E-027

# Wikipedia を用いた概念ベース構築法の提案
# Proposal of concept base construction method using Wikipedia

船岡 篤† 土屋 誠司‡ 渡部 広一‡
Atsushi Funaoka Seiji Tutiya Hirokazu Watabe

## 1. はじめに

近年，人間と円滑なコミュニケーションをとる知的なロボットへの期待が高まっている．人間にはある単語から別の単語を連想する能力があり，人間同士のコミュニケーションは連想によって円滑に行われていると考えられる．そこでロボットが人間と円滑に会話するには連想と同等なシステムが必要だと思われる．連想のためのシステムとして，語概念連想システムがある．このシステムは概念ベース[1]と関連度計算方式[2]によって構成されている．概念ベースは電子化された辞書などから機械的に構築するもので，ある語がそれを特徴づける語とその重要度を表す数値の対の集合によって定義される．

既存の概念ベースでは「ubuntu」といった専門性の高い用語や「e-sports」などの新語が含まれていないという問題がある．その問題を解決する方法として，大規模な Web 百科事典である Wikipedia を使用する概念ベースがある．Wikipedia はフリーで使用できるウェブ百科事典で，誰でも自由に編集・閲覧できる．日本語版の記事は約 110 万記事あり，大規模な概念を獲得することができる．また使用されている語の分野の包括範囲が広く，時事用語など近年使用される日常生活に定着した語も多く含まれている．そのため，Wikipedia を情報源とすることで，人間が日常生活で使用するより多くの語に対応した概念ベースの構築に役立つと考えられる．Wikipedia は辞書と違い，「開始」といった一般的な単語は記事本文中には存在するが記事タイトルとしては存在しない．以降このような概念を記事なしと呼ぶ．Wikipedia を用いるにあたって，記事なし概念の獲得が必要となる．本研究ではこの Wikipedia を用いた概念ベースを構築する手法を提案する．

## 2. 関連技術

### 2.1 概念ベース

概念ベースは概念・属性・重みで構成される．概念とはある単語，属性は概念の意味を表す単語であり，重みはその属性の重要度を表す．概念ベースの例を表 1 に示す．

表 1　概念ベースの例

| 概念 | 属性 |
|---|---|
| 医者 | (医師,0.34)(患者,0.11)(病院,0.08)・・・ |
| 病院 | (医院,0.25)(手術,0.11)(施設,0.04)・・・ |
| 治す | (治療,0.43)(医療,0.21)(病気,0.13)・・・ |

†同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

### 2.2 関連度計算方式

関連度計算方式とは，ある二つの概念間の関連の強さを定量的に表現する手法である[2]．関連度は 0 から 1 の間で変動し値が大きいほど関連は強くなる．概念 $A$，$B$ の持つ属性を$a_i$，$b_j$，重みを$u_i$，$v_j$とし，属性数を$L$個，$M$個とすると，概念 $A$，$B$は次のように表される．

$$A = \{(a_1, u_1),\ (a_2, w_2),\ \cdots,\ (a_L, w_L)\} \quad (1)$$

$$B = \{(b_{x_1}, v_{x_1}), \ldots,\ (b_{x_M}, v_{x_M})\} \quad (2)$$

これより，概念 $A$，$B$ の一致度$DoM(A,B)$と関連度$DoA(A,B)$は以下のように表される． なお，$a_i = b_j$は属性$a_i$と$b_j$が一致したことを表す．

$$DoM(A,B) = \sum_{a_i = b_j} min(u_i, v_j) \quad (3)$$

$$DoA(A,B) = \sum_i DoM(a_i, b_{x_i}) \times \frac{(u_i + b_{x_i})}{2} \times \frac{\min(u_i, v_{x_i})}{\max(u_i, v_{x_i})} \quad (4)$$

### 2.3 X-ABC 評価

基準概念を X と置き，この概念 X と関連が非常に強い概念 A，ある程度関連がある概念 B，まったく関連のない概念 C によって構成された 4 つの概念の組を用意する．例えば概念 X が「椅子」なら，概念 A に「腰掛け」，概念 B に「机」，概念 C に「喉」のような概念が与えられている．

概念 X と概念 A との関連度を DoA(X,A)とする．各概念においても同様としたとき，次の条件式を満たすとき正解とする順序正解率で評価する．

$$DoA(X,A) > DoA(X,B) > DoA(X,C) \quad (5)$$

### 2.4 形態素解析

形態素解析とは，文を意味を持つ最小単位の列に分割することである．本研究では，概念ベース構築のための元記事を形態素解析することで，概念や属性となる単語を抽出する．また，その際不必要となる助詞や助動詞などの雑音除去を行うためにもこの解析が必要となる．本研究では形態素解析ソフトとして MeCab[3]を用いる． MeCab はフリーの形態素解析ソフトで，解析に用いる辞書をユーザが指定，追加できる．本研究，では NEologd という，WEB 上の情報を用いて構築された，時事情報や専門用語などを含む約 310 万語の語彙を持つ辞書を利用する．

## 3. 概念ベースの構築手法

国語辞書から概念ベースを構築する手法と同様に概念を獲得すると，Wikipedia には記事が存在しない語がある．そこで語群が共起して出現する情報源から概念と属性を自動的に獲得する手法[4]を用いる．本研究では 2019 年 4 月時点の Wikipedia の記事データ約 114 万記事を情報源とする．

### 3.1 概念と属性の獲得

概念と属性は記事本文中より獲得する．まず 1 文を形態素解析し各自立語の基本形を得る． 出現した自立語をそれぞれ概念として獲得し，共起して出現した概念を属性として獲得する．ここで，パラメータとして窓幅[5]を設定し，概念の窓幅分の概念をその概念の属性として獲得する．

### 3.2 属性への重みづけ

獲得した属性に重みを付与する手法として，情報検索おいて広く利用されている$tf \cdot idf$を用いる．$tf$とは記事中に特定の語句が出現する回数を表している．$idf$とは各概念の特定性を表す手法であり，全概念の 1 次属性空間内において，対象となる概念を属性として持つ概念の総数から算出することができる．以上の$tf$と$idf$の値を利用し，ある概念$A$の属性$a$の重み$w(A,a)$を(4)式によって与える．$V_{all}$は全概念　数，$df(a)$は全概念の 1 次属性空間内で概念 $a$ を属性として持つ概念の数である．

$$idf(a) = \log_2 V_{all} / df(a) \ + \ 1.0 \qquad (6)$$

$$w(A,a) = tf(A,a) \times idf(a) \qquad (7)$$

## 4. 構築結果

構築した概念ベースの結果を表 2 に示す．既存 CB は国語辞書を情報源とした概念ベースであり，既存 WikipediaCB が既存 CB の手法で概念・属性を獲得し，記事なし概念は記事中より獲得し，その概念が出現する記事タイトルを属性として獲得したものである．Co-WikipediaCB が本研究で構築した概念ベースで，窓幅は後述する X-BC 評価で最も高い精度となった 28 である．

表 2　構築結果の比較

| | 概念数 | 平均属性数 |
|---|---|---|
| 既存 CB | 87,242 | 37.6 |
| 既存 WikipediaCB | 1,112,640 | 79.7 |
| Co-WikipediaCB(w=28) | 2,104,070 | 417.9 |

## 5. 評価

### 5.1 X-BC 評価

概念ベースが適切に単語間の関係を表現できているかを評価するために X-BC 評価手法を用いる．X-BC 評価手法とは，関連度の値を比較することで概念ベースを評価する手法である．ある基準概念 X と，概念 X と関連がある概念 B, 全く関連が無いであろう概念 C の 3 つの概念を 1 組として評価セットを人手で作成する．評価セットは表 7-9 のテストセットを用いる．概念 X と概念 B の関連度を DoA(X,B), 概念 X と概念 C の関連度を DoA(X,C)とし，関連度の値が式(7)を満たした場合を正解とする．

$$DoA(X,B) > DoA(X,C) \qquad (8)$$

この評価を全ての組で行い，正解となった組の割合を概念ベースの精度とする．結果を表 3 に示す．新たに構築した新聞記事概念ベースは表に窓幅の大きさを示す．

表 3　各概念ベースの X-BC 評価

| 概念ベース | 精度[%] |
|---|---|
| 既存 CB | 94.1 |
| 既存 WikipediaCB | 88.4 |
| Co-WikipediaCB(w=28) | 92.3 |

### 5.2 ヒット率

テストセットのうち概念 X，概念 B，概念 C の 3 つ全てが概念ベースに存在する割合をヒット率とする．表 4 に各概念ベースのヒット率を示す．

表 4　各概念ベースのヒット率

| 概念ベース | ヒット率[%] |
|---|---|
| 既存 CB | 57.3 |
| 既存 WikipediaCB | 97.6 |
| Co-WikipediaCB(w=28) | 90.3 |

## 6. 考察

構築結果より，概念数は既存 CB の約 24 倍，既存 WikipediaCB の約 2 倍となり，平均属性数も大幅に増加した．つまり，語彙数の問題を解決できたと考えられる．Wikipedia を用いることで多くの概念を獲得でき，また共起を用いることで記事本文中に出現する記事なし概念を獲得できたと考えられる．

X-BC 評価においては既存概念ベースと同程度の精度となり，辞書に掲載されているような基本的な語に関して語の関係を正しく表現できていると考えられる．

ヒット率においては既存概念ベースよりも上回り 33%向上し多くの語を獲得できていることがわかった．しかし既存 Wikipedia 概念ベースよりかは 7.3%低下した．ヒットしなかった語の例として「ミュージックステーション」や「ベルリン問題」がある．これらは，複合語であり，正しく 形態素解析できなかったゆえに概念として獲得できなかったためだと思われる．そのため，形態素解析の際に複合語処理を行うことでヒット率の上昇が期待できる．

## 7. おわりに

本稿では Wikipedia を用いることで語彙数の問題を解決し，共起範囲に窓幅を設け精度の向上を図った．その結果他の概念ベースと同程度の精度となった．また，語彙数の増加も見受けられた．窓幅をさらに広げることで，さらなる精度の上昇が見込まれる．

**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] 井筒大志，渡部広一，河岡司，“概念ベースを用いた連想機能実現のための関連度計算方式”，情報科学技術フォーラム FIT2002，pp.159–160，2002

[3] “MeCab -- 形態素解析器”，http://taku910.github.io/mecab/，京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所，2019-2-11 参照.

[4] 芋野美紗子，吉村枝里子，土屋誠司，渡部広一，“共起する情報群からの概念ベース自動生成手法”，信学技報，HCS2014-114，pp25-30，2015

[5] 山口修平，土屋誠司，渡部広一，“新聞記事を用いた大規模概念ベースの構築手法”，社会システムと情報技術研究ウィーク(WSSIT2019) 人工知能学会 知識ベースシステム研究会，2019