ＭＳＬＲパーザにおける未定義語処理の一検討

2C－6

伍井 啓恭[*1]　植木 正裕[*2]　相川 勇之[*3]　杉山 健司[*1]　田中 穂積[*2]
[*1]㈱日本電子化辞書研究所　[*2]東京工業大学　[*3]三菱電機㈱

1．はじめに

文脈自由文法(ＣＦＧ)モデルに基づく自然言語の解析手法について、文法記述の明確性と高速性の両立を期待する観点で、これまで多くの研究がされている。それらの中でも、一般化ＬＲ(ＧＬＲ)法は、処理効率、及び拡張性の面で優れている[1]。また、形態素解析と統語解析を統合したＭＳＬＲ*パーザ[3]が知られている。しかし、ＧＬＲ法における日本語を対象とした未定義語処理は、あまり研究されていない。

斎藤[4]は、ＧＬＲ法を用いて、エラーにより解析が行き詰まった場合にカテゴリの置換、挿入、及び読み飛ばしにより解析を続行する方式を提案した。エラーが起きない場合でも、非終端記号を仮定するギャップ埋め処理を提案している。しかし、この方式は、試行しなければならない場合の数が非常に増大する危険がある。

今井[5]は解析が失敗した場合にステージを以前にreduce した時点まで戻して処理する方式を提案している。しかし、今井[5]の対象は英語である。英語の場合、未定義語の存在範囲が明確であるが、日本語の場合は、単語そのものの境界が不明確なため、未定義語の混入が誤った単語切り出しを招く可能性もあり、より大きな問題になる。この解決手法は明らかになっていない。

ＧＬＲ法に基づく日本語解析の未定義語処理の研究として植木[3]がある。これは、字種の情報を用いて、未定義の固有名詞を推定するものである。この方式は固有名詞の処理に関して一定の成功を得ているが、他の多くの未定義語出現の場合にも対応しなければならない。

そこで我々は、ＧＬＲ法に対する未定義語処理の一般的な枠組を得るという目的と、未定義語獲得にコーパス上での統計的処理を適用し、その統計的処理とＣＦＧモデルに基づく自然言語パーザの処理を融合的に行なう手法を確立するという目的で研究を行なっている。本稿では、ＭＳＬＲパーザをベースとして、ＥＤＲ日本語辞書[7]とＥＤＲ日本語コーパス[7]を利用した実験を行ない、手法の有効性について確認をしたので報告する。

2．未定義語処理方針

解析精度向上のため、未定義語処理の実現について、統計的な情報の利用による次のような方針を考えた。

(1) まず、辞書引き時点で未定義語の可能性がある部分にマーキングをし、その情報を可能な限り利用する。これは植木の方式に相当する。
(2) 処理が生き詰まった時点で、過去に遡って、未定義語処理を行なう。
(3) 遡った時点からＬＲ表より遷移可能な非終端記号をリストアップし、未定義語の品詞を推定する。また、コーパスを用いて事前に得た、表記、及び品詞の統計情報から未定義語の範囲を推定する。
(4) また、未定義語の開始時点の選択に、ある種のヒューリスティクスを導入して統計処理の適用範囲を制限した。

3．実験

前述の方針に基づき、システム構成を図１のように考えた。本実験では未定義語処理プログラムが行なう処理を少数の例でシミュレートし有効性を確認した。

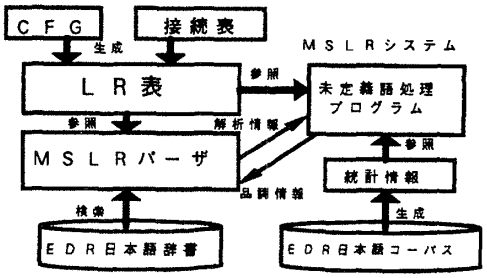


図１ ＭＳＬＲシステム構成図

辞書として、ＥＤＲ日本語辞書(約 25 万語)を用いた。文法は植木[3]の文節文法 904 ルールを用いた。ＬＲ表は植木の文法より生成し、接続表の制約を組み込んだ。尤度推定にはコーパスからの統計情報を利用する。本実験ではタグ付きコーパスであるＥＤＲ日本語コーパス(約 21 万例文)を使用した。n-gram 情報として形態素の 5-gram を長尾[6]の手法を応用して収集した。n=5 としたのは計算機の処理能力制限によるもので特に意味はない。以下、解析が行き詰まった場合の処理方法について説明する。(簡単のため n=3 で説明する。)

(1) 辞書には存在せずコーパスに存在する語の処理

解析失敗時までに作成されているグラフ構造化スタック(GSS)の各ステージからＬＲ表を検索し、それぞれ遷移可能な非終端記号リストを作成する。文末から表記の一致する形態素 n-gram 情報を検索する。その中で先頭のカテゴリが先のカテゴリリストとも一致するものを非終端記号候補として解析を続行する。解析後、最小コストのパスを最尤の結果とする。

上記を確認するための実験では、ＥＤＲコーパスからランダムに選んだ 100 文(n-gram 抽出からは除いた)をＭＳＬＲパーザで解析した。未定義語処理なしでは 15 文が解析に失敗し、15 文中の 17 形態素が辞書に存在しなかった。この手法を使うと 17 形態素中 16 形態素までは形

*Morphological and Syntactic LR

A study of unknown word processing in MSLR parser
Hiroyasu Itsui[*1], Masahiro Ueki[*2], Takeyuki Aikawa[*3], Kenji Sugiyama[*1], Hozumi Tanaka[*2].
[*1]Japan Electronic Dictionary Research Institute, LTD. [*2]Tokyo Institute of Technology. [*3]Mitsubishi Electric Corporation.

態素 n-gram からカテゴリと表記が一致し、情報の抽出ができた。できた例は平仮名表記が多く存在した。できなかった1例は「ムチ打たれた」である。

(2) 辞書にもコーパスにも存在しない語の処理

未定義語のうち、コーパスにも存在しない形態素の推定について「交差点で事故ったらしい。」という入力文を具体例として説明する。このとき、「事故る(動詞)」が無いため、解析に行き詰まる。このときの GSS を図2に示す。

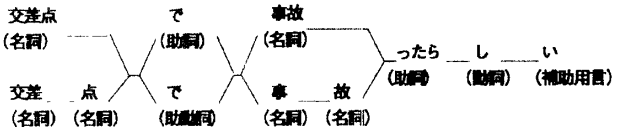


図2 解析失敗時の最長ＧＳＳ

文末からの部分文字列に表記が一致する形態素 n-gram のデータを検索する。文字列「事故った」はコーパス中には無いため完全一致では失敗する。そこで、後方の 2-gram は表記で入力文と一致をとる。先頭 1-gram は、品詞と表記の長さに置き換えて、品詞はＬＲ表からの推定品詞と一致をとり、表記の長さは未定義語の推定範囲の情報として用いる。また、推定品詞は、普通名詞、固有名詞、サ変名詞、形容詞、形容動詞、動詞の6品詞に制限した。この候補例を表2に示す。

表2 推定候補

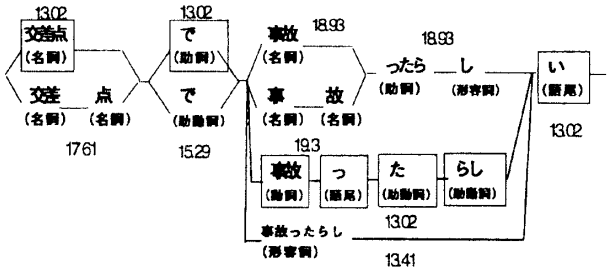
ステージ	推定表記	推定品詞	後方2gramより得た形態素列
9	し	(形容詞)	い (語尾) 。 (記号)
6	ったらし	(形容詞)	い (語尾) 。 (記号)
5	故ったらし	(形容詞)	い (語尾) 。 (記号)
4	事故	(動詞)	っ (語尾) た (助動詞)
4	事故ったらし	(形容詞)	い (語尾) 。 (記号)
3	で事故	(動詞)	っ (語尾) た (助動詞)
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

これらの候補各々について、形態素列のコーパス中での生起確率からコストを計算する。部分文字列Wの1つの形態素分割結果を $w_1 \cdots w_n$ とすると、生起確率 P(W) は、式(1)で近似できる。文Sが、m個の部分文字列で分割される場合のコスト C(S)は、式(2)よりコスト最小法に合わせて算出する。

$$P(W) = \max_{w_1 \cdots w_n \in W} \prod_{i=1}^{n} P(w_i | w_{i-(n-1)} \cdots w_{i-1}) \qquad (1)$$

$$C(S) = \sum_{j=1}^{m} \log\left(1/P\ (W_j)\right) \qquad (2)$$

また、効率化のためヒューリスティクスにより枝刈りをする。ここでは自立語の語頭は拗音の間では区切れないのでステージ6が刈られる。文全体のコスト評価の結果ステージ4で事故(動詞語幹)が最適解として選択される。図3にGSSの状態を示す。



注：数値はその候補を選択した場合の文の最小コスト値を示す

図3 未定義語処理後のＧＳＳ

以上により、入力文中に、辞書にはないような単語が含まれていても、その単語の範囲と品詞を推定できる。

4．おわりに

統計的情報を利用した処理をＭＳＬＲパーザに導入した。小規模な範囲での実験では、本処理方法が有効である見通しを得た。今後の課題として以下があげられる。

1. n-gram の生起確率を用いて確率の近似計算をした。構文木の確率計算方式として計算式を一般化する。ＬＲ表への確率の付与を検討する。
2. 形態素 n-gram ではスパースネスの問題がある。スムージング処理の導入、さらには、タグ付きコーパスより入手容易な大量のタグなしコーパスを利用可能なように方式を拡張する。
3. 未定義語があるにも拘わらず解析が終了してしまう場合がある。この場合の対処について検討する。

また、本研究は創造的ソフトウェア育成事業の「知識ベース増殖のためのソフトウェア」の一環として行なった。

参考文献

[1] 田中他：自然言語解析の新しい方法－ＬＲ表工学の提案(1)，人工知能学会研究会資料 SIG-J-9501-1(12/8)，(1995).

[2] H.Tanaka : Integration of Morphological and Syntactic Analysis based on LR Parsing, Journal of Natural Language Processing, 2, 2 ,pp.59-74(1995)

[3] 植木他：EDR 辞書を用いて日本語文の形態素解析と統語解析を行なうシステム，EDR 電子化辞書利用シンポジウム，pp33-39(1995).

[4] 斎藤：一般ＬＲ構文解析法におけるエラー処理，情報処理学会論文誌，Vol.37,No.8,pp.1506-1513 (1996).

[5] 今井他：一般化ＬＲ構文解析法による文中の複数箇所の誤りの検出と修正，言語処理学会第2回年次大会，pp.153-156，(1996).

[6] 長尾他：大規模日本語テキストのｎグラム統計の作り方と語句の自動抽出，自然言語処理 96-1(1993).

[7] ＥＤＲ電子化辞書仕様説明書 第2版(1995).