曖昧な情報要求の言語的表現の特徴に関する調査 : Q&A サイトへの投稿文を用いて

On the Features of Vague Queries and its Linguistic Characteristics

森高 楓[†]，松香 敏彦[†]
Kaede Moritaka, Toshihiko Matsuka

[†]千葉大学
Chiba University
22l1021u@student.gs.chiba-u.jp，matsuka@chiba-u.jp

概要

インターネット検索において，検索者が自身の必要とする情報やその入手方法を適切に把握していない(情報要求が曖昧な)場合がある．本研究では，Yahoo! 知恵袋に投稿された質問文を，質問者の情報要求が言語化されたものと捉えて分析し，曖昧な情報要求が言語化されたときの特徴を探索した．外部の web ページを参照する回答が寄せられた質問文中で出現頻度が高くなる語を発見したが，情報要求の曖昧さとの関係性は今後検討する必要がある．

キーワード：ウェブ情報検索, 情報要求

1. はじめに

インターネット上の情報量の増加により，人々が身近な問題解決に取り組む際に，ウェブ情報検索によって情報を集め利用することが一般的になった．その中で，ウェブ情報検索を行うユーザー自身でさえ，自分が必要としている情報を的確に言語化できないような状況や，検索システムに入力すべき適切なクエリが思いつかない状況下でもウェブ情報検索を行う場合がある。そのような状況下では，ユーザーは問題解決に役立つ情報を効率的に得られないことが指摘されている[1][2]．そのような検索に対する支援が求められるが，そのためには，支援の対象となるような曖昧な情報要求を検出する必要がある．そこで，本研究では，問題状況はある程度把握されているものの，その解決に必要な情報や，その情報を得るための方法を適切に把握していない状態を，「曖昧な情報要求」を持つ状態として，その情報要求が言語的に表現されるときの特徴を探索的に調査した．

言語化された情報要求とみなされてきたものの一つに，コミュニティ Q&A サイトに投稿された質問文がある[1]．本研究では特に，インターネットに存在する情報を参照する URL を含むような回答が寄せられた質問文に注目し，それらの質問文に共通する特徴を探索した．回答となる Web ページがインターネット上に存在するにもかかわらず質問が投稿された原因のひとつは，質問者の Web 検索によって問題解決に役立つ Web ページが発見・検索できなかったこと，すなわち質問者が明確な情報要ができなかった（曖昧な情報要求のみが可能であった）からだと考えられるからである．

2. 対象と方法

調査対象として，Q&A サイトである「Yahoo! 知恵袋」に過去に投稿され解決した質問から無作為抽出されたデータセット[3]を用いた．データセットに含まれている質問約 206 万件のうち，約 24 万件に，URL を含む回答 (以下「LA」とする) があった．LA があった質問 (以下「LA あり質問」) と，そうでない質問 (以下「LA なし質問」) それぞれから，質問者自身が選択した質問主題を示す「カテゴリ」の比を維持できるように 1 万件ずつを抽出し，分析の対象とした．

分析対象とした質問の全文に対し，形態素解析用辞書 UniDic, 形態素解析用エンジン MeCab を用いて形態素解析を行った．その結果をもとに，質問文の文字数と単語数の分布，および，質問文中の各単語（名詞・動詞・形容詞・副詞に限定し，日本語の文章に一般によく出現する語[4]・補助動詞を除いた）出現頻度を調べた．

3. 結果

各質問文の文字数と単語数は幅広く分布し, LA あり質問と LA なし質問の間に明確な差は見られなかった. (図 1，図 2).

一方で，質問文中の出現頻度が高い語については，LA あり質問と LA なし質問の間で違いがみられた．例えば，「教える」・「下さる」・「分かる」といった語は，いずれの質問文でも出現頻度が高かったが, LA あり質問でより多く見られた．また，「曲」のように LA あり質問でのみ出現頻度が特に高い語もあった． (図 3，図

4)

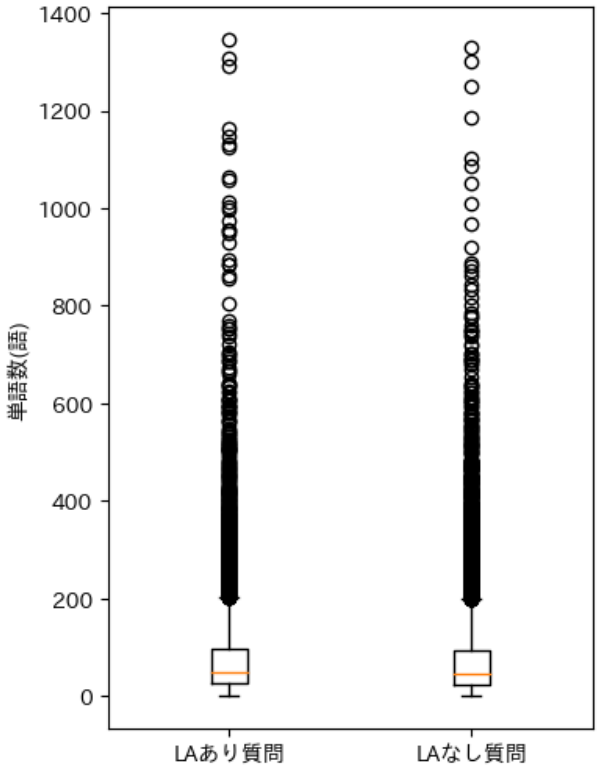

図 1　各質問文の文字数

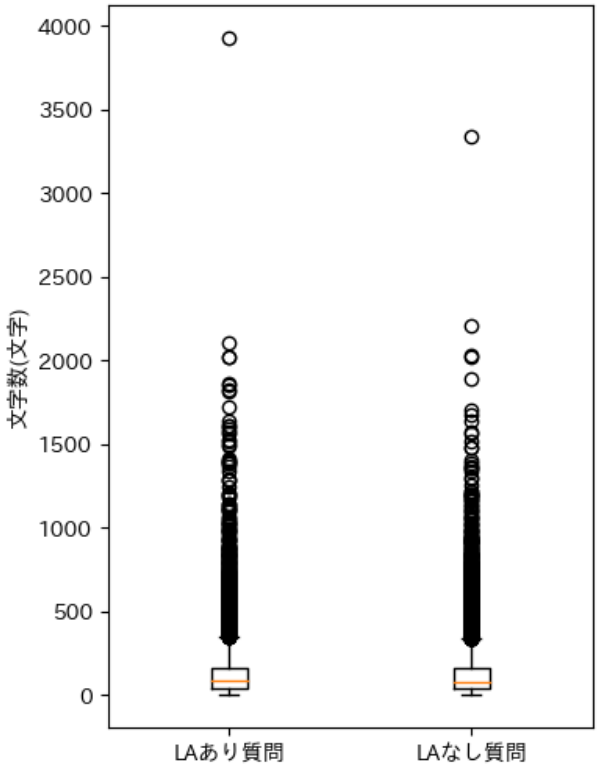

図 2　各質問文の単語数

4.　考察

調査の結果，Q&A サイトにおいて，外部の web ページを参照する回答が寄せられるような質問文について，文字数・単語数においては明確な特徴がみられなかったが，「教える」「下さる」「分かる」や「曲」といった単語の出現頻度が高いという特徴が発見された．

「教える」「下さる」といった語の出現頻度が高い理由として，丁寧に質問文が記述されている・事実を問うている，といった質問の仕方によるものが考えられる．また，「曲」という単語の出現頻度が高い理由としては，特定の音楽に関する情報を求める質問には外部の web ページを参照して回答しやすい，といった質問の主題によるものが考えられる．いずれも，今回発見されたような特徴が生じる理由を明らかにするためにはさらなる検証が必要であるが，外部の web ページを参照する回答が寄せられる要因として，質問者の情報要求の曖昧さ以外にも，質問の主題や質問の仕方といった要因がある可能性が示唆された．

また，今回発見された特徴のうち，質問者の情報要求が曖昧であったことに起因するものがあるかについても，さらなる検討が必要である．

謝辞

本研究では，国立情報学研究所の IDR データセット提供サービスにより LINE ヤフー株式会社から提供を受けた「Yahoo! 知恵袋データ（第 3 版）」を利用した．

文献

[1] 大塚 淳史・関 洋平・神門 典子・佐藤 哲司 (2011). “情報要求の言語化を支援するクエリ拡張型 Web 検索システムに関する一検討”,情報処理学会論文誌データベース,*4*(3),1-11. http://id.nii.ac.jp/1001/00078039/

[2] 和田 洋祐・相場 亮 (2009). “曖昧な情報要求に対する分類と推薦を用いた検索支援”,情報処理学会第 71 回全国大会論文集,271 - 272. http://id.nii.ac.jp/1001/00139022/

[3] LINE ヤフー株式会社 (2023). “Yahoo! 知恵袋データ（第 3 版）”,国立情報学研究所情報学研究データリポジトリ. https://doi.org/10.32130/idr.1.3

[4] 大島 裕明・中村 聡史・田中 克己 (2007). “SlothLib: Web 研究のためのプログラミングライブラリ”,日本データベース学会 Letters,*6*(1),113-116.
注：実際に使用したページを次に示す.
http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt

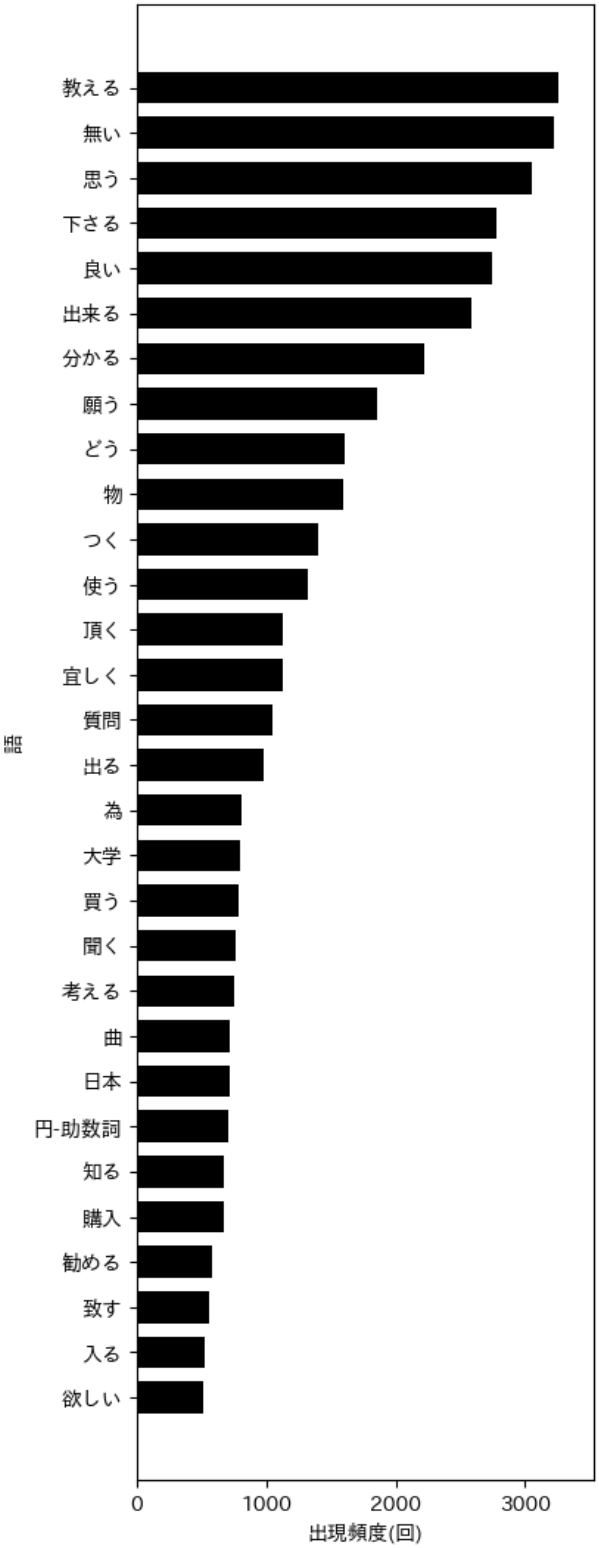

図 3　LA あり質問文中の単語出現頻度（上位 30 語）

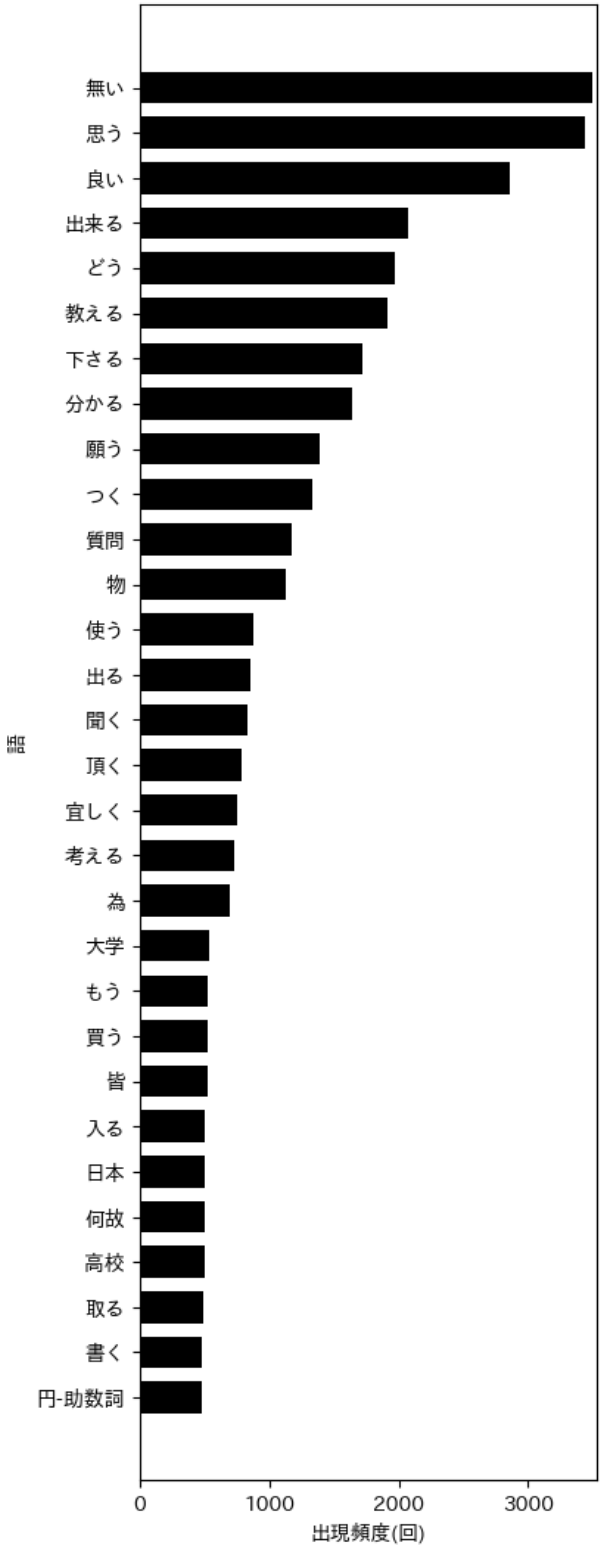

図 4　LA なし質問文中の単語出現頻度（上位 30 語）