社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74（9）
2008 / 12 / 9

PLSA言語モデル適応におけるアニーリングスケジュールの評価

加藤 正治[†,††] 小坂 哲夫[†] 伊藤 彰則[††] 牧野 正三[††]

† 山形大学 大学院理工学研究科
〒 992-8150 山形県米沢市城南 4-3-16
†† 東北大学 大学院工学研究科
〒 980-8579 仙台市青葉区荒巻字青葉 6-6-05
E-mail: †katoh@yz.yamagata-u.ac.jp

あらまし 潜在的意味解析 (PLSA) の学習においてアニーリングを行うことは局所最適解に陥ることを防ぐ意味で重要である．本報告では，アニーリングスケジュールを連続関数で定義することで明確化し網羅的な比較検討をする．提案法を「日本語話し言葉コーパス (CSJ)」の講演音声で評価しところ，増加関数に基づくアニーリングスケジュールで作成した言語モデルは，28.7% のパープレキシティ削減と 5.3% の単語誤り率の改善を得た．
キーワード 潜在的意味解析，アニーリングスケジュール，最尤推定，言語モデル適応

Evaluation of annealing schadule for PLSA language model adaptaion

Masaharu KATO[†,††], Tetsuo KOSAKA[†], Akinori ITO[††], and Shozo MAKINO[††]

† Graduate School of Science and Engineering, Yamagata University
4-3-16 Johnan, Yonezawa, Yamagata 992-8510 Japan
†† Graduate School of Engineering, Tohoku University
6-6-05 Aramaki aza Aoba, Aoba-ku, Sendai 980-8579 Japan
E-mail: †katoh@yz.yamagata-u.ac.jp

Abstract Probabilistic Latent Semantic Analysis (PLSA) is a powerful statistical laguage model. However the PLSA has the local maxima problem. To overcame this problem, the EM annealing algorithm has been proposed. In this paper, we designed annealing schedule β with some continuous functions. As a result, we found that increasing functions and square root functions are the best for annealing schedule. In the experiment, we obtain 28.7% perplexity reduction and 5.3% word error rate reduction.
Key words probabilistic latent semantic analysis, annealing, EM algorithm, language model adaptation

1. はじめに

統計的言語モデルである N-gram モデルは非常に強力であり，音声認識の分野でも重要な役割を果たしている．しかしながら，N-gram モデルは単語の生起が直前の $n-1$ 単語にのみ依存する言語モデルである．現実的に用いられる依存性の範囲は n 単語に限られ，n が 4 を越えることはほとんどない．そのため，大域的な単語の依存性を考慮できない．これに対して，長距離文脈情報を用いたさまざまな統計的言語モデルが提案されている．[1]

その中でも，文書と話題の関係に着目したモデルとして「確率的潜在意味解析」(PLSA:Probabilistic Latent Semantic Analysis) [2] がある．PLSA は隠れ属性である話題（topic）を基準に文書（document）から単語（word）の出力確率を推定するモデルである．PLSA に関連するこれまでの研究では，話題数は数十から数百であり，より大規模な実験は行われていない．大規模な比較実験の妨げとなるのは，計算量である．PLSA の計算は，記憶容量・計算量が話題数，文書数，語彙サイズの積に比例するため，大語彙を扱う場合に話題数を増やして実験することは計算量的に困難であった．我々は，これまでの研究で並列化による高速学習を実現している [3]．

パラメータの推定には EM アルゴリズム（expectation-maximization algorithm）を用いる．EM アルゴリズムは推定の過程で局所最適解に陥る場合がある．Tempered EM などのアニーリングスケジュールが提案されている．しかしながら従来の研究では，乗数を「徐々に小さく（大きく）していく」などの表現でしか語られておらず，網羅的な評価は報告されていない．

本報告では，アニーリングスケジュールで用いるアニーリングの乗数を連続関数で設計することを提案し，比較結果を報告する．

作成した言語もデルを「日本語話し言葉コーパス (CSJ)」の講演音声で評価する．まず，書き起こし文に対するパープレキシティによる評価を行う．次に，音声認識実験での評価を示す．

2. PLSA 言語モデル

2.1 PLSA 言語モデル

PLSA(Probabilistic Latent Semantic Analysis) は，大規模なコーパスにおいて，文書 (document) と単語 (word) の出現確率を「話題 (topic)」により特徴付けてモデル化する手法である．文書 $d \in \mathcal{D}$ が単語 $w \in \mathcal{W}$ を生成する確率は，話題 $t \in \mathcal{T}$ を用いて次の様に定式化される．

$$P(w|d) = \sum_{t \in \mathcal{T}} P(w|t)P(t|d) \tag{1}$$

ここでの PLSA 言語モデルは unigram 言語モデルとして定義される．N-gram 言語モデル（$N \geq 2$）の場合については unigram rescaling を用いる [4]．例えば，trigram 言語モデル（$N = 3$）の場合は次式で求める．

$$P(w_i|w_{i-2}^{i-1}, d) \propto \frac{P(w_i|d)}{P(w_i)} \cdot P(w_i|w_{i-2}^{i-1}) \tag{2}$$

2.2 EM 学習法に基づくパラメータ推定

PLSA のモデルは，EM アルゴリズムを用いて尤度を最大化するように計算される．

E-step

$$Q(t, w, d, r) = P(w|t)^{(r)} P(t|d)^{(r)} \tag{3}$$

$$P(t|w, d)^{(r)} = \frac{Q(t, w, d, r-1)}{\sum_{t' \in \mathcal{T}} Q(t', w, d, r-1)} \tag{4}$$

M-step

$$R(t, w, d, r) = N(w, d) P(t|w, d)^{(r)} \tag{5}$$

$$P(w|t)^{(r)} = \frac{\sum_{d \in \mathcal{D}} R(t, w, d, r)}{\sum_{w' \in \mathcal{W}} \sum_{d \in \mathcal{D}} R(t, w', d, r)} \tag{6}$$

$$P(t|d)^{(r)} = \frac{\sum_{w \in \mathcal{W}} R(t, w, d, r)}{\sum_{t' \in \mathcal{T}} \sum_{w \in \mathcal{W}} R(t', w, d, r)} \tag{7}$$

ここで，$N(w, d)$ は文書 d における単語 w の出現回数である．また，r は現在の学習回数である．

適応文 $\hat{d}$ が与えられたとする．$P(w|t)$ は文書に依存しないので，これを固定して (4)，(7) を用いて適応文 $\hat{d}$ に対する $P(t|\hat{d})$ を推定する．式 (1) より，適応後の unigram 確率 $P(w|\hat{d})$ を求める．

2.3 アニーリングスケジュール

実際の計算では局所最適解に落ち入ることを防ぐため，式 (4) の $P(t|d, w)$ に $\beta \quad (0 < \beta \leq 1)$ を乗じて計算する．

$$P(t|w, d)^{(r)} = \frac{Q(t, w, d, r-1)^{\beta(r)}}{\sum_{t' \in \mathcal{T}} Q(t', w, d, r-1)^{\beta(r)}} \tag{8}$$

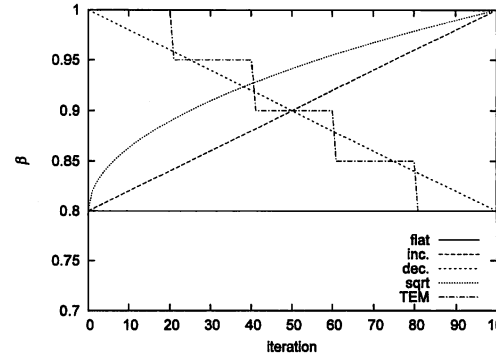

図 1 アニーリングスケジュール
Fig. 1 Annealing schedules.

3. アニーリングスケジュールの比較

3.1 連続関数に基づくアニーリングスケジュール

アニーリングスケジュールは，EM 推定のときに局所最適解に陥ることを防ぐために重要である．しかしながら，これまで用いられているアニーリングスケジュールは，段階的に乗数を引き下げる，または，複数の候補から最良の値を選択するといった指針しかなく，網羅的な比較は報告されていない．

ここでは，アニーリングスケジュールを連続関数で定義することで明確化する．また，初期値の違いも含めて比較する．関数は，一定 (flat)，単調増加 (inc.)，単調減少 (dec.)，増加関数 (sqrt) の 4 つを用いる．

一定値 $\beta^{(r)} = \beta^{(0)}$
単調増加 $\beta^{(r)} = \beta^{(0)} + (1 - \beta^{(0)}) \cdot (r/R)$
単調減少 $\beta^{(r)} = 1 - (1 - \beta^{(R)}) \cdot (r/R)$
増加関数 $\beta^{(r)} = \beta^{(0)} + (1 - \beta^{(0)}) \cdot \sqrt{(r/R)}$

$\beta^{(0)}$ を初期値とする．関数が単調減少のときはパラメータとして収束値 $\beta^{(R)}$ を与える．$\beta^{(r)}$ は第 r 回目の学習での乗数である．R は，収束までの繰り返し回数であり，ここでは学習回数と等しいものとする．

図 1 に学習回数 100 回，$\beta^{(0)} = 0.8$ のときのアニーリングスケジュール $\beta^{(r)}$ を示す．比較として，TEM も表示している．

3.2 アニーリングスケジュールの比較

学習テキストには，『日本語話し言葉コーパス』(CSJ: Corpus of Spontaneous Japanese) を用いる．

テキストはデータベース付属の XML から作成する．転記基本単位 (IPU:Inter-Pausal unit) 毎に文頭・文末記号を挿入し，SUW(Short-Unit Word) を単語として採用する．

学会講演，および，模擬講演の計 2668 講演を用いる．延べ単語数は 8486301 語である．語彙として学会講演・模擬講演それぞれで 2 回以上出現した単語を登録する．語彙サイズは 47309 語である．

基準言語モデルとして単語 trigram を作成する．bigram, trigram のカットオフは，それぞれ，1, 3 とする．

評価用のテキストとして学会講演男性話者の 10 講演 (test

表 1 アニーリングスケジュールの違いによる比較
Table 1 Effect of annealing schedules.

topic /iter.	$\beta^{(0)}$ ($\beta^{(R)}$)	APP flat	inc.	dec.	sqrt
100 /100	0.5	105.73	99.04	105.73	94.31
	0.6	105.73	96.04	105.31	92.58
	0.7	105.70	93.35	98.95	91.14
	0.8	97.97	91.11	95.48	90.45
	0.9	92.74	89.93	91.54	90.14
	1.0	91.49	—	—	—
1000 /1000	0.5	105.73	80.33	105.73	77.35
	0.6	105.74	79.33	105.74	76.44
	0.7	101.97	77.74	100.19	75.78
	0.8	95.97	76.39	92.52	75.44
	0.9	84.76	76.70	86.19	76.49
	1.0	82.84	—	—	—

表 2 TEM と連続関数を用いたアニーリングの比較
Table 2 Proposed annealling shedules and TEM.

topic	iter.	β		APP
100	100	TEM	1.0 → 0.8	97.17
		dec.	1.0 → 0.8	95.48
		sqrt	0.8 → 1.0	90.45
1000	1000	TEM	1.0 → 0.8	92.10
		dec.	1.0 → 0.8	92.52
		sqrt	0.8 → 1.0	75.44

set 1) を用いる．評価単語数は 30837 語である．基準言語モデルでの補正パープレキシティは 105.81 である．

PLSA 言語モデルの初期値として $P(t|d)$ には一様乱数を，$P(w|t)$ には unigram 確率 $P(w)$ を用いる．話題数と学習回数の組み合わせは (100,100) と (1000,1000) の 2 通りに設定する．適応テキストには正解テキストを用い，1 講演毎にモデルを適応する．

従来法との比較として，話題数 100，学習回数 100 回 ($\beta = 1.0, 0.95, 0.9, 0.85, 0.8$: 繰り返し各 20 回)，話題数 1000，学習回数 1000 回 ($\beta = 1.0, 0.95, 0.9, 0.85, 0.8$: 繰り返し各 200 回) を比較する．

適応文として正解テキストを用いる．アニーリングスケジュールと初期値の違いによる性能の比較を表 1 に示す．話題数 100，学習回数 100 回では，増加関数 (sqrt) では $\beta^{(0)} = 0.8$ で 90.45，単調増加 (inc.) では $\beta^{(0)} = 0.9$ のとき 89.93 と，最も良い結果を得られている．一定値 (flat) や減少関数 (dec.) では，$\beta^{(t)} < 1.0$ の条件における結果は $\beta^{(t)} = 1.0$ の場合よりも悪く．アニーリングの効果が得られていない．一定値 (flat) で初期値 $\beta^{(0)} \leqq 0.6$，単調減少 (dec.) で収束値 $\beta^{(R)} = 0.5$ とするとき，$P(t|d)$ はすべて一定値 (話題数の逆数) となり，$P(w|t)$ は各 t 毎に等しくなった．これは，平滑化の効果が大きすぎてモデルが学習されなかったためだと考えられる．最も良い結果は，話題数 1000，学習回数 1000 回，$\beta^{(0)} = 0.8$ で増加関数 (sqrt) を用いたときに得られ，補正パープレキシティの値は 75.44 であった．このときのベースライン言語モデルからの改善率は 28.7% である．

表 3 音響分析条件
Table 3 Acoustical analysis.

sampling	16kHz, 16bit
frame shift	8msec
analysis window	Hamming, 32msec
feature vector	logEng,MFCC(12), Δ logEng,ΔMFCC(12), $\Delta\Delta$ logEng$\Delta\Delta$MFCC(12) (39 order)

従来法 (TEM) と提案法の比較を行った結果を表 2 に示す．従来法を [TEM] の行に，従来法と同様に β を減少させる連続関数を [dec.] の行に，提案法で最も良かった増加関数を [sqrt] の行に示す．

結果として，提案法の増加関数 (sqrt) が最も良い結果となった．β を減少させる場合の比較では，従来法 (TEM) と単調減少 (dec.) の差はほとんどなかった．したがって，β をステップ状に減らすことは，β を一様に減らす場合と同等であると言える．

3.3 音声認識実験による評価

次に，提案法によって学習した PLSA 言語モデルを音声認識によって評価する．音響分析条件を表 3 に示す．ベースライン音響モデルは，CSJ 学会・模擬講演で学習した HM-Net 3000 状態 16 混合のモデルを利用する．[5] ベースライン言語モデルは，3.2 節と同じものを用いる．適応テキストにはベースライン言語モデルの認識結果を用い，1 講演毎に適応する．

評価には，学会講演男性 10 名を用いる．ベースライン認識システムの単語誤り率は 20.38% である．

話題数は 1000，学習回数は 1000 とする．適応には認識結果を用いる．

適応した言語モデルの補正パープレキシティを表 4 に単語誤り率を表 5 に示す．

最も良い結果は，単調増加 (inc.)，$\beta^{(0)} = 0.5$ のときに得られていて，WER は 19.30% で改善率は 5.3% である．PLSA 単独での性能評価として，ベースライン音響モデルと PLSA 適応を組み合わせた結果を表 6 に示す．認識性能改善の評価をするために，t-検定を用いる．講演毎に対応をとりの認識率の差で評価する．

$$t_0 = \frac{|\bar{d}|}{\sqrt{\sum_{i=1}^{n}(d_i - \bar{d})^2}/\sqrt{n(n-1)}} \tag{9}$$

ここで，講演 i での認識率の差を d_i とし，平均を $\bar{d}$ とする．自由度は $n-1$ である．危険率 5% (両側検定) で検定する場合，$t_0 > 2.2622$ であればは両者に有意差がある．最も良い結果は，単調増加 (inc.)$\beta^{(0)} = 0.5$ で得られている．適応前との比較では，$t_0 = 2.3758$ であり有意な差が認められる．

3.2 節で，補正パープレキシティでの評価が最も良い結果を得たのは増加関数 (sqrt)，$\beta^{(0)} = 0.8$ でり認識性能は 19.69% である．このときは，$t_0 = 2.1718$ であることから有意な差は認められない．

PLSA 言語モデル適応と MLLR 音響モデル適応とを併用する場合について評価する．ベースラインの認識結果 (単語誤り

表 4 認識結果を用いた PLSA 適応の APP

Table 4 Unsupervised PLSA adaptation(adjusted perplexity).

topic /iter.	$\beta^{(0)}$ ($\beta^{(R)}$)	APP			
		flat	inc.	dec.	sqrt
1000	0.5	105.72	83.94	105.72	81.87
/1000	0.6	105.72	82.72	105.72	81.04
	0.7	101.66	81.99	100.05	80.66
	0.8	96.02	81.20	92.86	80.74
	0.9	86.21	80.92	87.48	80.94
	1.0	85.90	—	—	—

表 5 認識結果を用いた PLSA 適応での単語誤り率

Table 5 Unsupervised PLSA adaptation (word error rate).

topic /iter.	$\beta^{(0)}$ ($\beta^{(R)}$)	WER [%]			
		flat	inc.	dec.	sqrt
1000	0.5	20.39	19.30	20.39	19.51
/1000	0.6	20.39	19.46	20.39	19.65
	0.7	20.32	19.58	20.21	19.75
	0.8	19.92	19.55	19.82	19.69
	0.9	19.82	19.60	19.72	19.79
	1.0	19.76	—	—	—

表 6 ベースライン音響モデルと PLSA 適応

Table 6 Baseline acostic model and PLSA.

n-gram	WER[%]	t_0
baseline	20.38	—
PLSA inc. $\beta^{(0)} = 0.5$	19.30	2.3758
PLSA sqrt $\beta^{(0)} = 0.8$	19.66	2.1718

表 7 MLLR 適応と PLSA 適応

Table 7 MLLR and PLSA adaptation.

n-gram	WER[%]	t_0
baseline	17.72	—
PLSA inc. $\beta^{(0)} = 0.5$	16.95	2.5967
PLSA sqrt $\beta^{(0)} = 0.8$	17.13	2.3096

率：20.38%）を MLLR 適応のラベルと PLSA 適応の文章として用いる．MLLR 適応は回帰クラスタの自動設定法 [6] を用い，フレームしきい値を 2000 とする．

MLLR 適応音響モデルと PLSA 適応の結果を表 7 に示す．ベースラインと比較すると，MLLR 適応での WER は 17.72% で改善率は率は 13.0% である．PLSA 適応と併用する場合の改善率は単調増加 (inc.)$\beta^{(0)} = 0.5$ のとき 16.9% である．このときの PLSA による改善率は 4.3% である．

また，3.2 節で補正パープレキシティに対して最も評価の良かった増加関数 (sqrt)$\beta^{(0)} = 0.8$ の場合は 15.9% の改善率を得る．このときの PLSA による改善率は 3.3% である．MLLR 適応をした場合の結果について，PLSA の認識性能改善の効果を評価する．PLSA を併用することでの改善率は 3.3% である．このとき，$t_0 = 2.3096 > 2.2622$ であり危険率 5% で有意な差が認められる．

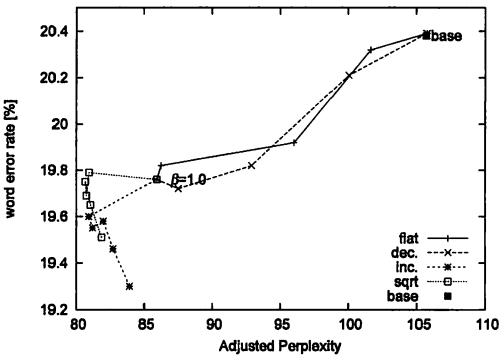

図 2 補正パープレキシティと単語誤り率の関係

Fig. 2 Adjusted perplexity and word error rate.

4. 考　　察

PLSA 言語モデルの EM 学習におけるアニーリングスケジュールを連続関数で設計した．設計した関数の中では，単調増加 (inc.) と増加関数 (sqrt) を用いると補正パープレキシティによる評価で有意差が認められた．一方，一定値 (flat)，単調減少 (dec.) では性能改善の効果はみられない．特に，一定値の関数の初期値 $\beta^{(0)}$，または，単調減少関数の収束値 $\beta^{(R)}$ を小さくすると PLSA の平滑化の効果が大きくなりすぎる．学習の終了段階で係数を小さくすることは適応効果を得られない．

単語誤り率と補正パープレキシティの関係を図 2 に示す．両者の評価は一致するとは言えない．特に，単語誤り率と補正パープレキシティの評価が良いところでばらつきが大きくなっている．例えば，増加関数 (sqrt) で初期値 $\beta^{(0)} = 0.8$ とした場合は，補正パープレキシティで有意な差が認められても，認識性能では有意な差は得られていない．ベースラインの音響モデルは話者間での性能のばらつきが大きく PLSA 単独の性能改善の効果が隠れてしまったと考えられる．

一方，音響モデルの適応を前提とした場合では増加関数 (sqrt) 初期値 $\beta^{(0)}0.8$ のとき PLSA 適応の有無に対して認識性能改善に有意な差が認められる．このとき，PLSA 適応モデルは単独の評価と同一ものを使用している．音響モデル適応と言語モデル適応の相乗効果が現れたといえる．

5. ま と め

連続関数によるアニーリングスケジュール設計法を提案し，CSJ 講演音声で提案法を評価した．アニーリングについては増加関数 (sqrt)，または，単調増加 (inc.) が良いことを示した．PLSA 言語モデルでは補正パープレキシティでベースラインのモデルより 28.7% の改善率を得ている．

音声認識実験では，PLSA 単独で 5.3% の認識率改善を得ている．MLLR 音響モデル適応を併用した場合の改善率は 16.2% であった．このときの，PLSA 言語モデル適応の効果は t-検定により有意な差が認められた．

文　献

[1] J.R. Bellegarda, "Statistical language model adaptation: review and perspectives," Speech Communication, vol.42, no.1, pp.93–109, jan 2004.

[2] T. Hoffmann, "Probabilistic latent semantic indexing," Proc. of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99), pp.50–57, 1999.

[3] 加藤正治，小坂哲夫，伊藤彰則，牧野正三，"PLSA 言語モデルの並列化による高速学習," 日本音響学会研究発表会講演論文集，2008，秋季 3-1-1.

[4] D. Gildea, and T. Hofmann, "Topic-based language models using EM," Proc. Eurospeech, 2003, 2003.

[5] 堀貴明，加藤正治，伊藤彰則，好田正紀，"状態クラスタリングによる HM-Net の構造決定法の検討," 電子情報通信学会論文誌. D-II，vol.J81-D2，no.10，pp.2239-2248，1998.

[6] 加納淳也，加藤正治，伊藤彰則，好田正紀，"話者照合における話者モデルの MLLR 適応の検討," 情報処理学会研究報告 1996-SLP-026，vol.99，no.108，pp.55-60，1999.