Word Association Mining and Analysis

Outline

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

Basic Word Relations: Paradigmatic vs. Syntagmatic

- 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"
- 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"
- 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)
  - 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?)

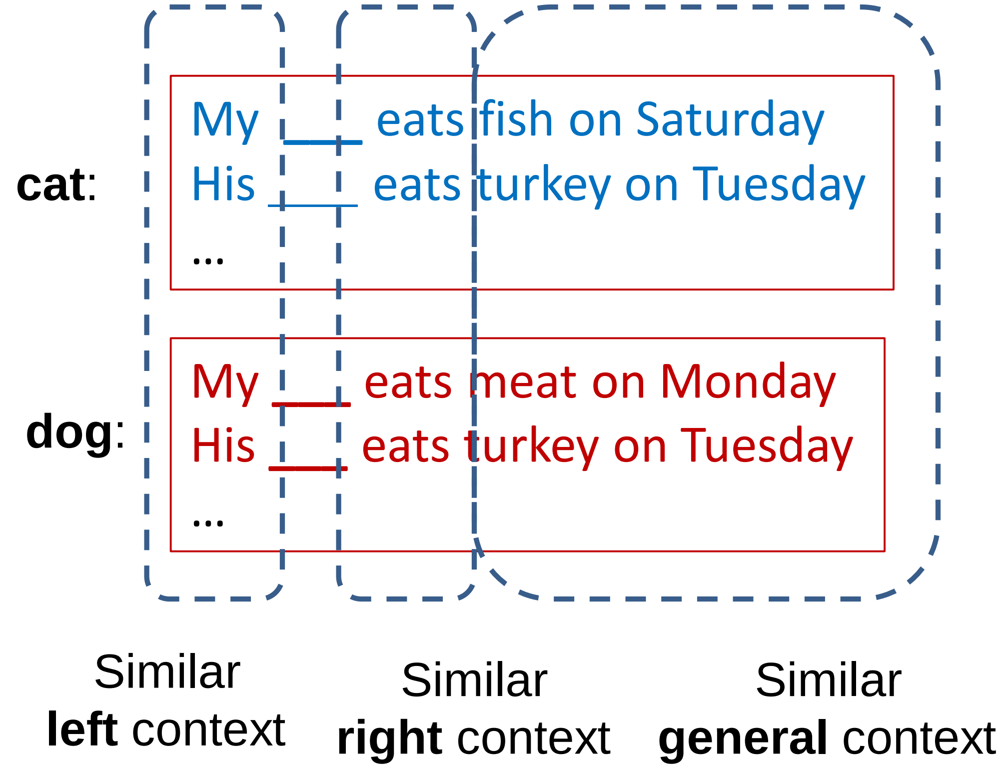
Outline

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

Mining Word Associations: Intuitions

Paradigmatic: similar context

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

My ___ **eats** ___ on Saturday
His ___ **eats** ___ on Tuesday
My ___ **eats** ___ on Monday
His ___ **eats** ___ on Tuesday
…

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

What words to the **right?**

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

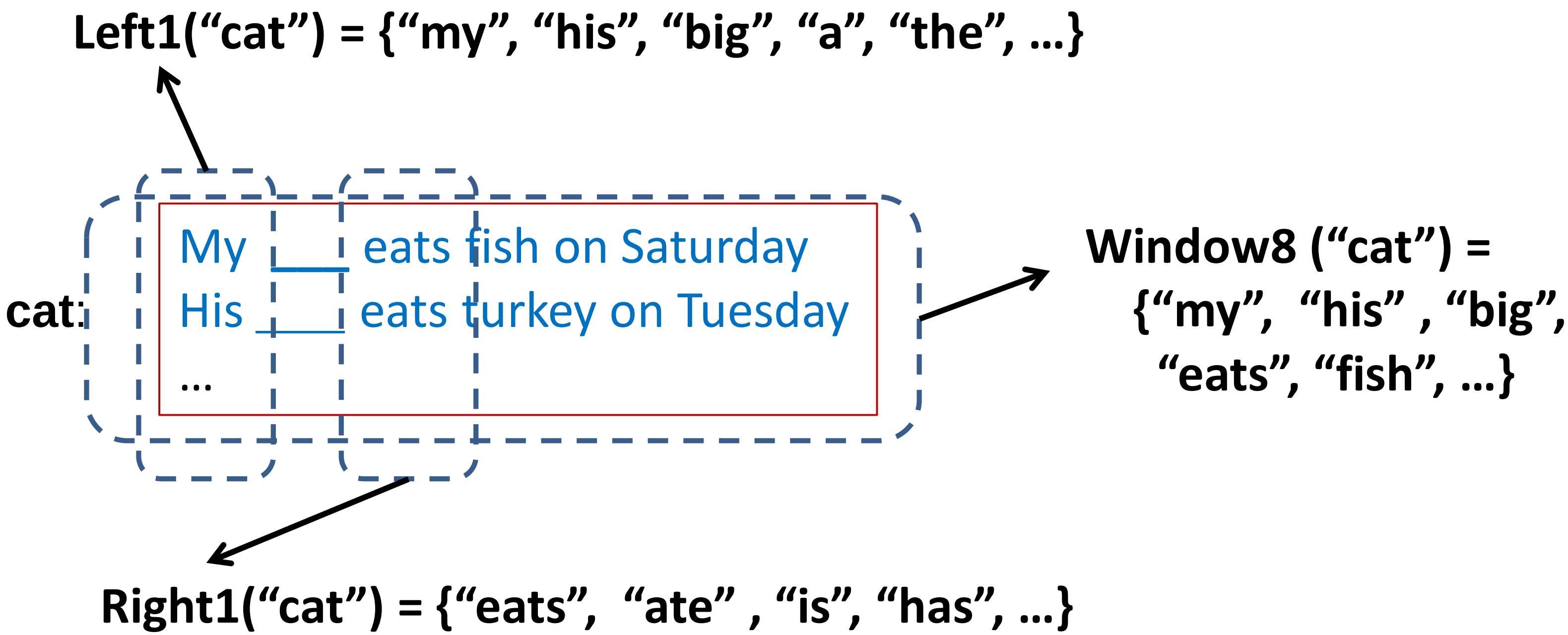
Mining Word Associations: General Ideas

- **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
- Paradigmatic related words tend to have syntagmatic relation with the same word $\rightarrow$ **joint discovery** of the two relations
- These ideas can be implemented in many different ways.

Outline

- What is a word association?
- Why mine word associations?
- How to mine word associations?
  - Paradigmatic Relation Discovery
  - Syntagmatic Relation Discovery

Word Context as "Pseudo Document"



Context = pseudo document = "bag of words"

Context may contain adjacent or non-adjacent words

Measuring Context Similarity

Sim (“cat”, “dog”) =

Sim (Left1(“cat”), Left1(“dog”))

+ Sim (Right1(“cat”), Right1(“dog”)) +

…

+ Sim (Window8(“cat”), Window8(“dog”)) = ?

High sim(word1, word2)

$\rightarrow$ word1 and word2 are **paradigmatically related**

Expected Overlap of Words in Context (EOWC)

Probability that a randomly picked word from d1 is w_i

Count of word w_i in d1

$$d1 = (x_1, \dots, x_N) \qquad x_i = \frac{c(w_i, d1)}{|d1|}$$

Total counts of words in d1

$$d2 = (y_1, \dots, y_N) \qquad y_i = \frac{c(w_i, d2)}{|d2|}$$

$$Sim(d1, d2) = d1.d2 = x_1 y_1 + \dots + x_N y_N = \sum_{i=1}^{N} x_i y_i$$

Probability that two randomly picked words from $d1$ and $d2$, respectively, are identical.

Would EOWC Work Well?

- Intuitively, it makes sense: the more overlap the two context documents have, the higher the similarity would be
- However
  - It favors matching one frequent term very well over matching more distinct terms
  - It treats every word equally (overlap on “the” is not as so meaningful as overlap on “eats”)

Improving EOWC with Retrieval Heuristics

- It favors matching one frequent term very well over matching more distinct terms

$\rightarrow$ **Sublinear transformation of Term Frequency (TF)**

- It treats every word equally (overlap on "the" is not as so meaningful as overlap on "eats")

$\rightarrow$ **Reward matching a rare word: IDF term weighting**

Adapting BM25 Retrieval Model for Paradigmatic Relation Mining

$$\boldsymbol{d1 = (x_1, \ldots, x_N)} \qquad BM25(w_i, d1) = \frac{(k+1)c(w_i, d1)}{c(w_i, d1) + k\left(1 - b + b * \frac{|d1|}{avdl}\right)}$$

$$x_i = \frac{BM25(w_i, d1)}{\sum_{j=1}^{N} BM25(w_j, d1)} \qquad b \in [0,1] \quad k \in [0, +\infty)$$

$\boldsymbol{d2 = (y_1, \ldots, y_N)}$ **$\boldsymbol{y_i}$ is defined similarly**

$$\boldsymbol{Sim(d1, d2) = \sum_{i=1}^{N} IDF(w_i) x_i y_i}$$

BM25 can also Discover Syntagmatic Relations

$\boldsymbol{d1 = (x_1, \ldots, x_N)}$ $$BM25(w_i, d1) = \frac{(k+1)c(w_i, d1)}{c(w_i, d1) + k\left(1 - b + b * \frac{|d1|}{avdl}\right)}$$

$$x_i = \frac{BM25(w_i, d1)}{\sum_{j=1}^{N} BM25(w_j, d1)}$$

$b \in [0, 1]$

$k \in [0, +\infty)$

IDF Weighted $\boldsymbol{d1 = (x_1 * IDF(w_1), \ldots, x_N * IDF(w_N))}$

The highly weighted terms in the context vector of word w are likely syntagmatically related to w

Summary of Paradigmatic Relation Discovery

- Main idea for discovering paradigmatic relations:
  - Collecting the context of a candidate word to form a pseudo document (bag of words)
  - Computing the similarity of the corresponding context documents of two candidate words
  - Highly similar word pairs can be assumed to have paradigmatic relations
- Many different ways to implement this general idea
- Text retrieval models can be easily adapted for computing similarity of two context documents
  - BM25 + IDF weighting
  - Syntagmatic relations can also be discovered as a "by product"

Outline

- What is a word association?
- Why mine word associations?
- How to mine word associations?
  - Paradigmatic Relation Discovery
  - Syntagmatic Relation Discovery

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 Monday
His dog **eats** turkey on Tuesday
…

My ___ **eats** ___ on Saturday
His ___ **eats** ___ on Tuesday
My ___ **eats** ___ on Monday
His ___ **eats** ___ on Tuesday
…

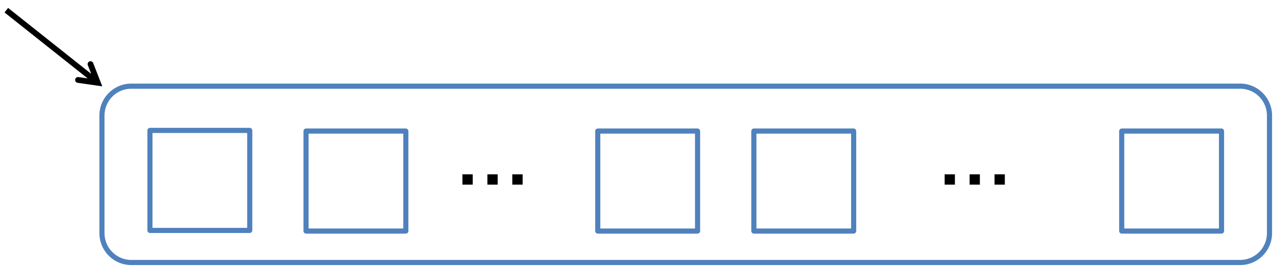
What words tend to occur to the **left** of "**eats**"?

What words 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 other?

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

Word Prediction: Formal Definition

Binary Random Variable:

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

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

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

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 X_w?

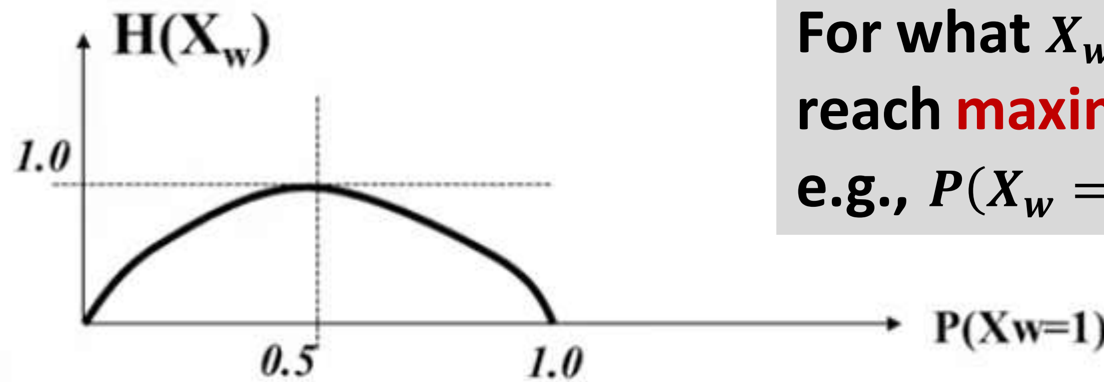
Entropy H(X) Measures Randomness of X

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

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

$$Define \ \ 0\log_2 0 = 0$$

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



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

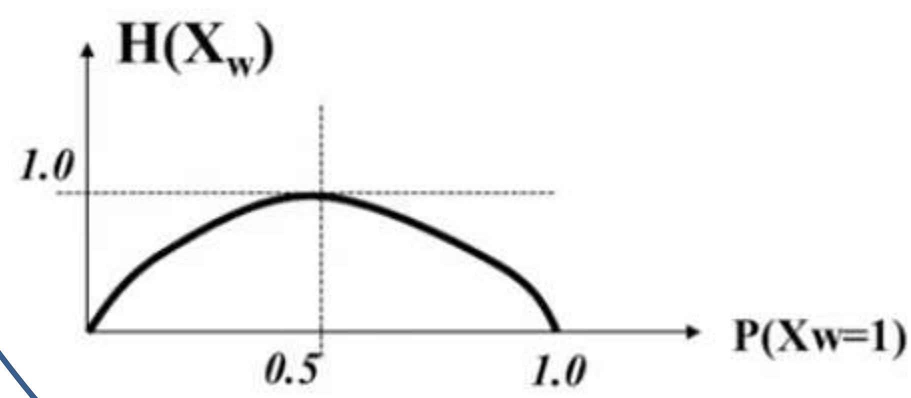
$$X_{coin} = \begin{cases} 1 & head \\ 0 & tail \end{cases}$$

Fair coin: $p(X=1) = p(X=0) = \frac{1}{2}$

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

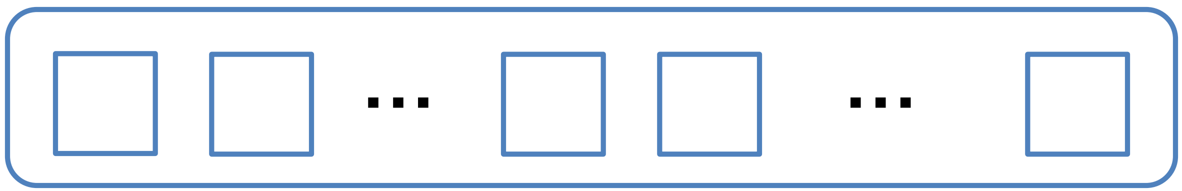
Completely biased: $p(X=1) = 1$

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



Entropy for Word Prediction

Is word $\boldsymbol{W}$ present (or absent) in this segment?



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

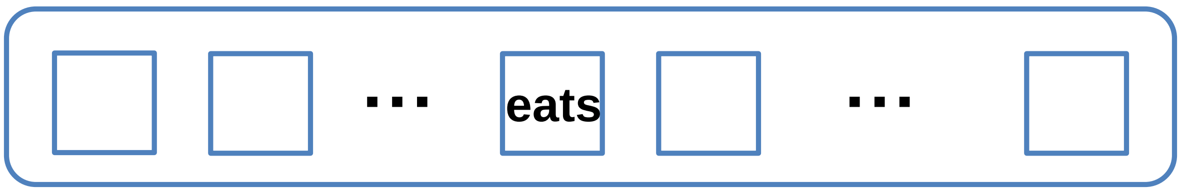
Which is **high/low?** $\boldsymbol{H(X_{meat}), H(X_{the}), H(X_{unicorn})}$**?**

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

High entropy words are harder to predict!

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,** $\boldsymbol{H(X_{meat})}$**?**

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

Conditional Entropy

Know nothing about the segment

$\boldsymbol{p(X_{meat}=1)}$

$\boldsymbol{p(X_{meat}=0)}$

Know "eats" is present $\boldsymbol{X_{eats}=1}$

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

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

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

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

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

Conditional Entropy: Complete Definition

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

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

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

What is the minimum possible value of $H(X|Y)$?

Conditional Entropy to Capture Syntagmatic Relation

$$H(X_{meat}|X_{eats}) = \sum_{u\in\{0,1\}} [p(X_{eats} = u)H(\mathbf{X}_{\text{meat}}|\,\mathbf{X}_{\text{eats}} = \mathbf{u})]$$

$$H(X_{meat}|X_{meat}) = ?$$

Which is smaller? $H(Xmea_t|X_{the})$ **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,** $H(Xmea_t)$**?**

Conditional Entropy for Mining Syntagmatic Relations

- For each word $w1$
  - For every other word $w2$, compute conditional entropy $H(Xw_1|Xw_2)$
  - Sort all the candidate words in ascending order of $H(Xw_1|Xw_2)$
  - 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 $H(Xw_1|Xw_2)$ and $H(Xw_1|Xw_3)$ are comparable, $H(Xw_1|Xw_2)$ and $H(Xw_3|Xw_2)$ are not!

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

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

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

Mutual Information:

$$\boldsymbol{I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)}$$

- Properties:
  - Non-negative: $I(X;Y) \quad 0$
  - 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**"?

$$I(X_{eats}; Xm_{eat}) = I(X_{meat}; X_{eats}) > I(X_{eats}; X_{the}) = I(X_{the}; X_{eats})$$

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

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.

MI and KL-divergence

$$I(X;Y) = H(Y) - H(Y|X)$$

$$= -\sum_{y\in\{0,1\}} P(Y=y)\log P(Y=y)$$

$$-\sum_{x\in\{0,1\}} P(X=x)H(Y|X=x)$$

$$= -\sum_{y\in\{0,1\}} P(Y=y)\log P(Y=y)$$

$$+\sum_{x\in\{0,1\}} P(X=x)\sum_{y\in\{0,1\}} P(Y=y|X=x)\log P(Y=y|X=x)$$

$$= \cdots$$

MI and KL-divergence (cont.)

$$H(X;Y) = H(Y) - H(Y|X)$$

$$= \cdots$$

$$= \sum_{x\in\{0,1\}} \sum_{y\in\{0,1\}} P(Y=y, X=x) \log \frac{P(Y=y|X=x)}{P(Y=y)}$$

$$= -\sum_{y\in\{0,1\}} \sum_{x\in\{0,1\}} P(Y=y, X=x) \log P(Y=y)$$

$$+ \sum_{y\in\{0,1\}} \sum_{x\in\{0,1\}} P(Y=y, X=x) \log P(Y=y|X=x)$$

$$= \sum_{x\in\{0,1\}} \sum_{y\in\{0,1\}} P(Y=y, X=x) \log \frac{P(Y=y, X=x)}{P(Y=y)P(X=x)}$$

$$= KLD(P(X,Y)||P(X)P(Y))$$

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 $w1$: $\boldsymbol{p(X_{w1} = 1) + p(X_{w1} = 0) = 1}$

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

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

Relation Between Different Probabilities

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 $w2$:

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

Constraints:

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

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

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

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

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)

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

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

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

	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

$\boldsymbol{count(w1)}$ **= total number of segments that contain** $\boldsymbol{w1}$

$\boldsymbol{count(w2)}$ **= total number of segments that contain** $\boldsymbol{w2}$

$\boldsymbol{count(w1, w2)}$ **= total number of segments that contain both** $\boldsymbol{w1}$ **and** $\boldsymbol{w2}$

Smoothing: Accommodating Zero Counts

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

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

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

	$W1$	$W2$
¼ PseudoSeg1	0	0
¼ PseudoSeg1	1	0
¼ PseudoSeg1	0	1
¼ PseudoSeg1	1	1
Segment_1	1	0
…		
Segment_N	0	1

Actually observed data

Smoothing: Add pseudo data so that no event has zero counts (pretend we observed extra data)

Summary of Syntagmatic Relation Discovery

- Syntagmatic relations 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

Questions?