情報処理学会論文誌 Vol.61 No.1 103–112 (Jan. 2020)

音声中の検索語検出における クエリの関連語を利用したリスコアリング方式

丹治 遥[1,a)] 小嶋 和徳[1] 李 時旭[2] 南條 浩輝[3] 伊藤 慶明[1]

受付日 2019年5月21日, 採録日 2019年10月3日

概要：音声中で検索したいキーワードが話されている箇所を特定する音声中の検索語検出（STD：Spoken Term Detection）の研究がさかんに行われている．検索精度向上のために，先行研究として高順位候補を含むドキュメント内のすべての候補の照合距離を有利にする方式等が提案されている．本論文では，クエリを含む講演内で話されるトピックの内容に関連してクエリと共起する単語をクエリの関連語と呼び，関連語は当該講演内に複数回出現すると仮定する．クエリの関連語を特定するため，本論文では Word2vec を用いた単語の分散表現が有効と考える．音声ドキュメントの単語認識結果中の各単語を Word Vector 化し，クエリの Word Vector と比較し，類似度を求めることでクエリの関連語を取得する．一方，未知語（OOV：Out-of-Vocabulary）クエリは単語認識結果に出現しないため Word Vector を算出できないため，本論文では Web 検索を併用する方式を採用し，クエリで Web 検索し得られたテキスト中の出現単語も Word2vec に用いてクエリの意味的情報を補い，未知語クエリの Word Vector を算出できるようにする．これにより，未知語クエリに対応させることができ，既知語（IV：In Vocabulary），未知語いずれのクエリでも関連語を的確に求められると考える．以上のようにして，クエリの関連語を特定し，関連語を含むドキュメント内のすべての候補の距離を有利にすることで検索精度の向上を図る．NTCIR-10，12 の Formal Run の 2 種のテストセットを用いて評価した結果，両テストセットで検索精度が向上した．また，先行方式と併用することでさらに精度が向上し，提案方式の有効性を確認できた．

キーワード：音声中の検索語検出，クエリの関連語，Word Vector，リスコアリング

Rescoring by Using Words Related to a Query for Spoken Term Detection

Haruka Tanji[1,a)] Kazunori Kojima[1] Shi-Wook Lee[2] Hiroaki Nanjo[3] Yoshiaki Itoh[1]

Received: May 21, 2019, Accepted: October 3, 2019

Abstract: We propose a rescoring method using words related to a query for spoken term detection (STD). In this paper, we assume that words associated with the topic in speech data and co-occurring with the query are called “words related to the query,” and that the related words appear multiple times in the speech data. To identify the words related to the query, we introduce distributed expression of words obtained by Word2vec, and first convert each word in the word recognition results of speech data into a word vector. Each word vector is then compared with a word vector of the query. Words related to the query are determined by calculating the degree of similarity between the two word vectors. However, a word vector of an out-of-vocabulary (OOV) query cannot be obtained in this manner, since OOV queries do not appear in word-recognition results. For such OOV queries, we perform a Web search using the query, whereupon texts including the query are extracted. Recognition results of the speech data and the extracted texts are then combined and used for training of Word2vec. Distances to all candidates in the document, including words related to the query, are used advantageously. Experiments are conducted to evaluate the performance of the proposed method using open test collections of the NTCIR-10 and NTCIR-12 workshops. For retrieval accuracy, an improvement of 3.2 points in mean average precision was achieved using the proposed method.

Keywords: spoken term detection, words related to the query, word vector, rescoring

[1] 岩手県立大学
Iwate Prefectural University, Takizawa, Iwate 020–0693, Japan
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki 305–8560, Japan
[3] 京都大学学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University, Kyoto 606–8501, Japan
[a)] g231q018@s.iwate-pu.ac.jp

1. はじめに

近年，Web 上の動画投稿サイトや Blu-ray Disc レコーダの利用が一般的となっており，動画データや音声データ等，音声を含む大量のビデオデータ（音声ドキュメント）を保存する機会が増加している．これにともない，大量のビデオデータからユーザが所望する特定のキーワードが話されている区間を検索する機能に対するニーズが高まっている．この機能の実現のため，ビデオデータ中の音声情報を用いて検索語（クエリ）を検索する音声中の検索語検出（STD：Spoken Term Detection）の研究がさかんに行われている．国立情報学研究所が主催する NTCIR Workshop 9 [1] が 2011 年に，NTCIR Conference 10 [2] が 2013 年に，NTCIR Conference 11 [3] が 2014 年に，NTCIR Conference 12 [4] が 2016 年に開催され，STD について様々な観点から評価された．

STD とは，音声ドキュメント内で 1 つまたは連続する複数の単語からなるクエリが出現する区間を特定するタスクである．一般的な STD システムでは，大語彙連続音声認識システムを用いて検索対象の音声ドキュメントをあらかじめ単語単位で認識し，その認識結果を用いて検索を行う．一方，音声認識システムの単語辞書に登録されていない単語が STD においてクエリとなると音声認識時の誤認識のために検索ができない．この未知語のクエリに対応するため，単語より小さい単位のサブワードレベルでの認識結果を用いて照合を行う方式が一般的である [5]．これはクエリのサブワード系列と検索対象音声ドキュメントのサブワード系列を照合し，その照合距離の小さい順に候補として出力するものである．ただし，同音異義語や音素列の近い別の語も照合距離が小さくなるため誤検出が増えるという問題がある．

音声ドキュメントは一般に話題，対話，セッション，講義，講演単位等で分けられており，NTCIR の評価セットにおいても講演ごとに分かれている．たとえば，クエリを「東北」とした場合，ある講演中に「東北」と話されていれば，その講演では「東北」に関する内容が話されていると想定できる．「東北」に関連した「仙台」「奥羽地方」「田舎」等の単語も話される可能性がある．本論文ではこのように同一講演内で話されるトピックの内容に関連してクエリと共起する単語をクエリの関連語と呼び，クエリおよびクエリの関連語はクエリが出現する講演内に複数回出現すると仮定し，関連語を抽出した後，その関連語を含む講演内のすべての候補の距離を有利にすることで検索精度の向上を図る手法に取り組む．

先行研究 [6] では，照合により得られた高順位候補は高い適合率を示し，クエリは特定のドキュメントに頻繁に出現する傾向が強いことから，高順位候補を含むドキュメントにはクエリが複数含まれていると仮定し，高順位候補を含むドキュメント内のすべての候補の照合距離が小さくなるよう補正を行うことで，検索精度の向上を実現した．文献 [7] では，高順位候補を含むドキュメントと内容が類似しているドキュメントにもクエリが複数含まれていると仮定し，類似ドキュメント内のすべての候補の照合距離が小さくなるよう補正を行うことで，検索精度のさらなる向上を実現した．これらはいずれもクエリが特定のドキュメントに頻出することを仮定した手法である．

情報検索においては，Web 上のテキストを利用し取得した関連語を利用するクエリ拡張（Query Expansion）が用いられているが，本研究もその一種と位置付けられる．本研究では関連語を取得するために Word2vec を利用し，これを STD に応用した点に特徴がある．一方，単語共起情報の利用や検索語を加工する等の以下の研究事例がある．

文献 [8] では，クエリとよく共起する単語が検索結果の候補の周辺に出現していれば，当該候補はクエリを含む正解発話である可能性が高いと考え，その候補の距離を有利にすることで検索精度の向上を実現した．この共起単語の情報は Web 検索により得られるテキストから取得していたが，共起単語を正しく得られない場合はリスコアリングできず精度が向上しなかった．文献 [9] では，検索結果の候補を含む講演の話題を調べ，クエリの説明文と講演の意味的な類似性を音声内容検索によって求め，類似する候補の距離を有利にすることで検索精度の向上を実現した．文献 [10] では，クエリの前または後に格助詞を付与したものをクエリの拡張語とし，拡張語で検索した結果を利用して元の検索結果を補正することで検索精度の向上を実現した．文献 [11] では，音声内容検索においてベクトル空間モデルとクエリ尤度モデルに単語共起情報を導入しモデルを拡張することで検索性能の向上を実現した．

本研究も上記のいずれの研究と同様クエリ拡張を利用した手法である．本研究は文献 [8] と同様，クエリとよく共起する語（関連語）が検索結果の候補付近に出現していれば，当該候補はクエリを含んでいる可能性が高いという仮定に基づき，照合距離の補正を行うものである．文献 [8] では Web 上のテキストから学習した共起単語情報と候補の単語信頼度に基づいた補正を行っているが，本研究では共起単語（関連語）を含む講演中のすべての候補の照合距離が有利になるよう補正する．具体的には，関連語を含む講演はその関連語の出現頻度が高いほどクエリを含んでいる可能性が高いと想定し，特定の講演にある特定の語が頻出するという先行研究 [6] の知見に基づき，関連語を含む講演内のすべての候補の照合距離を関連語の出現頻度の大きさに応じて有利にする処理を行うもので，この点において新規性を有すると考える．

本論文では，クエリの関連語を見つけるために Word2vec [12], [13] を用いる．Word2vec は単語間の関連性を表現できる単語の分散表現を求める手法であり，これに

より求まった各単語の特徴ベクトル（以降，Word Vector）を用いることで単語間の類似度を求めることができる．本論文では，音声ドキュメントを単語認識してその出現単語を Word Vector 化し，クエリと各単語の Word Vector を用いて類似度を計算し，クエリの関連語を求め，これを STD に用いる手法を提案する．一方，未知語クエリは音声認識結果には含まれないため，このような関連語の抽出方式では，未知語クエリの Word Vector を算出できない．本研究ではこの問題も扱う．具体的には Web 検索を併用する方式を採用する．Web 検索では検索単語に関するタイトルとスニペット（以降，Web テキスト）が複数出力される．この Web テキストには検索した単語が出現しており，その単語の意味や単語に関する話題等が含まれている．そこで，クエリ単語での検索結果の Web テキスト中の単語も Word Vector の学習に用いることで未知語クエリの単語的意味を学習し Word Vector を求めることができる．以上により，クエリと各単語の Word Vector を求め，クエリとの類似度を計算することで関連語が複数個得られる．

本論文では，以上のようにして選定した関連語を含む講演を抽出し，それらの講演内のすべての候補の照合距離を，その関連語の出現頻度の大きさに応じて有利になるよう補正（リスコアリング）する方式を提案し，その有効性を示す．

本論文の構成は次のとおりである．2 章では先行研究 [6] の高順位候補を含むドキュメント優先方式について，3 章では提案方式である Web 検索と Word Vector を用いたリスコアリング方式について述べる．4 章では提案方式の評価実験，先行研究 [6] との比較統合について述べる．5 章で結論を述べる．

2. 高順位候補を含むドキュメント優先方式

先行研究の高順位候補を含むドキュメント優先方式 [6] について概説する．1 章で述べたとおり高順位候補を含むドキュメントにはクエリが複数含まれていると仮定する．音声ドキュメントは講演音声ファイル（$\Omega = \Omega_1, \Omega_2, \Omega_3 \ldots$）で構成されているとし，まず，STD を行った結果を講演ごとに分類・順位付けを行う．たとえば，講演 Ω_i 内の高順位候補とクエリとの照合距離は小さい場合に，講演 Ω_i にはクエリを複数含んでいる可能性が高い．そこで，高順位候補の照合距離を用いて講演 Ω_i 内の下位の候補区間の照合距離に対して，以下の式 (1) により調整（リスコアリング）を行う．α（$0 \leq \alpha \leq 1$）は重み係数を表す．たとえば講演 Ω_i の j 番目の発話が Ω_i 内で k 位であった場合，このときの照合距離を $D(\Omega_i(j), k)$ とする．リスコアリング後の照合距離 $D'(\Omega_i(j), k)$ は，その候補区間の元々の照合距離 $D(\Omega_i(j), k)$ と 1〜T 位（$1 \leq t \leq T$）までの候補の照合距離 $D(\Omega_i(j), t)$ の平均を線形結合することで求められる．

$$D'(\Omega_i(j), k) = \alpha D(\Omega_i(j), k) + (1-\alpha)\frac{1}{T}\sum_{t=1}^{T} D(\Omega_i(j), t) \quad (1)$$

この手法はクエリが特定の講演に頻出するという仮定に基づくものであり，本研究の目指す「クエリは関連語と共起する」という情報を用いる手法とは異なるものである．とらえている観点が異なるため，提案手法との併用が有効と考える．このことは評価実験で示す．また，提案手法での関連語の選択において，特定の講演である語が頻出するという知見を応用する．

3. クエリの関連語を用いたリスコアリング方式

3.1 Word2vec

まず，本論文で用いた Word2vec について概説する．Word2vec [12], [13] とは，ニューラルネットワークを用いた単語の特徴ベクトル化，すなわち単語の分散表現を求める手法である．この分散表現は単語の概念を表す低次元の密なベクトルで表される．学習テキスト中の各単語を周辺の単語から予測するタスク（疑似的な単語予測のタスク）を設定し，テキストデータを用いてニューラルネットワークで学習する．中間層における各単語の特徴を表す低次元ベクトルがその単語の重みであり，これを抽出することによって，単語の概念を表すベクトルを獲得する．周辺の単語の重みベクトルの和を中間層の値とする（周辺単語から中心単語を推定する）モデルを Continuous Bag-of-Words（CBoW）モデルと呼び，周辺の単語のうちの 1 つに対する重みベクトルを中間層の値とする（中心単語から周辺単語を推定する）モデルを Skip-gram モデルと呼ぶ．いずれのモデルも，入力層と中間層をつなぐ重み行列，つまり各単語に対する重みベクトルの集合が最終的に生成する単語分散表現（Word Vector）となる．これにより，単語を意味的空間上の 1 点に対応させることができ，単語に対する意味的な計算が可能となる．本論文では処理時間を考慮し，学習が Skip-gram よりも高速な CBoW モデルを用いた．

3.2 アルゴリズム

次に，提案するリスコアリング方式について説明する．**図 1** にその処理図を示す．あらかじめ，検索対象の音声ドキュメントを音声認識システムを用いて単語認識し，得られた Triphone 系列を状態系列に変換する．クエリが与えられるとクエリを Triphone に変換した後状態系列にし，連続 DP（Dynamic Programming）照合を行うことで照合距離を求める．局所距離には状態間の音響距離 [5] を用いた．照合距離 $D(\Omega_i(j), k)$ が小さい順に候補として出力し，その結果を初期 STD 結果として保持する．この初期 STD 結果に提案方式を適用する．以下，図 1 の Step 1〜5 の処

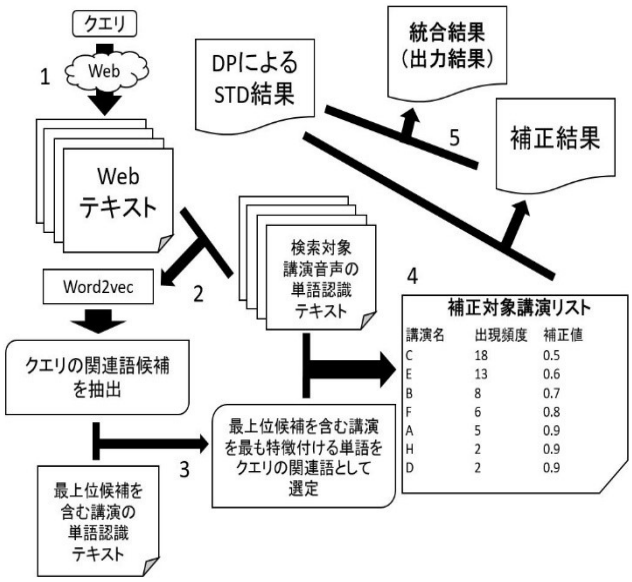

図 1 提案方式の全体図
Fig. 1 Overall view of the proposed method.

理手順について説明する．

Step 1：Web テキストの取得

まず，クエリで Web 検索し，その検索結果の上位 S 件分の Web テキストを取得する．この Web テキストで Word2vec を学習すること（Step 2）で，音声ドキュメントの単語認識テキストに出現しない未知語クエリの Word Vector を求めることができる．一方，S を大きくしすぎた場合や検索結果のリンク先の本文ページの文章まで取得した場合，処理に時間を要するため，本論文では，Web 検索結果中のタイトルとスニペットのみとし，$S = 100$ とした．

Step 2：Word2vec の学習

Word2vec の学習には検索対象の講演音声の単語認識テキストと Step 1 で取得した Web テキストを用い，両テキストに出現する単語の Word Vector を算出する．検索対象の講演音声の単語認識テキストを Web テキストとともに学習することで，クエリの意味的情報および OOV クエリの情報を補うことができ，より正確な関連度合いを学習できると考える．

Step 3：関連語の選定

Step 2 によりクエリと各単語の Word Vector を求め，クエリとの類似度を計算することで関連語が複数個得られる．一方，音声ドキュメントには出現せず Web テキストのみに出現する単語とクエリとの類似度が高くなるケースが考えられ，その場合，STD 精度を向上させるための関連語を適切に選定できないことが想定される．そのため，Word Vector を用いて求めた複数の関連語の中から，クエリの関連語として最もふさわしい単語を選定する必要がある．本論文では，選定する関連語は名詞に限定する．そこで，クエリと類似度の高い名詞を関連語と決定するのではなく，類似度の高い複数の名詞単語を上位 N（$= 100, 200, \ldots, 500$）個抽出し，クエリの関連語候補とする．文献 [14] で示されたように，検索結果における最上位候補は最も適合率が高く，最上位候補を含む講演はクエリを含んでいる可能性が高いため，そのクエリの関連語は講演中に出現している可能性が高い．たとえば，「トランプ大統領」というクエリであれば，当該講演中で「トランプ大統領」に関する内容が話されている可能性があり，「アメリカ」「政治」等の「トランプ大統領」の話題に関連した単語も話されている可能性がある．これらの単語はクエリ周辺の発話を特徴付ける単語と考えられる．そこで，最上位候補を含む講演を対象とし，抽出した N 個の関連語候補の中からクエリ周辺を特徴づける単語をクエリの関連語として選定する．具体的には，最上位候補を含む講演における複数の関連語候補に対しそれぞれの tf-idf 値を計算し，最も tf-idf 値の高い単語 1 個をクエリの関連語として選定する．複数個選定する場合も考えられるが，その検討は今後の課題とする．

Step 4：関連語を含む講演に対しリスコアリング

Step 3 で選定した関連語を含む講演はその関連語の出現頻度が高いほどクエリを含んでいる可能性が高い．これは，文献 [6] で示されたとおり，ある語は特定の講演で頻出する傾向があるためである．音声ドキュメントの単語認識結果に対し，関連語で完全一致に基づく文字列検索を行うことで，選定した関連語を含む講演を複数特定し，その特定した講演をリスコアリングの補正対象としリストに登録する．関連語の出現頻度が高いほどリスコアリング時の補正効果が大きくなるよう補正値を設定する．具体的には，リストに登録されている講演内のすべての候補に対して，以下の式 (2) により，照合距離が小さくなるように補正を行う．音声ドキュメントは講演音声ファイル（$\Omega = \Omega_1, \Omega_2, \Omega_3 \ldots$）で構成されており，$D(\Omega_i(j), k)$ はリスト内の講演 Ω_i の k 位の発話 $\Omega_i(j)$ の照合距離を表し，$\mathrm{new}\, D(\Omega_i(j), k)$ はリスコアリング後の照合距離を示す．$D(\Omega_i(j), k)$ に補正値 β（$0.5 \leq \beta \leq 0.9$）を乗じて補正する．

$$\mathrm{new}\, D(\Omega_i(j), k) = \beta \times D(\Omega_i(j), k) \tag{2}$$

この補正値の決め方は様々考えられるが，本論文では頻度順に 0.5，0.6，...，0.9 とし，頻度順位 5 番目以降の講演はすべて 0.9 とした．

Step 5：線形和統合

Step 4 では，関連語を含む講演を抽出したが，クエリを含まない講演が抽出されるケースが考えられる．その場合は正しく補正されず，検索精度が低下する．そこで，このように間違って補正されるケースを考慮し，リスコアリング結果の照合距離に対し，元の検索結果の照合距離と線形和統合することで，適切な照合距離となるよう調整する．統合は以下の式 (3) を用いて行う．γ（$0 \leq \gamma \leq 1$）は統合時の重み係数を表す．音声ドキュメントは講演音声ファイル（$\Omega = \Omega_1, \Omega_2, \Omega_3 \ldots$）で構成されており，統合後の照合

距離 new $D'(\Omega_i(j),k)$ は，Step 4 でリスコアリングした後の照合距離 new $D(\Omega_i(j),k)$ と元の照合距離 $D(\Omega_i(j),k)$ を線形結合することで求める．

$$\begin{aligned}&\text{new } D'(\Omega_i(j),k)\\&=\gamma\times\text{new } D(\Omega_i(j),k)+(1-\gamma)\times D(\Omega_i(j),k)\end{aligned}\tag{3}$$

4. 評価実験

4.1 実験条件

音響モデルと言語モデルの学習データには，CSJ [15] の学会講演と模擬講演をあわせた 2,702 講演から評価に用いる 177 講演を除いた 2,525 講演のうち，偶数講演（1,255 講演，約 287 時間）を使用した．音響モデルは 3 状態の Triphone で構成した．音声ドキュメントの認識には DNN-HMM を用いて単語単位で認識を行った．

DNN は Feedforward 型で構築し，各層を RBM として Pre-training を行った後に RBM を連結して Fine-tuning を行うことで学習した．DNN の学習に用いる音声特徴量は，40 次元の FBANK と Δ，ΔΔ の計 120 次元を用いた．音声特徴量の抽出条件は**表 1** のとおりである．この FBANK 120 次元を DNN の入力特徴量とし，中心フレームに前後 5 フレームを追加した 1,320 次元（11 フレーム × 120 次元）とした．これに合わせ，DNN の入力層のノード数を 1,320 とした．Kaldi [16] を用いて 3 状態の Triphone を作成し，状態数は今回 3,238 状態となった．DNN の出力はこの Triphone の 3,238 個の状態の事後確率とし，出力層のノード数を 3,238 とした．その他の学習条件は**表 2** に示すとおりである．

Web 検索エンジンは Google を使用した．Web 検索による Web テキストの取得は 2018 年 7 月 1 日に行った．Word Vector を用いた単語の特徴ベクトル化および類似度算出には，Python 用トピックモデリングライブラリの gensim [17] で実装されている Word2vec を用いた．Word2vec の学習パラメータは，ベクトル次元数：200，文脈窓長：5，単語の最低出現頻度：1，学習係数：0.05 とした．

検索（STD）には，CPU：Intel Core i7-980X，GPU：GeForce GTX 750 Ti，RAM：12 GB のマシンを使用した．

表 1 音声特徴量抽出条件
Table 1 Extraction conditions for speech features.

デジタル化	標本化周波数 16kHz 量子化 bit 数 16bit
特徴量	FBANK(40dim) + ΔFBANK(40dim) + ΔΔFBANK(40dim)
窓長	25 msec
フレームシフト	10 msec
窓関数	ハミング窓

表 2 DNN の学習条件
Table 2 Training conditions of DNN.

ノード数		入力層：1,320 隠れ層：2,048 出力層：3,238
隠れ層数		5
活性化関数		中間層：シグモイド関数 出力層：ソフトマックス関数
RBM	学習係数	0.004
	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
	エポック数	30

表 3 テストセット
Table 3 Test set.

	NTCIR-10	NTCIR-12
検索対象データ	SDPWS104 講演 （約 28.6 時間 40,746 発話）	SDPWS98 講演 （約 27.5 時間 37,782 発話）
クエリ	Formal Run : 100 種 (IV: 47, OOV: 53)	Formal Run : 113 種 (IV: 72, OOV: 41)

4.2 テストセット

評価には，**表 3** に示す NTCIR-10，NTCIR-12 で用いられた Formal Run テストセットを使用した．NTCIR-10 では音声ドキュメントワークショップの講演音声（SDPWS：Corpus of Spoken Document Processing Workshop）の 104 講演（約 28.6 時間，40,746 発話），NTCIR-12 では SDPWS の 98 講演（約 27.5 時間，37,782 発話）が検索対象音声ドキュメントとして用いられた．クエリには，NTCIR-10 Formal Run で使用された 100 クエリ，NTCIR-12 Formal Run（Single term）で使用された 113 クエリを用いた．正解情報は，NTCIR オーガナイザから提供されたものを用いた．パラメータ γ と N については，テストセット間での交差検証を行った．3.2 節のとおり，パラメータ β は $0.5 \leq \beta \leq 0.9$ の 0.1 刻みの値をとり，$S = 100$ とする．

4.3 評価指標

正解の判定は NTCIR 同様に発話単位で行い，クエリが発話内で 1 度以上話されていればその発話を正解とした．検索精度の評価には MAP（Mean Average Precision）を用いた．AP（Average Precision）は検索結果を上位から出力していき，正解が出力された時点での適合率を全正解

で平均したものである．各クエリで AP を求め，それらを全クエリで平均したものが MAP となる．AP，MAP はそれぞれ以下の式 (4)，(5) で求められる．クエリ $\boldsymbol{q}$ に対する正解発話数を $\boldsymbol{C_q}$，$\boldsymbol{M}$ は検索対象の総発話数，$\boldsymbol{\delta_i}$ はバイナリ関数で，検索結果の i 番目の発話が正解なら 1，不正解なら 0 となる．$precision(q,i)$ はクエリ q の i 番目の検索結果出力時点での適合率である．Q はクエリ数を表す．

$$AP(q) = \frac{1}{c_q} \sum_{i=1}^{M} \delta_i \times precision(q,i) \tag{4}$$

$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} AP(q) \tag{5}$$

4.4 提案方式の評価実験

パラメータ γ は 0.1 おきに，N は 100〜500 で変化させて実験を行った．結果を**図 2** と**図 3** に示す．図 2 は NTCIR-10，図 3 は NTCIR-12 のときの検索精度を示す．Baseline はリスコアリング方式適用前の結果を示す．それぞれのテストセットで以下のパラメータの組合せで最も検索精度が高くなった（Baseline からの向上値も示す）．

NTCIR-10：$\gamma = 0.5$，$N = 300$，
3.3 pt の向上（78.4% → 81.7%）

NTCIR-12：$\gamma = 0.7$，$N = 300$，
4.1 pt の向上（72.8% → 76.9%）

どちらのテストセットでも N は 300，γ は 0.5〜0.7 で高い精度が得られ，ほぼ同等のパラメータになったことから，本方式の頑健性を確認できた．

一方，NTCIR-12 ではすべての γ において，安定して検索精度が向上したが，NTCIR-10 では $\gamma = 1.0$（リスコアリング後）のとき Baseline から約 1.0〜2.0 pt 減少した．適切に関連語を選定できず，クエリを含まない講演を誤って補正するケースが多かったことが原因と考える．

Web テキストの容量は 1 クエリあたり平均で 0.03 MB だった．Web 検索結果の本文ページも取得すれば容量は増加するが，その分処理に時間を要する．本文ページまで取得した際の処理時間の変動についての調査は今後の課題とする．

検索精度が変化したクエリの内訳を**表 4** に示す．計 213 クエリ中 Baseline で AP が 100%で本方式適用後も 100%のクエリ（表 4 の 100% → 100%）を除いた 145 クエリのうち，74 クエリは検索精度が向上（うち OOV は 35 クエリ），50 クエリは低下（うち OOV は 29 クエリ），21 クエリは変化がなかった（うち OOV は 14 クエリ）．クエリごとに結果を考察すると，AP が向上したクエリで，たとえば「アーティキュレーション」（OOV）の AP は 21.4 pt 向上（67.7% → 89.1%）し，選定された関連語は「発音」であった．「アーティキュレーション」を含む講演で「発音」が複数出現しており，さらに Web テキスト中でその意味

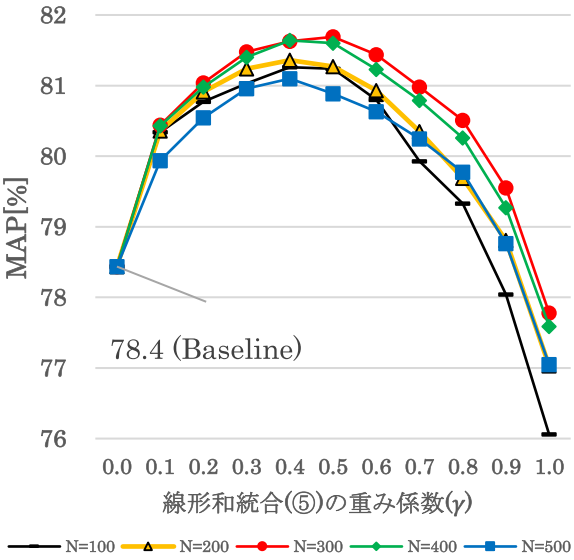

図 2 NTCIR-10 に提案方式を適用した結果

Fig. 2 The result of applying the proposed method to NTCIR-10.

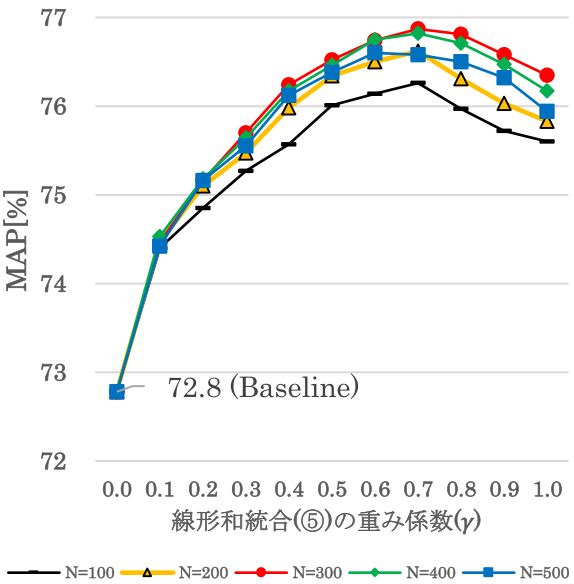

図 3 NTCIR-12 に提案方式を適用した結果

Fig. 3 The result of applying the proposed method to NTCIR-12.

を解説する記事が多かった．このため，これらの文章から単語的意味を学習できたと考える．「おはようございます」（IV）は AP が 45.0 pt 向上（50.0% → 95.0%）し，選定された関連語は「本日」であった．「おはようございます」を含む講演では冒頭の挨拶で「えーおはようございます，本日の…」のように「本日」と共起して出現している箇所が多く，Web テキスト中も同様に「えーおはようございます，本日のパーソナリティは…」等のように書かれている記事が多かったため，これらの文章から正しく学習できたと考える．AP が低下したクエリで，たとえば「API」（OOV）の AP は 16.6 pt 低下（36.1% → 19.5%）し，選定された関連語は「仕様」であった．Web テキスト中では API の仕様

表 4 提案方式により検索精度が変化したクエリ数
Table 4 Number of queries whose retrieval accuracy has changed by the proposed method.

	NTCIR-10			NTCIR-12		
	IV	OOV	計	IV	OOV	計
向上	20	17	37	19	18	37
低下	8	19	27	13	10	23
変化なし	1	6	7	6	8	14
100%→100%	18	11	29	34	5	39
合計	47	53	100	72	41	113
100%→100%を除いた合計	29	42	71	38	36	74

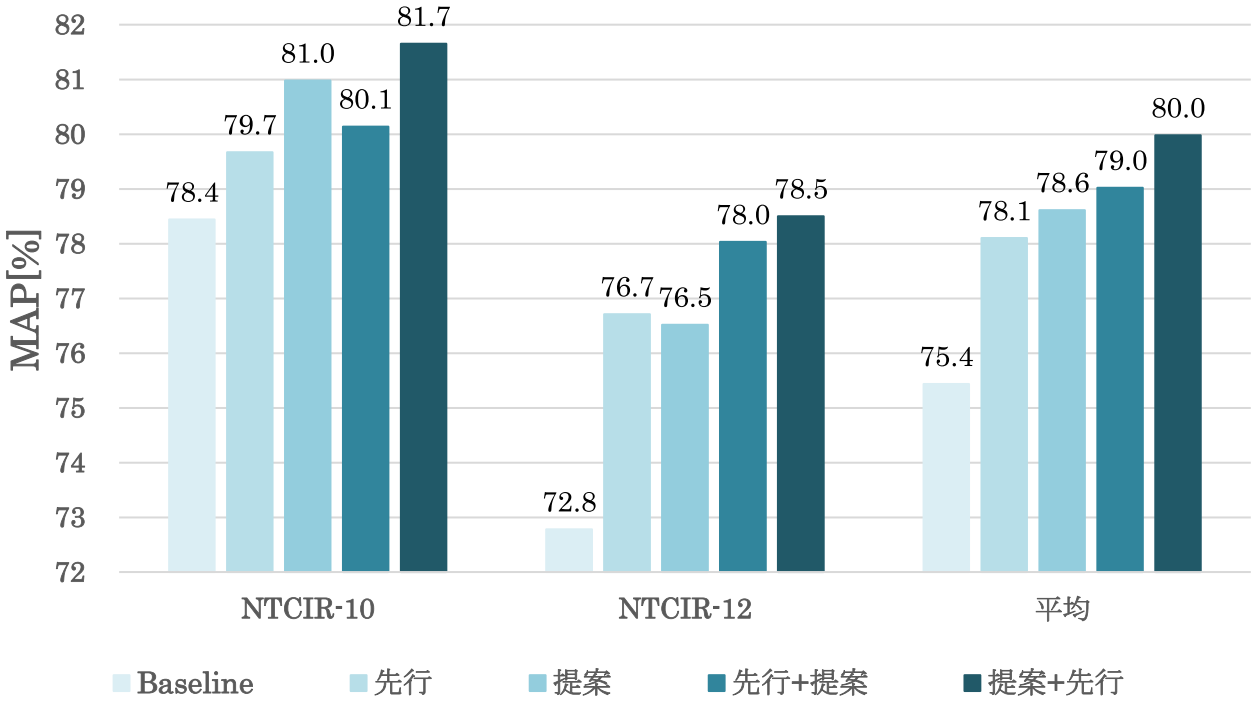

図 4 各方式の検索精度
Fig. 4 Retrieval accuracy of each method.

について書かれている記事が多かったが，「API」を含む講演には「仕様」がまったく出現していなかったため，正しく学習できなかったと考える．「アメリカ」（IV）は AP が 3.8 pt 低下（100.00% → 96.2%）し，選定された関連語は「セント」であった．これは，中国企業のテンセントの音楽配信子会社がアメリカに上場した記事が Web テキスト中に含まれており，その「テンセント」が形態素解析で「テン」と「セント」に分かち書きされ，「セント」が最上位候補を含む講演で tf-idf 値が高くなったことが原因と考える．

検索精度が向上，低下するクエリの共通点や法則性は現段階では確認できないため今後の課題とする．検索精度に変化がなかったクエリのなかで「キタチャンキタロボ」（OOV）は，Web テキスト中に出現せず，Word2vec で学習できなかったため本方式が適用できなかった．

4.5 先行研究との比較・併用実験

提案方式（3 章）と高順位リランキング方式（2 章），およびそれらを併用した方式との比較を行った．その結果を図 4 に示す．併用における+は適用順を示す．MAP は NTCIR-10 と NTCIR-12 のテストセット間で交差検証により求めた．

Baseline と各方式単体の検索精度を比較すると，先行方式，提案方式それぞれ以下のように MAP が向上した（括弧内は Baseline との比較を示す）．

NTCIR-10：先行 1.3 pt（78.4% → 79.7%），
提案 2.6pt（78.4% → 81.0%）
NTCIR-12：先行 3.9 pt（72.8% → 76.7%），
提案 3.7pt（72.8% → 76.5%）
平均：先行 2.7 pt（75.4% → 78.1%），
提案 3.2 pt（75.4% → 78.6%）

提案方式は先行方式と同じように高い STD の効果があり，検索精度の向上を実現した．

Baseline と比べ両テストセットの平均で，先行+提案で

表 5 補正対象講演リストの評価
Table 5 Evaluation of the list of lectures to be rescored.

	N = 100	N = 200	N = 300	N = 400	N = 500
適合率(%)	36.33	35.36	34.01	33.32	33.87
再現率(%)	75.85	73.15	72.12	71.88	71.99
F 値(%)	39.85	38.33	36.70	35.93	36.27
登録講演数	1.50	1.54	1.57	1.57	1.60

3.6 pt，提案+先行で 4.6 pt の向上となった．平均でも統合の効果が見られており，それぞれを単独で適用するよりも高い精度となった．先行+提案での改善が小さめなのは，交差検証においてテストセット間でのパラメータが大きく異なったためと考えられる．実際に先行+提案において最も検索精度が高くなったパラメータは，NTCIR-10 で $\gamma = 0.1$，$N = 400$，NTCIR-12 で $\gamma = 0.5$，$N = 300$ で，パラメータ γ に差があった．3.4 節で述べたとおり，NTCIR-10 では適切に関連語を選定できず，クエリを含まない講演を誤って補正するケースが多かったため，併用による大きな補正効果が得られず，γ は低い値に収束したと考えるが，詳細については今後の課題とする．

4.6 提案方式の処理時間計測

本方式は処理手順が多く，検索に時間を要すると想定される．3.2 節で示した Step1 から Step5 まで，すなわちクエリで Web 検索を行ってから統合結果を得るまでの一連の処理すべてを行ったときの時間を計測した（DP による初期 STD の検索時間は除く）．その結果を**図 5** に示す．図に示す処理時間は NTCIR-10 と NTCIR-12 の計 213 クエリの（1 クエリあたりの）平均の処理時間である．MAP は $\gamma = 1.0$ のときの NTCIR-10 と NTCIR-12 の平均を示す．N を 100 増やすごとに平均で 0.73 秒増加している．最も検索精度の高い $N = 300$ のときで 4.63 秒，そのうちの 2.18 秒は Step3 における関連語を選定するための tf-idf の計算時間であった．また，初期 STD の処理時間は 0.18 秒であった．これと合わせると，$N = 300$ のとき 4.81 秒となる．これは，検索時間としては長いため，処理時間の短縮が求められる．tf-idf の処理をせずに関連語を選定すればさらに処理速度が速くなるが，その具体的な手法の検討と検索精度への影響については今後の課題とする．また，Web 検索や Word2vec の学習に時間を要すると当初は想定していたが，Step1 の Web 検索を行ってから Step3 の関連語候補を抽出するまでの時間は $N = 300$ のとき NTCIR-10 と NTCIR-12 の平均で 2.07 秒であった．これは Web 検索ではタイトルとスニペットのみの取得に抑え，4.4 節のとおり 1 クエリあたりの Web テキストの容量は NTCIR-10 と NTCIR-12 の平均で 0.03 MB と非常に少量なサイズだったため，Web 検索処理は短時間に収まったと考える．また SDPWS104 講演および SDPWS98 講演の単語認識テキストの容量がともに約 2.5 MB と，これらもまた少量なサイズなため，検索対象の講演音声の単語認識テキストと Web テキストをともに Word2vec に用いても少ない学習量に収まり，Word2vec の学習においても処理が短時間に収まったと考える．

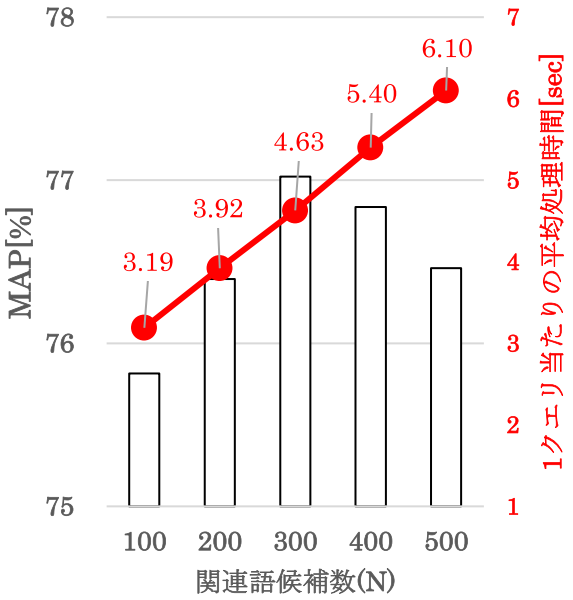

図 5 提案方式の処理時間
Fig. 5 Processing time of the proposed method.

4.7 補正対象講演リストの評価

提案方式では，選定した関連語を含む講演を複数特定し，その特定した講演をリスコアリングの補正対象としてリストに登録する処理を行った．そこで，補正対象講演リストに正解発話を含む講演が正しく登録されているか評価を行った．具体的には，補正対象講演リストの適合率，再現率，F 値，登録講演数を算出した．結果は**表 5** のとおりである．表の数値は NTCIR-10 と NTCIR-12 の計 213 クエリの（1 クエリあたりの）平均値である．

提案方式による検索精度向上結果（図 2，3）に反して，適合率，再現率，F 値すべて N を 100 増やすごとに数値が低下していき，$N = 400$ から 500 に増やしたときにわずかに向上した．登録講演数は N を 100 増やすごとにわずかに増えているが，数値が 1.50 から 1.60 に収まっていることから，リストには 1 クエリあたりおよそ 1，2 個講演が登録されていると考えられる．図 2，3 から考えると，適

合率，再現率は $N = 300$ のときに最良の結果になると推察できるが，この原因の調査は今後の課題とする．全体的に，適合率は 30%台，再現率は 70%台となっており，適合率の方が低い結果となっている．クエリを含む正解の講演は半分以上登録できている一方，クエリを含まない不正解の講演も半分以上登録されたと考えられる．関連語の選定方法を改善すればリストも改善されると考える．

4.8 再実験による関連語および検索精度の考察

本方式では関連語を取得するために Web 検索を併用する方式を採用した．Web の検索結果はつねに変動する可能性があるため，同じクエリで検索時期が異なれば関連語が変わり検索精度も変動すると考えられる．そこで前回の 2018 年 7 月 1 日に Web 検索での実験に加え，2019 年 7 月 23 日に Web 検索して再実験を行った．テストセットは NTCIR-10 を用いた．実験条件は 4.1 節と同様であり，関連語候補数 N は 4.4 節で最も検索精度が高くなった 300 とした．前回の検索精度は $\gamma = 0.5$ で Baseline から最大で 3.3 pt の向上（78.4% → 81.7%）となったが，今回は $\gamma = 0.4$ で Baseline から最大で 2.8 pt の向上（78.4% → 81.2%）となった．1 年後の Web 検索して得た情報を用いても検索精度は向上し，最適な γ の値に大きな違いはなかった．関連語を確認してみると，今回 100 クエリ中 63 クエリは前回と同じ関連語となった．

4.9 今後の課題

本方式では Web テキストと検索対象の講演音声の単語認識テキストの Word2vec 学習を行った．一方，検索対象のデータ量がさらに大きい場合を考えると，学習に時間がかかり検索に時間をさらに要すると想定される．Web テキストの取得件数 S を大きくしすぎた場合や検索結果のリンク先の本文ページの文章まで取得した場合等も同様である．今後は Word2vec の学習テキスト量に依存せず高速に処理する方式の検討を行いたい．

現状では取得する関連語候補を名詞のみに限定しており，他の品詞は考慮していない．動詞，形容詞等の単語も関連語候補として取得すればより正確な関連語の選定が行えると考える．今後は名詞以外の品詞単語も関連語候補として取得する場合の検索精度の変化を調査していきたい．

本論文では選定する関連語の数を 1 個としたが，1 個だけでなく複数個選定すれば関連語情報はさらに正確になり，より適切な補正対象講演を決定できると考える．今後は本方式において選定する関連語の適切な個数について調査を行いたい．

現状ではリスコアリング時の補正値 β を関連語の頻度順に 0.5，0.6，...，0.9 と手動で決めており，自動的な決め方ではない．補正する講演によって補正値を自動で決定することができればより適した照合距離になると考えられる．また，この補正値 β の与え方次第で検索精度が大きく変動すると考えられる．そのため，今後は関連語の頻度だけでなく様々な情報を用いて補正値 β を適切な値に自動で決定する方法を検討していきたい．

4.4 節で述べたとおり，関連語候補数 N が 300 のときに最も検索精度が高くなった．一方，N を 400 以上にすると検索精度が低下する．関連語候補数 N を大きくしすぎるとクエリと関連性のない単語が関連語候補となり，その単語が誤って関連語として選定されることが要因として考えられるが，詳細については今後調査していきたい．

5. おわりに

本論文では STD において Word2vec による単語の分散表現および Web 検索を用いることでクエリの関連語を選定し，その関連語情報を用いたリスコアリング方式を提案した．Web テキストおよび検索対象の単語認識結果を Word2vec で学習した結果からクエリの関連語候補を複数保持し，その関連語候補のうち最上位候補を含む講演を最も特徴付ける単語をクエリの関連語として選定し，その関連語を含むドキュメント内のすべての候補の距離を有利にすることで検索精度の向上を図った．

講演音声を対象とした NTCIR-10，12 の Formal Run の 2 種のテストセットを用いて評価した結果，両テストセットで検索精度が平均 3.2 pt 向上し，本方式の有効性を確認した．さらに提案方式を適用後に先行方式を適用することで平均 1.4 pt，トータル 4.6 pt の向上が得られ，提案方式と先行方式を併用することの有効性も確認できた．

提案方式のパラメータについては，2 つのテストセットでどちらでも関連語候補数 N は 300，線形和統合の重み係数 γ は 0.5〜0.7 付近のときに高い精度が得られ，ほぼ同等の値となった．一方，先行研究と併用時のパラメータの決め方に検討の余地があることが分かった．Web での検索時期が異なっても提案方式の効果が見られ，Web 検索に対する頑健性を確認した．

4.9 節でも述べたとおり，補正値 β の与え方次第で検索精度が大きく変動すると考えられる．今後は補正値 β を適切な値に自動で決める方法の検討について取り組む予定である．

謝辞 本研究の一部は JSPS 科研費 18K11358 の助成を受けたものです．

参考文献

[1] Akiba, T., Nishizaki, H., Aizawa, K., Kawahara, T. and Matsui, T.: Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop, *NTCIR-9 Workshop Meeting*, pp.223–235 (2011).

[2] Akiba, T., Nishizaki, H., Aikawa, K., Hu, X., Itoh, Y., Kawahara, T., Nakagawa, S., Nanjo, H. and Yamashita, Y.: Overview of the NTCIR-10 SpokenDoc-2 Task,

NTCIR-10 Workshop Meeting, pp.573–587 (2013).
[3] Akiba, T., Nishizaki, H., Nanjo, H. and Jones, G.I.F.: Overview of NTCIR-11 Spoken&Doc Task, *NTCIR-11*, pp.350–364 (2014).
[4] Akiba, T., Nishizaki, H., Nanjo, H. and Jones, G.J.F.: Overview of NTCIR-12 Spoken&Doc Task, *NTCIR-12*, pp.167–179 (2016).
[5] 岩田耕平，伊藤慶明，小嶋和徳，石亀昌明，田中和世，李 時旭：語彙フリー音声文書検索方式における新しいサブワードモデルとサブワード音響間距離の有効性の検証，情報処理学会論文誌，Vol.48, No.5, pp.1990–2000 (2007).
[6] 小嶋和徳，紺野和磨，田中和世，李 時旭，伊藤慶明：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スコアリング方式，電子情報通信学会 D，Vo1.J100-D, No.1, pp.70–80 (2017).
[7] 清水嘉乃，李 時旭，小嶋和徳，伊藤慶明：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，情報処理学会第 80 回全国大会，5Q-08, pp.2-393–394 (2018).
[8] 小田原一成，山下洋一：音声中の検索語検出における単語共起情報の利用，情報処理学会研究報告，2016-SLP-110, pp.1–6 (2016).
[9] 南條浩輝，川口達也：検索語の説明文による音声内容検索を利用した音声検索語検出，情報処理学会研究報告，Vol.2017-SLP-115, pp.1–6 (2017).
[10] 南條浩輝，前田 翔，吉見毅彦：音声検索語検出のための検索語拡張法，情報処理学会論文誌，Vol.58, No.10, pp.1735–1744 (2017).
[11] 川崎 祥，秋葉友良：単語共起を用いた偽陽性誤りに頑健な音声ドキュメントの検索モデル，日本音響学会春季研究発表会，3-Q5-5，pp.193–196 (2014).
[12] Mikolov, T., Sutskever, I., Ghen, K., Corrado, G. and Dean, J.: Efficient Estimation of Words and Phrases and their Compositionally, *Advances in Neaural Information Processing Systems 26*, pp.3111–3119 (2013).
[13] Mikolov, T., Ghen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *Processing of the International Conference on Learning Representations* (*ICLR*), pp.1–12 (2013).
[14] 丹治 遥，小嶋和徳，李 時旭，南條浩輝，伊藤慶明：音声中の検索語検出における最上位候補を含む講演及びその類似講演優先方式，日本音響学会春季研究発表会，2-Q-17, pp.185–186 (2018).
[15] National Institute for Japanese Language and Linguistics: Corpus of Spontaneous Japanese, available from ⟨http://pj.ninjal.ac.jp/corpus_center/csj/⟩.
[16] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Goel, O.N., Hannemann, M., Motlicek, P., Oian, Y., Schwarz, P., Silovsky, J., Stemmer, G. and Vesely, K.: The Kaldi Speech Recognition Toolkit, ASRU (2011).
[17] gensim topic modeling for humans, available from ⟨https://radimrehurek.com/gensim/index.html⟩.

丹治 遥

平成 30 年岩手県立大学ソフトウェア情報学部卒業．同年同大学大学院修士課程入学，現在に至る．

小嶋 和徳 （正会員）

平成 7 年秋田大学大学院鉱山学研究科電子工学専攻修士課程修了．平成 10 年岩手県立大学ソフトウェア情報学部助手．平成 20 年同大学講師．博士（工学）．電子情報通信学会，人工知能学会各会員．

李 時旭 （正会員）

平成 9 年韓国嶺南大学 M.Sc.（音声認識）．平成 13 年東京大学大学院工学系研究科情報通信工学専攻博士課程修了（工学博士）．同年産業技術総合研究所入所．現在，同研究所情報技術研究部門研究員．日本音響学会，韓国音響学会各会員．

南條 浩輝 （正会員）

平成 10 年京都大学工学部情報学科卒業．平成 13 年同大学大学院情報学研究科修士課程修了．平成 16 年同大学院情報学研究科博士後期課程修了．同年龍谷大学理工学部助手．平成 19 年同助教．平成 27 年 8 月より京都大学学術情報メディアセンター准教授．音声認識・理解，音声ドキュメント処理の研究に従事．電子情報通信学会，日本音響学会，日本バーチャルリアリティ学会，外国語教育メディア学会，IEEE 各会員．平成 21 年日本音響学会粟屋潔学術奨励賞受賞．

伊藤 慶明 （正会員）

平成元年東京大学大学院工学系研究科航空学専攻修士課程修了．同年川崎製鉄（株）に入社．平成 4 年技術研究組合新情報処理開発機構に出向．平成 12 年岩手県立大学准教授．平成 25 年より同教授．博士（工学）．人工知能学会，日本音響学会，電子情報通信学会，IEEE 各会員．