論文の内容の要旨

論文題目　TOWARDS PRACTICAL HPSG PARSING
主辞駆動句構造文法による実用的なテキストの解析にむけて

氏　名　鳥澤 健太郎

本論文では、言語学的に妥当な文法フォーマリズムである主辞駆動句構造文法(Head-driven Phrase Structure Grammar)の自然言語処理における実用化を目指した一つの研究について報告する。従来、文法の記述形式としては様々な枠組みが提案されてきたが、主辞駆動句構造文法はその中でももっとも洗練されたものとして捉えられている。この理由はこの文法枠組みが、言語学的な抽象化を経た一般的な文法記述から個々の語に特有のイディオマティックな言語表現に関する具体的な文法／辞書記述まで一様な表現形式で記述することを許すことにある。これにより、従来の純粋に言語学的な文法記述では困難であった計算機上での実現および現実のテキストの解析、すなわち構文解析への使用が比較的容易に可能となっている。また、もう一点、主辞駆動句構造文法の望ましい点としては、解析結果として構文解析木を出力するだけでなく、意味表現とよばれる一種のデータ構造を出力できることである。自然言語の表層の文とその意味するところの関係は非常に複雑であるが、主辞駆動句構造文法の枠組みが提供する表層のテキストから意味表現への写像は他の文法枠組みに比して非常に柔軟であり、アプリケーション作成の際には有利になるものと期待される。

しかしながら、主辞駆動句構造文法を用いた構文解析を機械翻訳などのアプリケーションで使用するには、依然として解決しなければならない二つの問題がある。

- 実際に計算機を使用してテキストを解析する際の速度
- 高い精度の構文木、意味表現を得るために必要な大量の辞書の作成

本論文では、これら二つの問題に関する研究を報告する。より具体的には、

- 主辞駆動句構造文法の文脈自由文法への近似的コンパイルにより構文解析を高速化する手法を提案する。高速化の効果は新聞の文などを使用した

実験によって実証される。

- テキストから語彙の意味的分類を自動学習するアルゴリズムを提案し、アルゴリズムが生成した語彙分類を、被験者の直感との整合性、及び主辞駆動句構造文法を用いる構文解析器と組み合わせた場合の解析精度向上という二つの観点から評価する。さらに学習アルゴリズムの性質として、可能な語彙の分類にある仮定を加えた場合に、このアルゴリズムが極限における同定という学習成功の基準をみたし、なおかつ conservative かつ consistent な学習アルゴリズムのクラスに属するということを示す。

まず、文脈自由文法への近似的コンパイルによる解析速度の向上であるが、この手法の正当性を一般的な形で示すのに十分な枠組みの形式化がこれまで存在しなかったため、とりあえず主辞駆動句構造文法の形式化を行った。具体的には、主辞駆動句構造文法で使用される表現形式である素性構造を操作するため、確定節プログラムというプログラミング言語の定式化を行った。このプログラミング言語は、実際に計算機上に実現されているプログラミング言語である LiLFeS の中核をなすものである。本研究ではこのプログラミング言語の解釈実行のアルゴリズムを与え、プログラムの意味を素性構造の集合として表現し、実行アルゴリズムの正当性を示す。また、文脈自由文法へのコンパイルは、文法の部分的な評価と考えることができるが、その部分評価を行う際に確定節プログラムの実行の停止を保証するため、遅延評価機構を導入しプログラミング言語を拡張する。先の場合と同様に、この拡張された言語に対しても、実行アルゴリズムを示し、その正当性を検証する。さらには、やはり、文脈自由文法へのコンパイル手法の正当性を示すのに必要な実行結果の単調性も証明する。

ついで、確定節プログラムを用いて主辞駆動句構造文法を定式化した。従来、主辞駆動句構造文法を計算機で実現する場合には、実現が容易なように単純化を行った上で行うのが常であった。本研究で提案する定式化では、より、枠組みの原点に近い形での記述を許す。また、この定式化では、素性構造への再帰的操作といった複雑な操作を文法に記述することをゆるしたうえで、確定節プログラムで言及した単調性が保存されることを示す。さらに、この定式化を用いて、構文解析のプロセスを定式化する。

以上の準備を行ったのち、文脈自由文法への近似的コンパイル手法を導く。主

辞駆動句構造文法を用いた構文解析が低速である第一の原因は、素性構造上の操作である単一化とよばれる操作が非常に高価であることである。したがって、解析の高速化を図る一つの手段は、解析時の単一化を可能な限り減らすことである。我々の文脈自由文法への近似的コンパイル手法は、コンパイルされた文脈自由文法によって、単一化を用いた本来の解析に先立ち、可能な構文解析木を近似的に生成する手段を提供する。これは、逆の捉え方をすれば、もとの主辞駆動句構造で受理され得ない構文木を単一化を用いることなく前もって除去することになり、構文解析に不可避な試行錯誤に要する単一化の回数を減らすことが可能となる。本論文では、まず、構文解析木を幹(trunk)と呼ばれる部分的な構文解析木に分割し、構文解析木がオリジナルの主辞駆動句構造文法を用いて受理できるか否かの判断を幹を用いて近似的に行えることを示す。ついで、この幹の集合を表現する手段として語彙項目オートマトンと呼ばれる概念を導入する。重要な点は、幹および語彙項目オートマトンは一つの単語のみに依存するということ、および、すべての可能な構文木を考慮しようとすると原理的には無限の幹が必要となるが、このような無限の幹の集合は有限の語彙項目オートマトンによって表現しえるということである。これにより、語彙項目オートマトンは入力文が与えられる以前に辞書のみから実際に生成できることが示せ、また、この生成結果を用いて構文解析木の受理を高速に判定できることが示せる。文脈自由文法へのコンパイルは、語彙項目オートマトンを変換することと見なせる。また、さらなる高速化を計るため、構文解析以前に生成された語彙項目オートマトンを用いて文法記述の中に含まれる確定節プログラムの実行を高速化する技法についても述べる。以上に述べた手法の効果は、複数の英語、および日本語の文法を用いた解析実験によって実証される。

ついで、意味的な語彙分類の学習手法についてであるが、ここで分類とは、語集合の分割のことをいう。分割された語の集合それぞれ、すなわち語のクラスはなんらかの意味的な共通の性質を持つ語を含むものと考える。このような語のクラスは、構文解析や、表層の文を意味表現に写像する際に必要となる。本研究では、語彙分類のテキストからの自動学習の定式化をおこない、まず、主辞駆動句構造文法を用いた統計的構文解析での精度向上によってその評価を行う。さらに将来の意味表現への写像の実現を考慮しつつ、如何に人間の意味的直感にあった語彙分類が得られるかという点についても評価を行う。自然言語

の単語の特徴として、一般的なクラスタリング手法の研究などでは見落とされがちな重要な点がいくつかある。その一つは、自然言語の語は曖昧性を持つということである。つまり、一つの語は複数の対象／概念を指示しうる。卑近な例では、「私」とは自分自身のことを指すだけでなく、「私する」という文脈では、「私物化する」という行為を指し示す。したがって、一つの語に対して、一通りのクラスを示すだけでは不十分である。逆の言い方をすれば、オーバーラップをゆるした語集合の分割を考慮する必要がる。また、もう一つ重要な点は、自然言語のユーザーすなわち一般の人々は、非常に大量の語の意味、つまりは語の所属するクラスを含む情報と考えられるもの、を比較的少数の文から学習できるということである。これを説明するための一つの仮説は言語以外の認識機構によって付随的な情報が得られ、未知の語に対する分類の助けとなるという仮説である。しかしながら、語の指し示すものは視覚、聴覚などの他の認識機構によって認識できるものだけとはかぎらず、この仮説で十分であるとはいいがたい。さらに、一般の人々の語の分類は、少なくとも基本的なレベルでは人によらず、収束している。各々の人々が語彙を獲得する環境が多いに異なることを考えれば、この事実は自明な方法で実現されるものではない。本研究での我々の出発点は、より正確な文法、ないしはテキスト解析の枠組みで必要な語彙の分類は、人間が使用しているであろうものとなんらかの形で一致しており、おそらく以上の３点の特徴を反映しているべきものであるというものである。より具体的には、言語中での語の使用の文脈、すなわち例文の集合から、語の意味的分類を学習するという前提に立ち、以上にあげた三点、つまり、語の曖昧性、少数の文からの大量の語彙分類の学習、および、学習結果の収束を考慮した学習モデルおよびアルゴリズムを定式化する。その際に、仮説として、曖昧でない語の存在、および、未知語のクラスを一意に決定できる文脈の存在を仮定する。この仮説は以上にあげた３点の語彙分類の特性に関して望ましい性質をもつ。特に、学習アルゴリズムの性質を、Gold によって定式化された「極限における同定」という概念を用いて分析する。また、この収束は、正の例のみを使っただけで可能であり、さらに conservative かつ consistent といわれる学習アルゴリズムで多項式時間のものが存在することを示す。さらにそのようなアルゴリズムで統計的手法を用いるものを示し、新聞一年分のデータを用いた実験結果を示す。