日本語音声会話文ラティスからのキーワード候補の抽出法

荒木 哲郎
福井大学 工学部

池原 悟
NTTコミュニケーション科学研究所

四方 啓智
福井大学 工学部

橋本 昌東
福井大学 工学部

1 はじめに

会話文は記述文に比べると、話す対象が特定化され音声などによって表現されることが多いために、文のつなぎとしての冗長語が挿入したり、いい直しや語が省略されたり、また倒置が行なわれ比較的語順が自由であることなどの特徴がある。そのため会話文解析においては、文法をベースとした記述文の解析技法をそのまま適用することは一般に難しい。

日本語音声会話理解システムを実現する一方式として、会話文中のキーワードを抽出しそれらを中心に意味理解を行う方法が考えられている[1][2]が、そこでは曖昧な会話文候補の中から如何に正しくキーワードを抽出するかが基本的な問題となっている。

一方、これまでに任意な日本語の記述文を中心とした日本語音声認識において、音響処理の結果出力された音節認識候補の曖昧さを、言語処理を用いて解消するのに、音節の2重マルコフ連鎖モデルを用いた方法が提案され、その有効性が示されている[3][4]。

本報告では、冗長語を含んだ音声会話文に対して、正解候補がすべて含まれる置換誤り型の音節文ラティスの中から、冗長語を除去し、名詞を中心としたキーワード候補を抽出方法を提案し、その有効性を実験を行って定量的に評価する。

2 音声会話文に含まれる冗長語の抽出法とキーワード音節候補の抽出法

2.1 基本的な定義

本論文では会話文として、冗長語やいい直しを等を一切含まないキーボード入力タイプと冗長語を含む音声入力タイプの2種類の会話文(国際会議の問い合わせに関する会話文)を用いる。

【定義1】 キーボード入力タイプの音節表記の会話文すべてに対して、定めた2重マルコフ連鎖確率を文マルコフ連鎖確率(SMP)、また文の中のキーワードの部分に限定した時のマルコフ連鎖確率を、キーワードマルコフ連鎖確率(KMP)と呼ぶ。但し、ここではキーワードを、名詞の単語に限定する。

またキーワードとしては、普通名詞および固有名詞を含む一般の名詞を対象とする。会話文単位に発声された連続音声に対して、音響処理の結果得られる曖昧な音節認識候補を表したものを、音節会話文ラティスと呼ぶ。本論文では、次の条件を満たす音節ラティスを対象とする。

【定義2】連続音声認識の音響処理においては、セグメンテーションは正しく行われ

（すなわち音節区間は正しく認識され、脱落・挿入誤りは無く候補の置換誤りだけが存在し）、正解候補はその中に必ず存在するものとする。このとき音節会話文ラテイスの音節列の中でもとの会話文のキーワードを構成する音節列を、キーワード音節列と呼ぶ。

会話文音節ラテイス及びそのキーワード候補の例を図1に示す。

2.2 冗長語の抽出法

音声会話文には、図1のように通常冗長語が含まれる。冗長語を含む場合と含まない場合との音声会話文に対するの2重音節マルコフ連鎖確率値(SMP)の変化を、図2に示す。同図より、冗長語の存在する所でマルコフ連鎖確率値が下落して様子がわかる。このような性質を用いて、冗長語を抽出する方法を示す。

【冗長語抽出法1】 2重音節マルコフ連鎖確率が、(n+2)回連続して落ち込むとき、その位置に長さnの冗長語が存在するとして抽出する。

【冗長語抽出法2】 冗長語抽出法1に加えて、さらに冗長語の種類が予めわかっているとして、冗長語のパターンマッチングを行ない、一致した冗長語候補の先頭位置から末尾位置まの間で、少なくとも一回2重マルコフ連鎖確率の値が閾値を下回るとき、この候補を冗長語として抽出する。

2.3 キーワードの抽出法

次に、上記の方法で求められた冗長語を除去し、冗長語なしの音声会話文に対して、キーワードを求める方法を述べる。すなわちマルコフ連鎖確率及び単語辞書引きをして、音節会話文ラテイスより、キワード候補を抽出する方法を示す。

【キーワード候補の抽出法】次の（1）と（2）を、音節会話文ラテイスの全てのtについて繰り返す。

（1）音節会話文ラテイスの位置tを先頭とし、最大長がkまで順に音節候補を組み合わせて得られる音節列をキーに単語辞書にアクセスし、品詞が名詞となる音節列を求める。

（2）キーワードマルコフ連鎖確率(KMP)を用いて、（1）で得られたキーワード音節列のマルコフ連鎖確率値を求め、大きい順にソートし第一位の候補を最尤な候補とする。

また音節候補ラテイスのすべての組合せから得られる音節文候補に対して、キーワード抽出を行なう方法を網羅的抽出法とよび、また音節文マルコフ連鎖確率(SMP)を用ビタービアルゴリズムを用いて求められる最ゆうな音節文候補に対してのみ、キーワード抽出を行なう方法を、文最ゆう型抽出法と呼ぶ。キーワード候補の抽出手順を図3に示す。本実験では、kは7として行なう。

3 実験結果

3.1 実験条件

1. マルコフ連鎖確率の統計データに用いる日本語文の種類と総文数:
   (a) 会話文（国際会議の受け付け）:3364文
   (b) 総文字数:89,397文字
   (c) 総キーワード(名詞)数:11,515
2. 入力の会話文数:50文(標本内データ)、50文(標本外データ)
   (a) 文の長さ:50以下
   (b) キーワード数:141(標本内データ)、128(標本外)

3. マルコフ連鎖確率辞書：キーワード音節マルコフ連鎖確率と音節文マルコフ連鎖確率

3.2 実験結果

2章で述べた冗長語抽出法を用いた実験結果を図4に、またキーワード抽出の実験結果を図5に示す。[1] 冗長語の抽出法の比較

図4より、音節マルコフ連鎖確率情報（落ち込みを調べる）のみによる方法（方法1）では、適合率＝40.9%、再現率=16.4％と小さいが、冗長語の種類が既知であるとして、冗長語のパターンマッチングを行なうことを併用すること（方法2）により、適合率＝79.8%、再現率=83.6%（方法1に比べて、適合率で39.8%、再現率で67.2％向上する）とかなり高い値が得られることがわかった。

[2] キーワードマルコフ連鎖確率を用いた音節会話文ラティスからのキーワード抽出精度

図5より、網羅型の抽出法では、10位内に82.0%の正解候補（再現率と適合率の積が最大となる条件で、共に42.2%）が得られること、また文最ゆう型の抽出法では、10位内に86.0%の正解候補（再現率と適合率が最大となる条件では、再現率=64.8%、適合率=32.4%）が得られることがわかった。

また本キーワード抽出法は、適合率は低いものの正しいキーワードの8割以上を抽出可能であることから、有効な方法であると考えられる。

4 おわりに

本論文では、冗長語を含む音声会話文に対して、文音節マルコフモデルを用いた冗長語の抽出法およびキーワードマルコフ連鎖確率と単語辞書引きを組み合わせたキーワード抽出方法を提案し、その有効性を実験的に評価した。その結果、以下のような知見を得た。

(1) 網羅型の抽出法では、10位内に82.0%の正解候補が得られること、また文最ゆう型の抽出法では、10位内に86.0%の正解候補が得られることがわかった。

(2) 音節文マルコフモデルに加えて、冗長語の種類が既知であるとして、冗長語のパターンマッチングを行なうことを併用するキーワード抽出方法により、適合率＝79.8%、再現率=83.6%が得られることがわかった。

（参考文献）

(1) 荒木、河原、西田、堂下:キーワード抽出に基づく意味解析による音声対話システム、信学技法 NLC91-51 pp25-32 (1992)

(2) 坪井、橋本、竹林:キーワードスポッテイングに基づく連続音声理解、SP91－95 pp33-40 (1992)

(3) 荒木，村上，池原:2重マルコフモデルによる日本語文節音節認識候補の曖昧さの解消効果，情処論，30,4,pp467-477 (1989)

(4) 村上，荒木，池原:日本語文音節入力に対して2重マルコフ連鎖モデルを用いた漢字かな交じり候補の抽出精度，信学論，D-II,J75-D-II,1,pp11-20 (1992)

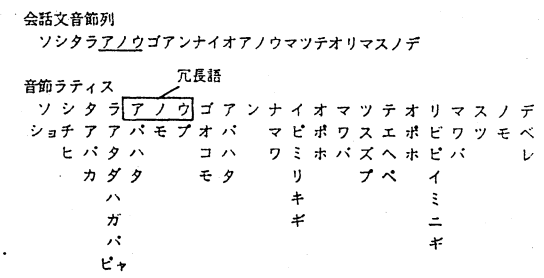

図1（A） 不要語を含む文の会話文音節ラティスの例

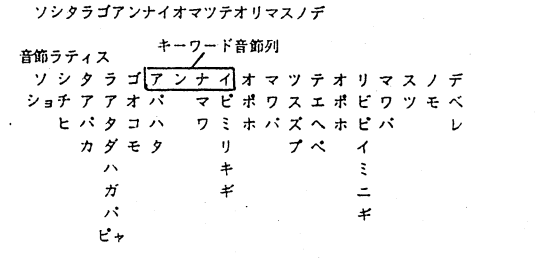

図1（B） 不要語を含まない文の会話文音節ラティスの例

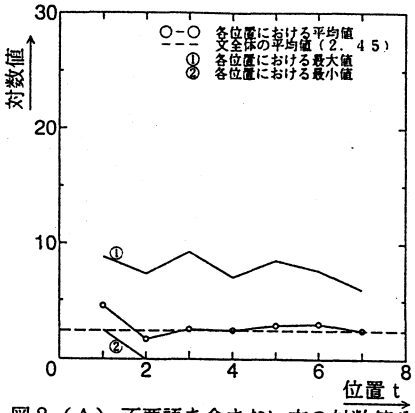

図２（Ａ）不要語を含まない文の対数値分布

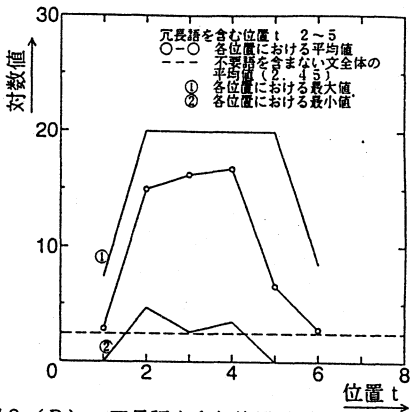

図２（Ｂ）　冗長語を含む位置付近の対数値分布

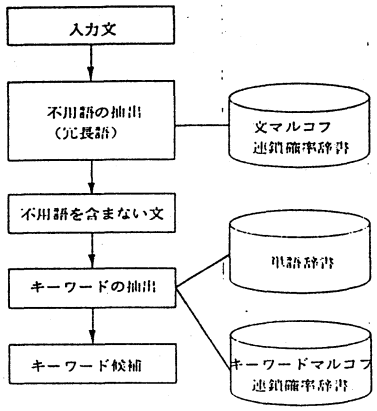

図３　　会話文の解析手順

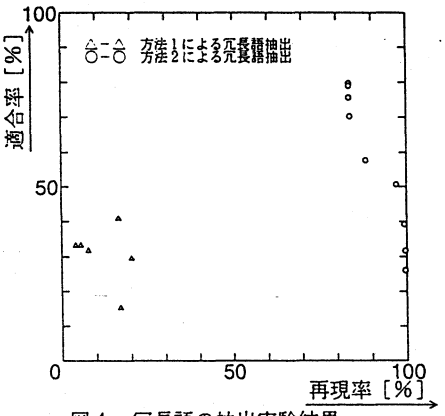

図４　冗長語の抽出実験結果

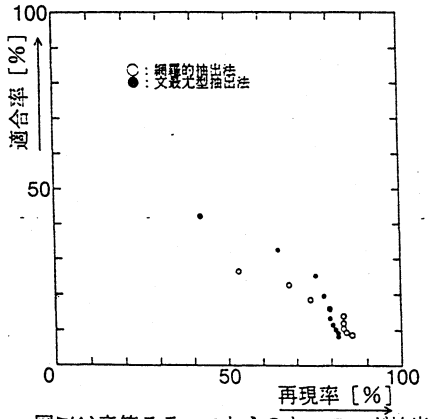

図5(A)音節ラティスからのキーワード抽出
再現率、適合率（標本外）

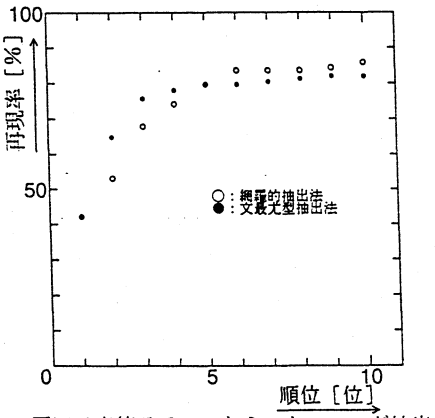

図5(B)音節ラティスからのキーワード抽出
（標本外）