意味関連辞書構築のための単語間関連度収集手法の検討

後藤 慎也　　鈴木 良生　　田添 丈博
鈴鹿工業高等専門学校　専攻科　電子機械工学専攻

1　はじめに

自然言語処理における文の意味解析のために、単語間の関係の知識が必要となる場合がある。例えば、比喩表現である「名詞＋のように＋形容詞＋名詞」という表現は、単語間の関係によって比喩性が異なる。それを判定するためには、名詞と形容詞の意味的な関係に関する知識が必要になる[1]。しかし、単語間の関係をまとめた「意味関連辞書」の作成には膨大なコスト（労力及び時間）が必要となってくる。また、格フレームや意味分類についての研究は多く存在する[2,3]が、単語間の関連の強さについて扱っているものはほとんどない。

本論文では、名詞と形容詞の関係性に着目することで、コーパスからの自動収集とアンケートによる手動収集を行い、それらの結果より比較・検討し、「意味関連辞書」を構築するための最適な収集手法について考察する。

2　意味関連辞書

図1のように関連があると思われる名詞と形容詞をそれぞれ結びつけ、その結びつきの強さを数値によって表し、辞書としてまとめたものを意味関連辞書と呼ぶ。また、矢印は関連の方向を示している。図1では名詞に対してどんな形容詞が関連しているかということを表している。

名詞と形容詞の組み合わせは膨大な量となるため、普通に一つずつ手入力していくと数年から十数年以上かかると考えられる。

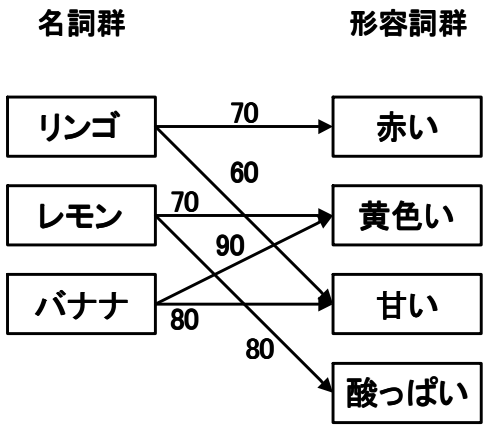

図1　意味関連辞書の例

3　コーパスからの自動収集

意味関連辞書を構築する際に最も問題となるのが、膨大なコストである。そこで、Web上から利用できる言語コーパスを利用し、自動収集の実験を行う。

3.1　実験方法

本実験には「少納言[4]」を用いる。少納言とは、国立国語研究所により構築されている、合計約1億500万語を持つ言語コーパスである「現代日本語書き言葉均衡コーパス」のための、Web上からの単語検索システムである。このシステムにより、コーパス中から名詞と形容詞の組み合わせが文脈中の前後40字以内で共起している件数を取得し、それぞれの組み合わせの関連度を測定する。また、関連度の計算にはJaccard係数を用いる。

膨大な量の名詞と形容詞のすべてについて実験を行うことは現実的には不可能である。

また、実験に用いる単語は、複数のジャンルの単語を用いることが望ましい。そこで、名詞と形容詞どちらも、5 種類のジャンルから 3 個ずつの単語を選び、15 個ずつを用いることとした。この実験に用いる名詞と形容詞を表 1 に示す。

3.2 実験結果

本実験において、名詞 15 個、形容詞 15 個より 225 組の AND 検索・OR 検索のデータを収集し、関連度を算出した。

収集データを名詞毎にまとめたときの、関連度上位 3 つを表 2 に示す。

3.3 考察

表 2 より、名詞毎に見たときに、最も関連度の高い組み合わせを見ていくと、「リンゴ」と「赤い」や「夏」と「暑い」など、多くの名詞において関連があると考えられる組み合わせが入力されていることが分かる。

しかし、「レモン」と「酸っぱい」に対して、「レモン」と「甘い」や、「冬」と「寒い」に対して、「冬」と「暑い」など、関連のある組み合わせと逆の意味の形容詞の組み合わせがいくつかあることが分かる。このような組み合わせの表れている表現として「寒い冬と暑い夏」といった対比や、「冬なのに暑い」といった表現が見られた。しかし、これらのような組み合わせの関連度が高い値となることは、比喩表現の判定への適用を考えたとき、あまり望ましくない。また、こういったデータは人手では入力されにくいことが考えられる。

しかし、多少の問題点は考えられるものの、大部分においては正しい結果が得られているといえる。

表 1　実験に用いる名詞及び形容詞

ジャンル	名詞		
果物	リンゴ	レモン	バナナ
花	タンポポ	アジサイ	チューリップ
季節	夏	秋	冬
鳥	ツル	ツバメ	スズメ
自然地形	山	川	海
	形容詞		
色	赤い	黄色い	白い
形	大きい	小さい	丸い
味	辛い	甘い	酸っぱい
印象	かわいい	かっこいい	かっこ悪い
温度	暑い	寒い	涼しい

表 2　少納言による実験結果

名詞	形容詞	AND検索	OR検索	関連度
リンゴ	赤い	38	5191	0.0073
	酸っぱい	10	1421	0.0070
	かわいい	15	3967	0.0038
レモン	酸っぱい	9	1547	0.0058
	黄色い	9	2520	0.0036
	甘い	10	3909	0.0026
バナナ	甘い	26	3619	0.0072
	黄色い	12	2243	0.0053
	白い	16	13426	0.0012
タンポポ	黄色い	9	1360	0.0066
	白い	11	12545	0.0009
	丸い	1	1160	0.0009
アジサイ	黄色い	11	1501	0.0073
	白い	13	12686	0.0010
	赤い	1	4336	0.0002
チューリップ	赤い	12	4358	0.0028
	黄色い	2	1543	0.0013
	白い	5	12727	0.0004
夏	暑い	539	19171	0.0281
	寒い	111	20210	0.0055
	白い	162	29717	0.0055
秋	白い	112	28065	0.0040
	暑い	50	17958	0.0028
	赤い	54	19761	0.0027
冬	寒い	408	12932	0.0315
	暑い	65	12664	0.0051
	白い	89	22809	0.0039
ツル	黄色い	4	2291	0.0017
	寒い	6	3918	0.0015
	白い	15	13467	0.0011
ツバメ	黄色い	1	1426	0.0007
	暑い	1	2444	0.0004
	白い	4	12610	0.0003
スズメ	小さい	7	7298	0.0010
	赤い	3	4437	0.0007
	黄色い	1	1614	0.0006
山	白い	610	113255	0.0054
	大きい	308	110952	0.0028
	小さい	207	108161	0.0019
川	白い	285	71806	0.0040
	大きい	159	69327	0.0023
	小さい	122	66472	0.0018
海	白い	410	79969	0.0051
	大きい	262	77512	0.0034
	小さい	169	74713	0.0023

4　アンケートによる手動収集

手動による収集として不特定多数からデータを少数ずつ得るような「アンケート」による実験を行った。

4.1　実験方法

入力者の負担を極力減らすため、いくつかの入力方法を検討した結果、お題となる名詞を提示し、ランダムに表示される 5 つの形容詞から選択してもらう形式が最も入力を行いやすかったため、この方法を採用した。

本実験では、スキップの選択肢を用意し、形容詞あるいはスキップを選択することで、名詞と形容詞がランダムに変更される。ただし、入力数が増えるのは、形容詞を選択したときだけである。既定の入力数に達するとアンケートは終了する。また、同じ実験中に同じ組み合わせは複数回入力できないようにしている。一度「リンゴ」に対して「赤い」を入力すると、次に「リンゴ」が出たときに選択肢として「赤い」は出ないようになっている。

実験協力者には入力の際の基準として以下のように示した。

1. 不適切な組み合わせ（関連度がまったくない）しかなければスキップ
2. 適切な組み合わせ（少しでも関連がある）の中で最も関連があると考えられる組み合わせを選ぶ

入力回数を関連度とし、実験に用いる単語はこれまで用いたものと同様の、表 1 に示す名詞と形容詞 15 個ずつの単語である。

5 人の実験協力者にそれぞれスキップを除いて 50 個ずつ入力を行ってもらうこととした。

4.2　実験結果

本実験によって、5 人の協力者から 250 個の入力データを得ることができた。

収集データを名詞毎にまとめたときの、関連度上位 3 つを表 3 に示す。ただし、「*1」は示したもの以外に 1 つ、関連度が同値のものが存在することを示す。

表 3　アンケートによる実験結果

名詞	形容詞	関連度
リンゴ	赤い	4
	かわいい	3
	甘い*3	3
レモン	黄色い	4
	酸っぱい	4
	小さい	2
バナナ	黄色い	3
	甘い	3
	白い*1	2
タンポポ	黄色い	4
	小さい	3
	白い	2
アジサイ	かわいい	3
	白い	2
	寒い	2
チューリップ	黄色い	4
	小さい	4
	赤い*1	3
夏	暑い	4
	赤い	2
	かわいい*3	1
秋	赤い	3
	黄色い	3
	涼しい	3
冬	寒い	5
	白い	3
	涼しい	2
ツル	大きい	3
	寒い	3
	白い*2	2
ツバメ	小さい	5
	かっこいい	4
	かわいい	3
スズメ	かわいい	4
	小さい	4
	丸い*2	2
山	涼しい	5
	大きい	4
	かっこいい*2	2
川	小さい	4
	涼しい	4
	寒い	3
海	大きい	4
	かっこいい	3
	暑い	3

4.3 考察

表 3 より、「リンゴ」における「赤い」や、「タンポポ」における「黄色い」、「夏」おける「暑い」、などのような、最も関連があると考えられる組み合わせが、過半数の入力者が選択していることが分かる。それ以外の組み合わせも、関連のあるものばかりであることが分かる。また、「冬」と「暑い」といった、正しい組み合わせと逆の意味の組み合わせが上位に来ていないことからも、人手による入力は表 2 と比較しても適切な結果であることは明らかである。

しかし、アンケート方式はコーパスからの自動収集に比べ、データの収集効率が悪い。これは、一人あたりの入力数を増やせば、入力者毎の負担が増え、入力者数自体を増やしにくくなってしまうためである。入力者のストレスを軽減するとともに、多くの入力者を集めることができ、収集効率を向上させるような工夫が必要となる。

5 収集手法の比較

自動収集と手動収集のそれぞれの特徴についてまとめる。

自動収集は多くのデータを短時間で収集することができ、収集効率が高い。データとしては、収集の際の粒度を細かくしていくことで大部分は正しいデータを得られるが、3 章の考察のように問題のあるデータも収集される。

手動収集は、それぞれ人が考えて入力を行うため、入力者が真面目に入力を行っている限りでは、正しい入力が行われる。しかし、入力に時間が掛かる、協力者が必要となる等、収集効率が低い。

それぞれの利点を生かし、欠点を補うために、自動収集と手動収集を組み合わせて使うことを考える。自動収集によって得られたデータを、アンケートの初期データとし、アンケートの結果によって、自動収集による収集データの問題点を改善できるのではないかと考える。また、アンケートにゲーム性を持たせることで、より収集効率を上げられると考える。

6 おわりに

本論文では、意味関連辞書の構築手法をデータの収集実験によって検討し、自動収集と手動収集の結果の比較検討を行った。今後は、より収集効率を上げることができるような、入力インターフェースとしてのゲームを考案し、試作する必要がある。

参考文献

1. 田添丈博, 椎野努: 比喩表現に属性が明示された場合の比喩性に与える影響とコンピュータモデルの検討　言語処理学会第 17 回年次大会 F4-4 (2011)
2. 村本英明, 鍜治伸裕, 吉永直樹, 喜連川優: 意味カテゴリに基づく語義曖昧性解消における Web 資源の活用について, 情報処理学会論文誌 Vol. 51 No. 10 pp.1234-1242 (2010)
3. 濱田慧, 笹野遼平, 柴田知秀, 河原大輔, 黒橋禎夫:分布類似度を用いた大規模格フレームの自動構築, 言語処理学会第 14 回年次大会　D3-5 (2008)
4. KOTONOHA「現代日本語書き言葉均衡コーパス」　少納言 http://www.kotonoha.gr.jp/shonagon/