Special Topics in Natural Language Processing

CS698O

Ashutosh Modi
CSE Department, IIT Kanpur

Lecture 4: Language Models 1
Jan 10, 2020

PROBABILITY REVIEW

Probability Review

Bayes Rule:

$$P(X,Y) = P(Y \mid X)P(X)$$
$$= P(X \mid Y)P(Y)$$

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

Probability Review

Bayes Rule:

$$P(X,Y) = P(Y \mid X)P(X)$$
$$= P(X \mid Y)P(Y)$$

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

Marginalization:

$$P(X) = \sum_{Y} P(X,Y)$$

$$P(Y) = \sum_{X} P(X,Y)$$

Probability Review

Bayes Rule:

$$P(X, Y) = P(Y \mid X)P(X)$$
$$= P(X \mid Y)P(Y)$$
$$P(Y \mid X) = \frac{P(X, Y)}{P(X)}$$
$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}$$

Marginalization:

$$P(X) = \sum_{Y} P(X, Y)$$
$$P(Y) = \sum_{X} P(X, Y)$$

Independence:

Given Joint Distribution:

$$P(Y_1, Y_2, \ldots, Y_N)$$

Suppose, Y_1 is independent of $\{Y_3, \ldots, Y_N\}$

$$\implies P(Y_1 \mid Y_2, Y_3, \ldots, Y_N) = P(Y_1 \mid Y_2)$$

Probability Review

Chain Rule:

$$p(X_1, X_2, \ldots, X_n) = p(X_n \mid X_{n-1}, \ldots, X_2, X_1) \times p(X_{n-1} \mid X_{n-2}, \ldots, X_2, X_1) \times$$
$$p(X_{n-2} \mid X_{n-3}, \ldots, X_2, X_1) \times \ldots\ldots$$
$$p(X_3 \mid X_2, X_1) \times p(X_2 \mid X_1) \times p(X_1)$$

Probability Review

Chain + Independence Rule:

Suppose, R.V.s occur in a sequence: $X_1, X_2, \ldots, X_n,$

Assume that every R.V. only depends on the previous one in the sequence then,

$$p(X_1, X_2, \ldots, X_n) = \ ?$$

Probability Review

Chain + Independence Rule:

Suppose, R.V.s occur in a sequence: $X_1, X_2, \ldots, X_n$,

Assume that every R.V. only depends on the previous one in the sequence then,

$$p(X_1, X_2, \ldots, X_n) = p(X_n \mid X_{n-1}) \times p(X_{n-1} \mid X_{n-2}) \times \ldots$$
$$p(X_3 \mid X_2) \times p(X_2 \mid X_1) \times p(X_1)$$

LANGUAGE MODELS

Based on Michael Collin's NLP Lecture Notes: http://www.cs.columbia.edu/~mcollins/lm-spring2013.pdf

Guess the next word

Lake Zurich is well known, its water is so transparent that _____

Guess the next word

Lake Zurich is well known, its water is so transparent that _____

The dog ran after the _____

Guess the next word

Lake Zurich is well known, its water is so transparent that _____

The dog ran after the _____

The book is in the _____

Guess the next word

Lake Zurich is well known, its water is so transparent that you

The dog ran after the mouse

The book is in the car

Guess the next word

Lake Zurich is well known, its water is so transparent that you

The dog ran after the mouse

The book is in the car

Language Model (LM) at work

$$p(w_k \mid w_{k-1}, w_{k-2}, \ldots, w_1)$$

Terminology

Vocabulary/Lexicon: Set of all tokens/words in the language

$$\mathcal{V} = \{\text{apple}, \text{banana}, \text{the}, \ldots, \text{zebra}\}$$

Sentence: Sequence of words

$$\mathcal{S} = w_1 w_2 w_3 \ldots w_n$$

$$\text{where } w_i \in \mathcal{V} \quad \forall i = 1, \ldots, n-1; \quad n \geq 1$$

$$\text{and } w_n = STOP$$

Terminology

Sentence: Sequence of words

$$\mathcal{S} = w_1 w_2 w_3 \ldots w_n$$

$$\text{where } w_i \in \mathcal{V} \quad \forall i = 1, \ldots, n-1; \quad n \geq 1$$

$$\text{and } w_n = STOP$$

the dog ran STOP
the cat is brown STOP
the class is tomorrow STOP
the the the STOP
apple STOP
STOP
…..

Terminology

Sentence: Sequence of words

$$\mathcal{S} = w_1 w_2 w_3 \ldots w_n$$

$$\text{where } w_i \in \mathcal{V} \quad \forall i = 1, \ldots, n-1; \quad n \geq 1$$

$$\text{and } w_n = STOP$$

the dog ran STOP
the cat is brown STOP
the class is tomorrow STOP
the the the STOP
apple STOP
STOP
.....

Sentence Set:

$$\mathcal{V}^\dagger = \{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \ldots\ldots\}$$

Language Model (LM)

LM Definition: Probability Distribution over sentences in $\mathcal{V}^\dagger$

Language Model (LM)

LM Definition: Probability Distribution over sentences in $\mathcal{V}^\dagger$

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

Language Model (LM)

LM Definition: Probability Distribution over sentences in $\mathcal{V}^\dagger$

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n) \geq 0 \qquad \forall w_i \in \mathcal{V}$$

$$\sum_{<w_1, w_2, \ldots, w_n> \in \mathcal{V}^\dagger} p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n) = 1$$

Language Model (LM)

LM Definition: Probability Distribution over sentences in $\mathcal{V}^\dagger$

$$p(w_1, w_2, \ldots, w_n)$$

$$p(w_1, w_2, \ldots, w_n) \geq 0$$

$$\sum_{<w_1, w_2, \ldots, w_n> \in \mathcal{V}^\dagger} p(w_1, w_2, \ldots, w_n) = 1$$

Why Language Models (LMs)?

- LMs used in several applications like speech recognition
- Neural LMs have shown SOTA results on many NLP tasks
- The LM techniques can be generalized to many other scenarios.

Language Model (LM)

LM Problem: How do we learn probability distribution over sentences?

$$p(w_1, w_2, \ldots, w_n) = ?$$

Probability Distribution

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

Naïve Approach: Count all sentences in a corpus.

Probability Distribution

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

Naïve Approach: Count all sentences in a corpus.

1. How many sentences of length ***n*** are there?

Probability Distribution

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

Naïve Approach: Count all sentences in a corpus.

1. How many sentences of length ***n*** are there? $|\mathcal{V}|^n$

Probability Distribution

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

Naïve Approach: Count all sentences in a corpus.

1. How many sentences of length ***n*** are there? $| \mathcal{V} |^n$

Assume number of words in the vocabulary is 40000 ($<<$ english vocabulary),

Suppose, each sentence is of length 10

Number of possible sentences $= 1.048576e + 46$

Probability Distribution

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

Naïve Approach: Count all sentences in a corpus.

1. How many sentences of length ***n*** are there? $| \mathcal{V} |^n$

Assume number of words in the vocabulary is 40000 ($<<$ english vocabulary),

Suppose, each sentence is of length 10

Number of possible sentences $= 1.048576e + 46$

Exponential Growth

Probability Distribution

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

Naïve Approach: Count all sentences in a corpus.

1. How many sentences of length ***n*** are there? $| \mathcal{V} |^n$

Assume number of words in the vocabulary is 40000 ($<<$ english vocabulary),

Suppose, each sentence is of length 10

Number of possible sentences $= 1.048576e + 46$

Exponential Growth

2. What about sentences never seen?

Probability Distribution

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

Naïve Approach: Count all sentences in a corpus.

1. How many sentences of length ***n*** are there? $| \mathcal{V} |^n$

Assume number of words in the vocabulary is 40000 ($<<$ english vocabulary),

Suppose, each sentence is of length 10

Number of possible sentences $= 1.048576e + 46$

Exponential Growth

2. What about sentences never seen?

$$p(w_1, w_2, \ldots, w_n) = \frac{C(w_1, w_2, \ldots, w_n)}{N}$$

Probability Distribution

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

Naïve Approach: Count all sentences in a corpus.

1. How many sentences of length ***n*** are there? $| \mathcal{V} |^n$

Assume number of words in the vocabulary is 40000 ($<<$ english vocabulary),

Suppose, each sentence is of length 10

Number of possible sentences $= 1.048576e + 46$

Exponential Growth

2. What about sentences never seen?

Poor Generalization

$$p(w_1, w_2, \ldots, w_n) = \frac{C(w_1, w_2, \ldots, w_n)}{N}$$

Probability Distribution

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

$$\begin{aligned} p(X_1, X_2, \ldots, X_n) = & \; p(X_n \mid X_{n-1}, \ldots, X_2, X_1) \times p(X_{n-1} \mid X_{n-2}, \ldots, X_2, X_1) \times \\ & \; p(X_{n-2} \mid X_{n-3}, \ldots, X_2, X_1) \times \ldots\ldots \\ & \; p(X_3 \mid X_2, X_1) \times p(X_2 \mid X_1) \times p(X_1) \end{aligned}$$

Probability Distribution: Independence Assumption

$$p(X_1 = w_1, X_2 = w_2, \ldots, X_n = w_n)$$

$$\begin{aligned} p(X_1, X_2, \ldots, X_n) = {} & p(X_n \mid X_{n-1}, \ldots, X_2, X_1) \times p(X_{n-1} \mid X_{n-2}, \ldots, X_2, X_1) \times \\ & p(X_{n-2} \mid X_{n-3}, \ldots, X_2, X_1) \times \ldots\ldots \\ & p(X_3 \mid X_2, X_1) \times p(X_2 \mid X_1) \times p(X_1) \end{aligned}$$

Assume, each word depends on only the previous word.

→ Each word is independent of the entire history given the previous word

Probability Distribution: Independence Assumption

$$p(X_1, X_2, \ldots, X_n) = p(X_n \mid X_{n-1}, \ldots, X_2, X_1) \times p(X_{n-1} \mid X_{n-2}, \ldots, X_2, X_1) \times$$
$$p(X_{n-2} \mid X_{n-3}, \ldots, X_2, X_1) \times \ldots\ldots$$
$$p(X_3 \mid X_2, X_1) \times p(X_2 \mid X_1) \times p(X_1)$$

Assume, each word depends on only the previous word.

→ Each word is independent of the entire history given the previous word

$$p(X_k \mid X_{k-1}, \ldots, X_2, X_1) = p(X_k \mid X_{k-1})$$

First Order Markov Assumption

Markov Assumption

First Order Markov Assumption

$$p(X_k \mid X_{k-1}, \ldots, X_2, X_1) = p(X_k \mid X_{k-1})$$

Markov Assumption

First Order Markov Assumption

$$p(X_k \mid X_{k-1}, \ldots, X_2, X_1) = p(X_k \mid X_{k-1})$$

Second Order Markov Assumption

$$p(X_k \mid X_{k-1}, \ldots, X_2, X_1) = p(X_k \mid X_{k-1}, X_{k-2})$$

Probability Distribution: Independence Assumption

$$p(X_n, X_{n-1}, \ldots, X_1) = p(X_n \mid X_{n-1}, \ldots, X_2, X_1) \times p(X_{n-1} \mid X_{n-2}, \ldots, X_2, X_1) \times$$
$$p(X_{n-2} \mid X_{n-3}, \ldots, X_2, X_1) \times \ldots\ldots$$
$$p(X_3 \mid X_2, X_1) \times p(X_2 \mid X_1) \times p(X_1)$$

Probability Distribution: Independence Assumption

$$p(X_n, X_{n-1}, \ldots, X_1) = p(X_n \mid X_{n-1}, \ldots, X_2, X_1) \times p(X_{n-1} \mid X_{n-2}, \ldots, X_2, X_1) \times$$
$$p(X_{n-2} \mid X_{n-3}, \ldots, X_2, X_1) \times \ldots\ldots$$
$$p(X_3 \mid X_2, X_1) \times p(X_2 \mid X_1) \times p(X_1)$$

$$p(X_n, X_{n-1}, \ldots, X_1) = p(X_n \mid X_{n-1}, X_{n-2}) \times p(X_{n-1} \mid X_{n-2}, X_{n-3}) \times$$
$$\ldots \times p(X_k \mid X_{k-1}, X_{k-2}) \times \ldots\ldots$$
$$p(X_3 \mid X_2, X_1) \times p(X_2 \mid X_1) \times p(X_1)$$

Probability Distribution: Independence Assumption

$$p(X_n, X_{n-1}, \ldots, X_1) = p(X_n \mid X_{n-1}, \ldots, X_2, X_1) \times p(X_{n-1} \mid X_{n-2}, \ldots, X_2, X_1) \times$$
$$p(X_{n-2} \mid X_{n-3}, \ldots, X_2, X_1) \times \ldots\ldots$$
$$p(X_3 \mid X_2, X_1) \times p(X_2 \mid X_1) \times p(X_1)$$

$$p(X_n, X_{n-1}, \ldots, X_1) = p(X_n \mid X_{n-1}, X_{n-2}) \times p(X_{n-1} \mid X_{n-2}, X_{n-3}) \times$$
$$\ldots \times p(X_k \mid X_{k-1}, X_{k-2}) \times \ldots\ldots$$
$$p(X_3 \mid X_2, X_1) \times p(X_2 \mid X_1) \times p(X_1)$$

$$p(X_n, X_{n-1}, \ldots, X_1) = p(X_1) \times p(X_2 \mid X_1) \times \prod_{i=3}^{n} p(X_i \mid X_{i-1}, X_{i-2})$$

Language Model

$$p(X_n, X_{n-1}, \ldots, X_1) = p(X_1) \times p(X_2 \mid X_1) \times \prod_{i=3}^{n} p(X_i \mid X_{i-1}, X_{i-2})$$

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

$where,$

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

N-Gram Model

N-Gram Model

Unigram Model

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

N-Gram Model

Unigram Model

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

Bigram Model

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

N-Gram Model

Unigram Model

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

Bigram Model

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

Trigram Model

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

N-Gram Model

Unigram Model

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

Bigram Model

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

Trigram Model

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

N-gram Model

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

Language Generation

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

$where,$

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

Language Generation

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

$$where,$$

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

STEP 1: Initialize $w_0 = w_{-1} = START, \quad i = 1$

Language Generation

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

$where,$

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

STEP 1: Initialize $w_0 = w_{-1} = START, \quad i = 1$

STEP 2: Sample (Generate) a word: $\quad w_i \sim p(X_i \mid X_{i-1} = w_{i-1}, X_{i-2} = w_{i-2})$

Language Generation

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

$where,$

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

STEP 1: Initialize $w_0 = w_{-1} = START, \quad i = 1$

STEP 2: Sample (Generate) a word: $\quad w_i \sim p(X_i \mid X_{i-1} = w_{i-1}, X_{i-2} = w_{i-2})$

STEP 3: If $\; w_i = STOP$

Language Generation

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

$where,$

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

STEP 1: Initialize $w_0 = w_{-1} = START, \quad i = 1$

STEP 2: Sample (Generate) a word: $\quad w_i \sim p(X_i \mid X_{i-1} = w_{i-1}, X_{i-2} = w_{i-2})$

STEP 3: If $\; w_i = STOP$

Terminate and Return sequence: $\; w_1, w_2, \ldots, w_{i-1}, STOP$

Language Generation

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

$where,$

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

STEP 1: Initialize $w_0 = w_{-1} = START, \quad i = 1$

STEP 2: Sample (Generate) a word: $\quad w_i \sim p(X_i \mid X_{i-1} = w_{i-1}, X_{i-2} = w_{i-2})$

STEP 3: If $\; w_i = STOP$

Terminate and Return sequence: $\; w_1, w_2, \ldots, w_{i-1}, STOP$

else:

Increment i and Go to STEP 2

Language Generation

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

$where,$

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

STEP 1: Initialize $w_0 = w_{-1} = START, \quad i = 1$

STEP 2: Sample (Generate) a word: $\quad w_i \sim p(X_i \mid X_{i-1} = w_{i-1}, X_{i-2} = w_{i-2})$

STEP 3: If $\; w_i = STOP$

Terminate and Return sequence: $\; w_1, w_2, \ldots, w_{i-1}, STOP$

else:

Increment i and Go to STEP 2

Language Model Example

I took the book when the book was available

Language Model Example

I took the book when the book was available

$$
\begin{aligned}
p(\text{I took the book when the book was available}) = {} & p(STOP \mid \text{available, was}) \\
& \times p(\text{available} \mid \text{was, book}) \\
& \times p(\text{was} \mid \text{book, the}) \\
& \times p(\text{book} \mid \text{the, when}) \\
& \times p(\text{the} \mid \text{when, book}) \\
& \times p(\text{when} \mid \text{book, the}) \\
& \times p(\text{book} \mid \text{the, took}) \\
& \times p(\text{the} \mid \text{took, I}) \\
& \times p(\text{took} \mid \text{I}, START) \\
& \times p(\text{I} \mid START, START)
\end{aligned}
$$

Language Model Estimation

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

Language Model Estimation

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

Let, $p(w_i \mid w_{i-1}, w_{i-2}) = \theta(w_i \mid w_{i-1}, w_{i-2})$

$\theta(q \mid r, s)$ is the model parameter corresponding to each trigram $\{s, r, q\}$

Language Model Estimation

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

Trigram Language Model

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

Language Model Estimation

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

Trigram Language Model

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

We would like to estimate $\theta(q \mid r, s)$

Language Model Estimation

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

We would like to estimate $\theta(q \mid r, s)$

Constraints:

$$\theta(q \mid r, s) \geq 0$$

$$\sum_{q \in \mathcal{V} \cup \{STOP\}} \theta(q \mid r, s) = 1$$

Language Model Estimation

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

We would like to estimate $\theta(q \mid r, s)$

Constraints:

$$\theta(q \mid r, s) \geq 0$$

$$\sum_{q \in \mathcal{V} \cup \{STOP\}} \theta(q \mid r, s) = 1$$

How many parameters in the model?

Language Model Estimation

How many parameters in the model?

$$\theta(q \mid r, s)$$

$$q \in \mathcal{V} \cup STOP$$

$$r, s \in \mathcal{V} \cup START$$

Language Model Estimation

How many parameters in the model?

$$\theta(q \mid r, s)$$

$$q \in \mathcal{V} \cup STOP$$

$$r, s \in \mathcal{V} \cup START$$

$$\text{About } \mid \mathcal{V} \mid^3 \text{ parameters}$$

$$\text{If } \mid \mathcal{V} \mid \sim 40000 \implies 6.4 \times 10^{13} \text{ parameters}$$

Language Model Estimation

How many parameters in the model?

$$\theta(q \mid r, s)$$

$$q \in \mathcal{V} \cup STOP$$
$$r, s \in \mathcal{V} \cup START$$

$$\text{About } \mid \mathcal{V} \mid^3 \text{ parameters}$$

$$\text{If } \mid \mathcal{V} \mid \sim 40000 \implies 6.4 \times 10^{13} \text{ parameters}$$

Number of parameters for n-gram mode $\sim \mid \mathcal{V} \mid^n$!!

Maximum Likelihood Estimate (MLE)

$$\text{Given a corpus,}$$

$$\mathcal{S}_1 = \{w_1^{(1)}, w_2^{(1)}, \ldots, w_n^{(1)}\}$$

$$\mathcal{S}_2 = \{w_1^{(2)}, w_2^{(2)}, \ldots, w_n^{(2)}\}$$

$$\vdots$$

$$\mathcal{S}_N = \{w_1^{(N)}, w_2^{(N)}, \ldots, w_n^{(N)}\}$$

Maximum Likelihood Estimate (MLE)

$$\text{Given a corpus,}$$

$$\mathcal{S}_1 = \{w_1^{(1)}, w_2^{(1)}, \ldots, w_n^{(1)}\}$$

$$\mathcal{S}_2 = \{w_1^{(2)}, w_2^{(2)}, \ldots, w_n^{(2)}\}$$

$$\vdots$$

$$\mathcal{S}_N = \{w_1^{(N)}, w_2^{(N)}, \ldots, w_n^{(N)}\}$$

$$p(\mathbf{S} \mid \Theta) = \prod_{k=1}^{N} \prod_{i=1}^{n} \theta(w_i^{(k)} \mid w_{i-1}^{(k)}, w_{i-2}^{(k)})$$

$$\Theta = \{\theta(q \mid r, s), \theta(r \mid s, t), \ldots\}$$

Maximum Likelihood Estimate (MLE)

$$\mathcal{L}(\Theta) = \prod_{k=1}^{N} \prod_{i=1}^{n} \theta(w_i^{(k)} \mid w_{i-1}^{(k)}, w_{i-2}^{(k)})$$

Maximum Likelihood Estimate (MLE)

$$\mathcal{L}(\Theta) = \prod_{k=1}^{N} \prod_{i=1}^{n} \theta(w_i^{(k)} \mid w_{i-1}^{(k)}, w_{i-2}^{(k)})$$

Optimization Problem

$$\underset{\Theta}{\operatorname{argmax}} \prod_{k=1}^{N} \prod_{i=1}^{n} \theta(w_i^{(k)} \mid w_{i-1}^{(k)}, w_{i-2}^{(k)})$$

Constraints

$$\theta(q \mid r, s) \geq 0 \qquad \sum_{q \in \mathcal{V} \cup \{STOP\}} \theta(q \mid r, s) = 1$$

Maximum Likelihood Estimate (MLE)

Optimization Problem $\underset{\Theta}{\text{argmax}} \prod_{k=1}^{N} \prod_{i=1}^{n} \theta(w_i^{(k)} \mid w_{i-1}^{(k)}, w_{i-2}^{(k)})$

Constraints $\theta(q \mid r, s) \geq 0 \quad \sum_{q \in \mathcal{V} \cup \{STOP\}} \theta(q \mid r, s) = 1$

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

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

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

Example

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

I am John STOP
I am out today STOP
John I am STOP
Mary I am STOP
The cat ran STOP
John and cat ran STOP
The cat ran after the mouse STOP

$$\theta(STOP \mid ran, cat) = ?$$
$$\theta(STOP \mid am, I) = ?$$
$$\theta(ran \mid cat, The) = ?$$
$$\theta(The \mid START, START) = ?$$

Example

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

I am John STOP
I am out today STOP
John I am STOP
Mary I am STOP
The cat ran STOP
John and cat ran STOP
The cat ran after the mouse STOP

$$\theta(STOP \mid ran, cat) = \frac{2}{3}$$

$$\theta(STOP \mid am, I) = \frac{2}{4}$$

$$\theta(ran \mid cat, The) = \frac{2}{2}$$

$$\theta(The \mid START, START) = \frac{2}{7}$$

Example

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

I am John STOP
I am out today STOP
John I am STOP
Mary I am STOP
The cat ran STOP
John and cat ran STOP
The cat ran after the mouse STOP

$$\theta(out \mid ran, cat) = ?$$

Example

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

I am John STOP
I am out today STOP
John I am STOP
Mary I am STOP
The cat ran STOP
John and cat ran STOP
The cat ran after the mouse STOP

$$\theta(out \mid ran, cat) = \frac{0}{3}$$

Example

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

I am John STOP
I am out today STOP
John I am STOP
Mary I am STOP
The cat ran STOP
John and cat ran STOP
The cat ran after the mouse STOP

$$p(S = \{John\ ran\}) = ?$$

Example

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

I am John STOP
I am out today STOP
John I am STOP
Mary I am STOP
The cat ran STOP
John and cat ran STOP
The cat ran after the mouse STOP

$$p(S = \{John\ ran\}) = \theta(STOP \mid ran, John) \times \theta(ran \mid John, START) \times \theta(John \mid START, START)$$

Example

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

I am John STOP
I am out today STOP
John I am STOP
Mary I am STOP
The cat ran STOP
John and cat ran STOP
The cat ran after the mouse STOP

$$p(S = \{John\ ran\}) = \underbrace{\theta(STOP \mid ran, John)}_{=indeterminate} \times \underbrace{\theta(ran \mid John, START)}_{=0} \times \underbrace{\theta(John \mid START, START)}_{=\frac{2}{7}}$$

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:

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

Zipf's Law

- Frequency of usage of a word is inversely proportional to its rank in the frequency table.

Zipf's Law

- Frequency of usage of a word is inversely proportional to its rank in the frequency table.
- Most frequent word occurs twice more often than second most frequent word and thrice than the third most frequent word.

Zipf's Law

- Frequency of usage of a word is inversely proportional to its rank in the frequency table.
- Most frequent word occurs twice more often than second most frequent word and thrice than the third most frequent word.

$$f_i \propto \frac{1}{i}$$

$$f_i = ci^k; \quad k = -1$$

$$log(f_i) = k\ log(i) + C$$

Zipf's Law

- Frequency of usage of a word is inversely proportional to its rank in the frequency table.
- Most frequent word occurs twice more often than second most frequent word and thrice than the third most frequent word.

$$f_i \propto \frac{1}{i}$$

$$f_i = ci^k; \quad k = -1$$

$$log(f_i) = k\ log(i) + C$$

- Word usage in languages follow *power law*
- Plot of *log frequency vs log rank* is linear

Summary

- Language Models provide a way to estimate probability of a sentence
- Markov assumptions make it possible to estimate the sentence probability
- Trigram models are trivial but work well in practice
- MLE for LM gives simple way to estimate probabilities in terms of word counts.
- MLE has overfitting problems.

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

- Next week
  - Language Models Smoothing
  - Language Model Evaluation
  - Neural Language Models