CS 481

Artificial Intelligence Language Understanding

January 12, 2023

Announcements / Reminders

- **Please follow the Week 01 To Do List instructions**

Plan for Today

- **Introduction to NLP**
- **Language basics**

Optional Textbook Online

Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

Here's our Dec 29, 2021 draft! This draft includes a large portion of our new Chapter 11, which covers BERT and fine-tuning, augments the logistic regression chapter to better cover softmax regression, and fixes many other bugs and typos throughout (in addition to what was fixed in the September draft, which added various missing sections (more on transformers, including for MT, various updated algorithms, like for dependency parsers, etc.).

Minor update: the Dec 29, 2021 draft had some errors in PDF files for chapters 9, 10, and 11, that made them unviewable in some PDF readers. On Jan 12, 2022 we updated the pdfs for the entire book and for those 3 chapters, so if you downloaded earlier, you might want to redownload. The content did not change from the Dec 29 2021 draft.

We've put up a list here of the amazing people who have sent so many fantastic suggestions and bug-fixes for improving the book. We are really grateful to all of you for your help, the book would not be possible without you!

Individual chapters are below; here is a single pdf of all the chapters in the Jan 12, 2022 draft of the book so far!
(This is the same exact content as the Dec 29, 2021 draft but with some PDF errors fixed that prevented Adobe Acrobat Reader from displaying some figures in chapters 9, 10, and 11)

Feel free to use the draft chapters and slides in your classes, the resulting feedback we get from you makes the book better!

As always, typos and comments very welcome (just email slp3edbugs@gmail.com and let us know the date on the draft)!

(Don't bother reporting missing refs due to cross-chapter cross-reference problems in the indvidual chapter pdfs, those are fixed in the full book draft)

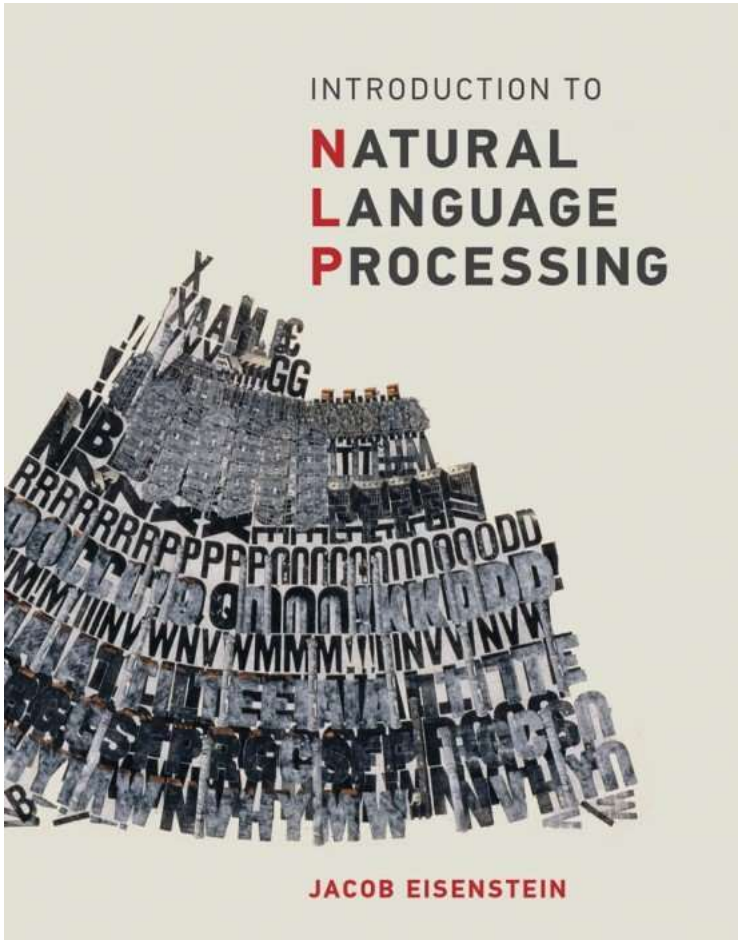
When will the whole book be finished? Don't ask.

If you need last year's December 2020 draft chapters, they are here; the September 2021 draft chapters are here.

	Chapter	Slides	Relation to 2nd ed.
1:	Introduction		[Ch. 1 in 2nd ed.]
2:	Regular Expressions, Text Normalization, Edit Distance	2: Text Processing [pptx] [pdf] 2: Edit Distance [pptx] [pdf]	[Ch. 2 and parts of Ch. 3 in 2nd ed.]
3:	N-gram Language Models	3: N-grams [pptx] [pdf]	[Ch. 4 in 2nd ed.]
4:	Naive Bayes and Sentiment Classification	4: Naive Bayes + Sentiment [pptx] [pdf]	[new in this edition]
5:	Logistic Regression	5: LR [pptx] [pdf]	[new in this edition]

Source: https://web.stanford.edu/~jurafsky/slp3/

Optional Textbook Online (Notes)



master ▾ gt-nlp-class / notes /

jacobeisenstein Update errata.md ... on Feb 8, 2021 History

..	
slides	3 years ago
eisenstein-nlp-notes-10-15-2018.pdf	3 years ago
eisenstein-nlp-notes-june-1.pdf	4 years ago
eisenstein-nlp-notes.pdf	3 years ago
errata.md	11 months ago
readme.md	12 months ago

readme.md

About these notes

This textbook was designed for the courses CS 4650 and CS 7650 ("Natural Language") at Georgia Tech. The latest version is eisenstein-nlp-notes.pdf.

Source: https://github.com/jacobeisenstein/gt-nlp-class/tree/master/notes

Natural Language Processing (NLP)

Definition:

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the **interactions between computers and human language**, in particular **how to program computers to process and analyze large amounts of natural language data**.

Involves:

- Speech processing
- Natural language understanding
- Natural language generation

Computers vs Language and Speech

- **Text processing:** engineering practices for transforming, normalizing, compressing or accessing textual data
- **Natural language understanding / processing:** the study of methods for exploiting or generating language represented as text, for practical tasks
- **Computational linguistics:** the use of computational tools to understand or learn the structure of human languages
- **Speech processing:** The study of methods for exploiting or generating language represented as audible waveforms, for practical tasks

Primary Reasons for NLP

- To enable human-machine communication
- To learn from written sources
  - most information (80-90% or more) in most organizations is in natural language (reports, order forms, bulletin boards, email, web pages, video, audio, etc.) and not in a traditional database!
  - most of that information is now digital
    - Estimate in 1998: ~60%
    - Now, more than 90%!
- To advance the scientific understanding of languages and language use

What Are Key NLP Goals?

- Long range perspective:
  - True understanding of natural language
  - Deep reasoning about texts
  - Real-time spoken dialogue / translation
- Engineering perspective:
  - Extract useful facts from documents
  - Search the web
  - Better spelling / grammar checking
  - etc.

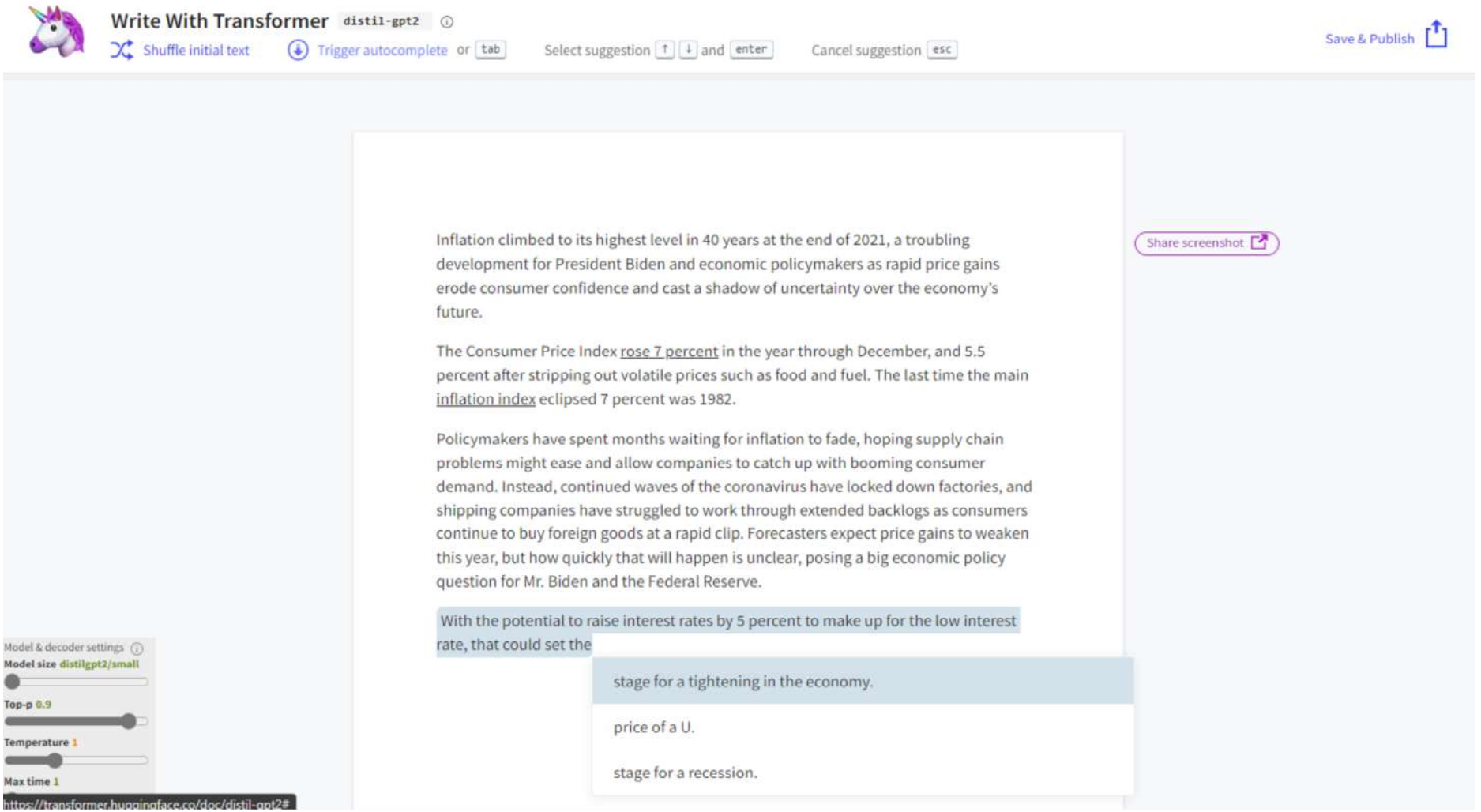
Core NLP Applications

- **Language modeling:** the task of predicting what the next word in a sentence will be based on history of previous words. Its goal is to learn the probability of a sequence of words appearing in a given language.
- **Text classification:** the task of bucketing the text into a known set of categories based on its content.
- **Information extraction:** the task of extracting relevant information from text.
- **Information retrieval:** the task of finding documents / data relevant to a specific query from a large collection.

Core NLP Applications

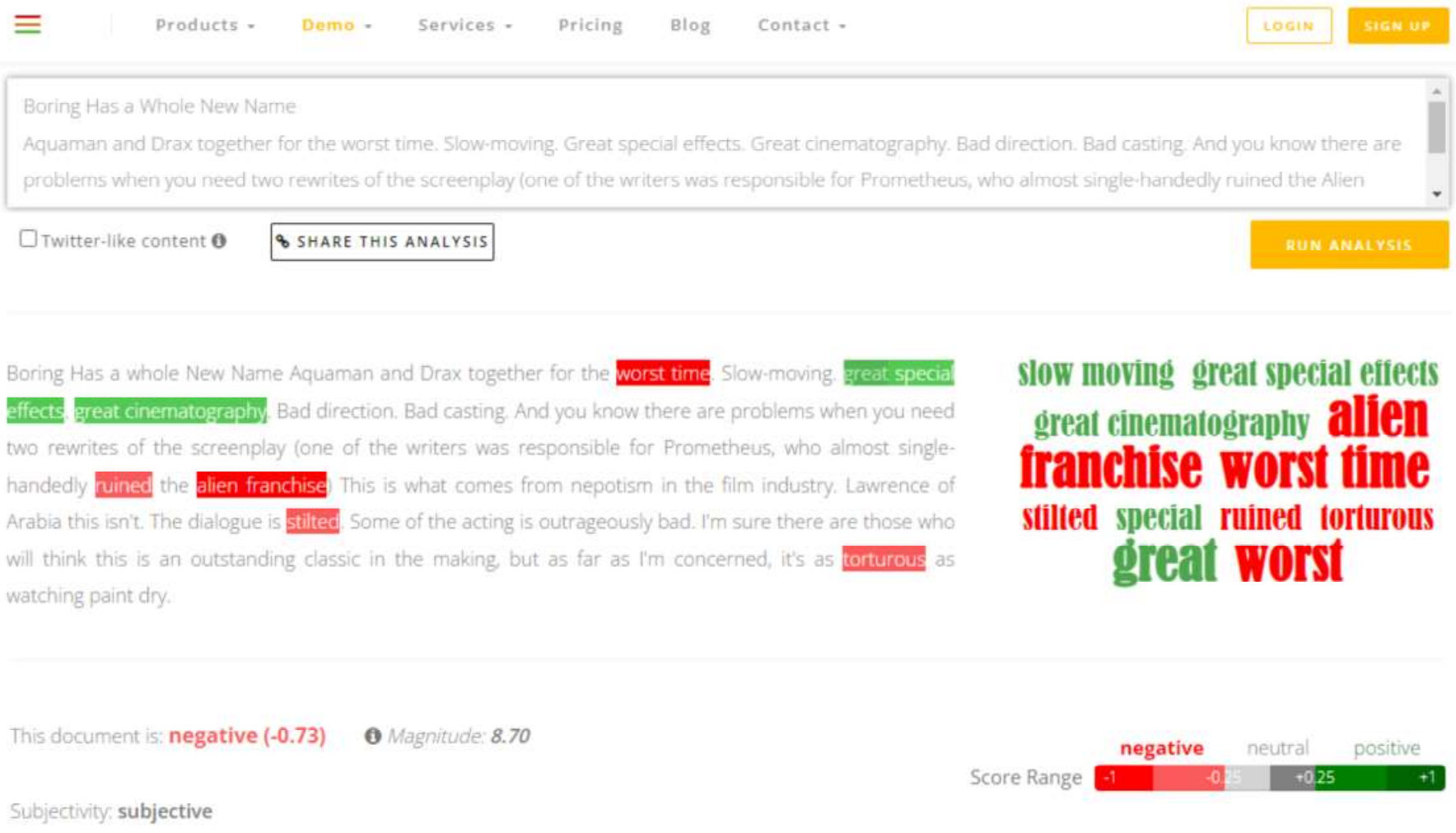
- **Conversational agent:** the task of building a dialogue systems that can converse in human languages.
- **Text summarization:** this task aims to create short summaries of longer documents while preserving the core content and meaning of the text.
- **Question answering:** the task of building a system that can answer questions posed in natural language.
- **Machine translation:** the task of converting a piece of text from one language to another.
- **Topic modeling**: the task of uncovering the topical structure of a large collection of documents.

Language Modeling Example



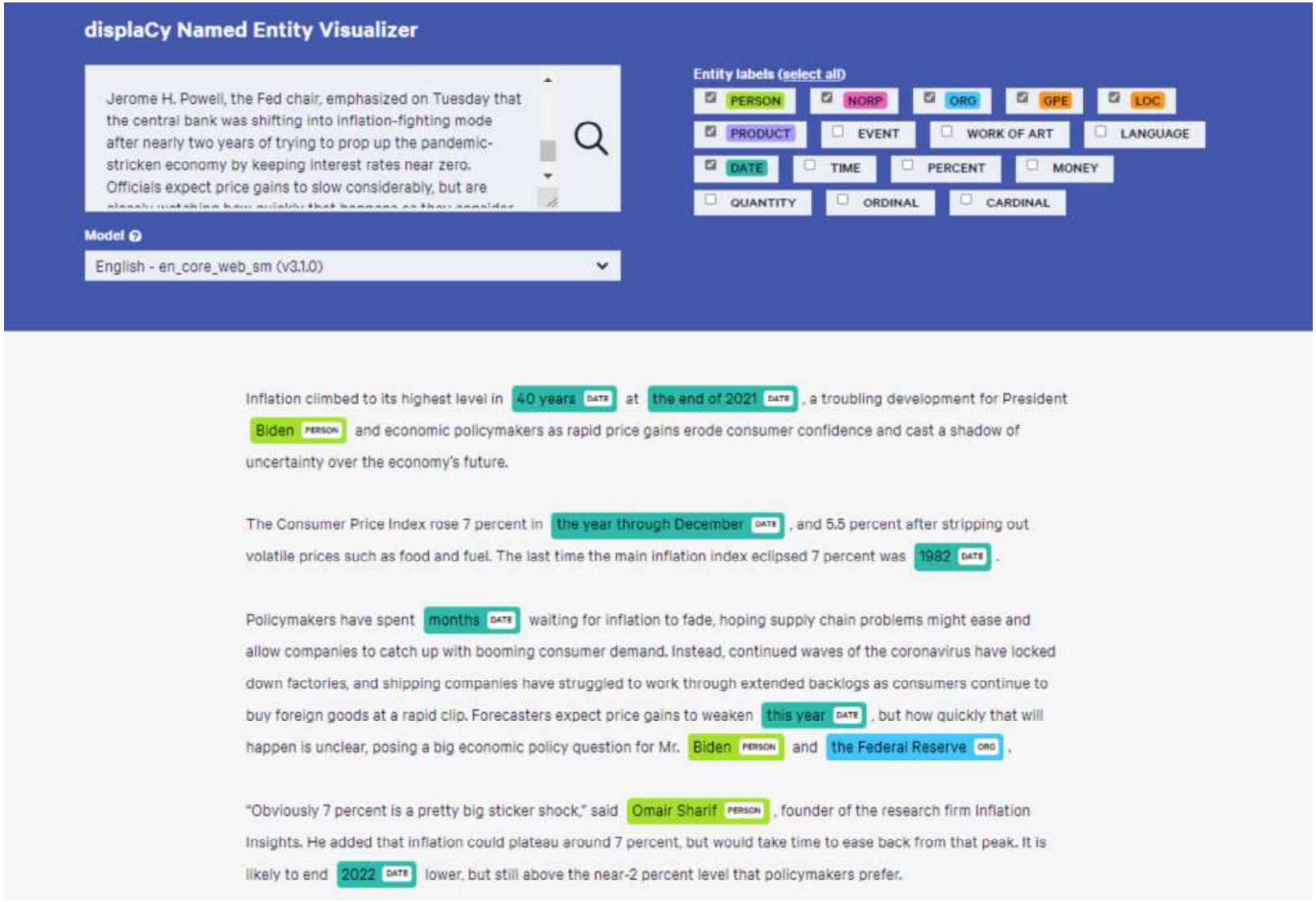
Source: https://transformer.huggingface.co/doc/distil-gpt2

Text Classification Example



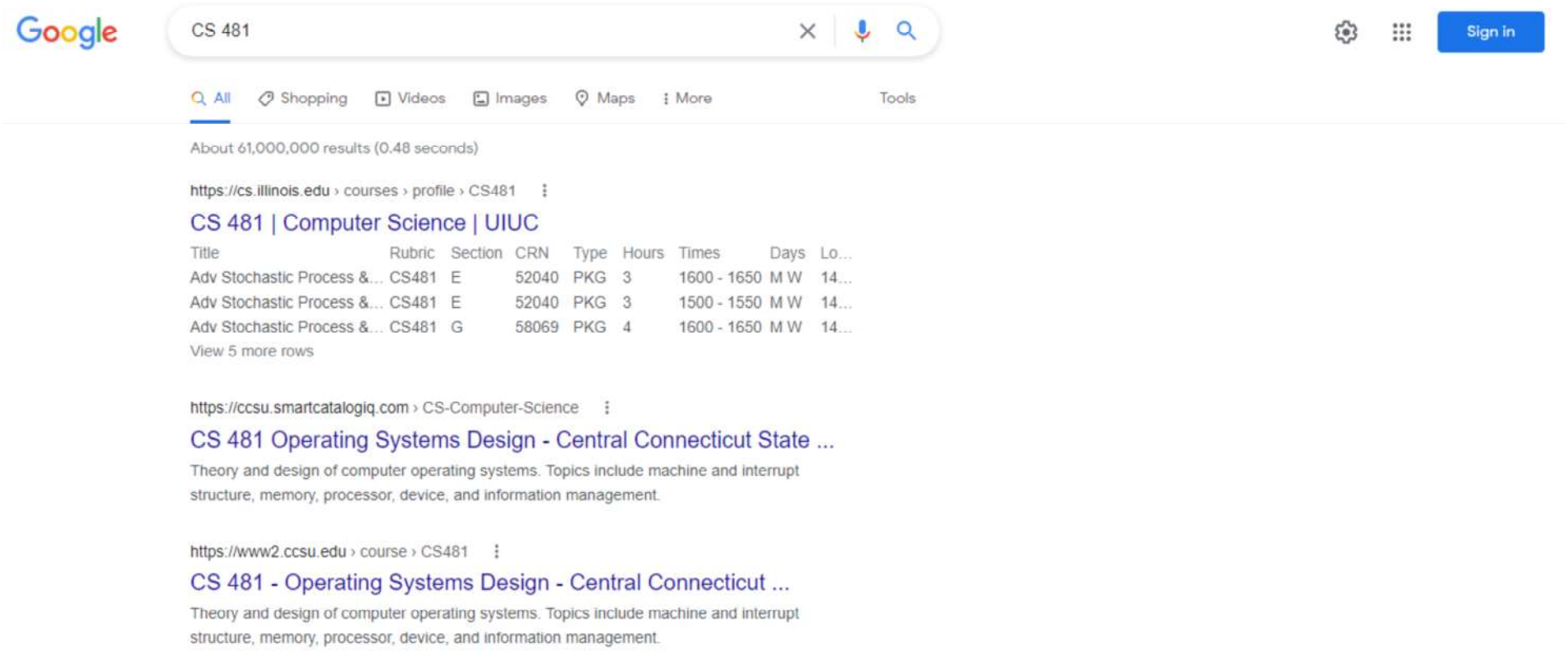
Source: https://text2data.com/Demo

Information Extraction Example



Source: https://explosion.ai/demos/displacy-ent

Information Retrieval/Search Example



Source: https://www.google.com/

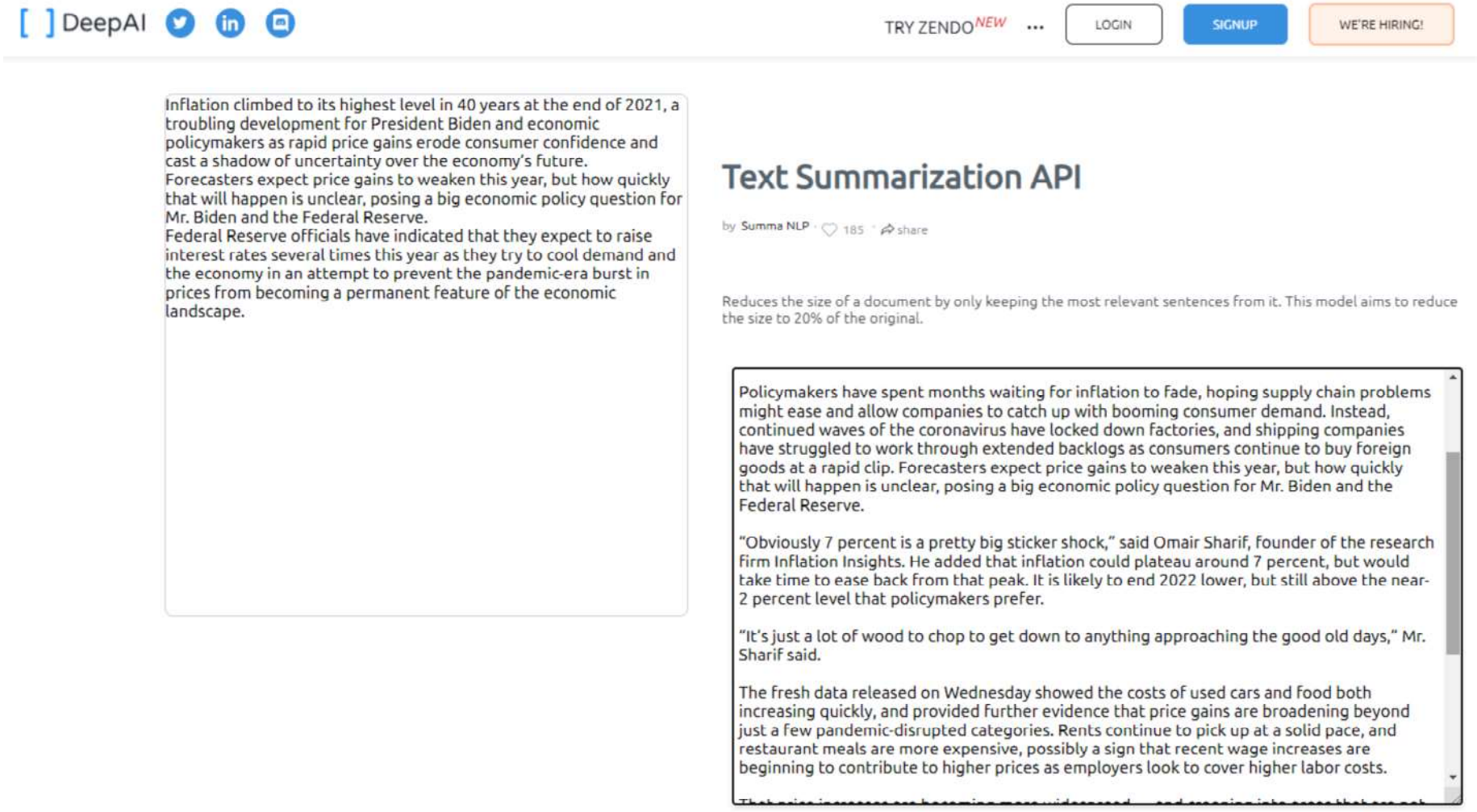
Conversational Agent Example

Meet Cimon, the floating AI astronaut

14,136 views • Feb 27, 2018 53 4 SHARE SAVE ...

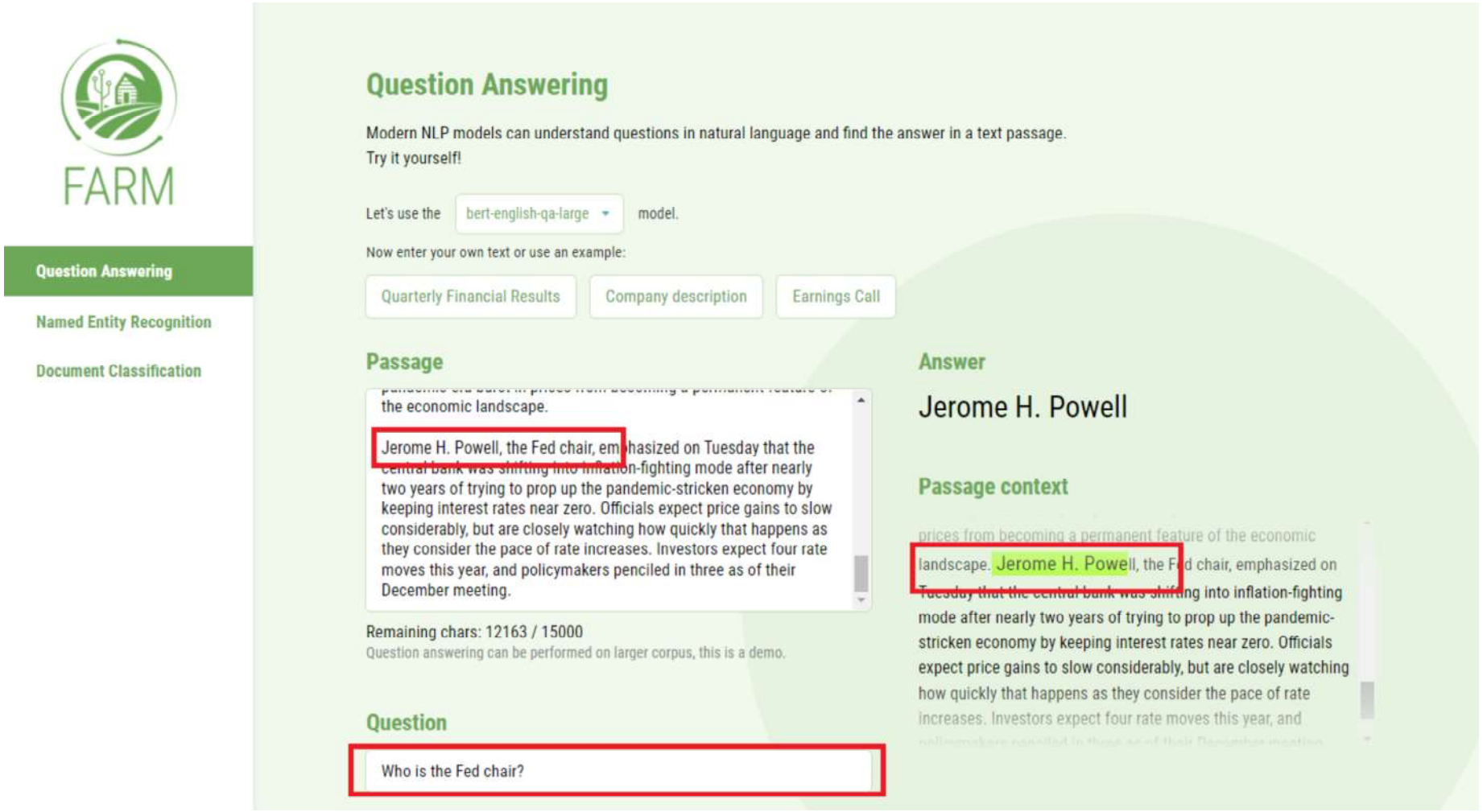
Source: https://www.youtube.com/watch?v=BRWJFQlNXHI

Text Summarization Example



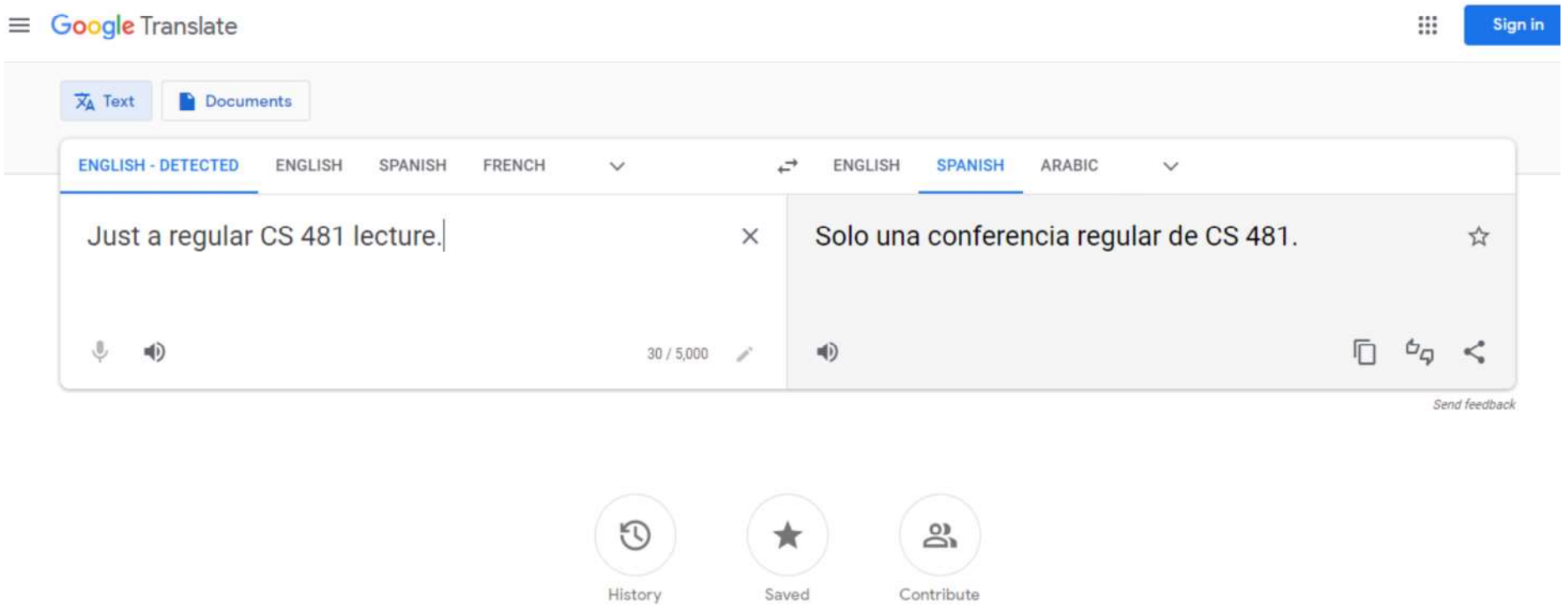
Source: https://deepai.org/machine-learning-model/summarization

Question Answering Example



Source: https://demo.deepset.ai/

Machine Translation Example



Selected Real-world NLP Applications

Applications	Text classification	Information extraction	Conversational Agent	Information retrieval	Question answering systems
General applications	Spam classification	Calendar event extraction	Personal assistants	Search engines	Jeopardy!
Industry specific applications	Social media analysis	Retail catalog extraction	Health record analysis	Financial analysis	Legal entity extraction

NLP Tasks in Order of Difficulty

What is Language?

Language is a structured system of communication that involves complex combinations of its consituent components, such as characters, words, sentences, etc.

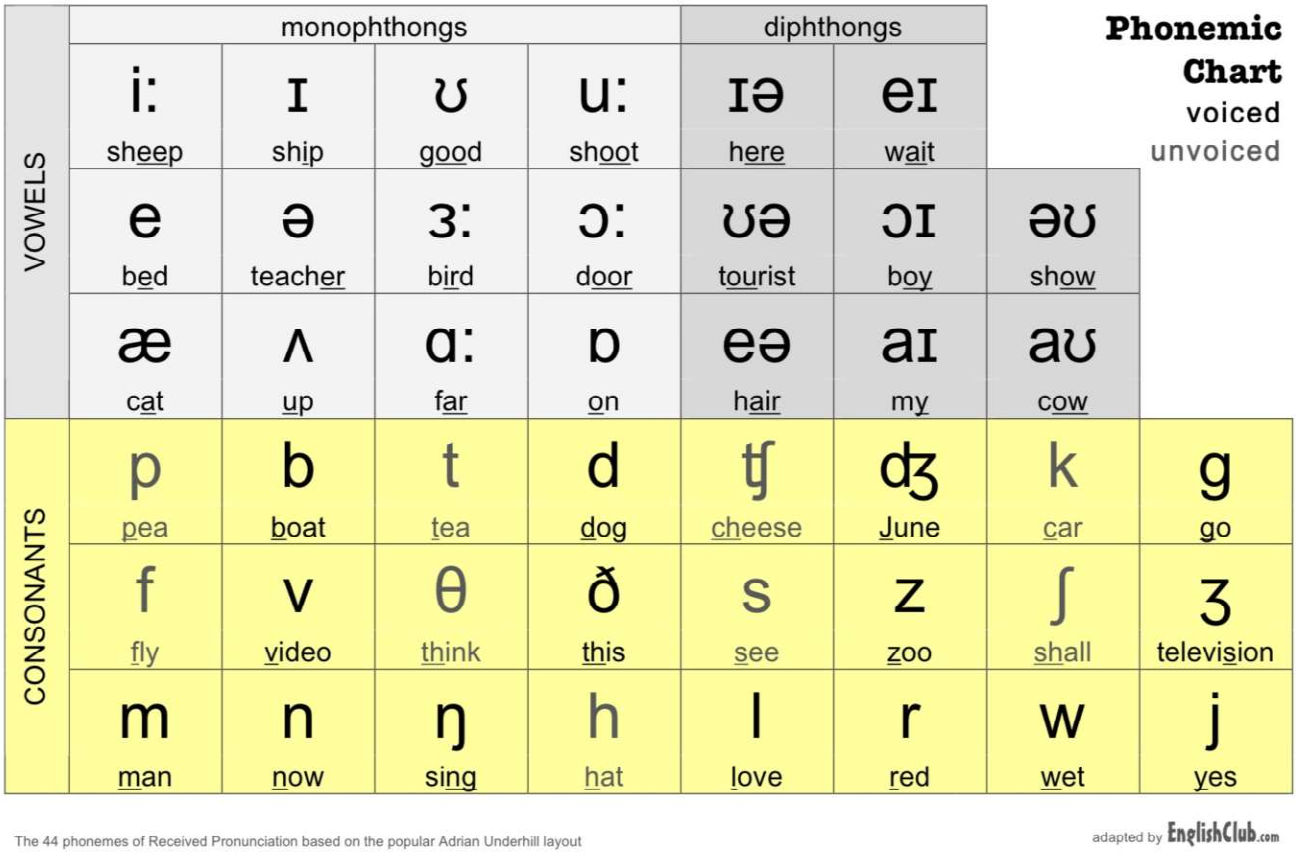
We can think of human language as composed of four major building blocks:

- **phonemes**
- **morphemes and lexemes**
- **syntax**
- **context / semantics**

Language: Phonemes

Phonemes are the smallest units of sound in a language.

They may not have any meaning by themselves.



Source: https://www.englishclub.com/pronunciation/phonemic-chart.htm

Language: Morphemes and Lexemes

A morpheme is the smallest unit of language that has meaning. It is a combination of phonemes. Not all morphemes are words. All prefixes and suffixes are morphemes.

unbreakble	*cats*
un + break + able	cat + s

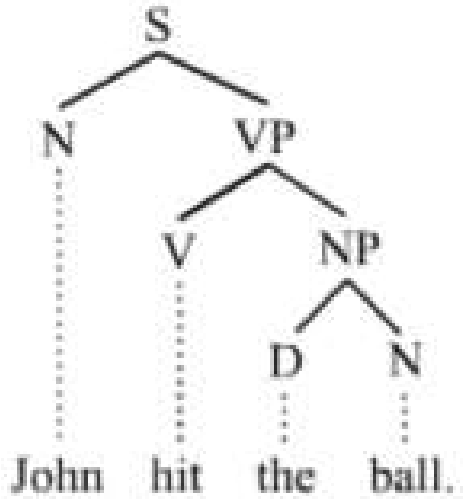
Lexemes are structural variations of morphemes related to one another by meaning. For example

run **and** *running*

belong to the same lexeme form.

Language: Syntax

Syntax is a set of rules used to construct grammatically correct sentences out of words and phrases in a language. A common approach to representing sentences is a parse tree.



Legend:

S for sentence

NP for noun phrase

VP for verb phrase

V for verb

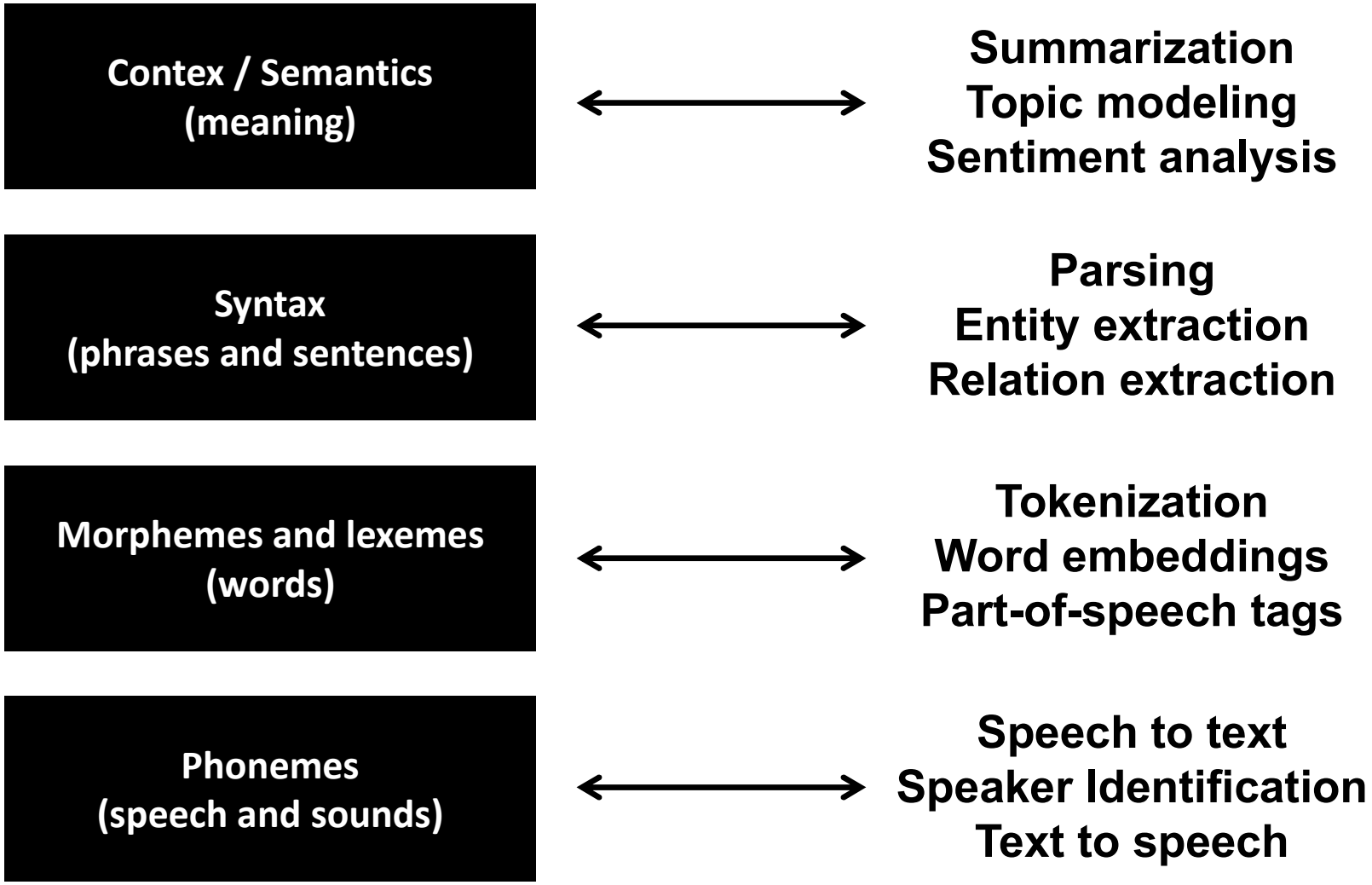
D for determiner, in this instance the definite article "the"

N for noun

Language: Context / Semantics

Context / semantics is how various parts in a language come together to convey a particular meaning. Context incluedes long-term references, world knowledge, and common sense along with the literal meaning of words and phrases.

Blocks of Language | Applications



Key NLP Tasks: Syntax

- **Lemmatization**: reducing the various inflected forms of a word into a single form for easy analysis.
- **Morphological segmentation**: dividing words into individual units called morphemes. undivided -> un - divided
- **Word segmentation**: dividing a large piece of continuous text into distinct units.
- **Part-of-speech tagging**: identifying the part of speech for every word.
- **Parsing**: grammatical analysis for the provided sentence.
- **Sentence breaking**: placing sentence boundaries on a large piece of text.
- **Stemming**: It involves cutting the inflected words to their root form.

Key NLP Tasks: Semantics

- **Named entity recognition (NER)**: determining the parts of a text that can be identified and categorized into preset groups. Examples of such groups include names of people and names of places.
- **Word sense disambiguation**: giving meaning to a word based on the context.
- **Natural language generation**: using databases to derive semantic intentions and convert them into human language.

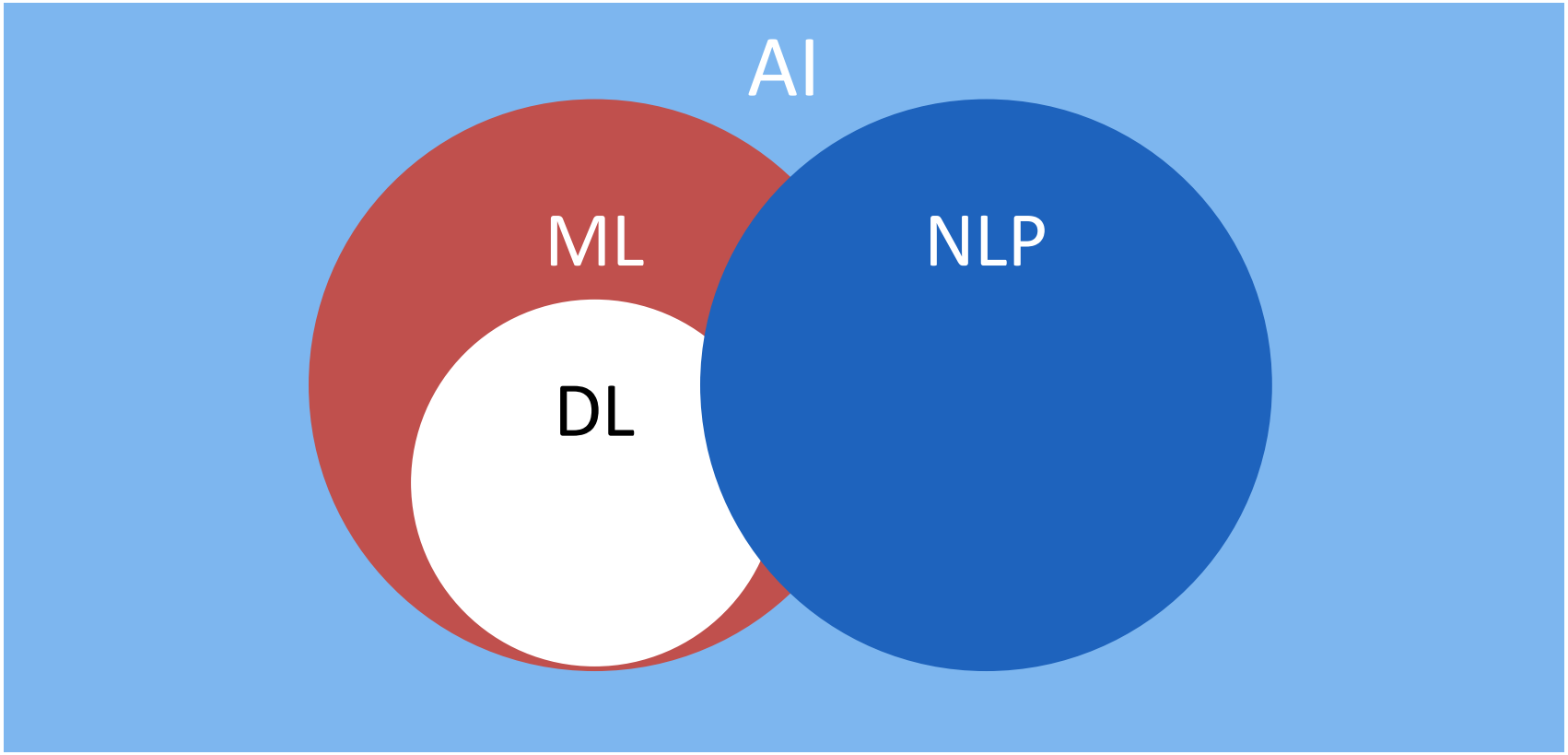
Why is NLP Hard?

- Complexity
- Ambiguity
  - "I made her duck"
- Common knowledge is required for understanding
- Fuzzy and probabilistic
- Creativity
- Diversity
- "Living" / evolving languages
  - neologisms, etc.

Approaches to NLP

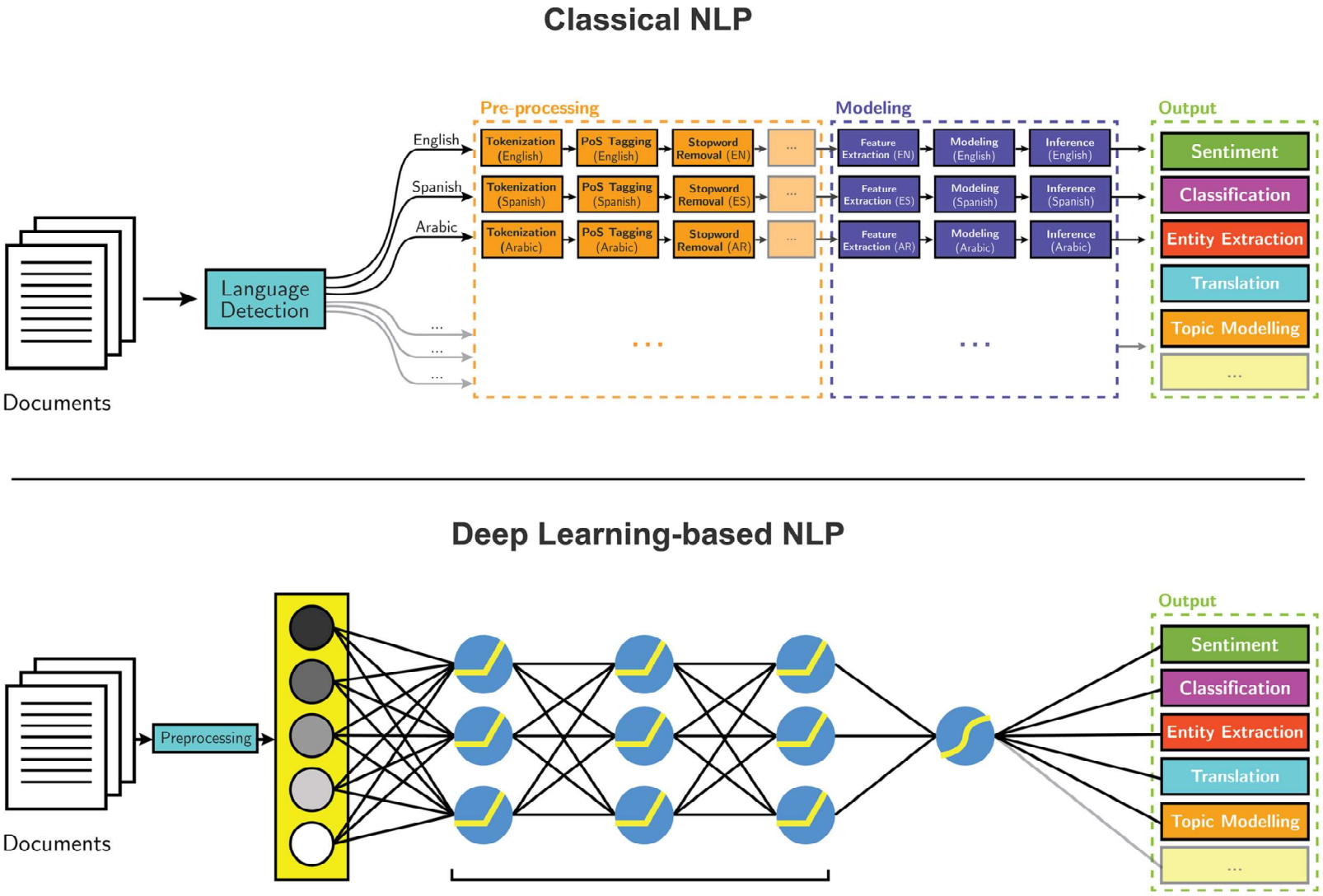
- Heuristics-based NLP
- Machine Learning NLP
  - Naive Bayes
  - Support Vector Machines
  - Hidden Markov Model
  - Conditional Random Fields
- Deep Learning NLP
  - Recurrent neural networks
  - Long-sort term memory
  - Convolutional neural networks
  - Transformers and autoencoders

AI vs ML vs NLP



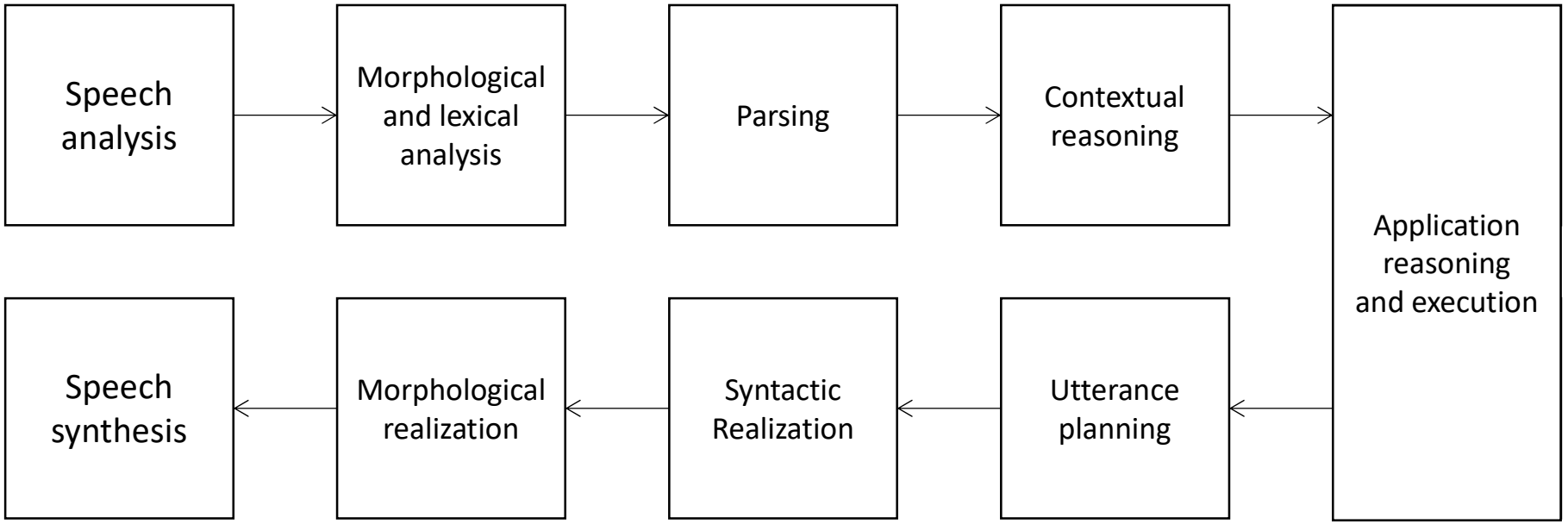
AI - Artificial Intelligence
ML - Machine Learning
DL - Deep Learning
NLP - Natural Language Processing

Classical NLP vs Deep Learning NLP

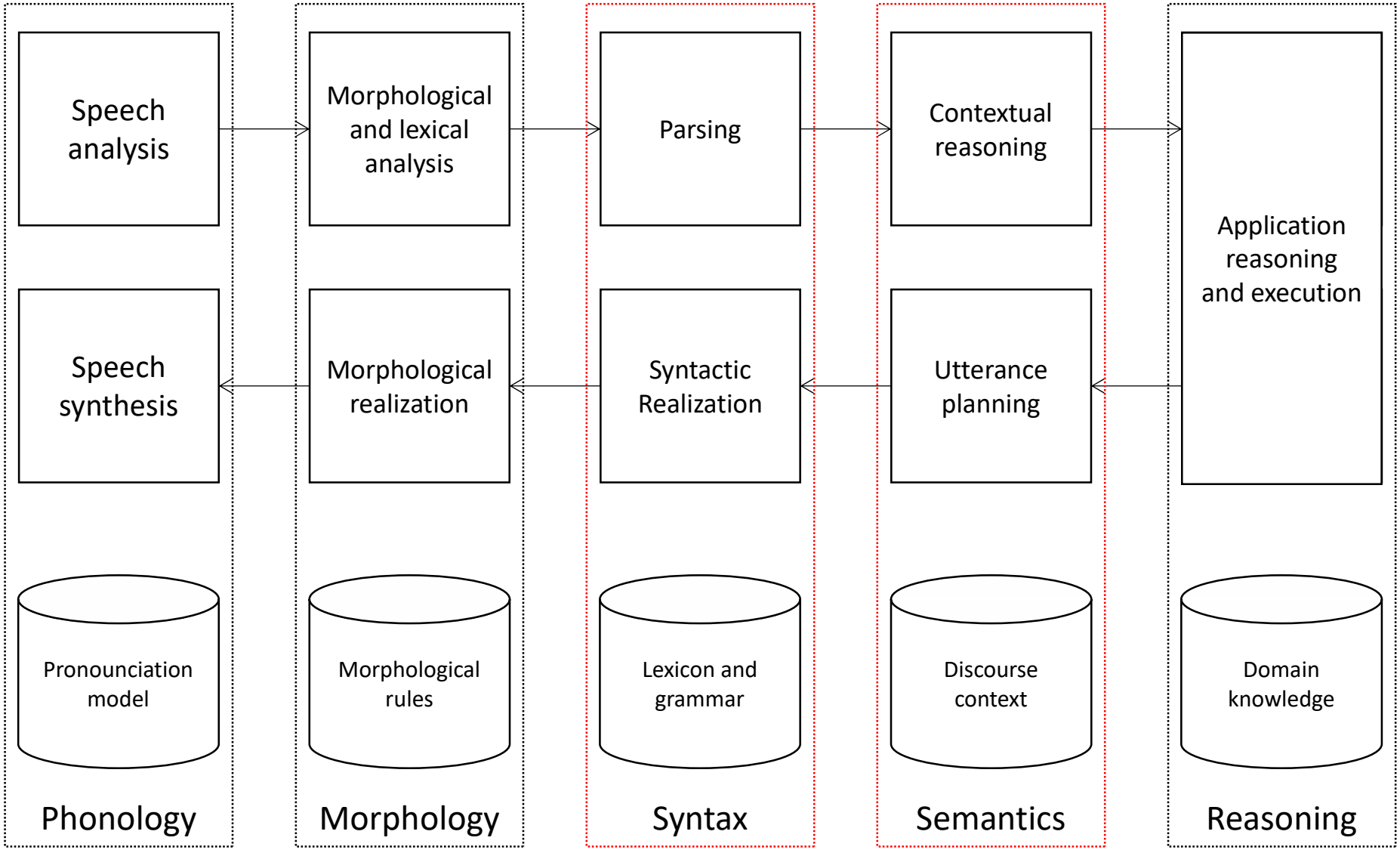


Source: https://www.oreilly.com/library/view/python-natural-language/9781787121423/6f015f49-58e9-4dd1-8045-b11e7f8bf2c8.xhtml

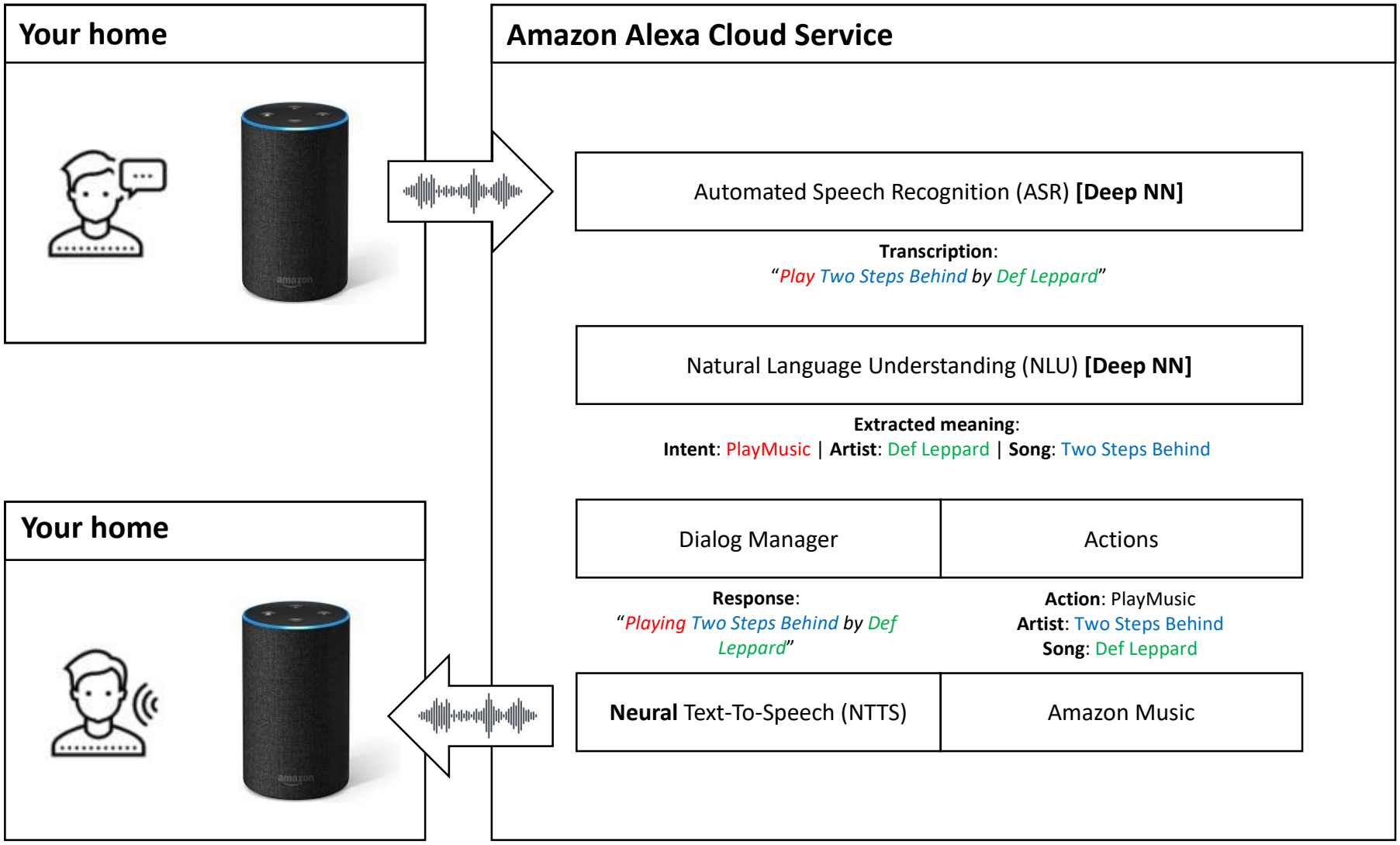
Basic NLP Spoken Language Pipeline



Basic NLP Spoken Language Pipeline



Voice Assistant: Alexa

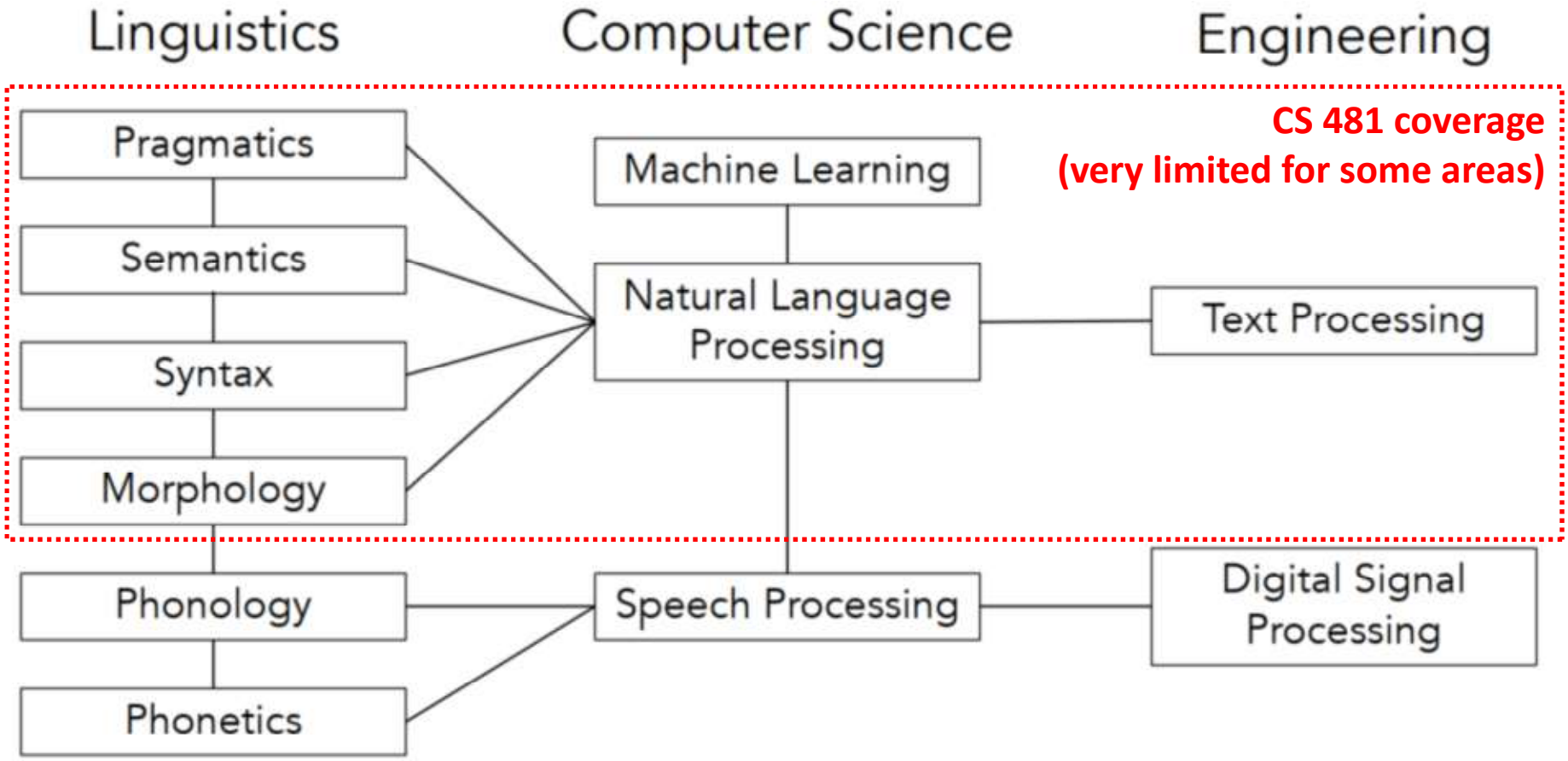


Generative Pre-trained Transformer 3

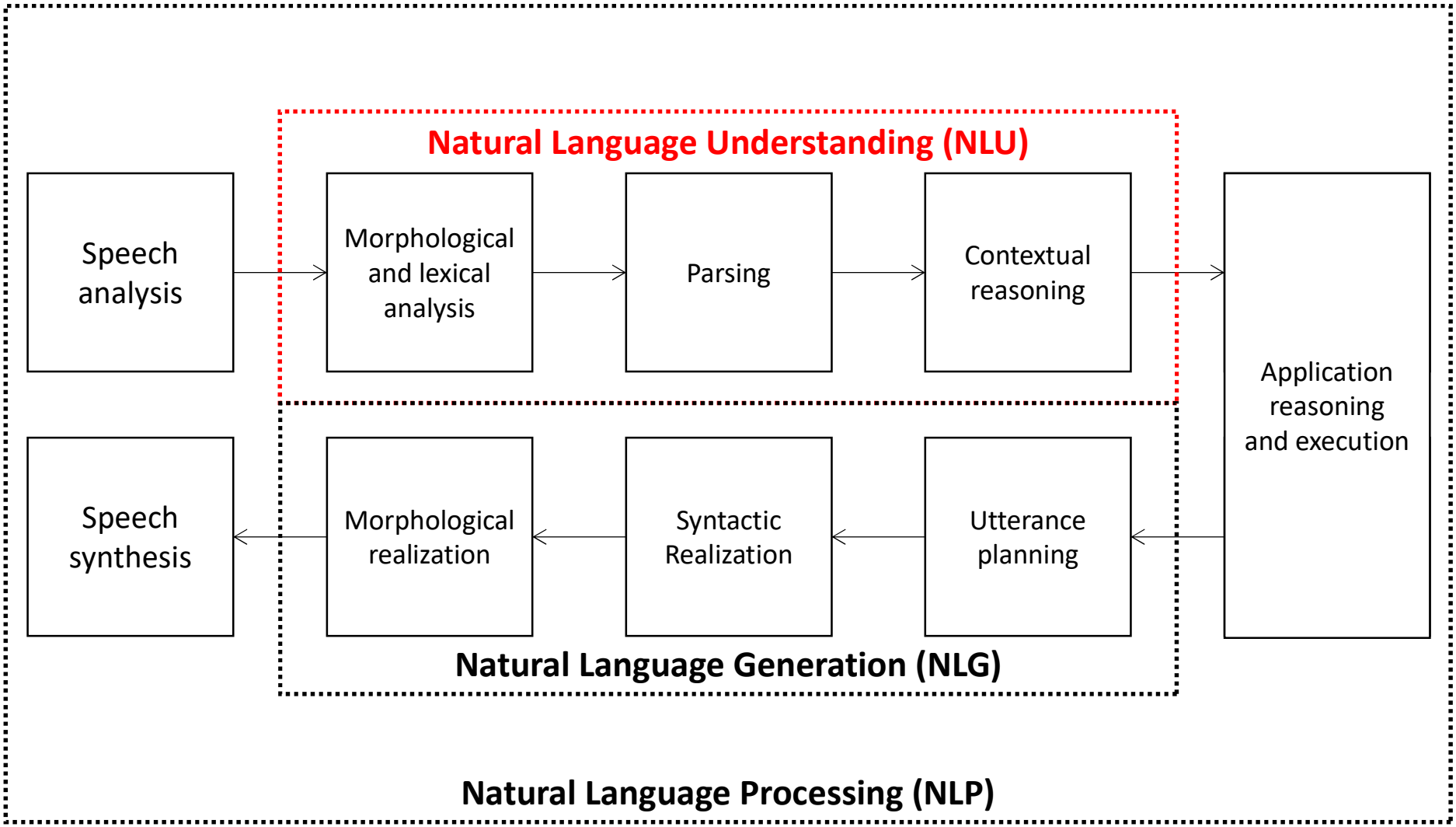
What is it?

Generative Pre-trained Transformer 3 (GPT-3) is an autoregressive **language model that uses deep learning to produce human-like text**. It is the third-generation language prediction model in the GPT-n series (and the successor to GPT-2) created by OpenAI, a San Francisco-based artificial intelligence research laboratory. GPT-3's full version has a capacity of **175 billion machine learning parameters**.

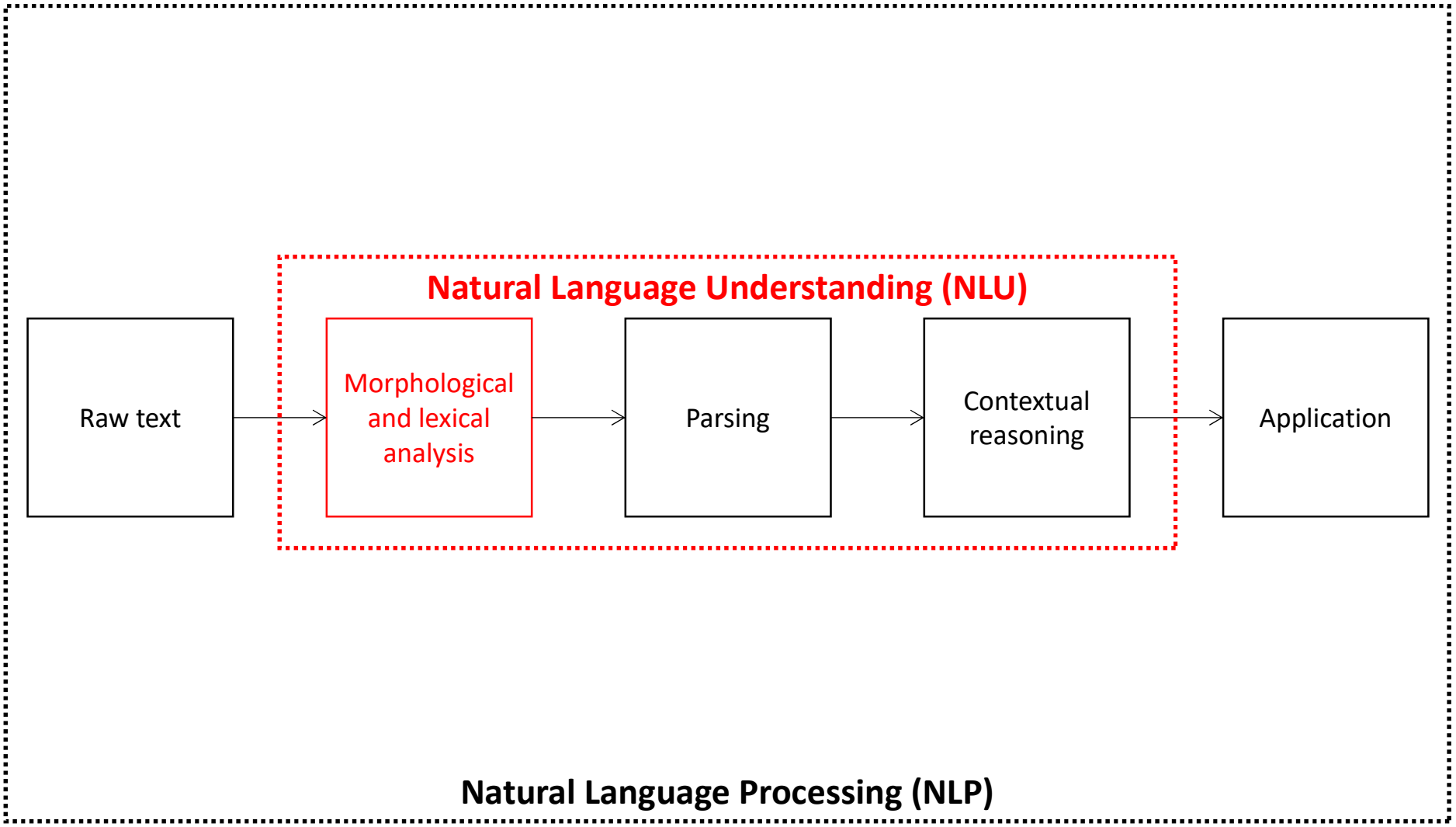
NLP vs. Adjacent Fields



Basic NLP Spoken Language Pipeline



Basic NLP Text Processing Pipeline



Common Lexical Categories

Lexical category	Definition*	Example
Adjective	A word or phrase naming an attribute, added to or grammatically related to a noun to modify or describe it	The ***quick red*** fox jumped over the ***lazy brown*** dogs.
Adverb	A word or phrase that modifies or qualifies an adjective, verb, or other adverb, or a word group, expressing a relation of place, time, circumstance, manner, cause, degree, etc.	The dogs ***lazily*** ran down the field after the fox.
Conjunction	A word that joins two words, phrases, or clauses	The quick red fox ***and*** the silver coyote jumped over the lazy brown dogs.
Determiner	A modifying word that determines the kind of reference a noun or noun group has, for example ***a***, ***the***, ***very***	***The*** quick red fox jumped over ***the*** lazy brown dogs.
Noun	A word used to identify any of the class of people, places, or things, or to name a particular one of these.	The quick red ***fox*** jumped over the lazy brown ***dogs***.
Preposition	A word governing, and usually preceding, a noun or pronoun and expressing a relation to another word or element in the clause	The quick red fox jumped ***over*** the lazy brown dogs.
Verb	A word used to describe an action, state, or occurence, and forming the main part of the predicate of a sentence, such as ***hear***, ***become***, and ***happen***	The quick red fox ***jumped*** over the lazy brown dogs.

* all definitions are taken from the New Oxford American Dictionary, 2nd Edition

Lexical Categories: Subcategories

Nouns:

- **common nouns represent classes of entities:**
  - *town*, *ocean*, *person*
- **proper nouns represent unique entities:**
  - *London*, *John*, *Eiffel Tower*
- **pronouns are nouns representing other entities (usually mentioned previously):**
  - *he*, *she*, *it*

Morphology

- Morphology is a study of the internal structure of words
- Words consist of:
  - lexeme (root form)
  - affixes (suffix, prefix)
- Morphology has two categories:
  - inflectional - does not create new lexemes (happier)
  - derivational - creates new lexemes (unhappy)
- Inlectional morphemes carry grammatical meaning (plural -s), but they do not change the meaning of the word

Suffix	Example	Verb
-ation	nomination	nominate
-ee	appointee	appoint
-ure	closure	close
-al	refusal	refuse
-er	runner	run

Suffix	Example	Adjective
-dom	freedom	free
-hood	likelihood	likely
-ist	realist	real
-th	warmth	warm
-ness	happiness	happy

Suffix	Example	Marked form
N/A	look	base form
-ing	looking	gerund form
-s	looks	third person singular
-ed	loooked	past tense form
-en	taken	past participle

Phrases

- Phrases consist of multiple words
- Phrases are rooted by at least one word of a particular type, but can also consist of words and phrases of other types
- Phrases can be combined to form clauses that are the minimal units to construct a sentence

Phrases and Clauses

Every sentence is constructed from phrases and/or clauses.

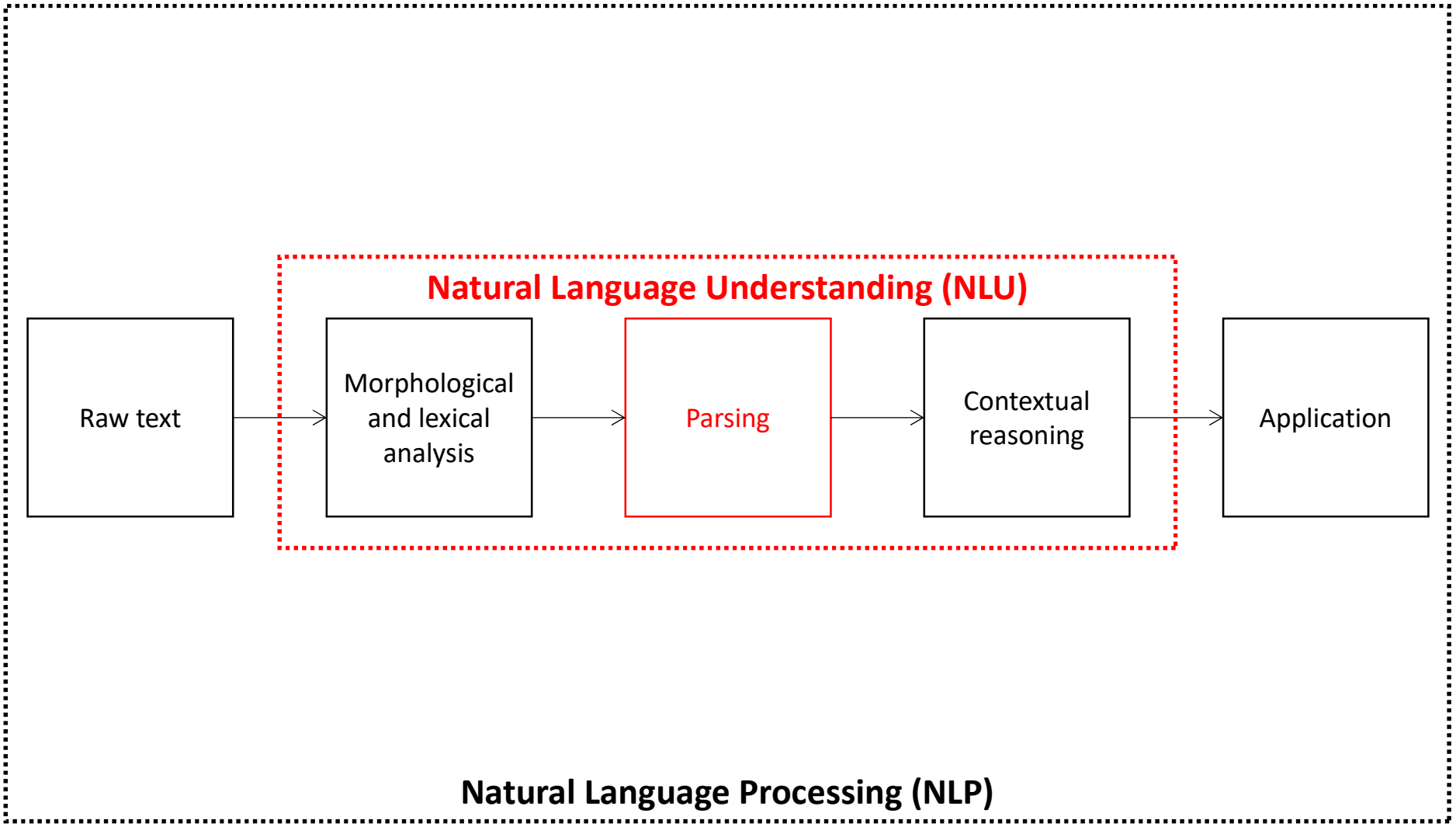
- **A phrase is a group of words, but it doesn't contain a subject and a verb.**
  - **Example:** *The big clock*
- **A clause is a group of words that contains a subject and a verb.**
  - **Example:** *The big clock chimed*

Phrase → Clause → Sentence

Common Phrasal Categories

Type	Example	Comments
Adjective	The *unusually red* fox jumped over the *exceptionally lazy* dogs.	The adverbs *unusually* and *exceptionally* modify the adjectives *red* and *lazy*, respectively, to create adjectival phrases.
Adverb	The dogs *almost always* ran down the field after the fox.	The adverb *almost* modifies the adverb *always* to create adverbial phrase.
Conjunction	The quick red fox *as well as* the silver coyote jumped over the lazy brown dogs.	Though this is somewhat of an exceptional case, you can see that the phrase *as well as* performs the same function as a conjunction such as *and*.
Noun	*The quick red fox jumped* over *the lazy brown dogs*.	The noun *fox* and its modifiers *the*, *quick*, and red create a noun phrase, as does the noun *dogs* and its modifiers *the*, *lazy*, and *brown*.
Preposition	The quick red fox jumped *over the lazy brown dogs*.	The preposition *over* and the noun phrase *the lazy brown dogs* form a prepositional phrase that modifies the verb *jumped*.
Verb	The quick red fox *jumped over the lazy brown dogs*.	The verb *jumped* and its modifier the prepositional phrase *over the lazy brown dogs* form a verb phrase.

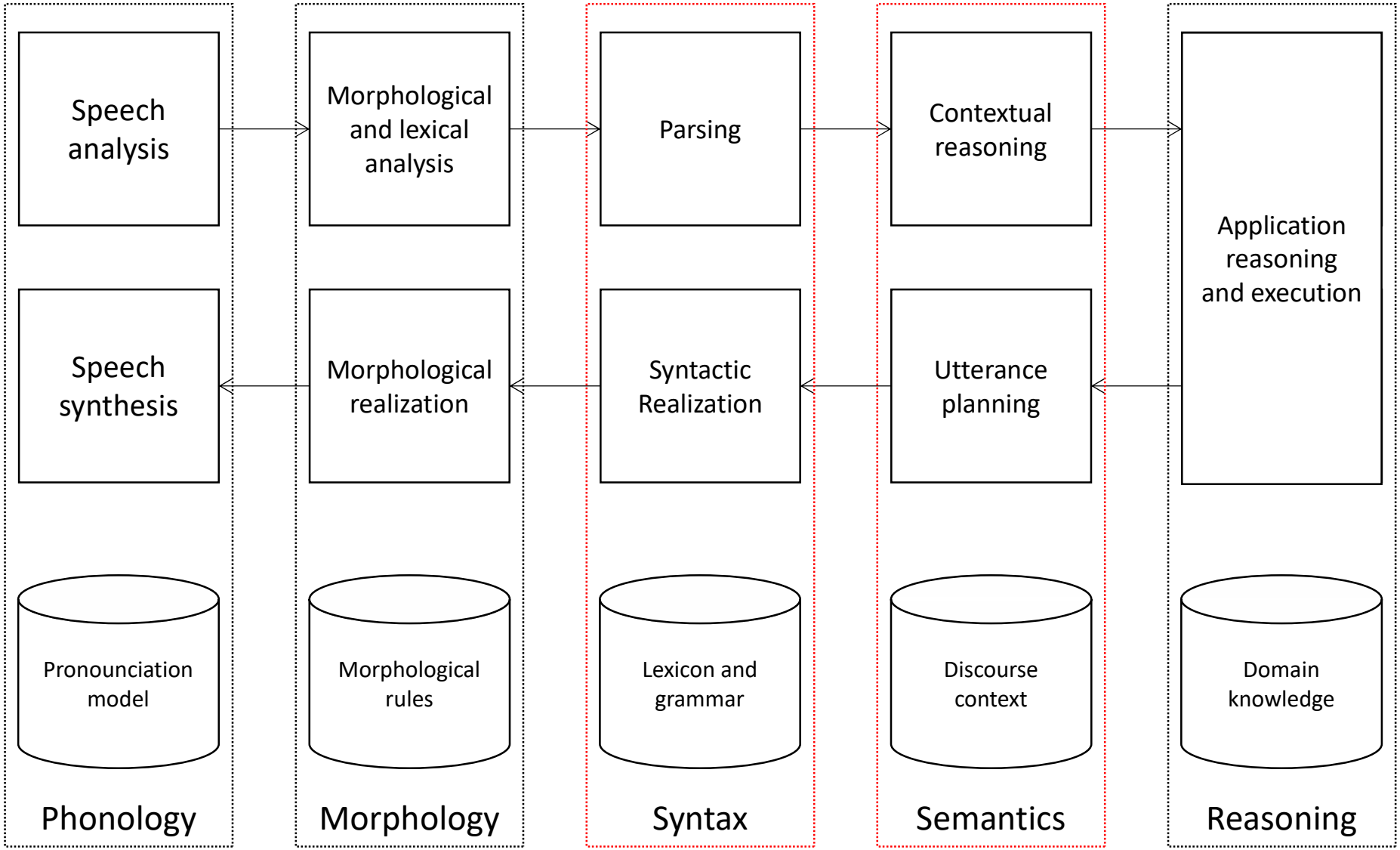
Basic NLP Text Processing Pipeline



Parsing

The task of determining the parts of speech, phrases, clauses, and their relationship to one another is called parsing.

Basic NLP Spoken Language Pipeline



Knowledge Levels / Forms for NLP

Level	Description
Phonetic and phonological knowledge	Concerned with how the words are related to sounds that realize them. Such knowledge is crucial for speech-based systems.
Morphological knowledge	Concerned with how words are constructed from the basic meaning units called **morphemes**.
Syntactic knowledge	Concerned with how words can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of what other phrases.
Semantic knowledge	Concerned with what the words mean and how these meanings combine in sentences to form sentence meanings. This is the study of context-independent meaning - the meaning a sentence has regardless of the context in which it is used.
Pragmatic knowledge	Concerned with how sentences are used in different situations and how use affects the interpretation of the sentence.
Discourse knowledge	Concerned with how the immediately preceding sentences affect the interpretation of the next sentence. This information is especially important for interpreting pronouns and for interpreting the temporal aspects of the information.
World knowledge	Includes the general knowledge about the structure of the world that language users must have in order to, for example, maintain a conversation. It includes what each language user must know about the other user's beliefs and goals.

(English) Syntax

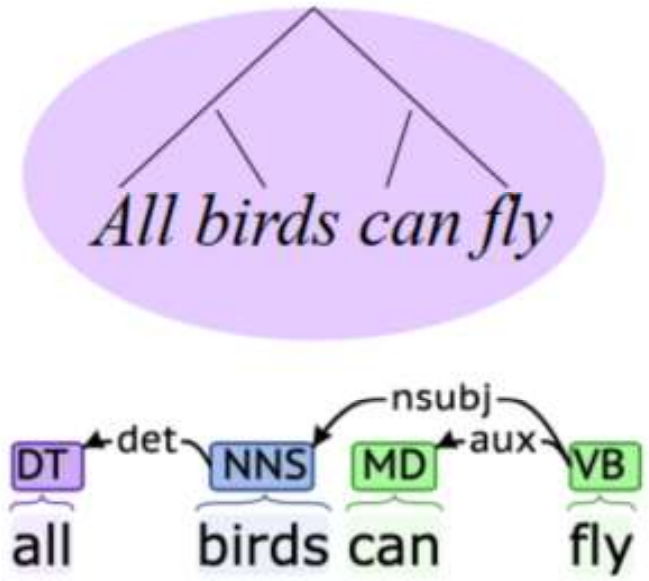
The structure of words and phrases within a sentence:

- **Different formalisms, coming from the American (phrase structure) and European (dependency grammar) structuralist traditions**

Applications:

- **Part-of-speech tagging**
- **Entity extraction**
- **Syntactic parsing (Context-Free Grammar)**
- **Syntactic parsing (dependencies)**

Examples:



Semantics

The representation of meaning in language:

- **at different levels: lexical, sentential, textual**
- **logical formalisms: reference and truth conditions**

Applications:

- **Word embedding / encoding**
- **Lexical resources**
- **Semantic role labeling**

Example:

$$\forall x\ [\text{bird}(x) \Rightarrow \text{fly}(x)]$$

Pragmatics

How language is used to achieve specific intentions:

- **conversational implicatures: how I interpret what you say because of what I assume you are trying to do**
- **speech acts**

Applications:

- **Speech act labeling**
- **Discourse structure parsing**
- **Dialogue systems**

Examples:

"I ate most of your cookies"

□

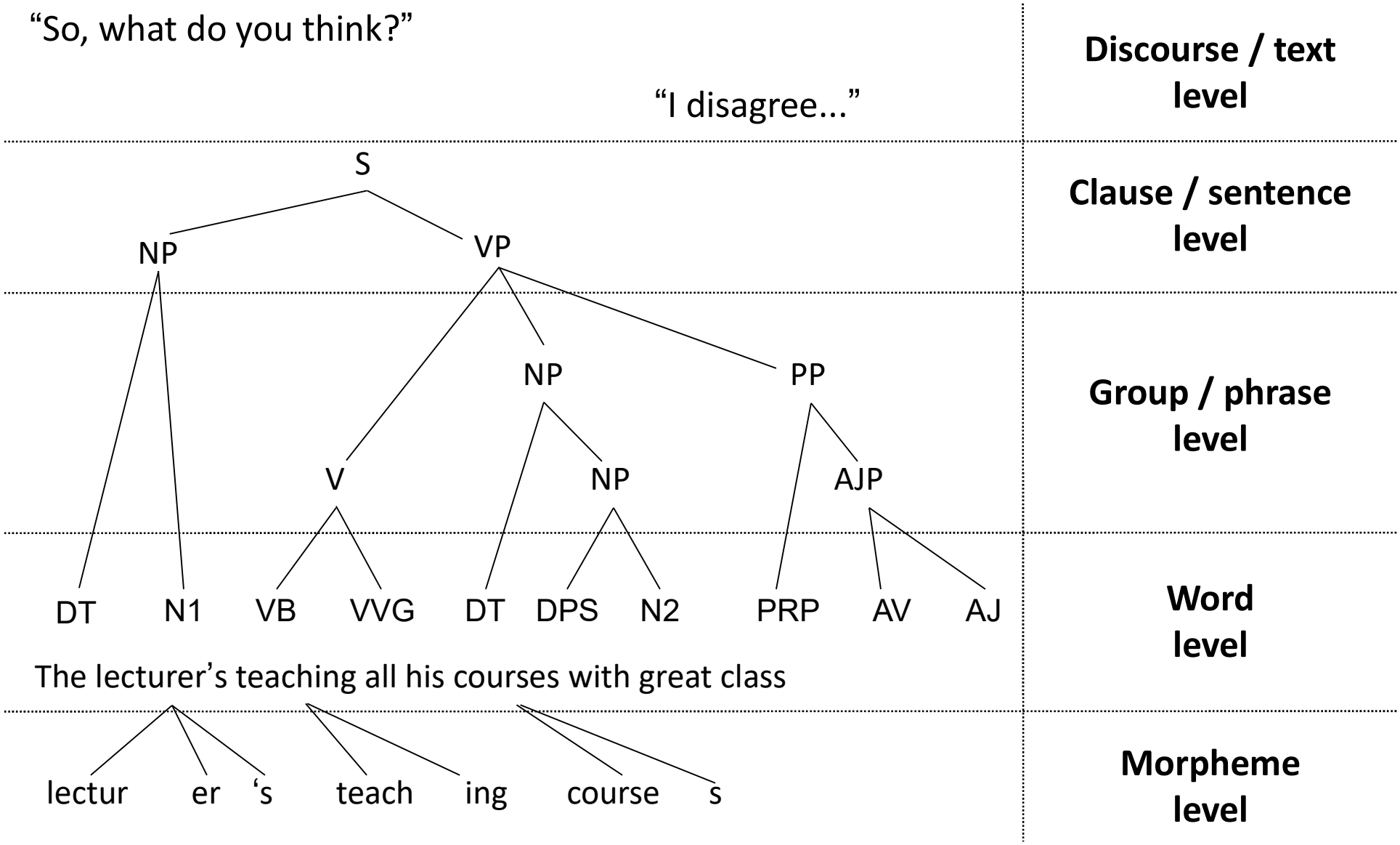
I did not eat all of your cookies

"Where does your brother live?"

□

I do not know where your brother lives

Structure / Rank Levels for NLP



Syntax - Semantics - Pragmatics

- Syntax: what is its "formal" relation structure?
- Semantics: what does it "mean"?
- Pragmatics: how is it "used"?

Consider a hypothetical "first" CS 481 sentence:

Sentence	Syntax	Semantics	Pragmatics
Language is one of the fundamental aspects of human behavior and is a crucial component of our lives.			
Green frogs have large noses.			
Green ideas have large noses.			
Large have green ideas nose.			

Knowledge - Rank Mapping: NLP Tasks

Rank / Domain	Syntax	Semantics	Pragmatics
Word	Parts-of-speech Morphology	Word senses Word similarity	Sentiment analysis
Group / Phrase	Shallow parsing	Named entity recognition Semantic role labeling	Deixis Coreference
Clause / Sentence	Parsing	Information extraction Entailment	Speech act interpretation Sentiment analysis
Discourse / Text	Rhetorical discourse structure	Text categorization Story understanding	Coherence Sentiment analysis

Representations and Understanding

Text understanding involves computing a representation of the meaning of sentences and texts.

- **The sentence itself is not enough to represent the meaning:**
  - **word *cook* has a verb and a noun sense,**
  - **word *catch* can mean a baseball move, a fish, etc.**

Representation Language Properties

Useful representation language have two properties:

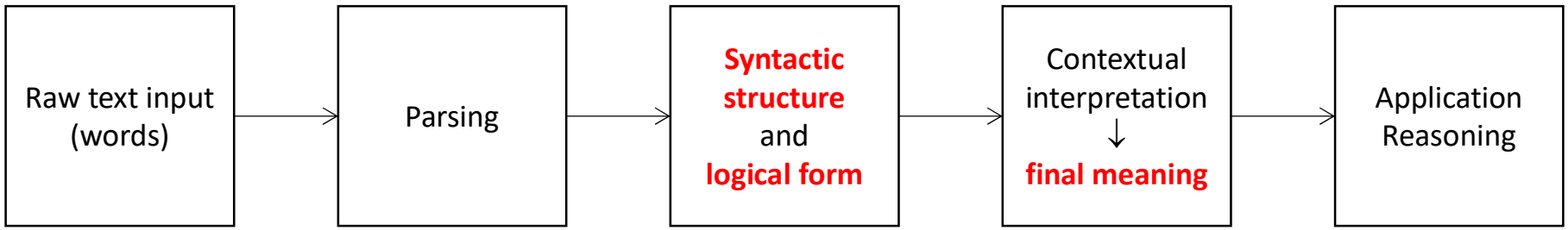
- **The representation must be precise and unambiguous. You should be able to express every distinct reading of a sentence as a distinct formula in the representation.**
- **The representation should capture the intuitive structure of the natural language sentences that it represents. For example:**
  - **sentences that are structurally similar should have similar structural representations, and**
  - **the meanings of two sentences that are paraphreses of each other should be closely related to each other.**

Syntax - Logical Form - Final Meaning

Representation can be realized using:

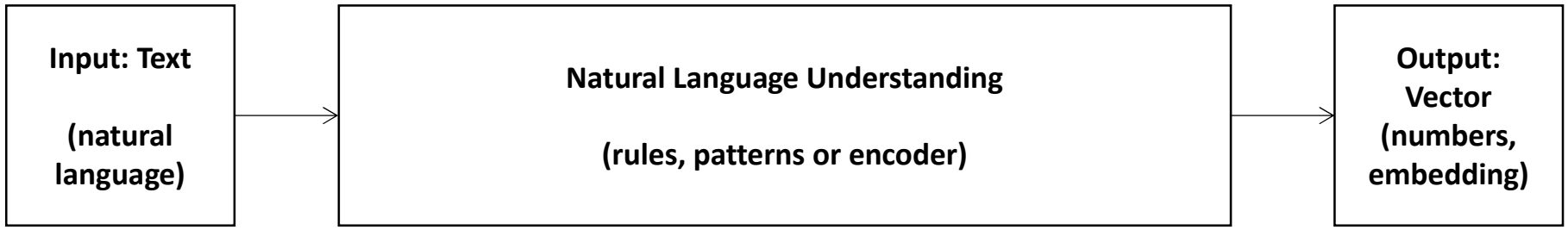
- **sentence syntactic structure: it indicates the way that individual words in a sentence**
  - **are related to each other**
  - **grouped together into phrases**
  - **modify other words**
  - **are of central importance**
- **the logical form: context-independent meaning of a sentence**
- **the final meaning: general knowledge representation meaning | context-dependent meaning**

NLU: Flow of Information



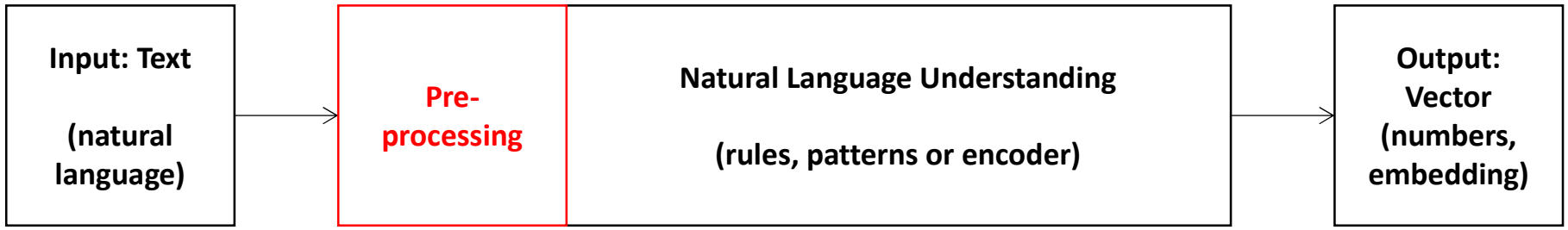
Automated Text Processing

The task of automatic processing of text is to extract a numerical representation of the meaning of that text. This is the natural language understanding (NLU) part of NLP. The numerical representation of the meaning of natural language usually takes the form of a vector called an embedding.



Automated Text Processing

The task of automatic processing of text is to extract a numerical representation of the meaning of that text. This is the natural language understanding (NLU) part of NLP. The numerical representation of the meaning of natural language usually takes the form of a vector called an embedding.



Exercise A: NLP

https://www.ibm.com/demos/live/natural-language-understanding/self-service/home

Exercise B: Sentiment Analysis

https://monkeylearn.com/sentiment-analysis-online/

https://text2data.com/Demo

Exercise C: Speech-to-text

https://cloud.google.com/speech-to-text

Exercise D: Text-to-speech

https://cloud.google.com/text-to-speech

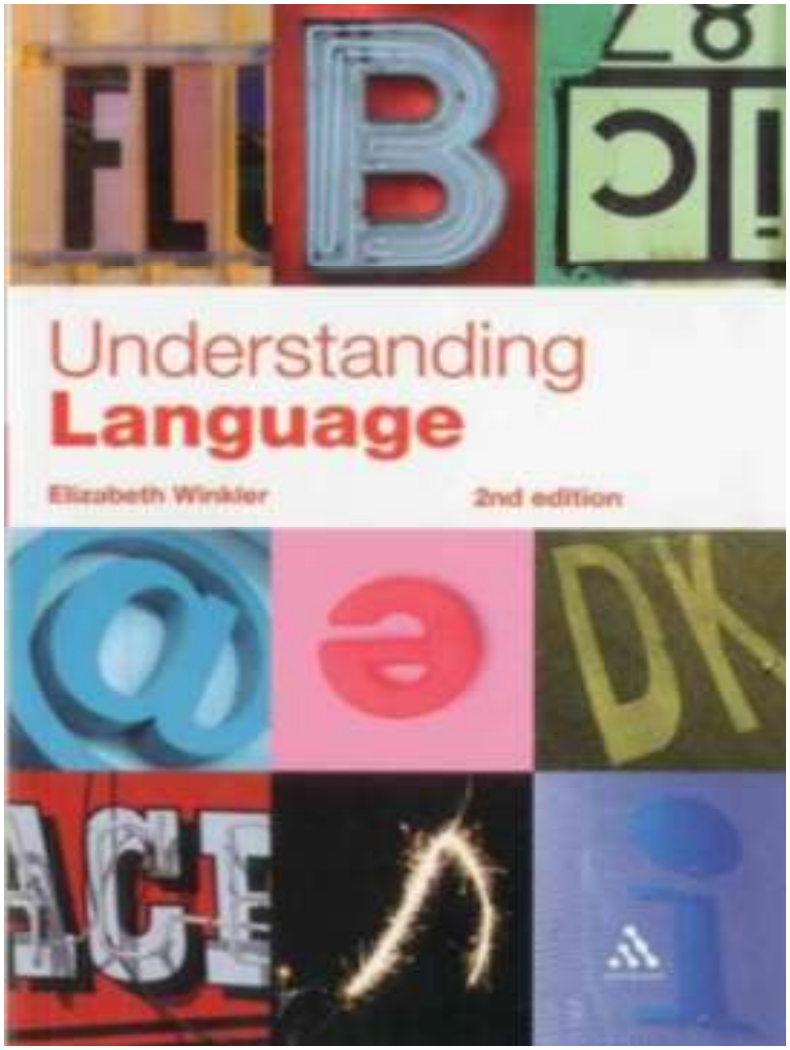
https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

https://www.ibm.com/demos/live/tts-demo/self-service/home

Exercise E: GPT-2

https://transformer.huggingface.co/doc/distil-gpt2

Optional Reading



Title:

Understanding Language (2nd edition)

Authors:

Elizabeth Winkler

ISBN:

9781441138965

Publisher:

Continuum Books

Published:

March 29, 2012