CS 463
NATURAL LANGUAGE PROCESSING

Dr. Saleh Haridy

2023-2024

CS 463: NLP

Course Code: CS 463 **Credits:** 3 (3 Lec., 1 Lab, 0 Tutorial)

Instructure: Dr. Saleh Haridy

Textbook:
- Jurafsky and Martin, “SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”, Third Edition, McGraw Hill, 2023.

References:
- Jacob Eisenstein, “Natural Language Processing”, November 13, 2018
- Manning and Schutze, "Statistical Natural Language Processing", MIT Press; 1st edition (June 18, 1999), ISBN: 0262133601

Online Resources:
- https://docs.python.org/3.10/tutorial/introduction.html
- https://www.nltk.org/
- https://www.nltk.org/book/

Lectures:

As per schedule

Evaluation and Grading Policy

Midterm Exam:	**20**
Two Quizzes	**20**
Mini Projects	**10**
Exercises	**10**
Final Exam:	**40**
Total	**100**

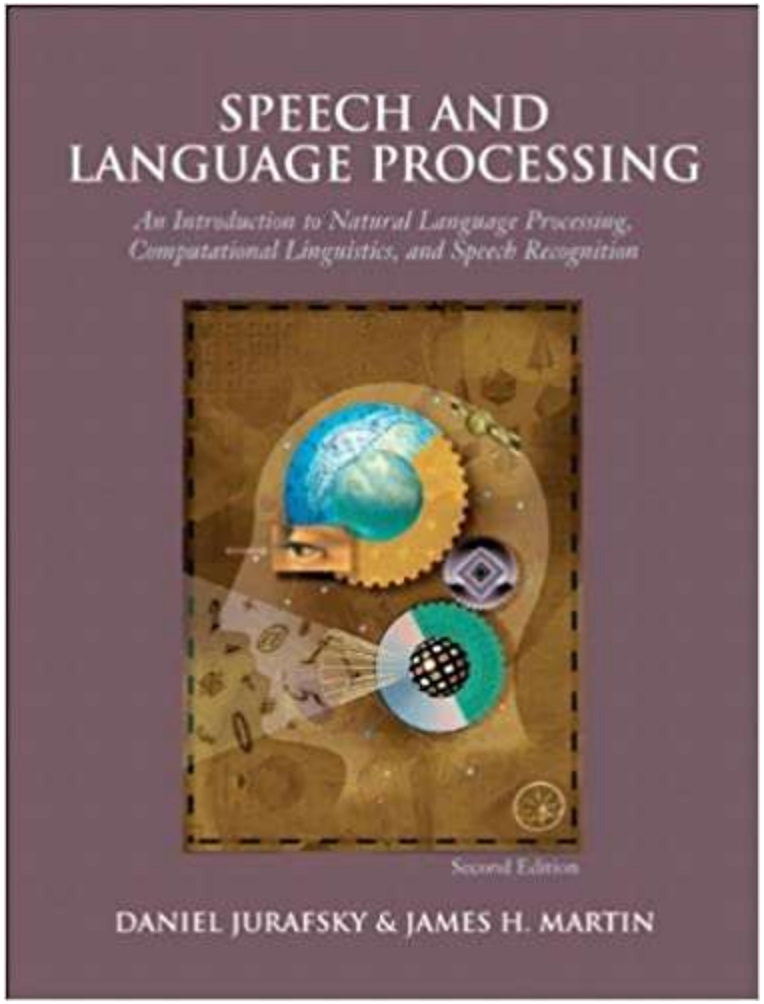
Course objectives

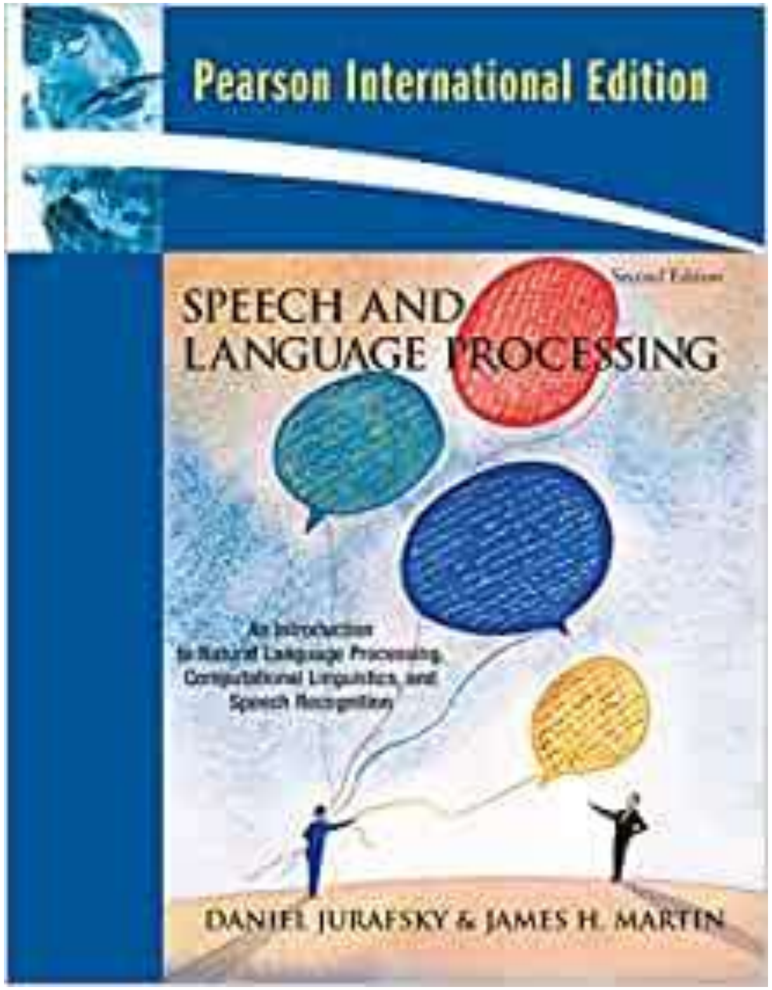
- To learn about Regular Expressions and Text Normalization and Edit Distance.
- To understand Language modeling using N-Gram and neural network models
- To design Text classification using Naïve Bayes, logistic regression, neural networks
- To understand vector semantics and embedding
- To understand Deep Learning language models and Chatbot

Course Contents

Week	Course Topics	Book's Chapter	Event Name
1	Introduction and Overview	1	
2	Regular Expressions, Text Normalization	2	
3	Minimum Edit Distance and Alignment	2	
4	N-gram Language Models	3	TEST 1
5	Naive Bayes and Sentiment classification	4	
6	Text classification using logistic regression	5	
7	Midterm Exam		
8	Vector Semantics and Embedding	6	
9	Neural Language Model I	7	
10	Neural Language Model II	7	
11	Part of Speech Tagging	8	TEST 2
12	Deep Learning Architectures for Sequence Processing	9	
13	Chatbots & Dialogue Systems I	24	
14	Chatbots & Dialogue Systems II	24	
15	FINAL EXAM		

Textbook


SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN


Pearson International Edition
Second Edition
SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
DANIEL JURAFSKY & JAMES H. MARTIN

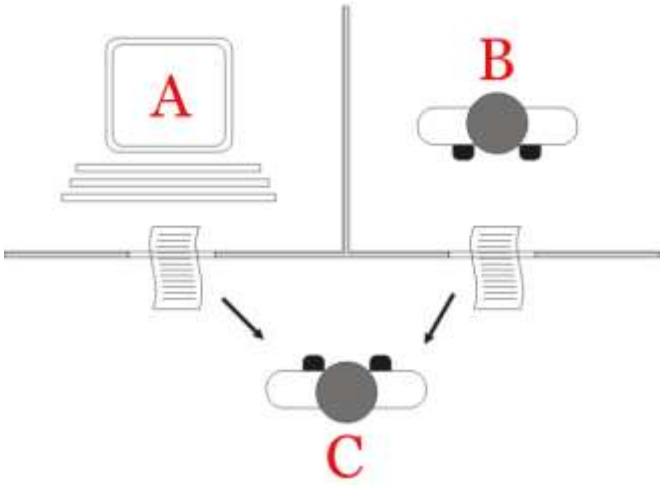
Week 1

Introduction to Natural Language Processing

What is Natural language processing?

- Natural language processing (NLP) refers to the branch of artificial intelligence or AI—concerned with giving computers the ability to **understand text** and **spoken words** in much the same way human beings can.
- Natural language processing has its roots in the 1950s when Alan Turing proposed what is now called the Turing test as a **criterion** of intelligence. The proposed test includes a task that involves the automated **interpretation** and **generation** of natural language.

Turing Test

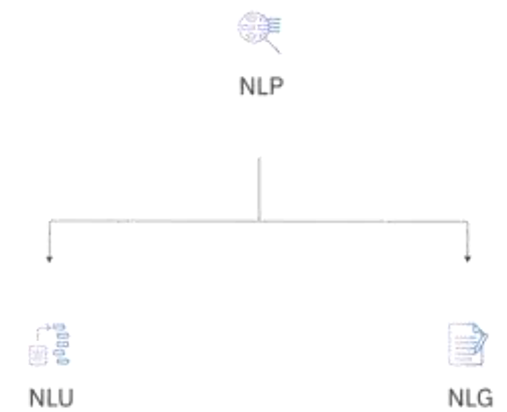


Human evaluator (C) judge natural language conversations between a human (A) and a machine (B) programmed to generate human-like responses

Natural language Processing/Understanding/Generation (NLP/NLU/NLG)?

- NL ∈{Arabic, Hindi, Spanish, Urdo, English, … Turkish}

- **Natural language understanding** (NLU) is a subset of natural language processing, which uses syntactic and semantic analysis of text and speech to determine the meaning of a sentence.
- **Natural language generation** NLG is the process of producing a human language text response based on some data input. This text can also be converted into a speech format through text-to-speech services.



Applications

- **Machine Translation** (Google translate,..)
- **Question Answering** (Information retrieval + NLP): IBM Watson
- **Dialogue Systems** (digital assistant)/Chatbots (casual conversation): Siri, Cortana, Alexa, Google Assistant,chatGPT,...
- **Text Summarization** (QuillBot , **SummarizeBot**, **Resoomer**,
- **Sentiment Analysis** (MonkeyLearn, Lexalytics, Brandwatch,…)
- ...

Machine translation

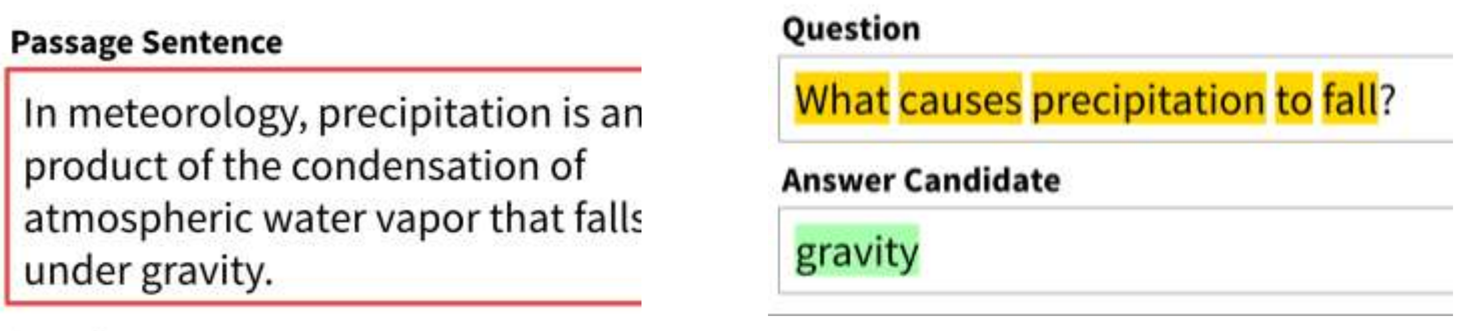
- Google Translate is an example of widely available NLP technology at work.
- Machine translation involves more than replacing words in one language with words of another.
- Effective translation has to capture accurately the meaning and tone of the input language and translate it to text with the same meaning and desired impact in the output language.
- A great way to test any machine translation tool is to translate text to one language and then back to the original.

Virtual agents and chatbots

- Virtual agents such as Apple's Siri, Google Assistant, Samsong Bixby, and Amazon's Alexa use speech recognition to recognize patterns in voice commands and natural language generation to respond with appropriate action or helpful comments.
- Chatbots can answer various questions asked during an interactive conversation.
- Interactive conversation means the system keeps a track of questions asked earlier and can engage in longer conversations.
- They have a sought of *memory* which helps answer in a more friendlier manner. Also, they retrieve information such as weather, stock prices from various sources. Hence, their ability is far beyond Q&A systems in this sense.

Question Answering (QA)

- The purpose of QA is to locate the text for any new question that has been addressed.
- It is programmed to answer questions only from a particular source of information of sometimes questions belonging to a common topic.
- They could be thought of a search engine which only works for a specific topic.

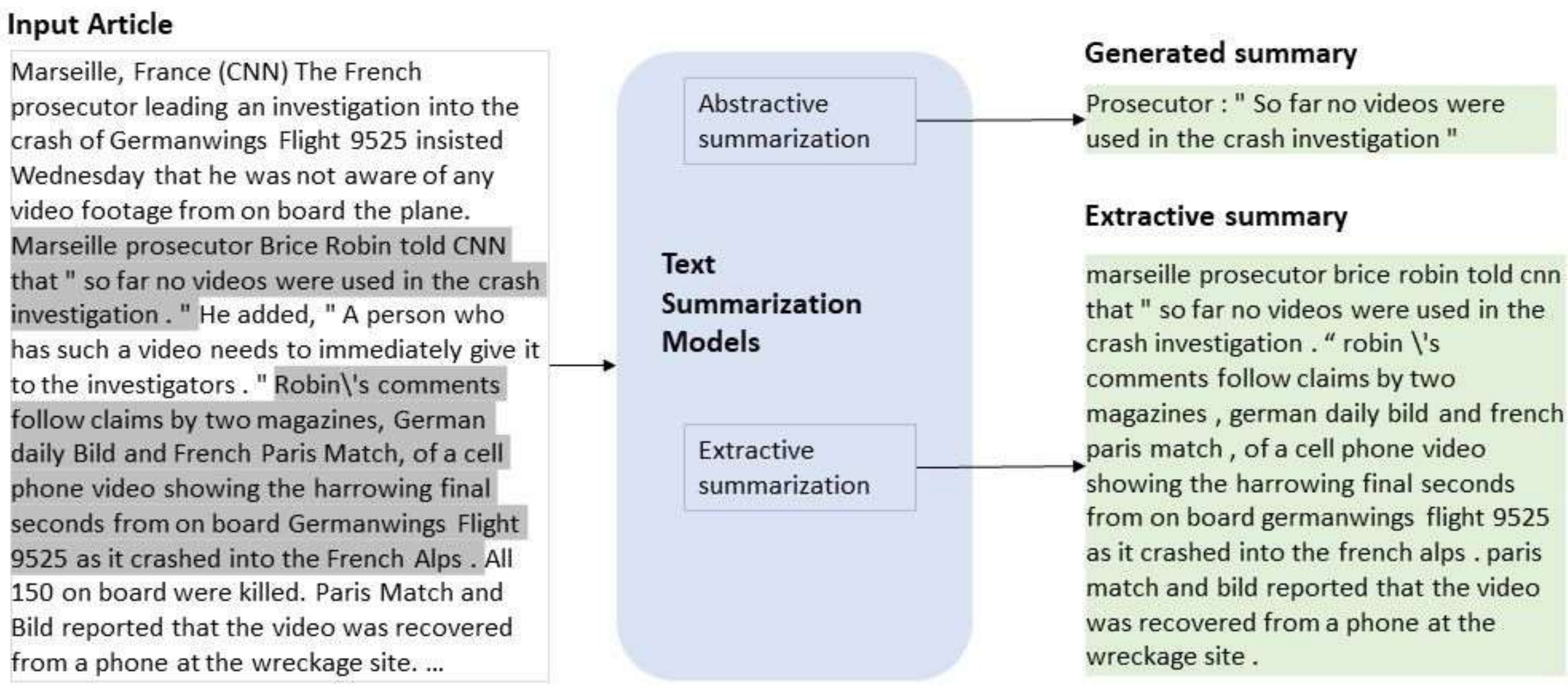


Sentiment analysis

- NLP is an essential business tool for uncovering hidden data insights from social media channels.
- Sentiment analysis can analyze language used in social media posts, responses, reviews, and more to extract attitudes and emotions in response to products, promotions, and events–information companies can use in product designs, advertising campaigns, and more.

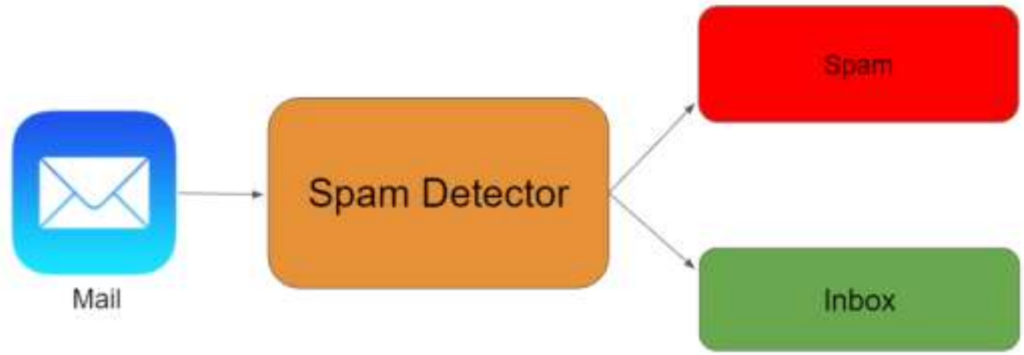
Text summarization

- Text summarization uses NLP techniques to digest huge volumes of digital text and create summaries and synopses for indexes, research databases, or busy readers who don't have time to read full text.



Spam detection

- Spam detection use NLP's text classification capabilities to scan emails for language that often indicates spam or phishing.
- These indicators can include overuse of financial terms, characteristic bad grammar, threatening language, inappropriate urgency, misspelled company names, and more.



Current situation of NLP technologies

mostly solved

Spam detection
- Let's go to Agra! ✓
- Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON			ORG		LOC
Einstein	met	with	UN	officials in	Princeton

making good progress

Sentiment analysis
- Best roast chicken in San Francisco!
- The waiter ignored us for 20 minutes.

Coreference resolution
Carter told Mubarak he shouldn't run again.

Word sense disambiguation
I need new batteries for my ***mouse***.

Parsing
I can see Alcatraz from the window!

Machine translation (MT)
第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)
You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)
Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase
- XYZ acquired ABC yesterday
- ABC has been taken over by XYZ

Summarization
- The Dow Jones is up
- The S&P500 jumped
- Housing prices rose

Economy is good

Dialog
Where is Citizen Kane playing in SF?

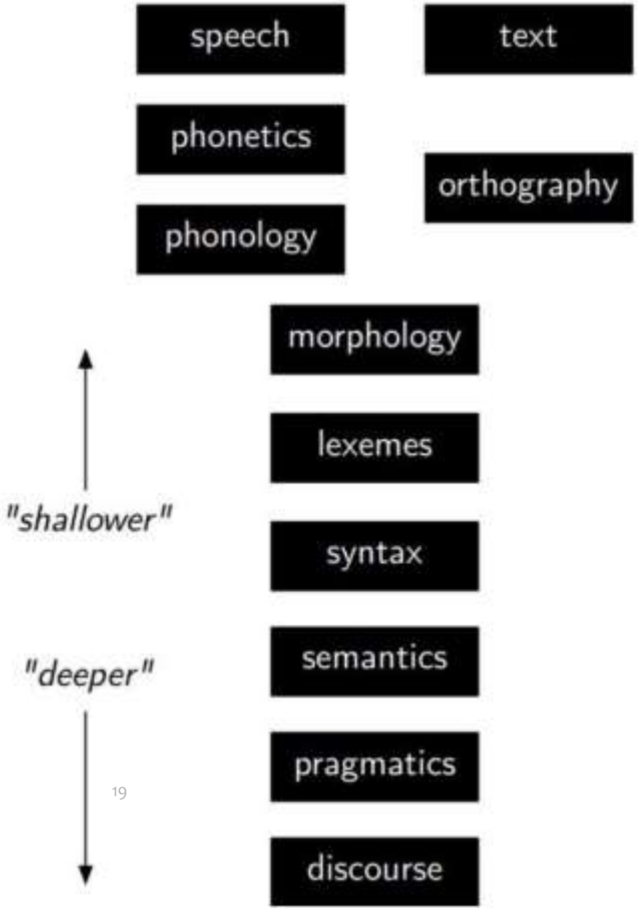
Castro Theatre at 7:30. Do you want a ticket?

How to Acquire Knowledge from Language?

- What distinguishes language processing applications from other data processing systems is their use of *knowledge of language*.
- Assume you write a program to chat with a human,
  - It should recognize words from audio signal, so it requires knowledge about phonetics and phonology.
  - Should know that “doors” is plural, it require morphological analysis of word.
  - It must be able to concatenate words properly to create a sentence. This require knowledge of a syntax analysis.
  - To answer a question, it should know the meaning of each words (lexical semantics), as well as compositional semantics
  - It should know the tone of speaker, so decide the action by using pragmatic or dialogue knowledge.
  - It makes use of knowledge about how words like that or pronouns like it or she refer to previous parts of the discourse (coreference resolution).

NLP tasks

- **Phonetics and Phonology**: knowledge about linguistic sounds
- **Morphology/lexemes**: Knowledge of meaningful component of word, it concerns the way words are built up from smaller meaning bearing units. .
- **Syntax**: Knowledge of the structural relationships between words, it concerns how words are put together to form correct sentences.



NLP tasks

- **Semantics**: knowledge of meaning, it concerns what words mean and how these meanings combine in sentences to form sentence meanings
- **Pragmatics**: knowledge of the relationship of meaning to the goals and intentions of the speaker. It concerns how sentences are used in different situations and how use affects the interpretation of the sentence.
- **Discourse**: knowledge about linguistic units larger than a single utterance. It concerns how the immediately preceding sentences affect the interpretation of the next sentence

Common NLP tasks

- **Morphological analysis**
- **Word segmentation (Tokenization):** Separate a chunk of continuous text into separate words. For a language like English or Arabic, this is fairly trivial, since words are usually separated by spaces. However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language.
- **Lemmatization**: The task of removing inflectional endings only and to return the base dictionary form of a word which is also known as a lemma. Lemmatization is another technique for reducing words to their normalized form. But in this case, the transformation actually uses a dictionary to map words to their actual form.
- **Stemming**: The process of reducing inflected (or sometimes derived) words to a base form (e.g., "close" will be the root for "closed", "closing", "close", "closer" etc.). Stemming yields similar results as lemmatization, but does so on grounds of rules, not a dictionary.

Part-of-speech tagging:

- Given a sentence, determine the part of speech (POS) for each word (grammatical category of a word). Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb or adjective; and "out" can be any of at least five different parts of speech.

I want to print Ali's word file

I (pronoun)
want (verb)
to (prep)
to(infinitive)
print (verb)
Ali (noun)
's (possessive)
word (adj)
file (noun)

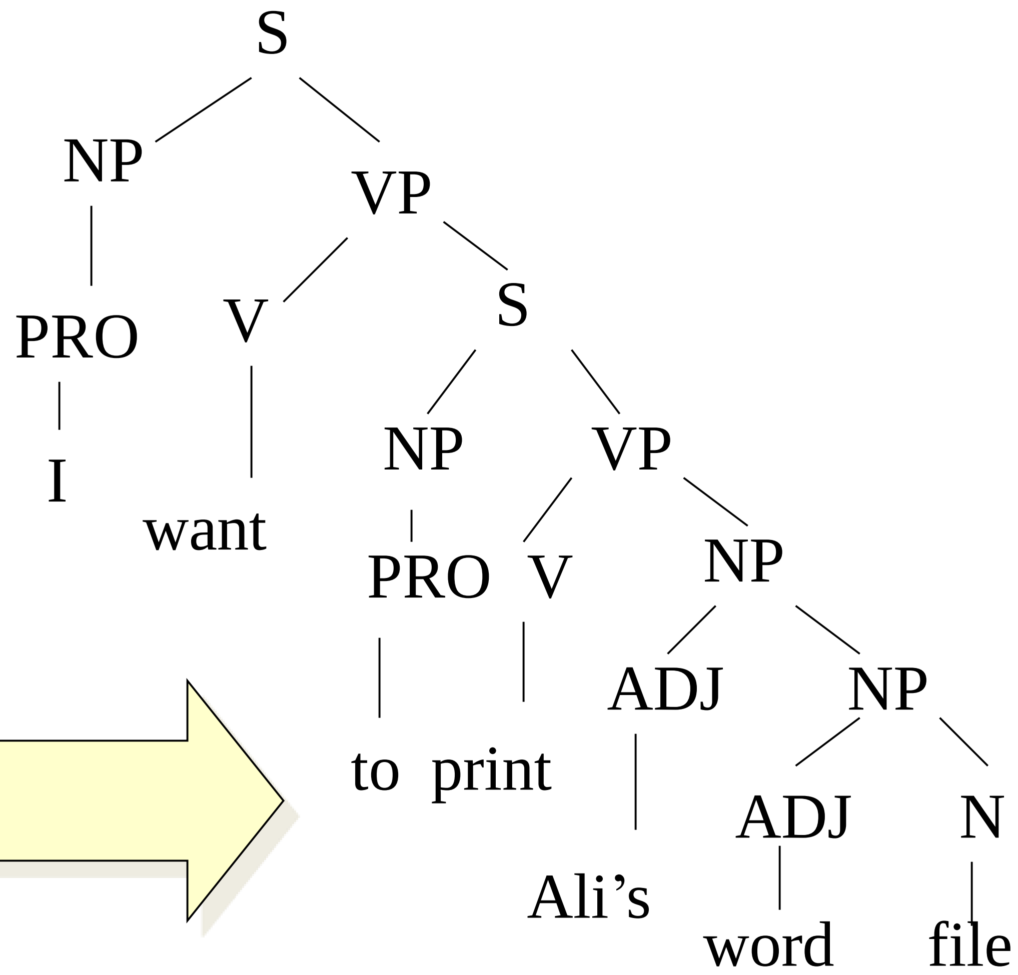
Syntactic analysis

Parsing: Determine the parse tree (grammatical analysis) of a given sentence. There are two primary types of parsing: dependency parsing and constituency parsing. **Dependency parsing** focuses on the relationships between words in a sentence (marking things like primary objects and predicates), whereas **constituency parsing** focuses on building out the parse tree using a probabilistic context-free grammar (PCFG)

- Assigning a syntactic and logical form to an input sentence
  - uses knowledge about word and word meanings (lexicon)
  - uses a set of rules defining legal structures (grammar)

Parse Tree

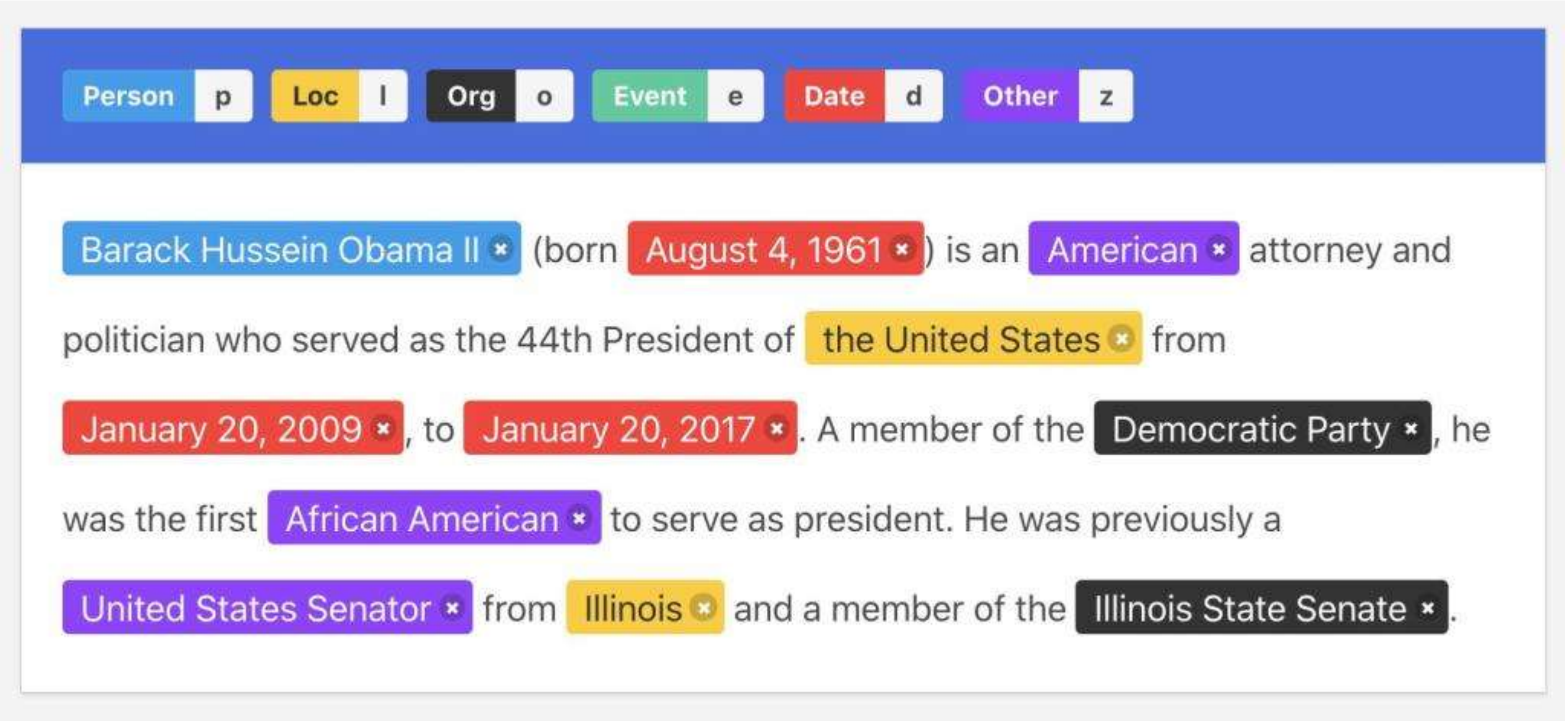
I (pronoun)
want (verb)
to (prep)
to(infinitive)
print (verb)
Ali (noun)
‘s (possessive)
word (adj)
file (noun)



Lexical semantics (of individual words in context)

- **Named entity recognition (NER):** Given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, organization).
- Although capitalization can aid in recognizing named entities in languages such as English, this information cannot aid in determining the type of named entity, and in any case, is often inaccurate or insufficient.
- For example, the first letter of a sentence is also capitalized, and named entities often span several words, only some of which are capitalized.
- Furthermore, many other languages in non-Western scripts (e.g. Chinese or Arabic) do not have any capitalization at all, and even languages with capitalization may not consistently use it to distinguish names.

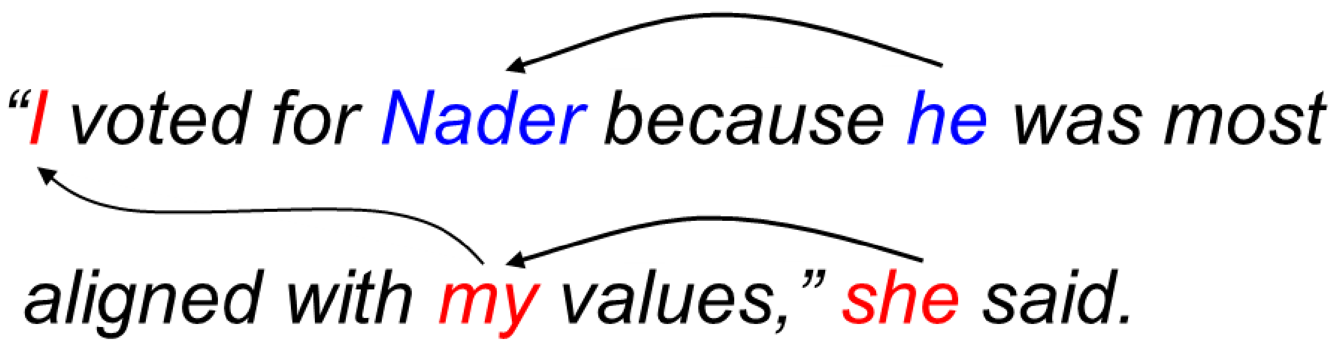
Example of the output of an NER tagger:



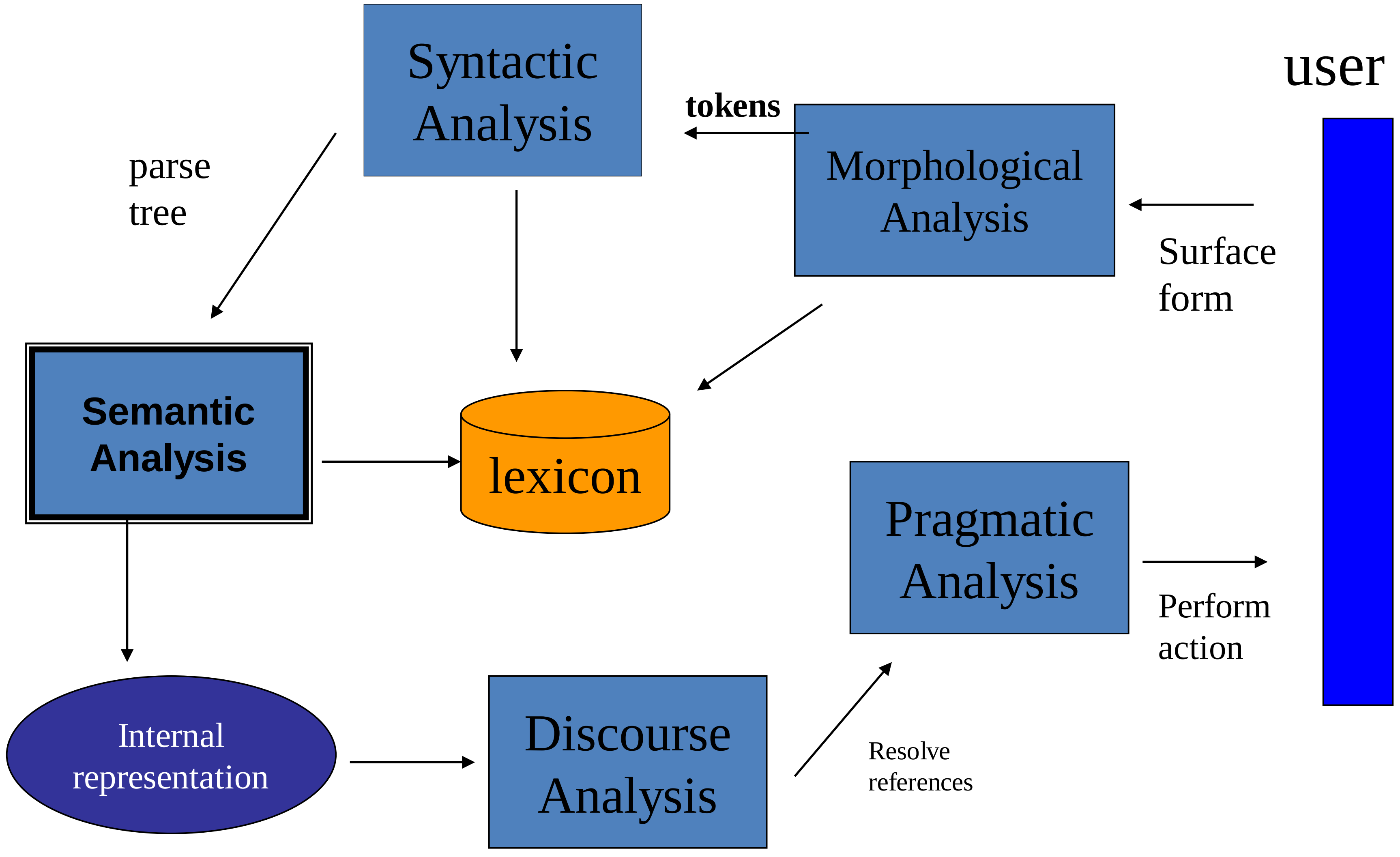
Discourse (semantics beyond individual sentences)

Coreference resolution

- Given a sentence or larger chunk of text, determine which words ("mentions") refer to the same objects ("entities").
- Anaphora resolution is a specific example of this task, and is specifically concerned with matching up pronouns with the nouns or names to which they refer.
- The more general task of coreference resolution also includes identifying so-called "bridging relationships" involving referring expressions.



NLP stages



Syntactic Analysis
tokens
Morphological Analysis
user
parse tree
Surface form
Semantic Analysis
lexicon
Pragmatic Analysis
Perform action
Internal representation
Discourse Analysis
Resolve references

Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations

Ambiguity

- More than one meaning for the same sentence
- Ambiguity at multiple levels
  - Word senses: **bank** (finance or river ?)
  - Part of speech: **chair** (noun or verb ?)
  - Syntactic structure: **I can see a man with a telescope**
  - Multiple: **I made her duck**

Ambiguity Example

I made her duck

لقد صنعت لها بطة

- I cooked waterfowl for her
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated duck

- لقد طهيت لها بطة.
- لقد قمت بطهي البطة التي تخصها
- لقد صنعت البطة (دمية) التي تمتلكها
- لقد جعلتها تخفض رأسها أو جسدها بسرعة.
- لوحت بعصا السحرية الخاصة بي وحولتها إلى طائر مائي غير متمايز

- First, the words ***duck*** and *her* are morphologically or syntactically ambiguous in their part-of-speech
- Duck can be verb or noun
- **Her** can be additive or possessive pronoun
- The word make is semantically ambiguous, it can mean create or cook

The Challenges of "Words"

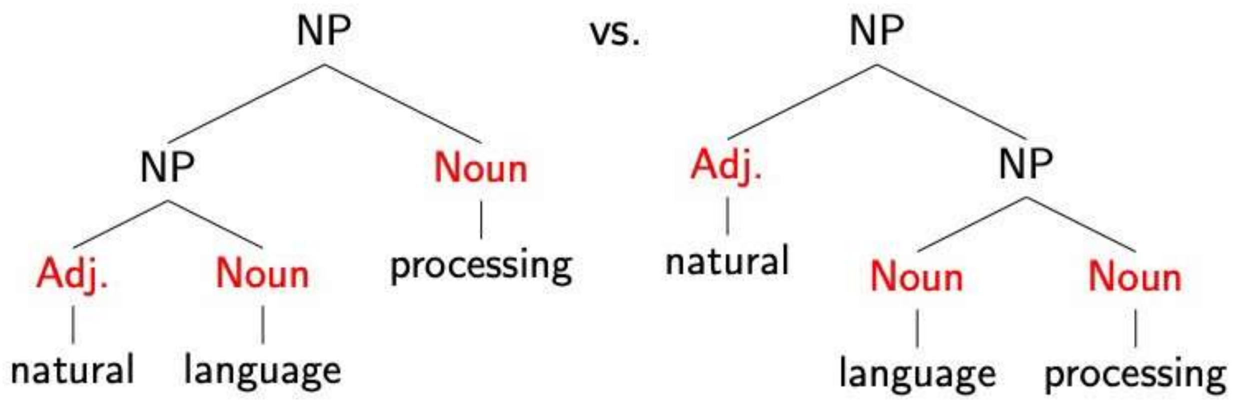
- Segmenting text into words
- Morphological variation
- Words with multiple meanings: bank, mean
- Domain-specific meanings: latex
- Multiword expressions: make a decision, take out, make up

Part of Speech Tagging

I know, right	shake my head			for	your		
ikr	smh	he	asked	fir	yo	last	name
!	G	O	V	P	D	A	N
interjection	acronym	pronoun	verb	prep.	det.	adj.	noun

				you		Facebook	laugh out loud
so	he	can	add	u	on	fb	lololol
P	O	V	V	O	P	^	!
preposition						proper noun	

Syntax


NP
vs.
NP
NP
Noun
Adj.
NP
Adj.
Noun
processing
natural
Noun
Noun
natural
language
language
processing

Morphology + Syntax

A ship-shipping ship, shipping shipping-ships

سفينة شحن السفن ، تشحن سفن الشحن

Syntax + Semantics

- We saw the woman with the telescope wrapped in paper.
  - Who has the telescope?
  - Who or what is wrapped in paper?
  - An even of perception, or an assault?

Dealing with ambiguity

- How can we model ambiguity and choose correct analysis in context?
  - Non-probabilistic methods return all possible analyses.
  - Probabilistic models return best possible analysis, i.e. most probable one according to the model.

But the “best” analysis is only good if our probabilities are accurate. Where do they come from?

Sparsity

- Sparse data due to Zipf's Law
- To illustrate, let's look at the frequencies of different words in a large text corpus
- Assume "word" is a string of letters separated by spaces
- Most frequent words in the English Europarl corpus (out of 24m word tokens)

any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

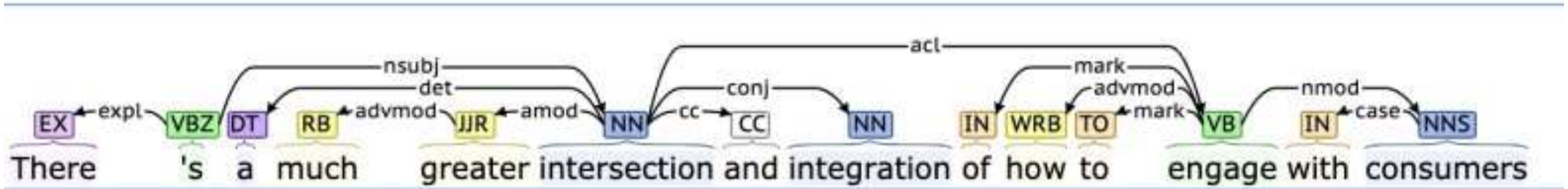
- Zipf's Law
  - Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words
  - This means we need to find clever ways to estimate probabilities for things we have rarely or never seen

Word counts

- Out of 93,638 distinct words (**types**), 36,231 (~40%) occur only once.
- Examples:
  - cornflakes, mathematicians, fuzziness, jumbling
  - pseudo-rapporteur, lobby-ridden, perfunctorily
  - Lycketoft, UNCITRAL, H-0695
  - policyfor, Commissioneris, 145.95, 27a

Variation

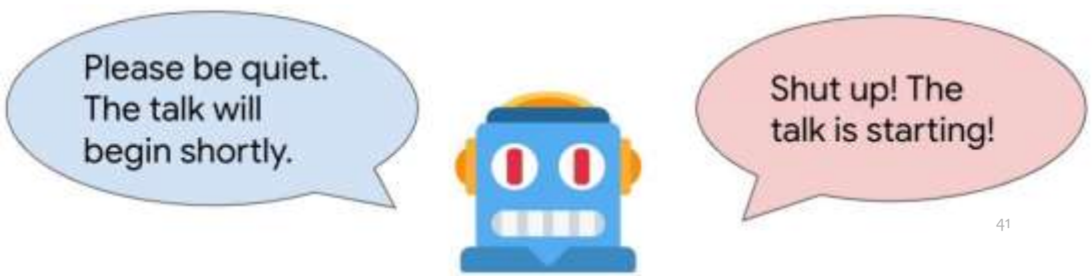
- Suppose we train a part of speech tagger or a parser on the Wall Street Journal…



- What will happen if we try to use this tagger/parser for social media?
  - *"ikr smh he asked fir yo last name so he can add u on fb lololol"*

Expressivity

- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:
  - *She gave the book to Tom* vs. *She gave Tom the book*
  - *Some kids popped by* vs. *A few children visited*
  - *Is that window still open?* vs. *Please close the window*

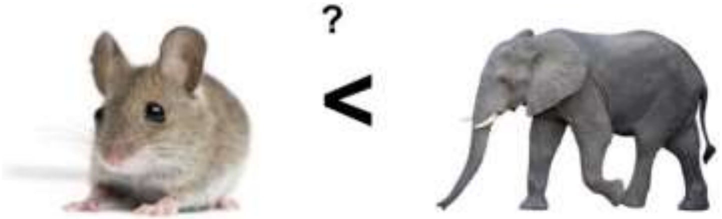


Unmodeled variables

- World knowledge
  - I dropped the glass on the floor and it broke
  - I dropped the hammer on the glass and it broke

“drink this milk.”

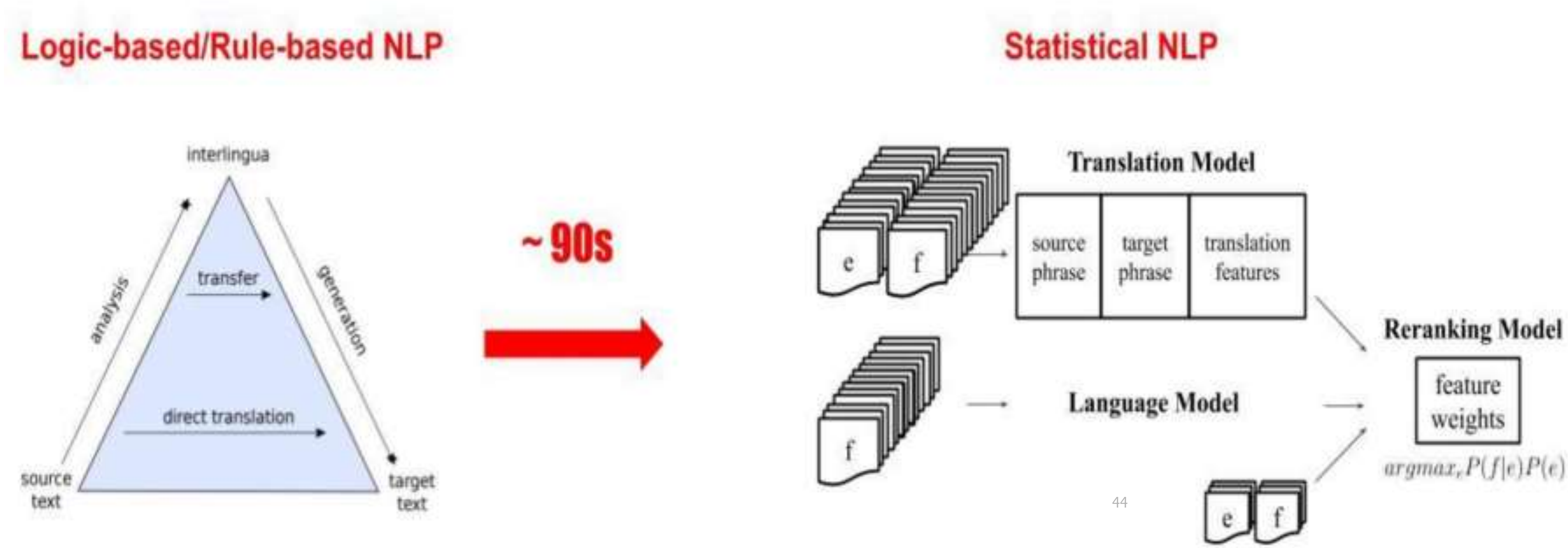
skater eats pavement



Unknown representation

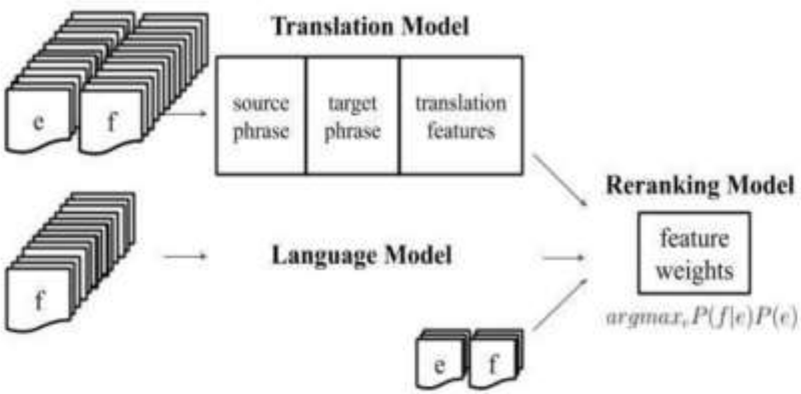
- Very difficult to capture what is RULES, since we don't even know how to represent the knowledge a human has/needs:
  - What is the "meaning" of a word, sentence, utterance?
  - How to model context?
  - Other general knowledge?

Symbolic and Probabilistic NLP



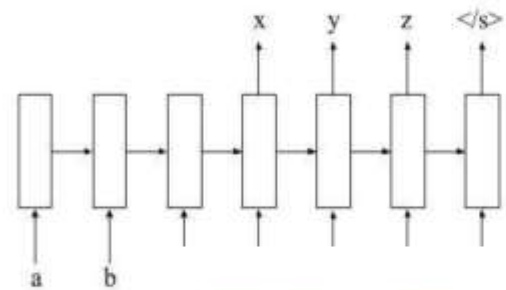
Probabilistic and Connectionist NLP

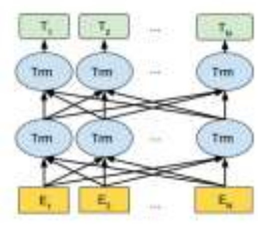
Engineered Features/Representations



~mid 2010s

Learned Features/Representations





NLP vs. Machine Learning

- NLP focuses on the understanding, processing, and generation of human language, while ML is a broader field that encompasses the development of **algorithms** and models that can learn from data and make predictions or perform tasks.
- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.