二言語対訳コーパスからの動詞の格フレーム獲得

宇津呂 武仁† 松本 裕治†† 長尾 眞†

自然言語処理のための大規模な意味辞書を構築するためには，人間のための辞書や大規模コーパスに含まれる自然言語の文を解析して，そこから意味辞書を構築する技術を確立することが重要となる．計算機で知識獲得を行う場合，全自動で知識が獲得されることが望ましいが，現在利用可能な情報が貧弱であるため，有用な知識を獲得するためには何らかの人間の介入が必要である．しかし，最終的に得られる結果が人間の主観的な判断の影響を受けないように，人間の介入は最小限に抑えたい．我々は，英語と日本語のように統語構造および語彙が異なる二言語間の翻訳例を構文解析して，その結果を二言語間で比較するというアプローチによって語彙的知識の獲得を行っている．そこでは，両言語の解析結果を比較することによって統語的および意味的曖昧性の両方が解消するため，単言語だけのアプローチに比べると人間の介入を大幅に抑えて語彙的知識を獲得できる．本論文では，二言語対訳コーパスから日本語の動詞の表層格フレームを獲得する手法について述べる．我々の手法では，システムと人間との相互作用は，動詞の複数の意味を類別する部分だけに許される．そこでは，システムが動詞の複数の意味を類別する手がかりをヒューリスティックスによって発見し，その妥当性を人間が判定するという形で相互作用が行われる．その際には，対訳例の英語の情報が有力な手がかりとなる．

Verbal Case Frame Acquisition from Bilingual Corpora

Takehito Utsuro,† Yuji Matsumoto†† and Makoto Nagao†

It is important to devise techniques of compiling semantic dictionaries from natural language corpora. Since currently available knowledge resources are too poor to extract lexical knowledge from natural language texts in a fully automatic way, human interaction is necessary to some extent, while it should be kept to a minimum in order to make the results stable. Our approach makes use of translation examples in two distinct languages that have quite different syntactic structures and word meanings (such as English and Japanese). In many cases, both syntactic and semantic ambiguities are resolved by comparing analyzed results of both languages, and it becomes possible to extract lexical knowledge with much less human interaction compared with monolingual approaches. This paper describes a method for acquiring surface case frames of Japanese verbs from bilingual corpora. The interaction is limited to the crucial points that the system detects for discriminating multiple senses of verbal case frames, then human gives some decision to it. The system reduces human interaction by virtue of English information as the heuristics for the discrimination.

1. はじめに

自然言語処理の技術を実用上有効なものとするためには，計算機処理のための大規模な意味辞書の構築が必要不可欠である．そのような大規模な意味辞書には，個々の言葉について，様々な種類の意味的あるいは語彙的な知識を記述しておく必要があり，例えば，(a)概念の集合および概念と表層語との関係．(b)概念間の階層関係あるいはシソーラス．(c)名詞概念の属性や動詞概念の格要素などの，概念間の非階層的関係．といった知識がこれに該当する．大規模な意味辞書の構築を実現するためには，特に，人間が読むための辞書や大規模なコーパスなどに含まれる自然言語の文を解析して，大規模な意味辞書を半自動で構築する技術を確立することが重要である．しかし，自然言語の文を解析する際には，1) **統語的曖昧性**および 2) **語彙の多義性**の問題が生じる．我々は，これらの問題を解消するために，英語と日本語のように統語構造および語彙が異なる二言語間の翻訳例を構文解析して，その結果を二言語間で比較するというアプローチを提案した[12),13)]．英語と日本語のように統語構造の異なる

† 京都大学工学部電気工学第二教室
Department of Electrical Engineering, Faculty of Engineering, Kyoto University
†† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Advanced Institute of Science and Technology, Nara

二言語の場合には，それぞれの言語が異なったタイプの統語的曖昧性を持つために，多くの場合，二言語間で構文解析結果を比較することによって統語的曖昧性が解消されると考えられる．また，概念が二言語間の中間言語的なものであると考えれば，二言語間の訳語のペアが持つ概念の集合は，単言語の表層語に対応する概念の集合の交わりとなると考えられるので，二言語間の訳語のペアを用いることによって概念レベルでの曖昧性が減少する[2),9)]．

このような考えに基づいて我々が提案した「二言語対訳コーパスからの語彙的知識獲得」[12),13)]の枠組を**図1**に示す．図1においては，まず，対訳コーパス中の対訳例の各単言語文を統語的な知識だけを用いて構文解析し，構文解析結果を表層語の依存構造とほぼ等価な素性構造に変換する．次に，対訳辞書の訳語の情報を参照することによって，二言語間で素性構造を照合する．照合の結果，構文解析結果の統語的曖昧性のうちのいくつかが解消する．例えば，例1の対訳例の場合は，英文中の前置詞句 "*on the hook*" について前置詞句付加の曖昧性があるが，二言語間の素性構造照合によって以下のような二言語間で照合された素性構造（以下，**対訳素性構造**と呼ぶ）が一意に得られる．

例1

英　語：I hung my coat on the hook.

日本語：私は上着をかぎにかけた．

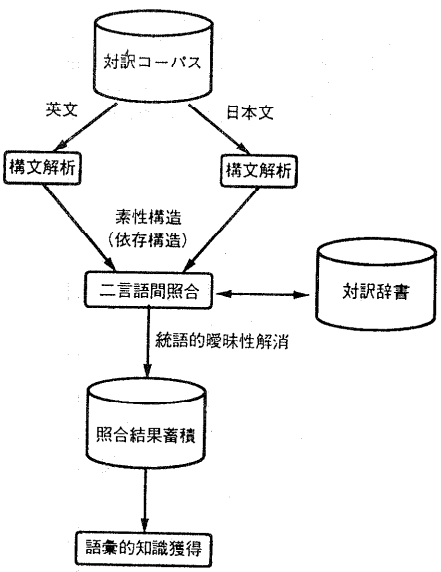

図1　二言語対訳コーパスからの語彙的知識獲得の枠組
Fig. 1　The framework of lexical knowledge acquisition from bilingual corpora.

$$\left[\begin{array}{l} pred : \langle hang,\ \text{かける}\rangle \\ tense : past \\ \langle subj,\ \text{は}\rangle : [pred : \langle I,\ \text{私}\rangle] \\ \langle obj,\ \text{を}\rangle : \left[\begin{array}{l} pred : \langle coat,\ \text{上着}\rangle \\ spec : my \end{array}\right] \\ \langle on,\ \text{に}\rangle : \left[\begin{array}{l} pred : \langle hook,\ \text{かぎ}\rangle \\ spec : the \end{array}\right] \end{array}\right]$$

この対訳素性構造においては，動詞は英語と日本語のアトミックな素性値のペア〈*hang*, かける〉で表現され，また素性名は，〈*subj*, は〉，〈*obj*, を〉，〈*on*, に〉のように英語と日本語の素性名のペアで表現される．格要素の名詞についても，同様の表現が用いられる．このような曖昧性の解消した素性構造を大量に蓄積すれば，そこから動詞の表層格フレームのような語彙的知識を獲得することが可能であると考えられる．

我々はこれまでに，計算機上で利用可能な和英辞典（講談社学術文庫の和英辞典[8)]）から約 40,000 対訳例を取り出し，対訳コーパスとして利用している．日英対訳文間で素性構造を照合することによって対訳例の統語的曖昧性を解消する手法を，この和英辞典の対訳例に適用した結果，日本語単独で一意に解析結果が得られる割合は約 20% であるが，二言語間で素性構造を照合すると，約 60% の割合で動詞の表層格構造が一意に決定した[14)]．本論文では，統語的曖昧性の解消した対訳素性構造を大量に蓄積した結果から，日本語の動詞の表層格フレーム* を獲得する手法について述べる．

一般に，計算機による知識獲得の枠組を設計する場合には，人手の介入がどれくらい必要かについて検討しなければならない．自然言語処理のための知識の獲得の分野では，意味辞書あるいは知識ベースなどを人手で構築するというアプローチによる研究がある[4)~6)]．このアプローチに伴う困難な問題としては，構築された意味辞書の記述に作業者の主観が含まれるために，辞書全体を通して記述が一貫しないなど，結果に不安定な面が出てくることが挙げられる．また，人手で行うために作業の量が膨大であり，意味辞書を拡張することも容易ではない．一方，計算機を利用してできるだけ自動的に意味辞書を構築するというアプローチの研究も行われている．このアプローチの例としては，人間のために書かれた辞書から概念間の階層関係あるいはシソーラスを抽出するという研究[10),11)]や，大規模なコーパスの文を構文解析して統計的な

* 以下，本論文中では，「格フレーム」という言葉はすべて「表層格フレーム」の意味で用いる．

データを蓄積し，そこから語彙的な知識を抽出するという研究[1),3)]がある．計算機を利用したアプローチをとることによって，意味辞書に対する客観的な記述が可能になり，また，大規模な意味辞書を容易に構築することが可能になると期待される．しかし，現在計算機上で利用可能な知識源には限界があるため，有用な語彙的知識を完全に自動で自然言語テキストから取り出すのは容易でないと考えられる．そこで，何らかの人間の介入が必要となるが，人間が必要以上に介入を行ってしまうと，最終的に得られる結果が人間の主観的な判断の影響を受け安定な結果を得るのが困難になるので，人間の介入は必要最小限に抑えるのが望ましい．このような観点から，我々は，知識を獲得する際に，システムが独力で判断を行うのが困難であるような重要なポイントにさしかかった時にだけ人間に質問を行い，人間がそれに対して何らかの判断を与えるという，限定された形の相互作用によって知識獲得を行うことにする．その際，システムと人間との相互作用としては，人間がシステムの示す少数の選択肢から一つを選ぶという形式のものだけを許すことにした．

対訳素性構造を大量に蓄積した結果から動詞の表層格フレームを獲得しようとする場合には，知識獲得の過程は，

1. **一つの動詞が持つ複数の意味・用法を類別する．**
2. **動詞のそれぞれの意味に対して，名詞シソーラスを用いて表層格フレームを構成する．**

という二つの部分に分けられる．この場合，一つの動詞が持つ複数の意味・用法を類別する部分が最も自動化困難な部分であると考え，**図2**に示す方法をとった．すなわち，システムは，一つの動詞について用例（対訳素性構造）の集合が与えられると，何らかのヒューリスティックスによって動詞の意味・用法の類別の手がかりを発見し，発見された手がかりが正しいかどうかを人間に質問し，人間がこれに答える．ここで，我々の方法では，日本語の動詞の意味・用法を類別する手がかりとして，英語の動詞や格ラベルを利用することができるので，単言語だけの場合と比べて質の高いヒューリスティックスを用いることができる．そして，最終的に正しいと判定された類別の手がかりによって，用例の集合を動詞の意味・用法の違いごとに部分集合に分割し，それぞれの意味について格フレームを構成する．

第2章では，名詞シソーラスおよび対訳素性構造のデータ構造について述べる．第3章では，動詞の複数の意味・用法を類別する方法について述べ，第4章では，格フレームを構成する方法を説明する．さらに第5章では，和英辞典の対訳例を集めた対訳コーパスから実際に動詞の表層格フレームを獲得する実験を行った結果について述べ，我々の手法を評価する．最後に，第6章で，全体のまとめおよび今後の発展について述べる．

2. データ構造

2.1 名詞シソーラス

我々は，表層格スロットの名詞に対する意味的制約を記述するために，名詞シソーラス中の意味カテゴリを用いる．名詞シソーラスは，各節点が意味カテゴリを表すような木構造としてとらえることができる．ここでは，SC を名詞シソーラス中の意味カテゴリの集合とし，$\preceq$ を意味カテゴリ間の上位・下位関係とし

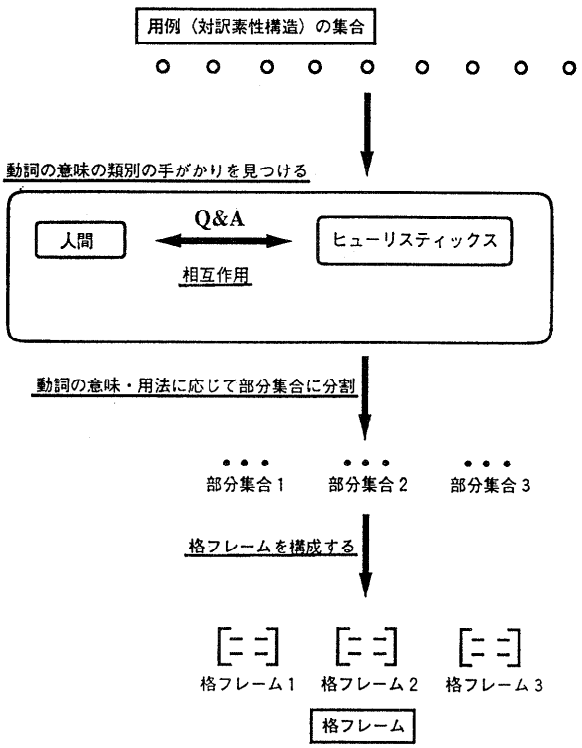

図2 格フレーム獲得の過程
Fig. 2 Process of acquiring case frames of a verb.

て，名詞シソーラスを有限半順序集合 $[SC, \preceq]$ で定義する．$[SC, \preceq]$ は木構造の根節点に相当する最大元 $\top$ を一つだけ持つ．また，二つの意味カテゴリ a および b の極小上界 (least upper bound) を $a \vee b$ によって表す．ここでは，任意の意味カテゴリ a および b について，極小上界が一意に求まると仮定する．

現在のところ，我々は，計算機上で利用可能な日本語のシソーラスとして，「分類語彙表」[7] を用いている．「分類語彙表」は6層の階層構造から構成され，階層構造の葉の部分に総数約 60,000 語の単語が分類されている．また，その体の類（名詞シソーラス）の部分には，総数約 45,000 語の名詞が分類されており，根節点の一つ下のレベルは，「抽象的関係」，「人間活動の主体」，「人間活動—精神および行為」，「生産物および用具」，「自然物および自然現象」という五つの意味カテゴリからなる．一つの単語が多義性を持つ場合には，その語にはシソーラス中の複数の意味カテゴリが割り当てられる．「分類語彙表」が動詞の格フレーム獲得の目的に十分耐えられるシソーラスであるとは言い難いが，現在入手可能な日本語のシソーラスの中では，最も正確で収録語数の最も多いシソーラスである．

2.2 対訳素性構造

対訳素性構造は，二言語間で素性構造を照合した結果得られる素性構造で，データ構造としては，アトミックな素性値および素性名として，二言語の訳語のペアを許した素性構造によって記述される[14]．動詞の格フレーム獲得の場合は，以下のような対訳素性構造を蓄積して格フレーム獲得を行う．

$$\begin{bmatrix} pred : \langle V_E, V_J \rangle \\ voice_e : Voice_E \\ aux_J : Aux_J \\ \langle l_{E1}, p_{J1} \rangle : \begin{bmatrix} pred : \langle N_{E1}, N_{J1} \rangle \\ sem : SEM_1 \end{bmatrix} \\ \vdots \qquad\qquad \vdots \\ \langle l_{En}, p_{Jn} \rangle : \begin{bmatrix} pred : \langle N_{En}, N_{Jn} \rangle \\ sem : SEM_n \end{bmatrix} \end{bmatrix}$$

ここで，V_J および V_E は対応する日本語および英語の動詞を表し，$Voice_E$ は英文の態を表し，また，Aux_J は日本文の助動詞を表す．また，$p_{J1}, \cdots, p_{Jn}$ は日本語の格ラベルを，$l_{E1}, \cdots, l_{En}$ はそれぞれ対応する英語の格ラベルを表す☆．$N_{J1}, \cdots, N_{Jn}$ は日本文の格要素の名詞を表し，$N_{E1}, \cdots, N_{En}$ はそれぞれ対応する英文の格要素の名詞を表す．ある日本語の名詞 N_{Ji} が複数の意味を持つ場合，この名詞は名詞シソーラス中の複数の（葉の位置の）意味カテゴリをもつ．そこで，各 SEM_i を名詞シソーラス中の意味カテゴリの集合によって表現し，これを各日本語名詞 N_{Ji} の**意味ラベル**と呼ぶ．

$$SEM_i = \{Sem_1, \cdots, Sem_k\}$$

用例中の名詞の意味ラベルから格要素の意味的制約を記述する場合には，各名詞の多義性を解消して正しい意味カテゴリ Sem_j を一つ選ばなければならない．

3. 動詞の複数の意味・用法の類別

本章では，まず日本語の動詞の複数の意味・用法を類別する手がかりとしてどのようなものが利用可能であるかを述べ，次に日本語の動詞の複数の意味・用法を類別する手法を提案する．

3.1 類別の手がかり

我々の枠組では，動詞の意味を類別する際に，格要素の日本語の名詞の意味ラベルだけでなく，英語の動詞や英語の格ラベルなどのいくつかの英語の情報が利用可能である．日本語の動詞の複数の意味・用法を類別する手がかりは，これらの情報を利用することによって得られる．

英語の動詞の違い

複数の意味を持つ日本語の一つの動詞を英語に訳す場合，異なった意味に対しては異なった英語の動詞が用いられることが多い．このような場合，対応する英語の動詞の違いが，日本語の動詞の複数の意味を類別する手がかりとなる．例えば，「かける」という日本語の動詞はいくつかの意味を持つが，次の例ではそれぞれの意味に対して異なった英語の動詞が対応する．

窓にカーテンをかける
　/hang curtains in a window
服に金をかける/spend money on clothes
音楽をかける/play some music

英語の格ラベルの違い

ある日本語の一つの動詞が複数の格フレームを持ち，それぞれの格フレームにおける動詞の意味・用法が異なる場合に，同じ格ラベルでマークされる格が，

☆ ここで格ラベルとは，表層格ラベルのことである．表層格ラベルとなり得るものは，日本語の場合は，格助詞・係助詞・述語接続助詞・引用助詞などであり，英語の場合は，"*subj*"・"*obj*"・前置詞・接続詞などである．日本語の格ラベルが係助詞「は」などの場合で，英語の情報を参照することによって「が」格または「を」格であることが一意に決定できる場合には，そのように変換しておく．

動詞の意味・用法に応じて異なった意味で用いられることがよくある．例えば，「書く」という日本語の動詞は，その用法の違いから二つの格フレームを持つが，どちらの格フレームも格助詞「に」でマークされる格を持つ．

黒板に字を書く

/write a character on the blackboard

友に手紙を書く/write a letter to a friend

この二つの「に」格は意味的に異なった用法で用いられており，「に」格の意味の違いが動詞「書く」の用法の違いと関係していると考えられる．ここで，英文の方を見ると，この例の場合は，この二つの「に」格が“*on*”および“*to*”という異なった前置詞に訳されており，日本語の動詞「書く」の用法の違いが英語における格ラベルの違いに反映されている．したがって，このような英語の格ラベルの違いが日本語の動詞の複数の意味を類別する手がかりとなる．

格要素の名詞の意味ラベルの違い

場合によっては，英語の情報を用いても，日本語の動詞の複数の意味・用法を類別する手がかりが得られないこともある．表層形は同じだが意味の異なる日本語の動詞が，英語でも同じ動詞に訳されることはあり得るし，日本語側で格ラベルは同じだが意味の異なる格が，英語でも全く同じ格ラベルを持つこともあり得る．例えば，日本語の「上げる」という動詞は，次のような二つの意味を持つが，どちらの意味に対しても英語の訳として“*raise*”という動詞が可能であり，また「上げる」の「を」格は“*raise*”の目的格に相当する．

帽子を上げる/raise one's hat

叫び声を上げる/raise a cry

したがって，英語の情報からは，この二つの意味を類別することはできない．しかし，「帽子」および「叫び声」という二つの格要素の名詞は，名詞シソーラス中では全く別の意味カテゴリに属するため，これらの名詞が持つ意味ラベルは集合として互いに素 (disjoint) となる．このように，英語の情報が役に立たない場合でも，格要素の名詞の意味ラベルを調べることによって，動詞の意味・用法の違いを検出できる場合がある．このような意味ラベルの違いは，名詞シソーラス中で適当なレベルを設定して，そのレベルでの意味カテゴリの違いとして検出するのが適切であると考えられる．ここで，名詞シソーラス中のどのレベルを用いるかについては，利用している名詞シソーラスを十分に吟味して決定する必要があるが，我々は初期設定として「分類語彙表」中の根節点より一つ下の五つの意味カテゴリを用いることにした．

3.2 類別の方法

ある日本語の動詞 V_J の格フレームを獲得する場合には，まず，素性 '*pred*' の値に V_J を含む対訳素性構造を集めて集合 S を作る．ここでの目的は，動詞 V_J の複数の意味・用法を類別し，それぞれの意味・用法に応じて対訳素性構造の集合 S を部分集合 $S_1, \cdots, S_l$ に分割することである．その際に必要な処理は，二つの対訳素性構造が用例として与えられた時に，そこに含まれる動詞の意味・用法が異なっているかどうかを判定することである．しかし，S のすべての要素の組合せについて，意味・用法の違いを調べるのは現実的ではない．そこで，ここでは，格に注目し，

二つの用例中の動詞の意味・用法が異なっている場合には，お互いに両立し得ないような二つの格があって，その二つの格の両立不可能性によって動詞の意味用法の違いが引き起こされる

と考える．そして，解くべき問題を「お互いに両立し得ないような二つの格の組合せを発見する」問題に置き換える．格の種類の数は高々知れているので，これによって問題が解きやすくなり，人間の介入も必要最小限に抑えられる．では，どのような格の組合せが両立不可能になるかであるが，ここでは，「一つの格助詞でマークされる表層格がいろいろな意味で使われる」ことに注目し，同じ格助詞でマークされていてしかも両立不可能であるような格の組合せを求めることにする．この際には，前節で述べたような英語の情報および格要素の名詞の意味ラベルを手がかりとして利用する．

そのためにまず，次のような**格記述** (Case Description) を素性構造の形で定義する．

$$\begin{bmatrix} pred : \langle V_E, V_J \rangle \\ \langle l_E, p_J \rangle : [sem : SEM] \end{bmatrix}$$

この格記述は，対訳素性構造から得られる格を記述するためのもので，英語の動詞 V_E および格ラベル l_E および日本語の格要素の名詞の意味ラベル SEM を含んだ形で表現される．一つの対訳素性構造中にはいくつかの素性ペア $\langle l_E, p_J \rangle$ が含まれるが，各素性ペアからはそれぞれ別の格記述が得られる．また，意味ラベル SEM は，格要素の名詞の意味ラベルの違いによって動詞の意味・用法の違いを検出するためのものなので，格要素の名詞の意味ラベルを名詞シソーラス

中の決められたレベルまで一般化したものを用いる．我々は現在，このレベルとして，「分類語彙表」中の根節点より一つ下の五つの意味カテゴリを用いている．したがって，この場合 SEM は，{抽象的関係}，{人間活動の主体}，{人間活動—精神および行為}，{生産物および用具}，{自然物および自然現象} の五つの集合のうちのどれかである．また，格要素の名詞の意味ラベルをこの5カテゴリまで一般化した結果，複数のカテゴリに渡る場合には，その格要素はそのいずれのカテゴリにも属するとする．たとえば，一般化の結果，意味ラベルが {人間活動の主体，人間活動—精神および行為} となる格要素は，SEM が {人間活動の主体} となる格記述と {人間活動—精神および行為} となる格記述の両方に属する．

このような格記述の間の両立不可能性を調べるために，前節で述べた手がかりをもとにした**ヒューリスティックス**によって両立不可能な格記述の組合せの候補を求め，その後，**人間との相互作用**によってその候補が正しいかどうかを決定する．両立不可能な格記述の組合せが求まれば，それをもとに対訳素性構造の集合 S を部分集合 $S_1, \cdots, S_t$ に分割する．以下で，これらの処理について説明する．

ヒューリスティックス

ここでは，格記述のうち，日本語の格ラベル p_J が同じものの組合せについて，前節で述べたように以下の(a)〜(c)の手がかりを適用し，両立不可能な格記述の組合せの候補を求める．

(a)二つの格記述 CD_1 および CD_2 が，異なる英語の動詞をもつ．

$$CD_1=\begin{bmatrix} pred:\langle V_{E1}, V_J\rangle \\ \langle l_{E1}, p_J\rangle:[sem: SEM_1]\end{bmatrix}$$

$$CD_2=\begin{bmatrix} pred:\langle V_{E2}, V_J\rangle \\ \langle l_{E2}, p_J\rangle:[sem: SEM_2]\end{bmatrix}$$

(b)CD_1 および CD_2 は同じ英語の動詞を持つが，日本語の格ラベル p_J に対して異なる英語の格ラベルを持つ．

$$CD_1=\begin{bmatrix} pred:\langle V_E, V_J\rangle \\ \langle l_{E1}, p_J\rangle:[sem: SEM_1]\end{bmatrix}$$

$$CD_2=\begin{bmatrix} pred:\langle V_E, V_J\rangle \\ \langle l_{E2}, p_J\rangle:[sem: SEM_2]\end{bmatrix}$$

(c)CD_1 および CD_2 は同じ英語の動詞および格ラベルを持つが，格要素の名詞の意味ラベルが異なる．

$$CD_1=\begin{bmatrix} prod:\langle V_E, V_J\rangle \\ \langle l_E, p_J\rangle:[sem: SEM_1]\end{bmatrix}$$

$$CD_2=\begin{bmatrix} pred:\langle V_E, V_J\rangle \\ \langle l_E, p_J\rangle:[sem: SEM_2]\end{bmatrix}$$

人間との相互作用

ここでは，システムがヒューリスティックスによって選んだ候補が，両立不可能な格記述の組合せとなっているかどうかを人間に判定させるために，次の三つの選択肢を用意する．その際の質問は，二つの格記述およびその用例となるいくつかの対訳例（対訳素性構造）を人間に示すことによってなされる．

Ans 1: Clue

CD_1 と CD_2 は両立不可能であり，V_J の一つの格フレームの中に同時に現れることはない．

Ans 2: Co-occur

CD_1 と CD_2 は異なる意味の格であるが，お互いに両立可能であり，V_J の一つの格フレームの中に共起可能である．

例えば，日本語の「書く」という動詞は助詞「で」でマークされる格を二つ持つ．この二つは様態を表す格と道具を表す格で，英語ではそれぞれ "*in*" および "*with*" という前置詞に相当する．

英語で書く/write in English

筆で書く/write with brush

この二つの格は，一つの文の中に共起可能であり，「書く」の意味・用法の違いとは関係ない．

Ans 3: Equivalent

CD_1 と CD_2 は意味的に見て同一の格であり，一つの格にまとめるのが正しい．

例えば，日本語の「見る」という動詞の次の二つの用例は，英語の動詞の違いのためにヒューリスティックスによって検出されるが，この二つは意味的に同一の用法と考えてもよいと思われる．

映画を見る/see a movie

試合を見る/watch a game

これらの質問に対する判定の結果が **Equivalent** になる可能性は，(c)，(b)，(a)の順に低くなると考えられる．最初に **Equivalent** になる格記述をまとめてしまうと，全体として質問の数が少なくなるので，これらの質問は，(c)，(b)，(a)の順になされる．

また，ヒューリスティックス(a)および(b)から分かるように，我々の手法では，英語の動詞や格ラベルが異なる格記述は，両立不可能な組合せの候補として必ず検出される．つまり，格要素の名詞の意味ラベルが同じ場合でも，英語の情報が優先される．ただし，

日本語の動詞の一つの意味に対して英語の動詞の訳語が幾つも存在する場合には，ヒューリスティックス(a)による検出結果が無駄になる割合が多くなる．

対訳素性構造の集合を部分集合に分割

ここではまず，人間によって **Equivalent** の関係にあると判定された格記述を一つにまとめる．この結果，すべての格記述の集合が **Equivalent** という関係によって同値類 $ECD_1, \cdots, ECD_p$ に分割される．実際には，この処理は，質問・応答を行っている最中に動的に行われる．次に，格記述の間で得られていた両立不可能関係を同値類の間の両立不可能関係に拡張する．これは，「ペア $\langle CD_1, CD_2 \rangle$ が両立不可能な関係にあり，ECD_1 が CD_1 を含み ECD_2 が CD_2 を含むならば，同値類のペア $\langle ECD_1, ECD_2 \rangle$ も両立不可能な関係にある」という形で行う．

この結果得られる両立不可能関係 $\langle ECD_1, ECD_2 \rangle$ は，二つの対訳素性構造（すなわち，ECD_1 に含まれる格記述を持つ対訳素性構造と ECD_2 に含まれる格記述を持つ対訳素性構造）の間の両立不可能性を表しているといえる．したがって，両立不可能な対訳素性構造を同じ集合に含まないという制約のもとで対訳素性構造の集合 S を部分集合 $S_1, \cdots, S_l$ に分割することができる☆．

4. 用例からの格フレームの構成

対訳素性構造の集合 S が，動詞の意味・用法の違いに応じて部分集合 $S_1, \cdots, S_l$ に分割されれば，次には各部分集合 S_i から表層格フレーム Fr_i を構成する．一般に日本語の動詞 V_J の表層格フレームは，次のような素性構造で表現できると考えられる．

$$\begin{bmatrix} pred : V_J \\ p_{J1} : [sem : SR_1] \\ \quad \vdots \qquad \vdots \\ p_{Jm} : [sem : SR_m] \end{bmatrix}$$

ただし，$SR_1, \cdots, SR_m$ は格要素の名詞の意味的制約を表す．各 SR_j は，格要素として許容される意味カテゴリを記述したもので，名詞シソーラスの意味カテゴリの集合によって表現される．ここでは，これまでの処理の都合上，表層格フレーム Fr_i を，格記述の同値類 ECD と格要素の名詞に対する意味的制約 SR のペア $\langle ECD, SR \rangle$ の集合によって表現する．

$$Fr_i = \{\langle ECD_1, SR_1 \rangle, \cdots, \langle ECD_n, SR_n \rangle\}$$

ただし，$ECD_1, \cdots, ECD_n$ は，部分集合 S_i 中のすべての対訳素性構造から得られる ECD である．また，ECD_j に対する意味的制約 SR_j は，S_i 中で ECD_j の格要素となる名詞をすべて集め☆，それらの名詞の意味ラベル $SEM_1, \cdots, SEM_N$ から以下の要領で計算する．

一般に，名詞は多義性を持つため各意味ラベル SEM_r $(1 \leq r \leq N)$ は名詞シソーラス中の意味カテゴリの集合で表現される．意味的制約 SR_j を計算する際には，各意味ラベルからそれぞれ意味カテゴリを一つだけ選んで N 個の意味カテゴリ $Sem_1, \cdots, Sem_N$ を求め，これらをすべて含むように意味的制約を求める．このときには，意味的制約 SR_j ができるだけ狭い範囲に抑えられるようにそれぞれの名詞の意味カテゴリを一つ選ぶ．ただし，この解を一般的に求めるのは容易ではないので，ここでは，「別の一つの名詞の意味カテゴリから最も近くなるように，その名詞の意味カテゴリを一つ選ぶ」というヒューリスティックスを用いている．また，意味的制約 SR_j の計算の際には，**図 3** に示すように名詞シソーラス中に上限を設け，その上限よりも下位の意味カテゴリによって意味的制約を記述する．したがって，意味的制約 SR_j は名詞シソーラス中の意味カテゴリの集合によって表現される．「分類語彙表」の場合は，現在のところ，上限として根節点の一つ下の 5 カテゴリを設けている．

5. 実験および評価

本章では，日本語の 16 個の動詞について対訳素性

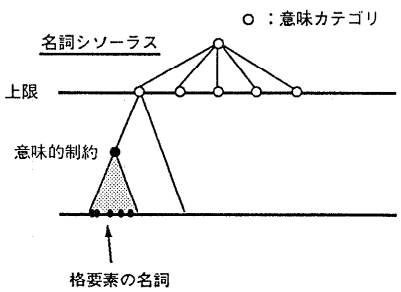

図 3 格要素の名詞に対する意味的制約の計算
Fig. 3 Calculating the semantic restriction of a case slot.

☆ この分割は，必ずしも一意に求まるとは限らない．複数の部分集合に含まれうる対訳素性構造は曖昧であるとみなして，現段階ではこれを無視している．

☆ ここで，ある日本語の名詞 N_J が ECD_j の格要素であるとは，ECD_j に現れる日本語の格ラベルを p_j として（これは一つの ECD_j に対してただ一つしかない），N_J が格ラベル p_j でマークされるということである．

構造を集め，表層格フレームを獲得する実験を行った結果について述べる．

5.1 表層格フレーム獲得の例

例として，「買う」という動詞の表層格フレームを48個の対訳素性構造から獲得する過程を説明する．ただし，説明の都合上，以下では簡略化した例を用いる．はじめに，「買う」という動詞について対訳素性構造を集めた結果から，格記述が**表1**のように得られる．ただし，この際には，一定回数以上現れた格記述のみを対象とする．この格記述をもとに，システムはヒューリスティックスによって「買う」の複数の意味を類別するための手がかりを検出し，その手がかりが正しいかどうかを人間に質問する．質問・応答の結果は**表2**のようになる．「買う」の場合，英語側で対応する三つの動詞 "*buy*", "*incur*" および "*appreciate*" の違いが「買う」の意味の違いになっている．この際の質問・応答の例は次のようになる．

質問（システム）：次の二つの格の間の関係は？

1: *buy-⟨obj, を⟩*

例文：私が家を買う/I buy a house.

彼が土地を買う/He buys a land.

2: *incur-⟨obj, を⟩*

例文：私が反感を買う/I incur antipathy.

彼が恨みを買う/He incurs enmity.

選択肢：（**a**）Clue（**b**）Co-occur（**c**）Equivalent

応答（人間）：⇒（**a**）Clue

表2の質問・応答の結果，「買う」の三つの格フレームが**表3**のように得られる（ここでは，簡単化のため格要素の名詞の意味的制約は「分類語彙表」の名詞シソーラスの根節点の一つ下の5カテゴリで記述してある．）．比較のために，情報処理振興事業協会（IPA）が公開している「計算機用日本語基本動詞辞書 IPAL (Basic Verbs)」[4] 中の「買う」の格フレームを**表4**に示す．「動詞辞書 IPAL (Basic Verbs)」には動詞861語の格フレームが含まれており，格要素の意味的制約を記述するために約20個の意味マーカが用いられている．IPALでは，「買う」は四つの格フレームを持っている．このうち，格フレーム1および2は英語の動詞 "*buy*" の意味に相当し，その違いは「に」格があるかないかだけである．我々の対訳コーパスには，「に」格を伴う「買う」の用例が出現しなかったため，「に」格を持つ格フレームは獲得されなかった．獲得された格フレームを見ると，IPALの格フレームと比べていくつかの格が欠けてはいるものの，「買う」の意味の類別についてはIPALとほぼ同等の結果が得られているといえる．

表1 「買う」の格記述
Table 1 Collected *Case Descriptions* of "買う".

	格ラベル	意味ラベル	頻度	例
1	*buy-⟨subj, が⟩*	人間	6	私
2	*buy-⟨obj, を⟩*	生産物	13	家, 車
		自然物	3	土地
3	*incur-⟨subj, が⟩*	人間	2	彼
4	*incur-⟨obj, を⟩*	人間活動	3	反感
5	*appreciate-⟨obj, を⟩*	人間活動	3	努力

表2 「買う」の意味の類別の手がかりに関する質問・応答
Table 2 Questions and answers about the clue to sense discrimination of "買う".

格記述の組合せ		判定
buy-⟨obj, を⟩-生産物	*buy-⟨obj, を⟩*-自然物	Equivalent
buy-⟨obj, を⟩-生産物・自然物	*incur-⟨obj, を⟩*-人間活動	Clue
buy-⟨obj, を⟩-生産物・自然物	*appreciate-⟨obj, を⟩*-人間活動	Clue
incur-⟨obj, を⟩-人間活動	*appreciate-⟨obj, を⟩*-人間活動	Clue

表3 獲得された「買う」の格フレーム
Table 3 Acquired case frames of "買う".

	格ラベル	意味的制約	格要素の例
1	*buy-⟨subj, が⟩*	人間	私
	buy-⟨obj, を⟩	生産物, 自然物	家, 土地
2	*incur-⟨subj, が⟩*	人間	彼
	incur-⟨obj, を⟩	人間活動	反感
3	*appreciate-⟨obj, を⟩*	人間活動	努力

5.2 獲得された格フレームの数—IPALとの比較

次に，16個の動詞について，獲得された格フレームの数をIPAL中に含まれる格フレームの数と比較した結果を**表5**に示す☆．表中で，Aは獲得された格フレームの数を，Iは IPAL 中に含まれる格フレームの数をそれぞれ表す．また，I∧¬Aは，IPAL中には含まれるが獲得されなかった格フレームの数を，A∧¬Iは，獲得されたがIPAL中には含まれない格フレームの数を表す．さらに，$X>Y$ は X の複数の格フレームが Y の一つの格フレームに対応することを

☆ このうち，平仮名表記してある動詞については，原則として，可能なすべての漢字表記の動詞が対応する．コーパス中においては，同じ意味の動詞が平仮名表記と漢字表記の両方の表記で現れ得る．またIPAL中の格フレームについても，対応するすべての動詞の格フレームを併せたものと比較する．

表し，具体的には n-1 によって，X の n 個の格フレームが Y の一つの格フレームに対応することを示す．また，「用例の数」は格フレーム獲得に用いられた対訳素性構造の数を，「質問回数/A」は質問・応答の回数を獲得された格フレームの数で割って正規化した数を表す．ここで，「質問回数/A」という数は，格フレームを一つ獲得するために，システムがヒューリスティックスによって格フレーム類別の手がかりの候補をいくつ検出したかを表す．

いくつかの動詞については，IPAL にはあるが我々の実験では獲得されなかった格フレームがある．これは，我々の対訳コーパスが十分な数の用例を含んでいないことが主な原因である．獲得された格フレームの数と IPAL の格フレームの数の差が最も大きかった例は，「かける」という動詞の場合で，IPAL の格フレームのうちの 28 個が獲得されなかった．また，獲得されたが IPAL 中には含まれない格フレームもいくつかあった．IPAL では，動詞の慣用的な用法が十分に分類されておらず，通常の用法の格フレームの例として慣用的な用法が記述されているが，我々の実験では，慣用的な用法は別の格フレームとして獲得された．実際，IPAL に含まれていなかった 12 個の格フレームの内，六つは慣用的な用法であった．

実験で獲得された複数の格フレームが IPAL の一つの格フレームに対応しているような例もあった．例えば，IPAL では「つける」という動詞の用法のうち，「刀に刃をつける」という用法と「理由をつける」という用法が「ある物に何かを加えたり装着したりする」という意味の同じ格フレームとして扱われているが，我々の実験ではこの二つの用法は分離された．ただし，このような二つの用法を分離するかどうかは，いずれも用法の分離の基準をどこに置くかに依存しており，極端な場合，IPAL とできるだけ同じ基準で用法を分離するように意識して実験を行うことも可能である．

一方，実験で獲得された一つの格フレームが，IPAL の複数の格フレームに対応するような例もあった．これは，システムのヒューリスティックスでは，動詞の複数の意味を類別する手がかりが検出できなかったことが原因である．例えば，「知る」という動詞の用法として，「噂を知る (“*know the rumor*”)」および「人生の苦労を知る (“*know the bitterness of life*”)」という二つの異なる用法がある．IPAL によれば，「知る」は，前者においては「何かに関して情報を得たり知識を持ったりする」という意味であり，後者においては「物事の本質を十分に理解する」という意味である．しかし，「分類語彙表」中では，「噂」も「苦労」も「人間活動」の意味カテゴリに属しており，現在のヒューリスティックスではこの二つの用法を類別する手がかりを検出するのが困難であった．これは，格要素の名詞の意味ラベルの違いを名詞シソーラスの根節点の一つ下の 5 カテゴリ間で検出しているため

表 4 IPAL の「買う」の格フレーム
Table 4 Case frames of “買う” in IPAL.

	格	意味マーカ
1, 2 (*buy*)	が	人間
	を	具体名詞, 抽象名詞
	(に	人間) 2 のみ
3 (*incur*)	が	人間
	を	精神
4 (*appreciate*)	が	人間
	を	性質, 精神, 動作

表 5 獲得された格フレームの数—IPAL との比較
Table 5 Comparison of number of case frames: acquired and IPAL.

A：獲得された格フレームの数　I：IPAL 中の格フレームの数
$X>Y$: X の複数の格フレームが Y の一つの格フレームに対応
(n-1: X の n 個の格フレームが Y の一つの格フレームに対応)

動詞	A	I	I∧¬A	A∧¬I	A>I	A<I	用例の数	質問回数/A
受ける	5	9	6	0	3-1	0	88	5
うつ	3	23	20	0	0	0	37	4.3
買う	3	4	0	0	0	1-2	48	10.7
かかる	3	29	25	0	0	1-2	35	2
書く	3	2	0	0	2-1	0	115	3
かける	7	35	28	1	0	1-2	59	2.6
聞く	3	6	2	0	2-1	1-2	85	4.7
来る	3	12	8	0	0	1-2	52	4
知る	1	5	2	0	0	1-3	69	6
つく	12	31	23	4	0	0	83	7.4
つける	10	14	6	1	2-1	0	108	7.3
出る	10	32	23	1	0	0	89	6
とる	10	29	18	2	0	1-6, $(1\text{-}2)^{\times 3}$	118	7
ひく	7	14	8	2	0	1-2	45	3.1
見る	1	13	10	0	0	1-3	106	22
持つ	5	12	4	1	0	$(1\text{-}3)^{\times 2}$, 1-2	69	8.4

で，ここでより細かい検出法を用いればいくらでも細かい意味・用法の類別が可能となる．しかし，格記述の分類を細かくすればするほど，格記述の数が多くなるので，いくらでも細かくすればよいというものではない．意味ラベルの違いの検出を名詞シソーラス中のどの意味カテゴリで行うかについては，今後，実験の結果から最適な意味カテゴリを求める必要がある．また，「とる」「持つ」といった動詞については，実験で獲得された一つの格フレームが IPAL の複数の格フレームに対応するという傾向が強く出ている．これは，「とる」の場合は "*take*"，「持つ」の場合は "*have*" というように，日本語の動詞の複数の意味に対して，英語の一つの動詞が対応し得ることが原因であり，これらの動詞については，格要素の名詞の意味カテゴリ中心の手法が必要であると言える．

5.3 二言語間素性構造照合の精度の影響

この実験では，二言語間で素性構造を照合した結果，照合結果の動詞の表層格構造が一意に決定したものを集めて，動詞の表層格フレームの獲得を行った．しかし，実際には，動詞の表層格構造が一意に決定したものの中に，単言語で係受けが誤っているものや二言語間での照合が誤っているものが含まれてしまう場合がある．したがって，これらの誤った対訳素性構造が，表層格フレーム獲得の際にどのような影響を及ぼすかが問題である．基本的には，誤った対訳素性構造は，同一の誤りパターンのものが一定回数以上発生しない限り，表層格フレーム獲得の際にも無視されるため，実際にはそれほど大きな悪影響を及ぼすことはない．また，同一の誤りパターンのものが一定回数以上発生した場合でも，それが二言語間での照合の誤りだけであり，日本語での係受けが正しい場合には，日本語の動詞の表層格フレームとして誤った結果が獲得されることはない．したがって，誤った対訳素性構造が含まれることによって，統計的に有効な情報の数が減少することはあるが，誤った対訳素性構造によって誤った表層格フレームが得られてしまうことは現実的には少ないと考えられる．

実際に，「買う」の格フレーム獲得の際に用いた 48 個の対訳素性構造について調べた結果，誤りと思われる対訳素性構造は 7 個 (14.6%) 含まれており，そのうち，日本語の表層格構造として誤った結果になっているものは 2 個 (4.1%) であった．また，これらの誤った対訳素性構造のうちで，同一の誤りパターンが一定回数以上発生したものはなかったので，表層格フレーム獲得には悪影響を及ぼさなかった．さらに，16 個の動詞全てについて調べた結果，表層格フレーム獲得に悪影響を及ぼした対訳素性構造が含まれていた動詞は三つしかなく，誤った対訳素性構造の数はいずれも 2 個以下であった．

今回の実験で用いた和英辞典の対訳例の文は，単文であるかあるいは単文でない場合も日英の文構造が似ていることが多かったため，比較的精度良く表層格フレームの獲得が行えたが，一般には対訳例の日英の文構造があまり似ていないことも多い．今後，一般の対訳コーパスから，動詞の表層格フレーム獲得をより効率良く正確に行うためには，二言語間の素性構造照合の精度を上げる必要がある．そのための対策としては，1) 単言語解析の精度を上げて，曖昧性の数を減らす．2) 特に単文以外の場合でも二言語間の照合を精度良く行うために，対応する動詞句同士を照合するというように，部分フレーズ主導型戦略による照合[14]を行う．3) 照合の際に，訳語の情報だけでなく，シソーラス中の類語などを用いて二言語間の単語対応に関する情報を増やす．といったことが考えられる．

6. おわりに

本論文では，対訳コーパスから動詞の表層格フレームを獲得する手法について述べた．我々の手法では，日本語の動詞の複数の意味・用法を類別する際に，英語の動詞や格ラベルが有用な情報となる．我々の手法によって，動詞の表層格フレームを獲得する実験を行った結果，対訳コーパスから動詞の表層格フレームを獲得することが可能であることがわかった．

動詞の複数の意味・用法を類別する方法としては，3.2 節で述べたように，格の間の両立不可能性を利用して用例の集合を分割する方法を用いた．この方法は，用例中の個々の格だけに注目した方法で，一つの用例中の複数の格の共起関係が最も重要な要因となって動詞の意味・用法が類別される場合には，これをシステムが内部で検出することができない．今後，複数の格の共起関係を積極的に利用する方法を考える必要がある．ただし，5.1 節の例からもわかるように，質問・応答の際には格記述と一緒にいくつかの用例が人間に提示されるので，人間は用例中の複数の格の共起関係を見て判断を下すことができる．

また，対訳コーパスに比喩的な文や誤った文などが含まれている場合には，適切な格フレームが獲得されない可能性がある．今後，ノイズを含んだ用例から最適

な格フレームを獲得するような手法を実現する必要があると考えられる.ただし,現在の我々の手法によれば,動詞の比喩的な用法を,通常の用法とは別の格フレームとして獲得することが可能であると考えられる.

現在のところ,シソーラスとしては,日本語の名詞シソーラスのみを用いているが,今後英語のシソーラスが利用可能となり,さらに,日英の間でシソーラスの対応がとれれば,二言語のシソーラスから得られる情報によって,日本語の名詞の多義性の解消がより精度良く行えるようになると思われる.

また,実験の結果から,今後十分な数の動詞について格フレームを獲得するためには,より大規模な対訳コーパスを構築する必要があることがわかった.我々は,これまでに,新聞の社説を英訳した対訳例を約21,000例(動詞の総数約70,000個)集めており,現在,これらの対訳例の文を解析して対応する動詞句を取り出しそこから動詞の格フレームを獲得するための手法の研究を行っている.この結果については今後報告する予定である.

謝辞 和英辞典のデータの使用を許可して下さった講談社編集部ならびに,和英辞典のデータを提供して下さった電子技術総合研究所の横山晶一氏および東京工業大学の田中穂積教授・徳永健伸氏に感謝いたします.

なお,本研究は,一部文部省重点領域研究「知識科学」の援助を受けている.

参考文献

1) Brent, M.: Automatic Acquisition of Subcategorization Frames from Untagged Text, *Proc. of the 29th Annual Meeting of the ACL*, pp. 209-214 (1991).
2) Dagan, I., Itai, A. and Schwall, U.: Two Languages are More Informative Than One, *Proc. of the 29th Annual Meeting of the ACL*, pp. 130-137 (1991).
3) Hindle, D.: Noun Classification from Predicate Argument Structures, *Proc. of the 28th Annual Meeting of the ACL*, pp. 268-275 (1990).
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 (Basic Verbs) 説明書 (1987).
5) 日本電子化辞書研究所:概念辞書(第2版), TR-012 (1989).
6) Lenat, D. et al.: *Building Large Knowledge-based Systems*, Addison-Wesley (1990).
7) 国立国語研究所:分類語彙表,秀英出版 (1964).
8) 清水 護,成田成寿(編):和英辞典,講談社学術文庫 (1979).
9) 徳永健伸,田中穂積:対訳辞書からの概念項目の自動抽出,人工知能学会誌, Vol. 6, No. 2, pp. 228-235 (1991).
10) 冨浦洋一,日高 達,吉田 将:語義文からの動詞間の上位-下位関係の抽出,情報処理学会論文誌, Vol. 32, No. 1, pp. 42-49 (1991).
11) 鶴丸弘昭,竹下克典,伊丹克企,柳川俊英,吉田将:国語辞典を用いたシソーラスの作成について,情報処理学会自然言語処理研究会, 83-16 (1991).
12) 宇津呂武仁,松本裕治,長尾 眞:二言語対訳コーパスからの語彙的知識獲得,「自然言語処理の新しい応用」シンポジウム論文集, pp. 104-114 (1992).
13) Utsuro, T., Matsumoto, Y. and Nagao, M.: Lexical Knowledge Acquisition from Bilingual Corpora, *Proc. of the 14th International Conference on Computational Linguistics*, Nantes, France, pp. 581-587 (1992).
14) 宇津呂武仁,松本裕治,長尾 眞:日英対訳文間の素性構造照合による統語的曖昧性の解消,情報処理学会論文誌, Vol. 33, No. 12, pp. 1555-1564 (1992).

(平成5年1月12日受付)
(平成5年3月11日採録)

宇津呂武仁(正会員)

1966年生. 1989年京都大学工学部電気工学第二学科卒業. 現在,同大学院博士後期課程在学中. 自然言語処理の研究に従事. 人工知能学会,日本ソフトウェア科学会, ACL 各会員.

松本 裕治(正会員)

1955年生. 1977年京都大学工学部情報工学科卒業. 1979年同大学院工学研究科修士課程情報工学専攻修了. 同年電子技術総合研究所入所. 英国インペリアルカレッジ客員研究員,(財)新世代コンピュータ技術開発機構研究員,京都大学工学部助教授を経て,1993年より奈良先端科学技術大学院大学教授. 自然言語処理,論理プログラミング等に興味を持つ.

長尾　眞（正会員）

1960年京都大学工学部電子工学科卒業．1962年同大学院修士課程修了．京都大学工学部助手，助教授を経て，1973年より同教授，現在に至る．1976年より国立民族学博物館併任教授．パターン認識，画像処理，自然言語処理，機械翻訳等の研究に従事．