同志社大学大学院　博士論文

語概念連想システムの構築とその応用

芋　野　美　紗　子

同志社大学大学院　工学研究科
情報工学専攻

2014 年 3 月

目 次

図 目 次

表 目 次

第1章　序論

効率的なWeb検索のための単語群に対する処理から，人と対するような会話を行うシステムまで，情報処理技術において自然言語を扱うテーマは数多く存在する．扱う自然言語の情報はただ1つの語の場合もあれば，膨大な文書集合の場合もあり様々であるが，それらを扱う多くの自然言語処理において処理対象の意味を理解するためのアプローチは必要不可欠であると考える．

語や文などが存在するとき，その意味とは「語や文を見たときに人間が理解する内容」の事だと考える．人間はある1つの語に対して，その語が指し示す事物・事象の内容を意味として覚え，理解することが出来る．また，ある文書について，記述されている内容を文書の意味として読み解くことが出来る．語や文の意味を理解できるからこそ，人間は「意味が同じ語」や「意味が類似している文書」といった判断も容易に行うことが出来る．

この意味の理解を情報処理技術においても行うためには，自然言語で表現される処理対象の意味を何らかの形で機械上に定義しなければならない．例えば語を意味によって分類することで上位・下位関係や同義・類義関係を記述するシソーラス [1]，概念（ノード）を関係（リンク）で結ぶ意味ネットワーク [2]，文書中の語群を出現頻度からベクトル化するベクトル空間モデル [3]，if-then形式の記述で条件文による推論の知識を記述するプロダクション規則 [4] など，既存の自然言語処理において語や文，概念の意味を理解するために様々な手法での「意味定義」が行われている．

自然言語を扱う以上，人間が行う意味の理解と機械が行う意味の理解に乖離があることは望ましくない．しかし前述したような各種の意味定義によって，人間のような自然言語に対する意味の理解が得られるかと考えたとき，そこには曖昧さや柔軟さが欠けているように感じられる．

提案されてきた多くの手法で定義される意味とは，例えば単語の語彙的な意味，つまり国語辞書の語義文により示される意味であったり，文書中に出現する語群の集合体で定義された処理対象に依存する意味が大半である．しかし人間がある一つの語に対して理解する意味は，単語の語彙的な意味だけではない．人間はある文書同士が類似しているかを，出現している語の集合が一致しているかどうかのみでは判断しない．例えば「雨」の語彙的な意味は「大気で水蒸気が冷えてできた雲の中で水滴ができ，それが地上に落ちる現象」だが，人間は「傘」「長靴」「カビ」「紫陽花」といった語彙的な意味とは関係のない語との間でも，意味の近さを理解することが出来る．また，文書中に出現する語それぞれに対しても，表記や語彙的な意味以外の観点で類似性を判断できる．また今まで提案されてきた意味定義においては，語や事物，概念間の関係性が明確に決められているものが多く，例えばシソーラスでは語と語の間に上位・下位や同義・類義といった関係性を定義する必要がある．しかし人間がある語と語に関連があると感じるとき，そこに必ず明確な関係性の定義があるとは言い切れない.「蝸牛」と「貝」は

シソーラスにおいて上位・下位関係に分類されるが,「蝸牛」と「紫陽花」をこのような体系で定義することは至極困難である．しかし人間にとっては「蝸牛」と「紫陽花」という語の間に何かしらの関連を見出すことは容易い．

語彙的な意味のみに依存するわけではなく，また分類・体系付けられた構造でも表現しきれない，ある種の「飛躍的な発想」の元で理解する自然言語の意味や関連を私達人間は無意識に扱っている．これは人間が持つ連想能力によって支えられている．連想とは，自身の知っている事柄，情報，概念といった多種多様の「知識」から，他の知識を思い浮かべることと定義する．人間が「紫陽花と蝸牛」の間に関連を見いだせるのは，これらの語が互いを連想させるからである．強いて，シソーラスや意味ネットワークのような「関係性の定義」に照らし合わせるならば,「連想できる」という関係性がこの語の間には存在している．人間が行う意味の理解を機械においても行うためには，この連想を表現する機構が不可欠であると考える．そこで本稿では,「連想できる」という曖昧な関係性を定義した知識ベースを構築することにより，人間らしい連想を機械上に実現するための語概念連想システムの提案および，それを用いた応用技術について述べる．

以下，第 2 章では連想を表現する機構である，語概念連想システムについて述べる．語概念連想システムは自然言語で表現されたある語に対して，そこから連想できる他の語の集合を属性として持たせることで概念化した「概念ベース」を核としたシステムである．ある概念から連想できる他の概念を想起する処理および，概念と概念の間の関連を定量的に表現することで意味の近さを測る関連度計算方式によって構成される．

第 3 章では概念ベースの構築手法について述べる．基本となる概念ベースを国語辞書の見出し語と語義文を用いて作成し，概念の意味を他の概念の集合，つまり属性で表す連鎖構造の知識ベースを構築する．さらに属性に対して個々の重要度を示す重みを付与するため，人手によるサンプル概念の属性の評価結果から属性の信頼度を算出する．この基本概念ベースに対して属性の選別をするためのルールを与え，これに従い概念にとって不適切な属性を削除することで概念ベースの精度向上を行う．しかし，概念ベースの情報源として国語辞書だけを用いる場合には，概念の不足や属性によって定義される意味の狭さが問題となる．そこで新聞記事を用いた概念および属性の拡張を行うことにより，概念ベースの知識をより充実したものとする．さらにサンプル概念の属性の評価結果を用いない新たな重み付け手法により，人手による作業を必要としない概念ベースの拡張を行う．また，シソーラスを新たな情報源とした属性の追加手法についても述べる．

第 4 章では前章までで述べた概念ベースに対して更に属性を追加し，概念の意味定義を拡張する手法を提案する．新たな属性の追加手法として，概念ベースの連鎖構造を利用した二次属性からの属性追加および Web 上の情報から属性を取得する手法を述べる．二次属性および Web 上から属性の候補となる語を取得した上で，各種の閾値設定により適切な属性を選択・追加することで概念ベースの精度向上を行う．

第 5 章では語概念連想システムの拡張として，複数語概念連想システムについて述べる．これは 2 章で述べる語概念連想システムにおける，ある語から連想できる他の語を想起する処理を複数語に対応させたシステムである．複数の概念を並列に見たとき，それらから想起されるべき概念を取得するシステムについて述べる．

第 6 章から第 8 章では，前章までで述べた語概念連想システムを自然言語を対象とした情報処理技術に応用する事例について述べる．

まず第 6 章ではロボットとの知的会話を視野に入れた新聞記事見出し文の意味具体化手法について述べる．人間らしい会話をロボットが行うためには，あいさつのような慣用的な表現だけでなく，質問や返答，提案，更には何かしらの話題についての雑談など様々な種類の発話が必要であると考え，そのうち時事情報を話題とした会話について着目をした．時事情報を容易に取得できる新聞の見出し文を知的会話のリソースとして扱うために，見出し文を特有の書式から会話に適した表現へと変換した上で，記事本文による意味の具体化を行う手法について述べる．

第 7 章では自然言語処理の分野で活発に議論される「言い換え」や「変換」処理に語概念連想システムを応用する手法として，新聞記事をロボットとの知的会話のリソースとして用いることを視野に入れた，記事中の難解な語を人間の会話に出現するレベルの平易な表現へ変換する手法を提案する．難解語を変換する際には，他の簡単な一語に変換する一語変換と，一語では説明できない難解語を文によって変換する N 語変換を併用することで人間にとって自然に感じる語句変換を実現している．一語変換の際には，難解語と変換を行う一語との意味的な近さを関連度により判断し，候補から適切な語を選択している．また，N 語変換においては難解語のもつ意味文と変換に用いる文との関連を，概念ベースに定義される語の意味知識を利用して算出することで適切な変換を行うことを目指した．

第 8 章では従来のタスク型の会話システムや膨大な対話例のコーパスに依存した応答生成とは違い，語概念連想システムによる語の連想機能およびそれを基盤として構築される常識判断システムを活用することで，人間の発話からそれに適した自律的な応答を生成する手法について述べる．提案手法では人間の発話中の場所に関する情報を起点として，場所での行動を連想して応答する「場所連想」，場所に存在する人や物への一般的な感覚による共感を応答する「形容詞連想」，人間の発話内容から連想できる他の話題を応答する「話題転換連想」の三つの処理を行い，人間が常識的と感じる応答を生成した．

第2章　語概念連想システム

2.1　はじめに

情報処理技術における自然言語の意味理解は，きちんと整理されたデータ群による意味定義を必要とする手法が一般的である．シソーラスは語の語彙的な意味における互いの関係を分類配列することで木構造による意味定義を提供している．意味ネットワークでは対象世界に存在する概念と概念を，その間の関係性を示すリンクで繋いだものの集合体で意味を定義する．これらの定義では互いに繋がる語間に対して明確な関係性がすでに付与されており，つまり逆に言えば明確に記述できない曖昧な関係性を表現することが難しい．しかし人間は語彙的な意味のみに依存するわけではなく，また分類・体系付けられた構造でも表現しきれない自然言語の意味や関連を理解することが出来る．例えば「雨」という語は「大気で水蒸気が冷えてできた雲の中で水滴ができ，それが地上に落ちる現象」を意味するが，我々は雨が降れば「傘」や「長靴」を使い,「災害」により「山」が「崩れ」て「川」が増水し,「湿気」で「カビ」が生え,「梅雨」が訪れると「紫陽花」が咲き「蝸牛」が這う…のように様々な語の間に関連を見出し，それらの意味の近さを理解することができる．

このような曖昧さ，柔軟さを持った意味の理解を行えるのは，人間が連想という能力を有しているためだと考える．連想とは，自身の知っている事柄，情報，概念といった多種多様の「知識」から他の知識を思い浮かべることと定義する．人間が「紫陽花と蝸牛」の間に関連を見いだせるのは，これらの語が互いを連想させるからである．

人間が「紫陽花」から「蝸牛」，もしくはその逆を連想できるのは，総括的な視点でそれぞれの語が表す内容を知識として保持しているためだと考える．人間が「紫陽花」という語の指し示す内容，つまり「紫陽花の意味」として持っている知識は語彙的なものだけではなく，自身が存在する現実世界において「紫陽花」とはどのような事物・事象であるか，という概念的なものだと考える．そこにはもちろん語彙的な意味も含まれるが，更に抽象的，例えば「紫陽花に蝸牛という組み合わせが一般的に共感される」という事実関係も「紫陽花」の意味として存在している．様々な事物・事象を概念的な意味で捉えることにより，単なる語彙的な意味の近さだけではない関連の有無を見出すことができる．

情報処理技術において事物・事象の概念化を行うための思想として注目されるのがオントロジー [5] である．オントロジーとは何かしらのタスクやドメイン領域における事物・事象を概念化するための構造を定義する事を指し，これにより知識の体系化，明示化，共有化が期待できるとされている．オントロジーにおいて定義される構造とは，対象とする世界において事物・事象がどのように存在しているかを表現するための枠組みであり，事物・事象そのものの分類や意味を定義するものではない．対象となる事物・事象の存在を説明するために必要十分なカ

テゴリーは何かを考察し，それを書き記したものがオントロジーである．

前述した，人間が「紫陽花」という語に対して持つ概念的な意味をオントロジーにより表現できるかを考えると，それは困難であると思われる．オントロジーは概念化する事物・事象の存在を説明するために必要十分なカテゴリーを如何に選び，それらをどのような構造に組み合わせるかが鍵となるが，そもそも人間自身が「現実世界において紫陽花とはどのような事物・事象であるか」を完全に理解しているとは思えない．語彙的な意味にも分類・体系付けられた構造にも依存しない,「連想できる」という曖昧な関係性を記述する最適なカテゴリーを創造することは非常に難しい．しかし人間のような連想を機械上で表現するためには，この曖昧な関係性を何らかの構造で定義する必要がある．

そこで本章で述べる語概念連想システムでは，現実世界に存在する事物・事象を表す自然言語の「語」を，曖昧な関係性を持っているという事実のみに着目して概念化し，その概念の知識集合である「概念ベース」を用いて人間のような連想と意味の理解を機械で表現することを目指している．

概念ベースでは，自然言語で表現される様々な「語」に対して人間が持つ意味，つまりその「語」が示すものが現実世界においてどのような事物・事象であるかを,「語」から連想できる他の「語」の集合を属性として付与することで概念化し，定義する．概念と属性の間にある関係性自体を表すラベルやカテゴリーは存在しない．人間が「紫陽花」と「蝸牛」に何かしらの関係性を見出すならば，その関係性が分からなくとも概念「紫陽花」の属性に「蝸牛」を含んでしまえばよい．これはオントロジーにおける概念化のための構造定義とはまったく違うアプローチである．

この概念ベースを用いることで，ある語から連想される他の語を想起する処理および，語と語の間の関連を定量的に表現することで意味の近さを測る関連度計算方式が構築され，これらの処理により語概念連想システムは構成される．語概念連想システムにある語を入力すると，その語から人間が連想できるであろう他の語を出力する．また，２つの語を入力することで，語間の関連を定量的に表現した関連度という値を出力する．

2.2　概念ベース

概念ベース [6–8] とは，様々な語の意味を他の語の集合（属性）により概念化し，その概念の知識を集めたものである．概念および属性は主に国語辞書，新聞記事，シソーラスから抽出される．属性はそれぞれ重みを持っており，それが概念にとっての属性の重要度を示す．任意の概念 A はその属性 a_i と重み w_j によって以下のように定義される．なお，概念 A の属性数は N 個とする．

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_N, w_N)\}$$

表 2.1 に概念定義の一例を示す.

概念ベースにおいて，属性となる語は全て概念としても定義される．つまり，それぞれの属

表 2.1: 概念と属性の例

概念	(属性，重み)
蝸牛	(腹足類，1.62)，(螺旋，0.715)，(梅雨，0.3594)，(食料，0.002)，…
紫陽花	(紫，0.603)，(雨，0.396)，(咲く，0.165)，(美しい，0.046)…
梅雨	(梅雨前線，1.69)，(紫陽花，1.14)，(季語，0.485)，(蝸牛，0.481)…

性からも，その属性の意味を定義する新たな属性を導くことが出来る．この構造を一例で表すと図 2.1 のようになる．

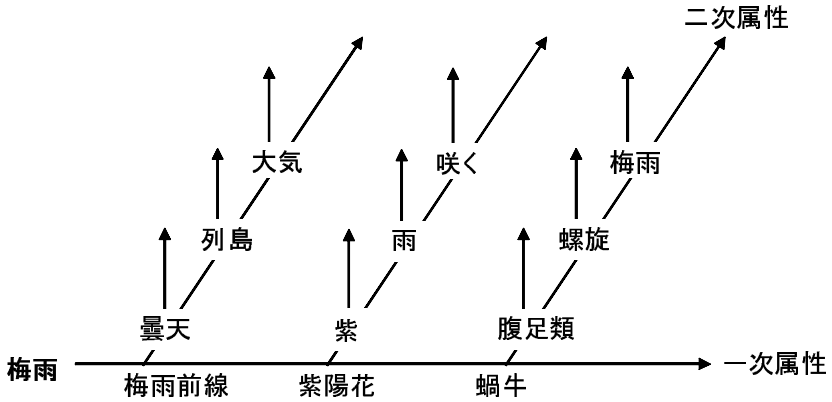


図 2.1: 概念「梅雨」を二次属性まで展開した場合の例

「梅雨」という概念には「梅雨前線，紫陽花，蝸牛…」などといった属性が付与されている．これらの属性を概念「梅雨」の一次属性と呼ぶ．さらに一次属性それぞれからも新たに属性を導くことができ，これを元の概念「梅雨」の二次属性と呼ぶ．つまり概念ベースは属性の連鎖的構造により定義されている．

表 2.1，図 2.1 に示した概念と属性の例からも分かるように，概念ベースにおける概念の定義は，その概念を示す語の語彙的な意味に留まらない．概念との明確な関係性を定義できない語でも，人間が何かしらの関連を見出せる場合には区別無く属性として付与をする．概念ベースのこの構造により，人間らしい連想の可能性を持った知識を構築することが出来る．

2.3 関連度計算方式

関連度計算方式 [9, 10] は 2 つの概念間に人間が見出す関連を定量的に表現する手法である．関連度計算方式ではまず，概念同士がもつ属性それぞれを意味が近いもの同士で対応付ける．そ

の上で対応付けられた属性同士の属性，つまり元の概念の二次属性同士を比較して定義された意味知識の類似度合いを算出する．

まず，属性の対応付けに用いる一致度について述べる．概念 A および概念 B の一次属性をそれぞれ a_i，b_j とし，それに対応する重みを u_i，v_j とする．それぞれの概念が持つ属性数を L 個と M 個 $(L < M)$ とすると，概念 A および概念 B は次のように定義される．

$$A = \{(a_i, u_i)|i = 1\sim\ L\}$$
$$B = \{(b_j, v_j)|j = 1\sim\ M\}$$

このとき，概念 A，B の一致度 $DoM(A, B)$ を以下の式で定義する．

$$DoM(A, B) = \sum_{a_i=b_j} min(u_i, v_j) \tag{2.1}$$

$$min(\alpha, \beta) = \begin{cases} \alpha & (\beta \geq \alpha) \\ \beta & (\alpha > \beta) \end{cases} \tag{2.2}$$

a_i=b_j は属性同士が表記的に一致した場合を示している．つまり一致度とは概念 A と概念 B 双方が共通して持つ属性の内，小さいほうの重みを足し合わせたものとなる．共通した属性は概念 A と概念 B でそれぞれ重み付与が成されている．この重みのうち，小さいほうの重み分は概念 A と概念 B 両方の属性に有効であると考えるためである．つまり一致度とは，双方の概念にとって有効な属性が持つ重みの和を示す数値である．関連度計算方式ではこの一致度を用いて，関連度を算出する概念同士の属性を対応付ける．

この一致度を利用して属性間の最も対応のよい組み合わせを決定した上で，関連度の算出を行う．概念 A と概念 B の関連度を算出することを考える．

まず関連度の算出を行う概念それぞれの属性のうち，双方に共通して存在するものを抽出する．今，概念 A と概念 B 双方に共通して存在する属性が t 個あったとすると，まずその t 個の属性は優先して抽出し，一致するもの同士で対応付けを決定する．この対応付けられた属性から，共通属性関連度 $DoA_{com}(A, B)$ を以下の式から算出する．

$$DoA_{com}(A, B) = \sum_{a_i=b_j} min(u_i, v_j) \tag{2.3}$$

共通属性関連度は式からも分かる通り，共通属性のみを用いた一致度の算出である．一致度では共通属性がもつ重みのうち，小さい方が採用される．ここで，用いられなかった大きい方の重みを以降の計算で活用するために，概念の再定義を行う．具体的には，$a_i = b_j$ の時にそれらの重みの関係が $u_i > v_j$ となる場合には属性 a_i の重みを $u_i - v_j$ に更新した上で属性 b_j を概念 B から除外する．逆に $u_i < v_j$ となる場合には属性 b_j の重みを $v_j - u_i$ に更新した上で属性 a_i を概念 A から除外する．この処理により，それぞれの概念には共通属性が存在しなくなる．この再定義された概念を概念 A' および概念 B' とする．

次に共通属性が除外された概念 A' および概念 B' の関連度 $DoA_{def}(A', B')$ を算出する．所持する属性数が少ない概念 A' を基準とし，その属性の並びを固定する．その上で概念 B' の一次

属性を概念 A' の各属性との一致度の和が最大になるように並び替える．このとき概念 A' の属性と重みを (a'_i, u'_i)，概念 B' の属性と重みを (b'_{x_i}, v_{x_i}) として次のように定義する．

$$A' = \{(a'_i, u'_i)|i = 1 \sim L - T_1\}$$
$$B' = \{(b'_j, v'_j)|j = 1 \sim M - T_2\}$$

このとき，共通属性を除去した概念間の関連度 $DoA_{def}(A', B')$ を以下の式によって定義する．$DoA_{def}(A', B')$ は対応が決定した属性間の重みの比率と重みの平均値を属性一致度に乗じることで，属性一致度を補正する．

$$DoA_{def}(A', B') = \sum_{s=1}^{L-T_1} DoM(a'_s, b'_s) \times \frac{min(u'_s, v'_s)}{max(u'_s, v'_s)} \times \frac{u'_s + v'_s}{2} \tag{2.4}$$

$$max(\alpha, \beta) = \begin{cases} \alpha & (\alpha \geq \beta) \\ \beta & (\beta > \alpha) \end{cases} \tag{2.5}$$

ここで $DoA_{com}(A, B)$ および $DoA_{def}(A', B')$ の合計を，概念 A と概念 B の関連度 $DoA(A, B)$ と定義する．

$$DoA(A, B) = DoA_{com}(A, B) + DoA_{def}(A', B') \tag{2.6}$$

関連度は $0.0 \sim 1.0$ の値をとり，値が高いほど関連の深い語であることを意味する．
表 2.2 に概念 A と概念 B の関連度計算の例を挙げる．

表 2.2: 関連度計算の例

概念 A	概念 B	関連度の値
梅雨	雨	0.3789
梅雨	眼鏡	0.0024
エスカルゴ	パセリ	0.0226
エスカルゴ	ペン	0.0031
紫陽花	蝸牛	0.0261
紫陽花	梅雨	0.0758
蝸牛	梅雨	0.0398

2.4 まとめ

本章では，自然言語を対象とした情報処理技術における人間の連想を表現する機構の必要性およびその困難さについて述べ，それらを機械上に表現することを目的とした語概念連想システムについて示した．

人間は関係性を明確に定義できない語間においても適切な関連を見出すことが出来る．これは人間が自身の知っている事柄，情報，概念といった多種多様の知識から他の知識を思い浮かべる「連想」という能力を有しているためである．この連想は，様々な事物・事象に対して人間が持つ概念的な意味の知識により実現している．この人間が持つ概念的な意味を機械上に定義するためには，語彙的な意味にも分類・体系付けられた構造にも依存しない,「連想できる」という曖昧な関係性を定義する必要がある．そのような曖昧な関係性に着目し，事物・事象を表す自然言語の「語」を概念化した概念ベースについて述べ，また，概念ベースを用いることで算出される関連度計算方式の詳細を記した．

第3章　概念ベースの構築方法

3.1　はじめに

人間は，語と語の間にある曖昧な関連を見出すことが出来る．例えば「紫陽花」と「蝸牛」という２語の間に，私達は何かしらの関連を見出す．しかしこの二語は語彙的な意味では類似しておらず，また二語の間に明確な関係性を持っているわけでもない．それでも私達がこの２語に対してなにかしらの関連を見出せるのは，例えば「雨が降る中，紫陽花の葉に付く蝸牛」のような，今までの経験や日常，世間の常識に裏打ちされた「連想できる」という関係性をこの２語それぞれの意味として持っているためである．

このような「連想できる」という曖昧な関係性により，現実世界に存在する事物・事象を表す自然言語の「語」を概念化したものが概念ベースである．概念ベースではある語に対して，その語から連想できる他の語を属性として付与することで，種々の語を「概念」として定義している．属性として様々な語を付与することで，人間のように曖昧な関連を見出すことが出来る [11].

本章ではこの概念ベースの構築について述べる．概念・属性の取得，属性の精錬，重みの付与，属性追加といった概念ベース構築のための各種処理について示し，それぞれの手法を用いた概念ベースの精度を算出する．

3.3 節では国語辞書を用いた基本概念ベースの構築について述べる．基本概念ベースは国語辞書の見出しを概念，その語義文中の語を属性とした概念ベースであり，属性を概念としても定義することで概念の連鎖的構造を持った概念ベースを構築する．

3.4 節ではルールによる属性の精練の手法とその効果を示す．概念と属性の間に関連があるか否かをルールに従い判別し，概念の意味定義に適切でない属性の削除を行うことで概念ベースの精錬を行う．

3.5 節では新聞記事による基本概念ベースの拡張手法について述べる．基本概念ベースは概念の属性として辞書の語義文のみが用いられており，そのため属性群は概念の直接的な意味を表す語が大半となっている．そこで新聞記事中での語と語の共起を利用し，何かしらの関連が見いだせる新たな語を概念，属性として追加する．また，概念ベースの連鎖的構造を利用した属性への重み付け手法についても述べる．

3.6 節ではシソーラスによる属性の追加手法を示す．語の階層的な関連を木構造により示したシソーラスを情報源として，より人間の感覚に近い属性を概念へ追加する手法について述べる．

3.2　概念ベースの精度評価手法

概念ベースの精度評価には $X-ABC$ 評価セット [12] を用いる．これはある基準概念 X と，この概念 X と関連が非常に強い概念 A，概念 A ほどではないが関連があると思われる概念 B，まったく関連のない概念 C によって構成された評価セットで，人手により作成されている．X・A・B・C の概念を 1 組として，500 組が評価セットとして存在している．表 3.1 に評価セットの例を示す．

表 3.1: X-ABC 評価セットの例

X	A	B	C
飲食店	食堂	米	青空
引き上げる	撤収	敗戦	恨み顔
仲買	仲介	市場	仕舞う
丸	円	図形	決まり
老女	おばあさん	皺	武芸
通り雨	雨	傘	製品
沿岸	海岸	船	練乳
伝導	伝える	熱	一人
...	...	...	...

概念 X と概念 A との関連度を $DoA(X,A)$，概念 X と概念 B との関連度を $DoA(X,B)$，概念 X と概念 C との関連度を $DoA(X,C)$ とする．そして表で示した評価セットにおける $DoA(X,C)$ の平均を $AveDoA(X,C)$ として，それぞれの概念間の関連度の値を比較することで概念ベースを評価する．この評価セットを用いた評価方法として，順序正解率と C 平均順序正解率がある．

順序正解率

各関連度に以下の関係がある場合を正解とする．

$$DoA(X,A) > DoA(X,B) > DoA(X,C) \tag{3.1}$$

この評価を全ての組に対して行った上で，正解となった評価データの組の比率を概念ベースの精度とする．

C 平均順序正解率

概念 X と関連がない概念 C との関連度 $DoA(X,C)$ は，本来 0.0 となるのが理想である．しかし関連度計算方式の特性上，概念 X と概念 C に一つでも共通した属性が存在すれば微小な値が算出されてしまう．そこで概念 C との関連 $DoA(X,C)$ を誤差とみなし，その平均 $AveDoA(X,C)$ を評価セット全体での平均誤差とする．そして $DoA(X,A)$，$DoA(X,B)$，$DoA(X,C)$ それぞ

れの関連度の間に平均誤差以上の差が存在していれば，人間の常識にそった関連度が算出されているとして正解と見なす．具体的には以下の式を満たす場合に正解と見なす．

$$DoA(X, A) - DoA(X, B) > AveDoA(X, C) \tag{3.2}$$

$$DoA(X, B) - DoA(X, C) > AveDoA(X, C) \tag{3.3}$$

$$AveDoA(X, C) = \frac{\sum_{i=1}^{set_{num}} DoA(X_i, C_i)}{set_{num}} \tag{3.4}$$

この評価を全ての組に対して行った上で，正解となった評価データの組の比率を概念ベースの精度とする．本章以降，C 平均順序正解率における評価セット数 $set_{num} = 500$ となっている．

3.3 国語辞書からの基本概念ベース構築

国語辞書から概念と属性を取得することで構築される，基本概念ベースの構築手法 [6, 7] について述べる．

この概念ベースでは国語辞書 [13–18] の各見出し語を概念と定義し，見出し語それぞれの語義文中の語を属性候補として抽出する．次にその属性候補に対して属性信頼度と呼ばれる指標による選別を行い，適切な属性のみを残す．この属性に対して，学習データによる重みを付与することで基本概念ベースは構築される．

3.3.1 国語辞書からの概念・属性の抽出

国語辞書には，見出し語とその意味を説明する語義文が存在する．この見出し語を概念と見なし，概念を説明する語義文中の自立語を属性として抽出する．属性それぞれには出現頻度による重みを付与する．

次に,「属性の持つ属性」および「概念を属性として持つ概念」を新たな属性として追加する．具体的な例で説明すると，まず概念「馬」を定義する際には，国語辞書の見出し語「馬」がもつ語義文から属性として「家畜，たてがみ，…」といった語を取得することができる．同様に国語辞書中の見出し語を概念化すると，属性として得られた「たてがみ」も概念として定義される可能性がある．概念「たてがみ」ももちろん語義文による属性付与がされているため，その属性を元の概念「馬」の属性としても用いる．これが「属性の持つ属性」の追加である.「概念を属性として持つ概念」の追加では,「馬」という語を属性としてもつ概念，例えば概念「競馬」がそれに当たるとすると，概念「馬」の属性にも「競馬」を追加する．

最後に,「概念化されていない属性」および「全ての概念に出現する属性」,「ある閾値以下の重みの属性」を削除する.「概念化されていない属性」とは，語義文中には出現したが見出し語としては辞書に無い語を指す．また,「全ての概念に出現する属性」は日本語特有の言い回しに

由来する語を指し，例えば「こと」や「もの」といった語である．この属性削除の後，所持属性数が閾値より少ない概念を概念ベースから削除する．

以上の処理により，国語辞書からの概念および属性の抽出が完了する．

3.3.2　属性信頼度の算出

属性信頼度とは，得られた属性が概念にとってどれぐらい信頼できる情報かを表す値で，人手によるサンプル評価を用いて算出される．この信頼度によって各属性をクラス分けすることで，そのクラスに従い後述する重み付けを行う．

信頼度の算出は，概念と属性の関連を見出すための 6 つの手掛かりと，ランダムに選んだサンプル概念 100 語の目視評価結果を組み合わせることで行う．

まずサンプル概念の持つ属性を「適切，どちらでもない，不適切」の 3 段階に分けて評価する．これは 3 名の目視によって行われ，誰にも不適切と判断されなかった属性を「適切属性」とする．

次に，6 つの手掛かりそれぞれについて，その手掛かりが合致したときに，属性がどれぐらい信用できるかを表す信頼度を設定していく．それぞれの手掛かりとその信頼度は以下のようになる．

概念と属性の一致

概念と属性の表記が完全一致している場合を指す．この場合，概念と属性の関連は疑いようが無いため，信頼度は 100%となる．

辞書中の出現によって付与された属性の重み

3.3.1 節において付与された属性それぞれの重みの値によって関連の度合いを測る．サンプル概念の属性が持つ重みの値から，重みと適切属性の割合を算出して，この割合を信頼度とする．例えば重み 0.02 以上 0.03 未満の属性のうち，適切属性の割合は 60%であった．この割合が重み 0.02 以上 0.03 未満の属性の信頼度となる．

概念と属性の関連度

概念と属性の間で関連度を算出し，その値から関連の度合いを測る．前述した重みと同じく，関連度の値と適切属性の割合を信頼度とする．

概念と属性の漢字一致

概念と属性の表記において漢字の一部が一致している場合を指す．サンプル概念においてこの手掛かりに合致する属性のうち，適切属性の割合は 90%となったため，これを信頼度とする．

相互属性

概念 A の属性 a_i を概念として見たとき，a_i の属性に A が存在する関係を指す．概念 a_i の属性 A の重みを用いて，前述した重みによる手掛かりに従った信頼度を得る．

語関係データにおいて定義される概念と属性の関係

国語辞書から構築された語の関係を示すデータにおいて，概念と属性が同義，類義，上位下位のいずれかの関係にある場合を指す．明確に関係を定義されているため，信頼度は 100%とする．

3.3.3 信頼度によるクラス分けと学習による重み付与

ある概念の属性の信頼度を算出し，その信頼度の値に従って属性のクラス分けを行う．各クラスの条件は表 3.2 のようになる．

表 3.2: 信頼度によるクラス分け

クラス	信頼度 (%)
信頼度 1	80 以上
信頼度 2	60 以上 80 未満
信頼度 3	40 以上 60 未満
信頼度 4	20 以上 40 未満
信頼度 5	0 以上 20 未満

このクラスごとに，重みの付与を実験的に行う．信頼度 1 の属性は重み 1.0，信頼度 1 かつ同義・類義・上位下位いずれかの関係にある属性は重み 1.0~ 16.0，信頼度 2~ 5 の属性に関しては重み 0.0~ 1.0 のパターンを用意し，組み合わせることで実験を行う．実験は 2 章 3.2 節で示した $X - ABC$ 評価セットを重み学習用に作成し，このセットにおいて C 平均順序正解率が最も高くなったときの重み付けを採用する．

以上の処理により，国語辞書からの基本概念ベースの構築が完了する．

3.3.4 基本概念ベースの精度評価

基本概念ベースの精度評価を図 3.1 に示す．

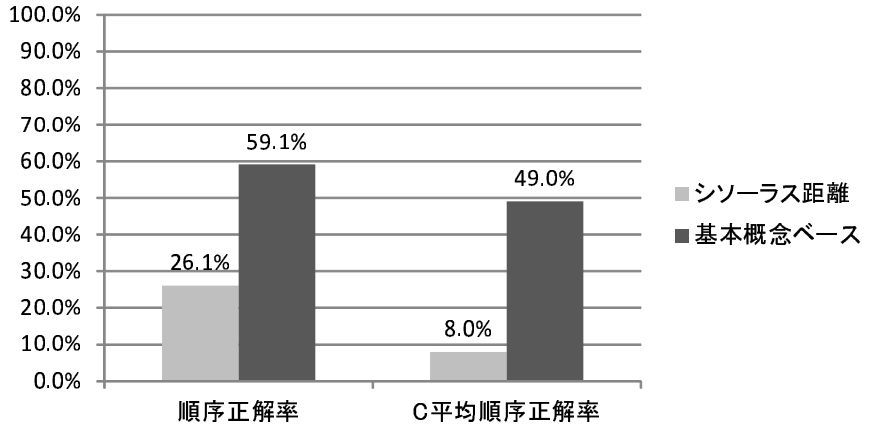


図 3.1: 基本概念ベースの評価結果

評価には 2 章 3.2 節で示した 500 セットの $X-ABC$ 評価セットを用いた．なお，比較対象としてシソーラス距離 [19] による類似度計算の結果を示した．順序正解率，C 平均順序正解率ともにシソーラス距離と比べて高い精度を示していることがわかる．評価セットは人手で作成されたものであり，この結果から概念ベースの構造は人間らしい関連性の判断により適していることが示される．

構築される基本概念ベースは，概念数が 33699 語，1 概念あたりの平均属性数が 45 語となった．表 3.3 に概念および属性の例を示す.

表 3.3: 定義された概念と属性の例

概念	属性	属性数
蝸牛	蝸牛, 隅, 殻, 争い, 巻き貝, 同体, 陸生, 触, 触角, 綱, 狂言, 対, 争う, 雌雄, 方, 先端, 異称, 二, 小国, 螺旋, 上がり, 知る, 寓話, 貝, 種類, 夏, 同じ, 体, 長い, 眼目, 目, 卵生, 一, 陸上, 明暗, 判別, 総称, 氏, 頭, 中, 渦巻き, 山伏, 柔軟, 若葉, 一群れ, 形, 喰う, 眼, 多い, 右, 湿気	51
紫陽花	紫陽花, 花, 萼, 葉, 低木, 改良, 球状, 対生, 色, 花弁, 咲く, 発達, 枝, 梅雨, 瘡, 青, 鋸, 解熱, 不実, 集散, 海岸, 空色, 紫, 生, 歯, 序で, 藍, 夏, 幹, 桜, 初夏, 中性, 密, 四つ, 品種, 叢生, 美しい, 頃, 集まる, 治療, 球形, 序に, 自生, 荒い, 変える, 観賞, 集まり, 装飾, 薬用, 芸人, 形造る	51
髑髏	髑髏, 野晒し, 晒し, 風雨, 頭, 頭蓋骨, 頭骨, 骨, 白骨, 意, 同じ, 肉, 頭髪, 首, 寮, 数える, 鳶職, 雨, 部分, 布, 曲がり, 上がり, 里人, 晒し, 木綿, 箏曲, 頭金, 風, 初め, 動物, 不承知, 脳, 出家, 気に入る, 考え, 首領, 親方	36
眼鏡	眼鏡, 見分ける, 目, 視力, 調整, 望遠鏡, 力, 遠眼鏡, 器具, 善悪, 輪, 女, 光線, 遠視, 不完全, 防ぐ, 鑑識, 近視, 保護, 乱視, 見, 江戸, 色, 物, 眼, 強い, 老眼, 眼識, 髪, 見える, 髻, 一, 良否, 髪型, 判断, 時代, 顕微鏡, 凹, 可否, 色眼鏡, 二分, 正しい, 狂う, 監視, 二, 誤る, 定める, 入る, 物事	49

なお，評価セットは人手で作成されているため，シソーラスおよび基本概念ベースに存在しない語句も含まれている．基本概念ベースでは $X-ABC$ 評価セット中の 2000 語中，167 語が概念として定義されていなかった．概念として定義されなかった語の一部を図 3.2 に示す.

飲食店, 有力者 , 御手洗 , 農作物, 新月, 二進法, 公用語, 弁護人, 量子, 回覧板, 芸術的, 量子, 実存 , 文学者, 師匠, 満員電車 , 義務教育 , じゃが芋 , 肉体労働, 北極圏 , 憧れ …

図 3.2: 概念定義されなかった語の一部

3.4 ルールによる属性精練手法

基本概念ベースの属性は国語辞書の語義文に出現する語から得られる語群であるが，属性として採用するか否かの選別は辞書中の出現頻度のみによって判断されている．この属性の選別にルールを定めることで，属性の精練を行う手法 [20,21] について述べる．

ルールによる選別では，まず基本概念ベースに対して 4 つのルール群による精練を行い，精錬用概念ベースを作成する．この精錬用概念ベースにより算出される関連度とルール群を組合せることにより属性のランク分けを行い，属性の選別を行う．

3.4.1 ルール群による精錬用概念ベース作成

ルール群を用いて属性の選別を行い，精錬用概念ベースを作成する．ルールは属性信頼度を算出する際に利用した手掛かりのうち,「概念と属性の漢字一致」「相互属性」「語関係データ」を用いる．さらに「シソーラスにおいて概念と属性が上位・下位・仲間の関係にある」というルールを加え，これらのルールのどれにも適合しなかった属性を削除する．この処理により作成された概念ベースを精錬用概念ベースとして利用し，基本概念ベースの属性の選別を行う．

3.4.2 関連度とルール群の組合せによる属性のランク分け

前節で作成された精錬用概念ベースを利用して，基本概念ベースに定義される概念と属性の関連度を算出する．この関連度を前節のルールに加え，基本概念ベースの属性のランク分けを行う．具体的には,「漢字一致」「相互属性」「語関係データ」「シソーラス」それぞれのルールに合致する属性と概念の関連度を算出し，それがある閾値以上ならば属性として適切と判断する高ランクに，そうでない場合は低ランクに分類する．また,「関連度」そのものもルールに加える．これはある概念 A の属性 a_i の関連度が閾値以上であった場合を適切と判断する．

ただし各ルールにより適切な閾値は異なるため，ここでは 3.3.2 節で述べたサンプル概念による適切属性を利用する．これにより，例えば「漢字一致」のルールに沿う適切属性の関連度がどれ位の値かを算出することが出来る．関連度の閾値を変化させることで，この適切属性をどれ位の割合で高ランクに選べるかをルールごとに実験して閾値を決定する．図 3.3 にランク分けの流れを示す.

ここで，$J1\_High$，$J1\_Low$ はそれぞれ「漢字一致」「相互属性」「語関係データ」「シソーラス」「関連度」のルールのうち，どれかに合致した属性の選別を行うための閾値を指す．前述したとおり，合致したルールにより閾値は異なる．$J1\_High$ はそれぞれのルールにおいて適切属性が 8 割を保つ閾値，$J1\_Low$ は 6 割を保つ閾値と設定している．また $J2$ は関連度を用いた特別ルールで,「概念 A と属性 a_i の関連度が閾値より高く，属性 a_i 以外の属性との関連度と比べて十分高い」場合に適切と判断する．このルールの閾値設定は適切属性が 3 割を保つ値としている．N は各属性の適合ルール数を指し，多くのルールに適合する場合にはランクを上位に上げる．以上のようなランク分けを行い，ランク C と判断された属性は不適合と判断して概念ベースから削除することで，ルール精練概念ベースが構築される．

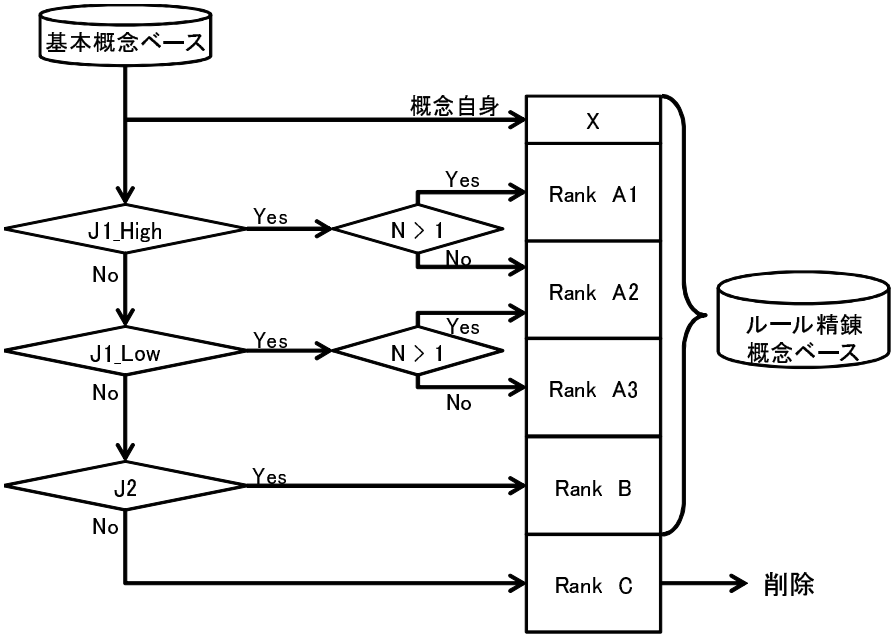


図 3.3: ルールによるランク分けの流れ

3.4.3　ルール精練概念ベースの精度評価

図 3.4 にルール精練概念ベースの精度評価を，表 3.4 に基礎概念ベースから駆除された属性の例を示す.

ルールによる属性の精練によって，概念ベースの精度が向上したことが分かる．ただし，概念数は基本概念ベースと変わらないため，本節の評価でも $X - ABC$ 評価セット中の 2000 語中，167 語が概念として定義されていない．この精練処理により，概念当たりの平均属性数は約 29 語となり，基本概念ベースと比較して 4 割弱の属性が削除されている.

表 3.4: 概念から削除された属性例

概念	削除された属性例
蝸牛	狂言, 争う, 二, 小国, 夏, 眼目, 明暗, 氏, 中, 山伏, 柔軟, 若葉, 喰う, 眼, 多い, 右, 湿気
紫陽花	発達, 梅雨, 瘧, 鋸, 解熱, 不実, 海岸, 生, 歯, 序で, 形造る, 密, 四つ, 叢生, 頃, 集まる, 治療, 荒い, 変える, 集まり
髑髏	同じ, 寮, 鳶職, 頭金, 気に入る
眼鏡	輪, 女, 防ぐ, 江戸, 強い, 髪, 髻, 髪型, 時代, 二分, 狂う, 誤る, 定める

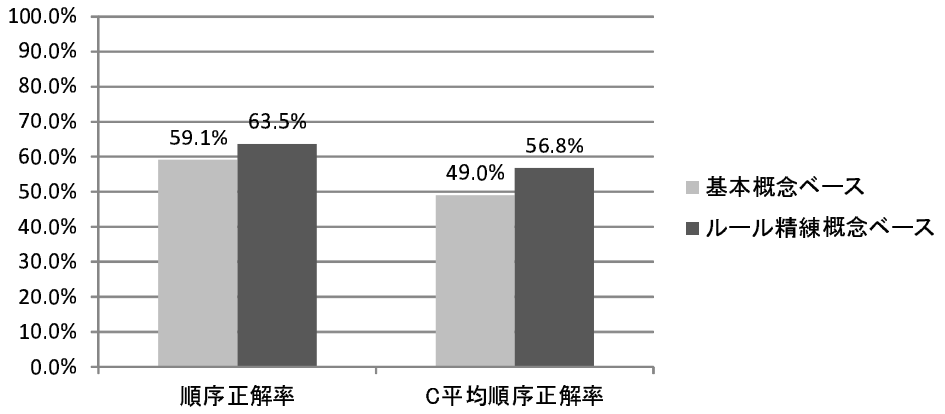


図 3.4: ルール精練概念ベースの評価結果

3.5 新聞記事によるルール精練概念ベースの拡張

ルール精練概念ベースを構築する際に利用した国語辞書には，見出し語が約 20 万語登録されている．しかし属性の取得範囲が語義文のみであるため，適切に属性が取得できない概念が多々存在し，結果として概念数は 33699 語に留まっている．そのため，人間が自然と考え付く語句によって構成される $X - ABC$ 評価セット中の語にも，概念化されていないものが存在する．これは人間が日常的に扱う語の意味知識を定義できていないということになり，知識が不足していることは否めない．また，国語辞書の語義文は見出し語の語彙的な意味を説明する文であるため，概念ベースが目指す連想のための知識表現には至っていない．

そこで，概念ベース構築の情報元として新聞を加えることでルール精練概念ベースの拡張を行う [22]．新聞記事中には日常で人間が用いる語が散乱しており，また記事中で互いに共起しあう語を概念と属性の関係に見立てることで，概念ベースの意味知識に必要となる「何かしらの関連がある語」を取得することが期待できる．

3.5.1 国語辞書と新聞記事を用いた概念と属性抽出

情報元として新聞記事を用いることで，国語辞書のみでは属性が付与されず，概念とならなかった見出し語に対しても概念化できる可能性が出てくる．そこでまず，国語辞書に記載されている見出し語のうち，概念化に適した語のみを抽出する．概念は単独で何かしらの意味をもつ必要があるため，ここでは名詞・動詞・形容詞・形容動詞いずれかの見出し語を概念化する語句として抽出する．これらの概念には，もちろん基本概念ベースにおいてすでに概念化されている語も含まれているが，新聞記事からの属性取得は基本概念ベースに存在するか否かを区別せずに行う．つまり基本概念ベースにおいてすでに定義された概念に関しては，新聞からさらに属性が追加されることになる．

新聞記事からの属性取得は新聞記事内での語と語の共起関係を利用する．ある一定の記事範囲において共起する語同士には，何かしらの関係があると考えてそれらを属性として取得する．共起を判別する記事範囲については句読点によって区切られた領域とし，その範囲で互いに共起する語同士をそれぞれ概念および属性の関係と見なす．新聞記事の例を図 3.5 に示す．

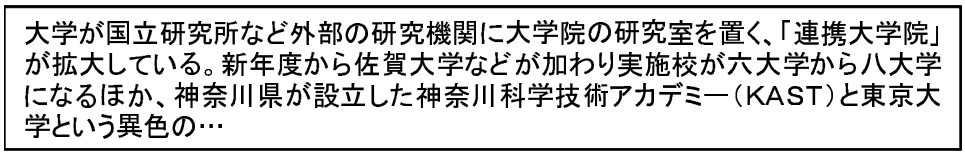


図 3.5: 新聞記事の例

句読点を範囲の区切りとするため，例では「大学が国立研究所など外部の研究機関に大学院の研究室を置く」が共起の範囲となる．この範囲中に存在する名詞・動詞・形容詞・形容動詞を形態素解析により抽出し，それぞれある 1 語を概念，それ以外を属性と見なすことで概念と属性のセットが出来る．表 3.5 に作成される概念と属性のセットの一部を示す．

表 3.5: 新聞記事から得られる概念と属性のセットの一部

概念	属性 1	属性 2	属性 3	属性 4	属性 5	属性 6	属性 7	属性 8
大学	国立	研究所	外部	研究	機関	大学院	研究室	置く
国立	研究所	外部	研究	機関	大学院	研究室	置く	大学
研究所	外部	研究	機関	大学院	研究室	置く	大学	国立
外部	研究	機関	大学院	研究室	置く	大学	国立	研究所
...	...	...	...	...	...	...	...	...

概念「大学」の属性として，共起の範囲に含まれる「大学」以外の語が取得されている．ただしこの時に新聞記事から抽出される概念および属性の語は，概念化する語として国語辞書から選ばれたもののみである．

3.5.2　新聞記事での出現頻度による擬似重みの付与

新聞記事から得られた属性には重みが付与されていない．そこで共起回数から擬似的な重みを付与する．

まず 3.5.1 節に示した属性に対して，概念との共起回数を付与する．共起回数は新聞記事全体を句読点で区切った範囲毎に概念と属性が存在する回数を数える．これにより，ある概念の一次属性の共起回数を合計すれば，それはその概念の新聞記事全体での出現頻度と等しくなる．共起回数を付与した概念ベースのイメージを図 3.6 に示す．

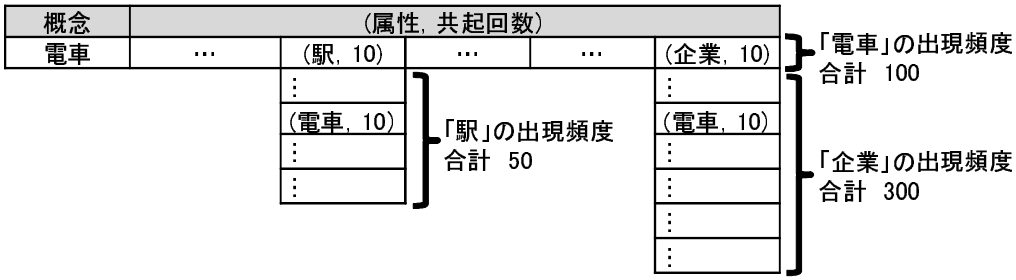


図 3.6: 共起回数を付与した概念ベースのイメージ

概念「電車」の属性である「駅」と「企業」について考える．この属性は共に概念「電車」との共起回数が 10 回である．そのため，共起回数をそのまま重みに利用した場合にはこれらの属性の価値は等しくなる．ここで属性「駅」を概念と見なして一次属性を取得し，その共起回数の合計，つまり「駅」の出現頻度を算出した結果 50 回だったとする．属性「企業」についても同じように出現頻度を算出すると 300 回だったとする．すると概念「電車」にとって，記事全体に対し出現頻度が多い「企業」という語との 10 回の共起と，あまり多く出現しない「駅」という語との 10 回の共起の価値が同じである事は妥当ではない．そこで属性に対する擬似重み付与の手法として，相互情報量を用いる．概念 A の属性 a に対する擬似重み W_anp は次の式で定義される．

$$W_anp = \log_2 \frac{q_{Aa}}{\sum_k q_{Ak} + \sum_k q_{ak} - q_{ak}} \tag{3.5}$$

ここで q_{Aa} は概念 A と属性 a の共起回数，$\sum_k q_{Ak}$ は概念 A の一次属性の共起回数の合計，つまり概念 A の出現頻度，$\sum_k q_{Ak}$ は属性 a を概念と見なした際の概念 a の一次属性の共起回数の合計，つまり概念 a の出現頻度を指す．分母において q_{ak} を減算しているのは，概念 A の一次属性中には属性 a が，概念 a の一次属性中には属性 A が存在するため，共起回数を重複して加算してしまうためである．以上の定義による擬似重みを属性に対して付与する．

3.5.3 ルールによる属性精練手法と概念ベース idf による重み付け

3.5.2 節において付与された擬似重みを用いて関連度を算出することで，3.4 節のルールによる属性精練のランク分けを行うことが可能になる．さらにこの精練により得られる属性のランクを利用して，重みの付与を行う．

まずサンプル概念に対してルールによる属性精練のランク分けを行い，それぞれのランクの適切属性の割合を調査する．表 3.6 に適切属性の割合を示す．

次に新聞から得られた属性も同じようにランク分けを行う．それぞれの属性の重みとして，表に示したランクごとの適切属性の割合を付与する．なお 3.4 節のルールによる属性の精練では，

表 3.6: ルール毎の適切属性の割合

ランク	適切属性の割合
X	1
$A1$	0.84
$A2$	0.74
$A3$	0.57
B	0.33
C	0.13

ランク C を不適合属性として削除したが，新聞記事による拡張は概念及び属性の拡充が目的であるため，削除は行わずに重みを小さくすることで対処する．

次に，概念ベース *idf* の算出を行う．*idf* とは対文書頻度の事であり，一般的には「様々な文書が存在する空間で，ある特定の文書にしか出現しない語句は重要である」という事を示す指標である．これは空間中の総文書数を，語句が出現する文書数で割ったものの対数によって表される．この考えを概念ベースに対応付けたものが概念ベース *idf* である．まず空間中の 1 文書を，1 概念が持つ属性群と考える．つまり概念ベースに定義されている概念の総数が全文書数となる．よってある概念 X の概念ベース *idf* は，全概念数を概念 X が属性として出現する概念の数で割ったものの対数で表される．具体的な式は以下のように定義される．

$$CV_N(X) = \log_2 \frac{V_{all}}{df_N(X)} \tag{3.6}$$

$CV_N(X)$ は N 次属性空間内における概念 X の概念ベース *idf* である．V_{all} は概念ベースに定義されている全概念数，$df_N(X)$ は N 次属性集合内において概念 X を属性として持つ概念の数である．概念ベースでは概念の持つ属性の範囲を，N 次展開により広げることが出来る．この重み付けでは概念ベース *idf* は 1 概念が持つ属性群の範囲を二次属性まで展開した状態で算出している．

上記の式より，概念ベース *idf* はその値が大きいほど概念ベース内での出現頻度が少ないということになる．そのため概念ベース *idf* の値が大きい属性は，その概念にとって重要な意味を持つと考えられる．例えば概念「色」を 2 次属性空間内で属性として持つ概念数と，概念「紫色」を属性として持つ概念数を比べた場合，後者のほうが少ない．つまりこの 2 次属性空間において,「紫色」という属性の方が「色」という属性よりも概念を強く特徴付けているといえる．

各々の属性の概念ベース *idf* と，ランク分けによって付与された重みを掛け合わせたものを属性の新たな重みとする．以上の処理により，新聞記事を用いた基本概念ベースの拡張が行われる．

3.5.4 新聞概念ベースの精度評価

新聞概念ベースの精度評価を図 3.7 に示す.

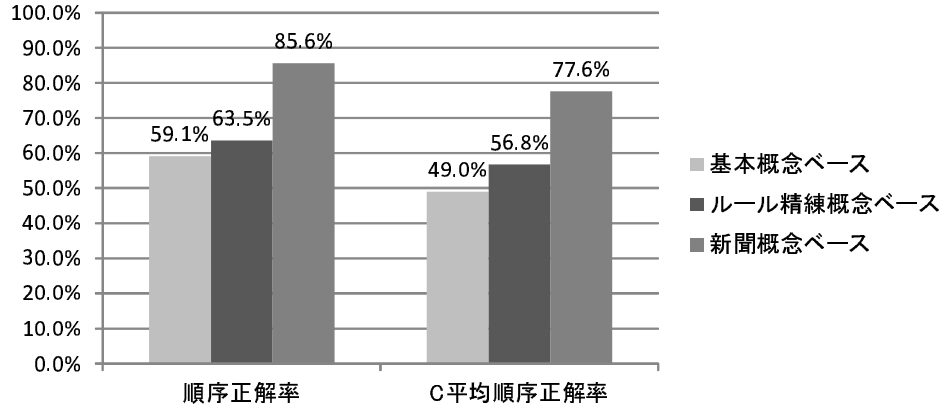


図 3.7: 新聞概念ベースの評価結果

概念及び属性の拡張により，新聞概念ベースの概念数は 87242 語，平均属性数は約 138 語となった．また，基本概念ベース，ルール精練概念ベースにおいて定義されていなかった 167 語についても全て概念化され，人間が日常で用いる語に足る知識ベースとなった．表 3.7 に追加された概念および属性の例を示す.

表 3.7: 追加された概念および属性例

概念	追加された属性例
蝸牛	軟体動物, 舞舞, 腹足類, 蛞蝓, 毎々, エスカルゴ, 蝸牛殻, ...
紫陽花	山茶花, 南天, 紫色, 紅, 桃, 独活, 浅葱, 落葉, 濃い, 毬, ...
髑髏	骨揚げ, 曝す, 脊椎, 雨風, 造血, 距, 革, ...
眼鏡	眼力, 鑑定, 真贋, 鑑別, 近眼鏡, 真偽, 裸眼, ...
エスカルゴ (追加概念)	エスカルゴ, 鮑, 蛤, 薊, 蝸牛, 養殖, 大型, 唐辛子, 大掛かり, 食用, フランス, 料理, 刷る, 使う
パセリ (追加概念)	オランダ芹, パセリ, 人参, 山葵, 西洋料理, 楓, 草, 緑色, 緑, 洋食, 黄, 香気, 越年, 葉, 競る, 裂ける, 夏, 細かい, 科, 薄い, 就ける, 似る, 使う, 在る

3.5.5　サンプルを用いない重み付け手法とその精度評価

前節までの手法で構築される概念ベースの重み付けには，人手によるサンプル概念の評価結果が必要だった．しかし新聞記事などの巨大な情報元を用いて概念ベースの規模が大きくなっていくと，少ないサンプルの結果に依存した重み付けでは適切な結果が得られ辛くなる．そこでサンプルに依存しない重み付け手法として，関連度と概念ベース *idf* による重み付けを行う．対象とする概念ベースは前節の新聞記事による拡張を行った概念ベースとする．重み付けにおいてサンプル概念が必要である部分は，3.5.3 節に示したランクごとの適切属性の割合である．提案する重み付けでは，この部分に概念と属性の関連度を用いる．つまりある概念 A の属性 a_i の新たな重み w_i は，概念と属性の関連度 $DoA(A, a_i)$ に 3.5.3 節で示した概念 A の概念ベース *idf* を掛け合わせた以下の式となる．

$$w_i = DoA(A, a_i) \times CV_N(a_i) \tag{3.7}$$

概念ベース *idf* の次元 N に関しては，N=1~ 4 までを実験的に行った結果，N=3 の場合が最も精度が良い結果となった．この重み付けによる精度評価結果を図 3.8 に示す．

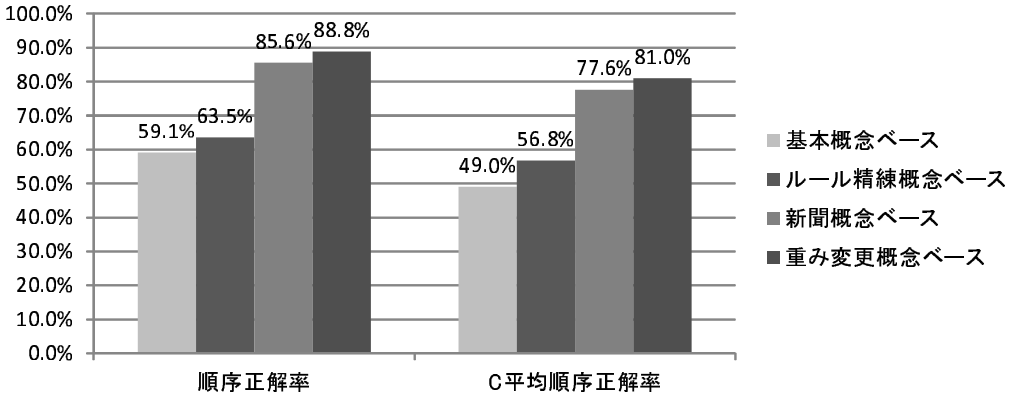


図 3.8: 重み変更概念ベースの評価結果

順序正解率，C 平均順序正解率ともに精度が向上しており，サンプル概念を用いない重み付けの有効性が示された．この重み付け手法は概念ベース以外の情報元を必要としないため，他の手法による属性追加においても汎用的に用いることが出来る．

3.6　シソーラスによる属性の追加手法

概念ベースの意味知識をより充実させるためには，属性に様々な分野・観点の語を持たせる必要がある．前節までの概念ベース構築においては，国語辞書の語義文から概念の直接的な意味を表す属性群を，新聞記事中の語句から何かしらの関連があると考えられる多様な属性群を取

得した．基本概念ベースと比較して，新聞による拡張後は概念，属性の数も大幅に増加し，精度も向上している．よって他の様々な情報元からも概念に対して何かしら関連がある語を取得し，属性数を増加させることで概念ベースの精度向上が期待できる．そこで本節ではシソーラス [1] を用いた属性の追加を行う手法について述べる [8]．シソーラスは人手により作成された語の関係を表す情報であるため，シソーラス上の関係を利用することにより人間の感覚に近い属性が取得できると考える．

3.6.1 シソーラス

シソーラスとは日本語語彙体系から作成されており，名詞の意味用法を木構造で表したものである．2710 個の意味属性 (ノード) の上位-下位関係，全体-部分関係が木構造で表され，ノードに属する名詞として約 13 万語のリーフが登録されている．シソーラスの一部を図 3.9 に示す.

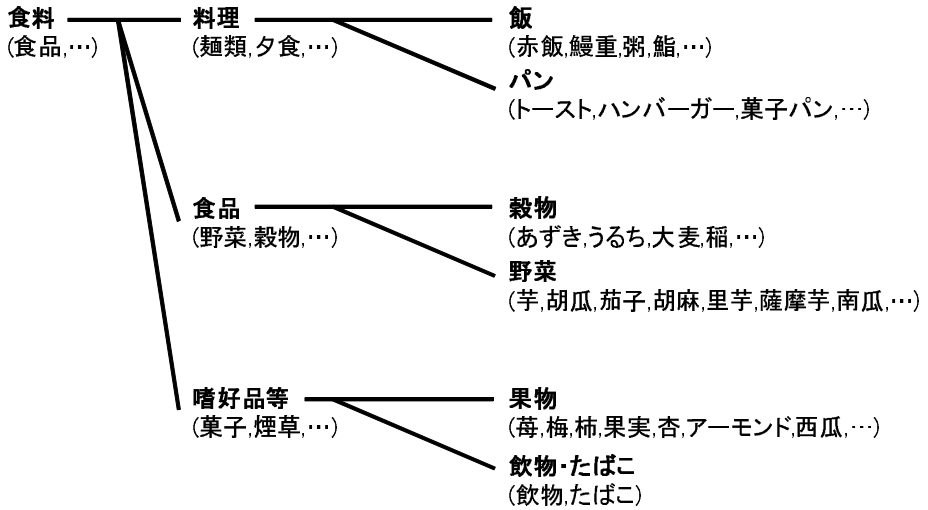


図 3.9: シソーラスの一部

太字で示した語がノード，ノードの下の括弧内がリーフを表しており，左側がより上位，つまり抽象的な語になっている．リーフ「芋」に着目すると，このリーフの一階層上位の関係にある語はノード「野菜」となる．また，一階層上位に共通のノードを持つ語同士を仲間関係にあると定義する．リーフ「芋」であれば,「胡瓜，里芋，南瓜…」といったような語が仲間関係になる．

3.6.2　属性候補の選別と追加

まず概念をシソーラスのノードもしくはリーフと対応付けた上で，上位・下位・仲間関係にあるノードもしくはリーフを属性候補として抽出する．この時，概念のシソーラス上での対応付けは基本的に表記一致により行う．また，シソーラス上に存在しない概念に関しては属性追加の対象としない．ただしシソーラス上のノードには特有の表記によるものがあり，そのままでは対応付けできない．そこで以下に示すルールによりシソーラス上の表記と概念表記を対応付ける．

記号の削除と分割

「乗り物 (本体 (移動 (陸圏)))」や「飲物・たばこ」のようにノード表記中に記号が存在する場合には，その記号を削除した上で概念との対応を取る．この時，前述のような括弧記号によるノードの詳細情報の記述に関しては，先頭の語句のみを抽出して対応を取る．後者の複数の語の併記に関しては記号により語を分割し，それぞれで表記一致を取る．

「等」の削除

「嗜好品等」のように，末尾に「等」が付いた語句の総括を表すノードに関しては,「等」を削除して対応を取る．

シソーラス上に対応付けられた概念の属性候補として，各概念から見て 3 階層を上限とした上位関係の語，1 階層を下限とした下位関係の語，ならびに仲間関係に当たる語を取得する．例えば概念「芋」の属性候補をシソーラスから取得すると，図 3.9 より上位関係の語ならば「野菜，食品，食料」といった語句が得られる．この条件に従い，1 概念あたりにシソーラス上から取得できる属性候補の数は表 3.8 のようになる．

表 3.8: 1 概念あたりの追加属性候補数

	最大	平均
3 階層までの上位関係	38	4.0
1 階層までの下位関係	639	36.9
仲間関係	1423	150.3
合計	1429	152.9

ここで下位関係，仲間関係より得られる属性候補は非常に数が多く，また仲間関係に関しては概念の意味知識としては適さない語も多く存在する．図 3.9 を見ると，概念「苺」の仲間関係には「アーモンド」が存在するが，人間はこの両者に対して深い関連は感じない．そこでこれらの属性候補に対して選別処理を行う．属性の選別には重みを利用する．シソーラスから取得した属性候補群に対して，3.5.5 節で述べたサンプルを用いない重み付け手法により重みを付与する．この重みの降順に属性候補を並べ，上から順に一定個数を追加することで適切な属性

のみを概念ベースに付与する.

3.6.3 シソーラス概念ベースの精度評価

シソーラスから属性追加を行った概念ベースの精度評価を図 3.10 に示す. 下位関係および仲間関係にあたる属性候補の追加個数に関しては実験的に精度を求め, 結果として下位関係の語は重み順に 10 個, 仲間関係は重み順に 15 個を属性として追加した場合が最も高い精度となった.

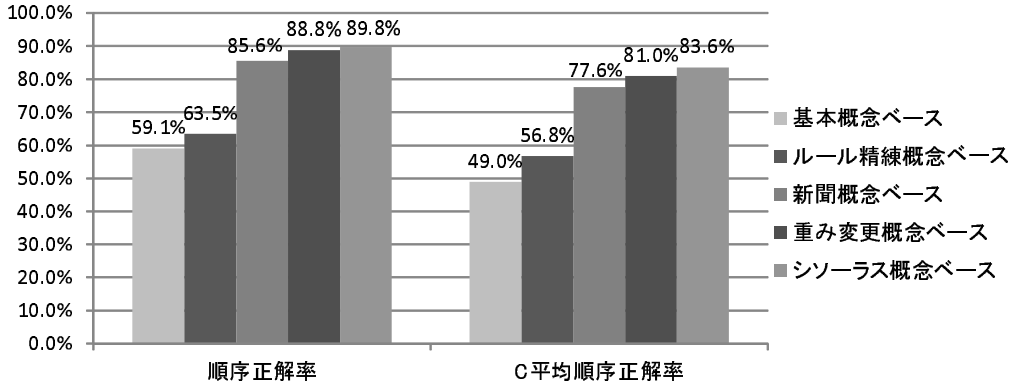


図 3.10: シソーラス概念ベースの評価結果

両評価において精度が向上しており, 適切な属性が追加されたことが分かる. 1 概念の平均属性数は約 37 個となり, 新聞概念ベースと比較して約 3 割の増加となった. 表 3.9 に追加された概念および属性の例を示す.

表 3.9: 追加された属性例

概念	追加された属性例
蝸牛	蜷, 法螺貝, 螺, 牡蠣, 飯蛸, 帆立貝, 子安貝, ...
紫陽花	花水木, 夾竹桃, 木槿, 石南花, 寒椿, 枸杞, ...
髑髏	骨揚げ, 曝す, 脊椎, 雨風, 造血, 距, 革, ...
眼鏡	膝蓋骨, 顱頂骨, 枯骨, 膝皿, 鼻骨, 骨幹, 腕骨, ...
エスカルゴ	帆立貝, 常節, 牡蠣, 法螺貝, 腹足類, 赤貝, 烏貝, ...
パセリ	野菜, 食料, 食品, 作物, 植物, 根芹, 水菜, 春菊, ...

3.7　各概念ベースの $X - ABC$ 評価における関連度の変化

3.3 節から 3.6 節までで述べた各概念ベースについて，$X - ABC$ 評価において算出される関連度を図 3.11 に示す.

$DoA(X, A)$ は評価セットの基準概念 X と概念 A の関連度を表す．他も同じく，それぞれ基準概念 X と概念 B，基準概念 X と概念 C との関連度を表しており，図には評価セット全体での平均値を示している．なお，基本概念ベースおよびルール精練概念ベースに関しては評価セット内の概念が存在しない場合のセットを除いた平均となっている．

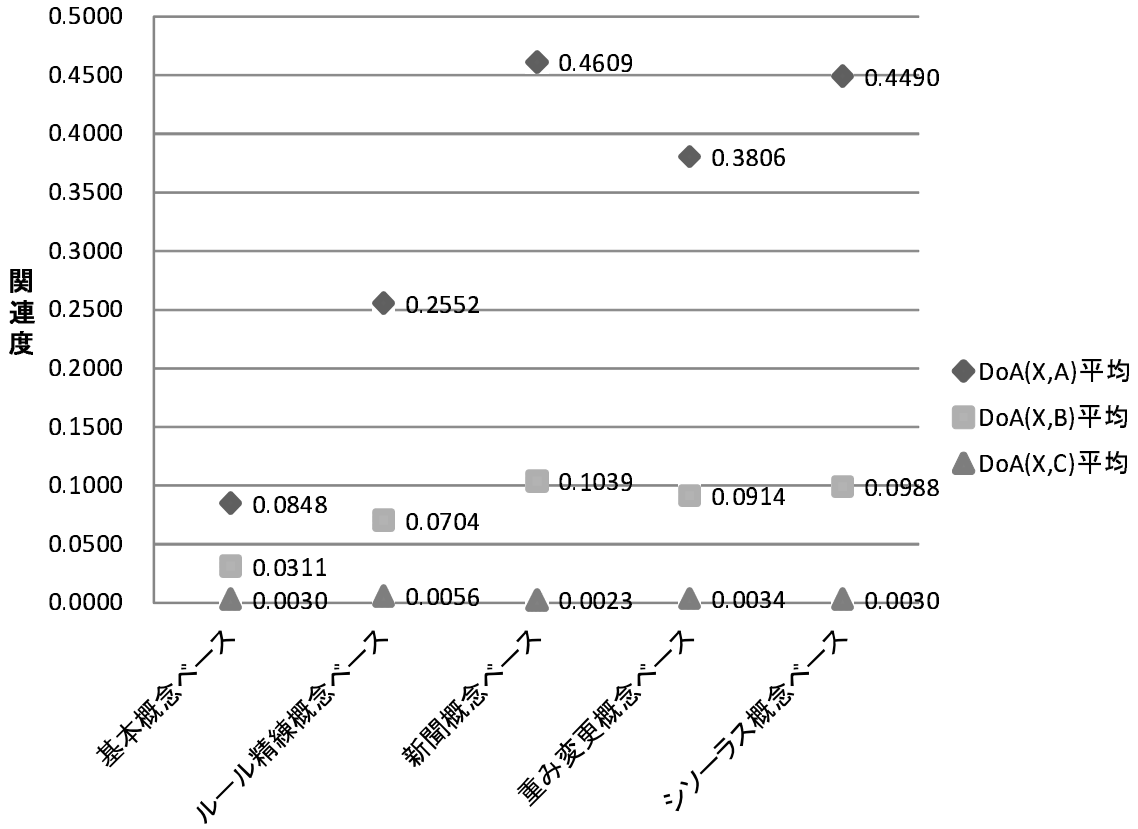


図 3.11: 各概念ベースの関連度

$X - ABC$ 評価セットは基準概念 X との関連が適切に表現できているかを評価するセットである．概念 A が最も基準概念 X と関連が強く，概念 B は概念 A ほどではないが人間なら関連を見出すことが出来る語である．そして概念 C は基準概念 X とまったく関係の無い語となっている．つまり基準概念 X との各関連度は，評価セット全体において概念 A とのものが高く，概念 B とのものが中程度，概念 C とのものが理想上 0.0 になり，かつそれぞれの関連度の値に大きな差があれば概念ベースの精度向上が認められる．

図 3.11 に示したそれぞれの関連度の平均値を見ると，基本概念ベースと比較して全ての概念ベースにおいて概念 A との関連度が飛躍的に高くなっており，各種拡張手法や精練手法が適切かつ大きな効果をもたらしていることが分かる．新聞概念ベースでは概念 A，概念 B それぞれとの関連度が大きく上昇しているが，これは概念および属性の数がこの時点で非常に増加したため，関連度計算方式において属性の合致や強い関連を持つ属性が存在する可能性が増えたためと思われる．

3.8 おわりに

本章では概念ベースの構築方法について述べた．基本概念ベースの構築では，国語辞書の見出し語とその語義文を利用して概念及び属性の定義を行った．概念の意味定義を他の概念の集合，つまり属性群により行うことで,「概念の意味定義の意味定義」を考えることが出来る連鎖的構造の知識ベースを作成した．さらに基本概念ベースの属性に対してルールを用いた精練を行うことで，より精度の高い概念ベースの構築を行った．

しかし国語辞書のみを情報元として構築された概念ベースには，概念の不足や属性により定義される意味知識の少なさが問題となる．そこで新聞記事から概念と属性の関係になり得る語群を抽出することで，概念ベースの拡張を行った．新聞記事において共起する語同士には何かしらの関連があると考え，それを属性とすることで概念の意味定義をより広げることが出来た．また，関連度と概念ベース *idf* を用いた手法による重み付けを行った．この重み付けは概念ベースの構造のみを利用しているため，それまでの重み付けに用いていた人手による適切属性の評価データを必要としない．これにより適切属性の評価データに依存せず，概念ベースへの属性追加に汎用的に用いることが出来る重み付けが行われた．

さらに概念ベースの意味知識をより充実させるために，シソーラスを用いた属性の追加を行った．人手により作成されたシソーラスの知識を概念ベースに追加することで，概念の持つ属性をより多角化させることが出来た．

第4章　二次属性およびWebからの属性追加

4.1　はじめに

概念ベースに定義される概念をより人間らしい意味に近づけるためには，様々な関連を見出せる語を属性として持たせる必要がある．前章において述べたシソーラスからの属性追加のように，様々な情報元から適切な属性を付与することで，概念ベースの精度を高めることが出来ると考えられる．そこで本章では概念ベースの精度向上を目的とした新たな属性追加手法について述べる [23,24].

属性追加のための情報元として，本章では二次属性およびWebを用いた．二次属性は前章において様々に選別・精練された一次属性から得られるため雑音の少なさが予想できるが，得られる語群は一次属性の意味定義を行う語のみであるため，新たな分野・観点の属性追加は期待できない．また，一次属性の数によっては二次属性数が膨大なものになることが予想され，重要な語のみを選別する必要がある．一方Webからの属性候補取得は，Web上に存在する様々な文書情報から属性を取得するため新たな分野・観点の属性取得が期待できる．しかしWeb上の情報は雑多であり，二次属性に比べて雑音が存在する可能性は高いと考えられるため，こちらも適切な属性候補の選別が必要不可欠である．

そこで適切な属性候補を選別するために，概念ベースidf，重み，関連度の3種類の値を用いて属性を選別する手法について検討，評価を行い概念ベースの属性拡充を行う．なお，属性を追加する概念ベースは前章の最後で述べたシソーラス概念ベース [8] とする．

4.2　追加属性候補の取得

本節では概念へ追加する属性の候補の取得について述べる．追加属性候補は二次属性およびWebより取得する．

4.2.1　二次属性からの追加属性候補の取得

ある概念の意味定義は自身が所持する一次属性によって成されている．この一次属性は前章で示したような属性信頼度やルールによる選別などを経て精練された語群であるため，そこから得られる二次属性からも概念にとって適切な語を得られる可能性が高い．そこで概念から全ての二次属性を展開し，展開元の概念への追加属性候補とする．図4.1に概念「冬」に対して二次属性から新たな属性を取得する様子を示す.

概念「冬」の一次属性である「冬季」から，さらに属性を導く．これが概念「冬」の二次属

性である．他の一次属性からも同様に属性を導き，得られた二次属性群から概念「冬」と関連が強い語，例えば「冬季オリンピック」を新たな属性とすることが出来る．

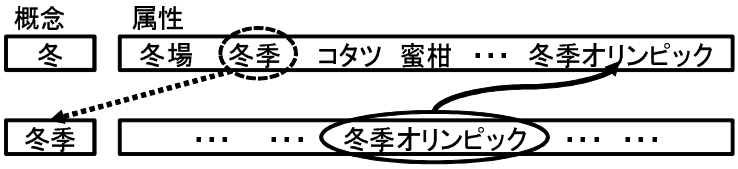


図 4.1: 概念「冬」の二次属性からの親属性取得

4.2.2　*Web* からの追加属性候補の取得

Web からの追加属性候補の取得には，www を用いた属性獲得手法を用いる [25]．これは www 検索エンジンを用いてある語の検索結果を取得し，そこから抽出した語群を用いて検索語の属性となり得る語を取得する手法である．これは概念ベースに存在しない語を擬似概念化する為に提案された手法だが，本章では概念ベースにすでに定義されている概念を検索語とし，得られた語句を追加属性候補とした．図 4.2 に本手法を用いて得られる属性の具体例を示す．

概念 ： **メニエール病**

属性 ： **めまい**, 症状, 耳鳴り, **難聴**, 内耳, 病気, **治療**, **耳**, **原因**, **吐き気**, **ストレス**, **純**, **灸**, **降板**, 発作, **診断**, **回転**, **加護**, **水腫**, **嘔吐**, **舞台**, **病名**, **悪化**, **リンパ**, **医学**, **リンパ液**, **検査**, **病院**, **突発**, **薬**

図 4.2: Web から得られる属性の具体例

「メニエール病」は現在の概念ベースに定義されている既存概念であり，黒字で示した属性が，現在の概念に属性として登録されていない語となっている．このように，*Web* 上の自立語から概念に関連のある新たな語を取得することができる．

4.3　追加属性候補の選別

属性追加を行うシソーラス概念ベースは，平均属性数が 37 個となっている．つまり平均二次属性数は 37 の 2 乗で 1369 個という膨大な量になる．もちろん，平均属性数よりも多くの属性を持つ概念も存在しており，二次属性から得られる属性候補の数は非常に多くなることが予想

される．

Web からの属性候補取得では，概念を検索にかけて得た結果を形態素解析にかけて自立語を取得している．そのため，ブログのように 1 ページに関連のない内容が複数含まれている場合，必要な語と雑音が属性候補として同時に取得されてしまう．

以上の点から，概念にとって重要な属性を多く追加するために選別を行う必要がある．そこで追加属性の候補に概念ベース *idf*，重み，関連度を用いて値付けを行い，それらに閾値を定めて選別を行う．

4.3.1 概念ベース *idf*

概念ベース *idf* とは，文書処理でよく用いられる $tf \cdot idf$ の考え方を概念ベースに適用したもので，概念ベース内での各概念の価値を算出する一つの方法である．概念ベース *idf* は，概念ベース全体を一つの文書空間として捉えることで算出し，値が大きいほど概念の意味定義に重要な語となる．概念ベース *idf* の具体的な算出式は 3 章 3.5.3 節に示した通りである．概念ベース *idf* の値に閾値を定め，一定以上ならば属性として重要であると判断し，追加を行う．

4.3.2 概念との関連度

追加する属性と，追加先の概念との関連度を利用する．属性とは概念の意味定義を行う語であるため，概念との関連性も高くなるのではないかと考えられる．そこで概念と属性候補の関連度を計算し，その値が一定以上ならば属性として追加を行う．関連度の計算には 2 章 2.3 節に示した関連度計算方式を用いる．

4.3.3 属性候補の重み

概念ベースの属性には，その重要性を意味する重みが付与される必要がある．そこで属性候補への重み付与を行った後，その重みに閾値を定めて一定以上ならば属性として追加する．重みは 3 章 3.5.5 節に示した関連度と概念ベース *idf* による算出を用いる．

4.4 精度評価

二次属性および *Web* から属性追加候補を獲得し，概念ベース *idf*，重み，関連度のそれぞれに閾値設定を行った上で選別した属性を概念ベースへ追加した．C 平均順序正解率による精度評価を以下に示す．まず，概念ベース *idf* に閾値を定めて属性選別を行った場合の精度評価を図 4.3，図 4.4 に示す．

二次属性からの属性追加では，概念ベース *idf* が 5.1 以上の場合のみを追加したときに，追加前に比べて 0.8%の精度向上が見られた．また，他の閾値でも追加前と比べて精度が向上している．*Web* からの属性追加では，概念ベース *idf* が 4.0 以上の時に 0.6%の精度向上となった．こ

ちらも精度は追加前と同じもしくは向上しているが，二次属性からの属性追加と比べて全体的に精度の向上率は低くなっている．

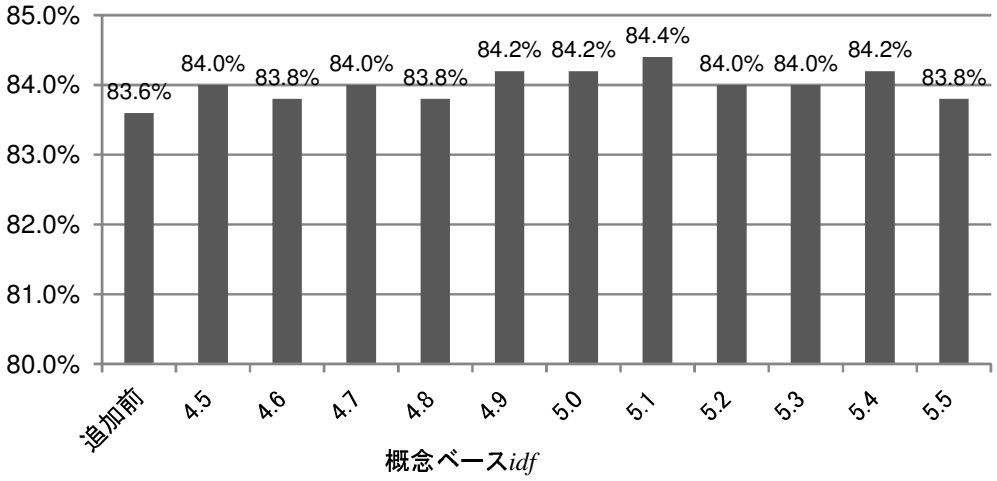


図 4.3: 二次属性からの追加精度 (閾値：概念ベース idf)

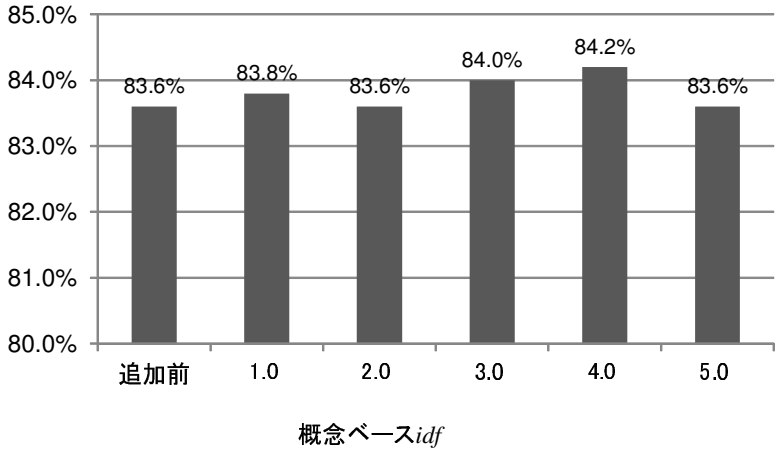


図 4.4: Web からの追加精度 (閾値：概念ベース idf)

次に概念と属性の関連度に閾値を定めて選別を行った場合の精度を図 4.5，図 4.6 に示す．

二次属性からの属性追加では，選別に関連度を用いた場合どの閾値でも精度が下がる結果となった．Web からの属性追加では，関連度が 0.004 もしくは 0.006 以上の場合に属性追加を行ったとき，精度が 0.4%向上した．また，他の閾値で追加を行った場合にも精度が同じもしくは向

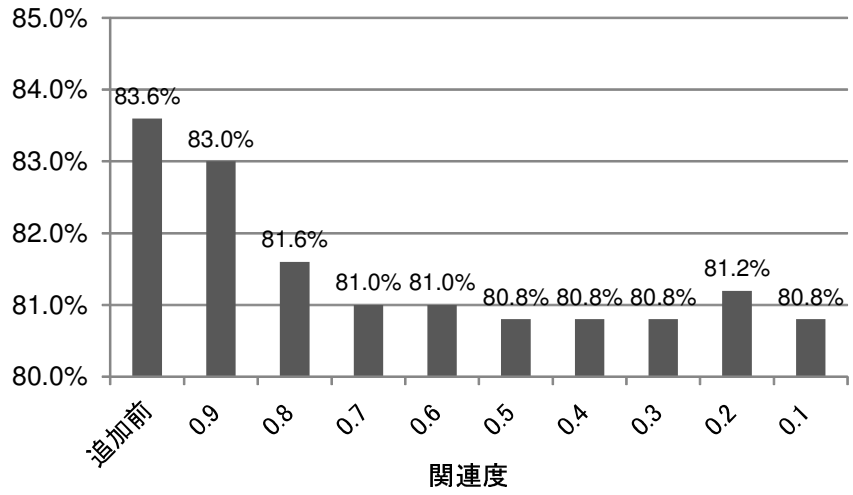


図 4.5: 二次属性からの追加精度 (閾値：関連度)

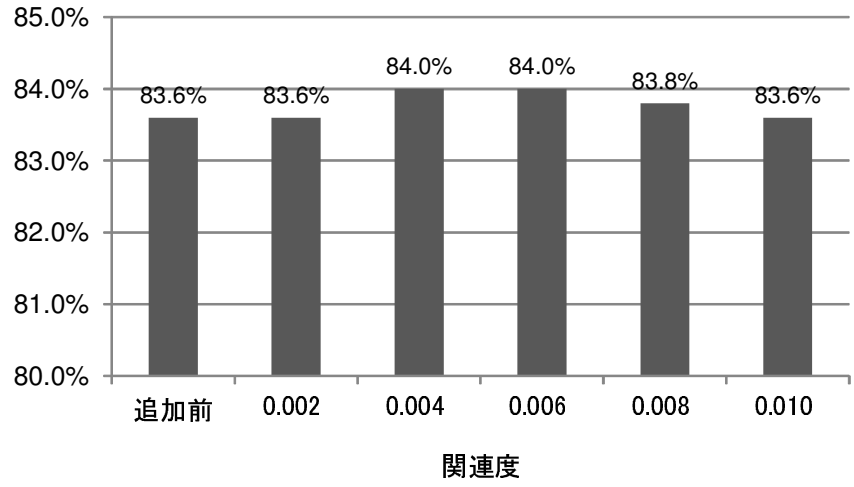


図 4.6: *Web* からの追加精度 (閾値：関連度)

上している．

最後に，重みに閾値を定めて属性選別を行った場合の精度を図 4.7，図 4.8 に示す．

二次属性からの属性追加では，選別に重みを用いた場合，関連度を用いた場合と同じようにどの閾値でも精度が下がった．*Web* からの属性追加では，重みが 0.005 以上の場合に属性追加を行ったとき，精度が 1.0%向上した．他の閾値でも，全体的に精度が向上している．

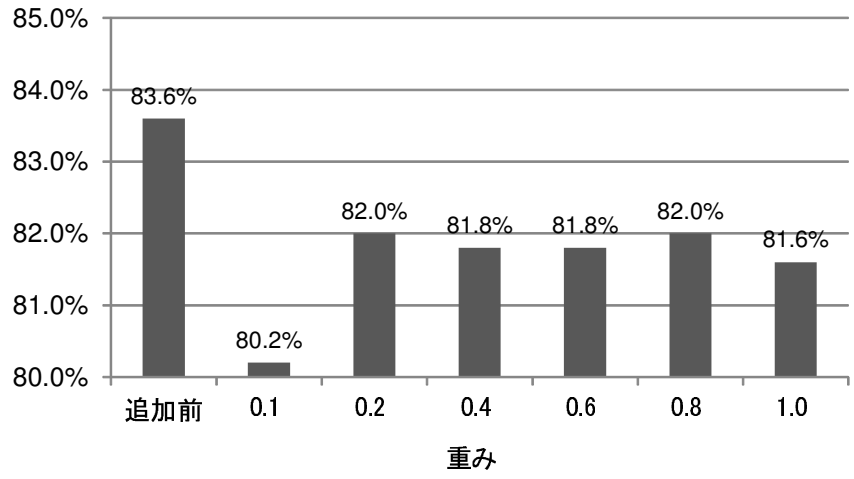


図 4.7: 二次属性からの追加精度 (閾値：重み)

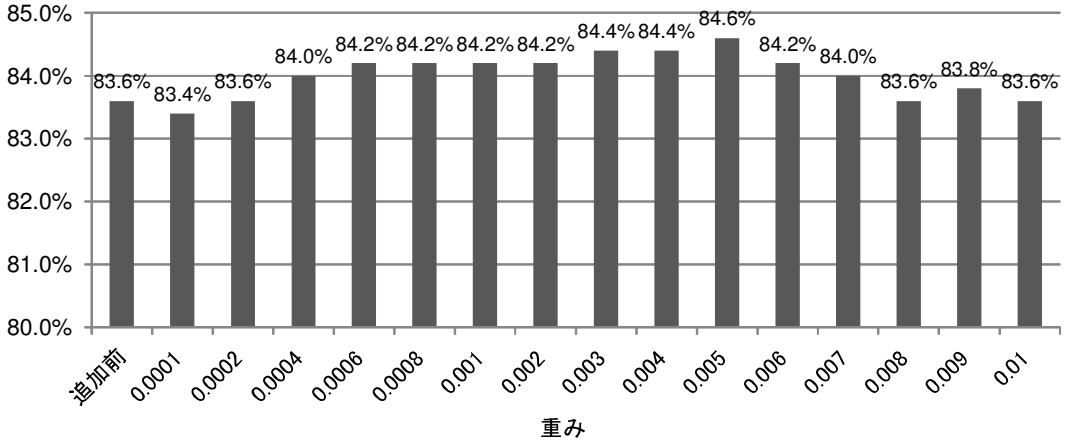


図 4.8: Web からの追加精度 (閾値：重み)

以上の結果から，二次属性からの属性追加と Web からの属性追加の両手法において精度の向上が見られることが分かった．よって二つの手法を統合した属性追加を行うことで，さらに精度向上を見込めるのではないかと考えられる．二次属性からの属性追加では概念ベース idf を閾値に，Web からの属性追加では重みを閾値にした場合がそれぞれの最高精度を出している．しかし単純に最高精度を出した閾値での結果を統合した場合が，統合後の最高精度になるとは限らない．そこで各手法で最高精度の閾値以外の組み合わせについても統合を行い，評価を行った．

Web からの属性追加の重み閾値を，最も精度の良かった 0.005 で固定した上で二次属性からの属性追加を概念ベース *idf* の閾値を変動させて行った結果が表 4.1，二次属性からの属性追加で最も精度の良かった概念ベース *idf* の閾値 5.1 を固定し，*Web* からの属性追加の重み閾値を変動させた結果が表 4.2 である．

表 4.1: 統合結果 (*Web* からの追加　重み閾値固定)

Web(閾値:重み)	二次属性 (閾値:概念ベース *idf*)	精度 (%)
0.005	4.5	84.2
	4.6	84.4
	4.7	84.6
	4.8	84.8
	4.9	84.4
	5.0	84.8
	5.1	84.8
	5.2	84.8
	5.3	84.4
	5.4	84.4
	5.5	84.6

表 4.2: 統合結果 (二次属性からの追加　*idf* 閾値固定)

二次属性 (閾値:概念ベース *idf*)	*Web*(閾値:重み)	精度 (%)
5.1	0.0002	85.2
	0.0004	85.6
	0.0006	85.4
	0.0008	85.4
	0.001	85.4
	0.002	85.4
	0.003	85.4
	0.004	85.0
	0.005	84.8
	0.006	84.8
	0.007	84.8

表 4.1 の結果より，まず二次属性からの属性追加では最高精度が 84.8%となり，追加前と比べて 1.2%の精度向上となった．また，二次属性からの属性追加のみと比べて 0.4%，Web からの

属性追加のみと比べて 0.2%の精度向上となった.

表 4.2 の結果より，Web からの属性追加での重み閾値を 0.0004 にした場合，精度が 85.6%となった．これは追加前と比べて 2.0%，二次属性からの属性追加のみと比べて 1.2%，Web からの属性追加のみと比べて 1.0%の精度向上となっている.

以上より，二次属性からの属性追加を概念ベース idf が 5.1 以上の場合に，Web からの属性追加を重みが 0.0004 以上の場合に行った結果，手法統合後の属性追加で精度が 85.6%となり，追加前と比べて 2.0%の精度向上となった.

4.5　構築された概念ベースの検証

ここまで，概念ベースに適切な属性を追加するための二つの手法と，それぞれについて属性選別を行った結果の精度を示した．本章では二次属性からの属性追加と Web からの属性追加の両手法について，どのような変化や特徴があるか以下の視点で検証を行った.

4.5.1　属性数の変化

概念が持つ属性数の変化と，手法を用いても属性が追加されない概念の数を調査した．表 4.3 に平均属性数と追加前との属性数の差を，表 4.4 に各手法において属性が一つも追加されなかった概念数を示す.

表 4.3: 属性追加による属性数変化

属性の情報源	平均属性数	追加前との差
二次属性	41.0	3.4
Web	43.5	5.9
統合	57.6	20.2

表 4.4: 属性が追加されない概念数

属性の情報源	属性が追加されない概念数
二次属性	17232
Web	4108
統合	634

二次属性からの属性追加では，表 4.4 に示すとおり属性が追加されない概念数が非常に多い．また，追加された属性数も Web からの属性追加と比べて少なくなっている．二次属性からの属性追加では，ある概念に追加する属性の候補を概念自身の一次属性から展開している．そのた

め展開した二次属性の中には既に概念の一次属性として存在している語も多く，結果として新たに追加される属性の数が少なくなっている．

4.5.2 追加される属性例

処理を行うことで実際に追加される属性の調査と，それぞれの手法によって得られる属性の特徴について検証と考察を行った．二次属性から得られる属性例を表 4.5 に，*Web* から得られる属性例を表 4.6 に示す．

表 4.5: 二次属性からの属性追加例

概念	追加された属性例
蝸牛	蛞蝓魚, 背負う
紫陽花	這松, 黐の木, 小梅, 芙蓉峰
髑髏	貝殻骨, 前膊骨, 親骨, 橈骨, 骨相学
眼鏡	鬚鬟, 老視, 光学ガラス
エスカルゴ	焼蛤, 舌平目
パセリ	体菜, ヴァニラ, 菊菜

表 4.6: *Web* からの属性追加例

概念	追加された属性例
蝸牛	季語, 外耳, 牛, 初音, 劇団, 宿, 聴覚, 葉
紫陽花	園芸, 秋色, 花言葉, 壁紙, 浴衣, イラスト, 雨, 素材, 季節, 観光
髑髏	吸血鬼, 骸, 吸血, 眼帯, パーツ, 水晶, 紋様, スカル, クローム, 復讐, 梵字, カード, ピンチ
眼鏡	コンタクトレンズ, 補聴器, レンズ, コンタクト, お洒落, セカンド, ブランド, 度, 本舗, 創業
エスカルゴ	芹, ピクルス, ドレッシング, 定食, ビタミン, 菜園, ミネラル, パスタ, 調理, 乾燥, 自家製, 栽培, 料理, 王子, 栄養, ドラマ, 工房, レシピ
パセリ	フレンチ, ガーリック, 日産, テープ, バター, クリーナー, 牧場, 食堂, レシピ, 団地, 大量, 歌詞, 中古

二次属性からの属性追加手法では概念や一次属性を包括する語や，細分化した具体的な語句が取得される傾向があった．例えば概念「髑髏」には，具体的な他の骨の種類が属性として追加されている．概念「エスカルゴ」に追加された属性「焼蛤」は，3 章 3.6.3 節においてシソーラスから追加される貝類の一次属性を細分化した語である．二次属性とは，概念がもつ一次属性それぞれの意味定義をしている語群である．N 次の属性を辿っていくことは，語の意味をよ

り詳しく展開していくことになる．そのため概念をより具体的に表した語や上位に位置する語などが属性として取得されやすくなっている．一方 Web からの属性追加では属性候補を Web 上から取得しているため，現在の概念が持つ属性と繋がりのない，新たな関連性の語を獲得できる可能性が高くなっている．概念「紫陽花」の属性として得られた「浴衣」は柄のデザインとして関連性が存在している．パセリの属性として得られた「バター」などは，パセリ自身の意味やパセリの上位関係にある野菜と直接関連は無いが，食べ合わせや調理方法においては非常に関連性の高い語である．

4.5.3　概念ベース *idf* と重みの傾向

二次属性からの属性追加では概念ベース *idf* を，Web からの属性追加では重みを閾値として属性追加を行った．この二つの値が，各手法でどのような特徴を持っているか検証を行った．そこで各手法で追加された属性の概念ベース *idf* と重みを，概念ごとに平均をとってその分布を調査した．まず図 4.9 に両手法で得た属性候補の概念ベース *idf* の分布を, 図 4.10 に概念ベース *idf* が 2.0 以上の部分を拡大した分布を示す.

二次属性から得られる属性候補に比べ，Web から得られる属性候補は概念ベース *idf* の値が小さいことがわかる．また二次属性から得られた属性候補は，概念ベース *idf* の値が広い範囲に分布している．そもそも概念ベース *idf* とは，概念ベース内での属性としての使用頻度から算出された値である．二次属性から得られる属性候補群は，概念ベース内からほぼ均等に抽出されるために概念ベース *idf* の値も広く分布している．しかし Web から得られた属性候補は Web 上で用いられる一般的な語群であるため，概念ベース内での語の重要性に特化した概念ベース *idf* では値の分布が偏ってしまうものと考えられる．つまり Web からの属性追加で得られた属性候補を選別する値として，概念ベース *idf* は不向きだといえる．総じて値が小さく分布も偏っているため，閾値がある一定より小さくなると一気に属性が追加されてしまう．

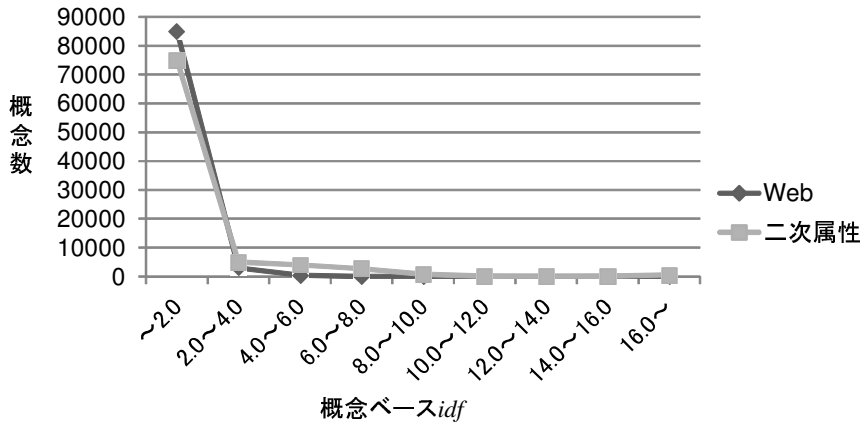


図 4.9: 概念ベース *idf* 毎の概念分布 (全体)

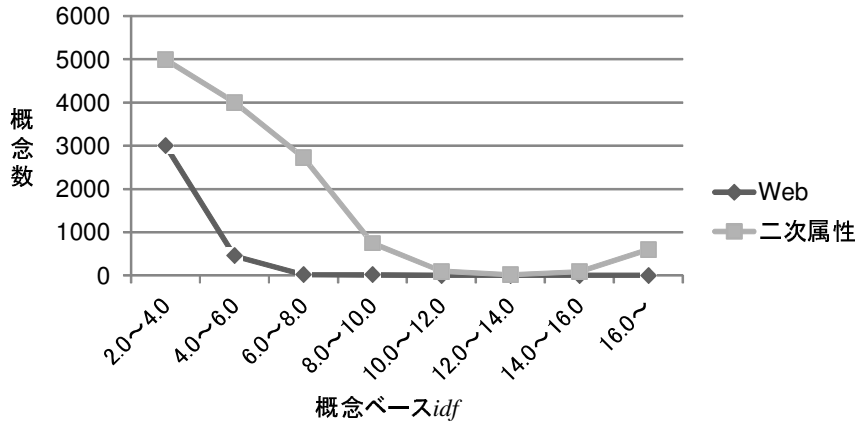


図 4.10: 概念ベース *idf* 毎の概念分布 (*idf* 値 2.0 以上)

次に図 4.11 および図 4.12 に両手法で得た属性候補の重みの分布を示す.

二次属性から得られた属性候補は，Web から得られた属性候補に比べて重みが大きな値に渡って分布している．これは，二次属性から得られた属性候補と，追加先の概念との関連度が大きな値を得やすいためである．関連度は概念同士の二次属性までを用いて算出しているため，もともと概念の二次属性に存在していた属性候補との関連度は大きく出やすい．そのため，二次属性から得られた属性候補の中で適切な属性と不要な属性両方に区別なく大きな重みが付与されやすくなっている．よって関連度や重みを二次属性から得た属性候補の選別に用いた場合，属性として必要な語だけを区別して得ることが難しい．一方 Web からの属性追加で得られる属性候補は，元の概念が持つ属性に関係のない語を多く得ることができる．そのため不当に大きな関連度や重みが付くことが少なく，必要な属性を区別して選別できていると考えられる．

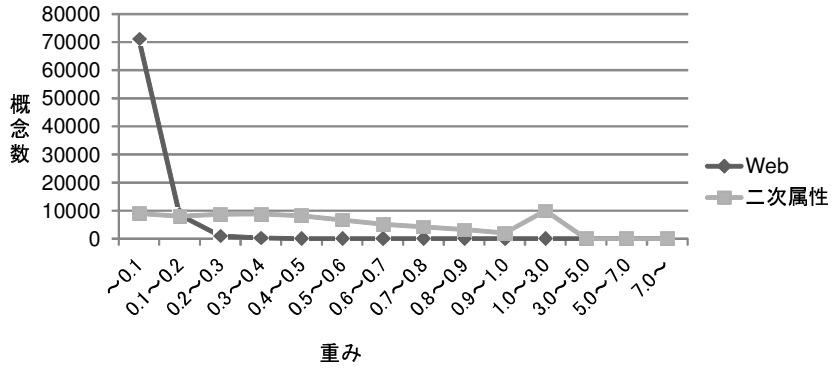


図 4.11: 重み毎の概念分布 (全体)

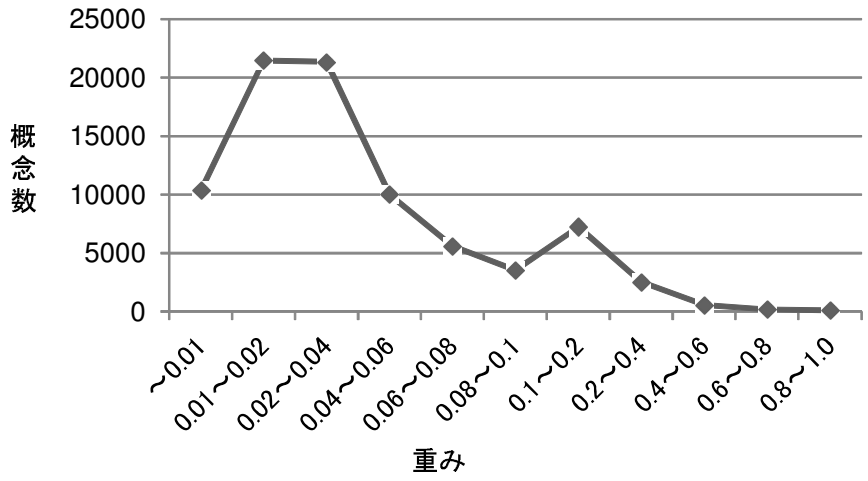


図 4.12: 重み毎の概念分布 (*Web* からの追加分　重み 0.01 以上)

4.6　おわりに

本稿では概念ベースに定義されている概念に対して新たな属性を追加する手法として，二次属性からの属性追加と *Web* からの属性追加の二つを提案した．提案手法の結果として，概念がもつ属性数の平均が約 37 個から約 58 個に増加し，概念ベース全体での属性数を増やすことに成功した．また，概念ベースの精度は 83.6%から 85.6%となり，属性の追加前と比べて 2.0%の精度向上を得られた．このことより，属性を追加することで概念が持つ意味を広げ，概念ベースの精錬が行われたことを示した．

二次属性からの属性追加では，概念と属性候補の間に高い関連度が算出されるため，属性選別の閾値として用い難いことが分かった．また，*Web* からの属性追加で得られた属性候補は，概念ベース *idf* の値が総じて低く，分布も偏っているために選別手法としては適さないと分かった．

提案手法はどのような媒体から作られた概念ベースに対しても，新たな属性を自動的に獲得することができる．それぞれの特徴にそった属性選別方法を示したことで，今後の概念ベースへの属性追加に汎用的に使用できると考えられる．

第5章　複数語概念連想システム

5.1　はじめに

本章では語概念連想システムの拡張として，複数語概念連想システムについて述べる．このシステムは語概念連想システムにおけるある語から連想できる他の語を想起する処理を複数の語に対応させたシステムである．

連想ゲームについて考える．このゲームはテーマとなるある 1 語を提示された被験者 A が，テーマを知らない被験者 B に対して「テーマを連想させるテーマ以外の語」を順次述べていく．述べられた語句が少ないうちに B がテーマを想起すれば多くの得点を得る，といったルールが一般的である．この連想ゲームを成り立たせているのは，ゲーム名の通り連想能力である．2 章において述べた語概念連想システムでは，概念ベースによって概念化された種々の語から，人間の連想に基づいた他の語を想起することができる．本章ではこれを拡張して，連想ゲームのように複数の語から連想される他の語を想起するシステムを提案する．

5.2　複数語概念連想システム

本章で提案するシステムの具体的な処理の流れを図 5.1 に示す．

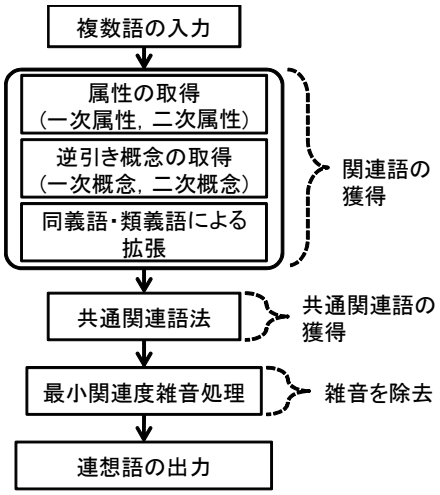


図 5.1: 複数語生成手法の流れ

複数語概念連想システムは，入力される複数の語から連想される別の語を生成する．提案システムでは概念ベースを利用して，入力される語それぞれについて何かしらの関係性を見出せる「関連語」を取得する．この関連語群から複数の入力語全てと関連を見出せる語（連想語）を選択する．また，関連度計算方式によって入力語と関連語との間の関連性を定量化し，正しい連想語を選択する．

入力は概念ベースにおいて概念化されている語，語数は 2 語以上とした．まず入力された各語と関連のある語（関連語）の獲得を行う．関連語は「属性の取得」「逆引き概念の取得」「同義語・類義語による拡張」の 3 つの処理により獲得する．入力語ごとの関連語を獲得した上で「共通関連語法」によりすべての入力語に共通して存在する関連語を共通関連語として獲得する．最後に共通関連語から雑音を除去するために「最小関連度雑音処理」を行い，連想語を出力する．

5.3　関連語の獲得

関連語とは入力された複数の語それぞれについて，人間なら何かしらの関連があると考える語を指す．例えば「針」という語にとって「釣り，縫い物，注射，刺す...」といった語は関連語であると言える．

関連語の獲得手法として，概念ベースを利用した「属性の取得」と「逆引き概念の取得」および語の同義・類義関係を示す関係語辞書を利用した「同義語・類義語による拡張」の 3 つを提案する．

5.3.1　属性の取得

入力された複数の語それぞれを概念として見たとき，その属性は入力語と何かしらの関連を持った語群となるはずである．そこで入力語の属性を取得し，それらを関連語と見なすこととした．

取得する属性の範囲について，まず一次属性は概念を意味定義する直近の語であるため，関連性は強いと考えられる．さらに一次属性の語を概念と見たとき，それらの意味定義を行う二次属性も元の入力語と関連する可能性がある．よって属性の取得においては，入力単語の一次属性のみを関連語とする場合と一次属性および二次属性を関連語とする場合の 2 つのパターンについて処理を行った．具体的な取得例として入力語「夏」からの属性の取得を図 5.2 に示す．

一次属性のみを関連語とする場合には「夏場，夏休み，海...」といった語が入力語「夏」の関連語となる．さらに二次属性を関連語とする場合は，例えば一次属性「夏場」を概念と見たときの属性「夏季，暑さ，太陽...」といった語群が関連語として加えられる．

二次属性を関連語とする場合，一次属性を多く持つ概念が入力語として与えられると関連語が膨大になるという問題がある．そこで二次属性に関しては取得個数の上限を 100 語とし，一次属性の重み上位 10 語それぞれから二次属性を 10 語取得することとした．100 語の打ち切り例を図 5.3 に示す．

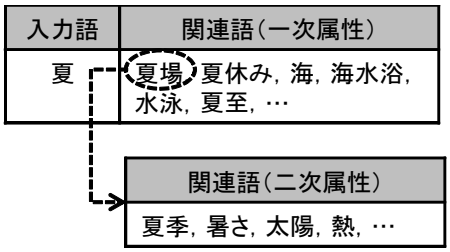


図 5.2: 属性の取得

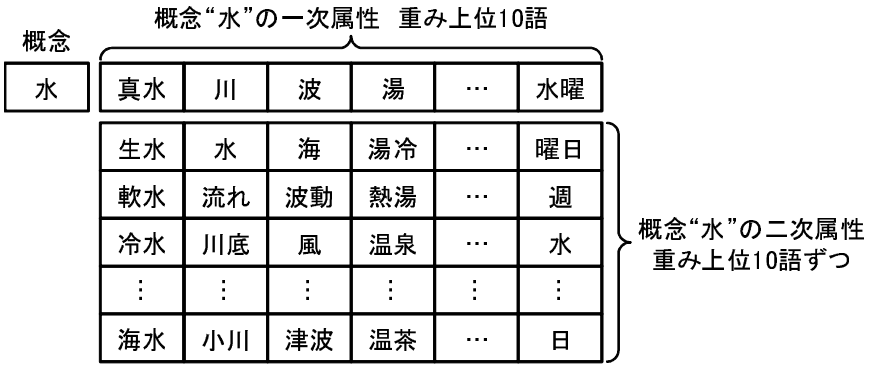


図 5.3: 二次属性の 100 語打ち切り

5.3.2 逆引き概念の取得

概念ベースは属性の連鎖構造により定義されている．この構造を利用して，概念からその属性を取得し，さらにその属性の属性を取得するといった順方向への属性展開とは逆に，属性から概念を展開することが出来る．つまり，ある概念 X について，X を属性として持つ概念 Y を取得することが出来る．このとき概念 Y を X の逆引き概念と呼ぶ．

属性として X を持つということは，この逆引き概念 Y の意味定義に X が用いられているということである．よって概念 X にとって，自身を属性として持つ逆引き概念 Y は関連性の強い語であると考えられる．そこで，入力語の逆引き概念を関連語として取得することとした．具体例として入力語「夏」からの逆引き概念の取得を図 5.4 に示す．

「夏」の逆引き概念，つまり「夏」を属性として持つ概念を関連語として取得する．ここでは「スイカ」や「競泳」といった概念の属性に「夏」が存在しているため，概念「夏」の逆引き概念としてこれらの語が得られ，関連語となる．

なお逆引き概念の取得においても，一次逆引き概念までの展開および二次逆引き概念までの展開の 2 パターンについて処理を行った．二次逆引き概念とは「入力語を属性として持つ概念（＝一次逆引き概念）を属性として持つ概念」という事になる．この二次逆引き概念に関しても

取得個数の上限を 100 語，一次逆引き概念の重み上位 10 語それぞれから二次逆引き概念を 10 語取得することとした．

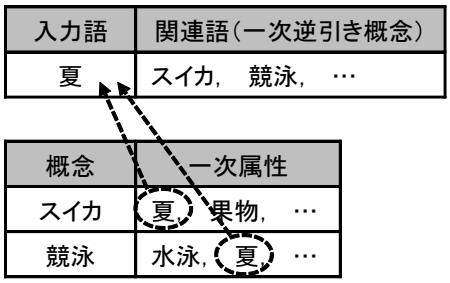


図 5.4: 逆引き概念の取得

5.3.3　同義語・類義語による拡張

入力語をその同義語・類義語により拡張した上で，その一次属性および一次逆引き概念を関連語として取得する．具体例として入力語「運動」の同義語・類義語による拡張を図 5.5 に示す．

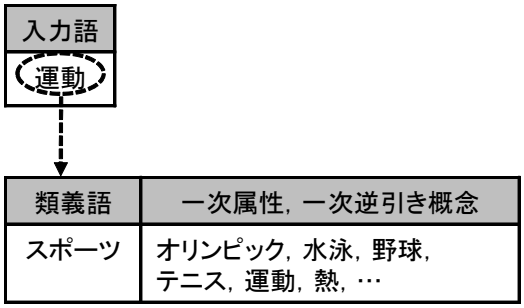


図 5.5: 同義語・類義語からの関連語取得

同義語・類義語は国語辞書の語義文に明記された見出し語の同義・類義語を抽出して用いる．図 5.6 に国語辞書における見出し語「運動」の語義文を示す．

語義文中には見出し語の類義語，同義語，反意語を表す記号が明記されている．このうち，辞書中に明記された全ての同義・類義関係にある語を抽出し，拡張に用いた．辞書全体から，同義語関係の語が 643 ペア，類義語関係の語が 31268 ペア取得された．

うんどう
【運動】
《名・自サ》☆１★めぐり動くこと。動き。「彼の数百年来依然として運動なき有様なりし人ゝが、⇔〈田口卯吉・日本開化小史〉」☆２★〔理〕物体が時間の経過に従って、その空間的位置を変える現象。「運動の法則」☆対★⊆☆１★☆２★⊇静止。☆３★〔健康の増進・維持のために〕体力を使って体を動かすこと。「海水浴は追って実行する事にして、運動丈は取り敢ずやる事に取り極（き）めた⇔〈夏目漱石・吾輩は猫である〉」「遊歩（うんどう）に便宜（びんぎ）なる場所とも見えねば、⇔〈坪内逍遥・当世書生気質〉」☆参★ひゆ的に、頭をはたらかすことにもいう。「頭の運動」☆類★スポーツ。☆４★ある目的を達するために人に呼びかけたり働きかけたりすること。「学生運動」「またオフラアナの長官からロプウヒンを威嚇（いかく）するように運動したのだ⇔〈大仏次郎・地霊〉」☆類★奔走。尽力。

図 5.6: 見出し語「運動」の語義文

5.4 共通関連語法

前節で述べた各手法により得た，入力語ごとの関連語を比較し，共通する関連語（共通関連語）を取得する．各入力語が共通して持つ関連語は，入力語の全てと関連を持つ語と判断できる．そこで共通関連語法により得た共通関連語を最終的な出力候補とする．具体例として「夏，水，運動」の 3 語における共通関連語法の処理を図 5.7 に示す.

入力語	関連語
夏	夏場, 夏休み, 海, 海水浴, **水泳**, 夏至, **熱**, **競泳**, …
水	真水, 生水, 川, 海, 波, 湯, **熱**, **水泳**, 遊泳, **競泳**, …
運動	競技, **競泳**, 練習, 波, 力学, 体育, 動的, …

類義語	一次属性, 一次逆引き概念
スポーツ	オリンピック, **水泳**, 野球, テニス, 運動, **熱**, …

入力語：夏　水　運動
共通関連語：水泳　競泳　熱

図 5.7: 共通関連語法

入力語「夏，水，運動」のそれぞれが図 5.7 に示したような関連語を保持しており，そのうち共通関連語は下線太字で示した「水泳，競泳，熱」となる．

5.5　最小関連度雑音処理

複数語概念連想システムにおける最終的な出力は，入力された各語全てから連想される語である．前節までの手法によって得られた共通関連語は入力全てと何かしらの関係がある語だが，全ての語と強い関連を持っている保証は無い．そこで最小関連度雑音処理では各入力語と共通関連語との関連度を利用して，共通関連語における雑音の除去を行う．

ある入力語 A, B および共通関連語 C があったとき，A と C の関連度，B と C の関連度をそれぞれ算出し，小さい方の値を共通関連語 C の最小関連度と定義する．この最小関連度に閾値を設定し，閾値以下の最小関連度となる共通関連語を雑音として除去する．なお最小関連度の閾値は実験的に求めた 0.05 を用いた．

具体例として「夏，水，運動」という 3 語の入力語から得られた共通関連語「水泳」と「熱」についての最小関連度雑音処理を図 5.8 に示す．

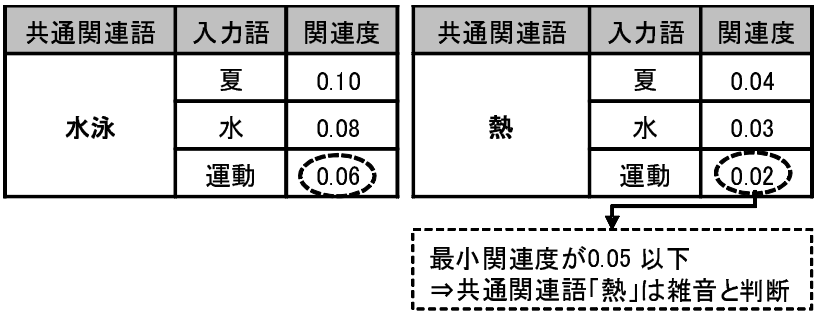


図 5.8: 最小関連度雑音処理

入力語「夏，水，運動」のそれぞれと共通関連語「水泳」と「熱」の関連度を算出する．まず「水泳」については入力語「運動」との関連度 0.06 が最小関連度となり，これは閾値 0.05 より大きいので雑音と見なさず，除去されない．「熱」についても同じく関連度を算出すると，入力語「運動」との関連度 0.02 が最小関連度となる．これは閾値より小さい値であるため，入力語「運動」と共通関連語「熱」の間の関連が希薄であると判断できる．よって共通関連語「熱」は雑音であると判断し，最終的な出力に含まない．以上の処理によって共通関連語の雑音除去を行い，最終的な出力である連想語を得る．

5.6　複数語概念連想システムの評価

評価はアンケートによって作成した入力語と連想語の組み合わせ 100 セットを用いて行った．このテストセットは被験者 2 名に対して連想ゲームを実施することで作成した．

まず，テーマとなる語を 1 語，被験者 1 名（被験者 A）にのみ指定する．テーマを提示された被験者 A は，テーマを知らない被験者 1 名（被験者 B）に対して「テーマを連想する語」を一語ずつ提示していく．被験者 B がテーマにたどり着いた時点で連想ゲームは終了する．この

とき，被験者 A が B に対して提示した語群を，複数語概念連想システムへの入力語と見なして提案手法より連想語を生成する．

次にテストセット作成とは別の被験者 3 名に対して入力語および連想語を提示し，手法が出力する語が人間の連想に即しているか否かを評価した．精度は 3 名中 2 名以上が正解とした語の割合として算出した．また，擬似的な再現率として連想語中にテーマとして提示した語が出力されている割合を算出した．再現率を擬似的としたのは，本来は入力から人間が思いつく可能性のある全ての語句をテストセットとして保持しない限り，完全な再現率の算出が出来ないためである．

関連語の取得方法の違いにより 2 種類の連想語生成手法を作成し，それぞれについて評価を行った．関連語の取得方法のパターンは次に示す通りである．

A：**一次属性および一次逆引き概念**
B：**A＋二次属性および二次逆引き概念**

なお，同義語・類義語による拡張は双方のパターンに適用する．図 5.9 に評価結果を示す．

結果として A のパターンでは精度が，B のパターンでは擬似再現率がそれぞれ優位な結果となっていることが分かる．精度は A が 5.3%，再現率は B が 17.0%高い結果となっており，優位性は B の方が高い．また F 値を算出すると A が 0.630，B が 0.681 となり，結果として二次属性および二次逆引き概念を用いた手法が良い評価となる．

提案手法で出力された連想語の例を表 5.1 に示す．

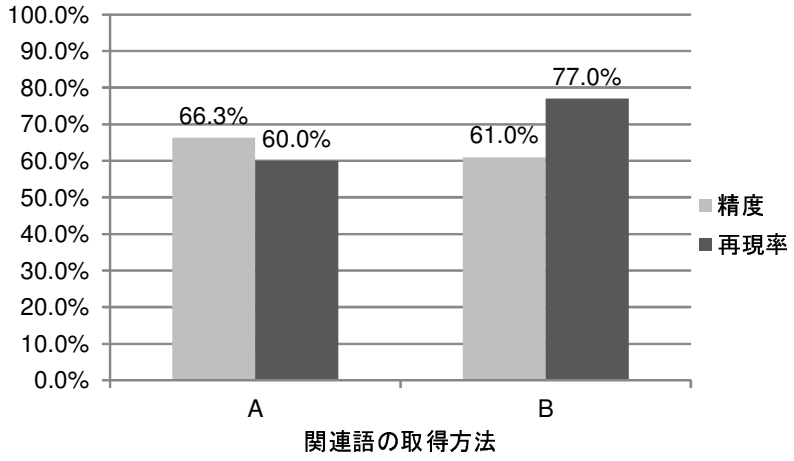


図 5.9: 評価結果

表 5.1 において，連想語の太文字は連想ゲームにおけるテーマを示す．また×は連想語が出力されなかったことを示し，括弧内はその際のテーマである．

表 5.1: 出力された連想語の例

入力語	連想語
海，遊び	**海水浴**，泳ぐ，涼み，水遊び
火力，風力，水力	動力，出力，入力，電力，原動機，**発電**，動輪，起動力，リフト，　機帆船，起電力，油圧
甘い，菓子，洋	洋菓子，**ケーキ**，生菓子，駄菓子，カステラ，茶菓，酒，水飴
学校，通学，ランドセル	**小学生**，児童，小学校，学生，学ぶ
魚，米，酢	食品，散らし鮨，**鮨**，食料，握り鮨，酒
図，道，紙	×（地図）
火事，車	×（消防車）

「魚，米，酢」という語から，それぞれ単独では連想され辛い「鮨」を出力することが出来ている．また,「甘い，菓子，洋」の例のように，単独では意味を定義し辛い形容詞を含む入力からも，テーマ「ケーキ」が得られている．ただしこの例では駄菓子や酒といった誤った語も出力されている．これは入力語それぞれを同等の価値で，単独に扱っている点に問題が有ると考えられる．人間ならば「洋」という語を「菓子」に付随させることで「洋菓子」という語を得，さらにそれを用いて連想を行うと考えられる．

5.7　おわりに

本章では概念ベースと関連度計算方式による語概念連想システムの拡張として，複数語概念システムの提案を行った．概念ベースに定義される語の知識を活用することで，入力される単語から関連のある別の語を想起することが可能となった．そしてその中から入力単語全てと関連する語を取得し，関連度計算による雑音処理を行うことで，複数の語から連想される語の提示を行った．提案手法では精度 61.0%，擬似再現率 77.0%という結果で人間が複数の語から自然と連想する語の生成が行われ，語概念連想システムによる人間らしい連想機構の一端を示した．

第6章　Webニュース記事本文を利用した見出し文の意味具体化手法

6.1　はじめに

前章までで述べた語概念連想システムを自然言語を対象とした情報処理技術に応用する事例として，本章ではロボットとの知的会話を視野に入れた新聞記事見出し文の意味具体化手法について述べる．

近年，人間と円滑なコミュニケーションが行える知的ロボットの実現に向けて様々な研究が行われている．人間同士のコミュニケーションはその多くが会話によってなされており，将来的にはロボットに対しても人間のような会話能力が求められると考える．人間が行う会話の種類は様々であり，あいさつのような慣用的な表現だけでなく，質問や返答，提案，更には何かしらの話題についての雑談など様々である．そのうちの１つとして，新聞やテレビから得られる時事情報を話題とした会話について考える．

人間の会話中に時事情報が話題として出現することは珍しくない．今朝起きた鉄道事故を会話の話題に用いることもあれば，政治家の不祥事を話題に論を交えることもあるだろう．これらの時事情報はテレビやラジオ，新聞といった媒体を通じて提供される．特に新聞は近年ネット上での公開が一般的であり，多くの時事情報が手軽に閲覧できるようになっている．この新聞によって与えられる時事情報を話題として提供することで，ロボットから時事情報を用いた発話を行うことが出来るのではないかと考える．

新聞を利用した発話をロボットに行わせる最も簡単な方法は，新聞中の文を発話テンプレートに埋め込むといった手法である．例えば記事から抜き出した一文や，記事を端的に表した見出し文を発話テンプレートに埋め込むことで，時事情報を話題とした発話を形成することができる．ここで，本章では新聞記事の見出し文に着目する．見出し文は一つの新聞記事の内容を端的に要約した一文であり，これをロボット会話の為のリソースとして用いることで，時事情報を話題とした発話文を生成できるのではないかと考える．

しかし新聞記事の見出し文は，その端的さゆえに具体的な情報が往々にして省略されている．例えば“ オリンパス元社長，社長職復帰断念 ”という見出し文からは，オリンパス元社長とはいったい誰なのか，いつ断念することを決めたのかといった情報が省略されている．また，見出し文には体言止め，助詞や動詞の欠落といった表現が多用されており，そのままでは新聞記事内容を示す自然な一文にはならない．また文の途中の空白や三点リーダによる見出し文内容の変化や，コロン記号によるテーマの提示といった見出し文特有の書式も多く表れる．

そこで見出し文を特有の書式から自然な表現に変換し，さらに記事本文を用いて意味の具体化を行う手法を提案する．意味の具体化については“ いつ ”,“ だれが ”,“ どこで ”という具

体的な情報に当たる語句を本文中から抜き出し，追加や置換を行う．

6.2 提案手法の流れ

提案手法では Web ニュースサイトから得られる見出し文に対して語句の追加や置換を行い，意味の具体化を図る．また，体言止めや助詞，動詞の欠落といった表現や見出し文特有の書式に関しては，解析と変換のルールを定めることで処理を行う．

図 6.1 に提案手法の流れを示す.

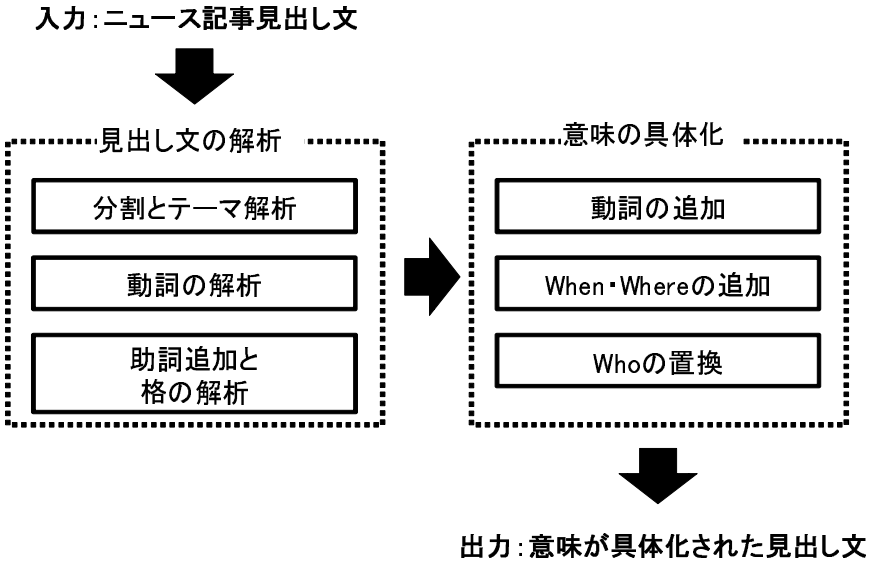


図 6.1: 提案手法の流れ

入力は各新聞社（朝日・毎日・読売）の Web ニュースサイトから取得したニュース見出し文とする．まず見出し文に対して，ルールに従った解析を行う．解析では見出し文特有の書式，体言止め，助詞の欠落に対して処理を行った上で，見出し文の各語句を Who，What，When，Where の格に分類する．

次に，解析結果から見出し文で欠けている情報を判断し，その見出し文が示すニュース記事本文中の語句から情報の取得・補完を行うことで意味の具体化を行う．具体化の処理として，まず見出し文の解析結果で動詞，When，Where のいずれかにおいて語句が分類されなかった場合に，それらの分類に当たる語句を記事本文から追加する．さらに，Who に分類される語句が見出し文中に存在した場合は記事本文からより具体的な情報を取得した上で Who の置換を行う．

6.3 見出し文の解析

見出し文で欠けている情報が何であるかを調べるため，見出し文の解析を行う．まず見出し文特有の書式に対して分割およびテーマの解析による処理を行う．その上で見出し文を日本語係り受け解析器「南瓜」[26] により語句に分け,「動詞の解析」および「助詞追加と格の解析」を行う．

6.3.1 分割とテーマ解析

「オリンパス前社長、関与認める　東京地検が任意聴取」のような見出し文では文の前半と後半で話の内容が変化している．このような見出し文に関しては前後で分割を行い，それぞれについて具体化処理を行う．各新聞社（朝日・毎日・読売）の Web ニュースサイトから取得したニュース見出し文において用いられる分割の表現を調査し，結果として全角空白もしくは三点リーダが存在する場合を分割の条件とした．また,「インフルエンザ：患者数、都市部を中心に倍増」のように，記号「：」を用いて記事内容を示す主要な語句を示す記述も見出し文特有の書式として存在している．この記号「：」の前に存在する語句をテーマと定義し，記号「：」以降の文のみを見出し文として扱う．ただし，後述する意味の具体化が行われた後に，テーマおよび「に関して」という語句を先頭に追加する．例えば「１２年度予算案：”はやぶさ２”に３０億円」の場合はテーマが「１２年度予算案」となる.「”はやぶさ２”に３０億円」の部分に対して意味の具体化が行われた後,「１２年度予算案に関して」という語句が先頭に付与される．

6.3.2 動詞の解析

見出し文において動詞がどの語句に当たるかを解析する．品詞解析には形態素解析器「茶筅」を用い，動詞もしくはサ変名詞と判断された語句を見出し文の動詞にあたる部分と判断する．ただし，サ変名詞と判断された語句については語尾に「する」を付与することで体言止め表現を無くす処理を行う．また,「養成へ」「停止か」のようにサ変名詞の後ろに助詞「へ，か」が続く表現は「養成するかもしれない」のように変換を行った上で動詞のあたる部分と判断する．例えば「パレスチナ民兵１人死亡」という見出し文があった場合には，サ変名詞である「死亡」に「する」を接続して「死亡する」をこの見出し文の動詞にあたる語句と判断する．

6.3.3 助詞追加と格の解析

見出し文の格（Who，What，When，Where）の情報を分類する．まず，前節で述べた処理で判明した動詞とその直前の語の間に助詞が存在しない場合はその補完を行う．具体的には Web から自動構築された大規模格フレームシステム [27] を用いて動詞と直前の語を繋ぐ助詞を検索し，最も頻度の高い助詞を選択する．

図 6.2 に助詞追加の例を示す.

例えば「パレスチナ民兵１人死亡」という見出し文は前節の処理から動詞が「死亡する」で

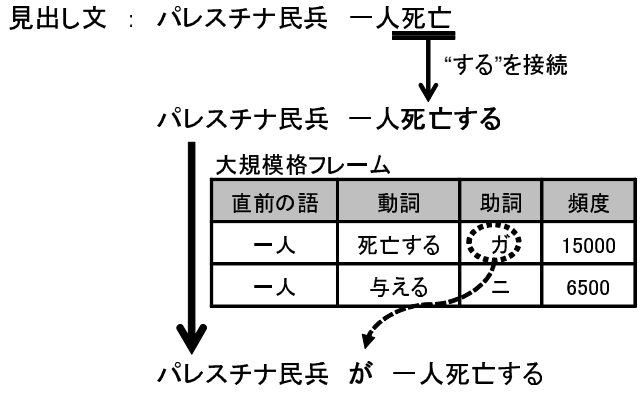


図 6.2: 助詞追加の例

あると分かる．大規模格フレームにおいて，直前の語である「１人」と「死亡する」を繋ぐ頻度が最も高い助詞は「が」である．よって「パレスチナ民兵１人（が）死亡（する)」という助詞の追加を行う．

助詞追加を行った上で，表 6.1 に示す分類規則に基づいて格の情報の解析を行う．なお，表中で用いている上位ノードの検索はシソーラス [1] により行った．

表 6.1: 分類規則

条件	分類
語句中に助詞「が」が含まれる	Who
語句中に助詞「は」が含まれる	Who
語句中に読点「、」が含まれる	Who
語句中に助詞「を」が含まれる	What
語句中に助詞「に」が含まれる	Whom
語句末尾の上位ノードに”場所”が存在	Where
語句末尾の上位ノードに”時間”が存在	When

ここで表 6.1 に示した規則のうち,「文節に読点「、」が含まれる」という条件で Who 格への分類が行われた際には，自然な文へ変換するために読点を助詞「が」へ置換する事とした．

「パレスチナ民兵１人死亡」という見出し文は,「パレスチナ民兵/１人死亡」という語句に分けられる．補完処理により「１人（が)」という語句に助詞「が」が含まれるため，これが Who であると分類される.「パレスチナ民兵」のようにどの分類にも含まれない語句が存在した場合には，南瓜によりその語句の係り受け関係を調査した上で，係り先の語句へ接続する．この例では「パレスチナ民兵」は「１人死亡」に係っており，これらを接続して「パレスチナ民兵１

人（が)」の部分をまとめて Who と判断する.

6.4 意味の具体化

特有の書式に対する解析および Who，What，When，Where の分類を行った見出し文に対して,「動詞の追加」「When・Where の追加」「Who の置換」の三つの処理を行い，見出し文の意味を具体化する.

6.4.1 動詞の追加

6.3.2 節に示した動詞の解析において，動詞にあたる語句が存在しなかった場合には動詞の追加を行う．見出し文が示す記事本文中から動詞を全て取得し，その中から見出し文に適した動詞となる語句を選択する．見出し文末尾の語と記事本文中の動詞全てとの組み合わせで大規模格フレームを検索し，双方を繋ぐ助詞が存在し，かつ頻度が最も高い組合せを調査する．例えば「浦和東・菊池が３発」という見出し文の場合，末尾の語「発」と記事本文中の動詞全てとの組合せで大規模格フレームを検索する．結果，最も頻度の高い組合せは助詞「を」によって動詞「決める」と接続される組合せであったため,「浦和東・菊池が３発を決める」のように動詞を追加する.

6.4.2 When・Where の追加

6.3.3 節において，When もしくは Where に分類される語句が存在しなかった場合には When・Where の追加を行う．記事本文中の語句の，シソーラス上における上位ノードを調べ，その中にノード「時間」が存在する場合には When，ノード「地名」もしくはノード「場所」が存在する場合には Where に追加する．複数の語が条件に該当する場合には，記事本文中に最も早く出現した語を追加対象とする．また，When の追加に関しては”…月…日”のような具体的な日時の記述が記事中に存在する場合には，それらを優先して追加する．具体的には数字の後ろに「月」もしくは「日」がある語句を表記一致により検索し，存在する場合にはその語句を When と判断する.

6.4.3 Who の置換

6.3.3 節において Who に分類される語句が得られた場合に，記事本文中の語句を用いて置換を行うことで見出し文における主体の意味具体化を図る．記事本文中の全語句を置換の候補とし，そこから最も Who の置換に適切な語句の選択を行う.

まず，動詞の解析もしくは動詞の追加により得られた見出し文の動詞と，Who に分類された語句および置換候補語句それぞれとの，Web 検索における共起ヒット件数を取得する．このとき「置換候補語句と動詞」の共起ヒット件数が「Who に分類された語句と動詞」の共起ヒット

件数と比べてあまりにも小さい場合にはその候補語は置換に適さないと判断する．具体的には「置換候補語句と動詞」の共起ヒット件数が「Who に分類された語句と動詞」の共起ヒット件数の一割に満たない場合は置換候補語句から除外した．

次に，置換候補語句と Who に分類された語句との関連度を算出する．関連度の算出にはそれぞれの概念が持つ属性と重みが必要となるが，新聞記事中の語句の多くは固有名詞や複数の語句の集合などであり，概念ベースに定義されていない未定義語である場合が多い．そこで，未定義語に対して Web 上から属性を取得する手法を利用してこれらの語句の概念化を行う．置換候補語句と Who に分類された語句との関連度が高いほど，それらの語句の間の関連性が高く置換に適していると判断できる．また，この関連度に下限の閾値を定めることで関連が無い語句による置換が行われないようにする．なお，関連度の下限値は実験により 0.1 と設定した．図 6.3 に Who の置換処理の具体例を示す．

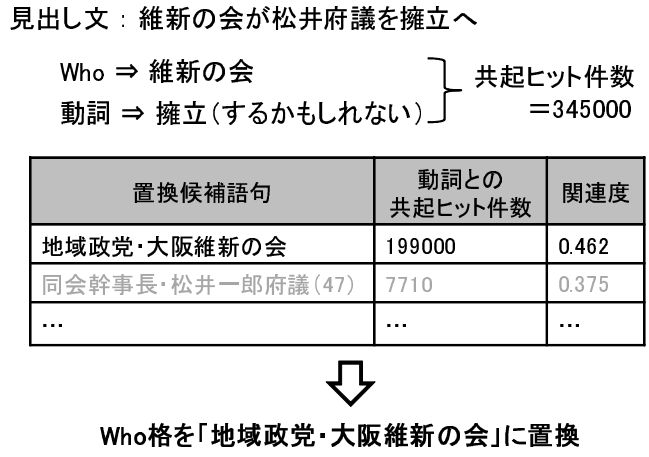


図 6.3: Who の置換処理の具体例

「維新の会が松井府議を擁立へ」という見出し文から，Who に分類される語として「維新の会」，動詞として「擁立（するかもしれない）」という語句が得られる．これらの語句の Web 検索における共起ヒット件数は 345000 件となった．次に記事本文中から得られた置換候補語句である「地域政党・大阪維新の会」と「同幹事長・松井一郎府議」それぞれの語句と動詞との共起ヒット件数を調査すると，後者の置換候補語句のヒット件数が 345000 件の一割に満たないため，候補から外れる．残った置換候補語句との関連度は 0.462 となっており，これは下限値 0.1 より大きい．よってこの置換候補語句と Who に分類された語句の間に十分な関連があると見なして置換を行う．結果として Who に分類される語句を「地域政党・大阪維新の会」に置換する．

6.5 評価

評価は毎日，朝日，読売の Web ニュースサイトから取得した各 40 文，計 120 文の見出し文を用いて行う．この見出し文 120 文と提案手法により意味の具体化を行った出力文のセットを被験者 3 名に提示し,「見出し文の解析」「動詞の追加」「When・Where の追加」「Who の置換」それぞれの処理が正しく行われ，意味が具体化された出力を得られたかの判断を行った．なお，被験者への提示の際には提案システムのどの処理が行われたかもあわせて知らせている．2 名以上が出力文を妥当とした場合に正解とした．また，不正解となった出力に関しては「見出し文の解析」「動詞の追加」「When・Where の追加」「Who の置換」の三つの処理のうち，どの部分に不備があったかをあわせて調査した．

図 6.4 に評価結果を示す.

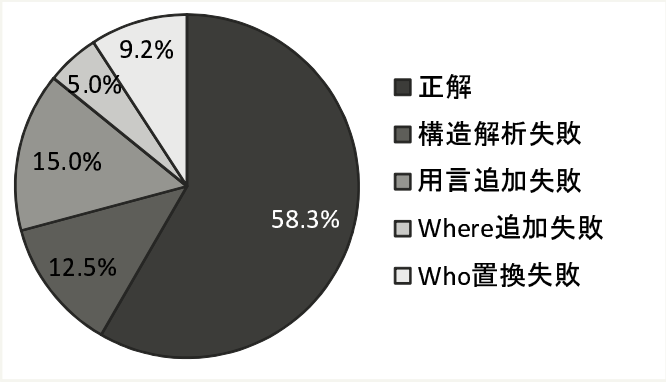


図 6.4: 評価結果

結果として評価文全体の 58.3%の見出し文について，意味の具体化を行うことができた．表 6.2 に意味具体化の例を示す.

中島氏が離党届提出　八ツ場ダム建設再開に抗議

見出し文「那覇西が初戦敗退」の例では,「那覇西が何なのか」「何において敗退したのか」「どこの出来事なのか」といった種々の情報が欠落しており，見出し文そのものから得られる情報は非常に少ない．しかし提案システムにより意味の具体化が行われた結果，那覇西が高校名であること，サッカーの試合において敗退したこと，試合は埼玉で行われたことが 1 文で理解できることが分かる．

「民主・斎藤恭紀議員、離党表明…追随の動きも」の例では，三点リーダによって分割された前半部分に関して,「２７日午前」という When の追加や,「離党を表明する」という助詞の追加およびサ変名詞の処理による適切な変換が行われている．しかし後半部分に関しては,「追随の動きも挙げる」というサ変名詞の処理において，時制の違和感が生じている．追加される動詞の選択は問題なく行えているが，文全体の時制にあわせた末尾処理が必要であることがわかる．

表 6.2: 意味具体化結果

見出し文	意味具体化後
オリンパス前社長、関与認める 東京地検が任意聴取	菊川剛・前社長（７０）が２１日の 捜索前関与を認める． 東京地検特捜部が任意聴取する．
民主・斎藤恭紀議員、離党表明… 追随の動きも	民主党の斎藤恭紀衆院議員が ２７日午前離党を表明する． 追随の動きも挙げる．
那覇西が初戦敗退	那覇西が 第９０回全国高校サッカー選手権大会第２日 埼玉県の埼玉スタジアムで初戦敗退する．
中島氏が離党届提出	民主党の中島政希衆院議員が２４日 党本部に離党届を提出する．

6.6　おわりに

本章では語概念連想システムの応用事例として，ロボットとの知的会話を視野に入れた新聞記事見出し文の意味具体化手法について述べた．新聞の見出し文は記事の内容を端的に要約した一文であるが，その端的さゆえに具体的な情報が省略されており，また見出し文特有の書式も多く存在する．そこで見出し文の解析と変換のルールを定めることで処理を行い，見出し文に対して語句の追加や置換を行うことで意味の具体化を行う手法を提案した．意味の具体化のための処理を提案し，「Who の置換」においては語概念連想システムにおける関連度計算方式を用いることで，見出し文中の語句の置換にふさわしい語句を記事本文から取得することができた．結果として 120 文の見出し文のうち，58.3%の見出し文で意味の具体化を行うことができた．

第7章　新聞記事中の難解語を平易な表現へ変換する手法

7.1　はじめに

本章では，一般的な自然言語処理分野において活発に議論される処理に語概念連想システムを活用する事例を示す．具体的には「言い換え」や「変換」といわれる処理分野に着目し，新聞記事中に出現する難解な語を人間の会話に出現するレベルの平易な表現へ変換する手法を提案する．

新聞を会話リソースとすることで，時事情報を話題とした発話文を生成できる可能性については前章において述べた．前章においては見出し文に着目した処理を提案したが，本章では新聞記事の活用について考える．新聞記事を会話リソースとして見たときに問題になるのが新聞記事表現の難解さである．新聞のように公に対して公開される文章は短い文で端的に内容を表すため，馴染みの薄い難解な言葉，俗にいう「堅い」言葉を多く使う．これらの言葉は文章として読むには違和感はないが，会話に用いるには自然ではないことが多い．例えば「貸与する」という言葉は会話では「貸す」という言い方をするほうが自然である．また，一般的にはそう難解ではない言葉，例えば「落下した」という言葉も会話ということを考えると「落ちた」のような更に易しい表現の方が馴染みやすいと感じる．つまり会話に用いられる言葉と新聞といった公的な文中に用いられる言葉の間には，同じ意味を表すにしても難易度や馴染みの深さに違いがある．ロボットの発話にもこのような語の違いに配慮しなければ，人間はロボットとの会話に違和感を覚えてしまう．そこで本章では新聞記事を人とロボットの会話に利用することを想定して，記事中の難解語を馴染みのある別の平易な表現に変換する手法を提案する．

語の難解さ，平易さの判断には [28] で報告されている単語親密度を用いる．これは語の「馴染み深さ」を定量化した数値であり，新聞記事に用いられる語と一般的な会話に用いられる語の間にある単語親密度の差を調査することで新聞記事中の難解語を自動的に判断，平易な表現への変換を可能とする．また，変換処理を行う上で重要な意味の保持に関しては，人間の連想能力を模倣した語概念連想システムを用いることでそれを実現する．語と語，文と文の間の意味関係を柔軟に表現することを目指した語概念連想システムの機構を利用することで，変換前の記事が持つ意味を考慮した変換を行うことができる．さらに，変換後の記事をより人間にとって違和感の無いものとするために，人間が自然に行う語の変換に則った処理を提案する．人間は語の変換を行う際に 1 語を別の 1 語でこともあれば文章を用いて変換することもあるという考えのもと，語をそれと同じもしくは近い意味の別の 1 語に変換する 1 : 1 の変換処理（1 語変換）および語を文によって表現する 1 : N の変換処理（N 語変換）の双方を組み合わせた変換を行うことで人間が自然だと感じる語の変換を目指す．

7.2　関連研究と提案手法の特徴

文中の語を他の表現に変換に関する研究は数多くなされており，平易な表現への変換技術そのものとしての研究 [29] や Web 検索への利用を目的とした複数パターンの変換の生成 [30]，利用者の言語能力に配慮した平易化 [31–33]，会話への利用 [34] といった形で報告が成されている．これらの研究においても，語の表現を変換するためのアプローチとして 1 : 1 の変換処理および語を文によって表現する 1 : N の変換処理が挙げられている．

1 : 1 の変換については，例えば [31] では児童向け新聞の記事と一般の新聞記事との間でベクトル空間モデルによるマッチングを取り，同一内容の記事の対から 1 語対 1 語の変換対を作成している．また [30] では Web を用いて入力された文字列中の語の変換候補を生成している．変換対象となる語（名詞，形容詞，動詞，カタカナ語）を入力から取り除いた文字列を用いて Web 検索を行うことで，変換対象の語があった場所に入る他の語を取得することが出来る．対して [29] の報告では，国語辞典の定義文を変換に用いる 1 : N の変換処理が報告されている．定義文を変換に適した形の文に整形するルールを策定し，日本語として違和感の無い変換を行うことを目指している．

これらの変換処理はそれぞれ，1 : 1 の変換処理および 1 : N の変換処理を単独で行っているが，本章で提案する手法はこの双方を組み合わせることでより人間の思考に沿った変換処理を提案できると考える．人間がある語の変換を行う際には，まず別の 1 語に言い換えることができないかを考える．これは変換の対象となる語の同義語や類義語によって行うことが可能である．しかし同義語や類義語を持たない語も数多く存在することを考えると，この 1 : 1 の変換では不十分である．また私たちの行う会話では，1 つの語の変換に文を用いる場合も多々考えられる．これは分かりにくい語が出現した場合にその語の「意味を説明する」ことで語の変換を行っている．例えば「明言」という語ならば，同じ意味を持つ一語を探すよりも「はっきりと言い切る」という文による変換が自然である．1 つの語に対して文，つまり N 個の語による変換という機能が無ければ，人間の会話に近い自然な変換はできない．

[32] や [33] では，本章と同じく 1 : 1 の変換と 1 : N の変換の組み合わせについて述べられている．例えば [32] では対象となる文章を自治体の Web ページに固定し，人手による変換対の作成によって語の変換を実現している．変換対はシソーラスや国語辞典の定義文を人の目で参照して作成しており，よってある語を変換するための対は 1 語である場合もあれば短い文の場合もある． [33] では文化遺産に関する説明文を平易化することを目的として，そのための変換パターンの解析を行っている．この中では専門用語に対して文章による変換で補足を行うパターンや，外来語を同じ意味の日本語へ変換するといった手法により説明文を平易化できると報告している．これらの手法では変換対や変換のためのパターンが人手により作成されるため高精度を期待できるが，それに伴う労力も非常に大きい．また，変換対象を固定しているため作成した変換対やパターンの汎用性に欠けると考えられる．本章の提案手法では変換のための語や文を既存の辞書資源から自動的に選択するため，労力や汎用性の点で優位性があると考える．

国外でも語の変換に着目した評価型ワークショップ [35,36] が開催され，[37–39] といった研究が報告されている．[35] では文章中の英単語 1 語を別の語で変換するというタスクが設定されており，例えば [37] では変換のための語を得るために $N-gram$，語の出現頻度，Web ヒット

件数，さらには変換前の文章を他言語に翻訳した後，再度英語に翻訳するなどの様々な手法を組み合わせることで語にポイント付けを行い，変換を実現している．[36] では [35] で示された 1 語の変換に際してより平易な語を選択するというタスクになっている．例えば [38] では [37] のポイント付けを基礎とし，さらに [40,41] で定義された心理言語学的モデル，例えば語の具体性やイメージアビリティといった側面でスコア付けを行ったデータを用いて平易性の判断を行っている．[39] では語の平易性の判断材料として様々なコーパス内における出現頻度や語の長さを用いている．これらのタスクにおいても変換の処理は $1:1$ のものが大半であり，英文による変換は行われていない．また，[36] のタスクでは人手で用意された変換の候補となる語に対して平易性のランク付けをすることで変換を行っており，変換の候補となる語の選出は行っていない．候補の選出処理は [37] によって報告されているが，この手法は [35] における総括でも述べられている通り，変換に必要なリソースや処理過程が非常に複雑なものとなっている．前述したとおり，本章の提案手法では $1:1$ および $1:N$ の変換手法を組み合わせることで人間の自然な変換を実現する点，語概念連想システムを用いることで変換のための語や文を既存の辞書資源から自動的に選択できる点でこれらの研究と比べて優位性があると考える．

7.3 難解語の変換手法の概要

図 7.1 に提案する語変換処理の概要図を示す.

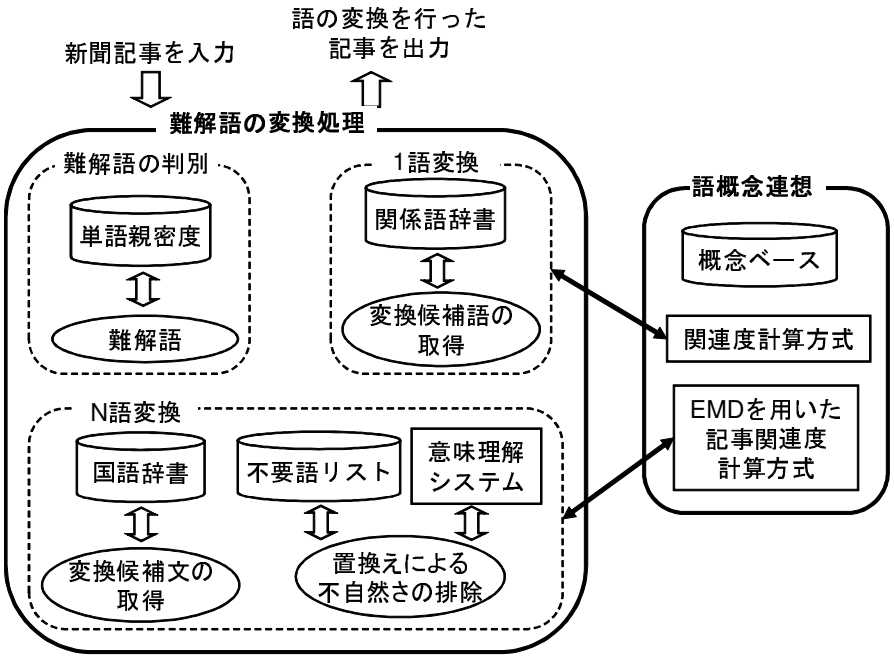


図 7.1: 語変換処理の概要図

本章で提案する難解語の変換手法は人間が自然に行う語の変換に沿い，1 : 1 および 1 : N の変換処理を組み合わせることで行う．入力は新聞記事とし，語の変換処理は句点を区切りとする記事中の 1 文ずつで行う．入力された記事中から会話に適さない馴染みの薄い語（難解語）を判別し，別の平易な語もしくは文に変換する．提案手法は同義語，類義語を用いた 1 : 1 の変換処理（1 語変換）と，1 つの語を文で変換する 1 : N の変換処理（N 語変換）によりこれを実現する．また，各変換処理において人間の連想能力を模倣した語概念連想システムを用いることで，語の表記に依存しない柔軟な語の変換を行う．語と語，文と文の意味的な近さを考慮した変換を行うことで，人間の常識に沿った語の選択や多義性の解消を図ることが出来る．

難解語の判別には単語親密度 [28] を用いる．単語親密度とは単語に対する馴染みの度合いを主観的に評価した値であり，数値が高いほどより馴染みのある単語であることを示す．これは 18 歳以上の被験者 40 名に対して単語を提示し，1 から 7 までの数字で馴染みがあるか否かを評価した結果を平均化することで算出される．表 7.1 に単語親密度の一部を示す．

表 7.1: 単語親密度の例

単語	単語親密度
あいさつ	6.59
心配	6.44
危ぶむ	4.72
サリドマイド	4.03

表 7.1 に示した通り，例えば「あいさつ」のようにごく一般的な語は単語親密度が高く，万人にとって馴染みの深い語であることがわかる．一方「サイドマイド」（睡眠薬の一種）は専門的な用語であり，一般的には馴染みが薄く単語親密度も低い値となる．日常的に使用する語とは，万人にとって馴染みのある語であると考える．新聞記事中に現れる「危ぶむ」という表現は，一般的な会話ならば「心配する」程度の表現の方が違和感なく馴染みやすい．つまり馴染みの度合いが高い語ほど会話への利用に適していると考えられる．また単語親密度が高ければ高い語ほど，その語を文字として提示された場合と音声として提示された場合の双方で語彙判断の反応時間が短く認知の誤りも少ないという結果が報告されており [42–44]，この事からも単語親密度が高い語ほど会話への利用に適した平易な語であるといえる．そこで本章では単語親密度が低い語を変換すべき難解語とみなし，平易な表現への変換処理を行う．

難解語と置き換える平易な表現は語の同義・類義関係を示した関係語辞書 [7,45] および国語辞書から得る．1 語変換においては国語辞典から自動構築された関係語辞書を用いて難解語の同義語・類義語を取得し，これらを変換に用いる語の候補（変換候補語）とする．辞書における同義語はある語と同じ意味を持つ別表記の語，類義語は類似の意味を持ち言い換えることの出来る語と定義されているため，これらの語を用いることで 1 語変換を行うことができる．N 語変換では国語辞書を用いて難解語の意味を説明する定義文を取得し，これらを変換に用いる文の候補（変換候補文）とする．

N 語変換に用いる変換候補文は辞書に記載されているそのままの形で語の変換を行うと，出力される文が日本語として不自然な場合がある．例えば辞書の定義文に出現する「転じて」や「~ の別名」といった言い回しは，そのままの形で変換に用いた場合に不自然さを発生させる要因となる．このような変換に必要の無い語を不要語と定義し，不要語リストを用いてこれらの削除を行う．また，元の記事中の語を定義文で変換した際に What や Who といった文中の情報が重複することによって不自然さが発生する場合がある．そこで意味理解システム [46] を用いて不自然さを排除した上で難解語を文に変換し，会話に適した語句で構成された文を出力する．

7.4 EMD を用いた記事関連度計算方式

提案手法では語の変換を行う際に，元の語と変換の候補となる語（変換候補語）との間の意味的な近さを考慮するために語概念連想システムを用いる．具体的には，まず 1 語変換においては複数得られる可能性のある同義語・類義語の中から最も元の語に近い意味を持つ変換候補語を選別するために関連度計算方式を用いる．次に N 語変換においては多義性の解消のために EMD を用いた記事関連度計算方式を用いる．これは難解語が多義語であった場合に辞書の定義文が複数取得されるため，文書間の関連性を定量化することで元の記事と最も関連の強い定義文を判別し，意味の特定を図るものである．

以下に EMD を用いた記事関連度計算方式について述べる．

EMD を用いた記事関連度計算方式は，ヒッチコック型輸送問題 [47](需要地の需要を満たすように供給地から輸送を行う際の最小輸送コストを解く問題) で計算される距離尺度である EMD を文書検索へ適用したもので，2 つの記事間の関連性を定量的に表現することが可能であり [48] によりその有用性が報告されている．

EMD とは 2 つの離散分布があるときに一方からもう一方の分布への変換を行う際の最小コストを指す．離散分布はそれを構成する要素と重みの対の集合で表現され，コスト算出の際には変換前の離散分布の要素が持つ重みを供給量，変換先の離散分布の要素が持つ重みを需要量と考え，要素間の距離を供給量，需要量にしたがって重みを運送すると考える．できるだけ短い距離で，かつ需要量に対して効率的に重みを運送する経路が EMD となる．これを文書検索に適応させる際には，文章中の自立語（名詞，動詞，形容詞）を要素として捉え，自立語の集合を離散分布と考える．ある文章の離散分布を違う文章の離散分布へ変換すると考えると，その際のコストが最小となる文章が元の文章に最も近い文章となり文書検索へ適用することが可能となる．EMD を用いた記事関連度計算方式について，以下の図 7.2 に示すような簡略図を用いて説明する．

ある文書 A と B があったとき，文書 A を文書 B に変換する際のコストを考える．それぞれの文書を文中の自立語 $Word_{Ai}, Word_{Bj}$ の離散分布と考える．まず自立語それぞれには重みの付与を行うが，本章では tf・idf の考え方を用いた．

語の網羅性である tf は，文書 A 中に出現する語 $Word_{Ai}$ の頻度 $tfreq(Word_{Ai}, A)$ を文書 A 中のすべての語数 $tnum(A)$ で割ったものを利用する．算出式は以下のようになる．

$$tf(Word_{Ai}, A) = \frac{tfreq(Word_{Ai}, A)}{tnum(A)} \tag{7.1}$$

次に語の特定性である idf については，3 章 3.5.3 節に示した概念ベース idf [49] を用いた．$CV_N(Word_{Ai})$ を N 次属性空間内における概念 $Word_{Ai}$ の概念ベース idf と定義すると，自立語 $Word_{Ai}$ へ付与する重み w は次のような式で定義される．

$$w = tf(Word_{Ai}, A) \times CV_3(Word_{Ai}) \tag{7.2}$$

つまりある自立語の重みは，自立語の網羅性 tf と自立語の概念ベース idf を掛け合わせることで与えられる．

このようにして文書 A, B 共に自立語への重みを付与する．ここでは例として図 7.2 のように重みが付与されたとする．

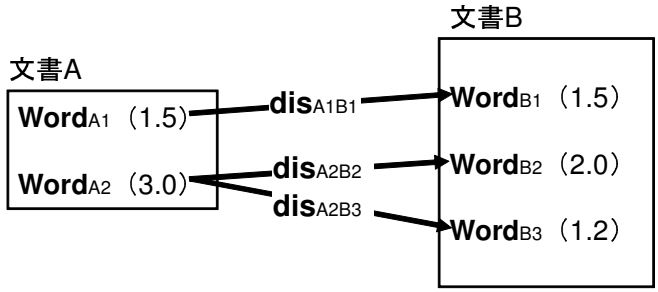


図 7.2: EMD による記事関連度計算方式

EMD では変換コストの算出を行う際に離散分布を構成する要素同士の距離を用いる．EMD を用いた記事分類方式ではこの距離を自立語同士の関連性であると考え，一致度によってこれを求める．$Word_{A1}$ と $Word_{B1}$ の距離 dis_{A1B1} は次の式で表される．

$$dis_{A1B1} = 1 - DoM(Word_{A1}, Word_{B1}) \tag{7.3}$$

一致度は関連性が高いと値が大きくなるため，1 から引いた値を距離としている．ここで $Word_{A1}$ と $Word_{B1}$ の間の変換コスト $cost_{A1B1}$ は次の式で算出される．

$$cost_{A1B1} = dis_{A1B1} \times 1.5 \tag{7.4}$$

これは $Word_{A1}$ と $Word_{B1}$ の距離に重みを掛けたものである．$Word_{A1}$ と $Word_{B1}$ が持つ重みは同じく 1.5 であるため供給量と需要量が合致し，$Word_{A1}$ からの重みの運送はこの時点で終了する．同様にコストの計算を行っていき，最終的にすべての運送経路のコストを足し合わせたものが EMD となる．図 7.2 の例では EMD は次のように表される．

$$EMD = cost_{A1B1} + cost_{A2B2} + cost_{A2B3} \tag{7.5}$$

$$cost_{A1B1} = dis_{A1B1} \times 1.5 \tag{7.6}$$

$$cost_{A2B2} = dis_{A2B2} \times 2.0 \tag{7.7}$$

$$cost_{A2B3} = dis_{A2B3} \times 1.0 \tag{7.8}$$

以上のような式で算出された EMD の値の最小値を最適化計算で求めて文書間の類似性を算出している.

7.5 語の変換処理の流れ

語の変換処理では入力された文から難解語を自動的に判別し，関係語データ [45], [7] による馴染みのある語への変換，もしくは国語辞書による文への変換を行う．具体的な処理の流れを図 7.3 に示す.

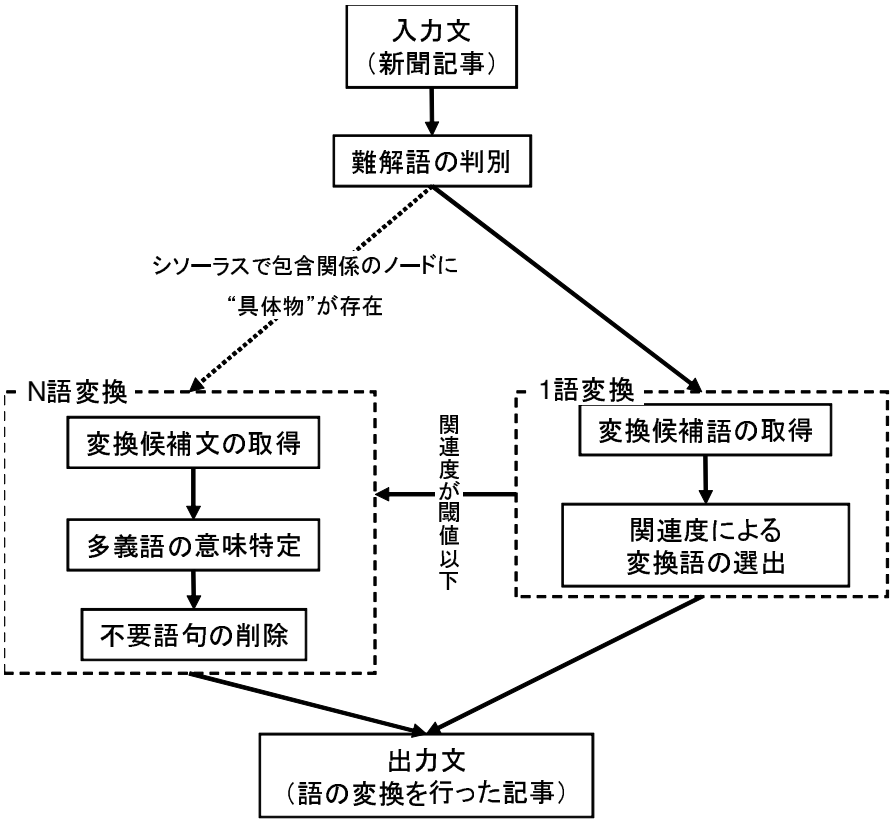


図 7.3: 語の変換処理の流れ

まず入力文を構成する単語の内，馴染みのない語を単語親密度の閾値により判別し，難解語とする．この難解語をシソーラス [1] 上で検索し，難解語を意味的に包含するノードの中にノード名「具体物」が存在する場合には N 語変換を，それ以外の場合には 1 語変換を先に行う．こ

れは具体的な物を示す単語は別の 1 語に変換することが困難であるため，シソーラスにより具体物と判断できる語に関しては N 語変換のみによって変換を行うためである．例えば「サリドマイド」のように具体的な薬品名を別の 1 語に変換することを考えると，物質を示す化学式や化合物名などが挙がる．それらは平易な表現とは言いがたく，そもそも難解な具体物の別称が平易であることは少ないと考えられる．この場合ならば「睡眠薬の一種」という文による変換を行えば自然でかつ平易な表現となる．

ノード「具体物」を上位に持たない語は，まず 1 語変換の処理を行う．ここでは語の同義，類義関係を示した関係語辞書から難解語の同義語および類義語を取得することで変換候補語を得る．これら変換候補語と難解語との関連度を算出し，最も高い関連度の候補語を用いて変換を行う．ただし，この際の関連度には下限値を設定し，最大関連度が閾値以下の場合には 1 語変換によって得られた候補語の信憑性が薄いと判断して N 語変換へ処理を移す．

N 語変換では国語辞書から変換候補文を取得して変換を行う．難解語が多義性を持つ場合には複数の変換候補文を取得することになるため，元の記事中で使われている意味をもつ変換候補文を記事関連度計算方式により判別する．また，難解語をそのまま変換候補文に変換した場合，辞書特有の言い回しや記事全体での情報の重複などにより元の文が不自然になる場合がある．そこで元の文と変換候補文との比較を行い，不要語句の削除を行うことで変換による不自然さを排除する．これらの処理を行った上で得られる文を用いて新聞記事中の 1 文を変換する．

7.6　難解語の判別

まず入力された新聞記事から，変換すべき馴染みの薄い語を判別する処理を行う．入力された新聞記事を句点（”。”もしくは”．”）を区切りとして 1 文ごとの記事文に分割して処理を行う．1 文に対して形態素解析を行い，各単語の単語親密度に閾値を定めることによって馴染みの有無を判断し，馴染みの無い単語を難解語とする．本章では会話のための資源として新聞記事を用いることを背景としているため，単語の馴染み深さの基準は「一般的な会話で使われる単語であるか否か」とする．この基準の作成には日本語話し言葉コーパス [50] を用いた．

7.6.1　閾値の決定

日本語話し言葉コーパスとは日本語による発話音声を大量に収集したデータベースである．収録されている発話音声中の語数は約 750 万語，時間は約 66 時間分となっている．発話音声には一般的な対話や学会講演といった様々なデータが収録されているが，このうち対話の音声を用いて「一般的な会話で使われる単語」の調査を行った．表 7.2 にデータの一部を示す．

単語親密度の閾値を決定するために，表 7.2 に示したような日本語話し言葉コーパスの対話データを構成する単語 2000 語と，新聞記事中の単語 2000 語とを無作為に抽出し，それぞれの単語親密度の平均と分布を調査した．その結果，新聞記事における単語親密度の平均が 5.74，標準偏差は 0.70，対話データにおける単語親密度は平均が 6.05，標準偏差が 0.66 となった．対話において用いられる語の単語親密度の平均の方が，新聞記事より高い値になっている．この

事から会話に利用するには新聞記事中の単語は馴染みが薄いことがわかる．

表 7.2: 日本語話し言葉コーパスの例

話し手	対話文
A	うちは妹が二人居て
B	ええ
A	で，五人家族なんですよ　女ばかりだからね　よく喋る
B	三人娘　姦しいみたいな
A	そうそう　お母さんも　よく　喋るしね
B	お父さん　大変ですね

新聞記事における単語親密度のデータ群（A とおく）と対話データにおける単語親密度のデータ群（B とおく）が，お互いにできるだけ他方の分布に属さないような値を閾値とすれば,「一般的な会話で使われる単語」を判別する閾値になると考えられる．そこで確率密度関数を用いて最適な閾値の調査を行った．確率密度関数は以下の式によって求める．

$$f_{(x)} = \frac{1}{\sqrt{(2\ \pi\ \sigma^{2})}} \mathrm{e}^{\frac{(\mathrm{x}-\mu)^{2}}{\sigma^{2}}} \tag{7.9}$$

ここで μ は単語親密度の平均，σ は標準偏差である．ある閾値があった時に，A に属するデータが閾値を越える確率および B に属するデータが閾値を越える確率を算出し，双方の和が最も小さい時の閾値を A と B を区切る最適な値とした．その結果，新聞記事に用いられる単語と一般的な会話で使われる単語の単語親密度による閾値は 5.82 となった．よって，入力された新聞記事中の単語の内，単語親密度が 5.82 以下の単語を難解語と判別し，語の変換処理を行うこととした．

7.6.2 閾値の評価

前節で決定した閾値が，人間と同じレベルで馴染み深い語と難解語を判別できるかの評価を行った．単語親密度が 5.82 よりも大きい，つまり馴染み深いと判断された 200 語と，単語親密度が 5.82 以下，難解語と判断された 200 語を新聞記事からランダムに取得し，それらを人間の目視で評価した．評価は被験者 3 名（男性 2 名，女性 1 名）で行い，それぞれの語が会話に出現する語としたときに難解と感じるか，平易と感じるかの判断を行った．なおこのとき，被験者には評価を行う合計 400 語が単語親密度の閾値以上であるか否かは知らせていない．多数決により 2 名以上が難解と感じた語は「人が難解と感じる語」，2 名以上が平易と判断した語を「人が平易と感じる語」とした．単語親密度の閾値によって馴染み深いと判断された 200 語については「人が難解と感じる語」であった場合に×,「人が平易と感じる語」であった場合に○と評

価する．単語親密度の閾値によって難解語と判断された 200 語については,「人が難解と感じる語」であった場合に○,「人が平易と感じる語」であった場合に×と評価する．表 7.3 に閾値の評価結果を示す.

表 7.3: 閾値の評価結果

	○	×
人が難解と感じる語	83.0%	17.0%
人が平易と感じる語	99.5%	0.5%

各評価者 2 名ずつの kappa 係数はそれぞれ 0.729，0.668，0.790 であった．結果として,「人が難解と感じる語」を 83.0%の精度で難解語であると判断できた．また,「人が平易と感じる語」に関しては 99.5%の精度で馴染み深い語，つまり変換の必要がない語であると判断することができた.

7.7　1 語変換

1 語変換では 1 つの単語をより平易な別の 1 つの単語に変換する．難解語の同義語・類義語を取得してこれらを変換候補語とし，その中から変換に最も適した語を選択する．本章における変換に適した語とは，変換前の語と比べて平易であり，かつ意味が同じ語である．平易であるかどうかの判断は単語親密度により行う．また，変換前と意味が同じ語を適切に選択するために関連度計算方式を用いた手法を提案する.

7.7.1　変換候補語の取得

変換候補語には難解語の同義語・類義語を用いる．これにより難解語と同じもしくは近い意味を持つ別の単語群を得ることができる．同義語・類義語の取得には関係語辞書を用いた．関係語辞書とは国語辞書に記載されている定義文から，見出し語の同義語，類義語といった関係語を自動的に抽出した辞書である．関係語の抽出手法に関しては [45] および [7] において示されている．定義される関係語の例を表 7.4 に示す.

表 7.4: 関係語辞書の例

単語	同義語	類義語
懸念	心配	不安
付近	近所，そば	周辺
協議	会議	相談

この辞書から得られる同義語，類義語を 1 語変換における変換候補語とする．表 7.4 に示したように，1 つの単語に対して複数の同義語・類義語が定義されている場合があるため，変換候補語は複数の単語群となる．

7.7.2 単語親密度と関連度による変換語の選出

変換候補語の語群から 1 語変換に適切な変換語を選出する．選出には変換候補語の単語親密度および，難解語と変換候補語との関連度を用いる．まず同義語・類義語として得られた語のうち，7.6 章で述べた閾値 5.82 以上の単語親密度を持つ語を選出する．これは単語親密度が高く馴染みが深いと判断される語であるほど，平易な変換に適すると考えられるためである．しかし単語親密度は馴染みの深さのみを表現する数値であり，語と語の意味の近さに関しては考慮されていない．変換を行う以上，難解語と最も意味の近い語が選出されるべきである．そこで語の意味を定量化する手法として，関連度計算方式を用いる．単語親密度が閾値以上である変換候補語の中から，元の難解語との関連度が最も高い語を選出することで「平易性がある語のうち，最も意味が近い語」を変換語とすることが出来る．具体的な変換候補語の選出方法について，記事の一部を用いて説明する (図 7.4).

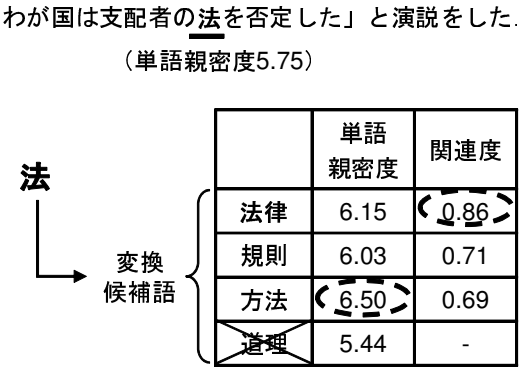


図 7.4: 変換語の選出

この文の中で「法」という語の単語親密度は 5.75 であり，これは 7.6 章で述べた閾値 5.82 を下回るため難解語となる.「法」の同義語・類義語から「法律」「規則」「方法」「道理」という 4 つの変換候補語が得られる．これら変換候補語から，最も適切な変換語を選択する．

まずそれぞれの単語親密度を見ると,「道理」は単語親密度が 5.44 となり閾値 5.82 に達していないため変換語から外れる．ここで各変換候補語と元の難解語「法」との関連度を算出し，最も関連度が高い語を変換語として選出する．この例では単語親密度が最も高い「方法」ではなく，関連度の最も高い「法律」が変換語として選ばれることになる．

7.7.3　1 語変換から N 語変換へ移行する条件

難解語と変換候補語との関連度に閾値を定め，閾値を越える変換候補語が存在しない場合には N 語変換を行う．これは関連度が低いということは難解語と変換候補語との関連性が薄く，変換には不適切であると判断できるためである．

関連度の閾値設定は概念ベースの評価方法である $X-ABC$ 評価 [12] を参考にして行う．具体的な評価方法に関しては 3 章 3.2 節に示した通りである．

このテストセットは被験者実験によって作成されており，つまり人間の感覚に合致した評価セットになっている．人間の自然な感覚を反映しているこの評価セットにおいて同義，類義関係と判断された $X-A$ 間の関連度は，本提案手法における難解語と変換候補語との関連性の有無を判断する閾値に値すると考えた．

評価セットは 500 組存在するため，$X-A$ 間の関連度も 500 個の値が算出される．そこから人間が同義，類義と感じる語同士の関連度を意味する値として平均値を算出した．これは [12] において用いられている評価式の中で，まったく関連のない $X-C$ 間の関連度の平均値を「関連がない語の間で算出される関連度」として用いる考え方に倣い，同義，類義関係にある $X-A$ 間の関連度の平均値を「同義，類義関係の語の間で算出される関連度」とした．$X-A$ 間の平均値は 0.34，分散は 0.04，$X-C$ 間の平均値は 0.002，分散は $9.43\times10^{-}6$ であった．よって提案手法では，難解語と変換候補語との関連度が $X-A$ 間の平均値である 0.34 より低かった場合には N 語変換へ処理を移行する．

7.7.4　1 語変換の評価

評価は朝日新聞 [51] のネット上に掲載された記事からジャンルに関わらずランダムに取得した記事中の難解語 200 語を対象として行った．なお，これらの記事は 7.7.3 節で用いたものとは異なる記事である．またこれらの記事中の難解語 200 語はすべて 1 語変換の処理によって変換が行われる難解語である．元の記事と変換後の記事を提示し，変換結果が適切であるか否かを難解語 1 語ずつに対して評価を行った．評価は被験者 3 名（男性 2 名，女性 1 名）により行い，評価内容は変換前と比べて変換後の記事が平易になっているかという平易性および変換前の記事と変換後の記事で意味が変わっていないかという意味保持性の 2 つについて行った．平易性に関しては平易であるか否かの 2 パターン（○，×）の評価，意味保持性に関しては意味が完全に同じである，意味は通じるが違和感を感じる，意味が違っているという 3 パターン（○，×，△）の評価を行い，双方とも最終的に 3 名の評価の多数決で評価を決定した．

関連度の有用性を示すための比較として，単語親密度のみを変換後の決定基準として用いた場合の 1 語変換についても同様の評価を行った．平易性の評価結果を図 7.5 に，意味保持性の評価結果を図 7.6 に示す．

各評価者 2 名ずつの kappa 係数は平易性の評価では 0.756，0.821，0.796 であった．意味保持性の評価では 0.615，0.767，0.625 となった．結果として関連度を用いた提案手法では平易性が 72.0%，意味保持性が 88.0%となり双方とも単語親密度のみの場合と比べて高い評価となった．これにより難解語と変換候補語の関連性を考慮することが 1 語変換に有効であることが分かる．

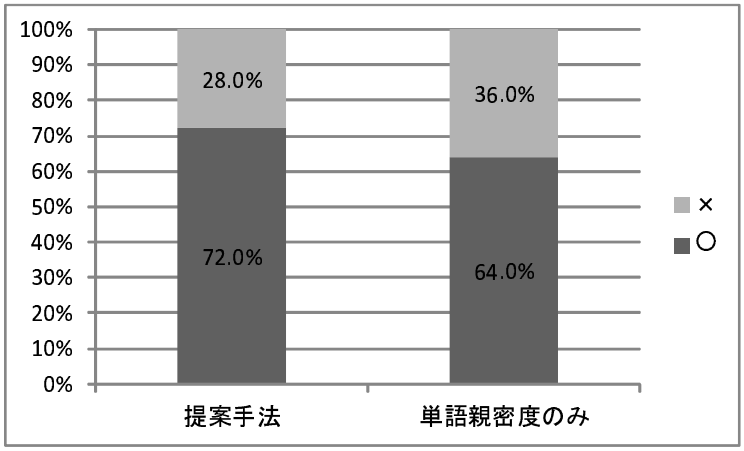


図 7.5: 1語変換の評価（平易性）

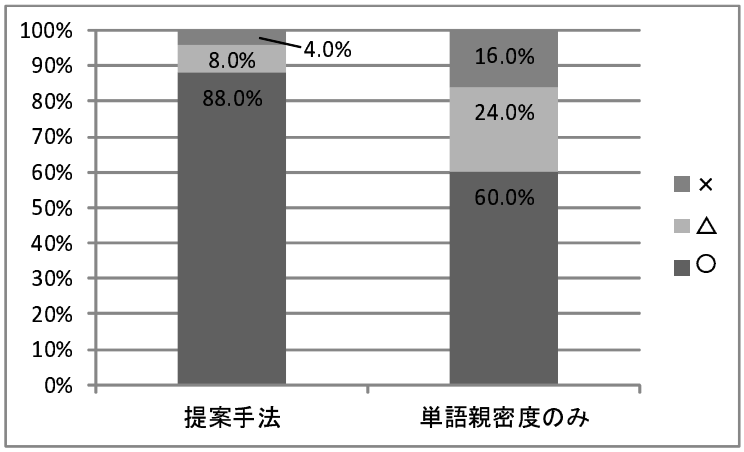


図 7.6: 1語変換の評価（意味保持性）

表 7.5 に具体的な変換例を示す．なお，評価に△もしくは×が存在する変換後の記事文については，括弧内に想定される正解例を示した．

2つ目の例にある「筋」という語は変換候補語として「線，血管，理屈」が得られる．このうち最も単語親密度が高い語は「線」であるが，関連度を考慮することで「理屈」が選ばれた．

「規定」という語の変換に関しては,「約束」と「ルール」の2つの変換候補語から「約束」が選択された．どちらの変換候補語が選択されても元の記事の意味は損なわないが,「約束」という語では私的な決め事というニュアンスが強いため「平易に変換されたが（平易性○），意味に違和感がある（意味保持性△）」という評価になり,「ルール」の方がより適切と判断された．

表 7.5: 1 語変換の例

元の記事文	変換後の記事文	評価
容認する考えを示した	**認める**考えを示した	平易性○, 意味保持性○
問題の**筋**が違う	問題の**理屈**が違う	平易性○, 意味保持性○
SNS を活用した 採用の**手法**	SNS を活用した採用の**テクニック** （SNS を活用した採用の**方法**）	平易性○, 意味保持性△
2002 年に制定した **規定**に基づく	2002 年に制定した**約束**に基づく （2002 年に制定した**ルール**に基づく）	平易性○, 意味保持性△
進行している**模様**	進行している**パターン** （進行している**様子**）	平易性×, 意味保持性×
横断中の 52 歳の男性 と**接触した**	横断中の 52 歳の男性と**コンタクトした** （横断中の 52 歳の男性と**当たった**）	平易性×, 意味保持性×

7.8 N 語変換

1 語変換では変換ができない場合，つまり 1 つの語では説明できない語を相手に伝える際に人間はその語の意味を文で伝える．そこで N 語変換では 1 つの単語を N 語の単語群，つまり文で変換することで 1 語変換ができない難解語の変換を行う．

まずシソーラスにおいて難解語の包含関係にあるノードに「具体物」が存在する場合には 1 語変換が不可能であると判断し，N 語変換を行う．例えば「サリドマイド」という具体物は一般的に馴染みの薄い語であるが,「催眠薬の一種」という文章で変換されることでその内容を理解することが出来る．このように具体的な物を示す語は，同じ意味を持つ別の 1 語に変換するよりも具体物の説明を文章で行う方が馴染みのある表現になる．また 7.7.3 節に示したように 1 語変換における変換候補語の関連度が閾値以下の場合にも，1 語変換では適切な変換を行えなかったと判断して N 語変換を行う．

7.8.1 変換候補文の取得

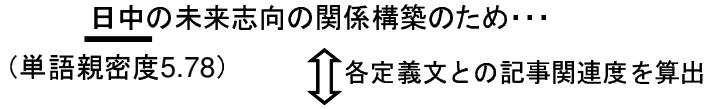
N 語変換では国語辞書 [17] に記載された語の定義文を，変換を行うための文（変換候補文）として利用する．国語辞書の定義文は語の意味を説明する文であるため，これを利用することで難解語の意味を損ねることなく N 語による変換が可能になる．また，定義文が端的かつ正しい日本語表現で記されているため，変換後の記事表現が煩雑にならないと考えられる点で，N 語変換の資源として国語辞書は適当である．

本章で使用した国語辞書には 238,000 語の見出し語とその定義文が格納されている．このうち，固有名詞および単一で意味を成さない代名詞，助詞の見出し語を省いた 94,544 語の見出し語と定義文を N 語変換に用いた．

7.8.2 多義語の意味特定

難解語が多義語であった場合，それぞれの意味から辞書の説明文が得られるため変換候補文が複数取得される．そこで適切な文を選択するために記事関連度計算方式を用いて難解語が含まれる元の文に意味が近い変換候補文を選択して変換を行う．図 7.7 に具体的な変換候補文の選択方法を示す.

「日中」という語には図に示すように 2 つの意味が定義文として記載されており，多義語である．このような多義語の場合は，辞書のそれぞれの定義文と，難解語を含む元の記事文との間で記事関連度の算出を行い，値の高い変換候補文を語の変換に用いる．例の場合では「日中」は「日本と中国」という候補文が選択され，記事は「日本と中国の未来志向の…」と変換される．



見出し語	定義文	記事関連度
日中	日の出ている間	0.09
	日本と中国	0.15

図 7.7: 多義語の意味特定の具体例

7.8.3 不自然さの排除

辞書の定義文の中には，そのままの形で N 語変換に用いると日本語として不自然になってしまうものがある．例えば「財政再生計画を策定する」という文中の「策定」は単語親密度が 3.16 の難解語であり，1 語変換では関連度が閾値より大きい変換候補語が得られず，N語変換が行われる語である．この時，辞書における「策定」の定義文「政策や計画などを考えて決めること」をそのまま語の変換に用いてしまうと「財政再生計画を政策や計画などを考えて決めること」となり，日本語として不自然である．このような変換によって起こる不自然さの排除方法として，不要語の削除と記事中の情報の重複排除を行う.

まず，不要語の削除について述べる．不要語とは辞書によく出現する言い回しのうち，変換を行う際には必要の無い語の事を指す．この不要語を人手で判断してリスト化したものが不要語リストである．図 7.8 に具体的な不要語の一覧を示す.

例えば「蜀魂」という語の定義文は「ホトトギスの別名」となっているが，実際に「蜀魂」という語を変換する際に必要となる語は「ホトトギス」の部分のみである．このように辞書の定義文に存在する不要な言い回しは変換の際に削除する.

不要語を削除した後に記事中の情報の重複排除を行うが，これには意味理解システム [46] を

すなわち，　転じて，　など，　こと，　さま，　ある
～の異名，　～の別名，　～の古名
～の謙譲語，　～の尊敬語，　～の丁寧語
あるいは，　もしくは，　または

図 7.8: 不要語の一覧

利用する．このシステムは入力された文を，6W1H（Who，What，When，Where，Whom，Why，How）と用言の 8 種類に分類する．意味理解システムの出力例を図 7.9 に示す.

入力文）妹が昨日，本屋で母に絵本を買ってもらった

Who	What	When	Where	Whom	Why	How	用言
妹	絵本	昨日	本屋	母			買う

図 7.9: 意味理解システム

入力文の「誰が」にあたる語は「妹」であり，これが意味理解システムでは Who に分類される．このシステムで元の記事文と辞書から得た変換候補文をそれぞれ処理し，分類が重複した場合には不自然にならないように不要部分を削除する．具体的な例を図 7.10 に示す.

図 7.10 の例では元の記事文「財政再生計画を策定する」と不要語を削除した変換候補文「政策や計画を考えて決める」の 2 文である．ここで元の記事文と変換候補文の間で分類に重複があった場合，どちらか一方を用いて出力する文を作成する．具体的には難解語ではない部分で分類の重複が起こった場合には元の記事文を，難解語の部分で分類の重複が起こった場合には変換候補文を用いる．図 7.10 を見ると What の重複は難解語ではない部分であるため，元の記事文である「財政再生計画」が選択される．逆に用言での重複は難解語の部分であるため，変換候補文である「考えて決める」が選択される．このようにして分類の重複を排除した上で，変換を行い結果を出力する．図 7.10 の例では最終的に「財政再生計画を策定する」という元の記事文が「財政再生計画を考えて決める」と変換される．

7.8.4　N 語変換の評価

N 語変換の評価は朝日新聞の記事からランダムに取得した記事中の難解語 200 語を用いて行った．この 200 語は 7.7.4 節で述べた 1 語変換の評価とは異なる記事から得られる難解語であり，すべての語で N 語変換が行われる．元の記事と変換後の記事を提示し，変換結果が適切であるか否かを難解語 1 語ずつに対して評価を行った．評価は被験者 3 名（男性 2 名，女性 1 名）により行い，評価内容は変換前と比べて変換後の記事が平易になっているかという平易性および変

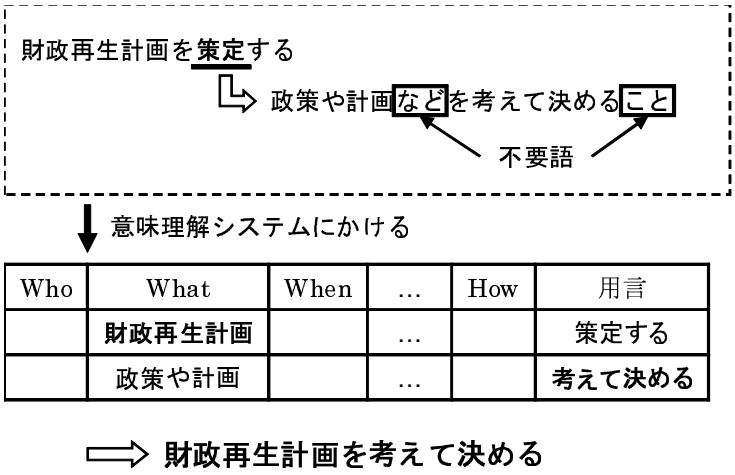


図 7.10: 格重複の排除

換前の記事と変換後の記事で意味が変わっていないかという意味保持性の 2 つについて行った．平易性に関しては平易であるか否かの 2 パターン（○，×）の評価，意味保持性に関しては意味が完全に同じである，意味は通じるが違和感を感じる，意味が違っているという 3 パターン（○，×，△）の評価を行い，双方とも最終的に 3 名の評価の多数決で評価を決定した．

評価としてベクトル空間モデル [3] を用いた N 語変換との比較を行った．変換候補文を選択する際に，提案手法では記事関連度計算方式により多義語の意味特定を行うが，比較手法ではベクトル空間モデルによりこれを行う．

ベクトル空間モデルは文書検索などの分野で広く使われる手法であり，文書を構成する語に重みを付与し，文書をベクトルとして表現する．文書検索の際には検索課題と検索対象の文章それぞれのベクトルが作る角度の余弦によって類似性を算出し，検索課題に近い文書を判断する．まず，変換対象語を含む元の記事文 o と変換候補文 d を，文中の語に付与された重み w で次のようなベクトルで表現する．

$$o = (w_{o1}, w_{o2}, \cdots, w_{oM}) \tag{7.10}$$

$$d = (w_{d1}, w_{d2}, \cdots, w_{dM}) \tag{7.11}$$

なお，ここでの M は変換対象語を含む元の記事文 o と変換候補文 d に現れる語の総数である．重み w の付与には様々な方法がとられるが，本章では $tf \cdot idf$ の考え方を用いた．元の記事文 o に対する変換候補文 d の得点 $s_o(d)$ は以下の式により定まる．

$$s_o(d) = \frac{\sum_{j=1}^{M} w_{dj} w_{oj}}{\sqrt{\sum_{j=1}^{M} w_{dj}^2} \sqrt{\sum_{j=1}^{M} w_{oj}^2}} \tag{7.12}$$

以上で示したベクトル空間モデルによる N 語変換および記事関連度計算方式による N 語変換を行い，結果を比較した．平易性の評価結果を図 7.11 に，意味保持性の評価結果を図 7.12 に示す．

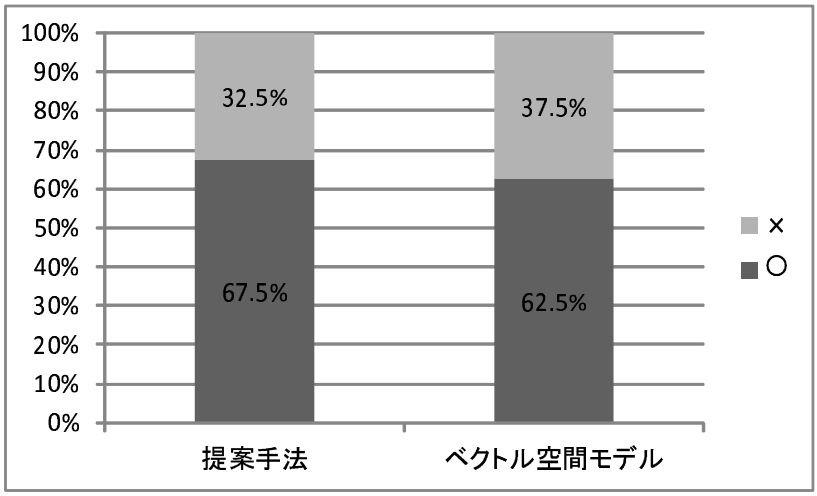


図 7.11: N 語変換の評価（平易性）

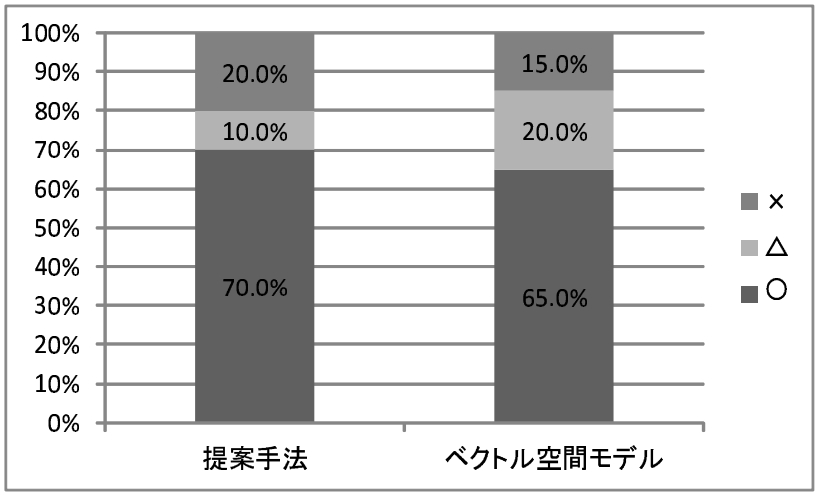


図 7.12: N 語変換の評価（意味保持性）

提案手法である記事関連度による N 語変換の評価は平易性で 67.5%，意味保持性で 70.0%となった．これにより記事関連度計算方式による辞書の定義文と元の記事文との対応付けが有効であることが分かる．また，7.8.3 節で述べた不自然さの排除処理について，不要語の削除は全

体の 21.0%, 記事中の情報の重複排除については 15.0%の変換において行われた.

表 7.6 に具体的な変換例を示す. 評価に△もしくは×を含む変換後の記事文については, 括弧内に想定される正解例を示した.

表 7.6: N 語変換の例

元の記事文	変換後の記事文	評価
他行よりも 高めの現行の金利	**よその銀行**よりも 高めの現行の金利	平易性○, 意味保持性○
態度が**急変**した	態度が**急激に変化**した	平易性○, 意味保持性○
事件の**経緯**をよく知る	事件の**経過**をよく知る （事件の**入り組んだ事情**を よく知る）	平易性○, 意味保持性△
識者はこう見ている	**物事に対して 正しい判断をくだす 力のある人**は こう見ている	平易性○, 意味保持性△
政治と金が**絡む**	政治と金が **他の物の周りに巻きつく** （政治と金が **密接に結びつく**）	平易性×, 意味保持性×
国の**根幹**	国の**根と幹** （国の**最も重要なところ**）	平易性×, 意味保持性×

1 つ目の例にある「他行」という語は, 同義語から「外出」という変換候補語を得るが, 元の記事は銀行の金利に関する内容であるためこれは誤りである. しかし関連度が閾値以下となるため, 1 語変換は行わずに N 語変換で処理される. 辞書では「他行」という語は「よその銀行」という正しい変換が行える定義文があるため, ○の評価となっている.

2 つ目の例にある「急変」という語は多義語であり,「急激に変化すること」という意味と「急に起こった変事」という意味が存在する. 元の記事内容は暴行事件における状況説明であり, 変換候補文としては「急激に変化すること」が正しい. 本手法で提案した記事関連度計算方式を用いた N 語変換ではこの「急変」の変換が正しく行えたが, 比較として行ったベクトル空間モデルによる変換ではもう一方の変換候補文である「急に起こった変事」が選ばれた. また,「急激に変化すること」という定義文は「こと」が不要語として削除され, 自然な表現に変換することができている.

「絡む」の変換に関しては多義性を適切に判別できなかった例である.「絡む」という語の意味としては「他の物の周りに巻きつく」は正しいが, 文脈から考えると「密接に結びつく」の方が適切である. この変換に関してはベクトル空間モデルを用いた場合にも同じく「他の物の周りに巻きつく」という変換を行ってしまっていた.

「識者」の例では N 語変換により長い文の形に変換されたことで, 言い回しとして冗長であ

り違和感があるという点から意味保持性の評価が△となった.

7.9　評価と考察

人間が違和感を感じない語の変換を行うためには，人間と同じく 1 : 1 の変換と 1 : N の変換を組み合わせることが必要である．それを踏まえ，ここまでで提案してきた 1 語変換および N 語変換の手法を統合した手法を，本章で提案する難解語の変換手法としてその評価を行う．変換処理を統合したことによる有効性を示すため，提案手法，難解語を無理やり 1 語変換のみで変換した場合，N 語変換のみで変換した場合の 3 種類の変換について評価を行った．評価実験の方法および評価結果について以下に述べる．

評価には朝日新聞から取得した 50 記事からランダムに選んだ記事文を利用した．全単語数は 1567 語，うち単語親密度の閾値によって難解語と判断された語は 249 語である．

まず被験者 3 名（男性 2 名，女性 1 名）に対して記事中の全単語を提示し，それぞれの語について会話に出現する語としたときに難解と感じるか，平易と感じるかの判断を行った．多数決により全単語を難解と感じる語，平易と感じる語に分類し，人が変換すべきと判断した語と変換しなくてよいと判断した語の判別を行った．

次に評価に用いた記事を変換前と変換後のセットにして被験者に提示し，平易な表現となっている方の記事を選択させる．この際，被験者にはどちらが変換前でどちらが変換後かは示さない状態で記事を提示し，表現が分かりやすいと感じる方を選択させた．提案手法により変換された後の記事が平易であると選ばれた場合に○，変換前が選ばれた場合に×の評価とする．

最後に同じ被験者 3 名に対して変換前と変換後の記事を各々がどちらの記事であるか示した上で，変換後の記事が意味的に欠損したり違和感のある表現になっていないかという意味の保持性について評価を行った．意味が保持され，違和感もない場合には○，何らかの違和感を感じる場合には△，意味が違っていたり，日本語表現としておかしいといった意味が保持されていない場合に×の評価とした．

手法の評価は，人が変換すべきと判断した語と変換すべきでないと判断した語に分けて示す．まず人が変換すべきと判断した語については，変換手法により語が適切に変換されたか否かを評価した．なお，人が変換すべきと判断した語は 222 語，そのうち提案手法によって変換された語は 206 語であった．平易性に関する評価結果を図 7.13 に，意味保持性に関する評価結果を図 7.14 に示す.

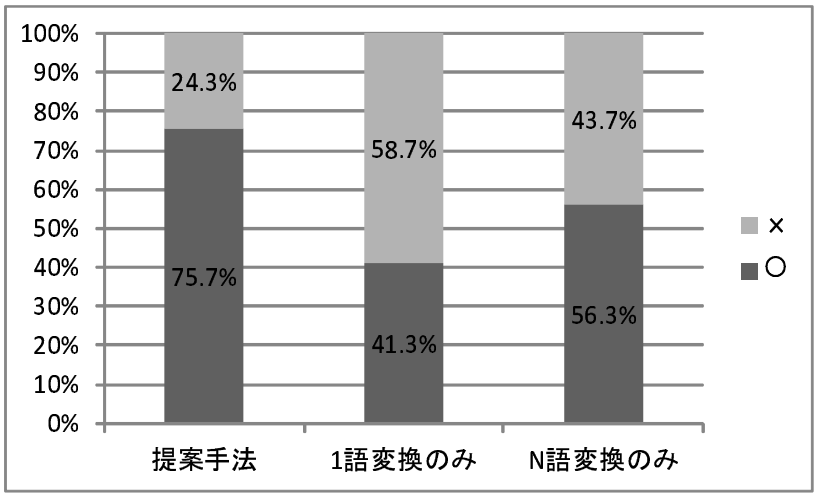


図 7.13: 変換すべき語の評価結果（平易性）

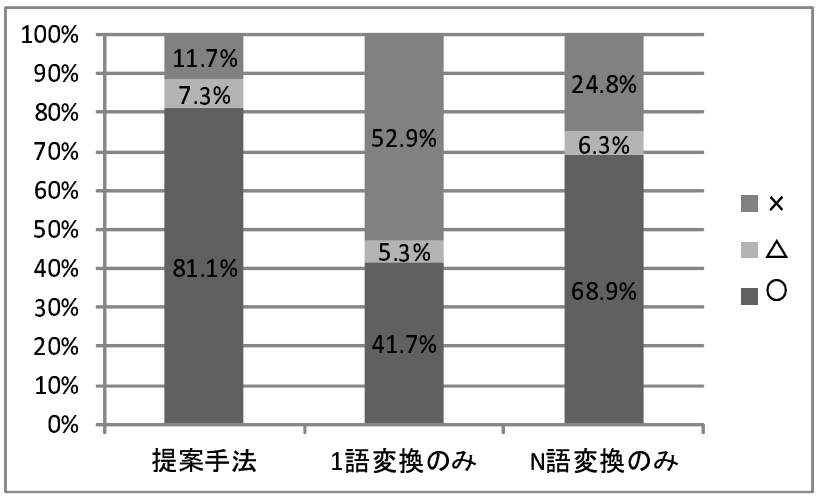


図 7.14: 変換すべき語の評価結果（意味保持性）

提案手法では平易性の評価が 75.7%，意味保持性の評価が 81.1%となった．難解語となった 249 語について 1 語変換および N 語変換のどちらで処理が行われたかの内訳は，1 語変換によって処理された難解語が 76 語，N 語変換によって処理された難解語が 173 語であった．

1 語変換のみで変換を行った場合，平易性で 58.7%，意味保持性で 52.9%が×の評価となった．これは，1 語変換では同義語および類義語が存在しない場合は変換することが出来ないため，変換すべき語の多くが変換できず難解語のまま残ってしまったためである．今回の評価で

はすべての難解語のうち 48.1%にあたる 99 語が変換不可となった.

N 語変換のみの場合は辞書に定義された語であれば変換可能であるため，変換不可の語は全体の 7.3%，15 語に留まったが，こちらも提案手法と比べて平易性，意味保持性共に評価は低くなっている.

次に変換すべきでないと判断した 1345 語についての評価を表 7.7 および表 7.8 に示す.

人が変換すべきでないと判断した語に関しては，変換が行われないもしくは変換されてしまったが平易である，意味が保持されている場合にそれぞれ○の評価としている．すべての手法において，○の評価は高くなっている．1 語変換のみの場合には，前述したとおり変換がそもそも不可能である難解語が多かったため，変換すべきで無い語の多くが変換されないままとなり○の評価が高くなっている.

表 7.7: 変換すべきでない語の評価結果（平易性）

	提案手法	1 語変換のみ	N 語変換のみ
○	97.5%	99.9%	98.0%
×	2.5%	0.1%	2.0%

表 7.8: 変換すべきでない語の評価結果（意味保持性）

	提案手法	1 語変換のみ	N 語変換のみ
○	98.4%	99.9%	98.4%
△	0.4%	0.0%	0.4%
×	1.2%	0.1%	1.2%

以下に実際の変換例を示す．表 7.9 に 1 語変換のみを用いた場合と提案手法の変換例を，表 7.10 に N 語変換のみを用いた場合と提案手法の変換例を示す．なお，括弧の中は各変換の評価を示す.

「送還」という難解語は同義，類義語が得られず，1 語変換のみでは変換することができない．N 語変換では送還の意味として「送り返すこと」が存在するため正しい変換が行える.

「維持」の例では 1 語変換を行うと類義語から「持つ」という変換候補語が得られる.「持つ」は確かに「維持」を平易に変換したものだが，文脈から不自然であると判断されて意味保持性は×となった．これを提案手法で変換すると N 語変換により「保ち続ける」という変換がされ，双方とも○の評価となった.

最後の例では難解語が 2 つ存在している．1 語変換のみの場合には「破片」という難解語が「かけら」に変換されるが,「落下」に関しては変換候補語が得られずに変換できない．提案手法では N 語変換により「落下」に関しても「下に落ちること」という定義文から変換が可能となり，結果として表 7.9 のような変換が行えた.

表 7.9: 1 語変換のみと提案手法の比較

元の記事文	提案手法	1 語変換のみ
送還してほしい	**送り返して**ほしい （平易性○，意味保持性○）	- （変換されず×）
状態を**維持して**いる	状態を**保ち続けて**いる （平易性○，意味保持性○）	状態を**持って**いる （平易性○，意味保持性×）
岩の**破片**などが **落下して**きても	岩の**かけら**などが （平易性，意味保持性○） **下に落ちて**きても （平易性○，意味保持性○）	岩の**かけら**などが （平易性○，意味保持性○） 落下してきても （変換されず×）

表 7.10: N 語変換のみと提案手法の比較

元の記事文	提案手法	N 語変換のみ
虚偽の申告	**嘘**の申告 （平易性○，意味保持性○）	**真実ではないと知りながら真実であるかのようにみせる**申告 （平易性○，意味保持性△）
双方の運転手から	**両方**の運転手から （平易性，意味保持性○）	**関係しているあちらとこちら**の 運転手から （平易性○，意味保持性△）
国内で初めて**承認され**	国内で初めて**認められ** （平易性，意味保持性○）	国内で初めて **その事柄が正当であると判断され** （平易性×，意味保持性○）

表 7.10 に示した例を見ると，N 語変換のみを用いた場合の変換結果は意味的には元の記事文と相違ない．しかしこれらの表現が会話中に現れると想定すると多くの人は不自然であると感じる．N 語変換は文章による変換であるため，この変換が多用されると変換後の記事が冗長であると感じやすく，結果として意味保持性において違和感を感じてしまい△の評価となる場合や，平易性の無い×の評価となっている．一方，提案手法ではこれらの記事は 1 語変換によって変換され，その変換結果は意味を損ねず，かつ違和感もないことがわかる．以上のことから，1 語変換および N 語変換を組み合わせることによってより人間にとって違和感の無い変換が行えることが分かる．

提案手法では難解語の判別に 7.6.1 節で示した閾値を用いている．本評価では人が変換すべきと判断した 222 語のうち 206 語が難解語と判別されており，つまり変換すべき難解語 16 語がこの閾値では難解語と判断されていない．人が難解と感じる語を 100.0%判断できることを重視す

る場合には親密度の閾値を上げればよいが，閾値が高すぎると記事中の多くの語が難解語と判断されると考えられる．各変換処理によってそれらの語が全て別表現に変換されると変換後の記事が冗長になる可能性がある．そこで本評価において人が変換すべきと判断したが，提案手法では難解語と判別されなかった語のうち最も高い親密度であった「汚染」という語を基準として評価を行い，提案手法との比較を行った．具体的な親密度の閾値は 6.15 である．平易性に関する評価結果を図 7.15 に，意味保持性に関する評価結果を図 7.16 に示す.

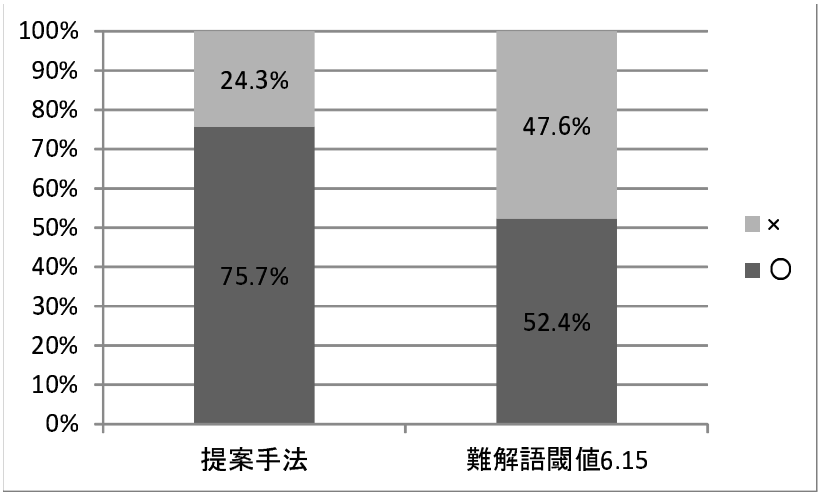


図 7.15: 難解語判別の閾値変更による評価結果（平易性）

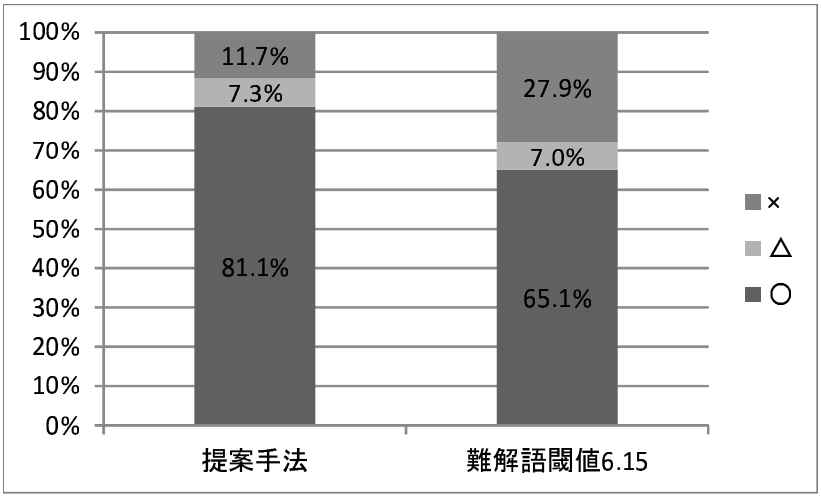


図 7.16: 難解語判別の閾値変更による評価結果（意味保持性）

結果として，難解語判別の閾値を 6.15 とした場合には平易性，意味保持性ともに評価が下がる結果となった．表 7.11 に提案手法による出力と難解語判別の閾値変更を行った場合の出力の比較を示す.

難解語判別の閾値を高くしたことで，提案手法では変換されなかった「視野」や「交差点」といった語が難解語と判断されている．これらの語は目視では変換しなくて良い語と判断されていた語である．これらが難解語と判断されることによって，例えば「視野」という語は N 語変換により「視線を固定したままの状態で見ることのできる範囲」と変換されている．これは意味としては正しいが，元の「視野」という 1 語の表現と比べて非常に冗長であるため，平易性において評価が×となっている.「交差点」の評価も「視野」と同様に冗長な表現で平易性を失っている．しかし同じ記事中の「右折」という語は，人が変換すべきと判断したが提案手法では変換されなかった 16 語の 1 つであり，これは「右へ曲がる」という平易な表現へ変換することが出来ている.

閾値を上げることで，人が変換すべきと判断したが提案手法では変換されなかった 16 語に対しての変換は行われたが，それに伴い人が変換しなくてよいと判断した語に対しても多くの変換が行われた．具体的には新たに 88 語が難解語となったが，これらの語は人の判断では変換せずとも平易であるとされており，変換を行うことで逆に平易性が損なわれやすい結果となった．このことより，7.6.1 節において示した閾値の設定は有効であると考える.

表 7.11: 提案手法と難解語判別の閾値変更の出力比較

提案手法	難解語判別の閾値変更後
一部には火を付けた者もいた	**ひとまとまり**には火を付けた者もいた （平易性×，意味保持性×）
売り**注文**が出ている	売り**リクエスト**が出ている （平易性×，意味保持性△）
国内で初めて認められ	**一国の領土内**で初めて認められ （平易性×，意味保持性○）
視野の中心が暗くなったりする	**視線を固定したままの状態で見ることのできる範囲**の中心が暗くなったりする （平易性×，意味保持性○）
政府間で固まっていた**日程**	政府間で固まっていた**物事を行うときの予定** （平易性×，意味保持性○）
交差点を**右折した**際に	**二本以上の線，特に街路が交わっている所**を （平易性×，意味保持性○） **右へ曲がった**際に （平易性○，意味保持性○）
命令に 3 回**違反した**	命令に 3 回**従わなかった** （平易性○，意味保持性○）

以上の評価結果より，会話中に出現する語の単語親密度によって難解語の判別を行い，1 語変換と N 語変換を組み合わせることで平易な表現への変換を行う提案手法の有効性を示した.

7.10　おわりに

本章では，語概念連想システムを一般的な自然言語処理分野において活発に議論される処理に活用する事例として，新聞記事中の難解な語を会話に適した平易な表現へ変換する手法を提案した．変換の際には人間が行う語の変換処理に沿い，1 つの語を別の 1 語で変換する 1 語変換および文章で変換する N 語変換を組み合わせることでより人間にとって自然な変換が行えることを示した．1 語変換では難解な語を同義語・類義語により別の 1 語へ変換し，N 語変換では 1 語では変換できないような語や具体物を示す語について文による変換を行った．この二つの変換を組み合わせた変換手法を提案，評価してその有効性を示した．また，処理の各段階において語概念連想システムにより語と語，文と文の関連性の有無を判断することで変換前の語から変換後表現にかけて意味の保持を行った．最終的な結果として新聞記事 50 件，249 語の変換を行い，変換すべき難解語を結果として変換すべき難解語を 75.7%の精度で平易な表現に，81.1%の精度で正しい意味を保持した表現に変換することが出来た．

第8章　知的会話における連想応答の生成手法

8.1　はじめに

質問や返答，提案，雑談のような種々の会話は相手と自身の発話のやり取りによって形成される．ロボットと人間においても自然な会話を行うためには，ロボット側が独立して発話を行うのではなく，相手の発話を受けて適切な発話を返すことが必要と考えられる．そこで本章ではロボットに対して人間のような会話能力を持たせるために，相手の発話を受けてそれに適する知的な応答を生成する手法について述べる．

会話システムに関する研究としては，タスク型のシステムや膨大な応答パターンのコーパスを用いたシステムなどが一般的である．例えばタスク型のシステムでは，旅館予約システム [52]，秘書システム [53] といったものが報告されている．これらはタスク達成のために必要な情報を得るための質問や応答をルールとして定義し，それに従った会話を行う．そのためルールから外れた発話への対応や，タスク以外の話題を持つ会話の実現は難しい．また，応答パターンのコーパスを用いたシステムとして文献 [54] などが報告されている．これは人間とのチャット環境に置かれ続けることで膨大な応答パターンを保有していき，そこから適切な応答のルールを学習することでタスクに依存しない雑談的な会話を生成することを目指している．しかしパターンの学習によって行われるシステム側の発話はあくまで過去に存在した応答の焼き直しであり，相手の発話を受けてその内容から適切な応答を自律的に生成している訳ではない．

本章で述べる手法は，従来のタスク型の会話システムや膨大な対話例のコーパスに依存した応答生成とは違い，人間の発話からそれに適した自律的な応答を生成することを目指す．特定のタスクに依存しない表現や思考を持つ人間の発話に対応し，その発話に対して常識的かつ過去の焼き直しではない自律的な応答を生成する．これを実現するためには人間が持つ言葉の知識や，人間らしい常識的な判断の機構を会話システムへ組み込む必要がある．そこで語概念連想システムによる語の連想機能および，それを基盤として構築される常識判断システムを会話システムに活用することで，自律的な応答を生成する手法を提案する．

8.2　常識判断システム

常識判断システムは，人間が意識的もしくは無意識的に行っている「常識的な判断」を機械上に構築することを目的として構築された技術である．

人間は様々な事柄に関して常識的な知識を持ち合わせており，この知識を応用することで事象を常識的に判断することが出来る．つまり様々な事柄に関する常識的な知識を機械上に体系づけすることで，常識的な判断をシステムすることが出来ると考えられる．しかし，現実世界

には「様々な事柄」は無尽蔵に存在しており，これら全てを知識として体系づけることは不可能である．そこで常識判断システムでは，常識知識を複数の観点（量，時間，感覚，感情，場所など）に類別し体系づけ，さらに各観点において少数の代表的な語とそれに関する知識のみを定義する．この代表的な語はもちろん現実世界の様々な事柄を網羅は出来ていないが，定義されなかった事柄を，定義された知識に対応付けることでこれを解決する．これには人間が持つ言葉の知識とそれを用いた連想を行う語概念連想システムを用いている．本章ではこの常識判断システムを人間の発話内容から常識的な応答を生成するために活用する．前述した常識の観点のうち，提案手法では場所および感覚に関する常識判断システムを用いた応答文の生成を行う．

8.2.1　場所判断システム

場所判断システム [55] は場所を表現する語からその場所に存在する人や物，行われる事象を想起する.「神社」や「病院」のように，ある人・物・事象が実際に存在する具体的な場所を表す語を対象とし，これらの語を場所語と定義する．現実世界において場所語は多種多様に存在しているが，このうち代表的な場所語に関する知識のみが場所判断知識ベースに格納されている.

場所判断知識ベースには，その場所に存在する人や物を表す場所主体語と，その場所で行われる事象を表す場所目的語が各々の代表的な場所語と関連付けられて登録されている．場所語のうち，知識ベースに格納される語を「代表語」として，これら代表語をノードまたはリーフとするシソーラス構造により代表語と場所主体語，場所目的語の関係を効率よく表現する．ノードとなる語を「分類語」と定義し，分類語および代表語はそれぞれ場所主体語と場所目的語を知識として持つ．分類語の持つ場所主体語と場所目的語はシソーラスの下位ノードおよびリーフに継承される構造となっている．図 8.1 に場所判断知識ベースのイメージを示す.

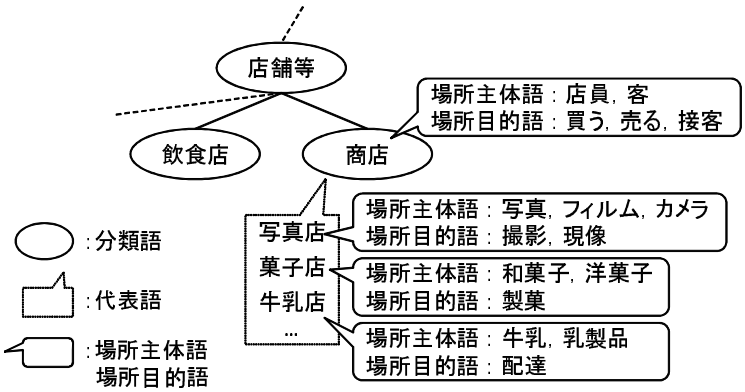


図 8.1: 場所判断知識ベースのイメージ

例えば分類語「商店」は場所主体語として「店員，客」という知識を持っている．これらの場所主体語は木構造の下位に継承されるため,「商店」に分類される代表語「写真店」は自身が

持つ場所主体語「写真，フィルム，カメラ」に加えて「店員，客」という語を得る．実際の場所判断知識ベースの例を表 8.1 に示す.

表 8.1: 場所判断知識ベースの一部

場所語	場所主体語	場所目的語
神社	神主，鳥居，•••	参拝，祈願，•••
病院	医者，患者，•••	診察，入院，•••

場所判断システムの処理として，まず入力された語が場所語であるか否かの判断を行う．入力語 X が場所判断知識ベースに登録されている分類語 P に対して意味的な関連性が極めて強い場合，すなわち，未知語 X と分類語 P との関連度が定義した閾値を越える場合，入力語 X は場所語であると判断する．これを未知語処理手法と呼ぶ.

次に場所語であると判断された語について，場所主体語と場所目的語を想起する．場所語のうち場所判断知識ベースに格納されている既知語については，知識ベースに登録されている場所主体語と場所目的語を出力する．場所判断知識ベースに格納されていない未知語については，未知語処理手法により関連が強いと判断された分類語に属する代表語との関連度計算を算出し，最も関連が強い代表語に未知語を対応付ける．そして，対応付けられた代表語に登録されている場所主体語，場所目的語との関連度を算出し，その値があらかじめ設定したある閾値を越える場合のみ，未知語に関連付ける場所主体語，場所目的語として出力する．システムの出力例を表 8.2 に示す.

表 8.2: 場所判断システムの出力例

入力	場所語か	場所主体語	場所目的語
牛乳	×	−	−
本屋	○	客，店長， 本，雑誌，•••	売る，買う， 接客，•••
時計	×	−	−

8.2.2 感覚判断システム

感覚判断システム [56] は名詞に対して，人間が常識的に想起でき，特徴付けられる感覚に関する語を取得する．この「感覚」とは五感（視覚・聴覚・嗅覚・味覚・触覚）の刺激によって得られる感覚を指す．感覚判断システムは名詞とその特徴である「感覚」の関係を日常的な名詞の知識ベース (感覚判断知識ベース) を構築することによって明確にし，必要な感覚語を取得する.

感覚判断知識ベースは人間の五感で感じる感覚の知識である「感覚語」を集めた知識ベースである．場所判断知識ベースと同じく，シソーラス構造のノードを「分類語」と定義し，各分類語における代表的な語を「代表語」として格納している．代表語および分類語はそれぞれの感覚を表現する語（感覚語）を知識として持ち，分類語の持つ感覚語はシソーラスの下位ノードおよびリーフに継承される．感覚判断知識ベースのイメージを図 8.2 に示す.

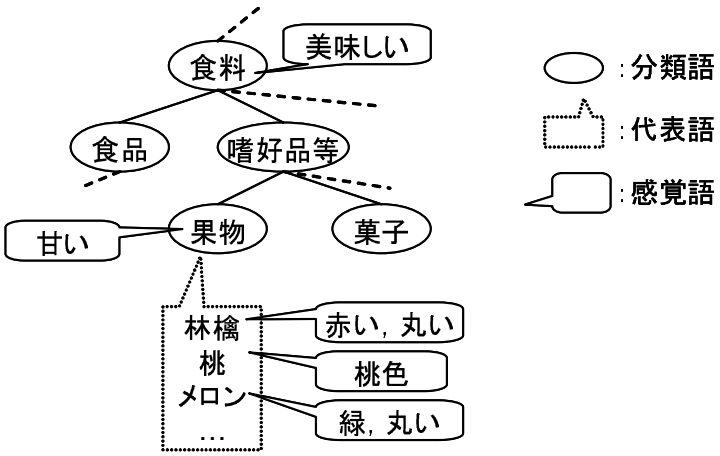


図 8.2: 感覚判断知識ベースのイメージ

分類語「果物」は「甘い」という感覚語を持っている．感覚語は木構造の下位に継承されるため,「果物」に分類される代表語「林檎」は自身が持つ感覚語「赤い，丸い」に加えて「甘い」という感覚語を得る．実際の感覚判断知識ベースの例を表 8.3 に示す.

表 8.3: 感覚判断知識ベースの例

分類語	代表語	感覚語
果物	林檎	赤い，丸い，甘い，美味しい
果物	桃	桃色，甘い，美味しい
果物	メロン	緑，丸い，甘い，美味しい
…	…	…

感覚判断システムの処理として，まず入力された語が感覚知識ベースにおいて分類語もしくは代表語として存在する場合には，それらが持つ感覚語を出力として提示する．感覚判断知識ベースにない未知語に対しては，関連度計算を用いて感覚判断知識ベース内に含まれる高い関連度を持つ代表語を取得し，その語群の属するシソーラスのノードから分類語としての感覚を取得する．さらに，概念ベースを用いて未知語の一次属性に現れる感覚語を未知語固有の感覚

として取得する．表 8.4 に未知語「パンダ」における固有の感覚語取得の例を示す.

表 8.4: 未知語固有の感覚取得の例

概念	属性
パンダ	熊，動物，**白**，ライオン，自然，生きる，チベット，ぬいぐるみ，足，**黒**，山，中国，**大きい**，林，竹，…

未知語「パンダ」では，その属性から「白・黒・大きい」などを未知語固有の感覚語として取得することができる．感覚判断システムを用いた例を表 8.5 に示す.

表 8.5: 感覚判断システムの使用例

概念	感覚語
林檎	赤い，甘い，丸い
夕焼け	眩しい，赤い，美しい
騒音	煩い

8.3 連想応答の生成

8.2 節で述べた場所および感覚の常識判断システムを用いて，人間の発話内容から連想される常識的な応答（連想応答）を生成する．応答は人間の発話中の場所に関する情報を起点として，場所での行動を連想して応答する「場所連想」，その場所に存在する人や物への一般的な感覚による共感を応答する「形容詞連想」，そして人間の発話内容に関連はあるが異なる話題を応答する「話題転換連想」の三つの処理により生成する.

8.3.1 場所連想

場所連想では人間の発話中の場所に関する情報，つまり場所語から連想される行動についての応答を生成する．例えば人間が「美術館へ行きました」と発話した場合，美術館で行う常識的な行動を連想することで「絵画を見てきたのですか？」という応答を返す.

まず人間の発話を 7 章 7.8.3 節において述べた意味理解システムによって 6W1H（Who，What，When，Where，Whom，Why，How）と用言に分類する．この時，応答の起点となる場所語が Where に分類されており，かつ What に分類される語が存在しない，つまり「何をしたのか」が分からない場合に場所連想による応答生成を行う．例えば,「美術館へ行きました」という発

話文の場合，Where に「美術館」，用言に「行く」という語が分類される．この時 What に分類される語がないため，場所連想により美術館で何をしたのかという応答を連想する．

処理として，まず発話が「場所を訪れた」事に関する内容であるかの判断を行う．これは，例えば「遊園地を所有している」という発話の場合は Where に場所語である遊園地が分類されるが，この発話は遊園地を訪れたという内容ではない．そのため，遊園地での行動についての応答は不適切である．そこで「行く」や「訪れる」のように場所を訪れたことを主内容とする動詞を格納した場所動詞知識ベースを構築し，意味理解システムで用言に分類される語がこの知識ベースに存在する場合に人間の発話が「場所を訪れた」ことに関する内容だと判断して応答生成を行う事とする．場所動詞知識ベースに格納されている動詞は「行く，訪れる，訪問する，出発する，出かける，連れて行く，飛び立つ，帰る，帰宅する，戻る，近づく，入る，来る」の全１３語である．

「美術館へ行きました」という発話の場合，用言に分類された「行く」という語は場所動詞知識ベースに格納されている．そのため，この発話は「場所を訪れた」事に関する内容であると判断される．

次に，Where に分類された場所語を場所判断システムの入力とし，場所主体語および場所目的語を取得する.「美術館へ行きました」という発話の場合には,「美術館」が場所判断システムの入力となる．出力例を表 8.6 に示す.

表 8.6:「美術館」に対する場所主体語と場所目的語の出力例

場所語	場所主体語	場所目的語
美術館	館長，芸術家，彫刻，美術品，絵画，作品，芸術品，工芸，書芸，フレスコ画	展示，閲覧，鑑賞，観る，見る

この場所主体語および場所目的語から「何をしたのか」という応答を生成するが，この時，場所主体語から人物に関わる語を，場所目的語から受け身をとる頻度が高い語を省く．人物に関わる語，例えば表 8.6 における「館長」や「芸術家」といった語は日常会話の話題としてはあまり使用されないと考えられる．人物に関わる語の判別には 3 章 3.6.1 節に示したシソーラスを用い，場所主体語のうち「人物」ノードに含まれる語を除く．また，受け身をとる語は発話を行った人間自身の行動に繋がりづらいため削除することとした．受け身をとる頻度が高い語に関しては Web から自動構築された大規模格フレームシステム [27] を用いて判別を行う.「美術館」から得られた場所主体語および場所目的語を用いた処理例を表 8.7，表 8.8 に示す.

大規模格フレームシステムに用言「展示」，名詞「絵画」を与えると，表 8.7 よりガ格が最多，表 8.8 よりそのガ格において受動態が最多であることがわかる．そこで,「展示」は受身で用いられることが多いと考えられるため，このような語を削除する.

最後に Where に分類された場所語と，場所主体語群で関連度の算出を行い，最も高関連度の主体語を取得する．さらに，取得された主体語と場所目的語群とで同じく関連度の算出を行い，

表 8.7: 格に対する頻度数

名詞 (入力)	用言 (入力)	格 (出力)	件数 (出力)
絵画	展示	ガ格	376
		ヲ格	267
		ノ格	47
		カラ格	2

表 8.8: 「絵画が展示」に対する態の頻度

名詞 (入力)	用言 (入力)	格 (入力)	態と件数 (出力)
絵画	展示	ガ格	能動態 28 件 受動態 348 件 使役 0 件

最も高関連度の目的語を取得する．これらの取得された主体語および目的語を用いて場所連想の応答を生成する．

図 8.3 に「美術館へ行きました」という発話に対する場所連想の処理例を示す.

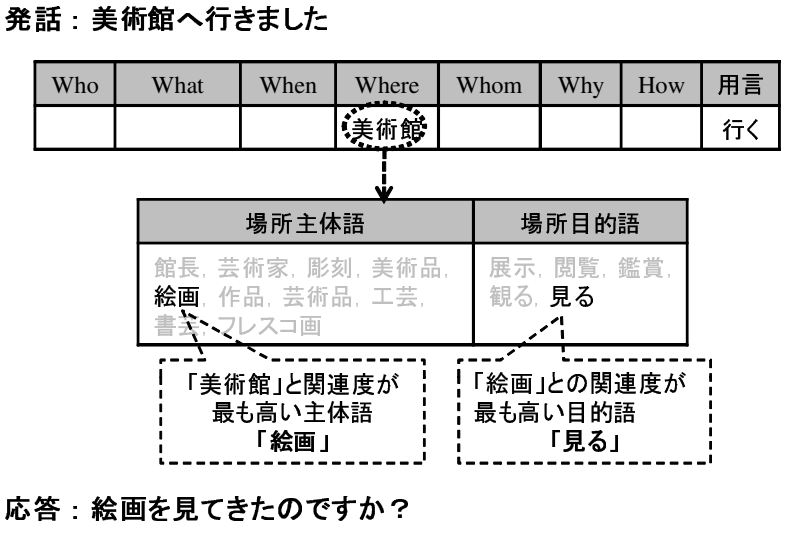


図 8.3: 場所連想の処理例

不必要な語を除いた場所主体語群と場所語をそれぞれ関連度計算した結果,「美術館－ 絵画」

の関連度が最も高かった．そのため，場所主体語「絵画」を選択する．更に，選択した主体語と場所目的語をそれぞれ関連度計算し，結果として「絵画－ 見る」の関連度が最も高かった．このため，連想応答に用いる語として「絵画－ 見る」を選択する．この処理が，人間の発話で分からなかった「何をしたのか」を連想したことになる．

この選択した名詞と用言の組を用いて，8.3.4 章にて後述する語尾変換および適用条件に合致するテンプレートを利用して応答文を作成する．この手法によって,「美術館へ行きました」という人間の発話に対し,「絵画を見ましたか？」や「絵画を見てきたのですか？」といった連想応答文を作成する．

8.3.2　形容詞連想

形容詞連想では人間の発話中の場所に存在する人や物への常識的な感覚による共感を示す応答を生成する．図 8.4 に「草原でタンポポを見ました」という発話に対する形容詞連想の処理例を示す.

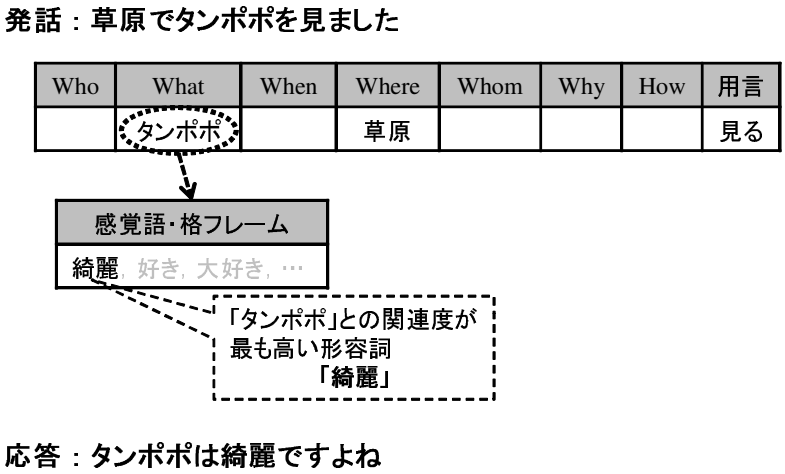


図 8.4: 形容詞連想の処理例

まず，場所連想と同じく人間の発話を意味理解システムにより分類する．ここで形容詞連想では場所連想と違い，What に分類される語が存在する場合にも処理を行う．形容詞連想では場所に存在する人や物から感覚を連想するが，What に分類される語が存在する場合には人間の発話中にすでに人や物が含まれていることになる．よって What に分類される語から感覚を連想することで，応答を生成することが出来る．What に分類される語が存在しない発話に関しては，場所連想と同じ手法を用いて場所語から場所主体語を取得することでこれらを得る．

次に What に分類された語もしくは取得された場所主体語を用いて，感覚判断システムにより適切な感覚語を取得する．また，補助的に Web から自動構築された大規模格フレームシス

テムを用い，上位 30 件の形容詞を取得する．取得できた感覚語および形容詞と What に分類された語との関連度を計算し，最高関連度をもつ形容語を決定する．例の場合,「タンポポ− 綺麗」の組が最も高い関連度となり，連想応答に用いる語としてこれらが選択される．この選択した名詞と形容語の組を用いて，語尾変換を行い，適用条件に合致するテンプレートを利用して「タンポポは綺麗ですよね」のような応答を作成する．

8.3.3 話題転換連想

話題転換連想では人間の発話内容に関連はあるが異なる話題を持った応答を生成する．具体的には，人間の発話内容から「どこで何をしたのか」が判明している場合に，その場所でできる他の行動について問いかけることで話題の転換を狙う．図 8.5 に「カジノでポーカーをしました」という発話に対する話題転換連想の処理例を示す．

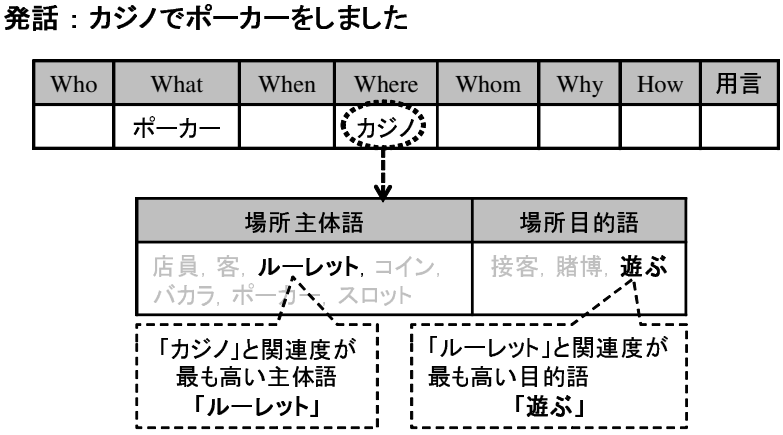


図 8.5: 話題転換連想の処理例

まず，場所連想と同じく人間の発話を意味理解システムにより分類し，更に場所語「カジノ」から場所主体語および場所目的語を取得する．ここで場所主体語のうち，人間の発話内容から判明している「ポーカー」を不要語として削除する．これにより，人間の発話と関連はあるが，異なる話題を持った主体語のみを残すことが出来る．

次に不要語を削除した場所主体語群と場所語をそれぞれ関連度計算し，関連度が最も高かった「カジノ− ルーレット」の組を選択する．更に，選択した主体語と場所目的語をそれぞれ関連度計算し，結果として「ルーレット− 遊ぶ」の関連度が最も高かった．このため，連想応答に用いる語として「ルーレット− 遊ぶ」を選択する．この選択した名詞と用言の組を用いて，語尾変換を行い，適用条件に合致するテンプレートを利用して「ルーレットで遊びましたか?」のような応答文を作成する．

8.3.4　応答文テンプレート

応答文テンプレートとして，以下の表 8.9 に示すテンプレートパターンを知識ベース化しておく．

表 8.9: テンプレートパターン

適用条件	テンプレートパターン
用言が「動詞」 または「サ変接続」 の場合	・{ 場所主体語 } を (に/は)　{ 場所目的語 } てきたのですか？ ・{ 場所主体語 } を (に/は)　{ 場所目的語 } のですか？ ・{ 場所主体語 } を (に/は)　{ 場所目的語 } ましたか？
用言が 「形容詞」の場合	・{ 場所主体語 } を (に/は)　{ 状態語 } ですよね ・{ 場所主体語 } を (に/は)　{ 状態語 } でしたか？ ・{ 場所主体語 } を (に/は)　{ 状態語 } のですか？

助詞の選択では大規模格フレーム検索を用いる．大規模格フレーム検索では名詞と用言の用法を検索することができるため，名詞に場所主体語，用言に場所目的語及び状態語を指定し，名詞と用言の二語の関係から複数の用言の用法と頻度を取得することができる．最も高い頻度をもつ用法を助詞として獲得する．

また，用言の語尾変換は過去形とし，テンプレートに沿うように動詞を過去形に変換する．動詞の過去形は,「連用形」＋「た」が基本形であり，上一段活用の「着る」を例にすると，連用形は「着」となるため「着」＋「た」となる．しかし，五段活用「買う」の場合，連用形は「買い」になるので,「買う」の過去形は「買い」＋「た」となる.「買う」の過去形は「買った」であるので,「買い」を連用形の音便形を用いて,「買っ」と変換する必要がある．

以上のように，上一段活用動詞や下一段活用動詞では音便処理は必要ないが，五段活用動詞については音便処理が必要となる．五段活用動詞における，過去形とするための音便の種類を表 8.10 に示す．未然形の種類によって分類される．

表 8.10: 音便の種類

未然形	音便	変換語	具体例
カ，ガ	イ音便	い	書く→書い＋た
サ	イ音便	し	探す→探し＋た
ア，ラ，ワ	促音便	っ	買う→買っ＋た
ナ，バ，マ	撥音便	ん	遊ぶ→遊ん＋だ

8.4 評価と考察

提案した連想による応答手法が応答文として適切であるかを評価する．評価データは中学英語テキスト [57–62] の日本語訳文を用いる．中学校で学習する英語はコミュニケーションの基礎を養うことを目標としてカリキュラムおよびテキストが作成されており，基礎的なコミュニケーション会話の指標になりうると考えられる．そこでこれらのテキストより，連想応答の生成条件に合致する会話文を，場所連想，形容詞連想，話題転換連想それぞれの評価用に各 100 文，計 300 文抽出して評価データとした．

評価データを入力とし，提案した連想応答手法を用いて出力した応答文を被験者 3 名にて評価した．入力文と応答文を評価者に対して同時に提示し，作成された応答文が常識的か非常識かの判別を行う．この結果，3 名中 2 名以上が常識的としたときを「○」，3 名中 1 名が常識的としたときを「△」，3 人全員が非常識としたときを「×」として評価を行った．

場所連想，形容詞連想，話題転換連想のそれぞれに対して評価を行った結果を図 8.6 に示す．

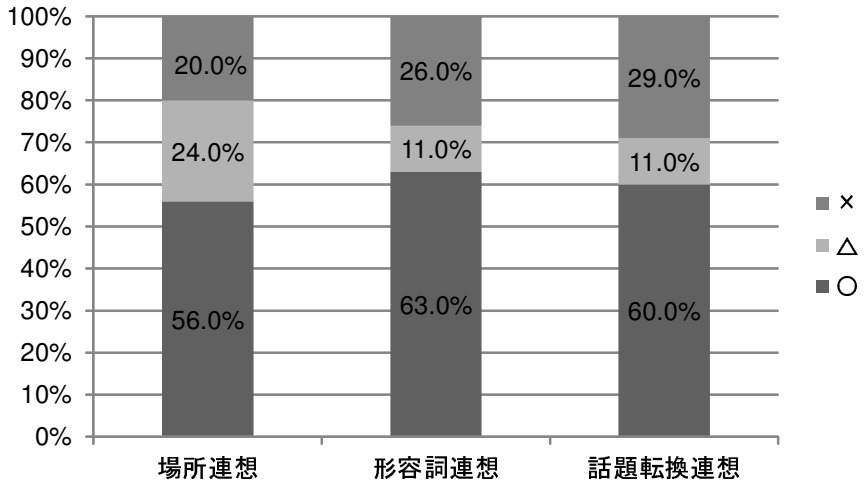


図 8.6: 連想応答評価結果

「○」と「△」を合わせた割合を精度とすると，場所連想は 80%，形容詞連想は 74%，話題転換連想は 71%の精度となった．

提案手法では，名詞と動詞または名詞と形容語の組み合わせ候補のうち，人間の発話中の語と最も関連度の高い語の組で応答の生成を行っている．そこで場所連想の処理過程を例に，関連度が最も高い語の組を第 1 位とし，降順に第 2 位から第 7 位の語の組において評価を行うことで，関連度による選択が正しいかを調査した．結果を図 8.7 に示す．

図 8.7 を見ると，関連度が低い組み合わせにより生成された応答は○，△の評価ともに精度が下がっている．関連度が高い 1 位の語の組み合わせが最も良い評価を示しており，これにより人間の発話中の語と関連度が高い語の組であるほど適切な応答を生成できていることが分かる．

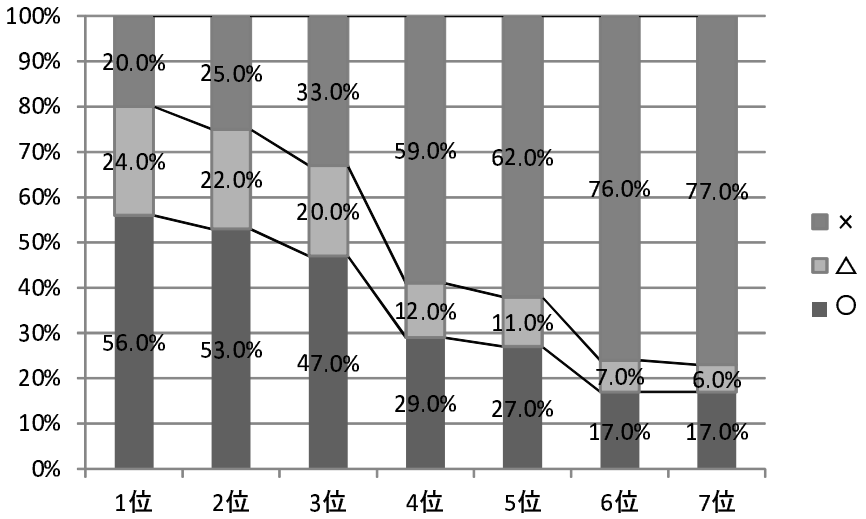


図 8.7: 関連度の高さ順による組み合わせ評価

場所連想の成功例，失敗例を表 8.11 に示す.

表 8.11: 場所連想の成功例と失敗例

	入力文	出力文
成功例	カジノへ 行きました	ポーカーで 遊びましたか？
失敗例	図書館へ 行きました	蔵書を 閲覧してきたのですか？
	教会を 訪れました	聖書を 信仰してきたのですか？

失敗例として,「図書館へ行きました」という入力文において，場所語「図書館」から得られる場所主体語に適切な語「本」があるにも関わらず,「図書館」との関連度が最も高い「蔵書」が選択された．そのため，雑談のような会話ではあまり用いられない「蔵書」を用いた応答文が生成されてしまい，不自然な文と感じさせる結果となった．また,「教会を訪れました」という入力文においては，連想によって,「聖書」や「信仰する」を導いたものの，教会で行う具体的な行動に結びつかず，それらを用いた適切な応答が作成できなかった．このように，場所に存在する物（場所主体語）や行動（場所目的語）を組み合わせるだけでは適切な応答ができない入力文が存在することがわかった.

形容詞連想の成功例，失敗例を表 8.12 に示す.

表 8.12: 形容詞連想の成功例と失敗例

	入力文	出力文
成功例	動物園で 象を見ました	象は 大きいですよね
失敗例	洞窟で コウモリを見ました	コウモリは 有名ですよね
	喫茶店で 珈琲を飲みました	珈琲は 黒かったですか？

形容詞連想では，形容する語を取得する方法として感覚判断システムと大規模格フレームシステムを複合的に用いたが，対象とする単語によっては適切な形容語をほとんど取得できない場合が存在した．例えば,「洞窟でコウモリを見ました」という入力文に対して，対象とする単語「コウモリ」から取得した形容語には適切な形容語がほとんど存在しておらず，その中で最も関連度が高かった「有名」が選択された．関連度に閾値を設けることで，適切な形容語が存在しない場合は連想を行わないという処理が必要になると考えられる．

「喫茶店でコーヒーを飲みました」の例では,「コーヒー」から得られる形容詞中に「美味しい」が取得できたにも関わらず，関連度が最も高い「黒い」が取得されてしまった．事物にとって当然の事柄に関する形容詞，例えば色や形を表すような語は会話において非常に優先度が低いと考えられ，応答候補としての優先度を低くする必要があると考えられる．

話題転換連想の成功例，失敗例を表 8.13 に示す．

表 8.13: 話題転換連想の成功例と失敗例

	入力文	出力文
成功例	喫茶店でケーキを 食べました	紅茶を 飲みましたか？
失敗例	教会で聖書を 読みました	十字架を 信仰しましたか？
	劇場で映画を 見ました	証明を 鑑賞しましたか？

「教会で聖書を読みました」という入力文に対して「十字架」を連想することができたものの,「十字架」に対する適切な行動が場所目的語に存在せず，それらを用いた適切な応答が作成できなかった．このように，場所に存在する物（場所主体語）や行動（場所目的語）を組み合わせるだけでは適切な応答ができない入力文が存在することがわかった．

8.5　おわりに

本章ではロボットによる知的会話について考察し，タスクや大規模な用例コーパスに依存せず，相手の発話を受けてそれに適する自律的な応答を生成する手法について述べた．語概念連想システムにより提供される語の連想機能および，それを基盤として構築される常識判断システムを会話システムに活用することでより豊かな応答を行う連想応答についての提案を行った．提案手法に基づいて構築したシステムにより検証を行った結果，場所連想は 80%，形容詞連想は 74%，話題転換連想は 71%の精度という評価を得た．提案手法により得られる常識的な応答文をロボットに発話させることで，機械が会話の内容を理解していると人間に対して感じさせると期待できる．

第9章　結論

本論文では，人間らしい連想を機械上に実現するための語概念連想システムの構築と，それを用いた応用技術の提案を目的として，概念ベースの構築・精錬手法とその精度評価および複数語概念連想システム，新聞記事見出し文の意味具体化手法，新聞記事中の難解語変換，そして知的会話における連想応答の生成手法について述べた．

第 2 章では自然言語処理における人間の連想を表現する機構の必要性およびその構築の困難さについて述べた．その上で，連想を機械上に表現することを目的とした語概念連想システムと，それを構成する概念ベースおよび関連度計算方式の構造について示した．人間が持つ語の意味とは，語そのものの語彙的な意味だけではなく，他の様々な関係性を内包した概念的なものである．この概念的な語の意味により，人間は様々な事物・事象に対して単なる語彙的な意味の近さだけではない関連の有無を見出すことができる．しかしこの関係性は非常に曖昧であり，シソーラスや意味ネットワーク，オントロジーのような構造による定義は困難である．そこで概念ベースでは，自然言語で表現される様々な「語」に対して人間が持つ意味を,「語」から連想できる他の「語」の集合を属性として付与することで概念化し，定義する．語概念連想システムではこの概念ベースを活用し，ある語を入力するとその語から人間が連想できるであろう他の語を出力する．また，２つの語を入力することで，語間の関連を定量的に表現した関連度という値を出力する．

第 3 章では概念ベースの各種構築手法および精練手法について述べた．国語辞書の見出し語を概念，語義文中の語を属性とした基本概念ベースの構築では，人手によるサンプル概念の評価結果を用いた属性への重み付与を行い，結果としてシソーラス距離による評価結果と比べて高い精度を得た．精度は順序正解率 59.1%，C 平均順序正解率 49.0%となった．この基本概念ベースに対してルールによる属性選別を行い，属性の約 4 割を不適合として削除したルール精練概念ベースでは，順序正解率 63.5%，C 平均順序正解率 56.8%を得た．しかしこれらの概念ベースは概念数が約 3 万語強と，一般的な国語辞書の語彙量と比べて少ない．そこで新聞記事における語と語の共起を利用することで，新たな概念及び属性の追加を行った新聞概念ベースの構築について述べた．新聞概念ベースではルール精練概念ベースに用いたルールを利用して属性の選別を行う．さらに，概念ベースの構造を利用した関連度と概念ベース *idf* による重み付け手法により，人手によるサンプル概念の評価結果を用いない概念ベースの構築手法が確立された．精度は順序正解率 88.8%，C 平均順序正解率 81.0%と大きく向上し，知識の拡充が行えることを示した．さらにシソーラスを用いた属性の追加手法について述べた．シソーラスは人手で作成された大規模な知識ベースであり，信頼性の高い属性を得ることができる．概念に対して新しい属性および重みの付与を行った結果，順序正解率 89.8%，C 平均順序正解率 83.6%の概念ベース構築が行われた．

第 4 章では前章までで述べた概念ベースに対して，さらに属性を追加するための手法を提案した．概念ベースの二次属性および Web 上の情報から概念に追加する属性候補を取得した．これらの候補の選別を行うために，概念ベース *idf*，概念と属性の関連度，属性の重みそれぞれに閾値を定めて多数の実験を行い，精度の変化を調査した．また，二次属性と Web の各々から得られる属性を統合して追加するための閾値に関しても実験を行った．結果として二次属性の追加には概念ベース *idf*，Web からの属性追加には重みを閾値として利用することで精度が向上した．それぞれの属性候補取得の情報元と各種閾値の関係を調べるため，概念ベース *idf* と重みの分布を調査した．結果，二次属性は概念ベース内から均等に抽出されるため，概念ベース *idf* の分布が広い一方，Web から得られた属性候補の概念ベース *idf* は値が小さく，閾値の変動により追加属性数が大幅に増加する傾向が見られた．逆に重みは関連度を用いて算出するため，二次属性からの属性候補に大きな重みが付きやすく，閾値として適さない．最終的な属性追加後の精度として C 平均順序正解率 85.6%となり，シソーラスから属性を追加した概念ベースと比べて 2.0%の精度向上を得た．

第 5 章では語概念連想システムの拡張として，ある語から連想できる他の語を想起する処理を複数語に対応させたシステムについて述べた．与えられた複数語から，それぞれの語に関連する関連語を取得し，そのうち共通する語を連想語として出力する手法を構築した．関連語の取得では概念ベースの連鎖的構造および同義語・類義語を利用することで様々な語を連想語候補として得ることができた．これらの連想語候補から関連度計算方式を用いて適切な連想語を選出した．結果として，出力された連想語の精度 61.0%，擬似再現率 77.0%を得た．

第 6 章では語概念連想システムを自然言語を対象とした情報処理技術に応用する事例として，ロボットとの知的会話を視野に入れた新聞記事見出し文の意味具体化手法について述べた．人間が行う自然な会話の 1 つとして新聞やテレビから得られる時事情報を話題とした会話について考え，新聞の見出し文をロボットの会話リソースとして用いることを考えた．この時，見出し文には体言止めや助詞の欠落といった特有の書式が多く表れ，また端的さゆえに具体的な情報が往々にして省略されている．そこで見出し文を特有の書式から会話に適した表現へと変換した上で，記事本文による意味の具体化を行う手法を提案した．結果として見出し文 120 文の内，58.3%の見出し文について意味の具体化を行うことができた．

第 7 章では自然言語処理の分野で活発に議論される「言い換え」や「変換」処理に語概念連想システムを応用する手法を述べた．ロボットが一般的な人間が行うような会話能力を提供するための一端として，新聞記事を会話リソースになり得る難易度へ変換するための処理を提案した．具体的には新聞記事中に出現する，会話には適さない難解な語を単語親密度により判別し，それらの語句を会話に適した難易度の語および文に変換する．難解語と会話に適する語との閾値は，話し言葉コーパス中に出現する単語と新聞記事中の単語それぞれの単語親密度の分布から導出した．提案手法において，一単語を他の一単語に置き換える変換を一語変換と呼び，ここでは関連度計算方式により難解語と変換の候補となる語との関連度を算出することで適した語の選別を行った．一方，文による変換を N 語変換と呼び，ここでは難解語の用いられている文と変換に用いる文との関連性を，語概念連想システムを活用した EMD を用いた記事関連度計算方式により定量化することで適切な変換を行った．結果として一語変換，N 語変換を組み合わせた提案手法では，意味を保持できているかの評価において 81.1%の精度，平易性を持

たせているかの評価において 75.7%の精度で新聞記事中の難解語の変換を行った．

第 8 章では従来のタスク型の会話システムや膨大な対話例のコーパスに依存した応答生成とは違い，語概念連想システムによる語の連想機能およびそれを基盤として構築される常識判断システムを活用することで，人間の発話文からそれに適した自律的な応答を生成する手法について述べた．提案手法では人間の発話中の場所に関する情報を起点として，場所での行動を連想して応答する「場所連想」，場所に存在する人や物への一般的な感覚による共感を応答する「形容詞連想」，人間の発話内容から連想できる他の話題を応答する「話題転換連想」の三つの処理を行い，人間が常識的と感じる応答を生成した．結果，場所連想は 80%，形容詞連想は 74%，話題転換連想は 71%の精度という評価を得た．

本論文では自然言語処理において一般的な手法，例えばシソーラス，意味ネットワークのような明確な関係性の定義に基づいた手法やオントロジーのような事物・事象を概念化するための構造を定義する手法とは違った視点を持って人間らしい言葉の意味の捉え方について考察し，またその考察に沿った技術である語概念連想システムの構築とその応用について述べた．

自然言語は「人間が話す言葉」であり，それを処理する上で重要なことは「人間らしい」ことであると考える．人間は明確に名づけられる関係性もなく，また語を構成する文字同士がどれぐらいの確率で共起するかを意識しなくても，言葉同士の関連を見出すことができる．もちろん，確率に基づいた言語処理や語間の明確な関係性の定義を構築する技術は自然言語処理において重要であり，またその有効性も様々に示されている．それらに加えて，語概念連想システムが提唱する「人間らしい曖昧性を表現する処理」が，今後の自然言語処理における重要な要素と成りえる事を示した．

謝辞

本論文は，私が同志社大学大学院 工学研究科 知識情報処理研究室に在籍していた期間に行った研究をまとめたものである．本研究を遂行するにあたり，多くの方々に多大なるご指導，ご支援を賜りましたことを心から感謝いたします．

河岡司先生には，大学への編入学のきっかけとなってくださったことをまず感謝いたします．先生が同志社大学で行っておられた研究内容や研究室の紹介を見たことが，私の人生の大きな転機となりました．研究室に配属されてからは，研究内容への細やかなご指導は勿論のこと，研究室での生活にも様々な面で助けていただき，また張り合いを与えてくださいましたこと，感謝いたします．研究室に配属されてすぐに「君は代表で概念班だ」と笑顔でご指示いただいたことは未だに忘れられず，様々な場面で話の種となっております事をお許しください．本当に楽しい１年間でした．有難うございました．

渡部広一先生には，本大学への編入当初から様々な面で大変お世話になりました．編入学試験の面接官が先生であられた事は，今から思うとなにやら縁のようなものを感じている次第です．研究室に配属されるまでは至極真面目なお方なのだろうと思い，先生と対するときは気を引き締めておりました．もちろん，今も先生を前にするときには失礼の無いように気をつけているつもりですが，当時と比べると若干の気の緩みは否めませんこと，お許しください．また，先生との珈琲を飲みながらの談義は私の研究室生活において非常に重要なものでした．時には研究の話，時には馬の話，時には電子遊戯の話と実に楽しい時間をすごさせていただいたこと，感謝いたします．あまり面白おかしい事ばかりではなく，真面目な謝辞も述べさせていただきたく思います．渡部広一教授からいただいたご指示，ご指導はいつも問題の核心を逃さぬもので，感嘆するばかりでした．疑問を曖昧にしない姿勢や問題の本質を見る視点を，少しでも吸収できればと学んでまいりました．これからも先生を目標として精進いたします．

土屋誠司先生には，大学院からご指導いただき，そして私が博士課程に進むきっかけを頂きましたこと感謝いたします．大学院の１年目から，研究，実験，研究室運営と様々な張り合いのある事柄を与えていただきました．また，種々の酒席で楽しい時間を過ごさせていただきましたことも，感謝いたします．銀杏の美味しさに目覚めたのは先生のお陰です．私も先生のように，楽しい酒席を作れるような人物を目指していきたいと思います．珈琲文化を研究室に浸透させることが出来たのは，先生からお貸しいただいた道具類のお陰であることも，重ねて感謝いたします．では，真面目な謝辞を述べさせていただきたく思います．土屋誠司准教授がご指導くださった研究発表や論文執筆，学生指導の心得は，実に的確で分かりやすく，感謝の念に堪えません．先生の広い洞察力や行動力を私も持てるよう，頑張ってまいります．

下原勝憲先生には，博士後期過程におけるセミナーを担当していただき，また公聴会におい

てご質問・ご感想をいただいたこと感謝いたします．院生になり先生の授業を始めて受けたとき，なんと優しげな方だろうと感じました．事実，急なお願いであったセミナー担当を快諾くださり，研究について丁寧に話を聞いてくださり，一対一でのセミナー発表にすら笑顔でお付き合いくださいました．心より感謝いたします．ありがとうございました．

吉村枝里子先輩には，研究室生活において様々な面でお世話になりました．研究や後輩指導に悩んだときには相談に乗っていただきました．愚痴めいた話も聞いていただきました．作業に行き詰ったときには気分転換の雑談に長々とお付き合いいただきました．昼食時には私のくだらない，実の無い話にも付き合ってくださいました．このように羅列いたしますと，本当に私が碌なことをしていない気がいたしますが，先輩にとっても楽しい時間が少しでもあったなら幸いです．有難うございました．

知識情報処理研究室の先輩，同期，後輩の諸氏には本当にお世話になり，またお世話をしたと自負している方もちらほらと居られます．どちらにせよ，私の研究室生活が満ち足りたものであったのは皆様のお陰です．本当に有難うございました．数少ない同期の奥田君，洞井君，山村君の三名は，様々な意味で特に印象深く残っております．人数が少ないにも関わらず，あまり群れることの無かった，しかし何故かやるときには団結できた不思議な関係が非常に心地よかったです．

最後に，両親に対して心からの感謝を述べさせていただきます．算数と理科が苦手で，国語が得意だった一人娘は，理系の高校へ行き，工学部へ編入し，博士課程にまで進みました．好きなことをすればよいと，私の人生を私に任せてくれた事を感謝いたします．情報系の技術者として父を尊敬いたしております．日々の家事を見事にこなしながら仕事を手伝う母を尊敬いたしております．長い学生生活を様々な面で支えてくれた事を感謝いたします．有難うございました．

関連図書

[1] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦：日本語語彙大系, 岩波書店 (1997).

[2] Collins, A. M. and Quillian, M. R.: Retrieval Time from Semantic Memory, *Journal of Verbal Learning and Verbal Behavior*, Vol. 8, No. 2, pp. 240–248 (1969).

[3] Salton, G. M., Wong, A. and Yang, C.: A Vector space model for automatic indexing, *Communications of the ACM*, Vol. 18, No. 3, pp. 613–6201 (1975).

[4] Rychener, M. D.: Control requirements for the design of production system architectures, *Proc. of the symposium on Artificial intelligence and programming languages*, pp. 37–44 (1977).

[5] 溝口理一郎, 人工知能学会：オントロジー工学 (知の科学), オーム社 (2005).

[6] 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).

[7] 小島一秀, 渡部広一, 河岡司：連想システムのための概念ベース構成法– 属性信頼度の考え方に基づく属性重みの決定, 自然言語処理, Vol. 9, No. 5, pp. 93–110 (2002).

[8] 北川晋也, 奥村紀之, 渡部広一, 河岡司：シソーラスの分類情報を利用した概念ベースの属性追加手法, 情報処理学会第 68 回全国大会講演論文集 4N-5 (2006).

[9] 荒木孝允, 奥村紀之, 渡部広一, 河岡司：比較対象概念の共通属性を重視する動的関連度計算方式, 同志社大学理工学研究報告, Vol. 48, No. 3 (2007).

[10] 渡部広一, 奥村紀之, 河岡司：概念の意味属性と共起情報を用いた関連度計算方式, 自然言語処理, Vol. 13, No. 1, pp. 53–74 (2006).

[11] 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).

[12] 奥村紀之, 土屋誠司, 渡部広一, 河岡司：概念間の関連度計算のための大規模概念ベースの構築, 自然言語処理, Vol. 14, No. 5, pp. 41–64 (2007).

[13] 大野晋, 浜西正人：類語国語辞典 4th edition, 角川書店 (1990).

[14] 武部良明（編）: 必携類語実用辞典, 三省堂 (1977).

[15] 三省堂編修所（編）: 必携用事用語辞典 4th edition, 三省堂 (1992).

[16] 見坊豪紀 : 三省堂現代国語辞典 2nd edition, 三省堂 (1992).

[17] 松村明, 三省堂編修所（編）: 大辞林, 三省堂 (1992).

[18] 新村出（編）: 広辞苑, 岩波書店 (1992).

[19] 長尾真 : 岩波講座ソフトウェア科学 15 自然言語処理, 岩波書店 (1996).

[20] 眞鍋康人, 小島一秀, 渡部広一, 河岡司 : 概念間の関連度やシソーラスを用いた概念ベースの自動精錬手法, 同志社大学理工学研究報告, Vol. 42, No. 1, pp. 9–20 (2001).

[21] 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錬手法, 信学技報 NLC2001-93, pp. 109–116 (2002).

[22] 橋本隆志, 渡部広一, 河岡司 : 新聞記事等の文書を用いた概念自動学習による概念ベース構築方式, 情報処理学会自然言語処理研究会資料 2000-NL-148-13, pp. 89–96 (2002).

[23] 芋野美紗子, 吉村枝里子, 土屋誠司, 渡部広一 : Web および二次属性を用いた属性追加手法の提案, 情報科学技術フォーラム FIT2011 F-007, pp. 393–396 (2011).

[24] Imono, M., Yoshimura, E., Tsuchiya, S. and Watabe, H.: Method to add new attributes to concepts by Web and second-order attributes, *Proc. of ICAI2013 Vol.I*, pp. 131–135 (2013).

[25] 辻泰希, 渡部広一, 河岡司 : www を用いた概念ベースにない新概念およびその属性獲得手法, 第 18 回人工知能学会全国大会論文集 2D1-01, pp. 1–4 (2004).

[26] 工藤拓, 松本裕治 : チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002).

[27] 河原大輔, 黒橋禎夫 : 高性能計算環境を用いた Web からの大規模格フレーム構築, 情報処理学会　自然言語処理研究会 171-12, pp. 67–73 (2006).

[28] 天野成昭, 近藤公久 : NTT データベースシリーズ日本語の語彙特性（第 1 期 CD-ROM 版）, 三省堂 (1999).

[29] 鍜治伸裕, 黒橋禎夫, 佐藤理史 : 国語辞典に基づく平易文へのパラフレーズ, 情報処理学会自然言語処理研究会 NL-144-23, Vol. 2001, No. 69, pp. 167–174 (2001).

[30] 熊本忠彦, 田中克己 : ２種類の共起辞書を用いた語彙的言い換えに基づく Web 検索システム, 人工知能学会論文誌, Vol. 23, No. 5, pp. 355–363 (2008).

[31] 西村健二, 田中成典, 北野光一, 田中裕一, 大林睦：児童向け新聞教材のための言い換え表現対の抽出に関する研究, 情報処理学会第 71 回全国大会, pp. 293–294 (2009).

[32] 中野智子, 遠藤淳, 菅原昌平, 乾健太郎, 藤田篤：Web サイトへのアクセシビリティ向上を目的とした難語の平易化, 信学技報 WIT2005-25, pp. 11–14 (2005).

[33] 藤沢仁子, 相原健郎, 神門典子：文化遺産に関する説明文の対象ユーザに合わせた言い換えの提案, 信学技報 NLC2006-2, pp. 7–12 (2006).

[34] 鍜治伸裕, 岡本雅史, 黒橋禎夫：WWW を用いた書き言葉特有語彙から話し言葉語彙への用言の言い換え, 自然言語処理, Vol. 11, No. 5, pp. 19–37 (2004).

[35] McCarthy, D. and Navigli, R.: SemEval-2007 task 10: English lexical substitution task, *Proc. of the 4th International Workshop on Semantic Evaluations*, pp. 48–53 (2007).

[36] Specia, L., Jauhar, S. K. and Mihalcea, R.: SemEval-2012 task 1: English Lexical Simplification, *Proc. of the First Joint Conference on Lexical and Computational Semantics*, pp. 347–355 (2012).

[37] Hassan, S., Csomai, A., Banea, C., Sinha, R. and Mihalcea, R.: UNT: SubFinder: combining knowledge sources for automatic lexical substitution, *Proc. of the 4th International Workshop on Semantic Evaluations*, pp. 410–413 (2007).

[38] Jauhar, S. K. and Specia, L.: UOW-SHEF: SimpLex: lexical simplicity ranking based on contextual and psycholinguistic features, *Proc. of the First Joint Conference on Lexical and Computational Semantics*, pp. 477–481 (2012).

[39] Sinha, R.: UNT-SimpRank: systems for lexical simplification ranking, *Proc. of the First Joint Conference on Lexical and Computational Semantics*, pp. 493–496 (2012).

[40] Wilson, M.: The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2., *Behavior Research Methods*, Vol. 20, No. 1, pp. 6–11 (1988).

[41] Gonzalez, H. S. and Davis, C.: The bristol norms for age of acquisition, imageability, and familiarity., *Behavior Research Methods*, Vol. 38, No. 4, pp. 598–605 (2006).

[42] 天野成昭, 近藤公久：音声単語の語彙判断に対する新密度の影響, 日本音響学会秋季研究発表会講演論文集 1, pp. 363–364 (1998).

[43] Cynthia, C. M., John, M., Eve, S. and Jennifer, Y.: Word familiarity and frequency in visual and auditory word recognition, *Journal of Experimental Psychology. Leaning, memory and cognition*, Vol. 16, No. 6, pp. 1084–1096 (1990).

[44] 天野成昭, 近藤公久：日本語の語彙特性 第 1 巻 単語新密度 増補, 三省堂 (2008).

[45] 小島一秀, 渡部広一, 河岡司：常識判断のための概念ベース構成法：概念間論理関係を用いた概念属性の重み決定法, 信学技報 AI2000-80, pp. 57–64 (2001).

[46] 篠原宜道, 渡部広一, 河岡司：常識判断に基づく会話意味理解方式, 言語処理学会第 8 回年次大会発表論文集, pp. 651–654 (2002).

[47] Hoffman, A. J.: On simple linear programing problems, *Proc. of Symposia in Pure Mathematics*, pp. 317–327 (1963).

[48] 藤江悠五, 渡部広一, 河岡司：概念ベースと Earth Mover's Distance を用いた文書検索, 自然言語処理, Vol. 16, No. 3, pp. 25–49 (2009).

[49] 奥村紀之, 小島一秀, 渡部広一, 河岡司：電子化新聞を用いた概念ベースの拡張と属性重み付与方式, 情報処理学会研究報告, pp. 55–62 (2005).

[50] 国立国語研究所：日本語話し言葉コーパスの構築法, 国立国語研究所 (2006).

[51] 朝日新聞社：asahi.com.

[52] 飯田善久, 梅津圭介：アスペクト指向プログラミングによる旅館予約システムの開発, 成蹊大学理工学研究報告, Vol. 43, No. 1, pp. 9–15 (2006).

[53] 谷垣宏一, チャクラボルティゴウタム, 白鳥則郎：自然言語インタフェースに基づいた電子秘書システムの構成, 情報処理学会研究報告, マルチメディア通信と分散処理研究報告 95(22), pp. 73–78 (1995).

[54] 木村泰知, 荒木健治, 桃内佳雄, 栃内香次：遺伝的アルゴリズムを用いた帰納的学習による音声対話処理手法, 電子情報通信学会論文誌 Vol.J84-D2, No. 9, pp. 2079–2091 (2001).

[55] 杉本二郎, 渡部広一, 河岡司：概念ベースを用いた常識場所判断システムの構築, 情報処理学会自然言語処理研究会資料 2003-NL-153, pp. 81–88 (2003).

[56] 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, 人工知能学会論文誌, Vol. 19, No. 2, pp. 73–82 (2004).

[57] *NEW HORIZON English Course 1,2,3*, 東京書籍 (2005).

[58] *NEW CROWN ENGLISH SERIES 1,2,3*, 三省堂 (2005).

[59] *SUNSHINE ENGLISH COURSE 1,2,3*, 開隆堂 (2005).

[60] *TOTAL ENGLISH 1,2,3*, 学校図書 (2005).

[61] *ONE WORLD English Course 1,2,3*, 教育出版 (2005).

[62] *COLUMBUS 21 ENGLISH COURSE 1,2,3*, 光村図書 (2005).

研究業績一覧

番号	題名	年月	発表した方法	著者
査読付き論文				
1	An Intelligent Method for Retrieval of Verbal Terms from the Web as Answers in Response to Complex Interrogative Sentences	2011. 7	Proc. of ICAI2011, Vol. I, pp.326-331	Hirokazu Watabe Misako Imono Eriko Yoshimura Seiji Tsuchiya
2	Emotion Judgment Method from a Meaning of an Utterance Sentence	2011. 9	KES2011,LNAI6881, pp.367-376	Seiji Tsuchiya Misako Imono Eriko Yoshimura Hirokazu Watabe
3	Necessary Tools Choice in a Particular Situation for Computer Conversation	2011. 9	KES2011,LNAI6881, pp.474-483	Eriko Yoshimura Misako Imono Seiji Tsuchiya Hirokazu Watabe
4	Emotion judgment method from a user utterance sentence	2012. 1	International Journal of Knowledge-Based and Intelligent Engineering Systems, Vol.16, No.1, pp.11-16	Seiji Tsuchiya Misako Imono Eriko Yoshimura Hirokazu Watabe

番号	題名	年月	発表した方法	著者
5	Automatic Detection of Illogical Adjective Phrase Based on Commonsense for Computer Conversation	2012. 1	International Journal of Knowledge-Based and Intelligent Engineering Systems, Vol.16, No.1, pp.3-10	Eriko Yoshimura Misako Imono Seiji Tsuchiya Hirokazu Watabe
6	Topic Word Extraction Using World Wide Web Search Rankings for Computer Conversations	2012. 7	Proc. of ICAI2012, Vol.I, pp.165-171	Eriko Yoshimura Misako Imono Seiji Tsuchiya Hirokazu Watabe
7	Calculating Degree of Association Incorporating Viewpoint Using a Concept-Base	2012. 7	Proc. of ICAI2012, Vol.I, pp.191-197	Hirokazu Watabe Misako Imono Eriko Yoshimura Seiji Tsuchiya
8	A Method for Generating Association Words from Several Other Words in an Association System	2012. 7	Proc. of ICAI2012, Vol.II, pp.730-734	Misako Imono Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe
9	Meaning Judgment Method for Alphabet Abbreviation Using the Association Mechanism	2012. 9	Proc. of KES2012, pp.209-218	Seiji Tsuchiya Misako Imono Eriko Yoshimura Hirokazu Watabe
10	The Degree of Association between Concepts Focusing on the Viewpoint	2012. 12	Computer Technology and Application, Vol.3, No.12, pp.801-807	Misako Imono Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

番号	題名	年月	発表した方法	著者
11	知的会話処理における連想応答手法	2013. 1	人工知能学会論文誌 SP-A, Vol.28, No.2, pp.100-111	吉村枝里子 芋野美紗子 土屋誠司 渡部広一
12	Proposing a Method of Generating Association Words from Multiple Words Based on Association System	2013. 4	Journal of Communication and Computer, Vol.10, No.4, pp.468-473	Misako Imono Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe
13	Inference of the Day Topic Word Using WWW Search Rankings for Computer Conversations	2013. 4	Journal of Communication and Computer, Vol.10, No.4, pp.513-524	Eriko Yoshimura Misako Imono Seiji Tsuchiya Hirokazu Watabe
14	新聞記事中の難解語を平易な表現へ変換する手法の提案	2013. 6	自然言語処理, Vol.20,No.2,pp.105-132	芋野美紗子 吉村枝里子 土屋誠司 渡部広一
15	Computer-based Method for Association Response in Autonomous Conversation	2013. 7	Proc. of ICAI2013, Vol.I, pp.118-123	Eriko Yoshimura Misako Imono Seiji Tsuchiya Hirokazu Watabe
16	Method to add new arrtibutes to concepts by Web and second-order attributes	2013. 7	Proc. of ICAI2013, Vol.I, pp.131-135	Misako Imono Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

番号	題名	年月	発表した方法	著者
17	Meaning Judgment Method for Alphabet Abbreviation Using Wikipedia and Earth Mover's Distance	2013. 7	Proc. of ICAI2013, Vol.I, pp.148-153	Seiji Tsuchiya Misako Imono Eriko Yoshimura Hirokazu Watabe
18	Association inference processing to extract knowledge sentence for question answering	2013. 7	Proc. of ICAI2013, Vol.II, pp.741-742	Hirokazu Watabe Misako Imono Eriko Yoshimura Seiji Tsuchiya
19	Method of Embodying the Meaning of Headlines using News Articles	2013. 9	Proc. of KES2013, pp.336-344	Misako Imono Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe
20	口語表現に対応した知識ベースと連想メカニズムによる感情判断手法	採録決定	人工知能学会論文誌, Vol.29, No.1	土屋誠司 鈴木基之 芋野美紗子 吉村枝里子 渡部広一
書籍				
21	Semantic Inference of Unknown Words by Positioning in Thesaurus Based on an Association Mechanism	2013.8	Pattern Recognition Methods and Application Chapter.9, pp.163-174	Eriko Yoshimura Misako Imono Seiji Tsuchiya Hirokazu Watabe

番号	題名	年月	発表した方法	著者
口頭発表（研究会）				
22	概念ベース精錬のための属性追加手法の提案	2011. 3	信学技報 AI2010-58, pp.1-6	芋野美紗子 吉村枝里子 土屋誠司 渡部広一
23	Web ニュース記事データを用いた見出し文の意味的具体化	2012. 3	情報処理学会研究報告 2012-ICS-166, No.1, pp.1-6	稲井聡 芋野美紗子 土屋誠司 渡部広一
24	嗜好に基づく時事情報推薦システムの構築	2013. 3	研究報告知能システム 2013-ICS-170, No.1, pp.1-6	山本達也 芋野美紗子 土屋誠司 渡部広一
25	ユーザの要求に応じたニュース記事の表形式要約システム	2013. 3	研究報告知能システム, 2013-ICS-170, No.9, pp.1-6	西口駿祐 芋野美紗子 土屋誠司 渡部広一
26	単語の共起情報に基づいた音声認識誤り単語の補正手法	2013. 3	信学技報 AI2012-44, Vol.112, No.477, pp.19-24	角地良太 芋野美紗子 土屋誠司 渡部広一
27	多義概念の属性を代表語に利用した概念ベースにおける多義性の解消	2013. 3	人工知能学会知識ベースシステム研究会資料, SIG-KBS-B203-02, pp.9-14	柳瀬秀夫 芋野美紗子 土屋誠司 渡部広一

番号	題名	年月	発表した方法	著者
28	百科事典を用いた概念ベースの構築	2013. 3	人工知能学会知識ベースシステム研究会資料 SIG-KBS-B203-03, pp.15-20	大竹慎吾 芋野美紗子 土屋誠司 渡部広一
29	クラスタリングによるデータ精錬を用いた脳波による感情判断方式	2013. 3	信学技報 AI2012-47, Vol.112, No.477, pp.37-42	泉啓太 芋野美紗子 土屋誠司 渡部広一
口頭発表（全国大会）				
30	重み配分に着目した概念ベースの精練	2009. 9	情報科学技術フォーラム FIT2009 F-062, pp.557-560	芋野美紗子 吉村枝里子 土屋誠司 渡部広一
31	テレビの視聴履歴を基にした時事情報提供システムの構築	2011. 9	情報科学技術フォーラム FIT2011 E-002, pp.199-200	山本達也 芋野美紗子 土屋誠司 渡部広一
32	文章を整理するための表自動生成手法	2011. 9	情報科学技術フォーラム FIT2011 E-006, pp.207-208	西口駿祐 芋野美紗子 土屋誠司 渡部広一
33	話題による音声認識誤り単語の補正手法	2011. 9	情報科学技術フォーラム FIT2011 E-023, pp.251-252	角地良太 芋野美紗子 土屋誠司 渡部広一
34	新聞記事からの複合語概念表記の獲得	2011. 9	情報科学技術フォーラム FIT2011 E-029, pp.269-270	柳瀬秀夫 芋野美紗子 土屋誠司 渡部広一

番号	題名	年月	発表した方法	著者
35	画像を用いた物体の詳細情報認識手法	2011. 9	情報科学技術フォーラム FIT2011 H-016, pp.137-138	八木亮 芋野美紗子 土屋誠司 渡部広一
36	脳波知識ベースを用いた感情判断方式	2011. 9	情報科学技術フォーラム FIT2011 J-050, pp.655-656	泉啓太 芋野美紗子 土屋誠司 渡部広一
37	Web および二次属性を用いた属性追加手法の提案	2011. 9	情報科学技術フォーラム FIT2011 F-007, pp.393-396	芋野美紗子 吉村枝里子 土屋誠司 渡部広一
38	語概念連想を用いた複数単語からの連想語生成手法の提案	2012. 3	言語処理学会第 18 回年次大会発表論文集 C2-1, pp.409-412	芋野美紗子 吉村枝里子 土屋誠司 渡部広一
39	質問文の意味を考慮した Wikipedia からの回答文抽出手法	2012. 9	情報科学技術フォーラム FIT2012 E-012, pp.183-184	森泰宏 芋野美紗子 土屋誠司 渡部広一
40	連想メカニズムを用いた直喩の意味解析法	2012. 9	情報科学技術フォーラム FIT2012 E-018, pp.195-196	鞠山大樹 芋野美紗子 土屋誠司 渡部広一
41	男女の脳波知識ベースを統合した脳波感情判断手法	2012. 9	情報科学技術フォーラム FIT2012 J-006, pp.419-420	森本麻代 芋野美紗子 土屋誠司 渡部広一

番号	題名	年月	発表した方法	著者
42	性別・年代別の嗜好情報を基にした話題語提供システム	2012. 9	情報科学技術フォーラム FIT2012 E-035, pp.239-240	南光 芋野美紗子 土屋誠司 渡部広一
43	Web ニュース記事本文を利用した見出し文の意味具体化手法	2013. 3	言語処理学会第 19 回年次大会発表論文集 A5-5, pp.508-511	芋野美紗子 吉村枝里子 土屋誠司 渡部広一
44	大規模格フレームを用いた概念ベースへの動詞属性の追加	2013. 9	情報科学技術フォーラム FIT2013 E-017, pp.219-220	小泉政弥 芋野美紗子 土屋誠司 渡部広一
45	主成分分析を用いた脳波感情判断システムの構築	2013. 9	情報科学技術フォーラム FIT2013 J-027, pp.437-438	森智洋 芋野美紗子 土屋誠司 渡部広一
46	概念ベースの二次属性を用いた直喩解析法	2013. 9	情報科学技術フォーラム FIT2013 E-018, pp.221-222	丸山礼文 芋野美紗子 土屋誠司 渡部広一
47	概念の多義性を考慮した属性構造化による概念ベースの構築	2013. 9	情報科学技術フォーラム FIT2013 E-019, pp.223-224	小川真路 芋野美紗子 土屋誠司 渡部広一