Introduction to the Shared Tasks

Marco Baroni, Stefan Evert and Alessandro Lenci

ESSLLI Distributional Semantics Workshop

Hamburg, August 5 2008

From word distributions to meaning

- Corpus-based Semantic Models (CSMs) have been claimed to be plausible models of human knowledge organization and learning:
    - “The dimensionality–optimizing method offers a promising solution to the ancient puzzle of human knowledge induction. It still remains to determine how wide its scope is among human learning and cognition phenomena. [...] We would suggest that applications to problems in conditioning, association, pattern and object recognition, contextual disambiguation, metaphor, concepts and categorization, reminding, casebased reasoning, probability and similarity judgment, and complex stimulus generalization are among the set where this kind of induction might provide new solutions” (Landauer and Dumais 1997: 235)

CSMs in cognitive research

- Measures of semantic similarity based on CSMs have been demonstrated to predict behavioral performance in various tasks
    - synonymy identification (Landauer and Dumais 1997)
    - text coherence (Landauer and Dumais 1997)
    - categorization (Burgess and Lund 1997)
    - semantic priming in lexical decision tasks (Lowe 2000, McDonald and Brew 2002, Vigliocco *et al.* 2004)
    - word substitution errors (Vigliocco *et al.* 2004)
    - child vocabulary acquisition (Li *et al.* 2004, Baroni *et al.* 2007)
    - etc.

Why yet another shared task?

- Careful and extensive empirical tests of the cognitive and linguistic plausibility of CSMs are still lacking
- Evaluation campaigns for semantics exist in NLP (cf. SEMEVAL), but

Why yet another shared task?

- Careful and extensive empirical tests of the cognitive and linguistic plausibility of CSMs are still lacking
- Evaluation campaigns for semantics exist in NLP (cf. SEMEVAL), but
  - focus on large-scale quantitative tasks

Why yet another shared task?

- Careful and extensive empirical tests of the cognitive and linguistic plausibility of CSMs are still lacking
- Evaluation campaigns for semantics exist in NLP (cf. SEMEVAL), but
    - focus on large-scale quantitative tasks
    - are oriented towards engineering applications rather than to linguistic and cognitive issues

Why yet another shared task?

- Careful and extensive empirical tests of the cognitive and linguistic plausibility of CSMs are still lacking
- Evaluation campaigns for semantics exist in NLP (cf. SEMEVAL), but
  - focus on large-scale quantitative tasks
  - are oriented towards engineering applications rather than to linguistic and cognitive issues
  - tend to focus on just one aspect, i.e. synonym identification

CSMs and synonym identification

- Synonym identification on the TOEFL (Test of English as a Foreign Language) test set is THE standard task for CSMs evaluation
- 80 items for which subjects must select the correct synonym among 4 candidates
  - target = furnish answers={supply, impress, protect, advise}
  - target = physician answers={chemist, pharmacist, nurse, doctor}
- Rapp (2004)
  - LSA: 92%
  - natives: 97.75%

Why yet another shared task?

- Human lexical semantic competence is a composition of multi-faceted abilities
  - classifying entities
  - describing their properties
  - recognizing the similarities among meanings
  - building the interpretation of complex expressions via composition of lexical meanings
  - drawing inferences, etc.
- different areas of the lexicons may dramatically vary for the semantic dimensions relevant for their organization
  - nouns: taxonomy, meronymy, functionality, etc.
  - verbs: causation, telicity, agency, manner, etc.

A shared task for CSMs

- Research teams have been invited to test their computational models on a variety of cognitively plausible semantic tasks
  - parallel to those that cognitive scientists use to design behavioral experiments aimed at investigating the human semantic memory
  - data set were extracted from resources commonly used for psycholinguistic experiments
  - linguistically relevant criteria were also taken into account
  - hard cases (e.g. polysemous words) were not avoided
- The goals of the shared task

A shared task for CSMs

- Research teams have been invited to test their computational models on a variety of cognitively plausible semantic tasks
  - parallel to those that cognitive scientists use to design behavioral experiments aimed at investigating the human semantic memory
  - data set were extracted from resources commonly used for psycholinguistic experiments
  - linguistically relevant criteria were also taken into account
  - hard cases (e.g. polysemous words) were not avoided
- The goals of the shared task
  - NOT competition!

A shared task for CSMs

- Research teams have been invited to test their computational models on a variety of cognitively plausible semantic tasks
  - parallel to those that cognitive scientists use to design behavioral experiments aimed at investigating the human semantic memory
  - data set were extracted from resources commonly used for psycholinguistic experiments
  - linguistically relevant criteria were also taken into account
  - hard cases (e.g. polysemous words) were not avoided
- The goals of the shared task
  - NOT competition!
  - understanding how different models highlight different semantic aspects

A shared task for CSMs

- Research teams have been invited to test their computational models on a variety of cognitively plausible semantic tasks
  - parallel to those that cognitive scientists use to design behavioral experiments aimed at investigating the human semantic memory
  - data set were extracted from resources commonly used for psycholinguistic experiments
  - linguistically relevant criteria were also taken into account
  - hard cases (e.g. polysemous words) were not avoided
- The goals of the shared task
  - NOT competition!
  - understanding how different models highlight different semantic aspects
  - evaluating how far we are from an integrated model

A shared task for CSMs

- Research teams have been invited to test their computational models on a variety of cognitively plausible semantic tasks
  - parallel to those that cognitive scientists use to design behavioral experiments aimed at investigating the human semantic memory
  - data set were extracted from resources commonly used for psycholinguistic experiments
  - linguistically relevant criteria were also taken into account
  - hard cases (e.g. polysemous words) were not avoided
- The goals of the shared task
  - NOT competition!
  - understanding how different models highlight different semantic aspects
  - evaluating how far we are from an integrated model
  - evaluating which aspects of semantics are beyond the reach of purely distributional approaches

The tasks

1. Modelling free association
2. Categorization
   - a Concrete nouns categorization
   - b Abstract/concrete nouns discrimination
   - c Verb categorization
3. Generation of salient properties of concepts

1. Modelling free association

- In psychology, free associations are the first words that come to the mind of a native speaker when presented with a stimulus word
  - stimulus (cue): *saddle*; response (target): *horse*
- They provide a window to investigate the mechanisms underlying the organization of the semantic memory
  - responses reflect the patterns of interconnection within the lexical-conceptual system

1. Modelling free association

- In psychology, free associations are the first words that come to the mind of a native speaker when presented with a stimulus word
    - stimulus (cue): *saddle*; response (target): *horse*
- They provide a window to investigate the mechanisms underlying the organization of the semantic memory
    - responses reflect the patterns of interconnection within the lexical-conceptual system
- Free association are measured with association norms
    - native speakers are presented with stimulus words and are asked to write down the first word that comes to mind for each stimulus
    - the strength of association between a stimulus (S) and response (R) is quantified by the percentage of test subjects who produced R when presented with S

- Co-occurrence hypothesis (Miller, 1969; Spence and Owens, 1990, Schulte im Walde and Melinger, in press)
  - semantic association is related to the textual co-occurrence of the stimulus-response pairs
    - first-order co-occurrence (collocations) – stimulus: *morse*; response: *code*
    - higher-order co-ccurrence (distributional similarity) – stimulus: *keep*; response: *retain*
- Evaluate free associations is a straightforward “baseline” interpretation of distributional similarity
  - association norms like CSMs produce a quantitative analysis of the association strength between word pairs
  - qualitative analysis of the type of the association between the words is missing

1. Modelling free association

Free associations and CSMs

- Three subtasks to test CSMs with free associations
  - a discrimination
  - b correlation
  - c response prediction
- For each subtask, training and test data were extracted from the Edinburgh Associative Thesaurus (EAT)

1. a – Free association discrimination

Description

- The goal is to discriminate between strongly associated and non-associated stimulus-response pairs
- Data were randomly sampled from three groups
  - **FIRST**: frequent first responses (given by more than 50% of test subjects) as strongly associated pairs
  - **HAPAX**: responses that were produced by a single test subject;
  - **RANDOM**: random combinations of headwords from the EAT that were never produced as a cue-target pair (in any direction)
- The task is to discriminate between the FIRST category and the other two
- Evaluation
  - classification accuracy (baseline is 66%)

1. a – Free association discrimination

Data set sample

- 3 X 100 stimulus–response pairs

stimulus	response	TYPE
retain	keep	FIRST
transplant	heart	FIRST
enquire	ask	FIRST
mobile	immobile	HAPAX
inhuman	violent	HAPAX
peace	piece	HAPAX
eventual	picket	RANDOM
coleman	collect	RANDOM
beatles	fork	RANDOM

1. b – Free association correlation

Description

- The goal is to use CSMs to predict free association strength for a given list of stimulus-response pairs.
- Association strength ranges from 0 to 1 (the highest value in the EAT is 0.91)
  - association strength in the data set is uniformly distributed across the full range
- Evaluation
  - linear correlation (Pearson) and rank correlation (Kendall) between predictions and the gold standard

1. b – Free association correlation

Data set sample

- 240 stimulus–response pairs

stimulus	response	ass. strength
morse	code	0.9082
ding	dong	0.8800
donor	blood	0.8788
holster	gun	0.8300
peel	orange	0.5464
grime	dirt	0.5408
similar	alike	0.1237
lettuce	green	0.1224
sweater	wool	0.1212

1. c – Response prediction

Description

- The goal is to use CSMs to predict the most frequent responses for a given list of stimulus words
- Stimuli were randomly selected from entries in the EAT database with a clearly preferred response
  - the association strength of the dominant response is ≥ 0.4, and at least three times as high as that of the second response
- Evaluation
  - CSMs can suggest up to 100 response candidates for each cue
  - the model score is the average rank of the correct response

1. c – Free association prediction

Data set sample

- 200 stimulus–response pairs

stimulus	response	strength 1^{st} resp.	strength 2^{nd} resp.
ache	pain	0.6162	0.0808
adequate	enough	0.4592	0.1122
adult	child	0.4583	0.0521
affair	love	0.4000	0.0632
aged	old	0.5579	0.0316
alter	change	0.5361	0.0928
amusing	funny	0.6224	0.0714
anatomy	body	0.4062	0.0729
apology	sorry	0.4792	0.0625

2. Categorization

- In categorization tasks, subjects are typically asked to assign experimental items - objects, images, words - to a given category or to group together items belonging to the same category
  - categorization presupposes an understanding of the relationship between the items in a category
- Categorization is a basic cognitive operation presupposed by further semantic tasks
  - inference
    - if X is a CAR then X is a VEHICLE
  - compositionality
    - $\lambda y : \mathit{FOOD} \lambda x : \mathit{ANIMATE}(\mathit{eat}, x, y)$
- “Chicken-and-egg” conundrum in the relationship between categorization and similarity (cf. Goodman 1972, Medin et al. 1993)

2. Categorization

- Categorization is operationalized as a clustering task
  - the recommended clustering algorithm was the *repeated bisections* algorithm implemented in CLUTO (Karypis, 2003)
- Clusters evaluation
  - entropy – whether words from different classes are represented in the same cluster (best = 0)
  - purity – degree to which a cluster contains words from one class only (best = 1)

Task 2.a - Concrete nouns categorization

Data set

- 44 concrete nouns belonging to 6 semantic classes
  - stimuli were extracted from McRae et al. (2005) Semantic Norms

Natural

bird

- chicken
- eagle
- duck
- swan
- owl
- penguin
- peacock

groundAnimal

- dog
- elephant
- cow
- cat
- lion
- pig
- snail
- turtle

fruitTree

- cherry
- banana
- pear
- pineapple

green

- mushroom
- corn
- lettuce
- potato
- onion

Task 2.a - Concrete nouns categorization

Data set

Artifact

tool

- bottle
- pencil
- pen
- cup
- bowl
- scissors
- kettle
- knife
- screwdriver
- hammer
- spoon
- chisel
- telephone

vehicle

- boat
- car
- ship
- truck
- rocket
- motorcycle
- helicopter

Task 2.b - Abstract/concrete nouns discrimination

Description

- Behavioral and neuropsychological evidence suggests that abstract and concrete concepts might be represented, retrieved and processed differently in the human brain
- Test data were extracted from the Medical Research Council (MRC) Psycholinguistic Database (Coltheart 1981)
  - CONC index in MRC (ranging from 100 to 700) summarizes the subjects' judgment about noun concreteness
  - high stability and replicability of the abstract vs. concrete discrimination judgements
- Abstract have a loose connection with perceptual dimensions and most of their semantics is probably distributionally driven
  - hard cases for embodied theories of concepts (cf. Barsalou, Glenberg, etc.)

Task 2.b - Abstract/concrete nouns discrimination

Data set

- 40 nouns divided into three "concreteness classes" (HIGH, MEDIUM and LOW)

High concreteness

- chicken
- eagle
- lion
- turtle
- banana
- onion
- potato
- bowl
- pencil
- telephone
- truck
- ship
- car
- bottle
- hammer

Task 2.b - Abstract/concrete nouns discrimination

Data set

Medium concreteness

- pollution
- invitation
- shape
- empire
- foundation
- fight
- smell
- ache
- ceremony
- weather

Task 2.b - Abstract/concrete nouns discrimination

Data set

Low concreteness

- jealousy
- truth
- hypothesis
- hope
- mercy
- mistery
- gratitude
- concept
- temptation
- pride
- belief
- insight
- wisdom
- luck
- distraction

Task 2.c - Verb categorization

Data set

- 45 verbs belonging to 9 semantic classes
  - classes are adapted from Vinson and Vigliocco 2007 and are consitent with well-known linguistic classifications (cf. Levin, WordNet, etc.)
  - verb classifications are highly multidimensional and controversial!

Cognition

communication

- suggest
- talk
- speak
- request
- read

mentalState

- evaluate
- remember
- know
- forget
- check

Task 2.c - Verb categorization

Data set

Motion

motionManner

- run
- fly
- drive
- walk
- ride

motionDirection

- arrive
- enter
- fall
- rise
- leave

changeLocation

- carry
- push
- move
- send
- pull

Task 2.c - Verb categorization

Data set

Body

bodySense

- listen
- smell
- feel
- look
- notice

bodyAction

- eat
- breathe
- drink
- smile
- cry

Task 2.c - Verb categorization

Data set

exchange

- acquire
- lend
- buy
- sell
- pay

changeState

- kill
- destroy
- repair
- die
- break

Task 3 - Property generation

- Concepts and meanings are commonly regarded as complex assemblies of properties (cf. semantic features, Qualia, etc.)
- Subjects show a remarkable degree of agreement in tasks that require enumerating the typical properties of a concept: *a dog* barks, has a tail, is a pet, etc.
- Psychologists have been collecting feature norms, i.e., speaker-generated lists of concepts described in terms of properties (cf. Mc Rae et al. 2005)

property	type	freq.
a_vehicle	superordinate	9
has_4_wheels	external_component	18
is_fast	systemic_property	9
used_for_transportation	function	19

Task 3 - Property generation

Task operationalization

- The goal is to compare CSMs with speaker-generated norms to evaluate their ability to characterize the properties of a target concept
- 44 target concepts (the same used in the Concrete Nouns categorization task)
  - the gold standard for the evaluation is represented by the top 10 properties for each concept in the McRae norms
- Evaluation
  - given the ranked output of a model, we compute precision for each concept with respect to this gold standard, at various *n*-best thresholds, and we average precision across the 44 concepts.
    - precision – number of properties produced by a model that match a property in the gold standard, out of the *k*-threshold (with k = 10, 20 or 30)

Task 3 - Property generation

Property expansion

- The word produced by the models might be just a variant of the property in the norms:
  - morphological variants (e.g. *fly* vs. *flying*)
  - (near)-synonyms (e.g. *feather* vs. *plume*)
- For each of the 10 properties of the gold standard we generated an expansion set, i.e. a list of single word expressions that seemed plausible ways to express the relevant property
  - synomyms of the original property were extracted from WordNet
  - inflectional and derivational variants were added

Task 3 - Property generation

Data set sample

duck

- behavior-_swims {swim, swimming, swims}
- has_feathers {feather, feathering, feathers, plumage, plume, plumes}
- has_feet {feet, foot, footed}
- has_webbed_feet {web, web-footed, webbed}
- behavior-_quacks {quack, quacks, quacking}
- has_a_bill {beak, bill, neb, nib, peak, peck, pecker}
- is_edible {comestible, eat, eatable, eaten, edible}
- hunted_by_people {hunt, hunted, hunting, hunts}
- lives_on_water {aquatic, lake, ocean, river, sea, water}
- behavior-_flies {aviate, flies, flight, fly, flying}

Conclusions and open issues

- CSMs need to be carefully evaluated and compared on different tasks
  - to explore the parameter space
  - to substantiate any claim concerning their cognitive plausibility
  - to evaluate their ability to tackle linguistically relevant issues
- Many issues are left out of the picture
  - tasks focussing on meaning compositions (e.g. V + O or Adj + N)
  - extended analysis on PoS differences in CSMs
  - increase the dialogue with formal approaches to meaning

Some references 1

L. Barsalou (2005). Situated conceptualization. In H. Cohen and C. Lefebvre (eds.), *Handbook of Categorization in Cognitive Sciences*, Elsevier: 619-550.

M. Baroni, A. Lenci and L. Onnis (2007). ISA meets Lara: A fully incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the ACL Workshop on Cognitive Aspects of Language Acquisition*, Prague, 29 June 2007.

A. Glenberg and D. Robertson (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, 43: 379-401.

J. Karlgren and M. Sahlgren (2001). From words to understanding. In Y. Uesaka, P. Kanerva and H. Asoh (eds.), *Foundations of real-world intelligence*, CSLI, Stanford: 294-308.

T.K. Landauer and S.T. Dumais (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2): 211-240.

P. Li, C. Burgess and K. Lund (2000). The acquisition of word meaning through global lexical co-occurrences. *Proceedings of the 31st Child Language Research Forum*: 167-178.

Some references 2

S. Schulte im Walde, A. Melinger (in press). “An in-depth look into the co-occurrence distribution of semantic associates”. *Italian Journal of Linguistics*, A. Lenci (ed.), *Special Issues on Distributional Models of Meanings*.

D.P. Spence, K.C. Owens (1990). “Lexical co-occurrence and association strength”. *Journal of Psycholinguistic Research*, 19. 317Ð330.

G. Vigliocco, D. Vinson, W. Lewis and M. Garrett (2004). Representing the meanings of objects and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, 48: 422-488.

G. Vigliocco, D. Vinson (2007). *Semantic Representation*. In G. Gaskell (ed.) *Handbook of Psycholinguistics*. Oxford: Oxford University Press