Cognitive Multimodal Processing: from Signal to Behavior

Alexandros Potamianos

School of ECE, National Technical Univ. of Athens, Greece

3rd Signal Processing Jam
(in honor of Prof. G. Karagiannis)
Athens, January 20, 2015

Acknowledgements

- *Elias Iosif, Georgia Athanasopoulou*: semantic representations, manifold semantic models, semantic-affective autoencoders
- *Nikos Malandrakis*: semantic-affective models, movie emotion tracking
- *Petros Maragos,George Evangelopoulos, Nancy Zlatintsi*: saliency-based video summarization

References

[1] E. Iosif and A. Potamianos. 2010. “Unsupervised semantic similarity computation between terms using web documents”. IEEE Transactions on Knowledge and Data Engineering.
[2] E. Iosif and A. Potamianos. 2013. “Similarity computation using semantic networks created from web-harvested data”. Natural Language Engineering.
[3] N. Malandrakis, A. Potamianos, E. Iosif and S. Narayanan. 2013. “Distributional Semantic Models for Affective Text Analysis”. IEEE Transactions on Audio, Speech and Language Processing.
[4] G. Athanasopoulou, E. Iosif and A. Potamianos. 2014. “Low-Dimensional Manifold Distributional Semantic Models”. In Proc. COLING.
[5] N. Malandrakis, A. Potamianos, G. Evangelopoulos and A. Zlatintsi, 2011. "A supervised approach to movie emotion tracking ,"in Proc. ICASSP
[6] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. 2013. "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," IEEE Transactions on Multimedia.

Talk Outline

- Motivation: cognitive semantic models
- Maxims of interaction
  - Attention and saliency
  - Common ground and concept representations
- From semantics to behavior
  - an example: semantic-affective models
- Dual system processing
- Multimodal fusion
- Grand challenges

Multimodal Signal and Interaction Processing

- From signal to semantics
- From signal to attitudes, behaviors and interaction
  - Affective computing, emotion recognition, sentiment analysis
  - Social signal processing (SSP): personality, status, dominance, persuasion, rapport etc.
  - Behavioral signal processing (BSP): socio-emotional state, cognitive state monitoring
- Challenges:
  1. Define, label and annotate the high-level behaviors associated with interaction (manual)
  2. Devise computational algorithms to analyze, classify or recognize behaviors (automatic)

List of Open Questions

1. How are concepts, features/properties, categories, actions represented?
2. How are concepts, properties, categories, actions combined (compositionally)?
3. How are judgements (classification/recognition decisions) achieved?
4. How is learning and inference (especially induction) achieved?

Answers should fit evidence by psychology and neurocognition!

Three Solutions

- Symbolic
  - cognition is a Turing machine
  - computation is symbol manipulation
  - rule-based, deterministic (typically)
- Associationism, especially, connectionism (ANNs)
  - brain is a neural network
  - computation is activation/weight propagation
  - example-based, statistical, unstructured (typically)
- Conceptual
  - intermediate between symbolic and connectionist
  - concepts are represented as well-behaved (sub-)spaces
  - computation tools: similarity, operators, transformations
  - hierarchical, semi-structured

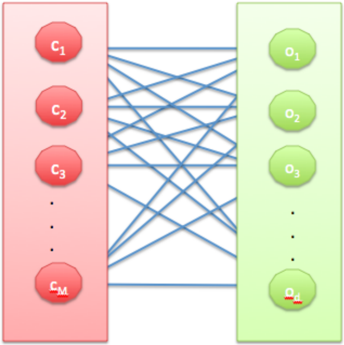
Properties of the Three Approaches

- Symbolic
  - Good for high-level cognitive computations (math)
  - Poor generalization power
  - Too expensive and slow for most cognitive purposes
- Conceptual
  - Excellent generalization power (intuition, physics)
  - Good for induction and learning; geometric properties (hierarchy, low dim., convex) guarantee quick convergence
  - Properties and actions defined as operators/translations
  - Still too slow for some survival-dependent decisions
- Connectionist (machine learning)
  - General-purpose, extremely fast and decently accurate
  - Computational sort-cuts create cognitive biases
  - Poor generalizability power due to high dimensionality and lack of crisp semantic representation

Representation Learning

- Properties of a classifier with good generalization properties [Bengio et al 2013]:
  - Low-dimensionality/Sparseness
  - Distributed representations/hierarchy
    - Depth and abstraction
  - Shared factors across tasks
- Examples: auto-encoders, manifolds, deep neural nets ...

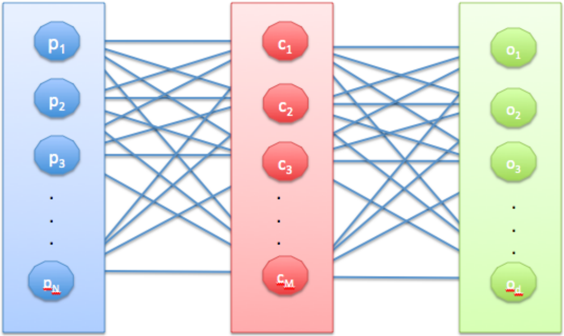
Classification



Class Space (M-dim)

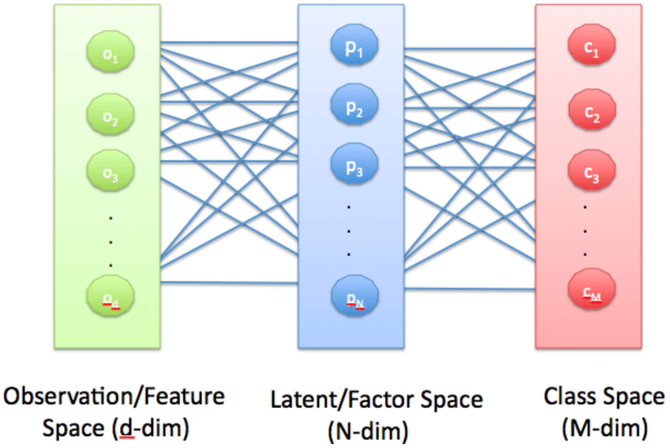
Observation/Feature Space (d-dim)

Latent Spaces and Causality

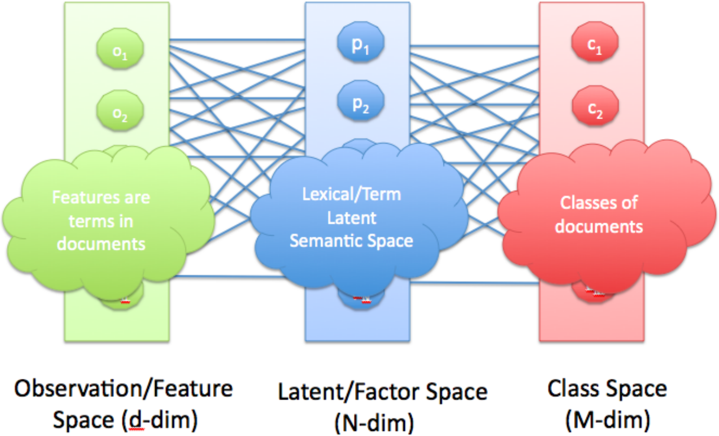


Latent/Factor Space (N-dim)	Class Space (M-dim)	Observation/Feature Space (d-dim)

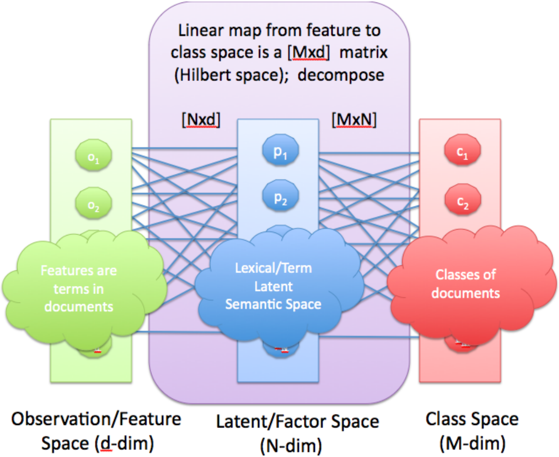
Latent Spaces and Dependencies



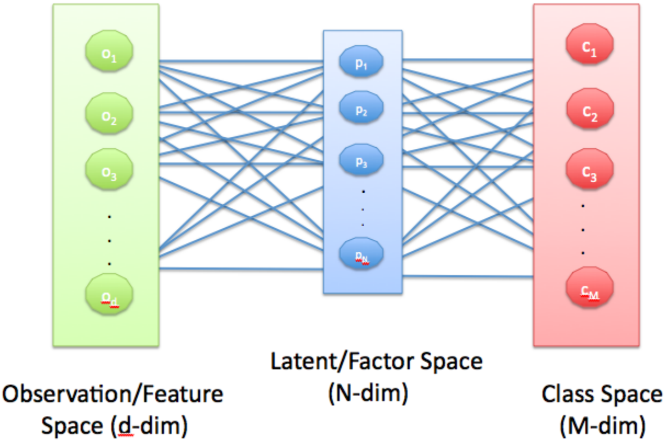
Example: Information Retrieval



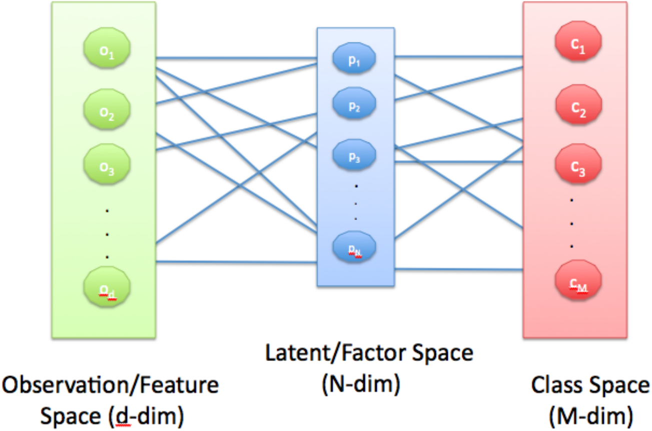
Linear Maps and Matrix Decomposition



Low Dimensionality



Sparsity


o1
o2
o3
od
p1
p2
p3
pN
c1
c2
c3
cM
Observation/Feature Space (d-dim)
Latent/Factor Space (N-dim)
Class Space (M-dim)

Other Common Modeling Mistakes (1)

- Are normed vector spaces (Banach) or Euclidean spaces (Hilbert) good?

  YES Fast convergence properties to unique fixed points
  NO Orthogonality and curse of dimensionality
  NO Tremendous waste of resources

Solution 1: Dimensionality reduction
Solution 2: Manifolds: union of low-dimensional sub-spaces that have good geometric properties
Solution 3: Use deep neural networks

Other Common Modeling Mistakes (2)

- Are all features, classes, latent representation elements born equal?

  YES They are all points in my vector space model

  NO There is hierarchy and abstraction

Solution 1: Use hierarchical models, e.g., hierarchical manifolds, decision trees
Solution 2: Use sets (activation areas) instead of points , e.g., sparse distributed memory
Solution 3: Use deep neural networks

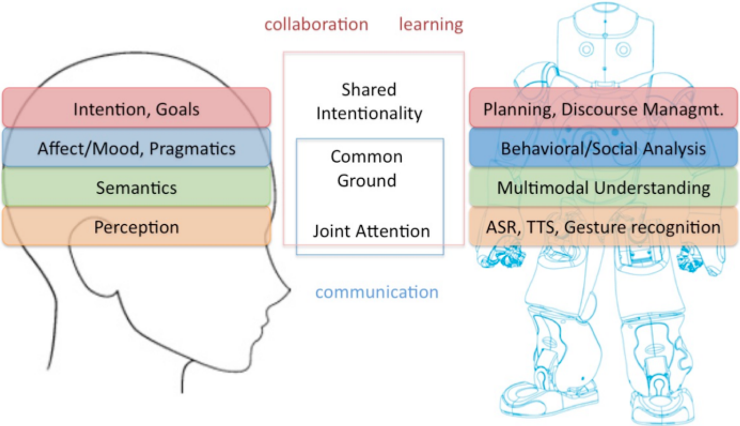
Representation Learning

- Properties of a classifier with good generalization properties [Bengio et al 2013]:
  - Low-dimensionality/Sparseness
  - Distributed representations/hierarchy
    - Depth and abstraction
  - Shared factors across tasks
- Examples: auto-encoders, manifolds, deep neural nets ...
- How to induce these properties in your classifiers:
  - Include as regularization term in training classifier criterion
  - Include properties directly in classifier design
  - Go deep and pray (dirty neural net tricks)

My Vision

- Cognitively-motivated semantic and behavioral models
  - Emphasis on induction not classification
  - Associations not probabilities/distance
  - Hierarchical manifold models not metric spaces
  - Multimodal not unimodal
  - Mappings between modalities/layers (hub architecture)
  - Other cognitive considerations, e.g., parallelism ...

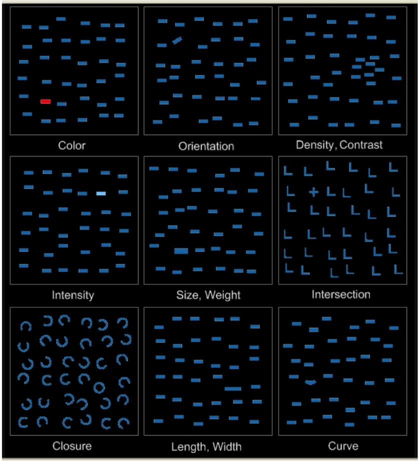
Maxims of Interaction



Cognition and Attention

- What grabs our attention?
  - Salient events
- Attention and Perception:
  - A simple perceptual algorithm
  - Quickly identify relevant (to survival) information
  - Bottom-up selectional attention: features extracted via low level signal processing
  - Fusion of top-down and bottom-up attention
- The attention/saliency relationship is used in multimedia production

Low-level visual features (from feng-gui.com)

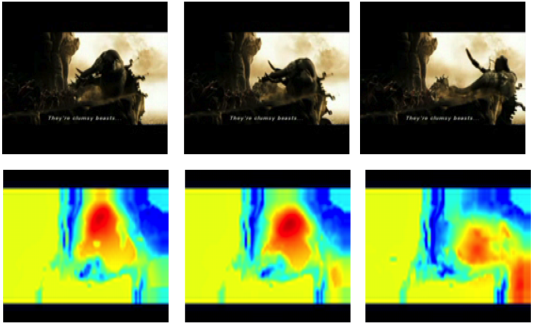


Bottom-up saliency estimation

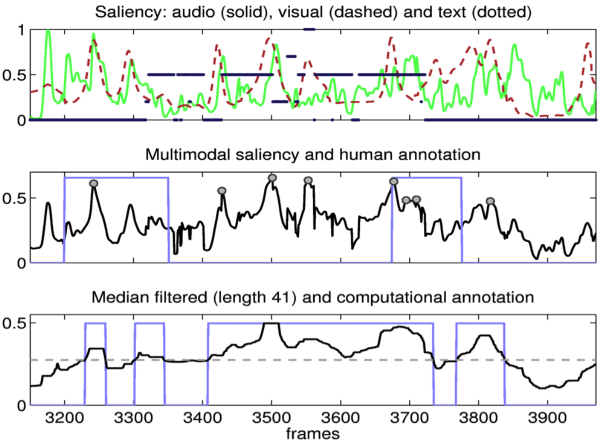
- Audio: rhythm, energy, change of frequency content, ...
- Images: color, orientation, density, intensity, size, weight ...
- Video: motion (direction, velocity), flicker
- Such low level features capture about 60-80% of “events” in each modality
- How do we capture the rest?
  - Multimodality (up to 90%)
  - Semantics (top-down selectional attention)
- High-performing computational algorithms for saliency estimation

Video summarization using audio-visual-text saliency

from [G. Evangelopoulos et al. 2013]



Video summarization using audio-visual-text saliency



Challenges

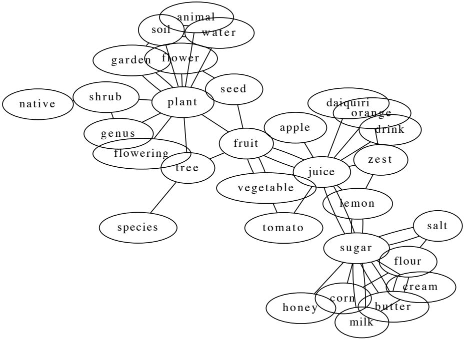
1. Extracting mid- and high-level features including incorporating semantics (scenes, objects, actions)
2. Fusion of features over time and over modalities
3. Computational models for the fusion of the bottom-up (gestalt-based) and top-down (semantic-based) attentional mechanisms
4. Applying these multimodal salient models to realistic human-human (especially) and human-computer interaction scenarios
5. Identifying the dynamics of attention and constructing joint (interactional) attention models

Constructing concept representations

- Word are associated with feature vectors
  - crisp, parsimonious representation of semantics
- Distributional semantic models (DSMs)
  - Semantic information extracted from word frequencies
  - Estimate co-occurence counts of word pairs or triplets
  - Estimate statistics of word context vectors
- Semantic networks
  - discovery of new relations via systematic co-variation
  - robust estimates – smoothing corpus statistics over network
  - rapid language acquisition

Example of Lexical Semantic Network

- Linked nodes: lexicalized senses and attributes
  - Informative for semantic similarity computation
- Computation of structural properties, e.g., cliques



Cognitive Considerations

Table 3.1
Some major differences between brains and digital computers

Brains	Computers
100,000,000,000 processing units	1–100 processing units
1000 operations/second	1,000,000,000 operations/second
Embodied	Abstract, disembodied
Fault tolerant	Frequently crashes
Graded, probabilistic signals	Binary, deterministic signals
Evolves and is self-organizing	Is explicitly designed
Learns	Is programmed

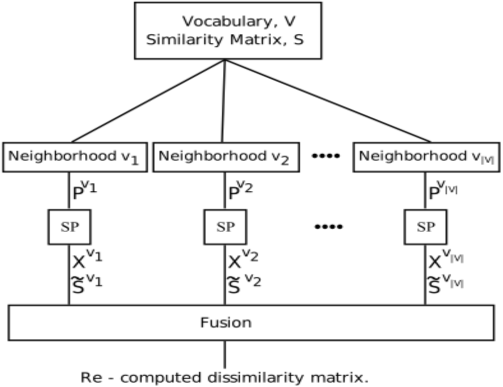
from [Feldman's book "From molecule to metaphor"]

Manifold DSMs

- Cognitive semantic space is fragmented in domains
- Sparse encoding of relations in each domain (manifold)
- Low-dimensional subspaces with good geometric properties
  - vs non-metric global semantic space
- Semantic similarity operation is performed locally in each subspace
- Decision fusion to reach semantic similarity score

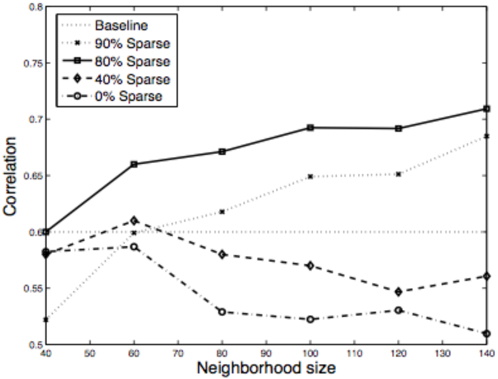
Manifold DSMs

from [Athanasopoulou and Potamianos, COLING 2014]



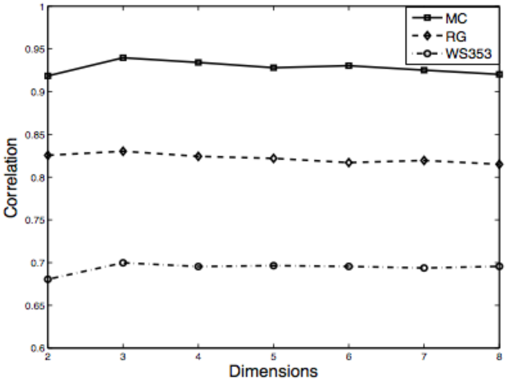
Sparse similarity matrices

- Correlation w. human ratings on the WS363 word semantic similarity task



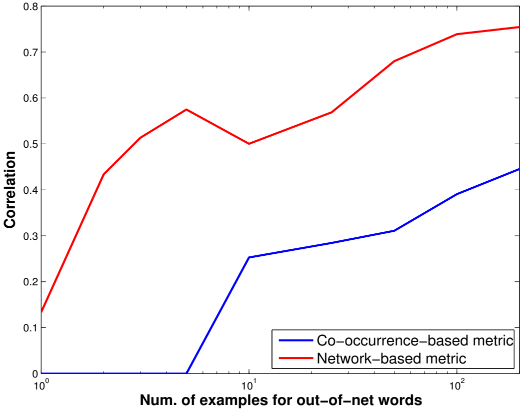
Effect of dimensionality

Very-low dimension in subspaces gives good or best performance!

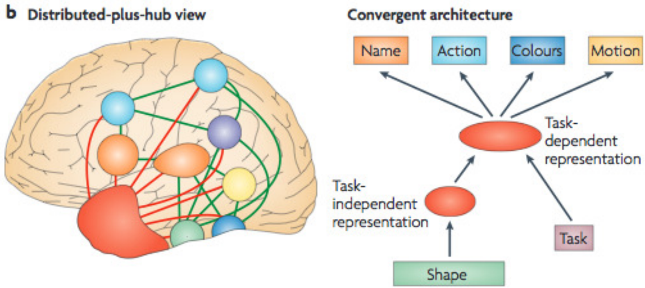


Lexical Acquisition using a semantic model

Learning the semantics of an unseen words from three web snippets!



Cogntive Maps



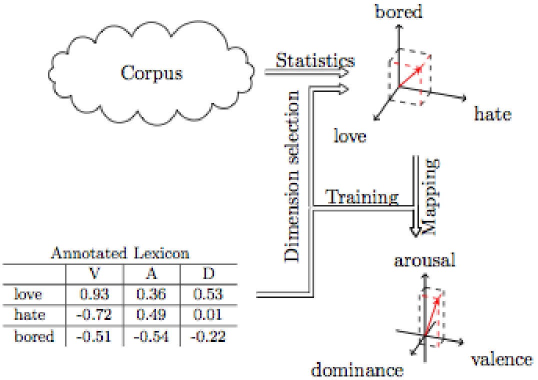
from [Patterson, Nestor and Rogers, 2007]

From Semantics to Behavior

Main idea: map from one representation space (semantics) to another, e.g., affect

- We present a method of expanding an affective lexicon, using web-based semantic similarity
- Assumption: semantic similarity implies affective similarity.
- Create a map from a semantic to an affective representation

Semantic-Affective Mapping



	V	A	D
love	0.93	0.36	0.53
hate	-0.72	0.49	0.01
bored	-0.51	-0.54	-0.22

Semantic-Affective Models

from [Malandakis et al 2013], extension of [Turney and Littman, 2002]

Assumption: the valence of a word can be expressed as a linear combination of the valence ratings of seed words weighted by semantic similarity and trainable weights a_i:

$$\hat{v}(t) = a_0 + \sum_{i=1}^{N} a_i \; v(w_i) \; d(w_i, t), \tag{1}$$

- t : a word or n-gram (token) not in the affective lexicon
- $w_1 ... w_N$: seed words
- $v(.)$: valence rating of a word or n-gram
- a_i : weight assigned to seed w_i
- $d(w_i, t)$: semantic similarity between word w_i and token t

Given

- an initial lexicon of K words
- a set of $N < K$ seed words

we can use (1) to create a system of K linear equations with $N + 1$ unknown variables:

$$\begin{bmatrix} 1 & d(w_1, w_1)v(w_1) & \cdots & d(w_1, w_N)v(w_N) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & d(w_K, w_1)v(w_1) & \cdots & d(w_K, w_N)v(w_N) \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} 1 \\ v(w_1) \\ \vdots \\ v(w_K) \end{bmatrix} \quad (2)$$

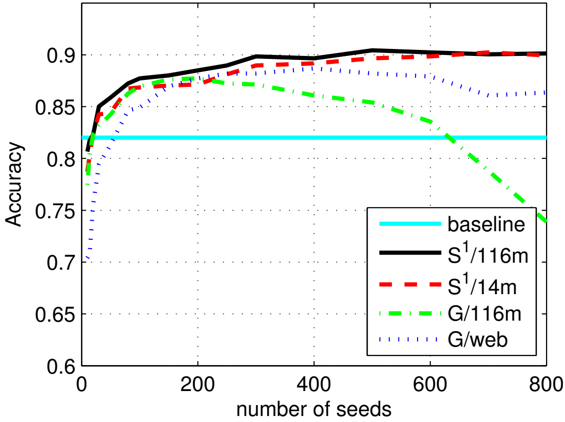
Solving with Least Mean Squares estimation provides the weights a_i.

Example, $N = 10$ seeds

Order	w_i	$v(w_i)$	a_i	$v(w_i) \times a_i$
1	mutilate	-0.8	0.75	-0.60
2	intimate	0.65	3.74	2.43
3	poison	-0.76	5.15	-3.91
4	bankrupt	-0.75	5.94	-4.46
5	passion	0.76	4.77	3.63
6	misery	-0.77	8.05	-6.20
7	joyful	0.81	6.4	5.18
8	optimism	0.49	7.14	3.50
9	loneliness	-0.85	3.08	-2.62
10	orgasm	0.83	2.16	1.79
-	w_0 *(offset)*	1	0.28	0.28

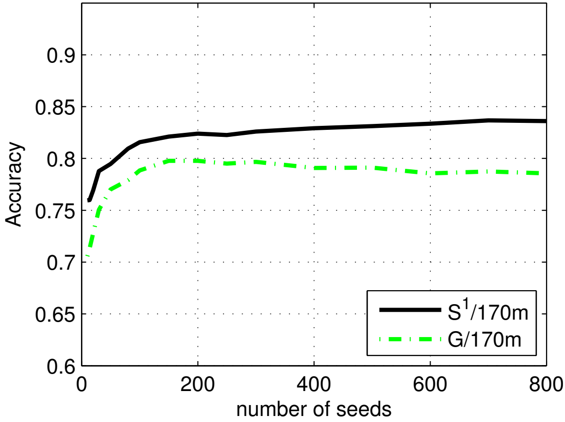
Word Polarity Detection (ANEW)

2-class word classification accuracy (positive vs negative)

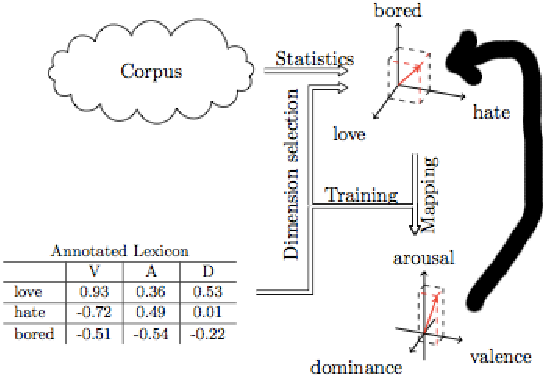


Word Polarity Detection (BAWLR)

2-class word classification accuracy (positive vs negative)



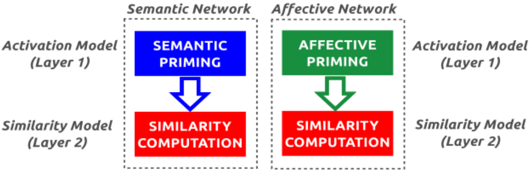
Cognitive Auto-Encoders



	V	A	D
love	0.93	0.36	0.53
hate	-0.72	0.49	0.01
bored	-0.51	-0.54	-0.22

The Semantics of Emotion (1)

- Semantic vs Affective Priming [Iosif and Potamianos, 2015]
- From Semantics to Affective Spaces and back



The Semantics of Emotion (2)

- Task: synonymy and antonymy pair detection
- Can semantic-affective auto encoders improve our semantic reprensetations?

Semantic relation	Baseline (random)	Feature types	
		Lexical (Lex1,Lex2,Lex3)	Affective (Aff1,Aff2,Aff3)
Synonymy	50%	61%	62%
Antonymy	50%	61%	**82%**

Classification accuracy for synonymy and antonymy: lexical vs. affective feature sets

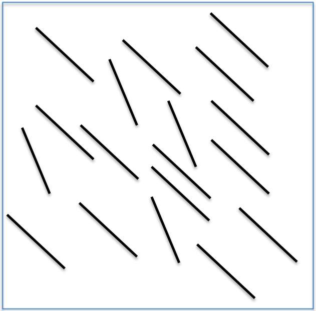
- Emotion carries important semantic information!
- Cognitive autoenconders show great potential in unlocking this information

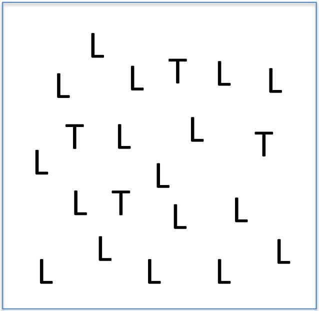
Dual-System Processing: System 1 vs System 2

- Using Kahneman's (and others) formalism:
  - System 1 (intuition): generates
    - impressions, feelings, and inclinations
  - System 2 (reason): turns System 1 input into
    - beliefs, attitudes, and intentions
- Associative relations reside in System 1
- But where do semantic relations reside?

Example

- Example from vision: system 1 vs system 2

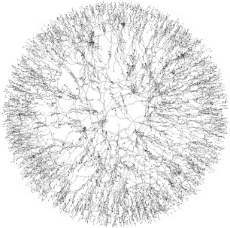




Proposed semantic similarity two-tier system

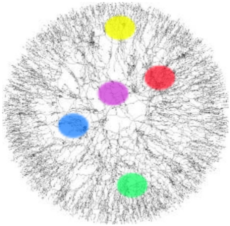
- Unifies the three approaches
- Fuzzy vs explicit semantic relations
- Word senses vs words vs concepts
- A two tier system
  - An associative network backbone
  - Semantic relations defined as operations on network neighborhoods (cliques)
- Consistent with system 1 vs system 2 view
- Furthermore we believe that the
  - underlying network consists of word senses, and
  - is a low dimensional semi-metric space

Lexical Network - Semantic Neighborhoods



Lexical Network

- Undirected graph $G = (N, E)$
  - Vertices N: words in lexicon L
  - Edges E: word similarities



Semantic Neighborhoods

- For word i create subgraph G_i
- Select neighbors of i
  - Compute $S(i, j), \forall j \in L, i \neq j$
  - Sort j according to $S(i, j)$
  - Select $|N_i|$ top-ranked j

Semantic Neighborhoods: Examples

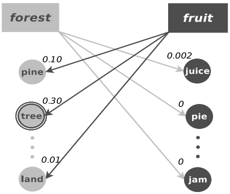
Word	Neighbors
automobile	**auto**, truck, **vehicle**, **car**, **engine**, bus, ...
car	truck, **vehicle**, travel, service, **price**, **industry**, ...
slave	slavery, beggar, **nationalism**, society, **democracy**, **aristocracy**, ...
journey	**trip**, holiday, **culture**, **travel**, **discovery**, **quest**, ...

- Synonymy
- Taxonomic: IsA, Meronymy
- Associative
- Broader semantics/pragmatics
- ...

Neighborhood-based Similarity Metrics: M_n

[from E. Iosif and A. Potamianos, 2013]

M_n metric: maximum similarity of neighborhoods



- Motivated by maximum sense similarity assumption
  - Neighbors are semantic features denoting senses
  - Similarity of two closest senses
- Select max. similarity: $M_n(\text{"forest"}, \text{"fruit"}) = 0.30$

Performance of web-based similarity metrics

- Task: similarity judgment (Miller-Charles dataset)
- Evaluation metric: correlation wrt to human ratings

Feature	Description	Correlation
context	AND queries	0.88
context	IND queries	0.55
context	IND queries: network	0.90

- Comparable to structured DSMs, WordNet-based approaches

Cognitive Fusion

- Types of fusion:
  1. Multimodal fusion, i.e., fusion between modality-specific processing outputs and multimodal outputs
  2. Fusion over time, i.e., how stimuli are integrated both within and across modalities
  3. Fusion of top-down (data-driven) and bottom-up (semantic) processing, or in general fusion between different layers of cognitive and computational processing

Challenge: go beyond simple algorithms that employ (weighted) averages of outputs (across time, modalities and processes) and design algorithms that make often highly non-linear fusion decisions depending on our cognitive state, behaviors and intentions

Grand Challenges

1. Annotation of the mid- and high-level behaviors associated with human-human and human-machine interaction
2. Attention and saliency modeling using mid- and high-level features (including semantics), as well as fusion model of top-down and bottom-up attentional mechanisms
3. From signal to semantics: use “big data” to construct distributed, low-dimensional semantic cognitive representations
4. From semantics to SSP/BSP labels: estimate mapping between semantics and other cognitive representation layers
5. Design models that are stateful and are able to predict cognitive biases, nonlinear logic, abrupt state transitions and surprise
6. Design multi-modal fusion algorithms that exhibit nonlinear behavior and depend on cognitive states, behaviors etc

Thank you