英語文パーザにおける表層的・統語的情報を用いた構造的曖昧性解消機構

斉藤仁史　　宮崎正弘

新潟大学大学院自然科学研究科

1　はじめに

英語文構文解析では、各単語の多品詞性や係り受け等による構造的曖昧性によって多くの解析木が生成される。品詞の曖昧性を抑制するために、接続表と単語出現頻度を用いて、構文解析の前処理としての形態素解析を行う。それによってパーザの処理量を減らし、構造的曖昧性を抑制する。構文解析は、文法記述量が少なく構造化 CFG を扱える逐次型のボトムアップチャートパーザ、”Schart Parser”[1] を用いて行う。ここで、動詞の文型制約による解析弧抑制処理、群動詞や連語の解析木決め打ち処理等をパーザに導入する。このように意味情報を用いずに可能な限り構造的曖昧性を減らす手法を考案し、その有効性を示す。

2　英語文形態素解析

各単語がもっている全ての品詞に対して解析処理を行うと、多品詞性やそれによる構造的曖昧性によって多くの解析木が生成される。そこで、品詞の曖昧性を抑制するために、構文解析の前処理として入力文の形態素解析を導入する。

形態素解析では、各単語において、前後の品詞同士の接続の可否や語形情報、頻度に基づいて品詞の絞り込みを行う。その手順を以下に示す。

1. 品詞同士の接続の可否を表す接続表の作成
2. 単語の語形・統語・頻度情報を英語辞書から獲得
3. 前方の単語と後方の単語との接続の可否を判定
4. 2,3 を文頭の単語から文末の単語まで繰り返し
5. 頻度係数による候補の絞り込み

2.1　接続表の作成

単語の品詞を名詞、動詞、BE 動詞、形容詞、副詞、人称代名詞主格、人称代名詞目的格、その他の代名詞、関係詞、前置詞、助動詞、接続詞、限定詞、不定詞の 14 個に、文頭、文末の概念を品詞として加え、品詞を計 16 個とした。ここで、限定詞は、冠詞、人称代名詞所有格、”this”、”some”などである。そして、未知語の品詞は名詞とする。

この 16 個の品詞を用いて、単語同士が接続できる場合は 1、できない場合は 0 の要素をもつ 2 次元配列表 (接続表) を作成した。その接続表の値は、文が平叙文なのか疑問文なのか、文中に特定の単語が出てくるかどうか、または単語の語形情報などによって変化する。

2.2　単語の語形・統語・頻度情報の獲得

EDR 電子化辞書 [2]、語形変化辞書、ユーザー辞書という三つの辞書を利用して、各単語の [”品詞”,”頻度”,”語形情報”] を獲得する。EDR 辞書で品詞と頻度を獲得する。ユーザー辞書は、他の二つの辞書が使えない場合、または使うまでもない場合に、決め打ち用の辞書として使用する。

2.3　接続の可否の判定

接続の可否を判定する処理は、(文の単語数＋ 1) 回行われる。これは、1 番目の単語の前に品詞が文頭である実体のない単語、最後の単語の後ろに品詞が文末である実体のない単語がそれぞれ存在していると考えているからである。

前方の単語、後方の単語のそれぞれの全ての組合せについて判定が行われ、接続が不可とされた後方の単

語の品詞は除去される。

判定の基準は接続表と単語の語形情報である。語形情報は具体的には、前方語の語形情報が○○○を含んでおり、かつ後方語の語形情報が△△△を含んでいる場合、後方語の品詞が×××である可能性を消すという形で利用する。例えば、前方の単語”is”の語形情報が[”BE 動詞単数形”] であり、後方の単語”play”の語形情報が [”動詞原形”,”名詞単数形”] である場合、BE 動詞単数形–動詞原形という接続はありえないので、”play”が動詞である可能性を除去する。

2.4 頻度係数による候補の絞り込み

最終的に文全体の品詞の候補は一つに絞りこまれないことがほとんどであり、第一候補だけを見て形態素解析が正解しているかを判断するのは無理がある。そこで、N 番目の候補までを有効とし、N 番目の候補までに全ての単語の正しい品詞が含まれていたら、形態素解析は成功であるとした。N というのは各文によって異なり、最大で文の単語数と同値、ただし頻度係数が第一候補の値より 2.3 以上小さい候補は除外とした。2.3 という値は $log(10)$ からきている。つまり、頻度が第一候補の頻度と比べて 10 分の 1 以下の候補が除外される。各候補の頻度係数を各単語の品詞の頻度の相乗平均と定義し、この数字を用いて各文の N を求め、候補の絞りこみを行った。

$$\text{頻度係数} = \frac{1}{n}\sum_{i=1}^{n}\log f_i$$

ただし
f_i：i 番目の単語におけるある品詞の頻度
n：文の単語数

2.5 形態素解析の正解率

平均単語数が約 7.5 であり、単文、重文、複文が混在する 300 文を入力文として、形態素解析の解析結果の正解率を求めた結果を表 1 に示す。保留というのは曖昧でどの品詞が正しいのか判断できない文であり、正解率を計算する際、例文数から除外している。失敗の原因としては、接続関係による正しい品詞の除去、頻度による正しい品詞が除去が半々といったところであった。

表 1: 形態素解析の解析結果 (正解率)

	例文数	成功	失敗	保留	正解率
合計	300	237	14	49	94.4%

3 英語文構文解析

構文解析は、逐次型のボトムアップチャートパーザ、”Schart Parser”[1] を用いて行う。”Schart Parser”の特徴を以下に示す。

- 解析を逐次型ボトムアップチャート法で行う
- 従来の CFG 規則に統語構造情報を埋め込むことができる構造化 CFG の導入
- 反復記号、選択記号、字面指定記号の導入

ボトムアップチャート法は、一般化 LR 法と比較すると余分な多義を発生する可能性がある。しかし、構造化 CFG や反復記号などの導入で、構造的曖昧性抑制と文法数削減の工夫がなされている。また、人間が見て解析途中の状態や文法記述が直感的にわかりやすいパーザである。

■構造化 CFG の例
```
(N1 N2 (N3 N4 N5))
```
■反復記号の例
```
(N1 N2 N3*)=(N1 N2 N3)|(N1 N2 N3 N3)...
```
■選択記号の例
```
(N1 [N2 N3])=(N1 N2)|(N1 N3)
```
■字面指定記号の例
```
(N1 N2:word N3)
```

図 1：”Schart Parser”の特徴

3.1 動詞の文型制約

英語には基本 5 文型と呼ばれる文法的な構造がある。その 5 文型も、補語や目的語に何を取るのかという点で細かく分類される。例えば、SVC の文型の文は、C にあたるものに名詞をとる場合や形容詞をとる場合がある。また、SVO の文型の文では、O にあたるものによってより細かな分類が可能である。これらは、動詞によって統語構造が大きく左右されていると言える。

そのため、構文解析のための文法を実装する際、このような文型の細かな分類毎に独立した文法を用意した。

単語の情報を獲得するために用いる EDR 辞書には、それぞれの動詞がどの文型を取り得るのかという情報が記述されている。形態素解析の段階でこの情報を獲得し、構文解析の際、動詞が取りうる文型によって読み込む文法を制限する。例えば、SVC の文型しか取り得ない動詞が文法を読み込むとき、SVC 用の文法のみを読み込み、SVO やその他文型用の文法は読み込まない。この処理によって、不要な解析弧の生成を抑制することが出来る。

3.2 動詞の変化形による制約

前節で述べた動詞の文型による制約に加え、動詞の変化形による制約を導入した。動詞には、原形、三単現形、過去形、過去分詞形、現在分詞形の五つの変化形がある。動詞句を生成する文法について、原形の動詞からなる動詞句、三単現の動詞からなる動詞句、というように五つの変化形全てのための文法を用意した。これによって、動詞句のための文法は 5 倍に増えたが、受動態や進行形の文など、特定の変化形の動詞しか取らない構文について独立した文法を用意することが可能となり、曖昧性を大きく減少させることが出来る。

3.3 群動詞決め打ち処理

英語には、動詞が副詞や前置詞、名詞と結びついて一つの動詞としての意味を表す場合があり、これを群動詞 (もしくは句動詞) と呼ぶ。全ての動詞表現から見たら群動詞はごく一部である。しかし、字面指定によって個別の文法を用意しておくには数が多すぎる。従って、通常の文法ファイルには群動詞の構造を正しく表せるような文法は用意しないこととする。そして、この結びつきを明確に表すために、文中に群動詞となる部分があった場合は文法ファイル外からの特別文法を呼び出す群動詞決め打ち処理を導入した。

具体的には、入力文の字面検索であらかじめデータベースに登録してある群動詞とマッチした場合、その群動詞の核となる基本動詞の文法を読み込む際に、特別文法を呼び出し、それのみを読み込む、というものである。図 2 は決め打ち処理を使用した場合としなかった場合の解析木の例を示している。

■使用した場合

```
|-vp0
   |-vpit00
     |-vit00--< look >
     |-pp0
       |-prep--< for >
       |-pobj
         |-np
           |-opron--< him >
```

■使用しなかった場合

```
|-vp0
| |-vpit00
|   |-vit00--< look >
|-advp
  |-pp2
    |-prep--< for >
    |-pobj
      |-np
        |-opron--< him >
```

図 2：群動詞決め打ち処理の例 (look for)

3.4 連語決め打ち処理

複数の単語が結びついて一つの意味を表すのは群動詞に限ったことではなく、前置詞やその他の品詞の単語でもありうる。また、特定の単語がキーとなる特殊な構文も多く存在する。それら一つ一つのための専用文法を文法ファイルに用意しておくのは、少なくわかりやすい文法という点で見ても、処理速度の点で見ても、得策ではない。そこで、これらについては群動詞決め打ち処理と同じように、特別文法による決め打ち処理を導入した。

4 評価

平均単語数が約 10.3 であり、単文、重文、複文が混在する 270 文を入力文として、構文解析の解析結果の正解率を求めた結果を表 2 に示す。

表 2: 構文解析の解析結果 (正解率)

	成功	F1	F2	T1	正解率
合計	226	13	31	3.94	83.3%

ただし
F1：解析が失敗する
F2：解析は成功するが、正しい構文木がない
T1：平均構文多義数

また、失敗となる場合（F1+F2）の原因を調査した結果を表 3 に示す。

表 3: 失敗の原因

	総数	R1	R2	R3	その他
合計	44	2	0	13	29

ただし
R1：形態素解析の副作用
R2：群動詞・連語処理の副作用
R3：PP-attachment 制限の副作用

正解率は 8 割を越えたが、平均構文多義数は約 4 とやや多めの結果となった。決め打ち処理の副作用が若干見られ、検討が必要である。

5 解析例

例文：`Time flies like an arrow.`
《結果数 002》

```
■結果 001 ■
|-s
  |-decl
    |-sbj
    | |-np
    |   |-n--< time >
    |   |-n--< flies >
    |-vp0
      |-vpdo10
        |-vdo10--< like >
        |-dobj
          |-np
            |-art--< an >
            |-n--< arrow >
■結果 002 ■
|-s
  |-decl
    |-sbj
    | |-np
    |   |-n--< time >
    |-vp1
    | |-vpit01
    |   |-vit01--< flies >
    |-advp
      |-pp2
        |-prep--< like >
        |-pobj
          |-np
            |-art--< an >
            |-n--< arrow >
```

6 おわりに

本稿では、”Schart Parser”に前処理の形態素解析と各種抑制処理を導入し、意味情報を用いずに可能な限り構造的曖昧性を減らす手法を考案し、その有効性を実験により確認した。しかし、あくまでも減らしただけであって、あらゆる文の構造的曖昧性を一つに絞りこめるわけではない。

今後の課題としては、以下のようなものが上げられる。

・形態素解析と決め打ちによる副作用の検証
・文法の見直しと群動詞・連語データベースの充実
・構文解析木の一層の絞りこみ

参考文献

[1] 川辺諭, 宮崎正弘; 構造を含む生成規則を扱える拡張型チャートパーザ–Schart パーザの実装–, 言語処理学会第 11 年次大会発表論文集 (2005)

[2] 株式会社日本電子化辞書研究所; EDR 電子化辞書仕様説明書 (1993)