マルコフモデルによる言い直し対象の文字列の検出について

荒木　哲郎*　　池原　悟**　　橋本　昌東*
*福井大学工学部
**ＮＴＴコミュニケーション科学研究所

1. はじめに

より制約のない発話の認識を考えた場合、発話自体が非文法的な場合があり、文法をベースとした従来の解析法をそのままでは適用できないという言語処理側の問題がある。発話を非文法的にしている要因としては、「えーと」に代表される冗長語の挿入、言い直し、語の省略や倒置があるが、これらは人間同士のコミュニケーションにも現れるものであり無視できない[1][2]。

これらのうち、本論文で対象とする言い直しについては、その音響的特徴や単語・品詞レベルの構造的特徴を用いて検出・訂正を行うことがこれまでに試みられている[2][3]。しかしながら、まだ言い直しを正しく解析する技術が確立されるには至っていない。

本論文では、これまでに音節マルコフモデルを用いた日本語文の解析法として提案されている、誤り検出法[4]及び仮文節境界の推定法[5]を拡張し、音節会話文の中から言い直しの対象となる文字列を検出する方法を提案し、その有効性を実験により定量的に評価する。

2. 日本語会話文における言い直しの特徴

ＡＴＲ言語データベースの旅行に関する会話文における言い直しの対象となっている文字列634個について、文字列の正しい文に対する出現位置とその形を調査した結果を以下に示す。ただし、本論文では、言い直しの対象となる文字列を（）で囲んで示す。また、言い直しの対象となる文字列の先頭位置及び末尾位置をそれぞれ開始点、終了点と呼ぶ。

2.1 言い直しの対象となる文字列の出現位置

言い直しの対象となる文字列の正しい文に対する出現位置を大別すると、次の3つのタイプに分かれる。

[1]文節境界に一致しているもの　80%
・日本の（芸能を）伝統芸能を...
・（い）一度発券しますと、...
[2]単語境界に一致しているもの　17%
・ホテルのお医者様（を）と一緒に...
・...をお待ち（し）致しております。
[3]単語境界にも一致しないもの　3%
・フィリッ（ト）プスと申しますが...
・...露天風（が）呂があると...

このことから、言い直しの対象となる文字列のうちの80%は、文節境界に一致した位置に出現していることがわかる。

本論文では、まず第1ステップとして[1]のタイプの文字列の検出を検討する。[2]のタイプ及び[3]のタイプの文字列の検出は、文節内においてさらに、その文字列を検出するという問題に帰着できると考えられるので、今後の課題としたいと思う。

2.2 言い直しの対象となる文字列の形態

今回対象とする[1]のタイプの文字列を、その形からさらに大別すると、次の2つのタイプに分かれる。

[A]文節を構成するもの　14%
・お席の（料金は）料金も...
・ホテルの（予約で）予約を...
[B]文節の断片を含むもの　66%
・よい（お座敷）お座敷を...
・（公園の前）公園の間を...

このことから、言い直しの対象となる文字列が文節境界に出現するもののうち66%は、途中で中断した形になっており、文節の断片を含んでいることがわかる。

3. マルコフ連鎖確率による言い直し対象の文字列の検出法

本論文では、言い直しの対象となる文字列の出現位置が文節境界に一致する[1]のタイプの場合について、その検出方法を述べる。検出の流れは図1に示すように、まず文の全体にわたって仮文節境界の推定法を適用する。ただし、この段階では開始点か終了点かそれ以外の文節境界かといった区別はなされない。そして、[1]のタイプのうちの[B]のタイプについては、さらに終了点を特定を試みる。

そこで、3.2節で[1]のタイプ全てに適用できる仮文節境界の推定法を、3.3節、3.4節でそのうちの[B]のタイプに適用できる終了点の検出法を述べる。

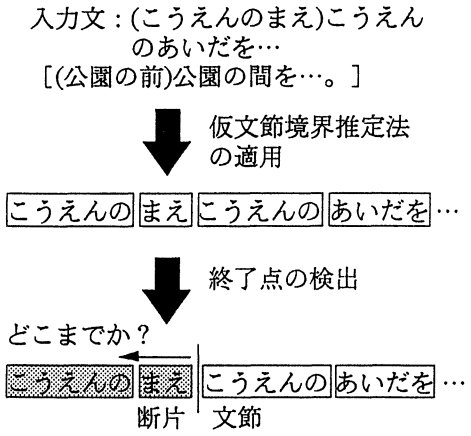

図１　言い直しの対象文字列の検出

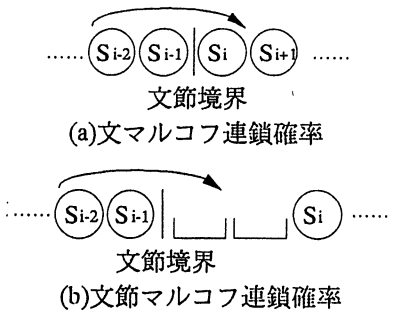

図２　文および文節マルコフ連鎖確率の学習法(2重マルコフ)

3.1 マルコフ連鎖確率

まず、言い直し対象の文字列を検出するために用いる2種類のマルコフ連鎖確率を定義する（図２）。

【文および文節マルコフ連鎖確率】
文単位の音節会話文について統計をとって求めたマルコフ連鎖確率を文マルコフ連鎖確率、文節単位の音節会話文についてのものを文節マルコフ連鎖確率と呼ぶ。 ■

3.2 仮文節境界推定の適用法

ここでは、記述文を中心とした日本語べた書きかな文の文節切りを行う方法として提案されている、文節マルコフ連鎖確率を用いた仮文節境界推定法を[1]のタイプの文字列が存在する会話文に適用して、言い直し対象の文字列の開始点・終了点及び文節境界を推定する方法を述べる。仮文節境界推定の基本的な判定法を図3に示す。

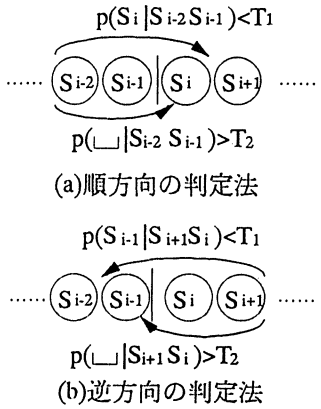

図３　仮文節境界推定法の基本的な判定法(2重マルコフ)

図1のように、[1]のタイプの文字列の開始点・終了点は、文節の断片と文節が接する境界となっている場合がある。そして、その文節の断片の中の文字列は、文節や単語として成り立たない文字列や誤った文字列を含むため、判定法によっては、文節境界では有効に働くが、開始点・終了点において有効に働かないものがある。そこで、順方向と逆方向の判定法をor条件で併用することを考え、この問題の解決を図る。

3.3 文マルコフ連鎖確率をによる言い直し対象の文字列の終了点の検出法

[B]のタイプの言い直し対象の文字列は、その終了点の前後の音節を組み合わせた音節列についての文マルコフ連鎖確率をとった場合、値が悪くなることが期待される。このことを用い、終了点を図4のように検出する。

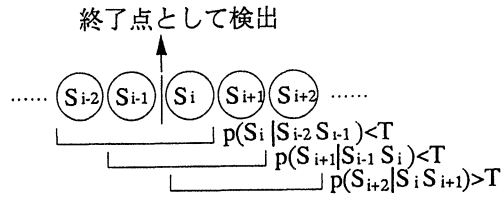

図4　文マルコフ連鎖確率による終了点の検出法(2重マルコフ)

3.4 文節マルコフ連鎖確率による言い直し対象の文字列の終了点の検出法

[B]のタイプの言い直し対象の文字列は、その終了点が文節の断片と文節の境界になっている。このことを用い、終了点を図5のようにして検出する。すなわち、後に続く文節の先頭文字列に空白を付加したマルコフ連鎖確率は良くなるが、文節断片の文字列に空白を付加しても確率は良くならない。

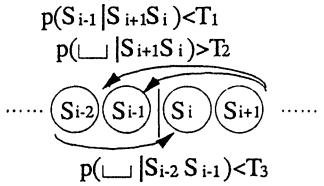

図5　文節マルコフ連鎖確率による終了点の検出法(2重マルコフ)

4. 実験条件

仮文節境界の推定法を適用する実験入力文としては、次のようなものを用いた。

[1]文数:標本外100文(ホテルへの問い合わせ)
[2]言い直し対象の文字列:[1]のタイプ
[3]文節境界数:728
[4]開始点数:89
[5]終了点数:107

言い直し対象の文字列の終了点検出の実験入力文としては、次のような文を用いた。

[1]文数:標本外80文(ホテルへの問い合わせ)
[2]言い直し対象の文字列:[B]のタイプ
[3]終了点数:83

5. 実験結果

仮文節境界の推定および終了点の検出の精度は、再現率・適合率で評価し、その積が大きいほど精度は良いものと考える。ただし、下の定義式における境界とは、実験(1)では言い直し対象の文字列の開始点・終了点、文節境界これら全てを、実験(2)では言い直し対象の文字列の終了点のみを対象とする。

$$再現率=\frac{設定された境界の正解再現数}{正解境界数}$$

$$適合率=\frac{設定された境界の正解再現数}{設定境界数}$$

5.1 実験(1):仮文節境界推定法の適用

[1]のタイプ全てに適用できる仮文節境界の推定について、結果を表1、図6に示す。同結果より、(1)どの判定法でも2重に比べて3重の方が精度が良い。言い直し対象の文字列の開始点・終了点の再現率にも注目するならば、(2)順方向と逆方向の判定法のor条件での併用において、再現率と適合率の積が最大のとき、言い直し対象の文字列の開始点の再現率82.0%・終了点の再現率73.8%となることがわかり、言い直し対象の文字列の開始点・終了点は、仮文節境界推定法でほぼ推定できることが分かる。しかし、どの境界が開始点・終了点かを特定することは難しい。

判定法	再現率(全体)	適合率(全体)	開始点再現率	終了点再現率
順方向(3重)	65.3	80.4	73.0	17.5
逆方向(3重)	67.7	86.6	36.0	70.9
順or逆(3重)	87.0	77.6	82.0	73.8

表1　主な推定結果

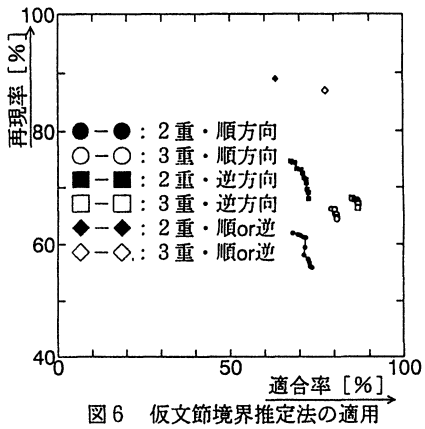

図6　仮文節境界推定法の適用

5.2 実験(2):言い直しの対象となる文字列の終了点の検出

音節マルコフモデルによっての検出が難しい開始点を求めるには、終了点を中心として言い直し対象の文字列と言い直しの結果の文字列のマッチングをとる必要がある。その第1ステップとなる終了点の検出の結果について述べる。終了点検出の第1の方法である文マルコフ連鎖確率を用いた方法について、結果を図7に示す。2重に比べて3重の方が精度が良く、再現率と適合率の積が最大のとき、再現率49.4%・適合率22.8%であった。

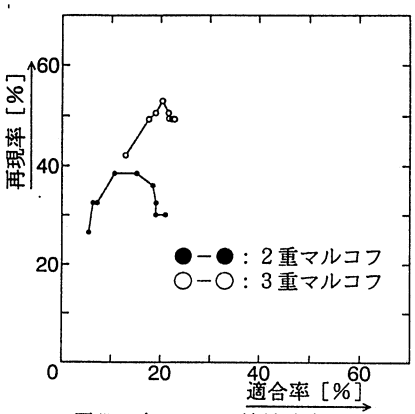

図7　文マルコフ連鎖確率による終了点の検出

終了点検出の第2の方法である文節マルコフ連鎖確率を用いた方法について、結果を図8に示す。2重に比べて3重の方が精度が良く、再現率と適合率の積が最大のとき、再現率60.2%・適合率39.4%であった。

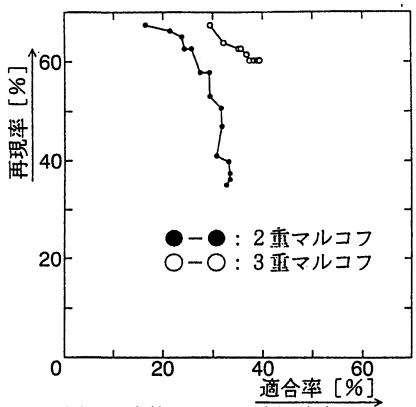

図8　文節マルコフ連鎖確率による終了点の検出

6. まとめ

本論文では、文及び文節マルコフモデルを用いて音節会話文から言い直し対象文字列を検出する方法を提案し、その有効性を実験により評価した。その結果、以下の知見を得た。

[1]仮文節境界推定法による言い直しの対象文字列の開始点・終了点及び文節境界の推定精度は、再現率87.0%・適合率77.6%(開始点の再現率は82.0%・終了点の再現率は73.8%)である。

[2]文マルコフモデルによる言い直し対象の文字列の終了点の検出精度は、再現率49.4%・適合率22.8%である。

[3]文節マルコフモデルによる言い直し対象の終了点の検出精度は、再現率60.2%・適合率39.4%である。

今後の課題としては、本手法の精度の向上、本論文で対象とした[1]のタイプの開始点の検出が挙げられる。

謝辞

本研究を行う上で、会話文データの調査に対してご協力いただいた、ＡＴＲ音声翻訳通信研究所データ処理研究室の森本室長並びに当研究室の方々に感謝致します。

参考文献

[1]村上,嵯峨山:"自由発話音声認識における音響的および言語的な問題点の検討",NLC 91-57,SP91-100

[2]中川,小林:"自然な音声対話における間投詞・ポーズ・言い直しの出現パターンと音響的性質",日本音響学会誌,vol.51,No.3

[3]ヒーマン,ローケンキム:"構造情報を用いた言い直しの検出",NLC95-56,SP95-51

[4]荒木,池原,塚原:"2重マルコフモデルによる日本語文の誤り検出並びに訂正法",NL94-7

[5]荒木,池原,土橋,笹島:"3重マルコフモデルによるべた書きかな文の仮文節境界の推定法",NL102-14