

## 語の共起情報を考慮した感覚連想メカニズムに関する研究

米谷 彩† 渡部 広一† 河岡 司†

人間はあいまいな情報を受け取った場合にも、適宜に解釈し、適切に会話を進めたり、適切な行動をとったりすることができる。これは、人間が長年の経験により、言語における知識を蓄積し、その基本となる概念に関する「常識」を確立しているからである。コンピュータをより人間に近づけるためには、常識を踏まえて物事を扱うことができる「常識的判断システム」の構築が必要不可欠であると考えられる。

「感覚判断システム」は、知識ベースや国語辞書等から自動的に構築された概念ベースを用いてある語とその特徴である「感覚」についての常識判断を可能にしている。感覚判断システムの機能には、形容詞と名詞、または複数の形容詞の組み合わせから具体的な名詞を連想する「感覚的連想」がある。本研究では語の共起情報を用いることにより感覚的連想の信頼性を向上することを目指す。

## Sensuous Association Mechanism using Coincidence Information

Aya KOMETANI† Hirokazu WATABE† Tsukasa KAWAOKA†

When we humans receive uncertain information, we interpret it suitably, to understand what the speaker is trying to say. This is possible because we have "commonsense" concerning the word, which is built up from knowledge that is stored from long time experience. In order to understand the meaning, we think that the construction of a "Commonsense Judgment System" is necessary. "Sensuous Judgment System" can associate the characteristics of a word based on our senses, such as an apple is red, a sweet food is a candy, By using a knowledge base consisting of basic words and a concept-base, which is automatically constructed from dictionaries. "Sensuous Association", one of the functions of the "Sensuous judgment system" is that associates the noun from the combination of noun and adjective, or two or more adjectives. This paper proposes a mechanism to raise the reliability of "Sensuous Association" by using "Coincidence Information".

### 1. はじめに

コンピュータを人間にとってより使いやすいものにするには、人間同士が日常行っているコミュニケーションの仕組みをモデル化し、コンピュータと人間とのインタフェースにもこれを取り入れることが望まれる。

人間は主に言葉によって意思疎通を行う。言葉をただの文字列としてではなく、その背後にある様々な情報を暗黙のうちに捉えることにより、適切な意味理解と言動を可能にしている。これは、人間が日々の生活の中で言語に関する知識を豊富に蓄積し、それぞれの概念に対する「常識」を確立しているからである。コンピュータを人間のよりよいパートナーとするためには、常識を踏まえた理解を行う「常識判断システム」の構築が不可欠である。我々の研究室では既に、量や時間、感情など様々な常識判断を行うシステムが確立されている。

「感覚判断システム」は、ある語とその特徴である「感覚」についての常識判断を実現することが目的である。感覚は形容詞・形容動詞（以下、形容詞類）で表される。感覚判断システムの機能の一つに、名詞と形容詞類、または複数の形容詞類の組み合わせから名詞を連想する「感覚的連想」がある。本研究では、感覚的連想のメカニズムの提案、評価、検討を行い、語の共起情報を用いることによりその信頼性を上げることに成功した。

### 2. 概念連想メカニズム

感覚判断システムで利用する概念ベースや関連度計算を用いた概念連想メカニズムについて説明する。概念連想メカニズムによって、語と語の関連の深さを数値で定量的に評価することが出来る。

#### 2.1 概念ベース

「概念ベース」<sup>[3]</sup>は複数の国語辞書や新聞から自立語を抽出して自動構築した、概念を表す語のデータベースである。

ある概念  $A$  を、 $A$  と関連が強いと考えられる「属性」 $a_i$  と、属性の重要度を表す「重み」 $w_i$  の集合で定義する。概念  $A$  の属性数を  $n$  個とすると次のように表される。

$$A = \{(a_1, w_1), (a_2, w_2), \dots, (a_n, w_n)\}$$

ここで、 $a_i$  を概念  $A$  の一次属性と呼ぶ。 $a_i$  もまた概念ベースに登録された語であるため、 $a_i$  の属性も同様に導くことができる。これを概念  $A$  の二次属性と呼ぶ。一つの概念は  $n$  次属性までをその構成要素として持つ。

概念ベースには約9万語の概念が登録されている。

#### 2.2 関連度計算

「関連度」<sup>[2]</sup>とは、概念と概念の結びつきの度合いを定量的に評価した値である。

2つの概念  $A$ 、 $B$  間の関連度は各概念を二次属性まで展開し、重みを使った計算によって最も対応のよい一次属性の組み合わせを求め、それらの一致する属性個数を評価することにより算出する。これを「関連度計算」と呼ぶ。

関連度は0～1の実数値をとり、値が1に近いほど概念間の関係が深いことを示す。表1に関連度の例を示す。

表1 関連度計算の結果

概念A	概念B	概念Aと概念Bの関連度
花	桜	0.44738
花	飛行機	0.00443

†同志社大学 工学研究科

Graduate School of Engineering, Doshisha University

### 3. 五感判断と知覚判断

感覚判断システムは従来の「感覚判断システム」<sup>[7]</sup>と、「知覚判断システム」<sup>[10]</sup>をまとめたものである。本稿では従来の感覚判断を「五感判断」として説明する。

五感判断では、五感の刺激に関する形容詞類と名詞の関係を明らかにしている。人間がある語から想起する形容詞類は五感に関するものに限らない。知覚判断では、そのような五感以外の形容詞類を「知覚」と呼んで扱っている。例を表2に示す。

表2 五感と知覚

五感に関する形容詞類	知覚に関する形容詞類
赤い(視覚)	優しい
煩い(聴覚)	懐かしい
甘い(味覚)	惨い
臭い(嗅覚)	愛しい
熱い(触覚)	危ない
...	...

五感判断と知覚判断は扱う語は異なるが、その知識構造やメカニズムはほぼ同一である。

### 4. 感覚想起

感覚はある語に対して人間が抱く一般的な印象のことであり、その言葉の指す事物が持つ特徴でもある。例えば、会話に「火事」という単語が出てくれば人間は「恐ろしい、危ない」と感じ、「林檎」ならば「赤い、丸い」という特徴をもつものだとわかる。「花」に対して「汚い」という形容詞が常識であることは人間なら誰しも理解できるが、これは花が本来「美しい」ものとして認識されているためである。

感覚判断システムではある語とその特徴に関する常識的な連想を行う目的としている。システムの主な機能は、名詞の入力に対して適切な形容詞類を出力すること(感覚想起)と、名詞と形容詞の組み合わせまたは複数の形容詞類の組み合わせに対して適切な名詞を出力すること(感覚的連想)である。

感覚判断システムによって以下のような応答が可能になる。

「妹の結婚が決まりました」  
 -- 「それはおめでたいですね」  
 「冷たい飲み物が飲みたい」  
 -- 「ジュースはいかがですか？」

これは入力「結婚」から「めでたい」という感覚が、「冷たい」と「飲み物」から「ジュース」という名詞が出力されるために成立するものである。入出力の例を表3に示す。

表3 感覚判断の入出力の例

	入力	出力
感覚想起	太陽	明るい, 眩しい
	赤ちゃん	可愛い, 弱い
感覚的連想	美しい, 生き物	孔雀, 玉虫, 蝶
	丸い, 赤い	西瓜, 林檎

#### 4.1 システムの要素

あ感覚判断システムを実現するためには、語とその特徴である「感覚」の関係についての知識が必要となる。五感判断に関する知識は「五感知識ベース」、知覚判断に関する知識は「知覚知識ベース」に登録した。

2つの知識ベースをまとめて「感覚知識ベース」として扱う。

#### 4.1.1 感覚語・知覚語

名詞から想起される感覚は形容詞類で表現できる。例えば、林檎からは赤い・丸い・甘いなどが想起できる。五感に関する形容詞類のうち、非日常的なものを省き同義語を一つにまとめ、99語を「感覚語」として定義した。

感覚語については、五感の種類によって2つのレベルに分類した。例えば、赤いは五感「視覚」に關係しており、その中でも「色」である。よって「赤い」は「視覚:色」に分類される。このような五感と五感詳細の組み合わせは計15種ある。例を表4に示す。

表4 感覚語

単語	感覚語	五感	五感詳細
林檎	赤い	視覚	色
地球	丸い	視覚	形

知覚語については、その語の持つ印象がよいものか悪いものかによって大別した。例えば、「懐かしい」はよい印象の言葉であり、「ずるい」は悪い印象の言葉である。それぞれに「+」「-」という情報を与えた。例を表5に示す。

表5 知覚語

単語	知覚語	正負情報
故郷	懐かしい	+
詐欺	ずるい	-

#### 4.1.2 感覚知識ベース

五感判断を例に感覚知識ベースと感覚想起の仕組みを説明する。

感覚想起には名詞とその感覚である形容詞類の関係を記した知識が必要である。しかし、すべての名詞について、対応する形容詞類との論理関係を記憶させることは困難であり効率が悪い。そこで、日常的によく使われる語をシソーラスのリーフから選び出し、「代表語」として知識ベースに持たせた。そして、代表語から想起される形容詞類を「感覚」として登録する。また代表語を、シソーラスを基に木構造を持つ「分類語」に分類し、それらにも感覚を与える。五感知識ベースのイメージを図1に示す。

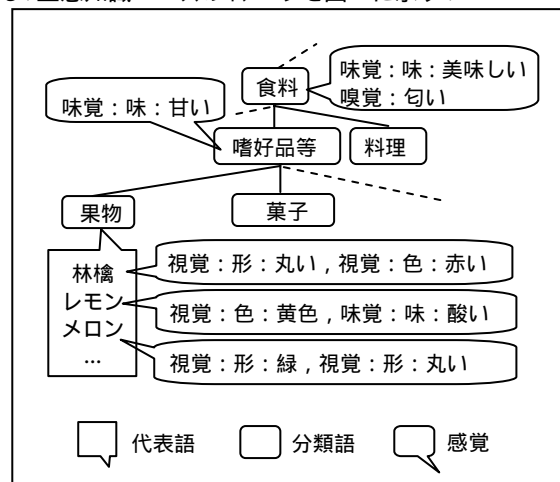


図1 五感知識ベース

知識ベースを木構造化することによって、上位ノードからの感覚の継承が可能となる。代表語固有の感覚と、分類語から継承された情報を含めて「感覚」として取得する。取

得た感覚の中に相反する意味の感覚が継承された場合は、下位にある情報を優先する。この五感知識ベースによって代表語と感覚、分類語と感覚の間に論理関係を持たせることができ、入力代表語であった場合にはその特徴である感覚を想起することが可能となった。感覚知識ベースの感覚情報はすべて人手によって作成され検証されたものである。したがってこの情報には間違いはなく、代表語に関する感覚想起は常に正確である。代表語、分類語のデータ数は表6のとおりである。

表6 感覚知識ベースのレコード数

	代表語	分類語
五感知識ベース	1294	179
知覚知識ベース	772	517

#### 4.2 未知語処理

感覚知識ベースを用いることによって、代表語に関する感覚の想起が可能となる。しかし、会話に出現する名詞の多くは感覚知識ベースに登録された代表語ではない。これらに対する感覚の想起も必要であり、それらの語に対する感覚判断が必要である。知識ベースに登録されていない単語を「未知語」と呼び、これらに関して感覚想起を行うメカニズムを「未知語処理」と呼ぶ。

感覚判断の未知語処理では2つの処理を行う。一つは、未知語が属する分類語を決定し、分類語の感覚を取得すること。もう一つは、概念ベースからその単語固有の感覚を取得することである。前者を「分類語特定」、後者を「感覚付加」と呼ぶ。

##### 4.2.1 分類語特定

未知語に対して感覚判断知識を利用するために、概念ベースや関連度計算から成る概念連想メカニズム(2章)を用いる。未知語について意味的な関連やその強さの度合いを評価し、知識ベース内の最も関連の強い分類語に代替することで感覚を取得する。

まず、未知語がシソーラスのどのノードに属するかを参照する。シソーラスに存在すると、属する分類語が即座に決定される。未知語がシソーラスに存在しない場合は、感覚知識ベースに登録されている代表語・分類語との意味的な関連の強さを関連度計算によって算出する。算出された関連度がある設定値を超えていれば、意味的に関連が強いと判断し、関連付ける。

この設定値は、概念ベース全体の中から無作為に抽出した概念Xに対して、「女性-婦人」などの極めて密な関係、「夕焼け-赤い」などの密な関係、「電車-眼鏡」などの疎な関係の3種類の語を概念ベースから人手で無作為に抽出し、4人が4人共にその関係が正しいと判断した200セット(計600)のデータを使用して得られたものである。ある概念Xと極めて密な関係、密な関係、疎な関係にある語の関連度の平均は、0.34、0.13、0.02であり、各関係の間には関連度に十分有意な差が見られた。

そこで、未知語と感覚知識ベース内の代表語との関連度が、極めて密な関係と密な関係の間値である関連度平均0.23を超えるものを、意味的に関連が強いと判断して対応付けた。対応付けられた単語が代表語であればその分類語を感覚知識ベースから取得し、対応付けられた単語が分類語であるとそれ自体を取得する。

ここで得られる分類語の集合から未知語の分類語を多数決によって決定する。多数決で決められない場合は、それらの内で最高の関連度のものを未知語の分類語とする(図2)。

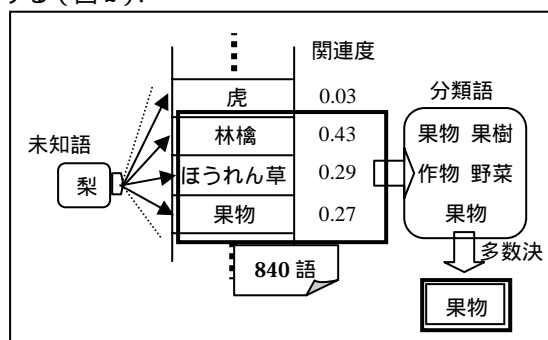


図2 分類語特定の例

未知語の分類語が決定されると、その分類語の感覚を感覚判断知識ベースから取得することができる。分類語特定処理が正確に行われると、未知語の分類語として適切な感覚を取得することができる。

##### 4.2.2 感覚付加

五感判断を例に、概念ベースを用いて単語固有の感覚を取得する処理について説明する。

概念ベースでは、概念をその語と関連の強いと考えられる語の集合である「属性」で意味を定義している。よって、属性に感覚語そのものや感覚語と関係の強い語が含まれると、それを感覚とみなすことができる。

未知語の一次属性、二次属性を調べると、感覚語そのものや、感覚を取得できそうな(代表語や分類語に該当する)語が含まれている場合が多いことがわかった。例えば表7では、概念「パンダ」の属性には感覚語である「大きい」や感覚に順ずる単語「白」や「黒」、さらに感覚「可愛い」を想起する「ぬいぐるみ」などが含まれている。

表7 概念と属性

概念	属性
パンダ	熊 動物 白 自然 生きる チベット 中国 ぬいぐるみ 足 黒 山 大きい 林 竹 ...

これら属性の情報を利用した幾通りかの処理により、未知語に対して単語固有の感覚を与えることができた。さらに、既知語(知識ベースに登録されている語)についても多くの感覚情報を得ることが可能になった。

#### 4.3 雑音排除

これまでの処理を経て得られた感覚情報には、間違いも含まれていた。特に、自動構築した概念ベースの情報を参照する感覚付加処理においては、雑音情報が多く取得された。この雑音を除去して出力の信頼性を上げる手法を次に説明する。

##### 4.3.1 信頼度合成方式

確からしさを「信頼度」と呼ぶ。信頼度は確率のような値であり、確率と同様の扱いが可能である。各処理の出力の精度を測定し、それを信頼度として設定した。

ある一つの感覚が複数の処理から取得された場合、つまり複数の信頼度を持つ場合はその信頼度を計算によって合成できる。入力に対して2つの知覚取得処理1,2から

信頼度 が得られたとき、この知覚の合成された信頼度は次のようになる。

$$P = p_1 p_2 / \{p_1 p_2 + (1 - p_1)(1 - p_2)\}$$

この式は次のように導かれる。ある感覚の信頼度が取得処理1, 2から と導かれているが、これは処理1では確率、処理2では確率 で適切であると判断されたことを意味する。この事象の空間は図のように表現できる。図3には4つの領域があるが、起こり得るのは両方が適切または不適切となる領域のみである。なぜなら、属性が適切であれば処理1, 2ともに知覚が適切である事象が発生し、属性が不適切であればともに不適切である事象が発生するからである。

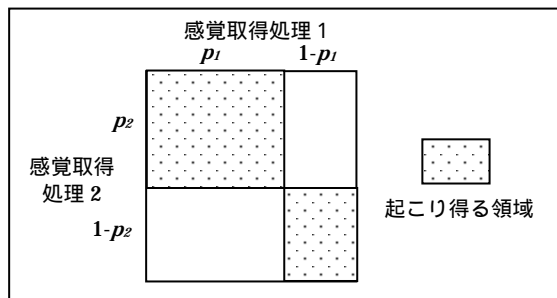


図3 2つの取得源の事象空間

信頼度合成方式によって、信頼性の高い複数の処理から同じ感覚情報が得られた場合にはその信頼度が高くなり、逆に、信頼性の低い処理からのみ得られた情報の信頼度は低くなった。したがって、信頼度に閾値を設けることにより、信頼性の高い情報と低い情報を切り分けることが出来ると考え、実験によって適切な閾値を求めた。

#### 4.3.2 正負情報

知覚判断においては、雑音を排除するために知覚知識ベースに登録した知覚語の「正負情報」を利用した。

よい(+)イメージの知覚の出力が期待されるとき、悪い(-)イメージの知覚が含まれるなら、それらは全て雑音になり、消去の対象である。同様に、悪いイメージの出力が期待された場合にはよいイメージの知覚は雑音になる。そこで、代表語が入力された場合にはその知覚の正負情報を調べ、出力の中に存在する矛盾する知覚を排除する処理を施す。但し代表語が+、-の両方のイメージを持つ場合には処理を行わない。正負情報の利用によって常識的、非常識でない出力も一部削られてしまったが、雑音を半減させた。

#### 4.4 感覚想起の評価

五感判断と知覚判断の感覚想起処理それぞれを、精度と再現率について評価した。

精度は出力の信頼性を表す値である。感覚想起によって出力された感覚に対して、正解なら「常識的(○)」、不正解なら「非常識(x)」、正解とも不正解とも言えない場合は「非常識でない(△)」の3段階評価を与えた。五感判断では全出力数に対する常識的又は非常識でない出力の割合を、知覚判断では全出力数に対する常識的な出力の割合を精度とした。

再現率は、人間が期待する出力の内どれだけをシステムが出力できるかを表す値である。五感判断、知覚判断のいずれも、人間が想起した感覚の個数に対するシステ

ムが出した常識的又は非常識でない感覚の個数の割合を再現率とした。

五感判断の評価データは日常よく使われる名詞をランダムに取り出した300語(代表語100語含む)、知覚判断の評価データも同様に取り出した200語(代表語40語含む)である。それぞれの評価結果を表8と表9に示す。

表8 感覚想起の評価(五感)

	精度(%)	再現率(%)
全体	96.3	64.7
既知語	98.2	84.4
未知語	91.7	37.4

表9 感覚想起の評価(知覚)

	精度(%)	再現率(%)
全体	82.0	47.2
既知語	91.4	70.8
未知語	78.6	41.3

精度から、五感判断・知覚判断の感覚想起処理は、既知語・未知語に関わらずいずれも入力された語に対してほぼ間違いのない感覚を出力することができると言える。また再現率から、人が期待する出力の約半数以上を返答できることがわかる。以上より、「感覚想起」については、名詞から適切な感覚を想起するという目的において、十分な成果が得られていると言える。

#### 5. 感覚的連想

感覚判断システムの重要な機能の一つに「感覚的連想」がある。感覚想起が名詞から形容詞類を連想するのに対し、感覚的連想では逆に形容詞類を含む情報から具体的な名詞の連想を可能にする。

形容詞類と総称的な名詞の組み合わせから具体的な名詞を連想することを「感覚的連想A」、または複数の形容詞類の組み合わせから具体的な名詞を連想することを「感覚的連想B」とする。入出力の例を表10に示す。

表10 感覚判断の入出力の例

	入力	出力
感覚的連想A	危ない, 動物	毒蛇, 豹, 猛獣
	美しい, 人	美女, 佳人, 天女
感覚的連想B	可愛い, 小さい	赤ちゃん, 兎, 金魚
	丸い, 赤い	西瓜, 林檎, トマト

##### 5.1 感覚的連想メカニズム

感覚的連想を実現する仕組みについて説明する。

###### 5.1.1 感覚的連想A

感覚的連想Aでは、まず入力された形容詞を感覚語(知覚語)に置換した後、以下の4つの手法によって具体的な出力を得る。

- 感覚知識ベースの分類語に属する語のうち、入力形容詞を感覚として持つものを取得
- 入力形容詞の属性のうち、入力名詞のシソーラスでのノードに属するもの取得
- 入力名詞の属性のうち、感覚想起の結果が入力形容詞になるもの取得
- 入力名詞のシソーラスのノードに属する名詞のうち、感覚想起の結果が入力形容詞になるもの取得

例えば入力に「冷たい」と「飲み物」が与えられた場合、

分類語「飲物」に属する語のうち「冷たい」を感覚にもつもの取得

「冷たい」の属性で、「飲み物」のシソーラスでのノードに属するもの取得

「飲み物」の属性で、感覚想起の結果が「冷たい」になるもの取得

「飲み物」のシソーラスのノードに属する名詞で、感覚想起の結果が「冷たい」になるもの取得

### 5.1.2 感覚的連想 B

感覚的連想 B では、まず入力された形容詞を感覚語(知覚語)に置換した後、以下の 2 つの手法によって具体的な出力を得る。

代表語と分類語に属する語のうち、入力形容詞全てを感覚としてもつもの取得

入力形容詞の属性に含まれる名詞の内、重複するもの取得

例えば、入力に「赤い」と「甘い」が与えられた場合は

代表語と分類語に属する語のうち「赤い」と「甘い」を感覚としてもつもの取得

「赤い」と「甘い」の属性に含まれる名詞の内、重複するもの取得

### 5.2 単純感覚的連想

5.1 節で説明した手法を「単純感覚的連想」と呼び、出力の精度を調べた。精度は、出力された名詞が入力セットに対して妥当であるか否かを判断し、全出力に対する妥当な出力の割合で求めた。データは、人間が具体的な名詞を想起しうる形容詞類と名詞の組み合わせ 90 セット(感覚的連想 A)、形容詞類 2 語の組み合わせ 30 セット(感覚的連想 B)を用意した。評価は五感に関する形容詞類についての連想と、知覚に関する形容詞類についての連想に分けて行った(表 11, 表 12)。

表 11 感覚的連想 A の評価

	処理	処理	処理	処理
五感	99.6%	78.9%	79.2%	85.7%
知覚	98.9%	90.1%	63.4%	78.8%

表 12 感覚的連想 B の評価

	処理	処理
五感	100%	59.2%
知覚	100%	44.9%

感覚的連想 A・B ともに、知識ベースだけを用いる処理の出力はほぼ正確であり、概念ベースを参照するその他の処理は、出力に雑音が含まれていた。

感覚的連想 A の間違った出力を調べたところ、入力された形容詞類と対応していないもの(例:美しい+人 元締め, 管理人)、入力された名詞と対応していないもの(例:高い+建物 木)に分けられた。このうち形容詞と対応していない間違いが全体の 78%であった。これは、概念ベースの属性に不適切な形容詞類や、そのような形容詞類を想起させる名詞が含まれているためだと思われる。対処法として、概念ベースの属性を感覚の観点から精練すること、または入力語と出力語の関連の深さを判断し、関連が低い出力を雑音として排除することが考えられる。

4章で記した感覚想起においては、概念連想メカニズムを用いることにより信頼性の高い連想を実現している。そこで、感覚的連想においても関連度計算を用いることにより出力の信頼性向上を目指すことにした。

## 6 4 つの関連度計算方式

感覚想起で用いた概念連想メカニズムは、関連度計算に「意味関連度計算方式」<sup>[6]</sup>を採用したものである。意味関連度計算方式は概念の属性の一致度合いから関連性を判断しているため、連想により導き出せるような語の間の関連性判断は上手くできないと考えられる。例えば、人間が「花束 - 美しい」の方が「狸 - 美しい」よりも関連が強いと思うのは、「花束」と「美しい」が共に使われる頻度が「狸」と「美しい」が共に使われる頻度よりも高いためであると考えられる。感覚的連想の雑音排除に用いるためには、このような対になる語の関連性を考慮に入れた関連度計算方式が望ましいと考えられる。

我々の研究室では、意味関連度計算方式の他にも幾つかの関連度計算手法が考案されている。それぞれの関連度計算方式についてその特徴を簡単に記す。

### 6.1 意味関連度計算方式

人が概念間の関連性を判断する場合、「自動車 - 車」「電車 - 汽車」といったように意味的に近いかが重要な判断基準となると考えられる。意味関連度計算方式は意味的にどれだけ近いかを判断することで概念間の関連性を評価する計算方式である。

2つの概念間の意味関連度は、各概念を二次属性まで展開し、一致する属性と重みによって求める一致度を使い計算する。一致度は 0~1 の実数値をとる。具体的には、一致する二次属性を調べ、その重みを使って計算する一致度の和が最大になる一次属性の組み合わせを作る。

### 6.2 共起関連度計算方式

共起関連度方式は概念の共起情報から関連性を判断する方式である。意味的な近さを判断するだけの意味関連度計算方式では、「花束 - 美しい」といった連想により導き出せるような関連性の判断は難しい。そこで対になって出現する頻度の高い概念間には関連があることを考慮に入れた関連性評価法について示す。

#### 6.2.1 表記的共起関連度計算方式

ある語の対が複数の文書で共出現する頻度が高ければ高いほどその語の間には関連があると考えられる。語の対が共出現する頻度から関連性を判断する表記的共起関連度の考え方を以下に示す。

$$Co(A, B) = \left( \frac{df(A \& B)}{df(A)} + \frac{df(A \& B)}{df(B)} \right) / 2$$

$df(A \& B)$ : 概念 A, B が共に一次属性に出現する概念数

$df(A)$ : 概念 A が一次属性に出現する概念数

$df(B)$ : 概念 B が一次属性に出現する概念数

例えば、 $df(\text{自動車} \& \text{運転})$  は「自動車」と「運転」を共に属性に持つ概念数を数えるが、「車」と「運転」を共に持つ概念は対象にならない。このように、意味的に近い語との共起は考慮せず、「自動車 - 運転」という「表記」の対がどれだけ出現するかを評価する。

### 6.2.2 意味的共起関連度計算方式

概念Aと概念Bの関連が強いほどそれぞれの概念の意味特徴を表す属性集合内に対象とする語が数多く出現すると考えられる。そこで、属性集合内に対象とする語がどれだけ出現するかで関連性を評価する。但し、概念ごとに属性に出現する回数に幅があるため、出現回数だけの評価では不十分と判断し、稀に出現する語を重要とする *idf* 値を利用する。

対象とする属性集合は三次属性まで展開する。また、情報検索の分野でよく用いられる稀に出現する語を重要とする *idf* 値を利用して評価する方法を以下の式で定義する。

$$CoCalc(A, B) = \left( \frac{b_1}{Aznum1} + \frac{a_1}{Bznum1} \right) / 2$$

$$CoCalc2(A, B) = \left( \frac{b_2}{Aznum2} + \frac{a_2}{Bznum2} \right) / 2$$

$$CoCalc3(A, B) = \left( \frac{b_3}{Aznum3} + \frac{a_3}{Bznum3} \right) / 2$$

*AznumN*: 概念AのN次属性数

*BznumN*: 概念BのN次属性数

*a<sub>N</sub>*: 概念Aが概念BのN次属性内に出現する回数

*b<sub>N</sub>*: 概念Bが概念AのN次属性内に出現する回数

### 6.2 意味共起関連度計算方式

意味共起関連度計算方式は、意味関連度計算方式と意味的共起関連度計算方式を複合利用することによって、よりよい関連性の実現を目指したものである。評価式は、以下のように定義されている。

$$MCR(A, B) = \frac{MR(A, B) + Cw \times CR(A, B)}{1 + Cw}$$

*MR(A, B)*: 意味関連度

*CR(A, B)*: 共起関連度

*Cw*: 共起関連度重み

共起関連度に重み *Cw* を付与し、意味関連度に足し合わせ意味共起関連度を求める。

### 6.3 各関連度計算方式の比較

感覚的連想の雑音排除に用いる関連度計算方式を決定するため、感覚に関するデータを用いて4つの手法の評価を行った。

#### 6.3.1 評価用データ

感覚知識ベースの代表語から、任意の概念X(基準概念)と、Xに対応する感覚を概念Aとして取り出した。さらに、Aの反対感覚としての概念B、Xと関連の低い感覚を無作為に選んだ概念Cとを取り出し、4つの概念(X-A,B,C)を1セットとして用意した。五感に関するデータが573組、知覚に関するデータが729組の合計1302組の評価用データを作成した(表13)。評価用データはすべて概念ベースで定義されている概念のみで構成されている。これを用いて各手法の精度を調査した。

表13 評価データ(五感判断)

X	A	B	C
羽毛	軽い	重い	暗い
...	...	...	...

### 6.3.2 評価方法

概念Xと概念Aとの関連度 *Rel(X,A)*, 概念Xと概念Bとの関連度 *Rel(X,B)*, 概念Xと概念Cとの関連度 *Rel(X,C)* を求める。そこから  $Rel(X,A) > Rel(X,B) > Rel(X,C)$  を満たすときを正解として精度(順序正解率)を求める。評価用データ全体の順序正解率が高ければ人間の感覚に近い判断ができていくことになる。

全データの *Rel(X,A)*, *Rel(X,B)*, *Rel(X,C)* の平均値とX-ABC順序正解率を以下に示す。

表14 X-A,B,C順序正解率(五感)

	(X,A)	(X,B)	(X,C)	正解率(%)
意味	0.127	0.036	0.003	76.4
表記的	0.200	0.043	0.009	<b>83.9</b>
意味的	0.042	0.009	0.002	<b>83.9</b>
意味共起	0.087	0.023	0.002	82.7

表15 X-A,B,C順序正解率(知覚)

	(X,A)	(X,B)	(X,C)	正解率(%)
意味	0.056	0.009	0.003	42.0
表記的	0.058	0.009	0.003	<b>89.8</b>
意味的	0.010	0.002	0.001	<b>82.6</b>
意味共起	0.034	0.005	0.002	56.9

表14と表15から、どの方式も  $(X,A) > (X,B) > (X,C)$  であることがわかる。また、順序正解率では表記的共起関連度と意味的共起関連度が高い値を示している。

順序正解率の他に、全データの *Rel(X,A)*, *Rel(X,B)* それぞれにおける変動係数を求めた。変動係数は標準偏差を平均値で割ることによって求められる、データのばらつきを評価する尺度であり、雑音とそうでない情報を切り分けるのに適した方式を選ぶ基準とする。また、正解と間違いの関連度の差が大きく出る方式ほど望ましいため、各方式の  $(X,A)/(X,B)$  の値も求めた。

表16 変動係数と  $(X,A)/(X,B)$  の値(五感)

	変動係数 (X,A)	変動係数 (X,B)	$(X,A)/(X,B)$
意味	1.087	1.353	3.493
表記的	<b>0.784</b>	<b>1.167</b>	<b>4.604</b>
意味的	1.566	1.371	<b>4.779</b>
意味共起	1.090	1.231	3.725

表17 変動係数と  $(X,A)/(X,B)$  の値(知覚)

	変動係数 (X,A)	変動係数 (X,B)	$(X,A)/(X,B)$
意味	2.223	3.497	6.321
表記的	<b>1.780</b>	2.988	6.104
意味的	1.977	2.253	6.255
意味共起	2.108	3.186	6.255

表16から、 $(X,A)/(X,B)$  の値からは表記的共起関連度と意味的共起関連度が比較的良好な値を示している。一方、変動係数に関しては表記的共起の値が小さく、つまり関連度のばらつき具合が少ないことがわかる。

以上より、感覚的連想の閾値に用いるには表記的共起関連度方式が最も適していると判断し、これを採用することにした。



## 7 共起情報利用による精度向上

表記的共起関連度方式を用いた概念連想メカニズムによって、感覚的連想の出力の信頼性を上げる試みを行った。

### 7.1 閾値調査

入力された形容詞類と出力された名詞との関連度を求め、ある関連度以下の情報を不適切な出力として排除する。そのためには、出力を信頼性の高い情報と低い情報(雑音)とに切り分ける閾値を設ける必要がある。

閾値の調査には、5.2 節の感覚的連想Aの評価セットに含まれる形容詞類と、セットから出力された名詞を用いた。出力を正解と不正解に分け、それぞれについて入力された形容詞類との関連度を求めた。図4に入力が「白い」と「服」だった場合の例を示す。

白い + 服		ウエディングドレス, 白無垢, 白衣, 上着	
	出力	「白い」との表記的共起関連度	
正解	ウエディングドレス	0.128	
	白無垢	0.216	
	白衣	0.120	
不正解	上着	0.024	

図4 入力形容詞と出力名詞の表記的共起関連度

閾値調査用のデータは、正解群で652セット(五感:378, 知覚:274), 不正解群で132セット(五感:71, 知覚:61)用意した。

全てのデータを評価したときの正解群, 不正解群それぞれの平均関連度を表17に示す。

表17 全データの正解群と不正解群の平均関連度

	正解群	不正解群
五感	0.183	0.071
知覚	0.086	0.085

表17より、五感については正解群と不正解群の平均関連度に有意な差が見られるため、適切な閾値を設けることが出来ると考えられる。一方、知覚は正解群と不正解群それぞれの平均に殆ど差が見られず、適当な閾値を設けるのは困難と思われる。これは、正解群では表記的共起関連度が0になったものが多かったためである。

閾値の設定には、閾値と出力の関係を表すデータが必要である。そこで、閾値を0から2.0まで0.01刻みに変化させたときの出力の変化をグラフにした(図5, 図6)。

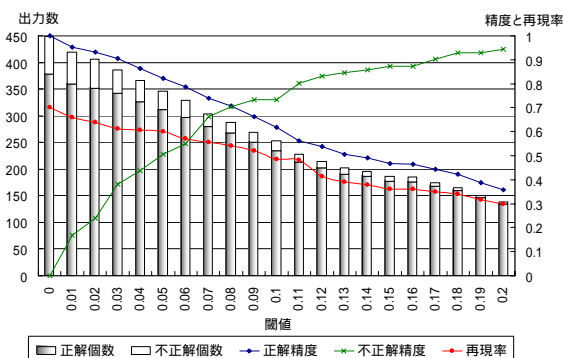


図5 閾値と出力の関係(五感)

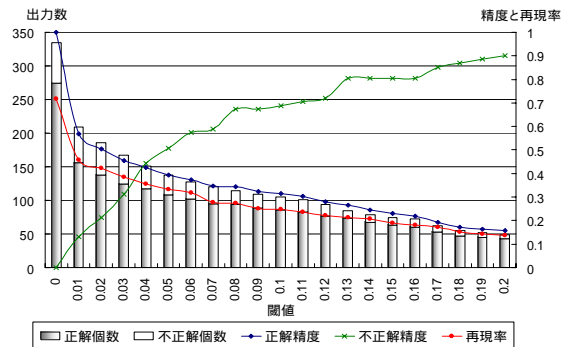


図6 閾値と出力の関係(知覚)

グラフでは、各閾値における正解と不正解の個数, 正解精度(その閾値で残った正解の割合)と不正解精度(その閾値で排除された不正解の割合), 再現率を表している。再現率は、人間が期待する出力のうちシステムがどの程度を出力できたかを表す値である。形容詞類と名詞のセットから連想できる具体名詞をアンケートによって集め、2人以上が解答した名詞を期待される出力として評価した。

図5のグラフから、五感に関する感覚的連想では閾値を上げるほど不正解精度が上がり、出力に占める不正解の数が減少することがわかる。しかし、閾値を上げるほど出力中の正解もある程度減少し、正解精度は低下する。このような状況では、正解個数が正解精度, 不正解精度が共に最も高くなる閾値が望ましいと考えた。そこで、グラフにおいて両データの交点付近である閾値0.08を採用した。この閾値での再現率は0.54であり、期待される出力の約半分を満たすことができる。

一方、図6のグラフからは、知覚に関する感覚的連想では閾値を上げるほど不正解は減って精度は上がるものの、正解も著しく減少することがわかる。つまり、閾値によって出力を絞り込んでも信頼性を上げる効果は低いとわかった。この理由と対処法については7.3節で考察する。

### 7.2 共起情報利用による精度向上

7.1節で設けた閾値によって、出力の精度がどう変化かを検証した。用いたデータは4.2節で用いた、形容詞類と名詞の組み合わせ90セットと形容詞類2語の組み合わせを20セットである。閾値による絞込みは感覚的連想Aの処理～、感覚的連想Bの処理に対して行った。出力を閾値0.08で絞ったときの各処理の精度を表18, 表19に示す。

表18 感覚的連想Aの評価(閾値0.08)

	処理	処理	処理	処理
五感	99.6%	97.8%	92.0%	96.0%

表19 五感形容詞類に関する感覚的連想Bの評価(閾値0.08)

	処理	処理
五感	100%	73.9%

閾値を設けない場合(単純感覚的連想)と比較して、どの程度の精度が向上したかをグラフに示す(図7)。

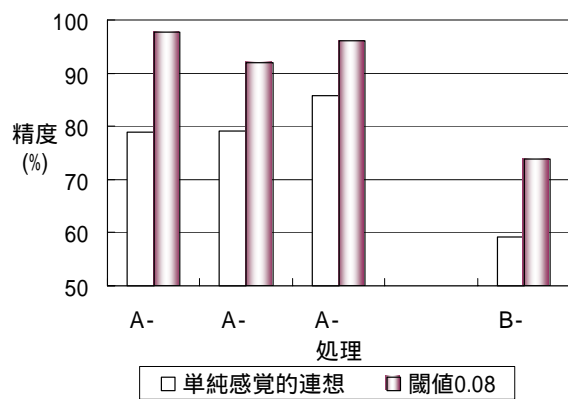


図7 閾値による雑音排除の効果

図7より、いずれの処理も10%以上の精度向上が確認できる。また、閾値を設けた場合の再現率の変化をグラフに示す。(図8)

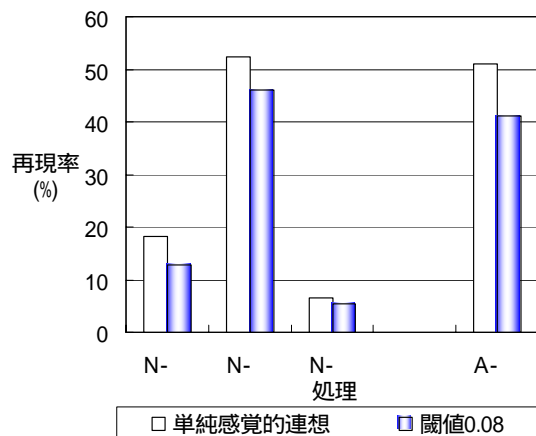


図8 閾値による再現率の変化

閾値設定後の再現率は閾値を設けない場合の70%程度であり、出力を極端に減らすことなく精度を上げることが可能であるとわかった。

以上より、五感に関する感覚的連想においては、表記的共起関連度を用いた閾値を設けることで出力の信頼性を上げることが成功したと言える。この、共起情報を利用した感覚的連想を感覚判断システムの標準機能として組み込んだ。

## 7.2 共起情報利用による精度向上

知覚に関する感覚的連想においては、五感のように表記的共起関連度による適当な閾値を設けることができなかった。これは、評価セットの形容詞類と名詞の表記的共起関連度が0になるものが多く、全体の精度に影響したためである(表17参照)。関連度が0になるセットを省いた評価データでの結果を参考までに示す(図9)。

図9のグラフから、知覚に関する形容詞類でも0以外の値で関連度が計算できた場合には適当な閾値の設定が可能だと思われる。

表記的関連度がうまくとれないのは、形容詞類を概念とする語の属性の不足や、概念の属性に形容詞類が含まれている数が少ないことが理由として考えられる。今後、感覚判断システムの信頼性を上げていくためには、感覚に

関する属性の質をよくすることが求められる。具体的には、属性の追加や削除といった作業が必要である。

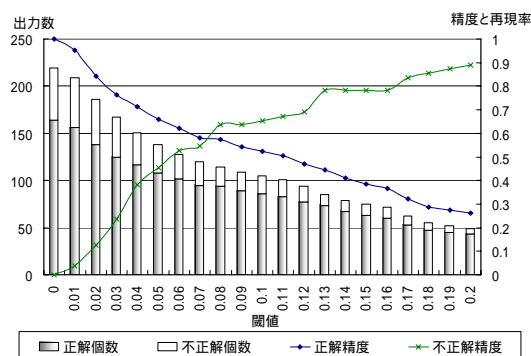


図9 関連度が0以外のデータを用いた閾値と出力の関係

## 7 共起情報利用による精度向上

本研究の成果は、感覚判断システムの「感覚的連想」機能について、そのメカニズムの評価・検討を行い、語間の共起情報を用いることによって出力の信頼性を向上させたことである。またその過程において、4つの関連度計算方式の調査を行った結果、名詞と形容詞類の関連の深さを計るには表記的共起関連度が適しているとわかった。

本稿で提案した共起情報を用いた感覚的連想によって、感覚判断システムの常識判断の幅が広がり、ひいてはコンピュータによる自然な会話の実現に貢献したと思われる。

## 参考文献

- [1]笠原 要, 松澤 和光, 石川 勉, “国語辞書を利用した日常語の類似性判別”, 情報処理学会論文誌, Vol.38, No.7, pp.1272-1283, 1997
- [2]渡部 広一, 河岡 司 “常識的判断のための概念間の関連度評価モデル”, 自然言語処理, Vol.8, No.2, pp.39-54, 2001
- [3]広瀬幹規, 渡部広一, 河岡司: “概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精練手法”, 信学技報, NLC2001-93, pp.109-116, 2002
- [4]土屋 誠司, 小島 一秀, 渡部 広一, 河岡 司: 常識的判断システムにおける未知語処理方式, 人工知能学会論文誌, 17巻, 6号, B, pp.667-675, 2002
- [5] NTTコミュニケーション科学研究所監修, 「日本語語彙体系」, 岩波書店, 東京, 1997
- [6]井筒 大志, 渡部 広一, 河岡 司: 概念ベースを用いた連想機能実現のための関連度計算方式, 情報科学技術フォーラム, FIT2002, E-39, pp.159-160, 2002
- [7]堀口 敦史, 渡部 広一, 河岡 司: 常識的感覚判断システムの構築, 情報処理学会知能と複雑系研究会資料, 2002-ICS-130, pp.31-36, 2002
- [8]渡部 広一, 堀口 敦史, 河岡 司 “常識的感覚判断システムにおける名詞からの感覚想起手法”, 人工知能学会誌, Vol.19, No.2, pp.73-82, 2004
- [9]青田 正宏, 渡部 広一, 河岡 司: 概念の意味・表記と共起情報を用いた関連度計算方式, 同志社大学理工学研究報告, Vol 45, No.1, pp.23-34, 2004
- [10]米谷 彩, 渡部 広一, 河岡 司: 常識的知覚判断システムの構築, 第17回人工知能学会全国大会論文集, 3C1-07, 2003