語の対訳知識を用いた対訳テキストの文対応付けアルゴリズム

森本 康嗣　梶 博行
（株）日立製作所 中央研究所

事例ベース翻訳で必要となる事例ベース構築のため，日英対訳テキストを文単位で対応付けるアルゴリズムの検討を行った。文中の対訳関係を持つ単語数と全単語数の比により日本文と英文の相関度を定義し，ＤＰマッチングによりテキスト全体における文単位の対応関係を決定する。さらに，相関度において，単語の出現頻度情報を用いて単語に重み付けを行った。また，局所的なヒューリスティックスを用いて対訳テキストをブロックに分割することにより，高速化を図った。実験の結果，正解率が約９５％という結果が得られた。また，ブロック分割により相関度計算回数が約３５％に減少，頻度情報の利用により正解率が約２％向上することを確認した。

１．はじめに

ここ数年，実際のテキストから言語データを抽出するアプローチが盛んになっており，対訳コーパスの重要性が高まっている。通常，対訳テキストには，原文と訳文間の対応関係は付与されていない。そこで，対訳テキストを文単位に対応付ける研究が行われている。Brown[1]は，文長に基づいて英語，仏語からなる対訳テキストを対応付けている。また，Chen[2]は，統計的に得られる語の対応関係を用いて対訳テキストを対応付ける手法を提案している。Utsuro[3]は，対訳辞書と統計情報を組み合わせて対応付けを行う手法を提案している。

本報では，対訳辞書を用いた対訳テキストの文対応付けアルゴリズムについて述べる。本報で述べるアルゴリズムでは，単語の出現頻度情報を利用して対応付け精度の向上を図っていること，局所的なヒューリスティックスを用いることにより処理の高速化を図っていることに特徴がある。

２．対訳テキスト対応付けアルゴリズム

２．１ 対訳テキスト対応付けの定式化

本報告では対訳テキスト対応付けを，次のように定式化する。

（１）日本文の列$J=\langle J_1, J_2, J_3, \ldots, J_m\rangle$，英文の列$E=\langle E_1, E_2, E_3, \ldots, E_n\rangle$からなる対訳テキストを仮定する。このとき，日本文$J_i$と英文$E_j$が対訳関係にある場合に得点が高くなるような相関度$R(J_i, E_j)$を定義する。

（２）配列$A(i,j)=R(J_i, E_j)$において，$A(1,1)$から$A(m,n)$に至るパスを考える。その中で，相関度に基づいた得点が最高のパスを正しい対応付けとして選ぶ（図１）。

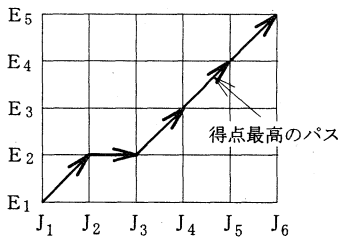


図１．対訳テキスト対応付けの定式化

パスＰは配列Ａの要素の列で表現する。図１の例の場合，次のようになる。

$P=\langle A(1,1), A(2,2), A(3,2), A(4,3), A(5,4), A(6,5)\rangle$

２．２ 対訳文の相関度

本報告では，対訳辞書を用いて求められる日本文と英文中の単語の対訳関係により，日本文と英文の相関度Ｒを定義する。

$$R=\frac{R_j+R_e}{2} \quad \text{ただし，} \quad R_j=\frac{m_j}{C_j}, \quad R_e=\frac{m_e}{C_e} \qquad (1)$$

Aligning Sentences in Parallel Texts using Bilingual Dictionary
Yasutsugu MORIMOTO, Hiroyuki KAJI
Hitachi, Ltd. Central Research Laboratory

ここで，R_j，R_eをそれぞれ，日本文相関度，英文相関度と呼ぶ。m_jおよびm_eは，それぞれ，日本文中あるいは英文中で，他方の文中の少なくとも一つの単語と対訳関係を持つ単語の数である。また，C_j，C_e は，それぞれ日本文，英文中の全単語数である。ただし，機能語は処理の対象から除外している。

2．3 頻度情報による相関度の改良

出現頻度の高い単語は，対訳文として誤った対に対しても相関度を大きくする要因となる。そこで，単語の出現頻度を求め，相関度に反映することにした。改良された相関度Rを次のように定義する。

$$R = \frac{R_j + R_e}{2} \quad \text{ただし}$$

$$R_j = \frac{\sum_{w_j \in M_j} \frac{1}{f(w_j)}}{C_j}, \quad R_e = \frac{\sum_{w_e \in M_e} \frac{1}{f(w_e)}}{C_e} \qquad (2)$$

ここで，M_j，M_eは，それぞれ日本文，英文内で対訳関係を持つ単語の集合，C_j，C_eは，それぞれ日本文，英文の全語数，w_j，w_eはそれぞれ日本文単語，英文単語である。$f(w)$は，単語wが出現した文の数を表す。単語が出現した文の数によって単語の頻度情報を表すのは，同じ文に複数回出現する単語を考慮したためである。また，頻度情報を取得する範囲を，相関度計算対象である日本文と英文の対(J_i, E_j)を含む幅nの範囲に限定する。具体的には，次の範囲において頻度情報を取得する。

日本文の範囲：$J_{i-(n-1)/2} \sim J_{i+(n-1)/2}$
英文の範囲　：$E_{j-(n-1)/2} \sim E_{j+(n-1)/2}$

単語の頻度情報を求める範囲を限定するのは，次の理由による。単語の出現位置の分布には偏りがあるため，ある単語が多く現れる箇所とまれにしか現れない箇所がある。テキスト全体において多く現れる単語であっても，ある箇所の近辺で少ししか現れなければ対応付けの手掛かりとなりうる。

2．4 パスの得点の定義

単語の対訳関係を用いて定義された相関度に基づき，対訳テキストの対応付けを行う。パスの得点を次のように定義し，DPマッチングによって得点最高のパスを求める。

・パスの得点の定義

パス上の対訳関係の相関度の平均をパスの得点score(i,j)とする。すなわち，A(1, 1)からA(i, j)までのパスの得点は次のようになる。

$$score(i,j) = \frac{\sum_{(Jk,El) \in P} A(k,l)}{|P|} \qquad (3)$$

ここで，Pは，パス上の文の対の集合である。

2．5 ブロック分割による高速化

2．5．1 ブロック分割処理

語の対訳関係に基づく対応付けは，形態素解析，対訳辞書に基づいた語の対応付けなどの処理を伴うため，相関度の計算に時間がかかる。そこで，ブロック分割による高速化手法を検討した。

実際の対訳テキストを分析すると，対応付けが容易である箇所とそうではない箇所が混在している。このとき，対応付けが容易である箇所を先に対応付けて，テキスト全体をより小さいブロックに分割することができれば，その後の対応付けの計算量を大幅に減少させることができる。

ここで，対応付けが容易にできる箇所として，1対1の対応関係が連続している箇所が挙げられる。そこで，1対1の対応関係の候補を見つけ，これが連続している箇所があれば，対応付けるという局所的なヒューリスティックスを用いる。

・日本文J_iにおけるテキストの分割可能条件

（1）ある日本文J_iに対し，日本文相関度R_jおよび英文相関度R_eが共に最大となるE_jを探し，これが見つかればE_jを対応する英文の候補とする。ここで，R_j，R_eは，2．2の式(1)あるいは2．3の式(2)を用いる。

（2）J_iの対応英文候補がE_jであり，J_{i+1}の対応英文候補がE_{j+1}であるとき，日本文J_iにおいてテキストが分割可能であるという。このとき，(J_i, E_j)を分割点と呼び，(J_i, E_j)の直後で対訳テキストを分割できる。

ここで，日本文相関度R_jと英文相関度R_eを別々に用いるのは，１対１に対応する文を判別するためである。

この分割可能条件を満たす文をテキストの中心から探索し，見つかった分割点でテキストを２個のブロックに分割する。そして各ブロックに対して，同じ手続きを再帰的に適用することによってブロック分割を行う。なお，ブロック分割処理は次のときに停止する。

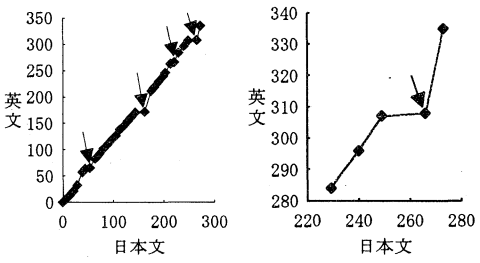
・ブロック内に分割点がない。
・ブロックの大きさが閾値以下になった。

２．５．２ 誤分割点削除処理

ブロック分割処理で用いるヒューリスティックスは，局所的な制約条件に基づいているため，誤分割を発生させる場合がある。そのため，誤分割点を削除する方法を検討した。誤分割点削除処理は，次のような仮定に基づいている。

・正しい対訳関係を示すA上のパスは，テキスト全体を見れば，A(1,1)とA(m,n)を結ぶ直線に沿っている。

実際の対訳テキストをブロック分割した結果を図２に示す。図中，プロットが分割点であり，矢印が誤分割点を示す。ある誤分割点の近辺を拡大したものを，図２(b)に示す。



(a)　(b)

図２．ブロック分割の結果

図２を見ると，正しい分割点が，全体としては直線上に沿っていることと，誤分割点の前後でグラフの傾きが大きく変化している。以上より，誤分割点の削除条件を，次のように定めた。

・分割点候補の削除条件

各ブロック内の日本文文数と英文文数の比をｒとする。ある分割点の前後のブロックのｒを，それぞれ，r_1，r_2とし，ｒの平均値をr_{av}，ｒの標準偏差をσとするとき，次の条件を満たす分割点を削除する。

・$r_1 < r_{av} - \sigma$ あるいは $r_1 > r_{av} + \sigma$ かつ $r_2 < r_{av} - \sigma$ あるいは $r_2 > r_{av} + \sigma$
・$(r_1 - r_{av}) \times (r_2 - r_{av}) < 0$

３．評価

３．１ ブロック分割の効果

３．１．１ ブロック分割精度の評価

ブロック分割処理によって検出された分割点および誤分割削除処理後の分割点の精度を評価した。対象テキストを表１に示す。

表１．対象文書

文書番号	種類	日本文文数	英文文数
文書１	特許明細書	92	102
文書２	特許明細書	182	227
文書３	特許明細書	274	336
文書４	マニュアル	670	693
文書５	マニュアル	934	865

頻度情報の有無によるブロック分割精度を評価したところ，頻度情報を取得する範囲ｎが広い程，精度が良くなる傾向が見られた。しかし，範囲ｎが広い程，相関度計算コストが増大するため，ブロック分割では頻度情報を利用しない相関度を採用した。

表２に，誤分割点削除処理前後の分割点の精度および過剰に削除された正しい分割点数を示す。

表２．ブロック分割の精度

	誤分割削除前	誤分割削除後	過剰削除数
文書１	0.0 %	0.0 %	0
文書２	0.0 %	0.0 %	1
文書３	2.9 %	0.0 %	2
文書４	10.0 %	1.4 %	1
文書５	2.6 %	0.0 %	1
平均	3.1 %	0.3 %	1

ブロック分割の誤りは，対応付け精度を大幅に悪化させるため，特に少ないことが望ましい。結果を見ると，誤った分割点を全て削除することはできなかった。しかし，誤り率は全体の対応付けの誤り率と比較して十分小さいので，十分な分割精度が得られていると考える。

過剰に削除される分割点は，精度には大きな影響を与えない。ブロックが大きくなるため，処理速度を低下させる可能性があるが，実験の程度であれば特に問題ないと考える。

3．1．2 高速化に関する評価

ブロック分割しない場合とブロック分割した場合の相関度の計算回数の比較を行った。その結果を表3に示す。なお，ブロック分割には，頻度情報なしの相関度を用いた。ブロック内の相関度の計算回数は，頻度情報の有無に無関係である。また，ブロック分割ありの場合の計算回数は，ブロック分割の計算回数と各ブロック内の計算回数の和である。

ブロック分割を行うことにより，ブロック分割を行わない場合と比較して，平均約３５％に相関度計算回数が減少した。

表3．ブロック分割による高速化の効果

	相関度	計算回数	計算回数の
	分割あり	分割なし	比
文書1	1749	914	52.3 %
文書2	14756	6480	43.9 %
文書3	25814	6971	27.0 %
文書4	87857	23805	27.1 %
文書5	164374	39643	24.1 %
平均	−	−	34.9 %

3．2 対応付け精度の評価

ブロック分割後のブロック内の対応付けにおいて，頻度情報を利用しない場合と利用した場合の対応付け精度を比較した。結果を表4に示す。

頻度情報を取得する範囲ｎについては，ｎが２０を越えると精度が悪化していく傾向がみられ，ｎが２０以内のときには，特に傾向が見られなかった。そこで，処理速度を考慮し，ｎを３とした。また，ブロック分割では，両方とも頻度情報なしの相関度を用いた。単語の頻度情報を利用することにより，平均約２％正解率が向上した。

表4．対応付けの正解率

	頻度情報なし	頻度情報あり
文書1	97.8 %	98.9 %
文書2	94.5 %	96.2 %
文書3	92.7 %	95.9 %
文書4	87.6 %	90.3 %
文書5	92.2 %	96.2 %
平均	93.0 %	95.5 %

4．おわりに

対訳テキストの文対応付けアルゴリズムの検討，評価を行った。本アルゴリズムは，対訳辞書を用いて，テキスト単位で対応している日英の対訳テキストを文単位で対応付ける。特許出願明細書，ソフトウェア製品のマニュアルの対訳テキストを対象に正解率が平均約９５％という結果を得た。本アルゴリズムの特徴は，以下の点である。

（１）相関度を定義する際，単語の頻度情報に基づいて対訳関係を持つ単語に重み付けを行った。頻度情報を用いない場合と比較して，平均で約２％対応付け精度が向上することを確認した。

（２）ヒューリスティックスを用いて入力テキストを小さなブロックに分割することにより，高速化を図った。さらに，誤ったブロック分割点を削除する方法を提案した。ブロック分割を行わない場合と比較して，相関度計算回数が約３５％に減少し，分割精度は９９．５％であった。

参考文献

[1] Brown, P. F., Lai, J. C., Merce, R. L.: "Aligning Sentences in Parallel Corpora", Proceeding of 29th Annual Meeting of the ACL, pp. 169-176 (1991).

[2] Chen, S. F.: "Aligning Sentences in Parallel Corpora", Proceedings of 31st Annual Meeting of the ACL, pp. 9-16 (1993).

[3] Utsuro, T., et al.: "Bilingual Text Matching using Bilingual Dictionary and Statistics", Proceedings of the 15th International Conference on Computational Linguistics, pp. 1076-1082 (1994).

Ellipsis Generation in Communicative Dialogues

Kristiina Jokinen
Graduate School of Information Science, Nara Institute of Science and Technology*

1 Introduction

A common approach to the generation of elliptical utterances is to construct a semantic representation for a full sentence and then remove those concepts that are already known to the user or otherwise clear in the context (cf. [8, 3]).

We raise two objections to this approach. First, from the point of view of communication in general (Communicative Activity Theory, [1]), communicators share the assumption that everything the partner says develops the joint purpose of the dialogue, and it is not necessary to explicitly refer to all information that is to be communicated.[1] Consequently, if an elliptical contribution appropriately conveys the new information, is a *conversationally* full contribution, regardless of its *syntactic* incompleteness.

Second, it is a fallacy that conceptual specifications mostly correspond to propositions to be realised as clause-like chunks. As pointed out by [4], the underlying content of a text cannot be expressed as a set of composable facts, since the facts stand in relations and dependencies: whether a fact is explicitly expressed or not, depends not only on whether the hearer knows the fact, but also on a complex reasoning process with respect to the context.

In this paper we present a new way to plan conversationally appropriate contributions and generate elliptical utterances in natural language dialogue systems. We regard contributions as *referring* expressions: their generation must ensure that the correct conceptual situation is identified. Only new information is necessary in the contribution, and related information is added to make the reference accurate, valid, consistent and free from false implicatures. The planning process is governed by pragmatic rules which determine successfulness of an utterance in a given dialogue context. Success is measured by a preference function which partially orders the possible contributions with respect to how successfully they convey the communicative goal. This approach also tackles the problematic area between content planning and surface generation: it builds a bridge over the "generation gap" by using communicative knowledge to interleave reasoning about information content with reasoning about linguistic expression in planning conversationally adequate contributions.

The paper is organised as follows. We first discuss the distinction between explicit and implicit information and their relation to ellipsis. We then introduce the response planner algorithm, and finally we work through an example that shows how the framework has been applied in an implemented dialogue manager.

2 Explicitness and Implicitness

According to *Communicative Activity Theory* ([1]), speakers behave as rational motivated agents and trust the partner to behave in a similar way. In [5], this is formalised into two principles on which the speakers' communicative competence is based on:

(1) **The Responsiveness Principle:** Report the new information that results from the evaluation of the partner's contribution.

(2) **The Minimalism Principle:** Add contextual information only as needed to convey the whole goal, to avoid false implicatures, and to obey syntactic constraints.

The Responsiveness Principle accounts for the fact that communication takes place at all. The Minimalism Principle (a variation of Grice's Maxim of Quantity) accounts for elliptical contributions.

We make separate distinctions between explicit and implicit information on the one hand and between elliptical and complete sentences on the other hand. Explicitness and implicitness deal with concepts to be communicated to the partner, while ellipsis deals with grammatical realisation.

A relevant concept is a concept which is a part of the conceptual representation of the contribution.

An explicit concept is a relevant concept which is lexically realised on the surface level. NewInfo must always be explicit.[2]

An implicit concept is a relevant concept which is not lexically realised on the surface level, but can be inferred from the context.

*This work was started at the Centre for Computational Linguistics, UMIST, Manchester, UK, and continued at Computational Linguistics Laboratory, Nara Institute of Science and Technology, where the author is a JSPS visiting researcher. I would like to thank Yuji Matsumoto, Graham Wilcock and John Phillips for helpful and stimulating discussions on the topic.

[1] If the speaker repeats facts already known in the immediate dialogue situation, an implication can be drawn that there is an important reason why the facts are repeated.

[2] If NewInfo is unrealisable, replanning must take place since the system is unable to express the result of the evaluation.

Ellipsis denotes syntactic incompleteness. A contribution is elliptical if some of the syntactically obligatory arguments of the main verb are not lexically realised (*Rent.*), or if it does not contain a main verb (*In Bolton. Where? 12*).

Concepts are represented as world model concepts, and a contribution can realise a concept if there is a mapping from the concept to a lexical predicate. Elliptical realisation is subject to the linguistic constraints of the particular language. A concept which could be implicit in the conceptual representation may explicitly appear in the surface contribution, if it is required by language specific syntactic constraints. Conversely, implicit responses need not be elliptical. For instance, *S2* in (3) carries implicit information that the car hire companies are located in Bolton, and that the user's wish to rent a car is linked to the system's ability to give information about car hire companies. The user can infer the link from the initial setting of the dialogue, but if this were in doubt, the link would be made explicit to prevent the user from making false implicatures.

(3) Welcome to the Electronic Yellow Pages Information Service System. Please enter your request.
U1: I want to rent a car.
S1: Where?
U2: In Entwistle
S2: Where is Entwistle?
U3: In Bolton.
S2: Ok. Here is a list of car hire companies: < list >

3 The Response Planner

The Dialogue Manager (DM) and Natural Language Engine (NLE) share a Conceptual Lexicon (CL) which maps between NLE semantic predicates and World Model concepts. Thus DM reasons on language-independent conceptual representations, while NLE operates only on linguistic information and in generation. DM gives a fully specified semantic representation to NLE for surface generation.[3] The key resource of the Dialogue Manager is the Context Model, a dynamic knowledge base containing information about contributions, discourse referents, Central Concept, NewInfo, goals, and expressive, evocative and evoked attitudes. The DM also accesses the application backend (in our case: Yellow pages database), world model knowledge base, and the communicative principles. A more detailed description of the content of the different components is given in [5]. In the World Model every instantiated conceptual object has a unique index and the conceptual objects are organised into a subsumption hierarchy. Explicit concepts are also used to refer to a disjunction D of the concepts A and B (`disj(D,A,B)`), a set S of referents of the concept type C (`setOf(C,S)`), and a cardinality C of the set X (`cardinality(X,C)`). Designated concepts are mapped to application model headings.

We assume that the user goal has been recognised, expressive and evocative attitudes inferred, and that the system has formulated its own goal with a specified NewInfo. More detailed description of goal formulation can be found in [5].

The response planner produces the minimal representation for a system intention which successfully refers to the concepts to be communicated, includes NewInfo and conveys no false implicatures. The algorithm is based on four Relevance Criteria (cf. [7]): the contribution must be Accurate (represent the speaker's goal truthfully), Valid (indicate that the partner's evocative attitudes have been addressed), Consistent (the concepts must form a connected graph in the World Model and be linguistically realisable), and Free From False Implicatures (FFI, the contribution must not trigger unwanted implicatures). The Relevance Criteria are preference functions which define a partial order among the possible contributions, and the preferred contribution is the one which is among the maximal elements of each preference function.

The algorithm resembles Reiter's algorithm [7] to generate successful referring expressions for object-type entities. Reiter's algorithm is based on a user's domain knowledge, and determines a minimal set of attributes which are to be included in the object description, so that the description distinguishes the intended object from other objects in the context, is minamal and and free from false implicatures.[4] [6] formalises conversational implicature as a preference function which orders object descriptions according to their ability to successfully refer to the intended object. The function is decomposed into separate preference rules that cover each type of implicature, and the description which is the maximal element under the preference function is considered free from false implicatures. It is assumed that the preference rules do not conflict.

Our algorithm differs from Reiter's in two respects: we allow preference functions to conflict, and we exploit the fact that a partial order may have several maximal elements. We do not use Minimality as a preference

[3] The chosen task division between the two components is radical and more research is needed to draw appropriate and practical border-line between the reasoning and generation processes. This paper assumes that DM is the main system component, and it has indirect access to linguistic information via CL. It has control over such border-line tasks as the check if a concept is linguistically realisable, if some linguistic constraint requires the planned set of concepts to be augmented with more concepts, if a particular lexical element carries some extra connotations, and the choice between ambiguous lexical entries.

[4] Minimality refers to the minimal number of conceptual componets of a contribution and not to the length of actual surface expression, thus the task is polynomial, as pointed out by [6].

criterion as such, but have incorporated it into the basic setting of the task: any of the contributions rendered maximal by a preference function can be used as a successful referring expression, and the smallest one is chosen only if it is maximal according to other preference functions as well. The maximal element satisfies all the Relevance Criteria simultaneously and can be found in the intersection of the maximal elements of the four relevance criteria. If such an element cannot be found, the goal cannot be successfully expressed in the dialogue context.

The algorithm is given below, with the following abbreviations: *Agenda* is the set of chosen concepts, GC is the set of goal concepts, NI is NewInfo, DR is the set of (known) discourse referents, EEC is the set of concepts that form the content the partner's explicit evocative attitudes, IEC is the set of concepts that form the content the partner's implicit evocative attitudes, LRC is the set of linguistically required concepts, and LUC is the set of linguistically unrealisable concepts (that need to be replaced by sub- or superconcepts).

(1) Initialise Agenda with NewInfo.

(2) Check Accuracy: AccAgenda contains those GCs which are not known in the context:
$AccAgenda = Agenda \cup GC \setminus DR.$

(3) Check Validity:
(a) If $sysGoal = want(s, know(u, P))$, check the difference between explicitly evoked user expectations and the planned response:
- Collect the concepts of the previous explicit evocative user attitudes (EEC).
- $Val = EEC \backslash GC$
- If $Val = 0$, then check the user's evocative intentions:
If $userGoal = want(u, know(u, P))$, then AccAgenda matches the evocative intentions and $ValAgenda = AccAgenda$.
If $userGoal = know(s, P)$, then evocative intentions and the evoked response have no common concepts since NI is based on the re-evaluation of the previous user goals, and the GCs which are not evoked by the evocative intentions must be added: $ValAgenda = AccAgenda \cup (GC \backslash EEC)$
- If $Val \neq 0$, then $ValAgenda = AccAgenda \cup Val$.
(b) If $sysGoal = want(s, know(s, P))$, check the difference between implicitly evoked user expectations and the planned response:
- Collect the concepts of the previous implicit evocative user attitudes (IEC).
- $Val = IEC \backslash GC$
- If $Val = 0$, then no unreachable implicit concepts: $ValAgenda = AccAgenda$
- If $Val \neq 0$, then unknown concepts are added (if the user responded with an elliptical utterance, the implicit concepts are known)
$ValAgenda = AccAgenda \cup (Val \backslash DR)$ (4) Check Consistency:
(a) Connectedness: there exists a path between each concept in the ConcAgenda
If a path exists, then $Connected = 0$
If not, then $Connected$ = intermediate concepts that make the graph connected
(b) Linguistic constraints:
- Map concepts in $ValAgenda$ to semantic predicates via Conceptual Lexicon. If such a mapping exists, $Conc = 0$. If no mapping, then
if $type(Concept) = object$, map from super- or subconcept, and if this mapping exists, $Conc = ValAgenda \backslash \{Concept\} \cup \{super/subConcept\}$, else fail.
if $type(Concept) = event$ map from partOfPlanConcept. If this mapping exists, $Conc = ValAgenda \backslash \{Concept\} \cup \{partOfPlanConcept\}$, else fail.
- Grammaticality constraints:
LRC = obligatory arguments of the main verb
LRC = full complements of explicit arguments
$ConcAgenda = ValAgenda \cup Connected \cup Conc \cup LRC$
(5) Check Freedom From False Implicatures (FFI):
if `shift(PrevCC,CurrCC) & notClosed(PrevCC)`,
$FFIAgenda = ConcAgenda \cup \{CurrCC\}$
if `CurrNI = PreviousNI & notSurprise(NI)`,
$FFIAgenda = ConcAgenda \cup \{CurrCC\}$
(6) Send FFIAgenda to the surface generator to realise.

4 An Example

Consider example 3. The context after the first user utterance are given in Fig. 1. The relevant part of the World Model contain the dashed concepts in Fig. 2.

EXPRESSIVE ATTITUDES OF INPUT:	
explicit	want(u,know(s,[user(u),wantE(w,u,r), rentE(r,u,c),car(c)]))
EVOCATIVE ATTITUDES OF INPUT:	
explicit	want(u,want(s,know(s,[user(u),wantE(w,u,r), rentE(r,u,c),car(c)])))
EVOKED ATTITUDES FOR RESPONSE:	
	know(s,[user(u),wantE(w,u,r), rentE(r,u,c),car(c)]),
SYSTEM GOAL FOR RESPONSE:	
	want(s,know(s,[user(u),wantE(w,u,r), rentE(r,u,c),car(c),location(r,_)]))
CENTRAL CONCEPTS:	
	topic(1,rentE(r,u,c))
	topic(2,rentE(r,u,c))
DISCOURSE REFERENTS:	
	dr(wantE(w,u,r)), dr(rentE(r,u,c)), dr(car(c)), dr(user(u))

Figure 1: The content of the Context Model after the first user contribution *I want to rent a car.*

The system goal is to know what is the location of the renting event.[5] NewInfo is `location(r,_)` and Agenda is initialised with this. The Accuracy rule prefers this Agenda over other possibilities, and so does

[5]The database contains several car hire companies in several locations, and the system wants to know which of them are given to the user.

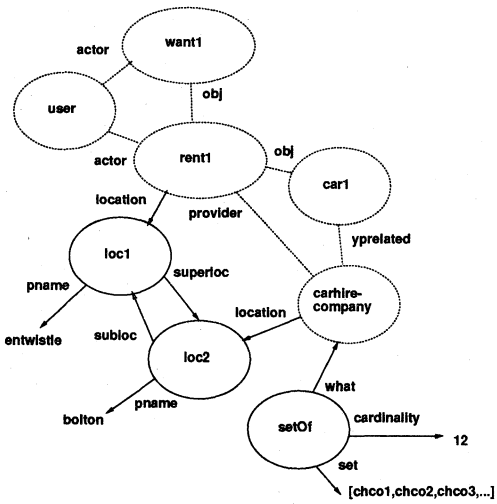


Figure 2: Relevant part of World Model for dialogue 3.

Validity Consistency and FFI. It is chosen as the preferred representation and given to the surface generator to realise.

The situation is different when the system plans a similar type of question $S2$. NewInfo is `superloc(e,_)`. The Agenda, initialised with this concept, is again preferred by the Accuracy and Validity rules. Consistency fails since there is no mapping from `superloc(e,_)` to a lexical predicate, and the superconcept, `location(e,_)`, is tried instead. This mapping is possible, and the concept `superloc(e,_)` is replaced by the concept `location(e,_)` in ConsAgenda. However, FFI does not prefer this Agenda because of false implicatures: the topic is shifted from `rent(r,u,c)` to `location(r,e)` and since the previous topic is not closed (the dialogue can continue with the topic), the current topic must be explicit to prevent the user from interpreting NewInfo as related to the previous topic (the assumption is that elliptical contributions continue the previous topic unless this is closed). Also, since the system does not want to convey surprise of the given location and still repeats NewInfo about location, the topic is added to the Agenda. The preferred FFIAgenda is thus `{location(e,_),pname(e,entwistle)}`, which is realised as *Where is Entwistle?*

5 Limitations and future work

The presented framework provides an intuitively appealing approach to response generation. It regards ellipses as natural utterances which are generated as a side-effect of the speaker fulfilling the obligations of communicative responsiveness. The approach can be compared to [8] who want to generate cooperative responses by over-answering yes-no questions. However, our system is not designed from the point of view of over-answering to extend a response with information that would prevent follow-up questions, but rather, to provide an appropriate and relevant responses on the basis of communicative activity principles. Thus we interleave the planning and realisation processes so that the decisions on the relevant concepts to be included in the response and the realisation of these concepts is flexibly controlled, and

Future work will include extension of the model to more complex 'elliptical' constructions like gapping, and refinement of the pragmatic rules. Also the criterion of consistency, the border-line between DM and NLE, needs more attention.

References

[1] J. Allwood. *Linguistic Communication as Action and Cooperation.* Department of Linguistics, University of Gothenburg, 1976. Gothenburg Monographs in Linguistics 2.

[2] J. G. Carbonell and P. J. Hayes. Recovery strategies for parsing extragrammatical language. *Computational Linguistics*, 9(3-4):123–146, 1983.

[3] N. Green and S. Carberry. A hybrid reasoning model for indirect answers. In *Annual Meeting of the Association for Computational Linguistics*, pages 58 – 65, 1994.

[4] H. Horacek. An integrated view of text planning. In R. Dale, E Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation*, pages 29–44. Springer-Verlag, Berlin, 1992.

[5] K. Jokinen. *Response Planning in Information-Seeking Dialogues.* PhD thesis, University of Manchester Institute of Science and Technology, 1994.

[6] E. Reiter. The computational complexity of avoiding conversational implicatures. In *Proceedings of the ACL*, pages 97–104. 28th Annual Meeting, University of Pittsburgh, Pittsburgh, Pennsylvania, 1990.

[7] E. Reiter. Generating descriptions that exploit a user's domain knowledge. In R. Dale, C. Mellish, and M. Zock, editors, *Current Issues in Natural Language Generation*, pages 257–285. Academic Press, London, 1990.

[8] W. Wahlster, H. Marburger, A. Jameson, and S. Buseman. Over-answering yes-no questions: Extended responses in a NL interface to a vision system. In *Proceedings of the 8th IJCAI Conference*, pages 643–646, 1983.

文章一括処理による係り受け関係の解析

佐々木美樹　　坂本仁
沖電気工業（株）関西総合研究所ＡＩプロジェクト

1　はじめに

自然言語を機械的に処理する場合、文脈を考慮しなければ正しい係り先の決定が行なえないような係り受け関係が少なくない。しかし、従来は、係り受け関係を一文毎に個別に解析するだけであり、解析した結果を他の係り受け関係の処理に利用していない。このため、人間が読めば文章内に明らかな根拠があるのに誤って解析したり、同様の係り受け関係に対して正しく解析する場合と誤って解析する場合が混在したりするなどの問題がある。本発表では、文章全体から抽出した係り受け情報を利用した係り受け関係の解析を試みる。特に、再現性が高く係り受け情報が多く得られる語句単位の係り受け関係として、日本語の名詞句の係り受けと、英語のing-formの係り受けとを例として、文章を一括処理した係り受け関係の解析について述べる。

2　従来の手法の問題点

文章一括処理を行なわない場合は、係り受け関係を一文毎に個別に解析するだけであり、推定した結果を他の係り受け関係の推定に反映しない。例えば、日本語の名詞句では、「アメリカの日本への反発は．．．」という文においては、この文だけで「アメリカ」の係り先は「日本」ではなく「反発」であると意味から推定が可能であるが、「アメリカの業界への反発は．．．」という文においては、「アメリカ」の係り先が「反発」であると推定する手がかりはない。同じ文章内の「アメリカの反発は強く．．．」という文では「アメリカ」の係り先が「反発」であっても、それが手がかりになっていなかった。

また、英語では、辞書の情報により品詞を推定した後に文の解析を行なっているため、文章によって語の品詞や用法が変化することに対応できない。例えば、名詞句「radioactivity monitoring system」「technique underlying Ethernet」はどちらも「名詞＋ing-form＋名詞」の形をしているが、「radioactivity monitoring system」は「放射能監視システム」と訳すのが正しく「monitoring」は前置修飾であり、「technique underlying Ethernet」は「イーサネットの基礎となる技術」と訳すのが正しく「underlying」は後置修飾である。これらの修飾方向は意味から推定が可能であっても、この部分を構文解析によって正しく決定することは多品詞解消が正確に行なわれているという前提が必要であり困難であった。

文章から情報を抽出し曖昧性を解消する試みとして、構文解析の出力においてパックされている係り受け関係を展開して処理することによって前置詞句の曖昧性を解消する手法[1]や、入力表現との間のシソーラスに基づいた意味距離によって前置詞句の係り先のあいまい性を解消する手法[2]が提案されている。

本手法では、計算機で文章から抽出し処理できる表層的な情報だけで、係り受け解析の精度を向上させることを試みる。

3 解析方法

3.1 根拠となる関係の定義

3.1.1 日本語の名詞句

ある語の係り先は、何通りも考えられる。例えば、「アメリカの業界への反発」の場合、「アメリカ」の係り先は「業界」と「反発」が考えられる。しかし、「業界」は「反発」に係るほかない。この係り受けは語の意味を推測するまでもなく確実であると思われる。

よって、係り先の候補が1つしか存在しない、1番最後の係り受け関係を「確実な」係り受け関係とする。これを「根拠となる」関係とする。

更に、名詞の性質を定義する。根拠となる係り受け関係X→Yに対して、ある名詞が、Yである場合よりもXである場合が多いならば、Xを「前置性質」の名詞であるとする。逆に、Yである場合の方が多いならば、Yを「後置性質」の名詞であるとする。

3.1.2 英語のing-form

語が多品詞を持った状態で「根拠となる」関係の抽出を行なう。形態素解析結果が名詞・代名詞・数字・未知語のいずれかを持つ語を根拠抽出時の名詞とする。形態素解析結果が冠詞である語を根拠抽出時の冠詞とする。形態素解析結果が形容詞のみである語を根拠抽出時の形容詞とする。

「名詞＋ing-form＋名詞」の場合は、この部分だけでing-formの係り先を決定することが困難である。しかし、「形容詞＋ing-form＋名詞」の場合は、前の語が形容詞であるため、ing-formは後置修飾ではないと判断できる。「冠詞＋ing-form＋名詞」の場合でも同様である。そこで、「{冠詞｜形容詞}＋ing-form＋名詞」を「前置修飾の根拠となる」関係とする。同様に、「名詞＋ing-form＋形容詞」、「名詞＋ing-form＋冠詞」の場合は、この部分でのing-formは前置修飾ではないと判断できる。そこで、「名詞＋ing-form＋{冠詞｜形容詞}」を「後置修飾の根拠となる」関係とする。

更に、動詞の性質を定義する。あるing-formが、後置修飾の根拠となる関係である場合が多いならば、「後置修飾の性質」の動詞であるとする。前置修飾の根拠となる関係である場合が多いならば、「前置修飾の性質」の動詞であるとする。根拠となる関係により修飾方向が決定したing-formを「根拠となる動詞」とする。

3.2 根拠となる関係による係り受け関係の決定

3.2.1 日本語の名詞句

1番目の名詞をA、2番目の名詞をB、3番目の名詞をCとおいた「名詞＋名詞＋名詞」の3語組に対して、A→Bにのみ根拠となる係り受け関係がある場合はA→B、A→Cにのみ根拠となる係り受け関係がある場合はA→Cにする。A→BとA→Cの両方に根拠となる係り受け関係がある場合、または、A→BとA→Cの両方ともに根拠となる係り受け関係がない場合には、直後の語に係る係り受け関係を優先するように、A→Bにする。

名詞の性質を適用する場合は、Bが前置性質の名詞であるならば、A→BはBの性質上B→Cに比べ成立しにくい弱い関係であるといえるので、A→Cにする。Bが後置性質の名詞であるならば、A→BはBの性質上B→Cに比べ成立しやすい関係であるといえるので、A→Bにする。

3.2.2 英語の ing-form

根拠となる動詞の性質を適用する。ある ing-form が、後置修飾の性質の動詞であるならば、後置修飾にする。前置修飾の性質の動詞であるならば、前置修飾にする。根拠となる動詞がない場合には、前置修飾にする。

4 実験

4.1 日本語の名詞句

日本語では名詞句の係り受け関係として「名詞＋名詞＋名詞」の３語組を対象とした。１番目の名詞をＡ、２番目の名詞をＢ、３番目の名詞をＣとする。計算機関係の文章56000文から、曖昧さがある３語組約8000組を計算機処理により抽出し、４組に１組の割合で抽出した1980組のデータに対して、Ａ→Ｂの係り受けがあるかないか、Ａ→Ｃの係り受けがあるかないか、人手で正解を付与する作業を行なった。

適用した根拠は、以下の様である。

対象	根拠となる関係の種類	根拠となる関係を適用した数
Ａ→Ｂ	約 40000	509/1980
Ａ→Ｃ	約 40000	273/1980

係り先が一致した割合は、以下の様になった。

対象	係り受けがある割合	根拠となる関係を適用	根拠となる関係と名詞の性質を適用
Ａ→Ｂ	82 % (1624/1980)	95 % (483/509)	−
Ａ→Ｃ	33 % (648/1980)	51 % (140/273)	61 % (62/102)

4.2 英語の ing-form

英語では ing-form の係り受け関係として「名詞＋ing-form＋名詞」を対象とした。計算機関係のマニュアル、英字新聞、生産技術関係のレポートから、「名詞＋ing-form＋名詞」の係り受け関係を計算機処理により抽出し、ing-form が前置修飾か後置修飾である適当に抽出したデータに対して、人手で用法を付与する作業を行なった。

文章の分野	文章の語数	抽出した「名詞＋ing-form＋名詞」数(種類)	データ数
マニュアル	約 7688000	15992(3857)	277
新聞	約 1989000	4807(4335)	810
レポート	約 56000	255(211)	217

適用した根拠は以下の様である。

文章の分野	根拠となる関係の種類	根拠となる動詞の種類	動詞の性質を適用した数
マニュアル	3631	426	266/277
新聞	3911	889	727/810
レポート	134	63	143/217

修飾方向が一致した割合は、以下の様になった。

文章の分野	前置修飾である割合	動詞の性質を適用	動詞の性質を適用した正解率
マニュアル	78 % (216/277)	92 % (245/266)	91 % (253/277)
新聞	55 % (445/810)	84 % (611/727)	81 % (658/810)
レポート	83 % (180/217)	96 % (137/143)	87 % (189/217)

5 考察と課題

英語のing-formでは、どの分野の文章においても効果が確認された。動詞の性質を適用したデータが多く、動詞の性質の利用度は高い。根拠となる動詞が適用された割合も高く、効率がよい根拠といえる。文章毎のing-formの修飾方向の傾向に関係なく、修飾方向が一致した割合は向上し、文章間の差は減少しており、本手法は広く有効な手法であるといえる。特に、英語においては、語が多品詞を持った状態で係り受け関係を解析できるという点で、実用性が高い。今回は、「名詞＋ing-form＋名詞」において、ing-formが前置修飾でも後置修飾でもないデータは除いてあるが、今後は、根拠によってそれらを解析することができるかどうか試みていく。

日本語の名詞句では、根拠となる関係が適用された範囲で効果が確認された。実験で用いた3語組で、A→Cの係り受けがあるかないかを判定するという観点においては、A→Cの係り受けがある割合が低いため、すべてA→Cでないとする方がよい結果になるが、A→Cである係り受け関係を絞り込むという観点において、根拠となる関係と名詞の性質を導入した本手法は充分有効であるという見通しが得られた。今回は、根拠となる関係が2語対であるため、英語のing-formの動詞1語の場合に比べ、根拠となる関係を適用したデータが少なく全体の効果に寄与しないと思われるが、根拠となる関係の数を増やすことで適用する範囲を広げるようにして対処していきたい。A→BでありかつA→Cである場合が存在するため、修飾方向を解析する英語のing-formの場合に比べ、一方の解析結果をもう一方の解析に利用することができないという点で難しさがあった。今後は、A→BとA→Cを組み合わせた新たな根拠を探っていく。

6 おわりに

日本語の名詞句、英語のing-formを対象として、文章一括処理による係り受け関係の解析の手法を試みた。本手法は、意味情報や人手を必要とせず、表層的な情報によって文章全体で整合を取り、係り受け関係の正確さを向上させるなどの優れた点があるので、従来の係り受け関係解析を補強する手段として確立を目指したい。

今後は、語句レベルの解析から節や文レベルの解析に拡張していきたいと考えている。

参考文献

[1] 那須川：”文脈制約を利用した曖昧性解消”, 人工知能学会第7回全国大会. (1993).

[2] 隅田, 古瀬, 飯田：”英語前置詞句係り先の用例主導あいまい性解消”, 電子情報通信学会論文誌 Vol.J77-D-IINo.3 (1994).

先行詞探索による文章内容の分類

近藤恵子　東京工芸大学
上里福美　東京工芸大学

1.概要

文脈理解には、辞書、意味解析など多くの問題がある。この解決の一つの糸口として、文章をその内容により分類することを考えた。分類により辞書の範囲をしぼることが可能となり、また意味解析についても複数の選択肢に対する有効な条件となりうる。分類に当たっては、その使用されている名詞に着目した。文章中、重要な名詞は意味的には繰り返し使用される。しかし、実際文章を書くに当たっては繰り返し表現を避けるため、指示詞に変えられていることが多い。[1]では、文章中の指示詞を先行詞に置換する手法を提案した。今回はこの手法を用い、文章中の指示詞を先行詞に置換した後、文章に含まれる一般名詞の使用頻度を調査する。これにより、文章における各名詞の重要度を量る。辞書には予め、その名詞の属する分類内容、例えば「芸術」「数学」「経済」などの分類情報を記録しておく。各名詞の重要度と、辞書に記された分類内容から、文章の内容がどの分類に最も近いかを検討する。

2.先行詞探索手法

2.1.概要

文脈理解の手法構築にあたっては、指示詞の照応する先行詞を如何に検出するかが重要な課題となる。[2]では先行詞照応条件について幾つかの仮説を示し、検証を試みているが、具体的な検討は行っていない。そこで、その仮説に基づいた先行詞照応手法の構築と精度について検討した。

2.2.仮説

[2]では、照応について、以下の4つの仮説があげられている。

(仮説1)照応は段落にまたがることがない。
(仮説2)照応は交差することがない。
(仮説3)複文および重文の先頭を除く節中の照応に対する先行詞は、その節および、その節の上位の節には含まれない。
(仮説4)単文中の照応、および、重文の先頭節で、しかも埋め込み節でない節中の照応に対する先行詞は、それより前の文（段落をまたがってもよい）にある。

2.3.探索

探索は、以下の規則に即して行った。

（1） 複文、重文を単文に直す。
（2） 単文中の指示詞の場合は、前文より探索をする。
（3） 重文の第1文の指示詞の場合は前文より探索する。
（4） 一文による複文の下位文の指示詞の場合は、前文より探索する。
（5） 重文の第2文以後の下位文の指示詞の場合は、重文の前の文より探索する。
（6） 探索中に、前の指示詞に行き当たった時は、その先行詞まで飛び、そこから前へと解析を進める。
（7） 段落の始めまで遡っても照応する先行詞がない時、探索は失敗する。

2.4.結果

仮説のみの探索規則によるシミュレーションの正答率は、約33％であった。これには複数の原因が考えられ、その問題点についての改良を行った。指示詞に前もって与えた単数、複数などの条件は優先させつつ、ある程度の自由度を持たせた。連体詞の指示詞の場合、被修飾語が同じもの、もしくは関連したものが有効であるという条件を付加した。また、倒置文を検出、一文として訂正する。以上の改良により、正答率は約83％にアップした。

3.使用回数の調査

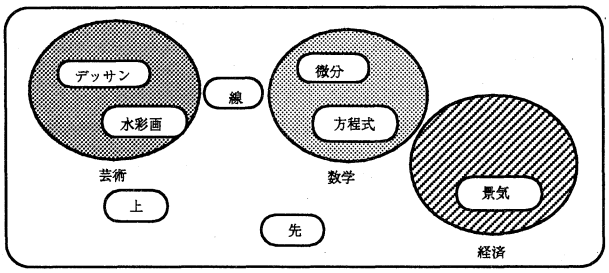
指示詞を先行詞に置き換えたことにより、語の面からの文意が明確になった。これは、指示詞によって意味的に繰り返し使用されていた語が、単語として明らかにされたことによる。この置換された後の文章における単語の使用頻度の高

さは、置換以前の文章中における各単語の意味的な重要度と連動する割合が高い。
今回の分類の手法はキーワードによる方法のため、調査は名詞のみを対象とした。手順を以下に示す。

（1）文章の頭より名詞を探索し、発見した名詞にはマーキングをする。
（2）文章の終わりに向かいながら、名詞を検索する。
（3）発見した名詞が過去にマーキングされていれば個数を数え、そうでなければ、また新たにマーキングを行う。
（4）これを、文章の終わりまで繰り返す。
（5）結果、文章中に使用されたすべての名詞に対して、使用回数が判明する。

4.分類方法

使用回数を利用し、分類を行う。分類は「芸術」「経済」「工業」「数学」等程度の範囲とした。
分類のキーワードとなる名詞には、辞書により分類情報を付した。キーワードとは明らかにその分野の専門用語である名詞を指す。例えば、「デッサン」という名詞には「芸術」の分類情報が付され、分類のキーワードとなる。「景気」という名詞には「経済」の分類が付され、キーワードとなる。この辞書のモデルは以下のように図示される。



名詞分類のモデル1

このキーワードの分類情報により、各分類の確率を計算する。計算は、以下の方法により行う。

$$\text{名詞Aの重要度} = \text{名詞Aの文章全体での使用回数} \qquad (1)$$

$$\text{「芸術」分類確率} = \frac{\Sigma\,(\text{「芸術」のキーワード名詞Aの重要度})}{\text{すべての分野に対するキーワードの出現回数}} \qquad (2)$$

この式は、その特定の文章中に現われたすべての分野のキーワードの使用回数のうち、「芸術」に関するキーワードの使用回数がどれほどであるかを表わしている。分子は文章中に使用された「芸術」に対するキーワードの出現回数と言い替えることもできる。他の分野に対しても、同様の計算でその確率を求めることができる。

文章名	総名詞数	キーワード使用回数		
		芸術	数学	経済
文章A	302	21	4	0

文章Aのキーワード使用回数調査の1例

この例においては、分類は以下の計算により行われる。これを全分野について計算し、最終的に最も確率の高いもの、この文章Aについては「芸術」に分類される。

「芸術」分類確率 ＝ 21／（21＋4＋0） ＝ 0.84
「数学」分類確率 ＝ 4／（21＋4＋0） ＝ 0.16
「経済」分類確率 ＝ 0／（21＋4＋0） ＝ 0

5.分類方法の改良

この方法ではキーワードの少ないものに対しては、精度が低くなる恐れがある。特に、専門語を避けて書かれた初心者向けの入門書について、その傾向が強い。そのため、専門性の多少低い名詞についても、分類に活用することを考えた。先の使用回数の調査において、キーワード以外の名詞に対してもその使用回数は明らかになる。この使用回数は、その文章に対するその名詞の重要度を表わしているという点では、キーワードも他の名詞も同様であると言える。しかし、

キーワードと他の名詞との明確な違いは、キーワードはある特定の分類にのみ含まれ、それ以外の名詞はいくつもの分野にわたり使用されるという点である。そのため使用頻度が如何に高くとも、分類に利用するに際してはキーワードと同等の重要度は持っていないと言える。よって（1）式は成立しない。
ここで、キーワード以外のその名詞が、分類に際してどれほど重要となるか、その重要度に重みを加重することを考えた。重みは学習により与え、式としては次のように書かれる。

名詞Aの分類αに対する重要度 ＝
名詞Aの文章全体での使用回数 × 名詞Aの分類αに対する重み （3）

5.1 重みの学習

重みは、キーワード以外のある特定の名詞Aの使用頻度が、分野によってどれだけ片寄っているかにより示す。「線」というキーワードでない名詞を例に考える。まず、各種の分野の文章について、先の分類方法により使用回数を調査し、計算、比較、分類を行う。この分類結果のモデルを以下に示す。ここでは特に「線」という名詞についてのみを扱った。

分類	文章名	総名詞数	「線」使用回数	文類別総名詞数	文類別「線」使用回数	文類別「線」使用頻度
芸術	文章A	302	32	458	39	0.085
	文章C	156	7			
数学	文章B	123	13	233	16	0.069
	文章E	110	3			
経済	文章D	108	0	108	0	0

分類結果のモデルの1例

文類別「線」の使用頻度とは、各分野での総名詞数に対する「線」の使用回数の割合を表わしている。「線」の「芸術」に対する重みは「線」使用頻度の全分野の合計のうち、「芸術」における使用頻度が占める割合により求められる。

「線」の「芸術」に対する重み ＝ 0.085／(0.085＋0.069) ＝ 0.55

同様の計算により、「数学」に対する重み〈0.45〉、「経済」に対する重み〈0〉を得る。この学習結果は新たな入力文章に対して働き、次の入力文章Fに「線」という名詞が含まれていた時、その分類の計算は、以下のように行われる。

文章名	総名詞数	キーワード使用回数			「線」使用回数
		芸術	数学	経済	
文章F	101	34	2	0	13

文章Fの使用頻度調査

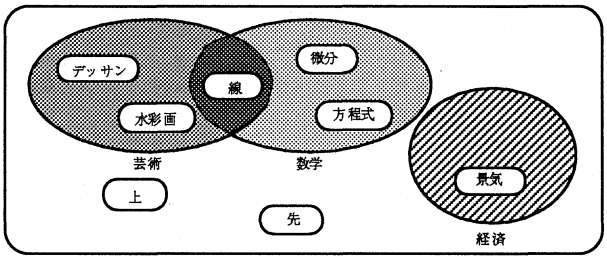
$$「芸術」分類確率 = \frac{34+(13\times 0.55)}{34+2+0+13} = 0.84$$

$$「数学」分類確率 = \frac{2+(13\times 0.45)}{34+2+0+13} = 0.16$$

よって、文章Fは「芸術」に分類される。
この結果よりさらに重みの学習をさせる。上の『分類結果のモデルの1例』の表にさらにこの文章Fを加えることにより、分類「芸術」の総名詞数は〈559〉、「線」使用回数は〈52〉、よって使用頻度は〈0.093〉となり、重みも以下のように変化する。

「線」の「芸術」に対する重み ＝ 0.093／(0.093＋0.069) ＝ 0.57

これにより他の分野に対する重みも変化し、「数学」に対する重みは〈0.43〉となる。
このときに使用した辞書のモデルは以下のように図示される。一つの分野の円の中に収まっている「水彩画」「方程式」などの名詞はキーワードであり、複数の分野の重なりにある「線」のような名詞は重みを加重することにより、二次的なキーワードとなりうる。

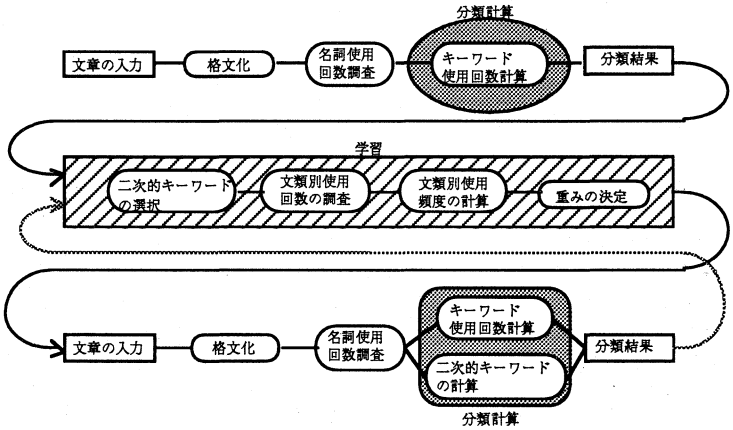


名詞分類のモデル2

このような二次的キーワードは複数の分野にまたがり使用されるが、ある程度の片寄りのあるものが望ましい。そのため、選択に当たっては各分野にわたる多くの文章の分類を行った後、蓄積された名詞の使用頻度の情報を元に、片寄りの大きい名詞を選ぶ必要がある。

6.構成

以上の手順を流れ図により示す。



手順の流れ模式図

7.まとめ

文章の分類に当たっては、キーワードの使用頻度が手がかりとなりうる。また、それ以外の名詞に対しても、重みを付加することにより、二次的なキーワードとなりうるものがある。この重みは各分野にわたる学習を繰り返すことにより、より精度の高い設定が可能となり、分類をより確実なものとすることができる。

参考文献

（2）「先行詞探索手法」
近藤恵子、上里福美　電子情報通信学会 1994年秋季大会講演論文集 情報・システム　D-65,p68,1994
（1）「日本語文章における照応・省略現象の基本的検討」
藤沢伸二、増山繁、内藤昭三　情報処理学会論文誌　Vol.34　No.9,1993

日本語文章における名詞の指示対象の推定

村田 真樹　長尾 眞
京都大学工学部 電気工学第二学科

1 はじめに

日本語文章における名詞の指す対象が何であるかを把握することは，対話システムや高品質の機械翻訳システムを実現するために必要である．そこで，本研究では主題・焦点，名詞の指示性，経験的規則などの情報を用いて名詞の指示対象を推定する．このとき指示詞や代名詞やゼロ代名詞の指示対象も推定する．日本語には冠詞がないことから，二つの名詞が照応関係にあるかどうかを判定することが困難である．これに対して，我々は冠詞に代わるものとして名詞の指示性[1]を研究しており，これを用いて名詞が照応するか否かを判定する．例えば，名詞の指示性が定名詞ならば既出の名詞と照応する可能性があるが，不定名詞ならば既出の名詞と照応しないと判定できる．さらに，名詞の修飾語や所有者の情報を用い，より確実に指示対象の推定を行なう．

2 名詞・代名詞等の指示対象の推定方法

本研究では，以下で述べる名詞の指示性と修飾語と所有者の三つの条件をすべて満足するときのみ照応すると解析する．

表層表現から名詞の指示性を推定し[1]これを利用して指示対象を推定する．指示性が定名詞句の場合は前方にある同一名詞を指示対象とする．指示性が定名詞句以外の場合は，前方の主題と焦点[2]から指示対象を探し，指示性の推定における定名詞句でない度合[3]と主題・焦点の重みと指示対象との距離の三つの情報を組み合わせることにより照応するか否かを決定する．本来は定名詞句以外の場合は既出の名詞を指示することはないが，名詞の指示性の推定を誤ることがあり実際には定名詞句の可能性があるためこのような処理を行なう．

修飾語を持つ名詞については，同じ修飾語を持つ同一名詞であることを照応する条件とする．例えば，以下の文章中の「左の頬」は修飾語の異なる「右の頬」と照応しない．

さて，隣の家に瘤のあるおじいさんがもう一人住んでおりました．
このおじいさんの瘤は右の頬にありました．
（中略）
天狗達は，前の晩に来たおじいさんから取った瘤をそのおじいさんの左の頬に付けてしまいました． (1)

所有者が推定できる名詞の場合は，同じ所有者を持つ同一名詞であることを照応する条件とする．例えば，以下の文章中の「頬」は所有者が同じ「おじいさん」であることから照応する．

さて，おじいさんには（おじいさんの）左の頬に瘤がありました．
それは人の拳ほどもある瘤でした．
まるで（おじいさんの）頬を（おじいさんが）膨らませているかの様に見えるのでありました． (2)

所有者の推定は，意味素性[4]が動物の一部を意味するPARである名詞に対してのみ行なう．その名詞が存在する文の主語かそれまでの主題の中から意味素性がHUM(人間)かANI(動物)のものを探し出して，それを所有者とする．

以上は同一名詞の場合に限って説明したが，以下の例のようにある名詞を末尾に含む名詞がその名詞の指示対象となることがある．

オーストリアのレルヒ少佐が、日本の陸軍将校に、本格的にスキーを教えはじめたのは、明治の末のことである。少佐は日本の軍人たちを前にしてまず「メテレスキー」と号令した。 (3)

このような場合も解析できるように，本研究では以上の方法での名詞が同一であるという条件を末尾に含むという条件に変更する．ただし，末尾に含むという条

[1] 本研究では名詞の指示性として総称名詞，定名詞，特定性不定名詞，不特定性不定名詞を考えている(総称名詞，定名詞，不定名詞の定義は文献[1]を参照のこと．特定性不定名詞句は話者が指示対象を認識している不定名詞句とし，不特定性不定名詞句は話者が指示対象を認識していない不定名詞句とする．[2])．総称名詞，定名詞，不定名詞の判定は文献[1]で行ない，不定名詞における特定性・不特定性の判定は“「(名詞A)が存在しない」ならば名詞Aは不特定性”という表層表現を利用した決定的な規則により行なっている．文献[1]の利用においては以下のように規則を変更した．同一名詞が前方にある場合定名詞などに得点を与える規則があったがこれを省いた．また，意味素性[3]がPARの場合定名詞に得点を加えるという規則を追加した．

[2] 本研究での主題と焦点はそれぞれ表1，表2により定義する．

[3] 文献[1]による指示性の推定では得点を用いており，定名詞句でないと推定した場合得点から定名詞句でない度合が得られる，その度合が小さいほど照応しやすくする．

[4] 本研究では名詞意味素性辞書[3]を用いる．

条件部 ⇒ { 提案 提案 ..}
提案 := (指示対象の候補　得点)

図 1: 候補列挙規則の表現

条件部 ⇒ (得点)

図 2: 候補判定規則の表現

件の場合は同一名詞の場合に比べ照応しにくくなるようにしておく．

指示詞については種類が多く，それぞれに対して詳細に規則を作ることによって指示対象を推定する．代名詞は会話文章中によく現れるので，会話文章の話し手や聞き手を把握することで[5]，指示対象を推定する．ゼロ代名詞は，主題や焦点と格フレームによる選択制限によって推定する [4] .

3 名詞等の指示対象を推定する枠組

3.1 推定の手順

本研究での名詞の指示対象の推定は，名詞の解析の手がかりとなる複数の情報をそれぞれ規則にし，これらの規則を用いて指示対象の候補に得点を与えて，合計点が最も高い候補を指示対象とすることによって実現する．

まず，解析する文章を構文解析・格解析する [5]．その結果に対して文頭から順に文節ごとにすべての規則を適用して指示対象を推定する．規則には，指示対象の候補をあげるための候補列挙規則とその列挙された複数の候補に対して適用する候補判定規則の二種類がある．候補列挙規則は図 1，候補判定規則は図 2の構造をしている．

図中の「条件部」には文章中のあらゆる語やその分類語彙表 [6] の分類番号や IPAL の格フレーム [7] の情報や名詞の指示性の情報や構文解析・格解析の結果の情報などを条件として書くことができる．「指示対象の候補」には指示対象の候補とする名詞の位置もしくは「特定指示として導入」などを書くことができる．「特定指示として導入」のときは，個体を特定指示として新たに導入する．これは特定性不定名詞句 [1] などのように，既出の個体を指示せず談話に新たに特定指示として個体を導入する場合に利用される．「得点」は指示対象としての適切さの度合を表している．

指示対象の推定は条件を満足した規則により与えられる得点の合計点で行なう．まずすべての候補列挙規則を適用し得点のついた指示対象の候補を列挙する．このとき同じ候補を列挙する規則が複数あれば得点は加えてまとめる．次に列挙された指示対象の各候補に対してすべての候補判定規則を適用して，各候補ごとに得点を合計する．最も合計点の高い指示対象の候補を指示対象と判定する．最も合計点の高い指示対象の候補が複数個ある場合は，一番初めに出された指示対象の候補を指示対象とする．

3.2 指示対象の推定に用いる規則

名詞の解析のための規則

名詞の解析は候補列挙規則のみで行なった．候補列挙規則は 9 個作成したが，そのうちの主要なものを適用順序に従って以下に示す．

名詞の解析のための候補列挙規則

1. 推定した名詞の指示性が定名詞の場合で，その名詞を末尾に含み修飾語や所有者が同じ名詞 A が前方にある場合 (ただし，固有名詞の場合は修飾語や所有者の条件を無視する．また，末尾に限らず含まれればよいとする．)
   {(名詞 A 20)}
2. 名詞の指示性が総称名詞の場合
   {(総称指示として個体導入 10)} [6]
3. 名詞の指示性が不特定性の不定名詞の場合
   {(不特定指示として個体導入 10)} [6]
4. 名詞の指示性が総称名詞でも不特定性の不定名詞でもない場合
   {(特定指示として個体導入 10)}
5. 指示性が定名詞以外の場合に適用される．d は文献 [1] によって推定した指示性により決まる値である．定名詞の得点を越える得点を総称名詞と不定名詞が持たない時，$d = 0$．定名詞の得点より 1 点高い得点を総称名詞か不定名詞が持つ時，$d = -3$．定名詞の得点より 2 点高い得点を総称名詞か不定名詞が持つ時，$d = -6$．定名詞の得点より 3 点以上高い得点を総称名詞か不定名詞が持つ場合はこの規則は適用されない．
   { (修飾語や所有者が同じで重みが w で n 個前[7] の同一名詞の主題 $w - n + d + 4$)
   (修飾語や所有者が同じで，今解析している名詞を末尾に含む重みが w で n 個前の主題 $w - n - 5 + d + 4$)
   (修飾語や所有者が同じで重みが w で n 個前の同一名詞の焦点 $w - n + d + 4$)
   (修飾語や所有者が同じで，今解析している名詞を末尾に含む重みが w で n 個前の焦点 $w - n - 5 + d + 4$)}
   主題や焦点の定義と重みは表 1，表 2のとおりである．

[5] 会話文章の話し手や聞き手の推定は，その会話文章の発話動作を表す用言のガ格とニ格をそれぞれ話し手，聞き手とすることによって行なう．会話文章の発話動作を表す用言は，その会話文章に「と言った．」などがつけばそれとし，そうでない場合は前文の文末の用言とする．

[6] 総称指示もしくは不特定指示として導入された個体は，他の名詞から指示されないようにしている．

[7] 主題が何個前かを調べる方法は，主題だけを数えることによって行なう．主題がかかる用言の位置が今解析している文節よりも前の場合はその用言の位置にその主題があるとして数える．そうでない場合はそのままの位置で数える．

表 1: 主題の重み

表層表現	例	重み
ガ格の指示詞・代名詞・ゼロ代名詞	(太郎 が) した.	21
名詞 は / には	太郎 はした.	20

表 2: 焦点の重み

表層表現(「は」がつかないもので)	例	重み
ガ格以外の指示詞・代名詞・ゼロ代名詞	(太郎 に) した.	16
名詞 が / も / だ / なら / こそ	太郎 がした.	15
名詞 を / に /, /.	太郎 にした.	14
名詞 へ / で / から / より	学校 へ行く.	13

指示詞や代名詞やゼロ代名詞の解析のための規則

指示詞や代名詞やゼロ代名詞を解析するため規則を100個ほど作成したが，そのうち主要なものを以下に示す．

指示詞や代名詞やゼロ代名詞の解析のための候補列挙規則

1. 「それ / あれ / これ」や連体詞形態指示詞の場合で，その指示詞の直前の文節に用言の基本形か「～とか」などの例を列挙するような表現がある場合
   {(例を列挙するような表現 40)}
2. 「それ / あれ / これ」や連体詞形態指示詞の場合
   {(前文，もしくは，指示詞の前方の同一文内に逆接接続助詞か条件形を含む用言がある場合はその用言 15)}
3. 名詞形態指示詞か「その / この / あの」の場合
   {(重みが w で主題と焦点を合わせて数えて n 個前にある，同一文中か前文の主題 $w-n-2$)
   (重みが w で主題と焦点を合わせて数えて n 個前にある，同一文中か前文の焦点 $w-n+4$)}
4. 一人称の代名詞の場合 {(話し手 25)}
5. ガ格の省略の場合のデフォルト規則
   { (重みが w で n 個前の主題 $w-n*2+1$)
   (重みが w で n 個前の焦点 $w-n+1$)}

指示詞や代名詞やゼロ代名詞の解析のための候補判定規則

1. 「ここ / そこ / あそこ」であって，指示対象の候補となった名詞が場所を意味する意味素性 LOC を満足する時，10点を与える．
2. ソ系の連体詞形態指示詞の場合に，それが係る名詞Bの用例「名詞Aの名詞B」[8] を検索し，名詞Aと指示対象の候補となった名詞の類似レベルにより得点を与える．
3. 代名詞の場合に，指示対象の候補となった名詞が意味素性 HUM を満足する時，10点を与える．
4. 指示対象の候補となった名詞と格フレームの格要素の用例の名詞との類似レベルにより得点を与える [4] .

3.3 名詞の指示対象の推定例

名詞の指示対象を推定した例を図3に示す．これは図中の下線部の「火」の解析を正しく行なったことを示している．

その時お爺さんはあまり遠くない所にある空き地に火が燃えているのに気が付きました。
赤い顔をして、鼻の青い、恐ろしい目付きの五六人の男が、火の周りに立っているのを見ました。

候補	(前文の) 火	総称指示として導入
2番目の規則		10
5番目の規則	12	
合計	12	10

指示性の推定結果

指示性	不定名詞	定名詞	総称名詞
得点	1	2	3

5番目の規則 $= w-n+d+4$
$= 15-4-3+4=12$

図 3: 名詞の指示対象の推定例

4 実験と考察

指示対象の推定を行なう前に構文解析・格解析を行なうが，そこでの誤りは人手で修正した．格フレームはIPALの辞書のものを用いたが，IPALの辞書にない用言に対しては人手で格フレームを作成した．本研究による方法で名詞，指示詞，代名詞，ゼロ代名詞の指示対象を解析した実験結果を表3に示す．名詞の解析精度は文中に指示対象が存在する名詞についてのものである．これは照応する名詞に注目したためである．ゼロ代名詞の解析精度は指示対象が存在するか否かがあらかじめわかっていると仮定して解析した時の精度である．

また，本稿であげた各手法の有効性を確かめるために指示性の利用の仕方をかえて表4の対照実験を行なった．表のように，本研究の規則による方法では適合率と再現率がともに均等に良かった．これは本研究の規則が指示性を適切に利用していることを意味している．指示性が定名詞句と推定された名詞句のみが照応するとした方法では再現率が悪い．これは指示性の推定の時に定名詞句であるのに他の名詞句と誤って推定し，照応しないとシステムが解析したためである．また，同一名詞はすべて照応するとした方法では，適合率が悪い．これは，同一名詞があってもそれらは照応するとは限らないということを意味している．

[8] この用例には EDR の共起辞書 [8] を用いる．

表 3: 本研究の実験結果

テキスト	文数	名詞の解析	指示詞の解析	代名詞の解析	ゼロ代名詞の解析
学習サンプル	204	85% (130/153)	87% (41/ 47)	100% (9/ 9)	86% (177/205)
テストサンプル	184	77% (89/115)	86% (42/ 49)	82% (9/11)	76% (159/208)

各規則で与える得点は学習サンプルにおいて人手で調節した.
学習サンプル { 例文 (43 文), 童話「こぶとりじいさん」全文 (93 文)[9], 天声人語一日分 (26 文), 社説半日分 (26 文), サイエンス (16 文)}
テストサンプル { 童話「つるのおんがえし」前から 91 文抜粋 [9], 天声人語二日分 (50 文), 社説半日分 (30 文), サイエンス (13 文)}

表 4: 名詞の解析における対照実験の結果

指示性が定名詞句と推定された名詞句のみ照応する	本研究の規則	指示性を用いない	修飾語・所有者の条件を用いない	修飾語・所有者の条件のみ用いる	同一名詞はすべて照応する	末尾に含む名詞とすべて照応する
学習サンプル						
92%(117/127) 76%(117/153)	82%(130/159) 85%(130/153)	72%(123/170) 80%(123/153)	65%(138/213) 90%(138/153)	64%(131/205) 86%(131/153)	52%(134/260) 88%(134/153)	47%(130/279) 85%(130/153)
テストサンプル						
92%(78/ 85) 68%(78/115)	79%(89/113) 77%(89/115)	69%(79/114) 69%(79/115)	58%(92/159) 80%(92/115)	58%(103/178) 90%(103/115)	47%(102/218) 89%(102/115)	44%(106/240) 92%(106/115)

表中の上段と下段はそれぞれ適合率と再現率を表す. 評価に適合率と再現率を用いたのは, 先行詞がない名詞をシステムが誤って先行詞があると解析することがあり, この誤りを適切に調べるためである. 適合率は先行詞を持つ名詞のうち正解した名詞の個数を, システムが先行詞を持つと解析した名詞の個数で割ったもので, 再現率は先行詞を持つ名詞のうち正解した名詞の個数を, 先行詞を持つ名詞の個数で割ったものである. 「同一名詞はすべて照応する」以外の推定では, 先行詞は今解析している名詞を末尾に含むものであることを条件としている. 「指示性が定名詞句と推定された名詞句のみ照応する」は名詞の解析のための 5番目の規則をけずったものに相当し, 「指示性を利用しない」は 1番目の規則をけずり, 5番目の規則の d をけずりこの規則がいずれの指示性の時でも適用されるようにしたものに相当する. 「修飾語・所有者の条件のみ用いる」は 1番目の規則がいずれの指示性の時でも適用されるようにし, 5番目の規則をけずったものに相当する.

修飾語句や所有者を利用して指示対象の絞り込みを行なったが, これが有効に働いている. しかし, 修飾語句が異なっていても照応する場合があり, このような場合は解析を誤った.

そこでおじいさんは近くの大きな杉の木の根元にある穴で雨宿りをすることにしました.
(中略)
次の日, このおじいさんは山へ行って,
杉の木の根元の穴を見つけました. (4)

この例の下線部の「穴」は同一の穴であり照応するが, 修飾語の文字列が異なっているため照応しないと誤って解析された. このような場合についても解析できるようにするには, 異なる表現であっても同じ意味であることを把握できるようにする必要がある.

5 おわりに

本研究での手法は主に名詞の指示対象の推定に名詞の指示性や修飾語や所有者を用いることであった. 実験を通じて, これらを用いることの有効性を示した. 名詞の指示性の推定精度が向上すると名詞の指示対象の推定精度が向上すると考えている. そこで, 名詞の指示性の推定精度を向上させる研究を行なう必要がある.

参考文献

[1] M. Murata and M. Nagao, Determination of referential property and number of nouns in Japanese sentences for machine translation into English, *Proceedings of the 5th TMI*, (1993), pp. 218–225.

[2] 井上和子, 山田洋, 河野武, 成田一, 名詞, 現代の英文法, 第 6 巻, (研究社, 1985).

[3] 渡辺靖彦, 黒橋禎夫, 長尾眞, IPAL 辞書と分類語彙表を用いた単語意味辞書の作成, 情報処理学会第 45 回全国大会予稿集, 6F-8, (1992).

[4] 村田真樹, 長尾眞, 用例を用いた日本語文章におけるゼロ代名詞の指示対象の推定, 「IPAL シンポジウム'95」論文集, (情報処理振興事業協会 技術センター, 1995), pp. 63–66.

[5] S. Kurohashi and M. Nagao, A method of case structure analysis for japanese sentences based on examples in case frame dictionary, Vol. E77-D, No. 2, (1994), pp. 227–239.

[6] 国立国語研究所, 分類語彙表, (秀英出版, 1964).

[7] 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).

[8] (株) 日本電子化辞書研究所, EDR 電子化辞書 日本語共起辞書評価版第 2.1 版, (1994).

[9] 中尾清秋, こぶとりじいさん 他 鶴の恩がえし、きき耳ずきん, 英訳「日本むかしばなし」シリーズ, 第 7 巻, (日本英語教育協会, 1985).

日本語文章の文脈構造の評価方法

小野 顕司　住田 一男
(株)東芝　研究開発センター　情報・通信システム研究所　第三研究所

1　要旨

高精度の自動抄録や要約システムの実現のためには、文章の大局的な構造を解析する文脈解析が不可欠である。文脈構造解析の評価は、文脈構造の記述方法(文間の修辞的関係の種類など)を定めてそれを被験者に学習させ、幾つかの文章に対してその文脈構造を記述させ、共通する部分を正解として、システムの解析結果と照合するのが普通である。

しかし、記述方法の定め方によっては、ある種の文脈構造（論旨の構造）を表現できない。

また、文脈構造の記述方法を学習するには個々の修辞関係がカバーする範囲などについて被験者間で認識を一致させる必要があるが、しかしその調整の仕方によっては、記述が誘導的に、あるいは文章の解釈の押しつけになる危険性をもっている。

本稿では被験者に文脈構造を記述させることなく、文脈構造を評価する方法を提案し、従来方法と同等以上の評価能力があることを実験的に示す。

2　はじめに

我々は、文脈構造を文と文が、順接、逆接、例示といった３０数種類の修辞的関係によって相互に関連づけられたものとして捉え、その解析方法について研究してきた[1] [2] [3] [4]。

読者の主観が反映する重要文の判定に比べて、この修辞的構造の把握は言語的であり、被験者間での一致が高いと思われたが、実際被験者に構造を記述させてみるとおおきなばらつきがあった。

文章によってははっきりした文脈構造がない文章もあるし、記述された構造は、被験者の理解の度合いや解釈をも反映するので、かならずしも記述の完全一致が望ましい訳ではない。

しかし、構造の記述方法の不備による記述のばらつきもかなりあるように感じられた。

日本語論説文を対象とした文脈構造の研究としては、[5] [6] [7] などがあるが、文脈構造を何らかのかたちで外在的に表現し評価しようとする限り同じ問題が発生する。

また、特定の構造表現方法に文脈構造の評価方法が依存していると、異なる表現方法を持つシステム間での比較ができないし、将来表現方法を改善したときに、データを全て作り直さなくてはならない。

そこで、本稿では文脈構造を記述することなく、文脈構造を評価する方法を提案する。

3　評価方法

本実験では、文章を文と接続表現の連接したものと捉える。そして、そのサブセットを生成し、原文と比べて文と文のつながり方が矛盾しないかを判定する。

以下に具体的に方法を述べる。

例えば、図1に示す5文の段落を考える。この文章は、下線を引いた4つの接続的な表現を a,b,c,d とし、それらを除いた文を 1,2,3,4,5 としたとき

[1 a 2 b 3 c 4 d 5]

という系列で表現することができる。この系列から、順序を変えないで、また、文と接続表現が交互するという条件で、任意の部分系列を生成する。例えば、

[1 b 5]

という系列を考える。これは、

"最近の日本語ワードプロセッサは小形・軽量、低価格となり、「Ｒｕｐｏ」などのパーソナルワードプロセッサが一般家庭にまで普及しつつある。また、文書の作成を支援する知能をもったワードプロセッサが開発されるであろう。"

という文章に相当する。この文章に対して、２つの文の意味関係が原文と同じであるか判定する。この場合は、同じであるとは見做しがたい。次に

[1 c 5]

という系列を考える。これは

"最近の日本語ワードプロセッサは小形・軽量、低価格となり、「Ｒｕｐｏ」などのパーソナルワードプロセッサが一般家庭にまで普及しつつある。今後は、文書の作成を支援する知能をもったワードプロセッサが開発されるであろう。"

という文章に相当する。この文章の場合は、原文と意味関係はほぼ同じであると見做せる。

今の例では２文間の関係であったが、実際の文章では３文以上の間での結束性が問題となる場合もある。従って、部分系列としては、２文からなるものだけでなくあらゆる場合を考える。原文が５文である場合、以下の 55 の部分系列が存在する。

[1], [2], [3], [4], [5]
[1 a 2] [1 a 3], [1 b 3] [1 a 4], [1 b 4], [1 c 4] [1 a 5],[1 b 5], [1 c 5],[1 d 5]
[2 b 3] [2 b 4], [2 c 4] [2 b 5], [2 c 5], [2 d 5] [3 c 4] [3 c 5], [3 d 5] [4 d 5]
[1 a 2 b 3] [1 a 2 b 4], [1 a 2 c 4] [1 a 2 b 5], [1 a 2 c 5], [1 a 2 d 5]
[1 a 3 c 4], [1 b 3 c 4] [1 a 3 c 5], [1 a 3 d 5], [1 b 3 c 5], [1 b 3 d 5]
[1 a 4 d 5], [1 b 4 d 5], [1 c 4 d 5] [2 b 3 c 4] [2 b 3 c 5], [2 b 3 d 5]
[2 b 4 d 5], [2 c 4 d 5] [3 c 4 d 5]
[1 a 2 b 3 c 4] [1 a 2 b 3 c 5], [1 a 2 b 3 d 5] [1 a 2 b 4 d 5], [1 a 2 c 4 d 5]
[1 a 3 c 4 d 5], [1 b 3 c 4 d 5] [2 b 3 c 4 d 5]
[1 a 2 b 3 c 4 d 5]

同様に、原文が３文である場合は、８個、４文である場合は、21個、６文である場合は、144 個の部分系列が存在する。

ただし文章中に同じ接続表現がある場合には、部分系列の個数は激減する。

このようにして、１つの段落について、数十から百数十の部分系列が生成され、それぞれについて原文との論旨の整合性が被験者によって判定される。

これらの判定結果には種々の文脈規定要素（接続表現、話題の分布や推移）が反映されており、総体として、この段落の文脈構造を反映しているとみなせる。

以下では、被験者間の判定の一致度について実験した結果を述べる。

1: 最近の日本語ワードプロセッサは小形・軽量、低価格となり、「Ｒｕｐｏ」などのパーソナルワードプロセッサが一般家庭にまで普及しつつある。
2: 一方、機能の面では、当社のかな／漢字変換入力方式はＪＷ－８Ｄ、べた入力かな／漢字変換ができるようになった。
3: また文書を編集校正し印刷する機能の豊富さはすでに満足な程度に達しており、使用者の希望どおりの文書ができるようになった。
4: 今後は個々の機能についてマンマシンインタフェースの面から改良されていくだろう。
5: また文書の作成を支援する知能をもったワードプロセッサが開発されるであろう。
東芝レビュー vol.41 No.7, page 616 (1986) より抜粋

図 1: テキスト例

4 評価方法の評価

東芝レビューおよび情報処理学会誌から採取した3, 4, 5, 6 文からなる段落各 10 個計４０個に対して、全ての部分系列とそれに対応する文章を作成し、原文との論旨の整合性を３人の被験者に○, △, ×の３段階評価で判定させた。その後、被験者間での一致度を計算した。結果を表 1に示す。表中”完全一致”は、△判定があった場合をカウントしていない。”一致”は、△判定のものをも含んでいる。

文の数が増えるに連れて、一致度が下がっている。３文のときの完全一致度、および５文のときの値が悪いのは、偶然的な理由によるものと思われる。

従来方式による被験者間での一致度を表 2に示す。この実験は、高校国語教科書および新聞社説から採取した4, 5 文からなる段落各 20 個計４０個に対して、３人の被験者に直接構造を記述させ、その記述の一致度を計算したものである。構造の記述方式は、内容的にまとまる文と文を括弧でくくるというものであり、文と文の間の、あるいはそれらがまとめられたものの間の修辞的関係についてはみていない。

従来の評価方法との違いを列挙すると、以下のようになる。

- 被験者は文脈構造の記述方式を学習する必要がない。
- 従って、記述方式の学習の過程で文章の解釈が誘導的になる危険がない。
- 判定結果の一致度は、従来と同等か、若干良い。この改善は、構造記述方法の不備による記述のばらつきが減ったためである。
- ただ、１つのパラグラフの判定のために、数十から百数十の文章の判定をしなければならなず、実験に必要な労力が大きい。

5 おわりに

本稿では被験者に文脈構造を記述させることなく、文脈構造を評価する方法を提案した。本稿で述べた方法は、被験者が文脈構造の記述の仕方を勉強する必要がなく、また、直接構造を記述させる方法より、被験者間の一致度が若干高い。

	完全一致	一致
3文	51 %	67 %
4文	58 %	65 %
5文	43 %	57 %
6文	53 %	61 %

表 1: 本評価方式による結果の一致度

	完全一致	一致
4文	40 %	63 %
5文	30 %	55 %

表 2: 従来方式による結果の一致度

部分系列群に対する被験者の判定結果には、種々の文脈規定要素（接続表現、話題）が総合的に反映されている。また、構造の解析方法や記述方法が将来変化しても、データとして利用することができる。また、本方式は、明示的な接続表現が無い場合でも同様に行なうことができる。

文脈構造解析システムを評価する際には、そのシステムが解析した結果から文章を生成させ、その文章と上述の判定結果を比較することになる。その方法については、文献[8]などを参考にしてこれから検討する予定である。

今回は1段落内での文脈構造の評価をおこなったが、今後は段落間での構造評価についても本方式を適用してみたい。

参考文献

[1] 小野，浮田，天野：文脈構造の分析，情処研資 NL70-2, 1989.

[2] 木下，小野，浮田，天野：日本語テキスト理解における文脈構造抽出法，「談話理解モデルとその応用」シンポジウム，pp.125-136, 1989.

[3] 小野，住田，浮田，天野：文章の分割と文脈構造の解析，第43回情処全大論文集 3，pp.251-252，1991.

[4] Sumita,K.,Ono,K.,Chino,T.,Ukita,T., and Amano,S.: A Discourse Structure Analyzer for Japanese Text, Proceedings of International Conference on Fifth Generation Computer Systems, Vol.2, pp.1133-1140, 1992.

[5] 辻井潤一：論説文における文脈構造，日本学術振興会　文字言語・音声言語の知能的処理第152委員会第7回研究会資料 7-1, 1988.

[6] 福本，安原：文の連接関係解析に基づく文章構造解析，情処研資 NL88-2, pp.9-16, 1992.

[7] 間瀬，大西，杉江：説明文の抄録生成について，信学技報, Vol.89, No.457, NLC89-40, pp.5-12, 1990.

[8] 佐久間まゆみ編集: 文章構造と要約文の諸相、くろしお出版、 Frontier series 4, 1989.

動詞の多義性解消における格の弁別能力と集中度の有効性について

藤井 敦*　　秋山 典丈　　徳永 健伸　　田中 穂積

東京工業大学 工学部

本論文は、システムに蓄えられた例文と入力文との間で、対応する格が取る名詞の類似度に基づいて動詞の多義性を解消する従来の手法に、「格の弁別能力」と「格の集中度」という二つの尺度を導入した新しい手法を提案する。「格の弁別能力」とは格が動詞の語義を弁別する能力を表す尺度であり、取り得る名詞の例 (事例) が語義によって異なる格ほど語義の弁別能力は大きくなる。動詞の多義性解消において弁別能力の大きい格をより重視する。「格の集中度」とは、格が取る事例の意味的な広がりを表す。広がりが大きい格は様々な名詞を取りやすいことから、その格が持つ事例と入力文中の名詞との間に、より大きな類似度を割り当てる。いくつかの動詞について新聞記事を中心とするコーパスを用いて実験を行い、二つの尺度を考慮した場合に動詞の多義性を解消する精度が向上することが確認された。

1　はじめに

動詞の多義性解消は自然言語解析における重要な処理の一つである。機械翻訳における訳語選択や、係り受け解析における名詞句の係り先の決定においても動詞の多義性解消が重要な役割を担う。

動詞の語義ごとに格フレームを記述したIPAL動詞辞書 (以下IPALと略す) [1] の例文とシソーラスを用いて動詞の多義性を解消する新しい手法が黒橋らによって提案されている [2]。これは、入力文とIPALに登録された例文の間で対応する格が取る名詞の類似度を計算して最も類似度の高い格フレームを選択する方法である。本研究では、この方法に二つの新しい尺度を導入して動詞の多義性を解消する精度 (解析精度) の向上を試みた。

第一に、動詞の語義を弁別する能力はガ格やヲ格などの各々の格によって異なることが指摘されている [3]。また、同様の概念として「かかりの広さ」が提案されている [4, 5]。これは、かかり要素がどの程度述部を限定するかという概念であり、述部を限定する程度が強いほどかかりの広さは狭くなる。しかし述部の持つ意味的曖昧性までは解消しない。本手法では、格が動詞の語義を弁別する能力を表す尺度として「格の弁別能力」を導入する。格の弁別能力とは、「かかりの広さ」を動詞の語義の弁別にまで拡張したものであり、動詞の語義を限定する力が強い格ほど弁別能力は大きくなる。

直観的には、格が取り得る名詞が、動詞の語義によって差が無い格ほど弁別能力は小さくなる。例えばIPALにおいて動詞「かける」には34個のサブエントリがある[1]。「彼が眼鏡をかける」という文の「かける」は、そのうちの一つ「ひも状のような物を身体に付ける」のサブエントリに属する。この例文の動詞の語義を決めるのは「彼」よりも「眼鏡」であると考えられる。これは、「彼」という単語が「かける」の別の語義 (水をかける、橋をかけるなど) のガ格にも現れることができるので、「彼」だけでは動詞の語義を決めるのが決められないのに対して、「眼鏡」は他の語義のヲ格には現れにくいので動詞の語義を決定しやすいためだと考えられる。つまり、この例では動詞「かける」のヲ格はガ格よりも弁別能力が高いと考える。

第二に、格が取り得る名詞間の意味的な広がりを表す尺度として「格の集中度」を導入する。名詞間の意味的な類似度が高いほど集中度が大きくなり、逆に類似度が低いほど、つまり広がりが大きいほど集中度は小さくなる。集中度が小さい格は様々な意味の名詞を取る傾向があるといえる。

黒橋らの手法では、システムが持っている格ごとの名詞の例 (以下、事例と記述する) の数が不十分である場合に、入力文中の名詞が、複数の語義の (対応する格の) 事例との間で同じ類似度となり、多義性が解消できないという問題が生じる。例えば、動詞「くわえ

*fujii@cs.titech.ac.jp

[1] 漢字の表記の違いを含む。

る」についてIPALでは5つのサブエントリが用意されている。そのうちの二つ「ある事に対して何らかの見解を付け足す」(語義1とする)と「これまであった物にさらに新しく何かを付け足す」(語義2とする)のヲ格について考えてみる。IPALではヲ格の事例として、それぞれ「配慮・解釈・検討・批判・説明」、「塩・新人・彼・(ゴッホの)絵」が挙げられている。今「得点をくわえる」(語義1に属する)という入力があったとき、「得点」と先に挙げた語義1、語義2の事例との類似度を分類語彙表を用いて計算すると、どちらの語義に対しても類似度が0となる[2]。

格の集中度を考慮すると、語義1の事例は互いに意味的な類似度が高いのに対して、語義2の事例には意味的な類似度が低く[3]、集中度が小さい。そこで、語義2のヲ格は様々な意味の名詞を取る傾向があると考え、「得点」と語義2の事例との間に、より大きい類似度を割り当てる。

本稿では、2節で黒橋らのアルゴリズムについて概説し[2]、3節で格の弁別能力について、4節では格の集中度について説明する。さらに5節ではコーパスを用いた評価実験について述べる。

2 基本アルゴリズム

動詞の多義性解消に関して、黒橋らは、格フレーム辞書としてIPAL、シソーラスとして分類語彙表[6]を用いた方法を提案している[2]。そのアルゴリズムを概説する。

(1) 入力文の格要素と格フレームの事例の対応付けを行う。

(2) 対応付けられたそれぞれの格要素について、入力文の名詞と格フレームの事例との間の類似度を計算する。類似度の値は分類語彙表における二つの名詞の分類コードの一致するレベルによって決まる。一致するレベルと類似度との関係を表1に示す。

(3) 式1に従って格フレームと入力文との対応の評価値($score$)を計算し評価値の最も高い格フレームを選択する。

$$score = \begin{cases} 0 & (\text{if } l > n) \\ sum \times \sqrt{\frac{1}{n}} \times \sqrt{\frac{n}{m}} & (\text{otherwise}) \end{cases} \quad (1)$$

n: 対応付けされた格要素数

l: n + (入力文側の対応付けられていない格要素における必須要素数)

[2]黒橋らの手法による

[3]分類語彙表による

m: n + (格フレーム側の対応付けられていない格要素における必須要素数)

sum: 対応付けされた格要素の類似度の和

l,m,nを図1に示す。長方形は名詞を表す。ガ・ニ・ヲ・ヘ・ヨリ格を必須要素とする(ただしニ・ヨリを伴う格要素が時を表す場合は任意要素とする)。

表1: 名詞XとYの間の分類コードの一致レベルと類似度との関係

一致するレベル	0	1	2	3	4	5	6	一致
類似度 ($sim(X,Y)$)	0	0	5	7	8	9	10	11

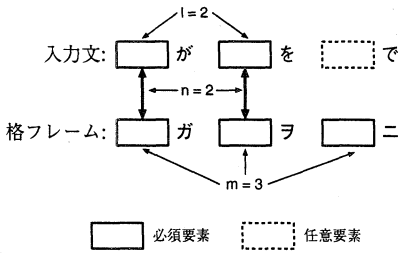


図1: 入力文と格フレームの対応付け

3 格の弁別能力

本論文で扱う記号を示す。

- $N_{s,k}$: 語義s, 格kの事例の集合
- $N^i_{s,k}$: $N_{s,k}$のi番目の要素
- $|N_{s,k}|$: $N_{s,k}$の要素数
- I_k: 入力文中の格kが取る名詞
- $sim(X,Y)$: 名詞XとYの間の分類コード(表1)に基づく類似度
- $d(k)$: 格kの弁別能力
- $c(s,k)$: 語義s, 格kの集中度

格の弁別能力とは格が動詞の語義をどの程度限定するかを表す尺度であり、格が取る事例が語義によって差が無いほど弁別能力は小さくなる。格の弁別能力は各動詞の格ごとに計算する。

n個の語義を持つ動詞の格kの弁別能力$d(k)$を式2で与える。ただし、分類コードがレベル5まで一致する単語どうしを共通要素($N_{i,k} \cap N_{j,k}$)とする。

$$d(k) = \frac{1}{{}_nC_2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{|N_{i,k}| + |N_{j,k}| - 2|N_{i,k} \cap N_{j,k}|}{|N_{i,k}| + |N_{j,k}|} \quad (2)$$

$N_{s,k}$の中でI_kとの間の類似度が最も高い事例を$E_{s,k}$とする。入力文に現れた格i $(i = 1 \ldots K)$について類似度の加重和を計算し、これを式1におけるsumとする。

$$sum = \sum_{i=1}^{K}(exp(\alpha \cdot d(i)) \times sim(I_i, E_{s,i})) \quad (3)$$

αは定数である。評価値($score$)の計算は式1と同じである[4]。

5節の実験に用いた動詞について、IPALの事例を用いて計算した格の弁別能力の値を表2に示す[5]。

表2: 格の弁別能力

動詞	ガ	ニ	ヲ	大小関係
のる	0.895	0.950	—	ガ < ニ
わかる	1.00	0.480	—	ニ < ガ
とる	0.661	—	0.989	ガ < ヲ
おさめる	0.463	1.00	0.992	ガ < ヲ < ニ
くわえる	0.625	0.939	0.972	ガ < ニ < ヲ
あたえる	0.876	0.944	0.978	ガ < ニ < ヲ

4 格の集中度

格の集中度とは格が取る複数の事例が意味的にどの程度類似しているかを表す尺度であり、各語義の格ごとに計算する[6]。

$N_{s,k}$中の事例の全ての組について類似度を計算し、その平均を語義s, 格kの集中度と考え、式4で与える。$m = |N_{s,k}|$とする。

$$c(s,k) = \begin{cases} \dfrac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} sim(N_{s,k}^i, N_{s,k}^j)}{{}_mC_2} & (\text{if } m > 1) \\ \text{最大値} & (\text{if } m = 1) \end{cases} \quad (4)$$

$m = 1$すなわち事例が一つしかない場合、集中度は類似度の最大値(表1では11)とする。

集中度が小さい格は、より多くの名詞を取りやすいので、分類語彙表の分類コードを用いて計算した類似度(2節参照)を格の集中度によって補正し、集中度の小さい格により高い類似度を与える。nを語義数とするとき、I_kと$E_{s,k}$の間の補正後の類似度$sim'(I_k, E_{s,k})$を式5で計算する。βは定数である。

$$sim'(I_k, E_{s,k}) = sim(I_k, E_{s,k}) + \beta\frac{1/c(s,k)}{\sum_{i=1}^{n} 1/c(i,k)} \quad (5)$$

5 評価

5.1 実験

本研究で提案した手法を評価するために、新聞記事を中心とするコーパスに含まれる動詞の多義性解消の実験を行った。コーパスはあらかじめ人手によって係り受けと語義を決定し、データとして用意する。実験に用いたデータを表3に示す。

表3: 実験に用いたデータ

動詞	語義数[7]	データ数
のる	10 (8)	126
わかる	5	60
とる	29 (18)	84
おさめる	8	108
くわえる	5 (4)	168
あたえる	4	528

これらのデータを動詞ごとに6等分し、そのうちの一つをテストデータ、残りの5つを訓練データ群とする。初回はIPALの例文を与えてテストデータの解析精度を調べる[8]。以降、訓練データ群をシステムに一つずつ与えて、事例・格の弁別能力・集中度を更新しながらテストデータの解析精度を調べる。図2に実験の概略を図示する。

テストデータを変えて同様の操作を6回行い、訓練回数ごとに平均を計算する。以上の操作によって、各動詞ごとに訓練回数と解析精度との関係を求める。

実験に用いた全ての動詞の多義性解消の精度(解析精度)を平均したグラフを図3に示す。グラフの横軸はシステムの訓練回数、縦軸は解析精度(%)を表す。比較のために以下の4通りの方法で実験を行った。グラフの番号はこれらの手法の番号に対応している。

1. 格の弁別能力・集中度の両方を考慮
2. 格の弁別能力のみを考慮
3. 格の集中度のみを考慮
4. どちらも考慮しない(黒橋らの手法)

[4]ただし、入力文中の全ての格の名詞が例文の名詞と完全に一致した場合は格の弁別能力や集中度は考慮しない。

[5]必須要素のみ示す

[6]弁別能力は語義ごとではなく動詞ごとに計算することに注意。

[7]IPALにおける語義数で、「乗る」と「載る」のような漢字の表記の違いも含む。括弧内の数字は実験に使用したコーパスに出現した語義数。

[8]語義を一意に特定でき、尚かつそれが正解であるものを正解とし、(解析精度) = (正解数)÷(テストデータ数)で計算する。

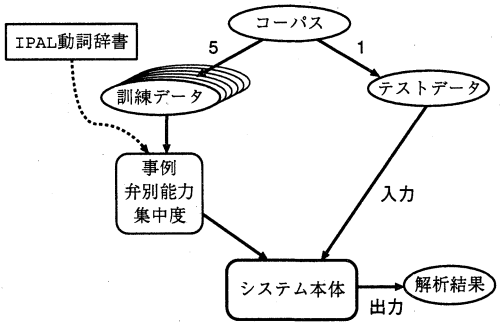


図 2: 実験の手順

個々の動詞ごとに 0〜5 回の訓練における解析精度の平均を計算した例を表 4に示す。1 から 4 までの番号は図 3における各手法と対応している。最後の列は手法 1 と 4 の解析精度の差を表している。

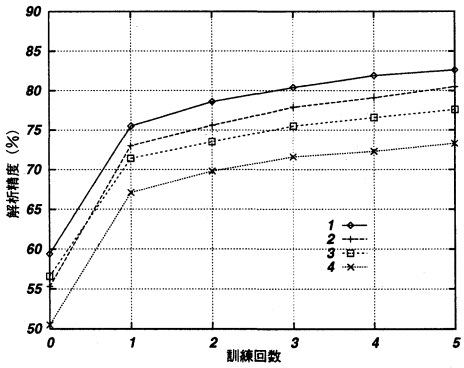


図 3: 訓練回数と解析精度 (%) との関係

表 4: 動詞ごとの解析精度 (%)

動詞	4	3	2	1	1 − 4
のる	73.8	79.5	74.9	79.6	5.8
わかる	50.8	56.9	55.8	65.0	14.2
とる	47.8	50.0	63.7	64.9	17.1
おさめる	61.4	61.6	65.3	65.4	4.0
くわえる	73.8	79.2	78.8	81.2	7.4
あたえる	70.1	75.0	76.9	79.5	9.4

5.2 考察

今回の実験によって以下のことが確認された。

- 本研究で提案した格の弁別能力・集中度の導入によって動詞多義性解消の解析精度が向上する。表 4において、最大で 17.1% 精度が向上した (動詞「とる」)。
- 本手法は、事例データが少ない初期の段階から解析精度を向上させることができる。図 3において、手法 1 では 1 回の訓練で手法 4 の最終的な解析精度を上回っている。事例ベースのシステムにおいて事例データが十分得られないような場合に本手法は有効である考えられる。
- 事例データの増加とともに解析精度が向上する。

6 おわりに

本研究では、従来の動詞多義性解消の手法に格の弁別能力と集中度を導入し、評価実験を行うことでその有効性を示した。

今後の研究課題として、本論文中のα、β等の定数の選び方も含め、弁別能力や集中度の計算式についてさらに検討する予定である。

コーパスでは格の省略が頻繁に起こる。格の省略をどのように扱うかについても検討の余地がある。

また、分類語彙表は名詞の多義性についてあまり考慮されておらず、規模としても大きくないことから、より大規模なシソーラスを利用して名詞の類似度をより精密に計算することなども今後の課題である。

参考文献

[1] 情報処理振興事業協会技術センター. 計算機用日本語基本動詞辞書 IPAL, 1987.

[2] Sadao Kurohashi and Makoto Nagao. A Method of Case Structure Analysis for Japanese Sentences Based on Examples in Case Frame Dictionary. *IEICE TRANSACTIONS on Information and Systems*, Vol. E77-D, No. 2, pp. 227–239, February 1994.

[3] 益岡隆志. 命題の文法—日本語文法序説. くろしお出版, 1987.

[4] 佐伯哲夫. 現代日本語の語順. 笠間書院, 1975.

[5] 徳永健伸, 岩山真, 乾健太郎, 田中穂積. 日本語語順の推定モデルとその応用. 情報処理学会 自然言語処理研究会, Vol. 81, No. 2, pp. 9–16, 1991.

[6] 国立国語研究所（編）. 分類語彙表. 秀英出版, 1964.

幾何的モデルにおける空間的な量に関する形容詞の意味解釈

山田 篤
奈良先端科学技術大学院大学 情報科学研究科
e-mail: yamada@is.aist-nara.ac.jp

1 はじめに

本稿では、具象空間内に存在する対象の属性の形容に用いられる、「高い」、「深い」、「大きい」、「長い」等の空間的な量をあらわす形容詞の意味の取り扱いについて述べる。我々はこれまでに、自然言語記述と現実世界との対応関係への興味から、日本語の情景描写を対象として、言語表現からの対象世界の幾何的モデルの再構成 (可視化) について研究を行なってきた[3, 2]。そこでは、対象の空間的な配置に焦点をあて、場所を表す句によって示される対象間の空間的関係の抽出と、その幾何的モデル上での解釈を検討した (図1)。本稿

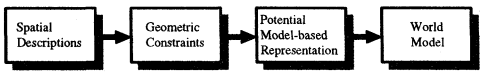


図 1: 言語表現からの幾何的モデルの再構成過程

では、この過程において言語表現として先に挙げたような形容詞が対象の形容に用いられた場合に、それが実際の幾何的モデルとどのように対応するかについて検討を加える。

一般に、言語表現における形容詞の出現形態としては体言修飾 (**大きい**象)、用言修飾 (**高く**飛ぶ)、述語 (象は**大きい**、その象は**大きい**) 等があり、本来はそれぞれの場合についてその取り扱いについて検討を加えるべきであるが、本稿では特に、形容詞と名詞との関連について、名詞によって表されている対象が特定の形容詞によって修飾される属性と、その対象のもつ幾何的な量との関連を問題にする。より、具体的なタスクとしては、

- 指示対象の同定 (例えば、「高いビル」といわれた場合に幾何的モデルの中からそれに合うものを探す)
- 指示通りの変形 (例えば、「そのビルは高い」といわれた場合に指示に合うように対象を変化させる)
- 質問への応答 (例えば、「そのビルは高いか」ときかれた場合に判断を下す)

を想定している。

本稿で対象としている形容詞を表1に示す。これらは計算機用日本語基本形容詞辞書 IPAL[6] 中、意味分類として「空間的広がり；量」をもつとされているもの (16個、反義語毎にペアにすると8対) である。

表 1: 空間的量を表す形容詞

量＋	高い	深い	厚い	太い	広い	長い	大きい	遠い
量－	低い	浅い	薄い	細い	狭い	短い	小さい	近い

2 従来の研究

岡田 [5, 4] は属性概念を差に関する概念として捉え、形容詞の分類を行なった。この分類によれば、表1にあげた形容詞はすべて単純概念であり、概念構造としては主体と比較対象のみを要求するもっとも単純な形となる (例外は「遠い」、「近い」で、これらは更に起点、目標を要求する)。また、内容としては、すべて「場所」と「形」に分類されており、このうち、

1. 場所のみを表すもの…遠い、近い
2. 形のみを表すもの…厚い、薄い、太い、細い、広い、狭い、長い、短い、大きい、小さい
3. 両方を表しうるもの…高い、低い、深い、浅い

となっている。形容詞の場合、比較という視点は重要であり、何を比較対象として選ぶか (比較基準) が問題となるが、より具体的なレベルでは主体と比較対象それぞれの何を比較するのかを明らかにする必要がある。

また、形容詞に関する別の問題として、多義性の判別があげられる。例えば、「高い」は必ずしも空間的量を表すとは限らない。「高い万年筆」はおそらく価格に

ついていっているものと考えられる。内海ら[7, 8]は形容詞-名詞対について、形容詞がもつ複数の語義の中から語義の偏向、被修飾語、文型に関する情報を用いて一つを選択し、それに基づいて被修飾語の属性値集合を変更する方法を提案している。この結果、例えば「深い」の基本義に「#空間的奥行+」というものがあった場合にこれに修飾された「皿」の属性値集合が変更され、「#空間的奥行+」になる。ここで特に問題となるのは形容詞に対する被修飾語(の種類)の分布をその語義毎に知る必要があるという点である。これは大規模なコーパスに基づく分析によってある程度明らかになるかもしれない。また、対象に与えられた属性値の解釈については何も言及していない。処理のある段階において、「深い」の多義の中から「#空間的奥行+」を選び出すことは必要であるが、実際には、「#空間的奥行+」の測り方は対象によって異なり得るし、場所に関する読みと形に関する読みの区別も必要となる。

3 場所に関する形容詞の取り扱い

まずはじめに、先の分類で場所を表すとされている形容詞について、その幾何的モデル上での取り扱いについて考えてみる。これらに共通することは、幾何的モデル上で関与するのは対象の位置パラメータであるということである。

「遠い」、「近い」 例として、「遠いビル」を取り上げる。これは明らかにビルの場所に関する言明であるから、幾何的モデルにおいて関連するパラメータはビルの位置を表すパラメータである。一方で、基準となる場所(「～から」あるいは「～に」に相当)が存在し、この基準となる位置と対象の位置の間の距離が考慮される空間的量となる。これが何らかの判断基準に基づいて、大であると判定されれば「遠い」、小であると判定されれば「近い」となる。

「高い」、「低い」(場所) 「高い」(「低い」)の場所に関する用法も同様の構造をもつ。例えば「高い雲」では、雲の位置パラメータが関与し、これに対して基準面(多くの場合は地平面)から上(重力の向きとは逆)の方向に測った距離が考慮される空間的量となる。

「深い」、「浅い」(場所) 「深い」(「浅い」)の場所に関する用法についてもほぼ同様であるが、「高い」(「低い」)とは距離を測る向きが異なる。例えば「深い震源」では、震源の位置に対して基準面(この場合も地平面)から下(重力の向き)の方向に測った距離が考慮される空間的量となる。

4 寸法に関する形容詞の取り扱い

先の分類で形を表すとされていたものは、幾何的モデル上ではすべて対象自身が持つ寸法に関する量と関係する。現代形容詞用法辞典[9]からこれらの形容詞の空間的用法に関わる語義を抜き出し、表2に示す。

表2: 空間的寸法に関する形容詞の語義
(現代形容詞用法辞典[9]による)

高い	空間的に基準となる面からの距離が遠い様子 (鉛直方向の距離が一般的)
深い	空間的に基準面からの距離が大きい様子 (鉛直方向、基準面から下へ向かう場合が一般的)
厚い	ものの一面から他の面までの距離が長い様子
太い	細長く伸びているものの幅がある様子 (断面積、幅)
広い	空間的に面積が大きい様子
長い	空間的に連続しているものの二点間の距離が大きい様子 (特定の基点は考えない) (直線でも、曲がりくねっていてもよい) (鉛直方向については位置を変更できること)
大きい	形態などが大である様子

これを関連する量に注目して整理すると、

- 距離に関するもの…高い、深い、厚い、太い、長い
- 面積に関するもの…太い、広い
- その他…大きい

となる。

4.1 単純幾何的意味

今、図2に示すような直方体状の物体を対象として形容詞と対象の空間的な寸法の間の関係について簡単な思考実験をしてみる。上に挙げたそれぞれの形容詞

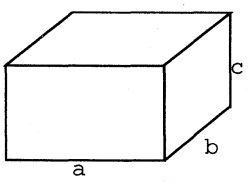


図2: 直方体状の物体に付随する3つの寸法

と関連する量はどれか。直観的には「高い」に対して

はc、「深い」はc、「厚い」はc、「太い」はなし、「広い」はaとb、「長い」はa、「大きい」はaとbとcがそれぞれ関与するように思われる。ここでとった判断基準を分析してみると、以下のようになる。

高い: 垂直方向の寸法
深い: 垂直方向の寸法
厚い: 3つのうち最も小さい寸法
太い: 細長さが要求されるため使えない
広い: 最も大きい面の面積
長い: 最も大きい寸法
大きい: 全体の体積

これはこの判断が絶対に正しいというものではない。また、これらは対象の形状と現在の配置のみに基づいた判断であるが、実際には、その対象によって下される判断がある。例えば、単純に「高い」、「低い」とは「高さ」に関する比較であり、「高さ」とは垂直方向の長さの差であるとすると、3次元空間内の物体は必ずなにがしかの垂直方向の「高さ」を持つことになり、それに基づいて「高い」か「低い」かが言えてしまうことになる。しかし、人間が横になっていても重力方向の量を測るわけではないし、ペンについて高いか低いかという言い方をすることもない。このように考えてくると、対象による違いを適格に捉える必要があることがわかる。

4.2 対象に関する知識

形容詞と関連する対象の寸法を判断する際に、具体的な対象に関してどのような知識が必要となるかについて検討する。例として机を形容してみる。

高い机: 天板を上にして標準的に置かれた場合の垂直方向の距離
深い机: 不可[1]
厚い机: 不可[2]
太い机: 不可
広い机: 天板の面積
長い机: 正置した状態で使用者から見た横方向の距離
大きい机: それが占める空間の全体的な体積

種々の対象について同様の検討を加えた結果を整理して以下の着目点を得た。

[1]ここでは、「?奥行きが深い机」といった関係節による修飾によって測り方が特定される場合を含めない。
[2]提喩を認めると、これは天板の厚い机という解釈がありうるが、机全体としての「厚さ」は考えられない。

対象の基準的な見方 対象によっては基準的な見方が存在する。例えば、机に対しては正置された状態の対象に対面している典型的な使用者の見方がこれにあたる。これは、空間記述における投影的な関係を理解する際にとられる内在的/外在的/直示的用法 [1] とも関連する。これらの区別は、前後左右といった方向をもとにして対象間の位置関係を捉える際に、その方向をどのようにして得るかによって生ずる。例えば、「車の前のボール」に対して、その位置関係を

直示的: 話者から見た車との関係
内在的: 車自身の向きに関連した関係
外在的: 車の動きの向きに関連した関係

として捉えるという違いを表す。

今、対象が内在的な向きを持っていれば、それによって参照フレームが定まる。対象が内在的な向きを持たないか、それが参照フレームの確立に用いられないならば、対象への到達可能性、動き、近傍の物体、地球の重力などの外在的な向きが対象に付与される。それらに全く関係なく観測者の位置を基準にして参照フレームが定まるのが直示的な用法である。これらの場合に、視点が対象の外部にあるか、内部にあるか、また、対象の大きさや顕現性、可動性などがフレームの決定に影響を与えることが知られている。さらに、重力の存在によって垂直方向の次元が特権的な方向となり、特に垂直方向では直示的用法が使えない。

機能との関連性 同じ天板に注目しているにも関わらず、天板が広い机は「広い机」といえるのに天板が厚い机は「厚い机」といえない理由は、机の機能との関連性の有無である。

特定化の傾向 机の垂直方向の寸法は特に「高い」、「低い」という観点から表現されるので、同じ量を「長い」、「短い」で形容することはない。

可動性 対象を容易に動かすことができる場合に、その対象の内在的なフレームによって用いられる。従って「高い本」は、現在の配置状況における重力方向に基づいて測られるのではなく、本の固有のフレームによって、垂直方向が決められる。

逆に、容易に動かし得る対象がたまたま置かれている状況における重力方向の寸法でもって「高い」、「低い」は判定されない。よってたとえ鉛筆が机の上にたっているような状況でもその長軸方向の距離は「高い」ではなく「長い」で形容される。

4.3 各形容詞に関する検討

紙幅の制約から、代表的なものについて簡単に述べる。

「高い」、低い」(形状) 典型的な配置をした場合の対象の下端から上端までの距離について言う。スキーマとして下から上への方向性を持つ (図3)。山のようなものであれば、常に重力の方向に一致。人間では、たとえ重力方向に対して別の配置をしていたとしても、直立した状態での下端から上端までの距離が対象となる。鼻など、垂直方向が要件ではない例がある。

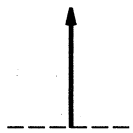


図 3: 「高い」に関するスキーマ

「深い」、「浅い」(形状) 典型的な配置をした場合の対象の上端から下端までの距離。日本語においては凹形をしたものの開口部から測る場合が一般的[3]。湖の場合は、地表面を基準面として、その対象の底までの距離を測る。雪の場合は、雪の表面から最下部 (この場合はこれが地表面) までの距離。押し入れは鉛直方向ではない例。開口面を基準として、その対象の突き当たりまでの距離を測る。これについては容器のスキーマ (図4) の利用が考えられる。

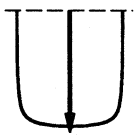


図 4: 容器のスキーマ

5 おわりに

本稿では、具象空間内に存在する対象の属性の形容に用いられる空間的な量をあらわす形容詞の意味の取り扱いについて、幾何的モデルとの対応をとった場合に、形容されている対象のどのような量と関係するかについて検討を行なった。対象の単純な幾何的属性のみから出発して、実際の言語の用法を見てみることによって必要となる対象知識を分析した。具体的には

1. ある形容詞がどのような名詞を修飾するか、
2. どんな量を修飾するか、
3. それは幾何的モデルにおいて何に対応しているか

ということが問題となった。現在、この検討をもとにして形容詞-名詞対の処理のための対象知識の体系化と計算機上でのプロトタイプシステムの作成を行ないつつある。また本稿では述べていないが、実際の量の判定や可視化の際の指示通りの変形については、我々が従来から用いているポテンシャルモデルを用いて定式化する予定である。

[3] 英語の deep には、「奥行き」を表す用法 (the desk is 2 meters deep) があり、必ずしも凹形を要求しない。

参考文献

[1] G. Retz-Schmidt. Various views on spatial prepositions. *AI Magazine*, Vol. 9, No. 2, pp. 95–105, 1988.

[2] A. Yamada, T. Yamamoto, H. Ikeda, T. Nishida, and S. Doshita. Reconstructing spatial image from natural language texts. In *Proc. COLING-92*, pp. 1279–1283, 1992.

[3] 山田, 網谷, 星野, 西田, 堂下. 自然言語における空間描写の解析と情景の再構成. 情報処理学会誌, Vol. 31, No. 5, pp. 660–672, 1990.

[4] 岡田. 自然言語および図形理解のための属性概念の分析. 情報処理学会論文誌, Vol. 26, No. 3, pp. 497–504, 1985.

[5] 岡田. 自然言語および図形理解のための属性概念の分類. 情報処理学会論文誌, Vol. 26, No. 1, pp. 25–31, 1985.

[6] 情報処理振興事業協会技術センター. 計算機用日本語基本形容詞辞書 IPAL (Basic Adjectives). 1990.

[7] 内海, 堀, 大須賀. 自然言語処理のための形容詞の意味表現. 人工知能学会誌, Vol. 8, No. 2, pp. 192–200, 1993.

[8] 内海, 堀, 大須賀. 比喩を含む言語表現の理解. 人工知能学会誌, Vol. 8, No. 2, pp. 201–211, 1993.

[9] 飛田, 浅田. 現代形容詞用法辞典. 東京堂出版, 1993.

授受表現による視点を利用した日本語の複文の意味解析に関する一考察

西澤 信一郎　　中川 裕志
横浜国立大学工学部　　横浜国立大学工学部

1 はじめに

本稿では，接続助詞「ので」による日本語の順接の複文について，節間のゼロ代名詞照応の問題を取り上げ，意味役割や語用論的役割などを用いた意味解析の立場から論じる．この問題を扱うためには，複文における節間の意味的なつながりを把握する必要がある．ここでは，「主節により記述される内容を結果とすると，その原因は必ず存在し，『ので』による従属節があるのならばそこにその原因を求めることができる」と捉えることにする．その上で，**動機保持者による主観的な因果性**[1]あるいは**複文における視点の一貫性**[2]という語用論的制約を用い，複文の意味解析が可能であることを示す．また，この結果を，文献[3]にあるような意味解析システムに応用することによって，従属節が1)主観述語，2)「〜がる」，3)受動態，4)授受補助動詞を用いた表現，であり，主節がa)意志的動作を表す動詞，b) a)+授受補助動詞による表現，であるような，接続助詞「ので」による順接複文の解析が計算機上で可能であることを述べる．

2 視点を含む順接の複文

2.1 視点

視点は，文献[4]などで詳しく述べられているとおり，授受(補助)動詞や主観表現，受動態などによって文中に頻繁に導入される要素である．本稿では，主に授受補助動詞によって文中に導入される視点について扱う．これは，文献[4]にしたがって次のように定義する．

定義 1 (視点) 視点は，その文で述べられている事象を話し手が記述する際に，自己同一視化の対象とする人物を指し示す語用論的役割である．

また，授受補助動詞では，授受動作による利益を受ける人物が導入されるので，これを受益者という語用論的役割により扱う．本稿では，授受補助動詞の代表的な語として「やる(あげる)」「くれる」「もらう」を対象とするが，これらの語における視点や受益者は次のように設定される．

制約 1 (視点および受益者) 授受補助動詞「やる(あげる)」「くれる」「もらう」では，視点及び受益者の指示対象が以下のように決定される．

「やる(あげる)」の場合 視点は主語を指示対象とし，受益者は一般に「〜(のため)に」に続いて記述される目的語を指示対象とする．

「くれる」の場合 受益者は一般に「〜(のため)に」に続いて記述される目的語を指示対象とし，視点も受益者と同一の対象を指示する．

「もらう」の場合 受益者は主語を指示対象とし，視点も受益者と同一の対象を指示する．

さらに，「もらう」では**依頼者**という語用論的役割が次のように設定される[5]．

定義 2 (依頼者) 授受補助動詞「もらう」が「A が B に 〜てもらう」のように用いられた場合，「A が B に動作の依頼をし，その依頼が聞き入れられたことにより，A は B から利益を受け取る」という意味が生じる．この A を**依頼者**という語用論的役割で表す．

視点は，一文中において一貫していることが要求される．文献[4]では特に断わっていないが，これは本稿で扱うような複文においても要求される，非常に強い制約となる．

制約 2 (複文における視点の一貫性) 従属節と主節それぞれの視点は一貫していなければならない．

なお，授受補助動詞以外の表現でも次の制約のように視点が導入されることがある．

制約 3 (授受表現以外での視点) 主観形容詞や主観動詞による主観表現では，経験者(主語)が視点の対象となり，「〜がる」という表現では観察者もしくは経験者(主語)のいずれかが視点の対象となる．また，受動態では，その被影響者(主語)が視点の対象となる．

2.2 複文中での視点の効果

本稿では，複文の節間の関係を，「主節から従属節へ対して，主観的動機もしくは一貫する視点を要求する」としている．そこで，主節の記述形式を，授受補助動詞が含まれない場合とそれが含まれる場合とにわけ，文の解釈に視点が及ぼす効果を述べる．

主節に授受補助動詞が含まれない場合

この場合，主節は従属節に主観的動機を要求する．これは，**動機保持者**という語用論的役割によって表される[1]．

定義 3 (動機保持者) 順接複文における動機保持者とは，従属節で記述される状況によって，主節中で記述される何らかの動作もしくは状態を引き起こすに十分の動機を持つ人物を指す．

従属節中では，動機保持者は従属節の記述形式によって**表 1** に示すような[1]参照関係をとる[1, 2]．
例えば，次のような文の解釈を考える．

[1]以降で，従属節もしくは主節で設定される各種役割を記述するために，"役割名[設定された節]"という表記を用いる．例えば，"経験者[従属節]"は，従属節中の経験者を示す．

表 1: 従属節中での動機保持者の参照関係

記述形式	動機保持者になり得る役割
授受補助動詞	受益者 [従属節]
受動態	被影響者 [従属節]
～がる	観察者 [従属節]
主観形容詞	経験者 [従属節]
主観動詞	観察者 [従属節]

(*1*) a. 外出を認めてくれたので，散歩に出掛けた．

b. 受益者 [従属節] = 動機保持者
受益者 [従属節] = 視点 [従属節]

c. 受益者 [従属節] = 動機保持者 = 動作主 [主節]

従属節に「くれる」があるので，(*1*b) のように動機保持者及び視点が従属節中に設定される．しかし，主節には授受補助動詞がないので，主節からは動機保持者が要求され，「散歩に出掛けたのは，外出を認めてもらった人」という，(*1*c) のような解釈となる．

主節に授受補助動詞が含まれる場合

この場合は，主節に設定される視点が，**制約 2** に従って，従属節に視点を要求する．そのため，従属節中にも授受補助動詞がある場合には，それによって設定される視点と主節の視点とが一致するような解釈が得られることになる．一方，従属節中に授受補助動詞がない場合には，それによる視点が存在しない．この場合には，**制約 3** に従って従属節での視点が決定し，これと，主節の視点とが一致するような解釈が得られると考えられる．

さらに，以下のような現象が観察される．

1. 主節の補助動詞が「やる (あげる)」のときには，補助動詞によって設定される受益者の指示対象が複文中 (従属節中) に存在しない場合があり，そのような文は解釈が困難になる．

2. 補助動詞が「くれる」では，受益者と視点とが一致するため，1. のようなことはない．しかし，「くれる」という受身的な表現では動機という概念が生じにくい．そのため，動機保持者による節間の関係がつくれず，3. の「もらう」に比べ，文が不自然になる場合がある．

3. 補助動詞が「もらう」の場合，2. と同じように，受益者と視点が一致するため，1. のようなことはない．
また，「もらう」では，**定義 2** のような依頼者が「依頼をするための動機」を持つと考えられる．
そこで，「主節での依頼者 (これは受益者と同じ対象を指す) が，その動機を従属節に求める」ように動機保持者による解釈が可能となる．
この解釈は，視点の一貫性と対立しないため，従属節と主節とを結ぶ意味的なつながりが二通り得られることになり，それによって，2. よりも解釈の容認度が高くなるものと思われる．

例文を挙げてみる．

(*2*) a. 疲れたので，仕事を手伝ってもらった．

b. 経験者 [従属節] = 動機保持者
経験者 [従属節] = 視点 [従属節]

c. (主節の主語 =) 受益者 [主節] = 視点 [主節]

d. 視点 [従属節] = 視点 [主節]
経験者 [従属節] = 動機保持者 = 依頼者 [主節]

主節の「もらう」によって視点および依頼者が主節中に設定され，従属節に対してそれぞれ視点及び動機保持者を要求する．その結果，(*2*d) の解釈となり，視点及び動機保持者それぞれによる関係が対立しないので，非常に容認しやすい解釈となる．

3 制約変換による解析システム

「ので」による順接の複文の意味理解システムとしては，文献 [3] にあるような，制約論理プログラミングの手法を用いたシステムがある．2.2で述べたような各役割の関係は，各役割を変数とした時に，その変数の持つ制約とみなすことができるため，本稿で述べているような複文の意味解析も，基本的には文献 [3] と同様の枠組で扱うことができると考えられる．

ここでは，上記のシステムで視点の関係をどのように扱うかについて述べ，さらに (*1*) を例にとり，解析の手順についても述べる．

なお，本稿で述べるシステムは，Prolog 上で動作するものであり，文献 [3] でのシステムとほぼ同じ構成である．ただし，動詞及び形容詞の意味辞書については，IPAL を参考にした[2]．

3.1 視点の制約

2.2で述べたような複文を解析する際には，概ね以下のように制約を用いればよい．

1. 主節に視点が設定されていない場合には，従属節側で**表 1** に示した関係を用い，さらに動機保持者と主節側の動作主とを関係づける制約を用いる．

2. 主節に視点が設定されている場合には，**制約 1** もしくは**制約 3** に示す視点と主節の視点とを一致させる制約を用いる．

このように，節間の制約として，動機保持者に関する制約と視点の一貫性に関する制約の二通りが考えられること，また，本稿では触れていないが，この二つの制約は独立でかつ視点の一貫性に関する制約の方が高い優先度を持つこと [2]，などがわかっており，

[2]IPA の「計算機用日本語基本動詞辞書」および「計算機用日本語基本形容詞辞書」の情報を基に，一部変更した辞書を作成し，本システムにおける意味辞書として用いた．

文献[3]のように，節間の制約を接続助詞「ので」の辞書項目として記述してしまうと無理を生じると考えられる．

そこで，複文に関する制約のうち，**表1**，**制約3**にある，従属節内で局所的に決定する制約については，接続助詞「ので」の辞書項目として記述し，動機保持者もしくは視点により節間を関係付ける制約については，該当する文法規則により呼び出される制約として記述することにする．つまり，節間を関係付ける制約を適用するタイミングを意図的に制御することになり，視点の一貫性に関する制約を優先的に適用し，さらにその結果と対立しない範囲で動機保持者に関する制約も適用する，という処理を行なうことが可能となると考えられる．

「ので」の意味辞書

以上のように，従属節内での動機保持者や視点に関する関係は，「ので」の語彙項目となる．そこで，「ので」の意味辞書を図1のようにする．辞書項目中の近接素性の値が，従属節の内容を示す素性構造となる．この値は，従属節の記述形式により異なるので，Prologの述語 constraint_NoDe/3 により制約として記述する．

```
のだ(助動詞,F,ダ列タ系連用テ形,
 [F=[主辞:[品詞:助動詞,
           接続関係:順接,
           修飾関係:従属節,
           節情報:Clause],
     見出し語:ので,
     近接:Adj,
     意味:Sem],
  constraint_NoDe(Adj,Clause,Sem)]).
```

図1: 接続助詞「ので」の意味辞書

```
constraint_NoDe(
  [[主辞:Clause*[品詞:Pos,
                 態:能動],
    意味:[意味主辞:Sem*[事象:[soa:[受益者:Motiv,
                                視点:PoV]]]]],
  Clause,
  [意味主辞:Sem*[接続:ので,修飾関係:従属節,
                 接続関係:順接,
                 語用論的役割:[動機保持者:(+),
                               視点:(+)],
                 動機保持者:Motiv,視点:PoV],
   意味修飾辞:[]]) :-
member(Pos,[補助動詞1a,補助動詞1b,補助動詞2b]).
```

図2: 従属節内の制約記述の例

図2は，従属節に授受補助動詞がある場合の制約である．第一引数が従属節となる事象の意味素性を表す．また，第三引数が，この制約を満足する場合の従属節の意味素性の値である．変数 Motiv および PoV の参照関係により，“受益者$^{[従属節]}$ = 動機保持者”という制約および授受補助動詞による視点が従属節の視点となることが示されている．

文法規則からの制約呼び出し

さらに，節間を結びつける制約は，文法規則より補強項を通じて呼び出される制約として記述する．例えば，主節に授受補助動詞がない場合は，“動機保持者 = 動作主$^{[主節]}$”という制約が用いられるが，これは，Prologの述語として図3のように記述される．

```
constraint_Sub_to_Main(_,
 [品詞:動詞,態:能動],
 [意味主辞:[語用論的役割:[動機保持者:(+),視点:(+)],
           動機保持者:Motiv]],
 [事象:[soa:[動作主:Motiv]]]).
```

図3: 節間の制約の例

第二引数が主節の形式であり，能動態の動詞であることを示す．また，第三引数が従属節の意味素性の値であり，第四引数が主節の意味素性の値である．なお，第一引数は現段階では使用していない．

ここでは，変数 Motiv による参照関係によって，主節の動作主が従属節の動機保持者と一致する，という制約が記述されている．

3.2 複文の解析例

ここでは，例文(*1*)「外出を認めてくれたので，散歩に出掛けた」を例にとり，本システムにおける解析がどのように行なわれるのかについて述べる．なお，最終的な解析結果は図4のような素性構造となり，実際のシステムでもこれに相当する出力を得ることができる．

従属節部分の解析

まず「外出を認めてくれたので」という従属節部分が解析される．このとき，従属節内の制約として図2に示した制約が用いられる．その結果，図5に示す素性構造が従属節の意味素性の意味主辞の値として求まる．

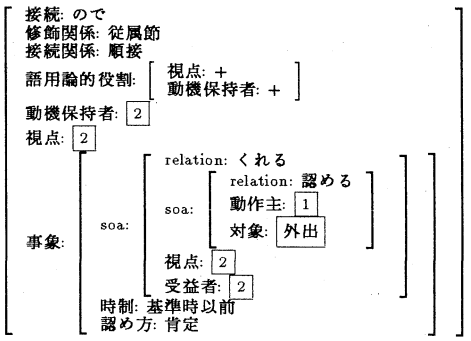


図5: 従属節の意味主辞素性

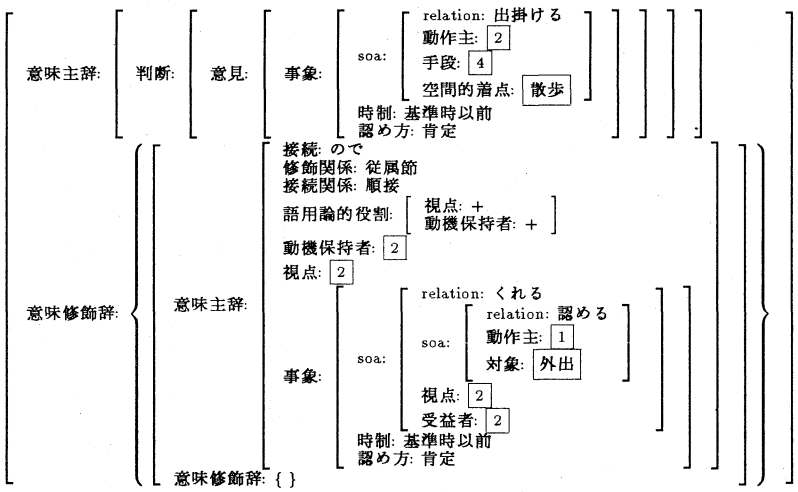


図 4: 単一化による複文の素性構造

主節及び複文全体の解析

次に，主節部分「散歩に出掛けた」が解析され，図 6 の意味素性値を得る．そして，素性構造の単一化により，図 4 の素性構造が得られる．この時，図 2 に示した制約が用いられ，これを満足するように素性構造が変換されている．この結果，図 4 のタグ 2 に表されるように，“受益者[従属節] = 動機保持者 = 動作主[主節]” という関係が求まる[3]．

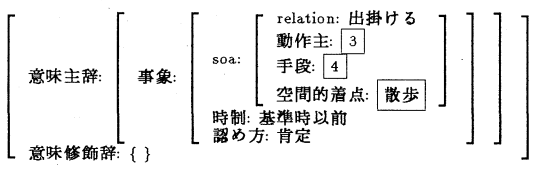


図 6: 主節の意味素性

4 おわりに

本稿では，「ので」による順接複文の意味的な解釈に対して視点が与える効果について述べた．また，その結果に制約論理プログラミングの手法を応用し，複文の意味理解システムを計算機上に試作した．

[3]図 4 において，タグ 1 は「認める」の動作主の意味素性を参照し，タグ 4 は「出掛ける」の手段の意味素性を参照する．また，散歩 および 外出 はそれぞれ「散歩」「外出」という名詞句の意味素性を参照するタグである．

なお，本稿では視点を扱ったため，文献 [3] での議論と比較して，「話し手」もしくは実際の文脈で「話し手」となる人物，という要素を解析結果の中へ導入しやすいと考えられる．しかし，本稿ではあくまでも解析の対象を「文中における意味役割もしくは語用論的役割の間の関係の解析」に限っている．実際に，文脈への登場人物など実世界の要素と各役割との対応までをも扱おうとするならば，文脈情報やさらに詳しい語彙知識，いわゆる「常識」と呼ばれるような知識なども必要となると考えられ，今回の検討対象からは外れるが，さらに精密な意味解析を行なうためには，今後扱う必要がある問題だと思われる．

参考文献

[1] Hiroshi Nakagawa and Shin'ichiro Nishizawa. Semantics of Complex Sentences in Japanese. In *COLING-94*, pp. 679–685, 1994.

[2] 中川裕志. 動機と視点の係わり –「ので」「のに」で接続された日本語複文の場合 –. 認知科学 vol.2 no.2 に採録決定.

[3] 西沢信一郎, 中川裕志. 主観的動機に関する意味および語用論的制約を利用した日本語複文の理解システム〜「ので」「のに」による接続を中心として〜. 自然言語処理, Vol. 2 Number 1, pp. 19–38, jan 1995.

[4] 久野 暲. 談話の文法. 大修館書店, 東京, 1978.

[5] 仁田義雄. ヴォイス的表現と自己制御性. 仁田義雄(編), 日本語のヴォイスと他動性, pp. 31–57. くろしお出版, 1991.

形態素で残る曖昧性を考慮した日本語文の係り受け解析

秋山 典丈 藤井 敦 徳永 健伸 田中 穂積
東京工業大学 工学部

1 はじめに

日本語の係り受け解析をする上で問題となる曖昧性は二種類ある．第一は文節が持つ係り属性・受け属性が複数あり得るという曖昧性，第二は属性は同じだが係る場所が複数あるという曖昧性である．第一の一つの文節が複数の係り・受け属性を持つという曖昧性は，その文節を構成する形態素の持つ属性の曖昧性から生じている．例えば，文節を構成する格助詞「で」は直前の名詞と結合して道具格を構成する場合と，判定詞（あるいは助動詞）として直前の名詞と結合して名詞述語を構成する場合があるので，係り属性・受け属性が一意に決定できない．従来の係り受け解析の研究では，第二のいわゆる「係り受け」の場所に関する曖昧性は考慮しているが，第一の文節の属性の曖昧性については未考慮だったり，考慮はしていても二つの曖昧性を区別していなかったりするため，その後の処理で正しい結果が得られないことがあった．

本研究ではこの文節の持つ複数の属性に着目し，係り属性・受け属性の組で決まる「文節の種類」という文節の属性を定義し，係り受けの関係を文節間ではなく「文節の種類」間の関係として捕える方法を提案する．それにより文節が複数の係り属性・受け属性を持つことを，一つの文節に複数の「文節の種類」を割り当てることによって表現できるので，従来考慮されていなかった，形態素解析でも残る係り受け解析時の曖昧性を処理することが可能になる．

また，構文的制約を二段階で適用することにより前述の二つの曖昧性を解消する方法を考案した．最初の段階で，「文節の種類」の曖昧性を圧縮した状態で制約を適用し，係り受けの曖昧性を部分的に解消する．次の段階では，残った「文節の種類」の曖昧性を展開した後に，展開された個々の「文節の種類」の割り当てのセットに対し制約を適用する．二段階に分けることにより制約の適用の重複を避け，効率的な解析を実現している．そして構文的制約の有効性を実験により確認した．

本研究では以上のように形態素解析で解消不可能であった文節間の係り受けに関する二つの曖昧性を考慮した係り受け解析の新しい手法を提案する．

2 文節の種類

一つの文節には**複数**の「文節の種類」が割り当てられる可能性があり，その時，その文節は属性に関して曖昧であり，文全体として『いくつかの「文節の種類」の割り当てのセット』を持つことになる．この「文節の種類」は，文節の持つ係り・受け属性の違いによって分類しており，12種類ある．「文節の種類」間の係り受け可能性は表1で表され，これによって自動的に係り受けの候補が列挙できる．表で横(左)の「文節の種類」は縦(上)の「文節の種類」に係りうるとよむ．

表1: 「文節の種類」の係り・受けの接合表

係り＼受け	格	主題	用言	名詞述語	連用修飾	連用修飾名詞	連体修飾	連用接続	並列名詞	名詞接続	用言修飾	名詞述語修飾
格			○	△			△				○	△
主題			☆	☆							☆	☆
用言	○	○	●	●		○	○		○		●	●
名詞述語	○	○	●	●		○	○		○		●	●
連用修飾			○	○							○	○
連用修飾名詞			○	○							○	○
連体修飾	○	○		○		○			○			○
連用接続			○	○							○	○
並列名詞	●	●		●			●					●
名詞接続	○	○		○			○					○
用言修飾			○	○							○	○
名詞述語修飾			○	○							○	○

表1中で○は通常の係り受け，●は並列の係り受け，△は通常の係り受けだが純粋な名詞が受けて（用言の基本連用形・副詞でなく），格が係る場合，格は「ガ・ハ・モ・ニ・デ・ヨリ」に限られ，☆は通常の係り受けだが，連体形にしか係らないことを意味する．

この「文節の種類」は基本的には以下に示す形態素によって分類できる．

格: 名詞・動詞の連用形 + 助詞 (ハ, ガ, ヲ,...) など．

主題: 名詞・動詞の連用形 + 助詞 (ハ・モ)．主題は非交差の制約が適用不可能で，用言の連体形には係らないことを除くと格と同じである．名詞が時相名詞・形式名詞の場合その文節は主題に決定される．

用言: 動詞・形容詞・形容動詞．次に述べる名詞述語以外の用言．

名詞述語: 名詞 + である，だ，な，など．判定詞[益岡92]で終わる述語．連体修飾を受けることが用言と区別した理由である．

連用修飾: いわゆる副詞（連体修飾を受けない）．

連用修飾名詞: 副詞的名詞 (連体修飾を受ける副詞), 時相名詞, 時相名詞 + から, 時相名詞・形式名詞 + に. 「結果, 場合, 今日, 七月,...」など.

連体修飾: 名詞 + 連体助詞 (ノ), 連体詞など.

連用接続: 文接続の接続詞. 「また, そして, あるいは」など. この文節の直前には用言・名詞述語が来る.

並列名詞: 名詞 + 並列助詞 (ト, ヤ,...)・、(読点) など.

名詞接続: 単語接続の接続詞. 「また・そして・あるいは」など. この文節の直前には体言が来る.

用言修飾: 『用言』 + 助詞など (連用修飾 (副詞) 節). 「降っているので」など. 助詞は, 「つつ・まま・ながら・ように・けれども」などがある. 他に以下の場合も用言修飾に分類する.

(i.) 「動詞の基本連用形 + 助詞 (ニ)」(目的を表す.) 「ボールを蹴りに公園に行った.」

(ii.) 「『機・機会・教訓・契機・皮切り・最後』等の名詞 + 助詞 (ニ)」

(iii.) 「『段々・次々』等の名詞 + 助詞 (ト)」

名詞述語修飾: 『名詞述語』 + 助詞など (連用修飾 (副詞) 節). 「大降りなので」など. 名詞述語に続く助詞は用言の連用修飾に同じ.

一方, 形態素として曖昧性がなく一つに決まっているが, これに複数の「文節の種類」を割り当てなければならないことがある (表 2). このような文節については複数の「文節の種類」を割り当てる.

表 2: 複数の「文節の種類」を持つ文節

形態素	「文節の種類」
サ変名詞 + 、(読点)	並列名詞 / 用言
名詞 + ニ (助詞)	格 / 並列名詞
名詞 + デ (助詞)	格 / 名詞述語
名詞 + ト (助詞)	格 / 並列名詞 / 連用修飾名詞
名詞 + ノ (助詞)	連体修飾 / 格
接続詞かつ並列助詞	名詞接続 / 連用接続
スグ (副詞)	連用修飾 / 連体修飾
名詞 + カラ (助詞)	格 / 連体修飾
名詞 + 、(読点)	並列名詞 / 名詞述語

3 構文的制約

本稿では形態素のみで得られる情報を用いて係り受けに関する構文的制約を考えた. この制約は以下の性質を持たねばならない.

- 適切なものを確実に適用し,(「文節の種類」を使用することにより文節の属性の違いを考慮にいれているので, 係り受け関係にスコア付け [山下 93] をする必要がなく, 係るか係らないか一意に決定できる.)
- 意味を考慮しないと解決できない係り受けの曖昧性は残し, それ以外については可能な限り構文的制約で曖昧性を解消する.

この制約の評価は 5 節で行う. 制約は効率性を考えて, 複数の「文節の種類」を持つ文節における曖昧性を展開する前と後で二回適用する. 初回の制約で消去された「文節の種類」については曖昧性を展開する必要がなくなる.

例えば「名詞の並列は格になっている文節や読点の後ろに係ることはない.」という制約を考える. 「名詞 + ト」と「名詞 + ノ」の文節を考える. 後者の「名詞 + ノ」の文節は連体修飾と格という二つの「文節の種類」を持つことがある. この文節がどちらの種類に属するかが分からない場合は, 上記の制約を適用できない. したがって後者の文節はなかったものとして後続する文節に制約を適用する.

以下の表 3 が制約を表している. 4 節で実際に解析例を示し, その際に適用する制約についてはそこで説明する. 展開に無関係の制約は, 初回の制約では厳密に適用できないので展開した後で再び厳密に適用する.

表 3: 制約の種類

展開前に適用する制約	展開に無関係の制約
ガについて, 読点のない副詞, 括弧に対する非交差, 連体修飾句の制限, 従属節について	名詞の並列, 二重格について, 並列でない連用修飾句, 主題について, 読点無しの格について, 連用修飾句の制限, 用言の連用形, 非交差

4 解析の流れ

実際の解析は, 以下の順で進む.

(i.) 形態素列から文節を生成.

(ii.) 文節に「文節の種類」を付与.

(iii.) 可能な係り受けの候補を列挙.

(iv.) 構文的制約の適用.

(v.) 「文節の種類」の曖昧性の展開.

(vi.) 再び構文的制約の適用.

なお, 文節の生成については省略する.

以下の文を例にとって説明する.

「体力の尽きるまでインドを旅行、見聞を広めた。」

4.1 「文節の種類」の付与

例文から文節を生成し, 文節に「文節の種類」を割り当てた状態が図 1 である.

4.2 可能な係り受けの候補を列挙

表 1 に従い, 可能な係り受けの候補を列挙する (図 2).

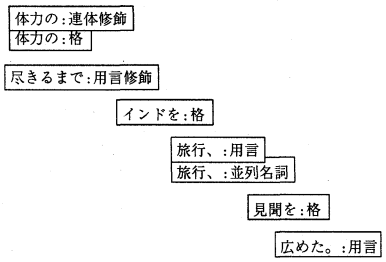


図 1: 「文節の種類」の付与

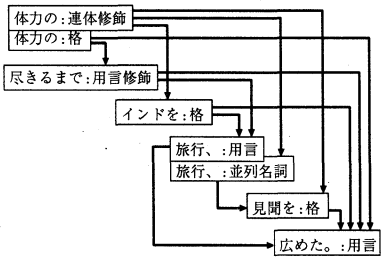


図 2: 係り受けの列挙

4.3 構文的制約の適用

次に最初の構文的制約を適用する. 図 2 に制約をかけた図が図 3 である.

- “体力の:連体修飾” から “旅行:並列名詞”, “見聞を:格” への係りは構文的制約「連体修飾句の制限」により消去,
- “体力の:格” から “広めた:用言” の係りは制約「ガについて」により消去,
- “インドを:格” から “広めた:用言” の係りは制約「二重格について」により消去.

また, この段階で “インドを:格” は “旅行:用言” にしか係らず, “旅行:並列名詞” は存在不可能な「文節の種類」であることが分かるので消去する. この例で使用した制約について説明する.

連体修飾句の制限 以下の二つの制約がある.

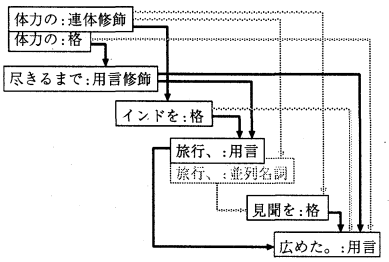


図 3: 初回の制約後

- 読点を伴わない連体修飾句は, 読点をこえては係らない.
- 読点を伴う連体修飾句は直後の係りうるものには係らない.

ガについて 読点を伴わない「名詞 + ガ」のガ格の文節は, すぐ次に係りうるものに係る.

二重格について 読点を伴わない「ガ」以外の格をみつけ, 間に動詞をはさまずにくる次の格が同じ格の場合は, その次の格はその次の動詞にだけ係る. 読点を伴う場合は並列の可能性があるのでこの制約は適用しない.

4.4 「文節の種類」の曖昧性の展開

図 3 を「文節の種類」の曖昧性で展開すると二つに分解される (図 4).

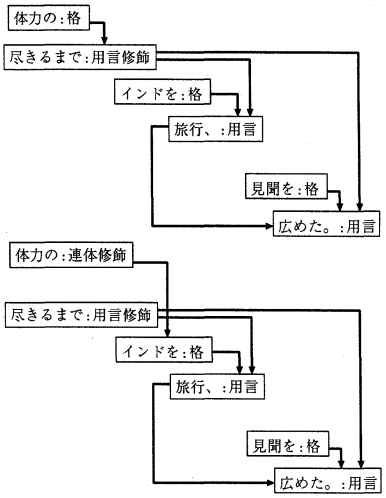


図 4: 曖昧性の展開後

4.5 再び構文的制約の適用

図 4 に制約をかけた図が図 5 である. 上では制約「無読点の格・連用修飾」を適用, 下では “体力の:連体修飾” が “インドを:格” にしか係らないことから制約「非交差」を適用している. 結局, 下の「文節の種類」のセットは存在しないことになり, 上のセットが正しいことが分かる. 以下に適用した制約の説明をする.

無読点の格・連用修飾 読点を伴わない格・連用修飾は, その後に用言と読点とがあった場合にそれらより後ろには係らない.

非交差 ある文節の係る文節が一つしかない場合, 主題を表す文節を除いて, それと交差する係り受けを排除する.

従来の日本語の係り受け解析に関する研究は非交差の原則を使用しているものがほとんどである. しか

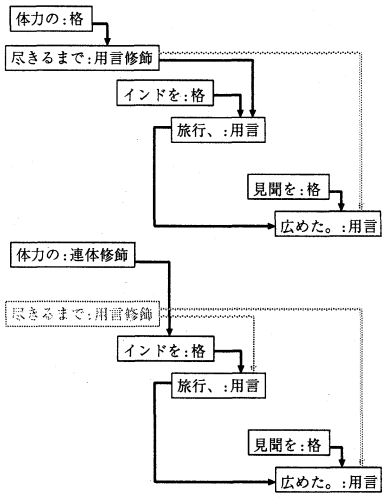


図 5: 二回目の制約後

し, 日本語には厳密には非交差の原則は適用できない. [児玉 87] . 本研究では,

- 「私は母に学校へ行くと言った.」

(「私は」(主題)「行く」と「母に」「言った.」が交差している.) このような文を考え, 主題については非交差の原則は適用しない.

5 評価

本システムを計算機上に実装して実験をした.

新聞記事・教科書からのコーパスを使い, 50 ～ 140 字からなる文 50 文を係り受け解析した. 表 4 は解析した文についての概要である. 表 5 は文と文節数と係り受け数のそれぞれの段階ごとの推移である.

表 4: 解析した文

解析した文	50
失敗した文	0
文字数/文	63.7
文節数/文	14.9
実行時間/文	3 秒

表 5: 制約の有効性

	セット数	文節の種類数	係り受け数
初期状態 (4.2節)	685	12127	49491
初回の制約後 (4.3節)	655	10343	22596
二回目の制約後 (4.5節)	585	9170	15555

50 文に対して, 最終的に 585 の係り受け解析結果を得ている (表 5). このことは 1 文当たり 12 弱の係り受けの曖昧性が残されていることを意味している. この残された曖昧性の数は意味解析にとって, さほど大きな負担となる曖昧性の数ではない. また, 初回の制約で 49491 の係り受けが 22596, すなわち約 54 %が消去され, 展開後でも 22596 の係り受けが 15555, つまり約 31 %が消去されている. これらの結果から細かい「文節の種類」の導入と構文的制約が十分有効であるといえる.

解析して残った曖昧性を人間が分析したところ, 曖昧な係り受けが残されているのは並列構造や文節「名詞 + の」の連続した部分であり, これらの曖昧性解消については語と語の類似度等を用いた意味解析 ([黒橋 92] など) が必要となるもので, 人間でも構文的な情報だけでは曖昧性の解消が不可能であることが分かった. また, 表 6 は各制約がどのくらい適用されているかを示した表である. 表 6 の総計の値 1779 と, 表 5 の 49491 − 22596 = 26895 が一致しないのは, 前者は圧縮された状態であるのに対し, 後者は展開された状態だからである. 「連体修飾」や「無読点の格・連用修飾」の制約が多く適用されている.

表 6: 各々の制約の有効性

制約	展開前	展開後
格充足	60	-
括弧非交差	55	-
ガ	181	-
ハ	9	-
連体修飾	781	-
トとニ	0	-
主題	24	-
格	69	-
従属節	1	-
文節の消去	4	-
連用修飾	146	6
用言連用形	16	401
無読点	301	1965
動詞の連続	32	24
名詞並列	8	338
非交差	92	2735
総計	1779	5469

6 おわりに

本稿では細かい文節の種類を導入し, 従来考慮されていなかった形態素解析で残る係り属性・受け属性の曖昧性を考慮した文節間の係り受けの解析手法, および構文的制約を提案し, その有効性を確認した.

今後は意味解析により, さらに係り受けに関する曖昧性を解消し, また形態素解析と意味解析を係り受け解析を通して統合することにより, 効率的・精密な処理を行うことを目指す.

参考文献

[益岡 92] 益岡隆志, 田窪行則. 基礎日本語文法—改訂版—. くろしお出版, 1992.

[黒橋 92] 黒橋禎夫, 長尾眞. 長い日本語文における並列構造の推定. 情報処理学会論文誌, Vol. 33, No. 8, pp. 1022–1031, 1992.

[山下 93] 山下晃司, 安原宏. 形態素情報による日本語の係り受け解析. 情報処理学会 自然言語処理研究会, 11 1993.

[児玉 87] 児玉徳美. 語順の普遍性. 山口書店, 1987.

「の」による名詞化構文の認知言語学的考察

堀江　薫

東北大学大学院国際文化研究科異文化間教育論講座
980-77 仙台市青葉区川内
e-mail: khorie@intcul.tohoku.ac.jp

Abstract: This paper attempts a semantic and cognitive characterization of syntactic constructions involving nominalizer *no*('の') in Modern Japanese, as contrasted with other competing nominalization strategies, in view of the notions of Grammaticalization, Markedness and Cognitive Complexity.

1．はじめに

日本語の助詞「の」は、節を名詞に転換する「名詞化辞」としての機能を持ち、多様な構文で使用される。現在まで、名詞化辞「の」の生起する個々の構文の統語的・意味的特徴を解明しようとする各論的研究は生産的に行われてきた（Kuno 1973, レー・バン・クー 1988, Kuroda 1992, 他多数）。その反面、名詞化辞としての「の」の生起する様々な構文に共通する特徴を明らかにしようという方向性を持った研究はそれほど多く見られなかった（Tsubomoto 1981, 国広1992, Horie 1993ab, 1994, 佐治1993, 成田1994などを除く）。

本論文では、まず「の」による名詞化構文を概観し、現代日本語において、それらの構文が、他の名詞化構文とどのような競合関係にあるかを考察する。最後に「の」による名詞化構文に共通する意味的・認知的な特徴は何か、また「の」による名詞化構文が他の名詞化構文と比較して広範囲で使用されるのは何故かという問題を、「文法化」、「有標性」および「認知的複雑さ」という概念を用いて考察する。

2．「の」による名詞化構文と他の名詞化構文の競合関係

名詞化辞「の」が関与している構文には下記のようなものがある。

(a) 補文：[あの人が今日来る]のを知ってた？
(b) 主要部内在型関係節：[くりが落ちている]のを拾って食べた。
(c) 分裂文：[私が生まれた]のは、満州です。
(d) 「ので」・「のに」節：[あの人が誘った]ので／のに、太郎は来なかった。
(e) 「のだ」文：地面がぬれている。[雨が降った]のだ。

現代日本語において、文を名詞節に変換する際に用いられる名詞化辞には、「の」の他に、国語学で形式名詞とよばれる「こと」・「ところ」の2つがある。以下にそれぞれの補文化辞としての用例を示す。

（1）[山田が泣いていた]こと を思い出した。
（2）[あの男が入っていく]ところ を目撃した。

この他に助詞に述語の連体形が直接接続するという「直接名詞化」（'direct nominalization'; Martin 1975の用語）が行われることもある。

（3）[雨が降っていた]（Ø） にもかかわらず、たくさんの人が集まった。

この名詞化は、「の」や「こと」のような明示的な名詞化のマーカーを欠いているところから、本論文では「ゼロ名詞化」と呼び、関与する名詞化辞を「ゼロ」と呼ぶことにする。例文では、ゼロ名詞化辞を、上記（3）のように「Ø」で示す。

ここで注目すべきことは、「の」以外の名詞化辞は、いずれも、「の」とは生起する環境を共有し競合関係にあるが、お互い同士（例えば「こと」と「ゼロ」）は相補分布をなし、競合関係にないということである。

（1'）[山田が泣いていた]こと／の を思い出した。（「ところ」・「ゼロ」不可）
（2'）[あの男が入っていく]ところ／の を目撃した。（「こと」・「ゼロ」不可）
（3'）[雨が降っていた]（Ø）／の にもかかわらず．．．(「こと」・「ところ」不可)

3．「の」による名詞化構文の意味的・認知的特徴

前節で述べたような名詞化辞「の」の広範囲にわたる分布は、「の」それ自体の固有の性質と切り離せないものと考えられる。では、その固有の性質とは何であろうか。以下では、他の名詞化辞「こと」・「ところ」・「ゼロ」との比較で、「の」による名詞化構文の意味的・認知的な特徴を明らかにしていくことにする。まず、前節の（1'）ー（3'）の例文をもう一度検討してみよう。

まず最初に（1'）と（2'）における「こと」と「ところ」の使用条件を検討する。名詞化辞「こと」と「ところ」に共通しているのは、両者とも、それぞれに『事実・命題』、『場所』という語彙的な意味を有する普通名詞が、通時的に、『名詞化』という文法的な機能（あるいは意味）を担うようになった、すなわち「文法化」（'grammaticalization'; cf. Hopper and Traugott 1993;国語学でいう「形式化」に近い用語）した形式だということである。しかし「こと」と「ところ」は、現代日本語において完全に語彙的な意味を喪失したわけではなく、非常に抽象度の高い意味ではあるが、依然として語彙的な意味を保持している。この、残存している語彙的な意味は、「こと」と「ところ」の名詞化辞としての機能に一定の制約を与えていると考えられる。（1'）において「こと」が容認可能なのは、「こと」がもつ『事実』という語彙的な意味が、「思い出す」という述語の意味と整合するものだからである。一方（2'）において「ところ」が容認可能なのは、「ところ」がもつ、『時空間・状況』といった語彙的な意味が、「目撃する」という述語の意味と整合するものだからである。（1'）における「ところ」および（2'）における「こと」の使用が容認不可能であるのは、主節の述語との意味的な整合性に欠けているためで

ある。紙幅の関係で詳述できないが、「こと」および「ところ」は、補文以外にも、前節であげた「主要部内在型関係節」（「ところ」のみ）、「分裂文」といった構文でも使用されることがあるが、その場合も、あくまで主節の述語との間に意味的な整合性が成立することが不可欠な条件である。

では、「こと」と「ところ」という語彙項目から文法化した名詞化辞と比べて、「の」はどのような意味的特徴をもっているのだろうか。重要なことは、「の」という名詞化辞自体が、固有の語彙的意味を持たず、名詞化する節に対して『名詞化』という文法的な意味（または機能）のみを担っているということだと思われる。この、語彙的意味の欠如という「の」の（負の）意味的特性は、「こと」や「ところ」といった語彙的意味を有する名詞化辞に比べて、共起する述語との選択制限がより少ない、ということを保証するものである。「対立する言語的単位において複雑で一般的でない特徴をもつ方が有標、単純で一般的な特徴をもつ方が無標となる」(大塚・中島1987, p.692)という言語学の「有標性(markedness)」の概念を用いると、「の」は、「こと」や「ところ」と比べて、音節数の少なさという形態の面からも、また語彙的意味の欠如という意味の面からも、「無標の(unmarked)」項であるといえる。

では、前節の例文（3'）で競合していた「ゼロ」名詞化辞と「の」に関しては、有標性の観点からどのようなことがいえるだろうか。「有標性」の観点からすると、「ゼロ」名詞化辞は、無標のように一見思われる。何故なら、「ゼロ」名詞化辞は、「の」と同じく、「名詞化」という文法的な意味（または機能）以外に何らの語彙的な意味を担わないばかりでなく、「の」と異なり、明示的な言語形式として存在しないからである。しかしながら、実際には「ゼロ」名詞化辞は決してその使用範囲において「の」を凌駕することはなく、現代日本語における名詞化の手段としては決して無標とはいいがたいのである。

このことは、「有標性」を決定する際に、単に形態上の複雑さのみを基準にしては決められないということを示している。具体的には、Battistellaが主張するように、「（言語形式の実際の）分布（'distribution'）」や「生起の自由さ('freedom of occurrence')」という概念が、個別言語における有標性を決定する際に重要な役割を果たしているようである("Distribution within a language plays an important role in the determination of language-particular markedness values. Unmarked terms are distinguished from their marked counterparts by having a greater freedom of occurrence and a greater ability to combine with other linguistic elements; Battistella 1990, p.27)。分布および生起の自由さに関していうと、現代日本語において、「ゼロ」名詞化辞は、限られた数のイディオム的表現や少数の助動詞・接続助詞の前を除いては、殆ど用いられない。したがって、分布および生起の自由さを基準にすると、「ゼロ」が有標、「の」が無標ということになる。このことによって、有標性の決定において、「形態上の複雑さ」と「生起の自由さ」の両方の基準が競合した場合、後者が前者よりも優先される、ということがわかる。

ここまでの議論で、有標性の観点からみた場合、「の」は、競合する「こと」・「ところ」・「ゼロ」という３つの名詞化辞と比べた場合、無標な名詞化の手段だということがわかった。では、ある言語形式が、他の形式に比べて無標であるということは、認知的にどのようなことを意味するのであろうか。この点に関しては、「有標の範疇は、無標の範疇に比べて注意・心理的労力・処理時間の点で、認知的により複雑である傾向がある("The marked category tends to be cognitively more complex - in terms of attention, mental effort or processing time - than the unmarked

one.")」というGivónの「認知的複雑さ(Cognitive complexity)」の概念(Givón 1990, p.947)が有用であると思われる。「こと」と「ところ」に関していえば、残存している語彙的な意味が「の」に比べてより多くの「注意・心理的労力・処理時間」を要求するのであり、「ゼロ」に関していえば、生起頻度の低さが、「の」に比べて、同定する上でより多くの「注意・心理的労力・処理時間」を要求するのであろう。この認知的複雑さの低さこそが、「の」の使用上の制限の低さにつながり、他の名詞化辞よりも広範囲な環境での生起・使用を可能たらしめていると考えられる。

4．おわりに

本論文では、現代日本語の名詞化辞「の」が、「名詞化」という文法的な意味のみを担っており、競合する「こと」・「ところ」・「ゼロ」の３つの名詞化辞に比べて形態または生起の自由さという観点からみて無標であるため、認知的複雑さの度合いが低く、そのことが広範囲な環境における「の」の使用を可能にしていることを論じてきた。今後、古代日本語・韓国語との対照で、上記の主張をさらに一般性の高いものにすることを課題としている。

謝辞　本研究は平成７年度伊藤謝恩育英財団日本研究助成および文部省科学研究費補助金重点領域研究（２）(No. 06232211)による研究成果の一部である。

参考文献

Battistella, Edwin L. 1990. *Markedness*. New York: State University of New York Press.

Givón, Talmy. 1990. *Syntax* II. Amsterdam: John Benjamins.

Hopper, Paul. and Elizabeth Traugott. 1993. *Grammaticalization*. Cambridge UP.

Horie, Kaoru. 1993a. From zero to overt nominalizer *no*: a syntactic change in Japanese. *Japanese/Korean linguistics* 3. ed. by Soonja Choi. Stanford: CSLI. 305-21.

___.1993b. *A cross-linguistic study of perception/cognition verb complements: a cognitive perspective*. Doctoral dissertation, University of Southern California.

国広哲弥. 1992.「のだ」から「のに」・「ので」へ. カッケンブッシュ寛子他編. 日本語研究と日本語教育. 名古屋：名古屋大学出版会. 17-34.

Kuno, Susumu. 1973. *The structure of the Japanese language*. Cambridge, MA: MIT Press.

Kuroda, S.-Y. 1992. *Japanese syntax and semantics*. Dordrecht: Kluwer.

Martin, Samuel E. 1975. *A reference grammar of Japanese*. New Haven: Yale UP.

成田一. 1994. 連体修飾節の構造特性と言語処理. 田窪行則編. 日本語の名詞修飾表現. 東京：くろしお出版. 67-126.

大塚高信・中島文雄. 1987.新英語学辞典. 東京：研究社.

佐治圭三. 1993.「の」の本質. 日本語学. 10月号. 4-14.

Tsubomoto, Atsuro. 1981. It's all *no*. *CLS* 17. 393-403.

レー．バン．クー. 1988.「の」による埋め込みの構造と表現の機能．東京：くろしお出版.

格助詞「の」の分類と解析

横山 晶一　加藤 貴子　廣重 拓司
山形大学 工学部

1　はじめに

2つの名詞を格助詞「の」で結びつけた「AのB」という形の複合名詞は、日本語に非常によく用いられ、その表現も多様である [2]。これらを機械的に処理するためには、A、Bそれぞれの意味を調べて、その結合の仕方を分類する必要がある。本研究では、両方の名詞の意味素性に着目し、その組合せによって全体の意味（一般にはBで規定される）を分類した [1]。また、「BのA」という逆の形が成り立つかどうかにも注目して、その性質を分類した。これによって、「AのB」という形をした複合名詞を機械的に解析するための指針が得られた。なお、言語データとしては、「品詞列集成」 [5]、「語と語の関係解析資料」 [6] に出現する語を中心に調査した。

2　「AのB」の意味

「AのB」を用法によって分類すると、次のようになる。この分類は一部奥津 [3] を参考にしたものである。

(1) 属性

AとBの間に所有、所属、作成、部分と全体、場所、時間などの関係が成り立つものである。

例：「私の息子」（所有）、「学校の先生」（所属）

(2) 同格

AとBが身分などを表す場合には、「AのB」を「AであるB」、という形に言い換えることができる。

例：「主役の俳優」

(3) 動作

A、Bのどちらかがサ変動詞の場合には、それが自動詞か他動詞かによって、どちらかが主体または客体となり、全体として何らかの行為または動作を表す。この場合には、その動詞によって、他方の名詞の意味素性が規定される場合が多い。

例：「米の生産」、「実験の方法」

(4) 目的、対象

AがBの目的や対象になる。「Aを～するためのB」、「AについてのB」などと解釈できるが、曖昧性が生じやすい。

例：「哲学の講演」

(5) 慣用的表現

日常的に使われ、言い換えが困難な表現である。これについては本研究の対象外とする。

例：「噂の種」

3　意味素性による「AのB」の分類

「AのB」が前節のどの分類に属するかを機械的に判断するためには、名詞A、Bの各々についての意味素性を調べて、その組合せによって判断することが望ましい。そこで、名詞に関して次のような12の意味素性を導入する。この素性は、主としてIPAL基本動詞辞書 [4] の意味素性を参考にした。

表1：名詞の意味素性

素性名	略号	例
人間	HUM	先生、母
組織	ORG	学校、日本
物	OBJ	家、車
動物	ANI	犬、鳥
植物	PLA	花、木
生物の部分	PAR	頭、羽
場所	LOC	隣、公園
時間	TIM	去年、季節
数量	QUA	3人、5kg
性質	PRO	多様、長さ
動作	ACT	勉強、旅行
抽象名詞	ABS	態度、責任

表1に示すように、これらはごく一般的な意味素性として用いられているものである。「AのB」各々にこれらの意味素性をあてはめると、$12 \times 12 = 144$ の組合せになる。これらについて、意味的には中心となる名詞Bを基準にして、前節での分類を適用する。すべての結果について示すには紙数が足りないので、特に問題となる意味素性を含む一部の結果について以下に述べる。

3.1 Bが意味素性 HUM を持つ場合

Bが意味素性 HUM を持つ場合について、Aの意味素性、前節での分類、「BのA」(言えない場合には×。なお、これについては4節で述べる)、例とともに示す。

表2:「AのB」の分類(BがHUMの場合)

A	B	分類	BのA	例
HUM	HUM	所有	同格	私の息子
HUM		同格	所有	主役の俳優
ORG		所属	所有	学校の先生
OBJ		(Aが場所)	所有	山の男
		(目的)	所有	お茶の先生
ANI		同格	所有	犬のタロウ
PLA		所属	所有	お花の先生
PAR		Aが場所	一部	胸の赤ん坊
LOC		Aが場所	Bが場所	隣の人
TIM		Aが時間	Bが時間	去年の先生
QUA		Aが数量	Bが数量	12人の男
PRO		Aが性質	×	異色の先生
ACT		Aが動作	Bが動作	憧れの女性
ABS		状態	所有	ご機嫌の母

この表から明らかなように、名詞AがORGという素性を持つ場合には、所属になることが決定できる。同様に、ANI, PLAの場合も、全体の意味について、紛れなく決定することができる。

また、PAR以下では、全体の意味はBになるが、分類上は、名詞Aが重要になる。すなわち、Bの限定的な性質をAで示すことになる。たとえば、「12人の男」では、男の数量を前につけた「12人」で示しているし、「ご機嫌の母」では、母の状態を示している。

意味的な分類が困難なのは、名詞AがHUM, OBJの場合である。しかしながら、HUMの場合には、次のように、さらに細かい意味素性を与えることによって、ほとんどの場合には分類が可能である。

すなわち、

HUMのHUM:

- Bが身分、状態の性質、親族、人間関係 → 「所有」
- Aが身分、状態の性質、親族、人間関係 → 「同格」

AがOBJの場合には、Aが場所的な場合と、Aが目的を表す場合とがあるが、これは現状では意味素性を細かくしても区別することが困難である。

3.2 Bが意味素性 OBJ を持つ場合

前節でも少し触れたように、片方が意味素性OBJを持つ場合は、HUMの場合と違って問題が多い。表3に、名詞BがOBJの場合の、表2と同様の分類を示す。

表3:「AのB」の分類(BがOBJの場合)

A	B	分類	BのA	例
HUM	OBJ	(所有)	Aが場所	私の本
		(作成)	種類	正宗の名刀
		(目的)	Aが場所	障害者の施設
ORG		(所有)	×	学校の校旗
		(作成)	×	A社のシール
		(Aが場所)	×	日本の山
OBJ		一部	×	バナナの皮
OBJ		Aが場所	×	川の石
OBJ		材料	×	絹の布
OBJ		(種類)	種類	大根のみそ汁
		(目的)	×	電車の切符
ANI		種類	種類	犬の鎖
PLA		種類	Aが場所	木の家
PAR		Aが場所	×	指の糊
LOC		Aが場所	Bが場所	前の家
TIM		Aが時間	Bが時間	夕暮れの山
QUA		Aが数量	Bが数量	5kgの米
PRO		Aが性質	×	上等の洋服
ACT		(Aが動作)	×	愛用のギター
		(目的)	×	遊びの木馬
ABS		(種類)	×	西陣織の着物
		(目的)	×	夕飯の餃子

表3の分類で、( )でくくられたものは曖昧である。「OBJのOBJ」という分類も曖昧な場合がある(たとえば種類と目的に分類された場合)が、その他は、名詞Aと名詞Bとの関係を考えたり、意味素性を細分化することによって、次のように意味分類が可能である。

OBJのOBJ:

- AがBの部分 → 「一部」
- Aが自然物 → 「Aが場所」
- Aが自然物の加工物 → 「材料」
- それ以外の場合 → (種類)、(目的)

「OBJ の OBJ」とならない場合は、曖昧性がある場合が多い。たとえば、「HUM の OBJ」では、例にあげた3つの分類、例ともいずれに属することも可能である。すなわち、「私の本」と言った場合には、通常は「私が所有する本」であるが、私が作成した本という解釈も可能であり、(やや無理な解釈であるが)「私のための本」という解釈もできる。他の2つの例についてもこれは言える。したがって、意味素性を細かくしてもこの曖昧性は解消できない。

同様なことが「ORG の OBJ」についてもいえる。「A社のシール」といった場合、「A社」の所有、作成、場所のいずれになるかは、文脈に依存して決定されることが多く、単独の名詞句のみからは決められない。

また、名詞AがACTという意味素性を持つ場合にも、BがAの目的語になる場合(「愛用のギター」)、Aを動作の目的とする場合(「遊びの木馬」)などが曖昧になる。これは、関連する動詞(「愛用する」、「遊ぶ」)が自動詞か他動詞かを区別したり、動詞の結合価と名詞の意味素性との関係 [4] を考慮に入れることによって解決できる場合もあるが、一般的には「AのB」のみだけからでは難しい。同じことが、Aが意味素性 ABS を持つ場合にも言える。

3.3 Bが素性 ABS を持つ場合

表4に、Bが意味素性 ABS を持つ場合を示す。ほとんどの場合逆が言えないのがこの素性の特徴である(後述)。また、「OBJ の ABS」という素性関係では、他の意味素性関係と同様に、次のような細分化による区別が可能である。

OBJ の ABS:

- Bが行為、変化の結果もたらされるもの → 「原因」
- 「原因」以外のもの → 「目的」

その他の表4での曖昧な箇所、すなわち、「HUM の ABS」、「ORG の ABS」、「ANI の ABS」、「PLA の ABS」、「PAR の ABS」については、前節と同じように、「所有」、「作成」、「目的」の間で曖昧性が生じることが多い。しかしながら、BがOBJの場合とは異なり、名詞Bの性質によって、ある程度曖昧性を解消できると思われるものも存在する。

たとえば、名詞AがHUMの場合、表のAをすべて「父」で置き換えると、「父の試験」は意味を変えて「所有」になり、「父の俳句」は「作成」になる。また、「芭蕉」で置き換えると、「芭蕉の態度」は「所有」、「芭蕉の試験」はやや曖昧であるが、「目的」と解釈できる。「医師」で置き換えると、いずれも意味を変えずに、「医師の態度」は「所有」、「医師の俳句」は「作成」になる。すなわち、これらの例では、「父の試験」を除いては、統一的な扱いが可能と考えられる。そこで「態度」を「精神的なもの」という統一概念で取り扱い、「俳句」を「文芸」とすれば、この場合には通用するが、これがきちんとした細分化になっているかどうかは他のデータによる検証が必要である。

同様のことは、「ORG の ABS」、「PAR の ABS」についても言える。すなわち、名詞Bの「伝統」、「特徴」、「サイズ」などについて、もう少し意味を詳細に考察すれば、分類が明確化できる可能性がある。

「ABS の ABS」についても、非常に広範囲の意味をカバーしている ABS 同士の組合せであるので、両者の細分化によって分類を明確にできる可能性がある。

表4:「AのB」の分類(BがABSの場合)

A	B	分類	BのA	例
HUM	ABS	(所有)	×	父の態度
		(作成)	×	芭蕉の俳句
		(目的)	×	医師の試験
ORG		(所有)	×	学校の伝統
		(作成)	×	A局の番組
		(Aが場所)	状態	日本の不況
OBJ		原因	状態	車の事故
OBJ		目的	×	車の免許
ANI		(所有)	×	猫の特徴
		(作成)	×	犬の声
PLA		(所有)	×	花の名前
		(目的)	×	お花の作法
PAR		(所有)	×	顔の特徴
		(種類)	×	足のサイズ
		(Aが場所)	状態	顔のニキビ
LOC		Aが場所	×	部屋の条件
TIM		Aが時間	×	今の段階
QUA		Aが数量	B数量	30代の人口
PRO		Aが性質	×	悪性の風邪
ACT		(Aが動作)	×	結婚の意志
ABS		(状態)	×	本来の調子
		(種類)	×	喘息の持病
		(原因)	×	災害の犠牲
		(目的)	×	日程の細目

4 「BのA」の可能性

すでに述べたように、「AのB」と言ったときの名詞句の意味は、一般的にはBの意味である。そのため、「BのA」、つまり「AのB」の逆は、意味が変わったり言えなくなるものがほとんどであると考えられる。すでに示した表2、3、4の中に、「BのA」が言えるかどうかを記載した。ここでは、どのような場合に「BのA」という言い方ができるかどうかについて考察する。すでに述べた表と関連するもののみを以下に列挙する。

(1)「HUMのHUM」(表2参照)が「同格」ならば、Bが固有名詞である場合には、逆が言える。

妻の薫 → 逆:「所有」

主役の俳優 → 逆:×

(2)「ORGのHUM」は「所属」であるが、Bが特定個人である場合には、逆が言える。

学校の先生 → 逆:「所有」

アジアの難民 → 逆:×

(3)「OBJのHUM」で、OBJが場所ならば、それが自然物以外である場合には逆が言える。

車の男(自然物以外) → 逆:「所有」

山の男(自然物) → 逆:×

(4)「LOCのOBJ」(表3参照)は当然「Aが場所」になるが、Aが抽象的な場所の場合には逆が言える。

前の家(抽象的) → 逆:「Bが場所」

公園の滑べり台(具体的) → 逆:×

(5)「QUAのOBJ」、「QUAのABS」(表3、4参照)といったように、意味素性QUAが前に来る場合は、AがBの数量、個数、人数、値段等の場合には逆が言えるが、AがBのサイズ、順番の場合には逆が言えない。

18人の写真家(人数) → 逆:「Bが数量」

48キロ級の小林(サイズ) → 逆:×

(6)「ANIのOBJ」は「種類」を表すが、AがBの表層的な意味を表している場合には逆が言える。

くまのぬいぐるみ(表層的) → 逆:「種類」

犬の鎖 → 逆:×

このように、「BのA」の可能性を探ることによって、各語のさらに深い意味が追求できると考えられるし、また、すでに曖昧性を指摘した、所有、作成などの明確化の可能性も考えられる。

5 おわりに

格助詞「の」の意味を分類し、「の」の前後の名詞の意味素性を考慮することによって、いくつかの意味素性を持つ「AのB」については、素性を考慮するだけで機械的に意味を求める可能性が得られた。本論では詳しくは述べなかったが、「ANIのORG」、「OBJのPAR」のように、意味素性のみから排除できる組合せも存在する。これらを考慮することによって、無意味な解析を避けることができる。

また、一部のものについては、「AのB」という形からだけでは意味が決められないことも明らかになった。これらは、文の中で意味を考慮する必要があるが、文の中でどのような役割の時にどのような意味になるかは今後の検討課題である。

「BのA」という逆の形については、たとえば「ABSのABS」でほとんど逆が言えないということから、同じ意味素性同士の組合せの場合でも、順序が重要な役割を果たしているという考察が得られた。OBJ, ABSという意味素性の分類は非常に粗いもので、さらに細かい意味素性を検討する必要があるが、逆が言えないという観点から見て、両者の名詞の関係をさらに明らかにする必要がある。

本稿では、「の」の分類について意味素性という面からのみ扱ったが、従来から行われてきた係り受け関係や論理的な取り扱いもを取り入れることによって、さらに明確な指針が得られることも考えられる。ここで扱った手法のシステム化も含めて、今後の検討課題である。

参考文献

[1] 加藤 貴子:格助詞「の」の分類と解析に関する研究, 山形大学卒業論文(1995).

[2] 寺村 秀夫:日本語のシンタクスと意味III、くろしお出版(1991)

[3] 奥津 敬一郎:「ボクハウナギダ」の文法、くろしお出版(1991)

[4] 情報処理振興事業協会技術センター:計算機用日本語基本動詞辞書IPAL −辞書編− (1987)

[5] 電子技術総合研究所:新編 日本語品詞列集成(1979)

[6] 田中 康仁:語と語の関係解析資料(1991)

分類語彙表の増補とその利用

中　野　　洋（国立国語研究所）

１．はじめに

分類語彙表は、日本を代表するシソーラスである。多くの研究に用いられた。言語情報処理においても利用は古くからある。言語処理の高度化にともないその利用は増えているが、その意味を正しく理解されていないところがある。そこで、分類語彙表は、「①意味分類表である。②掲載してある語は例である。③多義語の語義すべてを配置したのではない。」ことを述べる。

『分類語彙表』を増補している。現在、増補の候補は82,828語である。「①増補の手順、②現在の語数分布」について述べる。

さらにこれまでになかったと思われる分類語彙表の１つの利用法として「①対照研究への利用 ②Semantic Count への利用」について述べる。

２．分類語彙表

国立国語研究所資料集６『分類語彙表』が昭和39年3月に刊行されていらい、現在29版をかさねる。研究所の刊行物の中ではもっとも発行部数が多い。一般の表現辞典としての利用が多いためだろうが、言語研究への利用も少なくない。宮島達夫・小沼悦(1992)は『分類語彙表』を言語研究に利用した論文１１９例を集めて解説している。この中には、たとえば日本語処理の道具（例えば辞書）としての使用は含んでいない。したがって、『分類語彙表』を直接間接に利用した研究はこの何倍、何十倍にのぼると思われる。

2.1 意味分類表である

分類語彙表は、語の意味分類表であって、事物の分類や概念の分類表ではない。また、これは人間用のシソーラスである。概念と意味の違いについては、柴田武(1988)を参照されたい。

言語処理でも概念と語の意味を区別して扱う方がよいことを示した報告がある。堺和宏ら(1988)は、意味処理用の辞書の構造を、言語に依存しない内容を記述した概念辞書と、言語に依存する情報の概念対応辞書、それに統語情報と概念に依存しない意味を記述する単語辞書の３層に分けることを提案している。

田中穂積ら(1987)は、分類語彙表などの「従来のシソーラスは、階層化されたもの相互が意味的にどの様な関係にあるかが不明確で曖昧なことが多い。たとえば階層化されたもの相互が上位／下位関係にあるのか、それとも部分／全体関係にあるのかがはっきりしない」ことを指摘し、これらが自然言語意味処理には不十分であり、階層関係を明確にしたシソーラスの作成が重要であると主張している。分類語彙表は言語処理用に作成したものではないのでただちに役に立たない面もあろうかと思う。しかし、下に示すように人間用には使いやすい。人間が自然言語を用いて機械を利用する場面などこの種のシソーラスが必要となる状況も生じると思う。

田中もいうようにシソーラス作成は、人間の長期にわたる注意深い作業となる。理想的には、人間にとって作成しやすい分類と作業、人間にも機械にも使える分類が望ましい。我々は従来の目的のために同じ方法で増補している。利用の側で工夫したり、作り替えたりできればよい。

分類語彙表は、人間用である。その分類は次のようになっている。たとえば、「におい(1.504)」の項目には、次の語例がある。

香(か・におい) かおり　芳香　香気　臭み
　臭気　異臭　悪臭　残り香　移り香
体臭　口臭　俗臭

この項目の中は、２つの段落に分かれている。「香(か・におい)」から始まる段落と、「体臭」に始まる段落である。前者は、においそのものを表す語であり、後者は何かのにおいを表す語である。この段落分けは形の上でも示してある。

ひとつの段落もいくつかに分かれる。前者は、「香(か・におい)」、「かおり」、「臭み」、「残り香」で始まる４つのグループに分かれそうである。後者は「体臭」と「俗臭」の２つのグループになろうか。これらのグループ分けの印はなにもない。しかし、利用者である我々には分かる

し、したがって検索も早くできる。

グループの中の語の並びにも意味がありそうである。たとえば、「臭み、臭気、異臭、悪臭」は、良くないにおいのグループだが、最初はより広い意味の語であり和語である。次に漢語で、さらに悪いにおいそのものを表す複合語で後ろ要素が「臭」となっている。これらの配列順序は必ずしも各項目に共通する規則にしたがって並んでいるのではない。しかし、それぞれを読むと分かる。もしこれを５０音順に並べると「悪臭、異臭、臭み、臭気」となってしまう。これと比べると人間には先の配列の方が検索が早いのが分かるだろう。

このような段落やグループ分け、またその並びに印がないから、あるいは明確な規則がないからといって、意味がないわけではない。さらに、機械処理に利用できないとも思わない。その分野の研究に期待したい。

2.2　語は例である。

『分類語彙表』の解説にあるとおり、語は例である。その項目に入る語すべてを掲載しているのではもちろんない。その項目をみて容易におもいつく語はのせていない。いくつかの語が組になっているような場合にもその一部をあげているだけである。「イギリス、アメリカ」はあるが、「スペイン」はない。「真北、真西」はあるが、「真南、真東」はない。

2.3 多義語

解説には、ある程度、多義を考慮したとある。増補においてはさらに考慮したが、完全ではない。したがって、多義語のすべての意味それぞれに番号をつけたのではない。もちろん、番号をつけなかったからといってその意味を認めないというわけではない。2.2のように語は例である。

３．分類語彙表の増補

解説によれば、収録語数はおよそ３万２千６百である（この数値は語彙表の延べ語数ではない）。これらの語は国立国語研究所報告21『現代雑誌九十種の用語用字』第一分冊の語彙表に掲げる使用率の高い語、さらに阪本一郎氏の『教育基本語彙』など日常生活でより基本的な役割をはたしている語である。これを研究に用い、あるいは言語処理に用いるには語が少ない。そこでこれを増補する。

3.1 増補の手順

増補は以下の手順で進めている。

①方針決定　体系の大きな変更は行わない。
　多義語を入れる。サ変語幹を用の類にも入れる。
②候補語の選択　全体に語を増やす。複合語・新語・多義語・慣用句・ｵﾉﾏﾄﾍﾟ・専門語など
③仮番号付け・段落および段落内の位置決定
④項目内の調整　語の追加削除・新段落作成
⑤項目間の調整　品詞分類間の調整ほか
⑥全体の見直し　機会あるごとに行う
⑦表記などの統一　ただし表記の基準を示すものではない。
⑧白表紙版公開　２分冊計 660頁。広く意見を聞く。
⑨公刊

現在は、③④⑤⑥の段階である。次回の白表紙版は、今年秋を目指している。

3.2 現在の語数分布

科研費を受けて作成した「『分類語彙表』形式による語彙分類表」(中野 1989、下表白表紙)の掲載延べ語数は５万２千弱だった。次の科研費「言語研究におけるｼｿｰﾗｽの利用法」（平成元～２年度）では 60,784語を得た。国語研内の課題「分類語彙表の増補」を経て、現在 82,828語（下表 671本）が増補の候補となっている。

この表の語数は、各項目に配置された語例の延べ語数である。多義語はいくつもの項目に配置されているからそれぞれ数えられている。たとえば、６７１本では、ある表記の語がひとつの項目だけに配置されたのは59,850語である。複数の項目に配置された語の項目数と語数は、それぞれ 2-8311、3-1455、4-294、5-80、6-31、7-12、8-8、9-4、10-5、18-1、21-1である。ちなみに 21の項目に配置された語は「する」である。

これらの掲載語の表記の異なり語数は、順に 33,472 47,826 70,052である。

さて、表の増補率をみれば、元の版からどれほど増補されたかがわかる。全体では白表紙が 1.42倍、６７１本が 2.25倍である。

品詞別に見れば、用の類が 3.91倍ともっとも多くなっている。相の類、体の類はそれぞれ1.9倍前後である。用の類が多い理由は、「サ変語幹＋する」の形の語を入れたからである。

分　類	掲載延べ語数			増補率	
	元版(FD)	白表紙	671本	白表紙	671本
体の類	26,984	40,227	54,591	1.49	2.02
抽象的関係	6,780	9,026	12,663	1.33	1.87
人間活動主体	3,272	5,020	7,127	1.53	2.18
人間活動	9,920	14,708	19,247	1.48	1.94
生産物	3,277	5,656	7,960	1.73	2.43
自　然	3,735	5,817	7,594	1.56	2.03
用の類	4,779	5,358	18,710	1.12	3.91
抽象的関係	2,153	2,380	8,145	1.11	3.78
人間活動	2,158	2,441	9,265	1.13	4.29
自　然	468	537	1,300	1.15	2.78
相の類	4,653	6,147	8,928	1.32	1.92
抽象的関係	2,212	2,899	4,316	1.31	1.95
人間活動	1,788	2,515	3,571	1.41	2.00
自　然	653	733	1,041	1.12	1.59
その他類	364	390	599	1.07	1.65
抽象的関係	99	105	127	1.06	1.28
人間活動	265	285	472	1.08	1.78
総計	36,780	52,122	82,828	1.42	2.25

最も増えたのは、2.32で 21語から 153語と7.29倍に増補された。相の類では 3.17で 19語から 103語と 5.42倍となった。体の類では 1.27で 142語から 546語と 3.85倍となった。 2.32(創作)の一部を次に示す。

［ＦＤ版の最初の段落］

1 *著わす 詠む 詠み込む 歌いあげる 焼き直す

［６７１本で上に直接対応する段落］

3　著す　著作する　著述する　撰述する
　執筆する[3152-1]　述作する　書く
　創作する　生む［傑作を～］　作る
　作文する　作詞する　詩作する　句作する
4　詠む　詠み込む　詠ずる　詠じる
　歌う［和歌を～］　歌い上げる
　即詠する　即吟する　偶詠する
　苦吟する　沈吟する　詠進する
5　焼き直す　翻案する　潤色する
　改作する　摸作する
　脚色する　劇化する　アレンジする
　文字化する[23150 1]

以上に示したものは、増補作業中のデータである。語数についても、語例、項目、表記なども変る可能性がある。

4．分類語彙表の利用

分類語彙表の解説には、このようなシソーラスの役割の一つとして情報処理での利用をあげている。実際、言語処理での利用は早かった。
現在でも大学や企業などのいろいろな研究機関で使われている。たとえば、手元にある奈良先端科学技術大学院大学の松本研究室年報 1993-94 Lingua vol.1-2 掲載の論文29件のうち参考文献に分類語彙表があげてあるものが12件ある。

以下では、最近発表者が行った語彙の対照研究での利用と計画中の雑誌の語彙調査でのSemantic Countについて述べる。

4.1 中国流行歌の日中対照研究

標題についての語彙の対照研究を行った。
この種の研究において何を手がかりとすべきかが問題となる。研究の単位である語の認定すらも共通となりえるかさえ疑問である。

この研究では、中国語とその逐語訳を用いた。次の例は、日本語訳で「会う」を用いた一聯と対応する中国語の一聯である。同じ漢字を用いているとはいえ、このように対訳でかつ同じ意味を表わす語の用例を集めなければ分析はすすみがたい。

中国語と日本語訳での用例

再過二十年我們来相会
二十年たったら我々はまた会いましょう

再来看望親愛的媽媽
また親愛なるママに会いに来る。

不見哥哥心憂愁
兄に会わぬと心が憂愁する。

我們再相逢
わたしたちは再び会おう

我們相約在那小木橋
私たちはあの小さい木橋で会うことを約束する

4.1.1 分類語彙表の意味番号による語彙の対照

日本語訳の語に分類番号を付ける。これを集計し、意味分類項目を語数順に並べたのが次表である。表をみると中国流行歌の内容が表れる。すな

わち、異なりの「地形・山野,植物名,川・湖」が多いことは自然現象が題材になっていることを、異なりの「対人感情」と延べの「われ・なれ・かれ,親・先祖」が多いことは人称代名詞や親が話題になっていることを示している。

日本語訳 異なり語数順			延べ語数順		
分類番号	項目名	異	分類番号	項目名	延
1.5240	地形・山野	25	1.2000	われ・なれ	497
1.5520	植物名	20	2.3420	行為	174
1.5250	川・湖	17	3.1000	こそあど	107
1.3020	対人感情	15	2.1527	往復	103
1.1950	一二三	14	2.1200	存在	100

実際、日本と中国の流行歌を比べると山野、川等を表わす語が日本のそれより多いことがわかる。

分類	異なり	延べ	中国	日本
山など	36	88	16.7	3.45
川など	17	62	11.7	2.33
海など	28	51	9.7	6.16

さらに日本語訳の分類番号によって、中国語の類義語を集めることができる。以下に語例を示す。

分類	語例
[波・潮]	海波,海浪,巨浪,驚涛,春潮,清波,波濤,碧波,浪,浪流,涛声
[海・島]	海,海峡,海上,海水,海風,海面,海洋,海疆,岸,重洋,大海,島,東海,南海,湾,戈壁灘,鼓波嶼
[川・湖]	黄河,河流,渓流,湖,湖水,湖面,江水,小河,清流,西湖,泉水,太湖,大江,長江,灘,澎湖

以上のように、対照の手がかりとして意味番号は有効である。他の言語の意味分類を利用してどう異なるかも興味深いテーマである。

4.2 語彙調査における Semantic Count

語彙の対照研究では、異なる言語間の対照だけでなく、同じ言語の異なる材料、たとえば異なる時代の言語材料、異なる分野、異なる作品の比較も重要な課題である。

人称代名詞の比較などのように品詞や語種などの比較はそれぞれにつけた品詞や語種情報によって分析できる。しかし、個々の意味分野の語の差異を知るには、意味情報をつける必要がある。今、我々は分類語彙表を約８万語にまで増補している。そのような語彙を配置できる意味分類体系を作成している。これを用いれば Semantic Count が可能になり、語彙分析がより深められる。

宮島達夫(1986)は、1906年から1976年までの雑誌『中央公論』８冊を比較して場所を表わす語が漢語から外来語に変ったことを明らかにした。たとえば「英国」を「イギリス」、「露国」を「ロシア」、「西洋」を「ヨーロッパ」という具合である。分析には分類語彙表の番号を用いている。

これまでにも、国立国語研究所の語彙調査『総合雑誌の用語（前編）』『雑誌９０種の用語用字』『高校教科書の語彙調査』『中学校教科書の語彙調査』(国語研 1957,62,83,86)には分類語彙表の番号を付けている。山崎誠(1989)は、高校と中学校の教科書のデータを用いて「意味別語彙集」を作成した。これらは見出し語に番号を付けて分析したものである。

Semantic Count とするためには、文脈中の語それぞれに分類番号を付けなければならない。しかし、『分類語彙表』に掲載されていない語が現れた場合、独自に番号を付与できるかどうか、また、慣用句や擬声語・擬態語などへの付与は難しく、さらにひとつの番号に決定できるかどうかが問題である。

困難な問題が山積しているが、得られる成果も大きいと考えられるので試行的な調査を計画しているところである。

参考文献

堺和宏,徳永健伸,奥村学,田中穂積(1988)「自然言語の意味処理用辞書の構成法」（情処技法 Vol.88,No.38 88-NL66）

柴田武(1988)「語の意味と概念と外界」（『日本語大百科事典』,講談社）

田中穂積,仁科喜久子(1987)「上位／下位関係シソーラスＩＭＩＭＡＰＩの作成（Ｉ）」（情処技法Vol.87,No.84 87-NL-64）

中野洋(1995)「中国における流行歌の語彙」(計量国語学19巻8号)

宮島達夫・小沼悦(1994)「言語研究におけるシソーラスの利用」（宮島達夫著『語彙論研究』,むぎ書房）

--(1986)『雑誌用語の変遷』(秀英出版)

山崎誠(1989)「意味別語彙集」（『高校・中学校教科書の語彙調査』,秀英出版）

格パタン分析に基づく日本語動詞の語彙知識獲得

大石 亨　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

近年、動詞の構文的な振舞いは、その動詞の意味によってある程度まで予測されること、すなわち、ある動詞がどのような項構造 (syntactic argument structure) を持つかは、その動詞の語彙的な意味 (lexical semantic structure) と密接な関連があるということが指摘されている [1, 4]。言い替えれば、同じ格構造を持つ動詞には意味的な共通性があり、逆に、同一の動詞でも格のとり方によってその表す意味に違いが出てくるということである。我々は、かつて、この格のパタンによる意味をあらかじめ定義しておくことによって、表層的な情報から動詞の深層格を獲得する試みを述べたことがある [3]。

本稿では、この深層格獲得実験を通して得られたデータに基づいて、動詞の意味構造 (semantic structure) を抽出し、その構造間の関係を明らかにする。

2 意味構造 (semantic structure)

現在の言語研究においては、深層格は動詞がとる名詞句に与えられる単なるラベルではなく、動詞の意味構造を形成する各要素の構造的な位置が持つ属性の一般化であると考えられている [2]。したがって、それぞれの深層格を動詞の意味構造に対応させることにより、深層格構造を動詞の意味構造に還元することができる。さらに、一つの動詞が複数の格構造を持つ場合には、その組合せを考えることにより、より詳細な意味構造を獲得することができる。

例えば、「言う」という動詞について言えば、「『君が好きだ』と言う。」のように引用のト格を伴う補文をとることや、「早く帰るように言う」のように、引用内容を表す「ヨウニ」を伴ったり、「つまらないことを言う」のように「こと」名詞をヲ格にとることから、思考・感情などの知的動作を表すことがわかる。さらに、「彼に対して言う」のように「ニタイシテ」や「ニムカッテ」を伴うことから、相手に対して働きかける動作であることがわかる。また、「嬉しく思う」といえるのに対して、「*嬉しく言う」とは言えないことから、それが伝達動詞であることがわかる。これは、伝達動詞の場合は、引用成分の主体が引用動詞に対して独立的に関わり、「言う」人と伝達内容が無関係なのに、思考動詞の場合は、被引用成分つまり、思考内容の主体 (主語) が引用動詞の主体 (主語) と同一だと原則的に保証されているために、主体を消すことができ、副詞的な修飾への置換が可能であるからと考えられる。「私は (彼が) 嬉しく思う」のようには言えない。このようにして、「言う」という動詞の意味をかなりの程度まで推定することができる。

以下では、動詞の持つ格パタンの組み合わせによって動詞をいくつかのカテゴリーに分類し、そのカテゴリーごとに当てはまる意味構造を考える。

3 格パタンの組み合わせによる動詞分類

我々は、EDR 共起辞書 [5] を利用して、次のような構造を持つデータを 158,888 件蓄積している。

$$\left[\begin{array}{l}\text{動詞：開設する}\\ \text{格要素：}\left[\begin{array}{ll}\text{が：} & \left[\begin{array}{l}\text{名詞：会社}\\ \text{概念関係子：}\textit{agent}\end{array}\right]\\ \text{を：} & \left[\begin{array}{l}\text{名詞：事務所}\\ \text{概念関係子：}\textit{object}\end{array}\right]\\ \text{に：} & \left[\begin{array}{l}\text{名詞：大阪}\\ \text{概念関係子：}\textit{place}\end{array}\right]\end{array}\right]\end{array}\right]$$

このデータから格助詞のパタンを抽出し、すべてのパタンに対しあらかじめ設定してあるパタン区分(表1)を割り当て、動詞ごとに、その動詞がどのパタンをいくつとるかを登録したインデックスファイルを作成した(表2)。

表 1: 表層格パタン

格パタン	記号	格パタン	記号
が	a	が, に	b
が, で	b1	が, において	b2
が, へ	b3	が, に対して	b4
が, によって	b5	が, から	c
が, と	d	が, から, に	e
が, から, へ	e1	が, から, と	e2
が, と, に	f	が, と, に対して	f1
が, を, と	g	が, を, に	h
が, を, で	h1	が, を, において	h2
が, を, へ	h3	が, を, に対して	h4
が, を, から	i	が, を, から, に	j
が, を, から, へ	j1	が, を, から, と	j2
が, を	k	が, について	k1

表 2: 動詞別格パタンファイル (一部)

動詞	用例数	格パタンの区分と数
走る	480	k=169 a=139 b=110 c=16 b3=12 b2=12 e1=6 h=5 e=3 h3=3 h2=2 b5=2
逃げる	95	a=42 b3=18 b=14 c=14 e1=2
飛ぶ	195	a=68 b=35 k=33 b3=20 c=16 b2=7 e=6 e1=4 b5=3 h=2
関係する	111	b=66 a=26 d=18
関連する	114	b=83 a=18 d=13
結びつく	107	b=58 d=33 a=16
駆けつける	54	b=21 a=18 c=7 b3=5 b2=3
出かける	205	b=129 b3=41 a=26 c=3 e=3

さらに、動詞ごとに格のパタンを用例の多い順に並び替え、用例数が全体の5% 以下の格パタンを削除したうえで、同一の格パタンの並びを持つ動詞をまとめることにより、表3に示すようなデータを得た。

表 3: 格パタンの並びと動詞の対応表 (一部)

格パタン	動詞
a	なくなる はっきりする 減る 減少する 上昇する 深まる 進展する 衰える 整う 切れる 絶える 増大する 足りる 薄れる 変化する
a b	おさまる そろう たまる のぼる みえる 近づく 現れる 広がる困る 収まる 終わる 出現する 出来る 成長する 存在する 達する 遅れる 定まる 登場する 努力する 燃える発展する 表れる 落ち着く 劣る
a b b1	でる 育つ 泣く 広まる 降る 合意する 止まる 出回る 進む成功する 続く 誕生する 定着する 動く 普及する 優れる 落ち込む
a b b1 c	生まれる
a b b1 k	光る 生じる
a b b3	動き出す
a b b3 c	延びる
a b c	うかがえる くる 見える 上がる 落ちる
a b c b1	あがる 成り立つ
a b c e	届く
a b d	からむ 共通する 思える 対立する 分かる 変わる 絡む
a b k	よみがえる 完成する 輝く 驚く 欠かす 吹く 漂う 付く
a b k b1	咲く 生きる 浮く
a b k b3	移動する
a b k b3 c	飛ぶ
a b k c	流れる
a b k c b3	飛び出す
a b k h	にじむ
a b1	ふえる まとまる 悪化する 安定する 下がる 行われる 高まる 死ぬ死亡する 成立する 生活する 増える 低下する 売れる 発足する 表面化する 崩れる 亡くなる

次に、同じ格パタンの並びを持つものをその比率によって分割した。同じ<a,b>というパタンの並びを持つ動詞であっても、a(ガ型) がb(ガ、ニ型) に対して極端に大きな値を持つものと、ほぼ同じ値を持つものではその意味構造に違いがあると考えられるからである。同様に、<b,a>という並びを持つ動詞も、bの値が大きく、<b>と同類と考えられるものと、bとaがほぼ同じ値を持つものに分割される。後者は、<a,b>でaとbが似た値を持つものと類似していることになる。表3からも明らかなように、一つの動詞がとる格パタンの数が多いものは、そのグループに含まれる動詞が少数であるのに対し、<a,b><a,b1>のように、格パタンの数が少ないものほど多くの動詞が集中する傾向がある。そこで、この分割処理は、二つのパタンしか持たないグループに対して行なった。

最後に、類似した格パタンの並びを持ち、意味が近いと考えられるものをまとめる処理を行なった。また、任意格による影響と考えられるものを修正し、5%以下で切り捨てたパタンでも有力な情報を提供するものや、未解

析コーパスから得られた情報も考慮して最終的な分類を行ない、それぞれのグループごとに意味構造を考えた[1]。以下に例を示す。

単独変化動詞

これは、格のパタンとしてa(ガ型)のみを持つものである。ガ格には、変化の主体がくる。

なくなる はっきりする 減る 減少する 上昇する 深まる 進展する 衰える 整う 切れる 絶える 増大する 薄れる 変化する 膨らむ

単純に変化することだけを表すのが、このタイプの意味である。「ない」→「なくなる」、「深い」→「深まる」、「薄い」→「薄れる」など、これらの動詞の多くが形容詞から派生したものであることからも推測できるように、変化の結果は語彙の中に含まれている。それで、semantic structureは、次のようになる。

$$[_{\text{Event}}\text{GO}_{\text{Ident}}([_{\text{Thing}}\alpha]_{\text{A}}, \begin{bmatrix} _{\text{Path}} & \text{FROM}([_{\text{State}}\text{BE}([\alpha], [\text{NOT AT}([\beta])])]) \\ & \text{TO}([_{\text{State}}\text{BE}([\alpha], [\text{AT}([\beta])])]) \end{bmatrix})] \quad (1)$$

ここで、Identはsemantic fieldと呼ばれる意味分野の一つであり、抽象的な移動、すなわち、状態変化を表している。FROMおよびTOの引数は、それぞれ変化の前後の状態を表すが、それらは語彙の中に含まれているために項としては実現されず、変化の主体であるGO_{Ident}の第一引数のみにlinking subscriptのAが付されている。これがガ格として表層に現れる。

単独動作動詞

a(ガ型)を最も典型的にとるが、その他の格パタンも同時に持つ動詞は、ガ格で動作の主体を表し、無変化的動きを表すものが多い。これらの動詞は対応する他動詞を持たないという特徴がある。また、主体に意志性があるものとないものがある。

<a,b>という並びを持つ動詞(任意的な原因を表すデ格がついて<a,b,b1>というタイプになる場合もある)には、

のぼる 近づく 現れる 出現する 登場する 表れる 成長する 達する 落ち着く 終わる 進む 動く 落ち込む 発達する

などがあり、これらはニ格で終点(goal)を表すものである。semantic structureは、

$$[_{\text{Event}}\text{GO}([_{\text{Thing}}]_{\text{A}}, [_{\text{Path}}\text{TO}(\left[\left\{ \begin{matrix} \text{PLACE} \\ \text{THING} \end{matrix} \right\}\right]_{\text{(A)}})])] \quad (2)$$

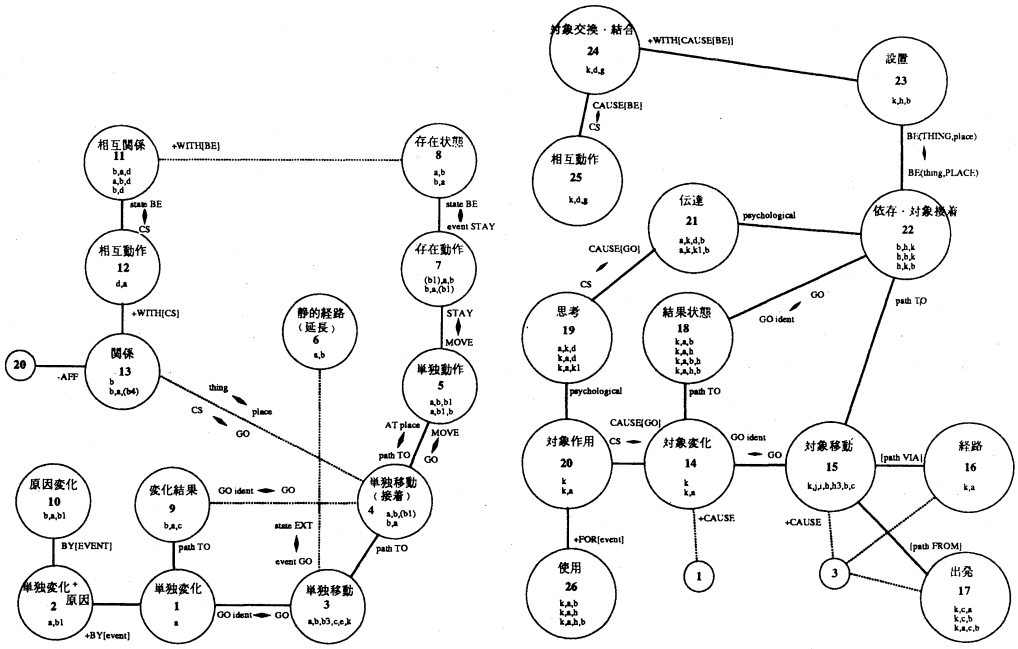
となる。

同様に、約30のグループに対してsemantic structureを設定したが、紙数の都合ですべてを記述することはできない。詳細は大石1995[6]を参照されたい。次節に、全体の構造を示す。

4 Semantic structureの関連図

次ページの右側の図は他動詞に、左側の図は自動詞に、それぞれ対応している。図中の○はカテゴリーを表し、カテゴリー名と、そのカテゴリーに属する動詞が持つ格パタンの並びを記載している。なお、○中の数字は各カテゴリーを識別するためのものであり、特に意味を持たない。カテゴリー同士をつなぐ線は、カテゴリー間の関係を表し、semantic structureの差異を付記している。ここでは、わかり易さのために円で区切って線でつないだが、もちろん、各カテゴリーに明確な境界はなく、関連するカテゴリーは相互に連続したものである。

[1]以下で用いる表現法は、Jackendoff1990[2]に則ったものである。



5 おわりに

本稿では、一つの動詞が複数の格パタンをとるときに、その組み合わせに基づいて動詞を分類し、その分類ごとに semantic structure を考える手法について述べた。詳細を述べることはできなかったが、対象とした 858 個の動詞のうち、約 80%以上の動詞を網羅している。ただし、このうちの約 200 個は<k,a>という並びを持ち、対象作用動詞に含まれるものであり、その意味は非常に多岐にわたっている。これらは格のパタンのみでは判断できないものであり、本手法の限界を示している。

また、分類できなかったもののうち、約 100 個は、場所や原因・目的を表す任意的なニ格やデ格などがつくことによって、本来あるべき位置とは別のところにまぎれこんだものである。さらに、動詞の多義性によって、複数の格パタンの並びが組み合わされたと思われるものもある。「行なう」「する」「なる」のような機能動詞や、自他両方に用いられる漢語サ変動詞などである。今後、さらに詳しい分析をしていく必要がある。

参考文献

[1] Grimshaw, J. *Argument Structure*, Vol. 18 of *Linguistic Inquiry Monograph*, MIT Press (1990).

[2] Jackendoff, R. *Semantic Structure*, MIT Press (1990).

[3] 大石亨, 松本裕治 格パタン分析を利用した深層格獲得手法について, 情報処理学会研究報告, Vol. 94, No. 78 (94–NL–104) (Nov. 1994).

[4] Pinker, S. *Learnability and Cognition: The Acquisition of Argument Structure*, MIT Press (1989).

[5] (株) 日本電子化辞書研究所 EDR 電子化辞書、日本語共起辞書評価版 (第 2.1 版) (1994).

[6] 大石亨 格パタン分析に基づく動詞の語彙知識獲得に関する研究, Master's thesis351202, 奈良先端科学技術大学院大学 情報科学研究科 (Feb 1995).

共起情報を用いた多義動詞の類別と名詞のクラスタリング

平岡 冠二　　松本 裕治
{kanji-h,matsu}@is.aist-nara.ac.jp
奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

コーパス内の共起情報に基づいて、語彙的知識を獲得する試みや、意味的に類似した語をクラスタリングする手法は、これまで数多く提案されてきた[3, 5]。一般には、

仮説 1 意味的に類似した語は同じ文脈に現れる

という仮説がたてられ、その有効性が認められるような実験結果も数多く見受けられる。しかし、そうした研究において語の多義性の問題は無視されることが多いのもまた事実である。特に、動詞に多義性がある場合、名詞を意味的にクラスタリングすることは困難となる。名詞の類似性はそれを支配する動詞を基準としているため、一つの動詞が複数の意味を持ちさまざまな名詞を支配するような現象は誤った語の類似性を導いてしまうことがある。

こうした問題に焦点を当てた研究も幾つかなされている。単言語コーパスの共起情報だけを用いて多義動詞の類別を行なうものに、文献[4]がある。この手法では、

仮説 2：多義は類似した語の集合で区別される

という仮説がたてられ、動詞を、共起する名詞を軸とする n次元ベクトル によって表した後、overlapping clustering(一つのエントリが複数のクラスタに属することを許す) によって多義動詞の意味を区別した動詞の分類を行なっている。

本稿で提案する多義動詞類別の手法も、以上に挙げた二つの仮説に基づくものであるが、多義動詞の類別と同時に名詞間類似度の精度を向上させる方法を用いている。以下の章では、その基本的アイデアと、処理の概要について述べる。

2 動詞の意味分割

多義語の意味は文脈に依存しており、特に多義動詞の場合、それが支配する名詞 (格要素) の意味によって決定される。

例 1
- 椅子に 腰を 掛ける。
- 椅子に 服を 掛ける。
- 彼は 証人に 立つ。
- 彼は 教壇に 立つ。

よって、格要素となる名詞を意味的に分類することにより、多義動詞の持つ複数の意味を類別することが可能であると考える。

上に挙げた例のようにどの格要素が多義動詞の意味を決定するかは一概にいうことはできないが、本研究では、動詞の「を格」に共起する名詞だけを対象に意味分類を行なう。この理由は、(1) 現時点では、形態素・構文解析における曖昧性を回避できない (2)「を格」の格要素については比較的正確に、かつ容易にテキストから抽出できる (3)「を格」は動詞の意味を決定する最も大きな要因と考えられるからである。

以上で述べた動詞分割の考え方をまとめると以下のようになる。

Step 1：動詞と その「を格」に出現する名詞との共起傾向から、名詞間の類似度を定義する。

Step 2：ある動詞 **v** が名詞の集合 **N** をその 「を格」に取った時、名詞類似度に基づいて **N** を意味的なクラスターに分類する。

Step 3：Step 2 の結果 m 個 のクラスターが得られた場合、仮説 2 に基づいて動詞 **v** を m 個 に分割し、それぞれを別の動詞と見なして Step 1 からを繰り返す。

3 名詞の意味分類

名詞の意味分類は名詞間の類似度を基に類似した名詞同士を集まり (クラスター) にすることで行なうが、本研究ではそのアルゴリズムとして代表的な非階層分類法である ISODATA(Iterative Self Organizing Data Analysis Techneques A)[9] を用いた。紙数の都合上、そのアルゴリズムの詳細については述べないが、クラスタリングの条件 (語のまとまり易さ) をコントロール可能であることを除けば、一般的な融合法や分裂法などと大差ないものである。

$$(2)\qquad sim(v,n_i,n_j)=\begin{cases} min(|\,MI_{を}(v,n_i)\,|,|\,MI_{を}(v,n_j)\,|) & : \begin{pmatrix} MI_{を}(v,n_i), MI_{を}(v,n_j) \text{ が} \\ \text{同じ符合である時} \end{pmatrix} \\ 0 & : \text{上記以外} \end{cases}$$

名詞の類似度は、動詞 v とその「を格」に出現する名詞 n の相互情報量 $MI_{を}(v,n)$ を基に導かれる [1, 6]。

$$(1)\qquad MI_{を}(v,n)=\log_2 \frac{\frac{f(v,n)}{N}}{\frac{f(v)}{N}\frac{f(n)}{N}}$$

N ：コーパス中の総文数
$f(v), f(n)$ ：v および n の出現頻度
$f(v,n)$ ：v の「を格」に n が出現する頻度

(1) 式 の値は、名詞の共起傾向を示すものと考え、同じ共起傾向を持つ名詞対 n_i, n_j に動詞 v から見た類似度 $sim_{を}(v,n_i,n_j)$ を (2) 式 のように、また、動詞全体から見た類似度 $SIM_{を}(n_i,n_j)$ を (3) 式 のように定義する。

$$(3)\qquad SIM_{を}(n_i,n_j)=\sum_v sim_{を}(v,n_i,n_j)$$

本手法では分割された動詞をそれぞれ別の動詞として扱うため、分割の処理が行なわれる度に $f(v)$ の値が小さくなり、結果として (3) 式 の値が急増してしまう。よって実際には、(3) 式 の値を [0-1] に単調変換し類似度全体の整合性を保っている。

4 多義動詞類別の手法

動詞分割に使用される名詞クラスターは、**仮説 1** に基づいているため、多義動詞の影響を受けている。よって、十分な類別精度を得るには、名詞間類似度から、多義動詞の影響を取り除かなければならない。

本手法では、動詞分割が行なわれる度に名詞間類似度を再計算することで、多義動詞の影響を類似度から徐々に取り除く方法を用いている。具体的に説明すると、まず、緩い (語がまとまり易い) 条件で名詞をクラスタリングし動詞分割を行なう (大まかな分割)。分割による共起情報の変更にともない名詞間類似度を再計算した後、前回よりクラスタリングの条件を厳しくしていくことで、名詞間類似度と動詞類別の精度を段階的に向上させる。以上の処理の概要を 図 1 に示す。

クラスタリングの条件を際限なく厳しくすれば、過剰な動詞分割を行なってしまうため、条件の上限を設定し

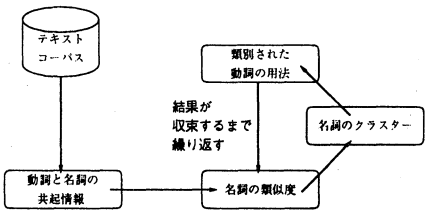


図 1: 類別処理の概要

ておかなければならないが、その設定に一般的指針があるわけではなく、経験的に得るしか方法はない。そこで本研究では、「分類語彙表」[8] を参照することにより動詞の過剰分割を抑制する方法を、ヒューリスティックとして用いる。

「分類語彙表」は、日本語の語彙を概念の階層構造によって分類したものであり、その体の類 (名詞シソーラス) には、約 45,000 語の名詞がその葉の位置に登録されている。ある名詞対 n_i, n_j に付けられた分類コードが一致するレベル (複数のコードがある場合は最大一致レベルのもの) に応じて、「分類語彙表」での名詞間距離 $D_{bgh}(n_i,n_j)$ を次のように定義する。

一致レベル	1	2	3	4	5	6
$D_{bgh}(n_i,n_j)$	16	8	4	2	1	0

過剰分割を抑制する基本的アイデアは極めて単純であり、「分類語彙表」の中で類似した名詞で動詞の分割を行なった場合、これを無視するだけである。以下に、その手順を示す。

「分類語彙表」での名詞間距離 $D_{bgh}(n_i,n_j)$ を用いて、名詞クラスター C の意味的集密度 d_C を以下のように定義する。

$$d_C=\frac{1}{\sharp C(\sharp C-1)}\sum_{n_i\in C}\sum_{n_j\in C} D_{bgh}(n_i,n_j)$$

ここで、$\sharp C$ は、クラスター C の要素の数であるが、「分類語彙表」 にない名詞に関しては あらかじめ除外しておく。

いま i 回目のループで、ある名詞クラスター C^{i-1} が m 個のクラスター $\{C_1^i, C_2^i, \cdots, C_m^i\}$

となり動詞分割に使われたとする。このとき、クラスター分割距離差分 Δ_{C^i} を以下のように定義する。

$$\Delta_{C^i} = d_{C^{i-1}} - \frac{1}{m}\sum_{k=1}^{m} d_{C_k^i}$$

$\Delta_{C^i} < 0$ であれば適切な動詞分割であるとしその結果を採用する。そうでなければ過剰な動詞分割とみなしてその時点での動詞分割結果を無効にする。

5 実験および結果

本研究で行なった実験は二種類で、どちらも同じデータ・条件で行なった。一つは、図 1 で表したものであり、もう一つは、これに先ほど述べたヒューリスティックを付加したものである (→ 図 2)。これ以後、前者を実験-1、後者を 実験-2 と呼ぶことにする。

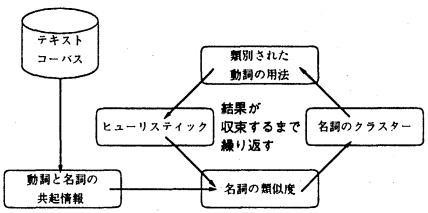


図 2: 実験-2 の概要

5.1 実験データ

基礎データとなる動詞と名詞の共起情報は、EDR 共起辞書 [2] と、朝日新聞の記事データ (総計 約６４万文) から約１８万組を得た。名詞間類似度はコーパス中に１０回以上出現した名詞 (6175 語) を対象に約１４０万組を獲得した。また類別実験は、コーパス中に３０回以上出現した動詞 (1768 語) を対象に行なった。

5.2 結果

実験結果の一例として、動詞 “引く” の類別結果を 表 5.2 に示す。表5.2 は、実験-1 と 実験-2 の結果を並べたものであり、2列目にある名詞は、動詞分割に使われた名詞クラスターの要素である。実際には、動詞 “引く” は 実験-1 で９つ、実験-2 では８つに分割された。

表 1: 「引く」 の類別結果

実験前		実験-1		実験-2	
引く	所得税	引く-1	所得税	引く-1	所得税
	年金		年金		年金
	経費		経費		経費
	人目	引く-2	人目	引く-2	人目
	注目		注目		注目
	血	引く-3	血		血
	水道		水道		水道
	注意		注意		注意
	⋮	⋮	⋮	⋮	⋮

6 評価および考察 (IPAL との比較)

類別結果を評価するために、情報処理振興事業協会の「IPAL 日本語基本動詞辞書」[7](以後 IPAL) に記載されている動詞の用法と、実験で得られた結果との比較を行なった。この評価結果の一例を 表 6 に示す。

表 2: IPAL との比較

I : IPAL 内で「を格」 を持つ用法数
E_i : 実験-i で得た動詞の多義性の数
$I \wedge E_i$: I の用法の一つが、E_i での一つの動詞と対応すると判断された数

	I	E_1	E_2	$I \wedge E_1$	$I \wedge E_2$
引く	10	9	8	4	4
走る	1	10	8	1	1
呼ぶ	5	18	18	2	2
置く	7	17	15	4	3

表 6 が示すように、一部の用法については類別に成功しているが、IPAL との一致率 ($I \wedge E_i / I$) は、それほど高いものではなく、最大で ５７%、平均では、実験-1・実験-2 共に ４５% であった (内容は異なる)。

ヒューリスティックを用いた 実験-2 の場合、過剰な動詞分割を押えることができた (一つの動詞につき約２つの分割を押えた) が、一致率を向上させるまでには至っていない。不一致の原因としては、二つの事が考えられる。一つは名詞間類似度が不足していたためであり、実際には類似した名詞であっても、その類似性が観測され

ず動詞の分割に使われてしまった例がほとんどであった。また、実験-2 では「分類語彙表」にない名詞が全体の 7 割強を占めており、十分な効果を発揮することができなかったこともある。この問題に関しては、コーパス量を増やすことで、ある程度対処できると考えられる。

もう一つは、名詞類似度に含まれるさまざまな影響を取り除けなかったことが考えられる。つまり、誤った類似度を導く原因は、動詞の多義によるものだけでなく、名詞の多義も関連しているということである。本研究では共起情報から名詞の類似度を導く際、名詞の意味を一つに固定しているが、一般に、名詞の意味は一つではなく、複数の局面(概念)を合わせ持っている。そして、名詞の意味はそれを支配する動詞によって決まるものであり、どのような動詞が、名詞のどの局面を引き出すかということは表層の情報から判断できず、結果として誤った距離を導いてしまう可能性がある。よって、今後名詞の多義を類別することが必要となるが、名詞の意味の違いをとらえるために動詞を固定したのでは全く解決にならない。こうした、動詞・名詞どちらか一方から他方を一元的に見る弊害を避けるには、双方の問題を段階的に解消していくしかないと考えられる。この点に関しては今後の課題である。

次に、用法の数だけを比較するために、IPAL から 20 の動詞を、用法の数が少ないものから多いものまでを選んだ。この 20 の動詞を横軸に並べ、IPAL ならびに 実験-1 での用法の数(E_1 の値)を縦軸に取ったグラフを 図 6 に示す。図 6 の中で、実線で書かれているのが IPAL

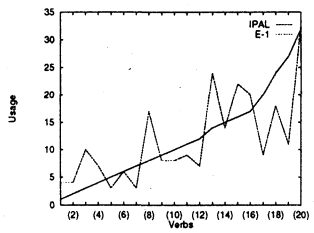


図 3: IPAL と 実験-1 での用例数の比較

での用法の数を表しており、破線は、実験-1 での用法の数を表している。実験-1 のグラフはかなりばらつきがあるものの、IPAL での用法の数が増えるにつれ、同様の増加傾向を示している。よって、多義動詞の用法の数に関しては、おおよその推定が可能であると考えられる。今後、このばらつきを押えより精度を高める必要があるが、これも先ほどと同様に、名詞の多義性の問題を避けて通ることはできない。

7 おわりに

コーパスを用いた語彙知識獲得の研究における問題の一つに、動詞が持つ多義性の問題がある。本論文では、動詞の「を格」に現れる名詞を意味的な集合にすることで、多義動詞の持つ複数の意味を分割する手法について述べた。この手法を用いた実験では、多義動詞の一部の意味については類別が可能であること、多義性の数に関してはおおよその推定が可能であることを示したが、大量のコーパスを必要とする性質上、十分な精度を示すことはできなかった。現在、今回使用したテキストデータの数倍の規模を持つコーパスが利用可能であるため、より大規模な実験が可能であるが、我々は本稿の手法を拡張することで、十分な多義動詞類別が可能であるとは考えていない。動詞の多義の問題は、名詞の多義を無視して解消できるものではないため、双方を段階的に解消していくべきであろう。

謝辞

EDR 電子化辞書データの使用を許可して下さった株式会社 日本電子化辞書研究所に感謝致します。

参考文献

[1] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):235–242, 1990.

[2] EDR. 日本語共起辞書評価版 *(第 2.1 版)*. (株) 日本電子化辞書研究所, 1994.

[3] Francesc Ribas Framis. An experiment on learning appropriate selectional restrictions from a parsed corpus. In *Proc. of COLING-94*, pages 769–7748, 1994.

[4] Fumiyo Fukumoto and Jun'ichi Tsujii. Automatic recognition of verbal polysemy. In *Proc. of COLING-94*, pages 762–768, 1994.

[5] Ralph Grishman. Generalizing automatically generated selectional patterns. In *Proc. of COLING-94*, pages 742–747, 1994.

[6] D. Hindle. Noun classification from predicate argument structures. In *Proc. of the 28th Annual Meeting of ACL*, pages 268–275, 1990.

[7] IPA. 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, 1987.

[8] 国研. 分類語彙表. 国立国語研究所 秀英出版, 1964,1993.

[9] 奥野 忠一, 久米 均, 芳賀 敏郎, and 吉澤 正. 多変量解析法. 日科技連, 1974.

自然言語における程度概念の離散モデル

亀井 真一郎　村木 一至
NEC 情報メディア研究所

1 はじめに

実世界における物事の度合、程度は連続的に変化する。自然言語が有限の語彙を使用して程度概念をどのように表現しているかに関しては、従来から種々の興味深い言語現象が指摘され、修飾の選択制限、否定との関係、会話の含意、論理語との関係などが研究されてきた [1, 2, 3, 4, 5, 6]。我々は「自然言語には程度概念を離散的にとらえる理解の枠組が存在する」という観点にたったモデルを導入し、程度概念にまつわる諸現象を統一的に取り扱うことを試みている [7, 8, 9, 10]。本稿ではこのモデルを主として程度概念の否定操作に適用してその有効性を説明する。

2 程度概念の特徴とモデル化する際の困難

自然言語には物事の程度を表す表現が多数存在する。例えば以下に示すのがその例である。

(1) a. always, often, sometimes, seldom, never
b. all, many, some, few, no
c. tall, short

(2) a. 常に, しばしば, 時々, めったに.. ない, 決して.. ない
b. 全て, 多く, いくらか, ほどんど.. ない, 全く.. ない
c. 高い, 低い

これらの語が指し示す現実の量は状況や話者の判断によって大きく変化する。「しばしば」と「時々」とを分ける明確な基準があるわけではなく、これらの言葉が発話されたときに指し示す具体的な量は聞き手にはわからない。にもかかわらず我々は日常これらの表現を多用し、かつその意味が他の表現と比べて特に曖昧だとは感じていない。程度表現の意味理解とは、これらの言葉の指す現実の量がわかるということではなさそうである。

程度表現はどの言語にも存在するが、従来は異言語の表現を比較する枠組が欠如していた。通常、日本語の「しばしば」と英語の「often」は意味的に対応すると考えられるが、そう考えて良い理由は自明ではない。両者とも現実世界に具体的対象を持たないから意味的対応の根拠を現実世界の量に求めることはできない。「しばしば」と「often」が対応するというのはどういう意味においてか、またその対応はどこまで厳密になりたつのか、それを記述するための枠組を定義する必要がある。

また程度表現は副詞、代名詞、限定詞、形容詞など品詞を越えて存在し、いくつかの共通した意味的性質を持っている。例えば程度強調詞類 (Intensifier) に修飾されるときの制限を考えてみる。often, seldom, many, few, tall, short は very に修飾されるが、all, no, always, never は very には修飾されない。反対に前者は almost に修飾されないが、後者は almost による修飾を受ける。また一方 some, sometimes は very にも almost にも修飾されない。しかしながら従来はこのような品詞とは独立な意味的性質を理解する枠組が欠けていた。

表 1: very、almost と程度表現の修飾関係

	very	almost
often, seldom, many, few, tall, short	+	−
all, no, always, never	−	+
some, sometimes	−	−

我々はこれらの問題に答えるため程度表現の新しい意味モデルを提案した [7, 8, 9, 10]。次節以降ではそのモデルの基本概念を説明する。

3 程度の基本概念素とリスト表現

強調詞の修飾制限は、品詞に独立な、程度に関する概念素の存在を示唆している。そこで我々は程度表現の基本的概念素として、「all」「many」「some」「few」「no」の意味から抽出した5つの意味素「A」「M」「S」「F」「N」を仮定した。さらに、これらの意味素を並べたリストを導入し、具体的な程度表現の意味を表すのに用いる。

(3) {A, M, S, F, N}

上記が程度概念の意味を表すための基本リストである。個々の程度表現、例えば「高い」「低い」の意味を「高さ」という軸と「多い」「少ない」という程度に分け、後者の部分をこのリストを用いて表現すると、表3の様になる。すなわち、各々の程度表現は一つ一つで独立に意味が定まるのではなく、リストで表現されるような枠組全体の中の相対的位置によって意味が定まるのだと考える。現実世界の程度は連続量であるが、それを言

語世界ではこのリストの値のように離散的にとらえていると考えて程度の概念素を仮定すると、程度概念と程度強調詞類の間の修飾の選択制限を表4のようにモデル化できる。これは程度表現がもつ意味的性質の一つを統一的に記述している。

表 2: 程度表現の意味のリスト表現 (1)

基本リスト	{A,	M,	S,	F,	N}
all, always	{A,	−,	−,	−,	−}
many, often	{−,	M,	−,	−,	−}
some, sometimes	{−,	−,	S,	−,	−}
few, seldom	{−,	−,	−,	F,	−}
no, never	{−,	−,	−,	−,	N}

表 3: 程度表現の意味のリスト表現 (2)

基本リスト	{M,	S,	F}
tall	{M,	−,	−}
not tall and not short	{−,	S,	−}
short	{−,	−,	F}

表 4: 程度強調詞類と程度概念素との修飾制限

程度強調詞	例	程度の概念素				
		A	M	S	F	N
a. 増幅語	very, extremely	−	+	−	+	−
b. 緩和語	somewhat, pretty	−	+	−	+	−
c. 弱化語	a little, slightly	−	+	−	+	−
d. 近接語	almost, nearly	+	−	−	−	+
e. 完結語	absolutely	+	−	−	−	+

4 程度概念の2重リスト表現

4.1 数量を含む質問とそれに対する答え

一般の程度表現を取扱う前に問題点を簡単化するため「数」を含む以下の例文を考える。

(4) I solved three of the problems.

この文は通常「問題をちょうど3問解いた」と解釈される。しかし例えばテストの合否のボーダーラインが3問であるような状況では、発話者が解いた問題数が実際には4問以上であっても上記の文はごく自然に発せられる。この事実は発話の「数」部分の意味表現が単なる「数」では不十分であることを示している。このことは下記のような質問とその応答の考察でより明確になる。

(5) A: Did you solve three of the problems ?
B: − Yes, in fact I solved four.
− No, I solved four.

興味深いことに上記の問いには Yes/No 両方の答が可能である。この現象を説明するために以下の5つの状態を考える。(1) 問題が全て解けた状態、(2) 解けた問題数が文中に表現されている数 (=この場合「3」) を越えている状態、(3) 解けた問題数が文中に表現されている数に一致している状態、(4) 解けた問題数が文中に表現されている数に達していない状態、(5) 全く問題が解けなかった状態。これらの状態を表現するために「A」「>n」「=n」「<n」「N」5つの概念素を導入し、下記のようなリスト表現でそれらの相対位置を表現する。

(6) $\{A, >n, =n, <n, N\}$

このリスト表現を用いると上記の5つの状態は表5の様に表される。

表 5: 数の意味のリスト表現

基本リスト	{A,	>n,	=n,	<n,	N}
all	{A,	−,	−,	−,	−}
>three	{−,	>n,	−,	−,	−}
three	{−,	−,	=n,	−,	−}
<three	{−,	−,	−,	<n,	−}
none	{−,	−,	−,	−,	N}

さて上述の質問文に Yes/No 両方の答えが可能であることを説明するため、文中の「数」部分の意味表現として以下のような2重リスト表現を導入する。

(7) $\left\{ \begin{matrix} -,-,=n,-,- \\ A,>n,=n,-,- \end{matrix} \right\}$

この2重リストの上段は「直接解釈リスト」であり、発話中にある数表現の値そのものを表している。下段は「可能性解釈リスト」であり、発話中には直接現れない解釈を表現している。この二つのリストの全体が発話された「数」部分の意味を表現するものと考える。この2重リストを用いると質問に対する答の可能性が図1のように単純な操作により求められる。図1の左の2重リストは質問文中にある数「3」を表現する。図の真中は実際の状況を表しているリスト表現 (実状況リスト) である。ここでは実際に解いた問題数「4」を示している。実際の状況は発話ではないので1重リストで表現される。さてこの二つのリストの共通部分 (intersection) を計算する。問いの2重リストの上段 (直接解釈リスト) と実状況

リストとを比較すると共通部分はない。したがってこの場合の答えは「No」になる。一方、質問の2重リストの下段(可能性解釈リスト)と実状況リストには共通部分(>n)が存在する。これは「Yes」の答えに対応する。2重リストの上下二つのリストの差の部分が、このモデルにおける「会話の含意」の可能性の明示的表現である。

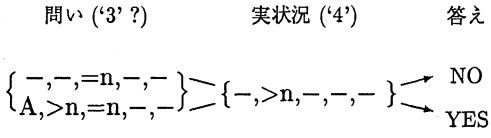


図 1: 数を含む問いに対する答えの可能性

4.2 数量を含む文に対する否定操作

本節では数を含む肯定文に対する否定文の意味を考察する。以下の文が文(4)に対応する否定文である。

(8) I didn't solve three of the problems.

この否定文にはいくつかの意味解釈が存在する。まずその一つは「解けなかった問題が3問ある」という解釈である。本論文のモデルを用いると、肯定文の意味表現から否定文のこの解釈の意味表現が図2のようにして導出できる。すなわち肯定文の2重リストの値の意味を「解けた問題の数」から「解けなかった問題の数」に変更するだけでよい。この操作は、この解釈で元の肯定文の「数」部分が否定語の作用領域の外にあることに対応している。

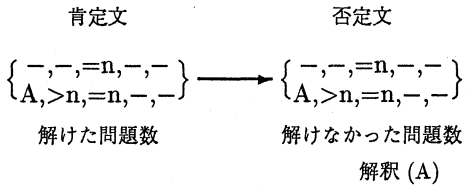


図 2: 数を含む否定文の解釈(A)

この否定文の他の解釈としては下記の対話例が示すように「解いた問題の数が文中に表現されている数(=3)に達しなかった」という解釈がありうる。

(9) A: Did you solve three of the problems?
B: No, I didn't (get to) solve three of the problems.

——— 解釈 (B)

図3に肯定文の意味表現からこの解釈を導出する操作を示す。まず肯定文の2重リストの上下各リストの反転リストを作成する。次にその二つの反転リストの共通部分をとる。それが求めるべき解釈の下段「可能性解釈リスト」である。この可能性解釈リストから両端の値(=AとN)を削除して得られたリストが求める解釈の上段「直接解釈リスト」であり、この両者を合わせた2重リストが求める解釈の意味表現である。二つの反転リストの共通部をとったこの解釈は、肯定文の「直接解釈」の否定と「可能性解釈」の否定の両方に合致する否定解釈である。この否定解釈の「直接解釈リスト」を求める際に両端の値を削除する操作は、この否定解釈が通常、解いた問題の存在自体および解けなかった問題の存在自体は否定していないことに対応している。

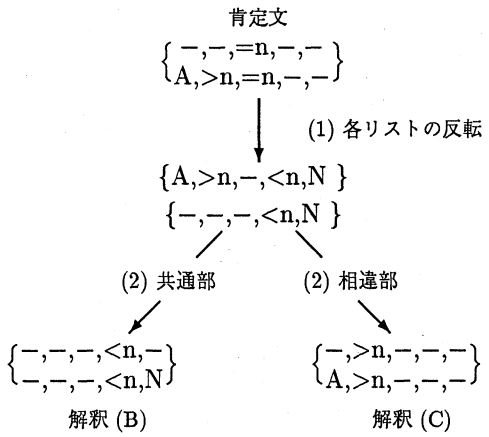


図 3: 数を含む否定文の解釈(B,C)

この否定文は下記の例文(10)が示すように「解いた問題の数が文中に表現されている問題の数を越えている」という解釈も可能である。この解釈を肯定文の意味表現から導出するには、図3において2つの反転リストの相違部分をとればよい。すなわちこの解釈は元の肯定文の「直接解釈」の否定だけしか満たしていない否定解釈である。なおこの否定文がこのように解釈される場合には、文が発話される際に数部分に強勢(ストレス)がおかれる。この言語現象はこの否定操作の中で二つのリストの相違部分をとる操作に対応している。

(10) A: Did you solve three of the problems?
C: No, I didn't solve THREE of the problems: I solved ALL of them.

——— 解釈 (C)

5 程度概念の否定

5.1 'Many' 'A Few' 'Few'

本節では一般の程度表現にこのモデルを適用しそれらの否定の意味解釈を導出する。「many」「few」「a few」の意味を2重リストを用いて表現したのが図4である。通常「few」は否定的な少なさを表し「a few」は肯定的な少なさを表すと言われるが、そのことを明示的に表現したのはこのモデルが初めてである。

(a) many $\begin{Bmatrix} -,M,-,-,- \\ A,M,-,-,- \end{Bmatrix}$

(b) a few $\begin{Bmatrix} -,-,-,F,- \\ A,M,S,F,- \end{Bmatrix}$

(c) few $\begin{Bmatrix} -,-,-,F,- \\ -,-,-,F,N \end{Bmatrix}$

図 4: 'Many' 'A Few' 'Few' の2重リスト表現

5.2 'not many' の解釈の導出

「many」の意味を表す2重リストに対して前節で導入した否定操作を適用させたのが図5である。反転リストの共通部分をとった2重リストは「not many」が「単なる Some ではなくそれ以下をも含意する」ことを表す。反転リストの相違部分をとったもう一方の2重リストは「not many」が「全て」を意味しうることを表すが、上段の「直接解釈リスト」が空であることから、これが通常行なわれない解釈であることがわかる。このようにこのモデルは「not many」の vagueness も ambiguity も解釈の妥当性も同時に導出することができる。

6 おわりに

本論文では程度概念の意味を2重リスト内の離散的な相対位置で表現するモデルを説明した。この2重リスト表現上で定義した否定操作は肯定文の意味表現を元にして否定文の全ての解釈を導出することができる。

程度表現の意味理解とは指し示される現実量がわかることではなく、理解の枠組(リスト表現)内の相対的位置がわかるということである。また「often」と「しばしば」が対応するという直観は、日本語と英語の程度概念が類似した構造をもち、かつこの二語の相対的位置が対応しているという事実に基づいている。なお程度の概念素は本稿で導入したものだけではない。重要なのは自然言語における程度概念の離散的な把握の枠組である。

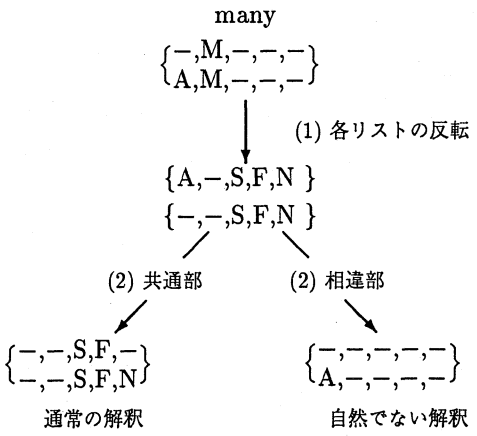


図 5: 'Not many' の解釈の導出

本稿の程度表現の扱いを論理の観点から見ると量化子の一般化に相当する。現在はこのモデルの他の論理演算子への適用を試みている。例えば従来から程度表現と論理演算子「OR」との類似性が指摘されているが「inclusive OR」と「exclusive OR」とを2重リストに対応させられるものと考えている。「OR」「AND」「ならば」など一階の述語論理の演算子を拡張させることで自然言語の推論に適した論理の構築を目指したい。

参考文献

[1] Grice, H. P. 'Logic and conversation' Syntax and Semantics 3. Academic Press. 1975.

[2] Horn, L. R. 'On the semantic properties of logical operators in English' The Indiana University Linguistics Club 1976.

[3] Bolinger, D. L. 'Degree words' Mouton. 1972.

[4] Gazdar, G. 'Pragmatics: Implicature, Presupposition, and Logical Form' Academic Press. 1979.

[5] 太田朗「否定の意味」大修館書店 1980.

[6] Hirschberg, J. B. 'A theory of scalar implicature' ペンシルバニア大学博士論文. 1985.

[7] 亀井、村木「程度表現のモデル化」電子情報通信学会 NLC 研究会 88-6. 1988.

[8] Kamei and Muraki 「An Explanatory Model of Degree Concept」自然言語理解と人工知能国際シンポジウム ISKIT'92. 1992.

[9] 亀井、村木「程度表現の意味モデル」情報処理学会第4 5回全国大会 予稿集 1E-10. 1992.

[10] Kamei and Muraki 「A Discrete Model of Degree Concept in Natural Language」 Coling-94, pp.775- 781. 1994.

属性比喩の理解のための計算モデル

内山 将夫　　板橋 秀一
筑波大学

1 はじめに

本稿では、修飾表現の情報理論的モデルを提案し、それに基づいた属性比喩の解釈を述べる。属性比喩の典型は、「狼のような男」という句である[1]。この句は、男を狼に喩えることにより、男の獰猛さを強調している。このように、喩詞(喩える言葉)と被喩詞(喩えられる言葉)との関係のうちでも、特に、両者の意味を構成する属性間の関係により意味が決まるような比喩を属性比喩という。

我々は、文献[2]で、情報理論に基づいた属性比喩の計算モデルを提案した。本稿では、そのモデルを拡張し、修飾表現全体のなかに属性比喩を位置付ける。

2 修飾表現の情報理論による定式化

次の二つの観点から修飾表現を特徴付ける。

- (他の修飾表現に対する)ニュース性、
- (被修飾語に対する)明瞭性。

まず、名詞の意味を定義し、次に、ニュース性と明瞭性とを定義する。

2.1 名詞の意味

名詞や名詞句の内包的意味[3]を概念と呼び、それを図1のように素性構造で定義する[1]。

〈概念〉:=(〈概念名〉　(〈属性〉+))
〈属性〉:=(〈属性名〉　((〈属性値〉〈確率〉)+))

図 1: 概念の定義

[1]概念を構成する属性と属性値及びそれに付与する確率の決定法は本稿の考察外である。

ある概念をCとし[2]、その属性F_iが属性値A_jからなるとすると、A_jの確率$p(C.F_i.A_j)$は次の式を満す。

$$p(C.F_i.A_j) \geq 0,$$

$$\Sigma_j p(C.F_i.A_j) = 1.$$

図2に概念の例を示す。また、概念Cの属性F_iのエントロピー$S(C.F_i)$は次式である。

$$S(C.F_i) = -\Sigma_j p(C.F_i.A_j) \lg p(C.F_i.A_j),$$

ただし、lgを2が底の対数とする。$S(C.F_i)$を概念Cの属性F_iの曖昧さと呼ぶ。

```
(狼      ((性格    ((荒荒しい . 0.9)(大人しい . 0.1)))
          (毛深い   ((yes . 0.99)(no . 0.01)))))
(男      ((性格    ((荒荒しい . 0.6)(大人しい . 0.4)))
          (毛深い   ((yes . 0.01)(no . 0.99)))))
(男の子  ((性格    ((活発な . 0.7)(大人しい . 0.3)))
          (外見    ((若い . 0.99)(老いた . 0.01)))))
(女の子  ((性格    ((活発な . 0.3)(大人しい . 0.7)))
          (外見    ((若い . 0.99)(老いた . 0.01)))))
```

図 2: 概念の例

一般に、修飾表現と被修飾語とは、外延が同じであり、内包が異なる。これは、同一の確率変数における確率分布の変動、つまり、修飾語を条件とする条件付き確率と考えることができる。すなわち、修飾表現NPと被修飾語Nとについて属性F_iの属性値A_jの確率は、修飾語をAPとすると、

$$p(NP.F_i.A_j) = p(N.F_i.A_j|AP)$$

という関係にあると解釈できる。これに基づいてニュース性と明瞭性とを定義する。

[2]混同の恐れがない限り、名詞や名詞句と概念とを同一の記号で表わす。

2.2 ニュース性

被修飾語Nについて属性F_iの属性値A_jの確率が$p(N.F_i.A_j)$のとき、修飾表現NPの属性F_iのニュース性$N(NP.F_i)$を次式で定義する。

$$N(NP.F_i) = -\Sigma_j p(NP.F_i.A_j) \lg p(N.F_i.A_j) - S(N.F_i)$$

このとき$p(N.F_i.A_j) \geq \epsilon > 0$を仮定する。$\epsilon$は最小の確率であり、本稿では$\epsilon = 0.01$である。

ニュース性は獲得された平均情報量と獲得を期待された情報量(エントロピー)との差である。例えば、被修飾語の属性において確率が$p(N.F_i.A_j)$であった属性値が修飾表現において確率1であったとすると、このときに獲得された情報量は$-\lg p(N.F_i.A_j)$である。しかし、一般には、修飾表現において属性値が一つには確定しない。そのときには、$N(NP.F_i)$の第一項は獲得された平均情報量となる。

これは、修飾表現により限定された外延が、どのくらい稀な事象かを表わしているので、(他の修飾表現に対する)ニュース性と呼ぶ。

2.3 明瞭性

被修飾語Nと修飾表現NPの属性F_iに関する曖昧さを、それぞれ、$S(N.F_i)$と$S(NP.F_i)$とする。$S(NP.F_i)$はNに修飾語を付加した結果の曖昧さであるので、$S(N.F_i)$と$S(NP.F_i)$との差は、修飾表現により与えられた被修飾語の内包に関する情報であり、曖昧さの減少、すなわち明瞭さの増加である。(被修飾語Nに対する)修飾表現NPの属性F_iの明瞭性$C(NP.F_i)$を次式で定義する。

$$C(NP.F_i) = S(N.F_i) - S(NP.F_i)$$

ニュース性と明瞭性は共に負にもなる。

2.4 属性比喩の特徴

属性比喩の特徴は、その内包的意味が喩詞と被喩詞から合成できることである。喩詞と被喩詞とはそれぞれ名詞であるので、その概念をV, Tとする。すると属性比喩Mを構成する属性F_iの属性値A_jの確率$P(M.F_i.A_j)$は次式であると仮定する。

$$p(M.F_i.A_j) = \frac{\alpha}{\alpha+\beta} p(V.F_i.A_j) + \frac{\beta}{\alpha+\beta} p(T.F_i.A_j) \quad (1)$$

αとβは、喩詞と被喩詞の属性値が比喩の属性値に与える影響(重み)を示す。この値は次の三つの要因により決まると仮定する。

- 語彙的要因

  同一概念でも属性値が異なれば重みが異なる場合がある。

- 構文的要因

  喩詞には典型的な言葉が使われるため、一般に、$\alpha \geq \beta$であろう。

- 文脈的要因

  比喩が使用される文脈によっては、α, βの値が語彙的要因と構文的要因だけでは決まらない。

本稿では$\alpha = \beta = 1$とする。

3 ニュース性/明瞭性と修飾表現との関係

図3にニュース性/明瞭性と修飾表現との関係を示す。図3は、属性値が二つである任意の属性について、被修飾語の属性の確率分布と修飾表現のそれとが一定の関係にあるときの、ニュース性(News)と明瞭性(Clarity)の値をプロットしたものである。

図3において、"Minimum"で示されるプロットは、被修飾語と修飾表現とが確率的に独立であるとし、それぞれの属性の確率分布をランダムに選択したときの三万組に対して、ニュース性が最小であったもののプロットである。

また、"Literal Expressions"は、被修飾語の属性値の確率分布をランダムに設定し、一つの属性値を選んで、その属性値の修飾表現における確率を1としたときのニュース性と明瞭性との関係である(五千組)。例えば、"赤い花"という修飾表現は、被修飾語の色に関する属性値のなかで"赤"が修飾表現において確率1となるため、この範疇に入る。このプロットにおいて、被修飾語の属性値のうちで、確率の大きいほうが、修飾表現において確率1ならば、そのときの

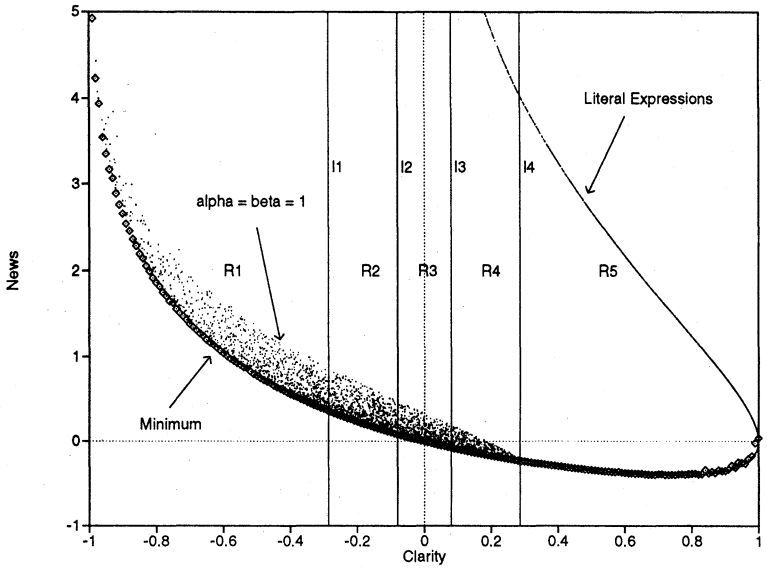


図 3: 断定的な表現と曖昧な表現におけるニュース性と明瞭性

ニュース性は最小値になり、"Minimum" の上に落ちる。したがって、"Literal Expressions" が示すプロットは、確率の小さいほうの属性値が修飾表現において確率 1 であったときのものである。なお、どちらの場合でも、修飾表現の曖昧さは 0 であるため、明瞭性の値は被修飾語の曖昧さ (≥ 0) に等しい。

"alpha = beta = 1" で示されるプロットは、修飾語と被修飾語の属性値の確率分布が独立であるとした五千組に対して、修飾表現の確率を (1) 式において $\alpha = \beta = 1$ として求め、ニュース性と明瞭性とを計算した結果である。

修飾表現の属性値のうちの一つが確率 1 となり、残りが 0 となるような修飾表現を断定的な表現と呼び、修飾表現の属性値の確率が (1) 式で決るようなものを混合表現と呼ぶ。すると断定的な表現では、どの属性値が選択されたかが問題なので、ニュース性が重要となる。(どの属性値が確率 1 となっても、明瞭性の値は変らない。) それに対して、混合表現では、どの属性値も特には選択されないので、ニュース性は重要ではない。明瞭性が重要となる。

図 3は、四本の直線により五つの領域に分割されている。直線 $l4$ は明瞭性が 0.28 の位置にある。被修飾語において、二つの属性値の確率のうち一つが 0.95 であり、修飾表現が断定的であるなら、明瞭性は 0.29 になる。また、混合表現 ("alpha = beta = 1" で示されるプロット) における明瞭性の最大値は 0.27 である。したがって、$l4$ は断定的な表現と混合表現とを区別する。$l4$ の右側、領域 $R5$ に位置する修飾表現が断定的な表現である。

直線 $l3$ は、明瞭性が 0.08 の位置にある。これは、被修飾語において、二つの属性値のうちの一つの確率が 0.99 であった場合の断定的な表現における明瞭性である。また、直線 $l2$ は縦軸に関して $l3$ と対称である。この二本の直線により区切られる領域 $R3$ に位置する表現は、修飾表現として不適当であると考えられる。例えば、図 2の概念によると、"若い男の子" や "老いた男の子" がそうである。前者は、修飾するほどの価値がなく (ニュース性が低く)、後者は表現の対象が考えにくい (ニュース性が高すぎる)。つまり、$R3$ には、当り前のことか有りえないことを述べた表現が位置する。

直線 $l1$ は、縦軸に関して $l4$ と対称である。対称である必然性はないが、これにより分けられる領域 $R1$ には、明瞭性の値が低いため修飾表現として不適当な

表現が位置すると考えられる。

残りの領域$R2, R4$が属性比喩の領域である。ここに位置する表現の明瞭性は適度なものである。それに加えて喩詞と被喩詞とが存在するという言語形式から、当該の表現が属性比喩として適当な表現であると判断される[3]。

4 属性比喩と明瞭性

前節で述べたように、属性比喩においては明瞭性が重要となる。表1に図2から構成される表現を明瞭性と共に示す。表現の属性値の確率は(1)式において$\alpha = \beta = 1$として求めた。

領域	表現.属性	明瞭性
R4	狼のような男.性格	+0.16
R3	男の子のような女の子.外見	0
	女の子のような男の子.外見	0
R2	男の子のような女の子.性格	-0.12
	女の子のような男の子.性格	-0.12
R1	男のような狼.性格	-0.32
	男のような狼.毛深い	-0.92
	狼のような男.毛深い	-0.92

表 1: 属性比喩と明瞭性

属性比喩の特徴のうち、次の二つが表1から読み取れる。

選択的強調 被喩詞の属性のうちの幾つかが比喩において強調される。強調される属性は、領域$R2$と$R4$に位置するものである。例えば、「狼のような男」では“性格”という属性が強調される。“毛深い”という属性は、「狼」と「男」とでは確率分布が大幅に違うため、領域$R1$に位置することになり、強調される属性とはならない。

意味の非対称性 「AのようなB」と「BのようなA」とでは意味が異る。これには明瞭性の非対称性と確率分布の非対称性とが関わる。「狼のような男」とは言うが「男のような狼」と言わないのは、前者においては、領域$R4$に“性格”という属性が位置するのに対して、後者では、全ての属性が$R1$に位置するためである。これは明瞭性の非対称性である。

「男の子のような女の子」と「女の子のような男の子」とでは、“性格”という属性が領域$R2$に位置する。これらの明瞭性に差はない。しかし、(1)式で計算される属性値の確率を見ると、「男の子のような女の子」では、“性格”という属性のうち“活発な”という属性値の確率が0.3から0.5に上っている。それに対して、「女の子のような男の子」では、“大人しい”という属性値の確率が上っている。これは確率分布の非対称性である。

5 おわりに

修飾表現の意味は、修飾語を条件とする、被修飾語の意味の条件付き確率であると仮定した。それに基づき、(他の修飾表現に対する)ニュース性と(被修飾語に対する)明瞭性とを定義した。次に、属性比喩の意味が喩詞と被喩詞の意味から合成できると仮定した。そして、属性比喩では、ニュース性と明瞭性の観点のうち、明瞭性が重要であることを示し、属性比喩の特徴のうちで、属性の選択的強調と意味の非対称性とを明瞭性により説明した。

本モデルの妥当性を実証することが今後の課題である。

参考文献

[1] 諏訪正樹、岩山真：“比喩の計算モデル”, 情報処理 Vol.34, No.5 (1993).

[2] 内山将夫、板橋秀一：“視点を考慮した比喩の理解”, 情処研報 95-NL-105-8, (1995)

[3] 白井賢一郎：「形式意味論入門—言語・論理・認知—の世界」, 東京：産業図書, (1985).

[3] “いかつい男のメガネの色”というような連体修飾表現もここに位置するであろう。この場合には、メガネの色は黒である確率が高くなると思う。

計算機によるトートロジーの意味理解 ～検出機構の検討～

滝澤 修
taki@crl.go.jp

郵政省通信総合研究所 関西先端研究センター
（651-24 神戸市西区岩岡町岩岡588-2）

１．はじめに

言外の意味を含む修辞を計算機に理解させる手法を確立することは，自然言語処理における重要な課題である．そのような修辞の一つに，「美しいものは美しい」「ゴミはゴミだ」などのトートロジー（tautology 同語反復，同義循環）がある．トートロジーは，自然な表現として日常の自然言語に多く出現する一方，文法に沿わない省略や暗黙の指示関係などを伴う場合が多いため，通常の自然言語処理では手に負えない場合が多い（注１）．そのため，自然言語理解研究の一つとして，トートロジーを検出（自動抽出）および理解する手法を検討する必要がある．しかしトートロジーは従来，言語学，心理学の分野における興味の対象であったにとどまり，自然言語処理の観点からあまり注目されていなかった．

本稿では，計算機によるトートロジー理解へのアプローチとして，検出機構の実現のために必要な事柄を検討する．

２．トートロジー検出の研究

2.1 検出と理解とを分けることの妥当性

トートロジー以外の代表的な修辞として，比喩（隠喩）がある．比喩的な表現とリテラルな表現とは本来連続的なものであり，理解して初めて比喩であることがわかる（検出される）．従って比喩の場合，検出と理解とを切り離すことは困難である．それに対してトートロジーは，同じ語の反復という明確な表層構造をしていることから，意味理解を行わずに検出だけを行うことが比較的容易と考えられる．従って検出と理解とをある程度切り分けて扱うことができる．

トートロジーを計算機によって意味理解するためには，トートロジーをまず検出し，次にその検出部分だけを独立した意味理解機構において別処理する，という処理の流れが考えられる．つまりトートロジー検出機構は，高度な自然言語理解機構における前処理に位置づけられる．

2.2 トートロジーの定義と，検出の研究方法について

トートロジーの厳密な定義は言語学の分野においても見あたらない．そのため「トートロジー」という呼称が指す言語表現の範囲は，言語研究者によってマチマチなのが現状である．例えば中村[1]は，「一切絶対ない」のような「不注意な重複表現」のこととしており，佐山ら[2]は，「ゴミはゴミだ」のような，中村がいう同義循環に相当する表現のこととしている．そこでまず，自然言語理解におけるトートロジーの定義を明確に与える必要がある．

筆者らがトートロジーを自然言語理解の対象にしている理由は，通常の意味理解手法では手に負えないような言語表現を減らすことにある．従って自然言語理解の立場からトートロジーを概念的に定義すると，「同語が反復される修辞のうち，通常の意味理解手法では手に負えないもの」のようになる（注２）．従って，トートロジーの定義を決めるためには，「通常の意味理解手法」を想定しておく必要があることになる．

以上に基づき，筆者らはトートロジー検出の研究の進め方として，表1に示した4つのステップを踏むことを考えた．すなわち，自然言語理解におけるトートロジーの定義を与えることを第3ステップまでに行い，それに基づいて第4ステップにおいてトートロジー検出機構を構築する（注3）．次章で，第1ステップの検出機構について検討した結果を述べる．

表1　トートロジー検出機構の研究の進め方

【第1ステップ】
トートロジーである可能性のある言語表現を，極力漏らさず検出する機構を作る．

【第2ステップ】
その機構を用いて，ある程度の規模のコーパスからトートロジー（である可能性のある言語表現）を自動抽出する．

【第3ステップ】
その抽出結果を検討し，想定した意味理解手法を参考にして，トートロジーの範囲（定義）を決める．

【第4ステップ】
その定義に従ったトートロジーのみを検出する機構を構築する．

3．トートロジー検出機構の検討

3.1　第1ステップにおけるトートロジーの条件

第1ステップにおけるトートロジーが満たす条件として，以下の3つを設定する．

(1) 反復語の存在
一文中に同じ語が近接して2つ出現していることを，トートロジーであるための最低条件とする（注4）．この語を「反復語」と呼ぶ．反復語を記号「R」で表すと，トートロジーは，「R～R」のように定式化できる．この「～」は，反復語にはさまれた語句を表す．この「R～R」をトートロジーの「出現形式」とする．

(2) 反復語の品詞
反復語は，名詞，形容詞，動詞など，それだけで意味をもつ語，すなわち自立語に限定される．助詞などの付属語は反復語にはならない．

(3) 反復語の基本形の一致
反復語は，例えば「送るだけ送る」において前出と後出の反復語が共に「送る」であるように，形態的に完全に一致する場合と，例えば「送るだけ送って下さい」において前出が「送る」で後出が「送って」であるように，基本形が一致しても活用変化などのため形態的には不一致の場合とがある．トートロジーは同義を循環させることによって含意するという意味上の効果を狙った言語表現であるので，活用変化などの意味的に無関係な要因は，トートロジーであるか否かには無関係と考えられる．そこで，反復語は基本形が一致することを条件とする．

これらの条件を満たす言語表現を，第1ステップにおけるトートロジーとする．

3.2　トートロジーの出現形式

3.1の条件に従うトートロジーの例をいくつか検討した結果，表2に示す出現形式が得られた．これらの出現形式の中には，反復語の間に副詞などの挿入が許容される場合がある（例「勉強する時はさすがに勉強するよ」，「法律はあくまでも法律だ」）．

表2　トートロジーの出現形式の例

出現形式	Rの品詞	文例
RはR	名詞など	ゴミはゴミだ 契約は契約なんだから それはそうだ それはそれとして おまじないはおまじないでも，よく効くのだ うちはうちで勝手にやる 大学は大学なりの問題を抱えている
RがR	名詞など	歳が歳だから
RもR	名詞など	歳も歳だから
RにnotR	動詞など	寝るに寝られない
RにR	名詞，動詞など *1	走りに走る 乗りに乗る
RにはR	動詞など	あるにはあった
R(な)らR	名詞	駄目だった(な)ら駄目（だった）でいいや 駄目なら駄目でいいや
RというR	名詞など	建物という建物が倒壊した
RといえばR	名詞，動詞，形容詞など	あるといえばある，ないといえばない *2 当り前といえば当り前だ
RといえるR	名詞など	成果といえる成果が無い
RらしいR	名詞など	成果らしい成果が無い
RことはR	動詞，形容詞など	あることはある，ないことはない *3 知らせることは知らせる *4
RものはR	形容詞など	やっぱり，美しいものは美しい
R時はR	動詞など	私だって，勉強する時は勉強するよ
R場合はR	動詞など	落ちる場合は落ちる
RだけR	動詞など	送るだけ送って下さい 聞くだけ聞いたら？

［備考］

*1 反復語は前出が名詞，後出が動詞

*2 反復語「ない」は，両方とも形容詞

*3 反復語「ない」は，前出は形容詞，後出は助動詞 → *2との違いに注意

*4 この文の解釈は以下の通り少なくとも2つ考えられる

解釈①　知らせるべきことは知らせるようにしろよ

解釈①の例：「知らせるのを忘れていました」に対する返答
「知らせることは知らせろよ」

解釈②　一応知らせるが，多分無駄だろう

解釈②の例：「彼にも知らせておいてくれ」に対する返答
「知らせることは知らせるけどね」

3.3 構想中の検出機構（第１ステップ）の概要

検出機構には，語の同定，すなわち形態素解析がまず必要となる．次に，トートロジーが「Ｒ～Ｒ」の出現形式であることを利用し，予め定めておいた出現形式と入力文とを照合する．その際，出現形式毎に反復語の品詞に関する制約を設ける（例えば表２に例示したＲの品詞に限定するとか）．また反復語の照合は，3.1の(3)で述べたように，活用形などでなく基本形で行う．

現在，出現形式の増強と制約条件等の分析，および照合方式の検討を進めている．なお，形態素解析はＪＵＭＡＮ[3]を用いている．

４．考察

入力文と出現形式との単純な照合だけでは，明らかにトートロジーでない言語表現まで検出してしまう場合がある．例えば出現形式「ＲはＲ」との照合をとると，「彼は彼の妻に電話した」も検出されてしまうことになる．これを避けるためには，単語間の構文上の関係なども条件に加えた機構が必要と思われる．

５．おわりに

計算機によるトートロジー理解へのアプローチとして，検出機構の実現のために必要な事柄を検討した．そしてトートロジーの出現形式を調査した．今後，出現形式の増強と制約条件等の分析，および照合方式の検討を進めていく．

【参考文献】

[1] 中村明:「日本語レトリックの体系」,p.144, 岩波書店(1991).
[2] 佐山公一,阿部純一:「日本語同語反復文の意味解釈」, 心理學研究, Vol.65, No.1, pp.25-33 (1994).
[3] 松本裕治他:「日本語形態素解析システムＪＵＭＡＮ使用説明書」, 京都大学工学部長尾研究室 (1993).
[4] R.W.Gibbs,Jr. and N.S.McCarrell: "Why Boys Will Be Boys and Girls Will Be Girls: Understanding Colloquial Tautologies", J. Psycholinguistic Research, Vol.19, No.2, pp. 125-145 (1990).

（注１）機械翻訳などでは，名詞述語文のように構文構造の単純なトートロジーに対しては，構文レベルでの逐語訳によってある程度の対応が可能である．例えば「男の子は男の子だ」に "A boy is a boy" という訳を与えて処理を終了しても問題は無いように思える．但しこの逐語訳として "A boy is a boy", "Boys will be boys", "Boys are boys" などが考えられるが，これらはそれぞれ理解し易さや受け取られる意味が異なると指摘されている[4]．このような異言語間のトートロジー理解の違いに対応するためには，逐語訳では限界があるといえる．

（注２）この定義自体がトートロジーかも知れない ^^;

（注３）表１の第２ステップにおいて，トートロジーがコーパス中にどれ位の頻度で出現するかを知ることもでき，トートロジーの計算機処理研究の意義を示すデータも得られることが期待できる．

（注４）「男はオスだ」のような同義循環も意味的にはトートロジーに含めるべきであるが，現段階では対象外とする．

動的意味論を用いた漸進的談話解釈モデル

松原 茂樹† 外山 勝彦‡ 稲垣 康善†

†名古屋大学工学部 ‡中京大学情報科学部

1 はじめに

漸進的解釈 (incremental interpretation) とは, 文あるいは談話を語単位でその出現順序に従って順次解釈することをいう. 人間が言語理解過程において漸進的解釈を行なうことは心理学の分野で実証されている [1], [12]. また, 計算言語学の分野でも, 漸進的解釈の計算モデルがいくつか提案されており, 自然言語の曖昧性を効率的に解消する方法を与えるのに用いられている [5], [8]. これらのなかで, 漸進的解釈の必要性は対話理解において一層顕著である [10]. しかしながら, これまで提案された漸進的解釈の枠組は談話を対象とするものでなく, それらの枠組でもって対話理解をモデル化することは難しく, 談話の漸進的解釈に関する基礎理論の確立が重要な課題として望まれている.

一方, 談話解釈には動的解釈が不可欠である [13]. 動的解釈とは, 言語表現と文脈との間の相互作用, すなわち, 言語表現の意図している意味が文脈に依存して決定され, かつ, その意味により新たな文脈が形成されることをいう. 近年, その重要性の認識に伴い, このような自然言語の動的解釈に関する研究は活発におこなわれている. なかでも動的意味論 (dynamic semantics) は, 論理式の意味を情報の更新と捉える意味論であり [4], 遷移系 (transition system) との関連で捉えることができる [2]. すなわち, 遷移系は状態の集合と状態間の到達可能関係との組で定義されるため, 論理式の動的な意味を遷移系における到達可能関係と見なせる. よって, 論理式を解釈するごとに状態が遷移するモデルを構成できる. このモデルにおいて状態を文脈と見なすと, 文脈に依存して論理式が解釈され, かつ, その解釈によって新たな文脈に遷移するという動的解釈を実現できる. さらに, 語に対して到達可能関係を規定すれば漸進的解釈が実現可能となる.

このような考え方に基づいて, 本報告では, 動的意味論を用いた談話の漸進的解釈モデルを提案する. 具体的には, 語に動的意味論を与えその上で遷移系を構成し, 談話を漸進的にかつ動的に解釈することにより, 談話を理解しながら文脈と相互作用することが可能なモデルを提案する.

本論文の構成は以下の通りである. まず2節では, 本報告で提案するモデルの基本的アイデアについて述べる. 3節で動的意味論を用いた漸進的解釈のモデル化について説明し, 4節で関連研究との比較を与える.

2 動的意味論と漸進的解釈

本節では, 動的意味論 [4] が漸進的解釈とどのように関連するのかについて述べる. 2.1節では, その準備として動的意味論とその遷移系について, 例を用いて説明する. なお, 動的意味論とよばれる枠組には, 談話表示理論 (Discourse Representation Theory:DRT)[7], 更新の論理 (Update Logic)[4], 動的述語論理 (Dynamic Predicate Logic: 以下 DPL)[3] などがあるが, 本報告では DPL を取り上げる. また, 2.2節では DPL を用いたときの漸進的解釈の問題点を指摘し, それに対して動的意味論を用いて漸進的解釈をモデル化するための基本的アイデアを示す.

2.1 動的意味論と遷移系

次の (1) は文にまたがる照応関係を含む談話である.

$$A\ man\ walks.\ He\ whistles. \quad (1)$$

談話 (1) を1階述語論理式へ構成的に変換すると,

$$\exists x(man(x) \wedge walk(x)) \wedge whistle(x) \quad (2)$$

となる. 以下では, 論理式 (2) の DPL における解釈について述べる [3].

まず, 2項組 $M = \langle D, F \rangle$ を構造という. ここで, D は個体の集合, F は n 引数述語記号 p に対して $F(p) \subseteq D^n$ を対応づける写像である. また, 付値関数 g は変数 x に個体 $g(x) \in D$ を割り当てる関数であり, $h[x]g$ で付値関数 h と g とでは変数 x 以外のすべての変数について同じ値を割り当てることを意味する. 以下のように構造ならびに付値関数によって論理式の意味を定める.

定義 2.1 G をすべての付値関数の集合とするとき, 論理式 ϕ に対して, $[\phi]_M \subseteq G \times G$ を対応づける意味評価 $[\ ]_M$ を次に定義する. 以下では添字 M を省略する.

$$[p(x_1,..,x_n)]=\{\langle g,h\rangle | h=g, \langle h(x_1),..,h(x_n)\rangle \in F(p)\}$$
$$[\phi \wedge \psi]=\{\langle g,h\rangle | \exists k : \langle g,k\rangle \in [\phi], \langle k,h\rangle \in [\psi]\}$$
$$[\exists x\phi]=\{\langle g,h\rangle | \exists k : k[x]g, \langle k,h\rangle \in [\phi]\}$$

定義 2.1 より, 論理式 (2) の意味は次のようになる.

$$[\exists x(man(x) \wedge walk(x,y) \wedge whistle(x)] = \{\langle g,h\rangle | h[x]g, h(x) \in F(man), h(x) \in F(walk), h(x) \in F(whistle)\} \quad (3)$$

論理式 (2) において $whistle(x)$ の変数 x は自由である. しかし, DPL の意味論によれば, $walk(x)$ の変数 x と同一の値が割り当てられる. これは照応が解決されたことを示す.

次に, 動的意味論に基づく遷移系について述べる [2]. 遷移系は論理式の集合をΦとするとき, 2項組 $T=\langle 2^G, \{[\phi]\}_{\phi\in\Phi}\rangle$ で定義される. $q\subseteq G$を状態とよぶとき, 遷移系Tは論理式の解釈による状態遷移を規定する.

定義 2.2 状態$q\subseteq G$のもとで論理式$\phi\in\Phi$を解釈したとき, 遷移後の状態$[\phi](q)\subseteq G$は遷移系Tにより,

$$[\phi](q)=\{h|g\in q, \langle g,h\rangle\in[\phi]\} \quad (4)$$

で与えられる. 従って, 各論理式に対しては次のように規定できる.

$[p(x_1,..,x_n)](q)=q\cap\{g|\langle g(x_1),..,g(x_n)\rangle\in F(p)\}$
$[\phi\wedge\psi](q)=[\psi]([\phi](q))\}$
$[\exists x\phi](q)=[\phi](\bigcup_{g\in q}\{h|h[x]g\})$

状態$q\subseteq G$で論理式(2)を解釈したとき, 遷移後の状態は定義2.2より,

$$\begin{aligned}&[\exists x(man(x)\wedge walk(x))\wedge whistle(x)](q)\\ &=\bigcup_{g\in q}\{h|h[x]g, h(x)\in F(man),\\ &\quad h(x)\in F(walk), h(x)\in F(whistle)\}\end{aligned} \quad (5)$$

となる. このように論理式の意味を付値関数間の関係と規定することによって, 論理式の解釈による状態遷移をモデル化できる.

2.2 動的意味論による漸進的解釈のモデル化

論理式の解釈により状態が遷移するという見方は, 一見, 漸進的解釈に適しているように思われる. ところが, 漸進的解釈では語を解釈するごとに状態が遷移する必要があるのに対して, 2.1節で示したような遷移系が与える状態の遷移は論理式の構造によって決定される. このためMilwardは, 動的意味論が漸進的解釈のための意味論として適切でないと主張している [9]. 以下では, 動的意味論と漸進的解釈との関連性について述べる.

論理式(2)の解釈による状態遷移について, 定義2.2より式(6)が成立する.

$$\begin{aligned}&[\exists x(man(x)\wedge walk(x))\wedge whistle(x)](q)\\ &=[whistle(x)]([\exists x(man(x)\wedge walk(x))](q))\end{aligned} \quad (6)$$

また, 仮に

$$[\exists x](q)=\bigcup_{g\in q}\{h|h[x]g\} \quad (7)$$

と定めると, 定義2.2の$[\exists x\phi](a)$は, $[\phi]([\exists x](q))$と表せる. よって, 式(6)より,

$$\begin{aligned}&[\exists x(man(x)\wedge walk(x))\wedge whistle(x)](q)\\ &=[whistle(x)]([man(x)\wedge walk(x)]([\exists x](q)))\\ &=[whistle(x)]([walk(x)]([man(x)]([\exists x](q))))\end{aligned} \quad (8)$$

が成り立つ. いま, 語wの意味を$[\![w]\!]$で表す. 談話(1)を構成する語の意味$[\![a]\!], [\![man]\!], [\![walks]\!], [\![whistles]\!]$を, それぞれ, $[\exists x], [man(x)], [walk(x)], [whistle(x)]$とする. このとき式(8)より,

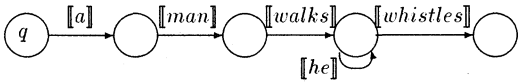


図 1: 談話(1)に対する状態遷移図

$$\begin{aligned}&[\exists x(man(x)\wedge walk(x))\wedge whistle(x)](q)\\ &=[\![whistles]\!]([\![walks]\!]([\![man]\!]([\![a]\!](q))))\end{aligned} \quad (9)$$

となる. これは, 語に動的意味論を与えることによって語の解釈による状態遷移が実現可能となることを示唆する(図1).

これらの考察から, 次節では, 語に対する動的意味論を規定し, その上で遷移系を構成することにより, 談話の漸進的解釈の形式的モデルを提案する.

3 談話の漸進的解釈モデル

本節では, 談話の漸進的解釈モデルについて述べる. まず, 3.1節で語に対する意味論を示す. 3.2節では漸進的解釈モデルを定義する.

3.1 語の意味

以下では, 自然言語の単語を語といい, そのすべての集合をWとする. Wには, コンマ(","), ピリオド(".")も含まれる.

定義 3.1 語はその範疇により次のように分類される. なお, 範疇Aに対して$B_A\subseteq W$でもって範疇Aの語の集合を示す.

(1)B_{CN}(普通名詞)= $\{man, boy, car, book, \ldots\}$
(2)B_{TN}(固有名詞)= $\{Sue, John, Mary, \ldots\}$
(3)B_{PN}(代名詞)= $\{she, him, it, \ldots\}$
(4)B_{IV}(自動詞)= $\{walks, whistles, ran, \ldots\}$
(5)B_{TV}(他動詞)= $\{loves, owns, saw, \ldots\}$
(6)B_{AD}(形容詞)= $\{new, pretty, happy, \ldots\}$
(7)B_{AR}(不定冠詞)= $\{a, an\}$
(8)B_{AE}(定冠詞)= $\{the\}$
(9)B_{CP}(コンマ, ピリオド)= {",", "."}

次に語の動的意味論を定義する.

定義 3.2 次のような組$M=\langle D, I\rangle$を**構造**という. ただし,

(1) Dは個体の空でない集合である.
(2) IはWを定義域とする写像で, $w\in B_{CN}\cup B_{TN}\cup B_{PN}\cup B_{IV}\cup B_{AD}$に対して$I(w)\subseteq D$, $w\in B_{TV}$に対して$I(w)\subseteq D\times D$.

このような構造でもって語に意味を与えるため, 変数, ならびに, 変数にDの元を対応づける付値関数を用いる. 以下では, 変数の集合をV, 付値関数の集合をG

とする．また，変数の空でない系列 $t \in V^n$ と付値関数 $g \in G$ との対 (t,g) を**状況**という．

例 3.1　x,y を変数，g,h を付値関数とするとき
$(x,g),(xy,h),(xyx,g)$
は状況である．

定義 3.3　$S \subseteq V^n \times G$ を状況の集合とする．また，$Range(g) \subseteq D$ で付値関数 g の値域を示すものとする．このとき，構造 M のもとで語 $w \in W$ に意味を与える写像 $[\![\ ]\!]_M : W \longrightarrow \mathcal{P}(S \times S)$ を次のように定義する．なお，以下では，構造を示す添字 M を省略する．

(1) $[\![w \in B_{CN}]\!]$
$=\{\langle(x,g),(y,h)\rangle | y = x, h = g, h(y) \in I(w)\}$
$\cup\{\langle(xy,g),(z,h)\rangle | z = y, h = g, h(z) \in I(w)\}$

(2) $[\![w \in B_{TN}]\!]$
$=\{\langle(x,g),(y,h)\rangle | y = x, h[y]g, h(y) \in I(w)\}$
$\cup\{\langle(xy,g),(z,h)\rangle | z = y, h[z]g, h(z) \in I(w)\}$

(3) $[\![w \in B_{PN}]\!]$
$=\{\langle(x,g),(y,h)\rangle | y = x, h = g,\}$
$\cup\{\langle(xy,g),(z,h)\rangle | z = y, h = g\}$

(4) $[\![w \in B_{IV}]\!]$
$=\{\langle(x,g),(y,h)\rangle | y = x, h = g, h(y) \in I(w)\}$
$\cup\{\langle(xy,g),(z,h)\rangle | z = y, h = g, h(z) \in I(w)\}$

(5) $[\![w \in B_{TV}]\!]$
$=\{\langle(x,g),(xy,h)\rangle | h = g, \langle h(y), h(z)\rangle \in I(w)\}$
$\cup\{\langle(xy,g),(xy,h)\rangle | h = g, \langle h(z), h(v)\rangle \in I(w)\}$

(6) $[\![w \in B_{AJ}]\!]$
$=\{\langle(x,g),(y,h)\rangle | y = x, h = g, h(y) \in I(w)\}$
$\cup\{\langle(xy,g),(z,h)\rangle | z = y, h = g, h(z) \in I(w)\}$

(7) $[\![w \in B_{AR}]\!]$
$=\{\langle(x,g),(y,h)\rangle | y = x, h[y]g\}$
$\cup\{\langle(xy,g),(z,h)\rangle | z = y, h[z]g\}$

(8) $[\![w \in B_{AE}]\!]$
$=\{\langle(x,g),(y,h)\rangle | y = x, h[y]g, h(y) \in Range(g)\}$
$\cup\{\langle(xy,g),(z,h)\rangle | z = y, h[z]g, h(z) \in Range(g)\}$

(9) $[\![w \in B_{CP}]\!]$
$=\{\langle(x,g),(y,h)\rangle | y \neq x, h = g\}$
$\cup\{\langle(xy,g),(z,h)\rangle | z \neq y, h = g\}$

このように語の意味を状況間の関係と定める．各語の意味は，語の解釈による状況から状況への遷移関係を示す．

例 3.2　語 Sue, loves, a, man の意味を次に示す．
$[\![Sue \in B_{TN}]\!]=\{\langle(x,g),(x,h)\rangle | h[x]g, h(x) \in I(Sue)\}$
$\cup\{\langle(xy,g),(y,h)\rangle | h[x]g, h(x) \in I(Sue)\}$
$[\![loves \in B_{TV}]\!]=\{\langle(x,g),(xy,g)\rangle | \langle g(x), g(y)\rangle \in I(loves)\}$
$\cup\{\langle(xy,g),(xy,g)\rangle | \langle g(x), g(y)\rangle \in I(loves)\}$
$[\![a \in B_{AR}]\!]=\{\langle(x,g),(x,h)\rangle | h[x]g\}$
$\cup\{\langle(xy,g),(y,h)\rangle | h[y]g\}$
$[\![man \in B_{CN}]\!]=\{\langle(x,g),(x,g)\rangle | g(x) \in I(man)\}$
$\cup\{\langle(xy,g),(y,g)\rangle | g(y) \in I(man)\}$

3.2　漸進的解釈モデル

以下では，談話の漸進的解釈モデルを定義する．また，モデルを用いた談話解釈例を示す．

定義 3.4　$S \subseteq V^n \times G$ を状況の集合，$[\![w]\!]$ を語 $w \in W \subseteq S \times S$ の意味とする．このとき，2 項組 $T = \langle 2^S, \{[\![w]\!]\}_{w \in W}\rangle$ を**漸進的解釈モデル**という．

以下では，状況の集合 $q \subseteq S$ を**状態**という．漸進的解釈モデルによる状態遷移は次のように定義される．

定義 3.5　$q \subseteq S$ を状態とする．状態 q で語 $w \in W$ を解釈したとき，遷移後の状態 $[\![w]\!](q)$ は漸進的解釈モデル T により，
$[\![w]\!](q) = \{j | i \in g, \langle i,j\rangle \in [\![w]\!]\}$
で与えられる．

$[\![w]\!](q)$ は語 w を解釈した時点での文脈を表現する．

例 3.3　照応関係を含む談話 Sue loves a man. She saw him の漸進的解釈例を示す[1]．なお，解釈前の状態を $q_0 = \{(x,f) | \exists v : f(v) \in F(man)\}$ とする．状態 q_0 は男の人がいるという文脈を表現する．

$q_1 = [\![Sue]\!](q_0)$
$=\{(x,g) | \exists v \exists f : g[x]f, g(v) \in F(man), g(x) \in F(Sue)\}$

$q_2 = [\![loves]\!](q_1)$
$=\{(xy,g) | \exists v \exists f : g[x]f, g(v) \in F(man), g(x) \in F(Sue), \langle g(x), g(y)\rangle \in F(loves)\}$

$q_3 = [\![a]\!](q_2)$
$=\{(y,h) | \exists v \exists x \exists f : h[x,y]f, h(v) \in F(man), h(x) \in F(Sue), \langle h(x), h(y)\rangle \in F(loves)\}$

$q_4 = [\![man]\!](q_3)$
$=\{(y,h) | \exists v \exists x \exists f : h[x,y]f, h(v) \in F(man), h(x) \in F(Sue), \langle h(x), h(y)\rangle \in F(loves), h(y) \in F(man)\}$

$q_5 = [\![.]\!](q_4)$
$=\{(z,h) | \exists v \exists x \exists y \exists f : h[x,y]f, h(v) \in F(man), h(x) \in F(Sue), \langle h(x), h(y)\rangle \in F(loves), h(y) \in F(man)\}$

$q_6 = [\![he]\!](q_5)$
$=\{(x,h) | \exists v \exists y \exists f : h[x,y]f, h(v) \in F(man), h(x) \in F(Sue), \langle h(x), h(y)\rangle \in F(loves), h(y) \in F(man)\}$

$q_7 = [\![saw]\!](q_6)$
$=\{(xu,h) | \exists v \exists y \exists f : h[x,y]f, h(v) \in F(man), h(x) \in F(Sue), \langle h(x), h(y)\rangle \in F(loves), h(y) \in F(man), \langle h(x), h(u)\rangle \in F(saw)\}$

[1] 本報告では，照応解決手法については述べない．詳細は [11] を参照されたい．

q_8 $=[\![him]\!](q_7)$
$=\{(y,h)|\exists v \exists y \exists f : h[x,y]f, h(v) \in F(man), h(x) \in F(Sue), \langle h(x), h(y)\rangle \in F(loves), h(y) \in F(man), \langle h(x), h(y)\rangle \in F(saw)\}$

ここで提案したモデルは語の解釈による状態遷移を与えるものであり, 談話を動的にかつ漸進的に解釈することができる.

4 従来の関連研究との比較

これまで提案されてきた漸進的解釈の計算モデルは, 主に, 自然言語の曖昧性解消に用いられた[6]. 統語的に曖昧な文に対してすべての可能な統語解析結果を求める方法では, 排除されるであろう多くの解析結果を生成することになり, 効率が悪い. よって, できる限り早い段階で曖昧性を解消するために漸進的解釈を用いる. Mellishは曖昧性を早期に解消するモデルを提案した[8]. ところがその枠組は厳密に左から右へ解釈が進行するモデルではなかった. そこでHaddockは, Mellishの枠組を洗練し, 定名詞句を漸進的に解釈するモデルを提案した[5]. Haddockのモデルでは, 語を解釈するごとに名詞句の指示対象に関する制約を生成, 蓄積し, 制約充足処理をおこなう. 指示対象が唯一に定まったとき, そこが名詞句の終了時点であるとみなす. よって, 名詞句の後に続く前置詞句が, 動詞を修飾するのか, あるいは, 名詞句を修飾するのかという曖昧性を早期に解消できる[1]. ただし, 制約充足処理は与えられた文脈に依存するが, 新たな文脈を構成することはない. よって, Haddockのモデルは動的解釈を実現していない.

それに対して本報告で提案した漸進的解釈モデルは, 談話理解で必要不可欠な動的解釈を実現する. 例えば, 不定名詞句の解釈によって新たな対象を文脈に導入することができる. すなわち, 文脈を動的に扱うことができる.

一方, Milwardはλ計算を用いて漸進的解釈をおこなうモデルを提案した[9]. Milwardのモデルでは, 状態をλ項で表現し, また, 語の意味もλ項で表す. 状態遷移は状態表現と語の意味表現とのλ変換に基づく. しかし, Milwardの枠組での解釈対象は一文の範囲に限られており, 談話を解釈することはできない. さらに, 不定名詞句, あるいは, 定名詞句を含む文に対する漸進的解釈については明らかでない.

それに対して, 本報告の漸進的解釈モデルは談話を対象とする. また, 不定名詞句および定名詞句についても妥当な解釈を与える.

5 むすび

柔軟な対話理解をモデル化するために, 談話の漸進的解釈に関する基礎理論の確立が望まれる. また, 談話理論としては, 解釈が動的なものが要求される.

本報告では, 以上の認識に基づき, 動的意味論を用いた談話の漸進的解釈の形式的モデルを提案した. 本モデルは, 談話を語の連鎖と捉え, 談話理解過程を語の動的解釈過程と見なすことで, 談話の漸進的解釈を実現する. このような観点から見て, 本報告で述べたモデルは談話の漸進的解釈に関する基礎理論として位置付けられる.

なお本報告で想定する談話とは, 情報が単調に増加するものである. しかし, 本来, そのような範疇に属さない談話も多くみられる. また実際, 人間が漸進的に解釈できない談話も存在する. このような談話をも含めた言語理解のモデル化は今後の課題である.

参考文献

[1] Altmann, G. T. M. and Steedman, M. J.: Interaction with Context During Human Sentence Processing. Cognition, **30**, pp. 191-238 (1988).

[2] Fernand, T.: Transition Systems and Dynamic Semantics, *LNAI*, **633**, pp. 232-251 (1992).

[3] Groendijk, J. and Stokhof, M.: Dynamic Predicate Logic, *Linguistics and Philosophy*, **14**, pp. 39-100 (1991).

[4] Groendijk, J. and Stokhof, M.: Two Theories of Dynamic Semactics, *LNAI*, **478**, pp. 55-64 (1990).

[5] Haddock, N. J.: Incremental Interpretation and Combinatory Categorial Grammar, *IJCAI'87*, pp. 661-663 (1987).

[6] Haddock, N. J.: Computational Models of Incremental Semantic Interpretation, Language and Cognitive Processes, **4**, (3/4), pp. 337-368 (1989).

[7] Kamp, H.: Discourse Representation Theory, What it is and Where it Ought to Go, Natural Language at the Computer., *LNCS*, **320**, pp. 85-111 (1988).

[8] Mellish, C. S.: *Computer Interpritation of Natural Language Descriptions*, Ellis Horwood Limited (1985). 田中 穂積(訳): 自然言語意味理解の基礎, サイエンス社 (1987).

[9] Milward, D.: Dynamics, Dependency Grammar and Incremental Interpretation, *COLING 92*, pp. 1095-1099 (1992).

[10] Milward, D. , and Cooper, R.: Incremental Interpretation: Applications, Theory, and Relationship to Dynamic Semantics, *COLING 94*, pp. 748-754 (1994).

[11] 松原, 外山, 稲垣: 発話の連続性に基づく談話理解モデルを用いた照応解決, 人工知能学会第8回全国大会, pp. 619-622 (1994,6).

[12] Sanford, A. J. and Garrod, S. C.: What, When, and How?:Quentions of Immediacy in Anaphoric Reference Resolution, *Language and Cognitive Processes*, **4**, No. 3/4, pp. 235-262 (1989).

[13] 白井 賢一郎: 自然言語の意味論 - モンタギューから状況への展開 -, 産業図書 (1990).

日本語マニュアル文における条件表現の語用論

森 辰則　　龍野弘幸　　中川 裕志

横浜国立大学 工学部 電子情報工学科
{mori,tatsuno}@forest.dnj.ynu.ac.jp, nakagawa@naklab.dnj.ynu.ac.jp

1 はじめに

今日，家庭向けの電化製品から，ビジネス向けの専門的な機器まであらゆる製品にマニュアルが付属している．これらの機器は，複雑な操作手順を必要とするものも多い．これを曖昧性を含まずに記述することが，マニュアルには求められている．また，海外向けの製品などのマニュアルで，このような複雑な操作手順を適切に翻訳することも困難である．我々は，このような問題の解決の基礎となる，マニュアル文の計算機による理解の手法について検討している．

一般に，表層表現から直接得られる情報には曖昧性が含まれる．特に，日本語では主語が頻繁に省略されるため，ゼロ代名詞の適切な指示対象を同定することが必要になる．

従来の研究は [安部 88] など記述対象や事象に関する領域知識を利用して，意味表現の表す物事に関する推論により，意味表現の未決定部分を決定するという方向であった．しかし，この方法を用いるには，大規模な領域知識ないしは常識知識をあらかじめ備えておく必要があるが，現在そのような領域・常識知識ベースは存在していない点が問題である．したがって，この問題に対処するためは，個別の領域知識にほとんど依存しない情報を用いることが必要となる．本稿では，この個別の領域知識にほとんど依存しない情報として，言語表現自身が持つ意味によって，その言語表現がマニュアル文に使用される際に顕在化する制約について考察する．この方法では，言語表現から知識表現に写像する過程で曖昧性を減らすことになる．ここで重要な点は，本稿での考察が個別のマニュアルが記述している個別領域を問題にしているのではなく，マニュアル文でありさえすれば，分野や製品を問わずいかなるマニュアル文にも通用する制約について考察しようとしている点である．

本稿では，特に，ゼロ代名詞の指示対象同定問題に対して，マニュアル文の操作手順においてしばしば現れる条件表現の性質を利用することを提案する．すなわち，マニュアルの操作手順に現れる条件表現についてその語用論的制約を定式化し，主に主語として現れる動作主に対応するゼロ代名詞の指示対象同定に応用する．

本稿で考える「動作主」は，動作の主格になるいわゆる動作主だけではなく，状態などの対象も含まれる．そこで，以下では [仁田 93] のいう「主 (ぬし)」とする．'主' はより広い分類であり，a) 対象に変化を与える主体，b) 知覚，認知，思考などの主体，c) 事象発生の起因的な引き起こし手，d) 発生物，現象，e) 属性，性質の持ち主を含む．

2 マニュアル文における基本的制約

まず，マニュアルを構成する最も基本的なオブジェクトおよびその言語的な役割は大別すると次のようになる．

制約 1 (言語的役割とオブジェクト)

話し手 メーカー (マニュアルライター).

聞き手 マニュアルの読み手である利用者.

第三者 システムの全体もしくは一部.

これらを考慮するとマニュアル文で用いられる人称は次のようになる．

制約 2 (人称)
一人称 = メーカー, 二人称 = 利用者, 三人称 = システム

次に，基本的な表現形式についての考察をする．マニュアルの基本的な構成は説明の仕方の説明，操作手順の説明，アフターサービスに関する説明等からなる．これら各々の文脈に現われる文は性質が異なる．操作手順の説明では，話し手の動作は既に完了しているが，説明の仕方の説明，アフターサービスに関する説明では，その限りではない．そこで，以下の考察では，マニュアルの主要部である操作手順の説明に現れる場合を考える．

3 条件表現の‘主’に関する制約

日本語の条件表現には，「れば」「たら」「なら」「と」があり，基本的性質は異なる[益岡 93]我々が調べた範囲では，マニュアル文では「たら」「なら」はあまり用いられていなかった．また，「れば」に比べて，「と」の出現頻度が高かった．以下の節ではそれぞれの場合について考察する．

3.1 「と」文の‘主’制約

[久野 73]によると，接続助詞「と」について，前件は先行条件を表し，後件は，その当然の結果，習慣的な結果，或いは不可避な結果を表すとある．また，[益岡 93]によると，「と」が未然の事態を表す場合，後件の事態が前件の事態に連動して起こるという意味において前件と後件の二つの事態が一体の事態であることが強調されている．このような性質から，「と」の後件は，命令・要求・決意を表せないとされる．よって，後件には基本的に事実の叙述や判断，推量の表現のみが許される．また，基本的にはマニュアル文では確実な物事のみを述べるものであり，物事の不確実さを表すような話し手の態度，特に判断，推量の表現は現れにくい．したがって，事実叙述のみが後件に現われると考えられる．事実叙述として現われ得るのは，ある動作の記述と，許可表現などによる何らかの動作に関連する状態記述である．

動作の記述を考える際に重要となるのが動詞の意志性，無意志性の問題である．動詞の意志に関する分類として，[IPA87]の分類に基づくと，‘主’が意図的に行ないうる動作を表す意志動詞と，‘主’による意図的な動作を表さない無意志動詞とがある．動詞の命令形が命令を表し，意志・推量形が意志・勧誘を表すものが，意志動詞であり，命令形が願望を表したり，意志・推量形が推量を表すのが無意志動詞である．無意志動詞は，無意志用法のみであるが，意志動詞は，意志用法のみのものと，意志用法，無意志用法の両方に使えるものの2種類がある．無意志動詞としては，「痛む」，無意志用法もある意志動詞としては，「落す」，意志用法のみの意志動詞としては，「捜す」などがある．

まず，意志用法の動詞が後件で使われる場合を考える．「と」文の後件には，先に述べたように依頼，勧誘表現は存在しない．そのため，動作手順の説明では，動詞の基本形つまり「る形」が用いられることがほとんどである．「る形」で動作主が聞き手の場合は，実質的に依頼表現になる．従って，「と」文では後件で依頼を表現できないため，‘主’は聞き手にはなり得ない．また，「と」文では，先に述べたように決意を表すことができない．「る形」で‘主’が話し手の場合意志を表すが，この用法も「と」文では存在しないため‘主’は話し手にはならない．‘主’が第三者の場合，「る形」では，依頼，意志等を表さないので，「と」文の性質には抵触しない．したがって，人称に関する制約より第三者であるシステムが後件の‘主’となる．例えば，

(1) 「取消キーを押すと，文書作成画面に戻ります．」[富士 88]

において，「文書作成画面に戻る」のはシステムである．

無意志用法の場合は，「る形」が意志，命令，依頼等を表さないので，意志用法の場合と異なる振舞いをする．例えば，「触れると，感電します．」の後件の‘主’は利用者になる．

以上に述べたように条件表現では‘主’の意志的動作か否かが重要となる．マニュアルには，もちろん，動作以外の状態記述もある．これらは，‘主’の意志とは無関係である．以下では，これらも含めて，意志用法の動作を「動作」，それ以外を「非動作」と表す[1]．

この分類の下で，「と」に関する制約は次のようになる．

制約 3 (「と」の後件の‘主’制約)
接続助詞「と」による複文構造において，後件の述部が動作であり非過去の場合には，その‘主’は3人称つまりシステムになる．

この制約の検証のために，接続助詞「と」が用いられているマニュアル文例を約400例ほど集め，制約3について調べた．その結果，調べた範囲では，これらの制約に違反する文はなく，制約の妥当性が確認された．

3.2 「れば」「たら」「なら」の使用例についての考察

まず，[益岡 93]による「れば」「たら」「なら」の意味を列挙しよう．

「れば」 時間を越えて成り立つ普遍的因果関係を表す．また，前件が状態表現の場合，仮定的表現に解釈されすい．

[1]具体的には，意志動作を表す動詞，依頼表現，「～てみる」，「～ておく」を「動作」とし，「非動作」を無意志動作を表す動詞(「なる」「ある」など)，可能表現，「～ている」，コピュラ(「～です」など)，形容詞，形容動詞とした．

「たら」 1) 時間の経過にともなって実現することが予想される事態を表すものと，2) 実現するかどうかが定かではないような事態が実現したことを仮定し，それにともなってどのような事態が実現するかを表現するもの，とがある．

「なら」 後件に表現の重点があり，前件を真と仮定した想定のもとで，後件で判断や態度の表明が行なわれる．また，「れば」「たら」に比べて前件と後件のつながりが弱い．

これらからまず分かることは，「と」の場合と異なり，仮定的表現となりうるため，後件で依頼表現が使用可能であることである．「れば」は普遍的因果性を表すため，その後件は前件の発生にともなって必然的に生じる結果であるから，原理的には話し手の態度が介入する余地がない．依頼も話し手の態度のひとつであるから，後件に依頼は現れない．つまり，「れば」の後件には基本的には依頼表現が現れないことを意味する． 後に示す実例文の分析でも「れば」接続の文では後件が依頼，つまり，利用者の動作のものは非常に少ない．ただし，前件が状態表現の場合は仮定的になる，とあることから，その場合は後件に依頼表現が現れる可能性がある．これに該当する例として次のものがある．

(2) 「ウィンドウを見る必要がなければ，ウィンドウをリサイズ・コーナを使用して小さくするのではなくアイコンにして下さい．」 [Sun92, p.63]

一方，前件で仮定が表現される「たら」「なら」の場合は次のように考えられる．仮定すること自体が既に話し手の態度の一種であり，その仮定下での後件の記述にも当然話し手の態度が含まれる．したがって，後件で話し手の態度のひとつである依頼が現れることは可能性が高い，と言える． (2) の「なければ」を「なかったら」や「ないなら」に代えてみれば分かるように，「たら」，「なら」も同様に依頼を表すことができるのことも，上の説明から予想されることである．

以上のように「れば」「たら」「なら」では後件で依頼表現を使用することができることから，「と」と同じ制約を導くことはできない．そこで，「れば」「たら」「なら」の使い分けを考えるために，主節つまり後件を次のような観点から分類する．

まず，操作手順の説明の場合と限定しているので，メーカーの動作は完了していると考えられる．従って，‘主’となりうるオブジェクトは利用者とシステムである．そして，「と」と同様に動作 / 非動作の観点から分類する．この分類より，可能な‘主’と動作 / 非動作の組合せは，「利用者の動作」「利用者の非動作」「システムの動作」「システムの非動作」の 4 つになる．この 4 つの状態をそれぞれの接続助詞で接続すると各々 16 通りの接続が考えられる．以下では，この分類に従って，「れば」，「たら」，「なら」を前件及び後件の性質により分類し考察する．

表に「れば」「たら」「なら」の分類を示す．例文数が 5% を越える組合せに印をつけた．全体を概観すると，「れば」と「たら」「なら」とでは，使用傾向が大きく違うことが分る．「れば」では後件が利用者の動作になりにくい．これは「と」の場合と同じである．つまり (2) の形式の文はほとんど現れないということを表している．逆に「たら」，「なら」では「れば」とは相補的に後件が利用者の動作になりやすい．また，全般的に，前件がシステムの動作である文が非常に少ない．これについては，現在のシステムのほとんどが，利用者の働きかけにより何か他の動作を行なったりある状態に移行したりするからであろう．前件がシステムの状態である文でも，そのシステム状態は利用者の動作に起因するものであるというタイプが多い．

「れば」文の場合，前件が利用者の動作である文が多い．これは，「れば」文の基本的性質である因果関係は，動作の方が表しやすいためと考えられる．さらに，前件がシステムの状態である文も，そのシステム状態は利用者によって引き起こされた結果であるという文が多い．この理由は，動作的側面を残しているため，上記の場合と同様の理由で「れば」で表しやすいからであろう．

3.3 暗黙規則

今までの考察から，「れば」「たら」「なら」についてのマニュアルにおける使用方法に関する傾向が得られた．特に，‘主’に注目すると文型と強い相関があることがわかる．そこで次の暗黙規則を立てることができる．まず，「れば」については，「と」とほぼ同様の分布になるので以下のようになる．

暗黙規則 1 (「れば」の後件の‘主’制約)
接続助詞「れば」による複文構造において，後件は利用者の動作を表さない．つまり，後件の述部が意志用法の動詞の場合には，その‘主’はシステムになる．

		後件				
前件	**れば**	利用者の動作	利用者の非動作	システムの動作	システムの非動作	合計
	利用者の動作	1例 0.4%	65例 *28.9%*	53例 *23.6%*	13例 *5.8%*	132例 58.7%
	利用者の非動作	4例 1.8%	12例 *5.3%*	0例 0.0%	1例 0.4%	17例 7.6%
	システムの動作	0例 0.0%	0例 0.0%	1例 0.4%	4例 1.8%	5例 2.2%
	システムの非動作	6例 2.7%	20例 *8.9%*	38例 *16.9%*	7例 3.1%	71例 31.6%
	合計	11例 4.9%	97例 43.1%	92例 40.9%	25例 11.1%	225例 100.0%

		後件				
前件	**たら**	利用者の動作	利用者の非動作	システムの動作	システムの非動作	合計
	利用者の動作	25例 *42.4%*	0例 0.0%	0例 0.0%	0例 0.0%	25例 42.4%
	利用者の非動作	6例 *10.2%*	1例 1.7%	0例 0.0%	0例 0.0%	7例 11.9%
	システムの動作	8例 *13.6%*	2例 3.4%	0例 0.0%	1例 1.7%	11例 18.6%
	システムの非動作	14例 *23.7%*	0例 0.0%	2例 3.4%	0例 0.0%	16例 27.1%
	合計	53例 89.8%	3例 5.1%	2例 3.4%	1例 1.7%	59例 100.0%

		後件				
前件	**なら**	利用者の動作	利用者の非動作	システムの動作	システムの非動作	合計
	利用者の動作	0例 0.0%	0例 0.0%	0例 0.0%	0例 0.0%	0例 0.0%
	利用者の非動作	8例 *88.9%*	0例 0.0%	0例 0.0%	0例 0.0%	8例 88.9%
	システムの動作	0例 0.0%	0例 0.0%	0例 0.0%	0例 0.0%	0例 0.0%
	システムの非動作	1例 *11.1%*	0例 0.0%	0例 0.0%	0例 0.0%	1例 11.1%
	合計	9例 100.0%	0例 0.0%	0例 0.0%	0例 0.0%	9例 100.0%

「たら」「なら」については，これと相補的な分布をしているので，以下のようになる．

暗黙規則 2 (「たら」「なら」の後件の'主' 制約)

接続助詞「たら」「なら」による複文構造において，後件は利用者の動作しか表さない．つまり，後件の'主'は利用者である．

前出の分布表から上記の暗黙規則の予測の正しさを調べてみると，「れば」に関する暗黙規則 1 は 95.1%，暗黙規則 2 は「たら」に対して 89.8%，「なら」に対しては，文例が少ないものの，100% 満たす．よって，これらの暗黙規則は十分妥当性を持っていると考えられる．

4 おわりに

マニュアル文に現われる条件表現「と」，「れば」，「たら」，「なら」について言語学的，実証的考察を行ない，その性質について述べた．また，その性質から，各条件表現の後件の'主'について，制約ならびに暗黙規則を提案し，十分妥当性を持つことを検証した．これらの制約や暗黙規則を利用することにより，マニュアル文から知識獲得に必要不可欠なゼロ代名詞の照応候補の絞り込みなどを効率よく行なえると期待される．

参考文献

[安部 88] 安部憲広．機構部品組み立てマニュアル理解システム．特定研究「言語情報処理の高度化」報告書，1988.

[久野 73] 久野すすむ．日本文法研究．大修館書店，1973.

[IPA87] 情報処理振興事業協会技術センター．計算機用日本語基本動詞辞書 IPAL(Basic Verbs) — 解説編 —．情報処理振興事業協会，3 月 1987.

[仁田 93] 仁田義雄．日本語の格を求めて．仁田義雄(編)，日本語の格をめぐって．くろしお出版，東京，1993.

[益岡 93] 益岡隆志．日本語の条件表現について．日本語の条件表現．くろしお出版，東京，1993.

マニュアル出典

[Sun92] Sun Microsystems, Inc. Desktop システム・ユーザ・ガイド，1992.

[富士 88] ハイテクノジーコミュニケーションズ（株）．OASYS Lite F・ROM 11/11 D 操作マニュアル，1988.

など計 16 冊

多次元尺度法を用いた語順パラメータの間の関係付け

Relations among Word Order Parameters Analyzed by Multi-Dimensional Scaling

江原　暉将

Terumasa EHARA

*NHK*放送技術研究所

NHK Science and Technical Research Laboratories

eharate@strl.nhk.or.jp

1．はじめに

世界には多くの言語があり、それらの間には類似点と相違点がある。言語類型論は、この点を明らかにする研究分野である。言語類型論の分野の中でも、語順に関する類型についての研究は初期のころから行われており、様々な事実が分かっている。しかし、従来の研究結果は、定性的なものであり、また、語順に関する様々な性質を、主として、個別に検討するものであった。本研究は、従来から調べられている語順に関する諸性質を、定量的かつ総合的に捉え、諸性質の間の関係を明らかにするとともに、語順類型から見て、世界の言語を特徴付けることを目的としている。そのために、多次元尺度構成法を用いる。

2．語順に関する諸性質とその数量化

本文で用いる語順に関する諸性質としては、［角田］で用いられている20の性質を用いた。表1にこれらの性質を示す。［角田］は、130の言語に対して、これらの諸性質がどのように実現されているかを定性的に示した。ここでは、定量的な解析を行うために、これらの諸性質を語順パラメータとして、数量化する。パラメータの値は-10ないし+10の範囲であり、各性質に対して、＋の方向は、その性質が日本語で、どのように実現しているかに応じて定めた。例えば、第1番目の語順パラメータP1である、平叙文における主語（S）と動詞（V）の相対順序については、日本語と同一のSVの語順が＋であり、VSの語順が－である。次に、注目している言語Lに対して、パラメータの値は、以下のように定める。

（1）Lの典型的な表現に対して、当該の性質が片方のみに実現しているときには、+10または-10の値を与える。例えば、日本語では、典型的にSVであるから、P1の値が+10である。また、ウエールズ語では、典型的にVSであるから、P1の値が-10である。

（2）Lに対して、当該の性質が片方のみに実現していない場合は、その程度に応じて、+10と-10の間の数値を与える。例えば、英語では、固有名詞と普通名詞の順序に対して、N-PrとPr-Nの両方の場合があるが、N-Prの方が多いため、英語に対するP9の値は-5とした。

（3）Lに対して、当該の性質がほぼ同等に生ずる場合には、値0を与えた。また、Lに対して、当該の性質が不明の場合や、その性質を考えることができない場合も0を与えた。例えば、中国語では、平叙文における目的語（O）と動詞（V）の語順として、OV

表1　語順に関する性質とそのパラメータ化

No	語順に関する性質	＋	その他	－
1	平叙文での主語（S）と動詞（V）	SV		VS
2	平叙文での目的語（O）と動詞	OV		VO
3	名詞（N）と側置詞（Ap）	N-Ap		Ap-N
4	所有格（G）と名詞	GN		NG
5	指示詞（Dm）と名詞	Dm-N		N-Dm
6	数詞（Nu）と名詞	Nu-N		N-Nu
7	形容詞（A）と名詞	AN		NA
8	関係節（Rel）と主名詞	Rel-N		N-Rel
9	固有名詞（Pr）と普通名詞	Pr-N		N-Pr
10	比較の基準（C）と形容詞	CA		AC
11	本動詞と助動詞（X）	VX		XV
12	副詞（D）と動詞	DV		VD
13	副詞と形容詞	DA		AD
14	疑問の印	SF		SI
15	一般疑問文でのSV倒置	no		yes
16	疑問詞の位置	SF	PiD	SI
17	特種疑問文でのSV倒置	no		yes
18	否定の印（Ng）と否定の対象（On）	On-Ng		Ng-On
19	条件節（Cc）と主節（Cl）	Cc-Cl		Cl-Cc
20	目的節（Cp）と主節	Cp-Cl		Cl-Cp

SI：文頭　SF：文末　PiD：平叙文の位置

と VO が、ほぼ同等に生ずるので、P2 の値が 0 となる。

（4）P16 については、疑問代名詞が平叙文と同一の位置に現われる場合、+1 を与える。

以上の方法で 126 の言語に対して 20 の語順パラメータの値を設定した。この設定に当たっては、多く［角田］によったが、一部［亀井］も参考にした。また、筆者独自の調査による部分も一部ある。付録 1 にその結果を示す。付録 1 では、値 10 を a で示してある。

3．多次元尺度法による解析

語順パラメータを用いて、言語 Li と Lj（i, j = 1,...,N ; N = 126）の間の距離が、R 次元（R = 20）ユークリッド空間での距離として計算できる。このようにして得られたN＊Nの距離行列Dに基づいて、Torgersonの多次元尺度構成法によって、内積行列Bが求められる。このとき、Torgersonの単純法で加算定数を計算し、用いた。この値は -0.89 であった。Bの固有値を図 1 に示す。第 2 主成分までで累積寄与率 53% を達成している。図 2 はN個の言語を第 2 主成分までの固有空間に配置した結果である。この空間を「語順空間」と名付ける。各言語に対する、第 1 および第 2 主成分の値を付録 1 の列 A と B に示す。

4．語順パラメータの間の関係

語順空間での各言語の配置から、以下の様にして語順パラメータの間の関係を求めることができる。パラメータ Pk に対して、L= {Li ¦ i = 1,...,N} を次の 3 個の部分に直和分割する。Lk+ = {Li ¦ Pk(Li)>0}、Lk0 = {Li ¦ Pk(Li) = 0}、Lk- = {Li ¦ Pk(Li) < 0}。語順空間でのLk+とLk-の平均位置をqk+、qk-とするとき、パラメータベクトルをqk = qk+ − qk-と定義する。ベクトルqkと横軸との成す角（ak）を求めて、Lk+とLk- の要素数と共に表 2 に示す。表 2 から、語順パラメータの集合P = {Pk ¦ k=1,...,R } を以下の 3 個の部分に分割する。$P1$ = {Pk ¦ 0≦ak≦30}、$P2$ = {Pk ¦ 60≦ak≦90}、$P3$ = {Pk ¦ 30< ak<60}。$P1$ の要素は第 1 主成分と相関が強く、13 個の語順パラメータを含む。また、$P2$ の要素は第 2 主成分と相関が強く、4 個の語順パラメータを含む。また、各部分に属するパラメータ同士はお互いに相関が強い。$P1$ に属する代表的なパラメータとして、P2（目的語と動詞の順序）を選ぶことができ、$P2$ に属する代表的なパラメータとしては、P7（形容詞と名詞の順序）を選ぶことができる。

これによって、言語は、目的語と動詞の順序によって第 1 に特徴付けられ、形容詞と名詞の順序によって第 2 に特徴付けられることが分かる。前者は、従来から言われていたことであるが、後者は、必ずしも明確ではなかった。［Comrie］によれば、目的語（O）と動詞（V）の順序と形容詞（A）と名詞（N）の順序には強い相関があるとされ、

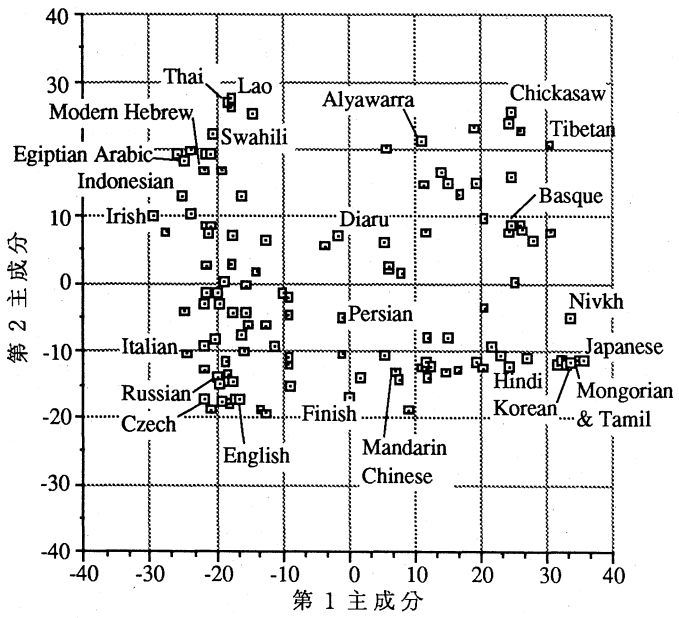


図 2　語順空間内での126言語の分布

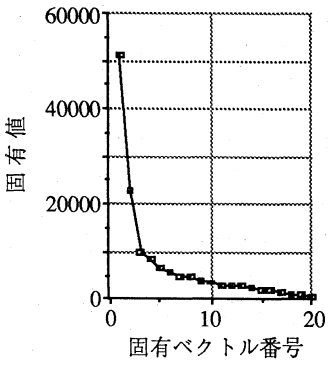


図 1　内積行列の固有値

OV N-Ap GN ANの組とVO Ap-N NG NAの組の２つが語順に関する主要な型であるとされているが、ここでの結果は、必ずしもそうとは言えないことを示している。

５．語順と思考パターンの関係

本節では、語順と思考パターンの関係についての仮説を述べる。前節で述べたように言語はOV型であるか、VO型であるかで、第１に特徴付けられる。そして、OV型である言語は、N-Apであり、GNである傾向が強い。逆に、VO型言語はAp-N、NGである傾向が強い。日本語は典型的なOV型であり、欧州の言語の多くはVO型である。

一方、文を主要素と従要素との関係を中心に解析する手法に依存解析（依存文法）がある。VO型言語の文を依存解析すると、主要素が従要素の前方にくることが多く、OV型言語の文を依存解析すると、主要素が従要素の後方にくることが多い。依存解析と同様な手法である係受け解析が日本語に対して存在する。依存解析も係受け解析も、解析の結果、従要素と主要素を有向弧で結んだ解析木を得るが、図３に示すように、依存解析では、弧の方向が主要素から従要素の方向を向き、係受け解析では従要素から主要素の方向を向く。この違いは単なる記法の差であろうか。筆者は、これを語順類型の影響ではないかと考える。つまり、弧の方向を示す矢印は本物の矢と同じに、前方から後方に向かって付けられるのが自然であるので、VO型言語の多い欧州で生まれた依存解析では、弧の方向が、主要素から従要素に向かい、OV型言語である日本語に対する係受け解析では従要素から主要素に向かったのではないかと考える。そして、この差は、実は思考パターンにも影響しているのではなかろうか。つまり、VO型言語の話者は、トップダウン（演繹的）に思考する傾向が強く、OV型言語の話者はボトムアップ（帰納的）に思考する傾向が強いのではないかと考える。あるいは、既に、このような関係のないことが証明されているのかもしれないが、もし、そうでないとすれば、こうした関係があるか否かを実証することは興味あることと思われる。

表２　語順パラメータの分類

Group No.	Parm. No	ak (deg.)	*L* k+	*L* k-
1	9	0.0	29	75
	3	0.9	51	60
	10	1.7	33	53
	2	3.6	55	68
	12	4.6	38	18
	11	4.7	39	50
	18	5.4	30	75
	8	6.3	21	83
	4	8.1	73	45
	20	10.2	35	15
	16	11.6	30	95
	1	17.9	99	26
	14	19.7	50	47
3	15	37.6	112	14
	17	40.4	109	17
	19	54.4	121	3
2	6	61.4	96	29
	5	64.0	83	38
	13	72.6	85	28
	7	83.8	65	60

６．おわりに

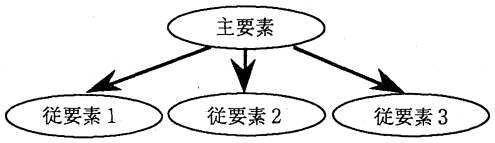
これまでに得られている語順類型に関する研究成果を、多次元尺度構成法を用いて定量的に調べた。その結果、第２主成分までで、累積寄与率53％が得られた。第１主成分は「目的語と動詞の順序」、第２主成分は「形容詞と名詞の順序」として意味付けられることが分かった。最後に、語順類型と思考パターンの間の関係に関する仮説を述べた。

参考文献

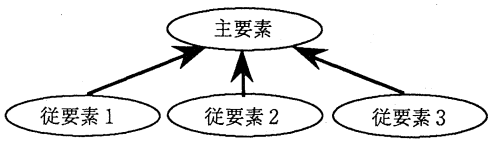
[Comrie] Comrie, B.：Language Universals and Linguistic Typology, University of Chicago Press, 1981.

[亀井] 亀井孝ほか（編著）：言語学大辞典１～４、三省堂、1988～92年。

[角田] 角田太作：世界の言語と日本語、くろしお出版、1991年。



(a) 依存木



(b) 係受け木

図３　依存木と係受け木

付録1　126言語に対する語順パラメータ値（1～20）と語順空間での位置（1軸:A、2軸:B）

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	A	B
1	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	1	a	7	a	a	35.8	-11.2
2	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	1	a	-7	a	a	33.5	-11.5
3	a	a	a	a	a	a	a	a	5	a	a	a	a	a	a	1	a	7	a	a	34.7	-11.2
4	a	a	a	0	a	a	a	a	-a	a	a	a	a	7	a	1	a	-7	a	-a	21.7	-9.3
5	a	a	a	7	a	a	a	a	a	a	a	a	a	7	a	-7	a	7	a	5	32.8	-11.6
6	a	a	a	a	a	a	a	a	a	a	a	0	a	a	a	1	a	0	5	a	32.3	-11.1
7	a	0	a	5	a	a	a	-a	a	-5	5	0	a	7	-a	-a	-a	-7	a	a	9.2	-18.5
8	a	-a	5	a	a	a	a	-5	5	-5	-a	0	a	-5	a	-a	a	-7	-5	a	0.2	-16.9
9	a	a	a	0	a	5	-5	a	-a	a	a	0	a	7	a	1	a	7	5	5	25.0	0.4
10	a	a	a	a	a	-a	-a	5	-a	a	a	a	a	7	a	-7	a	7	a	a	30.5	7.6
11	a	a	a	a	a	-a	-a	5	-a	a	a	a	a	7	a	-7	a	7	a	a	30.5	7.6
12	a	7	a	a	a	a	a	-a	a	0	a	a	0	a	-a	a	-7	a	-a		12.6	-12.2
13	a	-a	-a	-5	a	a	a	-a	-a	-a	-a	0	a	-8	a	-7	a	-7	a	-a	-19.9	-14.0
14	a	-a	-a	-a	a	a	a	a	-a	-a	-5	0	a	-a	a	-a	a	-7	5	0	-18.3	-14.5
15	a	-a	-a	-a	a	a	a	a	-a	-5	-a	-a	0	a	0	-5	-a	-5	-7	a	-a	-21.6	-17.1
16	a	-a	-a	-5	a	a	a	-5	-a	-a	-5	-a	a	-9	a	-a	a	-7	5	0	-17.6	-14.6
17	a	-a	-a	-5	a	a	a	a	-a	-a	-a	-5	0	a	-a	-5	-a	-5	-7	-5	-5	-20.7	-18.7
18	a	-a	-a	5	a	a	a	a	-a	-5	-a	-a	0	a	0	-5	-a	-5	7	5	5	-12.4	-19.2
19	a	-a	-a	5	a	a	a	a	-a	-5	-a	-a	a	0	-5	-a	-5	7	-5	-a	-18.1	-17.8
20	a	-a	-a	5	a	a	a	a	-a	-5	-a	-a	0	a	0	-5	-a	-5	7	5	0	-13.4	-18.8
21	a	-a	-5	-5	a	a	a	a	-a	-a	-a	-a	0	a	0	-5	-a	-5	-7	a	-a	-19.4	-17.4
22	a	-a	-5	-5	a	a	a	a	-a	-a	-a	-a	0	a	0	-5	-a	-5	7	5	-a	-17.3	-17.1
23	a	-a	-a	5	a	a	a	a	-a	-5	-a	-a	0	5	0	-a	-a	-a	7	5	-a	-16.5	-17.3
24	-a	-a	-a	-5	-a	a	-a	-a	-a	-a	-a	-a	-a	5	-a	a	-a	a	-a	5	-a	-29.6	10.2
25	-a	-a	-a	-5	-a	a	-5	-a	-a	-a	-a	-a	5	-a	a	-a	a	0	5	-a	-27.8	7.7
26	-a	-a	-a	-5	5	a	-5	-a	-a	0	a	-a	5	-a	a	-a	a	0	5	0	-14.0	1.8
27	a	-a	-a	-5	a	a	-5	-a	-a	-a	-a	-a	a	-a	-a	-a	-a	0	5	5	-21.9	-12.6
28	a	-a	-a	-5	a	a	-5	-a	-a	-a	-a	0	a	0	a	-a	-a	-7	5	-5	-20.1	-8.3
29	a	-a	-a	-5	a	a	-5	-a	-a	-a	-a	0	a	0	-5	-a	-5	-7	a	-a	-21.7	-9.2
30	a	-a	-a	-5	a	a	-5	-a	-a	-a	-a	-a	a	0	-a	-a	-a	-7	a	-a	-24.6	-10.3
31	a	-a	-a	-a	a	a	a	a	-a	-a	-a	-a	0	a	-a	a	-a	a	-7	a	0	-19.7	-14.9
32	a	-a	-a	-a	5	a	5	-a	-a	-a	-a	-a	a	0	a	-a	-5	-7	5	-a	-24.9	-4.2
33	5	-a	-a	-5	5	a	5	-5	-a	-5	-a	0	a	-a	a	-a	a	-7	5	5	-16.0	-10.1
34	a	a	-a	0	a	a	-5	-a	-5	0	-5	a	a	-a	a	-5	a	-7	a	-5	-0.9	-5.1
35	a	a	a	a	a	a	a	0	0	a	a	a	a	a	0	a	1	a	-7	a	-a	23.1	-10.8
36	a	a	a	a	a	a	a	0	a	a	a	a	a	-a	a	1	a	-7	a	-a	24.1	-12.1
37	a	a	a	a	a	a	a	0	a	0	a	a	a	5	a	0	a	7	a	0	26.9	-11.1
38	a	a	a	a	-a	5	-a	a	5	a	a	a	a	5	a	-5	-5	-7	a	-5	24.7	9.0
39	a	-a	-a	-a	-a	5	-a	-a	-a	-a	-a	-a	-a	-5	a	-5	a	0	a	-a	-24.9	18.3
40	-5	-a	-a	-a	-a	a	-a	-a	5	-a	0	0	-a	-a	a	-a	a	-a	a	-a	-21.8	16.9
41	a	a	-a	a	a	a	a	a	-a	0	5	a	a	0	a	1	a	0	a	-a	11.5	-11.6
42	a	-a	-a	-a	5	-a	5	-a	-a	0	-a	-a	0	a	a	-5	a	-a	5	-a	-17.3	7.0
43	a	-a	-a	-a	-a	-5	-a	-a	-a	-a	-a	-a	-5	-a	a	-a	a	-7	a	-a	-25.9	19.3
44	a	-a	-a	-a	-5	-a	-a	-a	-a	0	-a	-a	-a	-a	a	1	a	-7	a	-a	-20.6	22.3
45	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	1	a	7	a	a	35.8	-11.2
46	a	a	a	a	a	a	a	a	0	a	a	a	a	5	a	1	a	7	a	a	32.9	-11.7
47	a	a	a	a	a	a	a	a	0	a	a	a	a	5	a	1	a	0	a	a	31.7	-11.9
48	a	a	a	a	-a	-a	-a	a	0	a	a	a	5	a	a	1	a	7	a	0	30.6	20.7
49	a	a	a	a	0	-a	-a	0	-a	a	a	a	-a	a	a	1	a	7	a	0	24.1	22.8
50	a	a	a	a	a	-a	-a	a	5	0	a	a	a	a	a	1	a	-7	a	a	30.7	7.7
51	a	0	-a	a	a	a	a	a	a	0	-a	0	a	a	a	1	a	-7	a	0	6.8	-13.1
52	a	-a	-a	-a	-a	-a	-a	-a	-a	-a	-5	0	-a	a	a	1	a	-7	5	-a	-18.5	26.9
53	a	-a	-a	-a	-a	-a	-a	-a	-a	-a	-5	0	-a	a	a	a	a	-7	5	-a	-17.6	27.6
54	a	-a	-a	-a	-a	-a	-a	-a	-a	-a	-5	0	-a	a	a	-a	a	-7	a	-5	-17.8	26.1
55	a	-a	-a	-a	-a	5	-a	-a	-a	-a	-5	0	5	a	a	1	a	-7	5	5	-16.2	13.0
56	a	-a	-a	-a	-a	5	-a	-a	-a	-a	-a	0	5	5	-5	-a	a	-7	5	-a	-23.8	10.3
57	a	-a	-a	-a	-a	a	-a	-a	-a	-a	-a	0	-5	-a	a	-5	a	-7	5	-a	-25.1	12.9
58	-a	-a	-a	-5	5	a	5	-5	-a	-a	0	0	a	5	a	-a	a	-a	5	5	-15.4	-6.2
59	-a	-a	-a	-a	a	a	a	a	-a	-a	0	-a	0	-a	5	a	-a	a	5	-a	-21.4	-1.6
60	-a	-a	-a	-5	5	a	a	0	-a	-a	-a	0	-5	0	a	-a	a	-a	5	-a	-21.7	-2.9
61	-a	-a	-a	5	5	5	5	-a	-a	-a	0	-a	a	5	a	-a	a	-a	5	5	-15.7	-4.4
62	-5	-a	-a	-a	a	a	5	-a	a	-a	-a	0	a	0	a	1	a	0	a	-a	-16.3	-7.8
63	5	-a	-a	-a	a	a	5	-a	-a	-a	-5	0	a	-a	a	-a	a	-a	5	5	-18.7	-11.7
64	-a	-a	-a	-5	-a	-5	-5	-a	-a	-a	-a	0	-5	0	a	1	a	-a	5	-a	-23.9	19.8
65	-a	-a	-a	-5	5	-5	-a	-a	-5	-a	0	0	-a	a	-a	-5	a	-7	5	-a	-19.4	16.9
66	-a	-a	-a	-5	-a	a	-a	-a	-a	0	-a	0	-a	5	a	1	a	-7	5	-a	-21.3	19.3
67	-a	-a	-a	-5	-5	-a	-a	-a	0	0	0	0	-a	0	a	-5	a	-a	a	-a	-14.6	25.1
68	a	a	0	5	5	-5	5	-a	a	0	0	0	0	-a	a	-a	a	-a	a	-a	6.1	2.6
69	a	a	a	a	a	a	a	-a	a	0	0	0	-a	-a	-a	a	-a	a	-a	-a	5.4	6.3
70	a	a	a	5	a	-a	-a	0	-5	0	a	0	-a	0	a	-a	a	-a	5	-a	14.2	16.7
71	a	a	0	-5	-5	-a	-5	-a	-a	0	0	0	-a	0	a	-a	a	7	a	a	5.8	20.1
72	a	a	0	0	-5	-a	-5	0	0	a	a	0	0	0	-5	a	-a	a	-a	5	5	11.2	14.7
73	a	-5	0	5	5	5	-5	0	0	0	0	0	0	-a	-7	a	-a	a	-a	5	-a	-1.6	7.2
74	a	a	5	5	5	5	-5	0	-a	0	0	0	a	5	a	-a	a	-7	a	-a	7.9	1.4
75	-a	-a	0	a	5	5	5	-a	-a	0	-a	0	0	-7	a	-a	a	0	a	0	-9.1	-2.1
76	a	a	a	5	a	a	a	0	a	0	0	0	a	5	a	-a	a	7	5	5	20.1	-12.4
77	a	a	a	5	a	a	a	0	-a	0	0	a	a	5	a	-a	a	0	5	5	16.5	-12.7
78	a	-a	a	a	a	a	-a	-a	-a	0	0	-a	-a	-7	a	-a	a	0	5	-a	-3.7	5.7
79	a	a	a	a	a	a	-a	-a	a	a	a	a	-a	a	a	-a	a	7	a	0	26.0	8.7
80	a	a	a	a	-a	0	-a	0	5	a	0	0	0	-a	a	-a	a	a	a	0	19.3	15.0
81	a	a	a	a	-a	-a	-a	0	a	0	0	0	0	a	a	1	a	7	a	-a	19.1	23.0
82	-a	a	a	a	0	a	0	0	0	0	a	-a	-a	5	a	-a	-5	7	5	-5	11.7	7.8
83	-5	5	a	a	0	5	-5	0	0	a	a	0	-a	0	a	-a	a	7	a	-5	16.9	13.3
84	a	a	a	5	a	a	a	5	-a	a	5	a	a	5	a	-a	a	-a	a	-5	19.3	-11.5
85	a	a	a	a	a	5	-a	-a	5	a	a	a	0	5	a	1	a	7	a	-a	24.2	7.7
86	a	a	a	a	0	5	5	a	a	0	a	0	a	-a	a	-a	a	0	a	-a	20.7	-3.5
87	a	-a	0	5	a	a	a	0	0	0	0	0	5	-a	a	-a	-7	a	a		1.8	-14.0
88	a	a	0	5	-a	-a	-a	0	a	0	0	0	0	0	a	-a	a	7	0	-a	10.9	21.2
89	a	a	a	a	a	-a	-a	-a	a	0	a	0	0	-7	a	-a	a	-7	a	a	20.7	9.9
90	-a	-a	-a	0	5	a	5	-a	-a	0	-a	-5	0	5	a	-a	a	-a	5	-a	-20.0	-1.4
91	5	-a	-a	-5	-a	a	a	-a	-a	-a	0	0	-5	-a	a	-a	a	-a	a	-a	-21.6	2.5
92	a	-a	-a	-a	0	a	-a	-a	0	-a	0	0	a	-a	-a	-a	-a	0	a	-a	-19.7	-3.1
93	-a	-a	-a	-5	-a	a	a	-a	-a	-a	0	0	a	5	a	-1	a	0	a	-a	-18.9	0.4
94	a	-a	-5	a	a	a	-5	-a	-a	-a	-a	0	a	0	a	-a	a	0	5	-a	-12.4	-6.1
95	a	-a	-a	a	-a	a	-a	-a	-a	-a	-a	0	0	a	0	-a	a	0	5	-5	-12.7	6.4
96	-a	-a	-a	a	-a	a	-5	-a	0	-a	-a	a	-a	a	-a	a	-a	a	5	-5	-17.6	3.1
97	-a	-a	-a	a	-a	5	-5	-a	0	-a	-a	-a	a	0	a	-a	a	-a	5	-a	-21.1	7.4
98	a	-a	-a	a	-a	a	-a	-a	-a	-a	-a	0	a	5	a	-a	a	-a	a	-a	-17.4	-4.6
99	a	-a	-a	-5	a	a	a	-a	-a	-a	-a	0	a	0	a	-a	a	-7	5	-5	-18.3	-13.7
100	-a	-a	-a	a	-a	a	-a	-a	-a	-a	-a	0	a	7	a	-a	a	-a	a	-a	-20.7	8.7
101	-a	-a	-a	0	-a	a	-a	-a	-a	-a	-a	0	0	-a	a	-a	a	-a	5	-a	-21.0	19.1
102	5	-a	-a	-a	-a	a	-a	-a	-a	-a	0	0	a	0	a	-a	a	-a	5	-a	-21.6	8.7
103	-a	-a	-a	a	a	a	5	-a	-a	0	-5	a	a	0	a	-a	a	-7	a	-a	-11.3	-9.3
104	a	-a	-5	a	a	a	a	-a	-a	0	a	a	-a	a	-a	a	-a	-7	a	-a	-8.9	-15.1
105	a	a	a	0	a	a	a	-a	5	a	0	0	a	-a	a	-a	a	-7	5	-a	10.9	-12.4
106	0	0	a	a	a	a	a	-a	a	-a	a	0	a	a	-a	a	-a	a	0	5	0	7.5	-14.3
107	a	a	a	5	a	a	a	-a	-a	a	0	0	a	-a	a	-a	a	-7	5	5	12.0	-13.8
108	a	a	a	5	5	a	5	-a	0	0	0	0	a	-8	a	-a	7	5	0	11.8	-8.1
109	a	a	a	a	a	a	a	-a	0	a	0	a	a	-7	a	-a	a	-a	a	-a	14.7	-13.1
110	a	a	0	a	a	a	5	-a	-a	0	0	0	a	-7	a	-5	a	-a	a	0	5.4	-10.7
111	a	a	0	0	a	a	-a	0	0	0	0	0	a	0	-a	a	-a	a	-a	a	6.0	2.0
112	a	a	a	a	a	-a	-a	0	0	a	a	a	a	-9	a	1	a	0	a	5	28.0	6.6
113	a	a	a	a	a	-a	-a	-a	5	a	a	a	5	-a	a	-5	a	0	5	a	26.3	7.9
114	a	a	a	a	-a	-a	-a	0	a	0	a	a	-a	a	a	-a	a	0	5	5	24.7	25.6
115	a	a	a	a	-a	-a	-a	0	a	0	a	a	-a	a	a	-a	a	0	5	5	24.7	25.6
116	a	a	a	a	-5	-a	-a	0	-a	0	a	a	-a	a	a	1	a	7	a	a	24.2	24.1
117	a	a	a	a	-5	-a	-a	-a	0	a	a	a	a	a	a	-a	a	7	0	0	24.5	15.9
118	5	-a	0	a	a	a	a	-a	0	0	0	0	0	-5	a	-a	a	-7	5	5	-1.2	-10.5
119	a	-a	0	a	a	a	5	-a	-a	-a	0	0	a	-7	a	-a	a	-a	a	-a	-9.2	-11.8
120	a	-a	0	5	a	a	a	-a	0	0	-a	0	0	-a	a	-a	a	-a	a	-a	-9.0	-10.9
121	-a	-a	0	5	a	5	5	-a	-a	0	5	0	5	-a	a	-a	a	-a	5	-a	-9.1	-4.8
122	-a	-a	0	a	-a	a	a	-a	-a	0	0	0	a	5	a	-a	a	-a	5	-a	-10.1	-1.4
123	-a	-a	0	-a	-a	a	a	-5	-5	0	-a	0	a	a	a	-a	a	-a	5	-a	-15.7	-0.2
124	a	a	a	-a	-5	-5	-a	-5	-5	a	0	a	a	5	a	1	a	7	5	-5	15.1	15.1
125	a	5	a	5	a	a	5	5	-a	a	5	0	a	0	a	1	a	-5	5	-a	14.9	-8.1
126	a	a	a	a	a	-5	a	a	0	a	a	a	a	5	a	1	a	7	a	5	33.4	-4.9

1:Japanese 2:Korean 3:Mongolian 4:Evenki 5:Turkish 6:Mari 7:Hungarian 8:Finnish 9:Abkhaz 10:Adyghe 11:Kabardian 12:Georgian 13:Russian 14:Polish 15:Czech 16:Bulgarian 17:Serbo-Croatian 18:Swedish 19:Norwegian 20:Danish 21:German 22:Dutch 23:English 24:Irish 25:Welsh 26:Breton 27:French 28:Portuguese 29:Spanish 30:Italian 31:Esperanto 32:Rumanian 33:Modern Greek 34:Persian 35:Panjabi 36:Hindi 37:Bengali 38:Basque 39:Egyptian Arabic 40:Modern Hebrew 41:Tigrinya 42:Hausa 43:Yoruba 44:Swahili 45:Tamil 46:Kannada 47:Burushaski 48:Tibetan 49:Mizo 50:Burmese 51:Mandarin Chinese 52:Thai 53:Lao 54:Cambodian 55:Vietnamese 56:Malay 57:Indonesian 58:Tagalog 59:Ilokano 60:Kapampangan 61:Bikol 62:Palauan 63:Chamorro 64:Tongan 65:Samoan 66:Niuean 67:Maori 68:Warrungu 69:Kalkatungu 70:Diyari 71:Alyawarra 72:Warlpiri 73:Djaru 74:Kuniyanti 75:Amuesha 76:Jaqaru 77:Aymara 78:Guarani 79:Urubu-Kaapor 80:Canela 81:Piraha 82:Hixkarynana 83:Apalai 84:Quechua 85:Tuyucan 86:Tol 87:Highland Chontal 88:Walapai 89:Eastern Pomo 90:Mam 91:Ixil 92:Quiche 93:Pokomchi 94:Rabinal Achi 95:Cakchiquel 96:K'ekchi' 97:Jacaltec 98:Tojolabal 99:Chontal Mayan 100:Chorti 101:Copala Trique 102:Isthumus Zapotec 103:Pipil 104:Nahuatl 105:Yaqui 106:Papago 107:Hopi 108:Chemehuevi 109:Comanche 110:Luisenyo 111:Kiowa 112:Navajo 113:Slavey 114:Choctaw 115:Chickasaw 116:Omaha-Ponca 117:Dakota 118:Blackfoot 119:Atikamekw 120:Sahaptin 121:Nez Perce 122:Coast Tsimshian 123:Gitksan 124:Eskimo 125:Chukchi 126:Nivkh

日本語OCR後処理のための確率主導型誤り検出法と文法主導型誤り検出法の比較

渥美 清隆　　増山 繁
atsumi@smlab.tutkie.tut.ac.jp　masuyama@tutkie.tut.ac.jp
豊橋技術科学大学 知識情報工学系

1 はじめに

近年, パーソナルコンピュータ上で動作する, 性能のよい日本語OCRシステム(以下, 単にOCRと省略する)が利用可能になりつつあるが, それに伴ない, 文書入力作業は軽減されたものの, OCR出力後の誤り文字の検出および訂正という作業にかなりの時間を割かなければならなくなった. ところが, このような誤り文字を検出する作業は長時間にわたる集中力を要求し, 人間に強度の負荷を与える作業であり, コンピュータによる誤り検出および訂正の支援が求められている.

このコンピュータによる誤り検出については, すでにいくつかの論文が報告されており [1, 2, 3, 4, 7], 活発な議論がされているが, それぞれの論文で提案されている検出法を客観的に比較評価した論文は少ない. 本論文では, 現在提案されている代表的な2つの誤り検出法として, 確率主導型の誤り検出法と文法主導型の誤り検出法について, できるだけ客観的な比較を試み, その性質について比較検討する.

2 非文の定義とOCRの誤りの性質

本研究における非文とは, OCRにかける前の原文と比較して, OCRから出力された文が1文字でも置換, 欠落, 誤挿入されているものを言う. 誤りの種類とその変換過程を表1にまとめた. 表1で示した変換の逆変換を誤り訂正演算であると定義するとき, いかなる非文も誤り訂正演算の組み合わせにより原文に復元することができる [5]

ところで, 今回使用したOCRでは, 文字の欠落誤りが単独で起ることはなく, 必ず置換誤りと複合して起きた. これは, ビットマップのイメージ上に何らかの情報が存在する場合には, それを必ず何らかの文字に変換するためである. ところが, 文字の大きさには依存せずに変換を行うため, 2文字以上の列を1文字と認識して変換してしまう場合がある. これは広義の意味での置換ではあるが, 本研究では1文字単位の誤り解析を行うため, このような誤りは欠落と置換が同時に起った複合誤りとする.

表 1: 誤りの種類

正しい文	誤った文	種類
$\cdots w_i\ w_{i+1}\cdots$	$\cdots w_i\ x\ w_{i+1}\cdots$	挿入
$\cdots w_{i-1}\ w_i\ w_{i+1}\cdots$	$\cdots w_{i-1}\ w_{i+1}\cdots$	欠落
$\cdots w_{i-1}\ w_i\ w_{i+1}\cdots$	$\cdots w_{i-1}\ x\ w_{i+1}\cdots$	置換

入力文を $w_0\ w_1\ w_2\cdots w_{n-1}\ w_n$,
$w_0,\ w_1,\cdots,w_n,\ x \in \Sigma$,
Σ は終端記号集合とする.

3 各誤り検出法の概説

我々の知る限りにおいて, 現在提案されている誤り検出法としては, 大きく分けて2つある. 1つはコーパスから得た統計情報を利用し, 誤りを検出する手法(以下, 確率主導型とする)[1, 2] と, もう1つは対象の言語を規定する文法を用いて, トップダウン的な解釈に基づいて誤り検出をする手法(以下, 文法主導型とする) [3, 4, 7] である.

しかし, これら2つのモデルとも, OCRの内部の情報を利用した誤り検出, つまり, 認識文字の第1候補のみならず, 第n候補までの認識文字を利用した誤り検出の研究報告が多い. 我々の立場としては, OCR以外のより広い誤り検出にも利用できるよう

にするため, 認識文字を第1候補文字のみ利用する, つまり, OCRがテキストを完全に出力した後に誤り検出をする方法を採っている.

ここでは, この第1候補文字のみを利用した, それぞれのモデルについての誤り検出法の概略について, 以下に述べる.

3.1 確率主導型誤り検出法

確率主導型の誤り検出法は更にいくつかの種類に分類できるが [1, 2], ここでは2重マルコフモデルを用いた誤り検出法 [2] について述べる. 2重マルコフモデルは大きく2つの部分に分けることができる. 1つは統計情報を得る部分であり, もう1つは統計情報を適用する部分である.

統計情報を得る部分では, 学習対象文字列 $x = w_1 w_2 \cdots w_n$ があるとき, 条件付き確率 $P(w_{i+2}|w_i w_{i+1})$ を計算する. 先頭文字と末尾文字に関してはそれぞれ特殊文字を用意して計算する. 例えば先頭文字では $P(w_1|\$\$)$, \$ は先頭を表す特殊文字, として計算する.

次に, このようにして得られた統計情報に基づき, 入力文字列 $y = v_1 v_2 \cdots v_n$ を解析する. 2重マルコフモデルでは, この入力文字列から任意の3文字列 $v_i v_{i+1} v_{i+2}$ を取り出し, 先に得た統計情報から条件付き確率 $P(v_{i+2}|v_i v_{i+1})$ を求める. このとき部分文字列 $v_{j-2} v_{j-1} \cdots v_{j+k} v_{j+k+1}$ について, 足切値 T を定めるとき, 連続して確率 P が T 以下であるならば, 部分文字列 $v_j \cdots v_{j+k-1}$ は誤り文字列であると推定する.

この推定方法は, 文献 [2] で述べられている単独の欠落誤りに対応することができないが, 2節でも述べたように, 今回使用したOCRでは単独の欠落誤りは出現しないので, この推定方法でも十分である.

本研究で作成した2重マルコフモデルで使用するための条件付き確率の統計情報は, 学習対象文字列 x を $\$\$w_1, \cdots, w_{i-1} w_i w_{i+1}, w_i w_{i+1} w_{i+2}, \cdots, w_n \&\&$ (\$ は先頭を示す文字を, & は末尾を示す文字である), のように, 3文字列の要素に分割し, 最初の2文字が同一であるような要素をそれぞれ1つに集めて, 条件付き確率の計算を行うことによって得た. 学習に際して, 学習対象文字列は段落単位で与えた. これは, 入力文字列の先頭文字や, 末尾文字が誤り文字であった場合に誤り文字を検出することが難しいため, やや大きなブロック単位で学習させることにより, 文の先頭文字や末尾文字の誤り検出もできるだけ行えるようにした.

学習用コーパスとしては, 日経新聞社が販売している1990年,1992年の新聞記事をまとめたCD-ROMから, 誤り検出用に用いる文を除いた社説のすべてと, 同じく春秋 (コラム) のすべてをそれぞれ用意し, 学習させた. また, この2つのコーパスを合せたコーパスからも学習させた. 社説は2年分で3.14MByte, 春秋は同じく2年分で0.93MByteである. これらから得た条件付き確率の項目数は社説で394409件, 春秋で210283件, 両方を合わせた場合では521946件を得ることができた.

3.2 文法主導型誤り検出法

文法主導型の誤り検出モデルもいくつかの種類に分類できるが, ここでは, 我々が研究を進めている誤り検出手法 [5, 6, 7] について述べる.

入力文字列 x を受理する言語 L(G) の文法 G が, 文脈自由文法 $G = (N, \Sigma, S, P)$ で表現できるとき, この文法を誤り訂正用文法 G' に拡張することができる [5]. 具体的な拡張方法は紙面の都合上述べないが, 2節で述べたような誤り変換の種類にそれぞれ対応した誤り訂正演算と呼ばれる導出規則を定義する. この導出規則を最小回数 [5], あるいは準最小回数 [6] 使用することで解析木を作成し, 入力列にいかなる誤りが含まれていても, 文法的には正しい解析木を得ることができる [5]. この方法を応用し, 誤り訂正演算が適用された部分文字列を, 誤りが含まれている部分文字列であると推定することによって, 誤りを検出することができる.

ところが, この方法は複数の解析木を出力するために, 誤り推定文字列を一意に定めることができない. そこで, 今回はこの複数の解析木のうち, 文節の区切りとして, もっとも適切な区切りになっている解析木を人間の手によって選択し, その解析木に含まれている誤り訂正演算を適用してる部分文字列を

誤りが含まれている部分文字列であると推定した.

本研究で作成した文法主導型誤り検出プログラムは主として, 文献[6]を採用し, 形態素解析としても動作するように拡張した[7]. 形態素解析として使用する辞書はICOTが提供しているTRIE辞書ユーティリティ[8]に含まれている形態素解析辞書(約15万件)を利用した. 文法は独自に作成し, 誤り訂正用文法に拡張した. 作成した文法の導出規則数は686個, 終端記号の数は366個であり, これを拡張した文法については, 導出規則数は2145個, 終端記号の数は366個で拡張する前と同じである.

4 実験方法

本論文における実験の目的は, 確率主導型誤り検出法と文法主導型誤り検出法の比較, 特に, 2重マルコフモデルの誤り検出法と誤り訂正演算を含む文法を利用した誤り検出法について, 比較をする. しかし, ここで行うのは性能の比較評価ではなく, 性質の比較評価であることに注意していただきたい. これは, マルコフモデルは, ある一定以上の大きさのコーパスを用意することにより, ある程度の性能を得ることができるが, 文法を用いた方法では, 辞書のチューニングや導出規則のチューニングに大きく左右されるため, 単純には性能比較ができないためである.

主に比較する内容としては, 社説と春秋のそれぞれ1遍ずつのOCR出力後の文書を用意し, その文書の誤り検出をそれぞれの検出法で実行する. 2重マルコフモデルの学習内容と足切値に関して, 社説については, 社説のみの学習を利用し, 足切値を変化させて性能を調査した. また, 春秋については足切値を0.001に固定し, 学習内容を春秋のみ, 社説のみ, 春秋と社説のそれぞれを用意し実験した. その結果について, 2種類の文書間でどのように誤り検出の仕方が変化するのかを見る. OCRにかける文書は, 600dpiのレーザープリンタから印刷し, それをイメージスキャナに読み取らせる方法をとった.

今回使用したOCRはBIRDS社製のThe OCR 日・英 Ver.1.0を用い, イメージスキャナはEPSONのGT-6500(300dpi)を利用した. このOCRに文書をかけた結果, 社説の方の記事は909文字を認識し, 内870文字を正しく認識した. 認識率は95.8%であった. 春秋の方の記事は677文字を認識し, 内622文字を正しく認識した. 認識率は93.1%であった.

表2: 社説について, 2重マルコフモデルの足切値による違いと文法主導型との比較

	2重マルコフモデル			文法型
足切値(T)	0.0005	0.001	0.002	
適合率[%]	52.5	52.4	50.7	18.1
再現率[%]	79.5	84.6	87.2	43.6

表3: 春秋について, 2重マルコフモデルの学習内容による違いと文法主導型との比較

	2重マルコフモデル			文法型
学習内容	春秋	社説	春秋と社説	
適合率[%]	23.2	21.2	29.6	19.0
再現率[%]	87.8	85.7	85.7	40.8

評価基準としては, 適合率と再現率を採用する. 適合率, 再現率はそれぞれ以下のような計算式に基づくものする.

$$\text{適合率} = \frac{\text{正しく誤りを推定した文字数}}{\text{誤りを推定したすべての文字数}} \times 100\ [\%]$$

$$\text{再現率} = \frac{\text{正しく誤りを推定した文字数}}{\text{実際に誤りであるすべての文字数}} \times 100\ [\%]$$

5 実験結果と両モデルの比較

実験結果を表2と表3に示す.

表2, 3の通り, 2重マルコフモデルの性能がかなり良いことが分る. 特に社説では, もとの記事の文体の統一性があるために, 実用レベルでの性能があると言える. ところが, 春秋の方では, 2重マルコフモデルについて, 再現率はある程度の水準を保ってはいるが, 適合率に関しては, 社説に比べて, いずれもかなり劣る. これは, 春秋というコラムの中での文体の統一性が社説に比べてかなり乏しいためと思われる. また, 学習内容が増加するに従って, 適合率は

上る傾向にあり, 再現率は下る傾向にある. これは, あまりに大きなコーパスから学習させた場合, 本来の目的である誤り検出が行えず, 誤り文字を見逃す可能性が増加することを示唆する. このため, どの程度の学習を行わせればよいのか, 十分に検討する必要があるだろう.

文法主導型の誤り検出は, 今回の実験では辞書や文法のチューニングが十分でなかったために, あまりよい結果は得られなかった. 主な原因としては, 辞書が形態素解析用であったために, 単漢字などが登録されており, これが再現率を大きく下げる要因となっている. しかし, 社説に対する結果と, 春秋に対する結果について, ほとんど性能差が見られなかったことに注意する必要がある. これは, 辞書が特定の分野に偏っていなければ, ある程度推敲した文書ならば, 文体に関わらず安定した性能を得ることができるのではないかと考えられる. 今回は, 2種類の文書でしか実験を行わなかったので, この性質については追試の必要がある.

最後に共通した性質として, 両検出法とも実際の誤り文字の直後を誤り推定文字とすることがかなりあった. これは, 何らかの誤りを発見してはいるものの, 前後の正しい文字に引っぱられて, 正しく推定することができなかったためである. このような誤り文字を検出するためには, 別の方策が必要となるかもしれない.

6 むすび

本論文では, 確率主導型の誤り検出法と文法主導型の誤り検出法の比較実験を行い, 以下のことを確認した.

- 2重マルコフモデルによる誤り検出は, 高い再現率を得ることができる.
- 2重マルコフモデルの学習内容や, 入力する文章の文体によって, 確率主導型の誤り検出は大きな性能差, 特に適合率の差がある.
- 文法主導型の誤り検出は文体によらず, 安定した性能を得ることができる可能性がある.
- どちらの検出法も誤り文字の直後を誤り推定文字とすることが多い.

今後は, 文法主導型誤り検出法の辞書や文法のチューニングを行い, 性能を高めるとともに, どのような文書にも安定して高い性能を得るために, 確率主導型誤り検出法と, 文法主導型誤り検出法の両方を組み合わせた誤り検出法についても検討予定である.

参考文献

[1] 森, 阿曽, 牧野 : "2重マルコフモデルを用いた日本語文書認識後処理", 情処研報 94-NL-102-12, Jul. (1994).

[2] 荒木, 池原, 塚原, 小松 : "マルコフモデルを用いたOCRからの誤り文字列の訂正効果", 情処研報 94-NL-102-13, Jul. (1994).

[3] 田邉, 木谷 : "文字認識誤り指摘のための形態素解析の適用性検討", 情処研報 94-NL-102-1, Jul. (1994).

[4] 黄瀬, 白石, 高松, 福永 : "構文・意味解析を用いた文字認識後処理法", 信学論, Vol.J77-D-II, pp.2199-2209, Nov. (1994).

[5] Aho and Peterson : "A Minimum Distance Error-Correcting Parser for Context-Free Languages", *SIAM J.Comput.*, pp.305-312, Dec. (1972).

[6] 渥美, 増山 : "構文解析上の自由度をもった非文訂正法の一提案", 信学論, Vol.J76-D-I, pp.686-688, Dec. (1993).

[7] 渥美, 増山 : "複数候補を出力する非文訂正法のOCR出力の誤り訂正への応用とその候補選出について", 情処研報 94-NL-104-4, Nov. (1994).

[8] (財) 新世代コンピュータ開発機構 : "TRIE辞書ユーティリティ" (1991).

RWCにおける品詞情報付きテキストデータベースの作成

井佐原 均（電子技術総合研究所）　元吉 文男（電子技術総合研究所）
徳永 健伸（東京工業大学）　橋本 三奈子（情報処理振興事業協会）
荻野 紫穂（日本アイ・ビー・エム株式会社）　豊浦 潤（新情報処理開発機構）
岡 隆一（新情報処理開発機構）

1 はじめに

RWCP（新情報処理開発機構、Real World Computer Pertnership）では、平成6年度よりRWCデータベースワークショップを設置し、実世界に関するデータの収集と利用を目的として、テキスト・音声・画像からなるデータベースの作成を行なっている。また、これら3分野の融合として、マルチモーダルデータベースの作成を計画している。

実世界（Real World）における自然言語処理の研究を行なうためには、電子化されたテキストデータベースが必要不可欠である。また、技術の健全な発展と評価のためには、このようなテキストデータベースは広く公開され、研究者が同一のデータに対して実験を行なえることが必要である。

欧米では既にこのような研究・評価用のテキストデータベースが作成され、研究者や企業に対して、共有資源として広く公開されているが、日本においては、この種のデータの組織的な作成は未だ十分には行なわれていない。また、入手可能なデータは高価な場合が多い。

このような点を踏まえて、RWCデータベースワークショップのテキストグループでは、研究・評価用に公開することを前提として、言語情報を付加した現代日本語のテキストデータベースを作成することとした。

2 RWCテキストデータベースの基本的立場

RWCテキストデータベースは以下の条件を満たすことを目標に作成している。

1. 大規模であること
2. 現実のテキストを反映したbalanced corpusであること
3. 精密かつ正確な情報を付加したテキストデータベースであること
4. 必要に応じて対訳テキストデータベースについても検討すること

また、作成の基本理念は「研究・評価を目的としての無償公開」「作成時の協調・分散」「言語理論から独立した汎用性」である。

「公開」については学術目的であれば、誰もが使える共通の資源とすることがもっとも大切な条件であると考えた。ただし、著作権の問題がこの種の言語データ共有においては、常に問題となる。今年度は既に他機関が公開を行なっている、あるいは著作権を主張しない、ものについて、RWCとしての加工と公開を行なうこととした。来年度以降はRWCとして積極的に著作権者と交渉し、公開できるデータを増やしていく予定である。

なお、現段階では、完成したテキストデータベースの配布方法については、確定した案はない。小規模な範囲でのテスト利用からのフィードバックにより、テキストデータの収集・加工についての示唆を得ることから順次始めていく予定である。

「協調・分散」とは、テキストデータベースの共有化を目指す他組織と相互に連絡を取り、協調しながら、ただしお互いは独立して、テキストデータベースの作成を行なおうというものである。具体的には、ＩＰＡコーパス［1］を作成中の情報処理振興事業協会とは、言語情報の付与に用いる品詞体系を共通のものとし、また、互いに収集するテキストの重複を避け、全体としてバランスの取れた言語データを集積するように注意している。（社）日本電子工業振興協会（電子協）においては、その年度報告書のテキストデータベース化および共有化を進めている［2］が、これは未加工テキストであり、ＲＷＣではその未加工テキストを得て、品詞情報の付加を行なっている。このように分野、形態（未加工データと加工データ）、加工作業（品詞体系）のそれぞれについて、「協調・分散」を実現している。

「汎用性」の考え方は後述する品詞体系の設定にもっとも強く反映している。ここでは特定の言語理論に依存するのではなく、出来るだけ多くの情報を付与しておくことにより、利用者がどのような理論に基づく品詞体系を用いようとしている場合にも（比較的）容易に変換できるように品詞体系を作成した。

3　ＲＷＣテキストデータベースの概要

テキストデータベースは、未加工テキストを集めたものと、言語情報を付加したものとに大別される。今年度ＲＷＣで作成したものは、言語情報を付加したものである。

言語情報の付加としては、（1）単語分割、（2）各単語への読みや品詞の付加、（3）係受け構造等の構文情報の付加、（4）意味情報の付加、等が考えられるが、今年度は、（1）および（2）を対象とした。来年度以降は、これに加えて（3）の係受け情報を加えたデータベースの作成を予定しており、今年度はその書式の検討を行なった。

3.1　対象とするテキスト

ＲＷＣでは、今後も継続的にテキストデータの収集・加工を行なうことにより、このテキストデータベースを balanced corpus としていく予定である。その一部をなすものとして、今年度は主として「公開」の可能性に注目して対象テキストを選択した。

今年度、対象としたテキストは、以下に示す１１１７９文からなる。

通商白書（編集・発行 通商産業省）
　　平成４年度版（２３７７文）　　平成５年度版（３２１４文）　　平成６年度版（１７７１文）
１９９４年版 我が国産業の現状－図とデータでみる産業動向（通商産業大臣官房調査統計部 編）
　　マクロ編（２８７文）
電子協平成４年度機械翻訳システム調査委員会報告書「機械翻訳システムの実用化に関する調査研究」
　　（３５３０文）

これらのテキストはよくこなれたものであり、自然言語処理研究用のテキストとしては適切なものである。通商白書の一文は比較的長く、本文では平均で６０文字弱である。電子協の報告書ではこれよりも多少短く、４０文字強である。この値は概ね、新聞の社説と同等である。

3.2 品詞体系

ここでは、品詞体系作成の基本方針として、「汎用性」を考えた。単語に品詞を付与する場合には、ある程度主観的な判断が必要となる。品詞体系そのものについても、各自の立場によって、受け入れ難い場合もあろう。ここでは、利用者が自分の研究目的に合わせて（自由に）取捨選択あるいは変更して利用できるように品詞体系を設定した。

例えば、本品詞体系の品詞がいわゆる学校文法での取り扱いあるいは名称と異なるような箇所には、括弧付で学校文法の品詞を付加しておき、利用者による変更を容易にしている。また、この品詞体系では、「形容動詞」を認めているが、もしこれを「名詞」＋「助動詞」としたい場合には容易に変換が出来る。

「単語」分割においては、どのような単位で分割するかが問題となる。ＲＷＣテキストデータベースの作成に際しては、形態素解析ツールを利用し、構文情報を用いない範囲で処理を行なっているため、形態素単位での分割を行なった。したがって付与する「品詞」も形態素解析レベルで行なえるもの（を中心）とした。

この品詞体系は、ＴＨｉＭＣＯ（Tagset of High quality for Integrated Multi-usage Corpus Openly available to public）と名付けられた。ここでは、品詞は必要に応じて第１レベルから第５レベルまでに詳細化して記述される。第１レベルの品詞としては、以下の１２のものが挙げられている。

（１）名詞　（２）動詞　（３）形容詞　（４）形容動詞　（５）副詞　（６）連体詞
（７）接続詞　（８）助詞　（９）助動詞　（１０）感動詞　（１１）記号　（１２）その他

下位のレベルの分類については言語学的に妥当であり、かつ言語処理に有用と思われるような分類を加えてある。また、人手による修正を前提としているため、本データベースには基本的には、「正解」が記述されている。（人手による修正の段階でも）判断が分かれるような場合には、解釈を保留し曖昧性を残した形の品詞を設定している。たとえば、並立助詞か終助詞かが判断できない「か」に対しては「並立助詞／終助詞」という品詞を与えている。品詞体系の詳細については、参考文献［1］に詳しい。

3.3 データベースの書式

タブコードで区切られた次の項目を１レコードとして持つデータベースを作成する。

（１）分割された単語　（２）読み　（３）原形　（４）品詞分類

実際のデータベースの例を図１に示す。

3.4 作業手順

1. テキストの入手

   通商白書については、平成４年度版および５年度版については、手作業で入力を行なった。平成６年度版およびわが国産業の現状については電子化されたファイルを入手した。しかしながら、このファイルは電子化された最終稿ではあるが、出版物とは微妙に異なる点があり、比較修正した。電子協の報告書については、電子化されたものを入手した。

2. 前処理

   （ａ）一文単位に分割　（ｂ）文ＩＤの付与　（ｃ）特殊文字等の変更

3. 形態素解析

   日本ＩＢＭの形態素解析ツールＪＭＡを用いて自動解析した。ＲＷＣが採用した品詞体系がＪＭＡが元来用いていたものとは異なるため、後処理フィルターを作成し、品詞の置きかえを行なっている。

4. 人間による修正

修正作業は日本語学専攻の大学院生および学部学生が行なった。品詞情報付きテキストデータベースを作成するための援助ツールとして、Nemacs 上で動く（emacs-lisp で記述された）編集ツールを作成した。これは（1）レコードの分割と作成、（2）レコードの結合と削除、（3）品詞選択、の各機能を持つものである。これは上記の書式に基づくデータベースを対象にするツールであり公開可能である。

4 おわりに

実世界の自然言語の解析に関する研究と、その評価に用いるためのテキストデータベースの作成を平成6年度より開始した。これまでに作成したものは1万文強であり、テキストデータとしては分量的には極めて少なく全く不十分なものであるが、今後継続的に拡張していく予定である。

参考文献

[1] 橋本 三奈子、荻野 紫穂、徳永 健伸、元吉 文男、井佐原 均：ＩＰＡコーパスの概要、ＩＰＡシンポジウム'９５論文集（１９９５）

[2] 自然言語処理技術の動向に関する調査報告書、（社）日本電子工業振興協会自然言語処理技術委員会（１９９５）

ＷＧ０６：ｇｅｎｊｏｕ：０００００６０文ＩＤ

なお	ナオ	なお	接続詞			
、	、	、	記号			
最近	サイキン	最近	名詞	副詞可能		
の	ノ	の	助詞	格助詞		
動向	ドウコウ	動向	名詞			
を	ヲ	を	助詞	格助詞		
見る	ミル	見る	動詞	一段	見出し形	
と	ト	と	助詞	接続助詞		
、	、	、	記号			
（中略）						
ようやく	ヨウヤク	ようやく	副詞	助詞類接続		
産業	サンギョウ	産業	名詞			
活動	カツドウ	活動	名詞			
に	ニ	に	助詞	格助詞		
動き	ウゴキ	動き	名詞			
が	ガ	が	助詞	格助詞		
出	デ	出る	動詞	一段	連用タイ接続	
始め	ハジメ	始める	動詞	一段	連用タ接続	非自立
て	テ	て		助詞	接続助詞	
いる	イル	いる	動詞	一段	見出し形	非自立
。	。	。	記号			

図1　実際のデータの例