自然言語処理 92−11
(1992. 11. 20)

ハイパー・シソーラスとその学習

李 航

NEC C&C 情報研究所

〒216 川崎市宮前区宮崎 4-1-1

E-mail: lihang@ibl.cl.nec.co.jp

本稿では、意味知識の表現モデルであるハイパー・シソーラスを提案し、さらにその学習方式を提案する。従来の意味知識表現では、シソーラス、あるいは意味素性によって意味知識を表現する。しかし、それには意味知識を正確に表現できない問題点がある。本研究で提案するハイパー・シソーラスは、格フレームの観点から意味知識を表現し、その問題点を解決する。ハイパー・シソーラスを利用した意味解析と訳語選択は従来より質が向上する。

Hyper Thesaurus: A Representation Model for Semantic Knowledge

Hang LI

C&C Information Technology Research Laboratories, NEC

Miyazaki 4-1-1, Miyamae-ku, Kawasaki 216, Japan

E-mail: lihang@ibl.cl.nec.co.jp

This paper proposes a new model named hyper thesaurus for semantic knowledge representation. Classically, semantic knowledge is represented by using a thesaurus or semantic features, but these representations can not adequately describe semantic knowledge necessary for satisfatory natural language processing. In order to resolve this problem, a hyper thesaurus, which represents semantic knowledge using case frames and concept hierarchies, is proposed. A method for learning a hyper thesaurus is also shown.

1 はじめに

意味知識をどのように表現、獲得、利用するかが現在の自然言語処理における未解決の問題の一つである。本稿では、意味知識の表現モデルであるハイパー・シソーラスを提案し、さらにその学習方式を提案する。

「意味知識」という用語を正確に定義することが難しいが、本研究では、意味知識が「鳥が飛ぶ」、「飛行機が飛ぶ」、「ボールが飛ぶ」等が意味的に正常で、「森が飛ぶ」、「川が飛ぶ」等が意味的に異常であることを判断する知識であるとする。つまり、意味知識は格フレームを規定する知識であるとする。意味知識がなければ、自然言語解析では文をほぼ曖昧性なく解析すること[1]が困難である。また、意味知識がなければ、機械翻訳では訳語をより適切に選択することも不可能である。

選択制限と呼ばれる従来の自然言語の意味解析手法では、シソーラスあるいは意味素性によって意味知識を表現している。例えば、英語の動詞 fly の主語をシソーラス上の概念 bird、aircraft 等によって規定する。それらの概念の下位概念がすべて fly の主語になり得るとする。選択制限は自然言語解析における曖昧性の解消にはかなり有効である。しかし、選択制限によっては満足のいくほど自然言語の解析を行うことができないこともよく知られている。選択制限で利用する意味知識の表現に問題があると思われる。

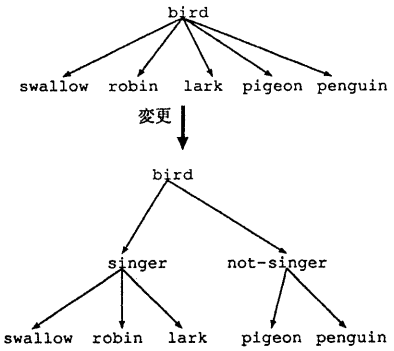

図 1: シソーラス

図1に示すシソーラスでは bird とその下位概念を示している。英語では鳥は種類によって動詞 sing の主語になれるものとなれないものがある。sing の主語になれる鳥とそれでない鳥を区別して記述するためには、シソーラスを図1のように変更する必要がある。しかし、そうすると、その他の動詞の格フレームの規定の記述に支障が出てくる。例えば、より正確に動詞 fly の主語を規定するためには、飛べる鳥と飛べない鳥があるので、異なる階層を定義する必要がある。単一の上位下位関係の構造をもつシソーラスですべての用法をうまく表現することが困難である [荻野 87][徳永 89]。できれば異なる視点によるシソーラスの構築が望ましい。本研究で提案するハイパー・シソーラスはまさにそのような「シソーラス」である。ハイパー・シソーラスは以下の特徴をもつ。

1. ハイパー・シソーラスは意味知識を表現するものである。
2. ハイパー・シソーラスはノードとリンクから構成される。ハイパー・シソーラスではノードが言語概念を表す[2]。名詞概念と動詞概念はそれぞれ階層を形成する。名詞概念と動詞概念の階層は概念の上位下位関係を表す。
3. 動詞概念が格フレームをもち、動詞概念の格フレームの格スロットから名詞概念へリンクが張られる。このようなリンクを格関係リンクという。格関係リンクには個体リンクとグループ・リンクの二種類のものがある。個体リンクには重みがつく。本研究ではこの重みをファジイとする。ある名詞概念と動詞概念のある格スロットが個体リンクでつながる時、その名詞概念が動詞概念のその格スロットに属することができることを意味する。その個体リンクにつくファジイは名詞概念が動詞概念のその格スロットの属する度合いを表す。ある名詞概念とその下位概念がすべて動詞概念のある格スロットに属し、しかもそれらの名詞概念が動詞概念のその格スロットに属する度合い (ファジイ) がすべて 1 である時、その名詞概念と動詞概念のその格スロットの間にグループ・リンクを張る。グループ・リンクは表現の効率を上げるためのものである。
4. ハイパー・シソーラスでは、言語概念はその用法のすべてによって規定されるとする。従って、動詞概念はその格フレームのすべてによって規定される。動詞概念の格フレームの格スロットはそれに属することができるすべての名詞概念、およびそれぞれの名詞概念のその格スロットに属するファジイによって規定される。名詞概念はその属することのできる格スロットのすべてによって規定される。
5. ハイパー・シソーラスは意味知識を動的に表現する。ハイパー・シソーラスでは意味知識は現在までに獲得した知識に基づいて作った「仮説」であるとする。ハイパー・シソーラスでは意味知識を格関係リンクの張り替えとファジイの変更によって修正することができる。

格フレームの観点から意味知識を表現することがハイパー・シソーラスの最大の特徴である。普通のシソーラスはその形が変われば、表現する意味知識の内容も変わる。しかし、ハイパー・シソーラスでは、(普通のシソーラスと対応する) 名詞概念と動詞概念の階層の形が変わっても、意味知識の表現効率が変わるだけで、意味知識の内容が変わらない。ハイパー・シソーラスはきわめて正確に人間のもつ意味知識を表現する[3]。

ハイパー・シソーラスを構築するのに膨大な量の意味知識の取扱いが必要である。機械が自動的、あるいは半自動的に意味知識を獲得することが望ましい。本稿ではハイパー・シソーラスにおける意味知識の学習方式を提案する。学習の際、テキスト・データから意味知識を獲得するだけでなく、テキスト・データから十分に獲得できない意味知識を人間に質問し、獲得する。

ハイパー・シソーラスを利用して、自然言語解析における意味解析、あるいは機械翻訳における訳語選択を行うことができる。ハイパー・シソーラスを利用したこれらの自然言語処理は従来より質が向上する。

[1] 曖昧性を完全に解消するためには文脈処理が必要である。

[2] 言語概念という用語の定義は後に述べる。

[3] 人間はほとんどの意味知識について共通の認識をもつ。ここでは意味知識についての人間の共通認識を意味知識の「真」の正解とする。

近年実例に基づく機械翻訳の研究が盛んである。ハイパー・シソーラスを利用した機械翻訳は、実例の一般化[4]を行っている点で実例に基づく機械翻訳の改善になる。

本稿の構成は以下の通りである。第2章でハイパー・シソーラスを提案する。第3章でハイパー・シソーラスの学習について述べる。第4章でハイパー・シソーラスの利用について述べる。第5章で決論について述べる。

2 ハイパー・シソーラス

2.1 従来の意味知識表現とその問題点

従来ではシソーラスあるいは意味素性によって意味知識を表現している。シソーラスによって意味知識をうまく表現できない例を第1章で挙げた。以下では意味素性によって意味知識をうまく表現できない例を示す。

意味素性による意味知識表現では、例えば、walksの主語が動物であることを意味素性 *+animal* によって表現し、stoneが動物でないことを意味素性 *−animal* によって表現する。「*A stone walks」という文を解析すると、stoneとwalkのもっている意味素性の間に矛盾があるので、この文が意味的に異常であることがわかる。しかし、例えば、英語の動詞takeに関して以下の表現ができる。

I often forget to take my umbrella. (1)

Shall I take your message to John? (2)

We usually take the children to school in the car. (3)

She has gone to the dentist to have a tooth taken out. (4)

表1: 名詞とその意味素性

umbrella	+arrangement,+keep_rain_off, +cloth_over_frame
message	+information, +passed
child	+human, +young
tooth	+bony_object,+in_the_mouth,+biting, +chewing

以上の例文におけるtakeの目的語になる名詞とそれらの名詞の意味素性を表1に示す[5]。表1からわかるように、それらの名詞には共通の意味素性がない。それらの名詞の共通の特徴を正確に記述するためには「動詞takeの目的語になれる」という意味素性 *can_be_taken* を定義する以外にはいい方法がない。そうすると、ほぼ動詞の数に比例する意味素性が必要になってくる。意味素性を用いるより直接動詞の格フレームと名詞の関係を記述したほうがよい。

シソーラス、あるいは意味素性による意味知識表現ではイエスかノーの記述をしかしないのも問題である。ある文の意味解釈がある度合いで妥当であることを表現したいことがある。例えば、

The pilot flied to Tokyo. (5)

英語の動詞flyには「飛行機を操縦する」意味と「飛行機に乗る」意味がある。従って、人間にとって上の文には二通りの解釈がある。「パイロットは飛行機を操縦し、東京にいった」という解釈と「パイロットは飛行機に乗って東京にいった」という解釈である。しかし、人間の場合、平均的な文脈においては前者の解釈が解釈される可能性が大きい。シソーラス、あるいは意味素性によって意味知識を表現する時に、その二つの解釈の一つを排除してしまうか、あるいは二つの解釈を同等に残してしまうかのいずれになる。度合いを含めた形で意味知識を表現する必要がある。

シソーラスが意味素性継承の階層であるとすれば、シソーラス、あるいは意味素性によって意味知識をうまく表現できないのは本質的に同じことである。

まとめて、従来の意味知識表現には以下の問題点がある。

1. シソーラスは異なる視点による意味知識の表現ができない。
2. 意味素性によって格スロットをうまく規定できない。
3. シソーラス、あるいは意味素性で度合いを表せない。

2.2 プロトタイプ理論

近年、言語学ではプロトタイプ理論 [Lakoff87][Taylor89] が注目を集めている。プロトタイプ理論は言語カテゴリを説明する理論である。言語カテゴリとは、ある条件を満たす言語的な概念の集合のことである。言語カテゴリは構文カテゴリと意味カテゴリからなる。

プロトタイプ理論の主な主張を以下にまとめる。

- プロトタイプ理論は哲学者ウィトゲンシュタイン (Wittgenstein) の影響を受けている。ウィトゲンシュタインは、「言葉の意味はその用法の集まりによって定義され、用法以外の方法で言葉の意味を定義することができない」と考えていた。プロトタイプ理論では言語カテゴリ、特に意味カテゴリは素性等によって規定できないと主張する。プロトタイプ理論では言語カテゴリを度合いのついたプロトタイプによって規定する。
- 言語知識 (Linguistic Knowledge) と世界知識 (Encyclopaedic Knowledge) は異なる知識である。言語知識は言語概念の関係を表す知識であり、世界知識は世界知識としての概念の関係を表す知識である。言語概念は言語表現によって表され、世界知識としての概念と関係をもつ。しかし、言語概念と世界知識としての概念は同一のものではない。例えば、言語概念のelephantは動詞「eat」の主語になれることや動詞「walk」の主語になれること等によって定義される。しかし、世界知識としての概念のelephantは「四本の足をもつ」ことや「長い鼻をもつ」こと等によって定義される。当然、世界知識としての概念のelephantが言語概念のelephantに影響を与えるが、その両者は同一のものではない。

[4]「実例の一般化」の定義は後に述べる。

[5]名詞の意味素性をLongman辞書の定義文に基づいて定義した。

- 単語の表す言語概念が互いに密接な関係をもち、全体が一つの家族を形成する。このことを家族的類似性 (Family Resemblance) という。

格フレームの格スロットも一種の意味カテゴリとみなすことができる。本研究ではプロトタイプ理論の考え方を格スロットに適用する。具体的には以下のことを主張する。

1. 格スロットを意味素性によって規定することができない。
2. 格スロットを言語概念と言語概念のその格スロットに属する度合いによって規定すべきである。

素直にこの主張に従えば、言語概念と言語概念の格スロットに属する度合いの表によって格スロットを規定し、意味知識とすべきである。しかし、そのような意味知識表現は表現の効率が悪いし、構築のコストも大きすぎる。如何に効率よく格スロットを言語概念と度合いによって規定するか、つまり、効率よく意味知識を表現するかが重要な課題になる。ハイパー・シソーラスはその課題を解決するためのものである。

2.3 ハイパー・シソーラスの提案

自然言語処理では、単語が複数の言語概念を表現すると考え、普通、単語のもつ複数の語義を単語の表現する複数の言語概念とする。しかし、プロトタイプ理論によれば、単語の表現する言語概念は一つの家族を形成し、原理的に分割できないものである。単語 (の表現する家族) を言語概念にしたほうがよいか、それとも単語の語義を言語概念にした方がよいかが今後の研究課題である。本研究では、単語 (の表現する家族) を言語概念とする。

本研究では言語概念の格スロットに属する度合いをファジイとする。これは格スロットを集合と見なすことができるからである。例えば、動詞 fly に関しては表 2に示す表現ができ、それぞれの表現に意味的な正しさを表わす評価値を与えることができる。これらの評価値を名詞概念の fly の主語スロットに属するファジイとする (図 2)。ファジイは主観確率の近似であると見なすことができる [仁木 91][李 92]。ファジイには取扱い易いというメリットがあるので、本研究では主観確率ではなく、ファジイを採用する。

表 2: 言語表現とその意味的な正しさの評価値

a bird flies	1
a swallow flies	1
a eagle flies	1
an aircraft flies	1
a helicopter flies	1
a ball flies	0.6

次に、ハイパー・シソーラスを提案する。ハイパー・シソーラスは意味知識を表現するものである。ハイパー・シソーラスはノードとそれをつなぐリンクからなる。ハイパー・シソーラスではノードが言語概念を表す。ハイパー・シソーラスの例を図 3に示す。

言語概念は名詞概念と動詞概念からなる。名詞概念と動詞概念はそれぞれ階層を形成する。階層が木構造であってもよいし、

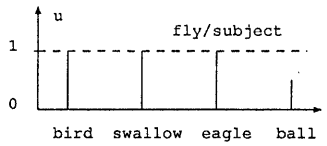

図 2: ファジイ

束 (Lattice) 構造であってもよい。動詞概念の階層と名詞概念の階層は世界知識を反映したものであるとする。これはプロトタイプ理論での言語知識が世界知識とつながりをもつという考えによる。名詞概念と動詞概念の階層が従来のシソーラスと対応するものであると考えることもできる。

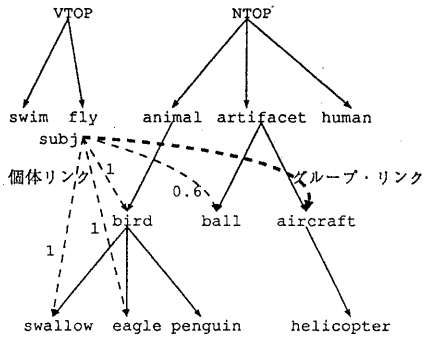

図 3: ハイパー・シソーラス

ハイパー・シソーラスでは動詞概念が格フレーム[6]をもつ。動詞概念の格スロットから名詞概念へリンクが張られる。このようなリンクを格関係リンクという。格関係リンクには個体リンクとグループ・リンクの二種類のものがある。個体リンクにはファジイがつく。ある名詞概念と動詞概念のある格スロットが個体リンクでつながる時、その名詞概念が動詞概念のその格スロットに属することができることを意味する。その個体リンクにつくファジイは名詞概念が動詞概念のその格スロットの属する度合いを表す。ある名詞概念とその下位概念がすべて動詞概念のある格スロットに属し、しかもそれらの名詞概念が動詞概念のその格スロットに属する度合い (ファジイ) がすべて 1 である時、その名詞概念と動詞概念のその格スロットの間にグループ・リンクを張る。グループ・リンクは表現の効率を上げるためのものである。ハイパー・シソーラスでは個体リンクによってあるカテゴリに属する「例外的な言語概念」を表現する。当然、ある名詞概念からある格スロットへの格関係リンクがない時、その名詞概念がその格スロットに属さないことを意味する。

動詞概念はそのもつ格フレームのすべてによって規定される。さらに、動詞概念の格フレームの格スロットはそれに属することができるすべての名詞概念、およびそれぞれの名詞概念のその格スロットに属するファジイによって規定される。名詞概念はその属することのできるすべての格スロットによって規定される。言語概念が数多くの概念とリンクされることがその言語概念を様々な視点からみることができることと対応する。

動詞概念の階層を定義する時に格フレームの継承を考慮する必要がある。Webster らは動詞の格フレームの格スロットが必須、禁止、自由であるかどうかによって動詞を分類し、動詞の

[6]格フレームは深層であってもよいし、表層であってもよい。

階層を定義した[Webster89]。Websterらによれば、格フレームによって行った動詞の分類は、ほとんど動詞の意味分類にもなる。Websterらと同じように動詞概念の階層を導入すればよいかも知れない。本研究では、さしあたり動詞概念がフラットの階層を形成するとする。動詞概念が必須格と自由格をもつとする。動詞概念の(仮想的な)トップ概念VTOPが自由格をもつとする。自由格は非単調的に継承される。つまり、子供ノードの格スロットの内容と親ノードの格スロットの内容の間に矛盾がある時に、子供ノードの格スロットの内容が優先される。動詞概念の階層をどのように定義すればよいかが今後の研究課題である。

ハイパー・シソーラスでは、ファジイによって概念の格スロットに属する度合い、ひいては言語表現の意味的な正しさを表現している。度合いは応用(例えば、例文5の解析)からの要請だけでなく、意味知識の記述からの要請でもある。自然言語の表現が特別な文脈においてはほとんど何でも「解釈可能」である。度合いは「平均的な文脈」において表現の正しさを表現している。意味知識を構築するのに度合いが欠かせないものである。

ハイパー・シソーラスは意味知識を動的に表現する。ハイパー・シソーラスでは、意味知識が静的なものであるとはせず、あくまでも現在までに獲得した知識に基づいて作った「仮説」であるとする。ハイパー・シソーラスでは学習によって意味知識を修正することができる。修正はリンクの張り替えとリンクのファジイの変更によって実現される。本研究では上位下位関係を表すリンクを静的なリンク(Static Link)、格関係を表すリンクを動的なリンク(Dynamic Link)ともいう。

格フレームの観点から意味知識を表現することがハイパー・シソーラスの最大の特徴である。普通のシソーラスはその形が変われば、表現する意味知識の内容も変わる。しかし、ハイパー・シソーラスでは、(普通のシソーラスと対応する)名詞概念と名詞概念の階層の形が変わっても、意味知識の表現効率が変わるだけで、意味知識の内容が変わらない。

ウィトゲンシュタインは次のことを言っている[岡田86]。

> 文法[7]は言葉の実際的用法の記述であるので、言語の営業簿と言える。従って言語の実際の業務処理に関することはすべてその営業簿からみてとることができる。

ハイパー・シソーラスはウィトゲンシュタインの営業簿であることがいえる。但し、ウィトゲンシュタインがいっている「言葉の用法」は、その言葉の使われる状況も含むので、ハイパー・シソーラスはまだ完全な営業簿ではない。

2.4 実現

ハイパー・シソーラスをPrologで実現した。実現のレベルでは、図3に示すハイパー・シソーラスにおける名詞概念の階層と動詞概念の階層を以下のように定義する。

```
% nnode(NodeName,Children)
nnode('NTOP',[human,animal,artifacet]).
nnode(human,[]).
nnode(animal,[bird]).
nnode(bird,[swallow,eagle,penguin]).
nnode(swallow,[]).
nnode(eagle,[]).
nnode(penguin,[]).
nnode(artifacet,[aircraft,ball]).
nnode(aircraft,[helicopter]).
nnode(helicopter,[]).
nnode(ball,[]).

% vnode(NodeName,Children)
vnode('VTOP',[fly,swim]).
vnode(fly,[]).
vnode(swim,[]).
```

さらに、動詞概念の格フレームを他の述語で定義する。論理的には動詞概念の格フレームは動詞概念のノードに含まれる。格フレームの変更は格フレームを表わす述語の変更によって実現される。

```
fly([[pred,fly],[subject,bird,1,i]]).
fly([[pred,fly],[subject,aircraft,1,g]]).
fly([[pred,fly],[subject,ball,0.6,i]]).
......
```

2.5 補足

格フレームの格スロットが共起関係にある時[8]、その格フレームを一まとまりにし、ハイパー・シソーラスで表現する必要がある。例えば、英語の動詞destroyの表現を表3に示す。

表3: 言語表現

The fire destroied the bridge.
The hurican destroied the village.
His speech destroied my hopes.
*The speech destroied the village.

destroyの主語と目的語の間に共起関係があることがわかる。よって、destroyの格フレームをまとめて表現する必要がある。

```
destroy([[pred,destroy],
         [subject,disaster,1,g],
         [object,area,1,g]]).
destroy([[pred,destroy],
         [subject,disaster,1,g],
         [object,building,1,g]]).
......
```

世界知識を利用しなければ、自然言語処理の質が向上できないといわれる。しかし、世界知識をどのようなモデルで表現すればよいかが大きな問題である。言語構造に反映された世界知識が自然言語処理にとって必要な知識である。ハイパー・シソーラスではそのような世界知識を言語構造によって表現している。例えば、

The sun rises in the east. (6)

[7]ウィトゲンシュタインのいう文法と普通我々のいう文法とは少し異なると思われる。

[8]厳密には、ほとんどの格フレームの格スロットが共起関係にある。共起関係がさほど強くない時、各格スロットが互いに独立していると仮定する。

```
rise([[pred,rise],[subject,sun,1,i],[in,east,1,i]).
```

「太陽が東から昇る」という世界知識を動詞概念 rise の格フレームによって表現する。慣用表現や定着した比喩表現も同じようにハイパー・シソーラスで表現することができるであろう。

2.6 拡張

今まで動詞概念の格フレームを考えてきた。しかし、複合名詞句等の場合、ヘッドとしての名詞概念が格フレームをもち、その格フレームの格スロットに修飾語としての名詞概念が入ることも考えられる。例えば、所有表現はそうである (表 4)。ハイパー・シソーラスでは名詞概念が格フレームをもつとすることもできる。そうすると、ハイパー・シソーラスの構造がもっと複雑になる。

表 4: 言語表現とその意味的な正しさの評価値

the teacher's house	1
the teacher's work	1
the teacher's arrival	1
the table's surface	0.8
the building's age	0.6
the sky's color	0.4
the doorway's height	0.4

ハイパー・シソーラスでは、名詞概念の動詞概念の格フレームの格スロットに属する頻度を表す確率を定義することもできる。確率は動詞概念と格スロットが決まった時に名詞概念のその格スロットに属する確率 $P(noun|verb, slot)$ であるとする。名詞概念と動詞概念の格スロットの間に張られる個体リンクに以上の確率をつけることができる。ファジイと確率の両方を用いて動詞と名詞間の格関係をもっと正確に表現することができる [李 92]。

2.7 メリットと問題点

ハイパー・シソーラスによる意味知識表現には以下のメリットがある。

1. ハイパー・シソーラスが意味知識を正確に表現できるので、ハイパー・シソーラスを利用した意味処理は信頼性が高い。
2. ハイパー・シソーラスを利用した意味処理はほとんど知識の検索だけですむので、処理のコストが高くない。
3. ハイパー・シソーラスでは動的なリンクの張り替えが自由にできるし、また、そのファジイの変更も自由にできるので、学習に適している。

ハイパー・シソーラスによる意味知識表現には以下の問題点がある。

1. 膨大な量の知識を取扱わなければならない。学習による意味知識の獲得の研究が必要である。
2. ファジイは主観的なものである。ファジイを定義するためのよりよい基準の研究が必要である。

3 ハイパー・シソーラスの学習

本章では、ハイパー・シソーラスの学習について述べる。

学習では、名詞概念と動詞概念の階層が与えられ、テキストからの実例の抽出と人間への質問により、格フレームの知識と格スロットの知識を獲得する。

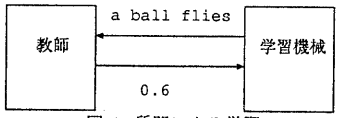

図 4: 質問による学習

まずテキストから実例をできる限り抽出し、実例から格フレームと格スロットの知識を獲得する。それから、人間への質問によって格スロットの知識をさらに獲得する。人間への質問によって、学習の質を高める。具体的には、テキストに出現していない表現、テキストから抽出する時に間違って抽出する可能性のある (と機械が思っている) 表現等が意味的に正しいかどうかを人間に質問する。人間がその表現の意味的な正しさを五段階で評価する (図 4)。また、個人の判断にはバラツキがあるので、複数の人間に質問し、その評価値の平均をとる方式も考えられる。質問の後、学習機械が質問の結果に基づいてハイパー・シソーラスを構築する (あるいは、構築し直す)。具体的には、名詞概念から動詞概念の格スロットへリンクを張ったり、リンクにファジイを設定したりする。結果的には意味知識が獲得される。

次に具体例で学習の過程を示す。例えば、図 3のハイパー・シソーラスの構築にあたって、「a swallow flies」をテキストから抽出できたとする。動詞概念 fly の格フレームをその例から獲得できることになる。さらに名詞概念 swallow が fly の主語になれることもわかる。次に学習機械が名詞概念の階層に基づいて、swallow の上位概念が fly の主語になれるかどうかを人間に質問する。「a bird flies」が意味的に正しいかどうかを質問し、1 が返ってきたとする。より正確に学習を行なうためには bird の下位概念の penguin と eagle が fly の主語になれるかどうかについても人間に質問する必要がある。学習の結果がほどほどに正しければよいという戦略をとるならば、その時、bird の下位概念がすべて fly の主語になれるとし、bird と fly の主語スロットをグループ・リンクで連結する。学習機械がさらに「an animal flies」が意味的に正しいかどうかを質問し、0 が返ってきたとする。そこで、「a swallow flies」に関わる意味知識の学習が終了する。

この学習方式では人間にできるだけ少ない数の質問をし、格フレームの格スロットを獲得することが重要である。今後、効率的な学習 (質問) アルゴリズムの開発を行なっていく予定である。

単語を言語概念とするハイパー・シソーラス、あるいは語義を言語概念とするハイパー・シソーラスのいずれも構築することが可能であると述べた。しかし、語義を言語概念とするハイパー・シソーラスの学習では、テキストから表層の単語をしか抽出できないので、人間が単語の語義分割をする必要がある。従って、語義を言語概念とするハイパー・シソーラスの学習はコストがかかる。

ウィトゲンシュタインは、言語と言語の織り込まれた活動の総体を言語ゲーム (Sprachspiel) と呼んでいる。ウィトゲンシュタインの考え方によれば、人間は一定の言語を話す人々との共

同体のもとに生まれ、そこで営まれる言語ゲームに参加する能力を身につけることを通して一人前の母国語話者となる。計算機にとっても、人間との相互作用によって言語知識を獲得していくことが重要であろう。

現在までの多くの自然言語処理のシステムの構築では、意味知識の記述には母国語話者を必ずしも必要としなかった。しかし、ハイパー・シソーラスの学習では、意味の正しさの判断には母国語話者が必要である。このことはハイパー・シソーラスでは「質」の高い意味知識を構築していることを裏付けている。

4 ハイパー・シソーラスの利用

本章ではハイパー・シソーラスを利用した自然言語処理について紹介する。

4.1 自然言語解析における優先度

[李 92] では自然言語解析における優先度を次のようにモデル化した。

$$\text{優先度} = [\text{意味ファジイ}, \text{意味確率}, \text{構文確率}] \qquad (7)$$

つまり、意味ファジイ、意味確率、構文確率のベクトルを優先度とする。特に、意味ファジイを利用することによって、信頼性の高い優先度を実現することができた。意味ファジイは、ハイパー・シソーラスにおけるファジイのことである。

4.2 機械翻訳における訳語選択

本研究では訳語ペアのハイパー・シソーラスを利用した機械翻訳を考案した。以下ではそれについて説明する。

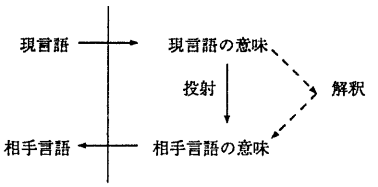

図 5: 翻訳のモデル

人間、あるいは機械の翻訳過程をおよそ図5に示すようにモデル化することができる。人間は翻訳を行う時に現言語を解析し、現言語の意味を抽出し、解釈を行う。さらに解釈の対応する相手言語の意味を選択し、相手言語を生成する。一方、解釈を含めた機械翻訳の実現が極めて困難なため、現在までの多くの機械翻訳システムは解釈を行わず、現言語意味から相手言語意味への投射によって翻訳を行っている。将来、解釈を含めた機械翻訳が実現できても、翻訳には現言語意味の保存が必要なので、現言語意味から相手言語意味への投射の利用はやはり欠かせない。言語間の意味の投射に関する研究が非常に重要なテーマである。本研究では、訳語ペアのハイパー・シソーラスによって、文単位の言語間の意味の投射を表現する。

訳語ペアのハイパー・シソーラスでは、名詞の訳語ペアと動詞の訳語ペアがそれぞれ現言語に従って階層を形成する。動詞の訳語ペアが格フレームをもつ。格フレームでは現言語と相手言語における格の対応も記述する。名詞の訳語ペアから動詞の訳語ペアの格スロットへ格関係リンクが張られる。訳語ペアのハイパー・シソーラスはやはり学習によって構築される。

以下では、例を通じて訳語ペアのハイパー・シソーラスの学習と利用について説明する[9]。「taro plays tenis」という実例が与えられたとする。この実例に基づいて、学習機械が人間に質問しながら、動詞の訳語ペア ⟨play, する⟩ の格フレームを獲得していく。訳語ペア ⟨play, する⟩ に関する意味投射の知識が獲得される。例えば、⟨play, する⟩ が以下のような格フレームをもつとする。

```
play-suru([[pred,[play,'する']],
          [[subject,'が'],[human,'人間'],1,g],
          [[object,'を'],[sports,'スポーツ'],1,g]]).
play-suru([[pred,[play,suru]],
          [[subject,'が'],[human,'人間'],1,g],
          [[object,'を'],[game,'ゲーム'],1,g]]).
```

その他に、⟨play, 演ずる⟩ と ⟨play, ひく⟩ の格フレームの学習もできたとする (図 6)。

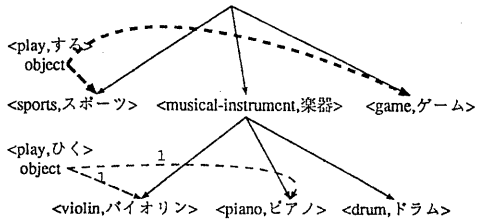

図 6: 訳語ペアのハイパー・シソーラス

いま「japanese play card」という文が入力されるとする。翻訳システムがそれぞれの単語による可能な訳語ペアをすべて見つけ出す。さらに訳語ペアの組合せによるすべての訳をつくり、それぞれの訳のファジイを計算する。論理積のファジイの定義に従い、訳を構成する格関係のファジイの最小値をその訳のファジイとする。

```
play-suru([[pred,[play,'する']],
          [[subject,'が'],[japanese,'日本人'],1],
          [[object,'を'],[card,'トランプ'],1]]).
```

という訳のファジイが 1 である。一方、

```
play-suru([[pred,[play,'する']],
          [[subject,'が'],[japanese,'日本人'],1],
          [[object,'を'],[card,'カード'],0.2]]).
```

という訳のファジイが 0.2 である。すべての訳の中からファジイのもっとも高い訳を選択する。つまり、整合性がもっともよい訳を選択する。ファジイのもっとも高い訳が複数ある時に動詞訳語ペアの出現頻度に従って、もっとも出現頻度の高い動詞訳語ペアによる訳を採用する。また、名詞句の訳語選択も似たような形で行える。

近年、実例に基づく機械翻訳 (Example Based Translation、以下 EBT) に関する研究が盛ん [佐藤 90][Sato90][隅田 91] であ

[9] この例は [佐藤 90] による。

る。EBTでは入力と実例のシソーラスにおける距離を計算し、もっとも距離の近い翻訳例を模倣し、訳を出力する。EBTでは抽象化された規則を設ける必要がなく、理想的には翻訳の例をデータベースに追加するだけで翻訳の向上を実現することができる。しかし、現在のEBTには以下の問題点がある。学習によって構築した訳語ペアのハイパー・シソーラスを利用した翻訳ではそれらの問題点を解決することができる。

1. 本研究では実例の適用範囲を明確にすることを一般化というが、現在のEBTでは実例の一般化を行っていない。現在のEBTでは、ある実例が与えられた時、その実例がそれ自身の翻訳にしか適用できない例なのか、それともある範囲にわたって適用できる例なのかを明確に記述する枠組みがない[渡辺92]。訳語ペアのハイパー・シソーラスでは、各翻訳パターンの適用範囲がはっきりと記述される。しかも、それぞれの適用範囲は離散的でなく、ファジイによって連続的に記述される。

2. 現在のEBTは計算コストが高い。入力と可能なすべての実例とベスト・マッチングを行うため、膨大な量の計算が必要である。一方、訳語ペアのハイパー・シソーラスを利用した翻訳では、処理がほとんどファジイ値の検索で済むので、計算量が少ない。

3. 現在のEBTは単一の上位下位関係構造のシソーラスを利用する。しかし、そのようなシソーラスの作成はできない。

訳語ペアのハイパー・シソーラスを利用した機械翻訳はEBTのメリットをそのまま保っている。つまり、複雑な規則を必要とせず、学習によって成績がますますよくなる。訳語ペアのハイパー・シソーラスを利用した機械翻訳は、EBTと較べると、学習のコストが高いことが欠点である。。

EBTは実例を用いて翻訳パターンを表す点で優れている。プロトタイプ理論の立場からみれば、実例は翻訳パターンの典型的な例にほかならない。しかし、実例を用いるだけでは不十分な場合が多く、その実例を一般化する必要がある。そういう意味では、訳語ペアのハイパー・シソーラスを利用した機械翻訳は現在のEBTの一つの改善である。

5 おわりに

本稿ではハイパー・シソーラス、ハイパー・シソーラスの学習、及びハイパー・シソーラスの利用について述べた。

本研究を幾つかの観点から評価することができる。

1. 意味知識の新しい表現モデルであるハイパー・シソーラスを提案した。

2. ハイパー・シソーラスにおける意味知識を実例と質問によって学習する方式を提案した。

3. ハイパー・シソーラスによる実例に基づく機械翻訳の改善策を提案した。

参考文献

[Angluin88] D. Angluin, *Queries and Concept Learning, Machine Leaning 2, 1988*

[Brent91] M. Brent, *Automatic Acquisition of Subcategorization Frames from Untagged Text, ACL91, 1991*

[Lakoff87] G. Lakoff, *Women,Fire,and Dangerous Things, What Categories Reveal about the Mind. The University of Chicago Press, Chicago and London, 1987*

[李 92] 李航, 優先度 = 意味ファジイ ∧ 意味確率 ∧ 構文確率, 情報処理学会自然言語処理研究会,*92-10,Nov.1992*

[仁木 91] 仁木直人, 確率の立場から見たファジィ理論, 日本ファジィ学会誌, *Vol.3, No.4, 1991*

[荻野 87] 荻野綱男, シソーラス作成の問題点, 言語, *Vol.6, No.5, pp.64-71, 1987*

[岡田 86] 岡田雅勝, ウィトゲンシュタイン, 人と思想 *76*, 清水書院, *1986*

[佐藤 90] 佐藤理史, *MBT1:* 実例に基づく訳語選択, 人工知能学会誌, *Vol. 6, No. 4, 1990*

[Sato90] S. Sato, M. Nagao, *Toward Memory-based Translation, COLING90, 1990*

[隅田 91] 隅田英一郎, 飯田仁, 用例主導型機械翻訳, 情報処理学会自然言語処理研究会, *82-5,1991*

[Taylor89] J. Taylor, *Linguistic Categorization, Prototypes in Lingustic Theory. Clarendon Press, Oxford, 1989*

[徳永 89] 徳永健伸, 奥村学, 田中穂積, 概念階層への視点の導入, 情報処理学会論文誌, *Vol 30, No 8, 1989*

[鶴丸 91] 鶴丸弘昭, 竹下克典, 伊丹克企, 柳川俊英, 吉田将, 国語辞典情報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*

[Utsuro92] T. Utsuro, Y. Matsumoto, M. Nagao, *Lexical Knowledge Acquisition from Bilingual Corpora, COLING92, 1992*

[Velardi89] P. Velardi, M.T.Pazienza, *Computer Aided Interpretation of Lexical Cooccurrences, ACL89, 1989*

[渡辺 92] 渡辺日出雄, 浦本直彦, *Example-Based Machine Translation* の問題点に関する考察, 情報処理学会第45回全国大会,*1992*

[Webster89] M. Webster, M. Marcus, *Automatic Acquisition of the Lexical Semantics of Verbs from Sentence Frames, ACL89, 1989*

謝辞

本研究の機会を与えて下さったNEC C&C情報研究所情報基礎研究部の中村勝洋部長に感謝いたします。本研究を進めるにあたって、有益なコメントを数多く頂いた安倍直樹主任(NEC C&C情報研究所情報基礎研究部)、伝康晴氏(京都大学工学部/ATR自動翻訳電話研究所)等、多くの方々に感謝します。