Internal Use Only

002

TR-IT-0183

統計的言語モデルの構築と MAP 推定を用いたタスク適応

Construction of Statical Language Model and Research of Task Adaptation with MAP Estimation

久木 和也† 政瀧 浩和 匂坂 芳典
Kazuya HISAKI† Hirokazu MASATAKI Yoshinori SAGISAKA

1996. 8

内容梗概

近年、連続音声認識における探索空間の減少を目的として、n-gram と呼ばれる統計的言語モデルが盛んに用いられている。本研究では、n-gram を実際に構築するために必要な種々の手法について述べ、実験によりそれらの手法を評価した。また、新聞などの一般的な大規模データから作成した n-gram を MAP 推定を用いて音声認識の特定のタスクに適応させる手法を提案し、評価・考察を行った。

ATR 音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories
†京都大学工学部情報工学教室
Department of Infomation Science, Kyoto University

© 株式会社 エイ・ティ・アール音声翻訳通信研究所
© 1995 by ATR Interpreting Telecommunications Research Laboratories

目次

1 はじめに 1

2 統計的言語モデルの構築 1

2.1 n-gram 1

2.2 平滑化とカットオフ 2

2.3 実験 3

2.4 考察 3

3 タスク適応 4

3.1 背景 4

3.2 MAP 推定法によるタスク適応 4

3.3 実験 6

3.4 考察 9

4 まとめ 11

1 はじめに

近年、音声認識に用いる言語モデルとして、テキストデータから単語間の統計的な情報を抽出した統計的言語モデルが盛んに使用されている。統計的言語モデルは、従来用いられていた構文規則に基づくモデルに比べ、特定の単語間の単語連接などの意味的情報も含んでおり、また文法などの知識を必要としないため構築が容易であるなどの利点がある。

音声認識に用いる統計的言語モデルとしては、n-gramが広く用いられているが、語彙が増えるに従いモデルが巨大になり、また、元のテキストデータに現れない連接が表現できないサンプリング誤差などの問題がある。これらの問題を解決するためには、信頼性の薄いデータを削除するカットオフ、および、元のテキストデータに存在しない連接に対しても確率を与える平滑化などの手法が必要となる。

また、音声認識ではシステムに合わせて発話内容を特定のタスクに限定する場合が多いが、対象のタスクに合致する大量のデータ収集は一般に困難であり、n-gramのパラメータを正確に推定することは容易ではない。そこで、まず新聞など、大量に得やすい一般的なテキストデータからモデルを構築し、それをシステムの対象とするタスクの小規模なデータで修正するタスク適応の手法により、この問題を解決する方法が考えられる。

本研究では、始めにカットオフ、および、平滑化の手法について比較検討し、さらに、MAP(最大事後確率)推定法を用いたタスク適応の手法を提案し、評価・考察を行った。

2 統計的言語モデルの構築

2.1 n-gram

n-gramは、図1のように、直前のn-1個の単語から次の単語への遷移確率を求めるモデルであり、言語を$(n-1)$重マルコフモデルで近似したものに相当する。

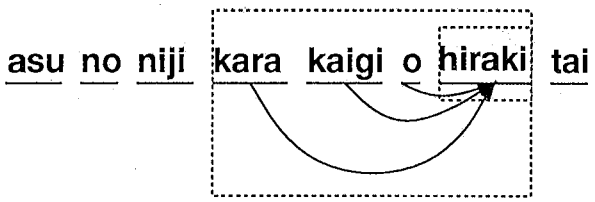

図 1: n-gram

単語系列$w_1, w_2, ..., w_m$の出現確率$P(w_1, ..., w_m)$は、

$$P(w_1, ..., w_m) = \prod_{i=1}^{m} P(w_i | w_{i-n+1}, .., w_{i-1})$$

として求める。

単語間の遷移確率は、単語列の頻度$n(w_1, \ldots, w_n)$とすると、最尤推定を用いて

$$P(w_n | w_1, \ldots, w_{n-1}) = \frac{n(w_1, \ldots, w_n)}{n(w_1, \ldots, w_{n-1})}$$

とするのが一般的である。

n-gram モデルは、n の値を大きくするほど長い単語系列を扱うことになり、より大局的な情報を含むことができるためにモデルの精度が向上すると考えられるが、パラメータ数が非常に大きくなり、また信頼性のあるパラメータを得るために必要なテキストデータの量が膨大になるために現実的には n をあまり大きくできない。逆に、n が小さい場合は少ないデータ量で信頼性のあるパラメータが得られるが、モデルとしての精度は低下する。本研究ではこの点を考慮して、実用上妥当と考えられる n=2(bigram) および n=3(trigram) を用いた。

2.2 平滑化とカットオフ

前節で述べたように、n-gram では一般的に最尤推定によって確率を計算するため、学習データに現れない単語遷移は文法・意味的に正しくても遷移確率が 0 になってしまうサンプリング誤差が発生するという問題がある。テキストデータの量を増やすことによりサンプリング誤差は小さくなるが、モデルが巨大になる問題が発生する。そこで、学習データに現れない遷移に対しても、ある基準に従って確率を与える平滑化の手法が必要となる。

代表的な平滑化の手法として、削除補間 [1] と back-off 平滑化 [2] が挙げられる。

削除補間は、n-gram の遷移確率 $P(w_i|w_{i-n+1},...,w_{i-1})$ を、

$$P(w_i|w_{i-n+1},...,w_{i-1}) = \sum_{j=0}^{n} \lambda_j P(w_i|w_{i-j+1},...,w_{i-1})$$

のように、0,1,2,..,n-gram の線形結合で表すことにより、n-gram の遷移確率が 0 になる場合を (n-1)-gram 以下で補う手法である。

back-off 平滑化は、まず一定の式に基づいてテキストデータに存在する n-gram の遷移確率値を削り、それによって得られた確率値をテキストデータに存在しない n-gram の遷移確率として分配する手法である。確率の分配は、(n-1)-gram の遷移確率に比例して行う。

代表的な back-off 平滑化では、頻度 r の n-gram の種類を n_r とし、単語列 $w_1,...,w_n$ のテキストデータでの頻度を $c(w_1,...,w_n)$ とすると、n-gram (頻度 $c(w_{i-n+1},...,w_i) = r$) の平滑化後の確率 $\tilde{P}(w_i|w_{i-n+1},...,w_{i-1})$ を、

$$\tilde{P}(w_i|w_{i-n+1},...,w_{i-1}) = \frac{(r+1)\cdot n_{r+1}}{c(w_{i-n+1},...,w_{i-1})\cdot n_r}$$

と減少させ、それによって生じる遷移確率の余剰分 $1-\sum_{w_i:c(w_{i-n+1}...w_i)>0} \tilde{P}(w_i|w_{i-n+1},...,w_{i-1})$ を

$c(w_{i-n+1},...,w_i) = 0$ なる n-gram に対し、

$$\tilde{P}(w_i|w_{i-n+1},...,w_{i-1}) = \frac{1-\sum_{w_i:c(w_{i-n+1}...w_i)>0} \tilde{P}(w_i|w_{i-n+1},...,w_{i-1})}{1-\sum_{w_i:c(w_{i-n+1}...w_i)>0} \tilde{P}(w_i|w_{i-n+2},...,w_{i-1})} \cdot \tilde{P}(w_i|w_{i-n+2},...,w_{i-1})$$

のように再帰的に分配する。本研究ではこの式を用いている。

学習データ量を多くすることによっても、サンプリング誤差が減少し n-gram の精度を向上することができる。しかし、それに伴い語彙が増加し、また、出現頻度の小さい単語遷移が増大することにより、記憶容量の制限から計算機上で扱うことが困難になる。出現頻度

の小さい単語遷移は学習テキスト中に偶然出現したものであり、信頼性は低いと考えられる。したがって、一定頻度以下の単語遷移をモデルから削除(カットオフ)し、それらの遷移確率は出現しない単語遷移と同様に平滑化により再推定することで、n-gramの精度を大きく低下させずにパラメータ数の削減を図ることが考えられる。

2.3 実験

前節で示した平滑化およびカットオフの効果を評価するため、実際のn-gramに対して適用し、テストセットパープレキシティを求めた。使用したデータはATRの旅行対話データベースで、語彙サイズ6,263, 単語総数354,354である。

平滑化の手法として、削除補間およびback-off平滑化を用いた時のそれぞれのパープレキシティ値を表1に、カットオフのしきい値を0～4まで変化させたときのパープレキシティおよびn-gramのデータサイズを表2に示す。ただし、平滑化を行わない場合に遷移確率が0となるのを防ぐために、一定値以下の確率を底上げ(flooring)している。flooringの値は0-gram/100($=1.597\times10^{-6}$)とした。

表1: 平滑化によるパープレキシティの変化

	平滑化なし	削除補間	back-off
bigram	22.483	20.158	17.399
trigram	99.959	14.234	10.211

表2: カットオフによるパープレキシティおよびn-gramのサイズの変化

cutoff	0	1	2	3	4
bigram	17.399	20.179	21.320	22.394	23.259
size	760KB	339KB	218KB	168KB	138KB
trigram	10.211	13.110	13.724	14.020	14.227
size	2397KB	603KB	357KB	267KB	215KB

2.4 考察

表1より、削除補間よりback-off平滑化の方がより低いパープレキシティを示すことが分かる。これは、削除補間では各n-gramの重みが一定であるのに対し、back-offでは存在するn-gramから減少させて得られた確率によって変動し、非線型に重みづけされることによってより精度の高い平滑化が行われているからであると考えられる。また、bigramよりもtrigramの方が、削除補間とback-off平滑化とのパープレキシティの差が大きいことから、学習データがより疎の時にback-off平滑化は削除補間よりも効果的に平滑化が行なわると考えられる。したがって、平滑化の手法としてはback-off平滑化が削除補間より有効であると考えられる。

次に、n-gramのカットオフのしきい値とパープレキシティおよびn-gramのサイズの関係について見ると、n-gramのサイズはカットオフしきい値におおよそ反比例することが分かる。しきい値1でもサイズがほぼ半分になるので、パラメータ減少の効果は大きいといえる。一方、パープレキシティはしきい値を上げるごとに増大するが、bigramではしきい値1で、trigramではしきい値4で削除補間のカットオフなしのパープレキシティとほぼ等しくなる。したがって、頻度1～2程度の単語遷移はカットオフすることにより、言語モデルとしての精度を大きく下げることなくモデルのサイズを大幅に減少させる可能であることが明らかになった。

3 タスク適応

3.1 背景

連続音声認識を用いたシステムは、使用するタスクを限定するのが一般的である。これは、語彙を少なくし、特定タスクの言語モデルを構築する方が高い認識率が期待できるからである。

特定のタスク向けの言語モデルを構築するには、その対象タスクのテキストデータを用意する必要がある。これは通常対話データの書き起こしなどから得ることができるが、書き起こしは人手で行うためデータを大量に得ることが困難であり、生成されたモデルの精度が低くなるという問題がある。そこで、新聞などの容易に得られる大規模テキストデータから一般的な言語モデルを構築し、それを小量のタスク向きデータを用いて目的のタスクに適応させることにより、言語モデルのタスクに対する精度を向上させることを考える。

タスク適応の手法には、大規模データのn-gramとタスク依存データのn-gramの頻度を混合する方式[3]や、大規模データからタスクが近いと思われる文章を取り出しタスク依存データを補う方式[4]などが提案されている。本研究では、MAP(Maximum A Posteriori: 最大事後確率)推定法[5]を用いて、統計的言語モデルのタスク適応を行う手法を提案する。

3.2 MAP推定法によるタスク適応

MAP推定法とは、学習データに対するモデルパラメータの事後確率が最大になるようにパラメータを決定する手法であり、パラメータを決定することによる誤り率を最小にする。

学習データ x に対してモデル θ を推定する式は、以下のようになる。

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} P(\theta|x)$$

これをBayes則を用いて書き換えると、

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \frac{P(x|\theta)P(\theta)}{P(x)} = \underset{\theta}{\operatorname{argmax}} P(x|\theta)P(\theta)$$

となり、最尤推定の式に $P(\theta)$ を乗じた形になる。この $P(\theta)$ は、モデル自身の事前確率を表す。考え方を変えると、これは元のパラメータを分布 $P(\theta)$ ととらえ、その分布上で観測データ x によってパラメータを修正しているともみなせる。

そこで、パラメータの事前分布を大規模データ (以下「適応元データ」と記述) から推定し、観測データとしてタスク向きの小規模データ (以下「適応先データ」と記述) を用い、適応を行うことを考える。この時、最尤推定との違いであるモデルの事前確率 $P(\theta)$ の設定が重要となってくるが、本研究ではパラメータを n-gram の確率値の集合とし、事前分布をそれの m(n-gram の個数) 次元の Gauss 分布とする。頻度を用いず、確率値とするのは、モデルの大きさによる影響を排除するためである。

まず、事象 (n-gram) が 2 個のみの場合について考える。

モデルパラメータを以下のように定義する。

- 適応元データのモデル：
  $\theta_0 = \{p_0, 1-p_0\}$
- 適応先データのモデル：
  $X = \{x, 1-x\}$
- 適応後のモデルパラメータ：
  $\theta = \{p, 1-p\}$

ここで、モデルの事前分布を平均が θ_0 となる Gauss 分布で表し、モデルが与えられた時の X の出現確率を平均が θ となる Gauss 分布で表すと、

$$P(\theta) \sim N(\theta_0, \sigma_0^2) \propto exp[\frac{(p-p_0)^2}{2\sigma_0^2}]$$

$$P(X|\theta) \sim N(\theta, \sigma^2) \propto exp[\frac{(x-p)^2}{2\sigma^2}]$$

となる。ここで、X を観測した時にモデルが θ である事後確率は、

$$P(\theta|X) \propto exp[\frac{(p-p_0)^2}{2\sigma_0^2} + \frac{(x-p)^2}{2\sigma^2}]$$

$$\propto exp[\frac{1}{2}\{(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2})p^2 - 2(\frac{p_0}{\sigma_0^2} + \frac{x}{\sigma^2})p\}]$$

となる。この式は p に関する Gauss 分布となるので、最大値を求めるためにこの分布の平均を求めることにすると、$P(\theta|X) \sim N(\theta_M, \sigma_M^2)$ $(\theta_M = \{p_M, 1-p_M\})$ とすると、

$$P(\theta|X) \propto exp[\frac{(p-p_M)^2}{2\sigma_M^2}]$$

$$\propto exp[\frac{1}{2}(\frac{1}{\sigma_M^2}p^2 - 2\frac{p_M}{\sigma_M^2}p)]$$

となる。これより、

$$\frac{1}{\sigma_M} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}$$

$$\frac{p_M}{\sigma_M} = \frac{p_0}{\sigma_0^2} + \frac{x}{\sigma^2}$$

より、

$$p_M = \frac{\sigma_0^2 x + \sigma^2 p_0}{\sigma_0^2 + \sigma^2} = \lambda\, p_0 + (1-\lambda)x \quad (\lambda = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2})$$

となる。事象が任意の個数ある場合でも、互いの事象を独立とみなせば同様な結果が得られる。

以上の式を用いて適応を行うにはモデルの分散を求める必要があるが、そのためには十分多くのモデルを作成できるだけのデータを用意しなければならず、現実には計算不可能である。そこで、適応後の n-gram の確率が適応元データと適応先データの確率の線形結合になっていることに着目する。見方を変えると、λ は適応元データのタスクにおける確からしさを表していると考えられるため、削除補間のパラメータ学習に用いられる held-out アルゴリズムを用いて λ を学習することが可能である。これは、適応先テキストデータを順次読み込み、出現するデータが適応元データに属する確率 (データが適応元に属して出現する確率 / そのデータが出現する確率) を求め、それを全体にわたって平均したものを新たに λ として繰り返し学習する手法である。λ の学習の式を以下に示す。ただし、p_0 は読み込んだ n-gram の適応元データにおける確率であり、x' は適応先データにおける確率である。

$$\hat{\lambda} = \frac{1}{N} \sum \frac{\lambda\, p_0}{\lambda\, p_0 + (1-\lambda)x'}$$

本来の held-out アルゴリズムでは、適応先データのモデル構築に使用したデータと λ の学習に使用するデータを分離する必要があるが、削除補間アルゴリズムでは分離する代わりに適応先モデルから学習の際に出現する自分自身のデータを削除 (頻度を 1 減らして計算) することによって、学習が正しく行われるようにする。上の式で x' となっているのは適応先データから現在読み込んでいる自分自身のデータを削除していることを示している。

λ の値は、先の式より n-gram ごとに独立に存在し得るが、パラメータ数が多くなり λ の推定精度が低下する。一方、全ての単語遷移に対して同一の λ を用いてタスク適応を行った場合、タスクに頻出する特定の単語遷移の確率を上げることができなくなり、適応の効果が小さくなる可能性がある。

ここで、一般的にタスクと語彙との関連性が高いことから、ある n-gram のタスクにおける確からしさはその中に含まれる単語に負うところが大きいと考えられるため、λ を n 単語列の先頭の単語の関数とする。つまり、n-gram $w_1, ..., w_n$ において、単語 w_1 が同じ n-gram は同一の λ を用いる。これによって、タスク適応の精度を下げずに λ の数を削減し、λ の推定精度を向上させることを考える。ただし、n-gram の先頭が学習セットに出現しない単語である場合、テキスト全体で学習を行った λ を用いて補間する。

3.3 実験

前節で述べたタスク適応法を用いた実験を行った。使用したデータは、2 章と同じ ATR の旅行対話データベースで、データベース全体で得られた n-gram を、分けられたトピックに適応させることにする。

使用した各データの語彙数および単語総数を表 3 に示す。学習セットおよびテストセットは、データベースの全対話から、ランダムに選択した $\frac{1}{4}$ の対話データをテストセットとし、残りのデータを学習セットとすることで振り分けた。ただし、各トピックから最低 1 対話がテストセットに選択されるようにした。

表 3: 使用データの語彙数および単語総数

トピック	学習セット		テストセット	
	語彙サイズ	単語総数	語彙サイズ	単語総数
A	2169	118225	1518	40766
B	690	8802	500	3263
C	4046	135447	2522	45838
D	1405	15494	835	5147
G	808	8376	500	3164
H	693	8075	517	2845
I	976	10415	620	3554
J	366	2693	232	801
K	892	8655	526	2608
O	1518	19306	1063	7213
R	549	5382	326	1406
S	549	5474	318	1428
T	453	2687	262	847
U	585	3581	365	1229
V	368	1742	202	546

学習セットの全体を適応元データとし、学習セットのそれぞれのトピックを適応先データとしてタスク適応を行い、それによってテストセットパープレキシティを計算したものを表4および5に示す。n-gram 自身の補間には back-off 平滑化を用いた。なお、n-gram のカットオフは行っていない。

表 4: タスク適応によるパープレキシティの変化 (bigram)

トピック	適応元	適応先	適応後	単一λ	λ
A	15.221	13.632	13.688	13.661	0.576
B	17.630	13.534	12.383	11.911	0.434
C	23.861	23.319	22.966	23.249	0.882
D	30.587	29.476	24.790	24.249	0.730
G	21.023	14.953	14.157	13.765	0.562
H	25.031	14.916	14.843	14.113	0.456
I	24.686	16.542	16.370	15.794	0.577
J	33.448	17.649	16.584	15.707	0.537
K	22.426	17.648	16.266	15.445	0.619
O	24.427	18.732	18.054	17.527	0.574
R	18.809	13.682	12.103	11.434	0.469
S	14.471	12.124	10.046	9.977	0.455
T	27.264	16.637	16.578	14.984	0.643
U	25.970	16.840	15.173	14.147	0.570
V	37.281	21.010	22.638	17.685	0.680

表 5: タスク適応によるパープレキシティの変化 (trigram)

トピック	適応元	適応先	適応後	単一λ	λ
A	10.022	9.405	9.540	9.651	0.905
B	9.556	8.247	7.469	6.967	0.672
C	17.017	17.118	16.548	16.645	0.941
D	21.259	22.766	19.124	18.552	0.916
G	12.081	9.748	9.538	9.088	0.805
H	13.714	9.089	10.152	9.369	0.769
I	13.177	10.489	10.904	10.447	0.877
J	19.387	14.136	15.016	13.610	0.890
K	12.964	11.318	10.604	10.067	0.859
O	13.145	11.010	11.063	10.718	0.834
R	11.745	9.365	8.769	7.789	0.728
S	8.441	7.690	6.813	6.428	0.768
T	15.601	12.691	12.843	12.123	0.927
U	15.194	11.089	11.577	9.988	0.819
V	25.536	18.465	19.994	16.775	0.888

表中のλは、全体で学習を行った平均のλであり、単一λとは、そのλのみを適応に用いてパープレキシティの計算を行ったものである。

3.4 考察

まずタスク適応後のパープレキシティと、適応先データのパープレキシティを比較すると、bigramではトピックA,Vで増加した他は全て減少し、trigramでは増加したトピックが8ありほぼ半数を占めた。これより、bigramでは十分に適応能力があるとみなせるが、trigramでは不十分であると考えられる。ただし、trigramでのパープレキシティ増加分の平均が0.589、減少分の平均が1.055と減少する場合の方が変化量が大きいことから、全体的には適応が成功することによる利益の方が大きいと考えられる。

しかし、適応後のパープレキシティを全体で学習を行ったλを使用した場合と比較すると、一部(bigramで1例、trigramで2例)を除き、全体で学習したλを用いた方がパープレキシティが低くなっている。これより、λを語彙に応じて分離したことによる効果がほとんどなく、パープレキシティ減少の主因は適応先データにない遷移が適応元データによって補間されることであると思われる。

単語ごとのλの値を見ると、低頻度の単語に対してはλが1や0など極端な値をとっていることが多いことから、低頻度の単語に対してはλの学習量が不十分になるとともに学習の際の単語遷移の削除による影響が大きくなるために、λを精度良く推定することができなくなっていると考えられる。そこで、低頻度の単語に対するλを学習時にマージし、学習量を増やすことで信頼性を向上させるなどしてλの推定法を改善することが必要となる。

trigramで適応が不十分だったのは、trigramではデータがより疎になるために、bigramよりn-gramの削除の影響が大きく、λの学習が不十分になってしまったためと思われる。

次に、適応先データの信頼性とタスク適応によるパープレキシティの減少量および全体で学習したλの値との関係を調べる。適応先データの信頼性はそのテストセットパープレキシティに関係する(大きいほどモデルが粗く、精度が低い)と考えられるため、bigramにおける適応先データのパープレキシティとパープレキシティ減少量およびλとの関係を図2と図3に示す。いずれも、適応先データのパープレキシティを横軸にとっている。

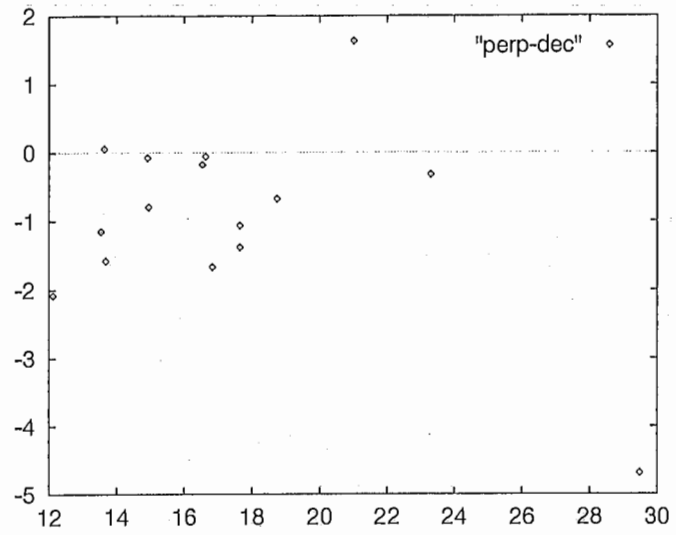

図 2: 適応先データのパープレキシティと適応による減少量の関係

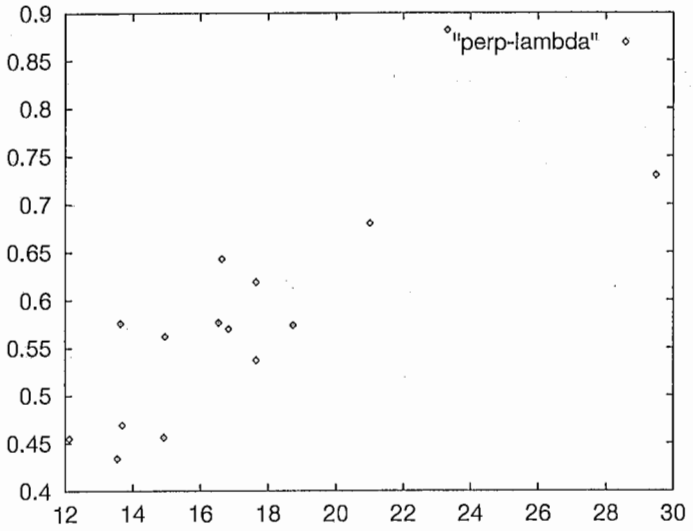

図 3: 適応先データのパープレキシティと λ の関係

これらより、適応先データのパープレキシティが大きいほど λ も大きくなり、適応先データの信頼性が薄い場合に適応元データの重みを増しているが、適応先データのパープレキシティと適応によるパープレキシティ減少量との間にはあまり相関関係が見られないことが分かる。すなわち、λ の値は適応先データの信頼性に応じてほぼ正しく変動しているが、それがパープレキシティの減少にあまり結び付いていない。これは、先に述べたように全体で学習した λ の信頼性が高くても低頻度の単語に対する λ の信頼性が低いために全体で適応効果を打ち消しているとも考えられるが、MAP 推定の事前分布の与え方に問題があると考えることもできる。そこで、λ の推定法を改善しても適応能力が弱い場合、事前分布の与え方を変更することも必要になると思われる。

4 まとめ

本研究では、まず平滑化やカットオフなど、統計的言語モデルを構築し音声認識に適応するために必要となる種々の手法について実験を行い、比較検討した。その結果、平滑化の手法としては削除補間より back-off 平滑化の方がよりよい結果を示した。また平滑化に合わせて適切なしきい値でカットオフを施せばモデルの精度を大きく低下させることなくパラメータ量が大幅に削減できることが可能であることが分かった。

次に、統計的言語モデルを音声認識タスクに適応させるために、パラメータ推定に MAP 推定法を、パラメータの事前分布に Gauss 分布を用いたタスク適応法を提案し、実際に適応実験を行った。その結果、bigram では大半のデータでパープレキシティが減少したが、データの補間による効果が主で、タスクに頻出する n-gram の重みを増加させるなどの効果は得られなかった。低頻度の単語に対して適応係数の学習量が不十分であると思われることから、適応係数の学習・推定法の改善が必要である。改善した場合でも効果がなければ、事前分布の変更を考慮する必要がある。

参考文献

[1] F.Jelinek, R. L. Mercer. Interpolated estimation of Markov Source Parameters from Sparse Data. Proc. Workshop Pattern Recognition in Practice, pp.381-37,1980.

[2] Slava M. Katz. Estimation of probabilities from Sparse Data for the Language model Component of a Speech Recognizer. IEEE Trans. on Acourstics, Speech and Signal Proc., Vol. ASSP-35, pp.400-401, 1987.

[3] 伊藤彰則, 代島直人, 丸山敦, 加藤正治, 好田正紀. 大語彙言語データベースからの N-gram 構築とタスク適応の検討. 情報処理学会研究報告, SLP-11-5, 1995.

[4] Pablo Fetter, Alfred Kaltenmeier, Thomas Kuhn, Peter Regel-Brietzmann. Improved Modeling Of OOV Words In Spontaneous Speech. ICASSP96-1-136, Vol.1, pp534-537, 1996.

[5] 松岡達雄, Chin-Hui Lee. 最大事後確率推定法 (MAP 推定法) によるオンライン話者適応. 信学技報 SP93-133, 1994.