Speech Processing Applications in Quaero

Sebastian Stüker
Introduction

- Quaero is an innovative, French program addressing multimedia content

- Speech technologies are part of the Quaero program:
  - Researched for many decades: Speech Recognition, Machine Translation, Information Retrieval
  - Reached level of maturity that is suitable for application

- Scientific content is disseminated through lectures–live and recordings
  - How to find relevant content
  - How to reduce content to the most important information
  - How to bridge the language barrier

- Access to lecture content can be facilitated through technology
  - Automatic Indexing, Information Retrieval and summarization
  - Speech-to-speech translation
Quaero

- 5 year French research Program started in 2008 with 26 partners and 200 Mio € budget (100 Mio € funding)

- 5 Application Projects:
  - Digital Enrichment of Content: Digitalization and semantic annotation of content
  - Digital Media Management
  - Personalized Video
  - Search Engine: search text, video, images, music
  - PC and mobile portals: access and interaction with TV

- 2 Technology Projects
  - Core Technology Cluster: underlying technologies
  - Corpus Project: corpora for technology development
Speech Applications in Quaero

- **Automatic Speech Recognition (ASR)**
  - 7 languages: EN, FR, GE, RU, SP, GR, PL
  - 4 partners: KIT, LIMSI, RWTH, VR

- **Speaker Diarization**
  - Cross show speaker diarization
  - Political speaker tracking

- **Language Recognition**

- **Dialog and Vocal Interface**

- **Emotion Recognition**

- **Machine Translation**

- **Speech-to-Speech Translation**

Progress driven by the concept of coopetition
Automatic Speech Recognition (ASR)

- **Applications for ASR:**
  - Interaction with machines (robots, appliances, computers)
  - Human-to-Human Interaction (speech translation)
  - Observation of users

- **Speech recognition is difficult due to variability:**
  - Speaker, speaking style, speaking rate
  - Acoustic environment, varying microphones

- **Speech recognizers learn automatically:**
  - From transcribed collections of audio data
  - From large collections of text data
Machine Translation (MT)

- Automatic Translation between Languages
  - **Rule Based Approach**: Write translation rules by hand
  - **Statistical Machine Translation**: Learn translations from large corpora of parallel texts

Difficulties of Machine Translation:
- Word reordering
  - „Which one do you prefer“ -> „Welches bevorzugen Sie?"
- Word fertilities
  - „Querría leche caliente“ -> „I would like hot milk“
- Ambiguities
  - „bank“: financial institute or the edge of a river
  - „Time flies like an arrow“ (7 meanings)
Speech-to-Speech Translation

Speech-to-Speech Translation:
- Merge Speech Recognition, Translation and Speech Synthesis
- Translate lectures, videos, meetings, dialogs simultaneously
- Provide translation in an unobtrusive way
**Offline vs. Online Recognition**

**Offline**
- The complete material can be examined before the actual recognition:
  - Segmentation can use more complex techniques
  - Gaussian Selection for speed-up
- Can run with a high single CPU real-time factor:
  - Better recognition results
  - Speed up via parallel processing of different material on several CPUs
  - Upper bound for processing time only given by available computing resources, time limit for availability, rate of generation of new data
- Can run in several passes
  - Speaker adaptation
  - System combination
- Thus: Combination of additional knowledge and more processing power/time leads to better recognition rate
Offline vs. Online Recognition

Online (Simultaneous)
- Material to be recognized not known a-priori
  - All processing has to be done as the material arrives
  - Segmentation has to be done on the fly
- Should run in (near) real time
  - Continuous result output
  - As low a latency as possible
- Should run with low latency
  - Minimum amount of latency necessary for recognition, and especially the translation (correct translation only possible if complete sentence is known ?!)
- Need for speed and online data “generation” lead to additional challenges/decreased performance
Offline vs. Online Recognition

Applications

- Offline
  - Translation of databases in advance
  - Needs sufficient processing power,
  - Rate of newly added material needs to be low enough
  - Time lag between addition of video and availability of translation must be acceptable

- Online
  - Simultaneous translation
  - Either in the room, or if broadcasted over the web
  - Also for databases:
    - Translation on demand: Translate only when needed, save that translation for later use
Academic Lectures

- Academic Lectures happen live
  - Either in English or in native language of the country
  - Often international audience

- University lectures increasingly recorded and distributed electronically:
  - OpenCourseWare (OCW) by MIT: [http://ocw.mit.edu/index.htm](http://ocw.mit.edu/index.htm)

- iTunes U Service: 75,000 files (videos, PDF etc.) from > 12 countries
Academic Lectures

- Not only courses but also scientific dissemination/training through talks and lectures

Examples

- Conference talks / Lectures:
  - www.ted.com
  - www.videolectures.net
Advancing Science with Speech Technology

Finding relevant information:
- Use ASR to automatically transcribe the content of a lecture
- Use information retrieval and question answering to find relevant information fast

Summarizing content:
- Use automatic summarization technology to concentrate content for faster intake

Overcoming the Language Barrier
- Currently “broken” English the academic lingua franca
- Not as widely spoken as thought
- Large language diversity, but also important
- Use speech translation technology to overcome the barrier
Language Diversity

- 5,000 to 7,000 languages exist in the world today
- Languages are frequently dying
- Speech translation systems exist only for a fraction of the languages in the world
- Languages Selected: Rich, populous, politically relevant
- Training of translation systems requires large amounts of annotated training data
- Speech translation now performs well enough for use in real-life applications
- Language diversity related to biological diversity
- In order to cover more, maybe even all, languages in the world, fast and cheap methods for creating ASR systems in new, under-resourced languages are needed.
The need for Translation

- Lectures are being held in all kind of languages
- Even command of English as a lingua franca is not as wide spread as commonly assumed
- Human Translator costs are very high and therefore not feasible for lectures

English language knowledge (not mother tongue)
Importance of Language Diversity

- “Language is the pinnacle achievement of mankind” (Vivianne Redding, European Commission)
- Language is inherently linked with culture
- The way we speak / the way our language is structured influences the way we think (e.g. see research on cognition by Lera Boroditsky at UCLA)
- Diversity in thinking and reasoning
  - Large and diverse knowledge encoded in different languages
- Language Diversity as important for human well-being and prosperity, as biodiversity in nature
Advantages, Challenges of Lectures

Advantages

- Monologue, no/limited interaction
- Prepared speech, varying levels of spontaneity
- Sufficiently good recording setup, wide bandwidth, close talking or at least table microphone

Challenges

- Background Music
- Teasers
- Sound effects
- Question-/Answering Interaction
- Wide variety of topics–open domain
Lecture Translation Showcase

Simultaneous Translation of Lectures/Conference Talks

- Available Language Directions
  - English <-> Spanish
  - English <-> German

- Architecture
  - Concatenation of ASR and MT:
    - Passing single best output
    - Only word sequence, no other information
  - Different set-ups possible:
    - Completely server based
    - Or distributed computing
      - ASR on client
      - Translation on server
      - Distribution of the result via WWW
Lecture Translator “Classic”

Translation Server (optional)
Lecture Translator “Web”

Public Address
Subtitle Display
Targeted Audio

Recognition

Translation Server
Web Server

Personal Web Enabled Devices

Local Audience
Remote Audience

Web Server
Screenshot of Lecture Translator Web Interface

ISL comet livecast for lecture translator

Alex Waibel: one two one two
...good morning ladies and gentlemen
...welcome to today's presentation of the k i t simultaneous lecture translation system
our system consists of the speech recogniser that runs on a laptop
...and a remote translation unit that is hosted on server in our institute
...it is capable of online translation with very low lagged and high quality is our aim for
the near future
...to provide online lexicons ration has the service at k i t
...to enable foreign students from various origins to follow all lectures given here

Alex Waibel: uno dos uno dos
...bueno
...bienvenida a la de hoy de la k creo t sistema nuestro sistema consiste en el discurso
recoger que corre de un
...y una remota unidad que es ubicada sobre servidor en nuestro instituto
...es capaz de on-line con muy baja retraso y alta calidad es nuestro objetivo para el futuro
...para proporcionar on-line lexicons racion tiene el servicio a k creo t
"super bowl"

- All
- Français
- English
- 中文
- العربية
- español
- Русский

Select sources

In this video:

People Organization Location

Radical Imam MSNBC contributor, Washington editor of the nation and New star of Fox and friends. Thank you, Chris to say about a Super Bowl. That does not have the Patriots in it. It is coming up on tonight's show just months.

I would not be so crass as to blatantly rip off the Daily Show's idea of a moment of Zen. I have to admit that in reviewing some of the more insanely bad right wing American coverage of Egypt to be able to talk about it with Christie's today this out of context moment I think actually qualifies as kind of Zen.
View4You Demo
Jibbigo: Speech-to-Speech Translation on the iPhone

- First Speech-to-Speech Translation Software completely on an iPhone
- Translates natural speech between English and Spanish
- All components run on the iPhone
- No network connection needed
Conclusion

- Quaero is an innovative, French program addressing multimedia content

- Speech technologies are part of the Quaero program:
  - Automatic Speech Recognition
  - Speech translation
  - Information Retrieval

- Scientific content is disseminated through lectures:
  - Live lectures
  - Recordings and databases of lectures

- Access to lecture content can be facilitated:
  - Automatic Indexing
  - Information Retrieval and summarization
  - Speech-to-speech translation to overcome the language barrier