

LA LINGUISTICA COMPUTAZIONALE E LE POSSIBILITA' CHE OFFRE

A. ZAMPOLLI

Università di Pisa e CNR, Pisa.

Ho riportato la distinzione tra software didattico e non, per dire che ciò di cui parlerò oggi: il software non didattico. In particolare il problema è questo. La linguistica computazionale ha creato dei sistemi e degli strumenti. Come il professor Porcelli ha detto sono orientati ad assistere il ricercatore nello studio del linguaggio: il filologo, lo storico, tutti quelli che hanno a che fare con informazione espressa in forma di linguaggio naturale.

Il problema che ci poniamo noi è quello di sapere se e come gli strumenti che la linguistica computazionale crea possano essere utili in qualche modo nell'insegnamento sia della lingua madre che delle lingue straniere.

Mostrerò rapidamente qualche esempio di qualche sistema, che non è inteso direttamente all'insegnamento, ma che noi speriamo, chiediamo se può essere utilizzato o no. La risposta può essere data solo dai glottodidatti, dagli insegnanti stessi. Quando noi veniamo a queste conferenze è perché noi cerchiamo una collaborazione e le competenze che noi non abbiamo.

Ci sono varie definizioni della linguistica computazionale. Noi adottiamo quella più semplice: qualsiasi utilizzazione di calcolatori per elaborare i dati che concernono il linguaggio naturale (per esempio inchieste sociolinguistiche) o per analizzare i dati che descrivono il linguaggio naturale: lessici, grammatiche e strutture di vario tipo o che presentano il linguaggio naturale come ad esempio i testi.

Vorrei sottolineare un fatto importante per capire la situazione in cui ci troviamo oggi, e il fatto che negli ultimi tre o quattro anni c'è stata un'esplosione delle richieste di metodi e di strumenti della linguistica computazionale per attività pratiche. Cosa sono queste applicazioni? Esse mirano ad assistere

o a sostituire un uomo nell'eseguire alcune operazioni, la cui esecuzione ha un interesse pratico, economico. Esempi sono la traduzione, la documentazione, il dialogo uomo-calcolatore, l'accesso a basi di dati linguistici, la sintesi e la comprensione del parlato, l'automazione degli uffici, ecc.

Oggi sapete che si parla della cosiddetta società dell'informazione e praticamente una grande quantità di informazione è prodotta attraverso il linguaggio naturale e attraverso il calcolatore.

Questo ha portato le autorità sia a livello internazionale, che a livello di alcuni dei paesi più avanzati, a formulare il concetto di industria delle lingue. Essa è concepita come l'insieme delle attività a scopo pratico di cui agli esempi che precedono. Chiaramente è un mercato estesissimo. Per esempio si calcola che solamente la Comunità Europea traduca 100 milioni di pagine all'anno (!). Se fate il conto del costo di traduzione di una cartella in 10 o 12 lingue, vi rendete conto di cosa questo rappresenti.

Non c'è dubbio che esiste un mercato potenziale dell'industria delle lingue. Il problema è che la tecnologia che abbiamo non è in condizioni ancora, oggi, di fornire tutti i prodotti che potrebbero servire per rispondere a una richiesta di questo tipo.

Alcune autorità, la CEE, il Consiglio d'Europa, il Governo giapponese, il Governo tedesco in parte il CNR e via via, hanno capito questa situazione. L'esistenza di un mercato. Che il mercato è un mercato che dovrebbe consentire all'Europa una posizione di vantaggio nei confronti degli altri due grandi blocchi – Stati Uniti e Giappone – perché l'Europa, a differenza di quelli, ha un'esperienza multilinguistica. Hanno però capito che per affrontare un mercato di questo tipo è necessario promuovere la ricerca e mettere a punto alcune cose, per esempio descrivere esaurientemente le lingue come sono oggi in tutti i loro aspetti (lingua comune, lingua tecnica). Per fare questo occorrono alcuni strumenti, per esempio la raccolta di corpora monolingue, di corpora multilingue, di basi di dati lessicali mono e multilingue, ecc.

Se questo è vero, vuol dire che avremo l'occasione oggi di inserirci in un movimento di questo tipo e chiederci quante cose possono essere utilizzate tra quelle prodotte per l'insegnamento delle lingue.

Concludo con tre rapidi esempi.

Un aspetto centrale della linguistica computazionale è quello di creare degli algoritmi capaci di utilizzare delle regole, di fare l'analisi o la generazione di frasi, di testi di linguaggio naturale in genere. Questo è certamente il settore che ha interessato fin da principio gli insegnanti che si occupavano di tecnologie per l'insegnamento delle lingue. Però secondo me è quello che ha

avuto un minor numero di applicazioni. Il motivo è che prima di tutto i metodi che abbiamo non sono capaci di trattare qualsiasi frase. Questo è uno dei motivi per cui ancora oggi i sistemi per l'analisi della risposta libera non ci sono. Ma soprattutto il problema è quello della complessità dei formalisti e degli algoritmi che fanno queste cose. Ci sono delle cose semplici che potrebbero essere utilizzate subito e che sono state citate. È possibile certamente creare tutto un insieme di esercizi per verificare o per insegnare la morfologia.

Qui avete un esempio di spagnolo. Un'opera di spagnolo, che credo sia "Cronaca di una morte annunciata" con tutte le analisi che il calcolatore suggerisce a livello morfologico e con il suggerimento di disambiguazione degli omografi. So che ci sono già degli esperimenti in corso di utilizzazione di queste cose per l'insegnamento della morfologia.

A livello di sintassi la cosa è più complessa. Se per esempio vedete un grafico esso rappresenta la descrizione delle regole per il gruppo nominale in italiano. In un programma di traduzione automatica, che stiamo svolgendo per conto della Comunità, questo serve per analizzare una frase semplice e dire...questo è il soggetto, questo è il predicato, questo è il complemento...

È chiaro che questi sistemi sono ad elevata complessità, però credo che si potrebbe cercare di identificare delle applicazioni anche in questo settore, che ci potrebbe essere un aiuto nella verifica automatica di esercizi. Si potrebbe cercare di fare qualcosa di più per l'analisi della risposta libera; ma soprattutto io ho visto delle applicazioni non tanto nello studio per l'apprendimento della lingua, ma quanto per lo studio della grammatica e per lo studio scientifico di una lingua. Per esempio dare a uno studente un programma capace di generare delle frasi applicando delle regole, dargli il programma, dirgli di scrivere delle regole di generazione e far sì che egli veda quali regole escono.

Un altro settore in cui io credo che vi siano delle possibilità di collaborazione sono i lessici. Noi distinguiamo tre tipi di risorse. Suppongo che alcuni di voi avranno visto fuori la dimostrazione della Olivetti con il compact-disc. È un disco compatto su cui c'è un certo numero di vocabolari. Sapete che la maggior parte dei vocabolari oggi è prodotta già in fotocomposizione. Per esempio noi a Pisa abbiamo l'Oxford Advanced Learn Dictionary, il Longman, il Collins, il Casati ecc.

Ci sono delle tecniche abbastanza semplici, che avete visto anche fuori, che aiutano semplicemente in una consultazione migliore del contenuto del vocabolario. Il concetto è che per un uso normale del vocabolario, o conoscete la parola o non trovate l'informazione. Se avete la possibilità di accedere al

vocabolario in qualsiasi punto, potete dare una parola, un'espressione e il calcolatore la cerca all'interno di tutte le voci.

La linguistica computazionale fa qualcosa in più e riesce a trasformare vocabolari di questo tipo in basi di dati lessicali o in basi di conoscenze lessicali. Praticamente le basi dati lessicali sono l'equivalente del contenuto del vocabolario, in cui però le informazioni sono strutturate e identificate chiaramente. Si sa cosa è un esempio, cosa è una definizione, e via via.

Le basi di conoscenze lessicali sono qualcosa in più e sono quelle che servono per esempio per la traduzione meccanica, per l'analisi del linguaggio, ecc. Oltre a quello che è il contenuto del vocabolario, ci sono delle informazioni implicite date dal vocabolario.

Per esempio, analizzando le definizioni, è possibile costruire delle tassonomie concettuali, cioè dire: queste parole appartengono allo stesso campo lessicale. E d'altra parte noi inseriamo delle informazioni in più, ad esempio – per i verbi – qual è la sottocategorizzazione, la struttura dell'argomentazione e cose di questo tipo.

Lo schema cui si giunge oggi è praticamente di avere un dizionario monolingue, un dizionario monolingue in un'altra lingua, connessi attraverso una situazione di lavoro che noi chiamiamo lessicografia, ma che potrebbe essere utilizzata dallo studente, connessa ai corpora di testi nella lingua 1, nella lingua 2.

E parliamo anche di corpora contrastivi o paralleli, cioè vogliamo avere all'interno di questi corpora alcuni testi che esistono sia nella lingua 1 sia in traduzione nella lingua 2.

Allora, voglio dire che i corpora a disposizione oggi crescono continuamente. Ne possiamo citare alcuni: il Trésor de la Langue Française ha 200 milioni di parole; il Cobit, costruito per la Collins dall'Università di Birmingham ne ha 20; quello che stiamo costruendo noi per l'Italiano ne dovrebbe avere 20 o 30 a breve scadenza.

Quindi praticamente per tutte le grandi lingue di cultura oggi c'è un corpus di riferimento.

A questo punto chiederei ai miei collaboratori di mostrare un esempio di cosa si può fare su un lessico o su dei corpora e concluderei con una proposta rapida di collaborazione.