Lecture 8: Multimodal semantics & compositional semantics

Multimodal distributional semantics

Compositional semantics

Compositional distributional semantics

Uses of word clustering and selectional preferences

Widely used in NLP as a source of lexical information:

- Word sense induction and disambiguation
- Parsing (resolving ambiguous attachments)
- Identifying figurative language and idioms
- Paraphrasing and paraphrase detection
- Used in applications directly, e.g. machine translation, information retrieval etc.

Outline.

Multimodal distributional semantics

Compositional semantics

Compositional distributional semantics

Multimodal semantics

Intuition: Humans learn word meanings from linguistic, perceptual and sensory-motor experience

This includes:

- linguistic input (text or speech)
- visual input (images and videos)
- other sensory modalities: taste, smell, touch etc.
- motor control and its simulation

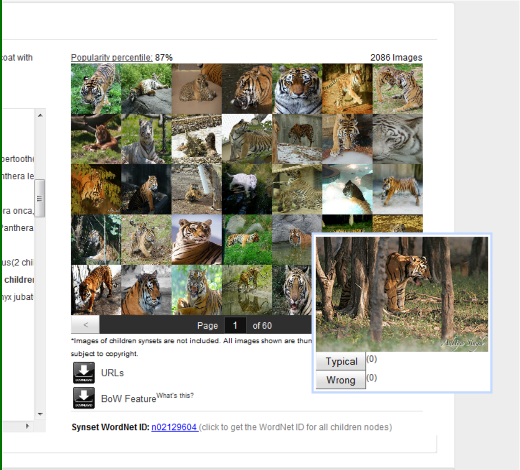
Multimodal semantics in NLP today mainly focuses on building word representations from text, images and (recently) videos.

Obtaining language+vision representations

1. Need a visual corpus
   - ImageNet
   - Yahoo! Webscope Flickr 100M
   - etc.
   - ...*or* use an image search engine
2. Need a way to extract visual features:
   - bag-of-visual-words models
   - convolutional neural networks (CNNs)
3. Need a way of combining visual and linguistic information
   - various fusion strategies

ImageNet

- Animals
  - Birds
  - Fish
  - Mammal
  - Invertebrate
- Scenes
  - Indoor
  - Geological formations
- Sport activities
- Materials and fabric
- Instrumentation
  - Tools
  - Appliances
  - …
- Plants
  - …



Slide credit: Fei-Fei Li

Bag-of-visual-words models

Elia Bruni, Nam Khanh Tran and Marco Baroni (2014).
Multimodal distributional semantics.

General intuition:

- inspired by bag-of-words
- train on a corpus of images, e.g. ImageNet
- break images into discrete parts — visual words
- ignore the structure
- represent words as vectors of visual words

Obtaining visual words

Given a corpus of images:

- Identify keypoints (corner detection, segmentation)
- Represent keypoints as vectors of descriptors (SIFT)
- Cluster keypoints to obtain visual words
- Bag of visual words – ignore the location

Representing linguistic concepts

- Retrieve images for a given word, e.g. *dog* (from a corpus or the Web)
- identify keypoints in each of the images
- map to visual words
- represent words as vectors of co-occurrence with visual words

Combining text and visual words

Example task: word similarity estimation, e.g. using *cosine*

1. Feature level fusion:
    - concatenate textual and visual feature vectors
    - dimensionality reduction (some approaches) – map the features into the same low dimensional space, e.g. using SVD or NMF
    - estimate similarity of the vectors
2. Scoring level fusion:
    - estimate similarity for textual and visual vectors separately
    - take a mean of the similarity scores

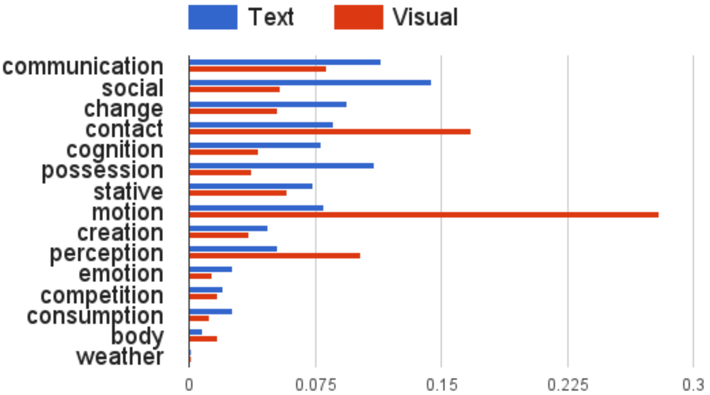
Tasks and applications

- word similarity estimation
- predicting concreteness (via image-dispersion)
- selectional preference induction
- bilingual lexicon induction
- metaphor detection
- lexical entailment $\approx$ hypernym identification

Multimodal models outperform the linguistic ones in all of these!
But...

- work quite well for nouns and adjectives
- more difficult to extract visual features for verbs

How is visual data different from linguistic data?



Verb classes in Yahoo! Webscope Flickr 100M and BNC corpora

Biases in the data

- Textual corpora: abstract events and topics
- Image corpora: concrete events / actions, also topic bias
- Videos: extended actions, states

The next big questions

1. What semantic information do we learn from the images?
2. Which words benefit from visual information?
3. Other modalities:
   - auditory and olfactory perception (some work done)
   - motor control — really tough one!

Outline.

Multimodal distributional semantics

Compositional semantics

Compositional distributional semantics

Compositional semantics

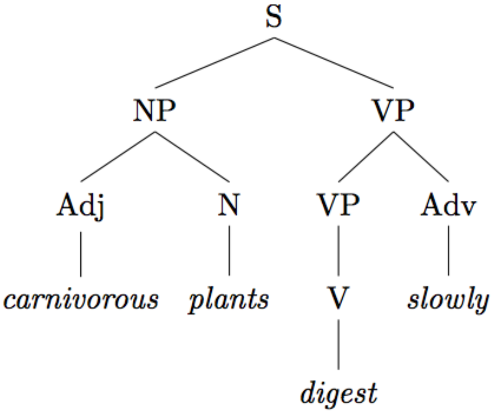
- Principle of Compositionality: meaning of each whole phrase derivable from meaning of its parts.
- Sentence structure conveys some meaning
- Formal semantics: sentence meaning as logical form

  Kitty chased Rover.
  Rover was chased by Kitty.

  $$\exists x, y[\text{chase}'(x, y) \wedge \text{Kitty}'(x) \wedge \text{Rover}'(y)]$$

  or $\text{chase}'(k, r)$ if k and r are constants (*Kitty* and *Rover*)
- Deep grammars: model semantics alongside syntax, one semantic composition rule per syntax rule

Compositional semantics alongside syntax



Semantic composition is non-trivial

- Similar syntactic structures may have different meanings:

  it barks
  it rains; it snows – pleonastic pronouns

- Different syntactic structures may have the same meaning:

  Kim seems to sleep.
  It seems that Kim sleeps.

- Not all phrases are interpreted compositionally, e.g. idioms:

  red tape
  kick the bucket

  but they can be interpreted compositionally too, so we can not simply block them.

Semantic composition is non-trivial

- Elliptical constructions where additional meaning arises through composition, e.g. logical metonymy:

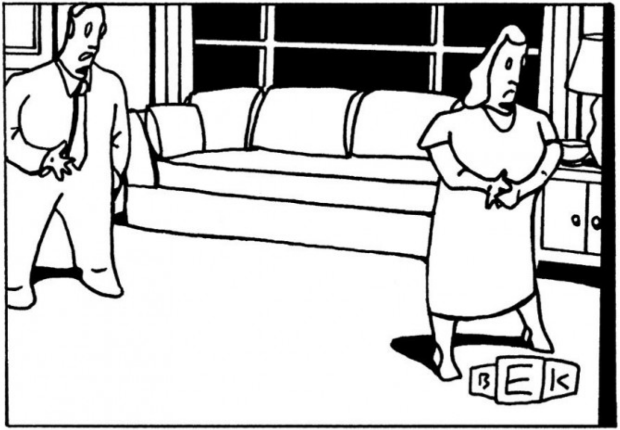
  fast programmer
  fast plane

- Meaning transfer and additional connotations that arise through composition, e.g. metaphor

  *I cant **buy** this story.*
  *This sum will **buy** you a ride on the train.*

- Recursion

Recursion



"Of course I care about how you imagined I thought you perceived I wanted you to feel."

Outline.

Multimodal distributional semantics

Compositional semantics

Compositional distributional semantics

Compositional distributional semantics

Can distributional semantics be extended to account for the meaning of phrases and sentences?

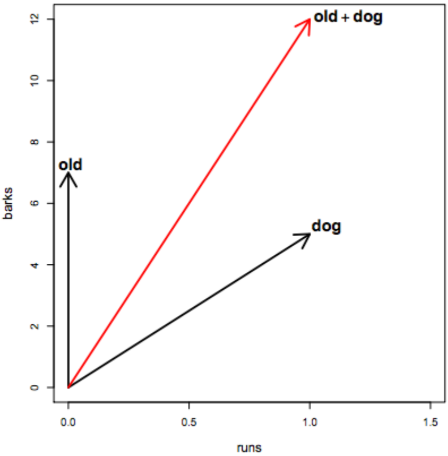
- Language can have an infinite number of sentences, given a limited vocabulary
- So we can not learn vectors for all phrases and sentences
- and need to do composition in a distributional space

1. Vector mixture models

Mitchell and Lapata, 2010. *Composition in Distributional Models of Semantics*

Models:

- Additive
- Multiplicative



Additive and multiplicative models

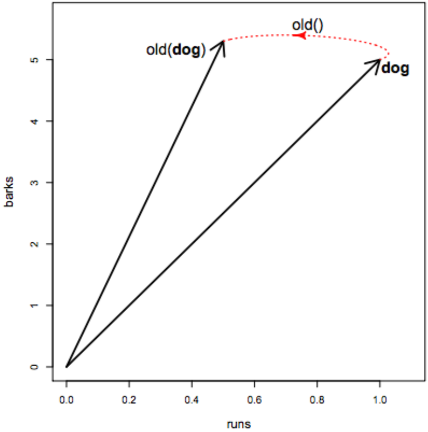
				additive		multiplicative	
	dog	**cat**	**old**	**old + dog**	**old + cat**	**old ⊙ dog**	**old ⊙ cat**
runs	1	4	0	1	4	0	0
barks	5	0	7	12	7	35	0

- correlate with human similarity judgments about adjective-noun, noun-noun, verb-noun and noun-verb pairs
- but... commutative, hence do not account for word order
  John hit the ball = *The ball hit John*!
- more suitable for modelling content words, would not port well to function words:
  e.g. *some dogs; lice and dogs; lice on dogs*

2. Lexical function models

Distinguish between:

- words whose meaning is directly determined by their distributional behaviour, e.g. nouns
- words that act as functions transforming the distributional profile of other words, e.g., verbs, adjectives and prepositions



Lexical function models

Baroni and Zamparelli, 2010. *Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space*

Adjectives as lexical functions

$$\textit{old dog} = \textit{old}(\textit{dog})$$

- Adjectives are parameter matrices ($\mathbf{A}_{old}$, $\mathbf{A}_{furry}$, etc.).
- Nouns are vectors (**house**, **dog**, etc.).
- Composition is simply **old dog** $= \mathbf{A}_{old} \times$ **dog**.

OLD	runs	barks
runs	0.5	0
barks	0.3	1

$\times$

	dog
runs	1
barks	5

$=$

	OLD(dog)
runs	$(0.5 \times 1) + (0 \times 5) = 0.5$
barks	$(0.3 \times 1) + (5 \times 1) = 5.3$

Learning adjective matrices

1. Obtain a distributional vector $\mathbf{n}_j$ for each noun n_j in the lexicon.
2. Collect adjective noun pairs (a_i, n_j) from the corpus.
3. Obtain a distributional vector $\mathbf{p}_{ij}$ of each pair (a_i, n_j) from the same corpus using a conventional DSM.
4. The set of tuples $\{(\mathbf{n}_j, \mathbf{p}_{ij})\}_j$ represents a dataset $\mathcal{D}(a_i)$ for the adjective a_i.
5. Learn matrix $\mathbf{A}_i$ from $\mathcal{D}(a_i)$ using linear regression.

Minimize the squared error loss:

$$L(\mathbf{A}_i) = \sum_{j \in \mathcal{D}(a_i)} \|\mathbf{p}_{ij} - \mathbf{A}_i \mathbf{n}_j\|^2$$

Polysemy in lexical function models

Generally:

- use single representation for all senses
- assume that ambiguity can be handled as long as contextual information is available

Exceptions:

- Kartsaklis and Sadrzadeh (2013): homonymy poses problems and is better handled with prior disambiguation
- Gutierrez et al (2016): literal and metaphorical senses better handled by separate models
- However, this is still an open research question.