Introduction to Natural Language Processing

COMP61332: Text Mining
Week 1
Riza Batista-Navarro

Natural Language Processing

A research field focussed on creating software systems with knowledge about **natural (human) language**

Interdisciplinary: makes use of theories from **Linguistics**, adopts an **Engineering** approach

Aimed at **human-like understanding of language** (but not yet there)

Contributing disciplines

Linguistics: formal models of language, linguistic knowledge

Computer Science: representations, efficient processing, state machines, parsing algorithms, probabilistic models, dynamic programming, machine learning

Mathematics: formal automata theory, computational modelling

Psychology: psychologically plausible modelling of language use

Is NLP difficult? Why?

Natural language (NL) is full of **ambiguity**

- having more than one possible interpretation
- humans not always aware of it, but machines need to deal with it

Many types of ambiguity:

- Phonological
- Lexical
- Syntactic
- Semantic

Source:
https://medium.com/@nstjohn10/Turning-ambiguity-into-action-a-framework-52043fbedd6a

Why is NLP difficult?

Phonological

multiple interpretations due to **how it sounds** (important in speech processing)

e.g., *"I will be [writing|riding] this weekend."* (writing a piece of text or horseback riding?)

Lexical

multiple interpretations due to **a word having multiple senses**

e.g., *"I am going to the bank."* (financial entity or river?)

Why is NLP difficult?

Source: *https://americanenglish.state.gov/*

Syntactic

- due to **a word having more than one possible part of speech**

  e.g., *"I saw her duck."* (animal [noun] or bend down [verb]?)

- or, due to **prepositional phrase attachment**

  e.g., *"I saw the man on the hill with the telescope."* (who has the telescope?)

Why is NLP difficult?

Source: https://infinitewellbeing.co.uk/

Semantic

multiple possible interpretations unless knowledge of the world is available

e.g., *"The children ate the cookies because they were very hungry."* (What does "they" refer to?)

- to a human: "children" (obviously)
- to a machine: "children" or "cookies" (unless it knows that cookies cannot feel hunger, or that eating usually follows being hungry)

Two major approaches to NLP

Symbolic

- Rule- and dictionary-based systems
- Captures linguistic knowledge in rules written by experts

Statistical/Machine learning-based

- Data-driven
- Use of large amounts of (labelled) textual data to train systems, discover patterns

Hybrid approaches

Comparison

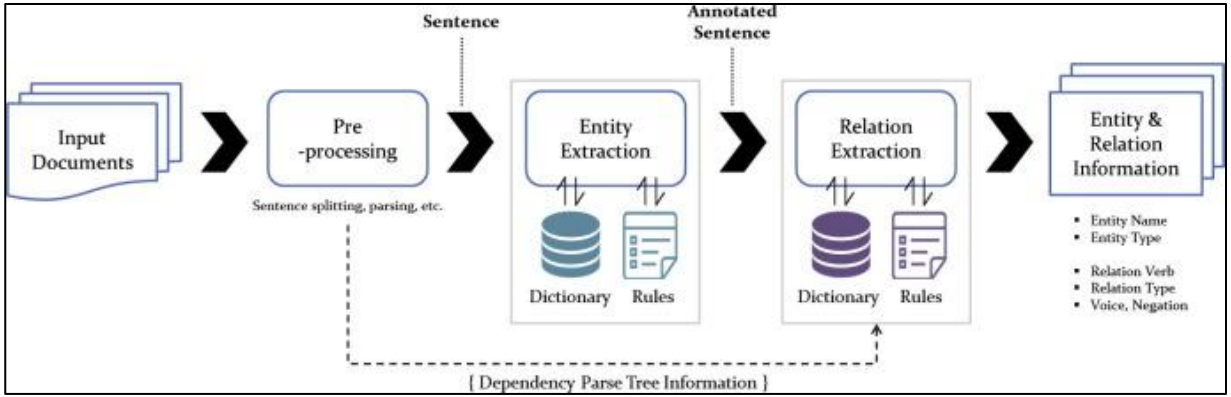
Symbolic	Statistical/Machine learning
✓ Expert knowledge yields highly precise results	✓ Can generalise well on unseen examples
✕ Shortage of experts	✕ Need people for labelling
✕ Laborious rule writing, dictionary preparation	✕ Time consuming and laborious labelling
✕ Domain adaptation problematic	✕ Must retrain for new domain
✓ Results can be interpreted	✕ Often cannot inspect/change models
✓ Good when labelled data is hard to obtain	✓ Good where dictionaries are unavailable

NLP Pipelines

A ‘complete’ NLP system is usually a pipeline of components

Each component tackles a specific problem, e.g.,

- Sentence segmentation
- Tokenisation
- Part-of-speech tagging
- Parsing
- Information extraction



Source: https://www.sciencedirect.com/science/article/pii/S1532046415001756

NLP and Text Mining

Natural language processing

- considers language structure, linguistics
- requires knowledge of grammar

Text mining

- not necessarily looking at grammar
- possibly relying only on frequencies to find correlations

NLP-based Text Mining

- TM that is driven by NLP