特定ドメイン向けローカル検索用のサジェスト提示に向けた分析

鈴木 琴音[1]　岩本 和真[2]　安藤 一秋[1]
[1]香川大学創造工学部　[2]香川大学大学院創発科学研究科
{s21t343, s24g351, ando.kazuaki}@kagawa-u.ac.jp

概要

検索システムの支援の 1 つであるサジェスト機能は, 膨大な検索ログを利用して実装されているため, ローカル文書を対象とした検索システムに転用できない．この課題を解決するために，本研究では，検索対象文書内の単語類似性に基づくサジェスト機能について検討する．本稿では，事前分析として，Wikipedia データで学習された Word2Vec モデルと，検索対象文書集合の特定ドメインデータで追加学習した Word2Vec モデルを用いて，特定ドメインデータの追加学習による単語ベクトルの分布や類似性の変化を定量・定性的に分析する．分析の結果，追加学習したモデルは，特定ドメインにおける語彙の関係性を捉えられており，検索対象文書のドメイン特性を活かしたサジェスト機能の実現可能性を確認した．

1 はじめに

検索システムには, Google や Mirosoft Bing のように大規模でパブリックな文書集合を対象とするものと，特定の組織や部署，個人のローカル環境などに保存された文書を対象とするものがある．前者と後者では, 利用するユーザ数に圧倒的な差が存在する．また, 生成 AI の発展に伴い，ローカルな文書集合を外部情報として活用する Retrieval-Augmented Generation（RAG）[1]が注目されており，近年は，ローカル文書検索の重要性が増加している．

検索システムを利用する上で，ユーザが求める情報を的確に取得するには，適切なクエリを入力する必要がある．しかし，クエリの表現方法は多様であり，検索タスクへの慣れやドメイン知識の有無・深さなどによって，ユーザの入力するクエリが必ずしも求める文書に適しているものとは限らない．この問題に対して，検索システムは，適切なサジェストを提供することで, ユーザのクエリ選択を支援する．ユーザ数が膨大な検索システムのサジェスト機能は, 膨大な検索ログを活用することで実現されている．一方，大規模な検索ログの収集が期待できない検索システムには，同様な手法に基づくサジェスト機能が実装できない．

この課題を解決するため，本研究では，検索対象文書集合内の単語類似性に基づくサジェスト機能について検討する．ユーザの検索ログではなく，検索文書に焦点を当てることで，大規模な検索ログの収集が困難な検索システムにおけるサジェスト機能の実現が期待できる．ローカル文書検索を対象とした場合，検索対象文書の数が限られているため，検索対象文書を活用するコストは大きくないといえる．

本稿では，ローカル文書集合として，学術論文集合を想定する．そして，ローカル文書検索におけるサジェスト機能の検討に向け，汎用的な単語埋め込みモデルである Word2Vec モデル[i]を，学術的な専門性を含む人工知能学会全国大会[ii]（JSAI）の論文概要データで追加学習した Word2Vec モデルを用いて，特定ドメインデータの追加学習による単語ベクトルの分布や単語間類似性の変化を定量・定性的に分析する．

2 関連研究

検索クエリの改善やクエリ拡張，サジェストに関する研究は，いくつか先行事例がある[2, 3, 4, 5, 6].

鹿島ら[5]は，検索クエリの精度向上を目的として，Word2Vec を用いた検索クエリ置換手法を提案した．鹿島らの手法では，クエリの類似語や関連語を生成し，元のクエリを置換することで検索結果を改善する．実験の結果，元のクエリで検索した結果よりも精度が向上し，特に曖昧な単語や同義語が含まれるクエリで顕著な効果を発揮したと述べている．

出永ら[6]は，論文検索を対象に，適合性フィードバックを活用してクエリを拡張する手法を提案した．ユーザからの適合性フィードバックに基づき，適合

[i] https://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

[ii] https://www.ai-gakkai.or.jp/

文書では高頻度，非適合文書では低頻度の単語を重要語としてスコア付けし，拡張クエリ候補としてランキングしてユーザに提示する．実験の結果，適合性フィードバックを用いたクエリ拡張により検索性能が向上したと述べている．

3 単語間類似度の変化の分析

検索対象文書内の単語類似性に基づくサジェスト機能を検討する準備として，Word2Vec と，3 年分の JSAI 論文概要（1,152 件）で Word2Vec を追加学習したモデルを用いて，特定ドメインデータの追加学習による単語間類似度の変化について，定量・定性的に分析する．

3.1 サジェストデータの構築

論文検索する際，実際に利用されるクエリを収集して分析に利用する．以下，分析で用いるクエリデータの構築手順を説明する．

自然言語処理に関する研究に従事している学生 10 名（学部 4 年生 3 名，大学院生 7 名）に対して，「自然言語処理」と「BERT」を第一クエリとして提示する．10 名には，自身の研究テーマの関連論文を検索する状況を想定してもらったうえで，第二クエリとして入力する単語を回答してもらい，第一クエリと第二クエリのペアを収集する．なお，第二クエリは，Word2Vec のトークンに分割して利用する．そして，第二クエリを第一クエリに対するサジェストと見なし，以降，第一クエリに対する正解サジェストとして利用する．

収集後，トークン分割した結果，「自然言語処理」に対するサジェストとして 32 語，「BERT」に対するサジェストとして 37 語が得られた．表 1 に収集したサジェストの例を示す．表 1 より，「対話コーパス」や「系列長」など，自然言語処理に関連する単語が確認できる一方，「うつ病」や「公衆衛生」など一般的に使われる単語なども確認できる．このことから，学術論文をローカル文書として想定した場合においても，多様な単語がサジェストになりえることがわかる．

3.2 追加学習による単語間類似度の変化に注目した分析

Word2Vec（ベースモデル）と 3 年分の JSAI 論文概要（1,152 件）で Word2Vec を追加学習したモデル

表 1：収集したサジェストデータの例

自然言語処理	BERT
対話コーパス	雑談対話
記事推薦	記事推薦
アスリート	誹謗中傷
公衆衛生	系列長
Word2Vec	感情分析
うつ病	うつ病
検知	ファインチューニング

（追加学習モデル）を用いて，それぞれの第一クエリに対するサジェスト間の類似度ランキングの変化を確認する．評価には，次式の MRR（Mean Reciprocal Rank）を用いる．

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{Rank_i} \quad (1)$$

ここで，|Q|は，各第 1 クエリに対する正解サジェストの総数，$Rank_i$は，第一クエリと Word2Vec モデル内のすべての単語間のコサイン類似度を求め，類似度が高い順にランキング（類似度ランキング）し，正解サジェストが最初に出現した順位とする．つまり，Word2Vec モデル内のすべての単語をサジェスト候補と見なした場合，収集した正解サジェストが類似度ランキングの何位に出現するのかに注目して，分析する．また，収集した正解サジェストが必ずしも高い順位に集中してランキングされるとは限らない．そこで，第一クエリとサジェスト間の類似度範囲を確認するため，サジェスト候補として出力する類似度の上限閾値 α を設定し，α を変動させた際の MRR の推移を確認する．つまり，類似度が閾値 α 以上の各モデル内の単語をサジェス候補から除外して，閾値毎の MRR を算出する．上限閾値 α は，1.0～0.0 間を 0.05 間隔で減少させる．

図 1 は，閾値を変化させた際の第一クエリと Word2Vec 内の全単語とのコサイン類似度に基づく MRR の推移を示す．図 1 より，追加学習モデルの「BERT」は，他のモデルを大きく上回る MRR となった．特に，閾値 α が 0.7 付近のとき，MRR が最大で約 0.007 に達した．一方，「自然言語処理」は，閾値 α が 0.55 付近で MRR が最大となり，この範囲において一定の性能向上が確認できる．

しかし，ベースモデルについては，全体的に MRR が低く，閾値 α の変化による特徴は見られなかった．つまり，正解サジェストがサジェスト候補の中に埋

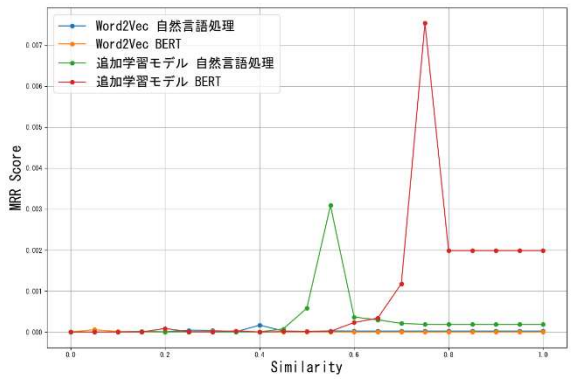

図 1：閾値を変化させた際の MRR の推移

もれてしまっていることを意味する．追加学習モデルは，特定ドメインデータで追加学習したことにより，ベースモデルよりドメインに関連する単語間の類似度が高くなるため，サジェスト性能が向上したと考えられる．

「自然言語処理」と「BERT」の MRR が最高値になる閾値が異なることから，閾値の設定がモデル性能に影響を与えるといえる．また，単語ごとに最適な閾値が異なる傾向がある，つまり，クエリごとに閾値を変更する必要があるため，単一の閾値では最適な結果を得ることは難しい可能性がある．

3.3 考察

ベースモデルと追加学習モデルの MRR の分布を比較した際，追加学習モデルには，ある一定の閾値によって MRR が急激に向上する現象が見られた．この要因を調査するため，まず，第一クエリとサジェスト間の類似性を比較し，特定ドメインデータの追加学習が単語間の関係性にどのような変化を与えるか調査する．

まず，第一クエリと各サジェスト間の類似度を測り，モデルごとの箱ひげ図で分布を確認する．ベースモデルと追加学習モデルによる類似度分布を図 2 に示す．図 2 より，ベースモデルでは「自然言語処理」の類似度が 0.2〜0.4 付近に，「BERT」は類似度が 0.15 から 0.3 付近に分布している．どちらも類似度は全体的に低い．一方，追加学習モデルでは，「自然言語処理」の類似度が 0.35〜0.6 付近に，「BERT」は類似度が 0.45〜0.8 に分布している．よって，特定ドメインデータの追加学習によって，特定ドメインに関連するクエリとサジェスト間の類似度が高くなったといえる．

次に，追加学習モデルにおいて，類似度が高い階級に集中する単語群が MRR を向上させる要因になっているかを確認するため，各モデルの類似度別の単語数の分布を比較する．

図 3 と図 4 は，第一クエリ「自然言語処理」，「BERT」と各モデル内のすべての単語間の単語類似度別の単語数分布を示す．図 3 より，ベースモデルの単語数分布において，最も単語数が多い類似度階級は 0.15，追加学習モデルでは 0.3 である．図 4 では，ベースモデルの単語数分布で最も単語が多い類似度階級は 0.2，追加学習モデルでは 0.35 である．よって，どちらのモデルにおいても単語数の分布は，低い類似度階級に集中している．

上記の分析に基づいて，追加学習モデルのほうが MRR の値が高く，追加学習モデルにのみ，閾値による MRR の変化が大きかった要因を考察すると，追加学習によって，特定ドメインに関係する単語間の類似度が高くなり，ほかの単語より関係性が強調できたと考えられる．追加学習が特定分野における単語間の関連性を強化する傾向を示したことから，検索対象の文書集合に関連するドメインデータでモデルを強化することで，検索ログに頼ることなく，有効なサジェストを提示できる可能性がある．

しかし，図 2 より，ベースモデルと比較して，追加学習モデルの四分位範囲が広くなっていることが確認できる．これは，類似度を用いてサジェストを選択する際，ノイズとなる語をサジェストする可能性が増えることを意味する．この課題を解決するには，単語ベクトルの類似度のみではなく，ユーザ情報といったほかの情報を組み合わせることで，個々に最適なサジェスト提示が可能になると考えられる．

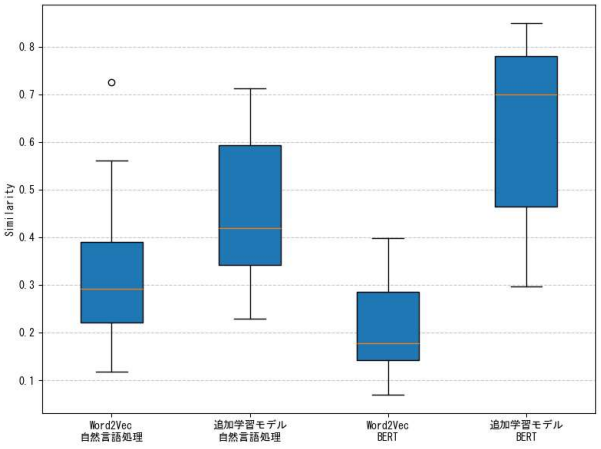

図 2: ベースモデルと追加学習モデルの箱ひげ図による類似度分布

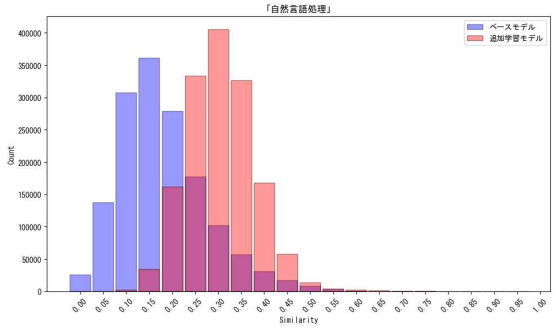

図 3:「自然言語処理」を対象とした
単語類似度階級における単語数分布

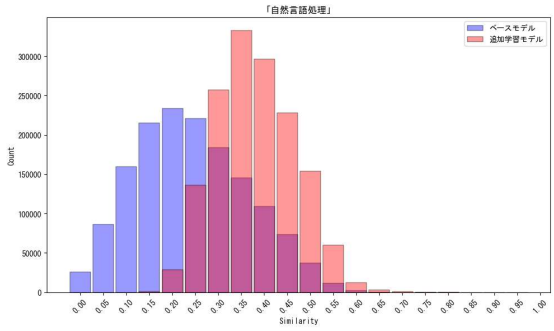

図 4:「BERT」を対象とした
単語類似度階級における単語数分布

4 おわりに

本稿では，ローカル文書検索におけるサジェスト機能の検討に向け，Wikipedia データで学習された Word2Vec モデルと，検索対象文書集合の特定ドメインデータで追加学習した Word2Vec モデルを用いて，特定ドメインデータの追加学習による単語ベクトルの分布や類似性の変化を定量・定性的に分析した．分析の結果，追加学習したドメインに関連するクエリとサジェスト間の関係性が向上することを確認した．これにより，ユーザの検索ログに依存せずに，ローカル環境で検索対象文書のドメイン特性を活かしたサジェスト機能を実現できる可能性が確認できた．

今後の課題として，単独ドメインではなく，複数ドメインのデータセットを活用した追加学習手法を検討する．また，追加学習による分布範囲の拡大により，ノイズ語がサジェストされる可能性もあるため，この影響を確認したうえで解決法を検討する．さらに，類似度閾値を動的に調整する手法やユーザ情報といったほかの情報を用いたサジェスト機能の組み合わせなどについても検討する．

参考文献

1. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. Proc. of the 34th International Conference on Neural Information Processing Systems (NIPS '20). pp.9459-9474. 2020.
2. 染谷 瑛進，加藤 誠．クエリ自動補完のための文書コレクションからのクエリログ生成．第 16 回データ工学と情報マネジメントに関するフォーラム．9 pages. 2024.
3. 田貝 奈央，加藤 誠．大規模言語モデルを用いたクエリ拡張による会話ログの検索．第 16 回データ工学と情報マネジメントに関するフォーラム．5 pages. 2024.
4. 堀 憲太郎，大石 哲也，峯 恒憲，長谷川 隆三，藤田 博，越村 三幸．Wikipedia からの拡張クエリ生成による Web 検索とその評価．第 20 回セマンティックウェブとオントロジー研究会．pp.13-1-13-8. 2009.
5. 鹿島 好央，北山 大輔．Word2Vec と Web 検索を用いた検索クエリ置換手法．DEIM Forum 2017 論文集, 8 pages, 2017.
6. 出永 悠真，福田 悟志，冨浦 洋一． 論文検索における適合性フィードバックを用いたクエリ拡張支援．FIT2019 論文集，第 2 分冊, pp.219-223, 2019.