情報処理学会論文誌 Vol.54 No.7 1967–1977 (July 2013)

音声入力型情報検索における ベイズリスク最小化音声認識のための 単語重要度の自動推定

古谷 遼[1,a)] 七里 崇[1] 南條 浩輝[2,b)]

受付日 2012年11月30日, 採録日 2013年4月5日

概要：音声入力型の情報検索のためのベイズリスク最小化音声認識の研究を行う．音声入力型情報検索では，ベイズリスク最小化音声認識は検索に影響が大きい単語を重要視し，検索に致命的な音声認識誤りを避けることを目的として行われる．その際，検索への影響が大きい単語にその影響度を反映した重要度を設定することが重要であるものの，そのような重要度の自動決定手法はこれまでに存在しなかった．この問題に対し，本論文では，音声入力型情報検索における音声認識のための単語重要度の自動推定手法を提案する．具体的には，検索要求のテキストとその音声認識結果および検索の正解ラベルの 3 点を推定のための学習データとし，単語重要度を推定する手法を提案する．重要度推定用のデータについて，人手による準備を必要とする方法（教師あり推定），一部を必要としない方法（半教師あり推定），および必要としない方法（教師なし推定）のそれぞれを提案し，複数の検索システムを用いてそれらの有効性を調べた．教師ありおよび半教師あり推定の結果から提案する推定の枠組みが正しく動作をすることを示した．さらに教師なし推定の結果から，提案手法を用いることで音声入力型情報検索の音声認識にとって有効な重要度を決定できることを示した．

キーワード：ベイズリスク最小化音声認識，情報検索，単語重要度，自動推定

Automatic Estimation of Word Importance for Minimum Bayes-risk Decoder in Spoken Query-based Information Retrieval

Ryo Furutani[1,a)] Takashi Shichiri[1] Hiroaki Nanjo[2,b)]

Received: November 30, 2012, Accepted: April 5, 2013

Abstract: Minimum Bayes-risk (MBR) based automatic speech recognition (ASR) oriented for spoken query-based information retrieval (IR) is addressed. In a spoken query-based IR system, MBR decoding (ASR) is taken aiming to reduce fatal ASR errors on IR. For such ASR, although an importance should be assigned to each word according to its influence on IR, no automatic estimation methods of such importance are proposed. In this paper, we propose an automatic estimation of word importance, which requires 1) text queries, 2) corresponding spoken queries and their ASR results, and 3) list of documents to be retrieved (correct documents for each query). Three kinds of estimation methods are proposed; supervised, semi-supervised, and unsupervised methods, and they are evaluated with several IR systems. We confirmed that our method is reasonable from results of the supervised and the semi-supervised estimations, and confirmed that the unsupervised method can determine appropriate word importance for ASR of a spoken query-based IR.

Keywords: minimum Bayes-risk decoding, information retrieval, word importance, automatic estimation

[1] 龍谷大学大学院理工学研究科
Graduate School of Science and Technology, Ryukoku University, Otsu, Shiga 520–2194, Japan
[2] 龍谷大学理工学部
Faculty of Science and Technology, Ryukoku University, Otsu, Shiga 520–2194, Japan
[a)] furutani@nlp.i.ryukoku.ac.jp
[b)] nanjo@nlp.i.ryukoku.ac.jp

1. はじめに

音声認識をフロントエンドに持つ情報検索，すなわち音声入力型の情報検索のための音声認識について研究を行う [1], [2]．音声入力型の情報検索システムにおいては，バックエンドの情報検索システムが高性能であっても，音

声認識で認識誤りが発生した場合，その影響を受けて検索性能が低下する．この問題に対し，現在提案されている手法の 1 つに，音声認識と情報検索を一体としてシステム化する方法がある [3], [4]．しかし，情報検索および音声認識システムに関する深い知識がない場合には，このような手法は採用できない．さらに，既存のテキストベースの情報検索システムを変更せずに，フロントエンドに音声認識を加えたい場合には，上記の手法は採用できない．これらの背景から，音声認識誤りによる影響を受けにくい音声入力型の情報検索を実現するための方式が求められている．

情報検索には，検索の観点から重要な語句（たとえばキーワード）とそうでない語句が存在する．このため，情報検索のフロントエンドとしての音声認識ではすべての語句を同等に扱うことは適切でなく，重要語句を優先的に扱い，それらの音声認識誤りを少なくすることが重要である．音声認識の評価は重要語句がどの程度認識されたかという観点で行う必要がある．このような研究の 1 つに，情報検索のキーワードの音声認識誤り，すなわちキーワード誤り率に着目して情報検索の性能向上を狙う方法がある [5]．この方法は，あらかじめ情報検索のキーワード集合を定義でき，かつ各キーワード間には差がないタスクには十分である．しかし，キーワード間に差がある場合，たとえばキーワードの重みを用いるベクトル空間モデルに基づく検索システムなどでは，すべてのキーワードを同等に扱うことが適切とは限らず，誤ると影響が大きいキーワードを優先的に音声認識することが重要である．これらの問題に対して，我々は重み付き単語誤り率（WWER: Weighted Word Error Rate）を提案している [6]．これは，情報検索に対する情報損失の観点から各語句に異なる重要度を与え，重要度に基づき音声認識の評価を行う尺度である．我々はこれまでに，ベイズリスク最小化（MBR: Minimum Bayes-Risk）の枠組みに基づいて WWER を最小化する音声認識を行うことが情報検索システムの性能向上に効果的であることを示している [6]．

MBR 音声認識をフロントエンドに持つ情報検索において音声認識誤りの影響を少なくするためには，音声認識時に各単語に対して適切に重要度を与えることが重要である．適切な単語重要度とは，それらから算出される WWER と音声認識誤りによる検索性能の低下の間に高い相関が得られるような重要度である．両者の高い相関は，前段の MBR 音声認識において WWER 最小の仮説を探索することが後段の検索システムでの性能向上につながりやすいことを示す．しかし，そのような単語重要度を人手で設定することは容易ではない．後段の情報検索システムの内部が既知であれば，情報検索における単語重要度を音声認識時に用いることも考えられる．従来は，後段の情報検索システムを熟知した開発者がこのような重要度を経験的に決定していた [6] ものの，この単語重要度が音声認識において適切である保証はない．後段の情報検索システムの内部が未知であれば，そもそもこのアプローチを用いることができない．これらのことより，単語重要度の自動決定手法が望まれている．

このような背景に基づき，本論文では情報検索を目的とした MBR 音声認識のための単語重要度の自動推定手法を提案する．具体的には，音声認識誤りによる検索性能の低下率と WWER が等しくなるように単語重要度を決定する手法を提案する．提案手法は情報検索システムの入出力にのみ着目する手法であり，本手法により，情報検索システムについての知識を持たなくても適切な単語重要度の決定が可能となる．このような研究はこれまでには行われておらず新しい．

本論文の構成について述べる．2 章で WWER および WWER 最小化音声認識について述べる．3 章で単語重要度を自動で推定する手法について述べる．また，そのための学習データの自動生成についても述べる．4 章で単語重要度の自動推定についての実験を行い，本手法で適切な単語重要度が推定できることを示す．5 章で結論を述べる．

2. 単語重要度に基づく音声認識の枠組み

2.1 重み付き単語誤り率

重み付き単語誤り率（WWER）は式 (1) で定義される [6]．

$$\mathrm{WWER} = \frac{V_I + V_D + V_S}{V_N} \tag{1}$$

V_I は挿入誤り単語の重要度の合計を，V_D は削除誤り単語の重要度の合計を，V_S は置換誤り区間の単語重要度の合計を，V_N は正解文の単語重要度の合計を表す．なお，誤り単語を同定する際には，単語誤り率（WER）を求める際と同様に DP マッチングの結果を用いるため，すべての単語の重みを等しく設定したとき，WWER は WER と一致する．

WER と WWER の比較例を**図 1** に示す．それぞれの認識文で下線が付された単語は音声認識誤り単語を表す．WER で評価を行った場合は認識文 2 が良い候補と評価される．ここで，各語の情報検索に与える影響の大きさについて，“オーロラ” は 10，“発生” は 3，“条件” は 2，その他の単語は 1 であることが分かっていたとする．この場合，認識文 2 は “オーロラ” という重要度の高い単語が誤っているため，情報検索への影響が大きい．しかし，このような重大な誤りを WER では正しく評価することができな

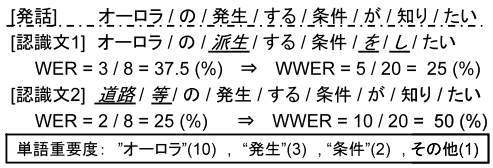

図 1 WER と WWER の比較例
Fig. 1 Comparison of WER and WWER.

い．一方，誤り単語の重要度に着目する WWER で評価を行った場合は，“オーロラ” という単語が正しく認識されている認識文 1 を良い候補であると評価できる．このように情報検索の観点から音声認識結果を評価する際には，情報検索への影響を反映した単語重要度に基づく WWER が適切であることが分かる．

2.2 ベイズリスク最小化音声認識

WWER の定義が定まったときに WWER を最小化する音声認識は，式 (2) で表されるベイズリスク最小化（MBR）の枠組みで定式化できる [6]．

$$\hat{W} = \arg\min_{W} \sum_{W'} l(W, W')^{\lambda_1} P(W', X)^{\lambda_2} \tag{2}$$

ここで，$l(W, W')$ は仮説 W' を仮説 W に誤った際の損失を求める損失関数を表し，$P(W', X)$ は入力信号 X と W' の同時確率（音声認識スコア）を表す．λ_1，λ_2 は損失関数および確率の重みパラメータである．WER 最小化を目的とした場合は，損失関数として WER，もしくは WER 定義式の分子に相当する最小編集距離（Levenshtein Distance）を用いればよいことが知られている [7], [8]．同様に，重要度の高い単語の誤り最小化，すなわち WWER 最小化を目的とした場合は，損失関数として式 (1) で定式化される WWER，もしくは WWER の分子を用いればよいことが示されている [6]．

これらのことは，情報検索のための音声認識には，適切な単語の重要度の決定が課題であることを表している．本論文では，このような重要度の推定手法を述べる．

3. 単語重要度の自動推定

3.1 適切な単語重要度

単語の重要度は，その語の音声認識誤りが情報検索の性能に与える影響の程度に基づいて設定することが重要である．具体的には，WWER と検索性能の低下が等しくなるように単語の重要度を設定することが重要である．ここで我々は，情報検索の性能低下の評価尺度のために検索性能低下率（IRDR: Information Retrieval performance Degradation Ratio）を定義する（式 (3)）．

$$\mathrm{IRDR} = 1 - \frac{H}{R} \tag{3}$$

R と H はそれぞれ，書き起こしと音声認識結果の検索要求を用いた際の検索性能を表す．したがって，IRDR は音声認識誤りが情報検索に与える影響の割合を表す．この IRDR を WWER から予測できれば，WWER 最小化音声認識により検索性能の向上が得られると考えられる．このような WWER を定義できる単語重要度が適切である．

3.2 単語重要度の自動推定

WWER から IRDR を予測できるような単語重要度を人手で設定することは困難である．これに対し，本論文では，想定される検索要求の発話データ，その書き起こし，および検索の正解集合の 3 つのデータを用いて，単語重要度を自動推定する手法を提案する．これにより，既存のテキスト入力型情報検索システムを容易に音声入力型情報検索システムに拡張することができる．単語重要度の自動推定の概要を**図 2** に示す．具体的な手順を以下に示す．

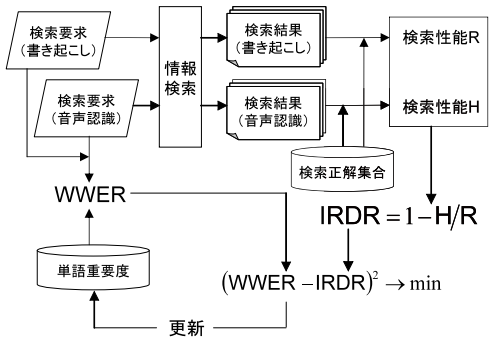

図 2 単語重要度の自動推定の概要
Fig. 2 Overview of word importance estimation.

(1) 学習データとして検索要求の音声認識結果 W'_m と誤りのない書き起こし W_m のペアを N 件用意し，各 m（$m = 1, \ldots, N$）に対して手順 (2)〜(6) を行う．
(2) W_m の検索の正解データを用意する．
(3) 検索の正解データを用いて，W_m で検索を行ったときの検索性能 R_m を得る．
(4) 検索の正解データを用いて，W'_m で検索を行ったときの検索性能 H_m を得る．
(5) R_m と H_m から，式 (3) に基づき IRDR_m を求める．
(6) W_m と W'_m から，W'_m に含まれるすべての誤り単語とそれらの頻度，および W_m に含まれるすべての単語とそれらの頻度を求める．
(7) $\mathrm{IRDR}_m > 0$ となる各 m について IRDR_m と $\mathrm{WWER}_m(\boldsymbol{x})$ が等しくなるように各単語の重要度 x_k を推定する．ここで，$\boldsymbol{x}$ は単語の重要度から構成されるベクトルであり，x_k は k 番目の単語の重要度を表す．

実際には，すべての m について WWER と IRDR を完全に一致させることは不可能である．そのため，単語重要度推定の枠組みにおける手順 (7) は，WWER と IRDR との平均二乗誤差 $F(\boldsymbol{x})$ を最小化する問題として定義する（式 (4)）．

$$\begin{aligned} F(\boldsymbol{x}) &= \sum_m \mathrm{ERR}_m(\boldsymbol{x})^2 \\ &= \sum_m \left(\mathrm{WWER}_m(\boldsymbol{x}) - \mathrm{IRDR}_m\right)^2 \\ &= \sum_m \left(\frac{E_m(\boldsymbol{x})}{C_m(\boldsymbol{x})} - \mathrm{IRDR}_m\right)^2 \end{aligned} \tag{4}$$

ここで，m は手順 (1) で用意した検索要求の ID である．

$E_m(\boldsymbol{x})$ は式 (1) の分子にあたり，音声認識誤りした単語の重要度の合計を表す．$C_m(\boldsymbol{x})$ は式 (1) の分母にあたり，正解文の単語の重要度の合計を表す．$E_m(\boldsymbol{x})$ および $C_m(\boldsymbol{x})$ はそれぞれ式 (5) で定義される．

$$E_m(\boldsymbol{x}) = \sum_{k=1}^{K} e_{m,k} x_k, \quad C_m(\boldsymbol{x}) = \sum_{k=1}^{K} c_{m,k} x_k \tag{5}$$

$e_{m,k}$ は k 番目の単語が検索要求 W'_m において誤った回数を表し，$c_{m,k}$ は k 番目の単語が検索要求 W_m（正解文）に出現した回数を表す．この $e_{m,k}$ および $c_{m,k}$ は手順 (6) で求められるものである．

次に，各単語の重要度の最適値を求める方法について述べる．本研究では，$F(\boldsymbol{x})$ を最小化するために最急降下法を用いる．具体的には，初期値としてすべての単語の重要度を適当な正の値（典型的には 1）とし，各単語の重要度 x_k を WWER と IRDR の平均二乗誤差が収束するまで，もしくは繰返し回数が一定数に達するまで繰り返し更新する．x_k の更新は式 (6) に基づいて行う．

$$x_k^{(t+1)} = \begin{cases} x_k^{(t)} - \alpha \cdot \dfrac{\partial F}{\partial x_k^{(t)}} & \text{if } \beta_{min} \le x_k^{(t+1)} \le \beta_{max} \\ x_k^{(t)} & \text{otherwise} \end{cases} \tag{6}$$

なお，

$$\frac{\partial F}{\partial x_k^{(t)}} = 2 \sum_m \mathrm{ERR}_m\left(\boldsymbol{x}^{(t)}\right) \cdot \frac{\partial \mathrm{ERR}_m}{\partial x_k^{(t)}} \tag{7}$$

$$\begin{aligned} \frac{\partial \mathrm{ERR}_m}{\partial x_k^{(t)}} &= \left(\frac{E_m\left(\boldsymbol{x}^{(t)}\right)}{C_m(\boldsymbol{x}^{(t)})} - \mathrm{IRDR}_m \right)' \\ &= \frac{E'_m\left(\boldsymbol{x}^{(t)}\right) \cdot C_m\left(\boldsymbol{x}^{(t)}\right) - E_m\left(\boldsymbol{x}^{(t)}\right) \cdot C'_m\left(\boldsymbol{x}^{(t)}\right)}{C_m\left(\boldsymbol{x}^{(t)}\right)^2} \\ &= \frac{1}{C_m\left(\boldsymbol{x}^{(t)}\right)} \left(e_{m,k} - \frac{E_m\left(\boldsymbol{x}^{(t)}\right) \cdot c_{m,k}}{C_m\left(\boldsymbol{x}^{(t)}\right)} \right) \\ &= \frac{e_{m,k} - c_{m,k} \cdot \mathrm{WWER}_m\left(\boldsymbol{x}^{(t)}\right)}{C_m\left(\boldsymbol{x}^{(t)}\right)} \end{aligned}$$

ここで，$x_k^{(t)}$ は t 回目の更新時の x_k の値を表し，$\boldsymbol{x}^{(t)}$ は $x_k^{(t)}$ から構成されるベクトルである．α は単語重要度の更新幅を調整するパラメータを表す．β_{min} および β_{max} はそれぞれ単語重要度の最小値と最大値を表す．β_{min} および β_{max} を設定しないと，一部の単語にきわめて大きい重要度やきわめて小さい重要度が設定されてしまい，実際のベイズリスク最小化音声認識における仮説に対して $\mathrm{WWER} \fallingdotseq 0$ や $\mathrm{WWER} \gg 1$ となって音声認識性能を低下させる可能性が増大する．β_{min} および β_{max} は，これらを防ぐ目的で用いる．本手法は WWER と IRDR の間に強い正の相関が得られるように単語重要度を推定する手法である．なお，k 番目の単語が学習データに含まれない場合は，式 (7) の $e_{m,k}$ および $c_{m,k}$ は 0 となり，その単語の重要度 x_k は更新されず初期値のままとなる．

3.3 単語重要度推定のための学習データ

提案する単語重要度の自動推定手法には以下の 3 つのデータが必要である．

- 想定される検索要求の発話データ（の音声認識結果）
- 発話データの正しい書き起こし
- 各検索要求の正解集合

本論文では，これらの 3 つの学習データの作成方法が異なる以下の 3 種類の単語重要度推定手法を提案する．

- 学習データをすべて人手で作成する教師あり推定（3.3.1 項）
- 学習データを部分的に人手で作成する半教師あり推定（3.3.2 項）
- 学習データをすべて自動で作成する教師なし推定（3.3.3 項）

3.3.1 単語重要度の教師あり推定手法

想定される検索要求の発話データの音声認識結果，その正しい書き起こし，および検索の正解集合の 3 つのデータをすべて人手で作成し，単語重要度を自動推定する手法を本論文では教師あり推定と表記する．教師あり推定では学習データ作成のコストが非常に大きいが，信頼できる単語重要度を推定することができる．検索対象が小規模のシステムであり，検索要求のデータを十分に獲得できている場合に，教師あり推定を実現可能と考えられる．

3.3.2 単語重要度の半教師あり推定手法

検索対象が大規模なシステムであれば，検索の正解集合を人手で作成するコストは高く，教師あり推定は困難である．これに対し，検索の正解集合を自動で生成することで学習データ作成のコストを軽減し，検索対象が大規模なシステムに対しても単語重要度を自動推定することを可能とする手法も提案する．具体的には，想定される検索要求の発話データの音声認識結果とその正しい書き起こしは人手で与え，検索の正解集合を人手で与えずに単語重要度を自動推定する手法を提案する．本論文では，この手法を半教師あり推定と表記する．半教師あり推定の手順は，3.2 節で示した単語重要度推定手順の (2) から (4) を以下で置き換えたものである．

(2) W_m に対する検索の擬似正解データを生成する．
(3) 検索の擬似正解データを用いて，W_m で検索を行ったときの検索性能 R_m を得る．
(4) 検索の擬似正解データを用いて，W'_m で検索を行ったときの検索性能 H_m を得る．

半教師あり推定は，タスク・ドメインが限定されており，検索要求が十分に推測できる場合に実現可能と考えられる．また，すでに稼働している情報検索システムでは，過去に入力された検索要求を学習データとして利用できるため，想定される検索要求の集合を生成するコストは高くなく，このような場合にも半教師あり推定は利用可能である．半教師あり推定では，学習データ作成のコストは教師あり

推定よりも小さい反面，推定された単語重要度の信頼度は教師あり推定よりも低くなると考えられる．

3.3.3 単語重要度の教師なし推定手法

Web 検索などのオープンタスクで，かつ検索対象も大規模であるような検索システムの場合，想定される検索要求を作成することが困難であり，半教師あり推定も困難である．このような背景に基づき，重要度の推定に必要な 3 つのデータをすべて自動で生成して推定を行う手法を提案する．本論文では，この手法を教師なし推定と表記する．教師なし推定の手順は，3.2 節で示した単語重要度推定手順の (1) から (4) を以下で置き換えたものである．

(1) 学習データとして テキスト擬似検索要求 W_m とそれに対応する擬似音声認識結果 W'_m のペアを N 件用意し，各 m ($m = 1, \ldots, N$) に対して手順 (2)〜(6) を行う．
(2) W_m に対する検索の擬似正解データを生成する．
(3) 検索の擬似正解データを用いて，W_m で検索を行ったときの検索性能 R_m を得る．
(4) 検索の擬似正解データを用いて，W'_m で検索を行ったときの検索性能 H_m を得る．

教師なし推定を用いることで，タスク・ドメインおよび検索対象の規模にかかわらず，単語重要度の推定が可能となる．

3.4 単語重要度推定のための学習データの自動生成方法

3.4.1 検索の擬似正解データの自動生成方法

半教師あり推定と教師なし推定では，各検索要求に対する検索の正解集合を人手で与えない．検索の真の正解集合の代わりに，書き起こし文での検索結果のうち，上位のもの数件を擬似正解データとして用いる．この代用は，音声認識誤りが含まれる文での検索結果を，書き起こし文での検索結果で評価することを意味する．IRDR は書き起こし文を用いて検索を行ったときの結果と，音声認識誤りが含まれる文を用いて検索を行ったときの結果の異なり方を表す尺度として定義されているため，この代用は妥当である．

3.4.2 テキスト擬似検索要求と擬似音声認識結果の自動生成方法

教師なし推定では，検索要求の発話と書き起こしを与えない．そのために，想定される検索要求（テキスト）とそれに対応する擬似音声認識結果を自動生成する手法が必要である．これには様々な手法が考えられるが，本論文では検索要求のテキストをテンプレートから擬似的に作成し，ランダムに単語を挿入，削除，置換することで擬似的な音声認識結果を生成する．この手法で生成される擬似的な音声認識結果は実際の音声認識結果とは異なる可能性が高いが，単語重要度の推定には単語の誤りが含まれているペアがあれば十分と考える．これは，本手法は，ある単語が別の単語に置き換えられたときに検索に及ぼす影響から重要度を推定するものであり，実際の音声認識誤りを反映していなくても重要度が推定できるものであることと，ランダムに大量の誤りパターンを生成することで，様々な語の重要度を推定できると考えられるためである．

4. 単語重要度推定の実験

4.1 テストコレクション

本研究では，NTCIR-9 SpokenDoc [9] テストコレクションを用いる．これは，日本語話し言葉コーパス（CSJ）[10] の講演 2,702 件を検索対象とするタスクである．

本研究で用いたデータセットの構成を以下に示す．

- 検索対象文書：CSJ の講演の書き起こしテキスト（2,702 講演）
- 利用者の検索要求を記述した「検索課題」：39 課題（dry run 用）
- 利用者の検索要求を記述した「検索課題」：86 課題（formal run 用）
- 検索課題を満たす「正解文書のリスト」

4.2 検索システムの構成

情報検索システムはベクトル空間モデルに基づく文書検索システムとし，GETA [11] を用いて構築した．索引語には名詞と動詞の基本形を用いた．本研究では，検索要求 $\boldsymbol{Q}$ が与えられたとき，すべての文書 $\boldsymbol{D}_i$ について $\boldsymbol{Q}$ との類似度 $\mathrm{Sim}(\boldsymbol{Q}, \boldsymbol{D}_i)$ を算出し，類似度が高い順に上位 1,000 件を出力する．本研究では，提案手法が異なる検索システムで有効であることを調査するためにベクトルの類似度尺度として，SMART [12]，TFIDF，TF の 3 種類を採用し，異なる 3 種類の検索システムを構築した．

SMART，TFIDF，TF はそれぞれ式 (8)，(9)，(10) で定式化される．

$$\mathrm{Sim}\,(\boldsymbol{Q}, \boldsymbol{D}_i) = \mathrm{SMART}\,(\boldsymbol{Q}, \boldsymbol{D}_i) = \sum_{t=1}^{T} (Q_t \cdot D_{i,t}) \quad (8)$$

$$Q_t = \begin{cases} \dfrac{1 + \log\,(\mathrm{qtf}_t)}{1 + \log\,(\mathrm{avqtf})} \cdot \log \dfrac{N}{n_t} & \text{if } \ \mathrm{qtf}_t > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$D_{i,t} = \begin{cases} \dfrac{1 + \log\,(\mathrm{tf}_{i,t})}{1 + \log\,(\mathrm{avtf})} \cdot \mathrm{Norm} & \text{if } \ \mathrm{tf}_{i,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{Norm} = \frac{1}{(1 - \mathrm{slope}) \cdot \mathrm{pivot} + \mathrm{slope} \cdot \mathrm{utf}_i}$$

$$\begin{aligned} \mathrm{Sim}\,(\boldsymbol{Q}, \boldsymbol{D}_i) &= \mathrm{TFIDF}\,(\boldsymbol{Q}, \boldsymbol{D}_i) \\ &= \sum_{t=1}^{T} \left(\mathrm{qtf}_t \cdot \frac{\mathrm{tf}_{i,t}}{\mathrm{dtf}_i} \cdot \log \frac{N}{n_t} \right) \end{aligned} \quad (9)$$

$$\begin{aligned} \mathrm{Sim}\,(\boldsymbol{Q}, \boldsymbol{D}_i) &= \mathrm{TF}\,(\boldsymbol{Q}, \boldsymbol{D}_i) \\ &= \sum_{t=1}^{T} \left(\mathrm{qtf}_t \cdot \frac{\mathrm{tf}_{i,t}}{\mathrm{dtf}_i} \right) \end{aligned} \quad (10)$$

ここで，$\mathrm{tf}_{i,t}$ は $\boldsymbol{D}_i$ 中での単語 t の出現数，avtf は $\boldsymbol{D}_i$ における単語の出現数の平均を表す．dtf_i は $\boldsymbol{D}_i$ 中での総単語数を表す．pivot は 1 文書中の異なり単語数の平均，utf_i は $\boldsymbol{D}_i$ 中の異なり単語数を表す．slope は補間係数であり，本研究では 0.2 とした．qtf_t は，$\boldsymbol{Q}$ 中での単語 t の出現数，avqtf は $\boldsymbol{Q}$ に含まれる単語の出現数の平均を表す．N は検索対象の文書集合の全文書数を表し，n_t は単語 t を含む文書の数を表す．

4.3 検索の評価尺度

本研究では，提案手法が検索の評価尺度が異なっても頑健に動作することを示すために検索性能の評価尺度として，11 点平均精度（11ptAP: 11-point Average Precision）[13]，補間なし平均精度（AP: noninterpolated Average Precision）[13]，DCG（Discounted Cumulative Gain）[14] の 3 種類を用いた．

11 点平均精度は式 (11) で定式化される．

$$\mathrm{11ptAP}_{Q_k} = \frac{1}{11}\sum_{i=0}^{10} IP_{Q_k}\left(\frac{i}{10}\right) \tag{11}$$
$$IP_{Q_k}(x) = \max_{x \le R_{Q_k}(t)} P_{Q_k}(t)$$

ここで，$R_{Q_k}(t)$ と $P_{Q_k}(t)$ は，それぞれ Q_k に関する検索順位 t における再現率と精度を表す．$IP_{Q_k}(x)$ は，再現率レベルが x 以上の精度 $P_{Q_k}(t)$ の最大値を表す補間精度である．

補間なし平均精度は式 (12) で定式化される．

$$\mathrm{AP}_{Q_k} = \frac{1}{N_{Q_k}}\sum_t I_{Q_k}(t)\, P_{Q_k}(t) \tag{12}$$

ここで，$I_{Q_k}(t)$ は検索要求 Q_k に関する検索順位 t の文書が適合文書であれば 1 を，そうでなければ 0 を返す関数を表す．$P_{Q_k}(t)$ は，Q_k に関する検索順位 t における精度を表す．N_{Q_k} は Q_k の適合文書の総数を表す．

DCG は式 (13) で定式化される．

$$\mathrm{DCG}_{Q_k}(t) = \begin{cases} G_{Q_k}(1) & \text{if } t = 1 \\ \mathrm{DCG}_{Q_k}(t-1) + \dfrac{G_{Q_k}(t)}{\log(t)} & \text{otherwise} \end{cases} \tag{13}$$
$$G_{Q_k}(t) = \begin{cases} r & \text{if } d_{Q_k}(t) \in R \\ i & \text{otherwise} \end{cases}$$

ここで，$d_{Q_k}(t)$ は Q_k に関する検索結果の t 番目の文書を表す．R は，適合と判定された文書集合（正解データ）を表す．R に該当しない検索結果は不適合とする．本研究では，各適合度の利得の度合いは $(r, i) = (2, 0)$ とした．また，対数関数の底は 2 とした．

本研究では，IRDR 計算のために 11 点平均精度，補間なし平均精度，DCG をそれぞれ用いる．検索要求 Q_k の音声認識誤りによる検索性能低下率 IRDR_{Q_k} を 11 点平均精度を用いて計算する例を式 (14) に示す．

$$\mathrm{IRDR}_{Q_k} = 1 - \frac{\mathrm{11ptAP}_{Q'_k}}{\mathrm{11ptAP}_{Q_k}} \tag{14}$$

ここで，$\mathrm{11ptAP}_{Q_k}$ は検索要求 Q_k の書き起こしを用いて検索を行った場合の 11 点平均精度を表し，$\mathrm{11ptAP}_{Q'_k}$ は検索要求 Q_k の音声認識結果 Q'_k を用いて検索を行った場合の 11 点平均精度を表す．補間なし平均精度，DCG を用いて IRDR を計算する場合も同様に行う．

4.4 実験データ

4.4.1 教師あり推定のための学習データ

単語重要度の教師あり推定の学習データ生成には以下を用いた．

- NTCIR-9 SpokenDoc dry run 用の検索課題の集合（$\boldsymbol{W}_d$）
- NTCIR-9 SpokenDoc dry run 用の検索課題の読み上げ音声の音声認識結果の集合（$\boldsymbol{W}'_d$）
- NTCIR-9 SpokenDoc で提供されている正解文書集合（$\boldsymbol{C}$）

検索要求の発話データの正しい書き起こしの集合 $\boldsymbol{W}_d$ には，NTCIR-9 SpokenDoc の dry run 用の検索課題 39 件を用いた．

検索要求の発話データの音声認識結果の集合 $\boldsymbol{W}'_d$ には，dry run 用の 39 件の読み上げ音声 546 発話（男性 10 名，女性 4 名の合計 14 名 × 39 件 = 546 発話）の音声認識の結果を用いた．このとき，14 名の音声認識結果のうち，認識結果が同じものはそれに対応する書き起こしとともに学習データから取り除いた．この音声認識結果を得るための音声認識システムとして，JNAS コーパスから学習した triphone モデル（文献 [15] のもの），CSJ [10] の講演 2,702 件の書き起こしから学習した 3-gram 言語モデル（語彙サイズ約 2 万）からなるものを用いた．

検索の正解集合 $\boldsymbol{C}$ には，NTCIR-9 SpokenDoc の正解文書リストを用いた．NTCIR-9 SpokenDoc では正解ラベルとして適合（R）ラベルと部分適合（P）ラベルが人手により付されており，本研究ではこのうち適合（R）ラベルが付けられた文書のみを正解文書と見なした．

4.4.2 半教師あり推定のための学習データ

単語重要度の半教師あり推定の学習データ生成には以下を用いた．

- NTCIR-9 SpokenDoc dry run 用の検索課題の集合（$\boldsymbol{W}_d$）
- NTCIR-9 SpokenDoc dry run 用の検索課題の読み上げ音声の音声認識結果の集合（$\boldsymbol{W}'_d$）
- $\boldsymbol{W}_d$ の各検索結果から生成した擬似正解データの集合（$\boldsymbol{C}_{\boldsymbol{W}_d}$）

検索要求の発話データの正しい書き起こしの集合 $\boldsymbol{W}_d$ と

音声認識結果の集合 $\boldsymbol{W}'_d$ は，教師あり推定と同一のものである．

検索の正解集合 $\boldsymbol{C}_{\boldsymbol{W}_d}$ には，3.4.1 項の手法に基づき，書き起こしテキスト $\boldsymbol{W}_d$ の各検索課題の検索結果を用いた．なお，本実験での検索タスクと異なる音声検索タスク（NTCIR-3 WEB 検索タスク [14]）を用いた予備実験に基づき，検索結果の上位 20 件を正解集合とした．

4.4.3 教師なし推定のための学習データ

単語重要度の教師なし推定の学習データ生成には以下を用いた．

- 検索要求のテンプレートと特徴語から生成した擬似検索要求の集合（$\boldsymbol{W}_p$）
- 擬似検索要求に対しランダムに誤りを発生させて生成した擬似音声認識結果の集合（$\boldsymbol{W}'_p$）
- 擬似検索要求 $\boldsymbol{W}_p$ の各検索結果から生成した擬似正解データの集合（$\boldsymbol{C}_{\boldsymbol{W}_p}$）

擬似検索要求の集合 $\boldsymbol{W}_p$ はテンプレートと特徴語を用いて 10,000 件作成した．検索要求のテンプレートは，NTCIR-3 WEB 検索タスク [14] の音声入力検索課題 47 件をもとに作成した．具体的には，検索要求中の一部を特徴語をあてはめるためのブランクに置き換えることで作成した．擬似検索要求生成のための特徴語は，4.4.1 項で述べた音声認識システムの単語辞書中の単語（本実験では名詞）とした．これは，音声入力型情報検索システムでは，音声認識辞書中の単語のみを検索に用いる語とするので，それらの重要度が推定されれば十分なためである．実際に検索要求からテンプレートを作成し，擬似検索要求を生成した例を**図 3** に示す．

擬似音声認識結果の集合 $\boldsymbol{W}'_p$ として，3.4.2 項の手法に基づきランダムに単語を挿入，削除，置換することで，擬似検索要求 $\boldsymbol{W}_p$ の各文ごとに 5 種類の擬似音声認識結果を生成し，これを用いた（合計 50,000 文）．挿入および置換する単語リストは 4.4.1 項で述べた音声認識システムの単語辞書に登録されているすべての単語とした．擬似音声認識結果は 50,000 文の WER の平均が 50%となるように生成した．実際に生成された擬似検索要求と擬似音声認識結果の例を**図 4** に示す．

元の検索要求：	サルサ を 踊れる よう に なる 方法 が 知り たい
テンプレート：	<> に なる 方法 が 知り たい
擬似検索要求：	農地 に なる 方法 が 知り たい

図 3 擬似検索要求生成の例
Fig. 3 Example of quasi-query generation.

擬似検索要求 1：	研修生 の 歴代 の 文句 が 知り たい
擬似音声認識結果 1：	健在 の インテリア 文句 が 知り たい
擬似検索要求 2：	修道院 の 歴代 の 越後湯沢 が 知り たい
擬似音声認識結果 2：	修道院 の 歴代 の うんそう が 知り たい

図 4 擬似検索要求と擬似音声認識結果の例
Fig. 4 Example of quasi-queries and quasi-ASR results.

検索の正解集合 $\boldsymbol{C}_{\boldsymbol{W}_p}$ には，3.4.1 項の手法に基づき，擬似検索要求 $\boldsymbol{W}_p$ の各検索課題の検索結果を用いた．なお，半教師あり推定と同様に，異なるタスクでの予備実験に基づき，検索結果の上位 20 件を正解集合とした．

4.5 評価データ

評価データ作成のための検索要求には以下の 2 種類の合計 125 件を用いた．

- NTCIR-9 SpokenDoc dry run 用の検索課題 39 件（クローズド評価用検索要求）
- NTCIR-9 SpokenDoc formal run 用の検索課題 86 件（オープン評価用検索要求）

前者は教師ありおよび半教師あり推定で単語重要度の推定に用いた検索要求の書き起こし $\boldsymbol{W}_d$ と同一であり，提案する推定の枠組みが適切に動作するかの評価に用いる．これらの検索要求に対して，3.4.2 項の手法を用いてランダムに単語の削除，挿入，置換を行い，擬似的に音声認識誤りを含む検索要求を作成した．挿入および置換する単語リストは 4.4.1 項で述べた音声認識システムの単語辞書から作成した．その際，WER がおよそ 10%，20%，30%，...，100%となるようにし，各 WER でそれぞれ 100 件作成した．これは，評価データに何らかの偏りが発生することを防ぐためである．こうして，クローズド評価用検索要求の評価データとして書き起こしテキストと誤りを含むテキストの 39,000 件のペアを，オープン評価用検索要求の評価データとして 86,000 件のペアを準備した．このペアで検索を行って IRDR を計算し，IRDR > 0 のペアを評価データとした．IRDR ≤ 0 となるのは，音声認識誤りが含まれているにもかかわらず偶然に検索性能が低下しなかった/向上したケースと解釈できるものであり，本 WWER はこれらのようなものを予測するものではないため，評価データから除いた[*1]．検索の正解集合には，NTCIR-9 SpokenDoc の正解文書のリストのうち，適合度 R のものを用いた．

単語重要度推定の評価は，書き起こしテキストと誤りを含むテキストの各ペアに対して，推定された単語重要度に基づく WWER を計算し，それとそのペアの検索性能の低下率（IRDR）との相関を求めることで行った．高い正の相関が得られるほど，WWER による IRDR の予測精度が高いことを示しているため，適切な単語重要度が得られているといえる．なお，IRDR の最大値は 1 となるため，WWER と IRDR の相関を求める際に，WWER > 1 の場合は WWER $= 1$ として計算した．本論文ではいくつか実験を行っているが，WWER > 1 となるようなペアの割合は，平均で 1 割程度であった．

[*1] 実際に IRDR ≤ 0 のものは，学習時にも取り除いている．3.2 節手順 (7) 参照．

4.6 単語重要度推定の評価

4.6.1 実験条件

検索の類似度 3 種類と検索の評価尺度 3 種類の組合せ合計 9 種類の検索システムで単語重要度推定の実験を行った．単語重要度推定は以下の 3 種類を行った．

- 教師あり推定
- 半教師あり推定
- 教師なし推定

本実験では，WWER と IRDR の平均二乗誤差（式 (4)）が収束するまで，もしくは繰返し回数が一定数に達するまで学習を繰り返した．本実験では，単語重要度の更新前と更新後において，それぞれの WWER と IRDR の平均二乗誤差を計算し，その差が 10^{-5} を下回ったときに平均二乗誤差が収束したと判断した．単語重要度推定の最大繰返し回数は 10,000 回とした．単語重要度の初期値としてすべての単語の重要度を 1 とした．この状態で WWER を計算すると，WWER は WER と一致するため，この初期値とした．単語重要度の推定時のパラメータ（式 (6)）として，α は 0.001，β_{min} は 1，β_{max} は 10 と設定した．これらの決定は，異なるタスク（NTCIR-3 WEB 検索タスク [14]）を用いた予備実験の結果，具体的には，β_{min} に対する β_{max} の比を 10 程度とし，最大繰返し回数と α の積を β_{max}（$= 10$）程度となるように決定すればよいことに基づいている．

比較のために，ベースラインとして以下の 2 種類の単語重要度を設定した．

- すべての単語に 1（“一様” と表記，このとき WWER = WER）
- 内容語（名詞と動詞）に 10，機能語（それ以外）に 1（“品詞” と表記）

提案手法とベースラインの比較のために，それぞれの単語重要度を用いて計算した WWER と IRDR の相関係数の差の検定を有意水準 1%で行った．このとき，Bonferroni 法を用いて 2 つのベースラインとの検定を行った．具体的には，2 つのベースラインそれぞれとの検定を行い，調整化された有意水準 $\alpha' = 0.005$ を用いた．相関係数の差の検定手法には，Meng-Rosenthal-Rubin Method [16] を用いた．

4.6.2 教師あり推定と半教師あり推定の結果

教師あり推定および半教師あり推定の実験結果を**表 1** に示す．表中の太字は，相関係数が，2 つのベースラインのいずれと比較しても有意に高い場合を示す．まず，クローズド評価用の検索要求を用いた評価について述べる．SMART 検索システム（検索評価は 11 点平均精度）では，一様な単語重要度を用いた場合，すなわち WER と IRDR の相関係数は 0.52 であった．品詞ごとに一様な単語重要度を用いた WWER と IRDR の相関係数は 0.53 であり，一様な単語重要度の場合と同等の結果となった．検索システムで索引語として用いられている内容語に高い重要度を与えているにもかかわらず，一様な重要度を用いた場合と同等である．このことは，人手で単語重要度を決定することが困難であることを示している．これに対して，教師あり推定で決定した単語重要度を用いた WWER と IRDR の相関係数は 0.66 であった．半教師あり推定で決定した単語重要度を用いた WWER と IRDR の相関係数は 0.63 であった．教師あり推定で得られた単語重要度に基づく WWER と IRDR の相関は，9 つすべてのシステムで 2 つのベースラインに比べて高く，有意水準 1%で有意差があった．また，半教師あり推定を用いることで，9 つのシステムのうち 7 つで一様な重要度および品詞ごとに一様な重要度を用いる場合よりも高い相関が得られた（有意水準 1%で有意差あり）．このことは，学習データの検索要求に含まれる単語に対して，適切な単語重要度が推定できることを示している．実際に，クローズド評価用検索要求 39 件に含まれる単語は，学習データ中（$\boldsymbol{W}_d$ もしくは $\boldsymbol{W}'_d$）にすべて含まれている．

表 1 WWER と IRDR の相関（教師あり推定と半教師あり推定）
Table 1 Correlation between WWERs and IRDR (supervised/semi-supervised estimation).

検索システム			単語重要度の決定方法							
			クローズド評価用の検索要求で評価				オープン評価用の検索要求で評価			
検索		評価	一様	品詞	教師あり推定	半教師あり推定	一様	品詞	教師あり推定	半教師あり推定
SMART	+	11ptAP	0.52	0.53	**0.66**	**0.63**	0.53	0.53	0.53	0.53
SMART	+	AP	0.53	0.54	**0.66**	**0.63**	0.53	0.54	0.53	0.53
SMART	+	DCG	0.51	0.52	**0.65**	**0.63**	0.50	0.50	0.50	0.50
TFIDF	+	11ptAP	0.47	0.47	**0.61**	**0.58**	0.49	0.49	0.49	**0.50**
TFIDF	+	AP	0.46	0.47	**0.62**	**0.58**	0.50	0.49	0.50	0.50
TFIDF	+	DCG	0.45	0.46	**0.58**	**0.54**	0.46	0.46	0.46	0.46
TF	+	11ptAP	0.44	0.43	**0.49**	0.44	0.47	0.47	**0.48**	0.46
TF	+	AP	0.45	0.45	**0.51**	0.45	0.48	0.48	0.48	0.47
TF	+	DCG	0.44	0.44	**0.50**	**0.46**	0.45	0.44	**0.47**	**0.48**

すべての単語重要度が 1 の場合を “一様”，内容語（名詞と動詞）の重要度に 10，機能語（それ以外）の重要度に 1 の場合を “品詞” と表記．
太字は 2 つのベースラインのいずれと比較しても有意に高い相関係数

表 2 WWER と IRDR の相関（教師なし推定）
Table 2 Correlation between WWERs and IRDR (unsupervised estimation).

検索システム			単語重要度の決定方法		
検索		評価	一様	品詞	教師なし推定
SMART	+	11ptAP	0.53	0.53	**0.59**
SMART	+	AP	0.53	0.54	**0.60**
SMART	+	DCG	0.51	0.51	**0.54**
TFIDF	+	11ptAP	0.49	0.48	**0.59**
TFIDF	+	AP	0.49	0.49	**0.60**
TFIDF	+	DCG	0.46	0.46	**0.51**
TF	+	11ptAP	0.46	0.46	0.46
TF	+	AP	0.47	0.47	0.47
TF	+	DCG	0.44	0.44	0.44

すべての単語重要度が 1 の場合を “一様”，内容語（名詞と動詞）の重要度に 10，機能語（それ以外）の重要度に 1 の場合を “品詞” と表記．
太字は 2 つのベースラインのいずれと比較しても有意に高い相関係数

次に，オープン評価用の検索要求を用いた評価について述べる．教師あり推定，半教師あり推定のいずれの結果も，2 つのベースラインとの間に，相関係数の差は見られなかった．これは，本推定手法は，学習データの検索要求に含まれない単語に対してそもそも重要度を推定できない（初期値のままとなる）ものであるためである．実際に，オープン評価用検索要求 86 件に含まれる単語のうち，学習データ中（$\boldsymbol{W}_d$ もしくは $\boldsymbol{W}'_d$）に含まれているものの割合は 35.6%と低かった．しかし，学習データ外の検索要求が入力された際であっても，このことに由来する悪影響がないことが分かる．

これらの結果は，単語重要度の推定手法が期待どおりの動作をすることを示しており，提案する枠組みが適切であることを示している．

4.6.3 教師なし推定の結果

教師なし推定の実験結果を**表 2** に示す．表中の太字は，相関係数が，2 つのベースラインのいずれと比較しても有意に高い場合を示す．教師なし推定においては，クローズド評価用の検索要求とオープン評価用の検索要求の区別がない（クローズド評価用の検索課題 39 件を学習データとして使用していない）ため，これらをすべて用いて評価を行った．なお，教師なし推定の学習データとして用いる擬似検索要求の中に，評価データに含まれる検索要求の書き起こしおよび擬似音声認識結果と偶然に一致するものはなかった．SMART 検索システム（検索評価は 11 点平均精度）では，一様な単語重要度を用いた場合の WWER（WER），品詞ごとに一様な単語重要度を用いた WWER と IRDR の相関係数はいずれも 0.53 であった．これに対して，教師なし推定で決定した単語重要度を用いた WWER と IRDR の相関係数は 0.59 と高かった．9 つのシステムのうち 6 つで最も高い相関係数が得られ，有意水準 1%で有意差があった．このことは，一様な単語重要度および品詞ごとに一様な単語重要度よりも，教師なし推定を行うことで得られた単語重要度から IRDR をより近似する WWER を得ることができることを示している．また，評価用検索要求合計 125 件に含まれる単語のうち，学習データ中（$\boldsymbol{W}_p$ もしくは $\boldsymbol{W}'_p$）に含まれているものの割合は 96.7%であった．このことは，ほぼすべての単語が学習データに含まれており，それらに対して適切に重要度が推定できたことを示している．実際に SMART 検索システム（検索評価は 11 点平均精度）での教師なし推定により得られた単語重要度の例を**図 5** に示す．文書特定能力が高い単語である “講演” や “音声” には，他の単語よりも高い重要度が設定されていることが分かる．

講演	音声	の	特徴	について	知り	たい
3.25	1.95	1.00	1.00	1.00	1.00	1.00

図 5 教師なし推定で得られた単語重要度の例（SMART で検索，11 点平均精度で評価するシステム）
Fig. 5 Example of word importance from unsupervised estimation.

表 3 検索要求の書き起こしでの検索性能
Table 3 IR performance by text query.

検索	評価		
	11ptAP	AP	DCG
SMART	0.54	0.51	0.71
TFIDF	0.36	0.33	0.57
TF	0.08	0.07	0.24

4.6.4 実験結果の考察

単語重要度推定の実験を行った結果，検索類似度として SMART もしくは TFIDF を用いたシステムでは適切な単語重要度が推定できたものの，検索類似度として TF を用いたシステムでは IRDR と高い相関を持つ WWER を定義できるような単語重要度を推定できないケースが多かった．この原因として，検索システム自体の性能が考えられる．クローズド評価用およびオープン評価用すべての検索要求の正しい書き起こし（音声認識誤りなし）を用いて検索を行った際の検索性能を**表 3** に示す．SMART 検索システムの検索性能は，11ptAP，AP，DCG でそれぞれ 0.54，0.51，0.71 であった．TFIDF 検索システムの検索性能は，11ptAP，AP，DCG でそれぞれ 0.36，0.33，0.57 であった．これに対し，TF 検索システムの検索性能は，11ptAP，AP，DCG でそれぞれ 0.08，0.07，0.24 と，他のシステムと比較して検索性能が低いことが分かる．元々の検索性能が低いシステムでは検索結果の信頼性が低く，それを用いた単語重要度の推定は難しいと考えられる．

4.7 単語重要度推定が情報検索に与える効果

WWER と IRDR の相関が高くなることにより，どの程

度情報検索の性能向上が得られるかは，音声認識システムや検索システムの構成などによって異なるため一概に述べることはできない．ただし，我々はこれまでに音声言語処理システムにおいて，音声認識性能と言語処理性能の間により強い相関を持つように単語重要度を定義して WWER 最小化音声認識を行うことが言語処理の性能向上につながることを示しており [17], [18]，特に元々の音声認識性能が低い場合に WWER 最小化音声認識の効果が見られることを示している [18]．また，ベイズリスク最小化音声認識は，認識率が低い際に有効であることが Schlüter らによって示されている [19]．これらのことから，本手法で単語重要度を推定して WWER 最小化音声認識を行うことで，特に音声認識性能が低いような音声に対して，情報検索への影響が小さくなるような音声認識を行えることが期待できるといえる．

5. おわりに

情報検索用の音声認識のための，適切な単語重要度の自動推定手法を提案した．具体的には，想定される検索要求の発話データ，その書き起こし，および検索の正解集合の 3 つのデータを用いて自動で単語重要度を推定する手法を提案した．教師ありおよび半教師あり推定の結果から，提案する単語重要度推定の枠組みが適切であることを示し，教師なし推定により適切な単語重要度を設定できることを示した．今後は，様々な音声入力型情報検索システムを用いて本手法で単語重要度を推定し，実際に WWER 最小化音声認識を行って評価を行っていく予定である．

謝辞 本研究は科研費基盤研究（B）（21300066）の助成を受けたものである．

参考文献

[1] 翠 輝久，河原達也：限定されたドメインにおける質問応答機能を備えた文書検索・提示型対話システム，情報処理学会研究報告，2006-SLP-62, pp.69–74 (2006).
[2] 桐山伸也，広瀬啓吉，峯松信明：話題知識を導入した文献検索音声対話システム，電子情報通信学会論文誌，Vol.J85-D-II, No.5, pp.863–876 (2002).
[3] Hori, C., Hori, T., Isozaki, H., Maeda, E., Katagiri, S. and Furui, S.: Deriving Disambiguous Queries in a Spoken Interactive ODQA System, *Proc. IEEE-ICASSP*, pp.624–627 (2003).
[4] 翠 輝久，駒谷和範，清田陽司，河原達也：音声対話によるソフトウェアサポートのための効率的な確認戦略，電子情報通信学会論文誌，Vol.J88-D-II, No.3, pp.499–508 (2005).
[5] Matsushita, M., Nishizaki, H., Utsuro, T. and Nakagawa, S.: Improving Keyword Recognition of Spoken Queries by Combining Multiple Speech Recognizer's Output for Speech-driven WEB Retrieval, *IEICE Trans. Inf. & Syst.*, Vol.E88-D, No.3, pp.472–480 (2005).
[6] 南條浩輝，河原達也，七里 崇：音声理解を指向したベイズリスク最小化枠組みに基づく音声認識，電子情報通信学会論文誌，Vol.J91-D, No.5, pp.1314–1324 (2008).
[7] Goel, V., Byrne, W. and Khudanpur, S.: LVCSR rescoring with modified loss functions: A decision theoretic perspective, *Proc. IEEE-ICASSP*, Vol.1, pp.425–428 (1998).
[8] Stolcke, A., König, Y. and Weintraub, M.: Explicit word error minimization in N-best list rescoring, *Proc. EUROSPEECH*, pp.163–166 (1997).
[9] Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T. and Matsui, T.: Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop, *NTCIR-9*, pp.223–235 (2011).
[10] Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, *Proc. ISCA & IEEE-SSPR*, pp.7–12 (2003).
[11] 高野明彦，西岡真吾，今一 修，岩山 真，丹羽芳樹，久光徹，藤尾正和，徳永健伸，奥村 学，望月 源，野本忠司：汎用連想計算エンジンの開発と大規模文書分析への応用，入手先 ⟨http://geta.ex.nii.ac.jp/pdf/itx002.pdf⟩ (2002).
[12] 小作浩美，内山将夫，井佐原均，河野恭之，木戸出正継：WWW 検索における複数検索結果の統合処理とその評価，情報処理学会論文誌：データベース，Vol.44, No.SIG 8(TOD 18), pp.78–91 (2003).
[13] 北 研二，津田和彦，獅々堀正幹：情報検索アルゴリズム，共立出版，ISBN 4-320-12036-1 (2002).
[14] 江口浩二，大山敬三，石田栄美，神門紀子，栗山和子：NTCIR-3 WEB：Web 検索のための評価ワークショップ，*NII Journal*, No.6, pp.31–56 (2003).
[15] 河原達也，武田一哉，伊藤克亘，李 晃伸，鹿野清宏，山田篤：連続音声認識コンソーシアムの活動報告及び最終版ソフトウェアの概要，情報処理学会研究報告，2003-SLP-49, pp.325–330 (2003).
[16] Meng, X.L., Rosenthal, R. and Rubin, D.B.: Comparing correlated correlation coefficients, *Psychological Bulletin*, Vol.111, No.1, pp.172–175 (1992).
[17] Nanjo, H. and Kawahara, T.: A New ASR Evaluation Measure and Minimum Bayes-Risk Decoding for Open-domain Speech Understanding, *Proc. IEEE-ICASSP*, pp.1053–1056 (2005).
[18] Shichiri, T., Nanjo, H. and Yoshimi, T.: Minimum Bayes-Risk Decording with Presumed Word Significance for Speech Based Information Retrieval, *Proc. IEEE-ICASSP*, pp.1557–1560 (2008).
[19] Schlüter, R., Nussbaum-Thom, M. and Ney, H.: On the relation of Bayes Risk, Word Error, and Word Posteriors in ASR, *Proc. INTERSPEECH*, pp.230–233 (2010).

古谷 遼

2011 年龍谷大学理工学部情報メディア学科卒業．2013 年同大学院修士課程修了．在学中，音声認識に関する研究に従事．

七里 崇

2007 年龍谷大学理工学部情報メディア学科卒業．2010 年同大学院理工学研究科修士課程修了．在学中，音声認識・理解の研究に従事．2007 年情報処理学会音声言語情報処理研究会学生奨励賞受賞．2010 年日本音響学会学生優秀発表賞受賞．

南條 浩輝 （正会員）

1999 年京都大学工学部情報学科卒業．2001 年同大学院情報学研究科修士課程修了．2004 年同博士後期課程修了．同年龍谷大学理工学部助手．2007 年より同助教．音声認識・理解の研究に従事．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，IEEE 各会員．2008 年度日本音響学会粟屋潔学術奨励賞受賞．