自然言語処理 121－5
（1997. 9. 11）

ＭＳＬＲパーザにおける未定義語処理の一方式

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]〒101 東京都千代田区神田佐久間河岸７８番地１第二阿部ビル
[*1]{itsui,kenji}@edr.co.jp

要旨
自然言語解析において入力文中に現れる未定義語の扱いは、大きな課題の１つである。特に単語を分かち書きしない日本語では、未定義語の範囲を推定するのが困難である。そこで我々は形態素解析、及び統語解析を統合したことを特長とするＭＳＬＲパーザをベースに未定義語処理の改良を試みた。未定義語の推定に、ＬＲ表の状態遷移から得られる品詞情報と、コーパスから予め抽出したvari-gramの統計情報を用いて処理を行なった。ＬＲ表は人手で作成したＣＦＧルールから生成した。辞書として、ＥＤＲ日本語辞書[7]、統計情報抽出用のコーパスとして、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した。
キーワード 未定義語、コーパス、vari-gram、ＭＳＬＲ

A method of unknown word detection 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.
[*1]Daini-abe building 78-1,Kanda-sakumagashi, Chiyoda-ku tokyo 101,Japan
[*1]{itsui,kenji}@edr.co.jp

Abstract
Unknown words cause a great problem in natural language analysis. It is difficult to find out unknown words in Japanese because it has no blank between words. This paper proposes an unknown word deduction method that is implemented in MSLR (Morphological and Syntactic LR Parser) developed at Tokyo Institute of Technology. Our method can determine range and part-of-speech of unknown words by using LR action table and vari-gram statistics from EDR corpus. Experiments have shown the effectiveness of our method.
key words unknown word, corpus, vari-gram, MSLR

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。ネットワーク上の情報の中でも特に大量で重要なものとして、自然言語で記述された文書の情報がある。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。

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した自然言語処理をするためには、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、形態素解析、及び統語解析をして、文の形態素情報、構文構造情報を取り出す処理が重要となる。

本稿では、単語の間に空白を置かない言語である日本語を対象とし、未定義語（入力文中に含まれる辞書に存在しない単語）を抽出する新しい方法を提案する。

次にこの方法をＭＳＬＲシステム[3]に組み込み、未定義語抽出の実験を行ない、提案した方法の有効性を報告する。

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 手続き的な未定義語の処理

手続き的に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。この処理では辞書検索失敗時点で処理がなされるため、正しい未定義語が得られない場合がある。

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]がある。解析過程を多段階に分けて横形探索することにより未定義語抽出の再現率を高めているが、適合率を高める手段については明らかにしていない。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。長単位の付属語列を精度よく扱うことに課題がある。

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。

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がある[4]。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可能性がある。これについてスコア付けによる試行回数を制限する方法が提案されているが具体的にギャップに対してどのようなスコア付与をすべきかについては明らかでない。

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が提案されている[5]。この例では対象は英語である。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を考慮する必要がある。

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。

2.4 コーパスに基づく日本語形態素解析

近年、日本語においても大量のコーパスが整備され、これに伴いコーパスを用いて確率的な言語モデルを構築し、形態素解析システムに取り込む研究がされている[15][16][17]。従来 n-gram (bi-gram や tri-gram)の固定長の並びを扱うモデルが主流であった。しかし、bi-gram では2つの形態素間の接続を扱うだけで、それ以上の言語現象を扱うのが困難であり、n の値を大きくするとパープレキシティは減少すると予測されるが、カバー率の低下や標本規模の減少により統計的信頼性が失われるという課題がある。

これに対して最近 vari-gram の言語モデルを用いた研究がされている[18]。これは言語モデルの複雑さにあわせて任意の長さの連接規則を適用するものである。類似な処理に連語の登録手法があるが、単語から品詞へクラスを拡大する等の上位階層性を含めたモデル化可能である部分が異なる。

[18]では、手作業で規則を構築してるため、段階的にシステムの精度を向上できるが、連鎖がより長くなってくると人手で規則を構築するのに限界が生じてくる。

2.5 我々の未定義語処理

我々は、ＣＦＧモデルに基づく解析手法として、ＧＬＲ法を用いた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 vari-gram の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。ＧＬＲ法の一般性と拡張性を失うことなく、n-gram と比較して標本規模を抑えることが可能な vari-gram の情報を自動的にコーパスから抽出することにより効率的に精度の高いシステムを構築した。

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。

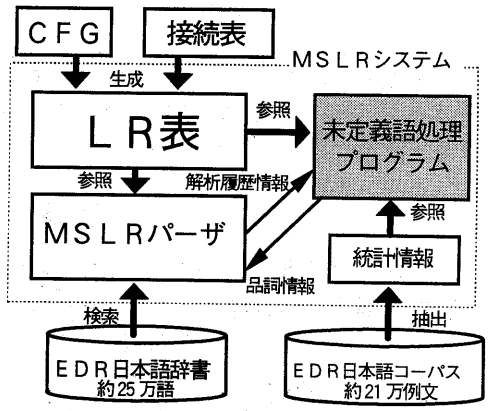

図1 ＭＳＬＲシステム構成図

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（図２）。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。

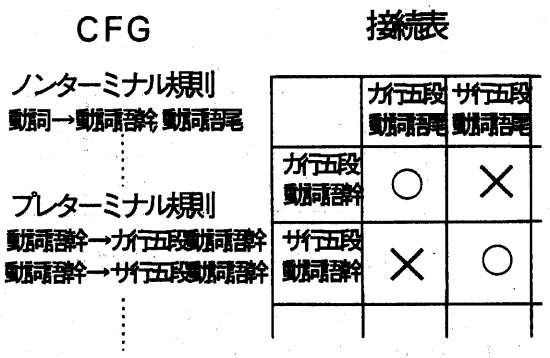

図２ 文法と接続規則の関係

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 vari-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、及び入力文の表記と vari-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、未定義語処理アルゴリズムを説明する。

尤度推定にはコーパスから得た vari-gram 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)[14]を使用した。vari-gram 情報として形態素の 1～5-gram を収集した。最大値を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。図３に具体例で示す。スタックトップが状態 i において、辞書に

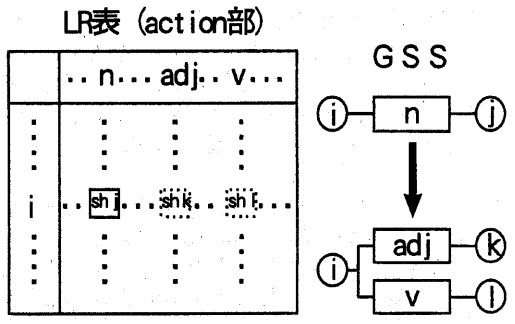

図３ ＬＲ表からのカテゴリ推定

品詞が名詞しかなかったとする。この場合にアクション sh j で解析がすすめられ解析が行き詰まっている。そこで、状態 i において名詞以外に遷移可能な品詞を探す。ここでは形容詞と、動詞が

見つかる。このような品詞を予めリストアップしておき、後vari-gramの処理で照合する。最終的に未定義語候補として残ったものはＧＳＳの情報として追加し解析を再開する。

文末から表記の一致する形態素 vari-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。

(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。このときの最長のGSSを図4に示す。

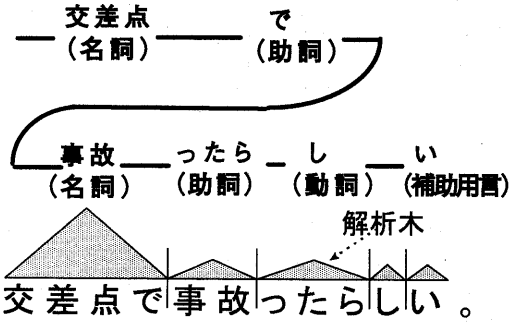

図4 解析失敗時の最長ＧＳＳ

解析の行き詰まった位置から文頭方向に遡って未定義語の先頭位置を推定する。

部分文字列に表記が一致する形態素を vari-gram のデータから検索する。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、形容詞、形容動詞、動詞の 4 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)
⋮	⋮	⋮	⋮

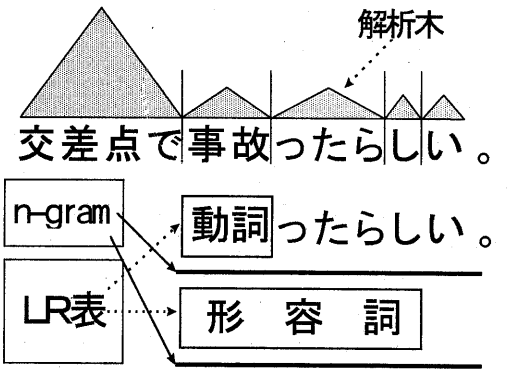

図5 未定義語推定全体構成

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。

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

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

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。

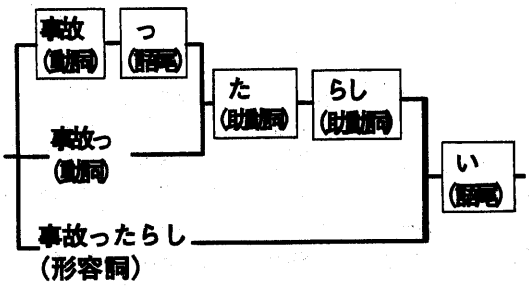

図3 未定義語処理後追加されたＧＳＳ

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．実験

人手によってタグ付けされているＥＤＲコーパスからランダムに選んだ100文(1文の平均文字長 38.2 文字)2412 形態素をテストセットとした。vari-gram 情報はテストセットを除いたＥＤＲコーパスより抽出し直した。

本実験では、ＥＤＲ日本語単語辞書に表記と品詞双方が登録されていない形態素を未定義語と定義した。品詞体系はＥＤＲコーパスとあわせてテストセット中の未定義語を抽出し、これを正解のセットとした。テストセット中には２０６個所に１５７語の未定義語が存在した。表2に正解セットの品詞別の全形態素数と含まれていた未定義語数を示す。

表2 テストセットに含まれる未定義語数

品詞	名詞	動詞	形容動詞	形容詞	その他
形態素数	899	274	26	26	1187
未定義語数	120	21	3	2	11

次に、テストセットを未定義語処理ありのＭＳＬＲシステムに入力し、出力された結果のうち表記と品詞の双方がマッチするものをカウントした。カウント結果を適合率、再現率で表3に示す。適合率、再現率の定義は下記とした。

適合率：システムの出力した正解未定義語数／システムの推定した未定義語数

再現率：システムの出力した正解未定義語数／テストセットの正解未定義語数

MSLR 正解、MSLR 出力、未定義語数がそれぞれ、システムの出力した正解未定義語数、システムの推定した未定義語数、テストセットの正解未定義語数である。

表3 未定義語推定精度

	適合率	再現率	MSLR 正解	MSLR 出力	未定義語数
出現数	79.9	55.8	115	144	206
語数	69.5	46.5	73	105	157

7．考察

再現率が低い値を示している。これは、入力文中に未定義語が存在するにもかかわらず解析が途中終了しない場合があり、この場合には、本提案の方法の処理対象とならないためである。

実験では、ＥＤＲコーパスの形態素の区切りと品詞双方に一致しなければ正解としなかったため厳しい評価基準であったが、表記と品詞の双方をＬＲ表から予測される未定義語の品詞情報とvari-gram から予測される未定義語の範囲、及び品詞の情報を用いることによる有効性が確認できた。

8．おわりに

統計情報を利用した新しい未定義語処理方法をＭＳＬＲシステムに導入した。本実験により、本手法の有効性を確認した。

我々は、本手法により未定義語を獲得し、語彙ベースに蓄積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

(注) 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。

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、解析が途中終了しないために未定義語の抽出に失敗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。

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).

[15] 森他:形態素 bi-gram と品詞 bi-gram の重ね合わせによる形態素解析，情報処理学会自然言語処理研究会 112-6，pp.37-44 (1996).

[16] 竹内他:ＨＭＭによる日本語形態素解析システムのパラメータ学習，信学技報 NLC95-9 (1995).

[17] 永田:ＥＤＲコーパスを用いた確率的日本語形態素解析，E.D Ｒ電子化辞書利用シンポジウム，pp.49-56 (1995).

[18] 北内他:日本語形態素解析システムへの可変長連接規則の実装，言語処理学会第３回年次大会発表論文集，pp.437-440 (1997).