

連載講座



自然言語処理入門—V—

文法と辞書を作ろう†

岡田直之† 中村順一†

機械処理に必要な文法データならびに辞書データは、どのようにして作成すればよいのであろうか。最終回は、この問題に取り組んでみよう。

6. 言語データ

自然言語処理におけるデータの重要性が一般的に認識され始めたのは、1980年代に入ってからのことである。そのため言語データを蓄積したり管理したりする理論や技術は、まだ十分に成熟しているとはいえない。自然言語処理の研究者や技術者は、現在、この大きな課題に取り組んでいる最中である。

言語データを作成する上で最大の問題は、データの規模が膨大であるという点である。文法データに関しては、かなりの数の規則を作っても、また新たな言語現象に遭遇してしまう。辞書データについても、文法データと同様の問題があるだけでなく、表層上明確な形式をもたないデータであるためテキストから抽出しにくいという、やっかいな問題がある。

初級英語を素材として、6.1では実際に文法データを、また6.2では辞書データをそれぞれ手作業で作成する。これらを通じて基本的な言語データを実際に蓄積する技術を身につけよう。最後に6.3では、機械辞書の作成やその自動化など、今後の課題を考える。

6.1 文法データ

まず初めに、表-1に示す英文テキストをご覧願いたい。典型的な初級英語の例文であるが、さてそれらに対しどのように取り組めば要領良く文法規則を抽出できるであろうか？ さらに表-1を発展させ、通常の英文まで考慮に入れて文法体系

表-1 英文テキスト

1. This is a bird.
2. Is this a book?
3. Yes, it is.
4. This isn't a boat.
5. Do you have a pen?
6. No, I don't.
7. I don't have an orange.
8. Oh, you are the cowgirls!
9. She often plays with Mika on Sunday.
10. Ellen is smart.
11. Which is your guitar?
12. What does she make?
13. When do you play tennis?
14. Look at Mr. Reed.
15. There is a house in the picture.
16. We can see many ships in the harbor.
17. The South Island is larger than the North Island.
18. Ellen said, "I have a question, Dad."
19. You look prettier in a skirt.
20. How pretty you look!
21. He will come to Seattle with his parents next month.
22. Will you show me around the city later?
23. He became a student at that school.
24. He gave his friend the cartoon.
25. English is understood by many people.
26. The pass was so high that he could not score a goal.
27. They called their capital Cusco.
28. You asked me to open the window.
29. I know how to swim.
30. Taro got a big world globe from a man who worked with his father.
31. She takes care of me, because my mother died a long time ago.
32. When I saw your sister's blue eyes, I remember my girl friend.
33. We learn that we have to use both the mind and the eyes when we look at things.

注. 中学校英語教科書「ニュープリンス 1-3」の中から、各レッスンで課題となっているような文を主として取り出した。

を構築するには、どのように取り組めばよいのであろうか？

一般に、膨大なデータを整理するとき不可欠なことは、作業の中に1本の線が通っていることである。この線に沿って作業を進めていけば、漏らしたり重複したりせず、あるいは複雑に入

† Introduction to Natural Language Processing: Development of Grammar and Dictionary by Naoyuki OKADA and Jun-ichi NAKAMURA (Department of Artificial Intelligence, Kyushu Institute of Technology).

†† 九州工業大学情報工学部知能情報工学教室

り組んだ理象を解き明かすことができる, という
 拠り所をもつことである. そのための有力候補
 は, 基本的なものと応用的なもの, あるいは標準
 的なものと特殊なもの, という線であろう. そこ
 で, 基本ならびに標準について少し考えておこう.

基本の条件

1. 対象世界の物事をとらえるとき, 出発点とな
 る性質をもつ.

2. 出発点の性質は, 対象世界において重要であ
 り, それ以上分解すると物事としての性質が損な
 われる.

3. 対象世界の物事は, それらを合成したり変形
 したりして構成される.

標準の条件

1. 対象世界の物事を比較するときに, ひな型
 (手本) となっている.

2. 形式が単純で, 入り組んでいない.

3. 広い範囲にわたって適用できる.

以下では, これらの線に沿いながら, 句構造文
 法 (PSG) のデータを作成しよう. 前回の 5.3.2
 の最後でも触れたが, 解析と生成では文法規則に
 違いがある. 以下では, 解析の立場から取り組も
 う. なお, 実際に文法を作成する場合には, 言語
 学で提案されている理論を大いに参考にすべきで
 ある. しかし, ここでは文法の工学的な作成方法
 を実感してもらうという点から, 細かな点につ
 いては, 言語学的に異論がある規則もあえて採用
 する.

6.1.1 文法カテゴリ

文法の世界では, 構文カテゴリと語彙カテゴリ

表-2 構文カテゴリ

記号	意味	例
s	文	This is a bird. Do you have a pen?
np	名詞句	pen, my girl friend
vp	動詞句	is smart, play tennis
ajp	形容詞句	smart, very pretty
adp	副詞句	often, a long time ago
prp	前置詞句	in the harbor, by many people
infp	不定詞句	to open the window
thatc	that-節	that we have to use the mind

を使って規則を表現する. これらのカテゴリの定
 め方は文法体系に大きな影響を与える. そこでま
 ず中学の1-3年版の教科書に目を通してみよう.

そうすれば必要な構文カテゴリと語彙カテゴリを
 表-2 と表-3 に示すものにまとめることができよ
 う. 記号は, できるだけ各カテゴリの英文名を思
 い起こしやすいような省略形を用いている.

6.1.2 基本構文と構文要素の規則

基本構文

基本/応用という観点から, 構文上まず注目さ
 れるのは単文である. 単文は述語としての動詞を
 一つしか含まず, 構造が簡単である. 学校文法では
 単文の構造が5文型として一応整理されている.
 この5文型が機械処理にも適しているかどうかは,
 実際にシステムを作成してみないと分からない.
 しかし英語教育における長い経験から, これ
 らが標準的な英文であることは事実である. この
 ことを意識して, 表-1 の中から単文を取り出し,
 表-4 のように整理した. ただし単文でも, 疑問
 文や否定文など変形しているものは除いている.

それでは, 表-4 に沿って句構造規則を作ろう.

表-3 語彙カテゴリ

記号	意 味	例	記号	意 味	例
nou1	普通名詞	boat, orange	pr	前置詞	in, with
nou2	固有名詞	Ellen, Cusco	coj1	従属接属詞	when, because
prn1	人称代名詞	I, you	coj2	等位接続詞	and, but
prn2	指示代名詞	this, it	ipn	疑問代名詞	which, who
vrbl	動詞	see, come	iad	疑問副詞	why, when
vrbl	be- 動詞	is, are	rpn	関係代名詞	which, who
aux	助動詞	will, can	that	that- 節マーカー	that
be	be- 助動詞	is, are	to	不定詞マーカー	to
have	have- 助動詞	have	not	否定マーカー	not
do	do- 助動詞	do	int	感嘆マーカー	oh
det1	不定冠詞	a, an	yn	yes/no- 応答	yes, no
det2	定冠詞など	the, this	com	コンマ	,
adj	形容詞	pretty	prd	ピリオド	.
adv	副詞	often	qst	疑問符	?

表-4 基本的な単文

[標準]

S+V

9. She (often) plays (with Mika on Sunday).
21. He (will) come (to Seattle with his parents next month).

S+V+C

1. This is a bird.
10. Ellen is smart.
19. You look prettier (in a skirt).
23. He became a student (at that school).

S+V+O

16. We (can) see many ships (in the harbor).

S+V+O+O

24. He gave his friend the cartoon.

S+V+O+C

27. They called their capital Cusco.

[特殊]

15. There is a house in the picture.
17. The South Island is larger than the North Island.

注. 標準文では、副詞句、前置詞句、助動詞など動詞の修飾詞はかっこでくくっている。

まず、文は名詞句と動詞句からなる、という骨組みは共通なので、その規則を作ろう。

- s --> np vp. (1)

次に5文型の動詞句に注目して、その構造を記述してみよう。とりあえず、動詞の修飾詞は除く。

- vp --> vrb1. (2)

- vp --> vrb2 np. (3)

- vp --> vrb2 ajp. (4)

- vp --> vrb1 np. (5)

- vp --> vrb1 ajp. (6)

S+V が(2)に、そして S+V+C が(3)~(6)に対応していることは、すぐお気づきであろう。be-動詞は重要な振舞いをするので、vrb2 として区別してとらえている。

次に、S+V+O の動詞句を表現しよう。

- vp --> vrb1 np. (5')

形としては(5)と同じ内容である。つまり表-2の構文カテゴリでは、補語が名詞句の場合、S+V+C は S+V+O と区別できない。そこで(5)の規則は(5')と共通とし、区別が必要になれば6.2で議論する意味素性を用いることにしよう。

最後に、S+V+O+O と S+V+O+C、それぞれについて記述しよう。

- vp --> vrb1 np np. (7)

- vp --> vrb1 np np. (7')

前記と同様の事情で、(7)を共通の規則としよう。

【問6.1】 表-1 では、中学校教科書の調査の際、大切な例文を見落としている。そのため(1)~(7)の規則でとらえられない基本的かつ標準的な文がある。その例を示し、対応する規則を作りなさい。

解答例

一般に、言語データの調査において、規模が大きくなるにつれ多少の見落としや重複は避けられない。その場合大切なことは、抛り所としている線上で論理的にそれが発見できることであろう。学校文法で学習した経験から、補語には名詞句に加えて形容詞句もあることを知っている。(3)と(5)に関しては、それぞれ(4)と(6)でそのことが考慮された。しかし(7)に関しては、S+V+O+Cにおける名詞句だけがとらえられた。そこで、形容詞句を補う必要がある。

例文

S+V+O+C: He found everything new.

規則

- vp --> vrb1 np ajp. (8)

以上により、基本的かつ標準的な構文規則ができた。いわゆる5文型では、各構文要素の機能を重視するのに対して、(1)~(8)では表-2と表-3のカテゴリに従って動詞句の型を整理した。

次に、前置詞句や副詞句など動詞を修飾する要素を、少し考えよう。表-4における動詞の修飾詞を見ると、

- vp --> vp adp. (9)

- vp --> vp prp. (10)

という二つで often を除いてすべて処理できる。たとえば21の例文に関しては、まず will come が vp としてまとまった後(その規則はまだ示されていない)、その vp と to Seattle が(10)でまとまり、次にその vp と with his parents が再び(10)でまとまる。さらにその vp と next month とが、(9)によってまとまる。

最後に、特殊な単文15と17を集めて、“用例集”を作ろう。

- s2 --> there vrb2 np prp. (11)

- s2 --> np vrb2 adj than np. (12)

これらのほかにも、特有の構造をもつ単文が出現するであろう。そこで、特殊な単文はカテゴリ s2 で表しておく。s と s2 をまとめて、次を得る。

- sent --> (s; s2). (13)

ただし, (s; s2) は, 二つの要素のいずれかを選ぶことを指示しており, 実質的には sent → s と sent → s2 という二つの規則と同じである。

構文要素

次に, 単文を構成する要素の規則を考えよう。基本的な名詞句から始める。表-1 の例文を見れば, 次の規則が得られる。

np → (noul; nou2; prn1; prn2). (14)

np → (det1; det2) np. (15)

np → ajp np. (16)

特に説明の必要はなからう。

少し複雑な名詞句としては, 例文にはないが, 前置詞句や不定詞句をとともう, 次のようなものがあるだろう。

np → np prp. (17)

np → np infp. (18)

次に, 助動詞に関する規則を作ろう。助動詞は, 例文 16 や 21 から, 以下のように動詞を修飾する。

vrbl → (aux; be; have) vrbl. (19)

vrbl → (aux; have) vrbl. (20)

ただし aux は時制や様相の助動詞, be は進行形の助動詞, have は完了形の助動詞である。以上のほかにも, 形容詞句, 副詞句, 前置詞句などの規則が必要であるが, 紙面の都合で割愛する。

6.1.3 複合構文の規則

再び表-1 に戻ろう。次は, 基本構文を変形したり合成したりして得られる, “複合構文” の規則について調べよう。

複合の型

学校文法では, 受け身文, 疑問文, 否定文などの変形操作あるいは重文や複文などの合成操作を通じて, 複雑な文の構造を説明する。まずこの考えに沿って, 表-1 の未処理の文を整理してみよう。その結果を表-5 に示す。“基本的ではない”ものが案外簡単な文にもあることがよく分かる。

表-5 では, 複合文を大きく“変形文”と“合成文”とに分けた。変形文は, 文字どおり一つの文を変形したもの, また合成文は二つ以上の文を合成したものである。さらに合成文は, “埋め込み文”と“結合文”とに分けた。埋め込み文では, 一つの文のある要素に他の文(あるいは述語を含む句または節)が埋め込まれる, また結合文では, 二つの文がある形式で結合する。たとえば

表-5 複合文

[変形]

受け身

25. English is understood by many people.

疑問

22. Will you show me around the city later?

2. Is this a book?

5. Do you have a pen?

11. Which is your guitar?

13. When do you play tennis?

12. What does she make?

否定, 感嘆, 命令他

7. I don't have an orange.

4. This isn't a boat.

8. Oh, you are the cowgirls!

20. How pretty you look!

14. Look at Mr. Reed.

6. No, I don't.

3. Yes, it is.

[合成]

埋め込み

28. You asked me to open the window.

29. I know how to swim.

18. Ellen said, “I have a question, Dad.”

結合

30. Taro got a big world globe from a man who worked with his father.

31. She takes care of me, because my mother died a long time ago.

32. When I saw your sister's blue eyes, I remember my girl friend.

埋め込み+結合

33. We learn that we have to use both the mind and the eyes when we look at things.

[特殊]

26. The pass was so high that he could not score a goal.

表-5 の 28 においては, ask の第 2 目的語に to open the window が埋め込まれており, また 31 では she takes care of me と my mother died a long time ago の二つが結合している。結合文では, 分解すると元の文が完全な形で得られるが, 埋め込み文では分解すると埋め込まれる側が不完全なものとなる。

このような複合文の性質を記述するには, しっかりした見通しに従って, まず文法理論を決めなければならない。第 2 回の 2.1.2 で述べたように, 変形操作に比較的忠実な変形文法 (TG), あるいは PSG を拡張した一般化句構造文法 (GPSG), さらに機械処理を意識した単一化文法 (UG) などがある。以下では, 読者に手軽に実験システムを作成していただくことを意図して, 若干拡張は

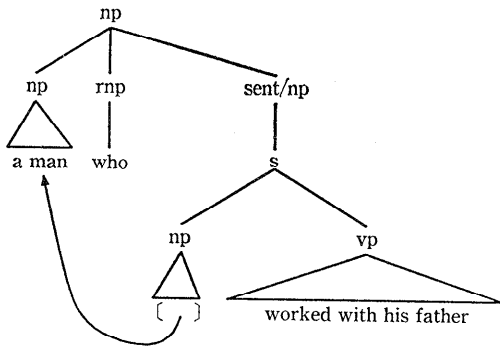


図-1 “痕跡”をもつ節の構造

するが、PSG を基本にして確定節文法 (DCG) 形式で変形や複合をとらえよう。簡単のため、形態素などの細かい処理は省略する。

変形型

変形型の規則をとらえる前に、まず PSG の一つの拡張として、痕跡 (trace) とよばれる処理を導入しよう。図-1 に具体的な例を示している。例文 30 にあるように関係代名詞などでは、名詞句 np が先行詞となって文 sent の外へ出たため、sent の内部では当該の np 部分が空白 [] となる。したがって、書換え規則を用いて通常の解析を進めると、失敗に終わる。そこで、“sent/np” で np が欠けている sent を表すことにしよう。痕跡処理は、このように規則に記述してあった場合に、欠けた部分をうまく扱うための拡張である。標準の DCG では扱えないが、6.2.3 で紹介するシステムであれば扱うことができる。

痕跡処理 “/” を利用すると、図-1 の関係代名詞は、次のように記述できる。

np → np rnp sent/np. (21)

では、早速、/ を用いて受け身文 (例文 25) の規則を作ろう。受け身変形は、次の操作をとまう。

English is understood by many people.

<== Many people understand English.

そこで、受け身の解析として以下の規則を作る。

sentp → np be vp/np by np. (22)

つまり、能動文の vp は目的語としての np を含むが、受け身文ではそれが主語として先行し、vp の中では空白となる。なお、受け身文としてのカテゴリを “sentp” と表現している。

[問 6.2] 疑問変形を考えよう。例文 5 や 22 のような will などの助動詞を用いる yes-no 型疑

問文は、次の操作をとまう。

Will you show me around the city?

<== You will show me around the city.

Do you have a pen?

<== You have a pen.

これらに対する規則を示せ。また、when などの疑問詞で始まる、例文 13 の場合はどうか？

解答例

will などの aux を含む sent では、語順を入れ換えるだけでよく、次の規則を得る。

sentq1 → aux sent/aux. (23)

一方、aux や be-動詞を含まない場合は、do を付加しなければならない。

sentq1 → do sent. (24)

次に疑問副詞などの場合は、いま得られた sentq1 に対して以下のような痕跡を考えるとよい。

When do you play tennis?

<== Do you play tennis []?

すると規則は、次のようになる。

sentq → iad sentq1/adp. (25)

合成型

合成型は、比較的簡単である。埋め込み型では、np に不定詞 (例文 28) や that-節 (例文 26) などを導入すればよい。

np → (infp; thatc). (26)

結合型の場合は、接続詞とコンマに注意して、たとえば次のように合成する。

csc → coj1 sent. (27)

sent → csc com sent. (28)

最後に、ピリオドなど文としての形式を整える。

sentence → sent prd. (29)

以上、基本的には PSG に基づいて文法データを作成した。特に、基本/応用及び標準/特殊という切り所全体を見通すことに心がけた。これらの規則がとらえている“内容”は、GPSG, LFG など文法理論が異なっても、共通すると期待できよう。しかし、表-1 程度の文であっても案外複雑な規則が必要となったが、取り扱うテキストが大きくなるにつれ、基本と応用、あるいは標準と特殊の境界が不明確になり、必要な規則が増大してくる。そこで、文法規則は特にもたずに、“用例集”の考えを押し進めて、巨大なコーパス (書類、文献、資料などの集成) をデータとしてもち

解析や翻訳に利用するという方法もある¹⁾。さらには、規則の適用の確からしさを確率でとらえる手法なども提案されている²⁾。

6.2 辞書データ

辞書には、意味データと語彙データが必要である。意味データとして特に大切なのは、語の概念を構成する意味素性と、概念と概念の間の結びつきである。また語彙データとしては、文法及び意味情報をどのような形式で単語辞書に盛り込むかが大切である。以下では、これらを解説しよう。

6.2.1 意味データ

語の概念は人が幼いころからしだいに概念を形成していく過程（すなわち学習過程）に注目すると、単純概念 (simple concept) と連結合成概念 (interconnected-synthesized concept) とに大別できる、と筆者らは考えている（参考文献4）参照）。単純概念は、現実世界の指示対象と直接結びつく具象的な概念で、たとえば“花”，“走る”，“長い”などである。それに対し連結合成概念は、いくつもの単純概念が結合したりあるいは単純概念から派生したりして得られる抽象的な概念で、たとえば“職業”，“売る”，“ずるい”などである。

意味素性

第2回の2.2.1では、名詞の意味素性について例を示した。ここでは、動詞と形容(動)詞についてみてみよう。表-6をご覧願いたい。単純概念の一覧表を示している。動詞が事象を表し、形容(動)詞が事象の属性を表すとして分類した。単純概念が現実世界の事象と直接結びつくという意味が、具体例を通じておおよそ理解できるであろう。この表は、次の作業によって得られた。

類似の意味内容の語を集めた辞典として、シソーラスとよばれるものがある。国立国語研究所で編纂されたシソーラス“分類語彙表”には、日常の言語生活で基本的な動詞と形容詞（形容動詞を含む）が、それぞれ、およそ4,300と2,300収録されている（ただし旧版³⁾）。これらの概念をそれぞれ対象世界とみなして、その中から類似した概念をいくつも集めては、なぜ類似しているのかを考察する。たとえば事象の場合，“走る”，“歩く”，“行く”，“進む”などは，“位置の変化”という点で共通している。最初は、小さなグループに対してこのような特徴抽出を繰り返す。次に、ある程度の数の特徴が得られると、バランス

表-6 単純事象／属性概念の意味素性

事 象 概 念		属 性 概 念	
意味素性	例	意味素性	例
心理的		心理的	
精神の変化	悲しむ	感情	うれしい
感覚の変化	匂う	感覚	寒い
物理的		物理的	
位置の変化	入る	場所	深い
向きの変化	裏返す	向き	斜めだ
形の変化	こわれる	形	丸い
質の変化	腐る	質	堅い
量の変化	増える	量	多い
光の変化	光る	光	暗い
色の変化	赤らむ	色	赤い
熱の変化	冷やす	熱	熱い
力・勢いの変化	締める	力・勢い	強い
音の変化	響く	音	やかましい
出現・消滅	現れる	出現・消滅	あらわだ
開始・終了	始まる	開始・終了	だしぬけだ
時間の変化	遅れる	時間	早い
状態		状態	
継続	続く	継続	たえだえだ
有様	そびえる	有様	うっとうしい
抽象	基づく	抽象	等しい
その他	食べる	その他	さようだ

を考えて、大きな特徴を分解したり、逆に小さなものを統合したり、あるいは偏ったとらえ方を修正したりする。このような試行錯誤を、対象世界の概念がすべていずれかのグループに属するようになるまで繰り返し、積み上げた結果が表-6である。

この手法の長所は、少なくとも対象とした範囲の概念をカバーしている、未知のものが出現したときは同様の手法で更新できる、などである。連結合成概念についても調査し、学術・芸術、宗教、言語、社会などの意味素性を見出している。

概念間の関係

概念の種々の結びつきの中でも、上位と下位の階層は、大切とされている。この問題を、次の例を用いて考察しよう。

C = {川, 川底, 上流, 瀬, 激流, 大川}

先頭の要素“川”に対して、残りの5つの要素は、次のような連体修飾詞で把握できる。

川底：底である，川の部分

上流：源に近い，川の部分

瀬：浅い，川の部分

激流：激しく流れる，川

大川：大きな，川

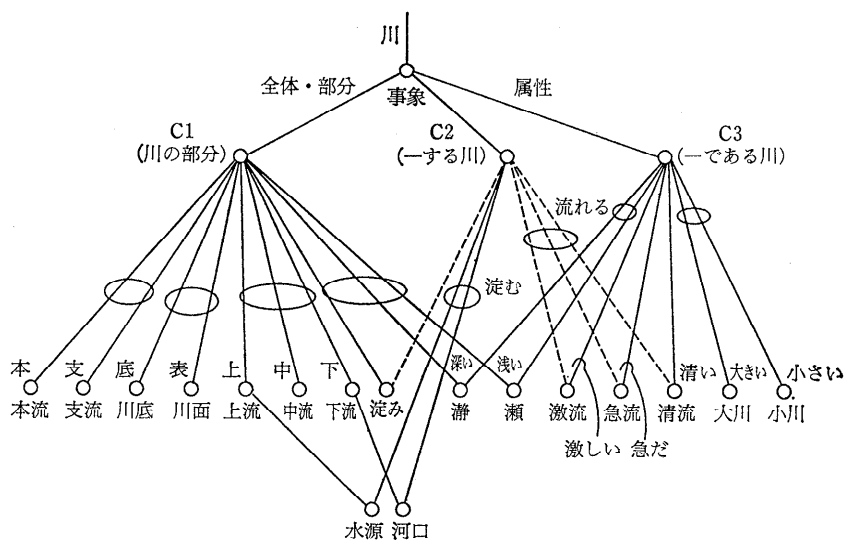


図-2 概念ネットワーク

すなわち5つの要素は、連体修飾子を通じて川の下位概念となっている。図-2は、この考え方を押し進めて作成した概念ネットワークである。

【問 6.3】 概念の間に階層をもたらず性質はいくつかの観点がある。大川は、川というクラスの一つのメンバである。この関係をISAと呼び、ISA(大川, 川)のように表現する。では、瀬と川の関係はどのように把握すればよいだろうか？

解答例

瀬は川の部分である。そこで、全体と部分の関係をPART-OFと呼び、PART-OF(瀬, 川)のように表すと分かりやすい。

6.2.2 語彙データ

実用的な言語処理システムでは、少なくとも数万語の語彙項目を必要とする。語彙項目の形式は文法理論に大きく左右される。ここでは、6.1の文法規則及び前項の意味素性を用いる、実験レベルの項目を例示しよう。

初めに、個々の語に対する項目を、次のような語彙規則(書換え規則)で記述しよう。記述形式は、第3回の4.2.2のDCGの説明を思い出してほしい。なお、形態素解析の詳細情報には立ち入らない。

nou1([sem/pen1], nou1(pen))

→ [pen].

(1)

[sem/pen1]は、penが多義であるときその中の第1番目の意味へのポインタ(別の項目への指示)

である。

prn1([sem/she1], [mph/sg3_1], prn1(she))

→ [she].

(2)

mphは形態素情報を意味し、sg3_1は単数、3人称、第1格を表す。

vrbl([sem/take2], vrbl(take_care_of))

→ [take, care, of].

(3)

動詞(特に熟語)の例である。形容詞も同様である。

adj([sem/small1], adj(small))

→ [small].

(4)

次に、上に述べた意味ポインタ semの指示する、具体的内容をいくつか示そう。

sem(pen1, [tool*writing]).

(5)

[]内に、意味素性が示されている。この例では、“道具(tool)の一種であり書くためのもの(writing)”ということになる。

sem(drive1,

[evt, [operation], [sfrm/pl_3],

[Agt, [agt, [human], sbj],

[Obj, [obj, [movable], obj1]]).

(6)

格構造が示されている。evtは事象そのものを記述しており、ここでは単にoperationということで、深い意味構造には立ち入っていない。なお、sfrm/pl_3はこの事象がobjを必要とする型であることを意味している。またAgt, Objでは、それぞれが文中主語(sbj), 目的語(obj1)で表現されることを示している。

SENTENCE Taro drives a sports car.

STRUCTURE TOTAL 1

No. 1 time : 82 msec

```

[-sentence
  [-sent
    [-s
      [-np
        [-nou2 -- taro
      ]
      [-vp
        [-vrb1
          [-drive
          ]
          [-suffix -- es
          ]
        ]
        [-np
          [-det1 -- a
          ]
          [-np
            [-ajp
              [-adj -- sports
              ]
              [-np
                [-nou1 -- car
                ]
              ]
            ]
          ]
        ]
      ]
    ]
  ]
  [-prd --
    [drive,
      [evt, [operation], [sfrm/pl_3]],
      [taro, [agt, [human], sbj]],
      [car,
        [obj, [movable*thing], obj]],
        [sports, [att, [sports], mdf], [car, [pos, [thing], obj]]],
        [a, det1]
      ]
    ]
  ]
]

```

図-3 解析結果の例(1)

sem(small1, [att, [shape], [pos, [thing]]).

(7)

att は, small1 という属性そのものを記述している。なお pos の部分は, その属性をもつ主体が thing であることを意味する。

6.2.3 実験

それでは, 本章の文法及び意味規則に基づく実験結果*をご覧にいれよう。図-3 は, 本講座でおなじみの例文を解析したものである。SENTENCE は入力文を表し, STRUCTURE は解析木を表現している。TOTAL は解析に曖昧性のあるばあい, その個数を示す。また No. はそのうち何番目の解析であるか, そして time はその解析に要する時間を表す。解析木の下に, “リスト形式”で意味解析の結果が示してある。car を sports が修飾 (mdf) しているため, その部分が少し詳しくなっている。

もう一つ, やや長い目の文の解析例を図-4 に示しておく。構文的に従属節 csc の構造がきちんと解析され, その意味も, リストの6行目において, die を核とする従属節は “disappearance” という “reason” で主節の take_care_of を修飾 “mdf” していることが解析されている。なお, 初回の講座の冒頭で示した実験例も, もう一度見直して欲しい。

以上, 紙面の都合で 2, 3 の例しか表示できな

SENTENCE She takes care of me, because my mother died a long time ago.

STRUCTURE TOTAL 2

No. 1 time : 160 msec

```

[-sentence
  [-sent
    [-s
      [-np
        [-prnl -- she
        ]
        [-vp
          [-vrb1
            [-take
            ]
            [-suffix -- es
            ]
            [-care
            ]
            [-of
              [-np
                [-prnl -- me
                ]
              ]
            ]
          ]
        ]
      ]
      [-com -- ,
      ]
      [-csc
        [-coj1 -- because
        ]
        [-sent
          [-s
            [-np
              [-det4 -- my
              ]
              [-np
                [-nou1 -- mother
                ]
              ]
              [-vp
                [-vrb1
                  [-die
                  ]
                  [-suffix -- ed
                  ]
                ]
                [-adp
                  [-a
                  ]
                  [-long
                  ]
                  [-time
                  ]
                  [-ago
                  ]
                ]
              ]
            ]
          ]
        ]
      ]
    ]
  ]
  [-prd --
    [take_care_of,
      [act, [social], [sfrm/pl_3]],
      [she, [agt, [human], sbj]],
      [i, [obj, [human], obj]],
      [die,
        [reason, [disappearance], mdf],
        [evt,
          [disappearance],
          [sfrm/pl_1],
          [form/csc]
        ],
        [mother, [agt, [human], sbj], [my, det4]],
        [a_long_time_ago, [att, [time], mdf]]
      ]
    ]
  ]
]

```

図-4 解析結果の例(2)

かったが, この実験により本章の言語データが期待どおりの振舞いをする事が確認できた。なお, 文法データ, 語彙データ及び実験システムの詳細については参考文献 1) を, また意味データの詳細については参考文献 4) を参照してほしい。

6.3 今後の課題

6.3.1 辞書データの蓄積と自動作成

最初にも述べたが, 自然言語処理システムを実用的に使用しようとする, 膨大な言語データを準備する必要がある。特に辞書データは, “文中に辞書にない語があると極端に処理がむずかしくなる” という点から処理システムの能力を直接左右する重要なものである。また, その内容も, 4. や 5. で示したような簡単なものではなく非常に複雑である。このため, 各所で語彙・意味データを蓄積し機械辞書を作成する努力が精力的に行われている。大規模な辞書の作成のプロジェクトとして日本電子化辞書研究所の辞書⁴⁾が, また詳細な辞書記述として情報処理振興事業協会の計算機用日本語基本動詞辞書 IPAL⁵⁾が知られている。

機械辞書に必要な情報の例として, Mu 機械翻訳プロジェクト⁶⁾で作成された動詞辞書を図-5 に

* 本実験には, 東京工業大学田中穂積教授らによって開発された自然言語処理システム LangLab を用いた。記して謝意を表する。

- 1 ((\$ 見出し番号 "V0035500-01")
- 2 (\$ 更新年月日 "850213")
- 3 (\$ 見出し情報 (\$ 見出し語 "はんだづけする")
- 4 (\$ 語尾字数 2) (\$ 読み "はんだづけする")
- 5 (\$ 異形語 "はんだづけする"))
- 6 (\$ 形態素情報 (\$ 形態品詞 動)
- 7 (\$ 動詞活用型 サ変)
- 8 (\$ 前接情報 2) (\$ 後接情報 13))
- 9 (\$ 構文-意味情報 (\$ 分野コード 電気)
- 10 (\$ 構文品詞 動詞)
- 11 (\$ 格パターン V1
- 12 (\$ アスペクト 瞬時)
- 13 (\$ 態 受身 可能 'である') (\$ 意志 有)
- 14 (\$ 格支配情報
- 15 ((\$ 表層格 が) (\$ 深層格 主体)
- 16 (\$ 名詞意味マーカ OH) (\$ 必須性 1))
- 17 ((\$ 表層格 を) (\$ 深層格 対象)
- 18 (\$ 多詞意味マーカ OM OA) (\$ 必須性 1))
- 19 ((\$ 表層格 に) (\$ 深層格 受け手)
- 20 (\$ 名詞意味マーカ OA) (\$ 必須性 1))))))

図-5 Mu システムの辞書における“はんだづけする”の記述

示そう。この例は、“はんだづけする”という動詞に関する記述である。1-2 行には、辞書内容を管理するための情報が記述してある。小人数で作成する実験的な辞書であれば、このような情報は不要であろうが、多人数で作成する場合には、このような情報も必要となる。

3-5 行には見出しに関する記述がある。漢字表記の語には読みを付加しておくとか何とか都合がよい。また、5 行目のように“異形語”の指示も必要である。特に日本語では、漢字を使うかどうかやカタカナ書きの場合の表記の変化形（たとえば、“コンピューター”と“コンピュータ”）があるので、辞書を完全なものにすることは容易でない。

6-8 行は形態素解析に必要な情報が記述してある。第 2 回の 3. の説明を思い出してほしい。ただし、前後の語に関する制限は具体的な語ではなく、2 や 13 のように、識別番号で示されている。最後の 9-20 に構文及び意味に関する内容が表現されている。特に 14-20 行は 2.2.2 で説明した格構造に相当するものである。この例であれば、

格助詞	格	選択制限
が	主体 (動作主)	OH (人)
を	対象	OM (部品及び材料), OA (生産物)
に	受け手	OA (生産物)

という格構造が表現されている。

表-7 IPAL 辞書における“掛ける”の意味分類 (部分)

意味記述 1	一面に何かを覆う。
文例 1	彼女は食卓にテーブルクロスを掛けた。
意味記述 2	ある道具の作用を何か・どこかに及ぼす。
文例 1	彼女は毎日廊下に雑布を掛けている。
意味記述 3	ひも状の物を何かに巻き付ける。
文例 1	彼女はプレゼントにリボンを掛けた。
意味記述 4	ひも状のような物を身体に付ける。
文例 1	彼女は首に真珠のネックレスを掛けている。
意味記述 5	機械を操作して音が出るようにする。
文例 1	あの喫茶店はモダンジャズを掛けている。
意味記述 6	身体の部分进行他の物・所に接触させる。
文例 1	彼は椅子に腰を掛けている。
...	...
意味記述 23	設置する。
文例 2	燕が軒下に巣を掛けた。
意味記述 24	物の表面を滑らかにしたり、その一部を削る目的である物を働かせる。
文例 1	大工が角材にかんなを掛けた。
意味記述 25	異なる品種の動植物を交ぜ合わせて、新種を作り出す。
文例 1	トマトにじゃがいもを掛けたものをポマトといいます。

これまでは説明を簡単にするため、語の一つの意味だけを考えていた。しかし、語にはいくつもの意味があるのが普通である。表 7 を見てほしい。これは、IPAL (Basic Verbs) での“掛ける”の意味分類の一部である。この辞書では、“掛ける”の意味を 25 種類に分類しており、それぞれに対して格構造に相当する情報が記述されている。その他、辞書の内容・構成の詳細については参考文献 8) に詳しい。

2, 3 の例に示しただけであるが、実用的な自然言語処理に利用できる辞書の作成は非常に手間のかかることが想像できるだろう。そこで、市販の用例集や辞書など、既存の生のデータから言語データを自動抽出しようという研究が行われている^{7), 8)}。ここでは、その具体例として、語と語の

表-8 LDOCE 中の名詞の定義文の例

ace	a <i>person</i> of the highest class or skill in something
comedian	an <i>actor</i> who tells jokes or does amusing things to make people laugh
telescope	a tubelike scientific <i>instrument</i> used for seeing distant objects by making them appear nearer and larger—see picture at OPTICS
alum	a chemical <i>substance</i> used in medicine and preparing leather
abbey	the <i>group</i> of <i>people</i> living in such a building
travel	the <i>act</i> of travelling

表-9 自動抽出した上位下位関係の連鎖の例 (person)

person	
accountant	a <i>person</i> whose job is to keep and examine the money accounts of businesses
CPA	certified public <i>accountant</i>
ace	a <i>person</i> of the highest class or skill in something
actor	a <i>person</i> who takes part in something that happens
comedian	an <i>actor</i> who tells jokes or does amusing things to make people laugh
comedienne	a female <i>comedian</i>
extra	an <i>actor</i> in a cinema film who has a very small part in a crowd scene
ham	an <i>actor</i> whose acting is unnatural, esp. with improbable movements and expr
mime	an <i>actor</i> who performs without using words

意味関係を辞書から抽出する手法について紹介しよう^{9),10)}.

英々辞書での名詞の説明を表-8 に示す。これは、LDOCE という辞書¹¹⁾ の例である*。これらの説明文は、“語を定義しようとしている”ということから定義文と呼ばれる。

定義文には大きく分けて二つのパターンがある。

1. 表の上部の例のように、より一般的な意味をもつ語に付加的な説明を加えて限定するというパターン。“ace” は “person” の一種であるが “highest class” である。“comedian” も “person” の一種であるが、“jokes” などと言う。

2. 表の下部のように、関連のある語を用いて説明するというパターン。“abbey” は、“people” のグループ (group)。“travel” は、旅行する (traveling) 行為 (act)。

1 は、上位下位関係で、問 6.3 で説明した ISA である。LDOCE の定義文を調査したところ、このような上位下位関係は、簡単なパターンで比較的精度良く抽出できることが分かった。たとえば、“person” に関係した上位下位関係を辿れば、表-9 に示す語を自動的に抽出することができた。“comedienne” は女性 (female) の “comedian” で、“comedian” は “actor” の一種で、“actor” は “person” の一種である。

また、2 に関して、LDOCE の定義文では、表-10 に示す語が用いられていることも分かった。これらを手がかりとすることにより、語と語の意味関係を自動抽出することができる。さらに、より詳細に定義文を分析すれば、図-6 に示すような、意味ネットワーク (第 1 回の 1.2.2 参

表-10 LDOCE 中の機能語と意味関係

意味関係	機 能 語
(a) 上位下位	one, any, either, pair, type(s), kind(s), sort(s), piece(s), bunch, number, lot
(b) 部分全体	part(s), side, top, base, block
(c) メンバ	set, member, group, class, family
(d) 行為	act, way, action
(e) 状態	state, condition
(f) 量	amount, sum, measure
(g) 質	quality, degree
(h) 形状	form, shape
(i) その他	mixture, branch

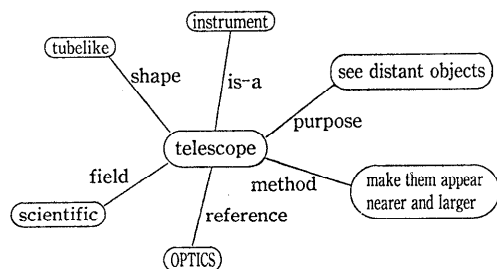


図-6 LDOCE の定義文に含まれる複雑な意味関係の例

照) も作成できる可能性がある。この例は、表-8 の “telescope” の定義文をうまく解析できれば求めることができるであろう。

このような語と語の意味関係は、精度の高い自然言語処理に必要となるので、日本電子化辞書研究所でも主として手作業で作成を進めている (参考文献 6))。既存の辞書やハンドブックなどから自動的に抽出できれば、自然言語処理用の辞書の作成のため、強力な手助けとなるであろう。

6.3.2 自然言語処理の課題

70 年代後半から 80 年代にかけて我が国の自然言語処理ではワードプロセッサ (“ワープロ”), そして機械翻訳システムという、大きな目標があった。研究者や技術者はこれらを課題として研

* ここで示す例は、1978 年発行の初版の例である。現在販売されている 1987 年発行の新版とは多少異なる。

究, 開発に精力的に取り組んだ。それでは 90 年代の大目標とは一体何であろうか? それを考えるために, 筆者らが現状の課題と考えているものを示そう。一つは, 技術的な不十分さであり, もう一つは, 応用対象の不明確さである。

ワープロで長いひらがな列を漢字に変換させると, 一応変換するが, まだまだ完全ではない。これは, 仮名漢字変換では複雑な文法や意味を使用していないためである。翻訳システムは, かなり複雑な文法や意味を用いているが, それでも十分なものではない。より精度の高い解析を行う手法をまだまだ研究する必要がある。これには, 本講座で説明した構文・意味解析はもちろんのこと, より深い意味や文脈, 知識を用いた処理方法を考えなければならない。例文や確率を用いる, 超並列計算手法を用いるといった新しいアプローチの研究も始まっている。また, 書かれた文章だけでなく, 人間同士やコンピュータと人間の対話を扱う手法に関する研究もさらに力を入れなければならない。

一方, どのように応用するかも問題である。機械翻訳やそれを音声認識・合成と結合した自動通訳は, すばらしい応用である。もし, 人間に近い能力をもったシステムが開発できれば, 問合せへの応答や相談など, 応用の夢は広がる。しかし, 現在の技術で作成できる自然言語処理システムの能力とはかなり隔たりがある。したがって, 現状及び近い将来実現できる適切な応用を見つける努力が求められている。

自然言語処理は理論的にも技術的にもかなり力をつけてきた。また, 研究・開発に必要な各種ツールやデータの整備や新しいアプローチによる挑戦も進んでいる。今は, かつてのワープロや翻訳システムに匹敵する大目標を見いだす個別的努力と模索とがなされている状況といえよう。このようにときにこそ, 特に若手の研究者や技術者のすばらしい発想力や開発意欲が強く望まれる。

以上, 初心者の方を対象として 5 回にわたる自然言語処理講座を展開した。初回到述べたように, 読者が自然言語処理の世界を一通り理解でき, また自ら処理システムを作成する力を身につけることができたなら, 筆者らの大きな喜びである。最後に, 読者の中から将来この分野に取り組んでいただける方の現れるであろうことを願っ

て, 講座を閉じたい。

文 献

- 1) Sato, S. and Nagao, M.: Toward Memory-Based Translation, Proceedings of COLING 90, Helsinki (1990).
- 2) Black, E., Lafferty, J. and Roukos, S.: Development and Evaluation of a Broad-Coverage Probabilistic Grammar of English-Language Computer Manuals, Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (1992).
- 3) 国立国語研究所 (編): 分類語彙表, p. 362, 秀英出版 (1972).
- 4) 日本電子化辞書研究所: EDR 電子化辞書仕様説明書, (株)日本電子化辞書研究所 (1983).
- 5) 情報処理振興事業協会: 計算機用日本語基本動詞辞書 IPAL (Basic Verbs), 情報処理振興事業協会 (1992).
- 6) Nagao, M., Tsujii J. and Nakamura, J.: The Japanese Government Project for Machine Translation, Computational Linguistics, Vol. 11, No. 2-3, pp. 91-110 (1985).
- 7) 鶴丸弘昭: 単語間の上位下位関係の自動抽出, 情報処理学会, SG-FIS 3-1 (1986).
- 8) Wilks, Y. et al.: Providing Machine Tractable Dictionary Tools, Machine Translation, Vol. 5 (1990).
- 9) Nakamura, J. and Nagao, M.: Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation, Proc. of COLING 88, pp. 459-464 (1988).
- 10) Nakamura, J. and Okada, N.: Towards Lexical Knowledge Base Construction by Co-Operation of Human and Machine, Information Modelling and Knowledge Bases IV, (H. Kangassalo, H. Jaakkola, K. Hori, T. Kitahashi ed.) pp. 159-170, IOS Press (1993).
- 11) Procter, R.: Longman Dictionary of Contemporary English, Longman Group Limited, Harlow and London, England (1978).

参 考 文 献

- 1) 岡田直之: 自然言語処理入門, p. 155, 共立出版 (1991).
6. で述べた文法規則及び実験システムについて, プログラムやデータのリストも含めて詳しく説明されている。
- 2) Winograd, T.: Language as a Cognitive Process, Vol. 1, Syntax, Addison Wesley (1983).
付録に基本的な文法データが記載されている。
- 3) 益岡隆志, 田窪行則: 基礎日本語文法, くろしお出版 (1989).
日本語の文法現象が分かりやすく整理されている。

- 4) 岡田直之：語の概念の表現と蓄積，p. 160，電子情報通信学会（1991）。

動詞，形容詞（形容動詞を含む）及び名詞についての詳細な意味分析が示されている。

- 5) 大野 晋，浜西正人：角川類語新辞典，角川書店（1981）。

シソーラスの一つで，語彙の世界を図書館分類学的に細かく整理している。

- 6) 電子化辞書研究所（編）：概念辞書，TR-020（1990）。

研究所で開発を進めている大規模概念辞書の概要が述べられている。

- 7) 長尾 真（監）：日本語情報処理，電子情報通信学会（昭60）。

少し古いが，日本語情報処理全般が概説されており，実用レベルでの機械辞書の作成，探索，実例などについても述べられている。

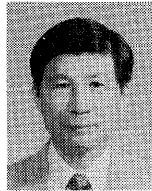
- 8) 野村浩郷編：言語処理と機械翻訳，講談社サイエンスフィク（1991）。

機械翻訳からみた機械化辞書に関して，内容，問題点など詳しい解説が第3章に述べられている。

参考ソフト

今回の実験システムで用いた LangLab は，日本デジタルイクップメント社のユーザグループ DECUS のパブリックドメインで利用に供せられている。

（平成6年1月13日受付）



岡田 直之（正会員）

1964年東海大学工学部卒業。1966年九州大学大学院工学研究科修士課程修了。同年同工学部助手，1976年大分大学工学部助教授，1978年同教授を経て，現在九州工業大学情報工学部教授。工学博士。人工知能の研究に従事。電子情報通信学会，人工知能学会各会員。



中村 順一（正会員）

1979年京都大学工学部卒業。1982年同大学院工学研究科博士後期課程中退。同年京都大学工学部助手，1989年九州工業大学情報工学部助教授。工学博士。自然言語処理，音楽情報処理の研究，計算機ネットワークの管理に従事。電子情報通信学会，ソフトウェア科学会，日本認知科学会，Association for Computational Linguistics 各会員。

