Broadcast Speech Transcription and Translation

A. de Gispert, X. A. Liu, W. J. Byrne, M. J. F. Gales & P. C. Woodland
Cambridge University Engineering Department
{ad465,xl207,wjb31,mjfg,pcw}@eng.cam.ac.uk

1 INTRODUCTION

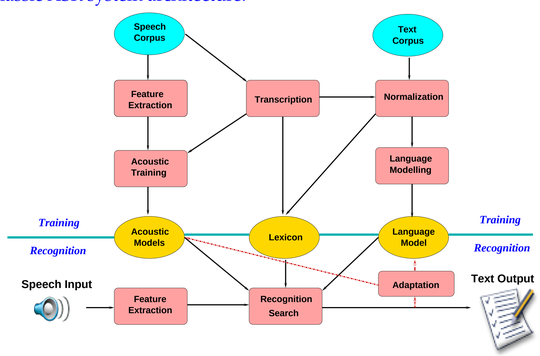
- Key research areas for speech translation are:
  - Speech Recognition (ASR) / audio Transcription.
  - Statistical Machine Translation (SMT).
  - Integration between ASR and SMT.
- System development involves:
  - Developing statistical models for speech and language tasks.
  - Parameter estimation from large data sets.
  - Finding the best hypothesis / speech segment etc.
- Both ASR and SMT can be formulated using a Source-Channel model:

Transcription	Translation
Input - an utterance A	Input - a foreign sentence F
Output - a transcription $\widehat{W}$	Output - an English sentence $\widehat{E}$
$\widehat{W} = \text{argmax}_W P(W\|A)$	$\widehat{E} = \text{argmax}_E P(E\|F)$
$= \text{argmax}_W \frac{P(A\|W)P(W)}{P(A)}$	$= \text{argmax}_E \frac{P(F\|E)P(E)}{P(F)}$
$= \text{argmax}_W \underbrace{P(A\|W)}_{\substack{\text{Acoustic}\\\text{Model}}} \underbrace{P(W)}_{\substack{\text{Source}\\\text{Language Model}}}$	$= \text{argmax}_E \underbrace{P(F\|E)}_{\substack{\text{Translation}\\\text{Model}}} \underbrace{P(E)}_{\substack{\text{Source}\\\text{Language Model}}}$

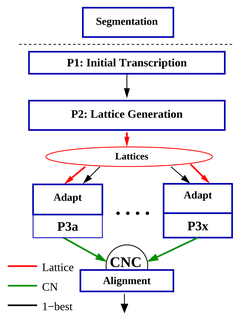
Both use Maximum A Posteriori search using models learned from data.

2 SPEECH TRANSCRIPTION

- Classic ASR system architecture:



- Statistical models typically trained using:
  - more than 1500 hours of speech
  - more than 2500 millions of words of text
  - continuous density HMMs using cepstral features
  - N-gram language models
  - maximum likelihood and discriminative training
  - parameter tying and smoothing
  - normalization and adaptation techniques
  - combining outputs from complimentary systems
- Broadcast speech transcription system:
  - Find speech and cluster into homogeneous regions
  - Search for best possible word strings for each segment using acoustic and language models
  - Multiple recognition passes: **lattices** constrain search space in later stages
  - Unsupervised adaptation "tunes" models to speech
  - Can run system faster if "prune" search more

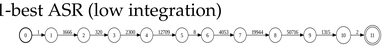


3 ASR AND SMT INTEGRATION

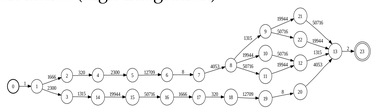
- Sentence boundary detection:
  - SMT systems require well-defined sentential structures.
  - Post-processing ASR system outputs to locate sentence boundaries.

Original STT outputs: 各位 好 欢迎 您 收看 中文 国际 频道 的 今日 关注 今天 呢 我们 要 讨论 的 是 伊朗 核 问题.
Post-processed STT outputs:
Sentence 1: 各位 好 欢迎 您 收看 中文 国际 频道 的 今日 关注.
Sentence 2: 今天 呢 我们 要 讨论 的 是 伊朗 核 问题.

- Translating multiple recognition outputs:
  - feeding ASR modelling information into translation.
  - flexible decoding framework for translation.
  - translate 1-best ASR (low integration)



  - translate ASR lattice (high integration)



4 GENERATIVE MODEL OF SPEECH TRANSLATION

- Noisy channel model for speech translation
  - Translation from target to source: search for most probable source sentence to have generated the target sentence

Target Speech	*Target Sentence*	*Target Phrase*	*Source Phrase*	*Source Sentence*
$\mathbf{A}$	$\longleftarrow \mathbf{t}_1^{\mathbf{J}}$	$\longleftarrow \mathbf{v}_1^{\mathbf{R}}$	$\longleftarrow \mathbf{u}_1^{\mathbf{K}}$	$\longleftarrow \mathbf{s}_1^{\mathbf{I}}$
Models	$\mathbf{P}(\mathbf{A}\|\mathbf{t}_1^{\mathbf{J}})$	$\mathbf{P}(\mathbf{t}_1^{\mathbf{J}}\|\mathbf{v}_1^{\mathbf{R}})$	$\mathbf{P}(\mathbf{v}_1^{\mathbf{R}}\|\mathbf{u}_1^{\mathbf{K}})$	$\mathbf{P}(\mathbf{u}_1^{\mathbf{K}}\|\mathbf{s}_1^{\mathbf{I}})$	$\mathbf{P}(\mathbf{s}_1^{\mathbf{I}})$
FSMs	$\mathcal{L}$	Ω	Φ	$\mathbf{W}$	$\mathbf{G}$
	ASR Word Lattice	*Target Phrase Segmentation Transducer*	*Phrase Translation, Reordering Transducer*	*Source Phrase Segmentation Transducer*	*Source Language Model*

  - The final translation is given by

$$\widehat{\mathbf{s}_1^{\mathbf{I}}} = \text{argmax}_{\mathbf{s}_1^{\mathbf{I}}} \{ \max_{\mathbf{t}_1^{\mathbf{J}},\mathbf{v}_1^{\mathbf{R}},\mathbf{u}_1^{\mathbf{K}},\mathbf{K}} \mathbf{P}(\mathbf{A},\mathbf{t}_1^{\mathbf{J}},\mathbf{v}_1^{\mathbf{R}},\mathbf{u}_1^{\mathbf{K}},\mathbf{s}_1^{\mathbf{I}}) \} .$$

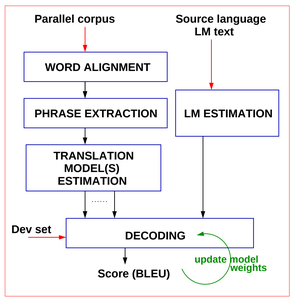
- Implementation with Weighted Finite State Transducers
  - Translation is performed using libraries of standard FST operations
  - Architectures may be limited, but easy to work on components
  - No special-purpose decoder required
  - Standard best-path search through the translation FST $\mathcal{T}$

$$\mathcal{T} = \mathbf{G} \circ \mathbf{W} \circ \mathbf{\Phi} \circ \mathbf{Q}$$

  - Efficient implementation with OpenFST (`www.openfst.org`)

5 TRANSLATION SYSTEM TRAINING

- Sentence-aligned parallel corpora
  - Arabic↔English ∼ 150M words (6M sent.)
  - Chinese↔English ∼ 240M words (10M sent.)
- English monolingual data
  - Used for first-pass and rescoring LM
  - Currently ∼ 5 billion words
- Standard training stages:



  - Minimum Error Training
    - Find optimal models weights
    - Optimize evaluation metric (BLEU) on development set
  - Lattice Rescoring and MBR
    - Rescoring LM: zero cut-off 5-gram
    - Phrase segmentation 2-gram
    - Model 1 lattice-to-string alignment
    - Minimum Bayes Risk decoding of final N-best lists

6 RESULTS

- Text and speech translation performance
  - **BLEU**: Automatic metric that measures similarity between hypothesis and a set of human-translated references
  - Dev/Test size: ∼ 60k words
  - Scores: 4 references for text, 1 reference for audio

Task	Condition	DEV SET	TEST SET
Arabic→English	clean text	54.02	53.70
Chinese→English	clean text	31.07	32.24
Chinese→English	audio reference	18.60	18.10
	ASR 1best	–	16.59

- ASR lattice translation yields slight gains (without ROVER)
- Arabic-English examples

wrfD Hmzp kSf Asm Al$rkp Al<srA}ylyp wtHdyd mA <*A kAn TlbhA qd rfD >w qbl.	
Hamza refused to disclose the name of the Israeli company, and to determine whether the request had been rejected or before.	Ref: Hamza refused to disclose the name of the Israeli company and to reveal whether its its application was accepted or refused.
wH*r AlwzrA' AlmjtmEwn bTlb mn flsTyn mn AstmrAr Al>n$Tp Alnwwyp Al<srA}ylyp...	
the ministers warned the assembled at the request of the Israeli nuclear activities outide ...	Ref: The meeting ministers warned, following a request by Palestine , against the continuation of the Israeli nuclear activities outside...

- Main challenges
  - Reordering. Allowing any phrase movement is not feasible, and constraints produce wrong word order
  - Fluency. Translation often omits connectors
  - Need for more structure (ie. syntax) ?