Natural Language Processing

1. Introduction and logistics

Paolo Torroni

Fall 2022

Notice

Credits

The present slides are largely an adaptation of existing material, including:

- slides by Jurafsky & Martin
- slides by Manning
- slides by Potts & MacCartney
- slides by Black & Mortensen

I am especially grateful to these scholars.

Downloading and ~~sharing~~

A copy of these slides can be downloaded from virtuale and stored for personal use only. Please do not redistribute.

Table of Contents

Welcome to
Natural Language Processing

ELIZA

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
It's true. I am unhappy.
DO YOU THINK COMING HERE WILL HELP YOU NOT TO
BE UNHAPPY
I need some help, that much seems certain.
WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
Perhaps I could learn to get along with my mother.
TELL ME MORE ABOUT YOUR FAMILY
My mother takes care of me.
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU
My father.
YOUR FATHER
You are like my father in some ways.

J Weizenbaum, ELIZA: a computer program for the study of natural language communication between man and machine, Communications of the ACM, January 1966

Chirpy Cardinal

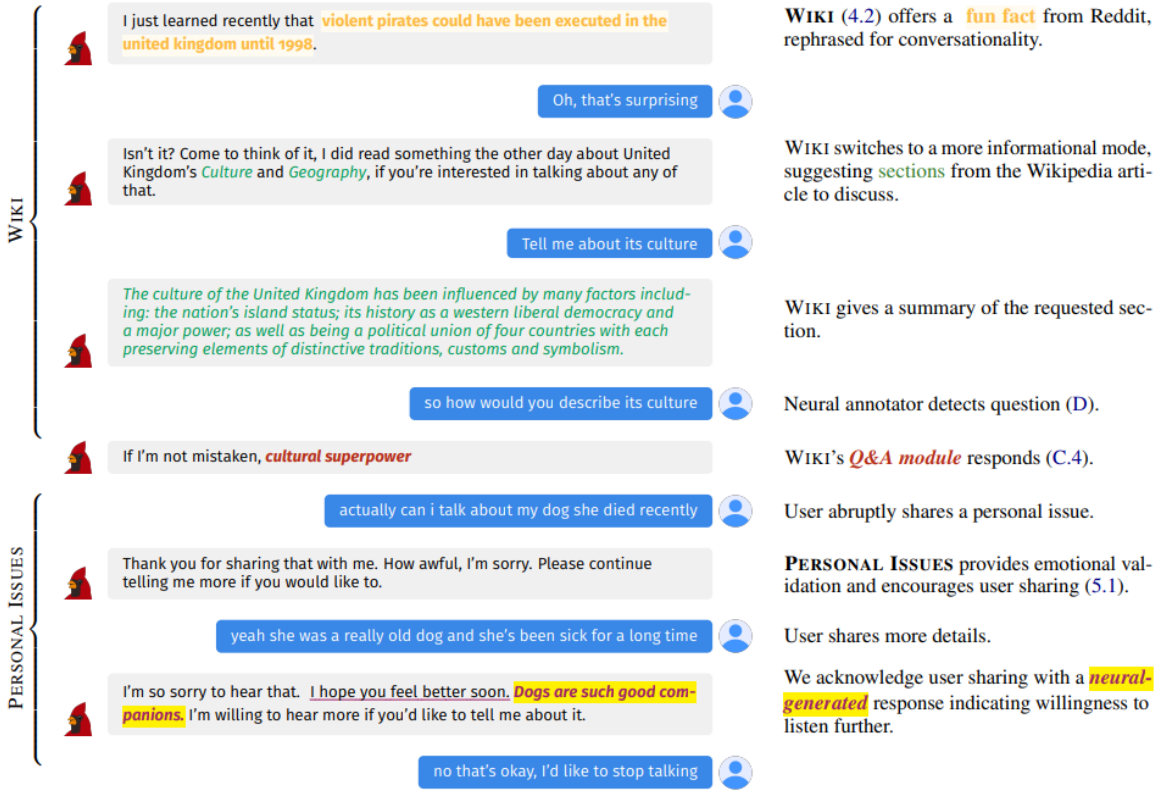


Figure 1: An example dialogue. To respect users' privacy, this is a conversation with an author, not a real user. We model dialogue as a series of subconversations (left side), whose *topics* are parsed and tracked by a neural entity linker. Each subconversation is managed by a RG, whose "scaffolded" architecture comprises hand-written treelets (plain text) incorporating numerous dynamic elements, ranging from ***neural generation*** to *retrieval from Wikipedia* to **neurally rephrased fun facts**. Prefix-based generation provides controllability, especially for sensitive topics like personal issues.

E A Chi et al., Neural Generation Meets Real People: Building a Social, Informative Open-Domain Dialogue Agent, arXiv 2022

What is NLP?

Automating the analysis, generation, and acquisition of human ("natural") language

- Analysis (or "understanding" or "processing" . . .)
- Generation
- Acquisition

Why NLP?

- Answer questions using the Web
- Translate documents from one language to another
- Do library research; summarize
- Manage messages intelligently
- Follow directions given by any user
- Fix your spelling or grammar
- Flag fake news or hate speech
- Write poems or novels
- Listen and give advice
- Estimate public opinion
- Read everything and make predictions
- Interactively help people learn, help disabled people, help refugees/disaster victims, document or reinvigorate indigenous languages, ...

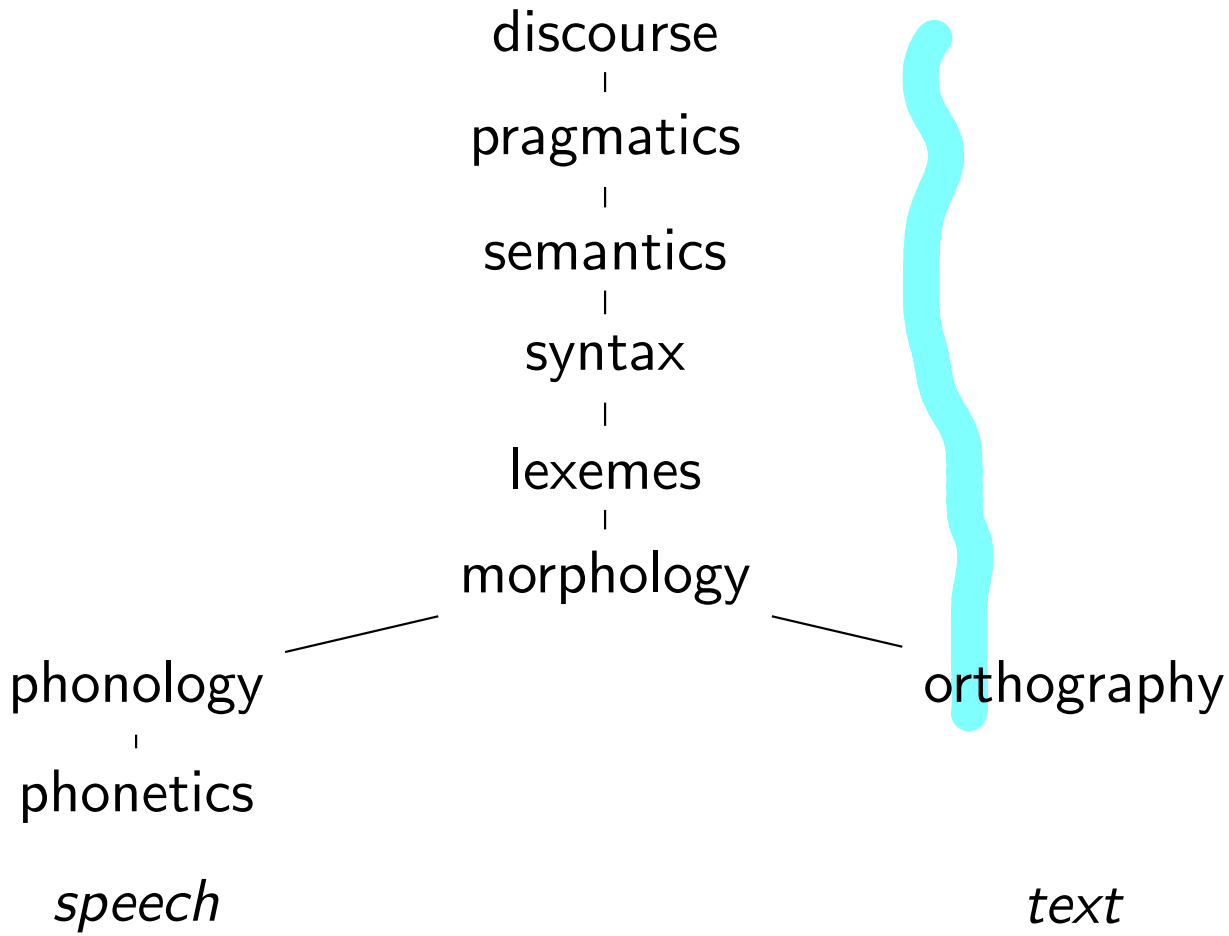
What is NLP? (more detail)

Automating the analysis, generation, and acquisition of human ("natural") language

- Analysis (or "understanding" or "processing" ...): input is language, output is some **representation** that supports useful action
- Generation: input is that **representation**, output is language
- Acquisition: obtaining the **representation** and necessary algorithms, from knowledge and data

Representation?

Levels of Linguistic Representation



Why It's Hard

- Input is likely to be noisy
- Linguistic representations are **theorized constructs**; we cannot observe them directly
- Difficult to obtain training data for each aspects
- The mappings between levels are extremely complex
- Appropriateness of a representation depends on the application

dataset → corpus / corpora } (linguistic) resources

Ambiguity

- Each string may have many possible interpretations at every level
- Correct resolution of the ambiguity depends on the **intended meaning**, which is often inferable from context
- People are good at linguistic ambiguity resolution
- Computer not so
  - How do we represent sets of possible alternatives?
  - How do we represent context?

Complexity of Linguistc Representatons

- **Richness**: there are many ways to express the same meaning, and immeasurably many meanings to express. Lots of words/phrases
- Each level interacts with the others
- Tremendous diversity in human languages
  - Languages express the same kind of meaning in different ways
  - Some languages express some meanings more readily/often

Models for NLP

Model: an abstract, theoretical, predictive construct. It includes:

- a (partial) representation of the world
- a method for creating or recognizing worlds
- a system for reasoning about worlds

NLP uses *many* tools for modeling

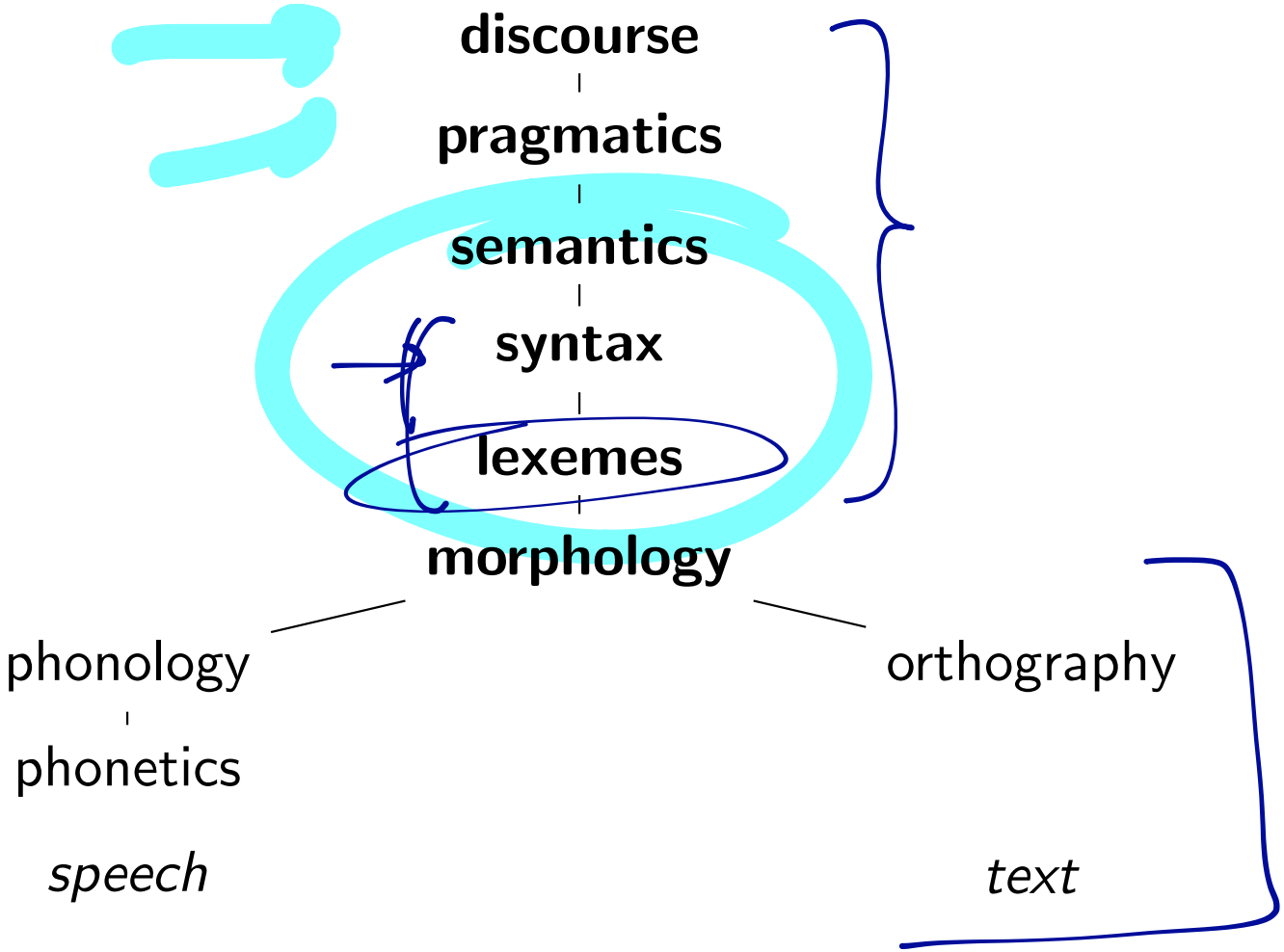
Surprisingly shallow models work fine for some applications

This course is meant to introduce some formal tools that will help you navigate the field of NLP

Many Related Disciplines

- **Natural Language Processing** is a branch of AI focused on the technology of processing language (includes NLU, NLG)
- **Computational Linguistics** is focused on the technology to support/implement linguistics; e.g., study of language change over time
- **Information Retrieval** is about obtaining information system resources that are relevant to an information need from a collection of those resources; e.g., Web search engines
- **Text Mining & Big Data Analytics** are subtypes of Data Mining, focused on search and retrieval of information from large bodies of text; e.g., information discovery; biomedical text mining supporting clinical needs

Let's Examine Some of the Levels



Morphology

Analysis of words into meaningful components

Spectrum of complexity across languages

- *Analytic* or *Isolating* languages (e.g., English, Chinese)
- *Synthetic* languages (e.g., Turkish, Finnish, Hebrew)

Examples

- evlerindekilerin
- 没人需要空格来分隔单词
- unfriend, Obamacare, Bill's

Lexical Analysis

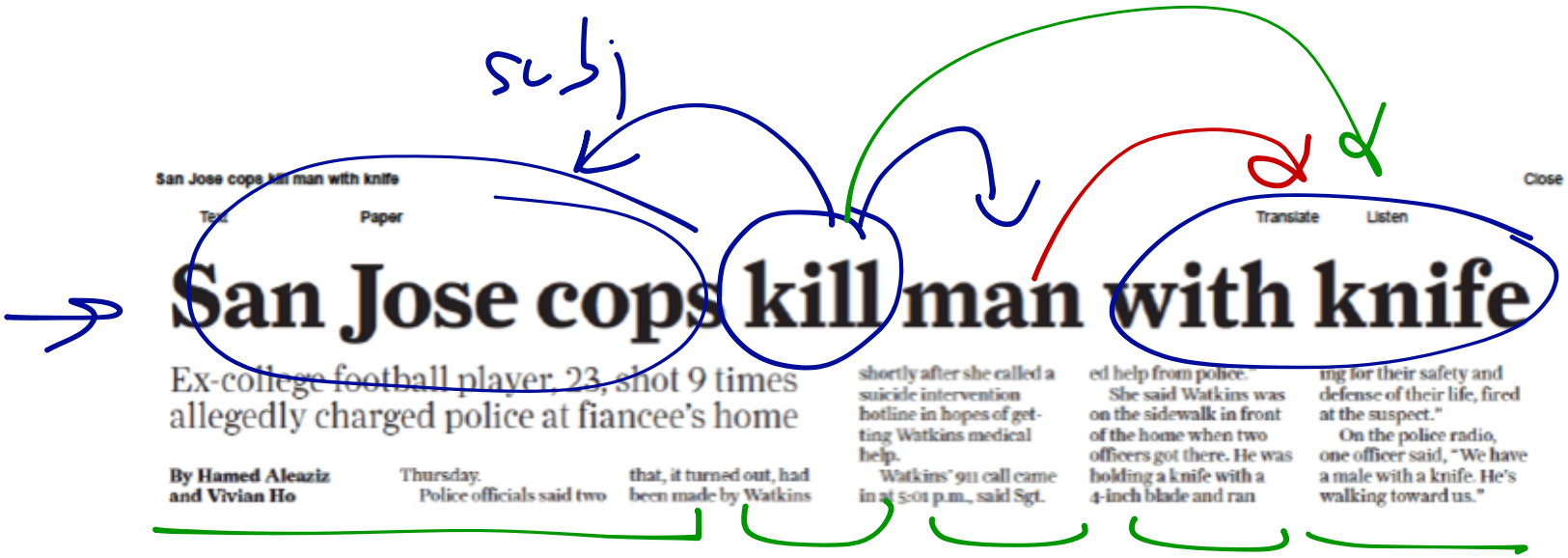
- Normalize and disambiguate words
- Words with multiple meanings: bank, mean
  - Extra challenge: domain-specific meanings
- Multi-word expressions
  - make ... decision, take out, make up, ...
- For English, part-of-speech tagging is one very common kind of lexical analysis

Syntax

- Transform a sequence of symbols into a hierarchical or compositional structure
- Closely related to linguistic theories about what makes some sentences well-formed and others not. For example:
  - I want a flight to Tokyo
  - I want to fly to Tokyo
  - I found a flight to Tokyo
  - I found to fly to Tokyo

Ambiguities

Prepositional phrase attachment ambiguity



Examples taken from http://web.stanford.edu/class/cs224n/

Ambiguities

Prepositional phrase attachment ambiguity

Examples taken from http://web.stanford.edu/class/cs224n/

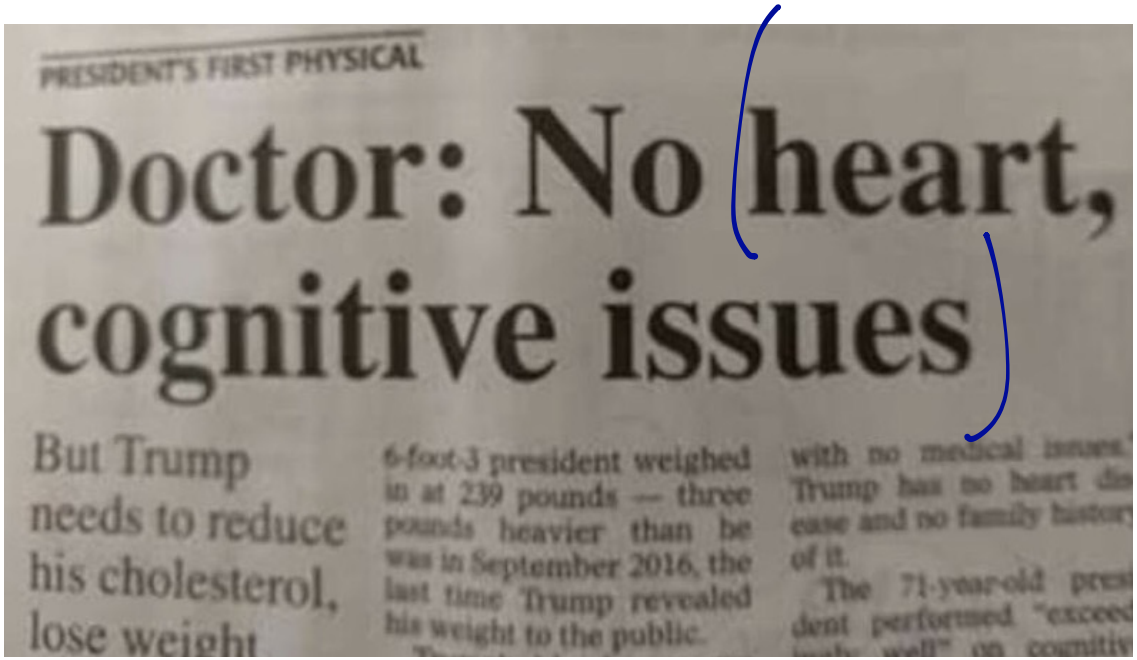
Ambiguities

Verb phrase attachment ambiguity

Examples taken from http://web.stanford.edu/class/cs224n/

Ambiguities

Coordination scope ambiguity


PRESIDENT'S FIRST PHYSICAL

Doctor: No heart, cognitive issues

But Trump needs to reduce his cholesterol, lose weight

6-foot-3 president weighed in at 239 pounds — three pounds heavier than he was in September 2016, the last time Trump revealed his weight to the public.

with no medical issues.' Trump has no heart dis ease and no family history of it.

The 71-year-old presi dent performed "exceed ... well" on cognitiv

Examples taken from http://web.stanford.edu/class/cs224n/

Ambiguities

Adjectival modifier ambiguity

MENTORING DAY

Students get first hand job experience

By Gale Rose
grose@pratttribune.com

Eager students invaded businesses all over Pratt experience what it would be like to work at those 40 businesses. They asked questions and got some hands on experience with imal Veterinarian Clinic for their business. Students got a tour of the facility, learned what happens in an examination. imals. Patton likes al kinds of animals and said she learned a lot from the experience. Watching the snake eat the mouse im-

Examples taken from http://web.stanford.edu/class/cs224n/

Ambiguities Explode Combinatorially

What possible syntactic analyses?

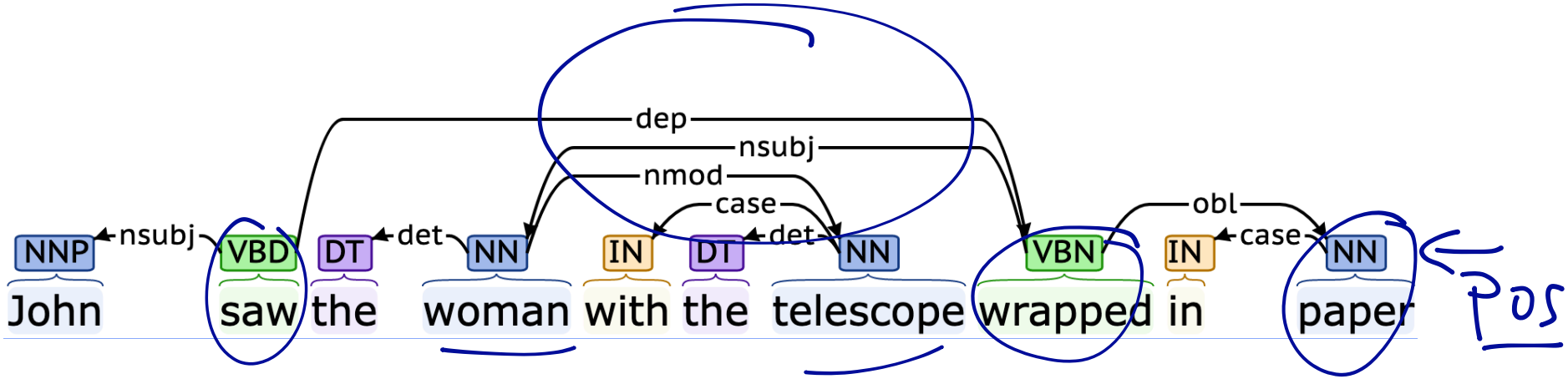
- John saw the woman with the telescope wrapped in paper
- John saw the woman with the telescope wrapped in paper
- John saw the woman with the telescope wrapped in paper
- John saw the woman with the telescope wrapped in paper

...

Ambiguities Explode Combinatorially

What possible syntactic analyses?

- John saw the woman with the telescope wrapped in paper



Credits: Stanford Core NLP online dependency parser https://corenlp.run/

Semantics

In this country a woman gives birth every 15 minutes.
Our job is to find that woman, and stop her. *(Groucho Marx)*

- Mapping of natural language sentences into domain representations
  - Robot command, database query, formal logic expression, ...
- Scope ambiguities
  - A seat is available to every customer
  - A web site is available to every customer
- Going beyond specific domains is a goal of AI

Pragmatics

- Any *non-local* meaning phenomena
  - "Can you pass the salt?"
    - "Are you 18?" "Yes, I'm 25."

Discourse

I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation.

Five score years ago, a great American, in whose symbolic shadow we stand today, signed the Emancipation Proclamation. This momentous decree came as a great beacon light of hope to millions of Negro slaves who had been seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of their captivity.

But one hundred years later, the Negro still is not free. One hundred years later, the life of the Negro is still sadly crippled by the manacles of segregation and the chains of discrimination. ....

(Martin Luther King, Jr.)

- Structures and effects in related sequences of sentences
- Texts, dialogues, multi-party conversations

In This Course...

- Syntactic and lexical analysis
  - edit distance, noisy channel model, part-of-speech tagging, grammars and parsing
- Semantics
  - language models, vector semantics, embeddings
- Tools
  - basic text processing tools
  - neural architectures for sequence processing: RNNs, CNNs, sequence-to-sequence models, attention, transformers
- Applications
  - spelling correction, text classification, machine translation, information extraction, question answering, dialogue systems and chatbots, sentiment analysis and argument mining

Course Organization and Logistics

Teaching Staff and Audience

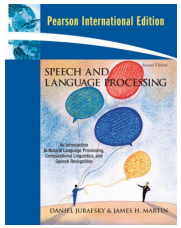
Teaching staff

- Instructor: Paolo Torroni
- Tutors: Andrea Galassi & Federico Ruggeri
- Language Technologies Lab

Intended audience

- 2nd-year students of the Masters' Degree in AI
- Anyone with an interest in NLP and who is familiar with
  - statistical and mathematical methods for AI
  - machine learning and deep learning concepts, methodologies and tools
  - the Python programming language
  - knowledge representation frameworks and methods

Learning Resources



- Course slides
- Textbook
  - Speech and Language Processing, by Dan Jurafsky and James H. Martin. 3^{rd} Ed. online draft.
- Selected articles, blogs, and other online resources
- Further reference textbooks
  - Natural Language Processing, by J Eisenstein. Online draft.
  - Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit, by S Bird, E Klein, and E Loper.
  - Natural Language Processing with Transformers, by L Tunstall, L von Werra, and T Wolf. O'Reilly.
  - Natural Language Processing with PyTorch, by D Rao and B McMahan. O'Reilly.
- Software
  - NLTK, Colab, PyTorch, Huggingface

Assessment

- Two **assignments**
  - Each graded on 6-point scale
  - If on-time: + 0,5 points (1 week grace period for one late assignment)
  - Exceptionally: + 0,5 points if solution is particularly clever
  - To be done **in groups of 3/4**
- A **project**
  - Methodology and implementation (10 points)
  - Report (4 points)
  - Oral discussion (4 points)
  - To be done **in groups of 3/4**
- Groups formation: **virtuale forum**
- Changing from one group to another is permitted (even welcome)
- Final grade is sum of all points earned
  - Cum laude if sum ≥ 31

Project Work & Thesis

- Project Work in NLP (3 cfu). Some options:
  - explore a foundational aspect
  - explore NLP methods for low-resource languages
  - work on a challenging application
  - work on a research topic
  - prepare a literature survey on a research topic
  - significantly expand project discussed in this course
- Thesis abroad
  - watch out for grants
  - helpdesk: International mobility office.
- Internship + thesis at a company
- Work with our group on a research topic

Contacts

- **Email**: p.torroni@unibo.it, a.galassi@unibo.it, federico.ruggeri6@unibo.it
  - *Hint*: To maximize answer speed, keep all teaching staff in Cc, and write "NLP course" in the Subject
- Office time on Teams by appointment (email)
- Phone: 051 2093767
- Contact details of teaching staff on Virtuale

Questions?