日英機械翻訳における意味解析のための単語辞書

横尾 昭男[*1] 宮崎 正弘[*2] 阿部 さつき[*3] 池原 悟[*4] 白井 諭[*1] 細井 純子[*5]

[*1]ＮＴＴコミュニケーション科学研究所 [*2]新潟大学 工学部
[*3]ＮＴＴアドバンステクノロジ（株） [*4]鳥取大学 工学部 [*5]フリー

1 はじめに

機械翻訳システムにおいて高品質な訳文を得るためには，正確な解析を行い，各単語の文法的・意味的情報を正しく決定する必要がある。これを実現するため，筆者らは，名詞の意味の体系化と，それに基づく単語辞書の構築を進めてきた。現在までに，新聞記事に出現する文を解析できるようにすることを目的として，約３０００カテゴリに分類された意味体系に基づいた４０万語に及ぶ単語辞書を構築してきた。ここでは，意味属性体系構築の基本的な考え方，単語辞書に収録した情報について報告する。

2 意味属性体系化の方針

2.1 名詞の意味体系化のねらい

従来，同義語や類義語を語義により体系的に分類・整理し，階層的な木構造にしたものして，分類語彙表[1]や角川類語新辞典[2]など人間用のシソーラスがある。これらは，近年電子化され計算機処理に使えるようになっているが，以下の点で計算機処理向きではない。

(1)分類観点が必ずしも明確でなく，上位－下位関係や全体－部分関係が識別子なしに混在していたりする。また，角川類語新辞典では，種々の連想関係をもつものが混在している。さらに，角川類語新辞典では，図書の十進分類法に基づき，大，中，小，最小分類の４階層に分類しているが，単語の意味分類を，このような固定的な枠組みにあてはめることは自然でない。

(2)計算機処理では，語義だけでなくその語義で表される対象概念の種々の見方，捉え方が必要である。たとえば，「学校」は「組織」「建物」「場所」など種々の見方，捉え方があるが，１つの分類項目（シソーラス上のノード）にしか掲げられていない。また，分類語彙表では原則として単語の多義は考慮せず，多義語を最も基本となる分類項目にのみ掲げている。

一方，計算機処理を前提としたものに，分類観点の明確なシソーラスの構築を目指した東工大シソーラス[3, 4]，概念を分類・体系化したＥＤＲ概念辞書[5, 6]などがある。これらは，上記の(1),(2)の問題点を解決しようとするものである。しかしながら，前者は分類項目数はかなり大きい（約１０００）が，収録語数が少ない試作段階のものであり，後者は収録語数が２０万語と大規模であるが，概念分類項目が約８００とやや少なく，日英機械翻訳における日本語用言を訳し分ける規則を記述するには，少々粗すぎる。語彙統計によれば，日常，普通の人が使用する単語数は３０００語で８０％のカバー率となること，英会話でよく使われる単語数は２０００～３０００語とみられること，通常使いこなされる漢字数は２０００～３０００であることなどを参考に，分類項目数は３０００を目標にした[7]。

以上の考えに基づき，対象の持つ特殊性を捨象する立場から概念化の視点(単語意味属性)を約３０００に分類し，名詞の意味属性体系を構築した。

2.2 名詞の意味体系化のポイント

名詞の意味属性の基本構成を検討するにあたり，以下の点に考慮して体系化を行った。

(1)分類観点について

上位－下位関係（is-a 関係）の他に，全体－部分関係（has-a 関係）にも着目して，対象を概念化する際の視点を階層的な木構造形式にまとめた。なお，意味属性間の関連が is-a 関係か has-a 関係かを示す識別子を木構造の枝に付与することにより，分類観点を明確にした。

階層的な木構造を基本構成とすることにより，以下の利点がある。

- 用言の文型パターンにおける名詞の意味制約条件を必要に応じて下位の意味属性を用いて細かく記述したり，上位の意味属性を用いて粗く記述できるようになる。
- 任意の意味属性のすぐ上位にある意味属性が高々１個しか存在しないことにより，意味属性体系の効率的な探索が可能となる。
- 上位の意味属性（ノード）の性質（属性）を下位の意味属性（ノード）に伝搬・継承できるようになり，下位の意味属性を定義するための記述量を削減できる。
- 上位の基本的な分類体系を崩すことなく，必要に応じて最下位の分類を細分化することにより，階層的な意味属性を拡張できる。

(2)固有名詞の扱い

固有名詞は，種類も多様で語数も多い。固有名詞を含む複合名詞の解析などでは，固有名詞について，一

般名詞意味属性より細かい精度の意味属性の分解能が必要になるため，部分的に細分化した別の意味属性体系とした。

なお，一般名詞意味属性により名詞の意味制約条件を記述した用言の文型パターンを用いて，格要素に固有名詞を含む文の意味解析を可能とするため，固有名詞にも固有名詞意味属性に対応する一般名詞意味属性を付与する。

(3)多義の扱い

単語の多義を考慮し，多義語に複数の意味属性を付与する。たとえば，「木」には，「樹木」⊂「植物」と「木材」⊂「人工物」の二つの意味属性を付与する。

(4)種々の対象の見方，捉え方の扱い

その語義で表される対象概念に種々の見方，捉え方があるものには，複数の意味属性を付与する。たとえば，「本」には，「物体」と「内容（情報）」の二つの捉え方があるので，「本」⊂「人工物」と「本」⊂「抽象物」の二つの意味属性を付与する。

(5)用言や用言性名詞の扱い

動詞から名詞に転生したサ変動詞型名詞や連用形名詞については，動作を表すのものは「事」の下位の意味属性を付与し，状態や関連を表すものは「抽象的関係」などの下位の意味属性を付与する。

また，形容詞，形容動詞から名詞に転生した「形容詞語幹＋さ，み…」型の名詞やいわゆる形容動詞語幹については，このような名詞の表している属性値に対応する属性（「性質」「状態」など）を示す意味属性を付与する。

(6)分類項目である意味属性の記述

階層的な木構造で構成された名詞意味属性体系のノードにあたる意味属性の名は，概念化の視点を表すのに最も適切と思われる単語（名詞）を用いて表現する。なお，通常の単語は多義性があるのに対して，意味属性を表す単語は，一語一義で使用している。

2.3 名詞の意味属性体系の基本構成

上記2.2節のポイントに基づき構築した，一般名詞と固有名詞の意味属性体系の基本構成について述べる。

(1)一般名詞意味属性体系の具体的構成

固有名詞以外の名詞を，外部世界に実在する実体（具体）と人間の頭の中に観念として存在する実体（抽象）に大きく二分した。具体は，また人間活動の主体となるもの（主体：人間の他に，人間の集合体としての組織，疑似人間＜準人間＞としての神仏などを含む），人間活動の具体的な場所となるもの（場所），人間活動の対象となるもの（具体物）に三分した。抽象は，また人間活動の対象となるもの（抽象物），動的属性を固定的に実体化して捉えたもの（事），実体や属性間の種々の関係や静的属性を固定的に実体化して捉えたもの（抽象的関係）に三分した。

さらに，上記の主体，場所，具体物，抽象物，事，抽象的関係は，人間用シソーラスの分類観点を参考にしながら細分化した。

このようにして，図1に示すような上位分類をもち，全体で２７００余りの意味属性をもつ木構造（最大１２段）で，一般名詞意味属性体系を構成した。

(2)固有名詞意味属性体系の具体的構成

固有名詞を地名（国際地域名，国名，行政区画名，地方名，自然地形名の他，施設・建造物名，交通路名，天体名なども含む），人名（姓，名，有名人名の他，神仏名のような準人名を含む），組織名（公共機関/企業/学校などの機関名，組合/会などの団体名など），その他の固有名詞（年号/時代など時の名，事件/行事/現象などのコトの名，言語/宗教/作品・出版物/理論・方式/法律/商品などモノの名，動物や乗り物などの愛称）に大分類した。

次に，これらを細分化することにより，図2に示すような全体で１２６の意味属性をもつ木構造（最大９段）で，固有名詞意味属性体系を構成した。

3 単語辞書

3.1 見出し語の収録条件

日本語の語は漢字，ひらがな，カタカナ，英数字など種々の字種で表記される。とくに，漢字には通常複数の読みがあり，熟語や固有名詞では特殊な読みを持

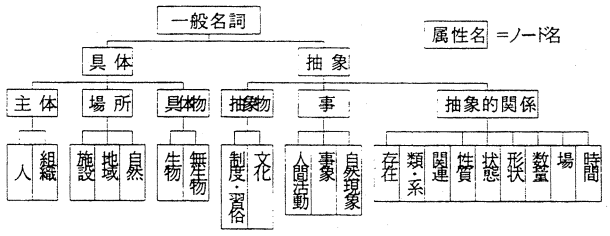

図１　一般名詞意味属性体系（上位４段まで）

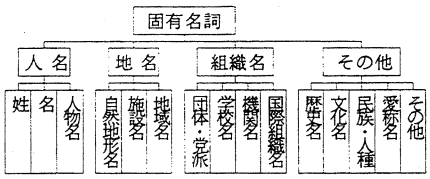

図２　固有名詞意味属性体系（上位３段まで）

つ場合も多く，同形語も少なくない。また，漢字は造語力が強く，複合語を自由に作り出せる。さらに，日本語においては，明確な正書法が確立されていないため，表記上のゆれがある。このような特徴をもつ日本語の語をどのような基準で単語辞書の見出し語として収録したかについて述べる。

(1)一般語・固有名詞・専門用語などの収録基準

単語辞書に収録する語彙は，日英翻訳実験の対象として選んだ新聞記事などのような現代国語の記述文で使用される語を対象とし，以下のような考え方に基づき一般語１２万語，固有名詞２０万語，専門用語（電気電子・情報関連用語）５万語，その他（時事用語）３万語の合計４０万語を収録した。

(a)一般語

現代国語の記述文で使用される基本語として，国語辞典に収録されているような一般的な用語を収録した。収録語の品詞は名詞，動詞，形容詞，形容動詞，副詞，連体詞，接続詞，感動詞，接辞，助動詞，助詞，記号と，多岐に渡る。

(b)固有名詞

日英翻訳実験の対象として選んだ新聞記事などに現れる地名，人名，組織名，その他の固有名詞を収録した。

(c)専門用語

実験対象の新聞記事（電気電子・情報関連）の専門用語（複合名詞などが多い）を収録した。

(d)その他

実験対象の新聞記事などに頻出する時事用語（複合名詞などが多い）を収録した。

(2)長単位語の扱い

日本語では，造語力の強い漢字により複合語が限りなく作り出される。とくに，専門用語や固有名詞には，複合語の形態をしたものが多い。しかし，このような複合語をあらかじめ単語辞書に収録しておくことはできない。そこで単語辞書には原則として語基や接辞などの短単位語を収録し，複合語や派生語などの長単位語は，解析処理により短単位語の組合せに分割することとした。日英変換辞書には，英語として適切な訳を生成するため，単位語だけでなく複合語も収録し，複合語の内部構造の解析により生成された部分複合語を基に，複合語内の単位語を組み合わせて辞書引きを行う[8]。ただし，以下のような語については，例外的に長単位で単語辞書に収録した。

(a)短単位語の組合せに分割できない長単位語

(b)３つ以上の語基が対等の関係で結合した並列語，並列語などの圧縮表現である縮退語

(c)国語辞典に子見出し語や派生語などとして収録されている一般用語

(d)格助詞相当の連語，法情報を表す連語

(e)全体が固有名詞となる複合固有名詞の一部

(f)専門用語の複合語

(3)同形語の扱い

品詞や読みの異なる同形語は，原則として別見出し語として単語辞書に収録した。なお，読みの同じ同形の固有名詞（例：清水＜姓と地名＞）は語数が多いことから，収録情報を圧縮して一つの見出し語として単語辞書に収録し，収録語数の削減による単語辞書のコンパクト化を図った。

(4)活用語の扱い

規則的な活用を行うものは，不変化部分と変化部分に分離し，別見出し語として単語辞書に収録した。助動詞，カ変動詞など，不規則な活用を行うものは，すべての活用形を単語辞書に収録した。

(5)表記のゆれの扱い

送りがなのゆれ，同一辞書内の表記上のゆれは，すべての可能な形態を収録した。異なった字種間の表記上のゆれは，表記されることの多い形態を収録した。

漢字の異体字と代表字体間のゆれ，カタカナ外来語の表記のゆれは，代表字体のみを収録し，異体字を変換テーブルで変換することで対処する。

(6)用言性名詞の扱い

名詞化するサ変動詞語幹や形容動詞語幹は，用言性名詞として単語辞書に収録し，用言化する場合，処理によりサ変動詞や形容動詞を生成する。

五段動詞から名詞化したものは，連用形名詞を収録した。一段動詞から名詞化したものは，単語辞書に収録せずに動詞不変化部分から処理により生成する。

3.2 意味属性の付与

単語辞書に収録された名詞，用言に対して，2.3節で述べた名詞の意味属性を付与する必要がある。固有名詞に対して固有名詞意味属性と一般名詞意味属性，固有名詞以外の名詞や用言に対して一般名詞意味属性を付与する必要があるが，膨大な名詞，用言に対してどのようにして適切な意味属性を付与するかが大きな問題となる。ここでは，このような問題をどのように解決し，単語辞書に収録された多くの名詞，用言に意味属性を付与したかについて述べる。

(1)固有名詞への意味属性の付与

固有名詞は種類も多様で語数も多いが，固有名詞の見出し語は，その種類毎に収集したので，その段階ごとに固有名詞意味属性を付与した。

固有名詞に一般名詞意味属性を半自動的に付与するため，固有名詞意味属性と一般名詞意味属性の対応表を作成し，それに基づき固有名詞に一般名詞意味属性

のデフォルト値を機械的に付与した。なお，機械的に付与された意味属性が正しいか否かを人手でチェックし，必要ならば意味属性の修正，削除を行った。

(2)固有名詞以外の名詞への意味属性の付与

固有名詞，専門用語以外の基本的な名詞（以下，「基本名詞」）については，既に分類語彙表や角川類語新辞典など人間用のシソーラスで体系的な意味分類がなされている。そこで，このような人間用のシソーラスを利用して基本名詞に一般名詞意味属性を半自動的に付与するため，一般名詞の意味属性体系における２７００余りの意味属性が，これらの人間用シソーラスのどの分類番号に対応するか検討し，意味属性との対応表を作成した。そして，この対応表を基に機械的に意味属性を付与した後，意味属性の追加，削除，詳細な分類への変更，順序の変更などを行った。

名詞相当の複合語や派生語，あるいは，専門用語については，これらの語の最後部の単語が全体の意味を表す主名詞となりやすいことに着目して，[9]に示す手順で一般名詞意味属性を半自動的に付与した。

この方法でも一般名詞意味属性が付与されない基本名詞や専門用語については，既に一般名詞意味属性を付与された基本名詞を参考にして，人手で一般名詞意味属性を付与した。

4 今後の課題

このようにして作成した単語辞書を実際の機械翻訳システムで使用した場合，かなりの精度で構文パターンの選択が可能であるが[7]，それでも，さらに細かい分類を必要とすることがある。そこで，今後は以下の観点を考慮しながら，意味属性体系の拡張を行う。

(1)意味属性名の定義

意味属性名は，概念化の視点を表すのに最も適切と思われる単語（名詞）を用いて記述されているが，本来曖昧性のある名詞では意味属性名を十分に表現しきれない。曖昧性のない明確な形式で意味属性名を記述する方法を検討する必要がある。

(2)分類の多観点化と明確化

意味属性体系における分類観点は上位－下位関係，全体－部分関係であるが，これ以外に静的属性，機能，構成要素など種々の分類観点が存在する。今後，意味属性体系における分類の多観点化について検討し，語を種々の分類観点から分類するとともに，観点による語と語の関係の変化を扱えるようにし，分類観点をより明確にした多次元シソーラスを構築する必要がある。

(3)意味属性の細分化と類語間の弁別特性の付与

現状の意味属性体系では，一つの意味属性内に種々の観点から見た語が混在している。今後，名詞句の意味解析や名詞の訳語選択などに意味属性体系を適用し，名詞と名詞間の意味的関連を扱えるようにするには，種々の観点から見た語が混在する意味属性を細分化して分類観点を明確化するとともに，一つの意味属性に含まれる類語間の語義や用法の差異を弁別できるような弁別特性を付与した類語弁別ネットワークを構築する必要がある。

(4)形容詞・形容動詞や副詞への意味属性付与に伴う意味属性体系の拡張

現状の意味属性体系では，実体（名詞）の静的属性の値（属性値）を表現する形容詞・形容動詞については，このような属性値に対応する属性を示す意味属性を付与することにしているが，一つの意味属性内に種々の属性に対応する語が混在している場合があり，分解能が十分でない。また，現状の意味属性体系に組込まれていない副詞について，その意味的用法の分析を進め，どのように意味属性体系に組込むかについて検討する必要がある。

(5)意味属性体系の拡張に伴うノード番号の拡張性

現状のノード番号はルートノードを‘1’とする連番が振られており，あるノードを細分化したり，新しいノードを木構造内に新設する場合，ノード番号をどのように付与するかが問題となる。今後，意味属性体系を拡張する場合，ノード番号をどのように拡張すべきかについて検討する必要がある。

5 おわりに

日英機械翻訳の研究のための意味属性体系の構築の基本的な考え方と単語意味辞書に収録した情報について述べた。今後は，意味属性体系の拡張について研究を進めるとともに，単語辞書の収録語数，収録情報の拡張を図って行く予定である。

参考文献

[1] 国立国語研究所：分類語彙表，秀英出版 (1964)
[2] 大野，浜西：角川類語新辞典，角川書店 (1981)
[3] 田中，仁科：上位/下位関係シソーラス ISAMAP1 の作成［Ⅰ］，情報処理学会研究報告．NL64-4, PP.25-34 (1987)
[4] 田中，仁科：上位/下位関係シソーラス ISAMAP の作成［Ⅱ］，情報処理学会研究報告，NL64-5, PP.35-44 (1987)
[5] 日本電子化辞書研究所：概念辞書（第２版），EDR Technical Report, TR-012 (1989)
[6] 日本電子化辞書研究所：概念辞書（第３版），EDR Technical Report, TR-020 (1990)
[7] 池原，宮崎，横尾：日英機械翻訳のための意味解析用の知識とその分解能，情報処理学会論文誌，Vol.34, No.8, PP.1692-1704 (1993)
[8] 宮崎，池原，横尾：複合語の構造化に基づく対訳辞書の単語結合型辞書引き，情報処理学会論文誌，Vol.34, No.4, PP.743-754 (1993)
[9] 池原，白井，横尾，ボンド，小見：日英機械翻訳における利用者辞書の意味属性の自動推定，自然言語処理，Vol.2, No.1, PP.3-17 (1995)