情 報 学 基 礎 39－7
（1995．9．14）

帰納学習アルゴリズムと階層型クラスタリング手法を用いた概念シソーラスの自動構築及び更新

山崎 毅文
NTT コミュニケーション科学研究所

マイク バッザーニ
カリフォルニア大学 アーバイン校

概念シソーラスの構築は、機械翻訳、テキスト検索等自然言語処理システムにおいて重要なタスクである。本稿では、与えられた事例集合から、機械翻訳システム向け翻訳ルールの学習と概念シソーラスの構築とを同時に行なう手法を提案する。概念シソーラスの構築では、既存のシソーラスを利用しない場合／更新する場合の２つの手法を提案する。利用しない場合は、まず、帰納学習アルゴリズムの適用により翻訳ルールの学習を行なう。次に学習されたルールからカテゴリー間類似度行列を作成し、クラスタリング手法の適用により、シソーラスを構築する。シソーラスを更新する場合は、既存シソーラスを一旦カテゴリー間類似度行列に変換することにより、シソーラスを利用しない場合と同様の手法が適用できる。実験結果により、本手法で構築されたシソーラスが、機械翻訳タスクにおいて有用であることが確認された。

Acquiring and updating a semantic hierarchy through induction and clustering

Takefumi Yamazaki
NTT Comunication Science Laboratories
1-2356 Take Yokosuka-shi Kanagawa 238-03 Japan
yamazaki@nttkb.ntt.jp

Michael J. Pazzani
University of California, Irvine, USA
pazzani@ics.uci.edu

This paper addresses the problem of constructing a semantic hierarchy for Japanese-English translation systems. The creation of a comprehensive hierarchy is one important step in this system because it is used to bias the learning of rules that indicate the English translation of Japanese verbs. We propose two methods of constructing a hierarchy: acquiring a hierarchy from scratch and updating a hierarchy. When acquiring a hierarchy from scratch, translation rules are learned by an inductive learning algorithm in the first step. A new hierarchy is then generated by applying a clustering method to internal disjunctions of the learned rules and new rules are learned under the bias of this hierarchy. When updating an existing manually-constructed hierarchy, we take advantage of its node structure. We report experimental results showing that the semantic hierarchies generated by our method yield learned translation rules with higher average accuracy.

1 はじめに

階層型知識の一つであるシソーラスは、多くの自然言語処理システムにおいて、重要／不可欠なものとなっている。例えば、多義である動詞の意味を同定する問題においては、分類語彙表やEDRシソーラス等が有効利用されている[黒橋 92][村木 95]。

一方、シソーラスが、自然言語処理システムに利用された場合、得られる結果はシソーラスの質に大いに依存する。我々が、機械学習アルゴリズムによる機械翻訳システム向け翻訳ルールの生成を試みた場合にも、学習されるルールの質が、学習時に背景知識として用いられるシソーラスの質に依存するという問題点があった[Almuallim et al. 1994]。既存のシソーラスは、人手作成であるため、対象となるシステムにとって、必ずしも最適な意味体系である保証はないことから、シソーラスの自動構築が望まれていた。

本稿では、帰納学習アルゴリズムと階層型クラスタリングの併用による、概念シソーラスの一つである名詞意味属性体系の自動獲得及び洗練手法を提案する。

提案する自動獲得手法では、まず帰納学習アルゴリズムを用いて、事例集合から翻訳ルールを求める。その際、背景知識としての意味属性体系は利用しない。次に、前段で得られたルールの条件部に現れる意味属性の出現頻度から相互情報量に基づいて、意味属性間の類似度を計算し、意味属性間の類似度行列を求める。最後に、この類似度行列から、クラスタリング手法を用いてクラスター木を作成し、これを意味属性体系とする。

一方、洗練手法では、まず既存の意味属性体系から、ノードの深さ等の情報を用いて、意味属性間の類似度行列を求める。次に、自動獲得手法と同様の方法で、意味属性間類似度行列を求める。次に、これら2つの行列を適度に混合し、一つの類似度行列を作成し、この行列からクラスタリング手法を用いて、意味属性体系を得る。

本稿では、まず、対象とする翻訳システムにおける知識ベースの概要について述べる。次に、帰納論理アルゴリズムを利用した、翻訳ルールの学習手法について述べる。次に、学習された翻訳ルールから、クラスタリング手法を用いて、意味属性体系を作成する方法について述べる。次に、意味体系構築において、既存意味体系を利用した場合の手法を、意味体系の修正手法として述べる。さらに、人工データ、実データを使った実験を行ない、本提案手法の有効性を検証する。

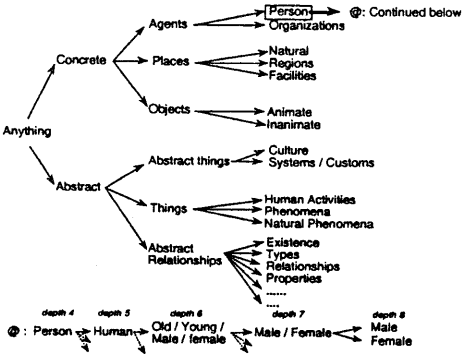

図 1: 一般名詞意味属性体系の一部

2 翻訳システムにおける知識ベースの構成

NTTで開発中の日英機械翻訳システムALT-J/E[Ikehara et al. 90][池原 91]を対象に、翻訳システムにおける知識ベースの構成について述べる。

本翻訳システムは大別して、3つの知識ベース: 意味属性体系、名詞辞書、翻訳ルールから構成されている。

図1に示すように、**意味属性体系**は、木構造をした一種の概念シソーラスであり、各ノードは、意味属性で記述されている。ノード間を結ぶエッジは、意味属性間のis-a関係を表す。現在、本体系は、12の階層、2715個のノードで構成されており、これらのノードは790個の中間ノードと、1925個の末端ノードにより構成されている。

名詞辞書は、約40万語の名詞を収録し、全ての名詞に対し、各々一つないしは複数(通常は、複数)意味属性が割り当てられている。例えば、名詞"コック"に対しては、[人]と[職業]の2つの意味属性が割り当てられいる。

翻訳ルールは、日本語文パターンと英文パターンの対応関係を規定しており、日英翻訳は、基本的にこのルールに基づいて実行される。一般に、図2に示すように、一つの日本語動詞に対して、通常、複数の英語動詞の対応づけがなされる。本ルールは、図で示すように、ルール条件部に日本文パターンを、ルール実行部に英語動詞を持つ。ここで、日本文パターンは、一つの日本語動詞、格要素(『主格』、『目的格』等)の主名詞が持つべき意味属性条件から構成される。ルールの条件部に表れる、[人][時間]等は、意味属性の一つである。ALT-J/Eでは、現在、10,000を越える翻訳ルールを保持している。

IF		THEN	
J-Verb	= "使う"	主格	$= N_1$
N_1 (主格)	≡ [主体]	E-Verb	="spend"
N_2 (目的格)	≡ [時間] or [金銭]	目的格	$= N_2$
IF		THEN	
J-Verb	= "使う"	主格	$= N_1$
N_1 (主格)	≡ [主体]	E-Verb	="employ"
N_2 (目的格	≡ [人々]	目的格	$= N_2$
IF		THEN	
J-Verb	= "使う"	主格	$= N_1$
N_1 (主格)	≡ [主体] or [人工物]	E-Verb	="use"
N_2 (目的格)	≡ [名詞一般]	目的格	$= N_2$

図 2: 日本語動詞 "使う" の翻訳ルールの例

3 帰納学習システム FOCL を用いた翻訳ルール学習

3.1 学習すべきタスク

本章では、背景知識として意味属性体系を用いて、事例集合から翻訳ルールを獲得するタスクについて述べる。

まず、学習アルゴリズムに与えられる訓練事例の形式について述べる。例えば、訓練事例として、日本語文:『社長が金を使う。』、英語文:『The president spends money』のペアが与えられたとする。

以上のペアから、ALT-J/E による構文解析ツールを利用して、各文の要素を特定し、以下のような結果を得る。

⟨[J-VERB = 使う , 主格 = "社長",
目的格 = "金"], E-VERB = spend).

これをこのまま訓練事例として用いるのでなく、文の要素の名詞を、その名詞が持つ意味属性に置き換えたものを訓練事例とする。前述の通り、名詞は、一般に一つないしは複数個の意味属性を持つ。例えば、"社長" の意味属性は、[長] と [管理職] であり、また、"金" の意味属性は、[金銭]、[貨幣]、[金属] であることが名詞辞書により、解る。よって、最終的に日本語動詞『使う』に対するルールを学習する際に学習アルゴリズムに与えられる形式は、次のような形になる。

⟨ [主格 ≡ { [長], [管理職] },
目的格 ≡ { [金銭], [貨幣], [金属] }],
E-VERB = spend),

ここで、$N \equiv S$ は、文の要素 N が、意味属性の集合 S で構成されることを意味する。上記の例は、主格や目的格等の文要素に割り当てられる意味属性の数は、2 以上であるという意味において、意味的曖昧性を持つ。一般に、学習対象の日本語動詞を決めた段階で、訓練事例の形式は、以下の通りになる。

$$\langle [\ N_1 \equiv \{a_1, a_2, \cdots\}, \\ N_2 \equiv \{b_1, b_2, \cdots\}, \cdots \\ N_n \equiv \{c_1, c_2, \cdots\}], \textit{E-Verb} \rangle$$

ここで、N_i は、主格、目的格等の文の要素を表し[1]、a_i, b_i, c_i は、意味属性を表す。

学習アルゴリズムのなすべきタスクは、上述した形式で与えられる訓練事例から、翻訳ルールにおける格要素の持つべき意味属性を同定することである。言い換えれば、学習対象の日本語動詞を一つ決めた時に、与えられた文脈に応じて、その日本語動詞を適当な英語動詞に割り当てるように、ルールの条件部、各文要素が取るべき意味属性を同定することである。

3.2 拡張版 FOCL の適用

3.2.1 FOCL における知識表現

我々は、翻訳ルール学習に、一階述語論理式を対象とする学習システム FOCL [Pazzani & Kibler 1992] を用いた。FOCL は、事前に与えられる述語の集合を用いて、ある決められたクラスを特徴づけられるように、ホーン節の集合を構築する。FOCL における節及びリテラル選択戦略は、基本的に FOIL [Quinlan 1990] で用いられたものと同様である。即ち、全ての正事例が少なくとも一つの節でカバーできるまで、節を生成する。節を構成するリテラルは、候補となるリテラル集合の内、最も情報量 (Information gain) の値が最も大きいリテラルを選択することで決まる。

まず、翻訳ルールを一階述語論理式の形式で表現できるように、必要な述語を定義する。意味属性間の一般／特殊関係 (いわゆる、is-a 関係) を表す述語 is-a(Term, Category) を導入する。is-a(Term, Category) は、Term にバインドされる意味属性が、Category にバインドされる意味属性と is-a の関係にある時に、真

[1] N_i は、深層格を表す変数である。現在、14 種類の深層格が定義されている。

の値を取る。さらに、必要な述語として、前章で述べたような意味曖昧性を扱えるように、is-a 述語を拡張し、2項述語 one-isa(TermSet, CategorySet) を新たに定義した。この one-isa(TermSet, CategorySet) は、TermSet の少なくとも一つの値が、CategorySet の少なくとも一つの値と is-a 関係にある時に、真の値を取る。この one-isa 述語を用いて、翻訳ルールを記述するに適した、選言的リテラル(Internal Disjunction) の形式で、ルールを記述することが可能になる。例えば、日本語動詞『使う』が、英語動詞『spend』に訳されることを示す翻訳ルールは、is-a, one-isa 述語を用いて、次のように、表現できる。

使う (Subject, Object, Everb) :-
isa(Subject, {[主体]}),
one-isa(Object, {[時間], [金]}), Everb = spend.

3.2.2 山登り法による選言的リテラルの学習

FOCL において、前述した one-isa 述語が扱えるように、FOCL のリテラル選択機構を変更した。リテラル選択戦略は、以下の通りである。各変数 V 毎に、まず、one-isa(V, c_1),...,one-isa(V, c_n) の中で、情報量が最大である、one-isa(V, c_i) を選ぶ。次に、少なくとも、一つの正事例を満足する意味属性の集合から、c_i と is-a 関係にある意味属性を除く。次に、得られた意味属性集合の要素 c_j に対して、one-isa(V, $\{c_j, c_i\}$) の情報量を計算し、one-isa(V,c_i) の情報量のものと比べ、情報量が増える間、リテラルの項を付加する。新たな項の付加は、山登り法に基づいて、行なわれ、最終的に、one-isa(V,$\{c_1, ..., c_n\}$) という形式の述語で表現されるルールが得られる。

4 階層型クラスタリング手法を用いた意味属性体系の構築 / 修正

4.1 相互情報量に基づく意味属性の類似度計算と意味属性体系の構築

以上述べたように、意味属性体系上の意味属性は、翻訳ルールの条件部を記述するのに、適度な一般化を与えるという意味で、本学習タスクでは、重要な意味を持つ。望ましい意味属性体系とは、適度な一般性をもつ意味属性ノードを中間ノードとして持った体系である。ここで、我々は、意味属性体系を構築するタスクを、予め末端ノードが与えられた際、適当な中間ノードを生成するタスクとして設定した。

意味属性体系構築の手順を以下に示す。まず、意味属性体系を用いずに、前章で述べた方法に従って、FOCL によって翻訳ルールを学習する。学習されたルールの条件部の選言的リテラル上に (即ち、リテラル one-isa(V, $\{c_1,..., c_n\}$) の第2項 $\{c_1,..., c_n\}$)、共起して出現する意味属性の出現頻度を利用して、意味属性間の類似度を求める。意味属性間の類似度は、出現頻度から相互情報量を計算することにより求める。意味属性 c_i、c_j が、学習された全てのルールに出現する確率を各々 $p(c_i)$、$p(c_j)$ とし、また意味属性 c_i、c_j が共起して出現する確率を $p(c_i\&c_j)$ とすると、意味属性 c_i、c_j の相互情報量は、$log(p(c_i\&c_j)/p(c_i)p(c_j))$ によって、計算され、これを両属性間の類似度とする。

次に、計算した相互情報量を基に、意味属性間の類似度行列[2]を作成する。さらに、本行列に階層的クラスター分析手法の一つである、群平均法 [田中 & 脇本 1994] を適用して、意味属性体系を構築した。一つのクラスタが、3つ以上の要素を保持できるように、標準的な群平均法に対し、若干の変更は行なった。クラスタリングアルゴリズムの概要を、表1に示す。

4.2 既存意味属性体系の構造を利用した意味属性体系の更新

前章では、事例のみから意味属性体系を構築する方法を提案したが、完全な体系を構築するためには、かなりの数の事例数が必要である。一方、人間によって予め構築された体系があるなら、それを有効利用した方が、少ない事例でより質の高い体系が得られることが期待される。本章では、既存意味属性体系を利用した意味属性体系の更新について述べる。意味属性体系の更新手法は、既存意味属性体系とその体系を用いて学習した翻訳ルール集合とを入力とする。手法の概要は以下の通りである。

[2]本方式で求められた行列は、2つの項目が無関係の時に値が0となるので、通常非類似度行列と呼ばれる。値が大きいほど、該当する2つの項目の関連が強いと見なされる。それに対し、類似度行列とは、2つの項目が全く同一である場合には、値が0となり、値が大きいほど、2つの項目が無関係であることを意味する。

Step 1: 類似度行列中で、値が最も大きい属性ペアを選ぶ。
Step 2: 選ばれたペアの一方の属性の列ベクトル中で、
前段階で選んだペアの値と同一の値を持つ、
他の意味属性を全て列挙する。
Step 3: 前段階で選んだ意味属性を一つにまとめる中間
ノードを生成する。
Step 4: 新しく生成したノードと、他の意味属性との
類似度を再計算する。
新しいノードと意味属性との類似度を、
新しいノードを生成したノード群と
対象意味属性との平均値とする。
Step 5: 類似度行列に対して、新しいノードを付加し、
新しいノード生成に利用された意味属性を
削除し、また前段階で再計算した結果の値を
反映させることにより類似度行列を再構成する。
Step 6: 類似度行列が、空になれば、止める。
それ以外は、step 1 に戻る。

表 1: 群平均法に基づく階層型クラスタリング

まず第一に、既存の体系から体系上の末端ノード間の類似度行列を求める。このノード間類似度計算は、まず対象である2つの末端ノードからルートに向かって、共通のノードを求める。次に、各々のノードからその共通ノードに辿りつくに要した is-a リンクの数のうち、小さい方の値の逆数をノード間の類似度とする。ここで、この類似度行列を、E と名付ける。

第二に、既存の体系を用いてルールを学習し、そのルールに表れる属性体系の中間ノードを、属性体系を使って全て末端ノードの選言の形に置き替える。次に、前章で述べた方法と同様に、それらのルールから、意味属性間の類似度行列を求める。この類似度行列を、R と名付ける。

第三に、これらの行列を足し合わせる。足し合わせる際には、まず、行列の要素を、値域が [0,1] になるように、正規化しておく。混合された類似度行列 C は、重み付けのパラメータ p [3]を用いて、次のように記述できる。

$$C_{ij} = pE_{ij} + (1-p)R_{ij}$$

最後に、この行列 C から、先に述べたクラスタリング手法で、体系木を作成する。提案した意味属性体系構築／更新手法の概要図を、図 3に示す。

5 実験結果

5.1 人工データと実データ

提案手法によって生成される意味属性体系を評価する実験を行なった。実験には、2種類のデータセットを用いた。一つは、人工データであり、ALT-J/E システムで現在用いられている意味属性体系と翻訳ルールを基に作成した。 事例生成手法の詳細は、文献 [金田 et al. 95] を参照して頂きたい。日本語動詞を訳し分ける英語動詞ルール一つに対し、300 事例を生成した。今回、24 種類の日本語動詞を扱った。各々が、3～6個の英語動詞ルールから構成されており、合計約 100 個の英語動詞ルールを対象としたので、全部で、約30,000 個の事例を生成し、利用した。

もう一方のデータセットは、実データであり、日英の対訳コーパスは、日英表現辞典 [Keene 91] 等の辞書から、単文化して、収集した。収集したデータを、ALT-J/E のツールを用いて構文解析した上で、学習データを生成した。収集されたデータは、約 48,000 文であり、5000 種類の日本語動詞を含んでいた。これらの日本語動詞のうち、事例数が 30 以上ある日本語動詞 30 個を選んで、実験した。これらの日本語動詞は、約 120 個の英語動詞ルールに相当し、実験に利用された事例数は、合計で約3,400 個であった。

実データは、次の点で、人工データよりも、扱いが難しい。まず第一に、実データは、主格や目的格等の属性が、通常、2 つ以上の値を持ち、曖昧性を含んでいる点である。第二に、100% 正確なパーザは、現在の所ないので、パージングされた結果に誤りを含む可能性があり、生成される訓練データに、ノイズが含まれる可能性が大きい点である。

[3]我々の実験では、両行列が同じ重みになるよう、p を 0.5 に設定した。

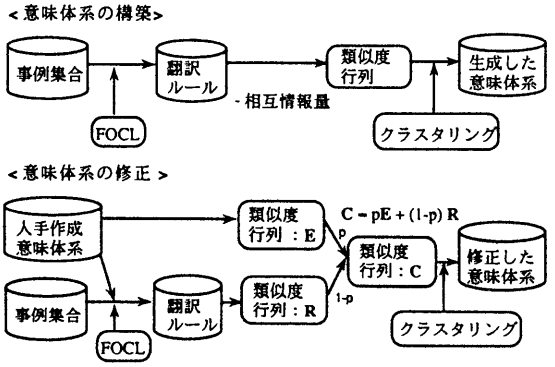

図 3: An overview of our proposed methods

5.2 人工データを用いた実験

5.2.1 意味属性体系の構築

正解率は、10 クロスバリーデーションを用いて評価した[4]。表 2 に、3 つの条件 (1. 意味属性体系を使用しない。2. 人手作成の体系を利用。3. 提案手法で求めた体系を使用。) で、学習されたルールの正解率を示す。24 日本語動詞の正解率を平均すると、意味属性体系を利用しない場合、85.9% の正解率を示し、一方、人手作成の体系を利用した場合は、97.4% の正解率を示す。($t(23) = 5.26, p < .0001$[5])。さらに、提案手法で求めた体系で学習した結果は、90.6% であり、意味属性体系を利用しない場合より、良い結果が得られた。($t(23) = 4.13, p < .001$)。本結果は、翻訳ルール学習において、意味属性体系が背景知識として有効であることを示している。

図 5に、提案手法によって生成した意味属性体系の一部を示す。図において、"hxxx"は、新たに生成されたノードを示す。今回の実験で、生成された全ノードの数は、415 個であった。生成された体系は、人手で作成された体系の一部を再生している。本結果より、生成された体系は、意味的に尤もらしい体系であることが解る。図 4に、日本語動詞『飲む』『焚く』に対する翻訳ルールの学習結果を示す。本結果から、図 5で表現されたように、新しく生成されたノードが、学習したルール上で利用されていることが解る。

表 2: 正解率: 人工データの場合

正解率 (%)				
意味体系使用せず	人手作成意味体系	生成した意味体系	ノイズ入り意味体系	修正した意味体系
85.9	97.4	90.6	93.3	97.1

飲む (Subject, Object, Everb) :-
one-isa(Object, {h294}), Everb = take.

焚く (Subject, Object, Everb) :-
one-isa(Object, {h107, 液体燃料, 気体燃料 }),
Everb = burn.

図 4: 生成された意味属性体系を用いて学習したルールの一部

[4]まず、各日本語動詞について、10 % の事例集合をテスト事例として取っておく。次に、残りの 90% の事例を学習事例として、帰納学習アルゴリズムによって、意味属性体系を用いずに、翻訳ルールを求め、この学習したルールから提案手法を用いて、意味属性体系を得る。さらに、この得られた意味属性体系を用いて、帰納学習アルゴリズムによって、再び翻訳ルールを求め、このルールの精度を、先に取っておいたテスト事例を用いて測定する。再び、テスト事例を違う事例集合に設定し、上記の手続きを 10 回繰り返す。

[5]t-test の結果を表す

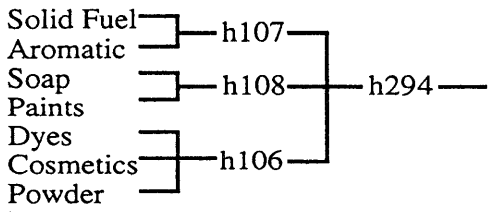

図 5: 生成された意味属性体系の一部

表 3: 正解率: 実データの場合

正解率 (%)			
意味体系使用せず	人手作成意味体系	生成した意味体系	修正した意味体系
83.9	84.9	86.4	87.2

5.2.2 意味属性体系の修正

属性体系の修正手法を評価するため、人手作成の体系上の幾つかのノードを削除することにより、人為的にノイズを含む属性体系を構築した。本実験では、人手作成体系を用いて学習したルールに表れた中間ノードを削除することにより、構築した。この意味属性体系を"ノイズ入り体系"と名付ける。

表2に、本実験の結果を示す。"ノイズ入り体系"で学習したルールは、"人手作成体系"で学習したのよりも、精度が低くなっており、中間ノードを削除した影響が表れている。"ノイズ入り体系"を修正した体系で、学習したルールの正解率は、平均して、97.1% であり、"ノイズ入り体系"で学習したものよりも、精度は高い ($t(23) = 3.672, p < .005$)。

5.3 実データを用いた実験

人工データを利用した時と同様の方法で、10クロスバリデーションによる実験を行なった。表3に結果を示す。人工データの時との違いは、"人手作成体系"をそのまま"ノイズ入り体系"と見なし、修正手法の対象としたことである。

表3によると、"提案手法で生成した体系"の方が、"人手作成体系"よりも、やや良い精度を示すルールを生成した ($t(29) = 1.7, p < .1$)。また、"修正手法で生成した体系"は、"人手作成体系"よりも、良い精度を示すルールを生成した ($t(29) = 3.796, p < .001$)。以上の結果は、提案手法が有望であることを示したが、実データを用いた結果は、人工データを用いたものに比べて、精度の点で劣っている。これは、実データとして、十分な数のデータが利用できなかったため、実データが潜在的に持つ曖昧さに対処できなかったためと考えられる。今後、新聞データ等、他の対訳コーパスの利用により、実データの充実を図りたい。

6 関連研究

クラスタリングは、従来概念クラスタリングの名の元で、研究されてきたが (例えば、[Fisher 1987]; [Michalski et al. 1986])、これらの手法は、直接的に今回の我々のタスクに適用できない。概念クラスタリングは、クラスタリングの対象が、特徴ベクトルで表現されているものを対象としているのに対して、一方、我々の対象タスクでは、与えられる情報が、クラスタリング対象の特徴ベクトルではなく、対象間の類似度であるためである。

自然言語処理の分野においては、類似語辞書の自動構築の研究が幾つか報告されている。Hindle [Hindle 90] は、動詞とその目的語との相互情報量を用いて、動詞とその目的語に表れる名詞間の類似度を求めた。Grenfenstette [Grefenstette 1992] は、対象分野で表現される名詞の修飾子を元に、名詞をクラスタリングする手法を提案した。しかし、両者とも、類似語の同定を主眼としており、シソーラス構築の際の尺度として、類似度を利用していない。

構成的帰納学習の分野では、その多くが、学習した結果を用いて新たな項を生成し、その新たに生成された項を用いて、学習される概念の記述をコンパクトにするものである。(例えば、Fringe [Pagallo et al. 1990]) 一方、本提案手法で用いた方法は、主旨は構成的帰納学習の分野と同じであるが、新たな項の生成の方法が異なり、学習によって得られた翻訳ルールの条件部に表れる内的選言に着目し、頻繁に共起するものを新たな項として定義する点が、従来手法と異なる点である。

7 まとめ

本稿では、クラスタリングを用いた、意味属性体系の構築／修正手法を提案した。事例集合からの翻訳ルール獲得には、一階述語論理式を対象とする学習アルゴリズムFOCLを用いた。学習したルールとクラスタリング手法を用いて、生成された中間ノードが、翻訳ルール学習における有益なバイアスを与えることを実験的に示した。さらに、人工データ及び実データを用いた実験で、提案した手法によって構築／修正した意味属性体系が、学習される翻訳ルールの精度を向上させることを示した。

一方、今後の課題として次のような問題が残されている。

現在の手法によって、新しく生成される中間ノードは、h1, h2 等単なるシンボルであり、人間にとって理解できる意味のある名前が付けられおらず、人間が陽に意味付けをする必要がある。また、シソーラス修正手法では、2種類の類似度行列を混合したが、一方は相互情報量による値、もう一方はシソーラス上の距離による値であるという意味で、2つの行列の値のスケールが異なっている。今後は、このスケールの違いを考慮した、混合方法を検討する必要がある。また、本手法では、全ての末端ノードが事前に与えられることを前提としていた。本問題を解決する手段として、他のクラスタリング手法(例えば、[Hindle 90])によって、名詞のグループ化を行ない、意味属性体系の末端ノードを構築する手法が考えられる。

謝辞

本研究は主に、筆者が、カリフォルニア大学 アーバイン校滞在中に行なったものである。本研究の機会を与えて頂いた、前NTTコミュニケーション科学研究所 河岡 司所長(現 同志社大学教授)に深謝します。本稿を作成するに当たって、有益なコメントを頂いた、NTTコミュニケーション科学研究所 金田重郎グループリーダ、春野雅彦研究員に感謝します。

参考文献

[Almuallim et al. 1994] Almuallim, H., Akiba, Y., Yamazaki T., Yokoo, A., and Kaneda, S. 1994. A tool for the Acquisition of Japanese-English Machine Translation Rules Using Inductive Learning Techniques. *The 10th Conference of Artificial Intelligence for Applications.*

[Fisher 1987] Fisher, D. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning* 2:139-172.

[Grefenstette 1992] Grenfenstette, G., 1992 SEXTANT: Exploring unexplored contexts for semantic extraction from syntactic analysis, *Proceedings of the 30th Annual Meeting of ACL*, 324–326.

[Hindle 90] Hindle, D. 1990 Noun classification from predicate argument structures, *Proceedings of the 28th Annual Meeting of ACL*, 268–275.

[Ikehara et al. 90] Ikehara, S., Shirai, S., Yokoo, A. and Nakaiwa, H. 1990. Toward an MT System without Pre-Editing—Effects of New Methods in ALT-J/E. *Proc. of MT Summit-3.*

[池原 91] 池原 悟, 宮崎 正弘, 横尾 昭男, "日英機械翻訳のための意味解析辞書", 信学技報, *NLC 91-19*,1991.

[金田 et al. 95] 金田 重郎, 秋葉 泰弘, 石井 恵, "事例に基づく英語動詞選択ルールの修正型学習手法", 言語処理学会第一回大会, *p333-336*,1995.

[黒橋 92] 黒橋 貞夫, 長尾 眞, "格フレームの選択における意味マーカと例文の有効性", 情処研究会, *NLC 91-11*,1992.

[Keene 91] Donald Keene, "日英表現辞典(改定版)", 朝日出版, 1991.

[Michalski et al. 1986] Michalski, R., and Stepp, R. 1986 Conceptual clustering of structured objects: A goal-oriented approach. *Artificial Intelligence* 28:43-69.

[村木 95] 村木 一至, 落合 尚良, "動詞語義例文と概念辞書を用いた動詞意味選択", 情処学会第 *50* 回全国大会, *3R-4*, 1992.

[Pagallo et al. 1990] Pagallo, G. & Haussler, D. 1990. Boolean feature discovery in empirical learning. *Machine Learning*, 5:71-100.

[Pazzani & Kibler 1992] Pazzani, M., and Kibler, D. 1992. The utility of knowledge in inductive learning. *Machine Learning*, 9(1):57-94.

[Quinlan 1990] Quinlan, J. R. 1990. Learning logical definitions from relations. *Machine Learning*, 5(3).

[田中 & 脇本 1994] 田中 豊, 脇本 和昌 "多変量統計解析法", 現代数学社, 1994.