Will Distributional Semantics Ever Become Semantic?

Alessandro Lenci

University of Pisa

7^{th} International Global WordNet Conference
January 28, 2014 - Tartu

IN SUPREMÆ DIGNITATIS · 1343 ·

CoLing Lab
Computational Linguistics Laboratory

What is Distributional Semantics?

Distributional semantics is predicated on the assumption that linguistic units with certain semantic similarities also share certain similarities in the relevant environments.

If therefore relevant environments can be previously specified, it may be possible to group automatically all those linguistic units which occur in similarly definable environments, and it is assumed that these automatically produced groupings will be of semantic interest.

Paul Garvin, (1962), "Computer participation in linguistic research", *Language*, 38(4): 385-389

What is Distributional Semantics?

Distributional semantics is predicated on the assumption that linguistic units with certain semantic similarities also share certain similarities in the relevant environments.

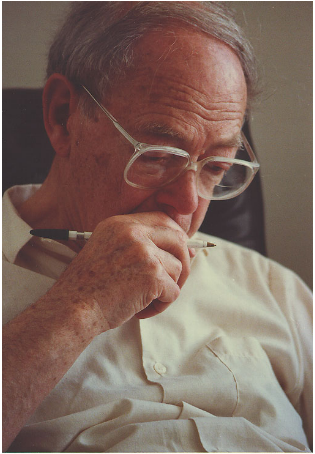
If therefore relevant environments can be previously specified, it may be possible to group automatically all those linguistic units which occur in similarly definable environments, and it is assumed that these automatically produced groupings will be of semantic interest.

Paul Garvin, (1962), “Computer participation in linguistic research”, *Language*, 38(4): 385-389

The Pioneers of Distributional Semantics

Distributionalism in linguistics

Zellig S. Harris



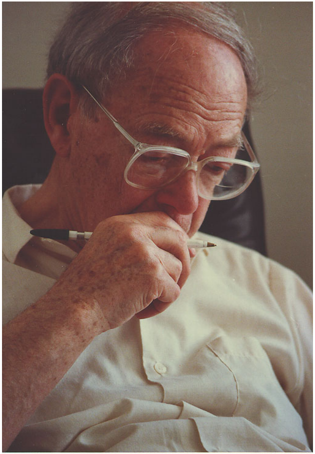
To be relevant [linguistic] elements must be set up on a distributional basis: x and y are included in the same element A if the distribution of x relative to the other elements B, C, etc. is in some sense the same as the distribution of y.

(Harris 1951: 7)

The Pioneers of Distributional Semantics

Distributionalism in linguistics

Zellig S. Harris



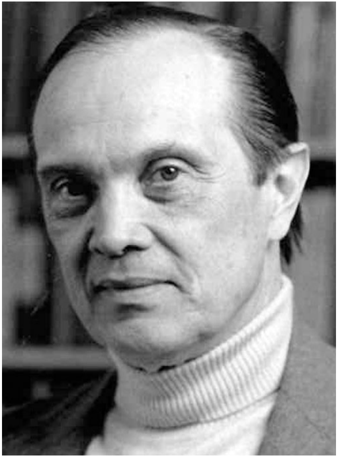
If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference of distribution.

(Harris 1954: 156)

The Pioneers of Distributional Semantics

Distributionalism in cognitive science

George A. Miller



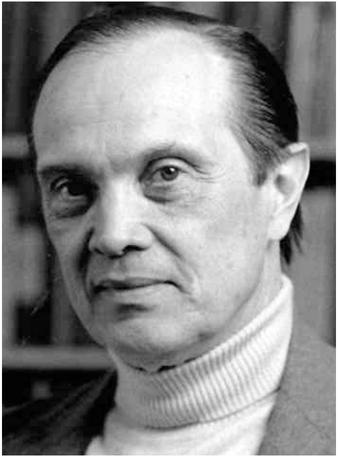
A linguist defines the distribution of a word as the list of contexts into which the word can be substituted; the distributional similarity of two words is thus the extent to which they can be substituted into the same contexts. [. . .] Several psychologist have invented or adapted variations on this distributional theme as an empirical method for investigating semantic similarities.

(Miller 1967: 572-573)

The Pioneers of Distributional Semantics

Distributionalism in cognitive science

George A. Miller



The contextual representation of a word is knowledge of how that word is used. [. . .] That is to say, a word's contextual representation [. . .] is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts. [. . .] Two words are semantically similar to the extent that their contextual representations are similar.

(Miller and Charles 1991: 5)

The Theoretical Foundations of Distributional Semantics

The Distributional Hypothesis

Lexemes with similar distributional properties have similar meanings

- Distributional semantic models (DSMs) are computational methods that turn the Distributional Hypothesis into an experimental framework for semantic analysis:
  - extract from corpora and count co-occurrences of lexical items with linguistic contexts
  - represent lexical items geometrically with distributional vectors built out of (a function of) their co-occurrence counts
  - measure semantic similarity with distributional vector similarity

The Theoretical Foundations of Distributional Semantics

The Distributional Hypothesis

Lexemes with similar distributional properties have similar meanings

- Distributional semantic models (DSMs) are computational methods that turn the Distributional Hypothesis into an experimental framework for semantic analysis:
  - extract from corpora and count co-occurrences of lexical items with linguistic contexts
  - represent lexical items geometrically with distributional vectors built out of (a function of) their co-occurrence counts
  - measure semantic similarity with distributional vector similarity

The Theoretical Foundations of Distributional Semantics

The Distributional Hypothesis

Lexemes with similar distributional properties have similar meanings

- Distributional semantic models (DSMs) are computational methods that turn the Distributional Hypothesis into an experimental framework for semantic analysis:
    - extract from corpora and count co-occurrences of lexical items with linguistic contexts
    - represent lexical items geometrically with distributional vectors built out of (a function of) their co-occurrence counts
    - measure semantic similarity with distributional vector similarity

The Theoretical Foundations of Distributional Semantics

The Distributional Hypothesis

Lexemes with similar distributional properties have similar meanings

- Distributional semantic models (DSMs) are computational methods that turn the Distributional Hypothesis into an experimental framework for semantic analysis:
  - extract from corpora and count co-occurrences of lexical items with linguistic contexts
  - represent lexical items geometrically with distributional vectors built out of (a function of) their co-occurrence counts
  - measure semantic similarity with distributional vector similarity

The Theoretical Foundations of Distributional Semantics

The Distributional Hypothesis

Lexemes with similar distributional properties have similar meanings

- Distributional semantic models (DSMs) are computational methods that turn the Distributional Hypothesis into an experimental framework for semantic analysis:
  - extract from corpora and count co-occurrences of lexical items with linguistic contexts
  - represent lexical items geometrically with distributional vectors built out of (a function of) their co-occurrence counts
  - measure semantic similarity with distributional vector similarity

From Contexts . . .

. . . dig a [hole. The	**car**	drove away] leaving behind . . .
. . . to directly [drive the	**car**	wheel angle] 3. Force . . .
. . . celebrity status, [drove fast	**cars**	and partied] with some . . .
. . . but there [are police	**cars**	that chase] you. Each . . .
. . . world of [money, fast	**cars**	and excitement] and, under . . .
. . . to pet [the family's	**cat**	and dog,] who tended . . .
. . . and then [wanted a	**cat**	to eat] the many . . .
. . . murmur is [detectable. The	**cat**	often eats] and drinks . . .
. . . behaviour of [a domestic	**cat**	playing with] a caught . . .
. . . have never [seen a	**cat**	eat so] little and . . .
. . . bank, children [playing with	**dogs**	and a] man leading. . . .
. . . sure you [encourage your	**dog**	to play] appropriate chase . . .
. . . Truth, Lord: [yet the	**dogs**	eat of] the crumbs . . .
. . . vegetable material [and enzymes.	**Dogs**	also eat] fruit, berries . . .
. . . hubby once [ate the	**dog**	food and] asked for . . .
. . . were back [at the	**van**	and drove] down to . . .
. . . go down [as the	**van**	drove off.] As he . . .
. . . heavy objects, [driving transit	**vans**	, wiring plugs] and talking . . .
. . . of the [fast food	**van**	being located] outside their . . .
. . . each of [the six	**van**	wheels , and] also under . . .

. . . to Distributional Vectors

$$
\begin{array}{l}
 \\
\textbf{car} \\
\textbf{cat} \\
\textbf{dog} \\
\textbf{van}
\end{array}
\begin{array}{c}
\begin{array}{cccccccc}
\textit{dog} & \textit{drive} & \textit{eat} & \textit{fast} & \textit{play} & \dots & \textit{the} & \textit{wheel}
\end{array} \\
\left(\begin{array}{cccccccc}
0 & 3 & 0 & 2 & 0 & \vdots & 2 & 1 \\
1 & 0 & 3 & 0 & 1 & \vdots & 2 & 0 \\
0 & 0 & 3 & 0 & 2 & \vdots & 2 & 0 \\
0 & 3 & 0 & 1 & 0 & \vdots & 3 & 1
\end{array}\right)
\end{array}
$$

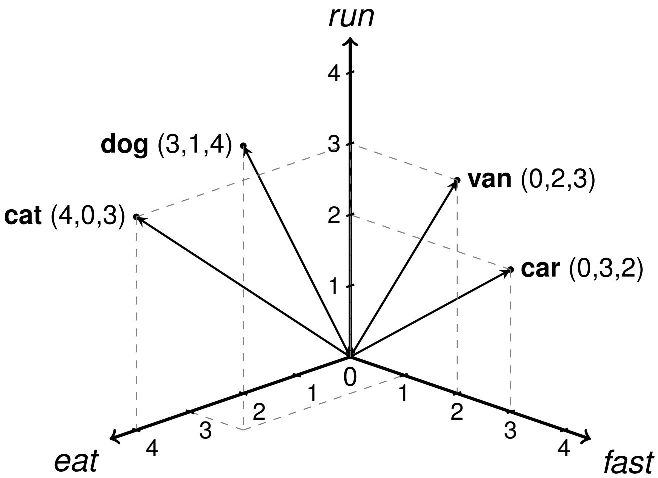
co-occurrence matrix

. . . to Distributional Vectors

$$
\begin{array}{l}
 \\
\textbf{car} \\
\textbf{cat} \\
\color{red}{\textbf{dog}} \\
\textbf{van}
\end{array}
\begin{array}{c}
\begin{array}{cccccccc}
\textit{dog} & \textit{drive} & \textit{eat} & \textit{fast} & \textit{play} & \dots & \textit{the} & \textit{wheel}
\end{array} \\
\begin{pmatrix}
0 & 3 & 0 & 2 & 0 & \vdots & 2 & 1 \\
1 & 0 & 3 & 0 & 1 & \vdots & 2 & 0 \\
\color{red}{\mathbf{0}} & \color{red}{\mathbf{0}} & \color{red}{\mathbf{3}} & \color{red}{\mathbf{0}} & \color{red}{\mathbf{2}} & \color{red}{\vdots} & \color{red}{\mathbf{2}} & \color{red}{\mathbf{0}} \\
0 & 3 & 0 & 1 & 0 & \vdots & 3 & 1
\end{pmatrix}
\end{array}
$$

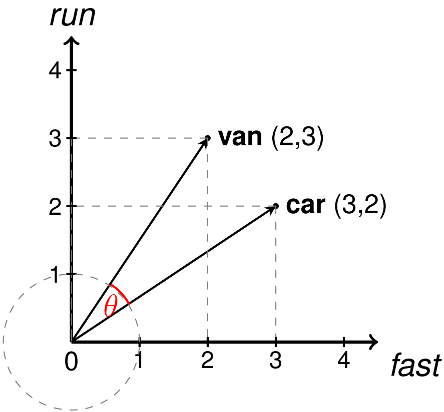
co-occurrence matrix

. . . to Distributional Vectors



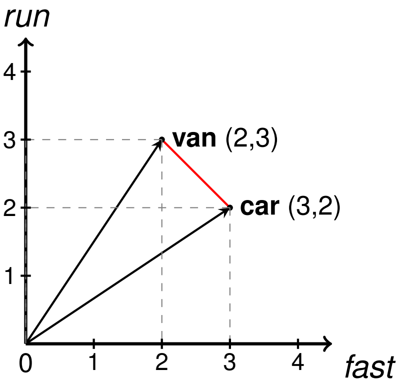
Measuring Vector Similarity

Cosine $$\frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$



Measuring Vector Similarity

Euclidean distance $\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$



From Vector to Semantic Similarity

- The Distributional Hypothesis predicts that words with similar distributional vectors are semantically similar

	car	*cat*	*dog*	*van*
car	1			
cat	0.33	1		
dog	0.60	0.94	1	
van	0.92	0.50	0.76	1

cosine similarities

From Vector to Semantic Similarity

- The Distributional Hypothesis predicts that words with similar distributional vectors are semantically similar

	car	*cat*	*dog*	*van*
car	1			
cat	0.33	1		
dog	0.60	0.94	1	
van	0.92	0.50	0.76	1

cosine similarities

Types of DSMs

- Linguistic contexts
  - text regions, linear (window-based) collocates, syntactic collocates, etc.
- Context weighting
  - raw co-occurrence frequency, entropy, tf-idf, association measures, etc.
- Distributional vector construction
  - matrix models, topic models, Random Indexing, neural embeddings, etc.
- Vector similarity measure
  - cosine, euclidean distance, LIN measure, etc.

Types of DSMs

- Linguistic contexts
  - text regions, linear (window-based) collocates, syntactic collocates, etc.
- Context weighting
  - raw co-occurrence frequency, entropy, tf-idf, association measures, etc.
- Distributional vector construction
  - matrix models, topic models, Random Indexing, neural embeddings, etc.
- Vector similarity measure
  - cosine, euclidean distance, LIN measure, etc.

Types of DSMs

- Linguistic contexts
  - text regions, linear (window-based) collocates, syntactic collocates, etc.
- Context weighting
  - raw co-occurrence frequency, entropy, tf-idf, association measures, etc.
- Distributional vector construction
  - matrix models, topic models, Random Indexing, neural embeddings, etc.
- Vector similarity measure
  - cosine, euclidean distance, LIN measure, etc.

Types of DSMs

- Linguistic contexts
  - text regions, linear (window-based) collocates, syntactic collocates, etc.
- Context weighting
  - raw co-occurrence frequency, entropy, tf-idf, association measures, etc.
- Distributional vector construction
  - matrix models, topic models, Random Indexing, neural embeddings, etc.
- Vector similarity measure
  - cosine, euclidean distance, LIN measure, etc.

The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity
- The Distributional Hypothesis is primarily a conjecture about word meaning
- Distributional semantics is based on a holistic and relational view of meaning
- Distributional semantics is based on a contextual and usage-based view of meaning
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts
- Distributional representations are quantitative, continuous, gradable and distributed

The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity
- The Distributional Hypothesis is primarily a conjecture about word meaning
- Distributional semantics is based on a holistic and relational view of meaning
- Distributional semantics is based on a contextual and usage-based view of meaning
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts
- Distributional representations are quantitative, continuous, gradable and distributed

The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity
- The Distributional Hypothesis is primarily a conjecture about word meaning
- Distributional semantics is based on a holistic and relational view of meaning
- Distributional semantics is based on a contextual and usage-based view of meaning
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts
- Distributional representations are quantitative, continuous, gradable and distributed

The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity
- The Distributional Hypothesis is primarily a conjecture about word meaning
- Distributional semantics is based on a holistic and relational view of meaning
- Distributional semantics is based on a contextual and usage-based view of meaning
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts
- Distributional representations are quantitative, continuous, gradable and distributed

The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity
- The Distributional Hypothesis is primarily a conjecture about word meaning
- Distributional semantics is based on a holistic and relational view of meaning
- Distributional semantics is based on a contextual and usage-based view of meaning
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts
- Distributional representations are quantitative, continuous, gradable and distributed

The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity
- The Distributional Hypothesis is primarily a conjecture about word meaning
- Distributional semantics is based on a holistic and relational view of meaning
- Distributional semantics is based on a contextual and usage-based view of meaning
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts
- Distributional representations are quantitative, continuous, gradable and distributed

Weak and Strong Distributional Hypothesis

Lenci (2008)

Weak Distributional Hypothesis

An empirical method for semantic analysis

- word meaning (whatever this might be) is reflected in linguistic distributions
- by inspecting a relevant number of distributional contexts, we may identify those aspects of meaning that are shared by words that have similar contextual distributions

applications: lexicography, ontology and thesauri learning and population, word sense disambiguation, relation extraction, question answering, Information Retrieval, etc.

Weak and Strong Distributional Hypothesis

Lenci (2008)

Weak Distributional Hypothesis

An empirical method for semantic analysis

- word meaning (whatever this might be) is reflected in linguistic distributions
- by inspecting a relevant number of distributional contexts, we may identify those aspects of meaning that are shared by words that have similar contextual distributions

applications lexicography, ontology and thesauri learning and population, word sense disambiguation, relation extraction, question answering, Information Retrieval, etc.

Weak and Strong Distributional Hypothesis

Lenci (2008)

Strong Distributional Hypothesis

A cognitive hypothesis about the form and origin of semantic representations

- word distributions in context have a specific causal role in the formation of the semantic representation for that word
- the distributional properties of words in linguistic contexts is an explanatory factor of human semantic competence

applications models of semantic memory (e.g. semantic priming, categorization, etc.), word learning, semantic processing, etc.

Weak and Strong Distributional Hypothesis

Lenci (2008)

Strong Distributional Hypothesis

A cognitive hypothesis about the form and origin of semantic representations

- word distributions in context have a specific causal role in the formation of the semantic representation for that word
- the distributional properties of words in linguistic contexts is an explanatory factor of human semantic competence

applications: models of semantic memory (e.g. semantic priming, categorization, etc.), word learning, semantic processing, etc.

Is Distributional Semantics Semantic at All?

- The answer can be based on a particular "pre-conception" of meaning

As Wittgenstein says, 'the meaning of words lies in their use.'

J. R. Firth (1951), "Modes of meaning"

Semantics with no treatment of truth conditions is not semantics

D. Lewis (1970), "General semantics"

To know the meaning of a sentence is to know its truth conditions

I. Heim and A. Kratzer (1998), *Semantics in Generative Grammar*

- The answer can be based on the types of semantic facts that distributional semantics is able to explain

Is Distributional Semantics Semantic at All?

- The answer can be based on a particular “pre-conception” of meaning

  As Wittgenstein says, ‘the meaning of words lies in their use.’

  J. R. Firth (1951), “Modes of meaning”

  Semantics with no treatment of truth conditions is not semantics

  D. Lewis (1970), “General semantics”

  To know the meaning of a sentence is to know its truth conditions

  I. Heim and A. Kratzer (1998), *Semantics in Generative Grammar*

- The answer can be based on the types of semantic facts that distributional semantics is able to explain

Is Distributional Semantics Semantic at All?

- The answer can be based on a particular “pre-conception” of meaning

  As Wittgenstein says, ‘the meaning of words lies in their use.’
  J. R. Firth (1951), “Modes of meaning”
  Semantics with no treatment of truth conditions is not semantics
  D. Lewis (1970), “General semantics”
  To know the meaning of a sentence is to know its truth conditions
  I. Heim and A. Kratzer (1998), *Semantics in Generative Grammar*

- The answer can be based on the types of semantic facts that distributional semantics is able to explain

Distributional Semantics and its Boundaries

Success stories

- Semantic similarity
  - synonymy, categorization, etc.
- Selectional preferences
  - semantic typing, co-composition, etc.
- Context-based semantic phenomena
  - sense-shifts, gradience, world knowledge integration, etc.
- Figurative language
  - analogy, metaphor, etc.
- Cognitive modeling
  - semantic priming, similarity judgements, thematic fit, etc.

Distributional Semantics and its Boundaries

Terrae incognitae

- Function words
  - negation, quantification, logical connectives, discourse particles, ecc.
- Intensionality
  - tense, aspect, modality, etc.
- Reference and coreference
  - indexicals, anaphora, etc.

Distributional Semantics and its Boundaries

Current challenges

- Polysemy
  - sense induction, regular polysemy, etc.
- Compositionality
  - adjectival modification, predicate-argument structures, etc.
- Semantic relations
  - hypernymy, antonymy, etc.
- Inference
  - lexical entailments, presuppositions, implicatures, etc.

From Semantic Similarity to Semantic Relations

- Similar words differ for the type of relation holding between them
  - *dog* is very similar to both *animal* and *cat*, but *animal* is an hypernym and *cat* is a coordinate (co-hyponym)
- DSMs provide a quantitative correlate of semantic similarity (relatedness), but do not discriminate between different types of semantic relations
  - cf. WordNet instead provides a "typed" semantic space

From Semantic Similarity to Semantic Relations

- Similar words differ for the type of relation holding between them
  - *dog* is very similar to both *animal* and *cat*, but *animal* is an hypernym and *cat* is a coordinate (co-hyponym)
- DSMs provide a quantitative correlate of semantic similarity (relatedness), but do not discriminate between different types of semantic relations
  - cf. WordNet instead provides a “typed” semantic space

Semantic Relations

- Paradigmatic semantic relations (Lyons 1977, Cruse 1986, Fellbaum 1998, Murphy 2003)

synonymy	*sofa* - *couch*
hyperonymy	*dog* - *animal*
co-hyponymy	*dog* - *cat*
antonymy	*dead* - *alive*
meronymy	*wheel* - *car*

Distributional Neighbors from the BNC

dog (window size= 2)	
cat	0.77
horse	0.67
fox	0.65
pet	0.63
rabbit	0.61
pig	0.57
animal	0.57
mongrel	0.56
sheep	0.55
pigeon	0.54
deer	0.53
rat	0.53
bird	0.53

good (window size= 2)	
bad	0.68
excellent	0.66
superb	0.48
poor	0.45
improved	0.43
improve	0.43
perfect	0.42
clever	0.42
terrific	0.42
lucky	0.41
smashing	0.41
improving	0.41
wonderful	0.41

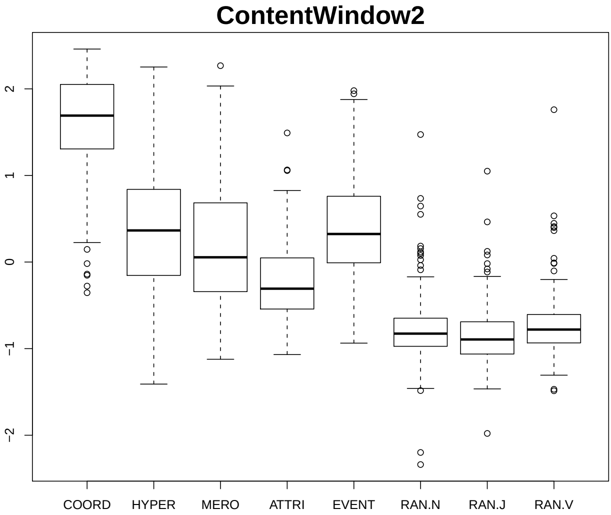
BLESS

Baroni and Lenci (2011)

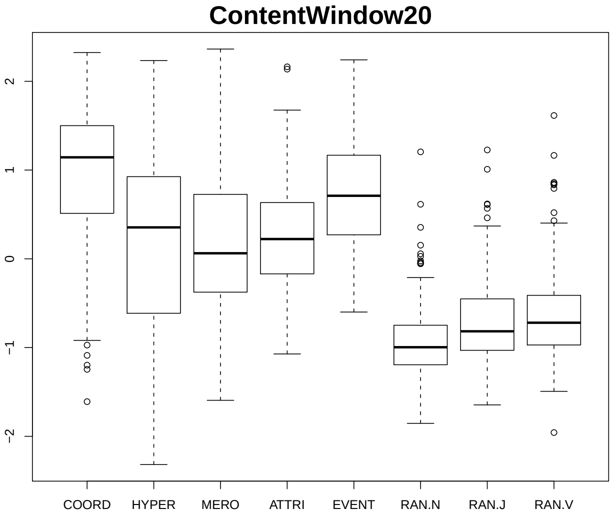
- BLESS is formed by 26,554 tuples expressing a relation between a (target) concept and a relatum (concept)
  - 200 basic-level nominal concrete concepts, 8 relation types, each instantiated by multiple relata (nouns, verbs or adjectives)
  - relata extracted from various resources (WordNet, ConceptNet, Wikipedia, corpora, etc.)

target concept	relation	relata
rabbit	HYPER	*animal, chordate, mammal, ...*
guitar	COORD	*violin, trumpet, piano, ...*
beaver	MERO	*fur, head, tooth, ...*
sword	ATTRI	*dangerous, long, heavy, ...*
butterfly	EVENT	*fly, catch, flutter, ...*
villa	RAN.N	*disease, assistance, game, ...*
donkey	RAN.V	*coincide, express, vent, ...*
hat	RAN.J	*quarterly, massive, obvious, ...*

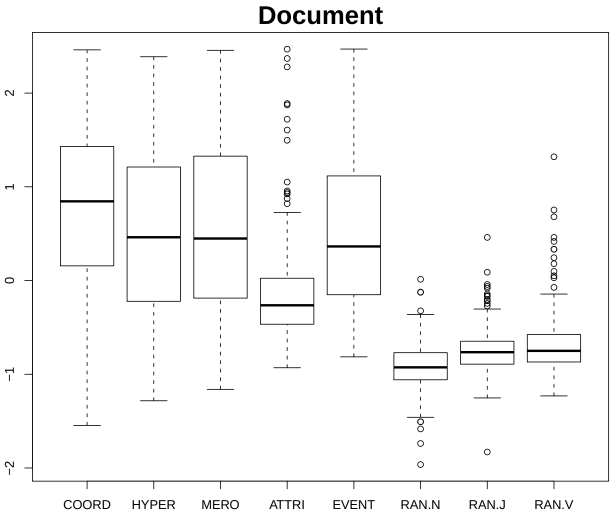
DSMs and Semantic Relations



DSMs and Semantic Relations



DSMs and Semantic Relations



Distributional Approaches to Semantic Relations

- Pattern-based approaches (Hearst 1992)
  - select as linguistic contexts lexico-syntactic patterns that express a given semantic relation between lexical items

hypernymy	*x is a kind of y*
	y such as x
antonymy	*x but not y*
	x or y

  - the pair $\langle a, b \rangle$ is an instance of relation R, if a and b are frequently linked by the patterns expressing R

Distributional Approaches to Semantic Relations

- Analogy-based approaches (Turney 2006, 2012, 2013; Baroni and Lenci 2010, Mikolov et al. 2013)
  - select instance pairs of a given semantic relation *R*
    - $\textit{Hypernymy} = \{\langle \textit{dog}, \textit{animal} \rangle, \langle \textit{tulip}, \textit{flower} \rangle, \langle \textit{cypress}, \textit{tree} \rangle \ \ldots\}$
  - $\langle a, b \rangle$ is an instance of relation *R*, if $\langle a, b \rangle$ is analogically similar to the instances of *R*
    - *dog*:*animal* = *car*:*vehicle*

Relational Similarity

Turney 2006, Baroni and Lenci 2010

- Measure relational similarity between word pairs in a pair-pattern co-occurrence matrix

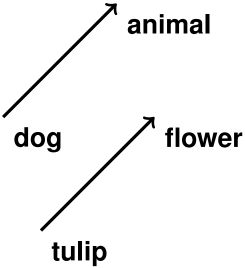
$$
\begin{array}{r c}
 & \begin{array}{cccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 \end{array} \\
\begin{array}{r} \langle \textbf{dog}, \textbf{animal} \rangle \\ \langle \textbf{tulip}, \textbf{flower} \rangle \\ \langle \textbf{cypress}, \textbf{tree} \rangle \end{array} &
\begin{pmatrix}
4 & 0 & 3 & 2 & 1 & \vdots & 2 & 3 \\
0 & 1 & 2 & 0 & 3 & \vdots & 1 & 0 \\
1 & 2 & 3 & 1 & 0 & \vdots & 2 & 0
\end{pmatrix}
\end{array}
$$

Neural embeddings

Mikolov et al. 2013

- Pairs of words sharing a particular relation are related by the same constant offset between their neural embeddings
  - distributional vectors built with a Recursive Neural Network

animal - **dog** + **tulip** = **flower**



Semantic Relations and Distributional Semantics

- Most of existing approaches are partially supervised
  - pattern selections, “seed” instances of relations, etc.
- Modeling semantic relations requires us to explain the entailments they license
  - X is a dog $\Rightarrow$ X is an animal
  - X is a dog $\not\Rightarrow$ X is a cat

Semantic Relations and Distributional Semantics

- Most of existing approaches are partially supervised
  - pattern selections, “seed” instances of relations, etc.
- Modeling semantic relations requires us to explain the entailments they license
  - X is a dog $\Rightarrow$ X is an animal
  - X is a dog $\nRightarrow$ X is a cat

Hypernymy in Distributional Semantics

- Hypernymy is an asymmetric relation
  - X is a dog $\Rightarrow$ X is an animal
  - X is an animal $\not\Rightarrow$ X is a dog
- Hypernyms are semantically broader terms than their hyponyms
  - extensionally broader
    - *animal* refers to a broader set of entities than *dog*
  - intensionally broader
    - *animal* has more general properties than *dog* (e.g. bark)
    - superordinates are *less informative* than basic level concepts (Murphy 2002)

Hypernymy in Distributional Semantics

- Hypernymy is an asymmetric relation
  - X is a dog $\Rightarrow$ X is an animal
  - X is an animal $\nRightarrow$ X is a dog
- Hypernyms are semantically broader terms than their hyponyms
  - extensionally broader
    - *animal* refers to a broader set of entities than *dog*
  - intensionally broader
    - *animal* has more general properties than *dog* (e.g. bark)
    - superordinates are *less informative* than basic level concepts (Murphy 2002)

An Extensional Approach to Hypernymy in DSMs

- Since the class (extension) denoted by a hyponym is included in the extension denoted by the hypernym, hyponyms are expected to occur in a subset of the contexts of their hypernyms

Distributional Inclusion Hypothesis (DIH) (Kotlerman et al. 2010)

if *u* is a semantically narrower term than *v*, then a significant number of salient distributional features of *u* is included in the feature vector of *v* as well

Directional Similarity Measures Based on the DIH

WeedsPrec (Weeds & Weir, 2003; Weeds et al., 2004)

$$WeedsPrec(u, v) = \frac{\sum_{f \in F_u \cap F_v} w_u(f)}{\sum_{f \in F_u} w_u(f)} \tag{1}$$

ClarkeDE (Clarke 2009)

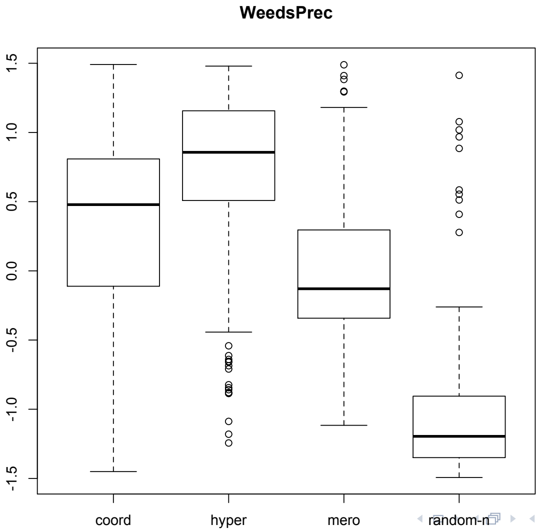
$$ClarkeDE(u, v) = \frac{\sum_{f \in F_u \cap F_v} min(w_u(f), w_v(f))}{\sum_{f \in F_u} w_u(f)} \tag{2}$$

Distributional Memory (Baroni and Lenci 2010)

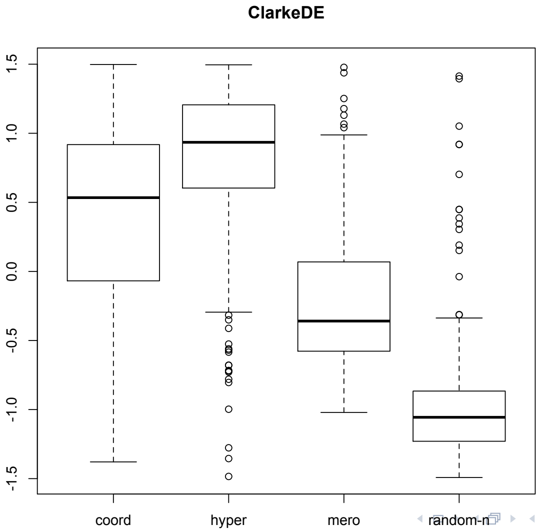
- Distributional features are syntactically typed collocates:

  `subj_intr-sing`, **`obj-read`**, **`subj_tr-read`**, etc.
- The context weighting function is Positive Local Mutual Information (LMI)
- The Distributional Memory corpus
  - 2.830 billion tokens resulting from concatenating
    - ukWac, about 1.915 billion tokens of Web-derived texts
    - English Wikipedia, a mid-2009 dump of about 820 million tokens
    - British National Corpus, about 95 million tokens
  - the corpus was tokenized, POS-tagged and lemmatized with the TreeTagger, and dependency-parsed with the MaltParser

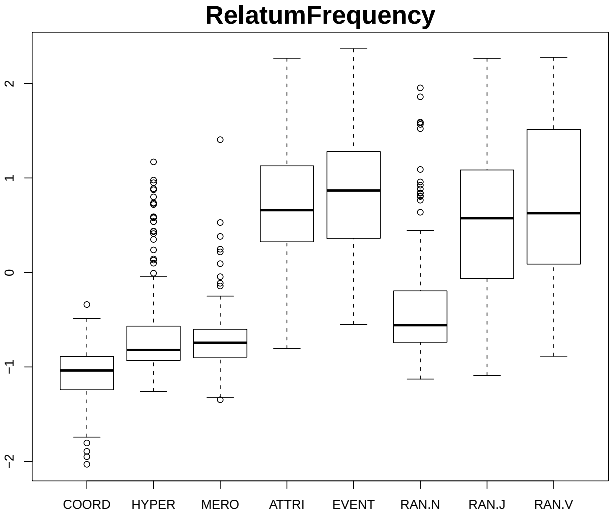
Directional Similarity Measures on BLESS



Directional Similarity Measures on BLESS



Directional Similarity Measures on BLESS



Evaluation with Average Precision (AP)

- AP combines precision, relevance ranking and overall recall
- For each similarity measure, AP is computed with respect to the 4 BLESS relations
  - the best possible score (AP = 1) for a given relation (e.g., HYPER) corresponds to the ideal case in which all the relata belonging to that relation have higher similarity scores than the relata belonging to the other relations
- For each relation R, AP is computed for each of the 200 BLESS target concepts

$$AP(u, R) = \frac{\sum_{r=1}^{|R|} (P(r) * rel(r))}{|R|} \quad (3)$$

$$rel(r) = \begin{cases} 1 & \text{if the word at rank } r \text{ has a relation } R \text{ with } u \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Evaluation with Average Precision (AP)

measure	COORD	HYPER	MERO	RANDOM-N
Cosine	0.79	0.23	0.21	0.30
WeedsPrec	0.45	0.40	0.31	0.32
ClarkeDE	0.45	0.39	0.28	0.33

Mean AP values for each semantic relation reported by the different similarity scores

New directional similarity measure

Lenci and Benotto (2012)

invCL (inverse ClarkeDE)

$$invCL(u,v) = \sqrt{ClarkeDE(u,v) * (1 - ClarkeDE(v,u))} \quad (5)$$

- A broader term should also be found in contexts in which the narrow term is not used
- If v is a semantically broader term of u, then the features of u are included in the features of v, but the features of v are also not included in the features of u, so that:
  1. a significant number of the u-contexts are also v-contexts
  2. a significant number of v-contexts are not u-contexts.

New directional similarity measure

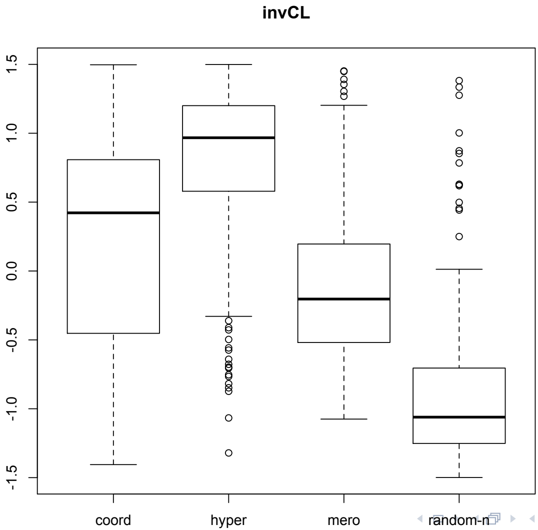
Lenci and Benotto (2012)

$invCL$ (inverse ClarkeDE)

$$invCL(u,v) = \sqrt{ClarkeDE(u,v) * (1 - ClarkeDE(v,u))} \quad (5)$$

- A broader term should also be found in contexts in which the narrow term is not used
- If v is a semantically broader term of u, then the features of u are included in the features of v, but the features of v are also not included in the features of u, so that:
  1. a significant number of the u-contexts are also v-contexts
  2. a significant number of v-contexts are not u-contexts.

Directional similarity measures



Evaluation with Average Precision (AP)

measure	COORD	HYPER	MERO	RANDOM-N
Cosine	0.79	0.23	0.21	0.30
WeedsPrec	0.45	0.40	0.31	0.32
ClarkeDE	0.45	0.39	0.28	0.33
invCL	0.38	0.40	0.31	0.34

Mean AP values for each semantic relation reported by the different similarity scores

An Intensional Approach to Hypernymy in DSMs

- The intension (concept) expressed by a hypernym includes more general properties than the intension of its hyponyms
  - *animal*: move, eat, is alive, etc.
  - *dog*: bark, has fur, has four legs, etc.

Distributional Informativeness Hypothesis (DIH) (Santus et al. 2014)

The most typical linguistic contexts of a hypernym are less informative than the most typical linguistic contexts of its hyponyms

An Intensional Approach to Hypernymy in DSMs

Santus et al. (2014)

- For every word w we identify the N most associated contexts c
- For each selected context c we calculate its entropy:

$$H(c) = \sum_{i=1}^{n} p(c, f_i) \log_2 p(c, f_i)$$

- For each w we calculate a generality index E_w as the median among the entropies of its N contexts
- We compare the semantic generality between two words w_1 and w_2:

$$SLQS(w_1, w_2) = 1 - \frac{E_{w_1}}{E_{w_2}}$$

- If $SLQS(w_1, w_2) > 0$, then w_1 is semantically less general than w_2

An Intensional Approach to Hypernymy in DSMs

Santus et al. (2014)

- Experiment 1 - identifying the hypernym in the 1277 hypernymy-related pairs in BLESS
  - given a hyponym - hypernym pair (w_1, w_2), the hypernym is correctly identified iff $SLSQ(w_1, w_2) > 0$

	SLQS	baseline (freq)
accuracy	87.00%	66.09%

- Experiment 2 - discriminating hypernyms from other semantic relations
  - SLQS is combined here with the cosine (hypernyms are both very similar but differ for semantic generality)

	COORD	HYPER	MERO	RANDOM-N
baseline	0.51	0.40	0.38	0.17
*SLQS*Cosine*	0.27	0.59	0.35	0.24

An Intensional Approach to Hypernymy in DSMs

Santus et al. (2014)

- Experiment 1 - identifying the hypernym in the 1277 hypernymy-related pairs in BLESS
    - given a hyponym - hypernym pair (w_1, w_2), the hypernym is correctly identified iff $SLSQ(w_1, w_2) > 0$

	SLQS	baseline (freq)
accuracy	87.00%	66.09%

- Experiment 2 - discriminating hypernyms from other semantic relations
    - SLQS is combined here with the cosine (hypernyms are both very similar but differ for semantic generality)

	COORD	HYPER	MERO	RANDOM-N
baseline	0.51	0.40	0.38	0.17
*SLQS*Cosine*	0.27	0.59	0.35	0.24

An Intensional Approach to Hypernymy in DSMs

Santus et al. (2014)

- Experiment 1 - identifying the hypernym in the 1277 hypernymy-related pairs in BLESS
  - given a hyponym - hypernym pair (w_1, w_2), the hypernym is correctly identified iff $SLSQ(w_1, w_2) > 0$

	SLQS	baseline (freq)
accuracy	87.00%	66.09%

- Experiment 2 - discriminating hypernyms from other semantic relations
  - SLQS is combined here with the cosine (hypernyms are both very similar but differ for semantic generality)

	COORD	HYPER	MERO	RANDOM-N
baseline	0.51	0.40	0.38	0.17
*SLQS*Cosine*	0.27	0.59	0.35	0.24

An Intensional Approach to Hypernymy in DSMs

Santus et al. (2014)

- Experiment 1 - identifying the hypernym in the 1277 hypernymy-related pairs in BLESS
  - given a hyponym - hypernym pair (w_1, w_2), the hypernym is correctly identified iff $SLSQ(w_1, w_2) > 0$

	SLQS	baseline (freq)
accuracy	87.00%	66.09%

- Experiment 2 - discriminating hypernyms from other semantic relations
  - SLQS is combined here with the cosine (hypernyms are both very similar but differ for semantic generality)

	COORD	HYPER	MERO	RANDOM-N
baseline	0.51	0.40	0.38	0.17
*SLQS*Cosine*	0.27	0.59	0.35	0.24

The Antonymy Conundrum

the reality of contextual representations has been argued so far with only a bare mention of the most likely source of counter-examples, namely, antonymous adjectives. The problem is easily stated. Antonyms have contrasting meanings - not just zero semantic similarity, but negative similarity, if that is possible. Yet they seem to be freely substitutable for one another: If the referent of a head noun has a particular attribute, the noun could take either polar value of that attribute.

Miller and Charles (1991: 25)

- Cosine similarity is not able to discriminate between synonyms and antonyms

	good
bad	0.68
excellent	0.66

The Antonymy Conundrum

the reality of contextual representations has been argued so far with only a bare mention of the most likely source of counter-examples, namely, antonymous adjectives. The problem is easily stated. Antonyms have contrasting meanings - not just zero semantic similarity, but negative similarity, if that is possible. Yet they seem to be freely substitutable for one another: If the referent of a head noun has a particular attribute, the noun could take either polar value of that attribute.

Miller and Charles (1991: 25)

- Cosine similarity is not able to discriminate between synonyms and antonyms

good	
bad	0.68
excellent	0.66

The Antonymy Conundrum

- Antonyms are strongly associated in the mental lexicon (Deese 1964, 1965)

The Co-Occurrence Hypothesis (Miller and Charles 1989, Fellbaum 1995)

Semantically opposed lexemes tend to co-occur in the same sentences

- Many empirical validations (Fellbaum 1995, Jones et al. 2012):

  All creatures great and small
  Rich and poor alike
  A matter of life or death
  Will the danger increase or decrease!

The Antonymy Conundrum

- Antonyms are strongly associated in the mental lexicon (Deese 1964, 1965)

The Co-Occurrence Hypothesis (Miller and Charles 1989, Fellbaum 1995)

Semantically opposed lexemes tend to co-occur in the same sentences

- Many empirical validations (Fellbaum 1995, Jones et al. 2012):

  All creatures great and small
  Rich and poor alike
  A matter of life or death
  Will the danger increase or decrease!

Antonyms in Distributional Semantics

- Mohammad et al.(2012) use an analogy-based approach combined with the Co-Occurrence Hypothesis
- Kim et al. (2013) apply neural embeddings to reconstruct intermediate values (e.g., *angry*) in adjective scales, given the antonyms (e.g., *furious* – *happy*)

Antonyms in Distributional Semantics

- The Co-Occurrence Hypothesis is not enough to identify antonyms
  - If two words are antonyms, they tend to co-occur, but...
    ...many words pairs that tend to co-occur are not antonyms
- Modeling antonyms requires us to explain the entailments they license

complementary	X is a alive $\Rightarrow$ X is not dead
	X is not alive $\Rightarrow$ X is dead
contrary	X is furious $\Rightarrow$ X is not happy
	X is not furious $\not\Rightarrow$ X is happy

Antonyms in Distributional Semantics

- The Co-Occurrence Hypothesis is not enough to identify antonyms
  - If two words are antonyms, they tend to co-occur, but...
    ...many words pairs that tend to co-occur are not antonyms
- Modeling antonyms requires us to explain the entailments they license

complementary	X is a alive $\Rightarrow$ X is not dead
	X is not alive $\Rightarrow$ X is dead
contrary	X is furious $\Rightarrow$ X is not happy
	X is not furious $\not\Rightarrow$ X is happy

Antonyms in Distributional Semantics

- The Co-Occurrence Hypothesis is not enough to identify antonyms
  - If two words are antonyms, they tend to co-occur, but...
    ...many words pairs that tend to co-occur are not antonyms
- Modeling antonyms requires us to explain the entailments they license

complementary	X is a alive $\Rightarrow$ X is not dead
	X is not alive $\Rightarrow$ X is dead
contrary	X is furious $\Rightarrow$ X is not happy
	X is not furious $\nRightarrow$ X is happy

Adjectival Antonyms in Semantic Spaces

Benotto and Lenci (2014)

Distributional Negation Hypothesis

If two adjectives are antonyms, each adjective is distributionally similar to the negation of the other

- The adjective *alive* is expected to share many contexts with NOT-*dead*
- Adjectives and their negation are represented with distributional vectors
  - $\mathbf{w}^{+}$ vector derived from all the positive occurrences of *w* in the training corpus
    - e.g. *The wood seemed alive, yet silent except for a wren*
  - $\mathbf{w}^{-}$ vector derived from all the negative occurrences of *w* in the training corpus
    - e.g. *The patient is not alive in the morning*

Adjectival Antonyms in Semantic Spaces

Benotto and Lenci (2014)

Hypothesis if w_i and w_j are antonyms,
$(cosine(w_i{}^+, w_j{}^-) \vee (cosine(w_i{}^-, w_j{}^+)) > cosine(w_i{}^+, w_j{}^+)$

Test set 83 pairs of antonyms (frequency of each antonym and its negation > 50)

automatic	conscious
comfortable	uncomfortable
obligatory	optional
drunk	sober
horizontal	vertical
national	international

Evaluation Accuracy of 77.10%

Adjectival Antonyms in Semantic Spaces

Benotto and Lenci (2014)

Hypothesis if w_i and w_j are antonyms, $(cosine(w_i{}^{+}, w_j{}^{-}) \vee (cosine(w_i{}^{-}, w_j{}^{+})) > cosine(w_i{}^{+}, w_j{}^{+})$

Test set 83 pairs of antonyms (frequency of each antonym and its negation > 50)

automatic	conscious
comfortable	uncomfortable
obligatory	optional
drunk	sober
horizontal	vertical
national	international

Evaluation Accuracy of 77.10%

Adjectival Antonyms in Semantic Spaces

Benotto and Lenci (2014)

Hypothesis if w_i and w_j are antonyms, $(cosine(w_i{}^{+}, w_j{}^{-}) \vee (cosine(w_i{}^{-}, w_j{}^{+})) > cosine(w_i{}^{+}, w_j{}^{+})$

Test set 83 pairs of antonyms (frequency of each antonym and its negation > 50)

automatic	conscious
comfortable	uncomfortable
obligatory	optional
drunk	sober
horizontal	vertical
national	international

Evaluation Accuracy of 77.10%

Will Distributional Semantics Ever Become Semantics?

- Distributional semantics was born as an empirical method to measure semantic similarity with corpus-based distributional statistics
- The range of addressed phenomena has enormously increased, but distributional semantics has still to prove its plausibility as a general semantic model
  - we have barely scratched the surface of key phenomena like compositionality, inference, semantic relations etc.
- Distributional semantics is able to address phenomena that are very problematic for other semantic theories
  - context-sense shifts, gradience, usage-based aspects of meaning, etc.

Will Distributional Semantics Ever Become Semantics?

- Distributional semantics was born as an empirical method to measure semantic similarity with corpus-based distributional statistics
- The range of addressed phenomena has enormously increased, but distributional semantics has still to prove its plausibility as a general semantic model
  - we have barely scratched the surface of key phenomena like compositionality, inference, semantic relations etc.
- Distributional semantics is able to address phenomena that are very problematic for other semantic theories
  - context-sense shifts, gradience, usage-based aspects of meaning, etc.

Will Distributional Semantics Ever Become Semantics?

- Distributional semantics was born as an empirical method to measure semantic similarity with corpus-based distributional statistics
- The range of addressed phenomena has enormously increased, but distributional semantics has still to prove its plausibility as a general semantic model
  - we have barely scratched the surface of key phenomena like compositionality, inference, semantic relations etc.
- Distributional semantics is able to address phenomena that are very problematic for other semantic theories
  - context-sense shifts, gradience, usage-based aspects of meaning, etc.

Will Distributional Semantics Ever Become Semantics?

- Two major issues:
  - To what extent semantic phenomena can be formulated in terms of vector similarity?
  - Which aspects of semantic knowledge can be decoded from linguistic distributions?
- New prospectives come from integrating distributional semantics with other models of meaning, to establish a division of semantic labour that capitalizes on complementary strengths

Will Distributional Semantics Ever Become Semantics?

- Two major issues:
  - To what extent semantic phenomena can be formulated in terms of vector similarity?
  - Which aspects of semantic knowledge can be decoded from linguistic distributions?
- New prospectives come from integrating distributional semantics with other models of meaning, to establish a division of semantic labour that capitalizes on complementary strengths

Computational Models of Language Meaning in Context

Dagstuhl Seminar, November 10 – 15, 2013

Organizers Hans Kamp, Alessandro Lenci, James Pustejovsky

Aims Probing the limits of distributional semantics and fostering new synergies with other semantic frameworks

Collaborators

Marco Baroni, Giulia Benotto, Gianluca Lebani, Qin Lu,
Magnus Sahlgren, Enrico Santus, Sabine Schulte im Walde

Thank You!!!

Tänan!!!

Grazie!!!

References

- M. Baroni and A. Lenci (2010), “Distributional Memory: A General Framework for Corpus-Based Semantics”, *Computational Linguistics*, 36(4): 673–721
- M. Baroni and A. Lenci (2011), “How we BLESSed distributional semantic evaluation”, in *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, Edinburgh
- G. Benotto and A. Lenci (2014), “Antonyms and Negation in Distributional Semantics”, in preparation
- W. Charles and G. A. Miller (1989), “Contexts of Antonymous Adjectives”, *Applied Psycholinguistics* 10: 357–375
- D. Clarke (2009), “Context-theoretic semantics for natural language: an overview”, in *Proceedings of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics*, Athens, Greece: 112–119
- D. A. Cruse (1986), *Lexical Semantics*, Cambridge University Press, Cambridge
- J. Deese (1964), “The Associative Structure of Some English Adjectives”, *Journal of Verbal Learning and Verbal Behavior* 3: 347–357
- J. Deese (1965), *The Structure of Associations in Language and Thought*, Johns Hopkins Press, Baltimore, MD
- C. Fellbaum (1995), “Co-occurrence and antonymy”, *International Journal of Lexicography*, 8(4): 281–303
- C. Fellbaum (ed.) (1998), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA
- J. R. Firth (1951), “Modes of meaning”, in *Essays and Studies*, The English Association, Oxford
- P. Garvin, (1962), “Computer participation in linguistic research”, *Language*, 38(4): 385-389
- Z. S. Harris (1951), *Methods in Structural Linguistics*, University of Chicago Press, Chicago, IL
- Z. S. Harris (1954), “Distributional structure”, *Word*, 10(2-3): 146–162
- M. Hearst (1992), “Automatic acquisition of hyponyms from large text corpora”, in *Proceedings of COLING 1992*, Nantes, France: 539–545

References

- I. Heim and A. Kratzer (1998), *Semantics in Generative Grammar*, Blackwell, Oxford
- S. Jones, M. L. Murphy, C. Paradis and C. Willners (2012), *Antonyms in English: construals, constructions, and canonicity*, Cambridge University Press, Cambridge
- J.-K. Kim and M.-C. de Marneffe (2013), "Deriving adjectival scales from continuous space word representations", in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington: 1625–1630
- L. Kotlerman, I. Dagan, I. Szpektor, and M. Zhitomirsky-Geffet (2010), "Directional distributional similarity for lexical inference", *Natural Language Engineering*, 16(04): 359–389
- A. Lenci (2008), "Distributional semantics in linguistic and cognitive research", in Lenci A. (ed.), *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science*, special issue of the *Italian Journal of Linguistics*, 20(1): 1–31
- A. Lenci and G. Benotto (2012), "Identifying hypernyms in distributional semantic spaces", in *Proceedings of *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1*, Montreal: 75–79
- D. Lewis (1970), "General Semantics", *Synthese*, 22(12): 18–67
- J. Lyons (1977), *Semantics. Volume 1*, Cambridge University Press, Cambridge
- T. Mikolov, W. Yih, and G. Zweig (2013), "Linguistic Regularities in Continuous Space Word Representations", in *Proceedings of NAACL-HLT 2013, Atlanta, Georgia*: 746–751
- S. M. Mohammad, B. J. Dorr, G. Hirst and P. D Turney (2012), "Computing Lexical Contrast", *Computational Linguistics*, 39(3): 1–60
- G. A. Miller (1967), "Empirical methods in the study of semantics", in *Journeys in Science: Small Steps – Great Strides*, University of New Mexico Press, Albuquerque: 51–73
- G. Murphy (2002), *The Big Book of Concepts*, MIT Press, Cambridge MA
- M. L. Murphy (2003), *Semantic Relations and the Lexicon. Antonymy, Synonymy, and the Other Paradigms*, Cambridge University Press, Cambridge

References

- G. Miller and W. Charles (1991), "Contextual correlates of semantic similarity", *Language and Cognitive Processes* 6(1):1–28
- E. Santus, A. Lenci, Q. Lu and S. Sabine Schulte im Walde (2014), "Chasing Hypernyms in Distributional Spaces with Entropy", submitted.
- Peter D. Turney (2006), "Similarity of Semantic Relations" in *Computational Linguistics*, 32(3): 379–416
- Peter D. Turney (2012), "Domain and Function: A Dual-Space Model of Semantic Relations and Compositions", *Journal of Artificial Intelligence Research*, 44: 533–585
- Peter D. Turney (2013), "Distributional Semantics Beyond Words: Supervised Learning of Analogy and Paraphrase", *Transactions of the Association for Computational Linguistics*, 1: 353–366
- J. Weeds and D. Weir (2003), "A general framework for distributional similarity", in *Proceedings of the EMNLP 2003*, Sapporo, Japan: 81–88.
- J. Weeds, D. Weir and D. McCarthy (2004), "Characterising measures of lexical distributional similarity", in *proceedings of COLING 2004*, Geneva, Switzerland: 1015–1021.