
ANNO ACCADEMICO 2024/2025

Tecnologie del Linguaggio Naturale

Teoria - Di Caro

Altair's Notes



DIPARTIMENTO DI INFORMATICA

CAPITOLO 1	SEMANTICA COMPUTAZIONALE	PAGINA 5
1.1	Introduzione	5
	Semantica Computazionale — 5 • Origini del NLP — 6 • Word Sense Induction — 6 • Comprensione del Linguaggio Naturale — 7	
1.2	Definizioni e Ricerca Onomasiologica	8
	Definizione delle Definizioni — 9 • Semasiologia e Onomasologia — 9	
CAPITOLO 2	TEORIE DEL SIGNIFICATO	PAGINA 11
CAPITOLO 3	CLUSTERING E TOPIC MODELLING	PAGINA 13
3.1	Panoramica	13

Premessa

Licenza

Questi appunti sono rilasciati sotto licenza Creative Commons Attribuzione 4.0 Internazionale (per maggiori informazioni consultare il link: <https://creativecommons.org/version4/>).



Formato utilizzato

Box di "Concetto sbagliato":

Concetto sbagliato 0.1: Testo del concetto sbagliato

Testo contenente il concetto giusto.

Box di "Corollario":

Corollario 0.0.1 Nome del corollario

Testo del corollario. Per corollario si intende una definizione minore, legata a un'altra definizione.

Box di "Definizione":

Definizione 0.0.1: Nome delle definizioni

Testo della definizione.

Box di "Domanda":

Domanda 0.1

Testo della domanda. Le domande sono spesso utilizzate per far riflettere sulle definizioni o sui concetti.

Box di "Esempio":

Esempio 0.0.1 (Nome dell'esempio)

Testo dell'esempio. Gli esempi sono tratti dalle slides del corso.

Box di "Note":

Note:-

Testo della nota. Le note sono spesso utilizzate per chiarire concetti o per dare informazioni aggiuntive.

Box di "Osservazioni":

Osservazioni 0.0.1

Testo delle osservazioni. Le osservazioni sono spesso utilizzate per chiarire concetti o per dare informazioni aggiuntive. A differenza delle note le osservazioni sono più specifiche.

1

Semantica Computazionale

1.1 Introduzione

1.1.1 Semantica Computazionale

La semantica computazionale (fig: 1.1) può essere divisa grossolonomicamente in tre parti:

- **Semantica lessicale:** consiste nello studio di *come* e *che cosa* denotano le parole di una lingua. Si analizzano:
 - *Significato letterale.*
 - *Polisemia:* parole con più significati.
 - *Relazioni semantiche:* sinonimia, antonimia, iponimia, etc.
 - *Composizione del significato.*
- **Semantica formale:** studia i modelli logico-matematici che definiscono formalmente i linguaggi. L'obiettivo è definire il significato in termini di condizioni di verità.
- **Semantica statistico-distribuzionale:** approccio computazionale e quantitativo al significato che combina metodi statistici e intuizioni linguistiche (in particolare il fatto che il significato delle parole possa essere inferito dalla loro distribuzione sui testi). Si analizzano grandi corpora per costruire rappresentazioni vettoriali delle parole (*embeddings*), in cui la vicinanza tra vettori (solitamente si usa la *cosine similarity*) riflette la somiglianza semantica.

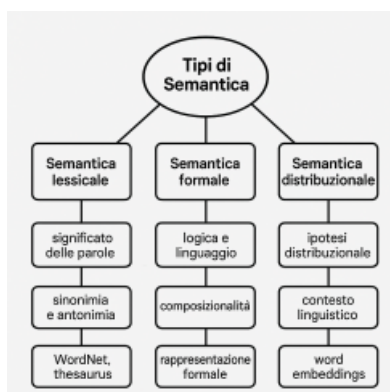


Figure 1.1: Tipi di semantica.

1.1.2 Origini del NLP

Inizialmente la linguistica computazionale e l'elaborazione del linguaggio naturale si occupavano del *question answering* (Q&A) ossia permettere a una macchina di leggere un testo (*what*) e rispondere a domande poste da un utente (*why*) attraverso l'impiego di codice e di risorse linguistiche (*how*). Con il passare del tempo le domande sono diventate sempre più complesse e variegiate riguardando: fatti specifici, richieste di elenchi, definizioni, motivazioni, elenchi, etc.

Proprio per questo motivo è emersa una nuova area di ricerca appositamente per la caratterizzazione delle domande. L'obiettivo è quello di costruire tassonomie per modellare ogni possibile sfaccettatura che una domanda possa avere. Tutto ciò per aumentare l'efficacia dei sistemi di Q&A che necessitano, in primo luogo, di comprendere la tipologia di domanda.

Negli ultimi, grazie allo sviluppo dell'*intelligenza artificiale generativa* (e dei modelli GPT, LLaMA, etc.), il Q&A si è evoluto. Questi modelli possono estrarre domande da un testo e *generare* risposte articolate, sintetiche o creative. Possono:

- Rispondere a domande complesse.
- Gestire dialoghi multi turno.
- Tradurre le domande e le risposte.
- Spiegare le proprie risposte.

Note:-

Quindi è ancora più importante determinare il contenuto della domanda in modo da evitare *allucinazioni*. Storicamente il Q&A è sempre stato un task complesso, ma nel periodo post-ChatGPT sta vedendo il suo apice mediante il meccanismo chiamato *prompting*.

Gran parte della ricerca non si limita al Q&A, ma anche a ciò che ci sta dietro o in parallelo. Per esempio PoS, NeR tagging, iperonimi, word sense disambiguation, etc. Oppure casi come quello del *suggeritore automatico*, presente nelle tastiere degli smartphone, che usa un modello statistico.

1.1.3 Word Sense Induction

Definizione 1.1.1: Word Sense Induction

La Word Sense Induction (WSI) è il task che riguarda l'identificazione del senso di una parola polisemica in una frase, all'interno di un determinato contesto.

Questo task ha problemi di:

- *Specificità*: molti sensi attribuiti alle parole non vengono utilizzati perché troppo specifici e sono solamente *rumore*. Questo è criticato da vari studiosi che sostengono sia necessario aggregare alcuni sensi troppo simili (e.g. in WordNet).
- *Copertura*: ci sono molte zone di linguaggio non coperte.
- *Soggettività*: nonostante le decisioni siano prese collettivamente c'è sempre una componente soggettiva.

Differenze tra Word Sense Induction (WSI) e Word Sense Disambiguation (WSD):

- La disambiguazione ha necessità di un dizionario/sense inventory (e.g. WordNet) che contiene tutti i possibili sensi per ogni parola. Nel WSI non esiste un dizionario.
- Nel WSI ci si basa sull'effettivo uso della parola in grandi quantità di dati.
- La WSD, essendo fatta da linguisti, è basata sulla grammatica. La WSI è basata sull'uso delle parole, anche sgrammaticato.
- La valutazione nel WSD è semplice (per esempio usando synsets gold), ma criticabile come visto in precedenza. Nel WSI è un po' più complicato.

Corollario 1.1.1 Pseudo-word

Il metodo della Pseudo-word è una tecnica per valutare algoritmi di WSI in assenza di risorse semantiche o annotazioni di senso. L'idea è quella di simulare l'ambiguità lessicale creando artificialmente delle parole ambigue e poi testare se il sistema è in grado di distinguerne i sensi sottostanti.

Fasi della Pseudo-word:

1. **Merging:** concatenazione di parole reali. Consiste nel fondere due o più parole esistenti in una parola ambigua. Queste parole devono avere significati distinti e usi in contesti diversi.
2. **Substitution:** sostituzione nei contesti. Tutte le occorrenze delle parole originali vengono sostituite nei testi dalla nuova parola create nel merging.
3. **Clustering:** identificazione dei sensi. Si applica un algoritmo di clustering (e.g. k-means, DB-SCAN, o modelli basati su embeddings) sulle rappresentazioni contestuali delle Pseudo-word per scoprire gruppi di usi distinti (sensi).
4. **Cluster-to-Class evaluation:** valutazione. Si valuta la qualità dei clusters ottenuti confrontandoli con le parole originali

Vantaggi e Limiti:

- ✓ Non sono richieste annotazioni manuali o risorse linguistiche.
- ✓ Può essere usato su grandi corpora in maniera automatica.
- ✗ I sensi creati non riflettono ambiguità reali.
- ✗ I contesti potrebbero essere troppo distinti e causare *overfitting*.

1.1.4 Comprensione del Linguaggio Naturale

- **Dizionari Elettronici:**
 - **Potere espressivo:** medio-alto, forniscono relazioni semantiche ricche.
 - **Scalabilità:** medio-bassa, solitamente sono costruiti manualmente (quindi difficilmente estendibili).
 - **Sorgente:** curata manualmente da esperti linguisti.
 - **Ambiguità e Soggettività:** ambiguità ridotta, soggettività media (su accezioni meno comuni).
- **Property Norms:**
 - **Potere espressivo:** alto per concetti concreti.
 - **Scalabilità:** bassa perché richiede raccolta tramite esperimenti psicologici o annotazioni.
 - **Sorgente:** spesso da studi cognitivi o crowd-sourcing/mechanical turk.
 - **Ambiguità e Soggettività:** alta, perché le persone non sono linguisti.
- **Frames:**
 - **Potere espressivo:** alto, cattura le strutture sintattico-semantiche.
 - **Scalabilità:** media, possono essere ampliati con annotazioni automatiche, ma richiede risorse linguistiche robuste.
 - **Sorgente:** tipicamente linguistica, contributi da annotatori esperti.
 - **Ambiguità e Soggettività:** media, perché c'è spesso intervento umano, ma si rimedia con strumenti automatici.
- **Senso Comune:**
 - **Potere espressivo:** molto alto, copre inferenze, aspettative sociali e causalità.

- *Scalabilità*: media, dato che servono molte persone.
- *Sorgente*: crowd-sourcing, scraping, machine learning.
- *Ambiguità e Soggettività*: molto alta, molte conoscenze sono implicite, culturali o controverse.
- *Visual Attributes*:
 - *Potere espressivo*: medio, utile per oggetti visibili, ma limitato ad aspetti percettibili.
 - *Scalabilità*: medio-alta con dataset di immagini annotate.
 - *Sorgente*: dati visivi + annotazioni.
 - *Ambiguità e Soggettività*: medio-alta, le percezioni visive sono soggettive.
- *Word Embedding*:
 - *Potere espressivo*: molto alto in contesti distribuzionali.
 - *Scalabilità*: molto alta, addestrabili su grandi corpora.
 - *Sorgente*: dati testuali in grande scala.
 - *Ambiguità e Soggettività*: medio-alta, dipendono dal contesto, dalla lingua e possono riflettere eventuali bias.
- *Corpus Manager*:
 - *Potere espressivo*: dipende dal corpus, utile per esplorare usi reali del linguaggio.
 - *Scalabilità*: alta, può gestire milioni/miliardi di parole.
 - *Sorgente*: testi reali.
 - *Ambiguità e Soggettività*: media, i dati sono "grezzi", quindi l'ambiguità linguistica è intrinseca.

Risorsa	Potere espressivo	Scalabilità	Sorgente	Ambiguità / Soggettività
Dizionari elettronici	Medio-alto	Medio-bassa	Manuale	Bassa
Property norms	Alto	Bassa	Esperimenti/Crowd	Alta
Frames	Alto	Media	Annotatori esperti	Media
Common-sense knowledge	Molto-alto	Alta	Crowd-sourcing/ML	Molto alta
Visual Attributes	Medio	Medio-alta	Immagini annotate	Medio-alta
Word/sense embeddings	Molto-alto	Molto alta	Testi su larga scala	Alta
Corpus manager	Variabile	Alta	Corpora reali	Media

Figure 1.2: Schema delle risorse.

1.2 Definizioni e Ricerca Onomasiologica

Nella linguistica la nozione di *definizione* è fondamentale. Una definizione è un testo progettato per guidare il lettore (o il *consumer*) verso un possibile significato associato a un termine all'interno di un contesto. Però bisogna tener presente che la relazione tra termini e significati non è univoca: un singolo termine può essere associato a una molteplicità di significati. Ogni significato può essere descritto da una definizione specifica o da più definizioni complementari.

Esistono diversi tipi di definizioni:

- *Genus-differentia*: identificano una categoria generale (*genus*) e specificano caratteristiche distintive.
- Definizioni basate su esempi.
- Definizioni tramite riferimenti ad altri concetti o termini già noti.
- Definizioni costruite tramite parafrasi, sinonimi o descrizioni operative.

Note:-

La qualità di una definizione dipende da chiarezza, accuratezza terminologica, adeguatezza rispetto al pubblico di riferimento, coerenza con il dominio e capacità di disambiguazione in contesti ambigui.

Definizione 1.2.1: Ricerca Onomasiologica

Partendo da un concetto o da un significato si deve identificare i termini o le espressioni linguistiche che possono denotarlo.

1.2.1 Definizione delle Definizioni**Domanda 1.1**

Come si descrive un concetto?

Domanda 1.2

Quali caratteristiche sono più importanti?

Domanda 1.3

Che relazione c'è tra un termine da definire e il suo gruppo semantico più generale?

Domanda 1.4

Come si scrive una definizione? Come si valuta la qualità di una definizione? Quanto si è d'accordo? Quanto si utilizza lo stesso linguaggio e la stessa terminologia? Come varia ciò tra concetti astratti/concreti e generici/specifici?

1.2.2 Semasiologia e Onomasologia

Nella lessicografia si possono distinguere due approcci alla relazione tra forma e contenuto:

- *Approccio semasiologico*: si parte da un termine linguistico (una parola o una locuzione) e si vogliono determinare i possibili significati. Si tratta dell'analisi correlata ai tradizionali dizionari.
- *Approccio onomasiologico*: si parte da un concetto, un'idea o una definizione e si vogliono individuare i termini linguistici che possono esprimere ciò. Questo processo è anche noto come *lessicalizzazione di un concetto*.

Aspetti collegati alla ricerca onomasiologica:

- *Dizionari analogici*: permettono la ricerca a partire da un concetto.
- *Tip-of-the-tongue*: fenomeno per cui il parlante ha in mente un concetto ma non riesce a esprimerlo.
- *Meccanismo del Genus-differentia*: supporta la costruzione di descrizioni concettuali per una risalita onomasiologica.

2

Teorie del Significato

3

Clustering e Topic Modelling

3.1 Panoramica

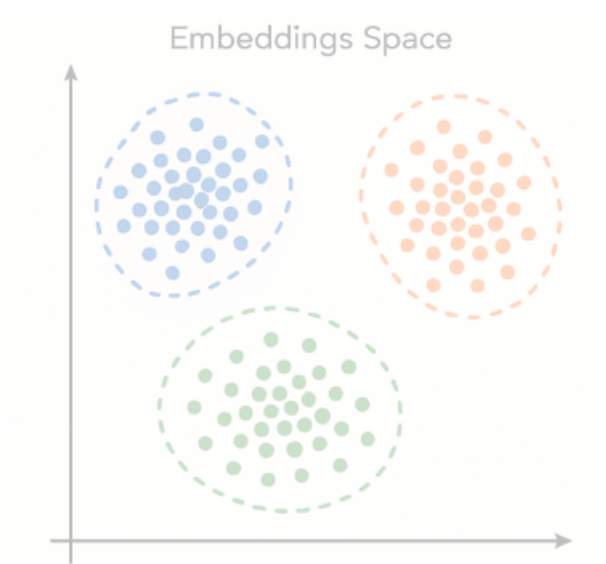


Figure 3.1: Clustering basato su embeddings.

