Multilingual Natural Language Processing

Winston Wu

9/3/2020

Today

- Course Logistics
- Introductions
- Puzzle
- Language Models
- Upcoming Events and Homework

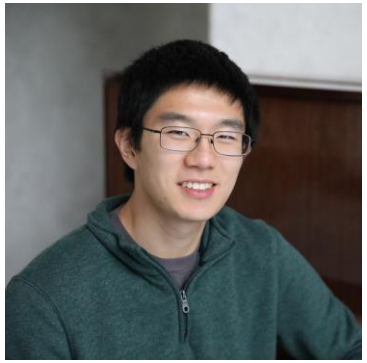
Course Logistics

EN.500.111(32) HEART: Multilingual Natural Language Processing

- Time: Thursdays 5-6:15pm
- Location: Online!

Instructor: Winston Wu

- Email: wswu@jhu.edu
- Office: Hackerman 226



By the end of the semester, you'll …

- Have basic knowledge of
  - Natural language processing
  - Linguistics
  - Historical linguistics
  - Machine learning
  - Machine translation
- Be able to think algorithmically and approach NLP problems
- Be familiar with data science tools and methodologies
- Appreciate the broad diversity of languages in the world

Topics

- General linguistics
- Orthography
- Historical linguistics
- Language identification
- Language modeling
- Tokenization and segmentation
- Word Embeddings
- Machine translation
- Low-resource NLP

About This Course

- Hopkins Engineering Applied Research Tutorials (HEART)
  - Small, ~12 person seminars
  - Designed to introduce freshmen to research
  - Taught by advanced PhD students

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

- Related courses
  - HEART: Breaking the Language Barrier: Recent Advances in Neural Machine Translation (Shapiro)
  - Introduction to Human Language Technology (Various)
  - Natural Language Processing (Eisner)
  - Machine Translation (Koehn)

What do I expect from you?

- This class is graded S/U
- Come to class, participate, and ask questions!
  - Piazza
- Collaboration: do assignments and project in pairs
- Assignments
  - Playing around with notebooks
  - Readings
- Language in 10 minutes
- Final project: run your own NLP experiment

Expectations for Online Classes

- Please rename yourself so we all know who we are
- Please use video if you can
- Mute yourself when you are not speaking
- Participate when requested
- Join and participate in breakout rooms when requested
- Lectures will not be recorded
- Course materials will be posted on the course webpage
- Outside of class, ask and answer questions on Piazza

Introductions

- Name
- Major
- Where are you from?
- What languages are you familiar with?
- What are your interests outside of school?
- What is language?

What is Multilingual NLP?

What is language?

What is Multilingual NLP?

What is natural language?

(and what does it mean to know a language?)

What is Multilingual NLP?

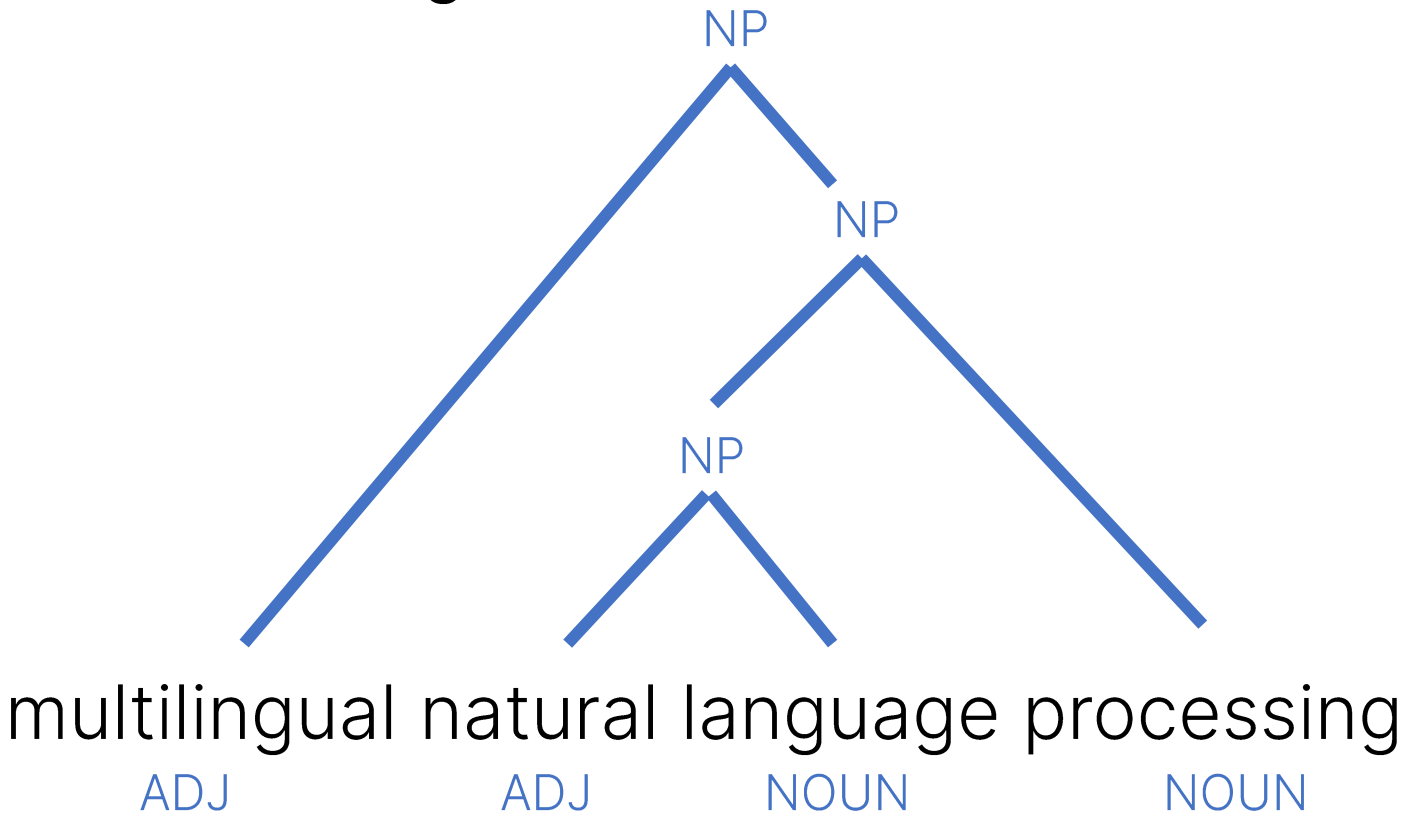
What is natural language processing?
(and where have we seen it in our lives?)

What is Multilingual NLP?

What is multilingual natural language processing?

What did we just do?

- Noun Phrase Chunking
- Parsing



Programming

- We will be using the julia programming language
  - Easy to learn
  - Learn concepts rather than specific implementation
  - Python is the dominant language in NLP research
    - So we will occasionally use Python libraries in Julia

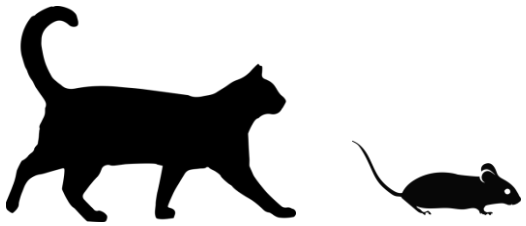
- We will use Pluto.jl notebooks
  - Responsive notebook
    - Immediately see the results of your code, in all relevant cells!
  - What you see is the *entire* state of the notebook
  - Source file is just a regular Julia file

Let’s do a puzzle!

Wheel of Fortune Cookies

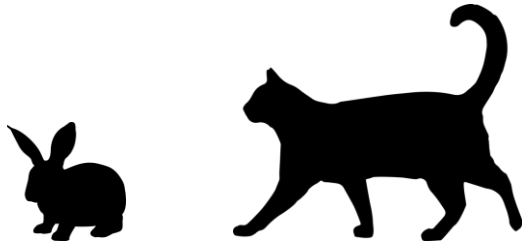
- http://nacloweb.org/resources/problems/2014/N2014-D.pdf

Question



The cat chased the mouse

The cat chased the rabbit



Which sentence is more likely?

How do you know?

How would a computer know?

Language Modeling

- A language model tells us how likely a sequence of words is
  - Everyone has one in their head!

- Why study language modeling?
  - Language is how humans encode knowledge about the world
  - There is a lot of language out there
    - Books, the internet
    - 7000+ languages
  - Knowing language can help solve other problems
  - State-of-the-art GPT-3 has gotten a lot of hype recently

Language Modeling

Probability!

- A language model tells us how likely a sequence of words is

- Simple language model: count how many times the sentence appears in a corpus

$$P(\text{phrase}) = \frac{\text{count(phrase)}}{\text{total \# of phrases}}$$

  - P(The cat chased the mouse) = 0.0000004
  - P(The cat chased the rabbit) = 0

- A probability of zero means the sentence is impossible?!

N-grams

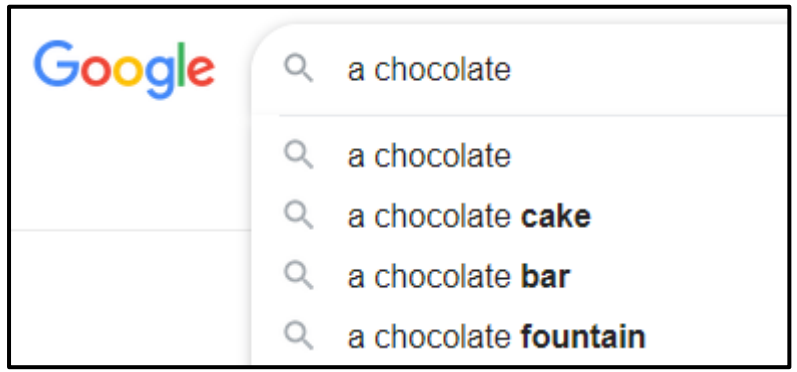
- Break down the sentence into a sequence of words

The cat chased the rabbit

≈ [The cat chased, cat chased the, chased the rabbit]

- Kinda naïve, but it is simple, fast, and works surprisingly well!
- Markov assumption: a word depends only on the previous couple of words, not the entire sentence

~~The tiny elephant ate~~ a chocolate _____



Comparing Sentences

A: The cat chased the mouse

B: The cat chased the rabbit

Suppose P(The cat chased the mouse) = 0

P(The cat chased the rabbit) = 0

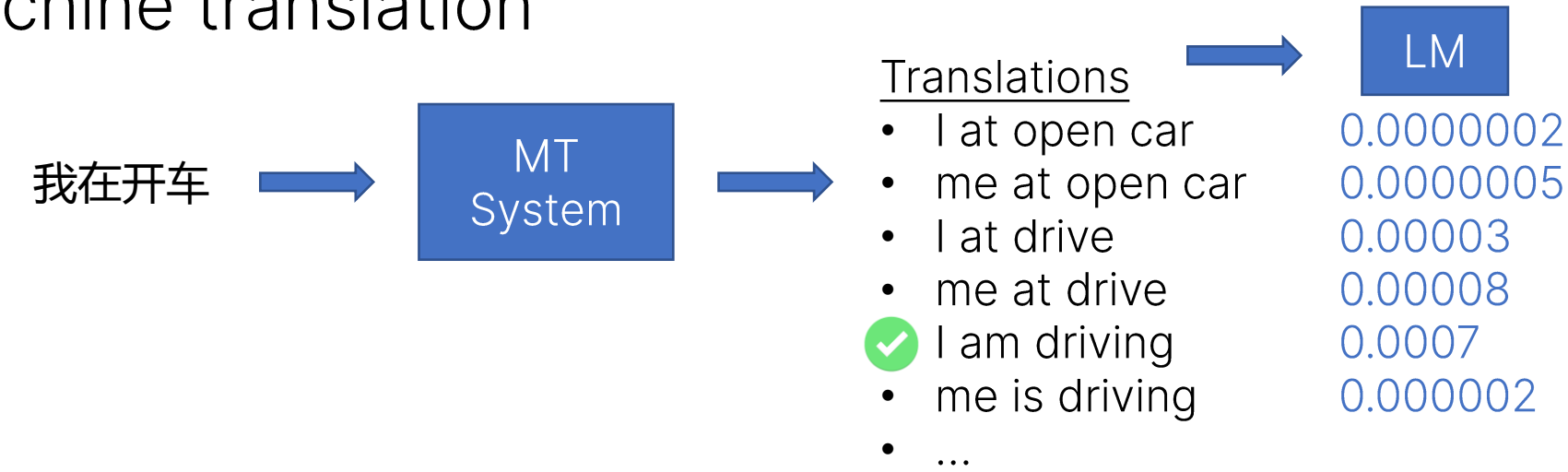
Which sentence is more likely?

A		B
P(chased the mouse)	>	P(chased the rabbit)

Applications

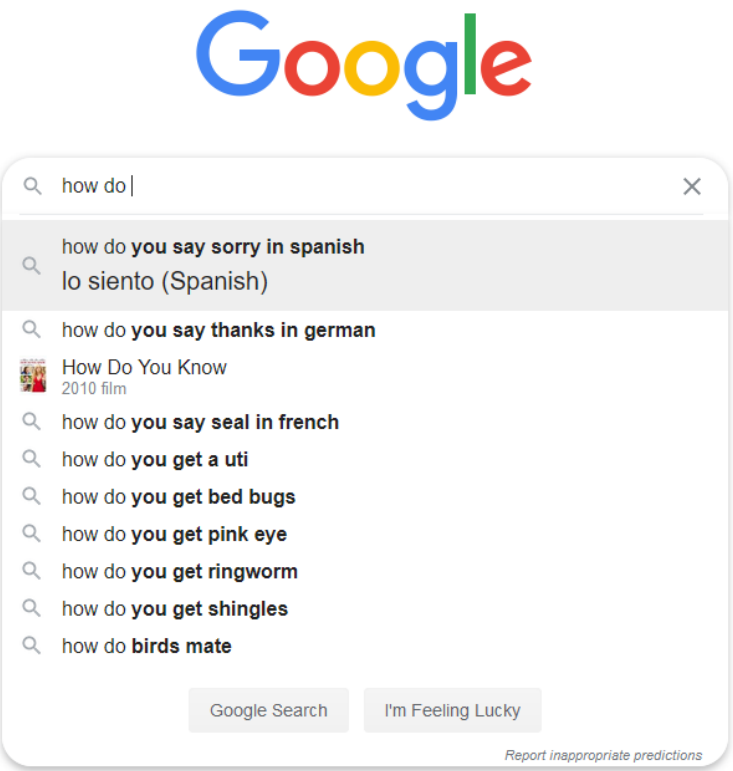
Language models can rank the output of other systems

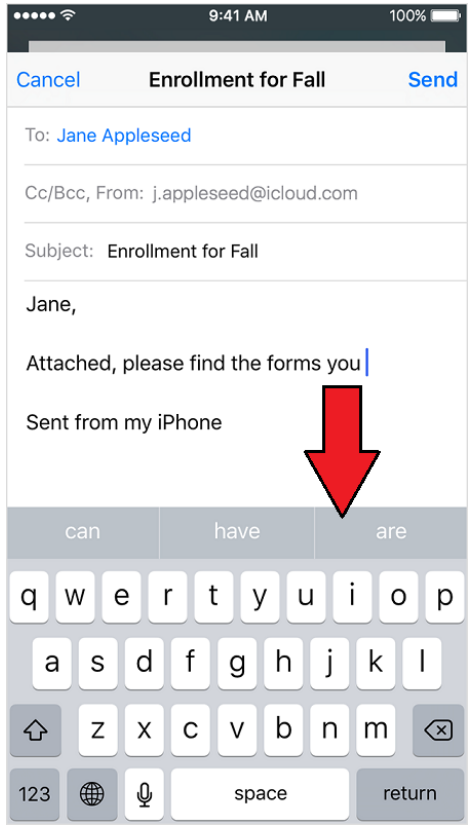
- Speech recognition
  - “recognize speech” vs “wreck a nice beach”
- Machine translation



Prediction

- Language models can also be used to predict text

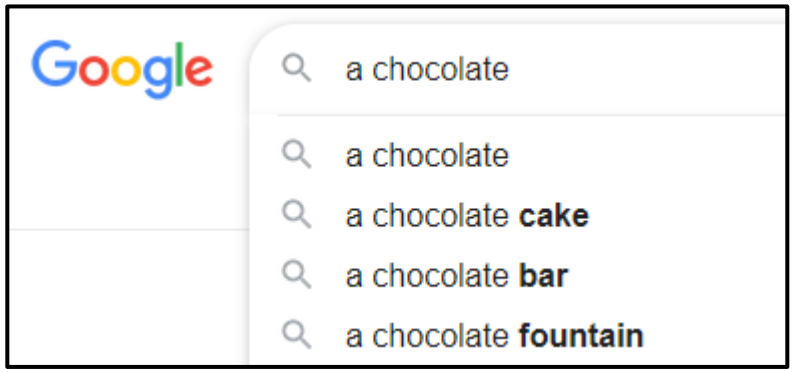




Prediction with n-gram LMs

$$\text{next word} = \underset{\text{word} \in \text{vocabulary}}{\textit{argmax}}\ P(\text{word} \mid \text{context})$$

$$P(\text{word} \mid \text{context}) = \frac{\text{count}(\text{context} + \text{word})}{\text{count}(\text{context} + ***)}$$



Example:

- context = "a chocolate"
- Compute P(a | a chocolate), P(aardvark | a chocolate), P(zygote | chased the), etc. for all words in the vocabulary
- Pick the word that maximizes this probability

Let’s code this up!

- You should have already downloaded Julia and Pluto
- Let’s try them out!

Topics

- Tokenization
- Frequency statistics
- N-grams
- Language model prediction

Reading

Try out GPT-3 in the links here:

- https://www.reddit.com/r/GPT3/comments/id9hru/list_of_free_sitesprograms_that_are_powered_by/

Just for fun:

- https://www.reddit.com/r/SubSimulatorGPT2/

Read this critique of GPT-3:

- https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

Scheduling

- Julia Tutorial
- Office Hours
  - Fill out the when2meet on Piazza

- Mid-semester checkup (late-October)

- December 3rd for project presentations?

Upcoming Events

- JHU Virtual Student Involvement Fair
  - Thursday, Sept. 3 (9-11am)
  - Friday, Sept. 4 (2-4pm)
  - Saturday, Sept. 5 (9-11pm)

- HopHacks
  - Register at https://hophacks.com/
  - Sept. 11-13

Survey Questions

- At the end of every class, I will ask some survey questions
  - Get you to recall and think about the material
  - Give me feedback
- They're quick! Should only take a minute ("minute paper")
- Send me a private message on Zoom

Survey Questions

1. Briefly describe what is a language model.
2. What was most interesting about today's topics?
3. What were you most confused about?
4. Any other questions, comments, concerns, suggestions?