

大規模言語モデルを活用したアルファベット表記とカナ表記間の変換手法

廣森 聡仁¹⁾, 鎗水 徹¹

1) 大阪大学 OUDX 推進室

hiromori.akihiro.oict@osaka-u.ac.jp

Mapping Strategies for Japanese Kana and English Names Using LLMs

Akihiro Hiromori¹⁾, Toru Yarimizu¹

1) OUDX Promotion Unit, Osaka University

概要

組織内に蓄積されたデータの統合と分析の重要性が高まっており、人名データに対するクレンジング処理は重要な課題の一つとなっている。特に、外国人名のアルファベット表記とカナ表記の相互変換は、言語間の発音差異や多様な表記方法のため簡単ではない。本取組では、大規模言語モデル (LLM) を活用し、人名に対するアルファベット表記とカナ表記を変換する手法を提案する。表記の変換に際し、直接 LLM に問い合わせるのではなく、Wikipedia に掲載された人物の記事に基づき、アルファベット表記とカナ表記の対応関係を抽出し、これに基づき CNN による変換モデルを構築することで、9 割程度の精度で、アルファベット表記とカナ表記の対応関係を導出できることを示した。

1 はじめに

近年、組織におけるデータ活用の重要性が一層増しており、各種システムに蓄積されたデータを効果的に統合及び分析することが求められており、そのための基盤として、ETL (Extract, Transform, Load) 処理は不可欠なものとなっている。抽出 (Extract) の段階では、組織の各種システムや外部のデータソースから必要なデータを収集し、複数の異なるフォーマットや構造のデータを取り扱うために、高度な抽出技術が求められる。変換 (Transform) の段階では、データの品質と一貫性を確保することを目的とし、抽出したデータを統一された形式や構造に整え、具体的には、データのクレンジングや不整合の修正、データ型の統一、集計や計算などを施し、ロード (Load) の段階では、変換後のデータをデータウェアハウスに格納する。この ETL 処理において、取扱いが難しいデータのの一つとして、人名データが挙げられ、漢字・カナ・アルファベット表記の扱い、全角及び半角の統一など、人名データに対する特有のクレンジング処理が必要不可欠である。特に、外国人名のアルファベット表記とカナ表記の相互変換は、言語間の発音差異や表記の多様性から一貫性のある変換が難しく、その結果、あるアルファベット表記に対応するカナ表記が複数存在し、データ

の一貫性や検索性に影響を及ぼすことが懸念される。また、アルファベット表記にはアクセント記号や特殊文字が含まれることがあり、文字コードの制限やフォントの問題によりデータが正確に保存されないことも懸念される。学生の国際化が一層すすむ大学においても、外国人に関する情報を適切に扱うことが求められ、ある人名のアルファベット表記とカナ表記に対し、正確な相互変換を行うためには、様々な国における人名を網羅した変換ルールが必要となるが、そのような変換ルールは膨大なものとなることに加え、継続的なメンテナンスは簡単なものではないが見込まれる。

本取組では、この課題を解決するために、大規模言語モデル (LLM) を活用したアルファベット表記とカナ表記間の変換手法を提案する。OpenAI による ChatGPT [1] や Anthropic による Claude [2] などに代表される LLM は多言語の理解に優れており、英語だけでなく様々な外国人名や多文化的な名前に対する処理が期待できるが、その利用に際し、クラウドとして提供される LLM に対し、組織で保持する個人データを第三者に渡すことは必ずしも好ましいことではない。また、モデルの学習には大量のテキストデータが必要とし、Meta 社の llama [3] のような、オープンソースのモデルによる推論にも大規模な GPU を必要とするため、個々の組織での LLM の運用は容易な

のではない。そこで、個別の人名に対してではなく、日本語版 Wikipedia に掲載されている人物の記事に対し、LLM による高度な自然言語処理能力により、アルファベット表記とカナ表記を抽出し、その対応関係を CNN により、アルファベット表記とカナ表記に特化したモデルとして構築し、これにより、人名のアルファベット表記とカナ表記間の変換を実現する。このモデルに対する評価実験を実施した結果、9 割程度の精度で、アルファベット表記とカナ表記の対応関係を導出できることを示した。

2 関連研究

アルファベット表記とカナ表記の対応について、これまでにも多数の取組が行われてきており、文献 [4] においては、スペイン語によるアルファベット表記とカナ表記との対応について、スペイン語の文字と音声の対応に基づき、子音や母音の表記、長音や促音、撥音の対応方法が提案されており、例えば、カ行は「ca, qui, cu, que, co」、ハ行は「ja, ji, ju, je, jo」と対応するなど、スペイン語に馴染みのある形式を使用されている。また、一つの名前に対して複数のカタカナ表記が存在するため、すべての表記を網羅した辞書を作成するのは困難であることから、アルファベット表記とカタカナ表記の間の対応規則を自動的に生成する方法も提案されている [5]。この手法では、与えられた表記を分割し、長い表記と短い表記に基づいた新たな規則を生成し、母音や子音の対応に関する知識を最小限に抑えることで、効率的なルール生成を実現している。一方、検索エンジンを利用して人名からカタカナの読みを自動的に取得する手法が提案されている [6]。従来の人名辞書や漢字辞書では、異なる読み方を持つ人物や、著名でない人物の名前の読みを適切にカバーできないことが課題となっているが、本手法では、検索エンジンを用いて人名に対するウェブ上の情報を収集し、パターンマッチングとフィルタリングを通して人名の読みを抽出し、複数の読み候補が得られた場合は、職業や所属といった属性情報を使って同姓同名の人物を区別し、それぞれに適した読みを付与することで、約 3 万件のデータに対し、74.6% の適合率、80.9% の再現率を達成することを示している。

また、人名に対する処理に機械学習を活用する取組も実施されており、[9] においては、読みにくい人名（いわゆる「キラキラネーム」）の言語的特徴を分析し、これを基にキラキラネームを自動判定する手法を提案している。キラキラネームに共通するいくつかの

特徴を用いてサポートベクターマシン (SVM) を使用した判定モデルを構築し、1 万件の名前を対象に精度 81.79%、再現率 91.84% という結果を示している。また、外国人名のカタカナ表記を自動的に推定する取組も実施されている [8]。リオ・オリンピックでは、英語アルファベットで提供される外国人の名前を日本語のカタカナに変換する際、専用のトランスリタレータを構築していたが、2020 年の東京オリンピックに向けて、国ごとの特性に合わせた学習を行うことで、カタカナ表記の精度を向上できることを示した。同様に、国籍情報を活用した人名のカタカナ表記自動生成手法 [7] が提案されており、この取組では、RNN への入力として、国籍情報を加えることで、言語ごとに異なる発音を適切に区別することで、従来の機械翻訳手法や RNN 単体の手法よりも精度が向上し、国によって異なる発音を正確にカタカナに変換できることを示している。

3 提案手法

本取組では、(1) アルファベット表記とカナ表記の対応を導出し、(2) その対応に基づき、CNN モデルを構築することで、様々な国の人名に対するアルファベット表記とカナ表記の対応関係を把握する。

まず、(1) アルファベット表記とカナ表記の対応においては、著名な人物に関する記事を多数掲載している Wikipedia の記事から、クラウドによる LLM を活用し、アルファベット表記とカナ表記の対応関係を抽出する。Wikipedia では、保持する記事のデータベースを提供しており、このデータベースに含まれる記事を段階的に処理していき、アルファベット表記とカナ表記の対応を抽出している。本取組では、2024 年 3 月 20 日時点のデータベースが保持する約 330 万件の記事に対し、GPT4-o mini を利用し、タイトルのみに基づき、その記事が人に関わるものであるかを判定し、その結果、約 30 万件の記事を人に関わるものとして判定された。それらの記事に対し、下記のように、タイトルだけでなく、Wikipedia の記事内容もプロンプトに加え、アルファベット表記とカナ表記を導出するよう、GPT4-o mini を利用し、約 20 万件の記事に対し、アルファベット表記とカナ表記の対応を得た。ここで対応を得られなかった記事は、そもそも人名の記事ではないものや、姓、名、ミドルネームの区切りが曖昧であり、LLM により個々の表記が得られなかった記事などが含まれる。

Please determine if "{word}" is a person.
If it is a person, please provide the kana
name and the english name base on the
following html string.

\HTML String:

{wiki}

(2) CNN モデルの構築においては、上記の手続きで得られた対応関係に基づき、アルファベット表記からカナ表記への変換を行うために、CNN モデルを構築した。まず、約 20 万件の記事に紐づく人名に紐づくアルファベット表記とカナ表記について、姓・名・ミドルネームに分割し、個々の単語に紐づく、アルファベット表記とカナ表記を学習用データセットとして用意した。8 割程度の単語については、対応する表記が一意に定まっているが、例えば、Andrew に対しては、'アンドレ'、'アンディ'、'アンドルー'、'アンディー'、'アンドラウ'、'アンデレ'、'アンドリュウ' など、Vladimir に対しては、'ブラディミール'、'ブラジミル'、'ブラジーミル'、'ヴラディミア'、'ヴラジミール'、'ウラディミール' など、多様な表記が存在し、単純な変換が難しいことがうかがえた。このデータセットを訓練用 (60%)、検証用 (20%)、テスト用 (20%) に分割し、あるアルファベット表記に対応するカナ表記を導出するモデルを CNN により構築し、CTC 損失 (Connectionist Temporal Classification Loss) に基づき、モデルを学習した。その結果、データセットに含まれるテスト用データに対し、約 9 割程度の正解率を達成した一方、人名からだけでは正しく表記を導出できなかったものも多かった。文献 [8, 7] に指摘されているように、アルファベット表記からカナ表記を導出するというタスクに対し、人名だけでは不十分であり、国籍に代表されるように、個人に紐づく他の情報が必要となることが伺えた。

4 まとめと今後の課題

本取組では、大規模言語モデル (LLM) を活用し、人名に対するアルファベット表記とカナ表記を変換する手法を提案した。この手法では、Wikipedia に掲載された人物の記事に基づき、アルファベット表記とカナ表記の対応関係を抽出し、これに基づき CNN による変換モデルを構築することで、9 割程度の精度で、アルファベット表記とカナ表記の対応関係を導出できることを示した。

今後の展望として、学習データセットとして、人名だけでなく、国籍情報を含む形で Wikipedia の記事を学習することや、Transformer に基づくアテンション機構を備えた Seq2Seq モデルに基づくモデル構築に取り組み、推定精度の向上を目指す。一方で、データセットとして利用する日本語版 Wikipedia の記事は、組織内の人のデータとの偏りがあることが懸念され、特定の言語や文化圏の名前に対する精度は必ずしも高くないことが想定される。Wikipedia の他の言語の記事を含め、データセットを拡充するなど、多言語対応を一層強化し、モデルの公平性と汎用性の向上を図っていきたい。

参考文献

- [1] OpenAI, GPT-4 Technical Report, <https://cdn.openai.com/papers/gpt-4.pdf>, 2023.
- [2] anthropic, Introducing the next generation of Claude, <https://www.anthropic.com/news/claude-3-family>, 2024.
- [3] Meta Llama team, The Llama 3 Herd of Models, <https://ai.meta.com/research/publications/the-llama-3-herd-of-models/>, 2024.
- [4] 野田 尚史, 高澤 美由紀, スペイン語アルファベットによる日本語音声表記, 国立国語研究所論集, Vol.19, pp.139-166, 2020.
- [5] 尾上 徹, 梅村 恭司, 岡部 正幸, アルファベット表記とカタカナ表記の対応規則の生成, 情報処理学会プログラミング・シンポジウム予稿集, Vol.52, pp.11-20, 2011.
- [6] *酒巻 智宏, 大向 一輝, 丹 英之, 武田 英明, 検索エンジンを用いた人名読みの推定, 2010 年度人工知能学会全国大会 (第 24 回), Vol.24, 2C2-4, 2010.
- [7] 宮崎 太郎, 熊野 正, 今井 篤, 国籍情報を用いた人名の音訳, FIT2016 (第 15 回情報科学技術フォーラム), E-018, 2016.
- [8] 安江 祐貴 佐藤 理史 松崎 拓也, 外国人名のカタカナ表記自動推定システムの改良, 言語処理学会 第 23 回年次大会 発表論文集, Vol.19, pp.139-166, 2017.
- [9] 山西 良典, 大泉 順平, 西原 陽子, 福本 淳一, 人名の言語的特徴の分析に基づくキラキラネーム判定, 日本感性工学会論文誌, Vol.15, pp.31-37, 2016.