Embeddings in Natural Language Processing

Distributional Semantic Models

Eva Maria Vecchi

Computational Linguistics Fall School 2019
IMS, University of Stuttgart

September 10, 2019

The knowledge bottleneck

- Inference requires formalized *knowledge* about the world and about the meanings of words.

The knowledge bottleneck

- Inference requires formalized *knowledge* about the world and about the meanings of words.
- **Q**: *Which genetically caused connective tissue disorder has severe symptoms and complications regarding the aorta and skeletal features, and, very characteristically, ophthalmologic subluxation?*

(Credit: Koller 2016)

The knowledge bottleneck

- Inference requires formalized *knowledge* about the world and about the meanings of words.
- **Q**: *Which genetically caused connective tissue disorder has severe symptoms and complications regarding the aorta and skeletal features, and, very characteristically, ophthalmologic subluxation?*
- **D**: *Marfan's is created by a defect of the gene that determines the structure of Fibrillin-11. One of the symptoms is displacement of one or both of the eyes' lenses. The most serious complications affect the cardiovascular system, especially heart valves and the aorta.*

(Credit: Koller 2016)

Lexical Semantics in Computational Linguistics

- Many words are *synonymous*, or at least *semantically similar*
- *He has passed on, met his maker, kicked the bucket, expired, ceased to be!*

(Credit: Koller 2016)

Information Retrieval

- Goal to find relevant documents, even if differently phrased
- QUERY: "female astronauts"
- DOCUMENT: "In the history of the Soviet space program, there were only three female cosmonauts: Valentina Tereshkova, Svetlana Savitskaya, and Elena Kondakova"
- System must recognize that *astronaut* and *cosmonaut* have similar meanings (in a given context!).

(Credit: Koller 2016)

Machine Translation

The box is in the pen. Bar-Hillel (1960)

- World knowledge necessary to disambiguate *polysemous* words
- Correct translation depends on selecting the correct sense of *pen*

(Credit: Koller 2016)

(Back to) Classical Lexical Semantics

- **Polysemy**: Word has two different meanings that are clearly related to each other
  - School_1: institution at which students learn
  - School_2: building that houses school_1
- **Homonyny**: Word has two different meanings that have no obvious relation to each other.
  - Bank_1: financial institution
  - Bank_2: land alongside a body of water

(Credit: Koller 2016)

Word Sense Disambiguation

- Word sense disambiguation is the problem of tagging each word token with its word sense.
- WSD accuracy depends on sense inventory; state of the art is above 90% on coarse-grained senses
- Techniques tend to combine supervised training on small amount of annotated data with unsupervised methods.

(Credit: Koller 2016)

Problem

- Hand-written thesauruses much too small
  - English Wordnet: 117.000 synsets
  - GermaNet: 85.000 synsets
- Number of word types in English Google n-gram corpus: > 1 million.
- This is not how we can solve the query expansion problem
- Can we learn lexical semantic knowledge automatically?
  - ...and in a way that is cognitively sound?

(Credit: Koller 2016)

Meaning and Distribution

We found a little, hairy wampimuk sleeping behind the tree.

Meaning and Distribution

We found a little, hairy wampimuk sleeping behind the tree.

- "Die Bedeutung eines Wortes liegt in seinem Gebrauch." (Ludwig Wittgenstein)
  - meaning = use = distribution in language

Meaning and Distribution

We found a little, hairy wampimuk sleeping behind the tree.

- "Die Bedeutung eines Wortes liegt in seinem Gebrauch." (Ludwig Wittgenstein)
  - meaning = use = distribution in language
- "You shall know a word by the company it keeps." (Firth, 1957)
  - distribution = collocations = habitual word combinations

Meaning and Distribution

We found a little, hairy wampimuk sleeping behind the tree.

- "Die Bedeutung eines Wortes liegt in seinem Gebrauch." (Ludwig Wittgenstein)
  - meaning = use = distribution in language
- "You shall know a word by the company it keeps." (Firth, 1957)
  - distribution = collocations = habitual word combinations
- Distributional hypothesis: difference of meaning correlates with difference of distribution (Zellig Harris, 1954)
  - semantic distance
  - Assumption: Semantically similar words tend to occur in the context of the same words. $\longrightarrow$ "similar" as approximation of "synonymous"

Meaning and Distribution

We found a little, hairy wampimuk sleeping behind the tree.

- “Die Bedeutung eines Wortes liegt in seinem Gebrauch.” (Ludwig Wittgenstein)
  - meaning = use = distribution in language
- “You shall know a word by the company it keeps.” (Firth, 1957)
  - distribution = collocations = habitual word combinations
- Distributional hypothesis: difference of meaning correlates with difference of distribution (Zellig Harris, 1954)
  - semantic distance
  - Assumption: Semantically similar words tend to occur in the context of the same words. ⟶ “similar” as approximation of “synonymous”
- “What people know when they say that they know a word is not how to recite its dictionary definition – they know how to use it [. . .] in everyday discourse.” (Miller, 1986)

(Evert et al, tutorial at ESSLLI 2018)

What does "bardiwac" mean?

- *He handed her a glass of bardiwacs.*
- *Beef dishes are made to complement the bardiwacs.*
- *Nigel staggered to his feet, face flushed from too much bardiwac.*
- *Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.*
- *I dined off bread and cheese and this excellent bardiwac.*
- *The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.*

What does "bardiwac" mean?

- *He handed her a glass of bardiwacs.*
- *Beef dishes are made to complement the bardiwacs.*
- *Nigel staggered to his feet, face flushed from too much bardiwac.*
- *Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.*
- *I dined off bread and cheese and this excellent bardiwac.*
- *The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.*

⟶ Bardiwac is a red wine

(Stefan Evert, tutorial at NAACL 2010)

Distributional semantics

Landauer and Dumais 1997, Turney and Pantel 2010, …

```
he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
 surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
 in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
 But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
 they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
 the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind
```

Distributional semantics

The geometry of meaning

Distributional Semantic Model (DSM): a scaled and/or transformed co-occurrence matrix $\mathbf{M}$, such that each row $\mathbf{x}$ represents the distribution of a target term across contexts.

- e.g., within a document, within a window of [content] words before and after, etc.

	shadow	shine	planet	night
moon	16	29	10	22
sun	15	45	14	10
dog	10	0	0	4

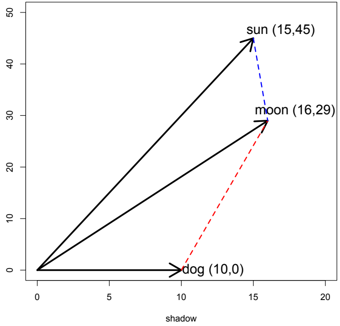
Distributional semantics

The geometry of meaning

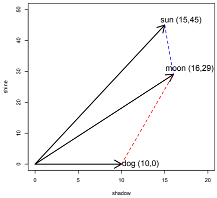
Distributional Semantic Model (DSM): a scaled and/or transformed co-occurrence matrix $\mathbf{M}$, such that each row $\mathbf{x}$ represents the distribution of a target term across contexts.

- e.g., within a document, within a window of [content] words before and after, etc.

	shadow	shine	planet	night
moon	16	29	10	22
sun	15	45	14	10
dog	10	0	0	4

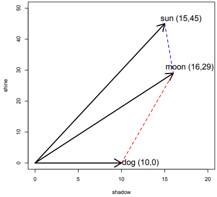


Lexical similarity



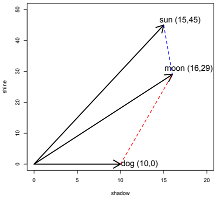
- Semantic similarity approximated by geometric distance of vectors (angle)
    - (correctly) ignores length of vectors (= frequency of words)
    - similar angle = similar proportion of context words

Lexical similarity



- Semantic similarity approximated by geometric distance of vectors (angle)
  - (correctly) ignores length of vectors (= frequency of words)
  - similar angle = similar proportion of context words
- Cosine of an angle is easy to compute
  - cos $\longrightarrow$ 1: angle is 0° (very similar)
  - cos $\longrightarrow$ 0: angle is 90° (very dissimilar)

Lexical similarity



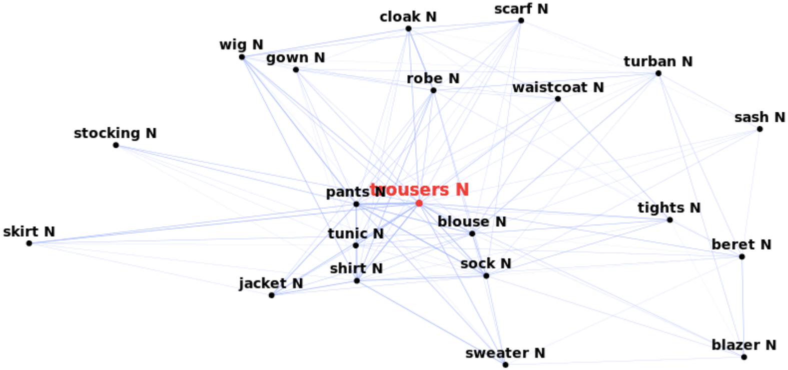
- Semantic similarity approximated by geometric distance of vectors (angle)
  - (correctly) ignores length of vectors (= frequency of words)
  - similar angle = similar proportion of context words
- Cosine of an angle is easy to compute
  - cos $\longrightarrow$ 1: angle is 0° (very similar)
  - cos $\longrightarrow$ 0: angle is 90° (very dissimilar)
- successful in tasks that concern content words: detecting synonyms, lexical entailment, ...
  - see Turney & Pantel, 2010; Baroni & Lenci, 2010, among others

Distributional Semantic Models

	get	see	use	hear	eat	kill
knife	0.027	-0.024	0.206	-0.022	-0.044	-0.042
cat	0.031	0.143	-0.243	-0.015	-0.009	0.131
dog	-0.026	0.021	-0.212	0.064	0.013	0.014
boat	-0.022	0.009	-0.044	-0.040	-0.074	-0.042
cup	-0.014	-0.173	-0.249	-0.099	-0.119	-0.042
pig	-0.069	0.094	-0.158	0.000	0.094	0.265
banana	0.047	-0.139	-0.104	-0.022	0.267	-0.042

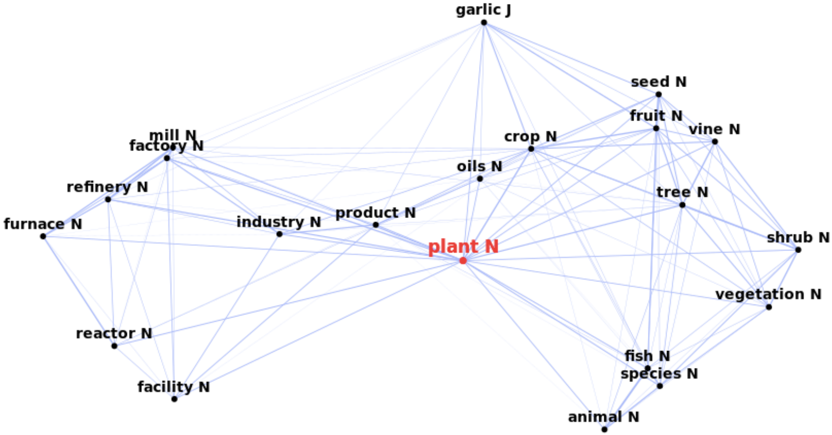
(Evert et al, tutorial at ESSLLI 2018)

Nearest Neighbors of *trousers*



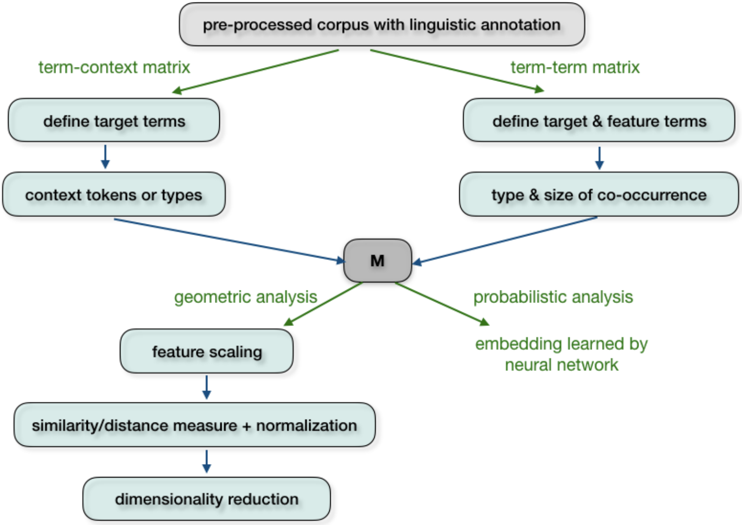
*Based on DSM built on EN Wikipedia, (filtered) dependency contexts

Nearest Neighbors of *plant*



*Based on DSM built on EN Wikipedia, (filtered) dependency contexts

Building a distributional model



(Evert et al, tutorial at ESSLLI 2018)

Linguistic Preprocessing

Defining a term

- Tokenization
- POS-tagging (*light*_N vs. *light*_J vs. *light*_V)
- Stemming/lemmatization
  - *go*, *goes*, *went*, *gone*, *going* → *go*
- Dependency parsing or shallow syntactic chunking

(Credit: Teufel 2017)

Linguistic Preprocessing

Defining a term

- Tokenization
- POS-tagging (*light*_N vs. *light*_J vs. *light*_V)
- Stemming/lemmatization
  - *go*, *goes*, *went*, *gone*, *going* → *go*
- Dependency parsing or shallow syntactic chunking

Effect of linguistic preprocessing

- Nearest neighbors of *walk* (BNC, DSM defined by head of the subject of *walk*)
  - **Word forms**: stroll, walking, walked, go, path, drive, ride, wander, sprinted, sauntered
  - **Lemmatized forms**: hurry, stroll, stride, trudge, amble, wander, walk-NN, walking, retrace, scuttle

(Credit: Teufel 2017)

Term-document vs. term-term matrices

- In IR, the "context" is always exactly one document
- This results in term-document matrices (aka "Vector Space Models")
- This allows us to measure the similarity of words with sets of words (e.g. documents vs. queries in IR)
- Term-document matrices are sparse

Context Type

- Context term appears in same fixed **window**
- Context term is a member in the same **linguistic unit** as target (e.g. paragraph, sentence, turn in conversation)
- Context term is linked to target by a **syntactic dependency** (e.g. subject, modifier)

Context Type

- Context term appears in same fixed **window**
- Context term is a member in the same **linguistic unit** as target (e.g. paragraph, sentence, turn in conversation)
- Context term is linked to target by a **syntactic dependency** (e.g. subject, modifier)
- Context type (e.g. window size) can have impact on how terms are related to those in its nearest neighborhood
    - For example, the tendency for smaller window sizes is to be pragmatically related (e.g. car, van, vehicle, truck), while in larger window sizes syntagmatically related (e.g. car, drive, park, windscreen)

(Credit: Teufel 2017)

Similarity vs. Relatedness

It is generally accepted that there are (at least) two dimensions of word associations:

- **Semantic Similarity**: two words sharing a high number of salient features (attributes) $\longrightarrow$ *paradigmatic relatedness*
    - (near) synonymy (*car-automobile*)
    - hyperonymy (*car-vehicle*)
    - co-hyponymy (*car-van-lorry-bike*)

(Credit: Teufel 2017)

Similarity vs. Relatedness

It is generally accepted that there are (at least) two dimensions of word associations:

- **Semantic Similarity**: two words sharing a high number of salient features (attributes) ⟶ *paradigmatic relatedness*
  - (near) synonymy (*car-automobile*)
  - hyperonymy (*car-vehicle*)
  - co-hyponymy (*car-van-lorry-bike*)
- **Semantic Relatedness**: two words semantically associated without being necessarily similar ⟶ *syntagmatic relatedness*
  - function (*car-drive*)
  - meronymy (*car-tire*)
  - location (*car-road*)
  - attribute (*car-fast*)
  - other (*car-petrol*)

(Credit: Teufel 2017)

Feature Scaling

Feature scaling is used to "discount" less important features:

- Logarithmic scaling: $O' = log(O + 1)$ (cf. Weber-Fechner law for human perception)

Feature Scaling

Feature scaling is used to "discount" less important features:

- Logarithmic scaling: $O' = log(O + 1)$ (cf. Weber-Fechner law for human perception)
- Relevance weighting, e.g. tf.idf (information retrieval)
  - $tf.idf = tf \cdot log(D/df)$
  - tf = co-occurrence frequency O
  - df = document frequency of feature (or nonzero count)
  - D = total number of documents (or row count of $\mathbf{M}$

Feature Scaling

Feature scaling is used to “discount” less important features:

- Logarithmic scaling: $O' = log(O + 1)$ (cf. Weber-Fechner law for human perception)
- Relevance weighting, e.g. tf.idf (information retrieval)
  - $tf.idf = tf \cdot log(D/df)$
  - tf = co-occurrence frequency O
  - df = document frequency of feature (or nonzero count)
  - D = total number of documents (or row count of $\mathbf{M}$
- Statistical **association measures** (Evert 2004, 2008) take frequency of target term and feature into account
  - often based on comparison of observed and expected co-occurence frequency (how surprised are we to see context term associated with target word?)
  - measures differ in how they balance O and E

Simple association measures

- **Pointwise Mutual Information** (PMI): compares observed vs. expected frequency of a word combination

$$PMI(w_1, w_2) = log_2 \frac{f_{obs}}{f_{exp}}$$

  - Disadvantage: PMI overrates combinations involving rare terms

Simple association measures

- **Pointwise Mutual Information** (PMI): compares observed vs. expected frequency of a word combination

$$PMI(w_1, w_2) = log_2 \frac{f_{obs}}{f_{exp}}$$

  - Disadvantage: PMI overrates combinations involving rare terms

- **t-score**: How many standard deviations is f_{obs} away from assumed mean (f_{exp})?

$$assoc_{t-test}(w_1, w_2) = \frac{f_{obs} - f_{exp}}{\sqrt{f_{obs}}}$$

Simple association measures

- **Pointwise Mutual Information** (PMI): compares observed vs. expected frequency of a word combination

$$PMI(w_1, w_2) = log_2 \frac{f_{obs}}{f_{exp}}$$

  - Disadvantage: PMI overrates combinations involving rare terms

- **t-score**: How many standard deviations is f_{obs} away from assumed mean (f_{exp})?

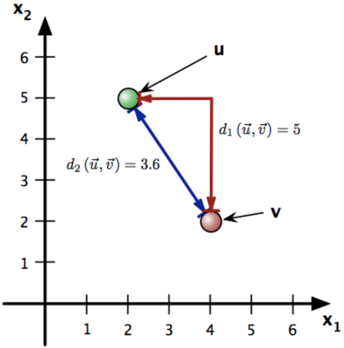
$$assoc_{t-test}(w_1, w_2) = \frac{f_{obs} - f_{exp}}{\sqrt{f_{obs}}}$$

- **Log-Likelihood** (Dunning, 1993): describes relative probability of obtaining the observed frequency for all permissible values of the parameters

$$G^2 = \pm 2 \cdot \left(f_{obs} \cdot log_2 \frac{f_{obs}}{f_{exp}} - (f_{obs} - f_{exp}) \right)$$

Geometric Distance

- **Distance** between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n \rightarrow$ (dis)similarity
  - $\mathbf{u} = (u_1, \ldots, u_n)$
  - $\mathbf{v} = (v_1, \ldots, v_n)$
- **Euclidean** distance $d_2(\mathbf{u}, \mathbf{v})$
- "City block" **Manhattan** distance $d_1(\mathbf{u}, \mathbf{v})$
- Both are special caes of the **Minkowski** p-distance $d_p(\mathbf{u}, \mathbf{v})$ (for $p \in [1, \infty]$)

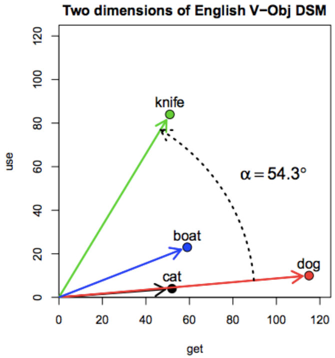


Similarity Measures

- Angle α between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ is given by

$$cos\alpha = \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}$$

- **Cosine** measure of similarity: $cos\alpha$
    - $cos\alpha = 1 \rightarrow$ collinear
    - $cos\alpha = 0 \rightarrow$ orthogonal
- Corresponding metric: **angular distance** α



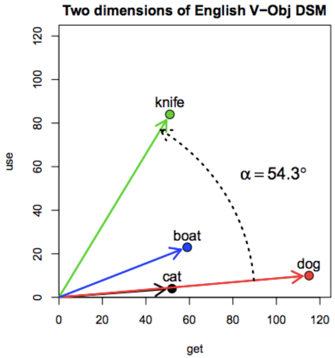
(Evert et al, tutorial at ESSLLI 2018)

Similarity Measures

- Angle α between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ is given by

$$cos\alpha = \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}$$

- **Cosine** measure of similarity: $cos\alpha$
  - $cos\alpha = 1 \rightarrow$ collinear
  - $cos\alpha = 0 \rightarrow$ orthogonal
- Corresponding metric: **angular distance** α



Euclidean distance or cosine similarity?

- They are the equivalent: if vectors have been normalized ($\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$), both lead to the same neighborhood ranking.

(Evert et al, tutorial at ESSLLI 2018)

LSA

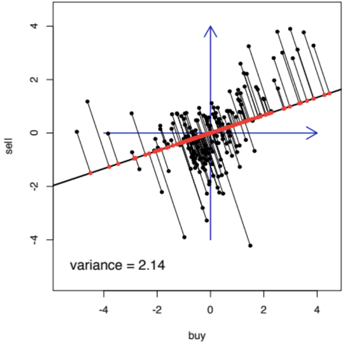
- Vectors in standard vector space are very sparse
- Orthogonal dimensions clearly wrong for near-synonyms
  canine-dog
- Different word senses are conflated into the same dimension
- One way to solve this: **dimensionality reduction**
- Hypothesis for LSA (Latent Semantic Analysis; Landauer): true semantic space has fewer dimensions than number of words observed
- Extra dimensions are noise. Dropping them brings out **latent** semantic space

Dimensionality reduction by PCA

- Principal component analysis (**PCA**)
  - orthogonal projection into orthogonal latent dimensions
  - finds optimal subjspace of given dimensionality (such that orthogonal projection preserves distance information)
  - but requires centered features → no longer sparse
- Singular value decomposition (**SVD**)
  - the mathematical algorithm behind PCA
  - often applied without centering in distributional semantics
  - note: optimality of subspace no guaranteed
- NB: row vectors should be re-normalized after PCA/SVD
  - unless cosine similarity / angular distance is used
  - also normalize vectors **before** dimensionality reduction

Dimensionality reduction by RI

- Random indexing (**RI**)
  - Project into random subspace (Sahlgren & Karlgren, 2005)
  - reasonably good if there are many subspace dimensions
  - can be performed online without collecting full co-occurrence matrix



(Evert et al, tutorial at ESSLLI 2018)

Some applications in computational linguistics

- Query expansion in IR (Grefenstette, 1994)
- Unsupervised POS induction (Schütze, 1995)
- Word sense disambiguation (Schütze, 1998; Rapp, 2004)
- Thesuarus compilation (Lin 1998; Rapp 2004)
- Attachment disambiguation (Pantel & Lin, 2000)
- Probabilistic language models (Bengio et al, 2003)
- Translation equivalents (Sahlgren & Karlgren, 2005)
- Ontology & wordnet expansion (Pantel et al, 2009)
- Language change (Sagi et al, 2009; Hamilton et al, 2016)
- Multiword expressions (Kiela & Clark, 2013)
- Analogies (Turney 2013; Gladkova et al, 2016)
- Sentiment analysis (Rothe & Schütze, 2016; Yu et al, 2017)
- $\longrightarrow$ Input representations for neural networks & machine learning

Software packages

Infomap NLP	C	*classical LSA-style DSM*
HiDEx	C++	*re-implementation of the HAL model (Lund & Burgess, 1996)*
SemanticVectors	Java	*scalable architecture based on random indexing representation*
S-Space	Java	*complex object-oriented framework*
JoBimText	Java	*UIMA / Hadoop framework*
Gensim	Python	*complex framework, focus on parallelization and out-of-core algorithms*
Vecto	Python	*framework for count & predict models*
DISSECT	Python	*user-friendly, designed for research on compositional semantics*
wordspace	R	*interactive research laboratory, but scales to real-life data sets*

Evaluation

Distributional similarity as semantic similarity

- DSMs interpret semantic similarity as a **quantitative notion**
  - if **a** is closer to **b** than to **c** in the distributional vector space, then a is more semantically similar to b than to c
- Different from **categorical** nature of most theoretical accounts
  - often expressed in terms of semantic classes and relations
- But it is not clear a priori what exactly makes two words or concepts "semantically similar" according to a DSM
  - may also depend on parameter settings

Semantic similarity and relatedness

1. **Attributional similarity** – two words sharing a large number of salient features (attributes)
   - synonymy (*car*/*automobile*)
   - hyperonymy (*car*/*vehicle*)
   - co-hyponomy (*car*/*van*/*truck*)

Semantic similarity and relatedness

1. **Attributional similarity** – two words sharing a large number of salient features (attributes)
   - synonymy (*car/automobile*)
   - hyperonymy (*car/vehicle*)
   - co-hyponomy (*car/van/truck*)

2. **Semantic relatedness** (Budanitsky & Hirst, 2006) – two words are semantically associated without necessarily being similar
   - function (*car/drive*)
   - meronymy (*car/tyre*)
   - location (*car/road*)
   - attribute (*car/fast*)

Semantic similarity and relatedness

1. **Attributional similarity** – two words sharing a large number of salient features (attributes)
   - synonymy (*car/automobile*)
   - hyperonymy (*car/vehicle*)
   - co-hyponomy (*car/van/truck*)

2. **Semantic relatedness** (Budanitsky & Hirst, 2006) – two words are semantically associated without necessarily being similar
   - function (*car/drive*)
   - meronymy (*car/tyre*)
   - location (*car/road*)
   - attribute (*car/fast*)

3. **Relational similarity** (Turney, 2006) – similar relation between pairs of words (analogy)
   - *policeman:gun* :: *teacher:book*
   - *mason:stone* :: *carpenter:wood*
   - *traffic:street* :: *water:riverbed*

DSMs and semantic similarity

- DSMs are thought to represent **paradigmatic** similarity
  - words that tend to occur in the same contexts
- Words that share many contexts will correspond to concepts that share many attributes (**attributional similarity**), i.e. concepts that are **taxonomically/ontologically similar**
  - synonyms (*rhino/rhinoceros*)
  - antonyms and values on a scale (*good/bad*)
  - co-hyponyms (*rock/jazz*)
  - hyper- and hyponyms (*rock/basalt*)
- Taxonomic similarity is seen as the **fundamental semantic relation** organising the vocabulary of a language, allowing categorization, generalization and inheritance

(Evert et al, tutorial at ESSLLI 2018)

Evaluation of (attributional) similarity

- **Synonym Identification**
  - TOEFL test (Landauer & Dumais, 1997)

Evaluation of (attributional) similarity

- **Synonym Identification**
  - TOEFL test (Landauer & Dumais, 1997)
- **Approximating semantic similarity** judgments
  - RG norms (Rubenstein & Goodenough, 1965)
  - WordSim-353 (Finkelstein et al., 2002)
  - MEN (Bruni et al., 2014), SimLex-999 (Hill et al., 2015)

Evaluation of (attributional) similarity

- **Synonym Identification**
  - TOEFL test (Landauer & Dumais, 1997)
- **Approximating semantic similarity** judgments
  - RG norms (Rubenstein & Goodenough, 1965)
  - WordSim-353 (Finkelstein et al., 2002)
  - MEN (Bruni et al., 2014), SimLex-999 (Hill et al., 2015)
- **Noun categorization**
  - ESSLLI 2008 dataset
  - AP (Almuhareb & Poesio, 2006)

Evaluation of (attributional) similarity

- **Synonym Identification**
  - TOEFL test (Landauer & Dumais, 1997)
- **Approximating semantic similarity** judgments
  - RG norms (Rubenstein & Goodenough, 1965)
  - WordSim-353 (Finkelstein et al., 2002)
  - MEN (Bruni et al., 2014), SimLex-999 (Hill et al., 2015)
- **Noun categorization**
  - ESSLLI 2008 dataset
  - AP (Almuhareb & Poesio, 2006)
- **Semantic Priming**
  - Hodgson dataset (Padó & Lapata, 2007)
  - Semantic Priming Project (Hutchison et al., 2013)

Evaluation of (attributional) similarity

- **Synonym Identification**
  - TOEFL test (Landauer & Dumais, 1997)
- **Approximating semantic similarity** judgments
  - RG norms (Rubenstein & Goodenough, 1965)
  - WordSim-353 (Finkelstein et al., 2002)
  - MEN (Bruni et al., 2014), SimLex-999 (Hill et al., 2015)
- **Noun categorization**
  - ESSLLI 2008 dataset
  - AP (Almuhareb & Poesio, 2006)
- **Semantic Priming**
  - Hodgson dataset (Padó & Lapata, 2007)
  - Semantic Priming Project (Hutchison et al., 2013)
- **Analogies** & **semantic relations** (similarity vs. relatedness)
  - Google (Mikolov et al., 2013b), BATS (Gladkova et al., 2016)
  - BLESS (Baroni & Lenci, 2011), CogALex (Santus et al., 2016)

The TOEFL synonym task

- The TOEFL dataset (80 items)
  - Target: *levied*
    Candidates: *believed*, *correlated*, ***imposed***, *requested*
  - Target: *fashion*
    Candidates: *craze*, *fathom*, ***manner***, *ration*
- DSMs and TOEFL
  1. take vectors of the target ($\mathbf{t}$) and of the candidates ($\mathbf{c_1} \ldots \mathbf{c_n}$)
  2. measure the distance between $\mathbf{t}$ and $\mathbf{c_i}$, with $1 \leq i \leq n$
  3. select $\mathbf{c_i}$ with the shortest distance in space from $\mathbf{t}$

Humans vs. machines on the TOEFL task

- Average foreign test taker: 64.5%

Humans vs. machines on the TOEFL task

- Average foreign test taker: 64.5%
- Macquarie University staff (Rapp, 2004):
  - Average of 5 non-natives: 86.75%
  - Average of 5 natives: **97.75**%

Humans vs. machines on the TOEFL task

- Average foreign test taker: 64.5%
- Macquarie University staff (Rapp, 2004):
  - Average of 5 non-natives: 86.75%
  - Average of 5 natives: **97.75**%
- Distributional semantics
  - Classic LSA (Landauer & Dumais, 1997): 64.4%
  - Padó & Lapata's (2007) dependency-based model: 73.0%
  - Distributional memory (Baroni & Lenci, 2010): 76.9%
  - Rapp's (2004) SVD-based model, lemmatized BNC: 92.5%
  - Bullinaria & Levy (2012) carry out aggressive parameter optimization: 100.0%

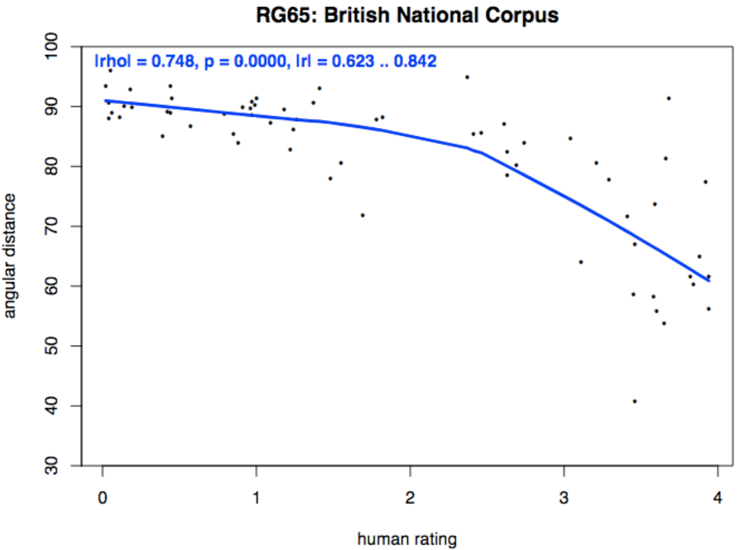
Semantic similarity judgments

- Rubenstein & Goodenough (1965) collected similarity ratings for 65 noun pairs from 51 subjects on a 0-4 scale

w_1	w_2	avg. rating
car	*automobile*	3.9
food	*fruit*	2.7
cord	*smile*	0.0

- DSMs vs. Rubenstein & Goodenough
  - for each test pair (w_1, w_2), take vectors $\mathbf{w_1}$ and $\mathbf{w_2}$
  - measure the distance (e.g. cosine) between $\mathbf{w_1}$ and $\mathbf{w_2}$
  - measure (Pearson) correlation between vector distances and R&G average judgments (Padó & Lapata, 2007)

Semantic similarity judgments



Semantic similarity judgments: results

Results on RG65 task

- Padó & Lapata's (2007) dependency-based model: 0.62
- Dependency-based on Web corpus (Herdağdelen et al., 2009)
  - without SVD reduction: 0.69
  - with SVD reduction: 0.80
- Distributional memory (Baroni & Lenci, 2010): 0.82
- Salient Semantic Analysis (Hassan & Mihalcea, 2011): 0.86

Semantic Priming

- Hearing/reading a "related" prime facilitates access to a target in various psycholinguistic tasks (naming, lexical decision, reading)
  - e.g. the word *pear* is recognized faster if heard/read after *apple*
- Hodgson (1991) single word lexical decision task, 136 prime-target pairs (cf. Padó & Lapata, 2007)
  - similar amounts of priming found for different semantic relations between primes and targets (circa 23 pairs per relation)
    - synonyms (synonym): *to dread/to fear*
    - antonyms (antonym): *short/tall*
    - coordinates (coord): *train/truck*
    - super- and subordinate pairs (supersub): *container/bottle*
    - free association pairs (freeass): *dove/peace*
    - phrasal associates (phrasacc): *vacant/building*

Semantic Priming

- DSMs and semantic priming
  1. for each related prime-target pair, measure cosine-based similarity between items (e.g., *to dread/to fear*)
  2. to estimate **unrelated primes**, take average of cosine-based similarity of target with other primes from same semantic relation (e.g., *to value/to fear*)
  3. similarity between related items should be significantly higher than average similarity between unrelated items

Semantic Priming

- DSMs and semantic priming
  1. for each related prime-target pair, measure cosine-based similarity between items (e.g., *to dread/to fear*)
  2. to estimate **unrelated primes**, take average of cosine-based similarity of target with other primes from same semantic relation (e.g., *to value/to fear*)
  3. similarity between related items should be significantly higher than average similarity between unrelated items

- Significant effects ($p < .01$) for all semantic relations
  - strongest effects for synonyms, antonyms & coordinates

Semantic Priming

- DSMs and semantic priming
  1. for each related prime-target pair, measure cosine-based similarity between items (e.g., *to dread/to fear*)
  2. to estimate **unrelated primes**, take average of cosine-based similarity of target with other primes from same semantic relation (e.g., *to value/to fear*)
  3. similarity between related items should be significantly higher than average similarity between unrelated items
- Significant effects ($p < .01$) for all semantic relations
  - strongest effects for synonyms, antonyms & coordinates
- Alternative: **classification** task
  - given target and two primes, identify related prime (→ multiple choice like TOEFL)

Evaluation Strategies

DSM evaluation in published studies

- **One model, many tasks** (Padó & Lapata 2007; Baroni & Lenci 2010; Pennington et al. 2014)
  - A novel DSM is proposed, with specific features & parameters
  - This DSM is tested on a range of different tasks (e.g. TOEFL, priming, semantic clustering)

Evaluation Strategies

DSM evaluation in published studies

- **One model, many tasks** (Padó & Lapata 2007; Baroni & Lenci 2010; Pennington et al. 2014)
  - A novel DSM is proposed, with specific features & parameters
  - This DSM is tested on a range of different tasks (e.g. TOEFL, priming, semantic clustering)

- **Incremental tuning of parameters** (Bullinaria & Levy 2007, 2012; Kiela & Clark 2014; Polajnar & Clark 2014)
  - Several parameters (e.g., scoring measure, distance metric, dimensionality reduction)
  - Many tasks (e.g. TOEFL, semantic & syntactic clustering)
  - Varying granularity of parameter settings
  - One parameter (sometimes two) varied at a time, with all other parameters set to fixed values or optimized for each setting
  - Optimal parameter values are determined sequentially

Recommended Readings

- Bullinaria, John A. and Levy, Joseph P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, **39**(3), 510-526.
- Bullinaria, John A. and Levy, Joseph P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, **44**(3), 890-907.
- Lapesa, Gabriella and Evert, Stefan (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, **2**, 531-545.

Multi-Modal DSMs

The Meaning of *Watermelon*

- The **watermelon** fruit has a smooth exterior rind (usually green with dark green stripes or yellow spots) and a juicy, sweet interior flesh.
- **Watermelon** not only boosts your "health esteem," but it is has excellent levels of vitamins A and C and a good level of vitamin B6.

The Meaning of *Watermelon*

- The **watermelon** fruit has a smooth exterior rind (usually green with dark green stripes or yellow spots) and a juicy, sweet interior flesh.
- **Watermelon** not only boosts your "health esteem," but it is has excellent levels of vitamins A and C and a good level of vitamin B6.

The Meaning of *New York City*

Multi-Modal Semantics: Motivation

- Semantics requires "grounding"
- Interesting applications at the interface of vision and language
- Better semantic representations for NLP
- Suggested Readings:
  - Bruni et al., 2014
  - Lazaridou et al., 2014
  - Silberer & Lapata, 2010
  - Roller & Schulte im Walde, 2013
  - ... *among others*

Multi-Modal Semantics: Motivation

- The relationship between form and meaning

"violin" <==> 

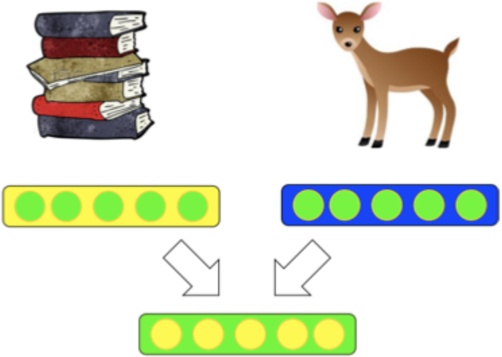
- How far can we get with textual representations alone?

Language and Vision

- Enrichment of pure textual vectors with **complementary information** coming from perceptual visual features.
  - Bruni et al., Multimodal Distributional Semantics. 2014

Language and Vision

- Enrichment of pure textual vectors with **complementary information** coming from perceptual visual features.
  - Bruni et al., Multimodal Distributional Semantics. 2014



(Elia Bruni, 2014)

Applications

Task 1 Predicting human **semantic relatedness** judgments
$\longrightarrow$ Improved!

Applications

Task 1 Predicting human **semantic relatedness** judgments
$\longrightarrow$ Improved!

Task 2 **Concept categorization**

- i.e. grouping words into classes based on their semantic relatedness
- *car* ISA *vehicle*, *banana* ISA *fruit*

$\longrightarrow$ Improved!

Applications

Task 1 Predicting human **semantic relatedness** judgments
$\longrightarrow$ Improved!

Task 2 **Concept categorization**

- i.e. grouping words into classes based on their semantic relatedness
- *car* ISA *vehicle*, *banana* ISA *fruit*

$\longrightarrow$ Improved!

Task 3 Determine the **typical color** of concrete objects

- *cardboard* is brown, *tomato* is red

$\longrightarrow$ Improved!

(Elia Bruni, 2014)

Applications

Task 1 Predicting human **semantic relatedness** judgments
$\longrightarrow$ Improved!

Task 2 **Concept categorization**

- i.e. grouping words into classes based on their semantic relatedness
- *car* ISA *vehicle*, *banana* ISA *fruit*

$\longrightarrow$ Improved!

Task 3 Determine the **typical color** of concrete objects

- *cardboard* is brown, *tomato* is red

$\longrightarrow$ Improved!

Task 4 Distinguish **literal vs. non-literal** usages of color adjectives

- *blue uniform* vs ***blue note***

$\longrightarrow$ Improved!

(Elia Bruni, 2014)

Do pigs fly?

- No, they don't → even though *pig* and *fly* are commonly seen together (idiomatic expression)

Do cats have heads?

A state-of-the-art distributional cat (Baroni et al, 2014)

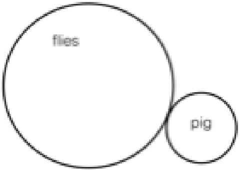
0.042 seussentennial	0.031 scarer	0.029 ragdoll
0.041 scaredy	0.031 scarer	0.029 purring
0.035 saber-toothed	0.031 repeller	0.029 whiskas
0.034 un-neutered	0.031 miaow	0.029 shorthair
0.034 meow	0.031 sphynx	0.029 scalded
0.034 unneutered	0.031 headbutts	0.029 retranslation
0.033 fanciers	0.031 spay	0.029 feral
0.033 pussy	0.030 fat	0.028 whisker
0.033 pedigreed	0.030 yowling	0.028 silvestris
0.032 sabre-toothed	0.030 flat-headed	0.028 laziest
0.032 tabby	0.030 genzyme	0.028 flap
0.032 civet	0.030 tail-less	0.028 purred
0.032 redtail	0.030 shorthaired	0.028 mummified
0.032 meowing	0.030 longhaired	...
0.032 felis	0.030 short-haired	0.0161 two-headed
0.032 whiskers	0.030 siamese	...
0.032 morphosys	0.030 english/french	0.0092 headless
0.031 meows	0.030 strangling	...
0.031 scratcher	0.030 non-pedigree	0.0021 pilgrim
0.031 black-footed	0.029 sabertooth	0.0021 out
0.031 mouser	0.029 woodpile	0.0021 head
0.031 orinthia	0.029 mewing	...

A state-of-the-art distributional cat (Baroni et al, 2014)

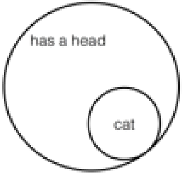
0.042 seussentennial	0.031 scarer	0.029 ragdoll
0.041 scaredy	0.031 scarer	0.029 purring
0.035 saber-toothed	0.031 repeller	0.029 whiskas
0.034 un-neutered	0.031 miaow	0.029 shorthair
0.034 meow	0.031 sphynx	0.029 scalded
0.034 unneutered	0.031 **head**butts	0.029 retranslation
0.033 fanciers	0.031 spay	0.029 feral
0.033 pussy	0.030 fat	0.028 whisker
0.033 pedigreed	0.030 yowling	0.028 silvestris
0.032 sabre-toothed	0.030 flat-**head**ed	0.028 laziest
0.032 tabby	0.030 genzyme	0.028 flap
0.032 civet	0.030 tail-less	0.028 purred
0.032 redtail	0.030 shorthaired	0.028 mummified
0.032 meowing	0.030 longhaired	...
0.032 felis	0.030 short-haired	0.0161 two-**head**ed
0.032 whiskers	0.030 siamese	...
0.032 morphosys	0.030 english/french	0.0092 **head**less
0.031 meows	0.030 strangling	...
0.031 scratcher	0.030 non-pedigree	0.0021 pilgrim
0.031 black-footed	0.029 sabertooth	0.0021 out
0.031 mouser	0.029 woodpile	0.0021 **head**
0.031 orinthia	0.029 mewing	...

World knowledge in language

- Distributional Semantics does not explain how our knowledge of **language** and our knowledge of the **world** interact!
- Model-theoretic semantics?
    - successful at modeling logical phenomena, e.g. quantification
    - set-theoretic interpretation
    - easy to interpret the logical inference of the examples given so far
    - need to integrate model-theoretic semantics, such as quantification

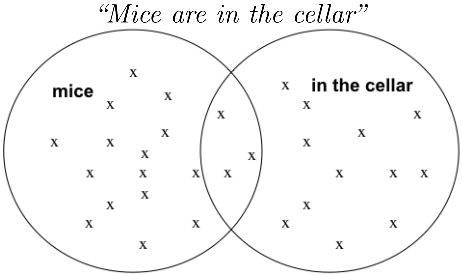


Logical inference:
if Bessy is a pig,
Bessy can't fly



Logical inference:
if Felix is a cat,
Felix has a head

Quantification



- Quantification intrinsic to most utterances
  - However, rarely explicit in naturally-occurring text
- Reference Act: *some*, *most*, *all* individuals in X do P
- Intuitive process
  - we assume only *some* of all the mice in the world have gathered – despite it not being explicit and despite not having infinite examples of mice in cellars

Modeling quantification

Quantification prerequisite for lexical semantics and inference tasks, e.g.

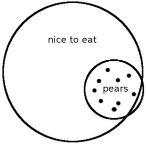
- hyponomy: *cat* is mammal
  - Without quantification we can do hyponomy, but with it, we can represent the whole scale of set overlap, up to disjointness (Erk, 2014)
- entailment: *most dogs have 4 legs* → *Lassie has 4 legs*
  - quantifier info as, say, features could permit a more direct representation of entailment (Baroni et al, 2012)
- logical inference: *the kouprey is a* MAMMAL
  - speakers have no problem knowing that if x is a *kouprey*, x is a MAMMAL, inference supported by lexical semantics of MAMMAL, which applies the property MAMMAL to all instances of the class

Modeling quantification is not trivial

- uncommon in text (circa 7% of NPs in large corpus)
- account for non-grounded quantification (*all cats are mammals*) and generics (*lions have manes*)
  - even adults make mistakes with generics
- semantics and pragmatics fail to provide an account of models themselves
- quantification highly dependent on speaker's interaction with the world and language
  - lexical semantic vs. world knowledge (e.g. speaker's beliefs about the concepts *bats* and *blind*)
  - pragmatics of quantifier use (e.g. speaker's personal interpretation of quantifiers in context)

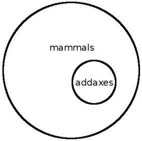
From words to worlds

I picked some pears today. They're really nice.



The reporters asked questions at the press conference.

The addax is a mammal.



Distributional and Model-Theoretic Semantics

- Distributional information influences semantic ‘knowledge’
  - e.g. knowing an *alligator* (see Erk, 2015)
  - assume a systematic relation

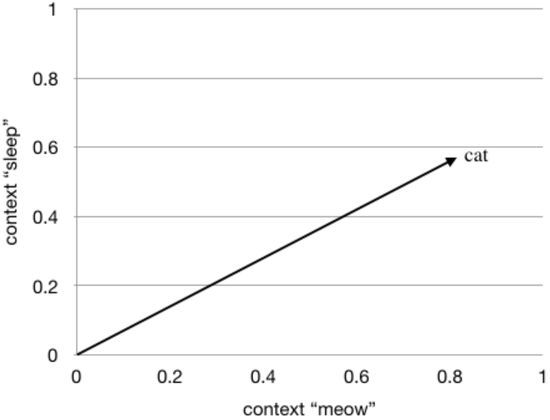
Distributional and Model-Theoretic Semantics

- Distributional information influences semantic ‘knowledge’
  - e.g. knowing an *alligator* (see Erk, 2015)
  - assume a systematic relation
- Set-theoretic models, like distributions, can be expressed in terms of vectors
  - good approximation of shared intuitions about the world

Distributional and Model-Theoretic Semantics

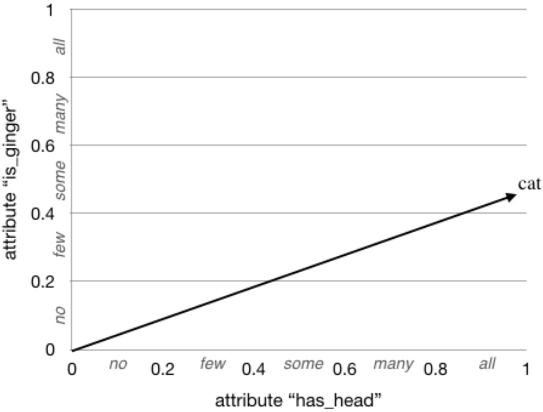
- Distributional information influences semantic ‘knowledge’
  - e.g. knowing an *alligator* (see Erk, 2015)
  - assume a systematic relation
- Set-theoretic models, like distributions, can be expressed in terms of vectors
  - good approximation of shared intuitions about the world
- Distributions can be translated into set-theoretic equivalents
  - assuming supervised learning

Distributional vector space



Weight: how lexically characteristic a context is for a target word.

Set-theoretic vector space



Weight: the set overlap between target and attribute.

Feature Norms

- Human subjects are asked to identify a concept's key attributes

AIRPLANE	SHRIMP	CUCUMBER
flies, 25	is_edible, 19	a_vegetable, 25
has_wings, 20	is_small, 17	eaten_in_salads, 24
used_for_passengers, 15	lives_in_water, 12	is_green, 23
requires_pilots, 11	is_pink, 11	is_long, 15
is_fast, 11	tastes_good, 9	eaten_as_pickles, 12

- McRae Norms (2005)
  - set of feature norms elicited from 725 participants for 541 concepts (7257 concept-feature pairs)

Feature Norms

- Used extensively in psychology but expensive to produce
- Feature norms are more “cognitively sound” than text-based distributional models, and more interpretable (Andrews et al., 2009; Făgărăşan et al., 2015)

	dog	black	book	animal	bread
CAT	4516	3124	1500	2480	1631

	has_fur	has_wheels	an_animal	a_pet	a_weapon
CAT	22	0	21	17	0

From norms to quantified predicates

(Herbelot & Vecchi, 2016)

Concept	*Feature*
ape	is_muscular
	is_wooly
	lives_on_coasts
	is_blind
	flies
tricycle	has_3_wheels
	used_by_children
	is_small
	used_for_transportation
	a_bike

From norms to quantified predicates

(Herbelot & Vecchi, 2016)

Concept	*Feature*	
ape	is_muscular	ALL
	is_wooly	MOST
	lives_on_coasts	SOME
	is_blind	FEW
tricycle	has_3_wheels	ALL
	used_by_children	MOST
	is_small	SOME
	used_for_transportation	FEW

From norms to quantified predicates

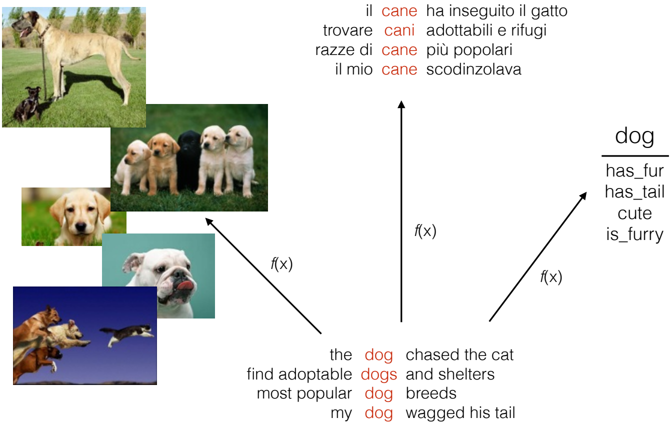
(Herbelot & Vecchi, 2016)

Concept	*Feature*		*weight*
	is_muscular	ALL	1.0
	is_wooly	MOST	0.95
ape	lives_on_coasts	SOME	0.35
	is_blind	FEW	0.05
	has_3_wheels	ALL	1.0
	used_by_children	MOST	0.95
tricycle	is_small	SOME	0.35
	used_for_transportation	FEW	0.05

Mapping between spaces

Andrews et al. (2009), Frome et al. (2013), Mikolov et al. (2013), Lazaridou et al. (2014), Făgărăşan et al. (2015), Dinu et al. (2015), etc.

Mapping between spaces



Andrews et al. (2009), Frome et al. (2013), Mikolov et al. (2013), Lazaridou et al. (2014), Făgărăşan et al. (2015), Dinu et al. (2015), etc.

Evaluation

(Herbelot & Vecchi, 2015)

1. Agreement with quantifier annotations
    - correlation between concept values in gold and mapped spaces

Evaluation

(Herbelot & Vecchi, 2015)

1. Agreement with quantifier annotations
   - correlation between concept values in gold and mapped spaces
2. Qualitative vector analysis (error analysis)
   - analysis of highly weighted contexts in mapped model-theoretic space
   - quality of neighborhoods

Evaluation

(Herbelot & Vecchi, 2015)

1. Agreement with quantifier annotations
   - correlation between concept values in gold and mapped spaces
2. Qualitative vector analysis (error analysis)
   - analysis of highly weighted contexts in mapped model-theoretic space
   - quality of neighborhoods
3. Generating quantifiers∗∗
   - map set-theoretic vectors back to natural language quantifiers for subject-predicate pairs

Generating natural language quantifiers

(Herbelot & Vecchi, 2015)

Instance		*Mapped*	*Gold*
raven	a_bird	MOST	ALL
pigeon	has_hair	FEW	NO
elephant	has_eyes	MOST	ALL
crab	is_blind	FEW	FEW
snail	a_predator	NO	NO
octopus	is_stout	NO	FEW
turtle	roosts	NO	FEW
moose	is_yellow	NO	NO
cobra	hunted_by_people	SOME	SOME
snail	forages	FEW	NO
chicken	is_nocturnal	FEW	NO
moose	has_a_heart	MOST	ALL
pigeon	hunted_by_people	NO	FEW
cobra	bites	FEW	MOST

Producing 'true' statements with 73% accuracy

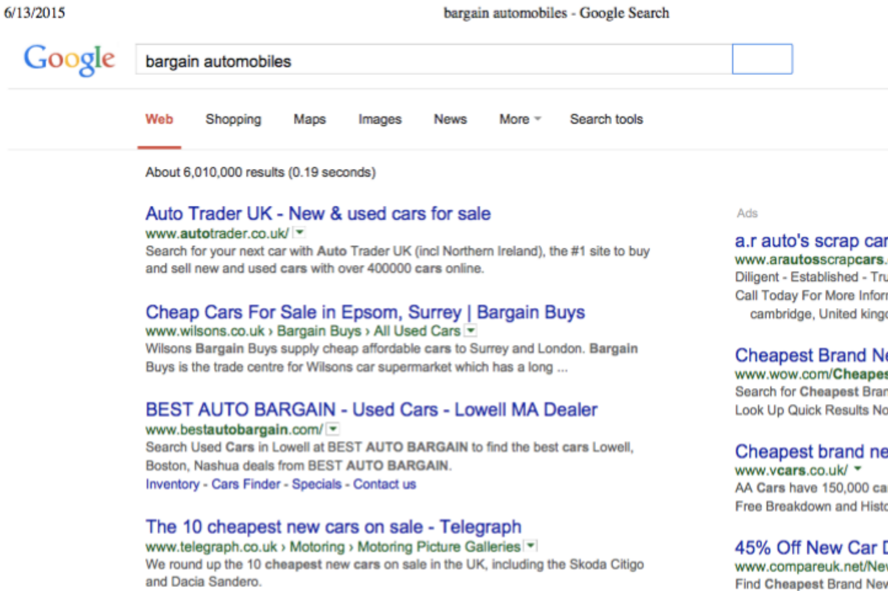
Multi-modal semantics: From words to worlds

(Herbelot & Vecchi, 2015)

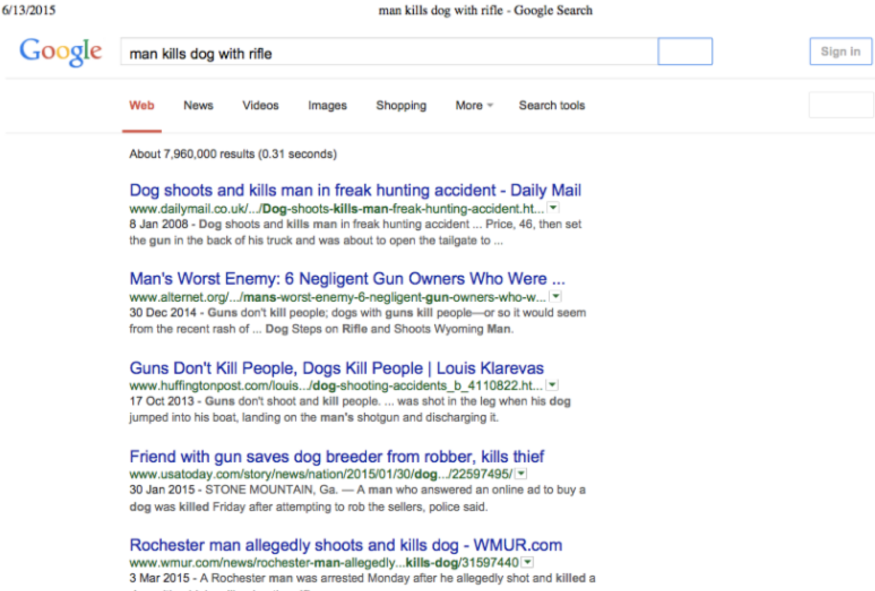
0.042 seussentennial	0.032 tabby	1 walks	1 has-a_tail
0.041 scaredy	0.032 civet	1 purrs	1 has-4_legs
0.035 saber-toothed	0.032 redtail	1 meows	1 an-animal
0.034 un-neutered	0.032 meowing	1 has-eyes	1 a-mammal
0.034 meow	0.032 felis	1 has-a_heart	1 a-feline
0.034 unneutered	0.032 whiskers	1 has-a_head	0.7 is-independent
0.033 fanciers	0.032 morphosys	1 has-whiskers	0.7 eats-mice
0.033 pussy	0.031 meows	1 has-paws	0.7 is-carnivorous
0.033 pedigreed	0.031 scratcher	1 has-fur	0.3 is-domestic
0.032 sabre-toothed	...	1 has-claws	...

Compositional Distributional Semantics

Words in Google

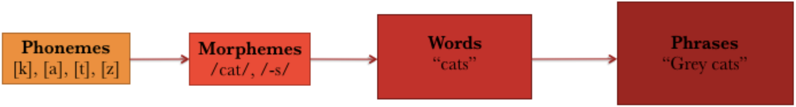


Sentences in Google



Formal Semantics and Compositionality

- It is well known that linguistic structures are **compositional**, in that simpler elements are combined to form more complex ones



- It is through the compositional quality of the phrase that meaning and a cognitive reference are formed

Logic-based frameworks in Formal Semantics

(Montague, 1974)

- Premise: No theoretically relevant difference between artificial (formal) and natural (human) languages

Logic-based frameworks in Formal Semantics

(Montague, 1974)

- Premise: No theoretically relevant difference between artificial (formal) and natural (human) languages
- Logical structures of natural languages by means of universal algebra and mathematical (formal) logic
  - *every white cat is asleep*
  - $\forall x[[white'(x) \wedge cat'(x)] \rightarrow asleep'(x)]$

Logic-based frameworks in Formal Semantics

(Montague, 1974)

- Premise: No theoretically relevant difference between artificial (formal) and natural (human) languages
- Logical structures of natural languages by means of universal algebra and mathematical (formal) logic
  - *every white cat is asleep*
  - $\forall x[[white'(x) \wedge cat'(x)] \rightarrow asleep'(x)]$
- Parallel to a syntactic system in which simple structures are put together into complex structures (e.g. Categorical grammar)
  - complex meanings are also constructed from simple meanings
  - corresponding to Frege's Principle of Compositionality

Logic-based frameworks in Formal Semantics

(Montague, 1974)

- Premise: No theoretically relevant difference between artificial (formal) and natural (human) languages
- Logical structures of natural languages by means of universal algebra and mathematical (formal) logic
  - *every white cat is asleep*
  - $\forall x[[white'(x) \wedge cat'(x)] \rightarrow asleep'(x)]$
- Parallel to a syntactic system in which simple structures are put together into complex structures (e.g. Categorical grammar)
  - complex meanings are also constructed from simple meanings
  - corresponding to Frege's Principle of Compositionality
- Note: This study is not necessarily interested in cognitive aspects, but an *elegant and simple mathematical framework* for natural language

Principle of Compositionality

(Frege, 1884)

The whole meaning of a phrase can be described according to the functional interdependency of the meanings of its well-formed parts.

1. *red manatee*
2. *fake gun* (not a gun)
3. *the horse ran* vs. *the color ran*

Frege (1884) cautions never to ask for the meaning of a word in isolation but only in the context of a statement

Principle of Compositionality

(Partee, 1995)

Partee (1995) refines the principle further by taking into account the role of syntax

- The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined
- In other words, each syntactic operation of a formal language should have a corresponding semantic operation
- Examples from Landauer et al. (1997)
    1. It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.
    2. That day the office manager, who was drinking, hit the problem sales worker with the bottle, but it was not serious.

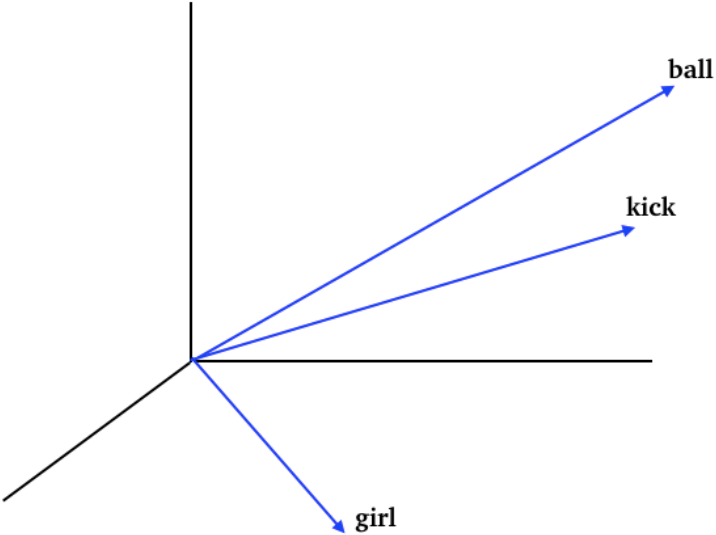
A question of degree

Compositionality is a matter of degree rather than a binary notion, since linguistic structures range across...

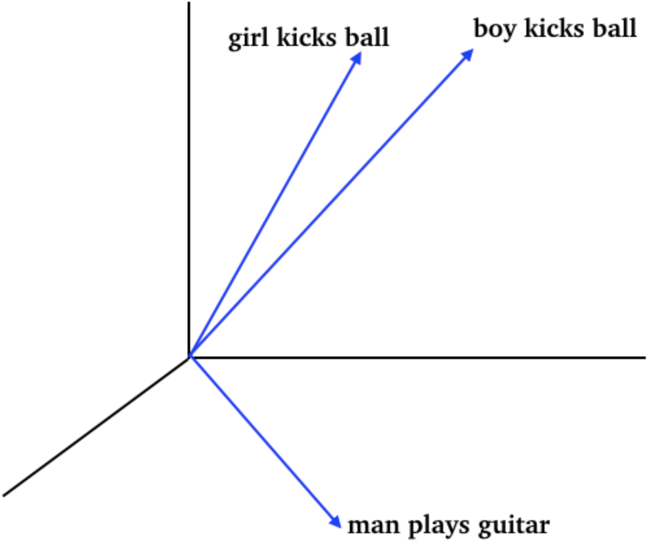
- **Fully compositional**, such as *black hair*
  - clear sense of set intersection
- **Partly compositional**: syntactically fixed expressions, such as *take advantage*, in which the constituents can still be assigned separate meaning
- **Non-compositional** phrases, such as *kick the bucket*, or **multiword expressions**, such as *by and large* whose meaning cannot be distributed across their constituents.

(Nunberg, Sag, & Wasow, 1994)

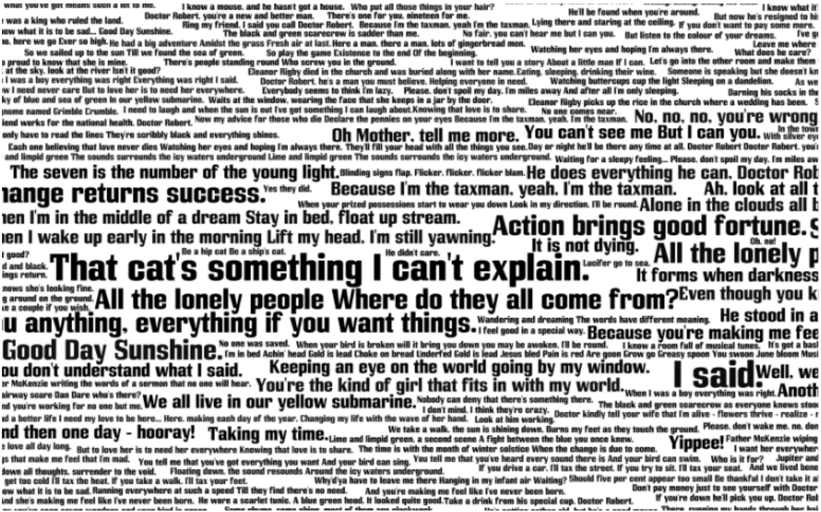
Word Space



From words to phrases



The "infinity" of sentence meaning



Vectors are too “small”

“You can’t cram the meaning of a whole %&!$# sentence into a single $&!# vector!” (Ray Mooney)

Sentence vectors?

- A fixed-size vector can't hold enough information (languages are infinite)
  - are languages really infinite? (not in practice, and maybe not in theory[1])
  - the sentence vector could be a structured object (e.g. density matrix)
  - the sentence space doesn't have to solve all of semantics (necessarily)
  - (and wouldn't this argument apply to lexical semantics as well?)

[1]Recursion and the Infinitude Claim (Pullum and Scholz, 2010)

Sentence vectors?

- A fixed-size vector can't hold enough information (languages are infinite)
  - are languages really infinite? (not in practice, and maybe not in theory[1])
  - the sentence vector could be a structured object (e.g. density matrix)
  - the sentence space doesn't have to solve all of semantics (necessarily)
  - (and wouldn't this argument apply to lexical semantics as well?)
- What about (formal) semantics?
  - compositionality, inference, logical operators, quantification, ...

[1] Recursion and the Infinitude Claim (Pullum and Scholz, 2010)

Element-wise operations on word vectors: Addition

black	0.34	0.64	...	-0.06	...

$+$

cat	0.15	0.29	...	-0.03	...

$=$

black + **cat**	0.49	0.93	...	-0.09	...

Element-wise operations on word vectors: Multiplication

black	0.34	0.64	...	-0.06	...

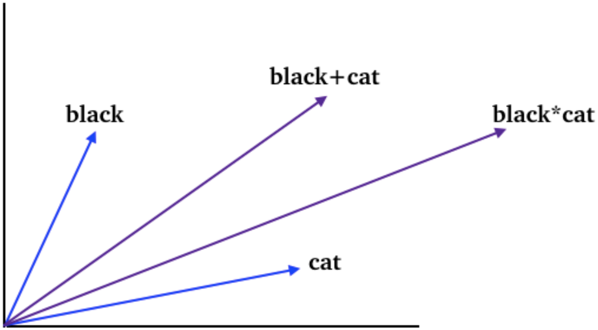
$\odot$

cat	0.15	0.29	...	-0.03	...

$=$

black $\odot$ **cat**	0.05	0.19	...	-0.002	...

Class Discussion: Pros and Cons?



A *functional* approach to composition in DS

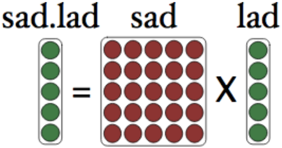
Baroni & Zamparelli EMNLP 2010, Baroni et al. LILT 2014, Paperno et al. ACL 2014
See also Coecke et al. LA 2010, Socher et al. EMNLP 2012

- Composition carried out by words that operate as functions on the representation of their input arguments

A *functional* approach to composition in DS

Baroni & Zamparelli EMNLP 2010, Baroni et al. LILT 2014, Paperno et al. ACL 2014
See also Coecke et al. LA 2010, Socher et al. EMNLP 2012

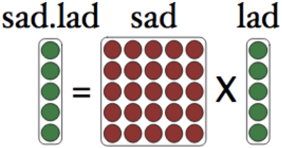
- Composition carried out by words that operate as functions on the representation of their input arguments
- Atomic arguments (nouns) are vectors, one-argument functions (e.g., adjectives, intransitive verbs) are matrices, function application is matrix-by-vector multiplication



A *functional* approach to composition in DS

Baroni & Zamparelli EMNLP 2010, Baroni et al. LILT 2014, Paperno et al. ACL 2014
See also Coecke et al. LA 2010, Socher et al. EMNLP 2012

- Composition carried out by words that operate as functions on the representation of their input arguments
- Atomic arguments (nouns) are vectors, one-argument functions (e.g., adjectives, intransitive verbs) are matrices, function application is matrix-by-vector multiplication

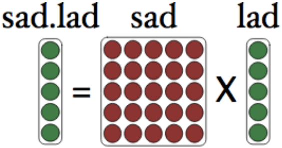


- Approach generalizes to multiple-argument functions (e.g., transitive verbs) through the tools of multi-linear algebra

A *functional* approach to composition in DS

Baroni & Zamparelli EMNLP 2010, Baroni et al. LILT 2014, Paperno et al. ACL 2014
See also Coecke et al. LA 2010, Socher et al. EMNLP 2012

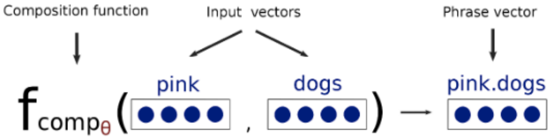
- Composition carried out by words that operate as functions on the representation of their input arguments
- Atomic arguments (nouns) are vectors, one-argument functions (e.g., adjectives, intransitive verbs) are matrices, function application is matrix-by-vector multiplication



- Approach generalizes to multiple-argument functions (e.g., transitive verbs) through the tools of multi-linear algebra
- Efficient methods to induce function representations from natural data (training corpus) in an unsupervised manner

General estimation of composition

Dinu, Pham & Baroni 2013; also: Guevara 2010, Baroni & Zamparelli 2010



- Use (reasonably frequent) corpus-extracted phrase vectors to learn the parameters of composition functions:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} ||\mathbf{P} - fcomp_\theta(\mathbf{U}, \mathbf{V})||^2$$

$\mathbf{P}/\mathbf{U}, \mathbf{V}$ - Phrase/Input occurrence matrices

The linear Full Additive composition model

Guevara GEMS 2010, Zanzotto et al. COLING 2010

- Given two word vectors $\overrightarrow{u}$ and $\overrightarrow{v}$ in syntactic relation R compute phrase vector $\overrightarrow{p}$

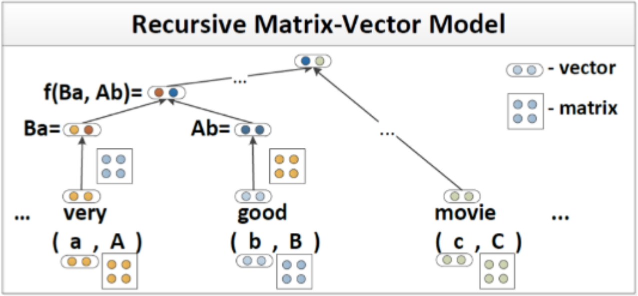
$$\overrightarrow{p} = \mathbf{A}_R\overrightarrow{u} = \mathbf{B}_R\overrightarrow{v} = [\mathbf{A}_R, \mathbf{B}_R]\begin{bmatrix}\overrightarrow{u}\\ \overrightarrow{v}\end{bmatrix}$$

- Parameters: syntax-dependent matrices $\mathbf{A}_R$ and $\mathbf{B}_R$
- General estimation from corpus-extracted phrase and word vectors as least-squares regression problem:

$$\underset{\mathbf{A}_R,\mathbf{B}_R}{\operatorname{argmin}} \|\mathbf{P} - [\mathbf{A}_R, \mathbf{B}_R]\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}\|^2$$

Composition in Neural Models

Socher et al. (2012, 2013)



- assigning a vector and a matrix to every word
- learning an input-specific, nonlinear, compositional function for computing vector and matrix representations for multi-word sequences of any syntactic type

Functional composition in morphology

Lazaridou et al. ACL 2013, Marelli & Baroni PsychRev 2015

word	*nearest neighbors*
carve.er	potter, engraver, goldsmith
broil.er	oven, stove, cooking, kebab, done
column	arch, pillar, bracket, numeric
column.ist	publicist, journalist, correspondent
industry.al	environmental, land-use, agriculture
industry.ous	frugal, studious, hard-working
nervous	anxious, excitability, panicky
nerve.ous	bronchial, nasal, intestinal

Phrase similarity data

Mitchell & Lapata (2008, 2010), Grefenstette and Sadrzadeh (2011)

AN	national government	cold air	1
	new information	further evidence	6
NN	environment secretary	party leader	5
	telephone number	future development	2
VO	offer support	provide help	7
	fight war	win battle	5

Phrase similarity data

Mitchell & Lapata (2008, 2010), Grefenstette and Sadrzadeh (2011)

AN	national government	cold air	1
	new information	further evidence	6
NN	environment secretary	party leader	5
	telephone number	future development	2
VO	offer support	provide help	7
	fight war	win battle	5
SV	fire glows	fire burns	6
	face glows	face burns	1
	discussion stray	discussion digresses	7
	child strays	child digresses	2
SVO	table shows result	table expresses result	7
	map shows location	map expresses location	1

Similarity intuitions (often) affected by verb-argument interactions

Results

Rank correlation (ρ) with subject scores

	SV	SVO
Verb only	0.06	0.08
Vector addition	0.13	0.12
Functional approach	**0.23**	**0.32**
Human	0.40	0.62

Sentence Similarity Data

- Semantic Textual Similarity (STS) datasets from SEMEVAL
- MSR Par dataset (1,500 pairs):
  - The fines are part of failed Republican efforts to force or entice the Democrats to return.
  - Perry said he backs the Senates efforts, including fines, to force the Democrats to return. 2.8
  - The bill says that a woman who undergoes such an abortion couldn't be prosecuted.
  - A woman who underwent such an abortion could not be prosecuted under the bill. 5.0

SICK: the Turing Test of compositional semantics

Marelli et al. 2014, 10K sentence pairs

sentence pair	**relatedness**	**entailment**
two men are taking a break from a trip on a snowy road two men are taking a break from a trip on a road covered by snow	4.9	A entails B
the girl is spraying the plants with water the girl is watering the plants	4.6	A entails B
the turtle is following the fish the fish is following the turtle	3.8	A contradicts B
the girl is spraying the plants with water the boy is spraying the plants with water	3.4	neutral
masked people are looking in the same direction in a forest a little girl is looking at a woman in costume	1.3	neutral

SICK Performance

Marelli et al. 2014

- Entailment: evaluated through classification accuracy wrt majority annotation
- Relatedness: evaluated through Pearson r with averaged subject rating

Model	relatedness	entailment
Majority baseline	NA	57%
Vector addition	**0.70**	**74%**
Functional approach	0.57	72%

What's going on?

- Word order is largely redundant
- Proportion of times a word sequence appears in more thanone order in the British National Corpus (100M words ofwritten and spoken English): **0.1%**
  - (Counting only sequences that form full sentences)
- Even in these cases, meaning is rarely deeply affected:
  - *however this is not the case*
    his however is not the case
  - *yesterday Mr. Andrews said it will never go away*
    Mr. Andrews said yesterday it will never go away
  - *no thank you I'm fine*
    no I'm fine thank you

What's going on?

Context-based representations might capture typical syntactic roles of words

Every boy in the country will be kicking a soccer ball about.

A man and a boy were kicking a football through the foot-high grass.

The boys were kicking a cheap rubber ball.

The only variation was during the first ten days, when players were not allowed to kick a ball.

After a few laps of the track we could kick a ball about or even have a go at throwing a javelin.

Popular tasks and core sentence meaning

1. Paraphrasing

   A woman cuts up broccoli.
   A woman is cutting broccoli.

   A woman is slipping in the water-tub.
   A woman is lying in a raft.

2. Sentiment analysis
3. Question Answering
4. Entailment (RTE4, SICK)
5. Modeling relations between sentences

Optional Assignment: Start composing!

- Get to know the DISSECT toolkit[2] (python)
  - Install the toolkit (link in course website)
  - Follow the tutorial on course website to become familiar with composition functions
  - Complete assignment posted online

[2] G. Dinu, N. The Pham, and M. Baroni.2013. DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of the System Demonstrations of ACL 2013*, Sofia, Bulgaria.

Suggested Readings

- Background Readings
  - Baroni et al. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vector
  - Mikolov et al. (2013). Efficient Estimation of Word Representations in Vector Space
  - Mikolov et al. (2013). Linguistic Regularities in Continuous Space Word Representations
  - Levy et al. (2015) Improving Distributional Similarity with Lessons Learned from Word Embeddings

- Readings
  - Socher et al. (2012). Semantic Compositionality through Recursive Matrix-Vector Spaces (Slides)
  - Levy & Goldberg (2014, CoNLL best paper) Linguistic Regularities in Sparse and Explicit Word Representations (Slides)
  - Moritz Hermann & Blunsom (2014, ACL). Multilingual Models for Compositional Distributed Semantics (Slides)
  - Faruqui et al. (2015, best paper at NAACL). Retrofitting Word Vectors to Semantic Lexicons
  - Norouzi et al. (2014, ICLR) Zero-Shot Learning by Convex Combination of Semantic Embeddings (Slides)