Automatic Speech Recognition (CS753)

Lecture 20: Discriminative Training for HMMs

Instructor: Preethi Jyothi

Mar 30, 2017

Discriminative Training

Recall: MLE for HMMs

Maximum likelihood estimation (MLE) sets HMM parameters so as to maximise the objective function:

$$\mathcal{L} = \sum_{i=1}^{N} \log P_\lambda(X_i | M_i)$$

where
X_1, ..., X_i, ... X_N are training utterances
M_i is the HMM corresponding to the word sequence of X_i
λ corresponds to the HMM parameters

What are some conceptual problems with this approach?

Discriminative Learning

- *Discriminative models* directly model the class posterior probability or learn the parameters of a joint probability model discriminatively so that classification errors are minimised
    - As opposed to *generative models* that attempt to learn a probability model of the data distribution
- [Vapnik] "***one should solve the (classification/recognition) problem directly and never solve a more general problem as an intermediate step***"

[Vapnik]: V. Vapnik, Statistical Learning Theory, 1998

Discriminative Learning

- Two central issues in developing discriminative learning methods:

    1. Constructing suitable objective functions for optimisation

    2. Developing optimization techniques for these objective functions

Maximum mutual information (MMI) estimation: Discriminative Training

- MMI aims to directly maximise the posterior probability (criterion also referred to as conditional maximum likelihood)

$$
\begin{aligned}
\mathcal{F}_{\mathrm{MMI}} &= \sum_{i=1}^{N} \log P_\lambda(M_i|X_i) \\
&= \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|M_i)P(W_i)}{\sum_{W'} P_\lambda(X_i|M_{W'})P(W')}
\end{aligned}
$$

- P(W) is the language model probability

Why is it called MMI?

- Mutual information $I(X, W)$ between acoustic data X and word labels W is defined as:

$$
\begin{aligned}
I(X,W) &= \sum_{X,W} \Pr(X,W) \log \frac{\Pr(X,W)}{\Pr(X)\Pr(W)} \\
&= \sum_{X,W} \Pr(X,W) \log \frac{\Pr(W|X)}{\Pr(W)} \\
&= H(W) - H(W|X)
\end{aligned}
$$

where H(W) is the entropy of W and H(W|X) is the conditional entropy

Why is it called MMI?

- Assume H(W) is given via the language model. Then, maximizing mutual information becomes equivalent to minimising conditional entropy

$$
\begin{aligned}
H(W|X) &= -\frac{1}{N}\sum_{i=1}^{N} \log \Pr(W_i|X_i) \\
&= -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\Pr(X_i|W_i)\Pr(W_i)}{\sum_{W'} \Pr(X_i|W')\Pr(W')}
\end{aligned}
$$

- Thus, MMI is equivalent to maximizing:

$$
\mathcal{F}_{\mathrm{MMI}} = \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|M_i)P(W_i)}{\sum_{W'} P_\lambda(X_i|M_{W'})P(W')}
$$

MMI estimation

How do we compute this?

$$\mathcal{F}_{\text{MMI}} = \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|M_i)P(W_i)}{\sum_{W'} P_\lambda(X_i|M_{W'})P(W')}$$

- Numerator: Likelihood of data given correct word sequence
- Denominator: Total likelihood of the data given all possible word sequences

Recall: Word Lattices

- A word lattice is a pruned version of the decoding graph for an utterance
- Acyclic directed graph with arc costs computed from acoustic model and language model scores
- Lattice nodes implicitly capture information about time within the utterance

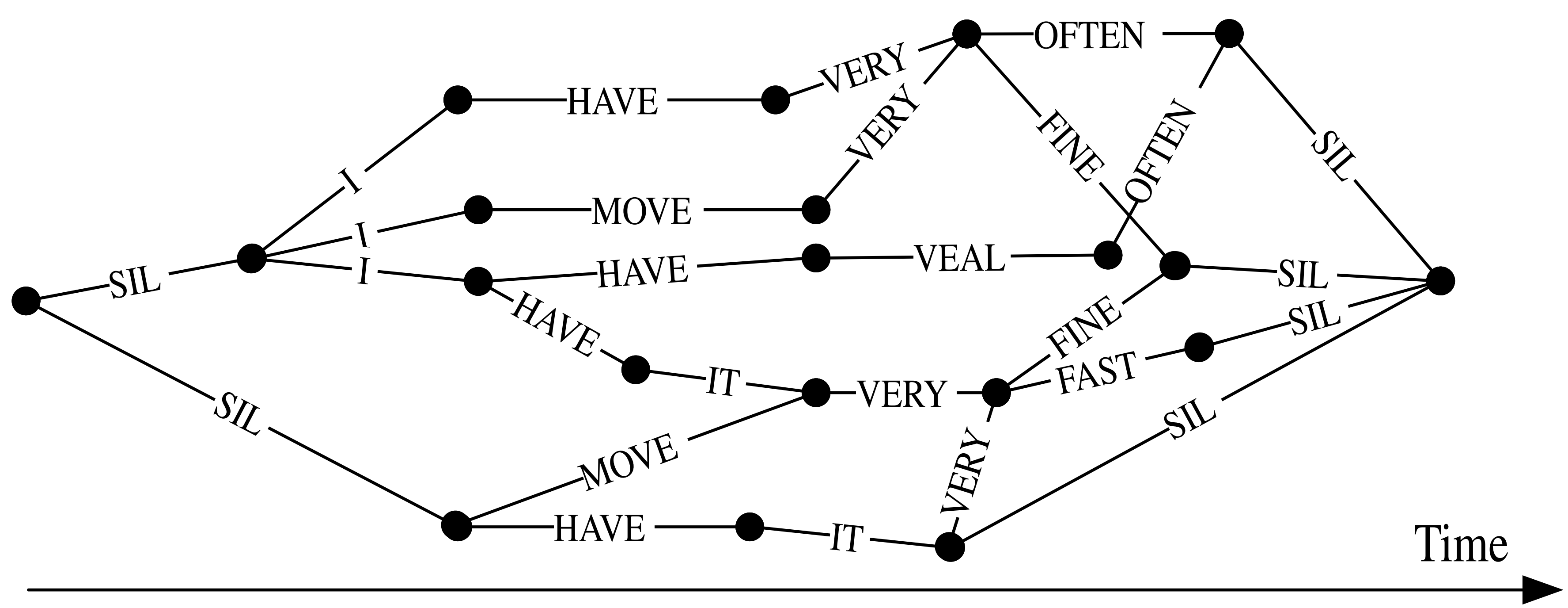

Image from [GY08]: Gales & Young, Application of HMMs in speech recognition, NOW book, 2008

MMI estimation

How do we compute this?

$$\mathcal{F}_{\text{MMI}} = \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|M_i)P(W_i)}{\sum_{W'} P_\lambda(X_i|M_{W'})P(W')}$$

- Numerator: Likelihood of data given correct word sequence
- Denominator: Total likelihood of the data given all possible word sequences
    - Estimate by generating lattices, and summing over all the word sequences in the lattice

MMI Training and Lattices

- Computing the denominator: Estimate by generating lattices, and summing over all the words in the lattice

- Numerator lattices: Restrict G to a linear chain acceptor representing the words in the correct word sequence. Lattices are usually only computed once for MMI training.

- HMM parameter estimation for MMI uses the extended Baum-Welch algorithm [V96,WP00]

- Like HMMs, can DNNs also be trained with an MMI-type objective function? Yes! (More about this next week.)

[V96]:Valtchev et al., Lattice-based discriminative training for large vocabulary speech recognition, 1996

[WP00]: Woodland and Povey, Large scale discriminative training for speech recognition, 2000

MMI results on Switchboard

- Switchboard results on two eval sets (SWB, CHE). Trained on 300 hours of speech. Comparing maximum likelihood (ML) against discriminatively trained GMM systems and MMI-trained DNNs.

	SWB	CHE	Total
GMM ML	21.2	36.4	28.8
GMM MMI	18.6	33.0	25.8
DNN CE	14.2	25.7	20.0
DNN MMI	12.9	24.6	18.8

[V et al.]:Vesely et al., Sequence discriminative training of DNNs, Interspeech 2013

Another Discriminative Training Objective: Minimum Phone/Word Error (MPE/MWE)

- MMI is an optimisation criterion at the sentence-level. Change the criterion so that it is directly related to sub-sentence (i.e. word or phone) error rate.
- MPE/MWE objective function is defined as:

$$\mathcal{F}_{\mathrm{MPE/MWE}} = \sum_{i=1}^{N} \log \frac{\sum_{W} P_\lambda(X_i|M_W)P(W)A(W,W_i)}{\sum_{W'} P_\lambda(X_i|M_{W'})P(W')}$$

where $A(W, W_i)$ is phone/word accuracy of the sentence W given the reference sentence W_i i.e. the total phone count in W_i minus the sum of insertion/deletion/substitution errors of W

MPE/MWE training

$$\mathcal{F}_{\text{MPE/MWE}} = \sum_{i=1}^{N} \log \frac{\sum_{W} P_\lambda(X_i|M_W)P(W)A(W,W_i)}{\sum_{W'} P_\lambda(X_i|M_{W'})P(W')}$$

- The MPE/MWE criterion is a weighted average of the phone/word accuracy over all the training instances
- $A(W, W_i)$ can be computed either at the phone or word level for the MPE or MWE criterion, respectively
- The weighting given by MPE/MWE depends on the number of incorrect phones/words in the string while MMI looks at whether the entire sentence is correct or not

MPE results on Switchboard

- Switchboard results on eval set SWB. Trained on 68 hours of speech. Comparing maximum likelihood (MLE) against discriminatively trained (MMI/MPE/MWE) GMM systems

	SWB	%WER redn
GMM MLE	46.6	-
GMM MMI	44.3	2.3
GMM MPE	43.1	3.5
GMM MWE	43.3	3.3

[V et al.]:Vesely et al., Sequence discriminative training of DNNs, Interspeech 2013

How does this fit within an ASR system?

Estimating acoustic model parameters

- If *A:* speech utterance and O_A: acoustic features corresponding to the utterance *A,*

$$W^* = \arg\max_W P_\lambda(O_A|W)P_\beta(W)$$

- ASR decoding: Return the word sequence that jointly assigns the highest probability to O_A

- How do we estimate λ in $P_\lambda(O_A|W)$?
    - MLE estimation
    - MMI estimation
    - MPE/MWE estimation

Covered in this class

Another way to improve ASR performance:
System Combination

System Combination

- Combining recognition outputs from multiple systems to produce a hypothesis that is more accurate than any of the original systems

- Most widely used technique: ROVER [ROVER].

  - 1-best word sequences from each system are aligned using a greedy dynamic programming algorithm

  - Voting-based decision made for words aligned together

  - Can we do better than just looking at 1-best sequences?

bbn1.ctm	there’s	a	lot	of	@	like	societies	@	@	ruin	engineers	and	lakes
cmu-isl1.ctm	there’s	the	labs	@	@	like	societies	@	for	women	engineers	i	think
cu-htk2.ctm	there’s	the	last	@	@	like	societies	@	true	of	engineers	and	like
dragon1.ctm	was	@	alive	@	the	legal	society	is	for	women	engineers	and	like
sri1.ctm	there’s	a	lot	of	@	like	society’s	@	@	through	engineers	@	like

Image from [ROVER]: Fiscus, Post-processing method to yield reduced word error rates, 1997

Recall: Word Confusion Networks

Word confusion networks are normalised word lattices that provide alignments for a fraction of word sequences in the word lattice

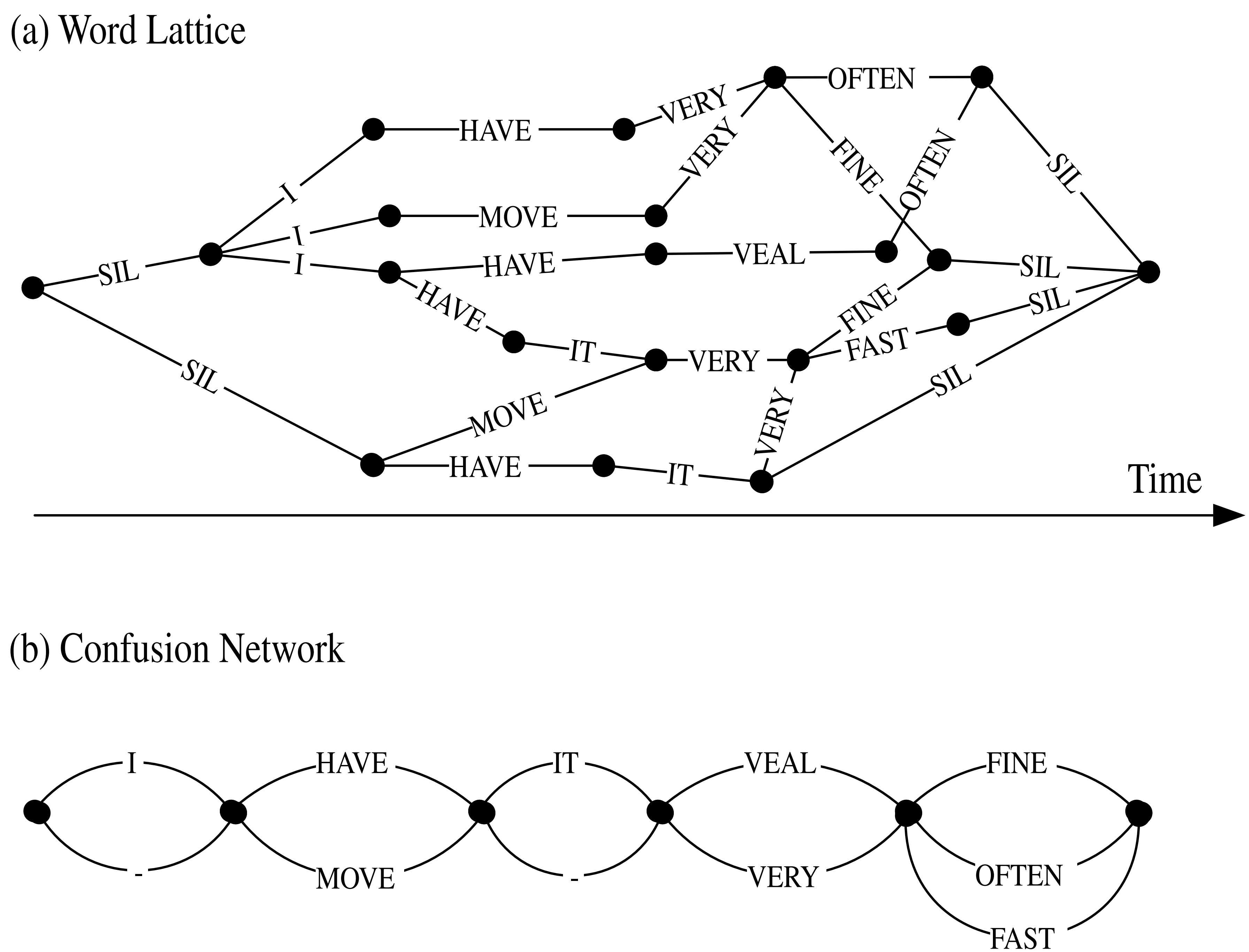

Image from [GY08]: Gales & Young, Application of HMMs in speech recognition, NOW book, 2008

System Combination

- Combining recognition outputs from multiple systems to produce a hypothesis that is more accurate than any of the original systems

- Most widely used technique: ROVER [ROVER].

    - 1-best word sequences from each system are aligned using a greedy dynamic programming algorithm

    - Voting-based decision made for words aligned together

    - Could align confusion networks instead of 1-best sequences

bbn1.ctm	there's	a	lot	of	@	like	societies	@	@	ruin	engineers	and	lakes
cmu-isl1.ctm	there's	the	labs	@	@	like	societies	@	for	women	engineers	i	think
cu-htk2.ctm	there's	the	last	@	@	like	societies	@	true	of	engineers	and	like
dragon1.ctm	was	@	alive	@	the	legal	society	is	for	women	engineers	and	like
sri1.ctm	there's	a	lot	of	@	like	society's	@	@	through	engineers	@	like

Image from [ROVER]: Fiscus, Post-processing method to yield reduced word error rates, 1997