NLP - LEZIONE 19

DEL 12/12/2019

RAPPR. DISTRIBUITA I

SIMILARITÀ TRA FRASI

Siamo interessati a CALCOLARE la SIMILARITÀ tra due date frasi

$$f_1 = w_1 \dots w_n$$

$$f_2 = k_1 \dots k_m$$

Per fare questo vogliamo iniziare cercando una eventuale relazione di PARAFRASI tra f_1 e f_2. Questa relazione sussiste quando f_1 e f_2 stanno indicando la STESSA INFORMAZIONE.

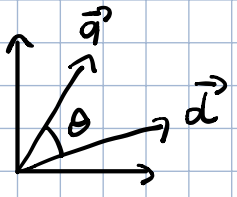
Notiamo però che il voler lavorare con le frasi intere può essere più utile lavorare con le SOTTOFRASI delle due frasi. Il concetto di SINONIMIA è infatti più CHIARO rispetto a quello della PARAFRASI.

> PARAFRASI := Relazione tra due FRASI che hanno lo stesso significato.
>
> SINONIMIA := Relazione tra due PAROLE che hanno lo stesso significato.

Notiamo che studiando IR (INFORMATION RETRIEVAL) per calcolare la similarità tra due frasi abbiamo

i) Rappresentato le FRASI come VETTORI in uno spazio vettoriale.

ii) Utilizzato il PRODOTTO SCALARE per calcolare l'angolo θ tra i due vettori.



Con il passo i) abbiamo trasformato una RAPPR. DISCRETA ($P = w_1 \dots w_n$) in una RAPPR. CONTINUA ($\vec{d} = (v_1, \dots, v_m)$). Le rappresentazioni CONTINUE sono anche dette DISTRIBUTED REPRESENTATIONS.

Q: Che RELAZIONI intercorrono tra le rappr. DISCRETE e quelle DISTRIBUITE?

Nella trasformazione dalla rappr. DISCRETA → DISTRIBUITA utilizzata in IR si PERDONO le informazioni relative all'ORDINE delle parole nelle frasi.

In generale si passa da rappr. DISCRETA a quelle DISTR. per poter lavorare su degli SPAZI METRICI. Molti algoritmi di APPRENDIMENTO funzionano su spazi metrici.

Andiamo quindi a vedere come si passa da rappr. DISCRETE a rappr. DISTRIBUITE nel contesto del NLP.

BAG OF WORDS ~ (14:00 min)

Sia $P = w_1 \dots w_n$. Vogliamo capire come ottenere

$$F(P) = P(w_1 \dots w_n) = \vec{P}$$

Nel modello BAG OF WORDS procediamo utilizzando la COMPOSIZIONALITÀ. In particolare definiamo prima la serie di VETTORI ONE-HOT per ogni PAROLA DISTINTA

$$F(w_1) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\vdots$$

$$F(w_m) = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

ONE-HOT VECTORS, formano la BASE ORTONORMALE dello spazio vettoriale.

(LOCAL DISTRIBUTED REPRESENTATION, PLATE 1995)

Date la frase $P = w_1 \dots w_n$ procediamo come segue:

$$F(w_1 \dots w_n) = \sum_{i=1}^{n} F(w_i)$$

È poi possibile aggiungere un PESO al contributo di ogni parola (vedi tf-idf ranking in IR).

Come detto prima, il modello BAG OF WORDS rede informazioni nell'ordine delle parole. Se volessimo CODIFICARE in una rappr. DISCRETA la frase $P = w_1 \dots w_n$ potremmo procedere come segue

$$F(w_1 \dots w_n) = \underbrace{F(w_1)}_{\substack{\text{CODIFICA} \\ w_1}} \oplus F(w_2) \oplus \dots \oplus F(w_n)$$

$\oplus$: OPERAZIONE DI CONCATENAZIONE

Con questa rappr. mantengo l'ordine delle parole in quanto mantengo SEPARATE le rappr. delle varie PARTI. Notiamo inoltre che se devo rappr. n simboli, con questa rappr. necessito $\log n$ bits per simbolo. Con la rappr. distribuita invece necessito n bits per simbolo. Questo tipo di rappr. è proprio quella che viene utilizzata per MEMORIZZARE i dati nei CALCOLATORI moderni.

La rappr. simbolica però non è METRICA, e quindi non posso calcolare la DISTANZA tra due ELEMENTI.

OSS: La rappr. DISTRIBUITA è chiamata così perché l'informazione dei SIMBOLI DISCRETI non si trova più in un POSTO UNICO ma è SPARSA in tanti posti.

FUNCTIONAL COMPOSITIONALITY ~ (35:00 min)

La FUNCTIONAL COMPOSITIONALITY è ottenuta applicando una funzione alle parti e poi mettendo tutto insieme. La funzione applicata alle PARTI e la funzione applicata al TUTTO potrebbero essere DIVERSE, come nel modello BAG OF WORDS.

La CONCATENATIVE COMPOSITIONALITY è ottenuta CONCATENANDO la CODIFICA di ogni parola. Con questo tipo di rappr. NON PERDIAMO informazioni sui singoli simboli utilizzati.

Notiamo che la composizionalità utilizzata nel BAG OF WORDS model NON è concatenativa, in quanto perdiamo l'ordine delle parole.

Ci poniamo quindi i seguenti QUESITI:

Q1) Quanto CONCATENATIVA può essere una composizionalità FUNZIONALE?

Q2) Utilizzando una rap. DISTRIBUITA, come facciamo a RIDURRE la DIMENSIONE dei vettori che utilizziamo?

OSS: L'ANALISI delle differenze tra i vari tipi di COMPOSIZIONALITÀ fu introdotta inizialmente dalle stesse che hanno sviluppato la teoria delle RETI NEURALI.

Si era interessati a capire se le rappr. DISTRIBUITE contenevano più o meno INFORMAZIONI rispetto a quelle DISCRETE.

Alcuni (FODOR e PYLSHYN) sostenevano che non c'era alcuna differenza tra le due rappr. Un altro (?) diceva che l'UNICA diff. è il fatto che nel mondo DISTRIBUITO le informazioni non sono esplicitamente e questo ci permette di fare una SINGOLA operazione MATRICIALE che prende TUTTE le informazioni ASSIEME.

SEMANTICA DISTRIBUZIONALE (PRESENTAZIONE FEDERICO & PAOLO) ~ (49:00)

Lo SPAZIO METRICO DISTRIBUZIONALE è stato introdotto per cercare di DIMINUIRE la DIMENSIONE dei vettori con cui si lavora. Esso è basato sulla seguente IPOTESI LINGUISTICA, detta anche IPOTESI DISTRIBUZIONALE:

> due PAROLE sono SIMILI se tendono a comparire negli STESSI DOCUMENTI.

OSS: LO SPAZIO METRICO DISTRIBUZIONALE È DIVERSO da quello DISTRIBUITO, e trova le sue origini nelle IDEE di due LINGUISTI: FIRTH e HARRIS.

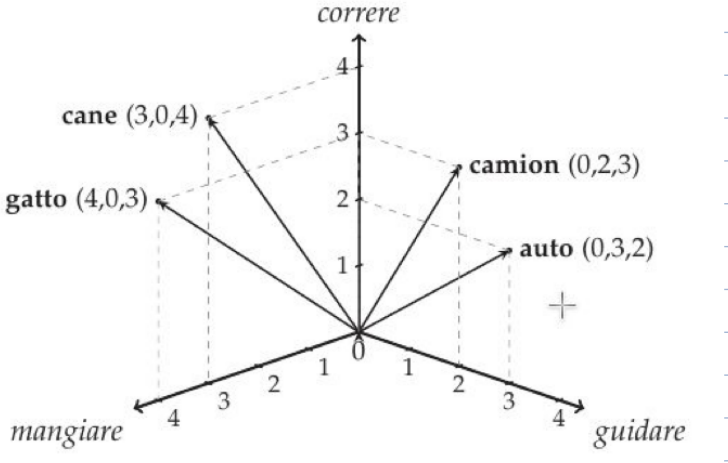
FIRTH → "Il SIGNIFICATO delle PAROLE è definito dal CONTESTO in cui vengon utilizzate."

HARRIS → "Due PAROLE son SIMILI se si PRESENTANO in CONTESTI SIMILI."

L'idea dietro all'IPOTESI DISTRIBUZIONALE è che la PRESENZA di una parola INFLUENZA le parole vicine alla parola: "Il CANE abbaia" VS "Il GATTO miagola".

Lo spazio metrico DISTRIBUZIONALE funziona quindi definendo un insieme di CONTESTI, che formano la base dello spazio vettoriale. Si procede quindi associando ad ogni PAROLA un VETTORE.

In questo esempio i contesti son
B = {MANGIARE, CORRERE, GUIDARE}



Formalmente uno SPAZIO SEMANTICO DISTRIBUZIONALE è definito come quadrupla $S = \langle T, B, M, F \rangle$, con:

- $T :=$ {PAROLE TARGET w}
- $B :=$ la BASE che contiene i CONTESTI LINGUISTICI sui quali viene valutata la SIMILARITÀ.
- $M :=$ Matrice di CO-OCCORRENZA
- $F :=$ METRICA utilizzata per misurare la DISTANZA tra i vettori.

L'aspetto FONDAMENTALE nella costruzione di questi nuovi è la SCELTA dei CONTESTI, ciascuno con la rispettiva GRANULARITÀ, che utilizziamo per COSTRUIRE i vari VETTORI.

OSS: Notiamo che utilizzando CONTESTI ALLARGATI "grandi" potrei trovare SIMILI parole OPPOSTE come "SALIRE LE SCALE", "SCENDERE LE SCALE".

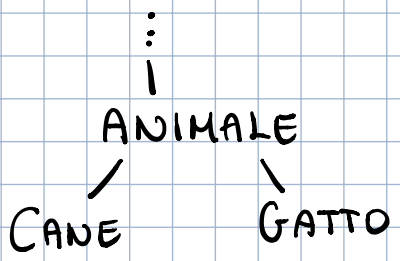
Negli SPAZI DISTRIBUZIONALI devo quindi stare ATTENTO a:

i) I VETTORI che rappr. le varie PAROLE.

ii) La FINESTRA utilizzata per CONTESTUALIZZARE le varie parole. Più GRANDE è la finestra, e più le parole saranno SIMILI per ragioni diverse dalle loro PROPRIETÀ.

WORD EMBEDDING

Il WORD-EMBEDDING è la TRASPOSIZIONE nelle RETI NEURALI del fatto che la similarità tra due parole DIPENDE dalla LUNGHEZZA DEL CONTESTO e dal CORPUS che utilizzo per fare TRAINING.



COTOPY

CANE e GATTO hanno entrambi correlate simili in quanto non la SPECIALIZZAZIONE della STESSA CLASSE SUPERIORE.

Rispetto a quanto fatto prima con lo SPAZIO METRICO DISTRIBUITO, con lo SPAZIO METRICO DISTRIBUZIONALE abbiamo cambiato i VETTORI e abbiamo DIMINUITO la DIMENSIONE dello SPAZIO.

Iniziamo con

$$F(w) = \vec{w} \in \mathbb{R}^n \quad (\text{ONE-HOT VECTOR})$$

Poniamo adesso utilizzare la MATRICE DI CO-OCCORENZA $W_{d \times n}$ definita nello spazio metrico DISTRIBUZIONALE come segue

$$\underset{d \times m}{W} \cdot \vec{w} = \vec{d} \in \mathbb{R}^d \quad (\text{MODELLO DISTRIBUZIONALE})$$

dove,

- m := # di PAROLE
- d := # di CONTESTI
- $(w_{i,j})$:= # di volte in cui la PAROLA j appare nel CONTESTO i.

Notiamo che $\vec{d} = W \cdot \vec{w}$ è un esempio di WORD EMBEDDING in quanto stiamo PASSANDO da un mondo all'altro tramite una TRASFORMAZIONE LINEARE.

Notiamo anche che NON SEMPRE ad OGNI DIMENSIONE è associato un SIGNIFICATO PRECISO. Può succedere infatti che ad una DIMENSIONE associamo una COMBINAZIONE LINEARE di più parole.

Per cercare di DIMINUIRE d l'idea è stata quella di trovare le righe della MATRICE W che non LINEARMENTE DIPENDENTI. Questa idea è implementata nella PCA (PRINCIPAL COMPONENT ANALYSIS)

La PCA forma un insieme di tecniche che utilizzano gli AUTOVETTORI della MATRICE per ridurre il # di DIMENSIONI.

La SVD (SINGULAR VALUE DECOMPOSITION) è la più famosa tecnica di PCA.

OSS: Andando a diminuire il # di dim. non potenzialmente PERDERE L'IDENTITÀ UNIVOCA dei SIMBOLI.

Ad esempio i SINONIMI potrebbero essere mappati nello STESSO ELEMENTO.

OSS: Alcuni esempi di RETI NEURALI che utilizzano WORD-EMBEDDING sono:

- ELMO
- BART
- CONCEPT-NET