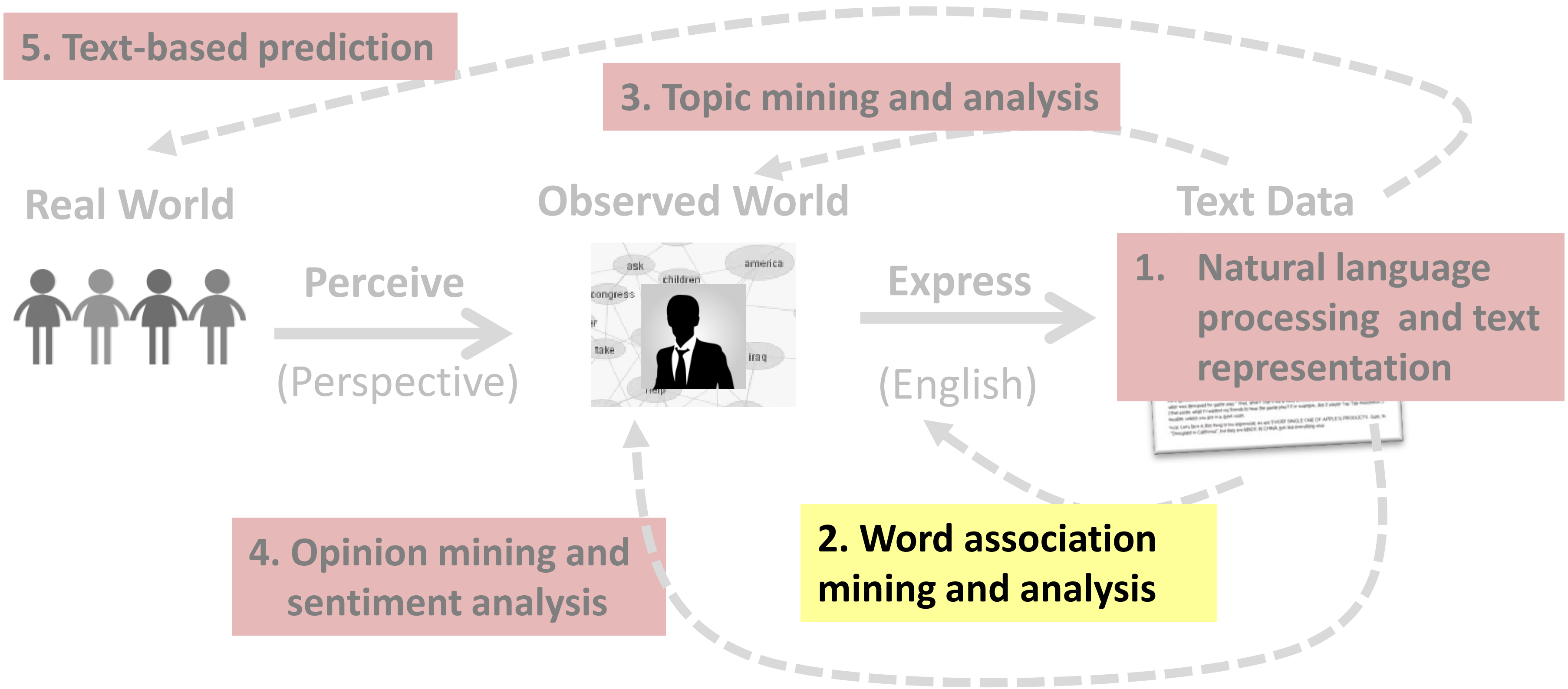
Syntagmatic Relation Discovery: Entropy

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

Syntagmatic Relation Discovery: Entropy



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

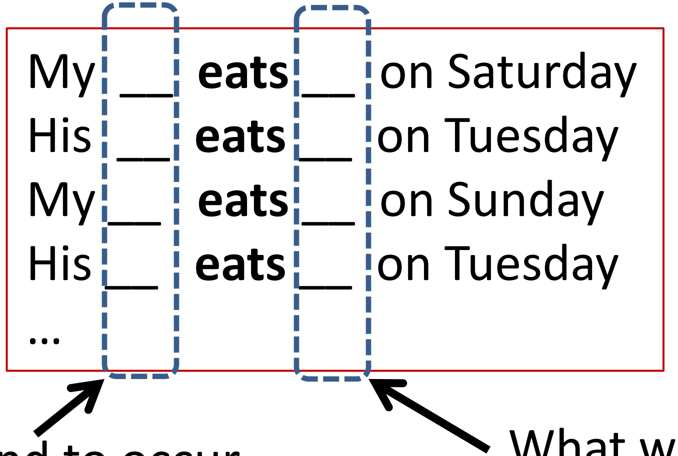
4. Opinion mining and sentiment analysis

2. Word association mining and analysis

Syntagmatic Relation = Correlated Occurrences

Whenever "**eats**" occurs, what **other words** also tend to occur?

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



My __ **eats** __ on Saturday
His __ **eats** __ on Tuesday
My __ **eats** __ on Sunday
His __ **eats** __ on Tuesday
…

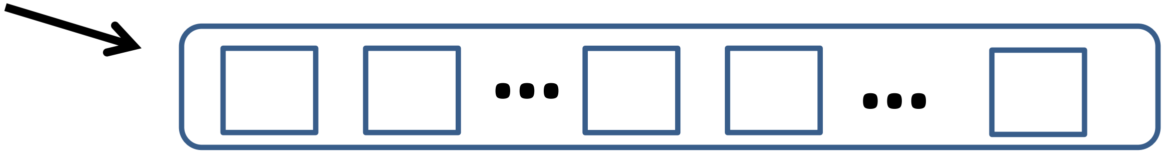
What words tend to occur to the **left** of **"eats"?**

What words are to the **right?**

Word Prediction: Intuition

Prediction Question: Is word **W** present (or absent) in this segment?

Text Segment (any unit, e.g., sentence, paragraph, document)



Are some words easier to predict than others?

1) W = "meat"   2) W="the"   3) W="unicorn"

Word Prediction: Formal Definition

Binary Random Variable :
$X_w \in \{0, 1\}$

$$X_w = \begin{cases} 1 & \text{w is present} \\ 0 & \text{w is absent} \end{cases}$$

$$p(X_w = 1) + p(W_w = 0) = 1$$

The more random X_w is, the more difficult the prediction would be.

How does one quantitatively measure the “randomness” of a random variable like Xw?

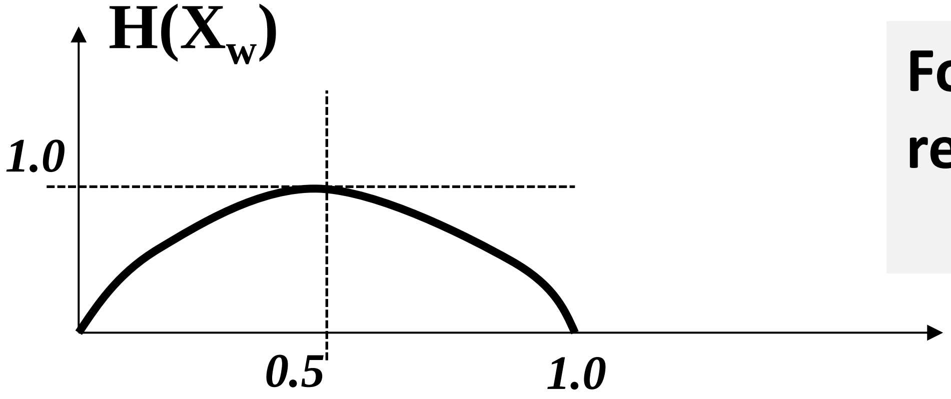
Entropy H(X) Measures Randomness of X

$$H(X_w) = \sum_{v\in\{0,1\}} -p(X_w = v)\log_2 p(X_w = v)$$

$$X_w = \begin{cases} 1 & w \text{ is present} \\ 0 & w \text{ is absent} \end{cases}$$

$$= -p(X_w = 0)\log_2 p(X_w = 0) - p(X_w = 1)\log_2 p(X_w = 1)$$

Define $0\log_2 0 = 0$



For what X_w, does $H(X_w)$ reach maximum/minimum?

E.g., $P(X_w=1)=1$? $P(X_w=1)=0.5$?

P(Xw=1)

or equivalently P(Xw=0) (Why?)

Entropy H(X): Coin Tossing

$$H(X_{coin}) = -p(X_{coin} = 0)\log_2 p(X_{coin} = 0) - p(X_{coin} = 1)\log_2 p(X_{coin} = 1)$$

X_{coin}: tossing a coin

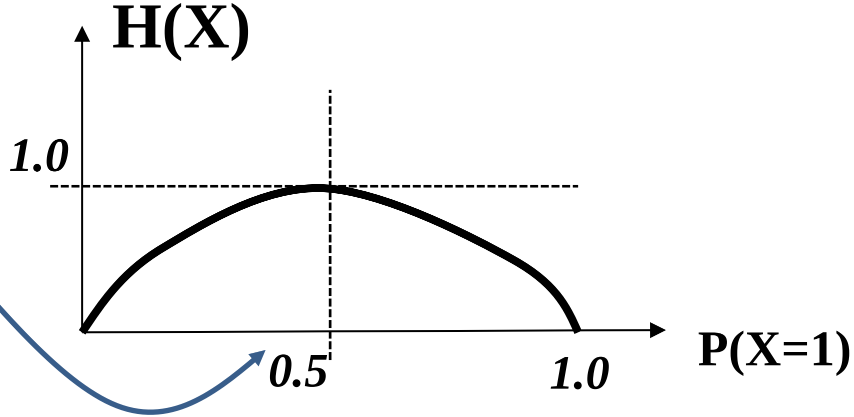
$$X_{coin} = \begin{cases} 1 & \text{Head} \\ 0 & \text{Tail} \end{cases}$$

Fair coin: p(X=1)=p(X=0)=1/2

$$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

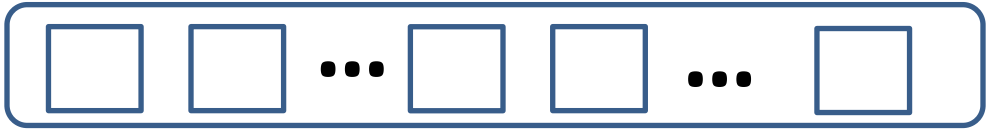
Completely biased: p(X=1)=1

$$H(X) = -0 * \log_2 0 - 1 * \log_2 1 = 0$$



Entropy for Word Prediction

Is word **W** present (or absent) in this segment?



1) W = “meat” 2) W = “the” 3) W = “unicorn”

Which is **high/low**? **H(X_{meat}), H(X_{the}), or H($X_{unicorn}$)?**

H(X_{the})≈0 → no uncertainty since p(X_{the}=1) ≈ 1

High entropy words are harder to predict!