

Sequence Models

Introduction to
Artificial Intelligence
COS302
Michael L. Littman
Fall 2001

Administration

Exams enjoyed Toronto.
Letter grades for programs:
A: 74-100 (31)
B: 30-60 (20)
C: 10-15 (4)
?: (7)
(0 did not imply "incorrect")

Shannon Game

Sue swallowed the large green __.
pepper frog
pea pill
Not:
idea beige
running very

"AI Complete" Problem

My mom told me that playing
Monopoly® with toddlers was a
bad idea, but I thought it would
be ok. I was wrong. Billy
chewed on the "Get Out of Jail
Free Card". Todd ran away with
the little metal dog. Sue
swallowed the large green __.

Language Modeling

If we had a way of assigning
probabilities to sentences, we
could solve this. How?
Pr(Sue swallowed the large green cat.)
Pr(Sue swallowed the large green odd.)
How could such a thing be learned
from data?

Why Play This Game?

Being able to assign likelihood to
sentences a useful way of
processing language.
Speech recognition
Criterion for comparing language
models
Techniques useful for other
problems

Statistical Estimation

To use statistical estimation:

- Divide data into equivalence classes
- Estimate parameters for the different classes

Conflicting Interests

Reliability

- Lots of data in each class
- So, small number of classes

Discrimination

- All relevant distinctions made
- So, large number of classes

End Points

Unigram model:

$$\Pr(w \mid \text{Sue swallowed the large green } ___\text{.}) = \Pr(w)$$

Exact match model:

$$\Pr(w \mid \text{Sue swallowed the large green } ___\text{.}) = \Pr(w \mid \text{Sue swallowed the large green } ___\text{.})$$

What word would these suggest?

N-grams: Compromise

N-grams are simple, powerful.

Bigram model:

$$\Pr(w \mid \text{Sue swallowed the large green } ___\text{.}) = \Pr(w \mid \text{green } ___\text{.})$$

Trigram model:

$$\Pr(w \mid \text{Sue swallowed the large green } ___\text{.}) = \Pr(w \mid \text{large green } ___\text{.})$$

Not perfect: misses "swallowed".

pillow crystal caterpillar
iguana Santa tigers

Aside: Syntax

Can do better with a little bit of knowledge about grammar:

$$\Pr(w \mid \text{Sue swallowed the large green } ___\text{.}) = \Pr(w \mid \text{modified by swallowed, the, green } ___\text{.})$$

pill	dye
one	pineapple
dragon	beans
speck	liquid
solution	drink

Estimating Trigrams

Treat sentences independently.

Ok?

$$\Pr(w_1 w_2)$$

$$\Pr(w_j \mid w_{j-1} w_{j-2})$$

$$\Pr(\text{EOS} \mid w_{j-1} w_{j-2})$$

Simple so far.

Sparsity

Pr(w| comes across)

as 8/10 (in Austen's works)

a 1/10

more 1/10

the 0/10

Don't estimate as zeros!

Can use Laplace smoothing, e.g., or
back off to bigram, unigram.

Unreliable Words

Can't take much stock in words
only seen once (*hapax
legomena*). Change to "UNK".

Generally a small fraction of the
tokens and half the types.

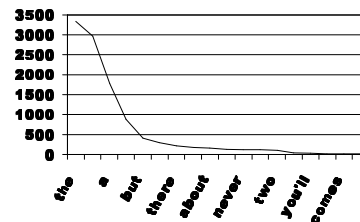
The boy saw the dog.

5 tokens, 4 types.

Zipf's Law

Frequency is proportional to rank.
Thus, extremely long tail!

Word Frequencies in Tom Sawyer



Using Trigrams

Hand me the ___ knife now .

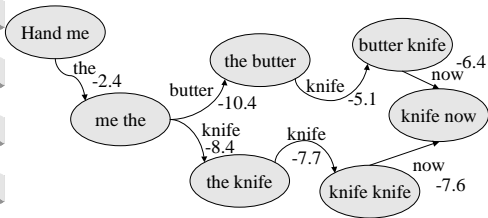
butter

knife

Counts

me the	2832670
me the butter	88
me the knife	638
the knife	154771
the knife knife	72
the butter	92304
the butter knife	559
knife knife	7831
knife knife now	4
butter knife	9046
butter knife now	15

Markov Model



General Scheme

$$\Pr(w_j = x \mid w_1 w_2 \dots \text{EOS})$$

$$= \Pr(w_1 w_2 \dots x \dots \text{EOS}) / \sum_x \Pr(w_1 w_2 \dots x \dots \text{EOS})$$

Maximized by $\Pr(w_1 w_2 \dots x \dots \text{EOS})$

$$= \Pr(w_1 w_2) \dots \Pr(x \mid w_{j-1} w_{j-2}) \Pr(w_{j+1} \mid w_{j-1} x) \Pr(w_{j+2} \mid x w_{j+1}) \dots \Pr(\text{EOS} \mid w_{n-1} w_n)$$

Maximized by $\Pr(x \mid w_{j-1} w_{j-2}) \Pr(w_{j+1} \mid w_{j-1} x) \Pr(w_{j+2} \mid x w_{j+1})$

Mutual Information

$\log(\Pr(x \text{ and } y) / \Pr(x) \Pr(y))$
 Measures the degree to which two events are independent (how much "information" we learn about one from knowing the other).

Mutual Inf. Application


Measure of strength of association between words
 levied: imposed vs. believed
 Reduces to simply
 $\Pr(\text{levied} \mid x) = \Pr(\text{levied}, x) / \Pr(x)$
 $= \text{count}(\text{levied and } x) / \text{count}(x)$
 "imposed" has higher score.

Analogy Idea

Find a linking word such that a mutual information score is maximized.
 Tricky to find the right word.
 Unclear if any word will have the right effect.
 traffic flows through the street
 water flows through the riverbed

What to Learn

Reliability/discrimination tradeoff.
 Definition of N-gram models
 How to find most likely word in an N-gram model
 Mutual Information



Homework 7 (due 11/21)

- 1. Give a maximization scheme for filling in the two blanks in a sentence like “I hate it when ___ goes ___ on me.” Be somewhat rigorous to make the TA's job easier.**
- 2. more soon**