# 表記ゆれが文埋め込みモデルに及ぼす影響についての考察

佐々木峻 [1] 山本大輝 [1]
[1] アクロクエストテクノロジー株式会社
{sasaki, yamamoto}@acroquest.co.jp

## 概要

近年の RAG (Retrieval-Augmented Generation) や検索システムでは，文埋め込みを用いた検索手法が注目されている．従来はキーワード一致に基づく TF-IDF ベースの検索が主流であったが，大規模事前学習モデルの発達に伴い，文埋め込みを用いた検索が盛んに使用されるようになった．しかしながら，文埋め込みがキーワードベースの検索に対して，どのような利点や欠点を持つかはまだ十分に明らかになっていない．特に日本語検索においては，多様な表記ゆれが検索精度に影響を与える一因となる．そこで本研究では，単語レベルの表記ゆれをクエリに与えた際，どのように文埋め込みベースの検索精度が変化するのかを検証する．実験では，JMTEB ベンチマークのうち JAQKET および JaGovFaqs-22k を用いて，同義語辞書によるクエリ変換を実施し、変換前後のクエリを使用した場合の検索精度への影響を分析した．さらに，文埋め込みに同義語を拡張したクエリ手法も提案し，その有効性を検証する．

## 1 はじめに

近年，大規模言語モデルの発展とともに RAG（Retrieval-Augmented Generation）システムが増え，文埋め込みベースの検索技術が急速に注目されている．従来のキーワードマッチングによる検索 (TF-IDF 等) は簡便でありながらも，単語表記のわずかな違いに対するロバスト性が弱いという課題があった．一方，文埋め込みを用いた検索では，埋め込み空間で類似性を測ることで，単純な文字列一致にとらわれず関連文書を取得できる強みがあると期待されている．しかしながら，文埋め込みモデルが実際にキーワード検索に対してどの程度優位性を持ち，特に日本語特有の表記ゆれに対してどう影響を受けるのかは明確になっていない．

本論文では，日本語検索における表記ゆれの一形態である同義語や送り仮名の違いなどに注目し，クエリ単語を同義語に変換した場合の検索精度を評価する．実験では，JMTEB[5] に含まれる QA データセット（JAQKET, JaGovFaqs-22k）を用いて，クエリ内単語を同義語辞書に基づき変換した後の検索精度を比較する．さらに，表記ゆれを抑制するために，クエリ内単語を拡張する手法（「同義語拡張クエリ」）を導入し，表記ゆれの影響をどの程度低減できるかを議論する．

- 「サンプル」 vs 「サンプル（半角）」
- 「東京大学」 vs 「とうきょう大学」
- 「申し込み」 vs 「申込み」
- 「不具合」 vs 「バグ」 (同義語)

**図 1** 日本語における表記ゆれの例

本稿の構成は以下の通りである．まず 2 節でキーワード検索における課題を整理し，3 節で関連研究を概説する．続いて 4 節で実験設定および実験方法を述べ，5 節で表記ゆれによる性能変化と，その抑制手法の検証結果を示す．最後に 6 節で結論と今後の展望を述べる．

## 2 キーワード検索における課題

日本語のキーワード検索では，全角・半角や送り仮名の違い，同義語の混在など「表記ゆれ」がしばしば問題となる図 1 に表記ゆれの例を示す．

全角・半角や送り仮名のゆれは，検索エンジンや形態素解析ライブラリの前処理によってある程度吸収することが可能な場合が多い．しかし，同義語や言い換え表現などは，事前に辞書を整備しキーワードクエリの拡張を行わない限り正しくヒットしない場合がある．これら表記ゆれを手動で整理することはコストが高く，また管理や更新を行うのも容易ではない．そのため，表記ゆれに対してある程度ロバストな検索手法のニーズが高まっている．文埋め込み手法は，キーワードが多少異なっていても同じ文脈を共有できる可能性があるため，その有効性が期待される．

# 3 関連研究

## 3.1 文埋め込みモデル

近年の文埋め込みモデルは，Transformer ベースの事前学習モデルを活用し，対照学習 (contrastive learning) による大規模事前学習を施した後，高品質なデータセットで微調整 (fine-tuning) するという流れが主流である．多言語対応モデルである multilingual-e5 ファミリーは，Wikipedia や CC News などの大規模コーパスを使い，NLI や MS MARCO といったデータセットで微調整を行うことで，多言語の文類似度推定や検索タスクで高い性能を示す [1].

一方，日本語特化のモデルとして名古屋大学が開発した ruri モデル群 (ruri-base, ruri-large など) がある．[2]．これらは日本語のコーパスを主体に学習されており，日本語の細やかな表現やドメインへの適応力が期待される．本研究ではこの 2 種類のファミリー (multilingual-e5, ruri) のモデルを使用し，それぞれ base および large を比較対象とする．

## 3.2 文埋め込みモデル評価

文埋め込みモデルを評価する研究としては Ada と BERT のベクトル検索時の性能について言及した馬らが実施した検証 [3] がある. この検証では BERT を使用した文埋め込みと OpenAI 社が提供している ada-002 を使用した文埋め込みによる検索結果の違いについて比較している. 彼らの研究では、まず旅行領域に特化した BERT モデル（J-BERT）と SentenceBERT（S-BERT）を構築し、宿の口コミデータを用いて事前学習と微調整を行った。一方、ada-002 は汎用的な文埋め込みモデルであり、特定ドメインへの微調整は行われていない. 評価のため、クエリを「視点の広さ」「知識量の多さ」「逸脱表現に対する受容性」の 3 つの尺度で分類し、それぞれのモデルの性能を分析した。その結果、ada-002 は以下の点で劣ることが示された.

1. **視点の広さ**: 長いクエリに含まれる複数の要素を適切に解釈し、検索結果に反映する能力が限定的である.
2. **知識量の多さ**: 旅行ドメインに特化した知識が不足しており、専門用語や地理的情報の理解に課題がある.

一方、ada-002 は以下の点で優れていることが確認された.

1. **逸脱表現に対する受容性**: 非定型的な日本語表現や顔文字、絵文字、英単語を含むクエリに対しても柔軟に対応し、適切な検索結果を提供する能力が高い。

このことから文埋め込みには学習元になるデータや、Fine Tuning 時の手法によって精度に影響を出す側面が変化することがわかる. そのため、本論文で検証する際には日本語に特化したモデルと多言語に対応したモデルを比較することで、対応の変化を検証する. また、矢野らの研究 [4] では、クエリの形式による埋め込み結果の検索精度の違いについて考察されている。文埋め込みモデルは文章を訓練データとしているため、キーワードの羅列よりも文章から埋め込みを生成することで、検索精度が向上することが示されている. このことから埋め込みモデルは入力形式の変化に対してはロバスト性が高くない可能性が高いことが示されている.

# 4 実験

## 4.1 使用するデータセット

本研究では JMTEB に含まれる日本語 QA データセットのうち，**JAQKET** および **JaGovFaqs-22k** を用いる.

**JAQKET**　クイズ形式の QA を題材にしたデータセットであり，クエリ（クイズ文）と Wikipedia 記事が紐づいている．本研究では JMTEB の検索タスク向けの Validation 用データをそのまま使用する.

**JaGovFaqs-22k**　官公庁サイトから取得されたよくある質問 (FAQ) をベースにした Q-A ペアを集約したデータセットである．官公庁の FAQ は，言葉遣いが比較的定型的であるため、表記ゆれの影響を調べるうえで有用である。

## 4.2 使用する文埋め込みモデル

次に，検索タスク用の文埋め込みモデルとして，以下の 4 モデルを使用する.

1. **multilingual-e5-base**
2. **multilingual-e5-large**
3. **ruri-base**
4. **ruri-large**

multilingual-e5 は多言語大規模コーパスで対照学習

されたモデルで，日本語を含む複数言語を扱うにもかかわらず高い精度を示している．ruri ファミリーは名古屋大学で作成された日本語特化モデルであり，特に日本語の多様な文体に強い可能性があると考えられる．

### 4.3 表記ゆれ検証用データ作成

検証のため，まず上記データセットに含まれるクエリ群に対して，**同義語辞書**を使い単語を自動変換したクエリを作成する．ここでは Sudachi[6] で公開されている同義語辞書を利用し，クエリ中の単語を辞書に存在する語彙があればランダムに同義語に書き換える．書き換えた後のクエリを元のクエリと区別し，

$$Q_{\text{org}} = \{q_{\text{org},1}, q_{\text{org},2}, \ldots\}, \quad Q_{\text{syn}} = \{q_{\text{syn},1}, q_{\text{syn},2}, \ldots\}$$

と定義する．なお，同義語が一切見つからないクエリは除外する．

### 4.4 同義語による影響の計測

同義語変換前後のクエリ群 $Q_{\text{org}}$ と $Q_{\text{syn}}$ を文埋め込みモデルでベクトル化したうえで，コサイン類似度に基づき Top-$k$ (ここでは $k = 5$) 件を文書コーパスから取得する検索を行う．

このとき，クエリ $q_{\text{org},i}$ の正解ドキュメントが Top-$k$ に含まれるインデックス集合を $I_{\text{org}}$，変換後 $q_{\text{syn},i}$ の正解が Top-$k$ に含まれるインデックス集合を $I_{\text{syn}}$ とし，それぞれのサイズを集計する．特に，両方で Top-$k$ に正解を含むクエリ数 $|I_{\text{org}} \cap I_{\text{syn}}|$（和 (AND)）と，片方のみ正解するクエリ数 $|I_{\text{org}} \oplus I_{\text{syn}}|$ (排他的論理和 (XOR)) を調べることで，

$$\frac{|I_{\text{org}} \oplus I_{\text{syn}}|}{|I_{\text{org}} \cap I_{\text{syn}}|}$$

がどの程度になるかを指標とする．この値が大きいほど，同義語変換によって検索結果が大きく変化していると言える．

## 5 結果

表 1 に JAQKET，JaGovFaqs-22k を対象にした実験結果の例を示す．

上記より，以下の傾向が見られる．

1. どのモデルでも，同義語の変換による「ヒットのずれ」は一定程度（5%から 15%程度）存在する．
2. base モデルより large モデルの方が XOR 値が低く，表記ゆれの影響を受けにくい．すなわち，検索精度が高いモデルほど表記ゆれに対して頑健である可能性がある．

**表 1** 表記ゆれ（同義語変換）による検索精度影響度 (Top-5 での正解 hit 数)

| データセット | モデル | XOR | AND | XOR/AND |
|---|---|---|---|---|
| JAQKET | e5-base | 67 | 440 | 0.1321 |
| | e5-large | 34 | 557 | 0.0575 |
| | ruri-base | 70 | 428 | 0.1406 |
| | ruri-large | 44 | 530 | 0.0767 |
| JaGovFaqs-22k | e5-base | 177 | 1817 | 0.0888 |
| | e5-large | 135 | 1976 | 0.0640 |
| | ruri-base | 202 | 1698 | 0.1063 |
| | ruri-large | 182 | 1813 | 0.0912 |

**表 2** 同義語拡張クエリによる Top-5 ヒット数の変化例 (AND に着目)

| データセット | モデル | AND(通常) | AND(拡張) |
|---|---|---|---|
| JAQKET | e5-base | 440 | **467** |
| | e5-large | 557 | **568** |
| | ruri-base | 428 | **450** |
| | ruri-large | 530 | **545** |
| JaGovFaqs-22k | e5-base | 1817 | **1899** |
| | e5-large | 1976 | **2033** |
| | ruri-base | 1698 | **1860** |
| | ruri-large | 1813 | **1946** |

日本語に特化した ruri モデルでは，英語コーパスを含む multilingual-e5 よりも，クエリ文の文脈をより正確に捉えられると考えられるが，今回の結果では表記ゆれによる影響度はむしろやや高く，特に base モデルで顕著だった．これは学習データにおける「同義表現」の多様性や，辞書のカバー範囲などが影響している可能性がある．

### 5.1 文埋め込み時の同義語拡張クエリの検証

同義語による性能変動が確認されたため，対策として「同義語をクエリ内に併記する」方式（同義語拡張クエリ）を試みた．具体的には，該当する単語を{元単語|同義語}という形で置き換える（例：「申し込み」→「{申し込み|申込み}」）．このようにクエリ側を拡張し，埋め込みモデルでベクトル化したときに，いずれの表現に対しても埋め込み空間で考慮されるように期待できる．

表 2 に同義語拡張クエリを用いた際の Top-5 ヒット数 (AND) の変化例を示す．

拡張後のクエリを用いると，どのモデルに対しても AND(変換前後いずれでも正解が Top-5 に含まれるクエリ数) が増加しており，表記ゆれの影響を緩和できることが示唆される．ただし，拡張クエリに

よって文字列長が増えるため，モデルのトークナイズで予期しない切り方をされる可能性もある．実運用では同義語辞書の整備が必要であるうえ，クエリが煩雑になるデメリットもあるため，適用範囲やコストとのトレードオフを慎重に検討する必要がある．

## 6 まとめ

本研究では，日本語の文埋め込み検索において，クエリの単語表記ゆれ，特に同義語変換が検索精度に与える影響を分析した．JAQKET，JaGovFaqs-22kを用いた実験から，どのモデルでも表記ゆれに対して一定程度の性能変動が生じる一方，モデルのサイズが大きくなるほどロバスト性が増す傾向にあることがわかった．さらに，同義語をクエリ上で併記する拡張手法により，ヒット数を増加させられる可能性が示唆された．

今後は，本研究で用いた同義語辞書以外の表記ゆれ（送り仮名揺れや外来語カタカナ表記揺れなど）に対しても同様の検証を行う予定である．また，検索の上流タスクである質問生成や多段階検索との組み合わせで，文埋め込みとキーワード検索をハイブリッドに組み合わせたアプローチについても検討していきたい．

## 参考文献


[1] “Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, Furu Wei” **Multilingual E5 Text Embeddings: A Technical Report**.

[2] “Hayato Tsukagoshi, Ryohei Sasano” **Ruri: Japanese General Text Embeddings**.

[3] “馬春鵬, 松田寛” **Ada or Bert: 検索における文埋め込み計算手法の比較研究**.

[4] “矢野 千紘 塚越 駿 笹野 遼平 武田 浩一” **日本語文埋め込みの文書検索性能と検索補助付き生成での評価**.

[5] “JMTEB: Japanese Massive Text Embedding Benchmark” https://huggingface.co/datasets/sbintuitions/JMTEB.

[6] Sudachi Project: https://github.com/WorksApplications/Sudachi (accessed 2024-12-30).