Natural Language Processing

Content: dialogsystems.ai

IEOR 135/290
UC Berkeley
Fall 2018

30 Minutes
Theory

1 Hour
Practice

Raise of hands:

1) Who has used NLP?
2) What examples come to mind?

Me from an NLP project perspective:

1) Radiology Hedging
2) Medical Records for Foresight
3) Suicide Hotlines
4) IT Service Center tools

NLP @ Berkeley:

SCET Alexa Fellowship

Dan Klein: Semantic Machines

https://blogs.microsoft.com/blog/2018/05/20/microsoft-acquires-semantic-machines-advancing-the-state-of-conversational-ai/

John Denero: Google Translate

The iPhone 1 keyboard is a fascinating study into NLP.

Natural Language Processing (NLP)

Humans speak/write, and listen/read. NLP tries to emulate this.

Natural Language Processing (NLP)

The field of CS concerned with understanding and generating words.

This isn’t easy:

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

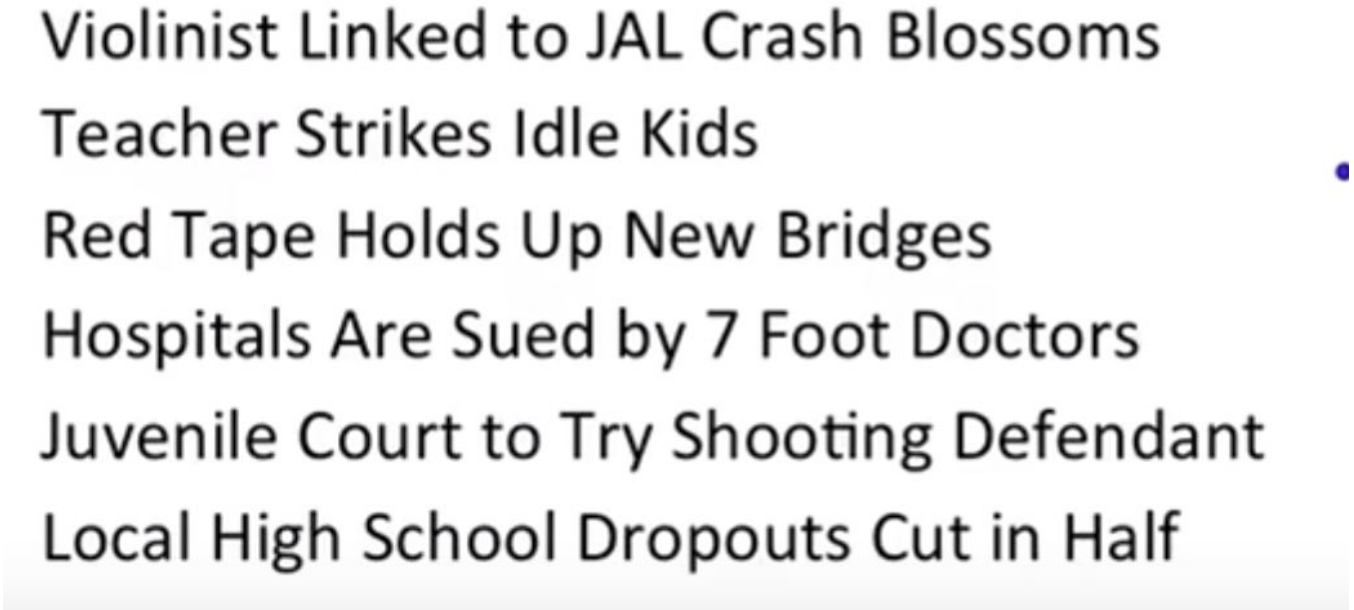
Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

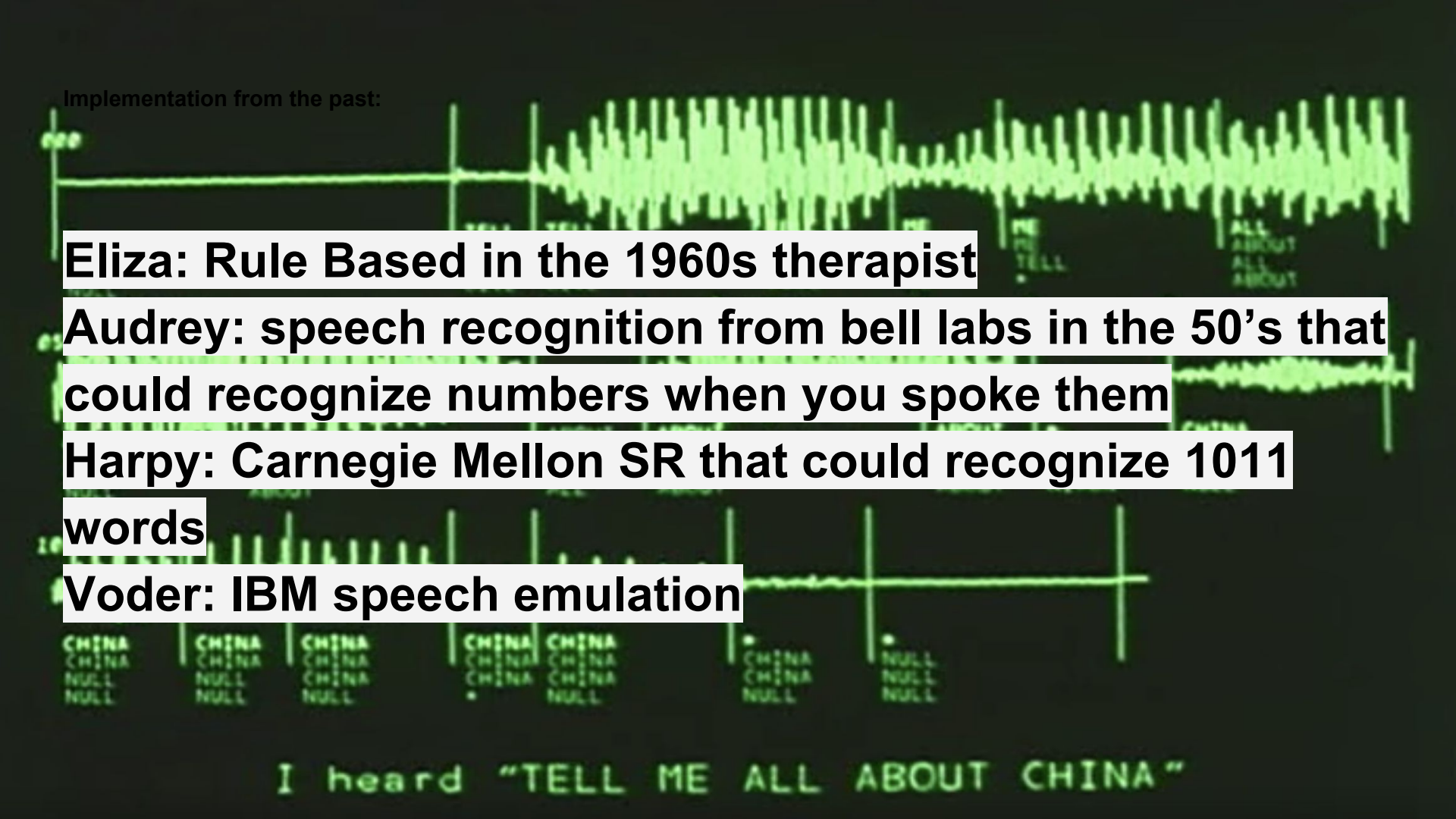
Where is *A Bug’s Life* playing …
Let It Be was recorded …
… a mutation on the *for* gene …

Natural Language Processing (NLP)

A common theme is dealing with ambiguity in word use.



Implementation from the past:



Eliza: Rule Based in the 1960s therapist

Audrey: speech recognition from bell labs in the 50's that could recognize numbers when you spoke them

Harpy: Carnegie Mellon SR that could recognize 1011 words

Voder: IBM speech emulation

Natural Language Processing (NLP)

Becoming increasingly relevant as textual data becomes more accessible and unstructured data becomes more useable.

https://github.com/niderhoff/nlp-datasets

Theme: we read and write

Reading

Implementations of Reading:

Analyzing unstructured data such as written medical records, finding data science insights
Using OCR with NLP to read medical notes or notes from the field

Implementations of Reading:

Speech recognition
Sentiment Analysis
Topic Modeling

Theme: we read and write

Implementations of Writing:

Beyond Chatbots:
Text search (not just on Google)
    Keyword search
Word Suggestions (while typing)
Spelling Suggestions

Implementations of Writing:

Text to Speech
Speech to Text
Machine Translation of languages
Captioning (YouTube)

Methods used to accomplish reading and writing:

Tokenization

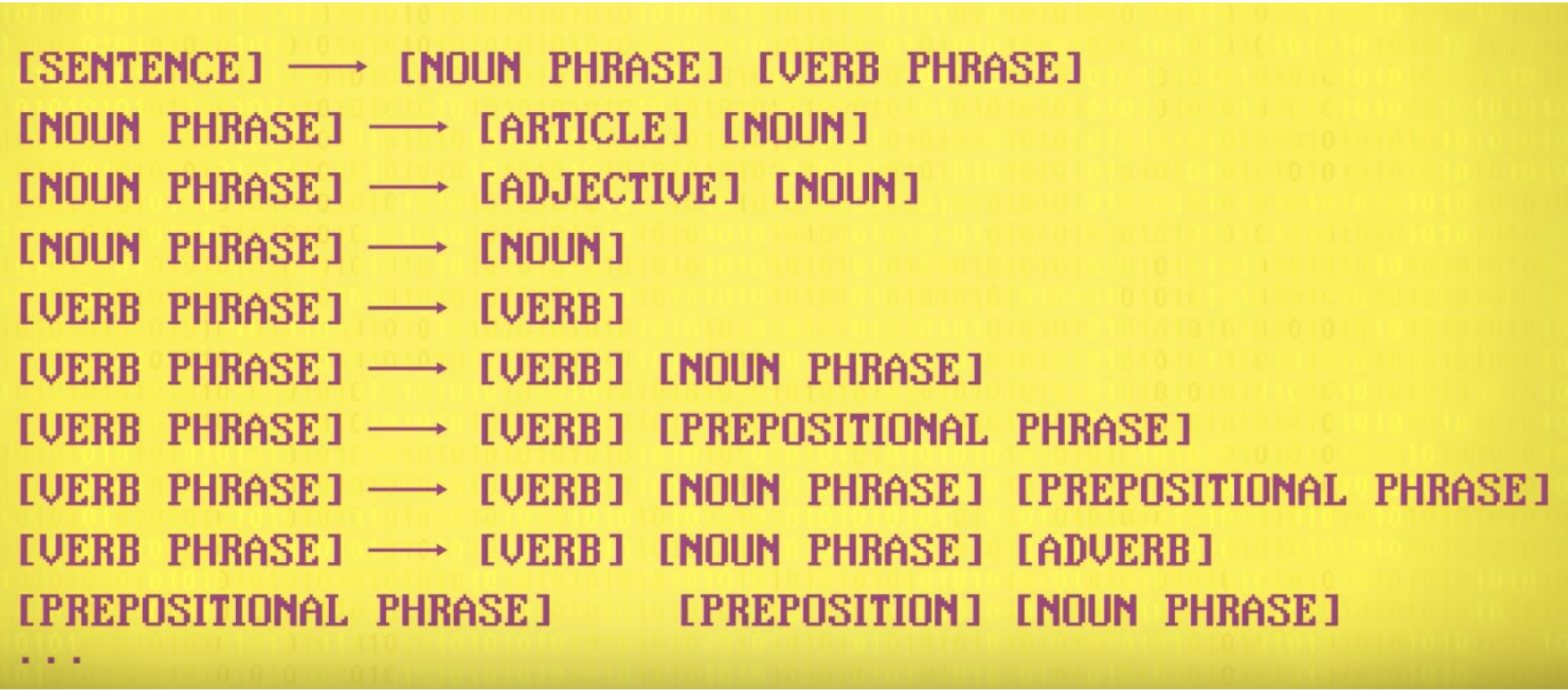
Stemming

Lemmatization

Stemming


Affectation
Affects
Affections
Affected
Affection
Affecting

Phrase Structuring



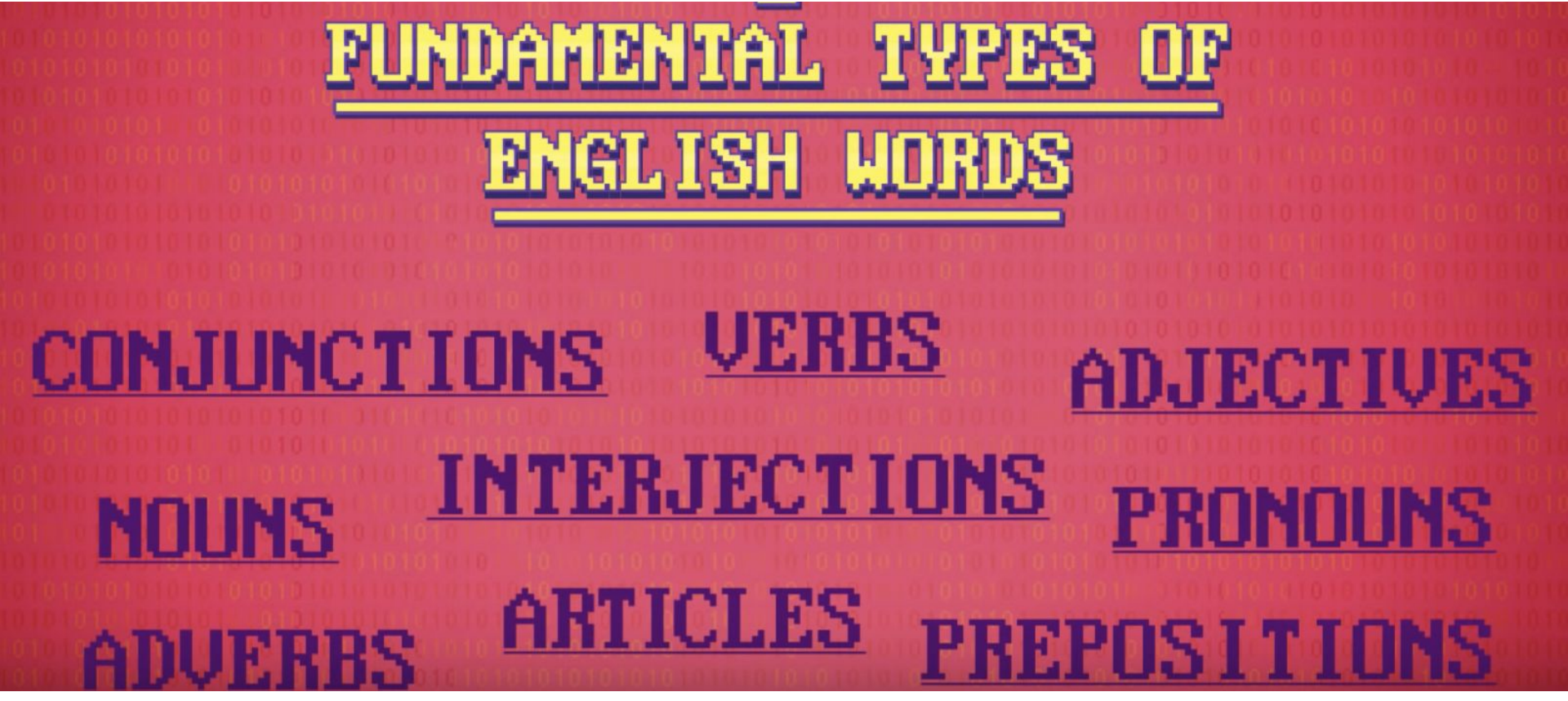
Methods used to accomplish reading and writing:

Lemmatization

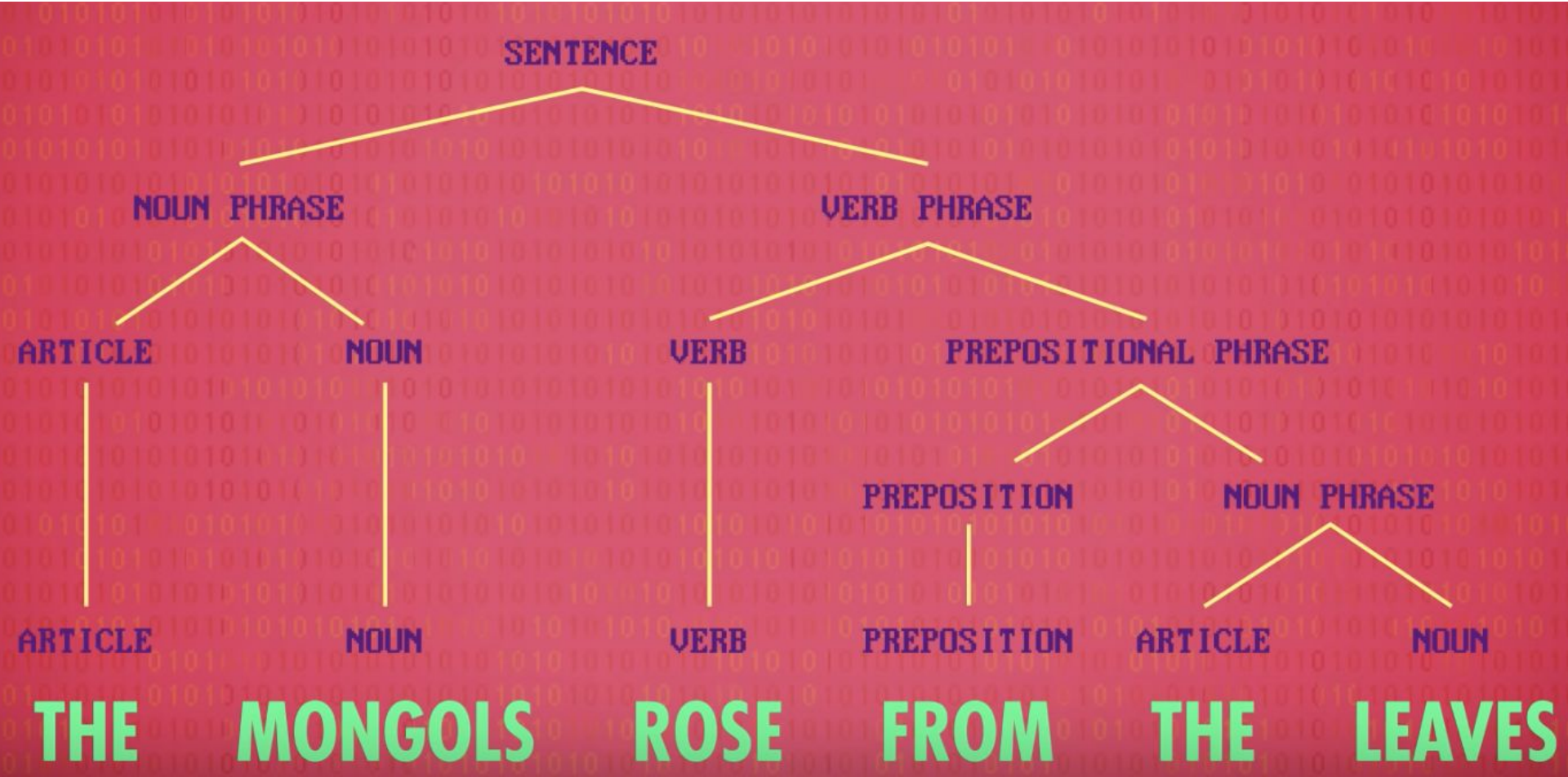
Part of Speech Tagging

Named Entity Recognition

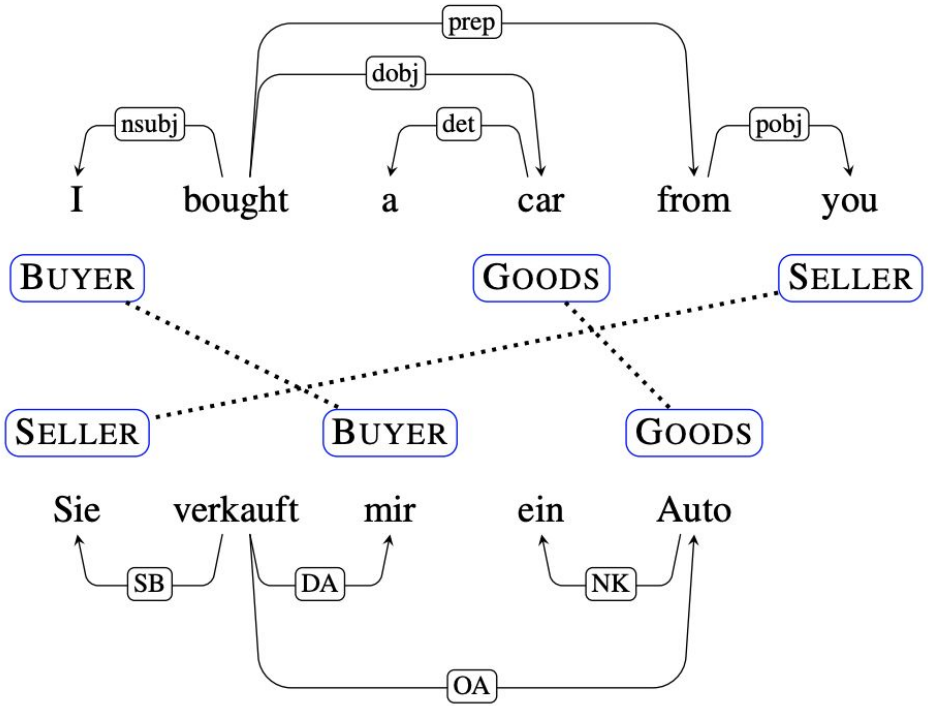
Part of Speech Tagging



Lemmatization:



Machine Translation



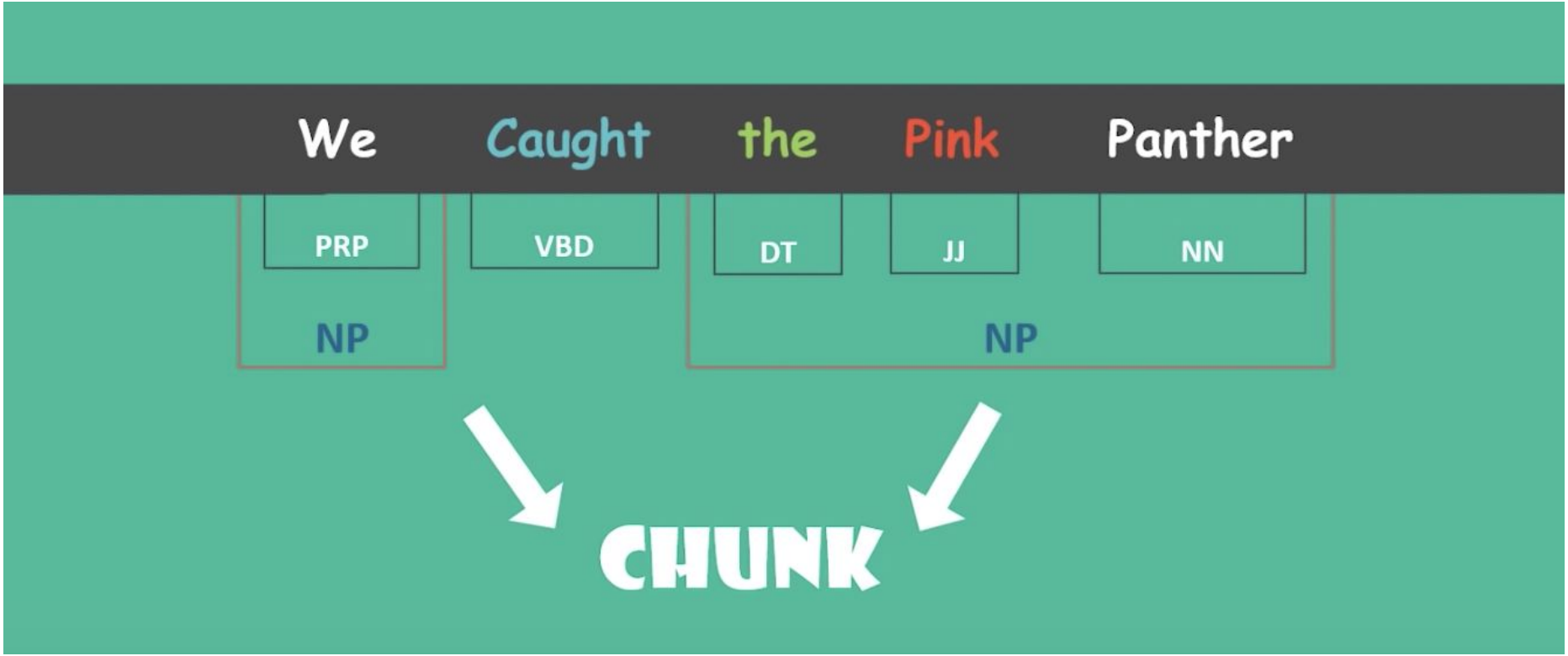
Methods used to accomplish reading and writing:

Chunking

Defining a document and a corpus

Identifying and removing ‘stop words’

Chunking



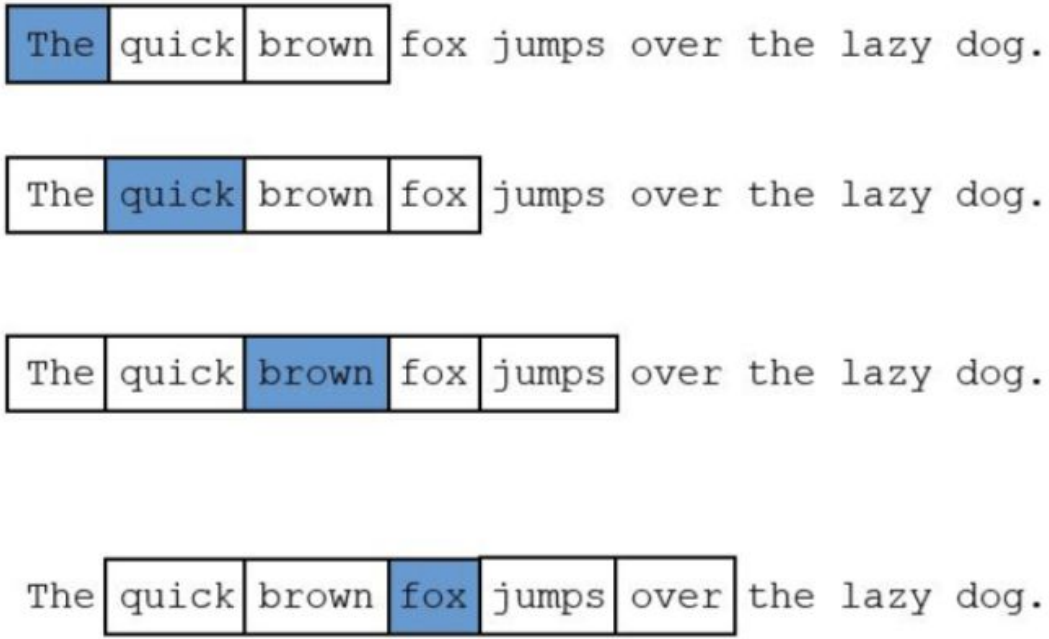
Methods used to accomplish reading and writing:

N-grams

Document Matrices

N-Grams

One assumption: Context ~ Meaning:



Document Matrices

	Hamlet	Macbeth	Romeo & Juliet	Richard III	Julius Caesar	Tempest	Othello	King Lear
knife	1	1	4	2		2		2
dog	2		6	6		2		12
sword	17	2	7	12		2		17
love	64		135	63		12		48
like	75	38	34	36	34	41	27	44

Term Frequency (TF)
X
Inverse Document Frequency (IDF)

Representing a word by how frequently it occurs, accounting for frequency overall.

TF: # times word occurs: $(tf_{t,d})$

IDF: count of documents the word is in (D_t), among total documents (N): $\log N/D_t$

$$tfidf(t,d) = tf_{t,d} \times \log \frac{N}{D_t}$$

You don’t have to start totally from scratch:

Pandas
Scikit Learn
Natural Language Toolkit
Google Knowledge Graph > Named Entity Recognition
Word2Vec : wikipedia


jupyter

Current Status of what's

80%-100% solved

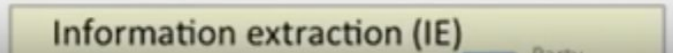
Spam detection, part of speech tagging, named entity recognition

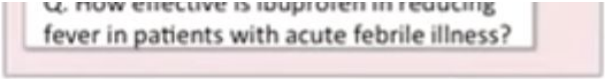
10%-50% solved

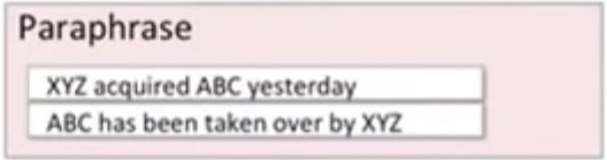
Sentiment analysis, Coreference resolution, word sense disambiguation, parsing, machine translation, information extraction

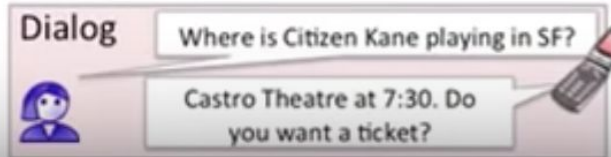
<.01%-10% solved:

Question answering, paraphrasing (the way it's 50% solved in CV), summarization, dialog









Theme: we read and write

Reading and Writing: Dialog Systems

An environment that facilitates the exchange of conversation.

Dialog Systems

When a phrase such as X occurs, what is likely to occur following that?
How can you know?

1) Dialog State Classification
2) Embeddings conversational concepts such estimated response time

Dialog Act Classification

Tag	Example	%
STATEMENT	*Me, I'm in the legal department.*	36%
BACKCHANNEL/ACKNOWLEDGE	*Uh-huh.*	19%
OPINION	*I think it's great*	13%
ABANDONED/UNINTERPRETABLE	*So, -/*	6%
AGREEMENT/ACCEPT	*That's exactly it.*	5%
APPRECIATION	*I can imagine.*	2%
YES-NO-QUESTION	*Do you have to have any special training?*	2%
NON-VERBAL	*<Laughter>,<Throat_clearing>*	2%
YES ANSWERS	*Yes.*	1%
CONVENTIONAL-CLOSING	*Well, it's been nice talking to you.*	1%
WH-QUESTION	*What did you wear to work today?*	1%

Dialogic Features	**Relative Rank of importance**
total interruptions so far	1
interruptions	2
total button usages so far	3
total repetitions so far	4
repetition	5
button usage	6
total start over so far	7
start over	8

Natural Language Processing

content: dialogsystems.ai

IEOR 135/290
UC Berkeley
Spring 2019
vince@bartle.io

Credits to:

Intro to Text Analysis from: adashofdata.com

Introduction to NLP @ iSchool: https://www.ischool.berkeley.edu/courses/info/159

PBS: Linguistics Slides

Dan Jurafsky: Solved, Slightly Solved, Not Solved Slide