新聞記事のトライグラムによるモデル化と適応化

赤松 裕隆　　中川 聖一

豊橋技術科学大学 情報工学系
〒441 愛知県豊橋市天伯町雲雀ヶ丘 1-1

1 はじめに

音声処理技術，及び，自然言語処理技術の向上により，自然発話 (対話) 音声を認識するシステムが広く研究されている。このようなシステムでは，音声の音響的特徴を記述すべく音響モデルと，音声の言語的特徴を記述すべく言語モデルが必要となる。従来の研究例を見ると，音響モデルとしては隠れマルコフモデル (HMM) を，言語モデルとしては文脈自由文法を用いているものが多い。しかし，文脈自由文法のみでは自然発話のバリエーションを十分にカバーできないため，言語情報を統計的にモデル化するN-gram ベースの言語モデルが音声認識にも応用されている。

言語モデルを N-gram ベースで構築する場合 (ルールベースで記述するのとは異なり)，大量の学習データが必要となる。最近では各種データベースが幅広く構築され，言語モデルの作成に新聞記事などの大規模なデータベースを利用した研究が行なわれている [1]。

しかし N-gram はタスクに依存するのでタスクに関するデータベースを用いて構築される必要がある。

例えば、観光案内対話タスクを想定し、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行なっている [2, 3]。

本研究ではタスクへの適応化のために、同一ジャンルの過去の記事を用いる方法について検討した。

2 言語モデルの評価尺度

2.1 エントロピーとパープレキシティ

エントロピーとパープレキシティは共に、対象とする文集合の複雑さを定量的に示す指標で、その文集合が複雑なほど、それぞれの値は大きくなる。

言語モデル G において、文 (単語列)$W_i = {w_1}^{L_i}$の出現確率を $P(W_i)$ とすれば、文集合 $W_1, W_2, \cdots, W_N$のエントロピーは次式で求められる。

$$H(L) = -\sum_{i=1}^{N} P(W_i) \log P(W_i)$$

テキスト文の連接を $W = W_1 W_2 \cdots W_N = w_1 w_2 \cdots w_T$とすれば、テストセットのエントロピーは

$$H(L) = -\log P(W)$$

で示される。トライグラムを用いた場合、P(W) は

$$\begin{aligned} P(W) &= P(W_1)P(W_2)\cdots P(W_N) \\ &= P(w_1|*\ \#)P(w_2|\#\ w_1) \\ &\quad P(w_3|w_1\ w_2)\cdots P(w_T|w_{T-2}\ w_{T-1}) \end{aligned}$$

となる。(注:#は文頭を、*は文末を示す)

この時、一単語当たりのエントロピーは

$$H_0(L) = -\frac{\sum_i \log P(W_i)}{\sum_i L_i}$$

また、言語の複雑さ・パープレキシティは

$$PP = 2^{H_0(L)}$$

と定義される。

2.2 補正パープレキシティ

CMU SLM toolkit[4] では語彙に含まれないものは全て未知語のカテゴリにまとめられ、語彙に含まれる形態素と等価に未知語のカテゴリは扱われる。そのため語彙数のセットが小さい程 (カバー率が小さい程)、パープレキシティは小さくなるということになり好ましくない。そこで評価テキスト中に出現した未知語の種類 m と、未知語の出現回数 n_u を用いてパープレキシティを補正する [3, 5]。補正パープレキシティは

$$APP = (P(w_1 ... w_n) m^{-n_u})^{-\frac{1}{n}}$$

で与えられる。

2.3 面種別での学習と評価

タスク依存の言語モデルを構築する場合、ターゲットとするタスクに関するデータのみを用いて学習する方がよいと考えられる [7]。

学習と評価用のコーパスとして毎日新聞の1994 年版の記事を用いた。形態素解析には RWCP が提供している毎日新聞形態素解析データを、電総研の伊藤氏の作成した括弧除去ツールで加工し、使用している [6]。学習には 1 月から 11 月までの記事を用い、評価には 12 月の記事を用いた。毎日新聞には全部で 13 面種に分類

表 1: トレーニングデータ

面種	形態素数	形態素種類数	全面種で学習, 面種別で評価				面種別で学習と評価			
			PP	APP	未知語の種類	cover 率	PP	APP	未知語の種類	cover 率
国際	1312728	32171	25.14	72.62	27382	89.77 %	17.05	34.52	27171	93.19 %
経済	1910712	41310	24.81	77.08	36406	89.36 %	16.28	35.86	36310	92.59 %
家庭	1305652	43182	27.10	117.38	38365	86.35 %	19.50	55.53	38182	90.25 %
スポーツ	1953408	36943	25.26	100.98	32214	86.88 %	17.52	36.79	31943	92.97 %
社会	5078734	74074	22.91	91.41	69077	87.76 %	18.17	62.27	69074	89.11 %
全面種	23454347	165755	24.66	102.72	160755	88.28 %	———	———	———	———

表 2: テストデータ

面種	形態素数	形態素種類数	全面種で学習, 面種別で評価				面種別で学習と評価			
			PP	APP	未知語の種類	cover 率	PP	APP	未知語の種類	cover 率
国際	125675	10084	41.11	101.02	6189	89.86 %	52.67	99.46	5702	92.76 %
経済	173612	13142	42.25	116.76	8959	89.00 %	52.98	111.64	8433	91.88 %
家庭	113454	12041	42.29	143.92	8236	86.66 %	55.38	142.47	7659	89.62 %
スポーツ	160158	11642	44.13	144.03	7932	87.04 %	51.16	109.72	7236	91.55 %
社会	443885	24712	36.75	126.07	19835	87.73 %	40.70	123.27	19751	88.96 %
全面種	2051584	56210	41.53	155.88	51213	87.99 %	———	———	———	———

されているが、「社説」、「科学」、「読書」などの面種にはデータが少な過ぎるので、面種別の結果は省いてある。登録した形態素数は 5000 で、N-gram の学習と評価に CMU SLM toolkit [4] を使用した。

その結果を表 1,2に示す。なお、20000 形態素による全面種のカバー率は約 96.40 % で、全面種で学習して全面種のテストデータで評価した場合、PP = 71.61, APP = 105.15 であった。

面種別の言語モデルは学習データが少な過ぎてパープレキシティがトレーニングデータでは小さ過ぎテストデータでは大きく、全面種の言語モデルを用いた場合の方がテストデータに対してはパープレキシティは小さい。そのため、全面種の記事で構築した言語モデルをターゲットとする面種に適応化する手法をとる。

3 適応化法

新聞記事では数日間に渡って関連のある記事が載っていることがある。そこで記事の評価時に、過去の数日間の記事で言語モデルを適応化しておけば、適応前より精度のよい言語モデルが出来ると考えられる。

ここで、N-gram 言語モデルの適応化には MAP 推定 (最大事後確率推定)[2, 8, 9] を用いる。適応化サンプルを与えた後の推定値は次式のようになっており、常に推定前の条件確率と現在与えたサンプルとの間で、サンプル数で重み付けされた線形補間の形になっている。

$$prob = \frac{\alpha \cdot N_0 \cdot prob_0 + N_1 \cdot prob_1}{\alpha \cdot N_0 + N_1}$$

α 重み
N_0 標準言語モデルの総数
N_1 適応化サンプルの総数
$prob$ MAP 推定後の条件確率 (N-gram 確率)
$prob_0$ 標準言語モデルでの条件確率
$prob_1$ 適応化サンプルでの条件確率

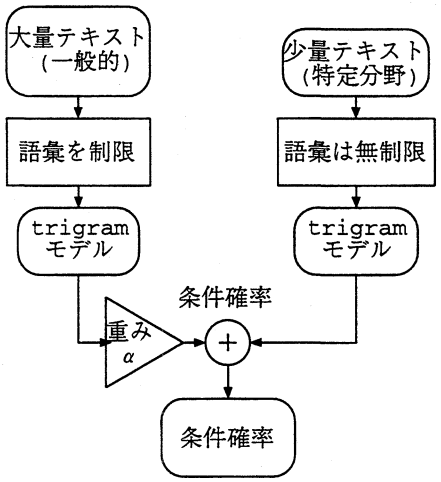

図 1: MAP 推定のブロック図

今回の実験では標準言語モデルと適応化サンプルのモデルの 2 つを構築しておき、バックオフを行なってスムージングした 2 つの条件確率を用いて MAP 推定を行なっている。この過程

のブロック図を図1に示す。標準言語モデルでは学習サンプルで出現頻度の高い形態素5000に限定した。適応化サンプルでは語彙を限定せず、全ての形態素を語彙リストに登録した。そのため、2つのモデルの語彙リストは独立している。

4 評価実験

4.1 実験方法

実験手順を以下に示す。

1. 標準言語モデルを構築する。
2. 標準言語モデルでテストデータのパープレキシティ(複雑さ)を求める。
3. 標準言語モデルを面種別の適応化サンプルでMAP推定し、同様にテストデータのパープレキシティを求める。

4.2 実験結果

実験結果を表3,4を示す。これらの表より

- 適応化前より適応化後の方がパープレキシティが小さくなること
- 5日より14日間の適応化サンプルの方がパープレキシティが小さくなること
- 6カ月前の数日間より直前の数日間の記事での適応化の方がパープレキシティが小さくなること

が分かる。

4.3 適応化サンプルの量とパープレキシティの関係

国際面とスポーツ面で適応化サンプルの期間を5,14日,1,2,3,6カ月にして求めたパープレキシティと補正パープレキシティを図??に示す。これを見ると、適応化サンプルの量が多くなるほど、パープレキシティが小さくなること、日数が多くなるにつれてパープレキシティが飽和していくことが分かる。

5 むすび

6カ月前の数日間の記事より直前の数日間の記事で適応化した方がパープレキシティが小さくなった。このことは言語モデルがジャンルだけでなく時間にも依存するものであることを示すものである。MAP推定による適応化を行なうことによってパープレキシティが下がる。このことは言語モデルの適応化がうまくいっていることを示している。

ただ、適応化サンプルの量を多くするほどパープレキシティが小さくなる傾向があり、N-gramベースでの言語モデルを少量サンプルで適応化させることは限界があると考えられる。

参考文献

[1] 大附克年, 森岳至, 松岡達雄, 古井貞熙, 白井克彦:「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,pp.63-68(1995-12)

[2] 伊藤彰則, 牧野正三:「対話音声認識のための事前タスク適応の検討」, 情報処理学会研究報告 95-SLP-14-13,pp.91-98(1996-12)

[3] 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」, 情報処理学会研究報告 95-SLP-6-7,pp.43-50(1995-5)

[4] R.Rosenfeld: "The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation", Proc. ARPA Spoken Language Systems Technology Workshop, pp.47-50(1995)

[5] 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)

[6] 伊藤克亘, 松岡達雄, 竹沢寿幸, 武田一哉, 鹿野清宏:「大語彙連続音声認識のためのテキストデータ処理」, 音響学会秋季論文集,3-3-10,pp.105-106(1996)

[7] 松永昭一, 山田智一, 鹿野清宏:「音節連鎖情報のタスク適応化」, 情報処理学会第42回全国大会(2),6D-5,pp.114-115 (1991-3)

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

[9] 政瀧浩和, 匂坂芳典, 久木和也, 河原達也:「MAP推定を用いたN-gram言語モデルのタスク適応」, 信学技報 SP96-103,pp.59-64(1997-1)

表 3: パープレキシティー (面種別)

記事・面	適応前 PP	直前の記事で適応 5日分 PP	α	直前の記事で適応 14日分 PP	α	6カ月前の記事で適応 5日分 PP	α	6カ月前の記事で適応 14日分 PP	α
国際	41.11	39.08	0.04	38.22	0.07	40.27	0.05	39.75	0.09
経済	42.25	40.72	0.06	39.59	0.10	41.41	0.08	40.89	0.20
家庭	42.29	41.25	0.05	40.81	0.09	41.64	0.07	41.44	0.20
スポーツ	44.12	39.66	0.03	37.62	0.04	42.51	0.05	41.51	0.09
社会	36.75	35.33	0.20	34.27	0.20	36.24	0.30	35.65	0.40

表 4: 補正パープレキシティー (面種別)

記事・面	適応前 APP	直前の記事で適応 5日分 APP	α	直前の記事で適応 14日分 APP	α	6カ月前の記事で適応 5日分 APP	α	6カ月前の記事で適応 14日分 APP	α
国際	74.93	59.32	0.04	51.93	0.07	65.56	0.05	57.63	0.09
経済	85.98	66.87	0.06	55.71	0.10	71.07	0.08	60.62	0.20
家庭	96.04	79.11	0.05	66.16	0.09	81.01	0.07	69.00	0.20
スポーツ	100.99	62.70	0.03	50.84	0.04	76.58	0.05	62.89	0.09
社会	90.30	59.64	0.20	47.08	0.20	64.57	0.30	51.03	0.40

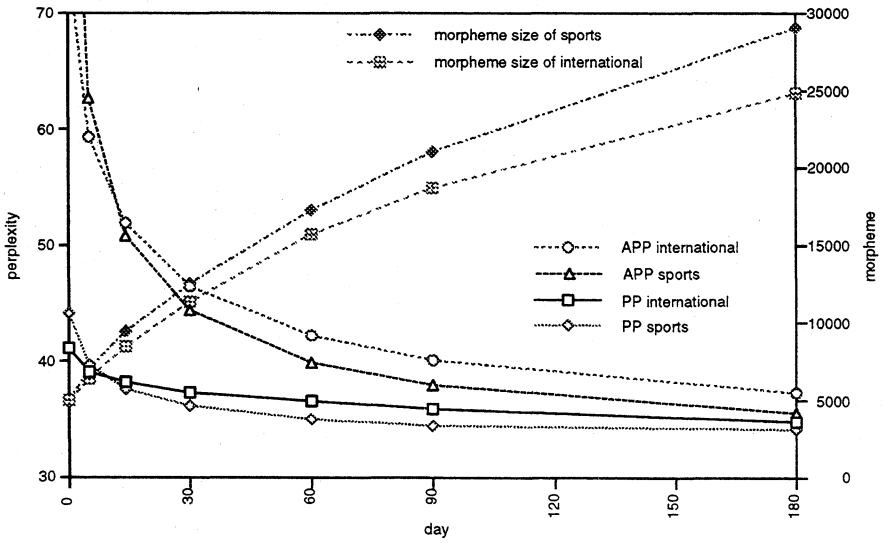

図 2: MAP 推定の日数とパープレキシティの関係