クラウドソーシングにより収集した語釈文を基にした単語の基本度推定

岡久 太郎* 久保 圭† 水谷 勇介† 河原 大輔† 黒橋 禎夫†
*京都大学 人間・環境学研究科 †京都大学 情報学研究科

okahisa.taro.35v@st.kyoto-u.ac.jp, kaykubo.ktu@gmail.com,
{mizutani, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

1 はじめに

我々は、未知の概念を理解する際、既知の概念を組み合わせることによって、その新たな概念を把握している。例えば、「投手」が表す概念を理解するためには、その前提として「野球」「投げる」「選手」等の概念を知っている必要がある。このような単語の表す概念同士の依存関係は、特定の単語の語釈文中にどのような単語が使用されるかという定義-被定義関係として考えることができる。「投手」であれば、「野球で打者にボールを投げる選手」という語釈文中に見られる単語「野球」「打者」「ボール」「投げる」「選手」が定義語として被定義語「投手」を定義している。このように、定義語は被定義語よりも通常、基本的な単語として考えることができる。

これまで、単語間の定義-被定義関係を考慮した研究として、辞書の語釈文を用いて、シソーラスの構築や単語の概念的基本度を数値化した研究が存在した。しかしながら、これらの研究は、辞書の語釈文を用いているため、辞書特有の文体が反映されてしまったり、多義語における中心的語義と周辺的語義の区別が難しい等の問題がある。

本稿では、クラウドソーシングによって収集した語釈文を基に、語彙項目間の定義-被定義関係を反映した、単語の概念的基本度を表す指標である**定義スコア (definition score)** を提案する。

2 先行研究と本研究の位置付け

これまで、単語の概念的基本度はシソーラスという形で開発されてきた。自然言語処理分野では、日本語辞書の見出し語 (=被定義語) と語釈文を用いて、シソーラスを構築する試みがなされている (e.g., [1])。このような研究は、大規模な語彙に対して、概念的上下関係を見いだせるものの、シソーラスの性質上、異なる枝分かれ先に位置する単語全ての概念的基本度を比較することは難しい。そのため、あらゆる語の基本度を検討するためには、シソーラスとは異なる形で単語の表す概念同士の基本度を考慮する必要がある。

一方、同じく辞書の語釈文をもとに、単語の基本度を数値化することを目指した研究も存在する。野呂らは、辞書学において提案されている、あらゆる単語の語釈文を記述できる語彙 (=定義語彙) を選定するために、以下の 2 つの仮説を立て、日本語辞書の見出し語と語釈文を基に、単語毎にスコアを付与した [3]。

(1) a. 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
b. 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

この研究では大規模な語彙に対して、定義-被定義関係を反映したスコアを与えていると言える。しかし、高スコア語の中に「転」(70,778 語中スコア順位 7 位) がある等、辞書の語釈特有の表現が高いスコアを得てしまうという問題も抱えている。

本研究では、先行研究のように辞書を用いるのではなく、クラウドソーシングを用いて、複数人から語彙項目の語釈文を収集し、そのデータを用いて、語の表す概念の基本度を算出する。本提案手法は、複数の一般人が記述したデータを用いることで、これまでの辞書の語釈文を基にした単語の概念的基本度推定よりも、我々の直感に近い指標を提供することが可能であると言える。次節以降では、定義スコアの具体的な算出方法を述べる。

3 方法

本節では、クラウドソーシングを用いた定義スコアの算出方法について詳述する。まず、3.1 節でクラウドソーシングによる語釈文の収集方法と結果の処理方法について説明する。次に、3.2 節において、具体的な定義スコアの算出方法について述べる。

3.1 クラウドソーシングによる語釈文の収集と結果の処理方法

本研究では、語釈文の収集において、Yahoo!クラウドソーシング[1]を用いた。クラウドソーシングは 9 回に分けて実施し、[2] の単語難易度を参考に選出した 11,936 語について各 10 人からの語釈文を集めた。1 人の作業者は、1 回のクラウドソーシングごとに最大 10 タスクを行うことが可能であり、1 タスクあたり 5 単語の語釈文[2]を書くことが求められた。9 回のクラウドソーシングを通して、合計 3,024 人がタスクに参加した。得られた語釈文の一例を下に挙げる。

(2) 「見通し」の語釈文

- 先の様子を想像する。
- この後どうかるかの予想。
- およそ、多分、そうだろう、という事。
- これから起こるであろう状況を推測する道。
- 先の見解。
- みとおすこと。
- 先の段取り。
- 未来。
- 予想されること。
- 視界。

今回のタスクは、自由記述式のものであるため、コピーアンドペーストを繰り返すようなチート回答を予め排除するためのチェック問題を設定することが難しい。そこで、今回は収集したデータの整備において以下の処理を行った。

(3) a. 同一の作業者 ID の作業者が複数の単語について同一の回答を記述した場合、それらの回答を削除した。

b. 呈示した単語と同一の単語のみを記述している場合、その回答を削除した。

また、回答中には、チート回答ではないものの、基本度の推定において問題となる以下のような語釈文が存在した。

(4) **庵 (いおり)** 日本家屋の離れ。お茶室など。
享受 (きょうじゅ) 受け入れて、楽しむこと。「自由を○○する」「情報を○○する」など。

(4) の回答は、第 1 文が「庵」の語釈となっており、第 2 文の「お茶室など」は「庵」の具体例である。また、(4) の回答も、「享受」の語釈は第一文のみであり、第 2 文は被定義語の使用例である。そのため、第 2 文以降の文末に「など」「等」と記述されている文は削除し、それ以前の文を使用することとした。

以上の処理を行った上で、各語釈文について形態素解析器 Juman++[3] を用いて、形態素解析を行った。その上で、人名・地名を除いた名詞、動詞、形容詞、副詞に限定し、被定義語 11,936 語に含まれているもののみを抽出した。その結果、全 111,215 の語釈文に含まれるトークン数 291,776 の定義語が得られた。

3.2 定義スコアの算出方法

本研究では複数人に記述させた語釈文をデータとして使用するため、1 つの被定義語の語釈文内に同一の語が複数出現することが少なくない。提案する定義スコアは、単語間の相対的な概念的基本度を数値化したものであり、基本的には、野呂ら [3] における仮説 (1) を本研究においても想定しているが、野呂ら [3] で用いられている単語参照グラフ[4]に対応した隣接行列では、そのような頻度の多寡を考慮することは難しい。

そこで、本研究では、形態素解析によって得られた各々の定義語とそれに対応する被定義語のペアである被定義語-定義語ペア 291,776 組[5]について、以下のようなアルゴリズムを適用した。

(5) 定義スコア算出アルゴリズム

1. 被定義語 w_i の定義スコア $d(w_i)$ それぞれに、0.0 から 10.0 の範囲でランダムに初期値を与える。
2. 被定義語 w_i と定義語 w_j の被定義語-定義語ペアの各組について、$d(w_i) < d(w_j)$

[1] https://crowdsourcing.yahoo.co.jp/
[2] タスクの説明画面では、「語釈文」という用語は使用せず、「次のことばをわかりやすく説明してください」という指示によって、各語を説明させた。
[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++
[4] 辞書の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で得た単語参照グラフを構築し、それ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) のスコアとしている。
[5] 頻度の影響を考慮するために、同一ペアの重複を許している。

を満たさない場合、T を $d(w_i)$ から減じ、T を $d(w_j)$ に加える。(T の初期値は 5)

3. 上記の修正率が 6%を上回る場合は、T を 0.8 倍し、2 以降の手順を繰り返す。

本アルゴリズムにおいて、数の多い被定義語-定義語ペア、すなわち、ある被定義語の語釈文に多く用いられる定義語が、全体の修正率に関わっていると言える。これにより、ある単語の語釈文として多く得られた回答、すなわち、当該の単語の中心的な語義ほど、定義スコアの最終的な値に影響を及ぼすと言える。

4 結果

(5) のアルゴリズムを適用した結果、10 回の繰り返しで修正率が 6%を下回った。定義スコアの上位、中位、下位の単語とそれらの 98 億文 Web コーパスにおける頻度順位を表 1 に示した[6] 。また、定義スコアの値と頻度の常用対数をプロットしたグラフを図 1 に示す。算出された定義スコアから、以下のことが分かった。

(6) 定義スコアは、頻度のみでは散らばりを見せる複数の単語の同義・類義関係を反映している。

これは、定義スコアが複数人によって記述された語釈文に基づいた指標であることに起因する。すなわち、同じ単語の同じ語義であっても、回答者によって同義語・類義語を用いて、異なる記述をしている場合が多々ある。(2) の「見通し」であれば、「先の様子を想像する」と「この後どうなるかの予想」[7]は同一の語義を説明していると言えるが、ここに見られる「先」(12.16) と「後」(12.01)、「想像」(11.30) と「予想」(10.81) は同義・類義関係にある単語である (括弧内は各語の定義スコア)。このように、1 つの被定義語に対して、同義・類義関係にある複数の単語を定義語とすることができるため、頻度に関係なく、単語が表している概念に対応したスコアを与えることができたと考えられる。

具体例として、定義スコア 13.26 の「一人一人」が定義語としてどのように使用されているかを見たい。以下に、「一人一人」が語釈文中に使用されている被定義語と得られた語釈文の一例を挙げる。

(7) **各々 (12.29):** ひとりひとり。
個人 (12.28): 一人一人。
個々 (12.10): 一人一人。
各自 (11.89): ひとりひとり。
互い (10.43): 一人一人。
各人 (10.16): 一人一人。
投票 (4.64): 何かを決める際に、それを決める人たちが、どうするかの候補を立て、一人一人、それぞれの意思について、自分の決めた候補についての意思を、何らかの集計ができる方法で、集める時の、集め方。
配分 (3.66): 一人一人の取り分。
人格 (3.44): ひとりひとりの人の性格。
配役 (3.01): 演劇など役柄のあるものを一人一人役をつける事。
配付 (2.77): 一人一人に配ること。
取り分ける (2.37): 一人一人に分け与える。
点呼 (2.42): 人員がそろっているかどうか、ひとりひとりの名を呼んで確かめること。

(7) を見ると、「各々」「個人」「個々」「各自」「互い」「各人」については、「一人一人」が単独で語釈文として記述されている。さらに、定義スコアに注目すると、やはりこれらの語はいずれも 10 を超えており、他の語とは大きく差が見られる。これらの事実から、「一人一人」とほぼ同義であると捉えられている語に関しては、「一人一人」と同程度の定義スコアを有しているのに対して、「一人一人」が概念説明の一部としてのみ機能している語に関しては、「一人一人」よりも低い定義スコアを有していることが分かる。

また、「一人一人」が単独で語釈文として使われている語でコーパス頻度が確認できる語の頻度順位は、「各々」5223 位、「個人」185 位、「各自」4737 位、「互い」6257 位、「各人」8970 位であることから、定義スコアは頻度のみでは散らばりを見せる複数の単語の同義関係をも反映したものとなっていることが分かる。

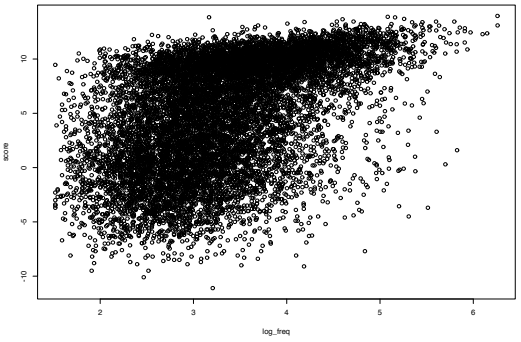

図 1: 縦軸に定義スコアを、横軸にコーパス頻度の常用対数を取った散布図

[6]今回、スコア順位と頻度順位について検討した単語は、11,936 語の中で、コーパスの形態素解析結果としても認められた 11,521 語に限定した。また、頻度順位はその 11,521 語を母数としている。

[7]実際の語釈文は (2) にあるように「この後どう “か” るかの予想。」とタイプミスが含まれているが、ここの議論では分かりやすさのため修正した。

スコア順位	単語	頻度順位	スコア順位	単語	頻度順位	スコア順位	単語	頻度順位
1	する	2	4000	敷設	9585	9000	肩車	11281
2	場所	186	4000	泊める	10006	9000	申	6174
3	事柄	6336	4002	業	1754	9000	暗躍	9430
4	違う	139	4003	改心	9732	9000	大敵	7556
5	何	106	4004	縦断	7775	9000	投書	11134
6	或る	734	4005	塩味	7722	9000	改訂	5514
7	方向	850	4006	発車	7892	9000	特訓	6854
8	行動	561	4007	最新だ	594	9000	灘	8179
9	人	15	4007	本箱	7816	9000	云々	4174
10	自分	28	4009	余地	5698	9000	肩身	9943
11	実行	1698	4010	滑稽だ	8735	9000	沙魚	10709
12	分かる	51	4011	苛々	2611	9000	配線	3980
13	やる	77	4011	細やかだ	5378	9000	渡航	7418
14	一人一人	2303	4013	分担	5253	9000	昴	7448
15	状態	156	4014	濃淡	8852	9000	亜鉛	6841
16	決める	441	4014	見栄	5582	9000	納め	8471
17	物	401	4014	作者	2031	9000	筆順	10256
18	持つ	55	4014	メロディ	5047	9000	駅前	2390
19	姿	671	4014	獲物	7653	9000	売れ行き	6135
20	自身	962	4019	皆無だ	6100	9000	花屋	8983
21	存在	406	4019	変哲	9612	9000	考察	3280
22	多い	63	4019	一睡	11125	9000	遺跡	3632
23	有る	1	4022	親戚	3872	9000	増産	8527
24	者	854	4023	樹脂	3252	9000	ザラザラ	8970
25	物事	3519	4023	決心	6409	9000	軽快だ	6058
26	出来る	25	4025	星空	5268	9025	烏	5616
27	出来事	679	4026	孤独だ	5625	9026	延焼	11281
28	多く	387	4027	磨く	2384	9027	本校	2666
29	事	70	4028	縮こまる	11110	9028	古風だ	9280
30	使う	35	4029	近代	4624	9028	秋分	10228
31	変わる	167	4029	打撃	3982	9028	漁業	7230
⋮			⋮			⋮		

表 1: 定義スコア順位と頻度順位

5 おわりに

本稿で提案した定義スコアは、これまでの辞書の語釈文をデータとした単語の基本度推定と異なり、クラウドソーシングにより専門家ではない一般人が記述した自然な語釈文をデータとした。また、複数人に語釈文を書かせることによって、中心的な語義ほど多くの人が言及するため、語義の中心性が頻度という形で反映された。さらに、同一の語義に関する語釈文においても、人によって異なる語彙項目を用いて説明することがあるため、同義関係にある複数の定義語について、被定義語との関係性を考慮することができた。

定義スコアは、Longman Dictionary of Contemporary English (LDOCE) [4] 等に見られる定義語彙選定の指標や外国人日本語学習者が優先して学ぶべき単語の選定等といった様々な応用可能性を有していると言える。また、今後は単語の基本度を利用した文章の専門性の推定等にも取り組んでいきたい。なお、語釈文および定義スコアについては公開予定である。

謝辞

本研究は京都大学と (公財) 日本漢字能力検定協会の研究プロジェクト「人工知能 (AI) による漢字・日本語学習研究」のもとで実施された。(公財) 日本漢字能力検定協会からの研究助成に感謝致します。

参考文献

[1] 正津康弘, 白井清昭, 徳永健伸, 田中穂積. 国語辞典の語釈文の解析と語義のシソーラスへのマッピング. 人工知能学会全国大会論文集, No. 13, p. 4 pages, 2001.

[2] 水谷勇介, 河原大輔, 黒橋禎夫. 日本語単語の難易度推定の試み. 言語処理学会第 24 回年次大会, pp. 670–673, 2018.

[3] 野呂智哉, 徳田雄洋. 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. 言語処理学会第 13 回年次大会, pp. 626–629, 2007.

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