7B-05

新語の獲得に対応した概念ベースの構築方式

三品　賢一†　　土屋　誠司‡　渡部　広一‡

同志社大学大学院理工学研究科†　　同志社大学理工学部‡

1. 研究背景

会話ロボットに必要な能力として，会話から様々な事柄を連想する能力が挙げられる．連想技術に有用な資源として概念ベース[1]がある．従来の概念ベースの構築方式では，形態素解析器が新語を検出できないために，新語を概念として扱うことができなかった．実用的な連想技術を実現するためには，新語の概念を扱える必要がある．そこで本研究では，単語分かち書き辞書 mecab-ipadic-NEologd[2]を用いた形態素解析を導入し，概念価値に基づく重みづけを用いた概念ベース構築方式を提案する．これにより，日々生まれる新語を概念ベースに取り込むことが容易になる．性能評価のため，本稿では従来研究で用いられた百科事典からの新語の概念化を行う．

2. 関連技術

2.1. 概念ベース

概念ベースとは，電子化された国語辞典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．ある語を概念として定義し，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する．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，式(1)のように定義される．

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

ここで，概念自身が持つ属性を 1 次属性と呼ぶ．すべての属性は概念としても登録されているため，1 次属性からも属性を導くことができる．この導かれた属性を，もとの概念に対する 2 次属性と呼ぶ．また 1 次，2 次属性の集合を 2 次属性空間と呼ぶ．概念ベースは，n次の属性の連鎖集合の構造となっている．

2.2. 関連度計算方式

関連度計算方式[3]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．

Construction of Concept-Base Corresponding to Acquiring Neologism

†Graduate School of Science and Engineering, Doshisha University

‡Faculty of Science and Engineering, Doshisha University

3. 従来研究

白石らは，時事用語や専門用語を概念として獲得するために，複合語の概念・属性を考慮した百科事典・国語辞典からの概念ベース構築方式を提案している[4]．この方式では，概念は国語辞典や百科事典[5]の見出し語から，属性は見出し語の説明文から獲得する．概念として獲得する語は“名詞”，“動詞”，“形容詞”，“複合語”，“アルファベットを含む語”，“カタカナを含む語”となる．概念として獲得する“複合語”は，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”となる語となる．“接頭詞＋名詞＋名詞”のように，上記の規則が連続していれば，それらも一つの複合語となる．

属性の重み付けには，概念ベース$tf \cdot idf$を用いる．ある概念Aの属性aの重み$W(A,a)$は以下の式(2)で求める．

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

ここで$tf_{A,n}(a)$は概念Aのn次属性空間に概念aが出現する頻度，V_{all}は概念ベース内の概念の総数，$df_n(a)$はn次属性空間に概念aを属性として持つ概念の数を表す．白石らは構築した概念ベースの性能評価実験の結果に基づき，$\mathrm{n} = 2$を用いている．

4. 提案手法

白石らによる複合語の抽出処理を行うことで，新語を概念として獲得できる可能性がある．しかし，白石らの手法では適切な複合語が抽出できるとは限らない．例えば，“コンピュータ犯罪防止が大きな問題となっている”という文からは，“コンピュータ犯罪防止”を一つの複合語として抽出する．百科事典の見出し語として“コンピュータ犯罪”が記載されていることからも，“コンピュータ犯罪”までを一つの複合語として抽出すべきである．そこで本研究では，形態素解析器 MeCab[6]と，単語分かち書き辞書 mecab-ipadic-NEologd を用いて，概念と属性を

獲得し，概念ベースを構築した．この辞書を用いることで，固有名詞や複合名詞などの長い単語を1単語として分かち書きすることができる．上記の“コンピュータ犯罪防止”は，“コンピュータ犯罪”と“防止”に分かち書きすることができる．本辞書はWebから新語を含む様々な語を収集して構築されるため，新語の検出性能の向上が期待できる．また，属性の重みづけは従来研究と同様に，概念ベース$tf \cdot idf$を用いた．

5. 評価実験

概念や属性の獲得方法を変えることによる概念ベースの性能の変化を評価するため，以下の内容の実験を行った．

5.1. 概念ベースの構築

まず複数の国語辞典から基本となる概念ベースを構築した．そして，この概念ベースに対し，百科事典から概念の追加を行った．白石らの複合語抽出手法を用いて概念と属性を獲得し，構築した概念ベースをCB_{comp}，提案手法を用いて構築した概念ベースを$CB_{neologd}$とする．

5.2. X-ABC評価

X-ABC評価では，任意の基準概念をXとし，概念Xと関連が強い概念をA，関連がある概念をB，関連がない概念をCとする．概念X, A, B, Cからなる組を複数用意し，次の条件式を満たすものを正解とする．

$$DoA(X,A) > DoA(X,B) > DoA(X,C) \quad (3)$$

ここで$DoA(X,A)$は概念XとAの関連度を表す．評価実験では，基準概念として百科事典の見出し語をランダムに選び，人手で概念A, B, Cを設定した200組の評価データを作成した．

5.3. 実験結果

実験結果を表1に示す．CB_{comp}に比べて$CB_{neologd}$では精度が9.5%向上した．また有意水準5%で符号検定を実施したところ，$Z = 2.959182 > 1.96$となり，有意に差があることが確認できた．

表1：X-ABC評価の正解率（%）

CB_{comp}	$CB_{neologd}$
43.5	53.0

6. 考察

基準概念Xを“ＪＩＣＡ”，概念Aを“組織”，概念Bを“協力”，概念Cを“布団”としたときに，CB_{comp}を用いた評価では成功し，$CB_{neologd}$では失敗していた．原因は，CB_{comp}での“ＪＩＣＡ”の属性には“無償資金協力”などが含まれており，“協力”単体での属性は存在していなかった．$CB_{neologd}$では“協力”単体での属性が存在していたため，“ＪＩＣＡ”と“協力”の関連度が上昇し，評価に失敗していた．このようなケースの改善方法としては，概念（見出し語）とシソーラス距離が近い属性の重みを大きくすることが考えられる．今後の課題として，シソーラス距離を考慮した重み付け手法を検討し，精度の向上を図る．

7. まとめ

本研究では新語の獲得に対応した概念ベースの構築方法として，単語分かち書き辞書 mecab-ipadic-NEologd を用いた概念と属性の獲得を行い，概念ベースを構築する手法を提案した．評価実験では精度が 9.5%向上し，提案手法が有効であることを示した．

謝辞

本研究の一部は，JSPS科研費 JP16K00311の助成を受けて行った．

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] 佐藤敏紀，橋本泰一，奥村学，“単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討”，言語処理学会 第23回年次大会 発表論文集，pp.875-878，2017.

[3] 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.

[4] 白石卓也，芋野美紗子，土屋誠司，渡部広一，“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”，情報科学技術フォーラム FIT2014，pp.213-214，2014.

[5] 「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.

[6] 工藤拓，“MeCab : Yet Another Part-of-Speech and Morphological Analyzer”，http://mecab.sourceforge.net/，2005.