音声ドキュメントを対象とした情報検索

秋葉友良 (メディア科学リサーチセンター, セマンティックアーカイブコア)

1 はじめに

音声認識技術の高度化に伴い、本来コミュニケーションの手段として用いられてきた話しことばによる音声を、知識や技術を伝達するメディアとして利用することが可能になってきた。我々が特別な道具を要すること無く日常的に発する音声を、年々大容量化するストレージに蓄え、高速化するネットワークを介して容易にアクセス可能な「文書」として扱うことができれは、ことばを利用した人間の知的活動は飛躍的に拡大するであろう。このような「音声ドキュメント処理」にとって鍵となる研究課題は、検索、要約、コンテンツ生成、など多岐にわたる。本稿では音声ドキュメントの検索に焦点を当て、筆者が関わってきた音声ドキュメント検索評価用テストコレクション構築活動、および検索手法開発・評価の研究活動について報告する。

2 音声ドキュメント検索評価用テストコレクション

音声ドキュメントの検索は、米国 NIST 主催の評価型ワークショップ TREC において、1997 年から 2000 年の間 Spoken Document Retrieval(SDR) Track にて、大規模な評価実験が行われている。TREC SDR では、英語の放送ニュース音声を対象に、最終的には約 557 時間の音声ドキュメントを対象とした評価用テストコレクションが構築された。また、TREC SDR テストコレクションの検索クエリを翻訳することによって、ヨーロッパ各国の言語から英語への言語横断検索テストコレクションが構築され、ヨーロッパにおける言語横断検索評価型ワークショップである CLEF(Cross Language Evaluation Forum) において利用されている。

音声処理の分野において「検索」という日本語は、既知の語を見つけることを目的とした、いわゆる「キーワード検索」の意味で使われることが多い。しかし、情報検索や言語処理分野で主な研究対象であるのは、検索者の情報要求を表す表現 (たとえば、キーワードリストや自然言語文) からそれに適合する文書を検索する「アドホック検索」である。前述の TREC SDR においても、初年度は「既知語の検索」(Known Item Search) をタスクとしたが、2 年目以降はアドホック検索をタスクとしてきた。

TREC や CLEF などでの音声ドキュメント検索の研究活動に対し、日本では長年の間、音声ドキュメント検索用のテストコレクションが存在していなかったが、2006 年 4 月に組織された情報処理学会音声言語情報処理研究会の「音声ドキュメント処理ワーキンググループ」の活動として、「日本語話し言葉コーパス」(以下、CSJ と略す) を対象とした音声ドキュメント検索評価用テストコレクション (以下、CSJ テストコレクション) が構築されつつある [3, 1]。CSJ テストコレクションと TREC SDR テストコレクションとの比較を表 1 に示す。

Table 1　TREC SDR と CSJ テストコレクションの比較

	TREC9 SDR	CSJ
言語	英語	日本語
対象文書	ニュース音声	講演音声
時間	557 時間	623.6 時間
文書数	21,754	2,702 (30,762)
単語/文書	169	2,324.9 (204.2)
検索クエリ	50	39
書き起し	低品質 (WER 10.3%)	高品質
音声認識 WER	26.7%	21.4%

(CSJ の括弧内数字は、30 発話単位を文書とした場合。)

3 統計的翻訳モデルを利用した音声ドキュメントのアドホック検索

音声ドキュメントを対象としたアドホック検索は、大語彙連続音声認識を用いて対象音声ドキュメントをテキストへと変換すれば、既存のテキストを対象とする検索手法がそのまま適用可能である。その際に問題となるのは、まず音声認識誤りの影響による自動書き起しテキストの劣化の問題がある。特に、現在の典型的な大語彙連続音声認識は数万語の認識辞書で構成されており、辞書にない語を認識することはできない。したがって、そもそも認識辞書に含まれない語は、検索質問中に現れてもそれを含む文書を見つけることができない。この問題に対し、音声認識による自動書き起しと人手書き起しの間の差異を「翻訳」によって補完する検索手法を開発した [4, 5, 2]。

提案手法は、音声認識による自動書き起しテキストに対し、人手書き起しされた場合に現れるであろう語を使って直接索引付けすることにより、自動書き起しと人手書き起しの差異を補完する。この索引付けには、自動書き起しテキストに現れる単語 e が、人手書き起しテキストにおいて単語 f として現れる確率 $t(f|e)$ を利用する。この確率を、統計的機械翻訳の用語に倣って単語翻訳確率と呼ぶ。

3.1 単語翻訳確率の推定

単語翻訳確率 $t(f|e)$ の推定には、音声認識結果の自動書き起しテキストと、人手による書き起しテキストのペアによるパラレルテキストを用いる。

まず、パラレルテキストの両サイドを形態素解析し単語列を得る。次に、この単語列ペアに対し、編集距離を指標とする DP マッチングを適用する。その結果から、自動書き起しと人手書き起しが完全一致する単語どうしについてのみアライメントを抽出し、これを初期アライメントとする。残りの単語、すなわち自動書き起しと人手書き起しが一致しない単語間については、可能なアライメントを列挙し、それぞ

れに断片的な回数でアライメントが出現したとして、部分的なアライメントを与える。これら単語アライメントをパラレルテキスト全体で収集し、最尤推定によりパラメータ推定を行なった。

3.2 単語翻訳確率を用いた音声ドキュメントの索引付け

ある音声ドキュメント D の自動書き起しに現れる単語集合を E_D とする。自動書き起しテキストからそのまま索引付けする場合、単語 $e \in E_D$ の D での単語頻度 $\mathrm{TF}_E(e, D)$ を元にした統計情報を利用し、例えば TF-IDF などの単語重みを利用して索引付けが行なわれる。一方、D を人手書き起しテキスト (単語集合 F_E) で索引付けする場合、同様に、単語 $f \in F_D$ の単語頻度 $\mathrm{TF}_F(f, D)$ を元に索引付けが行なわれる。この $\mathrm{TF}_F(f, D)$ の期待値 $E(\mathrm{TF}_F(f, D))$ は、単語翻訳確率 $t(f|e)$ を用いて、以下のように求めることができる。

$$E(\mathrm{TF}_F(f, D)) = \sum_{e \in E_D} t(f|e)\mathrm{TF}_E(e, D) \quad (1)$$

さらに、スムージングのため、$TF_E(e, D)$ との線形補間を行なう。

$$\tilde{\mathrm{TF}}_F(f, D) = \lambda E(\mathrm{TF}_F(f, D)) + (1 - \lambda)\mathrm{TF}_E(f, D) \quad (2)$$

この $\tilde{\mathrm{TF}}_D(f, D)$ を用いて、音声ドキュメントの自動書き起しテキストの索引付けを行う。ただし、閾値 α を設けて、α 以下の $\tilde{\mathrm{TF}}_D(f, D)$ となる単語では索引付けを行わない。

以上により正解テキストに含まれると期待される単語で索引づけされた音声ドキュメントに対し、既存のアドホック検索手法を適用し、音声ドキュメントの検索を行なう。

3.3 翻訳モデルの学習事例

パラレルテキストから学習された高頻度のアライメントの具体例を以下に示す (ただし、自動書き起し→人手書き起しの順)。

- 同音異義語

  感染→観戦, 解放→開放, 式→四季, 創造→想像, 下降→加工, ここ→個々, そこ→底

- 発音が類似

  研究→言及, 要素→様相, 構成→個性, 計算→欠損, 実験→事件, 情報→譲歩, 加工→確保, 父→土

3.4 実験結果

2 節で述べた CSJ テストコレクションを用いて、提案手法の評価を行った。音声ドキュメントを音声認識して得られた認識候補の 1 位のみ (1-best)、および、1 位から 10 位まで (10-best)、を索引付けに利用した従来手法との比較を行った。また、人手書き起しテキスト (認識誤り無しに相当) で索引付けした場合とも比較を行った。検索対象の文書は、講演音声を先頭から 15、30、60 発話で切り出すことで自動的に作成した。検索性能の評価尺度には、0.0 から 1.0 まで 0.1 刻みの各再現率レベルにおける補間精度を平均した 11 点平均精度を用いた。実験結果を図 1 に示す。

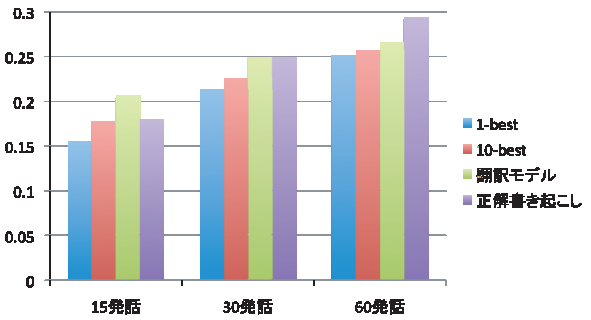

Fig. 1　11 点平均精度

正解発話区間 15,30,60 発話のすべてのタスク設定において、提案手法は従来手法を上回る性能を示した。性能向上は、短い発話区間を正解とした場合に顕著である。特に、15 および 30 発話区間を正解とした場合、提案手法は人手書き起しテキストを用いた場合より高い性能を示した。一方、60 発話区間を正解としたタスクにおいては、ベースライン手法との性能差は比較的小さい。

検索対象文書の語数が少ない場合、提案法は文書表現の拡張として働き、特に有効に機能したと考えられる。その効果は認識の複数候補を使う場合 (ベースライン 10-best) よりも大きい。一方、文書長が十分大きい場合は、その文書中の索引語だけで多様性が表現できるため、逆に誤った索引語を登録することによるノイズの影響が大きくなったと考えられる。

発表論文

[1] T. Akiba, K. Aikawa, Y. Itoh, T. Kawahara, H. Nanjo, H. Nishizaki, N. Yasuda, yoichi Yamashita, and K. Itou. Test collections for spoken document retrieval from lecture audio data. In *Proceedings of International Conference on Language Resources and Evaluation*, 2008.

[2] T. Akiba and Y. Yokota. Spoken document retrieval by translating recognition candidates into correct transcriptions. In *Proceedings of International Conference on Speech Communication and Technology*, 2008. (to appear).

[3] 秋葉, 相川, 伊藤, 河原, 南條, 西崎, 安田, 山下, 伊藤. 音声ドキュメント検索テストコレクションの試作と基本検索性能評価. 第 1 回音声ドキュメント処理ワークショップ講演論文集, pp. 73–80, 2007.

[4] 秋葉, 横田. 翻訳モデルに基づく講演音声ドキュメントのアドホック検索. 日本音響学会秋季研究発表会講演論文集, 2007.

[5] 秋葉, 横田. 認識候補から正解テキストへの翻訳モデルに基づく講演音声ドキュメントのアドホック検索. 第 2 回音声ドキュメント処理ワークショップ講演論文集, pp. 79–84, 2008.