Natural Language Processing

Dr. Muneendra Ojha

Assistant Professor

Department of Information Technology

Indian Institute of Information Technology - Allahabad

Natural Language Processing

The Bird's Eye View

01

Introduction to NLP

Machine vs Natural Language

- Machine languages are ***designed***
  - rules came first
  - people started using language once the rule set was complete
- Natural languages have ***evolved***
  - usage comes first
  - rules (grammar) formalized later

Machine vs Natural Language

- Machine languages are ***designed***
  - Highly structured
  - Rigorous
  - Precise syntax
  - Rules to create exactly defined concepts
  - Fixed vocabulary

Machine vs Natural Language

- Natural languages have ***evolved***
  - Messy
  - Ambiguous
  - Chaotic
  - Sprawling
  - Constantly in flux

Natural Languages

- Language or text underpins most of our communications.
- Our language presents our cultural production.
- Internet is mostly text.
- Language is how we store almost all of our knowledge.
- Our very thoughts are largely built upon language.

02

Challenges of NLP

Human languages are ambiguous

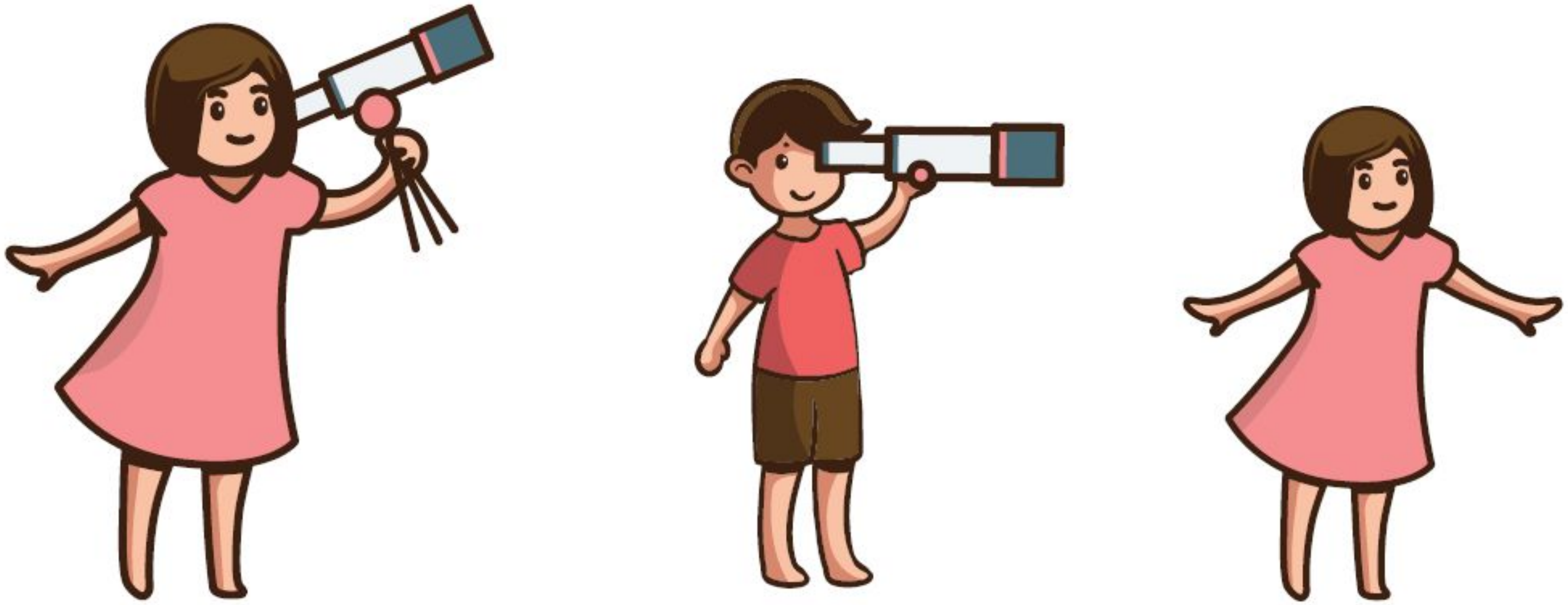
- The interpretation of natural language sentences are often not unique.

- Both structures (how sentences are formed) and semantics (what sentences mean) can have ambiguities in human language.

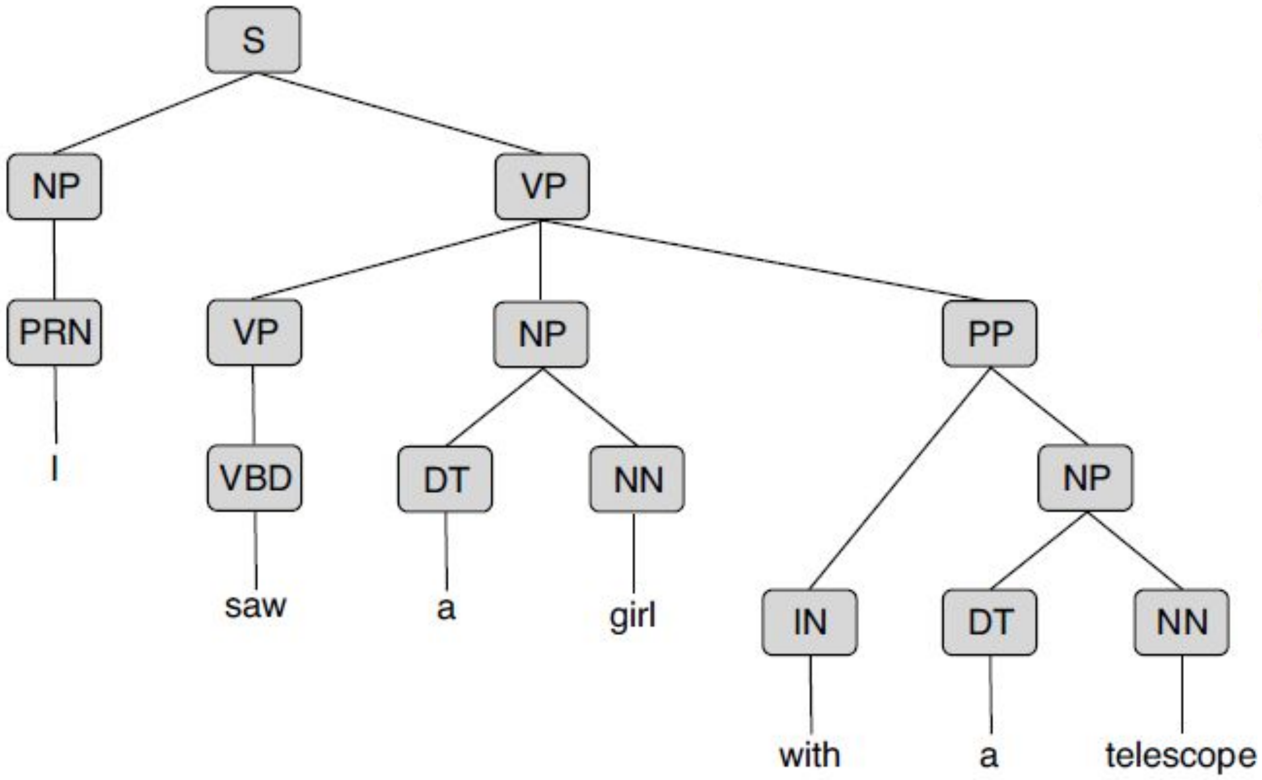
- Further, the context in which a sentence has been said also defines the meaning that is to be interpreted.

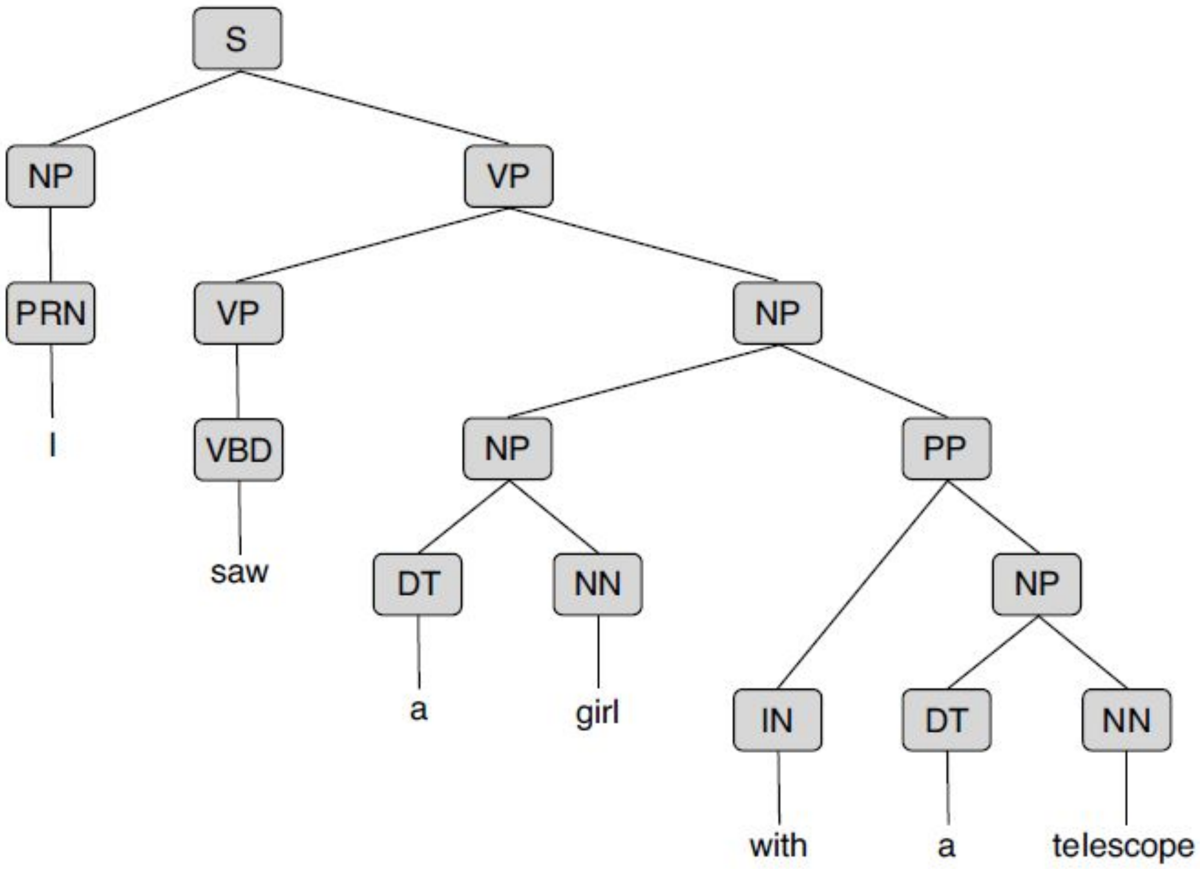
Human languages are ambiguous

- *Syntactic ambiguity*
  - I saw a girl with a telescope



NLP Task: PARSING





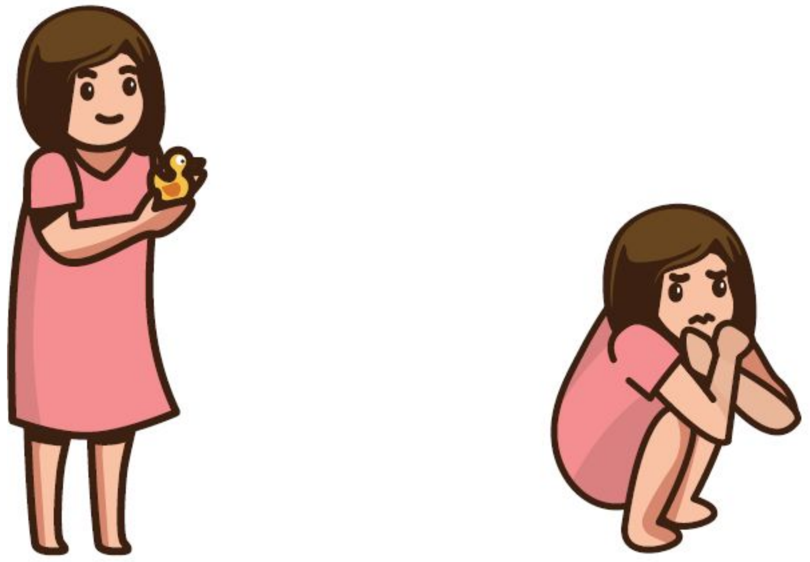
I saw a girl with a telescope

Human languages are ambiguous

- Lexical ambiguity:
  - a word that could be used as a verb, noun, or adjective.
- Semantic ambiguity
  - I saw a bat

Human languages are ambiguous

- *Contextual ambiguity*
  - machines don't understand the context outside of the sentence that is being translated
  - "I saw her duck."



Meaning: What and When?

- “A very good product!”: ***Positive***
- “I can’t think of a single good reason to use this product”: ***Positive?***
- “The product is not only cheap but also very good!”: ***Negative?***
- “Very bad. It crashes all the time!”: ***Negative***
- “It’s not bad.”: ***Negative?***
- “It’s very badly made.”: ***Negative***
- “I always wanted this feature badly!”: ***Negative?***

Single word in a language have two completely opposite meanings?

A long list of challenges…

- Contextual words and phrases and homonyms
  - *I **ran** to the store because we **ran** out of milk.*
- Synonyms
- Irony and sarcasm
- Ambiguity
- Errors in text or speech
- Colloquialisms and slang
- Domain-specific language
- Low-resource languages
- Lack of research and development


03
History of NLP

Making sense of Natural Language

- Ability to understand natural language has long eluded machines.
- Early attempts to build NLP systems were made through “Applied Linguistics.”
- Engineers and linguists would handcraft complex sets of rules to perform basic machine translation or create simple chatbots
- The famous ELIZA program from 1960s used pattern matching to sustain very basic conversation.

Rebellious language

- Handcrafted rules remained dominant approach till 1990s.
- But language is a rebellious thing and not easily pliable to formalization.
- After several decades of effort, the capabilities of these systems remained disappointing.
- Towards late 1980s, faster computers and greater data availability started making a better alternative viable.

A clever engineer

- When you find yourself building systems that are big piles of ad-hoc rules, as a clever engineer, you're likely to start asking:

"Could I use a corpus of data to automate the process of finding these rules, instead of having to come up with them myself??"

- And just like that, the state of the art of NLP graduated to doing machine learning.

Machine learning approaches

- Towards late 1980s, we started seeing machine learning approaches.
- Earliest ones were based on decision trees.
- Then statistical approaches started gaining speed.
- Over time, learned parametric models fully took over.
- Frederick Jelinek, an early speech recognition researcher, joked in the 1990s:

"Every time I fire a linguist, the performance of the speech recognizer goes up."

1990s to 2010s…

- The toolset of NLP—decision trees, logistic regression—only saw slow evolution from the 1990s to the early 2010s.
- Most of the research focus was on feature engineering.
- When Francois Chollet (the creator of Keras, and author of one of our reference books) won his first NLP competition on Kaggle in 2013, his model was based on decision trees and logistic regression.

Advent of Deep Learning

- Around 2014–2015, things started changing at last.
- As Deep Learning came into limelight, multiple researchers began to investigate the language-understanding capabilities of recurrent neural networks, in particular LSTM
- LSTM is a sequence-processing algorithm from the late 1990s that had stayed under the radar until then.

Keras and Bidirectional LSTM

- In early 2015, Keras made available the first open source, easy-to-use implementation of LSTM, just at the start of a massive wave of renewed interest in recurrent neural networks.

- Then from 2015 to 2017, recurrent neural networks dominated the booming NLP scene.

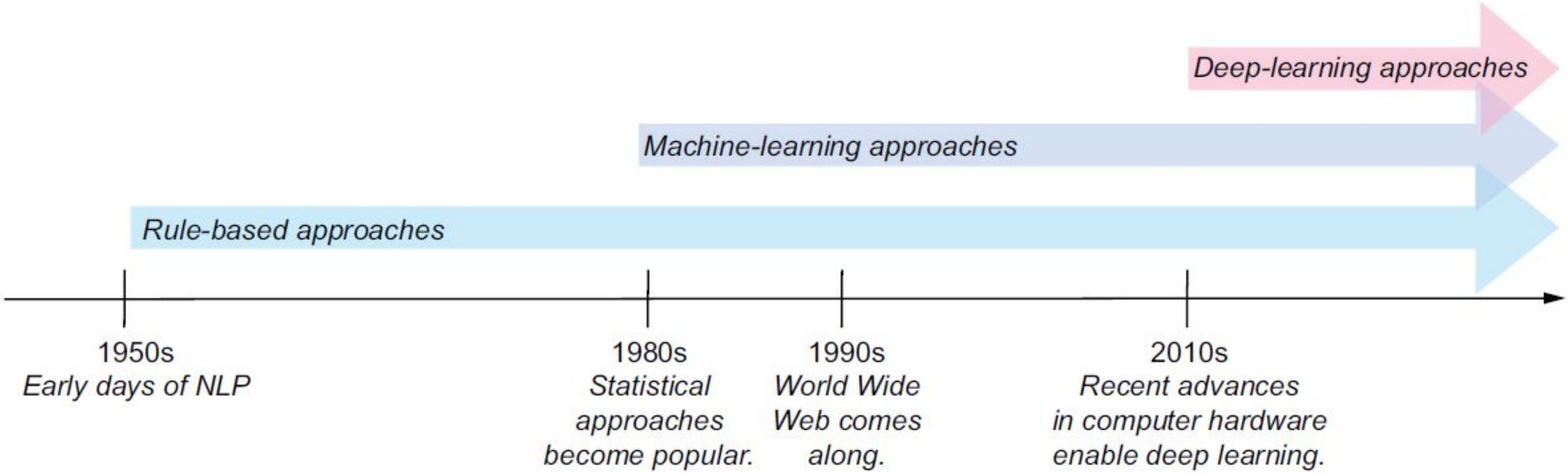
- Bidirectional LSTM models, in particular, set the state of the art on many important tasks, from summarization to question-answering to machine translation.

Finally we had transformers

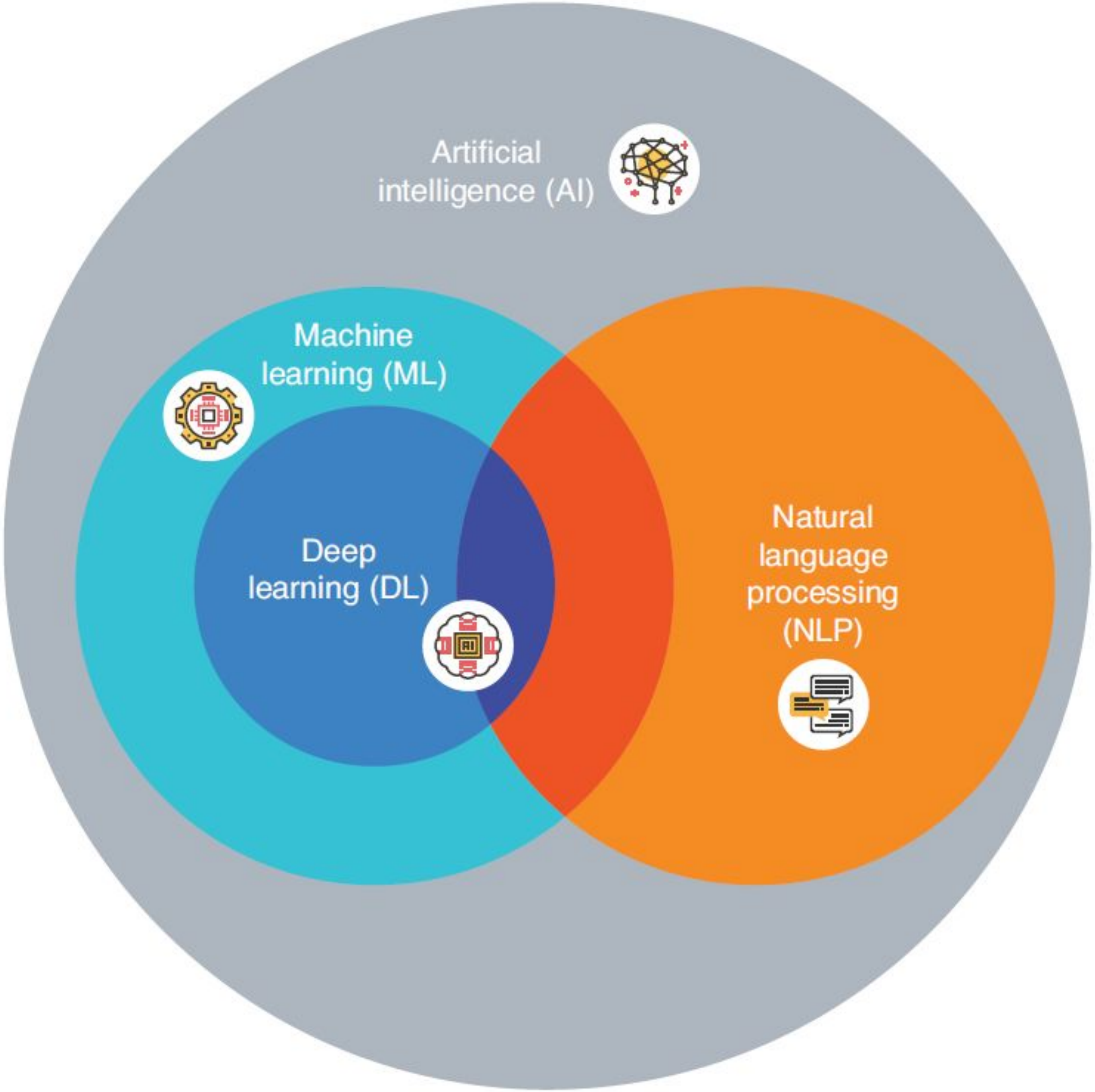
- Around 2017–2018, a new architecture rose to replace RNNs: the ***Transformer***,

- Transformers unlocked considerable progress across the field in a short period of time, and today most NLP systems are based on them.

three approaches for NLP



AI, ML and DL



Artificial intelligence (AI)
Machine learning (ML)
Deep learning (DL)
Natural language processing (NLP)

04

Modern NLP applications

Modern NLP is about…

- … using machine learning and large datasets to give computers the ability to ingest a piece of language as input and return something useful.

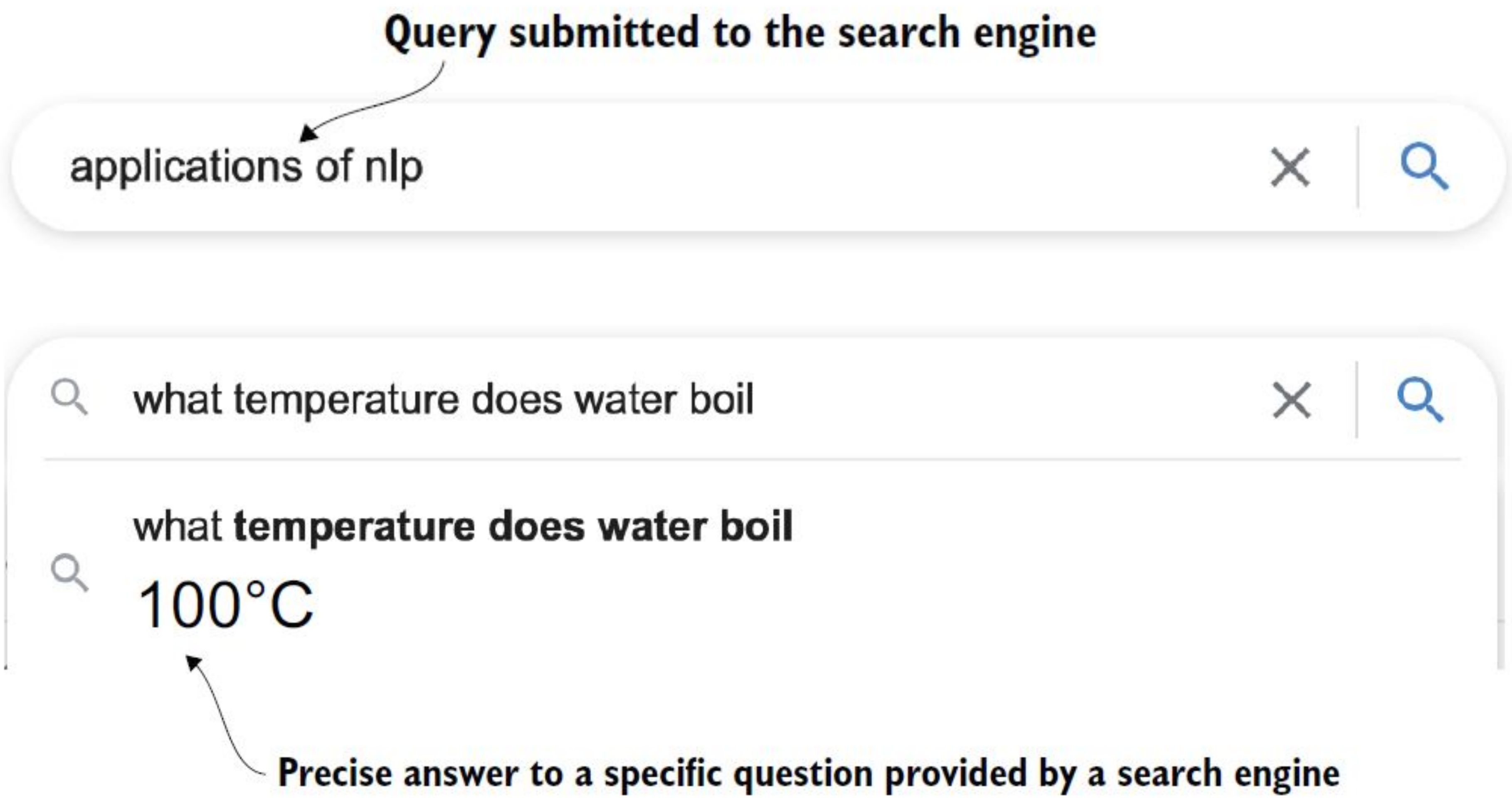
- Like predicting the following:
  - "What's the topic of this text?" (***text classification***)
  - "Does this text contain abuse?" (***content filtering***)
  - "Does this text sound positive or negative?" (***sentiment analysis***)
  - "What should be the next word in this incomplete sentence?" (***language modeling***)
  - "How would you say this in German?" (***translation***)
  - "How would you summarize this article in one paragraph?" (***summarization***) etc.

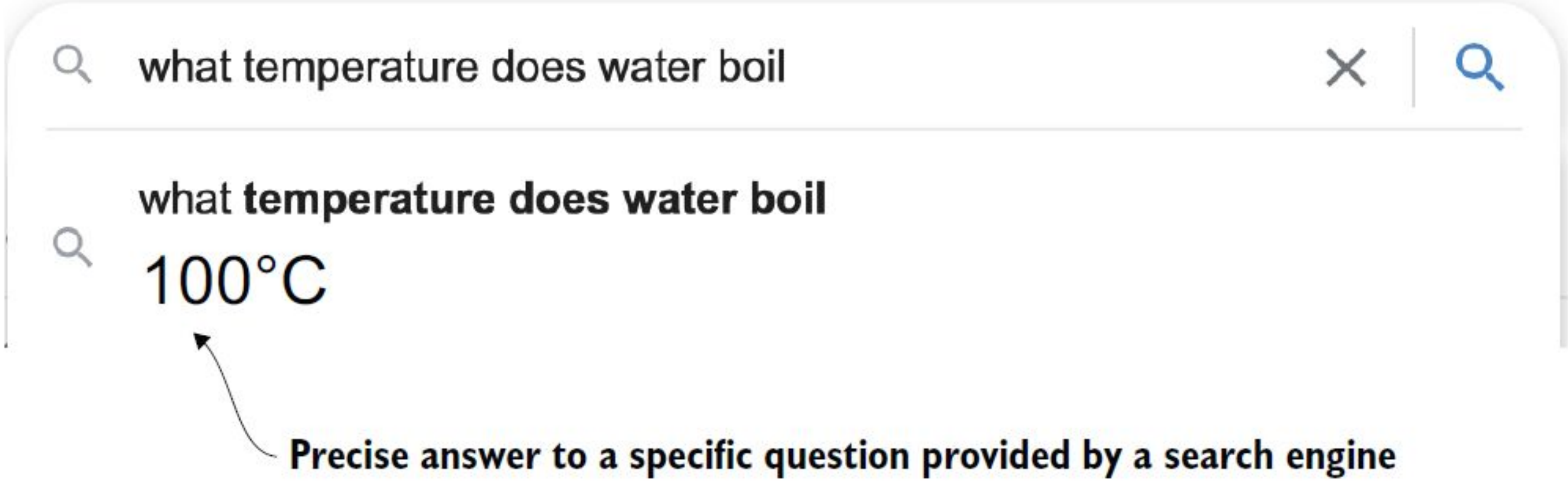
Typical tasks

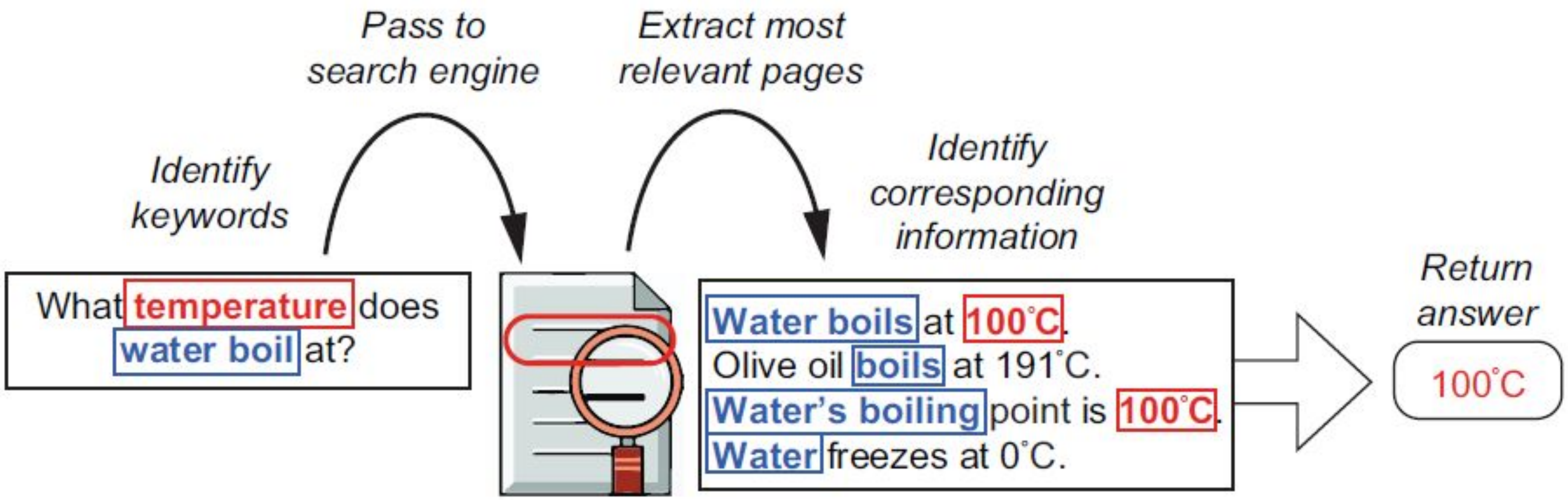
- ***Information search***

Query submitted to the search engine

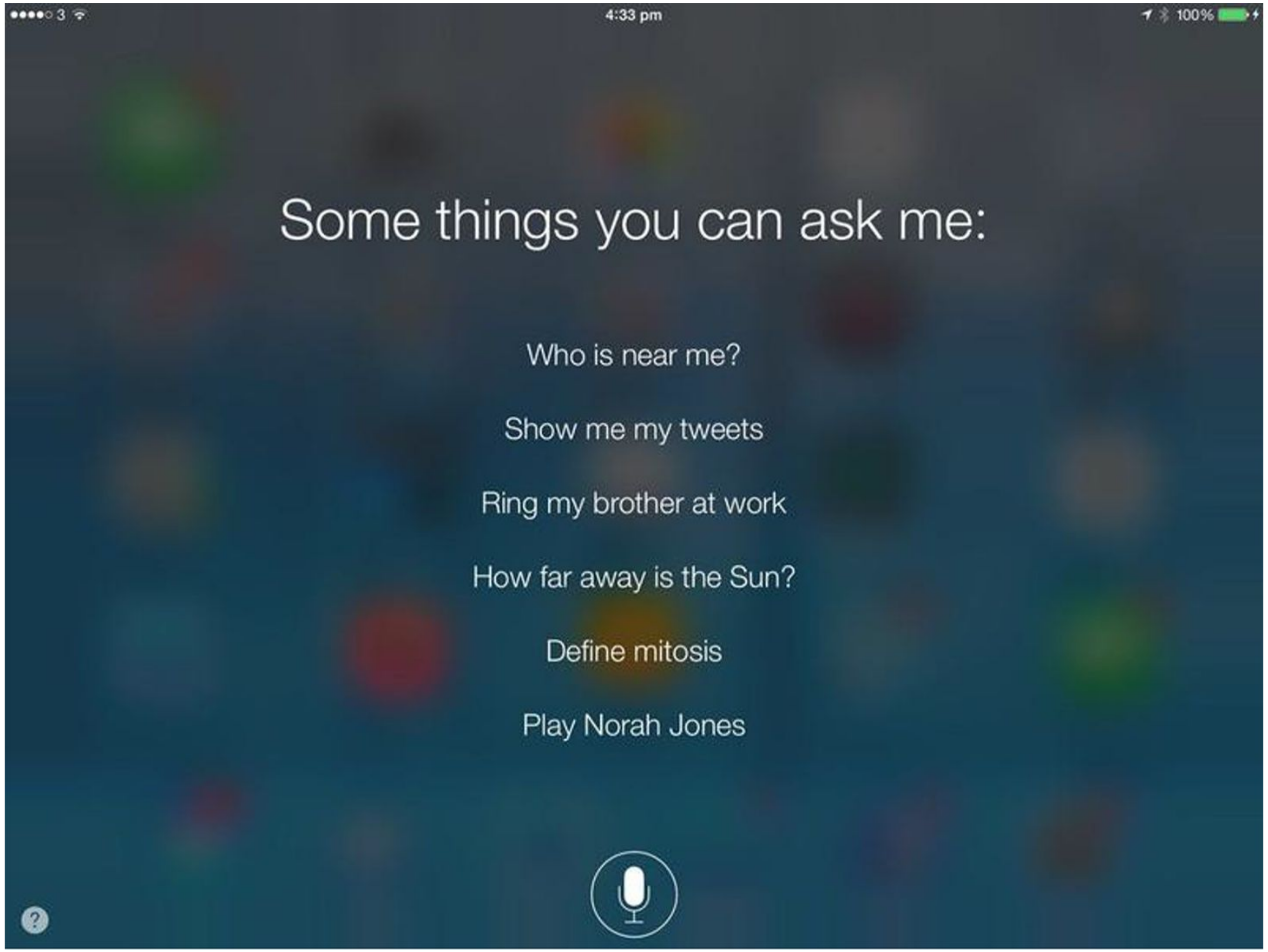
Information search



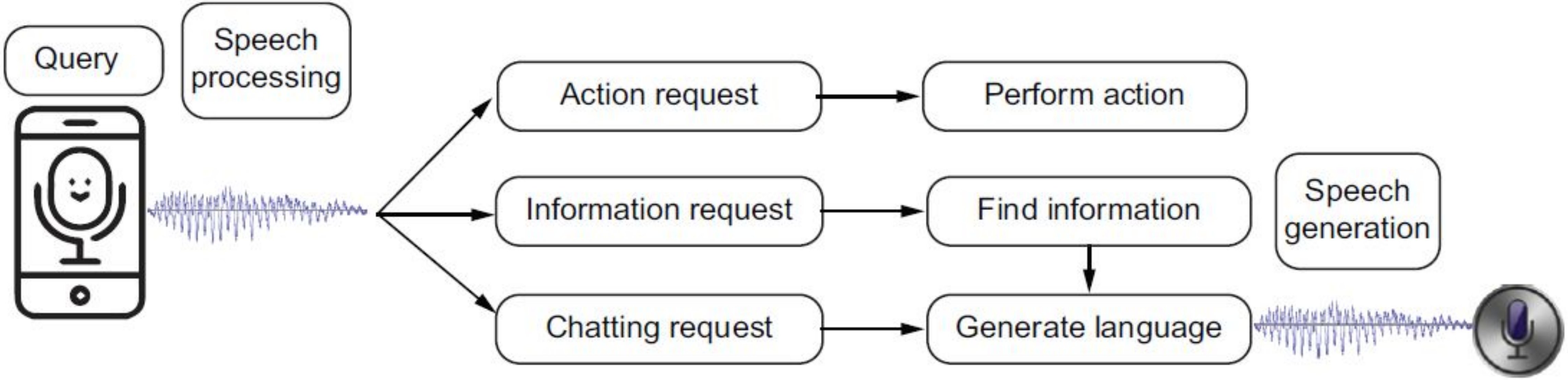

what temperature does water boil
what temperature does water boil
100°C
Precise answer to a specific question provided by a search engine


Pass to search engine
Extract most relevant pages
Identify keywords
Identify corresponding information
Return answer
What temperature does water boil at?
Water boils at 100˚C.
Olive oil boils at 191˚C.
Water's boiling point is 100˚C.
Water freezes at 0˚C.
100˚C

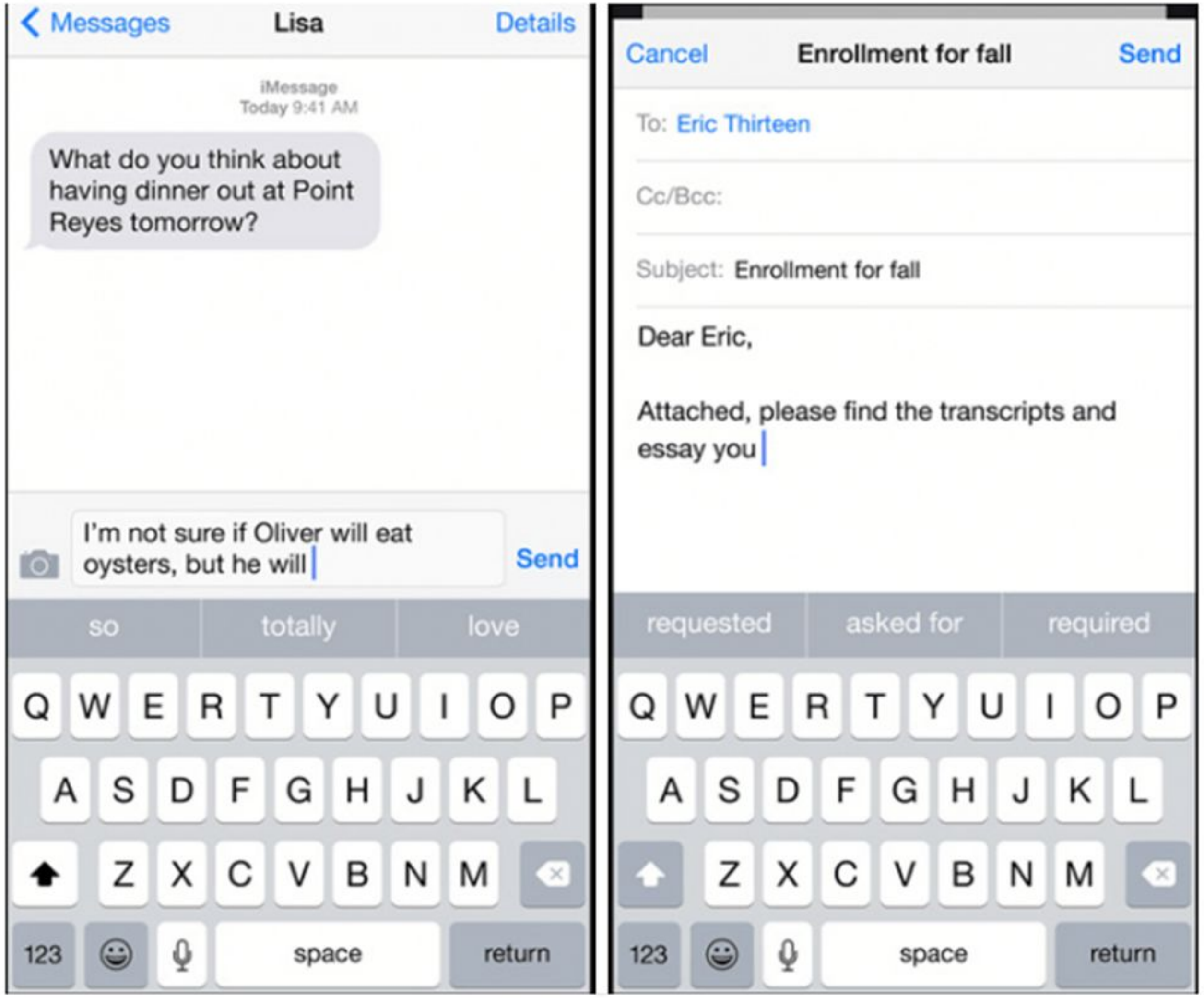
Conversational agents

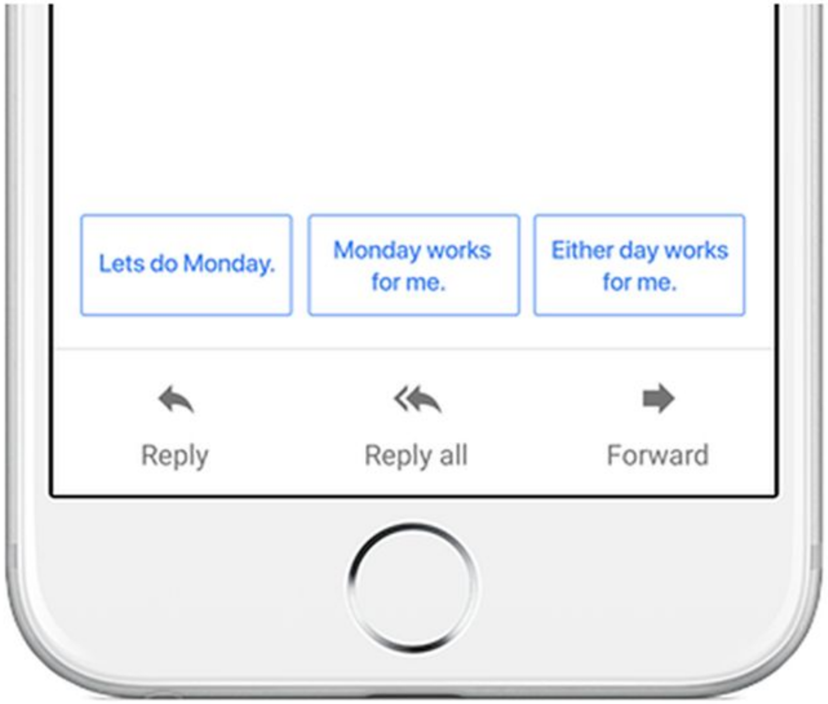


Pipeline of an intelligent virtual assistant

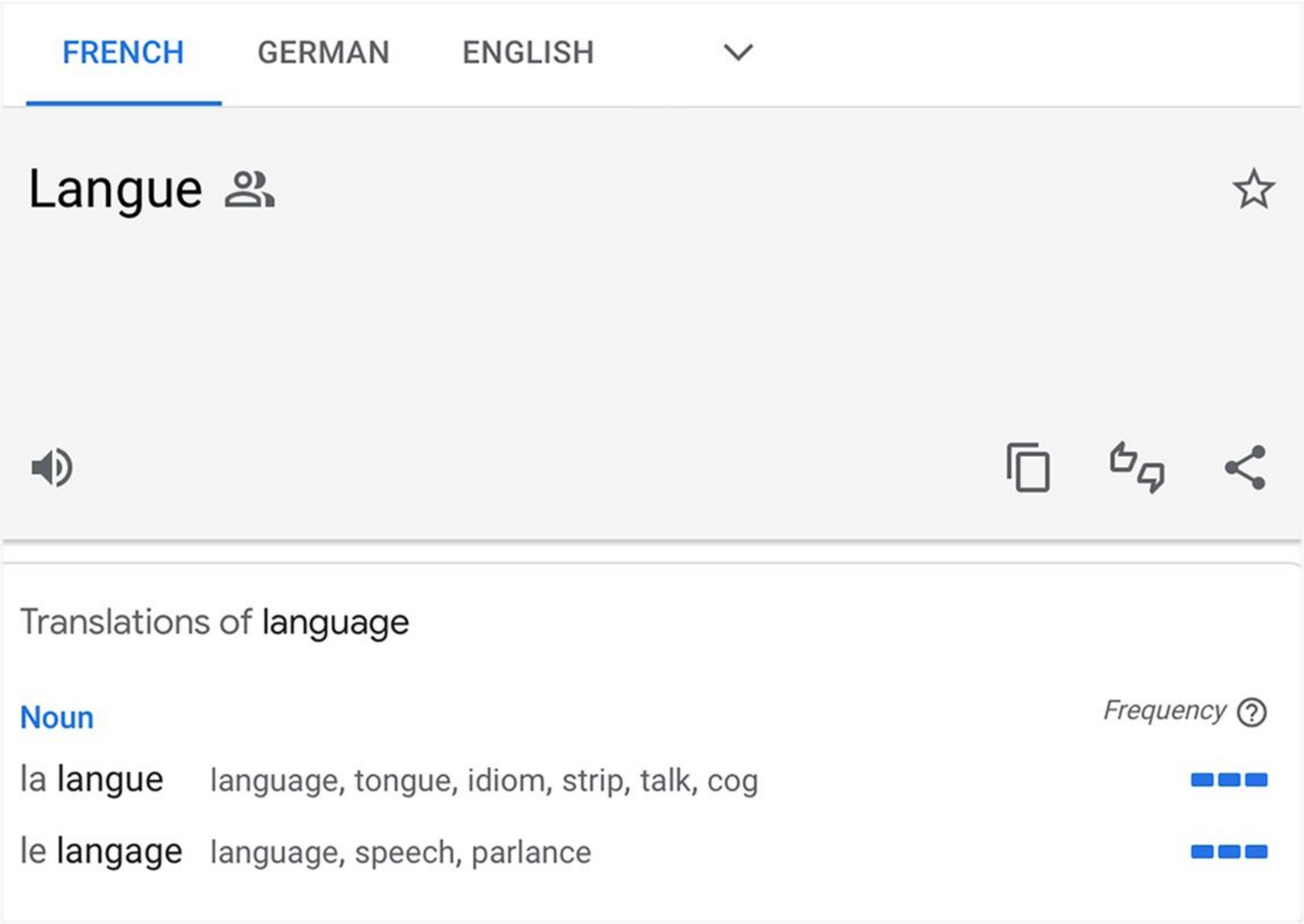


Text prediction and language generation



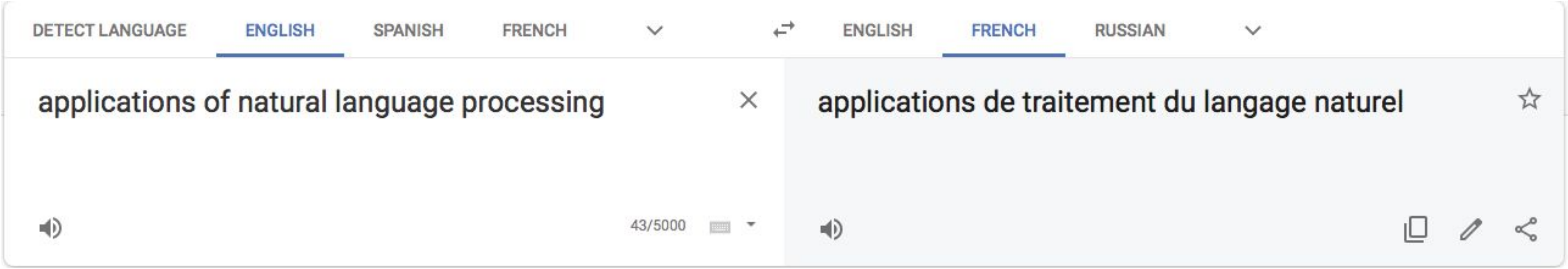


Machine translation



Several translations for the English word ***language*** to French (Google Translate)

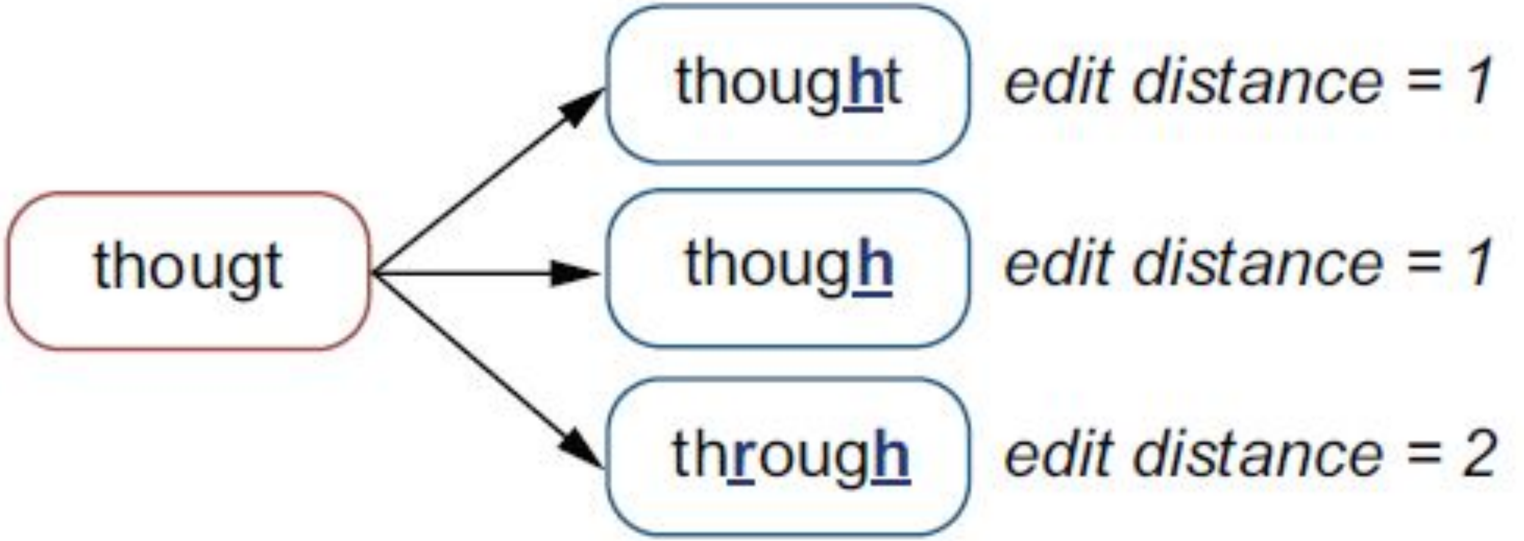
Machine translation



Phrase translation between English and French for

“applications of natural language processing”

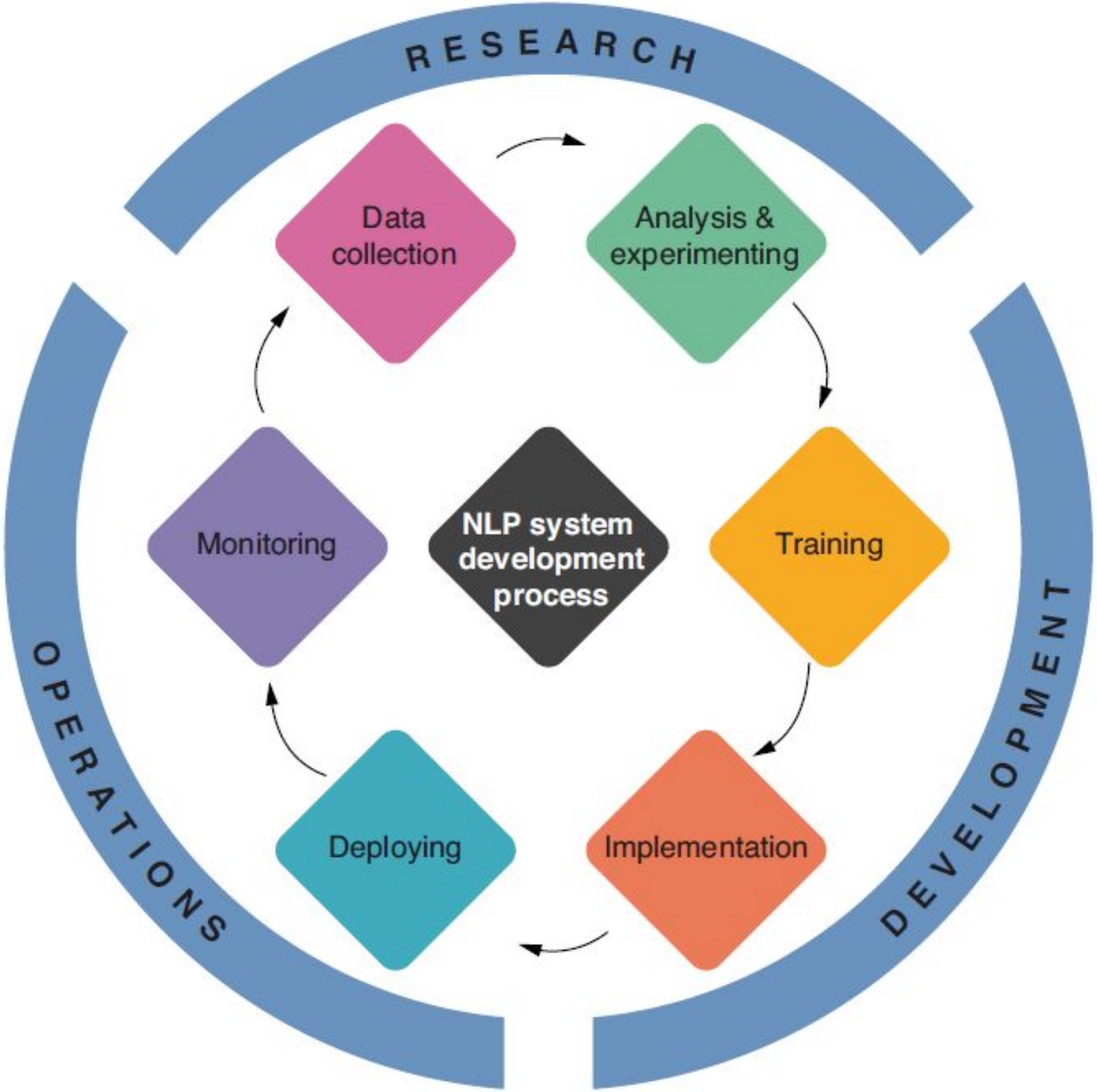
Spell- and grammar checking



1. *Rule-based with the use of a dictionary*
2. *Use of machine learning*
3. treat it as a machine-translation problem

05

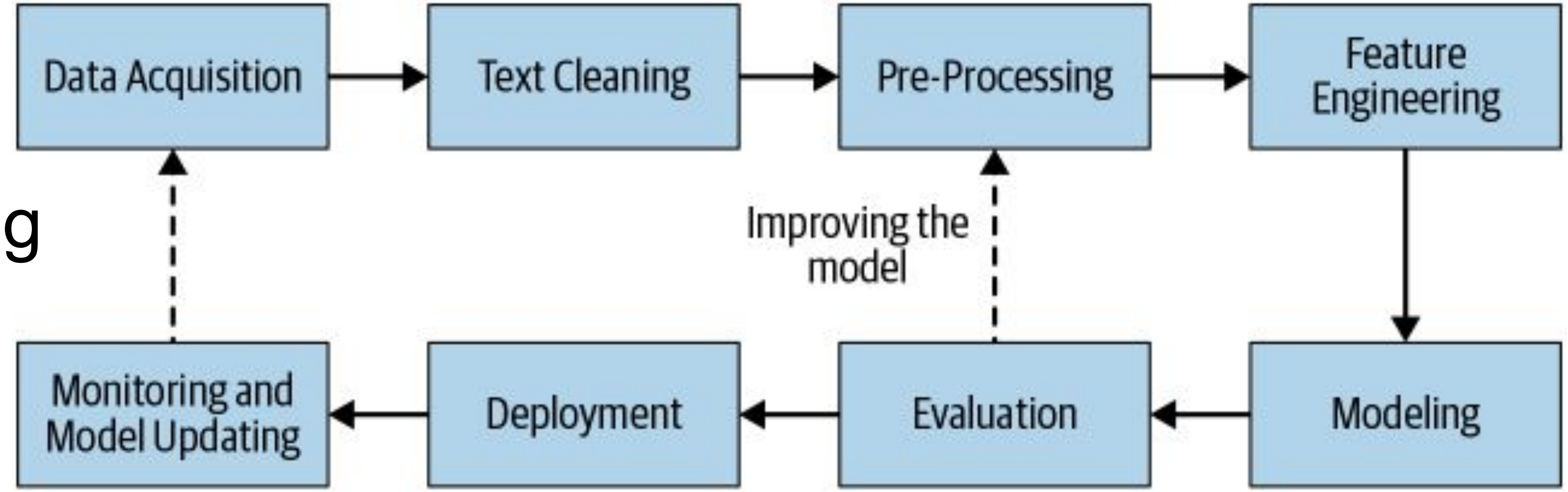
Building NLP Applications

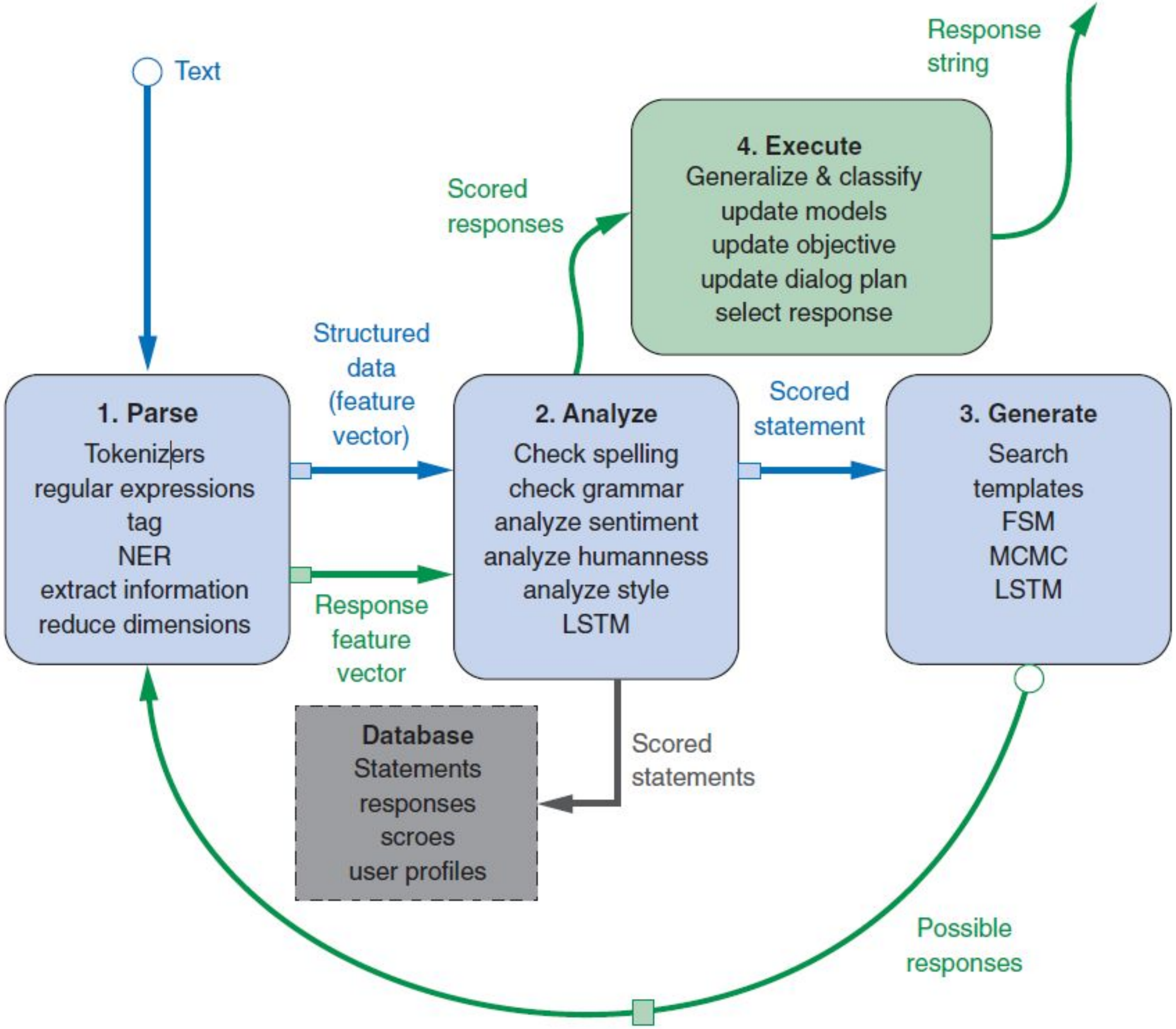


The development cycle of NLP applications

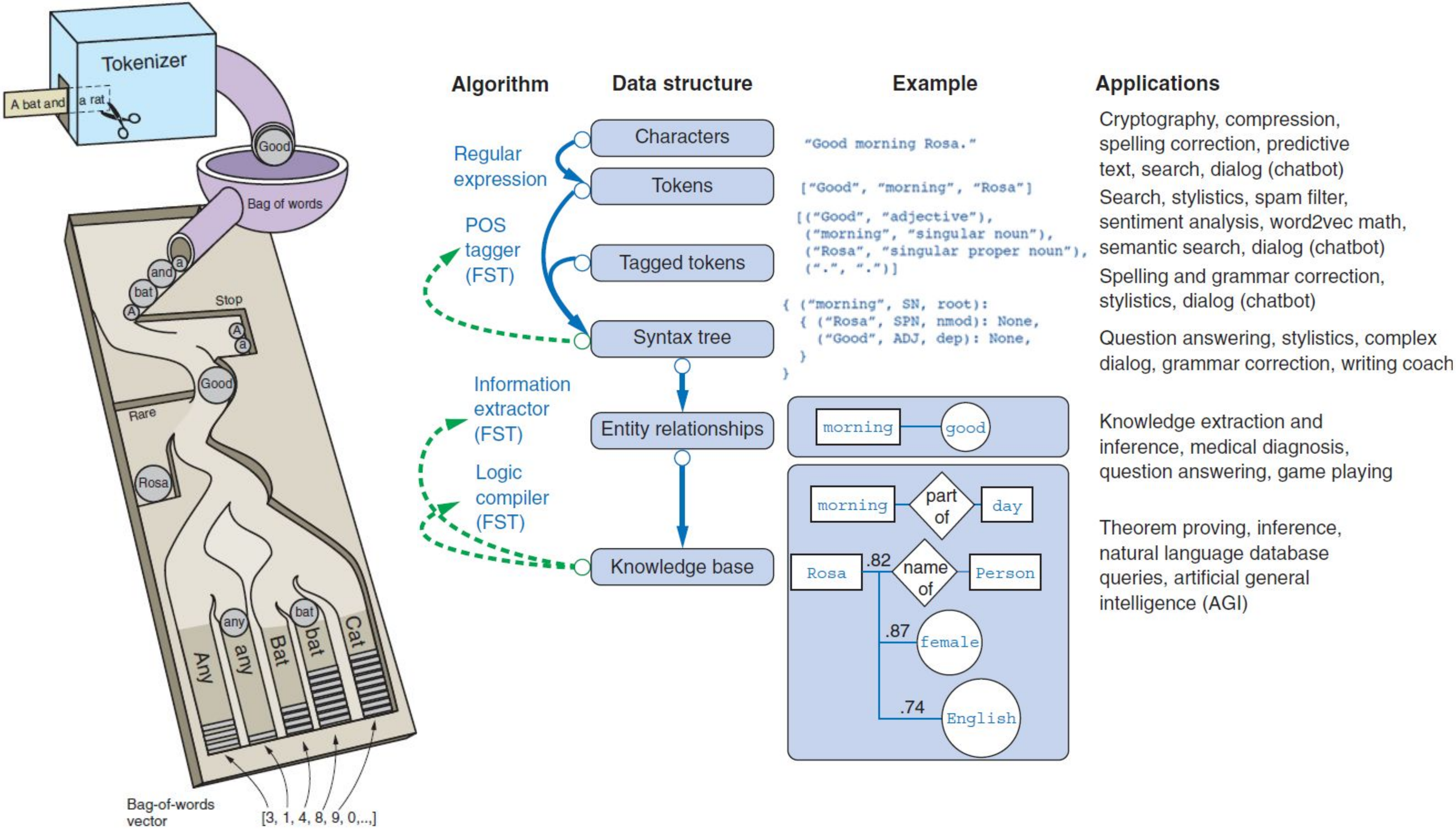
Data driven NLP system

1. Data acquisition
2. Text cleaning
3. Pre-processing
4. Feature engineering
5. Modeling
6. Evaluation
7. Deployment
8. Monitoring and model updating





A typical Chatbot recurrent pipeline

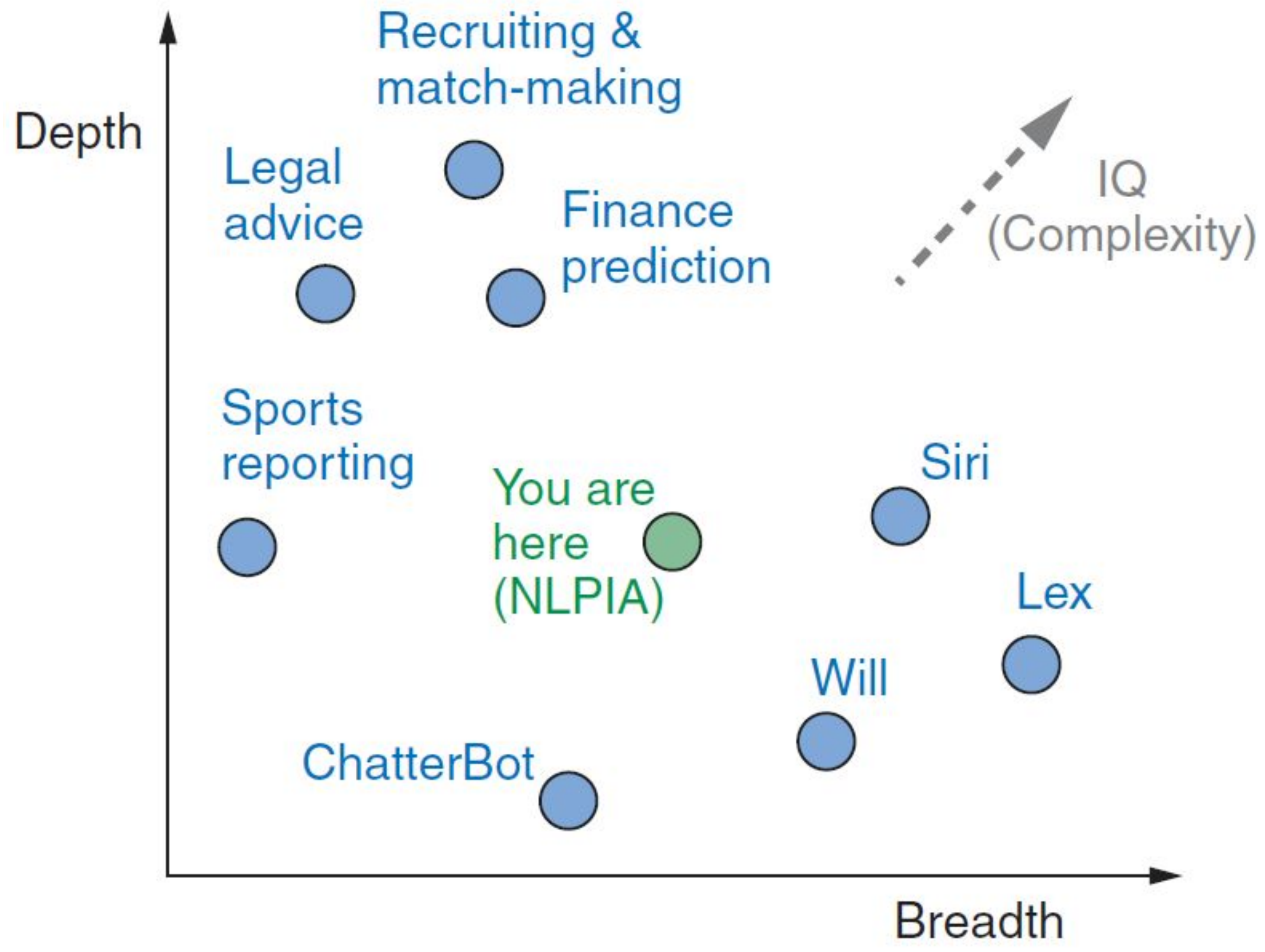


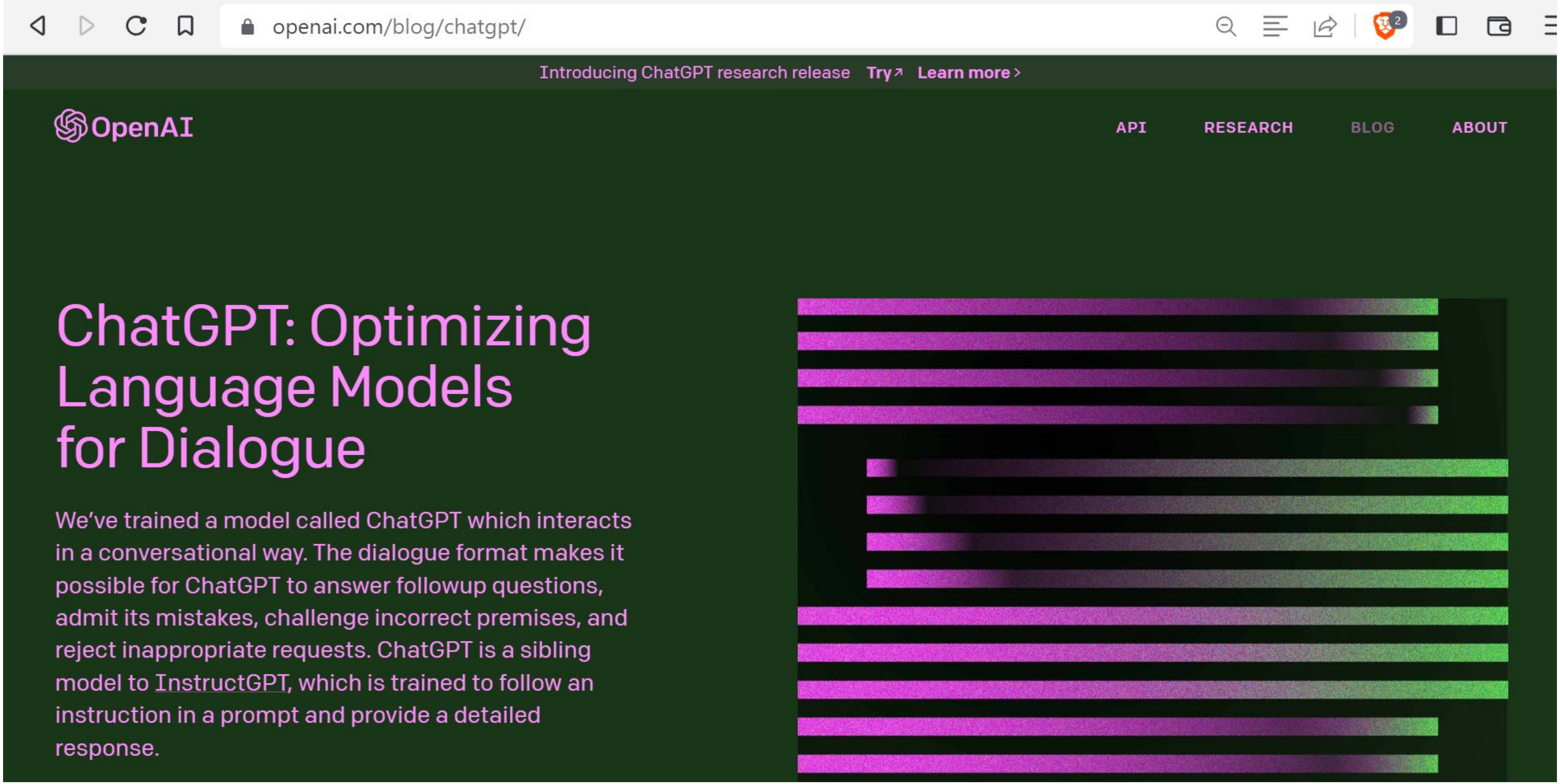
Tokenizer
A bat and
a rat
Good
Bag of words
a
and
bat
A
Stop
A
a
Good
Rare
Rosa
any
bat
Any
any
Bat
bat
Cat
Bag-of-words vector
[3, 1, 4, 8, 9, 0,...]
Algorithm
Data structure
Example
Applications
Characters
"Good morning Rosa."
Cryptography, compression, spelling correction, predictive text, search, dialog (chatbot)
Regular expression
Tokens
["Good", "morning", "Rosa"]
Search, stylistics, spam filter, sentiment analysis, word2vec math, semantic search, dialog (chatbot)
POS tagger (FST)
Tagged tokens
[("Good", "adjective"), ("morning", "singular noun"), ("Rosa", "singular proper noun"), (".", ".")]
Spelling and grammar correction, stylistics, dialog (chatbot)
Syntax tree
{ ("morning", SN, root): { ("Rosa", SPN, nmod): None, ("Good", ADJ, dep): None, } }
Question answering, stylistics, complex dialog, grammar correction, writing coach
Information extractor (FST)
Entity relationships
morning
good
Knowledge extraction and inference, medical diagnosis, question answering, game playing
Logic compiler (FST)
Knowledge base
morning
part of
day
Rosa
.82
name of
Person
.87
female
.74
English
Theorem proving, inference, natural language database queries, artificial general intelligence (AGI)

06

Future of NLP

Natural language IQ





openai.com/blog/chatgpt/

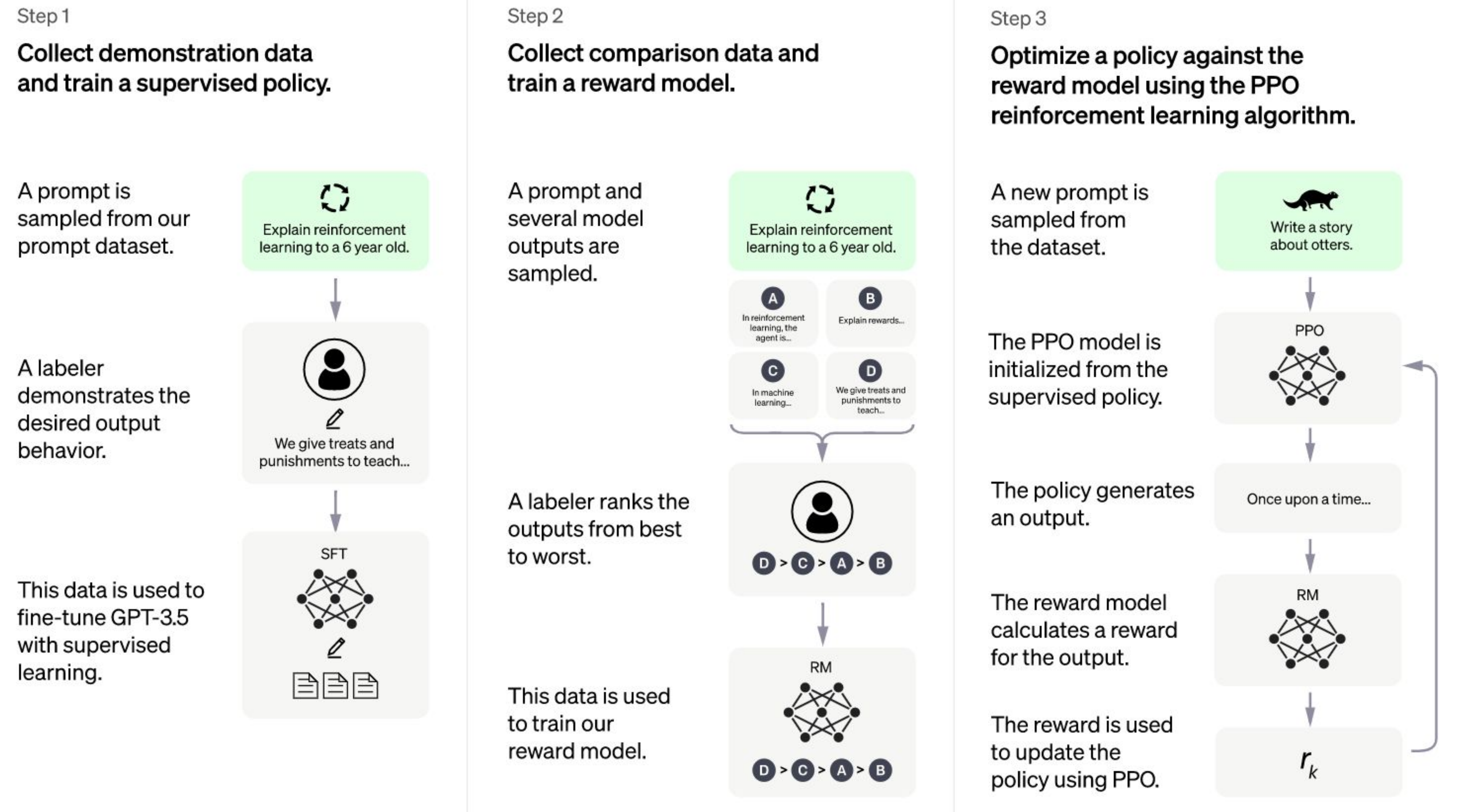
Introducing ChatGPT research release Try↗ Learn more >

OpenAI

API RESEARCH BLOG ABOUT

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.


Step 1
Collect demonstration data and train a supervised policy.
A prompt is sampled from our prompt dataset.
Explain reinforcement learning to a 6 year old.
A labeler demonstrates the desired output behavior.
We give treats and punishments to teach...
This data is used to fine-tune GPT-3.5 with supervised learning.
SFT
Step 2
Collect comparison data and train a reward model.
A prompt and several model outputs are sampled.
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
A labeler ranks the outputs from best to worst.
D > C > A > B
This data is used to train our reward model.
RM
D > C > A > B
Step 3
Optimize a policy against the reward model using the PPO reinforcement learning algorithm.
A new prompt is sampled from the dataset.
Write a story about otters.
The PPO model is initialized from the supervised policy.
PPO
The policy generates an output.
Once upon a time...
The reward model calculates a reward for the output.
RM
The reward is used to update the policy using PPO.
r_k

What are the topics in Natural Language Processing course

Some common topics that are covered in a course on Natural Language Processing (NLP) include:

- Language modeling and text generation
- Part-of-speech tagging and named entity recognition
- Syntax and parsing
- Semantics and sentiment analysis
- Text summarization and information extraction
- Machine translation and multilingual NLP
- Word sense disambiguation and coreference resolution
- Dialogue systems and conversational agents
- Ethics and bias in NLP

These topics are often covered using a combination of theoretical concepts and practical applications, and may include hands-on projects or programming assignments to give students experience implementing NLP techniques using various libraries or frameworks.

What are common school projects in NLP?

There are many different types of projects that students in an NLP course might work on, but some common project ideas include:

1. Text classification: This project might involve training a machine learning model to classify text into different categories, such as spam vs. non-spam emails or positive vs. negative movie reviews.
2. Named entity recognition: This project might involve developing a program that can identify and extract named entities from text, such as people, organizations, and locations.
3. Text generation: This project might involve training a language model to generate new text in a specific style or on a specific topic.
4. Sentiment analysis: This project might involve training a model to classify text as having a positive, negative, or neutral sentiment.
5. Machine Translation: This project can include training a machine learning model to translate from one language to another.
6. Text Summarization: This project can include training a model to generate a summary of a given text that tries to convey the most important information of the text.
7. Chatbot : This project can include implementing a chatbot system that can answer questions or engage in a conversation.
8. Word Embeddings: This project can include training a model to learn a vector representation of words.

These are just a few examples of the many types of NLP projects that students might work on. The specific projects will depend on the resources available and the skill level of the students, but the idea is to use a common NLP libraries and frameworks to apply the NLP concepts learned.

Regenerate response

07

Text and Reference Books

Text Books

- Ekaterina Kochmar - Getting Started with Natural Language Processing-Manning (2022)
- Hobson Lane, Hannes Hapke, Cole Howard - Natural Language Processing in Action_ Understanding, analyzing, and generating text with Python-Manning Publications (2019)
- Masato Hagiwara - Real-World Natural Language Processing_ Practical applications with deep learning-Manning Publications (2021)
- Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana - Practical Natural Language Processing_ A Comprehensive Guide to Building Real-World NLP Systems-O'Reilly Media, Inc. (2020)
- Stephan Raaijmakers - Deep Learning for Natural Language Processing-Manning Publications (2022)

Reference Books

- Neural network methods for natural language processing by Goldberg, Yoav
- Taweh Beysolow II - Applied Natural Language Processing with Python. Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing-Apress (2018)
- Thushan Ganegedara - Natural Language Processing with TensorFlow_ The definitive NLP book to implement the most sought-after machine learning models and tasks, 2nd Edition-Packt Publishing (2022)
- Denis Rothman - Transformers for Natural Language Processing_ Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more-Packt Publishing Ltd
- François Chollet - Deep Learning with Python-Manning Publications (2021)
- Paul Azunre - Transfer Learning for Natural Language Processing-Manning Publications (2021)