2005－SLP－57（4）
2005／7／15

生成文書モデルを用いた文書読み上げ音声認識

中里 理恵 †　　貞光 九月 †　　冨山 良介 †　　山本 幹雄 †　　板橋 秀一 ††

† 筑波大学大学院 システム情報工学研究科
連絡先: myama@cs.tsukuba.ac.jp
†† 独立行政法人 産業総合研究所

概要　生成文書モデルは文書全体に対して文書らしさの確率を付与できる。我々は生成文書モデルによる確率を用いて、文書全体を読み上げた音声を全体で最適化し、認識するシステムを検討する。局所的言語モデルとの融合方法としては unigram rescaling 法と文書尤度法の 2 種類を提案する。生成文書モデルとして DM(Dirichlet Mixtures) モデルと LDA(Latent Dirichlet Allocation) モデルを用い、文書全体を読み上げた音声データに対する認識実験を行った。実験結果から、生成文書モデルによる精度の改善効果、パープレキシティと精度の関連、キャッシュモデルを加えた場合の効果について検討した。また、具体的な認識結果例文を用いて本手法の特徴を分析した。

Read Documents Recognition using Generative Text Models

Rie NAKAZATO †,　Kugatsu SADAMITSU †,　Ryosuke TOMIYAMA †,
Mikio YAMAMOTO †,　Shuich ITAHASHI††

†Graduate School of Systems & Information Engineering, University of Tsukuba
††National Institute of Advanced Industrial Science and Technology

Abstract　Generative text models give the probability to a document as a whole. We investigated a system to recognize all sentences in a read document maximizing a global score including the document probability. To integrate generative text models with a local language model (a trigram model), we used two methods: the unigram rescaling method and the document likelihood method. Experimental results are given for read speeches of all sentences in some documents, using the Dirichlet mixture (DM) model and the latent Dirichlet allocation (LDA) model as a generative text model. We report effectiveness of generative text models, the relationship between test-set perplexities and accuracy, additional effect of the use of cache models, and an analysis of examples of recognition results.

1　はじめに

従来の n-gram モデルは「文」の確率モデルである。これに対して文書の生成モデル (以下、生成文書モデルと呼ぶ) は「文書」全体の確率モデルである (上田 2004)。すなわち、生成文書モデルは文書全体に対して文書らしさの確率 (文書確率) を付与できる。本稿ではこの文書確率を用いることで、文単位の最適化ではなく、文書全体に対する最適化を行う音声認識手法を検討する。

n-gram モデルで文書全体の確率をモデル化しようとすると、全単語の短距離条件付確率の積でモデル化するしかない。これでは、低い出現確率の単語が多数出現するとどうしても確率が下がってしまう。しかし、実際は低い出現確率の単語であっても関連の高い単語が並ぶのならば文書確率は高いといえる。一方、高い出現確率の単語であっても関連のない単語が多数並んだ場合は文書としての一貫性はなく、文書確率は低くなるはずである。

例えば図 1 に示すように、高い出現確率を持つ単語「イラク」と、比較的確率が低い単語「投手」と「野球」の 3 つの単語をそれぞれ含む文を考える。文としての確率は上から中, 中, 高である。次に文書としての確率を考える。文書 A の「野球, 投手」、文書 B の「投手, イラク」を含む文書を n-gram モデルで評価すると、上から中, 高の確率となるであろう。しかし、現実には文書 B よりも文書 A の方がより文書らしいと言える。これをモデル化するのが生成文書モデルである。従来から研究が行われてきたキャッシュモデルやトリガーモデル等の長距離依存言語モデルも広い文脈を利用しているが、文書全体の確率をモデル化しているわけではない。明示的に文書の確率を導入したものが一連の生成文書モデルである。

以下、今回の実験で使用した 2 つの生成文書モデルを説明した後に、これを利用した音声認識手法を述べる。最後

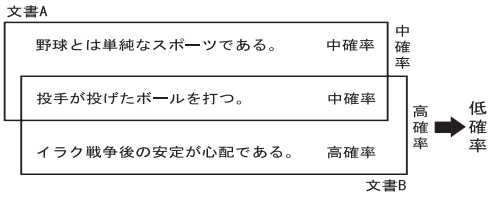

図 1: 生成文書モデルのイメージ

に音声認識実験の結果を示し、精度の改善効果、パープレキシティと精度の関連、キャッシュモデルを加えた場合の効果について検討する。また、具体的な認識結果の例文を挙げて本手法の特徴および今後の課題を考察する。

2 生成文書モデル

2.1 混合ディリクレ (DM) モデル

4 節の実験で生成文書モデルとして用いる混合ディリクレ (以下、DM) モデルと LDA モデルについて簡単に紹介する。まず DM モデルを説明するが、ここでは例を用いて直感的に説明する。分布の式やパラメータ推定法などの詳細は文献 (山本他 2003; 貞光他 2004) を参照。

DM モデルは、各文書中に出現する単語の割合 (すなわち出現確率) を確率ベクタとする、その上で定義される確率分布である。「割合」とは、すなわち各ベクタの値の和が 1 となるような空間であり、「単体」と呼ばれる。単体上の代表的な分布がディリクレ分布であり、DM モデルはこれを有限混合したモデルである。

図 2 は「携帯」「電話」「日本」の 3 単語について、毎日新聞 1999 年版の各記事中における割合を 3 次元の単体上にマップした図である (3 単語のいずれも出現しない記事は除いた)。各点が一つ一つの記事を表す。横軸が「携帯」の割合、縦軸が「電話」の割合である。「日本」の出現割合を示す軸がないが、他の 2 単語の出現割合より「日本」の出現割合は自ずと決まる。つまり、左下に寄るほど「日本」の出現割合が高いことを意味する。また、同じ点に複数の文書が存在する場合、その文書数が読み取れないため、出現割合には正規分布によるわずかなノイズを加え分散させている。この図より、頂点および辺に文書が密集しており、単独あるいは 2 単語で共起する場合が多いことが分かる (これは一般的な傾向である)。また、図中の実線で囲った部分にも文書が多いが、これは「携帯」と「電話」が同頻度で出現している文書が多いことを意味する。「携帯」と「電話」は「携帯電話」として出現する文書が多いため同数となりやすい。実線で囲った部分から「電話」側および「携帯」側に逸れた箇所にも注目されたい。図中破線で囲っている「電話」側の文書数は密になっている。これは「携帯電話」と「電話」と「日本」が共起した文書と捉えることができる。逆に「携帯」側の文書数は疎である。「携帯電話」と「携帯」と「日本」が共起することが少

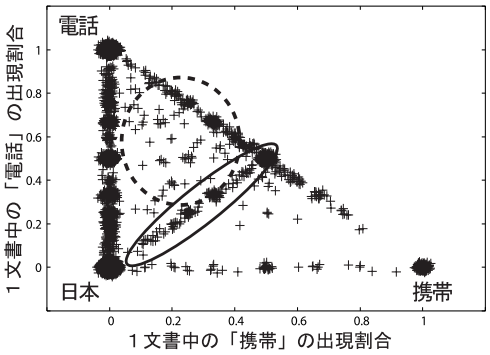

図 2: 「携帯」「電話」「日本」の出現割合の実測値

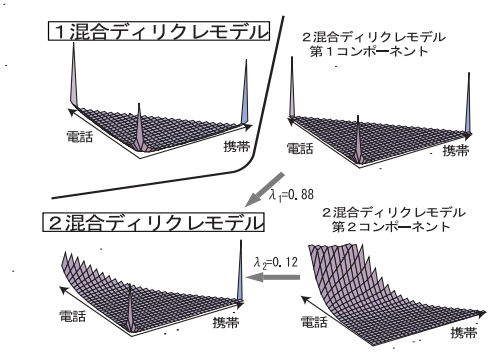

図 3: 「携帯」「電話」「日本」の DM モデルによるモデル化

ないのも、直感とあっている。以上のような文書の密度が高いところの文書の確率を高くし、密度の低いところの文書の確率を低くするような確率分布を形成できればよい。残念ながら、単一のディリクレ分布は非常に単純な分布形しかとれないため混合分布が必要となる。

図 2 のデータを 1 混合ディリクレモデルと 2 混合ディリクレモデルでモデル化した結果が図 3 である。現実のデータが頂点と辺に集中しているため、頂点と辺では無限大になるような密度分布が形成される (図では見易さのため途中で打ち切っている)。なお、無限大に発散しない辺を持つディリクレ分布は 2 混合ディリクレモデル中の第 2 コンポーネントの「日本」-「携帯」、「日本」-「電話」の辺のみである。1 混合ディリクレ分布は、各頂点において出現頻度が高いという傾向をうまく捉えているが、図 2 中の破線部の共起をモデル化できていない。2 混合ディリクレモデルの場合は、第 1 コンポーネントが 1 混合のモデルとほぼ同じで、第 2 コンポーネントが図 2 中の破線部の共起だけをうまくモデル化できている。これらを混合させた 2 混合ディリクレモデルは、1 混合ディリクレモデルよりも図 2 の精密なモデル化ができている。

2.2 LDA モデル

DM モデルは複数のトピックを考えるが、与えられた文書をその中のいずれか一つのトピックから生成されたとする。そのため、複数のトピックを同時に含むような文書を

モデル化できない。複数トピックから生成された文書をモデル化する手法が LDA(Blei et al. 2001) である。複数トピックが扱えることから、DM モデルよりも精密なモデル化が期待できるが、一つ一つのトピックを単純な多項分布でモデル化している点と、文書確率がトピックの混合割合に対する単一のディリクレ分布に依存しているため、言語モデルへの応用ではやや劣る (貞光他 2004)。

3 文書確率を用いた音声認識

3.1 文書確率の利用法の概要

次節の実験では文書確率を用いた 2 種類の音声認識手法を比較するが、どちらも基本的にはベースラインシステムが各発話に対して出力した 100-best の候補から rescoring 法を用いて最も良い候補を選択する方法を用いている。違いは文書確率の利用方法である。1 つ目は従来のトリガーなどの長距離依存言語モデルと同じように、認識が終わった部分を長い履歴とみなし、生成文書モデルを用いて後続の文の確率を推定する方法である。具体的には unigram rescaling 法を用いた (Gildea and Hofmann 1999)。2 つ目は、文書を前から順に認識するのではなく、文書全体で最適化を図る方法である。具体的には各文の 100-best の認識結果から、音響尤度と言語尤度の合計に文書尤度を加えた値が最大となるように認識文の組み合わせを選ぶ。

3.2 unigram rescaling 法

生成文書モデルを利用するオーソドックスな方法は、認識が終わった部分すべてを履歴として用いて、次の文の確率をベイズ学習的に推定する方法である。生成文書モデルの多くは unigram モデルをベースとして用いているため、unigram rescaling 法によって trigram に統合する (Gildea and Hofmann 1999)。認識が終わった部分を履歴データ h として生成文書モデルより unigram 確率 $P(w_i)$ の事後分布の期待値 $P(w_i|h)$ を求め、以下の式で trigram 確率 $P(w_i|w_{i-n+1}^{i-1})$ を変更する (山本他 2003)。

$$P(w_i|h, w_{i-n+1}^{i-1}) \propto \frac{P(w_i|h) \times P(w_i|w_{i-n+1}^{i-1})}{P(w_i)}$$

LDA の場合は文献 (三品他 2004b) を参照のこと。

またキャッシュモデルの効果を見るために、以下の式で unigram rescaling 法にキャッシュを組み込んだ。

$$P_{+cache}(w_i|h, w_{i-n+1}^{i-1}) = (1-\lambda_c)P(w_i|h, w_{i-n+1}^{i-1}) + \lambda_c\Big((1-\lambda_{cu})P_{bi}(w_i|w_{i-1}) + \lambda_{cu}P_{uni}(w_i)\Big)$$

$P_{bi}(w_i)$,$P_{uni}(w_i)$ がそれぞれ bigram と unigram のキャッシュである (Jelinek 1997)。λ_c はキャッシュモデル全体に対する重み、λ_{cu} はキャッシュモデル内における unigram の重みである。4 節の実験では $\lambda_c = 0.1$,$\lambda_{cu} = 0.99$ を用いた。

3.3 文書尤度法

新しい方法として、音響尤度と従来の言語尤度に加え、文書尤度を直接利用する文書尤度法を提案する。従来の方法が基本的に発話文を 1 文 1 文正確に認識していくことを目指していたのに対し、この方法では文書全体で最適な認識結果を得ることを目指す。

文書中の各発話の認識結果を s_i、N 文からなる文書全体の認識結果を $S = s_1, s_2, ..., s_N$ とし、音響尤度を $A(s_i)$, 言語尤度を $L(s_i)$, 文書尤度を $D(S)$, とし、以下のように rescoring スコア $R(S)$ を定義する。

$$R(S) = \sum_i A(s_i) + \lambda_1 \sum_i L(s_i) + \lambda_2 N \times D(S)$$

まず各発話ごとに N-best の候補を用意し、認識候補文の組み合わせに対して上記のスコアで最適化する。しかし、この方法では認識候補文の組み合わせが爆発的な数になるので、greedy サーチ法を今回は用いた。最初に各発話ごとに第 1 位の候補を組み合わせた文書を初期値とし、各発話ごとに尤度が改善する限り候補文を入れ替えていく。どの発話の候補文を入れ替えても尤度が改善しなくなった時点で終了する。

4 実験

4.1 実験条件

生成文書モデルとして DM と LDA モデル、記事読み上げ音声に対して Julius(河原他 2005; julius 2003) によって 100-best 候補を各文毎に出力し、rescoring 法による文書音声認識実験を行った。DM と LDA モデルの学習条件について次に示す。

- 学習データ：毎日新聞 1995〜1999 年 5 年分 505,433 記事 (156,518,030 単語)(毎日新聞社 1995〜1999)
- 形態素解析：chasen2.2.6(ipadic2.4.1),chawan2.09, suuji_syori1.0,PostProcess1.22 (鹿野 2001)
- 語彙：1995〜1999 年 出現頻度上位 20k・60k
- 学習テキスト：1995〜1999 年 5 年分
- 混合数：10・20・50・100・200

Julius は以下の条件である。

- デコーダ：Julius-rev3.4.2
- 音響モデル：CSRC 付属音響モデル (tri2000x16 性別依存)
- 言語モデル：bigram, 逆向き trigram, 語彙は 60k, discounting は Witten Bell カットオフなし (DM と LDA モデル作成で用いた同じ新聞記事 5 年分より CMU SLM toolkit(Clarkson and Rosenfeld 1997) で作成)

記事読み上げ音声は以下を用いた。

- テストデータ：新聞読み上げコーパス JNAS
- 話者:女性 5 名 (fp01〜fp05) 男性 5 名 (mp01〜mp05) の計 10 名

- 記事数：15 記事 (発話者男女各 2 名) 24-66 文/記事 計 1209 文 (19,878 単語), 単語カバー率は 20k で 96.3%, 60k で 99.2%

4.2 パープレキシティの評価

語彙が 60k の場合の各混合数におけるテストセットパープレキシティを図 4 に示す。テストデータは上記の JNAS の 15 記事である。LDA モデルより DM モデルの方がパープレキシティの値が小さく、言語モデルとしての性能が良いことが分かる。また両モデルにキャッシュモデルを加えるとさらに性能が良くなるが、DM と LDA モデルのパープレキシティの値がほぼ等しくなる。これは DM モデルがキャッシュモデルを含んだモデルであるためだと考えられるが、詳しい考察は今後の課題である。

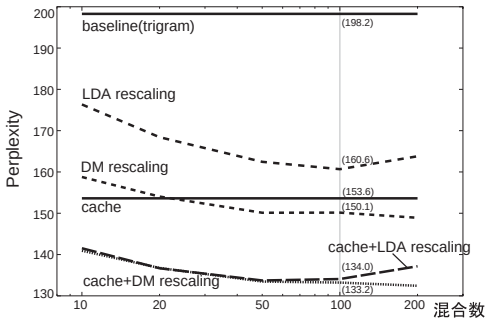

図 4: 言語モデルの性能評価

4.3 認識率による評価

ベースラインとの精度を比較した結果を図 5(語彙 20k)、図 6(語彙 60k) に示す。ここでのベースラインは言語尤度として単純な trigram のみを用いた場合のことである。図中には生成文書モデル 2 種 (DM,LDA) と利用方法 2 種 (unigram rescaling, 文書尤度法) の組合わせで 4 本の結果を示した。なお、文書尤度法のスコア式中の λ_2 で表される文書尤度の重みは 0.1 に統一した (中里他 2005)。oracles とは 100-best により得られた候補の中から正解を見て認識精度を最大とする候補を選んだ場合の結果である。20k の場合、文書尤度法の精度は向上しているが、rescalling ではあまり向上しなかった。60k は全手法がベースラインより上回る結果となった。これは 20k は未知語が多い分、誤認識が多いため、rescaling 法の場合特に文書の前半部分で誤認識単語が大きなノイズとなったと考えられる。一方、文書尤度法は全文書、全候補を同時に考慮するため、ノイズ的な効果が薄まったと解釈できる。

また 60k は 20k に比べてトピックを捉えるために有効な内容語が多く含まれるため、文書モデルの効果が出たと考えられる。例えば「ナリタブライアン」は競馬のトピックを端的に表す内容語であるが、20k では未知語となってしまう。言語モデルの性能が良い DM モデルの方が LDA

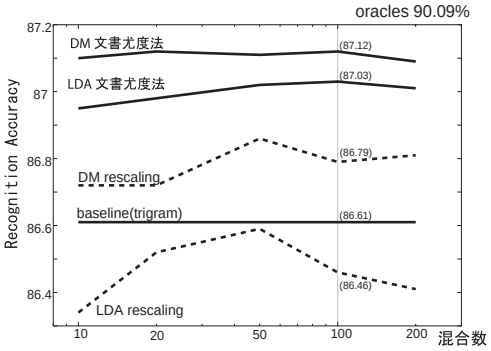

図 5: 精度の比較 (20k)

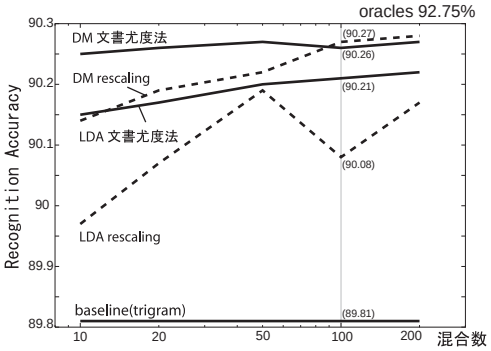

図 6: 精度の比較 (60k)

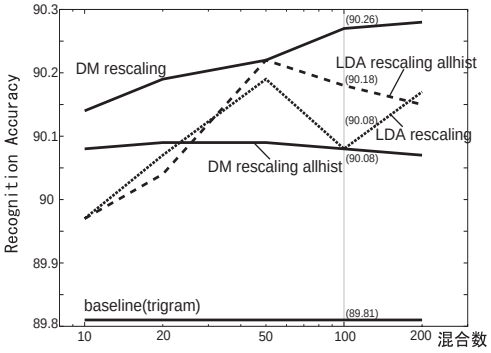

図 7: 精度の比較 (allhist、60k)

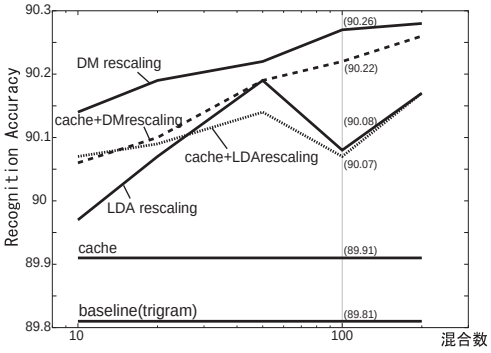

図 8: 精度の比較 (キャッシュモデル、60k)

モデルより精度でも高い値を示していることは予想通りである。

図 7 は unigram rescaling 法で用いるヒストリとして最初に与えられた N-best の第 1 位候補全てを用いて、ただ一度適応した場合の精度である (以下、図 7 では allhist と記す)。LDA モデルの場合は精度があまり変化しないが、DM モデルの場合は下がってしまった。この方法は誤認識結果を保存する方向に働くため、第 1 位候補の質に大きく影響を受けてしまう。また概観したところ、文書の前半部分では効果が現れている部分もあるので、さらに検討が必要である。

図 8 は rescaling 法にキャッシュモデルを加えた結果である。キャッシュモデルを加えるとパープレキシティによる言語モデルの評価では良い性能を示すが (図 4)、精度の向上は見られなかった。キャッシュだけでもパープレキシティでは DM モデルと同程度の性能であるが、精度は低かった。今回はキャッシュの重みの調整をパープレキシティで行ったが、キャッシュのようなヒューリスティックス的な方法は、このような調整に大きく性能が依存する場合がある点が問題である。

4.4 パープレキシティ減少率と WER 減少率の相関

ベースラインから unigram rescaling を使用した 60k、100 混合の場合における DM と LDA モデルのパープレキシティ減少率と WER(Word Error Rate) 減少率の相関を図 9 に示す。図中の点は各記事に対するプロットであり、直線は回帰直線である。生成文書モデルとして PLSA を用いた unigram rescaling 法では、パープレキシティが大幅に減少したものの WER にはほとんど効果が見られなかったという報告 (Gildea and Hofmann 1999) があるが、DM と LDA モデルにおいてはある程度の効果が見られた。パープレキシティの減少率が 25〜30%を越えた辺りからは確実に WER が減少しているので、今後も良い性能の言語モデルを開発することにより、音声認識の性能を更に向上することが可能であると信じる。

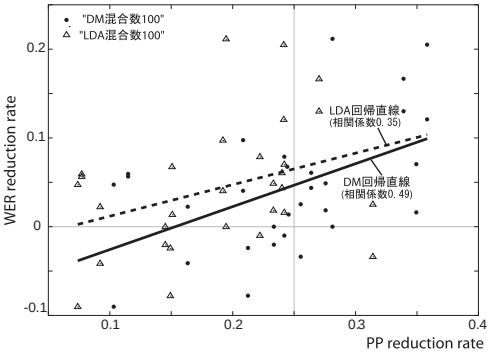

図 9: パープレキシティ減少率と WER 減少率の相関 (100 混合)

4.5 認識結果の詳細な考察

生成文書モデルがどのように認識性能に影響を及ぼしたかを DM モデルを用いた 60k、100 混合の場合について分析した。結果を表 1 に示す。

表 1 の「DM による改善 (悪化)」とは、ベースラインでは不正解 (正解) であったが、DM モデルによって正解 (不正解) となった箇所のことである。「oracles のみ正解」とは、oracles は正解であるが、ベースラインと DM モデルでは不正解となった文である (数が多いため、文を数えた)。

「仮名・漢字」は意味的には等しいが仮名を漢字に、あるいは漢字を仮名としてしまった箇所 (僕 ⟷ ぼく、既に ⟷ すでに、等) であり、「同音異義語」はその名の通り。「同音異義語に近い」は同音異義語ではないが比較的音が近い箇所、「その他」は助詞の違いなどを示す。

まず「仮名・漢字」であるが、DM モデルによって改善・悪化した箇所がほぼ同数であるため、ランダムに起こり得た結果であると推測できる。「oracles のみ正解」が 50 文あることから、今後 100-best の rescoring によって改善できるうちの約 25%は意味がないことが分かる。

また「同音異義語」は DM モデルによって最も改善が期待できると考えていたが、全体の約 12%に留まった。以下に例を挙げる。

- DM モデルによって改善した例
  - id :nf942-034
  - 正解文:ただ**棋士**によってアマチュアへの接し方は指導将棋だけでなく自宅教室を開く ...
  - tri :ただ**岸**によってアマチュアへの接し方は指導将棋だけでなく自宅教室を開く ...
  - DM :ただ**棋士**によってアマチュアへの接し方は指導将棋だけでなく自宅教室を開く ...
- DM モデルによって悪化した例
  - id :nf952-071
  - 正解文:**週末**のホームステイで驚いたのは ...
  - tri :**週末**のホームステイで驚いたのは ...
  - DM :**終末**のホームステイで驚いたのは ...
- oracles のみ正解した例
  - id :nm911-030
  - 正解文:レコードが出ると JRA から関係者にレコード**賞**として賞金とメダルが贈られる
  - tri :... 関係者にレコード**商**として ...
  - DM :... 関係者にレコード**商**として ...
  - ora:... 関係者にレコード**賞**として ...

次に「同音異義語に近い」であるが、DM モデルによって改善することができた箇所の 6 割を占めている。この傾向は同音異義語を訂正するスペルチェッカ (三品他 2004a) とは別な、音声認識ならではの特徴と言える。文書のト

表 1: DM モデルによる認識性能の分析結果

	仮名・漢字	同音異義語	同音異義語に近い	その他	計
DM による改善	15 14.7%	12 11.8%	66 64.7%	9 8.8%	102(箇所) 100%
DM による悪化	14 31.1%	8 17.8%	17 37.8%	6 13.3%	45 (箇所) 100%
oracles のみ正解	50 24.6%	30 14.8%	56 27.6%	67 33.0%	203(文) 100%

ピックを的確に捉えることにより大きく改善することができたと考えられる。以下に例を挙げる。

- DM モデルによって改善した例
  - id :nf911-017
  - 正解文:なぜ**タイム**面で逆転現象が起こるのか
  - tri :なぜ**財務**面で逆転現象が起こるのか
  - DM :なぜ**タイム**面で逆転現象が起こるのか
- DM モデルによって悪化した例
  - id :nf913-076
  - 正解文:第二戦では**速球**を決め球に二三振を ...
  - tri :第二戦では**速球**を決め球に二三振を ...
  - DM :第二戦では**昇給**を決め球に二三振を ...
- oracles のみ正解した例
  - id :nf911-023
  - 正解文:長距離レースでその傾向が強くこういう場合**ペース**は遅くなる
  - tri :... こういう場合**レース**は遅くなる
  - DM :... こういう場合**レース**は遅くなる
  - ora :... こういう場合**ペース**は遅くなる

「その他」では「は ⟷ が」、「は ⟷ も」などによる助詞による違いが多く見られた。

今後の課題として、DM モデルによって悪化した及び oracles のみが正解した「同音異義語」と「同音異義語に近い」箇所は、本手法により改善の余地があると考える。これを克服すれば 1%程度の精度向上が見込める。

5 おわりに

生成文書モデルを利用した音声認識手法として、unigram rescaling 法、文書尤度法を検討した。生成文書モデルとして DM と LDA モデルを用いた実験により、生成文書モデルを利用する方法は精度を改善できることを示した。

参考文献

David M. Blei, Andrew Y. Ng and Michael I. Jordan (2001). "Latent Dirichlet Allocation." In *Neural Information Processing Systems*, Vol. 14.

P.R.Clarkson and R.Rosenfeld (1997). "Statistical Language Modeling Using the CMU-Cambridge Toolkit."

D.Gildea and T.Hofmann (1999). "Topic-based Language Models Using EM." In *In Proc. of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*

F. Jelinek (1997). "Statistical Methods for Speech Recognition."

Julius (2003). http://julius.sourceforge.jp/.

上田修功 (2004). "テキストモデルの最前線 (1),(2)." 45 (3), pp. 282–289.

河原達也 李伸晃 (2005). "連続音声認識ソフトウェア Julius." 人工知能学会誌, 20 (1), 41–49.

貞光九月, 待鳥祐介, 山本幹雄 (2004). "混合ディリクレ分布パラメータの階層ベイズモデルを用いたスムージング法." 情報処理学会研究報告, 2004-SLP-53(1).

鹿野清広, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄 (2001). "音声認識システム." オーム社

中里理恵, 貞光九月, 山本幹雄, 板橋秀一 (2005). "文書確率を用いた文書読み上げ音声認識." 日本音響学会春季講演論文集, pp. 47–48.

毎日新聞社 (1995〜1999). "CD-毎日新聞'95〜'99 データ集."

三品拓也, 貞光九月, 山本幹雄 (2004a). "確率的 LSA を用いた日本語同音異義語誤りの検出・訂正." 情報処理学会論文誌, 45 (9), pp. 2168–2176.

三品拓也, 山本幹雄 (2004b). "確率的 LSA に基づく *n*gram モデルの変分ベイズ学習を利用した文脈適応化." 電子情報通信学会誌 D-II, J87 (7), pp. 1409–1417.

山本幹雄, 貞光九月, 三品拓也 (2003). "混合ディリクレ分布を用いた文脈のモデル化と言語モデルへの応用." 情報処理学会研究報告, 2003-SLP-48(5).