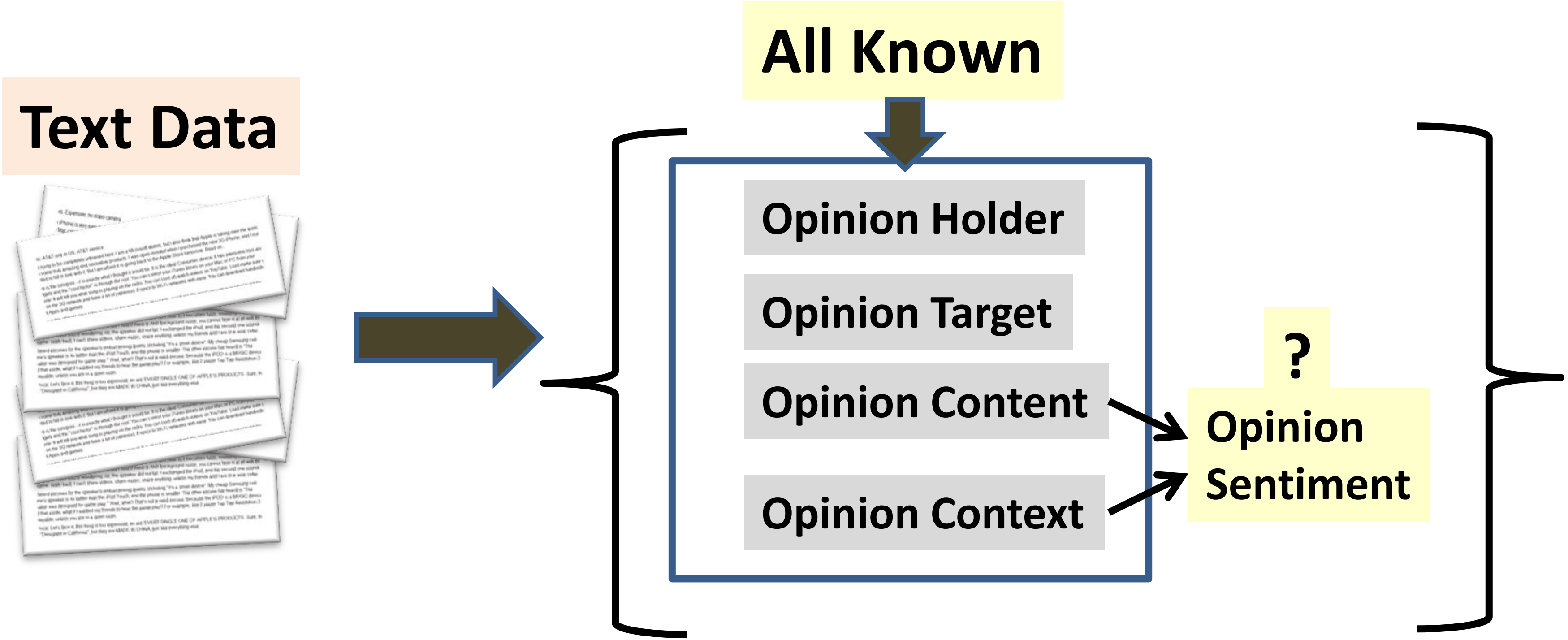
Opinion Mining and Sentiment Analysis: Sentiment Classification

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

Sentiment Classification



Text Data

All Known

Opinion Holder

Opinion Target

Opinion Content

Opinion Context

?

Opinion Sentiment

Sentiment Classification: Task Definition

- Input: An opinionated text object
- Output: A sentiment tag/label
  - Polarity analysis: e.g., categories = {positive, negative, neutral}, or categories ={5, 4, 3, 2, 1}
  - Emotion analysis (beyond polarity): e.g., categories ={happy, sad, fearful, angry, surprised, disgusted}
- A special case of text categorization! ➔ Any text categorization method can be used to do sentiment classification
- Further improvement comes from
  - More sophisticated features appropriate for sentiment tagging
  - Consideration of the order of the categories (e.g., ordinal regression)

Commonly Used Text Features

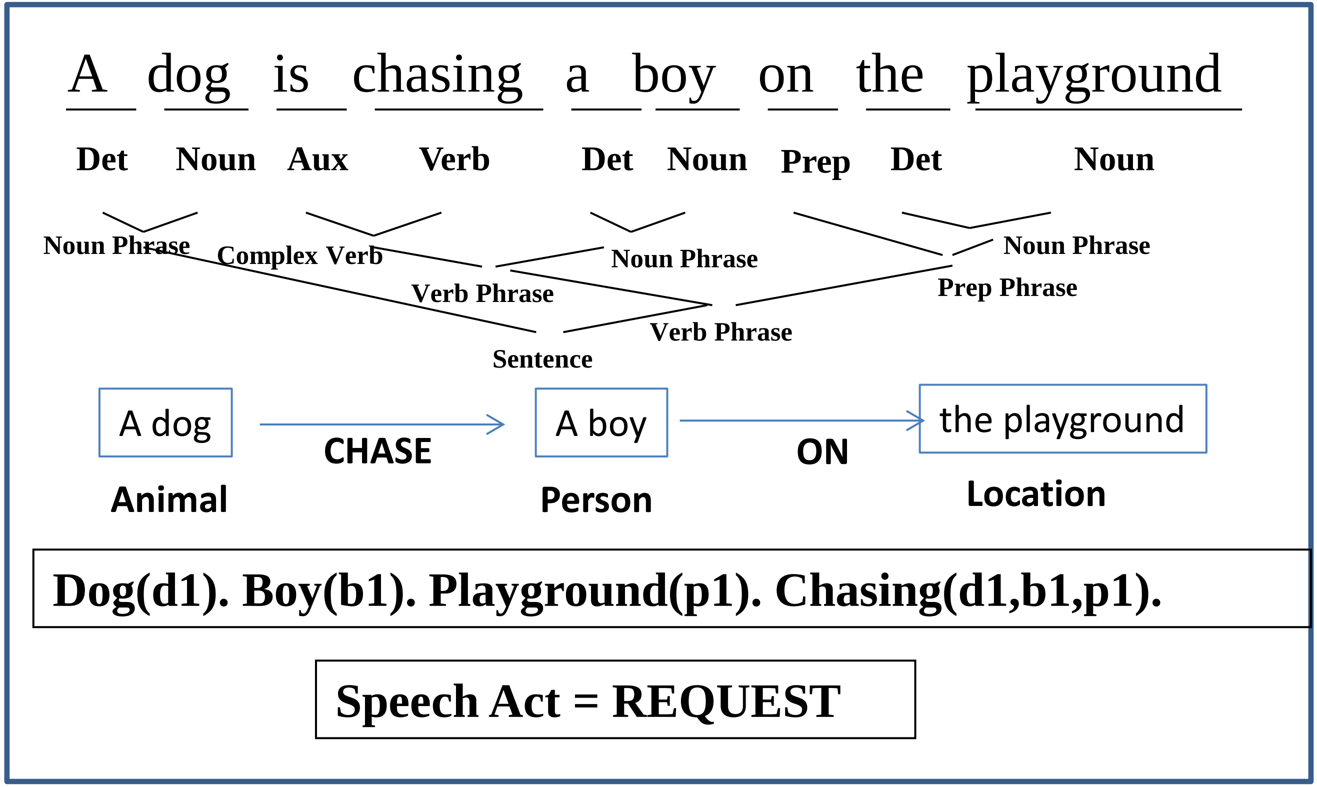
- Character n-grams: can be mixed with different n’s
  - General and robust to spelling/recognition errors, but less discriminative than words
- Word n-grams: can be mixed with different n’s
  - Unigrams are often very effective, but not for sentiment analysis (e.g. , “it’s not good” or “it’s not as good as”)
  - Long n-grams are discriminative, but may cause overfitting
- POS tag n-grams: mixed n-gram with words and POS tags
  - E.g., “ADJECTIVE NOUN” or “great NOUN”

Commonly Used Text Features (cont.)

- Word classes
  - Syntactic (= POS tags)
  - Semantic Concept: e.g., thesaurus/ontology, recognized entities
  - Empirical word clusters (e.g., cluster of paradigmatically or syntagmatically related words)
- Frequent patterns in text (e.g., frequent word set; collocations)
  - More specific/discriminative than words
  - May generalize better than pure n-grams
- Parse tree-based (e.g., frequent subtrees, paths)
  - Even more discriminative, but need to avoid overfitting
- Pattern discovery algorithms are very useful for feature construction

NLP Enriches Text Representation with Complex Features

A dog is chasing a boy on the playground



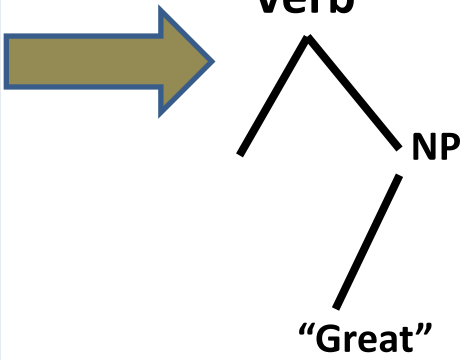
Dog(d1). Boy(b1). Playground(p1). Chasing(d1,b1,p1).

Speech Act = REQUEST

"great NOUN"

"Verb Adv Adj"

...



Feature Construction for Text Categorization

- Feature design affects categorization accuracy significantly
- A combination of machine learning, error analysis, and domain knowledge is most effective
  - Domain knowledge → seed features, feature space
  - Machine learning → feature selection, feature learning
  - Error analysis → feature validation
- NLP enriches text representation → enriches feature space (more likely overfitting!)
- Optimizing the tradeoff between **exhaustivity** and **specificity** is a major goal

high coverage (frequent)

discriminative (infrequent)