Word Association Mining and Analysis

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

Word Association Mining & Analysis

5. Text-based prediction

3. Topic mining and analysis

Real World

Observed World

Text Data

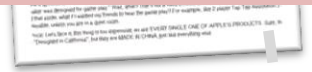
Perceive

(Perspective)

Express

(English)

1. Natural language processing and text representation



4. Opinion mining and sentiment analysis

2. Word association mining and analysis

Outline

- What is a word association?
- Why mine word associations?
- How to mine word associations?

Basic Word Relations: Paradigmatic vs. Syntagmatic

聚合关系　　组合关系

- Paradigmatic: A & B have paradigmatic relation if they can be substituted for each other (i.e., A & B are in the same class)
  - E.g., "cat" and "dog"; "Monday" and "Tuesday" 有聚合关系，名词-类词
- Syntagmatic: A & B have syntagmatic relation if they can be combined with each other (i.e., A & B are related semantically)
  - E.g., "cat" and "sit"; "car" and "drive" 有组合关系，可以组成一个句子
- These two basic and complementary relations can be generalized to describe relations of any items in a language

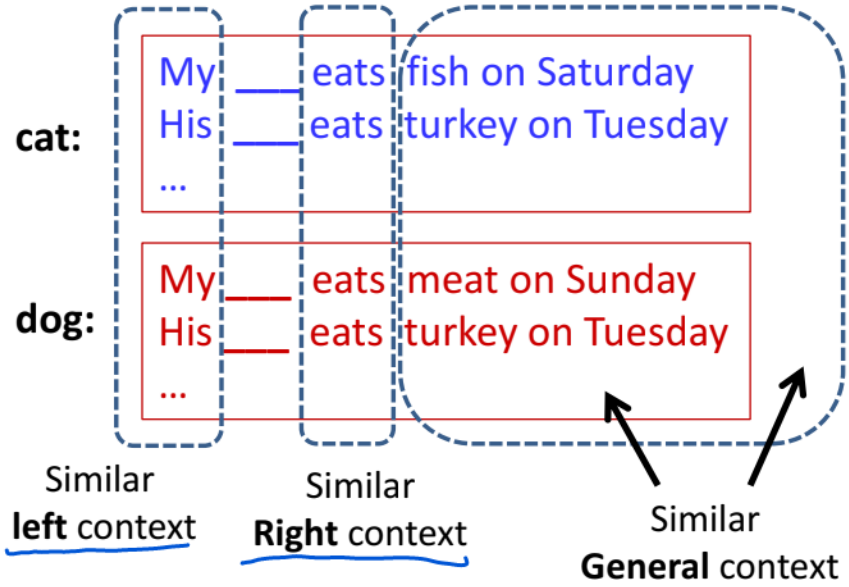
Why Mine Word Associations?

- They are useful for improving accuracy of many NLP tasks
  - POS tagging, parsing, entity recognition, acronym expansion
  - Grammar learning
- They are directly useful for many applications in text retrieval and mining
  - Text retrieval (e.g., use word associations to suggest a variation of a query) query expansion, recommand the query to user
  - Automatic construction of topic map for browsing: words as nodes and associations as edges
  - Compare and summarize opinions (e.g., what words are most strongly associated with "battery" in positive and negative reviews about iPhone 6, respectively?)

Mining Word Associations: Intuitions

Paradigmatic: similar context

My **cat** eats fish on Saturday
His **cat** eats turkey on Tuesday
My **dog** eats meat on Sunday
His **dog** eats turkey on Tuesday
…

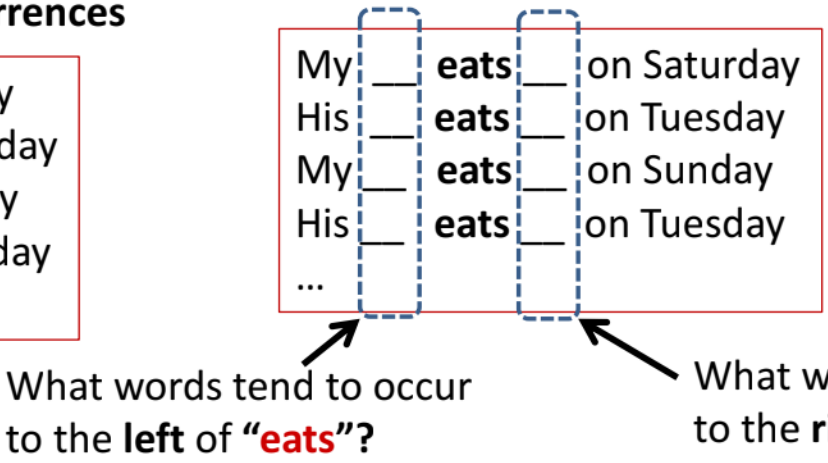


How similar are context ("cat") and context ("dog")? ✓
How similar are context ("cat") and context ("computer")? ✗

Mining Word Associations: Intuitions

Syntagmatic: correlated occurrences

My cat **eats** fish on Saturday
His cat **eats** turkey on Tuesday
My dog **eats** meat on Sunday
His dog **eats** turkey on Tuesday
…



Whenever "**eats**" occurs, what **other words** also tend to occur?
How helpful is the occurrence of "**eats**" for predicting occurrence of "**meat**"? ✓
How helpful is the occurrence of "**eats**" for predicting occurrence of "**text**"? ✗

Mining Word Associations: General Ideas

- **Paradigmatic**
  - Represent each word by its context
  - Compute context similarity
  - Words with **high context similarity** likely have paradigmatic relation
- **Syntagmatic**
  - Count how many times two words occur together in a context (e.g., sentence or paragraph)
  - Compare their co-occurrences with their individual occurrences
  - Words with **high co-occurrences but relatively low individual occurrences** likely have syntagmatic relation
- Paradigmatically related words tend to have syntagmatic relation with the same word ➔ **joint discovery** of the two relations
- These ideas can be implemented in many different ways!