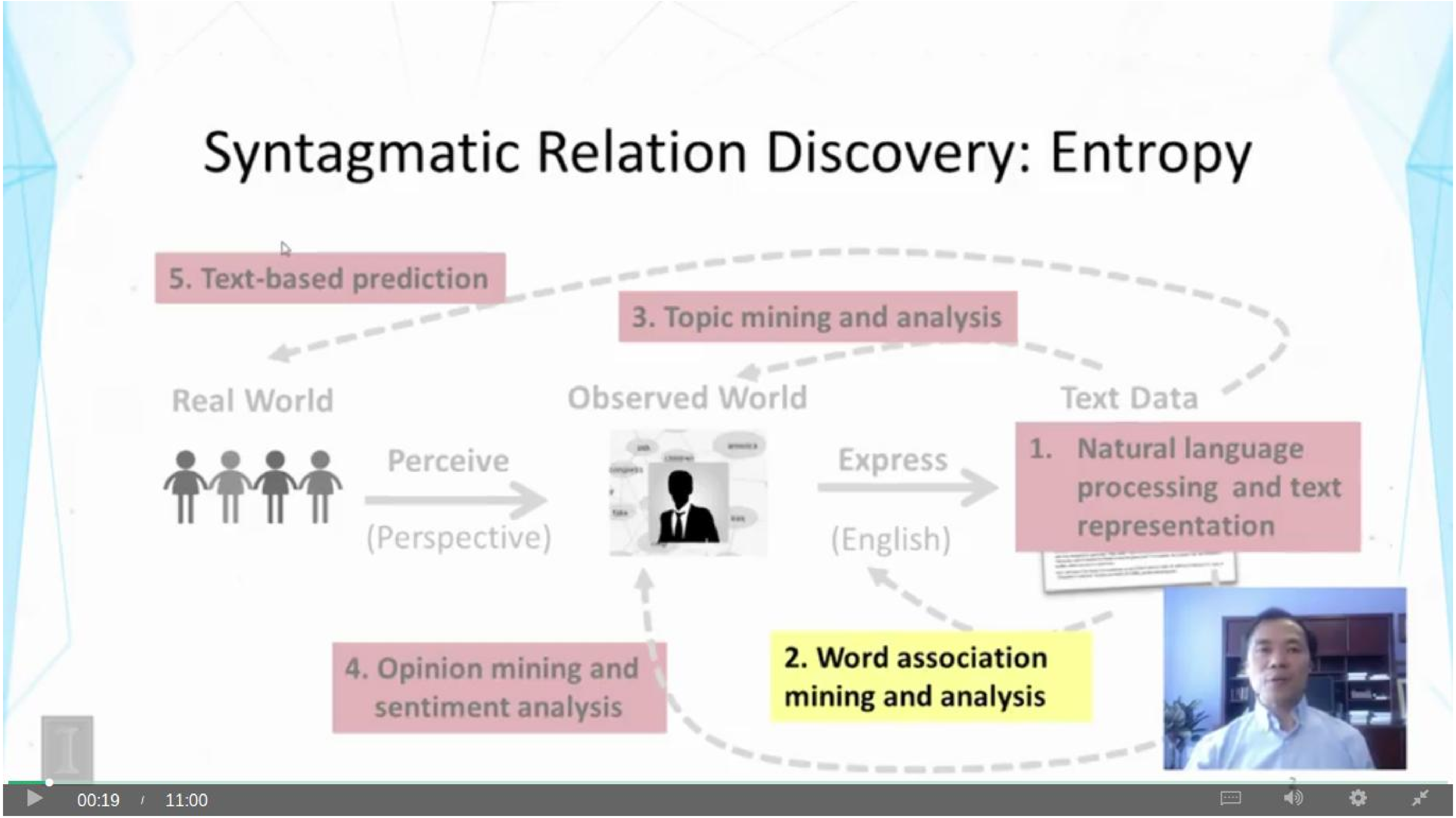
Syntagmatic Relation Discovery: Entropy

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

00:09 / 11:00


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
00:19 / 11:00

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**"?

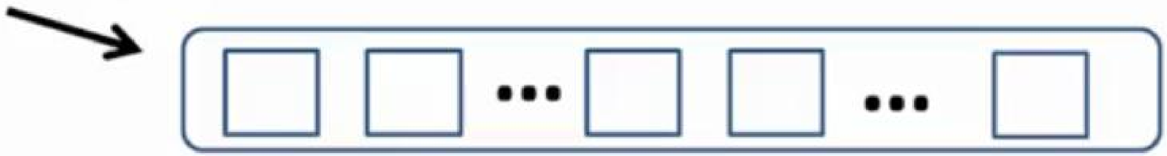
What words are to the **right?**

00:38 / 11:00

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 :
$X_w \in \{0, 1\}$

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

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

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?

03:01 / 11:00

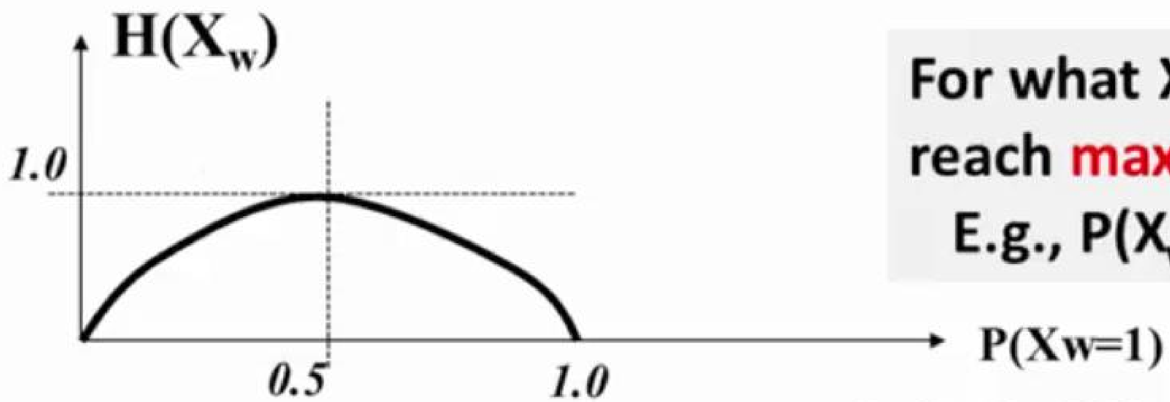
Entropy H(X) Measures Randomness of X

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

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

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

Define $0\log_2 0 = 0$



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

For what X_w, does $H(X_w)$ reach maximum/minimum?
E.g., $P(X_w=1)=1$? $P(X_w=1)=0.5$?

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

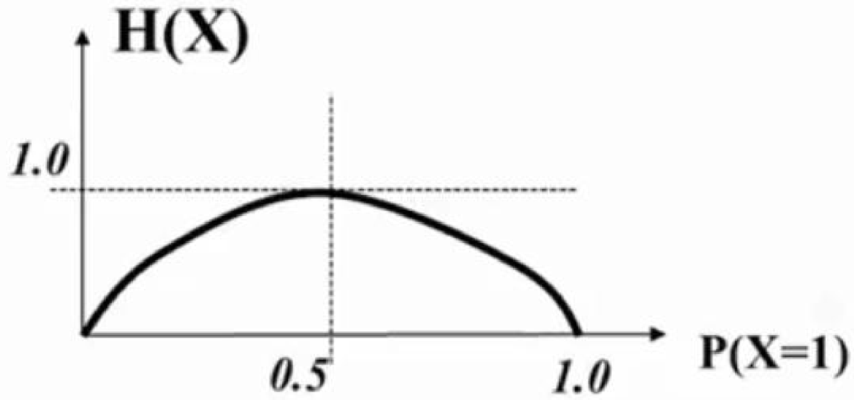
$$X_{coin} = \begin{cases} 1 & \text{Head} \\ 0 & \text{Tail} \end{cases}$$

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

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

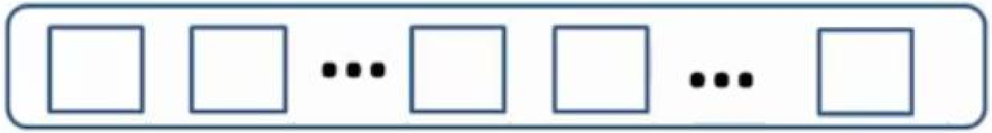
Completely biased: p(X=1)=1

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



Entropy for Word Prediction

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

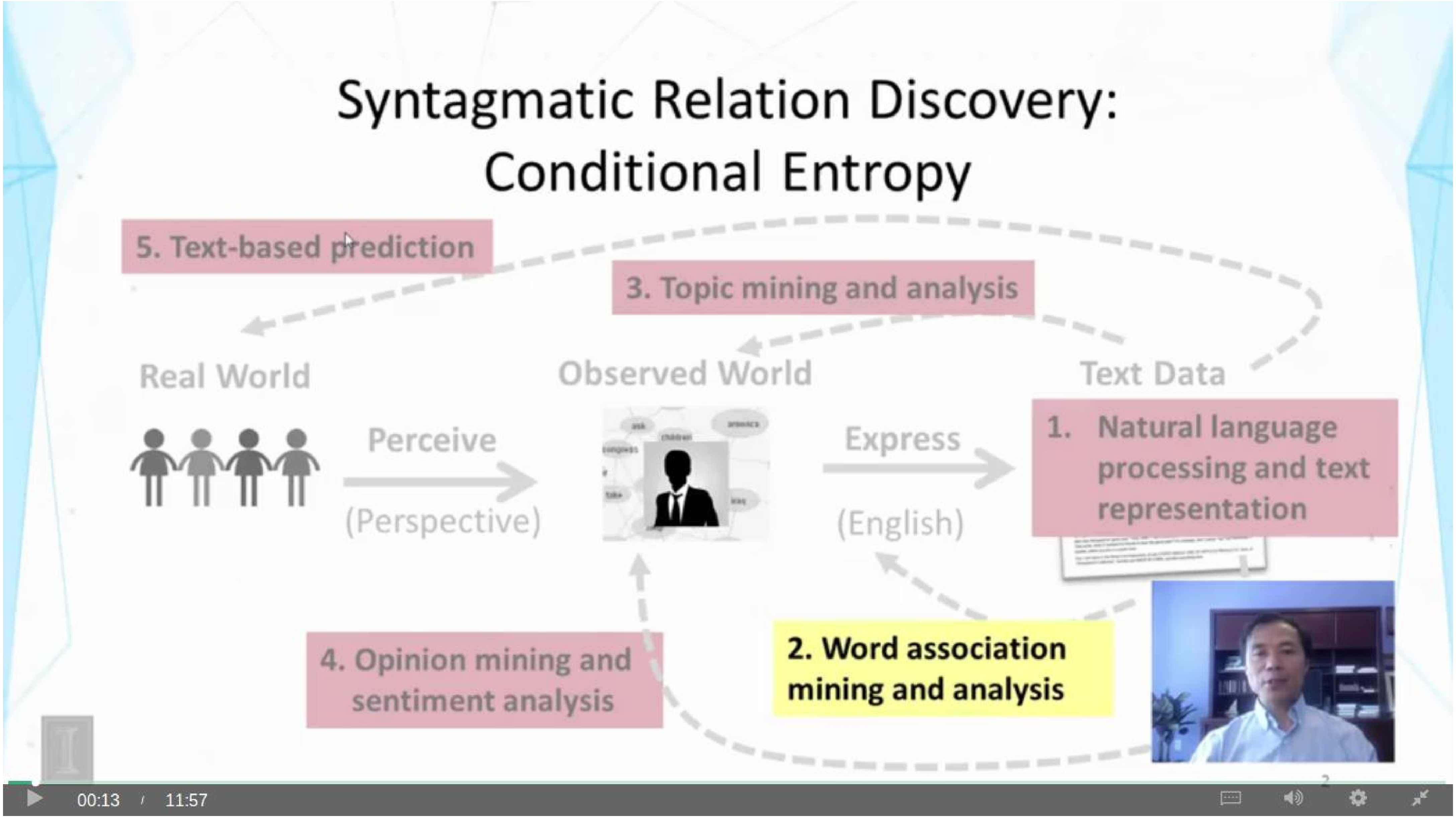


1) W = "meat" 2) W = "the" 3) W = "unicorn"

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

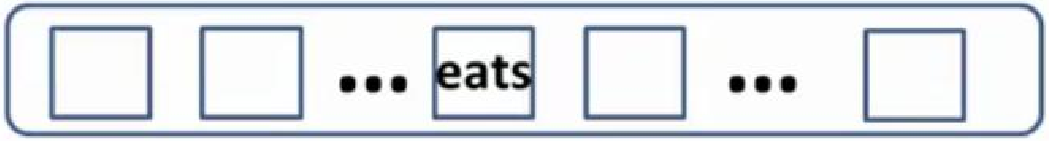
$H(X_{the}) \approx 0$ → no uncertainty since $p(X_{the}=1) \approx 1$

High entropy words are harder to predict!


Syntagmatic Relation Discovery:
Conditional 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
00:13 / 11:57

What If We Know More About a Text Segment?

Prediction question: Is "**meat**" present (or absent) in this segment?



Does presence of "**eats**" help predict the presence of "**meat**"?
Does it **reduce** the uncertainty about "meat", i.e., $\mathbf{H(X_{meat})}$?

What if we know of the absence of "eats"? Does it also help?

Conditional Entropy

Know nothing about the segment | **Know "eats" is present (X_{eats} =1)**

$$p(X_{meat}=1) \dashrightarrow p(X_{meat}=1 \mid X_{eats}=1)$$

$$p(X_{meat}=0) \dashrightarrow p(X_{meat}=0 \mid X_{eats}=1)$$

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

$$\begin{aligned} \underline{H(X_{meat} \mid X_{eats}=1)} = &-p(X_{meat}=0 \mid X_{eats}=1)\log_2 p(X_{meat}=0 \mid X_{eats}=1) \\ &- p(X_{meat}=1 \mid X_{eats}=1)\log_2 p(X_{meat}=1 \mid X_{eats}=1) \end{aligned}$$

$H(X_{meat} \mid X_{eats}=0)$ **can be defined similarly**

Conditional Entropy: Complete Definition

$$H(X_{meat} \mid X_{eats}) = \sum_{u \in \{0,1\}} [p(X_{eats} = u)\, H(X_{meat} \mid X_{eats} = u)]$$

$$= \sum_{u \in \{0,1\}} [p(X_{eats} = u) \sum_{v \in \{0,1\}} [-p(X_{meat} = v \mid X_{eats} = u) \log_2 p(X_{meat} = v \mid X_{eats} = u)]]$$

In general, for any discrete random variables X and Y, we have H(X) $\geq$ H(X|Y)

What's the minimum possible value of H(X|Y)?

03:53 / 11:57

Conditional Entropy to Capture Syntagmatic Relation

$$H(X_{meat} \mid X_{eats}) = \sum_{u \in \{0,1\}} [p(X_{eats} = u)\, H(X_{meat} \mid X_{eats} = u)]$$

$$H(X_{meat} \mid X_{meat}) = ?$$

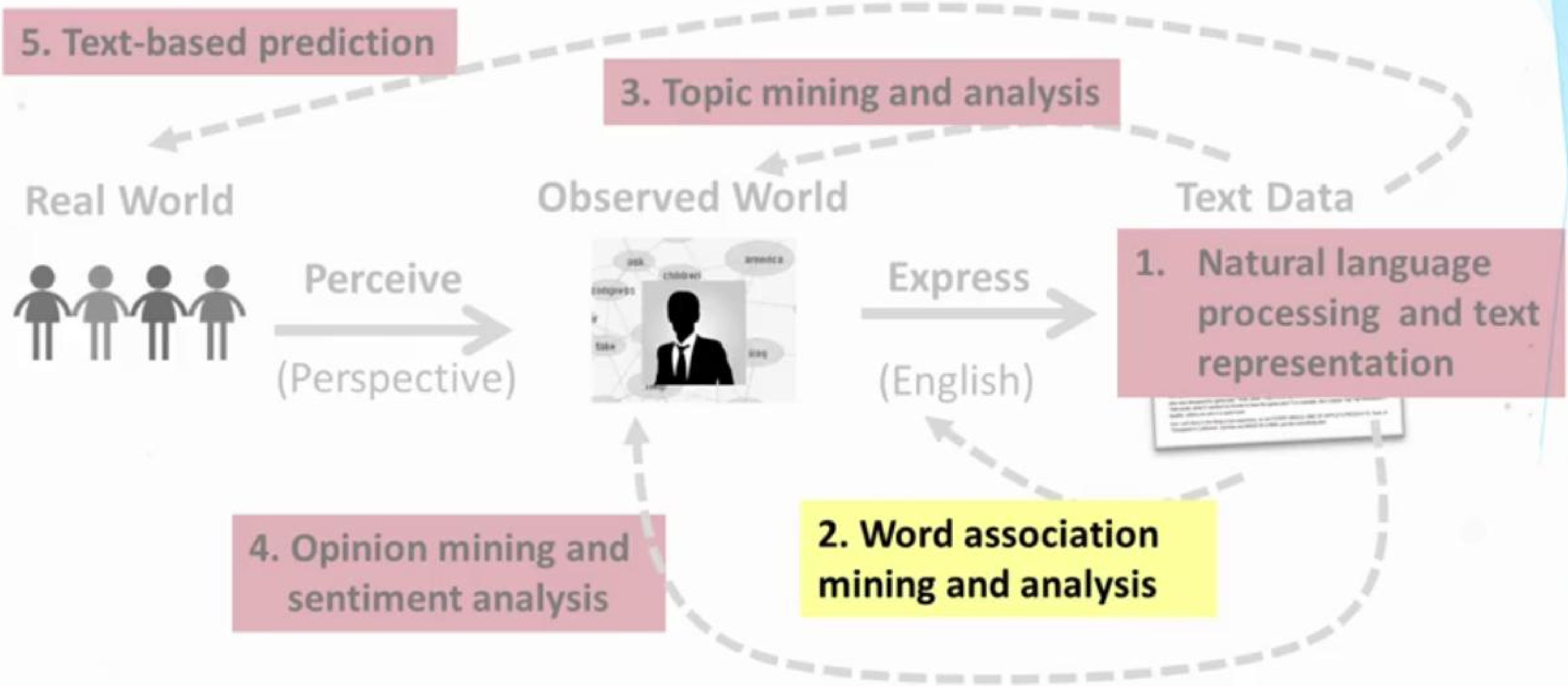
Which is smaller? $H(X_{meat}|X_{the})$ or $H(X_{meat}|X_{eats})$?
For which word w, does $H(X_{meat}|X_w)$ reach its minimum (i.e., 0)?
For which word w, does $H(X_{meat}|X_w)$ reach its maximum, $H(X_{meat})$?

Conditional Entropy for Mining Syntagmatic Relations

- For each word W1
  - For every other word W2, compute conditional entropy $H(X_{W1}|X_{W2})$
  - Sort all the candidate words in ascending order of $H(X_{W1}|X_{W2})$
  - Take the top-ranked candidate words as words that have potential syntagmatic relations with W1
  - Need to use a threshold for each W1
- However, while $H(X_{W1}|X_{W2})$ and $H(X_{W1}|X_{W3})$ are comparable, $H(X_{W1}|X_{W2})$ and $H(X_{W3}|X_{W2})$ aren't!

How can we mine the strongest K syntagmatic relations from a collection?

Syntagmatic Relation Discovery: Mutual Information



Mutual Information I(X;Y): Measuring Entropy Reduction

How much reduction in the entropy of X can we obtain by knowing Y?

Mutual Information: $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

Properties:
- Non-negative: $I(X;Y) \geq 0$
- Symmetric: $I(X;Y) = I(Y;X)$
- $I(X;Y) = 0$ iff X & Y are independent

When we fix X to rank different Ys, I(X;Y) and H(X|Y) give the same order, but I(X;Y) allows us to compare different (X,Y) pairs.

Mutual Information I(X;Y) for Syntagmatic Relation Mining

Mutual information: $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

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

Which **words** have high mutual information with "**eats**"?

$$I(X_{eats}; X_{meats}) = I(X_{meats}; X_{eats}) \quad > \quad I(X_{eats}; X_{the}) = I(X_{the}; X_{eats})$$

$$I(X_{eats}; X_{eats}) = H(X_{eats}) \geq I(X_{eats}; X_w)$$

Rewriting Mutual Information (MI) Using KL-divergence

The observed joint distribution of X_{W1} and X_{W2}

$$I(X_{w1};X_{w2}) = \sum_{u\in\{0,1\}} \sum_{v\in\{0,1\}} p(X_{w1}=u, X_{w2}=v) \log_2 \frac{p(X_{w1}=u, X_{w2}=v)}{p(X_{w1}=u)p(X_{w2}=v)}$$

The expected joint distribution of X_{W1} and X_{W2} if X_{W1} and X_{W2} were independent

MI measures the divergence of the actual joint distribution from the expected distribution under the independence assumption. The larger the divergence is, the higher the MI would be.

Probabilities Involved in Mutual Information

$$I(X_{w1};X_{w2}) = \sum_{u\in\{0,1\}} \sum_{v\in\{0,1\}} p(X_{w1}=u, X_{w2}=v) \log_2 \frac{p(X_{w1}=u, X_{w2}=v)}{p(X_{w1}=u)p(X_{w2}=v)}$$

Presence & absence of w1: $p(X_{W1}=1) + p(X_{W1}=0) = 1$

Presence & absence of w2: $p(X_{W2}=1) + p(X_{W2}=0) = 1$

Co-occurrences of w1 and w2:

$$\underbrace{p(X_{W1}=1, X_{W2}=1)}_{\text{Both w1 \& w2 occur}} + \underbrace{p(X_{W1}=1, X_{W2}=0)}_{\text{Only w1 occurs}} + \underbrace{p(X_{W1}=0, X_{W2}=1)}_{\text{Only w2 occurs}} + \underbrace{p(X_{W1}=0, X_{W2}=0)}_{\text{None of them occurs}} = 1$$

Relations Between Different Probabilities

Presence and absence of w1: $p(X_{w1}=1) + p(X_{w1}=0) = 1$

Presence and absence of w2: $p(X_{w2}=1) + p(X_{w2}=0) = 1$

Co-occurrences of w1 and w2:

$$p(X_{w1}=1, X_{w2}=1) + p(X_{w1}=1, X_{w2}=0)+p(X_{w1}=0, X_{w2}=1) + p(X_{w1}=0, X_{w2}=0) = 1$$

Constraints:

$$p(X_{w1}=1, X_{w2}=1) + p(X_{w1}=1, X_{w2}=0) = p(X_{w1}=1)$$

$$p(X_{w1}=0, X_{w2}=1) + p(X_{w1}=0, X_{w2}=0)= p(X_{w1}=0)$$

$$p(X_{w1}=1, X_{w2}=1) + p(X_{w1}=0, X_{w2}=1) = p(X_{w2}=1)$$

$$p(X_{w1}=1, X_{w2}=0) + p(X_{w1}=0, X_{w2}=0) = p(X_{w2}=0)$$

Computation of Mutual Information

Presence and absence of w1: $\boxed{p(X_{w1}=1)} + p(X_{w1}=0) = 1$

Presence and absence of w2: $\boxed{p(X_{w2}=1)} + p(X_{w2}=0) = 1$

Co-occurrences of w1 and w2:

$$\boxed{p(X_{w1}=1, X_{w2}=1)} + p(X_{w1}=1, X_{w2}=0) + p(X_{w1}=0, X_{w2}=1) + p(X_{w1}=0, X_{w2}=0) = 1$$

$$\boxed{p(X_{w1}=1, X_{w2}=1)} + p(X_{w1}=1, X_{w2}=0) = \boxed{p(X_{w1}=1)}$$

$$p(X_{w1}=0, X_{w2}=1) + p(X_{w1}=0, X_{w2}=0) = p(X_{w1}=0)$$

$$\boxed{p(X_{w1}=1, X_{w2}=1)} + p(X_{w1}=0, X_{w2}=1) = \boxed{p(X_{w2}=1)}$$

$$p(X_{w1}=1, X_{w2}=0) + p(X_{w1}=0, X_{w2}=0) = p(X_{w2}=0)$$

We only need to know $p(X_{w1}=1)$, $p(X_{w2}=1)$, and $p(X_{w1}=1, X_{w2}=1)$.

Syntagmatic Relation Discovery: Mutual Information

Part 2

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

Estimation of Probabilities (Depending on the Data)

$$p(X_{w1}=1)=\frac{count(w1)}{N}$$

$$p(X_{w2}=1)=\frac{count(w2)}{N}$$

$$p(X_{w1}=1, X_{w2}=1)=\frac{count(w1,w2)}{N}$$

	W1	W2	
Segment_1	**1**	**0**	Only W1 occurred
Segment_2	**1**	**1**	Both occurred
Segment_3	**1**	**1**	Both occurred
Segment_4	**0**	**0**	Neither occurred
…			
Segment_N	**0**	**1**	Only W2 occurred

Count(w1) = total number segments that contain W1
Count(w2) = total number segments that contain W2
Count(w1, w2) = total number segments that contain both W1 and W2

Smoothing: Accommodating Zero Counts

$$p(X_{w1}=1)=\frac{\text{count}(w1)+0.5}{N+1}$$

$$p(X_{w2}=1)=\frac{\text{count}(w2)+0.5}{N+1}$$

$$p(X_{w1}=1,X_{w2}=1)=\frac{\text{count}(w1,w2)+0.25}{N+1}$$

<u>Smoothing</u>: Add pseudo data so that no event has zero counts (pretend we observed extra data).

	W1	**W2**
¼ PseudoSeg_1	**0**	**0**
¼ PseudoSeg_2	**1**	**0**
¼ PseudoSeg_3	**0**	**1**
¼ PseudoSeg_4	**1**	**1**
Segment_1	**1**	**0**
…		
Segment_N	**0**	**1**

Actually observed data

Summary of Syntagmatic Relation Discovery

- Syntagmatic relation can be discovered by measuring correlations between occurrences of two words .
- Three concepts from information theory:
  - Entropy $H(X)$: measures the uncertainty of a random variable X
  - Conditional entropy $H(X|Y)$: entropy of X given we know Y
  - Mutual information $I(X;Y)$: entropy reduction of X (or Y) due to knowing Y (or X)
- Mutual information provides a principled way for discovering syntagmatic relations.