ANNO ACCADEMICO 2024/2025

Tecnologie del Linguaggio Naturale

Teoria - Di Caro

Altair's Notes

UNIVERSITÀ DI TORINO

DIPARTIMENTO DI INFORMATICA

Premessa

Licenza

Formato utilizzato

Box di "Concetto sbagliato":

Concetto sbagliato 0.1: Testo del concetto sbagliato

Testo contente 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 definizione

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 grossolonamente 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 somigliarità semantica.

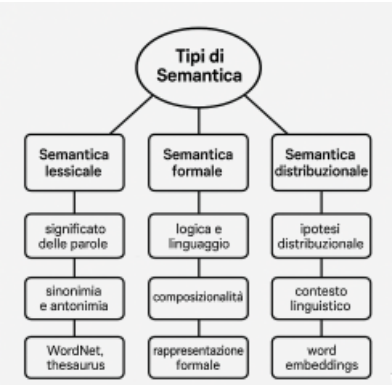

Figure 1.1: Tipi di semantica.

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 variegate 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 valutazionw 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 ilioni/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

2.1 Panoramica

Inizialmente, i tasks di NLP, richiedevano un approccio mirato:

- Venivano costruiti sistemi a regole, grammatiche o modelli statistici.
- Si utilizzavano risorse linguistiche annotate manualmente (dizionari, corpora taggati, etc.).
- Si progettavano algoritmi ad hoc per ciascun compito (parsing sintattico, disabiguazione semantica, traduzione automatica, etc.).
- Si ricorreva all'apprendimento supervisionato o semi-supervisionato, con ingegnerizzazione manuale delle features.
- La valutazione richiedeva spesso il coinvolgimento umano: annotatori, crowdsourcing, etc.

Note:-

Questo modo pre-LLM ha permesso la crescita del NLP e contribuito alla costruzione di molte risorse linguistiche (WordNet, FrameNet, etc.).

Negli ultimi anni l'avvento dei Large Language Model (LLM), modelli basati su reti neurali profonde e addestrati su enormi quantità di testo, ha causato uno shift di paradigma. Oggi molti tasks non richiedono più modelli separati: un singolo modello, ben progettato e promptato, è in grado di riassumere, tradurre, fare sentiment analisys, Q&A, etc.

2.1.1 Definizioni di Base

Alcune definizioni di base:

- *Lessico:* corrisponde al dizionario, ovvero a tutti gli elementi che si hanno a disposizione per costruire una frase.
- *Sintassi:* studia come gli elementi del dizionario possono essere collegati tra loro attraverso una struttura che permette di costruire frasi.
- *Semantica:* interpretazione di una struttura lessico-sintattica a cui si attribuisce un significato.
- *Pragmatica:* disciplina linguistica che si occupa del rapporto tra parole e contesto.
- *Ambiguità:* proprietà del linguaggio che permette di esprimere e comunicare con un numero basso di parole. Tuttavia aumenta la difficoltà nella comprensione di parole con più interpretazioni.

- *Polisemia:* fenomeno per cui una parola può esprimere più significati.
- *Omonimia:* fenomeno per cui una stessa forma ortografica e fonologica esprime più significati.

Altri aspetti del linguaggio:

- *Comunicazione:* strumento per condividere i significati all'interno della nostra mente.
- *Convenzione:* meccanismo con cui si veicola il contenuto semantico attraverso dei simboli.
- *Granularità:* dimensione che caratterizza i modi con cui vengono concettualizzate le situazioni che si vogliono descrivere, muta il significato della parola in base a dei dettagli.
- *Soggettività:* il linguaggio è un'approssimazione delle immagini mentali, quindi è soggetto a errori.
- *Similarità:* meccanismo innato che permette di inferire il significato di un termine sconosciuto riconducendolo a un termine conosciuto.
- *Esperienza personale:* insieme di tutti gli eventi della vita di una persona che formano la conoscenza di un singolo individuo.
- *Senso comune:* convenzioni che stabiliscono il significato che la collettività dà ad alcuni termini.
- *Cultura:* il significato di alcune parole è legato alla convenzione della cultura nella quale ci si trova.

Note:-

Queste definizioni creano un'ontologia, non c'è interesse per il significato specifico dei singoli concetti, ma al significato condiviso che gli si attribuisce.

2.2 Il Significato delle Parole

Filoni di pensiero:

- *Primitive:* per rappresentare il significato di una parola lo si frammenta in piccoli contenuti semantici atomici.
- *Relazioni:* il significato di una parola non è frutto di combinazioni atomiche di primitive universali, ma nasce dalla relazione con altre parole. Nessuna parola ha un significato intrinseco se non impiegata all'interno di un contesto lessicale.
- *Composizioni:* una parola prende significato sia quando è inserita in un contesto sia quando è composta con altre parole vicine.

2.2.1 Triangolo Semiotico

Definizione 2.2.1: Triangolo Semiotico

Il Triangolo Semiotico (fig: 2.1) è un modello del significato per cui qualsiasi concetto che si ha in mente è rappresentabile attraverso un triangolo i cui poli indicano rispettivamente il concetto, il referente e la rappresentazione.

Note:-

Il referente è anche chiamato fenomeno o istanza.
Il concetto è anche chiamato significato o interpretazione.
La rappresentazione è anche chiamata segno, termine, simbolo.

Corollario 2.2.1 Concetto

Corrisponde a ciò che si ha in mente senza utilizzare una convenzione.

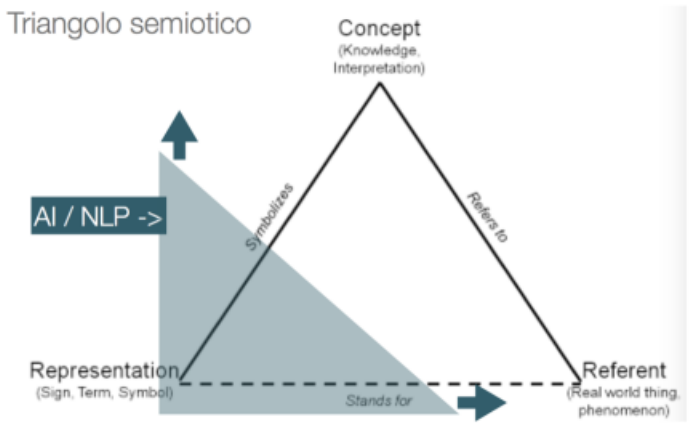

Figure 2.1: Triangolo semiotico.

Corollario 2.2.2 Rappresentazione

Si utilizza un simbolo convenzionale per comunicare il concetto.

Corollario 2.2.3 Referente

Un'istanza del concetto, ossia un elemento nel mondo reale.

Note:-

Per esempio il concetto di "gatto" in italiano e in inglese è lo stesso, ma la sua rappresentazione cambia (gatto/cat). Il referente è un qualsiasi gatto.

Domanda 2.1

Dove si collocano IA e NLP in questo triangolo e in quale direzione si muovono?

L'unico punto da cui si può partire è la rappresentazione perché nessun sistema informatico può prendere un concetto direttamente dalla nostra testa. Dall'insieme di testi presenti nel web si cerca di creare una concettualizzazione e si cerca di muoversi verso i referenti.

2.3 Multilinguismo e Granularità

2.3.1 Multilinguismo

Una delle sfide del NLP moderno è quella di trattare una pluralità di lingue, il ché è un'arma ha doppio taglio: è più difficile, ma fornisce molte più sfumature di significato. Analizzare un testo in più lingue permette di:

- Capire quali sono le informazioni semantiche più importanti, più certe o maggiormente condivise.
- Migrare informazioni semantiche da una lingua all'altra.

Osservazioni 2.3.1

Possibili problemi:

- I testi in lingue rare sono molto difficili da gestire, anche considerando che la maggior parte del web e dei testi scientifici è in inglese.
- Le lingue hanno sfumature che non possono essere tradotte direttamente.

2.3.2 Granularità

La Granularità può essere a livello di:

- *Parola:* complessità elevata.
- *Chunk:* composizione di parole (e.g. aggettivo + nome).
- *Discorso:* come per i chatBOTs.
- *Documento:* sistemi di sommarizzazione:
  - Estrattivi: estrapolano dal testo le parti più significative.
  - Astrattivo: generano nuove frasi a partire dal documento.
- *Collezione di documenti:* estrapolare gli argomenti principali (topic modelling).

2.4 Costruzione del Significato

2.4.1 Pustejovsky

Pustejovsky propose una teoria chiamata *generative lexicon* che utilizza una struttura basata su:

- *Argument Structure:* per esprimere il legame tra sintassi e semantica del concetto. In altre parola come mappare ciò che si vuole esprimere su un concetto mediante l'uso di lettere, parole e grammatica.
- *Event Structure:* per esprimere tutti i tipi di evento che coinvolgono quel concetto.
- *Qualia Structure:* per esprimere come sono definite le caratteristiche (qualia) di un concetto.
- *Inheritance Structure:* per collocare il concetto all'interno di una tassonomia per inferirne il significato.

Pustejovsky sostiene che per poter ragionare semanticamente in maniera precisa e completa abbiamo bisogno di formalizzare tutte queste strutture.

Definizione 2.4.1: Qualia

La qualia ha 4 ruoli:

- Costitutivo: esprime la parte di composizione del soggetto, riguarda il peso, la dimensione e le parti che lo compongono.
- Formale: esprime le caratteristiche che distinguono un concetto dagli altri dello stesso dominio.
- Telico: l'obiettivo o la funzione del concetto, il suo ruolo comportamentale.
- Agentive: tutte le entità (spesso umane) che rappresentano l'origine del concetto.

Note:-

Si tratta di una teoria formale in cui si assegna a ciascun elemento un ruolo e una struttura in base ai concetti definiti da Pustejovsky. In questo modo ogni frase può essere analizzata in modo formale. Il problema di questa teoria e la sua complessità che la rende difficile da implementare.

2.4.2 Hanks

Patrick Hanks enuncio una teoria del significato più semplice di quella di Pustejovsky: la *teoria delle valenze*. Questa teoria si basa sul concetto che il verbo sia la radice del significato: non esiste un'espressione di significato senza un verbo.

Definizione 2.4.2: Valenza

La valenza (fig: 2.2) è la cardinalità degli oggetti che compongono la struttura di cui il verbo è la radice. Un verbo può essere transitivo o intransitivo a vari livelli.

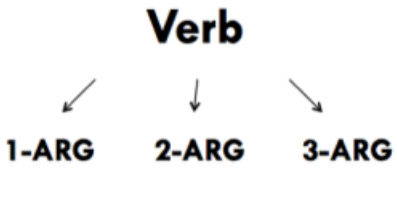

Figure 2.2: Valenze.

Note:-

Ogni valenza rappresenta un numero di argomenti chiamati slot e ogni possibile valore che possono assumere e chiamato filler.

Dati un verbo e una valenza si hanno (fig: 2.3):

- *Collocazione:* la combinazione di tutti i possibili filler.
- *Semantic Type:* delle macrocategorie che servono per raggruppare i vari filler.

Figure 2.3: Le due righe in verde rappresentano una valenza sintattica.

Domanda 2.2

Quali sono i Semantic Type? Quale deve essere il grado di generalizzazione?

- Non sempre si hanno sufficienti dati per tutte le parole, alcune sono rare e difficili da analizzare.
- I termini nei dati possono non sovrapporsi anche se sono simili.

3

Pre-LLM: Storia, Concetti e Task