ANNO ACCADEMICO 2024/2025

Tecnologie del Linguaggio Naturale

Teoria

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

Introduzione alle Tecnologie del Linguaggio Naturale

1.1 Prologo

La prima parte del corso sarà incentrata sulla linguistica computazionale generale, in cui ci si soffermerà sugli aspetti più tradizionali e linguistici[1]. In questa parte verrà anche trattato il parsing. Nella seconda parte si andranno a studiare la semantica lessicale e le ontologie. Infine, nella terza parte del corso si andrà a studiare NLP statistico e distribuzionale.

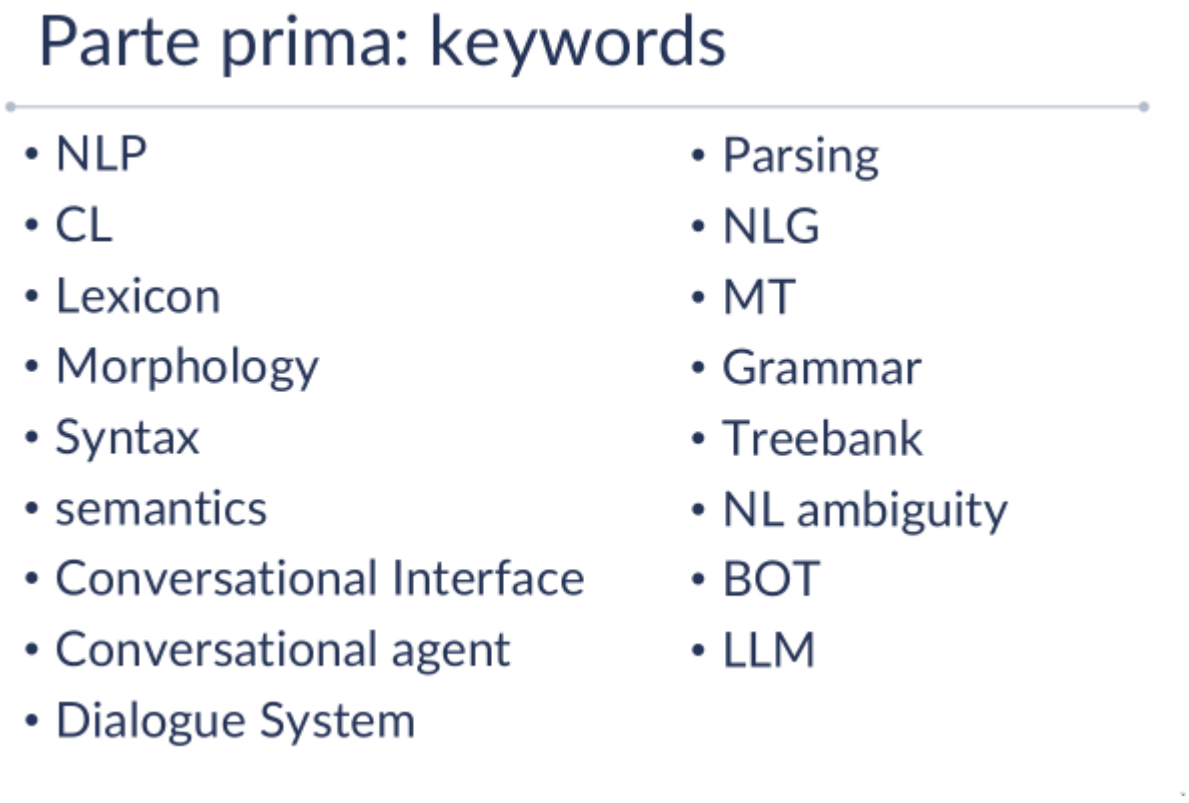


Figure 1.1: Il giorno prima dell'esame bisogna sapere cosa significano tutte queste parole :3

[1] Libro di riferimento: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. La prima e la seconda edizione, perché Jurafsky non riesce a finire il draft della terza :(

Le 4 ere della linguistica computazionale:

1. 1940 - 1969: primi tentativi.
2. 1970 - 1992: formalizzazione.
3. 1993 - 2012: apprendimento automatico.
4. 2013 - 2018: deep learning.

Note:-

Tutto cambiò nel 2018, quando NLP fu il primo successo su larga scala di rete neurale autosupervisionata.

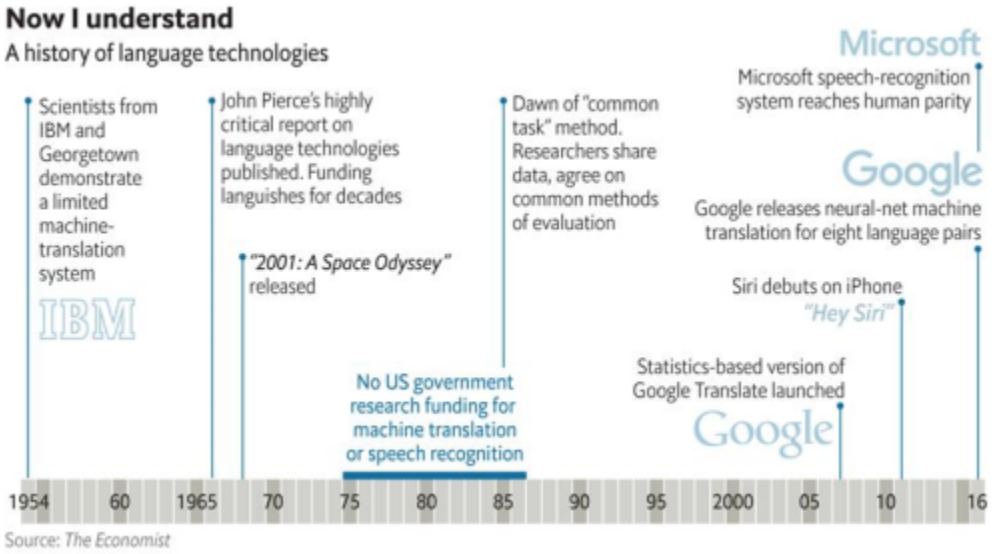


Figure 1.2: Il passato delle tecnologie del linguaggio naturale.

1.1.1 La Complessità del Linguaggio Naturale

C'è un legame tra linguaggio umano e intelligenza. Già Turing sosteneva che se si potesse parlare in un certo modo si fosse intelligenti (test di Turing). La differenza tra il linguaggio umano e un linguaggio di programmazione è l'*ambiguità*: C o Java non sono ambigui.

Il linguaggio umano:

- *Discretezza* (esistenza di elementi):
  - Api: Ritmo, orientamento, durata.
  - Esseri umani: Fonemi, morfemi, parole.
- *Ricorsività*:
  - Scimpanze: Gesti atomici.
  - Uomo: Gianni vede Pietro, Maria vuole che Gianni veda Pietro, Paolo crede che Maria voglia che Gianni veda Pietro.
- *Dipendenza dalla struttura*:
  - Non "una parola dietro l'altra" ma c'è una struttura: La ragazza parte, I ragazzi di cui mi ha parlato la ragazza partono.
- *Località*:
  - Gianni lo ha guardato.
  - Gianni ha detto che Pietro lo ha guardato.

Intelligenza e linguaggio nel il test di Turing:

- Possono le macchine pensare?
- Se riesco a parlare come un essere umano allora penso.
- Gioco dell'imitazione: un giudice deve capire se quello che ha davanti è un uomo oppure un computer.

Note:-

Ci sono una serie di obiezioni a questo test: teologia, matematica, coscienza, etc.

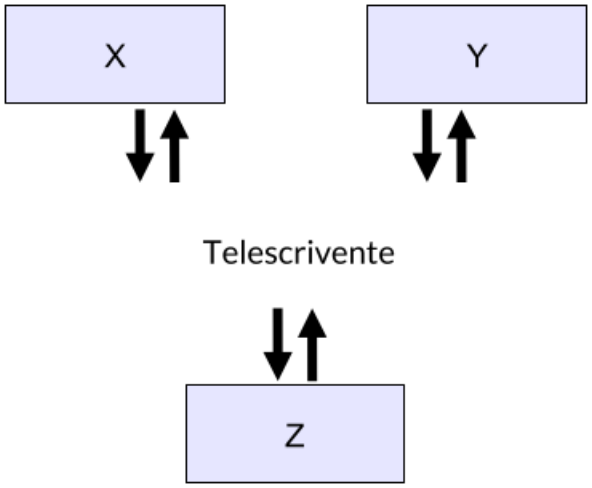


Figure 1.3: Il gioco dell'imitazione.

Nel 1966, Weizenbaum crea Eliza. Una macchina in grado di "comprendere" e ingannare gli esseri umani.

Note:-

Il punto debole del test di Turing e di Eliza è il giudice: se è coinvolto emotivamente potrebbe far passare un computer per un essere umano[a].

[a] Blade runner moment

Definizione 1.1.1: Winograd Schema

Evoluzione del Turing test: un test a scelta multipla che utilizza domande con una specifica struttura. In questi test gli esseri umani sono molto bravi a rispondere, i computer no.

Note:-

Rimuove il giudizio, quindi tecnicamente più accurato.

Corollario 1.1.1 Captcha

Un test di Turing inverso per capire se l'interloquitore è umano. Non c'è linguaggio, ma riconoscimento cognitivo.

Corollario 1.1.2 Voight-Kampff Test

Test in Blade runner basato sulle emozioni, evoluzione del test di Turing.

1.1.2 I Livelli di Conoscenza del Linguaggio

HAL 9000, in "2001: Odissea nello spazio" mostra un esempio di comunicazione.

Domanda 1.1

Come fa HAL a rispondere?

- Riconoscimento vocale.
- Comprensione del linguaggio naturale.
- Generazione del linguaggio naturale.
- Sintesi vocale.
- Recupero ed estrazione di informazioni.
- Inferenza.

Livelli della conoscenza:

1. Il suono: HAL deve essere in grado di analizzare e produrre dei segnali audio che contengono le parole: foni e fonemi.
2. Le parole: HAL deve essere in grado di riconoscere le singole parole.
3. Raggruppare le parole: HAL deve essere in grado di distinguere la struttura della frase.
4. Significato: HAL deve conoscere il significato delle singole parole e deve essere in grado di comporre questi significati per trovare il significato complessivo della frase.
5. Contesto e scopi: HAL deve avere delle conoscenze del mondo che gli permettono usare il linguaggio in maniera contestuale: *I'm afraid, I can't* invece di *I won't*.
6. Conversazione: HAL deve avere deve essere in grado di conversare, dando delle risposte e facendo delle domande pertinenti al discorso.

A ogni livello corrisponde una parte del linguaggio:

1. Fonetica e Fonologia: lo studio del suono della lingua.
2. Morfologia: lo studio delle parti significative delle parole.
3. Sintassi: lo studio sulla struttura e sulle relazioni tra le parole.
4. Semantica: lo studio del significato.
5. Pragmatica: lo studio di come il linguaggio è usato per compiere goal. Il passivo serve per mettere in luce/enfatizzare alcune parti della frase.
6. Discorso: lo studio delle unità linguistiche rispetto alla singola dichiarazione.

Note:-

Jurafsky è un chad nerd.

1.1.3 Strutture Linguistiche e Ambiguità

Analizzando i vari livelli si trovano diverse *strutture linguistiche*.

Definizione 1.1.2: Struttura Linguistica

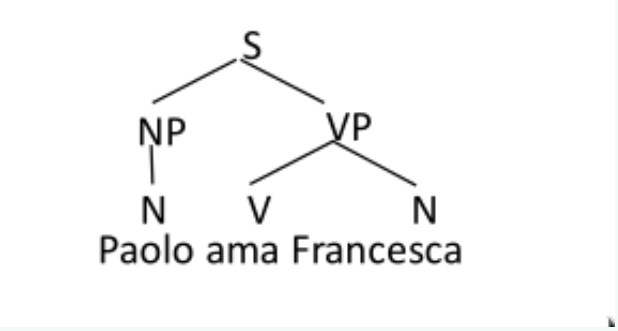
Una struttura è un insieme su cui è definita una relazione:

- Relazione fonetico-fonologica sull'insieme dei foni-fonemi.
- Relazione morfologica sull'insieme dei morfemi.
- Relazione sintattica sull'insieme delle parole.
- Relazione semantica sull'insieme dei significati delle parole.
- Relazione pragmatica sull'insieme dei significati delle parole e sul contesto.
- Relazione "discorsale" sull'insieme delle frasi.

Note:-

Ci sono relazioni tra i componenti della frase. Inoltre le relazioni cambiano a seconda della lingua.

Esempio 1.1.1 (Struttura Sintattica)



Definizione 1.1.3: Ambiguità

Il linguaggio naturale presenta frasi che possono essere interpretate in modi differenti.

Esempio 1.1.2 (Ambiguità)

"I made her duck"

- Ho cucinato una papera per lei.
- Ho cucinato una papera che apparteneva a lei.
- Ho creato una papera con la stampante 3D e gliel'ho data a lei.
- Ho fatto abbassare la sua testa.
- In Harry Potter[a]: Ho trasformato lei in una papera.

[a] Rowling merda.

Osservazioni 1.1.1

- Le parole "duck" e "her" sono morfologicamente ambigue nella loro parte del discorso. "Duck" può essere un verbo o un nome, "her" può essere un pronome dativo o possessivo.
- Il verbo "make" è sintetticamente ambiguo: può essere transitivo o intransitivo.
- Inoltre "make" è anche semanticamente ambiguo: può significare creare o cucinare.
- In una frase parlata c'è un altro livello per cui "her" può essere udito come "eye" e "make" come "maid".

Note:-

Essere ambigui permette di essere brevi e coincisi.

Altre proprietà notevoli del linguaggio:

- Linguaggio non standard, evolve nel tempo.
  - Scialla bros → chill → è easy.

- Segmentazione.
  - Il treno Torino San Remo.
- Locuzioni, spesso l'interpretazione non è composizionale.
  - Pollica verde.
- Neologismi.
  - Twettare[2]
- Conoscenza del mondo.
  - Lucia e Carola erano sorelle.
  - Lucia e Carola erano madri.
- Meta-linguaggio.
  - La prima cosa bella ha avuto un grandissimo successo.

1.1.4 Lo Stato dell'Arte

- 1976: In Canada un sistema riesce a stampare due bollettini meteo in due lingue diverse.
- BabelFish, di Yahoo, era un sistema "a regole" di trascrizione automatica, basato su Systran.
- 2011: IBM costruisce un supercomputer per battere un essere umano a Jeopardy, Watson.
- Tecnologie vocali: Speech Recognition, TextToSpeech, HTML5 Speech API (pagine web vocali).

[2] Musk merda.

Note:-

Dopo sette milioni e mezzo di anni Pensiero Profondo fornisce la risposta: "42"[a].

[a]Guida Galattica per gli Autostoppisti.

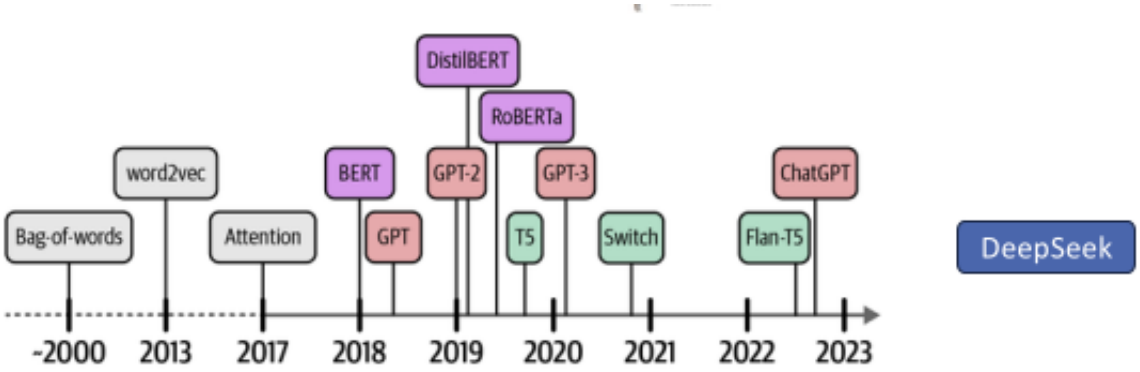


Figure 1.4: LLM. Tratto da "Hands-On Large Language Models", uscito nel Dicembre del 2024.

Note:-

Well, Deepseek è open source e funziona meglio di ChatGPT (a patto che non chiedi cosa sia successo a piazza Tienanment nel 1989).

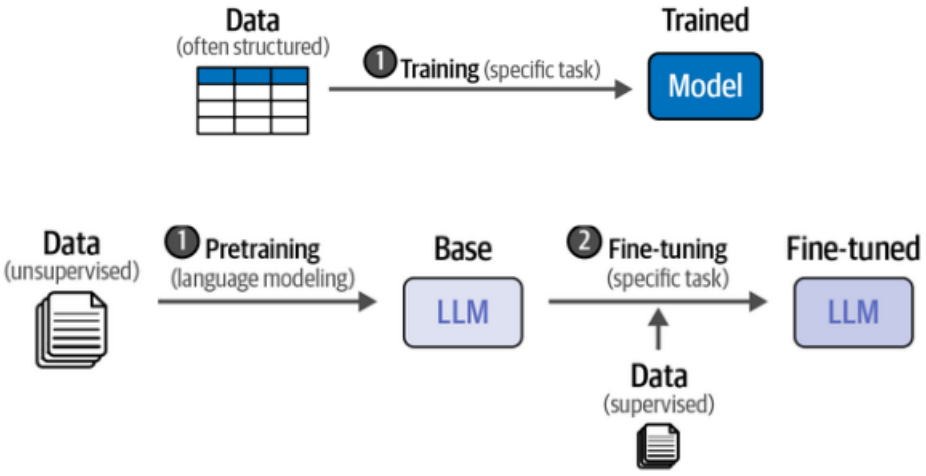


Figure 1.5: Shifting di paradigma dovuto al Machine Learning.

Definizione 1.1.4: AI Generativa

Modello di linguaggio di reti neurali multi-task basate sui transformer addestrati su una grande quantità di dati utilizzando self training e feedback umano.

- Modello di Linguaggio: Text prediction → T9.
- Multi-task: Google Translator, Siri.

Domanda 1.2

Come fare un LLM (M. Lapata)?

1. Collezionare una grande quantità di dati.
2. Chiedere al LLM di predirre la nuova parola in una frase.
3. Ripetere il tutto.

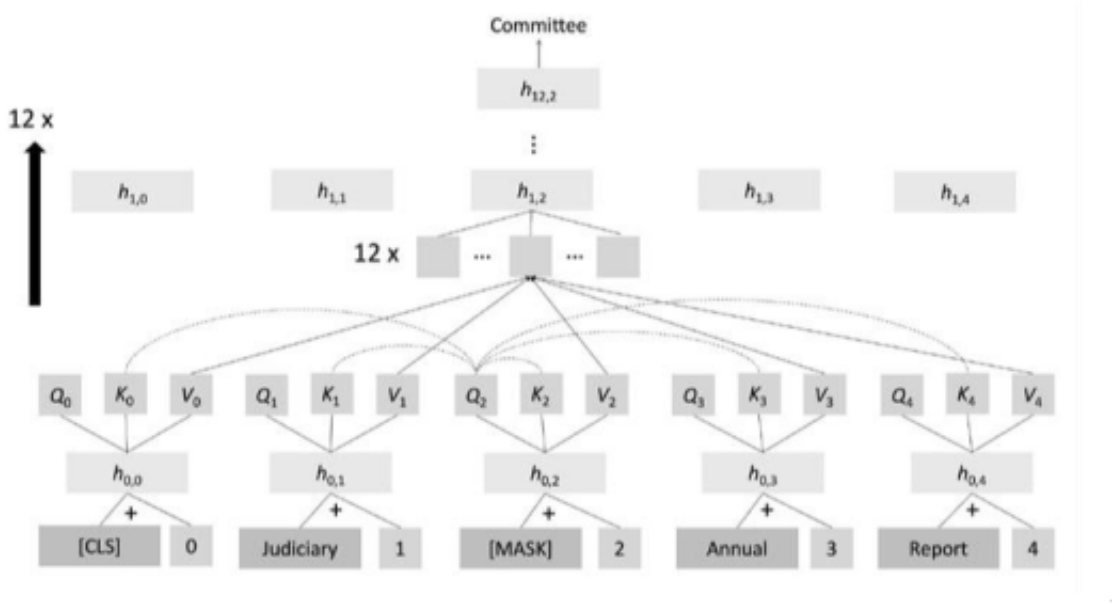


Figure 1.6: Auto addestramento di una rete neurale.

Domanda 1.3

Come usare un LLM?

- Sintonizzazione a grana fine.
- Prompting.

Si può usare un LLM per:

- Search Engine.
- Writer/Code assistant.

Note:-

Noam Chomsky odia questi sistemi. Secondo lui servono per evitare l'apprendimento.

DeepSeek:

- Apprendimento rinforzato automatico (senza essere umani).
- Meno costoso → politicamente importante.

***Il problema fondamentale*:** Convertire una frase o un testo in una forma che permetta l'applicazione di meccanismi di ragionamento automatico.

1.2 I Livelli Linguistici

1.2.1 Da Frase a Significato

Problema: Convertire una frase o un testo in una forma che permetta l'applicazione di meccanismi di ragionamento automatico.

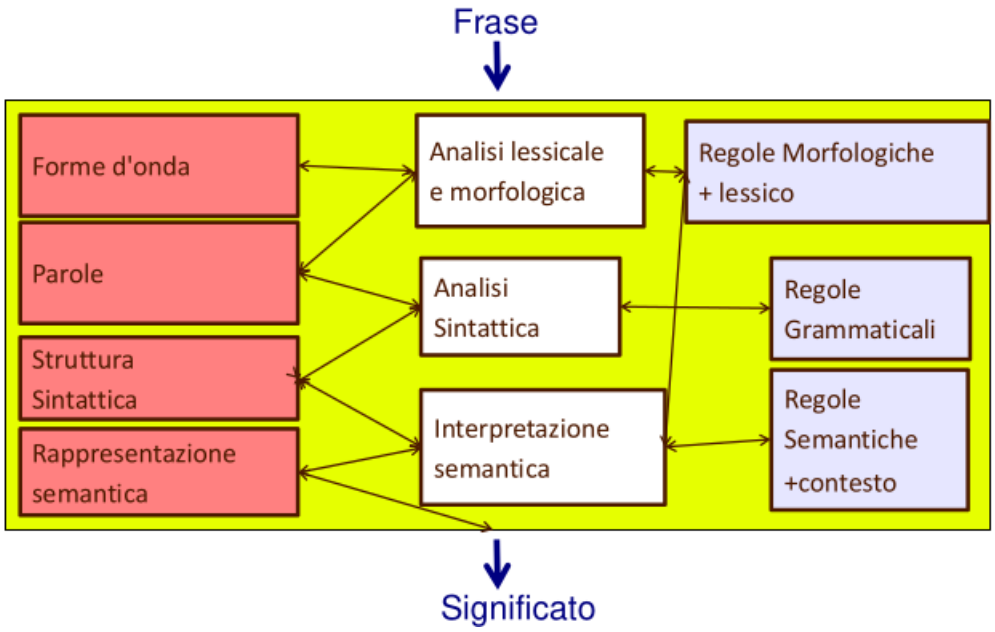


Figure 1.7: Passaggio da frase a significato.

Note:-

Però la situazione non è così semplice. Bisogna capire come funzionano i moduli e come comunicano

Nella linguistica computazionale c'è una divisione tra *regole* e *statistica*:

- Rules-driven.
- Data-driven.

Note:-

Steedman sostiene che i due aspetti dovrebbero convivere tra loro (2008). Evitare il regionamento tribale.

Domanda 1.4

Quando finisce una frase?

Definizione 1.2.1: Sentence splitting

Task in cui si deve capire quando una frase finisce.

- "!", "?" → Okay, pongono fine alla frase.
- ".":
  - Fine frase.
  - Abbreviazione (Doc., Mx.).
  - Numeri (0.2).

Domanda 1.5

Quindi come si costruisce un classificatore binario che decida EoS (End of String) o not EoS?

- Si possono scrivere regole a mano:
  - Espressioni regolari.
  - Tokenizer (FA) e regole.
- Addestrare un sistema di machine learning.

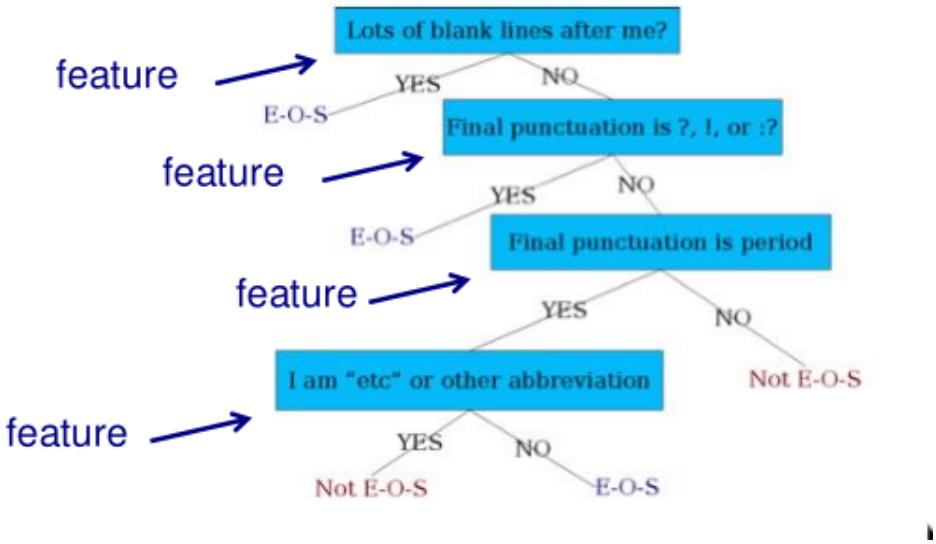


Figure 1.8: Albero di decisione.

Features più complesse:

- Caso di parole con ".".
- Caso di parole dopo ".".
- Features numeriche:
  - Lunghezza di parole con ".".
  - Probabilità che una parola con "." avvenga alla fine della frase.
  - Probabilità che una parola dopo "." avvenga all'inizio di una frase.

Domanda 1.6

Cos'è davvero un albero di decisione?

- Una serie di IF-THEN-ELSE incapsulati.
- Due possibilità per costruirlo:
  - *By-hand*: solo in contesti semplici.
  - *Machine learning*: su un training corpus.
- Il punto cruciale è la scelta delle features.

Osservazioni 1.2.1

- In questo corso ci concentreremo sullo studio delle feature linguistiche.
- In alcuni casi l'approccio by-hand verrà privilegiato poiché è didatticamente più chiaro/semplice e poiché è più semplice verificarne la fondatezza cognitiva mediante introspezione.

Domanda 1.7

Nei sistemi end-to-end cosa sono le feature linguistiche?

Definizione 1.2.2: Features Linguistiche Neurali

L'architettura neurale, ovvero il numero e il tipo di connessioni, codifica in maniera *implicita* le features linguistiche.

Note:-

La ricerca, in questo caso, si focalizza su quale scelta architetturale è più adatta alla modellazione implicita del fenomeno linguistico e alla creazione del corpus di training.

1.2.2 Il Livello Morfologico e l'Analisi Lessicale

Il lessico è fondato sul concetto di *parola*.

Domanda 1.8

Che cos'è una parola?

- Intuitivamente è una sequenza di caratteri delimitata da spazi o punteggiatura.
- Sequenze di più parole, Es. passammela = passa a me essa.
- Le parole hanno un significato unitario (semantica lessicale), ma volte sequenze di parole hanno un significato unitario. Es. di corsa, by the way.
- In altre lingue il problema è più grave.
  - In tedesco: Lebenversicherungsgesellschaftangestellter = impiegato di una società di assicurazione sulla vita.
  - In inglese: Wouldn't? = Would not.

Presenza di suffissi:

- CAPITANO (forma non declinabile).
- CAPITAN + O (nome o aggettivo o forma del verbo capitanare).
- CAPIT + ANO (forma del verbo capitanare).

Note:-

Non c'è una forma giusta a priori, ma c'è una forma giusta in base al contesto.

Definizione 1.2.3: Forme Composte

Generalmente una parola contenuto più una (o più) parole funzione.

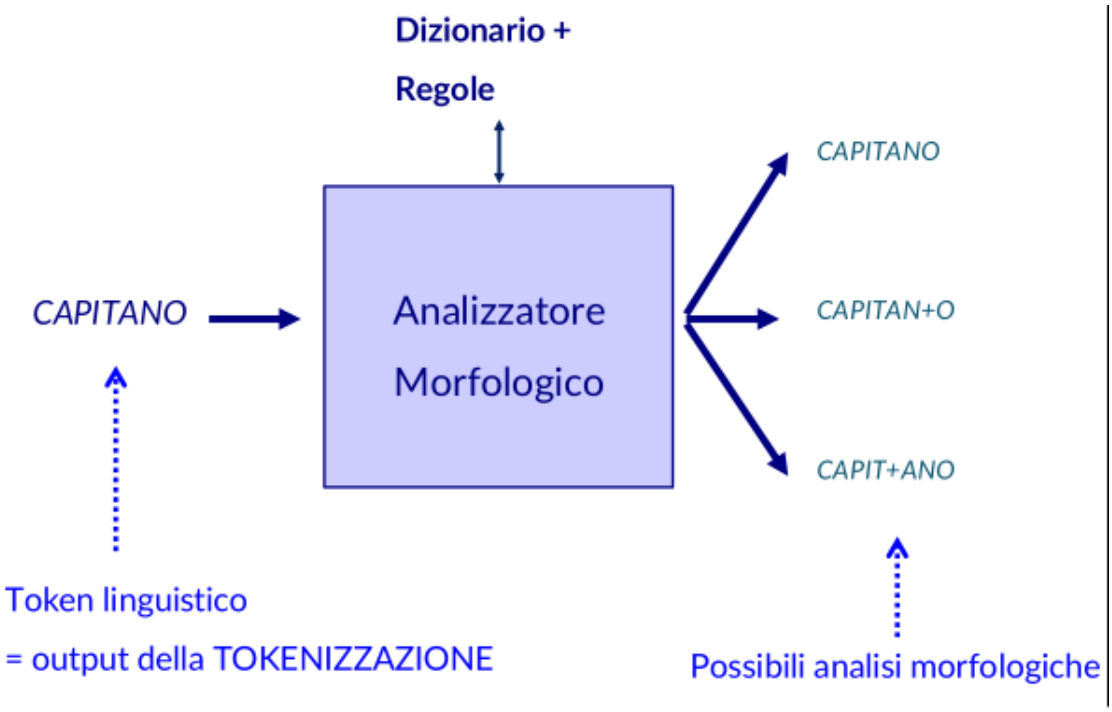


Figure 1.9: Analizzatore morfologico.

Esempio 1.2.1 (Forme composte)

STAMPAMELO:

- STAMP è una radice verbale.
- A è un suffisso verbale.
- ME e LO sono forme pronominali[a].

[a]Per triggerare gli Alt-Right.

Definizione 1.2.4: Forme Multiple

Le diversi componenti sono nel dizionario ma la semantica non è composizionale.

Esempio 1.2.2 (Forme multiple)

- Più o meno: puntatore tra le parole per recuperare la giusta semantica.
- Prendere un abbaglio: rimandare all'interprete semantico.

Definizione 1.2.5: Lemmatizzazione

Trasformare un lemma in forma normale.

Note:-

La forma normale non è stabile nel tempo.

Definizione 1.2.6: Stemming

Estrarre la forma radice (detta tema) da una parola.

Definizione 1.2.7: Paradigmatico

Si cambia una parte della parola con una equivalente si ha una frase morfologicamente corretta.

Definizione 1.2.8: Sintagmatico

I rapporti che intercorrono tra gli elementi che si succedono nella frasei rapporti che intercorrono tra gli elementi che si succedono nella frase.

Nome:

- Persone, oggetti, luoghi.
- Proprietà sintagmatiche:
  - Comparire dopo gli articoli.
  - Avere un possessivo.
  - Avere un singolare o un plurale.
- Comuni, propri, di massa, contabili.

Verbo:

- Eventi, azioni, processi.
- Molte forme morfologiche.
  - Tempo.
  - Modo.
  - Numero.
- Tante categorie (ausiliari, modali, copula, etc.).

Aggettivi:

- Proprietà.

Avverbi:

- Modificano qualcosa, spesso verbi, ma anche altri avverbi o intere frasi.

Note:-

Nomi, verbi, aggettivi e avverbi sono *di contenuto*, che puntano a oggetti reali.

Definizione 1.2.9: Classi Aperte

Classi che aumentano o scompaiono nel tempo costantemente (nomi, verbi, aggettivi, avverbi).

Definizione 1.2.10: Classi Chiuse

Classi che aumentano o scompaiono con tempi lunghissimi.

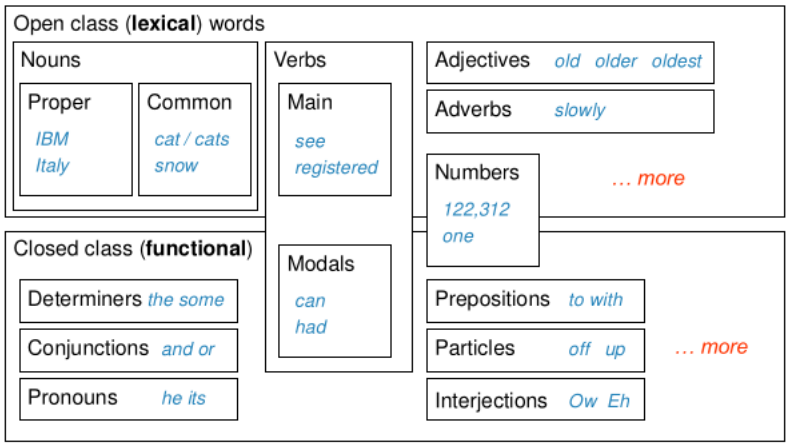


Figure 1.10: Parti aperte e parti chiuse.

Google Unviversal PoS: 12 PoS: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and particles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (a catch-all, e.g. abbreviations and foreign words).

1.2.3 Il Livello Sintattico

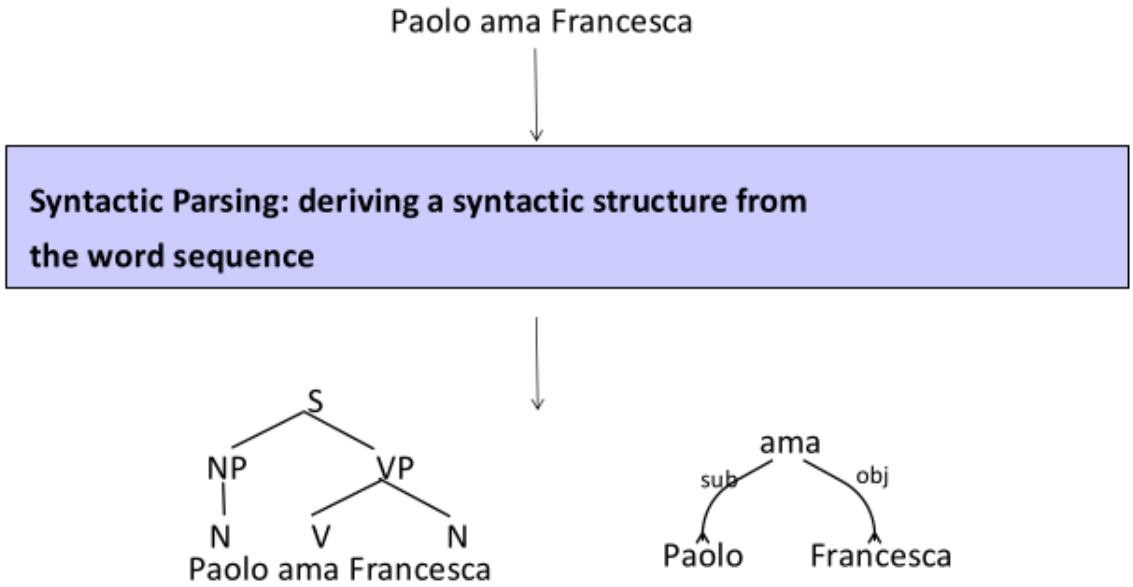


Figure 1.11: Parsing sintattico.

Note:-

Le due alternative derivano da prospettive diverse:

- Quella a sinistra è la struttura sintagmatica (o a costituenti).
- Quella a destra è la struttura a dipendenze.

Entrambe le alternative sono equivalenti.

1.2.4 Il Livello Semantico

1.2.5 Il Livello Pragmatico e del Discorso

2
Test2