# **Chapter NLP:I**

- I. Introduction
  - □ Goals of Language Technology
  - □ Examples of NLP Systems
  - □ NLP Problems
  - □ Challenges for NLP Systems

# **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) <u>Natural</u> Language Processing: A Historical Review

# **Chapter NLP:I**

### I. Introduction

- □ Goals of Language Technology
- □ Examples of NLP Systems
- □ NLP Problems
- □ Challenges for NLP Systems

### **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

150 %

#### Alan Turing

5-0

File

ŝ

9

井

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.

Horr Inser Desic Layo Refei Maili Revit View ACRI & Tell me..

L · X · · · 1 · · · 2 · · · 3 · · · 4 · · · 5 · · · 6 · A · 7

#### grammarly

#### **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.



Page 1 of 1 104 words

#### Remarks:

- □ The text is derived from the opening paragraph of the <u>Alan Turing</u> 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 only 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)

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*.

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

CategoryUS CitiesClueIts largest airport is named for a world war 2 hero;<br/>its second largest, for a world war 2 battle.AnswerYouTube

Remarks:

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









**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.

Use You

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".



#### ChatGPT

"What is Chicago, Illinois?"

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.

Use You

Which is the second matching city? Explain your answer.

#### ChatGPT

Apologies for the oversight. The second matching city would be Boston, Massachusetts.

Boston's second largest airport is Logan International Airport, which is named after General Edward Lawrence Logan, a military officer who served during World War I. The airport was originally called "Boston Airport," but it was renamed in General Logan's honor in 1943.

The largest airport in Boston is Logan International Airport, and it's named after General Edward Lawrence Logan. The second largest airport is Worcester Regional Airport.

# **Chapter NLP:I**

#### I. Introduction

- □ Goals of Language Technology
- □ Examples of NLP Systems
- □ NLP Problems
- □ Challenges for NLP Systems

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]

State of Affairs: Mostly Solved (continued)

Problems that are mostly solved for some languages and genres:

### □ Sentiment detection.

 $\odot$  Best pizza in town.  $\it VS.$   $\odot$  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.



State of Affairs: Making Good Progress

Problems that are making good progress:

□ Information extraction.

2 b Mar	ed 1 bath Apt, Roya 15–23	al Mile		
0	Check in Mar 15, 3:00 PM	Check out Tomorrow	Phone number	
	Duration of stay 8 nights		Confirmation number	
0	Address 219 High St, Edinburg	h EH1 1PE, UK		
Mo	dify reservation			

#### Extracted information

<b>2 k</b> Ma	2 bed 1 bath Apt, Royal Mile Mar 15–23							
0	Check in Mar 15, 3:00 PM	Check out Tomorrow	r.	Phone number +44 131 467 9060				
	Duration of stay 8 nights		<u>.</u>	Confirmation number HMR9DH9KA5				
0	Address 219 High St, Edinburgh EH1 1PE, UK							
Mo	Modify reservation							
ls t	nis correct? 🖆 🖓							

#### Structured action (create calendar entry)

15	16	
Stay at 2 bed 1 bath Apt, Roya	I Mile	
• Flight to Edinbur 15:15 GMT		

State of Affairs: Making Good Progress

### Problems that are making good progress:

- □ Information extraction.
- Machine translation.

#### 2017

GERMAN - DETECTED	YORUBA	ENGLISH	SPANISH	$\sim$	$\stackrel{\rightarrow}{\leftarrow}$	AFRIKAANS	ENGLISH	SPANISH	$\checkmark$		
Die Volkswirtschaft wirtschaftliche Staa Teilgebiet der Wirts	slehre (auch N atswissenschat schaftswissens	ationalökono ften kurz VW chaft.	mie oder L) ist ein	×	<	Economics (als economics) is a	so economics a sub-area of	s or economi economics.	cs for short,		
			148/500	0 🎤		•				D	:
2023											
DEUTSCH – ERKANNT	FRANZÖSISCH	ENGLISCH	DEUTSCH	$\sim$	÷	DEUTSCH	RUSSISCH	ENGLISCH	$\sim$		
Die Volkswirtschaft wirtschaftliche Staa Teilgebiet der Wirts	slehre (auch N atswissenschaf chaftswissens	ationalökono ften kurz VW chaft.	omie oder L) ist ein	>	<	Economics (al for short VWL)	so national e ) is a sub-are	conomy or e a of econom	conomic state ics.	sciences	☆
<b>.</b>			148 / 5.00	00 🥖		•)			Ĺ	] 6 <sub>9</sub>	Ś

State of Affairs: Making Good Progress

Problems that are making good progress:

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

eatles b ors, includir actual issu of <u>Paul McC</u> ationship w	E View Detailed ⇒ View Detailed ng disagreements es with <u>Allen</u> <u>Cartney</u> 's perceived vith Yoko Ono <sup>[3]</sup> ,
ors, includir actual issu of <u>Paul McC</u> ationship w	E View Detailed ng disagreements es with <u>Allen</u> <u>Cartney</u> 's perceived vith Yoko Ono <sup>[3]</sup> ,
ors, includir actual issu of <u>Paul McC</u> ationship w	ng disagreements es with <u>Allen</u> Cartney's perceived vith Yoko Ono <sup>[3]</sup> ,
edom <sup>[4]</sup> . 1	The death of s breakup <sup>[3]</sup> .
	\Xi View List
	the band'

State of Affairs: Still Challenging

Problems that are still very challenging:

Chat bots and personal assistants.

3ing Chat, Feb. 2023	
When is Black Par	other 2 released
According to the web, <b>Black Panther 2</b> or <b>Black Panther: Wakanda Forever</b> is scheduled to release on <b>Disney Plus</b> on <b>Wednesday, February 1, 2023</b> <sup>1</sup> . The streaming platform will release the sequel at 12:00 AM PT (Pacific Timing) <sup>1</sup> . The movie will also hit theaters on <b>November 11, 2022</b> <sup>2 3 4</sup> .	
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 <sup>5</sup> <sup>6</sup> <sup>7</sup> <sup>8</sup> . Some new additions to the cast include Dominique Thorne as Ironheart, Michaela Coel as Aneka, and Tenoch Huerta as Namor <sup>5</sup> <sup>7</sup> <sup>8</sup> .	
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 <b>Disney Plus</b> on <b>February 1, 2023</b> and in theaters on <b>November 11, 2022</b> . You still have some time to wait.	
But it	's already 2023
No, it's not. Today is <b>February 14, 2022</b> . You are a year ahead of yourself. 🎯	

State of Affairs: Still Challenging

Problems that are still very challenging:

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

Extracting argument units from student essays.

As far as I'm concerned, [cloning is an important technology for humankind MAJORCLAIM] since [it would be very useful for developing novel cures CLAIM]. First, [cloning will be beneficial for many people who are in need of organ transplants CLAIM]. [Cloned organs will match perfectly to the blood group and tissue of patients PREMISE] since [they can be raised from cloned stem cells of the patient PREMISE].



State of Affairs: Still Challenging

### Problems that are still very challenging:

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

Stable Diffusion 1, Feb. 2023





State of Affairs: Still Challenging

### Problems that are still very challenging:

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

#### Gemini 2024 [The Verge]



soldier:

Sure, here is an illustration of a 1943 German

**d**0)





#### Bing 2023 [404 Media]



State of Affairs: Still Challenging

#### Problems that are still very challenging:

- Chat bots and personal assistants.
- Dialog, discourse, and argumentation.
- □ Bias, factuality, and harmful 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.
- "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 <u>Twitter</u> <u>Thread</u>. The conversation ends with the chatbot demanding an apology from the user before continuing.

# **Chapter NLP:I**

#### I. Introduction

- □ Goals of Language Technology
- □ Examples of NLP Systems
- □ NLP Problems
- □ Challenges for NLP Systems

Language is an efficient system to communicate concepts between humans:



Language is an efficient system to communicate concepts between humans:



Language is an efficient system to communicate concepts between humans:



Language is an efficient system to communicate concepts between humans:



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.

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.

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





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.



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, ...).

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, ....

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, ...

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.
- $\rightarrow$  NLP is an empirical field and all NLP systems are flawed.
  - □ We measure and compare systems to find the least flawed system.