Automatic Transcription for a Web 2.0 Service to Search Podcasts

Jun Ogata, Masataka Goto, and Kouichirou Eto
National Institute of Advanced Industrial Science and Technology (AIST)

Motivation

- The amount of speech data is increasing rapidly on the web
  - Podcasts have become popular and widespread

- A novel Web 2.0 application based on ASR (automatic speech recognition)

**Automatic transcription of podcasts**

Overview

- Podcast search service based on ASR
  - Users can search, read, and annotate podcasts
  - Growing need for full-text speech retrieval service
  - Existing podcast retrieval services
    - Podcast search service based on ASR
    - Users can search, read, and annotate podcasts
    - Growing need for full-text speech retrieval service
    - Existing podcast retrieval services
      - (Podscope and EveryZing (PodZinger))
        - Hide full-text ASR results
        - Users have no means of correcting ASR errors
    - PodCastle
      - Allow full-text ASR results to be accessed by both users and external search services
      - Allow users to cooperate with each other to improve ASR performance
    - First instance of Speech Recognition Research 2.0
      - [Goto, Ogata, and Eto: Interspeech2007]

**What are podcasts?**

- Audio programs distributed on the web
  - Podcast = RSS syndication feed + MP3 files
  - Updated episodes are automatically downloaded

**Metadata**
  - Title: CNN News Update
  - Description: The latest news happening in the U.S. and around the world.

**Episode 1**
  - Title: CNN News Update (8-21-2007 7 AM EDT)
  - MP3: http://rss.cnn.com/...08-21-07-7AM.mp3

**Episode 2**
  - Title: CNN News Update (8-21-2007 6 AM EDT)
  - MP3: http://rss.cnn.com/...08-21-07-6AM.mp3

**Episode 3**
  - Title: CNN News Update (8-21-2007 5 AM EDT)
  - MP3: http://rss.cnn.com/...08-21-07-5AM.mp3

- (New episodes can be added at arbitrary intervals)

**Three functions**

- Searching function
  - Full-text search of ASR results
  - List of episodes containing a search term is displayed together with text excerpts

- Reading function
  - View the full-text ASR result to understand the contents
  - Each word is colored according to the degree of ASR reliability (confidence measure)
  - Full-text can be accessed by external search engines (e.g., Google)

- Annotating function
  - Add "annotations" to correct ASR errors
    - Select the correct candidate from the candidate list
    - Type in the correct text
  - Candidate list is generated by using a confusion network
Various types of speech
• Noisy speech, narrow-band speech, speech with music, etc.

Acoustic features
• 12 MFCCs, energy, and delta and delta-delta

Training data
• 600 hours of presentation speech in the Corpus of Spontaneous Japanese (CSJ)

Triphone HMM (4513 tied states, 16 Gaussians/state)

Various tasks and wide domains
• Very difficult to specify the topic and vocabulary

Out-of-vocabulary issue
• The latest topics and words appear frequently in newly added episodes on a daily basis

Up-to-date language model
• Use of daily updated text from web-news sites

Static corpora
Newspaper LM  CSJ LM  Web news LM

Up-to-date LM (trigram, 152163 words)

A mechanism to keep a speech recognizer always up-to-date is needed
Improving Methods on the Basis of Web 2.0

Use of Web 2.0 knowledge sources and contents

A huge amount of multimedia contents and annotations are generated and accumulated by a number of (anonymous) users. These resources are steadily increasing and updated on a daily basis.

Language model adaptation using RSS

RSS provides useful data:
- publishing date, titles, and accompanying text descriptions

Use of the text data in RSS for topic adaptation

1. Parse RSS title and description using a Japanese morphological analyzer
2. Retrieve text documents related to the topic of the episode from the web using a search engine
   (Only nouns are used as search terms)
3. Build a topic-dependent LM from the retrieved documents
4. Interpolate the topic-dependent LM with the baseline up-to-date LM

* Yahoo! web search API is used to search terms

B, D, and E: improved
A, C: only slight improvements
No useful information for the topic was given in the RSS

Acquiring pronunciations from the web

Many technical terms and coined words
- e.g., PodCastle, Wikipedia...

How to obtain their pronunciations?

A Web 2.0 service "Hatena diary keyword"
- Publishes a list of new keywords including explanations and pronunciations
- Over 200000 keywords are registered at present (about 300 words are added daily)

(A web service which provides full-text searching of podcasts)

Investigation

The pronunciations were not obtained for 11.5% (17438/152163) of words in our LM
The correct pronunciations can be obtained for 22.9% (3997/17438) of those words

Acoustic model training based on error corrections by users

Error corrections
- Some users may thoroughly correct all errors
- Most users will correct only errors related to the part according to their interests

The use of error correction results to improve ASR is a considerable challenge

Investigation in acoustic model training
- The amount of speech data was limited
- The corrected transcriptions were not perfect (about 50% of errors were corrected by users)

Performance of MLLR-MAP training

Future work

- Larger-scale experiments
- Training techniques so as to make best use of user corrections