MEDEA: Merging Event knowledge and Distributional vEctor Addition

Ludovica Pannitto, Alessandro Lenci
University of Pisa

CLIC-it 2018 - Turin, 10-12 December 2018

Compositional Distributional Semantics

Composing word representations into larger phrases and sentences notoriously represents a big challenge for distributional semantics (Lenci 2018).

Various approaches have been proposed ranging from simple arithmetic operations on word vectors (Mitchell e Lapata 2008), to algebraic compositional functions on higher-order objects (Baroni, Bernardi e Zamparelli 2014; Coecke, Clark e Sadrzadeh 2010), as well as neural networks approaches (Socher, Manning e Ng 2010; Mikolov et al. 2013.)

Vector addition still shows the best performances overall, its success being quite puzzling from the linguistic and cognitive point of view.

Psycholinguistic evidence shows that lexical items activate a great amount of **generalized event knowledge** (GEK) (Elman 2011; Hagoort e Berkum 2007; Hare et al. 2009), and that this knowledge is crucially exploited **during** online language processing, constraining the speakers' expectations about upcoming linguistic input (McRae e Matsuki 2009).

In this framework, **sentence comprehension** is phrased as the **identification of the event that best explains the linguistic cues used in the input** (Kuperberg e Jaeger 2016).

MEDEA

The purpose of MEDEA is to **integrate** vector addition with Generalized Event Knowledge activated by lexical items.

MEDEA is directly inspired by the model in Chersoni, Lenci e Blache 2017 and relies on two major assumptions:

- lexical items are represented with embeddings within a network of syntagmatic relations encoding prototypical knowledge about events;
- the semantic representation of a sentence is a structured object incrementally integrating the semantic information cued by lexical items.

MEDEA consists of two main components:

- a **Distributional Event Graph** (DEG) that models a fragment of semantic memory activated by lexical units;
- a **Meaning Composition Function** that dynamically integrates information activated from DEG to build a sentence semantic representation.

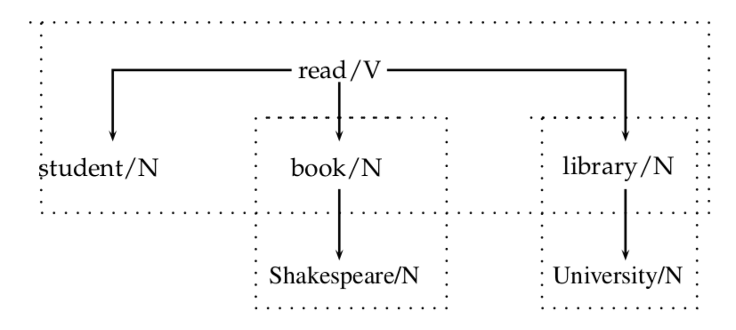
DEG

We assume a broad notion of *event*, corresponding to any **configuration of entities, actions, properties, and relationships**. An event can as complex as a whole sentence (e.g., *The student read a book*), but also a simpler relation holding in a noun phrase (e.g., *heavy book*).

We expect DEG to keep track of each event automatically retrieved from corpora, thus also containing information about **schematic or underspecified events**.

Events are **cued by all the potential participants**, depending on the distributional statistical association between the event and the participant.

Building DEG



A (hyper-)relation is added to the graph for each subset of each group extracted from the sentence

DEG: properties

Each node (i.e., lexeme) of DEG is associated with its distributional vector, therefore DEG can be queried on two parallel tiers:

- as a traditional distributional model, thus retrieving the most similar nodes to *w* (i.e., its paradigmatic neighbors), using a standard vector similarity measure like the cosine;
- retrieving the closest associates of *w* (i.e., its syntagmatic neighbors), using the weights on the graph edges.

para. neighbors	essay/N, anthology/N, novel/N, author/N, publish/N, biography/N, autobiography/N, nonfiction/N, story/N, novella/N
synt. neighbors	publish/V, write/V, read/V, include/V, child/N, series/N, have/V, buy/V, author/N, contain/V

Meaning Composition Function

We model sentence comprehension as the **creation of a semantic representation** SR, which includes two different yet interacting information tiers, that are equally relevant in the overall representation of sentence meaning:

- the **lexical meaning** component (LM), which is a context-independent tier of sentence meaning that accumulates the lexical content of the sentence, as traditional models do;
- an **active context** (AC), which aims at representing the most probable event, in terms of its participants, that can be reconstructed from DEG portions cued by lexical items.

Lexical Meaning

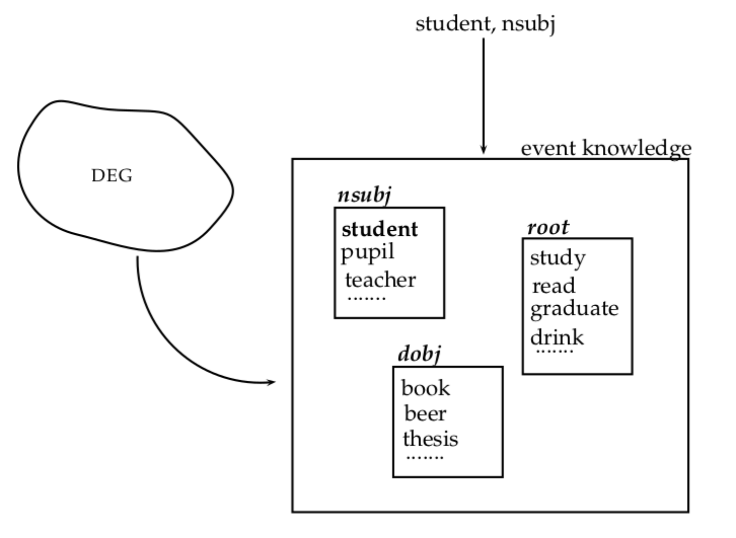
$$\begin{array}{ccccccc} \text{student} & & \text{read} & & \text{book} & & \\ \downarrow & & \downarrow & & \downarrow & & \\ \overrightarrow{student} & \oplus & \overrightarrow{read} & \oplus & \overrightarrow{book} & \longrightarrow & \overrightarrow{LM} \end{array}$$

Active Context

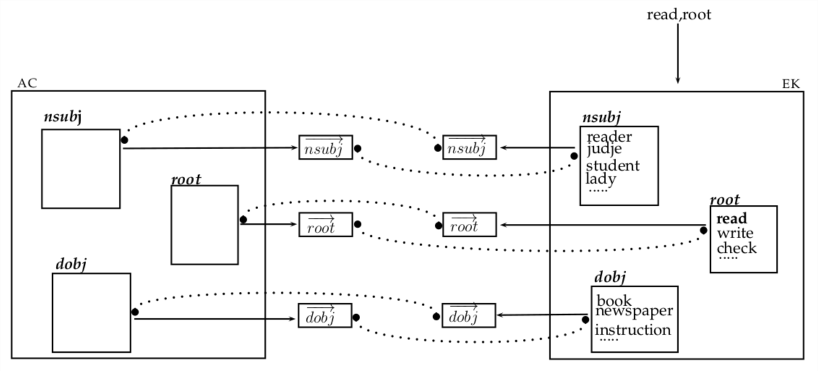
Each lexical item in the input activates a portion of GEK that is integrated into the current AC through a process of mutual re-weighting that aims at **maximizing the overall semantic coherence of the** SR.

- The AC is initialized empty;
- The sentence is processed incrementally, each time a new *lexeme - syntactic role* pair $\langle w_i, r_i \rangle$ (e.g., *student* - *nsbj*) is encountered, expectations about the set of upcoming roles in the sentences are generated from DEG;
- These expectations are then weighted with respect to what is already in the AC, and the AC is similarly adapted to the newly retrieved information.

Active Context: `retrieve`



Active Context: `merge`



The final semantic representation of a sentence consists of two vectors

- the **lexical meaning vector** ($\overrightarrow{LM}$)
- the **event knowledge vector** ($\overrightarrow{AC}$), which is obtained by composing the weighted centroids of each role in AC.

Evaluation

We wanted to evaluate the contribution of activated event knowledge in a **sentence comprehension task**.

Among the many existing datasets, we chose RELPRON (Rimell et al. 2016), a dataset of subject and object relative clauses, and the *transitive sentence similarity dataset* (TSS, Kartsaklis e Sadrzadeh 2014):

- intermediate level of grammatical complexity (i.e., they involve complete sentences)
- have fixed length structures featuring similar syntactic constructions (i.e., transitive sentences)

RELPRON 1,087 pairs, split in development and test set, made up by a *target* noun labeled with a syntactic role (either *subject* or *direct object*) and a *property* expressed as a *head noun* + *relative clause*.

(1) a. OBJ treaty/N: document/N that delegation/N negotiate/V
    b. SBJ treaty/N: document/N that grant/V independence/N

TSS dataset 108 pairs of transitive sentences, annotated with human similarity judgments. Each transitive sentence in composed by a triplet *subject verb object*.

(2) a. government use power
    b. authority exercise influence

(3) a. team win match
    b. design reduce amount

We tailored the construction of the DEG to this kind of simple syntactic structures, restricting to the case of relations among pairs of event participants.

In MEDEA, the SR is composed of two vectors:

- $\overrightarrow{LM}$, as the **sum of the word embeddings** (as this was the best performing model in literature, on the chosen datasets);
- $\overrightarrow{AC}$, obtained by **summing up all the weighted centroids of triggered participants**. Each *lexeme - syntactic role* pair is used to retrieve its 50 top s-neighbors from the graph. The top 20 re-ranked elements were used to build each weighted centroid.

inventory - document that store maintains

- the head noun *document* is encountered: its vector is activated as event knowledge for the *object* role of the sentence and constitutes the contextual information in AC against which GEK is re-weighted;
- *store* as a subject triggers some *direct object* participants, such as *product*, *range*, *item*, *technology*, etc. If the centroid were built from the top of this list, the cosine similarity with the target would be around 0.62;
- *s-neighbours* of *store* are re-weighted according to the fact that AC contains some information about the target already, (i.e., the fact that it is a document). The re-weighting process has the effect of placing on top of the list elements that are more similar to *document*. Thus, now we find *collection*, *copy*, *book*, *item*, *name*, *trading*, *location*, etc., improving the cosine similarity with the target, that goes up to 0.68; iv.) the same happens for *maintain*: its *s-neighbors* are retrieved and weighted against the complete AC, improving their cosine similarity with *inventory*, from 0.55 to 0.61.

Task

RELPRON - we produced the compositional representation of each property in terms of $\textsc{sr}$, and then ranked for each target all the 518 properties of the development set, according to their similarity to the target

$$s = cos(\overrightarrow{target}, \overrightarrow{LM}) + cos(\overrightarrow{target}, \overrightarrow{AC}) \qquad (1)$$

TSS - we evaluated the correlation of our scores with human ratings with Spearman's ρ

$$s = cos(\overrightarrow{LM_1}, \overrightarrow{LM_2}) + cos(\overrightarrow{AC_1}, \overrightarrow{AC_2}) \qquad (2)$$

Results - RELPRON

	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**

Results - Transitive sentences

	transitive sentences dataset		
	LM	AC	LM+AC
sbj	0.432	0.475	0.**482**
root	0.525	0.547	0.**555**
obj	0.628	0.537	0.**637**
sbj+root	0.**656**	0.622	0.648
sbj+obj	0.653	0.605	0.**656**
root+obj	0.732	0.696	0.**750**
sbj+root+obj	0.732	0.686	0.**750**

Conclusioni

Wrap-up

We provided a basic implementation of a meaning composition model, which aims at being **incremental** and **cognitively plausible**.

While still relying on vector addition, our results suggest that distributional vectors do not encode sufficient information about event knowledge, and that, in line with psycholinguistic results, activated GEK plays an important role in building semantic representations during online sentence processing.

Ongoing work

Our ongoing work focuses on refining the way in which this event knowledge takes part in the processing phase and testing its performance on more complex datasets: while both RELPRON and the transitive sentences dataset provided a straight forward mapping between syntactic label and semantic roles, more naturalistic datasets show a much wider range of syntactic phenomena that would allow us to test how expectations jointly work on syntactic structure and semantic roles.

Thank you :)

References

Asher, Nicholas et al. (2016). «Integrating Type Theory and Distributional Semantics: A Case Study on Adjective–Noun Compositions». In: *Computational Linguistics* 42.4, pp. 703–725.

Baroni, Marco, Raffaela Bernardi e Roberto Zamparelli (2014). «Frege in Space: A Program of Compositional Distributional Semantics». In: *Linguistic Issues in Language Technology* 9.6, pp. 5–110.

Blacoe, William e Mirella Lapata (2012). «A comparison of vector-based representations for semantic composition». In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, pp. 546–556.

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)*.

Coecke, Bob, Stephen Clark e Mehrnoosh Sadrzadeh (2010). *Mathematical Foundations for a Compositional Distributional Model of Meaning*. Rapp. tecn.

Elman, Jeffrey L (2011). «Lexical knowledge without a lexicon?». In: *The mental lexicon* 6.1, pp. 1–33.

Hagoort, Peter e Jos van Berkum (2007). «Beyond the sentence given». In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1481, pp. 801–811.

Hare, Mary et al. (2009). «Activating event knowledge». In: *Cognition* 111.2, pp. 151–167.

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.

Kuperberg, Gina R e T Florian Jaeger (2016). «What do we mean by prediction in language comprehension?». In: *Language, cognition and neuroscience* 31.1, pp. 32–59.

Lenci, Alessandro (2018). «Distributional Models of Word Meaning». In: *Annual Review of Linguistics* 4, pp. 151–171.

Manning, Christopher D. et al. (2014). «The Stanford CoreNLP Natural Language Processing Toolkit». In: *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010.

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.

Mikolov, Tomas et al. (2013). «Distributed representations of words and phrases and their compositionality». In: *Advances in neural information processing systems*, pp. 3111–3119.

Mitchell, Jeff e Mirella Lapata (2008). «Vector-based Models of Semantic Composition.». In:

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

Socher, Richard, Christopher D Manning e Andrew Y Ng (2010). «Learning continuous phrase representations and syntactic parsing with recursive neural networks». In: *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*. Vol. 2010, pp. 1–9.