2E-5

自然言語処理と概念の体系化について(2)

田中康仁（姫路短期大学）
吉田将（九州工業大学）

1. はじめに

「自然言語処理と概念の体系化」と題して情報処理学会第37回（昭和63年後期）全国大会で発表した。今回の発表は前回の発表の続編である。

自然言語の理解や，機械翻訳システムの開発をめざすには，人間がどのようにして考え，どのようにして情報を得ているか？　例えば翻訳を行うときに辞書はどのように利用されているか，どのような辞書を準備しているか，又蓄積された知識とはどんなものか？　これらを機械処理におきかえるためには何を準備しなければならないか等を考え，語の背景にある概念の分析を行い，概念の記述，記号化，体系化を行わなければならない。この作業は大変であるが実現しなければならない。

2. 概念の具体化

「概念」は個々の事物の共通性だけを取り出したものであり，「何か」ということを表わしたものである。

ここでは電子計算機で処理するという立場から，電子計算機で取り扱える方法，定義，記号化を考える。

また，言葉を機械処理するという立場から考えたい。ここでは1つの例を使って考え，概念の具体的定義方法を考えてみたい。又，1つの定義についての例が示されると，この部分は重要でないとか，このようにすればさらに表現がより良くなるということがあろう。それらの批判によりさらに発展したいと考えている。

「自動車」を例に取り上げて考えてみる。

「自動車」を何と読み，どのように仮名で表すかとか，品詞は何かという「言葉の面からの定義がある」

「自動車」は何らかの目的を持って作られたものであるから，利用目的などを考えることができる。

「自動車」の上位概念，下位概念は何か？

「自動車」の例をあげることによって概念を説明することも可能である。これを何らかの基準で分類することもできる。

構成要素をもって概念を説明することもできる。

また特徴，性能，性質……等の事柄も説明するために必要であるかもしれない。又概念を述べるにあたって修飾する語がある。大きい，小さい，美しい……，これらについても調べなければならない。

概念を表わす語を使用するにあたって共起する動詞等はあまり多くなく限られたものである。これらを集めることも重要である。機械翻訳においては格関係の定義も調べなければならない。

概念は語によって表現されるが，概念と語の関係の少しのずれにより各種の同義語が発生する。同義語をどのように表現するか，また概念は対立する概念，ある概念の反対の概念，否定などが考えられる。

概念が細分化したり，別の概念を含んでしまってしまい統合化される場合がある。これらも研究の対象にしなければならない。

概念は語によって表現されるが，各国語によって表現の仕方が異なるし，各国語の語が同一の対象概念を各国語で表わしてはいない。この区別も明確にしなければならない。

マンマシン・インターフェイスの面から考えると人間が読み理解することができる文章形式の表現も必要である。概念の発生した由来，起源や歴史等も必要であろう。

3. 利用面から

概念を表わす語を具体化し定義するのであるが，これについては詳細であればあるほど良いと考える。しかし，詳細になればなるほど，複雑になるし，ファイルの容量，データの収集の労力，費用も増大する。

収集したデータは常にメインテナンスを行っておかなければならない。このため利用面からこの程度で良いとする割り切りが必要である。

4. 概念の体系化

概念を表わす語はどのようなものから成り立っているか項目をあげる。

これは一つの提案と考えていただきたい。

Natural Language Processing and Systematization of Concept (2)

Yasuhito Tanaka, Himeji College
Sho Yoshida, Kyushu Institute Technology

語のもつ内容

1. 語の属性
   語の表記，品詞，発音，アクセント，仮名表記
2. 上位概念，下位概念
2.1 長単位用語，専門用語
3. 概念の特徴を表現するもの
   目的，特徴，性能，性質，利用方法，製造方法，…。
4. 具体的例による表示
5. 概念の構成要素，部分，全体の関係
6. 対立，反対，否定の概念
7. 同義語
8. 修飾語
9.1 比較関係（大小，高低）
9.2 順序関係
10. 動詞との共起関係
10.1 格関係（結合価）
11. 訳語（各国語との対比）
   各国語の語と概念の含包関係
12. 関連語
13. 語の転用
13.1 具体物 → 抽象的表現
13.2 慣用表現
13.3 連想による関係
14. 文章による表現，説明
15. 由来，起源，歴史
16. その他

4. データの収集作業

概念の定義にあたって必要なデータの一部として朝日新聞84日分の中から'の'の共起関係を集め整理している。約30万件のKWICより'の'の共起関係の抽出を始めた。

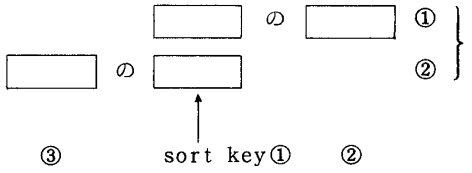
今後の研究に期待していただきたい。

概念構造の自動作成の方法として次のような方法を考えている。

(1) 'の'を中心とした新聞等のKWICから集める。

(2) 同一のものは一つにまとめ頻度情報を付加する。

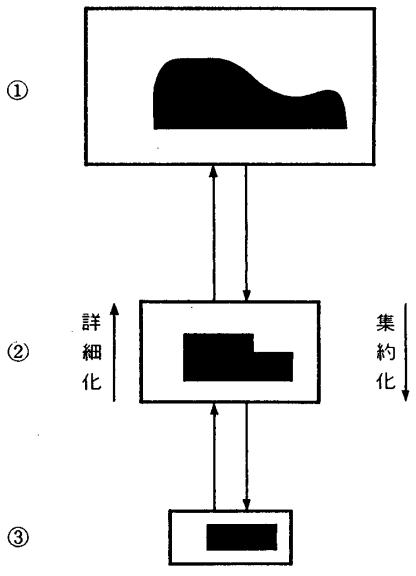
(3) 集められた共起情報を2倍に拡大し前接語，後接語を中心に分類する。



(4) 同一の語の集まりに対して次のような区分を付ける。
(i) (3)の①か②の区分
(ii) 概念区分 (1) ……
(2) ……
(3) ……
⋮
(iii) カテゴリー付けを行う。
概念を構成している詳しい内容についてカテゴリー付けを行う。

(5) 'の'を中心とした共起情報としても使いたいし，概念構造のデータとしても使えるようにしておきたい。そのようにしてDataの追加，修正，変更が容易に出来るようにもする。

大量のデータを集めると何かカテゴライズ化を行いその中から一般的な規則をみつけたいと考えるのであるが，その規則は小さな事象を捨象してできあがっているため注意しなければならない。
詳細な機械翻訳システム等を作成する場合には，一般的規則と個々の事象，例外の事象を持っておかなければならないであろう。



カテゴライズ化と問題点