Vol.2016-ICS-183 No.5
2016/3/16

百科事典と国語辞典による概念ベースの構築

白石卓也[†1]　芋野美紗子[†2]　土屋誠司[†2]　渡部広一[†2]

人間は自身の経験や学習から蓄積した語の知識（常識）と，語を他の語と関連付けることができる能力（連想）によって語の意味を語間の曖昧な関連性によって理解することができると考えられる．この連想をコンピュータに持たせることができれば，人間のような語の意味理解の実現に近づくと考えられる．そこで我々は，概念ベースと関連度計算方式を機軸とした語概念連想システムを構築している．概念ベースでは一つの語を概念とし，概念の意味を表す属性とその重要性を表す重みの対の集合で定義されている．既存の概念ベースは国語辞典が情報源であるため，時事用語や専門用語などの言葉が欠如しており，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ある．そこで，時事用語や専門用語を多く含む百科事典を情報源とした概念ベースがある．しかし，百科事典を用いた概念ベースでは概念・属性の獲得手法に不足があるため，概念の網羅性に問題があると考えられる．以上より本研究では，時事用語や専門用語だけでなく基本的な語と幅広い語を網羅するために，百科事典と国語辞典を情報源とした概念ベースの構築をした．その結果，本研究では概念を 220463 個獲得でき，既存の百科事典を用いた概念ベースより概念数を 191701 個多く獲得でき，概念の網羅性を向上することができた．また，X-ABC 評価での精度が 74.33%となり 22%向上することができた．

Construction of Concept-Base with Japanese Language Dictionaries and Encyclopedia

TAKUYA SHIRAISHI[†1]　MISAKO IMONO[†2]
SEIJI TSUCHIYA[†2]　HIROKAZU WATABE[†2]

1. はじめに

近年，情報処理技術の発展は目覚しく，その技術を用いた各種情報処理システムは人間社会のあらゆる分野に活用され，もはや欠かすことのできない存在となっている．しかし，それらのシステムの発展の多くは高性能化，多機能化によるもので，複雑化した利用方法はユーザにとって負担となるケースもある．あらゆるユーザにとって利用しやすいシステムを実現するためには，ユーザが特別な知識や技能を用いずともそのシステムを利用できる必要がある．そこで，人間が日常生活で使用する自然言語を用いて利用できるシステムを構築することができれば，ユーザの負担を軽減できると考えられる．

自然言語を対象とした処理をコンピュータで実現するためには，その対象となる語の意味をコンピュータが理解する必要がある．人間が理解している語の意味とは，体系的に整理された明確な知識だけではなく，明確には記述できない語と語の曖昧な関係性で表現された知識でもあると考えられる．人間が曖昧な関連性によって語の意味を理解することができるのは，自身の経験や学習によって蓄積した語の知識（常識）と，語を他の語と関連付けることができる能力（連想）によるものであると考えられる．これらの語に関する「常識」と「連想する能力」をコンピュータ上で実現することができれば，人間のような語の意味理解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，語の常識を用いて意味を判断する「常識判断システム」[1]とそれを支える「語概念連想システム」[1]を構築している．

語概念連想システムとは，概念ベース[2]を中核として語の連想や意味理解を実現するためのシステムである．概念ベースには様々な語（概念）が，それを特徴付ける語（属性）とその重要度を表す数値（重み）の対の集合によって定義されている．

既存の概念ベースでは，概念や属性を獲得する際に，電子化された国語辞書を主な情報源としている．国語辞書には基本的な語が多く収録されているが,「ねじれ国会」や「知的財産権」のような時事用語や専門用語などの近年使用され定着した語が欠如している．つまり，既存の概念ベースは人間が日常的に使用する言葉を十分に網羅しているとは言えず，知識に偏りが生じ，日常生活で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，時事用語や専門用語を多く含む百科事典から概念・属性の獲得を目指した百科事典を用いた概念ベース[3]がある．しかし，百科事典を用いた概念ベースでは概念・属性の獲得手法に不足があるため，概念の網羅性に問題があると考えられる．

以上より本稿では，時事用語や専門用語だけでなく基本的な語と幅広い語を網羅するために，百科事典と国語辞典を情報源とした概念ベースの構築を目指す．

†1 同志社大学大学院　理工学研究科
Graduate School of Science and Engineering, Doshisha University

†2 同志社大学　理工学部
Faculty of Science and Engineering, Doshisha University

ⓒ2016 Information Processing Society of Japan

2. 語概念連想システム

語概念連想システムでは概念ベースやシソーラスなどの知識を利用し，想起語処理と未知語処理を提供している．想起語処理とは与えられた語からその語に対して関連の強い語を想起する処理である．また，未知語処理とは常識判断システムが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）に置き換える処理であり，この未知語処理によってシステムが保持しておく知識を最小限に抑えることができる．我々は，これらの処理を概念ベースと関連度計算方式[4]によって実現している．以下，概念ベースと関連度計算方式について述べる．

2.1 概念ベース

概念ベースとは，電子化された国語辞書や新聞記事などから機械的に構築され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m個の属性a_iと重みw_i（>0）の対によって次のように表現される．

$$\text{概念}A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_m, w_m)\} \quad (1)$$

概念ベースの特徴として，属性を成す単語群も概念ベースの中で概念として定義されている点がある．概念Aの意味定義を行う属性a_iを、概念Aの一次属性と呼ぶ．この属性a_iを概念とみなして更に属性を導くことができ，概念a_iから導かれた属性a_{ij}を，元の概念Aの二次属性と呼ぶ．概念ベースの具体的な例を表 1 に示す．

表 1　概念ベースの例

概念	(属性,重み)
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), …

2.2 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．このように概念と概念の関連を定量化することで，コンピュータに語の関連を判断させることが可能となる．関連度計算方式にはお互いの概念が持つ属性の一致度と重みを利用する重み比率付き関連度計算方式を使用する．

2.2.1 一致度

ある概念A，Bにおいて，その属性をa_i，b_j対応する重みをu_i，v_jそれぞれ属性が L 個，M 個（$L \leq M$）とすると，概念A，Bはそれぞれ

$$A = \{(a_1, u_1), (a_2, u_2), \cdots, (a_L, u_L)\} \quad (2)$$

$$B = \{(b_1, v_1), (b_2, v_2), \cdots, (b_M, v_M)\} \quad (3)$$

となる．このとき，概念Aと概念Bの属性一致度$DoM(A,B)$を以下のように定義する．

$$DoM(A,B) = \sum_{a_i = b_j} \min(u_i, v_j) \quad (4)$$

$$min(u_i, v_j) = \begin{cases} u_i (u_i \leq v_j) \\ v_j (u_i > v_j) \end{cases} \quad (5)$$

ここで，$a_i = b_j$は属性同士が一致した場合を示している．つまり，一致度とは概念Aと概念Bそれぞれの属性を比較して一致した場合に，小さい方の重みを選択して足し合わせた合計値ということになる．これは，小さい方の重みは互いの属性の重みの共通部分となっているので，概念Aと概念Bどちらにも有効な重みだと考えられるためである．このとき各概念の重みの総和が 1 になるように正規化する．よって，一致度は 0.0 から 1.0 の実数値をとる．

2.2.2 重み比率付き関連度計算方式

重み比率付き関連度は次の手順で求める．2.2.1 項で述べた概念A，Bにおいて，まず属性数の少ない方の概念Aを基準とし，その属性の並びを固定する．その上で概念Bの属性を概念Aの各属性との一致度の和が最大になるように並び替える．このときの概念Bの属性と重みを(b_{Xi}, v_{Xi})として次のように定義する．

$$B = \{(b_{X1}, v_{X1}), (b_{X2}, v_{X2}), \cdots, (b_{XM}, v_{XM})\} \quad (6)$$

これらの概念についての重み比率付き関連度$DoA(A,B)$を次の式で定義する．

$$DoA(A,B) = \sum_i DoM(a_i, b_{Xi}) \times \frac{(u_i + v_{Xi})}{2} \times \frac{min(u_i, v_{Xi})}{max(u_i, v_{Xi})} \quad (7)$$

関連度$DoA(A,B)$は，属性の一致度に，属性間の重みの比率と平均値を乗じた値となる．

3. 情報源

概念ベースを構築するためには機械的に処理できる情報源が必要となる．そこで電子化百科事典「現代用語の基礎知識」[6]と電子化国語辞典「岩波国語辞典」[7],「広辞苑」[8]を情報源として用いる．本稿では，これらの辞書の見出し語から概念，説明文からその属性を獲得する．

3.1 百科事典

電子化百科事典「現代用語の基礎知識」に収録されている語は，時事問題，人文科学，自然科学，社会科学など，掲載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日常生活に定着した語も多く含まれている．そのため，この辞書を概念ベースの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る．本稿では 1991 年版から 2009 年版の電子化百科事典を処理の対象にした．

この百科事典は図 1 に示すように見出し語(英字表記)，カテゴリ，見出し語の説明文というように規則的な表記構造をしている．

ⓒ2016 Information Processing Society of Japan

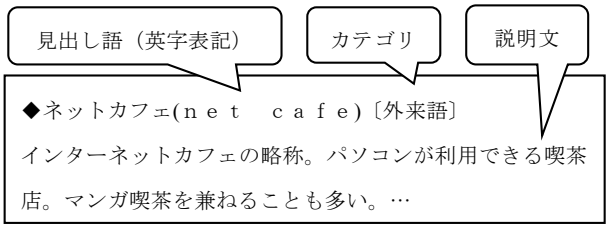

図 1　「現代用語の基礎知識」の構造

図 1 で示した百科事典の構造の見出し語には説明文だけでなくカテゴリが付随している．概念・属性の獲得を行う際に不適切なカテゴリの見出し語を除外する．不適切なカテゴリとして説明文が図や数値データ，見出し語が顔文字など記号で構成されたものがある．例えば，カテゴリ〔世界の国旗〕内では見出し語と図のみであるため，概念・属性の獲得には適さないカテゴリであると考えられる．

3.2 国語辞典

百科事典とは，人間の知識の全般を解説したものであるため，基本的な語彙を収録していない．そのため，基本的な語彙を補完するために国語辞典を概念ベースの情報源とする．電子化国語辞典「岩波国語辞典」，電子化国語辞典「広辞苑」を用いる．これらの国語辞書は図 2，図 3 に示すように読み，見出し語（英字表記），説明文というように規則的な表記構造をしている．

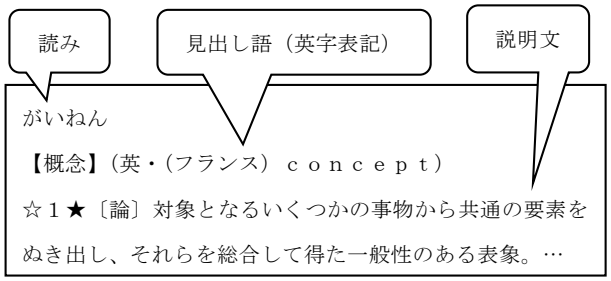

図 2　「岩波国語辞典」の構造

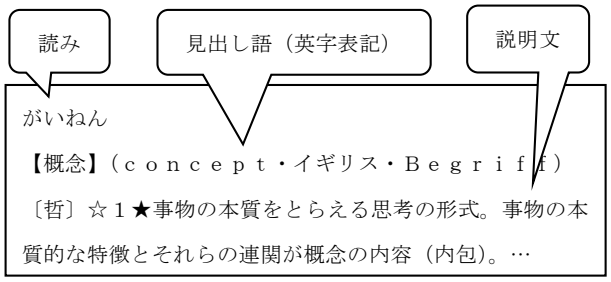

図 3　「広辞苑」の構造

4. 百科事典と国語辞書による概念ベース構築

本章では，百科事典と国語辞典を用いた概念ベースの情報源および構築手法，構築結果について述べる．

4.1 概念の獲得

本稿で用いる情報源の見出し語には単語や複合語だけでなく「地球の温暖化」といった句も収録されているため，形態素解析を行い，品詞を特定することで見出し語の選別を行う．形態素解析の処理では，文章を語が意味を成す最小単位である形態素に分解し，形態素それぞれの品詞や活用を判別する．形態素解析を行うためのツールとして茶筌[9]を使用する．ひらがなや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がある．形態素の品詞情報のみを手がかりにした場合，獲得できない語が存在する．不適切に分解される例を表 2 に示す．

表 2　「リコメンド」の形態素解析結果

形態素（表層）	品詞
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

本稿では概念として獲得する語の分類を茶筌の出力結果が「名詞」，「動詞」，「形容詞」となる語および 3.2.2 項で述べる「複合語」とする．また，「アルファベットを含む語」や「未知語」，「カタカナを含む語」についても考慮する．

「アルファベットを含む語」は，「ＤＮＡ」や「ＤＶＤレコーダー」といった略語や新語などが日常的に使用されているため概念として登録する必要があると考えられる．「カタカナを含む語」は「コンプライアンス」といった外来語が日常的に使用されているため，カタカナ部分が表 4 のように不適切に分解されたとしても概念として登録する必要があると考えられる．アルファベットやカタカナで構成されている語の多くは茶筌の辞書に存在しないため「未知語」となるが，意味のある語と考え，概念として獲得する．

「ひらがな」で構成されている語は茶筌の出力結果で感動詞や助詞，助動詞などを含んでしまうものが存在し語か句の判別ができないため本稿では考慮しない．

4.1.1 カタカナ・アルファベットを含む語の獲得手法

見出し語を形態素解析し，カタカナ表記の形態素が前後で並んでいる場合，それらの表記を連結し 1 つの形態素とする．これにより，表 6 の「リコメンド」ように形態素解析で不適切に分解される語を 1 語として獲得することができる．また，「アール・ヌーヴォー」のように「・(中黒)」がある場合も表記を連結する．英字も同様に表記を連結する．連結したこれらの語の品詞を「名詞-一般」とする．

例えば，「デジタルテレビ放送」は「デジタル+テレビ+放送」と形態素解析され，カタカナ表記の形態素を連結すると「デジタルテレビ+放送」となる．このように，連結後に複数の形態素に分割されている場合，4.3.2 項の複合語の獲得手法に移行する．

4.1.2 複合語の獲得手法

見出し語を形態素解析し，形態素を連結した結果，1 語となった見出し語を複合語として概念に獲得する．本稿では複合語を獲得する際，形態素解析した結果の品詞の前後関係が以下の表 3 に示すときに前後の形態素を連結する．

ⓒ2016 Information Processing Society of Japan

表 3 は各辞書よりランダムに選出した見出し語 4000 件を形態素解析することで得た.

表 3　複合語の結合規則

品詞（前部）	品詞（後部）
接頭詞-名詞接続	名詞-一般 名詞-サ変接続 名詞-形容動詞語幹 名詞-固有名詞
名詞-一般	名詞-一般 名詞-サ変接続 名詞-形容動詞語幹 名詞-ナイ形容詞語幹 名詞-副詞可能 名詞-固有名詞 名詞-接尾-一般 名詞-接尾-サ変接続 名詞-接尾-形容動詞語幹 名詞-接尾-地域 名詞-接尾-助数詞 動詞-自立（連用形）
名詞-サ変接続	名詞-一般 名詞-サ変接続 名詞-形容動詞語幹 名詞-ナイ形容詞語幹 名詞-副詞可能 名詞-固有名詞 名詞-接尾-一般 名詞-接尾-サ変接続 名詞-接尾-形容動詞語幹 名詞-接尾-地域 名詞-接尾-助数詞
名詞-形容動詞語幹	名詞-一般 名詞-サ変接続 名詞-形容動詞語幹 名詞-ナイ形容詞語幹 名詞-副詞可能 名詞-接尾-一般 名詞-接尾-サ変接続 名詞-接尾-形容動詞語幹 名詞-固有名詞
名詞-副詞可能	名詞-一般 名詞-サ変接続 名詞-形容動詞語幹 名詞-副詞可能 名詞-接尾-一般 名詞-接尾-サ変接続 名詞-接尾-形容動詞語幹
名詞-副詞可能	名詞-一般 名詞-サ変接続 名詞-形容動詞語幹 名詞-副詞可能 名詞-接尾-一般 名詞-接尾-サ変接続 名詞-接尾-形容動詞語幹
名詞-ナイ形容詞語幹	助動詞（特殊・ナイ）
名詞-固有名詞	名詞-一般 名詞-サ変接続 名詞-ナイ形容詞語幹 名詞-副詞可能 名詞-固有名詞 名詞-接尾-一般 名詞-接尾-サ変接続 名詞-接尾-形容動詞語幹 名詞-接尾-地域
接頭詞-数接続	名詞-数
名詞-数	名詞-数 名詞-接尾-助数詞
接頭詞-動詞接続	動詞-自立
動詞-自立（連用形）	名詞-一般 名詞-サ変接続 名詞-副詞可能 名詞-接尾-一般 動詞-自立 形容詞-非自立（アウオ段）
動詞-自立（体言接続特殊２）	名詞-一般 名詞-サ変接続 名詞-接尾-一般

「名詞-固有名詞」は 「名詞-固有名詞-一般」,「名詞-固有名詞-組織」,「名詞-固有名詞-地域-一般」,「名詞-固有名詞-地域-国」,「名詞-固有名詞-人名-一般」,「名詞-固有名詞-人名-姓」,「名詞-固有名詞-人名-名」の総称とする.

結合された複合語の品詞は後部のものを選択する．ただし, 後部の品詞が「名詞-接尾-一般」のときは「名詞-一般」,「名詞-接尾-サ変接続」のときは「名詞-サ変接続」,「名詞-接尾-形容動詞語幹」のときは「名詞-形容動詞語幹」,「名詞-接尾-地域」のときは「名詞-固有名詞-地域-一般」,「名詞-一般」と「動詞-自立（連用形）」が連結された複合語は「名詞-サ変接続」とする．3 語以上の形態素で構成される複合語についても同様に前後関係を参照し結合を繰り返すことで複合語として獲得する.「アジア自由貿易圏」を例に形態素結合の流れを図 4 に示す.

ⓒ2016 Information Processing Society of Japan

①	アジア（名詞-固有名詞）＋自由（名詞-形容動詞語幹）＋貿易（名詞-サ変接続）＋圏（接尾-名詞-一般）
②	アジア（名詞-固有名詞）＋自由貿易（名詞-サ変接続）＋圏（接尾-名詞-一般）
③	アジア（名詞-固有名詞）＋自由貿易圏（名詞-一般）
④	アジア自由貿易圏（名詞-一般）

図 4　形態素結合の例

図 4 では，先頭の「アジア（名詞-固有名詞）」と次の「自由（名詞-形容動詞語幹）」は表 3 より連結不可なので，「自由（名詞-形容動詞語幹）」とその次の形態素「貿易（名詞-サ変接続）」と連結し，「自由貿易（名詞-サ変接続）」となる．末尾の形態素「圏（接尾-名詞-一般）」まで連結作業を行った後，先頭の形態素に戻り，「アジア（名詞-固有名詞）」と「自由貿易圏（名詞-一般）」の連結する．

4.2 統合

百科事典の見出しには，「アール・ヌーヴォー（ａｒｔ　ｎｏｕｖｅａｕ）」や「アール・ヌーボー（ａｒｔ　ｎｏｕｖｅａｕ）」といったように表記ゆれが存在している．また，複数の辞書を用いているため，広辞苑では「アールヌーヴォー（ａｒｔ　ｎｏｕｖｅａｕ）」と表記されている．これらの語は同音・同義であるので，表記によって異なる概念にすべきでないと考えられる．そのため，獲得した概念に表記ゆれが存在するとき，それらの概念を統合し 1 つの概念として扱う必要があると考えられる．例えば，「アール・ヌーヴォー」のような外来語のカタカナ表記であれば，見出しより「ａｒｔ　ｎｏｕｖｅａｕ」という英字表記を手がかりとして概念の統合をすることができる．見出しの英字表記が同一であるが，カタカナ表記の見出し語（概念）の表記が異なる場合，それらを同一の概念として扱う．

4.3 属性の獲得

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，より適切な属性を獲得できると考えられる．以下に説明文から複合語を獲得するための手法を述べる．

4.3.1 最長前方一致による属性の獲得

説明文の先頭から 1 文字ずつ全概念と前方一致検索を行う．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．現在の文字列に表記一致する概念が存在しなければ，文字列の先頭の次の文字より同様の作業を行う．

4.3.2 形態素を用いた最長前方一致による属性の獲得

説明文を形態素解析し，先頭の形態素から 4.3.1 項のカタカナ･アルファベットを含む語の獲得手法および 4.3.2 項の形態素結合規則に沿って後部の形態素 1 つを結合する度に全概念と前方一致検索を行う．その中で形態素の結合数が最大になる複合語（概念）を属性として獲得し，獲得した複合語の次の形態素から同様の作業を文末まで繰り返す．現在の複合語に表記一致する概念が存在しなければ，複合語の結合を解き，その先頭の次の形態素より同様の作業を行う．

4.4 重みの付与

属性の選別により獲得した語に重みを付与する手法として，概念ベース$tf \cdot idf$を用いる．概念ベース$tf \cdot idf$とは，情報検索における索引語の重み付け手法として広く利用されている$tf \cdot idf$[10]の考え方を概念ベースに適用したものである．

概念ベースtfとは，概念ベース内における各概念中の属性の網羅性を表す尺度である．概念ベースを仮想的な文書集合として捉えることで算出する．概念ベースでは，各概念をn次の属性連鎖集合によって定義している．したがって，概念aのn次属性空間内において対象となる属性aが出現する頻度$tf_n(A,a)$を算出する．例えば，概念「自動車」が図 5 に示すような属性を持つ場合，属性「走る」の 2 次属性空間内頻度（tf_2(自動車, 走る) の値）は 3 となる．

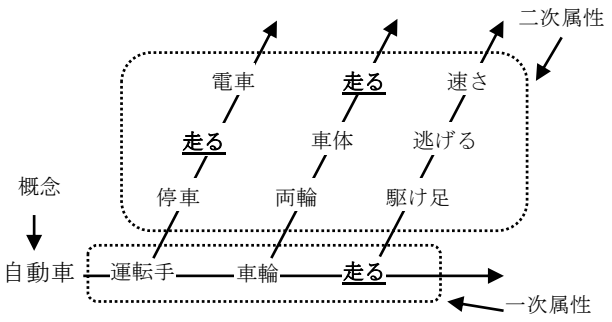

図 5　概念「自動車」の属性

概念ベースidfとは，概念ベース内における各概念の特定性を表す尺度である．全概念のn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．頻出する概念に比べて，稀に出現する概念は概念をより特徴付ける概念（属性）である考えられる．例えば，概念「人」を概念のn次属性空間内に属性として持つ概念は多く，概念「人力車」をn次属性空間内に属性として持つ概念は少ない場合，この文書空間内では概念「人力車」の方が概念「人」より概念を特徴付けることができる．n次属性空間内における概念aの概念ベースidfの値$idf_n(a)$を(8)式によって定義する．V_{all}は概念ベースに定義されている全概念数，$df_n(a)$は全概念のn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．以上の概念ベースtfと概念ベース idfの値を利用し，ある概念Aの属性aの重み$w(A,a)$を(9)式によって与える．

$$idf_n(a) = \log_2 \frac{V_{all}}{df_n(a)} \tag{8}$$

$$w(A,a) = tf_n(A,a) \times idf_n(a) \tag{9}$$

使用する属性空間（n）の設定によって算出される概念ベ

ースtfと概念ベースidfの値が異なる. そこで最も重み付けの精度がよい属性空間を実験的に検証する必要がある.

4.5 構築結果

百科事典と国語辞典から構築した概念ベース中, 4.3.1 項の最長前方一致によって属性を獲得した概念ベースを概念ベース *A*, 4.3.2 項の形態素を用いた最長前方一致によって属性を獲得した手法で属性獲得した概念ベースを概念ベース *B* と定義する. 表 4 にそれぞれの概念ベースの概念数と平均属性数を示す.

表 4 構築した概念ベースの概念数と平均属性数

	A	*B*
概念数	220463	220463
平均属性数	33	26

5. 評価

本稿で使用した概念ベースの評価を行った.

5.1 *X-ABC* 評価

X-ABC 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, *X* と関連が非常に強い概念 *A*, 概念 *A* ほどではないが関連がある概念 *B*, 全く関連が無いであろう概念 *C* の 4 つの概念を 1 組として評価セットを人手で作成する. 表 5 に百科事典の見出し語から作成した 300 組の評価セットの例を示す.

表 5 *X-ABC* 評価テストセット

X	*A*	*B*	*C*
アンパイア	審判	プロ野球	気管支喘息
会社説明会	合説	就職	太平洋時代
官僚主義	官僚制	国家公務員	ゼリー飲料

概念 *X* と概念 *A* との関連度を$DoA(X,A)$, 概念 *X* と概念 *B* との関連度を$DoA(X,B)$, 概念 *X* と概念 *C* との関連度を$DoA(X,C)$とする. それぞれの概念間の関連度の値を比較することで概念ベースを評価する. 概念 *X* と関連がない概念 *C* との関連度$DoA(X,C)$は, 本来 0.0 となるのが理想である. しかし関連度計算方式の特性上, 概念 *X* と概念 *C* それぞれの二次属性を比較して 1 つでも共通した属性が存在すれば微小な値が算出されてしまう. そこで概念 *C* との関連$DoA(X,C)$を誤差とみなし, その平均$AveDoA(X,C)$をテストセット全体での平均誤差とする. そして$DoA(X,A)$, $DoA(X,B)$, $DoA(X,C)$それぞれの関連度の間に平均誤差以上の差が存在していれば, 人間の常識に沿った関連度が算出されていると見なす. 以下の式を満たすとき正解とする.

$$DoA(X,A) - DoA(X,B) > DoA(X,C) \quad (10)$$

$$DoA(X,B) - DoA(X,C) > DoA(X,C) \quad (11)$$

$$AveDoA(X,C) = \frac{\sum_{i=1}^{m} DoA(X_i,C_i)}{m} \quad (12)$$

5.2 評価結果

本稿で構築した概念ベースへの重み付与時の使用属性空間毎, 関連度算出時の最大使用属性数毎に評価した. 最大使用属性数を 5 個から 50 個まで 5 個刻みで評価を行った. 百科事典と国語辞典を用いた概念ベース *A*, *B* の最大精度の比較を表 6 に示す.

表 6 構築した概念ベースの最高精度比較

A	*B*
69.33	74.33

表 6 より百科事典と国語辞書を用いた概念ベースでは概念ベース *B* の精度が高いことがわかる.

重みの付与時の使用属性空間数毎, 関連度算出時の最大使用属性数毎の概念ベース *A* の結果を表 7, 概念ベース *B* の結果を表 8 に示す.

表 7 概念ベース *A* の精度評価結果

使用属性数	1 次（%）	2 次（%）	3 次（%）
5	10.33	25.67	25.00
10	48.33	66.33	60.33
15	56.67	**69.33**	68.33
20	45.33	65.33	68.33
25	42.00	60.67	61.33
30	39.67	61.00	62.33
35	40.00	57.67	61.67
40	42.67	56.00	59.67
45	42.67	52.00	60.67
50	40.67	50.00	58.67

表 8 概念ベース *B* の精度評価結果

使用属性数	1 次（%）	2 次（%）	3 次（%）
5	11.00	27.00	35.67
10	49.00	63.67	35.33
15	57.33	**74.33**	33.67
20	50.33	72.33	33.33
25	48.00	70.67	30.33
30	47.33	65.67	29.00
35	47.67	68.00	28.67
40	50.33	67.33	26.33
45	49.67	66.67	25.67
50	53.00	67.00	27.00

表 7 と表 8 より, 百科事典と国語辞典を用いた概念ベースでは, 概念ベース *B* に 2 次属性空間を用いた概念ベース *tf・idf* で重みを付与し, 最大使用属性数 15 個で関連度を算出したときに精度が最も高くなることがわかる.

6. 考察

本稿の実験から得られたデータをもとに, 獲得した概念, 属性, *X-ABC* 評価結果についての考察を述べる.

6.1 獲得した属性についての考察

概念ベース *A*, *B* の属性をみる. 概念ベース *A*, *B* の概念「ネットカフェ」の属性の評価を表 9 に示す. また, 最も

精度が高くなった 2 次属性空間を用いた概念ベース *tf・idf* で重みを付与したときの重み降順 15 位までの属性を太字で示す．評価は 3 人で行い 2 人が適切と判断した属性を適切であるとした．

表 9　概念「ネットカフェ」の属性と評価

	適切	不適切（雑音）
A	**マンガ喫茶**，**ネットカフェ**，施設，ネット，**インターネットカフェ**，環境，**アミューズメントセンター**，料金，深夜，**ネットカフェ難民**，**仮眠**，フリーター，**喫茶店**，**仮泊**，食事，接続，**アカウント**，利用，**パソコン**，**インストール**，**オンラインゲーム**	ニュース，**急増**，安い，長時間，没頭，手軽，住居，郊外，比率，自宅，**ホームレス**，韓国，略称，可能，人気，寸前，複合，兼ねる，死亡，率，理由，果て，円，年，時間，日本，中国，替り，多い，型，**割引料**，プレ，金，倒れ，持ち，整，込め，高い，隣，断，増，ホ，ラ，イ，てい，とも，ば，め，こ，ら，い，し，する
B	**マンガ喫茶**，**ネットカフェ**，施設，**ネット**，**インターネットカフェ**，環境，**アミューズメントセンター**，料金，深夜，**ネットカフェ難民**，**仮眠**，フリーター，**喫茶店**，**仮泊**，食事，接続，**アカウント**，利用，**パソコン**，**インストール**，**オンラインゲーム**	ニュース，**急増**，安い，長時間，没頭，手軽，住居，郊外，比率，自宅，**ホームレス**，韓国，略称，可能，人気，寸前，複合，兼ねる，死亡，率，理由，果て，円，年，時間，日本，中国，替り，多い，型，割引，持ち込む，殖える，高い，整う，倒れる，ば，いる，おる，する

表 9 より，概念「ネットカフェ」の属性数は概念ベース *A* が 74 個（適切 21 個），概念ベース *B* は 61 個（適切 21 個）と概念ベース *A* の方が不適切な語が多いことがわかる．また，概念「ネットカフェ」の重み降順 15 位までの属性の内，概念ベース *A* では適切 12 個，概念ベース *B* は適切 13 個であることから，概念ベース *B* の方が関連度を算出する際により適切な属性が使用されたと考えることができる．

表 9 での概念ベース *A* の概念「ネットカフェ」の重み降順 15 位までの属性で不適切とされた「割引料」は概念ベース *B* の概念「ネットカフェ」の属性には存在しない．これは見出し語「ネットカフェ」の説明文中にある「割引料金」が概念ベース *A* の属性獲得手法では「割引料」と「金」に，概念ベース *B* の属性獲得手法では「割引」と「料金」に分割されたためである．「割引料」とは金融用語であり，「割引料金」とは意味が異なる．概念ベース *A* の「割引料」と「金」より，概念ベース *B* の「割引」と「料金」の方が説明文中の「割引料金」の意味に近い分割（属性獲得）ができていると考えられる．

これらのことより，4.3.2 項の形態素を用いた最長前方一致による属性獲得手法の方が 4.3.1 項の最長前方一致による属性獲得手法よりも適切に説明文を分割して属性を獲得できる手法であることがわかる．また，適切に説明文を分割して属性を獲得できる手法を用いたことで，概念ベース *B* は概念ベース *A* より *X-ABC* 評価の精度が高くなると考えられる．

6.2 重みの付与についての考察

1 次属性空間を用いた場合，概念の 1 次属性中における属性の出現回数は 1 回であるため，概念ベース *tf* の値は 1 になる．そのため，1 次属性空間を用いた場合は，概念ベース *idf* の値のみによって重みを付与していることになる．概念ベース *idf* の値は概念ベース中の多くの概念の属性に出現するほど小さくなる．ここで，概念ベース *B* の 1 次属性空間での概念ベース *idf* の値昇順 10 位までの概念とその全概念の 1 次属性中の出現回数，概念ベース *idf* の値を表 10 に示す．

表 10　概念ベース *idf* の値昇順 10 位までの概念

概念	出現回数	*idf*
する	118841	0.891504
いる	52021	2.08337
ない	41235	2.4186
もの	33999	2.69697
一	32851	2.74653
年	27012	3.02886
人	24077	3.19481
二	17397	3.66363
中	15292	3.84969
的	15223	3.85621

表 10 より，概念ベース *idf* の値が小さくなるほど，その概念は概念ベース全体においても，他の概念を特徴付けるのに適さないと考えられる．しかし，概念ベース *idf* の値のみでは，概念の特定性しか考慮できず，その属性が概念にとって重要な属性であるかを示すことができないと考えられる．そのため，1 次属性空間を用いた場合，*X-ABC* 評価の精度が概念ベース *tf* の値が属性毎に異なる 2 次属性空間を用いた場合よりも低くなったと考えられる．

2 次属性空間を用いた場合，「する」や「いる」といった「ネットカフェ」の意味特徴を表すには不適切な属性が下位のまま，「パソコン」や「喫茶店」といった「ネットカフェ」の意味特徴を表す属性が 1 次属性空間を用いた場合よりも上位に存在している． 概念ベース *idf* の値のみでは，「パソコン」という「ネットカフェ」の意味特徴を表す属性が「比率」という「パソコン」より「ネットカフェ」の意味特徴を表すといえない属性より下位にある．また，「する」や「いる」といった属性が下位にあることより，2 次属性空間を用いた場合も 1 次属性空間を用いた場合と同様，概念ベース *idf* の値が小さい概念は他の概念を特徴付けるのに適さないといえる．

ⓒ2016 Information Processing Society of Japan

2 次属性空間を用いた場合, 2 次属性まで展開するため，概念ベース *tf* の値が属性毎に異なる．情報検索における *tf* は「文書空間で何度も繰り返し言及される概念は重要な概念である」という仮定のもとに索引語の頻度で網羅性を考慮する尺度である．それを基に概念ベース *tf* は「*n* 次属性空間で何度も繰り返し出現する属性は重要な属性である」という仮定のもとに *n* 次属性空間での出現回数によって属性の網羅性を考慮する尺度である．一方で,「一般に頻度の高い語は, 文書を特徴付ける上であまり役に立たない」[10] と言われている．2 次属性空間を用いた概念ベース *tf* の値は「する」や「いる」といった属性の値が大きくなっている．概念を特徴付けるのに適さない属性の数値が大きいことがわかる．そのため，概念ベース *tf* と概念ベース *idf* のの積により属性・重みの並び順のバランスがよくなり精度が高くなったと考えられる．

3 次属性空間を用いた場合，大きい概念ベース *tf* の値に概念ベース *tf*・*idf* の値が引きずられ，不適切な属性の重みが大きくなったと考えられる．そのため，不適切な属性の関連度に対する影響が他の属性空間を使用した場合よりも大きくなり，精度が低くなったと考えられる．

1 次属性空間から 3 次属性空間での属性の並びと重みのより, 各使用属性空間で概念ベース *idf* の値が, 概念の特徴付けに役に立つ数値であると考えられる．そこで，概念ベース *idf* のみで重みを付与した概念ベース *B* の使用属性空間毎に *X-ABC* 評価結果の最大精度を表 11 に示す.

表 11　概念ベース *idf* を重みとした概念ベース *B* の精度

1 次	2 次	3 次
57.33	68.33	73.33

表 11 より重みの付与に使用する属性空間が大きくなるほど，*X-ABC* 評価結果が高くなることがわかる．これにより，概念ベース *idf* が示す概念の特定性も使用する属性空間が大きいほど精度が高くなると考えられる．

表 11 と表 6 より，2 次属性空間を用いた場合は概念ベース *tf*・*idf* による重み付与をした場合の方が概念ベース *idf* のみで重みを付与した場合よりも精度が高いことがわかる．逆に，3 次属性空間を用いた場合は概念ベース *tf*・*idf* による重み付与をした場合よりも概念ベース *idf* のみで重みを付与した場合の精度が高いことがわかる．

これにより，2 次属性空間を用いた場合では，概念ベース *idf* で概念の特定性のみではなく概念ベース *tf* で属性の網羅性も考慮でき精度が高くなったが，3 次属性空間を用いた場合では概念ベース *tf* が足を引っ張り精度が低くなったと考えられる．これは概念ベースでの雑音となりやすい概念ベース *idf* の値が小さい属性（概念）の概念ベース *tf* の値が高くなりやすく，さらに，使用属性空間が大きくなるほど出現頻度が爆発的に大きくなり，概念ベース *tf* と概念ベース *idf* のバランスが崩れてしまうためであると考えられる．

これらのことから，概念ベース *idf* の値が小さい概念および，概念ベース *tf* の値が高い属性を閾値や外れ値を用いて除去することで，本稿で構築した概念ベースの精度向上が期待できる．

7. おわりに

本稿では百科辞典と国語辞書を用いた概念ベースの構築手法について述べた．概念を網羅的に集めるためには，属性の網羅性もまた重要である．このことから新しい語が登録された概念ベースの情報源には幅広い知識によって概念ベースを構築するべきであると考えた．

概念ベース *tf*・*idf* による重みの付与では 2 次属性空間を使用したときが最も精度が良かった．また，本稿では概念・属性の獲得手法，情報源を見直すことで，*X-ABC* 評価手法での精度が既存の百科事典を用いた概念ベースよりも向上した．考察より，概念ベース *tf*・*idf* で重みを付与する際，概念ベース *tf*（属性の網羅性）と概念ベース *idf*（概念の特定性）のバランスが重要であると考えられる．また，概念ベース *tf* または概念ベース *idf* を用いて雑音となる概念，属性を削除することで，本稿で構築した概念ベースの精度向上が見込めると考えられる．

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] 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] 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.
[3] 大竹慎吾，芋野美紗子，土屋誠司，渡部広一．百科事典を用いた概念ベースの構築．人工知能学会知識ベースシステム研究会資料，SIG-KBS-B203-03，pp.15-20，2013.
[4] 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.
[5] 奥村紀之, 荒木孝允, 渡部広一, 河岡司．概念属性の動的評価に基づく概念関連度計算方式．情報処理学会，E-033，pp.223-226，2006.
[6] 「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.
[7] 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.
[8] 新村出．広辞苑第四版，岩波書店，1996.
[9] ChaSen -- 形態素解析器，http://chasen-legacy.sourceforge.jp/，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)
[10] 徳永健伸（編），“情報検索と言語処理”，東京大学出版会，1999.

ⓒ2016 Information Processing Society of Japan