TR-I-0134

文と文の結束性を捕らえるための知識

Local Cohesive Knowledge

工藤育男

Ikuo KUDO

1990 . 2.20

概要

(概要の内容)

対話文を文脈処理をするときには、割り込みや挿入などにより文脈そのものが、乱れてしまう。このような乱れに対しても、文と文との関係を正しく捕らえるためには、何らか知識を計算機に与えておく必要がある。ここでは、結束性ということに着目し、文脈を捕らえることをこころみる。まず、結束性に関する定義をおこない、この定義を使って、結束性を捕らえるための知識(Local Cohesive Knowledge)を生成する。そして、この知識を機械翻訳システムに応用し、文脈に起因する省略や照応、代用表現の処理への応用する。

ATR Interpreting Telephony Research Laboratories
ATR 自動翻訳電話研究所

"Local Cohesive Knowledge"

by *Ikuo KUDO*

(ATR Interpreting Telephony Research Laboratories,

Sanpeidani Inuidani Seika-cho Soraku-gun , Kyoto, 619-02, Japan)

Abstract

In a natural dialogue, there are many disturbances in the context level because of interruptions and inserted sentences. In spite of such phenomena, cohesion is a very important idea for understanding the context correctly. In our approach, cohesive knowledge which judges cohesion between sentences is given to the system and then the knowledge is used to find cohesion in disarranged context. It is also applied to interpret anaphora, ellipsis and pro-forms in the context. In order to do so, we define the knowledge and use its definition to make knowledge from linguistics database almost automatically.

目次

(注意)本稿は、情報処理学会自然言語処理研究会資料76-7、工藤育男「文と文の結束性を捕らえるための知識」に加筆·修正したものである。

1. はじめに

文脈処理を考える上で、結束性(Cohesion)と一貫性(Coherence)[1],[2],[3]という二つのキーワードがある。このキーワードを計算言語的な立場から、次のように捉えなおす。結束性とは、文と文が結び付いているか、否か、という二項関係を規定したもので、その二項がどんな関係で結び付いているかは規定しない。一方、一貫性とは、文がどのような関係で連なっていくかを規定するもので、主に、二項以上のロジカルな関係づけを規定する。すなわち、結束性と一貫性の違いは、(1)二項関係のみを規定するか、否か、(2)その関係を規定するか、否かであると考える。

一貫性という考え方は、生成の研究において、有益な役割を果たしてきた。というのは、文章を要約したり、すじみち立った文章を構成する場合には、話の論理的展開としての一貫性が保たてることが重要になるからである。それに対し、解析の場合は、一貫性よりも、結束性の方が重要になる。というのは、計算機の入力には、論理的な一貫性などない場合が多いからである。つまり、一貫性に基づいた文脈解析を行っても、一貫性のない文章が入力された場合には、解析不能になってしまうからである。それに対し、結束性に基づいた文脈解析を行う場合には、一貫性のない文章が入力された場合にも、解析不能になることはない。結束性の方が、一貫性のよりも、条件が緩いからである。

実際の会話では、応答の順序や方法が、必ずしも規則的でないため、文脈の乱れが生ずる(2章)。これらの現象は、省略·照応や代用表現などを処理する上で大きな影響を与える。この問題を的確に処理するためには、入力文が、文脈上のどの文に対する発話であるのか(結束性)を捕らえる必要がでてくる。そのためには、予め、計算機に何らかの形で、結束性に関する知識を与えておく必要がある。そのために、まず、結束性に関する知識(制約)を、構文情報と対応づけながら、形式的な定義(3.1)を与える。この定義に基づいて、言語データベースを利用して、これらの知識を半自動的に生成することをこころみる(3.4)。そして、この知識(制約)を、省略の補完や照応や代用表現の解釈に応用する(3.2)。

2.対話における文脈的現象

翻訳システムをつくる際には、省略·照応·代用表現などの問題を処理する必要がある。省略·照応·代用表現などの現象は、文脈に依存して決まる現象と、文脈以外の知識が深く関与している現象とがある。まず、その二つの現象を区別しておく。

2.1 文脈的現象

(1) 照応: 照応関係を機械翻訳システムのおいても捕らえておいた方がいい理由は、訳語選択のところで効いてくるからである。ATRで目標としている会話文のコーパスにおいては、一つの動詞に対して、平均4、5個の訳語の候補がある。

照応関係と言っても、例1に示すような文脈に依存している現象だけではない。例2に示すように、主人と課長が同一人物であることを指していると分かるのは、単に、言語的な情報だけからでは、判断できず、この発話状況を理解するための、言語外知識が必要になる。ここで

(例1)文脈依存の照応

「もし、講演をされたいのなら、論文を送って下さい」

「それは、いつまでに、送ればよいのですか。」

(例 2)状況依存の照応

「田中の妻ですけれども、主人はいますか。」

「課長は、食事にいきました。」

(例 3) 文脈依存の省略と文末の冗長性による省略

「クレジット·カードの名前を教えてください。」

「すみません。[∅ヲ]持っていないのですが。」

(例 4) 代用表現

「登録料は円で支払ってもよいですか。」

「ドルでお願いします。」

(例5)Referent Transfer

「京都ロイヤルホテルの申込みは、まだ、募集していますか。」

「京都ロイヤルホテルは、締め切りました。」

図1 文脈的現象の例

は、このような言語外知識に依存した問題までは言及せず、文脈に依存した範囲内で解決できる問題のみを扱う。

(2) 省略: 省略の起こり方にも、大きく二種類ある。一つは、文脈的現象による省略であり、トピックスの省略がその典型である。例えば、例3における「持っていないのですが」の文における目的格の省略が良い例である。もう一つの省略は、日本語の文末の冗長性による省略(4),(5)である。日本語の文末に、省略情報が隠されているため、明示的に述べる必要がない場合である。例3の「教えて下さい」の主語の省略などがこれに当たる。

(3)代用表現: 代用表現というのは、例4の「お願いします」が、これに当たる。ここでは、「支払って欲しい」という意味で使われている。この代用表現が指している意味は、文脈に依存する。

(4)Referent Transfer: 一文では、意味が通じないが、ある文脈では自然な文というのがある。その代表例が、referent transferである。例5の「京都ロイヤルホテル」という語は、「京都ロイヤルホテルの申込み」という意味で用いられている。この語の解釈は、文脈の中できまる。

ここで扱う問題は、文脈に依存した範囲内で解決できる省略の補完、照応関係および代用表現の同定、Referent Transferなどの文脈的現象である。

2.2 対話における文脈的ロバスト性

自然な対話においては、しばしば、文脈が複雑化する。文脈が複雑化するのは、次の理由によるものである。

(1)発話の順序の乱れ: 例えば、質問と答えの順序関係は、いつも、一定であるとは限らない。二つの質問を一つの発話の中で行い、それぞれの質問に対する答えがなされた場合、答えの文

(例6)二重Question(クロスの例)

「ホテルの名前を教えて下さい。京都駅から近いですか。」

「京都ホテルと京都ロイヤルホテルと京都プリンスホテルです。どれも駅の近くです。」

(例7)二重Question(ネストの例)

「会議の聴講をしたいのですが、どのような手続きをすればよいのですか? 参加料は、いくらですか?」

「登録料は100ドルです。すぐに登録申込み書を、お送り致します。」

(例8)並列句

「それから、御名前と連絡先の電話番号を、御聞きしたいのですが。」

「はい、会社名は、ATRです。電話番号は、06-211-5111です。名前は、木田裕司と申します。」

(例9)繰り返し

「学生でもこの会議に参加できますか?」

「勿論です。誰でも参加できます。」

(例10) 文脈的否定現象

「プロシーディングスは読みましたか?」

「プロシーディングスを送ってください。」

図2 文脈の乱れの例

の順番としては、二通り考えられる。(例6)では、質問と答えの順番が、クロスしているが、(例7)では、質問と答えの順番が、入れ弧になつている。また、サブ·ダイアローグが挿入されたり、ひどい場合は、途中で言い直したり、中断したりして、文のレベルでも乱れてしまう。

(2)文と文の対応関係が、一対一ではなく、一対多、多対一、多対多の関係になる。例えば、並列句(例8)や繰り返し(例9)により、質問と答えの対応関係は、一対多の対応になっている。

(3)文と文に結束性が存在する場合でも、その間に成り立つ関係は、複数の観点から関係を定義することができる。例えば、プランのような行為に対する関係の規定であるとか、命題の真偽値に関する規定であるとか。

[文レベルのロバスト性と文脈的ロバスト性]

ここで取り上げた例は、文脈が乱れる場合の一例にすぎない。しかし、このような乱れは、自然な対話においては、しばしば現れる。このような文脈上の乱れにも耐ええることを、文脈的ロバスト性(contextual robustness)と呼ぶことにする。文脈的ロバスト性を実現するためには、文脈上でどの文と文が結び付いているのかを捕らえる必要がでてくる。

2.3 従来の文脈処理アプローチの課題

省略の補完、照応関係の同定などに対処するための方法として、次の方法がある。

(1)シソーラスを用いたヒューリスティックス(6),(7)

このアプローチは、照応関係の同定(4)に利用されている。シソーラスを用いて、二つのカテゴリーがシソーラス上で照応関係とみなすことができるかどうかを判定する。この方法のいいところは、比較的な簡単な装置で、ある程度の問題に適用でき、実際のシステムに、インプ

リメントが可能であることである。

1)しかし、これらの方法の問題点は、文脈が乱れたときに、リカバリーが保証されいない。その理由は、文レベルでの対応関係についての情報がなく、名詞句レベルでの照応関係をチエックしているだけであるからである。2)また、シソーラスを構築する際に、セマンチック・カテゴリーをどのように設定すれば、どの程度の問題が解決できるのかも、あまり、明確になつていない。3)また、同じカテゴリーに属するものは、処理に困る。4)ただし、シソーラス上の情報だけを使ったこの方法では、照応関係の同定に利用できても、省略の補完には、あまりに多くの解を引き出してしまうため、日本語に多く見られる省略には、新たな方法を考えなくてはならない。

(2)Plan-basedな解析で用いる方法

この方法は、文脈を解析するときに、談話構造という構造を仮定して、文脈を処理(8),(9),(10),(11),(12)しようというところに特徴がある。この方法の利点は、対話のコントロール(13),(14)などに応用がきくことである。また、省略などの補完(10),(15)にも応用しようという試みもある。ただ、この方法で、一番問題になるには、処理対象が狭いことである。協調的な会話、目的指向型の対話といった型にはまった対話構造を処理の対象としており、予め、システムが持っている一貫性を有する決まり切った文脈しか処理できない点が問題である。

また、自然言語処理系とプランなどの推論するための表現系とのマッピングをどうはかるかという問題が生じる。特に、自然言語のもつシンタックスとの関係については、ほとんど、議論すらされていない状態である。

また、知識の生成方法についても、人手に頼って、アドホックな方法がとられているのが現状である。

3 結束性に基づく文脈処理

文脈的な乱れにも耐ええるためには、まず、どの文と文が結び付いているのかを判断する必要がある。つまり、文と文の結束性を捕らえる必要がある。ここでは、どの文と文が結び付くかを、二項関係(結束性に関する知識)として規定しておき、その知識と入力された文脈との整合を計ることにより、文脈を解析するものである。

まず、結束性に関する知識を構文情報との関係を含めて、形式的な定義を与える(3.1)。その理由は、人によるゆらぎをさけ、しかも、できるだけ網羅的に知識を抽出するためには、形式的な定義が必要になるからである。この知識の記述には、抽象化せずに、表層情報を利用する。抽象化せずに、表層情報を利用する理由は、制約条件を強く記述できること、構文情報との関係を明確にできるからである。また、表層情報で記述しても、実際に起こりえる組合せは少なく、記述量は問題にならない(3.4),(3.5)。この知識を、省略の補完や照応の同定、代用表現の解釈に応用する(3.2),(3.3)。

ただ、この方法では、対話の履歴については構成するが、履歴間にある構造解析までは行わない。構造解析については、必要に応じて、対話の履歴を解析すればいいと考える。

3.1 結束性の定義

ここでは、次のような組合せがある時に、結束性があるということにする。

a<X1, Y1, Z1>, b<X2, Y2, Z2>

a, bは動詞、X1, Y1, Z1は、動詞aの格要素で主格、目的格、第二目的格に対応する。今、動詞aの格要素と動詞bの格要素に、表1に示す関係が成立する場合、a<X1, Y1, Z1>とb<X2, Y2, Z2>との間に、結束性があるということにする。

(例10) 「論文を送りましょうか」
「論文を送って下さい」

この二つの文は、「論文」という名詞を仲立ちとして、表1のタイプ(1)により、結束性があるという。この二つの文を結ぶ結束性の知識を次のように書く。

送る<X1, 論文, Z1>, 送る<X2, 論文, Z2>

ここでは、結束性が成立するための条件を、結束性に関する知識(制約)と呼び、また、結束性の成立している動詞aとbの対のことを、結束性のある動詞対と呼ぶことにする。

結束性に関する知識は、結束性が成り立つための制約であるので、この制約を満たすものは、結束性がある。例えば、任意格が入った場合でも、結束性は成立する。

(例11) 「論文を会社に送りましょうか」
「論文を自宅に送って下さい」

結束性に関する知識は、二項関係が結ばれているか、否かのみを規定し、どんな関係で結ばれているかは規定していないので、(例10)も(例11)も同じ結束性に関する知識で、結束性があると判断される。

表1 結束性のタイプ

タイプ	
例	例文に対する結束性に関する知識
(1)同一名詞、同一動詞	
論文を送りましょうか 論文を送って下さい	送る＜X1,論文,Z1＞,送る＜X2,論文,Z2＞
(2)同一名詞、類義語、もしくは、反義語の動詞対	
論文を送りましょうか 論文を送付して下さい	送る＜X1,論文,Z1＞,送付する＜X2,論文,Z2＞
(3)同一名詞、異なり(上記(1),(2)以外)動詞対	
論文を読みましたか 論文を送って下さい	読む＜X1,論文＞,送る＜X2,論文,Z2＞
(4)類義語の名詞、同一動詞	
申込み用紙を送りましょうか すぐに、登録用紙を送ります	送る＜X1,申込み用紙Z1＞,送る＜X2,登録用紙,Z2＞
(5)類義語の名詞、類義語、もしくは、反義語の動詞対	
申込み用紙を送って下さい 申込み案内を送付します	送る＜X1,申込み用紙,Z1＞,送付する＜X2,申込み案内,Z2＞
(6)類義語の名詞、異なり(上記(5),(6)以外)動詞対	
電話はありますか 電話番号は03-333-3333です	ある＜電話＞,です＜電話番号,Y2＞
(7)「AのB」の名詞、同一動詞	
申込みの用紙を送って下さい では、用紙を送ります	送る＜X1,用紙(申込み),Z1＞,送る＜X2,用紙,Z2＞ (....)は、修飾語を意味する
(8)「AのB」の名詞、類義語、もしくは、反義語の動詞対	
講演の論文を募集してますか いいえ、講演は締め切りました	募集する＜X1,論文(講演)＞,締め切る＜X2,講演＞
(9)「AのB」の名詞、異なり(上記(7),(8)以外)動詞対	
申込みの期限を教えてください 申込みは締め切りました	教える＜X1,期限(申込み),Z1＞,締め切る＜X2,申込み＞
(10)複合語の一部の名詞、同一動詞	
登録料金を送って下さい 料金をすぐに送金します	送る＜X1,登録料金,Z1＞,送金する＜X2,料金,Z2＞
(11)複合語の一部の名詞、類義語、もしくは、反義語の動詞対	
国際会議はいつ開催されますか 会議は、8月10日に開かれます	開催する＜X1,国際会議＞,開く＜X2,会議＞
(12)名詞の複合語の一部が一致するし、動詞が異なる(上記(10),(11)以外)場合	
申込み期限を教えてください 申込みは締め切りました 期限は、10日までです 申込みは、10日までです	教える＜X1,申込み期限,Z1＞,締め切る＜X2,申込み＞ 教える＜X1,申込み期限,Z1＞,です＜期限,Y2＞ 教える＜X1,申込み期限,Z1＞,です＜申込み,Y2＞
(13),(14),(15)は、(7),(8),(9)の共有する名詞が類義語になった場合、(16),(17),(18)は、(10),(11),(12)の共有する名詞が類義語になった場合である。	

この「送る」という語を「送付する」という類義語で置き換えても、結束性は成立する。これが、タイプ2である。

(例12)　　[論文を送りましょうか」
　　「論文を送付して下さい」

これら二つの文に関する結束性の知識を、次のように記述する。

送る＜X1,論文＞,送付する＜X2,論文＞

これに対し、動詞が類義語や反意語でない場合でも、結束性が成立する場合がある。タイプ3の場合である。

(例13)「論文を読みましたか」
　　「論文を送って下さい」

これら二つの文に関する結束性の知識を、次のように記述する。

読む＜X1,論文＞,送る＜X2,論文,Z2＞

タイプ1から3までは、同一名詞を仲立ちとした結束性に関する定義である。それに対し、タイプ4、5、6は、類義語の名詞をを仲立ちとした結束性に関する定義である。タイプ7、8、9は、構文情報がからむ場合である。「申込みの期限」の「申込み」のような「AのB」という名詞句の修飾語の名詞を共有して、結束性が成立する場合である。タイプ10、11、12は、「申込み期限」の「申込み」のように複合語の一部を共有して、結束性が成立する場合である。タイプ13、14、15は、タイプ7、8、9の共有する名詞が類義語になった場合、タイプ16、17、18はタイプ10、11、12の共有する名詞が類義語になった場合である。ここでは、以上に示したタイプに、かなりの現象が収まっているので、この範囲にとどめておく。

3.2 Local cohesive knowledge

結束性に関する知識を、トピックスに関する省略や照応関係の同定や代用表現の補完に応用する。結束性に関する知識を制約条件とし、もし、そのような状況が成立したなら、こう解釈しようという解釈部分を付与する。この制約条件と解釈部分を含めて、Local cohesive knowledge(図3)と呼ぶことにする。解釈の結果は"＝＞"で、示すことにする。

```
local cohesive knowledge（簡略表現）
    a＜X1,Y1,Z1＞,b＜X2,Y2,Z2＞,          ;制約条件
        ＝＞(解釈規定).
```

図3 Local cohesive knowledge

(1)照応関係の解釈:　結束性に関する知識の媒介となる名詞を代名詞で置き換えても、結束性は成立する。

(例13)「論文を送りましょうか」
　　「それを送って下さい」

その代名詞の意味するところは、媒介とする名詞を指す(それ＝論文)と解釈する。このことを、図4の(b)のように書く。

(2)トピックスに関する省略の解釈: 結束性に関する知識の媒介となる名詞が省略されても、結束性は成立する。

(例文14) 「論文を送りましょうか」
「[∅ヲ] 送って下さい」

その省略は、媒介とする名詞を指す(∅＝論文)と解釈できる。このことを図4の(c)のように書く。

```
local cohesive knowledge (簡略表現)
(a)送る<X1,論文,Z1>,送る<X2,論文,Z2>.          ;タイプ(1)
(b)送る<X1,論文,Z1>,送る<X2,それ,Z2>,          ;制約条件
      =>  それ＝論文.                          ;解釈規則
(c)送る<X1,論文,Z1>,送る<X2,∅,Z2>,             ;制約条件
      =>  ∅＝論文.                             ;解釈規則
```

図4 Local cohesive knowledgeの例(1)

(3) Referent transferの解釈:

(例文15) 「電話はありますか」
「電話は00-9933-999です。」

答えの文の「電話」は「電話番号」を意味で用いられている。Referent transferの例の一つである。この解釈をきめるためには、解釈部に、「電話番号」と記述しておけばよい(図5.a)。

(4)類義語の解釈: 類義語の解釈も、Referent transferの解釈と同様、local cohesive knowledgeの解釈部に個々の解釈を記述する。なぜ、このように個別的に解釈を記述するのかというと、そもそも解釈というものが、個々の文脈的な状況において、それぞれ異なり、ほとんどルール化しても意味がないからである。例えば、

「電話はありますか」
「番号は00-9933-999です。」

この文脈では、答えの文の「番号」は、「電話番号」を意味しているが、

「クレジット･カードはありますか」
「番号は00-9933-999です。」

```
local cohesive knowledge
簡略表現
  (a) ある<電話>,です<電話,Y2>          ;タイプ(3)
      =>電話＝電話番号
  (b) ある<電話>,です<番号,Y2>          ;タイプ(6)
      =>番号＝電話番号
  (c) ある<電話>,です<電話番号,Y2>      ;タイプ(12)
```

図5 Local cohesive knowledgeの例(2)

という文脈では、「番号」=「クレジット･カードの番号」を意味することになる。大切なのは、解釈を導きだすプロセスをルール化することではなく、解釈が成立するための条件をいかに書き下せるかということである。結束性に関する知識とは、文脈上における解釈を決めるためのかなり強い制約になっているので、個々の解釈を記述することができるのである。

3.3 文脈処理の方法

ここでは、結束性に基づく文脈処理の機構について提案する。システムの流れ図を、図6に示す。入力文を、Lexical-functional Grammar (LFG) (16),(17)LFG に基づいた文法規則、および、辞書を使って解析する。その結果、中間表現(F-structures)(16)が得られる。この中間表現は、文脈処理するには、情報量が多いので、図7に示すようなスケルトン(骨格となる情報のみの構造)に変換される。

解析されたスケルトンと履歴上のスケルトンでペアをつくる。スケルトンのペアと、"local cohesive knowledge"をユニフィケーションさせ、もし、"local cohesive knowledge"の制約が満たされるなら、そのペアは、結束性があると考える。その結果として、省略･照応、代用表現の解釈が得られる。もし、マッチする"local cohesive knowledge"がない場合には、そのペアは、結束性がないと判断し、他の組合せを処理する。履歴中の全ての組合わせをとるのではなく、localな範囲(現在のバージョンでは、一発話前の履歴まで)でチエックする。

例えば、図7に例3のスケルトンとLocal cohesive knowledgeを記した。答えの文の目的格が省略されているので、Local cohesive knowledgeの(a),(b),(c)の内の(c)の条件にマッチし、結果として、「クレジット･カード」が補完される。

[簡略表現と厳密化した表現]

ここでは、構文情報との関係を厳密に表現するために、LFGの記述法を採用する。というのは、先に示した簡略表現の方が、直感的に理解しやすいし、紙面も節約できるが、その反

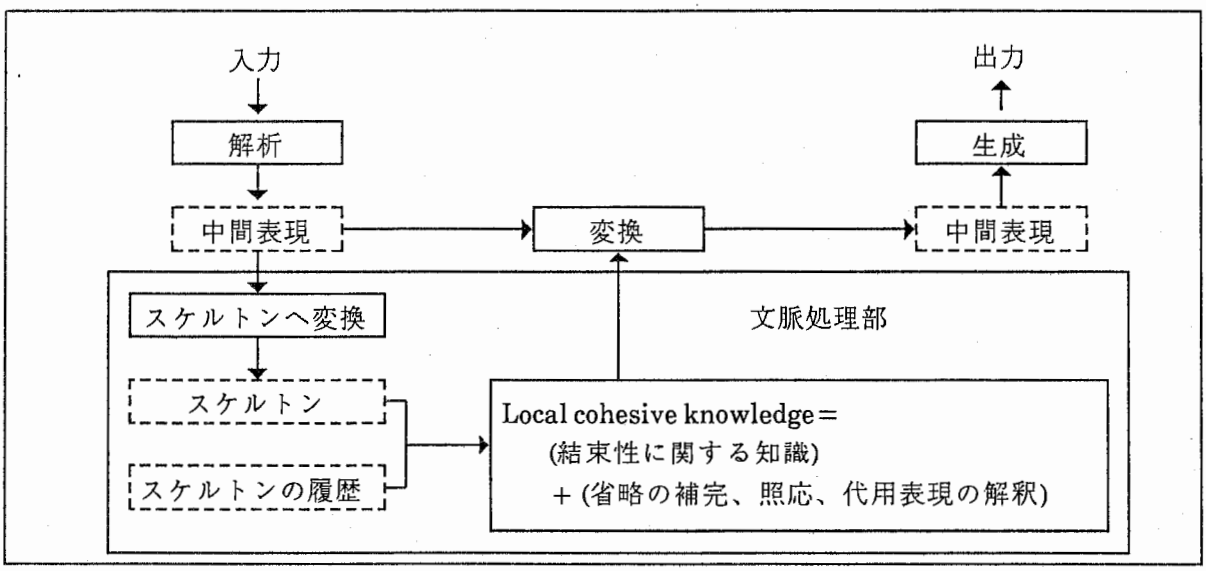


図6　対話文翻訳システムにおける文脈処理方法

(例3)　「クレジット･カードの番号を教えて下さい。」
　　「すみません。持っていないのですが。」

スケルトン(簡略表現)

　教える<∅, 番号(クレジット･カード),∅>,持つ<∅, ∅>

Local Cohesive knowledge

(a)教える<X1, 番号(クレジット･カード), Z1>,持つ<X1, クレジット･カード>.

(b)教える<X1, 番号(クレジット･カード), Z1>,持つ<X1, それ>,

　　=> それ=クレジット･カード.

(c)教える<X1, 番号(クレジット･カード), Z1>,持つ<X1, ∅>,

　　=> ∅=クレジット･カード.

図7. スケルトン(上)と local cohesive knowledge (下)

スケルトン(LFGの記述法);

「クレジット･カードの番号を教えて下さい」のスケルトン

　(f_1 PRED)='教える<(f_1 SUBJ),(f_1 OBJ2),(f_1 OBJ)>',

　　(f_1 SUBJ)=f_2, (f_2 PRED)=∅,

　　(f_1 OBJ2)=f_3, (f_3 PRED)=∅,

　　(f_1 OBJ)=f_4, (f_4 PRED)='番号',

　　(f_4 MOD)=f_5, (f_5 PRED)='クレジット･カード'.

∅は、省略されているものを意味する。MODは、修飾しているものを意味する。

「持っていないのですが」のスケルトン

　(f_{10} PRED)='持つ<(f_{10} SUBJ),(f_{10} OBJ)>',

　(f_{10} negative)=+,

　　(f_{10} SUBJ)=f_{11}, (f_{11} PRED)=∅,

　　(f_{10} OBJ)=f_{12}, (f_{12} PRED)=∅.

local cohesive knowledge (LFGの記述法);

　($\uparrow_1$ PRED)=$_c$教える<($\uparrow_1$ SUBJ),($\uparrow_1$ OBJ2),($\uparrow_1$ OBJ)>,　　(N.B)$\uparrow_n$はメタ変数

　　($\uparrow_1$ OBJ)=$_c\uparrow_3$,

　　　　($\uparrow_3$ PRED)=$_c$番号,

　　　　($\uparrow_3$ MOD)=$_c\uparrow_4$, ($\uparrow_4$ PRED)=$_c$クレジット･カード,

　($\uparrow_2$ PRED)=$_c$持つ<($\uparrow_2$ SUBJ),($\uparrow_2$ OBJ)>

　　(a) ($\uparrow_2$ OBJ)=$_c\uparrow_5$, ($\uparrow_5$ PRED)=$_c$クレジット･カード

　　or(b) ($\uparrow_2$ OBJ)=$_c\uparrow_5$, ($\uparrow_5$ PRED)=$_c$それ

　　=>($\uparrow_5$ ANAPHORA)=クレジット･カード,

　　or (c) ($\uparrow_2$ OBJ)=$_c\uparrow_5$, ($\uparrow_5$ PRED)=$_c$∅

　　=>($\uparrow_5$ ELLIPSIS)=クレジット･カード,

図8. LFGの記述法によるスケルトン(上)と local cohesive knowledge (下)

面、簡略表現には陽に表現されていない情報も存在しているからである。例えば、省略を補うと書いても、どの部分の省略なのか、明確ではない。それに対し、LFGの記述法では、一つ一つの制約が、独立した記述になっており、個々ばらばらに記述しても、メタ変数($\uparrow_n$)があるので、構文構造上のどの部分に対する制約であるのか、識別が容易である。図7の簡略表現を、LFG表現で示したのが、図8である。

図8におけるlocal cohesive knowledge中の

($\uparrow_1$ PRED)$=_c$教える<($\uparrow_1$ SUBJ),($\uparrow_1$ OBJ2),($\uparrow_1$ OBJ)>　　①

は、スケルトン中の

(f_1 PRED)='教える<(f_1 SUBJ),(f_1 OBJ2),(f_1 OBJ)>　　②

に対する制約になっている。①式における"$=_c$"は、制約(constraint)を意味し、入力に対する条件になっている。①式と②式がユニファイした際に、メタ変数($\uparrow_1$)は、②式のもつ実変数(f_1)に置き換えられる。図8の場合は、$\uparrow_1=f_1$, $\uparrow_2=f_{10}$, $\uparrow_3=f_4$, $\uparrow_4=f_5$, $\uparrow_5=f_{12}$に置き換えられる。結果として、

(f_{12} ANAPHORA)=クレジット·カード,

が得られる。"f_{12}"という実変数により、「持つ」の「OBJ」に対する省略であることが明記される。

このようにLFG表記をすることにより、結束性に関する知識と構文情報とのマッピング関係を明確にすることができる。

4 結束性に関する知識の生成について

4.1 結束性に関する知識の生成方法

ここで述べたアプローチは、知識を記述することにより、問題解決を計っていこうというアプローチであるので、知識を大量に生成する必要がある。そこで、知識の生産性ということが重要な課題になる。以下の三通りの方法を検討した。

(1)人手抽出する方法

データの収集は、キーボード会話(国際会議の申込みに関する会話が12会話、全体で約430センテンス、日英対応データ)について、作業フォームを決め分析をおこなった。

1) 選択基準が、人によって揺らぐ。すなわち、結束性の基準が、人の判断に依存しているため、不明確になってしまう。

2) 生産性が低い。対象としている会話データの中には、動詞は97,名詞190含まれており、約500対について整理をおこなった。この作業に、約2か月くらい要し、しかも、網羅的に知識を抽出できないという問題が残る。

(2)生成による方法

網羅的に知識を抽出するために、全ての可能な組合せを生成して、その組合せから、不適合なものを除去するという方法が考えられる。しかし、この方法には、二つの問題点がある。

1) まず、可能な組合せを生成して、不適合なものを除去するのは、人間には難しい。不適合なものを除去するためには、その組合せが起こらないことを考なくてはならないからである。いろんな場面を想定して、結局、判断がつかないことが多い。従って、人間にこのような判断させるのは、そもそも無理である。

2) また、コストが組合せ的にかかる。例えば、3000語程度の語彙を処理するシステムを構成しようとした場合、ATRのコーパスでは、その3000語彙の中に動詞が350、名詞が約1000含まれている。動詞と名詞の組合せが、35万とおりある。その内、実際に

$350 \times 1000 = 350{,}000$

$350{,}000 \times 3\% = 10500$

$350{,}000 \times 7\% = 24500$

動詞と共起する名詞の割合が、3%から7%であると過程すると、10500から24500の動詞、名詞のペアが作られる。これらの動詞、名詞のペアの組合せは、

$10500 \times 10500 = 1.10 \times 10^8 <$

生成される組み合わ$< 24500 \times 24500 = 6.00 \times 10^8$

11億とおりから60億とおりの場合を処理しなければない。それを処理するのに、1日1万とおり(現実には、かなり、この量をこなすのは、難しい)やったとしても、1万日から6万日かかることになる。

しかも、この作業は、大多数の可能性を振り落とす作業である。実際に可能性を全て列挙して、これらの作業を行ってみれば直ぐにわかることだが、起こりうる現象は、ごくわずかで

あることに気がつく。(結束性の知識に関するスパース性)結局、生成される現象と観測される現象の間に、大きなずれが生じていることになる。

(3)言語データ·ベースを利用する方法: この方法は、計算機で処理するため、コスト面、および、ゆれがなく、実用的な知識の抽出という意味で、他の二つの方法より優れている。言語データ·ベースを構築するコストをいれても、(2)の方法より、安上がりにできる。その理由は、(2)の方法では、数量の組合せ的な処理をしなければならないのに比べて、この方法は、観測されるデータに基づいて処理するため、そのドメインで比較的よく使われる知識が抽出できる。

[結束性に関する知識の生成方法]

3.1の形式的定義に基づいて抽出する。言語データ·ベースとして付加されている情報は、表2に示すようにある動詞の格要素に、どんな名詞がきているかというテーブルが構成されていればいい。任意の二つの動詞を選び、共通する名詞の存在を抽出すればよい。例えば、「参加する」と「行う」という動詞は、「会議」を共通の名詞としているので、「会議」を仲立ちとした結束性がある。表2からは表3に示す結束性に関する知識が抽出される。

表2 言語データ·ベースからの知識の抽出方法の例

動詞	名詞			
	主格	目的格	第二目的格	任意格
教える		参加料,資料,会議,プログラム,研究,申込み用紙,用紙,...		
参加する		会議,.プログラム		
持つ		アブストラクト,学生書,参加料,資料,感心,希望,興味,...		
送る		アブストラクト,参加料,料金,申込み用紙,用紙,.....	あて先,先生	
行う		フォーラム,発表,授業,会議,研究,セッション,....		午後,後,ホテル,今

表3 (表2)から抽出した知識の例

結束性のある動詞対		媒介となる名詞
教える	参加する	会議,プログラム
教える	持つ	参加料,資料
教える	送る	参加料,申込み用紙,用紙
教える	行う	会議,研究
参加する	行う	会議
持つ	送る	アブストラクト,参加料

[ATRのコーパス(18),(19),(20),(21),(22),(23)]

キーボードを媒介とした会話文が、60会話、述べ語数約7万語が収録され、データ・ベース化されている。異なり語数は、約3000語、その内動詞350語、名詞1000語が含まれている。

[結果]

結束性のタイプ(1),(2),(3)については、結束性のある動詞対が、3501とおり得ることができた。約1週間程度で処理が可能であった。タイプ(7),(8),(9)については、「AのB」というテーブルを作成して、結束性のある動詞対で217とおり抽出できた。タイプ(10),(11),(12)については、複合語のテーブルが完成していないので、抽出できていない。また、タイプ(4),(5),(6),(13),(14),(15),(16),(17),(18)については、類義語を含むので、機械的には、生成できていない。

この抽出された知識を調べてみると、関係づけが難しい現象も含まれている。例えば、

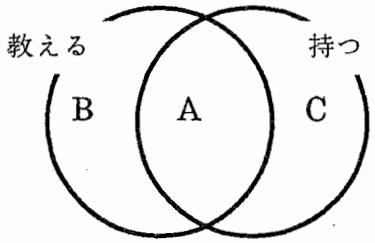
「発表の申込みについて教えてください」

「発表は、締め切られました」

このような現象は、数多く抽出される。このアプローチでは、結束性に関する知識として、記述しておけばよい。また、従来のアプローチで推論するにせよ、どのような現象が起こるのか抽出しておかなくては、ルール化できない。現象の生成ということが、重要な意味を持つと思われる。

[実際の抽出例]

「教える」と「持つ」という動詞について、コーパスから名詞構成要素を取り出したのが、図9である。「教える」と「持つ」の動詞対が、図9のAに示す名詞を媒介として、結束性を持っている。



A=[アメリカンエキスプレス、番号、電話、概要、学割、(御)質問、ホテル代、時間、カード、会議、観光料金、小切手、講演、口座、口座番号、クレジット·カード、申込書、申込み手続、申込み用紙、ペイパー、プログラム、論文、料金、差額、参加、参加料、参加料金、資料、食費、書類、宿泊費、取り引き銀行、登録費、登録申込み書、登録料、登録用紙、通訳、通訳者、割引、用件、要綱集、用紙、要約]

B=[アコモデーション、新田支店、あて名、あて先、ATR、バス、部数、電話番号、駅、演題、額、学生、学生割引、言語、議題、銀行、銀行振り込み、(ご)専門、(御)用件、払い戻し、平安神宮、ホテル、方法、住所、会場、会議場、会期、会社、会社名、観光、観光ツアー、関連、方、方々、為替、数、研究、研究発表、期限、期日、期間、金額、金閣寺、北大路、国際会議、込みぐわい、こと、言葉、講演者、交通、交通機関、キャンセル、キャンセル料、距離、京都ホテル、京都ロイヤルホテル、京都、京都駅、京都プリンスホテル、前、名勝、道順、目的、申込み、申込み方法、無料、名前、ネーム、二条城、送り先、(お)名前、大阪、旅館、旅行、、サービス、サイト·シーイング·ツアー、サイト·シーイング、先、参加者、専攻、セッション、世話、市バス、支払い、資格、氏名、締切り、市内観光、支社、支店、スピーカー、住友銀行、詳細、タクシー、手配、テーマ、点、手続、近く、ところ、取り引き、、東京、登録、討論、ツアー、追加料金、通訳電話、聴講、受付、運賃、話題、予定]

C=[アブストラクト、アメリカ·ドル、分野、ドラフト、ドル、同伴、英文、ファースト·サーキュラー、原則、学生書、現金、原稿、印鑑、会、感心、計画、希望、期待、子供、コピー、クラブ、局、興味、娘、日本円、プラン、レジストレーション、リコンファーム、両方、サマリー、責任、質問、施設、スライド、知識、登録料金、ツイン、所、円、要望、予約、全額],

図9.実際の抽出例

4.2 結束性に関する知識のスパース性

結束性に関する知識は、対となりうる組み合せは想像していたより少ない。そこで、その性質を統計的に調べてみると次のことが分かる。(1)動詞と動詞が、結び付く割合は、3割程度である。また、(2)ある動詞の格要素に成りうる名詞は、せいぜい、2~3割で、全ての名詞がとれるわけではない。

[動詞と動詞の結束率]

結束率とは、あるコーパスにおいて、ある動詞が、結束性のある動詞対をどの程度の割合で作っているかを示す数字である。

$$\text{ある動詞の結束率(\%)} = \frac{\text{その動詞を含む結束性のある動詞対数} \times 100}{\text{動詞の異なり語数}}$$

例えば、ATRのコーパスでは、「教える」という動詞は、97個の動詞と結束性のある動詞対を形成したので、結束率は、27.7%である。

「教える」の結束率 = 97×100/350 = 27.7(%)

表4 ATRコーパスにおける結束率 (%)　350動詞中の上位60個

動詞	結束率	個数
教える	27.7	97
する	24.9	87
送る	24.6	86
ある	23.4	82
持つ	22.6	79
なる	22.6	79
聞く	18.9	66
話す	17.4	61
お願いする	17.4	61
知る	17.4	61
含む	17.1	60
言う	16.0	56
伺う	16.0	56
開く	15.1	53
おっしゃる	13.7	48
取る	13.4	47
致す	13.1	46
待つ	13.1	46
行う	12.6	44
使う	12.3	43
考える	11.7	41
来る	11.1	39
分かる	10.9	38
扱う	10.6	37
存じる	10.6	37
入る	10.3	36
締め切る	10.3	36
募集する	10.0	35
書く	10.0	35
受ける	9.71	34

動詞	結束率	個数
出す	9.43	33
入れる	9.43	33
終わる	9.43	33
掛かる	9.14	32
書ける	9.14	32
答える	9.14	32
開催する	8.57	30
閉まる	8.29	29
会う	8.00	28
限る	8.00	28
付ける	8.00	28
引く受ける	7.71	27
居る	7.71	27
決める	7.71	27
伝える	7.71	27
外す	7.43	26
貰う	7.43	26
参加する	7.43	26
聴講する	7.43	26
近づく	7.14	25
運営する	6.86	24
始まる	6.57	23
祈る	6.57	23
断る	6.57	23
確かめる	6.57	23
立てる	6.57	23
喜ぶ	6.57	23
返す	6.29	22
見る	6.29	22
出る	6.00	21

このコーパスにおける結束率の最大値は、「教える」の結束率、27.7%であり、ある動詞が全ての動詞とは結びつかないことを示している。各動詞別の結束率を表4に示す。

このコーパスにおける動詞全体の結束率は、

$$\text{動詞全体の結束率(\%)} = \frac{\text{結束性のある動詞対の総数} \times 100}{\text{理論上の動詞対の組合せ}}$$

このコーパスからは、350個の動詞があり、その動詞の集合から3718個の結束性のある動詞対が生成されたので、

動詞全体の結束率(%)=

3718個×100/350×350 = 3.04(%)

ある名詞を媒介として、結束性のある動詞対を形成する割合は、このコーパスにおいては、3%程度であり、その割合は極めて低い。すなわち、結束性は、少なくとも、このコーパスにおいては、かなりスパースであることを示している。

[結束性の有無の相関図]

動詞間の結束性の有無を図示したのが、図11である。○は、動詞と動詞の間に結束性が成立している例、●は、現段階の観測データからは、動詞と動詞の間に結束性が成立していないことを示している。「教える」という動詞は、ほとんど、すべての動詞となんらかの名詞を介して、結束性が成立していることを示している。一方、その反対に、「構う」という動詞は、ほとんど全ての動詞と、名詞を介しては、結束していないことを示している。

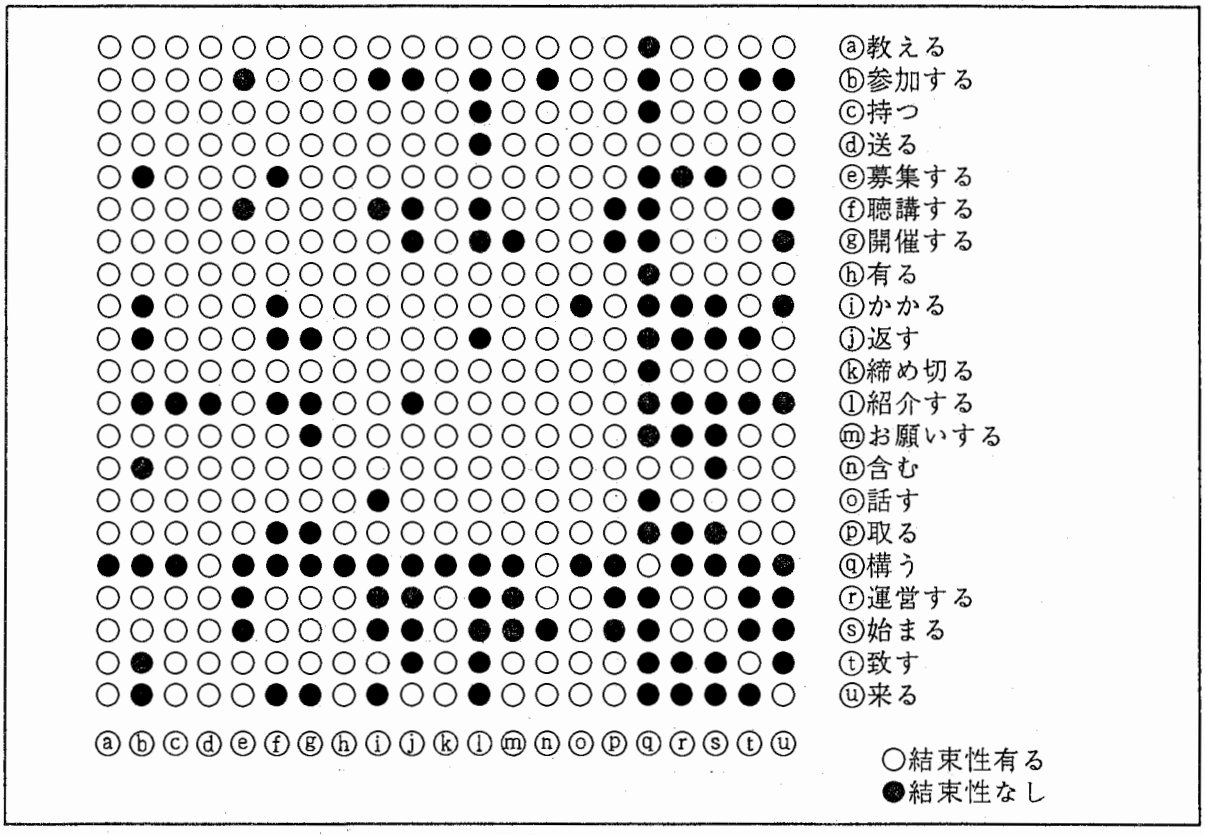


図11 結束性の有無の相関図

[動詞の各要素の個数]

また、全ての動詞が、結束性のある動詞対を形成しないのと同様に、全ての名詞が、ある動詞の格要素になるわけではない。このことを検証するために、1000語の名詞に、何個の割合で名詞が、格要素に成りうるかを調べた。その名詞数を表5のB欄に示した。格要素になる名詞の割合は、通常5%程度であり、最大でも、せいぜい2、3割程度であり、全ての名詞が格要素になるわけではないことが分かる。従って、結束性に関する知識は少なく、表層で記述しても、問題が起こらない。

4.3 格飽和率

結束率が低いのは、動詞間の性質ではなく、サンプリングに問題があるからだという考えもできる。そこで、このコーパスからどの程度の割合で知識を抽出することができたのかを示す尺度として、格飽和率を考える。表5のA欄に、実際にコーパスに現れた名詞の異なり語数(X)を、B欄に、1000語の中からその格に入りうる名詞数(Y)を、C欄に、格飽和率を示す。

表5 主な動詞の格飽和率

A＝出現単語数,B＝1000語中におけるその格に入りうる名詞数,C＝格飽和率(%)

動詞	A	B	C(%)
教える	153	260	58.8
参加する	2	2	100
持つ	68	88	77.3
送る	79	92	85.9
募集する	43	48	89.6
聴講する	8	13	61.5
開催する	19	23	82.6
ある	116	128	90.6
かかる	23	31	74.2
返す	16	16	100
締め切る	16	16	100
紹介する	11	25	44.0
お願いする	96	120	80.0
含む	35	45	77.8
話す	21	28	75.0
取る	18	36	50.0
構う	18	18	100
運営する	6	8	75.0
始まる	12	22	54.5
致す	20	112	17.9
来る	12	19	63.2
平均	37.7	54.8	74.2

$$\text{格飽和率(動詞、格)} = \frac{X \times 100}{Y}$$

例えば、「教える」という動詞の目的格には、153とおりの名詞がこのコーパスに出現した。1000語の名詞のうち260の名詞が入りうるので、

$$\text{格飽和率(教える, OBJ2)} = \frac{153}{260} = 58.8(\%)$$

の名詞について、すでに観測されたことを示している。表5にあげた、21種類の動詞の平均で、74.2(%)であるので、これらの格に対しては、よく飽和していると考えられる。

[課題]

(1)知識の抽出上の課題:

●現状では、「です」に関する結束性の処理が漏れている。

●類義語には、形式的な定義をあたえることができず、人間が意味的な判断をしなくてはならない。自動抽出への障害である。

●動詞の出現頻度が低く、格要素が省略されているものについては、漏れてしまう。

(2) シンタックスに関する制限

●現在のところ、単純名詞句、AのBの2種類のみ扱っている。今後、並列句、AのBとC、数詞、うめこみ文を含む名詞句の構造についても、検討していきたい。

(3)他の品詞の結束性

●構文情報に関する拡張、および、動詞と名詞以外の他の品詞を構成要素としたモデルについて、拡張したい。

(4)応用:

●対話翻訳システムの文脈処理のモジュールとしてインプリメント行い、評価を行っている。

●命題の管理機構への応用: 結束性に関する知識(Local cohesive knowledge)を応用し、対話の進行とともに、dynamicに命題の真偽値を管理するモデルを構築を行っている。ここで提案した文脈処理機構を、命題の管理機構のドライバーとして利用できる。

5.むすび

文脈が乱れたときも解析を行えるような文脈的なロバスト性を実現するためには、結束性という考え方が重要な働きをする。そのための知識として、結束性に関する知識を、構文情報との関連づけて、形式的定義を与えた。この定義に基づき言語データベースから効率よく生成する方法について述べた。その結果、文と文が文脈的に結び付く割合は、比較的低いことが分かった。我々のコーパスでは、その割合(動詞の結束率)は、最大のものでも28%程度、平均で3%程度であった。従って、結束性に関する知識を抽象化せずに、表層の情報を使って記述しても、現在の計算機環境を考えると記述量の問題はないといえる。

謝辞

この研究に際しまして、ご支援いただいたATR自動翻訳電話研究所樽松明社長、ならびに、飯田仁主任研究員、小倉健太郎主任研究員(現在NTT復帰)、有田英一男研究員(現在三菱電機復帰)、橋本一男研究員、篠崎直子研究員(現在退社)に感謝いたします。

参考文献

(1)Hobbs, Jerry.R:"Coherence and Co-reference", Cognitive Science, 3(1), PP.67-90.

(2)石崎、井佐原「文脈処理技術」情報処理学会誌、Vol.27, No8, 1986.

(3)情報処理:[大特集:自然言語理解」Vol30,No10,(1989).

(4)堂坂、小暮「対話参加者に関するゼロ代名詞の同定」情報処理学会第39回全国大会、5F-5,1989.

(5)堂坂浩二「対話登場人物を指示する日本語ゼロ代名詞の同定」ATR Technical Reports, TR-I-0117.

(6)野垣内、飯田「キーボード会話における名詞句の同一性の理解」情報処理学会自然言語処理研究会72-1, (1989.5.19)

(7)野垣内、飯田「対話における名詞句の同一性とその応用」ATR Technical Reports, TR-I-01094.

(8)土井、北橋「発話対に基づく対話解析」ATRワーキンググループ資料、1990.1.22.

(9)有田、飯田「日本語におけるタスク・オリエンテッドな対話の構造」電子情報通信学会 言語処理とコミュニケーション研究会資料、NLC87-10, (1987).

(10)有田、飯田「対話翻訳のための階層型プラン認識モデル」ATR Technical Reports, TR-I-0067.

(11)有田、飯田「目標指向型対話におけるドメイン知識の調査」ATR Technical Reports, TR-I-0068.

(12)飯田、有田「対話理解のためのプラン認識モデル」昭和63年人工知能学会全国大会第二回.

(13)佐川、杉原、杉江「柔軟な対応制御機構を持ったコンサルテーション・システム」情報処理学会論文誌 Vol29,NO4.(1988).

(14)加藤、中川「自然言語インターフェイスシステムにおける意図の把握と話題の管理」情報処理学会論文誌 Vol29,NO9.(1988).

(15)平井、北橋「語用論的制限を用いた日本語文の省略の補充」人工知能学会全国大会第一回論文集7-4 (1987).

(16) Kaplan, R.M.& Bresnan, J. 'Lexical-Functional Grammar: A Formal System for Grammatical Representation' In:Bresnan, J. (ed) 'The Mental Representation of Grammatical Relations', The MIT Press, Cambridge, Massachusetts, pp.173-281 (1982).

(17) Kudo, I. & Nomura, H. 'Lexical-functional Transfer: A Transfer Framework in a Machine Translation System Based on LFG', Proceedings of 11th International Conference on Computational Linguistics, Bonn, August, pp.112-114 (1986).

(18) Ogura Kentaro, Hasimoto Kazuo & Morimoto Tyuyosi (1988) :"An Integrated Linguistic

Database Management System",ATR Technical Reports, TR-I-0036.
(19) Kentaro Ogura, Hasimoto Kazuo & Morimoto Tyuyosi (1989) :"Object-oriented User Interface for a Linguistic Database" A working conference on Data and Knowledge base Integration, October 5-7,89, University of Keele, England.
(20)小倉、橋本、森元「言語データベース統合管理システム」情報処理学会 自然言語処理研究会69-4, (1988.12.6).
(21)小倉、橋本、森元「言語データ・ベース統合管理システム」ATR Technical Reports, TR-I-0036.
(22)橋本、小倉、森元「フレーム表現による検索機能を有する言語データベース管理システム」 Proceedings of Advanced Database Symposium' 89 , Kyoto Research Park, Kyoto, Japan, December 7-8, 1989.
(23)篠崎、水野、小倉、吉本「形態素情報利用解説書」ATR Technical Reports, TR-I-0077.