Baroni & Zamparelli

Nouns are vectors, adjectives are matrices

Representing adjective-noun constructions in semantic space

Swantje Tönnis & Nadine Theiler

Computational Semantics and Pragmatics
7 December 2012

Overview

Formal semantics vs distributional semantics

- Compositionality
- Treatment of adjectives

Sanne: Guevara's model

Adjectives as linear maps

- Idea: using co-occurrence information
- Implementation and experimental setup

Evaluation

- Predicting adjective noun vectors
- Comparing adjectives

Formal semantics vs distributional semantics

The two frameworks have *opposing* strengths and weaknesses:

Formal semantics vs distributional semantics

The two frameworks have *opposing* strengths and weaknesses:

Distributional Semantics

+ meaning as an abstraction over distributional information

Formal semantics vs distributional semantics

The two frameworks have *opposing* strengths and weaknesses:

Distributional Semantics

+ meaning as an abstraction over distributional information

Formal Semantics

− purely extensional notion of meaning

Formal semantics vs distributional semantics

The two frameworks have *opposing* strengths and weaknesses:

Distributional Semantics

+ meaning as an abstraction over distributional information

− not compositional

Formal Semantics

− purely extensional notion of meaning

Formal semantics vs distributional semantics

The two frameworks have *opposing* strengths and weaknesses:

Distributional Semantics

+ meaning as an abstraction over distributional information
− not compositional

Formal Semantics

− purely extensional notion of meaning
+ compositional

Formal semantics vs distributional semantics

The two frameworks have *opposing* strengths and weaknesses:

Distributional Semantics

+ meaning as an abstraction over distributional information
− not compositional

Formal Semantics

− purely extensional notion of meaning
+ compositional

Compositionality?

The meaning of a complex expression is determined by the meanings of its constituents and its syntactic structure.

Formal semantics vs distributional semantics

The two frameworks have *opposing* strengths and weaknesses:

Distributional Semantics

+ meaning as an abstraction over distributional information
− not compositional
+ focuses on *content words*

Formal Semantics

− purely extensional notion of meaning
+ compositional

Compositionality?

The meaning of a complex expression is determined by the meanings of its constituents and its syntactic structure.

Formal semantics vs distributional semantics

The two frameworks have *opposing* strengths and weaknesses:

Distributional Semantics

+ meaning as an abstraction over distributional information
− not compositional
+ focuses on *content words*

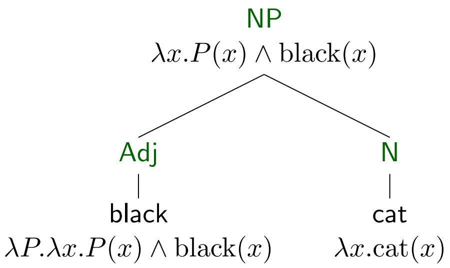
Formal Semantics

− purely extensional notion of meaning
+ compositional
+ focuses on *function words*

Compositionality?

The meaning of a complex expression is determined by the meanings of its constituents and its syntactic structure.

Formal semantics vs distributional semantics

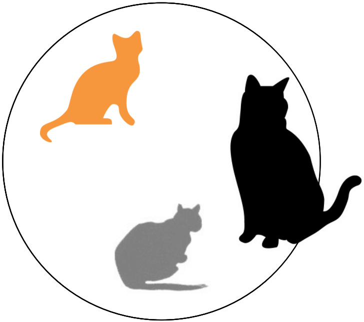


Compositionality?

The meaning of a complex expression is determined by the meanings of its constituents and its syntactic structure.

Formal semantics vs distributional semantics

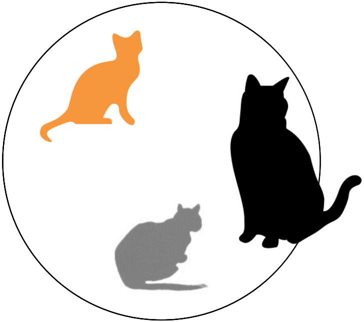
Treatment of adjectives in formal semantics



$\llbracket \text{cat} \rrbracket$

Formal semantics vs distributional semantics

Treatment of adjectives in formal semantics

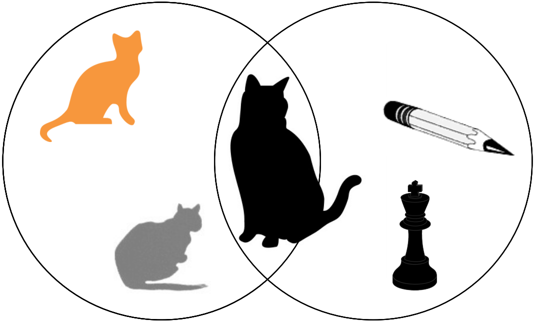


$[\![\text{cat}]\!]$

$[\![\text{black}]\!]$

Formal semantics vs distributional semantics

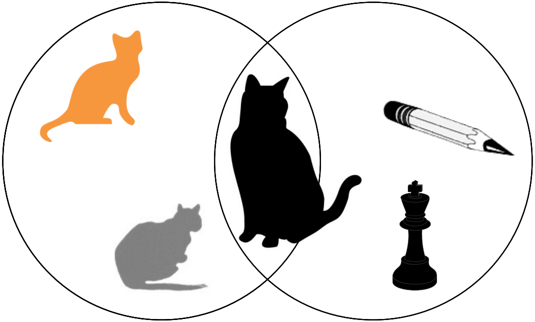
Treatment of adjectives in formal semantics



$$[\![\,\text{black cat}\,]\!] = [\![\,\text{black}\,]\!] \cap [\![\,\text{cat}\,]\!]\ ?$$

Formal semantics vs distributional semantics

Treatment of adjectives in formal semantics



$$[\![\text{black cat}]\!] = [\![\text{black}]\!] \cap [\![\text{cat}]\!]\ ?$$

But what about *fake*, *large* — and ultimately even *black*?

Formal semantics vs distributional semantics

Adjectives as functions

Most adjectives are non-intersective.

- Account for their meaning variation by viewing them as a *function*!

$$[\![\, \mathrm{NP} \,]\!] = [\![\, \mathrm{A} \,]\!]([\![\, \mathrm{N} \,]\!])$$

Formal semantics vs distributional semantics

Adjectives as functions

Most adjectives are non-intersective.

- Account for their meaning variation by viewing them as a *function*!

$$[\![\mathrm{NP}]\!] = [\![\mathrm{A}]\!]([\![\mathrm{N}]\!])$$

- can be sensitive to the noun
- need not return a subset of $[\![\mathrm{N}]\!]$

Formal semantics vs distributional semantics

Adjectives as functions

Most adjectives are non-intersective.

- Account for their meaning variation by viewing them as a *function*!

$$[\![\,\mathrm{NP}\,]\!] = [\![\,\mathrm{A}\,]\!]([\![\,\mathrm{N}\,]\!])$$

- can be sensitive to the noun
- need not return a subset of $[\![\,\mathrm{N}\,]\!]$

But how to construct these functions?

Formal semantics vs distributional semantics

Adjectives as functions

Most adjectives are non-intersective.

- Account for their meaning variation by viewing them as a *function*!

$$[\![\mathrm{NP}]\!] = [\![\mathrm{A}]\!]([\![\mathrm{N}]\!])$$

- can be sensitive to the noun
- need not return a subset of $[\![\mathrm{N}]\!]$

But how to construct these functions?

Adjectives as linear maps

Adjectives are *matrices*

They are *endomorphic linear maps* in noun space:

$$\overrightarrow{AN} = A \cdot \overrightarrow{N}$$

Adjectives as linear maps

Adjectives are *matrices*

They are *endomorphic linear maps* in noun space:

$$\overrightarrow{AN} = A \cdot \overrightarrow{N}$$

- If we know
  1. the context vector of the AN-pair $\overrightarrow{AN}$
  2. the context vector of the noun $\overrightarrow{N}$,

  then we can estimate the adjective matrix A.

Adjectives as linear maps

Adjectives are *matrices*

They are *endomorphic linear maps* in noun space:

$$\overrightarrow{AN} = A \cdot \overrightarrow{N}$$

- If we know
  1. the context vector of the AN-pair $\overrightarrow{AN}$
  2. the context vector of the noun $\overrightarrow{N}$,

  then we can estimate the adjective matrix A.
- This estimation is done by *partial least square regression.*

Adjectives as linear maps

Adjectives are *matrices*

They are *endomorphic linear maps* in noun space:

$$\overrightarrow{AN} = A \cdot \overrightarrow{N}$$

- If we know
  1. the context vector of the AN-pair $\overrightarrow{AN}$
  2. the context vector of the noun $\overrightarrow{N}$,

  then we can estimate the adjective matrix A.
- This estimation is done by *partial least square regression.*

In contrast to Guevara (2010):

The A matrices are specific to a single adjective.

Adjectives as linear maps

Experimental setup

- 2.83 billion token corpus

Adjectives as linear maps

Experimental setup

- 2.83 billion token corpus
- test set: 26,440 attested AN pairs

Adjectives as linear maps

Experimental setup

- 2.83 billion token corpus
- test set: 26,440 attested AN pairs
- semantic space:
  - co-occurrence matrix with sentence-internal co-occurrence counts

Adjectives as linear maps

Experimental setup

- 2.83 billion token corpus
- test set: 26,440 attested AN pairs
- semantic space:
  - co-occurrence matrix with sentence-internal co-occurrence counts
  - raw counts transformed into *Local Mutual Information* scores

Adjectives as linear maps

Experimental setup

- 2.83 billion token corpus
- test set: 26,440 attested AN pairs
- semantic space:
    - co-occurrence matrix with sentence-internal co-occurrence counts
    - raw counts transformed into *Local Mutual Information* scores
    - dimensionality reduction by *Singular Value Decomposition*: 40,999 $\times$ 300 matrix

Adjectives as linear maps

Experimental setup

- 2.83 billion token corpus
- test set: 26,440 attested AN pairs
- semantic space:
  - co-occurrence matrix with sentence-internal co-occurrence counts
  - raw counts transformed into *Local Mutual Information* scores
  - dimensionality reduction by *Singular Value Decomposition*: $40{,}999 \times 300$ matrix
  - semantic space also populated by adjectives and nouns not included in the AN test set

Evaluation

Predicting adjective noun vectors

Intuitively,...

... we want the predicted $\overrightarrow{AN}$ vectors to approximate the observed ones as closely as possible.

Evaluation

Predicting adjective noun vectors

Intuitively,...

...we want the predicted $\overrightarrow{AN}$ vectors to approximate the observed ones as closely as possible.

- Evaluate the system based on this:
  1. compute cosine of the predicted $\overrightarrow{AN}$ vector with all of the 41K vectors populating the semantic space
  2. rank these vectors by the obtained cosine values
  3. for each of the 26K observed $\overrightarrow{AN}$ vectors, check its position in the ranking

Evaluation

Predicting adjective noun vectors

method	*25%*	*median*	*75%*
alm	17	170	$\geq$1K
add	27	257	$\geq$1K
noun	72	448	$\geq$1K
mult	279	$\geq$1K	$\geq$1K
slm	629	$\geq$1K	$\geq$1K
adj	$\geq$1K	$\geq$1K	$\geq$1K

Table 3: Quartile ranks of observed ANs in cosine-ranked lists of predicted AN neighbors.

Evaluation

Predicting adjective noun vectors

method	*25%*	*median*	*75%*
alm	17	170	$\geq$1K
add	27	257	$\geq$1K
noun	72	448	$\geq$1K
mult	279	$\geq$1K	$\geq$1K
slm	629	$\geq$1K	$\geq$1K
adj	$\geq$1K	$\geq$1K	$\geq$1K

Table 3: Quartile ranks of observed ANs in cosine-ranked lists of predicted AN neighbors.

However...

For 27% of the *alm*-predicted $\overrightarrow{AN}$ vectors, the observed $\overrightarrow{AN}$ vector is not in the top-1K neighbourset.

Evaluation

In more detail...

The best results were obtained for high frequent adjectives:

new, great, American, large, different...

Evaluation

In more detail. . .

The best results were obtained for high frequent adjectives:

new, great, American, large, different. . .

- *new, large, different*:
  highly polysemous, bordering on function words!
- Can the model capture the polysemous nature of adjectives?
- Ideally, adjective meanings would arise only in combination with the noun they modify. Recall Pustejovsky's Generative Lexicon!

Evaluation

Dealing with polysemy

- Hope: certain weights affect only certain features

Evaluation

Dealing with polysemy

- Hope: certain weights affect only certain features
- Example: *green* could map concrete features to colour dimensions and abstract features to political dimensions

$$\underbrace{\begin{pmatrix} \omega_{\alpha 11} & \omega_{\alpha 12} & \omega_{\beta 13} & \omega_{\beta 14} & \omega_{\beta 15} \\ \omega_{\alpha 21} & \omega_{\alpha 22} & \omega_{\beta 23} & \omega_{\beta 24} & \omega_{\beta 25} \\ \omega_{\alpha 31} & \omega_{\alpha 32} & \omega_{\beta 33} & \omega_{\beta 34} & \omega_{\beta 35} \\ \omega_{\alpha 41} & \omega_{\alpha 42} & \omega_{\beta 43} & \omega_{\beta 44} & \omega_{\beta 45} \\ \omega_{\alpha 51} & \omega_{\alpha 52} & \omega_{\beta 53} & \omega_{\beta 54} & \omega_{\beta 55} \end{pmatrix}}_{\overrightarrow{\text{green}}} \quad \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{\overrightarrow{\text{chair}}} \quad \underbrace{\begin{pmatrix} 0 \\ 0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\overrightarrow{\text{initiative}}}$$

Evaluation

Dealing with polysemy

- Hope: certain weights affect only certain features
- Example: *green* could map concrete features to colour dimensions and abstract features to political dimensions

$$
\underset{\overrightarrow{\text{green}}}{\begin{pmatrix}
\omega_{\alpha 11} & \omega_{\alpha 12} & \omega_{\beta 13} & \omega_{\beta 14} & \omega_{\beta 15} \\
\omega_{\alpha 21} & \omega_{\alpha 22} & \omega_{\beta 23} & \omega_{\beta 24} & \omega_{\beta 25} \\
\omega_{\alpha 31} & \omega_{\alpha 32} & \omega_{\beta 33} & \omega_{\beta 34} & \omega_{\beta 35} \\
\omega_{\alpha 41} & \omega_{\alpha 42} & \omega_{\beta 43} & \omega_{\beta 44} & \omega_{\beta 45} \\
\omega_{\alpha 51} & \omega_{\alpha 52} & \omega_{\beta 53} & \omega_{\beta 54} & \omega_{\beta 55}
\end{pmatrix}}
\qquad
\underset{\overrightarrow{\text{chair}}}{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ 0 \\ 0 \\ 0 \end{pmatrix}}
\qquad
\underset{\overrightarrow{\text{initiative}}}{\begin{pmatrix} 0 \\ 0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}
$$

Evaluation

Dealing with polysemy

- Hope: certain weights affect only certain features
- Example: *green* could map concrete features to colour dimensions and abstract features to political dimensions

$$\underbrace{\begin{pmatrix} \omega_{\alpha 11} & \omega_{\alpha 12} & \omega_{\beta 13} & \omega_{\beta 14} & \omega_{\beta 15} \\ \omega_{\alpha 21} & \omega_{\alpha 22} & \omega_{\beta 23} & \omega_{\beta 24} & \omega_{\beta 25} \\ \omega_{\alpha 31} & \omega_{\alpha 32} & \omega_{\beta 33} & \omega_{\beta 34} & \omega_{\beta 35} \\ \omega_{\alpha 41} & \omega_{\alpha 42} & \omega_{\beta 43} & \omega_{\beta 44} & \omega_{\beta 45} \\ \omega_{\alpha 51} & \omega_{\alpha 52} & \omega_{\beta 53} & \omega_{\beta 54} & \omega_{\beta 55} \end{pmatrix}}_{\overrightarrow{\text{green}}} \qquad \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{\overrightarrow{\text{chair}}} \qquad \underbrace{\begin{pmatrix} 0 \\ 0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\overrightarrow{\text{initiative}}}$$

Evaluation

Problematic cases...

- often attributable to anomalous observed $\overrightarrow{AN}$ vectors
- model is worse at approximating the $\overrightarrow{AN}$ vectors of rare adjectives

SIMILAR			DISSIMILAR		
adj N	*obs. neighbor*	*pred. neighbor*	*adj N*	*obs. neighbor*	*pred. neighbor*
common understanding	common approach	common vision	American affair	Am. development	Am. policy
different authority	diff. objective	diff. description	current dimension	left (a)	current element
different partner	diff. organisation	diff. department	good complaint	current complaint	good beginning
general question	general issue	*same*	great field	excellent field	gr. distribution
historical introduction	hist. background	*same*	historical thing	different today	hist. reality
necessary qualification	nec. experience	*same*	important summer	summer	big holiday
new actor	new cast	*same*	large pass	historical region	large dimension
recent request	recent enquiry	*same*	special something	little animal	special thing
small drop	droplet	drop	white profile	chrome (n)	white show
young engineer	young designer	y. engineering	young photo	important song	young image

Table 4: Left: nearest neighbors of observed and *alm*-predicted ANs (excluding each other) for a random set of ANs where rank of observed w.r.t. predicted is 1. Right: nearest neighbors of predicted and observed ANs for random set where rank of observed w.r.t. predicted is $\geq 1K$.

Evaluation

Comparing adjectives

Since adjectives are no longer represented as vectors — how can we still compare them meaningfully?

Evaluation

Comparing adjectives

Since adjectives are no longer represented as vectors — how can we still compare them meaningfully?

Two methods:

1. represent adjective by the centroid of all $\overrightarrow{AN}$ vectors containing the adjective

   American adult, American menu... $\rightsquigarrow$ $\boxed{\textit{American}\ \text{N centroid}}$

2. unfold 300×300 matrix into 90K-dimensional vector

Evaluation

Comparing adjectives

Does this capture semantic similarity?

- Clustering adjectives:

white	nice	recent	big
black	excellent	new	huge
red	important	current	little
green	major	old	small
	appropriate	young	large

Evaluation

Comparing adjectives

Does this capture semantic similarity?

- Clustering adjectives:

white	nice	recent	big
black	excellent	new	huge
red	important	current	little
green	major	old	small
	appropriate	young	large

- Results:

input	*purity*
matrix	73.7 (68.4-94.7)
centroid	73.7 (63.2-94.7)
vector	68.4 (63.2-89.5)
random	45.9 (36.8-57.9)

Table 5: Percentage purity in adjective clustering with bootstrapped 95% confidence intervals.

Conclusion

- adjectives representable as matrices
- in line with their formal semantics treatment as functions
- learnable from co-occurrence data of adjective-noun pairs
- reliable predictions for adjective-noun vectors
- adjectives still comparable with regard to semantic similarity

Discussion / Open questions

- Can we really use centroids to represent polysemous adjectives?
- Is the model limited to attributive adjectives, or can it also be applied to predicative constructions?
- Baroni and Zamparelli claim that the model can naturally deal with recursion. They do not explicitly test this, though. So, can it?

References

- Marco Baroni & Roberto Zamparelli (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. pp. 1183–1193. ACL.
- Emiliano Raúl Guevara (2010): A regression model of adjective-noun compositionality in distributional semantics. In: *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*. pp. 33–37. ACL.