CS 445 - NATURAL LANGUAGE PROCESSING

Instructor: Dilara Keküllüoğlu
Fall 2024 – Week 1

Course Organization

- 2 lectures per week M 9:40-11:30, F 14:40-15:30
- SUCourse will be used actively (as I learn)
- Instructor: Dilara Keküllüoğlu, Room 1089, dilara.kekulluoglu@sabanciuniv.edu
- TAs: Ayşegül Rana Erdemli, Kerem Aydın, Semih Gülüm

Course Organization

- Office Hours: Tuesdays 15:00-17:00 (welcome to drop by anytime – for non-CS445 related things too)
- TA Office Hours: TBA
- Emails - Please do not expect a response outside of working hours and I will not expect that from you too. – e.g. do not send emails needing immediate response on Friday 8pm. You might need to wait until Monday afternoon.

Course References

- Reference: Daniel Jurafsky and James H. Martin, *Speech and **Language Processing: An Introduction to Natural Language Processing**, Computational Linguistics, and Speech Recognition* (3rd edition online). https://web.stanford.edu/~jurafsky/slp3/
- Recommended: Steven Bird, Ewan Klein, and Edward Loper, ***Natural Language Processing with Python***. https://www.nltk.org/book/
- Slide references:
  - **Foundations of Natural Language Processing Course**, University of Edinburgh (Ivan Titov, Alex Lascarides, Philipp Koehn, Sharon Goldwater, Shay Cohen, Khalil Sima'an),
  - **Speech and Language Processing** (Daniel Jurafsky and James H. Martin)

Course Structure

- 3 or 4 small **coding assignments** around the concepts taught in lectures (20%)
  - Basic Python knowledge is required
  - Will get more complicated as we go but will have same weight
- **Midterm Exam** (40%)
- **Group project** (40%)
  - Will share more information in the following lectures.
- Regular attendance is expected but no attendance checks will be done.

Course Structure

- After a few weeks, I plan to use some parts of the Friday session for **live coding** the concepts we learn.
- You have **three points** you can use for **late submissions**. Every day you are late for a coding assignment will remove a point from your balance. You can use these three points as you wish for the **coding assignments**.
- Letter grades will be assigned based on **curve**, considering the distribution of the grades in the class.

Course Outline

- Week 1: Introduction to Natural Language Processing
- Week 2: Regular Expressions and Text Normalization
- Week 3: Text Analysis and Edit Distance
- Week 4: Language Models
- Week 5: Text Classifications
- Week 6: Semantics, Embeddings, and Sentiment
- Week 7: Morphology, POS Tagging and Named Entities
- Week 8: Hidden Markov Models
- Week 9: Context-Free Grammars
- Week 10: Parsing
- Week 11: Content Review and Midterm
- Week 12: Deep Learning for NLP
- Week 13: Conversational Agents
- Week 14: NLP as a tool in Current Research

Learning Goals (Global)

1. **Describe** the statistical properties of text in natural language.
2. **Implement** programs that can process textual data and extract valuable information from it.
3. **Apply** well-known language processing techniques to text.
4. **Explain** the significance and principles of language modeling.
5. **Assess** the quality of natural language processing models applied to text.

Any questions about the course structure?

Tophat Exercise

- Exercise for me as I learn Tophat =)
- Please write one of your hobbies in the question (one word) so we can see which hobbies do we have in the class.

Learning Goals (Week 1)

1. **Understand** the course structure and information
2. **Define** natural language processing and the uses in our lives
3. **Describe** challenges in NLP
4. **Describe** corpora and its uses
5. **Decide** whether you want to take this class =)

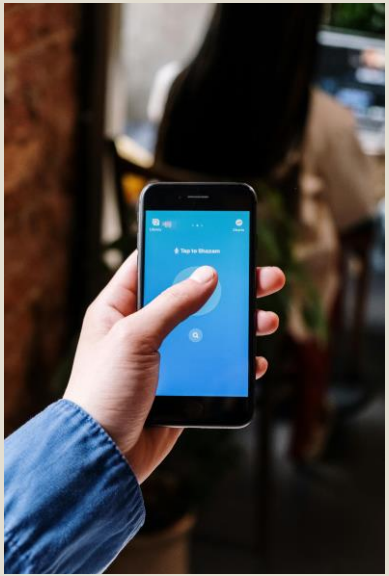
Tophat

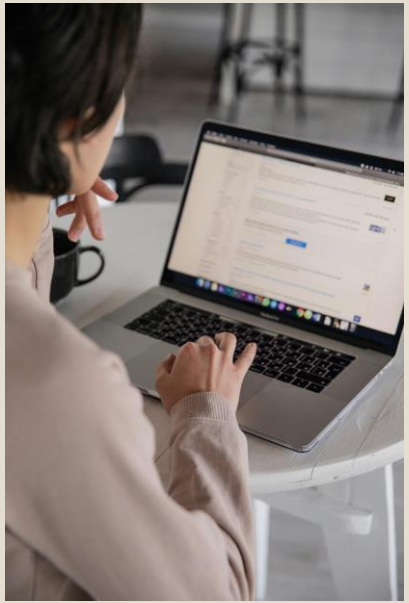
NLP provides a series of computational methods to analyze and generate languages people use to communicate.

What do you think Natural Language Processing is used for?

Some use cases

- Internet Search
- Sentiment analysis
- Speech recognition
- Machine translation
- Summarization
- …





Photos: Katrin Bolovtsova, cottonbro studio

Tophat

Which apps/products you use utilize NLP?

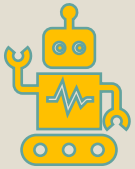
Popular apps with NLP

- Internet Browsers
- Youtube
- Netflix
- Twitter
- ...

Probably most of the apps you use!

Photos: pixabay

An example



A system that gives travel ideas.

System: Good morning! How can I help?

User: Can you give me ideas for my next trip? I think it will be in summer and I want to stay in Türkiye.

S: Do you like swimming or hiking more?

U: I love both actually but I like swimming more.

S: For swimming, you can choose one of the mediterranean coast cities. For example, Fethiye which also has nice beach trails.

…

An example - cont.

What should system **extract** from the text before recommendation?

- Timing - Season, length
- Activity preferences
- …

What kind of **problems** do you think the system will face?

- Supporting many scenarios
- Understand intent of the user correctly
- Many levels of structure in open text NLP systems

Words

This is a simple sentence.

WORDS

Example from Ivan Titov.

Morphology

This is a simple sentence.

be
3sg
present

WORDS

MORPHOLOGY

Example from Ivan Titov.

Part of Speech

DT VBZ DT JJ NN

This is a simple sentence.

be
3sg
present

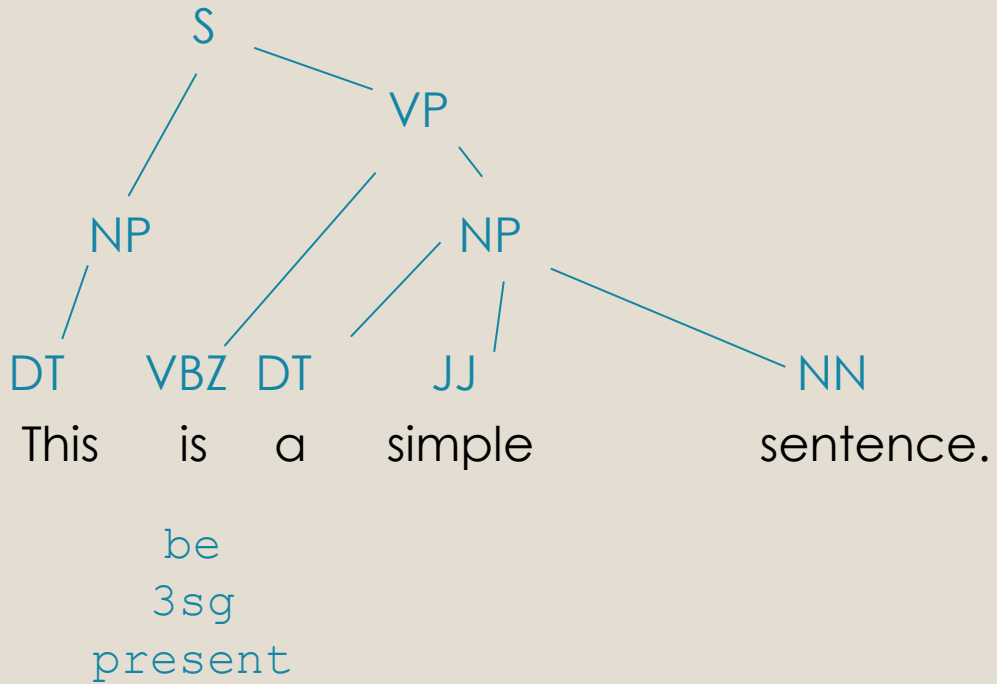
PART OF SPEECH

WORDS

MORPHOLOGY

Example from Ivan Titov.

Syntax



SYNTAX

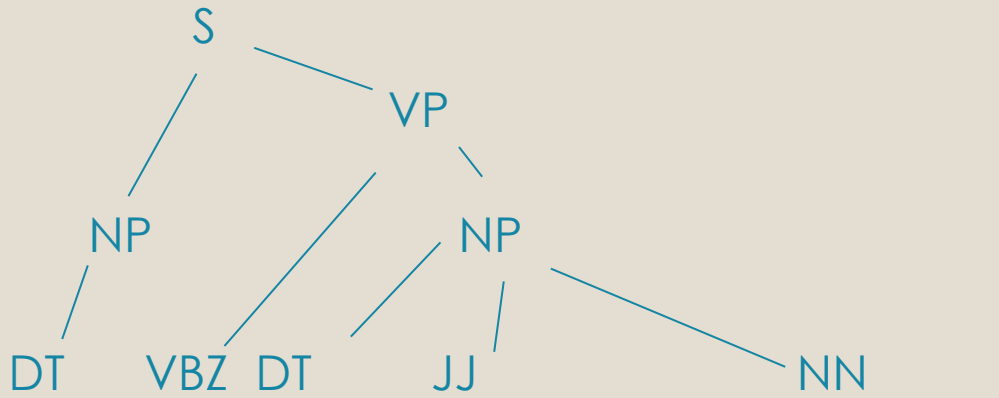
PART OF SPEECH

WORDS

MORPHOLOGY

Example from Ivan Titov.

Semantics



be
3sg
present

having
few
parts

string of words
satisfying the
grammatical rules
of a language

SYNTAX

PART OF SPEECH

WORDS

MORPHOLOGY

SEMANTICS

Example from
Ivan Titov.

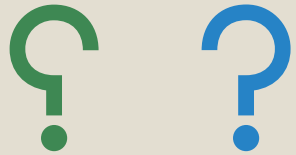
Why is NLP hard?

- Correctly identifying the **layers** of the sentences is not trivial.
- The language can be **variable** and **ambiguous**.

- A meaning can be said multiple ways. (Variability)
  - She took the bus.
  - She came here by bus.

- A word could mean multiple things. (Ambiguity)
  - She took the bus. - rode
  - She took my attendance. - recorded
  - She is taken with the cat's elegance. - fascinated

Ambiguity



- Homophones: hear - here
- Word sense: block (rock or neighbourhood)
- Part of speech: smell (verb or noun?)
- Syntactic structure: I saw a man with a telescope.
- Quantifier ambiguity: Every student did not pass the exam.
- Multiple meanings: I saw her duck.
- Reference: The girl told her mom about the fight. She was upset.
- Discourse: I will not cook today. Alice ordered take out.

Ambiguity Examples

Think about the following examples individually and decide what kind of ambiguity is present from the list.

- Bank
- British Left Waffles on Falkland Islands
- The meeting is cancelled. Nicholas is not coming to the office today.
- Duck
- John searches for a dog with microscope.

- **Homophones**: hear - here
- **Word sense**: block (rock or neighbourhood)
- **Part of speech**: smell (verb or noun?)
- **Syntactic structure**: I saw a man with a telescope.
- **Quantifier ambiguity**: Every student did not pass the exam.
- **Multiple meanings**: I saw her duck.
- **Reference**: The girl told her mom about the fight. She was upset.
- **Discourse**: I will not cook today. Alice ordered take out.

Ambiguity Examples

- Bank - Word Sense or part of speech
- British Left Waffles on Falkland Islands - Multiple meanings
- The meeting is cancelled. Nicholas is not coming to the office today. - Discourse
- Duck - Word sense or part of speech
- John searches for a dog with microscope. - Syntactic Structure

Zipf's Law

- Variability and ambiguity make NLP difficult
- Another challenge is word **sparsity** due to Zipf's Law
- Frequent words dominate the data (the, of, to, and, a)
- There are many words that rarely show up or show up just once
- There is a consistent pattern between the frequency and the rank of the words across languages.

Really interesting phenomenon of natural languages!!

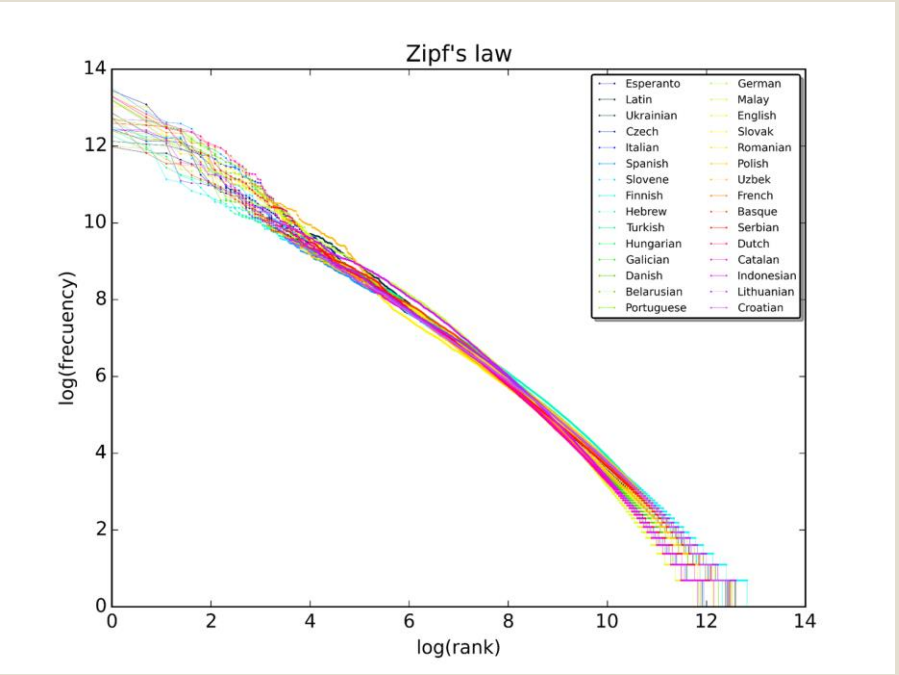
Zipf's Law

$$f \times r \approx k$$

- f = frequency of the word
- r = rank of the word
- k = constant number

Let's say most frequent word appears 10000 times in a document, the word in the second rank will appear around 5000 times in this document.

- Adding more documents will not solve sparsity problem.



SergioJimenez, A plot of the rank versus frequency for the first 10 million words in 30 Wikipedias (dumps from October 2015) in a log-log scale.

Robustness

- Another challenge!
- People have typos, use informal language, abbreviations that change with time
  - ijbol - (it was rofl in my time =))
  - Koreans adding ㅇ at the end of syllables - (OpenAI o1 seems to solve this!)
- Systems trained on perfect grammatical representation might fail on social media text.

Other Challenges

- Context Dependence
    - Meaning can change depending on the context. If you are in a hospital, he drew blood means different compared to a boxing match.
- Unknown Representation
    - It is not easy to represent this knowledge in a way that computers can reason. – even humans do not exactly know.
- Diverse Languages
    - Russian - really difficult morphological analysis
    - Japanese - no spaces between words, mixed alphabets

Challenges of NLP - Overview

1. Variability
2. Ambiguity
3. Sparsity
4. Robustness
5. Context dependence
6. Unknown representation
7. Diverse languages
8. ...

Corpora

- Collection of documents for NLP tools to use

- “corpus: a body of utterances, as words or sentences, assumed to be representative of and used for lexical, grammatical, or other linguistic analysis.” - dictionary.com

Corpora

- We use corpora to understand and model the languages
- Corpora can also have labels - language, author, source
  - annotations by people - sentiment, stance, sarcasm, etc.
- NLTK library in Python has corporas you can reach easily - you can try and play around to see

```
>>> import nltk
>>> from nltk.corpus import movie_reviews
>>> movie_reviews.words()
['plot', ':', 'two', 'teen', 'couples', 'go', 'to', ...]
>>>
```

What is the use of corpora?

To train and evaluate systems' performance

- Learning systems can be trained on the carefully curated corpora
- Evaluation and benchmarking should be performed on the same dataset to get reliable comparison.

Corpora Selection

- Selection of corpora for the task is important
- For sentiment analysis, you can use movie reviews corpus with sentiment labels
- For social media related tasks, it does not make sense to use news corpus vice versa
- Language of the corpora

Movie Reviews Corpus

'trees lounge is the directoral debut from one of my favorite actors , steve buscemi . he gave memorable performences in in the soup , fargo , and reservoir dogs . now he tries his hand at writing , directing and acting all in the same flick . the movie starts out awfully slow with tommy (buscemi) hanging around a local bar the " trees lounge " and him pestering his brother…

Category: positive

'say , tell me if you've seen this before : a crisis on-board a commercial airliner causes a stewardess to have to fly and land the plane herself . airport '97 anyone ? ray liotta is a psychotic serial killer being transported from new york to california on christmas eve . amazingly , on what would seemingly be a busy day of travel on one of the most flown routes , only about six other passengers are on the flight . anyway , they take off , liotta escapes and kills all the police and the pilots , and stewardess lauren holly locks herself in the cockpit to fly the plane . the story is beyond routine , the script is embarrassing (at one point , this jumbo jet is flying completely upside down) , the characters are worthless , and the performances are annoying . surprisingly , co-writer steven e . de souza actually wrote the first two " die hard " movies ! " turbulence " takes place at christmas time , yet the film was released a few days after the holidays . brilliant marketing , as no one cares about anything having to do with christmas after december 26th . the studio knew they had a bomb , and purposely dumped it out when the fewest number of people would see it .'

Category: negative

Brown Corpus

The Fulton County Grand Jury said Friday an investigation of Atlanta’s recent primary election produced ``no evidence‘’ that any irregularities took place…

('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'), ('Grand', 'JJ-TL'), ('Jury', 'NN-TL'), ('said', 'VBD'), ('Friday', 'NR'), ('an', 'AT'), ('investigation', 'NN'), ('of', 'IN'), ("Atlanta's", 'NP$'), ('recent', 'JJ'), ('primary', 'NN'), ('election', 'NN'), ('produced', 'VBD'), ('``', '``'), ('no', 'AT'), ('evidence', 'NN'), ("''", "''"), ('that', 'CS'), ('any', 'DTI'), ('irregularities', 'NNS'), ('took', 'VBD'), ('place', 'NN'), ('.', '.')

Corpora Characteristics

Corpora features will change depending on some factors.

- Who collected the data and for what purposes?
  - Researchers from US for medical research
  - Novels in Türkiye - for language studies
  - Biomedical companies - bacteria classification
- How was the data collected?
  - Web scraping
  - Scanning documents
  - Surveys
- When was the data collected?

Corpora Characteristics

- Who labeled the data? - Annotator bias
  - Language proficiency
  - Nationality
  - Occupation
- What is the language of the data?
  - Turkish, English, Russian, Japanese
- What is the domain of the data?
  - Health, Politics, Hobbies, News
- Can we use and distribute the data?
  - Was the collection ethical?
  - Can we use and distribute the results freely?

Biased Corpora

- Important to know all of these features might introduce bias to the dataset
- Consider a biomedical corpus created by collecting only the test results of one population - ages 18-25 in US.
    - Systems trained on this corpora will not work as well for another population - over 60 in Türkiye.
- People who collect and annotate bring their own bias into the dataset too.
- Dataset collected in 1980s will not be reflective of the status of our times.

Corpora curation and selection is an important problem.

Bias in NLP

Learning Goals (Week 1) - revisited

1. **Understand** the course structure and information
2. **Define** natural language processing and the uses in our lives
3. **Describe** challenges in NLP
4. **Describe** corpora and its uses
5. **Decide** whether you want to take this class =)

Following lectures

- Regular expressions
- Text normalization

Further resources

- Datasheets for Datasets (Gebru et al.) - https://www.microsoft.com/en-us/research/uploads/prod/2019/01/1803.09010.pdf
- Al & Bias Series, UCLA Institute for Technology (~18 minutes)
  - https://www.youtube.com/watch?v=8tfKdxo8Rj8
  - https://www.youtube.com/watch?v=FD-4yC95iZY
  - https://www.youtube.com/watch?v=xvb_A_qzXo4