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

2006－SLP－64 (40)
2006／12／22

情報量基準で語彙分割したPLSA言語モデルによる話題・文型適応

栗山 直人† 鈴木 基之† 伊藤 彰則† 牧野 正三†

†東北大学大学院工学研究科 〒980-9579 宮城県仙台市青葉区荒巻字青葉 6-6-05

E-mail: †{kriya,moto,aito,makimo}@makino.ecei.tohoku.ac.jp

あらまし PLSAは言語モデルの文脈適応に一般的に用いられる手法である．このPLSAの新しい利用方法を提案する．PLSA言語モデルの語彙を「話題語」「文型語」「汎用語」の３クラスに分割し,話題語PLSAモデルと文型語PLSAモデルを別々に学習・適応した後に３つのモデルを統合する．また新聞記事とCSJ間での品詞分類の出現パターン変化に基づいた,語彙分割基準の自動生成を提案する．評価実験では話題と文型の特徴が学習データで共起していないテキストについて，従来のPLSA言語モデルと比べ15.48%のperplexity削減が得られた．

キーワード 言語モデル，PLSA，話題適応，話者適応

Topic and style adaptation using vocabulary divided PLSA language model by criterion of information

Naoto KURIYAMA† Motoyuki SUZUKI† Akinori ITO† and Shozo MAKINO†

† Graduate School of Engineering, Tohoku University Aza-Aoba 6-6-05, Aramaki, Aoba-ku, Sendai-shi, Miyagi, 980-9579 Japan

E-mail: †{kriya,moto,aito,makimo}@makino.ecei.tohoku.ac.jp

Abstract PLSA (Probabilistic Latent Semantic Analysis) is one of promising language model adaptation methods. We propose a new way to combine PLSA and N-gram models by separating the vocabulary into three classes - 'topic'-related, 'style'-related and 'general'-related words. This method trains topic vocabulary PLSA model, style vocabulary PLSA model, and general vocabulary unigram model independently, and combines the three models. And we propose an automatic composing method of vocabulary divide criterion, using pattern of word-Class occurrence between newspaper and CSJ. The experimental result showed that the proposed method achieves 15.48% perplexity reduction than conventional PLSA model, about testset of which topic and style feature are not happen together in the training data.

Keyword Language model, PLSA, Topic adaptation, Speaker adaptation

1. はじめに

音声をテキストへ変換する書き起こし作業は，人手で行うと非常に労力を要し,自動化が求められている．

そこで大語彙連続音声認識を利用したディクテーション技術が研究され，ニュース音声などの音声認識に有利な音声データについては，既に高い精度が実現されている．しかし会議・講演などのディクテーションに必要な話し言葉認識については十分な性能が得られていない．話し言葉にはフィラーの出現や言い淀み言い直しがあり，文法的に正しく話すとは限らないことがその障害となっている．

そこで言語モデルを認識対象の話題や話し方の特徴に対して適応することが検討されている．言語モデルに話題を反映する手法には，適応対象に近い話題を持つテキストを学習データに重み付けして加える方法や，様々な話題ドメインに対応する言語モデルを予め用意し，目的の話題に最適な混合比を求めて混合する方法がある．後者の代表的な手法にPLSA[1]があり，[2][3]では高いPerplexity削減効果が報告されている．

2. 語彙分割を行ったPLSA言語モデル

2.1. PLSA言語モデル

PLSA(Probabilistic Latent Semantic Analysis, 潜在意味解析)とは，単語の出現頻度を基に，「話題」を，モデル化する手法である．文脈 h を反映した単語 w の出現確率 $P(w|h)$ が式（１）のように表される．$P(w|z)$ は単語 w に対する内部unigramモデル z が与える確率で，この内部モデルを潜在モデルと呼ぶ．潜在モデルはそれぞれが異なる話題を学習していて，目的の文脈 h に対して最適な混合比 $P(z|h)$ で混合することで，目標の話題に言語モデルを適応することができる．

この確率 $P(w|h)$ をunigram rescaling[2]によりtrigram

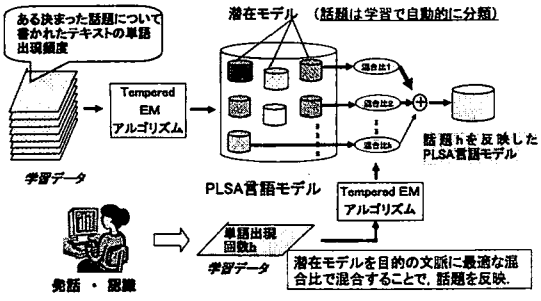

図１．PLSA 言語モデルの概要

$$P(w|h)=\sum_{z\in Z}P(z|h)P(w|z) \qquad (1)$$

$$P(w_i|h,w_{i-2},w_{i-1})\propto\frac{P(w_i|h)}{P(w_i)}P(w_i|w_{i-2},w_{i-1}) \qquad (2)$$

と併用して用いる．式を（２）に示す．

PLSA 言語モデルの学習は，特定の話題について書かれているテキストを大量に集め，それぞれの単語出現頻度を求め，それらに対し尤度最大化するような任意数の unigram を最尤推定することで行う．学習アルゴリズムには，Tempered EM アルゴリズム[1]を用いる．潜在モデルの混合比推定も同じアルゴリズムで行う．

2.2. 話題・文型の特徴分離

PLSA 言語モデルは元々は話題をモデル化するために提案されたものだが，実際に潜在モデルがどんな特徴を持っているのかどうか調べると，助動詞やフィラーの確率が特徴的に上昇した潜在モデルが発見できる．従って PLSA 言語モデルは話題だけでなく文型（話し方，文体）を学習する効果もあるといえる．そして話題と文型はほぼ独立な特徴と考えることができ，どのような話題・文型の組み合わせも可能である．

そこでこの二つの特徴を分離し，話題と話し方を別々に文脈適応することを考える．話題の PLSA と文型の PLSA を自由に組み合わせて使用できれば，あらゆる話題と文型が組み合わせ特徴を表現することができ，適応の柔軟性が大きく向上する．また学習データ中に現れなかった話題と文型の組み合わせに対しても適応が容易になり，学習データに対する依存性を小さくできると考えられる．

秋田らの研究[3]では，話題性をカバーするコーパスから話題 PLSA モデルを，話者性をカバーするコーパスから話者 PLSA モデルをそれぞれ学習し，仮認識結果によって文脈適応した後に重み付け混合して用いる方法を提案している．

これに対し本研究では，PLSA の語彙を話題の影響を受ける語彙「話題語」，文型の影響を受ける語彙「文型語」，そしてどちらの影響も受けない「汎用語」の３クラスに分割することで話題と文型を別々に適応する方法を提案する．

表１．人手で生成した語彙分割基準

話題語	名詞-一般，名詞-固有名詞，記号-アルファベット，名詞-副詞可能，名詞-サ変接続，名詞-形容動詞語幹，動詞-自立，chasen OOV
文型語	名詞-代名詞，名詞-数，名詞-非自立，接頭詞，名詞-特殊，名詞-接続詞的，名詞-ナイ形容詞語幹，名詞-接尾，動詞-非自立，動詞-接尾，形容詞，副詞，連体詞，接続詞，助詞-副助詞，助詞-終助詞，助動詞，感動詞，フィラー
汎用語	助詞-格助詞，助詞-接続助詞，助詞-係助詞，助詞-並立助詞，助詞-連体化，助詞-副詞化

2.3. 語彙の話題語・文型語・汎用語への分離

語彙を３クラスに分類するための分割基準は，形態素解析システム chasen[6]の品詞分類９０種類一つ一つについて，話題語・文型語・汎用語のどのクラスに属するか決定することで得る．

話題・文型の特徴分離は２クラスに分離するだけで実現できるが，ここではどちらの影響も受けない語彙を汎用語クラスとして分離する．理由として[2]では「出現頻度の特に高い単語は話題の影響を受けにくいため，PLSA が有効に働かない」と報告されている．格助詞などがこれに該当し，そのような高頻度の単語は話題にも文型にも影響されないと考えられる．さらに出現傾向が変化しない単語が文脈適応によって悪影響を受けることが，これまでの実験で分かっている[5]．そこで文脈の影響を受けない語彙「汎用語」は PLSA を用いない語彙クラスとして独立させることにした．

具体的な語彙の分割基準を表１に示す．これは人手で判断して決めた分割基準である．

2.4. 語彙分割 PLSA 言語モデルの概要

語彙分割を行った PLSA 言語モデルの与える確率は式（３）で表される．話題を表す語彙で構成されるモデル，文型を表す語彙だけで構成されるモデル，どちらの影響も受けない語彙で構成されるモデルの３つ（P_T, P_S, P_G）をそれぞれの語彙クラス出現確率で重み付け加算している．ただしある１つのクラスに属する単語は，他のクラスでは語彙に含まれないので，実質的には３つのモデルのどれかの確率が選択的に使用

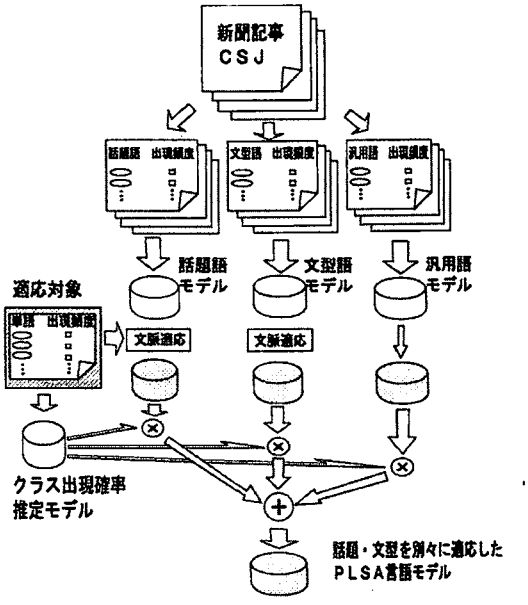

図２．語彙分割を行ったＰＬＳＡ言語モデルの概要

されることになる．語彙クラス出現確率は式（４），（５）で求められる．$N_h(w)$は適応ヒストリの単語数，$N_h(w,X)$はその中で語彙クラスXの単語の数を表す．

$$\begin{aligned} &P(w_i|h,w_{i-2},w_{i-1}) \\ &= P(T|h,w_{i-2},w_{i-1})P_T(w_i|h_T,w_{i-2},w_{i-1}) \\ &\quad + P(S|h,w_{i-2},w_{i-1})P_S(w_i|h_S,w_{i-2},w_{i-1}) \\ &\quad + P(G|h,w_{i-2},w_{i-1})P_G(w_i|w_{i-2},w_{i-1}) \end{aligned} \quad (3)$$

$$P(X|h,w_{i-2},w_{i-1}) = \frac{P(X|h)}{\sum_{w\in X}P(w)}\sum_{w_i\in X}P(w_i|w_{i-2},w_{i-1}) \quad (4)$$

$$P(X|h) = \frac{N_h(w,X)}{N_h(w)} \quad (5)$$

3. 情報量を基準とする語彙分割基準生成

3.1. はじめに

表１の語彙分割基準は人の直感的な判断によって話題・文型・汎用語を分割しているが，９０種類にわたる詳細な品詞分類について，３クラスのどの性質を持っているか的確に判断するのは難しく，主観に左右される部分が大きい．そこでより信頼性の高い分割基準を得るために，コーパス中での単語出現パターンの傾向から語彙分割基準を生成することを考える．

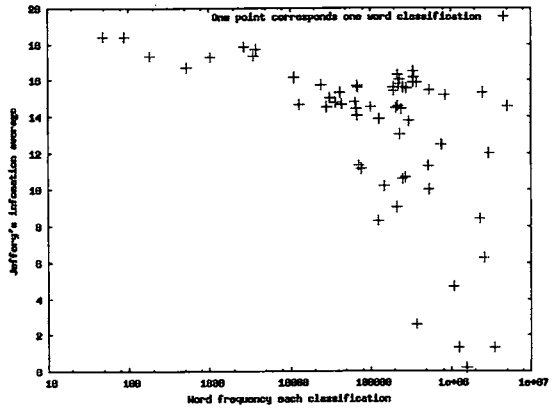

図３．Jeffery 情報量平均と形態素数の関係

3.2. Jeffery 情報量

Jeffery 情報量は確率分布 $p(k)$と $q(k)$の間の距離を測る尺度であり，式（６）で表される．$p(k)$と $q(k)$を入れ替えて求められる２つの Kullbuck-Leibler 情報量を加算したものと等しい．

これを用いて「ある学習記事(d)での品詞分類(C)単語の確率分布」と「学習データ全体で平均した品詞分類(C)単語の確率分布」の間の距離を測ることができる．式を（７）に示す．

$$J(p,q) = \sum_{k=1}^{n}(p(k)-q(k))\cdot\log\frac{p(k)}{q(k)} \quad (6)$$

$$J(d,D|C) = \sum_{w\in C}(P(w|C)-P(w|d,C))\log\frac{P(w|C)}{P(w|d,C)} \quad (7)$$

3.3. 汎用語の分離

この１つの記事と学習データ全体（つまり unigram）の間の Jeffery 情報量を，全学習データにわたって求め平均する．すると平均値が小さい品詞分類はどんなテキストでも出現パターンの変化が小さい品詞分類ということになり，つまり汎用語と考えられる．

９０種類の chasen 品詞分類それぞれについて式（７）の全記事平均を計算した結果を図３に示す．横軸は品詞分類ごとの形態素数で，プロット点一つが一つの品詞分類に対応する．また平均値の下から１０位までの品詞分類を表２に示す．

表２．Jeffery 情報量平均　下位１０品詞分類

品詞分類	Jeffery 情報量平均	形態素数	単語例
助詞-連体化	0.18	1.57M	の
助詞-格助詞-一般	1.31	3.53M	が，で，と・・・
助詞-係助詞	2.60	1.26M	こそ，しか，さえ・・・
助詞-格助詞-引用	2.69	0.37M	と
助詞-接続助詞	4.69	1.08M	けれど，のに，んで・・・
名詞-数	6.26	2.58M	１，一，百・・・
助詞-副詞化	8.27	0.13M	と，に
助動詞	8.42	2.33M	ござい，たい，でしょ・・・
助詞-並立助詞	9.05	0.21M	たり，とか，か・・・
名詞-非自立-一般	10.04	0.55M	つもり，もの，側・・・

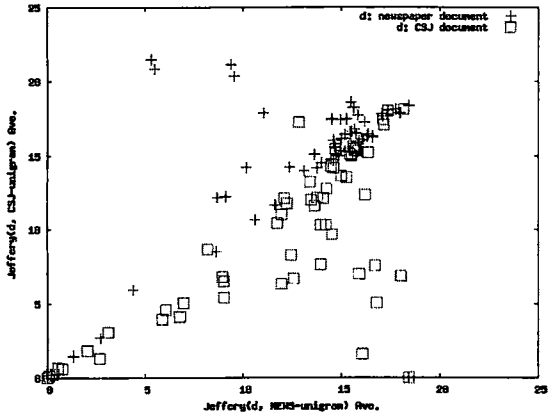

図４．新聞記事-CSJ に対する Jeffery 情報量の分布

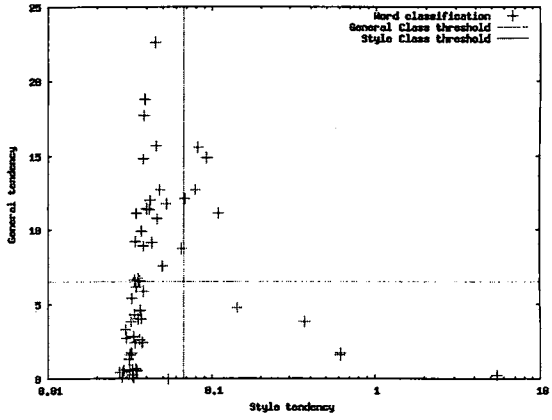

図５．汎用語らしさ・文型語らしさの分布

3.4. 文型語の分離

Jeffery 情報量を使い，文型語とそうでないものの識別を行う方法を考える．

文型語の例として助動詞を考える．新聞記事では文体が統制されているために，助動詞はどの新聞記事でもほぼ一定の出現パターンである．しかし CSJ では講演者の話し方，特に語尾の癖などによって出現パターンがばらつくことが考えられる．そこで文型の異なる２種類のコーパスクラス（新聞，CSJ）でそれぞれ unigram を作り，それぞれ新聞１記事に対する Jeffery 情報量の全記事平均，CSJ１記事に対する全記事平均を求めると，以下の４つの特徴量が得られる．

I.　新聞記事の unigram -新聞記事 d 間の情報量平均
II. 新聞記事の unigram - CSJ 記事 d 間の情報量平均
III.CSJ の unigram - 新聞記事 d 間の情報量平均
IV. CSJ の unigram - CSJ 記事 d 間の情報量平均

この４つの特徴量について，助動詞の例と対応付けると同じコーパスクラス間での値（I,IV）が小さく，違うコーパスクラス間（II,III）での値が大きくなる品詞分類が文型語であると考えられる．そこで I. ～IV. の４つの値を求め，傾向を調べた．

結果を図４．に示す．横軸は新聞の unigram に対する Jeffery 情報量，縦軸は CSJ の unigram に対する情報量で，そこに新聞記事 d の全記事平均を＋印（(x,y)=(I,III)），CSJ 記事 d 平均を□印（(x,y)=(II,IV)）でプロットしてある．一つの品詞分類について＋と□の２つの点がプロットしてある．

グラフ左上に分布した＋は新聞で一定に現れ，CSJ で記事の特徴によって出現パターンが変化していた品詞分類である．数詞・助動詞・動詞-接尾，動詞-非自立などがその傾向を示した．同じ品詞分類の CSJ 平均□は右下に大きく離れて分布する．これらが文型語であると考えられる．また＋と□が左下と右上に分かれる品詞分類も見られた．これは品詞分類自体の出現確率が新聞と CSJ で大きく異なるもので，フィラーなどがそうであった．

グラフ左下に分布したものはコーパスの種類によらず一定傾向で現れる，汎用語の性質が特に強い品詞分類と考えられる．格助詞・係助詞・接続助詞などがここに現れた．＋と□の間の距離は比較的小さい．

表３．Jeffery 情報量を基準に得られた語彙分割基準

話題語	名詞-副詞可能，名詞-サ変接続，名詞-接尾，形容詞-非自立，名詞-ナイ形容詞語幹，名詞-一般，名詞-形容動詞語幹，名詞-特殊-助動詞語幹，助詞-特殊，chasen OOV，名詞-固有名詞，形容詞-接尾，接頭詞，名詞-接続詞的，記号-アルファベット，名詞-引用文字列
文型語	助動詞，名詞-数，動詞-接尾，フィラー，動詞-非自立，代名詞，助詞-終助詞，連体詞，名詞-非自立-副詞可能，動詞-自立，接頭詞-名詞接続，助詞-格助詞-連語，接続詞，名詞-非自立-形容動詞語幹，名詞-接尾-特殊，副詞-助詞類接続，助詞-並立助詞，助詞-副助詞，副詞-一般，形容詞-自立，感動詞
汎用語	助詞-連体化，助詞-係助詞，助詞-格助詞-一般，助詞-格助詞-引用，助詞-接続助詞，助詞-副詞化，助詞-副助詞／並立助詞／終助詞，名詞-非自立-助動詞語幹，名詞-非自立-助動詞語幹，名詞-非自立-助動詞語幹，名詞-非自立-一般，名詞-接尾-形容動詞語幹

また右上に分布するものは比較対象によらず距離が大きいので，話題語であると考えられる．こちらも＋-□間距離が小さい．

3.5. 文型語らしさ・汎用語らしさ尺度

以上の考察から，図４上で同じ品詞分類についての＋・□点について，２点間の距離が文型語らしさ，２点間の中点から原点までの距離が汎用語らしさを反映していると考えられる．これに基づく文型語らしさ，汎用語らしさの計算式を式（８）（９）に示す．またその結果を横軸に汎用語らしさ・縦軸に文型語らしさでプロットしたものを図５に示す．

一部の品詞分類が *style(C)*, *general(C)* それぞれで極端に大きい値をとなり，どちらの値も小さい品詞分類が一箇所にまとまっている．

この図５を元に，文型語らしさ *style(C)* の閾値を 6.50 （名詞-副詞可能の直前），汎用語らしさ *general(C)* の閾値を 0.067 （助詞-並立助詞の直前）と設定して語彙分割の定義を決定した．図５中の縦線・横線はその閾値に対応する．ただし文型語かつ汎用語と判定された品詞分類は，汎用語とする．この結果に従い生成された分割基準を表３に示す．

$$style^2(C)=\left(\frac{1}{n_{NEWS}}\sum_{d\in NEWS}J(d\mid NEWS,C)-\frac{1}{n_{CSJ}}\sum_{d\in CSJ}J(d\mid NEWS,C)\right)^2 + \left(\frac{1}{n_{CSJ}}\sum_{d\in CSJ}J(d\mid CSJ,C)-\frac{1}{n_{NEWS}}\sum_{d\in NEWS}J(d\mid CSJ,C)\right)^2 \quad (8)$$

表４．実験条件

語彙数	30000+<UNK>
学習データ	毎日新聞２０００年度版から 63497 記事 （形態素数 26.9M） CSJ 学術講演・模擬講演・対話から 2580 記事 （形態素数 6.7M）
潜在モデル数	話題語モデル 100，文型語モデル 100，汎用語モデル 1
テストセット	学習データ類似テストセット： 毎日新聞２００１年度版から 50 記事 CSJ 学術講演・模擬講演から 47 記事
	話題・文型特徴混在テストセット： CSJ 模擬講演 S06 から 152 講演 「現在から過去数年の間に新聞・雑誌などで扱われたテーマ」

$$\frac{1}{general(C)}=\frac{1}{2}\left(\frac{1}{n_{NEWS}}\sum_{d\in NEWS}J(d\mid NEWS,C)+J(d\mid CSJ,C) + \frac{1}{n_{CSJ}}\sum_{d\in CSJ}J(d\mid CSJ,C)+J(d\mid NEWS,C)\right) \quad (9)$$

4. 評価実験

4.1. 実験条件

以上の方法で得られた情報量基準の語彙分割基準と表１の人手で生成した基準を用いて，語彙分割 PLSA 言語モデルの性能評価実験を行った．実験条件を表４に示す．言語モデルの学習データには広い話題をカバーし，書き言葉の文型をもつ新聞記事，そして話し言葉の文型・話者ごとの話し方の特徴を持つデータとして CSJ 講演書き下し文を用いる．Tempered EM アルゴリズムのアニーリングスケジュールは，話題語モデルで 140 回，文型語モデルで 184 回反復して学習を行い，学習初期は $\beta=1.0$ としてパラメータの収束が確認されるたびに β を段階的に減らし，学習終了時に $\beta=0.81$ となるようにする方法をとっている[4]．

テストセットには「学習データと類似したもの」と「学習データでは話題・文型が共起していないもの」の２種類を用いることにする．

潜在モデル混合数は話題・文型モデル共に 100 混合とした．これは同じ学習データ・テストセットで単一の PLSA モデルを用いた場合，性能飽和が見られた混合数である．この 100 混合の単一 PLSA と語彙分割 PLSA で perplexity を比較する．

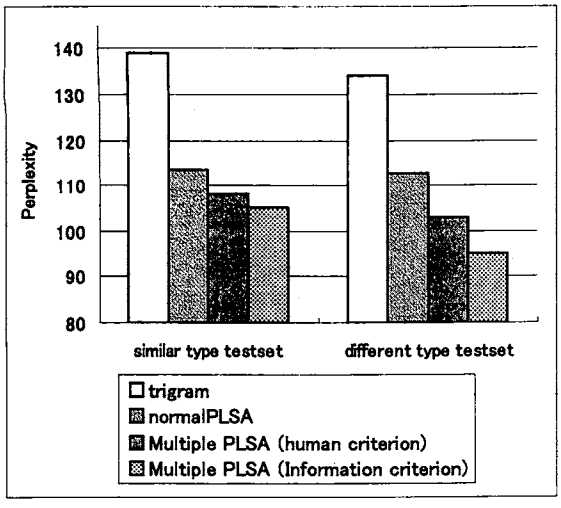

図６．perplexity による性能評価

4.2. 実験結果

実験結果を図６に示す．まず従来の単一ＰＬＳＡモデルによって学習データ類似テストセット(similar type testset)では trigram と比べ 18.52%，話題・文型の特徴が混在するテストセット(different type testset)では 16.11%の perplexity 改善が得られている．語彙分割ＰＬＳＡモデルを用いるとここから更に改善が得られ，Jeffery 情報量から生成した分割基準（表３）を用いた場合では，different type testset で従来の PLSA から更に 15.48%の perplexity 削減が得られた．学習データ類似テストセットでも 7.14%の改善が得られ，テストセットの種類によらず安定した性能を示している．

改善の理由としては，話題・文型の影響を受ける単語をそれぞれ分離したことにより，適応が行われやすくなったと考えられる．特に話題語モデルが話題に関連する単語だけに対して学習・適応されるため，適応の際のノイズが少なくなったことが効果的であると考えられる．

また汎用語に分類される高出現頻度単語を PLSA による適応の対象から除いたことにより，汎用語の確率が削られるのを防ぐことが有効であったと考えられる．

5. まとめ

PLSA 言語モデルの語彙を話題語，文型語，汎用語の３クラスに分割し，話題と文型を別々に適応することを試みた．Jeffery 情報量を用いて文型語らしさ・汎用語らしさを算出し，生成した語彙分割基準によって３つの言語モデルを学習した．評価実験では従来の単一 PLSA と比べて学習データ類似テストセットで 7.14%，話題・文型の特徴が学習データで共起していないテストセットでは 15.48%の Perplexity 削減が得られた．

文　献

[1] Thomas Hofmann,” Probabilistic Latent Semantic Analysis ” Uncertainity in Artificial Intelligence (1999)

[2] D.Glidea and T.Hofmann,” Topic-based langage models using EM ” EuroSpeech'99, pp.2167-2170 (1999)

[3] 秋田祐哉，河原達也，”話題と話者に関する PLSA に基づく言語モデル適応”，信学技報 NLC2003-61, SP2003-124，pp67-72 (2003)

[4] 栗山直人，鈴木基之，伊藤彰則，牧野正三，“PLSA 言語モデルの学習最適化と語彙分割に関する検討”，信学技報 SLP2006-060-8, pp37-42 (Feb.2006)

[5] 栗山直人，鈴木基之，伊藤彰則，牧野正三，“語彙を分割した PLSA 言語モデルによる話題・文型適応”，日本音響学会 2006 年秋季研究発表会講演論文集，2-2-2，pp55-56，(Sept.2006)

[6] 形態素解析システム茶筌 http://chasen.naist.jp/hiki/ChaSen/