リサーチデザインにおける言語資源の役割
―QAサイトコーパス(知恵袋データ†)の場合―

中渡瀬秀一　（国立情報学研究所）

2022

研究デザインにおける言語資源の役割とは？

概要

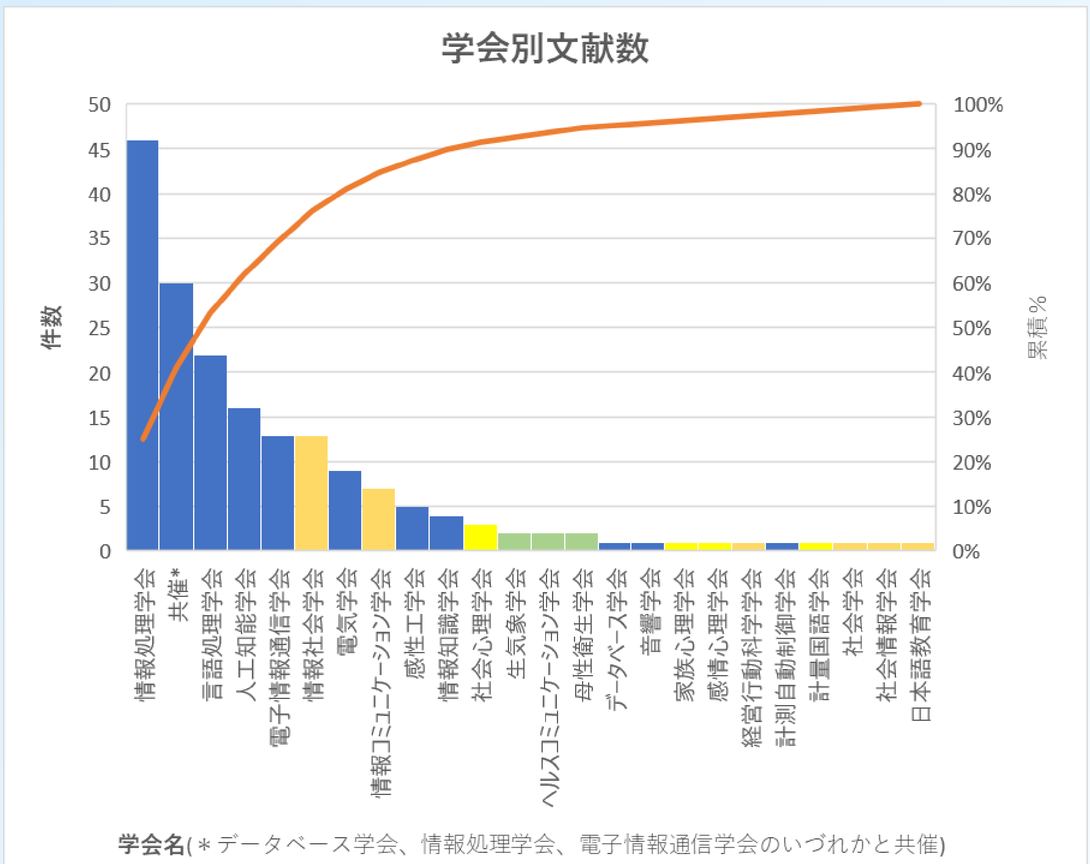
近年、言語学以外の様々な研究分野においても言語資源が活用されている。大学共同利用機関である弊所は研究者に言語資源など各種データ資源を配布することによって研究を支援する活動を行っている。
研究を促進する言語資源を開発するためには、言語資源が研究においてどのような貢献をしているのかを把握することが重要である。
上の背景から本研究では弊所が研究者に配布している言語資源（QAサイトのコーパスである**知恵袋データ**†）を対象にして、この資源の研究デザインにおける役割について調査した。本稿ではその調査結果を報告する。

方法：知恵袋データ†の利用契約者から毎年報告される研究成果文献一覧‡(2008年～2020年)をもとに国内学会系の媒体(論文誌・全国大会/シンポジウム/研究会の予稿集)に投稿された論文(184件) の中からオープンアクセス文献（88件）であるものを対象に以下の事項を調査した。
・学会別（分野別）の文献数
・研究デザインの類型とその型において言語資源の果たす役割
・知恵袋データに特有な研究
結果：知恵袋データは工学分野での利用が多いが人社系分野でも活用されている。研究デザインは2種類に大別される。一方はシステムや計算手法などの技術提案である。この場合、言語資源は技術を構成する要素(処理対象や制御情報) である。他方は社会的ニーズや現象の特徴などの知見を獲得することを目的とした分析である。この場合、言語資源は分析の対象である。

†Yahoo! 知恵袋データ（Yahoo!データセット）
2007年から国立情報学研究所がヤフー株式会社から提供を受けて研究者に提供しているデータセットでYahoo!知恵袋の質問と回答が含まれている。

1:Yahoo!知恵袋データを活用した成果文献数

・**Yahoo!知恵袋データ**を活用した研究成果発表数(2008年-2020年)を学会別に集計した。その結果、データ配布を告知した**工学系学会(情報処理分野)**での利用が大半を占めていたが、それ以外にも心理学(人文学)・社会学(社会科学)・衛生学系(医療関係)の分野の研究にも活用されている。



Yahoo!知恵袋データを活用した研究成果文献数(国内学会別)

2:研究デザインの型（技術提案型と分析型）

技術提案型

対象論文の大半を占める工学分野の研究に多く見られる研究プロセスは以下のように模式化される。このような研究のデザインの型を「技術提案型」とする。

１：目的 ⇒　２：技術提案 ⇒　３：評価/検証

・**過程**：まず研究の目的を設定する。次にその目的を実現する新技術（システム、技法、計算法等）または既存技術の改善方法を提案し評価実験を行う。最後に実験結果を分析することで提案の有効性を示すと共にその限界、残された課題、改善のための知見などを獲得する。

・**言語資源の役割**：知恵袋データは提案技術においてデータ処理の対象、または処理において参照される情報の役割を果たす。前者の技術例としては記事分類法、記事要約法、文章難易度計算法、意見抽出法などがある。また後者の例では質問応答技術、質問作成支援法、クエリ拡張推薦などがある。

分析型

工学分野の一部、また心理や医療分野の研究には上記とは異なる研究プロセスが見られた。それらは以下のように模式化される。このような研究のデザインの型を「分析型」とする。

１：目的 ⇒　２：対象分析 ⇒　３：結果解釈,知見獲得

・**過程**：まず研究の目的を設定する。次にその目的のために資料の分析を行う。最後に分析結果の解釈を行い有用な知見を獲得する。

・**言語資源の役割**：知恵袋データは分析対象となる資料である。分析の目的例としては、観光や健康情報に対するニーズ把握、ベストアンサーの特徴抽出、質問文分類のための知見獲得などが見られる。

3:知恵袋データの特徴と研究テーマ

知恵袋データは単なる言語データではなく、QAサービスを通じて生成された言語記録であるため以下のような特徴を持つ。

・質問とその回答との対話形式（ベストアンサー情報有）である
・内容に投稿時の言語使用実態や現実世界が反映されている
・質問はカテゴリに分類されている
・コミュニティ情報が含まれる（質問者と回答者の関係等）

上記の特徴が一般の言語資源としての活用の他にこのデータ特有の研究を可能にしている。以下に双方の研究例を挙げる。

・**一般言語資源としての研究例**：
　知識抽出、文章クラスタリング、トピックモデルや言語モデル作成、要約手法、テキスト難易度の計算、語の距離計算、概念階層の獲得、固有名抽出 など

・**知恵袋データの特徴を活かした研究例**：
[**QA情報**]
　ベストアンサー推定、質問構造モデリング、QAコミュニティ活性度計算、質問応答システム など
[**カテゴリ情報**]
　旅行情報や健康情報の分析、ドメインの形容語抽出 など
[**コミュニティ情報**]
　ベストアンサー推定 など

・**データの特徴と研究デザインの関係**
　QAサイトのデータに由来する特徴はQAサービスの改善や拡張に関わる技術提案の創出に貢献している。文が現実世界を反映し、カテゴリや日時情報を持つことは社会調査的な分析研究を可能にしている。このように言語データに付加された情報は研究課題やデザインの幅を広げている。そのため研究を促進する言語資源を開発する上でもこの点が重要になる。

参考資料：Yahoo! 知恵袋データ（提供：ヤフー株式会社、配布：NIIデータセット共同利用研究開発センター）https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr2/Y_chiebukuro.html
‡研究成果一覧　https://dsc.repo.nii.ac.jp/?action=repository_opensearch&index_id=22&count=20&order=16&pn=1

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
中渡瀬秀一　お問合せは ⇒ https://forms.gle/pBBeduzsoEYP2QsA6