Foundations of Natural Language Processing

Lecture 1: Introduction

Mirella Lapata
School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

Slides based on content from: Philipp Koehn, Alex Lascarides, Sharon Goldwater, Shay Cohen, Khalil Sima'an, Ivan Titov

What we'll do today

- Welcome to Foundations of Natural Language Processing!
- Make sure you are in the right class/room.
- We'll cover course logistics.
- We'll get started on *what* is NLP and *why* it is hard.

Background needed for this course

We assume you are familiar with most/all of the following:

- Basic Python programming
- Finite-state machines, regular languages, context-free grammars
- Dynamic programming (e.g., edit distance, Viterbi, and/or CKY algorithms)
- Concepts from machine learning (e.g., estimating probabilities, making predictions based on data)
- Probability theory (conditional probabilities, Bayes' Rule, independence and conditional independence, expectations)
- Vectors, logarithms, linear algebra, matrix operations
- Some basic linguistic concepts (e.g., parts of speech)

Where we are headed

INF2-iads discussed ideas and algorithms for NLP from a largely formal, algorithmic perspective. Here we build on that by:

- Focusing on real data with all its complexities.
- Discussing some of the NLP techniques in more depth.
- Introducing many tasks and technologies that didn't fit into the Inf2-iads story.
- By the end of the class, you'll know how to make your own ChatGPT.

Course organization

- Course organizer: Ivan Titov
- Lecturers: Mirella Lapata and Ivan Titov
- 3 lectures per week (Tue, Wed & Fri, 12:10–13.00)
- We will use Learn and Drupal for slides, lectures, labs, assignments, due dates, etc
- Labs: two groups, every two weeks (on Fridays, 13:10–14:30 or 14:30–16.00)
- Tutorials: in small groups, every two weeks (on Thursdays, 12:10–13:00, 13:10–14:30, or 14:30–16:00); tutorials start next week
- Course discussion forum: Piazza.

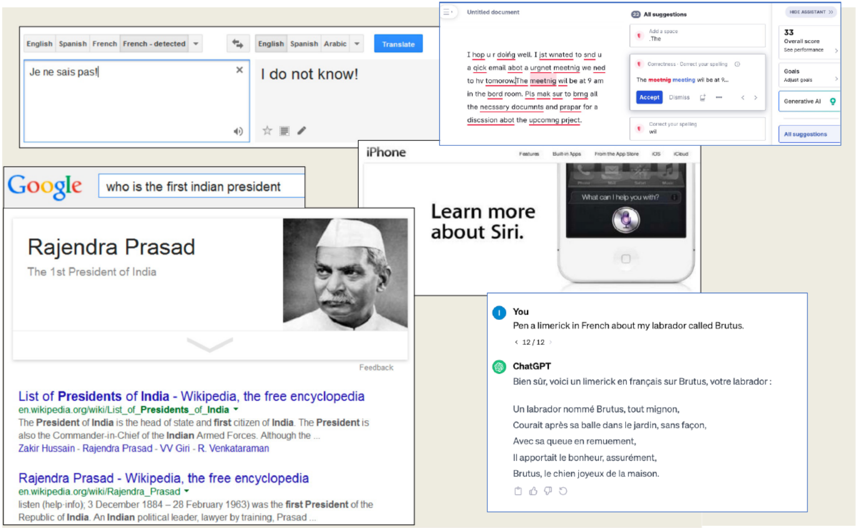
Check Learn for all the links and up-to-date information

Outside work required

In addition to attending lectures, you are expected to keep up with:

- Readings from textbook: Speech and Language Processing, Jurafsky and Martin: 3rd edition (online) and 2nd edition (paperback, International version, for chapters that aren't updated in 3rd ed).
- There will also be links to academic papers (recommended).
- Tutorials and quizzes.
- Lectures are being recorded. The audience is not in shot.
- Two assignments, worth 25%.
- Exam in April/May, worth 75% of final mark.

What is Natural Language Processing?



What is Natural Language Processing?

Applications

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

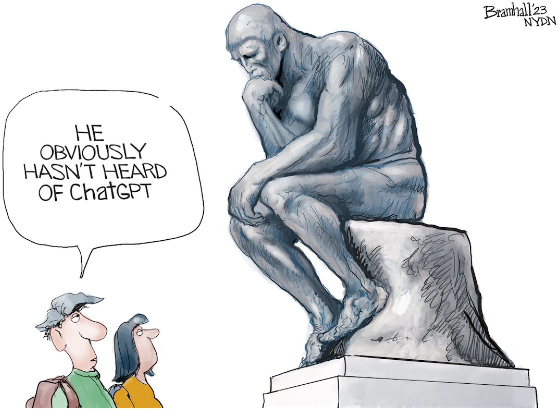
Core Technologies

- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Coreference resolution
- Named-entity recognition
- Word sense disambiguation
- Semantic Role Labeling
- ...

This course

NLP is a big field! We focus mainly on core ideas and methods needed for technologies in the second column (and eventually for applications).

- Core concepts in NLP
- Core linguistic problems and methodologies in NLP
- including machine learning, problem design, and evaluation methods



What does an NLP system need to “know”?

This is a simple sentence WORDS

- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

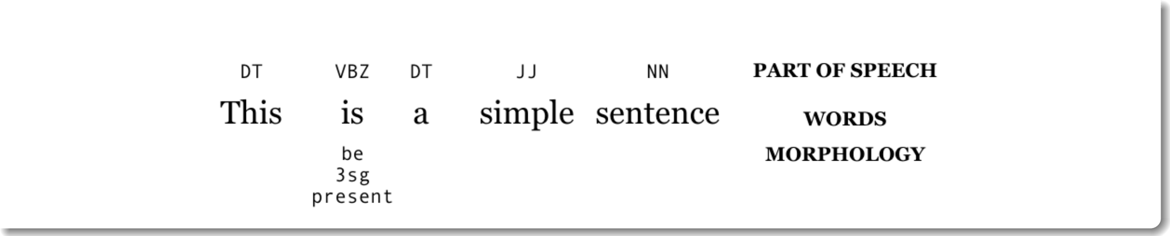
What does an NLP system need to “know”?

This is a simple sentence — WORDS

is: be, 3sg, present — MORPHOLOGY

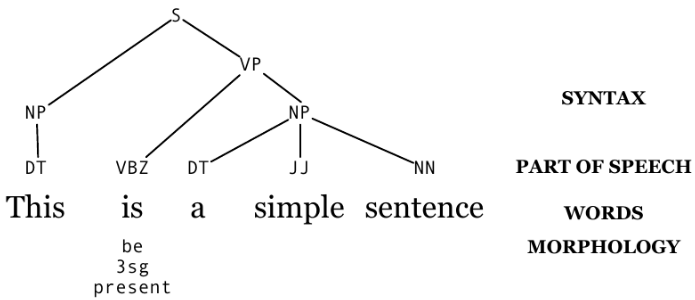
- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

What does an NLP system need to "know"?



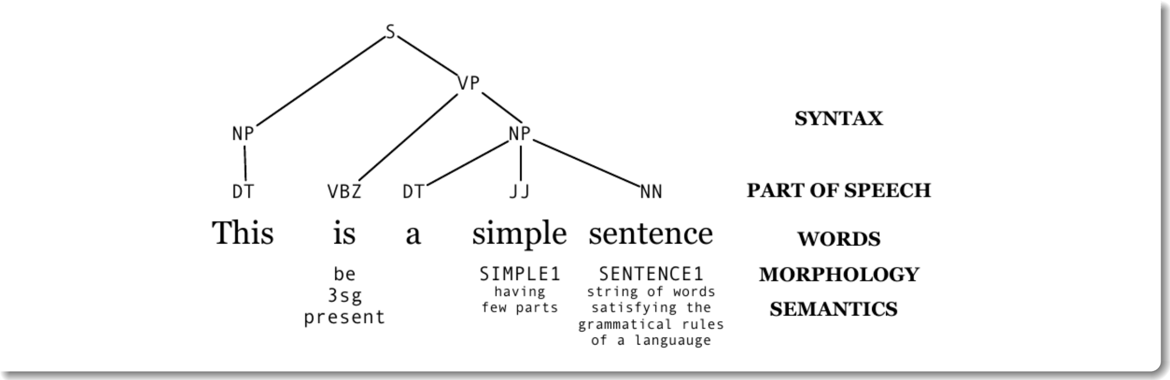
- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

What does an NLP system need to "know"?



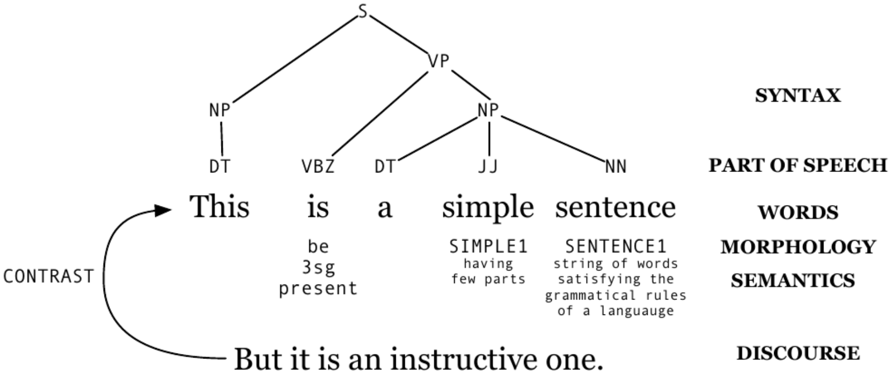
- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

What does an NLP system need to “know”?



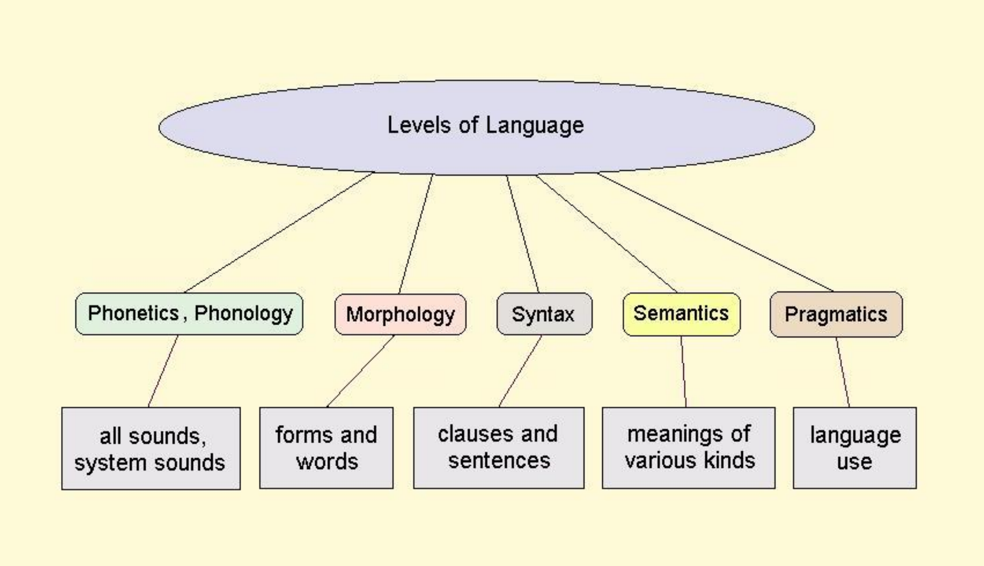
- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

What does an NLP system need to “know”?

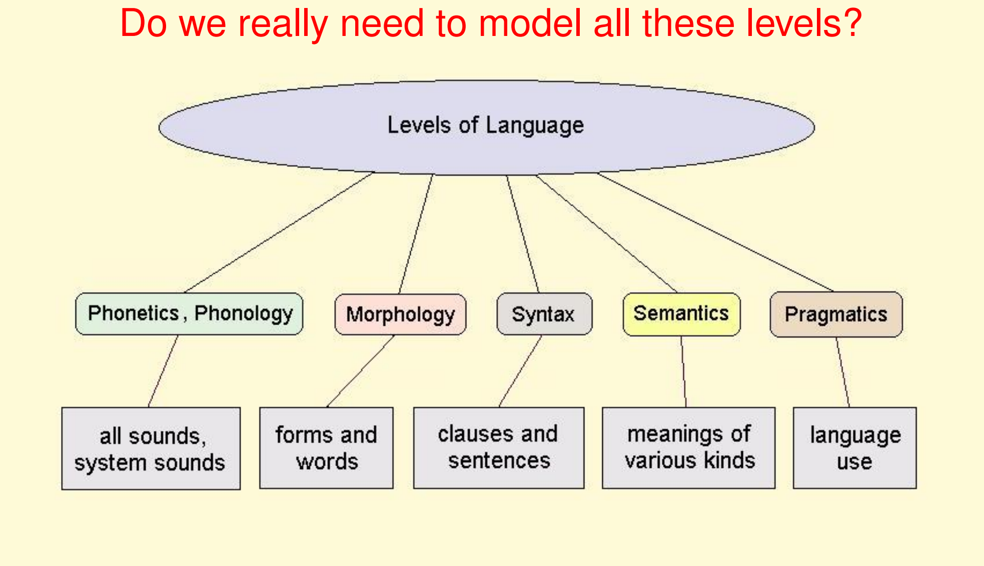


- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

Levels of Linguistics Analysis



Levels of Linguistics Analysis



Machine Reading Example

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCullock saw this bridge and decided to bring it to the USA. He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Machine Reading Example

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA. He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Machine Reading Example

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA. He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Machine Reading Example

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA. He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Machine Reading Example

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA. He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Machine Reading Example

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA. He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.
1. Who bought a bridge? 2. Where will the bridge be re-built? 3. How long will it take?

Machine Reading Example

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA. He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Machine Reading Example

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA. He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

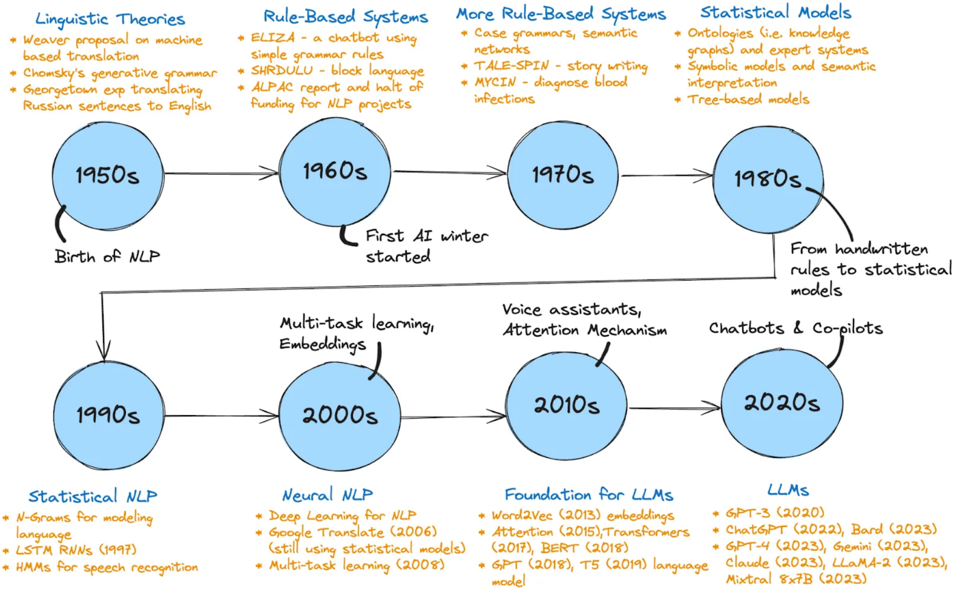
1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Machine Reading Example

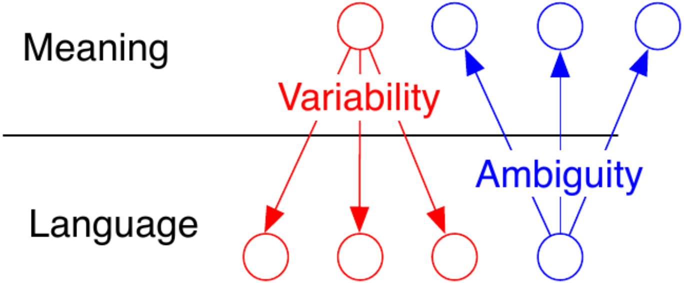
London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA. He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

History of NLP



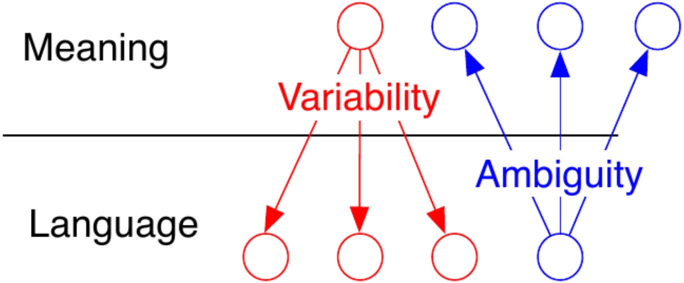
Why is NLP hard?



Variability:

He drew the house
He made a sketch of the house
He showed me his drawing of the house
He portrayed the house in his paintings
He drafted the house in his sketchbook

Why is NLP hard?



Ambiguity:

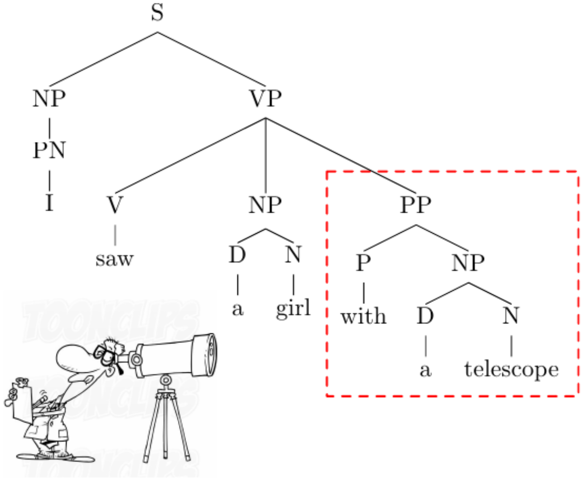
She drew a picture of herself	~ sketched, made a drawing of
cart drawn by two horses...	~ pulled
He drew crowds wherever he went ...	~ attracted
The driver slowed as he drew even with me	~ proceeded
The officer drew a gun and pointed it at ...	~ took out, produced

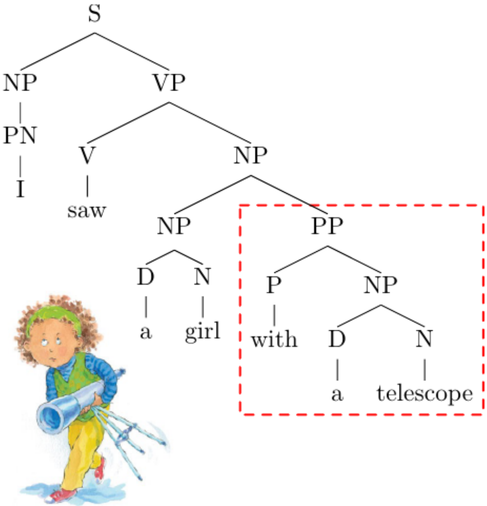
Why is NLP hard? Ambiguity at many levels

- Homophones: blew and blue
- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a girl with a telescope
  – We'll look into this in more detail!
- Quantifier scope: Every child loves some movie
- Multiple: I saw her duck
- Reference: John dropped the goblet onto the glass table and it broke.
- Discourse: The meeting is canceled. Emily isn't coming to the office today.

How can we model ambiguity, and choose the correct analysis in context?

Syntactic Ambiguity: Prepositional Phrase Attachment





Syntactic Ambiguity

Example with 3 preposition phrases, 5 interpretations:

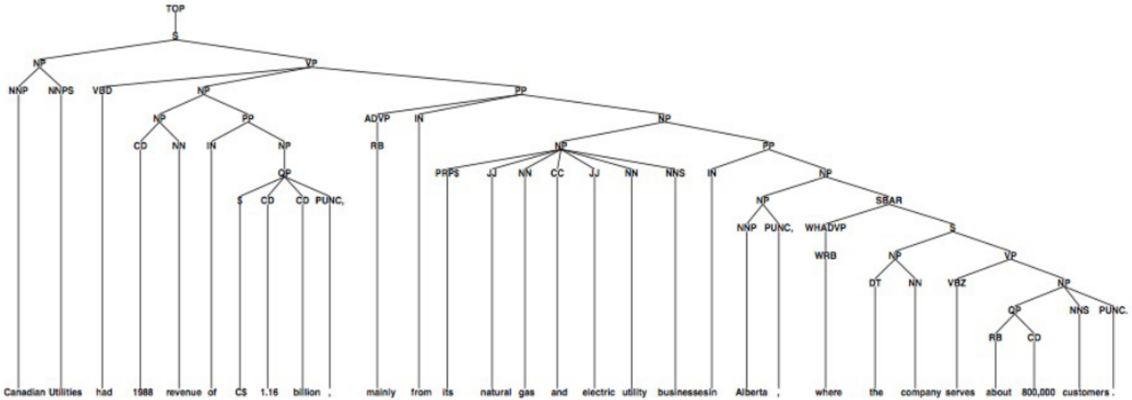
- Put the block ((in the box on the table) in the kitchen)
- Put the block (in the box (on the table in the kitchen))
- Put ((the block in the box) on the table) in the kitchen
- Put (the block (in the box on the table)) in the kitchen
- Put (the block in the box) (on the table in the kitchen)

The number of parses is an integer series, known as the Catalan numbers!

$$Cat_n = \binom{2n}{n} - \binom{2n}{n-1} \sim \frac{4^n}{n^{3/2}\sqrt{\pi}}$$

1, 2, 5, 24, 42, 132, 429, 1430, 4862, 16796, 58786, ...

A typical tree from a standard dataset (Penn treebank WSJ)



Canadian Utilities had 1988 revenue of $ 1.16 billion , mainly from its natural gas and electric utility businesses in Alberta , where the company serves about 800,000 customers .

Real Newspaper Headlines

Students Cook & Serve Grandparents

On Thursday, September 9, Gorman School hosted the first annual Grandparent's Day.

All Grandparents were invited to a school wide pancake breakfast. Upper grade students served as excellent chefs, as well as taking responsibility for serving the food and the clean up after-

- Ban on Nude Dancing on Governor's Desk
- Iraqi Head Seeks Arms
- Local High School Dropouts Cut in Half
- Red Tape Holds Up New Bridges
- Juvenile Court to Try Shooting Defendant
- Kids Make Nutritious Snacks

Collected by Chris Manning

Summary

- Course logistics
- What is Natural Language Processing (AI rock star!)
- Language consists of many levels of structure
- NLP is hard due to ambiguity at many levels

Next lecture: we discuss NLP challenges some more, and probabilistic modeling.