

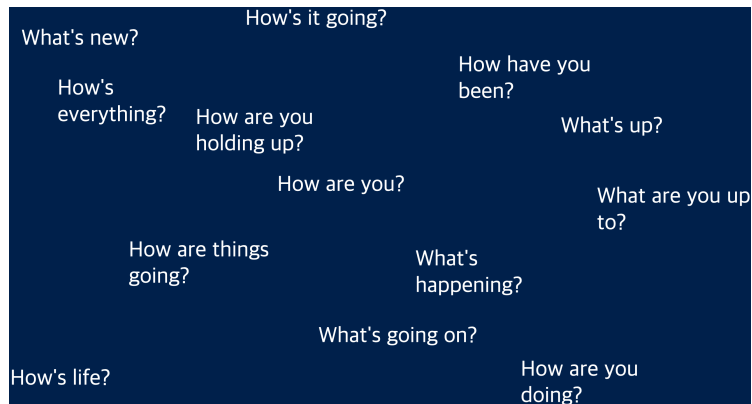
Natural Language Processing

What NLP Is and Why It Is Hard

Ramaseshan Ramachandran

Introduction

The goal of NLP is to enable computers to understand, interpret, and generate human language. Increasingly, the goal also includes conversing, reasoning, and acting on it.





Intended for Human Machine Communications

- ♦ Instructions are
 - ♦ Precise
 - ♦ Unambiguous
 - ♦ Mathematical equations

THE CHALLENGE OF LANGUAGE

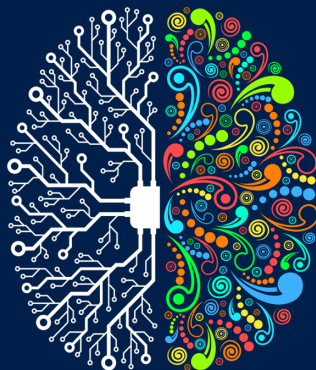
Human language is complex, ambiguous, and full of nuances that make it difficult for machines to process accurately.

- ▶ Local high school dropouts cut in half
- ▶ The king saw a rabbit with his glasses
- ▶ Juvenile court to try shooting defendant
- ▶ Safety experts say school bus passengers should be belted



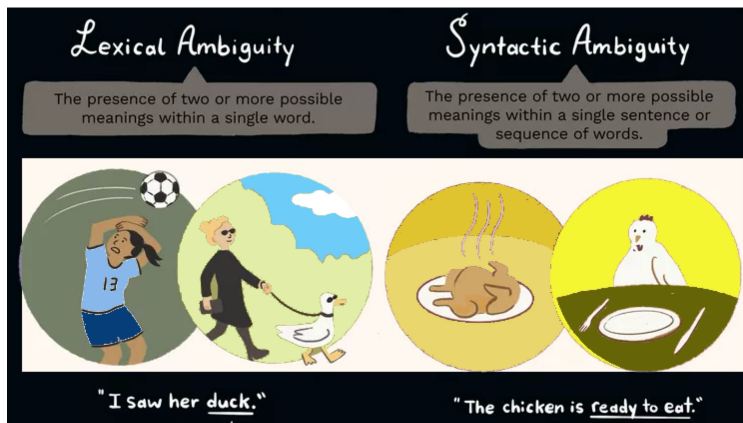
WHY IS NLP HARD? I

Creative and Analytical Representation of Ideas



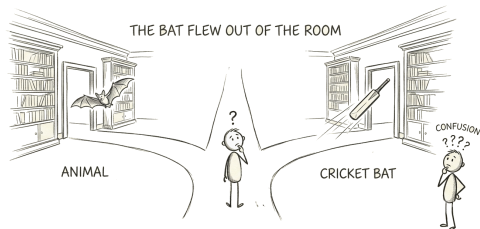
WHY IS NLP HARD? II

Ambiguous in nature

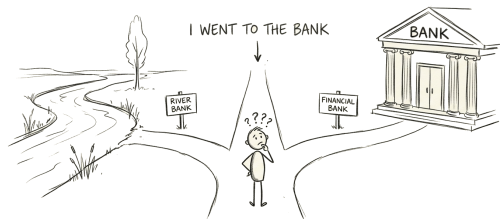


Source: Definition and Examples of Ambiguity in English(thoughtco.com)

ADDITIONAL EXAMPLES - LEXICAL AMBIGUITY

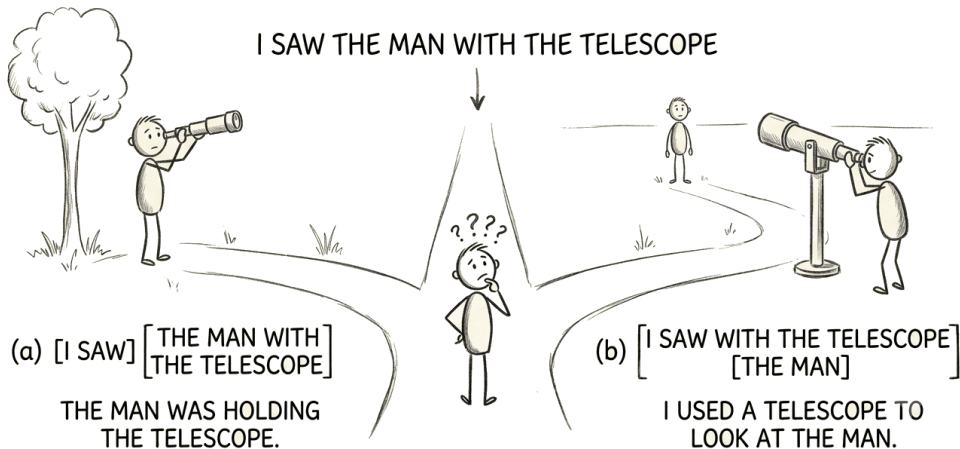


The **bat** flew out of the room



I went to the **Bank**

SYNTACTIC AMBIGUITY



English: I saw the man [using] a telescope.

Sanskrit: अहं दूरदर्शकेन नरम् अपश्यम्
(Ahaṁ dūradarśakena naraṁ apaśyam)

अहं	Ahaṁ	I	[Nominative]
दूरदर्शकेन	Dūradarśakena	using/by means of the telescope	[Instrumental case]
नरम्	Naraṁ	the man	[Accusative]
अपश्यम्	Apaśyam	saw	[Verb]

The telescope is the *instrument* of seeing: the case ending alone fixes the reading, with no word order to disambiguate.

THE SANSKRIT VERSION — ADJECTIVAL READING

English: I saw the man [who possessed] the telescope.

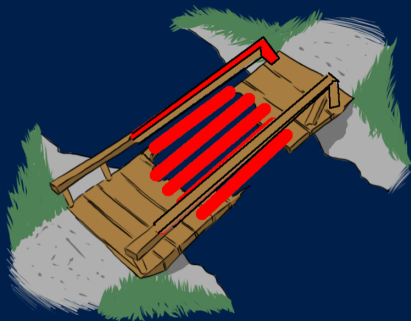
Sanskrit: अहं दूरदर्शकवन्तं नरम् अपश्यम्
(Ahaṁ dūradarśakavantam naram apaśyam)

अहं	Ahaṁ	I	[Nominative]
दूरदर्शकवन्तम्	Dūradarśakavantam	the one possessing a telescope	[Accusative adjective]
नरम्	Naram	the man	[Accusative]
अपश्यम्	Apaśyam	saw	[Verb]

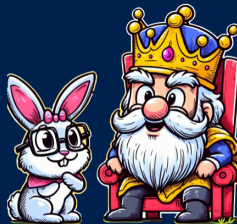
Because the modifier must agree with the man (**naram**) in gender, number, and case (Accusative), it cannot be mistaken for the instrument of the verb.

Ambiguous in nature

Red tape holds up new bridges



PRAGMATIC CONSIDERATIONS



Multiple representations of the same scenario

- ❖ The weather is extremely cold today
- ❖ It is freezing out there



SARCASM IS HARD EVEN FOR HUMANS

Irony, sarcasm (Yeah! right), double negatives, etc. make it difficult for automatic processing



Me: I am 25 years old

Mirror: Yeah, right

COMPLEX REPRESENTATION

Complex representation information (simple to hard vocabulary with uncommon usage of a sentence)

He was, in the way of most men, possessed of a rudimentary intelligence, his countenance ordinary, his bearing mild, with some weakness about the shoulders, his hair the color of ash; he spoke of the weather



The complex houses married and single soldiers and their families

TYPICAL TASKS

Information Retrieval	Find documents based on keywords
Information Extraction	Identify and extract personal name, date, company name, city..
Language generation	Description based on a photograph Title for a photograph
Text clustering	Automatic grouping of documents
Text classification	Assigning predefined categorization to documents Identify Spam emails and move them to a Spam folder
Machine Translation	Translate any language Text to another (when one language is unknown)
Grammar checkers	Check the grammar for any language

NLP IN EVERYDAY LIFE

- ▶ Sentiment analysis
- ▶ Search engines and question answering
- ▶ Content and news curation
- ▶ Machine translation
- ▶ Conversational assistants and chatbots (ChatGPT, Claude, Gemini)
- ▶ Speech-to-text: transcription of audio/video
- ▶ Code generation and writing assistance

Ability to harness information from a large corpus of text with no human intervention