Syntagmatic Relation Discovery

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

- Syntagmatic:
  - A and B have syntagmatic relation if they can be combined with each other (i.e., A and B are related semantically)
    - E.g., "cat" and "sit"; "car" and "drive"

Prediction of words in a context of another word

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**?

Prediction of absence and presence of a word

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"

Entropy for Word Prediction

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"

- **Entropy of word "the" is close to 0 because "the" may occur in every document.**
- **It is more difficult to predict words with higher entropy values.**

What if we know "eat" occurs in the segment?

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"

Illustration of Conditional Entropy

Know nothing about the segment — Know "eats" is present ($X_{\text{eats}} = 1$)

$p(X_{\text{meat}} = 1)$ ⇢ $p(X_{\text{meat}} = 1 | X_{\text{eats}} = 1)$

$p(X_{\text{meat}} = 0)$ ⇢ $p(X_{\text{meat}} = 0 | X_{\text{eats}} = 1)$

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

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

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

Conditional Entropy for Syntagmatic Relation Discovery

1. For each word w_1, enumerate all other words w_2 from the corpus.
2. Compute $H(X_{w1} | X_{w2})$. Sort all candidates in ascending order of the conditional entropy.
3. Take the top-ranked candidate words as words that have potential syntagmatic relations with w_1.

Limitation: values are not comparable across different words, so it can not be used to mine the strongest k syntagmatic relations

Mutual information for discovering syntagmatic relations

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**"?

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

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

Mutual Information

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 the definition of 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 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 $w2$:

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

Both $w1$ and $w2$ occur | Only $w1$ occurs | Only $w2$ occurs | None of them occurs

Constraints on probabilities in the mutual information function

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:

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

UNIVERSITY of DELAWARE UNIVERSITY of DE

Computation of Mutual Information

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 $w2$:

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

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

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

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

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

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

15

UNIVERSITY of DELAWARE UNIVERSITY of DE

Estimation of Probabilities involved in the definition of mutual information

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

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

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

	$w1$	$w2$	
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 $w2$

16

UNIVERSITY of DELAWARE UNIVERSITY of DE

Smoothing in estimation of probabilities for computing mutual information

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

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

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

	$w1$	$w2$
¼ PseudoSeg_1	0	0
¼ PseudoSeg_2	1	0
¼ PseudoSeg_3	0	1
¼ PseudoSeg_4	1	1

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

Segment_1	1	0
...		
Segment_N	0	1

Actually observed data

17