Vol.2010-NL-196 No.13
Vol.2010-SLP-81 No.13
2010/5/28

Spoken Term Detection のための テストコレクション構築とベースライン評価

西崎博光[†1]，胡新輝[†2]，南條浩輝[†3]，伊藤慶明[†4]，
秋葉友良[†5]，河原達也[†6]，中川聖一[†5]，松井知子[†7]，
山下洋一[†8]，相川清明[†9]

TREC において Spoken Document Retrieval（SDR:音声ドキュメント検索）の Track が 1996 年～2000 年に設定され，2006 年には NIST を中心に Spoken Term Detection（STD: 音声検索語検出）タスクが設定され，以降，海外では盛んに SDR，STD に関する研究が行われるようになった．情報処理学会音声言語情報処理研究会（SIG-SLP）で国内の音声ドキュメント処理研究の推進・活性化を目的として 2006 年に音声ドキュメント処理ワーキンググループを立ち上げ，これまでに SDR 評価用テストコレクションの構築を進めてきた．これに続き 2008 年から STD の評価用テストコレクションの構築を開始し，2009 年 10 月に SIG-SLP において中間報告を行った．本稿ではこれまでに構築してきたテストコレクションについての解説とベースライン評価について述べる．

Development of Test Collection for Spoken Term Detection and Its Baseline Evaluation

Hiromitsu Nishizaki[†1], Xinhui Hu[†2], Hiroaki Nanjo[†3], Yoshiaki Itoh[†4], Tomoyosi Akiba[†5], Tatsuya Kawahara[†6], Seiichi Nakagawa[†5], Tomoko Matsui[†7] ,Yoichi Yamashita[†8] and Kiyoaki Aikawa[†9]

Spoken Document Retrieval (SDR) was dealt with in one of tracks of TREC from 1996 to 2000. NIST supplied a task for Spoken Term Detection (STD) in 2006. Many researchers have been conducted as for SDR and STD after these projects. A working group for spoken document processing of SIG-SLP (Spoken Language Processing) in IPSJ also aimed to activate the researches for spoken document processing, and developed a test collection for SDR so far. The working group started to develop a test collection for STD last year, then, we presented the progress report at the SIG-SLP on Oct. 2009. This paper describes detail of the test collection and its baseline evaluation.

1. はじめに

近年，パソコンのマルチメディア環境，高速なインターネット，大容量の HDD-ビデオレコーダが普及し，撮り溜めた TV 放送，長期間に渡り録画した家庭用ビデオ，講義や教材用ビデオ，動画サイトでのビデオコンテンツなど，音声・動画を含んだマルチメディアコンテンツが増加・大容量化している．これに伴いこれらの大量のデータから見たい・聞きたい部分を検索したいという機能が求められるようになった．音声を含むデータに対しては，音声認識技術を適用してデータを検索する方式が有望であり，音声ドキュメント検索 (Spoken Document Retrieval: SDR) として既に様々な研究が行われてきている．

音声ドキュメント検索においては，ビデオや講義音声など音声を含むデータを音声ドキュメントと呼び，複数あるいは大量の音声ドキュメントがある中で，クエリに関連する内容を持つ音声ドキュメントを特定することを，アドホック（ad-hoc）音声ドキュメント検索，あるいは単に音声ドキュメント検索と呼ぶ．

検索の基本的な枠組みでは，まず音声ドキュメントを単語ベースで音声認識しておき，その認識結果である単語列に対してテキスト検索 [1] の技術を用いてドキュメントを特定する．性能を評価する際，音声認識では音声ドキュメントの「質」（発話の丁寧さや，録音の精度など）に主に影響されるが，音声ドキュメント検索では音声ドキュメントの「質」だけでなく「長さ」，「正解箇所の数」にも影響される（例えば，1 時間の音声ドキュメントから探す場合，10 時間の音声ドキュメントから探す場合，正解が全く含まれていない場合，これらの検索性能の比較は困難である）．このため音声ドキュメント検索では共通の音声ドキュメント，クエリ，正解に基づいて評価が行われることが望ましい．

TREC (Text REtrieval Conference) においては，Spoken Document Retrieval の Track が 1996 年の TREC-6 から取り上げられ，TREC-7～9 を経て 2000 年まで行われた [2]．これを機に海外では音声ドキュメント検索に関しての研究が推進・活性化された．

日本においても情報処理学会音声言語情報処理研究会(SIG-SLP)において国内の音声ドキュメント処理研究の推進・活性化を目的として 2006 年に音声ドキュメント処理

[†1] 山梨大学 University of Yamanashi
[†2] 情報通信研究機構 National Institute of Information and Communications Technology
[†3] 龍谷大学 Ryukoku University
[†4] 岩手県立大 Iwate Prefectural University
[†5] 豊橋技科大学 Toyohashi University of Technology
[†6] 京都大学 Kyoto University
[†7] 統計数理研究所 The Institute of Statistical Mathematics
[†8] 立命館大学 Ritsumeikan University
[†9] 東京工科大学 Tokyo University of Technology

Vol.2010-NL-196 No.13
Vol.2010-SLP-81 No.13
2010/5/28

ワーキンググループ（SDPWG）を立ち上げ，これまでに SDR 評価用テストコレクションの構築を進め，テストコレクションを公開している [3]．これは，日本語話しコーパス（CSJ [4]）を所有している研究機関であれば，研究目的に限り利用可能である[1]．

アドホック音声ドキュメント検索によりクエリと関連あるドキュメント群が特定できたとしても，その結果は一覧性・確実性に欠け，最上位のドキュメントでさえ，あるキーワードが含まれているかは実際に聞いてみないと確かめられない．検索語（1 個以上の単語からなる言葉）が話されている箇所を音声ドキュメント中から特定（検索語検出：Spoken Term Detection: STD）したいというニーズは音声ドキュメント検索において不可避である．

また，検索語が音声認識システムにおける未知語になる場合は多く [5]，未知語の検索機能は不可欠である． 米国規格協会（NIST）[6] が 2006 年に STD を新たなテーマとして設定して以降，未知語の検出を重視した STD の研究が盛んに行われるようになり，近年の音声関連の国際会議（INTERSPEECH，ICASSP，ASRU 等）でも音声ドキュメント検索（SDR，STD）のセッションが組まれている．

このような状況を踏まえ，日本語アドホック音声ドキュメント検索用テストコレクションに続き，SDPWG は日本語 STD 用テストコレクションの構築を 2008 年度から開始した．テストコレクションのβ版を策定し，その中間報告を 2009 年 10 月の音声言語情報処理研究会で報告した [7]．研究会で頂いたコメント・意見を参考にしながら，STD テストコレクション改定作業を行っている．

本稿では，主に，策定した STD テストコレクションの解説とベースライン評価について述べる．テストコレクション構築に当たっての方針，進捗状況，現状の課題等について説明する．また，本テストコレクションでは，CSJ を対象としたセットとなっている．STD では音声認識技術を利用するため，CSJ に収録されている講演音声の音声認識のやり方についての指針についても述べる．最後に，テストコレクションを用いた研究紹介を簡単に行う．

本報告を機に，幅広い専門分野の研究者からご意見を伺い，最終的なテストコレクションに反映させたいと考えている．

2. 日本語 Spoken Term Detection テストコレクション

2.1 Spoken Term Detection 概要

STD は，ある特定の検索語（1 単語以上から成る）が，音声ドキュメント群中の，どのドキュメントのどの位置に含まれているのかを特定するタスクである．

日本語以外の STD テストコレクションは， NIST が中心となって策定され，コンペティションが開かれている [6]．アラビア語（近代標準語とレバノン系），中国語（標準語），英語（米語）の 3 種の言語について，放送ニュース音声(Broadcast News)，電話での会話音声(Telephone Conversation)，会議音声(Roundtable Meeting)の 3 種の音声タイプについて，1～3 時間のデータが用意されている．

我々が構築する日本語テストコレクションでは，CSJ 収録講演を利用し，話し言葉音声を対象とする．データセットとしては 2 種類，すなわち全講演セット（約 600 時間），コア講演セット（44 時間）用いる．NIST のテストコレクションと異なり，音声認識が難しい話し言葉を対象に，検索対象データ規模も比較的大きいものとなっている．

2.2 テストコレクション策定のための基本的な方針

既に SDR，STD 研究を行っている研究者だけでなく，新たに関連する研究を始める研究者も利用できるよう，「SDR/STD 研究・開発を行う多様な利用者を想定し，複数かつ単純な検索語と正解セットの提供」を目指すこととした．

研究向けに提供・公開するものは以下の通りである．

① 検索語セット，
② 正解データ，
③ 音声認識システムで使用した音声認識辞書，
④ 音声認識用音響モデルと言語モデル，
⑤ 音声ドキュメントの認識結果．

本テストコレクションでは，CSJ を対象としているため，CSJ を所有している個人・団体であれば，誰でも利用可能である．また，音声認識環境が整っていない研究者，企業の方にも利用して頂きたいため，音声認識結果はもちろんのこと，音響モデル，言語モデル，音声認識辞書の提供も行う．

2.3 テストコレクション

2.3.1 音声ドキュメント

音声ドキュメントは CSJ を検索対象ドキュメントと想定している．CSJ は実際の学会等の講演音声と模擬講演，朗読音声等から構成されており，全部で 3302 の音声データが収録されている．

本テストコレクションでは，このうち，学会講演 987 講演，模擬講演 1715 講演の計 2702 講演，約 604 時間の音声ドキュメントを検索対象データとし，これを全講演セットと呼ぶ．また，全 2702 講演のうち，「コア」と称する 177 講演（学会講演 70，模擬講演 107），約 44 時間の音声ドキュメントをコア講演セットとする．コア講演には，豊富な研究用情報（アノテーション）が付与されているため [8]，これらの情報を利用した STD 研究も行うことが可能である．

1 詳細は http://www.cl.ics.tut.ac.jp/~sdpwg/index.php?csjsdrtc で公開予定．

Vol.2010-NL-196 No.13
Vol.2010-SLP-81 No.13
2010/5/28

2.3.2 検索語セット

音声ドキュメントに対し，検索語セットは以下セット案を現段階で策定している．

(1) 既知語検索語セット（付録 A を参照）
  a) 全講演用既知語セット：全講演セットを対象とした 100 検索語
  b) コア講演用既知語セット：コア講演セットを対象とした 50 検索語
(2) 未知語検索語セット：
  a) 全講演用未知語セット：50 検索語
  b) コア講演用未知語セット：50 検索語（付録 B を参照）
(3) 簡易性能付コア講演用 50 検索語（文献 [7] を参照）

以下，それぞれの検索語セットを設定するに当っての考え方を紹介する．

(1) 既知語検索語セット

a) 全講演用既知語セット：100 検索語

全音声ドキュメント 2702 講演を対象として，音声認識システムの語彙に登録されている検索語を 100 個用意した．検索語は 1 語（形態素）以上からなる内容語とした．実際の検索場面を想定し，その検索語が検索において意味がある言葉であるよう TF・IDF 値も検索語を選定する際に考慮した．検索語長に性能が左右されるため，現在 4～11 モーラの検索語を各 12 個，12 モーラの検索語を 4 つ用意した．付録 A に全リストを示す．付録 A の tf は総出現数，df は文書頻度を表している．

b) コア講演用既知語セット：50 検索語

CSJ のコア講演を対象とし，比較的小規模な音声データを検索対象とした．検索語は全て全講演用既知語セットに含まれているものである．現在 4～12 モーラの検索語をモーラ数毎に 8 個程度用意した(9 モーラ以上は適切なものが少ないため 8 個未満)．付録 A にリストを示す．

(2) 未知語検索語セット

音声認識辞書に含まれていない形態素を未知語と呼ぶ．1 つの検索語セットは 1 つ以上の形態素から構成されている．検索語を構成している形態素のうち，半数以上が未知語のものを未知語検索語セットとして採用する．

未知語は音声認識辞書の作り方によって異なる．辞書の作成方法については，2.3.3 節で述べる．

a) 全講演用未知語セット：50 検索語

全講演セットの中で，出現頻度が 2 回以下のタームを選定した．600 時間の音声から 1，2 語を見つけるため非常に難しいタスクである．このセットについては，現在改訂作業中であるため（3.3 節(1)を参照），付録には掲載していない．

b) コア講演用未知語セット：50 ターム

音声認識辞書はコア講演以外の講演データから制定されているため，コア講演を対象とした未知語セットでは，比較的高頻度語を採用できる．コア講演用未知語セットを付録 B に示す．

表 1．奇数・偶数セットのデータ分布

	講演数（学会，模擬）	時間（学会，模擬）	発話数	単語数
奇数セット	1,255（458，797）	281h（127，154）	16.7 万文	343 万語
偶数セット	1,270（459，811）	287h（130，157）	17.0 万文	351 万語
合計	2,525（917，1,608）	568h（257，311）	33.7 万文	694 万語

(3) 簡易性能付コア講演用 50 ターム

コア講演中の 49 講演約 19 時間を音声ドキュメントとして，既に文献 [9]で未知語として検索性能評価を行っている．この実験で用いた検索語，音声ドキュメントセットと性能を提示することにより，簡易に自分のシステムの評価する枠組みを提供する．

2.3.3 音声認識

本コレクションを用いた STD を行う際に利用する音声認識の指針について説明する．本コレクションを用いて研究を行う場合には，下記指針に準拠頂きたい．

文献 [7]で述べた音声認識は，クローズドな条件で学習した音響モデル，言語モデル，音声認識辞書を利用していた．今回，オープンな条件で，CSJ 講演音声の音声認識を行うため，次のような方針をとっている．

- 各講演音声には固有の ID 番号が付与されている．この ID の下 1 桁の数字が奇数番号であるか偶数番号であるかによって，2 分割する．便宜上，偶数セット，奇数セットと呼ぶ．
- 偶数セット，奇数セット，それぞれから音響モデル，言語モデルを学習する．認識辞書は，奇数・偶数セット双方に共通で出現する語を登録する．ただし，コア講演は各モデルの学習や辞書作成に用いない．
- 偶数セットの音声認識は奇数セットから学習したモデルを，奇数セットの音声認識には偶数セットから学習したモデルを利用する

奇数・偶数セットのデータを表 1 に示す．以下，詳細に説明する．

(1) 音声認識辞書

音声認識辞書に登録されている形態素数は約 27000 語である．

形態素の定義は，Chasen with UniDic[10]（今回，Chasen 2.4.4，UniDic 1.3.9 を利用している）に従う．CSJ の転記データをこの条件で形態素解析し，形態素に分割した．

奇数セット，偶数セットに共通で出現し，かつコアを除く全講演での出現頻度が 3 回以上のものを辞書に登録している．

(2) 言語モデル

言語モデルは，奇数・偶数セット毎に，(1)で形態素解析したテキストから Palmkit 1.0.32 を用いて単語 trigram モデルを学習する．

Vol.2010-NL-196 No.13
Vol.2010-SLP-81 No.13
2010/5/28

表 2．CSJ 講演の音声認識率（1-best） [%]

	単語正解率	単語正解精度
全講演セット	74.05	69.23
コア講演セット	76.68	71.93

(3) 音響モデル

奇数・偶数セット毎に，triphone ベースの HMM を学習する．使用する音声特徴量は，MFCC（12 次）＋ΔMFCC＋ΔΔMFCC＋ΔPower（1 次）＋ΔΔPower の合計 38 次元である．ガウス分布の混合数は 32，総分布数は 3000 である．

(4) 音声認識

音声認識エンジンには，Julius を用いた．Julius のオプションにより，N-best 出力やコンフュージョンネットワークの出力を行うことができる．研究用に提供できる認識結果には，10-best 出力とコンフュージョンネットワークの出力を含む予定である．

前述した条件による全講演セットとコア講演セットの音声認識率（1-best）を表 2 に示す．

2.3.4 正解の定義

文献 [7]では，検索語に対する正解の定義として，以下の 2 通りを提案した．

案 1 : 正解音声区間の中心と検索語検出区間の中心との差が 0.5 秒以内（NIST 基準）

案 2 : 検索語検出区間が正解フレーズ（発話）に含まれているか否か

案 1 の NIST 基準とする場合，CSJ には単語境界情報がないため，正解音声区間の境界時刻情報を付与する必要がある．しかし，現在は，まだその情報を用意できていないため，正解判定は案 2 を採用する．今後，案 1 の情報も付与していく予定である．

STD のアプリケーションを考えたとき，検索語検出区間を含むフレーズを出力し，ユーザはそれを聞くことによって内容を確認する．従って，案 2 の正解判定基準は，検索語が含まれているフレーズを特定できれば良いという考え方に基づく．

CSJ の転記テキストには，0.2 秒のポーズによって分割された単位で，書き起こしが行われている．本稿では，この単位をフレーズと呼んでいる．案 2 の場合は，検索語がどのフレーズに含まれているのかを検出するタスクと等価である．

現状では，この正解フレーズリストを提供する．

3. ベースライン検索性能

本テストセットのベースライン評価を行った．ベースライン評価は，単純な検索手法に基づいて行う．以下，既知語セットと未知語セット毎のベースラインについて説明する

3.1 既知語検索語セット

既知語の検索語セットの評価は，2.3.3 節で述べた条件により音声認識を行い，単純に検索語が完全に音声認識されているフレーズを検索結果として取り出すことで行った．全講演・コア講演セット双方とも同じ評価基準で行う．

表 3．既知語セットベースライン評価 [%]

	単語レベルの完全一致			音素レベル完全一致（参考）		
	再現率	精度	F 値	再現率	精度	F 値
全講演セット	49.7	87.6	63.5	50.7	89.8	64.8
コア講演セット	54.1	93.6	68.6	55.3	96.9	70.4

付録 A に各検索語のヒット数（Hit），再現率（Rec.），検索精度（Prec.）を示している．なお，今回は，検索語が 1 つも検出されなかった場合は，再現率，検索精度ともに 0 としている．また，全体の評価をまとめたものを表 3 に示す．参考までに，音素レベルでの完全一致に基づく検索評価結果も表 3 に併記している．表 3 の数値は，検索語毎の平均値である．

全講演セットよりもコア講演セットの方が音声認識率は高いため，検索性能としてはコアの方が良い．検索精度が高いのは，検索語に誤って誤認識される単語の事例が少なかったためである（検索語が検出できなかった場合も精度を 0 としているため，表 3 の精度は悪いように見える点に注意）．

また，当然ではあるが，単語レベルの完全一致マッチングよりも音素レベルでの完全一致（検索語と音声認識結果を音素表記に変換しマッチングさせている）の方が結果は良い．

3.2 未知語検索語セット

未知語評価では，単語レベルの完全一致を行うことができない．そこで，未知語のベースライン評価では，音声認識結果の音素表記系列に対する連続 DP によるスポッティングを行う．DP で用いる距離尺度は，Edit Distance（編集距離）を用いる．置換，挿入，脱落誤りに対して，それぞれ距離 1 と定義する．あるフレーズの音素系列と検索語の音素系列を連続 DP させ，距離が一定の閾値以下の場合に，そのフレーズに検索語が含まれていたと判断する．

実験結果を表 4 に示す．参考までに，音素レベルでの完全一致に基づく評価も併記する．また，コア講演の音声認識を行う際に用いる言語モデルを 2 通り試した．“単語 trigram”は 2.3.3 節で説明した言語モデルである．もう一方の“音節 trigram”は，言語モデルの学習テキストを音節列表記（例えば，“フーリエ”なら“fu u ri e”）に変換し，音節単位での trigram を学習した言語モデルである．

連続 DP で用いている閾値は，F 値が最適になる値を採用している．再現率で 50%弱の検出性能を得ることができた．単純な DP 手法でも約半数弱の未知語を検出することができている．また，湧き出し検出誤りも想定よりも小さい．

Vol.2010-NL-196 No.13
Vol.2010-SLP-81 No.13
2010/5/28

表 4. コア講演用未知語セットベースライン評価 [%]

音声認識に用いた言語モデル	連続 DP			音素レベル完全一致（参考）		
	再現率	精度	F 値	再現率	精度	F 値
単語 trigram※1	48.5	45.2	46.8	7.0	28.0	11.2
音節 trigram※2	62.6	56.8	59.6	7.2	26.0	11.3

※1 単語 trigram の音節認識率：正解率 86.5%，正解精度 83.0%

※2 音節 trigram の音節認識率：正解率 81.8%，正解精度 77.4%

単語 trigram を使った音声認識結果よりも，連続音素認識に近い音節 trigram を用いた音声認識結果の方が，より実際の発音に近い音素系列を出力することができる．単語の制約を外すだけでも，6 割強の未知語検出が可能なっている．

3.3 考察と課題

(1) 検索対象の音声ドキュメントについて

コア講演用未知語セットではベースラインの評価が F 値で約 47%であり，適当な難易度のタスクとなっている．しかし，全講演用未知語セットの場合，未知語を全 2702 講演から探すのは難しいすぎるという指摘もあり，1 講演から数箇所の正解を探すタスクで良いとの考え方もある．そこで， 以下のような，検索対象講演の時間数を変化させた性能評価方法を検討中である．

- 未知語検索語を含む講演　　→　ターゲット講演
- 未知語検索語を含まない講演　→　非ターゲット講演
- １つの未知語検索語に対しターゲット講演性能だけでなく，非ターゲット講演を追加した時の性能を提示(１時間，10 時間等)

(2) 正解セットについて

現状では，2.3.4 節で述べたように，案 2 を採用している．今後は，NIST の評価基準にも合わせるように，正解音声区間のラベリングを行う予定である．

(3) 検索性能評価について

①ベースライン性能

本報告により，現状のテストセットにおけるベースラインを示した．ベースラインを凌駕するための様々な研究アイデアが出てくることを期待したい．例えば，未知語セットについては，単語認識を行うよりも，連続音節（音素）認識を行った方が STD の性能を改善させることができる．

②評価指標

検索指標としては，現在の再現率，精度，F 値の他に NIST の STD 評価基準である ATWV（Actual Term Weighted Value）[6]を提示したいと考えている．

3.4 テストコレクションの公開

本テストコレクションは，近日 Web サイトにて公開予定である．公開サイトは，「http://www.cl.ics.tut.ac.jp/~sdpwg/」の予定である．アドホック検索のテストコレクションと合わせて利用されたい．

3.3 節で挙げた課題については，今後随時対応でき次第，公開していく予定である．

4. STD テストコレクションを使った研究例

2009 年 10 月に本テストコレクションの β バージョンを公開して以来，いろいろな研究機関で利用されている．

例えば，2010 年 2 月に豊橋技術科学大学で開催された音声ドキュメント処理ワークショップ（http://www.cl.ics.tut.ac.jp/~sdpwg/index.php?sdpws2010_program）では，STD に関する発表が 4 件，2010 年 3 月に電気通信大学で開催された日本音響学会春季研究発表会でも 4 件の発表があり，関心の高さが伺える．

5. まとめ

情報処理学会音声言語情報処理研究会における音声ドキュメント処理ワーキンググループが SDR 評価用テストコレクションの構築に続き STD の評価用テストコレクションの構築を進めている．本稿ではこのテストコレクションの構築とベースライン評価について述べた．

テストコレクションは，近日公開される予定である．今後，幅広い専門分野の研究者の方々に SDR/STD テストコレクションを利用して頂き，国内での本分野の活性化に寄与できると幸いである．

参考文献

1) 江口浩二, “情報検索のための確率的言語モデルに関する動向と課題,” 電子情報通信学会論文誌, Vol.J93-D, No.3, pp.157-169, 2010.
2) J. Garofolo, et al. ,“The TREC spoken document retrieval track: A success story,” Ninth Text Retrieval Conference (TREC-9), NIST, 2000.
3) Tomoyosi Akiba et al., “Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data,” IPSJ Journal, Vol. 50, No. 2, 1234–1245, 2009.
4) 国立国語研究所：日本語話し言葉コーパス，http://www.kokken.go.jp/katsudo/seika/corpus/
5) C. Parada et al. ,”Query-by-Example Spoken Term Detection For OOV Terms,”, in Proc. of the ASRU 2009, pp.404-409, 2009.
6) 2006 Spoken Term Detection Evaluation: http://www.itl.nist.gov/iad/mig/tests/std/2006/index.html.
7) 伊藤慶明他，”音声中の検索語検出のためのテストコレクション構築－中間報告－”，情報処理学会研究報告，Vol. 2009-SLP-78， No.4，8 pages，2009.
8) 前川喜久雄，“『日本語話し言葉コーパス』の外観”，Version 1.1，http://www.kokken.go.jp/katsudo/seika/corpus/releaseinfo/040/overview.pdf
9) 伊藤慶明他，“語彙制限のない音声ドキュメント検索における複数サブワードの統合－検索語彙に依存した検索性能推定指標の導入－,” 情処会論文誌, Vol. 50, 2, pp. 524‐533, 2009.
10) 形態素解析辞書 UniDic：http://www.tokuteicorpus.jp/dist/

Vol.2010-NL-196 No.13
Vol.2010-SLP-81 No.13
2010/5/28

付録 A　既知検索語セット：

全講演用 100 検索語（左），コア講演用 50 検索語（右）と検索性能

	For all spoken documents					For core documents				
WORD	tf	df	Hit	Rec.	Prec.	tf	df	Hit	Rec.	Prec.
12 モーラ	4 個					1 個				
国立 国語 研究 所	35	19	16	45.7	100.0	7	5	4	57.1	100.0
統計 数理 研究 所	10	8	5	50.0	100.0					
大 語彙 音声 認識	6	5	4	66.7	100.0					
談話 セグメント 境界	15	1	11	73.3	100.0					
11 モーラ	12 個					1 個				
音声 対話 システム	89	34	71	79.8	100.0	8	4	4	87.5	100.0
学習 指導 要領	59	18	44	72.9	97.7					
教師 なし 話者 適応	30	6	17	50.0	88.2					
コンビニエンス ストアー	30	21	18	60.0	100.0					
ウエーブレット 変換	25	3	5	20.0	100.0					
機械 翻訳 システム	24	7	13	50.0	92.3					
東京 ディズニー ランド	21	7	10	47.6	100.0					
線形 予測 分析	27	8	17	63.0	100.0					
人工 知能 学会	10	5	4	40.0	100.0					
自律 神経 失調	10	5	5	50.0	100.0					
トレード オフ の 関係	9	8	6	66.7	100.0					
環境 音 の 識別	8	1	0	0.0	0.0					
10 モーラ	12 個					2 個				
TF IDF	79	18	50	60.8	96.0					
ニューラル ネットワーク	69	15	38	53.6	97.4					
重要 文 抽出	45	12	32	66.7	93.8	7	5	5	71.4	100.0
阪神 大 震災	33	15	23	66.7	95.7					
サザン オール スターズ	32	4	13	40.6	100.0					
シドニー オリンピック	29	16	16	51.7	93.8					
最大 エントロピー	19	5	11	57.9	100.0					
天体 望遠鏡	14	3	4	28.6	100.0					
ワンパス トライグラム	13	1	3	23.1	100.0	13	1	3	23.1	100.0
宇宙 戦艦 ヤマト	11	1	0	0.0	0.0					
羊 たち の 沈黙	10	3	4	40.0	100.0					
ドル トン の 原子 説	7	1	0	0.0	0.0					
9 モーラ	12 個					6 個				
形態 素 解析	159	76	134	84.3	100.0	9	3	6	66.7	100.0
主 成分 分析	90	31	45	50.0	100.0					
原子 力 発電	44	10	29	59.1	89.7					
プロスペクト 理論	34	4	23	67.7	100.0					
音素 認識 率	32	7	27	81.3	96.3	16	1	11	68.8	100.0
ワーキング ホリデー	27	10	14	51.9	100.0					
ツー パス デコーダー	18	6	11	61.1	100.0	14	2	9	64.3	100.0
ヤ コ ビ 適応 法	16	1	0	0.0	0.0					
ヒト ゲノム 計画	14	3	0	0.0	0.0					
シラブル の 構造	9	1	7	77.8	100.0	9	1	7	77.8	100.0
キー ワード 抽出	14	5	11	64.3	81.8	5	1	4	80.0	100.0
エベレスト 街道	6	1	2	33.3	100.0	6	1	2	33.3	100.0
8 モーラ	12 個					8 個				
基本 周波 数	287	61	226	78.4	99.6	21	1	19	90.5	100.0
情報 検索	146	51	127	82.2	94.5	19	7	16	84.2	100.0
パープレキシティ	121	25	95	76.0	96.8	26	5	21	80.8	100.0
絶対 音感	86	3	29	33.7	100.0	38	1	22	57.9	100.0
就職 活動	72	29	42	58.3	100.0	5	4	3	60.0	100.0
インターラクション	61	7	52	78.7	92.3	18	3	14	66.7	85.7
英語 の 勉強	33	18	30	90.9	100.0					
平家 物語	22	4	11	50.0	100.0					
田園 都市 線	25	7	7	28.0	100.0					

Vol.2010-NL-196 No.13
Vol.2010-SLP-81 No.13
2010/5/28

	For all spoken documents					For core documents				
	tf	df	Hit	Rec.	Prec.	tf	df	Hit	Rec.	Prec.
沖縄 の 文学	19	1	6	31.6	100.0					
ウィザード オブ オズ	15	7	1	6.7	100.0	8	2	1	12.5	100.0
中央 林間	12	4	1	8.3	100.0	9	2	1	11.1	100.0
7 モーラ	12 個					8 個				
イントネーション	199	50	190	93.5	97.9	32	5	30	93.8	100.0
NHK	172	95	95	52.3	94.7	20	8	9	45.0	100.0
ATR	186	81	105	48.4	85.7	13	7	6	38.5	83.3
機械 翻訳	92	25	71	76.1	98.6	15	2	11	73.3	100.0
京都 大学	54	35	25	37.0	80.0					
交通 の 便	33	26	27	78.8	96.3	7	4	6	85.7	100.0
混合 重み	30	10	15	50.0	100.0	12	2	5	41.7	100.0
東南 アジア	44	28	4	9.1	100.0					
有声 休止	26	2	8	30.8	100.0	23	1	7	30.4	100.0
遅延 和 アレー	23	4	7	30.4	100.0					
ラジオ 体操	18	9	11	61.1	100.0					
お しゅうとめ さん	15	6	7	46.7	100.0	9	2	4	44.4	100.0
6 モーラ	12 個					8 個				
商店 街	208	75	138	63.9	96.4	21	9	10	47.6	100.0
大学 院	151	104	86	51.0	89.5	10	8	3	20.0	66.7
アナウンサー	132	45	86	59.9	91.9	27	5	18	66.7	100.0
研究 室	98	79	36	35.7	97.2	6	6	3	50.0	100.0
留学 生	95	32	74	70.5	90.5	16	3	11	62.5	90.9
ペット ボトル	36	19	23	61.1	95.7	9	1	5	55.6	100.0
世界 遺産	26	14	11	38.5	90.9					
オレンジ 色	23	18	22	91.3	95.5					
ADA ブースト	27	3	6	22.2	100.0					
産 婦人 科	15	10	1	6.7	100.0					
貝殻 虫	13	2	9	69.2	100.0	12	1	9	75.0	100.0
調音 地図	12	1	4	33.3	100.0	12	1	4	33.3	100.0
5 モーラ	12 個					8 個				
お 婆 ちゃん	244	100	202	72.5	87.6	27	6	20	70.4	95.0
アルバイト	234	102	224	86.8	90.6	11	6	11	90.9	90.9
ダイエット	166	35	143	73.5	85.3					
クラシック	85	29	62	65.9	90.3					
ラーメン 屋	60	25	37	58.3	94.6	10	3	5	50.0	100.0
コンクール	39	14	30	74.4	96.7	15	2	13	80.0	92.3
鼻 濁音	39	6	19	48.7	100.0	21	2	9	42.9	100.0
レントゲン	34	15	22	61.8	95.5					
ドラえもん	30	13	18	50.0	83.3					
予測 誤差	28	6	16	57.1	100.0	16	2	13	81.3	100.0
豊島 園	25	5	10	40.0	100.0	21	2	8	36.4	100.0
連想 語	16	1	0	0.0	0.0	16	1	0	0.0	0.0
4 モーラ	12 個					8 個				
地下 鉄	134	70	85	61.9	97.7	15	9	9	60.0	100.0
純音	119	21	79	56.3	84.8	15	3	4	20.0	75.0
世田谷	115	39	58	48.7	96.6	22	7	8	36.4	100.0
富士 山	92	38	0	0.0	0.0	13	6	0	0.0	0.0
鎌倉	65	26	68	84.6	80.9	13	4	8	61.5	100.0
青山	58	16	26	43.1	96.2	35	1	15	42.9	100.0
大田 区	45	11	22	11.1	22.7					
練馬 区	40	10	21	52.5	100.0	25	4	10	40.0	100.0
コサイン	30	18	11	30.0	81.8					
ベルギー	21	8	5	19.1	80.0	11	2	4	36.4	100.0
アイヌ 語	20	1	0	0.0	0.0					
ヤシ の 木	18	12	17	83.3	88.2					

※網掛けの単語は，文献 [7] で報告したリストから変更した検索語

Vol.2010-NL-196 No.13
Vol.2010-SLP-81 No.13
2010/5/28

付録 B　未知検索語セット：コア講演用 50 検索語と検索性能

モーラ	WORD	tf	df	Hit	Rec.	Prec.
13	石川島* 造船* 所	1	1	0	0.0	0.0
12	コンテキスト ディペンデント*	5	1	8	60.0	37.5
11	クリント* イーストウッド*	2	1	0	0.0	0.0
10（3個）	ボスニア・ヘルツェゴビナ*	1	1	1	100.0	100.0
	ユニバーサル* スタジオ	3	2	1	33.3	100.0
	ホテル ニューハンプシャー*	2	1	0	0.0	0.0
9（6個）	春桜亭* 円紫*	1	1	6	100.0	16.7
	談洲楼* 焉馬*	1	1	0	0.0	0.0
	竹取* 物語	5	1	3	60.0	100.0
	高島平* 駅	2	1	8	100.0	25.0
	タンチョウ* の 飛来* 地	2	1	2	100.0	100.0
	チトー* 大統領	2	1	2	100.0	100.0
8（8個）	スティーブン* キング	1	1	1	100.0	100.0
	名犬 ラッシー*	2	1	0	0.0	0.0
	駒沢* 公園	8	1	1	12.5	100.0
	まほろば* 連邦	5	1	3	60.0	100.0
	南大泉*	5	1	3	60.0	100.0
	伊曽保* 物語	2	1	7	100.0	28.6
	営団 赤塚*	1	1	2	0.0	0.0
	キラウエア* 火山	5	1	3	60.0	100.0
7（7個）	ユーゴスラビア*	7	1	1	14.3	100.0
	代々木 上原*	2	2	0	0.0	0.0
	釧路* 湿原*	3	2	2	33.3	50.0
	コザクラ* インコ	4	1	3	50.0	66.7
	奄美* 大島	1	1	2	0.0	0.0
	オスマン* トルコ	6	1	4	50.0	75.0
	奥穂高* 岳	1	1	0	0.0	0.0
6（8個）	光が丘*	11	3	12	90.0	75.0
	ノーベル* 賞	2	1	2	0.0	0.0
	西日暮里*	7	1	4	57.1	100.0
	常盤平*	7	1	9	100.0	77.8
	拝島* 駅	12	1	8	58.3	87.5
	本駒込*	3	1	0	0.0	0.0
	メーンランド*	2	1	0	0.0	0.0
	バンクーバー*	4	2	1	25.0	100.0
5（8個）	アルバニア*	9	1	85	100.0	10.6
	三河島*	3	1	17	100.0	17.6
	美堀町*	4	1	9	75.0	33.3
	屈斜路* 湖*	3	1	1	33.3	100.0
	スリーピー*	7	1	9	57.1	44.4
	ワイコロア*	6	1	40	83.3	12.5
	九品仏*	6	1	5	16.7	20.0
	NATO* 軍	3	1	39	66.7	5.1
4（7個）	那覇* 港	2	1	11	100.0	18.2
	ネパール*	28	1	6	7.4	33.3
	安保理*	5	1	38	60.0	7.9
	ヒマラヤ*	4	2	25	75.0	12.0
	知床*	15	1	175	71.4	5.7
	八潮* 市	7	1	60	14.3	1.7
	ケベック*	8	1	3	42.9	100.0

※ *マークがついている単語は未知語

※ 網掛けの単語は，文献 [7]で報告したリストから変更した検索語