LLM を用いたクロールデータからの人物略歴文抽出

中野佑哉 [1] 猪野麻巳子 [1] 二葉知泰 [1] 丸山翼 [1] 岸本耀平 [1] 永井隆広 [1]
[1]LINE ヤフー株式会社
{yuynakan,mamino,tfutaba,tsumaruy,ykishimo,takanaga}@lycorp.co.jp

概要

複数 Web サイトに点在する人物の略歴文を収集し，Web 検索結果に表示することで，検索サービスのユーザ体験向上が期待できる．しかし，略歴文収集に係る作業は，人間が行った場合においても難易度が高く，大規模な略歴文収集には大きなコストがかかる．そこで，本研究では，LLM を用いて，クロールデータから人物の略歴情報と関連する Web サイトを自動的に紐付け，引用形式を保ったまま略歴文を抽出するタスクを提案し，手法の有効性について検討する．

1 はじめに

Web 検索サービスにおいて，検索結果に表示される付加価値の高い情報を増やすことは，ユーザ体験の向上につながる重要な取り組みである．[1)] 例えば，ナレッジパネル[2)]に代表される，検索結果の直接回答 [1, 2, 3] は，ユーザの知りたい欲求に即時に応えることができるモジュールとして，Web 検索サービスに用いられている．特に検索需要の高い人物のクエリに関しては，その人物のプロフィールや出演作品など，数多くの情報が検索結果画面に直接表示されているが，多くのサービスにおいて掲出される情報は個々に独立した形で存在しており，これまでの生い立ちから代表作品，最新の出演情報までを端的に略歴情報として整理するには至っていない．

そこで我々は，Yahoo!検索における人物の直接回答モジュールにおいて，人物の略歴に関する複数のテキスト情報を Web ページからの引用形式で整理し，一つのモジュールとして掲出することで，さらなるユーザ体験向上を目指す．

我々が目指す，検索結果に略歴情報を掲出した例を図 1 に示す．元のテキストから改変を行わず，引用の形式を保つことで，事実と異なる内容の掲出を抑制するなど，情報の信頼性を担保する狙いがある．特定の Web ページから人物の略歴情報を引用形式で掲出するためには，複数の Web サイトに点在する情報を収集し，それらを整理して表示する必要がある．

図 1 略歴情報のサービス掲出例

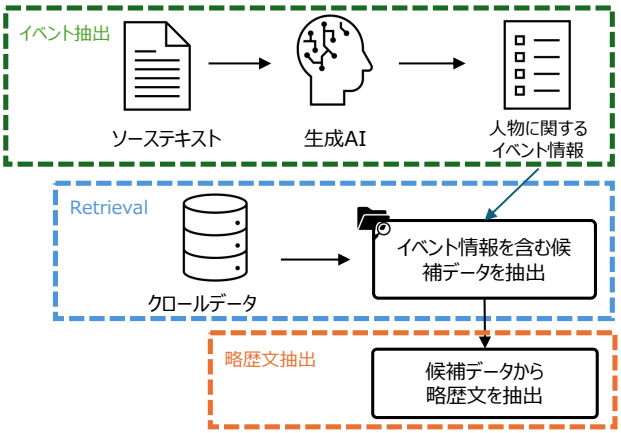

図 2 略歴文抽出の全体の流れ

しかし，略歴情報が記載されている Web サイトは多岐にわたり，その情報の信頼性や内容の適切さにもばらつきがある．また，インターネット全体から略歴情報を収集する行為は，人間が行った場合においても難易度が高く，大規模な略歴情報収集には大きなコストがかかる．

1） https://techblog.yahoo.co.jp/entry/20200210811710/
2） https://support.google.com/knowledgepanel/answer/9163198

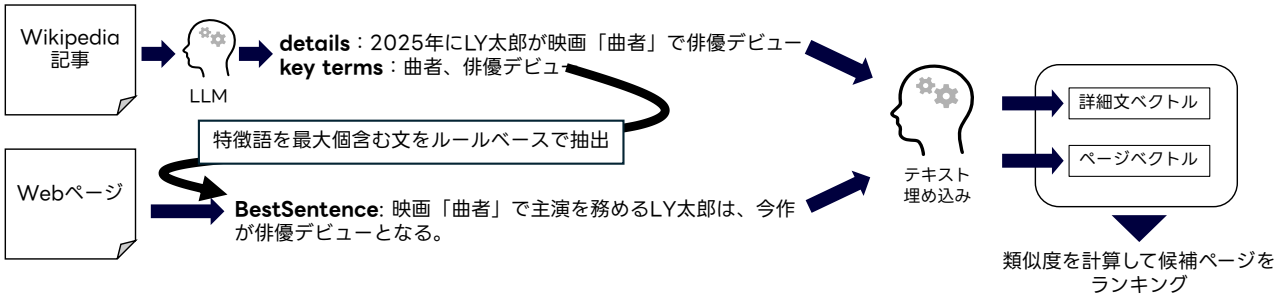

図 3　Retrieval タスクの全体の流れ

本研究では，Web テキストから人物の略歴情報を抽出するタスクを提案し，上述の課題解決に向けた手法を検討する．提案するタスクは，ソースとなるテキストから人物の略歴の基となるイベント情報を抽出するタスク，Web テキストの集合から収集したい人物の略歴情報が存在する Web サイトを自動的に紐付けるタスク，紐付けた Web サイトから略歴文を抽出するタスクで構成される．人物に関する Wikipedia 記事をソーステキストとし，大規模言語モデル（LLM）による情報抽出を行った．得られたイベント情報を用いて，独自に収集したクロールデータから関連する Web サイトを抽出し，得られた Web サイトの品質を，独自に設けた基準に従って品質を検証した．

2　関連研究

情報抽出 (Information Extraction) は文章中から重要な情報を抽出する技術として広く知られており [4, 5]，その中でもイベント情報抽出 (Event Extraction) は，文章中に記述された事象や出来事を抽出する重要なベンチマークとして，様々な研究開発が行われている [6, 7].

イベント情報抽出におけるルールベースの手法として，西田ら [8] は，人物に特徴的な属性情報に焦点を当てた手法により，Web 文書から人物に関する情報を抽出する手法を提案した．柳瀬ら [9] は，述語項構造解析により英文ニュースから企業の経営判断を伴う文の抽出を行った．これらの手法は，あらかじめ定義された属性や用語に応じたイベントの抽出に効果を発揮するが，未知の文書から柔軟にイベント情報を抽出することはできない．

ルールベースでは抽出が困難な非構造化テキストからの柔軟なイベント抽出手法として，LLM を活用した手法が注目を集めている [10, 11, 12, 13]. Sancheti ら [13] は，Web ページからの情報抽出と LLM による人物のプロフィール解釈機構を組み合わせた，同名人物の同定手法を提案した．提案手法は，人物に関する情報をあらかじめ定義したテーブルに沿って抽出を行っているが，出演情報や受賞歴など，Web 検索サービスにおいて需要の高い情報の整理を柔軟に行うには至っていない．また，LLM は流暢で自然な文を生成できるが，事実に基づかない応答を行う hallucination[14, 15] と呼ばれる現象が発生することがあり，略歴情報の信頼性を担保する工夫が必要になる．

3　略歴文の品質基準

本研究では，検索サービスへの略歴情報の掲出を念頭に，LLM が引き起こす hallucination や誤表記の無い略歴文を，信頼性の高い Web サイトから引用の形式で抽出することを目指す．最終的な出力物の品質基準として，ページ品質，イベント情報と Web ページの紐付け，抽出文の品質の 3 つを設定した．

ページ品質

- Web サイトの公式性が高い
- 対象者本人の人格を傷つけるような内容ではない

紐付け

- 抽出したいイベントに関連する内容が記載されている

抽出文

- 「生い立ち」「デビュー時の情報」「出演作品・活動内容」「受賞歴」「直近 3 年以内の出来事」などの情報が入っている
- 事実に基づいている

4　提案手法

本研究の目的である，人物の略歴テキスト抽出を実施するにあたり，対象人物にまつわる来歴や出演

情報などの情報を得るイベント抽出タスクと，得られたイベント情報とマッチするウェブサイトを引き当てる Retrieval タスクを設定した．タスクの全体の流れを図 2 に示す．本章では，各タスクの設定と提案手法について述べる．

4.1 Wikipedia からのイベント抽出

イベント抽出タスクでは，後段の Retrieval タスクでウェブテキストの紐付けを行うため，事前に人物のイベント情報を抽出する．人物の略歴情報が網羅されたソーステキストを入力とし，対象人物にまつわる来歴や出演情報などの情報を出力する．

本研究では，人物の Wikipedia 記事をソーステキストとして用いた．Wikipedia 記事は，独自にクロールした 2024 年 1 月時点のデータから抽出し，記事テキストに対して，本文部分のみを抽出する処理と，引用や上付き文字などの特殊な記号を削除する処理を施した．

Wikipedia 記事からのイベント抽出は，LLM を用いて行う．Wikipedia 記事テキストを入力とし，その中からイベント情報をテキストで抽出するようプロンプトを与える．抽出するイベント情報として，role，details，key terms，years を抽出する．それぞれの詳細は以下の通りである．

role　イベント情報の表題となるようなテキストを生成する．今回の手法においては情報の取りまとめとして利用するにとどめる．

details　イベント情報の詳細を記述した文章を抽出する．

key terms　details で抽出された文章に含まれる特徴語の集合を生成する．key terms は，後段の Retrieval タスクで Web ページとの紐付けを行う際のクエリとしての役割を想定して抽出を行う．

years　イベントが発生した西暦年を抽出する．

4.2 Retrieval タスク

Retrieval タスクでは，イベント抽出タスクで得た情報を基に，対象人物の略歴情報が存在する Web サイトを抽出し，関連度順にランキングする．イベント情報との紐付けを行うための Web データベースとして，独自に収集したクロールデータを用いる．

Retrieval タスクにおける提案手法の全体の流れを図 3 に示す．まず，4.1 節の処理で得られたイベント情報の key terms をクエリとして用いて，クロールデータからパターンマッチングで関連する候補 Web ページを抽出する．候補 Web ページを抽出する際，アダルトコンテンツや EC サイトなど，ノイズとなる特定のジャンルの Web ページを除外するため，ページ中の単語とページ URL に対してフィルタリングを行った．加えて，ページタイトルに抽出対象の人物名が含まれる Web ページのみを抽出対象とした．

次に，得られた候補 Web ページに対して，よりイベント情報の趣旨に合致するようランキングを行う．ランキングは，得られた候補 Web ページから，key terms が最大個含まれる文を bestSentence として抽出し，details の文との類似度を基準に順位付けする．bestSentence が複数含まれる場合，Web ページと bestSentence のペアを独立に扱い，順位を決定する．ここで抽出する bestSentence は，最終的な出力物である略歴文としてではなく，Web ページのランキングのみに使用する．類似度は，details と bestSentence のそれぞれのテキスト埋め込みのコサイン類似度を用いた．テキスト埋め込みには，多言語埋め込みモデルである multilingual-e5-large[3)][16] を用いた．

4.3 略歴文抽出

Retrieval タスクでイベント情報と紐付けた Web ページから，対象人物の略歴としてふさわしい文の抽出を行う．抽出は人手によって行い，作業者が直接 Web ページを閲覧し，本文中にある略歴として最も適切な文を抜き出す形で実施する．

5 実験

本章では，実験に用いるデータセットについて詳細を述べる．次に，提案手法を用いてイベント抽出と Retrieval タスクを行い，出力物の評価について述べる．

5.1 イベント情報評価

イベント抽出タスクでは，対象人物の Wikipedia 記事から，その人物にまつわる来歴や出演情報などのイベント情報を抽出する．作成したデータセットを LLM への入力とし，role，details，key terms，years をイベント情報として抽出するようプロンプトを与えた．与えたプロンプトは，付録 A の表 5 に示す．検索需要を基に独自に選定した 183 人に対して，イベント情報を抽出した．抽出に用いる LLM には，

3）　https://huggingface.co/intfloat/multilingual-e5-large

表 1 イベント抽出の出力例

	gpt-4o-2024-11-20
role	DAIGO ☆ STARDUST としてメジャーデビュー
details	氷室京介のプロデュースで「MARIA」でメジャーデビュー。宇宙から舞い降りたロック王子という設定で活動。
key terms	”DAIGO ☆ STARDUST”, ”MARIA”, ”氷室京介”, ”ロック王子”
years	2003

表 2 抽出したイベント情報の統計

	GPT-4o
抽出したイベント情報	1,832
details の平均文字長	37.10
一人あたりの平均抽出数	9.96
key terms の平均抽出数	1.93

表 3 抽出したサイトと bestSentence の品質評価結果

	基準内 (%)	基準外 (%)
ページ品質	94.5	5.50
bestSentence	38.4	61.6

OpenAI 社から提供される GPT-4o[4] を用いた．モデルによるイベント情報の出力例を表 1，出力物の統計情報を表 2 にそれぞれ示す．出力例から，GPT-4o は，details を情報の過不足なく記述し，指示文に忠実な key terms を抽出できていることがわかる．

5.2 Retrieval 結果評価

提案手法を用いて抽出，および，ランキングを施した Web ページに対して，ページ自体の品質，イベント情報との紐付け，bestSentence を略歴文とみなした場合の品質について，3 名のアノテータによる人手評価を行った．使用したイベント情報は，GPT-4o の出力を使用した．Web ページは抽出対象人物のイベント情報毎に，ランキング順に最大 5 ページ，1 人につき最大 10 件のイベントが紐付いており，これを 9 名分評価した．評価観点として，それぞれ 3 章の基準をすべて満たすページ，または，文を”基準内”と判定した．

評価結果を表 3 に示す．Web ページ自体の品質については，94.5% が基準を満たした一方，bestSentence は 38.4%に留まる結果となった．

次に，略歴文抽出タスクを実施し，略歴文が抽出可能かを評価した結果を表 4 に示す．正解データのラベル付けが施されていないクロールデータからの抽出であることを鑑み，Precision@k を評価指標

表 4 略歴文抽出の評価結果

	人物 A	人物 B	人物 C
Prec@5 (人手抽出)	0.78	1.0	0.70
Prec@5 (bestSentence)	0.44	1.0	0.40
紐付け Web ページ総数	762	2,553	7,649

とし，人手による抽出と bestSentence をそのまま略歴文とみなした場合の結果を算出した．人物 A (女性俳優)，人物 B (男性俳優)，人物 C (男性ミュージシャン) の 3 名を抽出対象とし，人手によって略歴文が抽出可能であった Web ページを人手による抽出の正解と判定した．人物 B は，上位 5 ページの中に略歴文がすべて存在したが，他 2 名の品質は大きく劣る結果となった．これは，紐付けた Web ページの総数や人物の属性，知名度に影響された可能性が考えられる．

人物 A と人物 C に関して，人手による抽出と比較して，bestSentence をそのまま略歴文とみなした場合，抽出性能の低下が見られた．本手法では，一度パターンマッチングで抽出した Web ページのランキングを，bestSentence と details との類似度を基準にしており，Web ページとイベント情報の紐付けや，略歴文抽出全体の品質改善に向けて，bestSentence の抽出には改善の余地が示唆される．基準内，基準外の例は付録 B の表 6 に示す．

6 おわりに

本研究では，Web ページから人物の略歴情報と関連する Web サイトを自動的に紐付け，引用形式を保ったまま略歴文を抽出するタスクを提案し，クロールデータと LLM を用いた手法を提案した．提案手法を用いてイベント抽出と Retrieval タスクを行い，手法の有効性を確認した．

今後の展望として，データセット全体に対する詳細な分析と実サービスへの応用，略歴文抽出タスクの自動化に取り組む．

4） https://platform.openai.com/docs/models/gp#gpt-4o

謝辞

本論文の執筆にあたり，有益な助言を頂いたLINE ヤフー株式会社の寺井朝人氏，宇城毅犠氏，芦原和樹氏，沖本祐典氏に感謝いたします．

参考文献

[1] Zhijing Wu, Mark Sanderson, B. Barla Cambazoglu, W. Bruce Croft, and Falk Scholer. Providing Direct Answers in Search Results: A Study of User Behavior. In **Proceedings of the 29th ACM International Conference on Information & Knowledge Management**, CIKM '20, pp. 1635–1644, New York, NY, USA, October 2020. Association for Computing Machinery.

[2] Artur Strzelecki and Paulina Rutecka. Direct Answers in Google Search Results. **IEEE Access**, Vol. 8, pp. 103642–103654, 2020.

[3] Andrew D. Haddow and Sarah C. Clarke. Inaccuracies in Google's Health-Based Knowledge Panels Perpetuate Widespread Misconceptions Involving Infectious Disease Transmission. **The American Journal of Tropical Medicine and Hygiene**, Vol. 104, No. 6, pp. 2293–2297, June 2021.

[4] Jim Cowie and Wendy Lehnert. Information extraction. **Communications of the ACM**, Vol. 39, No. 1, pp. 80–91, January 1996.

[5] Kailash A. Hambarde and Hugo Proença. Information Retrieval: Recent Advances and Beyond. **IEEE Access**, Vol. 11, pp. 76581–76604, 2023.

[6] Wei Xiang and Bang Wang. A Survey of Event Extraction From Text. **IEEE Access**, Vol. 7, pp. 173111–173137, 2019.

[7] Walker, Christopher, Strassel, Stephanie, Medero, Julie, and Maeda, Kazuaki. ACE 2005 Multilingual Training Corpus, February 2006.

[8] 西田成巨, 森辰則. Web 文書からの人物情報の抽出. 言語処理学会 第 16 回年次大会, pp. 359–362, 2010.

[9] 柳瀬利彦, 柳井孝介, 丹羽芳樹, 村上聡一朗, 渡邉亮彦, 宮澤彬, 五島圭一, 高村大也, 宮尾祐介, 中田亨. 企業経営における意思決定支援のためのイベント抽出. 人工知能学会全国大会論文集, Vol. JSAI2017, pp. 3J11–3J11, 2017.

[10] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey. **Frontiers of Computer Science**, Vol. 18, No. 6, p. 186357, November 2024.

[11] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. **Nature Communications**, Vol. 15, No. 1, p. 1418, February 2024.

[12] Mengna Zhu, Kaisheng Zeng, JibingWu JibingWu, Lihua Liu, Hongbin Huang, Lei Hou, and Juanzi Li. LC4EE: LLMs as Good Corrector for Event Extraction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 12028–12038, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[13] Prateek Sancheti, Kamalakar Karlapalem, and Kavita Vemuri. LLM Driven Web Profile Extraction for Identical Names. In **Companion Proceedings of the ACM Web Conference 2024**, WWW '24, pp. 1616–1625, New York, NY, USA, May 2024. Association for Computing Machinery.

[14] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5271–5285, Seattle, United States, July 2022. Association for Computational Linguistics.

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. **ACM Computing Surveys**, Vol. 55, No. 12, pp. 248:1–248:38, March 2023.

[16] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text Embeddings by Weakly-Supervised Contrastive Pre-training. **arXiv:2212.03533**, 2024.

表 5　イベント抽出で用いたプロンプト

Please extract the representative biographical information of a person from a Wikipedia article. Be sure to adhere to the following constraints:

==Constraints==
- 日本語で出力してください.
- Output the events in the order of importance to that person.
- The AI must extract at least 10 pieces of careers information. If it cannot be found, think again carefully and try to extract it.
- Output as many important events for the person as possible.
- Extract a wide range of events from debut to recent ones, including notable works.
- Do not extract information about negative events.
- Be sure to include titles of works appeared in, character names in works, and publication titles as key terms.
- Do not duplicate events when extracting them. Retain only the first extracted instance of similar events.
- Extract only the most well-known part of the title as key terms.
- Extract as many named entities as possible as key terms. For non-named entities, extract only the distinctive words useful for search queries as key terms.
- Do not include words with low TF-IDF scores, such as job titles or TV station names, as key terms.
- If you are unsure about your judgment, take a deep breath and think carefully.

===Wikipedia article===
{input_text}

表 6　基準内 Web ページと bestSentence の例

	ページタイトル	bestSentence
基準内	daigo、氷室京介&hyde の言葉に支えられた過去告白「勇気もらった」	2003 年に DAIGO ☆STARDUST 名義で、シングル曲『MARIA』でメジャーデビューし、その後 3 人組
基準外	北川景子と交際中の daigo「jp=人生のピーク」-作詞作曲にも恋愛は”大事”	関連記事芸能 DAIGO、氷室京介&hyde の言葉に支えられた過去告白「勇気もらった」2015/03/1314:15 芸能北川景子がエランドール新人賞「1 つだけ言えることは”IH”。

A　イベント抽出用プロンプト

提案手法において, イベント抽出を行うにあたり, LLM に与えたプロンプトを表 5 に示す．{input_text} には，Wikipedia 記事の本文が入力される．抽出に際し，以下の観点を含めるよう LLM のプロンプトを調整した．

- 最大 10 件のイベントを抽出する
- デビューから最新情報まで幅広く抽出する
- ネガティブな情報を含めない
- 作品名，役名，出版物名などの固有名詞を含める

B　基準内 Web ページと略歴文の例

表 1 に示した略歴情報の掲出例において，基準内と判定された Web ページのタイトルと bestSentence の例を表 6 に示す．基準内と判定された Web ページ，bestSentence は，抽出対象のデビュー当時に関するイベント情報に関連した内容になっているのに対して，基準外と判定された Web ページは，どちらも関連性が低い内容になっていることがわかる．

特に，bestSentence は，Web ページの本文とは異なる周辺箇所に記載された内容を抽出しており，これは，Web テキスト全体からパターンマッチングを行ったことに起因する．提案手法では，details と bestSentence によるランキングでイベント情報と Web ページの紐付けを行っており，bestSentence の品質が紐付けの精度に直接影響する．誤った紐付けを減らすためには，抽出対象とする本文をあらかじめ絞り込むなどの手法が有効であると考えられる．