

Homework

- For Wednesday 3/9:
- Get bigram and unigram frequencies from ANC website:
<http://americannationalcorpus.org/frequency.html>
- Implement two of the collocation-finding methods described in chapter 5
- Find the top n bigrams by each measure
- Write it up – a paragraph or two describing what's going on
- Midterm next week

Noisy channel model

- Information theory can be used for decoding messages using the *noisy channel* model
- We imagine communication along a low-quality telephone line:
 - sender creates a message S
 - message is transmitted over a communication channel which produces random changes to the message
 - receiver gets a message R
- The challenge is to guess what the original message was using knowledge of source and noise distributions

Information theory

- We can use conditional entropy $H(X|Y)$ to measure the amount of information conveyed by the communication channel
- $H(S|R)$ is a measure of how surprised you would be to find out that the message being sent is S , given that you received R
- For an ideal telegraph, $H(S|R) = 0$
- For an ideal cypher, $H(S|R) = H(S)$
- For most applications, $H(S|R)$ is somewhere in between

Noisy channel model

- Noisy channel model developed by Shannon to describe optimal error correcting codes
- For stochastic NLP, we imagine the observed text is the output of a noisy channel.
- The challenge is to find a *decoder* the average surprise in finding out what the underlying message was, given the observed text
- Formally, this comes down to minimizing $H(S|R)$ or maximizing $I(S; R)$

Noisy channel model

- First used for translation by IBM's T.J. Watson research lab in 1970's
- Many different problems can be posed as a noisy channel model
- Noisy channel model is applicable whenever we want to guess an unknown thing (e.g., a correctly spelled word) given a known thing (e.g, a correctly spelled word)
- We need a model of underlying message probabilities $P(S)$, plus a noise model $P(R|S)$

Noisy channel model

- Spelling correction
 - S = perfect text, R = text with errors
 - $P(S)$ = prob. of perfect text, $P(R|S)$ = error model
- Translation
 - S = Target language, R = Source language
 - $P(S)$ = prob. of target language, $P(R|S)$ = translation model
- Speech recognition
 - S = word sequence, R = speech signal
 - $P(S)$ = prob. of word sequence, $P(R|S)$ = acoustic model

Noisy channel model

- Translation:

Ik zit op de bank en kijk naar het televisie.

- We imagine that this sentence was actually produced in English and deformed by a noisy communication channel.
- To reconstruct the (hypothetical) English original, we need a model of the source probabilities $P(\text{English})$ and the error probabilities $P(\text{Dutch}|\text{English})$.

Noisy channel model

- We work backwards from the error probabilities $P(\text{Dutch}|\text{English})$ to get two possible English sources:

I'm at the bank watching television.

I'm sitting on the sofa watching television.

- One of these is much more likely as an English sentence than the other.

Noisy channel model

- Other alternatives have high $P(\text{English})$ and low $P(\text{Dutch}|\text{English})$:

Have a nice day.

- Or, low $P(\text{English})$ and low $P(\text{Dutch}|\text{English})$:

My hovercraft is full of eels.

- But given high enough $P(\text{Dutch}|\text{English})$, even a very low $P(\text{English})$ sentence might be chosen:

Mijn hovercraft zit vol paling.

Independence

- All statistical methods we looked at make the “i.i.d.” assumption
- This is clearly false for language, though we can sometimes fake it (bag of words model)
- Noisy channel models require a *language model* which assigns a probability to a text:

$$P(w_1, \dots, w_n) = \prod_{i=1}^n P(w_i)$$

- The bag of words assumption makes estimation of the model easy, but is also very unrealistic:

Also assumption bag but easy estimation is makes model of of, the the unrealistic very words.

Independence

- Instead, we can apply the chain rule:

$$\begin{aligned} P(w_1, \dots, w_n) &= P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times \dots \times P(w_n|w_1, \dots, w_{n-1}) \\ &= \prod_{i=1}^n P(w_i|w_1, \dots, w_{i-1}) \end{aligned}$$

- The fully independent model is a multinomial distribution with one parameter per word (conservatively, 20,000 parameters)
- The fully dependent model is a multinomial distribution with one parameter per word *per context* (something like 10^{220} parameters)
- By comparison, there are maybe 10^{79} atoms in the universe.