PLSA に基づくトピック HMM を用いた言語モデル構築の検討*

◎佐古 淳　滝口 哲也　有木 康雄（神戸大・工）

1　はじめに

我々は、野球実況中継を対象とした音声認識に取り組んでいる。野球実況中継では、試合の状況に応じて発話の特徴が変化することから、状況依存の言語モデルを用いることにより認識精度改善が期待できる。本稿では、状況依存の言語モデル構築法として PLSA(Probabilistic Latent Semantic Analysis) を用いた手法について検討を行う。ただし、野球実況中継では、文書の境界が曖昧であり、状況や話題が頻繁に移り変わる。本稿では、状況や話題の移り変わりを表現するモデルとして、PLSA における潜在トピックを HMM によって駆動する手法について報告する。

2　PLSA による言語モデル適応

PLSA[1] はテキストモデリングの一手法であり、文書を潜在トピックの混合によって表現する。すなわち、文書 d における単語 w の発生確率を、潜在トピック z を用いて、

$$P(w|d) = \sum_z P(w|z)P(z|d) \qquad (1)$$

のように表現する。PLSA では、EM アルゴリズムによってパラメータを学習するため、bi-gram または tri-gram 確率を学習するには莫大な計算量を必要とする。そのため、近似的に N-gram 確率を推定する手法として、unigram スケーリングによる言語モデル適応が提案されている [2]。$P(w_i|w_{i-1}w_{i-2})$ の適応後の確率は以下の式によって求められる。

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

3　トピック HMM

本研究におけるシステムの概略を図 1 に示す。まず、野球実況中継コーパスに対して PLSA を行い、各潜在トピックにおける単語の unigram 確率 $P(w|z)$ 及び、発話毎の潜在トピック重み $P(z|d)$ を求める。1 発話は 1 つの句点で区切られた範囲を基本とした 10〜20 単語程度で構成されている。学習コーパス中の全発話数は約 8 千発話であった。

次に、各発話の潜在トピック重み $P(z|d)$ を特徴量として HMM を学習する。HMM は ergodic HMM とする。野球中継において、状況・話題はイニング毎に初期化されるものと考えられる。そこで、発話をイニング毎に分割し、1 イニング 1 サンプルとして HMM 学習に用いた。初期段階では、HMM のすべての状態に同じ分布を与えておく。学習により、似た状況・話題についての発話が HMM の状態に偏って現れる。学習により得られた HMM の各状態における分布は、PLSA の潜在トピック重みである（図 2）。状態 s の潜在トピック重みを $P(z|s)$ で表す。ただし、学習後、$\sum_z P(z|s) = 1$ となるように正規化する。HMM の状態遷移確率は、状況・話題の遷移確率と捉えることができる。

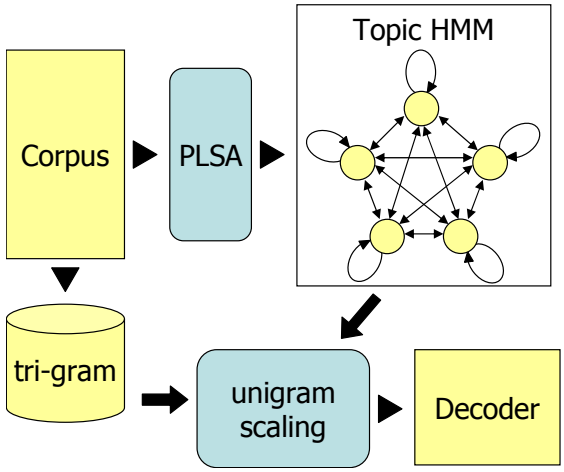

図 1: 提案手法の概略.

以上のようなトピック HMM を駆動しながら認識を行う。認識に用いる言語モデルは、単語列 $\mathbf{W}$、トピック HMM の状態系列 $\mathbf{S}$ として、以下のように定式化できる。

$$\begin{aligned} P(\mathbf{W}) &= \sum_{\mathbf{S}} P(\mathbf{W}, \mathbf{S}) \\ &= \sum_{\mathbf{S}} \prod_i P(s_i|s_1^{i-1}, w_1^{i-1}) P(w_i|w_1^{i-1}, s_1^i) \\ &\approx \sum_{\mathbf{S}} \prod_i P(s_i|s_{i-1}) P(w_i|w_{i-2}^{i-1}, s_i) \qquad (3) \end{aligned}$$

ここで、$P(w_i|w_{i-2}^{i-1}, s_i)$ は、式 2 より、

$$P(w_i|w_{i-1}^{i-2}, s_i) \propto \frac{P(w_i|s_i)}{P(w_i)} P(w_i|w_{i-1}^{i-2}, s_i) \qquad (4)$$

と表せる。

---

*Studies on language model construction using topic-hmm based on PLSA.
by Atsushi Sako, Tetsuya Takiguchi and Yasuo Ariki (Kobe Univ.)

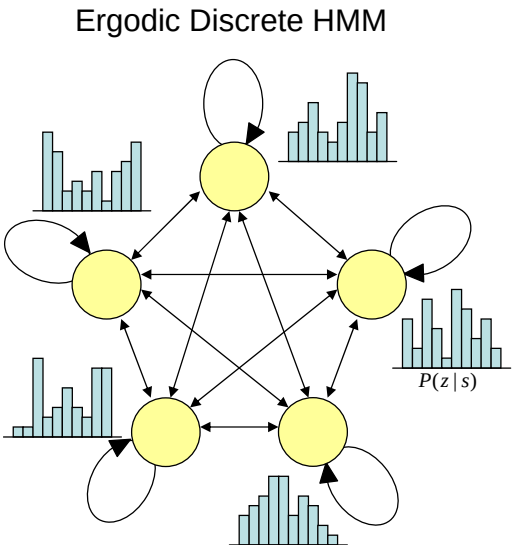

図 2: トピック *HMM*.

4 実験

以上のようなモデルを用いて認識実験を行う。テストセットは、野球実況中継音声である。音声はテレビではなくラジオのものを用いた。以下に実験条件を示す。

4.1 実験条件

音響モデルは，長母音化を考慮した音節 HMM（音節数 244）を用いた．ベースラインの音響モデルを，日本語話し言葉コーパス（CSJ:Corpus of Spontaneous Japanese）モニター版 [3] のうち男性話者 200 名の講演音声を用いて作成した．これに対し，MLLR+MAP によって実況中継音声（約 1 時間）の教師あり適応を行ったものを用いた．音響モデル適応に用いた音声は，テストセットと同一話者のものを用いた．1 状態あたりの混合分布数は 32，母音 (CV) は 5 状態 3 ループ，母音+子音 (CV) は 7 状態 5 ループである．サンプリング周波数は 16kHz，音響特徴量は 12 次元 MFCC と Δ MFCC，Δパワーの計 25 次元である．

ベースとなる言語モデルは、野球実況中継音声の書き起こしテキストから tri-gram モデルを作成した。書き起こしテキストは、話し言葉を書き起こしたものである。言語モデルの構築には Palmkit[4] を用いた。異なり単語数は約 2,300、コーパスサイズは約 8 万形態素であった。

デコーダーには Julius 3.5[5] を用いた．まず，通常のトライグラムモデルを用いて認識を行う．このとき，ベストパスだけではなく，十分に深い複数のパスを出力する．その後，提案手法の言語モデルを用いて N-best リスコアリングを行い，最終的な認識結果を得る．

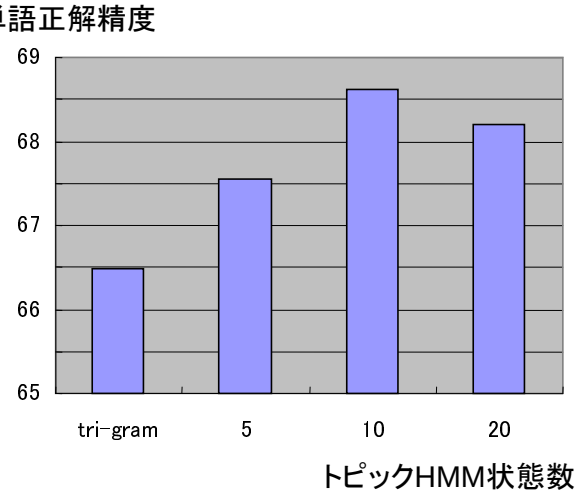

図 3: 実験結果.

以上のような認識条件・機構を用いて実験を行った。

4.2 実験結果

図 3 に結果を示す。tri-gram はトピック HMM を用いない通常の tri-gram による結果である。その他の数字はトピック HMM の状態数を表している。最も認識精度が高かったのはトピック HMM の状態数 10 のとき 68.6%であった。通常の tri-gram を用いた場合の 66.5%に対して 2.1%の精度向上が見られた。

5 おわりに

PLSA における潜在トピックを HMM によって駆動する言語モデルについて検討を行った。実験より、トピック HMM の状態数が 10 のとき最も良い認識精度が得られた。トピック HMM の状態数が多い方が微細にトピック毎の特徴をモデル化できる。ただし、状態数が多すぎるとトピックが学習データに依存しすぎてしまい、頑健性が低下するものと考えられる。今後、PLSA の潜在トピック数とトピック HMM の状態数との関係についても検討を行う必要がある。

参考文献

[1] T.Hofmann, “Probabilistic Latent Semantic Analysis”, In Proc. UAI, 1999.

[2] D.Gildea and T.Hofmann, “Topic-based Language Models using EM”, In Proc. Eurospeech, 1999.

[3] 古井貞熙，前川喜久雄，井佐原均.「『話し言葉工学』プロジェクトのこれまでの成果と展望」第 2 回話し言葉の科学と工学ワークショップ講演予稿集, pp.1-6, 2002 年 2 月.

[4] http://palmkit.sourceforge.net/

[5] 李晃伸，河原達也，鹿野清宏．“2 パス探索アルゴリズムにおける高速な単語事後確率に基づく信頼度算出法”, 情報処理学会研究報告, 2003-SLP-49-48, 2003-12.