Wikipediaを用いた概念ベースへの新概念追加手法

芋野 美紗子　　土屋 誠司　　渡部 広一
大同大学　情報学部
同志社大学　理工学部
{m-imono}@daido-it.ac.jp
{stsuchiy, hwatabe}@mail.doshisha.ac.jp

1　はじめに

人間は語句や文といった自然言語の情報からその意味を柔軟に理解することが出来る．ここでの意味とはいわゆる辞書に記載されている以外の，例えば「蝸牛」と「紫陽花」の間に何かしらの関連性を見出せるといった「飛躍的な意味の理解」も含まれる．

既存の自然言語処理の多くは，意味を理解するためのアプローチとして「明確な関係性の定義」を行っていると考える．たとえば語の意味を上位・下位関係や同義・類義関係を表現する木構造により定義したシソーラスや，現実の事物・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を作成するオントロジーなどが挙げられる．これらの技術によって多くの知識が体系化され活用されているが，前述したような「飛躍的な意味の理解」を行うことは困難である．シソーラスにおける「蝸牛」の上位ノードは「貝」,「紫陽花」の上位ノードは「樹木」であり，これらの間に関連性を見出すのは難しい．人間が理解する「飛躍的な意味」とは,「連想できる」というあいまいな関係性ではないかと考える．そこで「連想できる」という関係性のみに着目して様々な語句の知識を定義したものが概念ベースである．概念ベースでは自然言語で表現される様々な語句（概念）の意味を，その語句から連想できるほかの語句の集合（属性）によって定義する．概念と属性の間には関係性を示すラベルやカテゴリーは存在しない．このような概念ベースの考え方により語句の意味を定義することで，より人間らしい柔軟な意味の理解が可能になるのではないかと考える．

提案されている概念ベースに登録されている概念の数は 87,242 語であり，主に複数の国語辞書から作成されている．本稿では Wikipedia を知識源として概念ベースへ新たな概念を追加し，より人間が持つ知識に足る概念ベースを作成することを目指す．

2　語概念連想システム

語概念連想システムとは，人間のように柔軟な語句の意味理解を行うための機構である．語の意味を定義した「概念ベース」および関連性の定量化を行う「関連度計算方式」により語概念連想システムは構築される．本稿ではこの語概念連想システムの機構を用い，「場所語候補が確かに場所を表す語であるかの判断」および「要望文と場所語との関連性の定量的な判断」を行う．

2.1　概念ベース

概念ベース [1] は複数の電子化国語辞書などの見出し語を概念として定義し，人間が持つ概念への常識的な知識をモデル化した知識ベースである．ある概念の意味定義は，属性と呼ばれる他の概念群と属性それぞれの重要さを表す重みによってなされる．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
蝸牛	(腹足類,1.62)(螺旋,0.71)(梅雨,0.35)…
梅雨	(紫陽花,1.14)(季語,0.49)(蝸牛,0.48)…

概念ベースにおいて属性に現れる語句は，全て概念として定義されている．そのため，概念「蝸牛」の意味定義を行う属性「梅雨」も概念ベースにおいて意味定義がなされている．このような意味定義の連鎖的な繋がりにより，より人間らしい語句の意味定義が可能となる．

2.2 関連度計算方式

関連度計算方式 [2] は概念と概念の関連性を関連度とよばれる数値で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．関連度は 0.0 から 1.0 の値を取り，概念間の関連が強いほど大きな値を示す．関連度の具体例を表 2 に示す.

表 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

関連度は概念同士の属性の対応により算出される．互いが持つ属性の内，最も意味が近いもの同士の組を作った上でそれぞれの重みを用いて関連度を算出する.

3 Wikipedia からの概念追加手法

現在提案されている概念ベースの概念数は 87,242 語であり，これらは主に複数の国語辞書から作成されている．国語辞書と比べると Wikipedia[3] は登録語数が多く，容易に新たな概念を取得することができると考えられる．また Wikipedia における語義文は，いわゆる辞書的な意味だけではなく，関連ある様々な事物事象についての記載があり，より良い属性を容易に取得できるのではないかと考えられる.

3.1 概念・属性表記の取得

概念は Wikipedia の見出し語とし，それぞれの属性は見出し語が示すページに記載された語義文中の語句から取得する．ただし概念，属性共にその語句のみで意味があるものとするため，形態素解析の結果から自立語のみを抽出する．図 1 に見出し語「文字化け」から取得される属性例を示す.

概念及び属性の取得において，本稿では「数字のみ」「数字+数詞」について取得対象外とした．これは例えば「56」という数字の見出し語や「56 年」といった西暦を示す見出し語などを指す．追加される概念の候補数は 236,787，総概念候補数は 324,029 となった.

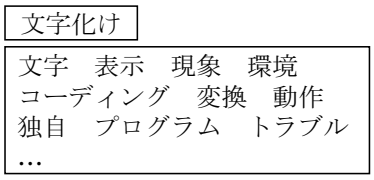

図 1: 取得される属性例

3.2 重み付与

取得した属性に対して重要度を示す重みを付与する．重みの算出には概念ベース TF 及び概念ベース IDF を用いる.

概念ベース TF による重みは，重みを付与する対象属性の N 次属性内における出現頻度から算出する．例えば概念「文字化け」の属性が図 2 に示すようなものだと仮定する.

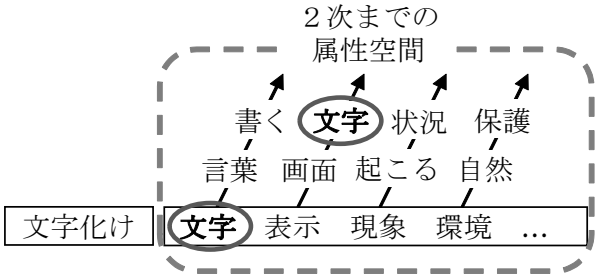

図 2: TF 算出の例

概念ベースにおいて属性も概念として定義されているため，それぞれの属性からもさらに属性を展開できる．概念「文字化け」が持つ属性を一次属性，一次属性から展開された属性を元の概念「文字化け」の二次属性と呼ぶ．本稿では属性の展開を二次までとし，この二次属性までの属性空間において重みを付与する対象属性が出現する頻度を概念ベース TF と定義する．図 2 の例において属性「文字」の概念ベース TF は 2 である.

概念ベース IDF による重みは，重みを付与する対象属性が付与されている概念の総数から算出する．例えば図 3 に示すような 3 つの概念及び属性が存在していると仮定する.

概念ベース中の全ての概念数を V_{all},「文字」を属性として持つ概念の数を df_1(文字) とすると，属性「文

字」の IDF 値は次の式で算出される．

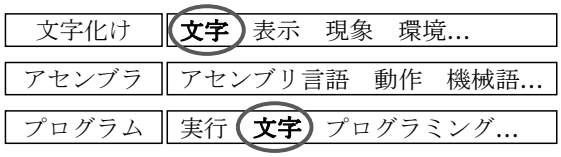

図 3: IDF 算出の例

$$IDF_1(\text{文字}) = \log_2 \frac{V_{all}}{df_1(\text{文字})} \quad (1)$$

図 3 の例であれば $\log_2 \frac{3}{2}$ となる．

最終的な重みの付与は「IDF のみ」「TF のみ」「$TF * IDF$」の 3 パターンを実施，結果を検証する．

4 評価方法

評価には $X - BC$ 評価セットを用いる．これはある基準となる概念 X，X との間に何かしら関連性のある概念 B，X とまったく関連の無い概念 C の 3 概念を 1 セットとした評価セットである．表 3 に評価セットの例を示す．

表 3: 評価セットの具体例

X	B	C
学術雑誌	権威	台風
暑中見舞い	西瓜	地球
文字化け	パソコン	歌人

評価の際には作成した各概念ベースを用いて関連度 $DoA(X,B)$ および $DoA(X,C)$ を算出する．概念 X と概念 B は何かしらの関連が見出せる概念同士であるため，関連の無い概念 C との関連度より大きくなることが望まれる．そこで $DoA(X,B)$ が $DoA(X,C)$ よりも大きい場合，そのセットを正解と見なす．3 概念のセットを人手で作成し，116 セット用意した上でこれらを用いて評価を行った．

5 評価結果

まずそれぞれの重み付与を行った概念ベースの評価結果を表 4 に示す．

表 4: 各重みによる評価結果

IDF	TF	$TF * IDF$
76.9 %	43.6 %	41.9 %

表 4 の結果より，概念ベース IDF による重み付与が最も高い精度となった．

IDF は値が小さいほど，多くの概念に属性として出現しているという事を指す．よってある閾値以下の IDF 値を持つ概念は Wikipedia における多くのページに出現する語句であり，各々の概念の意味定義において重要ではない雑音と考えられる．そこで表 4 に示した IDF 重みの概念ベースに対し，IDF の値に閾値を定めて概念の選別を行う．評価結果を表 5 に示す．

表 5: IDF 閾値による属性選別後の評価結果

IDF 閾値	可能セット数	精度（可能セット）
なし	116	76.9 %
3	115	76.5 %
4	110	76.4 %
5	95	82.1 %
6	33	81.8 %

表 5 の結果より，IDF の閾値を 5 と設定した場合に最も精度が高くなった．ただしテストセットに用いられている概念が IDF 閾値により削除されることが起こりえるため，精度は概念 X，B，C が全て削除されなかったセットのみを用いて算出している．3 概念全て残ったセットの数を可能セット数として示している．

テストセット中の概念が削除されることで閾値によりセット数に差が生じているため，表の精度をそのまま比較することは出来ない．そこで最も精度の高かった閾値 5 の際に残った 95 セットのテストセットを用いて，他の閾値での再評価を行った．結果を表 6 に示す．

表 6: 95 セットでの評価結果

IDF 閾値	精度
なし	80.0 %
3	77.9 %
4	78.9 %
5	82.1 %

最終的に IDF 閾値を 5 とした場合が精度 82.1 %となり最も良い結果となり，削除された概念数は 838 となった．IDF 閾値を 5 とした場合に削除された概念の例を図 4 に示す．

等　其の後　其の他　凡て 詳細　併せて　得る　数える 県　製品　豚　...

図 4: IDF 閾値で削除される概念の例

「等」「詳細」といった語は Wikipedia の文章構成上ほとんどのページに出現しており，意味定義にはふさわしくない概念として削除されている．「県」や「豚」といった語は一般性，使用度が高く削除されてしまっているが，概念として削除すべきかは疑問が残る．

最終的に概念ベースへ新たに追加された概念と属性の例を図 5 に示す．

概念	属性
プログラミング言語	言語，プログラミング，埋め込む，構文規則，規則，意味規則，定義，仕様，形式的，実装，生産性，向上，コンピュータ，変数，宣言文，実行
メーリングリスト	複数，人，電子メール，配信，同報，する，仕組み，特定，話題，関心，持つ，グループ，情報交換，する，場合，利用，する，多い
テーブルトークRPG	ゲーム，紙，鉛筆，サイコロ，道具，用いる，会話，ルールブック，記載，遊ぶ，対話，言葉，参加者，舞台，プレイヤー，キーパー，シナリオ
ヒップホップ	アメリカン，ヒスパニック，住民，コミュニティ，行う，文化，跳ぶ，アフリカ，音楽，ダンス，黒人，創造性，生まれる，スラング

図 5: 追加される概念と属性の例

変遷の激しい技術関連の用語や国語辞書などでは取得の難しい大衆文化に関する用語も Wikipedia からは容易に取得できる．属性に関してもページに記載されている語義文の量は充実しており，多くの情報を取得できている．

6　まとめ

本稿ではより人間らしい柔軟な語の意味理解を目指す概念ベースに対し，Wikipedia を用いて新たな概念を追加する手法について提案，検証を行った．Wikipedia の見出し語を概念とし，語義文から属性を取得することで自動的に概念の追加を行うことができる．属性への重みづけとして概念ベース IDF 値を用いた結果，精度は 117 セットのテストセットにおいて 82.1 %となり，Wikipedia から追加された概念の有効性を示せたと考える．ただし IDF 閾値による概念削除により精度の向上は見られたが，削除対象の概念は必ずしも適切とは言えず，精錬方法には改善の余地がある．

謝辞

本研究の一部は，JSPS 科研費 16K00311 の助成を受けて行ったものです．

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] Wikipedia [https://ja.wikipedia.org/wiki/]