

語彙定義のための日本語基本語彙リストの構築 および体系化に向けて

野呂 智哉 徳田 雄洋

東京工業大学 大学院情報理工学研究科

概要

語義情報を利用した自然言語処理では語義をどのように定義するかが重要となるが、その1つの方法として、少数の基本単語を利用して全ての語義を定義し、体系化することが考えられる。英語では、Longman や Oxford 等の辞書において、2,000 語から 3,000 語の基本単語を利用して全ての語義を定義している。日本語においても、基本単語の体系化に関する試みはいくつか行われているが、その基本単語で全ての語義を定義することを目的としていない、基本単語数が英語の場合と比較して多過ぎる等の問題がある。本研究は、語義の定義に必要な少数の基本単語を決定し、その基本単語で全ての語義を定義、体系化することを目的としている。本稿では、その構想について述べる。

1 はじめに

近年では、語義情報を利用した自然言語処理技術が注目されている。そのような技術では、語義をどのように定義し、体系化するかが重要となるが、それらを効率的に行うための1つの方法として、全ての語義を定義するために必要な少数の単語リストを用意し、その中の単語のみを利用して全ての語義の定義を行うことが考えられる。限られた単語で全ての語義を定義できれば、その語義をもとに効率的な体系化を行うことが期待できる。

英語では、Longman Dictionary of Contemporary English (以下、LDOCE と略す) [5] や Oxford Advanced Learner's Dictionary (以下、OALD と略す) [2] 等の辞書において、全ての語義を 2,000 語から 3,000 語の基本単語で定義している。一方、日本語では、限られた少数の単語で全ての語義を定義した辞書は、現在のところ存在しない。

日本語においても、基本単語の選定に関する試みはいくつか行われているが、それは、外国人に対する日本語教育等を目的として選定されたものであり、選定した単語のみを利用して全ての語義を定義することを直接の目的とはしていない。さらに、それらの多くは個人の主観によって選定されたものであり、体系的である一方、選定基準はその主観に大きく左右される。

客観的な選定手法として、言語資料を統計的に分析し、出現頻度や出現範囲(その語が出現する分野の広さ)をもとに基本単語を選定する手法がある。この手法は非常に単純であり、容易に選定することができるが、選定された単語リストが、全ての語義を定義するために必要十分な単語リストになるとは限らない。出現頻度の低い単語や特定の分野のみに出現する単語であっても、他の語義を定義する上で必要となる可能性もあれば、逆に、出現頻度が高くても、他の語義を定義する際には不要である(別の表現に換言できる)可能性もある。

別の手法として、単語親密度による選定がある [3]。これは、一般的に馴染みがある単語か否かをアンケートにより判定するものであり、その結果、約 28,000 語を基本単語として選定している。さらに、選定された基本単語を利用して語義の定義を行っているが、この 28,000 語という数は英

語の場合の 2,000 語や 3,000 語と比較して多過ぎる¹。出現頻度による選定の場合と同様、単語親密度による選定の場合においても、馴染みのある単語が必ずしも語義の定義に必要であるとは限らない。語義の定義に必要な単語の数は、28,000 語よりもっと少ないものと考えている。

本研究は、全ての語義の定義に必要な少数の基本単語を客観的手法により決定し、その基本単語で全ての語義を定義、体系化することを目的としている。限られた単語で全ての語義が定義できれば、語義情報を利用した自然言語処理技術に応用できるだけでなく、児童や外国人の日本語教育においても有用となると考えている。

本研究は、これから本格的に進めようとしているものであり、まだ具体的な結果が出ていないが、本稿では、本研究の構想を紹介し、当面の課題について述べる。

2 構想

本研究の流れは、以下のようになる。

1. 全ての語義を定義するために必要な基本単語の選定
 - (a) 選定手法の検討
 - (b) 基本単語数の決定
2. 基本単語による語義の定義

2.1 基本単語の選定

本研究では、基本単語を選定する際、内容語のみを対象とする。機能語の中にも、やや特殊なものが存在する可能性があるが、今回の検討の対象からは除外している。

基本単語の選定は、国語辞書の見出し語と語釈文を利用して行う。その手順は以下のとおりである。

1. 国語辞書の見出し語と語釈文から、各単語をノードとするグラフ (単語参照グラフ) を構築する
2. 構築したグラフをもとに、各ノード (単語) にスコア付けを行う
3. スコア順に単語をソートし、上位の単語を基本単語として採用する

単語参照グラフは、見出し語からその語釈文に含まれる各単語 (内容語) に対して有向枝をはることで構築する (図 1)。これにより、語義を定義する際の参照関係を表すグラフができる。図 1 の例は重みなしのグラフであるが、重み付きのグラフを構築することも考えられる。例えば、語釈文中の出現位置 (第 1 文か否か、文頭に近いか文末に近いか、など) にもとづくヒューリスティクスによって重みを変えることにより、関連の強さを反映したグラフができる。

こうして構築したグラフにおいて、他の単語から参照を多く受ける単語ほど基本単語としてふさわしいと考える。この仮定を反映する単純なスコア付けの方法として参照を受ける枝 (自身に向かってくる枝) の数をそのままスコアとすることが考えられるが、我々は、PageRank[4] や LexRank[1] の考え方を利用してスコア付けを行う。すなわち、単語参照グラフに対応する隣接行列の固有ベクトルを計算することにより、各単語のスコアを計算する。

¹実際は 28,000 語のうちの 17,000 語を使用して語義を定義しているが、それでも、まだ語数が多いと考えている。

単語	言語単位の一。文構成の最小単位で、特定の意味、文法上の機能を有するもの。
語彙	ある一つの言語体系で用いられる単語の総体。
文字	言語の伝達手段の一つとして使われる符号。点・線などを組み合わせたもの。

(大辞林より引用)

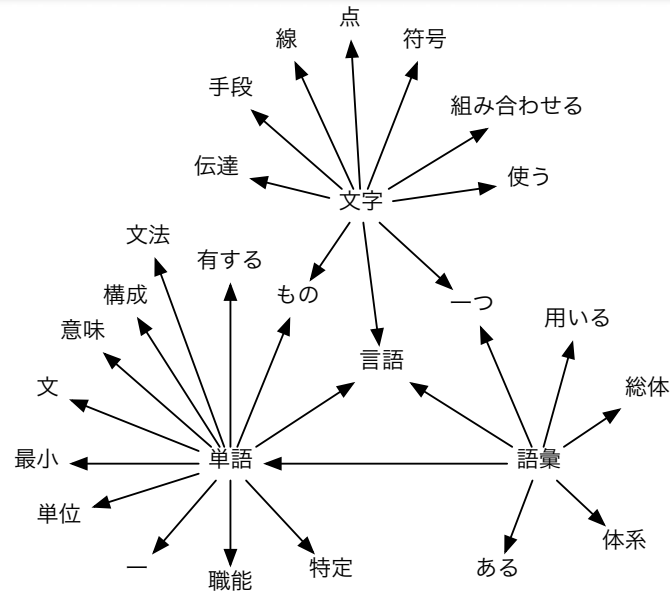


図 1: 単語参照グラフの構築

スコアを計算したら、各単語をスコア順にソートし、スコアの高い単語を基本単語とする。この際に問題となるのは、基本単語数をどのように決定するかである。本研究における基本単語とは、全ての単語を定義するために必要な単語であるため、一般的なテキストコーパスに出現する単語のうちのどのくらいをカバーするかを調べる等の方法は適切ではない。最終的には、次のステップで語義を定義しながら十分な基本単語数であるか否かを検討することになるが、事前に推定する方法については検討中である。

基本単語数の決定のほかに、選定された基本単語リストが妥当であるか否か(上述の手法で基本単語として適切な単語を決定できるか否か)を評価する必要がある。これも、最終的には、語義を定義する段階で検討することになるが、英語であれば、LDOCE や OALD のように既に基本単語を定めていない英語辞書に対して上述の手法を適用し、得られた基本単語リストと LDOCE や OALD が採用している基本単語リストを比較することにより、評価できるのではないかと考えている。

2.2 基本単語による語義の定義

基本単語を選定したら、それをもとに語義を定義する。その手順は、笠原ら [3] と同様の手順を想定している。

1. 既存の辞書の語釈文を初期状態とする。

2. 辞書の語釈文が基本単語と機能語のみで記述されている場合は、そのまま終了。

3. 基本単語以外の内容語が語釈文に含まれる場合は、語釈文を書き換える。

ここで、もし、基本単語と機能語のみで語義を定義できない単語があった場合、基本単語の選定手法、または基本単語数の決定方法の見直しが必要となる。

3 おわりに

本稿では、日本語の全ての語義を定義するために必要な基本単語リストの構築に関する研究の構想を紹介した。第1節で述べたように、全ての日本語の語義を限られた少数の単語で定義できれば、語義情報を利用した自然言語処理に利用できるだけでなく、日本語教育においても有用となる。実際、英語では、LDOCE や OALD は英語学習者にとって適した辞書とされ、さらにこれらの辞書を利用した自然言語処理の研究もされている。

本研究では、現在のところ、基本単語リストを構築する際、多義語を語義ごとに区別していない。すなわち、複数の語義を持つ単語が基本単語として選ばれた場合、どの語義であっても基本単語と見做される。しかし、実際には、同じ単語であっても、基本単語として必要な語義と不要な語義が存在するため、厳密には、これらを区別する必要がある。この問題についても、後に検討する必要があると考えている。

参考文献

- [1] Güneş Erkan and Dragomir R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, Vol. 22, pp. 457–479, 2004.
- [2] A. S. Hornby and Michael Ashby, editors. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, 2005.
- [3] 笠原要, 佐藤浩史, フランシスボンド, 田中貴秋, 藤田早苗, 金杉友子, 天野成昭. 「基本語意味データベース: Lexeed」の構築. 情報処理学会第159回自然言語処理研究会, pp. 75–82, 2004.
- [4] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford University, 1998.
- [5] Paul Proctor, editor. *Longman Dictionary of Contemporary English*. Longman, 2005.