6N-01

Sentence-BERT と ChatGPT を用いた Web 上の人物への NDLSH の付与

白川欣岳† 下倉雅行† 村上晴美†
大阪公立大学大学院情報学研究科†

1. はじめに

人物検索は Web 検索において重要な課題の一つである．Web 上の人物検索においては，同姓同名人物の混在などにより目的の人物にたどりつけないことや，人物を識別できないことがある．我々は Web 上の人物の選択と理解を支援するために，Web 上の人名検索結果の要約と可視化研究を行っている[1]．本研究では，Sentence-BERT と ChatGPT を用いて Web 上の人物に国立国会図書館の件名標目である NDLSH を付与する手法を検討した．NDLSH を付与することにより，人物にゴミの少ないキーワードを付与できるとともに，上位語，下位語，関連語を用いた探索的な検索が可能となる．

2. NDLSH

NDLSH（National Diet Library Subject Headings, 国立国会図書館件名標目表）は国立国会図書館が資料の主題検索のために作成した件名標目表であり，件名標目，標目よみ，ID，同義語，上位語，下位語，関連語，注記，分類記号(NDLC)，分類記号(NDC9)，分類記号(NDC10)，参照(LCSH)，参照(BSH4)，出典(BSH4)，出典，編集履歴，作成日，最終更新日の 18 項目で構成される．

3. 提案手法

Sentence-BERT を用いた手法と ChatGPT を用いた手法の 2 種類を提案する．

3.1 Sentence-BERT を用いた手法

Sentence-BERT（以下，SBERT）を用いて NDLSH の標目と人物（Web ページ）をそれぞれ 768 次元のベクトルで表現し，それらのコサイン類似度を計算することで該当人物に NDLSH を付与する．本研究では，日本語で事前学習された SBERT として，日鉄ソリューションズ株式会社所属の園部勲氏が公開しているモデルを利用した．

図 1 に手法の概要を示す．NDLSH ベクトル 3 種類，人物ベクトル 7 種類を組み合わせて 21 手法とし，人物に NDLSH を 10 件付与した．

Assigning NDLSH Headings to People on the Web Using Sentence-BERT and ChatGPT

†Yoshitaka Shirakawa, Masayuki Shimokura, Harumi Murakami, Graduate School of Informatics, Osaka Metropolitan University

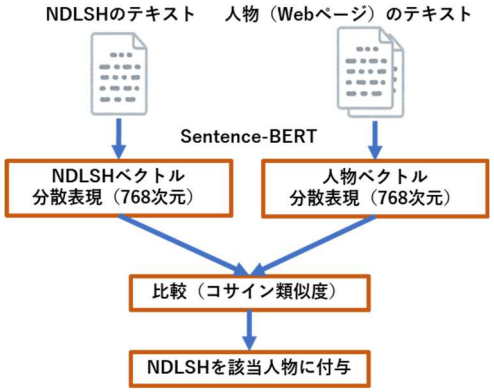

図 1 手法の概要

3.1.1 NDLSH ベクトル

NDLSH ベクトルは以下の 3 種類である．

- **(N1)** NDLSH の標目(単語)を SBERT にかけたベクトル（例「年縞」）
- **(N2)** Wikipedia の summary 部分を SBERT にかけたベクトル（例「年縞（ねんこう，英: varve）とは…」）
- **(N3)** (N-1)と(N-2)のベクトルの平均

3.1.2 人物ベクトル

人物ベクトルは以下の 7 種類である．

- **(P1)** 最上位ページの上から 800 字をベクトル化
- **(P2)** 最上位ページの上から 3 文をベクトル化
- **(P3)** 最上位ページの人名を含む 1 文をベクトル化
- **(P4)** ChatGPT(GPT-4.0)を用いて最上位ページを要約させた文章をベクトル化
- **(P5)** 各ページの上から 800 字をベクトル化
- **(P6)** 各ページの上から 3 文をベクトル化
- **(P7)** 各ページの人名を含む 1 文をベクトル化

ただし，(P5～P7)はページ毎に NDLSH を 10 件抽出し，その中でコサイン類似度が高い順に 10 件を人物に付与した．

3.2 ChatGPT を用いた手法

ChatGPT (GPT-4.0)を使用し，人物の最上位ページの HTML ファイルを読み込み，人物に関する内容を 500 字程度で要約させた．表 1 に ChatGPT に与えたプロンプトと，ChatGPT から得られた回答を示す．

回答のテキストを NDLSH と照合して頻度順に人物に付与した．今回は下倉ら[2]を参考に，NDLSH からは全角 1 文字のみ，() を含むもの，

ーー(ハイフン2つ)が含まれる語をあまり重要ではない，あるいは照合がうまくいかないと考えて除去している．また，NDLSH は文字列が長いほどより詳細な意味を持つと考え，長い文字列から順に照合している．例えば，「社会心理学」という文字列が文章中にある場合，NDLSH の「社会心理学」はカウントするが「理学」や「心理学」はカウントしない．

表1　ChatGPT のテキスト生成例

ChatGPT に与えたプロンプト	 このHTMLファイルを読み込んで野間佐和子に関する内容を500字程度で要約してください．
回答	野間佐和子氏は，講談社の6代目代表取締役社長で，日本雑誌広告協会の理事長や読書推進運動協議会の会長を務めていました．彼女は…

4.　評価実験

4.1 方法

20 の日本人氏名[3]を用いて，Google Web APIs でそれぞれ 50 件の検索を行い，検索結果から同姓同名の人物を手動で分離した 80 人分の Web ページの HTML ファイル[2]をデータセットとする．

著者らが人物につき一つの正解データを与え，正解データと付与された NDLSH が一致するか評価を行った．正解数は自動的に付与された NDLSH の中に正解がある人物数，正解率は 80 人中の正解数の割合，最上位正解数は自動的に付与された最上位の NDLSH が正しい人物数である．

4.2 結果と考察

正解数の上位5手法の結果を表2に示す．

表2　上位5手法の結果

ランク	手法	正解数	正解率	最上位正解数
1	ChatGPT	42	0.525	27
2	SBERT (N3×P5)	14	0.175	3
3	SBERT (N3×P7)	12	0.150	4
3	SBERT (N3×P4)	12	0.150	4
5	SBERT (N3×P6)	11	0.138	5

正解数を用いた評価では ChatGPT を用いた手法が最も良いことがわかった． SBERT を用いた手法の中では，NDLSH のベクトルは**(N3)**「(N-1)と(N-2)のベクトルの平均」を用いる方法が良かった．人物ベクトルは最上位ページだけでなく全ページを用いる方が良かった．

ただし，人物に NDLSH を付与する場合，件数が少なすぎると人物の理解や選択が難しい．ChatGPT を用いた手法では，人物に付与された件数は 1 人当たり平均 7.03 件であった．最少で1件，最多で16件付与された(SD=3.76)．

表3に上位2手法である ChatGPT 手法と SBERT 手法(N3×P5)で料理家の栗原はるみ氏に付与した結果の例を示す．正解は「料理」であり，ChatGPT 手法でのみ正解を得ているが，SBERT 手法では「フードコーディネーター」など，人物の理解に有用な料理に関する NDLSH が ChatGPT 手法より多く付与されている．

表3　ChatGPT と手法と SBERT 手法の比較

	ChatGPT 手法	SBERT (N3×P5)手法
1	サイ	精進料理
2	商品	懐石料理
3	食器	日本料理
4	販売	日本料理店
5	**料理**	お好み焼き
6		オーガニックフード
7		フードコーディネーター
8		カステラ
9		食事作法
10		うま味調味料

5.　関連研究

文書に統制語を付与する方法は機械学習を用いるものと用いないものに大別される．下倉ら[2,4]は機械学習を用いていないが本研究は機械学習の中でも深層学習を用いた手法を検討した．

6.　おわりに

Web 上の人物に SBERT と ChatGPT を用いて NDLSH を付与する手法を検討した．正解数による評価では ChatGPT を用いた手法の性能が最も良かったが付与件数が少ないことがある．SBERT を用いた手法は性能は劣るが人物の理解に有用な NDLSH を多く得られることがある．今後の課題としては，手法の改良を行うとともに，最上位以外の結果の評価，データセットの追加と評価実験などがあげられる．

謝辞

本研究は JSPS 科研費 19K12718 の助成を受けたものです．

参考文献

[1] 村上晴美，上田洋：Web 人名検結果の要約と可視化を目指して：2009 年度人工知能学会全国大会（第 23 回）論文集（2009）

[2] 下倉雅行，村上晴美：Web 上の人物への NDLSH の付与：2017 年度人工知能学会全国大会（第 31 回）論文集（2017）

[3]佐藤進也，風間一洋，福田健介，村上健一郎：実世界指向 Web マイニングによる同姓同名人物の分離：情報処理学会論文誌，データベース，Vol. 46，pp. 26-36（2005）

[4] 下倉雅行, 村上晴美: Web 上の人物への NDLSH の付与 - 2 種類のデータセットを用いた評価 -,情報処理学会第 86 回全国大会講演論文集 (2024)