Special Topics in Natural Language Processing

CS698O

Ashutosh Modi
CSE Department, IIT Kanpur

Lecture 5: Language Models 2
Jan 13, 2020

LM Review

$$p(X_n = w_n, X_{n-1} = w_{n-1}, \ldots, X_1 = w_1) = \prod_{i=1}^{n} p(X_i = w_i \mid X_{i-1} = w_{i-1}, X_{i-2} = w_{i-2})$$

$$where,$$

$$w_0 = w_{-1} = START; \quad w_n = STOP; \quad w_i \in \mathcal{V} \;\; \forall \; 1 \leq i \leq n-1$$

LM Review

$$p(X_n = w_n, X_{n-1} = w_{n-1}, \ldots, X_1 = w_1) = \prod_{i=1}^{n} p(X_i = w_i \mid X_{i-1} = w_{i-1}, X_{i-2} = w_{i-2})$$

$$where,$$

$$w_0 = w_{-1} = START; \quad w_n = STOP; \quad w_i \in \mathcal{V} \;\; \forall \; 1 \leq i \leq n-1$$

$$p(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} \theta(w_i \mid w_{i-1}, w_{i-2})$$

LM Review

$$p(X_n = w_n, X_{n-1} = w_{n-1}, \ldots, X_1 = w_1) = \prod_{i=1}^{n} p(X_i = w_i \mid X_{i-1} = w_{i-1}, X_{i-2} = w_{i-2})$$

$$where,$$

$$w_0 = w_{-1} = START; \quad w_n = STOP; \quad w_i \in \mathcal{V} \;\; \forall \, 1 \leq i \leq n-1$$

$$p(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} \theta(w_i \mid w_{i-1}, w_{i-2})$$

$$\theta(q \mid r, s) = \frac{C(q, r, s)}{C(r, s)}$$

$C(q, r, s) =$ Number of times trigram {s,r,q} occurs in the corpus

$C(r, s) =$ Number of times bigram {s,r} occurs in the corpus

LM Review

TRI-GRAM

$$\theta(q \mid r, s) = \frac{C(q, r, s)}{C(r, s)}$$

BI-GRAM

$$\theta(q \mid r) = \frac{C(q, r)}{C(r)}$$

UNI-GRAM

$$\theta(q) = \frac{C(q)}{N}$$

Problems with MLE

$$\theta(q \mid r, s) = \frac{C(q, r, s)}{C(r, s)}$$

Overfitting (Sparsity):

$\text{Many trigram estimates } (\theta(q \mid r, s)) \text{ are } 0$ Hapax Legomenon

Indeterminate Estimates:

$C(r, s) \text{ can be } 0$

Linear Interpolation

- The key idea is to rely on lower order statistics.
- Each n-gram has strengths and weaknesses
- For trigram estimates, also rely on bigram and unigram.

Linear Interpolation

- The key idea is to rely on lower order statistics.
- Each n-gram has strengths and weaknesses
- For trigram estimates, also rely on bigram and unigram.

$$\theta(q \mid r, s) = \lambda_1 \times \theta(q \mid r, s) + \lambda_2 \times \theta(q \mid r) + \lambda_3 \times \theta(q)$$

Linear Interpolation

- The key idea is to rely on lower order statistics.
- Each n-gram has strengths and weaknesses
- For trigram estimates, also rely on bigram and unigram.

$$\theta(q \mid r, s) = \lambda_1 \times \theta(q \mid r, s) + \lambda_2 \times \theta(q \mid r) + \lambda_3 \times \theta(q)$$

$$\lambda_i \geq 0 \quad \forall i \in \{1, 2, 3\}$$

$$\sum_{i=1,2,3} \lambda_i = 1$$

Linear Interpolation: Estimating Smoothing Parameters

$$\theta(q \mid r, s) = \lambda_1 \times \theta(q \mid r, s) + \lambda_2 \times \theta(q \mid r) + \lambda_3 \times \theta(q)$$

$$\lambda_i \geq 0 \quad \forall i \in \{1, 2, 3\}$$

$$\sum_{i=1,2,3} \lambda_i = 1$$

Linear Interpolation: Estimating Smoothing Parameters

$$\theta(q \mid r, s) = \lambda_1 \times \theta(q \mid r, s) + \lambda_2 \times \theta(q \mid r) + \lambda_3 \times \theta(q)$$

$$\lambda_i \geq 0 \quad \forall i \in \{1, 2, 3\}$$

$$\sum_{i=1,2,3} \lambda_i = 1$$

Use Development Set

Linear Interpolation: Estimating Smoothing Parameters

Use Development Set

$$\mathcal{L}(\lambda_1, \lambda_2, \lambda_3) = \sum_{q,r,s} C(q,r,s) \log \theta(q|r,s)$$

$$= \sum_{q,r,s} C(q,r,s) \log \left(\lambda_1 \times \theta(q|r,s) + \lambda_2 \times \theta(q|r) + \lambda_3 \times \theta(q)\right)$$

Linear Interpolation: Estimating Smoothing Parameters

Use Development Set

$$\mathcal{L}(\lambda_1, \lambda_2, \lambda_3) = \sum_{q,r,s} C(q,r,s) \log \theta(q|r,s)$$

$$= \sum_{q,r,s} C(q,r,s) \log \left(\lambda_1 \times \theta(q|r,s) + \lambda_2 \times \theta(q|r) + \lambda_3 \times \theta(q)\right)$$

$$\arg \max_{\lambda_1,\lambda_2,\lambda_3} \mathcal{L}(\lambda_1, \lambda_2, \lambda_3)$$

$$\lambda_i \geq 0 \quad \forall i \in \{1,2,3\}$$

$$\sum_{i=1,2,3} \lambda_i = 1$$

Linear Interpolation: Smoothing Parameters

$$\theta(q \mid r, s) = \lambda_1 \times \theta(q \mid r, s) + \lambda_2 \times \theta(q \mid r) + \lambda_3 \times \theta(q)$$

- Lambdas are indicative of confidence in each of the n-gram model
- For example, higher $\boldsymbol{\lambda}_1$ (= 1) indicates that trigram model alone itself will give a good estimate.
- In practice, there is dependence between $\boldsymbol{\lambda}$'s and n-gram counts.
- E.g. $\boldsymbol{\lambda}_1$ should be large when C(r,s) is large OR $\boldsymbol{\lambda}_1$ = 0 when C(r,s) = 0

Linear Interpolation: Smoothing Parameters

$$\theta(q \mid r, s) = \lambda_1 \times \theta(q \mid r, s) + \lambda_2 \times \theta(q \mid r) + \lambda_3 \times \theta(q)$$

$$\gamma > 0$$

$$\lambda_1 = \frac{C(r, s)}{C(r, s) + \gamma}$$

$$\lambda_2 = (1 - \lambda_1) \times \frac{C(r)}{C(r) + \gamma}$$

$$\lambda_3 = 1 - \lambda_1 - \lambda_2$$

Linear Interpolation: Bucketing

$$\theta(q \mid r, s) = \lambda_1 \times \theta(q \mid r, s) + \lambda_2 \times \theta(q \mid r) + \lambda_3 \times \theta(q)$$

- Smoothing parameters should vary depending on bigram count
- $\boldsymbol{\lambda}_1$ should be high when bigram count is high, $\boldsymbol{\lambda}_2$ should be high when unigram count is high
- This can be done via ***Bucketing***

Linear Interpolation: Bucketing

$$\theta(q \mid r, s) = \lambda_1 \times \theta(q \mid r, s) + \lambda_2 \times \theta(q \mid r) + \lambda_3 \times \theta(q)$$

$$\Pi : (r, s) \rightarrow \{1, 2, 3, \ldots, K\}$$

Function Π defines a partition of bigrams into K different subsets or buckets

Linear Interpolation: Bucketing

$$\theta(q \mid r, s) = \lambda_1 \times \theta(q \mid r, s) + \lambda_2 \times \theta(q \mid r) + \lambda_3 \times \theta(q)$$

$$\Pi : (r, s) \rightarrow \{1, 2, 3, \ldots, K\}$$

Example:

$$\Pi(r, s) = 1 \quad \text{if } C(r, s) > 0$$
$$\Pi(r, s) = 2 \quad \text{if } C(r, s) = 0 \text{ and } C(s) > 0$$
$$\Pi(r, s) = 3 \quad \text{otherwise}$$

Linear Interpolation: Bucketing

$$\theta(q \mid r, s) = \lambda_1 \times \theta(q \mid r, s) + \lambda_2 \times \theta(q \mid r) + \lambda_3 \times \theta(q)$$

$$\Pi : (r, s) \rightarrow \{1, 2, 3, \ldots, K\}$$

Example:

$$\begin{aligned}
\Pi(r,s) &= 1 \quad \text{if } 100 \leq C(r,s) \\
\Pi(r,s) &= 2 \quad \text{if } 50 \leq C(r,s) < 100 \\
\Pi(r,s) &= 3 \quad \text{if } 20 \leq C(r,s) < 50 \\
\Pi(r,s) &= 4 \quad \text{if } 10 \leq C(r,s) < 20 \\
\Pi(r,s) &= 5 \quad \text{if } 5 \leq C(r,s) < 10 \\
\Pi(r,s) &= 6 \quad \text{if } 2 \leq C(r,s) < 5 \\
\Pi(r,s) &= 7 \quad \text{if } C(r,s) = 1 \\
\Pi(r,s) &= 8 \quad \text{if } C(r,s) = 0 \text{ and } C(s) > 0 \\
\Pi(r,s) &= 9 \quad \text{otherwise}
\end{aligned}$$

Linear Interpolation: Bucketing

$$\Pi : (r, s) \rightarrow \{1, 2, 3, \ldots, K\}$$

Each bucket has it's own set of smoothing parameters: $\lambda_1^{(k)}, \lambda_2^{(k)}, \lambda_3^{(k)}$

Linear Interpolation: Bucketing

$$\Pi : (r, s) \rightarrow \{1, 2, 3, \ldots, K\}$$

Each bucket has it's own set of smoothing parameters: $\lambda_1^{(k)}, \lambda_2^{(k)}, \lambda_3^{(k)}$

$$k = \Pi(r, s)$$

$$\theta(q \mid r, s) = \lambda_1^{(k)} \times \theta(q \mid r, s) + \lambda_2^{(k)} \times \theta(q \mid r) + \lambda_3^{(k)} \times \theta(q)$$

Linear Interpolation: Bucketing

$$\Pi : (r, s) \rightarrow \{1, 2, 3, \ldots, K\}$$

Each bucket has it's own set of smoothing parameters: $\lambda_1^{(k)}, \lambda_2^{(k)}, \lambda_3^{(k)}$

$$k = \Pi(r, s)$$

$$\theta(q \mid r, s) = \lambda_1^{(k)} \times \theta(q \mid r, s) + \lambda_2^{(k)} \times \theta(q \mid r) + \lambda_3^{(k)} \times \theta(q)$$

$$\lambda_1^{(k)} \geq 0, \lambda_2^{(k)} \geq 0, \lambda_3^{(k)} \geq 0$$

$$\lambda_1^{(k)} + \lambda_2^{(k)} + \lambda_3^{(k)} = 1$$

Linear Interpolation: Bucketing Estimation

Use Development Set

Linear Interpolation: Bucketing Estimation

Use Development Set

$$\mathcal{L} = \sum_{q,r,s} C(q,r,s) \log \theta(q|r,s)$$

$$= \sum_{q,r,s} C(q,r,s) \log \left(\lambda_1^{(\Pi(r,s))} \times \theta(q|r,s) + \lambda_2^{(\Pi(r,s))} \times \theta(q|r) + \lambda_3^{(\Pi(u,v))} \times \theta(q) \right)$$

$$= \sum_{k=1}^{K} \sum_{q,r,s:\Pi(r,s)=k} C(q,r,s) \log \left(\lambda_1^{(k)} \times \theta(q|r,s) + \lambda_2^{(k)} \times \theta(q|r) + \lambda_3^{(k)} \times \theta(q) \right)$$

$$\lambda_1^{(k)} \geq 0, \lambda_2^{(k)} \geq 0, \lambda_3^{(k)} \geq 0$$

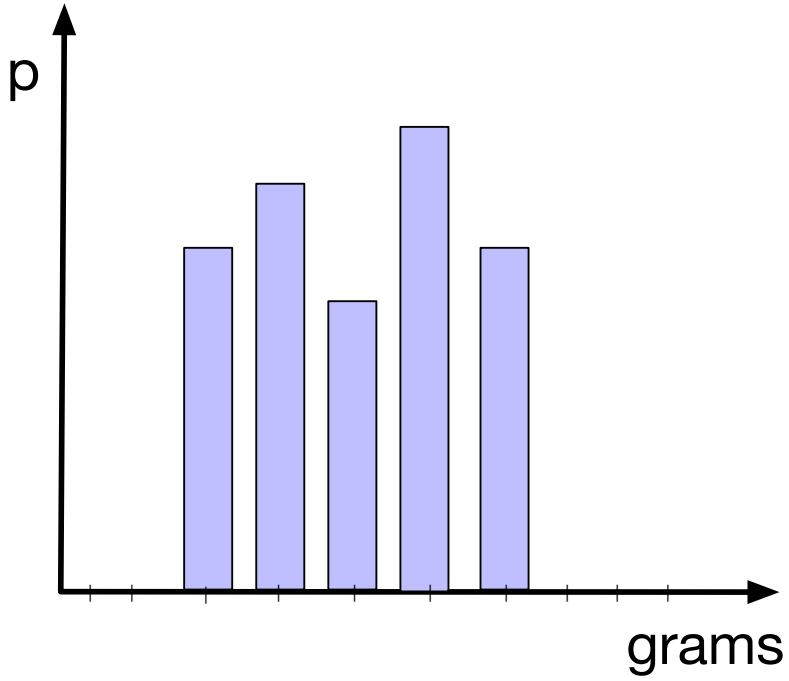
$$\lambda_1^{(k)} + \lambda_2^{(k)} + \lambda_3^{(k)} = 1$$

Discounting and Smoothing

The key idea is to take away some probability mass from non-zero counts and distribute to counts of grams never seen before.

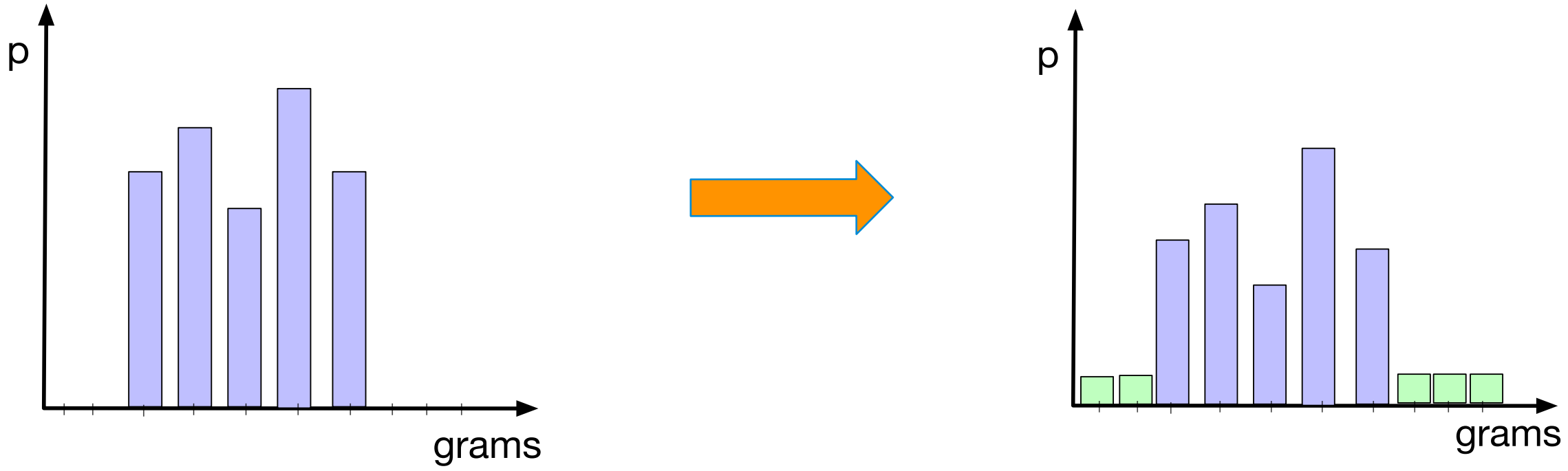
Discounting and Smoothing

The key idea is to take away some probability mass from non-zero counts and distribute to counts of grams never seen before.



Discounting and Smoothing

The key idea is to take away some probability mass from non-zero counts and distribute to counts of grams never seen before.



Discounting

The key idea is to discount off the non-zero counts

Discounting

The key idea is to discount off the non-zero counts

$$\theta(q \mid r) = \frac{C(q, r)}{C(r)}$$

Discounting

The key idea is to discount off the non-zero counts

$$\theta(q \mid r) = \frac{C(q,r)}{C(r)}$$

$$\text{For } C(q,r) > 0$$

$$C^*(q,r) := C(q,r) - \beta$$

$$0 \leq \beta \leq 1$$

Discounting

The key idea is to discount off the non-zero counts

$$\text{For } C(q,r) > 0$$

$$C^*(q,r) := C(q,r) - \beta$$

$$0 \leq \beta \leq 1$$

$$\theta(q \mid r) = \frac{C^*(q,r)}{C(r)}$$

Discounting

$$\theta(q \mid r) = \frac{C^*(q, r)}{C(r)}$$

What about the missing probability mass?

Discounting

$$\theta(q \mid r) = \frac{C^*(q, r)}{C(r)}$$

What about the missing probability mass?

$$\alpha(r) = 1 - \sum_{q:C(q,r)>0} \frac{C^*(q, r)}{C(r)}$$

Discounting

$$\theta(q \mid r) = \frac{C^*(q, r)}{C(r)}$$

What about the missing probability mass?

$$\alpha(r) = 1 - \sum_{q:C(q,r)>0} \frac{C^*(q, r)}{C(r)}$$

Divide the missing probability mass among bigrams with zero counts

Discounting

$$\mathcal{A}(r) = \{q : C(q, r) > 0\}$$
$$\mathcal{B}(r) = \{q : C(q, r) = 0\}$$

$$\theta_D(q|r) = \begin{cases} \frac{C^*(q,r)}{C(r)} & \text{If } q \in \mathcal{A}(r) \\ \alpha(r) \times \frac{\theta(q)}{\sum_{q \in \mathcal{B}(r)} \theta(q)} & \text{If } q \in \mathcal{B}(r) \end{cases}$$

Discounting: Trigrams

$$\theta_D(q|r,s) = \begin{cases} \frac{C^*(q,r,s)}{C(r,s)} & \text{If } q \in \mathcal{A}(r,s) \\ \alpha(r,s) \times \frac{\theta(q|r)}{\sum_{q \in \mathcal{B}(r,s)} \theta(q|r)} & \text{If } q \in \mathcal{B}(r,s) \end{cases}$$

$$\mathcal{A}(r,s) = \{q : C(q,r,s) > 0\}$$

$$\mathcal{B}(r,s) = \{q : C(q,r,s) = 0\}$$

$$C^*(q,r,s) := C(q,r,s) - \beta$$

Discounting Estimation

How to set the value of beta?

Use Development Set

Discounting Estimation

How to set the value of beta?

Use Development Set

$$\mathcal{L}(\beta) = \sum_{q,r,s} C(q,r,s) \log \theta_D(q|r,s)$$

$$\underset{\beta=\{0.1,0.2,\ldots,0.9\}}{\operatorname{argmax}} \mathcal{L}(\beta)$$

Laplace Smoothing (Add-one Smoothing)

The key idea is to add one to all counts and normalize

Laplace Smoothing (Add-one Smoothing)

The key idea is to add one to all counts and normalize

UNI-GRAM

$$\theta(q) = \frac{C(q) + 1}{N + \mid \mathcal{V} \mid}$$

N is total number of words in the corpus, and $\mid \mathcal{V} \mid$ is the vocabulary size

Laplace Smoothing (Add-one Smoothing)

The key idea is to add one to all counts and normalize

UNI-GRAM

$$\theta(q) = \frac{C(q) + 1}{N + \mid \mathcal{V} \mid}$$

BI-GRAM

$$\theta(q \mid r) = \frac{C(q, r) + 1}{C(r) + \mid \mathcal{V} \mid}$$

Laplace Smoothing (Add-one Smoothing)

The key idea is to add one to all counts and normalize

UNI-GRAM

$$\theta(q) = \frac{C(q) + 1}{N + \mid \mathcal{V} \mid}$$

BI-GRAM

$$\theta(q \mid r) = \frac{C(q, r) + 1}{C(r) + \mid \mathcal{V} \mid}$$

TRI-GRAM

$$\theta(q \mid r, s) = \frac{C(q, r, s) + 1}{C(r, s) + \mid \mathcal{V} \mid}$$

Laplace Smoothing (Add-one Smoothing)

The key idea is to add one to all counts and normalize

UNI-GRAM

$$\theta(q) = \frac{C(q) + 1}{N + \mid \mathcal{V} \mid}$$

BI-GRAM

$$\theta(q \mid r) = \frac{C(q, r) + 1}{C(r) + \mid \mathcal{V} \mid}$$

TRI-GRAM

$$\theta(q \mid r, s) = \frac{C(q, r, s) + 1}{C(r, s) + \mid \mathcal{V} \mid}$$

NOTE: In all cases, $\mid \mathcal{V} \mid$ is the vocabulary size

Evaluating LMs

How do we know if one LM is better than the other?

Evaluating LMs

Given the test set,

$$\mathcal{S}_1 = \{w_1^{(1)}, w_2^{(1)}, \ldots, w_{n_1}^{(1)}\}$$
$$\mathcal{S}_2 = \{w_1^{(2)}, w_2^{(2)}, \ldots, w_{n_2}^{(2)}\}$$
$$\vdots$$
$$\mathcal{S}_m = \{w_1^{(m)}, w_2^{(m)}, \ldots, w_{n_m}^{(m)}\}$$

What is the probability of the test set?

Evaluating LMs

Given the test set,

$$\mathcal{S}_1 = \{w_1^{(1)}, w_2^{(1)}, \ldots, w_{n_1}^{(1)}\}$$
$$\mathcal{S}_2 = \{w_1^{(2)}, w_2^{(2)}, \ldots, w_{n_2}^{(2)}\}$$
$$\vdots$$
$$\mathcal{S}_m = \{w_1^{(m)}, w_2^{(m)}, \ldots, w_{n_m}^{(m)}\}$$

What is the probability of the test set?

$$\prod_{i=1}^{m} p(S_i)$$

Evaluating LMs

What is the probability of the test set?

$$\prod_{i=1}^{m} p(S_i)$$

Let,

$$M = \sum_{i=1}^{m} n_i$$

Evaluating LMs

$$\prod_{i=1}^{m} p(S_i)$$

$$M = \sum_{i=1}^{m} n_i$$

What is the average log probability of the test set?

Evaluating LMs

$$\prod_{i=1}^{m} p(S_i) \qquad M = \sum_{i=1}^{m} n_i$$

What is the average log probability of the test set?

$$l = \frac{1}{M} log_2 \prod_{i=1}^{m} p(S_i)$$

Evaluating LMs

$$\prod_{i=1}^{m} p(S_i) \qquad M = \sum_{i=1}^{m} n_i$$

What is the average log probability of the test set?

$$l = \frac{1}{M} log_2 \prod_{i=1}^{m} p(S_i)$$

$$= \frac{1}{M} \sum_{i=1}^{m} log_2 \; p(S_i)$$

Evaluating LMs

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2 \ p(S_i)$$

$$\text{Perplexity} \ := 2^{-l}$$

Evaluating LMs

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2 \ p(S_i)$$

$$\text{Perplexity} \ := 2^{-l}$$

The *smaller* the value of perplexity,
better the language model is at modeling unseen data.

Perplexity

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2\ p(S_i) \qquad \text{Perplexity} := 2^{-l}$$

$$\text{Suppose, } N = |\ \{\mathcal{V} \cup STOP\}\ |$$

$$\theta(q \mid r, s) = \frac{1}{N}$$

Uniform Distribution

Perplexity

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2\ p(S_i) \qquad \text{Perplexity} := 2^{-l}$$

$$\text{Suppose, } N = |\ \{\mathcal{V} \cup STOP\}\ |$$

$$\theta(q \mid r, s) = \frac{1}{N}$$

Uniform Distribution

What is the perplexity of this LM?

Perplexity

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2 \; p(S_i)$$

$$\text{Perplexity} := 2^{-l}$$

$$\text{Suppose, } N = \mid \{\mathcal{V} \cup STOP\} \mid$$

Uniform Distribution

$$\theta(q \mid r, s) = \frac{1}{N}$$

$$\text{Perplexity} = N$$

Under uniform LM, perplexity is equal to vocabulary size

Perplexity

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2\ p(S_i)$$

$$\text{Perplexity} := 2^{-l}$$

Perplexity is a measure of effective vocabulary size under the LM

Perplexity: alternative formulation

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2\ p(S_i)$$

$$\text{Perplexity} := 2^{-l}$$

Perplexity is a measure of effective vocabulary size under the LM

$$\text{Perplexity} = \frac{1}{t}$$

$$\text{where, } t = \sqrt[M]{\prod_{i=1}^{m} p(S_i)}$$

Perplexity: alternative formulation

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2\ p(S_i)$$

$$\text{Perplexity} := 2^{-l}$$

Perplexity is a measure of effective vocabulary size under the LM

$$\text{Perplexity} = \frac{1}{t}$$

$$\text{where,}\quad t = \sqrt[M]{\prod_{i=1}^{m} p(S_i)}$$

t is the geometric mean of the trigram probability terms in a Trigram LM

Perplexity: alternative formulation

$$\text{Perplexity} = \frac{1}{t}$$

$$\text{where, } t = \sqrt[M]{\prod_{i=1}^{m} p(S_i)}$$

$$\prod_{i=1}^{m} p\left(S_i\right) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \theta\left(w_j^{(i)} | w_{j-1}^{(i)}, w_{j-2}^{(i)}\right)$$

t is the geometric mean of the trigram probability terms in a Trigram LM

Perplexity

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2\ p(S_i)$$

$$\text{Perplexity} := 2^{-l}$$

$$\text{Perplexity} = \frac{1}{t}$$

$$\text{where, } t = \sqrt[M]{\prod_{i=1}^{m} p(S_i)}$$

What is the perplexity

if any of trigram probability $\theta(q \mid r, s)$ in the test set is zero?

Perplexity

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2\ p(S_i)$$

$$\text{Perplexity} := 2^{-l}$$

$$\text{Perplexity} = \frac{1}{t}$$

$$\text{where, } t = \sqrt[M]{\prod_{i=1}^{m} p(S_i)}$$

What is the perplexity

if any of trigram probability $\theta(q \mid r, s)$ in the test set is zero?

$$\infty$$

Perplexity: practical values

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2\ p(S_i)$$

$$\text{Perplexity} := 2^{-l}$$

$$\text{Perplexity} = \frac{1}{t}$$

$$\text{where, } t = \sqrt[M]{\prod_{i=1}^{m} p(S_i)}$$

$$\mathcal{V} = 50000$$

Model	Perplexity
Unigram	955
Bigram	137
Trigram	74

Perplexity: practical values

$$l = \frac{1}{M} \sum_{i=1}^{m} log_2\ p(S_i)$$

$$\text{Perplexity} := 2^{-l}$$

$$\text{Perplexity} = \frac{1}{t}$$

$$\text{where, } t = \sqrt[M]{\prod_{i=1}^{m} p(S_i)}$$

$$\mathcal{V} = 50000$$

Model	Perplexity
Unigram	955
Bigram	137
Trigram	74

SOTA using deep learning models : perplexity of 21.8 with vocabulary size of 800K

Summary

- MLE for LM gives simple way to estimate probabilities in terms of word counts.
- MLE has overfitting problems.
- We looked at interpolation technique
- We looked at smoothing techniques
- Perplexity

References

1. Michael Collin's NLP Lecture Notes: http://www.cs.columbia.edu/~mcollins/lm-spring2013.pdf
2. Chapter 4, Speech and Language Processing, Dan Jurafsky and James Martin
3. A Bit of Progress in Language Modeling: https://arxiv.org/pdf/cs/0108005.pdf

- Next class
  - Other types Language Models
  - Neural Language Models