5Q-08

音声中の検索語検出におけるドキュメント間類似度を利用したリスコアリング方式

清水嘉乃† 小嶋和徳† 李時旭‡ 伊藤慶明†
岩手県立大学† 産業技術総合研究所‡

1. はじめに

TED(Technology Entertainment Design)[1]を始めとする利用可能な講義ビデオの増加に伴い，大量の音声データ中から特定の単語が発話されている区間を検索する音声中の検索語検出(STD: Spoken Term Detection)の研究が盛んに行われている[2][3].

一般的な STD システムでは，検索対象の音声データ(音声ドキュメント)を自動音声認識し，サブワード化しておく．クエリが与えられると，サブワードに変換し，音声ドキュメントとサブワードレベルでの照合を行い，距離の小さい区間順にユーザへ提示する．STD では誤認識等の要因による精度低下に対する改善が求められている．

検索結果において，正解区間は距離が小さく，その区間を含む音声ドキュメント内には他にも正解区間が含まれると考えられる．そこで先行研究[4]では，音声ドキュメント内の上位候補の距離を用いて各候補の距離を補正し，上位候補の距離が小さい音声ドキュメントの候補に対して優位にするリスコアリングを行い，精度向上を実現した．

一方，音声ドキュメント内に正解区間が含まれているにもかかわらず，上位候補の距離が大きい場合，補正効果が小さくなる可能性がある．そこで本稿では，内容の類似している音声ドキュメント間で共通してクエリ等のキーワードが発話されている可能性が高いと仮定し，高順位候補を含む音声ドキュメントと内容の類似している音声ドキュメント(類似ドキュメント)を特定し，その音声ドキュメント内の候補に対して距離補正を行うリスコアリング方式を提案する．実験により，先行方式との比較，及び併用による精度改善を検証する．

2. 先行方式[4]

STD 結果において，高順位候補を含む音声ドキュメント内には複数の正解区間が含まれており，高順位候補の距離が小さいほど正解区間が複数含まれている可能性が高いと考えられる．そこで，高順位候補の距離を用いて，その候補が含まれる音声ドキュメント内の候補に対して式(1)により補正を行う．

$$\mathrm{newD}_{\Omega_i k} = \alpha \mathrm{D}_{\Omega_i k} + (1-\alpha)\frac{\sum_{t=1}^{T} \mathrm{D}_{\Omega_i t}}{T} \quad (1)$$

音声ドキュメントΩ内でのk番目の順位の候補がΩ内でi番目の発話Ω_iであった場合，その距離を$\mathrm{D}_{\Omega_i k}$，補正後の距離を$\mathrm{newD}_{\Omega_i k}$とする． $\alpha(0 \le \alpha \le 1)$は補正係数である．

3. ドキュメント間類似度を利用したリスコアリング方式（提案方式）

高順位候補を含む音声ドキュメントには正解が含まれると仮定したように，音声ドキュメントと類似する音声ドキュメントにも正解区間が含まれていると仮定し，その類似ドキュメント内の候補に対して距離の補正を行う．

音声ドキュメント毎に類似度の高い音声ドキュメントをリスト化した類似ドキュメントリストを予め作成しておく．類似ドキュメントリストの作成手順は以下の通りである．

① 音声ドキュメントの単語認識結果を形態素解析
② Distributed Memory of Paragraph Vector(PV-DM)[5]を用いて各音声ドキュメントを特徴ベクトル化
③ 特徴ベクトルから各音声ドキュメント間の類似度を算出．類似度の高い順にリスト化

PV-DM は任意長の文章から固定長の特徴ベクトルを生成する教師なしニューラルネットワークであり，単語の意味や文脈/語順を考慮したベクトル化が可能である．

クエリが与えられると，まずそのクエリに対しての検索結果上位 N 件の候補を抽出し，それらの候補を含む音声ドキュメントを特定する．N 件の上位音声ドキュメントについて，事前に作成しておいた類似ドキュメントリストを参照し，上位音声ドキュメントとその類似ドキュメントの表中で類似度が高い音声ドキュメントから順に上位 M 件を補正対象音声ドキュメントとする．補正対象音声ドキュメントに対して取得順位ごとの補正係数を設定し，式(2)を用いて補正する．

$$\mathrm{newD}_{\Omega_i} = \alpha_m \times \mathrm{D}_{\Omega_i} \quad (2)$$

Ω_iにおける距離をD_{Ω_i}，補正後の距離をnewD_{Ω_i}，類似度が高い順に音声ドキュメントを取得した際，

A Re-scoring Method for Spoken Term Detection Using Simillarty between Apoken Documents.
†Shimizu Yoshino， Kazunori Kojima， Yoshiaki Itoh・ Iwate Prefectural University
‡Shi-wook Lee，・National Institute of Advanced Industrial Science and Technology

Ωの順位が$m(1 \leq m \leq M)$位であった場合の補正係数を$\alpha_m(0 \leq \alpha_m \leq 1)$とする．

4. 評価実験

4.1. 実験条件

音響モデル，言語モデルの学習には日本語話し言葉コーパスの全 2,702 講演から Core と呼ばれる 177 講演を除いた偶数講演(1,255 講演，約 287 時間)を用いた．自動音声認識には DNN-HMM を用いた．音響モデルは 3 状態の triphone を用い，状態共有により 3,009 状態とした．DNN への入力特徴量はフィルタバンク 120 次元とし，前後 5 フレームを追加した 1,320 次元（11 フレーム×120 次元），出力は各状態の事後確率とした．

表 1 に DNN，表 2 に PV-DM の学習条件を示す．

表 1　DNN の学習条件

ノード数		入力層 1320 隠れ層 2048 出力層 3009
隠れ層		5
RBM	学習係数	0.004
	モメンタム	0.9
	ミニバッチサイズ	256
	エポック数	10
DNN	学習係数	0.007
	ミニバッチサイズ	256

表 2　PV-DM の学習条件

次元数	400
窓長	20
単語の最低出現頻度	1
初期学習率	0.025
最終学習率	0.0001

4.2. テストセット

評価には NTCIR-10[2]と NTCIR-12[3]のテストセットを用いた．NTCIR-10 における検索対象は SDPWS 104 講演，クエリは Formal run 100 個を用いた．NTCIR-12 における検索対象は SDPWS 98 講演，クエリは Formal run Single Term 113 個を用いた．

各方式における閾値と補正係数は，NTCIR-10 と NTCIR-12 のテストセットでのクロスバリデーションにより決定し，オープンな評価とした．

5. 実験結果

図 1 に実験結果を示す．リスコアリング方式適用前を Baseline，各リスコアリング方式を併用して適用する場合，その適用順により，先行+提案あるいは提案+先行と示した．

各方式を単体で適用した場合の精度は，NTCIR-10 では提案方式，NTCIR-12 では先行方式が高くなった．方式を併用することで，両テストセットで精度が向上し，Average において，提案+先行で最も高い精度となり，有意水準 5%の T 検定において有意差を確認した．

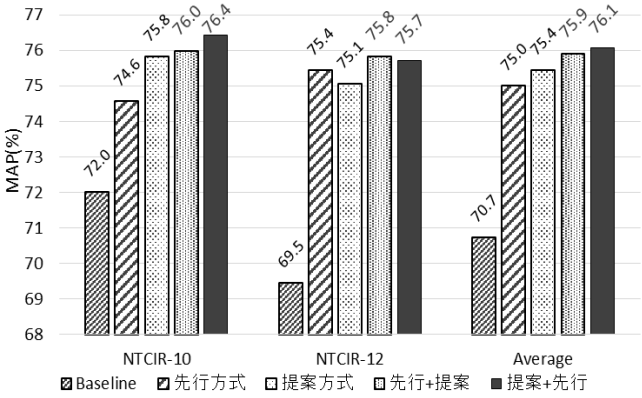

図 1　実験結果

提案方式における補正係数α_mは，補正対象音声ドキュメントに対してその類似度の順位ごとに 0.1 刻みで最適な MAP を示すように設定した．オープンな評価とするため，α_mは NTCIR-10 及び NTCIR-12 のテストセットでのクロスバリデーションで確定した．パラメータは両テストセットで 0.6～0.9 の間でほぼ同等となった．補正対象音声ドキュメントの再現率が高い検索結果の内，適合率が 20%以下の低適合率のものが複数存在していた．補正音声ドキュメント数を全検索結果に対して固定したのが原因と考えられる．これに伴い，最適な補正対象音声ドキュメントを抽出する方法について，今後検討する．

6. おわりに

提案方式により，Baseline からの精度向上，及び先行方式との併用による各リスコアリング方式を単体で適用した場合からの精度向上を確認した．今後は最適な補正対象音声ドキュメントを抽出する方法について検討する．

謝辞

本研究の一部は JSPS 科研費 15K00241，JP15593778 の助成を受けたものです．

参考文献

[1] Technology Entertainment Design, TED Conference, https://www.ted.com

[2] Tomoyoshi Akiba, et al., “Overview of the NTCIR-10 Spoken Doc-2 Task”, Proceedings of the NTCIR-10 Conference, 2013.

[3] Tomoyoshi Akiba, et al., “Overview of the NTCIR-12 Spoken Doc-2 Task”, Proceedings of the NTCIR-12 Conference, 2016.

[4] 小嶋和徳 他, 音声中の検索語検出における同文書内の高順位候補を利用したリランキング方式, 電子情報通信学会論文誌D Val.J100-D No.1, pp. 70-80, 2017.

[5] Quoc Lee, Tomas Mikolov, Distributed Representations of Sentences and Documents, Proceeding of The 31st International Conference on Machine Learning(ICML), pp.1188-11961, 2014.