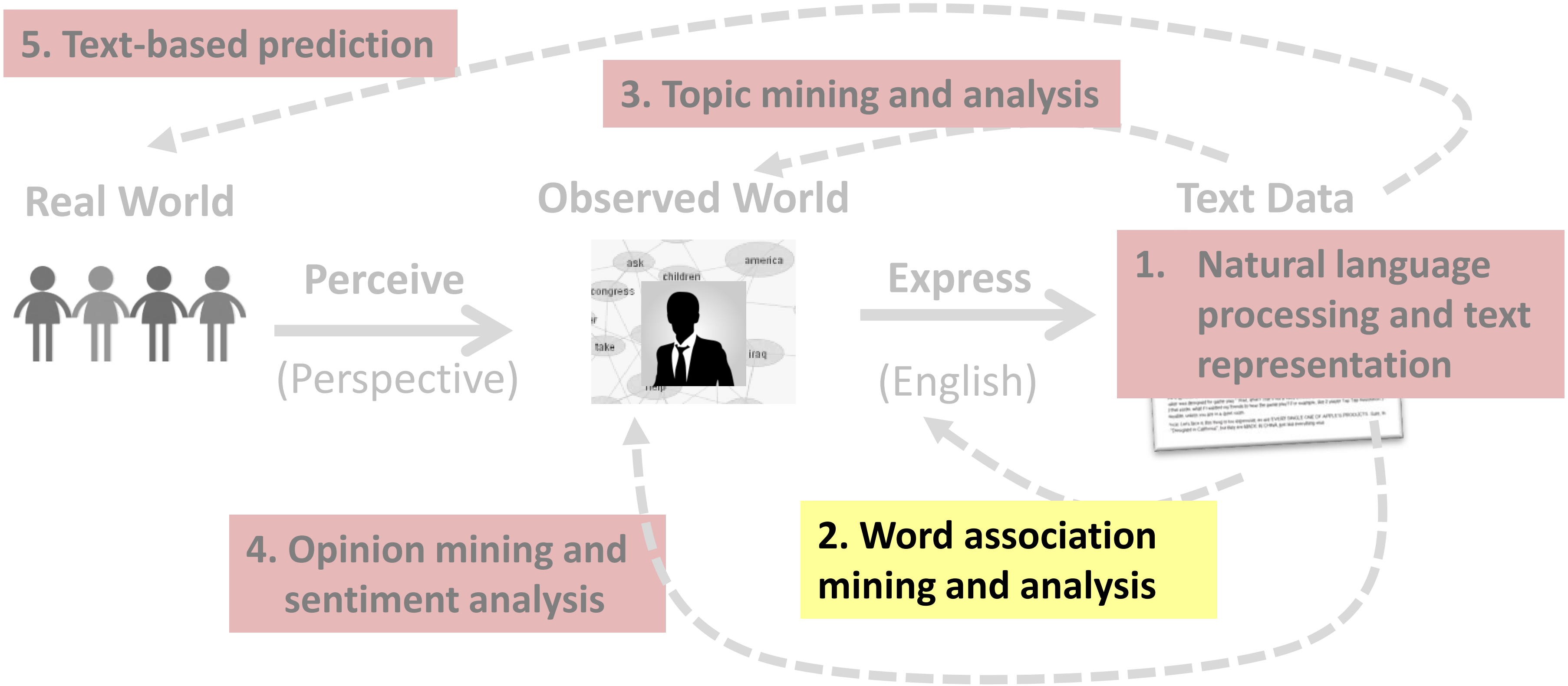
Syntagmatic Relation Discovery: Mutual Information

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

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**"?

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

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

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:

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

Relations Between Different Probabilities

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:

$$p(X_{W1}=1, X_{W2}=1) + p(X_{W1}=1, X_{W2}=0)+p(X_{W1}=0, X_{W2}=1) + p(X_{W1}=0, X_{W2}=0) = 1$$

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

Presence & 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 $p(X_{W1}=1, X_{W2}=1)$.

Estimation of Probabilities (Depending on the Data)

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

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

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

	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

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

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

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

	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

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

Summary of Word Association Mining

- Two basic associations: paradigmatic and syntagmatic
  - Generally applicable to any items in any language (e.g., phrases or entities as units)
- Pure statistical approaches are available for discovering both (can be combined to perform joint analysis).
  - Generally applicable to any text with no human effort
  - Different ways to define “context” and “segment” lead to interesting variations of applications
- Discovered associations can support many other applications.

Additional Reading

- Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999. (Chapter 5 on collocations)
- Chengxiang Zhai, Exploiting context to identify lexical atoms: A statistical view of linguistic context. Proceedings of the International and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-97), Rio de Janeiro, Brzil, Feb. 4-6, 1997. pp. 119-129.
- Shan Jiang and ChengXiang Zhai, Random walks on adjacency graphs for mining lexical relations from big text data. Proceedings of IEEE BigData Conference 2014, pp. 549-554.