知識ベース増殖のための日本語解析

伍井 啓恭[†] 植木 正裕[‡] 田中 穂積[‡]

[†] 日本電子化辞書研究所 [‡]東京工業大学

1. はじめに

インターネットをはじめとする計算機ネットワークが世界的な規模で拡大している。このような状況下では、ネットワーク上に存在する情報を自在に扱う技術が必要となる。ネットワーク上の情報の中でも特に量が豊富なのが、自然言語で記述された文書の情報である。これらを自在に利用するためには、情報検索、情報抽出、及び機械翻訳といった自然言語処理技術が重要となる。

しかし、ネットワーク上の文書情報は、従来の自然言語処理技術では扱いにくい以下の性質がある。

(1) 文書の分野が広範囲にわたる
(2) 新用語や新概念が発生する

これらの性質に対応するためには、各々の分野や、新しく発生する用語、及び概念に関する知識を新たに獲得する必要がある。しかし、これらの知識を人手で獲得するとなると、多大な労力を要してしまう。

そこで、広範な分野の文書、及び新用語や新概念を含む文書に対して、文書中の各文を機械的に意味のある単位（単語）に分割し、各単語の関係（構文構造）を取り出す解析処理が重要となる。

本稿では、対象を単語の間に空白を置かない言語（膠着語）の日本語とし、入力文中に含まれる辞書に存在しない単語（未定義語）の抽出法の一検討として、ＭＳＬＲシステムをベースとした未定義語抽出の実験を行ない、手法の有効性について確認したので報告する。

我々は、本手法により語彙知識を獲得し、知識ベースに蓄積するとともに、さらに漸進的に知識を獲得することを検討している。

2. 従来の未定義語処理

まず、従来の未定義語の処理と課題について述べる。

2.1 解析システム上の処理

解析システム上で未定義語を処理する方法として、辞書検索失敗位置からの最短文字列を未定義語として処理を続けるもの[9]がある。辞書検索失敗時点で処理するため正しい未定義語が得られない場合がある。

また、未定義語を考慮した解析は処理量を増大させることから、計算量の観点で処理量を軽減する試みとして、解析を多段階に行なうもの[10]、字種情報からの文節末の可能性を用いてコスト最小法の処理を効率化したもの[11]があるが、計算量と解析精度にはトレードオフがある。

さらに、形態素解析後の結果を用いて、未定義語を抽出するものとして、形態素解析結果の接辞等の情報を用いて固有名詞を抽出するもの[12]、コーパス近傍のパターンマッチ走査により複合語の抽出をするもの[13]がある。本来なら未定義語の存在により一意に決定することが難しいはずの形態素解析結果を無理に求めてそれを処理することになるので、対象を限定しない限り一般的な枠組みでの未定義語処理は難しいと思われる。

2.2 ＧＬＲ法を用いた処理

ＣＦＧモデルに基づく解析手法について、これまで多くの研究がされている。それらの中でも、ＧＬＲ法は、処理効率、及び拡張性の面で優れている[1]。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法を用いた解析において、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、このギャップ埋め処理は、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解析ステージを以前に 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、未定義語であっても単語境界が明確であるが、日本語の場合は、単語そのものの境界が不明確なため、未定義語の一部を含む長い単語として区切られたり、短い単語への分割などによって未定義語位置よりもかなり先まで解析が進んでしまう可能性もあり、より大きな問題になる。

2.3 統計的手法による未定義語の抽出

辞書を用いずにヒューリスティックスやコーパスなどから単語を推定する方法がある。

n-gram 統計を用いたもの[6],[7]、正規化頻度を用いるもの[8]があるが、文法や辞書の言語的な知識を用いていないため、精度良く単語抽出するためには、良質のヒューリスティックスや大量のコーパスが必要となる。

統計的手法による未定義語抽出処理は大量のコーパスからバッチ的に未定義語を獲得する処理であり、辞書を用いた解析時に生じる未定義語処理とは目的が異なるが、これらの辞書を用いないで得られた結果を用いて、形態素解析時に遭遇した未定義語候補の確からしさを確率的に推定できれば未定義語処理の精度を向上することが可能となると考えた。

我々は、ＣＦＧモデルに基づく解析手法として、冨田によるＧＬＲ法を用いるＭＳＬＲ (Morphological and Syntactic LR) システム[3]上に、未定義語処理の枠組を導入し、解析失敗までにパーザから得られた情報と、あらかじめコーパスから獲得した統計情報の両方を利用し、入力文中の未定義語の範囲と品詞の推定を行なう手法を検討した。

3．ＭＳＬＲシステムについて

未定義語処理の枠組を導入したＭＳＬＲシステムの構成について述べる。システムの構成を図1に示す。

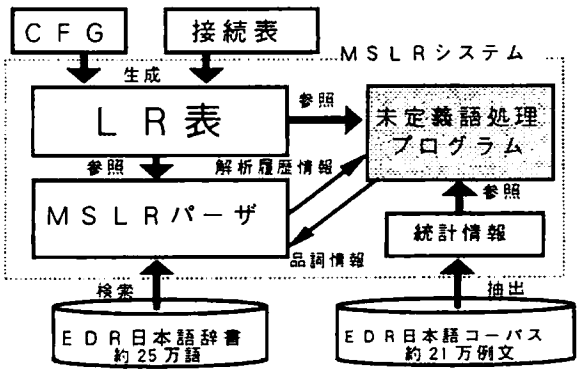

図1 ＭＳＬＲシステム構成図

ＭＳＬＲシステムは、辞書としてＥＤＲ日本語単語辞書(約 25 万語)[14]を用い、日本語の形態素解析・統語解析を行なうシステムである。解析は、ＧＬＲ法をベースにした方法で行なう。

ＭＳＬＲシステムでは、形態素レベルでの制約として形態素間の接続情報を、統語レベルでの制約としてＣＦＧ形式の文法を利用する。ＧＬＲ法による解析では、文法（約 900 ルール）からあらかじめ作成したＬＲ表を参照することで解析動作を決定する。ＭＳＬＲシステムでは、ＬＲ表の作成時に形態素間の接続制約を組み込むことで、ＬＲ表上に形態素レベルの制約と統語レベルの制約を統合し、パーザ自体には変更を加えることなく、形態素解析と統語解析の統合が行なえる。

ＥＤＲ日本語単語辞書の各単語には、左右１つずつ連接属性が付与されている。これは、例えば、用言の活用形のように、隣接する形態素との接続の違いを表すラベルである。ＥＤＲ辞書では、左右それぞれ約 100 の連接属性が定義されている。ＭＳＬＲシステムでは、品詞と左右連接属性を組み合わせることで、品詞をより細分化したものとして細品詞を定義している。

文法を作成する際に細品詞レベルまで記述を行なうと、ルール数も多くなり作成者の負担も大きくなる。ＭＳＬＲシステムでは、形態素間の接続制約を細品詞間の接続制約として定義することで、統語レベルでは区別する必要のない細品詞

をまとめて扱うことができる。すべての細品詞は、文法中の細品詞規則と呼ばれるユニットルールにより一意に決まる品詞カテゴリに分類される。文法はこのレベルの品詞カテゴリの上で記述される。ＬＲ表を作成する際に細品詞間の接続制約を組み込むことで、文法上は接続して見えるが、実際には接続が不可能な細品詞の組合せによる解析動作は削除される。

４．未定義語処理方針

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) 解析が行き詰まった時に処理を開始する。
(2) 品詞の推定には、LR 表上で解析動作が可能な品詞のリスト、および、n-gram から得られる品詞の情報を用いる
(3) 範囲の推定には、生成済みの構文木情報、および、入力文の表記と n-gram との一致する範囲の情報を用いる
(4) 字種情報などのヒューリスティックスを用いて、不必要な候補は削除する

５．未定義語処理手順

前述の方針に基づき、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。

尤度推定にはコーパスから得た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)[14]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) 辞書にはなく、コーパスにある語の処理

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)の各ステージのスタックトップの状態からＬＲ表を検索し、それぞれ遷移可能な品詞リストを作成する。文末から表記の一致する形態素 n-gram 情報を検索する。その中で先頭の形態素の品詞が先の品詞リストに含まれていれば推定品詞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システムで解析した。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。この例では、「事故る(動詞)」がないため、解析に行き詰まる。このときの GSS を図 2 に示す。

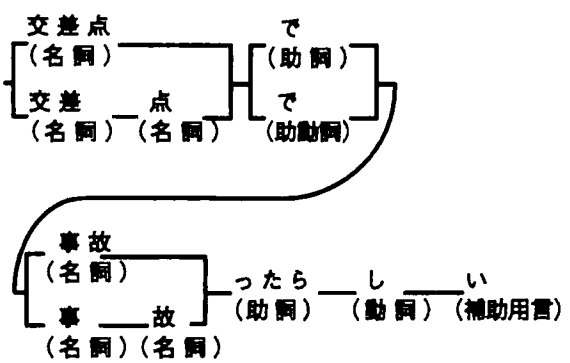

図２ 解析失敗時の最長ＧＳＳ

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」はコーパス中にはないため完全一致では候補が見つからない。

そこで、後方の 2 形態素は表記で入力文と一致をとる。先頭 1 形態素は、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範囲推定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１に示す。

表1　推定候補

ステージ	推定表記	推定品詞	後方2形態素より得た形態素列
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	(動詞)	た (助動詞) らし (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。

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

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

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促音の前では区切れないのでステージ6が刈られる。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。

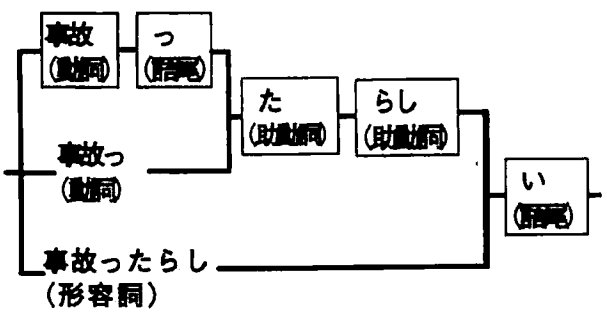

図3　未定義語処理後追加されたＧＳＳ

以上により、入力文中に、辞書にもコーパスにもない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。

6．おわりに

統計情報を利用した未定義語処理をＭＳＬＲシステムに導入した。小規模な実験では、本手法が有効であるという見通しを得た。

我々は、本手法により語彙知識を獲得し、知識ベースに蓄積するとともに、さらに漸進的に知識を獲得することを検討している。

今後の課題として以下があげられる。

1. 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拡張する。

2. 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開発」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] Tanaka, H., et al.: 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，情報処理学会自然言語処理研究会 96-1，(1993).

[7] 森他：nグラム統計によるコーパスからの未知語抽出，信学技報 NLC95-8，pp.7-12 (1995).
[8] 中渡瀬：正規化頻度による形態素境界の推定，情報処理学会自然言語処理研究会 113-3，pp.13-18(1996)
[9] 電子技術総合研究所推論機構研究室：拡張LINGOL，P.9 (1978).
[10] 大場他：未定義語を含む文の多段階構文解析解析法，情報処理学会自然言語処理研究会 70-4,pp.1-8 (1989).
[11] 吉村他：未登録語を含む日本語文の形態素解析，情報処理学会論文誌 Vol.30 No.3，pp.294-301 (1989).
[12] 木谷：固有名詞の特定機能を有する形態素解析処理，情報処理学会自然言語処理研究会 90-10，pp.73-80 (1992).
[13] 久光：文書走査を用いた複合名詞解析について，情報処理学会自然言語処理研究会 112-2，pp.7-14 (1996).
[14] ＥＤＲ電子化辞書仕様説明書 第２版 (1995).