Foundations of Natural Language Processing
Lecture 6a
Language Models: Evaluation I

Alex Lascarides

School of informatics

Recap: Language models

- **Language models** tell us $P(\vec{w}) = P(w_1 \ldots w_n)$: *How likely to occur is this sequence of words?*

  Roughly: *Is this sequence of words a "good" one in my language?*

- LMs are used as a component in applications such as speech recognition, machine translation, and predictive text completion.

- To reduce sparse data, N-gram LMs assume words depend only on a fixed-length history, even though we know this isn't true.

Evaluating a language model

- Intuitively, a trigram model captures more context than a bigram model, so should be a "better" model.
- That is, it should more accurately predict the probabilities of sentences.
- But how can we measure this?

Two types of evaluation in NLP

- **Extrinsic**: measure performance on a downstream application.
  - For LM, plug it into a machine translation/ASR/etc system.
  - The most reliable evaluation, but can be time-consuming.
  - And of course, we still need an evaluation measure for the downstream system!

- **Intrinsic**: design a measure that is inherent to the current task.
  - Can be much quicker/easier during development cycle.
  - But not always easy to figure out what the right measure is: ideally, one that correlates well with extrinsic measures.

Let's consider how to define an intrinsic measure for LMs.

A Straw Man: Accuracy

- Test corpus T is set of n-word sequences. For each sequence $w_1 \ldots w_n$ in T, LM observes $w_1 \ldots w_{n-1}$ and predicts $\hat{w}_n$.

- Accuracy:

$$\sum_{\vec{w} \in T} \frac{\hat{w_n} = w_n}{|T|}$$

Problem:

- $\hat{w}_n$ may be a perfectly good guess, even when $\hat{w}_n \neq w_n$.

A Better Measure: Entropy

- Definition of the **entropy** of a random variable X:

  $H(X) = \sum_x -P(x)\ \log_2 P(x)$

- Intuitively: a measure of uncertainty/disorder

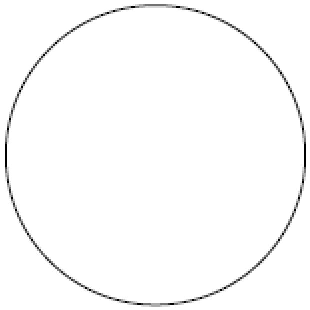
- Also: the expected value of $-\log_2 P(X)$

Entropy Example

One event (outcome)

$$P(a) = 1$$

$$H(X) = -1\log_2 1$$
$$= 0$$

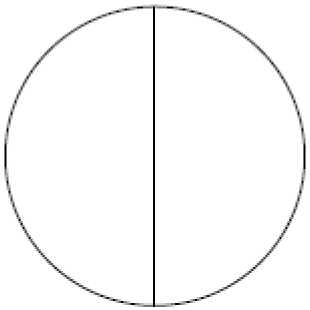


Entropy Example

2 equally likely events:

$$P(a) = 0.5$$
$$P(b) = 0.5$$

$$\begin{aligned} H(X) &= -0.5 \log_2 0.5 - 0.5 \log_2 0.5 \\ &= -\log_2 0.5 \\ &= 1 \end{aligned}$$

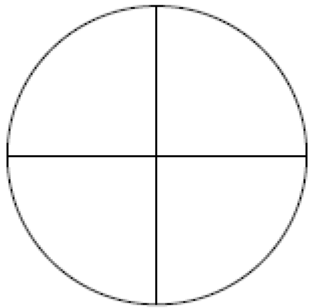


Entropy Example

4 equally likely events:

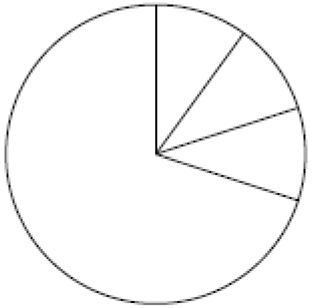
$$P(a) = 0.25$$
$$P(b) = 0.25$$
$$P(c) = 0.25$$
$$P(d) = 0.25$$

$$\begin{aligned} H(X) = & -0.25\log_2 0.25 - 0.25\log_2 0.25 \\ & -0.25\log_2 0.25 - 0.25\log_2 0.25 \\ = & -\log_2 0.25 \\ = & \; 2 \end{aligned}$$



Entropy Example

$P(a) = 0.7$
$P(b) = 0.1$
$P(c) = 0.1$
$P(d) = 0.1$



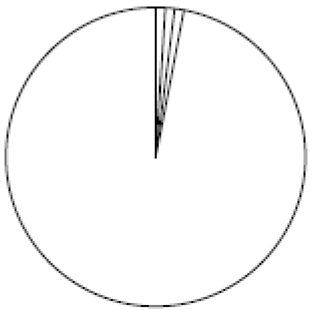
3 equally likely events and one more likely than the others:

$$
\begin{aligned}
H(X) = & -0.7\log_2 0.7 - 0.1\log_2 0.1 \\
& -0.1\log_2 0.1 - 0.1\log_2 0.1 \\
= & -0.7\log_2 0.7 - 0.3\log_2 0.1 \\
= & -(0.7)(-0.5146) - (0.3)(-3.3219) \\
= & \, 0.36020 + 0.99658 \\
= & \, 1.35678
\end{aligned}
$$

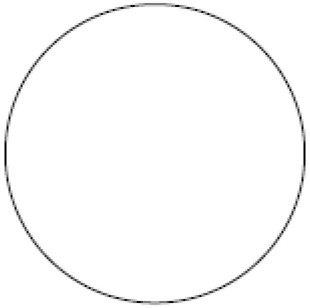
Entropy Example

3 equally likely events and one much more likely than the others:

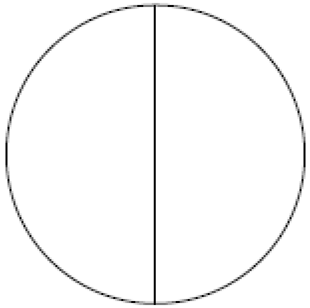
$$P(a) = 0.97$$
$$P(b) = 0.01$$
$$P(c) = 0.01$$
$$P(d) = 0.01$$



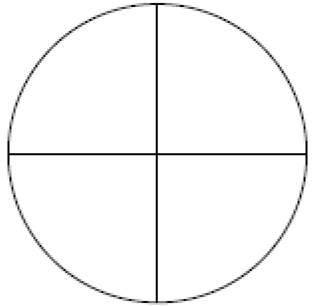
$$\begin{aligned}
H(X) &= -0.97\log_2 0.97 - 0.01\log_2 0.01 \\
&\quad - 0.01\log_2 0.01 - 0.01\log_2 0.01 \\
&= -0.97\log_2 0.97 - 0.03\log_2 0.01 \\
&= -(0.97)(-0.04394) - (0.03)(-6.6439) \\
&= 0.04262 + 0.19932 \\
&= 0.24194
\end{aligned}$$



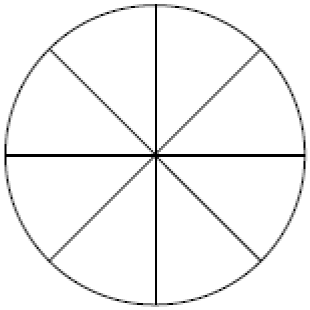
$H(X) = 0$



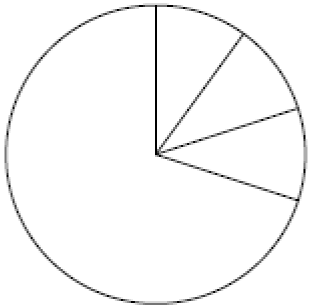
$H(X) = 1$



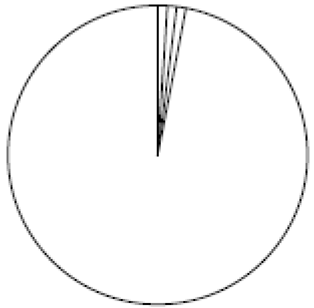
$H(X) = 2$



$H(X) = 3$



$H(X) = 1.35678$



$H(X) = 0.24194$

Summary

- We can't evaluate an LM with accuracy metrics.
- Entropy, however, measures confidence in the model's predictions, and this is an appropriate metric.
- There may be occasions where the model is confident, but wrong.
- But practical experience suggests entropy-based metrics correlate with extrinsic evaluation.

Next time: Evaluation II.