CS 2731 / ISSP 2230
Introduction to Natural Language Processing

Session 10: N-gram language models, part 1

Michael Miller Yoder

February 12, 2024

University of Pittsburgh School of Computing and Information

Course logistics

- Homework 2 is due **this Thu Feb 15**
  - Text classification
  - Written and programming components
  - Optional Kaggle competition for best LR and NN politeness classifiers
  - Ask questions and offer answers on the Canvas discussion forum
- Homework 1 grades should be out today
- Projects
  - Proposal and literature review is **due Thu Feb 22**
    - Instructions are on the project webpage
  - It's good to start the literature review early
  - Look for NLP papers in ACL Anthology, Semantic Scholar, and Google Scholar

Lecture overview: N-gram language models, part 1

- Language modeling
- N-gram language models
- Estimating n-gram probabilities
- Perplexity and evaluating language models

Core tasks and applications of NLP

CORE TASKS

- text classification
- representation learning

MODULE 2

- language models
- conditional language models

MODULE 3

- sequence labeling
- syntactic parsing

APPLICATIONS

- machine translation
- speech recognition & synthesis
- chatbots
- computational social science

Introduction to language models

Language Models Estimate the Probability of Sequences

Which of these sentences would you be more likely to observe in an English corpus?

- Hugged I big brother my.
- I hugged my large brother.
- I hugged my big brother.

Slide credit: David Mortensen

Language Models Estimate the Probability of Sequences

Which of following word would be most likely to come after “David hates visiting New...”

- York
- California
- giggled

Slide credit: David Mortensen

These are actually instances of the same problem: the language modeling problem!

Language Modeling is Tremendously Useful

LMs (language models) are at the center of NLP today and have many different applications

- Machine Translation
  P(high winds tonight) > P(large winds tonight)
- Spelling Correction
  P(about fifteen **minutes** from) > P(about fifteen **minuets** from)
- Text Input Methods
  P(i cant believe how hot you **are**) > P(i cant believe how hot you **art**)
- Speech Recognition
  P(recognize speech) > P(wreck a nice beach)

Slide credit: David Mortensen

The Goal of Language Modeling

Compute the probability of a sequence of words/tokens/characters:

$$P(\mathbf{w}) = P(w_1, w_2, w_3, w_5, \ldots, w_n)$$

$$P(\text{I, hugged, my, big, brother})$$

This is related to next-word prediction:

$$P(w_t | w_1 w_2 \ldots w_{t-1})$$

$$P(\text{York}|\text{David, hates, going, to, New})$$

Do you compute either of these? Then you're in luck:

You are a language model!

Slide credit: David Mortensen

N-gram language models

The Chain Rule Helps Us Compute Joint Probabilities

The definition of conditional probability is

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

which can be rewritten as

$$P(A,B) = P(A)P(B|A)$$

Slide credit: David Mortensen

The Chain Rule Helps Us Compute Joint Probabilities

If we add more variables, we see the following pattern:

$$
\begin{aligned}
P(A,B,C) &= P(A)P(B|A)P(C|A,B) \\
P(A,B,C,D) &= P(A)P(B|A)P(C|A,B)P(D|A,B,C)
\end{aligned}
$$

which can be generalized as

$$P(x_1, x_2, x_3, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \ldots P(x_n|x_1, \ldots, x_{n-1})$$

The Chain Rule!

Slide credit: David Mortensen

The chain rule to compute the joint probability of words in a sentence

$$P(w_1, w_2, w_3, \ldots, w_n) = \prod_i^n P(w_i | w_1 w_2 \ldots w_{i-1})$$

$$\begin{aligned}
&P(\text{now is the winter of our discontent}) = \\
&P(\text{now}) \times P(\text{is}|\text{now}) \times \\
&P(\text{the}|\text{now is}) \times P(\text{winter}|\text{now is the}) \times \\
&P(\text{of}|\text{now is the winter}) \times \\
&P(\text{our}|\text{now is the winter of}) \times \\
&P(\text{discontent}|\text{now is the winter of our})
\end{aligned}$$

Slide credit: David Mortensen

How Are We Estimating these Probabilities?

Could we just count and divide?

$$P(\text{discontent}|\text{now is the winter of our}) = \frac{\text{Count(now is the winter of our discontent)}}{\text{Count(now is the winter of our)}}$$

But this can't be a valid estimate! How many times in a corpus are either "now is the winter of our" or "now is the winter of our discontent" going to occur? This cannot be an estimate of their true probability.

Slide credit: David Mortensen

This May not Seem Very Helpful

Is $P(\text{discontent}|\text{now is the winter of our})$ really easier to compute than $P(\text{now is the winter of our discontent})$?

How is the chain rule helping us? A peak back at Naïve Bayes may provide a hint: **cheat**.

Slide credit: David Mortensen

Enter a Hero: Andrei Markov

Born	20 December 1978 (age 43) Voskresensk, Russian SFSR, Soviet Union
Height	6 ft 0 in (183 cm)
Weight	203 lb (92 kg; 14 st 7 lb)
Position	Defence
Played for	Khimik Voskresensk Dynamo Moscow Montreal Canadiens Vityaz Chekhov Ak Bars Kazan Lokomotiv Yaroslavl
Playing career	1995–2020

Slide credit: David Mortensen

Or, Rather, Andrey Markov

Born	14 June 1856 N.S. Ryazan, Russian Empire
Died	20 July 1922 (aged 66) Petrograd, Russian SFSR
Known for	Markov chains; Markov processes; stochastic processes
Fields	Mathematics, specifically probability theory and statistics
Doctoral advisor	Pafnuty Chebyshev

Slide credit: David Mortensen

Markov Did a Computational Linguistics

Interestingly, Markov's first application of his idea of **Markov Chains** was to language, specifically to modeling alliteration and rhyme in Russian poetry.

As such, he can be seen not only as a great mathematician and statistician, but also one of the forerunners of **computational linguistics** and **computational humanities**.

Slide credit: David Mortensen

Markov Showed that You Could Make a Simplifying Assumption

One can approximate

$$P(\text{discontent}|\text{now is the winter of our})$$

by computing

$$P(\text{discontent}|\text{our})$$

or perhaps

$$P(\text{discontent}|\text{of our})$$

- We only get an estimate this way, but we can obtain it by only counting simpler things: “our discontent”, “discontent”, “of our”, etc
- Ngram language modeling is a generalization of this observation

Slide credit: David Mortensen

This assumption is the Markov assumption

$$P(w_1, w_2, \ldots, w_n) \approx \prod_i P(w_i | w_{i-k} w_{i-1})$$

In other words, we approximate each component in the product:

$$P(w_i | w_1, w_2, \ldots, w_{i-1}) \approx P(w_i | w_{i-k} \ldots w_{i-1})$$

We will now walk through what this looks like for different values of k.

The Unigram Model ($k = 1$)

$$P(w_1 w_2 \ldots w_i) \approx \prod_i P(w_i)$$

The probability of a sequence is approximately the product of the probabilities of the individual words.

Some automatically generated sequences from a unigram model:

- fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, dollars, quarter, in, is, mass
- thrift, did, eighty, said, hard, 'm, july, bullish
- that, or, limited, the

What do you notice about them?

The Bigram Model ($k = 2$)

If you condition on the previous word, you get the following:

$$P(w_i | w_1 w_2 \ldots w_{i-1}) \approx P(w_i | w_{i-1})$$

Some examples generated by a bigram model:

- texaco, rose, one, in, this, issue, is, pursuing, growth, in, a, boiler, house, said, mr., gurria, mexico, 's, motion, control, proposal, without, permission, from, five, hundred, fifty, five, yen
- outside, new, car, parking, lot, of, the, agreement, reached
- this, would, be, a, record, november

Are these better?

Slide credit: David Mortensen

The Trigram Model

The trigram model is just like the bigram model, only with a larger *k*:

$$P(w_i | w_1 w_2 \dots w_{i-1}) \approx P(w_i | w_{i-2} w_{i-1})$$

The output of a trigram language model is generally **much** better than that of a bigram model **provided the training corpus is large enough**. Why do you need a larger corpus to train a trigram corpus than a bigram or unigram corpus?

N-gram models have trouble with long-range dependencies

In general, n-gram models are very impoverished models of language. For example, language has relationships that span many words:

- The **students** who worked on the assignment for three hours straight ***is/are** finally resting.
- The **teacher** who might have suddenly and abruptly met students **is/*are** tall.
- Violins are easy to mistakenly think you can learn to play ***them/quickly**.

Negative polarity: predict "some" vs "any"

- *I want **any**.
- I want **some**.
- I **don't** want **any**.
- *I think you said he thought we told them that she wants **any**.
- I think you said he thought we told them that she wants **some**.

Slide credit: David Mortensen

Ngram LMs Are Often Adequate

Nevertheless, for many applications, ngram models are good enough (and they're super fast and efficient)

Estimating n-gram probabilities

Estimating bigram probabilities with the maximum likelihood estimate (MLE)

MLE for bigram probabilities can be computed as:

$$P(w_i|w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$$

which we will sometimes represent as

$$P(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

Slide credit: David Mortensen

An example

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

$P(\texttt{I} \mid \texttt{<s>}) =$ $\quad$ $P(\texttt{Sam} \mid \texttt{<s>}) =$ $\quad$ $P(\texttt{am} \mid \texttt{I}) =$

$P(\texttt{</s>} \mid \texttt{Sam}) =$ $\quad$ $P(\texttt{Sam} \mid \texttt{am}) =$ $\quad$ $P(\texttt{do} \mid \texttt{I}) =$

Slide adapted from Jurafsky & Martin

More examples: Berkeley Restaurant Project sentences

can you tell me about any good cantonese restaurants close by

mid priced thai food is what i'm looking for

tell me about chez panisse

can you give me a listing of the kinds of food that are available

i'm looking for a good place to eat breakfast

when is caffe venezia open during the day

Slide adapted from Jurafsky & Martin

Raw bigram counts

Out of 9222 sentences

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Slide adapted from Jurafsky & Martin

Raw bigram probabilities

Normalize by unigrams:

Result:

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Slide adapted from Jurafsky & Martin

Bigram estimates of sentence probabilities

$$
\begin{aligned}
&P(\text{<s> I want english food </s>}) = \\
&P(\text{I}|\text{<s>}) \\
&\quad \times\ P(\text{want}|\text{I}) \\
&\quad \times\ P(\text{english}|\text{want}) \\
&\quad \times\ P(\text{food}|\text{english}) \\
&\quad \times\ P(\text{</s>}|\text{food}) \\
&\qquad = \ .000031
\end{aligned}
$$

Slide adapted from Jurafsky & Martin

Multiplication Considered Harmful

In reality, as was the case with NB classification, we do all of our computation in log space

- **Avoid underflow** Multiplying small probabilities by small probabilities results in *very small* numbers, which is problematic
- **Optimize computation** Addition is cheaper than multiplication

$$\log(p_1 \times p_2 \times p_3 \times p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

Slide credit: David Mortensen

The are high-performance toolkits for n-gram language modeling

- SRILM `http://www.speech.sri.com/projects/srilm/`
- KenLM `https://kheafield.com/code/kenlm/`

Slide credit: David Mortensen

Perplexity and evaluating language models

The Evaluation Process for ML Models

The goal of LM evaluation:

- Does our model prefer good sentences to bad sentences?
- Specifically, does it assign higher probabilities to the good/grammatical/frequently observed ones and lower probabilities to the bad/ungrammatical/seldom observed ones?

In ML evaluation, we divide our data into three sets: `train`, `dev`, and `test`.

- We train the model's parameters on the `train` set
- We tune the model's hyperparameters (if appropriate) on the `dev` set (which should not overlap with the `train` set
- We test the model on the `test` set, which should not overlap with `train` or `dev`

An evaluation metric tells us how well our model has done on `test`.

We Can Evaluate Models Intrinsically or Extrinsically

- **Extrinsic Evaluation** means asking how much the model contributes to a larger task or goal. We may evaluate an LM based on how much it improves machine translation over a BASELINE.
- **Intrinsic Evaluation** means measuring some property of the model directly. We may quantify the probability that an LM assigns to a corpus of text.

In general, EXTRINSIC EVALUATION is better, but more expensive and time-consuming.

Slide credit: David Mortensen

Extrinsic Evaluation of LMs

Best evaluation for comparing models A and B

- Put each model in a task (spelling corrector, speech recognizer, MT system)
- Run the task, get an accuracy for A and for B
  - How many misspelled words corrected properly?
  - How many sentences translated correctly?
- Compare scores for A and B

This takes a lot of time to set up and can be expensive to carry out.

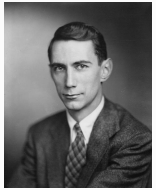
Perplexity is an intrinsic metric for language modeling

Perplexity evaluates the probability assigned by a model to a collection of text and is, thus, useful for evaluating LMs. Note:

- It is a rather crude instrument
- It sometimes correlates only weakly with performance on downstream tasks
- It's only useful for pilot experiments
- But it's cheap and easy to compute, so it's important to understand

Slide credit: David Mortensen

Intuition of Perplexity

The Shannon Game: 

- How well can we predict the next word?

  I always order pizza with cheese and ____

  The 33rd President of the US was ____

  I saw a ____

- Unigrams are terrible at this game. (Why?)

mushrooms 0.1

pepperoni 0.1

anchovies 0.01

....

fried rice 0.0001

....

and 1e-100

- A better model of a text is one which assigns a higher probability to the word that actually occurs

Slide adapted from Jurafsky & Martin

Deriving Perplexity for Bigrams

$$
\begin{aligned}
PP(\mathbf{w}) &= P(w_1 w_2 \ldots w_n)^{-\frac{1}{n}} && \text{Definition} \\
&= \sqrt[n]{\frac{1}{P(w_1 w_2 \ldots w_n)}} \\
&= \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i|w_1 w_2 \ldots w_{i-1})}} && \text{Chain Rule} \\
&= \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i)}} && \text{For Unigrams} \\
&= \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i|w_{i-1})}} && \text{For Bigrams}
\end{aligned}
$$

To minimize perplexity is to maximize probability!

Perplexity as branching factor

- Let’s suppose a sentence consisting of random digits
- What is the perplexity of this sentence according to a model that assign P=1/10 to each digit?

$$
\begin{aligned}
\mathrm{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\
&= (\frac{1}{10}^{N})^{-\frac{1}{N}} \\
&= \frac{1}{10}^{-1} \\
&= 10
\end{aligned}
$$

Slide adapted from Jurafsky & Martin

In general, a lower perplexity implies a better model.

Slide credit: David Mortensen

Lower perplexity = better model

Training 38 million words, test 1.5 million words, WSJ

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Slide adapted from Jurafsky & Martin

Questions?