

自然言語処理のための統合型 Web アプリケーション「品詞分析くん」の試作

岩井 憲一

滋賀大学教育学部

iwai@edu.shiga-u.ac.jp

自然言語研究には多くの優れた解析理論の展開がなされ、それらを実現したツールが構築・提供されている。しかし、ツールの大半が PC アプリケーションであり、初学者にとってはインストールの段階で躓いている場合も見逃せない。そこで筆者はいくつかのツールを Web アプリケーション化するとともに、グラフ表現による可視化機能を別途いくつか構築してこれら既存のものと合わせた統合型 Web アプリケーション「品詞分析くん」を試作した。本稿ではその概要について述べる。

1. はじめに

自然言語研究には多くの優れた解析理論やツールが構築・提供されている。しかし、ツールの大半が PC アプリケーションであり、初学者にとってはインストールの段階で躓いている場合も見逃せない。そこで筆者はいくつかの既存ツールの Web アプリケーション化を中心に、初学者が気軽に扱えるような支援サービスの実現を目指してきた。本稿では筆者がこれまでにやってきた取り組みの概要について述べる。

2. 研究の概要

2.1 研究の背景

自然言語処理ツールはオープンソースの PC アプリケーションが多く利用されており、ツールにデータを入力してその結果を可視化させて成果発表を行う形式が主流であった。

しかし、これらのツールを他のシステムに組み込んで新たな解析に流用したい場合は、ある程度プログラミングに精通する必要がある。そもそもツールを利用するためにインストールも覚束ない初学者は、容易に手を出すことは難しい。

そこで筆者はそのような初学者においても「必要性があればインストールも行うが、まずは一度使ってみよう」という声に応えるべく、既存のツールの Web アプリケーション化を試みた。

2.2 導入した解析ツールについて

Web アプリケーション化に際して、導入した既存の解析ツールについては、形態素解析ツールとして MeCab⁽¹⁾、係り受け解析ツールとして CaboCha⁽²⁾、単語のベクトル化による関連語導出ツールとして Word2Vec⁽³⁾の3つを採用した。

ツールの Web アプリケーション化についてはすでにいくつかのサイトで実現⁽⁴⁾されており、研究者向けにもクラウド型開発サイト⁽⁵⁾が提供されているが、本研究では初学者向けの統合型 Web アプリケーションサイトを目指している。

2.3 本研究のシステムについて

図1に筆者が構築した Web アプリケーションである「品詞分析くん」の入力インタフェースを示す。

品詞分析くん for Q&A + KuBOT + 共起

カテゴリ

質問題目

内容説明

表示初期化

図1 品詞分析くんの入力インタフェース

本システムの入力フォームはカテゴリ別で整理した質問題目とその内容説明についての自然言語処理を想定したものである。

このフォームに次のような内容を入力すると、本システムは、一連のツールの出力結果を続けて出力する。まずは以下(a)~(c)にて既存の解析ツールの応答結果(一部)を示す。

カテゴリ：三角形
 質問題目：合同条件
 内容説明：三角形の合同条件について
 教えて下さい。

(a) MeCab(形態素解析ツール)の場合

MeCab によって内容説明の例文を形態素解析した結果の一部を図2に示す。

分析結果：
 三角形 名詞,一般,*,*,*,三角形,サンカクケイ,
 サンカクケイ
 の 助詞,連体化,*,*,*,の,ノ,ノ

図2 Web アプリ化した MeCab の出力例(一部)

(b) CaboCha(係り受け解析ツール)の場合

CaboCha の場合は、MeCab によって分解された単語列間で係り受け関係のある単語対の列を出力する。その結果を図3に示す。

係り受け解析結果：

三角形 => 合同条件

合同条件 => 教える

図3 Web アプリ化した CaboCha の出力例

(c) Word2Vec(関連語導出ツール)の場合

Word2Vec の場合は, MeCab によって分解された各単語の関連語を出力する. その結果の一部を図4に示す.

関連語一覧：

三角形

(‘四角形’, 0.9194488525390625)

(‘正三角形’, 0.9083343744277954)

図4 Web アプリ化した Word2Vec の出力例(一部)

3. 本システムで実現した可視化機能について

3.1 比較グラフ可視化機能について

本節では論旨の都合から, フォームへの入力内容を以下のように変更する.

カテゴリ：育児

質問題目：育児中に飲みに行くこと

内容説明：9ヶ月の子供がいる母親です。現在育休中です。元々お酒が好きで、離乳食も三回食でミルクも飲むようになったのでお酒を解禁しました。

図1のインタフェースは, 本来は「質問と最適な回答の対」と「質問と最適ではない回答の対」を比較したいというねらいがあった. その過程で, 質問フォームに加えて回答フォームも用意しようとしたが, 質問フォームのみで, その結果を蓄積して比較するような簡素化した可視化機能の方が扱いやすいと考えた. この可視化機能はグラフ表現で出力しており, この出力グラフを「比較グラフ」と名付けた.

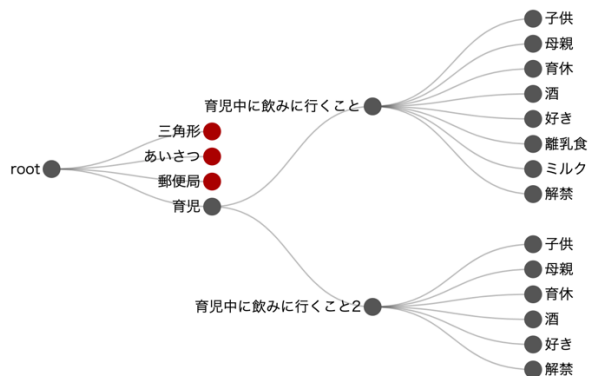


図5 比較グラフの出力例

図5に比較グラフによって可視化された例を示す. 各ノードをクリック(タップ)することで子ノードを展開・縮小することができる. 図5では比較の例を示しており, 質問題目のフォームに「育児中に飲みに行くこと」や「育児中に飲みに行くこと2」のように質問の名称を変えながら入力することで構成する単語列を比較することができる.

3.2 共起ネットワーク可視化機能について

MeCab によって生成された単語列から共起ネットワークを出力することができる. 本機能で生成された共起ネットワークの例を図6に示す. 本システムではブラウザのタブごとに各グラフを出力させる形式を採用しており, 出力された共起ネットワークを並べて比較させることで, その本質や違いに気づきやすくなると考えられる.

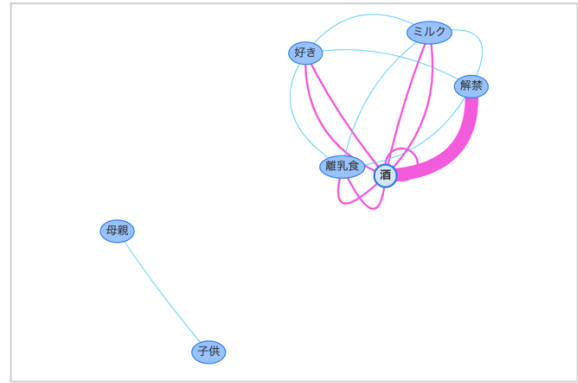


図6 共起ネットワークの出力例

4. おわりに

本稿では, 自然言語分析のための Web アプリケーション「品詞分析くん」におけるこれまでの成果について述べた. 本システムでは可視化機能は JSON 形式で出力されているので, 今後 Web API 化等さらなる改良を推し進めていく予定である.

参考文献

- (1)MeCab, <https://taku910.github.io/mecab/> (2020年11月22日現在).
- (2)CaboCha, <https://taku910.github.io/cabocha/> (2020年11月22日現在).
- (3)Word2Vec, <https://github.com/RaRe-Technologies/genism> (2020年11月22日現在).
- (4)形態素解析ウェブアプリ UniDic-MeCab, <http://www4414uj.sakura.ne.jp/Yasanichi1/unicheck/> (2020年11月22日現在).
- (5)テキスト解析- Yahoo!デベロッパネットワーク, <https://developer.yahoo.co.jp/webapi/jlp/> (2020年11月22日現在).