

べた書きかな文の仮文節境界の 補正方法

2L-9

荒木哲郎⁺池原悟⁺⁺土橋潤也⁺⁺:福井大学工学部⁺⁺:NTT情報通信網研究所

1. はじめに

日本語文解析には形態素解析、構文解析、意味解析などの各種レベルがあるが、べた書きの日本語文に対して最初に単語や文節などの単位に、分かち書きを行う形態素解析が基本的な処理として重要である。従来、漢字かな混じり文に対する分かち書き処理としては、高精度な技術が確立されている[1]。しかし、べた書きかな文の場合は、総当たり法でかな漢字変換等の処理により生成される、あらゆる単語候補列の組み合わせを考慮して解析を試みようとすると、一般に探索木が爆発するという問題が生じる。

このような問題を解決するために、[2]では、探索木の爆発を防ぎ、一定の処理時間内で日本語文の解析を終えるために、かな漢字変換を含めた形態素解析の対象となる範囲を仮文節として一時的に定める方法が提案されている。本論文では、このような仮文節境界に対し、辞書引きや品詞接続テーブルによる接続検定等を行うことによって、仮文節境界を補正する方法を提案する。

2. 仮文節境界推定モデルと補正方法

マルコフ連鎖確率値が文字間の結合力を表すことに着目し、図1に示されるような仮文節境界推定モデルにより、仮文節境界を決定する方法が提案され、その効果が示されている[2]。また、このように設定された仮文節境界の前後±1の範囲内に正解文節境界が存在する確率は、

約92%であることが確認されている(図2)。本論文では、設定された仮文節境界に対して、単語辞書引き及び品詞接続テーブルによって、文節先頭・末尾単語の存在性及び接続性を調べることで、誤って設定された仮文節境界をどの程度検出可能であるか(主として適合率の向上)を定量的に評価する。

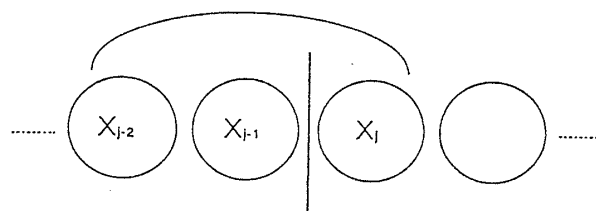


図1 仮文節境界推定モデル

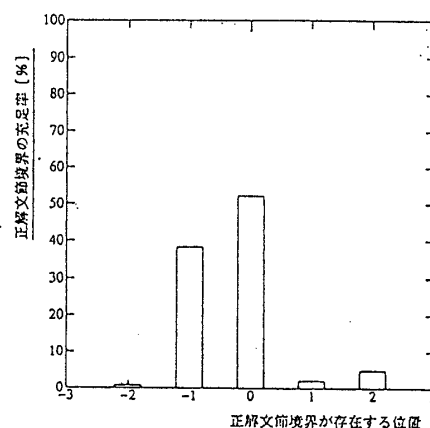


図2 設定仮文節境界の近傍に正解境界が存在する割合

3. 文節先頭及び末尾単語長並びに品詞接続情報

(1) 文節先頭・末尾単語長の分布

新聞記事による調査では、文節先頭にくる単語は、長さ6までのものが全体の99.9%、また文節末尾にくる単語は、長さ4までのものが98.9%を占めている。

(2) 文節先頭及び末尾品詞間の相互接続性

文節の先頭及び末尾にくる単語の相互接続可能な品詞の組み合わせ(品詞接続テーブル)の有効な数は、4872個であった。これは、500通り

A Method of Correcting Provisional Boundaries of "Bunsetsu"

Tetsuo Araki⁺ Satoru Ikehara⁺⁺ Junya Tsutiihashi⁺⁺:Faculty of Engineering, Fukui University⁺⁺:NTT Network Information Systems Laboratories

ある品詞活用形の組み合わせ500x500の1.9%にあたる。

4. 実験条件

(1) 入力データ

① 文の種類：新聞記事

② 字種：べた書き音節文

③ 総文章数：50文（384文節）

④ 総文字数：1977文字

(2) 使用辞書

① 仮文節境界設定：音節2重マルコフ連鎖確率

② 仮文節境界補正：単語辞書（15万語）

品詞接続テーブル

漢字2重マルコフ連鎖確率

5. 仮文節境界の補正実験結果

(1) 単語辞書引きによる単語候補の有無による仮文節境界の補正効果

文節先頭では先頭から1～6文字、文節末尾では末尾から1～4文字までのかな文字に対して、単語辞書引きを行い、単語候補の有無によって仮文節境界の補正実験を行った。その結果、文節先頭における1文字かなに対する単語候補がほとんどの場合存在することから、補正効果は適合率で約1%の向上にとどまった。

(2) 品詞接続テーブル情報による仮文節境界の補正効果

(1)の条件に加え、3.で求めた品詞接続テーブルによる補正結果を図3に示す。同図より、適合率で約4%の向上にとどまった。これは、辞書引きによって得られる単語候補の正確な認定処理が行われていないため、誤った単語候補間において、品詞の相互接続性が調べられる可能性があり、補正効果に限界が生じるためである。

(3) 漢字マルコフモデルを用いた先頭単語候補の認定による仮文節境界の補正効果

(1)、(2)の条件に加え、更に、正しい文節境界が仮文節境界の±1の範囲内に92%存在することから、仮文節境界に対して、±1の範囲の

先頭単語のすべてのかな漢字変換候補を、漢字マルコフによって順位付けを行い、上位1位又は5位以内に仮文節境界の先頭単語候補がくるか否かによって、その仮文節境界の補正を行う実験を行った（図3）。同図の結果より、5位以内で適合率約12%、1位のみで適合率約35%の向上が得られた。

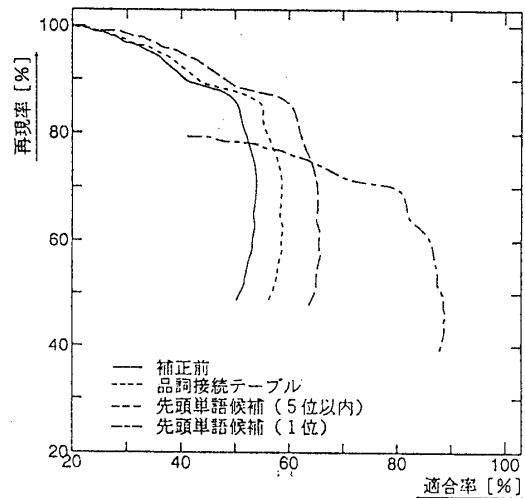


図3 仮文節境界の補正実験結果

6. おわりに

2重マルコフ連鎖確率によって推定された仮文節境界のうち、誤って設定された仮文節境界を、単語辞書引き及び品詞接続テーブルによって検出する方法を提案し、その有効性を定量的に評価した。その結果、単語の存在可能性及び品詞接続テーブルによる接続検定では、4%程度の向上と補正効果は小さいが、漢字マルコフによる文節先頭単語候補の認定を行うことによって、最大適合率で約35%の向上が得られた。今後の課題としては、誤った仮文節境界の検出だけでなく、正しい仮文節境界の設定方法が挙げられる。

参考文献

- [1] 宮崎, 大山: "日本文音声入力のための言語処理", Vol. 27, No. 11, pp1053-1061 (1986)
- [2] 土橋, 荒木, 池原: "2重マルコフ連鎖確率を用いたべた書き日本語文の文節境界推定", 信学会春季大会, Vol. 6, No. D-102, pp104 (1993)