# Natural Language Processing
# Probabilistic Language Models

Felipe Bravo-Marquez

June 13, 2023

# Overview

- The language modeling problem
- Trigram models
- Evaluating language models: perplexity
- Estimation techniques:
  1. Linear interpolation
  2. Discounting methods
- This slides are based on the course material by Michael Collins: `http://www.cs.columbia.edu/~mcollins/cs4705-spring2019/slides/lmslides.pdf`

# The Language Modeling Problem

- We have some (finite) vocabulary, say $\mathcal{V} = \{$the, a, man, telescope, Beckham, two, . . .$\}$
- We have an (infinite) set of strings, $\mathcal{V}^*$.
- For example:
  - the STOP
  - a STOP
  - the fan STOP
  - the fan saw Beckham STOP
  - the fan saw saw STOP
  - the fan saw Beckham play for Real Madrid STOP
- Where STOP is a special symbol indicating the end of a sentence.

# The Language Modeling Problem (Continued)

- We have a training sample of example sentences in English.
- We need to ”learn” a probability distribution $p$.
- $p$ is a function that satisfies:

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

$$p(x) \geq 0 \quad \text{for all } x \in V^*$$

- Examples of probability distributions:

$$p(\text{the STOP}) = 10^{-12}$$

$$p(\text{the fan STOP}) = 10^{-8}$$

$$p(\text{the fan saw Beckham STOP}) = 2 \times 10^{-8}$$

$$p(\text{the fan saw saw STOP}) = 10^{-15}$$

$$\ldots$$

$$p(\text{the fan saw Beckham play for Real Madrid STOP}) = 2 \times 10^{-9}$$

# Why on earth would we want to do this?

- Speech recognition was the original motivation.
- Consider the sentences: 1) recognize speech and 2) wreck a nice beach.
- These two sentences sound very similar when pronounced, making it challenging for automatic speech recognition systems to accurately transcribe them.
- Language models come into play to disambiguate the sentences by leveraging context and language patterns.
- The language model assigns probabilities to different word sequences based on their frequency of occurrence in a large corpus of text.
- In this case, when the speech recognition system analyzes the audio input and tries to transcribe it, it takes into account the language model probabilities to determine the most likely interpretation.
- The language model would favor $p$(recognize speech) over $p$(wreck a nice beach) since the former is a more common and coherent word sequence in the context of speech recognition.

# Why on earth would we want to do this?

- By incorporating language models, speech recognition systems can improve accuracy by selecting the sentence that aligns better with linguistic patterns and context, even when faced with similar-sounding alternatives.
- Related problems are optical character recognition, handwriting recognition.
- Acutally, Language Models are useful in any NLP tasks involving the generation of language (e.g., machine translation, chatbots).
- The estimation techniques developed for this problem will be VERY useful for other problems in NLP.

# Questions?

Thanks for your Attention!

# References I