情 報 学 基 礎 48－2
（1997. 11. 13）

N-gramを用いた対話文の言い直し表現の検出法

荒木 哲郎 （福井大学）　池原 悟 （鳥取大学）　三品 尚登 （福井大学）

自然な発話では言い直しという言語現象が生じ、対話文の解析を困難にしている。これまでにべた書きされた音節表記の対話文に対して、文節境界に挿入された、繰り返しタイプの言い直しの音節列(換言前音節列と呼ぶ)を検出する方法が提案され、その有効性が示されている。

本論文では、上記の検出法を換言前音節列が文節境界以外に現れる場合にも、適用可能なように改良する方法を提案する。すなわち、この方法は、繰り返し表現の音節列の候補を出現位置によらずに網羅的に抽出する処理と、抽出された候補が換言前音節列の候補であるか否かを判定する処理からなる。具体的には、第一段階では、n-gramモデルを用いて文中に現れる同一表現の音節列の組(n-gram)をすべて洗い出し、その後で連続して現れる同一のn-gram対を基に、換言前音節列となりうる音節列候補(拡大n-gramと呼ぶ)を求める。次の第二段階では、音節文マルコフ連鎖モデルを用いて、これらの拡大n-gramのうちでどれが換言前音節列かを決定する。その際、1文中に現れる1文字以上のn-gram音節列の対をすべて抽出する方法(第1の抽出法)と一つのn-gramの中に包含される部分列をすべて除去し最大のnn-gramの対を抽出する方法(第2の抽出法)二通りの方法を考える。

実際にATRの《旅行に関する対話文》データを用いて実験を行ないこれらの方法の有効性を評価した結果，次の知見を得た。

(i) 第1の網羅的な抽出法は、再現率は71.0～98%と高いが換言音節列以外の音節列が多く検出されるため適合率が16.0～36.6%と低い。一方、第2の抽出法は、再現率=81.0～91.0%および適合率=59.9～82.7%が両方ともに高い精度が得られること、

(ii) 第2の方法は、仮文節境界を用いて検出する方法と比べると、適合率は同程度で再現率が約5%程度高いことがわかり本手法の有効性が確認された。

A Method to Detect the Syllable Strings of Self-repair in Spontaneous Speech Using N-gram Model.

TETUO ARAKI , SATORU IKEHARA , NAOTO MISHINA

This paper proposes a method to detect the strings of self-repairs occuring in spontaneous speech,which are assumed to be represented with strings of syllables obtained correctly by accoustic processing, using n-gram model and Markov models of syllables.

This method is decomposed of the following two steps: The first step is to detect all of the pairs (caled as n-gram pair) with the same substrings appeared in a dialogu sentence using n-gram extraction methods. The second step is to determine whether n-gram pair is self-repair or not, using Markov models of syllables.

This method is experimentally evaluated to be effective using ATR database of dialogue.

1 はじめに

これまでに、べた書き音節文において、文節境界を手がかりとして繰り返しタイプの言い直し音節列を検出する方法が提案され、その有効性が示されている。しかし、この手法は文節境界以外に出現する言い直し表現や、また文節境界が誤って検出されたり未検出の場合には、検出ができないなどの問題がある。

本論文では、このような問題を解決するために、対話文の任意な位置に出現する繰り返しタイプの言い直し表現を検出可能なように改善する方法を提案する。従来、自然言語処理において用いられているn-gram統計を、任意の長さnに対して、大量の言語データから高速に求める手法が、長尾・森によって提案されており[1]、またこれを用いて自然言語表現に連続して現れる連鎖型の共起表現と離れた位置に現れる型の共起表現離散の文字列をすべて抽出する方法が、池原・白井によって示されている[2]。

本方法は、図2のように次の2つの処理から構成される。
(1) このようなn-gram抽出法を用いて、一文中に現れる連続して共起する連鎖型の繰り返し文字列および、離れた位置に共起する離散型の繰り返し文字列(n-gram)の組をすべて洗い出す処理と、
(2) これらの一対のn-gramの組から構成される範囲(拡大n-gramと呼ぶ)を決定し、これが言い直し表現の候補か否かを判定する2段階の処理からなる。

2 言い直し表現の定義と検出の考え方

2.1 言い直し表現の定義

言い直しとは、前に言ったことの誤りを訂正してもう一度言ったり、話の途中で言い淀んでしまってもう一度言うといった言語現象のことであり、具体的には次のようなものである
「はい，（かしこ）かしこまりました。」
「（わたし）わたくし鈴木が承りました。」

ここで、言い直しによって訂正される対象となる部分（括弧で示した部分）を《換言前音節列》と呼び、また言い直しによって訂正された部分（下線で示した部分）を《換言後音節列》と呼ぶ。

本論文では図1のように《換言後音節列》は《換言前音節列》に隣接するものとし、それぞれに同一な部分音節列が含まれるものとする。このような同一な部分音節列の組は、一つの対話文において連続して共起する場合(連鎖型の共起表現)と、離れた位置に共起する場合(離散型の共起表現)がある。その前に起きる同一部分文字列を、前方n-gramまた後に起きる同一部分文字列を後方n-gramと呼ぶ。また、このような2つのn-gramの組を用いて構成される文字列、すなわち前方n-gramの先頭から、後方n-gramの先頭までの範囲の文字列を、拡大n-gramと呼ぶ。拡大n-gramは、言い直し表現であるとき、換言音節列に相当する。

2.2 言い直し表現の検出に関する基本的な考え方

言い直し表現の検出方法は、図2のように2段階に分けて行なわれる。まず第一段階の処理では、n-gram方法を用いて対話文中に現れる同一音節列の組(n-gram)をすべて抽出する。その際、連鎖型または離散型の共起表現である一対のn-gramの組を抽出するに当たって、1つのn-gramに含まれる部分文字列の組を抽出するか否かによって、次の2通りの方法が考えられる。すなわち、
(i) 連鎖型および離散型の部分文字列の組をすべて抽出し拡大n-gramを求める方法、
(ii) 最大のn-gramの組を抽出し、それに含まれる部分文字列の組はすべて削除して拡大n-gramを求める方法
である。

次に第2段階では、前段階で求めた拡大n-gramのうち、どれが換言前候補になるかを、さらに2段階に分けて判定する。すなわち、前段階で求めた拡大n-gramの近傍で前方向および後方向のマルコフ連鎖確率値の落ち込みがあるかどうか、またその拡大n-gramを削除した位置で確率値の落ち込みがないかどうかによって、言い直しか否かを決定する。

3 言い直し音節列の検出方法

3.1 n-gramによる音節部分列の抽出法

一つの対話文におけるn-gramの求め方としては、次の2つの方法が考えられる。
(1) 連続的および離散的な同一の音節列の組をすべて求める方法、
(2) 包含されるすべての部分音節列のn-gramをすべて削除し、最大のn-gramだけを抽出する方法。

(2) の方法による n-gram の抽出の例を図 3 に示す。同図において、文＝ソンナニワルトハオモワ・・・] に対するソート前の原文番地ファイル、ソート後の拡張源番地ファイル、拡張原文番地ファイル等を用いて、1 文中に現れる n-gram を求め、拡大 n-gram を決定している。例えば連続的な前方および後方の 8-gram がそれぞれ《わるいとはおもわ》の場合には、両者より拡大 8-gram=《わるいとはおも》として求まることがわかる。

3.2 マルコフ連鎖モデルによる言い直し音節列候補の絞り込み法

マルコフ連鎖確率を用いて換言前音節列を検出する方法は、図 4 に示されるように、次の 2 通りの方法によって行なわれる。
(1) 拡大 n-gram が換言前音節列のときは前方向のマルコフ連鎖モデル (順方向) と後方向のマルコフ連鎖モデル (逆方向) を用いて、その連鎖確率値を調べると、その音節列の近傍 (マルコフ連鎖確率の値を求める際に、音節列の構成要素が影響を及ぼす範囲で、内部および先頭と末尾位置から前後に 3 文字まで離れた範囲) で、連鎖確率値が低下すると考えられる。また、
(2) 拡大 n-gram が換言前音節列のときはを原文から削除したあとの文に対して、順方向または逆方向のマルコフ連鎖モデルの連鎖確率値調べると、その削除位置の近傍で平均値のレベルに復帰すると考えられる。

以上より、拡大 n-gram が換言前音節列であるかが、上記の 2 つの条件を満たしていることを調べることにより判定できるが、拡大 n-gram に対するマルコフ連鎖確率の落ち込みを評価する方法としては、次の 2 つの方法を考える。
(a) 拡大 n-gram に対するマルコフ連鎖確率の平均値が、あるしきい値よりも小さいとき、
換言前音節列と判定する方法 (平均値評価法)。(b) 拡大 n-gram に対するマルコフ連鎖確率値の最小値が、あるしきい値を下回るとき、換言前音節列と判定する方法 (最小値評価法)。

3.3 言い直し音節列検出の精度の推定

第 1 段階および第 2 段階によって得られる言い直し音節列検出の精度の評価は，次に定義される適合率 P と再現率 R を用いて行なう。これらの積が大きいほど検出能力は高いと考えられるので、両者の積で検出方法の優劣を評価する.

言い直し音節列検出の適合率 P と再現率 R を次のように定義する.

$$P \equiv \frac{\text{正しく検出された言い直し音節列の数}}{\text{検出された音節列候補の総数}}$$

$$R \equiv \frac{\text{正しく検出された言い直し音節列の数}}{\text{もとの言い直し音節列の総数}}$$

4 実験結果と考察

4.1 実験条件

3.2.2 節で述べた各推定方法による文節境界の推定実験および、言い直しの検出実験を行うに当たって、以下に示すような実験入力文とマルコフ連鎖確率辞書を用いた.

1. 実験入力文
   - (a) 文の内容：旅行に関する会話
   - (b) 文の表記：文音節列
   - (c) 総文数：100 文（標本外，文節境界位置に出現する単純な繰り返しタイプの言い直しが少なくとも一つ存在する）
2. マルコフ連鎖確率辞書の統計データ
   - (a) データの内容：旅行に関する会話
   - (b) データの表記：文節音節列（空白記号付き，言い直しは含まない）
     - i. 総文数：6790 文
     - ii. 総音節数：236705 音節（空白記号を除くと 155345 音節）
3. マルコフ連鎖確率辞書のタイプ
   - (a) 種類：文節マルコフ連鎖確率
     - i. 次数：4 次（3 重）
     - ii. 方向のタイプ：順方向と逆方向

4.2 実験結果

4.2.1 n-gram による言い直し音節列候補の抽出結果

2 つの n-gram 抽出法による言い直し音節列候補の抽出実験結果を、表 2(適合率と再現率の積が最大の

場合) に示す。また適合率と再現率の変化を図 5 に示す。
また、抽出結果の例を表 3 に示す。

4.2.2 言い直し音節列の検出実験結果

(1) n-gram 抽出法の違いによる検出結果

第一の網羅的な抽出方法では、予想通り再現率が最大 (71.0-98.0%、) となるが、言い直し以外の音節列も一緒に検出されるために適合率はかなり低い値 (16.0-36.6%) となっている。一方、再現率 P と適合率 R の積 $P \times R$ の値では、第二の方法 (包含される部分列は除き、拡大 n-gram でその半分以上の音節列が前後で一致する) で最大となり、$P = 82.7$ %, $R = 81.0$%(積 $P \times R = 66.9$) という良い精度が得られることがわかった。

また、第一段階と第二段階の適合率の差は、第一の抽出方法で約 10-20%、また第二の抽出方法で約 20-30% であることから、音節文マルコフ連鎖確率によってかなり換言前音節列以外の音節列候補をかなり絞り込むことができることがわかった。

(2) マルコフ連鎖確率値の評価法の違いによる検出結果

表 3 および図 5 より、最小値評価方法は、平均値評価方法よりも再現率の面では優れているが、積 $P \times R$ においては平均値評価方法方が優れていることがわかった。これは、最小値評価方法では、再現率は優れているが、適合率が大きく低下するために総合的な評価としては劣ることになる。

(3) 仮文節境界を用いた検出方法との比較

仮文節境界を用いた検出方法による結果は、適合率 $P = 81.6$%, 再現率 $R = 75.5$ % (積 $P \times R = 61.6$) と比べると、適合率では同程度であるが、再現率では 5% 程度向上することがわかった。これは n-gram 方法により、換言前音節列候補が多く検出されることによるものである。

5 あとがき

本論文では、音響処理によって正しく出力されたべた書きの音節表記文に対して、任意の位置に現れる繰り返しタイプの言い直し音節列を検出する方法を提案した。

この方法は、最初に 網羅的にすべての n-gram を抽出する第一の方法、および内包される部分音節列を除いて最大の n-gram だけを抽出する第二の方法によって n-gram を求め、換言前と後の音節列に該当する拡大 n-gram を決定する。

次に、第 2 段階として音節マルコフ連鎖モデルを用いて、今求めた n-gram 候補が言い直しの音節列か否かを判定する方法を示した。

ATR の《旅行に関する対話文》データを用いて実験を行ないこれらの方法の有効性を評価した結果, 次の知見を得た。

(1) 第一の方法は再現率においてほぼ網羅的に抽出するが、換言前候補以外の候補も多く抽出するために、適合率はかなり低い。一方、第二の方法により再現率 81.0%(適合率 82.7%) の精度が得られることがわかった。

(2) 仮文節境界を用いた方法による言い直し音節列の検出結果より、再現率において約 5% 程度向上 (適合率は同程度) することが分かった。

なお, 今後の課題としては、n-gram 方法と仮文節境界を用いた方法の組合せによる言い直し表現の検出方法の研究などがが挙げられる.

参考文献

[1] Nagao, M. and Mori, S:*A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese*, The Proceedings of the 15th International Conference on Compitational Linguistics, pp611-615 (1994)

[2] 池原 悟, 白井 諭, 河岡 司: *N-gram* を用いた連鎖型狂気表現の自動抽出法, 言語処理学会 第 1 回年次大会, pp313-316 (1995)

[3] 荒木哲郎, 池原 悟, 橋本昌東 : 音節連鎖特性を用い対話文の言い直し表現の検出法, 情報処理学会情報学基礎研究会, (1997)

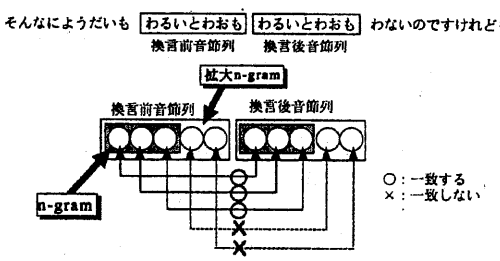

図1　n-gramによる拡大n-gramの定義

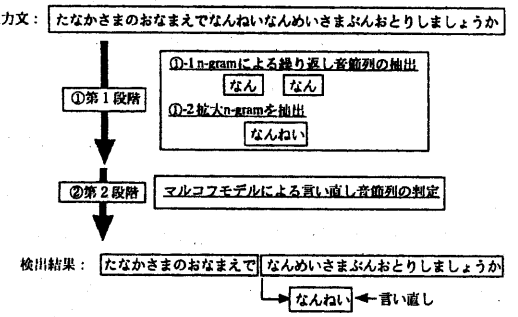

図2　言い直し音節列の検出手順

原文データ：そんなにようだいもわるいとわおもわるいとわおもわないのですけれども

原文番地ファイル

原文番地	文字列単語［ソートなし］（先頭部分）
1	そんなにようだいも
2	んなにようだいもわ
3	なにようだいもわる
4	にようだいもわるい
5	ようだいもわるいと
6	うだいもわるいとわ
7	だいもわるいとわお
8	いもわるいとわおも
9	もわるいとわおもわ
10	わるいとわおもわる
11	るいとわおもわるい
12	いとわおもわるいと
13	とわおもわるいとわ
14	わおもわるいとわお
15	おもわるいとわおも
16	もわるいとわおもわ
17	わるいとわおもわな
18	るいとわおもわない
19	いとわおもわないの
20	とわおもわないので
21	わおもわないのです
22	おもわないのですけ
23	もわないのですけれ
24	わないのですけれど
25	ないのですけれども
26	いのですけれども
27	のですけれども
28	ですけれども
29	すけれども
30	けれども
31	れども
32	ども
33	も

拡張原文番地ファイル

採否	抽出文字数	一致文字数	レコードNo	原文番地	文字列単語［ソートあり］（先頭部分）
○	6	6	1	19	いとわおもわないの
○	6	1	2	12	いとわおもわるいと
○	1	1	3	26	いのですけれども
○	1	0	4	8	いもわるいとわおも
×	0	0	5	6	うだいもわるいとわ
○	3	3	6	22	おもわないのですけ
○	3	0	7	15	おもわるいとわおも
×	0	0	8	30	けれども
×	0	0	9	29	すけれども
×	0	0	10	1	そんなにようだいも
×	0	0	11	7	だいもわるいとわおも
×	0	0	12	28	ですけれども
○	5	5	13	20	とわおもわないので
○	5	0	14	13	とわおもわるいとわ
×	0	0	15	32	ども
○	1	1	16	25	ないのですけれども
○	1	0	17	3	なにようだいもわる
×	0	0	18	4	にようだいもわるい
×	0	0	19	27	のですけれども
○	1	1	20	33	も
○	2	2	21	23	もわないのですけれ
○	9	9	22	16	もわるいとわおもわ
○	9	0	23	9	もわるいとわおもわ
×	0	0	24	5	ようだいもわるいと
○	7	7	25	18	るいとわおもわない
○	7	0	26	11	るいとわおもわるい
×	0	0	27	31	れども
○	4	4	28	21	わおもわないのです
○	4	1	29	14	わおもわるいとわお
○	1	1	30	24	わないのですけれど
○	8	8	31	17	わるいとわおもわる
○	8	0	32	10	わるいとわおもわな
×	0	0	33	2	んなにようだいもわ

拡張原文番地ファイル

採否	抽出文字数	一致文字数	レコードNo	原文番地	文字列単語［ソートあり］（先頭部分）
×	0	0	10	1	そんなにようだいも
×	0	0	33	2	んなにようだいもわ
○	1	0	17	3	なにようだいもわる
×	0	0	18	4	にようだいもわるい
×	0	0	11	5	ようだいもわるいと
×	0	0	5	6	うだいもわるいとわ
×	0	0	12	7	だいもわるいとわおも
○	1	0	4	8	いもわるいとわおも
○	9	0	23	9	もわるいとわおもわ
○	8	0	32	10	わるいとわおもわな
○	7	0	26	11	るいとわおもわるい
○	6	1	2	12	いとわおもわるいと
○	5	0	14	13	とわおもわるいとわ
○	4	1	29	14	わおもわるいとわお
○	3	0	7	15	おもわるいとわおも
○	9	9	22	16	もわるいとわおもわ
○	8	8	31	17	わるいとわおもわる
○	7	7	25	18	るいとわおもわない
○	6	6	1	19	いとわおもわないの
○	5	5	13	20	とわおもわないので
○	4	4	28	21	わおもわないのです
○	3	3	6	22	おもわないのですけ
○	2	2	21	23	もわないのですけれ
○	1	1	30	24	わないのですけれど
○	1	1	16	25	ないのですけれども
○	1	1	3	26	いのですけれども
×	0	0	19	27	のですけれども
×	0	0	24	28	ですけれども
×	0	0	9	29	すけれども
×	0	0	8	30	けれども
×	0	0	27	31	れども
×	0	0	15	32	ども
○	1	1	20	33	も

普通のn-gram

n-gram	方法1（連続型）	回数
9-gram	もわるいとわおもわ	2
8-gram	わるいとわおもわ	2
7-gram	るいとわおもわ	2
6-gram	いとわおもわ	2
5-gram	とわおもわ	2
4-gram	わおもわ	2
3-gram	おもわ	2
2-gram		0
1-gram	な	2
1-gram	い	4

拡大n-gram

n-gram	方法1（連続型）	回数
9-gram	もわるいとわお	2
8-gram	わるいとわおも	2
7-gram	るいとわおもわ	2
6-gram	いとわおもわ	2
5-gram	とわおもわ	2
4-gram	わおもわ	2
3-gram	おもわ	2
2-gram		0
1-gram	な	2
1-gram	い	4

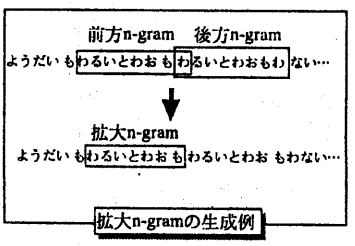

図3　n-gramによる同一文内の繰り返し音節列の抽出法の例

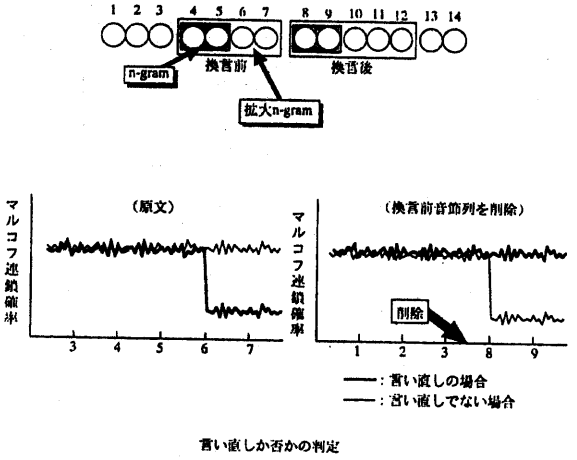

図4　マルコフモデルによる言い直しの判定

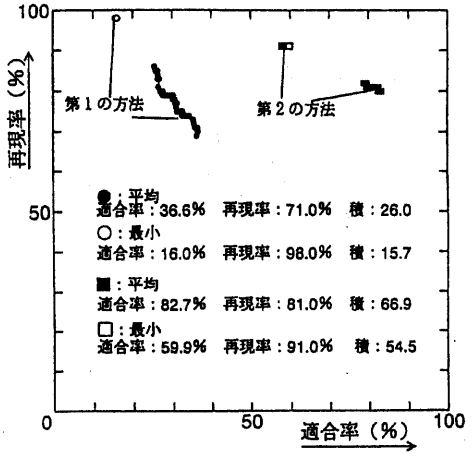

図5　各方式による言い直しの検出結果

表1　言い直し音節列の検出法

	マルコフモデルの種類	①換言前音節列を削除しない場合		②換言前音節列を削除した場合	
		言い直しでない場合	言い直しの場合	言い直しでない場合	言い直しの場合
1	順方向マルコフ連鎖確率	確率の落ち込み ×（なし） →検出しない	確率の落ち込み ○（あり） →検出する	確率の上昇 ×（なし） →検出しない	確率の上昇 ○（あり） →検出する
2	逆方向マルコフ連鎖確率	確率の落ち込み ×（なし） →検出しない	確率の落ち込み ○（あり） →検出する	確率の上昇 ×（なし） →検出しない	確率の上昇 ○（あり） →検出する

表2　各方式による言い直しの検出結果の比較

		第1段階			第2段階		
		P（%）	R（%）	P×R	P（%）	R（%）	P×R
第1の方法	平均	7.31	98.0	7.17	36.6(71/194)	71.0(71/100)	26.0
	最小	(98/1340)	(98/100)		16.0(98/611)	98.0(98/100)	15.7
第2の方法	平均	45.0	91.0	41.0	82.7(81/98)	81.0(81/100)	66.9
	最小	(91/202)	(91/100)		59.9(91/152)	91.0(91/100)	54.5

（注）P：適合率　R：再現率　P×R：積
（　）内は実際の検出数

表3　各方法による抽出結果の例

換言前音節列のタイプ	換言前音節列の例	第1の方法		第2の方法	
		換言前音節列	その他の音節列	換言前音節列	その他の音節列
①換言前音節列内の部分列の有無	ジャアソノ［ソイ］ツイ ンデオネガイシマス	◎ / ◎	× / ×	◎ / ◎	◎ / ◎
②換言前の音節列で内包でない音節列（排他的）	ホテルノマエニ［ノタク］サンノタク シイガ…	◎ / ◎	× / ◎	× / ×	× / ◎
③換言前音節列で内包で部分が共通	モウヒトツ［ツノ］ホウノカイシャデスケレドモ…	◎ / ◎	× / ×	× / ×	× / ×

（注）：■は換言前音節列と一致する拡大n-gram
▨は拡大n-gram■のもとになるn-gram
○はその他のn-gram

◎：正しい音節列の場合は検出し、誤った音節列の場合は検出しない
×：誤った音節列の場合は検出し、正解の音節列の場合は検出しない

第1段階 / 第2段階