Some Open Challenges for Spoken Language Processing

Lori Lamel

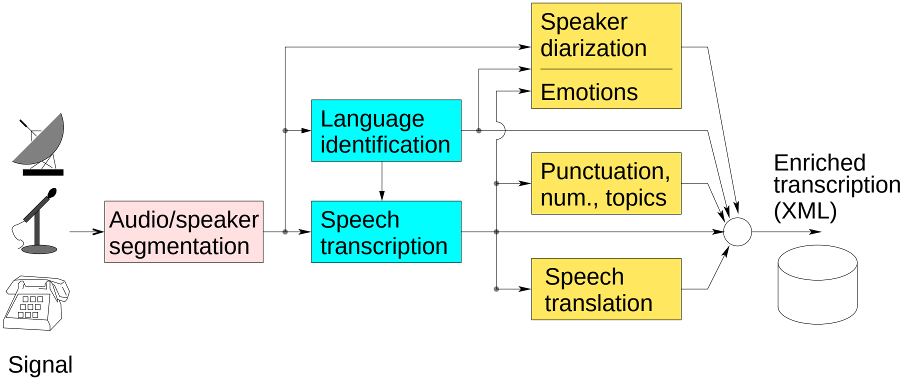
CHIST-ERA

Cork, September 6, 2011

Introduction

- Spoken language processing technologies are key components for indexing and searching audio and audiovisual documents
- Lots of information on web that is not in textual format
- Speech is ubiquitous
- Conversational systems (human-machine & human-human communication)
- Spoken language processing technologies
  - Speech-to-text transcription (STT)
  - Speaker diarization & recognition
  - Language identification
  - Spoken language dialog
  - Machine translation (MT)
- Applications: audiovisual media analysis, media monitoring, opinion monitoring, audiovisual archive indexing, captioning, question-answering, speech analytics, offline & online translation, social media, ...

Spoken Language Technologies

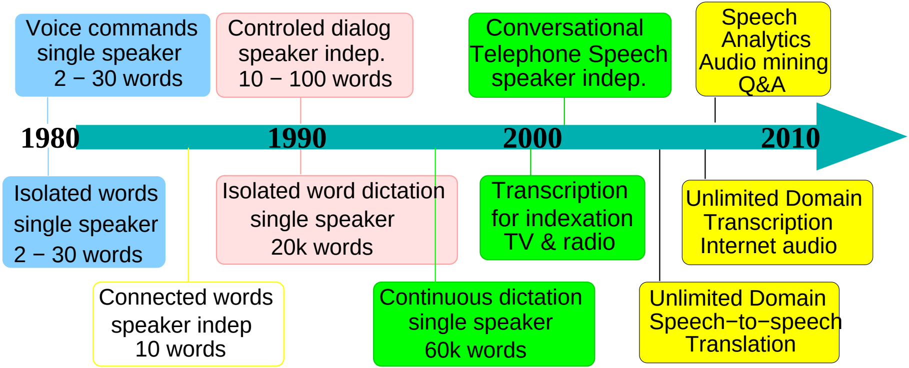


Some Open Challenges

- Providing 'equal' e-access for citizens
- Ubiquitous (intelligent) computing
- Developing generic models to remove task dependency
- Reduce development/porting costs for targeted applications (time & money)
- Automatic learning from unannotated data
- Use of context, keeping language models up-to-date
- Personalization
- Providing enchriched annotations for audio documents (speaker, language, topic, conditions, style, sentiment, state ...)

  CHIL vision: who what where when how (context aware)
- Close-to-real time translation of meetings, talks
  each person speaks and hears in their own language (initially key terms and concept), automatic identification of the persons who is talking
- Reduce gap between machine and human performances

30 Years of Progress



Indicative ASR Performance

Task	Condition	Word Error
Dictation	read speech, close-talking mic.	3-4% (humans 1%)
	read speech, noisy (SNR 15dB)	10%
	read speech, telephone	20%
	spontaneous dictation	14%
	read speech, non-native	20%
Found audio	TV & radio news broadcasts	5-15% (humans 4%)
	TV documentaries	20-30%
	Telephone conversations	20-30% (humans 4%)
	Lectures (close mic)	20%
	Lectures, meetings (distant mic)	50%
	Parliament	8%

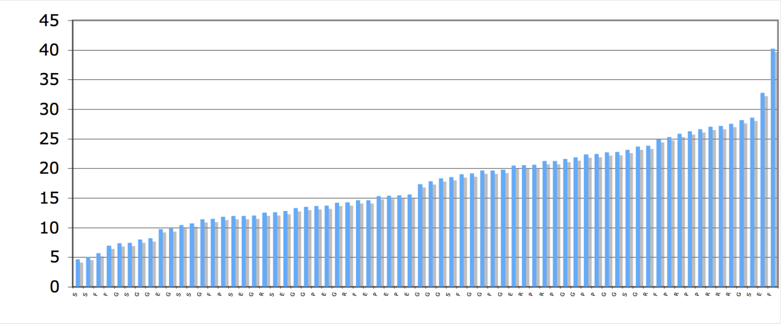
Why Is Speech Processing Difficult?

Text:	I do not know why speech recognition is so difficult
Continuous:	Idonotknowwhyspeechrecognitionissodifficult
Spontaneous:	Idunnowhyspeechrecnitionsodifficult
Pronunciation:	YdonatnowYspiCrEkxgnISxnIzsodIfIk∧lt
	YdonowYspiCrEknISNsodIfxk∧l
	YdontnowYspiCrEkxnISNsodIfIk∧lt
	YdxnowYspiCrEknISNsodIfxk∧lt

Important variability factors:

Speaker	*Acoustic environment*
physical characteristics (gender,	background noise (cocktail party, ...)
age, ...), accent, emotional state,	room acoustic, signal capture
situation (lecture, conversation,	(microphone, channel, ...)
meeting, ...)	

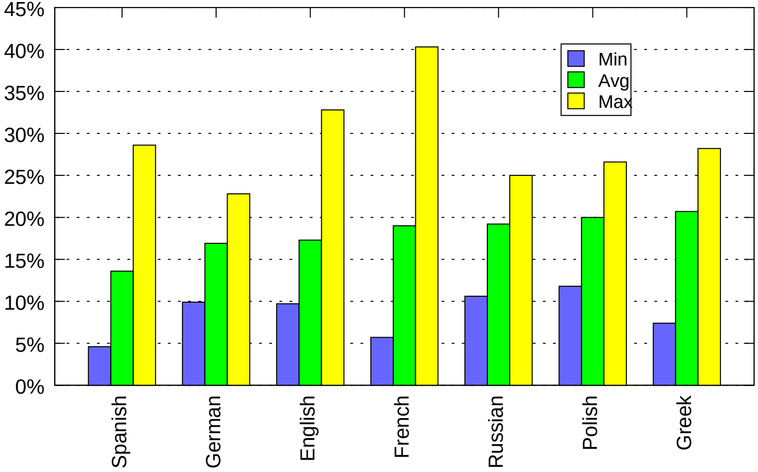
Quaero Eval10 - WER Variability



	English	French	German	Russian	Spanish	Greek	Polish
Best	9.7	5.7	9.9	10.6	4.6	7.4	11.8
Worst	32.8	40.3	22.8	25.0	28.6	28.2	26.6
Ave	17.3	19.9	16.9	19.2	13.6	20.7	20.0

WER versus Language

Mix of broadcast news and broadcast conversations
Lowest and highest document WER



Accent Adaptation

US English models (**H1**), Multi-accents models (**H2**)

ABC News Australia (sample #1)

H1: The winston alliances about three June (play)
H2: The western alliance is about to resume

ABC News Australia (sample #2)

H1: The nation safety terry general yacht who she (play)
H2: The NATO secretary general Jaap de Hoop Scheffer

France French models (**H1**), Multi-accents models (**H2**)

TV5 News Canada (sample #1)

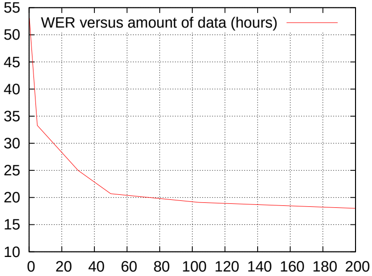
H1: mars devoir affecter ça va continuer cette d'ailleurs se regardent ...(play)
H2: absolument absolument *assister* ça va continuer cette pluie d'ailleurs si on regarde ...

System Development

- State-of-the-art speech recognizers use statistical models trained on
  - hundreds to thousands hours of transcribed audio data
  - hundreds of million to several billions of words of texts
  - large pronunciation lexicons
- Less e-represented languages
  - Over 6000 languages, about 800 written
  - Poor representation in accessible form
  - Lack of economic and/or political incentives
  - PhD theses: Vietnamese, Khmer [Le, 2006], Somali [Nimaan, 2007], Amharic, Turkish [Pellegrini, 2008]
  - Relative importance of textual vs audio data
  - SPICE: Afrikaans, Bulgarian, Vietnamese, Hindi, Konkani, Telugu, Turkish, but also English, German, French [Schultz, 2007]

Data for Model Training

- Data collection and transcription is costly
- How much does data bring?



BN data, ASR2000

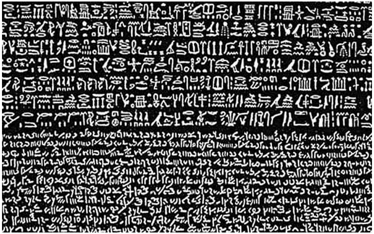
- Asymptotic behavior of the error rate
  - rapid progress on new problems (i.e. new data)
  - but slow progress on old problems (on average 6% per year)
- New data should cost less (need to learn to better use low cost data)
- Need more varied data

Machine Translation

- Text & speech translation
- Real-time speech translation (lectures, seminars, meetings, ...)
- Official documents (governmental, patents, documentation, ...)
- Some current research topics: *pivot translation, hierarchical model, syntax-based models, discriminanive word alignement, lexicalized reordering, POS-based reordering, long-range reorderings, multi-source translation, ...*
- Many proposed evaluation metrics: *Bleu, NIST, TER, TERp, HTER, Meteor, ...*
- NIST MetricsMaTr `http://www.nist.gov/itl/iad/mig/metricsmatr.cfm`
- Free online translation services illustrate the advances and deficiences of the state of the art
    - Can handle large volumes of data
    - Accuracy far below that of humans
- Highly subjective judgement of what is a good translation (adequacy, fluency)

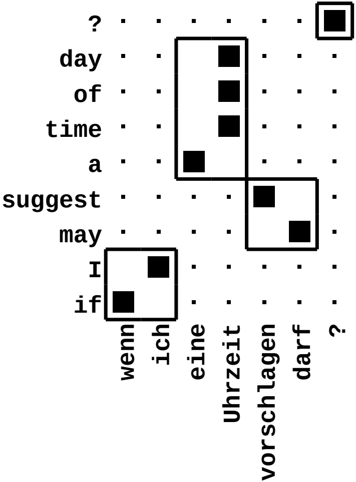
Machine Translation

- Statistical MT relies on translation models estimated on parallel texts
- Rosetta stone, European Parliament Plenary Sessions (EPPS), UN resolutions, Canadian parliament texts, ...
- Computationally expensive
- Need for spoken parallel documents



Using Parallel Texts

- Statistical MT uses parallel texts
- Alignment of sentences, phrases and words
- Reordering model, phrase translation table, target language model
- Adding knowledge (context, local/user/topic, linguistic)

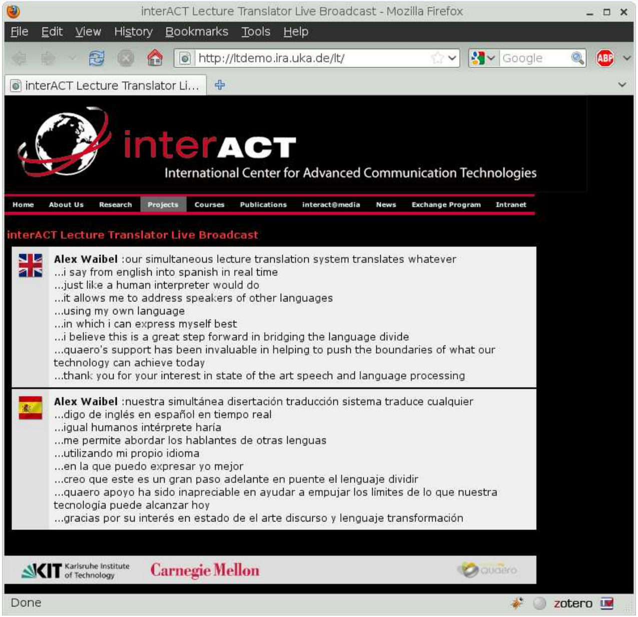


Quaero Euromatrix (from H. Ney)

- Joint effort between KIT, LIMSI and RWTH
- 22 languages , 462 pairs, English as pivot, 42 systems
- Training: EU laws (JRC), Europarl, UN resolutions, news commentaries
- Eval data: EU laws (Bleu scores)

	bg	cs	da	de	el	en	es	et	fi	fr	hu	it	lt	lv	mt	
bg	bg	46.0	41.9	41.2	39.9	56.8	46.9	34.0	33.9	48.2	34.9	45.6	33.3	36.9	32.2	4
cs	39.9	cs	42.9	43.3	40.1	57.5	48.3	35.5	35.0	49.7	35.3	47.8	34.4	37.5	31.9	4
da	38.2	46.3	da	45.0	39.7	56.3	48.3	35.9	36.2	49.9	35.3	48.0	34.1	37.7	30.9	4
de	31.2	44.7	43.4	de	38.7	54.5	46.6	34.6	35.2	48.5	34.9	45.9	33.5	36.3	30.0	4
el	39.1	45.2	42.3	41.8	el	54.0	49.4	32.8	33.3	50.4	33.4	48.8	31.9	35.4	30.8	4
en	46.7	53.3	50.0	47.5	45.2	en	55.5	40.5	39.4	51.4	40.6	54.8	38.9	43.1	43.5	5
es	40.1	47.7	45.1	44.5	43.1	59.5	es	35.2	35.5	54.9	35.2	52.2	33.8	37.1	32.5	4
et	35.0	40.3	37.7	39.6	33.2	51.8	41.3	et	33.7	43.5	32.6	40.3	33.8	36.7	27.2	3
fi	31.9	38.4	37.1	37.1	31.9	46.3	39.9	35.8	fi	40.3	34.7	38.6	32.6	34.4	25.4	3
fr	31.4	42.5	41.2	41.1	39.8	55.7	49.1	31.8	31.7	fr	30.7	48.7	31.6	34.4	28.2	4
hu	34.7	40.2	37.2	37.2	33.3	50.1	40.7	33.8	34.1	40.5	hu	39.2	32.0	35.0	27.4	3
it	40.5	48.3	45.3	45.2	43.7	59.9	53.0	35.9	36.1	55.5	35.2	it	34.4	37.8	32.8	4
lt	33.9	39.7	35.4	36.9	32.0	50.5	40.2	34.7	31.2	42.0	31.9	39.2	lt	38.5	26.8	3
lv	35.3	40.9	36.1	37.7	32.9	52.0	41.3	34.9	30.9	43.2	32.0	40.3	37.7	lv	27.0	3
mt	42.5	48.2	43.4	42.6	37.5	69.8	50.1	35.7	35.4	51.2	36.5	48.9	35.0	39.2	mt	4
nl	39.4	47.1	45.6	45.7	37.4	57.4	49.8	35.5	36.1	51.1	36.2	49.2	34.1	37.4	31.9	
pl	40.2	46.1	41.4	43.2	38.1	60.2	46.2	36.7	33.4	49.5	34.7	45.4	35.4	38.7	32.2	4
pt	40.1	47.5	45.0	44.4	43.4	59.8	54.2	35.5	35.4	55.7	34.6	52.5	33.9	37.2	32.4	4
ro	41.0	47.5	42.8	42.3	41.2	59.9	49.8	34.4	34.3	52.6	34.9	49.1	33.3	36.9	33.0	4
sk	40.8	49.9	42.8	41.8	39.2	59.4	47.2	35.0	34.6	47.9	35.9	45.9	34.4	38.1	33.2	4
sl	41.2	47.4	42.1	43.8	38.5	60.6	46.9	37.0	33.7	49.2	35.0	45.8	35.9	39.6	32.9	4
sv	37.6	45.9	44.8	43.4	39.4	58.0	47.4	35.0	35.6	48.5	34.8	46.5	33.4	36.6	31.5	4

Speech MT



interACT Lecture Translator Live Broadcast - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://ltdemo.ira.uka.de/lt/
Google
interACT Lecture Translator Li...
interACT
International Center for Advanced Communication Technologies
Home About Us Research Projects Courses Publications interact@media News Exchange Program Intranet
interACT Lecture Translator Live Broadcast
Alex Waibel :our simultaneous lecture translation system translates whatever
...i say from english into spanish in real time
...just like a human interpreter would do
...it allows me to address speakers of other languages
...using my own language
...in which i can express myself best
...i believe this is a great step forward in bridging the language divide
...quaero's support has been invaluable in helping to push the boundaries of what our technology can achieve today
...thank you for your interest in state of the art speech and language processing
Alex Waibel :nuestra simultánea disertación traducción sistema traduce cualquier
...digo de inglés en español en tiempo real
...igual humanos intérprete haría
...me permite abordar los hablantes de otras lenguas
...utilizando mi propio idioma
...en la que puedo expresar yo mejor
...creo que este es un gran paso adelante en puente el lenguaje dividir
...quaero apoyo ha sido inapreciable en ayudar a empujar los límites de lo que nuestra tecnología puede alcanzar hoy
...gracias por su interés en estado de el arte discurso y lenguaje transformación
KIT Karlsruhe Institute of Technology
Carnegie Mellon
quaero
Done
zotero

Audio Samples

- CHIL Seminar, spontaneous, far-field mike, non native

  I just give you a brief overview of [noise] what's going on in uh audio and why we bother with all these microphones and eh ...

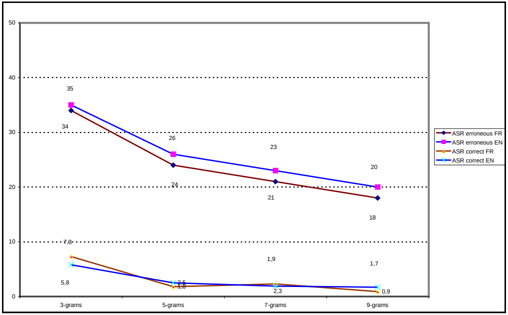
- Similar challenges to process interviews, focus groups, ...

Human ASR Benchmarks

- Human listeners significantly outperform machines on speech transcription tasks (5 to 6 times better than machines) [Greenberg, 1996; Lipmann, 1997; Pools, 1999]
- Variation handling: machines have trouble with rare events that are poorly modeled (pronunciation variants, disfluencies, ungrammatical sentences, noise, native and non-native accents etc.)
- Information sources
    - Humans use "higher-level" knowledge
    - Human listeners and ASR systems likely use different acoustic cues
    - Intrinsic spoken language ambiguities (language bias)
    - Simplified speech models (model bias)
- Speech Communication (2007) special issue on Bridging the Gap: Human Speech Recognition vs ASR
- Perceptual expts: Shinozaki & Furui, Vasilescu et al, 2009, 2011

Human Perceptual Tests

- Target words (acoustically poor, function words, 90% wrong) pose problems for humans: WER 21.5% French, 22.5% English



- Higher human error rate on stimuli with ASR errors
- Humans more errors on ASR deletions (poor acoustic information)
- Strong reduction of human WER with increasing context (3g→5g)

Data/Models/Knowledge (1)

- Better use of the data
- Semi- and unsupervised training methods
- Need to know when the machine is right or wrong (confidence scores)
- Ways to get cheap annotations:
  - Corrections from users: e.g. Nuance dictation, Google Translate
  - Crowd-sourcing, .e.g. Amazon Mechanical Turk
  - Use automatic systems to assist manual processing (virtous circle)
  - Web as training data (via IR and filtering techniques)
- Fast development methods (unsupervised testing)

- **Evaluation is a integral part of system development**

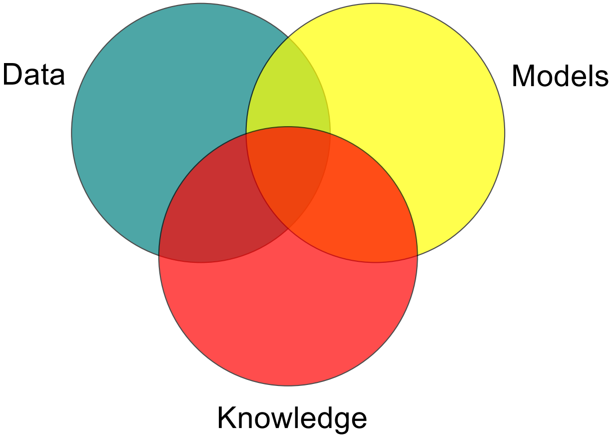
Data/Models/Knowledge (2)

- The same modeling techiques have been succesfully applied to a number of reasonably well e-resourced languages (with some language-specific adaptations)
- Some emerging research topics: *multi-layer perceptron based features, continuous-space language models, unsupervised training & adaptation, higher level knowledge sources, system combination...*
- Extension of language coverage (including low e-resourced languages)
- Automatic discovery of lexical and acoustic units
- Multilingual acoustic modeling to address training data limitations
- Class-based models (articulatory features)
- Automatic pronunciation discovery and better pronunciation models
- Detecting and handling language (code) switching

Data/Models/Knowledge (3)

- Extracting linguistic and paralinguistic knowledge from data
- Annotation of metadata (speaker, language, topic, emotion, style ...)
- Model adaptation: keeping models up-to-date
- Semantic modeling
    - Contextual understanding
    - Punctuation and prosodic features
    - Dialog, question-answering, opinion monitoring
- Reduce gap between machine and human performances (at least 20 years)
- Study of ASR errors & human perceptual experiments
- Cross-modal: using multiple information sources e.g., person identification in video: speaker diarization, OCR, face recognition, fusion

Summary



NIST MetricsMaTr

http://www.nist.gov/itl/iad/mig/metricsmatr.cfm

- Research challenge to promote development of innovative MT metrics that correlate will with human assessment of MT quality
- Drawbacks to the current evaluation methods
  - Automatic metrics primarily applied to English and utility for real applications unknown
  - Human assessments slow, subjective, costly, hard to standardize, require bilinguals
- Develop infrastructure for MT evaluation
  - bring together diverse community
  - to establish improved metrology
  - promote discussion and new perspectives for research

Thank you