ゲーム入力の収集による意味関連辞書の自動構築

後藤 慎也　　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
椎野 努
愛知工業大学　情報科学部　情報科学科

1　目的

自然言語処理における文章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。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（労働力及び時間）が必要となってくる。

本研究では、形容詞と名詞の関係性に着目することで、効率良く単語間関係のデータを収集するための方法を検証し、「意味関連辞書」を自動構築することを目的とする。

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、その結びつきを辞書としてまとめたものを意味関連辞書と呼ぶ。

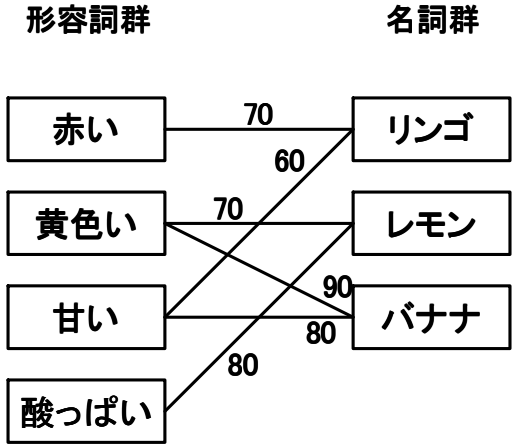

図1 意味関連辞書の例

意味関連辞書は、ある形容詞（名詞）からある名詞（形容詞）への関連性の度合（図1における線上の数値）に関するデータによって構成される。

この辞書の作成方法として、Web上に存在する文章よりデータを収集する方法が考えられる。しかし、この方法では適切なデータを収集することができない。例えば、「赤いリンゴ」のように形容詞と名詞が互いに連想されるようなものは少なく、「青いリンゴ」などの特殊の場合のほうが、より話題に上がりやすく、実際の関連性に比べて高くなってしまう。したがって、Web上からのデータの自動収集は適切な方法ではなく、データ収集には人による手入力の必要があると考える。

ところが、名詞・形容詞の組み合わせは膨大な量があり、手入力のコストはとても高いものとなる。

そこで本研究では、ゲームを利用することによって、データ入力に際してのストレスを緩和し、効率よく辞書を構築できるようなシステムを提案する。

3　連想ゲーム

3.1　概要

辞書構築のためのゲームとして、以下のような連想ゲームを作成した。

1. コンピュータがお題となる形容詞を提示する。
2. ユーザは、提示された形容詞より連想される名詞を入力する。
3. ユーザが名詞を入力、あるいはスキップすると、コンピュータは次のお題を提示する。
4. 制限時間が経過したら終了する。
5. 入力単語数に応じた得点が表示される。

このゲームにより、それぞれの形容詞での名詞の入力回数より、関連性のデータを収集することができる。図 2 に実行画面を示す。

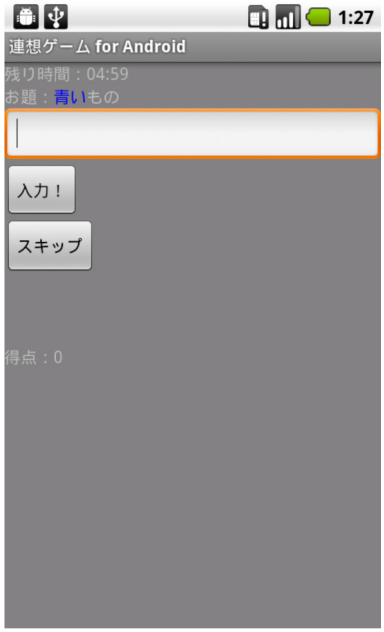

図 2 実行画面

6 種類の色を表す形容詞を、お題となる形容詞に使用し収集実験を行った。

3.2　実験結果

13 名の実験協力者より、197 パターンのデータを収集できた。表 1 から表 6 に収集されたデータを示す。但し、一度しか入力のなかった単語は省略する。

表 1　6 種類の色の形容詞での収集データ

形容詞　「黒い」		形容詞　「青い」	
単語	入力回数	単語	入力回数
髪の毛	5	海	13
炭	3	空	8
ゴキブリ	3	ブルーハワイ	2
海苔	2	地球	2
スーツ	2	信号	2
ゴマ	2		
夜	2		
鉛筆	2		
闇	2		
形容詞　「緑色の」		形容詞　「赤い」	
単語	入力回数	単語	入力回数
草	6	トマト	8
葉	6	血	6
キュウリ	6	林檎	5
木	5	太陽	3
森	3	リンゴ	3
カエル	2	火	3
ほうれん草	2	ポスト	3
ぴーまん	2	りんご	3
葉っぱ	2	消防車	2
ピーマン	2	いちご	2
		ち	2
		唐辛子	2
		とまと	2
形容詞　「白い」		形容詞　「黄色い」	
単語	入力回数	単語	入力回数
雲	8	レモン	7
雪	8	バナナ	6
紙	4	ピカチュウ	3
歯	4	ヒマワリ	2
ご飯	3	みかん	2
ホワイトボード	2	キリン	2
大根	2	チーズ	2
		レモン	2
		信号	2

3.3　考察

結果より、「海」「トマト」などの馴染みのある単語はより入力されやすくなっていることあわかる。しかし、ユーザにとって咄嗟に思い浮かばないような単語はほとんど入力されないため、データを収集できなくなる。また、入力文字数が多い単語は、ゲームの性質上入力されにくくなる。

これら問題点を解決するために、ユーザからの入力を、こちらの提示した単語を選択させる、という方法を用いることとした。

4　選択ゲーム

4.1　概要

単語を限定した場合でも、正しく単語間の関連が現れるかを確かめる実験を行った。

実験方法は、形容詞・名詞それぞれ 20 個ずつ用意し、それらの内の 5 個ずつをランダムに画面上に表示。その中から、関連のある名詞・形容詞の組を一組選ぶという方法で 100 組のデータを採った。

単語は、形態素解析エンジン「MeCab」の有している辞書より、形容詞・名詞ともに、辞書中に含まれている情報である頻出度順に 100 個の単語を抜出し、その中から一般的である単語を手作業で 20 個ずつ選んだ。

また、提示された単語内でどうしても組み合わせを作れない場合のために、スキップボタンを用意した。

図 1 に実行画面を示す。

図 3　実行画面

4.2　実験結果

この実験を異なる 3 人のユーザによって行った。その結果より重複した組み合わせを表 2 に示す。

表 2　重複した組み合わせ

形容詞	名詞	重複数
悪い	歴史	3
	政治	2
	経済	2
楽しい	スポーツ	8
	小学校	7
	野球	6
	学校	3
	高校	3
甘い	学校	4
	日本人	2
	小学校	2
強い	動物	5
	国家	4
	スポーツ	2
恐い	動物	3
厳しい	学校	3
	政治	2
	経済	2
	スポーツ	2
	家族	2
古い	教室	5
	歴史	3
	技術	3
	学校	2
	芸術	2
重い	国家	2
深い	歴史	5
	芸術	4
	技術	3
親しい	家族	6
	国家	2
多い	文書	3
	教室	3
	動物	2
難しい	経済	5
	文書	4
	行政	3
濃い	芸術	3
	経済	2
幅広い	芸術	3
	技術	2
	動物	2
忙しい	日本人	7
	行政	2
	スポーツ	2
面白い	スポーツ	4
	野球	4
	経済	3
	文書	3
	高校生	2
	技術	2
	芸術	2
優しい	日本人	5
	小学校	4
	家族	4
	技術	2
	行政	2
良い	家族	3
	日本語	2
	国家	2
	行政	2

4.3　考察

入力できる単語を限定することによって、連想ゲームと比べて「強い」「動物」などのような、連想はしにくいが限定的な条件において関連性があると考えられる組み合わせが多く入力されるようになった。

しかし、「古い」に対して一般的に関連が高いと考えられる「歴史」よりも「教室」の重複数が多いなど、意外な組み合わせの重複数が高くなっているものがある。これは、提示される単語によって結果が大きく左右されるためであると考えられ、データ数が増えれば改善されると考えられる。

また、形容詞と名詞を一度に複数個提示すると、視認性が落ち、組み合わせを探し難い。従って、形容詞は１つのみ提示し、名詞のみを見て組み合わせを探してもらうといった改善が必要となる。

5　まとめ

今回の選択ゲームでは、ゲームとしての楽しさ等はあまり考えられておらず、ゲームとしての、より効率的にデータを収集するという役割を果たすとは言えない。そこで、よりゲーム性の高いものとする方法として、入力方法をより直観的で簡単なものとする、スコア機能やランキング機能の実装などが考えられる。

また、別のゲーム案として「想起単語当てゲーム」がある。このゲームは、まずユーザに形容詞（名詞）を想起してもらい、辞書のデータベースより名詞（形容詞）提示していき、ユーザに関連度を判断してもらい、想起単語を推測するというゲームである。しかし、選択ゲームでの入力方式と少し異なるため検討が必要である。