3P−5

解析事例を用いた係り受け曖昧性の解消

ＮＴＴ情報通信網研究所

武石 英二

1．はじめに

柔軟な自然言語処理システムを実現するためには、対象とする文章から言語知識、解析知識を取得し、文章の固有性や特徴をシステムに反映するための枠組みが必要だと考えられる。

そのため、機械翻訳における変換規則の獲得や意味解析における意味表現の獲得の枠組みとして、解析・変換結果を蓄積し、それらを直接利用する手法が提案されている[1),2)]。一方、単語間の共起関係や係り受け関係から解析ルールを取得する手法も検討されている[3)]。しかし、前者の枠組みを統語構造の解析に適用したものは少なく、後者は係りと受けの２項間関係や動詞を中心とした構造を対象とするものが多いため、文全体の構造的な特徴を取得することは難しいと考えられる。

このような背景のもとで、解析例を用いた係り受け解析の検討を進めている。これは、係りと受けの２項間の関係を解析対象とし、その前後の文構造を解析対象をとりまく環境（単語の共起や文構造の固有性）として捉えようとするものである。

2．解析例を用いた係り受け解析の構成

係り受け解析方式の全体構成を図1に示す。これは、事例に基づく手法の一般モデル[4)]をもとにしたものである。点線内が実際に解析を行う部分であり、点線外が解析の成否をフィードバックする部分である。事例に基づく手法の精度は、事例データの固有性に依存する部分が大きいため、処理結果をフィードバックする機構が必須だと考えられる。本稿では、実解析部分のうちの［　］部分について述べる。

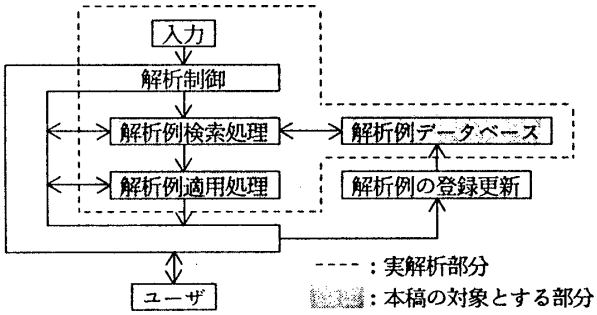

図1．解析例を用いた係り受け解析の構成

3．解析例の表現方法

入力となる文節列とその解析結果を対にして保存しておくだけでは、適用することのできる文が大幅に限定される。したがって、解析例そのものを一般化したり、解析例間を組織化することによって、適用することのできる範囲の拡大を図る必要がある。

本検討では、係り受けネットワークによる解析例の表現を採用した。解析例を、文節中の自立語をノード、主として付属語によって表現される係り受け関係をリンクとしたネットワークとしてとらえ、すべての解析例に対する係り受けネットワークを組み合わせたものを解析例全体と考える。また、正しい係り受け関係だけでなく、誤った係り受け関係もリンクとして保存する（実際には、既存の係り受け解析[5)]から出力される曖昧性を含んだ係り受け構造の正誤を人手によって判断して用いた）。以下では、正しい係り受け関係を示すリンクをリンクタイプＡ、誤った係り受け関係を示すリンクをリンクタイプＢと呼ぶ（図2）。

リンクℓが始点ノードｎ、終点ノードｎ'、係り受け関係ｋ、リンクタイプｔ、リンクの出現する解析例の全体｛ｓ｝から構成されると考えると、解析例の全体Ｘは、リンクの全体Ｘ＝｛ℓ｝、ℓ＝(ｎ, ｎ', ｋ, ｔ, ｛ｓ｝)とみることができる。

以下では、リンクℓの始点ノード、終点ノード、係り受け関係、リンクタイプ、リンクの出現文の全体を、それぞれｎ(ℓ)、ｎ'(ℓ)、ｋ(ℓ)、ｔ(ℓ)［＝ＡまたはＢ］、｛ｓ｝1と表す。

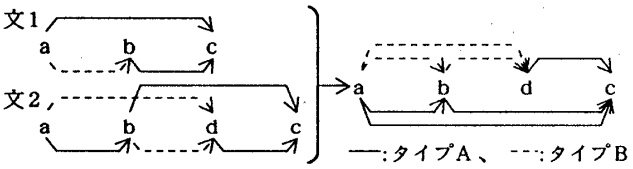

図2．解析例の表現方法

解析例全体をネットワークとして表現することによって、
- 着目する係り受けをもつ解析例を容易に取得することができる、
- 共通の部分構造をもつ解析例を容易に取得することができる、
- それによって、類似した解析例の中から適切な解析例を選択することができる、

という効果を期待することができる。また、正しい係り受け関係だけでなく、誤った係り受け関係も蓄積することによって、
- 不適切な解析例の適用を抑止することができる、
- 誤った係り受け関係を積極的に適用することによって、係り先の限定、特定を行うことができる、

という効果を期待することができる。

4．解析例による係り受け曖昧性の解消効果

３章で述べた解析例の表現方法を用いて、係り受け曖昧性の解消法について検討し、評価実験によって有効性の検証を行った。概要は次の通りである。

既存の係り受け解析の結果として、文節とその係り先文節（一般に複数個）のリストが出力される。係り先文節が複数個出力された文節に対して、その文節および係り先候補の文節をキーとして蓄積した解析例データからもっとも適切だと考えられる解析例を選択し、解析例中の係り受け関係を用いて係り先文節の特定を行う。また、曖昧性の解消は、文頭文節から順に１文節ずつ行う。

4．1．ノード間、リンク間、文間の類似度

（1）ノード間類似度

既存の解析処理に使用している日本語辞書[6)]には、単語の意味的な属性（意味カテゴリ：約２８００分類）が記述されている。この意味カテゴリを４０に粗く再分類したカテゴリを用いてノード間の類似度を求める。一般に１つのノードｎには、複数のカテゴリが付与されており、これをＣ(ｎ)で表す。２つのノードn_1とn_2の類似度Ｆを、ノードカテゴリの一致度によって次のように求める。

$$F(n_1, n_2) = \frac{\#\{C(n_1) \cap C(n_2)\} + \alpha f(n_1, n_2)}{\max\{\#C(n_1), \#C(n_2)\} + \alpha}$$

ここで、αは定数、#Ａは集合Ａの個数を示しており、$f(n_1, n_2)$はn_1とn_2の表記が一致する場合に１、一致しない場合に０をとる。

（2）リンク間類似度

２つのリンクℓ_1、ℓ_2の類似度Ｇを、次のように求める。

$$G(\ell_1, \ell_2) = \begin{cases} \dfrac{F(n(\ell_1), n(\ell_2)) + F(n'(\ell_1), n'(\ell_2))}{2} & [\ell_1, \ell_2\text{のリンクタイプ、係り受け関係が一致する場合}] \\ 0 & [\text{その他の場合}] \end{cases}$$

（3）文間類似度

処理対象文のノードの並びを$S = (n_1, \cdots, n_p)$、解析例を構成する文節に対応するノードの並びを$S' = (n'_1, \cdots n'_q)$とするとき、ＳとＳ'の文間類似度Ｈを次のように求める。

Resolution of Structural Ambiguity Using Example Data
Eiji TAKEISHI
NTT Network Information Systems Laboratories

$$H(S, S') = \frac{\sum_{i=1}^{p} \max_{j=1}^{q} \{F(n_i, n'_j)\}}{p}$$

4．2．解析例の選択、適用

処理対象文のノードの並びを$S=(n_1, \cdots, n_p)$、曖昧性解消の対象ノードをn_m、その係り先候補のノードをn_{s1}、$\cdots n_{sr}$、n_mからn_{s1}への係り受けを$\ell_{m,s1}$、その係り受け関係をkで表す。

（1）解析例の選択

「$F(n_m, n(\ell)) > \beta_1$、$F(n'_{si}, n'(\ell)) > \beta_2$なる$n'_{si}$ $[1 \leqq i \leqq s_r]$ が存在する、$k(\ell)=k$」…(※)

を満たすリンクℓをもつ解析例を適用候補として選択する。β_1、β_2は定数（しきい値）である。

（2）解析例の適用

選択された解析例を次の手順で適用する。

①選択された解析例のうち、処理対象文との文間類似度がもっとも大きい解析例を選択する。

②解析例内に(※)を満たすリンクが複数個（$\ell_1, \cdots, \ell_t$）存在する場合には、

$$G(\ell_{m,s\cdots}, \ell_v) = \max_{j=1}^{t} \max_{i=1}^{r} G(\ell_{m,si}, \ell_j)$$

を与えるリンクℓ_vをすべて選択する。

③ℓ_vのリンクタイプによって、n_mから$n_{s\cdots}$への係り受けの成否を決定する。

$t(\ell_v)=$Aのとき、$n_m \rightarrow n_{s\cdots}$：成立

$t(\ell_v)=$Bのとき、$n_m \rightarrow n_{s\cdots}$：不成立

4．3．評価実験

（1）実験対象データの概要

2種類の計算機マニュアル（1、2）を用いて曖昧性解消の評価実験を行った。マニュアル1の1章分（計算機の構成説明の部分）を解析対象文とした。解析対象文の概要を表1に示す。表1において多義数とは、曖昧性のある箇所の総数を示しており、候補2、3、4とは、それぞれ係り先の候補数が2、3、4である箇所数を示している。また、マニュアル2の対応する章を解析例データX、マニュアル2の他の1章分を解析例データYとし、解析例データX、X＋Yの2種類の解析例データを用いて実験を行った。解析例データの概要を表2に示す。表2において、各数値は個数を表しており、A、Bとはリンクにおけるタイプ A、Bの内数である。

表1．解析対象文の概要

総文数	多義数	候補2	候補3	候補4
67	95	78	12	5

表2．解析例データの概要

データ	総文数	ノード	リンク	A	B
X	67	322	635	516	119
X＋Y	133	560	1194	940	254

（2）実験内容

表3に示す3種類の実験を行った。×は解析例の間の順位付けを行わないことを示している。また、前節、前々節で導入した定数の値として、$\alpha=3$、$\beta_1=\beta_2=0$を用いた。

表3．評価実験内容

	解析例選択方法	解析例適用方法
実験1	表記一致	×
実験2	ノード間類似度F	×
実験3	ノード間類似度F	文間類似度H

（3）実験結果

実験結果を図3に示す。太線がデータX＋Yに対する結果を示しており、◎は正しい係り先を認定した（適用成功）件数、丸数字はタイプBのリンクが適用された内数、破線は同一類似度をもつ複数の矛盾するリンクが選択されたため、係り先が一意に決定しなかった件数を示している。また、折れ線は適用成功率を示している。

$$(\text{適用成功率}) = \frac{(\text{適用成功件数})}{(\text{適用成功件数}+\text{適用失敗件数})} \times 100$$

解析例間の順位付けを導入した場合、適用成功率をあまり低下させず（Xの場合で81．8%から80．4%、X＋Yの場合で87．0%から79．4%）に、選択成功数がXで20から61、X＋Yで24から66へ改善されたことから、ノード間類似度、文間類似度が有効に機能していることがわかる。

解析例データYの追加による効果については、選択成功件数、適用成功件数に増加が見られることから、ある程度の有効性を確認することができる。しかし、着目する係り受け曖昧性とあまり関連がない解析例が雑音となって適用が失敗する場合があり、また解析例数の増加に比べ、選択成功数の増加はそれほど大きくない。

リンクタイプBが適用された件数も比較的多い（32%から50%）ことから、誤った係り受け関係を保存することが有効であることがわかる。

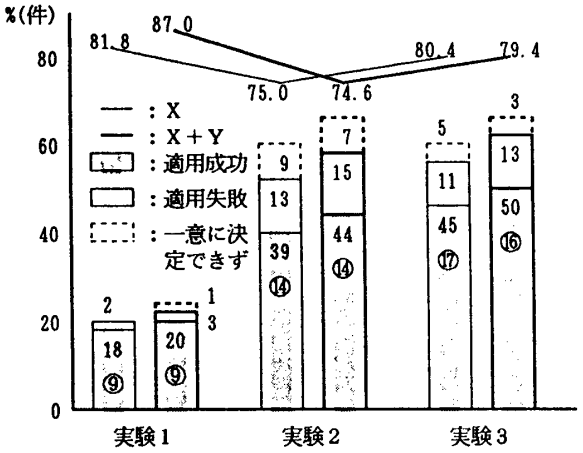

図3．評価実験結果

（4）考察

・解析例の選択、適用方法の改良

現在、粗分類して用いている意味カテゴリの利用方法の改良、リンクの出現頻度の導入によって、ノード間類似度の算出方法を精密化すること、解析例として保存された構造を利用することによって、文間類似度の算出方法を改善することが考えられる。

・解析の制御

どこから解析（曖昧性の解消）を行うかを決定する、解析例中の構造やすでに解析済みの部分構造を利用することによって、解析例の部分構造を適用したり複数の解析例を組み合わせて適用するなど、解析を制御する機構について検討する必要がある。

・解析知識の利用

現状では、たとえば「接続した機器」と「機器を接続する」の対応関係を認定できない、「AとBを利用する」に対して「Aを利用する」を認定できないなどの問題点がある。態やアスペクト、並列関係などの利用方法についても検討する必要がある。

5．おわりに

解析例を用いた係り受け解析の構成、解析例の表現方法および解析例を用いた係り受け曖昧性の解消効果について述べた。今後は、4章で考察した課題について検討を進めるとともに、2章で示した係り受け解析全体の構成についても具体化を進めていく予定である。

謝辞

本検討に対して数多くの貴重な助言を頂いたNTT情報通信網研究所、林良彦主任研究員に深く感謝致します。

参考文献

1）渡辺：文法的情報を含んだ翻訳例を用いるトランスファーシステム，情処第43回全国大会論文集，2H-4（1991）．

2）島津：コーパス解析に基づく事例ベースパーザ，情処人工知能研究会資料，75-10（1991）．

3）小松他：係り受け事例からの結合価フレーム獲得の試み，情処第43回全国大会論文集，3H-2（1991）．

4）小林：事例ベース推論の研究課題，情処人工知能研究会資料，75-4（1991）．

5）白井：日英翻訳システムALT－J／Eにおけるテーブル駆動型係り受け解析法，情処第34回全国大会論文集，5W-6（1987）．

6）池原他：言語における話者の認識と多段翻訳方式，情処論文誌，Vol.28，No.12（1987）．