CS230: Lecture 10
Sequence models II

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Today's outline

We will learn how to:

- Automatically **score an NLP model**
- **Improve Machine Translation** results with Beam search
- Build a **speech recognition** application

I. BLEU score

II. Beam Search

III. Speech Recognition

BLEU score

Neural Machine Translation

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

BLEU score

"The baby to be walking by him"

"The baby walks by himself"

Motivation:
Human Evaluation of MT are extensive but expensive.

Goal:
Construct a quick, inexpensive, language independent (and correlates highly with human evaluation) method to automatically evaluate Machine Translation models.

Centrale idea:
"The closer a machine translation is to a professional human translation, the better it is."

Kishore Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, 2002

BLEU score

Needs two ingredients:

- a numerical “translation closeness” metric
- a corpus of good quality human reference translations

In speech recognition:

a successful metric is *word error rate.* BLEU’s closeness metric was built after it.

BLEU score

Machine Translations	
Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party
Candidate 2	It is to insure the troops forever hearing the activity guidebook that party directs.

Human Translations	
Reference 1	It is a guide to action that ensures that the military will forever heed Party commands
Reference 2	It is the guiding principle which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

BLEU score

Unigram count as precision metric:

Machine Translation	
Candidate	the the the the the the the

Human Translations	
Reference 1	The cat is on the mat.
Reference 2	There is a cat on the mat.

$$\text{Standard Unigram Precision} = \frac{\text{\# MT words occurring in any reference HT}}{\text{\# MT words}} = 100\%$$

$$\text{Modified Unigram Precision} = \frac{\text{\# MT words occurring in any reference HT (clipped)}}{\text{\# MT words}} = 2/6 = 33.3\%$$

Kishore Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, 2002

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

BLEU score

Machine Translations	
Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party
Candidate 2	It is to insure the troops forever hearing the activity guidebook that party directs.

Human Translations	
Reference 1	It is a guide to action that ensures that the military will forever heed Party commands
Reference 2	It is the guiding principle which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

$$\text{Modified Unigram Precision (1)} = \frac{\text{\# MT words occurring in any reference HT (clipped)}}{\text{\# MT words}} = 17/18 = 94\%$$

$$\text{Modified Unigram Precision (2)} = \frac{\text{\# MT words occurring in any reference HT (clipped)}}{\text{\# MT words}} = 8/14 = 57\%$$

Kishore Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, 2002

BLEU score

Generalizing to modified n-gram precision metric:

$$\text{Modified n-gram precision} = \frac{\text{\# MT n-grams occurring in any reference HT (clipped)}}{\text{\# MT n-grams}}$$

Machine Translations	
Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party
Candidate 2	It is to insure the troops forever hearing the activity guidebook that party directs.
Human Translations	
Reference 1	It is a guide to action that ensures that the military will forever heed Party commands
Reference 2	It is the guiding principle which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

Modified bi-gram precision (Candidate 1) = 10/17
Modified bi-gram precision (Candidate 2) = 1/13

Kishore Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, 2002

BLEU score

Generalizing from a sentence precision, to a corpus precision:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{ngram \in C} Count_{clip}(ngram)}{\sum_{C' \in \{Candidates\}} \sum_{ngram' \in C} Count(ngram')}$$

BLEU Score

Figure 1: Distinguishing Human from Machine

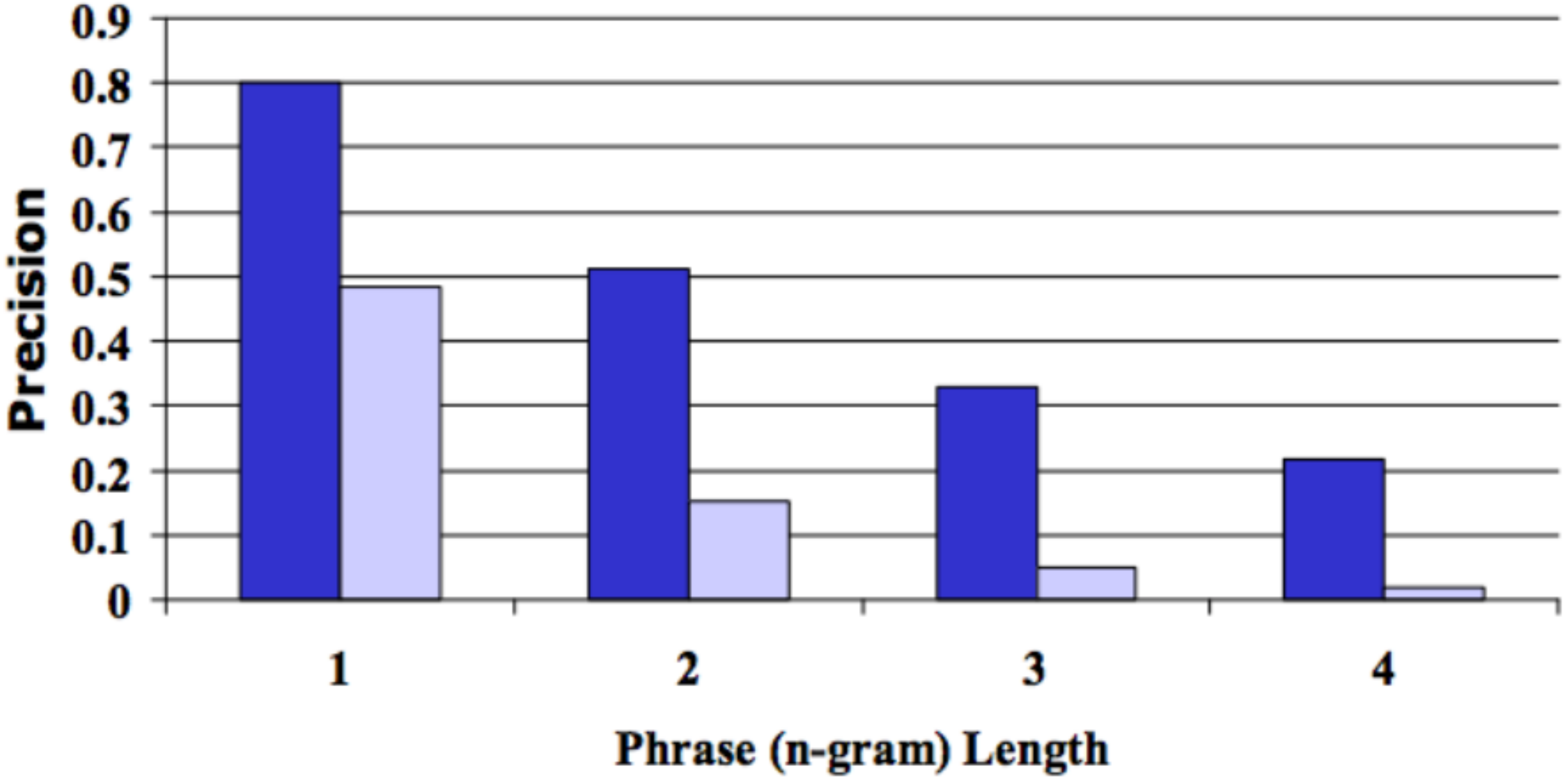
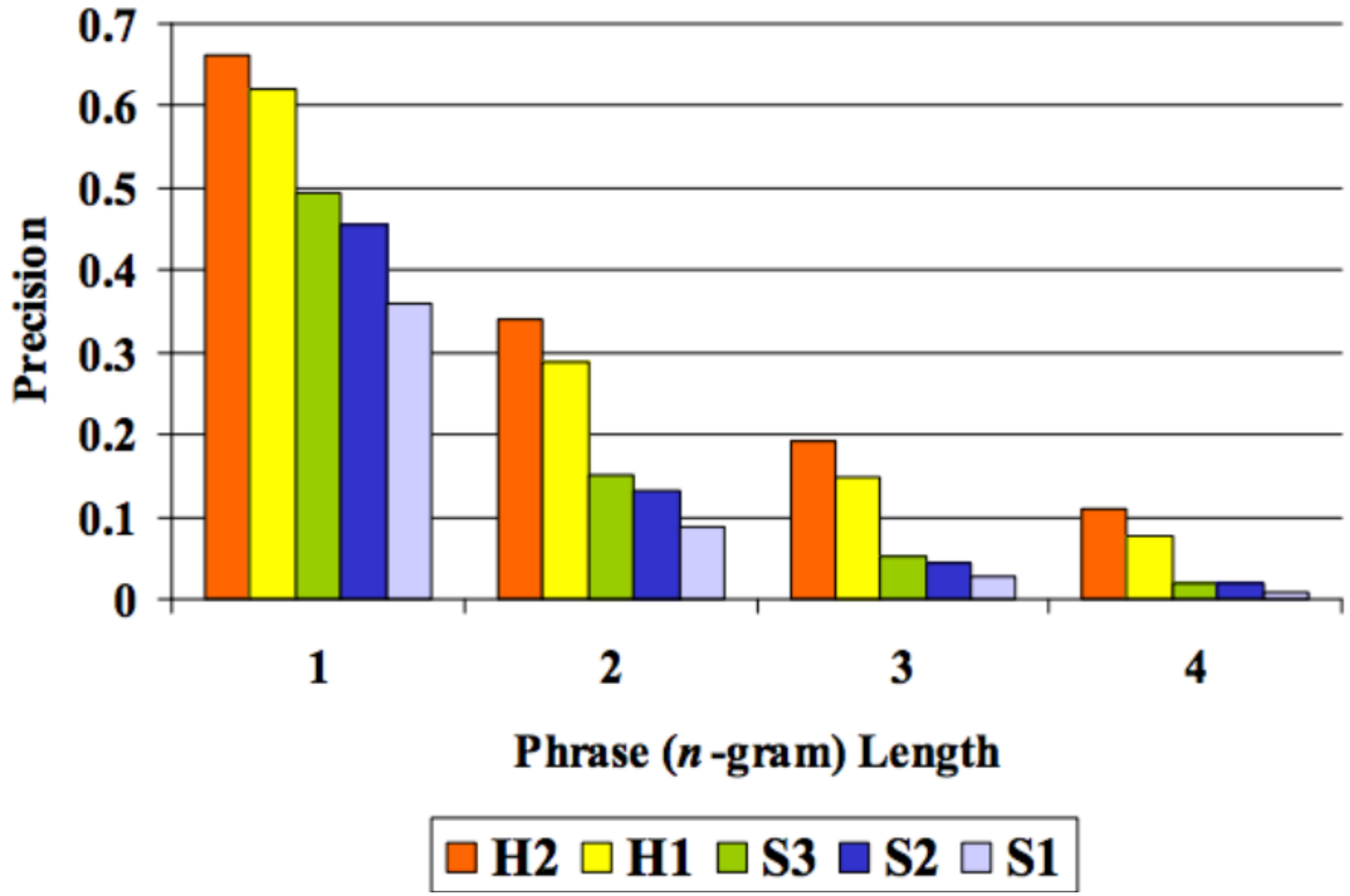


Figure 2: Machine and Human Translations



Geometric average of modified n-grams precision measures = $\exp(\sum_{n=1}^{N} w_n \log(p_n)) = p_1^{w_1} p_2^{w_2} \dots p_N^{w_N}$

Kishore Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, 2002

BLEU Score

Sentence length

too short translation

Machine Translations	
Candidate	of the

Human Translations	
Reference 1	It is a guide to action that ensures that the military will forever heed Party commands
Reference 2	It is the guiding principle which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

too long translation

Machine Translations	
Candidate 1	I always invariably perpetually do
Candidate 2	I always do

Human Translations	
Reference 1	I always do
Reference 2	I invariably do
Reference 3	I perpetually do

Conclusion: *longer translations are already penalized by the modified n-gram precision measure, not the too short translations*

$$\text{BP} = \text{Brevity penalty} = \text{decaying exponential} \sim \frac{\text{Total length of the machine's translation}}{\text{Total length of the candidate translation corpus}}$$

BLEU Score

BLEU definition:

$$BLEU = (p_1^{w_1} p_2^{w_2} \dots p_N^{w_N}) \cdot BP \quad \text{where} \quad BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{if } c <= r \end{cases}$$

log(BLEU) definition:

$$\log(BLEU) = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log(p_n)$$

Kishore Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, 2002

Beam search

Questions

- What is the main advantage of Beam search compared to other search algorithms?

It is fast, and requires less computations.

- What is the main disadvantage of Beam search compared to other search algorithms?

It may not result in the optimal solution in terms of probability.

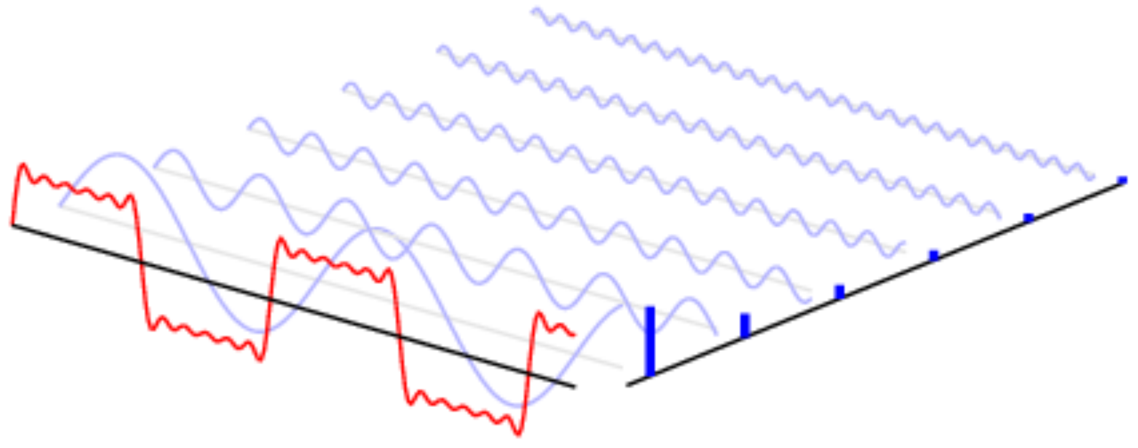
- What is the time and memory complexity of Beam search?

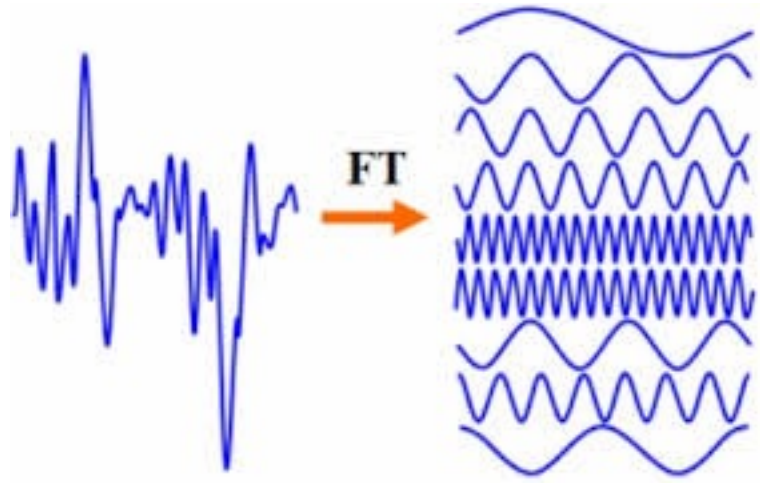
*It is O(b*Tx) in memory and O(b*Tx) in time.*

Speech Recognition Pipeline

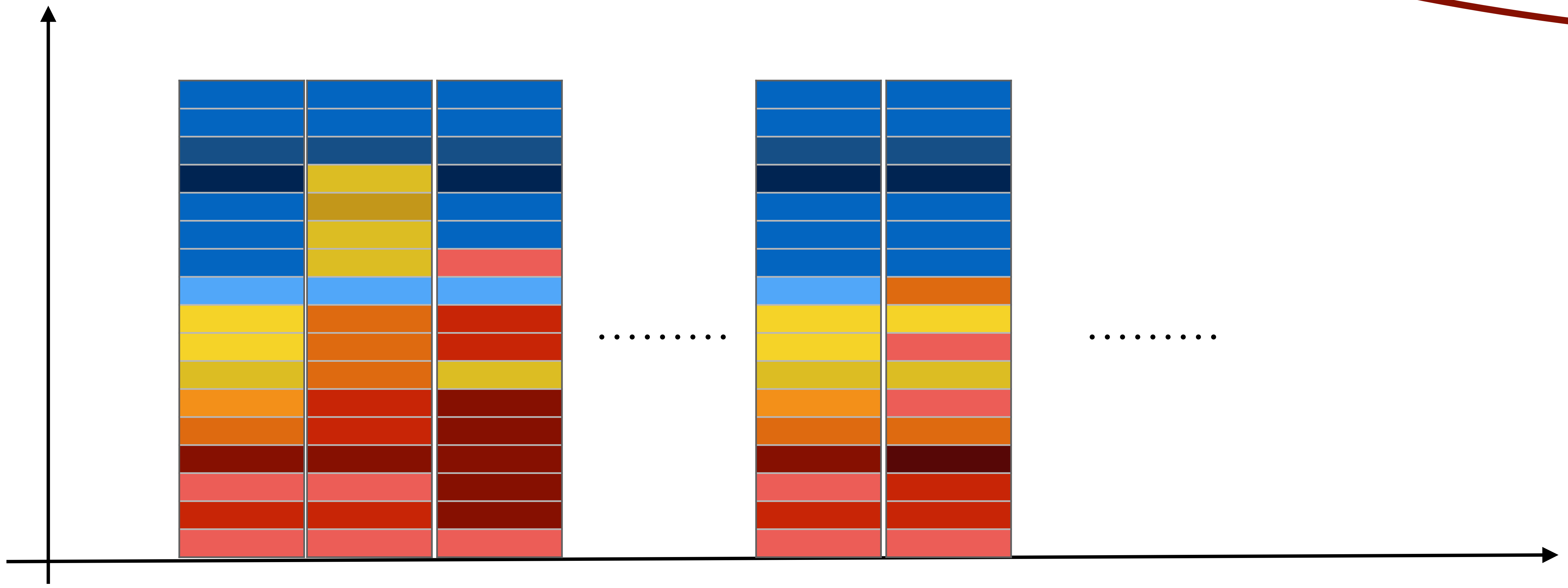
Audio Data:

X? → model? → Y?





frequency

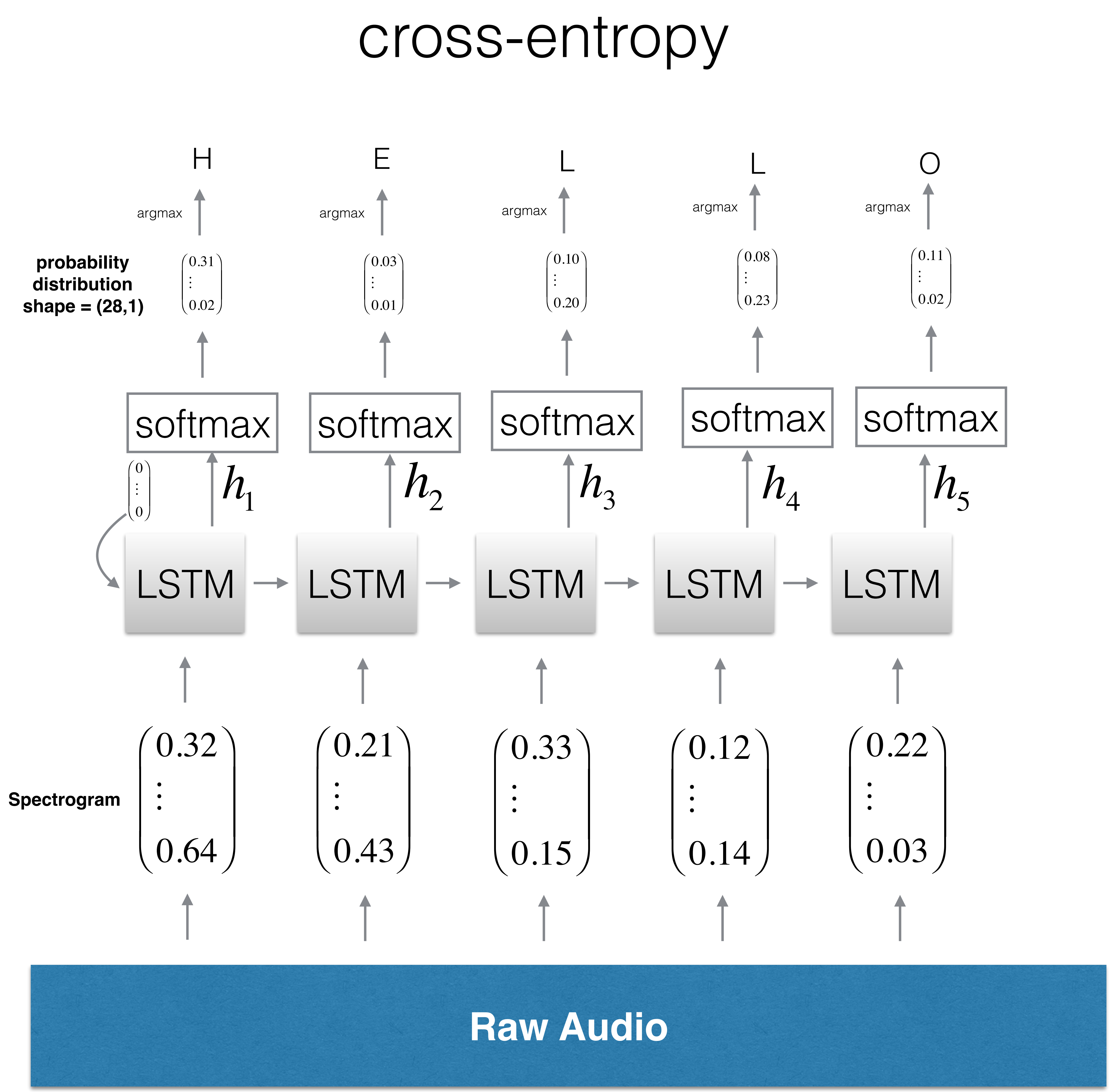


time

"Hello"

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

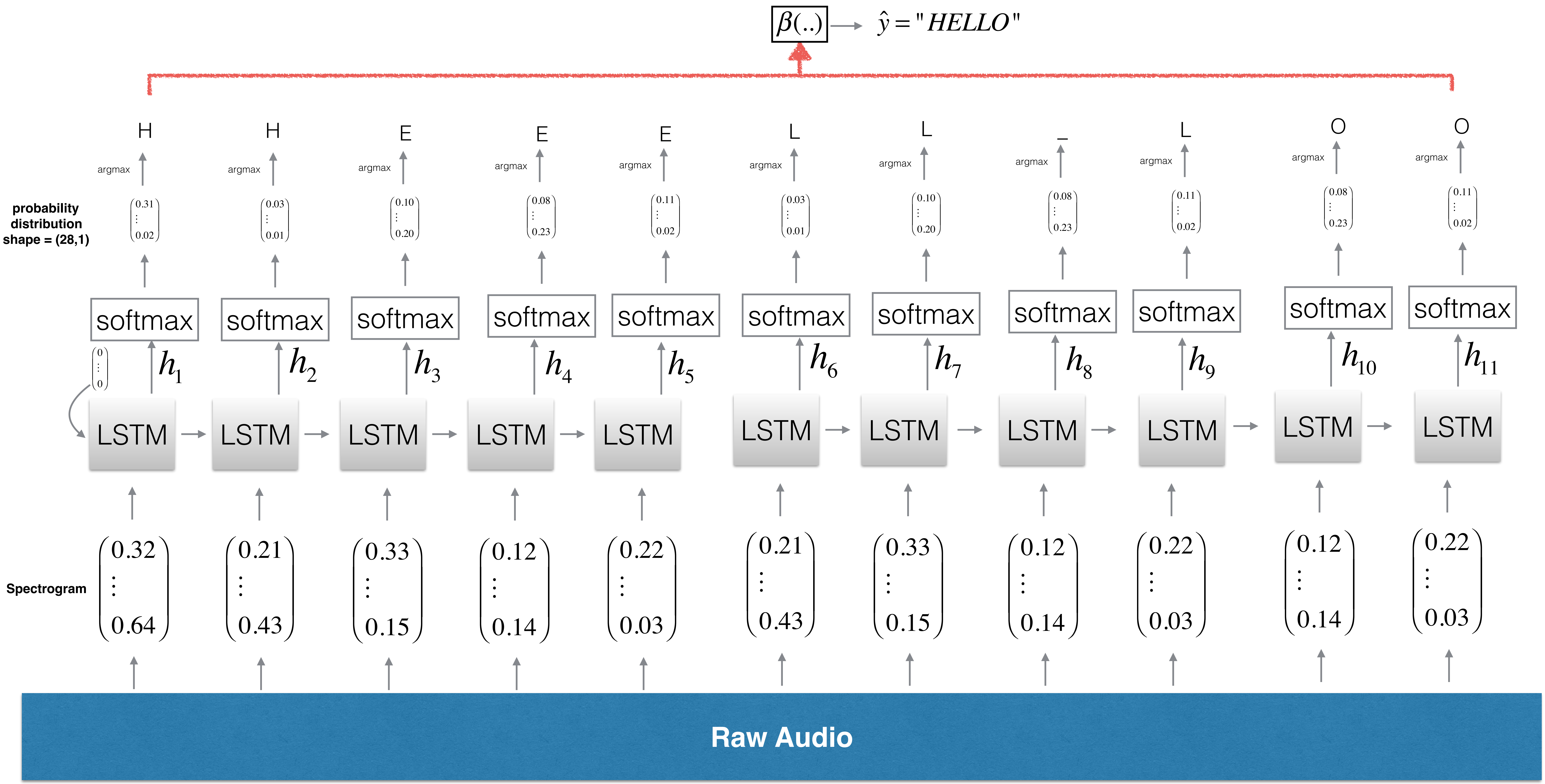
Speech Recognition



This never happens in practice because:

input length ≠ output length

Speech Recognition https://distill.pub/2017/ctc/



Graves et al., 2006, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Examples

$$\beta(HH\_EEEE\_LL\_LOO) = "HELLO"$$

$$\beta(H\_E\_L\_LOO) = "HELLO"$$

$$\beta(H\_\_LLL\_OO) = "HELO"$$

$$\beta(BBAA\_NA\_NN\_\_AA\_A) = "BANANAA"$$

Speech Recognition

Independence assumption

$$P(c_1|x) = \prod_{t=1}^{T_x} P(c_1^{\langle t \rangle}|x)$$

$$P(c|x) = \begin{bmatrix} 0.13 & HH\_E\_\_L\_LL\_OO \\ 0.04 & H\_EE\_L\_LL\_HO \\ 0.03 & HH\_E\_\_L\_LL\_OO \\ 0.01 & H\_EE\_\_L\_LL\_OO \\ 0.001 & H\_I\_ILL\_L\_OOOO \\ \vdots & \vdots \end{bmatrix}$$

$$P(y|x) = \sum_{c:\beta(c)=y} P(c|x)$$

$$P("HELLO") =$$

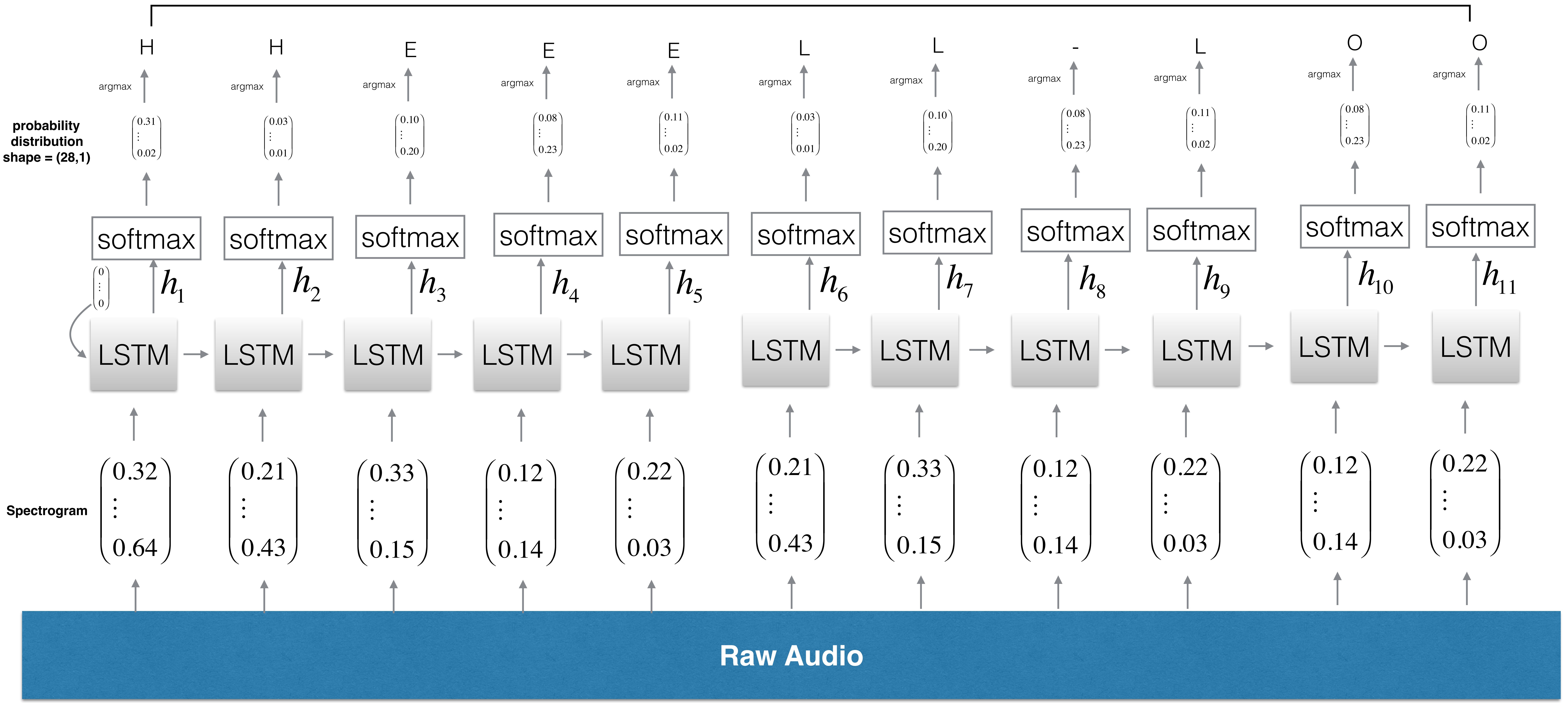
Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Speech Recognition

Loss?

$$\beta(..) \rightarrow \hat{y} = "HELLO"$$

$$L = -\sum_i \log P(\hat{y}^{(i)} | x^{(i)}) = -\sum_i \log\left(\sum_{c:\beta(c)=\hat{y}^{(i)}} P(c|x^{(i)}) \right)$$



Graves et al., 2006, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Speech Recognition

Inference?

BEAM SEARCH > MAX DECODING :)

Graves et al., 2006, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Speech Recognition

$$L = -\sum_i \log P(\hat{y}^{(i)} | x^{(i)}) = -\sum_i \log\left(\sum_{c:\beta(c)=\hat{y}^{(i)}} P(c | x^{(i)}) \right)$$

Implementations of CTC loss

tf.nn.ctc_loss(…)
Keras -> Custom loss

Graves et al., 2006, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Speech Recognition

Closing questions:

- How to incorporate information about the future?

A good way to efficiently incorporate future information in speech recognition is still an open problem.

- What's the consequence of $P(c_1|x) = \prod_{t=1}^{T_x} P(c_1^{\langle t \rangle}|x)$ (conditional independence)?

A model like CTC may have trouble producing such diverse transcripts for the same utterance because of conditional independence assumptions between frames.

But, on the other hand, it makes the model more robust to a change of settings.

- What is the problem with our output $\hat{y}$?

CTC model makes a lot of spelling and linguistic mistakes because P(y\x) directly models audio data. Some words are hard to spell based on their audios.

- Can you think of any practical applications leveraging this model?

Lipreading.

Assael et al., LipNet: end-to-end sentence-level lipreading, 2016
Hannun, "Sequence Modeling with CTC", Distill, 2017.
Chan et al.,2015, Listen, Attend and Spell