COLING 2025 • Abu Dhabi

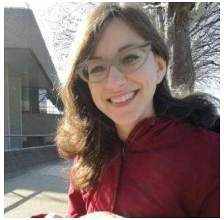
The 31st International Conference on Computational Linguistics

EnnCore

Neuro-symbolic AI Lab

Montague semantics and modifier consistency measurement in neural language models

Danilo Carvalho, Edoardo Manino, Julia Rozanova, Lucas Cordeiro, Andre Freitas

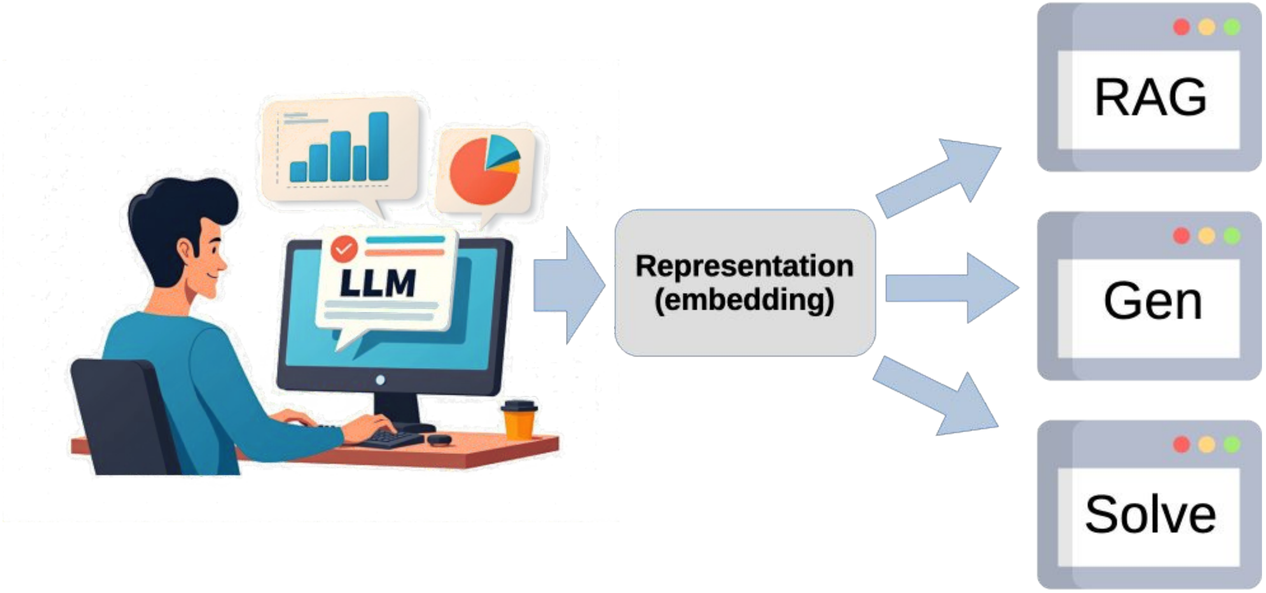
    

MANCHESTER 1824 The University of Manchester

idiap RESEARCH INSTITUTE

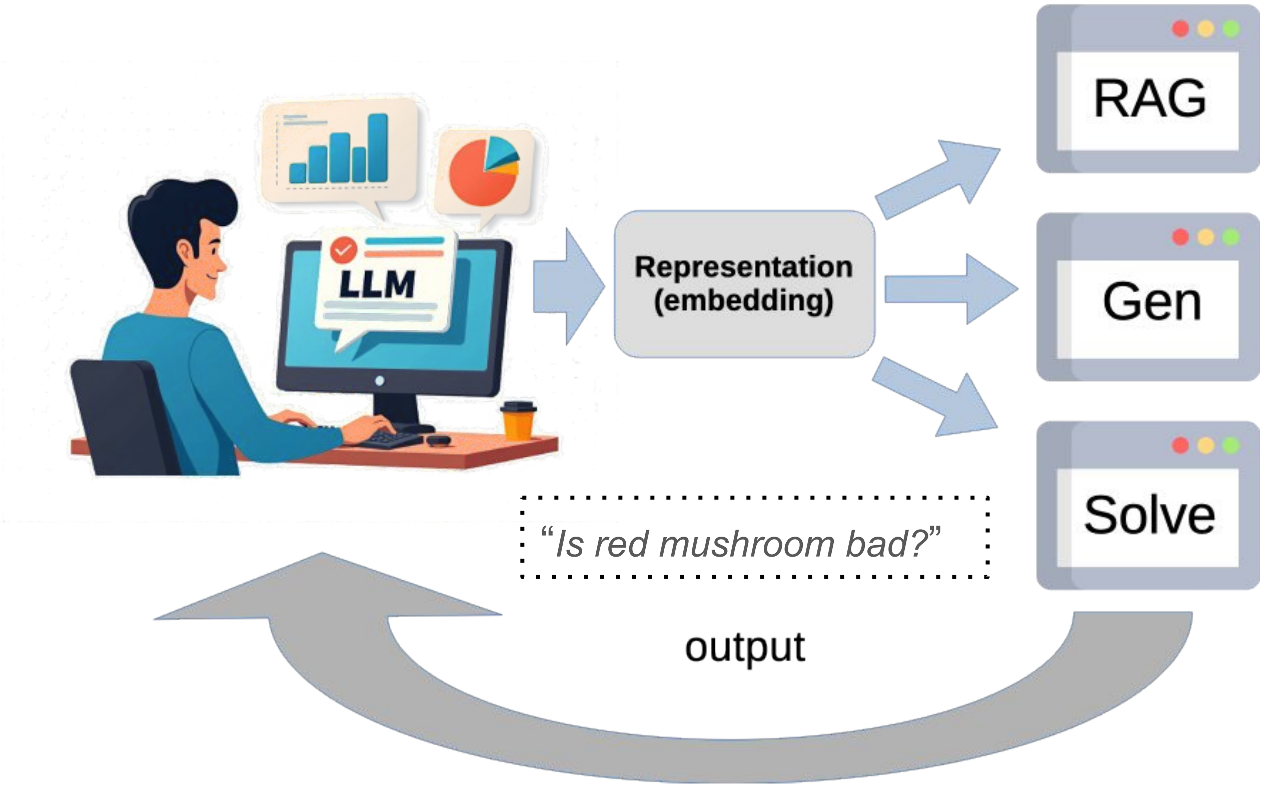
CANCER RESEARCH UK | Manchester Institute

Motivation



Expanding use of distributional representations
based on Neural Language Models

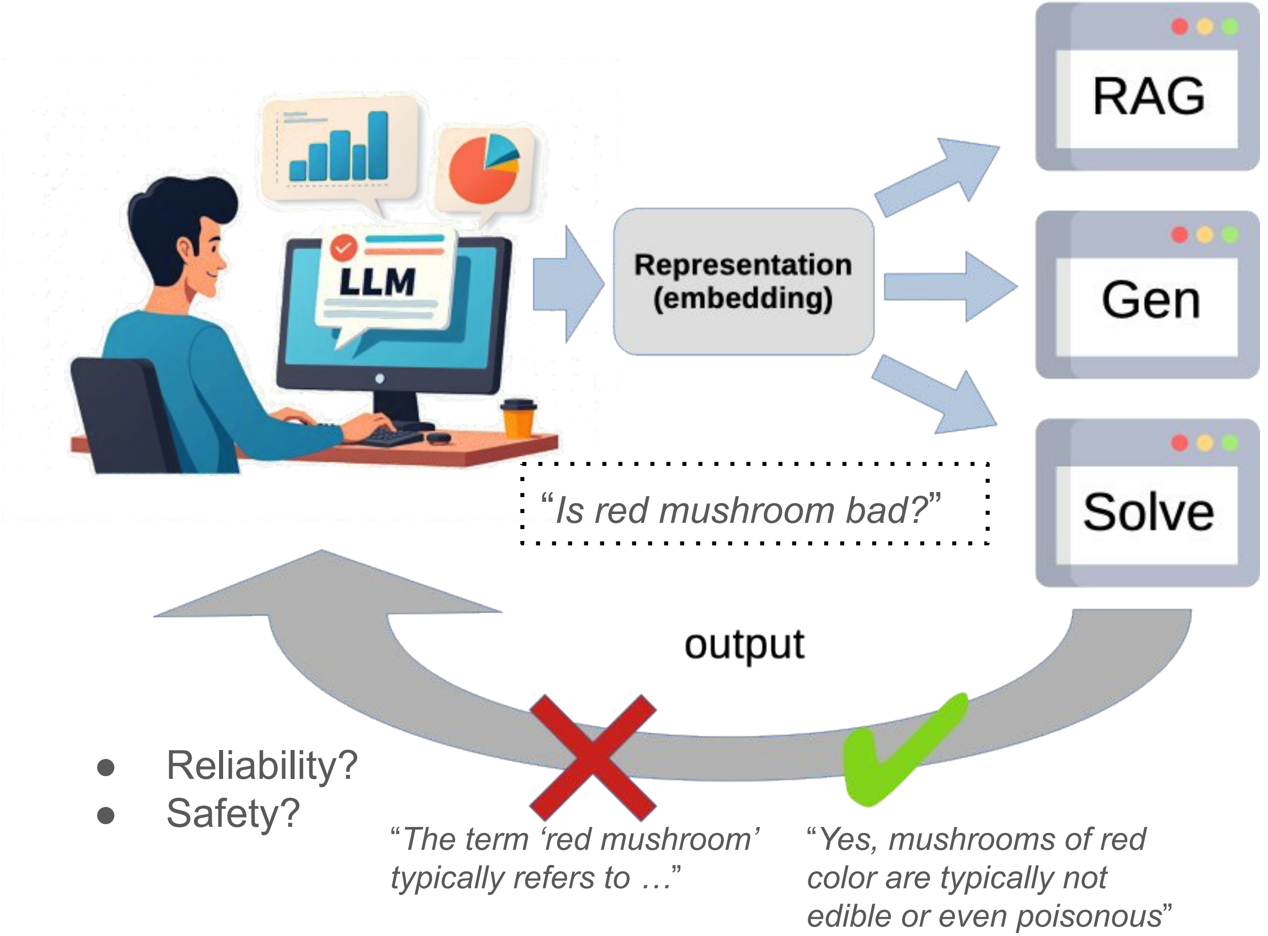
Motivation



Linguistic assumptions:

- Semantic alignment
- Pragmatics
- Compositionality

Motivation



- Reliability?
- Safety?

Linguistic assumptions:

- Semantic alignment
- Pragmatics
- Compositionality

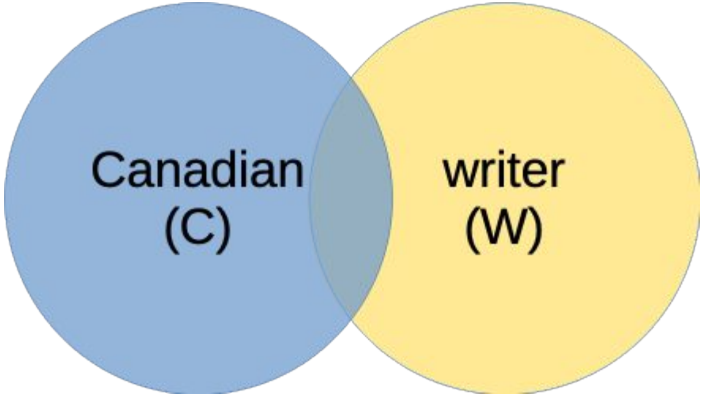
Capabilities/Limitations:

- Do embeddings capture essential compositional properties?

Case study:
Modifier phenomena

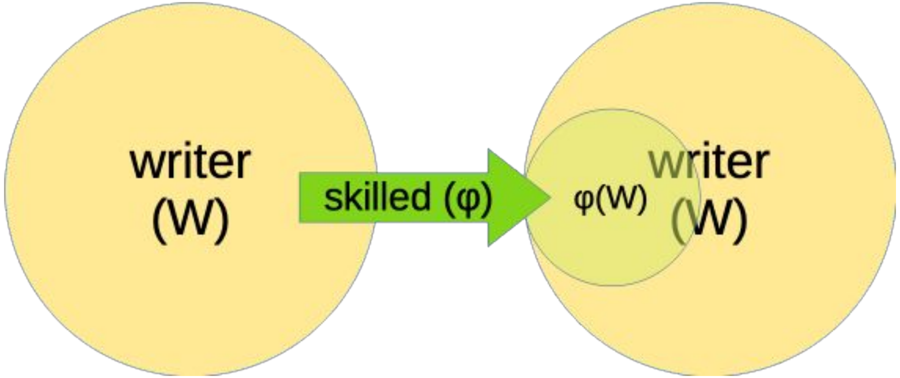
Modifier phenomena in NL

- Modification: a set of compositional principles regarding intensional interpretations from a Montagovian formalism (denotations).
- Adjective phrases being the object of analysis
- Adjective types:
    - Intersective (or extensional): describe the intersection of the noun denotation with one from the adjective itself.



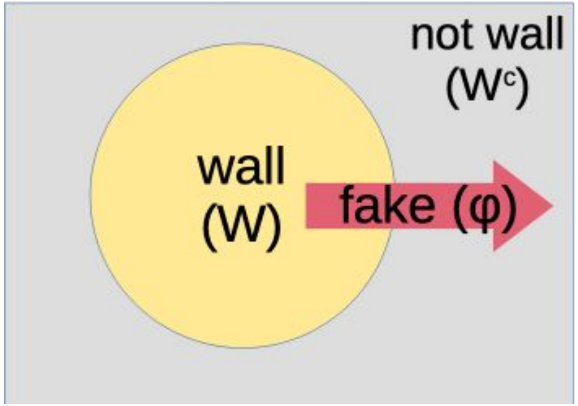
Modifier phenomena in NL

- Modification: a set of compositional principles regarding intensional interpretations from a Montagovian formalism (denotations).
- Adjective phrases being the object of analysis
- Adjective types:
  - Subsective (non-intersective): describe a strict subset of the noun denotation it modifies.



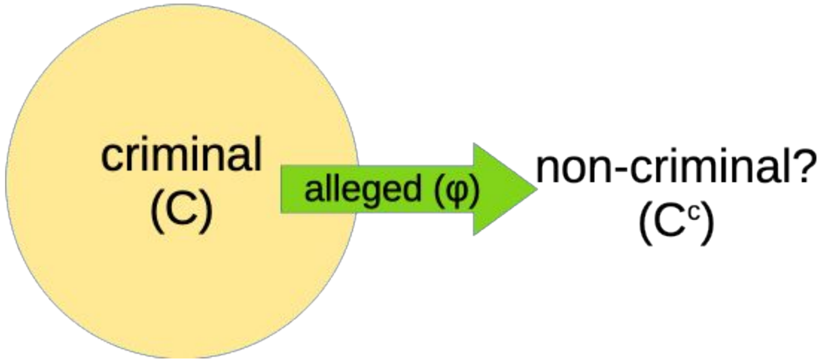
Modifier phenomena in NL

- Modification: a set of compositional principles regarding intensional interpretations from a Montagovian formalism (denotations).
- Adjective phrases being the object of analysis
- Adjective types:
  - Privative non-subsective: describe a set that is completely disjoint from the denotation of the noun it modifies.



Modifier phenomena in NL

- Modification: a set of compositional principles regarding intensional interpretations from a Montagovian formalism (denotations).
- Adjective phrases being the object of analysis
- Adjective types:
    - Plain non-subsective: describe a set that may or may not be a subset of the modified noun's denotation, depending on
      the adjective itself or the context.



Modifier phenomena in NL

- <u>Modification</u>: a set of compositional principles regarding intensional interpretations from a Montagovian formalism (denotations).
- Adjective phrases being the object of analysis
- Adjective types:
    - <u>Ambiguous</u>: can be applied as any of the previous categories, depending on the noun it modifies and the context.

      Example: in “big truck” the interpretation of “big” is intersective, while in “big fool” is subsective non-intersective.

Montague Denotations

We say that a noun n can be modified by an adjective a to form an adjective phrase: $p = an$

For example: in the phrase *p* = "Canadian writer", we have the following

Montague denotations (intensions): and corresponding sets (extensions):

$$n(x) = \lambda x.[writer(x)] \qquad N \equiv \{x \,|\, n(x) = \top\}$$

$$a(x) = \lambda x.[Canadian(x)] \qquad A \equiv \{x \,|\, a(x) = \top\}$$

$$p(x) = \lambda x.[a(x) \wedge n(x)] \qquad P \equiv A \cap N$$

Montague Denotations

On the other hand, if a is a non-intersective adjective, then the denotation of p involves functions over sets.

For example, the phrase p = “skilled writer” requires the following Montague denotations:

$$a(n, x) = \lambda n.\lambda x[skilled(n(x), x)]$$
$$p(x) = \lambda x.[a(W, x)]$$

where function a can discriminate whether x is a skilled writer, but has no concept of “skilfulness” in general. Accordingly, the corresponding sets (extensions) are:

$$P \equiv A \equiv \{x \,|\, p(x) = \top\} \subseteq N$$

Denotation Set Distance

Considering the intersective case: $P \equiv A \cap N$

The fact that *P* is a subset of both *A* and *N* and suggests the following distance relations between sets:

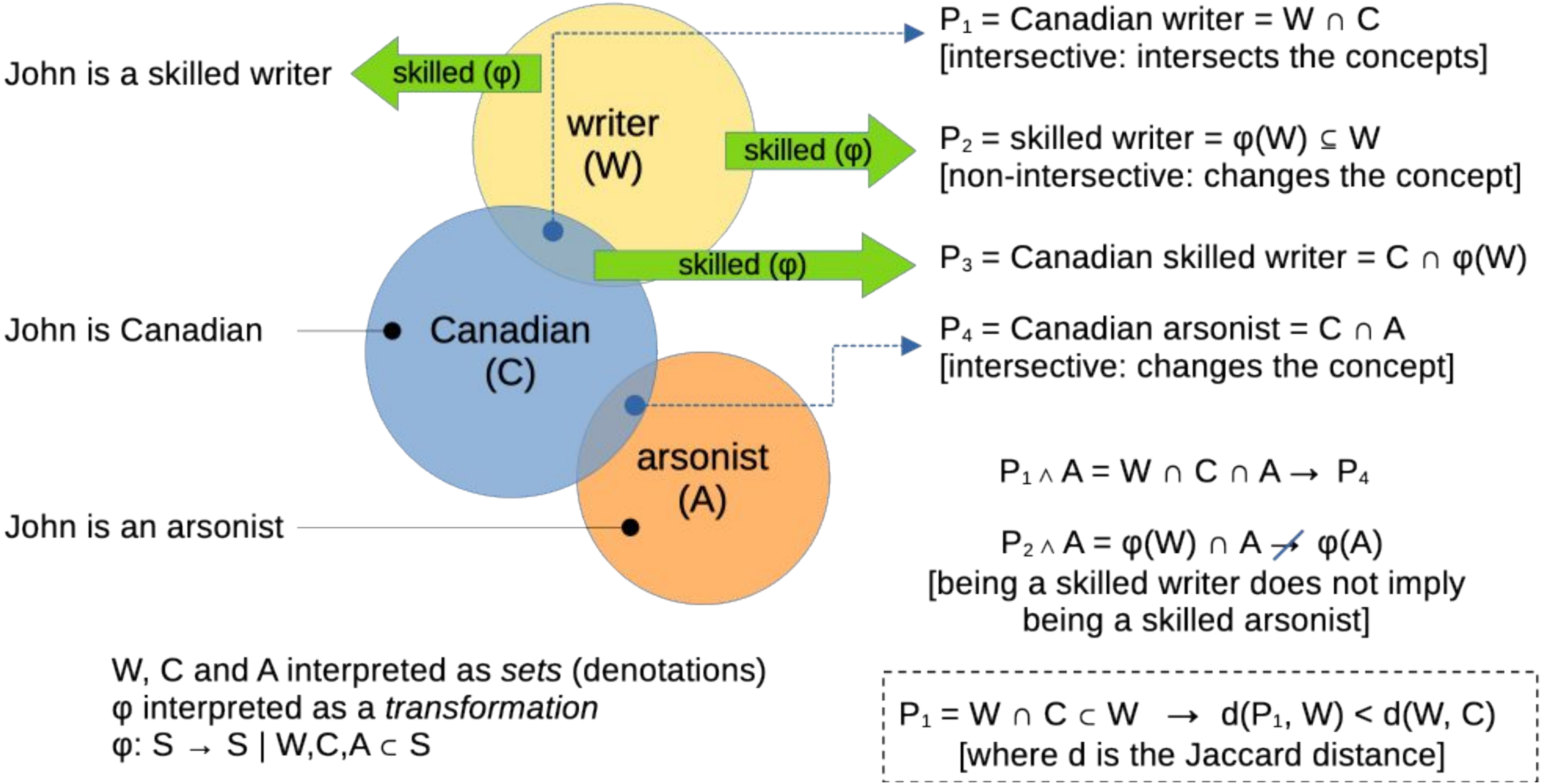
$$d(P, N) \leq d(N, A)$$

$$d(P, A) \leq d(N, A)$$

where is the Jaccard distance.

For longer phrases $p = a_1 \cdots a_k n$ with k adjectives, the distance relations can be generalised to:

$$d(P, A_i) \leq d(N, A_i) \quad \forall i$$


John is a skilled writer
skilled (φ)
writer (W)
skilled (φ)
skilled (φ)
P₁ = Canadian writer = W ∩ C
[intersective: intersects the concepts]
P₂ = skilled writer = φ(W) ⊆ W
[non-intersective: changes the concept]
P₃ = Canadian skilled writer = C ∩ φ(W)
John is Canadian
Canadian (C)
P₄ = Canadian arsonist = C ∩ A
[intersective: changes the concept]
arsonist (A)
John is an arsonist
P₁ ∧ A = W ∩ C ∩ A → P₄
P₂ ∧ A = φ(W) ∩ A ↛ φ(A)
[being a skilled writer does not imply being a skilled arsonist]
W, C and A interpreted as sets (denotations)
φ interpreted as a transformation
φ: S → S | W,C,A ⊂ S
P₁ = W ∩ C ⊂ W → d(P₁, W) < d(W, C)
[where d is the Jaccard distance]

On Neural Language Models

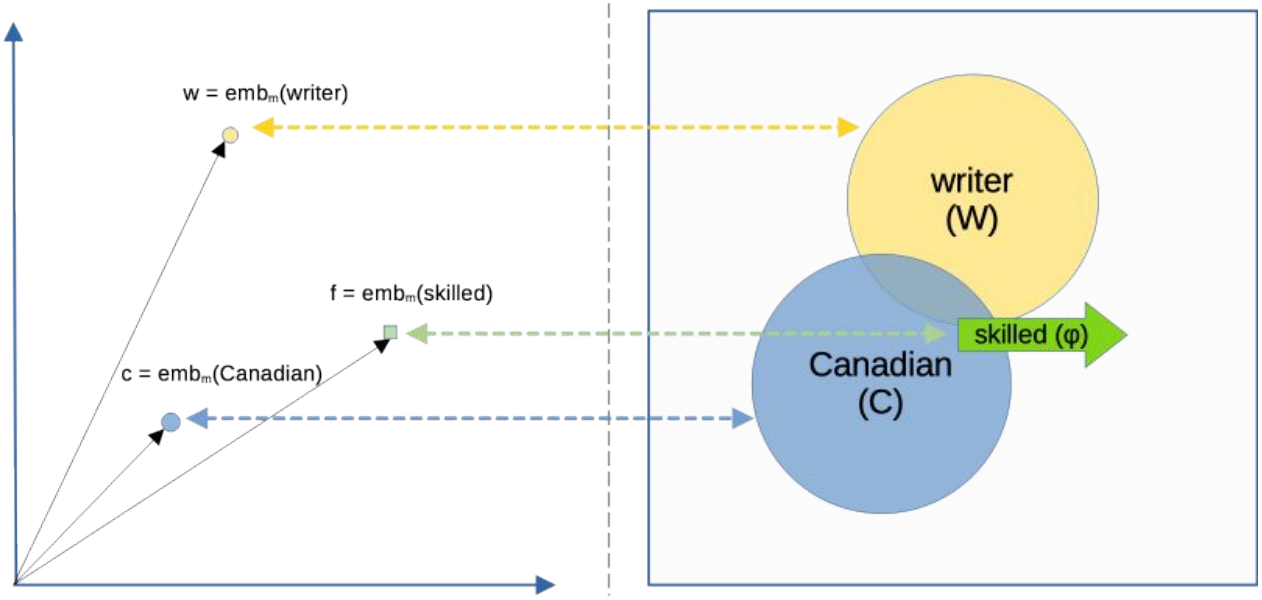
Our core hypothesis:

- If the phrase embedding correctly represents its denotation, we should observe some analogous inclusion relations between them.
- Since embeddings are defined in vector space, the inclusion relations must be replaced with another appropriate measure (e.g., cosine, Euclidean).

Distributional questions:

- Can we expect to observe a correspondence of these theoretical linguistic properties in neural language models that operate on dense vector spaces?
- To what degree can we observe evidence of the compositional effect of adjective modifiers?
  - Do contextual models differ from non-contextual ones in this regard?

Embedding-Denotation Analogy



Compositional intersectivity test

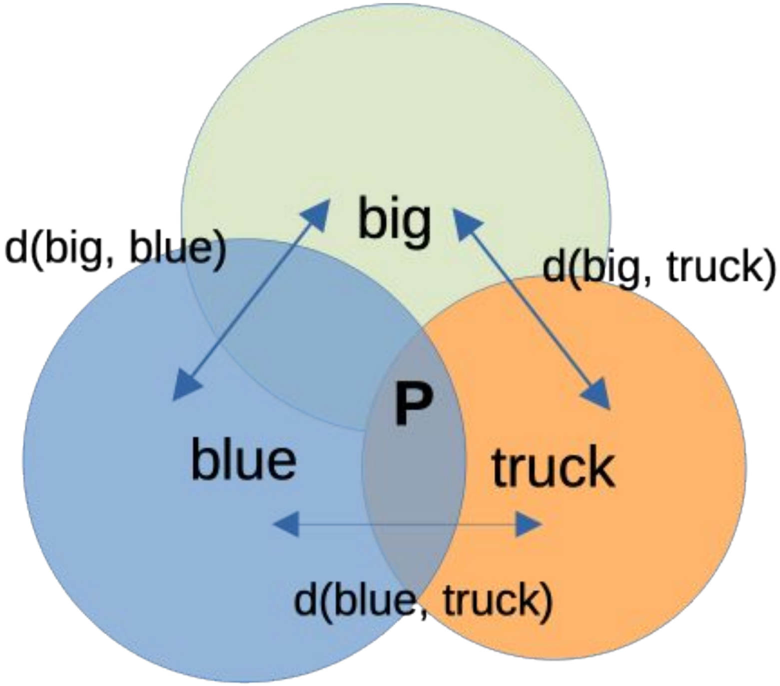
$E_{m,L}$ $\begin{cases} \text{dist}(\text{emb}_m(\text{Canadian writer}), c) \leq \text{dist}(c, w) \\ \text{dist}(\text{emb}_m(\text{Canadian writer}), w) \leq \text{dist}(c, w) \end{cases}$ ⟷ $\text{dist}(W \cap C, C) \leq \text{dist}(C, W)$, $\text{dist}(W \cap C, W) \leq \text{dist}(C, W)$

Compositional non-subsectivity test

$E_{m,L}$ $\{ \text{dist}(\text{emb}_m(\text{skilled writer}), f) \leq \text{dist}(\text{emb}_m(\text{skilled writer}), w)$ $\Delta\varphi(W) \leq \text{dist}(\varphi(W), W)$

Consistency Tests

- Testing intersectivity (single phrase):



$$d(P, big) \leq d(blue, big)$$

$$d(P, big) \leq d(truck, big)$$

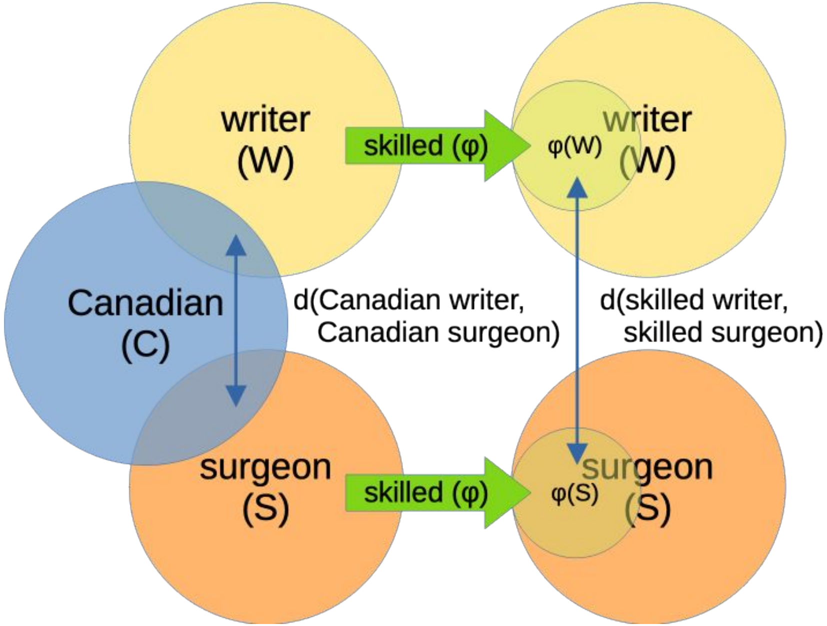
Same for the other words.

The consistency measure is then the expectation of those relations to be true when the adjectives are intersective.

Requires that the embedding of an adjective-noun phrase lies closer to each term than the distance between any pair of terms.

Consistency Tests

- Testing intersectivity (phrase pairs):



$$d(CW, CS) \leq (\varphi(W), \varphi(S))$$

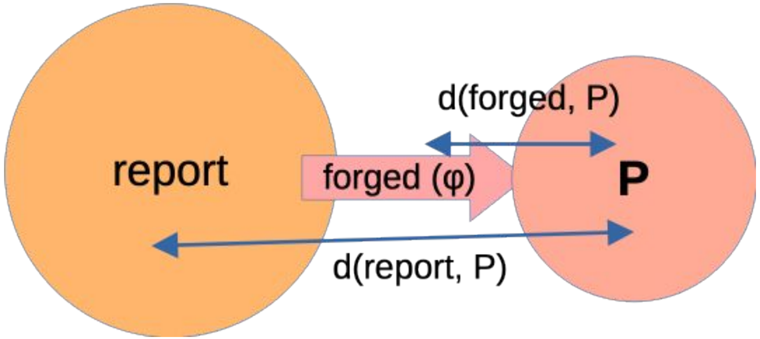
We expect a Canadian writer to have more in common with a Canadian surgeon than a skillful writer has with a skillful surgeon.

Requires adjective-noun phrases that share the same intersective adjective to be closer to each other than phrases with non-intersective ones.

Consistency Tests

- Testing non-subsectivity:

$$d(forged, P) \leq d(report, P)$$



Subsective composition guarantees P ⊆ [noun], whereas non-subsective composition does not.
→ embedding of P is closer to [noun] when the adjective is subsective.

Requires the adjective to “pull” the embedding of the whole phrase closer to them than the associated noun.

Experimental Setup

- Data: a collection of adjectives categorised by *Morzycki* (2016) and *Pavlick and Callison-Burch* (2016), augmented by a synonym for each instance, totalling 122 adjectives and 12 nouns.

Adjective Type	Set-Theoretic Definition	Examples	# of Adjectives
Subsective (Intersective)	$AN \subseteq N$ and $AN \subseteq A$	Red, Wild	22
Subsective (Non-Intersective)	$AN \subseteq N$ and $AN \not\subseteq A$	Skilful, Rare	12
Non-Subsective (Plain)	$AN \not\subseteq N$ and $AN \cap N \neq \emptyset$	Alleged, Disputed	54
Non-Subsective (Privative)	$AN \cap N = \emptyset$	Fake, Imaginary	28
Ambiguous	Contextually, one of the above	Old, Big	6

Experimental Setup

- Data: a collection of adjectives categorised by *Morzycki* (2016) and *Pavlick and Callison-Burch* (2016), augmented by a synonym for each instance, totalling 122 adjectives and 12 nouns.

- Phrases were generated by using a regular language defined by the expression *(adj)* + *noun*, where *adj* and *noun* are taken from the lists of adjectives and nouns respectively.

- The final dataset contains 44652 phrases.

Experimental Setup

- Models:
  - DPR (Karpukhin et al., 2020)
  - LaBSE (Feng et al., 2022)
  - Specter (Cohan et al., 2020)
  - OpenAI's text-embeddings-3-small [TE3-small] (OpenAI, 2024)
  - NV-Embed-v2 (Lee et al., 2024) [Ranked #1 in MTEB, Oct 2024]
  - Stella[en_1.5B_v5] ([@HuggingFace], 2024) [MTEB #3, Oct 2024]
  - Word2Vec (Mikolov et al., 2013)
  - Glove (Pennington et al., 2014)

CLS hidden state pooling

Closed-source

Specialised attention model

Non-contextual baselines

Results

Intersectivity experiment (single phrase)

Models	Adjective Type				
	S-I	S-NI	NS-Pl	NS-Pr	A
DPR	0.86	0.90	0.85	0.89	0.97
LaBSE	1.0	1.0	1.0	1.0	1.0
Specter	0.93	0.99	0.97	0.93	0.97
TE3-small	1.0	1.0	1.0	1.0	1.0
NV-Embed-v2	0.73	0.67	0.8	0.85	0.75
stella_en_1.5B_v5	1.0	1.0	1.0	1.0	1.0
Glove	1.0	1.0	1.0	1.0	1.0
Word2Vec	1.0	1.0	1.0	1.0	1.0

(two adjectives)

Models	Adjective Type Pair				
	(S-I, S-I)	(S-NI, S-I)	(NS-Pl, S-I)	(NS-Pr, S-I)	(A, S-I)
DPR	0.52	0.43	0.53	0.52	0.62
LaBSE	0.92	0.93	0.95	0.91	0.97
Specter	0.67	0.73	0.72	0.67	0.73
TE3-small	1.0	1.0	1.0	1.0	1.0
NV-Embed-v2	0.78	0.71	0.68	0.81	0.75
stella_en_1.5B_v5	1.0	1.0	1.0	1.0	1.0
Glove	1.0	1.0	1.0	0.94	1.0
Word2Vec	1.0	1.0	0.97	0.94	1.0

Notation: Ambiguous (A), Subsective-Intersective (S-I), Subsective Non-Intersective (S-NI), Plain Non-Subsective (NS-Pl), Privative Non-Subsective (NS-Pr).

Results

Intersectivity experiment (single phrase)

- Models with mean-pooling equivalent composition are universally intersective (vice-versa).
  - ➢ LaBSE, TE3-small and Stella are mean-pooling equivalent.
- Models without mean-pooling equivalent composition do not consistently capture adjective intersectivity.
  - ➢ On DPR, Specter and NV-Embed-v2, dist. relations don't correspond to adj. categorisation.

(two adjectives)

Models	Adjective Type Pair				
	(S-I, S-I)	(S-NI, S-I)	(NS-Pl, S-I)	(NS-Pr, S-I)	(A, S-I)
DPR	0.52	0.43	0.53	0.52	0.62
LaBSE	0.92	0.93	0.95	0.91	0.97
Specter	0.67	0.73	0.72	0.67	0.73
TE3-small	1.0	1.0	1.0	1.0	1.0
NV-Embed-v2	0.78	0.71	0.68	0.81	0.75
stella_en_1.5B_v5	1.0	1.0	1.0	1.0	1.0
Glove	1.0	1.0	1.0	0.94	1.0
Word2Vec	1.0	1.0	0.97	0.94	1.0

Notation: Ambiguous (A), Subsective-Intersective (S-I), Subsective Non-Intersective (S-NI), Plain Non-Subsective (NS-Pl), Privative Non-Subsective (NS-Pr).

Results

Intersectivity experiment (phrase pairs)

Models	Adjective Type Pair				
	(S-I, S-I)	(S-I, S-NI)	(S-I, NS-Pl)	(S-I, NS-Pr)	(S-I, A)
DPR	0.50	0.32	0.34	0.50	0.42
LaBSE	0.50	0.42	0.34	0.53	0.33
Specter	0.50	0.65	0.55	0.50	0.57
TE3-small	0.50	0.51	0.48	0.48	0.82
NV-Embed-v2	0.50	0.54	0.51	0.51	0.82
stella_en_1.5B_v5	0.50	0.75	0.64	0.58	0.91
Glove	0.50	0.66	0.69	0.70	0.47
Word2Vec	0.50	0.75	0.65	0.49	1.0

Notation: Ambiguous (A), Subsective-Intersective (S-I), Subsective Non-Intersective (S-NI), Plain Non-Subsective (NS-Pl), Privative Non-Subsective (NS-Pr).

- Each model places intersective emphasis in a different category of adjectives.
- Stella and the non-contextual baselines most closely approach the linguistically expected behaviour

Results

Non-subsectivity experiment

Models	Adjective Type				
	S-I	S-NI	NS-Pl	NS-Pr	A
DPR	0.46	0.37	0.48	0.54	0.39
LaBSE	0.36	0.31	0.51	0.33	0.19
Specter	0.48	0.31	0.49	0.57	0.33
TE3-small	0.81	0.75	0.74	0.77	0.39
NV-Embed-v2	0.84	0.79	0.79	0.83	0.81
stella_en_1.5B_v5	0.81	0.56	0.58	0.64	0.33
Glove	0.61	0.22	0.22	0.32	0.28
Word2Vec	0.55	0.21	0.34	0.49	0.0

Notation: Ambiguous (A), Subsective-Intersective (S-I), Subsective Non-Intersective (S-NI), Plain Non-Subsective (NS-Pl), Privative Non-Subsective (NS-Pr).

- None of the tested models behave according to the expectations given by the subsectivity formalism.
  - ➢ No significant differentiation for ‘NS’ categories.
- Larger models composition process largely emphasises adjectives instead of nouns.
  - ➢ Numerical behaviour hints at whether the model is more likely to choose intersective or non-intersective sense of ambiguous adjectives (e.g., “old”).

Conclusion

- Results indicate that current neural language models do not behave consistently according to expected behavior from the formalisms, w.r.t. intersective and subsective properties.
  - Models may not be capable of capturing the evaluated semantic properties of language.
  - Linguistic theories from Montagovian tradition are not matching the expected capabilities of distributional models.
- The proposed methodology is intended to be a stepping stone which can pave the way to a better understanding of LLMs latent spaces.
  - Other compositional properties to explore.
  - Linguistic properties need to be connected to NLP downstream task performance: Alignment of compositional semantics between inputs and expected outputs.

COLING 2025 • Abu Dhabi

The 31st International Conference on Computational Linguistics

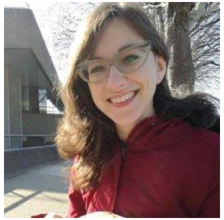
EnnCore

Neuro-symbolic AI Lab

Montague semantics and modifier consistency measurement in neural language models

Danilo Carvalho, Edoardo Manino, Julia Rozanova, Lucas Cordeiro, Andre Freitas





danilo.carvalho@manchester.ac.uk

MANCHESTER 1824 The University of Manchester

idiap RESEARCH INSTITUTE

CANCER RESEARCH UK | Manchester Institute

Consistency Tests

- Testing intersectivity (single phrase):

$$I_{m,p} \equiv d(emb_m(p), emb_m(t_i)) \leq d(emb_m(t_j), emb_m(t_k)) \qquad \forall i, j, k;\ \ j < k$$

$$E_{m,L}\{I_{m,p} = \top\}, \quad p \sim L$$

Requires that the embedding of an adjective-noun phrase lies closer to each term than the distance between any pair of terms.

- Testing intersectivity (phrase pairs):

$$II_{m,\{p\}} = d(emb_m(p_{a_1 n_1}), emb_m(p_{a_1 n_2})) \leq d(emb_m(p_{a_2 n_1}), emb_m(p_{a_2 n_2}))$$

$$E_{m,L^2}\{II_{m,\{p\}} = \top\}, \quad \{p\} \sim L^2$$

Requires adjective-noun phrases that share the same intersective adjective to be closer to each other than phrases with non-intersective ones.

Example: $d(Canadian\, writer, Canadian\, surgeon) \leq d(skillful\, writer, skillful\, surgeon)$

We expect a *Canadian writer* to have more in common with a *Canadian surgeon* than a *skillful writer* has with a *skillful surgeon*.

Consistency Tests

- Testing intersectivity (single phrase):

$$I_{m,p} \equiv d(emb_m(p), emb_m(t_i)) \leq d(emb_m(t_j), emb_m(t_k)) \qquad \forall i, j, k;\ \ j < k$$

$$E_{m,L}\{I_{m,p} = \top\}, \quad p \sim L$$

Requires that the embedding of an adjective-noun phrase lies closer to each term than the distance between any pair of terms.

- Testing intersectivity (phrase pairs):

$$II_{m,\{p\}} = d(emb_m(p_{a_1 n_1}), emb_m(p_{a_1 n_2})) \leq d(emb_m(p_{a_2 n_1}), emb_m(p_{a_2 n_2}))$$

$$E_{m,L^2}\{II_{m,\{p\}} = \top\}, \quad \{p\} \sim L^2$$

Requires adjective-noun phrases that share the same intersective adjective to be closer to each other than phrases with non-intersective ones.

- Testing non-subsectivity:

$$NI_{m,p} = d(emb_m(p), emb_m(a)) \leq d(emb_m(p), emb_m(n))$$

$$E_{m,L}\{NI_{m,p} = \top\}, \quad p \sim L$$

Requires the adjective to “pull” the embedding of the whole phrase closer to them than the associated noun.