E-029

新聞記事からの複合語概念表記の獲得
Acquisition of Compound Words in Newspaper Articles

柳瀬 秀夫† 芋野 美紗子† 土屋 誠司‡ 渡部 広一‡
Hideo Yanase Misako Imono Seiji Tsuchiya Hirokazu Watabe

1 はじめに

人には，ある語から関連性のある語を連想する能力があり，日常の会話で役立てている．この連想の能力をコンピュータに持たせることができれば，言葉を理解し人のように返答ができる会話システムの実現に近づくと考えられる．コンピュータが連想を行うためには人と同様に語と語の関係に関する知識を保持しておくことが必要だと考えられる．そこで，言葉の知識を一定の形式で集約し保持しておくことを目的に構築された知識ベースとして概念ベース[1]が存在する．

概念ベースには人が一般的に使用する語が登録されており約 12 万語が存在する．この概念ベースの特徴として，複数の形態素からなる語である複合語の登録数が少ないということが挙げられる．実際，40 人を対象としたアンケートにより収集された頻繁に使われる複合語 125 語の内, 約 80% の語が未登録であることが分かっている[2]. 複合語は会話や文章で常用されるため言葉の知識としての必要性が高い．本稿では複合語を概念ベースに登録することを前提とした，新聞記事からの複合語概念表記の獲得手法について述べる．

2 概念ベース

概念ベースは複数の電子国語辞書から機械的に構築された知識ベースに，新聞などから言葉を追加したものである．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．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(1)式のように定義される．

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

例を挙げると，概念「雪」は(2)式のように定義される．

概念「雪」＝{(雪,0.61),(白い,0.30),･･･,(下る,0.01)} (2)

3 複合語概念表記の獲得手法

複合語概念表記(以下，単に「複合語」と記す)を獲得する際には品詞情報を用いる．ただし品詞情報のみを利用したのでは，概念ベースに登録すべきでない語が数多く獲得されてしまう．そこで表記などの他の情報を利用することにより概念ベースに登録すべきでない語の除去を行った．以下に使用する品詞情報及び具体的な除去手法について示す．

3.1 使用する品詞情報

複合語を獲得するソースとして，「朝日新聞記事データ<邦文>2005 年版」[3](以下「朝日新聞」と記す)を使用する．茶筌[4]で「朝日新聞」に対し形態素解析を行い，表 1 のような前後関係にある 2 語の部分を複合語概念表記候補(候補語)として獲得する．さらに候補語は再度，形態素解析を行い品詞情報が不変のものを選出する．

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

表 1 候補語とする品詞の並びと例

品詞の並び	例
接頭語+名詞	不登校(不+登校)
名詞+名詞	電波時計(電波+時計)
名詞+接尾語	人間性(人間+性)
名詞+助動詞「ない」	申し訳ない(申し訳+ない)
動詞(連用形)+動詞	書き間違う(書き+間違う)
動詞(連用形)+形容詞-非自立	使いづらい(使い+づらい)

3.2 言葉として不成立な語の除去

「す旅」,「ど池」などの，言葉として不成立な語が形態素解析の誤りにより獲得されるためこれらの除去を行った．

まず，平仮名のみから成る複合語は形態素解析を誤っている可能性が高いとして候補語から除去した．また，接頭語や接尾語についても表記が平仮名だけの場合，一般的でない語を形成する傾向があるため除去した．例えば「み完成」といった語などがこれにあたる．

次に[接頭語+名詞], [名詞+名詞], [名詞+接尾語]の直前や直後に表記がカタカナのみから成る語がある場合は形態素解析を誤っている可能性が高いため除去した．例えば「スペシャルオリンピックス」という語は形態素解析により「スペシャル(名詞)＋ オリンピック(名詞)＋ ス(動詞)」のように区切られる．ここで「スペシャル ＋ オリンピック」の部分を抜き出すと存在しない語を獲得することになる．このように直前や直後に表記がカタカナのみから成る語がある場合，区切りを誤っていると考え候補語から除去した．

3.3 出現頻度の閾値による除去

候補語で出現頻度が低い語は一般的に使われにくい語であると考えられる．そこで出現頻度に閾値を設け，閾値未満の語は除去した．閾値には候補語で既に概念ベースに存在する語の出現頻度の平均値を用いた．

3.4 不要語リストによる除去

概念ベースには不要と考えられる語(不要語)を形成しやすいと判断できる語を集約したリスト(不要語リスト)を作成し除去を行った．例えば，接頭語の「同」という語を有する複合語は文章から切り離されると意味がなくなるため不要語だといえる．よってこの場合「同」という語を不要語リストに含める．このような語を「朝日新聞」を調査して集め，仮の不要語リストとした．これをアンケート調査し，仮の不要語リストに集めた語が形成する複合語を本当に不要語としてよいのかを確かめた．その上で最終的な不要語リストを決定した．不要語リストの例を表 2 に示す．

表 2 [接頭語+名詞]の不要語リスト

リスト名	グループ	例
不要接頭語	指示的語	同，各
	不成立語にしやすい語	御，ド
	数字	ひと，ふた
	時間の流れを表す語	旧，新
不要名詞	政治的略語	新，現，元

不要語リストは表 2 の他にも[名詞+名詞]のものが 3 つ，[名詞+接尾語]のものが 2 つ，動詞の部分のものが 1 つある．表中の「リスト名」とは複合語の構成部分を示しており，「グループ」とはリストの語がどのような不要語を構成するのかを示している．これらの不要語リストに存在する語をリスト名の部分に持つ語は除去を行う．

不要語リストは「朝日新聞」3 ヵ月分において除去された語の 100 語のサンプルを 5 人に見てもらい，それぞれの語が不要語であるかを多数決で評価しその割合により精度を求めた．評価結果は平均で 98.5%と高精度であった．

3.5 固有名詞の除去

人名や地名といった固有名詞は名称が同じものが複数ある場合や，その存在期間が短期間な場合があるため，一般性の高い語を収録する概念ベースには登録すべきでないと考え除去を行った．

まず，句読点以外の記号として“「」”や“『』”などがあるが，これらの記号で囲まれて存在する語は除去した．これは“「」”といった記号が，文書において作品タイトルなどの固有名詞を囲む形で存在する可能性が高いと考えられるためである．

次に，形態素解析において本来[固有名詞+普通名詞]と判断されるべきところを[普通名詞+普通名詞]と判断されたような語の除去を行った．例えば「三陸沖」は「三陸(名詞-一般)」と「沖(名詞-一般)」のように誤って区切られ，固有名詞と判断されない．そこで「朝日新聞」を検索して固有名詞に接続しやすい普通名詞を集約したものをリストとし，リストにある語を後部とする複合語は獲得しないこととした．例えば「沖」という語は「秋田沖」や「能登沖」という語に見られるように固有名詞に接続する性質がある．そこで後部に「沖」という語を有する複合語を除去することにすれば，「三陸沖」は獲得されないということになる．ただし，固有名詞に接続する頻度が高い名詞を集めるだけでは普通名詞にも高い頻度で接続する名詞までリストに含めることになる．よって固有名詞や普通名詞に接続する頻度の割合ごとにリストによる除去の精度を求め，複合語の獲得に用いるリストを決定した．

精度は「朝日新聞」3 ヵ月分において除去された語のサンプルを 100 語調査し，固有名詞と出現頻度が 1 回の語の割合により求めた．出現頻度が 1 回の語というのは 3.3 節を踏まえ，除去されても問題がないものと考えその割合を精度に含めている．リストによる除去の評価結果を表 3 に示す．表 3 より，精度として 7 割程度を維持している，固有名詞への接続率が 7 割以上の語を獲得のためのリストに含めることとした．

表 3 リストにより除去された語の内訳と精度

固有名詞への接続率（リスト）	固有名詞	出現頻度が 1 回の語	精度
9 割以上	34%	48%	82%
8 割以上 9 割未満	25%	45%	70%
7 割以上 8 割未満	20%	52%	72%
6 割以上 7 割未満	20%	47%	67%
5 割以上 6 割未満	6%	57%	63%

4 獲得結果

3 章の手法により「朝日新聞」1 年分から獲得した複合語の語数を表 4 に示す．獲得語数は計 9629 語であり，既に概念ベースにある語や表記違いの語を除くと 8133 語となった．

表 4 獲得結果

複合語	獲得語数
接頭語+名詞	292 語
名詞+名詞	6468 語
名詞+接尾語	1512 語
名詞+助動詞「ない」	17 語
動詞(連用形)+動詞	351 語
動詞(連用形)+形容詞-非自立	989 語

5 精度評価

まず，全獲得語の 9629 語に対して評価を行った．それぞれの複合語につき最大 100 語のサンプルを取り出し，概念ベースに登録すべき複合語(正解語)であるかを 5 人に判断してもらい多数決をとって評価した．表 5 に示す精度はその正解語の割合により求めたものである．[動詞+動詞]，[動詞+形容詞]については不適切な語が多少混ざっていたものの，いずれも高精度であるということがわかる．

表 5 獲得語の精度

複合語	精度	正解語の例
接頭語+名詞	100%	副社長，仮契約
名詞+名詞	100%	恒例行事
名詞+接尾語	100%	贈呈式，改革派
名詞+助動詞「ない」	100%	味気ない
動詞(連用形)+動詞	95%	握り返す
動詞(連用形)+形容詞-非自立	92%	割れにくい

次に，獲得時間について評価を行った．評価方法として，本稿の手法を実装したプログラムを利用した場合の所要時間と，人手で新聞記事から複合語を抜き出す作業の所要時間を比較した．まず，プログラムを利用した場合，「朝日新聞」1 年分から複合語を獲得するのに 11 分 14 秒を要した．比べて人手の場合には，人間が 1 時間に処理できる記事の量を実験して調べた結果，作業を連続して行った場合 97 日と 4 時間 36 分程度かかることが分かった．よってプログラムを利用すると人手よりも 12462 倍高速に処理できるということになる．

以上の精度評価から本稿の手法は，高精度かつ短時間で複合語を獲得することが可能であることがわかり，有効であることを示せた．

6 おわりに

本稿では新聞記事からの複合語概念表記の獲得手法を提案した．結果として，平仮名やカタカナという表記情報の利用，出現頻度の閾値設定，リストの使用によって除去を行うことにより，高い精度で複合語概念表記を獲得できた．

謝辞

本研究の一部は，科学研究費補助金（若手研究(B)21700241）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] 後藤敏貴，奥村紀之，渡部広一，河岡司，“概念ベースを用いた複合語の自動的属性取得法”，情報科学技術フォーラム FIT2005，pp123-125（2005）．

[3] 朝日新聞社，“朝日新聞記事データ ＜邦文＞ 2005 年版”，日本データベース開発株式会社．

[4] ChaSen -- 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/，（2011/ 1/ 27）．