Vol.2010-SLP-84 No.18
2010/12/21

NTCIR-9 SpokenDoc: 音声検索語検出と音声ドキュメント検索の評価枠組の設計

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

情報処理学会音声言語処理研究会 (SIG-SLP) の音声ドキュメント処理ワーキンググループは、音声ドキュメント検索のテストコレクションを構築し公開をしてきた。本稿では、これらのテストコレクションを基に情報検索評価型ワークショップ NTCIR-9 で実施することとなった、音声情報検索タスク “IR for SpokenDocuments (SpokenDoc)” のタスク設計について報告する。

NTCIR-9 SpokenDoc: Designing an Evaluation Framework for Spoken Term Detection and Spoken Document Retrieval

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

The Spoken Document Processing Working Group, which is organized in special interest group of spoken language processing (SIG-SLP), information processing society of Japan, have developed and released the test collections for evaluation of spoken document retrieval. This paper describes about the task design of “IR for Spoken Documents (SpokenDoc)” task, which will be carried out at NTCIR-9, an information retrieval evaluation workshop, based on the test collections.

1. はじめに

音声・画像・ビデオの記録・編集機器の拡大、およびインターネットをはじめとする情報通信網の発展により、誰でも気軽にコンテンツを作成・公開することが可能となり、マルチメディアコンテンツの増大が加速している。これらのコンテンツには、ファイル名やタイトル以外にはメタデータが付与されていないことが多く、従来のテキストベースの検索技術だけでは、目的のコンテンツにたどり着くことは困難である。一方、話し言葉を含むコンテンツの場合には、大語彙連続音声認識技術を利用することで言語情報を抽出し、テキスト検索技術を利用した検索が可能である。このような音声言語情報を対象とした検索技術は「音声ドキュメント検索」と呼ばれ、マルチメディアコンテンツの情報爆発時代に必要不可欠な技術になると考えられる[1)]。

このような背景のもと、情報処理学会音声言語処理研究会 (SIG-SLP) の音声ドキュメント処理ワーキンググループは、音声ドキュメント検索のテストコレクションを構築し、研究利用を条件に公開をしている。この度、これらのテストコレクションをベースとした音声ドキュメント検索タスクを、評価型ワークショップ NTCIR-9[2)] のコアタスクの一つとして実施することになった。NTCIR-9 音声ドキュメント検索タスク “IR for Spoken Documents (SpokenDoc)”[3)] では、実際の検索環境に近い条件 (自由発話音声を対象、未知語を含む検索課題) における共通タスクを設定し、日本で最初の大規模な音声ドキュメント検索タスクの評価を行う。

2. 関連研究

情報検索の分野で、開発したシステムをある程度限定した設定のもとで定量的に評価するためのデータセットをテストコレクションと言う。テキストを対象とした情報検索の分野で

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

は、TREC や NTCIR などの評価型ワークショップでの活動を中心に、多くのテストコレクションが積極的に構築されてきた。

音声ドキュメント検索のテストコレクションとしては、米国規格協会 (NIST) 主催の情報検索評価型会議 TREC(Text REtrieval Conference)[4] にて、最初に大規模な評価が行われた。1996 年の TREC-6 SDR Track では、Known Item Retrieval のテストコレクションが構築された。これは今日での、Spoken Term Detection タスクに相当する。その後、1997 年の TREC-7 から 1999 年の TREC-9 において、ニュース音声を対象とした内容検索のテストコレクションが構築された。最終的には、557 時間、約 2 万文書を対象としたテストコレクションが構築された[5]。

TREC SDR Track を引き継ぎ、言語横断検索に力を入れた欧州の評価型会議である CLEF(Cross Language Evaluation Forum)[6] では、ニュース音声を対象とした TREC SDR Track のデータを用いて言語横断の音声ドキュメント内容検索の評価 CLEF CL-SDR(Cross-Language Spoken Document Retrieval) が 2003〜2004 年に行われた。その後、インタビュー音声を対象とした検索タスク CLEF CL-SR(Cross-Language Speech Retrieval) が 2005〜2007 年に行われている[7]。

2006 年に米国規格協会 (NIST) が STD を新たなタスクに設定し[8]、共通の評価基盤が設定され、これを契機に STD 研究が活性化した。対象データは 3 時間程度の、ニュース音声、電話での会話、会議音声である。

日本においては、筆者らが属する情報処理学会音声言語情報処理研究会 (SIG-SLP) の音声ドキュメント処理ワーキンググループ[9] において、SDR および STD のテストコレクション構築が進められてきた。

3. タスク概要

音声ドキュメント処理 WG では、これまでに 2 種類のテストコレクション「CSJ 音声中の検索語検出テストコレクション」[10),11)]、「CSJ 音声ドキュメント内容検索テストコレクション」[12] を構築してきた。NTCIR-9 SpokenDoc では、これらのテストコレクションを拡張して、次の 2 つのサブタスクを行う。

Spoken Term Detection (STD) 単語あるいは数単語の列をクエリとして与え、音声ドキュメント中からクエリが現れる位置を特定するタスク。計算効率 (索引に必要な空間コスト、検索時間コスト、など) と検索性能 (精度と再現率) の 2 つの観点から評価を行う。

Spoken Document Retrieval (SDR) 文やキーワードリストなどの比較的長いクエリを与え、クエリと関連するパッセージあるいは講演を見つけるタスク。テキストを対象とした検索における内容検索に相当するが、検索対象が音声データである点が異なる。

4. 資　　源

4.1 音声ドキュメント

SpokenDoc タスクでは、国立国語研究所から公開されている「日本語話し言葉コーパス (Corpus of Spontaneous Japanese; CSJ)」[13] を音声ドキュメントとして利用する。タスク参加者は、CSJ を各自で保有している必要がある。CSJ に含まれるデータのうち、学会講演および模擬講演を検索対象文書とする。両講演データを合わせると、2702 講演となる。STD サブタスクでは、コアと呼ばれるサブセット 177 講演を対象とした評価も行う。

4.2 書 き 起 し

音声ドキュメント検索の典型的な手法は、まず対象音声データに対して音声認識を適用し、その結果として得られるテキストデータに対してテキストベースの検索手法を適用する、という流れに沿っている。前段の音声認識性能と、後段の検索性能を区別して評価するために、オーガナイザはリファレンスとして参加者共通で利用できる音声認識結果を提供する。これによって、参加者は自前で音声認識を行わなくてもタスクに参加可能となる。特に、テキスト処理分野の研究者が、誤りを含むテキストに対する検索手法に焦点を絞った参加が可能である。また、共通の認識性能のもとで、参加者の検索手法の比較が可能になる。

参加者が利用可能な書き起しデータは、以下の通りである。

- 人手書き起し
  CSJ には人手による書き起しテキストが附属している。正解人手書き起しを対象とした検索の性能を調べることで、検索性能の上限値を求めることができる。
- 参照自動書き起し
  オーガナイザから、以下の 2 種類の音声認識結果を提供する。
  - 単語ベース音声認識による書き起し
    自動書き起しを得るための音声認識の認識辞書には、CSJ の人手書き起しテキストを形態素解析して得られる約 27,000 語の形態素 (単語) を用いる。形態素の定義は、Chasen with UniDic-1.3.9 に従う。また、言語モデルには単語 3-gram を用いる。参加者には、認識結果と共に認識辞書が提供されるので、この書き起しに対するクエリ語の未知語と既知語の区別を行うことができる。

– 音節ベース音声認識による書き起し
音声認識の認識辞書には、日本語の全音節を登録する。言語モデルには音節の 3-gram を用いる。未知語の検索に焦点を当てた手法を試したい参加者は、この音節ベース書き起しを用いることで、クエリ語をすべて認識語彙外語と見なした実験を行うことができる。

これらの書き起しを得るため、CSJ 講演データ自体を学習データとして用いて、以下のような方針によりオープンな条件で音声認識を行った。

– CSJ 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 桁の数字が奇数か偶数かによって 2 分割する。これらを、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。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。

音声認識エンジンには Julius を用い、CSJ の転記基本単位 (IPU) を単位に認識を行った。認識結果として、10-best、コンヒュージョンネットワーク、ラティスの 3 種類の表現が出力されている。どの認識結果を使うかは、参加者の自由である。また、単語認識結果、音節認識結果の両方を組合せて使うことも可能である。

- 参加者による書き起し
参加者は自前の音声認識システムを用いて書き起しを作成することもできる。その際、既知語と未知語の条件をそろえるために、上記のオーガナイザが用意する単語ベース音声認識結果に用いた辞書を用いることが望ましい (必須ではない)。また、今後のテストコレクションを充実させるため、認識条件の詳細の説明と共に、得られた書き起しデータの提供を推奨する。

5. タスク定義

本節では、NTCIR-9 SpokenDoc タスク定義の基本的方針について説明する。詳しいタスク定義は、タスク Web ページ[3] で公開する文書を確認していただきたい。

5.1 検索対象の基本単位

CSJ の各講演は 200 ミリ秒以上のポーズで分割された転記基本単位 (Inter Pausal Unit; IPU) で書き起こしが行われている。SpokenDoc タスクでは、この IPU を正解判定の基本単位として用い、正解 IPU を見つけるタスクとして STD および SDR サブタスクを定義する。これにより、特定の語が発話された時刻および継続時間長の一致判定の問題、言語学的な文の単位の認定問題など、連続データである音声ドキュメントに付随する一致判定の揺れの問題を回避し、テキスト検索で利用されてきた離散データに対する様々な評価尺度の導入が可能となる。

5.2 STD サブタスク

クエリ

全講演用、コア講演用の 2 種類のクエリ語リストを提供する。参加者は、検索対象とする講演に応じて、クエリ語リストを選択し、検索結果を報告する。例えば、コア講演を対象とした STD タスクに参加する場合、コア講演用クエリ語リストを用いる。

結果提出

参加者は、一つのクエリ語リストに対し複数の検索結果を提出することができる。提出する一つの検索結果を run と呼ぶ。各 run の検索結果として、以下の 3 種類の情報を提出するものとする。

- run に関する情報
参加グループ ID、run の間の優先順位、検索対象とした音声ドキュメント (ALL または CORE)、使用した書き起しの別 (4.2 節)、を記述する。
- システムに関する情報
オフライン処理 (索引付け等) に関する情報として、使用したマシンスペック、処理に要した時間、索引のサイズ、を記述する。また、オンライン処理 (検索処理) に関する情報として、使用したマシンスペック、処理に要した時間、を記述する。また、用いた手法・システムの簡単な説明を記述する。
- 検索結果に関する情報
クエリ語リスト中の各クエリ語毎に、検出候補のリストを指定する。検出候補は、文書 ID、IPU の ID、検出の尤らしさを表すスコア、最終的な検出結果として出力するか否かのバイナリフラグ (YES または NO)、の 4 つ組で指定する。参加者は、自由に自らのスコア関数を定義してよいが、スコアの使用用途に注意する必要がある。オーガナイザはスコアを、値の大きい順に候補を選択して Recall-Precision 曲線を求めるために利用する。したがって、全クエリ語に対して共通の基準でスコアを設定する必要がある。バイナリフラグは、参加者の意図する最終的な検出結果の指定に利用する。参加者は、提出結果に対して Recall-Precision 分析が可能となるように、YES 判定の候補に加えて、十分な数の NO 判定の候補を結果に含めることが要求される。

評価指標

検出したIPUを単位としたRecallとPrecisionをクエリ毎に算出し、全クエリで平均した値を基本とした評価指標を用いる。バイナリフラグで参加者が指定した検出でのF-measure、F-measureが最大になるしきい値でのF-measure、などが分析のために利用される。

5.3 SDRサブタスク

クエリ

公開中のCSJ音声ドキュメント検索テストコレクション[12]に準じた検索トピックをクエリとして用いる。これらは、講演の一部に含まれる内容を探すクエリである。正解判定はIPUを単位とした可変長の区間に対して行われる。このクエリトピックに対して、次の2種類のタスクを設定する。

講演検索サブタスク　正解区間が含まれる講演を見つけるタスク。

パッセージ検索サブタスク　正解区間そのものを見つけるタスク。

両タスク共通で、一つのクエリセットを用いる。また、検索対象音声ドキュメントは、全2702講演とする。

結果提出

クエリセットに対し、参加者の提出する一つの検索結果をrunと呼ぶ。run毎に、以下の三種類の情報を提出するものとする。

- runに関する情報
  参加グループID、runの間の優先順位、使用した書き起しの別(4.2節)、対象タスク(講演検索 or パッセージ検索)、を記述する。
- システムに関する情報
  オフライン処理(索引付け等)に関する情報として、使用したマシンスペック、処理に要した時間、索引のサイズ、を記述する。また、オンライン処理(検索処理)に関する情報として、使用したマシンスペック、処理に要した時間、を記述する。また、用いた手法・システムの簡単な説明を記述する。
- 検索結果に関する情報
  クエリトピック毎に、検出候補のリストを優先度順で最大1000件指定する。検出候補は、講演検索サブタスクの場合は文書ID、パッセージ検索サブタスクの場合は文書IDとIPUのIDのペアで指定する。パッセージ検索サブタスクで指定するIPUは、検索結果の可変長区間(連続したIPU列)に含まれる、いずれかのIPU一つを指定する。

```
<RUN>
<SUBTASK>STD</SUBTASK>
<SYSTEM-ID>TUT</SYSTEM-ID>
<PRIORITY>1</PRIORITY>
<TARGET>CORE</TARGET>
<TRANSCRIPTION>REF-SYLLABLE</TRANSCRIPTION>
</RUN>
<SYSTEM>
<OFFLINE-MACHINE-SPEC>Xeon 3GHz dual CPU, 4GB memory</OFFLINE-MACHINE-SPEC>
<OFFLINE-TIME>18:35:23</OFFLINE-TIME>
...
</SYSTEM>
<RESULTS>
<QUERY id=SpokenDoc1-dry-CORE-001>
<TERM document=A01F005 ipu=0024 score=0.83 detection=YES />
<TERM document=S00M0075 ipu=0079 score=0.32 detection=NO />
...
</QUERY>
<QUERY id=SpokenDoc1-dry-CORE-002>
...
</QUERY>
</RESULTS>
```

図1　提出ファイルの例

評価指標

適合性判定により、各クエリトピックについて、可変長区間(IPU列)の集合が正解として与えられる。講演検索サブタスクについては、可変長区間が含まれる講演を正解文書とみなしてRecallおよびPrecisionを計算する。評価尺度としては、Mean Average Precision(MAP)を用いる。一方、パッセージ検索タスクでは正解文書が可変長区間となるため、文書検索タスクで用いられる評価指標がそのまま利用できない。現在のところ、いくつかの評価尺度の適用を検討中である。例えば、検索結果IPUが正解区間内に含まれる場合、その正解区間全体が検索されたと見なしてRecallおよびPrecisionを計算し、MAPを評価尺度とする評価手法[14]などを検討している。

5.4 提出ファイル

参加者は、run毎にタグ付けされた1つのファイルを提出する。提出ファイルの例を図1

に示す。

6. 評価スケジュール

2011 年 3 月までに、既存のテストコレクションを用いた dryrun(予行練習による結果提出) を行う。その後、4 月から 6 月までの間に formalrun(正式な課題配布と結果提出) を行う。2011 年 8 月に、参加者宛に評価結果の通知を行う。参加者は、評価結果を受けて成果報告会に向けた原稿を作成する。NTCIR-9 の成果報告ワークショップは、2011 年 12 月に予定されている。

7. ま と め

本稿では、NTCIR-9 SpokenDoc タスクのタスク設計について、オーガナイザの立場から基本的方針をまとめた。今後、参加者との議論を通して、より良い共通タスクの設計ができればと考えている。SpokenDoc への多数の参加を期待したい。

参 考 文 献

1) 秋葉友良：音声ドキュメント検索の現状と課題，情報処理学会研究報告，Vol.2010-SLP-82, No.10 (2010).
2) : 第 9 回 NTCIR ワークショップ. “http://research.nii.ac.jp/ntcir/ntcir-9/index-ja.html”.
3) : NTCIR-9 Core Task: ”IR for Spoken Documents (SpokenDoc)”. “http://www.cl.ics.tut.ac.jp/~sdpwg/index.php?ntcir9”.
4) : Text REtrieval Conference (TREC). “http://trec.nist.gov/”.
5) Garofolo, J.S., Auzanne, C. G.P. and Voorhees, E.M.: The TREC Spoken Document Retrieval Track: A Success Story, *Proceedings of TREC-9*, pp.107–129 (1999).
6) : Cross Language Evaluation Forum. “http://www.clef-campaign.org/”.
7) Pecina, P., Hoffmannová, P., Jones, G.J., Zhang, Y. and Oard, D.W.: Overview of the CLEF-2007 Cross-Language Speech Retrieval Track, *Advances in Multilingual and Multimodal Information Retrieval*, Berlin, Heidelberg, Springer-Verlag, pp. 674–686 (2008).
8) National Institute of Standards and Technology: Spoken Term Detection Evaluation Portal. “http://www.nist.gov/speech/tests/std/”.
9) : 音声ドキュメント処理ワーキンググループ Web ページ. “http://www.cl.ics.tut.ac.jp/~sdpwg/index.php?ntcir9”.
10) 西崎博光，胡　新輝，南條浩輝，伊藤慶明，秋葉友良，河原達也，中川聖一，松井知子，山下洋一，相川清明：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，情報処理学会研究報告，Vol.2010-SLP-81, No.13 (2010).
11) Itoh, Y., Nishizaki, H., Hu, X., Nanjo, H., Akiba, T., Kawahara, T., Nakagawa, S., Matsui, T., Yamashita, Y. and Aikawa, K.: Constructing Japanese Test Collections for Spoken Term Detection, *Proceedings of International Conference on Speech Communication and Technology*, pp.667–680 (2010).
12) 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, *Journal of Information Society of Japan*, Vol.50, No.2, pp.501–513 (2009).
13) Maekawa, K., Koiso, H., Furui, S. and Isahara, H.: Spontaneous Speech Corpus of Japanese, *Proceedings of International Conference on Language Resources and Evaluation*, pp.947–952 (2000).
14) 本田耕一郎，秋葉友良：講演音声を対象とした部分音声区間の内容検索タスクの設定とその検索手法の検討，日本音響学会春季研究発表会研究論文集，pp.185–188 (2010).