日英対訳情報を活用した用言意味属性の自動付与

中岩 浩巳
NTT コミュニケーション科学基礎研究所

関 嘉代
NTTアドバンステクノロジ

1. はじめに

機械翻訳システム等,自然言語処理システムにおける文脈処理では,文と文の関係を解析するのに,文の意味を用言の意味で代表させて,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,中岩らは用言の持つ語義と用法の関係に着目して,日本語用言の意味属性を分類し体系化した[1].この体系では,日本語の用言を,日本語とは言語族が異なり,対象概念の捉え方に大きな違いのある英語の用言と対比することにより,意味分類を行った[1].日本語と英語の用言の意味的対応関係については,図1に示すように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ALT-J/Eの日英構文意味辞書 [2][3]のような機械翻訳用辞書として整理されている.例えば,このALT-J/Eの辞書では,日本語用言とその格要素への意味的制約からなる日本語単文パターンと,それに対応する英語パターンの対が登録されている.そこで,このように整理された各用言毎のパターン対が持つ意味に対して対応する意味属性を付与することにより,日英言語間での用言の意味的多義を解消できる精度の分解能を持つことができた.

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策としては,次の2種類が考えられる.第1の策は,属性値付与の専門家により頭の中にある属性値を決定するための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である.第2の策は,既に属性値が付与済みの辞書情報やコーパス中の単語共起情報を用いて,辞書エントリーやコーパス中の特徴と属性値との相関関係を統計的に自動分析し,抽出された条件や決定木などの決定ルールをもとに,属性値が未付与の辞書エントリーの属性値を推測する手法である.前者は,専門家のノウハウが直接活用できるので,効率的で正確な属性値の付与が期待できる.しかし,この付与フローの作成は専門家による人手作業となるため,正確な付与フロー作成には人的・時間的コストがかかる点,誤りや不整合などが含まれる可能性がある点で問題がある.また第2の策は,付与済みの辞書エントリーに関するコーパス中の統計情報があれば,機械学習や統計処理により属性値の自動付与ルールが抽出できる点,統計情報から人間が気付かない現象も属性値との相関によりルールとして抽出できる点で有望である.しかし頻度が低いと属性や単語に対するルールの信頼性が低くなる点,属性値決定に有効な特徴量を注意深く選択しないと高い精度が得られない点で問題があった.

そこで,本稿では,この両者の手法の利点を活かしつつ欠点を克服した手法を考案するための手掛かりとして,この両者の手法で,日英機械翻訳用の構文意味辞書のエントリーに用言意味属性を付与させて,両者の認定精度を評価する.本検討により,用言意味属性を決定する上で有効条件を検討し,両手法の融合のための指針を得る.

[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
・N8(施設)で N2(動物)を 飼う. ＝>N8 raise N2
・N1(人 動物)が N2(食料 生物)を 食べる.
＝>N1 eat N2
[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,N8 等は結合価のラベル,括弧内は格の意味的制約)

[1] 同一の言語族内にある言語や歴史的,文化的に近い距離にある言語間の比較では,荒い語義分類になると予想され,それを,異なる言語族間の翻訳などに応用することは困難と考えられる.

2. 用言意味属性の分類[1]

用言意味属性の概要について述べる.本手法で採用する日本語の用言分類は以下の2つの観点から行った.

- **用言がもつ動的特性**

用言のもつ概念と談話場面に与える作用による分類

(1) 「持つ」 : <所有>
「開発する」 : <生成>

- **用言の格に対する関係**

用言の支配する格が用言に対してもつ役割による分類

(2)「完成する」→ N1 が 完成する:<N1 を生成する>
「開発する」→ N1 が N2 を 開発する :
<N1 が N2を生成する>

106種類に分類した用言の概念体系を図1に示す.

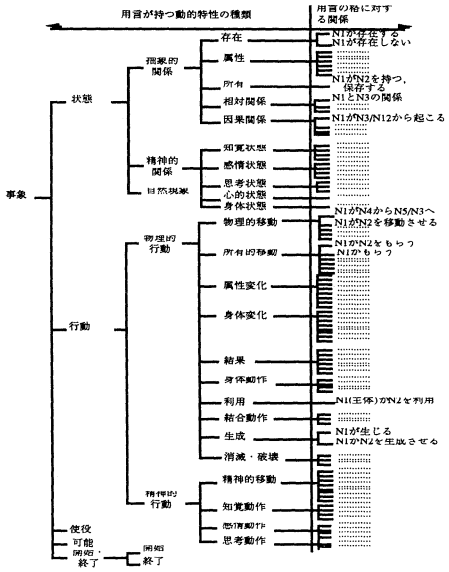

図2 用言意味属性体系

3．人手作成フローによる用言意味属性の付与

2章で説明した用言意味属性の個々の属性値は，もともと，専門家が頭の中で構文意味辞書の個々の辞書エントリーを分析し，意味を検討して決定して体系化したものである．よって，個々の属性値の意味やその決定条件は，専門家の頭の中に一番多くの情報が入っている．このため，属性に関して予備知識のない人でも辞書エントリーの情報を参照して容易に属性値を付与できるようにするには，この専門家の頭の中の情報を書き出して，既に属性付与済みの辞書情報と共に参照して決定するのが効率的であると考えられる．この属性値情報の書き出し形態としては，次の2種類が考えられる．

(1) 個々の属性値の定義や特長やその属性に所属する辞書エントリーの典型例を属性別に整理
(2) 専門家が属性値を付与する際に頭の中で活用する判断条件を判断木の形で付与フロー表として整理

前者は，個別の属性値について詳細に知りたい場合や数種類の属性値の候補から最適なものを1種類選ぶ場合には役立つ．しかし，属性の知識がない人が全属性値の候補から1種類選ぶ場合には適さない．これに対して後者は，属性の知識のない人でも判断木の各選択肢から当てはまるものを選択していけば最終的に属性値が決められる点で良い．しかし，人間が設定できる判断木の質問事項数は限られること，質問事項や選択肢が最適である保証が必ずしもあるとはいえないことなどから，付与された属性値が適切である保証がないという問題がある．

以上のように，2種類の手法には利点欠点が共存しているので，今回の検討ではこの2種類両方の形で情報を整理した．具体的には，(1)に相当する格属性の定義の説明に例を添付した表と，(2)に相当する属性値付与の専門家による図3の様な属性値付与フローを作成した．このフローは全てで39種類の質問項目からなり，各質問に答えていけば属性値が決まる仕組みになっている．

本付与フローを活用して用言意味属性を付与する際には，以下の手順で行うことを想定している．

[step 1] 日本語用言の品詞を判断．
[step 2] パターン全体を見て日本語用言の意味を判断．
[step 3] 付与フローに沿って選び属性値を決定．
[step 4] 適当な選択肢が無い場合には，国語辞典や英和辞典等を利用して，言い換え表現を検討．または，同じ日本語用言で別のパターン対に既に付与されている属性値を参考にして決定．

4．決定木自動学習による用言意味属性の付与

本章では，既に属性値が付与済みの辞書情報を用いて，辞書エントリー中の特徴と属性値との相関関係を統計的に自動分析し，抽出された決定ルールをもとに属性値が未付与の辞書エントリーの属性値を推測する手法について説明する．属性値の決定ルールを獲得手段としては，Quinlanが提案した決定木学習アルゴリズム ID3[4]をベースにして作成された決定木学習プログラム C5.0[5]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の条件に活用す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．

4．1　使用する特徴量

決定木学習アルゴリズムでは，特徴量と属性値の対の情報をもとに決定木を自動生成する．よって，生成された決定木の性能は，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．

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，2章で概説した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．よって属性値決定条件としては，パターンに含まれる格要素の種類が有望である．ALT-J/E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), N3(対象2;主に二格)のようにラベル付けされ，用言意味属性も，例えば，<N1 と N3 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，日本語と英語の等価表現の対の形で構成されているため，日本語では多義が有る場合でも，英語との対ではその多義が解消できる場合が多い[2]．よって，日本語側の意味的多義性の解消に英語パターン中の英訳語の種類の利用が有望である．これは，図3のフローを作成する際に行った属性値とパターンの関係の分析でも，例えば，英語パターン中の英訳語に"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る傾向にあるということから分かる．

(c) 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(人/動物)の具体的身体動作>や<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，この格への意味制約も特徴量として活用する．

(d) 日本語用言の意味カテゴリ

用言意味属性は，体系中で図2の左側に相当する用言が持つ動的特性の部分を見ると，属性値が個々の用言部分の意

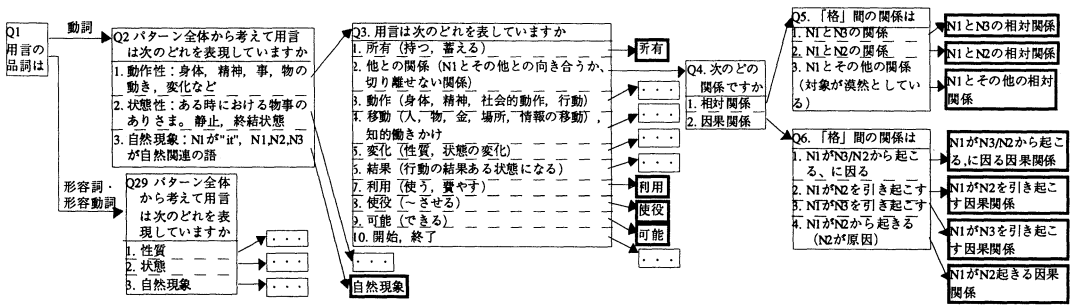

図3　属性値付与フロー（一部）

(太枠は，最終的に本フローにより決まる属性値，・・・はこの配下にまだ質問項目が続くとを示す)

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"や"運転する"のような<操作>の意味カテゴリのものが多い．よって，日本語用言そのものに付与されている意味カテゴリの種類も特徴量として活用する．

4．2　決定木学習手順

ここでは，日英構文意味辞書中の辞書エントリーから特徴量を抽出して，用言意味属性を決定するための決定木を作成する手順について説明する．手順は以下の通りである．

[step 1] 特徴量-属性値対のデータ抽出
図1の様な日英構文意味辞書の各辞書エントリーから4.1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，4．1(c)の格要素への意味的制約と(d)の意味カテゴリは意味を2718種類に分類し木構造で体系化したものである[2][3]．例えば，<操作>は意味カテゴリ上位概念として<仕事>-<労働>-<行為>-<人間活動>-<事>-<抽象>-<名詞>を持つ．よって，属性値を決定に有効なのは，この<操作>ではなくその上位概念かもしれない．このような問題を回避するために，意味カテゴリの特徴量は，Almuallimの手法[6]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した．

[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上記特徴量を決定木学習プログラムC5.0が受け取れるデータ形式に加工する．具体的には，全辞書エントリー中の特徴量を異なり別に集計し，設定した閾値回数以上出現した特徴量だけを，C5.0へ入力する特徴量として活用する．そして，その特徴量が個々の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のリストと属性値を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．

5．評価

3章で説明した人手作成属性値付与フローと，4章で説明した決定木学習プログラムC5.0による自動作成属性値付与決定木の性能を評価・比較するため，具体的な日英構文意味辞書の辞書エントリーを用いて，両者を評価した．

5．1　人手作成による属性値付与フローの評価

属性値付与の専門家でなくても本フローで容易に精度良く属性値が付与できるかを調べることにより，図3の属性値付与フローが適切に設計されているか評価する．

5．1．1　評価方法

評価は，以下の条件のもとで行った．

(a) 付与結果評価対象
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されたパターン対（P対）から無作為に選んだ100P対．この100P対中17P対には複数属性値が付与されていた．

(b) 属性値付与者（被験者）
パターン対作成の専門家だが用言意味属性の付与は行っていない者(A)と，本辞書とは全く関連のない者(B)の2名．この2名の結果を比較することで両者の熟練度の違いによる付与精度への影響を検討する．

(c) 評価で使用する資料
・属性値付与フロー（図3）
・属性値付与基準表
・付与対象の辞書パターン対リスト（100P対）
・辞典（国語辞典と英和辞典）

表1　属性値フロー表を用いた人手付与精度

被験者	付与精度	付与所要時間
A(P対熟練者)	78%	55分
B(P対未熟練者)	44%	83分

(d) 付与手順
被験者に用言意味属性，パターン対について簡単に事前説明した後，練習用に5P対に対して付与して作業手順を理解してから，本テストとして属性値を削除した100P対に属性値を付与した．1P対に付与する属性値の数に制限は設けなかった．

(e) 正解属性値
100Pに事前に付与されている属性値を正解とした．

5．1．2　評価結果

結果を表1に示す．この通り，精度，所要時間ともP対の作成になれているAの方が良い結果となった．よって，本属性値付与フローを用いた属性値の付与方法だけでは，その辞書内容の理解度に関係なく誰でも高い精度で付与できるとは言えない．しかし，用言意味属性の付与に関しては専門家でないAでも78%という精度で専門家と同じ属性値を付与することができたので，P対作成担当者が属性値付与を支援するための情報としては有効である．

この2名による属性値の付与結果を詳細に検討してみると，A，B両者の属性値の付与結果が一致したP対が，22%あった．一致したP対に付与された属性値は<属性><物理的移動><身体動作><物理的移動>などが多かった．よって，これらの属性値は，付与フローの記述が適切に行われているか，意味が捉えやすい属性であると言える．また，専門家の付与結果とは異なる属性値を付けた場合について詳細に分析すると，これらは次の5種類に分類できる．

(a) 意味の取り違え　　：A9件 B32件
AとBで大きく差が出た．これは，両者間のパターン対に対する知識の差によると推測される．よって，非専門家向きに辞書内容に関する資料や支援ツールの整備が必要．

(b) フロー表の不備　　：A4件 B11件
これは，作業時に用いた資料では属性決定の情報が不足することによる．よって，この誤りの原因を分析し，不足する用例や説明を追加していくことが必要．

(c) パターン対の情報不足　　：A3件 B3件
これは，格への意味制約がどの意味でもマッチすると条件を記述しているパターンへの属性付与の際に起こっている．よって，このパターン自体の意味制約の見直しが必要．

(d) 用言の格に対する関係の認定誤り　：A5件，B6件
図1の左側の用言がもつ動的特性の属性値は正しいが，図1の右側の用言の格に対する関係の属性値の選択で間違えた場合である．これは，格要素レベルでの個々の属性値の違いが分かるように，資料や支援ツールの改良が必要．

5．2　決定木自動学習による属性値付与の評価

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により属性値を付与するための決定木を自動学習し，その決定木を用いて実際に辞書エントリーに属性値を付与した際の精度を評価する．

5．2．1　評価方法

評価は，以下の条件のもとで行った．

(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，4．1で述べた特徴量の内，(a)格の種類は特徴量として常に使用し，(b)英語訳，(d)用言の意味カテゴリ及び(c)格への意味制約(N1格のみのパターンと，N1格とN2格のみのパターンの2種類)

は使用する場合としない場合で学習を行う（表2）．これにより特徴量の種類による認定精度の変化を分析する．

なお，各条件で学習対象となる辞書エントリーは，使用する特徴量情報が実際に含まれているものとした．例えば，格要素への意味制約を使用する場合には，N1 格のみからなる 4446 パターンと，N1 格と N2 格のみからなる 4260 パターンを学習対象とした．また，複数の属性値が付与された辞書エントリーは学習対象に加えなかった．

(b) 決定木学習プログラムの走行条件

決定木学習は 4 種類の学習データ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．また，4．2 [step 3]での特徴量の閾値回数による限定は，(b)英語訳と(d)用言の意味カテゴリで変化させその影響を調査する．

(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．これは，人手作成付与フローが専門家が全ての辞書エントリーの情報を検討対象に作成しているため，これと等価な実験条件とするためである．

(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．

5．2．2 評価結果

まず，使用する特徴量の種類に対する精度の評価結果を表3に示す（複数属性付与P対は加えず，閾値回数は1に統一）．これによると，(b)英語訳と(d)用言の意味カテゴリを用いた場合が最も高い値となった．また，付与対象が異なるので単純な比較はできないが，(b)，(d)に加え，(c)格への意味制約を利用した場合でも，(b)と(d)の場合より改善されなかった．これは，(c)が格要素の種類別に記述されており，適切な規則を獲得できるほど十分な量のデータが得られないこと，人手付与フローの評価と同様にパターン対の情報不足があることによるものと考えられる．

次に，使用する特徴量を表3の結果で最良であった(a)格の種類，(b)英語訳と(d)用言の意味カテゴリを使用した場合に固定し，閾値回数を(b)と(d)で変化させて，その付与精度を評価した（表4）．これによると，(b)=2，(d)=2まで閾値を増加させ入力データの特徴次元数を約 6 割に減らしても，付与精度は 70.4%と閾値を設定しない場合と同じであった．また，(b)の閾値を増やすと入力データの特徴時減数も決定木作成時間も大幅に減少する(作成時間は(b): (d) を 2:5 から 10:5 にすると 1/8 に減少)するが，付与精度は，(b)か(d)を使用しない場合に比べるとよりよい精度が得られている((b)未使用:58.8%，(d)未使用:61.0%に対し，閾値(b):(d)=10:5:65.4%）．以上から，(b)と(d)は使用して実際に使える計算機パワーに応じて，閾値を設定することが効率的で効果的な決定木作成手段である．

5．3 人手作成付与フローと決定木の比較評価

5．1と5．2の付与精度を比較し，両者の優位点，問題点を分析する．人手フローによる付与は人手作業を伴うにも関わらず，パターン対作成に従事していない一般の人間による作業では，44%の精度しか達成していない．これに対し，属性値付与済みの辞書情報を活用して決定木を自動作成する手法では，人手の作業なしでも最高 70.4%の精度を達成することができた．このことから完全自動での属性値の付与を考えると，この決定木学習プログラムを活用した手法の方が有効であるといえる．しかし，人手フローでもパターン対作成専門家による作業では，より高い精度を達成することができた(78%)．これは，付与フロー自体には，属性値付与専門家による属性値を決定する上での必須条件となる特徴量が記述されており，これをうまく活用すれば属性値付与に有効に働くことによると考えられる．

表2 決定木学習で用いる特徴量・辞書エントリー

使用する特徴量				学習対象辞書エントリー数
(a)格Pの種類	(b)英語訳	(c)格への意味制約	(d)用言の意味カテゴリ	
使用	未使用	未使用	使用	12601
	未使用	未使用	未使用	12601
	使用	未使用	使用	12601
	使用	使用(N1)	使用	3748
	使用	使用(N1,N2)	使用	3130

表3 使用する特徴量に対する決定木付与精度

使用する特徴量			入力特徴量次元数	作成決定木分岐数	付与精度
(b)英語訳	(c)格意味制約	(d)用言			
未使用	未使用	使用	1531	1252	59.8%
使用	未使用	未使用	4677	1705	61.0%
使用	未使用	使用	6190	2004	70.4%
使用	使用(N1)	使用	3748	1171	68.9%
使用	使用(N1,N2)	使用	3130	993	68.3%

表4 特徴量の閾値回数に対する決定木付与精度

閾値回数		入力特徴量次元数	作成決定木分岐数	付与精度
(b)英語訳	(d)用言			
1	1	6190	2004	70.4%
2	1	4021	2004	70.4%
2	2	3847	2004	70.4%
2	5	3603	1990	70.3%
5	1	2240	1783	67.7%
5	5	2022	1794	67.8%
10	5	1460	1640	65.4 %

6．まとめ

本稿では，機械翻訳システムの日英構造変換辞書に付与する用言意味属性を自動的に精度良く付与する手法の実現を目指して，専門家が人手で作成した属性値付与フローによる属性値付与方法と，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による属性値付与方法を提案し，その両者を性能評価した．本検討により，この2種類の利点を融合すれば，より正確な属性値付与処理系が実現できる見通しを得た．

今後は，この両者の手法をより厳密に評価するとともに，具体的に両手法を統合する技術について検討していきたい．また，属性値の付与支援という観点から人間の介入を前提としたよりよい付与環境についても検討していきたい．

参考文献

[1] 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，情報処理学会論文誌，Vol.38 No.2, pp.215-225 (1997).

[2] 池原,宮崎,横尾，日英機械翻訳のための意味解析辞書:電子情報通信学会言語理解とコミュニケーション研究会, Vol.NLC91 No.19 (1991).

[3] 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).

[4] Quinlan, J. R. : Induction of Decision Trees. Machine Learning, Machine Learning, Vol.1, No.1, pp.81-106 (1986).

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

[6] 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).