心理言語学的な意味属性の言語処理への適用可能性の検討

林 良彦 (大阪大学大学院言語文化研究科)

hayashi@lang.osaka-u.ac.jp

1　はじめに

単語の意味，あるいは，単語が指示する概念の捉え方の一つに，基本的な意味属性の組み合わせ (最も単純には集合) によりこれを規定しようという立場がある．本稿では，そのような立場に基づき，特に心理言語学的な観点から構築された意味属性の体系を工学的な言語処理へ適用する可能性とそのための問題点について基本的な検討を行う．

本稿の以下では，心理言語学的な意味属性に関連する研究について簡単に概観した後，本研究で着目する，McRae ら [4] による英語を対象とする概念・意味属性体系 (semantic feature norms) について紹介し，このような体系を言語処理に適用するための要件について議論する．次に，言語処理への適用可能性をもう少し具体的に検討するため，McRae の体系と Princeton WordNet (PWN)[5] との対応付けを行った結果について考察する．より具体的には，McRae の意味属性を構成する語と PWN における語彙概念の説明テキスト (gloss) 中の語とのオーバラップについて調査した結果を報告する．以上を踏まえ，今後の研究の方向性について議論する．

2　関連研究

基本的な意味要素 (意味属性) の組み合わせとして単語の意味を規定しようという考え方 (componenetial analysis) は古くから根強く存在し，近年のより現代的な語彙意味論のいくつかも，大きな意味ではその延長線上にあると見なすことができる．認知科学・心理言語学の分野においても，同様の考え方は一定のコンセンサスを得ており，特定の概念を規定するための意味属性の体系 (semantic feature norms) を経験的な手段により導こうとする試み ([4, 6] など) が多くなされてきた．これらの体系は，対象の概念に対して重要と考える性質 (properties) を被験者実験により収集し，得られた概念記述を実験者が整理するという手順により構築されてきた．

このような心理言語学的な概念・意味属性体系を言語処理的な観点から取り上げた研究は多くはないが，報告者が知る限り，以下のような研究例がある．

- Baroni ら [1] は，単語の持つ意味概念を意味属性の集合により規定するというモデルにおいて，意味属性タイプ+引数 という構造を持つ意味属性をコーパスから抽出し構造化する試みを行なっており，得られた結果を後述する McRae らの概念・意味属性体系 [4] と比較している．
- Kremer ら [3] は，多言語辞書において意味的に関係した語彙エントリを提供するという観点から，言語の違いを越えて特定のクラスの概念を規定する上で顕在性の高い意味・概念関係を心理実験的アプローチにより調べている．ここでも後述する McRae らの体系を基準として用い，これを多言語 (ドイツ語，イタリア語) に翻訳したものを実験で利用している．

3　McRaeらによる概念・意味属性体系

本稿では，心理言語学分野における体表的な概念・意味属性体系として，McRae らによる体系 [4] を取り上げる．この体系は，基本レベル (basic level)[1]に属すると考えられる，生物・無生物に関する概念 541 種類に対し，700 名以上の被験者により与えられた概念記述を整理して導いた 2,526 種類の意味属性からなっている．ここで，概念に対しては，それを曖昧性無く指示する英語名詞[2]が刺激語として被験者に提示された．公開されているデータには，各概念に対してどのような意味属性が何人の被験者によって付与されたかという基本的なデータのほか，意味属性への分類カテゴリの付与や，各種の統計データが含まれている．

[1]日常生活においてオブジェクトを分類する際に参照されていると考えられる概念階層におけるレベル (抽象すぎず詳細すぎない) を指す．

[2]後ほど議論するように，意味的な曖昧性の有無は基準とする意味の粒度に依存する．

意味属性の付与例: 例えば，英語名詞 ”accordion” が有するある語義 (実際は単義と考えてよい) によって指示される概念に対しては，{a_musical_instrument, associated_with_polkas, has_buttons, has_keys, inbeh_-_produces_music, is_loud, requires_air, used_by_moving_bellows, worn_on_chest} という意味属性集合が付与されている．各意味属性は，has_ や used_by_ などの一定の統制がなされた接頭辞部分と ”buttun”，”moving bellows” といった比較的自由な単語 (列) による部分からなる．本稿では，前者を属性タイプ[3]，後者を属性語と呼ぶ．本例の各意味属性の意味はほぼ自明と思われるが，inbeh_-_ という接頭辞は，非生物オブジェクトの振る舞いを表す．上記の例に明らかなように，a_musical_instrument のような概念の上位・下位関係，has_buttons などの全体・部分関係の他に，様態や機能を表す関係や，さらには，連想的な関係も含まれており，通常の語彙的オントロジーの言語資源には必ずしも含まれない豊かな情報が含まれている．

意味属性への分類の付与: 公開されているデータにおける付加的な情報として興味深いものに，意味属性への 2 つの体系による分類カテゴリの付与がある．そのうちの一つは，脳領域階層分類 (brain region taxonomy; 以下，BR 分類) [2] と呼ばれる体系によるもので，タクソノミー的関係 (taxonomic) のほか，encyclopaedic, function, smell, sound, tactile, taste, visual-colour, visual-form-and-surface, visual-motion の 9 種類のカテゴリからなる．例えば，先の意味属性の例においては，has_buttons には visual-form-and-surface, requires_air には encyclopaedic という BR 分類カテゴリが付与されている．

4 言語処理への適用可能性の要件

上記の例にも見られるように，心理実験的なアプローチにより導かれた概念・意味属性体系には，概念の規定において，心理的な顕在性の高い情報が含まれており，これを活用できれば，より高度な意味的処理が実現できる可能性がある．しかし，工学的な観点からは，これらの体系は小規模なものであり，また，特定の言語 (英語) に偏っているという大きな問題がある．

このような問題を克服するための究極的な方策は，より多くの心理実験を行うということであるが，コストなどの観点から現実的ではない．よって，既存の概念・意味属性体系の範囲を拡大し，さらに他言語へ展開する方法論を実現し，両者を併用していくことが必要となる．この意味で，[1] のように，コーパスから意味属性を抽出・構造化するとするアプローチは有用である．一方で，既存の意味知識・言語資源に対して，意味属性の情報を拡張しながら付加していくというアプローチも対案・相補的な方略として考えられる．

[3]公開されているデータからは，96 種の属性タイプを抽出したが，これらの中には誤りと思われるものもある他，整理統合が必要と思われるものも散見される．

5 PWN との対応付けに基づく概念・意味属性体系の調査

代表的な既存の言語資源として PWN[5] を選定したとき，McRae の体系から得られる意味属性情報がどの程度，PWN に含まれているかを調査した．

5.1 PWN synset への対応付け

McRae の体系における 541 種の対象概念に対応する英語名詞を見出し語として PWN を検索し，付与されている意味属性から推定される概念と対応すると判断できる PWN synset への対応付けを行った．[4] では意味的に曖昧性のない英語名詞を選択したとの記述があるが，PWN の検索において，複数の synset が対応付けの候補となったものは 372 件あり，対応付け候補が単一であったものは 169 件に過ぎなかった．また，主に PWN の概念区分の粒度が細かいことが要因で複数の候補 synset の中から対応する synset を一義に定めるのが困難なケースが 89 件あり，これらについては複数の synset への対応付けを許容した[4]．

5.2 属性語の synset gloss におけるオーバラップ

PWN においては，synset が規定する語彙概念の意味に対して英語による説明 (gloss) が付与されている．そこで，McRae の体系において付与されている意味属性から抽出した属性語が，対応する PWN synset の gloss においてどの程度のオーバラップをもつかを調査した．調査にあたっては，双方の言語表現に対して，解析器 TreeTagger[5]による lemmatization を施し，lemma におけるマッチングを行った．

[4]一方で，1 つの概念 (urn) に関しては，付与された意味属性に対応すると考えられる synset が存在しなかった．

[5]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

まず，502 種の概念/synset (92.8%) において，属性語のオーバラップが確認できた．特に，PWN における対応候補の synset が 1 件しか存在しない概念に対してその割合 (95.4%) は複数のものに対して有意に ($p < 0.05$) 高かった．

次に，属性語集合中で gloss とオーバラップする語の数の割合を尺度として候補の synset 集合をランキングするという設定により，オーバラップ状況の特性を調べた．この設定は一種の情報検索であり，評価尺度として情報検索で良く用いられる MAP (Mean Average Precision) を用いた．この結果，全ての属性語を用いた場合の MAP 値は 0.939 であり，BR 分類において taxonomic 以外に分類される意味属性から抽出した属性語のみを用いた場合の MAP 値は 0.936 であった．このように，MAP 値の差は極めて小さく，これは，そもそも非 taxonomic に分類される意味属性の割合が高い (McRae の体系において，token で 89.9%，type で 91.8%) という要因が大きいが，gloss において概念階層以外の様々な観点からの記述が行なわれている状況が裏付けられた．このことは逆に考えれば，語彙概念のテキストによる説明記述から心理言語学的な意味属性をある程度抽出できる可能性を示す．

5.3 周辺概念におけるオーバラップ

我々の関心は，ある概念に対して付与された意味属性がどの程度それに関連している概念にも適用出来るか，あるいは，どのような制限がありえるか，ということを明らかにすることにある．そこで,「周辺概念」における gloss とそもそもの概念に付与された属性語とにどの程度のオーバラップが見られるかを調査した．

より具体的には，対象概念に対応する PWN synset から任意の概念関係により 1-hop で結ばれた synset 集合を候補集合に加え，前節と同様の尺度によりランキングを行い，MAP 値を求めた．その結果，全ての属性語を用いた場合の MAP 値は 0.733 であり，非 taxonomic な意味属性から抽出した属性語のみを用いた場合の値は 0.695 であった．

まず，前節の結果と比べれば，両者の値とも減少している．しかし，この MAP 値の減少は悪いニュースではなく，周辺概念に対する gloss においても属性語がオーバラップしていることを示している．また，非 taxonomic な意味属性からの属性語のみを用いた場合の減少の度合いが前節の結果よりも大きい．これは，非 taxonomic な意味属性が「周辺概念」において相対的により多く成立している可能性を示す．

5.4 概念のタイプと概念関係の分布

既存の概念・意味属性体系を工学的に拡張しようとする試みにおいては,「どのような概念に対するどのような意味属性が，いかなる概念関係で結ばれた周辺概念にも拡張しうるか」を予測できることが望まれる．

概念のタイプに対する偏り： ランキングにおいて,「正解」である対象概念に対応する PWN synset が 1 位に選定されたもの (A 群:327 件; オーバラップする属性語がより多い)，2 位以下となったもの (B 群:213 件; オーバラップがより少ない) の 2 群に分け，概念のタイプに対する偏りを調査した．McRae の体系における 541 種類の概念は基本的にはフラットな集合をなしているが，以下の 3 つの基準によるグルーピングを行い，A 群/B 群による偏りが見られるかを調査した．

- PWN における lexicographer file 名: lexicographer file の名前は，”noun.artifact” のように，品詞名に接尾辞が付加した形になっており，この接尾辞を意味の大まかな分類を示すものとして利用する．
- クラスタリング 1: 公開されている McRae の体系のデータには，各概念に付与された意味属性の頻度分布を特徴ベクトルとした時の概念間の類似度のデータが含まれている．そこで，この類似度行列を入力とし，Ward 法によるクラスタリング (クラスタ数:30) を行い，クラスタラベルを概念タイプとした．図 1 に得られた結果の一部を示す．意味的にみて比較的良好なクラスタが得られている一方で，雑多なクラスタも多いというのが実際のところである．
- クラスタリング 2: 上記の類似度行列の代わりに BR 分類における 10 カテゴリの頻度分布を特徴ベクトルとした距離行列を算出し，同様のクラスタリングを行った．

その結果からは，いずれの場合も統計的に有意な偏りは確認できなかった (いずれも，$p > 0.2$) が，クラスタリング 1 において，2 つのクラスタ (図 1 の#5, #12) が A 群に有意に多いという結果が得られた．ただし，これらのクラスタにおいて要素の意味的な斉一性は高くはなく，選択するクラスタ数に依存して偶然に現れたと考えるのが妥当でる．以上から，いずれのの概念のタイプ分けによっても，周辺に属性が拡張されやすい概念のタイプを規定することは困難であり，おそらくは，属性が拡張されやすい/されにくい特定

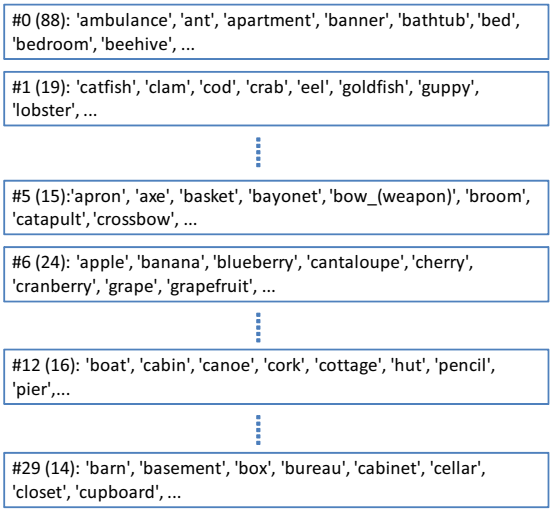

図 1: クラスタリング結果の一部 (括弧内は要素数)

表 1: 主要な概念関係の分布

概念関係	A 群	B 群
hypo (下位概念)	314	4,096
hype (上位概念)	123	1,485
mprt (被構成要素)	37	608
hprt (構成要素)	18	217
dmtc (ドメイン)	10	144
hmem (構成メンバー)	7	198
hasi (インスタンス)	2	176

のタイプの概念というものは存在しない可能性も想像される．

概念関係の偏り: 上記と類似の考え方により，正解の PWN synset よりも上位にランクされた周辺概念 synset (A 群) と下位にランクされたもの (B 群) の間に，概念関係 (対象の正解 synset とどのような関係で結ばれた周辺概念であるか) の観点において偏りがあるかを調べた．表 1 にその結果 (頻度上位のもののみ提示) を示す．A 群/B 群の概念関係の分布には統計的な有意差が見られた ($p = 0.001$) が，これは，hmem (holomyms-member; 構成メンバーの観点からの全体・部分関係)，hasi (has instance) の 2 つの概念関係において偏りが大きかったというのが要因で，この 2 つの関係を除外した場合の分布には有意差がなかった．このことより，hype/hypo (上位・下位関係)，mprt/hprt (構成要素の観点の全体・部分関係) といった主要な概念関係により結ばれる周辺概念にわたって，意味属性が拡張されうる可能性が分かる．

以上より,「どのような概念に対するどのような意味属性が，いかなる概念関係で結ばれた周辺概念にも拡張しうるか」という制約的な情報を得るには，少なくとも本稿で述べた範囲の分析では十分ではないと考えれる．

6 おわりに

以上に検討したように，心理言語学的なアプローチにより導出された概念・意味属性体系は言語処理への適用においても有望な豊かな情報を有しており，また，既存の言語資源との組み合わせにより適用範囲を展開できる可能性がある．一方で，概念に対する荒い意味的なグルーピングや概念間の関係タイプといった情報だけでは，意味属性を展開可能な範囲を制約として明確化することが困難であることも確認された．

今後は，概念･意味属性体系における，概念，意味属性の双方における適切な抽象化を図り，さらにはコーパスや Web などの言語資源からのデータを併用することにより，意味属性体系を必要な範囲で拡張し，周辺概念への展開可能性を計量化する手法について検討を進めていく．

謝辞: 本研究の一部は，科研費 (#24650123) による．

参考文献

[1] Marco Baroni, et al. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, Vol.34, pp.222–254.

[2] George S. Cree and Ken McRae. 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, Vol. 132, No. 2, pp.163—-201

[3] Gerhard Kremer, et al. 2008. Cognitively salient relaitons for multilingual lexicography. *Proc. of the Workshop on Cognitive Aspects of Lexicon* (COGLAEX 2008), pp.94–101.

[4] Ken McRae, et al. 2007. Semantic fieature production norms for a large set of living and nonliving things. *Behavior Research Methods*, Vol.37, pp.547–559.

[5] George A. Miller and Christiane Fellbaum. 2007. WordNet Then and Now. *Language Resources and Evaluation*, Vol.41, pp.209–214.

[6] David P. Vinson and Gabriella Vigliocco. 2008. Semantic feature production norms for a larget set of objects and events. *Behavior Research Methods*, Vol.40, pp.183–190.