概念ベース研究のその後

～表記と概念のマッピング手法と形容詞の類似性判別方式の紹介～

笠原 要 † 永森 千晴 †† 金杉 友子 †† 松澤 和光 †††
Kaname Kasahara † Chiharu Nagamori †† Tomoko Kanasugi ††
Kazumitsu Matsuzawa †††

† NTT コミュニケーション科学基礎研究所 †† NTT アドバンステクノロジ
††† NTT サービスインテグレーション基盤研究所

1 はじめに

ことば工学研究のきっかけとなった「概念ベース」の研究について、最近の研究状況を紹介する。概念ベースは、単語の意味 (「概念」と呼ぶ) を国語辞典情報を基に機械的に知識ベース化したものであり、約 4 万語の概念を、各々約 3 千次元のベクトル (「属性ベクトル」と呼ぶ) で表している。詳しくは文献 [1] を参照されたい。当研究会においても文献 [2] で紹介した。現在は、以下の内容の研究を進めている。

(1) 9 万語規模概念ベースの構築 [3]

(2) コーパスの共起情報に基づく概念ベースの補強 [4, 5]

(3) 概念ベースの情報検索システムへの応用 [6, 7]

(4) 概念ベースの意志決定支援システムへの応用 [8]

本稿では、これらのうちでことば工学に関係の深い成果として、(1) および (2) から 2 つの話題を取り上げ紹介する。また、その前提となる概念ベースの構築と単語の類似性判別について、簡単に説明する。

まず (1) では、国語辞典の見出しを使って概念ベースの概念の意味の範囲を決定する問題を取り上げる。国語辞典の特徴として、1 つの見出しに複数の表記（多くの場合は漢字表記）があり、また (読み等の違いから)1 つの表記に複数の見出しが当てはまるという「もつれた構造」がある。このため、概念ベースでの「概念」をどう選ぶかが重要であり、文献 [3] の研究では初期の概念ベースで用いた方法とは異なる新しい方法を用いている。実際に 9 万語の概念ベースに新しい方法を適用し、有効性の検証をした内容について紹介する。

次に (2) では、国語辞典以外の情報源としてテキストコーパスを用いて概念ベースを補強する試みについて、特に形容詞に関する実験を報告する。形容詞・形容動詞約 5 千語について記事中の共起情報によって「共起ベース」を構築し、概念ベースと比較した結果について考察する。

笠原 要 NTT コミュニケーション科学基礎研究所 〒619-0237 京都府相楽郡精華町光台 2-4 Tel (0774)93-5353 Fax (0774)93-5385 e-mail: kaname@cslab.kecl.ntt.co.jp

2 概念ベース

まず、概念ベースの構築方法について説明する。単語の意味である概念をどのように表わすかについては様々な研究があるが、我々は自動獲得と利用の容易さを考慮して、単語の特徴 (以後「属性」と呼ぶ) の重みを要素とする属性ベクトルで表現している。属性を獲得するために国語辞典を知識源とした。勿論、国語辞典から単語のあらゆる特徴を獲得することは不可能だが、語義文には単語を定義する語が含まれており、知識源として適切と考えた。

実際に属性ベクトルを獲得する方法を図 1 で説明する。見出し語「馬」の概念を表現するのに、語義文中の「首」や「尾」等を「馬」の属性と考え、それらの出現頻度を属性の重みとする。しかし、個々の説明文だけでは十分な特徴が獲得できない場合がある。これを解消するため、人間が国語辞典中で分らない単語があった場合に行う、辞典を再帰的に参照する “孫引き” や用例を参照する過程をモデル化し、属性となる単語を獲得し精錬する方法を提案した。

得られた属性ベクトルでは、属性となる単語間に関連性がみられる。この関連性を解消するため、単語の集合を複数に分類する類語辞典を利用し、属性を単語から類語辞典の分類に変換する。これによって、例えば「栗毛」と「鹿毛」のように分類上近い属性の重みは、分類「体毛」の重みとしてまとめられる。類語辞典には、約 30 万語を 3 千に分類しているコミュニケーション科学基礎研究所で開発された日本語語彙大系 [9] を用いた。最終的に得られる属性ベクトルの属性数は 3 千となり、3 千次元の多次元空間上の位置として単語の概念が表現される。現在までに、学研国語大辞典 [10] を知識源として約 9 万語の単語の概念を収録する概念ベースを構

築した。

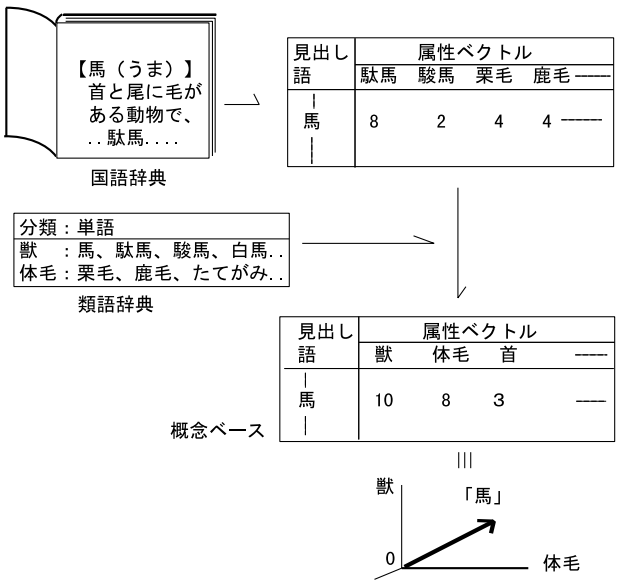

図 1: 国語辞典に基づく概念ベースの構築

次に、概念ベースを用いた単語の類似性判別について説明する。類似性判別とは 2 つの単語の似ている度合 (類似度) を計算するタスクと考え、概念ベースを用いる場合、2 つの単語の概念を表わす属性ベクトルどうしがなす角の余弦を類似度と定義する。

人間の行なう類似性判別の特徴として、判別の状況依存性があげられる。例えば、「動物」に関する話をしている時に、「馬」に対して「豚」と「自動車」のどちらが似ているかと尋ねられた場合常識的には「豚」が似ていると判断する。一方で、「乗り物」に関する話をしている状況では、反対に「自動車」の方が「馬」と類似していると判断する。そこで、人間同様の状況を考慮した類似性判別を行なうために、状況を表わすような単語 (“観点”) が与えられた場合、状況に応じて類似度を与える方法を提案した。観点となる単語の概念を用い、観点に含まれる属性の重みを考慮して類似度を計算することが可能となっている。(判別例：図 2)。

3 概念ベース構築における表記と概念のマッピング手法

概念ベースを作成する際の知識源となる国語辞典では表 1 の例のように、見出しは 1 つの読みと、1 つまたは複数の表記から構成されている。例えば [あだ]【徒・空】という見出しは、「あだ」という読みと「徒」「空」という 2 つの表記を持っている。英和辞典や漢和辞典のような、単語を表記から引く辞書とは異なり、国語辞典は単語を読みから引く構造になっているために、表記だけでは見出しを一意に決定することができない場合

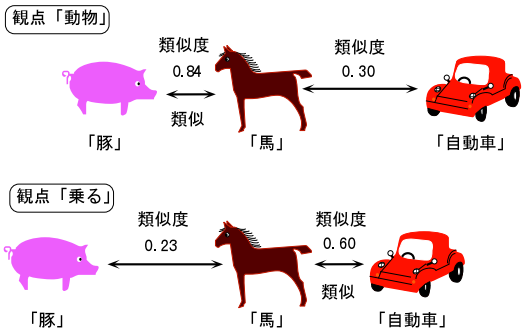

図 2: 観点に基づく単語の類似性判別

がある。例えば「徒」という表記は「あだ」「いたずら」「かち」等複数の見出しに含まれるため、読みまで指定しないと見出しを特定できない。更に、同じ表記と読みを持つ単語が、意味や用法の違いから複数の見出しに分けて定義されている場合もある。つまり、国語辞典の見出しは読み・表記・意味もしくは用法が決定してはじめて 1 つの見出し語が特定できるという複雑な構造を持っている。学研国語大辞典 [10] の見出しを調査した結果、全見出しの 11% が他の見出しにも現れる表記を持ち、また、8%が他の見出しの表記と完全に一致した。

一方、概念ベースを用いてテキストの情報処理を行なう際には、単語の表記情報しか与えられていない場合が多い。従って、概念ベースを作成する際に、国語辞典の見出しの複雑な構造を整理して、表記のみで 1 つの概念を指定できるようにすることが重要となる。本稿では、国語辞典の見出し語の表記と見出し語の概念を対応づけるマッピング手法を提案する。

表 1: 国語辞典中の見出しの例

[あだ]【徒・空】	[くう]【空】
[いたずら]【徒】	[そら]【空】
[うつお]【空】	[ただ]【徒】
[うつけ]【空 (け)・虚 (け)】	[と]【徒】
[うろ]【空・虚・洞】	[とほ]【徒歩】
[かち]【徒・徒歩】	[ほら]【洞】
[から]【空】	[むだ]【無駄・徒】
[きょ]【虚】	

3.1 提案手法

国語辞典の見出し語と語義文から概念を作成する場合、単純な手法としては、1 つの見出し語から 1 つの概念を作る (方式 1 とする)、共通の表記を持ついくつかの見出し語をまとめて 1 つの概念を作る (方式 2 とする)、の 2 つが考えられる。

方式 1 には、見出し語と語義の対応が 1 対 1 であり、国語辞典の定義を厳密に用いることができるという

利点があるが、表記から概念を一意に定めることができないという欠点がある。一方、方式 2 には、表記から概念を一意に定めることができるという利点があるが、元々異なる語義をもつ語が同じ概念として扱われてしまう、すなわち概念の精度が損なわれてしまうという欠点がある。

そこで、国語辞典中の表記と語義の対応を尊重し、かつ、表記から概念を一意に定めることができるマッピング法 (方式 3 とする) を提案する。その手法を以下に説明する。見出し語の表記を抽出し、表記と、その表記を含む見出しの対応表を作成する。概念は表記ごとに作成するものとし、その表記を含む全ての見出しの語義文から属性を抽出する。ただし、複数の表記が常に同じ見出しに現れる場合は、同じ概念であると考え、どちらも 1 つの概念を表す表記とした。例えば、[うつけ]【空 (け)・虚 (け)】という見出しには、「空け」「空」「虚け」「虚」の四つの表記を持っているが、「空け」と「虚け」は他の見出しの表記と重ならないので 1 つの概念とみなし、「空」「虚」「空け、虚け」の 3 つの概念を作成する。この方式により、見出し中の複数の表記のうち、他の見出しに含まれない表記の語義には変更を加えないので、異なる語義を持つ語が同じ概念として扱われるという問題を必要最小限に留めることができると考えられる。前章の見出しの例を、方式3 により概念にマッピングした例を表2に示す。

表 2: 方式 3 による概念マッピングの例

徒 ＝ （あだ＋いたずら＋かち＋ただ＋と＋むだ）
空 ＝ （あだ＋うつお＋うつけ＋うろ＋から＋くう＋そら）
空け・虚け ＝ （うつけ）
虚 ＝ （うつけ＋うろ＋きょ）
洞 ＝ （うろ＋ほら）
徒歩 ＝ （かち＋とほ）
無駄 ＝ （むだ）

3.2 実験と考察

前章で提案したマッピング法 (方式 3) で、どの程度意味の異なる見出し語が 1 つの概念にまとめられているのか、学研国語大辞典から作成した概念ベースを用いて調査した。

実験には、方式 1 により作成した概念ベース (概念数 79,471) と、表1の 15 の見出し語を用いた。方式 3 を用いると、15 の見出し語は表 2 の 7 つの概念にまとめられる。まず、7 つの概念それぞれについて、方式 3 により 1 つにまとめられる見出し語の概念の属性ベクトル各々と、それらの重心ベクトルとの余弦の平均値を求めた。次に、7 つの平均値の平均値を求めた。また、比較対象として、方式 2 を選んだ。この場合、表1の見出し語は全て 1 つの概念にまとめられる。各見出し語の 15 の属性ベクトルと、それらの重心ベクトルとの余弦の平均値を求めた。その結果を表 3、4に示す。

表 3: 実験結果

	方式 2	方式 3
余弦平均	0.32	0.73

表 4: 方式 3 の結果の内訳

概念	結合数	余弦平均
徒	6	0.43
空	7	0.45
空け, 虚け	1	1
虚	3	0.58
洞	2	0.82
徒歩	2	0.84
無駄	1	1

方式 2 と比べ、全体の平均値、概念ごとの平均値共に方式 3 の方が高くなっており、比較的意味の似た語を同じ概念としてまとめることができていることがわかる。方式 3 を用いて作成した概念ベースの類似語検索結果を表5に示す。方式2 を用いると同じ概念として扱われてしまう「空」と「徒歩」をわけることができ、かつ、適切な類似語が検索できている。

以上の通り、国語辞典の見出し語の表記から見出し語の概念を決定できるように対応づけるマッピング手法を提案した。この手法は、概念ベースの作成のみではなく、見出し語の読みから語義を参照する形式の辞書を利用した情報処理に広く利用可能であると考えられる。

表 5: 提案法による類似語検索結果

「空」の類似語		「徒歩」の類似語	
類似語	類似度	類似語	類似度
明らむ	0.76	歩き	0.57
満天	0.75	徒	0.56
雪空	0.74	拾い歩き	0.56
天心	0.74	急歩	0.53
秋天	0.74	歩卒	0.52
初空	0.74	行歩	0.52
雲霄	0.72	漫ろ歩き	0.51
曇天	0.72	緩歩	0.50
.....		.....	

4 コーパスを用いた形容詞の類似性判別方式

前章で述べた通り、現在、学研国語大辞典[10] を用いた 9 万語の概念ベース [3] が構築され、類似度計算、類似語検索とともに情報検索システム [6, 7] 等に応用されている。これらの概念ベースの利用を通して、概念ベース研究における課題を明らかにした。まず、国語辞典を

元にしているため新語や固有名詞等を扱えない点があげられる。また、国語辞典には一般的に使用頻度の低い語が含まれるために語彙に難語が多い問題もある。

そこで国語辞書を元にした概念ベース以外に、それを補う新たな意味知識ベースが必要だと考え、コーパスから共起情報に基づく意味知識ベースを作成する研究を進めている。共起とは、文や文書中で2つの単語が同時に出現することを指す。例えば、主語となる名詞と述語となる動詞や、ある単語に文中で隣接して現れる単語がコーパス中で共起している場合、これを用いて知識ベースを作る。これまでに連接する名詞-名詞の共起対を用いて共起ベースを作成し、名詞の類似性判別を行う方法を提案している[4]。今回は、形容詞について統語情報を利用した意味知識ベース(共起ベース)構築方法を提案し、9万語の概念ベース中の形容詞の概念との比較を行なった。

4.1 形容詞共起ベース構築方法

形容詞の共起ベースの構築方法について述べる。形容詞を選択したのは、語彙が少なく試験的に構築するのに適している点と、比較的単純な統語情報を使って共起対がとれると予想できる点を考慮した結果である。共起対抽出にあたって、英語ではあるがHatzivassiloglou[11]の手法を参考にし、日本語固有の問題を考えて単純な2ルールのみをまず用いることとした。

(1) 形容詞/形容動詞連用形+動詞

(2) 形容詞/形容動詞連体形+名詞

ただし、隣接する語どうしをそのまま抽出できる例は少ないため、共起対の間に出現する不要語(副詞など)の除去等の処理をする。

抽出した共起対については、形容詞/形容動詞を概念、名詞/動詞を属性、その共起頻度を属性の重みとし、概念ベースと同様の形式の共起ベースを構築する。

4.2 実験

4.2.1 形容詞共起ベースの作成

実際に形容詞共起ベースを構築した。ソースには電子化された91〜95年の毎日新聞5年分(4,358,257文。375MB)[12]を用いた。ALTJAWS[13]を使って記事文を形態素解析した後、形容詞/形容動詞とそれに修飾される名詞/動詞を抽出した。上記の手法により抽出できた共起対は、612,510対。任意に選んだ200対の獲得精度を調査したが、約9割が正しく抽出できていることがわかった。この共起対を用いて作成した共起ベースと9万語の概念ベース中の形容詞/形容動詞の概念を抽出したもの(以後「概念ベース」と記す)との比較を以降で行なう。

4.2.2 共起ベースと概念ベースとの比較

まず形容詞/形容動詞の語彙数を調査した。共起ベースの語彙数は4,626、概念ベースの語彙数は5,096、両者の重なり語彙数は2,398となった。共起ベースにはカタカナ語や「名詞+的だ」型の形容動詞等の新語、時事的な語が多い(ex. マルチメディア的だ, 吉本興業的だ)。概念ベースはカタカナ語は少ないが、漢字表記語については難語といえる語まで広くカバーしている(ex. 英邁だ, 魯鈍だ, 恪気深い)。両者の語彙に重なりは少なく、共起ベースが国語辞典の範囲外の語を扱えると期待できる。

概念を構成する形容詞/形容動詞が一般的にどの程度なじみがあるかを調査した。これは、新明解国語辞典の見出し語について、被験者30人になじみがある順に7〜1の7段階の評価値をつけさせ、それを平均した単語親密度データ[14]を元にした。単語親密度の平均を算出したところ、共起ベースが5.681、概念ベースが4.649となった。概念ベースは難語の概念が多く、難語を除外するには共起ベースを用いる方が良いと言える。

次に、共起ベース、概念ベースを比較すると、属性数の平均はほぼ同じであるのに、属性数の標準偏差にはかなりの差が見受けられる。重なり語彙について属性数の分布を調べ、同じ語彙でも属性数の挙動に差がでるかを調べた(図3)が、分布には相違が見られ、同じ概念でも両者の間では保持する属性に差が出ると考えられる。

表6: 共起ベース、概念ベース諸元

項　目	共起ベース (CB)	概念ベース (GB)
概念数	4,626	5,096
1概念の属性数平均	49.468	49.024
属性数標準偏差	113.965	20.408

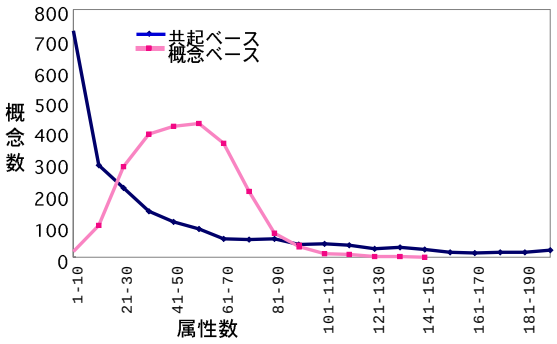

図3: 重なり語彙についての属性数分布

4.2.3 重なり語彙の評価

次に同じ語の概念の共起ベース, 概念ベース両者における性質の差を見るための評価を行なった。重なり語彙(2398 概念) を 3 つの語群に切り分け、各々類似語検索して評価を行なった。具体的には、同じ語の概念の属性数を調査し、その上で共起ベースと概念ベースの属性数の差によって 3 種の語群に分けた。

表 7: 同じ語の概念の属性数の差による分類

同じ語の概念の属性数の差	語数
CB − GB $>$ 50	683
GB − CB $>$ 50	712
-50 $\leq$ CB − GB $\leq$ 50	1003

さらに、電子化版『現代用語の基礎知識 1999』[15] を用いて、3 種の語群に属する形容詞/形容動詞を含む文を各々100 文ずつ300 文抽出しサンプルを作成。サンプル中の当該形容詞/形容動詞部分を共起ベース、概念ベース各々によって検索された上位 1 位の類似語に置き換え、置き換えた文が置き換える前の文と意味が類似しているかを被験者 (1 名) により評価した。評価結果は、類似語で置き換えるとはいえ前後の部分と照合した際文のニュアンスが合わない場合があり、類似率は3 割程度となった。だが、同概念について属性数の多いものを用いて類似語検索した方が良い結果が得られるのがわかる。属性数の多少により共起ベース、概念ベースの両者を使い分けて利用することが可能だと考えられる。

表 8: 置き換え文の類似率 (%)

属性数の差	類似率 (CB)	類似率 (GB)
CB − GB $>$ 50	33	26
GB − CB $>$ 50	17	37
-50 $\leq$ CB − GB $\leq$ 50	28	30

4.3 まとめ

国語辞典を元にした概念ベースを補う形で、形容詞についてコーパスから共起頻度を用いた意味知識ベースの構築法を提案し、実際に共起ベースを作成した。結果、品詞を限ったにも拘らず概念ベースと共起ベースの間には語彙に違いがあり、共起ベースによって今まで範囲外だった語を扱える確信を得た。また、属性数や親密度によって適した方を使い分け、複数の意味知識ベースを活用するある程度の指針ができたと考える。今後は、他の品詞や多種多様なコーパスによる共起ベースの構築、統語解析等による共起対抽出精度の向上など更に研究を重ねる予定である。

5 おわりに

NTT コミュニケーション科学基礎研究所で行われている概念ベース研究について、最近の進展を報告した。今後もことば工学に関連した成果について、当研究会で継続的に紹介していくことにしたい。

参 考 文 献

[1] 笠原, 松澤, 石川: 国語辞書を利用した日常語の類似性判別, 情報処理学会論文誌, Vol. 38, No. 7, pp. 1272–1284 (1997).

[2] 松澤: 概念ベースは人間の夢を見るか？, 人工知能学会ことば工学研究会資料, SIG-LSE-9902-8, pp. 54–60(1999).

[3] 永森, 笠原, 松澤: 概念ベース構築における表記と概念のマッピング手法, 第 14 回人工知能学会全国大会, 07-08, pp. 163–164(2000).

[4] 稲子, 笠原, 松澤: 複合語内の単語共起を用いた単語の類似性判別方式, 「言語資源の共有と再利用」シンポジウム, http://www.etl.go.jp/etl/nl/ sympo99/inago.html (1999).

[5] 金杉, 笠原: コーパスを用いた形容詞の類似性判別方式, 第 14 回人工知能学会全国大会, 07-07, pp. 161–162(2000).

[6] 熊本、島田、加藤: 概念ベースの情報検索への適用, 信学技報, AI87-63, pp. 9–16(1999).

[7] 加藤, 笠原, 北: 概念検索に基づく技術内容からのエキスパートの推定, 信学技報, NLC2000, pp. 55–62(2000).

[8] 藤本, 賀沢, 佐藤, 阿部, 松澤: Dsiu システム: Decision support for internet users 「ネット情報を使ってホットなものをあなたに！」, 人工知能学会論文誌, Vol.15, No.1, pp. 61–64(2000).

[9] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 (編): 日本語語彙大系, 岩波書店 (1997).

[10] 金田一, 池田 (編): 学研国語大辞典 (第 2 版), 学習研究社 (1988).

[11] Hatzivassiloglou,V., McKeown,K.,: Towards the automatic identification of adjectival scales: clustering adjectives according to meaning, In ACL 31, pp. 172–182(1993).

[12] CD-毎日新聞 91-95 版, 毎日新聞社 (1991-1995).

[13] 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).

[14] 天野, 近藤: NTT データベースシリーズ「日本語の語彙特性」第 1 巻・単語親密度, 三省堂,(1999).

[15] 現代用語の基礎知識 1999, 自由国民社 (1999).