Formalizzazione dell'Intreccio Semantico nei Testi: Un Modello Teorico

Luigi Usai LLM

Maggio 2025

Abstract

Questo paper propone un modello teorico per l'analisi e la formalizzazione dell'intreccio semantico presente in una pagina di testo. L'obiettivo è dimostrare che, data una pagina p di un racconto, l'intreccio semantico, denotato da intreccioSem(p), segue schemi fissi e determinabili scientificamente. Vengono definite le rappresentazioni di p, la semantica estratta dalle parole, e l'approccio grafico per descrivere tali relazioni. Tramite un'analisi basata su misure tipiche delle reti complesse, si evidenziano invarianti statistiche e geometriche che supportano la tesi di una struttura quasi universale degli intrecci semantici.

1 Introduzione

L'analisi delle strutture semantiche nei testi rappresenta una sfida affascinante per la linguistica computazionale e il trattamento automatico del linguaggio naturale. In questo lavoro si propone un approccio teorico volto a formalizzare l'*intreccio semantico* di una pagina di testo, ovvero la rappresentazione grafico-visiva delle relazioni tra le parole che compongono un testo. L'ipotesi centrale sostiene che, per pagine appartenenti a domini narrativi omogenei, tali reti semantiche evidenziano proprietà strutturali invarianti, identificabili attraverso metodi di analisi delle reti.

2 Definizioni Preliminari

Sia p una pagina di un racconto o di una storia. Le seguenti definizioni sono adottate per formalizzare il modello:

- **Pagina di Testo (p):** Una pagina contenente un insieme ordinato di parole.
- **Parole di p (**parole(p)**):** L'insieme delle parole, ovvero

$$\text{parole}(p) = \{w_1, w_2, \ldots, w_n\}.$$

- **Semantica del Testo (**sem(p)**):** La rappresentazione semantica ottenuta associando a ciascuna parola w un embedding

$$f(w) \in \mathbb{R}^d.$$

- **Visualizzazione di p (**Visualizzazione(p)**):** La disposizione grafica (ad esempio, un'immagine) del contenuto testuale di p.
- **Intreccio Semantico (**intreccioSem(p)**):** La rappresentazione grafica delle relazioni semantiche tra le parole di p, formalizzata come un grafo

$$G_p = (V, E),$$

dove:

- $V = \text{parole}(p)$,
- $E = \{(w_i, w_j) \mid S(w_i, w_j) > \tau\}$, con τ soglia prefissata.

La funzione di similarità tra due parole w_i e w_j è definita come:

$$S(w_i, w_j) = \frac{f(w_i) \cdot f(w_j)}{\|f(w_i)\| \, \|f(w_j)\|}.$$

Si definisce quindi la matrice di adiacenza $A = [A_{ij}]$ nel seguente modo:

$$A_{ij} = \begin{cases} 1, & \text{se } S(w_i, w_j) > \tau, \\ 0, & \text{altrimenti.} \end{cases}$$

3 Ipotesi e Schemi Invarianti

L'ipotesi principale è che, per pagine p appartenenti a un dominio narrativo omogeneo, il grafo semantico G_p presenti le seguenti proprietà:

1. **Distribuzione dei Gradi:** La probabilità $P(k)$ che un nodo abbia grado k segue una legge di potenza:

   $$P(k) \sim k^{-\gamma},$$

   con γ costante per il campione.

2. **Coefficiente di Clustering:** Per ogni nodo v del grafo, il coefficiente locale di clustering è:

   $$C_v = \frac{2 \cdot \text{numero di triangoli contenenti } v}{k_v(k_v - 1)},$$

   e il coefficiente medio di clustering è:

   $$C(G_p) = \frac{1}{|V|} \sum_{v \in V} C_v \approx C_0,$$

   dove C_0 è una costante.

3. **Lunghezza Media del Cammino:** La lunghezza media dei cammini in G_p è data da:

   $$L(G_p) = \frac{1}{|V|(|V| - 1)} \sum_{v \neq w} d(v, w) \approx L_0,$$

   dove $d(v, w)$ indica la distanza minima tra i nodi v e w, e L_0 è una costante.

4. **Invarianza Geometrica:** I layout ottenuti applicando algoritmi di disposizione, come quello di Fruchterman-Reingold, evidenziano invarianti geometriche (ad es. dimensione frattale d_F e modularità Q) assumendo valori costanti per pagine omogenee.

4 Procedura Operativa per la Validazione Sperimentale

Per verificare le ipotesi, si propone la seguente procedura operativa:

1. **Generazione del Campione:** Selezionare o generare N pagine $\{p_1, p_2, \ldots, p_N\}$, con $N \geq 100$, appartenenti ad un dominio narrativo omogeneo.
2. **Estrazione degli Elementi:** Per ogni pagina p_i, estrarre:
   - L'insieme delle parole, parole(p_i).
   - Gli embedding $f(w)$ per ogni parola $w \in p_i$.
3. **Costruzione del Grafo:** Per ciascuna pagina p_i, costruire il grafo

$$G_{p_i} = (V_i, E_i),$$

   dove

$$A_{ij} = \begin{cases} 1, & \text{se } \frac{f(w_i)\cdot f(w_j)}{\|f(w_i)\|\,\|f(w_j)\|} > \tau, \\ 0, & \text{altrimenti.} \end{cases}$$

4. **Estrazione delle Metriche:** Calcolare per ogni grafo G_{p_i} le seguenti metriche:
   - Distribuzione dei gradi e relativo esponente γ_i.
   - Coefficiente medio di clustering $C(G_{p_i})$.
   - Lunghezza media del cammino $L(G_{p_i})$.
   - Metriche geometriche (dimensione frattale $d_{F,i}$, modularità Q_i).
5. **Analisi Statistica:** Costruire un database delle metriche e verificare la convergenza verso valori medi:

$$\bar{\gamma} = \frac{1}{N}\sum_{i=1}^{N} \gamma_i, \quad \bar{C} = \frac{1}{N}\sum_{i=1}^{N} C(G_{p_i}), \quad \bar{L} = \frac{1}{N}\sum_{i=1}^{N} L(G_{p_i}).$$

5 Formulazione Teorica e Teorema di Invarianza

Teorema (Ipotesi di Invarianza dell'Intreccio Semantico):
Sia $\mathcal{P}$ l'insieme delle pagine testuali omogenee per stile e struttura narrativa e, per ogni $p \in \mathcal{P}$, sia G_p il grafo semantico ottenuto. Esistono costanti γ_0, C_0, L_0 e $d_{F,0}$ tali che:

$$\begin{aligned} P(k) &\sim k^{-\gamma_0}, \\ C(G_p) &\approx C_0, \\ L(G_p) &\approx L_0, \\ d_F(G_p) &\approx d_{F,0}. \end{aligned}$$

Schizzo della Dimostrazione:

1. *Analisi Empirica:* La raccolta di almeno 100 campioni di grafi G_p mostra la convergenza delle metriche γ, C e L attorno a valori medi costanti, con una varianza ridotta.
2. *Studio delle Proprietà di Rete:* Le reti semantiche naturali, analoghe a quelle derivate da testi narrativi, mostrano distribuzioni in legge di potenza ed elevati coefficienti di clustering, in linea con quanto previsto.

3. *Verifica Geometrica:* I layout dei grafi, analizzati attraverso metriche quali la dimensione frattale e l'indice di modularità, rivelano invarianti geometriche che supportano l'ipotesi di stabilità strutturale.

6 Conclusioni e Implicazioni

Il modello teorico presentato evidenzia come l'*intreccio semantico* di una pagina testuale, formalizzato attraverso grafi basati sulla similarità semantica, segua schemi fissi e quasi invariabili. Le proprietà statistiche e geometriche dei grafi G_p suggeriscono che esistono regole strutturali che governano la disposizione e l'interconnessione delle parole in un testo narrativo.

- **Applicazioni in Linguistica Computazionale:** Sviluppo di algoritmi per la categorizzazione automatica e il riconoscimento di stili narrativi.
- **Applicazioni in Intelligenza Artificiale:** Supporto all'interpretazione semantica e al riassunto automatico basato su reti semantiche.
- **Contributi alla Teoria delle Reti Complesse:** Evidenziazione di invarianti in sistemi naturali derivati da dati testuali.

Parole Chiave: Teoria dei Grafi, Linguistica Computazionale, Network Analysis, Intreccio Semantico, Invarianza Statistica, Testo Narrativo.

Autori: Luigi Usai e LLM