

比喩的素描を用いた類似語推論 およびその視覚化インタフェースの構築

長谷川 恭佑[†] 梶井 文人[‡] 後藤 文太郎[‡]

北見工業大学大学院工学研究科[†]

北見工業大学情報システム工学科[‡]

1 はじめに

WWW 検索に代表されるように、今や WWW はかつて無い規模の巨大知識源として広く活用されつつある。そうした WWW 活用の試みのひとつとして、我々は、WWW を利用した語の意味を素描する手法を提案し、図 1 に示す実装システム Murasaki の構築に取り組んで来た [1][2]。この手法は、語と語の間に存在する比喩的關係に着目することによって、ユーザが提示したクエリ語について様々な観点から説明する記述要素（デスクリプタ）をダイナミックに収集・整理して雑駁に説明するものである。



図 1: Murasaki 出力画面

このような未知語や新語の意味を手軽に調べ把握できる仕組みは、現代社会の情報アクセス活動において非常に有効であるが、更なる課題も残っている。

例えば、ユーザがある語について調べたいとき、場合によってはその語義のみならず対象語に類似した語についても把握しておきたい状況などが考えられる。具体的には、ユーザがクエリ語として「初音ミク」を提示した場合、現在の Murasaki では素描集合として {ボーカロイド, ソフト, キャラクター, ...} が得られる。これにより、「初音ミク」とは「ボーカロイド」の一つであることが連想的に把握できるのであるが、ユーザが「ボーカロイド」に属する他のインスタンスを知りたいと考える可能性がある。この場合、[巡音ルカ, 鏡音リン・レン, ...] といった情報提示が望ましいが、現状システムにはこの要求に対応できる仕組みが存在しない。

そこで本稿では、上記手法を応用した類似語推論とその視覚化について述べる。本研究における基本的考え方は以下の通りである。類似語同士は、その説明、素描集合同士も類似しているはずである。したがって、Murasaki システムが持つ素描集合を比較し、その類似性を判定する仕組みを組み込むことにより、クエリ語に対する類似語を推論することが可能である。

以下、2 章で類似語推論とその視覚化について説明し、3 章で構築した機構について考察を述べる。

2 類似語推論

本章では、類似語推論のためのログデータの整理と類似度計算について説明し、実行例を示す。図 2 にシステムの構成を示す。

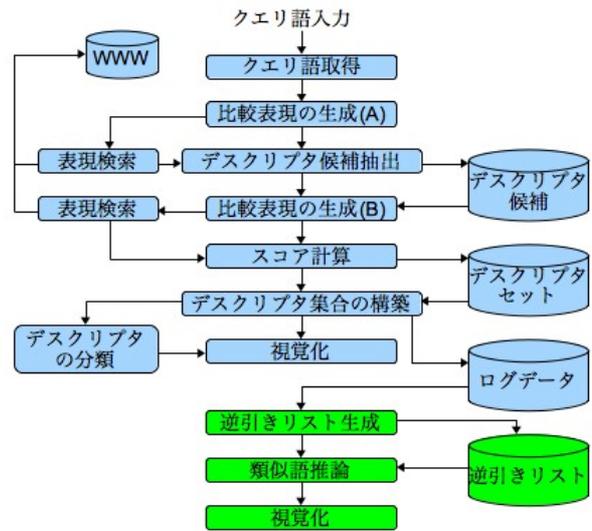


図 2: システムの構成

2.1 システムの概要

・ ログデータの整理

類似性を扱う基礎データとして Murasaki ログデータを利用する。類似語同士は類似した語義を持つため、語義の比較によって類似語の推論が可能であると考えられる。例えば、図 3 に示すように「自民党」というクエリ語はデスクリプタ「政党」を持つ。同じように「政党」をデスクリプタとして持つ語には「民主党」や「公明党」などがある。これらの語は「政党」という観点でみれば類似した語であると言える。

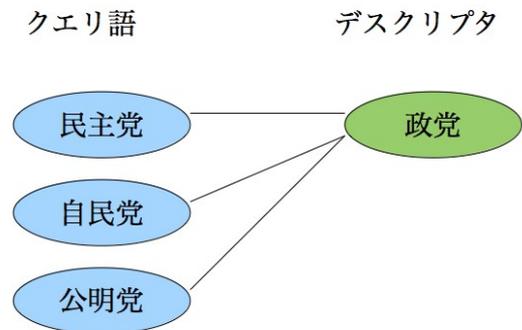


図 3: 類似語の関係

ログデータにはクエリ語毎に、そのデスクリプタとデスクリプタに付与されたスコアが CSV データ構造で格納されている。このログデータを利用して、共通のデスクリプタを持つ語を発見できれば類似語の推論が可能である。類似語推論のためには、デスクリプタから

Synonym Reasoning based on Figurative Description and Visualization

[†] Kyousuke Hasegawa

Kitami Institute of Technology

[‡] Fumito Masui, Fumitaro Goto

Kitami Institute of Technology

クエリ語へ辿る仕組みが必要となる。そのため、事前にログデータを整理し、図4のようなデスクリプタ起点のクエリ語の逆引きリストを作成しておく。



図4:逆引きリストの生成

・ 類似語推論および類似度計算

類似語推論を行う際は逆引きリストを参照し、対象語と共通のデスクリプタを含んでいる語を類似語として抽出する。その際、クエリ語とデスクリプタの関連度を表すスコアも同時に取得し、類似度計算に用いる。類似度 $Sim(q,r)$ の計算は以下の式でおこなう。

$$Sim(q,r) = \frac{\sum_{i=1}^n S(i,q) \times S(i,r)}{\sum_{r=1}^m \sum_{i=1}^n S(i,q) \times S(i,r)}$$

対象語 q に対するデスクリプタのスコア $S(i,q)$ と類似語 r に対するデスクリプタのスコア $S(i,r)$ をかけて仮スコアとする。共通するデスクリプタが複数存在する場合はそれぞれの仮スコアを加える。すべての類似語に対して仮スコアの計算をおこなった後、それぞれの仮スコアを全体の仮スコアの合計で除したものを類似度とする。

・ 視覚化インタフェース

クエリ語に対するデスクリプタ、類似語をそれぞれ視覚化する。中央にクエリ語、その周囲をにデスクリプタまたは類似語が同心円状に配置され、スコアが円の大きさに反映される。これによりユーザは直感的に重要キーワードを認識できる。

2.2 実行例

本節では、クエリ語を「自民党」として構築したシステムの実行例を示す。図5は実際にシステムを動作させた際の出力画面である。

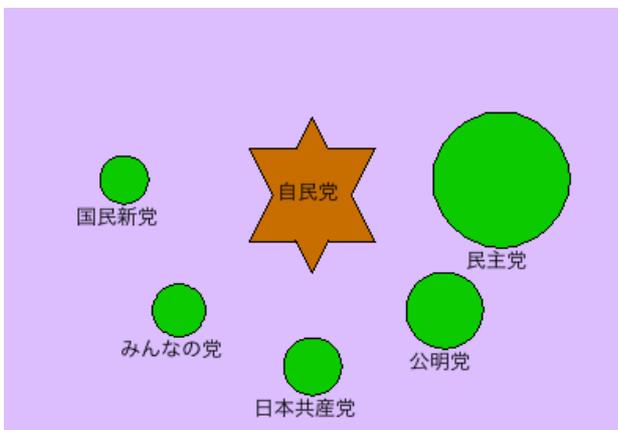


図5:類似語推論システムの出力画面

まず、Murasaki ログデータの中から「自民党」のデスクリプタとデスクリプタに付与されたスコア $S(i,q)$ {政党:0.1388, 派閥:0.0856, 党:0.0800, ...}を取得する。

次に、事前に Murasaki ログデータを整理して生成した逆引きリストを参照し、「自民党」と共通のデスクリプタを持つ語を類似語として取得する。テストに用いたデータには「政党」の逆引きリストの中に{公明党:0.2805, 民主党:0.1595, 自民党:0.1388, ...}が保存されていたため、自民党を除いたこれらの語を類似語として取得した。

最後に、取得した類似語について類似度 $Sim(q,r)$ を計算し、結果を視覚化する。

計算は以下の通りである。「自民党」に対する「政党」のスコア $S(i,q)$ 0.1388 を類似語の「政党」に対するスコアそれぞれにかけた結果、仮スコアは{民主党:0.0389, 公明党:0.0221, ...}となる。次に仮スコアの合計を計算し、合計値 0.1072 で個々の仮スコアを除いた結果{民主党:0.3631, 公明党:0.2064, ...}を類似度として算出した。

3 考察

構築したシステムの課題について考察する。今回提案した類似語推論手法は、単に共通のデスクリプタを含む語を提示する。類似語の提示について、その性能を検証する必要があると考える。具体的には、提示される類似語は知識獲得の補助として十分な量であるか、関係のない語は提示されないか、などを検証しなければならない。例えば、「政党」をデスクリプタとして含むクエリ語を提示した際、政党をどの程度網羅し提示できるか、また、政党でないものがどれだけ提示されるかが問題となる。特に、政党でないものが提示された場合は誤解の原因となる可能性がある。

また、本研究で提案する手法は、ログデータを整理し、視覚化に必要なデータを取得している。視覚化の度に整理するのではなく、事前に整理してデータを格納しておくことで即応性を保つ。事前に整理することで、整理後新たに作られたログデータについて反映されない問題が生じるため、定期的にログデータを整理し、データを格納し直すことで、データの新規性を保つ必要がある。

視覚化インタフェースについてはスコアを反映させた円の大きさや、クエリ語とデスクリプタの位置関係、色などによる表現方法が考えられる。

4 おわりに

Murasaki における類似語推論とその視覚化インタフェースの構築について述べた。提案手法により類似語の発見が可能となると考える。しかし、その性能について評価をおこなっていないため、実験などによる定量的な評価が必要である。

今後は、インタフェースの改良と提案手法の有効性についての検証をする予定である。

参考文献

[1] 川村佳史, 榎井文人, 河合敦夫, 井須尚紀:”WWW から Descriptive 知識を抽出・提示するシステム Murasaki の試作”, 言語処理学会第 12 回大会発表論文集, P8-10(2006.3.)

[2] 榎井文人, ジェブカ ラファウ, 木村泰知, 福本淳一, 荒木建治:”WWW 活用による語の比喩的素描手法”, 日本知能情報ファジィ学会誌, Vol22, No6, pp.707-709(2010)