語釈文記述のための日本語定義語彙の構築に関する一考察

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
{noro,tokuda}@tt.cs.titech.ac.jp

1 はじめに

存在するすべての語の意味を，限られた少数の語 (定義語彙) で記述し，体系化することは，言語教育や言語処理研究にとって重要である．例えば，児童や外国人がその言語を学習する際，定義語彙中の語の意味を理解すれば，それ以外のすべての語の意味を理解できる．また，語義情報を利用した自然言語処理においても，すべての語を効率的に体系化することが可能となる．

英語では，Longman Dictionary of Contemporary English (LDOCE) [8] や Oxford Advanced Learner's Dictionary (OALD) [4] 等の辞書において，すべての語釈文は 2,000 語から 3,000 語の定義語彙で記述されている．実際，これらの辞書は英語教育や言語処理研究に広く利用されている．

一方，日本語では，定義語彙を明確に定めた上で作成された国語辞書は存在しない．定義語彙に類似した概念に基本語彙があり，基本語彙に関する研究は，これまでに数多く行われている [6]．しかし，それらは主に児童や日本語を学習する外国人が知っている (べき) 語に関する研究であり，すべての語義を定義するために必要な語というより，むしろ，学習者が利用する辞書に見出し語として登録すべき語を選定することを直接の目的とした研究である．

従来研究における，辞書に見出し語として登録すべき「基本語彙」と，我々が注目する「定義語彙」は，類似すれどもまったく同じものにはならないと考えている．基本語彙は主に新聞・雑誌記事，日常会話，教科書等で使われる表現を対象として選定が行われるが，定義語彙は語釈文で使われる表現を対象とするため，「特に」,「〜の略」,「転じて (〜の転)」,「物事」等の語釈文特有の (表現に使われる) 語が基本語彙以外に必要になる．逆に，同義語や類義語は，語釈文を記述する際にはいずれか 1 つだけあれば十分である．例えば，「使う」,「用いる」,「使用する」,「利用する」の意味は類似しているため，この中から 1 つだけを定義語彙に採用すれば，それ以外は不要である．

我々は，既存の国語辞書の語釈文に使用されている語の分布をもとに，語釈文記述に必要な日本語定義語彙を構築することを検討している．本稿では，その手法を紹介し，評価実験によりその有用性，問題点について考察する．

2 日本語定義語彙の構築手法

国語辞書の語釈文に使用されている語の分布をもとに定義語彙を構築する．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，定義する語と定義される語の関係を表したグラフ (単語参照グラフ) を構築
2. 構築したグラフをもとに各語にスコア付けを行い，スコアの高い語を定義語彙に採用

2.1 単語参照グラフの構築

単語参照グラフは各語をノードとした有向グラフであり，見出し語からその語釈文中の各語と自身に対して有向枝をはることにより構築する (図 1．自身に対する有向枝は省略している)．多品詞語 (名詞の「余り」と副詞の「余り」) や同じ表記で読みが異なる語 (「小節」を「こぶし」と読む場合と「しょうせつ」と読む場合) を区別するため，語釈文中の表記だけでなく，品詞と読みも利用する．また，本研究では助詞，助動詞，固有名詞を定義語彙構築の対象から除外した．

図 1 の例は重み無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できる．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，等) にもとづくヒューリスティッ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を構築できる．

2.2 各語へのスコア付け

我々は，以下の仮定をもとにスコア付けを行う．

1. より多くの語の語釈文中で使用される語は定義語彙にふさわしい
2. 定義語彙にふさわしい語の語釈文中で使用される語は，使用されない語よりも定義語彙にふさわしい

言語	一定のきまりに従い音声や文字・記号を連ねて，意味を表すもの．
語彙	ある範囲の，あるいは広く一言語についての，語の総体．
言葉	意味を表すため，口で言ったり字に書いたりしたもの．
単語	意味の一まとまりを表し，文法上の働きをもつものとしての，言語の最小単位．

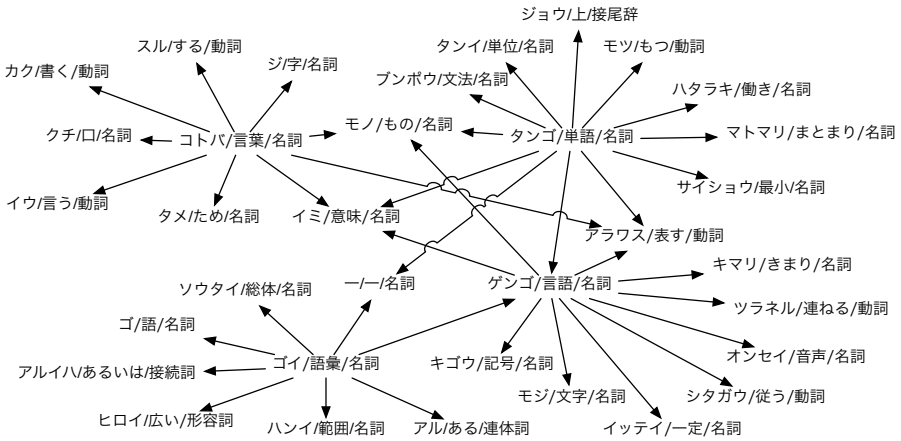

図 1　単語参照グラフ

1 つ目の仮定だけであれば，単語参照グラフにおいて参照を受ける枝 (自身に向かってくる枝) の数をそのままスコアとすれば良い．しかし，その語を語釈文中に使用する語の数だけでなく，どのような語から参照されているかも重要な要素であると考え，2 つ目の仮定を置いている．この 2 つの仮定を反映させるため，PageRank [7] や LexRank [1] の考え方を採用し，単語参照グラフに対応する隣接行列の固有値が 1 の時の固有ベクトルの各要素の値を各語 (各ノード) のスコアとする．

3　評価実験

提案手法によりどのような語が定義語彙として選ばれるかを確認するための実験を行った．データは岩波国語辞典コーパス [2] を使用した．岩波国語辞典コーパスは，岩波国語辞典第 5 版の全データに GDA タグ [3] 等の XML タグを付与したものであり，形態素，構文構造等のタグを含んでいる．本実験では，構文構造の情報は使用せず，形態素情報のみを利用する．

3.1　準備

岩波国語辞典コーパスには，GDA タグの他に岩波国語辞典特有のタグを含み，それにより見出し語，漢字表記，品詞，例文等の区別が容易になっている．本実験では，見出し語 (hd)，漢字表記 (orth)，品詞 (pos)，語釈文 (su)，語釈文中の各形態素 (n，v，ad 等) の情報を利用し，例文 (eg)，文法的解説 (gram)，対義語 (ant)，語源 (etym)，見出し参照 (sr) 等の補足的な情報は利用しない．また，語釈文中のかぎ括弧 “「」” で囲まれた文字列は引用的な表現が多く，定義語彙を考える上では適切ではないものが多いので，先に述べた補足的な情報を表すタグがついていなくても対象から除外する．

(例)　あかし【明石】「あかしちぢみ」の略．夏物の，ちぢみの絹織物．

提案手法の説明で述べなかった，辞書特有の問題点について具体的に述べる．

同一項目中に複数の漢字表記がある場合:　例えば，動詞「ひく」の項目では，「引く」，「弾く」，「曳く」，「牽く」，「挽く」，「碾く」，「轢く」，「退く」の表記が登録されている．このように，同一項目中に複数の漢字表記がまとめられている場合，単語参照グラフにおいても同一ノードとする．さらに，“「惹」も使う” のように語釈文中で別の表記の存在について言及している場合も同様に同一ノードとする．すなわち，同一項目中のすべての漢字表記は同一ノードとする (ただし，品詞が異なる場合は区別する)．

品詞変換:　語釈文中の各語には，コーパスアノテーション方針の品詞分類にもとづく品詞が付与されている．一方，見出し語の品詞 (pos タグで囲まれた部分) は岩波国語辞典が定める品詞分類にもとづく品詞である．この 2 つの品詞分類の違いを吸収するため，名詞，動詞，形容詞等，非常に粗い品詞分類を用意し，それぞれの品詞を変換する．

語釈文中と見出しでの表記の異なりの吸収: 「〜する こと」の「こと」は平仮名で表記することが一般的であり，語釈文中でも平仮名で表記されているが,「こと」の項目では漢字表記として「事」が記載されている．そのため，語釈文中に「こと」が出現した場合，それが「事」,「琴」,「古都」,「糊塗」のどれと同一であるかが分からない．その他に片仮名で表記されたり，送り仮名の違い，連濁による読みの変化も多く，これらをすべて別ノードとすると有用な結果が得られないと判断し，人手でノードの統合を行った．統合する基準は，(1) 平仮名や片仮名による表記，(2) 送り仮名の違い，(3) 連濁，の 3 つである．

以上の処理を行った結果，ノード数 70,778 個の単語参照グラフを半自動的に構築できた (1 項目あたり平均約 10 語) [*1].

スコア付け (固有ベクトル計算) は，Erkan ら [1] と同様，べき乗法により行った．また，random walk のための減衰係数 (damping factor) は 0.15，許容誤差 (error tolerance) は 10^{-4} とした．

3.2 実験結果

スコア付けによる上位 40 語を表 1 に示す．ただし，スコアは，1 位を 1.000 としている．

表 1 より,「ある (動詞，連体詞)」,「こと」,「もの」等のごく一般的な語だけでなく,「意」,「転」,「略」,「物事」,「特に」のように語釈文特有の (表現に使われる) 語も上位に現れることが分かる[*2].

逆に,「そういう」や「A」は定義語彙としては不適切であるように思われる．この 2 語は，岩波国語辞典には見出し語として登録されていないものである．語釈文中のみに出現し，見出し語として存在しない語は，単語参照グラフにおいて自身から他の語へ向かう枝が存在しないため，スコアが他に比べて高くなる傾向がある．

語釈文中のみに出現し，見出し語として存在しないと判断された語には，以下のようなパターンがある．

品詞の不一致: コーパスアノテーション方針の品詞分類と岩波国語辞典が定める品詞分類の違いを吸収するため，粗い品詞分類を用意し，それぞれを変換しているが，それでも一致しない場合がある．特に，語釈文中で接頭辞や接尾辞とされているものは，見出し語としては名詞や漢字母 (漢語の造語成分) であることが多い (語釈文中に「接尾語的に」のように記述されていることもある).

形態素区切りの不一致: 複合語を 1 語とするか 2 語以上に分割するかの違いにより，見出し語として存在しないと判断される場合がある．例えば，語釈文中では「かぎまわる」が 1 語となっているが，辞書の見出し語として「かぎまわる」は存在しない．表 1 の「そういう」はこれにあたる．

データ不具合による除外: コーパスデータの一部にフォーマットやアノテーションの不具合があり，それ

表 1 上位 40 語

	スコア	読み	表記	品詞
1	1.000	ある	有る, 在る, ある	動詞
2	0.7441	てき	的	接尾辞
3	0.6304	い	意	名詞
4	0.5806	ある	或る, ある	連体詞
5	0.5789	こと	事, こと	名詞
6	0.5048	する	為る, する	動詞
7	0.3710	てん	転	名詞
8	0.3113	もの	物, 者, もの	名詞
9	0.2377	一	一	名詞
10	0.2288	その	其の, その	連体詞
11	0.2002	ほう	方, ほう	名詞
12	0.1723	りゃく	略	名詞
13	0.1632	たつ	立つ, 建つ, たつ	動詞
14	0.1612	いる	居る, 処る, いる	動詞
15	0.1604	ひと	人, ひと, ヒト	名詞
16	0.1590	また	又, 復, 亦, また	接続詞
17	0.1532	つかう	使う, 遣う, つかう	動詞
18	0.1305	いく	行く, 往く, いく	動詞
19	0.1273	なる	成る, 為る, 生る, なる	動詞
20	0.1256	いう	言う, 云う, 謂う, いう	動詞
21	0.1146	どう	同	連体詞
22	0.1134	ものごと	物事, 物ごと, ものごと	名詞
23	0.1117	ご	語	接尾辞
24	0.1010	それ	其れ, それ	代名詞
25	0.09959	いがい	以外	接尾辞
26	0.09888	そういう	そういう	連体詞
27	0.09766	とき	時, 刻, とき	名詞
28	0.09362	二	二	名詞
29	0.08965	はっきり	はっきり	名詞
30	0.08926	いえる	言える	動詞
31	0.08869	とう	等	接尾辞
32	0.08866	じょうたい	状態	名詞
33	0.08142	けい	形	接尾辞
34	0.08113	えい	A	名詞
35	0.08038	とくに	特に, とくに	副詞
36	0.07861	あらわす	表す, 現す, 顕す, 著す, あらわす	動詞
37	0.07801	または	又は, または	接続詞
38	0.07711	いいあらわす	言い表す	動詞
39	0.07592	てん	点, てん	名詞
40	0.07540	ご	語	名詞

[*1] 使用したコーパスデータの一部にはフォーマットやアノテーション上の不具合があり，それらはすべて除外した.

[*2] 「意」,「転」,「略」はそれぞれ「〜の意」,「〜の転」,「〜の略」という表現で使われる.「〜の転」と似た表現に「転じて」があるが，動詞「転じる」も 67 位に入っている.

らは除外して実験を行った (脚注 (2)). 単語参照グラフ構築時には，ここで除外された語は見出し語として存在しないと判断される．

真に見出し語として存在しない: 例えば，「タオル地」の語釈文中に「輪奈 (わな)」という語があるが，これは見出し語としては存在しない．また，「書ける」，「説ける」等の可能動詞は，岩波国語辞典には見出し語として登録されていない．表 1 の「A」はこれにあたる．

データ不具合によるものは，将来，コーパスが修正されれば解決する．品詞，形態素区切りの問題は，単語参照グラフを構築する際に何らかの工夫が必要である．真に見出し語として存在しない語は，可能動詞以外は，一般的とは言い難い語がほとんどであり，単語参照グラフ構築時に除外しても問題ないと思われる．

4 関連研究

笠原らは，単語親密度により基本単語の選定を行っている [5]．これはアンケートにより馴染みのある語か否かを判定するものである．その結果，約 28,000 語を基本単語として選定し，そのうちの約 17,000 語を利用して語義を定義している．我々は，まだ具体的な定義語彙のサイズについて検討していないが，この語数は LDOCE や OALD で採用されている定義語彙の 2,000 語や 3,000 語と比較して多過ぎると考えている．また，馴染みがあるか否かと語釈文の記述に必要であるか否かは必ずしも同じではなく，実際に語釈文に使われている語をもとに定義語彙を構築する必要がある．

5 おわりに

本稿では，語釈文を記述するために必要な日本語定義語彙の構築手法を提案した．我々は，提案手法による上位の語を単純に取り出してそのまま定義語彙とすることができるとは考えていない．最終的には，人手によるチェックが必要となるが，その判断基準の 1 つとして提案手法が使えると考えている．

定義語彙により，すべての日本語の語義を定義できれば，言語処理研究や日本語教育にとって有用である．また，将来，新語の語義を定義する際にも，定義語彙を利用することにより語釈文の記述方法をコントロールできる．さらに，ある語釈文中では漢字で表記されているが，別の語釈文中では平仮名で表記されている等，表記のゆれに関する問題もある．定義語彙を構築する際に語釈文中での表記を決定しておくことで，この問題を解決できると考えている．

今後の課題を以下に挙げる．

- 本稿では，定義語彙のサイズについて言及していない．英語の場合，2,000 語から 3,000 語で語釈文を記述できるが，日本語でも同様のサイズで可能かどうかを検討する必要がある．
- 今回は，すべての語釈文中のすべての語を利用して単語参照グラフを構築したが，一般に，語釈文は，第 1 文が最も基本的な (語義に関する) 情報であり，後になるにつれて補足的な情報 (その語にまつわる歴史的背景，文法，語法等) になる．この性質を利用し，後の方の語釈文を無視してグラフを構築することにより，より定義語彙にふさわしい語が上位に入りやすくなると予想される[*3]．
- 定義語彙を構築できたら，実際にすべての語釈文をその定義語彙で記述しなおすことを考えている．その際，どの語を使って語義を定義するかだけでなく，どのような表現を利用すべきか (語釈文記述のための定義表現) についても検討する必要がある．

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] 橋田浩一. 岩波国語辞典のアノテーション — 照応・共参照・項構造 —, 2006. http://www.i-content.org/rwcDB/iwanami/doc/tag.html.

[3] 橋田浩一. GDA 日本語アノテーションマニュアル, 2005. http://i-content.org/gda/tagman.html.

[4] A. S. Hornby and Michael Ashby, editors. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, 2005.

[5] 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.

[6] 国立国語研究所. 日本語基本語彙 — 文献解題と研究 —. 国立国語研究所報告 116. 明治書院, 2000.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford University, 1998.

[8] Paul Proctor, editor. *Longman Dictionary of Contemporary English*. Longman, 2005.

[*3] 岩波国語辞典では，記号 "▽" の後に補足的な情報を記述しているが，コーパスではこの記号自体が削除され，補足説明を表すタグ (rem) も付与されていないため，区別できない．