Biographical Summary Generation from HTML Documents Using LLMs

Hiroshi Tanaka
Graduate School of Informatics
Osaka Metropolitan University
Osaka, Japan
nyankoroad4649@gmail.com

Harumi Murakami
Graduate School of Informatics
Osaka Metropolitan University
Osaka, Japan
harumi@omu.ac.jp

***Abstract*—This paper presents a method for extracting person attribute information from multiple HTML documents and then generating a biographical summary using large language models (LLMs). A biographical summary consists of a person attribute sentence (basic information) and a person achievement sentence (accomplishments). Person attribute sentence generation consists of a three-stage extraction process (extraction, merging, verification) using LLMs and template-based generation, while person achievement sentence generation consists of a two-stage extraction process (extraction, merging) using LLMs and LLM-based generation. To mitigate risks of hallucination, all extracted information includes source attribution and undergoes a verification process to filter out inaccurate extractions. Experimental evaluation using 80 persons and 832 HTML files demonstrates significant improvement in recall for occupation and affiliation compared to the prior rule-based method for person attribute sentence generation. A comparative experiment on person achievement sentence generation shows that our method with relevance annotations outperforms direct generation and basic extraction across automatic evaluation metrics.**

***Index Terms*—information extraction, large language models, HTML processing, person attributes, text generation**

I. Introduction

With the growth of the internet, identifying and understanding persons on the web has become increasingly important. Structured person information, such as that found on Wikipedia, greatly contributes to person understanding by providing concise summaries including proper reading of a name, birth date, birthplace, occupations, and affiliations. However, many individuals do not have a Wikipedia page, and for such persons, this information must be integrated from multiple HTML pages obtained through web search results.

Murakami et al. [1] proposed a method for generating biographical summaries from web search results using rule-based pattern matching with regular expressions. However, pattern matching could not handle the diverse expression formats found in HTML documents, resulting in low recall for occupations (24%), affiliations (59%), and positions (38%).

Recent advances in LLMs can address this limitation, since LLMs are able to understand context and handle diverse expression formats [2], [3]. However, applying LLMs to information extraction introduces the risk of hallucination output, where LLMs generate information not present in the source. For person information, such inaccuracies—e.g., incorrectly guessing names by misreading kanji characters or associating unrelated information with the target person—are particularly problematic.

We propose a method that replaces rule-based extraction with LLM-based extraction while also introducing source attribution and verification mechanisms to address the issue of hallucination. The study's main contributions include the following:

- A three-stage process (extraction, merging, verification) for accurate attribute extraction from multiple HTML documents
- Achievement sentence generation with relevance and interpretation annotations
- Source attribution and verification mechanisms to mitigate risks of hallucination

II. Related Work

Murakami et al. [1] analyzed Wikipedia person pages and designed a template for generating biographical summaries. They used regular expressions and morphological analysis to extract person attributes, along with majority voting for selection among multiple candidates. Our work directly extends this approach by replacing rule-based extraction with LLM-based methods.

Gur et al. [2] showed that LLMs fine-tuned with small amounts of task data could achieve high performance in HTML processing, completing 50% more tasks than prior models on the MiniWoB benchmark. Liu et al. [3] demonstrated that GPT-4 achieved 100% accuracy for email extraction and 98% for phone number extraction, significantly outperforming regex-based methods.

Developed through research on LLM-based text generation, STORM [4] automatically generates Wikipedia-like articles by collecting information from diverse perspectives. WikiBio [5] generates person introductions from structured infobox data. However, these methods target conventional Wikipedia articles or assume structured input, so they do not address the challenge of extracting person attributes from unstructured HTML and generating person-focused overview sentences. In contrast, our work focuses specifically on biographical summary generation from raw HTML documents.

III. Proposed Method

Our method generates biographical summaries from multiple HTML documents (roughly 10 per person on average). A biographical summary consists of a person attribute sentence and a person achievement sentence. Fig. 1 gives an overview of the system architecture.

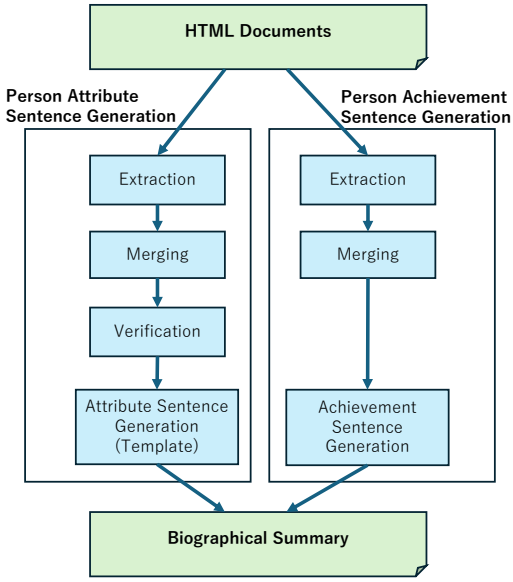

Fig. 1. Overview of biographical summary generation system.

A. Person Attribute Sentence Generation

The person attribute sentence describes basic person information (name reading, birth/death dates, birthplace, occupations, affiliations) and is generated through a three-stage extraction process (extraction, merging, verification) followed by template-based sentence generation.

1) Target Attributes: Following prior work [1], we extract six types of attributes to generate sentences like those in Wikipedia's opening sentence format, as shown in Table I.

TABLE I
Target Attributes with Examples

Attribute	Description	Example
Name reading	Name pronunciation	Kurihara, Harumi
Birth date	Year/month/day	1947/3/5
Death date	If deceased	–
Birthplace	Prefecture/city	Shizuoka, Shimoda
Occupation	General profession	Food researcher
Affiliation/Position	Affiliation-position pair	Yutori no Kukan Co., Ltd., CEO

2) Extraction Process: Multiple HTML files for a single person are processed by the LLM to extract attributes in JSON format. When the total token count exceeds a context limit, documents are split into sets and processed separately. Each extracted attribute includes its value, source HTML filename, and notes.

The extraction prompt incorporates three design principles: (1) *Prohibition of inference*: Readings of words/expressions guessed from kanji characters must not be extracted unless they are explicitly written elsewhere in the HTML document. Null values are permitted as a way to prevent forced guessing. (2) *Mandatory source attribution*: All extractions must include the source HTML filename for later verification. (3) *Priority ordering*: Array items are extracted in order of importance.

Table II shows an example of set-wise extraction for Asako Miura (28 HTML files split into 4 sets).

TABLE II
Set-wise Extraction Example

Set	#HTML	Extracted Occupations
1	6	researcher, professor, social psychologist
2	7	social psychologist, blogger
3	12	researcher
4	3	–

3) Merging Process: Extraction results from multiple sets are merged by the LLM into a single unified JSON file. Since each set processes different HTML files, the extracted values may vary in granularity and coverage. The LLM's natural language understanding is leveraged to integrate synonyms, normalize notation variations, and compare granularity levels.

The merging process follows three rules: (1) Prefer higher granularity (e.g., "Kwansei Gakuin University, Faculty of Letters" over "Kwansei Gakuin University"). (2) Remove duplicates by merging related items (e.g., "researcher" and "social psychologist" $\rightarrow$ retain the more specific "social psychologist"). (3) Sort by importance (most representative item first).

For the example in Table II, the merged result yields: "social psychologist, professor, blogger"; here, the generic "researcher" is subsumed by the more specific "social psychologist," and items are sorted by representativeness.

4) Verification Process: For each attribute in the merged result, the LLM cross-references the extracted value with the source HTML to determine whether to adopt, modify, or exclude the information. The verification process checks three criteria: (1) whether the value actually appears in the HTML, (2) whether it is an attribute of the target person (not another person mentioned on the page), and (3) whether it satisfies the attribute requirements (e.g., occupations should be general professions).

The LLM assigns a confidence score (1–10) and provides reasoning. Attributes scoring below 6 are excluded. Table III shows examples of verification decisions.

The Egawa case demonstrates a case of modification: HTML states "former pro baseball player," so "former" is added. The Kurihara case demonstrates a case of exclusion: While "lecturer" appears in the HTML, "NHK" is not ex-

TABLE III
VERIFICATION EXAMPLES

Person	Attribute	Value	Score	Action
Miura	Occupation	social psychologist	10	Adopt
Egawa	Occupation	pro baseball player	9	Modify*
Kurihara	Affiliation	NHK (lecturer)	0	Exclude

*Modified to "Former pro baseball player"

plicitly stated as her affiliation, and thus the attribute is not adopted.

5) Template-based Sentence Generation: The person attribute sentence is generated from verified attributes using a predefined template:

[Name] ([Name reading], [Birth date] - [Death date]) is a [Occupation] from [Birthplace]. [Affiliation] [Position].

When attributes are missing, complement rules ensure that natural sentences are maintained: no Birthplace → "Japanese"; no Occupation → "person"; no Affiliation → omit the second sentence.

B. Person Achievement Sentence Generation

While the person attribute sentence describes basic information using a template, the achievement sentence describes accomplishments and requires natural language generation, since each person's achievements are unique and cannot be templated. Achievement information extraction consists of a two-stage process: extraction and merging. The person achievement sentence is produced through the proposed LLM-based sentence generation.

Extraction: The LLM extracts achievement-related information from HTML documents and classifies them into five categories: field, occupation, achievements, evaluations, and features. Each item includes not only its value but also relevance (how the information relates to the person) and interpretation (what can be understood from it), which together help to filter out irrelevant information and support sentence generation.

Merging: Extracted items from multiple sets are merged using a point-based scoring system (rule-based, no LLM). Items extracted at higher-priority positions across multiple sets receive higher scores, reflecting their importance.

Achievement Sentence Generation: Based on the merged results, the LLM generates a natural 1–2 sentence summary that prioritizes high-scoring items.

C. Biographical Summary Examples

Fig. 2 shows an example of system output for a film critic. The biographical summary consists of the person attribute sentence and the achievement sentence describing his accomplishments. An infobox giving the attribute information is displayed to the right of the biographical summary.

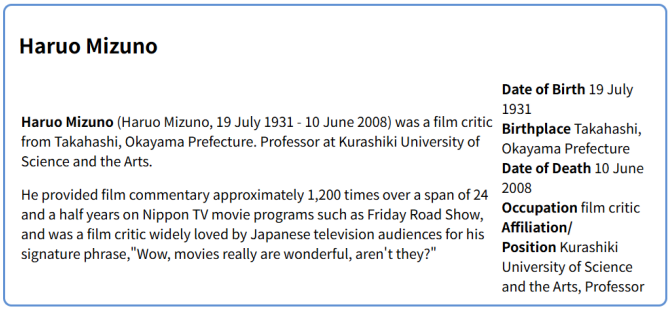

Fig. 2. Example output of biographical summary generated by proposed method.

IV. EXPERIMENTS

A. Dataset and Implementation

We used a dataset of 80 persons and 832 HTML files collected from web search results for 20 Japanese person names [6]. The persons include academics, artists, athletes, and professionals; in some cases, multiple persons share the same name (e.g., "Kazuhiko Kinoshita" corresponds to 16 different persons). Ground truth data were manually created following strict guidelines: Extract only information explicitly presented in HTML and record the source HTML filename for each value. We used GPT-5-mini as the LLM backend. For achievement sentence evaluation, Named Entity Match F1 was used to measure the overlap of named entities between generated and reference sentences. Named entities were extracted using GiNZA, a spaCy-based Japanese NLP library, and matched using both exact and partial string matching with NFKC normalization.

B. Attribute Extraction Results

Table IV shows the attribute extraction results. Compared to the prior rule-based method [1], recall for occupations improved from 24% to 84%. While the previous work evaluated affiliation (59%) and position (38%) separately, our method integrates these attributes, thus achieving a unified recall of 78%. The verification process successfully removed 5 inappropriate extractions, including misextracted SNS user IDs.

TABLE IV
ATTRIBUTE EXTRACTION RESULTS

Attribute	Precision	Recall	F1
Name reading	0.94	0.97	0.96
Birth date	1.00	1.00	1.00
Death date	1.00	1.00	1.00
Birthplace	0.87	1.00	0.93
Occupation	0.78	0.84	0.81
Affiliation/Position	0.77	0.78	0.77

C. Achievement Sentence Results

Of the 80 entries, 68 (85%) could be used to successfully generate achievement sentences, while 12 resulted in

empty generation due to insufficient achievement information in the source HTML. To evaluate the effectiveness of relevance/interpretation annotations, we compared three approaches (Table V).

- **Direct Generation**: HTML documents are directly input to the LLM without intermediate extraction or merging.
- **Basic Extraction**: Extraction of only values, without relevance/interpretation annotations, followed by merging and generation.
- **Proposed**: Extraction with relevance and interpretation annotations, followed by merging and generation.

TABLE V
COMPARISON OF ACHIEVEMENT SENTENCE GENERATION METHODS

Method	R-1	R-2	R-L	BERT	NE
Direct	0.51	0.28	0.38	0.83	0.46
Basic	0.50	0.29	0.40	0.84	0.42
Proposed	**0.54**	**0.38**	**0.45**	**0.87**	**0.58**

R-1/2/L: ROUGE-1/2/L, BERT: BERTScore F1, NE: NE Match F1

The proposed method achieved the highest scores in all metrics. Differences were particularly notable in ROUGE-2 (+0.09 over basic) and named entity match F1 (+0.16), indicating that the relevance and interpretation annotations contribute to the generation of more accurate and content-rich achievement sentences.

D. Discussion

1) Attribute Extraction: While name readings and dates achieved high F1 scores, occupation (F1=0.81) and affiliation/position (F1=0.77) showed relatively lower performance. This is attributed to the ambiguity of determining the primary occupation for a person with multiple roles (e.g., a university professor who is also a TV commentator) and the difficulty in correctly pairing affiliations with positions across multiple affiliations.

The verification process proved effective in removing hallucinated information: It filtered out cases where the LLM incorrectly extracted SNS user IDs as person names and, moreover, identified contradictions between HTML tags and body text. When source HTML lacked sufficient information, the system appropriately produced empty values rather than hallucinated content.

2) Achievement Sentence Generation: Of the 80 entries, 12 (15%) resulted in empty generation. Manual inspection confirmed these cases were appropriate: The HTML documents contained only basic profile information without notable achievements, demonstrating that the system can appropriately cancel generation rather than hallucinate content.

Results of the comparative experiment reveal that basic extraction provides marginal improvement over direct generation (BERTScore: 0.83→0.84), while adding relevance and interpretation annotations yields substantial gains (0.84→0.87). The named entity match F1 shows the largest gap (+0.16 over basic extraction), suggesting that annotations help the LLM identify the most important factual details and insert them in the generated sentences.

3) Limitations: This study has several limitations. First, evaluations were conducted by a single evaluator; consequently, multiple evaluators and inter-rater agreement verification are needed enhancements. Second, the dataset consists of only 80 persons; larger and more diverse datasets would better validate generalization. Third, the achievement sentence generation does not include a verification process, which would further improve the factuality of output sentences. Fourth, the current system processes only Japanese HTML documents; therefore, evaluation of multilingual sources remains future work.

V. CONCLUSION

We proposed a method for extracting person attributes from HTML documents and generating biographical summaries using LLMs. Our three-stage extraction process, featuring source attribution and verification, significantly improves recall for occupations (24%→84%) compared to the prior rule-based method for person attribute sentence generation. Affiliation and position, evaluated separately in the prior work (59% and 38%), are integrated in our method to achieve a unified recall of 78%. For achievement sentence generation, a comparative experiment demonstrated that our method with relevance and interpretation annotations outperforms both direct generation and basic extraction, achieving a BERTScore F1 of 0.87. Future work includes expanding the dataset, introducing multiple evaluators, and extending the system to handle multilingual HTML sources.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 19K12718, 25K15814.

REFERENCES

[1] H. Murakami, T. Konishi, and Y. Ura, "Generating Wikipedia-like biographical sentences from web people search results," in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 995–996, 2017.

[2] I. Gur et al., "Understanding HTML with large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2803–2821, 2023.

[3] Y. Liu, Y. Jia, J. Jia, and N. Z. Gong, "Evaluating LLM-based personal information extraction and countermeasures," in *34th USENIX Security Symposium (USENIX Security 25)*, pp. 1669–1688, 2025.

[4] Y. Shao et al., "Assisting in writing Wikipedia-like articles from scratch with large language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, 2024.

[5] R. Lebret, D. Grangier, and M. Auli, "Neural text generation from structured data with application to the biography domain," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1203–1213, 2016.

[6] S. Sato, K. Kazama, K. Fukuda, and K. Murakami, "Distinguishing between people on the web with the same first and last name by real-world oriented web mining," *IPSJ Transactions on Databases*, vol. 46, pp. 26–36, 2005. (in Japanese)