CSCI 5832
Natural Language Processing

Lecture 8

Jim Martin

Today 2/8

- **Review N-Grams**
- **Entropy/Models**
- **Parts of Speech and Tagging**

N-Gram Models

- **Assigning probabilities to sequences by**
  - **Using the chain rule to decompose the problem**
  - **Make some conditional independence assumptions to simplify things**
  - **Use smoothing and backoff to massage the counts into something that works.**

Back to Fish

	unseen (bass or catfish)	trout
c	0	1
MLE p	$p = \frac{0}{18} = 0$	$\frac{1}{18}$
c^*	$c^*(\text{unseen}) = 1 \times \frac{3}{1} = 3$	$c^*(\text{trout}) = 2 \times \frac{1}{3} = .67$
GT p^*_{GT}	$p^*_{\text{GT}}(\text{unseen}) = \frac{3}{18} = .17$	$p^*_{\text{GT}}(\text{trout}) = \frac{.67}{18} = \frac{1}{27} = .037$

That 3/18 is the main thing to remember about Good-Turing.

It's the probability mass we're reserving for the zero counts.

Good-Turing

- **But the basic Good-Turing approach is pretty broken when it comes to…**
  - **The other bigger buckets**
  - **And how to redistribute the mass among the zero counts**

Katz-Backoff:Trigram Case

$$P_{\text{katz}}(w_i|w_{i-2}w_{i-1}) = \begin{cases} P^*(w_i|w_{i-2}w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha(w_{i-1}w_i)P^*(w_i|w_{i-1}), & \text{else if } C(w_{i-1}w_i) > 0 \\ \alpha(w_i)P^*(w_i), & \text{otherwise.} \end{cases}$$

What Makes a Good Model?

- **Two answers:**
  - **Models that make your end application run better**
    - **In vivo evaluation**
  - **Models that predict well the nature of unseen representative texts…**

Information Theory

- **Who is going to win the World Series next year?**
- **Well there are 30 teams. Each has a chance, so there's a 1/30 chance for any team…? No.**
  - **Rockies? Big surprise, lots of information**
  - **Yankees? No surprise, not much information**

Information Theory

- How much uncertainty is there when you don't know the outcome of some event (answer to some question)?
- How much information is to be gained by knowing the outcome of some event (answer to some question)?

Information Theory

- This stuff is usually explained either in terms of betting or in terms of communication codes.
  - Number of bits needed to communicate messages on average
- Neither of which is terribly illuminating for language applications.

Aside on logs

- **Base doesn't matter. Unless I say otherwise, I mean base 2.**
- **Probabilities lie between 0 an 1. So log probabilities are negative and range from 0 (log 1) to –infinity (log 0).**
- **The – is a pain so at some point we'll make it go away by multiplying by -1.**

Entropy

- **Let's start with a simple case, the probability of word sequences with a unigram model**
- **Example**
  - **S = "One fish two fish red fish blue fish"**
  - **P(S) = P(One)P(fish)P(two)P(fish)P(red)P(fish)P(blue)P(fish)**
  - **Log P(S) = Log P(One)+Log P(fish)+…Log P(fish)**

Entropy cont.

- In general that's [redacted]
- But note that
  - the order doesn't matter
  - that words can occur multiple times
  - and that they always contribute the same each time
  - so rearranging… [redacted]

Entropy cont.

- One fish two fish red fish blue fish
- Fish fish fish fish one two red blue

[redacted]

Entropy cont.

- Now let's divide both sides by N, the length of the sequence:

[illegible]

- That's basically a per word average of the log probabilities

Entropy

- Now assume the sequence is really really long.
- Moving the N into the summation you get

[illegible]

- Rewriting and getting rid of the minus sign

[illegible]

Entropy

- **Think about this in terms of uncertainty or surprise.**
  - **The more likely a sequence is, the lower the entropy. Why?**

[illegible]

Entropy

- **Note that that sum is over the types of the elements of the model being used (unigrams, bigrams, trigrams, etc.), not the words in the sequence.**

Model Evaluation

- **Remember the name of the game is to come up with statistical models that capture something useful in some body of text or speech.**
- **There are precisely a *gazzilion* ways to do this**
  - **N-grams of various sizes**
  - **Smoothing**
  - **Backoff…**

Model Evaluation

- ***Given a collection of text and a couple of models, how can we tell which model is best?***
- **Intuition… the model that assigns the highest probability (lowest entropy) to a set of withheld text**
  - **Withheld text? Text drawn from the same distribution (corpus), but not used in the creation of the model being evaluated.**

Model Evaluation

- **The more you're surprised at some event that actually happens, the worse your model was.**
- **We want models that minimize your surprise at observed outcomes.**
- **Given two models and some training data and some withheld test data… which is better?**
  - **The model where you're not surprised to see the test data.**

Break

- **Quiz is Thursday.**
- **Next HW details coming soon.**
- **Shifting to Chapter 5**

Parts of Speech

- **Start with eight basic categories**
  - **Noun, verb, pronoun, preposition, adjective, adverb, article, conjunction**
- **These categories are based on morphological and distributional properties (not semantics)**
- **Some cases are easy, others are murky**

Parts of Speech

- **What are some possible parts of speech for *building*?**

Parts of Speech

- A quarantine in the Boca Raton building contaminated by deadly anthrax is set to be lifted.
- Dialogue is one of the powerful tools to building an understanding across differences and thereby leading to negotiation. It is an easy way to recognise ...
- The building project, which would be spread out over five years with schools most in need getting work first, would cost taxpayers with a ...
- Building for Independence, as its name indicates, demonstrates exactly what Canada's New Government is doing to support Canadians who are homeless or at ...
- The last time house building reached such high levels was in 1989 when over 191800 new homes were built.
- State lawmakers are considering building up a trust fund for schools so it will earn more money in the coming decades. ...

Tagging

- State/NN lawmakers/NNS are/VBP considering/VBG building/VBG up/RP a/DT trust/NN fund/NN for/IN schools/NNS so/IN it/PRP will/MD earn/VB more/JJR money/NN in/IN the/DT coming/VBG decades/NNS ./.

Parts of Speech

- **Two kinds of category**
  - **Closed class**
    - **Prepositions, articles, conjunctions, pronouns**
  - **Open class**
    - **Nouns, verbs, adjectives, adverbs**

Sets of Parts of Speech: Tagsets

- **There are various standard tagsets to choose from; some have a lot more tags than others**
- **The choice of tagset is based on the application**
- **Accurate tagging can be done with even large tagsets**

Penn Tagset

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	' or "
POS	Possessive ending	*'s*	"	Right quote	' or "
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	[, (, {, <
PRP$	Possessive pronoun	*your, one's*	)	Right parenthesis	],), }, >
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	. ! ?
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	: ; ... – -
RP	Particle	*up, off*			

Figure 5.6 Penn Treebank part-of-speech tags (including punctuation).

Tagging

- **Part of speech tagging is the process of assigning parts of speech to each word in a sentence… Assume we have**
  - **A tagset**
  - **A dictionary that gives you the possible set of tags for each entry**
  - **A text to be tagged**
  - **A reason?**

Three Methods

- **Rules**
- **Probabilities**
- **Sort of both**

Rules

- **Hand-crafted rules for ambiguous words that test the context to make appropriate choices**
  - **Early attempts fairly error-prone**
  - **Extremely labor-intensive**

Probabilities

- We want the best set of tags for a sequence of words (a sentence)
- W is a sequence of words
- T is a sequence of tags

[redacted]

Probabilities

- We want the best set of tags for a sequence of words (a sentence)
- W is a sequence of words
- T is a sequence of tags

[redacted]

Tag Sequence: P(T)

- **How do we get the probability of a specific tag sequence?**
  - **Count the number of times a sequence occurs and divide by the number of sequences of that length. Not likely.**
  - **Make a Markov assumption and use N-grams over tags...**
    - **P(T) is a product of the probability of N-grams that make it up.**

P(T): Bigram Example

- **<s> Det Adj Adj Noun </s>**

- **P(Det|<s>)P(Adj|Det)P(Adj|Adj)P(Noun|Adj)**

Counts

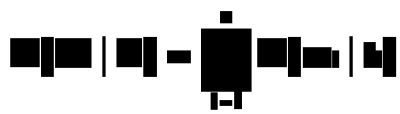
- **Where do you get the N-gram counts?**
- **From a large hand-tagged corpus.**
  - **For N-grams, count all the Tag_i Tag_{i+1} pairs**
  - **And smooth them to get rid of the zeroes**
- **Alternatively, you can learn them from an untagged corpus**

What about P(W|T)

- **First its odd. It is asking the probability of seeing "The big red dog" given "Det Adj Adj Noun"**
  - **Collect up all the times you see that tag sequence and see how often "The big red dog" shows up. Again not likely to work.**

P(W|T)

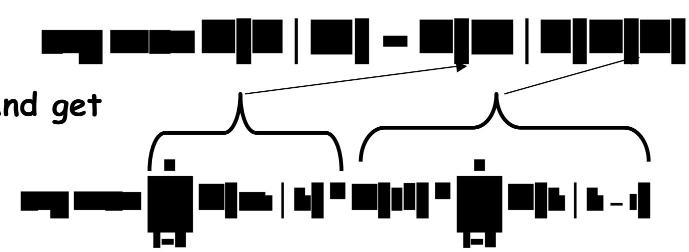
- We'll make the following assumption (because it's easy)… Each word in the sequence only depends on its corresponding tag. So…



- How do you get the statistics for that?

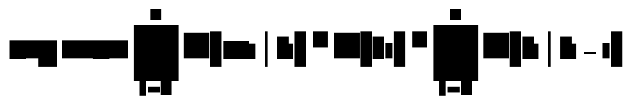
So…

- We start with
- And get



HMMs

- This is an HMM



- The states in the model are the tags, and the observations are the words.
  - The state to state transitions are driven by the bigram statistics
  - The observed words are based solely on the state that you're in

Performance

- This method has achieved 95-96% correct with reasonably complex English tagsets and reasonable amounts of hand-tagged training data.
- Forward pointer… its also possible to train a system without hand-labeled training data