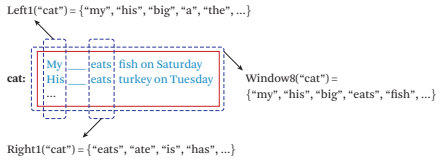
Paradigmatic Relation Discovery

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

Paradigmatic Relation

- Paradigmatic:
  - A and B have paradigmatic relation if they can be substituted for each other (i.e., A and B are in the same class)
    - E.g., "cat" and "dog"; "Monday" and "Tuesday"

Context of words convey semantics



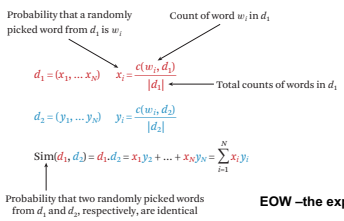
Context = pseudo document = "bag of words"
Context may contain adjacent or non-adjacent words

Multiple views of the context of a word can be used to compute similarity

Sim("cat", "dog") =
Sim(**Left1**("cat"), **Left1**("dog"))
+ Sim(**Right1**("cat"), **Right1**("dog")) +
...
+ Sim(**Window8**("cat"), **Window8**("dog")) = ?

High sim(word1, word2)
→ word1 and word2 are **paradigmatically related**

A similarity function for word contexts



EOW –the expected overlap of words

More about EOW

- 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. **Sublinear transformation of Term Frequency (TF)**
  - It treats every word equally (overlap on "the" is not as so meaningful as overlap on "eats"). **IDF weighting**

A different similarity function based on BM25

$d_1 = (x_1, \ldots x_N)$ $\quad \text{BM25}(w_i, d_1) = \dfrac{(k+1)c(w_i, d_1)}{c(w_i, d_1) + k(1 - b + b * |d_1|/\text{avdl})}$

$$x_i = \frac{\text{BM25}(w_i, d_1)}{\sum_{j=1}^{N} \text{BM25}(w_j, d_1)}$$

$b \in [0, 1]$
$k \in [0, +\infty)$

$d_2 = (y_1, \ldots y_N)$ $\quad y_i$ is defined similarly

$$\text{Sim}(d_1, d_2) = \sum_{i=1}^{N} \text{IDF}(w_i) x_i y_i$$

Summary

- Collecting the context of a candidate word to form a pseudo document
- Computing similarity of the corresponding context documents
- Word pairs with high similarity can be assumed to have paradigmatic relations