CS565: Intelligent Systems and Interfaces

NLP: An Introduction

Semester: Jan – May 2019

Ashish Anand

Associate Professor, Dept of CSE

IIT Guwahati

Administrative Information

- Course Website: https://aaiitggrp.github.io/2019cs565/

- Accept Canvas Invitation
  - Lecture Slides will be made available here
  - Assignment and Project Submissions
  - Online Interaction and email exchange

- Marks Distribution
  - Assignment – 10
  - Scribe – 10 [3 students independently prepare the notes]
  - Project – 40
  - Exams – 15 + 25

Objective

- Define NLP
- Discuss two school of thoughts
- Understand why NLP is hard
  - Ambiguity at multiple levels
  - Different levels of NLP

What do we mean by NLP/Computational Linguistics ?

- Natural Language – Written or spoken language used by humans. Example: Sanskrit, Hindi, English, German, …
- NLP – Computational methods to learn, understand & generate natural language content.
- Distinct fields study human language
  - Linguists, Speech Recognition, Computational Linguists, Computational psycholinguistics

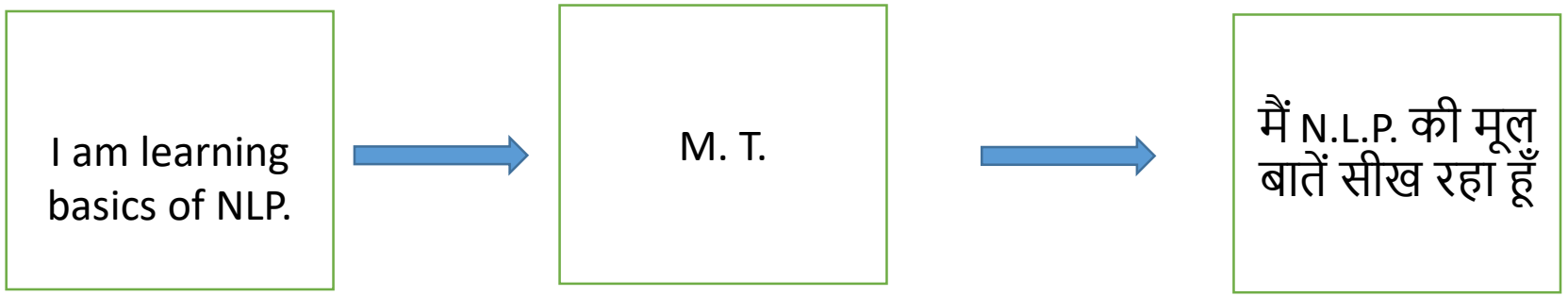
Three broad sub-areas

- Cognition
  - How do we acquire, comprehend and generate language ?
  - Good resource: http://www.mit.edu/~rplevy/teaching.html [Dept of Brain and Cognitive Sciences, MIT, USA]

- Natural Language Understanding [NLU]
  - Multiple layers

- Natural Language Generation [NLG]
  - Interlinked with NLU
  - Examples: MT, Abstractive Summarization, Chatbots/Conversational Agents

NLU and NLG



Natural Language Understanding

Natural Language Generation

Two broad paradigms to work with languages

- Rationalist
- Empiricist

Rationalist: Our brain is hardwired

- Primary objective: describe the language models of human mind (I-Language)
- Innate Language Faculty [Noam Chomsky]
- Significant part of knowledge is pre-coded

Rationalist: In Practice

- Focuses on
  - Rule based system and defining grammar
- Initial AI systems mimicked innate language faculty by trying to hardcode a lot of starting knowledge and reasoning mechanism
- Models: State Machines, Formal rule systems (Regular Grammar/CFG), Logic

Empiricist: Sense and experience in tandem with generic cognitive ability

- Primary objective: describe the language as it actually occurs (E-Language)

- Differs with rationalist in degree of belief about nature of precoded knowledge
  - Does assume generic ability of association, pattern recognition and generalization
  - Generic ability works in tandem with rich sensory inputs

Empiricist: In Practice

- Focuses on
  - Large collection of text and data-driven approaches
- Explores and uses common patterns in language use
- Appropriate Probabilistic, Statistical, Pattern-recognition and ML Models
  - Objective is to tune model parameters to learn the complicated and extensive language structure
  - We will see plenty of them during the course

Why NLP is Hard?

"WHAT IS YOUR LITTLE BROTHER CRYING ABOUT?"
"OH, 'IM—'E'S A REG'LAR COMP'TATIONAL LINGUIST, 'E IS."

http://specgram.com/CLIII.4/08.phlogiston.cartoon.zhe.html

Language is ambiguous

Example:

I made her duck

Time flies like an arrow.

- What is your inference of the two sentences?
- Whether all of them are meaningful/grammatically correct ?

Language is ambiguous

Examples: *I made her duck*

- Interpretations :
  - *I cooked duck for her*
  - *I cooked duck belonging to her*
  - *I caused her to quickly lower her body*

More examples of ambiguity

- Anne Hathaway vs. Warren Buffett's Berkshire Hathaway stock
  - When *Bride Wars* opened the stock rose 2.61%.

  [source: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1162/handouts/cs224n-lecture1.pdf]

- *Every Indian has a mother* vs. *Every Indian has a prime minister*

- *We gave the monkeys the bananas because they were hungry* vs. *We gave the monkeys the bananas because they were over-ripe*

Even more examples of ambiguity

- address, number
  - Pronunciation
- Fly, rent, tape
  - Part of speech
- ball, board, plant
  - Meaning

Types of Ambiguity

- Phonetic
  - My finger got number
- Morphological
  - Impossible vs important
  - Ram is quite impossible/ Ram is quite important
- Part of speech
  - Geeta won the first round
- Syntactic
  - Call Ram a taxi

Types of Ambiguity

- Pp attachment
  - The children ate the cake with a spoon.
- Cc attachment
  - Ram likes ripe apples and pears
- Sense
  - Ram took the bar exam
- Referential
  - Ram yelled at Shyam. He was angry at him
- Metonymy
  - Sydney called and left a message for Ram

Some other sources of difficulties

- Non-standard, slang, novel and short words
  - A360, +1-646-555-2223
  - Selfie, chillax
- Inconsistencies
  - junior college, college junior
- Parsing problems
  - Cup holder
- Metaphors, Humors, Sarcasm

Summary: why NLP is hard?

- Highly ambiguous at all levels
- Context is important to convey meaning
- Involves reasoning about the world

Different Levels of NLP

- Word
  - Phonetics and Phonology: study of linguistic sounds
  - Morphology: study of meaningful components of words [example]
- Syntax: structural relationship between words
- Semantic: study of meaning
  - Lexical semantics: study of meanings of words
  - Compositional semantics: How to combine words
- Pragmatics and Discourse: dealing with more than a sentence: paragraph, documents

References

- Chapter 1 [FSNLP]
- Chapter 1 [SLP – 2nd Ed.]
- Advances in natural language processing, J Hirschberg and C D Manning, Science 349 (6245), 261-266, 2015.