NLP 201
Introduction

Jeffrey Flanigan

University of California Santa Cruz
jmflanig@ucsc.edu

Fall 2024

Administrative

- Canvas website is up
- Assignment 0 out, due 4pm Friday

- We will send class announcements on Canvas
- Please turn on notifications in Canvas (off by default) or install the app
- Please use Canvas to communicate with Jeff or the TA
- We will try our best to respond within 12-24 hours during the weekdays

What is Natural Language Processing (NLP)?

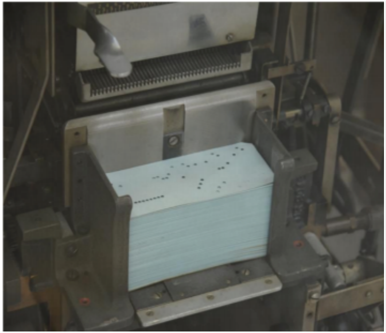
What is Natural Language Processing (NLP)?

- The set of methods for making human language accessible to computers (Eisenstein, 2018).

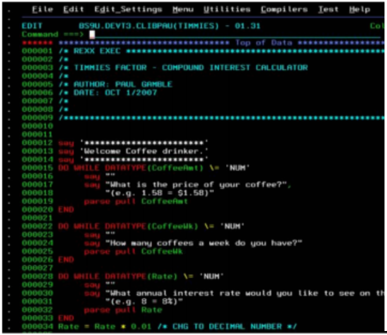
What is Natural Language Processing (NLP)?

- The set of methods for making human language accessible to computers (Eisenstein, 2018).
- Why do we want this?

Communication with Machines



∼ 50s-70s



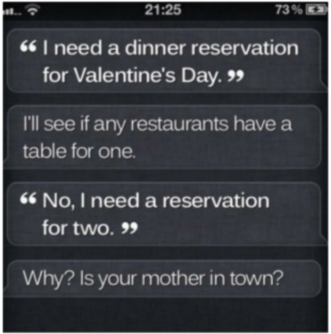
```
EDIT       BS9U.DEVT3.CLIBPAU(TIMMIES) - 01.31                     Col
Command ===>
****** ****************************** Top of Data ******************
/* REXX EXEC ********************************************************
/*
/* TIMMIES FACTOR - COMPOUND INTEREST CALCULATOR
/*
/* AUTHOR: PAUL GAMBLE
/* DATE: OCT 1/2007
/*
/*
/********************************************************************


say '************************'
say 'Welcome Coffee drinker.'
say '************************'
DO WHILE DATATYPE(CoffeeAmt) \= 'NUM'
     say ""
     say "What is the price of your coffee?",
         "(e.g. 1.58 = $1.58)"
     parse pull CoffeeAmt
END

DO WHILE DATATYPE(CoffeeWk) \= 'NUM'
     say ""
     say "How many coffees a week do you have?"
     parse pull CoffeeWk
END

DO WHILE DATATYPE(Rate) \= 'NUM'
     say ""
     say "What annual interest rate would you like to see on th
         "(e.g. 8 = 8%)"
     parse pull Rate
END
Rate = Rate * 0.01 /* CHG TO DECIMAL NUMBER */
```

∼ 80s



“ I need a dinner reservation for Valentine's Day. ”

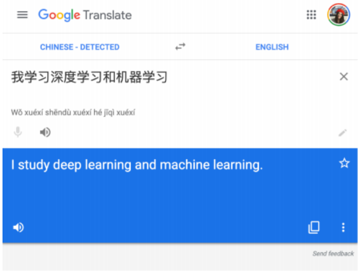
I'll see if any restaurants have a table for one.

“ No, I need a reservation for two. ”

Why? Is your mother in town?

∼ today

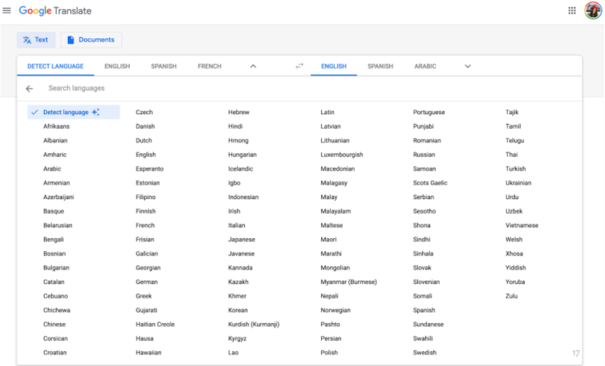
NLP Application: Machine translation



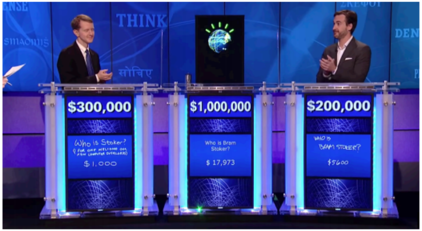
我学习深度学习和机器学习

Wǒ xuéxí shēndù xuéxí hé jīqì xuéxí

I study deep learning and machine learning.



NLP Application: Question Answering



- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England in the end of the 18th century?
- What do scientists think about the ethics of human cloning?

NLP has many end-user tasks (downstream tasks or applications)

- Machine translation
- Summarization
- Question answering
- Conversational agents
- Search (information retrieval)
- Recommender systems
- Document classification

Applications listed on the NLP wiki

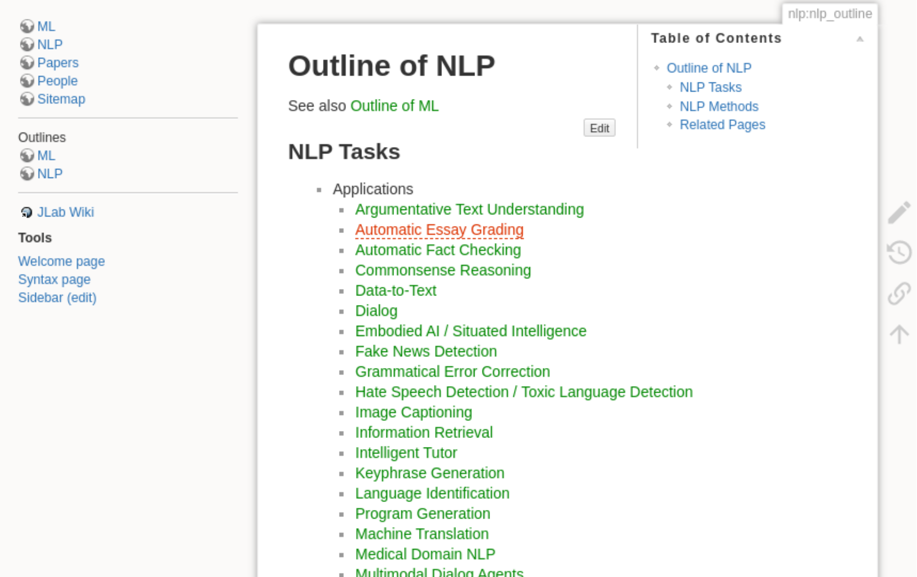


Table of Contents

- Outline of NLP
  - NLP Tasks
  - NLP Methods
  - Related Pages

Outline of NLP

See also Outline of ML

NLP Tasks

- Applications
  - Argumentative Text Understanding
  - Automatic Essay Grading
  - Automatic Fact Checking
  - Commonsense Reasoning
  - Data-to-Text
  - Dialog
  - Embodied AI / Situated Intelligence
  - Fake News Detection
  - Grammatical Error Correction
  - Hate Speech Detection / Toxic Language Detection
  - Image Captioning
  - Information Retrieval
  - Intelligent Tutor
  - Keyphrase Generation
  - Language Identification
  - Program Generation
  - Machine Translation
  - Medical Domain NLP
  - Multimodal Dialog Agents

NLP has many end-user tasks (downstream tasks or applications)

- Machine translation
- Summarization
- Question answering*
- Conversational agents*
- Search (information retrieval)
- Recommender systems
- Document classification

*These two tasks are **supertasks**.

Downstream tasks sometimes benefit from intermediate tasks

- Knowing a word's sense (i.e duck – animal vs duck – action) could help translate it. This is **sense disambiguation**. .
- Knowing if a word is a verb or noun (its part of speech) could help translate it (duck – noun vs duck – verb). This is **part-of-speech (POS) tagging**.
- Splitting text into sentences is often required before processing. This is **sentence segmentation**.
- Deciding what should count as a word ($100 vs $␣100 or it's vs it␣'s) (**tokenization**) usually has a very large effect on performance.

Examples of intermediate tasks

- Tokenization
- Language modeling
- POS tagging
- Lemmatization
- Synactic parsing
- Entity recognition
- Entity linking
- Relation extraction
- Semantic role labeling
- Semantic parsing
- Generation (from an intermediate representation)

Classes of tasks

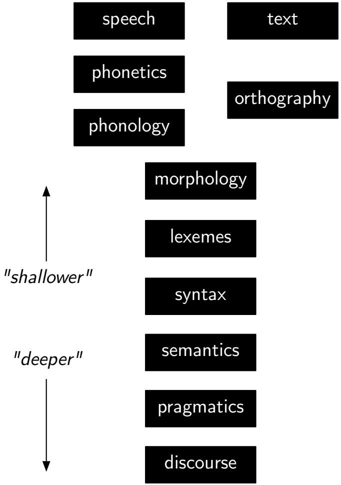
- Document Classification **binary or multi-label classification**
- Tagging **each token gets a label**
- Parsing **produce a tree or other structure over the words in sentence**
- Generation **produce a sentence from some representation of the desired output**
- Sequence-to-sequence **sequence of tokens to another sequence of tokens with possibly different length**

The traditional NLP pipeline

1. Tokenization **decide what is the unit of processing, usually words or subwords**
2. Morphological analysis **analysing the structure of the words**
3. Part-of-speech tagging
4. Syntactic Parsing
5. Semantic Parsing (optional)
6. Downstream task: classification, QA, summarization, etc **use information from previous stages in pipeline**
7. Generation (optional)

With deep learning, sometimes tasks are done **end-to-end**, without any intermediate steps.

Areas of Linguistics



speech

text

phonetics

orthography

phonology

morphology

lexemes

"shallower"

syntax

semantics

"deeper"

pragmatics

discourse

Areas of Linguistics

- Phonetics and phonology **the inventory of sounds, and how they are used in the language**
- Orthography **the writing system**
- Morphology **the study of words**
- Syntax **the study of how words go together form grammatical sentences**
- Semantics **the meaning**
- Pragmatics **the extra information beyond the meaning**
- Discourse **multiple sentences in order, either a monologue (one speaker) or dialog (more than one)**

Why is NLP hard?

Computers can understand programming languages.
Could we do the same thing for natural language?

Why is NLP hard?

“At last, a computer that understands you like your mother.”

Why is NLP hard?

“At last, a computer that understands you like your mother.”

- (3 Minutes) On your whiteboard or in the chat, write
  - as many ways of interpreting this sentence as you can think of
  - how would you (as a human) know which one to choose? what specifically would you use to decide?
- (3 minutes) Discuss with the person next to you or with the others in Zoom

Ambiguity

“At last, a computer that understands you like your mother.”

1. It understands you as well as your mother understands you.
2. It understands (that) you like your mother.
3. It understands you as well as it understands your mother.

1 and 3: Does this mean well, or poorly?

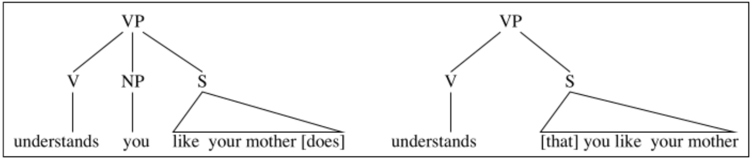
Ambiguity at Many Levels

At the acoustic level

1. "… a computer that understands you **like your** mother."
2. "… a computer that understands you **lie cure** mother."

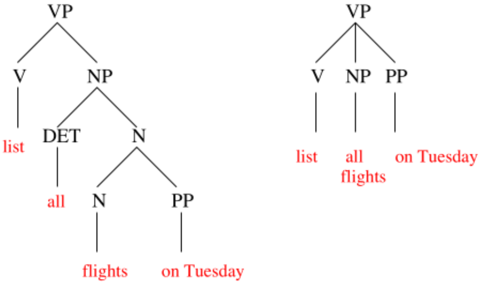
Ambiguity at Many Levels

At the syntactic level:



Different structures lead to different interpretations.

More Syntactic Ambiguity



Ambiguity at Many Levels

At the semantic (meaning) level:
Two definitions of "mother"

- a woman who has given birth to a child
- a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar

This is an instance of word sense ambiguity

More Semantic Ambiguity

At the semantic (meaning) level:

- They put money in the *bank*

  = buried in mud?

- I saw her duck with a telescope

Ambiguity at Many Levels

At the discourse (multi-clause) level:

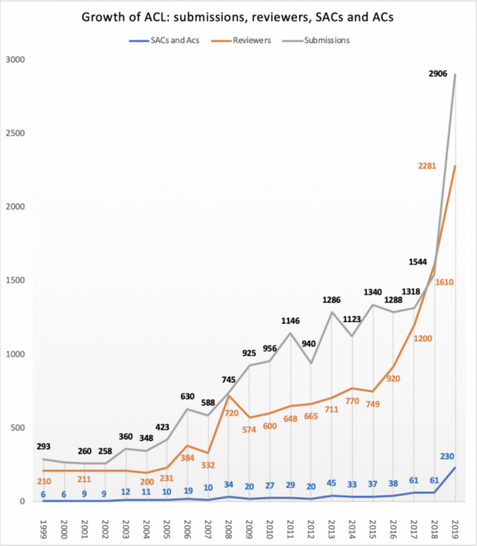
- Alice says they've built a computer that understands you like your mother
- But she . . .

  . . . doesn't know any details

  . . . doesn't understand me at all

This is an instance of anaphora, where she co-referees to some other discourse entity

Large growth in NLP in recent years



History of NLP

- 50's-90's: Rule-based methods
    - Drawbacks: time-consuming, usually doesn't scale (with some exceptions)
    - Example: SYSTRAN MT system, powered babelfish.com
- 90's-2010's: Machine learning and statistical methods
    - Drawbacks: usually lower performance than deep learning when lots of data
- 2014-present: Deep learning methods
    - Open issues: data-hungry, black-box, brittle, overfits quirks in datasets

History of NLP

- 50's-90's: Rule-based methods
  - Drawbacks: time-consuming, usually doesn't scale (with some exceptions)
  - Example: SYSTRAN MT system, powered babelfish.com
- 90's-2010's: Machine learning and statistical methods
  - Drawbacks: usually lower performance than deep learning when lots of data
- 2014-present: Deep learning methods
  - Open issues: data-hungry, black-box, brittle, overfits quirks in datasets
- Lots of progress, still a long way to go
- Older methods are important to know, can work better in certain situations. Breadth of knowledge helps drive progress

NLP applications are now commonplace

- Spam email filtering
- Google translate
- Built-in recommender systems (in Amazon, Ebay, Netflix, etc)
- Siri, Amazon Alexa
- Auto-completion suggestions
- Grammar checking
- Automatic essay grading (used by ETS)
- Inappropriate social media post filtering
- Fake news detection
- Lots you probably don’t even realize!

Ethics

- With widespread use, NLP has potential ethical issues such as
  - Bias
  - Censorship
  - Privacy
  - Security
- These issues are hot topics, very active area of research

Ethics

- Bias amplification: systems exacerbate real-world bias rather than correct for it
- Exclusion: underprivileged users are left behind by systems
- Dangers of automatic systems: automating things in ways we don’t understand is dangerous
- Unethical use: powerful systems can be used for bad ends

Relation of NLP to other fields

- Speech (both recognition and generation) are separate, not an NLP tasks
- Machine learning (computers learn from experience or examples)
- Linguistics (the study of language).
- Computational linguistics (CL)
    - Sometimes synonymous with NLP
    - CL often has larger emphasis on linguistics and linguistic theories

References I

Jacob Eisenstein. *Natural Language Processing*. 2018. URL https://github.com/jacobeisenstein/gt-nlp-class/raw/master/notes/eisenstein-nlp-notes.pdf.