C-012

テキストを含むデータに対するパイプライン型解析の有用性と日本語解析用ツールの開発
Usefulness of The Pipeline Type Analysis for Data Including Text and Development of a Tool for its Japanese Analysis

箕輪 弘嗣[1)]
Hirotsugu Minowa

1　はじめに

近年、機械学習、深層学習などの AI と言われる研究の発展、有用性の周知をきっかけに、データの重要性が認識され、パブリックデータなる情報公開が進められている。データの公開により、潜在的法則、知見を解明し有益な知見を得る事が狙いである。データの解析にはプログラミングスキルを要する。問題は、データ解析には知識以外にプログラミングスキルの習得を要し、データ解析の壁になっている。

そのため、近年、ノン・プログラミングツールなるプログラミングを必要とせず数値, テキスト, 地理情報 (GI) などのデータを解析できるツールの開発が進められている。例えば、Orange Canvas(以下 OC)[1], KNIME といったデータ解析ツールでは、Widget と称する単一処理機能を連結する事で多彩な解析を実現している。この処理をパイプライン型ツールと呼称するとする。一方、テキストに関しては言語ごとにツールを開発する必要があり、KHCoder や Text Mining Studio などの開発がなされている。

Orange Canvas らパイプライン処理の方が、解析の自由度が高く、多彩な解析を実現できる可能性がある。しかし、Orange Canvas ではテキスト解析の起点となる日本語の形態素解析ができない問題があった。そこで、筆者は日本語を形態素解析する Widget と、それら Widget らを有すアドオン (Addon) なる追加ソフトを簡単に導入する事ができる GUI ツールを開発したので報告する。

2　一般日本語文章の解析による可視化

開発した Widget を用い、多く解析されている共起ネットワークを実現する。対象は夏目漱石の「こころ」。その解析手順である Workflow を図 1 に示す。

Orange Canvas(v0.29.3 使用) へ Text mining, Textable, Network の Addon を導入する。加えて図 1 に示す、Filter(仮) と JA Morphological Parser(仮) の 2 つの開発したウィジットを内包するアドオン NLP を導入する。

「こころ」の電子データ 773_ruby_5968.zip を青空文庫サイト [2] より取得します。Zip ファイルを展開して得られた kokoro.txt ファイルをエディタで開き、冒頭のヘッダー部に該当する 16 行を削除したファイル kokoro_proc.txt を生成する。次に Text Files ウィジットで kokoro_proc.txt を OC 上に呼び出す。この際、Windows 以外の OS は文字コードを UTF8 といった適切値へ設定する必要がある。

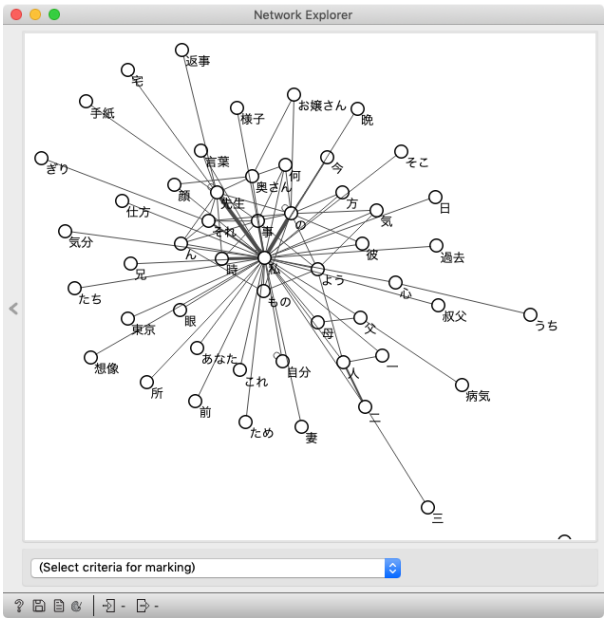

図 2　Orange Canvas による共起ネットワーク

次に Segment ウィジットで ‘Segment type‘に ‘Segment into lines‘を選び、対象文章における区分を行化する。Interchange ウィジットを用い、受け渡しデータ型を Segmentation 型から Corpus 型に形式変換する。「こころ」のテキストデータには、未だルビ (例:《わたくし》) やレイアウト指定文 (例:［＃５字下げ］一［＃「一」は中見出し］) が各所に残っている。それらを除去するために開発した Filter ウィジットを開き、「《.*》| [.*] |[「」、。]」と入力する。このコマンドは正規表現であり、「《と》囲まれた箇所、または、［と］で囲まれた箇所、または、「「」、。」のいずれかに一致する箇所の除去、を意味する。

次に JA Morphological Parser ウィジットを用い、各行

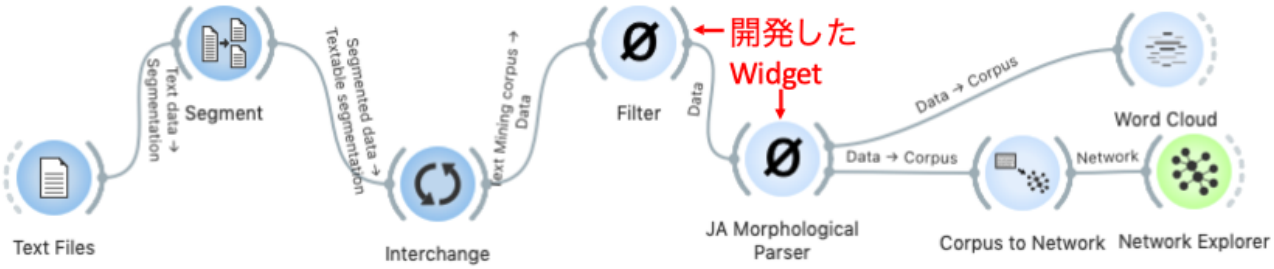

図 1　Workflow

1)　岡山商科大学

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

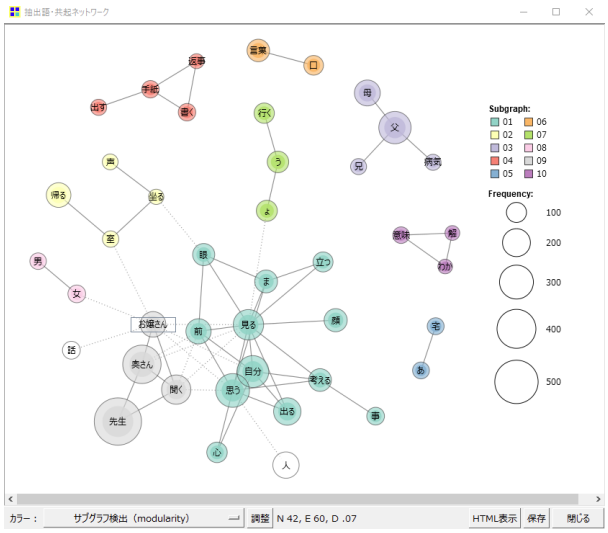

図 3　KHCoder による共起ネットワーク

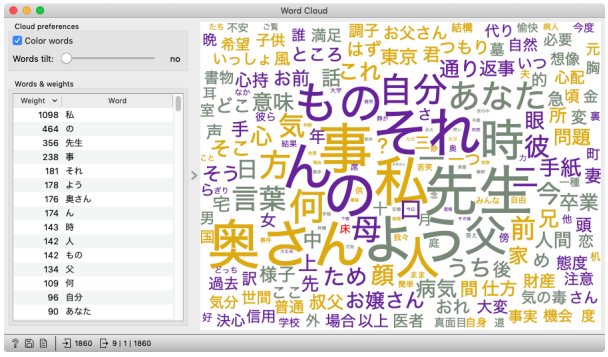

図 4　Orange Canvas による抽出語

内の文を形態素解析する。本ウィジットの形態素解析には Janome[3] を使用した、また、抽出語を「名詞」に限定した、理由は、単一単語だけでも比較的意味を推測できるからである。解析結果を Corpus to Network ウィジットへ渡す。そのウィジットのパラメータは 'Node type':Word、Threshold:7, Windows Size:1, Freq. Threshold:4 とした。その結果を Network Explorer へ渡す事で、共起ネットワーク解析の結果が得られる。OC の結果を図 2、KHCoder の結果を図 3 に示す。また、Word Cloud によって、抽出語が表示が可能である、OC の結果を図 4、KHCoder の結果を図 5 に示す。

両者の解析結果は異なっており、それらの差における問題がないか、近づける事ができるか、などを明らかにする必要がある。

3　OC Addon Free install helper の開発

本アプリ (図 6) は、筆者を含む他者が開発した非公式アドオンを Orange Canvas へインストールする負担を解決する。Orange Canvas へのアドオンのインストールは GUI 上で可能だが、それは、Orange Canvas に公式に登録されたアドオンに限られる。非公式 (公式以外) のアドオンの導入は、pip や conda といった Python のパッケージ導入用 CUI コマンドを用いる必要がある。しかし、これら CUI コマンドによるアドオン導入は、CUI に慣れない方から敬遠されてしまう恐れがあるためである (もちろん、本来なら、開発元で対応してもらえるのが最良であろう)。筆者はゼミ学生に使用してもらうため開発し

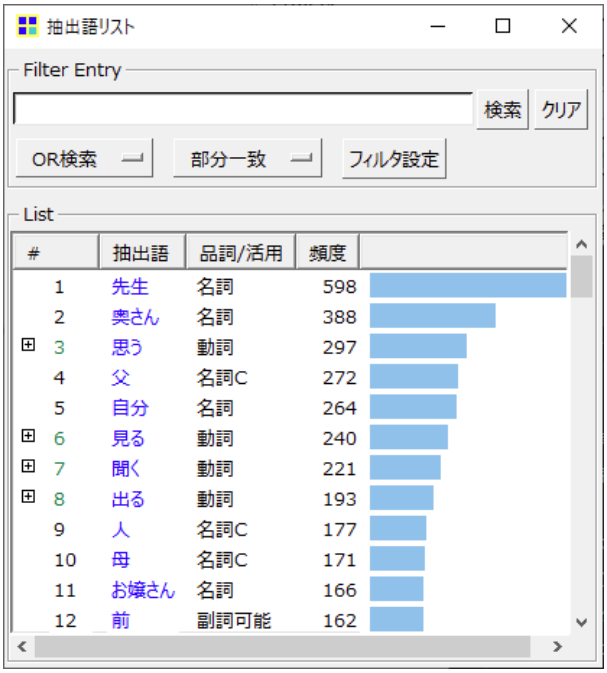

図 5　KHCoder による抽出語

た。アドオンのダウンロード先 URL は、アプリ同フォルダに作成される YAML 設定ファイルに、パッケージ詳細といったインストール源と共に記載されている。

本アプリの使い方はとても簡単である、GUI 上部から順にインストール先 PATH, インストール元の選択、インストールするパッケージ名とその Version を指定し、Install ボタンを押す事でインストールできる。

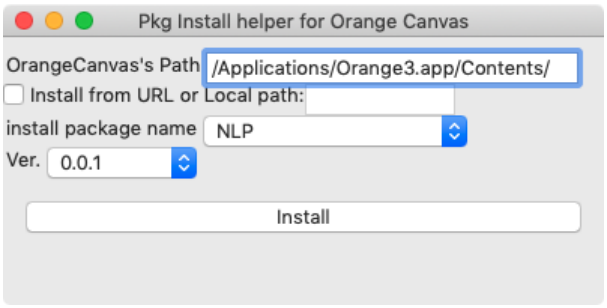

図 6　`OC installer helper` の画面

4　おわりに

本ソフトウェアは、筆者 Web サイト [4][5] いずれかからの配布を予定している。

参考文献

[1] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan, "Orange: Data mining toolbox in python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.

[2] "青空文庫　Aozora Bunko," https://www.aozora.gr.jp/.

[3] "Welcome to janome's documentation! (Japanese) — Janome v0.4 documentation (ja)," https://mocobeta.github.io/janome/.

[4] "HirotsuguMINOWA - Overview," https://github.com/HirotsuguMINOWA.

[5] "箕輪 弘嗣 (Hirotsugu Minowa) - マイポータル - researchmap," https://researchmap.jp/hirotsugu_minowa.

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.