4A−3

文節間文法を用いた未知語検出

石川永和† 伊藤彰則‡ 牧野正三†
†東北大学応用情報学研究センター ‡東北大学情報処理教育センター

1 はじめに

近年の自然言語処理システムでは文法をはじめとする各種言語情報は各単語ごとに分散記述, データベース化し, 維持・管理することが多い. しかしこのデータベース化にはコストがかかること, タスクごとに変更を迫られるなどの点から作成の自動化が望まれている.

これに当たっては大量のテキストを解析しなければならないが, 辞書未登録語（未知語）については検出を行ない, 言語情報を付与する必要がある.

本稿ではこの未知語検出に関する一方法を提案する. この方法は文節間の依存関係を表現した文節間文法を基礎とするもので, 前報告で述べた疑似文節を用いた未知語検出法に採り入れることにより検出率が向上した.

2 文節間文法

文節間文法は文節（を構成する要素）間の修飾関係を素性を媒介として表現したものである. 文を構成する各要素（文節・部分木等）には素性を元とするfeature,slotと呼ばれる2種類の集合が与えられる.featureは「それが修飾要素となる場合に結ぶことのできる修飾関係」,slotは「それが被修飾要素となる場合に結ぶことのできる修飾関係」を表す. 文中で隣接する2つの要素において, 文頭側のfeatureと文末側のslotに共通の素性が含まれる場合slot-fillingが生じ, 新たな部分木が生成される. この際, 新たな部分木のfeature, slotはそれぞれ文末側のfeature, slotと等しくなる. ただしfeatureについてはslot-fillingの起きた素性は取り除かれる. この解析の結果始端・終端がそれぞれ文頭・文末に一致し, 終端性を表す特別な素性（.Prop）をもつ文候補が文の解析結果と認定される. slot-fillingにより新たな部分木が生成される様子を図1に示す.

3 文節間文法を用いた未知語検出法

前報告では文節内文法を用いて疑似文節を生成し, 選択にコスト最小法を用いた未知語検出法を提案した. この際に文全体での統語的な整合性は考慮していなかった. ここでは検出に文節間文法を導入することにより構文的な情報を用いた. このモデルでは文法は各単語に記述されているので, 導入に当たっては未知語と仮定した語に対し, 解析用のfeature, slotを与える必要がある. これは未知語をどの品詞と仮定するかに応じてその品詞が共通に持つ素性を設定した. 本稿では未知語を名詞と仮定する. 疑似文節生成の際この素性のリストをもとに疑似文節のfeature, slotを生成し, 疑似文節を含んだ文節ラティスに対して文節間文法による解析を行なう. この結果得られた解析木に現れる疑似文節の実質語部分を未知語と推定する.

4 検出実験

この方法を用いた検出実験を行なった. 係り受け解析にはCYK法を用いた. 解析の際には生成された部分木にコストを設定し, コストの低いものを選択する. 新たに生成された部分木のコストは, 融合した要素のコストの和とする. さらに隣接の文節間の接続に対してコストを設定し, これも加えている. これらは文節のコストと文節の接続のコストより計算されるが, これらはある事象に注目した時の文節の生起確率・接続確率から決定した. 実験はいくつかのコスト設定法のもとで行なった. 結果を図2に示す. ここでは疑似文節と辞書登録語による文節候補（既知文節）の尤度の差を考慮し, 既知文節のコストを0とした場合の検出結果を示してある. さらに実験条件として1文中に現れる未知語数の制限も加えた. 現れる未知語数により解析木の枝刈りを行なった. この結果, 前報告で述べた構文情報を用いない検出法に比べ, 同様のコスト設定を用いた場合約7.6%検出率が向上し,76.5%となった. さらに未知語数の制限を行なうことにより最大

Detection of Unknown Words in the Morphemic Analysis for Construction of a Word Dictionary
Hisakazu ISHIKAWA†,Akinori ITO‡,Shozo MAKINO†
†Research Center for Applied Information Science,Tohoku University
‡Education Center for Information Processing,Tohoku University

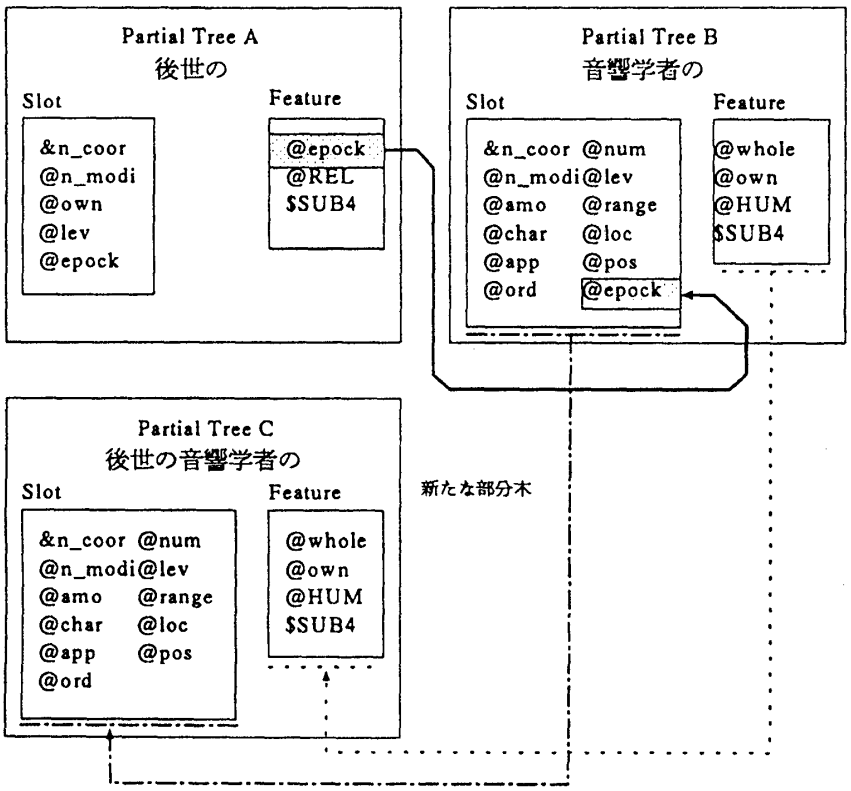

図 1: 新たな部分木の生成

85.5% の検出率を得た.

5 おわりに

本稿では, 疑似文節を用いた未知語検出に文節間文法を導入し, 統語解析を行いながら未知語の検出をする方法を提案した. いくつかの条件のもとで検出実験を行なった. 文節内文法のみを用いた検出法に比べ検出率の向上がみられた. 特に疑似文節と既知文節のコスト設定を変え, 未知語数の制限を加えた場合に大きく向上した.

参考文献

[1] 伊藤彰則: 「連続音声中からの文節の検出に関する研究」 東北大学審査修士学位論文, (1988)

[2] 伊藤彰則: 「タスクに依存しない日本語文音声の認識に関する研究」 東北大学審査博士学位論文, (1991)

[3] 石川, 伊藤, 牧野: 「文節オートマトンを用いた未知語検出法」 情報処理学会第 45 回全国大会, (1992)

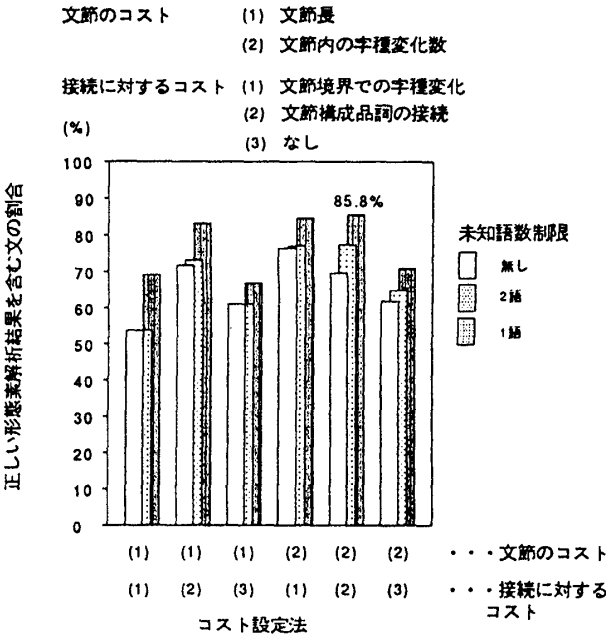

図 2: 文節間文法を用いた検出結果