

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

芋野 美紗子 吉村 枝里子 土屋 誠司 渡部 広一

同志社大学 理工学部, 大学院 工学研究科

{mimono, eyoshimura, tsuchiya, watabe}@indy.doshisha.ac.jp

1 はじめに

近年における情報分野の発展は目覚しく、世にあふれる様々な情報システムはもはや社会生活に無くてはならない存在である。また、ロボットの分野においては単純作業をこなすだけでなく、愛玩や介護といった人間のパートナーとしての存在への期待が高まっている。このような情報システムやパートナーとしてのロボットを人間が違和感なく扱うためには、万人にとって扱いやすいインターフェースが求められる。人間同士のインターフェースはその大半が会話によるコミュニケーションであり、今後はコンピュータやロボットにも人間らしいコミュニケーション能力が求められる。

人間は自然言語により柔軟な会話を行うが、これは人間が経験上持っている語の知識や常識知識、更にはある事柄から別の事柄を思い浮かべる連想能力といったものを駆使して行っている。本稿ではコンピュータに人間らしい連想能力を持たせるための一端として、複数単語からの連想語生成手法の提案を行う。例えば「針」という一語から関連する別の語を考えると、人間は「釣り、縫い物、注射、刺す...」といった語を連想することが出来る。しかし「針、糸」という二語になれば、おそらく「釣り、縫い物」といった語が連想されるだろう。これは「針」と「糸」という語についての知識から、双方に関連がある語のみを連想するためである。更に「針、糸、魚」という三語ならば「釣り」という語になると考えられる。

以上のような連想能力をコンピュータ上に実現するために、本稿ではすでに提案されている概念ベース [1] と関連度計算方式 [2] による語概念連想を用いた。概念ベースは人間が持つ語の知識をモデル化したものであり、これを利用して複数の語から別の関連語を取得する。さらに関連度計算方式によって入力語と連想される語との間の関連性を定量化し、正しい連想語を選択する。以上の機構を用いて複数単語からの連想語生成手法の提案を行い、コンピュータに人間らしい連想能力を持たせるための研究の一端を示す。

2 語概念連想

2.1 概念ベース

概念ベースは複数の電子化国語辞書などの見出し語を概念と定義し、見出し語の定義文中の自立語群を概念の意味を定義する属性として付与することで構築された知識ベースである。本稿で使用した概念ベースは自動的に概念および属性を構築した後に人間の常識に沿った属性の追加や削除を行ったものであり、87242語の概念が定義されている。

ある概念が持つ属性を、その概念の一次属性と呼ぶ。概念ベースでは属性を成す語も概念として定義されており、つまり属性を概念とみなして更に属性を導くことができる。属性から導かれた属性を、元の概念の二次属性と呼ぶ。概念ベースの具体例を表1に示す。

表 1: 概念ベースの具体例

概念	属性
夏	(夏場,0.34)(夏休み,0.11)(海,0.08)...
夏場	(夏季,0.25)(暑さ,0.18)(太陽,0.04)...
...	...

「夏」という概念が持つ属性「夏場」は、概念としても定義されている。この概念「夏場」の持つ「夏季、暑さ、太陽...」といった属性群が、元の概念「夏」の二次属性ということになる。

2.2 関連度計算方式

関連度計算方式は概念ベースに定義される概念と概念の関連性を定量的に表現する手法であり、その有効性が示されている [3]。以下に、概念 A と概念 B の関連度 $DoA(A, B)$ の算出方法について示す。

概念 A および概念 B の一次属性をそれぞれ a_i , b_i とし、対応する重みを u_i , v_i とする。それぞれが持

つ属性数が L 個と M 個 ($L \leq M$) とすると、概念 A , B はそれぞれ以下になる。

概念 $A = \{(a_1, u_1), (a_2, u_2), \dots, (a_L, u_L)\}$

概念 $B = \{(b_1, v_1), (b_2, v_2), \dots, (b_M, v_M)\}$

ここで一次属性の数が少ない概念 A の属性の並びを固定する。その上で概念 B の各一次属性を対応する概念 A の各一次属性との一致度 $DoM(A, B)$ の合計が最大になるように並べ替える。ただし、概念 A の属性と対応付けされなかった属性については無視する。

概念 $B = \{(b_{x1}, v_{x1}), (b_{x2}, v_{x2}), \dots, (b_{xL}, v_{xL})\}$

このとき、概念 A と概念 B の関連度 $DoA(A, B)$ を、 $DoA(A, B)$

$$= \sum_{i=1}^L DoM(a_i, b_{xi}) \times \frac{(u_i + v_{xi})}{2} \times \frac{\min(u_i, v_{xi})}{\max(u_i, v_{xi})}$$

と定義する。ここで $\min(u_i, v_{xi})$ は u_i と v_{xi} を比較して小さい値を、 $\max(u_i, v_{xi})$ は大きい値を指す。

なお、一致度 $DoM(A, B)$ は以下のように定義する。

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

$a_i = b_j$ は属性が表記的に一致した場合を示す。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、小さいほうの重みを足し合わせたものとなる。

3 連想語生成手法

連想語生成手法は、入力される複数単語から連想される別の語を生成するための手法である。具体的な処理の流れを図 1 に示す。

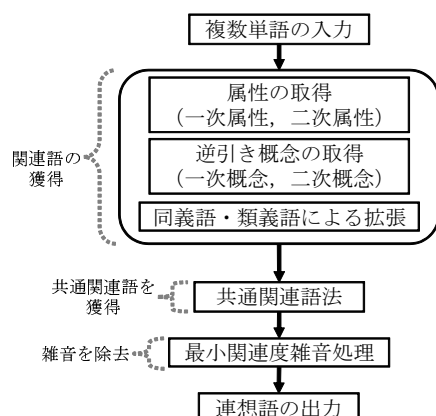


図 1: 連想語生成手法の流れ

入力は複数の単語群とし、語数に制限は無いが概念ベースにおいて定義されている語のみを対象とする。まず入力された各単語と関連のある語（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。

3.1 関連語の獲得

関連語とは入力された複数単語のそれぞれについて、何かしらの関連があると考えられる語である。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。

3.1.1 属性の取得

属性は概念の意味を定義する語であるため、概念と属性の間には何かしらの関連性があると考えられる。そこで属性の取得では入力された複数単語を概念として考え、その属性を関連語として獲得する。

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 2 に示す。

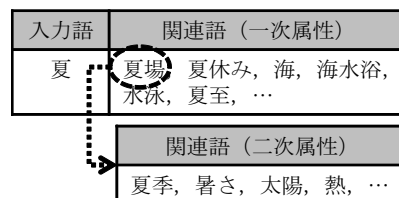


図 2: 入力語「夏」からの属性の取得

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 語と決定し、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。

3.1.2 逆引き概念の取得

ある概念 X について、 X を属性として持つ概念 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ため、これを関連語として取得する。具体例として入力語「夏」からの逆引き概念の取得を図 3 に示す。

入力語	関連語（一次逆引き概念）		
夏	スイカ	競泳	...

概念	一次属性		
スイカ	夏	果物	...
競泳	水泳	夏	...

図 3: 入力語「夏」からの逆引き概念の取得

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る。

なお逆引き概念の取得においても 3.1.1 節と同じく、二次逆引き概念までを関連語として取得するパターンについても処理を行った。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 3.1.1 節で述べたように取得個数の上限を 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。

3.1.3 同義語・類義語による拡張

同義語・類義語による拡張では入力語の同義語・類義語を概念として考え、その一次属性および一次逆引き概念を元の入力語の関連語とする。同義語・類義語については国語辞書から語の同義・類義関係を自動的に取得して作成した関係語辞書を用いる。関係語辞書は見出し語の 1 語に対して同義・類義関係の 1 語が

セットになって登録されており、同義関係が 389 セット、類義関係が 31268 セットとなっている。

3.2 共通関連語法

共通関連語法とは 3.1 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）を取得する手法である。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 4 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	スポーツ、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 、...

図 4: 共通関連語法の具体例

例では入力語「夏、水、運動」のそれぞれが図 4 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。

3.3 最小関連度雑音処理

前節までで得た共通関連語には人間の連想に沿わない、雑音が含まれている可能性がある。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中の雑音の除去を行う。

ある入力語 A 、 B および共通関連語 C があつたとき、 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 に示す。

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

共通関連語	入力語	関連度	共通関連語	入力語	関連度
水泳	夏	0.10	熱	夏	0.04
	水	0.08		水	0.03
	運動	0.06		運動	0.02

最小関連度が0.05以下
⇒共通関連語「熱」は雑音と判断

図 5: 最小関連度雑音処理の具体例

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処理は行われない。「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 3.2 節で得られた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。

4 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連想語となる 1 語と、その語を連想させる入力語を 2 語以上という条件で記述することで作成した。テストセットの入力語から提案手法によって連想語を生成し、精度と再現率による評価を行った。精度は全テストセットから出力される連想語について、評価者 3 名で正解および不正解の判断を行い、3 名中 2 名以上が正解とした語の割合として算出した。再現率はアンケートの際に得た連想語の 1 語が、提案手法により出力された連想語に含まれているかで算出した。

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 6 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

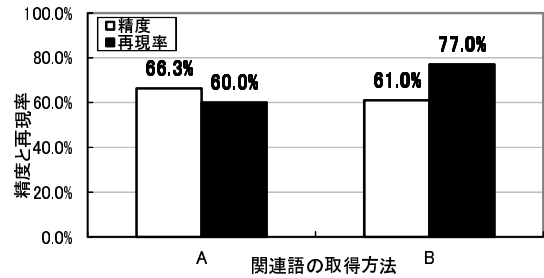


図 6: 精度と再現率

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。

以上のことから本稿における提案手法は B の関連語取得の方法を用いたものとし、精度 61.0%，再現率 77.0% となった。

5 おわりに

本稿ではコンピュータに人間らしい連想能力を持たせるための研究の一端として、概念ベースと関連度計算方式による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、複数単語から連想される語の提示を行った。

本稿の提案手法では精度 61.0%，再現率 77.0% という結果で人間が複数語から自然と連想する語の生成が行われた。これによりコンピュータにおける人間らしい連想機構の一端を示せたと考える。

参考文献

- [1] 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.
- [2] 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.
- [3] 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.