

語と語の関係について

—'に'について—

Acquisition of Knowledge Data for Natural Language (No. 1)

—From Asahi News paper—

—'N1'—

田 中 康 仁

吉 田 将

(姫路短期大学)

(九州工業大学)

YASUHIKO TANAKA

SHO YOSHIDA

Himeji College

Kyushu Institute Technology

あらまし 自然言語の分析によって知識データを収集する方法について述べる。朝日新聞記事データ84日分を用いた。

知識データによる多義性の解消方法について、多義性の問題点、多義性のための幾つか方法と問題の検討を行い、この中で特に語と語の関係による知識データが多義性の解消のために有効であることがわかった。

知識データの収集方法としては、格助詞'に'を中心とした新聞データのKWICを使い、その中から手作業で知識データを集めた。

約173.3万行のKWICを解析し、5.7万種類の語と語の関係の知識データを得た。この知識データを翻訳し、整理することにより機械翻訳の多義性の解消がはかられる。翻訳等に少し費用はかかるが解決の第一歩がつかめた。知識データをさらに収集し、整理し、新しい観点から文法規則の体系化を進めるべき時期に来ている。

Abstract This paper describes the results of considering the problems and some methods for solving the multivocal problems in words by using knowledge data. As a result, it was found that the knowledge data based on the relationship of words was especially effective in solving the multivocal word problems.

The knowledge data was gathered from partially analyzing general sentences by using a KWIC list with the kakujoshi (postpositional case auxiliary)"NI(に)" as its base.

From analyzing approximately 173,000 lines of the KWIC list 57,000 types knowledge data (relationships between words) were obtained.

By translating these knowledge data, and re-arranging them, the problem of multivocal words in machine translation can be solved. Though cost may be required to translate the data, the knowledge data obtained through this study has shown some possibility to act as a method for solving the problem of multivocal words.

The time has come to gather more knowledge data, re-arrange it and systemize the grammatical rules from a new aspect.

1. はじめに

機械翻訳の研究と実用化が進み、機械翻訳システムの使い易さに目が向けられヒューマン・インターフェイスの研究、開発が行われている。その結果使い易いエディター（プレエディット、ポストエディット）、データの検索蓄積に研究が進んでいる。

しかし、ポストエディットの原因となっている多義性の問題は解決しているのであろうか？あいまい性の問題は解決しているのであろうか？専門用語辞書等については解決しているだろうか。これら本質的な問題を解決する努力が研究者においては先で、それからポストエディットの問題を考えるべきである。ここでは語の持つ多義性の明確化のためには「語と語の関係による知識」を用いる方法と、この知識データの収集方法について具体的に述べる。

2. 多義性について

どのように多義性が発生するか具体的に考える。例を用いて説明する。

例 入る

- (i) 内へ入る enter, come(=go)in get in(=into) walk(=step) in(=into)
- (ii) 押しに入る break in, enter(a house)by, force, burglar, burglarize
- (iii) 加入する join, go into, subscribe to associate oneself with
- (iv) 入学する enter a school
- (v) 含む contain, hold
- (vi) 収容し得る can accommodate
- (vii) 収入がある have, get, receive
- (viii) 始まる begin, set in

「新和英大辞典」研究社より引用。

このように幾つもの意味を我々は文章中又は音声中から適切なものを判断している。この多義性の判別を計算機で行うとすれば「入る」という一語を操作しても解決できない。何らかの他の要素と組合せなければならない。そこでこの他の要素としてはどのようなものを考えればよいのであろうか。

3. 多義性の解消方法

多義性の解消方法にはどのようなものがあるのであろうか？

- (1) 語と品詞 (5) 格文法と意味マーク
- (2) 専門用語 (6) シノーラス
- (3) 複合語 (7) 語と語の関係（語の共起関係）
- (4) 慣用表現

色々考えられるが、ここでは特に語と語の関係について述

べる。

語は色々な語と結合するが、よく調べてみると特定の語との共起関係が多いことが判る。この共起関係の強いものを多量に集め利用すれば語の多義性が解消できる。

例 ストに入る → to go on strike

しゅんに入る → to have best season

ここでは、この語と語の関係についてのデータ収集方法を述べる。

4. 語と語の関係データの抽出

4-1 一般的方針

一つの語は無限に多くの語と結合することができるので、語の活動範囲や条件を明確にすることはできないのではないかという疑問が起る。また語自身も無限にあり、これらを全て調べあげることも大変な労力と時間がかかる。しかし、実際の語を調べてみると一つの語に関係する語は限られている。

人はある状況を見てそれにふさわしいある一定の表現をする。決して別の表現をしない。



のような状況をみれば“雷が鳴っている”という表現をする。

決して“雪が降っている”とは言わない。又簡単な表現形式を望む。

複雑で長い表現はもちいられない。

我々は物事や状況から意味の最小単位を表現することを知っている。この意味の最小単位を集めなければならない。

例えば、電話という語を考えてみると、電話の特性は通信の手段、物体、場所、………というように限られる。通信の手段としての機能、電話独特の特徴は電話独特のものである。これについては語と語の関係を数えあげることは簡単であり有限である。一般的な物体、場所としての語と語の関係を数えあげることは大変困難である。

併し、これらのうち主要なものは簡単にまとめることができる。語に特有な語や使用頻度の高い語と語の関係をテーブルにまとめ、その他のものはシステムにプリセットされたデフォルト値を用いるようにする以外に方法はないであろうか。

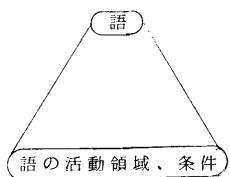
001 電話をかける	008 電話を作る
002 電話をきる	009 電話を製作する
003 電を持ち上げる	010 電話を組立てる
004 電話をこわす	011 電話を開設する
005 電話を握る	012 電話を引く
006 電を持つ	013 電話を撤去する
007 電話を改良する	014 電話を売る

- | | |
|--------------|-------------|
| 015 電話を販売する | 025 電話が鳴く |
| 016 電話を買う | 026 電話で伝える |
| 017 電話を購入する | 027 電話で話す |
| 018 電話を磨く | 028 電話で連絡する |
| 019 電話を受ける | 029 電話に出る |
| 020 電話を盗聴する | 030 電話の声 |
| 021 電話をかけなおす | 031 電話の部品 |
| 022 電話を待つ | 032 電話の金 |
| 023 電話を持たせる | 033 電話の料金 |
| 024 電話を聞く | 034 電話のベル |
| : | |

一つの語彙に関係する語彙は限られている。

表1 一つの語に関係する語は限られている。

語の持っている活動範囲、活動条件を明確化しなければならない。



1. 語の活動範囲を明確化する。
2. 語の活動条件を明確化する。

図1 語の活動状況

高いとか美しい……という語は使用頻度も高く、個別に語の活動範囲や条件を決めにくいものもある。これらについては一般的文法と“高い”とか“美しい”で最も多く使われる語の意味を含ませ、それ以外の場面で使用する特別の場合の高いとか、美しいという意味の使用条件を語と語の関係で規定しなければならない。使用頻度の低い語と語の関係については個別規則を使い、さらに一般文法を適用することになる。

自然言語の一般的知識をあげてみると次のようになる。これらを利用して総合的に自然言語解析、理解、生成を進めていかなければならない。

語の持っている知識をあげてみると次のようになる。

1. 語に関する属性
 - 1.1 語の属性
語、品詞、発音、アクセント、仮名表記
 - 1.2 長単位用語、専用用語
2. 語と語に関する属性（成立する前提条件）
 - 2.1 上位、下位の関係（シソーラス構造）

- | |
|-----------------|
| 2.2 反対語、否定語 |
| 2.3 部分、全体の関係 |
| 2.4 順序関係 |
| 2.5 比較関係（大小、高低） |

3. 語と語に関する属性(2)（成立する前提条件）

- 3.1 格による関係
- 3.2 文構成要素の共起による関係
- 3.3 儻用表現

4. 語と語に関する属性(3)（成立する前提条件）

- 4.1 連想による関係

5. 文の関係

- 5.1 因果関係（動詞と動詞）
- 5.2 場の設定条件
- 5.3 文と語の関係
- 5.4 文と文の連なりの関係

表2 語の持っている知識

4-2 知識データの収集方法

一般文の中から助詞、助動詞を利用し、KWCを用いて知識データを抽出する方法を利用した。

助詞、助動詞としては次のものを考えている。

が、を、に、へ、と、から、より、により、の、する、した、に対する、に関する………

KWCの例を次にあげてみる。

- | | | |
|--------------------|---------|------|
| 113403 隊で南東陵から第2登 | に 成功した。 | 197 |
| 113404 いて説得力のある記述 | に 成功した。 | フォー |
| 113405 二度目のタイトル防衛 | に 成功した。 | 次の防 |
| 113406 機器の協力を得て開発 | に 成功した。 | 従来の |
| 113407 景に青白く輝き、実験 | に 成功した。 | 発光雲 |
| 113408 時間同九時五分）登頂 | に 成功した。 | （13 |
| 113409 柳が判定勝ちで初防衛 | に 成功した。 | （YN |
| 113410 TKO勝ちし、初防衛 | に 成功した。 | （ロイ |
| 113411 秒KO勝ちし、初防衛 | に 成功した。 | （ロイ |
| 113412 東陵をたどり、第3登 | に 成功した。 | 4人の |
| 113413 ス）の四日間連続運動 | に 成功した。 | この磁 |
| 113414 完成、十九日、テスト | に 成功した。 | この実 |
| 113415 板の黒色部分）の開発 | に 成功した。 | これだ |
| 113416 ロドフの三隊員が登頂 | に 成功した。 | これで |
| 113417 一七五キロの高速試験 | に 成功した。 | これは |
| 113418 油を抽出する技術開発 | に 成功した。 | これは |
| 113419 戦”を挑み、割り込み | に 成功した。 | しかし、 |
| 113420 み、まんまと「城攻め」 | に 成功した。 | その公 |
| 113421 ぼうを」と訴え、再建 | に 成功した。 | もちろ |
| 113422 初のミサイル発射実験 | に 成功した。 | トライ |

113423	五日にかけて夜間登頂	に	成功した。	ネパー	③ データ数を多くすることができます。
113424	で勝ち、四度目の防衛	に	成功した。	プライ	④ 語と語の関係を網羅的に抽出できる。
113425	統一タイトルの初防衛	に	成功した。	レナー	⑤ 頻度情報も得られる。
113426	訪中、作品の買い上げ	に	成功した。	海外展	(2) 手作業を主体とした方式の特徴
113427	子を使ったモデル実験	に	成功した。	研究に	① 抽出作業でこまかい配慮ができる。
113428	の二度にわたって登頂	に	成功した。	四月二	② 再入力の費用がかかる。1件10円程度。
113429	北東慶から念願の登頂	に	成功した。	七九年	③ 17.3万件程度のKWICならばこの手法がよい。
113430	O勝ちし四度目の防衛	に	成功した。	石井は	④ 作業員の質が不均一であるため集められたデータの再分析、チェックを機械的に処理する方法を考えなければならない。このためには語の長さを適当な長さ以上は調査対象データとすることにより解決できる。しかし、個々のデータの検討もしなければならない。
113431	(十一人)が六日登頂	に	成功した。	中国登	⑤ データ収集の費用がかかる。しかし大学では学生に勤労奉仕させる場合や授業の演習の一部として行わせると安く行える。
113432	メラマンが入り、撮影	に	成功した。	中国領	抽出されたデータの一部を示す。
113433	ら七人)が四日、登頂	に	成功した。	登頂し	語と語の関係(に)
113434	化学変換装置」の開発	に	成功した。	冬の弱	
113435	員が無酸素で初登はん	に	成功した。	禿(か	
113436	の二度にわたって登頂	に	成功した。	二日に	
113437	タイプの化粧水の開発	に	成功した。	微生物	
113438	万キロワット)の受注	に	成功した。	歴史的	
113439	が二十一日、十五回転	に	成功した)写真。		
113440	字を書いてもらうこと	に	成功した)写真。		
113441	、こうした人材の育成	に	成功した「日本の経		
	KWICの例				

表3 助詞、助動詞を中心としたKWIC

ここでは格助詞「に」を選んだ。「に」を選んだ理由は「を」、「が」の知識データの収集が出来たので、次に使用頻度の高い格助詞として「に」を選んだ。対象としたデータは朝日新聞記事データ84日分を対象とした。この記事データの基礎的整備は東京大学工学部藤崎教授、亀田氏(現東京工科大学)等が行ったものである。この記事データを利用し、上述のようなKWICは機械的に容易に作成することができます。このKWICを基にして姫路短期大学の学生達にデータの抽出を行わせた延べ10数人の学生が約1ヶ月で作業を行った。データの抽出内容は図書カードに記入した。このカードを集め計算機の入力データとした。KWICを利用し、手作業で知識データを抽出した理由は和語が多いためとKWICの量があまり多くないためである。

手作業による大量の知識データの収集は単純作業の繰返しであり学問的価値が無いが、一旦集められ整理された大量の知識データは多くの人々に利用され、それから作り出される新しい知的な生産物は広く社会に利用され大きな意味を持ってくる。また、大量の知識データの多くの分野にわたる利用方法の研究も多くの人々に喜ばれ有意義である。

(1) 朝日新聞記事データの特徴

- ① 新聞は広い分野を対象範囲としている。
- ② 和語の語と語の関係を握りきれる。

③ データ数を多くすることができます。
④ 語と語の関係を網羅的に抽出できる。
⑤ 頻度情報も得られる。
(2) 手作業を主体とした方式の特徴
① 抽出作業でこまかい配慮ができる。
② 再入力の費用がかかる。1件10円程度。
③ 17.3万件程度のKWICならばこの手法がよい。
④ 作業員の質が不均一であるため集められたデータの再分析、チェックを機械的に処理する方法を考えなければならない。このためには語の長さを適当な長さ以上は調査対象データとすることにより解決できる。しかし、個々のデータの検討もしなければならない。
⑤ データ収集の費用がかかる。しかし大学では学生に勤労奉仕させる場合や授業の演習の一部として行わせると安く行える。
抽出されたデータの一部を示す。

語と語の関係(に)

Seq#				頻度
1	欧洲正面	に	移動する	1
2	下方	に	移動する	1
3	高所	に	移動する	1
4	国境	に	移動する	1
5	国連本部付近	に	移動する	1
6	山岳地帯	に	移動する	1
7	線沿い	に	移動する	1
8	大量	に	移動する	1
9	短期国債	に	移動する	1
10	同諸島近く	に	移動する	1
11	南側	に	移動する	1
12	別の場所	に	移動する	1
13	北	に	移動する	1
14	北東	に	移動する	1

表4 朝日新聞記事データより抽出した知識データ(後接語より分類)

Seq#			頻度	
1	トップ	に	あげる	1
2	トップ	に	ある	1
3	トップ	に	いる	1
4	トップ	に	おどり出る	1
5	トップ	に	かえる	1
6	トップ	に	なっている	2
7	トップ	に	なる	5

Seq#			頻度
8	トップ	に	混じる
9	トップ	に	座る
10	トップ	に	上がる
11	トップ	に	据える
12	トップ	に	選ばれる
13	トップ	に	働きかける
14	トップ	に	躍り出る
15	トップ	に	立つ
			3 4

表5 朝日新聞記事データより抽出し

た知識データ（先頭より分類）

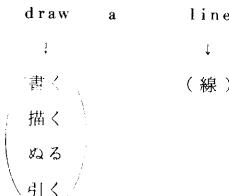
分析結果（に）

KW1Cの行数	1 7.3万件
収集したデータ（延データ）	8 3,403件
重複を取り除いたデータ	5 6,909件

(3) この方法の特徴

- ① 文を分析することによって得られた知識データであり、（作為的なデータではない）。
- 機械翻訳の訳しわけ等に適用すると知識データのヒット率が高くなる。
- ② 頻度情報が付いている。
- ③ 多くの語と語の関係が得られるため動詞の辞書が作りやすい。多くの例文を思い付きやすくなるため辞書が充実する。
- ④ 文の構文解析を行わずに得られる。
- 知識データを得るために文の構文解析をする方法が考えられる。さらに、構文解析の構文木を減らすために知識データを必要とする、という矛盾から抜け出せる。
- ⑤ 機械翻訳において訳文の生成がより適切に行なうことができる。

例



知識データに「線を引く」があるため、この訳語を優先させる。

⑥ ボトム・アップのアプローチである。

(4) 知識データ抽出時に判ったこと

- 1) 格助詞以外の「に」のデータが入った。格助詞の

「に」を中心としたデータの抽出を行ったのであるが、形容動詞の連用形が入ってしまった。これらについては形容動詞の語幹が名詞性が強いものについては知識データとして取り入れ、他のものは削除した。それ故このデータを取り扱う際には少し注意が必要である。

2) 助詞相当語

「に○ ○」となる助詞相当語がある。これについては整理された資料がないため知識データ抽出過程でまとめた。これは添付資料1で約 個の語をあげてある。添付資料2でこれらの語の使用例を例示した。

・助詞相当語の接続

名詞に接続するものと動詞に接続するものがある。

名詞	「に」から始まる助詞相当語	動詞
----	---------------	----

名詞	「に」から始まる助詞相当語	名詞
----	---------------	----

例

研究について述べる

会社に対する批判

・助詞相当語と格助詞、助詞

助詞相当語は格助詞、助詞に置き換えることが出来るがそれらは常にそのように出来るのか、ある動詞との関係についてのみ成立立つか調べておかなければならない。

について → を

・助詞相当語の頻度とシステムへの組込み

助詞相当語の使用頻度を調べどれだけのものを自然言語処理システム（特に機械翻訳システム）に組み込むかを考えなければならない。

・助詞相当語を分類すると次のようになる。

「に」より始まる助詞相当語の次に接続するものが、名詞（体言）か動詞（用言）かで区別できる。さらに次のようにもなる。

「に」より始まる助詞相当語

格助詞相当語	文型パターンに入れる
	文型パターンに入れない
助詞相当語	・助詞として独立の機能を認めるもの
	・助詞として独立の機能を認めないもの

・助詞相当語は同じ格助詞が二度使われる際に発生する場合がある。

例1

列車が東京駅で遅れている。……………①

列車が脱線で遅れている。……………②

①と②を合成する。

△列車が東京駅で脱線で遅れている。……………③

列車が東京駅で脱線によって遅れている。………④

例 2

太郎がガンで死んだ。……………⑤

太郎が自宅で死んだ。……………⑥

⑤と⑥を合成する。

△太郎がガンで自宅で死んだ。……………⑦

太郎がガンになって自宅で死んだ。……………⑧

このような文の合成時に発生するとも考えられる。どのような場合に使われているか今後の研究が必要である。

・助詞相当語の動詞と動詞の区別

助詞相当語は「に」に動詞が付いたものが多いこの動詞と文中の中心的役割をなす動詞の区別をしなければならない。

それ故、単純に文字列だけで助詞相当語であると判断は出来ない。

4 - 3 今後の課題

① 「語と語の関係」の知識データを全部翻訳する。

・1件当りの翻訳・チェック費用 5 0 0 円

・5.7万件の翻訳費用 2,850 万円

・1日当りの作業量(1人) 50 件／日

・延人日(3人で約2年間) 1,140 人日

費用が限られた場合としては頻度の高いものから翻訳する方法と、ある動詞から順次翻訳する方法が考えられる。一部翻訳した内容を最後に示す。

② 「の」、「から」、「へ」…等、「に」以外の助詞について「語と語の関係」の抽出を試みる。

③ この実験では集められなかったデータ等について、対象とする分野が異っていたためか、使われることが無くなってしまったか、等を検討する必要がある。

④ 機械翻訳システムや仮名漢字変換システム、音声や文字認識システムへ応用し、実用化する。

⑤ 語と語の関係でも多義性が判別できない場合が少し発生する。

これについては今後さらに検討しなければならない。

⑥ シソーラスとの照合

この5.7万種類の知識データと照合することによって不足している知識データを補充するとか、シソーラスの概念分類をさらに詳しく意味分類し、機械翻訳の多義語の判別、その他に役立つ。この資料は5.7万種類ある。これは膨大なデータであるから何かカテゴライズすべきであるという考え方がある。しかし、何かを

カテゴライズするとその例外が発生する。又、動詞をカテゴライズするとそれに対応する名詞もカテゴライズしたくなる。名詞に何か区分を付けようすると数十万の名詞があらわれる。労力の最適化も全体の作業の中から考えなければならない。ここではシソーラスとの照合による解決を提案する。そのため機械可読の大規模なシソーラスが提供されることを期待する。

① シソーラスとの照合の意味(I)

雪が降る
雨が降る → $\left\{ \begin{array}{l} \text{あられが降る} \\ \text{ひょうが降る} \end{array} \right.$ 雨、雪、あられ、ひょうが同一の意味マーカ上にあるか?

シソーラスとの照合による知識データの拡張意味マーカの確認

② シソーラスとの照合の意味(II)

語と語の共起関係とシソーラスとの照合は次のような意味がある。

- ・シソーラスの正しさの検証に役立つ
- ・意味マーカの細分化、統合化に役立つ
- ・訳し分けの判断と例外の抽出に役立つ

③ シソーラスとの照合の意味III

語と語の関係の知識データとシソーラスを組合せることにより、どの概念と動詞が結ばれるかを知ることができる。

A 1 と B という動詞の間に結合関係があるとわかれば A 1 が属する A 2 グループ内全部の語に B という動詞が結合するか否かを検討することができる。
もし A 1 個有のものであれば A 1 と B の結合とみることができる。
もし A 2 のグループ内の語と B が結合することがわかればそれは同一の訳語を取るか否かを調べる。さらに A 3 まで拡大し、B との結合を調べる。同様の方法を取りシソーラスの上位概念へと発展させて考える。語と語の結合を全ての語について調べること

語と語の共起関係

は出来ないので、該当する動詞との結合を語と語の関係とシソーラスとの照合により機械的に知り、その後、
語のシソーラス

図 2 語と語の共起関係と語のシソーラスとの関係

該当個所を局所的に詳細に調べる。このようにすると大巾に労力の削減をはかることができる。
このためからも語と語の関係の知識データでは前接語は基礎的概念語になるようにしている。

⑩ シソーラスと長単位用語

文の中に使われる語には複合語や長単位用語が多くもちいられている。これら長単位用語から基礎語

学校

大学
小学校
中学校
高等学校
各種学校
洋裁学校
専門学校
大学校

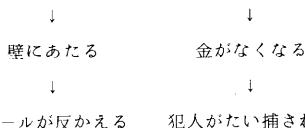
岡山大学
九州大学
京都大学
東京大学

を導く方法があるか、長単位用語の多くの語がシソーラスの体系に組込まれていなければならぬ。

図3 語と語の関係をさらに発展させるために小さい意味の集合を考える。小さい意味の集合に上位、下位の意味の集合を作る。

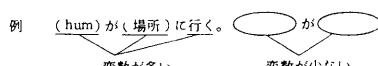
⑦ 今後の日本語の研究は単に単語を電子化し応用するにとどまらず、語と語の結合、日本語の概念結合、英語の概念結合、その他各國語の概念結合、その対比について研究を進め、電子化ファイルへと発展すべき時期にきている。これは点から線への発展である。線から面への発展は何かまだはっきりしないが次のようなものと確信している。幾何学では面を構成するものは2本の直線と直線と点の関係によって規定される。

例 1) ポールを投げる 2) 盗人(どろぼう)に入る



このような動詞句の部分連鎖の組が重要になる。動詞の因果関係、文の場の設定状況の解析とそのデータの蓄積が今後重要な課題である。

⑧ 結合価文法を考える人が多勢いるが、結合価文法は語と語の関係より変数が多い。変数が多ければ自然言語の中で一致率が悪くなる。一致率を上げるために



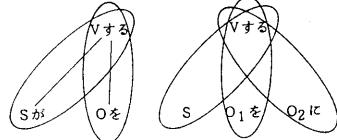
結合価の文型パターンを多くしなければならない。しかし文型パターンは2~3万程度である。この程度で一致率を向上させるためには意味マーカをあらくしな

ければならない。しかし、意味マーカを粗くすると機械翻訳の多義語の解消は出来ない。何を応用分野として考えるかにより文法も考えなければならない。

⑨ 結合価文法に一度に飛ぶのではなく次のように考えてみてはと考へる。

文を単語単位に分解するのではなく、文の中心をなす用語と各語を結びつけて分解する。このように分解すると意味がつかみ易くなる。

SがOをVする, SがO₁をO₂にVする



S→V O→Vのような素片を多くあつめてみると。これによって単語の持っている曖昧さを減らすことができる。この組合せを収集し体系化することが重要である。さらにS→V O₁→V O₂→Vのようなものでも曖昧さがあるものはO₁→O₂→Vのような連結を考えればよい。

⑩ 語と語の関係の知識データが多量に安価に入手可能になると自然言語の研究も新しい方向に進まねばならない。哲学の言葉に「量的拡大は質的変化をもたらす」とあるように次の発展が必要である。語と語の関係による知識データが安価に大量に入手可能になる時代を迎えた。

- ①文法の体系化(単純化、詳細化)をすることができる
 - ②構文解析における構文木の多発防止
 - ③機械翻訳の多義性の解消
 - 訳の向上がはかられる
 - ④文字認識、音声認識の精度向上をはかる
 - ⑤同音、同形異義語の判別を簡便する
 - ⑥自然言語処理の意味解析の発展をうながす
- この方法による知識データの収集は成功したが、今後各種の方法で知識データが増えると思われる。これについては次のことを考へなければならない。

5. 知識データの評価

知識データの収集方法が確立し、知識データが大量に収集できるようになってきた。今後は知識データの評価を行い、何が不足しているか、収集する知識データの重複はどの程度発生しているか、どのような分野の知識データが不足しているか等を検討しなければならない。

また集められた知識データの追加、修正が簡単に行えるような環境を作つてゆかねばならない。知識データ抽出作業は第一歩を進めた段階である。今後このデータを機械翻訳システムまで組込むとすると次のような段階を通らなければならない。

シソーラスについても評価を行う方法の確立が必要である。このための条件は何かをあげておかなければならぬ。

- ① シソーラスの語数
- ② シソーラスと応用分野
- ③ シソーラスの内容の公開、機械可読性
- ④ シソーラスと各種利用ユーティリティの整備
- ⑤ 知識の継承と推論システムの関係

等が考えられる。

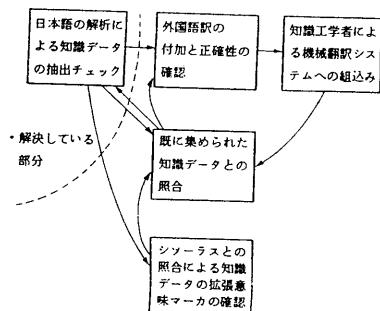


図3 語と語の関係の知識データが機械翻訳システムに組込まれるまでの作業プロセス

おわりに

機械翻訳の一つの大きな問題点である多義性の解消について知識データを利用することで明るい見通しを与えることができた。自然言語の分析は大変な仕事なので、なるべく規則による解決をはからうとするが、規則にはある限界があり、

細かい部分には効果がない。細かい部分を考えるにあたっては Bottom up による自然言語の解析と知識データの収集により、規則の大系化、再構築が必要である。この作業は大変根気のいる作業である。一面では理論的でない面があるが、これは次の step への発展のために通らなければならない道筋であると信じている。

ただ単純な知識データの収集と、問題の解決ではない。知識データを十分に集めれば体系化しやすく、何が主体か例外か判りやすくなる。

高品質の機械翻訳システムや文章理解システムを実現してゆくためには機械に知識をうめこまなければならない。この知識体系がシソーラス体系であり、概念体系である。

最後に、この研究のために朝日新聞 KW 1 C Tape を利用させて下さった東京大学工学部 藤崎教授、亀田氏に深く感謝の意を表します。

さらに、この研究の一部は文部省の科学研究費によって行った。

この「に」の知識データの校正にあたっては日本電子化辞書㈱の荻野孝野さんの助言により稻田順子さん(荻野の友人)に書いていただいた記して感謝の意を表す。又データ収集のために協力してくれた姫路短期大学の学生に感謝する。

1	～	に	～		3 4	～	にいえば	～			
2	～	にも	～	→～	に	3 5	～	にいたる	～		
3	～	には	～	→～	に	3 6	～	において	～		
4	～	にと	～		3 7	～	における	～			
5	～	にて	～		3 8	～	にかぎり	～			
6	～	にから	～		3 9	～	にかけて	～			
7	～	にして	～		4 0	～	にかわる	～			
8	～	につき	～		4 1	～	にくらべ	～			
9	～	につき	～		4 2	～	にくわえ	～			
1 0	～	につぐ	～		4 3	～	にさいし	～			
1 1	～	につけ	～		4 4	～	にしては	～	→～	にして	～
1 2	～	につれ	～		4 5	～	にしても	～	→～	にして	～
1 3	～	にでも	～		4 6	～	にそった	～			
1 4	～	にとり	～		4 7	～	にそって	～			
1 5	～	になく	～		4 8	～	にたいし	～			
1 6	～	になる	～		4 9	～	について	～			
1 7	～	により	～		5 0	～	につづき	～			
1 8	～	による	～		5 1	～	につづく	～			
1 9	～	に当り	～		5 2	～	につれて	～			
2 0	～	に致り	～		5 3	～	にできた	～			
2 1	～	に限り	～		5 4	～	にとって	～			
2 2	～	に限る	～		5 5	～	になって	～			
2 3	～	に際し	～		5 6	～	になった	～			
2 4	～	に対し	～		5 7	～	にはんし	～			
2 5	～	に比べ	～		5 8	～	にむけて	～			
2 6	～	に加え	～		5 9	～	によって	～			
2 7	～	に従い	～		6 0	～	によらず	～			
2 8	～	に続き	～		6 1	～	によると	～			
2 9	～	に伴い	～		6 2	～	にわたり	～			
3 0	～	に反し	～		6 3	～	にわたる	～			
3 1	～	にあたり	～		6 4	～	に対する	～			
3 2	～	にあたる	～		6 5	～	に対して	～			
3 3	～	にいうと	～		6 6	～	に代って	～			

参考文献

- (1) 田中康仁、吉田 将 自然言語の分析による知識データ情報処理学会自然言語処理学会
5 4 - 3 1 9 8 6 . 3
- (2) 田中康仁、吉田 将 自然言語の分析による知識データの収集「自然言語処理技術」
シンポジウム 1 9 8 4 . 1 1
- (3) 田中康仁、吉田 将 Acquisition of Knowledge Data by analyzing Natural Language
11th International Conference on Computational Linguistics COLING '86 1986. 8
- (4) 勝俣詮吉郎編 新和英辞典 研究社
- (5) 田中康仁 専門用語の自動抽出 第17回情報科学技術研究集会発表論文集 日本科学技術情報センター
1 9 8 0 . 1 0
- (6) 鈴木重幸、鈴木康之 日本語文法・連語論(資料編) 言語学研究会編 むぎ書房 1 9 8 3
(この資料は国語学の研究者が連語として取り扱い動詞の分類を行っている。)
- (7) 田中康仁 語と語の関係による知識データについて
「計量国語学と日本語処理」—理論と応用— 秋山書店 1 9 8 7 . 3
- (8) 田中康仁、吉田 将 知識データ(語と語の関係)に多義性の解消 情報処理学会自然言語処理
6 0 - 3 1 9 8 7 . 3
- (9) 田中康仁 語と語の関係解析用資料ーー“を”を中心とした。解説編(I)、(II)
文部省科学研究費特定研究「言語情報処理の高度化」 総括班 1 9 8 7 . 3
- (10) 田中康仁 語と語の関係解析用資料ー朝日新聞記事データ分析ーー“を”を中心としたー解説編、資料編(I)、(II)
文部省科学研究費特定研究「言語情報処理の高度化」 総括班 1 9 8 7 . 1 1
- (11) 首藤公昭、樋原斗志子 日本語の文構造のわく組みを与える表現ー機能カテゴリーと接続ルールー^{福岡大学総合研究所報第63号抜刷} 1 9 8 3 . 3
- (12) 首藤公昭、樋原斗志子 日本語の文構造のわく組みを与える表現ー構造的意味情報の整理ー^{福岡大学総合研究所報第63号抜刷}
- (13) 日本電子化辞書 「概念辞書(第1版)」 日本電子化辞書㈱ T R 0 0 7 1 9 8 8 . 1 1
- (14) 日本電子化辞書 「単語辞書(第2版)」 日本電子化辞書㈱ T R 0 0 6 1 9 8 8 . 1 1