LA LINGUISTICA COMPUTAZIONALE: UN CROGIOLO DI ESPERIENZE MULTIDISCIPLINARI

Fabio Tamburini
Dipartimento di Studi Linguistici e Orientali
Università di Bologna

In questo breve saggio, tenterò di delineare e caratterizzare i confini, peraltro molto sfumati, di una disciplina che, nata attorno alla metà del secolo scorso, affronta problemi estremamente attuali, specialmente in questi anni nei quali il trattamento dell'informazione è divenuto uno degli aspetti centrali della nostra vita quotidiana. Per far questo mi avvarrò di numerosi e preziosi contributi di studiosi che, negli anni, hanno analizzato e descritto attentamente l'evoluzione della disciplina [Ferrari, 2005; Key 2003; Lee, 2001; Mitkov, 2003; Spärck Jones, 2007].

Gli ovvi limiti di spazio mi costringeranno a tratteggiare brevemente argomenti che meriterebbero, e hanno meritato in passato, ben più ampie discussioni, e di questo mi scuso anticipatamente.

La Linguistica Computazionale (o *Natural Language Processing* – Trattamento Automatico delle Lingue) – d'ora in poi LC – è un settore fortemente interdisciplinare, e si occupa dell'elaborazione delle lingue, in ogni possibile aspetto, mediante l'uso di calcolatori. Dalla sua nascita alla fine degli anni '50, e dalla sua configurazione come disciplina autonoma, ha subito una crescita esponenziale in diverse direzioni arrivando ad attingere contributi da ambiti quali la linguistica, che produce i modelli teorici del linguaggio, la psicologia, che fornisce un'analisi dei processi cognitivi, la teoria dell'informazione, che analizza le modalità comunicative, la matematica e la statistica, che forniscono gli strumenti per esprimere tali modelli in modo computazionalmente trattabile e naturalmente l'informatica per quanto riguarda lo sviluppo degli algoritmi atti ad implementare i modelli teorici dei fenomeni linguistici.

Nelle aspettative comuni questa disciplina dovrebbe riuscire, in un futuro auspicabilmente prossimo, ma tuttavia non ben definibile, a produrre macchine in grado di interagire con gli esseri umani utilizzando il linguaggio naturale. Nella letteratura e cinematografia fantascientifica gli esempi di tali visioni sono numerosi (si pensi ad esempio all'elaboratore HAL9000 del film *2001 odissea nello spazio* o al robot C-3PO della saga di *Guerre stellari*). Tuttavia queste "macchine" sono ancora lontane, e gli esempi più sofisticati a nostra disposizione sono ben lungi dal poter soddisfare queste aspettative.

E' lecito quindi chiedersi: perché dopo più di cinquant'anni di intensa ricerca non si è ancora riusciti a centrare il bersaglio "imbrigliando" il linguaggio umano in opportuni modelli atti ad essere formalizzati e implementati con programmi per calcolatori? Dopo tutto un bambino si appropria di una lingua in pochi anni e senza seguire alcun corso sulla struttura e le caratteristiche della lingua stessa.

Molto del successo del linguaggio umano nei processi comunicativi deriva dall'innata abilità degli esseri umani nel gestire imprecisioni e ambiguità in modo efficiente, evincendo da un insieme estremamente ampio di stimoli e informazioni relative al contesto, testuale, situazionale ed emotivo, la corretta interpretazione e il senso del messaggio, un contesto ben più esteso delle poche parole utilizzate tradizionalmente nei sistemi di analisi testuale. La stessa forzata suddivisione dei task classici della LC (analisi morfologica, PoS-tagging, parsing, ecc…) mantenendo la divisione nei tradizionali livelli d'analisi a causa della mancanza di adeguate risorse modellistico-computazionali, introduce, spesso artificiosamente, un numero enorme di ambiguità, e quindi di problemi da risolvere computazionalmente, rendendo ogni task estremamente complesso da trattare con metodi automatici.

Questo ha generato negli anni situazioni estremamente frustranti e di tensione tra i linguisti computazionali, che in certi casi, hanno acquisito la reputazione di non essere in grado di far fronte alle sfide tecnologiche costruendo adeguati modelli teorici in grado di supportare la costruzione di sistemi performanti ed efficienti. Molti di questi problemi sono tuttavia dovuti al fallimento, o quantomeno allo scarso successo, degli studi nel campo dell'intelligenza artificiale (IA) degli ultimi 25/30 anni, campo di studi che ingloba, in qualche modo, la LC e che è sempre stato il principale riferimento della disciplina per attingere metodi e modelli.

Come ogni disciplina legata alle scienze naturali ha le sue sfide e i suoi "grandi problemi" da affrontare, così la linguistica computazionale si trova a dover fronteggiare alcune *grandi sfide*: *machine translation*, *information extraction*, *text summarisation*, *document retrieval and indexing*, *speech recognition, production and understanding*, *natural language interfaces*, *interactive dialogue systems*, *semantic Web*, sono solo alcuni dei problemi chiave che chiamano la disciplina e gli studiosi che vi lavorano a fornire soluzioni adeguate, sia teorico-modellistiche sia sperimentali e applicative.

Uno sguardo al passato

I primi approcci alla disciplina nascono da una moltitudine di contributi interdisciplinari che configurano un punto di vista empirico e distribuzionale sia in ambito linguistico [Firth, 1957; Harris, 1951] sia in ambiti più modellistico-tecnologici [Shannon, 1948; Turing, 1950; Weaver, 1949]. Questi studi, partendo dal concetto fondamentale che le co-occorrenze (o correlazioni) all'interno dei fenomeni sono fonti importanti di informazioni sulla lingua, delinearono una metodologia di indagine focalizzata principalmente sull'evidenza empirica di tali fenomeni, sulla *parole*, in termini saussuriani e su processi di analisi induttiva dei dati.

La proposta generativista [Chomsky, 1957] scardinò completamente questi primi approcci all'analisi linguistica con strumenti computazionali, spostando l'attenzione verso punti di vista razionalistici, basati prevalentemente su una visione del linguaggio come un oggetto formale, matematicamente descrivibile e in parte biologicamente determinato nel cervello umano, indagabile principalmente attraverso processi introspettivi e con metodi deduttivi che sono risultati estremamente adatti ad essere utilizzati con gli elaboratori. Le osservazioni di Chomsky sulla finitezza del materiale empirico a disposizione e la conseguente impossibilità di compiere indagini complete, frenarono prepotentemente gli studi condotti su base empirica [Chomsky, *ibid*] spostando la barra del timone della ricerca in LC verso metodologie basate su regole.

Nei primi anni '90 si è assistito ad un ritorno massiccio degli studi di carattere empirista, grazie all'introduzione di nuove metodologie di elaborazione dei dati empirici su base stocastico-statistica, al successo di tali metodologie nella costruzione di applicazioni reali, e a nuove indagini psicolinguistiche che hanno mostrato come l'apprendimento linguistico umano sia basato su criteri statistici molto più di quanto si pensasse fino a quel momento.

La Linguistica Computazionale oggi

Al giorno d'oggi l'approccio di gran lunga prevalente nell'elaborazione automatica del linguaggio naturale si colloca nel mezzo tra il *NLP* classico, basato su regole, di matrice razionalista e fondato sui metodi tipici dell'IA, e il *Machine Learning* (*ML*).

Le tecniche di *ML* affondano le loro radici in metodologie di analisi stocastico-statistiche estremamente sofisticate, in grado di costruire modelli del fenomeno in esame a partire da un'opportuna quantità di dati autentici (*corpora*), accuratamente annotati, che fungono sia da base statistica sia da insieme di esempi di una corretta gestione del fenomeno analizzato. La costruzione automatica dei modelli si avvale di complesse procedure statistiche che configurano veri e propri

processi di "apprendimento" guidato dai dati (si vedano [Callison-Bourne, Osborne, 2003] e [Alpaydin, 2004] per un quadro completo su questo tipo di approcci).

Le numerosissime metodologie che possono essere identificate, più o meno propriamente, nel dominio del *ML* (*neural networks*, *hidden markov models*, *maximum entropy models*, *kernel methods and support vector machines*, *decision trees*, *genetic algorithms*, ecc…) si configurano tutte come metodi in grado di apprendere, di derivare dai dati empirici, le caratteristiche fondamentali e le sfumature che definiscono il fenomeno in esame al fine di costruire riconoscitori o classificatori automatici per tale fenomeno. Possiamo trarre esempi di sistemi di successo ascrivibili a queste metodologie da ogni ambito di analisi linguistica: sistemi per il riconoscimento della lingua parlata (*Automatic Speech Recognition*), sistemi per l'etichettatura delle categorie lessicali (*PoS-tagging*), sistemi per la disambiguazione semantica a livello lessicale (*Word-Sense Disambiguation*), per l'analisi sintattica a livello di frase (*Parsing*) e molti altri ancora.

In realtà, negli anni '90 si è assistito ad una pericolosissima deriva della disciplina verso queste metodologie; il *ML* ha ottenuto risultati fino ad allora impensabili nella risoluzione dei problemi della LC, sorpassando di gran lunga le prestazioni ottenute, e ottenibili, da ogni sistema fondato su regole. Specialmente in ambito ingegneristico/tecnologico si è assistito a un progressivo scollamento dei progetti di ricerca dalle teorie linguistiche consolidate a favore di modelli stocastico/statistici che, molto spesso, non contengono alcuna esplicita informazione linguistica, allontanando la linguistica e i linguisti dal nucleo proprio della disciplina a favore di figure di studiosi decisamente più radicate in ambiti tecnologici.

Sintomatico e rappresentativo è l'amichevole scontro dialettico iniziato nel '98 da Jelinek con la celebre frase "*Whenever I fire a linguist, our system performance improves*" [Jelinek, 2005] che ha suscitato, molti anni più tardi, una simpatica, ma ferma, risposta di alcuni linguisti della scuola di Praga [Hajic, Hajicova, 2007]. A parte gli aneddoti che accompagnano ogni disciplina, il brusco cambio di orizzonte, venutosi a creare alle fine degli anni '90 e nei primi anni del nuovo secolo, ha radicalmente cambiato non solo le metodologie utilizzate nell'ambito degli studi, ma anche i rapporti tra le varie anime interdisciplinari che la LC possiede.

La domanda che si sono posti molti studiosi del settore è da certi punti di vista piuttosto inquietante: è possibile che la linguistica computazionale possa procedere senza il contributo dei linguisti? E' una domanda evidentemente provocatoria, sulla carta, ma, in pratica, in molti progetti di ricerca finalizzati alla produzione di applicazioni la componente "tecnologica" del team tende a procedere in maniera autonoma riuscendo in pochissimo tempo a produrre prototipi funzionanti che forniscono prestazioni allo stato dell'arte.

E proprio questi sono i principali vantaggi dell'applicazione di tecniche statistiche al *NLP* rispetto a quelle tradizionali basate su regole: sono veloci ed economiche da produrre, consentono una rapidissima costruzione di prototipi e sono molto robuste nel gestire l'immensa variabilità del linguaggio umano rispetto alle loro controparti basate su regole. Dato un opportuno corpus annotato da utilizzare come base empirica, spesso la produzione di un prototipo che mostra prestazioni già notevoli risulta essere un lavoro di pochi giorni.

Il punto cruciale, la domanda "vera" che è opportuno porsi è tuttavia un'altra: gli approcci stocastico-statistici e il *ML* sono in grado di risolvere compiutamente le grandi sfide che la disciplina ha di fronte? Negli ultimi anni è diventato ormai chiaro che, nonostante l'enorme passo in avanti che hanno consentito alla LC, queste metodologie, se non corroborate da opportune informazioni e modelli che solo le scienze linguistiche possono fornire, non potranno andare ancora molto lontano e certamente non risolveranno da sole i grandi problemi che ogni linguista computazionale sogna di poter affrontare e risolvere.

Queste metodologie infatti presentano anch'esse un tallone d'Achille piuttosto serio, per descrivere il quale è necessario fare un brusco salto indietro nel tempo, e precisamente negli anni '40, quando lo studioso George Zipf definì una legge empirica di grandissima importanza. La Legge Empirica di Zipf [1949] afferma che dato un qualsiasi corpus di testi autentici (ad esempio il corpus CORIS/CODIS [Rossini Favretti, *et al*. 2002]), la frequenza di ogni parola nel corpus è

inversamente proporzionale alla posizione (rango) che la parola assume nella lista che elenca tutte le parole del corpus in ordine di frequenza.

L'impatto che questa legge empirica ha sulla LC, e specialmente sui metodi basati sul *ML*, è dirompente: un numero estremamente limitato di parole della lingua – le parole grammaticali o vuote – compongono la maggior parte dei testi, mentre le parole lessicali o piene, molto più interessanti dal punto di vista dello studio della lingua, tendono ad assumere una frequenza molto più bassa. Questo porta alla necessità di costruire *corpora* enormi, al fine di avere un adeguato numero di esempi del fenomeno in esame dai quali ricavare regolarità ed eccezioni.

I metodi di *ML* basati su teorie stocastico-statistiche necessitano di una grande quantità di dati autentici arricchiti con annotazioni che mostrano la corretta etichettatura (o classificazione del fenomeno) per poter costruire automaticamente i modelli del fenomeno in esame; l'annotazione di tali risorse linguistiche viene di solito prodotta utilizzando procedure prevalentemente manuali, che molto spesso non consentono, per ragioni di tempi e costi, la produzione di risorse di dimensioni adeguate alla complessità dei problemi che si devono affrontare.

Tutte queste metodologie soffrono quindi del problema della "sparsità dei dati", ovvero di una cronica mancanza di dati empirici annotati affidabili coi quali costruire i modelli dei fenomeni linguistici. Se gli esempi di un determinato fenomeno sono insufficienti diventa impossibile distinguere un fenomeno non appartenente al sistema linguistico in studio da un fenomeno semplicemente raro.

La scarsezza di dati empirici forza quindi le metodologie di *ML* ad introdurre limitazioni fortissime nei modelli, limitazioni che, da un punto di vista strettamente linguistico, risultano inaccettabili. Com'è possibile, ad esempio, costruire un modello della lingua con tecniche a n-grammi [Manning, Schütze, 1999] considerando, per le ragioni sopra esposte, unicamente relazioni tra sequenze di tre parole (trigrammi)? Da un punto di vista linguistico equivale ad affermare che non vi è alcuna relazione tra una parola e le altre che compongono la stessa frase al di fuori di un cotesto composto da ±2 parole! D'altra parte il problema della sparsità dei dati non consente di allargare l'orizzonte molto oltre questo limite. Anche se, a dispetto di questi fatti, le prestazioni di tali modelli sono buone, il problema è molto serio e la necessità di risorse linguistiche accuratamente annotate (fonologicamente, morfo-sintatticamente, sintatticamente, semanticamente, ecc…) è pressante.

Il punto critico di questo tipo di imprese sembra quindi spostarsi sulle risorse linguistiche annotate; come possiamo costruire corpora annotati sufficientemente grandi e affidabili per supportare l'addestramento di tecniche di *ML*? Ma soprattutto, chi deve produrre tali risorse?

I contributi di [Hajic, Hajicova, 2007; Hajicova, 2006] esaminano accuratamente questi punti soprattutto in relazione al rapporto/contrasto tra le componenti linguistiche e ingegneristico-tecnologiche nel settore. La produzione di una risorsa linguistica annotata in modo affidabile richiede la definizione di accurate procedure di annotazione che siano la sistematica e consistente applicazione di una teoria linguistica sul fenomeno in esame. Ecco allora che l'apparente scollamento tra linguisti e "tecnocrati" sembra essere artificioso e, in realtà, tutte le informazioni utilizzate dai metodi di *ML* sono informazioni pre-elaborate dal linguista secondo una precisa teoria di riferimento che descrive accuratamente il fenomeno in studio.

D'altra parte considerare la linguistica computazionale come una scienza empirica implica forzatamente la necessità di poter compiere misure oggettive di natura quantitativa sui fenomeni. Senza considerare annotazioni aggiuntive, sono ben poche le misure che possono essere fatte direttamente sul testo e limitato risulta essere il numero dei problemi che possono essere efficacemente affrontati utilizzando il testo "puro" come unica fonte di informazione, escludendo ogni fase intermedia di comprensione e interpretazione da parte del linguista.

Seguendo il suggerimento di Hajic e Hajicova, è possibile immaginare il ruolo del linguista nella LC del 21-esimo secolo proprio come esperto che predispone, classifica, identifica ed evidenzia i fenomeni e i tratti che li definiscono al fine di produrre, tra le altre cose, risorse linguistiche ampie e coerenti rispetto ad una teoria di riferimento, da utilizzare come base per

l'addestramento dei metodi stocastico-statistici di *ML*. Il linguista è inoltre in grado di identificare quali sono i tratti maggiormente significativi nella definizione dei vari aspetti del fenomeno in esame, guidando il processo di modellizzazione automatica nella selezione di tali tratti.

Un ulteriore punto di contatto forte che vede i linguisti protagonisti riguarda la valutazione dei sistemi automatici di etichettatura. Da alcuni anni si è attestata l'importanza di oggettive campagne di valutazione che verifichino le reali prestazioni dei sistemi di annotazione automatica dei testi su insiemi di dati coerenti e comuni. Per troppo tempo si è assistito a "gare" sulle *performance* dei vari sistemi misurate in modo non omogeneo, con misure (metriche) differenti e su insiemi di dati diversi.

Una valutazione controllata, oggettiva e gestita da un unico organismo consente di valutare realmente le prestazioni delle differenti soluzioni del problema proposte dai vari studiosi, consentendo alla comunità una reale definizione dello "stato dell'arte", delle potenzialità e delle problematicità dei vari approcci.

Negli ultimi anni sono emerse numerosissime iniziative di questo tipo in ogni settore della LC, tanto che sarebbe impossibile elencarle nella loro totalità. Molte di queste campagne di valutazione si ripetono ormai da anni, con miglioramenti nelle procedure e con l'aumento della complessità dei *task*. A questo proposito mi sembra doveroso citare la prima campagna di valutazione sui prodotti della LC sviluppati per la lingua italiana – EVALITA 2007 – che comprendeva ben 5 task distinti, ai quali hanno partecipato globalmente 30 sistemi e altrettanti team di studiosi a livello internazionale.

La definizione di queste campagne richiede lo sviluppo di procedure e di misure linguisticamente motivate che vede il contributo dei linguisti come fondamentale.

Conclusioni

Questo settore di ricerca attrae i giovani proprio per la sua interdisciplinarità che mescola competenze umanistiche con i saperi relativi alle scienze naturali e tecnologiche.

La LC è al centro di una rivoluzione in questi anni. L'evoluzione delle macchine della prossima generazione deve necessariamente passare attraverso un'interazione linguistica e la comprensione e la gestione del linguaggio umano. Tutti i dispositivi che ci circondano, computer, cellulari, palmari, tutto il mondo della multimedialità, la stessa Internet, richiedono prepotentemente lo sviluppo di modelli adeguati al trattamento automatico delle lingue per raggiungere i livelli di efficienza nell'interazione persona-macchina che la società dell'informazione del 21-esimo secolo richiede, per la ricerca, la navigazione e l'estrazione delle informazioni in un ambiente multilinguistico e multiculturale.

Bibliografia

Alpaydin, E. (2004). *Introduction to Machine Learning*. MIT Press, Cambridge MA.

Callison-Bourne, C., Osborne, M. (2003). Statistical Natural Language Processing. In A. Farghaly (ed.), *A Handbook for Language Engineers*, CSLI Publications.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague, The Netherlands.

EVALITA 2007. http://evalita.itc.it/

Ferrari G. (2005). La ricerca in Linguistica Computazionale tra modelli formali ed analisi empirica, in G. Marotta (ed.), *Atti del Convegno di Studi in memoria di Tristano Bolelli, in Studi e Saggi Linguistici*, XL-XLI (2002-2003), Pisa, pp. 101-119.

Firth, J.R. (1957). A synopsis of linguistic theory 1930–1955. In the *Philological Society's Studiesin Linguistic Analysis*. Blackwell, Oxford, pp. 1–32.

Hajic, J., Hajicova, E. (2007). Some of Our Best Friend are Statiticians. In V. Matousek, P. Mautner (eds.), *TSD2007, LNAI 4629*, Springer-Verlag, pp. 2-10.

Hajicova, E. (2006). Old linguists never die, they only get obligatory deleted. *Computational Linguistics*, 32, 457-469.

Harris, Z. (1951). *Methods in Structural Linguistics*. University of Chicago Press.

Jelinek, F. (2005). Some of my Best Friend are Linguists. *Language Resources and Evaluation*, 39, 25-34.

Spärck Jones, K. (2007). Computational Linguistics:What About the Linguistics? Computational Linguistics, 33(3), 437-441.

Key, M. (2003). "Introduction", In Mitkov, R. (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: OUP.

Lee, L. (2004). "I'm sorry Dave, I'm afraid I can't do that": Linguistics, Statistics and Natural Language Processing circa 2001. In *Computer Science: Reflection on the Field, Reflection from the Field (Report of the National Academies' Study on the Fundamentals of Computer Science)*, pp 111-118.

Manning, C., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press.

Mitkov, R. (ed.) (2003). *The Oxford Handbook of Computational Linguistics*, Oxford: OUP.

Rossini Favretti R., Tamburini F., De Santis C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In Wilson, A., Rayson, P. and McEnery, T. (eds.), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Lincom-Europa, Munich. (2002), 27-38.

Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27:379–423 and 623–656.

Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, LIX:433–60.

Weaver, W. (1949). Translation. Memorandum. Reprinted in W.N. Locke and A.D. Booth (eds.), *Machine Translation of Languages*: Fourteen Essays, MIT Press, 1955.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least-Effort*. Addison-Wesley.