E-050

人名の特徴に基づく人名固有名詞概念の属性獲得手法

The Method of Acquisition of Person's Name Concept and Its Attribute Based on Feature of Personality

河部　匡剛†　　渡部　広一†　　河岡　司†
Masatake Kawabe　Hirokazu Watabe　Tsukasa Kawaoka

1. はじめに

近年，自然言語処理技術は急速に発展してきており，コンピュータとの自然な会話処理においては，人間らしい常識的な判断が求められている．常識的な判断とは，演算を含む会話文の意味理解を行う論理判断[1]，物の大小，長さ，広さ，重さ，場所，時間，速さに関する語の意味理解を行う量的判断[2]，時間に関する語の意味理解を行う時間判断[3]，場所に関する語の意味理解を行う場所判断[4]，また赤い，熱いといった感覚に関する意味理解を行う感覚判断[5]，そして，うれしい，悲しいといった感情を理解する感情判断[5]などである．

これら常識判断を実現するのが常識判断メカニズム[1]である．コンピュータが人間のように柔軟な判断をするにはコンピュータ上で人間の持つ連想処理を実現する必要がある．その連想処理の実現の為に，語の“概念”を概念の意味特徴を表す語（属性）とその重みの集合で定義した概念ベース[6]や，概念間の関連の強さを定量的に評価するための関連度計算方式[7]などで構成する連想メカニズムを利用している．図 1 に常識判断メカニズムと連想メカニズムの全体像を示す．

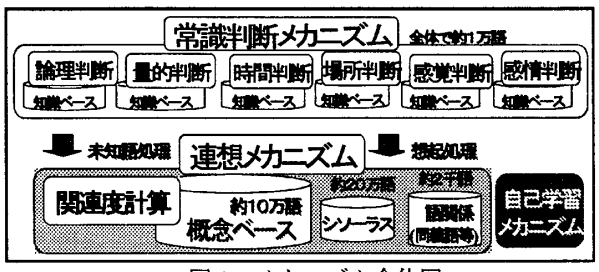

図 1　メカニズム全体図

概念ベースは，電子辞書や電子新聞などから機械的に構築され約 9 万語の概念が定義されている．しかし，日常的なコンピュータ会話において，概念ベースに定義されていない語（未定義語）が出現する可能性がある．未定義語とは，一般的な会話に出現する固有名詞や新語などである．本研究では特に，会話文において重要と考えられる人名固有名詞に関する属性を動的に取得する手法について提案している．

具体的には，ロボット型検索エンジン[8]を用いて未定義語に関する属性群を取得し，概念ベースに動的に定義する手法を提案する．これにより未定義語を概念として扱うことが可能になるため，未定義語も含めた任意の概念間の関連の強さを定量的に扱うことが出来る．

†同志社大学大学院工学研究科
Graduate School of Engineering Doshisha University

2. 連想メカニズム

連想メカニズムは，入力語に対し，関係の深い語を想起する想起語処理と，知識として所持していない語（未知語）を知識として所持している語（既知語）に置き換える未知語処理によって構成されている．想起語処理と未知語処理は概念ベースと関連度計算方式によって実現されている．この章では，概念ベース，概念ベースを用いた関連度計算方式について述べる．

2.1　概念ベース

概念ベースは，見出し語(概念)とその特徴を表す複数の語(属性)と属性の重要さ（重み）を対の組として集めた語の知識ベースである．概念Aは，見出し語Aと属性a_i，重みw_iにより以下のように定義する．

$$A=\{(a_1,w_1),(a_2,w_2),\ldots,(a_i,w_i),\ldots,(a_k,w_k)\} \quad (1)$$

なお，属性となる全ての語は概念ベースの概念として定義されている．そのため，属性からもさらに属性を得ることができる．このとき，はじめの概念より得られた属性を一次属性と呼び，一次属性の属性を二次属性と呼ぶ．このように，各概念は n 次元の属性連鎖集合として定義される(図 2)．

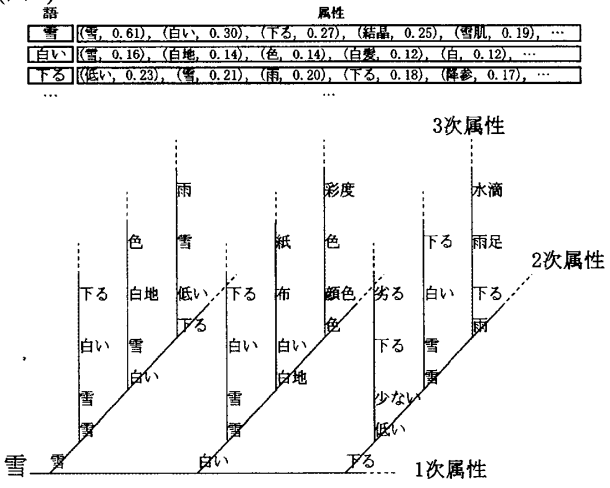

図 2　概念ベース

このように，概念とは，n 次元の属性連鎖集合で定義される属性空間である．また，単に「属性」と表記する場合は，各概念の一次属性を表すものとする．

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た語と語の関連の強さを，語と語の類似性だけでなく，共起など類似性以外の関連性も包含し，定量化する手法である．

関連度は 0 から 1 までの連続値をとり，関連の強い概念同士では高い値となり，関連の弱い概念同士では低い値となる．例えば，概念「医者」と「病院」の関連度は 0.72，概念「医者」と「太陽」の関連度は 0.04 となる．このように概念同士の関連の強さを定量化すれば，数値の大小比較によって，曖昧である概念間の関連性の強さをコンピュータに判断させることができるようになる．

3. Web を用いた未定義概念の属性獲得手法

3.1 オートフィードバック，拡張オートフィードバック

オートフィードバック(以下「AF」)[9]とは，未定義語の属性を，Webを用いて獲得する手法である．ロボット型検索エンジンGoogleを用いて未定義語に対する検索結果ページを取得し，テキスト情報から属性を獲得する．属性はすべて概念ベースに定義されている語である．AFにより獲得される属性の例を表 1 に示す．

表 1　AF により得られる属性

概念	属性 1	属性 2	属性 3	…
クレオパトラ	エジプト	真珠	古代	…
滝廉太郎	作曲	ピアノ	さくら	…

拡張オートフィードバック(以下「拡張 AF」)は,AF では対象としていなかった，未定義の複合語なども属性として含む語を獲得できるように拡張した属性獲得手法である．

- 「括弧の中の語を複合」
- 「名詞の連続は複合」
- 「アルファベットの連続は複合」

という複合語獲得ルールを持ち，AF とは異なる属性が取得される．例を表 2 に示す．*印は，概念ベースにない語である．

表 2　拡張 AF により得られる属性

概念	属性 1	属性 2	…
クレオパトラ	クレオパトラ*	シーザー*	…
滝廉太郎	荒城の月*	作曲	…
アガサ・クリスティー	名探偵ポワロ*	ポワロ*	…

3.2 AF，拡張 AF の問題点

拡張 AF には，複合語獲得ルールが失敗した語や，概念に対しほとんど関係のない語が多く獲得される．具体例を表 3 に示す．

表 3　不適切な属性の例

概念	属性
石川啄木	石川啄木石川啄木
アガサ・クリスティー	セブンアンドワイ

また，AF，拡張 AF では，検索エンジンの結果より得られた広い文章空間から属性を獲得するために，一般的な知識と言える語を含む文章の出現頻度は低くなり，その語が属性として得られにくい．例えば，「滝廉太郎」に対して「作曲家」といった，一般的な知識とも言える語を含む文章は，WWW 検索エンジンの結果から得られたテキスト文書空間内では出現頻度が低くなり，「作曲家」といった語は属性として得られにくい．

拡張 AF により取得した属性の多くは概念ベースに未定義であるため，その属性の属性(二次属性)を取得する際にさらに未定義語が出現するという連鎖が生じる．そのため，未定義語の二次属性を取得する際は AF を用いて未定義語が二次属性に含まれないようにする必要がある．

4. 人名固有名概念の属性獲得手法

AF より得られた属性と，拡張 AF 属性を選別した属性，特徴語を人名固有名詞の属性として取得する(図 3)．特徴語に関しては 4.2 節において述べる．

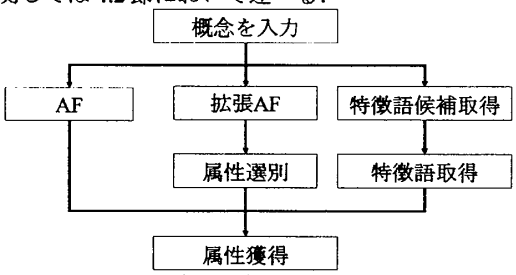

図 3　人名固有名詞概念の属性獲得の流れ

AF より得られる属性に関しては，属性はすべて概念ベースに含まれている語であるため信頼できる語が多いとみなし，選別を行わないものとする．

4.1 拡張 AF により得られる属性の選別

Web 共起率による選別，WebHit 件数による選別の両方の基準を満たす属性のみを獲得する．以下で各選別手法について述べる．

4.1.1 Web 共起率による選別

Web 共起率とは，Web 上における共出現頻度であり，WebHit 件数を基に以下のように定義する．WebHit 件数とは，Google を用いて Web 検索した際に Google が検索結果として取得した Web ページの数と定義する．

$$WebCoHit\ (X,A) = WebHit\ (X\ \&\ A)/WebHit\ (X) \quad (2)$$

WebCoHit(X,A)：語 *X* と語 *A* との Web 共起率

WebHit(X)：語 *X* の WebHit 件数

WebHit(X&A)：語 *X* と語 *A* との And 検索による WebHit 件数

Web 共起率に閾値を設け，値が閾値より低ければ，語と語の関連が弱いとみなし，属性から削除する．Web 共起率の例を表 4 に示す．

表 4　Web 共起率の例

概念 X	属性 A	Web 共起率
アガサ・クリスティー	セブンアンドワイ	0.001
石川啄木	石川啄木石川啄木	0.0005
桑田佳祐	サザンオールスターズ	0.333

4.1.2 WebHit 件数による選別

Google を用いて，属性として得られた語の WebHit 件数の値を取得し，その値によって選別を行う．WebHit 件数が少ない語というのは，Web 上の文章空間において言及さ

れている数が少なく，あまり見かけることがない語ということである．

WebHit 件数に閾値を設け，値が低ければその語は一般的ではない語であるとみなし属性から削除する． WebHit 件数の例を表 5 に示す．

表 5　WebHit 件数の例

入力語	WebHit 件数
アルキメデス	240,000
荒城の月	258,000
石川啄木石川啄木	195

4.2 特徴語獲得

人名に対し，「職業」，「地域」，「時代」，「国」などが特徴としてあげられ，各特徴に必要な語が存在する．これらの語を特徴語と定義する．例えば「手塚治虫」であれば，「漫画家」，「兵庫県」などの一般的な知識である語が特徴語としてあげられる．これらの語は，AF や拡張 AF では獲得されにくい(3.2 節)．そこで，4.2.1 節以下で述べる手法によってこれらを獲得する．

4.2.1 特徴語候補獲得

Wikipedia[10]はWeb上の百科事典である．誰でも閲覧，投稿，編集が自由に行え，多くの項目を保有している．特に説明文の第一段落は見出し語の要約になっている．そのため第一段落より自立語を取得し，特徴語候補とする．取得した例を表 6 に示す．取得した特徴語候補の中に入力概念に対し属性として不必要な語も含まれ，このまま属性として加えることはできないため，選別を行う必要がある．

表 6　特徴語候補の例

概念	候補 1	候補 2	候補 3	…
イチロー	1973 年	西春日井郡	野球選手	…
手塚治虫	日本	医師	兵庫県	…

4.2.2 特徴代表語リスト

職業，地域，時代，国それぞれの代表的な語をリストとして保持しておく．例えば，職業であれば{アナウンサー，税理士，評論家，…}といった語群である．

4.2.3 特徴語獲得の具体的手法

特徴語候補と，特徴代表語リスト内の語すべてとの関連度を計算し，その中で最大となった値が閾値を超えていれば，その候補語を特徴語として採用する．これを特徴語候補ごとに繰り返し，特徴語候補が特徴語として採用できるか判別する(図 4)．

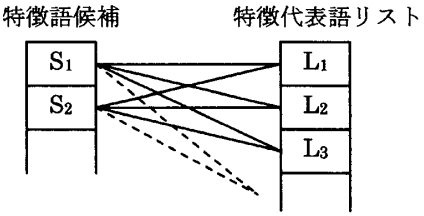

図 4　各特徴語候補と各特徴代表語リストの語との関連度計算

5. 実験・評価

評価データはアンケートより得られた人名固有名詞 100 語を用いる(表 7)．これらの語に対して AF 属性，拡張 AF 属性，特徴語を取得する．AF 属性，拡張 AF 属性により取得する属性は各 30 個とする．

表 7　評価用人名固有名詞概念の例

石川啄木	イチロー	クレオパトラ
手塚治虫	アルキメデス	デビッド・ベッカム
滝廉太郎	平将門	アガサ・クリスティー

評価には精度，再現率を用いる．精度とは獲得した属性中の正解語数の割合であり，再現率とは選別前の正解語数に対する選別後の正解語数の割合である．ここで正解語とは，4 人中 3 人が属性として適切であると目視評価で判断した語である．

5.1 拡張 AF 属性選別の閾値

5.1.1 Web 共起率の閾値

Web 共起率の閾値を変化させながら精度と再現率を調べた(図 5)．グラフ上で精度と再現率が交差する Web 共起率 0.12 を閾値として採用した．

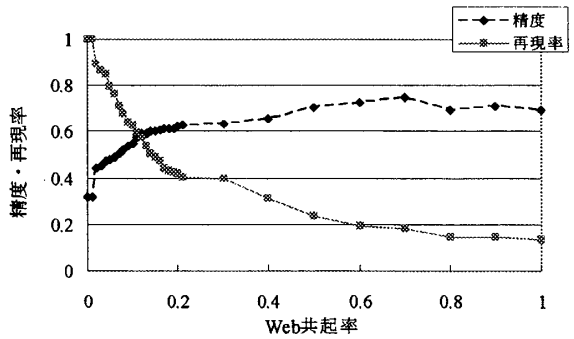

図 5　Web 共起率と精度・再現率

5.1.2 WebHit 件数と Web 共起率による属性選別

Web 共起率を 0.12 で固定し，WebHit 件数を変化させ精度と再現率を調べた(図 6)．

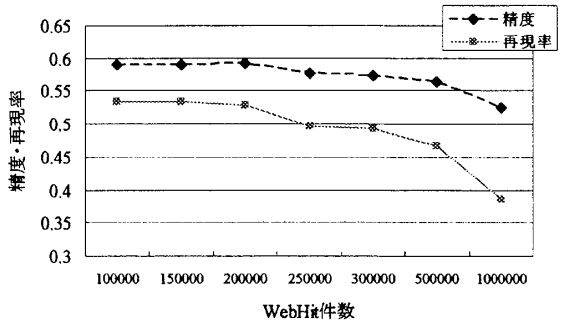

図 6　WebHit 件数と精度・再現率

WebHit 件数の閾値を 200,000 としたときに精度と再現率が高くなり，この値が閾値として適切であると考えられる．また，このとき精度 59.2%，再現率 52.9%，一概念あたりの属性は 8.5 個となった．

5.2 特徴語獲得における特徴語候補と特徴代表語リストとの関連度の閾値

特徴語候補と特徴代表語リストとの関連度をとり，関連度の閾値を変化させながら精度と再現率を調べる(図 5)．

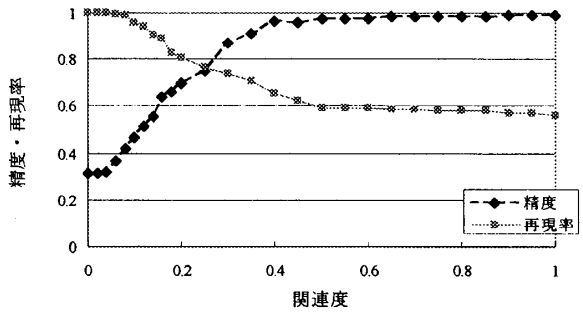

図 5　特徴語候補の関連度と精度・再現率

図 5 において閾値が 1 の時に精度が 1 となるのは，リストの語と特徴語候補とで表記一致した語のみを取得しているからである．閾値を上げることで，取得される属性の個数が減ってゆくため再現率は下がるが，良い属性しか取得されなくなるため精度は上がる．しかし閾値を上げすぎると属性の個数が減りすぎてしまうため，適切な閾値を設定する必要がある．グラフより，精度と再現率が交わる 0.25 の値を採用する．

5.3 手法のまとめ

拡張 AF による属性獲得手法，拡張 AF 属性の選別手法，提案手法により得られた属性のまとめを表 8 に示す．

表 8　既存の手法と提案手法のまとめ

手法	拡張 AF	拡張 AF 選別	特徴語	提案手法
精度	31.5%	59.2%	81.0%	52.6%
再現率	100.0%	52.9%	82.2%	81.3%
属性個数	30	8.5	2.8	41.3

拡張 AF 属性を選別することで，不適切な属性を削除し，拡張 AF 属性の精度を約 28%向上させることができた．

特徴語をあわせた提案手法では，精度 52.6%，再現率 81.3%，一概念あたりの属性数 41.3 個となった．本稿の提案手法により得られた属性の例を表 9 に示す．

表 9　提案手法により得られた属性（概念：手塚治虫）

AF	作品，アニメ，コミック，名作，図版，…
拡張 AF	手塚治虫，火の鳥，鉄腕アトム，講談社，…
特徴語	昭和，日本，漫画家，アニメーター

6. おわりに

本稿では，人名固有名詞の特徴を考慮した属性取得を行った．属性獲得手法の属性選別を行うことにより既存手法と比較して精度が向上した．また，既存手法では獲得されなかった特徴語を獲得することができた．

人名以外の未定義語に対しても本稿の提案手法を応用させることで，特徴を考慮した属性獲得が可能であると考えられる．

AF より得られる属性に関しては，属性はすべて概念ベースに含まれている語であるため信頼できる語が多いとみなし，選別を何も行わなかったが，これらの属性に対しても適切な選別手法を考案し，検討する必要があると考えられる．

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクト「人間と生物の賢さの解明とその応用」における研究の一環として行った．

参考文献

[1] 土屋誠司，小島一秀，渡部広一，河岡司，“ 常識的判断システムにおける未知語処理方式”，人工知能学会論文誌,Vol. 17, No. 6, pp. 667–675，2002.

[2] S. Tsuchiya, H.Watabe and T. Kawaoka, “A Quantitative Judgement System Based on an Association Mechanism for Natural Conversation with Computer”, AIA2007, pp.508-513, 2007 .

[3] 土屋誠司，渡部広一，河岡司，“ 連想メカニズムを用いた時間判断手法の有効性の検証”，信学技報，NLC2005-18，pp.33-38，2005.

[4] 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.

[5] 米谷彩，渡部広一，河岡司，“ 語の共起情報を考慮した感覚連想メカニズムに関する研究”，情報処理学会自然言語処理研究会資料，2005-NL-166，pp.63-70，2005.

[6] 奥村紀之，北川晋也，渡部広一，河岡司，“ 概念ベースの分析と精錬”，同志社大学理工学研究報告，Vol.46，No.3，pp.133-141，2005.

[7] 荒木孝允，渡部広一，河岡司，“共通・類似属性を考慮した概念間関連度計算方式”， 情報処理学会第 68 回全国大会講演論文集，4N-2，2006.

[8] Google：http://www.google.co.jp

[9] 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”，人工知能学会全国大会，2D1-01，2003.

[10] Wikipedia：http://ja.wikipedia.org