# PROBABILISTIC LANGUAGE MODELS

INTRODUCTION TO LANGUAGE MODELS

July 17-21, 2023

Ramaseshan Ramachandran

A Brief Introduction to probability
Probabilistic Language Model - Definition
Chain Rule
Markov Assumption
Target and Context words

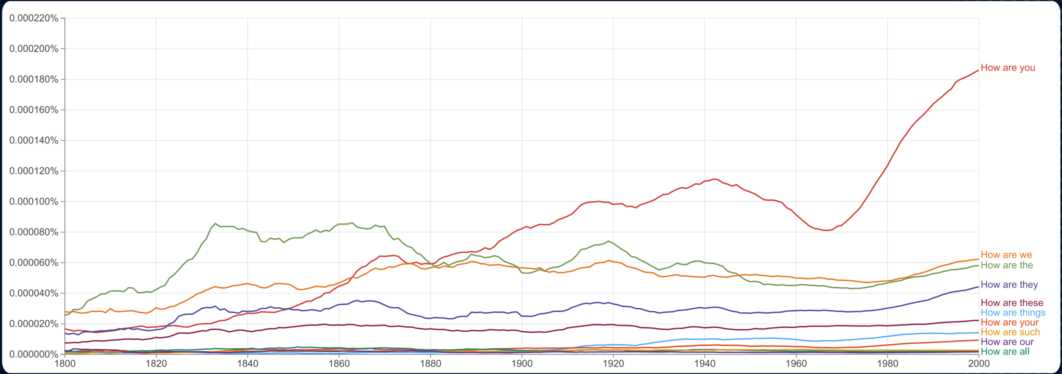
Language Modeling using Unigrams
Generative Model
Maximum Likelihood Estimate
Bigram Language Model
Bigram Language Model - Example
Perplexity
Curse of dimensions

# INTRODUCTION

How are _____? Can you guess the missing word?

# INTRODUCTION

How are ____? Can you guess the missing word?



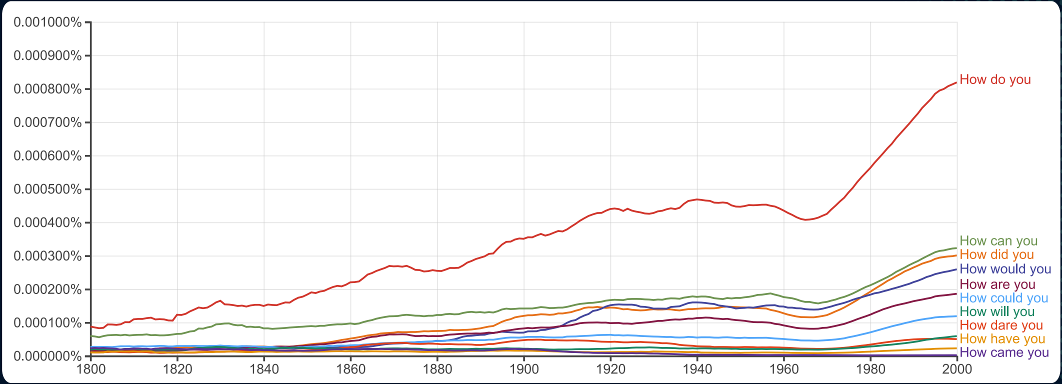

Source:Google NGram Viewer

# INTRODUCTION

How ____ you? Can you guess the missing word?

# INTRODUCTION

How _____ you? Can you guess the missing word?



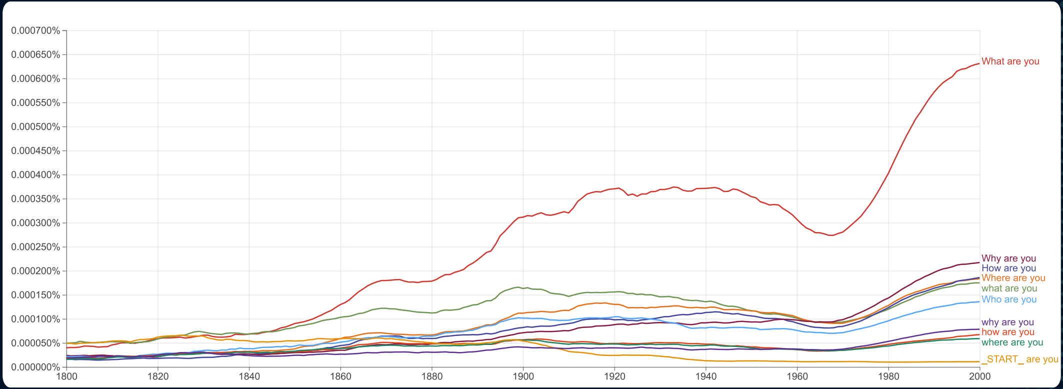

Source:Google NGram Viewer

# INTRODUCTION

_____ are you?

# INTRODUCTION

_____ are you?




Source:Google NGram Viewer

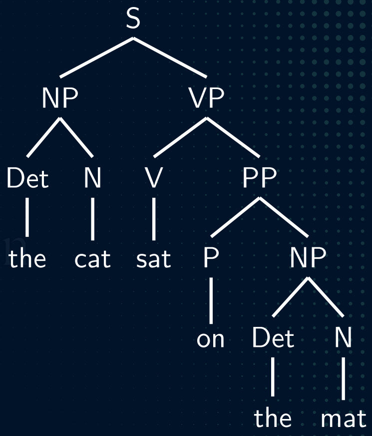
# INTRODUCTION

How do humans predict the next word?

- Domain knowledge
- Syntactic knowledge
- Lexical knowledge
- Knowledge about the sentence structure
- Some words are hard to find. Why?
- Natural language is not deterministic in general
- Some sentences are familiar or had been heard/seen/used several times
- They are more likely to happen than others, hence we could guess

# THE LANGUAGE MODEL

- Natural language sentences can be described by parse trees which use the morphology of words, syntax and semantics
- Probabilistic thinking - finding how likely a sentence occurs or formed, given the word sequence.
- In probabilistic world, the Language model is used to assign a probability $P(W)$ to every possible word sequence $W$.

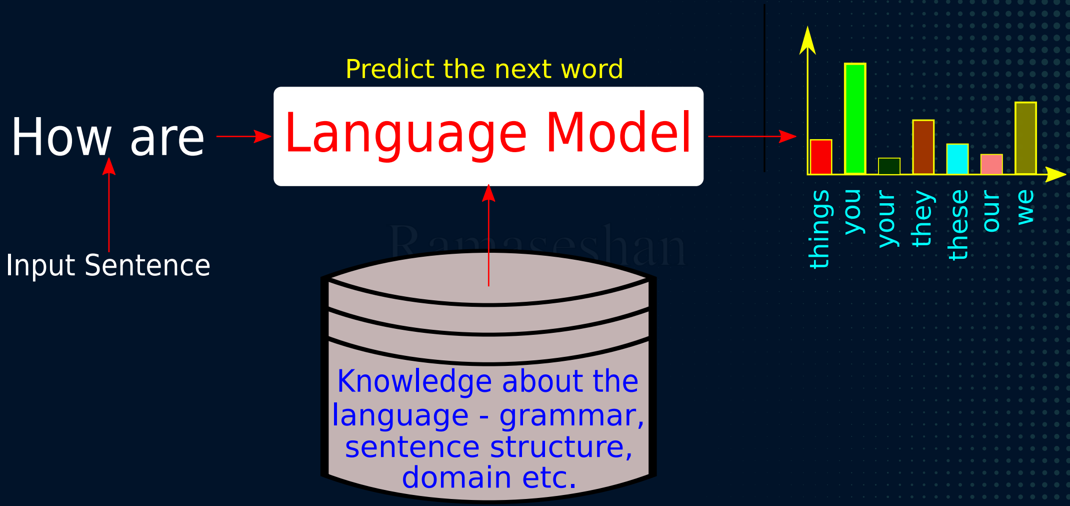





The current research in Language models focuses more on building the model from the huge corpus of text

# APPLICATIONS

| Application | Sample Sentences |
|---|---|
| Speech Recognition | Did you hear ***Recognize speech*** or Wreck a nice beach? |
| Context sensitive Spelling | One upon a ***tie***, ***Their*** lived aking |
| Machine translation | artwork is good $\rightarrow$ l'oeuvre est bonne |
| Sentence Completion | Complete a sentence as the previous word is given - GMail |



Predict the next word
How are
Language Model
Input Sentence
Ramaseshan
Knowledge about the language - grammar, sentence structure, domain etc.
things
you
your
they
these
our
we

# WHY PROBABILISTIC MODEL

- Speech recognition systems cannot depend on the processed speech signals. It may require the help of a language model and context recognizer to convert a speech to correct text format.
- As there are multiple combinations for a word to be in the next slot in a sentence, it is important for language modeling to be probabilistic in nature - judgment about the fluency of a sequence of words returns the probability of the sequence
- The probability of the next word in a sequence is real number $[0, 1]$
- The combination of words with high-probability in a sentence are more likely to occur than low-probability ones
- A probabilistic model continuously estimates the rank of the words in a sequence or phrase or sentence in terms of frequency of occurrence

# FORMAL DEFINITION

Let $\mathscr{V}$ be the vocabulary, a finite set of symbols or words. Let us use $\triangleleft$ and $\triangleright$ as the start and stop symbols and let them be the part of $\mathscr{V}$. Let $|\mathscr{V}|$ denote the size of $\mathscr{V}$.

Let $W$ be infinite sequences of words from the collection of $\mathscr{V}$. Every sequence in $W$ starts with $\triangleleft$ and ends with $\triangleright$. Then a language model is a probability distribution of a random variable $\mathscr{X}$ which takes values from $W$. Or p: $W \rightarrow \mathbb{R}$ such that



$$\forall x \in W, p(x) \geq 0 \text{ and} \tag{1}$$

$$\sum_{x \in W} p(X = x) = 1 \tag{2}$$

# PROBABILISTIC LANGUAGE MODEL

**Goal**: Compute the probability of a sequence of words

$$P(W) = P(w_1, w_2, w_3, ...w_n) \tag{3}$$

**Task**: To predict the next word using probability. Given the context, find the next word using



$$P(w_n | w_1, w_2, w_3, \ldots, w_{n-1}) \tag{4}$$

A model which computes the probability for (3) or predicting the next word (4) or complete the partial sentence is called as Probabilistic Language Model.

The goal is to learn the joint probability function of sequences of words in a language. The probability of $P(\text{The cat roars})$ is less likely to happen than $P(\text{The cat meows})$

# CHAIN RULE

Is it difficult to compute the probability of the entire sequence $P(w_1, w_2, w_3, \ldots, w_n)$?
***Chain rule*** is used to decompose the joint probability of a sequence into a product of conditional probability

$$
\begin{aligned}
P(W) &= P(w_1, w_2, w_3, \ldots, w_n) = P(w_1^n) && (5)\\
&= P(w_1)P(w2|w_1)P(w3|w_2, w_1)\ldots P(w_n|w_{n-1}, w_{n-2}, w_{n-3}, \ldots, w_1) && (6)\\
&= \prod_{k=1}^{n} P(w_k|w_1^{k-1}) && (7)
\end{aligned}
$$



- It is possible to $P(w|h)$, but it does not really help in reducing the computational complexity
- We use innovative ways to string words to form new sentences
- Finding the probability for a long sentence may not yield good outcome as the context may never occur in the corpus
- Short sequences may provide better results

## MARKOV ASSUMPTION

**Markov Assumption**: The future behavior of a dynamic system depends on its recent history and not on the entire history

The product of the conditional probabilities can be written approximately for a bigram as

$$P(w_k|w_1^{k-1}) \approx P(w_k|w_{k-1}) \tag{8}$$

Equation (8) can be generalized for an *n-gram* as

$$P(w_k|w_1^{k-1}) \approx P(w_k|w_{k-K+1}^{k-1}) \tag{9}$$

Now, the joint probability of a sequence can be re-written as

$$\begin{aligned}
P(W) &= P(w_1, w_2, w_3, \ldots, w_n) = P(w_1^n) && (10)\\
&= P(w_1)P(w2|w_1)P(w3|w_2, w_1)\ldots P(w_n|w_{n-1}, w_{n-2}, w_{n-3}, \ldots, w_1) && (11)\\
&= \prod_{k=1}^{n} P(w_k|w_1^{k-1}) && (12)\\
&\approx \prod_{k=1}^{n} P(w_k|w_{k-K+1}^{k-1}) && (13)
\end{aligned}$$

# TARGET AND CONTEXT WORDS

Next word in the sentence depends on its immediate past words, known as context words

$$P(w_{k+1} | \underbrace{w_{i-k}, w_{i-k+1}, \ldots, w_k}_{\text{Context words}})$$



n-grams

- unigram - $P(w_{k+1})$
- bigram - $P(w_{k+1} | w_k)$
- trigram - $P(w_{k+1} | w_{k-1}, w_k)$
- 4-gram - $P(w_{k+1} | w_{k-2}, w_{k-1}, w_k)$

# LANGUAGE MODELING USING UNIGRAMS

- All words are generated independent of its history $W, W_2, W_3, \ldots ..W_n$ and none of them depend on the other
- Not a good model for language generation
- It will have $|V|$ parameters
- $\theta_i = p(w_i) = \frac{c_{w_i}}{N}$, where $c_{w_i}$ if the count of the word $w_i$ and $N$ is the total number of words in the vocabulary
- It may not be able to pick up regularities present tin the corpus
- It is more likely to generate ***the the the the*** as a sentence than a grammatically valid sentence

# GENERATIVE MODEL

- Generates a document containing $N$ words using n-gram
- A good model assigns higher probability to the word that actually occurs

$$P(\mathbf{W}) = P(N)\prod_{i=1}^{N} P(W_i) \tag{14}$$



- The location of the word in the document is not important
- P(N) is the distribution over $N$ and is same for all documents. Hence it may be ignored
- $W_i$, to be estimated in this model is $P(W_i)$ and it must satisfy $\sum_{i=1}^{N} P(w_i) = 1$

# MAXIMUM LIKELIHOOD ESTIMATE

- One of the methods to find the unknown parameter(s) is the use of Maximum Likelihood Estimate
- Estimate the parameter value for which the observed data has the highest probability
- Training data may not have all the words in the vocabulary
- If a sentence with an unknown word is presented, then the MLE is zero.
- Add a smoothing parameter to the equation without affecting the overall probability requirements



$$P(\mathbf{W}) = \frac{C_{w_i} + \alpha}{C_W + \alpha|V|} \tag{15}$$

$$\text{If } \alpha = 1, \text{then it is called as Laplace smoothing} \tag{16}$$

$$P(\mathbf{W}) = \frac{C_{w_i} + 1}{C_W + |V|} \tag{17}$$

# BIGRAM LANGUAGE MODEL

- This model generates a sequence one word at a time, starting with the first word and then generating each succeeding word conditioned on the previous one or its predecessor
- A bigram language model or the Markov model (first order)is defined as follows:



$$P(\mathbf{W}) = \prod_{i=1}^{n+1} P(w_i|w_{i-1}) \tag{18}$$

where $\mathbf{W} = w_1, w_2, w_3, \ldots, w_n$

# BIGRAM LANGUAGE MODEL

- Estimate the parameter $P(w_i|w_{i-1})$ for all bigrams
- The parameter estimation does not depend on the location of the word
- If we consider the sentence as a sequence in time, they are time-invariant MLE picks up the word that is $\frac{n_{w,w'}}{n_{w,o}}$ where $nw, w'$ is the number of times the words $w_1, w'$ occur together and $n_{w,o}$ is the number of times the word $w$ appears in the bigram sequence with any other word
- The number of parameters to be estimated $= |V| \times (|V| + 1)$

# PROBABILISTIC LANGUAGE MODEL - EXAMPLE

Peter Piper picked a peck of pickled peppers
A peck of pickled peppers Peter Piper picked
If Peter Piper picked a peck of pickled peppers
Where's the peck of pickled peppers Peter Piper picked?

—

The joint probability of a sentence formed with $n$ words can be expressed as a product conditional probabilities - we use immediate context and not the entire history

$P(w_1|\langle\triangleleft\rangle) \times P(w_2|w_1) \times \ldots P(\langle E\rangle|w_n)$

and $P(w_{i+1}|w_i) = \frac{C(w_i.w_{i+1})}{C(w_i)}$

—

What is the probability of these sentences?
P(Peter Piper picked)
P(Peter Piper picked peppers)

| **Bigram** | **Frequency** |
|---|---|
| $\triangleleft$peter | 1 |
| peter piper | 4 |
| piper picked | 4 |
| picked a | 2 |
| a peck | 2 |
| peck of | 4 |
| pickled peppers | 4 |
| peppers $\triangleright$ | 1 |
| $\triangleleft$a | 1 |
| a peck | 1 |
| peck of | 1 |
| of pickled | 4 |
| peppers peter | 2 |
| ... | .. |
| $\triangleleft$... | 1 |

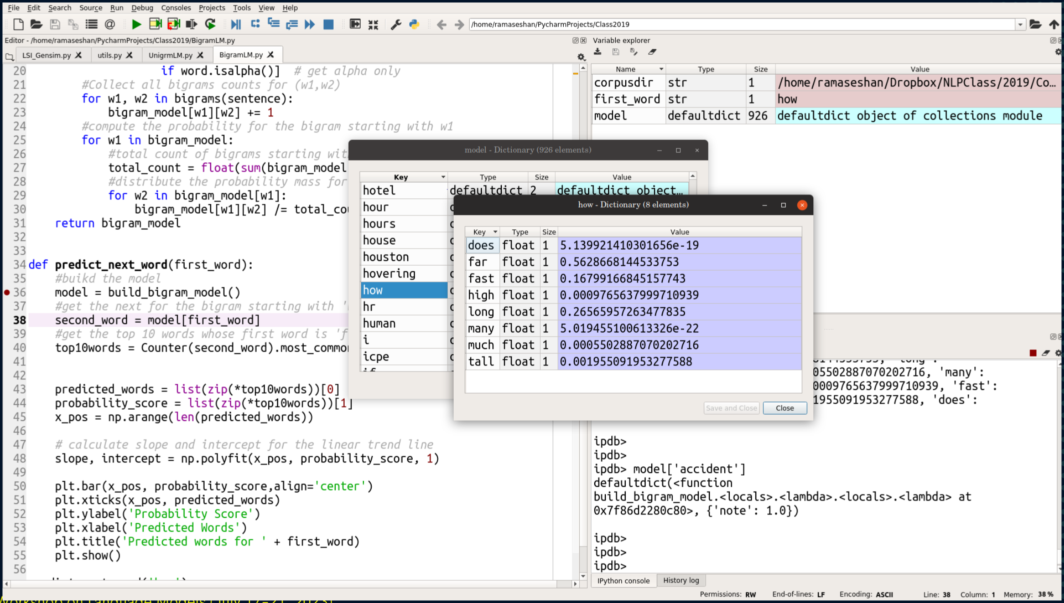
# BUILDING A BIGRAM MODEL - CODE

```python
#compute the bigram model
def build_bigram_model():
    bigram_model = collections.defaultdict(
        lambda: collections.defaultdict(lambda: 0))
    for sentence in kinematics_corpus.sents():
        sentence = [word.lower() for word in sentence
                    if word.isalpha()]  # get alpha only
        #Collect all bigrams counts for (w1,w2)
        for w1, w2 in bigrams(sentence):
            bigram_model[w1][w2] += 1
        #compute the probability for the bigram containing w1
        for w1 in bigram_model:
            #total count of bigrams conaining w1
            total_count = float(sum(bigram_model[w1].values()))
            #distribute the probability mass for all bigrams starting with w1
            for w2 in bigram_model[w1]:
                bigram_model[w1][w2] /= total_count
    return bigram_model
```

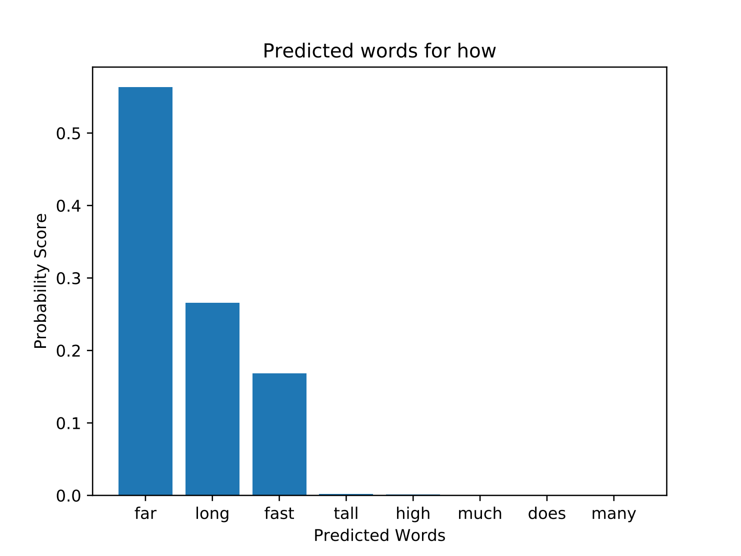
# BUILDING A BIGRAM MODEL - CODE

```python
def predict_next_word(first_word):
    #build the model
    model = build_bigram_model()
    #get the next for the bigram starting with 'word'
    second_word = model[first_word]
    #get the top 10 words whose first word is 'first_word'
    top10words = Counter(second_word).most_common(10)

    predicted_words = list(zip(*top10words))[0]
    probability_score = list(zip(*top10words))[1]
    x_pos = np.arange(len(predicted_words))

    plt.bar(x_pos, probability_score,align='center')
    plt.xticks(x_pos, predicted_words)
    plt.ylabel('Probability Score')
    plt.xlabel('Predicted Words')
    plt.title('Predicted words for ' + first_word)
    plt.show()

predict_next_word('how')
```

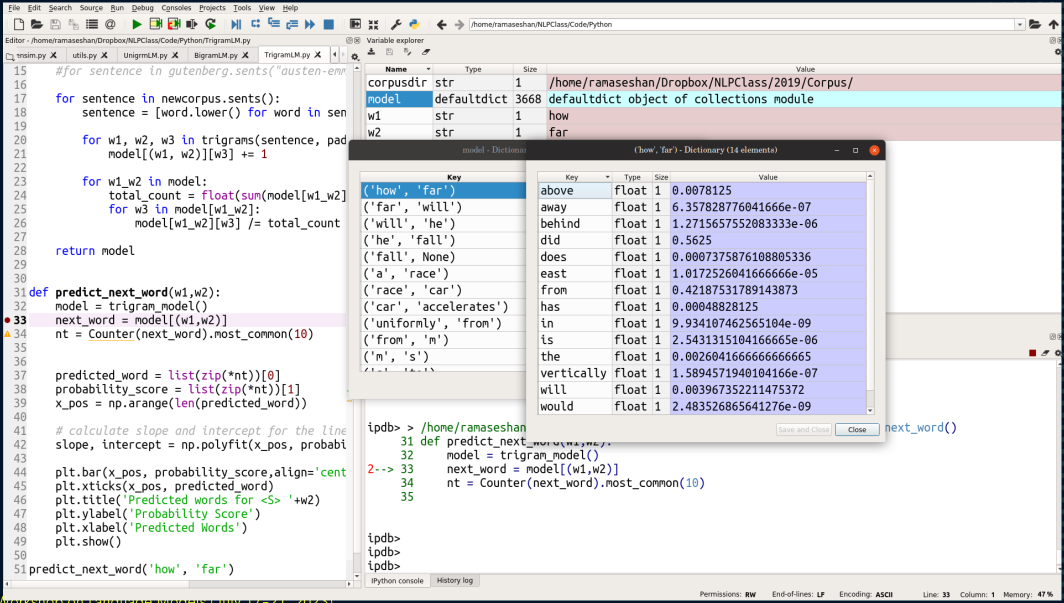
# MODEL PARAMETERS - BIGRAM EXAMPLE



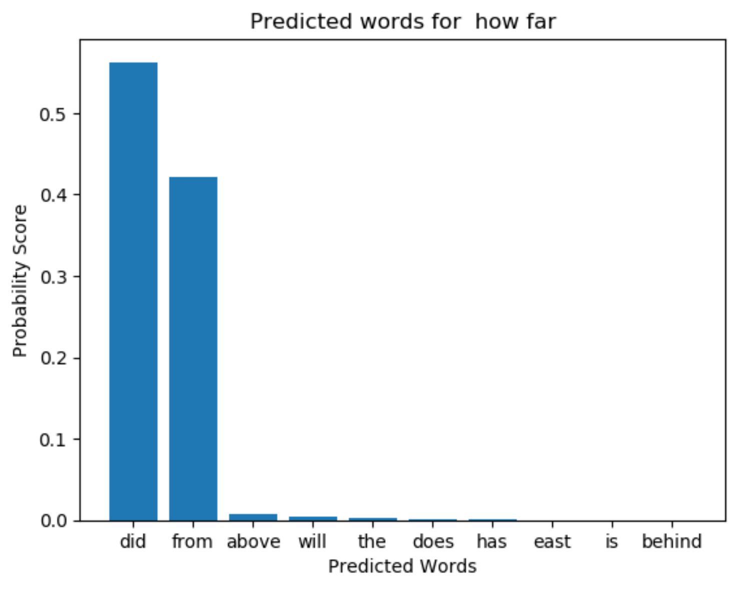
# BIGRAM MODEL - NEXT WORD PREDICTION



# MODEL PARAMETERS - TRIGRAM EXAMPLE



# TRIGRAM MODEL - NEXT WORD PREDICTION



# PERPLEXITY

Perplexity is a measurement of how well a probability model predicts a sample. Perplexity is defined as

$$\text{For bigram model, } PP(W_N) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}} \tag{19}$$



$$\text{For trigram model } PP(W_N) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1}w_{i-2})}} \tag{20}$$

A good model gives maximum probability to a sentence or minimum perplexity to a sentence

# UNKNOWN WORDS

- In a closed vocabulary language model, there is no unknown words or ***out of vocabulary words (OOV)***
- In an open vocabulary system, you will find new words that are not present in the trained model
- Pick words below certain frequency and replace them as OOV.
- Treat every OOV as a regular word
- During testing, the new words would be treated as OOV and the corresponding frequency will be used for computation
- This eliminates zero probability for sentences containing OOV

# CURSE OF DIMENSIONALITY

- A fundamental problem that makes language modeling and other learning problems difficult is the curse of dimensionality
- It is particularly obvious in the case when one wants to model the joint distribution between many discrete random variable
- If one wants to estimate the joint probability distribution of $10$ words in a language with a million words as vocabulary, then we need to estimate $10000000^9 \cdot (1000000 - 1) \approx 10^{60}$ parameters