論文の内容の要旨

論文題目
Local Structure Analysis of Japanese Using Lower-level Contextual Information
（低レベル文脈情報を用いた日本語局所構造解析）

氏　名　　久光　徹

インターネット等の普及により，昨今では膨大な量の電子化された文書が流通している．そしてそれらの情報を有効に活用するためには，自然言語処理技術が不可欠になりつつある．自然言語処理とは自然言語をその意味構造を反映したなんらかの形式的な構造へ変換する処理を指し，以下の諸段階からなる．すなわち，文を構成する各単語の認識（形態素解析），複数の単語からなるまとまった文法単位である「句」の同定とその内部構造の解析（ここでは便宜上句内構造解析と呼ぶ），句と句の間の文法的な関係の同定（構文解析），意味構造への変換（意味解析）である．本論文では，形態素解析と句内構造解析をあわせて局所構造解析と呼ぶ．局所構造解析は，殆どすべての自然言語処理における基盤技術であり，昨今では，機械翻訳だけでなく，情報検索，情報抽出，文書要約等においても不可欠の技術となりつつある．そして大量の開いた文書を扱わねばならない状況下では，高精度であるだけでなく，頑健で移植性の高い局所構造解析技術が強く求められている．本論文の目的は，日本語テキストを対象とする局所構造解析においてこれらの要求を満たすための新たな方法を提示し，その有効性を示すことである．

膠着型言語である日本語は，語自体の屈折変化や語順によらず，内容語に機能語が接続してできる文節と呼ばれる単位により，各語の文法的役割が示される．日本語の特性から，日本語形態素解析は，「入力文字列を辞書に記載された単語へ分割する処理」，日本語の句内構造解析は，「文節の同定とその内部構造の解析」と定義できる．ここで，語境界が空白で明示的に示される英語と異なり，日本語形態素解析においては語境界の同定と品詞付けを同時に行う必要があるため，最適単語列を求めるための探索空間が大きく，問題はより複雑である．また，語境界があらかじめ与えられないため，未知語の扱いも英語に比べて遥かに困難である．一方，日本語の句内構造解析においては，文節の境界の同定と名詞句以外の文節の内部構造の解析は容易である（例えば動詞句においては，動詞と屈折接辞間，屈折接辞同士の接続にきわめて強い言語的制約がある）が，名詞を含む文節中では，特に名詞連鎖からなる複合名詞が含まれているときには，複合名詞を構成する名詞間の依存構造の解析が必要であり，これは言語によらない困難な問題である．従って本論文では，句内構造解析に関しては，最も困難な複合名詞の構造解析に焦点を当てる．

日本語局所構造解析における手法は，言語的な知識にもとづき人手によりヒューリスティクス（ルールやコスト関数として実現）を構成する手法と，コーパスから自動的にデータを抽出してコスト関数や確率モデル等を構成する手法に大別できる．日本語形態素解析においては，前者の枠組みで，未知語がほとんど存在しないという条件下で，recall，precision 共に95～97%程度を達成するといわれており，後者の枠組みでも，例えば確率に基づく手法により同程度の精度が実現できるとしている（ただ，歴史的な経緯や，作成に必要なコーパスの量，可読性等の問題から，現時点ではルールに基づく形態素解析システムが多数を占めている）．一方，複合名詞の構造解析においては，ルールに基づく方法では，構成要素となる名詞の構文的・意味的情報を利用することにより，閉じた領域で，平均語基数3.4の複合名詞について95%程度の解析精度が報告されている．コーパスに基づく方法では，学習コーパスから適当な共起条件に基づいて抽出した名詞対を，シソーラス等で概念間の共起データに写像することにより概念間共起として学習し，これを用いて，3語(相当)の未知語を含まない複合名詞の構造解析において，日本語．英語ともに80%弱の精度を得ている．

しかし開いた文書に対応するため，従来の手法の枠組み内で，未知の文書への対応も考慮しつつ，より一層の精度向上を目指して解析方法を詳細化しようとすると，大きな困難に遭遇する．ルールに基づく枠組みでは，手法の精緻化

に伴い，ルールの作成・維持・管理に必要なコストが指数的に上昇する．コーパスに基づく枠組みでも，モデルの精緻化に伴い学習に必要なデータ量が指数的に増大し，"sparseness problem" が顕著になる．ここで，解析精度の向上のためには，単純なルールや確率モデルだけでなく，語に関するより高次の情報，すなわち，構文的，意味的，語用論的情報を用いるべきだという考え方もあるが，そのような高次の情報を用いようとすると，未知語に対して本質的に脆弱となってしまう．

ここで実際に局所構造解析の誤りを分析すれば，形態素解析にせよ，複合名詞の構造解析にせよ，多くの誤りは，必ずしも高次の情報でなく，周辺の単なる文字列や単語レベルの情報を手掛かりに解消できることがわかる．従って，従来の方法を精緻化・複雑化する方向とは全く異なり，「周辺の情報」を捉えて解析に利用するという新たな方向が考えられる．本論文では，この考えに基づき，従来の一文解析の枠組み内では取り扱えなかった，曖昧性や未知語処理に必要な「低レベルの文脈情報」を，解析対象の文の境界を越えて獲得し，従来型の手法と組み合わせて利用するパラダイムを提示する．ここで，「低レベル」とは，「対象とする解析の出力結果に含まれるレベルの情報を越えない」ことを意味する．例えば形態素解析においては，解析前の文字情報や，（誤りも込めて）形態素解析自身が出力した最適単語列，解の曖昧性の情報等が，この範疇に含まれる．また，ここでいう「文脈」とは，対象とする解析において，「低レベルの情報」を参照しうる一定範囲の文集合を指す．文脈内の低レベル情報を「（低レベル）文脈情報」と呼ぶ．参照する情報の種類と，「文脈」の広さの組み合わせについては様々な可能性があるが，例えば速度を重視する形態素解析では，より局所的かつ単純な情報を，複合名詞の構造解析では，より大域的な情報を参照することが自然であろう．以下本論文では，それぞれの処理に即して，具体的な組み合わせを検討する．

本論文の構成は以下のとおりである．まず，文境界を越えた文脈情報を有効に利用できるには，最も基礎となる初期形態素解析において，その解析精度・効率・頑健性が，ある一定の水準を越えていなければならない．これはそれ自体重要なテーマであるため，第２章において，現在多く用いられているルールに基づく形態素解析を念頭におき，解析アルゴリズム，コスト関数の体系的設定方法，辞書の見出しの最適化等について論じる．２章の内容により，簡潔・高精度・頑健であり，解析の曖昧性をコンパクトに保持し，文脈情報の情報を利用することを可能とする形態素解析が実現できる．

3，4章は本論文の中心となる部分であり，低レベルの文脈情報と，基本となる従来手法を組み合わせる方法を具体的に論じ，その有効性を検証する．3章では，形態素解析ついて，4章では，複合名詞の構造解析について述べる．開いた大量の文書の典型的な例として，新聞記事を題材とする．

3章では，2章で基礎を与えた形態素解析を基盤として，まず最も狭い文脈情報の利用形態である，形態素解析結果の書き換えルールを用いた後処理方法について述べる．書き換えルールは，日本語に適用するための改良を加えた誤り駆動型の自動学習により獲得する．後処理の精度を更に向上させるために，書き換えルールに加えて，「窓」と呼ぶ複数の文集合に対する解析結果をプールし，それらを相互に参照することにより，より高度な曖昧性の解消や，未登録語の同定も可能となることを示す．「窓」がすなわち「文脈」であり，参照する情報としては，窓内の各文の解析結果の曖昧性までを含む．これらの後処理により，未知語を比較的多数含む条件下で，recall，precisionともに，2～3%程度の向上が可能である．

4章では，漢字で書かれた名詞連鎖による複合名詞を対象とし，複合名詞の構造を規定する少数の基本ルール群とヒューリスティクスに加えて，複合名詞を構成する単語の共起情報を必要に応じて文脈中から獲得することにより，高精度かつ未登録語に対して頑健な構造解析が可能なことを示す．この際，二つの名詞間の共起は，4章で定義するテンプレートに含まれる二つの変数部分に二つの名詞がマッチすることにより定義し，質・量を兼ね備えた共起情報が獲得できることを示す．また，初期形態素解析において未知語の解析に失敗した場合も，テンプレートによる共起情報の獲得中，多くの場合文脈中の表層情報からその未知語自体が同定でき，同定された未知語を新たに共起情報抽出の対象に繰り入れることにより，未知語の解析誤りも修復される．この結果，頑健で移植性に優れた高精度な複合名詞解析が実現できる．精度は，3単語相当で88%を達成し，未知語の存在を考慮すれば，従来手法を大きく上回る．

5章では，局所構造解析の精度向上のためのもう一つの方法として，オフラインでの未知語獲得について論じる．具体的には，3，4章で示した方法では解析時に同定が困難と判明した，人名，社名等の固有名詞や，略称等を対象とし，単語同定のための少数のルールと，文書全体から得られる簡単な統計量を組み合わせて，新聞紙上に現れる未知語を，recall，precisionが各50%，95%前後で収集できることを示す．ここで，全文書を「文脈」，統計量を「文脈情報」とみなすことができ，この意味で5章は3，4章の延長線上にある．