CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computer Science

http://www.cit.ie

Natural Language Processing

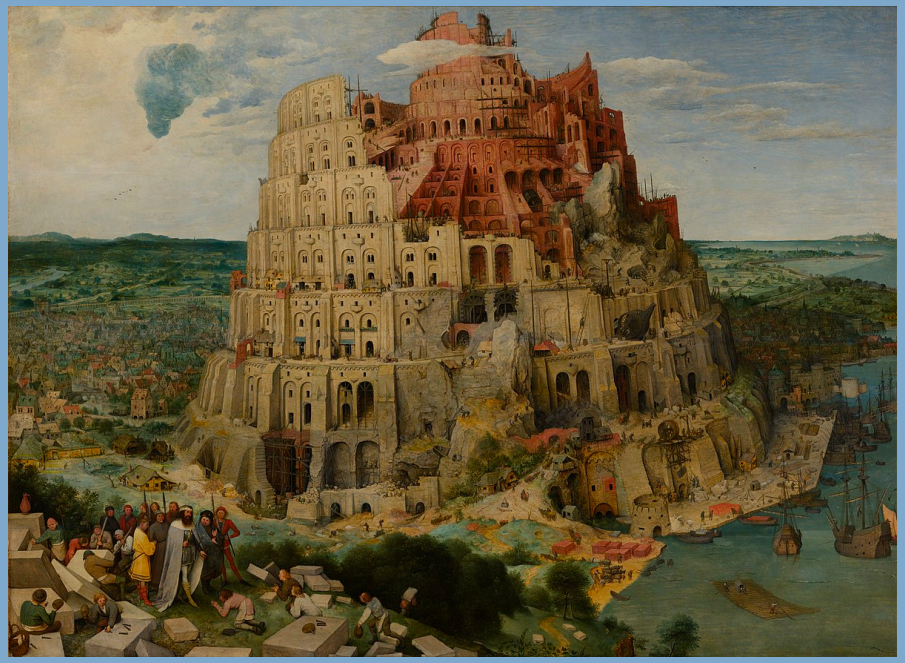
Week3: Statistical Language Modelling

Dr. Haithem Afli

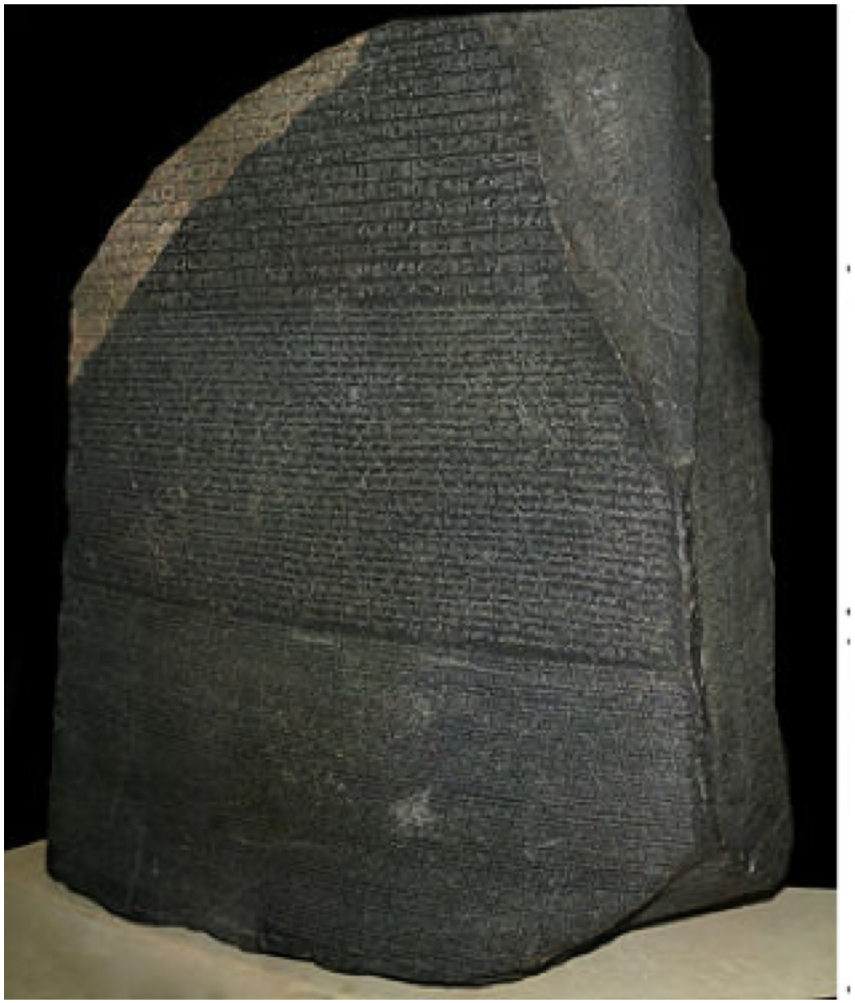
Haithem. afli@cit.ie

@AfliHaithem

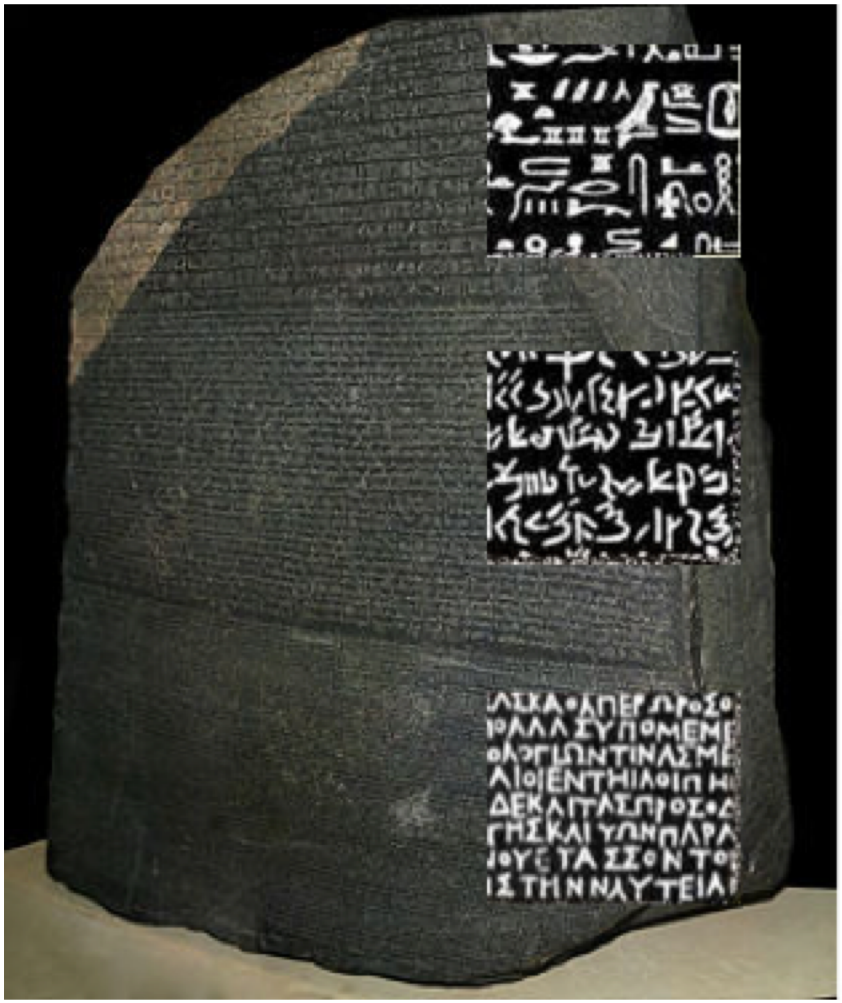
2020/2021



If you think the language industry is new



If you think the language industry is new, think again!



Rosetta Stone (British Museum)

Natural Language : An age-old industry ?

- For as far back as we can see, human has needed to communicate → so the origin of language industry is closely intertwined with the need of communication itself

The Tower of Babel and The House of Wisdom in Bagdad (Bait-al-Hikma)

Content

1. **What is a Language Model?**
2. N-gram Language Model
3. Smoothing
4. Evaluation
5. Management of Large Language Models
6. Other Approaches

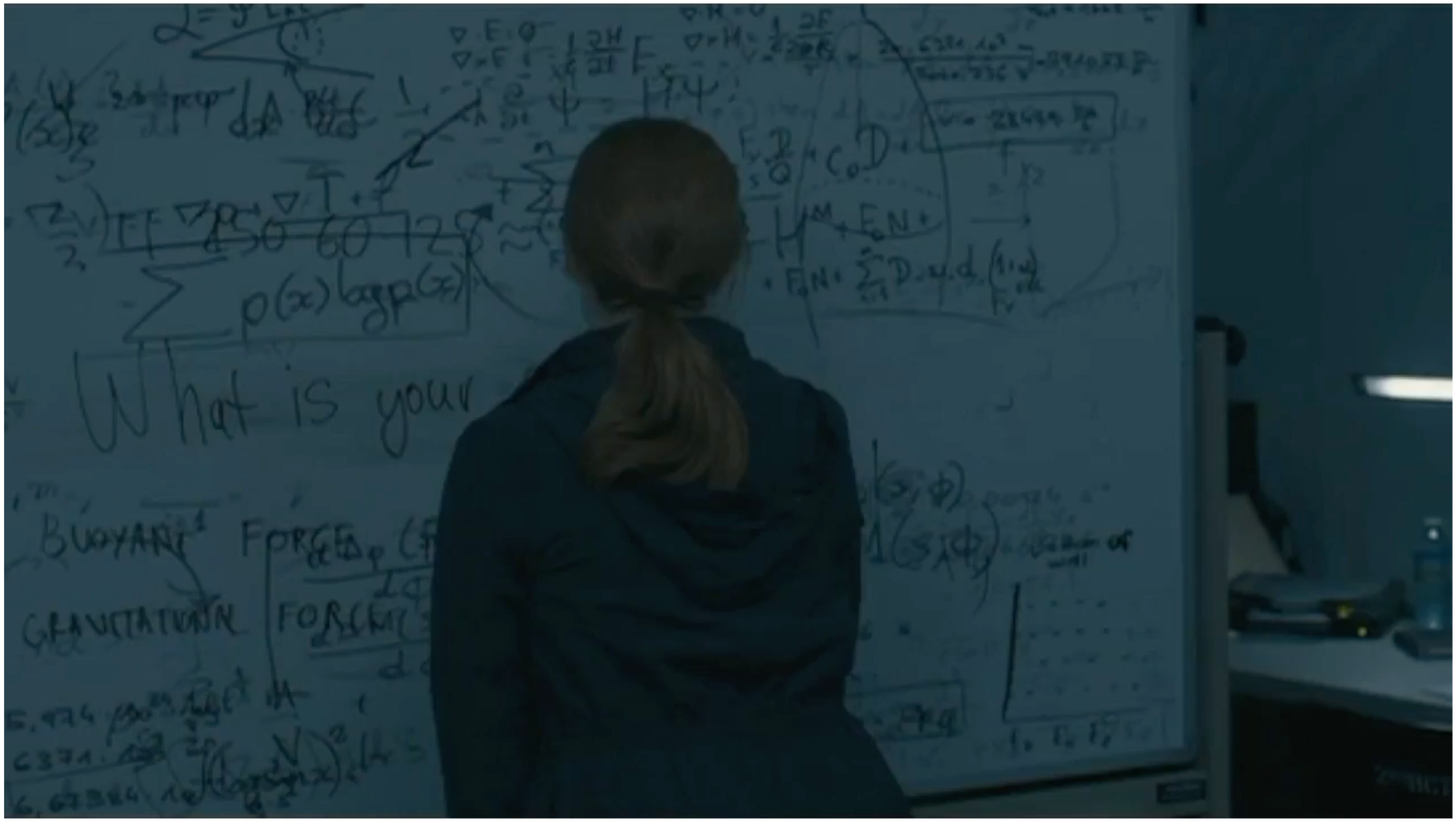
What is a language?

Can we define a language mathematically?

Deterministic Definition:
A language is the set of all the sentences we can say.

Probabilistic Definition:
A language is the probabilistic distribution of all possible sentences.

What is your purpose on Earth ?
Arrival (2016)



https://www.youtube.com/watch?v=bIuMmAXz8PM&ab_channel=SamHiggs

Formal Language

- Languages may contain infinite sentences
  - they cannot be defined by enumeration
- One way to define a formal language
  - define the grammar to generate all the sentences in that language
- Another way
  - build a machine to recognise if a sentence is valid in that language

Deterministic Definition

- Given a vocabulary $V = \{v_1, \dots, v_n\}$,
- We use V^* to denote the set of any sequence of words of V,

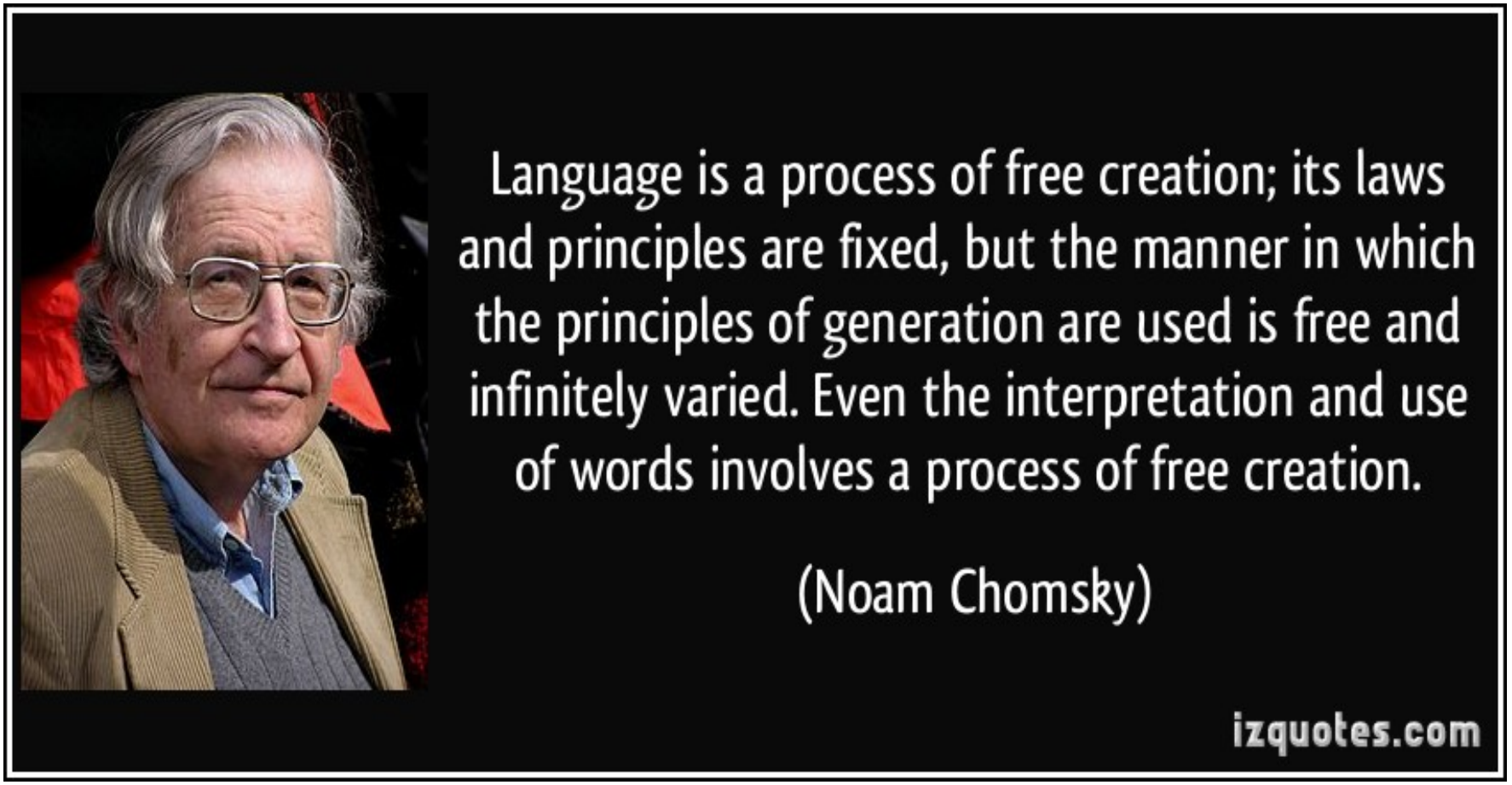
$$V^* = \{W | W = w_1, \dots, w_l\}, l \geq 1, w_i \in V$$

- Then a language can be defined as a subset of sentences:

$$L \subseteq V^*$$

- Any element of L is called a *sentence* in the language.

Chomsky Hierarchy



Probabilistic Definition

- Deterministic definition is not suitable for natural language in most cases
- For some sentences, not easy to say if it is a legal sentence in a certain language:
  - Child languages
  - Tweets
  - Slang
  - …

What is a language?

Can we define a language mathematically?

Deterministic Definition:
A language is the set of all the sentences we can say.

Probabilistic Definition:
A language is the probabilistic distribution of all possible sentences

Hard vs soft decision

Statistical Language Model

Defined as

$$p(s = w_1 \ldots w_n), \forall i: w_i \in V$$

The normalization condition

$$\sum_{s \in V^*} p(s) = 1$$

Statistical Language Model

- How can we estimate the probability of a sentence in a specific language?
- Unlike estimating the probability distribution of a dice, we cannot exhaust all the possible sentences in limited samples

Statistical Language Model

- How can we estimate the probability of a sentence in a specific language?
- Unlike estimating the probability distribution of a dice, we cannot exhaust all the possible sentences in limited sample

- Idea
  - break all sentences down to limited substrings (n-grams)
  - Estimate the probability of a sentence by these substrings
    - If a sentence has many plausible substrings then it might be a reasonable sentence

Simplest Language Model

- Simplest way to break down a sentence
  - split it to words
- Thus, the simplest language model

$$p(s = w_1 \ldots w_n) = \prod_{i=1}^{n} p(w_i)$$

- Here the probability of a sentence is just the multiplication of the probability of the words in the sentence
- This model is called **unigram language model**

Word Frequency

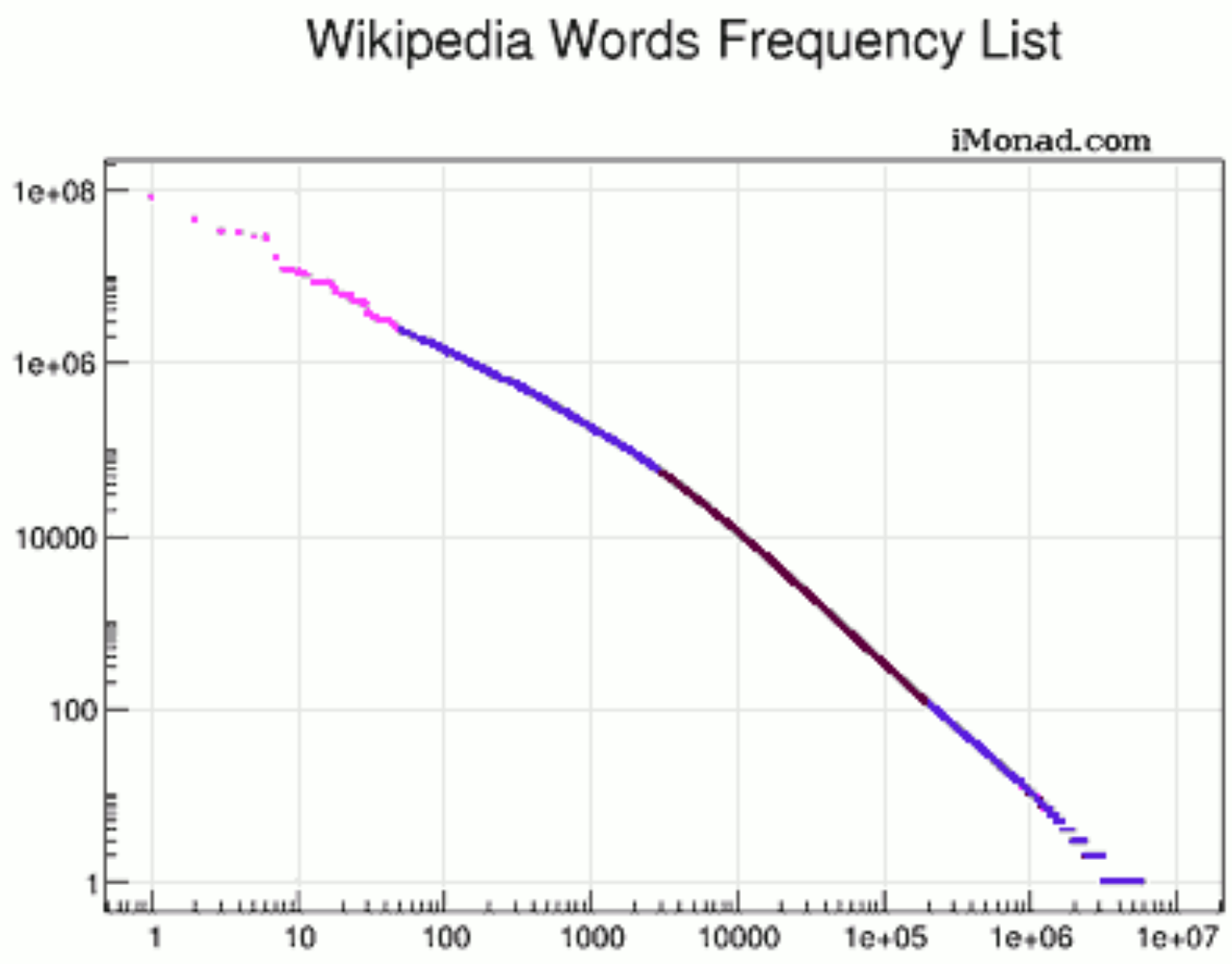
- *p(w)* is word frequency

$$p(w) = \frac{occurrences\ of\ w}{number\ of\ tokens}$$

Type	Occurrences	Rank
the	3789654	1st
he	2098762	2nd
[...]		
king	57897	1,356th
boy	56975	1,357th
[...]		
stringyfy	5	34,589th
[...]		
transducionalify	1	123,567th

Word Frequency

Word Frequency



Zipf's law

The frequency of any word is inversely proportional to its rank in the frequency table

Thus the **most frequent** word will occur approximately **twice** as often as **the second most frequent** word, **three times** as often as the **third most frequent** word, etc

Unigram Language Model

$$p(\text{"I am a student."}) = p(\text{"I"}) \times p(\text{"am"}) \times p(\text{"a"}) \times p(\text{"student"}) \times p(\text{"."})$$

- Problems of unigram language model
  - Unseen words

    $p(\text{"I am a } \textit{zdwi}\text{."}) = ?$
  - No word order

    $p(\text{"I am a student."}) = p(\text{"a I . student am"})$

Content

1. What is a Language Model?
2. **N-gram Language Model**
3. Smoothing
4. Evaluation
5. Management of Large Language Models
6. Other Approaches

N-gram Language Model

$$p(s = w_1 \ldots w_n) = p(w_1)p(w_2|w_1) \ldots p(w_n|w_1 \ldots w_{n-1})$$

$$= \prod_{i=1}^{n} p(w_i|w_1 \ldots w_{i-1})$$

N-gram Language Model

$$p(s = w_1 \ldots w_n) = p(w_1)p(w_2|w_1) \ldots p(w_n|w_1 \ldots w_{n-1})$$

$$= \prod_{i=1}^{n} p(w_i|\underbrace{w_1 w_2 \ldots w_{i-1}}_{\text{The whole history } (i-1 \text{ words})})$$

$$\approx \prod_{i=1}^{n} p\left(w_i|\underbrace{w_{i-(N-1)} w_{i-(N-2)} \ldots w_{i-1}}_{\text{The shortened history } (N-1 \text{ words})}\right)$$

N-gram Language Model

1-gram (unigram) Model:

$$p(s = w_1 \ldots w_n) = p(w_1) \times p(w_2) \times \cdots \times p(w_n)$$

2-gram (bigram) Model:

$$p(s = w_1 \ldots w_n) = p(w_1) \times p(w_2|w_1) \times \cdots \times p(w_n|w_{n-1})$$

3-gram (trigram) Model:

$$p(s = w_1 \ldots w_n) = p(w_1) \times p(w_2|w_1) \times p(w_3|w_1 w_2) \times \ldots \ldots \times p(w_n|w_{n-2} w_{n-1})$$

N-gram Language Model

$p(I\ don't\ like\ spiders\ that\ are\ poisonous)$

bigram

$p(I\ don't\ like\ spiders\ that\ are\ poisonous)$

trigram

N-gram Language Model

$$p(I\ don't\ like\ spiders\ that\ are\ poisonous) = p(I) \times p(don't|I) \times p(like|don't) \times p(spiders|like) \times p(that|spiders) \times p(are|that) \times p(poisonous|are)$$

bigram

$$p(I\ don't\ like\ spiders\ that\ are\ poisonous) = p(I) \times p(don't|I) \times p(like|I\ don't) \times p(spiders|\ don't\ like\ ) \times p(that\ |like\ spiders) \times p(are\ |spiders\ that) \times p(poisonous|that\ are)$$

trigram

Deal with Sentence Boundaries

- To better estimate the probabilities of words at sentence boundaries, we usually add <s> and </s> before and

after sentences

N-gram Language Model

$$
\begin{aligned}
p(&\text{<s>}I\ don't\ like\ spiders\ that\ are\ poisonous\text{</s>}) \\
&= p(I|\text{<s>}) \\
&\times p(don't|I) \\
&\times p(like|don't) \\
&\times p(spiders|like) \\
&\times p(that|spiders) \\
&\times p(are|that) \\
&\times p(poisonous|are) \\
&\times p(\text{</s>}|poisonous)
\end{aligned}
$$

bigram

$$
\begin{aligned}
p(&\text{<s>}I\ don't\ like\ spiders\ that\ are\ poisonous\text{</s>}) \\
&= p(I|\text{<s>}) \\
&\times p(don't|\text{<s>}I) \\
&\times p(like|I\ don't) \\
&\times p(spiders|\ don't\ like\ ) \\
&\times p(that\ |like\ spiders) \\
&\times p(are\ |spiders\ that) \\
&\times p(poisonous|that\ are) \\
&\times p(\text{</s>}|are\ poisonous) \\
&\times p(\text{</s>}|poisonous)
\end{aligned}
$$

trigram

Probability of an n-gram

Suppose we have the phrase “x y” (i.e. word “x” followed by word “y”)

p(y|x) is the probability that word y follows word x and can be estimated from a corpus as follows

$$p(y|x) = \frac{number\text{-}of\text{-}occurrences\ (\text{“}x\ y\text{”})}{number\text{-}of\text{-}occurrences\ (\text{“}x\text{”})}$$

bigram

Similarly, suppose we have the phrase “x y z”.

p(z|x y) is the probability that word z follows words x and y

$$p(z|x\ y) = \frac{number\text{-}of\text{-}occurrences\ (\text{“}x\ y\ z\text{”})}{number\text{-}of\text{-}occurrences\ (\text{“}x\ y\text{”})}$$

trigram

Estimation with n-gram LM

Trigram LM

History

the red (total: 225)

word	c.	prob.
cross	123	
tape	31	
army	9	
card	7	
,	5	

P(cross|the,red) =

Estimation with n-gram LM

Trigram LM

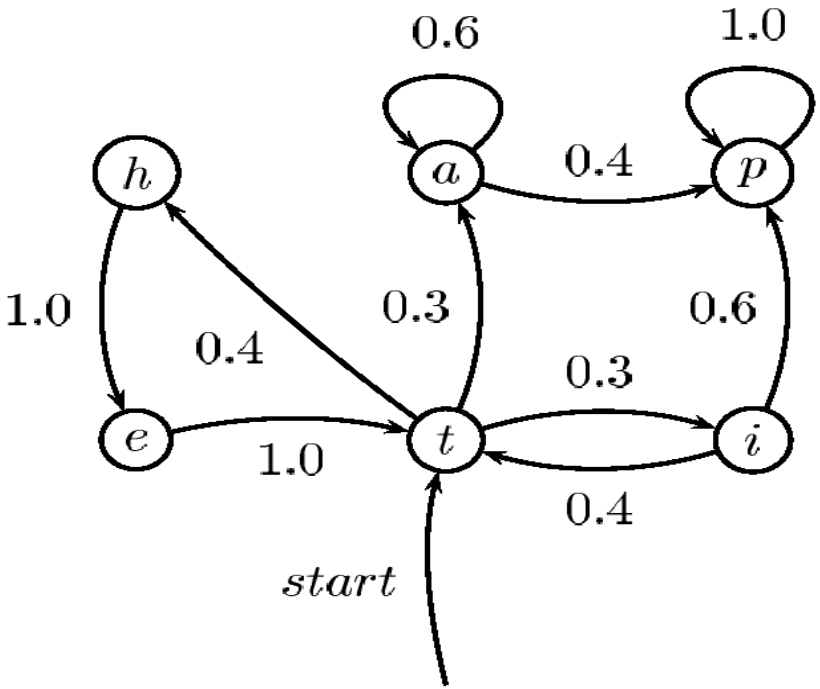
History

the red (total: 225)

word	c.	prob.
cross	123	0.547
tape	31	0.138
army	9	0.040
card	7	0.031
,	5	0.022

P(cross|the,red) = 123 / 225

A character-based bigram LM



$$p(\text{t} - \text{i} - \text{p}) = p(\text{t}) \times p(\text{i}|\text{t}) \times p(\text{p}|\text{i})$$
$$= 1.0 \times 0.3 \times 0.6$$
$$= 0.18$$

Unseen n-grams

- We have seen "$i\ like$" to in our corpus
- We have never seen "$i\ like\ to\ smooth$" in our corpus

$$p(smooth|i\ like\ to) = 0$$

- Any sentence that includes "$i\ like\ to\ smooth$" will be assigned probability 0
- Why is this a bad thing?

Discussion

Next Week

1. What is a Language Model?
2. N-gram Language Model
3. **Smoothing**
4. Evaluation
5. Management of Large Language Models
6. Other Approaches

Discussion

CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computer Science

http://www.cit.ie

Thank you

Haithem. afli@cit.ie

@AfliHaithem