Count-based Language Modeling

CMSC 473/673
UMBC

Some slides adapted from 3SLP, Jason Eisner
Outline

Defining Language Models

Breaking & Fixing Language Models

Evaluating Language Models
Goal of Language Modeling

\[ p_\theta([...text...]) \]

Learn a probabilistic model of text

Accomplished through observing text and updating model parameters to make text more likely
Goal of Language Modeling

Learn a probabilistic model of text

Accomplished through observing text and updating model parameters to make text more likely

$0 \leq p_{\theta}([...text...]) \leq 1$

$\sum_{t:t \text{ is valid text}} p_{\theta}(t) = 1$
Design Question 1: What Part of Language Do We Estimate?

$p_\theta([...text..])$

Is $[...text..]$ a

- Full document?
- Sequence of sentences?
- Sequence of words?
- Sequence of characters?

A: It’s task-dependent!
Design Question 2: How do we estimate robustly?

$p_\theta([...\text{typo-text}...])$

What if $[...\text{text}...]$ has a typo?
Design Question 3: How do we generalize?

What if \([... \text{text...}]\) has a word (or character or...) we’ve never seen before?
“The Unreasonable Effectiveness of Recurrent Neural Networks”
http://karpathy.github.io/2015/05/21/rnn-effectiveness/
“The Unreasonable Effectiveness of Recurrent Neural Networks”
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

“The Unreasonable Effectiveness of Character-level Language Models”
(and why RNNs are still cool)
http://nbviewer.jupyter.org/gist/yoavg/d76121dfde2618422139
Simple Count-Based

\( p(\text{item}) \)
Simple Count-Based

$p(item) \propto count(item)$

“proportional to”
Simple Count-Based

$p(item) \propto \text{count}(item)$

$$= \frac{\text{count}(item)}{\sum \text{any other item } y \text{ count}(y)}$$
Simple Count-Based

\[ p(\text{item}) \propto \text{count}(\text{item}) \]

\[ = \frac{\text{count}(\text{item})}{\sum_{\text{any other item } y} \text{count}(y)} \]

“proportional to” constant
Simple Count-Based

\[ p(\text{item}) \propto \text{count}(\text{item}) \]

sequence of characters $\rightarrow$ pseudo-words

sequence of words $\rightarrow$ pseudo-phrases
Shakespearian Sequences of Characters

![Graph showing the distribution of word lengths in Shakespearean text. The x-axis represents the length of words in characters, and the y-axis represents the count of occurrences. The graph shows a distribution skewed to the left, with most words being short.]
Shakespearian Sequences of Words

Count

"Sentence" Length (Words)
Novel Words, Novel Sentences

“Colorless green ideas sleep furiously” – Chomsky (1957)

Let’s observe and record all sentences with our big, bad supercomputer

Red ideas? Read ideas?
Probability Chain Rule

\[
p(x_1, x_2, \ldots, x_S) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \ldots p(x_S | x_1, \ldots, x_i) = \prod_{i} p(x_i | x_1, \ldots, x_{i-1})
\]
N-Grams

Maintaining an entire inventory over sentences could be too much to ask

Store “smaller” pieces?

p(Colorless green ideas sleep furiously)
N-Grams

Maintaining an entire *joint* inventory over sentences could be too much to ask

Store “smaller” pieces?

\[ p(\text{Colorless green ideas sleep furiously}) = p(\text{Colorless}) * \]
N-Grams

Maintaining an entire joint inventory over sentences could be too much to ask

Store “smaller” pieces?

\[ p(\text{Colorless green ideas sleep furiously}) = p(\text{Colorless}) \times p(\text{green} \mid \text{Colorless}) \times \]
N-Grams

Maintaining an entire joint inventory over sentences could be too much to ask

Store “smaller” pieces?

\[
p(\text{Colorless green ideas sleep furiously}) = \\
p(\text{Colorless}) * \\
p(\text{green | Colorless}) * \\
p(\text{ideas | Colorless green}) * \\
p(\text{sleep | Colorless green ideas}) * \\
p(\text{furiously | Colorless green ideas sleep})
\]
N-Grams

Maintaining an entire *joint* inventory over sentences could be too much to ask

Store “smaller” pieces?

\[
p(\text{Colorless green ideas sleep furiously}) = p(\text{Colorless}) \cdot p(\text{green} \mid \text{Colorless}) \cdot p(\text{ideas} \mid \text{Colorless green}) \cdot p(\text{sleep} \mid \text{Colorless green ideas}) \cdot p(\text{furiously} \mid \text{Colorless green ideas sleep})
\]
N-Grams

Maintaining an entire *joint* inventory over sentences could be too much to ask

Store “smaller” pieces?

\[
p(\text{Colorless green ideas sleep furiously}) = p(\text{Colorless}) \times p(\text{green | Colorless}) \times p(\text{ideas | Colorless green}) \times p(\text{sleep | Colorless green ideas}) \times p(\text{furiously | Colorless green ideas sleep})
\]
N-Grams

$p(\text{furiously} \mid \text{Colorless green ideas sleep})$

How much does “Colorless” influence the choice of “furiously?”
N-Grams

\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \]

How much does “Colorless” influence the choice of “furiously?”

Remove history and contextual info
N-Grams

\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \]

How much does “Colorless” influence the choice of “furiously?”

Remove history and contextual info

\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \approx p(\text{furiously} \mid \text{Colorless-green-ideas sleep}) \]
N-Grams

\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \]

How much does “Colorless” influence the choice of “furiously?”

Remove history and contextual info

\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \approx p(\text{furiously} \mid \text{ideas sleep}) \]
N-Grams

\[ p(\text{Colorless green ideas sleep furiously}) = \]
\[ p(\text{Colorless}) \times \]
\[ p(\text{green} \mid \text{Colorless}) \times \]
\[ p(\text{ideas} \mid \text{Colorless green}) \times \]
\[ p(\text{sleep} \mid \text{Colorless green ideas}) \times \]
\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \]
N-Grams

\[ p(\text{Colorless green ideas sleep furiously}) = \]
\[ p(\text{Colorless}) \times \]
\[ p(\text{green | Colorless}) \times \]
\[ p(\text{ideas | Colorless green}) \times \]
\[ p(\text{sleep | Colorless green ideas}) \times \]
\[ p(\text{furiously | Colorless green ideas sleep}) \]
Trigrams

\[
p(\text{Colorless green ideas sleep furiously}) = \]
\[
p(\text{Colorless}) * \]
\[
p(\text{green | Colorless}) * \]
\[
p(\text{ideas | Colorless green}) * \]
\[
p(\text{sleep | green ideas}) * \]
\[
p(\text{furiously | ideas sleep})
\]
Trigrams

\[ p(\text{Colorless green ideas sleep furiously}) = \]
\[ p(\text{Colorless}) \times \]
\[ p(\text{green | Colorless}) \times \]
\[ p(\text{ideas | Colorless green}) \times \]
\[ p(\text{sleep | green ideas}) \times \]
\[ p(\text{furiously | ideas sleep}) \]
Trigrams

\[
p(\text{Colorless green ideas sleep furiously}) = \\
p(\text{Colorless} \mid <\text{BOS}> <\text{BOS}>)* \\
p(\text{green} \mid <\text{BOS}> \text{Colorless})* \\
p(\text{ideas} \mid \text{Colorless green})* \\
p(\text{sleep} \mid \text{green ideas})* \\
p(\text{furiously} \mid \text{ideas sleep})
\]

*Consistent notation*: Pad the left with <BOS> (beginning of sentence) symbols
Trigrams

\[ p(\text{Colorless green ideas sleep furiously}) = \]
\[ p(\text{Colorless} \mid <\text{BOS}\rangle <\text{BOS}\rangle) \times \]
\[ p(\text{green} \mid <\text{BOS}\rangle \text{Colorless}) \times \]
\[ p(\text{ideas} \mid \text{Colorless green}) \times \]
\[ p(\text{sleep} \mid \text{green ideas}) \times \]
\[ p(\text{furiously} \mid \text{ideas sleep}) \times \]
\[ p(<\text{EOS}\rangle \mid \text{sleep furiously}) \]

*Consistent notation*: Pad the left with \(<\text{BOS}\rangle\) (beginning of sentence) symbols

*Fully proper distribution*: Pad the right with a single \(<\text{EOS}\rangle\) symbol
## N-Gram Terminology

<table>
<thead>
<tr>
<th>n</th>
<th>Commonly called</th>
<th>History Size (Markov order)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unigram</td>
<td>0</td>
<td>p(furiously)</td>
</tr>
</tbody>
</table>
# N-Gram Terminology

<table>
<thead>
<tr>
<th>n</th>
<th>Commonly called</th>
<th>History Size (Markov order)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unigram</td>
<td>0</td>
<td>p(furiously)</td>
</tr>
<tr>
<td>2</td>
<td>bigram</td>
<td>1</td>
<td>p(furiously</td>
</tr>
</tbody>
</table>
## N-Gram Terminology

<table>
<thead>
<tr>
<th>n</th>
<th>Commonly called</th>
<th>History Size (Markov order)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unigram</td>
<td>0</td>
<td>p(furiously)</td>
</tr>
<tr>
<td>2</td>
<td>bigram</td>
<td>1</td>
<td>p(furiously</td>
</tr>
<tr>
<td>3</td>
<td>trigram (3-gram)</td>
<td>2</td>
<td>p(furiously</td>
</tr>
</tbody>
</table>
## N-Gram Terminology

<table>
<thead>
<tr>
<th>$n$</th>
<th>Commonly called</th>
<th>History Size (Markov order)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unigram</td>
<td>0</td>
<td>p(furiously)</td>
</tr>
<tr>
<td>2</td>
<td>bigram</td>
<td>1</td>
<td>p(furiously</td>
</tr>
<tr>
<td>3</td>
<td>trigram (3-gram)</td>
<td>2</td>
<td>p(furiously</td>
</tr>
<tr>
<td>4</td>
<td>4-gram</td>
<td>3</td>
<td>p(furiously</td>
</tr>
<tr>
<td>$n$</td>
<td>n-gram</td>
<td>$n-1$</td>
<td>$p(w_i</td>
</tr>
</tbody>
</table>
N-Gram Probability

\[ p(w_1, w_2, w_3, \ldots, w_S) = \prod_{i=1}^{S} p(w_i | w_{i-N+1}, \ldots, w_{i-1}) \]
Count-Based N-Grams (Unigrams)

\[ p(\text{item}) \propto \text{count}(\text{item}) \]
Count-Based N-Grams (Unigrams)

\[ p(z) \propto \text{count}(z) \]
Count-Based N-Grams (Unigrams)

\[ p(z) \propto \text{count}(z) \]

\[ = \frac{\text{count}(z)}{\sum_v \text{count}(v)} \]
Count-Based N-Grams (Unigrams)

\[ p(z) \propto \text{count}(z) \]

\[ \text{count}(z) = \frac{\text{count}(z)}{\sum_v \text{count}(v)} \]
Count-Based N-Grams (Unigrams)

\[ p(z) \propto \text{count}(z) \]

\[ = \frac{\text{count}(z)}{W} \]

word type

number of tokens observed
The film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Word (Type)</th>
<th>Raw Count</th>
<th>Normalization</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>film</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>got</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>great</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>opening</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>went</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>on</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>become</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hit</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Count-Based N-Grams (Unigrams)**

The film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Word (Type)</th>
<th>Raw Count</th>
<th>Normalization</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>film</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>got</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>great</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>opening</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>went</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>on</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>become</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hit</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total: 16
Count-Based N-Grams (Unigrams)

The film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Word</th>
<th>Raw Count</th>
<th>Normalization</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td>film</td>
<td>2</td>
<td></td>
<td>1/8</td>
</tr>
<tr>
<td>got</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td></td>
<td>1/8</td>
</tr>
<tr>
<td>great</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td>opening</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td>and</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td>went</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td>on</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td>to</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td>become</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td>hit</td>
<td>1</td>
<td></td>
<td>1/16</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>
Count-Based N-Grams (Trigrams)

\[ p(z|x, y) \propto \text{count}(x, y, z) \]

order matters in conditioning

order matters in \textit{count}
Count-Based N-Grams (Trigrams)

\[ p(z|x, y) \propto \text{count}(x, y, z) \]

order matters in conditioning

\[
\text{count}(x, y, z) \neq \text{count}(x, z, y) \neq \text{count}(y, x, z) \neq \ldots
\]
Count-Based N-Grams (Trigrams)

\[ p(z|x, y) \propto \text{count}(x, y, z) \]

\[ = \frac{\text{count}(x, y, z)}{\sum_v \text{count}(x, y, v)} \]
Count-Based N-Grams (Trigrams)

The film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Context</th>
<th>Word (Type)</th>
<th>Raw Count</th>
<th>Normalization</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>The film</td>
<td>The</td>
<td>0</td>
<td></td>
<td>0/1</td>
</tr>
<tr>
<td>The film</td>
<td>film</td>
<td>0</td>
<td>1</td>
<td>0/1</td>
</tr>
<tr>
<td>The film</td>
<td>got</td>
<td>1</td>
<td></td>
<td>1/1</td>
</tr>
<tr>
<td>The film</td>
<td>went</td>
<td>0</td>
<td></td>
<td>0/1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a great</td>
<td>great</td>
<td>0</td>
<td></td>
<td>0/1</td>
</tr>
<tr>
<td>a great</td>
<td>opening</td>
<td>1</td>
<td>1</td>
<td>1/1</td>
</tr>
<tr>
<td>a great</td>
<td>and</td>
<td>0</td>
<td></td>
<td>0/1</td>
</tr>
<tr>
<td>a great</td>
<td>the</td>
<td>0</td>
<td></td>
<td>0/1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
the film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Context</th>
<th>Word (Type)</th>
<th>Raw Count</th>
<th>Normalization</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>the film</td>
<td>the</td>
<td>0</td>
<td>2</td>
<td>0/1</td>
</tr>
<tr>
<td>the film</td>
<td>film</td>
<td>0</td>
<td></td>
<td>0/1</td>
</tr>
<tr>
<td>the film</td>
<td>got</td>
<td>1</td>
<td></td>
<td>1/2</td>
</tr>
<tr>
<td>the film</td>
<td>went</td>
<td>1</td>
<td></td>
<td>1/2</td>
</tr>
<tr>
<td>a great</td>
<td>great</td>
<td>0</td>
<td>1</td>
<td>0/1</td>
</tr>
<tr>
<td>a great</td>
<td>opening</td>
<td>1</td>
<td></td>
<td>1/1</td>
</tr>
<tr>
<td>a great</td>
<td>and</td>
<td>0</td>
<td></td>
<td>0/1</td>
</tr>
<tr>
<td>a great</td>
<td>the</td>
<td>0</td>
<td></td>
<td>0/1</td>
</tr>
</tbody>
</table>
Outline

Defining Language Models

Breaking & Fixing Language Models

Evaluating Language Models
Maximum Likelihood Estimates

\[ p(\text{item}) \propto \text{count}(\text{item}) \]

Maximizes the likelihood of the training set

Do different corpora look the same?

Low(er) bias, high(er) variance

For large data: can actually do reasonably well
n = 1

, , land of in , a teachers The , wilds the and gave a Etienne any
two beginning without probably heavily that other useless the the

a different . the able mines , unload into in foreign the the be either other Britain finally avoiding , for of have the cure , the Gutenberg-tm ; of being can as country in authority deviates as d seldom and They employed about from business marshal materials than in , they
These varied with it to the civil wars, therefore, it did not for the company had the East India, the mechanical, the sum which were by barter, vol. i, and, conveniencies of all made to purchase a council of landlords, constitute a sum as an argument, having thus forced abroad, however, and influence in the one, or banker, will there was encouraged and more common trade to corrupt, profit, it; but a master does not, twelfth year the consent that of volunteers and […]

, the other hand, it certainly it very earnestly entreat both nations.

In opulent nations in a revenue of four parts of production.
n = 3

His employer, if silver was regulated according to the temporary and occasional event.

What goods could bear the expense of defending themselves, than in the value of different sorts of goods, and placed at a much greater, there have been the effects of self-deception, this attention, but a very important ones, and which, having become of less than they ever were in this agreement for keeping up the business of weighing.

After food, clothes, and a few months longer credit than is wanted, there must be sufficient to keep by him, are of such colonies to surmount.

They facilitated the acquisition of the empire, both from the rents of land and labour of those pedantic pieces of silver which he can afford to take from the duty upon every quarter which they have a more equable distribution of employment.
To buy in one market, in order to have it; but the 8th of George III.

The tendency of some of the great lords, gradually encouraged their villains to make upon the prices of corn, cattle, poultry, etc.

Though it may, perhaps, in the mean time, that part of the governments of New England, the market, trade cannot always be transported to so great a number of seamen, not inferior to those of other European nations from any direct trade to America.

The farmer makes his profit by parting with it.

But the government of that country below what it is in itself necessarily slow, uncertain, liable to be interrupted by the weather.
Maximum Likelihood Estimates

\[ p(\text{item}) \propto \text{count}(\text{item}) \]

Maximizes the likelihood of the training set

Do different corpora look the same?

For large data: can actually do reasonably well
0s Are Not Your (Language Model's) Friend

\[ p(\text{item}) \propto \text{count}(\text{item}) = 0 \rightarrow \]
\[ p(\text{item}) = 0 \]
0s Are Not Your (Language Model’s) Friend

\[ p(\text{item}) \propto \text{count}(\text{item}) = 0 \rightarrow p(\text{item}) = 0 \]

0 probability \(\rightarrow\) item is *impossible*

0s annihilate: \(x \cdot y \cdot z \cdot 0 = 0\)

Language is creative:
- new words keep appearing
- existing words could appear in known contexts

How much do you trust your data?
Add-$\lambda$ estimation

Laplace smoothing, Lidstone smoothing

Pretend we saw each word $\lambda$ more times than we did

Add $\lambda$ to all the counts
Add-\(\lambda\) estimation

Laplace smoothing, Lidstone smoothing

Pretend we saw each word \(\lambda\) more times than we did

\[
p(z) \propto \text{count}(z) + \lambda
\]

Add \(\lambda\) to all the counts
Add-$\lambda$ estimation

- Laplace smoothing,
- Lidstone smoothing

Pretend we saw each word $\lambda$ more times than we did

Add $\lambda$ to all the counts

\[ p(z) \propto \text{count}(z) + \lambda \]
\[ = \frac{\text{count}(z) + \lambda}{\sum_v(\text{count}(v) + \lambda)} \]
Add-\(\lambda\) estimation

Laplace smoothing, Lidstone smoothing

Pretend we saw each word \(\lambda\) more times than we did

\[
p(z) \propto \text{count}(z) + \lambda
= \frac{\text{count}(z) + \lambda}{W + V\lambda}
\]

Add \(\lambda\) to all the counts
The film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Word (Type)</th>
<th>Raw Count</th>
<th>Norm</th>
<th>Prob.</th>
<th>Add-(\lambda) Count</th>
<th>Add-(\lambda) Norm.</th>
<th>Add-(\lambda) Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>film</td>
<td>2</td>
<td>1/8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>got</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td>1/8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>great</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>opening</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>went</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>on</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>become</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hit</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>1</td>
<td>1/16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Add-1 N-Grams (Unigrams)
The film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Word (Type)</th>
<th>Raw Count</th>
<th>Norm</th>
<th>Prob.</th>
<th>Add-1 Count</th>
<th>Add-1 Norm.</th>
<th>Add-1 Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>film</td>
<td>2</td>
<td></td>
<td>1/8</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>got</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td></td>
<td>1/8</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>great</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>opening</td>
<td>1</td>
<td>16</td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>went</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>on</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>become</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hit</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Add-1 N-Grams (Unigrams)

The film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Word (Type)</th>
<th>Raw Count</th>
<th>Norm</th>
<th>Prob.</th>
<th>Add-1 Count</th>
<th>Add-1 Norm.</th>
<th>Add-1 Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>film</td>
<td>2</td>
<td></td>
<td>1/8</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>got</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td></td>
<td>1/8</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>great</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>opening</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>went</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>on</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>become</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hit</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Add-1 N-Grams (Unigrams)

The film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Word (Type)</th>
<th>Raw Count</th>
<th>Norm</th>
<th>Prob.</th>
<th>Add-1 Count</th>
<th>Add-1 Norm.</th>
<th>Add-1 Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>film</td>
<td>2</td>
<td></td>
<td>1/8</td>
<td>3</td>
<td></td>
<td>=1/10</td>
</tr>
<tr>
<td>got</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td></td>
<td>1/8</td>
<td>3</td>
<td></td>
<td>=1/10</td>
</tr>
<tr>
<td>great</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>opening</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>and</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>went</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>on</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>to</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>become</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>hit</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
<tr>
<td>.</td>
<td>1</td>
<td></td>
<td>1/16</td>
<td>2</td>
<td></td>
<td>=1/15</td>
</tr>
</tbody>
</table>

16 + 14*1 = 30

=1/15
Backoff and Interpolation

Sometimes it helps to use **less** context condition on less context for contexts you haven’t learned much
Backoff and Interpolation

Sometimes it helps to use **less** context condition on less context for contexts you haven’t learned much about

**Backoff:**

use trigram if you have good evidence otherwise bigram, otherwise unigram
Backoff and Interpolation

Sometimes it helps to use **less** context condition on less context for contexts you haven’t learned much about

**Backoff:**
use trigram if you have good evidence otherwise bigram, otherwise unigram

**Interpolation:**
mix (average) unigram, bigram, trigram
Linear Interpolation

Simple interpolation

\[
p(y \mid x) = \lambda p_2(y \mid x) + (1 - \lambda)p_1(y)
\]

\[0 \leq \lambda \leq 1\]
Linear Interpolation

Simple interpolation

\[ p(y | x) = \lambda p_2(y | x) + (1 - \lambda)p_1(y) \]

\[ 0 \leq \lambda \leq 1 \]

Condition on context

\[ p(z | x, y) = \lambda_3(x, y)p_3(z | x, y) + \lambda_2(y)p_2(z | y) + \lambda_1p_1(z) \]
Backoff

Trust your statistics, up to a point

\[ p(z|x, y) \propto \begin{cases} p_3(z \mid x, y) & \text{if count}(x, y, z) > 0 \\ p_2(z \mid y) & \text{otherwise} \end{cases} \]
Discounted Backoff

Trust your statistics, up to a point

\[ p(z|x, y) \propto \begin{cases} 
   p_3(z |x, y) - d & \text{if count}(x, y, z) > 0 \\
   \beta(x, y)p_2(z |y) & \text{otherwise}
\end{cases} \]
Discounted Backoff

Trust your statistics, up to a point

\[ p(z|x, y) = \begin{cases} 
    p_3(z|x, y) - d & \text{if count}(x, y, z) > 0 \\
    \beta(x, y)p_2(z|y) & \text{otherwise}
\end{cases} \]

- Discount constant
- Context-dependent
- Normalization constant
Setting Hyperparameters

Use a **development** corpus

Choose $\lambda$s to maximize the probability of dev data:

- Fix the N-gram probabilities (on the training data)
- Then search for $\lambda$s that give largest probability to held-out set:
Implementation: Unknown words

Create an unknown word token \(<\text{UNK}>\)

Training:

1. Create a fixed lexicon \(L\) of size \(V\)
2. Change any word not in \(L\) to \(<\text{UNK}>\)
3. Train \(LM\) as normal

Evaluation:

Use UNK probabilities for any word not in training
Other Kinds of Smoothing

Interpolated (modified) Kneser-Ney

Idea: How “productive” is a context?
How many different word types $v$ appear in a context $x, y$

Good-Turing

Partition words into classes of occurrence
Smooth class statistics
Properties of classes are likely to predict properties of other classes

Witten-Bell

Idea: Every observed type was at some point novel
Give MLE prediction for novel type occurring
Outline

Defining Language Models

Breaking & Fixing Language Models

Evaluating Language Models
Evaluating Language Models

What is “correct?”
What is working “well?”
Evaluating Language Models

What is “correct?”
What is working “well?”

Training Data
Dev Data
Test Data
Evaluating Language Models

What is “correct?”
What is working “well?”

- **Training Data**: acquire primary statistics for learning model parameters
- **Dev Data**: fine-tune any secondary (hyper)parameters
- **Test Data**: perform final evaluation
Evaluating Language Models

What is “correct?”
What is working “well?”

- Training Data: acquire primary statistics for learning model parameters
- Dev Data: fine-tune any secondary (hyper)parameters
- Test Data: perform final evaluation

DO NOT TUNE ON THE TEST DATA
Evaluating Language Models

*What is “correct?”*
*What is working “well?”*

**Extrinsic:** Evaluate LM in downstream task

- Test an MT, ASR, etc. system and see which LM does better
- Propagate & conflate errors
Evaluating Language Models

What is “correct?”

What is working “well?”

**Extrinsic:** Evaluate LM in downstream task

   Test an MT, ASR, etc. system and see which LM does better

   Propagate & conflate errors

**Intrinsic:** Treat LM as its own downstream task

   Use perplexity (from information theory)
Perplexity

Lower is better: lower perplexity ➔ less surprised

More outcomes ➔ More surprised
Fewer outcomes ➔ Less surprised
Perplexity

Lower is better: lower perplexity $\Rightarrow$ less surprised

$$\text{perplexity} = \exp\left(\frac{-1}{M} \sum_{i=1}^{M} \log p(w_i | h_i)\right)$$

$n$-gram history (n-1 items)
Lower is better: lower perplexity $\rightarrow$ less surprised

\[
\text{perplexity} = \exp\left(\frac{-1}{M} \sum_{i=1}^{M} \log p(w_i \mid h_i)\right)
\]

$\geq 0, \leq 1$: higher
Perplexity

Lower is better: lower perplexity $\Rightarrow$ less surprised

$$\text{perplexity} = \exp\left(\frac{-1}{M} \sum_{i=1}^{M} \log p(w_i \mid h_i)\right)$$

$\leq 0$: higher

$\geq 0, \leq 1$: higher
**Perplexity**

Lower is better: lower perplexity $\rightarrow$ less surprised

$$\text{perplexity} = \exp\left(\frac{-1}{M} \sum_{i=1}^{M} \log p(w_i \mid h_i)\right)$$

- $\leq 0$: higher
- $\geq 0, \leq 1$: higher
- $\leq 0$, higher
Perplexity

Lower is better: lower perplexity $\rightarrow$ less surprised

\[
\text{perplexity} = \exp\left(\frac{-1}{M} \sum_{i=1}^{M} \log p(w_i \mid h_i)\right)
\]

- $\geq 0$, lower is better
- $\leq 0$: higher
- $\geq 0$, $\leq 1$: higher
- $\leq 0$, higher
Perplexity

Lower is better: lower perplexity $\rightarrow$ less surprised

$$\text{perplexity} = \exp\left( -\frac{1}{M} \sum_{i=1}^{M} \log p(w_i \mid h_i) \right)$$

$\geq 0$, lower is better

$\leq 0$: higher

$\geq 0$, $\leq 1$: higher

$\leq 0$, higher

$\geq 0$, lower
Perplexity

Lower is better: lower perplexity $\Rightarrow$ less surprised

Perplexity = $\exp\left(\frac{-1}{M} \sum_{i=1}^{M} \log p(w_i \mid h_i)\right)$

- $\geq 0$, lower is better
- $\leq 0$: higher
- $\geq 0$, $\leq 1$: higher
- $\leq 0$, higher
- $\geq 0$, lower
Perplexity

Lower is better: lower perplexity $\Rightarrow$ less surprised

$$\text{perplexity} = \exp\left(\frac{-1}{M} \sum_{i=1}^{M} \log p(w_i \mid h_i)\right)$$

$$= \sqrt{\prod_{i=1}^{M} \frac{1}{p(w_i \mid h_i)}}$$

weighted geometric average
Perplexity

Lower is better: lower perplexity $\Rightarrow$ less surprised

$$\text{perplexity} = \sqrt[\text{M}]{\prod_{i=1}^{\text{M}} \frac{1}{p(w_i \mid h_i)}}$$

471/671: Branching factor
Outline

Defining Language Models

Breaking & Fixing Language Models

Evaluating Language Models