特定文書のクエリ補完に向けたGoogleサジェストの分析
Analysis of Google Suggestions for Query Completion of Specific Documents

鈴木 琴音[1]　安藤 一秋[1]
K. Suzuki[1]　K. Ando[1]
(香川大学[1])

1. はじめに

Google や Microsoft Bing をはじめとする Web 検索システムでは，ユーザが検索単語を入力すると，ニーズやトレンドといった全体ユーザおよび個人の検索ログに応じた単語候補を提示する．この機能は，検索効率の向上には有用であるものの, Web のような多数のユーザが利用し, 膨大な検索ログが活用できる状況でしか利用できない[1, 2]．そこで，本研究では，社内文書を対象とする RAG（Retrieval-Augmented Generation）や PC 内の文書検索のような特定文書を検索する（クエリログが少ない）状況において，ユーザあるいはシステムに適したクエリを補完する技術の実現を目指す．本稿では，事前調査として，Google サジェストに注目し，サジェストされる単語の特徴，キーワードの関係について分析する．

2. 使用データ

Wikidata から Wikipedia の見出し語を抽出して，検索クエリに利用する．検索クエリを Google サジェストに入力し, サジェストされる単語をクエリあたり最大 10 件取得する. 以降, これをサジェスト語とよぶ. サジェスト語は, 検索クエリに対して 1 語以上提示される場合があるため，2 単語ペアとなるように分割して分析に利用する．

3. サジェスト語の分析

3.1 品詞に基づく分析

検索クエリには，Wikipedia の見出し語の id を昇順に並べた場合の上位 1,000 件を利用する．そして，サジェスト語を形態素解析器（Mecab+ipadic+NEologd）で単語分割し, サジェスト語の品詞の傾向を分析する．

結果を表 1 に示す．サジェスト語は，名詞が提示される割合が圧倒的に多いことを確認した．

表 1　サジェスト語の品詞分析の結果

サジェスト語	件数
1単語(名詞)	8,867
1単語(名詞以外)	87
複合語	1,078

次に，各品詞において一単語で提示されたサジェスト語の出現頻度をカウントして，品詞別の頻出単語を確認する．なお, 名詞以外の件数は少ないため，名詞以外に集約する．

表 2 に品詞別の頻出単語を示す. 検証に用いたデータには，作者名が多く含まれていたため，「漫画」のキーワードが上位になった．また,「英語」「意味」「違い」など，日常的に Web 検索する際によく使う単語も上位に存在している．1 単語の名詞以外のサジェスト語については, データ数が少ないため傾向が把握できないが疑問詞系の単語の存在を確認した．

表 2　品詞別の頻出単語

名詞		名詞以外	
単語	出現回数	単語	出現回数
漫画	260	なぜ	7
英語	204	面白い	7
現在	168	読み	5
書籍	122	ありがとう	4
意味	90	怖い	4
違い	75	かつ	4

3.2 単語間類似度に基づく分析

検索クエリとサジェスト語の単語間類似度に傾向があるのか調査する．Wikipedia の見出し語 10,000 語に対するサジェスト語 103,661 を対象に，東北大学の Word2Vec2019（skip-gram）を使用して，検索クエリとサジェスト語間の単語類似度を求める．

実験の結果，単語間類似度が 0.5 より低いペアは 95,262 件あり，そのうち 46,337 件は類似度が 0 であった．また，8,393 件は類似度が 0.5 以上となり，そのうち 341 件は類似度が 0.8 以上であった．以上の結果から, 検索クエリとサジェスト語間の単語類似度は全体的に低い傾向であることを確認した．

4. おわりに

本稿では，Google サジェストで補完される単語の特徴を品詞と単語感類似度の視点で分析した．サジェスト語は名詞が多く，検索クエリとサジェスト語間の単語類似度は低い傾向を確認した．今後は，分析を継続すると共に，クエリログが少ない状況でもクエリ補完できる技術について検討する．

参考文献

[1] 染谷他, “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,” DEIM2024, 9 pages, 2024.
[2] 貴船他, “就活生のための Twitter を利用した Web クエリ拡張手法の提案,” FSS2023, pp.656-661, 2023.