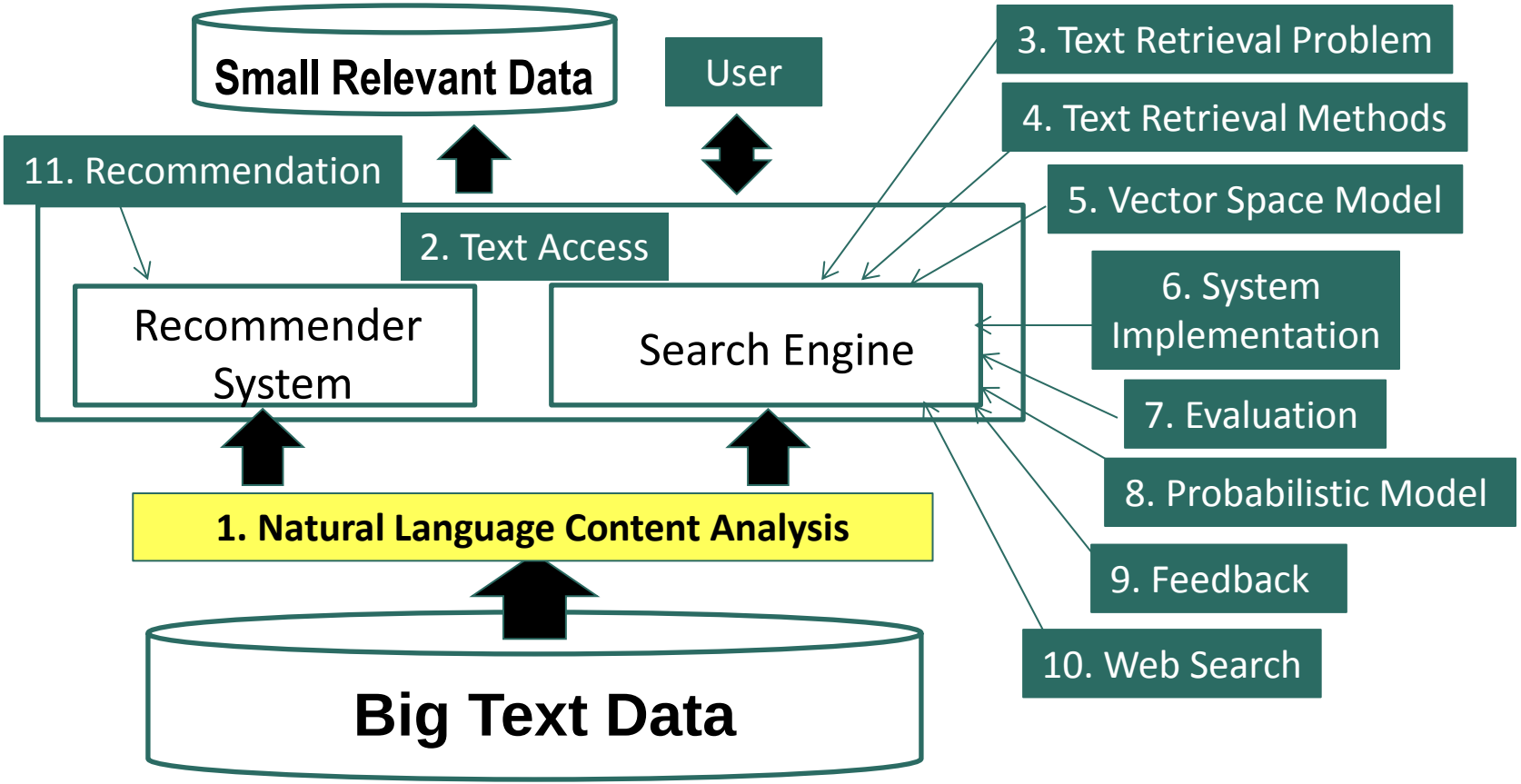
Text Retrieval and Search Engines

Natural Language Content Analysis

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

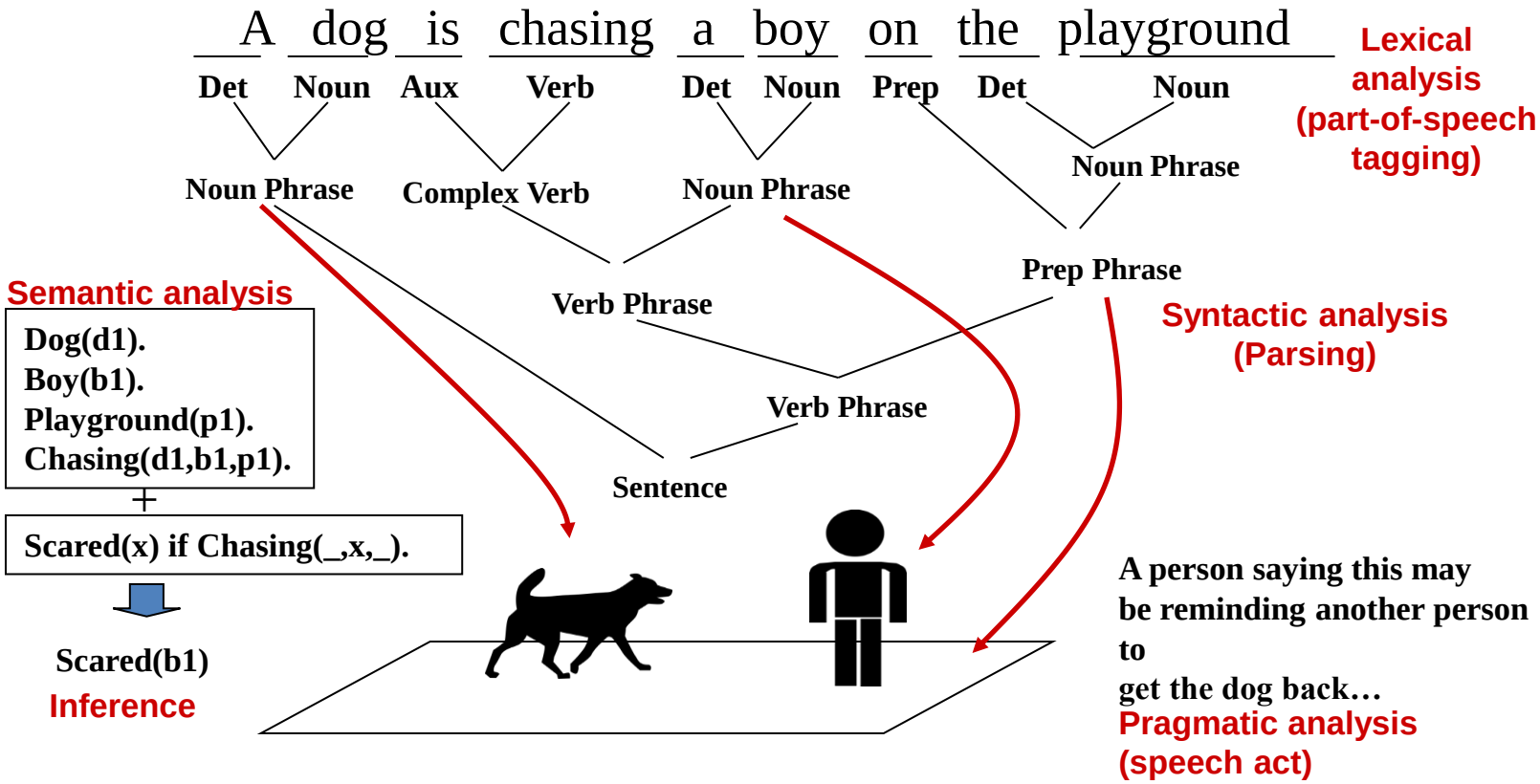
Course Schedule



Overview

- What is Natural Language Processing (NLP)?
- State of the Art in NLP
- NLP for Text Retrieval

An Example of NLP



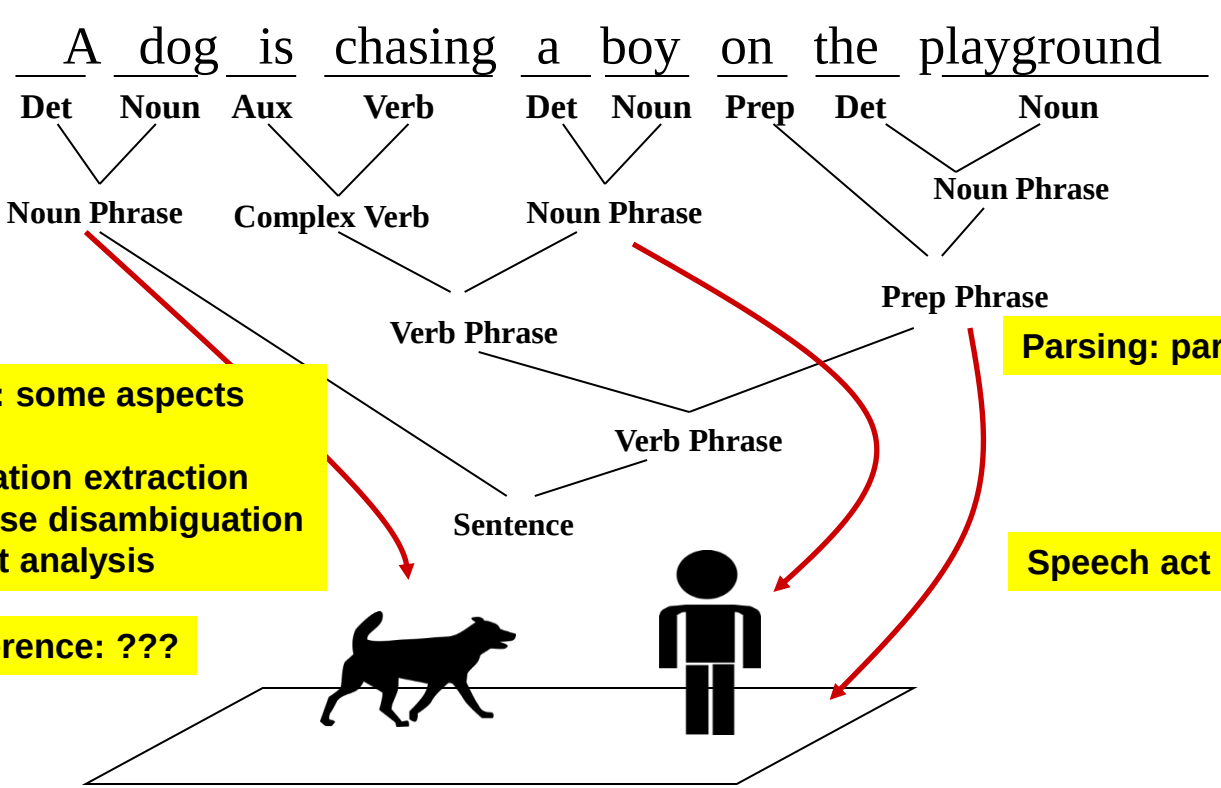
NLP Is Difficult!

- 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: E.g.,
  - "design" can be a noun or a verb (Ambiguous POS)
  - "root" has multiple meanings (Ambiguous sense)
- Syntactic ambiguity: E.g.,
  - "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 & general NLP tends to be "shallow" while "deep" understanding doesn't scale up

NLP for Text Retrieval

- Must be general robust & efficient ➔ Shallow NLP
- **“Bag of words” representation tends to be sufficient** for most search tasks (but not all!)
- Some text retrieval techniques can naturally address NLP problems
- However, deeper NLP is needed for complex search tasks

Summary

- What is Natural Language Processing (NLP)?
- State of the Art in NLP
- NLP for Text Retrieval

Additional Reading

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