COMPTER PARLER SOIGNER

Tra linguistica e intelligenza artificiale

Atti

Pavia, Collegio Ghislieri, 15-17 dicembre 2014

a cura di

Edoardo Maria Ponti – Marco Budassi

PaviaUniversityPress

Compter parler soigner : tra linguistica e intelligenza artificiale : atti : Pavia, Collegio Ghislieri, 15-17 dicembre 2014 / a cura di Edoardo Maria Ponti, Marco Budassi. – Pavia : Pavia University Press, 2016. – XIX, 119 p. ; 24 cm.

(Atti)

http://archivio.paviauniversitypress.it/oa/9788869520389.pdf

ISBN 9788869520372 (brossura)
ISBN 9788869520389 (e-book PDF)

In testa al front.: Ghislieri.

ISBN: 978-88-6952-038-9

Nella sezione *Scientifica* Pavia University Press pubblica esclusivamente testi scientifici valutati e approvati dal Comitato scientifico-editoriale.

In copertina: *cover del disco Golden Record, a bordo della sonda Voyager lanciata nel 1977, contenente saluti in 55 lingue e un messaggio vocale di Jimmy Carter*.

Prima edizione: luglio 2016

Pavia University Press – Edizioni dell'Università degli Studi di Pavia
Via Luino, 12 – 27100 Pavia (PV) – Italia
www.paviauniversitypress.it – unipress@unipv.it

Printed in Italy

Semantica distribuzionale. Un modello computazionale del significato

Alessandro Lenci – Università di Pisa – alessandro.lenci@unipi.it

1. Introduzione

I modelli semantici distribuzionali rappresentano il significato delle parole attraverso l'analisi statistica dei contesti linguistici in cui ricorrono. La similarità semantica dei lessemi è misurata sulla base della similarità delle loro distribuzioni contestuali costruite con informazioni estratte da corpora testuali. Nella semantica distribuzionale il significato è dunque una proprietà emergente dall'uso delle parole nei contesti linguistici. L'approccio distribuzionale all'analisi semantica è stato oggetto di intensa ricerca negli ultimi anni, sia in Linguistica Computazionale che nelle Scienze Cognitive. L'uso di informazioni distribuzionali estratte da corpora per misurare la similarità semantica tra termini lessicali è diventato un paradigma centrale nel Trattamento Automatico della Lingua (TAL). Lo sviluppo della semantica distribuzionale ha beneficiato della disponibilità sempre crescente di dati testuali e di una maggiore potenza di calcolo, che hanno consentito lo sviluppo di nuovi metodi di analisi dei contesti linguistici e la loro applicazione su larga scala, per sviluppare modelli realistici del lessico semantico utilizzabili sia in contesti applicativi, sia per modellazioni linguistiche e (neuro)cognitive.

Dopo un lungo percorso iniziato più di mezzo secolo fa, con periodi alterni di oblio e popolarità dovuti al succedersi di diversi paradigmi di ricerca dominanti nel TAL e in linguistica, la semantica distribuzionale ha raggiunto oggi una sua maturità. Molti algoritmi per la costruzione di spazi semantici distribuzionali e dati per la loro valutazione sono ora disponibili. Numerosi articoli di rassegna hanno contribuito a consolidare e approfondire gli aspetti metodologici dei modelli distribuzionali (Turney, Pantel, 2010; Baroni, Lenci, 2010). Negli ultimi anni, la ricerca ha concentrato molti dei suoi sforzi sullo sviluppo di metodi di ottimizzazione dei modelli per la loro applicazione a corpora di grandi dimensioni e sullo studio e valutazione dei parametri che determinano effetti significativi sulla qualità e sulla natura delle rappresentazioni semantiche costruite dai modelli distribuzionali. Tuttavia, questioni importanti rimangono ancora aperte, sia per giungere a una migliore comprensione del tipo di informazioni che questi metodi permettono di estrarre dai dati linguistici, sia per estendere la loro applicazione alla modellazione di nuovi fenomeni semantici.

Lo scopo di questo contributo è fornire una breve rassegna dello stato dell'arte della semantica distribuzionale nell'ambito dei modelli computazionali dell'analisi linguistica. Dopo una presentazione dei meccanismi principali alla base dei modelli distribuzionali e della loro valutazione sperimentale, illustreremo alcune sfide che questa metodologia di analisi semantica deve affrontare.

2. Principi e metodi di semantica distribuzionale

2.1. L'Ipotesi Distribuzionale

Il fondamento epistemologico della semantica distribuzionale è l'Ipotesi Distribuzionale: la similarità semantica si correla con la similarità delle distribuzioni nei contesti linguistici. Zellig S. Harris è normalmente ritenuto il pioniere dell'Ipotesi Distribuzionale (Harris, 1954). Infatti egli considerava la metodologia distribuzionale come l'unico approccio scientifico possibile allo studio del significato linguistico. Nei suoi ultimi lavori, Harris propose un metodo per classificare le parole sulla base dei loro contesti attraverso la raccolta e l'analisi delle relazioni di dipendenza sintattica, espresse in termini di operatori e argomenti (Harris, 1991).

A partire dagli anni Sessanta molte implementazioni dell'Ipotesi Distribuzionali sono state realizzate per la costruzione automatica di *thesauri* (G. Grefenstette, 1994). Un contributo cruciale allo sviluppo della semantica distribuzionale è infatti giunto dal *vector space model* nell'*Information Retrieval* (Salton et al., 1975), che ha permesso di apportare miglioramenti fondamentali alla metodologia originale di Harris rispetto sia alla natura dei dati che alla loro formalizzazione matematica, accelerando così la diffusione del paradigma distribuzionale in linguistica computazionale. Negli ultimi venti anni, la possibilità di applicare tale metodologia su ampia scala a corpora di grandi dimensioni ha fatto sì che l'approccio distribuzionale sia diventato il paradigma semantico di riferimento nel TAL.

2.2. La struttura dei modelli semantici distribuzionali

Nei modelli distribuzionali, le parole sono rappresentate come vettori costruiti a partire dalla loro distribuzione nei contesti linguistici, e la similarità tra le parole è approssimata attraverso la misura della distanza geometrica tra vettori. Il metodo standard per costruire modelli semantici distribuzionali è tipicamente formato da quattro fasi principali (Turney, Pantel, 2010):

1. per ciascuna parola target, vengono raccolti e contati i contesti generando così una matrice di co-occorrenza;

2. le frequenze sono poi trasformate in pesi statistici in grado di riflettere meglio l'importanza dei contesti;

3. la matrice che si ottiene è molto grande e sparsa, ovvero la maggior parte delle sue entrate è zero. Per tale motivo, vengono applicate tecniche matematiche per ridurre il numero delle sue dimensioni;

4. la similarità semantica delle parole target viene misurata attraverso la similarità dei corrispondenti vettori riga nella matrice.

I vettori costituiscono il principale strumento di rappresentazione matematica del contenuto lessicale nella semantica distribuzionale. Un vettore è una lista ordinata di numeri reali $(v_1, ..., v_n)$, nella quale ciascun valore v_i è l'i-esima componente del vettore. I vettori hanno interpretazioni geometriche. Se si assume un sistema di

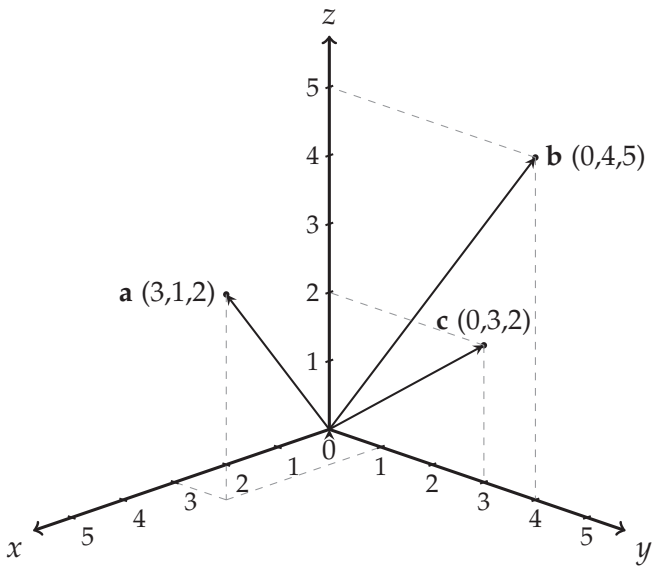

Figura 4. Vettori in uno spazio tridimensionale

assi cartesiani come quello nella Figura 4., i vettori **a**, **b**, e **c** identificano punti nello spazio e le loro componenti corrispondono alle coordinate dei punti sugli assi. Gli stessi vettori possono essere anche rappresentati da frecce che congiungono un punto origine a un punto finale. L'origine è fissata all'incrocio degli assi cartesiani, e le coordinate del punto finale corrispondono alle componenti del vettore. I modelli distribuzionali forniscono, dunque, una rappresentazione geometrica del lessico come uno spazio vettoriale semantico.

Ad esempio, supponiamo di aver contato in un corpus quante volte i nomi *auto*, *gatto*, *cane* e *camion* co-occorrono con i verbi *mangiare*, *guidare* e *correre*, ottenendo la seguente distribuzione di frequenza:

auto	*guidare*	3	*cane*	*mangiare*	3
auto	*correre*	2	*cane*	*correre*	4
gatto	*mangiare*	4	*camion*	*guidare*	2
gatto	*correre*	3	*camion*	*correre*	3

Rappresentiamo dunque *auto*, *gatto*, *cane* e *camion* con i seguenti vettori, indicando il vettore distribuzionale di un lessema con il lessema stesso in grassetto:

auto = (0, 3, 2)
gatto = (4, 0, 3)
cane = (3, 0, 4)
camion = (0, 2, 3)

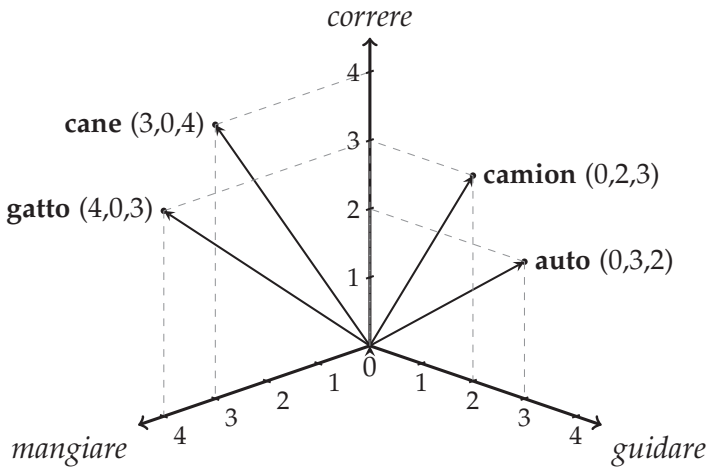

Figura 5. Vettori distribuzionali

La prima componente dei vettori è la frequenza di co-occorrenza con *mangiare*, la seconda con *guidare* e la terza con *correre* (zero indica il caso in cui una parola non ricorra mai in un dato contesto). Questi vettori corrispondono a punti (o frecce) nello spazio tridimensionale illustrato nella Figura 5.. Gli assi dello spazio sono etichettati con contesti linguistici (in questo caso specifico le parole *mangiare*, *guidare* e *correre*), e la posizione delle parole target nello spazio è determinata dalla loro frequenza di co-occorrenza. Come si è detto, invece delle frequenze è possibile usare vari tipi di pesi statistici (cfr. *infra*, Sezione 2.2.1.).

Una delle misure più comuni di similarità distribuzionale è il coseno dell'angolo θ tra due vettori:

$$\text{sim}_{\cos}(\mathbf{u},\mathbf{v}) = \cos(\theta) = \frac{\mathbf{u}\cdot\mathbf{v}}{||\mathbf{u}||||\mathbf{v}||} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}$$

I coseni dei vettori delle quattro parole target sono riportati nella tabella seguente:

auto	1			
gatto	0.33	1		
cane	0.44	0.96	1	
camion	0.92	0.50	0.66	1
	auto	*gatto*	*cane*	*camion*

Se due vettori sono geometricamente allineati sulla stessa linea e puntano nella medesima direzione, l'angolo tra di loro misura 0 gradi e il coseno è 1 (massima similarità); viceversa, se i due vettori sono indipendenti (ortogonali), il loro angolo è vicino a 90 gradi e il coseno è uguale a 0 (assenza di similarità). Il coseno

rappresenta dunque la similarità in termini geometrici e misura la similarità tra vettori con la loro vicinanza nello spazio. Nella Figura 5., **cane** e **gatto** sono molto vicini e puntano nella stessa direzione, perché hanno valori simili nelle stesse componenti. L'angolo che essi formano è dunque più piccolo di quello con **auto** e **camion**: più piccolo l'angolo, maggiore il coseno e la similarità tra i vettori.

I modelli distribuzionali hanno molteplici opzioni e possibilità di realizzazione, conseguenti a vari parametri in ciascuna fase del processo di costruzione. I valori di tali parametri possono modificare in maniera anche molto significativa la struttura dello spazio semantico.

2.2.1. Parametri

Un modello semantico distribuzionale riflette il comportamento delle parole nell'uso linguistico ed è quindi, per definizione, fortemente dipendente dal tipo di corpus che viene analizzato. Mentre negli anni Novanta venivano usati corpora specializzati di medie dimensioni per l'acquisizione e costruzione di *thesauri* distribuzionali (G. Grefenstette, 1994; Nazarenko et al., 2001), più di recente si è passati all'uso di corpora di grandi dimensioni, trasversali rispetto ai generi testuali e ai domini tematici. Sono stati usati in semantica distribuzionale corpora giornalistici o formati da articoli di enciclopedia (Peirsman, Geeraerts, 2009), corpora di riferimento bilanciati come il British National Corpus (Sadrzadeh, E. Grefenstette, 2011), grandi corpora ottenuti dal Web (Agirre et al., 2009), o combinazioni di corpora di varie tipologie (Baroni, Lenci, 2010). La tendenza a usare corpora di grandi dimensioni è principalmente motivata dalla necessità di aumentare la copertura delle risorse lessicali distribuzionali riducendo al contempo la sparsità dei dati, che notoriamente può avere effetti negativi sulla qualità degli spazi semantici.

Un altro parametro cruciale nell'implementazione dei modelli semantici distribuzionali è la definizione dei contesti. Tre tipi di contesti linguistici sono tipicamente usati: nei modelli basati su documenti (*document models)*, come la *Latent Semantic Analysis* (LSA) (Landauer, Dumais, 1997), le parole sono simili se appaiono negli stessi documenti o negli stessi paragrafi; nei modelli basati sulle parole (*word models)*, viene invece considerata una finestra di parole che ricorrono intorno ai lessemi target (Lund, Burgess, 1997; Sahlgren, 2008; Ferret, 2013); i modelli sintattici (*syntactic models)* sono vicini all'approccio originale di Harris, poiché usano come contesti le relazioni di dipendenza sintattica delle parole target (Curran, 2004; Padó, Lapata, 2007; Baroni, Lenci, 2010). I modelli basati sulle parole hanno un parametro aggiuntivo rappresentato dalla dimensione della finestra per la selezione delle parole contesto, finestra che può andare da poche parole a un intero paragrafo, mentre i modelli sintattici hanno bisogno di specificare le relazioni di dipendenza che sono selezionate per definire i contesti di co-occorrenza (Baroni, Lenci, 2010; Peirsman, Heylen et al., 2007). Alcuni esperimenti suggeriscono che i modelli sintattici tendano a identificare vicini distribuzionali che sono tassonomicamente correlati al target, principalmente co-iponimi, mentre i modelli basati sulle parole sono più orientati verso l'identificazione di relazioni associative (Van de Cruys, 2008; Peirsman, Heylen et al., 2007; O. Levy, Goldberg, 2014). La questione se i contesti sintattici forniscano veramente un vantaggio reale rispetto ai

modelli basati su finestre di parole è comunque ancora aperta. Una differenza molto più sostanziale esiste invece nei confronti dei modelli basati sui documenti, che sono fortemente orientati verso l'identificazione di vicini distribuzionali appartenenti a domini tematici ad ampio spettro (Sahlgren, 2006).

Altri parametri dei modelli distribuzionali hanno ricevuto particolare attenzione: i pesi statistici dei contesti e le misure di similarità dei vettori. Un'ampia gamma di scelte esiste per entrambi (Curran, 2004; Bullinaria, J. P. Levy, 2007), ma oggi la pratica più comune consiste nell'usare la *Positive Pointwise Mutual Information* come peso statistico per misurare la salienza dei contesti e il coseno come misura di similarità semantica. Questi infatti sono in grado di offrire le migliori prestazioni in vari tipi di esperimenti semantici (Turney, Pantel, 2010).

I vettori nella matrice di co-occorrenza forniscono una rappresentazione esplicita della distribuzione nei contesti (O. Levy, Goldberg, 2014). Ciascuna dimensione del vettore, infatti, corrisponde a un contesto specifico nel quale la parola target è stata osservata. I vettori di concorrenza espliciti hanno molte dimensioni (tipicamente nell'ordine delle centinaia di migliaia o anche più) e sono sparsi. Vari tipi di tecniche sono perciò usate per ridurre le dimensioni dei vettori e limitare così la complessità computazionale. L'approccio più comune consiste nel proiettare l'originale matrice sparsa in una matrice densa a ridotta dimensionalità usando metodi come *Singular Value Decomposition* (Landauer, Dumais, 1997), *Non-Negative Matrix Factorization* (Van de Cruys, 2010), e *Latent Dirichlet Allocation* (Blei et al., 2003). Crucialmente, le dimensioni dei vettori ridotti non corrispondono più a contesti espliciti, ma piuttosto a dimensioni semantiche 'nascoste', ovvero implicite, nell'originale distribuzione dei dati. Le tecniche di riduzione di matrici permettono di rimuovere il rumore nei dati estratti dai corpora e di sfruttare le ridondanze e correlazioni tra i contesti linguistici, migliorando così la qualità degli spazi semantici (Turney, Pantel, 2010). Un'alternativa molto popolare ed efficace è il *Random Indexing* (Sahlgren, 2006): invece di ridurre una matrice costruita in precedenza, rappresentazioni vettoriali con un numero ridotto di dimensioni sono costruite incrementalmente assegnando a ciascuna parola un vettore casuale che viene poi sommato ai vettori delle parole che co-occorrono con essa.

Molte ricerche sono state dedicate a comprendere l'impatto di questi parametri sulle performance dei modelli distribuzionali. Gli studi più recenti e comprensivi sono quelli di Lapesa, Evert (2014) e Kiela, Clark (2014), che analizzano un'ampia gamma di parametri quali il tipo di corpus, l'uso di metodologie di lemmatizzazione o *stemming* del testo, la tipologia dei contesti (dipendenze sintattiche contro co-occorrenze, direzione e dimensione della finestra, ecc.), i pesi statistici dei contesti, le misure di similarità e le tecniche di riduzione delle dimensioni dei vettori. Questi esperimenti permettono di individuare le configurazioni migliori dei parametri dei modelli distribuzionali in vari compiti semantici.

2.2.2. Contare o predire

I modelli che abbiamo descritto sopra usano un approccio alla costruzione delle rappresentazioni distribuzionali basato sul conteggio delle frequenze delle parole nei testi: le co-occorrenze estratte da corpora sono contate, poi pesate e infine

opzionalmente ridotte per costruire dei vettori densi. Per tale motivo sono anche definiti *count models*. Recentemente è apparsa una nuova famiglia di modelli distribuzionali basata su una tecnica di predizione: algoritmi neurali creano direttamente rappresentazioni distribuzionali dense e a bassa dimensionalità, imparando a predire in maniera ottimale i contesti di una parola target (Mikolov et al., 2013). Per tale motivo sono anche definiti *prediction models* e le rappresentazioni costruite da essi *embedding*, perché le parole sono *embedded* ('incassate') dentro uno spazio lineare a bassa dimensionalità formato da *feature* latenti. Vari tipi di regolarità linguistiche sono stati identificati negli spazi semantici formati dagli *embedding* neurali: ad esempio, il fatto che *re* e *regina* hanno la stessa relazione di genere di *uomo* e *donna* viene rappresentato attraverso le relazioni tra i rispettivi vettori nello spazio. In tal modo, il vettore di una parola (es. **regina**) può essere ricostruito dalle rappresentazioni delle altre parole attraverso una semplice aritmetica vettoriale (es., **re** − **uomo** + **donna**). Alcuni esperimenti hanno anche mostrato che i *prediction model* sono in grado di superare i *count models* in vari *task* (Baroni, Dinu et al., 2014).

Nonostante la loro crescente popolarità, la questione se gli *embedding* neurali siano una reale innovazione rispetto ai modelli più tradizionali è ancora lontana dall'essere risolta. Per esempio, le stesse regolarità catturate dai *prediction model* sono anche catturate dai modelli basati su conteggi di frequenza espliciti (O. Levy, Goldberg, 2014). Quando i parametri di questi ultimi sono accuratamente raffinati, non vengono osservate differenze significative nelle prestazioni tra le due tipologie di modelli (O. Levy, Goldberg, Dagan, 2015). È possibile che la ricerca futura dimostrerà se gli *embedding* offrano dei chiari vantaggi, ma per ora i due approcci non differiscono in maniera sostanziale per gli aspetti del significato che sono in grado di catturare. Sono semplicemente modi alternativi di costruire rappresentazioni distribuzionali.

3. Valutare i modelli semantici distribuzionali

La dicotomia classica tra metodi intrinseci e metodi estrinseci di valutazione nel TAL si applica anche alla semantica distribuzionale. Le valutazioni intrinseche hanno come obiettivo quello di misurare la qualità della risorsa stessa, confrontandola con i giudizi di valutatori umani o con simili risorse semantiche che possono essere utilizzate come standard di riferimento (*gold standard*). Le valutazioni estrinseche misurano invece il contributo specifico della risorsa per migliorare le performance di un sistema in cui é inserita.

La valutazione intrinseca dei modelli semantici distribuzionali viene realizzata attraverso la comparazione con risorse lessicali come i sinonimi del TOEFL (Landauer, Dumais, 1997), *thesauri* specializzati (G. Grefenstette, 1994), Wordnet (Curran, Moens, 2002; Padró et al., 2014; Anguiano, Denis, 2011), dizionari di sinonimi (Van der Plas, Tiedemann, Manguin, 2011), ecc. La valutazione intrinseca dei modelli di semantica distribuzionale è una questione complessa per molteplici ragioni. Prima di tutto, i modelli distribuzionali catturano una nozione molto ampia di vicinanza semantica (cfr. *infra*, Sezione 4.). Esiste dunque uno slittamento

inevitabile tra i risultati prodotti dai modelli computazionali e le risorse che invece contengono relazioni lessicali di tipo classico, quali dizionari dei sinonimi, *thesauri* e Wordnet. Un secondo tipo di problema è dovuto al fatto che i modelli semantici distribuzionali riflettono le specificità del corpus da cui sono costruiti, e possono dunque identificare relazioni che mancano in risorse lessicali di tipo generale. È infatti difficile, forse impossibile, verificare la validità di una relazione semantica fuori contesto (Muller et al., 2014). Allo scopo di superare tali limiti, sono state sviluppate risorse specificatamente orientate verso la valutazione dei modelli semantici distribuzionali. Uno degli standard di riferimento più usati è WordSim-353 (Finkelstein et al., 2002), con 353 coppie di parole inglesi associate a un punteggio di similarità semantica espresso da valutatori umani. Recentemente è stata anche realizzata una versione multilingue di questo *dataset* (Leviant, Reichart, 2015).

Per quanto riguarda la valutazione estrinseca, l'uso di *feature* distribuzionali è utile ogni volta in cui è necessario misurare la similarità tra parole o tra porzioni di testo più ampie. Vari esperimenti sono stati dedicati all'uso delle risorse distribuzionali nell'*Information Retrieval* per misurare la similarità tra le *query* (Alfonseca et al., 2009; Claveau, Kijak, 2015). Nei compiti di sostituzione lessicale (McCarthy, Navigli, 2007), i modelli distribuzionali sono usati per identificare potenziali sostituti di una parola prima del processo di disambiguazione (Fabre et al., 2014). La similarità distribuzionale è anche utilizzata per determinare il senso di una parola in un corpus (McCarthy, Koeling et al., 2007). Modelli di semantica distribuzionale si sono dimostrati efficaci in applicazioni di TAL più complesse come il *Textual Entailment* e la generazione automatica di riassunti (Cheung, Penn, 2013). Inoltre, gli *embedding* neurali sono stati usati con successo per migliorare sistemi di *Semantic Role Labeling* e *Named Entity Recognition* (Collobert, Weston, 2008).

4. Le sfide per la semantica distribuzionale

I modelli semantici distribuzionali sono stati oggetto di molte critiche, anche all'interno della comunità della linguistica computazionale. L'approccio esclusivamente induttivo al significato della semantica distribuzionale è molto utile dal punto di vista del TAL, ma rimane una questione aperta se le statistiche di co-occorrenza da sole siano sufficienti per affrontare aspetti semantici più profondi, oppure siano soltanto in grado di fornire un'approssimazione molto superficiale del significato lessicale (Sahlgren, 2008; Lenci, 2008; Koller, 2015).

L'Ipotesi Distribuzionale è di fatto una definizione della similarità semantica in termini di prossimità in uno spazio. La similarità semantica è però essa stessa una nozione molto vaga, che va dalla similarità tra parole alla similarità tra relazioni (Turney, 2006; Baroni, Lenci, 2010; Turney, 2013). La similarità semantica in senso stretto, come relazione tra parole che condividono simili tratti semantici (es. *auto* e *camion*), deve inoltre essere distinta dall'associazione semantica tra parole, come *auto* e *ruota* (Budanitsky, Hirst, 2006; Agirre et al., 2009). Questi due tipi di relazioni semantiche hanno proprietà molto differenti e tuttavia sono discriminate a fatica dai modelli semantici distribuzionali. Anche *gold standard*

come WordSim-353 contengono in realtà molte coppie di parole associate che non sono semanticamente simili in senso stretto (Agirre et al., 2009). Per affrontare questo problema, è stato recentemente sviluppato il *dataset* SimLex-999 allo scopo di valutare specificatamente la capacità di modelli semantici distribuzionali di catturare la similarità semantica piuttosto che relazioni di associazione semantica (Hill et al., 2015).

Un problema aggiuntivo è dato dal fatto che sia la similarità semantica che l'associazione semantica sono in realtà termini che coprono tipi molto diversi di relazioni lessicali. Per esempio, i sinonimi, i co-iponimi e perfino gli antonimi possono essere detti semanticamente simili perché condividono un elevato numero di tratti. Le associazioni semantiche dall'altro lato includono la metonimia, relazioni locative, o la semplice appartenenza al medesimo campo semantico (Morris, Hirst, 2004). Questa nozione ampia e graduata di associazione è al tempo stesso utile e problematica per le applicazioni del TAL, perché è molto difficile tracciare un limite chiaro tra le relazioni associative rilevanti e quelle non rilevanti (Sahlgren, 2008; Ferret, 2013). In generale, i vicini distribuzionali identificati dei modelli semantici distribuzionali hanno relazioni molto diverse con la parola target, suggerendo che i modelli semantici distribuzionali forniscono in realtà una rappresentazione 'a grana grossa' del significato lessicale. Recentemente i *dataset* BLESS (Baroni, Lenci, 2011) e EVALution (Santus, Yung et al., 2015) sono stati progettati proprio per valutare l'abilità dei modelli distribuzionali di discriminare differenti tipi di relazioni lessicali. Questa rappresenta un'importante area di ricerca nella semantica distribuzionale (Van der Plas, Tiedemann, 2006; Lenci, Benotto, 2012; Santus, Lenci et al., 2014; The Pham et al., 2015).

Una questione fondamentale è determinare quale tipologia di informazione semantica può essere catturata sulla base delle proprietà contestuali e quali parti del significato delle parole rimangono invece al di là delle possibilità dei modelli distribuzionali, a meno di non integrare le co-occorrenze contestuali con altre informazioni. Vari lavori recenti si focalizzano su questo problema. Gupta et al. (2015) mostrano che alcuni aspetti dell'informazione referenziale sono accessibili dal punto di vista distribuzionale, mentre Herbelot, Ganesalingam (2013) suggeriscono che l'informatività dei lessemi (ovvero la distinzione tra parole con maggiore o minore peso informativo) è difficile da ricavare su base distribuzionale. Zarcone et al. (2015) mostrano che non solo la prototipicità semantica di un argomento rispetto al predicato (il cosiddetto *thematic fit*), ma anche i vincoli determinati dal tipo semantico richiesto dal predicato possono essere catturati da modelli distribuzionali per rappresentare fenomeni di *coercion* e metonimia logica. In una prospettiva simile, lavori recenti propongono di creare un ponte tra semantica formale e distribuzionale (Guevara, 2011; E. Grefenstette, 2013), in maniera da combinare le rappresentazioni distribuzionali dei significati lessicali con le capacità inferenziali dei sistemi formali (Erk, 2013; Boleda, Erk, 2015).

4.1. Oltre le parole

La maggior parte dei lavori sulla semantica distribuzionale è dedicata all'analisi delle parole in isolamento, ma negli ultimi anni la ricerca si è anche focalizzata

sull'estensione di questi modelli per rappresentare unità semantiche più ampie, come sintagmi e frasi, seguendo due approcci principali. Il primo consiste nel considerare i sintagmi in aggiunta alle parole come elementi target a cui viene associata una rappresentazione vettoriale unitaria (es. Baldwin et al., 2003). Tuttavia questo rimane un approccio largamente minoritario a causa della scarsità di dati per ricostruire la distribuzione di unità complesse. La seconda opzione consiste nel modellare la composizionalità semantica all'interno del paradigma distribuzionale, sulla base dell'assunzione che l'informazione semantica sui sintagmi può essere computata combinando informazioni sulle rappresentazioni vettoriali dei loro componenti. Mitchell, Lapata (2010) propongono un modello generale per la semantica distribuzionale composizionale e analizzano diverse funzioni di combinazione vettoriale. Più di recente, è stato anche proposto un *task* di valutazione di aspetti legati alla semantica composizionale nell'ambito delle campagne SemEval (Marelli et al., 2014). Diversi tipi di unità sintagmatiche sono state esaminate, quali le coppie aggettivo-nome (Baroni, Zamparelli, 2010), verbo-nome (Mitchell, Lapata, 2010), e le frasi. Un altro settore di ricerca estremamente promettente riguarda l'uso di *feature* extralinguistiche per integrare i dati distribuzionali con altri tipi di informazione multimodale (Bruni et al., 2014).

5. Conclusioni e prospettive

La semantica distribuzionale è un paradigma scientifico giovane, ma nonostante la sua breve storia è stata in grado di guadagnare un grande successo nella comunità del TAL, con interessi crescenti nella ricerca linguistica e nelle scienze cognitive. Come è stato illustrato nelle sezioni precedenti, la varietà di modelli semantici distribuzionali sta aumentando rapidamente. Inoltre si è ottenuta una comprensione molto più profonda degli effetti e dei ruoli dei vari tipi di parametri rilevanti per le rappresentazioni distribuzionali. Certamente la semantica distribuzionale deve affrontare ancora molte sfide. Sotto molti punti di vista, i modelli semantici distribuzionali tuttora forniscono una rappresentazione molto grezza del significato, e i loro limiti (così come le loro potenzialità) devono essere tuttora esplorate. Al tempo stesso, il numero dei compiti semantici che sono affrontati da questi modelli è costantemente in espansione, andando ben al di là delle originali applicazioni dell'Ipotesi Distribuzionale per l'identificazione dei sinonimi. Tutto ciò rende questo ambito di ricerca estremamente vitale e promettente per ottenere una maggiore comprensione del significato linguistico.

Bibliografia

Agirre E., E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, A. Soroa (2009). "A study on similarity and relatedness using distributional and WordNet-based approaches". *Proceedings of Human Language Technologies: the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 19–27.

Alfonseca E., K. Hall, S. Hartmann (2009). "Large-scale computation of distributional similarities for queries". *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 29–32.

Anguiano E. H., P. Denis (2011). "FreDist: Automatic construction of distributional thesauri for French". *18ème conférence sur le traitement automatique des langues naturelles (TALN 2011)*, pp. 119–124.

Baldwin T., C. Bannard, T. Tanaka, D. Widdows (2003). "An empirical model of multiword expression decomposability". *Proceedings of the ACL 2003 workshop on multiword expressions: analysis, acquisition and treatment*. Vol. 18, pp. 89–96.

Baroni M., G. Dinu, G. Kruszewski (2014). "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 238–247.

Baroni M., A. Lenci (2010). "Distributional memory: a general framework for corpus-based semantics". *Computational Linguistics*, 36.4, pp. 673–721.

Baroni M., A. Lenci (2011). "How we BLESSed distributional semantic evaluation". *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pp. 1–10.

Baroni M., R. Zamparelli (2010). "Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space". *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1183–1193.

Blei D. M., A. Y. Ng, M. I. Jordan (2003). "Latent Dirichlet Allocation". *Journal of machine learning research*, 3, pp. 993–1022.

Boleda G., K. Erk (2015). "Distributional semantic features as semantic primitives – or not". Intervento all'AAAI Spring Symposium on Knowledge Representation and Reasoning.

Bruni E., N.-K. Tran, M. Baroni (2014). "Multimodal distributional semantics". *Journal of artificial intelligence research (JAIR)*, 49, pp. 1–47.

Budanitsky A., G. Hirst (2006). "Evaluating Wordnet-based measures of lexical semantic relatedness". *Computational linguistics*, 32.1, pp. 13–47.

Bullinaria J., J. P. Levy (2007). "Extracting semantic representations from word co-occurrence statistics: A computational study". *Behavior research methods*, 39 (3), pp. 510–526.

Cheung J. C. K., G. Penn (2013). "Probabilistic domain modelling with contextualized distributional semantic vectors". *Proceedings of ACL 2013*, pp. 392–401.

Claveau V., E. Kijak (2015). "Thésaurus distributionnels pour la recherche d'information et vice-versa". *Actes de la 13ème Conférence en Recherche d'Information et Applications (CORIA 2015)*.

Collobert R., J. Weston (2008). "A unified architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning". *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–167.

Curran J. R., M. Moens (2002). "Improvements in automatic thesaurus extraction". *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*. Vol. 9, pp. 59–66.

Curran J. R. (2004). "From distributional to semantic similarity". Tesi di dott. University of Edinburgh.

Erk K. (2013). "Towards a semantics for distributional representations". *Proceedings, 10th International Conference on Computational Semantics (IWCS-2013)*.

Fabre C., N. Hathout, L.-M. Ho-Dac, F. Morlane-Hondère, P. Muller, F. Sajous, L. Tanguy, T. Van de Cruys (2014). "Présentation de l'atelier SemDis 2014: sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés". *Actes de la conférence Traitement Automatique du Langage Naturel (TALN 2014)*, pp. 196–205.

Ferret O. (2013). "Identifying bad semantic neighbors for improving distributional thesauri". *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pp. 561–571.

Finkelstein L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin (2002). "Placing Search in Context: The Concept Revisited". *ACM Transactions on Information Systems*, 20.1, pp. 116–131.

Grefenstette E. (2013). "Towards a formal distributional semantics: simulating logical calculi with tensors". *Proceedings of the Second Joint Conference on Lexical and Computational Semantics.* Pp. 1–10.

Grefenstette G. (1994). *Explorations in automatic thesaurus discovery*. Norwell: Kluwer Academic Publishers.

Guevara E. (2011). "Computing semantic compositionality in distributional semantics". *Proceedings of the Ninth International Conference on Computational Semantics*, pp. 135–144.

Gupta A., G. Boleda, M. Baroni, S. Padó (2015). "Mapping conceptual features to referential properties". Intervento alla Conference on Empirical Methods in Natural Language Processing (EMNLP 2015).

Harris Z. S. (1954). "Distributional structure". *Word*, 10.2-3, pp. 146–162.

Harris Z. S. (1991). *A theory of language and information: a mathematical approach*. Oxford: Clarendon Press.

Herbelot A., M. Ganesalingam (2013). "Measuring semantic content in distributional vectors". *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Vol. 2, pp. 440–445.

Hill F., R. Reichart, A. Korhonen (2015). "SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation". *Computational Linguistics*, 41 (4), pp. 1–32.

Kiela D., S. Clark (2014). "A systematic study of semantic vector space model parameters". *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL*, pp. 21–30.

Koller A. (2015). "Top-down questions for distributional semantics". Intervento al Workshop on formal and distributional semantics.

Landauer T. K., S. T. Dumais (1997). "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge". *Psychological review*, 104.2, p. 211.

Lapesa G., S. Evert (2014). "A large scale evaluation of distributional semantic models: Parameters, interactions and model selection". *Transactions of the Association for Computational Linguistics*, 2, pp. 531–545.

Lenci A. (2008). "Distributional semantics in linguistic and cognitive research". *Italian journal of linguistics*, 20.1: *From context to meaning: distributional models of the lexicon in linguistics and cognitive science*, pp. 1–31.

Lenci A., G. Benotto (2012). "Identifying hypernyms in distributional semantic spaces". **SEM 2012: The First Joint Conference on Lexical and Computational Semantics (SemEval 2012)*, pp. 75–79.

Leviant I., R. Reichart (2015). "Judgment language matters: multilingual vector space models for judgment language aware lexical semantics". *ArXiv e-prints*. arXiv: 1508.00106 [cs.CL].

Levy O., Y. Goldberg (2014). "Linguistic regularities in sparse and explicit word representations". *Proceedings of the Eighteenth Conference on Computational Language Learning (CoNLL)*, pp. 171–180.

Levy O., Y. Goldberg, I. Dagan (2015). "Improving distributional similarity with lessons learned from word embeddings". *Transactions of the ACL*, 3, pp. 211–225.

Lund C., K. Burgess (1997). "Modelling parsing constraints with high-dimensional context space". *Language and cognitive processes*, 12.2-3, pp. 177–210.

Marelli M., L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, R. Zamparelli (2014). "Semeval-2014 task 1: evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment". *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 1–8.

McCarthy D., R. Koeling, J. Weeds, J. Carroll (2007). "Unsupervised acquisition of predominant word senses". *Computational Linguistics*, 33.4, pp. 553–590.

McCarthy D., R. Navigli (2007). "Semeval-2007 task 10: English lexical substitution task". *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 48–53.

Mikolov T., K. Chen, G. Corrado, J. Dean (2013). "Efficient estimation of word representations in vector space". *Proceedings of Workshop at ICLR 2013*, pp. 1–12.

Mitchell J., M. Lapata (2010). "Composition in distributional models of semantics". *Cognitive Science*, 34.8, pp. 1388–1439.

Morris J., G. Hirst (2004). "Non-classical lexical semantic relations". *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pp. 46–51.

Muller P., C. Fabre, C. Adam (2014). "Predicting the relevance of distributional semantic similarity with contextual information". *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pp. 479–488.

Nazarenko A., P. Zweigenbaum, B. Habert, J. Bouaud (2001). "Corpus-based extension of a terminological semantic lexicon". *Recent Advances in Computational Terminology*, pp. 327–351.

Padó S., M. Lapata (2007). "Dependency-based construction of semantic space models". *Computational Linguistics*, 33.2, pp. 161–199.

Padró M., M. Idiart, C. Ramisch, A. Villavicencio (2014). "Nothing like good old frequency: studying context filters for distributional thesauri". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 419–424.

Peirsman Y., D. Geeraerts (2009). "Predicting strong associations on the basis of corpus data". *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 648–656.

Peirsman Y., K. Heylen, D. Speelman (2007). "Finding semantically related words in Dutch. Cooccurrences versus syntactic contexts". *Proceedings of the CoSMO workshop at CONTEXT-07*, pp. 9–16.

Sadrzadeh M., E. Grefenstette (2011). "A compositional distributional semantics, two concrete constructions, and some experimental evaluations", *Quantum Interaction*, pp. 35–47.

Sahlgren M. (2006). "The word-space model". Tesi di dott. University of Stockholm.

Sahlgren M. (2008). "The distributional hypothesis". *Italian Journal of Linguistics*, 20.1, pp. 33–54.

Salton G., A. Wong, C.-S. Yang (1975). "A vector space model for automatic indexing". *Communications of the ACM*, 18.11, pp. 613–620.

Santus E., A. Lenci, Q. Lu, S. Schulte im Walde (2014). "Chasing hypernyms in vector spaces with entropy". *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 2, pp. 38–42.

Santus E., F. Yung, A. Lenci, C.-R. Huang (2015). "EVALution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models". *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pp. 64–69.

The Pham N., A. Lazaridou, M. Baroni (2015). "A multitask objective to inject lexical contrast into distributional semantics". Intervento al 53nd Annual Meeting of the Association for Computational Linguistics (ACL 2015).

Turney P. D. (2006). "Similarity of semantic relations". *Computational Linguistics*, 32.3, pp. 379–416.

Turney P. D. (2013). "Distributional semantics beyond words: supervised learning of analogy and paraphrase". *Transactions of the Association for Computational Linguistics (TACL)*, pp. 353–366.

Turney P. D., P. Pantel et al. (2010). "From frequency to meaning: vector space models of semantics". *Journal of artificial intelligence research*, 37.1, pp. 141–188.

Van de Cruys T. (2008). "A comparison of bag of words and syntax-based approaches for word categorization". *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pp. 47–54.

Van de Cruys T. (2010). "A non-negative tensor factorization model for selectional preference induction". *Natural Language Engineering*, 16.04, pp. 417–437.

Van der Plas L., J. Tiedemann (2006). "Finding synonyms using automatic word alignment and measures of distributional similarity". *Proceedings of the COLING/ACL, Main conference poster sessions*, pp. 866–873.

Van der Plas L., J. Tiedemann, J.-L. Manguin (2011). "Synonym acquisition across domains and languages", *Advances in distributed agent-based retrieval tools*. Berlin: Springer, pp. 41–57.

Zarcone A., S. Padó, A. Lenci (2015). "Same same but different: type and typicality in a distributional model of complement coercion". *NetWordS 2015 word knowledge and word usage*, pp. 91–94.