コーパスにおける文脈情報を利用した文法開発支援

川口 恭伸

北陸先端科学技術大学院大学 情報科学研究科

1997 年 2 月 14 日

キーワード： 文法開発, 仮説選択, 局所文脈情報.

本論文では、特定のコーパスに対し適用するように既存の文法をカスタマイズする作業を支援する方法を紹介する。

自然言語処理のアプリケーションにおいて実際に使用するために、われわれは使いやすく広い適用範囲を持った文法を必要としている。しかし、文法を開発することは、既存の文法を修正するとしても、時間を労力を要する作業である。一方、様々な形式のコーパスから自動的に文法を獲得する研究があるが、獲得された文法を実際のアプリケーションに組み込むためには修正が必要であることが多い。

文法修正の作業は、3 つに分類することができる。欠落した文法規則を推定して追加する**規則追加**、冗長な文法規則の削除を行なう**規則削除**、粒度の粗いカテゴリをより精密に言語使用と対応するいくつかのカテゴリに分割する**規則分割**である。本研究では、このなかでまず**規則追加**について扱った。規則追加では既に解析できている文が解析できなくなるといった致命的な悪影響が発生しないことと、文法の自動獲得で用いられている評価基準を参考にすることができるからである。

本研究の目標は、追加すべき規則の尤もらしい仮説を開発作業者に提示することである。それにはまず、解析不可能な文における部分解析結果と現在の文法をもとに、1 つの規則追加で解析が可能になるような規則の仮説を集め、その後で、尤もらしさを量る統計的評価値を用いて仮説の順位付けを行ない、上位いくつかの仮説を、開発作業者に提示する候補とする。

文法仮説にはインスタンス仮説と一般仮説という 2 つのレイヤーがある。前者は、文に対する不完全な解析木における欠落部分を埋めることができる不活性弧を表す。後者はインスタンス仮説に表される不活性弧を生成することができる文法規則を表す。

我々が採用した文法開発の枠組は、以下の 2 つの部分から構成される。

- 仮説生成 (規則ベースのアプローチ)
  現在の規則の欠落をうめるインスタンス仮説を多数生成する。

- 仮説選択 (コーパスベースのアプローチ)
  局所文脈情報による尤度評価値によって、生成されたインスタンス仮説を絞りこむ。

この枠組は、もともと Kiyono と Tsujii によって 1994 年に提案されたものである。

我々の使用した文法形式は、文脈自由型句構造文法である。この文法形式はシンプルで、多くの自然言語アプリケーションが採用している。仮説生成アルゴリズムは Kiyono と Tsujii が提案したものに基づいている。このアルゴリズムは、文法と部分解析結果の不活性弧ともとに多くのインスタンス仮説を生成する。生成された仮説はそれを含む文を完全な解析へと導くものであるが、その多くは間違った解析へ導いてしまう。規則ベースのアプローチでは正しい仮説を選択することができない。そこで、我々はコーパスベースのアプローチを用いて仮説選択を行なう。

仮説選択においては、仮説の尤もらしさを測るために、局所文脈情報を使用する。ここでいう局所文脈情報とは、あるカテゴリの直前直後にくるカテゴリのことである。我々のスコアは仮説の局所文脈情報と、同じラベルをもつカテゴリの正解構文木集合のなかでの局所文脈情報から、その仮説の尤もらしさを決定する。

例えば、インスタンス仮説 $hypo(A)$(A はカテゴリラベル) の左にカテゴリ L、右にカテゴリ Rがある場合には、$hypo(A)$ の尤もらしさは、L,A,Rという組が、正しい解析が行なわれている文の中で表れる頻度をもとに計算される。我々はこの L,A,Rのようなカテゴリのセットを “カテゴリのトライグラム” と呼ぶ。

我々のスコアの有用性を確かめるために、ある文法と、5000 文を超える解析不能文を用いて実験を行なった。この文法は括弧付きの EDR 英語コーパスから自動的に学習されたものであり、このコーパスの約 60%の文を解析することができる。

まずはじめに、我々は正しく解析された文からカテゴリのトライグラムを得る。実験コーパスは、解析が失敗した 5082 文である。仮説生成モジュールは 4730 文に対してインスタンス仮説を生成したが、そのうち 1513 文においては正しい仮説を含まなかった。そこで、この 1865 文は仮説生成における失敗とみなすことができる。残りの 3217 文は、多くの正しくない仮説と共に正しい仮説を持つ文である。

続いて、仮説に正解が含まれた 3217 文に対して、仮説選択の性能を調べた。結果として、3217 文のうちで 1362 文において、提示する仮説の候補数をスコアの上位 1 つだけに絞った場合に、正解を提示することができ、候補数を 10 に広げた場合、2634 文に対してせいかいを提示することができた。これは正解仮説が提示できた文の 81.9%をカバーする。しかし、一般にひとつの文には複数の仮説が存在するので、すべての正解仮説が上位にランクされたかたいうとそうではなかった。また、カテゴリトライグラムとして考慮するカテゴリの種類を語彙カテゴリのみに絞った場合よりも、すべてのカテゴリを考慮した場合の方が、やや高い精度を持つことがわかった。

本論文は5つの章から構成される。第1章は本研究に対する導入部である。まず第2章で、仮説生成の方法と、仮説選択の必要性と他の研究における仮説選択について述べる。そして第3章では、本研究で提案する環境の分布を用いた仮説選択の方法について述べ、第4章でこの仮説選択法の有効性についての実験と結果について述べる。最後に、第5章で本研究の結論と今後の課題を述べる。