A Large Arabic Broadcast News Speech Data Collection

Niklas Paulsson, Khalid Choukri, Djamel Mostefa, Denise DiPersio,
Meghan Glenn and Stephanie Strassel

Introduction

- Arabic Broadcast News collected from several satellite sources
- TV and radio from Arabic channels with different speaker styles
- Data collected by LDC and ELDA
- More than 4000 hours of satellite sources collected
- All recordings audited and a part selected for transcription
- Transcriptions followed the Quick Rich Transcription method
- Quick verification procedure
- More than 2400 hours of transcribed audio
- Transcripts produced as part of the DARPA GALE program:
  - To develop processing engines for analyzing huge volumes of speech and text
  - To deliver pertinent, consolidated and easy-to-understand forms to military personnel and monolingual English-speaking analysts

Auditing

- All complete recordings manually audited
- Selection of three 30 sec samples: beginning, middle, end
- Set of questions filled in using a web-based tool
- For each of the below questions the auditor fills in the answer:
  - Is there a recording?
  - Is the audio quality ok?
  - What is the language?
  - Is it from the intended program?
  - What is the data type?
  - What is the topic?
- Any files with poor quality or problems were rejected
- Audits from both the US and Morocco collection were regrouped
- A subset of all recordings that pass the audit selected for transcription

Data sources

- Data collected from satellite: TV and radio sources
- Two collection sites: US, Morocco
- Recordings from 16 Arabic news sources
- 140 hours audio / week collected
- Modern Standard Arabic and dialectal Arabic
- Broadcast News and Broadcast Conversation
- US collection sources:
  Abu Dhabi TV, Al Alam, Al Arabiyah, Al Iraqiyah, Aljazeera, Al Ordiniyah, Dubai TV, Kuwait TV, Lebanese Broadcasting Corp., Oman TV, Saudi TV, Syria TV
- Morocco collection sources:
  Al Hurra, Radio Sawa, Al Baghdadya, Yemen TV

Transcription and verification

- Selected audio data transcribed using XTrans tool
- Transcripts are verbatim, orthographic and time-aligned in Arabic script
- Transcripts follow “quick rich” transcription (QRTR) guidelines:
  - Time-stamped audio segments labeled for sentence type
  - Markup for limited speech phenomena
  - No vowels included
  - No lexicon
- Transcripts include:
  - Overlapping speech
  - Identification of speaker data (name, male/female, dialect status)
  - marked foreign language portions
  - Segmentation of sentence units – statement, question, incomplete
  - Speaker noise
- Each transcript reviewed during 18-minute quick verification stage, which checks:
  - Transcripts match the speech
  - Segmentation is correct
  - Orthography (transcript, speaker names)

Tools, format, post processing

- DVB-S streams captured from satellite
- Video and audio recorded in MPEG-2 format
- Post processing of the Morocco collection:
  - Automatically launched every day
  - Checks disk space for performing post processing
  - Check if the all programs have been recorded
  - Remove files shorter than 15 min
  - Add the date to each file
  - Converted into AVI (video) and WAV (audio)
  - Audio extracted from video and down-sampled
  - Audio stored in PCM (mono, 16 kHz, 16 bit)
  - Calculates MD5 checksum
  - Files compressed using the lossless scheme FLAC
  - Recordings are divided into subsets
  - Transferred daily to a server in Paris using RSync

Conclusion

- Collection of Arabic TV and radio sources
- Ongoing collection of more than 4300 hours
- Selected data transcribed using Quick Rich Transcription approach
- More than 2400 hours of audio transcribed to date
- Quick Verification of all transcripts