VIRGINIA TECH

Course Overview and Introduction

Lifu Huang

lifuh@vt.edu

Torgersen Hall, Suite 3160E

Today's lecture

- Course Admin:
  - How will we teach this course?
  - How will you be evaluated?

- Course Overview:
  - What is Natural Language Processing?
  - What topics will be covered in this course?

Class Information

- **Instructor**:
  - Lifu Huang (https://wilburone.github.io/)
  - Class Meets: 1:25-2:15pm @ DER 1014, Monday, Wednesday and Friday
  - Lectures: will be posted on Canvas before the class
  - Office Hours: 4-5pm @ Zoom, Monday and Wednesday
  - https://virginiatech.zoom.us/j/5961845863
- **Teaching Assistants**：
  - Shuaicheng Zhang (GTA)
  - Email: zshuai8@vt.edu
  - Office hours**:** TBD

Class Information

- **Prerequisites**
  - Data structures, algorithms, machine learning
  - Basic knowledge of deep learning, e.g., CNN, RNN, LSTM, Transformer
  - Programming skills in Python, Pytorch/Tensorflow
- **Course Website**
  - https://canvas.vt.edu/courses/136078/pages/natural-language-processing-fall-2021
- **Discussion Forum**
  - Piazza

Textbooks & Recommended References

- **Textbook Options**
  - *Jurafsky and Martin,* ***Speech and Language Processing.*** *(*Digital copy available for free from https://web.stanford.edu/~jurafsky/slp3/*)*
- **Other recommended resources**
  - *Deep Learning* (https://www.deeplearningbook.org/)
  - Online Course: *Natural Language Processing with Deep Learning* (http://web.stanford.edu/class/cs224n/)
  - Github: https://github.com/topics/natural-language-processing
  - Conferences: ACL, EMNLP, NAACL, AACL, COLING, EACL, ICLR, ICML, NeurIPS

Topic Coverage

- Review of Deep Learning Architectures
  - Long Short-Term Memory Networks (LSTM), Convolutional Neural Networks (CNN), Autoencoders, Transformer, Graph Neural Networks (GNN), etc.

- Natural Language Processing
  - Word embeddings, language models, contextualized word representations
  - Part-of-speech tagging, word sense disambiguation, sentiment analysis, parsing, syntactic/semantic parsing
  - Information extraction, machine translation, natural language generation, chatbots and dialogue systems, machine reading comprehension, natural language inference, misinformation, etc.

Course Work: Assignments, Exams and Course Project

- **Assignments** (60%)
  - Four Programming Assignments (in Python), equal weight (4 * 15%)
  - Homework assignments will be posted on Canvas
  - You will have two weeks to complete
- **Literature Review** (10%)
  - Read at least 10 NLP papers on a particular topic, and produce a written report that compares and critiques these approaches
  - Make sure you get a deeper knowledge of NLP by carefully reading the original papers, even if you don't build any actual systems
- **Final (Team) Project** (30%)
  - 1-3 students per team
  - Include source code, proposal/final presentation and report
  - Design and implement a novel approach for a particular task or topic
  - **Novelty** will be an important grading criteria

Other Policies

- **Regrading Requests**
  - Requests for regrading due to grading errors must be submitted to the TA within one week of the release of grades.
- **Late Homework Policy**
  - **No** late assignments will be accepted (sorry)
- **Final Letter Grade**

A:	93.3%–100%,	**A-:**	90.0%–93.3%,	**B+:**	86.6%–90.0%,	**B:**	83.3%–86.6%,	**B-:**	80.0%–83.3%,	**C+:**	76.6%–80.0%,
C:	73.3%–76.6%,	**C-:**	70.0%–73.3%,	**D+:**	66.6%–70.0%,	**D:**	63.3%–66.6%,	**D-:**	60.0%–63.3%,	**F:**	00.0%–60.0%.

- **Academic Integrity**
  - No cheating, no copy code from others or online
- **COVID**
  - **Masks are required at all times in class**

Questions for Short Discussion

- What to do apart from this course?
  - Pick a paper from the top conferences, reimplement it and reproduce the results
  - Pick a task/research project, and design ML/DL algorithms to solve the problem
  - Keep reading new papers (e.g., https://arxiv.org/list/cs.CL/recent) to learn about the new techniques or applications

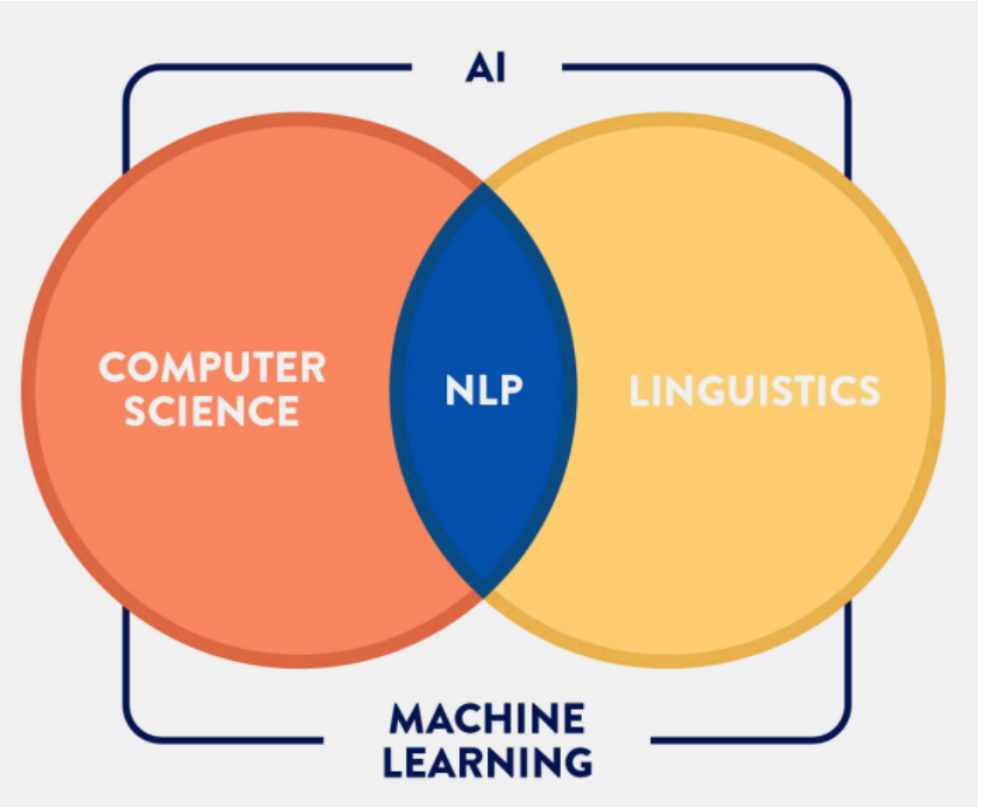
- Interested in working with me on NLP?
  - Weekly reading discussion meeting (maillist: nlp_reading_group)
  - Talk with me at Togersen 3160E

Course Overview

What is Natural Language Processing?

- The interdisciplinary field of Computer Science and Linguistics.

- NLP is the ability of a computer program to **understand human language** as it is spoken and written -- referred to as natural language



https://clevertap.com/blog/natural-language-processing/

Machine Translation

Google Research Philosophy Research Areas Publications People Tools & Downloads Outreach Careers Blog

PUBLICATIONS ›

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

CoRR, vol. abs/1609.08144 (201

Download Google Scho

Abstract

Google Translate Sign in

Text Documents

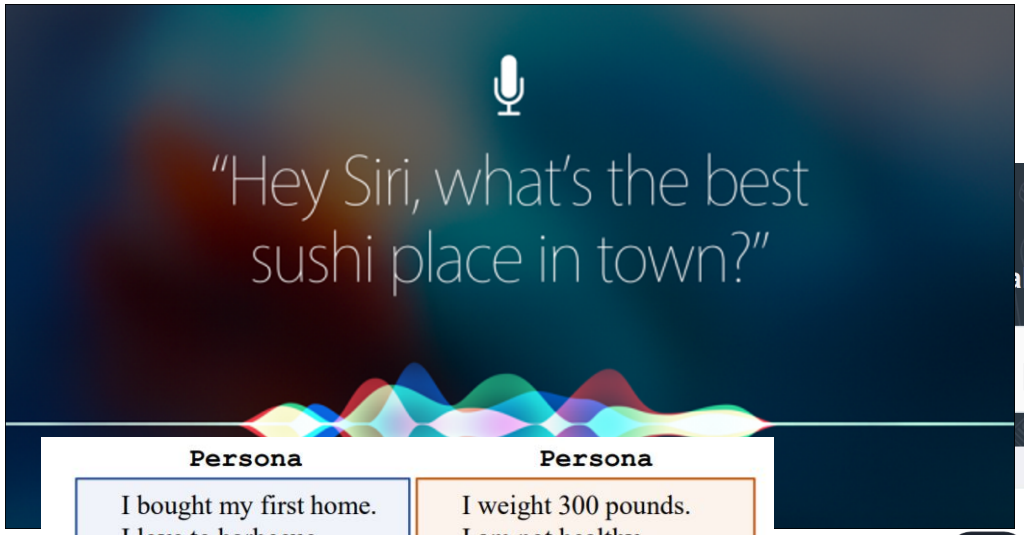
DETECT LANGUAGE ENGLISH SPANISH FRENCH ENGLISH SPANISH ARABIC

Search languages

Afrikaans	Danish	Hmong	Lithuanian	Romanian	Telugu
Albanian	Dutch	Hungarian	Luxembourgish	Russian	Thai
Amharic	✓ English	Icelandic	Macedonian	Samoan	Turkish
Arabic	Esperanto	Igbo	Malagasy	Scots Gaelic	Turkmen
Armenian	Estonian	Indonesian	Malay	Serbian	Ukrainian
Azerbaijani	Filipino	Irish	Malayalam	Sesotho	Urdu
Basque	Finnish	Italian	Maltese	Shona	Uyghur
Belarusian	French	Japanese	Maori	Sindhi	Uzbek
Bengali	Frisian	Javanese	Marathi	Sinhala	Vietnamese
Bosnian	Galician	Kannada	Mongolian	Slovak	Welsh
Bulgarian	Georgian	Kazakh	Myanmar (Burmese)	Slovenian	Xhosa
Catalan	German	Khmer	Nepali	Somali	Yiddish
Cebuano	Greek	Kinyarwanda	Norwegian	Spanish	Yoruba
Chichewa	Gujarati	Korean	Odia (Oriya)	Sundanese	Zulu

- Why it’s so difficult?

Dialog Systems, Chatbots, Digital Assistants



Persona	Persona
I bought my first home. I love to barbecue. I live in Springfield. I'm a writer.	I weight 300 pounds. I am not healthy. I am a man. I like The Godfather.

Hello how are you, I am new to the Springfield area.

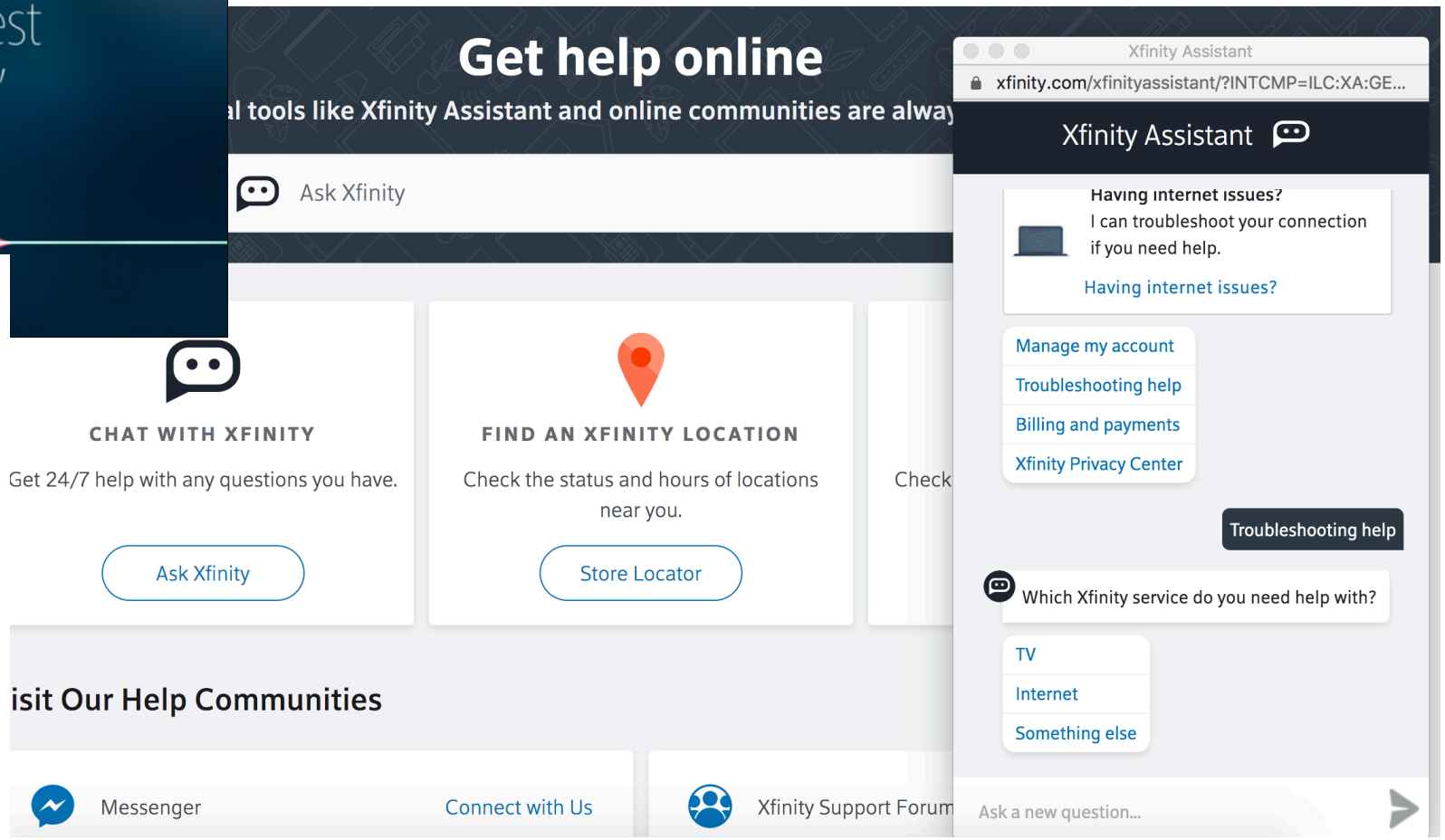
Hi! Seen any good movies lately?

I have been to the movies.

I love The Godfather, one of my favorites! Was that filmed?

I don't believe so. I don't watch movies more of a writer.

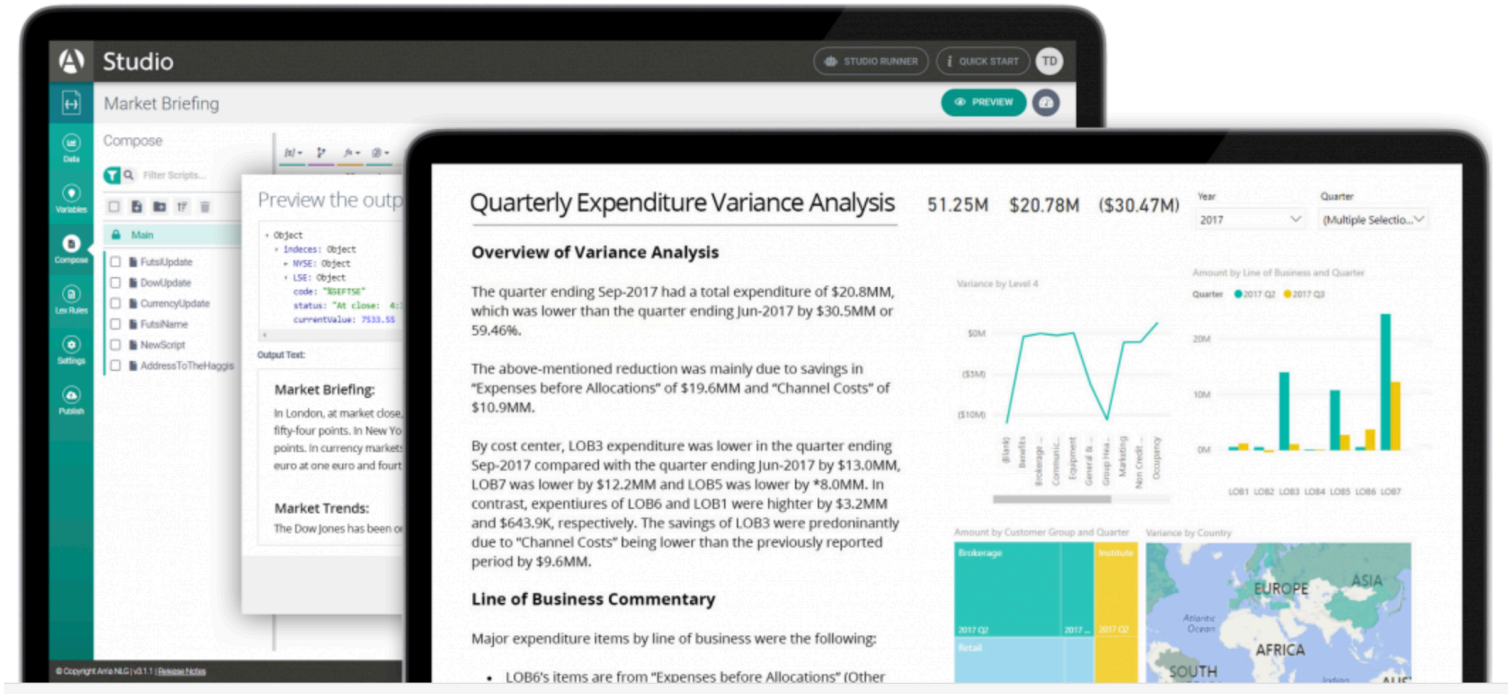
What do you write? Any diet books ? I am not very healthy.



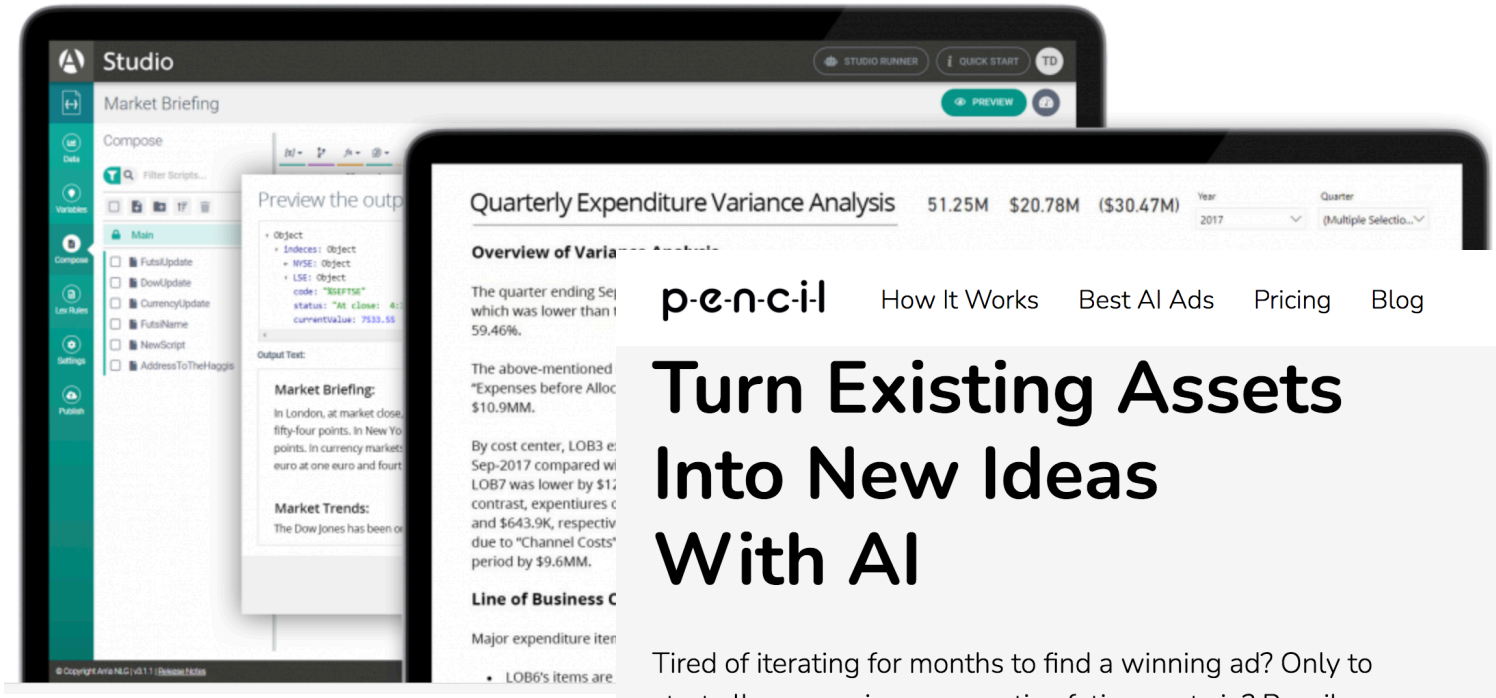
Sentiment/Opinion Analysis

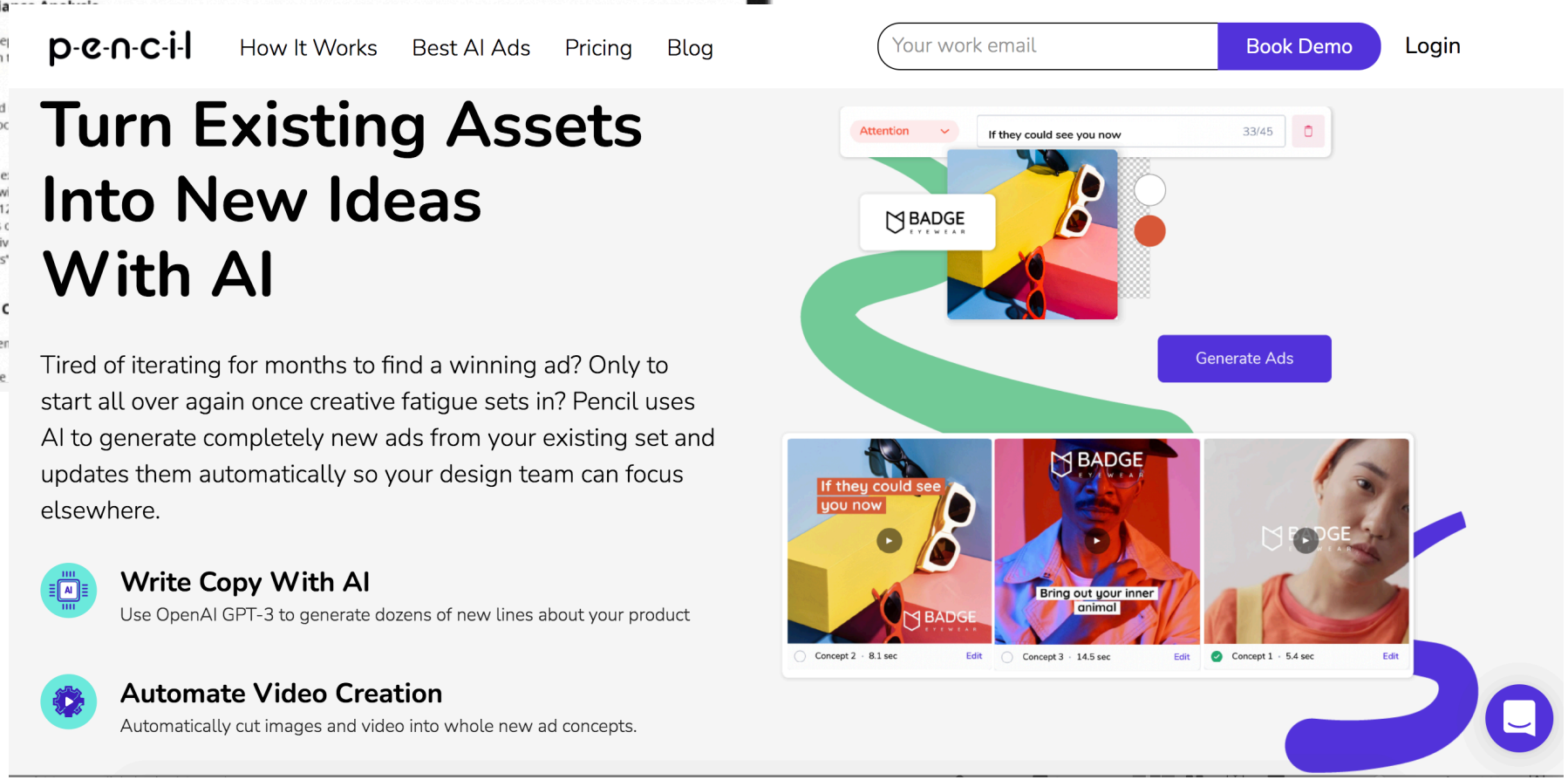
- Identify the emotional tone behind a body of text or other signals

Natural Language Generation



Natural Language Generation





What's the current state of NLP?

- Lots of commercial applications
  - Some applications are working pretty well already
  - Others not so much

- A paradigm shift around Deep Learning and AI more generally
  - Neural nets are powerful classifiers and sequence models
  - Public libraries (Tensorflow, Pytorch, etc.) and datasets make it easy for anyone to get a model up and running
  - "End-to-end" models put into question whether we still need the traditional NLP pipeline
  - We're still in the middle of this paradigm shift

What will you learn from this class?

- The core tasks (as well as data sets and evaluation metrics) that people work on in NLP
- The fundamental models and algorithms that have been developed for these tasks
- The relevant linguistic concepts and phenomena that will be encountered in these tasks

Fundamental Challenges in NLP

- Variety: a meaning can be expressed with various forms
  - Three young boys survived and are in **critical [Injure]** condition after spending 18 hours in the cold.
  - Today I was **let go [End Position]** from my job after working there for 4 years.
  - When we come back, media speculation run amuck over possible indictments at ***sixteen hundred Pennsylvania*** **[White House]** and the President 's scripted session with troops in Iraq .
- Ambiguity: a word may have multiple meanings
  - Still **hurts** ~~**[Attack]**~~ me to read this.
  - Stewart has found the road to fortune wherever she has **traveled** ~~**[Transport Person]**~~.
  -

Building a computer that ‘understands’ text: **The NLP pipeline**

Example

\More than a decade ago, Carl Lewis stood on the threshold of what was to become the greatest athletics career in history. He had just broken two of the legendary Jesse Owens' college records, but never believed he would become a corporate icon, the focus of hundreds of millions of dollars in advertising. His sport was still nominally amateur. Eighteen Olympic and World Championship gold medals and 21 world records later, Lewis has become the richest man in the history of track and field -- a multi- millionaire.

- What is this paragraph about?
- Who is Carl Lewis?
- Did Carl Lewis break any world records?

Task: Tokenization/Segmentation

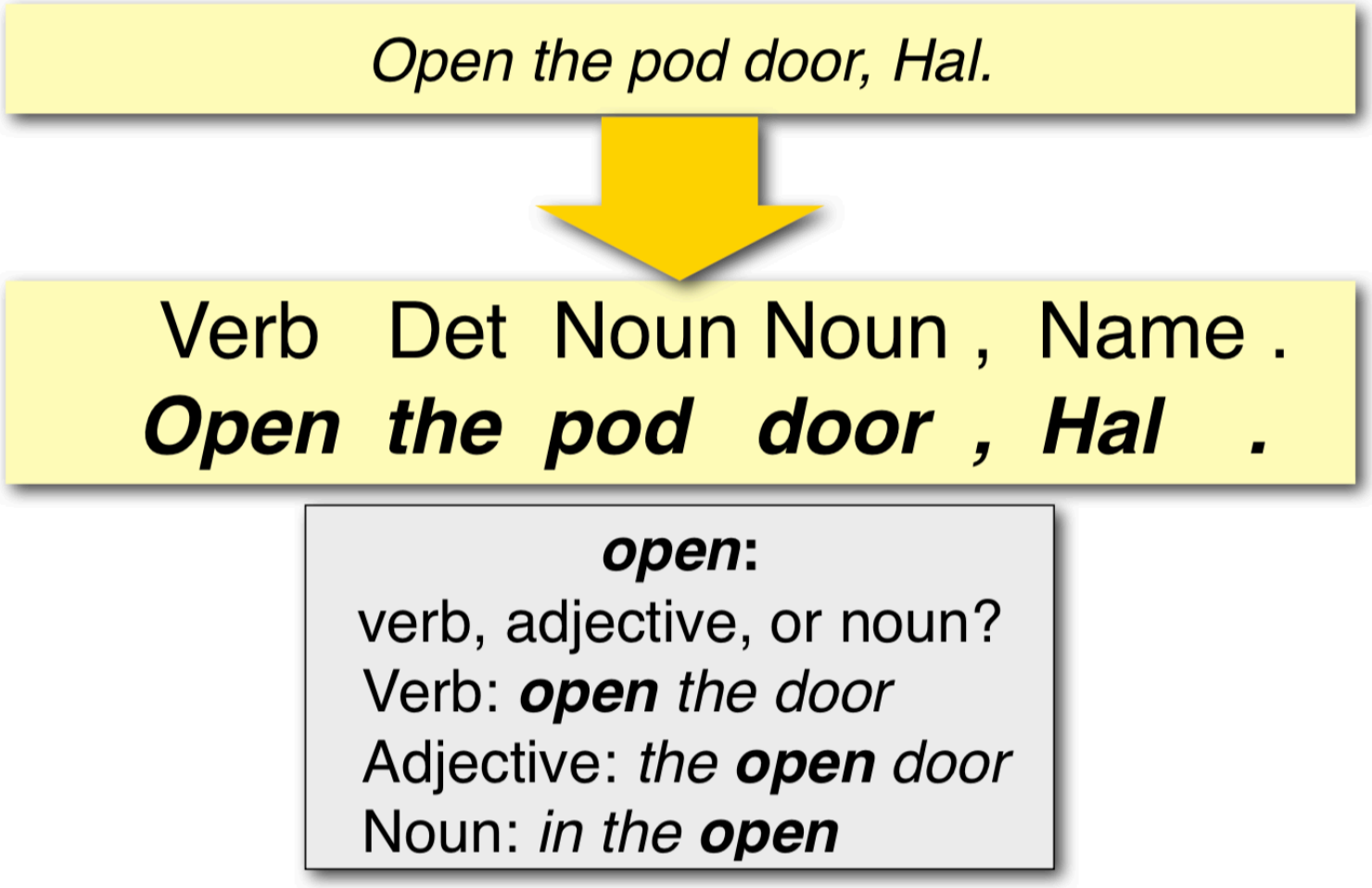
- Split the text into words and sentences
  - Languages like Chinese don’t have any spaces between words

    发生在中国中部河南省的洪水虽然正陆续退却，但灾害仍在持续。据《河南日报》消息，截至7月23日12时，省会郑州暴雨引发的洪涝和次生灾害已经导致51人遇难。

  - Even in English, this cannot be done deterministically

    There was an earthquake near D.C. You could even feel it in Philadelphia, New York, etc.

Task: Part-of-speech (POS) tagging

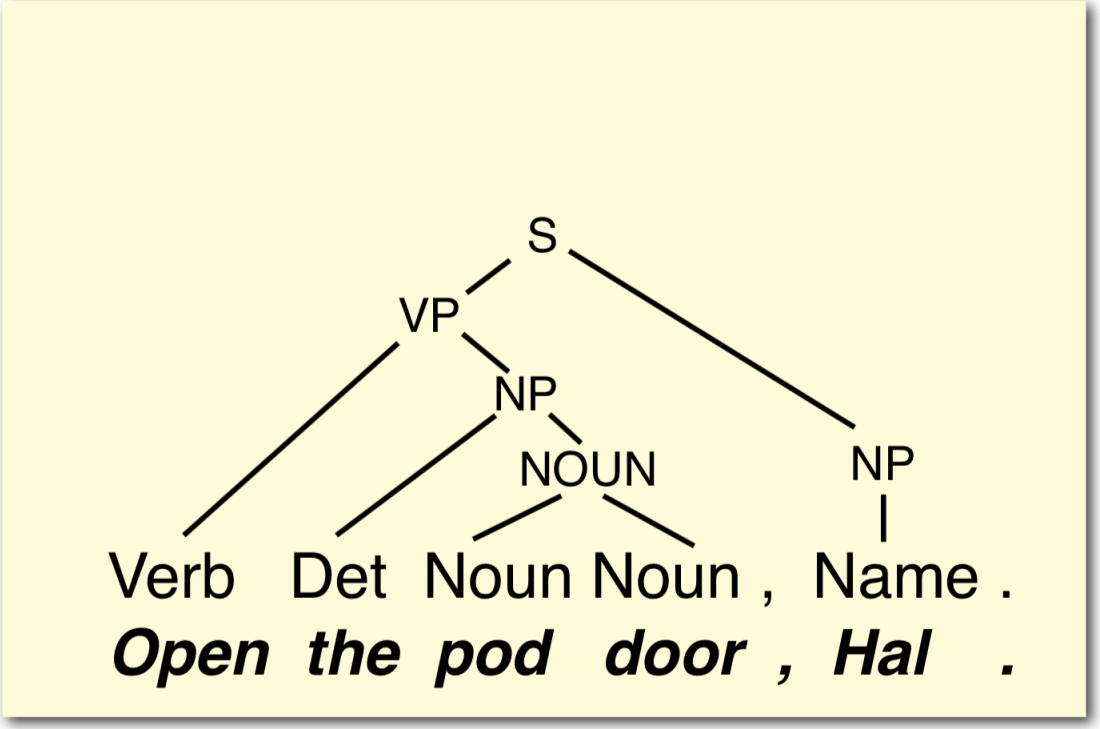


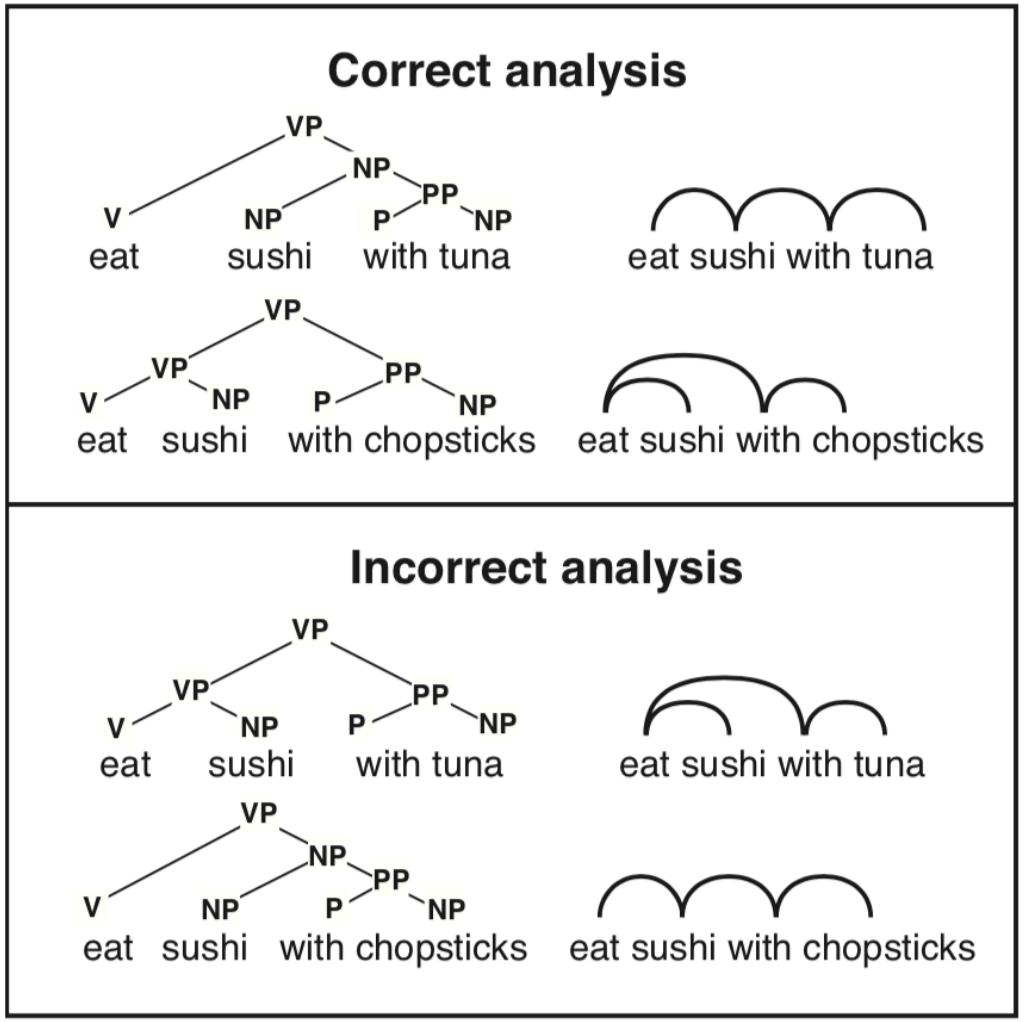
Task: Word Sense Disambiguation

"I made her duck"

- What does this sentence mean?
  - "duck": noun or verb?
  - "make": "cook X" or "cause X to do Y" ?
  - "her": "for her" or "belonging to her" ?
- Language has different kinds of ambiguity, e.g.:
  - Structural ambiguity:
    - "I eat sushi with tuna" vs. "I eat sushi with chopsticks"
  - Lexical (word sense) ambiguity:
    - "I went to the bank": financial institution or river bank?
  - Referential ambiguity
    - "John saw Jim. He was drinking coffee."

Task: Syntactic Parsing

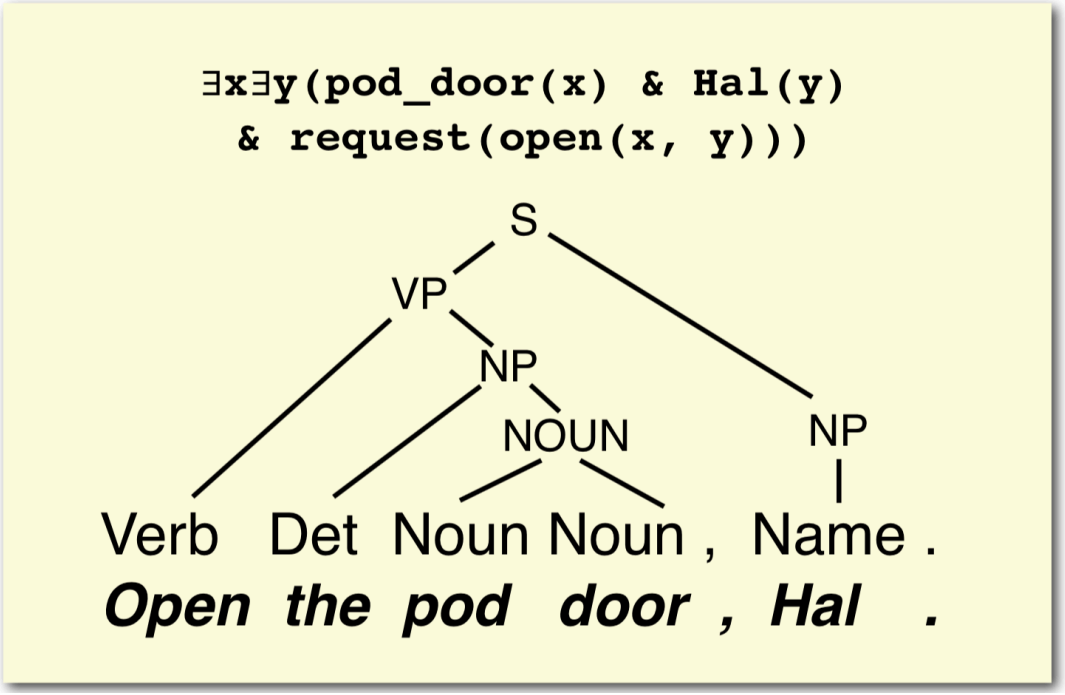




Observation: Structure Corresponds to Meaning

Task: Semantic Analysis

- Need a meaning representation
- Shallow semantic analysis: mapping to table structures (Information Extraction)
- Deep semantic analysis: formal logic
- Lexical semantics v.s. Compositional semantics
  - meanings of words v.s. meanings of sentences



Coreference Resolution

More than a decade ago, **Carl Lewis** stood on the threshold of what was to become the greatest athletics career in history. **He** had just broken two of the legendary Jesse Owens' college records, but never believed **he** would become a corporate icon, the focus of hundreds of millions of dollars in advertising. **His** sport was still nominally amateur. Eighteen Olympic and World Championship gold medals and 21 world records later, **Lewis** has become the richest man in the history of track and field -- a multi- millionaire.

A television advertisement with **Beamon** appeared before the final, featuring the record-holder saying, "**I** hope **you** make it, **kid**." So, when **Lewis** decided not to make any more attempts to try to break the record, **he** was loudly booed. When asked about those boos, **Lewis** said, "**I** was shocked at first. But after **I** thought about it, **I** realized that **they** were booing because **they** wanted to see more of **Carl Lewis**. **I** guess that's flattering."

Summary – NLP Pipeline

- An NLP system may use some of all of the following steps:
  - Tokenizer/ Segmenter
    - to identify words and sentences
  - POS tagger
    - to identify the part of speech of words
  - Word sense disambiguation
    - to identify the meaning of words
  - Syntactic/Semantic Parser
    - to obtain the structure and meaning of the sentence
  - Coreference Resolution
    - to keep track of the various entities and events mentioned in text

Summary – NLP Pipeline Assumptions

- Each step in the NLP pipeline embellishes the input with explicit information about its linguistic structure
  - POS tagging: parts of speech of word,
  - Syntactic parsing: grammatical structure of sentence,….
- Each step in the NLP pipeline requires its own explicit (“symbolic”) output representation
  - POS tagging requires a POS tag set (e.g. NN=common noun singular, NNS = common noun plural, …)
  - Syntactic parsing requires constituent or dependency labels (e.g. NP = noun phrase, or nsubj = nominal subject)

Summary – NLP Pipeline Shortcomings

- Each step in the pipeline relies on a learned model that returns the most likely predictions
  - It requires a lot of **annotated training data** for each step
  - Annotation is **expensive** and sometimes **difficult** (people are not 100% accurate)
  - These models are never 100% accurate
  - **Error propagation**: models make more mistakes if their input contains mistakes
- How do we know that we have captured the "right" generalizations when designing representations?
  - Some representations are easier to predict than others
  - Some representations are more useful for the next steps in the pipeline than others
  - But we won't know how easy/useful a representation is until we have a model that we can plug into a particular pipeline