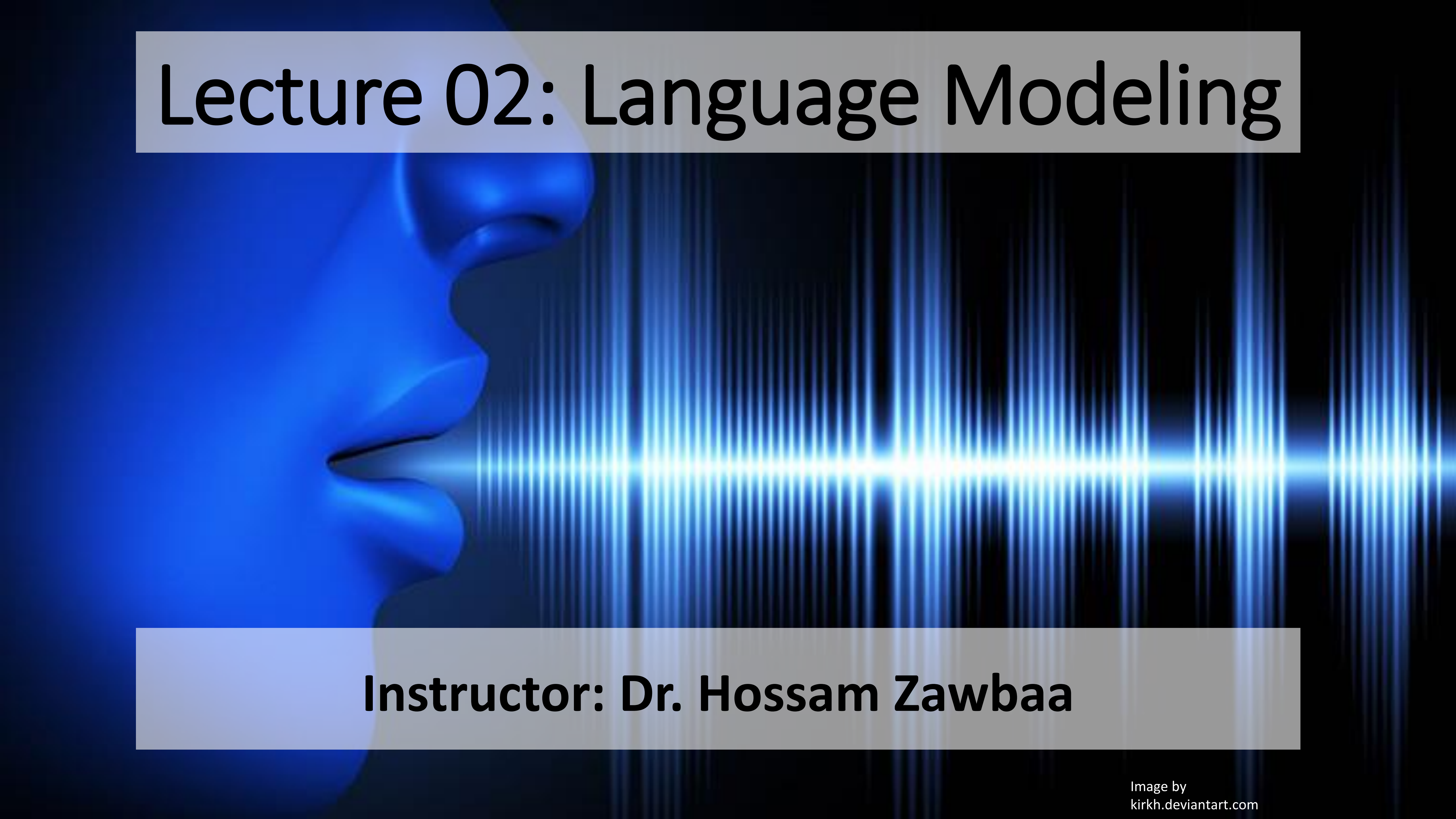
Lecture 02: Language Modeling



Instructor: Dr. Hossam Zawbaa

Image by
kirkh.deviantart.com

How many words?

- I do mainly business data processing
  - Fragments
  - Filled pauses
- Are **cat** and **cats** the same word?
- Some terminology
  - **Lemma**: a set of lexical forms having the same stem, major part of speech, and rough word sense
    - Cat and cats = **same lemma**
  - **Wordform**: the full inflected surface form.
    - Cat and cats = **different wordforms**

Moving toward language
Probability and part of speech tags

- What's the probability of a random word (from a random dictionary page) being a verb?

- How to compute each of these:
  - All words = just count all the words in the dictionary
  - # of ways to get a verb: number of words which are verbs!
  - If a dictionary has 50,000 entries, and 10,000 are verbs
    - **P(V) is 10000/50000 = 0.2**

How many words?

- ***Token*** is an **individual occurrence** of a linguistic unit in speech or words in a text.
- ***Type*** is the number of **distinct** linguistic unit in speech or words in a text.
- Thus, the sentence "a good food is a food that you like" contains **nine** tokens, but only **seven** types, as "a" and “food" are repeated.
- Switchboard-1 Telephone Speech Corpus (SWBD):
  - ~20,000 wordform types,
  - 2.4 million wordform tokens
- Brown et al (1992) large corpus
  - 293,181 wordform types
  - 583 million wordform tokens
- Let N = number of tokens, V = vocabulary = number of types
- General wisdom: V > O(sqrt(N))

Language Modeling

- We want to compute P(w1,w2,w3,w4,w5…wn), the probability of a sequence.
- Alternatively we want to compute P(w5|w1,w2,w3,w4): the probability of a word given some previous words.
- The model that computes P(W) or P(wn|w1,w2…wn-1) is called the **language model**.
- A better term for this would be “The Grammar”.
- But “Language model” or LM is standard.

Computing P(W)

- How to compute this joint probability:
  - P(“the”,”other”,”day”,”I”,”was”,”walking”,”along”,”and”,”saw”,”a”,”lizard”)
- Note: let’s rely on the Chain Rule of Probability

The Chain Rule of Probability

- Recall the definition of conditional probabilities

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

- Rewriting:

$$P(A \wedge B) = P(A|B)P(B)$$

- More generally

$$P(A,B,C,D) = P(A)\ P(B|A)\ P(C|A,B)\ P(D|A,B,C)$$

- In general

$$P(x_1,x_2,x_3,\ldots x_n) = P(x_1)\ P(x_2|x_1)\ P(x_3|x_1,x_2)\ \ldots\ P(x_n|x_1\ldots x_{n-1})$$

The Chain Rule Applied to joint probability of words in sentence

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)\ldots P(w_n|w_1^{n-1})$$
$$= \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$

- P("the big red dog was") = P(the) * P(big|the) * P(red|the big) * P(dog|the big red) * P(was|the big red dog)

Unfortunately

- There are a lot of possible sentences

- We'll never be able to get enough data to compute the statistics for those long prefixes:
  $P(\text{lizard}|\text{the,other,day,I,was,walking,along,and,saw,a})$

Markov Assumption

- Make the simplifying assumption
  - P(lizard|the,other,day,I,was,walking,along,and,saw,a) = P(lizard|a)
- Or maybe
  - P(lizard|the,other,day,I,was,walking,along,and,saw,a) = P(lizard|saw,a)

An example

- <s> I am Sam </s>
- <s> Sam I am </s>
- <s> I do not like green eggs and ham </s>

$$P(\texttt{I} \mid \texttt{<s>}) = \frac{2}{3} = .66 \qquad P(\texttt{Sam} \mid \texttt{<s>}) = \frac{1}{3} = .33 \qquad P(\texttt{am} \mid \texttt{I}) = \frac{2}{3} = .33$$

$$P(\texttt{</s>} \mid \texttt{Sam}) = \frac{1}{2} = 0.5 \qquad P(\texttt{<s>} \mid \texttt{Sam}) = \frac{1}{2} = 0.5 \qquad P(\texttt{Sam} \mid \texttt{am}) = \frac{1}{2} = .5$$

$$P(\texttt{do} \mid \texttt{I}) = \frac{1}{3} = .33$$

Maximum Likelihood Estimates

- The maximum likelihood estimate of some parameter of a model M from a training set T
  - Is the estimate
  - that maximizes the likelihood of the training set T given the model M
- Suppose the word Chinese occurs 400 times in a corpus of a million words (Brown corpus)
- What is the probability that a random word from some other text will be "Chinese"
- MLE estimate is 400/1000000 = .004
  - This may be a bad estimate for some other corpus
- But it is the **estimate** that makes it **most likely** that "Chinese" will occur 400 times in a million word corpus.

More examples: Berkeley Restaurant Project sentences

- mid priced thai food is what i’m looking for
- can you give me a listing of the kinds of food that are available
- i’m looking for a good place to eat breakfast
- when is caffe venezia open during the day

Raw bigram counts

- Out of 9222 sentences

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Raw bigram probabilities

- Normalize by unigrams:

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

- Result:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Bigram estimates of sentence probabilities

- P(<s> I want english food </s>) =

  p(i|<s>) x p(want|I) x p(english|want) x p(food|english) x p(</s>|food)

  = 0.24 x 0.33 x 0.0011 x 0.5 x 0.68

  = 0.000031

What kinds of knowledge?

- P(english|want) = 0.0011
- P(chinese|want) = 0.0065
- P(to|want) = 0.66
- P(eat | to) = 0.28
- P(food | to) = 0
- P(want | spend) = 0
- P (i | <s>) = 0.25

The Shannon Visualization Method

- Generate random sentences
- Choose a random bigram <s>, w according to its probability
- Now choose a random bigram (w, x) according to its probability
- And so on until we choose </s>
- Then string the words together
- <s> I

  I want

  want to

  to eat

  eat Chinese

  Chinese food

  food </s>

Unigram	• 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	• 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
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.

Shakespeare as corpus

- N=884,647 tokens, V=29,066
- Shakespeare produced 300,000 bigram types out of V^2= 844 million possible bigrams: so, 99.96% of the possible bigrams were never seen (have zero entries in the table)
- Quadrigrams worse: What's coming out looks like Shakespeare because it ***is*** Shakespeare