概念ベースを用いた複合語概念の合成

1R−4

永森千晴† 金杉友子† 笠原要‡ 松澤和光‡

†NTT アドバンステクノロジ(株) ‡NTT(株) コミュニケーション科学研究所

1 はじめに

我々は、人間が不完全な知識の下で行う概括的な判断を計算機上で実現する「アバウト推論」の研究を行なっており[1]、そのためには大規模な常識知識が必要であると考えている。そこで、常識を構成する日常的な言葉の意味(概念)を辞書から自動獲得して知識ベース(「概念ベース」と呼ぶ)を構築し、状況や文脈を表す単語(「観点」と呼ぶ)が与えられた時に概念間の類似性を判別する方式を提案した[2]。これまで、約4万の日常語の概念ベースを構築し、類似性判別方式の評価を行なった[3]。

単語間の類似度は、各概念を表すベクトルの余弦として求められる。単語の概念は、属性とその重みの対の集合により構成されているが、類似度を計算する際には、属性をシソーラスのカテゴリーに変換し(この操作を「圧縮」と呼ぶ)、カテゴリー数次元の空間上のベクトルとして表現する。この時、圧縮された個々の属性は独立しているとみなす。また、類似度の計算はベクトルを正規化した後に行なう。表1に、2つの概念の属性(圧縮前)と重みの一部を例として示す。

表1: 概念の例

概念「家」		概念「林檎」	
属性	重み	属性	重み
家	182	果実	39
家族	94	花	33
戸主	84	高木	32
住む	78	甘酸っぱい	32
家庭	76	寒地	28
祖先	74	果樹	26
人	70	林檎	22
家並み	68	品種	20
建物	63	春	20
家作	63	弁	19
.....		.....	

このような概念ベースの規模を拡張する方式の一つとして、ユーザがどのような語を入力しても、概念ベースを用いてその語の概念を近似的に表現するユニバーサル概念ベースの構想を提案した[4]。本稿ではその基本部分である、複合語の概念を合成する方式を提案する。

2 複合語概念の合成方式

複合語の構造は、言語学や辞書学における検討課題の一つであり、また、自然言語処理の分野においても大規模なコーパスによる分析[5]が行なわれているが、複合語の概念の合成については検討されていない。本稿では、複合語の概念の合成方式を検討する第一段階として、構造上最も単純な名詞2語で構成される複合語を対象とした合成方式を提案する。固有名詞と略語については、適切な概念を合成するには別の処理が必要と思われるため、今回の検討には含めない。

2.1 複合語の抽出

複合語の概念を合成する方式を検討する参考として、日常よく使われる複合語をコーパスから抽出した。対象とするコーパスとしては、「CD-毎日新聞95版」の1995年1月の記事1カ月分(7.8MB、延べ約180万語)を用いた。形態素解析[6]の結果、名詞、もしくは接辞が連続して出現している部分を対象とし、その内、概念ベース中の名詞2語の組合せで表現できる32,798種類の語(延べ109,138語)を複合語候補として抽出した。その複合語候補の内、出現頻度が高い語上位761語について人手による判定を行ない、594種類の複合語(延べ27,898語)を選出した。この594種類の複合語は、機械的に収集した複合語候補の集合に対し、語の種類では1.8%、延べの語数では26.5%を占めている。抽出した複合語の一部を以下に示す。

表2: 新聞記事より抽出した複合語(一部)

複合語	出現頻度	構成語1	構成語2
大震災	3799	大	震災
被災者	881	被災	者
被災地	585	被災	地
避難所	325	避難	所
敬称略	282	敬称	略
自治体	261	自治	体
県内	252	県	内
大地震	194	大	地震
会社員	194	会社	員
見直し	184	見	直し
救援物資	183	救援	物資
...	...	...	...

2.2 複合語概念の合成方式

複合語の概念を合成する方式として、複合語を構成する語(構成語と呼ぶ)の間の係り受け関係や意味関係を利用して、各構成語の概念のベクトルの大きさや方向を調節する方式が考えられる。しかし、複合語中の構成語間の関係は多様であり、また、文脈や状況によって変化する場合があることを考慮する必要がある。概念ベースを用いた類似性判別方式では、文脈や状況を「観点」として表現し、観点に応じて、状況にかかわる属性を強調した概念を表現することを特徴としている。そこで、観点を考慮した類似性判別を前提とし、2つの構成語の意

A Method for generating a concept of a compound word from concepts of daily-used words

Nagamori. C.†, Kanasugi. T.† Kasahara. K.‡, and Matsuzawa. K.‡

†NTT Advanced Technology ‡NTT Communication Science Laboratories

‡1-1, Hikarinooka Yokosuka-shi, Kanagawa 239 Japan

kaname@cslab.kecl.ntt.co.jp

味的な重み付けや関係付けを行なわずに合成を行なう方式を提案する。

２つの構成語（構成語$_1$, 構成語$_2$）から成る複合語の概念（概念(複合語)）を２つの構成語の概念の単純和として定義する。

$$概念(複合語) = \frac{概念(構成語_1) + 概念(構成語_2)}{|概念(構成語_1) + 概念(構成語_2)|} \quad (1)$$

例えば複合語「景気回復」の概念は、「景気」の概念と「回復」の概念の単純和で表現する。この合成の操作は、属性を圧縮し、各概念のベクトルを正規化した後に行なう。

2.3　接尾語を含む複合語の概念合成方式

「委員会」の「会」や「作曲家」の「家」などの接尾語の概念は、他の単語の概念と異なる性格があるため、別に扱う必要がある。接尾語として用いられる語の概念の大きな特徴として、属性数が多いことと、その語が接尾語として用いられる場合にはある程度意味が限定されることが挙げられる。概念ベース中の概念の平均属性数が44であるのに対して、2.1章で作成した複合語リスト中の接尾語20語の平均属性数は105であり、平均値の約2.4倍である。

表１には、概念「家」の属性が重みの大きい順に並べられている。「家」が「作曲家」という複合語の中で用いられる場合、表中最も重要な属性は「人」であると考えられる。しかし、「家」「家族」等、この場合不適切な属性が「人」より大きな重みを持つことになる。従って、接尾語の属性の中から不適切な属性を除く、あるいは必要な属性の重みを増すことが、接尾語の概念を適切に表現する際に有効であると予想される。接尾語の属性の選び方については、辞書の品詞情報を用いる方式、辞書中の見出し語の読みを用いる方式、シソーラスのカテゴリーを用いる方式などが考えられる。

このようにして作成した接尾語の概念ともう一方の構成語の概念の単純和によって、接尾語を含む複合語の概念を合成する。

3　実験

前章で提案した概念の合成方式の有効性を検討するために、簡単な予備実験を行ない、基本的な合成方式((1)式)について検証した。その例として、「景気回復」を取り上げ、概念「景気」「回復」から合成した「景気回復」の概念と概念ベース中の４万語の概念との類似度を計算することによって、「景気回復」の類似語を検索した。結果を表３に示す。観点を考慮しない場合にも、経済が良い状態を表す「好景気」が上位に現れており、適切な類似検索が行なわれていると考えられる。また、観点として「景気」と「回復」を選択し、どちらかの語を強調した場合には、それぞれの観点に関する語が検索結果の上位に現れており、観点に応じた検索結果であると考えられる。

また「伝統的」・「抜本的」のように用いられる、接尾語「的(てき)」を取り上げ、適切な属性を選択する方式を適用した。ここでは、シソーラス[6]を用い、「状態」と「性質」の下位に含まれる属性のみを接尾語「的」の適切な属性と仮定した。その結果、概念「的」の59個の属性の内、10個が適切な属性と選択された。この中には、これまで大きな重みを持っていた「的(まと)」に関する属性(目標, 弓, 発射 等)は含まれず、「性質」,「抽象的」など、接尾語の概念の属性としてふさわしいものが含まれており、接尾語の属性を選ぶという方式の有効性が期待できる。

表 3:「景気回復」の類似語

観点: なし		観点: 景気		観点: 回復	
類似語	類似度	類似語	類似度	類似語	類似度
回復	0.73	景気	0.91	回復	0.97
景気	0.73	前景気	0.86	取り直す	0.84
好景気	0.66	好景気	0.84	起死回生	0.78
景況	0.65	景況	0 83	復す	0.78
不況	0.63	空景気	0.82	巻き返し	0.77
好況	0.59	不況	0.82	持ち直す	0.76
活況	0.57	好況	0.78	反る	0.76
不景気	0.56	不景気	0.77	再起	0.76
戻す	0.50	活発	0.65	挽回	0.74
復す	0.50	豪勢	0.63	持ち	0.73
挽回	0.49	好調	0.59	弾力	0.70
持ち直す	0.49	ごね得	0.58	更生	0.70
再起	0.48	時化	0.57	生れ変る	0.69
.....		.....		.....	

4　おわりに

本稿では、概念ベース中の名詞２語を構成語とする複合語の概念の合成方式を提案し、予備検討を行なった。今後は新聞記事から抽出した複合語を用いた定量的な評価を行なう予定である。また、構成語が３語以上の複合語や用言を含む複合語に対する適用法についても検討して行きたい。

References

[1] 松澤, 石川, 湯川, 河岡: アバウト推論 －「常識的な推論」を目指して－, AI学会人工知能基礎論研究会, Vol. SIG-FAI-9401-1, pp. 1–8 (1994).

[2] 笠原, 松澤, 石川, 河岡: 観点に基づく概念間の類似性判別, 情報論文誌, Vol. 35, No. 3, pp. 505–509 (1994).

[3] 笠原, 松澤, 石川: 国語辞書を利用した日常語の類似性判別, 情処論文誌, Vol. 38, No. 7 (1997).

[4] 笠原, 松澤, 石川: ユニバーサル概念ベースの提案, 第55回情全大, Vol. 2R-05 (1997).

[5] 国立国語研究所: 電子計算機による新聞の語彙調査 (III), 秀英出版 (1972).

[6] Ikehara, S., Shirai, S., Yokoo, A. and Hiromi, N.: Toward an MT System without Pre-Editing -Effects of New Methods in ALT-J/E-, *MT Summit '91*, pp. 101–106 (1991).