# 類語辞典をつかう・類語を見つける（IRSME16045）

平成 29 年 1 月 19 日　原田 長州

文章を作成する際に、思いついた表現がしっくりこないことがある。思いついた表現に似た表現などを検索するときに利用するのが類語辞典である。

インターネットで公開されている類語辞典としては、Weblio 類語辞典[1]がある。

例えば、「いざこざ」で検索した場合の結果はこのようになる。（図 1）


Weblio 類語辞典のデータは、国立研究開発法人情報通信研究機構（NICT）の日本語 WordNet[2]にもとづくものである。それだけでなく、Weblio 独自の類語も含まれている。

この類語辞典を使う場面としては、言い換えの表現を探す場合や、ひとつの事柄を複数の言い方で表現するために類似表現を探す場合などが挙げられる。

**図 1 検索結果画面**
意義素と類語に分かれて表示される

## ■ 日本語 WordNet の類語について

日本語 WordNet の類語は、同じ概念を共有する同義であるとされる。同義かどうかは人間が判断している。現時点では日本語 WordNet のデータが公開されているだけであり、新しい版が公開される以外には類語を自動で追加する仕組みなどは提供されていない。

1 http://thesaurus.weblio.jp/
2 http://compling.hss.ntu.edu.sg/wnja/

## ■ 類語の追加

日本語 WordNet のページにおいて 1.1 版のデータが公開されている(本稿執筆時現在)。公開データがあるため、ユーザー側でデータを追加することができる（ただし、追記した独自の類語データを検索・表示するためには別のシステムが必要である)。

## ■ 類語を見つける

類語を見つけるためにはさまざまなアプローチが考えられる。今回は、特定の分野に関する文書を集め統計的に処理し、最終的に人間が判断するという方法を選択した。

収集した特定分野の文書には独自用語や独自の概念などが含まれている。類語と認識できると明確には考えられていなかった概念を視覚化・言語化することができる可能性がある。また、類語を探す段階で単純な用語の不統一に気づく場合もある。

1. **文書を集める**

   サンプルとして集めた文書を利用する。ファイルの形式は、テキスト形式ファイルやＣＳＶ形式ファイルにしておく。

2. **統計的処理をするために KHcoder[3]を使う**

   今回は KHcoder を利用してテキストのデータを分析する。このソフトは文書のデータを分析するフリーソフトウェアである。集計したデータを CSV 形式などで取り出すことや、ソフトウェアのソースコードも公開されている。このソフト単体で処理できない作業であっても、他のソフトと連携して対応することができる。

   KHcoder は、文章を単語単位に分割して分析できるようにする。単語 A と単語 B が同じ文章内で同時に現れる回数を数えて「共起ネットワーク図」（図２）を作成して視覚的に把握できるようになる。また、図３では、単語 A と単語 B の類似の割合をしめす Jaccard 係数の降順で並んでいる。

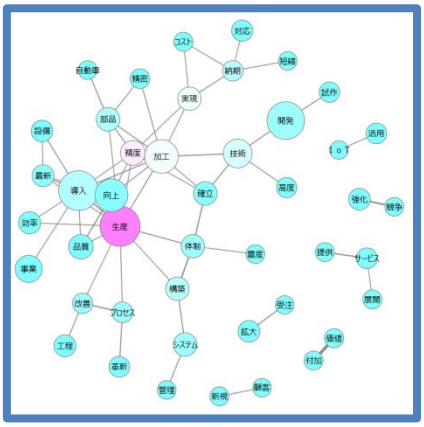


**図 2 共起ネットワーク図の例**

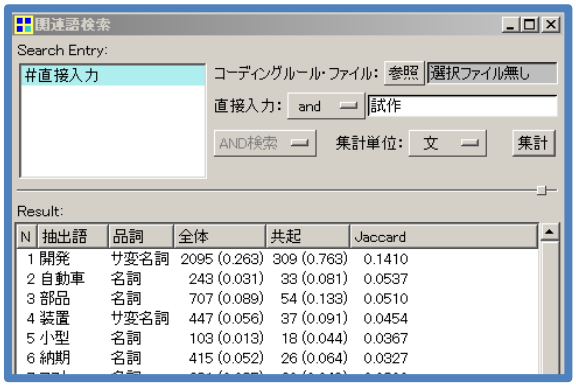


**図 3 関連語検索画面**

[3] http://khc.sourceforge.net/

このようなソフトを利用せずに、概念を整理・構築して追記することは可能であるが、処理スピードが上がらないだろう。

## ■ まとめ

本稿では、類語辞典を利用することから、利用しやすいライセンス形態で公開されている類語の拡張を目指して、類語を抽出する方法について考えた。フリーソフトウェアの KHcoder を利用すれば、文書から集計・分析ができるようになる。

数値とは異なり、扱いにくい文章をコンピューター上で処理できるようになることのメリットは大きい。(了)