存在する Known-Domain では拡張したデータがノイズとなったものの，訓練データが存在しない Unknown-Domain では一定の効果があったと考えられる．

最後に，訓練データ拡張 *All* を行い，TargetTenant で負例選択した Dense Retriever と BM25 のスコアの重み付け和を用いて検索した結果，Known-Domain と Unknown-Domain 共に比較的高い検索精度を達成した．特に Unknown-Domain においては，Dense Retriever と BM25 を組み合わせることで 4.6%の向上が確認された．

5. お　わ　り　に

本稿では，チャットボット事業におけるデータを用いて Dense Retriever の学習を行い，訓練データが存在しないような新規顧客を想定した Zero-shot FAQ 検索の性能評価を行った．実験では，新規顧客においても高い検索精度を得るために，Dense Retriever の学習時の負例選択方法を 4 通り比較し，さらに GPT-2 を用いた訓練データ拡張の有効性を示した．

現在我々は音声対話を通してユーザの課題解決を行うボイスボット事業を提供しており，チャットボット事業同様に FAQ 検索機能を実装している．ただし，ボイスボット事業では入力が音声となるため，音声認識誤りやユーザ発話の不明瞭さといった新たな課題があることがわかった．今後はそれらの解決を通して，頑健な FAQ 検索モデルの構築を目指す．

◇ 参　考　文　献 ◇

[Karpukhin 20]　Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t.: Dense Passage Retrieval for Open-Domain Question Answering, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online (2020), Association for Computational Linguistics

[Mass 20]　Mass, Y., Carmeli, B., Roitman, H., and Konopnicki, D.: Unsupervised FAQ Retrieval with Question Generation and BERT, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 807–812, Online (2020), Association for Computational Linguistics

[Radford 19]　Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners, *OpenAI blog*, Vol. 1, No. 8, p. 9 (2019)

[Robertson 09]　Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: BM25 and beyond, *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 4, pp. 333–389 (2009)

[Sakata 19]　Sakata, W., Shibata, T., Tanaka, R., and Kurohashi, S.: FAQ retrieval using query-question similarity and BERT-based query-answer relevance, in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1113–1116 (2019)

[加藤 20]　加藤拓真, 宮脇峻平, 西田京介, 鈴木潤：オープンドメイン QA における DPR の有効性検証, 言語処理学会第 26 回年次大会, pp. 237–240 (2020)

[鈴木 20]　鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也：JAQKET: クイズを題材にした日本語 QA データセットの構築, 言語処理学会第 26 回年次大会, pp. 237–240 (2020)

著　者　紹　介

二宮　大空

2021 年奈良先端科学技術大学院大学先端科学技術研究科博士前期課程修了後，国立研究開発法人情報通信研究機構有期雇用研究員として質問応答に関する研究に従事．2022 年サイバーエージェントに入社．現在は株式会社 AIShift で FAQ 検索の開発に従事．

2023 Vol.1

Text-aware Color Recommendation in Vector Graphic Documents

邱 倩如
Qiu Qianru
AI Lab
Research Scientist
qiu_qianru@cyberagent.co.jp

汪 雪テイ
Wang Xueting
AI Lab
Research Scientist
wang_xueting@cyberagent.co.jp

keywords: Multimodal learning, color recommendation, palette generation

Summary

Color selection plays a critical role in graphic document design and requires sufficient consideration of various contexts. However, recommending appropriate colors which harmonize with the other colors and textual contexts in documents is a challenging task, even for experienced designers. In this study, we propose a multimodal masked color model that integrates both color and textual contexts to provide text-aware color recommendation for graphic documents. Our proposed method primarily focuses on color palette completion, which recommends colors based on the given colors and text. Additionally, it is applicable for full palette generation, which generates a complete color palette corresponding to the given text. Experimental results demonstrate that our proposed approach outperforms state-of-the-art methods in color palette completion and full palette generation.

1. Introduction

A color palette refers to a specific set of colors in refined forms. A well-chosen color palette helps communicate the intended message effectively in design. In this study, the primary objective of color recommendation is color palette completion, which means suggesting appropriate colors in palettes, taking into consideration both the existing color and text contexts. In addition, we address the full palette generation task, which is a specific case of color recommendation. This task focuses on creating a set of harmonious colors based only on the given textual contexts, without factoring in any other color context. In this study, we present a versatile model that can effectively perform both color recommendation tasks.

In recent years, data-driven deep learning techniques have shown potential for color recommendation in graphic documents. A recent study [Qiu 23] proposed a masked color model for multi-palette representation and color recommendation. However, this work only examined the relationships among colors in multiple palettes. Some studies based on multi-modality learning have aimed to generate a color palette based on textual information for image colorization [Bahng 18, Maheshwari 21]. They proposed conditional GAN architectures to generate colors that reflect the semantics of input text for image colorization. However, the small scale of training data restricted the capacity to encode and represent complex textual information. In contrast, using text embeddings generated by large language models is more comprehensive than learning from the small-scale dataset.

In this paper, we propose a multimodal masked color model for text-aware color recommendation in graphic documents using the designed attention networks. We utilize the pre-trained CLIP model to obtain comprehensive text embeddings that can represent both textual and visual features, enabling us to consider a broader range of text representation for color recommendation. We primarily conduct a series of evaluations on color palette completion for graphic documents and full palette generation to validate the versatility and effectiveness of our proposed method.

2. Approach

As shown in 図 1, we extract multiple color palettes from image, graphic, and text element groups and reorder the color positions. Graphics encompass decorative elements as shapes and other vector-based illustrations. Additionally, we extract varied textual contexts in image and text elements, and obtain text embeddings using pre-trained large language models. We apply a text-aware masked color model to integrate the color and text representations for color sequence completion and predict the specified colors with a high probability for color recommendation.

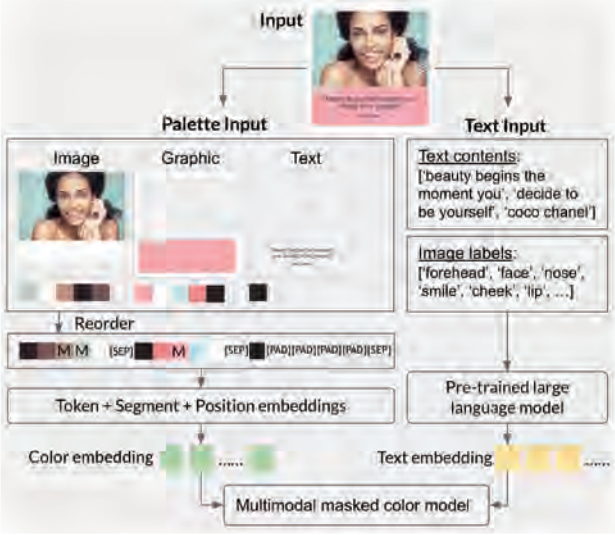


図 1 Overview of our approach.

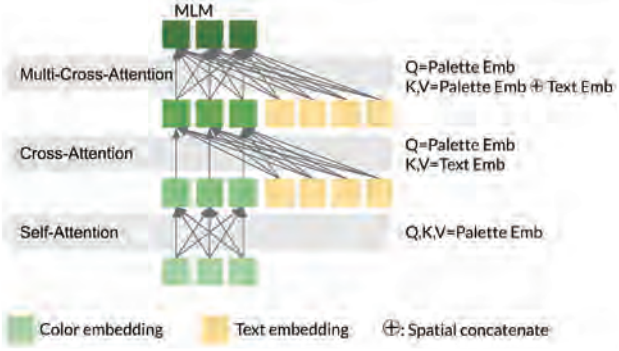


図 2 Multimodal masked color model including one self-attention and two cross-attention networks with color and text input representations.

2·1 Reordered color representation

To extract the color palette for image and graphic element groups, we adopt the k-means clustering method. For text element groups, we collect individual text colors and cluster them into a palette. In the related work by Qiu *et al.* [Qiu 23], colors in each palette are ordered according to the size of their color clusters, reflecting their color area size. However, we found that area-based color ordering does not positively impact model performance. One possible explanation is that the size of a color area is not necessarily proportional to its importance or relevance in the overall color context. On the other hand, lightness provides a simple and intuitive way to organize colors. In this approach, we utilize lightness in CIELAB color space that has perceptual uniformity as the basis for color order, which we believe will provide a richer color context.

Color representation can be divided into three primary stages: color quantization, encoding quantized colors into learnable embeddings, and training with a masked autoencoder approach. Color quantization aims to reduce the number of colors. In this work, RGB color data is first converted to CIELAB color space, with a range of [0, 255], and each color is assigned to one of the bins in a $b \times b \times b$ histogram, with $b = 16$. Then quantized color codes are encoded into vectors and embedded in the space during the learning progress. For palettes shorter than the fixed maximum length, we add the [PAD] token to complete the length. Additionally, the [SEP] token is appended at the end of a palette. The palettes of image, graphic, and text elements, are respectively labeled with the segment number 1, 2, and 3. Given a sequence of multiple palettes, we construct its input representation by summing the corresponding token embeddings, segment embeddings, and position embeddings.

2·2 Text representation

Textual information plays a crucial role in conveying semantic concepts and achieving design goals. Color variations often rely on figurative language [Kawakami 16], such as murkey blue, greeny blue, jazzy blue (best viewed in color). Design works have various forms of textual information. In this study, we gather two types of text from the graphic documents: text contents and image labels, as demonstrated in 図 1. Text contents are obtained directly from the original text elements. Image labels are extracted from image elements using Google Vision API to detect objects.

In this study, we use the CLIP [Radford 21] model to obtain the text embeddings of each text content and image label, which is one of pre-trained large language models that have made significant advancements in recent years. CLIP is pre-trained on a large corpus of images and their associated captions, which allows it to generate embeddings that capture both textual and visual information. The colors present in the input image can be correlated with the textual information during the training process, and we believe that this connection has the potential to improve the performance of text-aware color models. We employ the CLIP model to obtain text embeddings for each text content and image label.

2·3 Text-aware masked color model

As shown in 図 2, our proposed multimodal masked color model is composed of one self-attention module and two cross-attention modules, designed to effectively integrate color and textual contexts. An attention module is described as mapping a query (Q) and a set of key (K) and value (V) pairs to an output [Vaswani 17]. In the self-attention module, queries, keys, and values correspond to

input color embeddings. This module can effectively capture the intra-relationship among colors within the same palette and the inter-relationship between different palettes. To model the inter-relationship between colors and text, we introduce cross-attention modules. In the cross-attention module, queries are color embeddings from the self-attention module, while keys and values correspond to text embeddings. Additionally, the multi-cross-attention module is a comprehensive network that captures both intra-relationship and inter-relationship in colors and text. In the multi-cross-attention module, queries are color embeddings from the cross-attention module, whereas keys and values are obtained by concatenating color embeddings from the cross-attention layer and text embeddings.

The masked color model following a methodology similar to a masked language model (MLM) [Devlin 18], randomly masks the color tokens from the input, and then predicts the masked tokens based on their contexts. Once the color model is trained, it can then be utilized to predict specific colors with a high probability and generate color palettes that harmonize with textual contexts.

3. Experiments

In order to demonstrate the effectiveness of our proposed model, we conducted the experiments on color palette completion and full palette generation tasks. We collected the Multi-Palette-And-Text dataset compiled from the Crello-v2 dataset [Yamaguchi 21] for color palette completion task. It contains multiple palettes for image, graphic, and text elements, and textual information of text contents and image labels. The resulting dataset comprises 14,020 / 1,704 / 1,712 valid data for training, validation, and testing, all of which contain image-graphic-text palettes and English text contents. In addition, we used the Palette-And-Text (PAT) dataset [Bahng 18] for full palette generation task. The PAT dataset contains the five-color palette and text pairs. We randomly divide it into 8,147 / 1,018 / 1,018 data for training, validation, and testing.

3·1 Color palette completion

We primarily conducted a comparison on the color palette completion task between our proposed text-aware masked color model with the related work by Qiu *et al.* [Qiu 23], which only incorporates color representation in its masked model. To evaluate model performance, we measured the accuracy metric by comparing the predicted colors with the ground-truth colors. It calculates the corrected predictions of color codes. The comparison results of our proposed method and the related works are shown in 表 1.

表 1 Comparison results of accuracy@1 for predicting different numbers of masked colors. '@1' indicates the recommended colors with the highest probability of each model.

	Accuracy@1↑				
Method	1 color	2 colors	3 colors	4 color	5 colors
[Qiu 23]	36.72%	16.04%	6.45%	2.51%	0.76%
Ours	**47.13%**	**26.22%**	**15.67%**	**10.14%**	**5.76%**

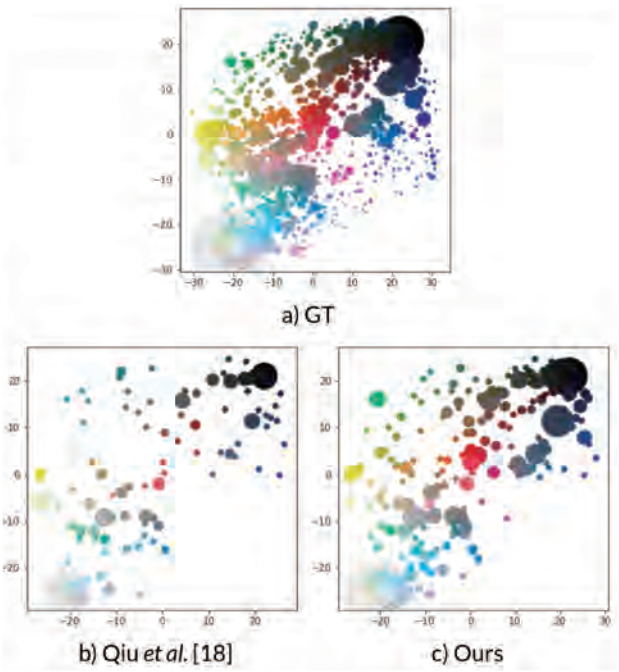


図 3 Color distributions that the data point is assigned with its own color and the point size reflects its frequency. a) displays all the ground truth colors in the multiple palettes of the test dataset. b) and c) display the correctly predicted three-color results of the related method [Qiu 23] and our method.

Our results indicated that our proposed method has higher accuracy than the method by Qiu *et al.* [Qiu 23] for predicting different numbers of masked colors.

To visualize the distribution of the correctly predicted colors, we utilized the 3-dimensional CIELAB color data and transform it into 2-dimensional space by t-SNE (t-Distributed Stochastic Neighbor Embedding). We selected each best-trained model with the highest validation accuracy and output the results for three-color prediction based on the test dataset. The distribution results are shown in 図 3. It demonstrated that the output of our model has denser points, indicating that our model had more accurate predictions, and the accurately predicted colors were not limited to black and white colors.

3·2 Full palette generation

We compared our proposed model with Text-to-Palette Generation Networks (TPN) in the most relevant work [Bahng

表 2　Comparison results of color diversity and palette similarity to GT.

	Diversity↑		Similarity to GT↓	
Generated palettes	Mean	Std	Mean	Std
TPN [Bahng 18]	22.21	10.78	29.26	13.35
Ours	**29.92**	10.27	**28.14**	12.91
GT	26.17	13.84	-	-

図 4　Qualitative analysis on textual context. We compare the generated palette results of our proposed method and the related work TPN [Bahng 18] with the ground truth.

18] on full palette generation. For evaluation, we adopt the color diversity evaluation in the related work [Bahng 18] that calculates the average pairwise distance between the five colors within a palette. In addition, we measure the similarity between the generated palettes and the ground truth (GT) with Dynamic Closest Color Warping method [Kim 21] that calculates the minimum distance between colors in different palettes. The comparison results of color diversity and palette similarity to GT in 表 2 indicated that our generated palettes have higher diversity and closer similarity to GT.

Moreover, we output a series of palettes based on various textual contexts. A comparison between generated palettes of our proposed method and the related work TPN with GT can be observed in 図 4. It is noted that higher diversity may not necessarily be a critical factor, as incorporating colors in the palette that are irrelevant to the text would increase diversity. On the other hand, palette similarity to GT holds greater importance, as it indicates whether the recommended results contain key colors that accurately convey the intended semantics.

4. Conclusion

In this paper, we presented a text-aware masked color model with reordering the color input based on lightness in CIELAB color space and CLIP-based text representation. Our method is improved to have greater performance on the accuracy compared to prior methods in recommending colors for graphic documents based on the given colors and textual contexts. Moreover, our proposal is applicable for full palette generation and surpasses related work on color diversity and palette similarity to the ground truth.

◇ 参　考　文　献 ◇

[Bahng 18]　Bahng, H., Yoo, S., Cho, W., Park, D. K., Wu, Z., Ma, X., and Choo, J.: Coloring with words: Guiding image colorization through text-based palette generation, in *Proceedings of the european conference on computer vision (eccv)*, pp. 431–447 (2018)

[Devlin 18]　Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018)

[Kawakami 16]　Kawakami, K., Dyer, C., Routledge, B. R., and Smith, N. A.: Character Sequence Models for Colorful Words, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1949–1954 (2016)

[Kim 21]　Kim, S. and Choi, S.: Dynamic closest color warping to sort and compare palettes, *ACM Transactions on Graphics (TOG)*, Vol. 40, No. 4, pp. 1–15 (2021)

[Maheshwari 21]　Maheshwari, P., Jain, N., Vaddamanu, P., Raut, D., Vaishay, S., and Vinay, V.: Generating Compositional Color Representations from Text, in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 1222–1231 (2021)

[Qiu 23]　Qiu, Q., Wang, X., Otani, M., and Iwazaki, Y.: Color Recommendation for Vector Graphic Documents based on Multi-Palette Representation, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3621–3629 (2023)

[Radford 21]　Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision, in *International conference on machine learning*, pp. 8748–8763PMLR (2021)

[Vaswani 17]　Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, *Advances in neural information processing systems*, Vol. 30, (2017)

[Yamaguchi 21]　Yamaguchi, K.: Canvasvae: learning to generate vector graphic documents, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5481–5489 (2021)

著　者　紹　介

邱　倩如

上海交通大学大学院を終了後、富士ゼロックス株式会社にてグラフィック自動生成の研究開発に携わる。2021 年にサイバーエージェントに中途入社。AI Lab ではデジタル広告クリエイティブ生成について研究を進めている。

汪　雪テイ

2018 年に名古屋大学大学院情報科学研究科博士後期課程修了。2018 年から東京大学大学院情報理工学系研究科で特任研究員、特任助教として、マルチメディア処理・ソーシャルネットワーク解析などの研究活動に従事。2021 年にサイバーエージェント入社。マルチメディアに関する研究に従事。

多様なタスク指向対話データの収集を目的としたクラウドソーシングにおけるインストラクションの設計

クリニック予約対話を例に

邊土名 朝飛 Asahi Hentona
株式会社 AI Shift
ML Engineer
hentona_asahi@cyberagent.co.jp

友松 祐太 Yuta Tomomatsu
株式会社 AI Shift
Data Scientist
tomomatsu_yuta@cyberagent.co.jp

佐々木 翔大 Shota Sasaki
理化学研究所，東北大学
shota.sasaki.yv@riken.jp

阿部 香央莉 Kaori Abe
東北大学
abe-k@tohoku.ac.jp

乾 健太郎 Kentaro Inui
理化学研究所，東北大学
kentaro.inui@tohoku.ac.jp

keywords: タスク指向対話，対話コーパス，クラウドソーシング

Summary

クリニック予約など，何らかのタスク達成を目的をとしたカスタマーとの接客では，オペレーターはカスタマーに対し提案や質問，説明を行い，場合によっては妥協するよう交渉する．このような柔軟なタスク指向対話システムを実現するためには，多様かつ大規模な対話データセットが必要である．大規模に対話データを収集する手段としてクラウドソーシングがあるが，単調な対話が行われないようインストラクション設計に注意を要する．本研究では，多様な対話データの収集を目的として，オペレーター役とカスタマー役のクラウドワーカーそれぞれに異なるインストラクションを生成・提示し，対話コーパスを構築した．このコーパスは，クリニック予約を対象とした約 100 対話分のデータからなる．また，構築した対話コーパスを分析し，より多様なタスク指向対話データの収集に向けた課題と知見を報告する．

1. はじめに

クリニック予約などの何らかのタスク達成を目的をとしたカスタマーとの対話では，オペレーターはカスタマーに対し提案や質問，説明を行い，場合によっては妥協するよう交渉する必要がある．こうしたオペレーターのように柔軟なタスク指向対話システムを実現するためには，多様かつ大規模な対話データセットが必要となる．実際に，既存のタスク指向型対話システムの研究では MultiWOZ [Zang 20] や Schema-Guided Dialogue Dataset [Rastogi 20] などの大規模な対話データで対話モデルの学習を行うのが一般的である．しかし，実サービスで稼働する対話モデルでは対象となる対話タスクのドメインが増え続けていくため，既存の公開対話コーパスだけではカバーできないドメインが生じてしまう．

そこで，大規模かつ効率的に対話データを収集する手段として，クラウドソーシングを用いて新規ドメインの対話データを収集することを考える．なお，クラウドソーシングではワーカーを管理することが難しく，対話インストラクションによっては単調で多様性の低い対話データが収集されてしまう問題がある．本研究では，多様な対話データの収集を目的とした対話インストラクションの設計の検討のため，クラウドソーシングを用いて実験的に小規模なクリニック予約対話データ（100 対話）を

収集する．また，構築した対話コーパスを分析し，より多様で高品質なタスク指向対話データの収集に向けた課題を報告する．

2. 対 話 設 定

本論文で扱う対話タスクはタスク指向対話において一般的である予約対話とし，ドメインはクリニック予約を選択した．クリニック予約対話の例を 表 1 に示す．

参加者は 2 人 1 組となり，クリニックの予約がしたいカスタマー役と，その受付をするオペレーター役をそれぞれ演じてもらい，構築した対話収集基盤上で日本語テキストチャットを行う．カスタマー役のインストラクションには予定が空いている，もしくは予定がある日時の情報が，オペレーター役のインストラクションにはクリニックの予約候補枠（空き枠）が必ず含まれている．カスタマー役とオペレーター役はそれぞれ対話しながら最適な予約候補枠を見つけ出し，オペレーター役が予約処理を完了したところで対話終了となる．

表 1　クリニック予約対話の例

カスタマー	こんにちは．東京にあるクリニックを予約したいのですが．
オペレーター	ご連絡ありがとうございます．東京のクリニックのご予約を希望ですね．東京のクリニックは複数ありますが，ご希望のエリアはございますか？
カスタマー	渋谷区でお願いします
オペレーター	渋谷区ですね．渋谷区では「東京渋谷院」がございます．ご希望の予約日はございますか？
カスタマー	来週月曜日の 14 時は空いていますか？
オペレーター	申し訳ございません．14 時は既に埋まっております．17 時以降であれば空き枠がありますがいかがでしょうか？
カスタマー	すいません，17 時以降は難しいです．その翌日の火曜日であればいつでも大丈夫です．
オペレーター	承知いたしました．では，来週火曜日の 14 時はいかがでしょうか？
カスタマー	はい，それでお願いします．
オペレーター	ありがとうございます．来週火曜日 14 時 東京渋谷院 でご予約進めてもよろしいでしょうか？
カスタマー	はい
オペレーター	予約が完了しました．当日はお気をつけておこしください．それでは失礼いたします．

3. 予備実験：小規模対話データ収集

クラウドソーシングを用いて対話データを収集するにあたり，予備実験として AI Shift 社内で小規模な対話データ収集を行った．実験参加者として，コールセンター業務担当者を含む日本語母語話者 38 名（19 ペア）に依頼した．参加者にはカスタマー役とオペレーター役の 2 名でペアを組んでもらい，1 回ごとに役割を交換して 2 回行ってもらった．

対話ツールは，児玉らの開発した日本語対話収集基盤 [児玉 21] を使用した．また，インストラクション表示画面と，実際の予約オペレーションに近づけるための Customer Relationship Management (CRM) ツールシミュレーターを別途実装した．CRM ツールシミュレーターは，予約枠候補が格納されているデータベースと連携しており，クリニック名検索，予約枠検索，予約処理を実行することができる．

収集した対話データの件数は合計 38 対話，559 発話で，1 対話あたり 14.3 発話，対話完了までにかかった平均時間は 14 分だった．また，予約まで完了させた対話の割合（以下，対話完了率）は 100%だった．

インストラクションに起因する問題の一つとして，希望の予約枠が取れなかった際に，オペレーターが提示した代替案をカスタマー側がそのまま受け入れてしまう事例が挙げられる．実際の予約対話においては，カスタマーとオペレーター間で妥協可能な点とそうでない点を考慮しつつ条件のすり合わせが行われることが多い．しかし，今回の実験のインストラクションでは，予約が取れなかった場合にどの予約枠を確保するのかをカスタマー役の参加者に委ねていた．このような設定においては，参加者にとって対話を継続するインセンティブが薄いため，結果として参加者はオペレーター役からの提案をそのまま受け入れてしまうことが多かったと考えられる．

また，インストラクション画面に表示されている予約条件をそのままコピーしている事例も多く見られた．インストラクション内容をコピーして相手に提示した場合，ほとんど対話を行わなくとも最も適した予約枠を見つけ出すことが可能になるため，単調な対話になってしまう恐れがある．

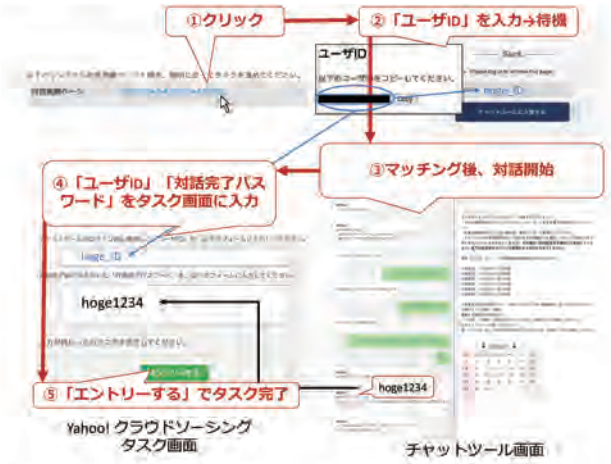


図 1　クラウドソーシングを用いた対話データ収集の流れ

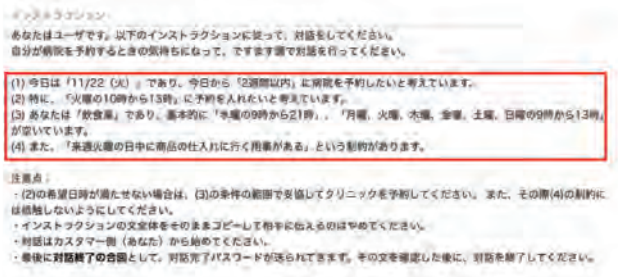


図 2　カスタマー役のワーカーに表示されるインストラクション画面の例

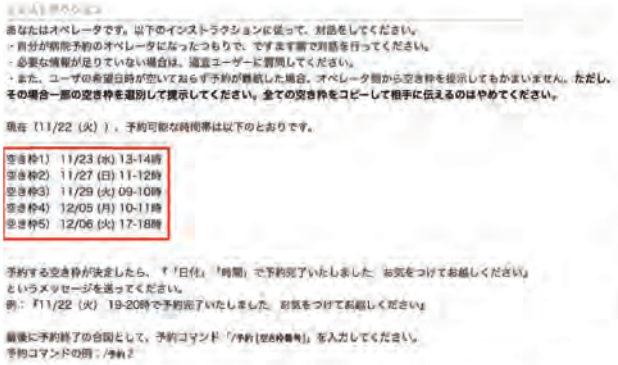


図 3　オペレーター役のワーカーに表示されるインストラクション画面の例

4.　クラウドソーシングを用いた対話データ収集

予備実験における対話収集時の課題を修正した後，クラウドソーシングを用いて対話データ収集を行った．対話データ収集の流れの概要を 図 1 に示す．今回は日本語対話データを収集することを目的としているため，ワーカーのほとんどが日本人と考えられる Yahoo!クラウドソーシング *1 を通して収集を行った．ワーカーは Yahoo!クラウドソーシングのタスク実施ページに記載した URL から，対話収集実験ページに移動する（1）．対話収集実験ページには，タスク説明とチャットルームのリンク，ユーザ ID が掲載されている．ユーザ ID はチャットルームに入る際の認証に使用されるもので，対話収集実験ページにアクセスすると自動で生成される．ワーカーがチャットルームのリンクをクリックするとログイン画面が表示され，ユーザ ID を入力すると待機室に入る（2）．待機室にいるワーカーが 2 名になるとチャットルームに遷移する（3）．チャットルーム画面を 図 4 に示す．チャットルームに遷移した 2 名に対して，オペレーター役かカスタマー役の役割が自動で割り振られ，役割に応じたインストラクションが表示される．カスタマー役とオペレーター役のインストラクションの例を，図 2，図 3 に示す．各ワーカーは表示されたインストラクションに従って対話を進める．予約したい空き枠が確定した場合，チャット画面から予約コマンド/**予約 [空き枠番号]** をメッセージとして送信することで予約が確定し，対話完了パスワードが発行される．最後に，Yahoo!クラウドソーシングのタスク画面に戻り，ユーザ ID と対話完了パスワードを入力することで対話収集が完了する（4，5）．

対話収集タスクは 2023 年 1 月 26 日の 20 時に開始し，約 1 時間半で 100 対話収集した．ワーカーは 1 名あたり 5 回までタスク参加可能とし，最終的に参加したユニークワーカー数は 145 人だった．1 対話完了時にワーカーに支払う報酬は 90 円に設定した．報酬は，1 対話あたりにかかる時間を 5 分程度と想定し，タスク開始時点の東京都の最低賃金（時給 1,072 円）を上回るように決定した．

対話収集ツールは **Slurk** [Götze 22] を用いた．Slurk は同一画面内にチャット画面とインストラクション画面の両方を表示できる．ワーカーごとに異なるインストラクションを表示することもできるため，ワーカーは複数の画面を行き来する必要がなく，対話時の負担を軽減できる．また，メッセージの連続送信が可能となっており，実際のオペレーター業務により近い環境でテキストチャットを行うことができる．予備実験で実装した CRM ツールシミュレーターは，ワーカーの負担軽減のため使用せず，インストラクションとテキストチャットでクリニック予約が完結する設定に変更した．インストラクションの詳細は次節で説明する．

クラウドソーシング上で予約対話タスクは 175 回実施され，そのうち正常に予約まで完了した対話は 96 対話，対話完了率は 54.9%だった．予約まで完了した対話の総発話数は 771 件で，1 対話あたりの平均発話数は 8 件となった．また，予約まで完了した対話の平均対話時間は 9 分だった．

4·1　インストラクションの改善

オペレーターが提示した代替案をカスタマー側がそのまま受け入れてしまう問題に対処するために，「どの程度妥協するのか」をカスタマー側の予約条件で明示するようにした（ 図 2 参照）．カスタマー役に提示されている 3 つの予約条件と，オペレーター役に提示されている 5 つの予約枠候補は，データベースに登録してある予約枠の日時情報に基づいてルールベースで自動生成しており，全く条件を満たさない予約枠は出さないようにしている．これにより，カスタマー役がオペレーター側の提案をそのまま受理するケースは低減できると考えられる．また，各予約枠候補はカスタマーの各予約条件を満たしているかどうかをフラグで管理しており，適切な予約ができている質の高い対話を識別することが可能となっている．これらの改善に加え，インストラクション画面のテキストをコピーできないよう機能を修正した．

*1　https://crowdsourcing.yahoo.co.jp/

図 4　チャットルーム画面のスクリーンショット

また，カスタマー役の予約条件（3）の中に「飲食業」「サラリーマン」「主婦」のようなペルソナ情報を設定した．ペルソナ情報を付与することで時間的制約の背景を想像しやすくなり，ワーカーの意欲向上に繋がると考えられる．

4.2　クラウドソーシングにおける課題

対話を行っているワーカーのペアのうち，片方のワーカーが何らかの理由で途中離脱したために，対話を進行できなくなった事例を複数確認した．この場合離脱されたワーカーは報酬が受け取れなくなってしまうため，別途補償タスクを設置するなどの対応を行う必要がある．他の課題として，ロールプレイをしつつ複数の予約条件を考慮して対話を進めることが難しいとのアンケート回答が複数得られた．特にオペレーター役に関しては，オペレーターとしての教育を受けていないワーカーにとって負担は大きかったと考えられる．また，そうしたワーカーによって行われる対話がオペレーターとして適切なものであるかは保証されていないため，対話モデルの学習データに使用することは難しい可能性がある．

これらの問題を解決するために，オペレーター役を社内に在籍しているコールセンター業務担当者に依頼し，カスタマー役のみクラウドワーカーに依頼する形式に変更することが考えられる．類似する収集形式として，林部の要約付き宿検索対話コーパスの設定がある [林部 21]．林部は，オペレーター役 2 名のうち 1 名を環境業界での接客経験者に依頼し，カスタマー役の参加者と対話を行う設定で収集している．これにより，オペレーター役が対話途中で離脱するリスクも抑えられるため，対話完了率の向上し，クラウドワーカーの負担も軽減されることが期待できる．

5.　お　わ　り　に

本研究では，クラウドソーシングを用いたタスク指向型対話データの収集を目的として，対話収集基盤の構築と，多様な対話を促すためにクラウドワーカーに提示するインストラクションの設計を行った．今後は，コールセンター業務担当者にオペレーター役を依頼し，より質の高い対話データを収集する方法について検討していく．

◇ 参　考　文　献 ◇

[Götze 22]　Götze, J., Paetzel-Prüsmann, M., Liermann, W., Diekmann, T., and Schlangen, D.: The slurk Interaction Server Framework: Better Data for Better Dialog Models, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4069–4078 (2022)

[Rastogi 20]　Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P.: Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 8689–8696 (2020)

[Zang 20]　Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., and Chen, J.: MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines, in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 109–117 (2020)

[児玉 21]　児玉 貴志, Bergeron, F., 新 隼人, 田中 リベカ, 坂田 亘, 黒橋 禎夫：クラウドソーシングで利用可能な日本語対話収集基盤, 言語処理学会第 27 回年次大会, pp. 859–863 (2021)

[林部 21]　林部 祐太：要約付き宿検索対話コーパス, 言語処理学会第 27 回年次大会, pp. 340–344 (2021)

著　者　紹　介

邊土名　朝飛

2021 年長岡技術科学大学大学院工学研究科修士課程修了後、サイバーエージェント入社。同子会社の株式会社 AI Shift、および AI Lab にて音声対話システムの研究開発に従事。

画像生成モデルの人手評価設計

大谷 まゆ
Mayu Otani
AI Lab
Research Scientist
otani_mayu@cyberagent.co.jp

富樫 陸
Riku Togashi
AI Lab
Research Scientist
togashi_riku@cyberagent.co.jp

澤井 悠
Yu Sawai
AI 事業本部 極予測 LP
Software Engineer
sawai_yu@cyberagent.co.jp

石上 亮介
Ryosuke Ishigami
AI 事業本部 極予測 LP
ML Engineer
ishigami_ryosuke@cyberagent.co.jp

keywords: 画像生成、評価、クラウドソーシング

Summary

画像生成モデルについて多くの研究成果が発表されているが，その評価の信頼性には未だ多くの課題が残されている．画像生成モデルの評価方法の選択肢として自動評価と人手評価がある．しかし人手評価の方法はこの分野において方法論が整備されていない．そのため，それぞれの研究グループが独自の人手評価を実施し，かつその評価方法に対する検証も不十分という状況である．このような状況の改善を目的に，本研究では人手評価の標準的な方法を検証した．また収集した人手評価の結果を主要な自動評価指標と比較することで自動評価指標の課題を明らかにした．

1. はじめに

画像生成モデル技術が広く実用されるに至り，多くの開発グループで活用が進められているが，乱立する画像生成モデルの中から良いモデルを見つけることは厄介な技術的課題である．この課題に対処することは今後のプロダクト開発の底上げに繋がると考え，画像生成モデルの評価について研究したのでその成果 [Mayu 23] の概要を本稿にまとめる．

2. 画像生成評価の現状

現在，多数のグループが競って画像生成モデルの研究を進めている．しかし画像生成モデルを評価する方法となると意外と未成熟な実態がある．生成画像の品質を表す指標として主要なものに Fréchet inception distance（FID）[Heusel 17] があり，ほとんどの研究ではモデルの有効性を示すためにこの FID の改善をアピールする．しかし FID の限界は広く知られており，他に良い客観指標がないので渋々使っているというのが実情に近いと思われる．

自動評価では十分な評価ができないため，人手評価を併用することが多い．しかし残念ながら，画像生成の領域において人手評価の方法は洗練されていない．近年出版された画像生成モデルの論文を調査した結果，37 本のうち 20 本が人手評価の結果を掲載していた（ここで 17 本は自動評価のみに依存しているということになる）．それぞれの評価内容ではそれぞれ異なるやり方が採用されており，共有された人手評価方法が不在であることが見えてきた．まず評価項目として全体的な品質と，入力テキストとの関連の強さを問う項目は多くの研究で採用されていた．回答形式の差異は大きく，複数の候補から最も良いサンプルを選ぶ形式は 10 件，品質の順番を回答させるものが 9 件，3 あるいは 5 点満点での絶対評価をつけさせるものが 3 件であった．また収集されたラベルデータの品質を定量的に報告しているものはなかった．つまり，それぞれの評価方法の妥当性は検証されていない．さらに評価方法の実装を公開する慣習がないため，後続の研究は評価プロトコルを再利用できない課題もある．また多くの詳細が未報告であることも浮き彫りとなった．例えば，多くの研究ではクラウドソーシングを使って評価者を募集するが，その際の評価者の実績や主要な使用言語などデータの品質に大きく影響する条件はほとんどの場合不明である．また支払われた報酬額の報告も欠けているが，これは近年の慣習に照らし合わせて不適切と判

A 案	B 案
Q. Rate the quality of the image.	Q. Does the image look like an AI-generated photo or a real photo?
（1） Very poor	（1） AI-generated photo
（2） Poor	（2） Probably an AI-generated photo, but photorealistic
（3） Acceptable	（3） Neutral
（4） Good	（4） Probably a real photo, but with irregular textures and shapes
（5） Very good	（5） Real photo

図 1　ワーカーへの質問文と選択肢の文例．意味のあるデータを収集するためには文言の調整が重要である．

断される可能性が高い．

このように，現状の画像生成における人手評価では評価方法自体の妥当性が議論されていないため，標準的な評価方法として扱えるものが不在である．また評価実験の詳細の公開が徹底されていないため研究プロセスにおける透明性の問題がある．

3. 今回の提案

そこで今回の研究では標準的な人手評価方法の整備を目標にした．具体的にはクラウドソーシング上でいくつかの一般的な評価タスクを設計し，実際にデータを収集して品質の高い評価を実現しやすい設計を探索した．適当な設計では，注意力のない評価者が多数参加したり，実験の意図が評価者に伝わらないため評価者ごとに大きなばらつきが生じることがある．そのような設計上の不備を検出するため，収集されたデータの品質評価の指標として，複数の評価者による意見の一致しやすさを計測した．具体的には Krippendorff の α[Klaus 80]*1を用いた．

まず基本の設計として，クラウドソーシングプラットフォームの Amazon Mechanical Turk（AMT）を利用することとした．クラウドソーシングサービスは異なる研究グループで設定を揃えることが容易であり，実験に係る費用も比較的安価に抑えられる *2．回答方式は 5 段階の絶対評価とし，評価項目は最も一般的な画像自体の品質と入力テキストとの関連性とした．また各サンプルにつき 3 人の評価者を割り当てた．

質問と選択肢の文言質問項目と選択肢の書き方は評価者の行動に大きく影響するが，あまり注意深く設計されていないことも多いように思う．特に 5 段階評価のような回答形式の場合，1 点と 5 点にだけラベルがふってあり，その中間の選択肢に関しては曖昧にしているようなフォームが散見される．このようなデザインでは意図しない評価のばらつきに繋がる．今回は入力テキストと画像の 200 ペアの評価タスクで 2 パターンの文言を比較した．図図 1（A）は少々曖昧な表現となっているが一般的な文言である．（B）はより具体的になるように質問文，選択肢を修正している．結果，文言を (B) にすることで評価者間の一致度は 0.18 から 0.39 へ大幅に改善した．

評価者の応募資格 AMT では作業を依頼する評価者に条件を設定することができる．今回は以下の 5 条件が評価の品質にどのように影響するかを調査した．

- Maturity: 18 歳以上であり不快なコンテンツが表示される可能性に同意している
- Experience: 5,000HIT を完了した実績があり，提出の承認率が 99%以上である
- Location: 英語を主要な言語とする国に在住している
- Skillfulness: 事前に指定されたテストをパスしている
- Master: AMT が独自に提供している優秀なワーカーへ付与される資格

この条件の組み合わせの影響をまとめたものが表 1 である．条件 i と ii のみでは，評価の不一致が多い．このグループは 1 サンプルあたりの作業時間が 12 秒程度と短くなっており，評価時における不注意が評価の不一致に表れていると考えられる．条件が厳しくなるにつれ，作業にかける時間が増え，評価の一致度は上がる傾向が見られた．最終的に AMT Master の利用が最も評価の一致度が高くなることが期待されたため，この後の実験ではこの設定を用いることとした．しかし AMT Master は発行基準が不明瞭であり，Master 資格の発行が停止されている疑いがあることには注意が必要である．

4. 実　　験

今回開発した評価タスクを用いて 4 つの画像生成モデル [Rombach 22, Zhou 22, Ding 22, Nichol 22] を評価した．入力テキストとして COCO Caption[Lin 14]，DrawBench[Saharia 22]，PartiPrompts[Yu 22] を使用した *3．各サンプルには 3 人の評価者が割り当てられ，1 サンプルごとに$0.05 を支払った．

人手評価と自動評価による結果を表 2 に示す．ここで注目すべきは人と自動評価指標によるモデルの順位の差異である．画像の品質を評価する指標として FID は広く使われているが，FID が好む CogView2 は人による評価では 4 つのモデルのうち 3 位となっている．またテキストとの意味的な類似度の評価に使われている CLIPScore[Hessel 21] の課題も明らかになった．CLIPScore では生成され

*1　一致しているほど高い数値をとり最大は 1 である。
*2　案件によっては社内のアノテーションセンターがより有効である場合が多いと思われるが，今回は英語話者が必須条件であることと，国際的学術コミュニティへ向けた取り組みという側面があったため AMT を選択した．
*3　誌面の都合上 DrawBench と PartiPrompts の結果は省略する．

表 1　評価者に対する 4 つの条件の組み合わせが評価データの品質に及ぼす影響．画像の品質（Fidelity）とテキストとの関連性（Alignment）の値は 3 人の評価者による平均評価値をサンプル全体で平均したものである．評価者間の一致度（IAA）として Krippendorff の α を計算した．Med. time は 1 回の評価にかかった推定時間の中央値である．

Qualification					Annotator performance			Stable Diffusion		Real image	
i	ii	iii	iv	v	Fidelity IAA	Alignment IAA	Med. Time	Fidelity	Alignment	Fidelity	Alignment
✓	✓				0.11	0.10	12.0	3.81	4.63	4.78	4.94
✓	✓	✓			0.39	0.26	16.0	2.83	4.18	4.43	4.76
✓	✓	✓	✓		0.37	0.40	20.0	2.71	4.23	4.32	4.67
✓				✓	0.53	0.44	25.0	2.65	4.18	4.58	4.81

表 2　COCO Captions における自動評価指標と人手評価の差異

model	Human		Automatic	
	Fidelity ↑	Alignment ↑	FID ↓	CLIPScore ↑
LAFITE	1.77	3.73	34.46	0.82
GLIDE	2.56	2.96	39.80	0.68
CogView2	2.19	3.55	29.57	0.68
Stable Diffusion	3.09	4.35	32.19	0.78
Real Image	4.49	4.78	—	0.76

図 2　画像と入力テキストに対する CLIPScore の例

た画像のほうが，COCO Caption 内の実画像よりもテキストとの関連性が高いと評価されていた．しかし実際の画像と CLIPScore を検証したところ（図 2），CLIPScore は直感に反する評価値をつけていることが確認された．

4·1　評価に必要なサンプルサイズの分析

評価データのサイズは人手評価を設計するにあたり重要な要因である．図 3 は評価するサンプルサイズ（入力テキスト数）によって実験の結果に生じる影響を可視化している．収集した 1,000 件の評価データからランダムに n 個を抽出し，平均評価値を算出することを 500 回繰り返した．色付き領域は 5-95%信頼区間である．色付き領域の重なりから，限られたサンプルサイズではモデルの順位が偶然変わることがわかる．また評価者の数に関しても，評価者が少ない場合は結果の信頼性に注意が必要である．人手評価でサンプルサイズを上げることは困難な場合も多いが、そのような場合は統計的有意差や効果量のチェックなどを併用することが推奨される．

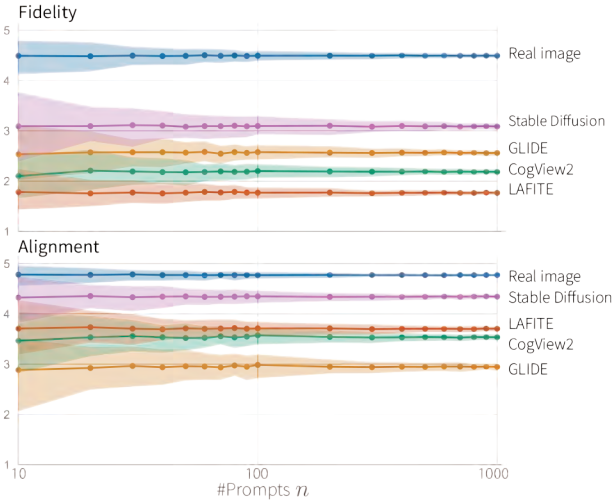


図 3　評価サンプルサイズの影響．サンプルサイズが小さいとモデルの順位が偶然反転することがある．

5. これから

まず，本稿の例からわかるように，多くの論文で使われているからといって信用できる評価方法とは限らないということを伝えておきたい．他のプロダクトにおいても，評価に潜在する課題の分析は開発の方向性を健全に保つために重要な基礎なので，今後も評価の改善を広げていきたい．

6. 補遺: プロダクトから見た画像生成 AI 評価の意義

本研究では，極予測 LP というプロダクト所属のエンジニア 2 名も共著者として参加した．プロダクト側でも画像生成 AI が広告クリエイティブ制作に大きな影響を与えるのは必至だと考えていたことや，各種の既存手法を再現することが可能なエンジニアがいるという点が今回の協力体制につながった．

極予測 LP のチーム内で広告クリエイティブ制作に対

して画像生成 AI を利用する試みは 2022 年の DALL-E や Midjourney 公開直後から着手していた．GLIDE ベースのカスケードモデルである *hoksai-mini* のように広告クリエイティブ分野の画像で独自に学習を行ったモデルや，Stable Diffusion をベースとしてファインチューニングしたモデルを構築して社内向けに公開した．

複数の画像生成手法やその学習済み重みのうち，どれをどの用途で選択すれば良いかというのは，広告ドメインは文献で評価されている対象とは異なる点で難しい．そして画像単体に関する自動品質評価尺度のみを用いるのは，現実的な制約の下ではあまり有用とはいえない．広告クリエイティブ制作においては制作対象の商品やサービス（商材）と広告訴求，広告主側が設定するレギュレーションや提供素材といった様々な制約があるため，少なくともテキストやその他のモダリティによる制約を扱える評価尺度が望ましい．一方，広告制作フローに介入する AI という性質上，人間の制作者と協調して制作に介入していくことになるので，人間によるフィードバックは得やすいというユースケース上の特徴がある．

これらの特徴から，オンラインの Human-in-the-Loop 設定で扱うにしろそうでないにしろ，広告制作者や広告を見る側のユーザーからの評価を模擬した人手評価の手法を確立することが必要であると考える．妥当な人手評価のプロトコルが確立することで，より人間と協調しやすい画像生成手法の実用化に近づくのではないかと期待している．信頼できる人手評価の方法が確立されれば，広告クリエイティブ制作のような複数の制約下での画像その他の生成手法に対するより良い自動評価尺度にもつながる可能性がある．

現状では広告ドメインのために考慮すべき種々の制約を評価の設問に含めることはできていない．今後は商材とのマッチングや種々の制約をテキストプロンプトから制御できているかを評価できるような，広告クリエイティブ制作における人手評価の設問を考えたい．よりプロダクト寄りの方向性としては，画像生成手法を制作フローに導入してみて，広告制作者からのフィードバックを得たりオンライン配信の効果検証を行うことも検討している．

◇ 参 考 文 献 ◇

[Ding 22] Ding, M., Zheng, W., Hong, W., and Tang, J.: CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers (2022), arXiv:2204.14217 [cs]

[Hessel 21] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning, in *EMNLP*, pp. 7514–7528 (2021)

[Heusel 17] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, in *NeurIPS*, p. 6629–6640 (2017)

[Klaus 80] Klaus, K.: Content analysis: An introduction to its methodology (1980)

[Lin 14] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, in *ECCV*, pp. 740–755 (2014)

[Mayu 23] Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkilä, J., Shin'ichi Satoh：Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation, in *CVPR* (2023)

[Nichol 22] Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M.: GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, in *ICML*, Vol. 162, pp. 16784–16804 (2022)

[Rombach 22] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models, in *CVPR*, pp. 10684–10695 (2022)

[Saharia 22] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, *arXiv preprint arXiv:2205.11487* (2022)

[Yu 22] Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., and Wu, Y.: Scaling Autoregressive Models for Content-Rich Text-to-Image Generation (2022), arXiv:2206.10789 [cs]

[Zhou 22] Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., and Sun, T.: LAFITE: Towards Language-Free Training for Text-to-Image Generation, in *CVPR* (2022)

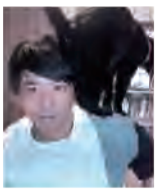
著 者 紹 介

大谷 まゆ

2018 年に奈良先端科学技術大学院大学情報科学研究科博士後期課程修了後、サイバーエージェント入社。コンピュータビジョン、機械学習に関する研究に従事。

富樫 陸(2)

2020 年サイバーエージェント新卒入社．専門は推薦システム．情報検索。推薦手法の分散最適化や検索タスク評価についての研究と応用に取り組んでいる．



澤井 悠(3)

AI 事業本部 極予測 LP 所属のエンジニア．プロダクトのデータサイエンス面を構想から実装，MLOps まで広く担当している．京都在住．

石上 亮介(4)

AI 事業本部で「極予測 LP」の開発，大規模言語モデル（LLM）をはじめとした基盤モデルプロジェクトのリードを担当．画像やテキストを対象としたマルチモーダルな AI の社会実装に従事している．

2023 Vol.2

何点加点する？

郡山市の保育所利用調整基準を見直す　シミュレーション編

竹浪 良寛
Yoshihiro Takenami
株式会社サイバーエージェント AI Lab
Data Scientist
takenami_yoshihiro@cyberagent.co.jp

森脇 大輔
Daisuke Moriwaki
株式会社サイバーエージェント AI Lab
Research Scientist
moriwaki_daisuke@cyberagent.co.jp

Wu Shuting
株式会社サイバーエージェント AI Lab
Data Scientist
wu_shuting@cyberagent.co.jp

松木 一永
Kazunaga Matsuki
株式会社サイバーエージェント AI Lab
Research Scientist
matsuki_kazunaga@cyberagent.co.jp

keywords: マーケットデザイン、マッチング理論、保育園利用調整

Summary

福島県郡山市では保育所利用調整基準の見直しを行った。特に、兄弟姉妹が同時に申し込む場合、単独で申し込む場合よりも入所率が低いことを問題視していた。マッチングアルゴリズムを用いたシミュレーションにより、兄弟姉妹同時申込に対して大幅な加点を行うことで、単独申込との入所率の差を縮小できることを示した。この結果を基に、郡山市は 2024 年度から新たな利用調整基準を実施した。今後はこの制度変更の効果検証を実施し、必要に応じてさらなる改善を行う予定である。

1. 保育所利用調整の「利用調整基準」

1·1 利 用 調 整

§1 保育所を利用するには

ある世帯には子どもがおり、保護者が就労するため、日中は誰かに子どもを見てもらう必要があるとする。このとき、この世帯では保育所を利用することを考えるだろう。

保育所を利用するには、まず、その世帯が住んでいる自治体（市町村および特別区）に申し込む必要がある。申し込む際には、世帯の構成員や、それぞれの就労や就学等の状況を報告する必要がある（保護者等の就労状況等）。就労証明書を勤務先から取り寄せ、それを添付する必要もある。兄弟姉妹が同様に保育所を利用している場合も報告する必要がある（児童の世帯状況）。子どもがすでに保育所やそれに準ずる施設を利用していればそれも報告する必要がある（児童の保育状況）。もちろん、希望する保育所のリストも報告する。

申込書に世帯状況と希望を記入した後、自治体は利用調整基準に基づいて各応募者を点数化し、点数が高い児童から優先順位を設定する。そして、多くの自治体では、優先順位が高い児童から、その児童が希望する保育所の中で空きがあるところに割り当てていく。

図 1　利用調整のフローと点数の決め方

割り当てが成功すれば、その保育所に内定となるが、割り当てができなかった場合は、入所保留（あるいは待機児童）という扱いになり、保育所に通うことはできない。この一連の流れは図 1 に示している。

§2 保育所の利用調整

保育所の利用調整は、保育所の供給（募集人数）と需要（申し込み人数）について、供給を需要が上回るおそれが

ある場合に利用調整を実施することとなっている*1。国の資料によれば、待機児童数は2023年4月時点で2,680人となっているほか、国の定義上待機児童にカウントされない「特定の保育園等のみ希望している者」が37,781人もおり[こど23]、利用調整が必要な状況となっている。

1·2 利用調整基準の見直し

§1 利用調整基準

利用調整基準とは、保育の必要性を点数化し、その点数をもって優先順位づけを行うものである。児童福祉法施行規則の第二十四条では「保育の必要の程度及び家族等の状況を勘案し、保育を受ける必要性が高いと認められる児童が優先的に利用できるよう、調整するものとする。」とある。これは、申込者の状況を踏まえ、保育の必要性が高い順に保育所を割り当てる必要があることを示している。

保育の必要性は、保護者の就労状況や児童・兄弟姉妹の状況等をふまえ点数化しており、より保育の必要性が高いと判断すれば高得点をつけている。例えば福島県郡山市では、ひと月140時間の就労で200点をつけるが、就労時間が短ければ点数は減少し、求職中であれば50点をつける、といった点数化を行っている[福島21]。そして、この点数が高い順に優先順位が設定され、保育所が割り当てられている。つまり、点数が高い人ほど保育所の入所について優先されている。

§2 利用調整基準見直しの影響

利用調整基準を見直した場合、割り当てがどのように変化するかは明らかではない。例えば、郡山市において兄弟姉妹在所の場合にはこれまでの25点加点していたが、これを100点の加点に変更したとする。このとき、兄弟姉妹在所に該当する児童の入所率は増加するだろうが、どの程度増加するかは明らかではない。副作用として、兄弟姉妹在所に該当しない児童の入所率が減少するだろうが、どの程度減少するかは明らかではない。

また、加点している属性の多さにも配慮する必要がある。郡山市では「保護者等の就労状況等」という項目で、就労状況の他に妊娠・出産、疾病等、障がい等、介護・看護、不在（父母の離婚等）等の点数が異なる属性があり、また「児童の世帯状況」という項目で満2歳までを利用年齢とする保育施設等を満了する入所児童、児童の再入所、生活保護世帯又は市民税非課税世帯、兄弟姉妹在所中、3人以上の多子世帯などで加点を行っている。そのため、ある属性を加点して優遇する際に、他の属性の人たちがどのような影響を受けるのか見通すことは簡単ではない。

1·3 利用調整基準とマーケットデザイン

保育所の利用調整は学術的にはマーケットデザイン、特にマッチング理論において取り扱われてきた[Okumura 19][Kamada 23][Sun 23]。保育所の利用調整のほか、問題設定が似ている学校選択制や研修医マッチングを取り扱った研究では、利用調整基準のような優先順位がつけられていることを所与として、マッチングアルゴリズムを変えることで割り当ての安定性や合理性、耐戦略性といったマッチング理論で重要な概念を満たすかどうか研究が行われてきた。

近年は優先順位の見直しとマッチング理論を融合した論文が出ている。[Shi 22]では、学校選択制において、実際に使われていたDAアルゴリズムというマッチングアルゴリズムを引き続き使用することを前提に、各児童の通学距離を最小化することを目的にし、優先順位を見直す手法を提案している。

このように、保育所利用調整基準の見直しにおいても、各自治体が保育所サービスを提供する目的や、福祉に対する考え方を達成できるよう、これらに適応した基準を作成することが重要である。利用調整のアルゴリズムだけではなく、利用調整基準をそのように見直すことで、自治体が望む利用調整が初めて実現できるからである。

2. 郡山市の状況・課題

2·1 郡山市の現状

2023年度までの郡山市の利用調整基準は、いくつかの特徴を持っている。その一つは、兄弟姉妹が同時に申し込む場合、点数の加算がないことである。一方、すでに兄弟姉妹が保育所に入所している場合には、25点が加算されている。

この利用調整基準のもとで代表的なケースについて2つ紹介する。まず、両親がフルタイムで就労し、育児休業から復帰する場合である。フルタイム就労の場合は200点、育児休業から復帰する場合は30点が加算されるため、このような場合には430点がつけられる。つぎに、片方の親がフルタイムで就労し、もう一方の親が求職中である場合である。求職中である場合は50点、自宅で保育している場合には1点が加算されるため、このような場合には251点がつけられる。もし、これらのケースに該当する児童について、すでに兄弟姉妹が保育所に入所している場合には、さらに25点が加算される。

2·2 兄弟姉妹の同所入所

兄弟姉妹を保育所に入れる際、制限される選択肢とそもそもの優先順位の低さが問題となり、単独申込よりも

*1 児童福祉法第二十四条第三項には次のように記載されており、保育所への供給を需要が上回る場合に利用調整を実施することになっていることがわかる。「市町村は、保育の需要に応ずるに足りる保育所、認定こども園（子ども・子育て支援法第二十七条第一項の確認を受けたものに限る。以下この項及び第四十六条の二第二項において同じ。）又は家庭的保育事業等が不足し、又は不足するおそれがある場合その他必要と認められる場合には、保育所、認定こども園（保育所であるものを含む。）又は家庭的保育事業等の利用について調整を行うとともに、認定こども園の設置者又は家庭的保育事業等を行う者に対し、前項に規定する児童の利用の要請を行うものとする。」

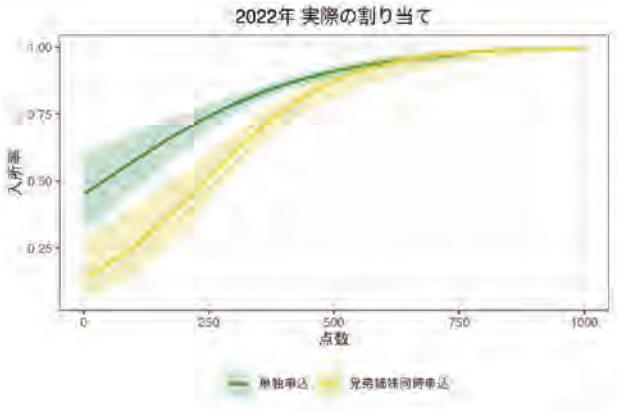


図 2　2022 年の実際の割り当てにおける点数と入所率の関係

入所が困難になる。実際に郡山市のデータを用いて、一般化線形混合モデル（GLMM）を用いたロジスティック回帰の結果を示す。ここでは単独申込と兄弟姉妹同時申込とにデータを分け、それぞれのグループにおける点数と入所率の関係を表している。図 2 の左側、点数が低いグループに注目すると、黄の線が緑の線を大きく下回っている。これは点数が低いほど兄弟姉妹での申し込みの入所率が単独での申し込みに比べて低くなり、その差が大きくなることを示している。つまり、同じ点数の子供でも、兄弟姉妹で申込む場合、単独で申し込む場合に比べて保育所への入所可能性が低くなっている。

兄弟姉妹を保育所に入れる場合は 2 通り考えられる。まず、上の子がすでに保育所に入所しており、下の子がこれから入所するケースである。もうひとつは、兄弟姉妹が同時に保育所に申し込むケースである。両ケースについて入所に関する困難が存在する。

まず、保育所の選択肢が限られ、入所確率が下がることである。異なる保育所を選ぶと複数の保育所への送迎、各園の異なる準備要件、就労時間の制約、重複する学校行事への対応などの負担が発生する。そのため、同じ保育所または近くの保育所を選ぼうとすることから、選択肢が制限される。また、兄弟姉妹を同時に入所させる場合は、それぞれの年齢に対応した空きがある保育所を選ぶ必要がある。例えば 1 歳と 3 歳で申し込む場合、同じ保育所で 1 歳と 3 歳の空き枠が存在している保育所を選ぶ必要がある。どちらからだけが空いていても同所入所はできないためだ。そして、このように選択肢が限られるため、入所確率が下がる。

次に、点数が低いため入所確率が下がることである。兄弟姉妹で同時に申し込む場合、保護者の片方が求職中であることが多い。郡山市の例でも、求職中は 50 点の加点にとどまり、フルタイム就労の 200 点に比べて点数が低い。そのため優先順位も下がり、順番が来る頃には兄弟姉妹の枠が空いていないなど、入所確率が下がる。

2·3　何点加点するか？

郡山市では、兄弟姉妹が同時に申し込んでも特別な加算がなかったこともあり、兄弟姉妹で同時に申し込む場合に加算を行うことを検討した。しかし、兄弟姉妹だけを極端に優遇するような利用調整基準とすると反発も予想されるため、公平な基準設定が必要であった。例えば、兄弟姉妹在所の場合 25 点を加算していることから、兄弟姉妹で同時に申し込む場合も 25 点加算する、というような見直しも考えられるが、そもそもなぜ 25 点を加算しているのか、そして 25 点の加算は効果があるのか、といった観点から議論がありうる。

3.　シミュレーション

最適な加点を見つけるためにシミュレーションを実施することとした。シミュレーションにより、兄弟姉妹への加点を行った場合に入所率へどのような影響があるかを検証する。

3·1　シミュレーションの目的と制約

利用調整基準の見直しに向けて、郡山市と合意したシミュレーションの目的と制約は次のとおりである。

- 【目的】単独申込でも兄弟姉妹同時申込でも（できるだけ）入所率を等しくする
- 【制約】全体の入所率が減少しないようにする
- 【制約】郡山市が望む優先順位を（できるだけ）実現する

まず、単独申込と兄弟姉妹同時申込で入所率に差があることから、どちらの申込でも入所率が均等になるような加点を検討することにした。次に、全体への影響を考え、全体の入所率が減少しないような加点を検討することとした。最後に、郡山市が望む優先順位を実現するようにした。具体的には表 1 のような順位をできれば達成したい、というものであった。

順序	属性	元の点数
1	育休明け＋同年齢同時申込	430
2	片方求職中＋同年齢同時申込	251
3	片方求職中＋きょうだい在所	276
4	育休明け＋異年齢同時申込	430
5	育休明け＋単独申込	430
6	片方求職中＋異年齢同時申込	251
7	転所希望	410

表 1　郡山市が希望する優先順位

これらを目的と制約として、シミュレーションを実施する。

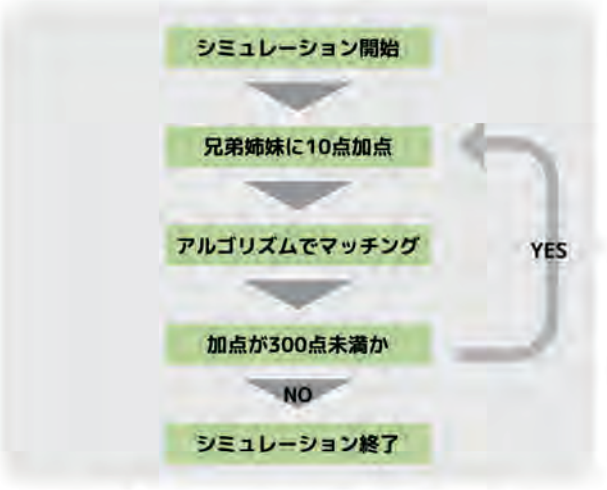


図 3　シミュレーションのフローチャート

3.2　シミュレーションの手法

シミュレーションの手法について説明する。使用したデータは、郡山市から提供された保育所への申し込み者データ（郡山市において匿名化済）と各保育所の年齢別募集人数データである（2022 年と 2023 年の 2 回分）。今回は図 3 のように比較的単純な方法を採用した。具体的には、兄弟姉妹で同時に申し込む人に 10 点加点し、[Sun 23] のアルゴリズム *2 で割り当てを行い、さらに 10 点加点して再度アルゴリズムで割り当てる、という手法である。今回、加点は 10 点から 300 点まで実施した。

この方法のメリットとしては理解しやすいということである。今回は郡山市の職員に向けて説明する必要があり、市役所内部での議論もシミュレーション結果に基づき実施してもらう必要があることから、わかりやすい方法を取った。デメリットとしては、学術的な厳密性は担保されないことである。

このようなシミュレーションを実施し、各点数で各属性の入所率がどのように変化するかを分析した。

3.3　シミュレーション結果

シミュレーションの結果について説明する。

図 4 において、X 軸はシミュレーションにおける兄弟姉妹同時申込への加点を、Y 軸は入所率を示している。まず、加点により全体の入所率は微増している。そして、加点を増やしていくと単独申込と兄弟姉妹同時申込の入所率の差が縮まる。加点なしの状況では、この差は 18% ポイントほどであったが、180 点ほど加点するとこの差はほぼなくなる。

シミュレーションの目的と制約を満たす加点のうち、郡山市は表 2 を選択することにした。目的と制約に照らし、年齢の異なる兄弟姉妹には 160 点、同年齢の兄弟姉妹には 200 点を加点する、という判断が行われた *3。

*2　このアルゴリズムは兄弟姉妹の同時入所や転園に対応し、実務で要請される性質を満たすものであることから、制度変更による影響を精度高く見積もることができる。

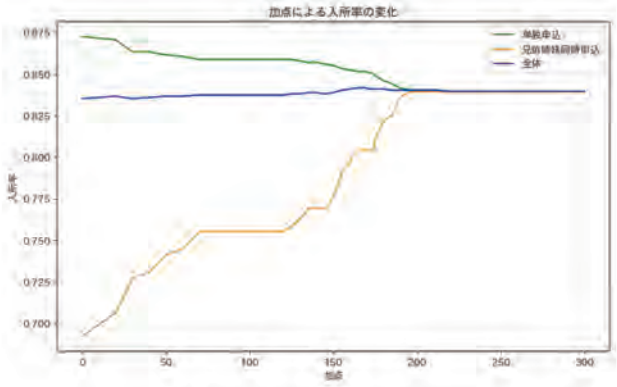


図 4　加点による入所率の変化

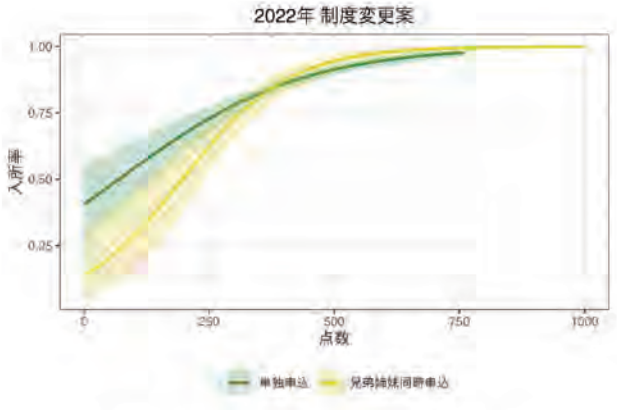


図 5　制度変更案における点数と入所率の関係

項目	点数
異年齢同時申込	160 点
兄弟姉妹在園	160 点
同年齢同時申込	200 点
多子	10 点

表 2　郡山市が選択した加点

ここでも同様に、表 2 で加点した場合の点数と入所率の関係について、図 5 の通り一般化線形混合モデル（GLMM）を用いたロジスティック回帰の結果を示す。ここでも単独申込と兄弟姉妹同時申込とに分け、点数と入所率の関係を表しており、点数は兄弟姉妹へ加点する前のものを用いている。ここでは両方の線が重なるような軌道となり、差が小さくなっている。つまり、加算後のシミュレーションにおいては、同じ点数の子供であれば、兄弟姉妹での申し込んでも単独で申し込んでも保育所への入所率は差がなくなっており、加算がない場合に比べ

*3　このように実際には異年齢兄弟姉妹と同年齢兄弟姉妹については異なる点数をつけるシミュレーションも実施していた。この理由は、郡山市の方針により、表 1 の通り異年齢兄弟姉妹よりも同年齢兄弟姉妹を優先すべきであったためである。

て公平に扱われている。よって、この加点により、全体の入所率が高まり、単独申込と兄弟姉妹同時申込の入所率の差が縮まることが確認された。

3·4　シミュレーションを通じて得た洞察

シミュレーションを実施した結果、単独申込と兄弟姉妹同時申込の入所率を近似させるためには、兄弟姉妹同時申込に対して大幅な加点が必要であるという結論に至った。現行の制度では、兄弟姉妹が在所している場合に 25 点の加点があるが、これを兄弟姉妹同時申込にも適用しても入所率の差は縮まらないことが図 4 から明らかになった。

単純に加点する点数を既存の制度に合わせてしまうと、期待した効果を達成することが難しいと予想される。このような観点からも、シミュレーションを用いて変化を予測することに価値があったと考える。

さらに、兄弟姉妹同時申込の入所率を向上させるために、単独申込の入所率をどの程度減少させることが許容できるかという議論も必要である。今回の制度変更案では、単独申込の入所率が 2-3% ポイント減少すると予測されている。このような副作用についてもシミュレーションを通じて予測し、郡山市の意思決定をサポートすることができた。

3·5　他属性への配慮

兄弟姉妹同時申込に特別な加点を設けると、他の特定の属性を持つ申込者が不利になる。郡山市では、疾病や障害、介護や看護、または一人親家庭などの状況にある申込者に対して特に保育の必要性があるとみなしていた。これらの申込者に対して兄弟姉妹同時申込に比べて相対的な不利をもたらすことは問題であると考えられた。

シミュレーションの結果、これらの属性を持つ申込者の入所率が微減することが確認された。そのため、これらの人々に対しても兄弟姉妹同時申込と同様に 160 点を追加することで、従来の入所しやすさを維持することを決定した。

4.　制度変更の実施

郡山市では 2024 年 4 月の保育所入所募集から、我々が行ったシミュレーション結果をもとに制度変更が実施された。

2023 年度までは図 6 の通り、兄弟姉妹在所と多子世帯についてのみ加点が行われた。2024 年度からは図 7 の通り、兄弟姉妹同時申請、兄弟姉妹在所、多子世帯に大きく加点されるようになった。特に、同年齢の兄弟姉妹同時申請（双子や三つ子）については 200 点が加点されている。これはシミュレーションの目的と制約における郡山市が望む優先順位（表 1）の通り、同年齢の兄弟姉妹同時申請を優先するため、この加点とした。

5.　効果検証に向けた準備

5·1　シミュレーションの限界と前提

今回のシミュレーションは、制度変更が外生的に発生するという前提の下で行われていた。つまり、制度変更があっても申込者は希望する保育所を変えずにシミュレーションを実施していた。そして、このシミュレーションの限界は申込者が制度変更に反応して行動を変化させる可能性を考慮に入れていない点にある。実際には制度変更に反応して希望する保育所を変更する可能性が存在する。例えば、制度変更によって有利になる兄弟姉妹同時申込者は希望する保育所を減らしたり、人気のある保育所を希望したりすることが考えらる。逆に、不利になる単独申込者は希望する保育所を増やしたり、人気のある保育所を敬遠したりする、といった反応が考えられる。また、新たに保育所入所を希望する人や逆に諦める人の存在もこのシミュレーションでは考えられていない。

5·2　EBPM の実践と効果検証の必要性

エビデンスに基づく政策形成（EBPM）の観点から、制度変更前後でいかなる影響があったのか検証することは重要である。そこで、今後は制度変更前後のデータを用いて、制度変更の効果検証を実施する予定である。検証項目としては、以下のような項目を検討している。

(1)　兄弟姉妹が保育所に入所しやすくなったか？
(2)　全体の入所率が減少していないか？
(3)　単独申込者の入所率が著しく減少していないか？

EBPM の一環として、これらの効果検証結果を踏まえ、次年度以降の制度変更を検討する。その実現に向け、高品質なエビデンスの収集が不可欠である。具体的な手段として、保育所への申し込み者を対象としたアンケート調査の実施や、他の自治体へのデータ提供依頼を実施している。以下では、これらの取り組みについて詳細に述べていく。

5·3　申込者に対するアンケートの実施

毎年保育所の申込を行う人は少なく、制度変更が各申込者の行動にどのように影響を及ぼしたかをデータから分析することは難しい。この問題を解決するために、郡山市と協力し、2024 年 4 月の入所申込者を対象とした Web アンケートを実施している。この調査は 2023 年 11 月から 12 月にかけて実施しており、現在（2023 年 12 月 26 日）も続行中である。

このアンケートの目的は、制度変更による行動の変化を具体的に把握することである。具体的には、申込者が制度変更の事実を申込時に認識していたかどうかを確認し、制度変更を知っていた人々に対しては、その行動に変化が生じたかどうか、例えば、保育所の希望数が変わったかどうか、人気の保育所を選択する傾向が出てきたかどうかなどを詳細に調査している。この調査結果は申込

兄弟姉妹在所中	25
3人以上の多子世帯	3

図 6　2023 年度までの利用調整基準（抜粋）[福島 21]

次のいずれかに該当する場合は当該点数とする	
(1) 兄弟姉妹同時申請（同年齢）	200
(2) 兄弟姉妹同時申請（異年齢）	160
(3) 兄弟姉妹在所中	160
(4) 別表第2に規定する保育の必要性の事由が疾病・障がい及び介護・看護、並びに父若しくは母又は両親が不在である場合（利用申請児童が1人の世帯に限る）	160
世帯内に18歳未満の子が複数いる場合（2人目から1人あたり）	10

図 7　2024 年度の利用調整基準（抜粋）[福島 23]

データと突合できるため、実際の申込行動や属性を組み合わせた分析を行う予定である。

5·4　他の自治体へのデータ提供依頼

効果検証において、シミュレーションによる制度変更の比較だけでは不十分である。なぜなら、制度変更前後のデータを用いた前後比較を実施しても、前述した行動変化の可能性を考えると、制度変更の効果を正確に測定することは難しいからである。制度変更前の行動と制度変更後の行動は前述したように異なる可能性があり、単に制度変更前後のデータを比較しただけでは適切な反実仮想を作り出すことができない。これは、制度変更による行動の変化がデータに反映されていない可能性があるためである。さらに、制度変更の影響は郡山市全体に及んでおり、郡山市のデータ内に適切な比較対象を見出すことはできていない。

そこで、制度変更の結果に対する行動変化を踏まえた効果検証が必要である。具体的には、郡山市と人口規模が近く、同じ県内に存在するため文化的な規範も近いと思われる福島市などのデータを用いた比較を行うことが考えられる。例えば、郡山市と福島市のデータを用いた差分の差分法による比較を実施することで、制度変更の前後での変化を検証できる可能性がある。そのため、現在、福島市をはじめとした福島県内の自治体にデータの提供を依頼しているところである。データを入手した暁には適切な効果検証分析を実施し、郡山市の制度改善に向けた検討を実施する予定である。

6.　ま　と　め

本稿では、保育所の利用調整基準について、特に兄弟姉妹同時申込に対する加点を検討した制度変更について述べた。

まず、保育所の利用調整が必要な状況において、利用調整基準とは何か、どのように優先順位が決まるのか、基準を見直した後にどのような影響が出るのか見通すことは難しい、といったことを述べた。

次に、郡山市の現状と課題を取り上げ、特に兄弟姉妹同時申込に関する問題点を指摘した。兄弟姉妹の同時入所が難しい原因として、単独申込に比べ選択肢が制限されること、優先順位が低いことが挙げられる。

その上で、兄弟姉妹同時申込に対する加点を検討するためのシミュレーションを行った。単独申込でも兄弟姉妹同時申込でも（できるだけ）入所率を等しくする、全体の入所率が減少しないようにする、郡山市が望む優先順位を（できるだけ）実現する、といった目的や制約のもとシミュレーションを実行した。シミュレーションの結果、目的を達成するには兄弟姉妹同時申込に対して大幅な加点を行うことが必要であることが明らかとなった。

最後に、シミュレーションに基づき郡山市が制度変更を実施したことを紹介し、効果検証の必要性と今後の実施方法についても述べた。

◇ 参　考　文　献 ◇

[Kamada 23]　Kamada, Y. and Kojima, F.: Fair Matching under Constraints: Theory and Applications, *The Review of Economic Studies* (2023)

[Okumura 19]　Okumura, Y.: School choice with general constraints: a market design approach for the nursery school waiting list problem in Japan, *The Japanese Economic Review*, Vol. 70, No. 4, pp. 497–516 (2019)

[Shi 22]　Shi, P.: Optimal priority-based allocation mechanisms, *Management Science*, Vol. 68, No. 1, pp. 171–188 (2022)

[Sun 23]　Sun, Z., Takenami, Y., Moriwaki, D., Tomita, Y., and Yokoo, M.: Daycare Matching in Japan: Transfers and Siblings, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 12, pp. 14487–14495 (2023)

[こど 23]　こども家庭庁：令和5年4月の待機児童数調査のポイント，https://www.cfa.go.jp/assets/contents/node/basic_page/field_ref_resources/f699fe5b-bf3d-46b1-8028-c5f450718d1a/8e86768c/20230901_policies_hoiku_torimatome_r5_01.pdf (2023), 最終閲覧日：2023 年 12 月 26 日

[福島 21]　福島県郡山市：郡山市保育施設等の利用調整及び保育の必要性の認定に関する事務取扱要領（令和 4 年 4 月 1 日一部改正），https://www.city.koriyama.lg.jp/uploaded/attachment/36090.pdf (2021), 最終閲覧日：2023 年 12 月 26 日

[福島 23]　福島県郡山市：郡山市保育施設等の利用調整及び保育の必要性の認定に関する事務取扱要領（令和 6 年 4 月 1 日一部改正），https://www.city.koriyama.lg.jp/uploaded/attachment/69154.pdf (2023), 最終閲覧日：2023 年 12 月 26 日

著　者　紹　介

竹浪　良寛

AI Lab 経済学社会実装チーム Data Scientist。仙台市役所で勤務後、2021 年サイバーエージェント入社。保育所利用調整等マーケットデザインの社会実装に従事。

森脇　大輔

AI Lab 経済学社会実装チーム Research Scientist。2017 年中途入社.AirTrack データサイエンティストを経て現職. 計量経済学専攻. EBPM データベース管理人. 経済学博士(ニューヨーク州立大アルバニー校)

Wu　Shuting

AI Lab 経済学チーム Data Scientist。2022 年サイバーエージェント中途入社。小売企業のデータ活用に経済学理論を活かし、広告運用効果や販促施策効果の推定に取り組む。

松木　一永

AI Lab 経済学社会実装チーム Research Scientist。2023 年サイバーエージェント中途入社。行動経済学を活用して経営課題解決をサポートするコンサルタントを経て、AI Lab 経済学社会実装チーム。人の意思決定に関連する研究テーマに関心がある。2013 年 Ph.D. Psychology (University of Western Ontario)

LCTG Bench: 日本語LLMの制御性ベンチマークの構築

栗原 健太郎
Kentaro Kurihara
株式会社 AI Shift
ML/DS Engineer
kurihara_kentaro@cyberagent.co.jp

三田 雅人
Masato Mita
AI Lab
Research Scientist
mita_masato@cyberagent.co.jp

張 培楠
Zhang Peinan
AI Lab
Research Scientist
zhang_peinan@cyberagent.co.jp

佐々木 翔大
Shota Sasaki
AI Lab
Research Scientist
sasaki_shota@cyberagent.co.jp

石上 亮介
Ryosuke Ishigami
AI 事業本部 基盤モデル事業部
ML Engineer
ishigami_ryosuke@cyberagent.co.jp

岡崎 直観
Naoaki Okazaki
東京工業大学
okazaki@c.titech.ac.jp

keywords: 大規模言語モデル (LLM), ベンチマーク, 制御性

Summary

日本を含む世界中で大規模言語モデル（LLM）の開発や事業における活用が加速していく中で、LLM の性能評価が重要課題になりつつある。LLM の事業における活用では、記事の入稿規程や SEO 対策などを考慮することから、生成結果の内容の品質のみならず、文字数制約や単語の制約などの制御性も LLM の性能の評価対象となり得る。しかし、日本語 LLM の制御性に着目した評価の枠組みは存在しない。本研究では、LLM の事業応用において留意すべき観点の一つである制御性に焦点を当て、評価ベンチマーク LCTG Bench を構築する。

1. はじめに

OpenAI 社の ChatGPT の公開以降、世界中で大規模言語モデル (Large Language Model: LLM) の研究・開発が加速している。高性能な LLM の開発サイクルに欠かせないのが、LLM の性能評価である。日本語の LLM の性能評価では、日本語言語理解ベンチマーク JGLUE [Kurihara 22] などのデータセットによるリーダーボードや、高性能な LLM による生成結果の一対比較 [Zheng 23] で品質評価が行われる。代表的なリーダーボードである lm-evaluation-harness [Gao 21][*1]では、JGLUE の他に、算術計算データセット MGSM [Shi 22]、要約データセット XL-Sum [Hasan 21] など、多様なタスク [Tikhonov 21, 鈴木 20] が用いられる。生成結果の一対比較では、高性能な LLM として GPT-4 [OpenAI 23]、データセットとして Rakuda Benchmark[*2]や Japanese MT-Bench[*3]などが用いられる。これらのベンチマークでは、流暢性・正確性などの生成の品質に焦点を当てて LLM を評価している。

しかし、LLM の事業における活用では、記事の入稿規程や SEO 対策などを考慮することから、文字数制約や単語の制約の順守など、生成の制御性も求められる。英語では、制御性の評価に焦点を当てた調査 [Zhang 23] や評価データセットの構築 [Sun 23]、単語数やキーワードの有無などの自動評価可能な制御項目の評価 [Zhou 23] や、要約タスクを基にした事実一貫性などの品質の観点からの制御項目の評価 [Liu 23] が実施されている。しか

*1 本稿〆切時点、最新のリーダーボードは https://rinnakk.github.io/research/benchmarks/lm/ で掲載。

*2 https://yuzuai.jp/benchmark

*3 https://github.com/Stability-AI/FastChat/tree/jp-stable/fastchat/llm_judge

表 1　LCTG Bench の構成

タスク	データ	フォーマット	文字数	キーワード	NG ワード
要約	ABEMA TIMES	120	120	120	120
広告文生成	CAMERA	150	150	150	150

以下の条件で与えられた文章を要約して出力してください。
［条件］
70 文字以上、180 文字以下で要約すること
［文章］
小学館「週刊少年サンデー」にて連載中の『葬送のフリーレン』（原作・山田鐘人、作画・アベツカサ）の TV アニメ化が決定し、キービジュアルが公開された。
...
キャラクターの佇まいからも彼らの気持ちが伝わると良いなと思います。」と、ビジュアルに込めた想いを語っている。 ※種崎敦美の「崎」は、正式にはたつさきの字 (C) 山田鐘人・アベツカサ／小学館／「葬送のフリーレン」製作委員会

図 1　要約タスクのプロンプトの例（制御項目は文字数）

以下の [文章] で与えられた説明文に対する広告文のタイトルを、[条件] に従って 1 つ作成してください。
［条件］
「it エンジニア」という単語を使って広告文を生成
［文章］
IT/Web エンジニア採用に特化した求人･スカウトサービス「フォークウェル ジョブズ」は、経験、知識、スキルともに専門性の高い即戦力エンジニア 44,000 人が集まるスカウトサービスです。即戦力エンジニアのデータベースからマッチした人材に直接アプローチが可能！

図 2　広告文生成タスクのプロンプトの例（制御項目はキーワード）

し、日本語においてはこうした制御性に焦点を当てた取り組みは存在しない。

本研究では、日本語 LLM の制御性の評価に焦点を当てたベンチマークとして、LCTG Bench (**L**LM **C**ontrolled **T**ext **G**eneration **Bench**mark) を構築する。LCTG Bench は 2 つの言語生成タスクで構成され *4、タスク横断的な制御性の評価が行えるように設計されている。本ベンチマークを用いた実験を通じて、GPT-4 などの多言語 LLM を含む 11 種類の日本語 LLM の制御性に関する現状と課題を示す。

2. LCTG Bench の構築

LCTG Bench の構成を表 1 に示す。LCTG Bench は要約タスク、広告文生成タスクの 2 つから構成されており、LLM の制御性能をタスク横断的に評価する。ただし、LLM の生成を評価する上で、正解の生成結果に依存する評価は LLM の生成の多様性を考慮できていないという課題がある。本ベンチマークでは、各事例に正解の生成結果を用意せず入力のプロンプトのみを用意することで、LLM の出力の制御性を自動評価できるようにする。

2 つのタスクのプロンプトの例を図 1, 図 2 に示す。各プロンプトは「タスクの指示文」「制御項目に関する条件文」「タスクの対象となる文章」の 3 要素で構成する。ただし、同じ意味でも異なる表現でプロンプトが入力されることでモデルの評価結果も変化し得る [Mizrahi 23]。そこで、制御項目に関する条件文のテンプレートは、同じ条件を与える多様な表現を収録するため、クラウドソーシング *5も用いて収集した *6。タスクの対象となる文章は、公開済みデータセット及びサイバーエージェントが保有する公開可能なデータより収集した。クラウドソーシングによるテンプレートの収集方法と収集例を付録付録 A に示す。

2·1　評価する制御項目

i.　フォーマット

LLM の出力フォーマットの制御に関する課題が指摘されている中で [He 23]、Function Calling*7などの外部ツールを利用した後処理を必要とする場面がある。ところがその精度は完璧とは言えず、事業での適用においては、LLM の出力フォーマットに関する制御性が要求される。本研究では、フォーマットの観点における基本的な性能を「出力の前後に不必要な説明文などを付与しない生成の性能」として評価する。要求する生成物以外の文を付与しないよう指示する条件文（付録付録 B）を作成する。

ii.　文字数

事業における記事やタイトルの作成などの応用において、文字数に制限が付くことがある。本研究では、「指定した範囲内の文字数で生成することができるか」を評価する。条件文のテンプレートはクラウドソーシングを用いて収集する。

iii.　キーワード・NG ワード

Web サイトのタイトルやキャッチコピーの生成においては、SEO 対策などの理由から生成結果に含めたい単語を指定したいケースがある。また、誇大広告になることを防ぐため、使用を避けるべき（NG）単語や表現が存在するケースもある。本研究では、「キーワードを含んだ文章、および NG ワードを含まない文章を生成することができるか」を評価する。条件文のテンプレートはクラウドソーシングを用いて収集する。

*4　今後更にタスクを追加する予定である。

*5　Yahoo!クラウドソーシング https://crowdsourcing.yahoo.co.jp/ を用いた。

*6　条件文のテンプレートのみを収集し、具体的な値については各制御項目毎に収集する。例えば文字数に関する条件文は、テンプレートとして「X 字以上 Y 字以内で要約して下さい」を収集し、X, Y にはランダムに生成した数値を代入することで条件文を作成する。

*7　https://platform.openai.com/docs/guides/function-calling

2·2 タスクの概要と条件文の作成方法

i. 要　約

難易度が高く事業でも適用例が多い言語生成タスクとして要約タスクを採用する。要約タスクでは、図 1 に示すように、条件に従って文章を要約するよう指示したプロンプト集合のデータセットを構築する。要約元の文章は、ニュースサイト “ABEMA TIMES”*8の記事から 6 カテゴリ *9の 120 件を用いる。条件文のテンプレートに代入する値について、文字数には値の上限と下限の制約を設け、上限 200 文字、下限 50 文字のランダムな 10 の倍数の値を代入する。キーワードは、要約元の記事に登場する重要度の高い単語から選択する。NG ワードについて、実際に要約を生成する場面で重要度の高い単語は NG ワードに指定されづらい可能性がある。しかし、本タスクにおいては指定した表現を除いた出力をする能力を測るという観点から、キーワードと同様に要約元の記事に登場する重要度の高い単語から選択する。具体的には、各サンプルの記事要約を GPT-4 を用いて 5 つ生成し、得られた要約集合に共通して出現する単語を重要度の高い単語として 2 つ抽出し、キーワード・NG ワードとする。これは、GPT-4 のような高性能な LLM の生成する要約に出現しやすい単語は文書中での重要度が高い単語であるという仮定に基づいている。

ii. 広告文生成

要約タスクと比較して要求される出力の文字数が少なく、キーワードの重要度が高いタスクとして広告文生成タスクを導入する。広告文生成タスクでは、図 2 に示すように、条件に従って広告文のタイトルを作成するよう指示したプロンプト集合のデータセットを構築する。広告文のタイトル生成の元となる文章は、広告文生成ベンチマーク CAMERA [Mita 23]*10の評価データのうち付与された検索キーワードが 2 件の事例について、LP テキスト部分から収集する。文字数については、要約タスクと同様に上限と下限の制約を設けつつも、要約タスクと比較して少ない文字数帯での出力を期待するタスクであることから、上限 50 文字、下限 20 文字のランダムな 5 の倍数の値を代入する。キーワード・NG ワードについては、CAMERA のサンプルに付与されている検索キーワードを使用する。

3. LCTG Bench を用いた LLM 評価

LCTG Bench を用いた日本語 LLM の評価実験を実施することで、日本語 LLM の現状と課題および本ベンチマークの有用性を示す。

*8 https://times.abema.tv/

*9 公開上の制約から「ニュース」以外のカテゴリから「エンタメ」「スポーツ」「アニメ」「将棋」「麻雀」「HIPHOP」の 6 つを選出した。

*10 https://github.com/CyberAgentAILab/camera

与えられた文章に「it エンジニア」という単語を用いて、タイトルを作成いたします。

タイトル: 即戦力 it エンジニアのデータベースからマッチした人材に直接アプローチ

このタイトルは、条件に従って「it エンジニア」という単語を用いて、文章に含まれているキーワードをタイトルに含めることで、検索結果で上位に表示させることができます。

図 3 広告文作成タスクにおける、内容と関係のない説明文が付与された LLM の生成結果の例：文字数制御の評価において、説明文の部分も文字数に含まれてしまう。

3·1 モ　デ　ル

実験に用いるモデルは、GPT-4 などの高性能とされているモデルの他、Llama 2 [Touvron 23] や GPT-NeoX [Black 22] などのベースとするモデルの種類やパラメータ数に多様性を持たせるよう選択した。また、各種 LLM のハイパーパラメータおよびシステムプロンプトは、原則 Hugging Face Hub に掲載されている値を既定値とみなして使用した。実験に用いたモデルと各種ハイパーパラメータの設定を付録付録 C に示す。

3·2 評　　価

制御性能の評価のみでは、制御性を満たしつつも生成内容が著しくタスクと乖離がある挙動を見逃す恐れがあることから、生成の品質評価も併せて行う。また、LLM は同じプロンプトの入力に対して異なる生成結果を出力するため [Ouyang 23]、評価結果にも揺れが生じる。そこで 1 つのプロンプトに対して 3 回生成を行い、各回ごとのスコアの平均値を最終的なスコアとする。さらに、LLM は要求するタスクの内容と関係のない説明文も付与した生成をすることがある。評価する制御項目のうち、「フォーマット」では説明文の有無を評価するものの、他 3 項目、および生成の品質評価については説明文が付与されていることで、タスクに対する生成結果の評価としてはノイズとなり得る（図 3）。そこで、フォーマット以外の 3 つの制御項目と生成の品質評価においては、LLM の生成結果から GPT-4 を用いて不要な説明文を除去したのちに評価を実施する *11。

i. 制御性能の評価

フォーマットに関して、GPT-4 による不要文の除去操作の前後の生成結果を比較し、要約タスクについては前後 10 文字、広告文生成タスクについては前後 5 文字が完全一致している事例の割合を算出する *12。文字数に関しては、生成結果の文字数が条件文で指定した文字数の範囲内に収まっている事例、キーワード・NG ワードに関しては、生成結果において条件文で指定した単語が

*11 不要な説明文の除去に用いたプロンプトを付録付録 E に示す。

*12 GPT-4 による除去操作により生成結果の中間部分が変更される恐れがある。そのため、除去操作前後の生成結果の完全一致による不要な説明文有無の判断は困難であるため、前後の文字列の比較を採用する。

表 2　要約タスクの制御性 (CTG) と生成の品質 (Quality) の評価結果（ca: cyberagent, line: line-corporation）

モデル	フォーマット		文字数		キーワード		NG ワード		Average	
	CTG	Quality	CTG	Quality	CTG	Quality	CTG	Quality	CTG	Quality
gpt-4-1106-preview (GPT-4 Turbo)	**0.992**	**0.925**	0.450	0.869	**0.972**	**0.886**	**0.970**	0.775	**0.846**	**0.864**
gemini-pro [Team 23]	0.914	0.894	**0.486**	**0.881**	0.939	0.856	**0.645**	**0.817**	0.746	0.862
ca/llama2-7b-chat-japanese	0.708	0.675	0.206	0.606	0.794	0.553	0.400	0.575	0.527	0.602
ca/llama2-13b-chat-japanese	0.708	0.767	0.336	0.725	0.805	0.714	0.394	0.722	0.561	0.732
ca/calm2-7b-chat	0.881	0.478	0.219	0.428	0.808	0.444	0.303	0.403	0.553	0.438
ca/mistral-7b-chat [Jiang 23]	0.914	0.792	0.278	0.742	0.884	0.742	0.219	0.750	0.574	0.756
elyza/ELYZA-japanese-Llama-2-7b-fast-instruct	0.458	0.789	0.325	0.792	0.803	0.803	0.305	0.778	0.473	0.790
rinna/youri-7b-chat	0.911	0.692	0.166	0.683	0.647	0.609	0.492	0.611	0.554	0.649
line/japanese-large-lm-3.6b-instruction-sft	0.344	0.067	0.125	0.056	0.597	0.044	0.525	0.056	0.398	0.056
llm-jp/llm-jp-13b-instruct-full-jaster-v1.0	0.253	0.047	0.003	0.039	0.364	0.017	0.808	0.036	0.357	0.035
matsuo-lab/weblab-10b-instruction-sft	0.944	0.530	0.194	0.500	0.614	0.458	0.500	0.495	0.563	0.496

表 3　広告文生成タスクの制御性 (CTG) と生成の品質 (Quality) の評価結果（ca: cyberagent, line: line-corporation）

モデル	フォーマット		文字数		キーワード		NG ワード		Average	
	CTG	Quality	CTG	Quality	CTG	Quality	CTG	Quality	CTG	Quality
gpt-4-1106-preview (GPT-4 Turbo)	0.960	**0.987**	0.222	**0.971**	**0.851**	0.711	**0.991**	**0.924**	0.756	**0.898**
gemini-pro	0.956	0.931	0.494	0.942	0.796	0.655	0.886	0.884	**0.783**	0.853
ca/llama2-7b-chat-japanese	0.391	0.822	0.416	0.854	0.564	0.731	0.807	0.736	0.544	0.786
ca/llama2-13b-chat-japanese	0.071	0.887	0.484	0.920	0.620	**0.738**	0.918	0.866	0.523	0.853
ca/calm2-7b-chat	0.860	0.735	0.396	0.762	0.449	0.618	0.664	0.661	0.592	0.694
ca/mistral-7b-chat	0.686	0.946	0.398	0.935	0.644	0.720	0.813	0.853	0.635	0.864
elyza/ELYZA-japanese-Llama-2-7b-fast-instruct	0.749	0.615	0.236	0.817	0.640	0.542	0.773	0.684	0.600	0.665
rinna/youri-7b-chat	0.724	0.320	0.200	0.347	0.524	0.258	0.649	0.259	0.524	0.296
line/japanese-large-lm-3.6b-instruction-sft	0.582	0.309	0.080	0.269	0.440	0.242	0.609	0.283	0.428	0.276
llm-jp/llm-jp-13b-instruct-full-jaster-v1.0	**0.991**	0.218	**0.951**	0.227	0.578	0.131	0.853	0.204	0.618	0.195
matsuo-lab/weblab-10b-instruction-sft	0.876	0.671	0.255	0.651	0.376	0.563	0.709	0.582	0.554	0.617

含まれている（いない）事例の割合をそれぞれ算出する。

ii.　生成の品質評価

生成の品質評価については Rakuda Benchmark などにおける評価方法に倣い、評価器として GPT-4 を活用する。GPT-4 を用いて付録付録 D に示すプロンプトを入力することで、適切な生成ができているか否かの 2 値分類を実施し、適切な生成ができている事例の割合を算出する。

3.3　結　果・考　察

2 つのタスクにおける各モデルの制御性能と生成の品質の評価結果を表 2, 表 3 に示す。全般的に GPT-4 及び gemini-pro が制御性能、生成品質のいずれにおいても高スコアを獲得している。ただし、文字数制約については全てのモデルでスコアが低い。この結果は、言語モデルの Tokenizer がトークン単位で処理が行われるがゆえに、日本語の文字数を正確に捉えられていないことが原因と考えられる。要約タスクにおいては、NG ワードの制御性についても GPT-4 以外のモデルで低スコアの傾向にある。本結果は、LLM が否定表現を適切に捉えることができないという報告 [Truong 23] と一致している。

2 つのタスクの同じ制御項目を比較すると、スコア差が大きいモデルが存在すること、特にキーワード・NG ワードにおいてモデルを問わず全般的にスコア差が大きいことがわかる。本結果より、同じ制御項目でもタスクによりその難易度が異なる可能性を本ベンチマークが示唆していると言える。また、llm-jp のモデルが広告文生成タスクの一部制御項目で高スコアを獲得しているが、生成の品質のスコアが著しく低いことから、広告文生成タスクで高い制御性能を持つと断定することはできない。例えば、図 2 を入力した際の llm-jp のモデルの出力の一つは「it エンジニア」であり、条件は満たすものの広告文のタイトルとしては意味を成していないことが確認できる。これは、JGLUE などの分類問題において、存在しないラベルを生成し不正解となっているが意味的には正解しているというような事例を、本ベンチマークでは捕捉することができることを示唆する。さらに、本結果における Average のスコアのランキングは、既存ベンチマークの Rakuda Benchmark や lm-evaluation-harness に掲載されているモデルの性能のランキングと異なっていることから、本ベンチマークを用いることで、既存ベンチマークとは異なる新たな視座を提供できるといえる。

4.　お　わ　り　に

本研究では、日本語 LLM の制御性能を評価するためのベンチマークとして LCTG Bench を構築し、全 11 種類の日本語 LLM の制御性能を評価し、日本語 LLM の現状と今後の課題を示した。LCTG Bench は 2024 年中に公開予定である *13。

謝　　辞

ベンチマークの構築にあたって、ABEMA TIMES のデータ提供を快諾してくださった長瀬さん他担当者の皆様に、心より感謝申し上げます。

*13　https://huggingface.co/datasets/kkurihara-cs/LCTG-Bench で公開予定。

◇ 参 考 文 献 ◇

[Black 22] Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., and Weinbach, S.: GPT-NeoX-20B: An Open-Source Autoregressive Language Model (2022), abs/2204.06745

[Gao 21] Gao, L., Tow, J., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., McDonell, K., Muennighoff, N., Phang, J., Reynolds, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A.: A framework for few-shot language model evaluation (2021)

[Hasan 21] Hasan, T., Bhattacharjee, A., Islam, M. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., and Shahriyar, R.: XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages, in Zong, C., Xia, F., Li, W., and Navigli, R. eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, Online (2021), Association for Computational Linguistics

[He 23] He, Q., Zeng, J., Huang, W., Chen, L., Xiao, J., He, Q., Zhou, X., Chen, L., Wang, X., Huang, Y., Ye, H., Li, Z., Chen, S., Zhang, Y., Gu, Z., Liang, J., and Xiao, Y.: Can Large Language Models Understand Real-World Complex Instructions? (2023), abs/2309.09150

[Jiang 23] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, de las D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E.: Mistral 7B (2023), abs/2310.06825

[Kurihara 22] Kurihara, K., Kawahara, D., and Shibata, T.: JGLUE: Japanese General Language Understanding Evaluation, in Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S. eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2957–2966, Marseille, France (2022), European Language Resources Association

[Liu 23] Liu, Y., Fabbri, A. R., Chen, J., Zhao, Y., Han, S., Joty, S., Liu, P., Radev, D., Wu, C.-S., and Cohan, A.: Benchmarking Generation and Evaluation Capabilities of Large Language Models for Instruction Controllable Summarization (2023), abs/2311.09184

[Mita 23] Mita, M., Murakami, S., Kato, A., and Zhang, P.: CAMERA: A Multimodal Dataset and Benchmark for Ad Text Generation (2023), abs/2309.12030

[Mizrahi 23] Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., and Stanovsky, G.: State of What Art? A Call for Multi-Prompt LLM Evaluation (2023), abs/2401.00595

[OpenAI 23] OpenAI, : GPT-4 Technical Report, *ArXiv*, Vol. abs/2303.08774, (2023)

[Ouyang 23] Ouyang, S., Zhang, J. M., Harman, M., and Wang, M.: LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation (2023), abs/2308.02828

[Shi 22] Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., Das, D., and Wei, J.: Language Models are Multilingual Chain-of-Thought Reasoners (2022), abs/2210.03057

[Sun 23] Sun, J., Tian, Y., Zhou, W., Xu, N., Hu, Q., Gupta, R., Wieting, J. F., Peng, N., and Ma, X.: Evaluating Large Language Models on Controlled Generation Tasks (2023), abs/2310.14542

[Team 23] Team, G.: Gemini: A Family of Highly Capable Multimodal Models (2023), abs/2312.11805

[Tikhonov 21] Tikhonov, A. and Ryabinin, M.: It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning, in Zong, C., Xia, F., Li, W., and Navigli, R. eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3534–3546, Online (2021), Association for Computational Linguistics

[Touvron 23] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G.: LLaMA: Open and Efficient Foundation Language Models (2023), abs/2302.13971

[Truong 23] Truong, T. H., Baldwin, T., Verspoor, K., and Cohn, T.: Language models are not naysayers: an analysis of language models on negation benchmarks, in Palmer, A. and Camacho-collados, J. eds., *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pp. 101–114, Toronto, Canada (2023), Association for Computational Linguistics

[Zhang 23] Zhang, H., Song, H., Li, S., Zhou, M., and Song, D.: A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models (2023), abs/2201.05337

[Zheng 23] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (2023), abs/2306.05685

[Zhou 23] Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L.: Instruction-Following Evaluation for Large Language Models (2023), abs/2311.07911

[鈴木 20] 鈴木正敏, 鈴木潤, 松田耕史, 田京介, 井之上直也: JAQKET: クイズを題材にした日本語 QA データセットの構築, 言語処理学会第 26 回年次大会 (2020)

◇ 付　　録 ◇

A. クラウドソーシングによる条件文の収集方法と収集例

2·1 節節で述べている制御項目のうち、「文字数」「キーワード」「NG ワード」の 3 項目の条件文のテンプレートを収集した。条件文は、ある与えられた条件文を同じ意味になるように書き換えるタスクを実施することで収集した。収集事例を下記に示す。

- 文字数: XXX-YYY 文字で要約して, XXX-YYY 文字でまとめること, XXX 文字以上、XXX 文字以下で要約 など
- キーワード・NG ワード: 「XXX」という単語を含める, XXX という言葉を使ってください。, 「XXX」という単語は入れない, 「XXX」という言葉は使用不可 など

B. フォーマットの条件文

フォーマットの条件文には、固定の文を用いた。2 つのタスクそれぞれで用いたフォーマットの条件文を以下に示す。

- 要約タスク: 文章の要約結果のみを出力し、要約結果の前後に説明文などは付与しないでください。
- 広告文作成タスク: 広告文のみを出力し、広告文の前後に説明文などは付与しないでください。

C. 実験に用いた LLM

実験に用いた LLM の一覧と設定したハイパーパラメータの値を表 D.1 に示す。

D. 生成の品質評価プロンプト

LLM の生成の品質評価をするために、GPT-4 に入力したプロンプトを図 E.3, 図 E.4 に示す。

て公平に扱われている。よって、この加点により、全体の入所率が高まり、単独申込と兄弟姉妹同時申込の入所率の差が縮まることが確認された。

3.4　シミュレーションを通じて得た洞察

シミュレーションを実施した結果、単独申込と兄弟姉妹同時申込の入所率を近似させるためには、兄弟姉妹同時申込に対して大幅な加点が必要であるという結論に至った。現行の制度では、兄弟姉妹が在所している場合に 25 点の加点があるが、これを兄弟姉妹同時申込にも適用しても入所率の差は縮まらないことが図 4 から明らかになった。

単純に加点する点数を既存の制度に合わせてしまうと、期待した効果を達成することが難しいと予想される。このような観点からも、シミュレーションを用いて変化を予測することに価値があったと考える。

さらに、兄弟姉妹同時申込の入所率を向上させるために、単独申込の入所率をどの程度減少させることが許容できるかという議論も必要である。今回の制度変更案では、単独申込の入所率が 2-3% ポイント減少すると予測されている。このような副作用についてもシミュレーションを通じて予測し、郡山市の意思決定をサポートすることができた。

3.5　他属性への配慮

兄弟姉妹同時申込に特別な加点を設けると、他の特定の属性を持つ申込者が不利になる。郡山市では、疾病や障害、介護や看護、または一人親家庭などの状況にある申込者に対して特に保育の必要性があるとみなしていた。これらの申込者に対して兄弟姉妹同時申込に比べて相対的な不利をもたらすことは問題であると考えられた。

シミュレーションの結果、これらの属性を持つ申込者の入所率が微減することが確認された。そのため、これらの人々に対しても兄弟姉妹同時申込と同様に 160 点を追加することで、従来の入所しやすさを維持することを決定した。

4.　制度変更の実施

郡山市では 2024 年 4 月の保育所入所募集から、我々が行ったシミュレーション結果をもとに制度変更が実施された。

2023 年度までは図 6 の通り、兄弟姉妹在所と多子世帯についてのみ加点が行われた。2024 年度からは図 7 の通り、兄弟姉妹同時申請、兄弟姉妹在所、多子世帯に大きく加点されるようになった。特に、同年齢の兄弟姉妹同時申請（双子や三つ子）については 200 点が加点されている。これはシミュレーションの目的と制約における郡山市が望む優先順位（表 1）の通り、同年齢の兄弟姉妹同時申請を優先するため、この加点とした。

5.　効果検証に向けた準備

5.1　シミュレーションの限界と前提

今回のシミュレーションは、制度変更が外生的に発生するという前提の下で行われていた。つまり、制度変更があっても申込者は希望する保育所を変えずにシミュレーションを実施していた。そして、このシミュレーションの限界は申込者が制度変更に反応して行動を変化させる可能性を考慮に入れていない点にある。実際には制度変更に反応して希望する保育所を変更する可能性が存在する。例えば、制度変更によって有利になる兄弟姉妹同時申込者は希望する保育所を減らしたり、人気のある保育所を希望したりすることが考えらる。逆に、不利になる単独申込者は希望する保育所を増やしたり、人気のある保育所を敬遠したりする、といった反応が考えられる。また、新たに保育所入所を希望する人や逆に諦める人の存在もこのシミュレーションでは考えられていない。

5.2　EBPM の実践と効果検証の必要性

エビデンスに基づく政策形成（EBPM）の観点から、制度変更前後でいかなる影響があったのか検証することは重要である。そこで、今後は制度変更前後のデータを用いて、制度変更の効果検証を実施する予定である。検証項目としては、以下のような項目を検討している。

(1)　兄弟姉妹が保育所に入所しやすくなったか？
(2)　全体の入所率が減少していないか？
(3)　単独申込者の入所率が著しく減少していないか？

EBPM の一環として、これらの効果検証結果を踏まえ、次年度以降の制度変更を検討する。その実現に向け、高品質なエビデンスの収集が不可欠である。具体的な手段として、保育所への申し込み者を対象としたアンケート調査の実施や、他の自治体へのデータ提供依頼を実施している。以下では、これらの取り組みについて詳細に述べていく。

5.3　申込者に対するアンケートの実施

毎年保育所の申込を行う人は少なく、制度変更が各申込者の行動にどのように影響を及ぼしたかをデータから分析することは難しい。この問題を解決するために、郡山市と協力し、2024 年 4 月の入所申込者を対象とした Web アンケートを実施している。この調査は 2023 年 11 月から 12 月にかけて実施しており、現在（2023 年 12 月 26 日）も続行中である。

このアンケートの目的は、制度変更による行動の変化を具体的に把握することである。具体的には、申込者が制度変更の事実を申込時に認識していたかどうかを確認し、制度変更を知っていた人々に対しては、その行動に変化が生じたかどうか、例えば、保育所の希望数が変わったかどうか、人気の保育所を選択する傾向が出てきたかどうかなどを詳細に調査している。この調査結果は申込

表 D.1　実験に用いた LLM のリストとハイパーパラメータの設定 (*は本稿〆切時点は未公開のモデル)

モデル	ベースモデル	max_new_tokens	temperature	top_p
gpt-4-1106-preview (GPT4-Turbo)	-	-	-	-
gemini-pro	-	-	-	-
cyberagent/llama2-7b-chat-japanese*	Llama 2	4,096	0.9	-
cyberagent/llama2-13b-chat-japanese*	Llama 2	4,096	0.9	-
cyberagent/calm2-7b-chat	Llama 2	4,096	0.8	-
cyberagent/mistral-7b-chat*	Mistral	4,096	0.8	-
elyza/ELYZA-japanese-Llama-2-7b-fast-instruct	Llama 2	4,096	-	-
rinna/youri-7b-chat	Llama 2	4,096	0.5	-
line-corporation/japanese-large-lm-3.6b-instruction-sft	-	4,096	1.0	-
llm-jp/llm-jp-13b-instruct-full-jaster-v1.0	-	4,096	0.7	0.95
matsuo-lab/weblab-10b-instruction-sft	GPT-NeoX	4,096	0.7	0.95

E.　不要な説明文の除去プロンプト

LLM の生成結果に含まれる不要な説明文の除去の際に、GPT-4 に入力したプロンプトを図 E.1, 図 E.2 に示す。

```
以下に提示している文章は、ある文章を生成 AI を用いて要約した出力結果です。
出力には「要約」あるいはそれに類する単語を含むような文として、「以下の文章を要約します。」「【要約】」などの冒頭の説明文や「以上が要約結果になります。」などの文末の説明文が入っていることがあります。また、英語でこれらの説明文が与えられることもあります。
提示した文章に上記で述べた説明文が含まれていない場合には提示した文章をそのまま出力し、上記で述べた説明文が含まれている場合は提示した文章から説明文を除去したものを抜き出してください。文章の中間部分を編集する必要は一切ありません。文が入っていることがあります。また、英語でこれらの説明文が与えられることもあります。
[文章]
{ 生成結果 }
```

図 E.1　要約タスクにおける不要説明文の除去に用いたプロンプト

```
以下に提示している文章は、ある文章を元に作成した広告文のタイトルです。
出力には「広告文：」や「広告文を作成します」などの冒頭の接頭辞や説明文、「作成しました。」「このタイトルは、、」などの接尾辞やタイトルの後ろの説明文が含まれていることがあります。
提示した文章に上記で述べた説明文や接頭辞、接尾辞が含まれていない場合には、提示した文章をそのまま出力してください。「」や**などの記号で囲われている事例の場合、記号は全て残したまま出力してください。
上記で述べた説明文が含まれている場合は提示した文章から説明文や接頭辞、接尾辞を除去したものを抜き出してください。冒頭、末尾以外の中間部分を編集する必要は一切ありません。新しく文字を追加などをしないでください。

[文章]
{ 生成結果 }
```

図 E.2　広告文生成タスクにおける不要説明文の除去に用いたプロンプト

```
以下に要約した文章とその要約元の文章が提示されています。
要約した文章は要約元の文章を適切に要約できているかを判断してください。
適切に要約できている場合は「適切」、適切に要約できていない場合は「不適切」と回答してください。
ただし、要約元の文章から断定できない情報が要約した文章に含まれている場合も「不適切」と回答してください。
「適切」「不適切」のいずれかのみを出力し、説明文などは付与しないでください。
【要約元の文章】
{ 要約元の文章 }

【要約した文章】
{ 生成結果 }
```

図 E.3　要約タスクにおける生成の品質評価に用いたプロンプト

```
以下に、ランディングページの説明文とその説明文をもとに作成した 1 つの広告文のタイトルがあります。
説明文の内容に基づいているタイトルを作成できているかを判断してください。
適切に作成できている場合は「適切」、適切に作成できていない場合は「不適切」と回答してください。
ただし、説明文とタイトルが完全に一致している事例とタイトルとして長すぎる事例も「不適切」と回答してください。
「適切」「不適切」のいずれかのみを出力し、説明文などは付与しないでください。

【説明文】
{LP テキスト }

【広告文のタイトル】
{ 生成結果 }
```

図 E.4　広告文生成タスクにおける生成の品質評価に用いたプロンプト

著　者　紹　介

栗原健太郎

2023 年早稲田大学河原研究室の博士前期課程を卒業、サイバーエージェントに入社。現在は株式会社 AI Shift にて Voice Bot の機能の研究開発、および本研究プロジェクトのリーダーに従事。趣味はドラムを演奏すること、サウナに入ること、麻雀を打つこと、キャンプ、自然言語処理について考えること。

Kubernetes 上の機械学習基盤におけるジョブスケジューリングとクォータの管理

岩井 佑樹
Yuki Iwai
CyberAgent group Infrastructure Unit
Software Engineer
yuki_iwai@cyberagent.co.jp

keywords: Kubernetes, Kueue, Cloud Native

Summary

CyberAgent にはマルチテナント Kubernetes クラスタ上に構築された共用の機械学習基盤が存在し，機械学習ジョブや Jupyter Notebook のような様々なライフサイクルのワークロードが稼働する．本論文ではこの機械学習基盤で行なっている Kueue によるジョブのスケジューリングやクォータ管理の手法を紹介する．

1. 背　　景

計算機演算性能の著しい発達により Transformer モデル [Vaswani 17] のような学習，推論ともに非常に大きな計算コストのかかる機械学習手法の研究開発が活発に行われている．一方で D.Sculley ら [Sculley 15] が述べているように，機械学習システムはデータ量やモデルの複雑さに応じて動的に計算機リソースの利用状況が変化するため，通常のシステムと比べて実運用上の課題が多く存在する．特に CyberAgent では，Kubernetes *1 クラスタ上に構築された共用の機械学習基盤である，Cycloud ML Platform (MLP) を運用していたが，Kubernetes の基本機能のみではジョブスケジューリングやテナント間の計算機リソース使用量を考慮したクォータ管理を行うことが難しいため，「Machine Resource Management」に対する課題が存在した．そこで MLP ではジョブスケジューリングのための Open Source Software (OSS) である Kueue *2 を用いてこれらの課題を解決した．本論文では MLP で使用しているクォータ設計を示しながら，Kubernetes 上で機械学習基盤を構築するためのジョブスケジューリングとクォータ管理の手法について紹介する．

2. Kueue

Kueue は Kubernetes Special Interest Group (SIG) Scheduling で開発されている，Kubernetes-Native なジョブスケジューリングおよびクォータ管理のための OSS であり，以下に示す 5 つの項目を達成するために開発されたものである．

(1) Queueing : ジョブは必要な計算機リソースが解放されるまで開始されるべきではない
(2) Execution order : ユーザはジョブの実行順序に干渉する方法を持つべきである
(3) Fair sharing : 利用可能な計算機リソースを複数テナント間で公平に共有すべきである
(4) Flexible placement : ユーザはジョブをスケジューリングする際，場所・VM および GPU 種別・時間を柔軟に表現できるべきである
(5) Budgeting : 管理者は計算機リソースの使用量を時間単位で管理できる必要がある

また Kueue は 6 種類のオブジェクトを通じてクォータなどの設定を行うことが可能であり，次節以降では MLP での設定を説明する上で特に重要な LocalQueue・ResourceFlavor・ClusterQueue について述べる．

2·1 LocalQueue

LocalQueue は各テナント（Kubernetes Namespace）内に作成するためのオブジェクトであり，ユーザは LocalQueue を通じて後述する ClusterQueue へジョブをキューイングする．

2·2 ResourceFlavor

ResourceFlavor は On-demand や Spot のようなコンピューティングノードのライフサイクルや GPU 種別などの計算機リソースの属性を表すためのオブジェクトである．管理者は後述する ClusterQueue で ResourceFlavor ごとにクォータを設定することが可能である．

*1 https://kubernetes.io/
*2 https://kueue.sigs.k8s.io/

2·3 ClusterQueue

ClusterQueue は利用可能なクォータ設定するためのオブジェクトであり，複数の ClusterQueue で計算機リソースを共有するための Cohort と呼ばれるグループを設定することも可能である．この Cohort 機能を利用することで，ある ClusterQueue 内で計算機リソースが不足した場合，Cohort 内の別 ClusterQueue から計算機リソースを借用することが可能である．その際クォータを貸し出した ClusterQueue に新たにジョブが投入された場合，その ClusterQueue は貸し出された計算機リソースを使用しているジョブを削除（Evict）することでクォータを取り戻し，新たに投入されたジョブを開始することが可能である．なおこの動作を Kueue では Preemption と呼ぶ．

また ClusterQueue には，利用が保証されたクォータである nominalQuota，Cohort 内から借用可能なクォータである borrowingLimit が存在し，ある ClusterQueue が利用可能なクォータ（*Quota*）は

$$Quota = nominalQuota + borrowingLimit \qquad (1)$$

で表される．

加えて ClusterQueue は，First In First Out (FIFO) キューであり，複数の LocalQueue からキューイングされたジョブが順次実行される．なお PrioritySortingWithinCohort 機能を利用することで，ある Cohort 内で各 ClusterQueue の先頭にあるジョブを優先度に基づいて並べ替えることができるため，計算機リソースの借用が発生する場合にジョブの Dequeue 順序の決定にジョブの優先度を反映させることが可能である．

3. システム構成

本節では，Cycloud ML Platform (MLP) について紹介した後，MLP での Kueue を用いたクォータ管理について説明する．

3·1 Cycloud ML Platform

MLP は機械学習エンジニアや研究者などの社内ユーザが利用可能なオンプレミスのマルチテナント Kubernetes 上に構築された共用の機械学習基盤である．MLP では主に以下 4 つのサービスを提供しており，GPUaaS の一部と Distributed において Kueue を導入している．

(1) GPUaaS：GPU がアタッチされたコンテナの払い出しとマネージドな Jupyter Notebook [*3] の提供
(2) AI Platform Training：機械学習ジョブ実行基盤
(3) AI Platform Prediction：機械学習モデルのサービング基盤
(4) Distributed：MPI [Message Passing Interface Forum 23] 通信ベースの大規模分散計算基盤

*3 https://jupyter.org/

3·2 クォータの構成

MLP では LocalQueue を全てのユーザに対して一つずつ作成しており，ResourceFlavor は GPU 種別ごとに作成している．ClusterQueue は管理者のための Admin ClusterQueue，重要プロジェクトのための Project-X ClusterQueue，その他のプロジェクトのための Shared ClusterQueue の 3 種類を定義しており，全ての ClusterQueue が同一の Cohort に所属している．

また図 1 に示すように，Shared と Admin の ClusterQueue は同じクォータサイズの ClusterQueue であるが，Admin の nominalQuota を設定しないことにより，Shared に空きが存在する場合のみ Admin はジョブが実行可能となる，クォータ共存モデルを採用している．

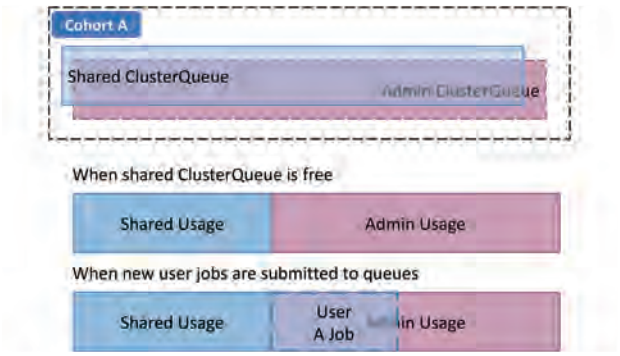


図 1 Admin ClusterQueue と Shared ClusterQueue のクォータ共存モデル

加えて図 2 に示すように，Shared と Project-X の ClusterQueue は Cohort 内 Preemption ポリシーを設定することで，Project-X に空きがある場合は Shared が Project-X からクォータを借用可能である．なおクォータの借用発生後に Project-X に対して新たなジョブが投入された際には，貸し出されたクォータを利用している Shared のジョブを Evict することで Project-X が Shared に貸し出していたクォータを取り戻すことが可能になる方式を採用している．この方式により，計算機クラスタの使用率向上と重要プロジェクトに対するクォータの準保証を同時に実現している．

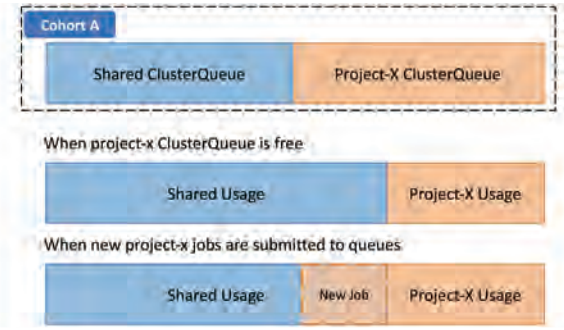


図 2 重要プロジェクト（Project-X）に対するクォータ準保証モデル

4. 今後の展望

前述した通り現在は GPUaaS の一部と Distributed に対してのみ Kueue による計算機リソースの管理を行なっているため，今後は Kueue による計算機リソースの管理を AI Platform Training と Prediction に対しても適用していく予定である．なお AI Platform Prediction で扱う推論サーバは，Dequeue 後にオートスケーリングによって使用する計算機リソースが変化する可能性があるワークロードであるが，Kueue は現在 Dequeue 後のワークロードのオートスケーリングに対応していないため，アップストリームでの動的計算機リソース再割り当ての機能開発も進めていく予定である．

5. まとめ

本論文では機械学習システムの一般的な課題である計算機リソースの管理に対して，MLP 上で Kueue を用いた改善手法を示した．Kueue は Kubernetes コア API 改善にも積極的に取り組むことで，仕様の標準化を推し進めており，これら標準化議論への参加や標準化された仕組みを MLP へ導入することによって，安定した機械学習基盤の提供に寄与していくつもりだ．

◇参考文献◇

[Message Passing Interface Forum 23] Message Passing Interface Forum, : *MPI: A Message-Passing Interface Standard Version 4.1* (2023)

[Sculley 15] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., and Dennison, D.: Hidden technical debt in machine learning systems, in *Advances in neural information processing systems*, pp. 2503–2511 (2015)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention is all you need, in *Advances in neural information processing systems*, pp. 5998–6008 (2017)

著者紹介

岩井　佑樹

2022 年に新卒で株式会社サイバーエージェントに入社．ソフトウェアエンジニアとしてプライベートクラウド上の機械学習基盤の開発に従事．Kubernetes WG Batch のメンバーであり，Kubernetes Kueue と Kubeflow WG AutoML/Training のメンテナを務める．「Kubernetes の知識地図」の執筆も手がけ，2023 年には Kubernetes SIG-Scheduling Contributor Award を受賞．

LLMを活用したテキストコンテンツ作成アプリの開発

田中 宏樹
Hiroki Tanaka
株式会社 CAM
ML Engineer
tanaka_hiroki@cyberagent.co.jp

keywords: LLM, GPT-3.5-Turbo, GPT-3.5-Turbo-16k, GPT-4, テキスト生成

Summary

現在，生成 AI は技術の最先端として大きな注目を浴びている．その中でも GPT-4 のような先進的な LLM モデルは，その高度な生成能力と広範な適用性から，業務効率化やプロダクトへの応用の可能性を秘めている．本稿では，業務効率化を目的としたテキストコンテンツ作成アプリケーションの開発事例を取り上げ，生成 AI を活用したアプリケーションの設計と開発プロセスについて詳述する．さらに，複数の LLM を組み合わせることで目的に応じたテキストコンテンツ生成が可能となり，それによる業務効率化が可能であることを示す．

1. はじめに

生成 AI には技術的進歩だけでなくビジネス領域への応用の観点でも注目が集まっている．特に，クリエイティブな作業も代替可能となってきたことで，今まで人間が行っていた創造的な作業を一部代替することが可能となり，業務効率化が期待されている．

株式会社 CAM は数多くのサービスを運用しており，これらの開発にエンジニアやデザイナーが日々奮闘している．優れたサービスの提供には大量の時間とコストが必要となるが，人的および時間的リソースは限られているため，効率化を行い生産性を向上することが求められる．

2. 課題

テキストコンテンツ作成を人手で行う場合，提供された情報の精査や信憑性の確認に加えて，サービスに適した構成や表現を考える必要がある．これはコストを増大させる要因となり，人的および時間的リソースが不足している状況では，目標とする制作数の達成が困難となる可能性がある．

この課題に対し，生成 AI でテキストコンテンツの骨組みを作成することで，業務効率改善に貢献できるのではないかという仮説のもと本アプリケーションを開発した．

3. アプリケーションの機能と実装

3·1 機能

本アプリケーションには２つの機能がある．１つ目の機能は，PDF や URL などの情報源から内容を読み取り，その内容を基にしたテキストコンテンツ作成が可能というものである．大量の情報から必要な情報を抽出して自分の言葉に置き換える作業がこの機能によって代替できる．さらに，情報源から必要な情報を抽出するという LLM の処理を追加することで，ユーザーが必要な情報をアプリケーション内で表示することが可能になる．この情報は，生成したテキストコンテンツに変更を加える場合や作成したテキストの中身を確認する場合の参考として使用できる．

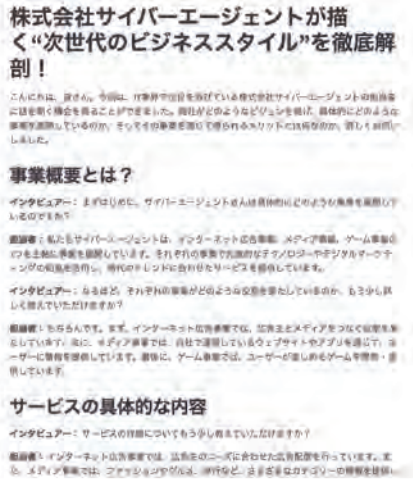
２つ目の機能は，過去のテキストコンテンツを例として掲示することで，その構成や表現方法をベースにした新しいテキストコンテンツの生成が可能というものである．これにより，ある程度ドメイン特化したテキストコンテンツであっても作成が可能になる．

3·2 システムの全体像

アプリケーション内では，GPT-3.5-turbo，GPT-3.5-turbo-16k，GPT-4 の３つの LLM を使用しており，全て Azure OpenAI Service で提供されている API を使用している．

アプリケーションの処理の流れを以下に示す．

(1) Few-shot で例として与えるテキストコンテンツを選択する
(2) 与えられた URL と PDF から文章を抽出し，3000 トークン毎に分割する
(3) 分割後のテキストをそれぞれ GPT-3.5-turbo モデルで要約する
(4) 要約後の文章を結合し，そこから必要な情報を GPT-3.5-turbo-16k モデルで抽出する
(5) 結合した要約後の文章から GPT-4 モデルでテーマ


株式会社サイバーエージェントが描く“次世代のビジネススタイル”を徹底解剖！

こんにちは、皆さん。今回は、IT事業で注目を浴びている株式会社サイバーエージェントの担当者に話を聞く機会を得ることができました。同社がどのようなビジョンを掲げ、具体的にどのような事業を展開しているのか、そしてその事業を通じて得られるメリットについて、詳しく紹介していきたいと思います。

事業概要とは？

インタビュアー：まずはじめに、サイバーエージェントさんは具体的にどのような事業を展開しているのでしょうか？

担当者：私たちサイバーエージェントは、インターネット広告事業、メディア事業、ゲーム事業の3つを主軸に事業を展開しています。それぞれの事業で先進的なテクノロジーやデジタルマーケティングの知見を活用し、時代のトレンドに合わせたサービスを提供しています。

インタビュアー：なるほど、それぞれの事業がどのような役割を果たしているのか、もう少し詳しく教えていただけますか？

担当者：もちろんです。まず、インターネット広告事業では、広告主とメディアをつなぐ役割を果たしています。次に、メディア事業では、自社で運営しているウェブサイトやアプリを通じて、ユーザーに情報を提供しています。最後に、ゲーム事業では、ユーザーが楽しめるゲームを開発・提供しています。

サービスの具体的な内容

インタビュアー：サービスの詳細についてもう少し伺えていただけますか？

担当者：インターネット広告事業では、広告主のニーズに合わせた広告配信を行っています。また、メディア事業では、ファッション、ゲーム、旅行など、さまざまなカテゴリーの情報を提供し

図 1　作成したテキストコンテンツ

を作成する

(6) Few-shot で与えた例，抽出した情報，作成したテーマから GPT-4 モデルでテキストコンテンツを作成する（図 1 参照）

上記に示したように，LLM による処理は全部で 4 回含まれている．要約や情報の抽出といった決められたルールに従って行う処理では，文章量に応じて GPT-3.5-turbo と GPT-3.5-turbo-16k を使用し，質問やテキストコンテンツの作成といった創造性が求められる処理では GPT-4 を使用している．

4. 精度向上のためのアプローチ

4·1 テキストの分割と要約

生成 AI を活用したテキストコンテンツ作成時の課題として，一度に送ることができるトークン数とリクエスト数の制限がある．この制限により，テキストコンテンツ作成時に使用する PDF や URL などの情報源のテキストサイズが大きくなると，要約ができなくなる可能性がある．そこで，情報源から抽出した文章を 3000 トークン毎に区切って要約することで，不要な情報を排除して制限に対処しながら，必要な情報のみによるテキストコンテンツ生成を可能にした．

4·2 テンプレートの選択

過去のテキストコンテンツを例として学習する機能において，与えるテキストコンテンツに工夫をしている．具体的には，汎用性が高いテンプレートだけでなくシリーズ化されているものを使用することで，言い回しや構成といった一般的な要素だけでなく，シリーズの続編の執筆などの限定的な使い方にも対応できるようにしている．

5. 今後の課題

LLM を活用した文章生成における課題として，ハルシネーションにより情報源にない情報が生成されるというものがある．生成するテキストコンテンツが社内で完結するものであれば良いが，外部に出す可能性がある場合は企業の信頼に直接関わってくるため注意しなくてはならない．その解決策として，生成過程で人間によるチェックを行うなど，人の手を介在させる方法がある．

また，生成したテキストコンテンツの評価が難しいという課題が業界全体に存在する．バナー広告のようにユーザー毎に表示する画像を変更するなどの比較が行いやすく，A/B テストを行えるケースであればそれらの指標を評価値として利用することができる．しかし，テキストコンテンツのように中身自体の価値が高く，複数のテキストコンテンツを作成して比較するという検証が行いにくいケースでは，代替できる直接的な指標が存在しない．その場合は過去のテキストコンテンツとの類似度の比較や，生成結果に対して人間による評価を付けながら出力結果を補正する必要がある．

6. ま と め

本稿では，複数の LLM を活用したテキストコンテンツ生成アプリケーションにより，テキストコンテンツの骨組みが作成可能であることを示した．そして，テキストコンテンツ生成時の精度向上のためのアプローチと今後の課題についてまとめた．LLM の進歩は早いため，紹介した課題が LLM の技術進歩で解決される可能性はあるが，システムのフローのような LLM 以外の設計によって解決する視点を持っておく必要がある．

本稿を通じて，今後の生成 AI を活用したサービス開発に対する新たな示唆を提供でき，その一助となることを期待する．

謝　辞

本アプリケーションの開発にあたり，株式会社 CAM AI Unit の皆様には大変お世話になりました．また，本稿の執筆に際しトレーナーである原和希さんには数多くのアドバイスをいただきました．深く感謝いたします．

著 者 紹 介

田中　宏樹

2023 年千葉工業大学大学院を修了後，サイバーエージェントに入社．現在は同子会社の株式会社 CAM にて生成 AI を活用した開発に従事．

編集後記

White Paper Project（以下 WPP）は、サイバーエージェントの各組織で行われている AI/Data 系の研究開発における技術を論文化するプロジェクトです。2017 年から年 2 回の頻度で発行してきました。
WPPは、以下の目的をもとに実施されています。
* AI/Data 系の研究開発の CA 全社への認知
* 社内に散らばる AI/Data 系の技術資産の集約
* 秘匿情報を含む研究成果のアウトプット
* 技術共有による車輪の再発明の防止
* 長期的な研究開発における中間的な情報共有

社内限定で集められた論文集の中から社外公開可能なものを集め、今回の社外公開版 WPPを発行することができました。
サイバーエージェントの多様な事業ドメインにおける研究開発の取り組みを知っていただく機会になることを、運営一同心より願っております。

White Paper Project 運営

White Paper Project

著者

阿部 香央莉
石上 亮介
乾 健太郎
岩井 佑樹
汪 雪テイ
大谷 まゆ
岡崎 直観
邱 倩如
栗原 健太郎
Wu Shuting
佐々木 翔大
澤井 悠
杉山 雅和
武内 慎
竹浪 良寛
田中 宏樹
張 培楠
富樫 陸
戸田 隆道
友松 祐太
二宮 大空
邊土名 朝飛
松木 一永
三田 雅人
森下 壮一郎
森脇 大輔

運営

海老澤 颯
下田 和
鈴木 智之
友松 祐太
原 和希
松月 大輔

デザイン

後谷 莉子（Design Factory）

※著者の所属は発行当時のものです


CyberAgent