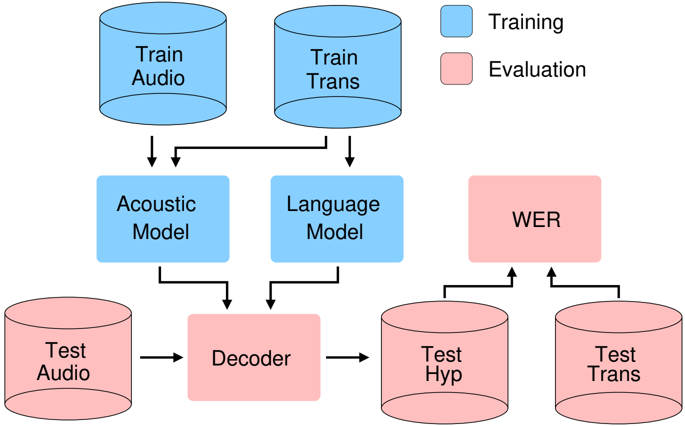
COM4511 Speech Technology: "Alternatives" to Supervised Learning

Anton Ragni

March 9, 2020

Supervised Learning

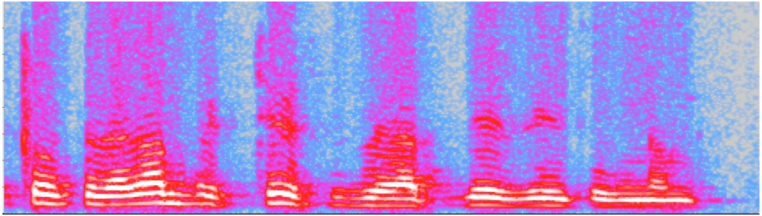
- Standard supervised learning setup



- Not always possible
  - new channel, domain, task
  - most of 7,000 languages spoken in the world lack supervised resources

Anton Ragni

Transcript Availability



`I want a train to Oxford ... uh ... I mean ... to Cambridge`

- Availability and quality of transcripts may vary
  - "accurate" verbatim human annotations

    `I want a train to Oxford um I mean Cambridge`
  - crowd-sourced verbatim human annotations

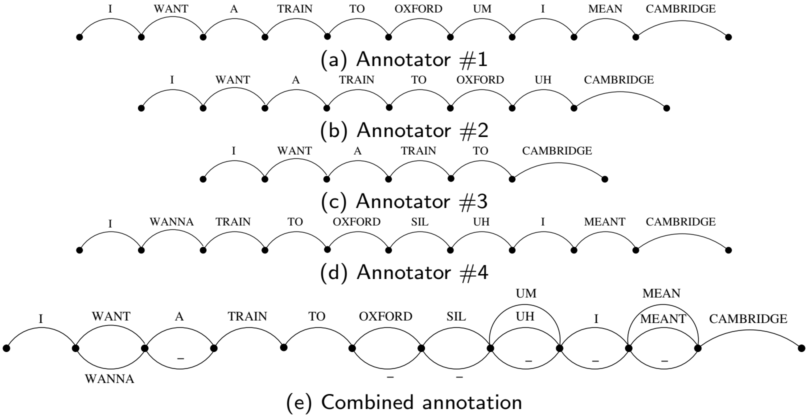
    `I want a train to Oxford uh Cambridge`
  - non-verbatim human annotations (subtitles)

    `I want a train to Cambridge`
  - automatic machine transcriptions

    `I wanna train to Oxford ... uh ... I meant Cambridge`
  - no transcriptions

Anton Ragni

Inter-Annotator Agreement



(a) Annotator #1

(b) Annotator #2

(c) Annotator #3

(d) Annotator #4

(e) Combined annotation

- Improve reliability of "annotation" using multiple "annotators"
  - non-trivial with sequence data
- Need to know how to deal with multiple annotations
  - options?

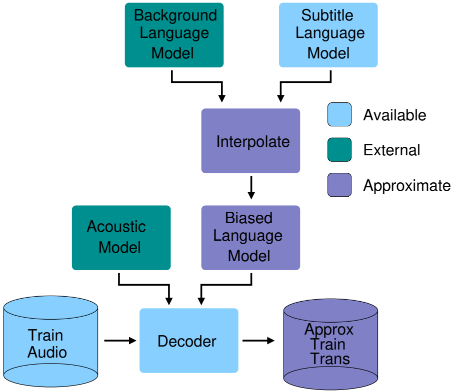
Anton Ragni

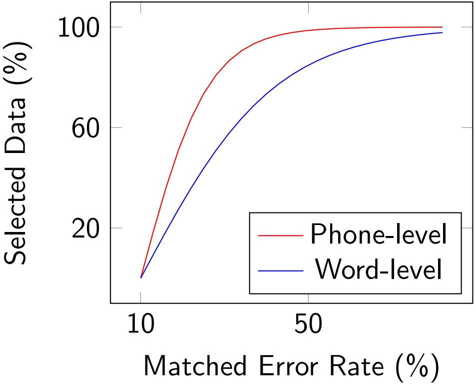
Audio Availability

- Availability and quality of audio may vary
  - quantity: zero, limited, large quantities
  - quality: matched, mismatched (different domain, channel)
- Example:
  - 8 kHz training audio, 16 kHz test audio and vice versa
  - one gender training audio, another gender test audio
  - read training audio, spontaneous test audio
  - 3 hours of training data, 1000000 hours of training data (total and per day)
  - given what you have learnt so far what is the best course of actions?

Anton Ragni

Lightly-Supervised Training





- Subtitles (closed-captions) not suitable for acoustic modelling (why?)
  - transcribe training audio with a biased language model

$$P_{\texttt{bias}}(w_l|\boldsymbol{w}_{l-1:l-n+1}) = \lambda P_{\texttt{sub}}(w_l|\boldsymbol{w}_{l-1:l-n+1}) + (1-\lambda)P_{\texttt{back}}(w_l|\boldsymbol{w}_{l-1:l-n+1})$$

  - assumes acoustic model is available!
- Use portion of derived transcripts for supervised training

Anton Ragni

Matched Error Rate

CC	—	AN	ELABORATE	MEAL
Hyp	BUT	DIDN'T	ELABORATE	—
Error	I	S	—	D

(a) Word-Level Alignment (100% error)

Code	Description
S	Substitution
D	Deletion
I	Insertion

(c) Error types

CC	–	–	A	N	–	–	E	L	A	B	O	R	A	T	E	M	E	A	L
Hyp	D	I	D	N	'	T	E	L	A	B	O	R	A	T	E	–	–	–	–
Error	I	I	S	–	I	I	–	–	–	–	–	–	–	–	–	D	D	D	D

(b) Grapheme-Level Alignment (60% error)

- Use Levenshtein (edit) distance to measure disagreement between CC and hypotheses

$$\text{Error Rate } (\%) = \frac{S + D + I}{N} \cdot 100\%$$

  - possible to compute at different levels (words, graphemes, phonemes)
  - discuss different tradeoffs

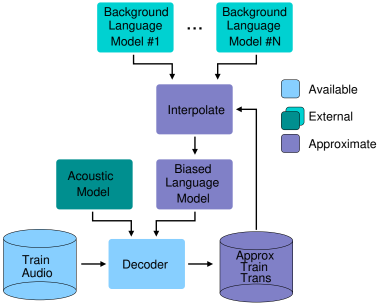
Example: English Broadcast News subtitles*

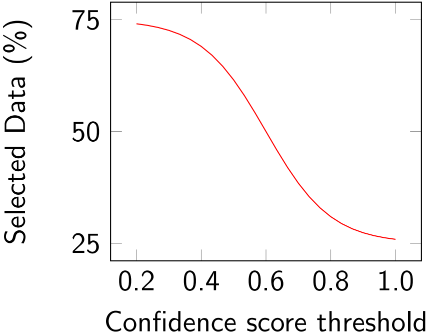
Type	Data (hrs)	WER (%)
Supervised	143	13.8
Lightly supervised	513	13.0
	743	12.4

- Use large quantity of subtitled audio to improve broadcast news transcription
  - external, seed, acoustic model trained on matched domain broadcast news audio

* H. Y. Chan and P. C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training", Proc. ICASSP, 2004.

Unsupervised Training

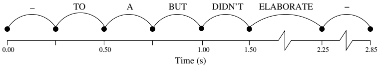




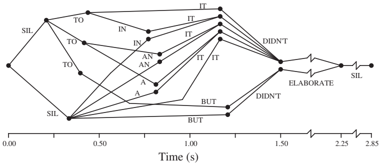
- Create biased language model using unsupervised methods
  - iterative transcription refinement
    1. initialise interpolation weights $\boldsymbol{\lambda}^{(0)}$
    2. transcribe training audio
    3. obtain new interpolation weights $\boldsymbol{\lambda}^{(1)}$ and repeat till convergence
  - maximise confidence score
- Use portion of derived transcripts for supervised training

Anton Ragni

Data Selection Strategies



(a) One path



(b) Multiple paths

- Use confidence scores to select audio with reliable transcripts
  - hard schemes: discard any file with average confidence below set threshold
  - soft schemes: use all files weighing any accumulated statistics with confidence scores
- Options available how much information to use
  - one path: limited ability to rectify transcription errors
  - multiple paths: enables mitigating transcription errors

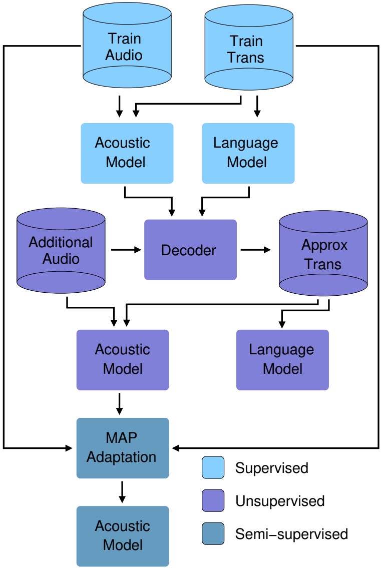
Example: Voice of America and Youtube Data for Limited Resource Languages

Language	Unsupervised Data (WER, %)		
	—	+News	+Youtube
Swahili	38.0	36.5	30.8
Tagalog	36.9	33.8†	—
Somali	57.9	54.3	50.8
Bulgarian	26.5	—	18.0
Lithuanian	27.5	—	21.4

US IARPA MATERIAL programme 2018—

- Use large quantity of web scrapped data to improve transcription accuracy
  - external narrow-band acoustic model
  - billions of words of web text data for language modelling
  - thousands hours of radio news and Youtube
- Different impact for different languages
  - discuss possible reasons

Semi-Supervised Training



- Initial supervised acoustic model
  - produces transcripts for unsupervised data
  - data quantity required varies
- Semi-supervised acoustic model
  - train on merged supervised and unsupervised data
  - alternatively, use MAP adaptation
- Forms of MAP adaptation
  - GMM-HMM:

$$\boldsymbol{\mu}_{j,m}^{\mathtt{map}} = \frac{\sum_{t=1}^{T} \gamma_{t,j,m}^{\mathtt{uns}} \boldsymbol{o}_t^{\mathtt{sup}} + \tau \boldsymbol{\mu}_{j,m}^{\mathtt{uns}}}{\sum_{t=1}^{T} \gamma_{t,j,m}^{\mathtt{uns}} + \tau}$$

    - compare to count smoothing (n-grams)
  - NN-HMM: fine-tune neural network weights

Anton Ragni

Example: Million Hours of Far-Field Data for Amazon Alexa

Type	Data (hrs)	WERR (%)
Supervised	7,000	0
Semi-supervised	100,000	~ 8
	1,000,000	~ 13

- Use vast data quantities to improve Amazon Alexa transcription accuracy
  - approximately 10% relative WER reduction with 100,000 hours
  - diminishing gains past the first 100,000 hours

S. H. K. Parthasarathi, N. Strom, "Lessons from building acoustic models with a million hours of speech",

Summary

- This lecture examined alternatives to supervised learning
  - varying transcript quality
  - varying audio quality and quantity
- Focus on approximate transcription schemes
  - lightly-supervised, semi-supervised and unsupervised learning
  - vary in terms of additionally available data
- Next lectures will look at different aspects of speech technology
  - speech synthesis
  - spoken dialogue systems

Anton Ragni