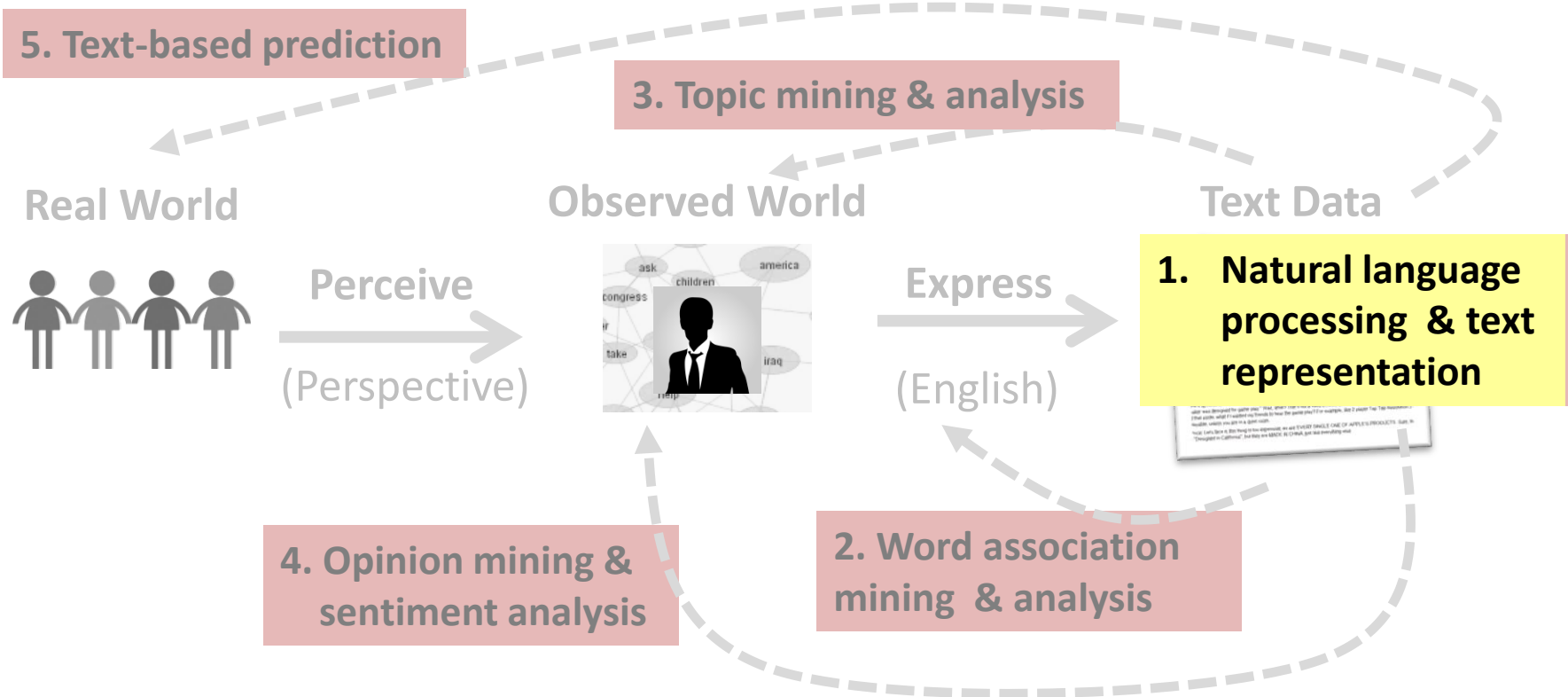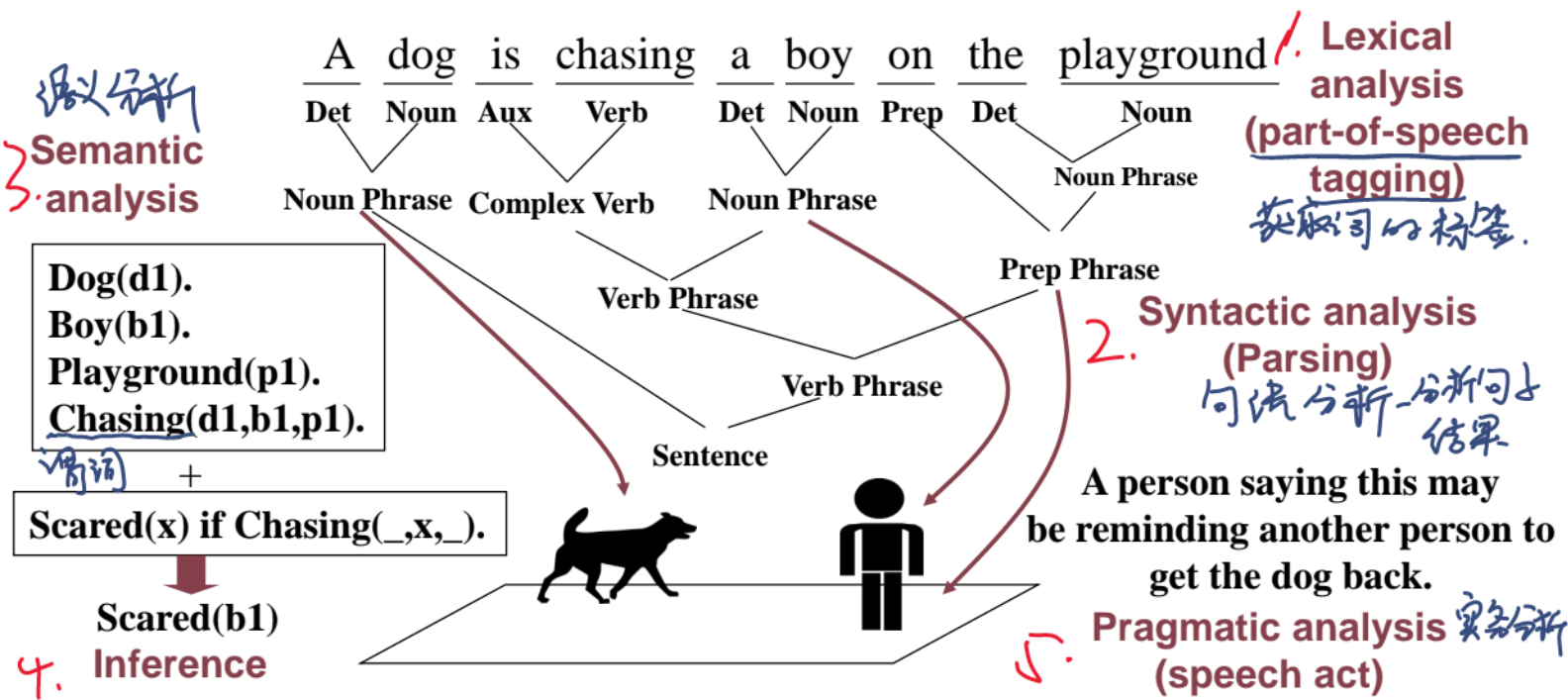# Natural Language Content Analysis

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

# Natural Language Content Analysis

**5. Text-based prediction**

**3. Topic mining & analysis**

Real World

Observed World

Text Data

**Perceive**

(Perspective)

**Express**

(English)

**1. Natural language processing & text representation**

**4. Opinion mining & sentiment analysis**

**2. Word association mining & analysis**

# Basic Concepts in NLP

A dog is chasing a boy on the playground

Det Noun Aux Verb Det Noun Prep Det Noun

Noun Phrase  Complex Verb  Noun Phrase  Noun Phrase

Prep Phrase

Verb Phrase

Verb Phrase

Sentence

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

+

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

Scared(b1)
Inference

语义分析
Semantic analysis

谓词

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

Pragmatic analysis
(speech act)
实际分析

Lexical analysis (part-of-speech tagging)
英顾词的标签.

Syntactic analysis (Parsing)
句法分析—分析句子结果

3

# 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*!
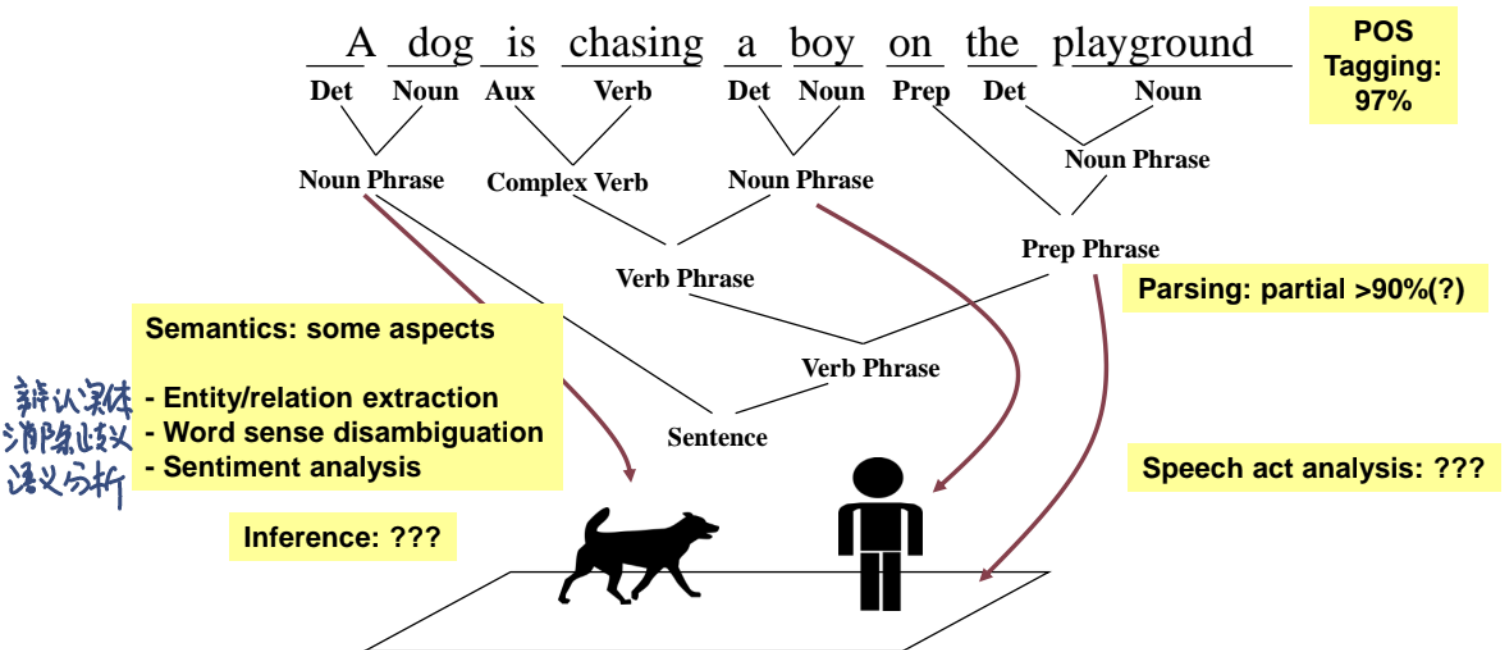  - Common sense reasoning is pre-required.

difficulties

# 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

A  dog  is  chasing  a  boy  on  the  playground

Det  Noun  Aux  Verb  Det  Noun  Prep  Det  Noun

**POS Tagging: 97%**

Noun Phrase  Complex Verb  Noun Phrase  Noun Phrase

Verb Phrase

Prep Phrase

**Parsing: partial >90%(?)**

**Semantics: some aspects**

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

Verb Phrase

Sentence

辨认实体
消除歧义
语义分析

**Inference: ???**

**Speech act analysis: ???**

6

# What We Can't Do

词性标注 · 100% POS tagging
- "He turned <u>off</u> the highway." vs "He turned <u>off</u> the fan."

歧义 · General complete parsing
- "A man saw a boy with a telescope."

深刻的
语义分析 · 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 要靠人工参与 进行训练

# Additional Reading

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