Che cos'è il Natural Language Processing video

Che cos'è il Natural Language Processing: in italiano chiamato anche elaborazione del linguaggio naturale o processamento del linguaggio naturale si tratta dello studio con metodi computazionali, quindi con algoritmi e computer del linguaggio naturale: cioè della lingua o delle lingue parlate da noi esseri umani.
Un esempio di linguaggio non naturale, ad esempio, è il linguaggio di programmazione. mentre il NLP si occupa del linguaggio naturale.

Ci sono diversi termini che si riferiscono a concetti simili o comunque tangenti.

Quindi cerchiamo di toglierci qualche dubbio terminologico sin dall'inizio. Si parla sia di linguistica computazionale *Computational Linguistics* sia di Natural Languagne Processing spesso usandoli come sinonimi, non ci sono delle definizioni scritte nella pietra, nella letteratura si trovano diversi punti di vista sulla questione. Un modo di distinguere i due approcci (se vogliamo) è quello di dire che la linguistica computazionale è lo studio del linguaggio naturale quindi linguistica effettuata con metodi computazionali mentre quando parliamo di elaborazione del linguaggio naturale (NLP) possiamo più porre l'accento sull'insieme delle **tecniche** computazionali che processano dei dati e nel caso particolare questi dati sono fatti da linguaggio naturale.

Il **Natural Language Processing** è realtà un intero campo di studio di applicazione di sviluppo ed è fatto da tanti Task: tanti sotto-problemi.
Storicamente ognuno di questi problemi ha delle intere comunità, e a volte degli eventi scientifici che se ne occupano in maniera più o meno interconnessa.
Un elenco di alcuni tra i Task più popolari del NLP (un elenco completo sarebbe molto più lungo di così)

Task della NLP troviamo:

- **L'analisi del sentimento**
- **Riconoscimento delle entità nominali**: quindi delle persone degli oggetti artefatti luoghi nel testo
- **Role Labeling** Quindi quando ci interessa sapere chi ha fatto cosa in un certo testo
- L'analisi sintattica
- L'analisi grammaticale e così via.

Dal punto di vista della linguistica, storicamente in linguistica tutti questi sotto problemi di analisi del linguaggio sono organizzati in una sorta di gerarchia dove la metafora vuole che vedete la freccia che va verso il basso indicava la superficie come superficie si intende il linguaggio vero e proprio quello che troviamo scritto o parlato mentre verso l'alto in questa sorta di di stack di livelli di analisi troviamo livelli via via **più astratti** dove andiamo a vedere fenomeni via via più lontani dalla superficie del linguaggio.
Una suddivisione di massima di questi livelli di analisi del linguaggio naturale sono dal basso verso l'alto quello:

della **morfologia** cioè lo studio della della forma delle parole delle frasi di morfemi e poi andando più verso la stazione la **sintassi, semantica pragmatica** e sulla destra della slide i Task che ho elencato lì come esempio Sono all'incirca orientati nella stessa maniera del grafico sulla destra.

Andiamo a vedere nel resto di questa lezione più in dettaglio i diversi livelli di analisi linguistica per darvi un'impressione di quanto **diversi** siano i **problemi** pur rimanendo l'input Se volete del problema lo stesso delle espressioni linguaggio naturale si possono guardare da diversi punti di vista e focalizzare l'attenzione su diversi fenomeni.

TOKENIZZAZIONE
Partiamo dal basso, dalla superficie quindi dei Task di NLP più vicini alla superficie del delle espressioni linguaggio naturale: la tokenizzazione è un tipico esempio di primo Task in questa gerarchia di analisi dato organizzazione è il problema di dividere un testo arbitrario in frasi e anche di dividere le frasi stesse in parole o comunque elementi che abbiano un senso unitario, tipicamente si parla di parole o di o di segni di punteggiatura ad esempio Emoji in qualche caso e qui c'è un esempio dove vediamo che la frase rappresentata come una stringa Viene divisa da un processo di tokenizzazione in una lista di token, si chiamano che in questo caso corrispondono proprio alle parole “This is a sentence” e l'ultimo elemento della lista È proprio il punteggiatura il punto.

Sulla tokenizzazione come su tutti gli altri tasti di cui parleremo ampia letteratura Cioè tutta una storia di approcci che sono si sono susseguiti negli anni e considerata è stata considerata un certo punto in anni recenti un problema Tutto sommato risolto Nel senso che gli algoritmi più nuovi riescono ad ottenere delle performance molto vicine alla perfezione anche se in realtà è stata un po' messa in crisi questa questa osservazione a seconda del tipo di input che noi diamo un tokenizzatore e in realtà degli approcci già molto semplici basati per esempio sul dividere le parole in base a allo spazio Oppure ai segni di punteggiatura ottengono una buona performance Il problema è questo è un tipo di uno scenario che si vede spesso in molti Task di nlp e che diciamo

l'80 E il 90% del problema è relativamente facile o molto facile come in questo caso ma il restante 10 o 20% delle casistiche diventa esponenzialmente difficile in problemi tipici nel caso della tokenizzazione sono per esempio le abbreviazioni gli acronimi i titoli Se pensate a dottor qualcuno con il segno di punteggiatura stabilire che quel punto alla fine dell'abbreviazione sia un la fine di una frase o semplicemente faccia parte del token precedente può essere un problema non banale e ho aggiunto lì una curiosità che in realtà i primi approcci basati proprio sul Deep Learning per la tokenizzazione sono già Cominciate ad apparire una decina di anni fa circa poi metterò alcuni materiali per approfondire tra i materiali del Corso.

LEMMATIZZAZIONE

proseguendo in questa carrellata veloce di alcuni Task di NLP vediamo la Lemmatizzazione.

La Lemmitazione è il problema Il Task di a partire dalla forma che si trova nel testo di una certa una certa parola trovare la sua forma base quello che si chiama il Lemma il Lemma è la versione di una parola che si trova nel dizionario.

Ad esempio, in italiano per i verbi il Lemma è dato dalla loro forma All'infinito. E qui di nuovo c'è un esempio vedete che il token is Che voce del verbo essere in inglese to be viene l'emmatizzato come B la lemmatizzazione è oltre a essere un problema interessante di per sé linguistica ha anche una sua funzione per livelli ulteriori di analisi ad esempio è un primo diciamo gateway nella direzione di passare dalle parole dei concetti perché sono i lemmi sono più astratti se vogliamo delle forme specifiche delle parole e anche in questo caso ci sono stati storicamente sono stati presentati tanti approcci i primi basati anche su semplicemente delle lunghe liste dei vocabolari o dei lemmari delle lunghe liste di forme di parole e associate ai loro lemmi e anche qui ci sono relativamente pochi casi che sono però molto difficili e tipici problemi sono quella della dell'ambiguità grammaticale delle parole, una certa parola può essere un verbo o può essere un sostantivo ad esempio a seconda del suo contesto.

Un algoritmo che faccia bene la Lemmatizzazione deve essere in grado di disambiguare i casi del genere.

Altri problemi possono essere quelli di incontrare nuove parole parole che in una risorsa o in un algoritmo magari non sono conosciuti e quindi bisogna in qualche modo generalizzare.

PART-OF-SPEECH TAGGING

il Task di *Part-of-speech tagging:* a parti del discorso è un po' un po' quello che alle scuole elementari Noi studiamo come analisi grammaticale.

Cioè quello di etichettare ogni parola di una frase di un'espressione con la sua parte del discorso che sono tipicamente sostantivo aggettivo verbo avverbio e cose del genere e

dico tipicamente perché non c'è una sola grammatica: in realtà ci sono sono state proposte tante diverse teorie della sintassi della grammatica e di conseguenza tanti set diversi di Label a cui fa riferimento in un algoritmo che faccia parte dello Speech tagging è stato sempre uno dei tasti principali soprattutto fino ai primi anni 2000 della linguistica computazionale ed è uno di quei Task sui quali riflettendoci bisogna porsi anche il problema di Come valutare la bontà di un algoritmo che risolve un Task del genere.

E stato fatto montare in uno scritto storico che se noi abbiamo un algoritmo che ottiene un'accuratezza del 95% in parte tagging e un numero che può sembrare molto alto ma in realtà in quanto il 5% vuol dire che in una frase di 20 parole che è una lunghezza normale per una frase in linguaggio naturale ce n'è statisticamente almeno una sbagliata Il che non è normalmente accettabile noi non possiamo accettare un programma che analizzi delle frasi sbagliandone praticamente sempre almeno commettendo almeno sempre un errore ad ogni frase.

PARSING

Altro Task tipico sul quale è stata pubblicata intere biblioteche di letteratura è quello del cosiddetto *parsing* più precisamente *parsing* in sintattico il sintattico è quel Task che ha come input sempre un'espressione linguaggio naturale ma come output ha

un intero albero un albero nel senso informatico del termine che ne rappresenta la struttura sintattica anche qui non entro troppo nel dettaglio delle diverse teorie che sono state che sono state proposte quindi l'esempio vediamo un albero che segue la il fondamento teorico della teoria delle Universal dipendentis dove

si vedono le relazioni tra le diverse parole perché il party in sintattico o sintassi e proprio lo studio delle relazioni tra diverse parole e della loro relazione gerarchiche in particolare motivo anche qui sono descritte formalmente come come un albero.

SENTIMENT ANALYSIS

vediamo ancora qualche altro Task più a livello astratto quindi qui siamo già nell'area linguistica della pragmatica dell'analisi del dell'intenzione comunicativa di un'espressione linguistica la Sentiment Analysis: un Task che negli ultimi anni è diventato sempre più popolare c'è molta attenzione anche da parte delle aziende di altre attori su la vera modelli con buone performance per sentimentalis e

si tratta di individuare a partire da un'espressione linguistica, un linguaggio naturale il suo sentimento nel senso proprio: il sentimento espresso da quell'espressione che può essere semplicemente inteso come un sentimento positivo o negativo o neutro oppure si può andare più in dettaglio ci sono molte varianti ovviamente di ognuno di questi Task in questo caso si può andare a vedere ad esempio il alcun tipo di sentimento espresso

nei confronti di certi specifici aspetti.
Questo è quello che avviene negli studi computazionali su le review per esempio di prodotti e di servizi quando vedete nel nelle pagine di piattaforme che propongono ad esempio alberghi oppure piattaforme di e-commerce spesso ci sono indicazioni su come avvengono percepiti certi servizi e prodotti ma divise per alcuni specifici aspetti

oppure si può andare più a grana fine nella nella scoperta delle Diverse emozioni espresse da un certo testo in linguaggio naturale. Quindi se un'espressione evoca paura o anticipazioni o gioia o rabbia e così via.

Dal punto di vista formale dal punto di vista computazionale sentimentale Esiste un tipico esempio ma ce ne sono tanti altri di text classification quindi classificazione del testo E questa è una tipologia di Task che ritornerà poi spesso perché sono tanti problemi che possono essere inquadrati se vogliamo in questa maniera.

TEXTUAL ENTAILMENT
ultimo esempio credo *Textual Entailment*: è un Task a livello semantico Quindi quando andiamo ad analizzare il significato delle parole contenute in una in una o più espressione linguaggio naturale e un Task molto importante.

Tuttora non risolto assolutamente nel senso che anche i migliori algoritmi i migliori sistemi anche di Deep Learning sono ancora abbastanza lontani dal dall'ottenere buone performance su questo tipo di Task ed è il Task di individuare automaticamente Se una frase rispetto a un'altra frase su antecedente e in contraddizione oppure segue logicamente oppure semplicemente non c'è correlazione tra tra queste due frasi.

c'è un esempio la premessa è che le classi sono spese le lezioni sono sospese in estate e quindi il Task è quello di esprimere un giudizio sulla frase: “There is no class in July” → non ci sono lezioni a luglio.
In questo caso c'è una relazione di Entailment perché la premessa porta logicamente come conseguenza alla conclusione e ma dal punto di vista computazionale non è assolutamente un problema banale, perché c'è bisogno che un sistema ad esempio sappia che “July” è il nome di un mese dell'anno, che quel mese si trova in estate, così via.
Quindi vedete che c'è tutto un livello di conoscenza extra che non si trova strettamente parlando nell'input linguistico Ma noi esseri umani l'abbiamo per la nostra esperienza.

Quindi se vogliamo che il sistema automatico sia in grado di risolvere lo stesso tipo di problemi di della lingua deve essere in qualche modo equipaggiato con questo tipo di conoscenza di diciamo di senso comune in questo caso o conoscenza del mondo e così via