1P－10

属性値に多価を許容する事例からの翻訳ルール学習手法

秋葉泰弘、アルモアリム・フセイン、山崎毅文、金田重郎
NTT情報通信網研究所

1．はじめに

近年、自然言語処理応用への学習アルゴリズム適用が試みられている。しかし、この分野では、学習事例の属性値が多価で、しかも、シソーラス等の巨大な背景知識が前提となることが多い。従って、属性値一価を前提とする、従来の機械学習アルゴリズムは、その適用が難しい。

本稿では、属性値に多価を許容できる、事例からのルール学習アルゴリズムを提案する。本アルゴリズムでは、背景知識の上下関係に基づく前処理により、属性値の多価性を許容している。本アルゴリズムを、機械翻訳の英語動詞選択ルールの学習に適用した結果、予め既存の意味解析処理により、多義性を１価に絞り込んだ場合と同等の学習性能を確認できた。

2．自然言語処理に学習アルゴリズムを適用する際の課題

日英機械翻訳ルール、動詞格フレーム等の意味解析ルールを、事例から学習する研究が試みられている。以下、NTTで開発を進めている日英機械翻訳システムALT-J/Eのための、翻訳ルール（正確には、ルール左辺部と英語動詞からなる部分翻訳ルール）を事例から学習する際の問題点を説明する。

2.1 翻訳ルール

翻訳ルールは、「王女が金を使う」といった日本文（単文）が与えられた時に、各単語の意味／格を絞り込むとともに、英語動詞を決定する。例えば、以下の形式を持つ。

IF	THEN
J-Verb ="使う"	
N1(が格) = [主体]	E-Verb = "spend"
N2(を格) = [金銭] or [時間]	

ここで、[主体]、[金銭]、[時間]等は、名詞の意味を分類する意味カテゴリで、ALT-J/Eの場合、この意味カテゴリは約３０００個あり、最大１２段のis-a 関係の階層的シソーラスを構成している。

2.2 事例からの学習における課題

機械学習における事例からの学習技術を利用するためには、事例をベクトル表現する必要がある。そのような学習事例を生成するために、日英対訳文（コーパス）、例えば、

＜王女が金を使う, The princess spends money＞

を係り受け解析すると、以下の様な情報が得られる。

J-Verb ≡ "使う"
N1(が格)≡ {[貴族]、[娘]、[女性]}
N2(を格)≡ {[財産]、[金属]、[曜日]、 [メダル]}
E-Verb ≡ "spend"

ここで、[資産]、[金属]等は、名詞の意味を表わす先の意味カテゴリであり、名詞辞書から得られる。

この様に、得られる情報は、属性値が多価であり、属性値に巨大な背景知識（シソーラス）を前提としている。

これに対して、既存の学習アルゴリズムは、属性値一価を前提とするものが多い。このため、学習前に、人手あるいは日英単語対訳辞書等で属性値を一価に絞る必要があった。また、背景知識の利用を前提とするため、ID3等の背景知識の利用を前提としない高速な学習アルゴリズムを適用できなかった。

3．多価を許容する学習アルゴリズム

先に述べた問題点に対処するために、本稿で提案するアルゴリズムでは、(1)背景知識を用いて、

Induction of Translation Rules from Ambiguous Training Examples.
Yasuhiro AKIBA, Hussein ALMUALLIM*, Takefumi YAMAZAKI, and Shigeo KANEDA
NTT Network Information Systems Laboratories,
1-2356, Take, Yokosuka-shi, Kanagawa-ken, 238-03, JAPAN
* *On leave from the Dept. of Information and Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia.*

学習事例を属性値一価の事例に変換し、次に、(2)ID3等の高速で背景知識を前提としないアルゴリズムにより学習を実行する。但し、上記(2)については説明は要しないと思われるので、以下、(1)について説明する。

まず、学習事例は、次の様な属性ベクトルで表わされるとする。但し、日本語側の動詞は、一個（例えば、「焼く」）に限定し、その動詞に関する事例のみをここでは考える。

$$< \begin{Bmatrix} N1 \equiv \{a1, a2, ----\} \\ N2 \equiv \{b1, b2, ----\} \\ 、、、、 \\ 、、、、 \\ Nn \equiv \{c1, c2, ----\} \end{Bmatrix} , \text{E-verb} >$$

この属性ベクトルは、日本語の単文、及び、それに対応した正しい英語動詞の対を表わしている。N1, N2等は、主語や目的語等の日本文の構成要素を表現し、a1, a2, b1等は、名詞の意味カテゴリである。「Nj ≡ S」は、日本文のNj成分の名詞が、意味カテゴリの集合Sを持つ事を示す。以下、多価を一価に変換する手順を示す。

Step I

各Nj毎に、全事例に出現する意味カテゴリ、ならびに、シソーラス上でその上位の意味カテゴリを取り出す。通常、この意味カテゴリの集合は、各Nj毎に異なったものとなる。

Step II

あるNjについて、上記の取り出された意味カテゴリの集合を {s1, s2, ,,} とする時、各学習事例のNjについて、以下の処理を行う。すなわち、当該事例のNjに出現したs、ならびにその上位概念であるsについては、その属性値を「1」とし、それ以外のsについては、その属性値を「0」として、{s1, s2, ,,} に関する属性値集合を作成し、これを事例の、Nj ≡ Sの代わりに置き換える。

全てのNjについて変換を行うと、事例は、バイナリ属性ベクトルに変換される。本法の特徴は、上記StepIで、現実のデータの偏りを利用して、学習事例の属性となる意味カテゴリを絞り込める点にある。

4. 評価と考察

日本語和語動詞6種類について、以下の2通りの学習実験を行った。評価は、テスト事例1個のCross-Validationによった。

ID3-NA（一価）

ALT-J/E自身の意味解析ルールを用いて、予め意味カテゴリを一価に絞り、本提案のアルゴリズムを適用。学習にはID3を利用。

ID3-A（多価）

各Njに関し、多義性を残し、本提案のアルゴリズムを適用。学習にはID3を利用。

表1に示す様に、いずれの動詞についても、両者は同等の学習性能を示した。

5. おわりに

本稿では、属性値に多価を許容する事例からのルール学習法として、背景知識の上下関係を利用して、多価をバイナリ属性に変換する手法を示した。本手法により、大きな背景知識が前提となる場合でも、現実のデータの偏りを利用して、高速な学習が可能となる。

実際に、和語動詞6動詞について、属性値に多価を許容した場合と、属性値を予め一価に絞った場合について、提案のアルゴリズムの学習性能を比較した。その結果、学習性能に差異はなく、多価の属性値から、正しい属性値が選択されている事を確認できた。なお、本提案のアルゴリズムは、翻訳ルールのみではなく、種々の意味解析ルールの学習に適用できると考えられる

文献

[H. Alumuallim et. al. 93] "Acquisition of Machine Translation Rules Using Inductive Leaining Techniques", IJCAI Workshop, 1993.

[Ikehara 90] "Toward an MT System without Pre-Editing-Effects of New Methods in ALT-J/E", Proc. of MT Summit-3, 1990.

[Quinlan 86] "Induction of Decision Trees", Machine Learning, 1 (1) : 81-106, 1986.

表1 評価結果

日本語動詞	英語動詞	事例数	ID3-NA	ID3-A
使う	use, spend, employ	80	93%	91%
飲む	drink, take, eat, accept	42	98	93
行う	conduct, play, hold	33	88	88
応じる	answer, enter, meet	30	87	90
焼く	burn, bake, roast, broil, cremate	27	89	93
解く	solve, undo, dispel	29	100	97
平均			92.5	92.0