Text Representation

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

Text Representation

5. Text-based prediction

3. Topic mining & analysis

Real World

Observed World

Text Data

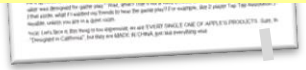
Perceive

(Perspective)

Express

(English)

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



4. Opinion mining & sentiment analysis

2. Word association mining & analysis

A dog is chasing a boy on the playground

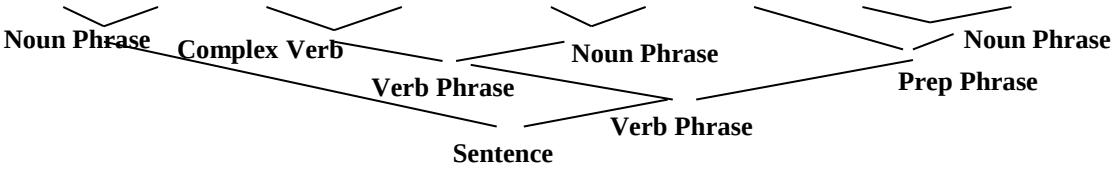
String of characters

A dog is chasing a boy on the playground

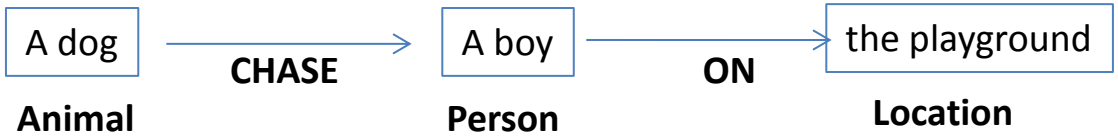
Sequence of words

Det Noun Aux Verb Det Noun Prep Det Noun

+ POS tags



+ Syntactic structures



+ Entities and relations

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

+ Logic predicates

Speech Act = REQUEST

+ Speech acts

Deeper NLP: requires more human effort; less accurate

Closer to knowledge representation

Text Representation and Enabled Analysis

This course

Text Rep	Generality	Enabled Analysis	Examples of Application
String		String processing	Compression
Words		Word relation analysis; topic analysis; sentiment analysis	Thesaurus discovery; topic and opinion related applications
+ Syntactic structures		Syntactic graph analysis	Stylistic analysis; structure-based feature extraction
+ Entities & relations		Knowledge graph analysis; information network analysis	Discovery of knowledge and opinions about specific entities
+ Logic predicates		Integrative analysis of scattered knowledge; logic inference	Knowledge assistant for biologists

Summary

- Text representation determines what kind of mining algorithms can be applied
- **Multiple ways** of representing text are possible
  - string, words, syntactic structures, entity-relation graphs, predicates…
  - can/should be **combined** in real applications
- This course focuses on **word-based representation**
  - **General and robust**: applicable to any natural language
  - **No/little manual effort**
  - **"Surprisingly" powerful** for many applications (not all!)
  - **Can be combined** with more sophisticated representations