Vol.2016-NL-226 No.4
Vol.2016-SLP-111 No.4
2016/5/16

非構造化文書からの用語検索における用語候補のリスコアリングの検討

森田　直樹[1,a)]　南條　浩輝[2,b)]　馬　青[3,c)]

概要：意味を表す文書表現 (説明文) を入力として与え，その説明文が示す語句（用語）を検索する用語検索を研究している．これまでに，辞書のような「見出し－説明文」という構造をもつ文書からではなく，そのような構造をもたない非構造化文書からの検索法を提案している．しかし，我々が以前に提案した方法では用語候補の上位に適切でない語が多数出力される問題があった．この問題に対して本論文では，出力された各用語候補に対して説明文との類似度を推定し，それに基づいて用語候補を並び替える手法を提案する．講演音声ドキュメントからの用語検索において，提案手法により，平均逆順位 (MRR) が向上することを確認できた．

キーワード：非構造化文書，用語検索，リスコアリング，パッセージ検索，関連語抽出

1. はじめに

意味を表す文章表現（説明文）からそれが示す語句（用語）を検索する用語検索について述べる．これまでに，辞書や Wikipedia などに代表される，見出しとその説明文を自身の構造として含んでいる文書（構造化文書）を検索対象とした用語候補の研究がなされている [1][2][3][4]．これは，通常の辞書引きとは逆の手順，すなわち定義文を検索キーとして見出し語を見つける研究である．しかし，新語や一般的でない専門用語は辞書に載っていないことが多く，これらの手法では探し出せないという問題があった．この問題に対し我々は，そのような語が存在することが多いマイクロブログや SNS，論文のような「見出し－説明文」という構造が存在しない文書（非構造化文書）を検索対象とした用語検索を研究している [5].

我々が提案している非構造化文書からの用語検索手法は，入力の説明文と意味的に類似していると思われる文書の一部（パッセージ）を選択し，そこから関連語を抽出して用語候補とするものである．これは説明文と似ている文の周辺に，ターゲットとなる用語が含まれているという仮定に基づくものである．実際に我々は，講演音声ドキュメントからの用語（地名とカタカナ語それぞれ 25 語）の検索を行い，本提案手法の仮定の妥当性を確認している [5]. ただし，用語候補の上位に適切でない語が多数出力され，求めている正しい回答（用語）の順位が低いという問題があった．これは，パッセージからの用語候補の抽出と順位付けが，パッセージと用語候補の類似度スコアのみに基づいて行われていることが主な原因であった．

この問題に対し，本論文では，抽出された用語候補それぞれに対し，説明文との類似性を考慮した上でスコアを修正する方法，すなわち用語候補のリスコアリング手法を提案する．

2. データ

2.1 非構造化文書

本研究は，「見出し－説明文」という構造が存在しない文書（非構造化文書）を検索対象としている．

非構造化文書には様々なものが考えられるが，検索対象として講演音声ドキュメントを採用する．これは日本語話し言葉コーパス [6] の学会講演 987 件と模擬講演 1715 件の合計 2702 件の講演の音声認識結果のテキスト（認識率 65 %から 95 %）[7] を検索対象とするものである．講演音声ドキュメントの特徴として「見出し－説明文」という構造だけでなく句読点や段落情報がないことが挙げられる．つまりどこまでが 1 文であるのかを示す手がかりもない．

[1] 龍谷大学理工学研究科
Graduate School of Science and Technology, Ryukoku University
[2] 京都大学学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto Univ.
[3] 龍谷大学理工学部
Faculty of Science and Technology, Ryukoku University
[a)] t15m007@mail.ryukoku.ac.jp
[b)] nanjo@media.kyoto-u.ac.jp
[c)] qma@math.ryukoku.ac.jp

Vol.2016-NL-226 No.4
Vol.2016-SLP-111 No.4
2016/5/16

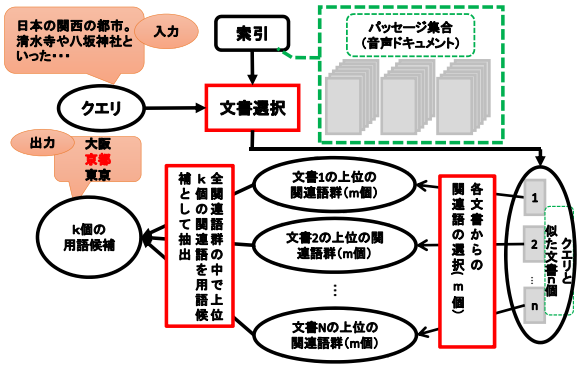

図 1　システムのイメージ図

したがって文書の一部の意味的に似たまとまりを見つけることも難しい．本論文では，[8] に基づき，無音（息つぎのポーズ）で区切られた音声を発話と定義し，10 発話をまとめて擬似的な意味のまとまりのパッセージとする．このパッセージのうち，説明文と近いものが見つかれば，そのパッセージ中に用語が含まれていると考え，そこから用語候補を取り出す．

2.2　検索クエリ

テストデータの用語は音声認識の辞書に含まれている単語リストの中から選択した地名とカタカナ語の単語をそれぞれ 25 個ずつである．検索に用いる説明文はその用語の説明を表す 3 文からなるものである．

以下に説明文の例を示す．

説明文の例

京都:
日本の関西の都市。清水寺や八坂神社といった寺や神社の名所多い。古都と呼ばれ歴史的価値のあるものが多い。
吉祥寺:
東京都武蔵野市。住みたい町ランキングに度々全国 1 位に。JR 中央線、京王井の頭線が通る。

3.　非構造化文書からの用語検索システム

我々が [5] で提案した非構造化文書からの用語検索システム (図 1）について述べる．これは，まず検索クエリ (説明文) を入力とし，その検索クエリと類似している講演音声ドキュメントの文書の一部（パッセージ）を選択する．次に選択したパッセージの中に正しい回答となる用語が含まれていると仮定し，そのパッセージ中の語それぞれにスコアをつけて用語候補とするものである．これは文書選択システム，関連語選択システム，用語の抽出システムの 3 つから構成されている．それぞれについて以下に述べる．

3.1　文書選択システム

文書選択システムはベクトル空間モデルに基づくものを用いる．これは検索対象（パッセージ）のベクトル表現と検索クエリ (説明文) のベクトル表現の相関量を計算して関連度（スコア）の高い順にパッセージを選択するものである．スコア付けするためには対象となる各パッセージのベクトル d_i ($1 \leq i \leq N$) とクエリのベクトル Q の類似度を求める必要がある．この類似度の算出には SMART[9] を用いる．具体的にはクエリ Q とパッセージ d_i での語 t_k の正規化出現頻度 q_{tk} および d_{i,t_k} を用いて，式 (1) で類似度 $SMART(Q, d_i)$ を与える．

$$SMART(Q, d_i) = \sum_{k=1}^{m} (q_{t_k} \cdot d_{i,t_k}) \tag{1}$$

ただし

$$d_{i,t_k} = \begin{cases} \dfrac{1 + \log(tf_{i,t_k})}{1 + \log(avtf)} & if \quad tf_{i,t_k} > 0 \\ \dfrac{}{(1 - slope) \cdot pivot + slope \cdot utf_i} & \\ 0 \qquad otherwise & \end{cases} \tag{2}$$

$$q_{t_k} = \begin{cases} \dfrac{1+\log(qtf_{t_k})}{1+\log(avqtf)} \log \dfrac{N}{n_{t_k}} & if \quad qtf_{t_k} > 0 \\ 0 \qquad otherwise & \end{cases} \tag{3}$$

ここでは，tf_{i,t_k} は d_i 中の t_k の出現数，$avtf$ は d_i における単語の出現回数の平均を表す．$pivot$ は 1 パッセージ中の異なり単語数の平均，utf_i は d_i 中の異なり単語数を表す．$slope$ は補間係数 (0.2) である．qtf_{t_k} は Q 中での t_k の出現回数，$avqtf$ は Q に含まれる単語の出現回数の平均を表す．N は検索対象パッセージ数を表す．n_{t_k} は t_k を含むパッセージ数を表す．

文書選択の精度向上のために，広域文書類似度を用いた手法 [10] を加えている [5].

3.2　関連語の選択システム

文書選択をした結果，選択された上位 n 件の各パッセージ d ($1 \leq d \leq n$) のそれぞれをクエリ Q_d とみなして式 (3) に基づいて $q_{t_k,d}$ の値を求め，この値の降順でパッセージごとに関連語を一定数（m 個）選択する．

3.3　用語の抽出システム

選択された各関連語 t_k について，$q_{t_k,d}$ の合計値 S_{t_k} を求めて（式 (4)），この値の降順で一定数（k 個）を抽出し，それを用語候補とする．

$$S_{t_k} = \sum_{d=1}^{n} q_{t_k,d} \tag{4}$$

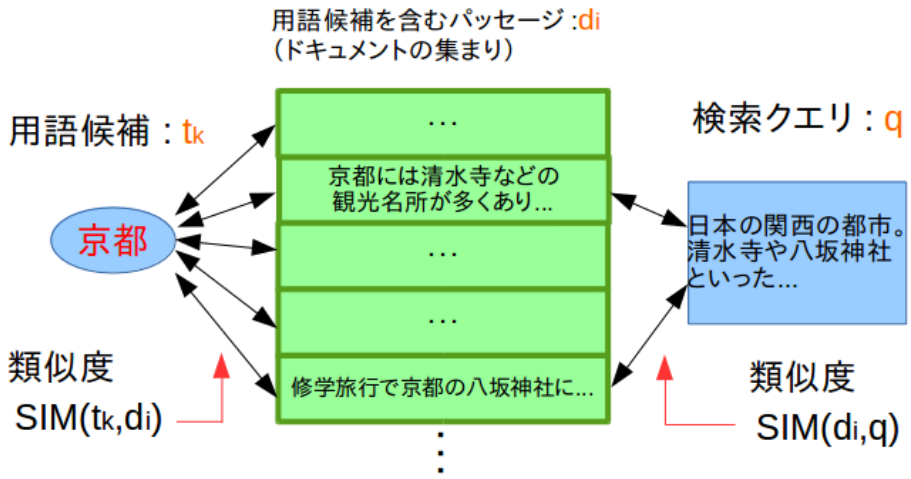

図 2 用語候補と検索クエリの推定

3.4 従来システムの問題点

式 (4) で定義する用語候補 t_k の用語候補らしさのスコア S_{t_k} は $q_{t_k,d}$，すなわちパッセージと用語候補の類似度だけで構成されている．このためパッセージが検索クエリと関連していないものである場合（文書選択の誤りの場合）や，パッセージのメインの記述が別の語に関する話題である場合などは，不適切な関連語が高い順位に表れてしまうという問題点がある．また検索クエリ（説明文）に含まれる語も候補に含まれるという問題点もある．本論文ではこれらの問題に対応する．

4. 用語候補のリスコアリングの手法

3 章で述べた問題点を解決するため，用語候補と検索クエリ（説明文）の類似度を考える．しかし，用語候補は「単語」であり，検索クエリは「文」である．本タスクは，「ユーザーがわからず，それを質問する」タスクであり，用語は検索クエリに含まれないため，これらの類似度は直接計算できない．そのため，次の仮定をおき，用語候補と検索クエリの類似度を推定する方法を提案する．すなわち，正しい用語を多く含む文書（パッセージ）とその説明文は似ていると仮定し，各用語候補を含むパッセージと説明文の類似度を用語候補を検索クエリの類似度としてみなす．この様子を，図 2 に示す．用語候補 t_k と類似度の高いパッセージ d_i を選択し，そのパッセージと検索クエリ q の類似度と用語候補とパッセージの類似度 $SIM(t_k, d_i)$ を，用語候補と検索クエリとの類似度 $SIM(d_i, q)$ の推定に用いる．

4.1 用語候補を含むパッセージの選択

用語候補を含むパッセージの選択について述べる．3 章と同様に SMART を用いて，ベクトル空間に基づきパッセージの選択を行う．各用語候補 t_k を Q だと考え，検索を行う．これにより，それぞれの用語候補 t_k に対して，用語候補 t_k 自身を含んだパッセージ d_i が SMART 順に選ばれる．このとき，用語候補 t_k とパッセージ d_i の類似度を $SIM(t_k, d_i)$ とする．次に選択されたパッセージ d_i と検索クエリ q との類似度 SMART の式によって求める．このパッセージ d_i と検索クエリ q の類似度を $SIM(d_i, q)$ とする．

4.2 用語候補の並び替え

用語候補と検索クエリの類似度 $SIM(t_k, q)$ を，4.1 節のようにして得られた用語候補とパッセージ，パッセージと検索クエリの 2 つの類似度である $SIM(t_k, d_i)$ と $SIM(d_i, q)$ を用いて，式（5）のように求める．

$$SIM(t_k, q) = \max_i((1 - dqw) \cdot \log(SIM(t_k, d_i)) + dqw \cdot \log(SIM(d_i, q))) \quad (5)$$

ここで dqw $(0 < dqw < 1)$ は，用語候補とパッセージ，パッセージと検索クエリの類似度の調整重みである．$SIM(t_k, q)$ はそれぞれの用語候補 t_k ごとに求め，この値を式 (4) の S_{t_k} に代用して，用語候補とする．この値だけでは順位が不安定なることが考えられるので，3.3 節で得られた 1 度目に得られた用語候補のスコアも用いることも提案する．すなわち，式 (6) に基づき $\hat{S_{t_k}}$ を求め，この値に基づいて用語候補を降順に並び替える．

$$\hat{S_{t_k}} = (1 - tqw) \cdot \log(S_{t_k}) + tqw \cdot SIM(t_q, q) \quad (6)$$

ここで，tqw $(0 \leq \text{tqw} \leq 1)$ は元のスコアと新しいスコアの調整重みである．

5. 実験

5.1 実験方法

検索クエリとして，3 つの文で構成される地名とカタカ

ナ語をそれぞれ 25 個の説明文を用意した．この説明文を検索クエリとして入力し，それに対する用語候補の出力を見る．上位 1000 個を出力し，正解となる用語の出現率，用語候補の中で上位何番目に正解となる用語が出力されたのかを，式 (7) で定義する平均逆順位 (MRR:Mean Reciprocal Rank) を用いて評価する．

$$MRR = \frac{1}{Q_N}\sum_{q=1}^{Q_N}\frac{1}{tRank_q} \tag{7}$$

ここで $tRank_q$ は検索クエリ q に対して，正解となる答えが用語候補として出力されたときの順位であり，Q_N は検索クエリの個数である．1000 件以内に見つからなかったときは $\frac{1}{tRank_q} = 0$ として計算する．

本実験では文書選択の際に上位何件のパッセージをとるかの n = 100，各文書から関連語をいくつとるかのパラメータ m = 100，用語候補の出力する数のパラメータ k = 1000 をとして，各検索クエリに対して用語候補を 1000 個出力し, その出力結果をリスコアリングすることで用語候補を並び替える実験を行った．

5.2 実験結果

5.2.1 クエリに含まれる単語の除外

これまでのシステムでは，検索クエリに含まれる単語が用語候補として出力されていた．本タスクでは用語はユーザがわからない語であるため検索クエリに含まれないと仮定できる．すなわち用語候補のうち検索クエリに含まれるものは用語でないといえる．従って，クエリに含まれる単語を用語候補から除外する．結果を表 1 と表 2 に示す．地名とカタカナ語の正解となる用語が候補として, 何番目に出力されたかを示している．

検索クエリと類似度の高い文書を選択し，関連語を抽出する際に，検索クエリに含まれる単語はスコアが高くなる．そのために検索クエリに含まれる単語が用語候補の上位を占めていた．検索クエリに含まれる単語を用語候補としないことで，カタカナ語用語検索では，上位 10 位までに含まれる語の数は変わらなかったが，地名用語検索においては，5 個から 10 個に増えた．地名用語検索において MRR が 0.0577 から 0.1674 に，カタカナ語用語検索において MRR が 0.0296 から 0.0505 と大きく精度が向上した．

5.2.2 用語候補とクエリの推定類似度を用いたリスコアリング

次に提案手法，すなわち用語候補とクエリの推定類似度を用いたリスコアリングの結果について述べる．表 3 に，地名用語検索の結果を示す．dqw と tqw を変化させたときの結果を示している．表 4 に，カタカナ語用語検索の結果を示す．

表 3，表 4 から，全体的にベースラインを上回る MRR となっていることがわかる．地名では $tqw = 0.8$，$dqw = 0.7$ のとき最大の MRR が得られ，0.2190 となった．カタカナ語では $tqw = 0.7$，$dqw = 0.1$ のとき最大の MRR が得られ，0.0865 であった．このときの結果を表 5，6 に示す．地名で

表 1 クエリに含まれる単語を除くリスコアリングの効果（地名）

地名	除外前	除外後
アメリカ	55	49
東京	122	114
中国	7	3
イギリス	25	21
京都	260	249
ドイツ	11	6
千葉	27	20
広島	10	5
スペイン	5	1
カナダ	592	583
群馬	16	10
八王子	488	479
エジプト	13	5
シドニー	6	1
メキシコ	19	14
名古屋	74	66
ラスベガス	6	2
成田	15	8
吉祥寺	39	27
シンガポール	128	121
静岡	*	*
イラン	*	*
モンゴル	*	*
高崎	8	4
熱海	29	23
正解出現率	88 % (22/25)	88 % (22/25)
MRR	0.0577	0.1674

正解となる用語の出現順位を表す
＊：順位が 1000 位以内に正解がなかった

表 2 クエリに含まれる単語を除くリスコアリングの効果（カタカナ語）

カタカナ語	除外前	除外後
ユーザー	164	158
コーパス	174	165
キーワード	9	5
カリキュラム	6	4
アルゴリズム	31	23
ノード	517	506
コスト	267	261
サンプル	*	*
ビット	808	799
ターゲット	843	838
プライド	257	252
コミュニケーション	16	12
スピーカー	116	107
マラソン	9	4
アーティスト	*	*
パスポート	*	*
サリン	6	3
スターバックス	*	*
デシベル	*	*
オリーブオイル	*	*
レントゲン	*	*
バイオリン	36	30
プリンター	96	86
コイル	*	*
マラリア	52	45
正解出現率	68 % (17/25)	68 % (17/25)
MRR	0.0296	0.0505

正解となる用語の出現順位を表す
＊：順位が 1000 位内に正解がなかった

表 3 用語候補のリスコアリングの効果（地名）

tqw ＼ dqw	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	**0.1885**	**0.1881**	**0.1877**	**0.1875**	**0.1881**	**0.1884**	**0.1887**	**0.1891**	**0.1686**
0.2	**0.1865**	**0.1870**	**0.1869**	**0.1886**	**0.1895**	**0.1898**	**0.1902**	**0.1897**	**0.1895**
0.3	**0.1692**	**0.1889**	**0.1891**	**0.1699**	**0.1721**	**0.1915**	**0.1928**	**0.1912**	**0.1908**
0.4	0.1515	0.1529	0.1518	**0.1726**	**0.1744**	**0.1944**	**0.1943**	**0.1947**	**0.1973**
0.5	0.1572	0.1561	0.1560	0.1563	0.1584	**0.1774**	**0.1978**	**0.1991**	**0.2008**
0.6	0.1417	0.1400	0.1387	0.1435	0.1464	0.1481	**0.1787**	**0.1823**	**0.1766**
0.7	0.1082	0.1275	0.1314	0.1419	**0.1691**	**0.1763**	**0.1874**	**0.1764**	0.1651
0.8	0.0957	0.0942	0.1332	0.1498	0.1545	**0.1705**	**0.2190**	**0.1882**	**0.2174**
0.9	0.0649	0.0725	0.0760	0.1090	0.1282	0.1394	**0.1768**	**0.1897**	**0.1888**
1.0	0.0252	0.0400	0.0408	0.0502	0.0563	0.0509	0.0417	0.0388	0.0401
Baseline	0.1674								

表 4 用語候補のリスコアリングの効果（カタカナ語）

tqw ＼ dqw	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	**0.0510**	**0.0509**	**0.0510**	**0.0510**	**0.0511**	**0.0510**	**0.0510**	**0.0510**	**0.0509**
0.2	**0.0536**	**0.0535**	**0.0532**	**0.0530**	**0.0509**	**0.0509**	**0.0509**	**0.0575**	**0.0576**
0.3	**0.0536**	**0.0536**	**0.0536**	**0.0534**	**0.0531**	**0.0529**	**0.0575**	**0.0575**	**0.0576**
0.4	**0.0538**	**0.0536**	**0.0536**	**0.0515**	**0.0511**	**0.0578**	**0.0563**	**0.0556**	**0.0544**
0.5	**0.0643**	**0.0579**	**0.0553**	**0.0511**	**0.0512**	**0.0564**	**0.0566**	**0.0568**	**0.0551**
0.6	**0.0639**	**0.0621**	**0.0554**	**0.0511**	**0.0510**	**0.0569**	**0.0552**	**0.0575**	**0.0577**
0.7	**0.0865**	**0.0846**	**0.0819**	**0.0611**	**0.0562**	**0.0532**	**0.0533**	**0.0591**	**0.0681**
0.8	**0.0726**	**0.0776**	**0.0734**	**0.0737**	**0.0540**	0.0425	0.0464	**0.0555**	**0.0581**
0.9	0.0443	0.0456	0.0441	0.0403	0.0347	0.0329	0.0347	0.0443	0.0451
1.0	0.0153	0.0217	0.0242	0.0188	0.0153	0.0136	0.0130	0.0128	0.0110
Baseline	0.0505								

値は MRR
赤字は Baseline よりも MRR が大きいもの

表 5 最も MRR が高い結果の用語候補順位 (地名)

地名	tqw=0.8,dqw=0.7
アメリカ	53
東京	109
中国	6
イギリス	107
京都	555
ドイツ	2
千葉	17
広島	12
スペイン	12
カナダ	555
群馬	2
八王子	540
エジプト	7
シドニー	1
メキシコ	4
名古屋	87
ラスベガス	1
成田	1
吉祥寺	10
シンガポール	128
静岡	*
イラン	*
モンゴル	*
高崎	2
熱海	12
正解出現率	88 % (22/25)
MRR	0.2190

正解となる用語の出現順位を表す
＊：順位が 1000 位以内に正解がなかった

表 6 最も MRR が高い結果の用語候補順位 (カタカナ語)

カタカナ語	tqw=0.7,dqw=0.1
ユーザー	244
コーパス	178
キーワード	5
カリキュラム	1
アルゴリズム	57
ノード	244
コスト	218
サンプル	*
ビット	433
ターゲット	379
プライド	351
コミュニケーション	16
スピーカー	86
マラソン	3
アーティスト	*
パスポート	*
サリン	3
スターバックス	*
デシベル	*
オリーブオイル	*
レントゲン	*
バイオリン	10
プリンター	28
コイル	*
マラリア	23
正解出現率	68 % (17/25)
MRR	0.0865

正解となる用語の出現順位を表す
＊：順位が 1000 位以内に正解がなかった

は候補の上位 10 位以内に出現した語の数は変わらなかったが，3 位以内の出現数は 4 個から 6 個へと増えている．カタカナ語では上位 10 位以内の出現数は 4 個から 5 個へ，3 位以内も 1 個から 3 個へと増えている．

これらの結果より，地名，カタカナ語ともに用語候補と検索クエリの推定類似度スコア $SIM(t_k, q)$ を利用する効果があること，そのスコア重み tqw は 0.7 から 0.8 程度のときに精度改善が大きいことがわかった．なお，元のスコアを統合せず $SIM(t_k, q)$ のみを用いた場合（式 (6) で $tqw = 1$ とした場合）は，Baseline よりも MRR 値が大きく低下している（表 3，表 4 最下段）．このことは，推定類似度スコア単体を使うのは効果がないことを示している．用語候補と検索クエリの推定類似度スコア $SIM(t_k, q)$ を求めるときに用いた用語とパッセージ，パッセージと検索クエリの統合重み dqw（式 (5)）に関しては，地名では 0.6 から 0.8，カタカナ語では 0.1 から 0.3 のとき効果が大きく，傾向が異なった．これらのパラメータの最適な値の決定法については今後検討していく予定である．

6. 結論

非構造化文書を検索対象とした用語検索において，用語候補のリスコアリングの手法を提案した．まず，検索クエリに含まれる単語が用語候補として出力されていた問題を解決した．これにより地名タスクで，MRR が 0.1674，カタカナ語タスクで 0.0505 と精度を向上することができた．次に用語候補と検索クエリの推定類似度を求め，その類似度を基にスコアを再度与えることで，用語候補を並び替える手法を提案した．MRR の向上を確認し，本提案手法は有効であることがわかった．提案方法は，用語候補とパッセージ，パッセージと検索クエリの類似度を結合して用語候補とクエリの類似度とするものであるが，この統合方法については，さらなる改善の余地があることを確認した．

謝辞 本研究は科研費（課題番号 25330368）の助成を受けた．文書選択システムの構築には GETA[11] を使用した．

参考文献

[1] 粟飯原俊介，長尾　真，田中久美子：意味的逆引き辞書『真言』，言語処理学会第 19 回年次大会発表論文集，pp. 406–409 (2013).
[2] 谷河息吹，馬　青，村田真樹：Deep Belief Network を用いた関連語・周辺語からの検索用語の予測，言語処理学会第 20 回年次大会発表論文集，pp. 547–550 (2014).
[3] Ma, Q., Tanigawa, I. and Murata, M.: Retrieval Term Prediction Using Deep Belief Networks, *The 28th Pacific Asia Conference on Language, Information and Computing (Paclic 28)*, pp. 338–347 (2014).
[4] 谷河息吹, 馬　青, 村田真樹：検索語の予測における Deep Learning と従来の機械学習との比較，言語処理学会第 21 回年次大会発表論文集，pp. 684–687 (2015).
[5] 森田直樹，南條浩輝，山本凌紀，馬　青：説明文を入力とした非構造化文書からの用語検索の検討，情報処理学会研究報告（第 17 回音声言語シンポジウム），2015-SLP-109, No. 16 (2015).
[6] 前川喜久雄：言語研究における自発音声，日本音響学会研究発表会講演論文集 (2001).
[7] Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y. and Itou, K.: Construction of a test collection for spoken document retrieval from lecture audio data, *IPSJ-Journal*, Vol. 50, No. 2, pp. 501–513 (2009).
[8] 西尾友宏，南條浩輝，吉見毅彦：講演音声ドキュメント検索のための擬似適合性フィードバック，情報処理学会論文誌，Vol. 55, No. 15, pp. 1573–1584 (2014).
[9] 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結合処理とその評価，情報処理学会論文誌 Vol.44 No.SIG 8（TOD 18），pp. 78–91 (2003).
[10] 南條浩輝，弥永裕介，吉見毅彦：広域文書類似度と局所文書類似度を用いた講演音声ドキュメント検索，情報処理学会論文誌，Vol. 53, No. 6, pp. 1654–1662 (2012).
[11] 汎用連想計算エンジン GETA：`http://geta.ex.nii.ac.jp`.