Event Knowledge in Compositional Distributional Semantics

Ludovica Pannitto
Corso di Laurea Magistrale in Informatica Umanistica, Università di Pisa

Oct 11th 2018

Rappresentazioni Semantiche

La competenza linguistica permette ai parlanti di capire e produrre un numero illimitato di espressioni linguistiche nuove e complesse.

La comprensione di tali espressioni avviene grazie alla costruzione di rappresentazioni semantiche, necessarie per **supportare il ragionamento umano** su eventi e situazioni che vengono descritti tramite il linguaggio.

(1) Dopo l'atterraggio, **il pilota ha spento il motore**.

(2) Dopo il rally, **il pilota ha spento il motore**.

Outline

Ci aspettiamo che le risorse computazionali riescano a modellare fenomeni di questo tipo, che costituiscono il nucleo dell'utilizzo del linguaggio da parte dei parlanti.

L'obiettivo di questo lavoro è di impegare **metodi usage-based** in un modello di **composizionalità** che sia **linguisticamente motivato e ispirato a modelli cognitivi**.

Contents

Background

Significato e contesto

Il fatto che il significato delle parole possa essere inferito dall'uso è qualcosa di intuitivamente vero[1].

(3) Le porse un bicchiere di *bardiwac*.

(4) Nigel barcollava, con il viso arrossato per il troppo *bardiwac*.

(5) Malbec, uno dei vitigni meno conosciuti di *bardiwac*, risponde bene al sole australiano.

(6) Ho cenato con pane e formaggio, e questo eccellente *bardiwac*.

(7) I drink erano delizioni: *bardiwac* rosso sangue e un Rhenish dolce e leggero.

[1] Esempi di Stefan Evert, dal British National Corpus, tradotti per questa presentazione

Ipotesi Distribuzionale

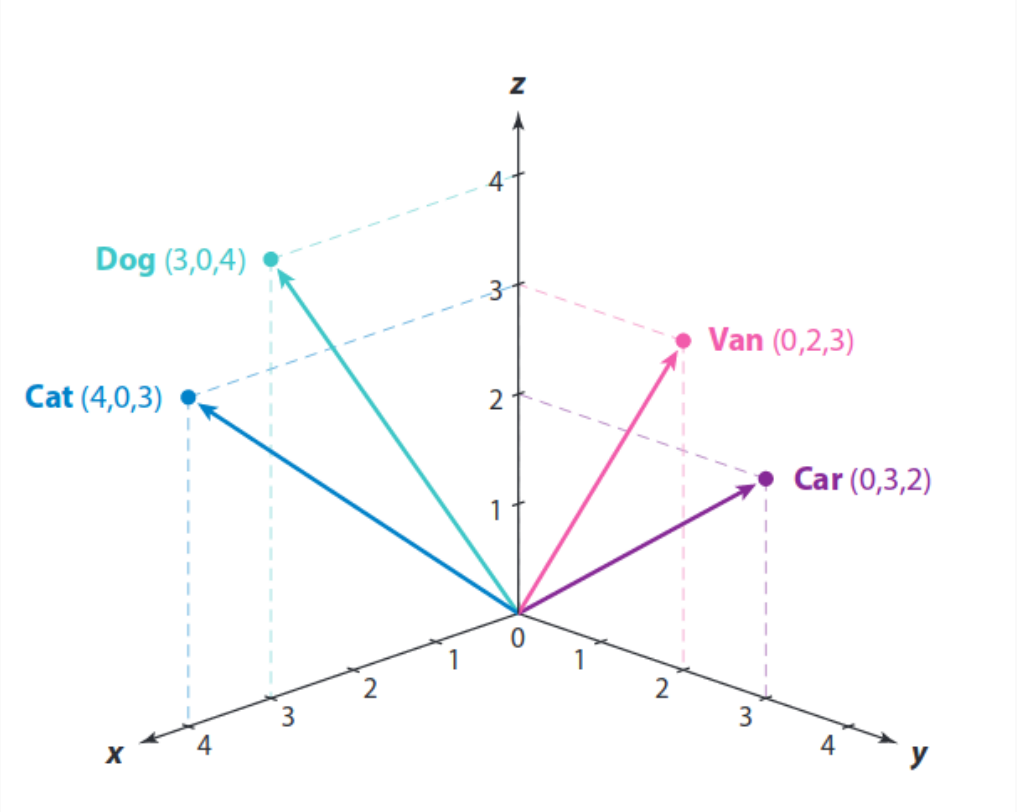
"lexemes with similar linguistic contexts have similar meaning" (Lenci 2008).

Il **distribuzionalismo** è una teoria generale del significato che prende ispirazione dalla filosofia del linguaggio (Wittgenstein 1953), ma viene sviluppata in campi quali la psicologia comportamentista (Deese 1966) e le scienze cognitive (Rubenstein e Goodenough 1965; Miller e Charles 1991).

"what people know when they know a word is not how to recite its dictionary definition – they know how to use it (when to produce it and how to understand it) in everyday discourse" (Miller e Charles 1991)

Rappresentazioni distribuzionali

Mentre le rappresentazioni della semantica formale sono discrete e categoriche, le rappresentazioni distribuzionali sono tipicamente **scalari** e **distribuite**.



Composizionalità fregeana

La tradizionale distinzione tra sintassi e semantica ha portato alla formulazione di una teoria della composizionalità semantica **sintatticamente trasparente**.

- tutti gli elementi che concorrono al significato della frase si trovano nelle rappresentazioni semantiche delle parole che la compongono
- il modo in cui le rappresentazioni sono combinate dipende esclusivamente dalla struttura sintattica della frase, né la struttura interna delle rappresentazioni né la pragmatica prendono parte alla composizione

Composizionalità fregeana

La tradizionale distinzione tra sintassi e semantica ha portato alla formulazione di una teoria della composizionalità semantica **sintatticamente trasparente**.

- tutti gli elementi che concorrono al significato della frase si trovano nelle rappresentazioni semantiche delle parole che la compongono
- il modo in cui le rappresentazioni sono combinate dipende esclusivamente dalla struttura sintattica della frase, né la struttura interna delle rappresentazioni né la pragmatica prendono parte alla composizione

$$\vec{p} = f(\vec{u}, \vec{v}, R, K) \tag{1}$$

Modello additivo

(Molto) in breve:

$$\vec{p} = \vec{u} + \vec{v} \tag{2}$$

Potenziali problemi:

- il significato di un'espressione complessa non è facilmente approssimato dal significato delle sue parti (*mangiare la mela*, *mangiare la polvere*)
- l'ordine delle parole non è preso in considerazione (*cane morde uomo*, *uomo morde cane*)
- questioni tecniche: vettori che vivono in diversi sottospazi finiscono per essere concatenati

Accettabile vs. Plausibile

Il problema della composizionalità è stato affrontato distinguendo tra frasi **possibili** e **impossibili**[2]:

(8) Il musicista suona il flauto a teatro.

(9) * Il nominativo suona la mappa nella tazza.

[2] Esempi da Chersoni 2018, tradotti per questa presentazione.

Accettabile vs. Plausibile

Il problema della composizionalità è stato affrontato distinguendo tra frasi **possibili** e **impossibili**[2]:

(11) Il musicista suona il flauto a teatro.

(12) * Il nominativo suona la mappa nella tazza.

La prima classe include frasi **tipiche** e **atipiche**, il cui status ha destato molto interesse recentemente.

(13) Il giardiniere suona le nacchere nella caverna.

[2] Esempi da Chersoni 2018, tradotti per questa presentazione.

Composizione in due fasi

La tradizionale classificazione binaria è stata a lungo supportata da un modello di composizionalità che prevede **due fasi** per l'interpretazione del linguaggio:

- calcolo del significato della frase in modo indipendente dal contesto
- integrazione del significato nel discorso o con informazione di tipo pragmatico

Questo modello è stato smentito dalle evidenze empiriche, che attribuiscono a frasi impossibili (violazioni semantiche) e a frasi implausibili (violazioni pragmatiche) lo stesso status cognitivo.

Lessico e Generalized Event Knowledge

*"in terms of the possible entities that participate in such events, knowing that a **waitress** is involved, for example,invokes a certain type of eating event."*
*"Instrument nouns can cue certain types of eating, as in eating with a **fork** versus eating with a **stick**." (McRae e Matsuki 2009)*

Lessico e Generalized Event Knowledge

> *"in terms of the possible entities that participate in such events, knowing that a **waitress** is involved, for example,invokes a certain type of eating event."*
> *"Instrument nouns can cue certain types of eating, as in eating with a **fork** versus eating with a **stick**." (McRae e Matsuki 2009)*

Il lessico mentale è organizzato come una **rete di mutue aspettative**, che influenzano la comprensione fornendo indizi linguistici per l'attivazione di conoscenza sugli eventi.
La comprensione di una frase è l'**identificazione o creazione** dell'evento che spiega meglio l'input linguistico.

Modello

Il nostro modello [3] presenta due componenti:

- un **Grafo Distribuzionale degli Eventi** (DEG) che modella un frammento della memoria semantica, ed è attivato dalle unità lessicali
- una **Funzione di Composizione del Significato**, che integra dinamicamente l'informazione attivata da DEG per costruire la rappresentazione della frase

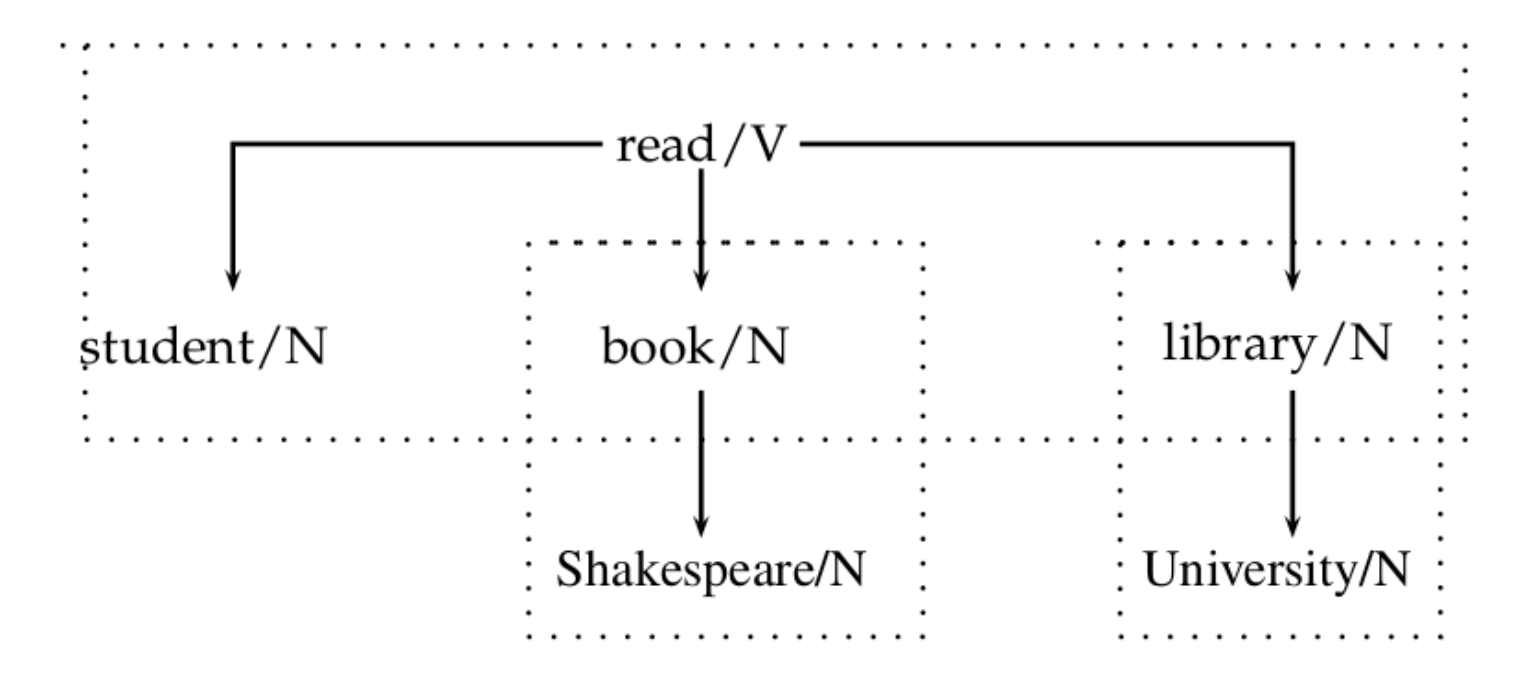
[3] seguendo l'approccio presentato in Chersoni, Blache e Lenci 2016; Chersoni, Lenci e Blache 2017; Chersoni et al. 2017

DEG

Con **evento** ci riferiamo a qualsiasi **relazione tra entità che costituiscono un evento, stato o situazione**.

Ci aspettiamo che la struttura dati tenga traccia di ogni evento (estratto automaticamente da corpora), e che gli eventi possano essere attivati da tutti i loro potenziali partecipanti, con un peso che dipende dall'associazione distribuzionale tra l'evento e il partecipante.

DEG: costruzione



Per ogni **sottoinsieme** di ogni **gruppo** estratto dalla frase, una relazione viene aggiunta al grafo.

DEG: proprietà

La struttura che ne risulta è:

- un **ipergrafo pesato** (*student read book*)
- un **multigrafo etichettato** (*student read book* vs. *book read student*)

DEG: proprietà

La struttura che ne risulta è:

- un **ipergrafo pesato** (*student read book*)
- un **multigrafo etichettato** (*student read book* vs. *book read student*)

Poiché i nodi (lessemi) del grafo sono rappresentati da **vettori distribuzionali**, DEG può essere interrogato su due livelli:

- come un **modello distribuzionale classico** (*book* $\rightarrow$ *essay*, *anthology*, *novel*, *author*, *publish*, *biography*, *autobiography*, *nonfiction*, *story*, *novella*)

DEG: proprietà

La struttura che ne risulta è:

- un **ipergrafo pesato** (*student read book*)
- un **multigrafo etichettato** (*student read book* vs. *book read student*)

Poiché i nodi (lessemi) del grafo sono rappresentati da **vettori distribuzionali**, DEG può essere interrogato su due livelli:

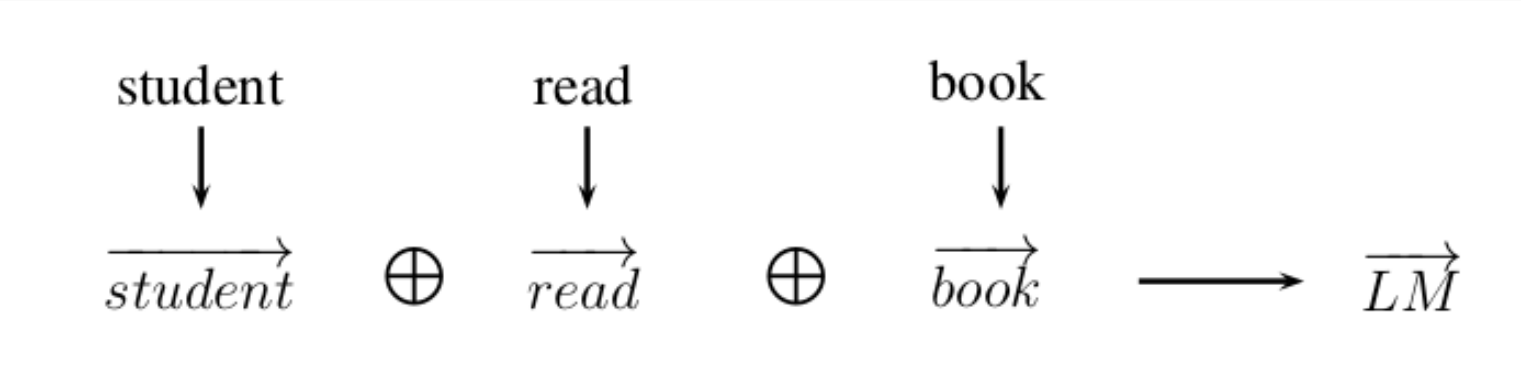
- come un **modello distribuzionale classico** (*book* $\rightarrow$ *essay*, *anthology*, *novel*, *author*, *publish*, *biography*, *autobiography*, *nonfiction*, *story*, *novella*)
- per ottenere i **vicini sintagmatici** di un lessema (*book* $\rightarrow$ *publish*, *write*, *read*, *include*, *child*, *series*, *have*, *buy*, *author*, *contains*)

Funzione di composizione

La comprensione di una frase è stata modellata come la creazione di una rappresentazione semantica SR, che condenti due diversi livelli informativi:

- una componente di **significato lessicale** (LM), che corrisponde alla rappresentazione *contex-independent* della frase e simula i modelli tradizionali
- l'**Active Context** (AC), il cui obiettivo è quello di rappresentare l'evento in termini dei suoi partecipanti, che possono essere ricostruiti da DEG a partire dagli item linguistici presenti nell'input

Lexical Meaning



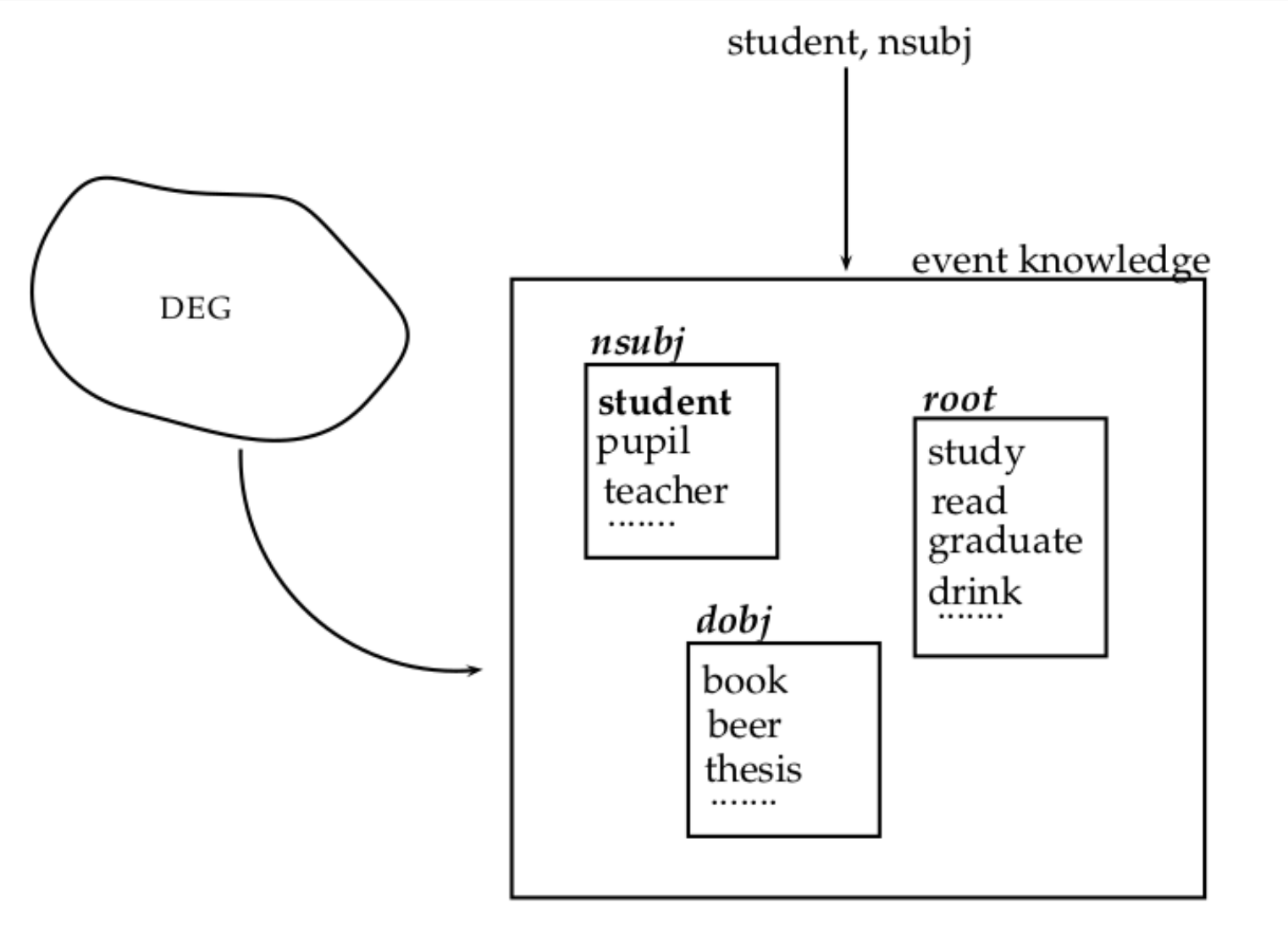
Active Context

L'AC rappresenta l'insieme di *aspettative sugli eventi linguistici* nella frase.

Implementa tre operazioni:

- Inizio di un nuovo processo (`inizialize`);
- **Elaborazione** di un nuovo input (`retrieve`);
- **Integrazione** della nuova informazione con i dati esistenti (`merge`).

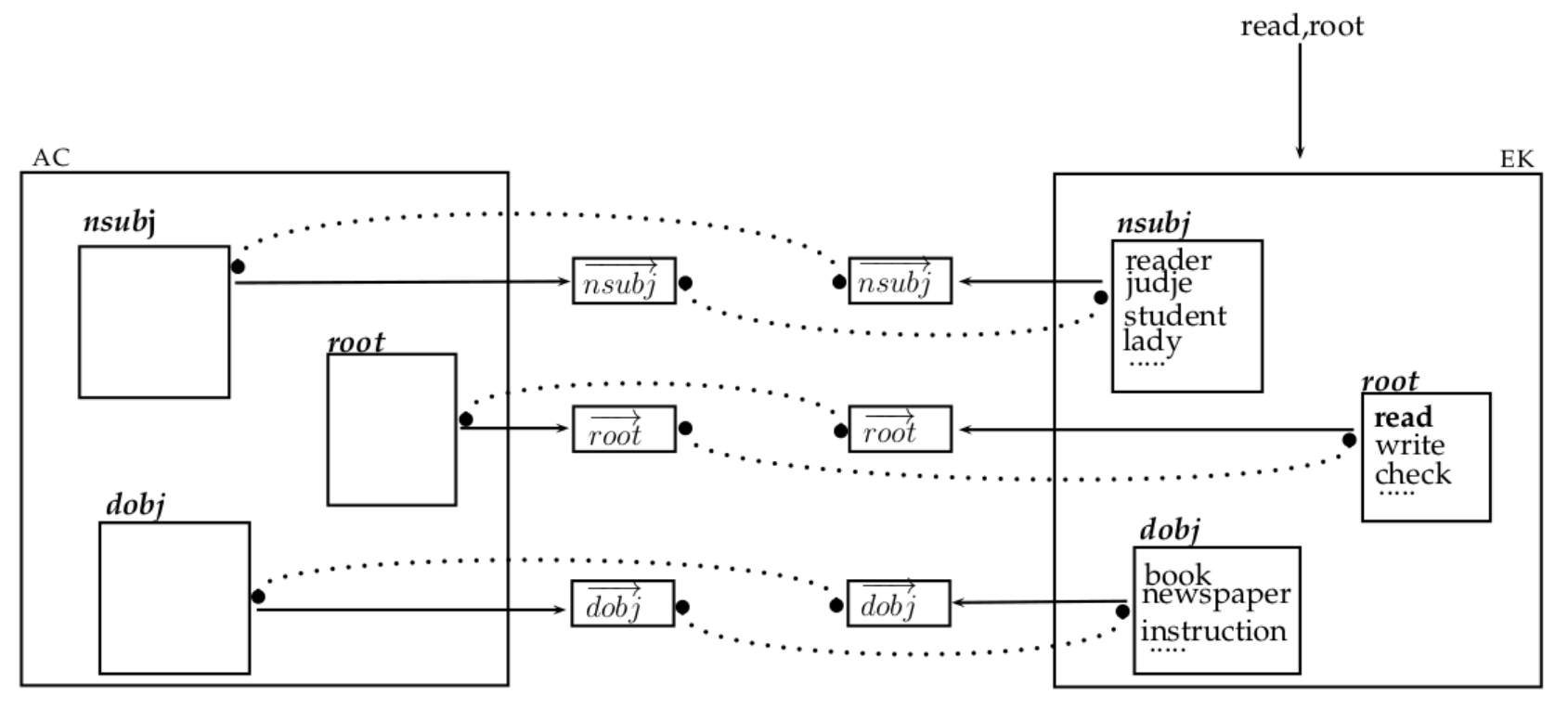
Active Context: `retrieve`

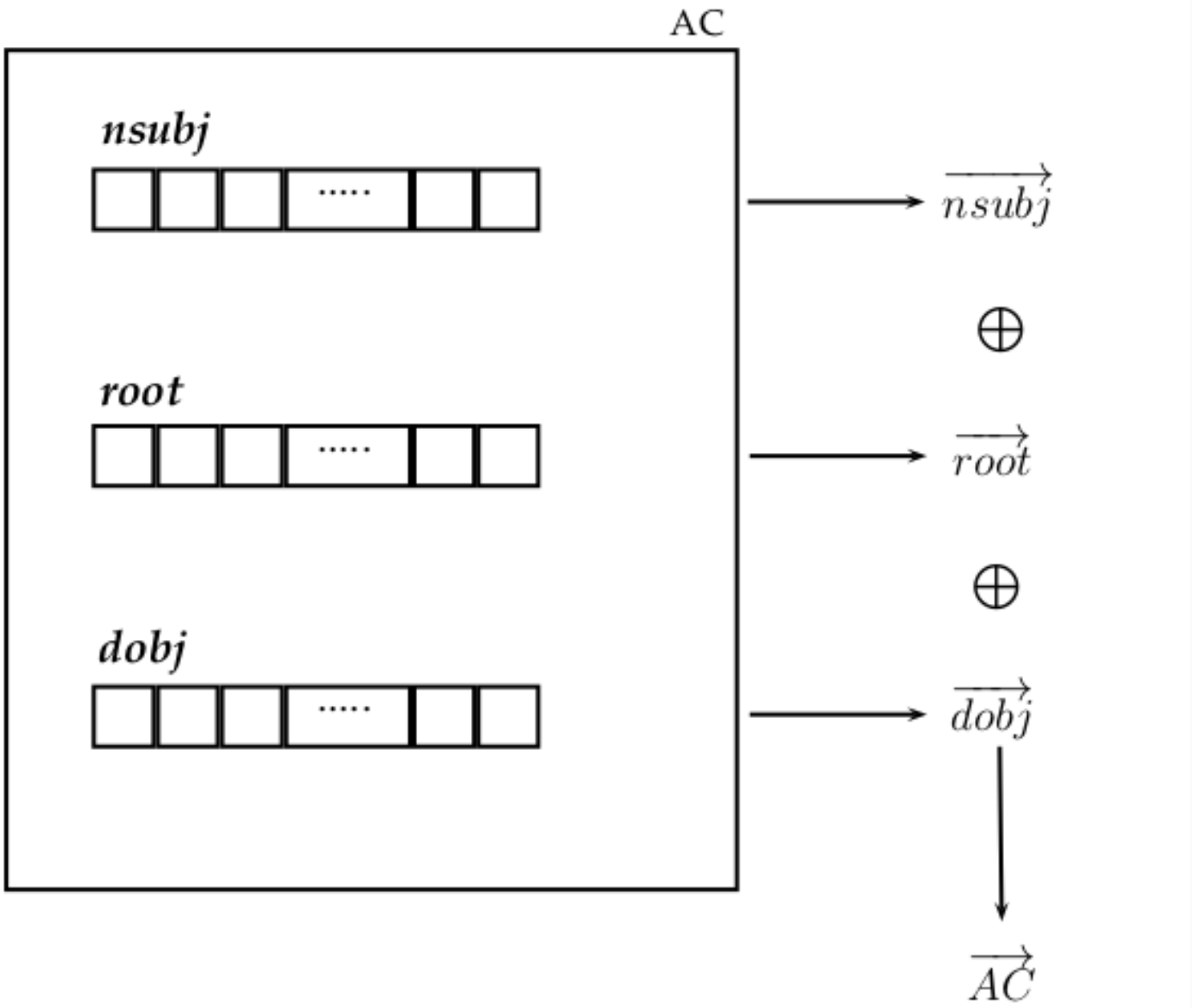


Active Context: `merge`

L'obiettivo del processo di integrazione è quello di **massimizzare la coerenza semantica** dell'evento.

L'interazione con l'informazione pre-esistente avviene tramite un **processo bidirezionale**, che permette ai partecipanti più salienti di fluttuare verso le posizioni più prominenti di AC, mentre quelli meno adatti vengono via via scartati.


read,root
AC
EK
nsubj
root
dobj
$\overrightarrow{nsubj}$
$\overrightarrow{root}$
$\overrightarrow{dobj}$
$\overrightarrow{nsubj}$
$\overrightarrow{root}$
$\overrightarrow{dobj}$
nsubj
reader
judje
student
lady
.....
root
read
write
check
.....
dobj
book
newspaper
instruction
.....


AC
nsubj
root
dobj
$\overrightarrow{nsubj}$
$\overrightarrow{root}$
$\overrightarrow{dobj}$
$\overrightarrow{AC}$

Valutazione

L'obiettivo degli esperimenti era valutare il contributo dell'**event knowledge** attivata in un task di comprensione semantica

Abbiamo implementato una versione ridotta del framework e testato il modello su due dataset (Rimell et al. 2016; Kartsaklis e Sadrzadeh 2014 che avessero:

- livello intermedio di complessità grammaticale
- lunghezza fissa e simili costruzioni sintattiche

RELPRON

RELPRON (Rimell et al. 2016) è composto da 1087 coppie (divise in development + test set) formate da un nome **target** etichettato con una **relazione sintattica** (*soggetto* o *oggetto diretto*) e una **proprietà** espressa da una frase relativa (*head noun*, *verbo*, *argomento*):

RELPRON

RELPRON (Rimell et al. 2016) è composto da 1087 coppie (divise in development + test set) formate da un nome **target** etichettato con una **relazione sintattica** (*soggetto* o *oggetto diretto*) e una **proprietà** espressa da una frase relativa (*head noun*, *verbo*, *argomento*):

OBJ	treaty/N: document/N that country/N sign/V
OBJ	treaty/N: document/N that government/N violate/V
SBJ	treaty/N: document/N that end/V war/N
SBJ	treaty/N: document/N that grant/V independence/N

Task

La valutazione è stata condotta su un task di **ranking**: il sistema ideale dovrebbe ordinare, per ogni termine, tutte le proprietà presenti nel dataset in modo che quelle corrispondenti al termine siano in testa alla lista.

I risultati sono espressi in termini di *Mean Average Precision* (MAP).

$$MAP = \frac{1}{N} \sum_{i=1}^{N} AP(t_i) \tag{3}$$

dove $AP(t_i)$ è l'*Average Precision* per il termine t_i:

$$AP(t) = \frac{1}{P_t} \sum_{k=1}^{M} Prec(k) \times rel(k) \tag{4}$$

Rappresentazione semantica

Ogni proprietà di RELPRON è stata rappresentata come una tripla

$$((hn, r), (w_1, r_1), (w_2, r_2)) \tag{5}$$

Rappresentazione semantica

Ogni proprietà di RELPRON è stata rappresentata come una tripla

$$((hn, r), (w_1, r_1), (w_2, r_2)) \tag{5}$$

Per ogni modello, è stata costruita una rappresentazione $\text{SR} = (\text{LM}, \text{AC})$, dove:

- la componente AC è considerata vuota per i modelli che non prendono in considerazione event knowledge (*baselines*)
- la componente LM è considerata vuota per i modelli basati solo su event knowledge

Rappresentazione semantica

Ogni proprietà di RELPRON è stata rappresentata come una tripla

$$((hn, r), (w_1, r_1), (w_2, r_2)) \tag{5}$$

Per ogni modello, è stata costruita una rappresentazione $\text{SR} = (\text{LM}, \text{AC})$, dove:

- la componente AC è considerata vuota per i modelli che non prendono in considerazione event knowledge (*baselines*)
- la componente LM è considerata vuota per i modelli basati solo su event knowledge

Consideriamo 6 modelli in totale, creati a partire da diversi sottoinsiemi degli elementi lessicali della proprietà:

(hn, r), (verb, s),(arg, t)

Rappresentazione semantica

Ogni proprietà di RELPRON è stata rappresentata come una tripla

$$((hn, r), (w_1, r_1), (w_2, r_2)) \tag{5}$$

Per ogni modello, è stata costruita una rappresentazione $\text{SR} = (\text{LM}, \text{AC})$, dove:

- la componente AC è considerata vuota per i modelli che non prendono in considerazione event knowledge (*baselines*)
- la componente LM è considerata vuota per i modelli basati solo su event knowledge

Consideriamo 6 modelli in totale, creati a partire da diversi sottoinsiemi degli elementi lessicali della proprietà:

(hn, r), (verb, s),(arg, t)

Rappresentazione semantica

Ogni proprietà di RELPRON è stata rappresentata come una tripla

$$((hn, r), (w_1, r_1), (w_2, r_2)) \tag{5}$$

Per ogni modello, è stata costruita una rappresentazione $\text{SR} = (\text{LM}, \text{AC})$, dove:

- la componente AC è considerata vuota per i modelli che non prendono in considerazione event knowledge (*baselines*)
- la componente LM è considerata vuota per i modelli basati solo su event knowledge

Consideriamo 6 modelli in totale, creati a partire da diversi sottoinsiemi degli elementi lessicali della proprietà:

(hn, r), (verb, s),(arg, t)

Rappresentazione semantica

Ogni proprietà di RELPRON è stata rappresentata come una tripla

$$((hn, r), (w_1, r_1), (w_2, r_2)) \tag{5}$$

Per ogni modello, è stata costruita una rappresentazione $\text{SR} = (\text{LM}, \text{AC})$, dove:

- la componente AC è considerata vuota per i modelli che non prendono in considerazione event knowledge (*baselines*)
- la componente LM è considerata vuota per i modelli basati solo su event knowledge

Consideriamo 6 modelli in totale, creati a partire da diversi sottoinsiemi degli elementi lessicali della proprietà:

(hn, r), (verb, s),(arg, t)

Rappresentazione semantica

Ogni proprietà di RELPRON è stata rappresentata come una tripla

$$((hn, r), (w_1, r_1), (w_2, r_2)) \tag{5}$$

Per ogni modello, è stata costruita una rappresentazione $\text{SR} = (\text{LM}, \text{AC})$, dove:

- la componente AC è considerata vuota per i modelli che non prendono in considerazione event knowledge (*baselines*)
- la componente LM è considerata vuota per i modelli basati solo su event knowledge

Consideriamo 6 modelli in totale, creati a partire da diversi sottoinsiemi degli elementi lessicali della proprietà:

(hn, r), (verb, s),(arg, t)

Rappresentazione semantica

Ogni proprietà di RELPRON è stata rappresentata come una tripla

$$((hn, r), (w_1, r_1), (w_2, r_2)) \tag{5}$$

Per ogni modello, è stata costruita una rappresentazione $\text{SR} = (\text{LM}, \text{AC})$, dove:

- la componente AC è considerata vuota per i modelli che non prendono in considerazione event knowledge (*baselines*)
- la componente LM è considerata vuota per i modelli basati solo su event knowledge

Consideriamo 6 modelli in totale, creati a partire da diversi sottoinsiemi degli elementi lessicali della proprietà:

(hn, r), (verb, s),(arg, t)

Rappresentazione semantica: LM

La componente del significato lessicale (LM) di ogni SR è stata costruita tramite **somma di vettori** delle parole presenti nella proprietà.

Rappresentazione semantica: AC

Per ogni proprietà:

- un AC è inizializzato vuoto
- le componenti sono processate nell'ordine in cui si trovano nella frase originale
- quando una coppia (w, r) viene presa in considerazione, i 50 vicini sintagmatici più salienti vengono estratti da DEG, e di questi i 20 elementi in cima alla lista prendono parte al centroide pesato

Esempio: *inventory, document that store maintains*

- *document* viene incontrato per primo, e il suo vettore è attivato come event knowledge per il ruolo di *object* della frase ($cos(\overrightarrow{inventory}, \text{AC}) = 0.530$)

Esempio: *inventory, document that store maintains*

- *document* viene incontrato per primo, e il suo vettore è attivato come event knowledge per il ruolo di *object* della frase ($cos(\overrightarrow{inventory}, \text{AC}) = 0.530$)
- *store* viene incontrato con il ruolo di *soggetto*, e le sue aspettative per il ruolo di *oggetto* vengono recuperate da DEG. La lista include *product*, *range*, *item*, *technology*, etc. ($cos(\overrightarrow{inventory}, \text{EK}(\text{store})) = 0.62$)

Esempio: *inventory, document that store maintains*

- *document* viene incontrato per primo, e il suo vettore è attivato come event knowledge per il ruolo di *object* della frase ($cos(\overrightarrow{inventory}, \text{AC}) = 0.530$)
- *store* viene incontrato con il ruolo di *soggetto*, e le sue aspettative per il ruolo di *oggetto* vengono recuperate da DEG. La lista include *product*, *range*, *item*, *technology*, etc. ($cos(\overrightarrow{inventory}, \text{EK}(\text{store})) = 0.62$)
- gli *s-neighbours* di *store* vengono ripesati rispetto all'AC, che contiene *document*. Gli elementi più simili fluttuano alla testa della lista, dove ora troviamo *collection*, *copy*, *book*, *item*, *name*, *trading*, *location*, etc., ($cos(\overrightarrow{inventory}, \text{AC}) = 0.68$);

Esempio: *inventory, document that store maintains*

- *document* viene incontrato per primo, e il suo vettore è attivato come event knowledge per il ruolo di *object* della frase ($cos(\overrightarrow{inventory}, \text{AC}) = 0.530$)
- *store* viene incontrato con il ruolo di *soggetto*, e le sue aspettative per il ruolo di *oggetto* vengono recuperate da DEG. La lista include *product*, *range*, *item*, *technology*, etc. ($cos(\overrightarrow{inventory}, \text{EK}(\text{store})) = 0.62$)
- gli *s-neighbours* di *store* vengono ripesati rispetto all'AC, che contiene *document*. Gli elementi più simili fluttuano alla testa della lista, dove ora troviamo *collection*, *copy*, *book*, *item*, *name*, *trading*, *location*, etc., ($cos(\overrightarrow{inventory}, \text{AC}) = 0.68$);
- ...

Sistema di punteggio

$$s = cos(\overrightarrow{target}, \overrightarrow{LM}) + cos(\overrightarrow{target}, \overrightarrow{AC}) \tag{6}$$

Sistema di punteggio

$$s = cos(\overrightarrow{target}, \overrightarrow{LM}) + cos(\overrightarrow{target}, \overrightarrow{AC}) \tag{6}$$

In generale, date due rappresentazioni semantiche $SR_1 = (\overrightarrow{LM_1}, \overrightarrow{AC_1})$ e $SR_2 = (\overrightarrow{LM_2}, \overrightarrow{AC_2})$, il punteggio finale viene calcolato secondo la formula:

$$s = sim(\overrightarrow{LM_1}, \overrightarrow{LM_2}) + sim(\overrightarrow{AC_1}, \overrightarrow{AC_2}) \tag{7}$$

dove $sim(\vec{x}, \vec{y})$ è una misura di similarità distribuzionale (es. *coseno*).

Risultati

	RELPRON		
	LM	AC	LM+AC
verb	0,18	0,18	**0,20**
arg	0,34	0,34	**0,36**
hn+verb	0,27	0,28	**0,29**
hn+arg	0,47	0,45	**0,49**
verb+arg	**0,42**	0,28	0,39
hn+verb+arg	0,51	0,47	**0,55**

Conclusioni

Wrap-up

Abbiamo fornito un'implementazione di un modello di significato composizionale, che mira ad essere **incrementale** e **cognitivamente plausibile**.

Anche se la **somma di vettori** continua a svolgere un ruolo importante nel modello, i nostri risultati suggeriscono che:

- i vettori distribuzionali non codificano, da soli, sufficiente informazione su **event knowledge**
- l'event knowledge gioca un ruolo significativo nella costruzione delle rappresentazioni semantiche **durante** il processing di espressioni linguistiche

Futuri sviluppi

- molti dei parametri del framework non sono stati esplorati (ruolo dei vicini paradigmatici, diverse funzioni di peso e aggregazione...)
- DEG è stato costruito in versione *ridotta*, non sono stati utilizzati contesti *congiunti* ma ogni relazione è stata trattata indipendentemente
- nel grafo completo gli stessi eventi, che sono codificati come *"supernodi"*, potrebbero essere rappresentati distribuzionalmente rispetto al loro vicinato sul grafo
- il framework è stato valutato solo su dataset piccoli e controllagti: una valutazione con dati più vari e complessi è necessaria per valutare l'effettivo miglioramento
- un ambito interessante riguarda l'integrazione di informazione non linguistica nel modello, e in generale l'estensione di questo metodo a modelli distribuzionali di dati percettivi in senso più generale

Grazie :)

References

Chersoni, Emmanuele (2018). «Explaining complexity in Human Language Processing: a Distributional Semantic Model». Tesi di dott.

Chersoni, Emmanuele, Philippe Blache e Alessandro Lenci (2016). «Towards a Distributional Model of Semantic Complexity». In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pp. 12–22.

Chersoni, Emmanuele, Alessandro Lenci e Philippe Blache (2017). «Logical Metonymy in a Distributional Model of Sentence Comprehension». In: *Sixth Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pp. 168–177.

Chersoni, Emmanuele et al. (2017). «Is Structure Necessary for Modeling Argument Expectations in Distributional Semantics?». In: *12th International Conference on Computational Semantics (IWCS 2017)*.

Deese, James (1966). *The structure of associations in language and thought*. Johns Hopkins University Press.

Kartsaklis, Dimitri e Mehrnoosh Sadrzadeh (2014). «A Study of Entanglement in a Categorical Framework of Natural Language». In: *Proceedings of the 11th Workshop on Quantum Physics and Logic (QPL)*. Kyoto ,Japan.

Lenci, Alessandro (2008). «Distributional semantics in linguistic and cognitive research». In: *Italian journal of linguistics* 20.1, pp. 1–31.

McRae, Ken e Kazunaga Matsuki (2009). «People use their knowledge of common events to understand language, and do so as quickly as possible». In: *Language and linguistics compass* 3.6, pp. 1417–1429.

Miller, George A e Walter G Charles (1991). «Contextual correlates of semantic similarity». In: *Language and cognitive processes* 6.1, pp. 1–28.

Rimell, Laura et al. (2016). «RELPRON: A relative clause evaluation data set for compositional distributional semantics». In: *Computational Linguistics* 42.4, pp. 661–701.

Rubenstein, Herbert e John B Goodenough (1965). «Contextual correlates of synonymy». In: *Communications of the ACM* 8.10, pp. 627–633.

Wittgenstein, Ludwig (1953). «Philosophical investigations (GEM Anscombe, trans.)». In: