音声言語情報処理 14－13
(1996. 12. 13)

対話音声認識のための事前タスク適応の検討

伊藤 彰則　好田 正紀

山形大学 工学部 電子情報工学科
〒992 米沢市城南 4 丁目 3-16
{aito,kohda}@ei5sun.yz.yamagata-u.ac.jp

あらまし　連続音声認識のための統計的言語モデルを構築するためには，大量の言語データが必要となる．しかし，特定のタスクの言語データを大量に集めるには大きな人手と時間がかかる．そこで本稿では，既存の大量の言語データに特定タスクの言語データを少量混合することによって，N-gram 言語モデルの性能の改善を試みる．ここでは特定タスクとして観光案内対話タスクを想定し，さまざまな大量データと混合したときの性能を調査した．その結果，ある程度タスクドメインに共通性があるデータの場合に性能が改善できることがわかった．また，形態素単位の N-gram を構築する場合，大量データと小量データの形態素解析の一貫性が重要であることが明らかになった．

キーワード　連続音声認識，統計的言語モデル，N-gram，タスク適応

Task adaptation of a stochastic language model for dialogue speech recognition

Akinori Itō and Masaki Kohda

Faculty of Engineering, Yamagata University
Jōnan 4-3-16, Yonezawa-shi, 992 Japan
{aito,kohda}@ei5sun.yz.yamagata-u.ac.jp

Abstract　A stochastic language model (SLM) is indispensable for continuous speech recognition. Generally, large corpus of the task domain is required to make a good SLM. When making a SLM for a specific task domain, it is ideal to obtain large number of sentences of the domain. But it takes large time and effort to collect linguistic data of a specific domain, especially of a spoken dialog domain. In this paper, we investigated possibility of making a good N-gram SLM using small corpus of a specific domain with task independent large corpus. Sightseeing information dialog task was chosen for the specific task, and we examined several kinds of corpora for task independent corpus. We carried out experiments to measure perplexity of the adapted N-gram model. From the experiments, it is found that the adaptation improved perplexity of the model when the task domain of the small and large corpora are similar. The results also showed that the coherence of morphemic analysis of the small and large corpora greatly affects the perplexity of the adapted model.

key words　continuous speech recognition, stochastic language model, N-gram, task adaptation

1 はじめに

高精度な連続音声認識の実現のためには，良い言語モデルが不可欠である．近年は，N-gramをはじめとする統計的言語モデルの利用が盛んになってきている．

統計的言語モデルの構築のためには，大量の言語データの存在が前提となる．特定のアプリケーション向きの言語モデルを作成するためには，その特定のアプリケーションで用いられる言語表現を大量に集めてモデルを作るのが理想的である．しかし一般には大量の言語データを収集するのは容易ではなく，多くの時間と人手を必要とする．新聞記事等，電子化された大量のデータが容易に入手できる分野では，それを用いた言語モデルの構築が行なわれている[1]．しかし，特定の分野について，新聞記事から構築したモデルが有効であるとは限らない．

本稿では，このような場合を想定し，事前に小量の特定分野のテキストを用い，これと大量で一般的なテキストを混合することによって，N-gram言語モデルの適応を行なう手法について検討する．

2 タスク適応

タスク適応の方式は，現在のところ大きく3つに分類することができる．1つは，認識結果を用いて言語モデルを逐次更新していく方法である[2, 3]．これをここでは「逐次型タスク適応」と呼ぶ．2つめは，あらかじめいくつかの分野を想定しておき，入力がどれにあたるかを推定する方法である[4]．これをここでは「混合型タスク適応」と呼ぶ．3つめは，適応したい分野のテキストを少量用意しておき，それを使ってモデルを変更する手法である[5]．これを「事前型タスク適応」と呼ぶ．これらの方法は相補的なものであり，互いに組みあわせて使うことができる．実際，事前型適応と逐次型適応を組み合わせる試み[6]や，逐次型適応と混合型適応の組み合わせ[7]も提案されている．

今回は，これらの適応の中で，事前型タスク適応について検討する．事前型タスク適応の中でも，ある程度広いタスクの中からその中の特定タスクに適応する場合と，まったく違うタスクに適応する場合が考えられる．本稿では，大量テキストと小量テキストのタスクがまったく異なる場合を想定し，このような場合にそもそも適応が可能かどうかについて検討を行う．

本稿では，タスク独立な大量テキストを「タスク独立テキスト」，特定タスクの小量テキストを「適応テキスト」と呼ぶことにする．また，評価に用いる特定タスクのテキストを「評価テキスト」と呼ぶ．このとき，タスク適応の手順は次のようになる．

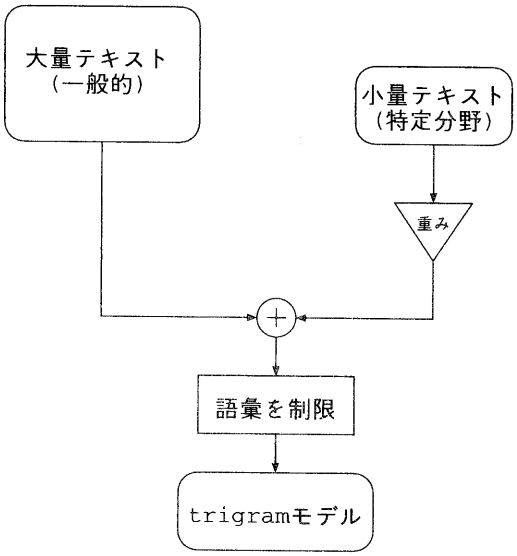

図 1: 言語モデルのタスク適応

1. 適応テキストを，重みつきでタスク独立テキストに加える．重みをwとすると，適応テキスト中でm回出現した単語は，タスク独立テキスト中でwm回出現した単語と同等に扱われることになる．
2. このようにしてできたテキストの中で，出現頻度が一定回数未満の単語を，未知語を表す記号に置きかえる．(語彙の制限)
3. テキストから統計を取り，N-gramモデルを構築する．

これらの処理のブロック図を図1に示す．

タスク独立テキストと適応テキストの混合については，MAP推定やEMアルゴリズムなどを用いる方法が提案されている[8]．MAP推定[8]では，適応テキストから求めた単語zの出現確率を$f(z)$，タスク独立テキストから求めた確率を$f'(z)$，適応テキストの単語数をm，タスク独立テキストの単語数をm'とするとき，

$$P(z) = \left(\frac{m}{m+\epsilon m'}\right) f(z) + \left(\frac{\epsilon m'}{m+\epsilon m'}\right) f'(z) \quad (1)$$

として求めている．今回我々が用いた方法では，タスク独立テキストと適応テキストの大きさの組み合わせごとにwを変えて適応実験を行い，最

適値を求めた．この方法は，unigramの場合

$$P(z) = \left(\frac{wm}{wm+m'}\right) f(z) + \left(\frac{m'}{wm+m'}\right) f'(z) \quad (2)$$

となり，$\epsilon = w^{-1}$とするとMAP推定に一致する．bigram以上のモデルでは，$c(s)$,$c'(s)$をそれぞれ適応・タスク独立テキストでの単語列sの出現回数としたとき，MAP推定の場合

$$P(z|x) = \left(\frac{m}{m+\epsilon m'}\right)\frac{c(xz)}{c(x)} + \left(\frac{\epsilon m'}{m+\epsilon m'}\right)\frac{c'(xz)}{c'(x)} \quad (3)$$

今回用いたモデルの場合

$$P(z|x) = \left(\frac{wc(x)}{wc(x)+c'(x)}\right)\frac{c(xz)}{c(x)} + \left(\frac{c'(x)}{wc(x)+c'(x)}\right)\frac{c'(xz)}{c'(x)} \quad (4)$$

となり，若干異なっている．

N-gramのモデルとしては形態素単位trigramを用い，linear back-off [10]によって平滑化を行なっている．

音声認識における話者適応実験では，適応のためのデータの大きさだけを変えて実験を行うことが多い．これに対して，今回の実験では，適応テキストとタスク独立テキストの大きさの両方を変えながら適応実験を行う．これは，そもそもタスク独立テキストを「大量」に使うことに意味があるかどうかを調べるためである．

3 語彙と評価尺度

この実験では，タスク独立テキストと適応テキストとを混合した後，語彙を制限している．また，評価尺度として，補正パープレキシティ[9]を用いている．ここでは，その意味について考察してみる．

ただでさえ少ないテキストコーパスについて語彙を制限することは，確率推定をますます不利にするように見える．しかし，ここで語彙制限を用いることには，次のような効果がある．タスク独立テキストは適応テキストと全く異なるタスクの文章であるため，そこに出現する多くの語彙は，適応先のタスクにおいてはあまり意味がない．そこで，出現頻度の低い単語を未知語にすることで，タスク独立テキスト中のタスク依存な単語を排除することができる．同時に，このような低頻度単語には名詞(特に固有名詞)が多く含まれるため，低頻度単語を未知語にすることで，タスク独立な発話の骨格のようなものが抽出できる可能性がある．例えば，後述する音響学会コーパス3000文セットにおいて，頻度20未満の単語を<UNK>という記号に置きかえると，

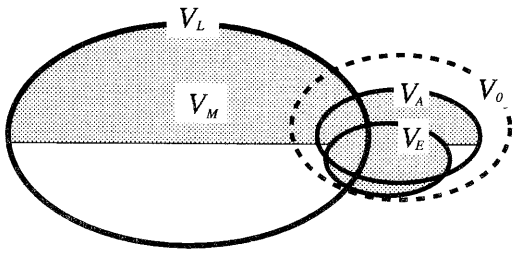

図 2: テキストと言語モデルの語彙の関係

置換前：

今日 の 昼食 な ん で す けど
どんな タイプ の レストラン が いい
で す か

置換後：

<UNK> の <UNK> な ん で す けど
どんな <UNK> の <UNK> が いい で す
か

のようになり，タスクに依存しない表現だけが残る．評価テキスト中の未知語は，この<UNK>と同等に扱われる．評価テキスト中の未知語にも固有名詞が多いため，うまく語彙を選べば，タスク独立テキストの未知語部分に評価テキストのタスク依存部分を「はめる」ことができる．

ここで，テキストと言語モデルの語彙について考察してみる．タスク独立テキストに出現する単語の集合をV_L，適応テキストに出現する単語の集合をV_A，評価テキストに出現する単語の集合をV_Eとする．また，単語集合Vの中でn回以上出現した単語の集合を$S_n(V)$とする．この時，補正パープレキシティによる評価は，言語モデル全体の語彙V_Mを

$$V_M = S_n(V_L \cup V_A) \cup V_E \quad (5)$$

とした場合に相当する．この研究の目的は，特定タスク用の言語モデルを作成することであるから，言語モデルの理想的な語彙をV_0とすると，$V_0 \supseteq V_A \cup V_E$である．これらの関係を図2に示す．

ここで，問題は次のように整理される．

1. 適応テキストに出現しない語彙$V_0 - V_A$をどうやって見つけるか．
2. 学習サンプルのない語彙の確率をどう推定するか．

1つめの問題については，$V_0 - V_A$ のうち，「高頻度の単語であって，たまたま V_A に出現しないもの」のみを推定しようとしている．すなわち，タスク独立テキストのうち高頻度単語だけを抽出することによって，多くのタスクに共通な語彙だけを抽出しようとしている．この意味では，言語モデルの語彙を

$$V_M = S_n(V_L) \cup V_A \cup V_E \qquad (6)$$

とすることも考えられる．式(5)と(6)の2つを予備実験によって比較した結果，式(6)による語彙の設定には問題があることがわかった．式(6)において n をある程度大きく設定すると，V_A の中の低頻度単語の確率よりも未知語の確率の方が大きくなり，それが悪影響となってパープレキシティが増加する現象が見られた．このため，以下の実験では式(5)を用いている．これについては，低頻度単語を未知語とする枠組み自体に検討の余地があるので，今後の検討課題としたい．2つめの問題については，学習テキストに出現せず，評価テキストに出現する語彙

$$V_U = V_E - (V_E \cap S_n(V_L \cup V_A)) \qquad (7)$$

の確率として，未知語の確率を再配分している．すなわち，コンテキスト x における単語 $w_u \in V_U$ の確率として，

$$P(w_u|x) = \frac{P(\texttt{<UNK>}|x)}{|V_U|} \qquad (8)$$

としている．ここで，$P(\texttt{<UNK>}|x)$ は，コンテキスト x で未知語が出現する確率を表す．

4 実験に用いたコーパス

適応・評価テキストとして，日本音響学会データベースの模擬対話のうち，京都観光案内に関する4対話(KIT0001D ～ 0004D)を用いた．このうち KIT0001D(「京都1」と呼ぶ)を評価用として固定し，KIT0002D ～ 0004D(「京都2～4」)を適応用に用いた．タスク独立テキストとして，音響学会データベース44対話のうち京都観光案内を除く40対話(ASJ)，ATR連続音声データベースの国際会議に関するキーボード対話(ATR)，EDRコーパス(EDR)の3種類を用いた．これらのコーパスは，形態素解析基準が互いに異なっている．これらのコーパスの諸元を表1に示す．

表 1: 実験に用いたコーパス

	文セット	文数	単語数	単語種
EDR	10	10	204	113
	20	20	416	218
	50	50	1179	514
	100	100	2332	927
	300	300	6948	2149
	600	600	13925	3605
	1200	1200	26979	5765
	2500	2597	63437	9375
	5000	5195	125960	15226
	10000	10390	251767	23578
ATR	50	50	556	176
	100	100	1189	300
	300	300	3815	667
	600	600	7989	1126
	1200	1200	17212	1739
	2500	2500	35173	2580
	5000	5000	69677	3438
ASJ	50	50	1753	356
	100	100	3842	629
	200	200	8798	906
	400	400	13229	1135
	800	800	19059	1482
	1500	1500	43978	1965
	3000	3000	96779	3593
評価	京都1	117	1542	328
適応	京都2	117	1254	288
	京都2+3	248	2856	471
	京都2+3+4	341	3976	569

5 評価実験(1)— 適応の効果

これまで述べた方法でモデルを作成し，評価実験を行なった．ここで行なった実験は，大きく分けて3つある．1つは，タスク独立テキストを用いてモデルを作り，京都1を評価する実験である．2つ目は，適応テキスト(京都2～4)を用いてモデルを作り，京都1を評価する実験である．そして最後に，タスク独立テキストと京都2～4を組みあわせた実験を行なった．

5.1 タスク独立テキストのみを用いた場合

ASJ, ATR, EDRの各コーパスについて，語彙を変えながらモデルを作成した．この結果を図3に示す．3つのコーパスの中では，ASJコーパスがもっとも良い結果となった．タスク独立テキストと適応テキストはどのコーパスの場合もタスク独立であるが，ASJコーパスは対話テキストであるため，広い意味でのタスクが共通しているためであろう．また，形態素解析基準が同じであることも要因の一つと考えられる．これらの要因については後で検討を行う．EDRコーパスでは，5000文セットから300回以上出現した単語を語

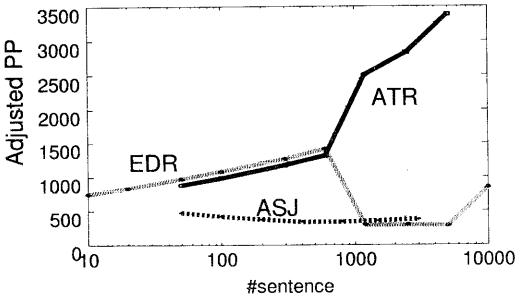

図 3: タスク独立テキストのみによる補正パープレキシティ

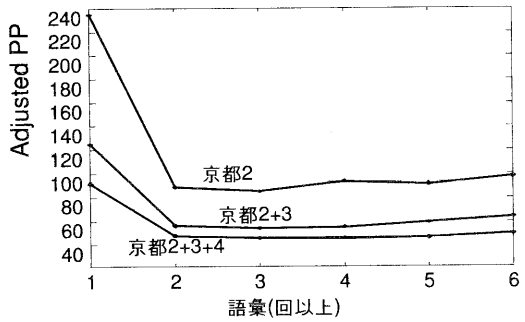

図 4: 適応テキストのみによる補正パープレキシティ

彙としてモデルを作成したときの性能が最も良く，パープレキシティは 280.3 であった．なお，5000 文セットで 300 回以上出現した単語は 37 種類で，出現頻度は全体の 44% に相当した．

5.2 適応テキストのみを用いた場合

適応テキストである「京都 2」「京都 2+ 京都 3」「京都 2+ 京都 3+ 京都 4」の 3 つを用いてモデルを作成し，評価を行なった．この場合も，語彙を変化させながら実験を行ない，最適な語彙を探した．この結果を図 4に示す．ここで横軸は語彙 (回以上)，縦軸は補正パープレキシティである．最適な点を抜き出すと，次のようになる．

テキスト	語彙 (回以上)	語彙数	補正 PP
京都 2	3	88	85.51
京都 2+3	3	142	54.32
京都 2+3+4	4	146	45.69

5.3 タスク独立・適応テキストを共に用いた場合

次に，タスク独立テキストと適応テキストを両方用いた場合について調べてみた．この場合は，適

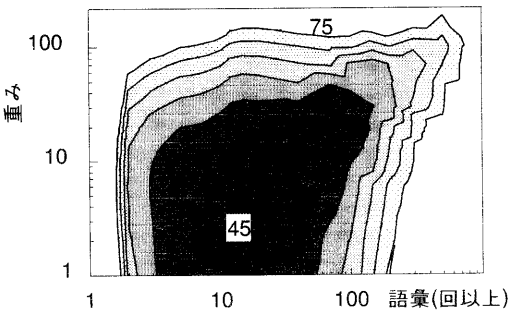

図 5: 語彙，重みとパープレキシティ

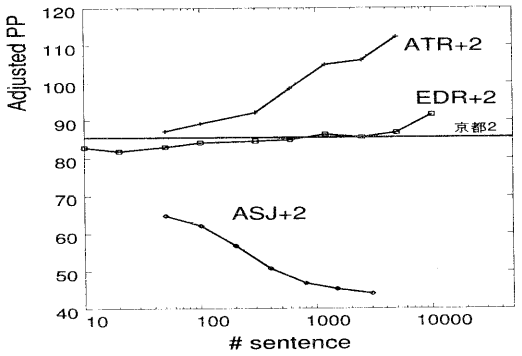

図 6: タスク独立テキストの文数に対する最適なパープレキシティ (京都 2 で適応)

応テキストの重みと全体の語彙がパラメータになる．ある語彙と重みについて，補正パープレキシティがどのようになるかを図 5に示した．この図は，ASJ コーパス 3000 文セットに対して京都 2 を適応させた結果を示している．横軸は語彙となる単語の最低出現頻度，縦軸は重みを表わしている．この図から，語彙と重みの最適値に一定の関係があることが見てとれる．

各文セットについての最適な補正パープレキシティを図 6,7,8に示す．横軸はタスク独立テキストの文数，縦軸は補正パープレキシティである．図中の直線は，適応テキストのみを用いた結果である．この結果から，タスク独立テキストとして ASJ コーパスを使った場合と，それ以外の場合で性質が大きく違うことがわかった．タスク独立テキストとして ASJ コーパスを使った場合は，タスク独立テキストの量を増やすほどパープレキシティが低下していく．これに対して，EDR と ATR コーパスを使った場合には，タスク独立テキストの文数を増やすことによって逆にパープレキシティが増加した．この場合，EDR および

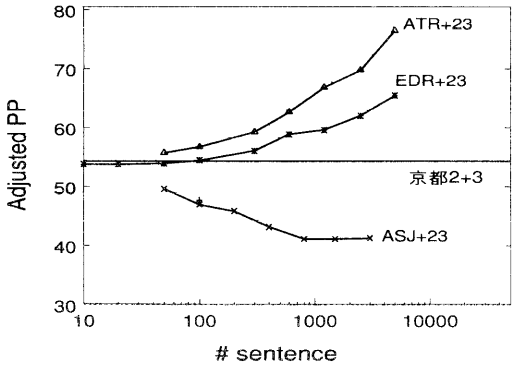

図 7: タスク独立テキストの文数に対する最適なパープレキシティ (京都 2+3 で適応)

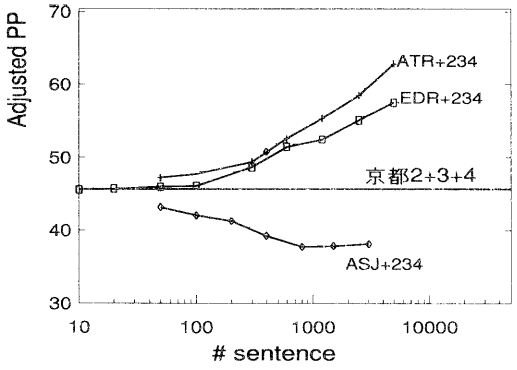

図 8: タスク独立テキストの文数に対する最適なパープレキシティ (京都 2+3+4 で適応)

ATR コーパスから作られるモデルは，語尾や助詞など，非常に一般的な部分のモデルとしてしか使われていないようである．EDR コーパスにおいて，20 文セットで 5 回以上出現した単語を語彙とする (「京都 2」を適応させた場合の最適値) という条件では，EDR コーパスの中で未知語以外として残るのは，ほとんどが助詞や語尾などであった．

6 評価実験 (2)— 形態素解析の影響

評価実験 (1) では，タスク独立テキストの種類によって，適応時のパープレキシティが大きく違うという結果になった．その原因として，2 つ考えられる．1 つは，タスク独立テキストと適応・評価テキストのタスクの差である．ここで用いた 3 つのタスク独立テキストは，適応・評価テキストとの独立性について，ある程度の差がある．これを表にしてみると，次のようになる．

表 2: 再解析前後の単語数

コーパス	文数	再解析前		再解析後	
		単語数	単語種	単語数	単語種
ATR	5000	69677	3438	65716	3069
EDR	10390	251767	23578	201425	24969

コーパス	書き / 話し言葉	メディア
ASJ	話し言葉	対話 (書き起こし)
ATR	(話し言葉)	キーボード対話
EDR	書き言葉	新聞・雑誌

適応・評価テキストはもともと ASJ コーパスの一部であるから，ASJ と種類が一致している．ATR コーパスは完全な話し言葉ではないが，ある程度適応・評価テキストに近いと思われる．EDR コーパスは，適応・評価テキストとかなり異なる文章である．この差が適応時のパープレキシティの差となって現れた可能性がある．

もう 1 つの可能性は，形態素解析基準の違いである．ASJ・ATR・EDR の 3 つのコーパスは，それぞれ違う基準で形態素解析されている．形態素解析基準の違いの多くは助詞や活用語尾などであるが，これらの中には出現頻度の高いものが多いため，パープレキシティに大きく影響する可能性がある．

この 2 つの可能性のうち，形態素解析基準による影響について調べる実験を行った．ASJ コーパスを解析した形態素解析基準[11]を用いて，ATR コーパスおよび EDR コーパスを解析し，この結果を用いて評価実験を行ってみた．ASJ コーパスは，自動形態素解析の結果を目視で修正したものであるが[11]，ATR および EDR コーパスについては，文節数最小基準で自動解析した結果をそのまま用いた．そのため，形態素解析誤りがある程度含まれていると思われるが，特にチェックは行っていない．ATR と EDR コーパスについて，再解析前後の単語数の違いを表 2に示す．

このようにして適応実験を行った結果を図 9に示す．このグラフは適応に京都 2 を用いた場合で，横軸はタスク独立テキストの文数，縦軸は補正パープレキシティである．適応に京都 2+3，京都 2+3+4 を用いた場合も，ほぼ同じ傾向が見られた．ATR コーパスについては，形態素解析基準を適応・評価テキストとそろえることで，パープレキシティが大幅に減少した．また，再解析前はタスク独立テキストの量を増やすことでパープレキシティが増加していたが，再解析を行うことで，タスク独立テキスト量の増加とともにパープレキシティが減少するようになった．ただし，そ

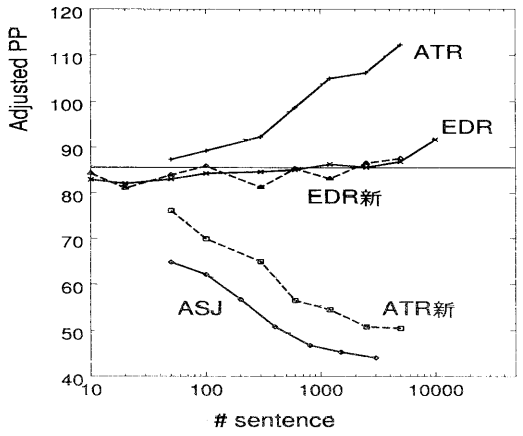

図 9: 再形態素解析後のパープレキシティ (京都 2 で適応)

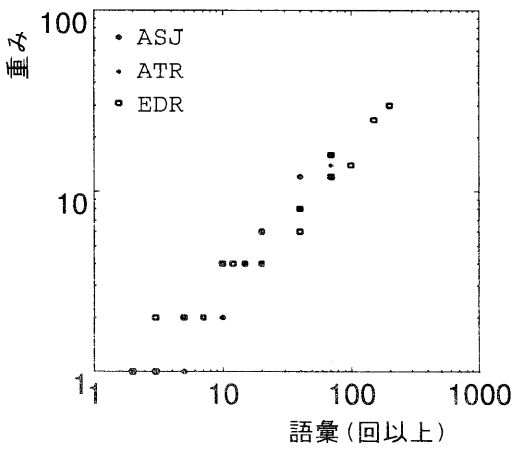

図 10: 各タスク独立テキストについての最適な語彙と重み (best 3)

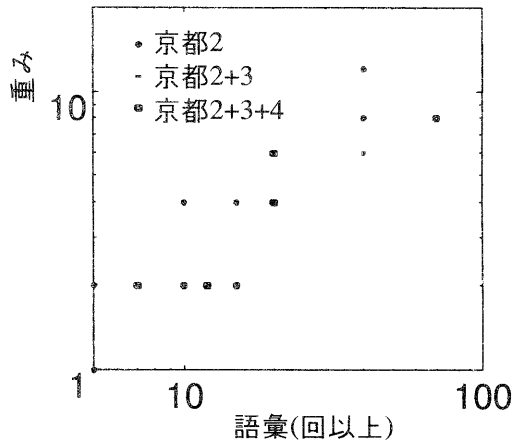

図 11: 各適応テキストについての最適な語彙と重み (best 3)

れでも ATR と ASJ コーパスの間にはパープレキシティの差が見られるが，この差の原因はよくわかっていない．この部分が「対話の書き起こし」と「キーボード対話」の差の影響なのかもしれない．EDR コーパスについては，再解析をしてもパープレキシティの改善は見られなかった．このことから，対話の言語モデルを作成するためには，タスク独立テキストとして対話コーパスを使う必要があると言えそうである．

7 最適な語彙と重みについて

今回の実験では，各タスク独立テキストと適応テキストの条件ごとに，語彙と重みを変化させて最適値を探しているが，これらのパラメータは自動的に決定できる方が望ましい．重みについては，EM アルゴリズムを用いて自動決定する方法が提案されている[8]．ここでは，パラメータを変化させて求めた最適値がどのようになっているかを分析してみる．

図 10は，ASJ, ATR, EDR の各タスク独立テキスト (ATR,EDR は再形態素解析したもの) に京都 2 を適応させた場合，各文セットでの語彙と重みの組みあわせのうち，補正パープレキシティが低かった組みあわせ 3 つをプロットしたものである．組みあわせを 1 つにしなかった理由は，語彙と重みの変化幅が粗かったため，組みあわせが最適値からずれている可能性があるためである．このグラフから，最適な語彙と重みの間には 1 次元的な関係がありそうだということがわかる．語彙 n[回以上]，重み w について一次回帰直線を計算すると，

$$w = 1.33 + 0.157n \qquad (9)$$

となった (相関係数 0.973).

タスク独立テキストを ASJ に固定し，適応テキストを変えた場合の各文セットでの語彙と重みの組みあわせのうち，補正パープレキシティが低かった組みあわせ 3 つをプロットしたものを図 11に示す．図 10と比較すると，傾向が不明確になっている．上と同じく一次回帰直線を計算すると，

$$w = 0.687 + 0.157n \qquad (10)$$

となり (相関係数 0.854)，傾きは図 10と同じになった．

8 まとめ

大量のテキストに特定分野の小量テキストを混合することで，タスクに合った統計的言語モデルを作る方法について検討を行なった．その結果，ある程度近い文体のテキストであれば，適応の効果が見られることがわかった．また，形態素単位のN-gramの場合，元のテキストに対する形態素解析の一貫性が重要であることが明らかとなった．

今後は，適応テキストと大量テキストのタスクの差の影響，最適な語彙と重みの自動的な決定法などについて調べていきたい．

参考文献

[1] 大附他：「新聞記事を用いた大語彙連続音声認識の検討」，信学技報 NLC95-55, SP95-90 (1995-12)

[2] R. Kuhn et al. "A Cache-Based Natural Language Model for Speech Recognition" IEEE Trans. PAMI, vol. 12, No. 6, pp. 570-583 (1990)

[3] R. Rosenfeld: "A hybrid approach to adaptive statistical language modeling", Proc. ARPA Human Language Technology Workshop, pp. 76–81 (1994-3)

[4] R. Iyer et al.: "Language modeling with sentence-level mixtures", Proc. ARPA Human Language Technology Workshop, pp. 82–87(1994-3)

[5] A. I. Rudnicky: "Language modeling with limited domain data", Proc. ARPA Spoken Language Systems Technology Workshop, pp. 66–69 (1995-1)

[6] S. Matsunaga *et al.*: "Task adaptation in stochastic language models for continuous speech recognition", Proc. ICASSP-92, vol. I, pp. 165–168 (1992)

[7] R. Iyer et al.:"Modeling Long Distance Dependence in Language: Topic Mixture vs. Dynamic Cache Models", Proc. ICSLP-96, pp. 236–239 (1996)

[8] M. Federico: "Baysian Estimation Methods for N-gram Language Model Adaptation", Proc. ICSLP-96, pp. 240–243 (1996)

[9] J. Ueberla, "Analysing a simple language model – some general conclusion for language models for speech recognition", Computer Speech and Language, vol. 8, No. 2, pp. 153–176 (1994-4)

[10] P. Placeway et al.: "The Estimation of Powerful Language Models from Small and Large Corpora", Proc. ICASSP-93, vol. II, pp. 33–36 (1993)

[11] 伊藤彰則，牧野正三：「音声認識のための文節構造モデルとその制約について」，情処学会研究報告 95-SLP-6-7, pp.43-50 (1995-5)