Distributional Compositionality
Compositionality in DS

Raffaella Bernardi

University of Trento

February 14, 2012

Acknowledgments

Credits: Some of the slides of today lecture are based on an earlier DS courses taught by Marco Baroni.

Distributional Semantics

Recall

The main questions have been:

1. What is the sense of a given *word*?
2. How can it be induced and represented?
3. How do we relate word senses (synonyms, antonyms, hyperonym etc.)?

Well established answers:

1. The sense of a word can be given by its use, viz. by the *contexts* in which it occurs;
2. It can be induced from (either raw or parsed) corpora and can be represented by *vectors*.
3. *Cosine similarity* captures synonyms (as well as other semantic relations).

From Formal to Distributional Semantics

New research questions in DS

1. Do all words live in the same space?
2. What about compositionality of word sense?
3. How do we "infer" some piece of information out of another?

From Formal Semantics to Distributional Semantics

Recent results in DS

1. From one space to multiple spaces, and from only vectors to vectors and matrices.
2. Several Compositional DS models have been tested so far.
3. New "similarity measures" have been defined to capture lexical entailment and tested on phrasal entailment too.

Multiple semantics spaces

Phrases

All the expressions of the same syntactic category live in the same semantic space.
For instance, ADJ N ("special collection") live in the same space of N ("archives").

important route	*nice girl*	*little war*
important transport	good girl	great war
important road	big girl	major war
major road	guy	small war
red cover	*special collection*	*young husband*
black cover	general collection	small son
hardback	small collection	small daughter
red label	archives	mistress

Multiple semantics spaces

Problem of one semantic space model

	and	of	the	valley	moon
planet	$> 1K$	$> 1K$	$> 1K$	20.3	24.3
night	$> 1K$	$> 1K$	$> 1K$	10.3	15.2
space	$> 1K$	$> 1K$	$> 1K$	11.1	20.1

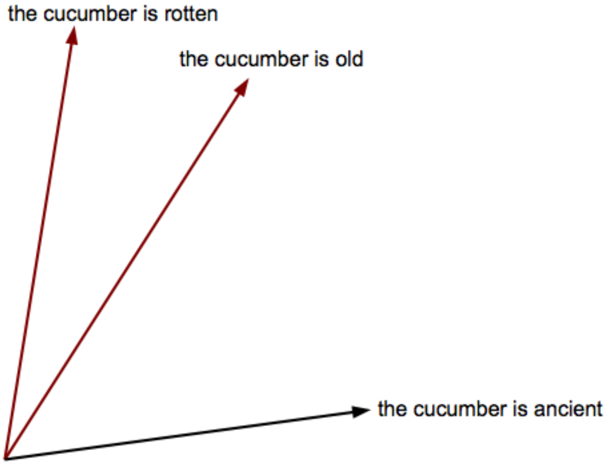
"and", "of", "the" have similar distribution but a very different meaning:

"the valley of the moon" vs. "the valley and the moon"

the semantic space of these words must be different from those of eg. nouns ("valley', "moon").

Compositionality in DS: Expectation

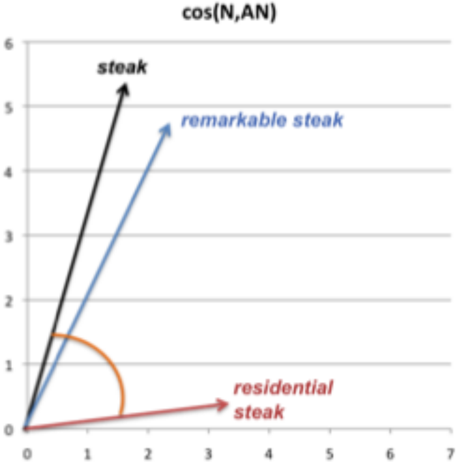
Disambiguation



Mitchell and Lapata 2008, Erk and Padó 2008

Compositionality in DS: Expectation

Semantic deviance

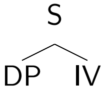


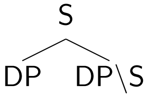
Compositionality in Formal Semantics

Verbs

Recall:

- an intransitive verb is a set entities, hence it's a one argument function. $\lambda x.\texttt{walk}(x)$;
- transitive verb: set of pairs of entities, hence it's a two argument function: $\lambda y.\lambda x.\texttt{teases}(y,x)$.





The function "walk" selects a subset of D_e.

Compositionality in Formal Semantics

Adjectives

Syntax:

$$[_{N}\ \mathrm{ADJ}\ \ \mathrm{N}] \qquad [_{N}\ \mathrm{N/N}\ \ \mathrm{N}]$$

ADJ is a function that modifies a noun:

$$(\lambda Y.\lambda x.\mathtt{Red}(x) \wedge Y(x))(\mathtt{Moon}) \rightsquigarrow \lambda x.\mathtt{Red}(x) \wedge \mathtt{Moon}(x)$$

$$[\![\mathrm{Red}]\!] \cap [\![\mathrm{Moon}]\!]$$

Compositionality: DP IV

Kintsch (2001)

Kintsch (2001): The meaning of a predicate varies depending on the argument it operates upon:

The horse run vs. the color run

Hence, take "gallop" and "dissolve" as landmarks of the semantic space,

- "the horse run" should be closer to "gallop" than to "dissolve".
- "the color run" should be closer to "dissolve" than to "gallop"

(or put it differently, the verb acts differently on different nouns.)

Compositionality: ADJ N

Pustejovsky (1995)

- red Ferrari [the outside]
- red watermelon [the inside]
- red traffic light [only the signal]
- ..

Similarly, “red” will reinforce the concrete dimensions of a concrete noun and the abstract ones of an abstract noun.

Compositionality in DS

Different Models

	horse	run	horse $+$ run	horse $\odot$ run	run(horse)
gallop	15.3	24.3	39.6	371.8	24.6
jump	3.7	15.2	18. 9	56.2	19.3
dissolve	2.2	20.1	22.3	44.2	12.4

- Additive and/or Multiplicative Models: Mitchell & Lapata (2008), Guevara (2010)
- Function application: Baroni & Zamparelli (2010), Grefenstette & Sadrzadeh (2011)
- For others, see Mitchell and Lapata (2010) overview.

Compositionality as vectors composition

Mitchell and Lapata (2008,2010): Class of Models

General class of models:

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

- $\vec{p}$ can be in a different space than $\vec{u}$ and $\vec{v}$.
- K is background knowledge
- R syntactic relation.

Putting constraints will provide us with different models.

Compositionality as vectors composition

Mitchell and Lapata (2008,2010): Constraints on the models

1. Not only the ith components of $\vec{u}$ and $\vec{v}$ contribute to the ith component of $\vec{p}$. Circular convolution:
$$p_i = \Sigma_j u_j \cdot v_{i-j}$$
2. Role of K, e.g. consider the argument's distributional neighbours Kitsch 2001:
$$\vec{p} = \vec{u} + \vec{v} + \Sigma \vec{n}$$
3. Asymmetry weights pred and arg differently:
$$p_i = \alpha u_i + \beta v_i$$
4. the ith component of $\vec{u}$ should be scaled according to its relevance to $\vec{v}$ and vice versa. multiplicative model
$$p_i = u_i \cdot v_i$$

Compositionality: DP IV

Mitchell and Lapata (2008,2010): Evaluation data set

- 120 experimental items consisting of 15 reference verbs each coupled with 4 nouns and 2 (high- and low-similarity) landmarks
- Similarity of sentence with reference vs. landmark rated by 49 subjects on 1-7 scale

	Noun	Reference	High	Low
The	fire	glowed	burned	beamed
The	face	glowed	beamed	burned
The	child	strayed	roamed	digressed
The	discussion	strayed	digressed	roamed
The	sales	slumped	declined	slouched
The	shoulders	slumped	slouched	declined

Table 1: Example Stimuli with High and Low similarity landmarks

Compositionality: DP IV

Mitchell and Lapata (2008,2010): Evaluation results

Models vs. Human judgment: different ranging scale.

Additive model, Non compositional baseline, weighted additive and Kintsch (2001) don't distinguish between High (close) and Low (far) landmarks.

Multiplicative and combined models are closed to human ratings. The former does not require parameter optimization.

Model	High	Low	ρ
NonComp	0.27	0.26	0.08
Add	0.59	0.59	0.04
Weight Add	0.35	0.34	0.09
Kintsch	0.47	0.45	0.09
Multiply	0.42	0.28	0.17
Combined	0.38	0.28	0.19
Human Judg	4.94	3.25	0.40

See also Grefenstette and Sadrzadeh (2011)

Compositionality as vector combination: problems

Grammatical words: highly frequent

	planet	night	space	color	blood	brown
the	$>$1K	$>$1K	$>$1K	$>$1K	$>$1K	$>$1K
moon	24.3	15.2	20.1	3.0	1.2	0.5
the moon	??	??	??	??	??	??

Composition as vector combination: problems

Grammatical words variation

	car	train	theater	person	movie	ticket
few	$>$1K	$>$1K	$>$1K	$>$1K	$>$1K	$>$1K
a few	$>$1K	$>$1K	$>$1K	$>$1K	$>$1K	$>$1K
seats	24.3	15.2	20.1	3.0	1.2	0.5
few seats	??	??	??	??	??	??
a few seats	??	??	??	??	??	??

- There are few seats available. negative: hurry up!
- There are a few seats available. positive: take your time!

Compositionality in DS: Function application

Baroni and Zamparelli (2010)

Distributional Semantics (e.g. 2 dimensional space):

N/N: matrix

red	d1	d2
d1	$n1$	$n2$
d2	$m1$	$m2$

N: vector

	moon
d1	$k1$
d2	$k2$

Function app. by the matrix product and returns a vector:
$red(\overrightarrow{moon}) = \sum_{i=1}^{n} red_i \ moon_i$

N: vector

	red moon
d1	$(n1, n1) \cdot (k1, k2)$
d2	$(m1, m2) \cdot (k1, k2)$

=

N: vector

	red moon
d1	$(n1k1) + (n2k2)$
d2	$(m1k1) + (m2k2)$

Compositionality in DS: Function application

Learning methods

- Vectors are induced from the corpus by a lexical association co-frequency function. [Well established]
- Matrices are learned by regression (Baroni & Zamparelli (2010)). E.g.:

  "red" is learned, using linear regression, from the pairs (N, red-N).

n and the moon shining i	f a large red moon , Campana
with the moon shining s	, a blood red moon hung over
rainbowed moon . And the	glorious red moon turning t
crescent moon , thrille	The round red moon , she 's
in a blue moon only , wi	l a blood red moon emerged f
now , the moon has risen	n rains , red moon blows , w
d now the moon rises , f	monstrous red moon had climb
y at full moon , get up	. A very red moon rising is
crescent moon . Mr Angu	under the red moon a vampire
...	...

Compositionality in DS: Function application

Learning matrices

red (R) is a matrix whose values are unknown (I use capitol letters for unknown):

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

We have harvested the vectors $\vec{moon}$ and $\vec{army}$ representing "moon" and "army", resp. and the vectors $\vec{n_1} = (n_{11}, n_{12})$ and $\vec{n_2} = (n_{21}, n_{22})$ representing "red moon", "red army". Since we know that e.g.

$$R\ \vec{moon} = \begin{bmatrix} R_{11}moon_1 + R_{12}moon_2 \\ R_{21}moon_1 + R_{22}moon_2 \end{bmatrix} = \begin{bmatrix} n_{11} \\ n_{12} \end{bmatrix} = \vec{n_1}$$

taking all the data together, we end up having to solve the following multiple regression problems to determine the R values (r_{11}, r_{12} etc.)

$$\begin{aligned} R_{11}moon_1 + R_{12}moon_2 &= n'_{11} \\ R_{11}army_1 + R_{12}army_2 &= n'_{21} \end{aligned}$$

$$\begin{aligned} R_{21}moon_1 + R_{22}moon_2 &= n'_{12} \\ R_{21}army_1 + R_{22}army_2 &= n'_{22} \end{aligned}$$

which are solved by assigning weights to the unknown (Baroni and Zamparelli (2010) have not used the intercept).

Compositionality in DS: ADJ N

Comparison Compositional DS models

Baroni & Zamparelli 2010 have

- trained separate models for each adjective;
- (a) composed the learned matrix (function) with a noun vector (argument) by matrix product ($\cdot$) – the adjective weight matrix with the noun vector value;
- composed adjectives with nouns using: (b) additive and (c) multiplicative model –starting from adjective and noun vectors;
- harvested vectors for "adjective-noun" from the corpus;
- compared (a) "learned_matrix $\cdot$ vector_noun" ("function application") vs. (b) "vector_adj $+$ vector_noun" vs. (c) "vector_adj $\odot$ vector_noun";
- shown that – among (a), (b), (c) – (a) gives results more similar to the "harvested vector_adj-noun" than the other two methods.

Compositionality in DS: ADJ N

Observed ADJ N vs. Composed ADJ(N): (a) when observed and composed are close

To double check the validity of the functional approach, the results of the matrix product has been compared to the vectors observed (induced) from the corpus:

adj N	*observed neighbor*	*predicted neighbor*
common understanding	common approach	common vision
different authority	different objective	different description
different partner	different organisation	different department
general question	general issue	general issue
historical introduction	historical background	historical background
necessary qualification	necessary experience	necessary experience
new actor	new cast	new case
recent request	recent enquiry	recent enquiry
small drop	droplet	drop
young engineer	young designer	young engineering

Compositionality in DS: ADJ N

Observed ADJ N vs. Composed ADJ(N): (b) when observed and composed are far

adj N	*observed neighbor*	*predicted neighbor*
American affair	American development	American policy
current dimension	left (a)	current element
good complaint	current complaint	good beginning
great field	excellent field	great distribution
historical thing	different today	historical reality
important summer	summer	big holiday
large pass	historical region	large dimension
special something	little animal	special thing
white profile	chrome (n)	white show
young photo	important song	young image

From Formal to Distributional Semantics

FS domains and DS spaces

- FS:
  - Atomic vs. functional types
  - Typed denotational domains
  - Correspondence between syntactic categories and semantic types
- Could we import these ideas in DS?
  - Vectors vs. matrices — Seems promising
  - Typed semantic spaces
  - Correspondence between syntactic categories and semantic types

Compositionality in DS: next steps

Summing up

- DS research has obtained satisfactory results on content words by evaluating them on different lexical semantic tasks.
- New research is "importing" in the DS framework some of the understanding achieved within the FS school.

To tackle compositionality in DS a better understanding of grammatical words should be reached.

FS starting point is logical entailment between propositions, hence it's based on the referential meaning of sentences ($D_t = \{0, 1\}$).

All domains are partially ordered, e.g.:

- $D_t = \{0, 1\}$ and $0 \leq_t 1$,
- $D_{e \to t} : \{student, person\}$,
  s.t. $[\![student]\!] = \{a, b\}$ and $[\![person]\!] = \{a, b, c\}$,
  by def: $[\![student]\!] \leq_{e \to t} [\![person]\!]$ since
  $\forall \alpha \in D_e\ [\![student]\!]([\![\alpha]\!]) \leq_t [\![person]\!]([\![\alpha]\!])$,

Entailment
Entailment in DS

- Lexical entailment: already some successful results.
- Phrase entailment: a few studies done.
- Sentential entailment: none.

Entailment

DS success on Lexical entailment

Cosine similarity has shown to be a valid measure for the synonymy relation, but it does not capture the “is-a” relation properly: it’s symmetric!
Kotlerman, Dagan, Szpektor and Zhitomirsky-Geffet 2010 see is-a relation as “feature inclusion” (where “features” are the space dimensions) and propose an asymmetric measure based on empirical harvested vectors. Intuition behind their measure:

1. Is-a score higher if included features are ranked high for the narrow term.
2. Is-a score higher if included features are ranked high in the broader term vector as well.
3. Is-a score is lower for short feature vectors.

Very positive results compared to WordNet-based measures.
They have focused on nouns.

Entailment

Entailment at phrasal level in DS

Baroni, Bernardi, Do and Shan (EACL 2012):

- Dagan et. al. measure
  - does generalize to expressions of the noun category, tested on N1 $\leq$ N2 and ADJ N1 $\leq$ N1.
  - does not generalize to expressions of other categories, tested on QPs.
- FS different partial order for different domains; DS different partial orders for different semantic spaces.

Entailment

SVM learned QP entailment

Quantifier pair	Correct
many $\models$ several	19%
many $\models$ some	86%
each $\models$ some	99%
most $\models$ many	18%
much $\models$ some	88%
every $\models$ many	87%
all $\models$ many	88%
all $\models$ most	85%
all $\models$ several	99%
all $\models$ some	99%
both $\models$ either	2%
both $\models$ some	56%
several $\models$ some	76%
Subtotal	*77%*

Quantifier pair	Correct
many $\not\models$ most	65%
many $\not\models$ no	52%
both $\not\models$ many	73%
both $\not\models$ most	94%
both $\not\models$ several	15%
either $\not\models$ both	62%
many $\not\models$ all	97%
many $\not\models$ every	98%
few $\not\models$ many	20%
few $\not\models$ all	97%
several $\not\models$ all	99%
some $\not\models$ many	49%
some $\not\models$ all	99%
some $\not\models$ each	98%
some $\not\models$ every	99%
several $\not\models$ every	99%
several $\not\models$ few	94%
Subtotal	*79%*

P: 77%, R: 77%, F: 77%, A: 78%**

Entailment

Partially ordered spaces

The results show that:

- DS models do contain information needed to detect the entailment relation among other categories too, e.g. QP.
- Not the same dimensions/not the same relations among dimensions are at work for different partial orders ($\leq_{QP}$ vs. $\leq_N$)

Questions: which are the dimensions involved in the entailment relation for the various categories? Can we hope to find an abstract definition based on atomic and function types as in FS?

Conclusions

Ideas imported from FS into DS

(a) Meaning flows from the words;
(b) "Complete" (vectors) vs. Incomplete words (matrices);
(c) Meaning representations are guided by the syntactic structure.
(d) Different partial order for different semantic spaces

Conclusions

What else?

(a) What's the meaning of grammatical work?

(b) What's the meaning of a sentence?

(c) What's the meaning of "entities", e.g., "John".

(d) Which is the DS representation corresponding to a higher order function, e.g. QP?

(e) What's the linear algebra operation corresponding to lambda abstraction – how can structure be de-composed in a DS representation (e.g. relative clauses)?

We are currently working on (a) and we will address some of these questions within the 5 year EU project: COMPOSES (`http://clic.cimec.unitn.it/composes/`).

References

- M. Baroni and R. Zamparelli (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. Proceedings of EMNLP
- M. Baroni, R. Bernardi, Q. Do, Ch. Shan (2012) Entailment above the word level in distributional semantics. Proceedings of EACL.
- E. Grefenstette and M. Sadrzadeh Experimenting with transitive verbs in a DisCoCat. GEMS 2011.
- E. Guevara (2010). A regression model of adjective-noun compositionality in in distributional semantics. Proceedings of GEMS.
- Kintsch Predication. (2001) Cognitive Science, 25(2): 173–202.
- J. Mitchell and M. Lapata (2008). Vector-based models of semantic composition. Proceedings of ACL.
- J. Mitchell and M. Lapata (2010). Composition in distributional models of semantics. Cognitive Science 34(8): 1388–1429