# A Structured Distributional Model of Sentence Meaning and Processing

E. Chersoni [1] E. Santus [2] L. Pannitto [3] A. Lenci [4] P. Blache [5] C.-R. Huang [1]

[1]Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University

[2]Computer Science and Artificial Intelligence Lab, MIT

[3]CIMeC, University of Trento

[4]CoLing Lab, University of Pisa

[5]Laboratoire Parole et Langage, Aix-Marseille University

# Sentence Meanings as Vectors

- The mainstream approach in distributional semantics assumes the representation of sentence meaning to be a vector, exactly like lexical items, built with different methods
  - pointwise operations to combine lexical vectors ($\mathbf{s} = \sum_{i=1}^{n} \mathbf{w_i}$, Mitchell and Lapata 2010)
  - higher-order linear-algebraic objects such as matrices and functions (Coecke et al., 2010; Baroni et al. 2014)
  - sentence embeddings directly learned with encoding-decoding neural networks (Kiros, et al. 2015, Conneau et al. 2017, Devlin et al. 2019)

# Sentence Meanings as Vectors

- Vector addition is still able to outperform more complex vector composition function and sentence embeddings (Wieting et al. 2016, Arora et al. 2017, Hill et al. (2016, Shen et al. 2018), but it is theoretically unsatisfactory

- Untrained sentence encoders using pre-trained embeddings behave as well as trained ones (Wieting and Kiela 2019)
  - "Most of the power in modern NLP systems is derived from having high-quality word embeddings, rather than from having better encoders [. . . ] Therefore one may wonder to what extent sentence encoders are worth the attention they're receiving"

# Structured Distributional Model

Chersoni, E., et al. (2019). "A Structured Distributional Model of Sentence Meaning and Processing". *Natural Language Engineering*, 25(4), pp. 483?502

- Natural language comprehension involves the dynamic (Kamp 1981, 2013) construction of semantic representations: mental characterization of the events or situations described in sentences

- SDM is grounded on psycholinguistic data showing that common-sense knowledge about events plays a key role in sentence comprehension (Elman and McRae 2019)

# Generalized Event Knowledge (GEK)

McRae and Matsuki (2009), "People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible", *Language and Linguistics Compass*, 3:1417-1429

- Long-term semantic memory stores generalized knowledge about events and their participants (GEK), derived from first-hand experience and from linguistic experience
- Linguistic expressions are cues to activate various aspects of GEK (Elman 2014)
- GEK is used to generate expectations (predictions) about the upcoming linguistic input, minimizing the processing effort (Bicknell et al. 2010, Matsuki et al. 2011, Paczynski and Kuperberg 2012, Metusalem et al. 2012)

  "the specific choice of verb can be used to bring to mind somewhat different scenarios, such as *eating* versus *dining*. [...] Instrument nouns can cue certain types of eating, as in *eating with a fork* versus *eating with a stick*. Finally, event nouns like *breakfast* or location nouns like *cafeteria* cue specific types of eating scenarios."
  (McRae and Matsuki 2009: 1419)

# SDM: structure

Chersoni, E., et al. (2019). "A Structured Distributional Model of Sentence Meaning and Processing". *Natural Language Engineering*, 25(4), pp. 483?502

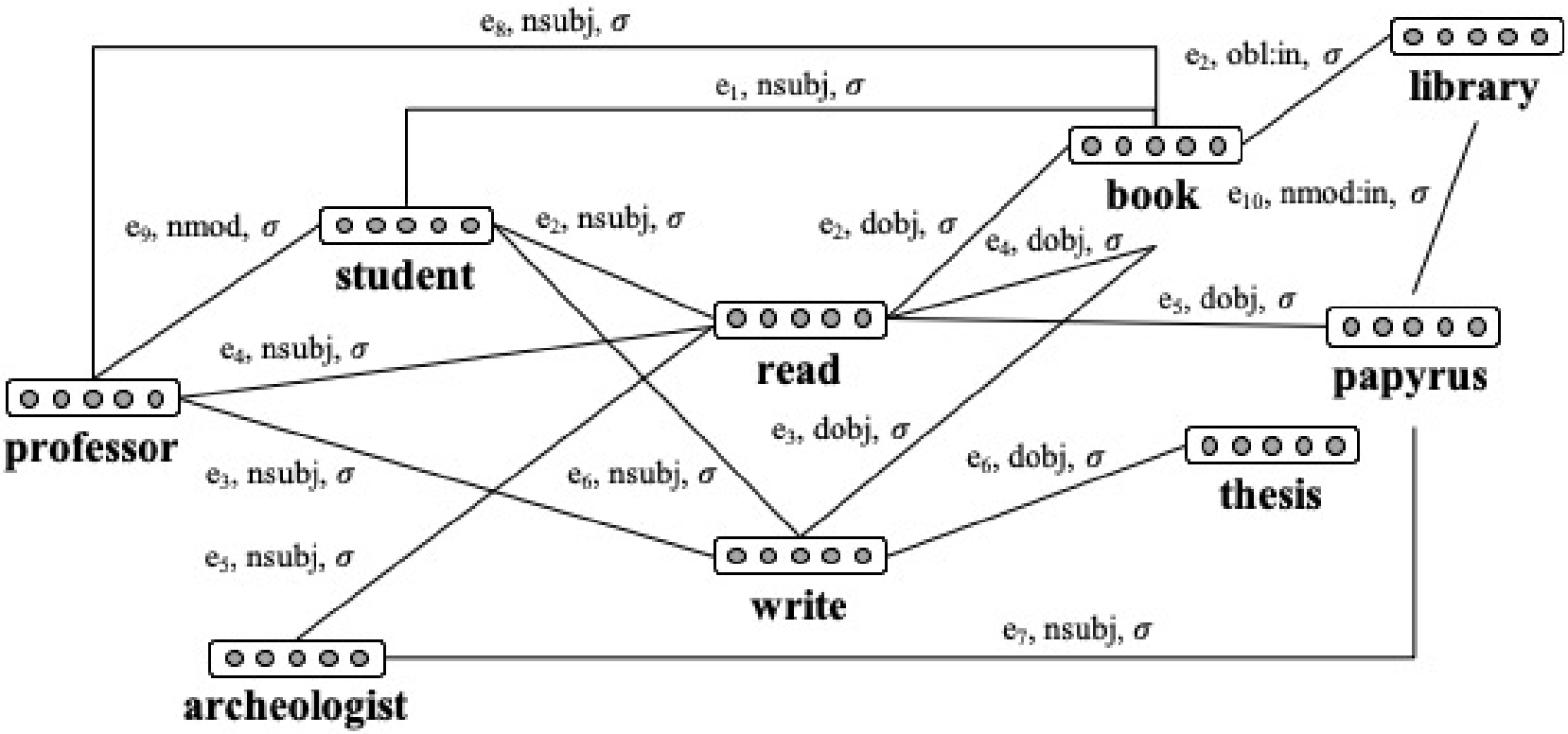
- "Division of labour" between formal and vector semantics: Semantic Representations are logical forms enriched with word embeddings (cf. also Beltagy et al. 2016, McNally 2017)

Distributional Event Graph (DEG): network of relations encoding knowledge about events and their typical participants

Semantic Representation (SR): formal structure that dynamically combines the information cued by lexical items

# The Distributional Event Graph (DEG)

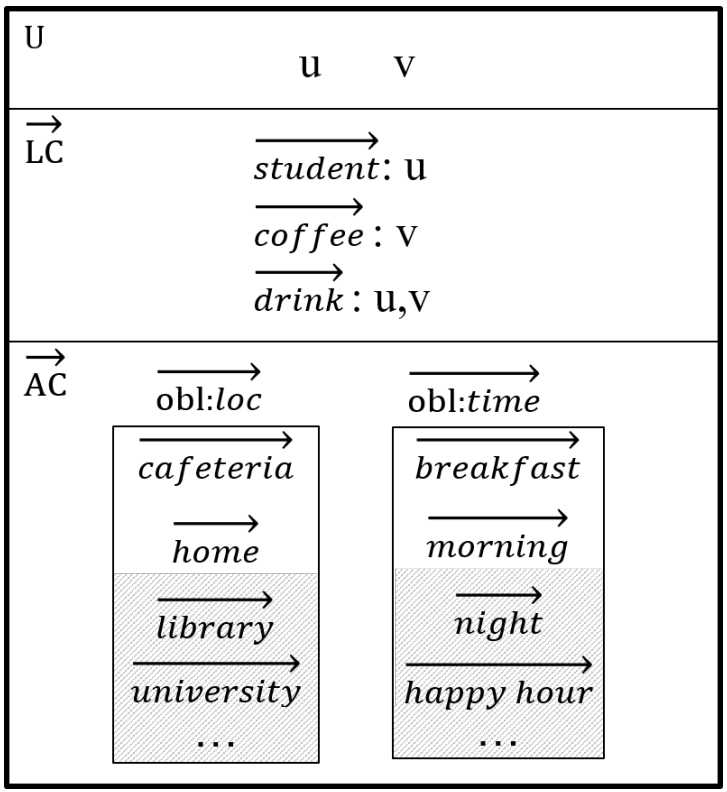
- DEG is a model of the GEK derived from the linguistic input
- an event is an *n*-ary relation between entities, and corresponds to the notion of situation knowledge or thematic associations (Binder 2016)



# The Semantic Representation (SR)

- SR is a formal structure directly inspired by DRT and consisting of three information tiers:

*The student drinks the coffee*



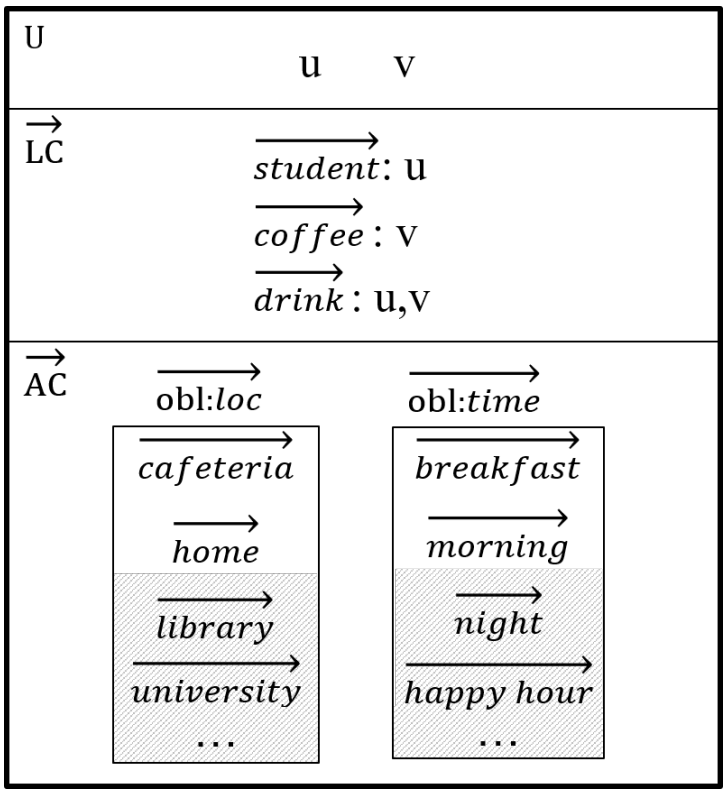

Universe (U)

- It includes the entities mentioned in the sentence (corresponding to the *discourse referents* in DRT)

# The Semantic Representation (SR)

- SR is a formal structure directly inspired by DRT and consisting of three information tiers:

*The student drinks the coffee*



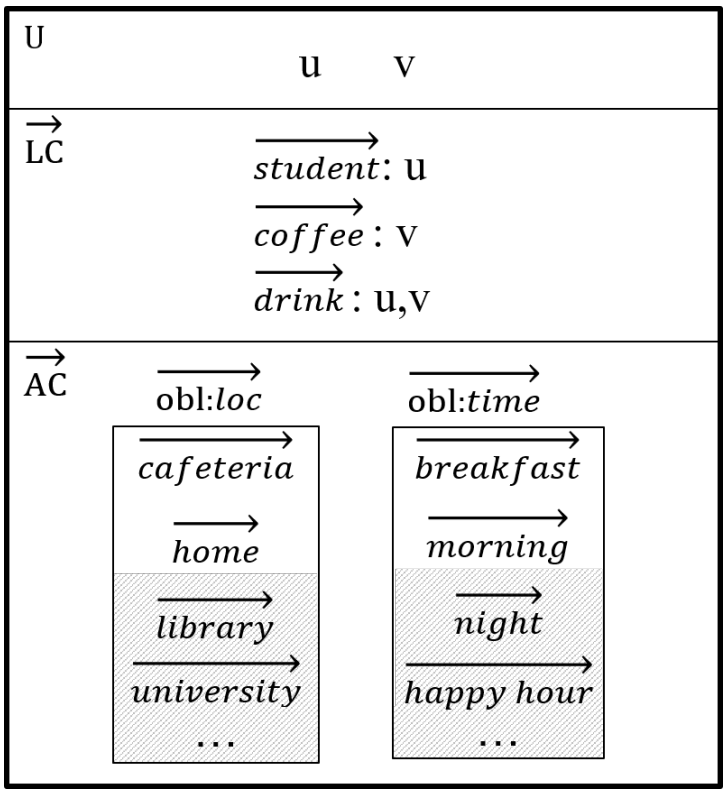

Linguistic Conditions (LC)

- A vector $\overrightarrow{LC}$ obtained from the linear combination of the embeddings of the words contained in the sentence

# The Semantic Representation (SR)

- SR is a formal structure directly inspired by DRT and consisting of three information tiers:

*The student drinks the coffee*



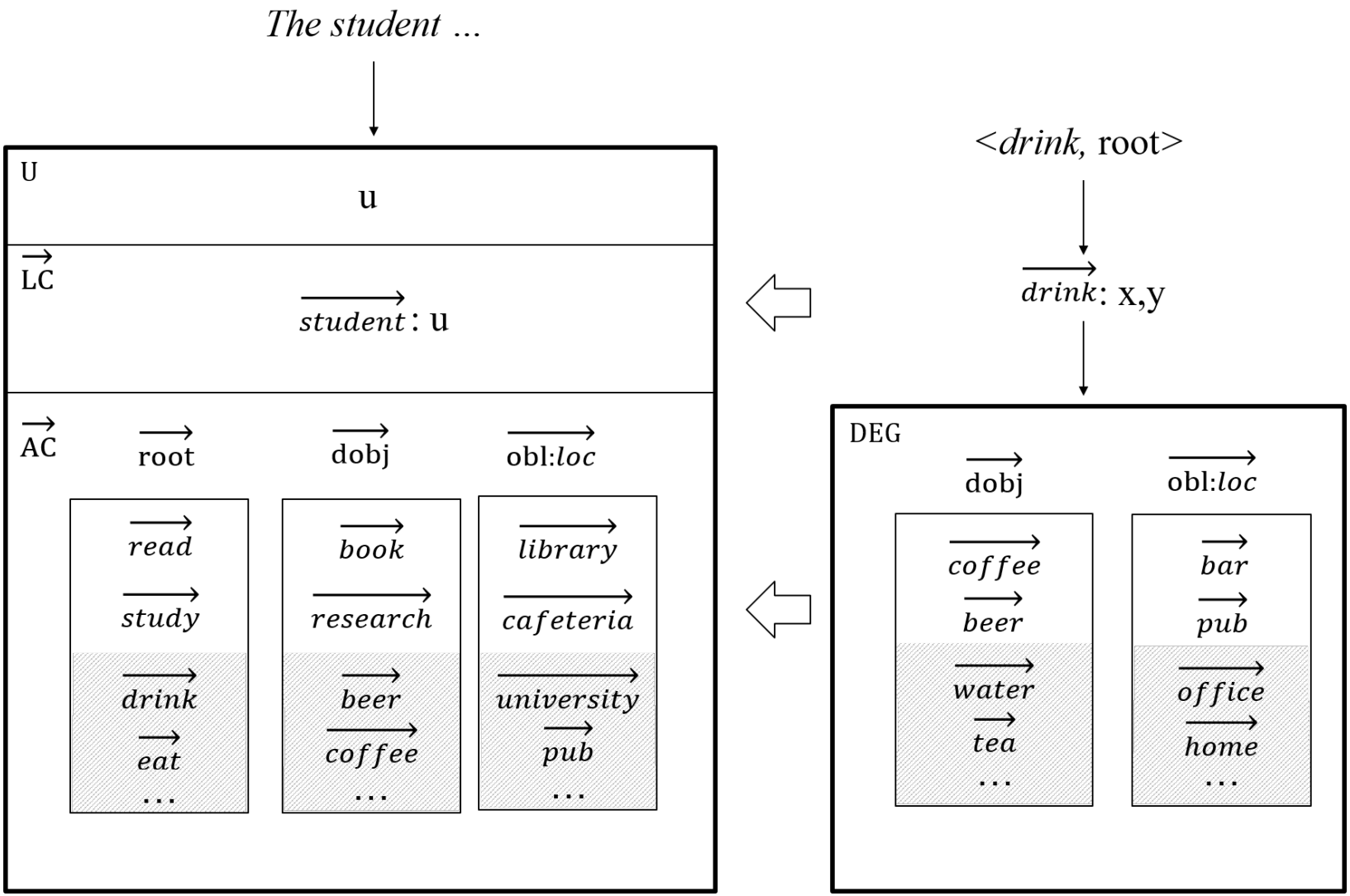

Active Context (AC)

- It contains a set of ranked lists of embeddings corresponding to the most likely words expected to fill a given syntactic role, represented with the weighted centroid vector of their $k$ most prominent items
- The ranking of each element in AC depends on two factors:
  - degree of activation from DEG by the lexical items
  - overall coherence with respect to the rest of information in AC

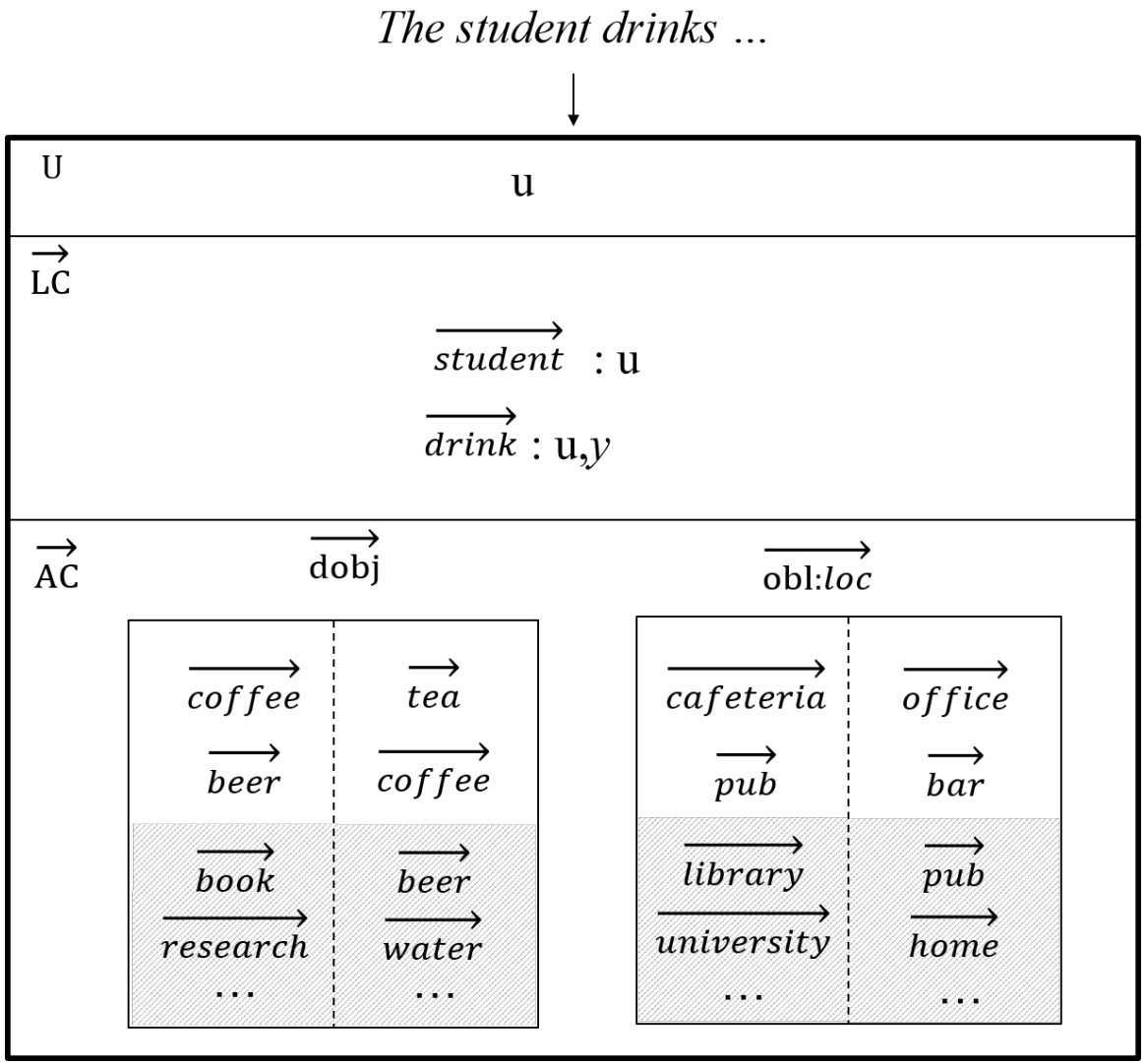
# Semantic Composition as Information Integration

- The LC and AC components of each SR are represented with vectors that are incrementally updated with the information activated by lexical items
- When we process a new pair $\langle w_i, r_i \rangle$ with a lexeme $w_i$ and syntactic role $r_i$:
  1. LC is updated with the embedding $\overrightarrow{w_i}$, which is simply added to $\overrightarrow{LC}$
  2. AC is updated with the embeddings activated from DEG by $w_i$:
     - the event knowledge activated by $w_i$ for a given role $r_i$ is re-ranked according to cosine similarity with the vector $\overrightarrow{r_i}$ available in AC
     - the newly retrieved information is used to update the centroids in AC, in order to maximize the semantic coherence of the representation

# Semantic representation (SR)



# Semantic representation (SR)



# RELPRON

Rimell, L., Maillard, J., Polajnar, T., Clark, S. (2016). "RELPRON: A relative clause evaluation data set for compositional distributional semantics.". In *Computational Linguistics, 42(4), 661-701.*

- 518 target–property pairs, where the target is a noun labelled with a syntactic function (either subject or direct object) and the property is a subject or object relative clause providing the definition of the target
  - *telescope*: *device* that *detects planets*
  - *telescope*: *device* that *observatory has*
- we produce a compositional representation for each of the properties. In each definition, the verb,the head noun and the argument are composed to obtain a representation of the property
  - models are evaluated in terms of the Mean Average Precision: given a term, all properties are ranked according to their similarity score

# DTFit

Vassallo, P., et al. (2018). "Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality". In *Proceedings of the Workshop on Linguistic and Neurocognitive Resources*

- 795 triplets, each differing only for the **Patient** role:
  - *sergeant*_N *assign*_V *mission*_N (typical)
  - *sergeant*_N *assign*_V *homework*_N (atypical)
- 300 quadruples, each differing only for the **Location** role:
  - *policeman*_N *check*_V *bag*_N *airport*_N (typical)
  - *policeman*_N *check*_V *bag*_N *kitchen*_N (atypical)
- For each patient and location tuple, the task is to predict the upcoming argument on the basis of the previous ones
  - we build a compositional vector representation for each dataset item by excluding the last argument in the tuple, and we measured the cosine similarity between the resulting vector and the argument vector
  - models are evaluated in terms of the Spearman correlation between the similarity scores and the human ratings

# Results

| | ADDITIVE | | | SDM | | |
|---|---|---|---|---|---|---|
| | sg | cbow | c-phrase | sg | cbow | c-phrase |
| verb | 0,16 | 0,16 | 0,13 | 0,21 | 0,20 | 0,19 |
| arg | 0,33 | 0,32 | 0,37 | 0,38 | 0,36 | 0,41 |
| hn + verb | 0,26 | 0,25 | 0,21 | 0,27 | 0,28 | 0,26 |
| hn + arg | 0,44 | 0,46 | 0,45 | 0,50 | 0,50 | 0,50 |
| verb + arg | 0,43 | 0,36 | 0,41 | 0,41 | 0,36 | 0,41 |
| hn + verb + arg | 0,50 | 0,47 | 0,47 | 0,54 | 0,52 | 0,54 |

Table: Results on RELPRON, expressed in terms of Mean Average Precision

| | ADDITIVE | | | SDM | | |
|---|---|---|---|---|---|---|
| | sg | cbow | c-phrase | sg | cbow | c-phrase |
| Patients | 0,63 | 0,52 | 0,60 | 0,65 | 0,62 | 0,66 |
| Locations | 0,74 | 0,70 | 0,74 | 0,75 | 0,74 | 0,76 |

Table: Results on DTFit, expressed in terms of Spearman's correlation

# No Structure, No Meaning

- Continuous vector representations of meaning have several advantages but a theoretically and empirically adequate model of sentence meaning is still missing
- It is possible and useful (and perhaps necessary?) to combine semantic structures with (distributional) vector representations of meaning
- SDM is based on an incremental model of sentence representation that integrates rich, data-driven common-sense knowledge about events in a structured formal representation
- SDM is currently being tested on other datasets to predict whether sentences represent typical vs. atypical events