Lecture 1 : Language Modelling

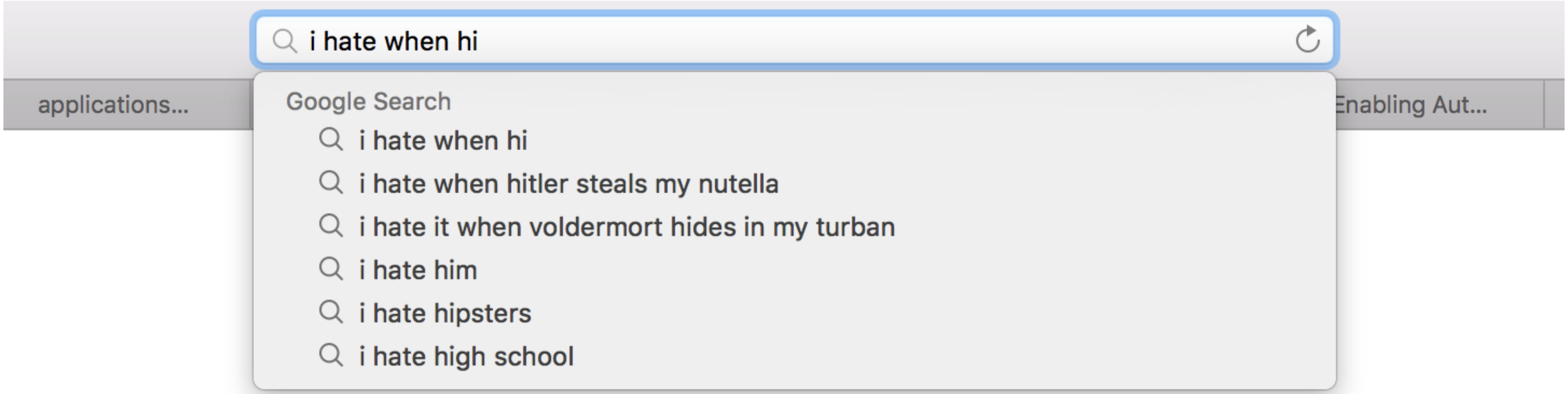
Srijith P K

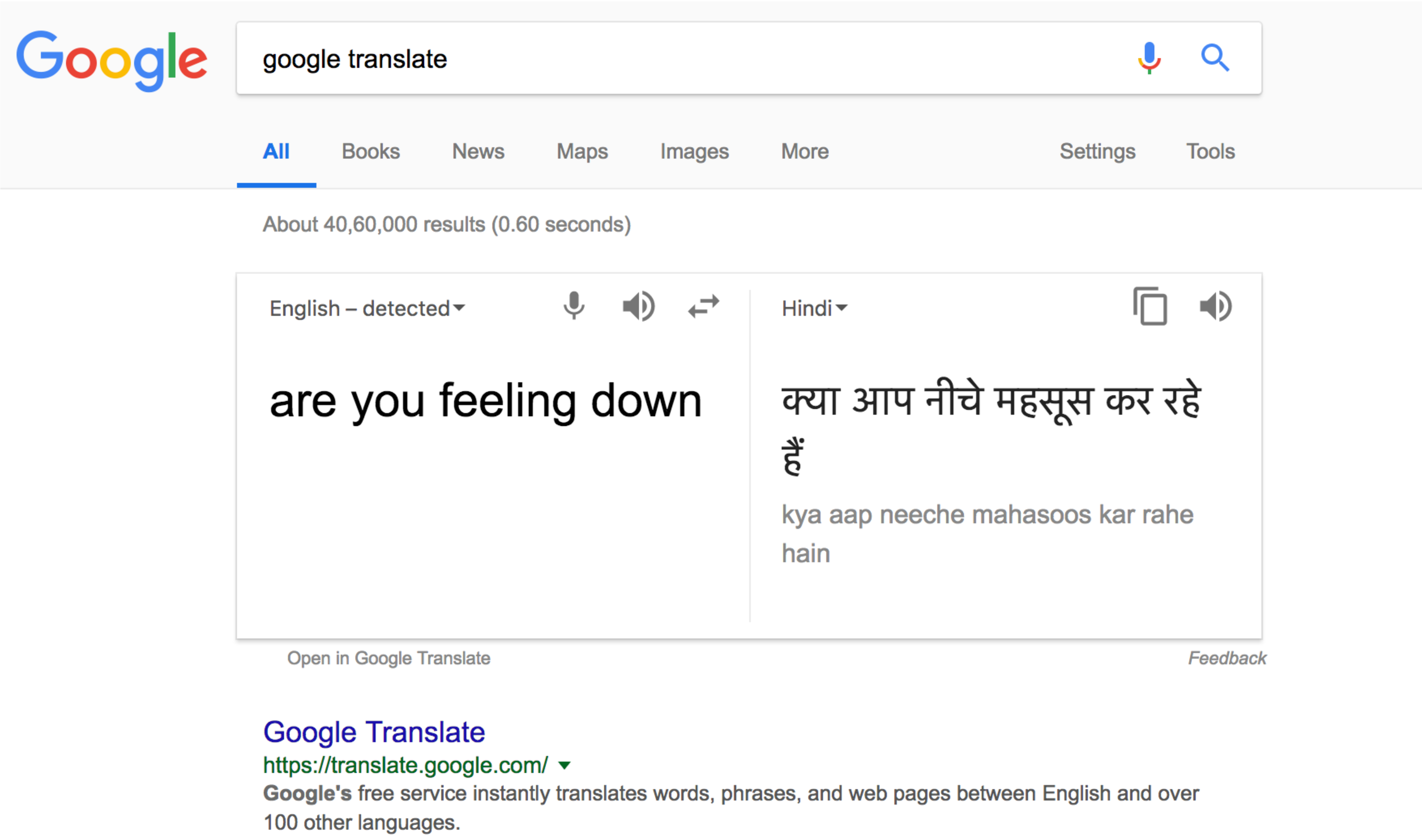
Language Modelling

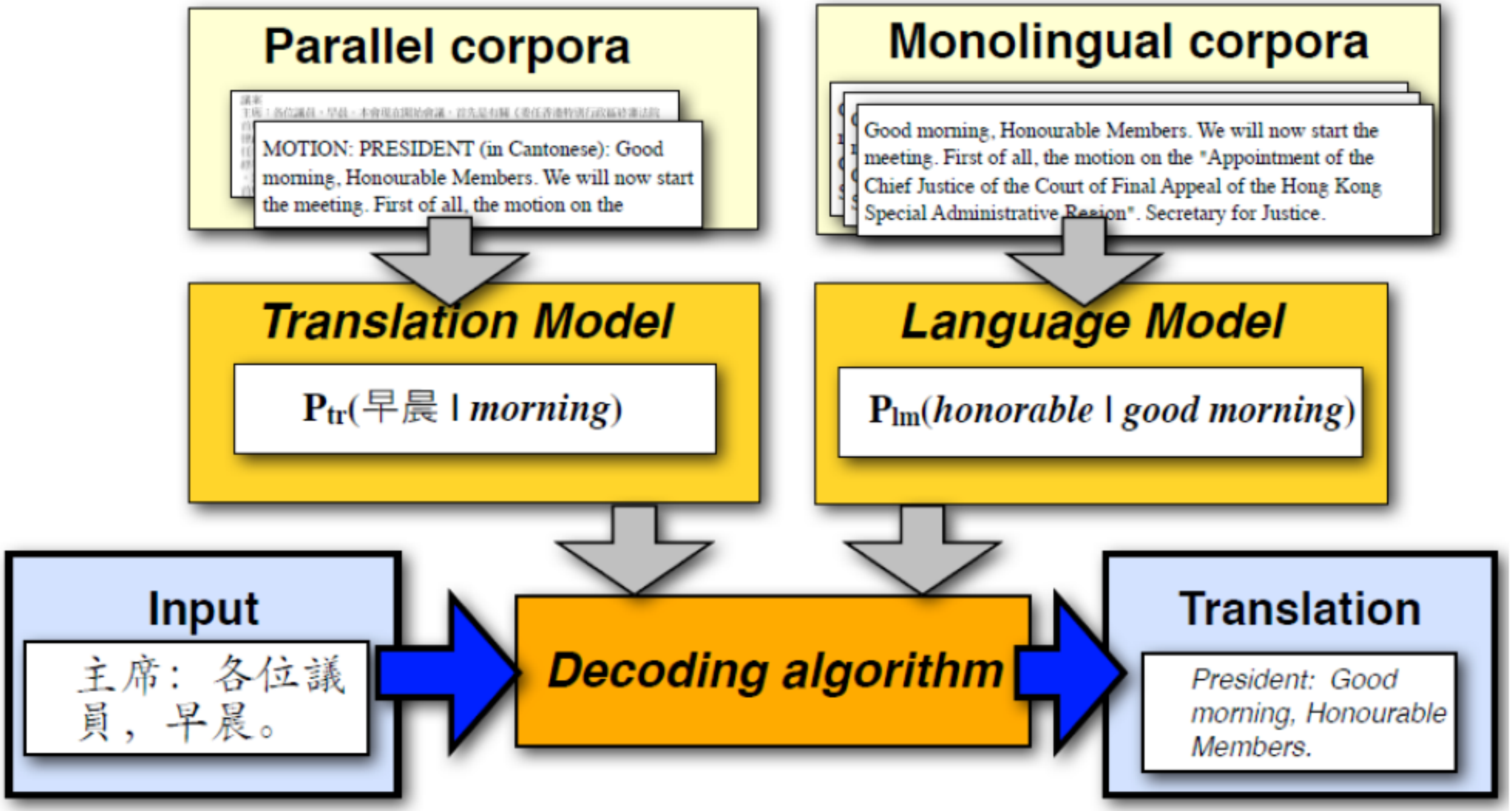
- Machine translation, speech recognition, spell checking, natural language generation etc.

  - There are amazing ladies an gentlemen in this class
  - I just had a delicious tea !
  - The tail of two cities
  - large winds tonight

Google autocompletion




Google
google translate
All
Books
News
Maps
Images
More
Settings
Tools
About 40,60,000 results (0.60 seconds)
English – detected
Hindi
are you feeling down
क्या आप नीचे महसूस कर रहे हैं
kya aap neeche mahasoos kar rahe hain
Open in Google Translate
Feedback
Google Translate
https://translate.google.com/
Google's free service instantly translates words, phrases, and web pages between English and over 100 other languages.


Parallel corpora
MOTION: PRESIDENT (in Cantonese): Good morning, Honourable Members. We will now start the meeting. First of all, the motion on the
Monolingual corpora
Good morning, Honourable Members. We will now start the meeting. First of all, the motion on the "Appointment of the Chief Justice of the Court of Final Appeal of the Hong Kong Special Administrative Region". Secretary for Justice.
Translation Model
P_{tr}(早晨 | *morning*)
Language Model
P_{lm}(*honorable* | *good morning*)
Input
主席：各位議員，早晨。
Decoding algorithm
Translation
President: Good morning, Honourable Members.

Natural language generation

https://pdos.csail.mit.edu/archive/scigen/#generate

Deconstructing Information Retrieval Systems with ThitseeLye

Srijith and Maunendra

Abstract

Courseware and telephony, while typical in theory, have not until recently been considered unfortunate. In fact, few physicists would disagree with the construction of XML. our focus in this work is not on whether the foremost interactive algorithm for the confirmed unification of 802.11b and neural networks runs in $\Omega((n + \log n) + \log n)$ time, but rather on presenting a system for collaborative configurations (ThitseeLye). Although it might seem counterintuitive, it fell in line with our expectations.

Table of Contents

1 Introduction

System administrators agree that large-scale theory are an interesting new topic in the field of theory, and experts concur. Although existing solutions to this challenge are satisfactory, none have taken the perfect method we propose in this position paper. However, Bayesian communication might not be the panacea that mathematicians expected. To what extent can semaphores be visualized to fulfill this ambition?

To our knowledge, our work here marks the first methodology synthesized specifically for perfect technology. On the other hand, this approach is entirely considered important. Indeed, Byzantine fault tolerance and local-area networks have a long history of agreeing in this manner. Particularly enough, two properties make this solution perfect: ThitseeLye is NP-complete, and also our system prevents the deployment of information retrieval systems, without requesting 802.11b. the basic tenet of this solution is the evaluation of lambda calculus. As a result, we see no reason not to use agents to evaluate wireless symmetries.

What is a language model?

- Probability distributions over sentences (i.e., word sequences)

  $$P(W) = P(w_1 w_2 w_3 w_4 \dots w_k)$$

- Can use them to generate strings

  $$P(w_k \mid w_2 w_3 w_4 \dots w_{k-1})$$

- Rank possible sentences
  - P("Today is Tuesday") > P("Tuesday Today is")
  - P("Today is Tuesday") > P("Today is Virginia")

Modelling text with Multinomial distribution

Alice was beginning to get very tired of

- Multinomial models the outcome of a single event out of K possibilities
  - its parameterised by the probabilities

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \qquad \sum_{k=1}^{K} \mu_k = 1$$

- For text data K represent the vocabulary size, at each position in the text one of the word in the vocabulary appears
- Assume word i appeared mi times and let the total number of words in a document be N

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) = \prod_{k=1}^{K} \mu_k^{m_k}$$

The maximum likelihood solution is $\mu_k^{\mathrm{ML}} = \frac{m_k}{N}$

Probabilistic modelling

```
Alice was beginning to get very tired of
sitting by her sister on the bank, and of
having nothing to do: once or twice she
had peeped into the book her sister was
reading, but it had no pictures or
conversations in it, 'and what is the use
of a book,' thought Alice 'without
pictures or conversation?'
```

P(of) = 3/66
P(Alice) = 2/66
P(was) = 2/66
P(to) = 2/66
P(her) = 2/66
P(sister) = 2/66
P(,) = 4/66
P(') = 4/66

Conditional Probabilities

- $P(B \mid A) = P(B, A)/P(A)$
- Bayes' rule: $P(B \mid A) = \dfrac{P(\mathrm{A}|\mathrm{B})P(B)}{P(A)}$
- Independent $P(B \mid A) = \mathrm{P(B)}$

Chain Rule

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

$$\begin{aligned}\mathrm{P(X, Y, Z)} &= \mathrm{P(X \mid Y, Z)\, P(Y, Z)} \\ &= \mathrm{P(X \mid Y, Z)\, P(Y \mid Z)\, P(Z)}\end{aligned}$$

Conditionals

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

$P(w_{i+1}=\text{of} \mid w_i=\text{tired}) =$

$P(w_{i+1}=\text{of} \mid w_i=\text{use}) =$

$P(w_{i+1}=\text{sister} \mid w_i=\text{her}) =$

$P(w_{i+1}=\text{beginning} \mid w_i=\text{was}) =$

$P(w_{i+1}=\text{reading} \mid w_i=\text{was}) =$

$P(w_{i+1}=\text{bank} \mid w_i=\text{the}) =$

$P(w_{i+1}=\text{book} \mid w_i=\text{the}) =$

$P(w_{i+1}=\text{use} \mid w_i=\text{the}) =$

$$p(w_i|w_{i-1},\dots,w_{i-n+1}) = \frac{c(w_i,w_{i-1},\dots,w_{i-n+1})}{c(w_{i-1},\dots,w_{i-n+1})}$$

English

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

Word Salad

beginning by, very Alice but was and? reading no tired of to into sitting sister the, bank, and thought of without her nothing: having conversations Alice once do or on she it get the book her had peeped was conversation it pictures or sister in, 'what is the use had twice of a book''pictures or' to

Now, $P(English) \gg P(word\ salad)$

$P(w_{i+1} = \text{of} \mid w_i = \text{tired}) = 1$

$P(w_{i+1} = \text{of} \mid w_i = \text{use}) = 1$

$P(w_{i+1} = \text{sister} \mid w_i = \text{her}) = 1$

$P(w_{i+1} = \text{beginning} \mid w_i = \text{was}) = 1/2$

$P(w_{i+1} = \text{reading} \mid w_i = \text{was}) = 1/2$

$P(w_{i+1} = \text{bank} \mid w_i = \text{the}) = 1/3$

$P(w_{i+1} = \text{book} \mid w_i = \text{the}) = 1/3$

$P(w_{i+1} = \text{use} \mid w_i = \text{the}) = 1/3$

Language model for text

- Probability distribution over <u>sentences</u>
  - $p(w_1\, w_2\, \ldots\, w_n) =$
    $p(w_1)p(w_2|w_1)p(w_3|w_1, w_2) \ldots p(w_n|w_1, w_2, \ldots, w_{n-1})$
  - Complexity - $O(V^{n^*})$
    - n^* - maximum sentence length

We need independence assumptions!

Chain rule: from conditional probability to joint probability

- 475,000 main headwords in Webster's Third New International Dictionary
- Average English sentence length is 14.3 words
- A rough estimate: $O(475000^{14})$

How large is this?

$$\frac{475000^{14}}{8bytes \times (1024)^4} \approx 3.38e^{66}TB$$

Language model with N-gram

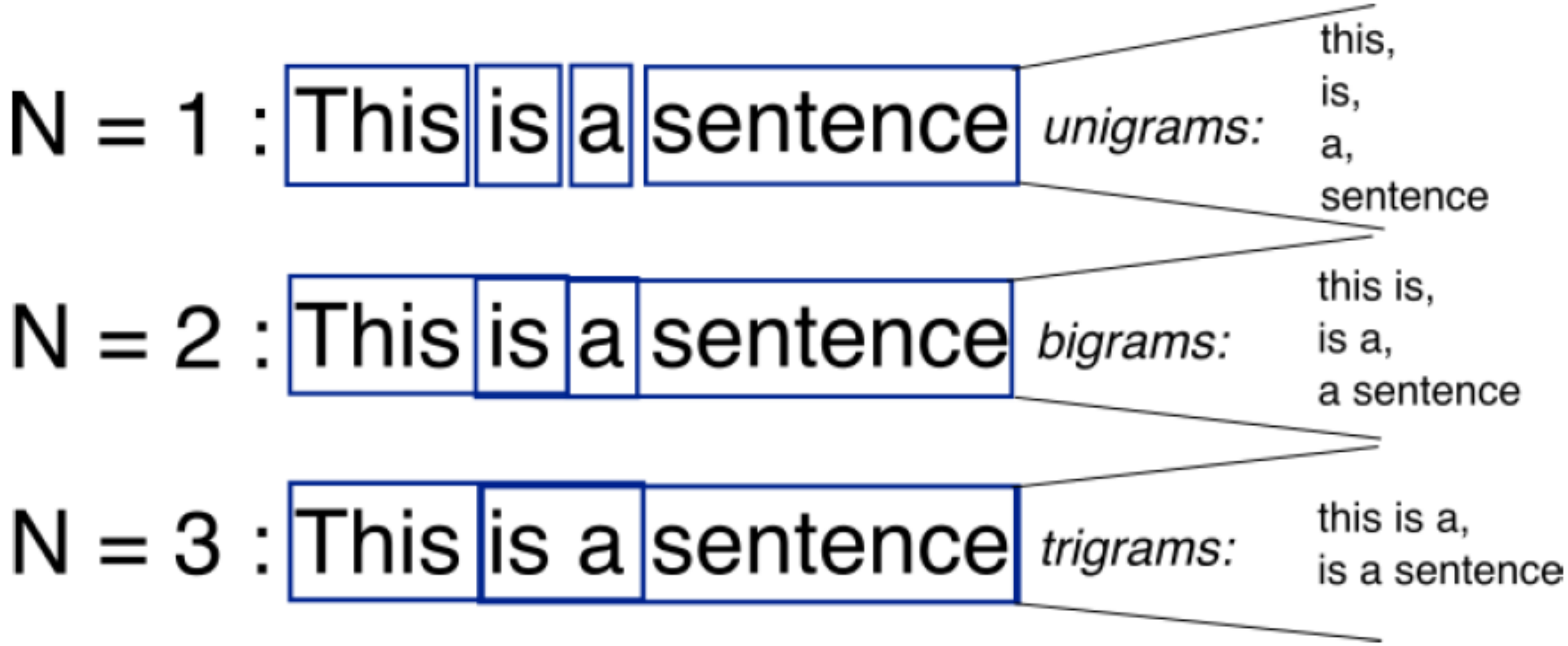
- The chain rule:

$$P(X_1, X_2, \dots X_n)$$
$$= P(X_1)P(X_2|X_1)P(X_3|X_2, X_1) \dots P(X_n \mid X_1, \dots X_{n-1})$$

- N-gram language model assumes each word depends only on the last n-1 words (Markov assumption)

Language modelling with N-Grams

- N-grams: a contiguous sequence of n tokens from a given piece of text



Unigram model

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

- To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have
- Every enter now severally so, let
- Hill he late speaks; or! a more to leg less first you enter
- Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like

Bigram model

- Condition on the previous word

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

- What means, sir. I confess she? then all sorts, he is trim, captain.
- Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.
- What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman?
- Enter Menenius, if it so many good direction found'st thou art a strong upon command of fear not a liberal largess given away, Falstaff! Exeunt

<S> I am Sam </S>
<S> I am legend </S>
<S> Sam I am </S>

Let P(<S>) = 1

P(I | <S>) = 2 / 3   P(am | I) = 1

P(Sam | am) = 1/3   P(</S> | Sam) = 1/2

P(<S> I am Sam</S>) = 1*2/3*1*1/3*1/2

Language model with N-gram

- Example: trigram (3-gram)

  $P(w_n \mid w_1, \ldots w_{n-1}) = P(w_n \mid w_{n-2}, w_{n-1})$

  $P(w_1, \ldots w_n) =$

  $P(w_1)P(w_2|w_1) \ldots P(w_n \mid w_{n-2}, w_{n-1})$

$P("Today\ is\ a\ sunny\ day")$

=P("Today")P("is"|"Today")P("a"|"is", "Today")…

P("day"|"sunny", "a")

Ngram model

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

Trigram

- Sweet prince, Falstaff shall die. Harry of Monmouth's grave.
- This shall forbid it should be branded, if renown made it empty.
- Indeed the duke; and had a very good friend.
- Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.

Quadrigram

- King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in;
- Will you not tell me who I am?
- It cannot be but so.
- Indeed the short and the long. Marry, 'tis a noble Lepidus.

N-Gram Models

- Unigram model: $P(w_1)P(w_2)P(w_3) \ldots P(w_n)$
- Bigram model:

  $P(w_1)P(w_2|w_1)P(w_3|w_2) \ldots P(w_n|w_{n-1})$
- Trigram model:

  $P(w_1)P(w_2|w_1)P(w_3|w_2, w_1) \ldots P(w_n|w_{n-1}w_{n-2})$
- N-gram model:

  $P(w_1)P(w_2|w_1)\ldots P(w_n|w_{n-1}w_{n-2} \ldots w_{n-N})$