E-004

複合語の概念・属性を考慮した百科事典および国語辞書による概念ベースの構築
Construction of Concept-Base with Japanese Language Dictionaries and Encyclopedia Based on Concept and Attribute of Compound Words

白石 卓也† 芋野 美紗子‡ 土屋 誠司‡ 渡部 広一‡
Takuya Shiraishi Misako Imono Seiji Tsuchiya Hirokazu Watabe

1. はじめに

人間はある語から関連性のある語を連想する能力があり，これを会話で役立てている．この連想する能力をコンピュータに持たせることができれば，自然な会話ができるコンピュータの実現に近づくと考えられる．そこで我々は，人間の常識を判断するシステムとそれを支える語概念連想システムを構築している．語概念連想システムを実現するために語の意味を理解するための概念ベース[1]や語と語の関連性の強さを計るための関連度計算方式[2]を用いている．

既存の概念ベース[1]は「知的財産権」のような時事用語や専門用語などの言葉が欠如している．現状では人間が日常的に使用する言葉を網羅できておらず，日常生活に必要な語を連想できないという問題が考えられる．そこで，百科事典を用いて時事用語や専門用語が登録された概念ベースの構築を目指す．

2. 概念ベース

概念ベースとは電子化された国語辞書や新聞記事などから自動的に構築した知識ベースである．ある語を概念と定義し，概念の意味特徴を表す語（属性）とその重要さを表す数値（重み）の対の集合によって定義している．ある概念 A は n 個の属性 a_i と重み w_i（>0）の対によって(1)式のように定義される．

$$A=\{(a_1, w_1),\ (a_2, w_2),\ \cdots,\ (a_n, w_n)\} \quad (1)$$

ここで，概念自身が持つ属性を一次属性と呼ぶ．概念ベースでは全ての属性は概念としても登録されているため，一次属性からも属性を導くことができる．ここで導かれた属性を元の概念に対する二次属性と呼ぶ．同様に任意の次元まで属性を導くことができる．つまり概念ベースは，n 次の属性の連鎖集合の構造になっている．

3. 関連度計算方式

関連度計算方式とは概念ベースにある 2 つの概念の関連の強さを定量的に表現する手法である．算出された数値を関連度と呼ぶ．関連度は 0.0 から 1.0 までの実数値で表現され，関連度が大きいほど概念間の関連が強いといえる．

4. 既存の百科事典概念ベース

時事用語や専門用語を多く含む電子百科事典「現代用語の基礎知識」[3]を情報源として利用した百科事典概念ベース[4]がある．この情報源は，見出し語，属するカテゴリ，説明文というような規則的な表記構造をしている．見出し語を概念としたとき，その説明文中の各単語は見出し語の特徴を表す語であるため属性とすることできる．

† 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

この百科事典には単語，複合語だけでなく「地球の温暖化」といった句も収録されているため，見出し語の選別を行っている．情報源から概念と属性を獲得する際に茶筌[5]を使用して形態素解析を行い，単語の品詞を特定している．また，百科事典は国語辞書にある基本的な語を見出し語に持っていないため，百科事典のみでは十分な語を包括できない．そのため，国語辞書を用いた既存の概念ベースに百科事典から獲得した概念・属性を追加することで言葉を補っている．

概念として獲得する語を茶筌の出力結果が「名詞」，「動詞」，「形容詞」となる語および「名詞のみで構成される複合語」としている．また，茶筌の出力結果が「固有名詞」，「助詞」，「記号」，「英数字」となる語を含む語を概念から除外している．しかし，この獲得手法では動詞や形容詞を含む複合語や「未知語」（茶筌の辞書に未登録の語）を獲得できない．さらに，カタカナで構成されている語は茶筌によって不適切に分解される場合（図 1）があるため，獲得できない語が存在する．

リ	名詞-固有名詞-一般
コメ	名詞-一般
ン	名詞-非自立-一般
ド	名詞-一般

図 1 「リコメンド」の茶筌出力

属性を獲得する際には複合語を考慮していない．よって，既存の概念・属性の獲得手法は適切でないといえる．

5. 百科事典と国語辞典による概念ベースの構築

本研究では複合語を考慮し，百科事典「現代用語の基礎知識」と国語辞書「岩波国語辞書」[6]から概念・属性を獲得することで概念ベースの構築を行う．

5.1 カテゴリによる見出し語の除外

4 章で述べたように情報源の見出し語にはカテゴリが付随している．概念の選別を行う前に概念候補に不適切なカテゴリの見出し語を除外する．例えば〔ヨーロッパ各国データ『世界事典』〕というカテゴリの見出し語「ルーマニア」の説明文は，面積や人口といったデータであり概念・属性の獲得には適さないカテゴリである．

5.2 概念の選別

概念として獲得する語を，茶筌の出力結果が「名詞」，「動詞」，「形容詞」となる語および「複合語」とする．また，既存手法では獲得しなかった「アルファベットを含む語」や，茶筌によって不適切に分解される「カタカナを含む語」についても獲得する．

本稿では「複合語」を茶筌の出力結果の前後関係が「接頭詞＋名詞」，「名詞＋名詞」，「名詞＋接尾詞」，「名詞+助動詞『ない』」，「動詞（連用形）＋動詞」，「動

詞＋形容詞」という前後関係のある語とする[7]．また，小選挙区（小（接頭詞）＋選挙（名詞）＋区（接尾詞））のような 3 語以上で構成される複合語についても同様に前後関係を見て複合語として獲得する．

「DNA」といったアルファベットの語は日常的に使用されているため，「アルファベットを含む語」を概念として獲得する．多くのアルファベットの語は茶筌の登録辞書に存在しないため「未知語」となるが，アルファベットの語は意味のある語と考え，概念として獲得する．「カタカナを含む語」は外来語など日常的に使用されているため概念として獲得する．上記のように選別を行うことで，既存手法より多くの概念を獲得できる．

5.3 属性の獲得手法

属性は概念として獲得した見出し語の説明文から獲得する．その説明文に複合語が存在するとき，その複合語を属性に使用することで，より適当な属性を獲得できる．属性を獲得する際，概念として獲得した語を使用する．複合語の属性を獲得する手法として，説明文の先頭から 1 文字ずつ全概念と表記一致の検索を行う．その中で文字数が最大になる文字列（概念）を属性として獲得し，獲得した文字列の次の文字から同様の作業を文末まで繰り返す．

5.4 重みの付与

属性の重み付けには *tf・idf* [8]の考え方を概念ベースに応用した概念ベース *tf・idf* を用いる．ある概念 *A* の属性 *a* の重み *W*(*A*, *a*)は以下の(2)式で求められる．

$$W(A,a) = tf(a_n) \times \log_2 \frac{V_{all}}{df_n(a)} \qquad (2)$$

ここで $tf(a_n)$とは概念 *A* の *n* 次属性内に概念 *a* が出てくる頻度，V_{all} は概念ベース内の全概念の総数，また $df_n(a)$は *n* 次属性空間内に概念 *a* を属性として持つ概念の数を表す．本稿では実験的に求めた *n*=2 で付与した重みを用いる．

6. 精度評価

6.1 *X-ABC* 評価

X-ABC 評価では，任意の基準概念を *X* と置き，この概念 *X* と関連が強い概念 *A*，関連がある概念 *B*，関連がない概念 *C* によって構成された概念の組を用意する．そこで，概念 *X* と *A* の関連度を *DoA*(*X*,*A*)とし，各概念に対しても同様とする．*DoA*(*X*,*C*)は関連のない概念同士の関連度であるため，理想としては 0 であるが，概念間の属性の表記一致を用いるため計算方式上 0 にするのは困難である．そのため，テストデータ全体での *DoA*(*X*,*C*)の平均を *AveDoA*(*X*,*C*)とし，これを関連度の誤差として用いる．次の条件式を満たすものを正解とし評価を行う．テストデータは 500 組用意した．

$$DoA(X,A) - DoA(X,B) > AveDoA(X,C) \qquad (3)$$

$$DoA(X,B) - DoA(X,C) > AveDoA(X,C) \qquad (4)$$

既存の百科事典を用いた概念ベース（*CB_A*），本稿で構築した概念ベース（*CB_B*）に対して評価を行った．関連度計算には概念の属性を使用するため，使用属性数を変更すると関連度も変化する．そこで重み上位順から使用する属性数を 10～40 個に変更しながら評価を行った．表 1 に評価結果を示す．

表 1　*X-ABC* 評価による精度（%）

	10 個	20 個	30 個	40 個
CB_A	41.6	68.2	70.4	68.6
CB_B	34.2	42.0	40.0	40.4

6.2 目視評価

ランダムに 100 概念を取得し，その重み上位 30 の属性が妥当であるかを評価する． *CB_A* と *CB_B* について評価を行った．被験者は 3 名で 2 名以上が正しいとした属性の割合を精度とする．表 2 に評価結果，表 3 に正解を○，不正解を×として評価の例を示す．

表 2　目視評価による精度（%）

CB_A	*CB_B*
50.2	60.0

表 3　概念「電子認証」の属性の一部

CB_A	*CB_B*
セキュリティ（○）	コンピュータ暗号（○）
フロッピーディスク（○）	デジタル署名（○）
ホスト（×）	ホスト・コンピュータ（○）

7. 考察

本稿で構築した概念ベースは，概念を約 128580 獲得でき，既存手法の 115093 より多くなる結果となった．また，目視評価より本手法の属性の精度が高くなる結果となった．これは，表 3 の「ホスト」と「ホスト・コンピュータ」のように茶筌で区切られた語ではなく，複合語を考慮して属性を獲得したため，適切な属性が多くなったと考えられる．

X-ABC 評価では，既存手法より低い精度となった．このため，重み付け手法を検討する必要があると考えられる．また，平均属性数が 35 となり，既存手法の 43 より少ないことから属性の追加についても検討する必要があると考えられる．

8. まとめ

複合語を考慮して概念・属性の獲得した結果，概念数の増加，高い精度の属性の獲得ができた．しかし，精度評価の結果より，本稿で構築した概念ベースの精度向上には重み付けや属性の獲得手法が課題であると考えられる．

謝辞

本研究の一部は，科学研究費補助金（若手研究（B）24700215）の補助を受けて行った．

参考文献

[1] 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.

[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] 「現代用語の基礎知識」編集部（編），“現代用語の基礎知識 1991～2009”，自由国民社，2009.

[4] 大竹慎吾，“百科事典を用いた新概念ベースの構築”，同志社大学工学部知識工学科修士論文，2013.

[5] 茶筌 - 形態素解析器，奈良先端科学技術大学院大学 情報科学研究科自然言語処理学講座(松本研究室), http://chasen-legacy.sourceforge.jp/，2013/6/19.

[6] 西尾実，岩淵悦太郎，水谷静夫，“岩波国語辞典第五版”，岩波書店，1994.

[7] 柳瀬秀夫，“新聞記事からの複合語概念表記の獲得”，情報科学技術フォーラム FIT2011，pp.269-270，2011.

[8] 徳永健伸（編），“情報検索と言語処理”，東京大学出版会，1999.