

マルコフモデルによる音節文ラテイスからのキーワード候補 の絞り込み

荒木 哲郎⁺ 池原 悟⁺⁺ 四方 啓智⁺

+福井大学

++NTTコミュニケーション科学研究所

Abstract

日本語音声会話文の意味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中心にして意味理解を行なう方法が考えられているが、ここでは曖昧な会話文の中からキーワードを正しく抽出し、キーワード間の関係を調べるの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文節(記述文)の音声認識において、音響処理の結果出力される曖昧な音節認識候補(音節ラテイス)の中から、音節の2重マルコフ連鎖モデルを用いて最も確からしい候補列を求める方法が提案され、その有効性が示されている。

本研究では、置換誤りを対象とし正解候補がすべて与えられている日本語会話文の音節ラテイスの中から、名詞(普通名詞および固有名詞を含む)を中心としたキーワード候補を、音節の2重マルコフ連鎖モデルと単語辞書引きによって求める方法を提案し、その有効性を実験によって定量的に評価した。その結果、以下の知見を得た。

1. 置換型のどんな入力文の音節ラテイスに対しても、2重マルコフ連鎖モデルのViterbiアルゴリズムを用いて得られる最優秀な音節候補の文字正解率は、平均で77.4%(標本外データ)-99.3%(標内データ)であり、文の長さの増加に伴い若干減少する。またマルコフ連鎖モデルを用いて得られる第一位文候補の正解率は、上記の文字正解率を基準にして文の長さ対応に求めた平均文正解率と比べると、文の長さに関係なく6-10%程度正解率は高いことがわかった。
2. 本手法によるキーワード候補の抽出精度は、再現率と適合率の積が最大となる条件下で、再現率および適合率とも、35.4%(標本外)-81.0%(標本内)であり、また対象とする候補順位を増加させることにより、再現率(カバー率)はさらに10%程度向上し、54.21-97.9%となることがわかった。

A Method to Find Syllable String Candidates of Keywords from Syllable Sentence Lattice Using Markov Model

Tetsuo ARAKI⁺ Satoru IKEHARA⁺⁺ Akinori SIKATA⁺

+Fukui University

++NTT Network Information Systems Laboratories

For the research of speech understanding systems, it is a basic problem to find correctly keywords of ambiguous spoken dialogue sentences and to study the relation among them.

On the other hand, 2nd-order Markov model of syllable characters is known to be useful to solve this ambiguity of syllable candidates of the string which is basic syntactic unit called a phrase (usually called a "bunsetsu").

This paper proposes a method to find syllable string candidates of keywords from syllable sentence lattice using Markov model. The effectiveness of this method is shown using experimental results.

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、文のつながりとしての冗長語が挿入したり、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、置換誤りを対象とし、正解候補がすべて含まれる音節文ラテイスの中から、名詞および固有名詞を中心としたキーワード候補を抽出するにあたって、[3] のマルコフモデルによる方法を拡張し、また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。 ■

本論文では、キーワードマルコフ連鎖確率の種類としては、2種類の会話文と新聞記事文(記述文)の3種を用意している。それぞれのキーワード2重マルコフ連鎖確率のエントロピーは、会話文が各々1.58及び1.91であり、新聞記事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。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、音節会話文ラテイスと呼ぶ。本論文では、次の条件を満たす音節ラテイスを対象とする。

【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。

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、キー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。■

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。

4 実験結果

4.1 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:

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

(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 of 会話文数 (海外旅行の受付):100 文
(このうち42文をキーワード抽出実験に用いる)

1. 文の長さが10文字以下の文:21 文
2. 文の長さが10-20文字以下の文:26 文
3. 文の長さが20-30文字以下の文:29 文
4. 文の長さが30文字以上の文:24 文

(4) マルコフ連鎖確率辞書: キワード音節マ
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
文候補の正解率

入力文の音節ラテイスの各音節候補を、重み
付けせずすべて均等に見做して得られる全ての
音節の文候補に対して、文マルコフ連鎖確率(標
本内と標本外の会話文および記述文に対する)
の Viterbi アルゴリズムを用いて得られる最ゆ
うな音節候補の文字正解率を求めると、次のよ
うに文の長さの増加に伴って減少し、標本内会
話文のマルコフモデルの場合、平均で99.3%で
あった。

(1)10文字以下の時:

100%(標本内)、99.7%(標本外)

(2)10文字から20文字までの時:

99.5%(標本内)、77.6%(標本外)

(3)20文字から30文字までの時:

99.3%(標本内)、76.7%(標本外)

(4)30文字以上の時:

99.0%(標本内)、75.3%(標本外)

また音節文ラテイスから得られるすべての音
節文候補に対して、文マルコフ連鎖モデルを用
いて得られる最ゆうな文候補の正解率を、音節
文字正解率から文の長さに応じて求めた平均文
正解率と比較した結果を図4に示す。同図より、
文字正解率より求めた平均の文正解率に比べ、
文マルコフ連鎖モデルより得られた文正解率の
方が、6-10%程度高いことがわかった。

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

音節文ラテイスから、標本内と標本外の会話
文および記述文(新聞記事文)のキーワードマ
ルコフ連鎖モデルを用いて、図2の手順に従っ
てキーワード候補を求めた時、文の長さ毎に適
合率 P と再現率(カバー率) R の積が最大とな
る条件下で求めた P と R の値を図3に示す。同
図より、標本内と標本外の会話文のキーワード
マルコフ連鎖モデルを用いた場合のキーワード
抽出は、最大で再現率が35.4(標本外)-81(標本
内)%、適合率が35.4(標本外)-81(標本内)%で
あった。

さらにそれぞれのキーワードマルコフ連鎖モデルを用いて求められるキーワード候補の順位を増加させた時の再現率(カバー率)の値を、図4(標本内キーワード)、図5(標本外キーワード)、および図6(記述文キーワード)に示す。同図より、標本内キーワードマルコフ連鎖モデルの場合には、97.9%のカバー率が得られ、また標本外のキーワードマルコフ連鎖モデルの場合には、54.2%のカバー率が得られることがわかった。

また標本内データに対しては高い正解率を示しているが、標本外の場合が低いのは会話文の学習が十分でないことを示している。そこで名詞に関するキーワードの学習を新聞記事データを用いて行なうことを考え、記述文のキーワードマルコフ連鎖確率を用いた場合、再現率が、81.1%まで向上することがわかった。

5 おわりに

本論文では、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、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、以下のような知見を得た。

1. 置換型のどんな入力文の音節ラテイスに対しても、2重マルコフ連鎖モデルのViterbiアルゴリズムを用いて得られる最優秀な音節候補の文字正解率は、平均で77.4%(標本外データ)-99.3%(標本内データ)であり、文の長さの増加に伴い若干減少する。またマルコフ連鎖モデルを用いて得られる第一位文候補の正解率は、上記の文字正解率を基準にして文の長さ対応に求めた平均文正解率と比べると、文の長さに関係なく6-10%程度正解率が高いことがわかった。
2. 本手法によるキーワード候補の抽出精度は、再現率と適合率の積が最大となる条件

下で、再現率および適合率とも、35.4%(標本外)-81.0%(標本内)であり、また対象とする候補順位を増加させることにより、再現率(カバー率)はさらに10%程度向上し、54.21-97.9%となることがわかった。今後は、キーワードの学習方法を研究していくと共に、さらにキーワードを動詞などに拡張してキーワード間の関係を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 pp25-32 (1992)
- (2) 坪井、橋本、竹林:キーワードスポッティ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)

正解文

ハイアノウドイツイオウストリアトイウコトデスネ

音節ラテイス

ハイ	ア	ノウ	ドイ	ツイ	オウ	スト	リア	ト	イ	ウ	コト	デ	ス	ネ		
タ	ビ	バ	モ	ブ	ロ	ビス	ボ	ラ	ツ	オ	ビ	バ	オ	ベ	ツ	メ
カ	リ	タ			ノ	リ	ブ			イ	タ			リ		
ア	キ				キ					ミ				キ		
	ギ				ギ					ニ				ギ		

図1 音節会話文ラテイス及びそのキーワード候補の例

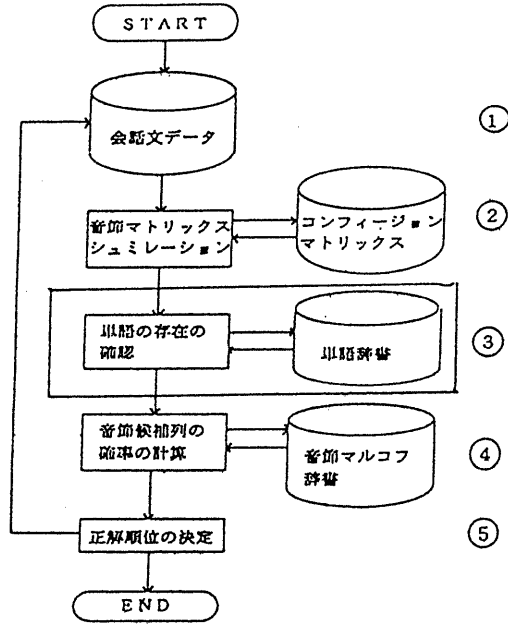


図2 キーワード候補の抽出手順

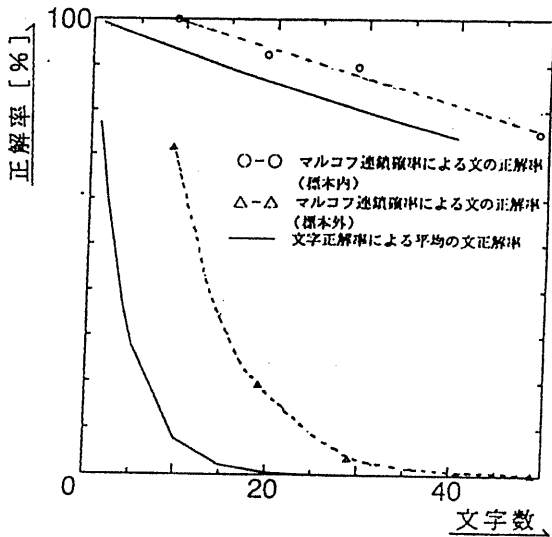


図3 音節マルコフモデルを用いた文候補の正解率

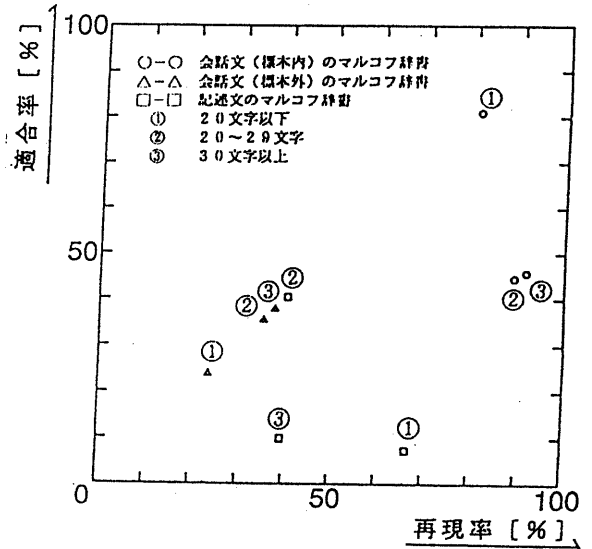


図4 音節会話文ラティスからのキーワード抽出の適合率、再現率

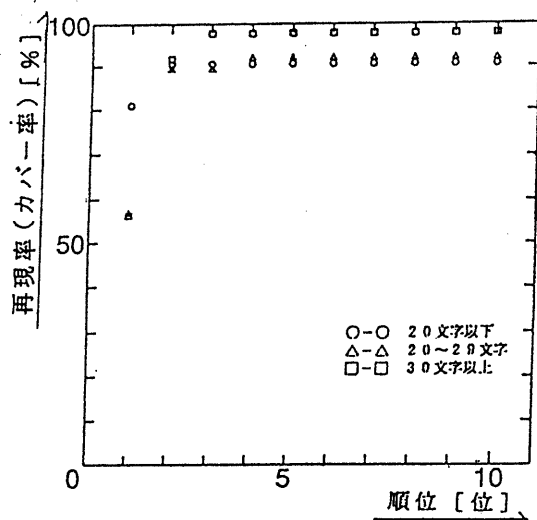


図5 音節ラティスからの
キーワード抽出
(会話文標本内
キーワードマルコフ)

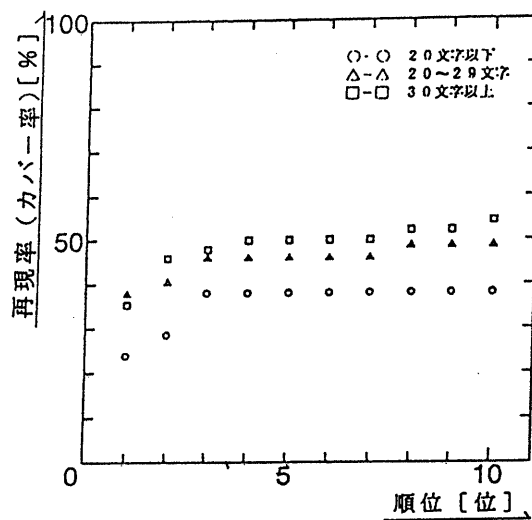


図6 音節ラティスからの
キーワード抽出
(会話文標本外
キーワードマルコフ)

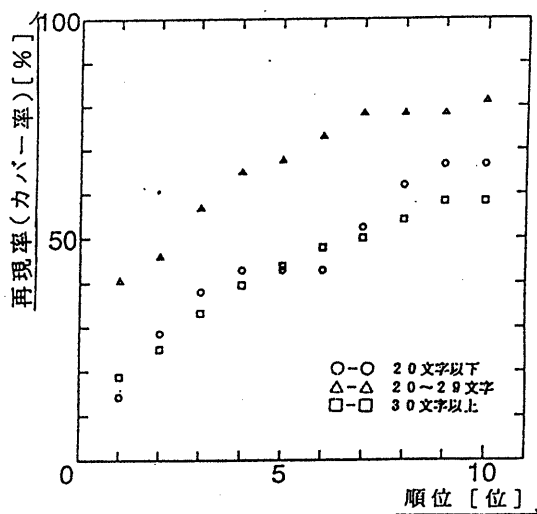


図7 音節ラティスからの
キーワード抽出
(記述文標本外
キーワードマルコフ)

文1 エウトイチオウワタシノホウノヨテイトイタシマシテハ

順位	キーワード候補	マルコフ連鎖確率のコストC*
1	ホウ	0.997400
2	ホウ	0.997400
3	ワタシ	1.486508
4	ヨテイ	1.813994
5	イチ	1.906296
6	オウ	2.031278
7	オウ	2.031278
8	ヒマ	2.211895
9	マシ	2.608135

文2 エウトドイツガヨツカカンデオウストリアニミツカカングライタイザインタイトオモウ

順位	キーワード候補	マルコフ連鎖確率のコストC
1	ホウ	0.997400
2	ドイツ	1.505130
3	オウストリア	1.559042
4	イツ	1.662796
5	タイザイ	1.678329
6	オウス	1.764122
7	ミツカ	1.950894
8	オウ	2.031278
9	ヨツカ	2.150894
10	ナン	2.211895
11	ナン	2.211895
12	グタイ	2.330987
13	キタイ	2.330987
14	ライ	2.374914
15	ガイ	2.388209
16	ナカ	2.608135
17	ナカ	2.608135
18	ダイ	2.688617

* $C = -\sum \log P$

図8 抽出されたキーワード候補の例