確率付決定木を用いた日本語構文解析

柏岡 秀紀 , 河田 康裕 , 金城 由美子 , Andrew Finch , Ezra Black
ATR音声翻訳通信研究所

1 はじめに

自然言語処理の解析は、形態素、構文、意味の大きな3つのレベルに分割することができ、個々の処理に関して研究が進められることが多い。構文解析の入力には、形態素解析済のデータが期待されることが多いが、構文レベルでの情報を利用しなければ、形態素レベルでは、判断できない状況もある。そのために、各処理を統合した処理機構が提案されている [1, 2, 3]。

各レベルの処理に統計情報を利用した研究も盛んに行われている。統計的言語モデルとして、n-gramや確率文脈自由文法を利用した研究が多い。これらのモデルは、学習が容易であり、比較的パラメータが少なく調整できる利点がある。しかし、n-gramでは、隣接した局所的な情報に着目しているため、離れた部分の情報を有効に活用することができない[1]。また、確率文脈自由文法では、文法に示された固定の確率値により構文構造を判断するために、統語的には同じ語の並びであっても、語の意味的な差異による構造の差を処理することが困難である[2]。本稿では、決定木の枠組みを統計的言語モデルとして利用することで、隣接した情報に加えて、離れた部分の特徴や語の特徴などの多様な語法の特徴を用いた処理機構を提案する。

これまでに、統計的言語モデルとして決定木モデルを利用することで辞書を必要としない形態素解析を提案しており、構文解析機構を同様のモデルを用いて実現することにより、柔軟に、形態素および構文の各処理を統合した処理機構を実現できる。以下、利用している知識として、文法、語法の性質について述べ、解析機構について説明し、これまでに行った予備実験と英語に対する実験との比較を通じて、本稿での解析機構について議論する。

[1]一定距離をおいたbigramの利用なども提案されているが、柔軟に離れた距離での特徴を利用することは困難である。

[2]幾つかの解決法 [4, 5, 6] が提案されている

2 利用する知識

一般に日本語の構文解析 [7] では、係り受け関係を捉えることが多く、本稿では、素性構造付文脈自由文法で記述している。また、決定木で利用される特徴として、語を構成している部分的な文字列の特徴や語あるいは品詞等の接続情報などを利用している。

2.1 文法

本稿では、16種類の素性を用いた素性構造付文脈自由文法を利用している。素性としては、カテゴリ(cat)、名詞の型(n_type)、動詞の型(v_type)などがあり、品詞タグは、基本的なカテゴリ、名詞や動詞の型、活用の種類や活用形の組み合わせにより表現され(表1)、約200種類におよぶ。文法は、この品詞タグをベースに、約100種類の統語規則から構成される(表2)。

表1: 品詞タグの例

RULE:	格助詞
EXAMPLE:	に, を, が, で, と, から, まで, へ, について, の, によって, って, より, として, にて
PARENT:	cat p bar 0 pos 助詞 p_type 格

また、この文法は、対話データでの表現を捉えられるように、考慮されており、いわゆる通常の文法という意味では、非文扱いされるものに対しても、処理できるように考慮されている [8]。

2.2 語法の性質(特徴)

前節で述べた文法の制約のみでは、形態素解析済の入力に対する構文構造だけでも、多くの解析候補がある。さらに、形態素解析の候補を考慮すると、膨大な量の候補が存在することになる。このような候補から正しい構造を選択するには、処理の過程で適切な選択をしていく必要があり、その選択を行うために、様々な特徴を利用する。

表 2: 文法規則の例

RULE:	VP_np+ap
DESCRIPTION:	VP_np+ap → NP AP
EXAMPLE:	[vp [np あなたさえ] [ap よければ]]
PARENT:	cat v bar 2 pos vp_np+ap v_type normal
CHILD1:	cat n bar 2 pos np_np+ 副 n_type V*
CHILD2:	cat CAT bar 2 pos V* n_type V* v_type V* a_type A katuyou V* katuyoukei V*
CONDITION:	not((A=unspec)|(CAT=s))

利用している特徴には、1) 語 (あるいは句、節) に関する特徴、と 2)(文内の) 文脈に関する特徴があり、形態素、構文のレベルで統計的な尤度にしたがって有効に活用される。

語に関する特徴

単語自身の持つ特徴であり、その部分的な綴り、文字数、構成文字種等に関する特徴や、品詞タグが付与された後には、品詞タグの持つ各素性の値が特徴として利用される[3]。単語自身が持つ語彙の特徴は、単語に対するタグを決めるのに非常に有効な情報であるとともに、構文構造のある範囲での主辞となる語の情報としても利用される。本稿で扱う語彙の情報は、見出し語として辞書より得られる情報を利用することもできるが、基本的には、先に述べたような語を構成している部分的な文字列や文字数等により特徴づけられているためにいわゆる「未知語」という概念がない[4]。

文内文脈に関する特徴

現在の処理対象に関する特徴だけでなく、その直前の語の特徴や、接続する文字の特徴、あるいは、処理対象の 2 つ前の語や品詞についての特徴、文頭や、文末に関する特徴、処理対象の前で一番近くにある助詞の情報、そこまでの単語数などが特徴として利用される。このように、ある一定の定められた語の情報だけでなく、柔軟に距離が変化する特徴も利用することができる。また、今処理している文字列が、同一文内に現れているかどうか等も、特徴として利用できる。

これらの特長を、「語法の特徴」と呼び、記述するための枠組みを開発した。

3 確率付決定木による解析

すべての「語法の特徴」が、形態素、および構文解析で利用されるわけではない。「語法の特徴」は、学習用コーパスに現れる統計的な優位性を基準に、解析知識として効率的に利用するために決定木の枠組みの中で利用される。

本手法で用いる決定木は、文献 [9] で述べられている手法を一部変更したものである[5]。利用する決定木は、2 分木のものであり、各分岐点での判断に「語法の特徴」を利用する。だたし、2 分木であるため、3 個以上の値を持つ特徴を直接利用することができない。そこで、3 個以上の値を持つ語法の特徴を決定木の分岐点の情報として利用するために、特徴の各値に “0”、“1” からなる固有のビット列を与え、そのビット列内の特定のビットを一つの分岐点の情報として利用する。また、その特徴が有効な特徴であるかどうかの分岐も行っている[6]。

3.1 処理の流れ

形態素解析、構文解析を統合する処理機構を実現するために、以下のような二つの統合の仕方が考えられる。

処理 A 文末まで一旦、形態素解析を行い、各候補に対して順次構文解析を行う。

処理 B 左から右に文字単位で処理しながら形態素解析、構文解析を行う。

[3] 語ではなく句や節として文法規則でまとめられている場合には、その規則の親ノードの持つ素性の値が特徴として利用される。また、どのようなノードから構成されているか、句や節を構成している単語数なども特徴として利用される。

[4] 強いて、本手法で「未知語」を考える場合には、特徴を抽出した学習データに現れない語と捉えることができる。

[5] 枝がりに、minimal cost-complexity アルゴリズムを用い、スムージングには、Forward-Backword アルゴリズムを用いた。

[6] 特徴が有効かどうかは、その特徴を判断することができるかいなかによる。例えば、文頭の単語を処理する場合に、直前の単語に関する特徴は利用できないため、有効な特徴ではない。

処理Aでは、形態素解析時に構文情報を有効に利用できるという統合での利点がなく、処理Bでの統合が望まれる。比較のために、現状の解析機構では、二つの処理手法を切り替えて利用できるようにした。

本手法では、その統計的言語モデルとして決定木モデルを採用し、解析処理のために、以下の3種類の決定木を、学習データを用いて構築する。

単語認識の決定木 単語としての妥当性を判断するための決定木

品詞付与の決定木 品詞タグの候補を選択するための決定木、

文法規則適用の決定木 適用する構文規則を選択するための決定木

解析処理では、この3つの決定木を以下の3つの状態で利用し、各状態で処理が行われる。

1. 単語認識の状態
   現在の処理対象となっている文字列が単語として妥当かを判断する。妥当な場合、単語ノードとした状態を品詞付与の状態とするとともに、次の文字を取り込んだ文字列を処理対象とする単語認識の状態をつくり、各状態の処理を行う。妥当でないと判断された場合、次の文字を取り込んだ文字列を処理対象とする単語認識の状態をつくり、各状態の処理を行う。

2. 品詞付与の状態
   現在の処理対象となる単語ノードに適切な品詞タグを[7]、タグノードとする構文規則適用の状態とするとともに、次の一文字を処理対象とした単語認識の状態を作り、各々の状態について処理を行う。

3. 構文規則適用の状態
   現在の処理対象となるノードから前のノードを参照し、適用できる構文規則に対して、ルールノードを設定し、構文規則適用の状態とするとともに、次の一文字を処理対象とした単語認識の状態を作り、各々の状態について処理を行う。

単語認識の状態では、単語ノードの作成に対して、処理文字列が単語として妥当かどうかを判断する決定木が利用される。ここで利用される決定木は“単語認識の決定木”であり、単語として現れる確率値が計算され、その値により、妥当か否かの判断がなされる。品詞タグを付与する場合には、“品詞付与の決定木”が利用され、確率付で一定以上の値を持つ品詞タグ候補が与えられる。構文規則適用についても、文法的に適用できる規則に対して、“文法規則適用の決定木”により適用できる規則の内、一定以上の値を持つ規則に対して、ルールノードが設定される。また、複数の状態の処理については、スタックデコーダアルゴリズム[10]を利用することで、処理の効率化をはかっている。

[7]複数の品詞候補がある場合は、その候補分の状態をつくり、処理を行う。

4 実験

本手法での実験を行うために、上記の処理機構を構築し、ATR音声翻訳通信研究所で収集した旅行会話について学習データを作成中である。現在のところ、利用できる学習データが少なく、また、2.2節で述べた特徴についても、構文規則適用の決定木に有効な特徴の整備が充分でないために、形態素解析と構文解析を統合したシステムとしての評価はできていないが、形態素解析を単独で動作させた場合の実験結果と、英語での品詞付与と構文解析を統合したシステムでの実験結果について以下に述べる。

日本語形態素解析

ATRで収集した旅行会話において、2.1節で述べた翻訳システムで利用されている品詞体系で、表3に示すデータで実験を行った結果、形態素解析の正解率は、87.4%であり、単語分割の正解率は、93.1%であった。

表3: 形態素解析での実験用データ

学習データ	約22万語(1万5千文)
スムージング	約6万語(3千文)
評価用データ	約7万語(5千文)

英語構文解析

これまでに述べた処理機構とほぼ同様の処理機構で英語の構文解析を行った[11]。英語の場合、3節で述べた統合の仕方として統合Aを採用している。入力として正解タグを与えた場合と正解タグを与えなかった場合の結果を、表4、表5に示す。この実験では、英語の文法体系として非常に詳細な体系を用いており、意味概念を含む3000程度の品詞タグを持つものであり、品詞付与が非常に困難な問題となっている。英語の品詞付与の精度は、品詞付与だけで大量のデータから学習した場合に70%前後であり、構文解析機構として学

習データを減少して学習した場合には、60% 前後の精度と考えられる。

表 4: 正解タグを付与したテキストの入力

length	文数	top	top 20	cross
1-10	1044	81.8%	95.0%	89.1%
11-15	248	30.2%	72.6%	43.1%
16-23	201	17.4%	48.3%	28.4%

表 5: タグを付与していないテキストの入力

Length	完全一致		統語のみ一致	
	top	top 10	top	top 10
1-10	34.5%	40.1%	50.4%	62.3%
11-15	1.2%	3.6%	11.3%	25.6%

5 考察

日本語の形態素解析の誤りでは、複数の語をまとめたような分割が多く現れており、それらの品詞の接続では、構文としてまとまらないものが含まれている。これらの誤りは、統合した解析機構で解消することができると思われる。現状の解析機構では、文法規則適用の決定木で利用している語法の特徴が不十分であり、解析候補を絞り込むことができない場合が多いが、統合した解析機構で解析結果を出力した場合の形態素解析結果は、形態素解析のみの場合に比べ、数 % の精度の向上が見られている。

文法規則適用の決定木で利用される「語法の特徴」の整備については、構文木データを構築している作業で得られた知見や、文法作成過程で得られた知見をもとに、進めていく予定である。

6 まとめ

本稿では、確率付決定木を用いた日本語構文解析法を、形態素解析機構を統合した機構として提案した。本機構で取り込んだ形態素解析機構は、辞書を必要としないものであり、頑健な処理機構になっている。また、構文解析で利用する統計的言語モデルと、同じ決定木モデルを利用した機構であるため、統計的言語モデルとしても、統一のとれたものになっている。今後は、構文解析のための語法の特徴を整備することにより、精度の向上を目指す。

さらに、翻訳システムでの訳語選択等に有効になるような意味解析を、文法体系で利用している素性に意味概念を捉えた素性を取り込むことで、実現できるように豊かな情報を持つ文法体系に変更し、多様な特徴を利用した高精度の解析機構の実現を目指す。

参考文献

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 春野, 白井, 大山: “決定木を用いた日本語の係り受け解析”, 自然言語処理シンポジュウム'97 「実用的な自然言語処理に向けて」, http://www.csl.sony.co.jp/person/nagao/nlsym97/, 1997.

[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)