1.1 Risorse linguistiche nel web

1.2 A1 - Tipologie di ambienti per il NLP

Nella categoria “Ambienti per il Natural Language Processing” rientra una serie di strumenti, ricchi di funzioni, che permettono di effettuare analisi su singoli testi o interi corpora.

A seconda dei casi, il loro uso può essere didattico o professionale, anche se la distinzione è piuttosto vaga.

Tecnicamente, questi ambienti si distinguono in base alla modalità di fruizione e al linguaggio di programmazione con cui sono stati sviluppati.

Ecco alcune possibilità: clicca sulle diverse opzioni per saperne di più.

1.3 A2 - Supporto multilingue

Un elemento centrale degli ambienti per il Natural Language processing è il supporto multilingue, che non ha niente a che fare con la traduzione dell'interfaccia utente (che comunque nella propria lingua è più facile da usare).

Il supporto multilingue equivale alla possibilità di analizzare in modo adeguato testi e corpora in italiano o altre lingue diverse dall'inglese.

Tradizionalmente solo l'inglese dispone di un ampio repertorio di "risorse linguistiche", quali corpora, lessici, grammatiche, algoritmi di apprendimento compatibili, corpora annotati per fungere da "test set", algoritmi già addestrati in diversi campi (tokenizzazione, POS-tagging, estrazione di entità nominate, segmentazione e allineamento, traduzione automatica, ecc.).

Ancora oggi, diversi sistemi danno per scontato che si faccia riferimento alla sola lingua inglese. Mentre sono pochi gli esempi di ambienti open-source e open-data dotati di un livello accettabile di risorse linguistiche per un numero significativo di lingue diverse dall'inglese. Sono ancora meno quelli progettati fin dall'inizio avendo in mente il supporto multilingue.

Fortunatamente, grazie a una tradizione multi-decennale di studi nel campo dell'analisi dei testi e della linguistica computazionale, oltre all'inglese, ci sono diverse lingue con una discreta dotazione di risorse: innanzitutto francese, tedesco, olandese, giapponese e spagnolo. Poi, italiano, portoghese, ceco, ebraico e polacco.

Solo di recente, grazie anche agli interessi commerciali delle grandi web company, sono state prodotte estese risorse linguistiche - non necessariamente aperte - per il cinese, il coreano e l'arabo e le lingue del sub-continente indiano.

1.4 B1 - 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), che rispondono ad alcuni requisiti:

•sono conservate in formato digitale, per consentire l'elaborazione automatica;

•se possibile, sono corredate di strumenti di consultazione informatici;

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

•sono sottoposte a particolari elaborazioni, come la normalizzazione, l'annotazione

e l'allineamento.

Sono stati sviluppati numerosi strumenti di supporto alla costruzione e all'uso dei corpora. Gestiscono il reperimento dei testi, la loro organizzazione, la normalizzazione, la manutenzione, ecc.

Questi strumenti, in genere, sono fortemente orientati all'elaborazione statistica, necessaria per operare su grandi moli di dati.

1.5 C1 - Database online

I database online funzionali all'elaborazione del linguaggio naturale comprendono un insieme eterogeneo di risorse che fanno capo ad alcune tipologie:

•I lessici, repertori di parole e forme flesse.

•I dizionari enciclopedici, anche settoriali (per esempio i "gazetter", dizionari geografici da usare insieme a carte geografiche o atlanti), che contengono il significato, le modalità d'uso e le traduzioni delle parole.

•Le ontologie, rappresentazioni schematiche dell'insieme di concetti relativi a un dominio.

1.6 D1 - Strumenti di traduzione automatica

La traduzione automatica (o MT, "Machine Translation") è un vecchio sogno della fantascienza. E, già 50 anni fa, era uno dei principali campi di interesse della nascente intelligenza artificiale.

Dopo i passi da gigante degli ultimi anni, anche se non raggiunge la qualità di una buona traduzione professionale, la traduzione automatica è diventata una seria concorrente dei traduttori umani.

Ma non si tratta solo di concorrenza.

I traduttori professionali, per raggiungere un livello di produttività che gli consenta di restare sul mercato, usano correntemente strumenti di aiuto (o CAT, "Computer Assisted Translation") che a loro volta fanno un uso rilevante dei servizi di

traduzione automatica.

1.7 D2 - Uso didattico dei traduttori automatici

Sul possibile uso didattico dei traduttori automatici si è cominciato a ragionare da diversi anni.

Tuttavia, 10-15 anni fa era prevalente la preoccupazione dei docenti per un loro "uso improprio", in pratica per copiare, aggirando le verifiche di competenza.

Attualmente, anche sulla base dell'esperienza, si è più propensi a valutare in modo equilibrato i casi in cui l'uso di un sistema automatico può aiutare a migliorare le competenze di un traduttore.

O, addirittura, fornire al discente l'occasione per un primo livello di apprendimento, perché un confronto critico tra il testo originale e quello tradotto è sempre fonte di conoscenza.

Provare per credere!

In materia di apprendimento delle lingue ci sono anche posizioni più radicali.

Qualcuno, considerati i grandi progressi della traduzione automatica negli ultimi dieci anni (e quelli prevedibili per il prossimo futuro), si sta chiedendo se valga veramente la pena di dedicare energie e investimenti all'apprendimento delle lingue straniere (in particolare, dell'inglese).

1.8 D3 - Traduttori automatici sul mercato

La maggior parte dei sistemi di traduzione automatica sono accessibili in due modalità:

•online, attraverso una comune interfaccia web interattiva;

•tramite librerie software (API, “Application Programming Interface”) disponibili sul web, a cui si può accedere sia da applicazioni personali di Computer Assisted Translation, sia da servizi erogati online.

Di solito la traduzione interattiva, frase per frase, è libera e gratuita, mentre la traduzione di frasi o di interi documenti è a pagamento, anche se i costi sono in forte calo.

I principali traduttori automatici si distinguono per la qualità media della traduzione e per il numero di “coppie linguistiche” che sono in grado di gestire.

Da tempo, i traduttori di Google e di Microsoft sono in grado di effettuare traduzioni dirette, da lingua a lingua, a partire da oltre 100 lingue.

1.9 D4 - Come usare i traduttori automatici

Un traduttore automatico può essere usato in diversi modi. Per esempio:

•Tramite un plugin (cioè un’estensione) del browser, per tradurre un'intera pagina web o una sua porzione.

•Tramite un sito web che consenta di caricare interattivamente un documento, una frase o un breve testo, tradurlo in remoto e scaricarlo nel formato voluto.

•Tramite di uno strumento di Computer Assisted Translation in grado di interrogarlo.

•Tramite un linguaggio di programmazione come Python, Java, o PHP. in questo modo è anche possibile chiedere la traduzione di molte frasi o di testi lunghi.

C’è da aggiungere che, molto spesso, anche i motori di ricerca traducono le parole chiave inserite dall’utente per ampliare la ricerca a pagine scritte in una lingua diversa.

1.10 D5 - Come verificare l'efficacia dei traduttori automatici

L'utente ha la possibilità di fare qualche verifica per saggiare l'attendibilità del traduttore automatico. Per esempio:

•Può analizzare il testo tradotto e verificare in base alle proprie conoscenze se ci sono errori grossolani.

•Può far tradurre un numero consistente di frasi, nella stessa lingua, da due diversi traduttori e confrontare i risultati. Se concordano, tutto bene. Altrimenti può

ricorrere a un terzo traduttore.

•Può usare un traduttore per tradurre in una lingua straniera e poi un secondo traduttore per ritradurre nella lingua originaria

1.11 Mappa delle risorse linguistiche nel web