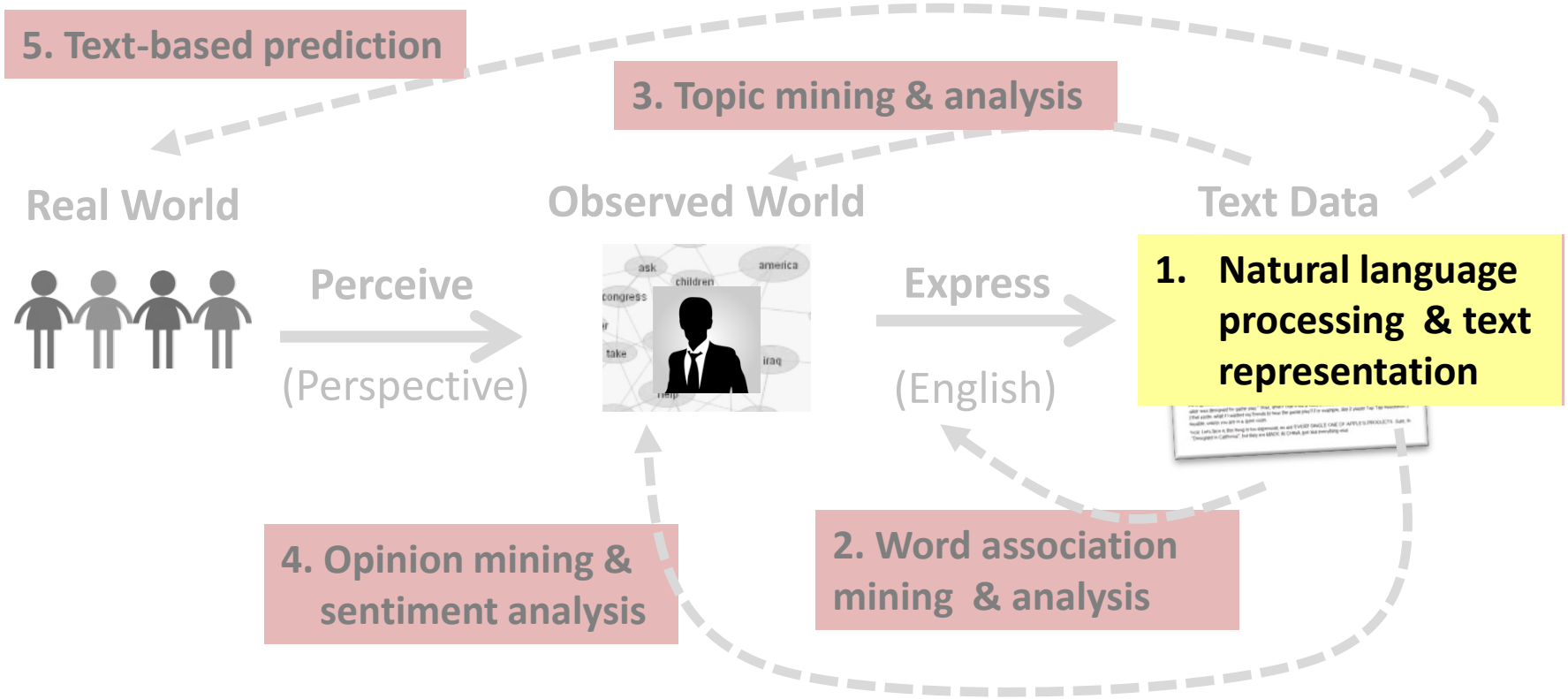
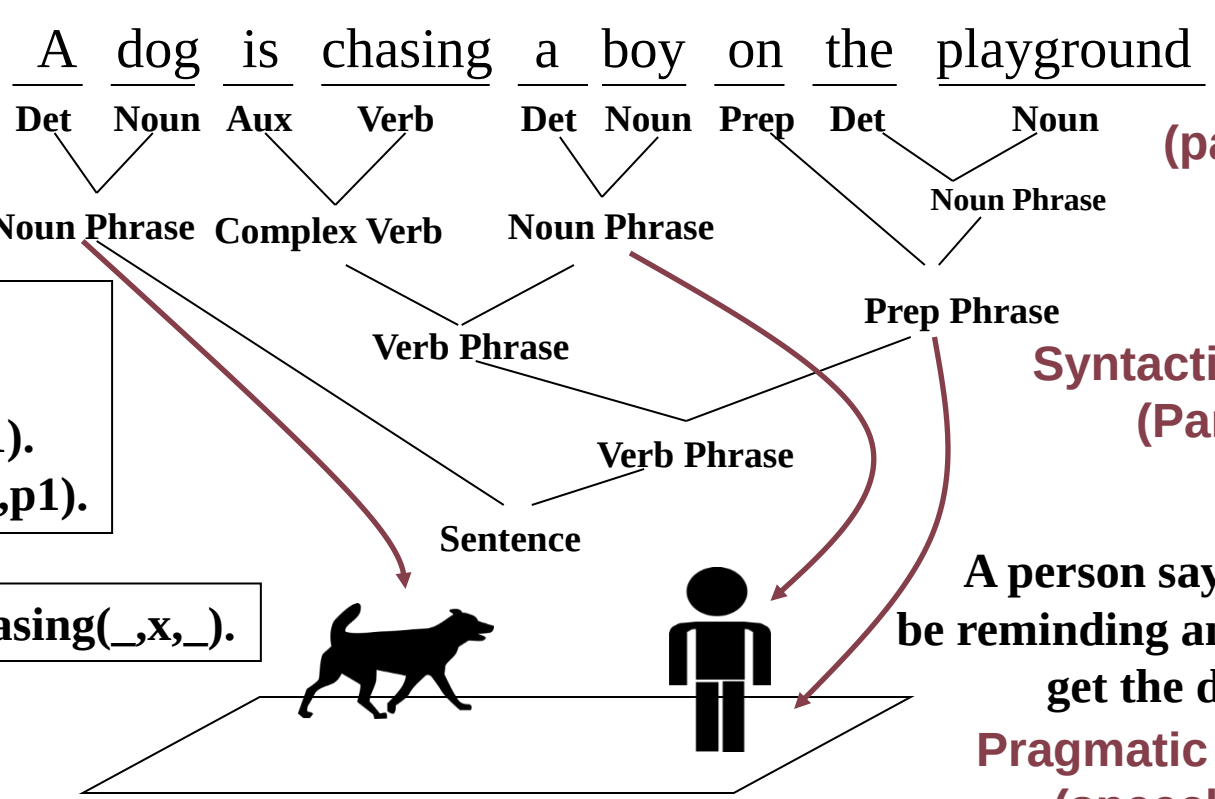
Natural Language Content Analysis

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Natural Language Content Analysis



Basic Concepts in NLP



Lexical analysis (part-of-speech tagging)

Syntactic analysis (Parsing)

Semantic analysis

Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).

+

Scared(x) if Chasing(_,x,_).

Scared(b1)
Inference

A person saying this may be reminding another person to get the dog back.

Pragmatic analysis (speech act)

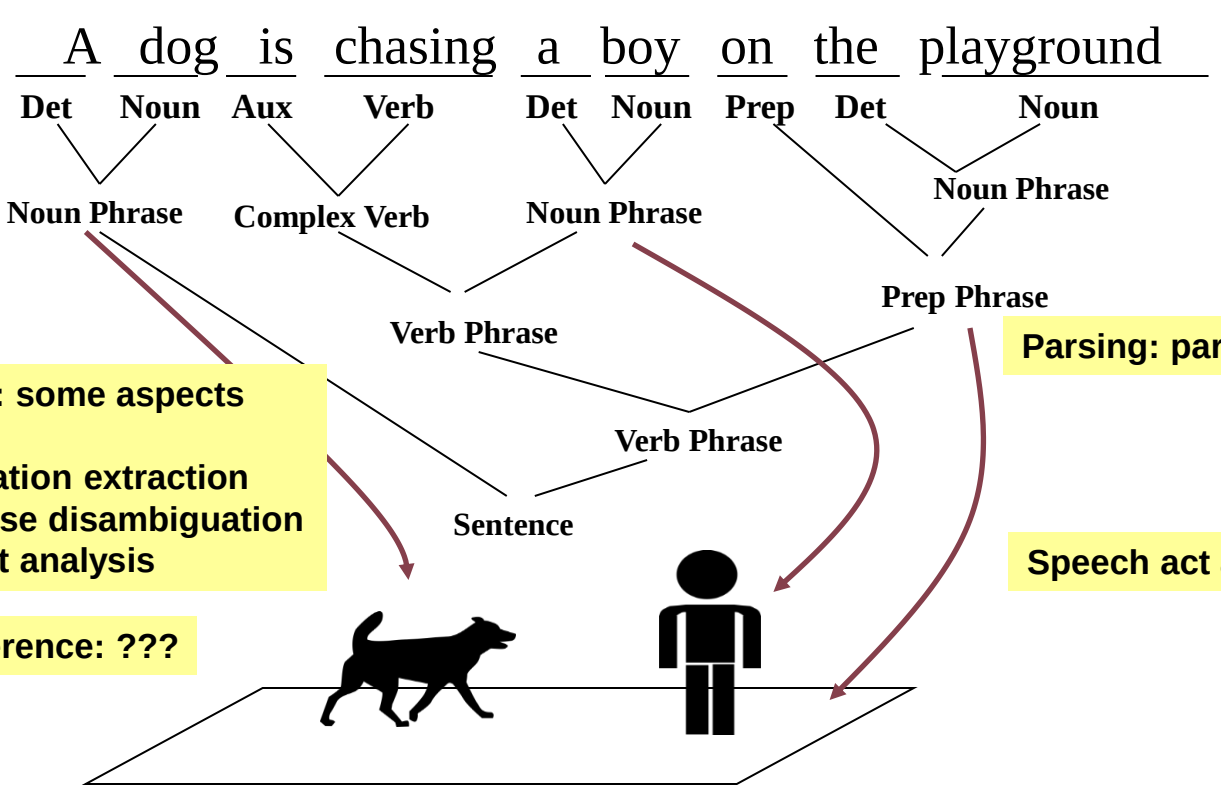
NLP Is Difficult!

- Natural language is designed to make human communication efficient. As a result,
  - we omit a lot of *common sense* knowledge, which we assume the hearer/reader possesses.
  - we keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve.
- This makes EVERY step in NLP hard
  - Ambiguity is a *killer*!
  - Common sense reasoning is pre-required.

Examples of Challenges

- Word-level ambiguity:
  - "design" can be a noun or a verb (ambiguous POS)
  - "root" has multiple meanings (ambiguous sense)
- Syntactic ambiguity:
  - "natural language processing" (modification)
  - "A man saw a boy *with a telescope*." (PP Attachment)
- Anaphora resolution: "John persuaded Bill to buy a TV for *himself*." (himself = John or Bill?)
- Presupposition: "He has quit smoking" implies that he smoked before.

The State of the Art



POS Tagging: 97%

Parsing: partial >90%(?)

Semantics: some aspects

- Entity/relation extraction
- Word sense disambiguation
- Sentiment analysis

Speech act analysis: ???

Inference: ???

What We Can't Do

- 100% POS tagging
  - "He turned off the highway." vs "He turned off the fan."
- General complete parsing
  - "A man saw a boy with a telescope."
- Precise deep semantic analysis
  - Will we ever be able to precisely define the meaning of "own" in "John owns a restaurant"?

Robust and general NLP tends to be *shallow* while *deep* understanding doesn't scale up.

Summary

- NLP is the foundation for text mining
- Computers are far from being able to understand natural language
  - Deep NLP requires common sense knowledge and inferences, thus only working for very limited domains
  - Shallow NLP based on statistical methods can be done in large scale and is thus more broadly applicable
- In practice: statistical NLP as the basis, while humans provide help as needed

Additional Reading

Manning, Chris and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press, 1999.