自然言語処理 135－18
（2000．1．28）

次発話予測モデルによる音声認識候補の再順序付け

岩本 秀明

ATR 音声翻訳通信研究所

音声認識の精度改善のために、発話間の単語共起を統計的に学習した言語モデルを利用することを提案する。このモデルによる改善方式は、直前の発話を前提として、発話間にまたがる単語対の生起確率から発話の尤度を求める、というものである。551 発話の音声認識結果に対し、提案モデルを適用する実験を行った。その結果、単語認識率の error rate を 2% 削減した。このうち、内容語が改変したのは、31 発話で、そのうち、23 発話を改善した。この内容語に対する改善率は、74.2% である。これに対して、改変があった全ての発話は 51 発話、そのうち改善されたものは 32 発話、その改善率は 62.7% である。この実験により、提案モデルの内容語への改善効果が高いことが明らかとなった。

Rescoreing Speech Recognition Candidates Using Next Utterence Prediction Model

Hideaki Iwamoto

ATR Interpreting Telecommunications Research Laboratories

For the purpose of improving speech recogintion rate, this paper proposes stocastic liguistic models which are trained on the basis of word cooccurrence between utterences. These models provide the improvement method which give score to an utterence by using its word cooccurrence probablities under the condition that its previous utterence has been occoured. We applied these proposing models to rescoring experiments for speech recognition output. As the result, we obtained 2% error rate reduction. And we found it effective to content word error correction.

1 はじめに

本研究は、発話間の単語共起を統計的に学習した言語モデルによる音声認識精度の改善を目的とする。

現在、キャッシュ・モデル、distant bi-gram、統語制約などの n-gram を超える長距離の言語モデルによる音声認識手法の研究が行われている。

キャッシュ・モデル [1] は、一度出現した単語は再び出現する確率が高い、という経験則を利用する。しかし、次発話予測モデルは、隣接発話間に限れば、キャッシュモデルが与える情報を含む。

n-gram、distant bi-gram および統語制約を最大エントロピー法 [2] により統合する手法がある。本稿で提案するモデルは、直前の発話に制限した distant bi-gram であるとともに、発話間に構成素境界をおいた統語制約という側面を持ち、それらを兼ね備えた効率的なアプローチとなっている。

現在では、さらに、音声入力から直接話題や発話意図を取り出す、音声理解の研究が盛んである。これらの研究のなかでも、発話内だけでなく、文脈情報を使った手法 [3][4] の効果があがっている。

本研究は、発話理解の研究とは異なり、発話そのものの認識精度を向上させることが目的である。以前の実験 [5] では、発話間の内容語共起と発話意図タイプを用いたが、発話意図タイプの効果があがらなかったため、比較的効果があった発話間の内容語共起を全ての単語共起へと拡張して、本稿では、内容語と機能語とを区別することの有効性を検証する。

2 発話間の単語共起に基づく言語モデル

2.1 発話間の単語共起

発話間の単語共起とは、連続する二つの発話に出現する単語の発話間にまたがる共起をいう。連続する二つの発話のうち、先行する発話を前発話 W^{N-1}、後続する発話を次発話 W^N とする。単語共起は、前発話の一単語と次発話の一単語との組合わせである。したがって、前発話の単語の数を $|W^{N-1}| = m$、次発話の単語の数を $|W^N| = l$ とすると、単語共起は $m \cdot l$ の組合わせがある。図1の例において、発話間の単語共起は、$3 \cdot 4 = 12$ の組合わせがあり、「フロント」(w_2^{N-1}) と「チェックイン」(w_1^N) などが発話間で共起している。

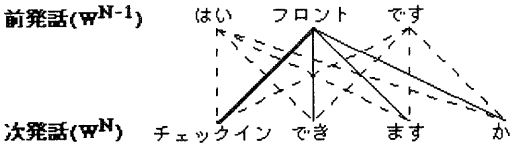

図 1: 発話間の単語共起

ここで、前発話で単語 w_i^{N-1} が生起することを条件とした、次発話で単語 w_j^N が生起する条件付き確率 $P(w_j^N|w_i^{N-1} \in W^{N-1})$ を学習データから以下のように求める。

$$P(w_j^N|w_i^{N-1} \in W^{N-1}) = \frac{f(w_j^N, w_i^{N-1})}{\sum_{W^{N-1} \in w_i^{N-1}} |W^N|}$$

ただし、$f(w_j^N, w_i^{N-1})$ は、w_j^N と w_i^{N-1} との同時生起頻度である。この学習データから得た統計値を発話間 bi-gram と呼ぶことにする。

2.2 次発話予測モデル

前節で述べた発話間の単語共起に基づいて、次発話予測モデルを定式化する。

A は音響特徴系列、C は文脈を表す。これらが条件として与えられたとき、発話 W の条件付き確率 $P(W|A,C)$ を最大化する単語系列 $\hat{W}$ を求める。

$$\begin{aligned}
\hat{W} &= arg\max_W P(W|A,C) \\
&= arg\max_W \frac{P(W,A,C)}{P(A,C)} \quad (1) \\
&= arg\max_W P(C)P(W|C)P(A|W,C) \quad (2) \\
&\simeq arg\max_W P(A|W)P(W|C)
\end{aligned}$$

式1における $P(A,C)$ および 式2の $P(C)$ は、$\hat{W}$ に依存しないため、これらかの式から省くことができる。また、A は C に依存しないものとして、$P(A|W,C)$ を $P(A|W)$ と近似する。最終的に、$P(A|W)$ が音響モデルを表し、$P(W|C)$ が次発話予測モデルを表す。

発話を $W = w_1^N w_2^N \cdots w_m^N$ として、次発話予測モデル $P(W|C)$ を以下のように定義する。

$$P(W|C) = P(w_1^N w_2^N \cdots w_m^N|C)$$

$$= P(w_1^N|C)P(w_2^N|w_1^N,C)\cdots P(w_m^N|w_1^N w_2^N\cdots w_{m-1}^N,C) \quad (3)$$

$$\simeq \prod_{j=1}^{m} P(w_j^N|C) \quad (4)$$

次発予測モデルでは、式 3 における発話内の直前の単語履歴は考慮しないこととする。

文脈 C を直前の発話 $W^{N-1} = w_1^{N-1}w_2^N\cdots w_l^N$ に限定し、式4の各項すなわち直前の発話を考慮した単語尤度を以下のように求める。

$$P(w_j^N|C) \simeq P(w_j^N|W^{N-1}) \simeq P(w_j^N|w_1^{N-1},w_2^{N-1},\cdots,w_l^{N-1}\in W^{N-1}) \simeq \begin{cases} \frac{1}{l}\sum_{j=1}^{l} P(w_j^N|w_i^{N-1}\in W^{N-1}) \\ \max_{i=1\cdots l} P(w_j^N|w_i^{N-1}\in W^{N-1}) \\ P(w_j^N|W^{N-1} \quad \arg\max_{i=1\cdots l} \log\frac{P(w_i^N|w_j^{N-1}\in W^{N-1})}{P(w_i^N)}) \end{cases} \quad (5)$$

式5のそれぞれは、

- 発話間 bi-gram の平均値
- 発話間 bi-gram の最大値
- 相互情報量最大の発話間 bi-gram

を表す。

3 次発話予測モデルの perplexity

音声認識の言語モデルと次発話予測の言語モデルとの比較 (4.3節で述べる) を perplexity(平均次単語分岐数) の観点から行う。このために、次発話予測モデルの学習にも、音声認識の言語モデルの学習データと共通のものを用いた。

3.1 学習・テストデータ

学習・テストデータの片側対話数、発話数および単語数を表 1に示す。片側対話とは、対話における二人の話者の片方のみの発話データである。テストデータの相手側対話は、Open の学習データには含まれ、Closed の学習データには含まれない。

表 1: データセット

	片側対話数	発話数	単語数
学習データ (Open)	7,099	87,533	1,336,642
学習データ (Closed)	7,099	87,495	1,333,729
テストデータ	42	551	6,236

3.2 音声認識の言語モデル

次発話予測モデルを多重クラス複合 n-gram とよばれる言語モデル [6] と比較する。このモデルは、先行単語のクラス C^f と後続単語のクラス C^t とで異なる体系のクラスを与えることを特徴とする。$W_{(j-1)}^N = w_1^N w_2^N\cdots w_{j-1}^N$ とすると、単語 w_j^N の尤度は、この言語モデルにより、以下のように与えられる。

$$P(w_j^N|W_{(j-1)}^N) \simeq P(w_j^N|C_j^t)P(C_j^t|C_{j-1}^f)$$

この言語モデル単体の perplexity は 21.39 である。

3.3 perplexity

次発話予測モデルのテストセット perplexity を表2に示す。「平均値」と「最大値」は、それぞれ、発話間 bi-gram の平均値および最大値により単語尤度を計算するモデルである。相互情報量最大の発話間 bi-gram により単語尤度を計算するモデルは、「最大値」のモデルと同じ値である。「単体」は、それらの予測モデルの単体での数値を表す。「混合」は、単語の尤度を求めるとき、音声認識の言語モデル $P(w_j^N|W_{(j-1)}^N)$ と次発話予測モデル $P(w_j^N|W^{N-1})$ とが与える単語尤度のうち、大きいものを選択する。

表 2: perplexity

	平均値		最大値	
	単体	混合	単体	混合
OPEN	181.98	18.19	95.23	16.61
CLOSED	144.46	17.73	63.83	13.97

4 評価実験

4.1 実験手順

次の手順で、次発話予測モデルを使って、再スコア付け実験を行った。

1. 音声認識の結果である単語をノードとするラティスとその直前の発話 W^{N-1} とを読み込む
2. 次発話予測モデルによりそのラティス上の各単語の尤度を求める
3. その尤度が既に与えられている音声認識モデルの尤度よりも高い場合、音声認識モデルの尤度と予測モデルの尤度とを適当に重みづけし、新たな尤度を求める
4. その尤度に基づいて、ラティスを探索し、1-best の新たな認識結果を得る

ここで、H は、発話内に限定されない発話履歴であるとする。認識モデルの尤度と予測モデルの尤度とは、以下のように λ により重み付けする。

$$P(w_j^N|H) = (1-\lambda)P(w_j^N|W_{(j-1)}^N)+\lambda P(w_j^N|W^{N-1})$$

ただし、$0 \leq \lambda \leq 1$ である。

4.2 実験条件

入力として、3.2 節で述べた言語モデルによる認識結果を用いる。学習データは、3.1 節で述べたものを用いる。学習データは、Open と Closed の 2 種類がある。次発話予測モデルも式5で述べた 3 種類のものを用いる。次発話予測モデルへの重み λ を

$$0, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, 1$$

とする。ただし、$\lambda = 0$ の場合は、再スコア付けにより得られる結果は認識結果そのものとなる。

4.3 実験結果

実験より得た 1best の単語正解率を表3 表4へ示す。図2へこれらを図示する。

平均値は、重み $\lambda = 1$ にした場合のみ、他のモデルの性能を上回ることがある。表2 においても、他のモデルに比較して、perlplexity が高い。

Open のデータでもっとも性能がよかったのが重み $\lambda = \frac{2}{3}$ の相互情報量モデルあり、認識率を 0.30 ポイント改善し、error rate を 2.1 % 削減した。

Closed のデータでもっとも性能がよかったのが重み $\lambda = \frac{2}{3}$ の最大値モデルあり、認識率を 1.32 ポイント改善し、error rate を 8.2 % 削減した。

表 3: Open データ

λ	平均値	最大値	相互情報量
0	83.9	83.9	83.9
1/3	84.0	84.1	84.1
2/5	84.1	84.1	84.2
1/2	84.0	84.1	84.1
3/5	83.8	83.9	83.8
2/3	83.7	83.7	83.6
1	82.5	82.3	82.2

表 4: Closed データ

λ	平均値	最大値	相互情報量
0	83.9	83.9	83.9
1/3	84.5	85.0	84.9
2/5	84.6	85.2	85.1
1/2	84.7	85.2	85.2
3/5	84.4	85.1	84.8
2/3	84.3	85.0	84.8
1	83.5	83.7	83.4

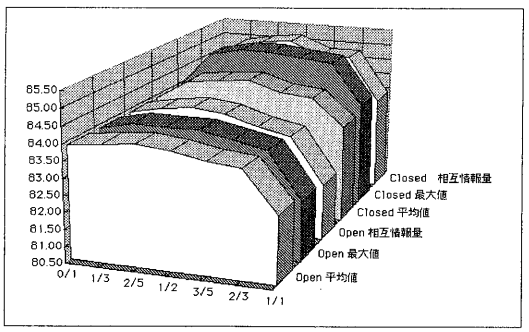

図 2: 実験結果

5 結果の分析

実験結果について分析する。この節で分析する結果は、以下の条件に基づいて得た。

- Open の学習データ
- 相互情報量最大の発話間 bi-gram に基づく次発話予測モデル
- 次発話予測モデルへの重み $\lambda = \frac{3}{7}$

この実験結果の単語認識率は、84.19 % である。音声認識結果そのままの認識率 83.87 % に比べ、認識率を 0.32 ポイントを改善し、error rate を 2 % 削減した。この実験結果 551 発話のうち、誤りがなかった、すなわち、単語正解率 100 % であった発話は 261 ある。誤りを含むにもかかわらず、今回の実験で改変のなかった発話は 231 である。改変のあった発話は 51 である。その分類を表5に示す。

5.1 内容語

話題語とは、表 5 にあげた、固有名詞、数詞、指示詞および形式名詞以外の内容語で、それらに比べて発話内容に関連が深いので、話題語という名称で分類した。これれらのなかで直接前発話の影響によって、改善されたものは、11 発話 19 単語ある。これらは、さらに、直前の発話に同じ単語が出現するものとそうでないものにわけることができる。前者は 5 発話 7 単語あり、後者は 6 発話 12 単語である。

例えば、以下の会話では、前発話の「移っ」により、「ございます」が「移ります」に改善された。

「もしよろしければこれから移っていただけますが」

「あそうですかでもちょっとゆっくりしてから移ります(ございます)」

ここで、「()」は認識誤りを、「□」は認識モデルよりも高い尤度を与えた発話間の単語共起を表す。

これに対して、次の会話では、前発話の「道」により認識誤り「方」を「角」に訂正している。

「ちょっとこの辺り詳しくないので駅までの道も教えていただけますか」

「はいホテルを出られまして北にひとつめの通りの角(方)にございます」

次の例も同様に、前発話の「熱」により認識誤り「風」を「風邪」に訂正している。

「お連れさまはお熱があるようでしょうか」

「いいや昨日から風邪気味だっていうのは聞いてたのですけど熱のほうはちょっとまだわからないのですけど」

表 5: 認識結果の改変内容

	改善		改悪		改善率 (%)	
	発話	単語	発話	単語	発話	単語
話題語	14	24	4	9	77.8	72.7
固有名詞	3	5	1	2	75.0	71.4
数詞	2	2	2	4	50.0	66.7
指示詞	3	5	0	0	100.0	100.0
形式名詞	1	1	1	1	50.0	50.0
内容語	23	37	8	16	74.2	69.8
機能語	5	6	6	7	45.5	46.2
言い淀み	4	4	2	3	66.7	57.1
広域の誤り	0	0	3	3	0.0	0.0
全体	32	47	19	29	62.7	60.3

5.2 機能語

機能語では、次発話予測モデルによるその削除が目立つ。

- 「フロントのヒラツジさんですね」を「フロントヒラツジさんですね」へ
- 「ファックスを送りたいのですが」を「ファックス送りたいのですが」へ

また、これも改悪例であるが、

- 「薬かなにかいただきたいのですけれども」を「お薬かなにかいただきたいのですけれども」へ

というように接頭辞の「お」が挿入されている。

これらの例は、正解から見ると改悪であるが、基本的に発話の意味内容を損なわない。

5.3 言い淀み

言い淀みは、それの削除や挿入については、機能語以上に発話の意味内容を損なわない。しかしながら、それを間違って他の内容語と認識することがあり、それは聞き手に余計な混乱を招く可能性がある。本実験では、認識結果「えー今日ここは２０７号室です」を「えーっとーここは２０７号室です」と改善した例がある。

5.4 広域の誤り

広域の誤りとは、「... お一夜の夕食だので空を予約したいので ...」「... 会議のいあとこわくないとこれがお見せたいのですけれども ...」など、認識結果をみても発話内容が推測できないものであり、それらを含めて 3 発話ある。これらに対し、次発話予測モデルが改善した例、また、正しい音声認識結果を広域の誤りに改悪した例、それぞれ一例存在する。表5の分類上はそれぞれ話題語の改善あるいは改悪に含んだ。改善例は、認識結果「... 手配の可否を連絡していただけますか」を

「... 手配の可否を連絡していただけますか」

とした。その直前の発話は、

「... 連絡してみますのでえっといつ頃までにお返事すればよろしいでしょうか」

となっている。ここでは、連続する発話における同一単語の出現がこの改善の原因となってはいない。

6 おわりに

次発話予測モデルを用いて、音声認識結果の再スコア付けを行った。単語認識率から見た全体の結果としては、わずかな改善効果しか認められなかったが、改善および改悪内容から見た結果については、次発話予測モデルの妥当性を示すことができた。本モデルでは、内容語についての改善効果が高く、機能語の改悪についても致命的なものは少ない。

今回分析した実験結果は、相互情報量に基づくモデルによる結果のみであった。しかしながら、単語正解率から見た3つのモデルの特徴が微妙に異なるため、改変内容についてのこれらのモデルの間での特徴が異なることが予想される。これらの特徴を明らかにすることによって、全体への効果が高いモデルを提案することがこれからの課題である。

参考文献

[1] Kuhn, K., Mori, R.De, "A Cashe-based Natural Language Model for Speech Recognition", IEEE Trans. Pattern Analysis and Machine Intelligence, vol.12, No.6, 1990

[2] Rosenfeld, Ronald, "A Maximum Entropy Approach to Adaptiove Statistical Language Modellling", Computer Speech and Langueage Vol.12, 1996

[3] Tanaka, H., Yokoo, A., "An Efficient Statistical Speech Act Type Tagging System for Speech Translation Systems", Proc of the 31th Annual Meeting of ACL, 1999

[4] Niimi, Y., Takinaga, N., Nishimoto, T., "Extraction of the Dialog Act and the Topic from Utterances in a Spoken Dialog System", Proc of ICSLP-98, 1998

[5] 岩本 秀明 , 妹尾 正身 , 荒川 直哉 , 横尾 昭男 , 森元 逞, "発話状況に基づく音声認識候補の再順序付け", 言語処理学会 第回 年次大会 , 1997

[6] 山本 博史 , 匂坂 芳紀 , "接続の方向性を考慮した多重クラス複合N—ｇｒａｍ言語モデル", 情報処理学会研究報告, SLP98-24, 1998