音節マルコフモデルによる日本語音節会話文ラテイスからのキーワード抽出法

7R－1

荒木　哲郎+　池原　悟++　四方　啓智+
+ 福井大学工学部　++ＮＴＴ情報通信網研究所

1　はじめに

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]。またこれまでに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[2][3]。

本報告では、音節ラテイス形式で与えられた音声会話文候補の中から、名詞を中心としたキーワードを抽出するにあたって、[2]のマルコフモデルによる方法と単語辞書引きを組み合わせた方法を新たに提案し、その有効性を実験を行って定量的に評価する。

2　会話文のマルコフ連鎖確率とそれを用いたキーワード音節候補の抽出法

【定義１】　音節表記の会話文及び記述文において、音節文全体についてのマルコフ連鎖確率を文マルコフ連鎖確率、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。∎

本報告では、会話文として２種類のものを用意しそれぞれについてキーワード２重マルコフ連鎖確率を求めたところ、そのエントロピーは、各々1.58及び1.91であった。これは新聞記事についてのキーワードマルコフ連鎖確率のエントロピー2.24に比べると、会話文はキーワードが限定されるため小さな値となっている。それぞれの飽和特性を図１に示す。

A Method to Detect Keywords of Japanese syllable Sentence Lattice Using Markov Model

Tetsuo ARAKI+　Satoru IKEHARA++　Akinori SIKATA+
+　Faculty of Engineering, Fukui University
++ NTT Network Information Systems Laboratories

【定義２】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、音節会話文ラテイスと呼ぶ。但し、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ものを、キーワード音節列と呼ぶ。∎

音節会話文ラテイス及びそのキーワード候補の例を図２に示す。次に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。∎

4　実験結果

4．１　実験条件

（１）日本語文の種類と総文数:

1. 会話文１（海外旅行の受付）:1,129
2. 会話文２（ＮＨＫのラジオ講座）:553
3. 新聞記事77日分:28,500

（２）総名詞数:

1. 会話文１（海外旅行の受付）:1,382
2. 会話文２（ＮＨＫのラジオ講座）:2,023
3. 新聞記事77日分:191,971

（3）入力の会話文数:110 文
（4）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

4．2　実験結果

[1] 音節文マルコフ連鎖確率を用いた音節会話文ラティスの候補絞り込み効果

会話文1、会話文2及び記述文の音節文マルコフ連鎖確率を用いて、音節会話文マトリックスから得られる音節候補文の絞り込み効果を図3に示す。

同図より、標本内会話文データについては、第3位までで１００%近くの正解文が求まるが、標本外データについては２０%程度の正解率しか得られていないことがわかる。

[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効果

会話文1、会話文2及び記述文のキーワードマルコフ連鎖確率を用いて、音節会話文マトリックスから求めたキーワード音節候補の抽出結果を図4に示す。

同図より、標本内会話文の場合のキーワード抽出は、再現率９０%、適合率３０%程度であった。さらに新聞記事データによるキーワードマルコフ連鎖確率を用いた場合、適合率は１０%程度と悪いが、再現率が、７０%程度得られることがわかった。

5　おわりに

本報告では、音節ラティス形式で与えられた音声会話文候補の中から、名詞を中心としたキーワードを抽出するにあたって、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方法を新たに提案し、その有効性を実験的に評価した。

その結果、音節会話文ラティスからのキーワード抽出において、記述文のキーワードマルコフ連鎖確率と単語辞書引きによる方法によって、再現率で７０%のキーワード候補が得られることがわかった。

今後は、適合率の向上させる方法を研究していくと共に、さらにキワードを固有名詞並びに動詞に拡張し研究していく予定である。

（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 (1992)
(2) 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)
(3) 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)

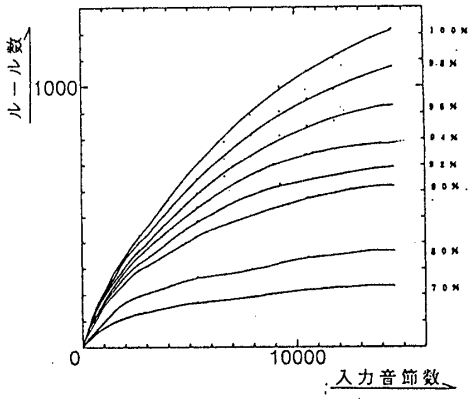

図１　２重マルコフ辞書（名詞）の飽和特性（会話文標本内）

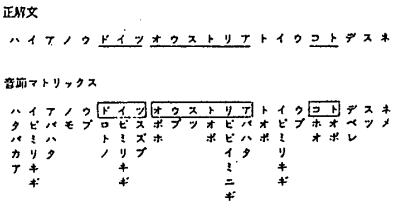

図２　音節会話文ラティス及びそのキーワード候補の例

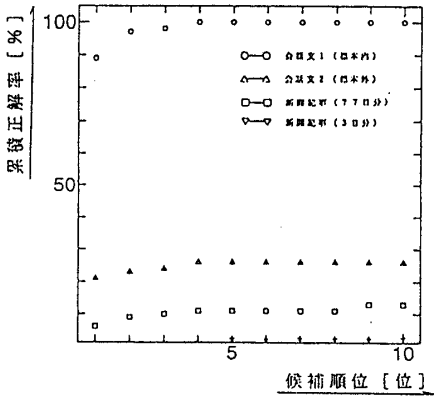

図３　音節文マルコフ連鎖確率を用いた音節会話文ラティスの候補絞り込み効果

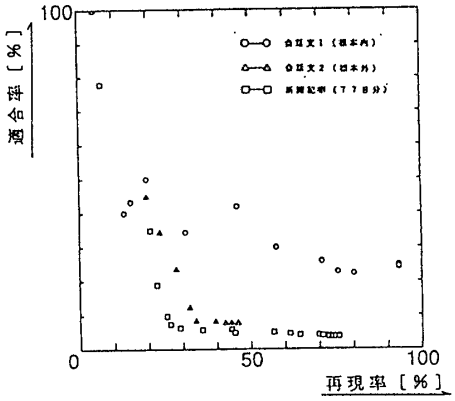

図４　キーワードマルコフ連鎖確率を用いた音節会話文ラティスからのキーワード候補