Vol.2010-SLP-84 No.20
2010/12/21

確率的言語モデルに基づく音声ドキュメント検索のための Web を利用したモデル拡張の検討

増村 亮[†1] 咸 聖俊[†1] 伊藤彰則[†1]

音声ドキュメントのアドホック検索に対する需要が近年増加している．音声認識を利用して音声ドキュメントをテキストへ変換することで既存のテキスト検索の技術が利用可能であるが，音声認識誤りにより，検索性能が大きく劣化することが知られている．この問題を解決するために，以前我々は認識誤りによって欠落してしまった単語を対象音声に関連する Web 文書を利用して補間する方法を提案した．本稿では新たに情報検索モデルとして近年注目されている確率的言語モデルに焦点を当てる．Web を利用した補間のアイデアを確率的言語モデルを利用する枠組みに組み込むために，我々は Web 検索ヒット数を用いた新たなスムージング方法，および Web 関連文書を用いた混合モデル化による文書モデル拡張手法を提案する．

Modeling Expansion using Web for Spoken Document Retrieval based on Probabilistic Language Model

Ryo Masumura ,[†1] Seongjun Hahm [†1]
and Akinori Ito[†1]

In recent years, there has been more and more demands for ad hoc retrieval of spoken documents. We can use existing text retrieval method by trascribing the spoken document into text using a Large Vocabulary Continuous Speech Recognizer (LVCSR). However, it is well known that the retrieval performance deteriorates severely by recognition errors. To solve this problem, we previously proposed a method which interpolate lacked words using relevant Web documents to the target spoken document. In this paper, we newly focus on probabilistic language model which is attracted attention as a information retrieval model. To introduce Web-based interpolation idea into language modeling approach, we propose new smoothing method using Web hit counts and mixture modeling method using relevant Web documents.

1. はじめに

近年のインターネットの発展とともに，我々はテキスト以外のマルチメディアコンテンツを利用する機会が増えている．しかし，動画など音声を含む多くのコンテンツ (音声ドキュメント) は，タイトル以外のメタデータが付与されていることが少なく，アドホックな情報検索は難しい問題となっていた．これに対して，近年大きな発展を見せる大語彙連続音声認識技術が注目を集めている．音声認識を用いて音声ドキュメントの自動ディクテーションを行うことで，通常の情報検索と同様のアプローチが利用可能となっている．

しかし大語彙連続音声認識技術を利用するにあたって，いくつかの問題が存在する．まずは認識誤りの問題である．音声認識結果には多くの認識誤りが存在する．そのため，音声にとって重要な単語の欠落が起こり得る．また，大語彙連続音声認識における大きな問題として未知語の問題がある．今日の音声認識器の登録語彙の大きさは有限であり，語彙外の単語，つまり未知語は必ず認識できないという問題がある．したがって，正確なディクテーションが行われた場合と比較して検索性能が大きく劣化してしまうことが知られている．

このような問題に対して，我々は音声認識で欠落してしまった単語を補間するアプローチに着目し，豊富な言語資源を有する World Wide Web から関連文書を集めることで欠落単語を補間する方法を提案している[1)]．以前我々は，古典的なベクトル空間モデルを検索モデルに用いて検討してきたが，本稿では，近年検索モデルとして注目される確率的言語モデルに新たに焦点を当てる[2)]．確率的言語モデルを用いるアプローチは，ベクトル空間モデルで用いられていた発見的手法を用いず，数理的に説明可能な枠組みを指向している．音声ドキュメント検索の分野でも近年導入が試みられていて，ラティスを利用する枠組み[3)]，翻訳モデルを利用する枠組み[4)]，トピックモデルに基づく枠組み[5)]，文書クラスタリングに基づく枠組み[6)] などの検討が行われている．そこで本稿では，Web を利用して欠落単語を補間するアイデアを確率的言語モデルによる検索モデルの枠組みに組み込む方法を検討する．

本稿は次のような構成とする．まず 2 章で確率的言語モデルによる情報検索について詳細を述べ，音声認識結果を利用する場合の問題について述べる．3 章では以上の問題を解決するために Web を利用したモデル拡張を提案する．そして 4 章で，テストコレクションによる検索実験を行い，本稿で提案した手法の有効性について述べる．

†1 東北大学大学院工学研究科
Graduate School of Engineering, Tohoku University

2. 確率的言語モデルによる情報検索

2.1 クエリ尤度モデル

確率的言語モデルによる情報検索は文書 D がクエリ Q に適合する確率 $P(D|Q)$ を求めることで実現できる．$P(D|Q)$ はベイズの定理から (1) 式のように表すことができる．

$$P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)} \propto P(Q|D)P(D) \propto P(Q|D) \tag{1}$$

$P(Q)$ は文書に依存しない項であり，また $P(D)$ は基本的に一様と見なせる．$P(Q|D)$ はクエリ尤度モデル (query likelihood model) と呼ばれ，あるクエリ Q はある文書 D に関する言語モデルからのランダムサンプリングによって生成したと想定する[2]．

文書 D に関する言語モデル (文書モデル) には，一般的に多項分布が用いられる[7]．多項分布に基づく文書モデルはユニグラム言語モデルとも呼ばれ，単語が独立に生起すると仮定する．文書 D の文書モデル θ_D からクエリ Q が生成される尤度，すなわちクエリ尤度 $P(Q|\theta_D)$ は (2) 式のように表される．

$$P(Q|\theta_D) = \frac{\Big(\sum_{w_i \in V} c(w_i, Q)\Big)!}{\prod_{w_i \in V} c(w_i, Q)!} \prod_{w_i \in V} P(w_i|\theta_D)^{c(w_i,Q)} \tag{2}$$

ここでは，クエリ Q に存在する語彙 $w_i \in V = \{w_1, w_2, ..., w_{|V|}\}$ が複数回出現することを考慮している．なお，$c(w_i, Q)$ は，クエリ Q において語彙 w_i が出現した頻度を表す．検索に依存しない部分を省略することで，(3) 式のように表現する．

$$P(Q|\theta_D) \propto \prod_{w_i \in V} P(w_i|\theta_D)^{c(w_i,Q)} \tag{3}$$

この $P(Q|\theta_D)$ に基づく文書ランキングによって，確率的言語モデルによる情報検索は実現される．よって，各文書モデルのパラメータ $P(w_i|\theta_D)$ の推定が重要となる．

2.2 ディリクレスムージングによる文書モデルのパラメータ推定

文書モデルのパラメータ推定において，最尤推定では零確率問題が発生してしまうため，一般的に事前分布にディリクレ分布を用いて MAP 推定が行われる．この方法をディリクレスムージングと呼ぶ．ディリクレ分布 $\mathrm{Dir}(\boldsymbol{\alpha}) = \mathrm{Dir}(\alpha_1, ..., \alpha_W)$ を事前分布として MAP 推定を行うと $P(w_i|\theta_D)$ は (4) 式のように推定できる．

$$P(w_i|\theta_D; \boldsymbol{\alpha})^{MAP} = \frac{c(w_i, D) + \alpha_i}{\sum_{k=1}^{W} c(w_k, D) + \sum_{k=1}^{W} \alpha_k} = \frac{c(w_i, D) + \alpha_i}{|D| + \sum_{k=1}^{W} \alpha_k} \tag{4}$$

ここでディリクレ分布のパラメータ α_i を，基底となる確率分布の係数倍として表現する．一般的に基底となる確率分布には，検索対象の文書コレクション C のモデル，つまりコレクションモデルのパラメータ $P(w_i|\theta_C)$ を用いて，(5) 式のようにおく．

$$\alpha_i = \mu P(w_i|\theta_C) \tag{5}$$

この式を (4) 式に代入して，(6) 式のように変形する．

$$P(w_i|\theta_D; \mu)^{MAP} = \frac{c(w_i, D) + \mu P(w_i|\theta_C)}{|D| + \mu} \tag{6}$$

このスムージングパラメータ μ の推定方法として，leave-one-out 尤度 (loo 尤度) を用いる方法がある[7]．つまり観測文書のそれぞれの単語の確率を，それ以外の単語を所与とする条件付き確率によって表し，それらの確率の積によって文書全体の出現確率を表す．文書コレクションに対する loo 対数尤度 $l_{-1}(C)$ は (7) 式となる．

$$l_{-1}(C) = \sum_{j=1}^{J} \sum_{i=1}^{W} c(w_i, D_j) \log\Big(\frac{c(w_i, D_j) - 1 + \mu P(w_i|\theta_C)}{|D_j| - 1 + \mu}\Big) \tag{7}$$

この $l_{-1}(C)$ を最大化するような μ を推定値とする．つまり (8) 式の最大化問題を解く．

$$\hat{\mu} = \arg\max_{\mu} l_{-1}(C) \tag{8}$$

これはニュートン法によって解くことが可能である．

2.3 音声認識を利用する場合の問題

音声ドキュメント検索では，音声認識結果から音声ドキュメントの文書モデルを推定する．音声認識技術を用いる際，前述の通り認識誤りおよびデコーダの未知語が大きな問題となる．確率的言語モデルを検索モデルとして用いる際，以下の二点に改善を求める必要がある．

第一に対処しなければならないのは零確率問題である．通常の情報検索では，コレクションモデルのパラメータの最尤推定値 $P(w_i|\theta_C)^{ML}$ を利用してスムージングを行うことで，全ての文書モデルに実際に文書コレクションに存在する全ての単語に対してなんらかの生成確率を与えることができた．しかし音声認識結果を用いる場合では，文書コレクションに対

しても単語の欠落が発生する．よって，コレクションモデルのパラメータに対してもスムージングを行うことが望ましい．

第二に，対象音声ドキュメントの話題を表す重要な単語であるにも関わらず，音声認識により欠落してしまった場合に何らかの対処が必要である．上述のようなスムージングにより零確率問題を回避できたとしても，重要単語に正しい生成確率が付与されていない場合，検索性能の大きな低下につながる．したがって，認識対象に出現しそうな単語を予測し，その単語の生成確率を補間することが望ましい．

3. Web を利用した確率的情報検索モデルの拡張

3.1 コレクションモデルに対するスムージング

文書モデルにディリクレスムージングを行う場合，通常 (6) 式のコレクションモデルのパラメータ $P(w_i|\theta_C)$ には最尤推定値 $P(w_i|\theta_C)^{ML}$ を用いる．しかし音声認識結果を扱う場合では，コレクションモデルに対してもスムージングを行うべきである．そこで，コレクションモデルにも事前分布としてディリクレ分布を導入する．

このディリクレ分布は，文書コレクションの全ての単語はもちろんのこと，あらゆる単語を生成するような多項分布に対する事前分布であることが望ましい．そこでディリクレ分布のパラメータに対する基底分布を導入する．本稿ではこの基底分布を表現するモデルを，グローバルモデル θ_G と定義する．我々はこのグローバルモデルのパラメータ $P(w_i|\theta_G)$ を Web を利用して求める方法を提案する．

Web 上の膨大な言語資源から単語頻度を計算することで，あらゆる単語に対してパラメータを与えるようなグローバルモデルを推定できる．しかし，膨大な言語資源の単語頻度に基づいてグローバルモデルを推定するのはあまりに非現実的であるため，既存の Web 検索エンジンから得られる単語の Web 検索ヒット数を代用する．Web 検索ヒット数で代用することは，文書ごとのブーリアン頻度を用いて Web 全体から推定することと等価と言える．単語 w の Web 検索ヒット数を $hit(w)$ とした時，十分大きい語彙空間 W に対してそれぞれの単語の Web 検索ヒット数を得ることで，(9) 式のようなグローバルモデルのパラメータを得る．

$$P(w_i|\theta_G)^{ML} = \frac{hit(w_i)}{\sum_{w \in W} hit(w)} \tag{9}$$

最尤推定したグローバルモデルのパラメータ $P(w_i|\theta_G)^{ML}$ を事前分布のパラメータの基底におき，係数 η を導入することで，コレクションモデルのパラメータを MAP 推定する．パラメータ $P(w_i|\theta_C;\eta)^{MAP}$ は (10) 式のように推定できる．

$$P(w_i|\theta_C;\eta)^{MAP} = \frac{c(w_i,C) + \eta P(w_i|\theta_G)}{|C| + \eta} \tag{10}$$

$$c(w_i,C) = \sum_{D \in C} c(w_i,D) \tag{11}$$

$$|C| = \sum_{i=1}^{W} c(w_i,C) \tag{12}$$

係数 η は前述の loo 尤度を用いることでニュートン法から推定可能である．このコレクションモデルのパラメータの MAP 推定値 $P(w_i|\theta_C;\eta)^{MAP}$ を (6) 式の $P(w_i|\theta_C)$ に用いる．これによってあらゆる単語に対して生成確率を付与でき，認識誤りや未知語などの問題により発生する零確率問題への対処が期待できる．

3.2 Web 関連文書を用いた重要単語の補間

次に，音声認識により欠落してしまった対象音声の重要単語に対する対処として，我々は対象音声に関連する文書を利用して，欠落した重要単語の補間を行う．そのためのアプローチとして，まず関連文書を収集して関連文書の多項分布モデルを作成する．

この関連文書の収集源として我々は Web を利用する．Web から収集する理由は，あらゆる話題に対しても関連文書を収集することが期待できるからである．しかし，実際に Web 上の全ての言語資源から関連文書を探すことは非常に困難であり，また Web データに逐次アクセスすること自体も効率的ではない．そこで本稿では，あらかじめ Web 空間から部分的にデータをサンプリングすることで構築した事前ダウンロードデータ群を利用する[8]．この事前ダウンロードデータ群は，様々な単語をキーワードとして Web 検索を行うことで収集した文書群であり，キーワードに用いた単語とダウンロードデータを対応付けて保持することで，Web キーワード検索と等価な方法で容易にデータを利用できる．この事前ダウンロードデータ群から以下の流れで関連文書の多項分布モデルを作成する．

step1: 最初に，事前ダウンロードデータ群から対象音声に関連がありそうな Web 文書を抽出する．具体的には，認識結果に出現したそれぞれの単語に対応付けられた Web 文書のみを抽出する．この抽出した文書群を，関連文書候補群 S とする．

step2: 次に，抽出した関連文書候補群 S から認識対象に関連する文書を選択する．そのために，KL ダイバージェンスを利用して認識結果との関連性を求める．それぞれの関連

文書候補に対して多項分布モデルを推定して，認識対象の文書モデルとの分布間距離を求め，ランキング化する．t 番目の関連文書候補の多項分布モデルを θ_{S_t} とすると，文書モデル θ_D を基準の分布とするので，KL ダイバージェンスは (13) 式に比例する．

$$-KL(\theta_D||\theta_{S_t}) \propto \sum_{w_i \in D} P(w_i|\theta_D) \log P(w_i|\theta_{S_t}) \tag{13}$$

この値が大きいほど，認識対象に関連する文書であると考えられる．

step3: 最後に，KL ダイバージェンスが高い順に N 文書選び，それらを関連文書群 $R = \{r_1, ..., r_N\}$ とする．この R を用いて関連文書モデルを推定する．関連文書モデルのパラメータに対しても文書モデルと同じディリクレ事前分布を導入しスムージングを行うことで，(14) 式のように求める．

$$P(w_i|\theta_R;\mu)^{MAP} = \frac{c(w_i,R) + \mu P(w_i|\theta_C)}{|R| + \mu} \tag{14}$$

$$c(w_i,R) = \sum_{j=1}^{N} c(w_i,r_j) \tag{15}$$

$$|R| = \sum_{i=1}^{W} c(w_i,R) \tag{16}$$

以上の流れで関連文書モデルを作成した後，欠落した重要単語の生成確率は関連文書モデルによって補うというアイデアのもと，文書モデル θ_D と関連文書モデル θ_R の線形補間による混合モデル化を行う．文書モデルのパラメータ $P(w_i|\theta_D)$ と関連文書モデルのパラメータ $P(w_i|\theta_R)$ 補間係数 $\lambda(\geq 0)$ を導入して線形補間する[9]．線形補間による混合モデル $P(w_i|\hat{\theta}_D)$ は (17) 式のように定式化する．

$$P(w_i|\hat{\theta}_D;\mu,\lambda) = \lambda P(w_i|\theta_D;\mu)^{MAP} + (1-\lambda) P(w_i|\theta_R;\mu)^{MAP} \tag{17}$$

これにより，音声認識結果から推定した文書モデルに，欠落してしまった重要単語を補間するようなモデル化を行うことができる．

4. 検索実験

4.1 テストコレクション

我々は検索評価実験のためのテストデータとして,「日本語話し言葉コーパス」を検索対象とした「CSJ テストコレクション」を用いた[10].「CSJ テストコレクション」は，CSJ の学会講演音声，模擬講演音声 (全 2702 講演) を検索するための 39 の検索クエリと，それぞれの音声に「適合」及び「部分適合」する講演音声 ID に関する情報が含まれている．テストコレクションには全 2702 講演に対する書き起こしも付属されているが，本稿では新たに音声認識器を構成して書き起こしを行った．その理由として，我々は欠落単語の補間を行うため，形態素解析器は，音声認識器の言語モデルの学習データを形態素解析したものと同一であることが望ましいからである．言語モデルには，テストコレクションを内包する CSJ3302 講演のデータからクローズドな単語 N-gram を作成した．この言語モデルを用いて，テストコレクション 2702 講演を自動で書き起こした．平均単語認識精度は 75.12%，平均未知語率は 0.23%となった．実験条件の詳細は表 1 に示す．

表 1 音声認識の実験条件
Table 1 Experimental cobdition of speech recognition

Decoder	Julius 4.1.2
Acoustic model	PTM triphone, comes with CSJ
Language mode	Forward bigram, backward trigram
LM smoothing	Witten-Bell
Training corpus	3302 lectures from CSJ (8209043 words)
# unigram	50000
Morphemic analyzer	ChaSen(ipadic+unidic)

なお評価には，それぞれの検索クエリに対して適合すべき正解の音声として「適合」と「部分適合」両方を使用し，検索対象は 1 つの講演を 1 ドキュメントとする．そして各検索クエリに対して，11 点平均補間適合率 (11pt AP) を求め，その平均値を評価する[10].

4.2 検索モデルの構築

本稿では，名詞のみ (ストップワードあり) を使用して検索モデルを構築するが，その際音声認識結果だけではなく，音声ドキュメントの正解文に対しても構築した．

まずグローバルモデルには，ipadic と unidic を混合した形態素解析器の辞書の全ての名詞 287715 単語に対して，それぞれの単語の Web 検索ヒット数を得ることで推定した．

次に関連文書モデルを作成する際，事前ダウンロードデータ群として，同様に形態素解析器の辞書の全ての名詞 287715 単語をそれぞれキーワードとして Web キーワード検索を行い，キーワードごとに 50URL の文書を取得することで約 1406 億単語のデータ群を構築した[8]．よって関連文書候補群の文書数は，認識結果に出現した名詞数 ×50 となる．この関連文書候補群から関連文書群を選択する際，t 番目の関連文書候補の多項分布 θ_{S_t} の推定に

表 2 パラメータの推定値
Table 2 Estimated values of each parameter

Smoothing method	μ (Correct)	η (Correct)	μ (Hypothesis)	η (Hypothesis)
Collection	186.79	-	394.38	-
Global	165.42	-	317.82	-
Collection+Global	186.25	15199.71	393.41	6642.18

は，MAP 推定 (グローバルモデルを事前分布のパラメータの基底に使用) を用い，文書モデルのパラメータは最尤推定値を用いて KL ダイバージェンスを計算した．

なお，以上の Web 検索には Yahoo! Japan の検索エンジンを用いた．

4.3 コレクションモデルに対するスムージングの効果

コレクションモデルに対するスムージングの効果について調査する．この実験では文書モデルの MAP 推定時に，(6) 式の $P(w_i|\theta_C)$ にどのパラメータを使用するかで比較を行う．まず従来法として，通常の文書検索で一般的に行われる $P(w_i|\theta_C)^{ML}$ を利用してスムージングする場合 (Collection)，また $P(w_i|\theta_C)$ の部分に直接 $P(w_i|\theta_G)^{ML}$ を利用してスムージングする場合 (Global)，そして提案法として，事前に $P(w_i|\theta_G)^{ML}$ を利用することでスムージング済みの $P(w_i|\theta_C;\eta)^{MAP}$ を利用してスムージングする場合 (Collection+Global) の 3 方法で比較を行う．各方法に対して，正解文 (Correct)，認識文 (Hypothesis) 両者に対して文書モデルを作成し，検索実験を行った．各パラメータの推定値を表 2，実験結果を図 1 に示す．

図 1 より，音声認識結果から文書モデルを推定する際 (Hypothesis) に着目すると，提案法 (Collection+Global) において 11ptAP で 0.4012 を得て，従来法 (Collection) と比較して 0.0233 ポイントの改善があった．実際に正解の書き起こしの文書コレクションの語彙には存在するが認識結果の文書コレクションには存在しない単語が約 1 万存在していたが，これらの単語に生成確率を付与することができていた．またグローバルモデルを文書モデル自体のスムージングに利用する (Global) と性能が劣化してしまうので，グローバルモデルをコレクションモデルのスムージングに適用すべきであることが分かる．なお正解文から文書モデルを推定する際 (Correct) にも提案法の効果があるのは，検索要求が「煙草」であったが文書コレクション内には「たばこ」しかなく，グローバルモデルによって零確率問題を回避できたからである．

4.4 Web 関連文書を用いた混合モデル化の効果

Web 関連文書を用いた文書モデルの混合モデル化の効果について調査する．本実験では前

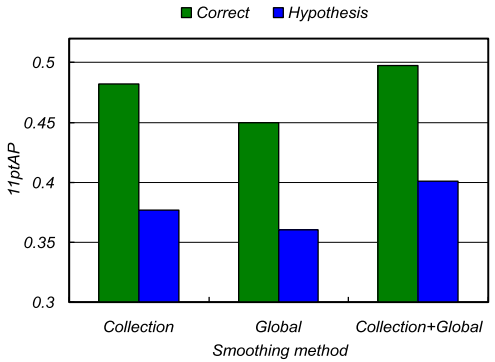

図 1 スムージング手法ごとの検索性能
Fig. 1 Retrieval perfoemace of each smoothing method

述の実験結果を反映させて，(6) 式および (14) 式の $P(w_i|\theta_C)$ には (10) 式を用いる．よって図 1 の (Hypothesis) の (Collection+Global) がこの実験における Baseline，また (Correct) の (Collection+Global) がこの実験における Correct となる．

まず (17) 式において $\lambda = 0$ とし，関連文書モデルのみで音声ドキュメントの文書モデルを構築した場合について調査した．それぞれの音声ドキュメントに対して，Web から収集した文書を KL ダイバージェンスでランキングした後，関連文書に関するモデルを推定するための関連文書数 N を変化させた際の結果を次の図 2 に示す．

図 2 より，ある程度多くの関連文書を用いて関連文書モデルを作成することで，$P(w|\theta_R)$ のみでも高い検索性能を実現できることが分かった．$N = 3000$ で最も高い検索性能が得られ，11ptAP で 0.3125 が得られた．関連文書候補群として抽出した文書数は，テストコレクション平均で約 11700 文書 (認識結果に出現した名詞数 $\times 50$) であったので，KL ダイバージェンスによるランキング後，上位約 1/4 程度の文書を関連文書群として用いればよいことが分かる．

次に，選択する関連文書数 $N = 3000$ と固定して，λ の値を変化させて各音声ドキュメントに対して文書モデルを構築し，検索実験を行った．その結果を図 3 に示す．

図 3 より，λ をある程度高く設定することで，Web 関連文書を用いた混合モデル化により検索性能が上昇することが分かる．これは認識結果から推定した文書モデルでは生成確率

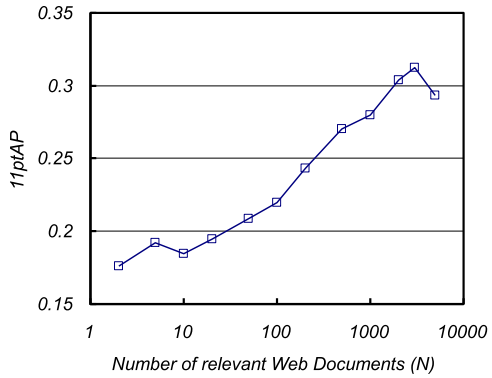

図 2 Web 関連文書数の変化による検索性能
Fig. 2 Retrieval performance with differenct number of relevant Web documents

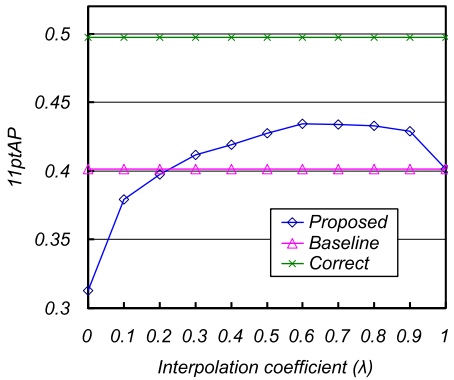

図 3 手法ごとの検索性能
Fig. 3 Retrieval performance of each method

が低かった重要単語に対して，関連文書モデルにより生成確率を補間できたことに起因すると考える．$\lambda = 0.6$ の時に 11ptAP で 0.4343 が得られ，混合モデル化を行う前と比較して，0.0331 ポイントの改善が得られた．

5. ま と め

本稿では，Web を利用して欠落単語を補間するアイデアを確率的言語モデルによる検索モデルの枠組みに組み込むために二つの方法を提案した．まず，零確率問題に対処するために，文書コレクションに対しても Web 検索ヒット数に基づく事前分布をおく方法を提案した．次に，重要単語に対する生成確率の補間のために，Web 関連文書を用いて音声ドキュメントの文書モデルを混合モデル化する方法を提案した．検索実験から，前者により約 0.0233 ポイントの検索性能の改善，後者によりさらに約 0.0331 ポイントの検索性能の改善を得た．

参 考 文 献

1) R.Masumura, A.Ito, Y.Uno, M.Ito and S.Makino, “ Document Expansion using Relevant Web Documents for Spoken Document Retrieval ”, In Proc. International Conference on Natural Language Processing and Knowledge Engineering, pp.612-619, 2010.
2) J.M.Ponte and W.B.Croft, “ A language modeling approach to information retrieval ”, In Proc. SIGIR 1998, pp.275-281, 1998.
3) T.K.Chia, H.Li and H.T.Ng, “ A Statistic Language Modeling Approach to Lattice-Based Spoken Document Retrieval ”, In Proc. Joint Meeting of the Conference on Empirical Methods in Natural Language Processing and the Conference on Natural language Learning, pp.810-818, 2007.
4) K.Honda and T.Akiba, “ Language Modeling Approach for Retrieving Passages in Lecture Audio Data ”, In Proc. International Conference on Language Resources and Evaluation (LREC 2010) , pp.1525-1535, 2010.
5) B.Chen, “ Latent Topic Modeling of Word Co-Occurrence Information for Spoken Document Retrieval ”, In Proc International Conference on Acoustic, Speech, and Signal Processing, pp.3961-3964, 2009.
6) X.Hu, R.Isotani, H.Kawai and S.Nakamura, “ Cluster-based Language Model for Spoken Document Retrieval Using NMF-Based Document Clustering ”, In Proc. Interspeech, pp705-708, 2010.
7) C.Zhai and J.Lafferty, “ A study of smoothing methods for language models applied to information retrieval ”, ACM TOIS, vol.22, no.2, pp.179-214, 2004.
8) 増村 亮, 咸 聖俊, 伊藤 彰則, “ 教師なし言語モデル適応のための Web Document を用いた単語のトピック表現 ”, 情報処理学会研究報告, Vol.2010-SLP-82-16, 2010.
9) X.Wei and W.B.Croft, “ LDA-based document models for ad-hoc retrieval ”, In Proc. SIGIR 2006, pp.178-185, 2006.
10) T.Akiba, K.Aikawa, Y.Ito, T.Kawahara, H.Nanjo, H.Nishizaki, N.Yasuda, Y.Yamashita and K.Ito, “ Test collections for spoken document retrieval from lecture audio data ”, In Proc. International Conference on Language Resourses and Evaluation, 2008.