

CS474 Natural Language Processing

Next...Language Modeling

- Introduction to generative models of language
 - » What are they?
 - » Why they're important
 - » Issues for counting words
 - » Statistics of natural language
 - » Unsmoothed n-gram models

IBM's Watson

<http://www-03.ibm.com/innovation/us/watson/what-is-watson/why-jeopardy.html>

What are *generative* models of language?

- Word prediction
 - *Once upon a...*
 - *I'd like to make a collect...*
 - *Let's go outside and take a...*
- Generative models can assign probabilities to strings of words

Why are word prediction models important?

- Augmentative communication systems
 - For the disabled, to predict the next words the user wants to “speak”
- Computer-aided education
 - System that helps kids learn to read (e.g. Mostow et al. system)
- Speech recognition
- Context-sensitive spelling correction
- ...

Why are word prediction models important?

- Can be used to assign a probability to the next word in an incomplete sentence
- Closely related to the problem of computing the probability of a sequence of words
 - Useful for part-of-speech tagging, probabilistic parsing, ...

The need for models of word prediction in NLP has not been uncontroversial

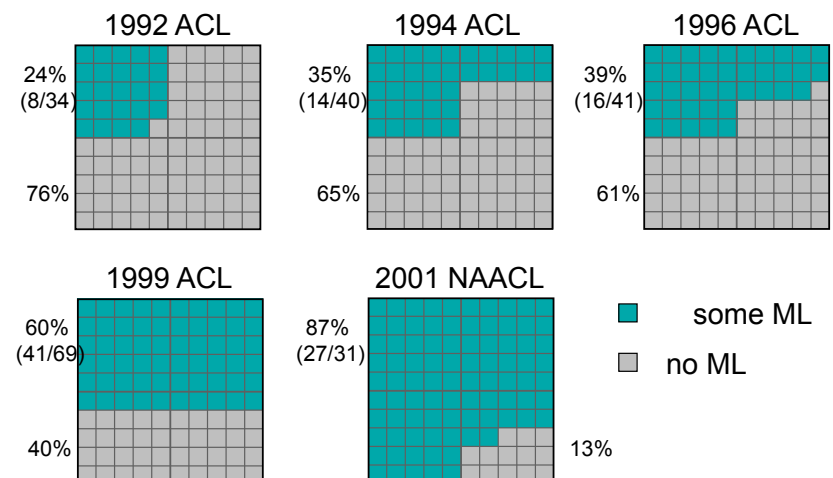
But it must be recognized that the notion “probability of a sentence” is an entirely useless one, under any known interpretation of this term. -Noam Chomsky (1969)

Every time I fire a linguist the recognition rate improves.
- Fred Jelinek (IBM speech group, 1988)

Paradigms in NLP

- Knowledge-based methods
 - Rely on the manual encoding of linguistic (and world) knowledge
 - » E.g. FSA's for morphological parsing, syntactic parsing
- Statistical/learning methods
 - Rely on the automatic acquisition of linguistic knowledge from corpora

Statistical/machine learning in NLP



Real life situations...

Word prediction gone awry



Woody Allen's "Take the Money and Run"

http://www.youtube.com/watch_popup?v=-UHOgkDbVqc&vq=medium#t=14

Word prediction gone amok

Seinfeld Sentence Finisher

- <http://www.youtube.com/watch?v=01teZKTYjQA&feature=related>


N-gram model

- Uses the previous N-1 words to predict the next word
 - 2-gram: bigram
 - 3-gram: trigram
 - 1-gram: unigram
- In speech recognition, these statistical models of word sequences are referred to as a **language model**

Goals

- Determine the next word in a sequence
 - Probability distribution across all words in the language
 - $P(w_n | w_1 w_2 \dots w_{n-1})$
- Determine the probability of a sequence of words
 - $P(w_1 w_2 \dots w_{n-1} w_n)$

Next...Language Modeling

- Introduction to generative models of language
 - » What are they?
 - » Why they're important
 -  » Issues for counting words
 - » Statistics of natural language
 - » Unsmoothed n-gram models

Counting words in corpora

- Ok, so how many words are in this sentence?
- Depends on whether or not we treat punctuation marks as words
 - Important for many NLP tasks
 - » Grammar-checking, spelling error detection, author identification, part-of-speech tagging
- Spoken language corpora
 - Utterances don't usually have punctuation, but they do have other phenomena that we might or might not want to treat as words
 - » I do uh main- mainly business data processing
 - Fragments
 - Filled pauses
 - » *um* and *uh* behave more like words, so most speech recognition systems treat them as such

Counting words in corpora

- Capitalization
 - Should *They* and *they* be treated as the same word?
 - » For most statistical NLP applications, they are
 - » Sometimes capitalization information is maintained as a feature
 - ◆ E.g. spelling error correction, part-of-speech tagging
- Inflected forms
 - Should *walks* and *walk* be treated as the same word?
 - » No...for most n-gram based systems
 - » based on the **wordform** (i.e. the inflected form as it appears in the corpus) rather than the **lemma** (i.e. set of lexical forms that have the same stem)

Counting words in corpora

- Need to distinguish
 - word types
 - » the number of distinct words
 - word tokens
 - » the number of running words
- Example
 - *All for one and one for all.*
 - 8 tokens (counting punctuation)
 - 6 types (assuming capitalized and uncapitalized versions of the same token are treated separately)