確率付決定木による言語解析

柏岡 秀紀, 金城 由美子
ATR音声翻訳通信研究所

1 はじめに

現在、自然言語処理では、形態素、統語、意味などの各段階での個別の処理手法が提案され、高精度のシステムが構築されつつある。日本語の形態素解析では、品詞、単語の連接知識や辞書を整備することで、高精度に動作させているシステム(JUMAN、茶筌、すもも)があり、また、統語解析でも、新聞記事などを対象にした精度のよいシステム(KNP)が提案されている。これらのシステムでは、各段階の処理を独立に行なっているため、形態素解析に構文情報を利用したり、統語解析に意味情報を利用したりすることは困難である。各段階における解析から得られる情報を相互に利用することは、解析精度の向上に効果がある[1]と考えられ、実際に形態素・統語解析を統合したシステム[2, 3, 4]がいくつか報告されている。

本稿では、確率付き決定木[9]を用いた形態素・統語解析の手法について報告し、形態素解析、統語解析を同時に行なうことによる効果を実験を通じて議論する。さらに、単語の意味情報を利用した解析、および、形態素解析による単語の意味属性の付与についても検討する。

2 決定木による言語解析

ここでは、決定木を利用して形態素解析と統語解析を処理する手法について述べる。決定木は、さまざまな属性を利用して対象となる事象を分類するものである。

形態素解析は、単語分割と単語への品詞付与という二つの処理から成る。単語分割は、「対象文字列を“単語”か“単語でない”かの二つに分類する処理」として、単語への品詞付与は、「対象となる単語をどの品詞に分類するかの処理」として、とらえることができ、このような分類する処理に対して決定木の利用は効果的な実現法である。

同様に統語解析についても、「係り受けの範囲によって分類する処理」、「処理範囲に含まれる語の係り受け関係で分類する処理」ととらえることにより、決定木を利用した処理が効果的な実現法となる。

このように、言語解析を分類という処理としてとらえた場合、分類するための基準として、どのような情報を利用するかが問題となる。従来の解析処理では、形態素解析であれば、品詞の接続情報や辞書の情報、統語解析であれば、係り受け関係と語の頻度情報のような限定された情報に着目して処理されていることが多い。これは、複数の情報を同時に利用しようとした場合に、各情報を利用する優先度の設定が困難なためと考えられる。適用する規則の優先度や、利用する情報の優先度を設定・調整するために、統計的な手法[5, 6, 7]が提案されている。

本稿で取り上げている決定木による解析も統計的な手法の一つである。決定木の各分岐点では、分類の指標となる特定の属性に着目し、対象の属性値により分類される。この分類のための属性をどのような順序で利用するかは、一定量の学習データにより、自動的に決めることができる。属性としては、多様なレベルの属性を利用することができ、従来の手法で問題になるさまざまな情報をうまく取り込める可能性がある。これらの属性を、本稿では“語法の性質”と呼ぶ。実際に、形態素解析では、字種、語の構成、語の接続関係、品詞の接続関係などの属性を利用した処理を行なっている。

3 形態素解析と統語解析

自然言語処理では、まず最初に形態素解析が行なわれ、続いて統語解析、意味解析を行ない、その結果を利用することで、様々な応用処理が実現される。

その第一ステップともいえる形態素解析は、2節でも述べたように、単語分割および品詞付与という二つの分類処理の組合せとして考えられる。

決定木を利用している本システムでは、それに対応する決定木を作成して処理しているが、品詞付与においては、2段階の分類処理として実現している。そのため、以下のような3種類の決定木を作成している。

1. 単語分割のための決定木
   対象とする文字列が単語であるか否かを判断する。

2. 品詞（大分類）を付与するための決定木
   名詞、動詞などの大まかな品詞分類を行なう。

3. 品詞（詳細分類）を付与するための決定木
   大分類で名詞であれば、普通名詞、固有名詞など、動詞であれば、活用形、活用型などの詳細な分類を行なう。また、意味的な属性を単語に付与する。

また、統語解析では、以下の2種類の決定木を作成し処理している。

1. 規則適用のための決定木
   処理する対象範囲に規則を適用するか否かを判断する。

2. 処理範囲決定のための決定木
   現在の対象範囲を広げて処理するか、新たな範囲を対象とするかを判断する。

以上の5種類の決定木を利用して、解析処理を進める。どの処理も、決定木を利用した、多様な情報から判断されるものであり、各決定木から得られる出力に対する信頼性は、相互に考慮しやすいものとなっている。実際には、スタックデコーダアルゴリズム [10] を利用して、5種類の決定木の利用回数に応じて比較する対象を限定し、より適正な判断が行なわれるようにしている。この5種類の決定木を利用する順序により、形態素解析のみを予め行なってから統語解析を行なう処理や、形態素解析と統語解析を同時に進める処理が可能となる。

4 実験

これまでに述べた決定木を利用する言語解析手法によるシステムにおいて、形態素解析、統語解析の処理の順序、また処理順序により利用できる情報の差による解析精度を比較するために、以下の実験を行なった。各解析処理では、決定木学習に利用できる“語法の性質”は同じとした。

- 形態素解析を行ない、その結果を利用した統語解析
  - 形態素解析結果のベストNを利用した統語解析
  - 人手による形態素解析結果を利用した統語解析
- 形態素解析と統語解析を可能な限り同時に処理した解析

対象としたテキストは、ATR音声翻訳通信研究所で収録した対話データで、翻訳システムのための品詞体系と、その品詞に基づき、対話データをカバーするように構築された文法を利用した [8]。決定木の学習に3000文、スムージングに3000文のデータを利用し、評価用には、1000文のテキストを利用した。

表1に、その結果を示す。

“単語分割”、“品詞情報”は、統語解析が出力した形態素の中で、正解形態素[1]と一致する割合を示している。また、“クロスブラケット”は、統語解析が出力したブラケットの中で、正解の統語構造の持つブラケットと交差するブラケットの割合、“完全一致”は、統語解析が出力したブラケットの中で、正解の統語構造のブラケットと利用している規則名まで完全に一致するブラケットの割合である。

“クロスのない文”は、統語解析が出力した解析結果において、文単位で正解の統語構造と交差するブラケットがない出力文の割合、括弧内は、入力文数に対する割合であり、“完全一致の文”は、文単位で統語解析の出力が正解の統語構造と完全に一致する割合、括弧内は、入力文数に対する割合である。“出力率”は、入力した文の数に対して、統語解析が解析結果を出力できた文の割合である。

現在のシステムでは、統語解析の際に、解析結果を出力しない入力文がある。これは、規則適用の決定木で利用している属性が不十分であり、決定木学習で適切な分類が行なわれていない状態で、正解候補の確率が閾値以下になってしまう場合である。実際に、学習データで、非常に稀な係り受け関係にあるものや、類似した係り受け関係が他に多数あるものが解析対象になる場合に、規則選択の閾値によって選ぶことができず、失敗している。

[1]正解形態素、正解の統語構造は、人手により形態素に分割され、統語構造を付与したデータである。

表 1: 解析結果の比較

	形態素解析を別途		形態素解析と同時
	正解形態素	複数の形態素解析	
単語分割	100	88.1	87.3
品詞情報	100	84.7	83.9
クロスブラケット	7.2	13.6	11.4
完全一致	87.0	71.0	73.4
クロスのない文	74.3 (64.8)	62.9 (56.0)	69.2 (54.0)
完全一致の文	66.3 (61.1)	48.5 (43.2)	53.8 (42.0)
出力率	92.1	89.1	78.0

5 意味属性の付与

本稿で述べた決定木を利用した処理では、特定の情報に重点をおいた処理と異なり、様々な情報を利用できる利点がある。そのため、意味属性を品詞に含めた詳細な品詞体系に対しても同様の処理で、形態素解析や統語解析を効果的に実現することができる。英語においては、本手法と同様の処理手法で、意味属性を考慮した非常に詳細な品詞体系での Tagger[12]、統語解析 [11] を実現している。

一般に、意味属性の付与を行なう場合、解析候補が多くなり、制限となる情報を効率的に利用することが困難であるため、絞り込みで失敗することが多いと考えられる。統語解析、形態素解析を同時に処理することにより、統語、形態素、意味の属性に関する制限を適宜取り込むことで、早期に適切な候補に絞り込める可能性がある。また、意味属性を付与することにより、様々な分野への応用が考えられる。

意味属性の付与に際して、他の解析に有効な情報を与えることを考慮して、現在のところ、以下のような属性を設定している。体言（主に名詞）の分類として、｛人、場所、組織、物、時｝の5種類の属性値を、また、用言を分類するために、6種類の属性値｛動作主、引用、共同、対象、経験者、目的｝を設定し、その組合せにより分類した。分類するに辺り、ATR で収録した対話データに照らし合わせることで、実際に利用されている表現を分類できるように考慮した。この分類により、用言と体言の連接（正確な共起）をとることができ、翻訳の精度や、解析の精度を向上させることが可能と思われる。

これらの意味属性を利用した処理を実現するために、体言（主に名詞）については、辞書にどのような属性値を持つかを記述し、用言についても、分類したテーブルを作成し、辞書的に利用することで、形態素解析、統語解析への影響を調べる予定である。これにより、解析処理の精度向上を目指し、効果的に利用できる意味属性を検討する。一方、意味属性の値を付与するための属性として、現状の形態素解析、統語解析から得られる情報を調べ、相補的に解析精度の向上を目指す。

6 考察

第4節で示した実験から形態素解析と統語解析を同時に処理することによる統語解析への効果を見ることができる。実際に、クロスブラケットの割合が減少し、完全一致の割合は増加しており、同時に処理することによる精度向上が見られる。ただし、出力率を考慮した場合に、この差異を効果としてとらえて良いかには、疑問が残る。また、形態素解析においては、同時に処理することによって、わずかではあるが精度が悪くなっているように見受けられるが、これも単純に比較できない問題であろう。表1に示した精度は、統語解析の出力があった場合のみを対象として精度を出しているため、構造的には誤っていても形態素解析は(文単位では誤りを含むが)、成功している場合が多くあり、その結果、形態素解析のみを行なう方が精度が高くなっているものと思われる。

これらの実験では、条件として、決定木学習に利用できる“語法の性質”は同じとしている。統語解析と形態素解析を同時に行なう場合は、形態素解析に統語情報を利用することができるのがメ

リットであるが、現在それに相当する“語法の性質”が整備できていない。本実験の形態素解析では明確な差異が見受けられないが、形態素解析に統語情報に関する“語法の性質”を組み込むことにより、同時に処理する場合の精度向上を見込むことができる。

また、本実験では、辞書を利用せずに形態素解析から統語解析までを行なっている。このような処理機構で、第5節で述べた意味属性を(辞書を必要とせずに)付与することができれば、未知語に対して意味的に類似する語を割り振ることで、解析精度の向上を期待することができる。また、検索システムなどにおいても、未知語を意味的に類似する語に置き換えることで、様々な処理に対応できると考えられる。

7 まとめ

本稿では、決定木を利用した言語解析手法について報告した。本手法では、形態素解析、統語解析を同時に処理することができ、同時に処理した際の効果について、実験を通じて考察した。また、意味属性を含む解析について検討し、実際に意味属性による単語分類を行なっている。

今後は、各処理に有効に働く“語法の性質”を整備するとともに、これらの意味属性による分類を利用することによる各解析処理の精度向上、および、各処理からの情報を利用した意味属性の付与の精度向上を目指すとともに、学習データ量と精度の関係についても、明確にしていきたい。

参考文献

[1] 田中穂積, 竹澤寿幸, 衛藤純司: “MSLR法を考慮した音声認識用日本語文法-LR表工学(3)-” 情報処理学会音声言語情報処理研究会, No. 15-25, pp. 145-150, 1997.

[2] 高橋, 柴山, 宮崎: “オブジェクト指向パーザPowerにおける構文的曖昧さの漸進的解消機構”, 言語処理学会第3回年次大会発表論文集, pp.197-200, 1997.

[3] 綾部, 徳永, 田中: “複数の接続制約のLR表への組み込みとそれによる解析の統合化”, 言語処理学会第3回年次大会発表論文集, pp.201-204, 1997.

[4] 植木, 徳永, 田中: “日本語解析システムMSLRの効率化に関する研究”, 言語処理学会第3回年次大会発表論文集, pp.209-212, 1997.

[5] Black,Jelinek,Lafferty,Magerman,Mercer, Roukos: “Towards history–based grammars: Using richer models for probabilistic parsing”, In Proceedings of the 31th Annual Meeting of the ACL, pp.31–37, 1993.

[6] Sekine,Grishman: “A corpus–based probabilistic grammar with only tow non–terminals” In proceedingsof the International Workshop on Parsing Technologies '95, 1995

[7] 白井, 乾, 徳永, 田中: “統計的日本語文解析における種々の統計量の扱いについて” 自然言語処理シンポジュウム'96「大規模資源と自然言語処理」, 1996.

[8] 河田, 金城, 柏岡: “日本語会話文の構文木付コーパス作成” 言語処理学会第4回年次大会発表論文集, 1998.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone: *“Classification and Regression Trees”*. Wadsworth & Brooks/Cole, Monterey, CA. 1984.

[10] F. Jelinek: “A fast sequential decoding algorithm using a stack” *IBM Journal of Research and Development*, 13:675–685. 1969.

[11] 柏岡,Black,Eubank: “詳細な文法を用いた統計的構文解析法” 情報処理学会第55回全国大会 講演論文集(分冊2), pp.356–357,(1997)

[12] Black,Eubank,Kashioka,Magerman,Saia,Ushioda: “Reinventing Part–Of–Speech Tagging” 自然言語処理 Vol.5, No.1, pp.3–23,(1998)