Distributional Lexical Semantics

Bridging the gap between semantic theory and computational simulations

Marco Baroni
http://clic.cimec.unitn.it/marco/

Stefan Evert
http://purl.org/stefan.evert/

Alessandro Lenci
http://www.ilc.cnr.it/~lenci/

What is lexical meaning?

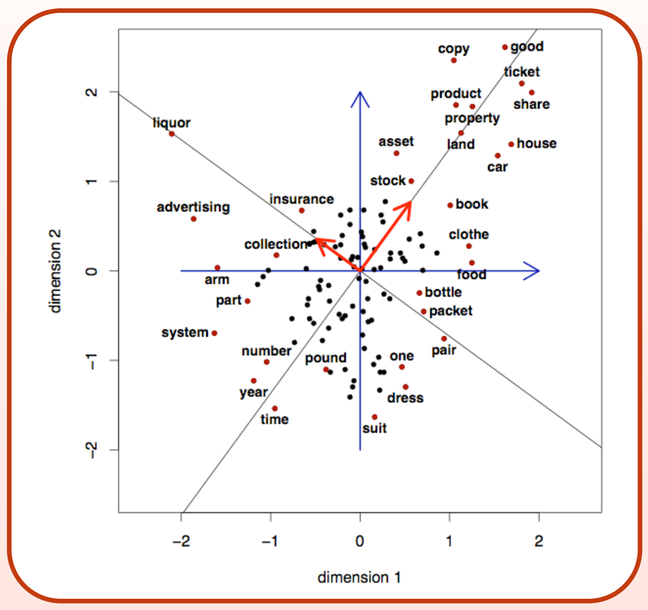
- Compositional semantics vs. lexical semantics
- Is linguistic meaning just a pointer from lexeme (after disambiguation) to a non-linguistic concept?
  - i.e., is semantic linguistic knowledge the same as conceptual knowledge?

$[\![\text{dog}_1]\!] =$

- What is a suitable formal representation of conceptual knowledge?
  - truth-conditional semantics
  - ontology of concepts and their semantic relations
  - definition in terms of characteristic properties
  - conceptual spaces (Gärdenfors), semantic differentials (Osgood)
  - neural activation patterns & associative memory

Word Space

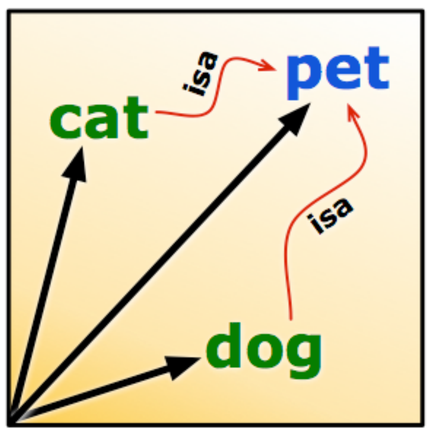
- Distributional hypothesis (Harris, Firth)
  - words with similar meaning should occur in similar contexts
  - can also be applied to non-linguistic concepts (and contexts)
- Vector space model based on
  - documents or textual units
  - cooccurring words
  - syntactic relations
  - lexico-grammatical patterns
- Model parameters
  - feature scaling & transform
  - metric (distance measure)
  - dimensionality reduction (SVD, RI, feature selection)
- Evaluation of word space
  - good at synonym test (TOEFL)
  - semantic fields & neighbours
  - technical applications: part-of-speech induction, word sense disambiguation and induction, information retrieval, language modelling, …



Distributional Lexical Semantics Workshop
(Hamburg, 4-9 Aug 2008)

Marco Baroni, Stefan Evert, Alessandro Lenci

- Link computational word space models to theoretical questions
- Key question: what kinds of semantic / conceptual knowledge do word space models encode?
  - semantic relations & ontological classification
  - defining properties of concepts
  - human free associations
- Non-competitive shared tasks (deadline: 4 Apr 2008)
  - analysis of word space & consequences for semantic theory
- Plausible model of human semantic knowledge should solve all tasks with uniform representation ➡ not a ML exercise



- Other interesting questions (topics for future research)
  - language/domain dependency
  - robust multi-purpose models
  - incrementality (↗LSA)
  - integrated sense disambiguation
  - represent compositional meaning of complex expression
  - context-dependence (linguistic and non-linguistic context)
  - polysemy

➲ **wordspace.collocations.de**

Shared tasks

- Task 1: **Free association**
  - Is word space a model of human associative memory?
  - Operationalisation: psychological association norms (USF, EAT)
    - discrimination between strongly associated words and random pairs
    - correlation with human norms (relative frequency of response)
    - prediction of most frequent human response for given cue word
  - Expected to require integration of first-order statistical associations (collocations) with higher-order associations (word space)

- Task 2: **Categorisation**
  - Classification of words/concepts into semantic categories
    - natural categories of concrete nouns (*animals*, *vegetables*, *artefacts*, …)
    - abstract vs. concrete nouns
    - semantic categories of verbs (based on Levin 1993)
  - Word space models used for clustering ➡ evaluate cluster purity

- Task 3: **Property generation**
  - Concepts characterised by certain typical properties
    - robust evidence from psychological property generation experiments
    - *e.g.* *dog* ➡ *barks*, *has a tail*, *is a pet/animal*, *is man's best friend*, …
  - Word space model must predict top 10 properties generated by human subjects (McRae et al. 2005)
    - either associated words or highly loaded features of the cue word
    - fuzzy comparison of properties: *has hair* matches *hair*, *hairy* and *fuzz*