1.1 Risorse linguistiche

1.2 Che cosa sono le risorse linguistiche?

Come sappiamo, molte delle operazioni che rientrano nell'analisi dei testi - come la suddivisione in token, il POS-tag e l'estrazione di entità nominate - oggi utilizzano programmi che contengono algoritmi di apprendimento automatico.

Questi algoritmi riescono a derivare le regole da applicare a partire da grandi moli di dati che fungono da esempio.

Si tratta di ricchi corpora, già elaborati da un operatore umano, costituiti da testi scientifici, articoli di cronaca o altri documenti rappresentativi di un determinato dominio di conoscenza.

I corpora rientrano nella categoria delle risorse linguistiche, che si usano per costruire, ampliare, rendere operativi e valutare modelli, algoritmi, componenti e sistemi per il trattamento automatico della lingua.

In generale, una risorsa linguistica è costituita da tre componenti:

1.un insieme di dati linguistici;

2.una serie di annotazioni o altre rappresentazioni formalizzate, come i POS-tag applicati ai singoli token;

3.uno strumento software che gestisce i dati.

Le risorse linguistiche sono costituite:

•da collezioni di dati linguistici originali, come le produzioni vocali o testuali che compongono un corpus,

•o dai dati che ne vengono derivati mediante opportune elaborazioni, come quelli che costituiscono un lessico.

I dati originali possono essere annotati allo scopo di descriverli o di darne una rappresentazione formalizzata.

Per esempio una registrazione vocale può essere accompagnata dalla sua trascrizione e in un corpus di testi i token possono essere annotati con i POS-TAG.

Inoltre le risorse linguistiche sono spesso accompagnate dagli strumenti usati per crearle, mantenerle o utilizzarle.

Oltre ai corpora, le risorse linguistiche comprendono lessici, dizionari enciclopedici, terminologie, ontologie e grammatiche.

Approfondiamole, partendo proprio dai corpora…

1.3 I corpora

Con “corpora” (plurale di “corpus”) si indicano raccolte di testi selezionati, orali o scritti, prodotti in contesti comunicativi reali: registrazioni di discorsi, articoli, documenti di varia natura.

C’è una branca della linguistica - chiamata, appunto, “linguistica dei corpora” - che usa strumenti di analisi quantitativa e statistica per esplorare le regolarità linguistiche che emergono dai testi e permettono di descrivere la struttura del linguaggio.

Ma non basta mettere insieme un po’ di testi per creare un corpus realmente utile.

Perché una raccolta di testi costituisca un corpus come lo intendiamo noi è necessario che:

•sia conservata in formato digitale, per consentire l’elaborazione automatica;

•se possibile, sia corredata di strumenti di consultazione informatici;

•sia molto ampia, per costituire un campione significativo del modo di parlare o scrivere in un determinato contesto e/o su un determinato argomento;

•sia sottoposta ad alcune particolari elaborazioni, come la normalizzazione, l'annotazione e l'allineamento.

Vediamo di cosa si tratta...

1.4 Caratteristiche dei corpora: la normalizzazione

I corpora, come abbiamo detto, sono patrimoni di grande valore per la linguistica e per questo è importante condividerli.

Condividere vuol dire poter confrontare i risultati e - perché no? - ammortizzare i costi.

Ma cosa significa "normalizzare" un corpus?

La normalizzazione può riguardare diversi aspetti, tra cui:

•la codifica, che deve seguire una modalità standard, che tipicamente è l'uso di Unicode con formato UTF-8;

•la granularità, cioè la dimensione tipica dei documenti, perché un articolo di giornale è diverso da un romanzo di centinaia di pagine;

•i criteri per individuare un documento all'interno del corpus;

•la modalità di annotazione, di cui parleremo fra poco.

1.5 La TEI, Text Encoding Initiative

Un'importante iniziativa di normalizzazione è la TEI, Text Encoding Initiative, lanciata nel 1987, che ha definito uno standard per la rappresentazione dei testi in formato digitale.

Il linguaggio usato per la rappresentazione è XML, che è simile al comune HTML con cui sono scritte le pagine web.

Tuttavia, a differenza di HTML, che è usato prevalentemente per specificare come un testo deve essere visualizzato su uno schermo, la TEI fa un uso più generale della marcatura: le etichette (chiamate anche "marche" o "tag") annotano singole parole o intere frasi per categorizzarle dal punto di vista semantico.

Per esempio, per indicare che un nome di persona si riferisce a un personaggio storico o che un toponimo rappresenta anche il nome con cui è nota una battaglia o un altro evento memorabile avvenuto in una certa data.

Il formato TEI include "tag" riferibili a 500 differenti componenti e concetti: parola, frase, carattere, segno, persona, ecc.

Ciascuno di questi “tag” è radicato in una o più discipline.

Così la Text Encoding Initiative viene incontro alle esigenze di diverse comunità scientifiche e professionali!

1.6 Caratteristiche dei corpora: l'annotazione

Annotare un corpus significa associarvi alcune informazioni che riassumono il risultato di elaborazioni precedenti.

Questa associazione può riguardare il corpus in quanto tale, ciascuno dei documenti che lo compongono o il testo dei singoli documenti.

È come se, partendo da un testo su carta, scrivessimo:

•in cima al foglio alcune informazioni generali (come la lingua in cui è scritto e il numero di parole);

•e accanto a ogni parola la sua analisi grammaticale.

E poi conservassimo il foglio con le annotazioni per usi futuri.

Ci sono diversi modi per annotare un testo. La principale distinzione è tra annotazione "in linea" e "fuori linea”.

L’annotazione “in linea” è simile a quella che abbiamo effettuato sul nostro testo su carta.

In un testo digitale si usa spesso la marcatura XML o il suo derivato HTML, in cui i frammenti di testo sono inseriti in “tag” di apertura e chiusura che ne indicano la funzione o il modo di visualizzarli.

Così, per esempio, questo simbolismo…

… indica che “La Divina Commedia” è il titolo del documento.

Si ha invece l'annotazione “fuori linea” quando l'informazione che arricchisce o interpreta il testo viene riportata a parte e posta in correlazione con i frammenti del testo (i singoli token, sequenze di token o intere frasi) mediante dei riferimenti chiamati "puntatori".

La pratica di annotare i corpora è sempre più comune, perché rendendo esplicite le informazioni che contengono, permette elaborazioni ancora più sofisticate.

1.7 Caratteristiche dei corpora: l'allineamento

L’allineamento è un’operazione che riguarda due corpora (o anche più) che hanno lo stesso contenuto informativo espresso in modi diversi, solitamente in lingue diverse.

Si tratta di un caso particolare di annotazione coordinata dei corpora, in cui si pongono in corrispondenza, ordinatamente, gli elementi di un corpus con quelli dell'altro.

L’allineamento può avvenire a diversi livelli:

•L’allineamento a livello di documento significa mettere in corrispondenza, uno a uno, i documenti costituenti.

Si creano così i corpora "paralleli", come le opere di Shakespeare in inglese e italiano, che consistono in coppie di documenti.

In cui, “The tragic history of Hamlet, prince of Denmark” corrisponde con “Amleto”, “Much Ado about nothing” con “ Molto rumore per nulla” e così via.

•Un allineamento a livello di frase ha una grana molto più fine. Assomiglia alle pubblicazioni dei classici con testo a fronte: dati due documenti, che devono essere già segmentati in frasi, possiamo mettere in corrispondenza le frasi del primo con quelle del secondo.

Sapremo così che “To be, or not to be, that is the question: …”, corrisponde con “Essere, o non essere, questo è il problema: …”.

· Un passo ulteriore consiste nell’allineamento a livello di token, che ovviamente si può fare solo con testi segmentati in token. Consiste nel mettere in corrispondenza sottoinsiemi di token di un testo con sottoinsiemi di token dell’altro.

Qui naturalmente le cose si complicano, perché di solito frasi corrispondenti in lingue diverse non hanno lo stesso numero di token o il loro ordine è diverso.

1.8 Corpora multilingue e corpora comparabili

Corpora paralleli possono essere molto utili per studiare le differenze tra due lingue, per costruire vocabolari, per estrarre terminologie.

Ma è anche possibile estendere l'operazione a più di due lingue, costruendo corpora multilingue.

In questi casi, allineare i corpora a livello di documenti può non essere complesso: ci sono molti esempi, come i documenti dell'Unione Europea, tradotti nelle diverse lingue comunitarie.

Ma, come è evidente, nascono una serie di problemi per l'allineamento a livello di frase e, ancora di più, di token.

Esistono poi corpora "comparabili", che contengono documenti in più lingue che hanno contenuto simile, ma non identico.

È il caso di Wikipedia, l'enciclopedia online in cui sono disponibili oltre 40 milioni di articoli bilingue sugli stessi argomenti, per 253 coppie di lingue.

1.9 I corpora nello studio del linguaggio

Come abbiamo visto più volte, i corpora ci forniscono la materia prima per trovare esempi concreti dei concetti relativi all'analisi dei testi e per verificare e "addestrare" gli algoritmi.

Vediamo ora di specificare meglio l'utilità dei corpora, a cominciare dalla lessicografia contemporanea dove rivestono un'importanza fondamentale.

Nella lessicografia sono usati, tra l'altro, per:

•associare ai lemmi a loro frequenza d'uso;

•identificare le concordanze, cioè le costruzioni tipiche in cui una certa parola si

presenta;

•cogliere le sfumature di senso delle parole in base ai contesti.

I corpora permettono di osservare l'uso effettivo di una lingua e di verificarne, su base statistica, le tendenze generali.

Infatti, quasi tutte le operazioni che effettuiamo sui testi, a diversi livelli di granularità (caratteri, token, sintagmi, frasi…), possono essere viste come strumenti per costruire "modelli" della "lingua" in generale o del particolare linguaggio usato da un certo autore o in settori particolari.

Per restare ai linguaggi settoriali, ti ricordo che estrarre i termini da un corpus di documenti prodotti dai membri di una certa comunità scientifica o professionale è il primo passo nella costruzione di una terminologia di settore.

1.10 I corpora e le lingue straniere

La disponibilità di corpora paralleli è di grande aiuto nell'insegnamento delle lingue e nel lavoro di traduzione.

Nell'insegnamento delle lingue, i corpora hanno due funzioni:

- supportano la costruzione di materiali didattici;
- permettono di osservare i contesti d'uso per inferire le proprietà di parole e di costruzioni più articolate.

Nel lavoro di traduzione, i corpora sono alla base della costruzione di modelli statistici di allineamento tra lingue diverse.

Analizzare la frequenza delle singole parole e la loro probabilità di occorrenza in specifici contesti lessicali o sintattici, facilita di molto sia la traduzione assistita, sia quella automatica.

Per fare un esempio, sapere che nel linguaggio italiano corrente quando si parla di funzionari pubblici è molto frequente l'accoppiamento tra "alto" e "papavero" aiuta a tradurre "alto papavero" con "top manager" piuttosto che con "high poppy"!

Per addestrare algoritmi di allineamento e traduzione automatica sono ancora molto usati gli "IBM alignment models" una sequenza di modelli di complessità crescente sviluppati negli anni '80 da un gruppo di ricercatori dell'IBM, veri pionieri del settore.

E, naturalmente, modelli statistici simili sono impiegati anche nel riconoscimento del linguaggio parlato...

1.11 I criteri di composizione dei corpora

In base al criterio di composizione, distinguiamo tipologie di corpora del tutto diverse.

Da una parte abbiamo le collezioni preesistenti (o comunque ben definite) di documenti che sono stati (o vengono ancora) prodotti regolarmente da istituzioni di diverso tipo.

Ne sono esempi la collezione storica di una rivista o di un quotidiano.

Abbiamo poi corpora i cui criteri di composizione riguardano la lingua, l'intervallo temporale, il genere, il settore o un autore.

Ne è un esempio il corpus di tutte le opere letterarie (narrativa, poesia, saggistica, ecc.) scritte in lingua Italiana nel XIII secolo a noi pervenute: non si tratta di un campione, ma dell'intera produzione letteraria che soddisfa i criteri di selezione.Ci sono infine altri corpora creati con criteri simili (lingua, genere, periodo, autore) ma a partire da una base di dimensioni eccessivamente ampie.

Per esempio, non sarebbe facilmente gestibile un corpus con tutta la letteratura italiana del XX secolo. In casi del genere, usiamo un corpus ridotto, che costituisce un campione rappresentativo dell'insieme.

Naturalmente, se disponiamo di un corpus in forma cartacea, dovremo prima digitalizzarlo.

Ma, per fortuna, un enorme raccolta di testi digitali esiste già...

1.12 Il web come corpus

Dal momento che la linguistica dei corpora richiede raccolte di testi autentici, non sorprende che molti autori vedano nel web la più ricca e accessibile fonte di

materiale linguistico oggi disponibile.Ci sono in proposito diversi approcci:

1. Il web come surrogato di un corpus.

Questo è l'approccio, ingenuo, di chi non ha idee chiare sul corpus di cui ha bisogno e non è consapevole della visione distorta che i motori di ricerca (e i loro algoritmi di indicizzazione) forniscono del reale contenuto del web.

2. Il web come supermercato.

Seguendo questo approccio, si costruisce un corpus in senso tradizionale, filtrando consapevolmente i contenuti del web con gli strumenti adatti.

3. Il web come corpus in senso proprio.

Qui si fa coincidere il corpus di interesse con il web.

4. Un mini-web usato come mega-corpus.

Con questo approccio, si costruisce una copia di una versione ridotta, ma rappresentativa, del web, da annotare come si vuole.

È un modo per cercare un equilibrio tra l'ingestibile estensione del web e le caratteristiche che vogliamo in un corpus, come la stabilità e la possibilità di eseguire annotazioni.

Ma, in pratica, come si fa?

1.13 Creare corpora dal web

Creare un corpus a partire dal web è soprattutto una questione di campionamento, che però presenta numerosi problemi teorici e pratici.

In pratica, sono quattro i passi da seguire:

1. Selezione degli indirizzi web (o URL) che fungono da "semi".
2. Reperimento delle pagine target mediante il "link following".
3. Pulizia dei dati.
4. Annotazione dei dati.

Cliccaci sopra per saperne di più!

1.14 Lessici, dizionari enciclopedici, terminologie e ontologie

Oltre ai corpora, le risorse linguistiche comprendono altri strumenti. Tra questi:

- I lessici, che sono repertori di parole e forme flesse.
- I dizionari enciclopedici, che forniscono il significato, le modalità d'uso e le traduzioni delle parole.
- Le terminologie, che permettono di studiare le parole semplici e composte usate in contesti specifici.
- Le ontologie, rappresentazioni schematiche dell'insieme di concetti relativi a un dominio.

E poi ci sono le grammatiche...

1.15 Le grammatiche

In linguistica si chiama "grammatica" un complesso di regole necessarie alla costruzione di frasi, sintagmi e parole di una determinata lingua. Di solito una grammatica supporta anche l'attività inversa, cioè l'analisi.

Una grammatica può avere più funzioni:

- la funzione "normativa" definisce come si deve o non si deve scrivere e parlare;
- la funzione "descrittiva" ci dice come scrivono o parlano certe classi di persone.

Inoltre, nel contesto della "teoria generativa del linguaggio", "grammatica" assume il significato di "modello della competenza linguistica di un parlante nativo".

Il termine "grammatica" può essere usato per riferirsi a diversi livelli della struttura di un testo.

La grammatica a livello sintattico, per esempio, consente di analizzare la frase in costituenti, individuando sintagmi nominali, verbali, preposizionali, ecc.

Un tipo di grammatica molto più semplice, che abbiamo già incontrato, è quella che consente di scomporre una frase in token: parole, segni di interpunzione, spazi e altri separatori. Questa grammatica può prendere la forma di una serie di espressioni regolari.

C'è da aggiungere che oggi, nel trattamento dei testi, l'impiego di algoritmi che "apprendono" le regole partendo da una grossa mole di dati tende a ridimensionare l'importanza delle grammatiche, intese, tradizionalmente, come espressione della competenza di un linguista umano!

1.16 Il problema della scarsità di risorse linguistiche

Oggi l'attività di ricerca e sviluppo nel trattamento automatico del linguaggio ha davanti un grande ostacolo.

È la carenza di risorse linguistiche adeguate. E, tra queste, soprattutto dei corpora.

Ricorda che ampi corpora di testi annotati sono alla base della qualità dei risultati dei metodi statistici e dell'addestramento di molti algoritmi per l'elaborazione del linguaggio naturale.

È una carenza di vecchia data (se ne parlava negli stessi termini quindici anni fa) ed è ancora più sentita per le lingue diverse dall'inglese. E, in particolare, per l'italiano.

Significa che nonostante il recente dilagare di assistenti vocali e altri sistemi digitali progettati per dialogare in linguaggio naturale con gli esseri umani, la strada da fare è ancora molta.