社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (17)
2006－ＮＬ－175 (17)
2006／9／13

純粋な逐次処理による構文解析での探索候補数の削減

狩野 芳伸　　宮尾 祐介　　辻井 潤一
東京大学情報理工学系研究科コンピュータ科学専攻

人間の言語処理についての心理学的な研究から、自然言語処理の研究で使われる多くのパーザと比べて、より解析時の構文候補数の少ない処理が可能なはずではないかと考えられる。自然言語処理の構文解析における研究は、より詳細な情報を導入する方向で進んできたが、本稿では、十分な情報を含んでいると仮定できる大規模解析済みコーパスから、情報を選択的に削減してゆくアプローチをとる。その上で、心理学的な知見から制約条件を導入し、逐次処理構文解析での候補分岐数を指標として情報を選択する手法と、実験計画を提案する。解析精度よりも直接的な計測が可能な指標を用いることで、情報の選択と指標との関係の分析が容易になることが期待できる。

Candidate Reduction in Syntactic Structure Analysis with Pure Incremental Processing

Yoshinobu Kano, Yusuke Miyao, Jun'ichi Tsujii
Department of Computer Science, Graduate School of Information Science and Technology, the University of Tokyo

Psycholinguistic researches suggest that humans are able to parse sentences with less candidates than the most NLP parsers. We take an approach to reduce information selectively from a large corpus which includes enough syntactic clues, while most of the syntactic parsers have developed to use more detailed information. Methods and experiment plans are proposed, in which psychological restrictions are introduced and candidate branch factors of the incremental parsing are used as an indicator. This indicator will make the analysis of the relation between the information selection and the indicator easier.

1. はじめに

構文解析で学習に用いられる新聞記事等から作成されたコーパスは、人間も高い精度で解析が可能であると期待できる。同時に、次節以降で説明するように、自然言語処理の研究で使われる多くのパーザと比べて、より解析時の候補数の少ない処理が可能なはずではないかと期待される。

人間の自然言語処理においてどのような情報が使用されうるかについては、心理学的な知見からある程度の制約を導ける。この心理学的な制約を導入することで、より少ない種類の情報でも構文解析が可能になり、構文候補数を削減できるのではないかと考える。

自然言語処理の研究では、「ある要因を導入したことにより、解析精度が向上した」という報告が多くみられる。しかし、各要因単独では精度の向上が観察されるが、複数の要因の組み合わせにおいてはそれほどの向上がみられない、といった結果が散見される。すなわち、各要因の解析精度に寄与する情報は重複していることが多く、また各要因がどのような相互作用をもつのかも必ずしも明確ではない。さらに、処理過程の多くの部分がブラックボックスになっており、ある程度以上の探求が困難な傾向があるように思われる。そこで、どの要因がどのような影響をもつのかという分析を容易にするために、解析精度よりも直接的に計測できる指標を使用したい。

一方で、絶対的な解析精度は、90%を超える報告も多い。そこで、高い解析精度を達成する際に学習データとして用いられたコーパスには、構文解析に必要な、多数のパラメータからなる十分な量の情報が含まれていると仮定する。その冗長性から、解析精度を損なうことなく、より小さな情報（パラメータ群）に削減することが可能ではないかと考える。

以下では、心理学的知見から導かれる制約を説明した上で、その制約を利用して対象コーパスからパラメータ数を削減する具体的な手法と実験計画を提案する。評価の指標としては、解析精度よりも直接的に計測できるものとして、構文候補の分岐数を用いる。

2. 心理学的制約

2.1. 心理学的知見

人間の言語処理が逐次的であることは、視覚的な入力処理における眼球運動の計測結果を見れば、あるいは聴覚的な入力処理を考えれば直感的にも明らかである。本稿では、

- 入力トークン列を順に処理する
- 処理過程で使用可能な情報はその時点までの入力列から導かれるものに限定される
- 処理過程で保持する情報量は入力に依存しない上限をもつ

ような構文解析処理を**純粋な逐次処理**と呼ぶことにする。たとえば、全ての入力を受け取った後ではじめて解析を行うものは、純粋な逐次処理には該当しない。処理過程で情報の取捨が行われる必要があるものとする。

埋め込み文の研究[1]からは、ある程度

以上の深さの埋め込み構造の処理が著しく困難であることが知られており、処理過程で保持する情報量の種類と上限を示唆している。

また、ガーデンパス文の研究[2]等から、逐次処理での解析過程で保持する必要のある構文候補数は、ほとんどの文章で非常に少ないことが期待される。すなわち、構文候補の選択で迷うように人工的に作られたガーデンパス文においても、限られた箇所で二つから数個程度の選択肢がありうるのみであり、ほとんどの部分では決定的である。ただし、人間の自然言語処理で利用されているであろう文をまたがるような文脈的な情報については、本稿の範囲外であり、このレベルの情報を扱わないことによる影響を考慮する必要がある。

2.2. Left-Corner文法

Left-Corner文法では、純粋な逐次処理が可能であり、埋め込み構造の深さと使用するスタック量が対応する[3]ため、上記のような心理学的制約を表現するのに適当な形式である。以下では、Left-Corner文法による構文解析を前提とする。

2.3. 心理学的制約

任意の CFG 規則を可逆な形で Left-Corner 文法に変換できるのが、Left-Corner 文法の特徴のひとつであり、何らかの方法でコーパスから抽出したCFG規則に対しLeft-Corner変換を適用する、という手順を経ることが多い。しかし、前述のように解析を容易にするため、もともとのコーパスにおけるパラメータの分布での関係を調べたい。

そこで、Left-Corner文法の解析処理過程の各ステップで使用・保持される情報を、ステップごとに条件として抽出し、**Left-Corner条件**と呼ぶことにする。ここでいうステップとは、Left-Corner文法による構文解析において「ある入力を受け取ってから次の入力を受け取るまでの処理」を指す。

Left-Corner 条件は、Left-Corner 変換前の CFG 構文木で考えた場合、Left-Corner カテゴリ・目標カテゴリ・スタックで構成されることになる。Left-Corner 条件が同一であるとは、Left-Corner 条件を構成するすべてのパラメータが一致するということである。

心理学的制約として、構文解析に用いる情報を Left-Corner 条件に限定することとし、この条件を次節のパラメータ削減における前フィルタとして用いる。その上で、解析精度を損なわない範囲で、逐次処理過程での構文候補分岐数を減らすことを指標とし、パラメータを減らすことを目標とする。

3. パラメータの削減

木構造の抽出できるコーパスの文データは、Left-Corner文法における逐次処理の順序に従い、Left-Corner条件からなる各処理ステップの一次元の配列として並べることができる。コーパスをこのような一次元配列の集合と考えたとき、同一のLeft-Corner条件をもつ二つ以上のステップからなる組が存在しうる。コ

ーパスを統計的に処理した場合、このステップの次にとりうるステップが二つ以上存在することになるので、このステップは構文解析における**分岐点**を形成する。すなわち、分岐点を形成するステップは複数の次ステップ候補をもつ。

コーパスの持つ情報が、構文解析に必要な情報を超えて冗長であれば、不要な分岐が存在する可能性がある。そのような不要な分岐を削減する方法を考える。

二つの文があって、両者が二つの分岐点を共有するとき、分岐点間のステップ群を**分岐区間**と呼ぶことにする。いくつかのパラメータを削除することで、分岐区間の対応するステップ群をそれぞれ同一のLeft-Corner条件とすることができれば、分岐数を削減することができる。

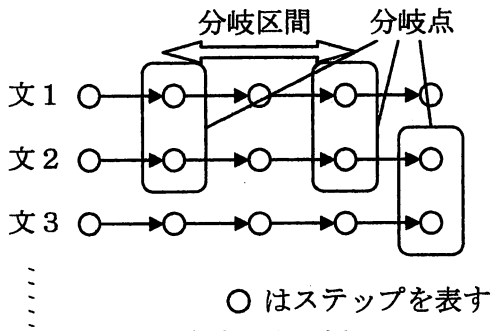

図1 分岐区間の例

しかし、単純にパラメータを削除していくと、本来必要なパラメータをも削除してしまう可能性がある。そこで、ある種のパラメータ群を削除対象外と指定した上で、

- 分岐区間で削除対象外のパラメータがすべて一致するものを候補とする
- 削除パラメータ数の少なさ、および分岐区間の短さ順に候補を処理する

ものとする。削除対象外とするパラメータの選定と、パラメータの削除をどこまで進めていくかが、重要な要素となることが予想される。

4. 実験計画

4.1. コーパス

英語の大規模解析済みコーパスであるPenn Treebank[4] をHPSGに変換したもの（以下HPSG Treebankと呼ぶ）を用いる。 HPSG Treebankからいくつかの素性を選択してCFGに変換しても十分高い精度で解析ができ、処理過程で保持すべき構文候補の数も比較的少ないことがわかっており[5]、前述の目的に必要な十分な情報を含んでいることが期待できる。

実験には、HPSG Treebank からいくつかの素性のみを残したものを用いる。これまでパラメータと呼んできたものは、HPSG の場合、素性ということになる。

4.2. 実験条件

実験は、以下のように異なる条件下でパラメータの削除を試み、評価を行い比較考察を行う。

- HPSG Treebank から導入する素性の種類：HPSG パーザでの解析結果を参考に、数種類のセットを用意する。
- 削除対象外とするパラメータ：品詞情報を削除対象外とする

が、Penn Treebank の品詞をそのまま用いる以外に、頻度の高い単語を別扱いとするものも試みる。

- Left-Corner 条件の使用範囲：Left-Corner 条件は、かなり深いスタック情報を含むため、スタック深さを限定した場合と、限定しない場合を試みる。

4.3. 評価

実験の評価は、以下の各点について行う。

- 分岐数の分布：出現頻度の著しく低い分岐は除外した上で、除外したものの数と分岐数の分布の変化を観察する。また、削除された分岐や残った分岐が妥当なものであるかをチェックする。
- 解析精度：パラメータを削減したデータで Left-Corner 文法の学習を行い、Treebank をパーズしブラケットレベルの解析精度とカバレッジを観察する。Left-Corner 文法は、二つのカテゴリとスタックに保持されるカテゴリの組み合わせでひとつの条件を構成するため、スパースネスが問題になるので、そのままでは絶対的な精度の高さは期待できない。相対的な解析精度の変化を測定する。
- 先行研究との比較：　Kleinらの研究[6]では、一般的なPCFGにおいて、語彙化（単語のもつ個別情報を導入する）を行わずにどこまで解析精度を向上させうるかを追求している。Kleinらが導入した要因の分布と、パラメータの削除により得られた分布を比較する。

5. まとめと今後の課題

本稿では、心理学的な知見から、純粋な逐次処理を行える Left-Corner 文法の使用を前提とし、Left-Corner 条件により絞込みをした上で、パラメータを削減することで構文候補の分岐数を削減する手法を提案した。

解析精度よりも直接的な計測が可能である、構文解析における分岐数を指標とすることで、パラメータと指標との関係の分析が容易になることが期待される。

今後の課題は、まずは実験計画の実行と評価である。その上で、検討すべき点を挙げる。

もともとのコーパスにおける各パラメータの分布に、期待されるような分布が含まれているのであれば、解析精度に寄与するものを残しつつ、パラメータの削減による分岐数の削減を成功させられるだろう。しかし、分布そのものを操作する必要がある場合は、パラメータの結合などを検討する必要があるかもしれない。

あるいは、今回は対象コーパスとして情報量の多い HPSG Treebank を選んだが、実験の前に何らかの変換を行う必要

はないか、あるいは他により適したコーパスがないか検討する必要があろう。構文木の形、特に、Penn Treebank から二分木化（binarization）する方法については、Left-Corner 文法との相性があると考えられる。

また、心理学的制約として挙げた埋め込み構造の深さについて、自己埋め込みの回数に上限をもたせる形で、パラメータ削除の評価指標に用いたいと考えている。

参考文献

1) Richard L. Lewis. Interference in short-term memory: The magical number two (or three) in sentence processing. *The Journal of Psycholinguistic Research*, 25(1):93-115, 1996.

2) Richard L. Lewis. Falsifying Serial and Parallel Parsing Models: Empirical Conundrums and An Overlooked Paradigm. *The Journal of Psycholinguistic Research* 29:241-248, 2000.

3) Mark Johnson. Finite-state approximation of constraint-based grammars using left-corner grammar transforms. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*. 619-623, 1998.

4) Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19:313-330, 1993.

5) Matsuzaki, Takuya. Private Communication. 2006.

6) Klein D., Manning C. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, 423-430, 2003.