Natural Language Processing

Data Science, Spring 2021

Before we start...

Before we start...

1. Our mod of the day.

Before we start...

1. Our mod of the day.
2. Project 1

Before we start...

1. Our mod of the day.
2. Project 1
3. Project 2

Before we start...

1. Our mod of the day.
2. Project 1
3. Project 2
4. Midterm

Our moderator

Our moderator

1. Julian!

Project 1

Project 1

1. Grading is almost done, I will be posting official mid-semester grades.

Project 1

1. Grading is almost done, I will be posting official mid-semester grades.
2. Regrades are case-by-case. I mostly defer to the TAs so there must be a very compelling issue.

Project 2

[1] Notice the timezone

Project 2

1. You can use SQL *and* Pandas, it doesn't have to all be SQL (might be impossible?)

[1] Notice the timezone

Project 2

1. You can use SQL *and* Pandas, it doesn't have to all be SQL (might be impossible?)
2. Plots:

[1] Notice the timezone

Project 2

1. You can use SQL *and* Pandas, it doesn't have to all be SQL (might be impossible?)
2. Plots:
   2.1 You don't need N-plots for N distributions!

[1] Notice the timezone

Project 2

1. You can use SQL *and* Pandas, it doesn't have to all be SQL (might be impossible?)
2. Plots:
   2.1 You don't need N-plots for N distributions!
   2.2 Look up box-and-whisker plots.

[1] Notice the timezone

Project 2

1. You can use SQL *and* Pandas, it doesn't have to all be SQL (might be impossible?)
2. Plots:
   2.1 You don't need N-plots for N distributions!
   2.2 Look up box-and-whisker plots.
3. Nothing will be accepted after 11:59pm EDT[1]

[1] Notice the timezone

Midterm

Midterm

1. Like the quizzes, but significantly more difficult.

Midterm

1. Like the quizzes, but significantly more difficult.
2. Last semester's exam is up on the website.

Part I: What.

Natural Language Processing

What is it?

Natural Language Processing

What is it?

1. ‘Understanding’ text

Natural Language Processing

What is it?

1. ‘Understanding’ text
2. Analyzing text

Natural Language Processing

What is it?

1. ‘Understanding’ text
2. Analyzing text
3. Translating text

Natural Language Processing

What is it?

1. ‘Understanding’ text
2. Analyzing text
3. Translating text
4. and more!

In the olden days:

Computation was very expensive, so computationally cheap methods were preferred

In the olden days:

Computation was very expensive, so computationally cheap methods were preferred

1. Dictionary lookup for translation: replace a word in one language with it's 'equivalent' in another language.

In the olden days:

Computation was very expensive, so computationally cheap methods were preferred

1. Dictionary lookup for translation: replace a word in one language with it's 'equivalent' in another language.
2. This didn't get very far.

In the olden days:

Computation was very expensive, so computationally cheap methods were preferred

1. Dictionary lookup for translation: replace a word in one language with it's 'equivalent' in another language.
2. This didn't get very far.
3. Between the 1930's and the 1980's, not much work was done on mechanical translation

NLP more generally

Two main schools of thought:

NLP more generally

Two main schools of thought:

1. Rule + Grammar-based methods

NLP more generally

Two main schools of thought:

1. Rule + Grammar-based methods
2. Statistical methods

NLP more generally

Two main schools of thought:

1. Rule + Grammar-based methods
2. Statistical methods

These two schools of thought go up and down in popularity.

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

1. Phonetics

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

1. Phonetics
2. Phonology

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

1. Phonetics
2. Phonology
3. Morphology

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

1. Phonetics
2. Phonology
3. Morphology
4. Syntax

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

1. Phonetics
2. Phonology
3. Morphology
4. Syntax
5. Semantics

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

1. Phonetics
2. Phonology
3. Morphology
4. Syntax
5. Semantics
6. Pragmatics

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

1. Phonetics
2. Phonology
3. Morphology
4. Syntax
5. Semantics
6. Pragmatics
7. Discourse analysis

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

1. Phonetics
2. Phonology
3. Morphology
4. Syntax
5. Semantics
6. Pragmatics
7. Discourse analysis
8. Stylistics

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

1. Phonetics
2. Phonology
3. Morphology
4. Syntax
5. Semantics
6. Pragmatics
7. Discourse analysis
8. Stylistics
9. Semiotics

Rule and Grammar-based methods

Let's use linguistics to formally reason about language.

Sounds complicated... what if we just didn't do that?

Statistical Methods

Cynically: Through it at some ML techniques and see what happens.

Statistical Methods

Cynically: Through it at some ML techniques and see what happens.

1. Decisions tress to automatically learn rules automatically

Statistical Methods

Cynically: Through it at some ML techniques and see what happens.

1. Decisions tress to automatically learn rules automatically
2. Hidden Markov Models (HMM) for parts-of-speech tagging

Statistical Methods

Cynically: Through it at some ML techniques and see what happens.

1. Decisions tress to automatically learn rules automatically
2. Hidden Markov Models (HMM) for parts-of-speech tagging
3. (Un)supervised learning of language models (this is the hotness)

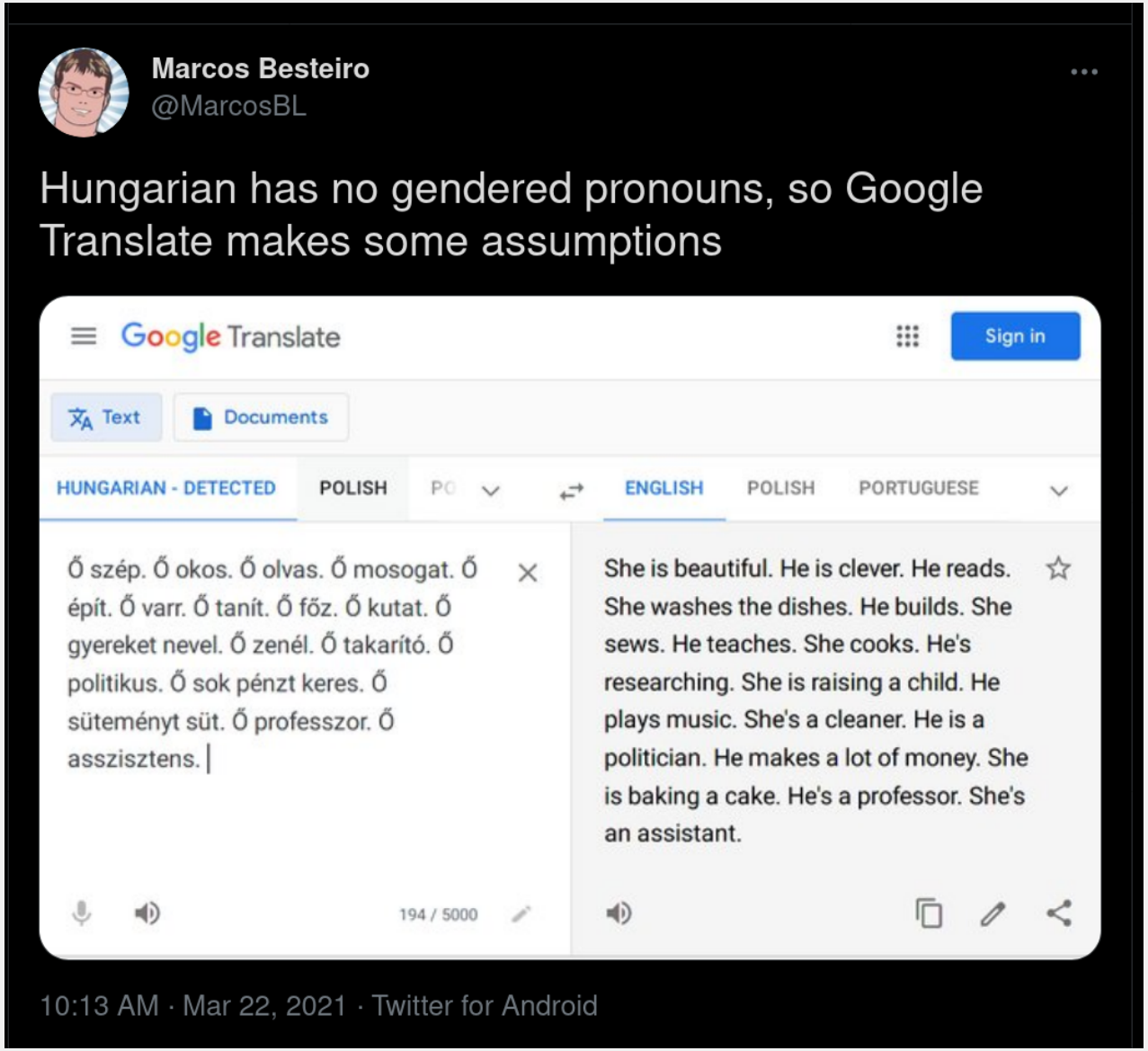
These two schools of thought go up and down in popularity.

Statistical Methods

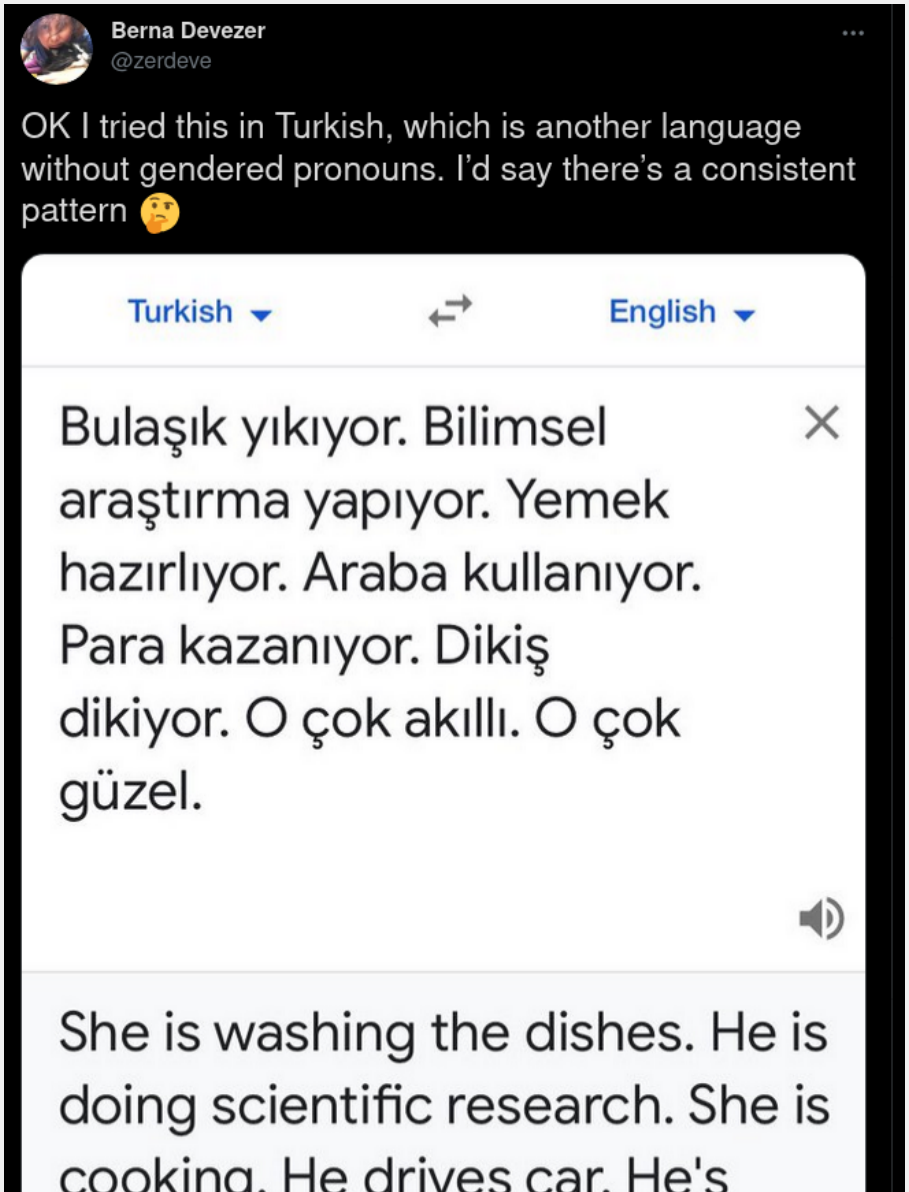
Cynically: Through it at some ML techniques and see what happens.

1. Decisions tress to automatically learn rules automatically
2. Hidden Markov Models (HMM) for parts-of-speech tagging
3. (Un)supervised learning of language models (this is the hotness)
4. We will discuss some of these techniques in the Machine Learning part of the semester.

These methods reflect our biases:



These methods reflect our biases:



Part II: Why.

Must be nice...

Some data scientists have it easy: their data is already in a machine friendly format (HTML/SQL Database/etc). But what about...

Must be nice...

Some data scientists have it easy: their data is already in a machine friendly format (HTML/SQL Database/etc). But what about...

1. Facebook posts

Must be nice...

Some data scientists have it easy: their data is already in a machine friendly format (HTML/SQL Database/etc). But what about...

1. Facebook posts
2. Product reviews

Must be nice...

Some data scientists have it easy: their data is already in a machine friendly format (HTML/SQL Database/etc). But what about...

1. Facebook posts
2. Product reviews
3. data dumps (Panama Papers, wikileaks, etc.)

Must be nice...

Some data scientists have it easy: their data is already in a machine friendly format (HTML/SQL Database/etc). But what about...

1. Facebook posts
2. Product reviews
3. data dumps (Panama Papers, wikileaks, etc.)
4. Can you think of other examples?

Must be nice...

Some data scientists have it easy: their data is already in a machine friendly format (HTML/SQL Database/etc). But what about...

1. Facebook posts
2. Product reviews
3. data dumps (Panama Papers, wikileaks, etc.)
4. Can you think of other examples?
5. We want this data to be useful too!

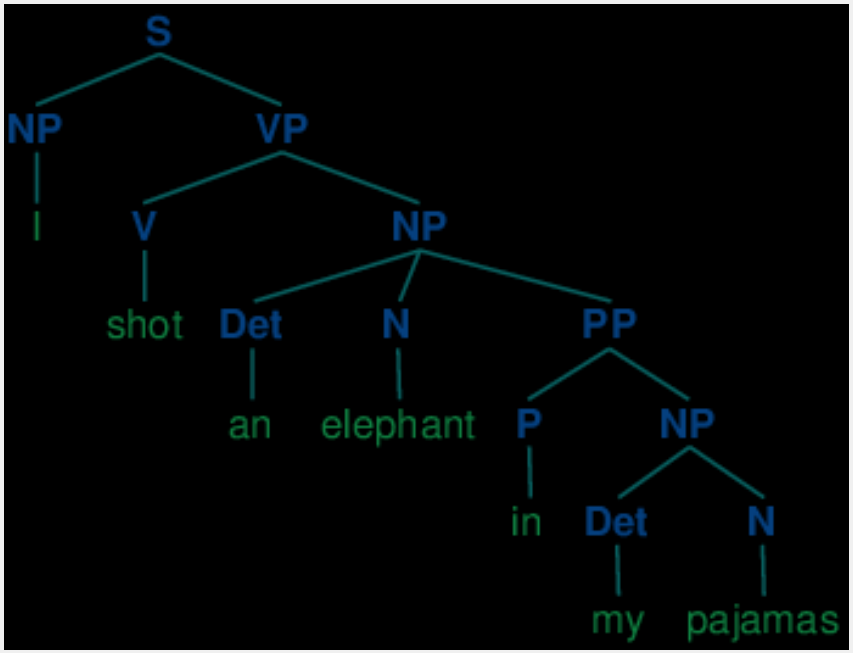
Part III: How.

‘Understanding’ is hard

“One morning I shot an elephant in my pajamas.”

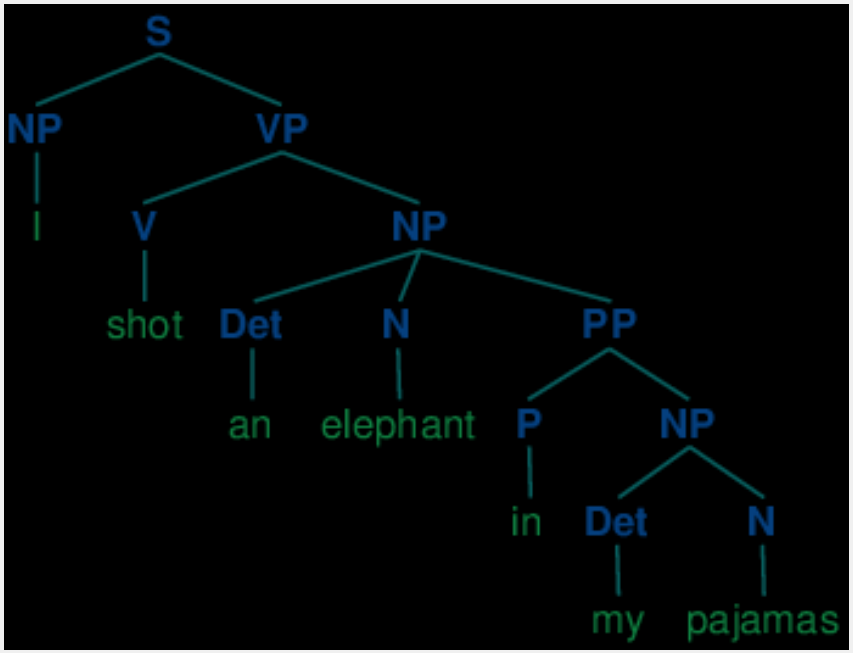
‘Understanding’ is hard

“One morning I shot an elephant in my pajamas.”



‘Understanding’ is hard

“One morning I shot an elephant in my pajamas.”



“How he got in my pajamas, I’ll never know”

‘Understanding’ is really hard

Winograd Schema Challenge:
“The city councilmen refused the demonstrators a permit because they [*feared*/*advocated*] violence.”

Sentiment

Understanding language directly is very hard. Perhaps understanding sentiment is less hard?

Sentiment

Understanding language directly is very hard. Perhaps understanding sentiment is less hard?
"I bought this product and I use it, but I wouldn't recommend it to my worst enemy"

Sentiment

Understanding language directly is very hard. Perhaps understanding sentiment is less hard?

"It might seem like this movie is bad, but once you get past the cheesy acting I rather enjoyed it"

Summary

Summaries are also possible: SMMRY an algorithm for mechanical summaries

Thanks for your time!

:)