決定木学習とヒューリスティックスを用いた用言意味属性の自動付与

1N－6

中岩　浩巳[*1]
NTT　コミュニケーション科学基礎研究所

関　嘉代[*2]
NTTアドバンステクノロジ

1．はじめに

機械翻訳等での自然言語の文脈処理では，文間関係を解析するのに，文の意味を用言の意味で代表し，用言間の意味的関係を追跡することがよく行われる．しかし，用言の種類が数万に上るため，その個々を用いてルール化することは必要な知識量の爆発を招き事実上不可能である．よって，用言の意味的用法を意味属性として縮退分類することが必要となる．

用言の分類に関しては従来から様々な研究がなされているが，我々は用言の持つ語義と用法の関係に着目して，図1に示すような原言語と目的言語の文型のパターン対形式からなる日英機械翻訳システムの日英構文意味辞書 [1]を対象に日本語用言の意味属性を 106 種類に分類し体系化した[2]（図2）．

しかし，この意味属性の辞書エントリーへの付与は，個々の属性の意味をよく理解していないと困難である．よって，不足する辞書エントリーの拡充や利用者がその利用目的に応じて利用者辞書として辞書エントリーを登録する際にも，体系を熟知した専門家による作業が必要となり大きな問題であった．

この問題を克服するには，(1)属性値付与の専門家が属性値付与フローを書き出しそれを用いて属性値を付与する手法，(2)属性値が付与済みの辞書情報を用い，辞書エントリーの特徴と属性値との相関を分析し抽出された決定ルールをもとに属性値を推測する手法が提案されているが，それぞれ単独では十分な付与精度が得られなかった[3]．

本稿では両者の利点を活かし欠点を克服するため，(2)の決定木学習を用いた属性値付与手法に(1)のヒューリスティックスも活用した手法を提案する．

2．決定木自動学習による用言意味属性の付与

本章ではまず(2)の，属性値が付与済みの辞書情報を用いて辞書エントリーの特徴と属性値との相関関係を自動分析し，抽出された決定ルールをもとに属性値を自動推測する手法について説明する．属性値の決定ルールの獲得手段は，Quinlan の決定木学習プログラム C5.0[4]を用いた．本プログラムでは，学習データとして，作成される決定木の分岐条件となる特徴量の値のリストと，その特徴値で実際に付与された属性値の対のデータを学習データとして入力し，属性値を決めるための決定木を出力する．

2．1　使用する特徴量

決定木学習アルゴリズムが生成する決定木の性能は，学習データ量と特徴量の種類に依存する．特に特徴量は，その種類により属性値を決定する効果が大きく違うので，特徴量は注意深く選択する必要がある．

[意味的結合価パターン変換辞書]
・ N1(主体)が　N2(文化 人間活動)を　暗記する．
　　＝>N1 learn N2 by heart
[慣用表現変換辞書]
・ N1(主体)は　背が高い　=>　N1 be tall

図1　日英構文意味辞書
(* N1,N2 等は結合価のラベル，括弧内は格への意味的制約)

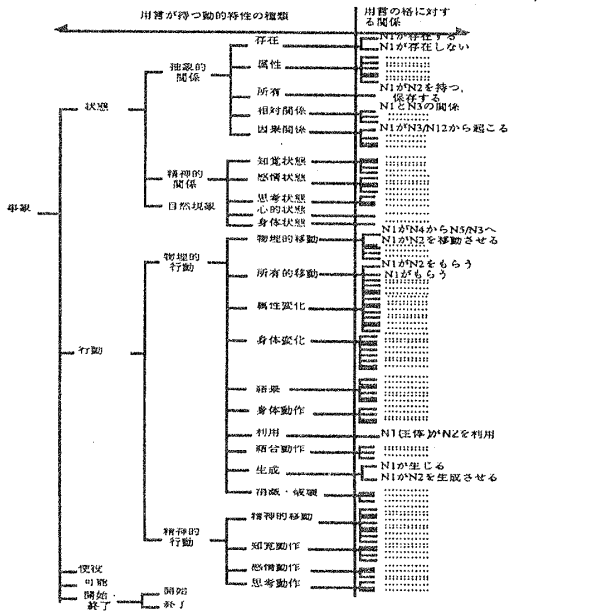

図2　用言意味属性体系

図1の様な日英構文意味辞書の日英パターン対に対して，用言意味属性の属性値を自動付与するため学習データとして選択した特徴量を以下に示す．

(a) 格パターンの種類

用言意味属性は，用言の格に対する関係の観点からも分類している．よって，このパターン中の格要素の種類が特徴量として有望である．図1の日英構文意味辞書では，格要素の種類は N1(動作主;主にガ格),N2(対象 1;主にヲ格)のようにラベル付けされ,用言意味属性も <N1 と N2 の相対関係>のように格ラベルが指定されている．よって，N1,N2 の様なパターン対の格ラベルの種類を特徴量として活用する．

(b) 英語パターン中の英訳語の種類

日英構文意味辞書は日英の等価表現対で構成されているため日本語の多義も英語との対ではその多義が解消できる場合が多い[1]．よって英語パターン中の英訳語の利用が有望である．このような傾向は，例えば"walk"や"eat"があると<N1(人/動物)の具体的身体動作>の属性値が付与されやすいという専門家の分析[3]でも示されている．

(c) 日本語用言の意味カテゴリ

用言意味属性は，属性値が個々の用言部分の意味だけから決まる場合がある．例えば，<身体動作>の用言は，"投げる"等の<操作>の意味カテゴリのが多い．この日本語用言の意味カテゴリも特徴量に活用する．

Automatic Addition of Verbal Semantic Attributes using Decision Tree Learning and Heuristics.
Hiromi Nakaiwa[*1] and Kayo Seki[*2]
NTT Communication Science Labs.[*1] and NTT Advanced Technology Corp.[*2]

2．2　決定木学習手順

決定木は下記の手順で作成する．

[step 1] 特徴量-属性値対のデータ抽出
日英構文意味辞書の各辞書エントリから 2.1 の特徴量と用言意味属性の対の情報を抽出する．

[step 2] 特徴量中の意味カテゴリの加工
step 1 で抽出された特徴量の内，2.1 (c)の意味カテゴリの特徴量を，Almuallim の手法[5]と同様に，そのカテゴリそのものだけではなく，その意味体系上の上位概念のカテゴリも特徴量として追加する．

[setp 3] C5.0 への入力学習データ形式への加工
各辞書エントリーの特徴量を C5.0 が受け取れるデータ形式に加工する．具体的には特徴量が辞書エントリーに現れるかを有無の 2 値で表現した特徴量リストと属性値を C5.0 への入力学習データする．

[step 4] C5.0 への学習データの入力と決定木学習
step 3 で作成された学習データを C5.0 に入力し，このデータの傾向を反映した決定木を作成する．

3．ヒューリスティックス

2 章の決定木学習に基づく手法は，属性値が付与済みの辞書エントリの統計的分析に基づき付与ルールが獲得されるため，属性値の付与傾向が直接ルールに反映される利点はあるが，属性値及び特徴量の頻度が低いとルールの信頼性が低くなるという問題があった．本章では，この低頻度による不適切な属性値付与の問題を回避するために，属性値別に，その属性値を付与するパターン対には必ず持っていなければならない特徴量を必須条件としてルール化した．具体的には，個々の属性値の定義で用いている下記の特徴量に基づく必須条件ルールを属性値別に属性値付与の専門家が作成した．

(a) 格パターンの種類（2．1(a)と同様）
例：<N1 と N3 の相対関係>→N1 と N3 を含む

(b) 日本語パターン中の格要素への意味的制約
用言意味属性には<N1(主体)が N2 を利用>のように，その属性の条件として格への意味制約を明示しているものがある．よって日本語パターンの格への意味制約が属性値を決定する場合があるので活用する．
例：<N1(主体)が N2 を利用>
→ N1 の意味制約が主体かその下位属性である

4．評価

属性値付与済み辞書情報を用いて決定木学習プログラム C5.0 により決定木を自動学習しそれを用いて辞書エントリーに属性値を付与した際の精度と，ヒューリスティックスも活用した際の効果を評価する．

4．1　評価方法

(a) 決定木学習で用いる特徴量・辞書エントリー
日英構文意味辞書の意味的結合価パターン辞書中で用言意味属性が付与済みパターン対から 2.1 で述べた 3 種類の特徴量を使用して学習を行った．複数属性値が付与されたエントリは，個々の属性値別に同じ特徴量を持つエントリとして学習対象に加える場合と，加えない場合の決定木を作成した．また，ヒューリスティックスの効果の検証用に，文献[3]の人手作成付与フローを用いた非専門家による付与評価で用いた 99 エントリに対するブラインドテストを行うため，これを含めず学習した決定木も作成した．

(b) 決定木学習プログラムの走行条件
決定木学習は 3 種類の特徴量を用いて特別なパラメータを設定していない C5.0 により行った．

(c) 決定木による属性値付与評価対象
決定木学習で用いた学習データと同じ全辞書エントリーに対し，学習決定木により属性値を 1 属性付与した．また，ヒューリスティックスの効果検証用の 99 エントリに対し，決定木が誤って付与した属性値がヒューリスティックルールの必須条件を満たさず排除できるかも検証した．この 99 エントリの評価では，99 エントリを決定木学習に用いなかったブラインドテストによる結果の評価も行った．

(d) 正解属性値
事前に付与されている属性値を正解とした．また，99 エントリに対しては，事前付与属性値とは異なるが専門家が見て適切である場合も評価する．比較のため 99 エントリに対する用言意味属性の非専門家（構文意味辞書は熟知）による人手付与フローを用いた付与評価結果も調査した[3]．

4．2　評価結果

評価結果を表 1 に示す．これによると，複数属性を含んで作成した決定木の方が全体的に高い認定精度が得られた．また，99 エントリへの付与結果では，ウィンドウで一致 78%適切を含むと 85%，ブラインドでも一致 57%適切を含むと 69%と，非専門家による人手付与の結果と同等以上の認定精度が達成できた．さらに，3 章のヒューリスティックスを活用すると認定失敗の約半数が回避できたことから，この有効性が実証できた．

表 1　属性値付与精度精度

決定木作成条件			学習対象	99 エントリ			
複数属性	99 エントリ	学習対象エントリ数	エントリ認定一致	一致	異なるが適切	ルールで排除	失敗
含む	含む	15136	70.8%	78%	7%	8%	7%
	含まない	15037	70.8%	57%	12%	19%	12%
含まない	含む	13512	67.9%	68%	8%	9%	15%
	含まない	13413	67.7%	53%	13%	13%	21%
99 エントリへの人手付与結果[3]				60%	18%	----	22%

5．まとめ

本稿では，用言意味属性の機械翻訳用日英構造変換辞書への効果的自動付与を目指して，属性値付与済みの辞書情報から機械学習により自動的に作成した決定木と属性値別ヒューリスティックルールを併用した属性値付与方法を提案し，その有効性を確認した．今後は，ヒューリスティックスを決定木学習の中での効果的に活用する手法を検討していく．

参考文献

[1] 池原，宮崎，白井，横尾，中岩，小倉，大山，林：日本語語彙大系，岩波書店，(1997).

[2] 中岩，池原：日英の構文的対応関係に着目した日本語用言意味属性の分類，情処論文誌，Vol.38 No.2, pp.215-225 (1997).

[3] 中岩，関：日英対訳情報を活用した用言意味属性の自動付与，言語処理学会第 5 回年次大会発表論文集，pp.201-204 (1999).

[4] Quinlan J. R. : http://www.rulequest.com/

[5] Almuallim, H. et.al: Two Methods for Learning ALT-J/E Translation Rules from Examples and a Semantic Hierarchy, Proc. of COLING-94, pp 57-63 (1994).