

UNIVERSITY of DELAWARE

UNIVERSITY of DE

Paradigmatic Relation Discovery

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

1

UNIVERSITY of DELAWARE

UNIVERSITY of DE

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"

2

UNIVERSITY of DELAWARE

UNIVERSITY of DE

Context of words convey semantics

Left1("cat") = {"my", "his", "big", "a", "the", ...}

cat:

My	eats	fish on Saturday
His	eats	turkey on Tuesday
...		

Window8("cat") = {"my", "his", "big", "eats", "fish", ...}

Right1("cat") = {"eats", "ate", "is", "has", ...}

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

3

UNIVERSITY of DELAWARE

UNIVERSITY of DE

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

4

UNIVERSITY of DELAWARE

UNIVERSITY of DE

A similarity function for word contexts

Probability that a randomly picked word from d_i is w_i

Count of word w_i in d_i

$d_1 = (x_1, \dots, x_n)$ $x_i = \frac{c(w_i, d_1)}{|d_1|}$ Total counts of words in d_i

$d_2 = (y_1, \dots, y_n)$ $y_i = \frac{c(w_i, d_2)}{|d_2|}$

Sim(d_1, d_2) = $d_1 \cdot d_2 = x_1 y_1 + \dots + x_n y_n = \sum_{i=1}^n x_i y_i$

Probability that two randomly picked words from d_1 and d_2 respectively, are identical

EOW –the expected overlap of words

5


UNIVERSITY of DELAWARE


UNIVERSITY of DE

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**

7


UNIVERSITY of DELAWARE


UNIVERSITY of DE

A different similarity function based on BM25


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


$$x_i = \frac{\text{BM25}(w_i, d_1)}{\sum_{j=1}^N \text{BM25}(w_j, d_1)} \quad \begin{matrix} b \in [0, 1] \\ k \in [0, +\infty) \end{matrix}$$

$$d_2 = (y_1, \dots, y_N) \quad y_i \text{ is defined similarly}$$

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

8


UNIVERSITY of DELAWARE


UNIVERSITY of DE

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

9