PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

LA GRAMMATICA E L'ERRORE

Le lingue naturali tra regole, loro violazioni ed eccezioni

a cura di
Nicola Grandi

Bononia University Press

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

Bononia University Press
Via Farini 37 – 40124 Bologna
tel. (+39) 051 232 882
fax (+39) 051 221 019

www.buponline.com
e-mail: info@buponline.com

© 2015 Bononia University Press

ISBN: 978-88-7395-982-3

I diritti di traduzione, di memorizzazione elettronica, di riproduzione e di adattamento totale o parziale, con qualsiasi mezzo (compresi i microfilm e le copie fotostatiche) sono riservati per tutti i Paesi.

Immagine di copertina: Grandville, *Les métamorphoses du jour*, 1869, tav. LXIII

Impaginazione: Sara Celia

Stampa: Global Print (Gorgonzola, Milano)

Prima edizione: marzo 2015

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

SOMMARIO

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

Modelli computazionali del linguaggio tra regole e probabilità

Alessandro Lenci

Università degli Studi di Pisa

Nel febbraio del 2011 il sistema informatico dell'IBM *Watson* ha partecipato al quiz televisivo *Jeopardy!* e ha sconfitto gli altri concorrenti umani (Ferrucci, 2012). La sfida ricorda da vicino quella del computer *Deep Blue* che nel 1996 vinse contro il campione di scacchi Kasparov, con l'importante differenza che la capacità di *Watson* risiede nel saper svolgere un compito (apparentemente) più facile e senza dubbio molto più comune che dare scacco matto. *Watson* è infatti in grado di rispondere a domande formulate in linguaggio naturale, grazie all'integrazione di alcune delle più sofisticate tecnologie per il Trattamento Automatico della Lingua (TAL). Il programma riceve in input una breve traccia linguistica (es. *The first man mentioned by name in the 'Man in the Iron Mask' is this hero of a previous book by the same author*) e deve decidere nel giro di una manciata di secondi se provare a rispondere scommettendo la cifra in palio, fornendo poi in caso positivo la riposta (es. *D'Artagnan*). *Watson* analizza linguisticamente la traccia iniziale identificandone la struttura sintattica e predicativa, genera un insieme di risposte potenziali sulla base degli indizi estratti dall'input e delle informazioni disponibili nella base di conoscenza del sistema, e infine produce la risposta con la maggiore probabilità di correttezza. Nonostante il successo mediatico ottenuto e le sue prestazioni certamente notevoli, *Watson* non

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

è molto diverso da altri sistemi per il TAL: entrambi sono dotati di *conoscenze linguistiche* attraverso le quali analizzano testi per estrarne il contenuto informativo allo scopo di rispondere a domande, come nel caso di *Watson*, oppure aumentare la naturalezza dell'interazione tra uomo e computer, tradurre testi in varie lingue, migliorare la ricerca e la gestione delle informazioni, ecc. Oltre a *lessici* con informazioni morfosintattiche e semantiche, le conoscenze linguistiche dei sistemi per il TAL comprendono *grammatiche computazionali* per analizzare automaticamente un numero potenzialmente illimitato di strutture linguistiche.

Il successo di *Watson* non sarebbe stato possibile fino a solo venti anni fa. In parte questo è dovuto all'enorme ampliamento delle capacità di calcolo e di memorizzazione dei computer, unitamente alla disponibilità di quantità prima inimmaginabili di informazioni digitali che consentono ai sistemi informatici di possedere le conoscenze necessarie per rispondere anche alle domande di un quiz televisivo[1]. Ma non è solo una questione di forza bruta di calcolo o di memoria. Il procedimento con cui *Watson* individua la riposta corretta è infatti intrinsecamente *statistico* e *probabilistico*. Il sistema decide di giocare, se la probabilità di trovare la risposta giusta supera una soglia di rischio che varia a seconda dell'andamento del gioco. Statistici sono molti dei moduli di analisi linguistica del sistema. Infine, la risposta stessa è probabilistica: *Watson* fornisce una serie di risposte ordinate secondo la loro probabilità di correttezza, calcolata dagli algoritmi del sistema integrando un insieme complesso di informazioni derivate dall'analisi linguistica dell'input e dalla base di conoscenza del sistema. Il successo di *Watson* è dunque il simbolo del predominio nella linguistica computazionale di ultima generazione dei *modelli statistici* rispetto a quelli tradizionali *a regole* che hanno invece rappresentato il paradigma dominante

[1] Watson può infatti contare su una mole enorme di informazioni acquisite automaticamente da enciclopedie e in particolare da Internet (es. da Wikipedia).

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

fino agli anni Ottanta del secolo scorso[2]. In realtà, i termini usati per definire questa opposizione sono fuorvianti. La nozione di regola non è infatti di per sé incompatibile con la statistica o la probabilità. La differenza risiede piuttosto nel *tipo di regole* che definiscono la grammatica.

Una grammatica è un modello delle *regolarità di una lingua*, che possiamo caratterizzare, con le parole di Zellig S. Harris, come *deviazioni dall'equiprobabilità* ("departures from equiprobability"):

> Each constraint that creates the partial order of words is a departure from randomness in this language universe, and yields a meaning. The information in a sentence or a discourse is thus formed by departures from equiprobability (Harris, 1991: 355).

Lo stato di equiprobabilità o *randomness* corrisponde alla situazione ipotetica in cui qualunque permutazione dell'ordine delle parole sia legittima e in grado di veicolare esattamente lo stesso contenuto di informazione. Il linguaggio è ovviamente una costante violazione di tale *randomness*: la sequenza *il cane ha morso un uomo* è una frase grammaticale, mentre la sua permutazione *cane il morso ha un uomo* non lo è. La sequenza *un uomo ha morso il cane* è altrettanto grammaticale della frase originale, ma veicola un contenuto informativo molto differente. Il modello tradizionale della grammatica rappresenta tali deviazioni dall'equiprobabilità attraverso la dicotomia tra strutture linguistiche *possibili*, ovvero grammaticali, ed *impossibili*, ovvero non grammaticali, usando la nozione di regola come strumento per tracciare i confini di tale partizione esclusiva. Questa visione 'a regole' della grammatica è ben descritta dalle parole di Edelman (2008: 247):

[2] Watson è più precisamente un sistema ibrido, poiché contiene al suo interno moduli a regole accanto a componenti statistiche, che comunque rappresentano l'asse portante della strategia di interpretazione della domanda e di generazione della risposta.

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

> For most of the 20th century, linguists assumed that grammars consist of algebraic RULES, of which there were supposed to be many fewer than the number of entries in the lexicon. On this view, the charge set upon the discipline of linguistics is to come up with a concise system of formal rules that jointly generate all the grammatical sentences in a given language, and none of the ungrammatical ones.

I sistemi a regole per il TAL incorporano questa visione algebrica della grammatica e svolgono i compiti di elaborazione linguistica (come associare a ogni parola la propria categoria morfosintattica, rispondere a una domanda, tradurre una frase o identificarne la struttura sintattica, ecc.) attraverso l'uso di regole che manipolano *strutture di simboli formali*. Il ruolo del linguista computazionale è dunque quello di individuare l'insieme di regole ottimale per la risoluzione di un particolare compito linguistico. Assegnare un'analisi a una frase è analogo a *dimostrare un teorema*: una frase *f* è analizzata come *a*, se e solo se *a* può essere ottenuta a partire da *f* attraverso una serie finita di applicazioni di regole della grammatica.

Le grammatiche variano per il metalinguaggio algebrico con cui vengono espresse le regole: sistemi di riscrittura sintagmatica, strutture ricorsive di coppie attributo-valore, grammatiche categoriali, automi, ecc. In ogni caso, le regole formali sono accomunate dal fatto di essere *discrete*, *qualitative* e *inviolabili* (*salvo eccezioni!*)[3]. Gli aspetti quantitativi o quelli relativi a dimensioni di variazione continua e graduale delle strutture linguistiche rimangono al di fuori del raggio di modellazione delle regole. La variazione nella *frequenza* di uso delle strutture linguistiche è un fattore che non viene rappresentato nelle regole formali. La grammatica opera in termi-

[3] Un'"eccezione" a questa affermazione è rappresentata dai modelli della grammatica basati su Optimality Theory, i cui vincoli (almeno nella versione non probabilistica della teoria) sono qualitativi, ma ordinati e violabili. Le violazioni dei vincoli grammaticali sono ammesse, purché vengano soddisfatti vincoli gerarchicamente dominanti (cfr. Lenci, 2005).

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

ni di opposizioni categoriali e qualitative: grammaticale vs. non grammaticale, nome vs. verbo, argomento vs. aggiunto, transitivo vs. intransitivo, animato vs. non animato, ecc. Un caso esemplare è la rappresentazione dei vincoli semantici che regolano le combinazioni predicato-argomento. I predicati corrispondono a funzioni associate a tipi semantici che specificano le categorie legittime di argomenti. Si consideri ad esempio la seguente coppia di frasi:

(1) a. *Il sasso ha ucciso l'uomo.*
b. **L'uomo ha ucciso il sasso.*

Un sistema simbolico può riconoscere il contrasto di grammaticalità di queste frasi avendo a disposizione le seguenti informazioni rappresentate come regole formali:

(2) a. $R(x{:}\alpha)\ A{:}\alpha \rightarrow R(A)$
b. *uccidere*: [SN_{ogg}: [+ANIMATO]]
c. *uomo*: [+ANIMATO]
d. *sasso*: [–ANIMATO]

La regola (2a) stabilisce una condizione generale sulla combinazione di predicati ed argomenti: un predicato $R(x)$ può essere applicato a un argomento A se e solo se il tipo semantico di A è equivalente a quello richiesto da R. La regola in (2b) specifica invece un vincolo di selezione lessicale del verbo *uccidere*: l'oggetto diretto deve essere animato. Una volta combinati i vincoli (2a, b) con le informazioni sui tipi semantici di *uomo* e *sasso* in (2c, d), un sistema computazionale a regole può derivare la grammaticalità di (1a) e la non grammaticalità di (1b).

Per essere grammaticale un'espressione linguistica non deve violare alcuna regola della grammatica, a meno che essa non venga dichiarata esplicitamente come *eccezione*. Le frasi seguenti, tratte dal corpus itWaC (Baroni *et al.*, 2009), rappresentano però chiare violazioni della regola (2b):

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

(3) a. *La burocrazia uccide le idee.*
b. *Il terrorismo uccide la democrazia.*
c. *Hai ucciso il mio amore.*

Poiché gli oggetti diretti non sono animati, queste frasi sono analizzabili solo specificando i nomi *idea*, *democrazia* e *amore* come eccezioni alla regola in (2b). Questa può essere concepita come una regola *default*, ovvero una regola generalmente valida *salvo particolari eccezioni*. Le strutture 'eccezionali' sono pertanto tipicamente rappresentate come un modulo qualitativamente distinto rispetto all'insieme potenzialmente aperto di strutture generate dalle regole della grammatica. Le eccezioni 'immagazzinate' in una struttura statica e finita (es. una base di conoscenza lessicale) si contrappongono al componente dinamico e autenticamente generativo delle regole. Questa architettura tradizionale dei sistemi per il TAL è del tutto simile a molti modelli tipici della tradizione simbolica razionalista in linguistica. Le esemplificazioni di segregazioni tra regole ed eccezioni sono innumerevoli, come ad esempio i modelli *dual route* della flessione in morfologia (Pinker / Ullman, 2002), o le teorie che collocano in sottosistemi qualitativamente differenti della grammatica le strutture sintattiche produttive e le strutture semi-idiomatiche lessicalizzate[4].

A differenza dei sistemi che rappresentano le regolarità della lingua con regole discrete e qualitative, i modelli statistici le rappresentano come *vincoli probabilistici*. La probabilità serve per modellare matematicamente eventi aleatori, che possono avere esiti diversi e con gradi variabili di incertezza sul loro accadimento. La probabilità è uno strumento quantitativo che ci consente di ragionare in una situazione di incertezza, facendo previsioni sul possibile verificarsi di un evento:

[4] Per una critica a tale dicotomia si veda Culicover / Jackendoff (2005), oltre che le teorie di tipo più dichiaratamente 'costruzionista', come Goldberg (1995) e Boas / Sag (2012).

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

> The argument for a probabilistic approach to cognition is that we live in a world filled with uncertainty and incomplete information. To be able to interact successfully with the world, we need to be able to deal with this type of information. […] The cognitive processes used for language are identical or at least very similar to those used for processing other forms of sensory input and other forms of knowledge. These processes are best formalized as probabilistic processes or at least by means of some quantitative framework that can handle uncertainty and incomplete information (Manning / Schütze, 1999: 15).

I valori della probabilità variano con continuità tra zero, che quantifica l'impossibilità di un evento, e uno, il valore assunto da un evento che accade con assoluta certezza. Lo spazio delle regole probabilistiche non è discreto, e le strutture linguistiche possibili in una lingua sono modellate con il *continuum* delle distribuzioni di probabilità, senza ridursi a pure opposizioni categoriali. Le probabilità dei vincoli della grammatica sono ricavate automaticamente dalla *distribuzione statistica* degli eventi linguistici osservati in corpora testuali. La frequenza di occorrenza di un evento linguistico viene infatti usata per stimare la sua probabilità[5]. Le regole della grammatica sono dunque rappresentate come generalizzazioni induttive che catturano *regolarità statistiche* presenti nell'uso linguistico. Nei sistemi probabilistici, il ruolo del linguista non è 'scrivere' le regole della grammatica, bensì *addestrare* il sistema a svolgere un dato compito, individuando la metodologia migliore che consenta al sistema stesso di estrarre dalla distribuzione statistica dei dati linguistici i vincoli e le regole per svolgerlo.

Le regole probabilistiche sono per loro intrinseca definizione violabili. Invero, superano la dicotomia stessa di regole ed ecce-

[5] Ad esempio, la probabilità di una parola *x* può essere stimata come il rapporto tra la frequenza di x in un corpus e il numero complessivo di parole nel corpus.

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

zioni, nella misura in cui strutture produttive e strutture eccezionali sono rappresentate nel medesimo spazio probabilistico. Ad esempio, le preferenze di selezione dei predicati verbali possono essere modellate attraverso una distribuzione di probabilità $P(N|V_r)$, che rappresenta la probabilità che un nome N sia l'argomento del verbo V nel ruolo grammaticale r (es. oggetto diretto). Questa probabilità può essere calcolata con il rapporto tra il numero di volte in cui N ricorre con V in un corpus con il ruolo r, e la frequenza totale con cui V ricorre con il ruolo r. Per esempio, $P(uomo|uccidere_{ogg})$ corrisponde alla probabilità di osservare *uomo* come oggetto diretto di *uccidere* e può essere stimata con il rapporto tra la frequenza di *uomo* come oggetto di *uccidere*, e il numero di volte con cui questo verbo ricorre con un oggetto diretto. Se utilizziamo il corpus itWaC per addestrare il nostro modello probabilistico delle preferenze di selezione di *uccidere*, otteniamo la seguente distribuzione di probabilità per gli oggetti diretti che compaiono nelle frasi (1) e (3)[6]:

(4) a. $P(uomo|uccidere_{ogg}) = 0{,}04$
b. $P(amore|uccidere_{ogg}) = 0{,}0016$
c. $P(idea|uccidere_{ogg}) = 0{,}0013$
d. $P(democrazia|uccidere_{ogg}) = 0{,}0008$
e. $P(sasso|uccidere_{ogg}) = 0$

La natura 'eccezionale' di *idea* come oggetto diretto di *uccidere*, rispetto alla 'regolarità' di *uomo*, viene rappresentata attraverso la differenza delle loro probabilità. L'impossibilità di *sasso* diventa solo l'estremo di un *continuum* che contempla anche casi poco probabili, ma pur sempre possibili. Mentre i sistemi di regole formali modellano la "departure from equiprobability" della grammatica con la dicotomia tra strutture possibili e impossibili, i modelli probabilistici sono invece in grado di riempire lo spa-

[6] Le probabilità sono state stimate utilizzando dati estratti da itWaC con la funzionalità "word sketch" su *Sketchengine* (http://www.sketchengine.co.uk/).

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

zio che intercorre tra questi due insiemi, individuando variazioni di probabilità all'interno dell'insieme delle strutture legittimate dalla grammatica.

I modelli probabilistici rappresentano ormai il paradigma dominante nel TAL. Quali sono le ragioni del loro successo? Vi sono prima di tutto due motivi più di natura 'tecnica', ma assolutamente non secondari. Poiché i modelli probabilistici si basano su grammatiche acquisite automaticamente dal sistema a partire dai dati statistici, essi consentono una maggiore rapidità di sviluppo rispetto ai modelli a regole che dipendono da grammatiche sviluppate manualmente. Inoltre, gli algoritmi per l'addestramento dei modelli statistici sono del tutto indipendenti dalla lingua. Un sistema può imparare a svolgere il medesimo compito linguistico in più lingue diverse, a patto che sia addestrato su dati adeguati. Lo stesso algoritmo per rappresentare le preferenze semantiche di *uccidere* può essere applicato a *kill* o *töten*, avendo a disposizione un corpus dell'inglese o del tedesco dal quale estrarre le statistiche necessarie per stimare le probabilità.

Il motivo più sostanziale della fortuna dei modelli statistici è legato alle loro migliori prestazioni nello svolgere compiti linguistici. I sistemi probabilistici hanno una maggiore robustezza nell'affrontare la variabilità della lingua. Le applicazioni per il TAL si trovano ad operare in condizioni ben lontane da quelle spesso idealizzate che sono fotografate da qualunque sistema di regole. La situazione di un sistema linguistico-computazionale in questo senso assomiglia molto a quella di un sociolinguista: entrambi devono gestire e modellare la variabilità dell'uso linguistico. Non è dunque un caso che sia in sociolinguistica che nel TAL si siano diffusi modelli di tipo probabilistico (si veda ad esempio la nozione di regola variabile illustrata da Berruto e da Cerruti nei loro contributi in questo volume). I sistemi per il TAL 'aperti' come *Watson*, ovvero non specializzati su un particolare dominio, sono sempre più spesso chiamati ad affrontare uno

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

spettro di variabilità linguistica amplissima, che va dalla prosa giornalistica fino ai microtesti su Twitter o Facebook. La variazione delle strutture della lingua è estremamente alta: essa riguarda tutti i livelli, dall'ortografia alla semantica, e difficilmente può essere gestita con insiemi di regole formali tradizionali. L'affermazione di Sapir (1921: 38) che "all grammars leak" è dunque ancor più vera per le grammatiche computazionali. Si pensi alle frasi in (3) che violano il vincolo sull'animatezza dell'oggetto di *uccidere*. Si può certamente replicare sostenendo che queste sono frasi metaforiche, ma ciò sposta semplicemente i termini della questione senza risolverla. Ugualmente non efficace è la soluzione di assumere che non si tratti dello stesso verbo *uccidere*. La semplice moltiplicazione dei sensi dei lessemi, oltre a non essere soddisfacente sul piano teorico (Pustejovsky, 1995), non aiuta un sistema computazionale ad affrontare il problema del loro riconoscimento.

Il vero tallone di Achille dei sistemi a regole è però la gestione delle *ambiguità*. Si consideri ad esempio il seguente esempio:

(5) *La banca centrale ha abbassato i tassi di interesse di tre punti per tre anni*.

I sistemi per il TAL scompongono l'analisi linguistica in una serie di fasi che comprendono la segmentazione del testo in input, l'analisi morfologica e la disambiguazione morfosintattica delle unità lessicali, l'analisi sintattica e infine l'interpretazione semantica. A ciascuno di questi livelli, (5) contiene molteplici casi di ambiguità: *centrale* può essere un nome, un aggettivo oppure anche l'imperativo del verbo *centrare* con un pronome clitico, *tasso* può riferirsi all'omonimo animale oppure a una quantità monetaria o essere una forma del verbo *tassare*, ecc. I sistemi a regole assegnano a una espressione linguistica *tutte le analisi compatibili con la grammatica*, ma non forniscono un criterio di scelta tra queste analisi. Ecco, ad esempio, quattro analisi sintattiche al-

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

ternative del sintagma verbale in (5), in realtà un sottoinsieme di tutte quelle teoricamente possibili[7]:

(6) a. [*ha abbassato* [*i tassi di interesse*] [*di tre punti*] [*per tre anni*]].
b. [*ha abbassato* [*i tassi di interesse* [*di tre punti*]] [*per tre anni*]].
c. [*ha abbassato* [*i tassi di interesse*] [*di tre punti* [*per tre anni*]]].
d. [*ha abbassato* [*i tassi di interesse* [*di tre punti* [*per tre anni*]]]].

Naturalmente l'analisi corretta in questo caso è (6a), ma le altre sono comunque legittime combinazioni di costituenti preposizionali in italiano: se l'input fosse *ha abbassato i tassi di interesse dei conti correnti per tre anni*, l'analisi corretta sarebbe (6b), con il SP *dei conti correnti* modificatore dell'oggetto diretto. I sistemi probabilistici sono in grado di risolvere il problema dell'ambiguità sfruttando un fatto fondamentale, ovvero che le analisi alternative, sebbene possibili, non sono tutte equiprobabili. Questo è ciò che rende la maggior parte delle ambiguità invisibili, dal momento che il contesto è generalmente in grado di fornirci informazioni sufficienti a scegliere l'analisi o interpretazione appropriata. Per esempio, la probabilità che *tasso* si riferisca a un animale è molto bassa dato il fatto che nella stessa frase si trovano parole come *banca* oppure *interesse*. I modelli probabilistici possono determinare qual è l'analisi più probabile di una struttura linguistica in un dato contesto, combinando informazioni sulle distribuzioni statistiche delle strutture linguistiche ricavate dai corpora. L'abilità dei sistemi probabilistici di risolvere e gestire le ambiguità nel linguaggio è uno dei motivi fondamentali che spiegano le prestazioni di un sistema come *Watson*.

[7] Le parentesi sono usate per indicare l'incassamento dei costituenti sintagmatici.

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

Naturalmente anche i modelli probabilistici della grammatica hanno il loro tallone di Achille. Questo è rappresentato dalla natura finita dei corpora sui quali vengono stimate le probabilità e dalla distribuzione zipfiana dei dati linguistici, che sono sistematicamente affetti da una rarità di attestazioni (Lenci *et al.*, 2005). In (4), P(*sasso*|*uccidere*$_{ogg}$) è uguale a zero, perché in itWaC *sasso* non ricorre mai come oggetto di *uccidere*. Dal momento che anche *cammello* non compare con questo verbo nel corpus, il modello assegna zero anche a P(*cammello*|*uccidere*$_{ogg}$), inferendo incorrettamente che *cammello* non è un argomento possibile di *uccidere*. Il fatto che le dimensioni di itWaC siano comunque ragguardevoli, con quasi di 2 miliardi di parole, mostra l'importanza del fenomeno della rarità dei dati linguistici. Casi come questo sono da sempre citati come argomenti contro la plausibilità dei modelli statistici in linguistica. In realtà, mostrano più limitatamente che la stima delle probabilità dei vincoli della grammatica richiede metodi più sofisticati di quelli usati in (4). I modelli di ultima generazione sono in grado di attenuare l'effetto negativo della rarità dei dati linguistici consentendo stime più accurate delle probabilità dei vincoli della grammatica, anche se il problema delle strutture grammaticali non osservate nei corpora continua comunque a gettare ombre sulle prestazioni e sulla plausibilità linguistica dei modelli statistici.

Watson rappresenta sicuramente un esempio delle grandi potenzialità dei sistemi attuali per il TAL, e in particolare del contributo offerto dai nuovi approcci di tipo statistico. Un'obiezione però sorge spontanea. I metodi probabilistici possono anche essere la migliore soluzione ingegneristica attualmente sul mercato per creare un sistema artificiale che vinca un quiz televisivo, senza per questo avere nessuna particolare rilevanza per lo studio del linguaggio. Si potrebbe anzi argomentare che l'approccio statistico è proprio un segno della distanza dei sistemi per il TAL rispetto alle modalità di elaborazione linguistica umana: i sistemi informatici sono costretti a perseguire strategie di anali-

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

si probabilistica, proprio perché non possiedono la competenza linguistica di un parlante nativo. In effetti, questa critica coglie parzialmente nel vero. Come si è detto sopra, anche i sistemi più evoluti per il TAL sono tuttora legati a una strategia sequenziale di elaborazione linguistica. Ad esempio, i moduli di analisi sintattica non hanno generalmente accesso a conoscenze di tipo semantico. Il problema dell'ambiguità di (5) non esisterebbe se il sistema fosse dotato di un maggiore grado di parallelismo nell'accesso all'informazione semantica e pragmatica, così come è tipico per l'elaborazione umana. In realtà, però, i modelli probabilistici non hanno solo una valenza ingegneristica. Chater *et al.* (2006) mostrano ad esempio il ruolo fondamentale dei modelli probabilistici per la comprensione dei processi cognitivi; la natura probabilistica dei processi umani di acquisizione ed elaborazione del linguaggio è invece bene illustrata e argomentata in Jurafsky (2003), Manning (2003) e Chater / Manning (2006). Uno dei vantaggi dei modelli probabilistici è proprio quello di consentire un elevato parallelismo dell'elaborazione del linguaggio, problematico invece per i modelli a regole che devono affrontare il complesso problema dell'ordinamento delle regole e della loro interazione. In questo senso, con tutte le differenze del caso, il metodo probabilistico con cui *Watson* elabora il linguaggio è forse più simile a quello umano di quanto non sembri a prima vista.

Anche ammesso che l'elaborazione del linguaggio sia probabilistica, è possibile che questa dimensione riguardi solo l'*uso* del linguaggio. Si potrebbe dunque continuare a modellare la competenza grammaticale con sistemi di regole formali discrete, limitando gli effetti probabilistici alla sfera dell'esecuzione. Sebbene questo tipo di ipotesi sia del tutto legittima, si scontra però con una ricca serie di evidenze empiriche che mostrano come molti fenomeni della grammatica, difficilmente riducibili ad effetti di *performance*, si collochino invece anch'essi nell'ambito del continuo (Manning, 2003; Fanselow *et al.*, 2006). Un caso esemplare è dato dalla distinzione tra complementi e aggiunti che gioca

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

un ruolo chiave in ogni descrizione sintattica. A partire da Vater (1978), la sua caratterizzazione come opposizione categoriale è stata spesso messa in discussione, dal momento che vi sono molte strutture che violano tutti i test sintattici normalmente utilizzati per decidere sullo status di aggiunto o di complemento, e che, dunque, si pongono come casi intermedi e non pienamente decidibili. Somers (1984) propone ad esempio che complementi e aggiunti formino un'opposizione scalare, e Dowty (2003) arriva ad ipotizzare che ogni sintagma dovrebbe essere simultaneamente rappresentato come argomento e aggiunto. Le evidenze empiriche suggeriscono che la differenza tra argomenti e aggiunti sia di natura gradiente e, dunque, più affine ad una modellazione continua di tipo probabilistico. Si può obiettare che spostarsi verso la dimensione probabilistica significhi annullare le opposizioni categoriali nell'indeterminatezza del caso. In realtà, come si è visto con le preferenze di selezione, la modellazione probabilistica è perfettamente compatibile con la presenza di estremi ben distinti e fortemente polarizzati. Non si deve dimenticare che certezza e impossibilità appartengono comunque allo spettro dei valori assegnati dalle distribuzioni di probabilità. Sostenere una rappresentazione probabilistica della distinzione tra argomenti e aggiunti non significa quindi negare che questa distinzione esista o che non ci siano casi incontrovertibili di complementi o di aggiunti. Significa piuttosto mettere la grammatica in grado di rappresentare anche casi di variazione graduale tra i due estremi. L'esplorazione delle potenzialità dei modelli probabilistici promette una maggiore capacità descrittiva della complessità, gradualità e variabilità dei fenomeni linguistici. Il loro successo nella realizzazione di strumenti per il TAL non può dunque essere confinato al semplice dominio dell'applicazione ingegneristica, ma apre interessanti prospettive anche per la descrizione linguistica e la modellazione cognitiva. Spesso i sistemi per il TAL sono stati tacciati di irrilevanza per lo studio del linguaggio, ma l'approccio probabilistico può gettare nuovi ponti tra teoria ed applicazione,

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

per una visione diversa della grammatica e della nozione stessa di regola.

Bibliografia

Baroni, M. / Bernardini, S. / Ferraresi A. / Zanchetta, E. (2009), *The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora*, "Language Resources and Evaluation" 43.3, 209-226.

Boas, H.C. / Sag, I.A. (2012), *Sign-Based Construction Grammar*, Stanford (CA), CSLI.

Bod, R. / Hay, J. / Jannedy, S. (eds.) (2003), *Probabilistic Linguistics*, Cambridge (Mass.), The MIT Press.

Chater, N. / Manning, C. D. (2006), *Probabilistic models of language processing and acquisition*, "TRENDS in Cognitive Sciences" 10.7, 335-344.

Chater, N. / Tenenbaum, J. / Yuille, A. (2006), *Probabilistic models of cognition: Conceptual foundations*, "TRENDS in Cognitive Sciences" 10.7, 287-291.

Culicover, P.W. / Jackendoff, R. (2005), *Simpler Syntax*, Oxford, Oxford University Press.

Dowty, D. (2003), *The Dual Analysis of Adjuncts/Complements in Categorial Grammar*, in E. Lang / C. Maienborn / C. Fabricius-Hansen (eds.) (2003), *Modifying Adjuncts*, Berlin, Mouton de Gruyter, 33-66.

Edelman, S. (2008), *Computing the Mind*, Oxford, Oxford University Press.

Fanselow, G. / Féry, C. / Schlesewsky, M. / Vogel, R. (2006), *Gradience in Grammar. Generative Perspectives*, Oxford, Oxford University Press.

Ferrucci, D. (2012), *Introduction to 'This is Watson'*, "IBM Journal of Research and Development" 56.3,4, 1-15.

Goldberg, A. (1995), *Constructions. A Construction Grammar Approach to Argument Structure*, Chicago (IL), Chicago University Press.

Harris, Z.S. (1991), *A Theory of Language and Information: A Mathematical Approach*, Oxford, Clarendon Press.

Jurafsky, D. (2003), *Probabilistic modeling in psycholinguistics: Linguistic comprehension and production*, in R. Bod / J. Hay / S. Jannedy (eds.) (2003), 39-95.

PDF concesso da BUP all'autore per l'espletamento delle procedure concorsuali

Lenci, A. (2005), *La sintassi tra ottimalità e probabilità. Soggetti e oggetti in una grammatica stocastica dell'italiano*, "Studi e Saggi Linguistici" 62, 43-87.

Lenci, A. / Montemagni, S. / Pirrelli, V. (2005), *Testo e computer. Elementi di linguistica computazionale*, Roma, Carocci.

Manning, C.D. (2003), *Probabilistic syntax*, in R. Bod / J. Hay / S. Jannedy (eds.) (2003), 289-341.

Manning, C.D. / Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge (Mass.), The MIT Press.

Pinker, S. / Ullman, M.T. (2002), *The past and future of the past tense*, "Trends in Cognitive Sciences" 6.11, 456-474.

Pustejovsky, J. (1995), *The Generative Lexicon*, Cambridge (Mass.), The MIT Press.

Sapir, E. (1921), *Language: An Introduction to the Study of Speech*, New York (NY), Harcourt Brace.

Somers, H. (1984), *On the validity of the complement-adjunct distinction in valency grammar*, "Linguistics" 22, 507-530.

Vater, H. (1978), *On the possibility of distinguishing between complements and adjuncts*, in W. Abraham (ed.) (1978), *Valence, Semantic Case and Grammatical Relations*, Amsterdam, John Benjamins, 21-45.