Information Theoretic Approaches to the Study of Language

Lecture 2

Matthew W. Crocker
Vera Demberg

Summer 2017

Language …

- Let random variables span over linguistic events:
  - Phonemes, letters, syllables, words …
- Such events are not uniformly distributed, and are conditioned by the linguistic context
  - We know this from linguistic theories at all levels
- To determine the entropy of language we need probabilistic models of linguistic events in context

Language Modeling

- Problem: determine the probability of next word, given the previous words
  - Find W_n which maximises $P(W_n|W_1 \ldots W_{n-1})$
- The better the model encodes the (probabilistic) structure of language, the more accurate it will be
- But building language models is difficult!

Language Modeling

- Find W_n which maximises P(W_n|W_1 ... W_{n-1})
  - Not practical, because of sparse data
  - Group contexts into equivalence classes, or “bins”
- Markov assumption: W_n depends on immediately preceding words:
  - N-grams: use *N-1* preceding words
- Problem: If Vocabulary is 20 000 then

N-Gram	Parameters
0th Order	20K
1st Order	400M
2nd Order	8T
3rd Order	1.6×10^{17}

Maximal likelihood estimation (MLE)

- Conditional Probabilities:

$$P(w_n \mid w_1 \ldots w_{n-1}) = \frac{P(w_1 \ldots w_n)}{P(w_1 \ldots w_{n-1})}$$

- Estimate P, using relative frequency:

$$P_{MLE}(w_1 \ldots w_n) = \frac{C(w_1 \ldots w_n)}{N}$$

- Counts are obtained from a training corpus

$$P_{MLE}(w_n \mid w_1 \ldots w_{n-1}) = \frac{C(w_1 \ldots w_n)}{C(w_1 \ldots w_{n-1})}$$

Entropy rate

Since information in a message depends on message length, we often normalize to the per-letter/per-word entropy rate:

$$H_{rate} = \frac{1}{n} H(X_1,...,X_n) = \frac{1}{n} H(X_{1n}) = -\frac{1}{n} \sum_{X_{1n}} p(x_{1n}) \log_2 p(x_{1n})$$

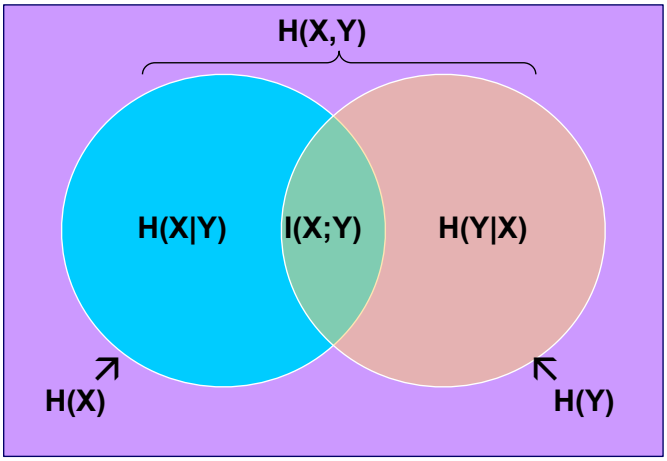
- "Language" is a stochastic process generating a sequence of tokens, *L=(Xi)* e.g., all the words you hear, utter, appear in Die Zeit, etc...
- We define the entropy of the language as the entropy rate for that process:

$$H_{rate}(L) = \lim_{n \to \infty} \frac{1}{n} H(X_1,..., X_n)$$

- Recall: $H(X_1,..., X_n) = H(X_1) + H(X_2 \mid X_1) + ... + H(X_n \mid X_1,..., X_{n-1})$
- Or, "the entropy rate of language is the limit of the entropy rate of a sample of the language, as the sample gets longer and longer" (Manning &Schütze)

Mutual Information

- Recall: chain rule for entropy:
  H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)
- Therefore: H(X) - H(X|Y) = H(Y) - H(Y|X) = I(X;Y)
- Mutual Information:
  The reduction in uncertainty for one variable due to knowing about another.



Mutual Information

$$I(X;Y) = H(X) - H(X \mid Y)$$
$$= H(X) + H(Y) - H(X,Y)$$
$$= \sum_x p(x)\log_2 \frac{1}{p(x)} + \sum_y p(y)\log_2 \frac{1}{p(y)} + \sum_{x,y} p(x,y)\log_2 p(x,y)$$
$$= \sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}$$

- Symmetric, non-negative measure of common information
- Measures the distance of a joint distribution from independence
- I(X;Y) = 0 when X,Y are independent
- MI grows as a function of both dependence and entropy

Simplified Polynesian

? Recall the following per-syllable distribution:
? Calculate I(V;C):

	p	t	k	
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
	1/8	3/4	1/8	

$$I(V;C) = H(V) - H(V \mid C)$$
$$H(V) = 2 \times \frac{1}{4}\log_2 4 + \frac{1}{2}\log_2 2 = \frac{3}{2}$$
$$H(V \mid C) = \frac{11}{8}$$
$$I(V;C) = \frac{12}{8} - \frac{11}{8} = \frac{1}{8}$$

$$H(V \mid C) = \sum_{c=p,t,k} p(C=c)H(V \mid C=c)$$
$$= \frac{1}{8}H(V \mid p) + \frac{1}{8}H(V \mid k) + \frac{3}{4}H(V \mid t)$$
$$= \frac{1}{8}H(\frac{1}{2},\frac{1}{2},0) + \frac{1}{8}H(\frac{1}{2},0,\frac{1}{2}) + \frac{3}{4}H(\frac{1}{2},\frac{1}{4},\frac{1}{4})$$

$$I(V;C) = \sum_{x,y} p(v,c)\log_2 \frac{p(v,c)}{p(v)p(c)}$$
$$= \frac{1}{16}\log_2 \frac{\frac{1}{16}}{\frac{1}{16}} + \frac{3}{8}\log_2 \frac{\frac{3}{8}}{\frac{3}{8}} + \frac{1}{16}\log_2 \frac{\frac{1}{16}}{\frac{1}{16}} + \frac{1}{16}\log_2 \frac{\frac{1}{16}}{\frac{1}{32}} + \frac{3}{16}\log_2 \frac{\frac{3}{16}}{\frac{3}{16}} + \frac{3}{16}\log_2 \frac{\frac{3}{16}}{\frac{3}{16}} + \frac{1}{16}\log_2 \frac{\frac{1}{16}}{\frac{1}{32}}$$
$$= \frac{1}{16} + \frac{1}{16} = \frac{1}{8}$$

Mutual Information

- (Average) Mutual Information: a measure of the reduction in uncertainty for one random variable due to knowing about another:

$$I(X;Y) = H(X) - H(X \mid Y)$$
$$= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

- Pointwise Mutual Information of two individual elements as a measure of association:

$$I(x',y') = \log_2 \frac{p(x',y')}{p(x')p(y')}$$
$$= \log_2 \frac{p(x' \mid y')}{p(x')} \qquad = \log_2 \frac{p(y' \mid x')}{p(y')}$$

Computing PMI

- Compute the probabilities using ML estimation:

$$I(x',y') = \log_2 \frac{p(x',y')}{p(x')p(y')}$$
$$= \log_2 \frac{\frac{c(w^1 w^2)}{N}}{\frac{c(w^1)}{N} \times \frac{c(w^2)}{N}} \qquad = \log_2 \frac{N \times c(w^1 w^2)}{c(w^1)c(w^2)}$$

- Simple example: $I(new,companies) = \log_2 \frac{14307676 \times 8}{15828 \times 4675} \approx .63$

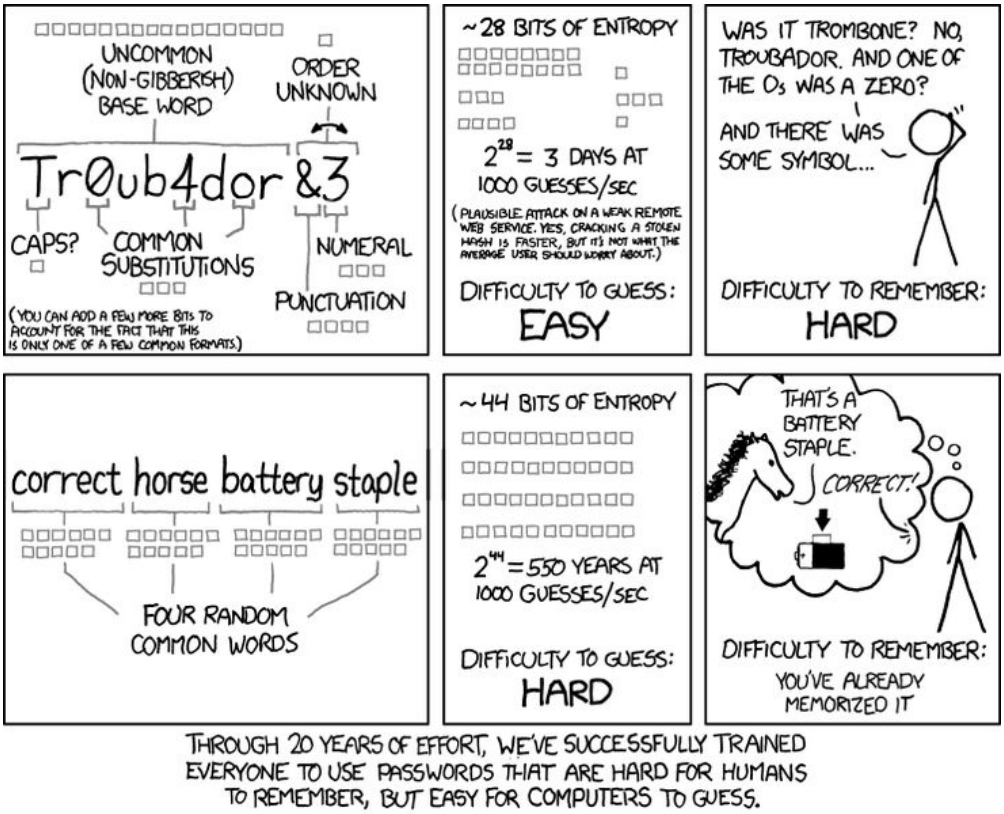
	W_1=new	$W_1 \neq$new	
W_2=companies	8	4667	4675
$W_2 \neq$companies	15820	14287181	14303001
	15828	14291848	14307676

More on Mutual Information

- MI can provide a ranking of possible collocations:

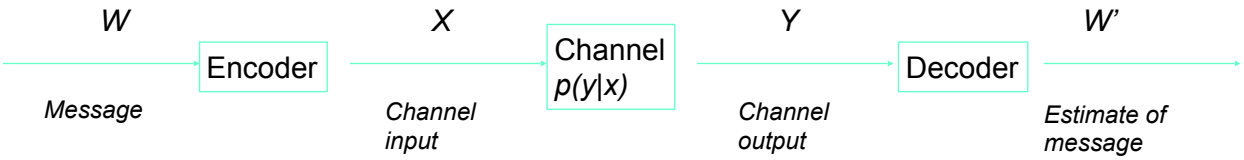
$$I(Bette, Midler) = \log_2 \frac{14307688 \times 20}{27 \times 41} \approx 17.98$$

I(w1,w2)	C(w1)	C(w2)	C(w1 w2)	Word 1	Word 2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	1569	20	time	last



Communication Systems (Shannon, 1948)

- Entropy tells us about the most efficient encoding of a message
  - What about the transmission of messages?
- Transmission can be modeled using a noisy channel:
  - A message W is encoded as a string X
  - X is transmitted through a channel according to a distribution p(y|x)
  - The resulting string Y is decoded, yielding an estimate of the message W'



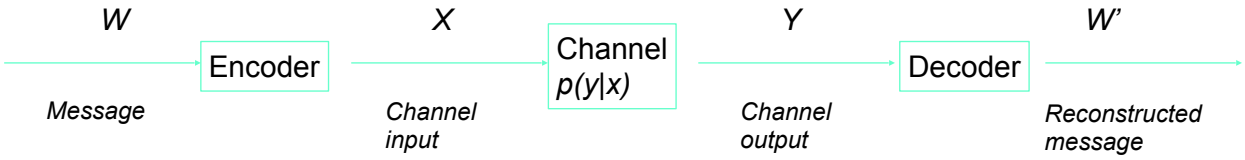
Channel Capacity

- The capacity of the channel is the number of bits on average that is can transmit, as determined by noise in the channel
- Discrete Channel: A discrete channel consists of an input alphabet X, an output alphabet Y, and a probability distribution p(y|x) the expresses the probability of observing symbol y given that x was sent.
- The *channel capacity* of a discrete channel is:

$$C = \max_{p(X)} I(X;Y)$$

- The capacity of a channel is the maximum of the mutual information of X and Y over all input distributions of the input p(x)

Noiseless Binary Channel



- Assume a channel whose input is reproduced exactly at the output

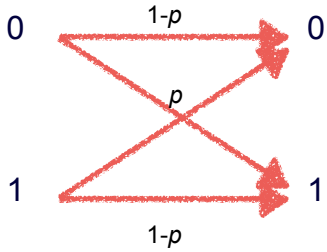
- Channel capacity of a noiseless binary channel:

$$C = \max_{p(X)} I(X;Y) = 1 \text{ bit}$$

- This maximum is achieved when p(0) = 0.5 and p(1) = 0.5
  - Since the uniform distribution maximizes entropy

Noisy Channel Model

- Recall: MI increases with entropy
- Entropy is maximised when I/O distribution is uniform:
  - So H(Y) = 1bit
- Confirm H(Y|X) = H(p):
  - Under what circumstances does this hold
  - If p=.5, H(p)=1, C = I(X;Y) = 1-1 = 0 *useless*
  - If p=0, H(p)=0, C = I(X;Y) = 1-0 = 1 *perfect*



- Channel capacity:

$$I(X;Y) = H(Y) - H(Y \mid X) = H(Y) - \sum_x p(x) H(Y \mid X = x)$$

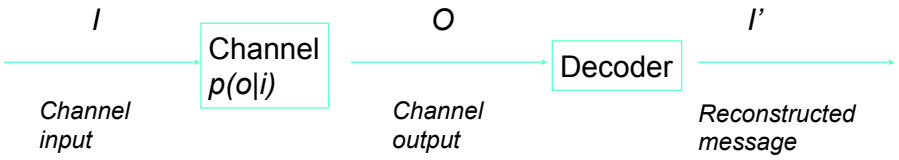
$$= H(Y) - \sum_x p(x) H(p) = H(Y) - H(p) \le 1 - H(p)$$

Therefore,

$$C = \max_{p(X)} I(X;Y) = 1 - H(p) \text{ bits}$$

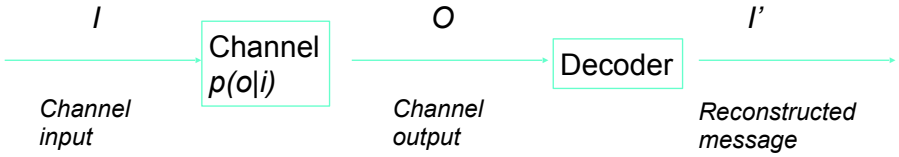
Applications of the Noisy Channel Model

- The noisy channel can be applied to decoding processes involving linguistic information.
- A typical way of formulating such a problem is:
  - Assume some linguistic input I;
  - I is transmitted through a noisy channel with the probability distribution p(o|i);
  - The resulting output O is decoded, yielding an estimate of the Input I'



Applying the Noisy Channel Model

- In most situations in linguistics, we cannot control the encoding:



- We want to find the most likely input for an observed output:

$$\hat{I} = \arg\max_i p(i \mid o)$$

- But, $p(i|o)$ is often difficult to estimate directly and reliably, so recall:

$$p(i \mid o) = \frac{p(o \mid i)\,p(i)}{p(o)}$$

- Therefore:

$$\hat{I} = \arg\max_i \frac{p(i)\,p(o \mid i)}{p(o)} = \arg\max_i p(i)\,p(o \mid i)$$

Widely used applications:

$$\hat{I} = \arg\max_{i} p(o \mid i)p(i)$$

Application	Input	Output	P(i)	P(o|i)
Speech recognition	Word sequences	Speech signal	Language model	Acoustic model
POS Tagging	POS sequences	Word sequences	Probability of POS sequences	P(word|tag)
OCR	Actual text	Text with mistakes	Langauge model	Model of OCR errors
Machine translation	Target language text	Source language text	Target language model	Translation model

Entropy and Language Models

- So far, we have used entropy to find the most efficient code for transmitting messages.
- Recall Simplified Polynesian:

6 Letters	bits	assumes
Naive	3	uniform distribution
Unigram	2.5	letter frequencies
Syllable	1.22	syllable bigram probabilities
TRUE	?	full conditional likelihood

- The more structure and regularities a model captures, the lower our uncertainty, or entropy, will be.
- We can use entropy as a measure of the quality of our models

Relative Entropy

- For two PMFs, p(x) and q(x), for an event space X, we can compute relative entropy as follows:

$$D(p \,\|\, q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)} \qquad = E_p\left(\log_2 \frac{p(x)}{q(x)}\right)$$

  - Also known as: Kullback-Leibler(KL) divergence

- KL-divergence compares the entropy of the two distributions

- KL-divergence between p and q is the average number of bits that are wasted by encoding events from a distribution *p* with a code based on distribution *q*.

  - Non-symmetric. Can you think why?

Example of KL-divergence

Example

For a random variable $X = \{0, 1\}$ assume two distributions $f(x)$ and $g(x)$ with $f(0) = 1 - r$, $f(1) = r$ and $g(0) = 1 - s$, $g(1) = s$:

$$\begin{aligned} D(f||g) &= (1-r)\log\tfrac{1-r}{1-s} + r\log\tfrac{r}{s} \\ D(g||f) &= (1-s)\log\tfrac{1-s}{1-r} + s\log\tfrac{s}{r} \end{aligned}$$

If $r = s$ then $D(f||g) = D(g||f) = 0$. If $r = \frac{1}{2}$ and $s = \frac{1}{4}$:

$$\begin{aligned} D(f||g) &= \tfrac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} + \tfrac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} = 0.2075 \\ D(g||f) &= \tfrac{3}{4}\log\frac{\frac{3}{4}}{\frac{1}{2}} + \tfrac{1}{4}\log\frac{\frac{1}{4}}{\frac{1}{2}} = 0.1887 \end{aligned}$$

Relative Entropy

Theorem: Properties of the Kullback-Leibler Divergence

1. $D(f||g) \geq 0$;
2. $D(f||g) = 0$ iff $f(x) = g(x)$ for all $x \in X$;
3. $D(f||g) \neq D(g||f)$;
4. $I(X;Y) = D(f(x,y)||f(x)f(y))$.

- Recall Mutual Information measures the distance of a joint distribution from independence, thus Mutual Information and Relative Entropy are related in the following way:

$$I(X;Y) = D(p(x,y) \,||\, p(x)p(y))$$

$$= \sum_{x,y \in X,Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Evaluating language models

- Often, we want to construct a probabilistic model of some linguistic phenomena.
  - Represent events (e.g. letters, words, or sentences that 'occur') by *X*
  - Assume some true probability distribution for X: *p(x)*
  - In building a model, *m*, of *p*, we want to minimise *D(p||m)*
- Cross entropy: *H(X,m) = H(X) + D(p||m)*

Cross Entropy

$$
\begin{aligned}
H(X,q) &= H(X) + D(p \,\|\, m) \\
&= -\sum_x p(x)\log_2 p(x) + \sum_x p(x)\log_2 \frac{p(x)}{m(x)} \\
&= \sum_x p(x)\log_2 \frac{p(x)}{m(x)} - p(x)\log_2 p(x) \\
&= \sum_x p(x)\log_2 p(x) + p(x)\log_2 \frac{1}{m(x)} - p(x)\log_2 p(x) \\
&= \sum_x p(x)\log_2 \frac{1}{m(x)}
\end{aligned}
$$

Cross-Entropy

- It's critical to evaluate cross-entropy on a corpus that is different from the training corpus. Why?
- The lower the cross-entropy, the better the model is a predicting (unseen) events in the language
- This implies that the model better encodes the probabilistic structure of the language
- But sparse data often means simple models outperform richer ones, in practice.

Road Map

- Next Wed: Tutorial on language models. Watch course web page for details!!
- Next: Do natural lexica represent a good code?
  - What is the role of ambiguity in the lexicon?
  - How might channel capacity and the noisy channel be applied to language?
- Then: How does the notion of information relate to on-line comprehension
- And … do speakers take comprehenders into account in the choices about linguistic encoding

"I blame entropy."