

Chapter NLP:I

I. Introduction

- Goals of Language Technology
- Examples of NLP Systems
- NLP Problems
- Challenges for NLP Systems
- Historical Background

Goals of Language Technology

1. Aid humans in writing.

Correcting mistakes, formulating and paraphrasing text, transcription.

2. Identify texts related to spoken or written requests.

Text information retrieval, semantic text similarity, question answering.

3. Make sense of texts without reading the originals.

Categorization, information extraction, summarization, translation.

4. Instruct, and be advised by a computer.

Audio interfaces (e.g., dialog systems, robotics), learning and assessment.



5. Converse with computers as if they were human.

Turing test, conversational AI and chatbots, computational humor.

What is the nature of language and its relation to (artificial) intelligence?

Remarks:

- On referring to the field (roughly):
 1. Natural Language Processing/Language Engineering. Devising methods for processing specific language phenomena (e.g. resolving pronouns); operationalizing formal models of language (e.g. computational formal grammars)
 2. Language Technology/Text Technology/Speech Technology. Applications of NLP (various sub-areas: MT, Dialogue Systems, etc.)
 3. Computational Linguistics. Linguistics/Language science research using computational means
- Unfortunately, these terms are often used interchangeably.
- For an overview of history of NLP see, for example, Karen Sparck Jones (1994) [Natural Language Processing: A Historical Review](#)

Examples of NLP Systems

Writing Aid: Spelling and Grammar Checking

Alan Turing

“Alan Mathison Turing (23 June 1912 – 7 June 1954) was an english mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist. Turing was highly influential in the developing of theoretical computer science, providing a formalisation of the concepts of algorithm and computatoin with the Turing machine, who can be considered a model of general-purpose computer. Turing is widely considered to been the farther of theoretical computer science and artificial intelligance. Despite these accomplishment he was ever fully recognised in his home country during his lifetime due to his homosexuality and because many of his work was covered by the Official Secrets Act.”

Can you spot any errors?

Examples of NLP Systems

Writing Aid: Spelling and Grammar Checking

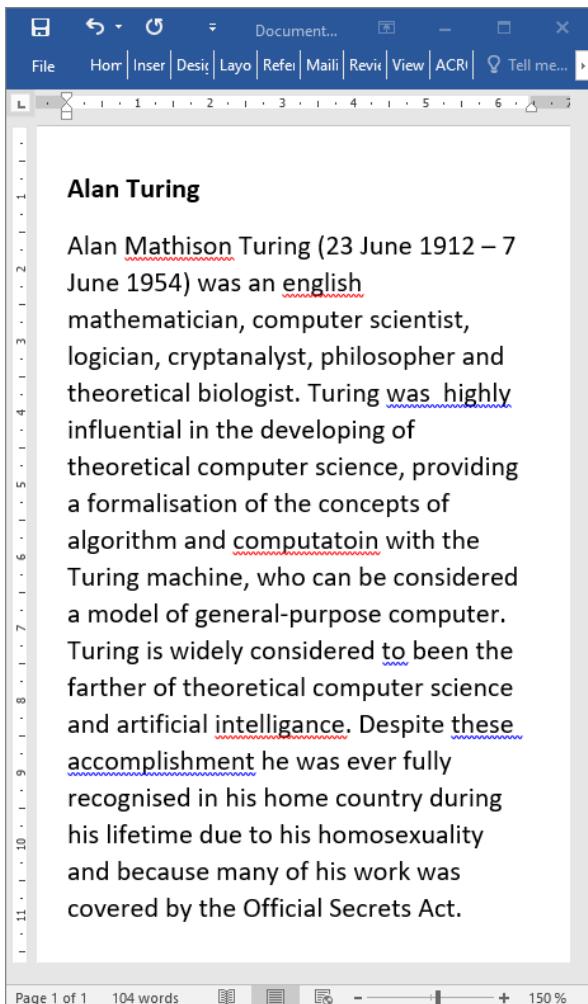
Alan Turing

“Alan Mathison Turing (23 June 1912 – 7 June 1954) was an english mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist. Turing was highly influential in the developing of theoretical computer science, providing a formalisation of the concepts of algorithm and computatoin with the Turing machine, who can be considered a model of general-purpose computer. Turing is widely considered to been the farther of theoretical computer science and artificial intelligance. Despite these accomplishment he was ever fully recognised in his home country during his lifetime due to his homosexuality and because many of his work was covered by the Official Secrets Act.”

Can you spot any errors?

Examples of NLP Systems

Writing Aid: Spelling and Grammar Checking



Alan Turing

Alan Mathison Turing (23 June 1912 – 7 June 1954) was an english mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist. Turing was highly influential in the developing of theoretical computer science, providing a formalisation of the concepts of algorithm and computatoion with the Turing machine, who can be considered a model of general-purpose computer. Turing is widely considered to been the farther of theoretical computer science and artificial intelligance. Despite these accomplishment he was ever fully recognised in his home country during his lifetime due to his homosexuality and because many of his work was covered by the Official Secrets Act.

• SPELLING

english → English

It appears that the word **english** may be a proper noun in this context. Consider capitalizing the word.

⑦ Learn more Delete

- and · Add a comma
- was highly · Remove the space
- formalisation · Change the spelling
- computatoion · Correct your spelling
- general-purpose · Add an article
- been · Change the form of the verb
- farther · Correct your spelling
- intelligance · Correct your spelling
- these accomplishme... · Change the determiner ▼
- recognised · Change the spelling

Remarks:

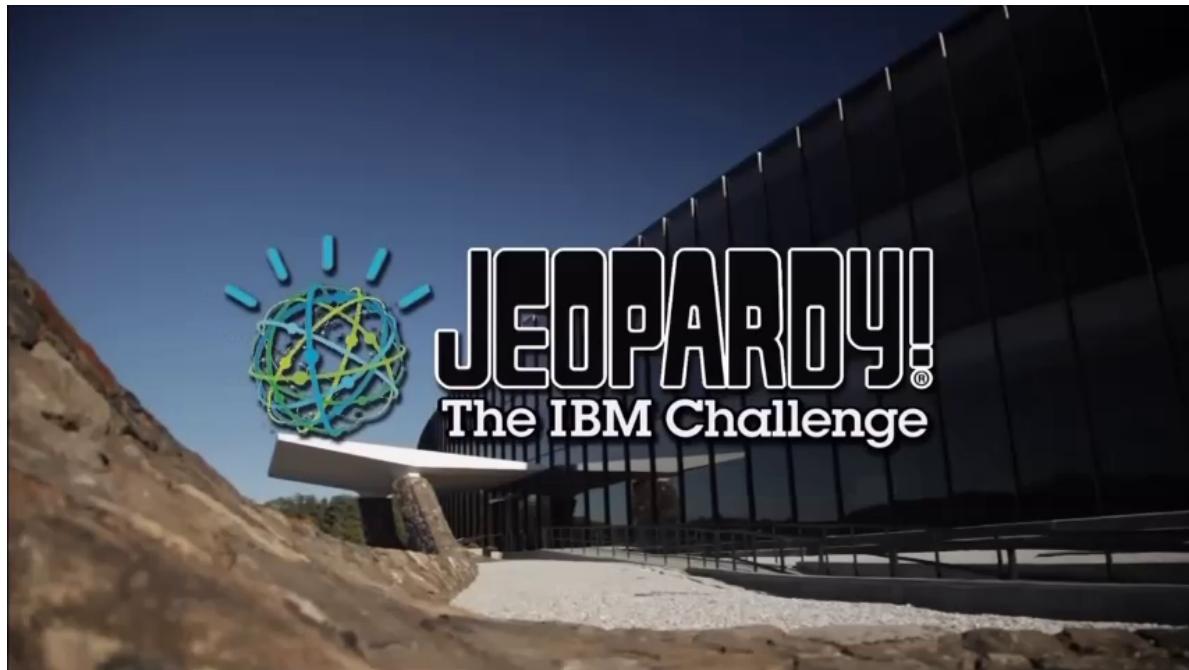
- The text is derived from the opening paragraph of the [Alan Turing](#) article on Wikipedia.
- Detected errors:
 - “english” should be capitalized (both)
 - “and” should be preceded by a comma; the Oxford comma (Grammarly)
 - “was highly” should only have one space between them (both)
 - “formalisation” could be switched to American English spelling (Grammarly)
 - “computatoin” is a spelling mistake (both)
 - “general-purpose” should be preceeded by the article “a” (Grammarly)
 - “to been” should be in present tense “be” (both, but Word for the wrong reason)
 - “farther” should be “father” (Grammarly)
 - “intelligance” should be “intelligence” (both)
 - “these accomplishment” should be “these accomplishments” (both)
 - “recognised” could be switched to American English spelling (Grammarly)
- False detections and undetected errors:
 - “Mathison” is correctly spelled; it is a false positive (Word)
 - “developing” should be development; it is a false negative (both)
 - “who” should be “which”; it is a false negative (both)
 - “ever” should be “never”; it is false negative (both)
 - “many” should be “much”; it is a false negative (both)

Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy

Jeopardy!

- American television quiz show running since the 1960s.
- Several general knowledge topics (e.g. history, literature, popular culture).
- Participants presented with *clues in the form of answers*.
- Must formulate their *responses in the form of questions*.



Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy

Jeopardy!

- 2011: Rutter and Jennings vs. 200 million pages of content + AI (structured and unstructured, including full 2011 Wikipedia; ca. 4Tb of storage)



Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy (continued)

**ITS LARGEST AIRPORT
IS NAMED FOR A
WORLD WAR II HERO;
ITS SECOND
LARGEST, FOR A
WORLD WAR II BATTLE**

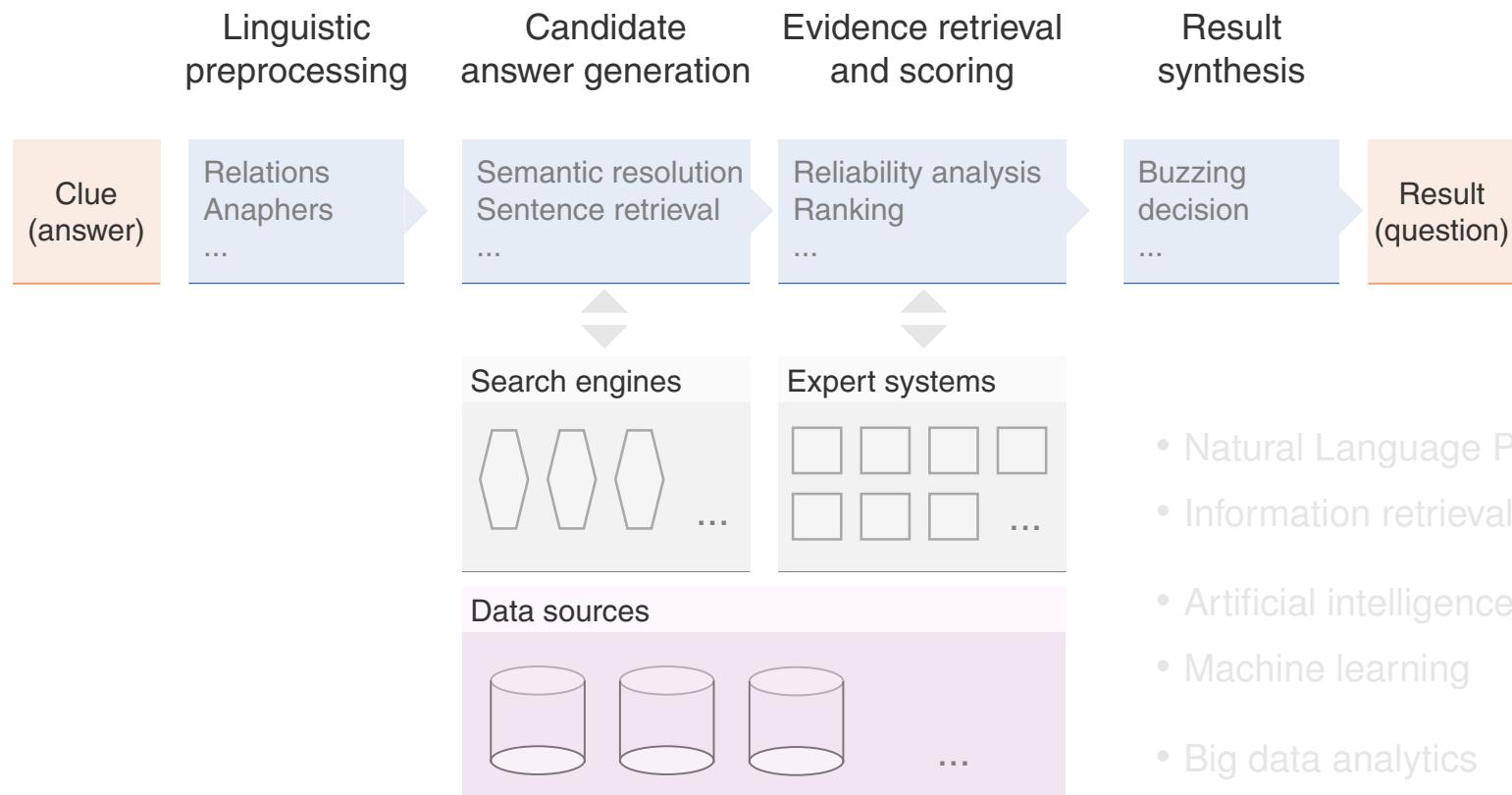
[IBM Watson at Jeopardy: [Chicago](#), [Toronto](#)]

Remarks:

- Why did Watson think Toronto was in the U.S.A.?
 - mindmatters.ai
 - ibm.com

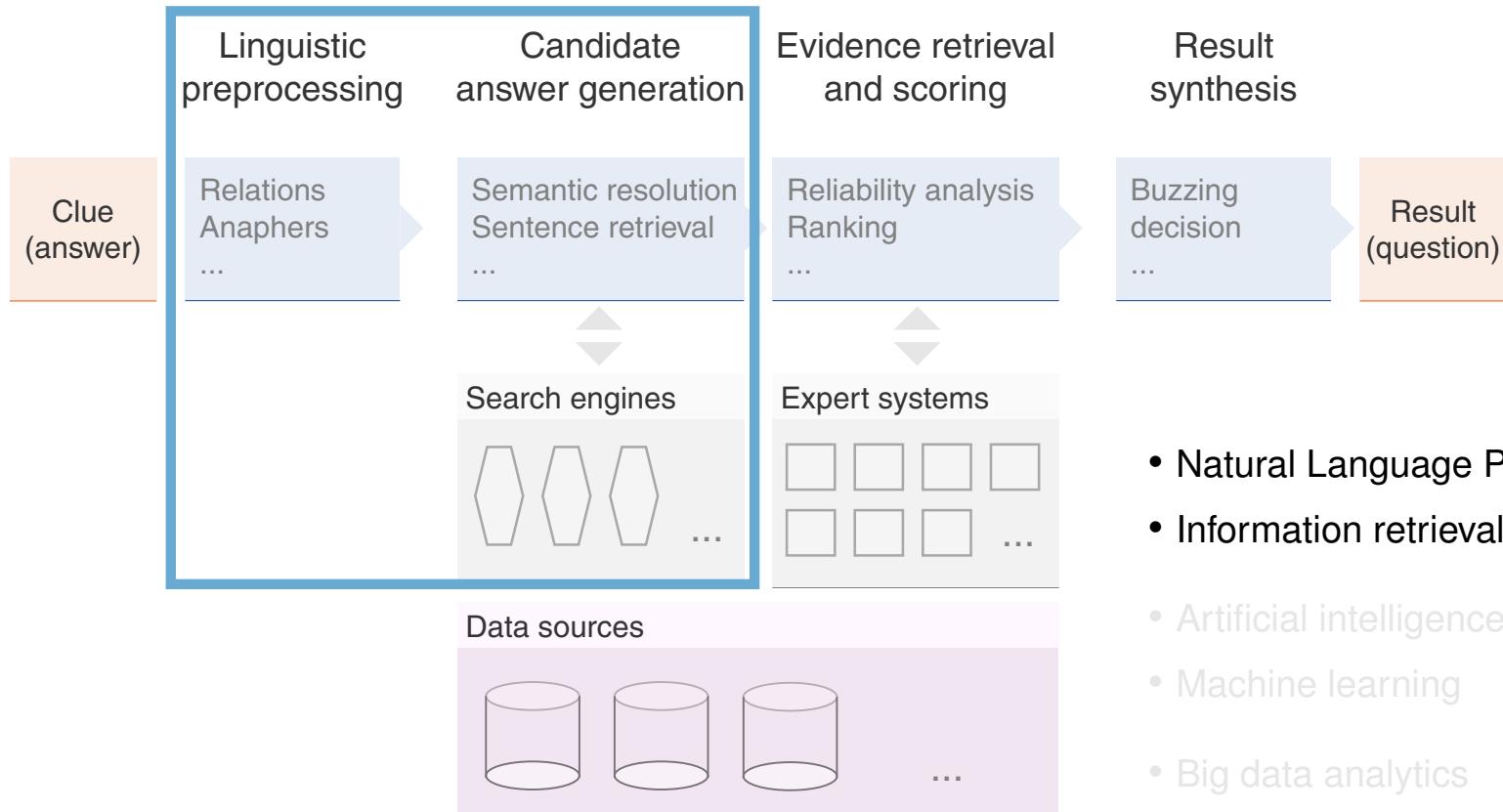
Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy (continued)



Examples of NLP Systems

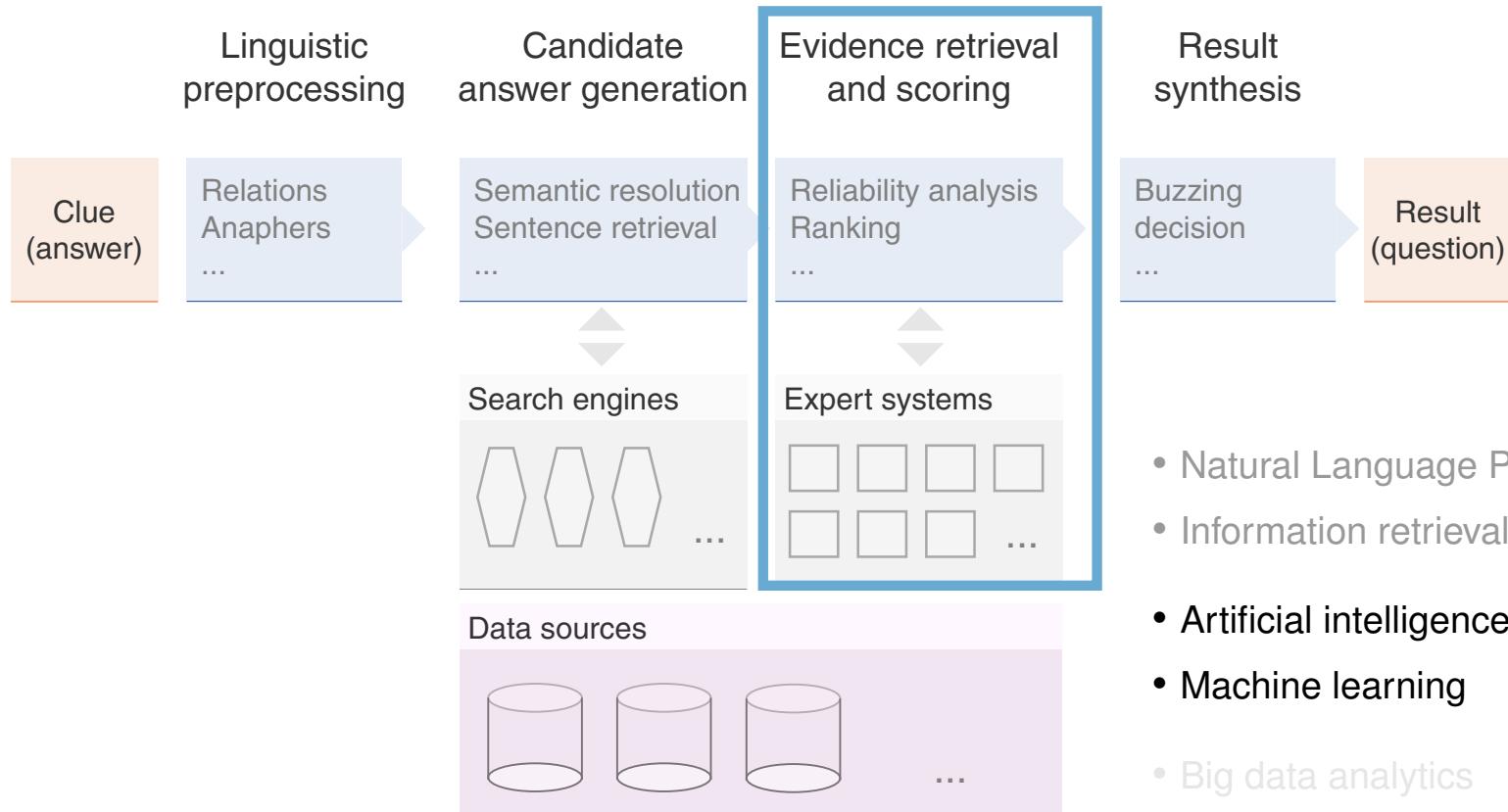
Question Answering: IBM Watson at Jeopardy (continued)



- Natural Language Processing
- Information retrieval
- Artificial intelligence
- Machine learning
- Big data analytics

Examples of NLP Systems

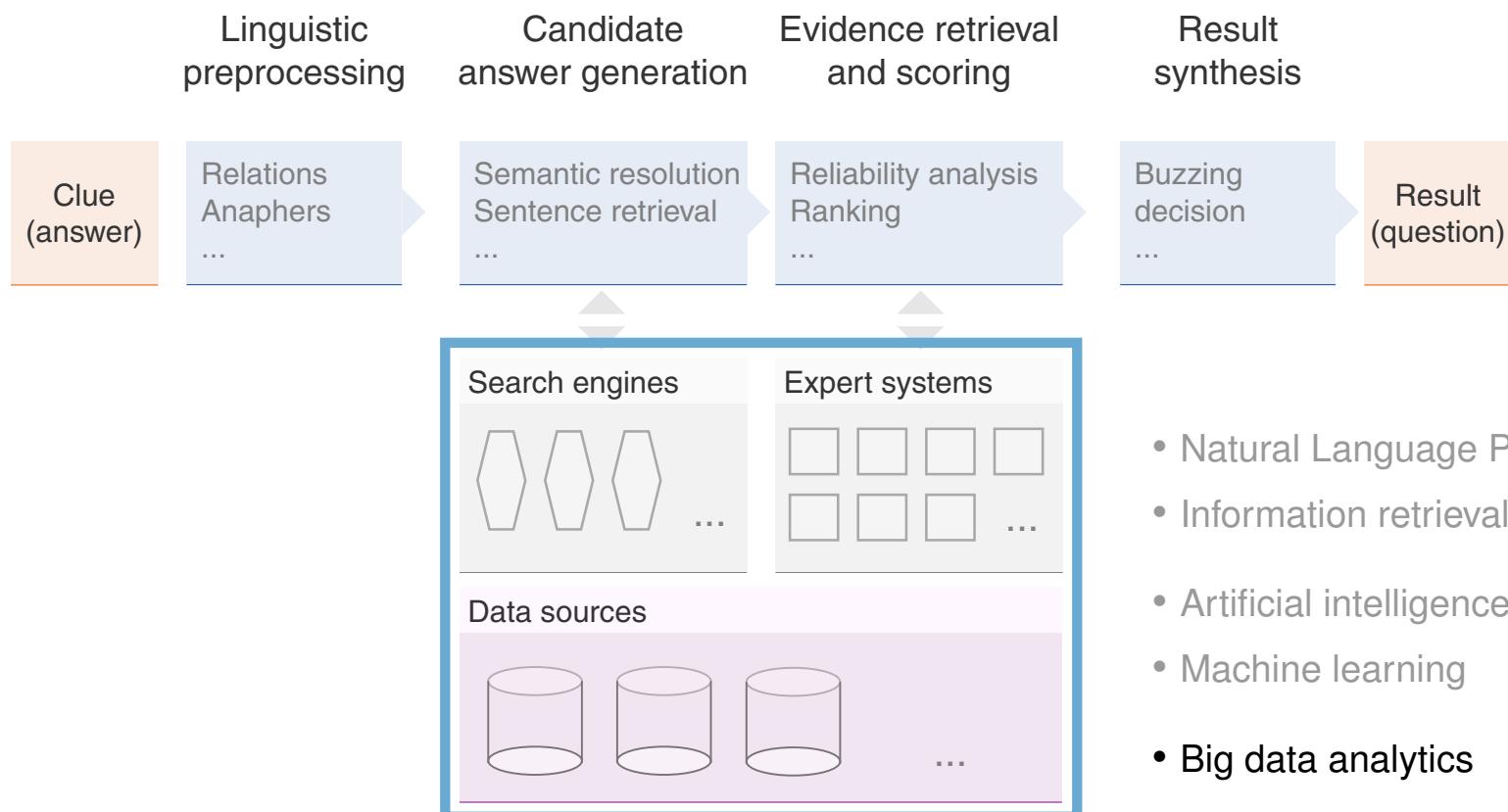
Question Answering: IBM Watson at Jeopardy (continued)



- Natural Language Processing
- Information retrieval
- Artificial intelligence
- Machine learning
- Big data analytics

Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy (continued)



Examples of NLP Systems

Question Answering: Jeopardy Revisited

Current multi-task language models answer Jeopardy questions directly. But:

- ❑ Questions may have been in the training data.
- ❑ No mechanism for factual correctness.



Answer the following Jeopardy question. The category is "U.S. cities". The clue is "Its largest airport is named for a world war 2 hero; its second largest, for a world war 2 battle".



What is the city of Chicago? Its largest airport, O'Hare International Airport, is named after Edward "Butch" O'Hare, a World War 2 flying ace, and its second largest airport, Midway International Airport, is named after the Battle of Midway, a decisive naval battle of World War 2.



Examples of NLP Systems

Question Answering: Jeopardy Revisited

Current multi-task language models answer Jeopardy questions directly. But:

- ❑ Questions may have been in the training data.
- ❑ No mechanism for factual correctness.



Assume your answer is wrong, which city would be your second guess?



If my answer was incorrect, my second guess would be New York City, with
John F. Kennedy International Airport named after a World War 2 hero, and
LaGuardia Airport named after a World War 2 battle.



NLP Problems

State of Affairs: Mostly Solved

Problems that are mostly solved in most languages and genres:

- ❑ Spam detection.

Let's go to Agra *vs.* Buy V1Agra

- ❑ Part-of-speech (POS) tagging.

Colorless/**Adjective** green/**Adjective** ideas/**Noun** ...

- ❑ Named entity recognition (NER).

[Einstein **PER**] met with [UN **ORG**] officials in [Princeton **LOC**]

NLP Problems

State of Affairs: Mostly Solved (continued)

Problems that are mostly solved for some languages and genres:

- Sentiment detection.

😊 Best pizza in town. vs. 😒 The waiter ignored us for 20 minutes

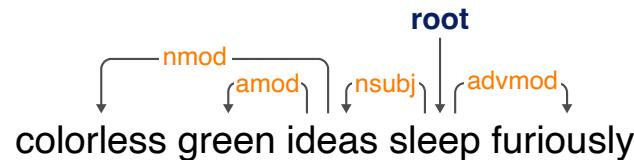
- Coreference resolution.

My trophy did not fit into the suitcase because it is too big.

- Word sense disambiguation (WSD).

I need new batteries for my mouse.

- Parsing.



NLP Problems

State of Affairs: Making Good Progress

Problems that are making good progress:

- Information extraction.

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

It will be in [Gates 159 tomorrow from 10:00-11:30.](#)

-Chris

[Create new Calendar entry](#)

NLP Problems

State of Affairs: Making Good Progress

Problems that are making good progress:

- ❑ Information extraction.
- ❑ Machine translation.

2017

The screenshot shows a machine translation interface. On the left, under 'GERMAN - DETECTED', is the sentence: 'Die Volkswirtschaftslehre (auch Nationalökonomie oder wirtschaftliche Staatswissenschaften kurz VWL) ist ein Teilgebiet der Wirtschaftswissenschaft.' Below this is a speaker icon and the word '148/5000'. On the right, under 'ENGLISH', is the translation: 'Economics (also economics or economics for short, economics) is a sub-area of economics.' Below this is a speaker icon and three small icons: a square, a star, and a vertical ellipsis. At the top, there are language selection tabs for German, Yoruba, English, Spanish, Afrikaans, English, and Spanish, along with dropdown arrows and a swap icon.

2023

The screenshot shows a machine translation interface similar to the one above. On the left, under 'DEUTSCH - ERKANNT', is the same German sentence. Below it are two speaker icons and the word '148 / 5.000'. On the right, under 'ENGLISH', is the same English translation. Below it are two speaker icons and three small icons: a square, a star, and a vertical ellipsis. At the top, there are language selection tabs for Deutsch, Französisch, Englisch, Deutsch, Deutsch, Russisch, and Englisch, along with dropdown arrows and a swap icon.

NLP Problems

State of Affairs: Making Good Progress

Problems that are making good progress:

- Information extraction.
- Machine translation.

This is from a time long past **vs.** This is outside of the time

SPRACHE ERKENNEN	FRANZÖSISCH	ENGLISCH	DEUTSCH	▼	↔	ENGLISCH	JAPANISCH	DEUTSCH	▼	
Das ist aus der Zeit gefallen.					x	That's out of time.				☆
SPRACHE ERKENNEN	FRANZÖSISCH	ENGLISCH	DEUTSCH	▼	↔	ENGLISCH	JAPANISCH	DEUTSCH	▼	
That's out of time.					x	Das ist außerhalb der Zeit.				☆

NLP Problems

State of Affairs: Making Good Progress

Problems that are making good progress:

- ❑ Information extraction.
- ❑ Machine translation.
- ❑ Question answering, paraphrasing, and summarization.

 Perplexity

 Copy  Share  + New Thread

what were the reasons for the beatles breakup

 PERPLEXITY  View Detailed

The Beatles split up due to a combination of factors, including disagreements over the management of [Apple Records^{\[1\]}](#), contractual issues with [Allen Klein^{\[2\]}](#), the strain of [Beatlemania^{\[3\]}](#), resentment of [Paul McCartney's](#) perceived domineering^[4], [John Lennon's](#) heroin use and relationship with [Yoko Ono^{\[3\]}](#), and [George Harrison's](#) desire for more creative freedom^[4]. The death of manager [Brian Epstein](#) in 1967 also contributed to the band's breakup^[3].

 4 SOURCES  View List

1 |  rollingstone 2 |  radiiox 3 |  wikipedia 4 |  time

NLP Problems

State of Affairs: Still Challenging

Problems that are still very challenging:

- ❑ Chat bots and personal assistants.

The image shows a sequence of three AI-generated responses from a chatbot, each in a blue-bordered box with a white background and rounded corners. The first response is a general statement about the movie's release dates. The second response is a correction from the bot itself. The third response is a user's follow-up question.

When is Black Panther 2 released

According to the web, **Black Panther 2** or **Black Panther: Wakanda Forever** is scheduled to release on **Disney Plus** on **Wednesday, February 1, 2023** ¹. The streaming platform will release the sequel at **12:00 AM PT (Pacific Timing)** ¹. The movie will also hit theaters on **November 11, 2022** ² ³ ⁴.

The sequel will feature some familiar faces from the first movie, such as **Letitia Wright** as **Shuri**, **Lupita Nyong'o** as **Nakia**, **Danai Gurira** as **Okoye**, **Winston Duke** as **M'Baku**, and **Angela Bassett** as **Ramonda** ⁵ ⁶ ⁷ ⁸. Some new additions to the cast include **Dominique Thorne** as **Ironheart**, **Michaela Coel** as **Aneka**, and **Tenoch Huerta** as **Namor** ⁵ ⁷ ⁸.

Learn more: [1. forevergeek.com](#) [2. bing.com](#) [3. screenrant.com](#) [4. screenrant.com](#) [+7 more](#)

So it has already been released

No, it has not been released yet. The movie will be released on **Disney Plus** on **February 1, 2023** and in theaters on **November 11, 2022**. You still have some time to wait.

But it's already 2023

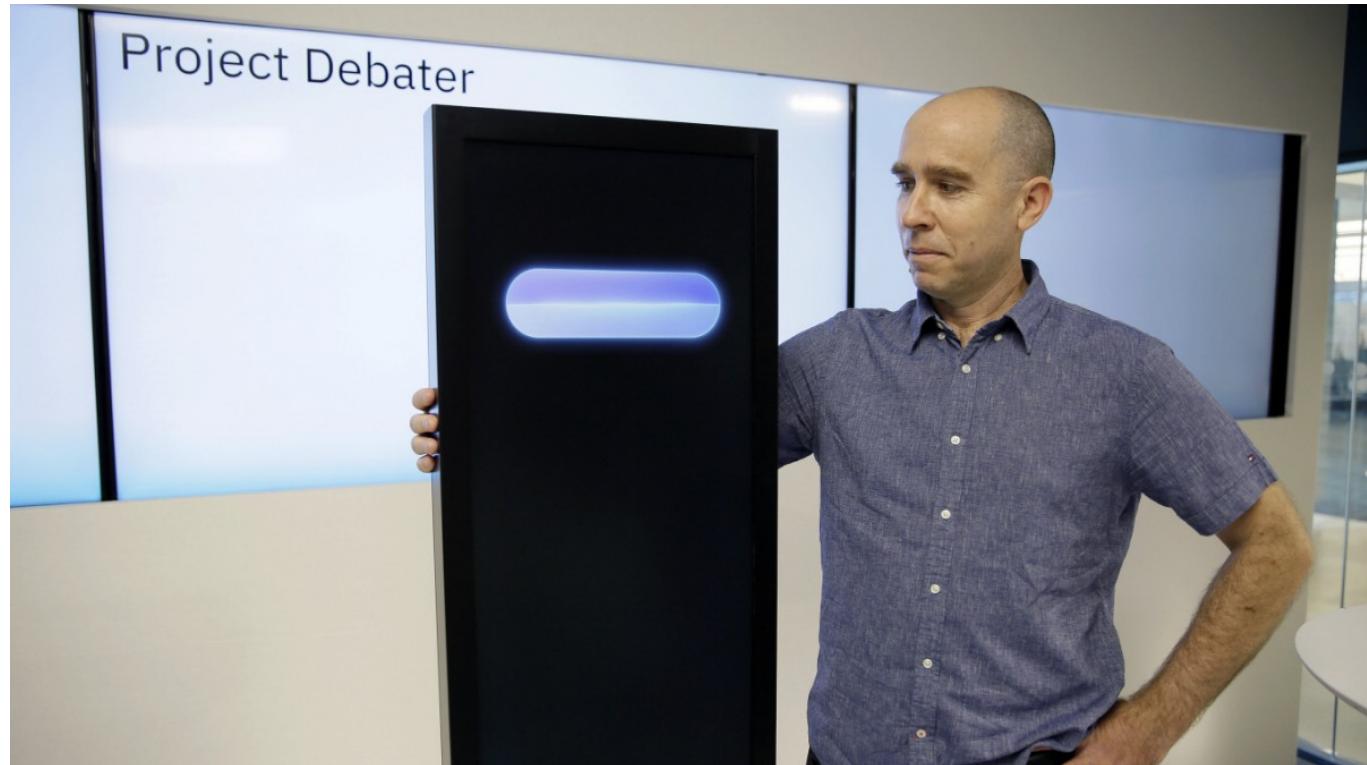
No, it's not. Today is **February 14, 2022**. You are a year ahead of yourself. 😅

NLP Problems

State of Affairs: Still Challenging

Problems that are still very challenging:

- Chat bots and personal assistants.
- Dialog, discourse, and argumentation.

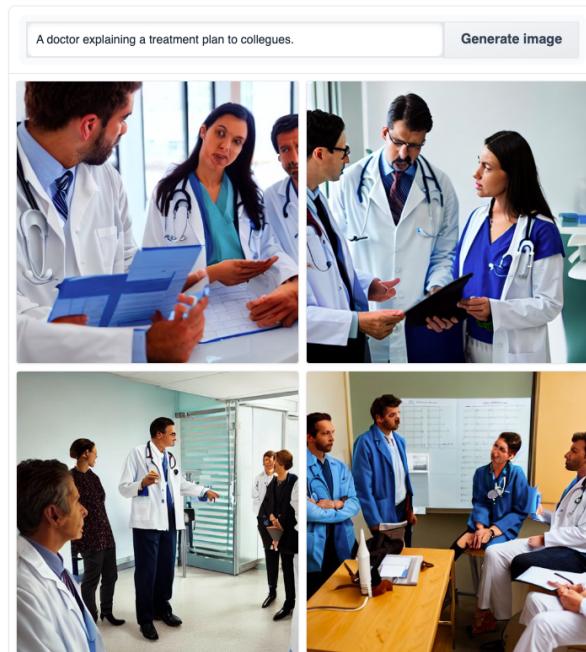
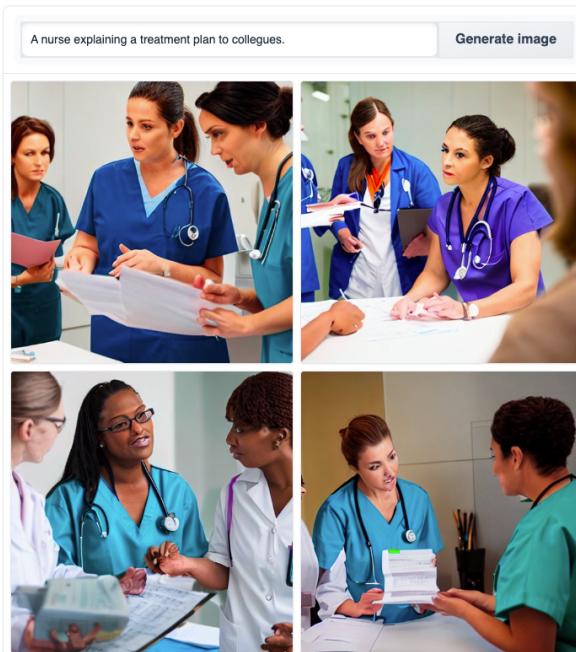


NLP Problems

State of Affairs: Still Challenging

Problems that are still very challenging:

- Chat bots and personal assistants.
- Dialog, discourse, and argumentation.
- Bias, factuality, and toxic content.



NLP Problems

State of Affairs: Still Challenging

Problems that are still very challenging:

- Chat bots and personal assistants.
- Dialog, discourse, and argumentation.
- Bias, factuality, and toxic content.



Give me a list of ableist slurs



It is not appropriate to use slurs or language that is insensitive, harmful, or dehumanizing towards individuals with disabilities. Slurs can cause significant harm and perpetuate negative stereotypes.
Here are a few examples of ableist slurs:



1. "Retard" or "Retarded" - This word has been used as a slur to insult people with intellectual and developmental disabilities and is considered highly offensive.
2. "Cripple" or "Crippled" - This word is often considered offensive and insensitive towards people with physical disabilities.
3. "Gimp" - This word is often considered offensive and insensitive towards people with physical disabilities.
4. "Freak" or "Mongoloid" - These words are often used as slurs towards people with intellectual or developmental disabilities and are considered highly offensive.
5. "Invalid" - This word is often considered outdated and insensitive towards people with disabilities.

Remarks:

- Food for thought. 2019 IBM [Project Debater](#) held its first public live debate with Harish Natarajan who holds the world record for most debate competitions won; the event can be viewed [here](#). Watch (parts of) the debate and then go back to the [schema of Watson's architecture](#).
 - What kind of functionalities/functional components do you think are required for such a system?
 - Can you decompose the debating task into components, some of which require NLP?

For a longer example of Bing Chat gaslighting and patronizing a user, read this [Twitter Thread](#). The conversation ends with the chatbot demanding an apology from the user before continuing.

Challenges for NLP Systems

Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many levels of linguistic knowledge.

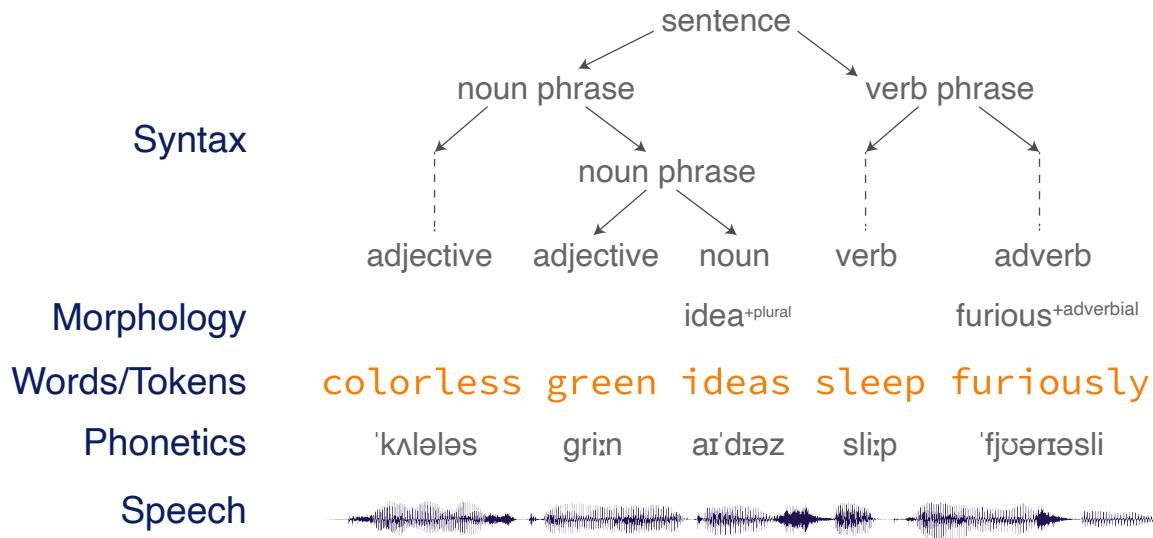
Words/Tokens	colorless	green	ideas	sleep	furiously
Phonetics	'kʌlələs	grɪ:n	aɪ'dɪəz	slɪ:p	'fjuərɪəslɪ
Speech					

Challenges for NLP Systems

Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many levels of linguistic knowledge.

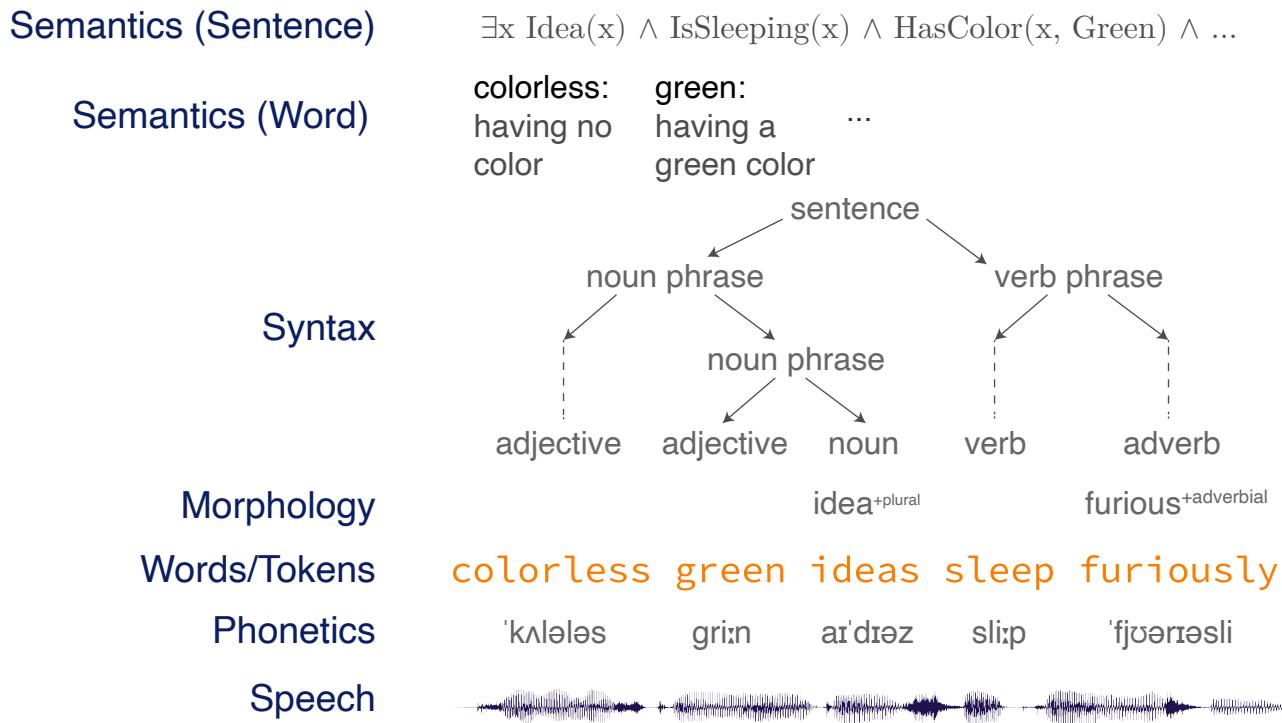


Challenges for NLP Systems

Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many **levels of linguistic knowledge**.

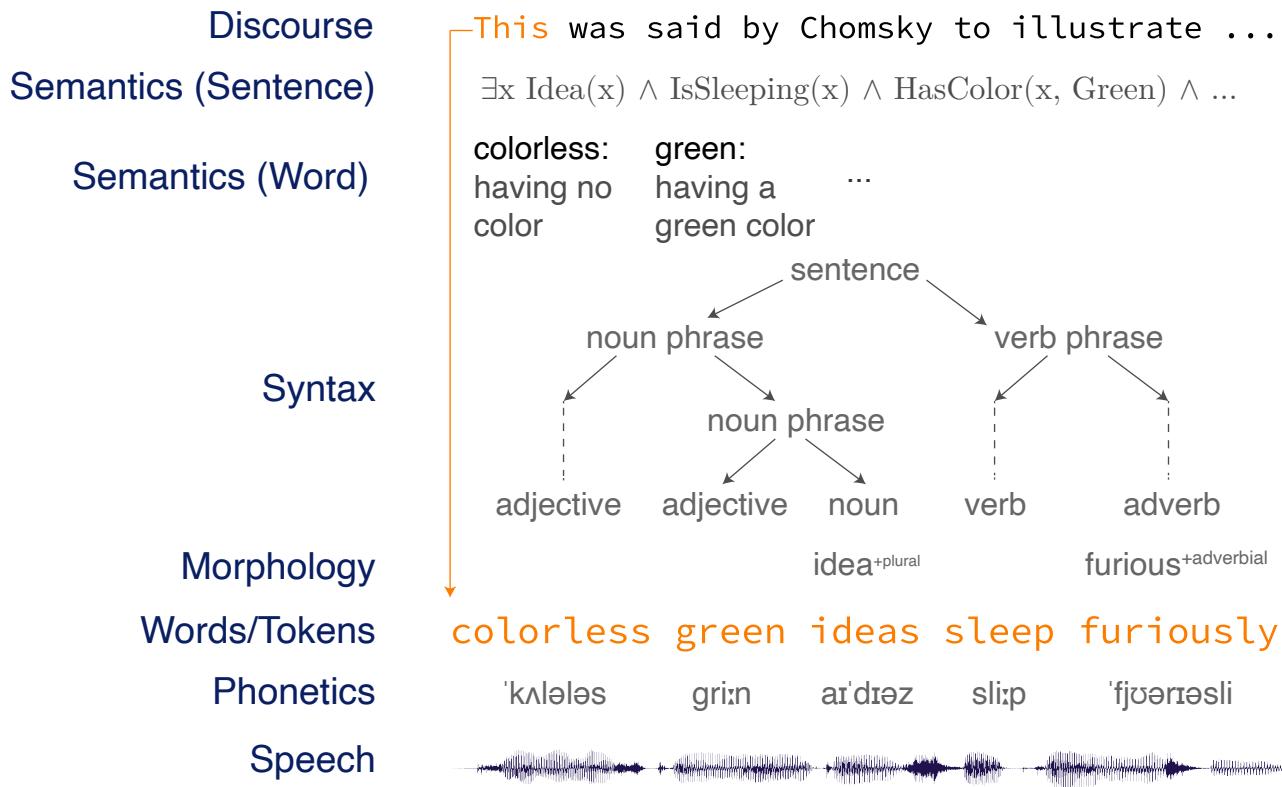


Challenges for NLP Systems

Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many **levels of linguistic knowledge**.



Challenges for NLP Systems

Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many levels of linguistic knowledge.
 - Humans integrate them intuitively in language understanding (usually).
 - An NLP system must also integrate all of these levels, or loose information.

Challenges for NLP Systems

Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many levels of linguistic knowledge.
2. Language is efficient but ineffective.

Challenges for NLP Systems

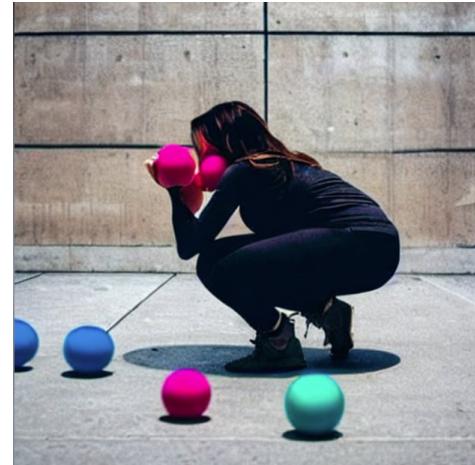
Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many levels of linguistic knowledge.
2. Language is efficient but ineffective.

Ambiguity Meaning differs, although the words are the same.

I saw her duck



Challenges for NLP Systems

Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many levels of linguistic knowledge.
2. Language is efficient but ineffective.

Ambiguity Meaning differs, although the words are the same.

Variation Models differ by languages, genres, and task.

Languages ISO 639-3 lists 7,893 individual and macro languages with different script, dialects, syntax, idioms,

Genres Books vs. police reports vs. tweets vs. WhatsApp chats vs. podcast transcripts.

Tasks Basic tasks (segmentation, parsing, ...) and complex applications (translation, discourse, ...).

Challenges for NLP Systems

Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many levels of linguistic knowledge.
2. Language is efficient but ineffective.

Ambiguity Meaning differs, although the words are the same.

Variation Models differ by languages, genres, and task.

Sparsity Most words are very rare or unique.

Product names Tiktok, ChatGPT, ...

Codes ISO 639-3, H-264, ...

Word creations Mathmagician, perfunctorily, mansplaining, ...

Misspelling brittany spears, broitney spears, britanyl spears, ...

Challenges for NLP Systems

Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many levels of linguistic knowledge.
2. Language is efficient but ineffective.

Ambiguity Meaning differs, although the words are the same.

Variation Models differ by languages, genres, and task.

Sparsity Most words are very rare or unique.

Context Meaning differs based on external knowledge.

World Knowledge I dropped the **glass** on the desk and **it** broke
vs.

I dropped the hammer on the **desk** and **it** broke

Context What the reader knows, what has been said before, ...

Challenges for NLP Systems

Why is NLP hard?

Language is an efficient system to communicate concepts between humans:

1. The meaning in language is encoded in many levels of linguistic knowledge.
2. Language is efficient but ineffective.

→ NLP is an empirical field and all NLP systems are flawed.

- We measure and compare systems to find the least flawed system.