NTCIR-10
SpokenDoc-2

IR (Information Retrieval) for Spoken Documents

Kiyoaki Aikawa\textsuperscript{1}, Tomoyosi Akiba\textsuperscript{2}, Seiichi Nakagawa\textsuperscript{2}, Hiroaki Nanjo\textsuperscript{3}, Hiromitsu Nishizaki\textsuperscript{4}, Yoichi Yamashita\textsuperscript{5}

\textsuperscript{1}Tokyo University of Technology, Japan
\textsuperscript{2}Toyohashi University of Technology, Japan
\textsuperscript{3}Ryukoku University, Japan
\textsuperscript{4}University of Yamanashi, Japan
\textsuperscript{5}Ritsumeikankan University, Japan
What is “SpokenDoc-2”? 

- Second round of the IR for spoken documents 
- Finding the information related to given a query from too much speech data

Speech data $\cong$ Transcribed text by automatic speech recognition

"information retrieval" $\cong$ Information retrieval (word-unit transcription) In for ma tion re trie val (sub-word-unit)

Problem Including a lot of errors!!

Participants of SpokenDoc-2 will challenge Information retrieval from very noisy text data

Techniques for SpokenDoc may be used for OCR or Machine Translated text retrieval
Target Speech Data

• Type of speech data
  – Broadcast news speech, podcast, **lecture speech**...
  – Having noisier words

• Databases
  – CSJ (Corpus of Spontaneous Japanese)
    • 2,702 lecture speeches, 628 hours
  – New target! **Real academic meeting lectures collection**
    • Over 70 speeches from the spoken document processing workshops
SpokenDoc-2 Task Overview

- Sub-task1: Spoken Term Detection
- Sub-task2: Spoken Document Retrieval
1. Finding utterances including a query term

QUERY TERM: “Information Retrieval”

Spoken documents collection

Search...

Utterances

...Spoken term detection...

...information retrieval...

...spoken document retrieval...

Find!

2. A query term is included or not included in a spoken document?
Spoken Document Retrieval

- Ad-hoc Information Retrieval from lecture speeches
- Finding the **passages** including the relevant information related to a given query topic
- **Query**
  - Text query
  - Spoken Query (optional)
Merits for Task Participants

• The organizers will provide the rich transcriptions by an automatic speech recognition system
  – Participants can focus on developing IR technique for noisy data, although you have not any techniques for speech recognition

• Participants will get the real lecture speeches:
  • Academic meeting lectures
  – All lectures have reference transcriptions
  – Some lectures have slide data and digitalized papers

All data is Japanese
We are welcome to join SpokenDoc-2

- **Schedule**
  - Mar. 2012: release of the task description
  - June 2012: release of the reference automatic transcription
  - Sept. 2012: dry-run
  - Nov. 2012: formal-run evaluation
  - Nov. 2012-Feb. 2013: relevance judgment
  - Feb. 2013: release of the evaluation results
  - May 2013: camera ready submission due
  - June 2013: NTCIR-10 workshop meeting

- **Contact**
  - SpokenDoc-2 organizers:
    - E-mail: ntcadm-spokendoc2@cl.ics.tut.ac.jp
  - Web (coming soon!)
  - Twitter: @spokendoc2