E-005

概念ベースを用いた複合語の自動的属性取得法

Automatic attribute acquisition method of compound word using Concept-Base

後藤 敏貴 † 奥村 紀之 † 渡部 広一 † 河岡 司 †
Toshitaka Goto Noriyuki Okumura Hirokazu Watabe Tsukasa Kawaoka

1. はじめに

人間は会話を行う際に，単語と単語の関連性を見出しながら理解し，返答している．この高度な会話システムをコンピュータ上で実現する手段として，連想メカニズムが提案されている．連想メカニズムは，ある単語（概念）に対して関連の深い語（属性），並びに各属性の重要性を示す重みの集合によって定義された概念ベース[1]によって実現されている．しかし，概念ベースには「記者会見」などのように複数の形態素で構成された複合語の概念が十分には登録されていない．そのため複合語概念を動的に生成する手法が必要となる．現在，複合語の属性取得手法として，「関連度優先法」と接尾語の置き換えによって実現されているが，多くの問題点が残されている．

本稿では，2 つの入力後から自然な連想を行う「二語連想システム」に基づく「共通概念法」や，シソーラス[4]を利用することにより，複合語の属性の質を高め，また，平均関連度法を利用することにより，不適切な属性の排除を行うことで，精度のよい複合語の属性取得方法を提案した．

2．概念ベースと関連度

2.1 概念ベース

概念ベースは，電子化された複数の辞書を形態素解析することで抽出した概念表記や属性によって機械的に構築された大規模知識ベースである．

ある概念Aをその語と関連が強いと考えられる語a_iと重みw_iの対の集合で定義する．

概念A＝｛ (a_1, w_1) ， (a_2, w_2) ， ･･･ (a_n, w_n) ｝

ここで，属性a_iを概念Aの 1 次属性と呼ぶ．また，属性a_iも概念ベースに登録されている 1 つの概念である．従って，a_iからも同様に属性を導くことができる．a_iの属性a_{ij}を概念Aの 2 次属性と呼ぶ．

2.2 意味関連度計算方式[2]

2 つの概念 A，B の意味関連度 $MR(A,B)$ は，概念 A，B の二次属性を利用し一致度を求め，一致度の和が最大になるように一次属性の組み合わせを作る．ここで，一次属性が完全一致する場合は特殊処理とする．概念ベースには約 9 万の概念が存在し，属性が一致することは稀である．従って，属性の完全一致を特別処理にすることにより，属性が一致した場合の評価を大きくすることができる．

概念 A，B の意味関連度 $MR(A,B)$ は，

$$MR(A,B)=\sum_{i=1}^{T} Match\ (a_i',b_i')\times(v_i'+u_i')\times\frac{1}{2}$$
$$\times(\min(u_i',v_i')/(\max(u_i',v_i'))$$
$$Match\ (a_i',b_i')=\sum_{a_{ip}'=b_{iq}'}\min(u_{ip}',v_{iq}')$$

$(u_{ip}', v_{iq}'$はa_i', b_i'の一致する一次属性の重み)

となる．

2.3 意味共起関連度計算方式[3]

意味的関連度計算方式と共起関連度計算方式[2] を合成して評価する方式を以下で定義し，この計算方式を意味共起関連度計算方式と呼ぶこととする．

$$MCR(A,B)=\frac{MR(A,B)+C_w\times CR(A,B)}{1+C_w}$$

$MR(A,B)$：意味関連度 $CR(A,B)$：共起関連度
Cw：共起関連度重み

意味共起関連度計算方式では，意味関連度MRの値に加え，共起情報から得られる共起関連度に定数重み Cw を乗じたものを付与することにより，演算を行う．実験により Cw=0.90 が最適とされているので，本稿ではこの値を用いる．

3. 複合語についての調査

3.1 複合語の構造

複合語とは，複数の形態素で構成されている語である．複合語を構成する各形態素を構成語と呼ぶ．複合語の構成例として表 1 のような語が挙げられるが，本研究では名詞のみを構成語とする複合語を対象とした．

表 1　複合語の構造

構造	例
名詞 ＋ 名詞	記者会見，総合運動公園
形容詞＋ 形容詞	青白い，ずる賢い
名詞 ＋ 形容詞	肌寒い，手厳しい
形容詞＋ 名詞	汚い海，広い空
動詞 ＋ 動詞	食べ歩く，張り詰める

3.2 概念ベースに含まれる割合

概念ベース中の概念約 9 万語を形態素解析して，複合語概念を抽出したところ，3 万 5556 語であった．しかし，「投資家」はないが「機関投資家」はあり，「合理化」はないが「産業合理化」はあるなど，複合語の観点では整理されていない．また，40 人にアンケートを行い，実際に頻繁に使われる複合語から 125 語について調査を行ったところ，「本格的」「基本的」「関係者」など，約 80%にあたる 101 語が概念ベース未登録であることがわかった．従って，概念ベースに未登録な複合語について対処する機

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

能が必要である．

3.3 複合語の構成品詞

概念ベースに定義されている複合語概念については，図1のような内訳となる．

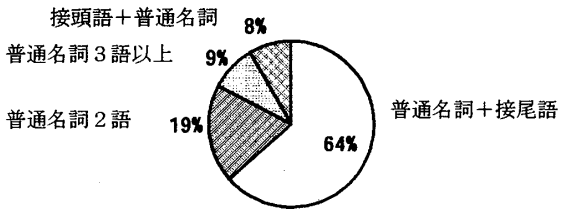

図１　複合語の種類

本研究では，「接頭語＋普通名詞」と「普通名詞＋接尾語」と「普通名詞２語」のみに限定し，3 語以上の構成語を持つ複合語については扱わないものとする．また，固有名詞や略語に関しては，固有名詞の多義性や略語から元の語を復元することが困難であることなど，特有の問題点があるため対象外とした．

4. 複合語属性の取得方法

4.1 関連度優先

複合語を形態素に分割し，構成語１と構成語２とする構成語 1 の属性と構成語 2 との関連度(MR)を取り，関連度の高い順に前の構成語の属性に採用する．構成語２の属性と構成語１も同様にする．これによって構成語の属性は片方の構成語に関係する語が優先される事になる．

4.2 接尾語の特徴を利用

接頭語・接尾語は以下の特徴から，別の語に置き換える必要があると考えられる．

① 接頭語・接尾語と成り得る語は，接頭語・接尾語以外の語と比べて属性数が多く，その属性は複数の意味が混在している．

② 接頭語・接尾語の概念として適切な属性の多くは「性質」「状態」「様子」「特徴」などを表している．

例えば，「基本的」の「的」に注目すると，「目当て」や「焦点」など様々な多義性があることがわかる．「基本的」の「的」という接尾語が表す属性としては，「性質」や「状態」等が正しいと考えられる．そこで 56 個の接頭語・接尾語について，置き換えのデータを作成した．この一例を表２に載せる．

表２　接頭語・接尾語の置き換え例

接頭語・接尾語	置き換え名詞
性	性質
剤	薬剤
症	症状
病	病気
費	費用

4.3 評価

「普通名詞＋普通名詞」と「接頭語＋普通名詞」と「普通名詞＋接尾語」の複合語を人手で 100 個を作成し，これを評価セットとする（表 3）．評価方法は目視で行い，生成した語の中で，複合語に対して適切な属性だと思われる語の割合を求め精度とする．

表３　評価データセット（一部）

預金通帳	老人ホーム	心霊現象
恋愛経験	研究機関	大衆文学
市場調査	ラーメン店	人道支援
プラスネジ	資格試験	市民権

精度は 46.7%，平均属性数は 26.3 個得られた．精度は 46.7%と低い値になり，さらなる向上の余地がある．関連度優先法では，「メロンパン」といったような構成語は，メロンの属性とパンに関連が見られない複合語に対して正しい属性が得られない．このため，新たな考案手法として二語連想システムによる属性取得手法，後部構成法を提案する．この手法については次章に説明する．

5. 構成語の特性を生かした属性取得方法

5.1 共通属性法

二つの構成語を概念とし，その属性を概念ベースから獲得する．両方の概念に含まれる共通する属性を出力する手法である．

5.2 共通概念法

共通属性法では，構成語の属性数が少なければ共通する属性が出力されない場合がある．その問題を解消するために，二つの構成語を属性と見なし，その二語も同時に属性として保持する概念を出力する手法である．この共通属性法と共通概念法を用いたものを二語連想と呼ぶこととする．

5.3 後部構成語の属性利用

後部構成語とは，複合語の後部の構成語である．例えば，「携帯電話」（「携帯」と「電話」が構成語）では，後部構成語は「電話」である．前部の構成後が後部の構成語の説明をするという特徴を利用した．

6. 二語連想と後部構成語属性利用の手法の評価

2 つの複合語属性取得手法の精度比較を行った．この結果を図 2 に示す．テストデータは 4.4 節で用いたものである．

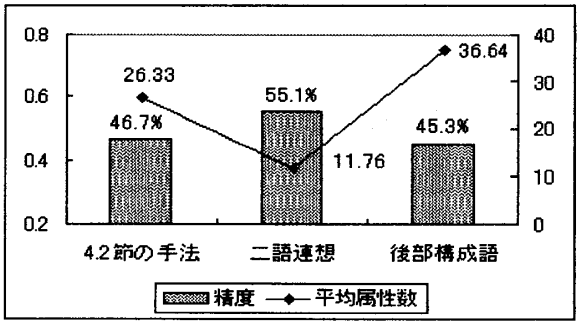

図２　4.2 節で述べた手法との精度比較

この結果，「改良手法」の精度は上昇したが，平均属性数は減少した．また，「後部構成語の属性利用」では精度は減少したが，平均属性数は増加した．この原因として，「後部構成語の属性利用」によりその属性が複合語に対して正しいものになっているとは限らず，概念ベースの質が高くない，不適切な連想語（雑音）が出力されるといったことが挙げられる．そこで，雑音を除去する（雑音除去）必要がある．また，雑音除去により属性数は減少してしま

うので，シソーラスを用いて属性数の増加を試みる．

7. 平均関連度法とシソーラスの利用.

7.1 平均関連度法を用いた雑音除去

雑音除去手法として関連度計算を用いた平均関連度法を用いた雑音除去手法を提案する．関連度計算手法には，意味関連度計算手法と意味的共起関連度計算手法を用いる．取得した属性から構成語の関連度を求め，平均を算出する（この平均を平均関連度と呼ぶ）．それぞれの平均関連度の平均を閾値にして雑音除去を行う．例として老人ホームの意味関連度を用いて求めた平均関連度の値を表 4 に示す．

表 4　意味関連度による平均関連度

	老人	ホーム	平均関連度
養老院	0.36	0.26	0.31
老人	1.00	0.07	0.54
本塁	0.03	0.11	0.07

算出した平均関連度の平均を求める．その平均を求めると 0.30 になる．この結果，平均関連度が 0.30 以上の養老院と老人が老人ホームの属性として採用される．

7.2 シソーラスの利用

シソーラスとは，索引、検索用の構造化された統制語彙集のことで，本研究では日本語語彙大系から作成された一般名詞の意味的用法を表す 2710 個の意味属性（ノード）の上位－下位関係，全体－部分関係が木構造で示されたものである．ノードに属する名詞として約 13 万語（リーフ）が登録されているものを用いている．複合語のそれぞれの構成語の共通親を検索して，そのステップ数が少なければ親ノードに関する属性を返す．ステップ数とは，そのリーフの親ノードを参照することで，一つ上の階層のリーフを参照することを 1 ステップとする．もし，ステップ数が多ければその複合語にはあまり関係がない構成語で構成されていると考えられるため，後部構成語の親ノードだけを用いる（5.3 節参照）．本研究では，検索ステップ数について目視で評価したところ，3 ステップ以内に共通親ノードが存在すると関連性があると見られたので，３ステップを閾値とする．

8. 雑音除去後の評価

8.1 上位の属性で精度評価

複合語の中には，属性が多く獲得できるものがあり，その属性には複合語に対して正しいものになっているとは限らず，不適切な属性（雑音）が含まれている可能性がある．そこで，属性が多く獲得できる複合語は，平均関連度の上位 15～30 個を獲得する．そのため，本研究では，目視で評価したところ，最も精度が高かった上位 20 個を用いている．テストデータは 4.4 節で用いたものを使用する．

8.2 目視による評価

9.1 節の評価結果を図 3 に示す．図中の R_wr は意味関連度計算手法を用いた平均関連度法による雑音除去手法で，R_co は意味的共起関連度計算手法を用いた平均関連度法による雑音除去手法である．TS はシソーラスによる属性追加方法を表す．max は生成した属性の中で関連度の値が上位である部分を抽出する手法で，数値は取得する属性数の上限を示す．例えば，max20 とは複合語 1 語に対して上位 20 個の属性を獲得することを表す．

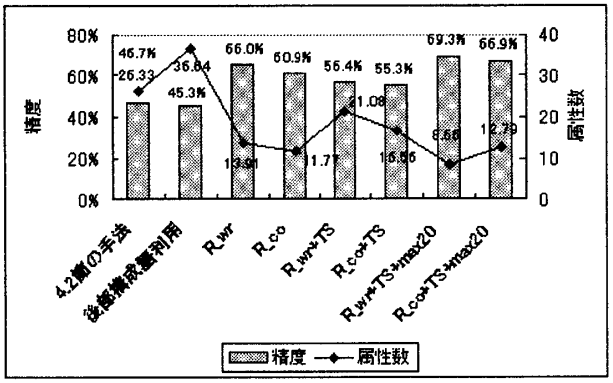

図 3　4.2 節の手法との精度比較

図 3 より，精度は「R_co＋TS＋max20」が最も良く，4.2 節ので説明した手法と比べて約 20%向上した．しかし，複合語 1 語に対する平均属性数は「R_co＋TS＋max20」では約 8 語とかなり減少し，4.2 節で説明した手法と比べると約 18 語減少している．いくら精度が良い属性が獲得できたとしても平均属性数が少ないと幅広い連想処理をうまく行うことができない．また逆に，平均属性数は多いが，精度が悪い場合，問題である．平均的に見て，精度がやや高く，平均属性数は約 12～15 個の手法が望ましい．これらより，精度と複合語 1 語に対する平均属性数の両面から見て，「R_wr＋TS＋max20」が有効な手法であるといえる．

9. おわりに

本稿では，会話や文章に頻繁に使われる複合語を対象にし，属性の生成法は「共通概念法・後部構成語の属性利用・シソーラスの利用」，雑音除去手法では「平均関連度法」を提案した．新しく属性数を増やすことにより，雑音も増加したが，「平均関連度法」を用いることで精度の良い属性を獲得することができた．構成語間の関連性を使用するより，有効な属性が取得でき精度の向上を実現した．

今後は概念ベースを精錬した上で，再度精度を評価し，また形容詞や動詞を含む複合語についても拡張し，より幅広い語に柔軟に対応していくことが望ましい．

謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクト“ 知能情報科学とその応用”における研究の一環として行った．

参考文献

[1] 広瀬幹規，渡部広一，河岡司：概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法，信学技報，NLC2001-93，pp.109-116，2002.

[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] 青田正宏，渡部広一，河岡司：概念の意味・表記と共起情報を用いた関連度計算方式．同志社大学理工学研究報告.vol45, No.1, pp.23-34, 2004.

[4] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦（編）日本語語彙大系．岩波書店，1997.