情報学基礎 28－4
（1992. 11. 9）

自然言語における意味処理

長尾 眞
京都大学工学部 電気工学第二教室

要旨

自然言語の意味には単語, 文, 文章レベルの意味などがあるが, 本論文では単語レベルでの意味について論じている. 自然言語処理の立場からは, 単語の意味は曖昧性の解消の為に必要で,20 種程度の意味素のシステムから,50 種, 500 種, さらには最近では 3000 ～ 4000 種のものが作られている. 一方, シソーラスも意味を表現しており, これと意味素の比較検討を行なった結果, シソーラスの方がよいということ, そしてシソーラスと用例との組み合せによって動詞の格構造の選択を行なうのがよいという結論を導いた. 最後に, これらのことから単語辞書に書き込むべき情報の種類を列挙して提案した.

Semantic Factors in Natural Language Processing

Makoto Nagao
Department of Electrical Engineering, Kyoto University

Abstract

Semantic marker systems in natural language processing can be classified by the number of semantic markers, such as 20 - 50, 200 - 500, and 3000 - 4000. For the disambiguation of possible plural structures of a sentence we need a very accurate semantic marker system which has 3000 - 4000 semantic markers. However, it is concluded in this paper that a thesaurus system is better than the semantic marker system for the disambiguation of a sentential meaning. Then the paper shows what kind of information is necessary for an electronic dictionary in the future.

1 はじめに

自然言語処理の立場から言語の意味を考えると次の4つに大きく分けることができるだろう. (1) 単語の持つ意味, (2) 文構造が表現する意味, (3) 文脈を考えた文の意味, (4) 文章全体の意味. これらの内で文章全体の意味として具体的なものは自動抄録であろう. その最も簡単なものはテキストからのキーワードの自動抽出を考えることができる. これに類するものとして最近筆者は目次情報をキーワードの代わり, あるいは抄録の代わりに用いることの妥当性を検証した[1]. これは科学技術論文などの場合に有効である. 特に目次の持つ章・節・項といった構造を保持した形でこれをフルテキストサーチの対象とすることによって従来のキーワード(著者のキーワード付けを含めて)を頼りにする検索よりもはるかによい検索結果が得られる可能性のあることを指摘した. そこで本論文では上記(1)および(2)の一部における言語の意味の取り扱いに焦点を当てて論じることにする.

2 意味素による意味表現と文の曖昧性の解消

私はお茶は飲むが, コーヒーは飲まない.

私はお茶は飲むが, 彼は飲まない.

という2つの文を考えよう. これらの文を純粋な句構造文法で解析すると, これらの文中の「... は」という句が主語として働いているのか, 目的語として働いているのかが一意に決定できず, いくつもの可能性を示すにとどまる. そこで言葉の持つ意味を何らかの形で明示的に表現し計算機で取り扱えるようにして, この問題を解決しなければならない.

格文法が最初にフィルモアによって導入された時は, 意味というものを明示的に示してはいなかったが, これを工学的に実現しようとすると意味というものを具体的に導入せざるをえなくなった. 例えば, 上記例文の「飲む」という動詞は, その主語として人間や他の動物を取るので, そういった種類の名詞には動物という意味素を与えるわけである. 「飲む」の目的語としてどのような単語が取られるかといえばコーヒーやジュースなどの液体か薬のような小粒の粒子あるいは粉のたぐいである. そこでこのような性質を持つ名詞には「液体」, 「粒子」, 「粉」といった意味素を与える.

このように意味素を多数設定し, 各名詞の持つ意味を考えて意味素を名詞に与える. そして各動詞がどのような格構造を取るかを表現する. 例えば, 図1は多くの意味を持つ動詞「掛ける」に対して与えられた格構造の例である. ここに示した情報処理振興事業協会の作成した日本語基本

【かける(掛ける) - サブエントリ 1】
意味: ひも状のような物を身体に付ける.
文型: N1ガ N2ニ* N3ヲ
文例: 彼女は首に真珠のネックレスを掛けている.
格要素の意味素と例:
N1 ガ [HUM] 彼
N2 ニ* [PAR] 肩, 手, 首
N3 ヲ [PRO] 眼鏡, じゅず, たすき

【かける(掛ける) - サブエントリ 2】
意味: 相手に働き掛けたり言葉を発したりする.
文型: N1ガ N2ニ N3ヲ
文例: 彼は彼女に声を掛けた.
格要素の意味素と例:
N1 ガ [HUM] 彼, 医者, 社長
N2 ニ [HUM] 彼女, 患者, 部下
N3 ヲ [ACT/LIN] 圧力 / 声, 言葉

注) '*' の付いた格要素は任意的なもの.

図 1: IPAL の格構造記述の例 [3]

動詞辞書[2](以下 IPAL と略称) は 861 語の動詞の格構造を記述するために図2に示す 19 個の意味素の体系を用いた.

3 意味素の粒度

意味素を導入してどのような名詞をどのような動詞と組にして使うことができるかは語の選択制限の問題として, もともとは Katz と Fodor によって導入された[4]. しかし, どのような意味素を設定すればよいかはこれまでほとんど議論されて来なかったといってよいだろう. 我々

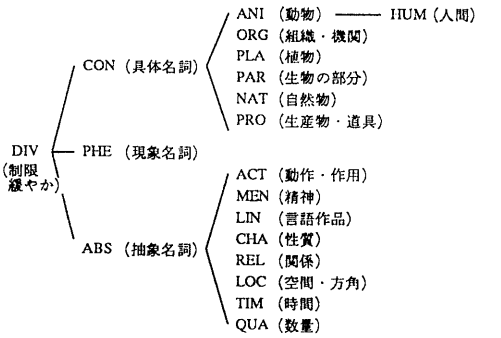


図 2: IPAL の意味素体系 [3]

が1982年から行なった機械翻訳プロジェクト(Muプロジェクト)では50余りの意味素を導入した[5]. 少ない数の意味素では曖昧な文のユニークな解析には役立たない. そこで意味素を増やせばよいというわけで, 100～200個の意味素を設定したとすると, 次にはこれらを使ってどこまで正確に名詞の意味素付与ができ, また格構造を作ることができるかという問題が出て来る. つまり微妙な意味の差を扱おうとすればするほど, この作業をする人が言語的に訓練されていなければならないのである. しかし膨大な数の単語に対するこの作業には複数の人(例えば10人以上)が従事することになるので, その人達の言語直感の差や作業の疲れなどから均質な結果をうることは非常に難しい. しかしNTTの言語グループで約3000個の意味素のシステムを作り, これを用いて格構造を作り, 非常に質の高い日本語文の解析のできるシステムを作り上げたのは大変すばらしいことである[6].

意味素を設定するときは, 思いついたものを単純に羅列すればよいというものではない. 何らかの形でこれを体系的に作って行く必要があり, そのためには図2にあるように木構造の体系として作り上げて行くことが必要であろう. このようにして意味素を作って行くとき, どのような基準で, どこまでの詳しさで意味素を設定するのがよいかが難しい問題であろう. これは結局多くの文章がどこまで正確に解析できるかという立場から, 多くの具体例に対してチェックをしてみる以外に方法はないだろう.

日本電子化辞書研究所(EDR)では現在約4000の意味素を設定して各単語の意味付け作業を行なっているという. EDRの場合もNTTの場合もその意味素体系の妥当性は多くの文についてチェックをしてみるという過程を通してその体系を完成させることになると思われる. このような場合, おそらく大半の意味素はある特別な言語表現の時にだけ使われるもので, しばしば有効に使われる意味素の数はごく限られたものであると思われる.

4 シソーラス

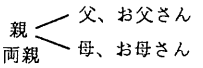
シソーラスはもともと文章を作るときにどの様な表現をとることができるかを便利に知るための辞書として19世紀中頃にRogetによって作られた. 例えば, 「頼む」という内容に対して「お願いする」, 「依頼する」, 「委嘱する」... といった表現があるということを示すことによって, 文章を書く人がその文脈などから良い表現を選ぶためのものであった.

これは本質的に同義語群を示すことであるが, これを同義語関係だけでなく, 上位語, 下位語の関係にまで拡大したものが今日広くシソーラスという語で呼ばれているものである. 例えば,

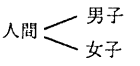
動物 ──人間 ── 大人 ── 親 ── 父

といった関係である. 日本語では分類語彙表がある. ほかに類義語辞典など種々のものが作られている.

シソーラスは原則として全ての単語を上位・下位概念と同義語概念によって分類したもので分類語彙表の場合には3万語あまりの語が分類されている. シソーラスを作る上での問題点はどのような観点から語を分類して行くかということであろう. 例えば,

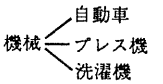


となるが, 一方では,

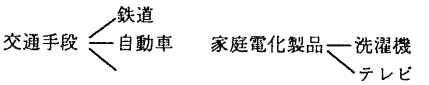


と分けてしまうと, 父は男子の下, 母は女子の下にはいり, 親の行き場所がなくなってしまう.

また観点(アスペクト)によっても分類がかわってくる. たとえば,



となるだろうが, 一方では



という関係にもなりうる.

かつては分類というのはある1つのものは分類体系全体の中の1箇所にしか入れないということが厳密な条件であったが, 計算機が発達してきた今日, 分類はかならずしもそのような性質を持たなくてよいというようになって来た. 必要と思われるだけの分類のアスペクトを次元とする多次元空間の中に各単語が位置づけられるようにすればよいという考え方である. そのためには後に述べるように各単語にあらゆる観点からの意味付け, あるいはあらゆる観点からの単語間の関係を明示することが必要となる.

5 意味素システムとシソーラスとの比較[7]

シソーラスを全ての単語に対して作るということは困難であるが, 同義語のグループは比較的安定に作ることができる. そして上位下位概念関係も一応納得の行くものは作れる. また, Roget以来今日まで, シソーラスについて

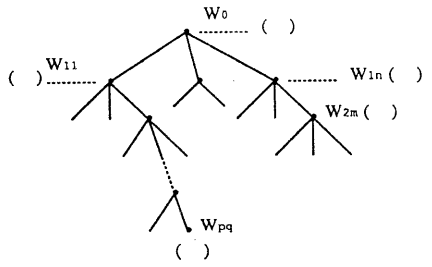


図 3: シソーラスの木

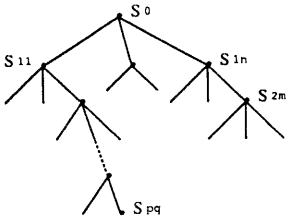


図 4: 意味素の体系

はかなりの経験の蓄積があるといえる．これに対して意味素の体系をどのように作ったらよいかについては経験が浅く，その方法は殆ど明らかにされていない．ここではこの問題について次のような考察を試みる．

まず，シソーラスは木の構造をしているので，これを図3のように示す．ここで木の各節点には 1 つの同義語の集合が対応していて，その集合の代表的な語が書かれているものとしよう．木の根に近い語ほど上位概念である．1 つの単語がいくつかの意味で使われている場合には，その数だけの同一単語がシソーラス木の適切な位置に入れられているものとする．

次に意味素のシステムについて考えよう．意味素の数が20~30 以下の場合はそれらの間に階層性を考えなくてもよいが，その数が増加して来ると木構造を作り，意味の上下関係でこの木を作らねばならなくなる (図 4)．特に数百以上,3000,4000 もの意味素の導入を考えるときにはそれは避けることのできないことである．

シソーラスや意味素の体系を作るということは文の解析や翻訳にこれを有効に使いたいということである．最も簡単な名詞句「a of b」あるいは「b の a」という表現における「of」の意味，あるいは「の」の意味を考えよう．of の意味を考えるということは「a of b」を日本語に翻訳するときに「b の a」と訳すべきか，「b である a」，「ab」などと訳すべきかの問題となる．「b の a」の場合には，この「の」を英語の「of」とすべきか，「at」，「on」，「in」，「for」などや「a's b」というような形とすべきかの問題である．

いま，与えられた句「a of b」が意味素のレベルで「Sa of Sb」(ここで Sa, Sb はそれぞれ a, b の持つ意味素) が許されているかどうかを調べることによって，もし許されているのであれば，その「Sa of Sb」の表現に対応させられている日本語表現が選択される．これによって of の意味が決定されたと考えられるのである．日本語の「の」についても同様である．

これに対して，シソーラスを用いて「a of b」の意味を決定し，その訳としてどのような日本語表現をとるかを決める過程を考えよう．そのためには「Wa of Wb」という英語表現が妥当なものとして存在し，それに対する訳も与えられているものとする．ここに，Wa, Wb はシソーラスにおいて a 及び b の同義語であるとする．

意味素やシソーラスを使って文の意味解析をするという場合のモデルはこのようなものとなるわけである．このように考えると，意味素による方法とシソーラスによる方法とにはほとんど完全な対応が存在する．そこで問題となるのは意味素の体系における意味素の数，あるいは意味の荒さ，つまり意味素の粒度である．シソーラスの場合は，とにかく 3 万 ~5 万語が木の深さにして 6~7 段の木構造に分類されている．これに対して意味素の数が例えば 5 0 の場合には，これで全意味分野をおおっているということは，これをシソーラスの木に対応させて考えれば，この木の上から 50 ノード (節) までのところを取ったということに対応する．もし意味素の数が 3000~4000 であれば，シソーラスの木の全てのノード (節) をとったことになるだろう．

意味素の体系を表現する木 (図 4) とシソーラスの木 (図 3) とは同じ形とは限らない．しかしシソーラスの木の各節点に対応する同義語集合には共通の意味が存在するから同義語としてまとめられているわけであるから，その意味が意味素の木の対応する節点の意味素であると考えることは妥当なことである．即ちシソーラスの木の各節点に書かれている代表的な単語が意味素の木の各節点の意味と見なしてよい．こう考えると意味素の木の詳しさがシソーラスの木の詳しさに等しいときはシソーラスによる方法と意味素の木による方法との意味処理上の能力は等しいが，意味素の数がそれにより少ない場合には意味素によるアプローチの方が能力が劣るということになるだろう．

6 例文, 例句を意味解析に用いることの優位性 [3]

文の意味解析を行なう場合を考えよう. 意味素の体系を用いる場合には意味素によって, どのような表現が許されているかを記述しておく. 格構造は最も典型的なものであり, 例えば図1の場合には

N(HUM) が N(PAR) に N(PRO) をかける. (1)

という形を取る. ここに, 例えば, N(HUM) は意味素 HUM を持つ名詞ということである. 「a of b」の場合も同様,

N(Sa) of N(Sb) : N(Sb) の N(Sa) (2)

the President of the US : 米国の大統領

これに対してシソーラスを用いる場合は, どういう表現が許されるかは例文, 例句によって表現する. すなわち,

彼女 が 肩 に たすき を かける. (3)

という例文からシソーラスの同義語あるいは下位語を調べて, 新しく与えられた例文がこの例文に対応するものかどうかを調べるのである. すなわち,

花子が肩にひもをかけている.

という文は上の文(3)と同種の文であると考える.

意味素を用いる場合もシソーラスを用いる場合も許される表現はどのようなものかを示す(1), (2), (3)といった表現の選択には十分な注意が必要である. このような例文に用いる意味素, あるいは単語はそれぞれの木において表現上許される最上位のものを用いねばならない.

さて, それでは意味素による方法と例文による方法は全く等しい能力を持つ方法かというと, シソーラスによる方法に優位性があるといった方がよいだろう. それは次のような理由からである.

(1) 膨大な数の意味素の体系はシソーラスのような具体的な単語のシステムを参照しなければ良いものが作れない.

(2) 意味的に許される表現を表記する時, 意味素のレベルだけで考えていては誤りをおかす危険性が非常に高い(上記(1), (2)などを例文を頭に浮かべずに書くことは困難である). これに対して例文を与える場合には作業がしやすい.

(3) 意味的に許される表現に不都合が生じたとき, シソーラスによる方法の場合には新しい例文を付加して行くだけでよく, その付加によってそれまでの体系に一切の変更を必要としない. 従って, 意味処理の精度を徐々に上げて行くことができる. これに対して意味素のシステムで行なう場合, もしその意味素のシステムがシソーラスのシステムと完全に一致している場合はよいが, そうでない場合には, 意味素の木そのものに変更を加えねばならないということが生じる. これは非常にコストのかかる変更となる.

以上のような考察の妥当性と, 意味素によるよりもシソーラスによる方が意味の区別をよりよくできるという実際の実験例をすでに別の論文で発表している[3].

7 辞書の持つべき情報

辞書はどのような情報を持つべきかについて考えてみよう. まず単語の形態的文法的情報としては, 発音, 単語の変化形, 品詞がある. 単語の変化形は言語によって種々のものがある. 英語では単数形, 複数形, 動詞の活用形, 時制による変化などがある. 日本語には男性が主として使う語, 女性が使う語, 尊敬, 対等など自分と相手との関係による単語の選択などに大きな特徴があるし, 同義語, 反対語, 上位概念語, 下位概念語, 派生語などの関係も大切である. すなわち, 1つの単語が直接的に関係を持つ他の語は何かをできるだけ多様な関係について明らかにする必要がある. 例えば, ごはんとめしの使い分け, ごはんとめしと食事の使い分けはコンピュータはもちろん, 外国人の場合にもよく分かるように記述しておく必要がある. 食事には丁寧さを表現する「お」がついてお食事といえるが, 「おごはん」, 「おめし」とは言えない等の情報も必要であろう. 動詞などに対しては先に述べた格文法に関する情報を与える必要がある.

多くの辞書はことばの意味を記述することにのみ専心し, そのことばがどのように使われるべきか, どのように使われてはならないかについては, ほとんど何らの情報も与えてくれない. これは母国語の人達を対象とし, ことばの使い方には問題が無いという暗黙の前提を置いているからであろう. しかし我々日本人でもかならずしも正しい日本語について十分な知識を持ち, その使用法に自信を持っているわけではない. 例えば, 「式典に参加して下さい.」という表現を丁寧にして「式典にご参加して下さい」という書き方をした時, これは日本語として正しいのか, 許されるのかどうか. 「式典にご参加下さい」というように, 「ご」を付けただけで, どうして「して」が取れてしまうのか. このようなことはどの辞書を見ても書かれていないのである. 敬語に関する専門書をいちいち見ないといけないようでは困るのである.

最近の辞書には派生語, 複合語などがかなり載せられるようになって来たことは喜ばしいことである. しかしもっと基本的なこと, たとえば接頭語, 接尾語の付き方や, ある言葉を文脈によって同義の違う単語に置き換えねばなら

ないことなどの微妙な用語法の説明が必要である．これは際限のないことだから書かないと全てをあきらめるというのは，良い辞書を作るという立場からは取るべき態度ではない．

辞書は言葉の意味を中心にして記述してある[8]．これがどのように表現されるべきものなのかについて考える必要がある．自然発生的に作られて使われて来ている言葉は，数学や物理学などの概念と違って明確な定義を与えることが非常に難しい．もしそのような定義が与えられるとしても，それが我々普通の人間が一度読んですぐ理解できるような平易な表現で書かれているかという問題もある．従って，言葉に対する定義は幾つもの異なった観点からできるだけ具体的な形で与えることが望ましい．それは少なくとも次のような要素を含んでいることが望まれる．

(1) 内包的定義　その言葉の内容・概念の本質を記述する．

(2) 外延的定義　その言葉の具体例を列挙することによって，それらに共通する本質を推測させる．

(3) 比較定義　その言葉の上位概念語 (類概念の語)，下位概念語 (種概念の語)，同義語，同位に位置する (並列する) 語，対立する概念の語，強い連想関係で想起される語などを示す．また，類似した語とどのように異なるかという意味上，用法上の差異を明確に示す．

(4) 要素構成的定義　その言葉の指すものがどのような要素から構成されているものか，それらの要素の相互関係は何かを示す．

(5) 性質・属性・機能・目的等の観点からの定義

(6) 関係的立場からの定義　他の概念との間に存在する原因・結果関係，部分・全体関係，前後・順序関係などを示す．

(7) 比喩的用法

(8) 歴史的説明　その語が生まれて来た経緯，その語の意味・用法の歴史的変遷を記述する．

こういったこと以外に，言葉の辞書の場合には，用例というものが大きな意味を持つ．言葉の用法を明示する方法として，文法的立場からの説明の他に，豊富な用例を示すことによって，その語の使い方を具体的に知ることが出来るからである．これは言葉の用法の外延的定義とみなすこともできる．上記の定義の中の比較定義の内容は独立したものとしてまとめられていることが多い．これはシソーラスであり，既に述べた．

参考文献

(1) 長尾真: 目次情報などを利用した図書・文献検索方式，情報の科学と技術,Vol. 42, No. 8 (1992.8).

(2) 計算機用日本語基本動詞辞書 IPAL(Basic Verbs) 説明書，情報処理振興事業協会技術センター (1987).

(3) 黒橋禎夫，長尾真: 格フレーム選択における意味マーカと例文の有効性について，情報処理学会自然言語処理研究会資料 91-11 (1992.9.18)

(4) Katz, J. , Fodor, J. : The structure of a semantic theory, Langauge, 39, 170-210.

(5) 長尾ほか: 科学技術庁機械翻訳プロジェクトの概要，情報処理,Vol. 26, No. 10, 1985.10.

(6) 池原，宮崎，横尾: 日英機械翻訳のための意味解析辞書，情報処理学会自然言語処理研究会資料,84-13, 1991.7.

(7) M. Nagao: Some Rationales and Methodologies for Example-based Approach, Proc. Workshop on Future Generation Natural Language Processing, UMIST, Manchester, July 30-31, 1992.

(8) 長尾真: 辞典形式での専門分野の知識の体系的構成法，人工知能学会誌,Vol. 7, No. 2, 1992.3.