Lexical Semantic Spaces: a Review

Alexandros Potamianos

Dept. of ECE, Technical Univ. of Crete, Chania, Greece

Problem Definition

- How are words/concepts represented and modeled
  - in the human brain
  - by machines
- Motivation
  - Accuracy in classification/recognition tasks (performance)
  - Efficiency: low energy, low storage, high speed
  - Generalization power for rapid learning and inference
- Disciplines involved:
  - Philosophy
  - Psychology
  - Linguistics
  - Computer Science: AI, NLP, Machine Learning
  - Cognitive Science 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

Bibliography

1. H. Plotkin, *Darwin Machines and the Nature of Knowledge*, Harvard University Press, 1997.
2. G. Murphy, *The Big Book of Concepts*, MIT Press, 2002.
3. P. Gardenfors, *Conceptual Spaces: the Geometry of Thought*, MIT Press, 2000.
4. D. Widdows, *Geometry and Meaning*, CSLI Publications, 2004.
5. T. T. Rogers and J. L. McClellan, *Semantic Cognition: A Parallel Distributed Processing Approach*, Bradford Books, 2006.
6. J. C. L. Ingram, *Neurolinguistics: An Introduction to Spoken Language Processing and its Disorders*, Cambridge U. Press, 2007.
7. Textbooks on NLP, AI, Machine Learning, Cognition, NeuroCognition ...

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!

Properties of the Three Approaches

Property	Symbolic	Conceptual	Connectionist
cognitive speed	very slow	slow	fast
machine speed	very fast	pretty fast	fast
cognitive accuracy	good	good	decent
machine accuracy	decent	good	good
dimensionality	high	low	high
representation	flat	hierarchical	distributed
interpretability	excellent	good	low
determinism	high	medium	low
reasoning (all data)	good	good	decent
compositionality	good	good	decent
induction/learning	poor	excellent	average

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

Some Terminology

- Words and words senses
- Concepts: nonlinguistic cognitive constructs
- Categories, basic categories, superordinates
  - Informativeness and distinctiveness (differentiation theory)
- Domains, e.g., color
- Functional role of parts-of-speech
  - Concepts typically represented by nouns
  - Properties typically represented by adjectives
  - Dynamic concepts represented by verbs
- Compositional semantics
  - Noun compounds (NN) and extensional analysis
  - Adjective-noun compounds (AN) and modifiers

Experimental Data I

- Priming, associations and similarity
- Evidence for cognitive mappings to low dim spaces
- Typicality effects
- Basic categories, hierarchical cognitive organization
- Categorical neurons (neurocognition)
- Similarity (Tversky et al)
  - Asymmetry in lexical similarity, e.g., Athens, NYC
  - Triangular inequality violation for similarity, esp. between classes and words
- Concept Learning
  - Category learning is 1-D
  - Prior knowledge does not impede statistical learning
  - Easier to learn conjunctive than disjunctive categories

Experimental Data II

- Category Induction (Ripp 1973, Osherson et al 1990)
  - $\{A \text{ has } p\} \Rightarrow \{B \text{ has } p\}$ conditioned on category C
  - depends on $sim(A, B)$
  - depends on typicality of A in C and diversity of A in C
  - second property develops between ages 5 and 8
  - adding more evidence $\{D \text{ has } p\}$ helps induction
  - but use $span(A, D)$ rather than $A \bigcup D$
  - inclusion fallacy (Tversky and Kahneman)
- Only a single word sense (category/domain) used in inductions
- Role of syntax/morphology in lexical/concept acquisition

Experimental Data III

- Children
  - can learn word meaning (approx.) from 2-3 examples!
  - posses similar cognitive mechanism for category learning and usage as adults
  - conceptual spaces exist but less developed
  - basic categories exist but might by superordinates to begin with
- Taxonomic bias
  - When name is given tendency to use categorical criteria
- Compositionality
  - People often overextend meaning to find appropriate meaning of noun compounds
- ....

Conceptual Spaces: Models from Psychology

- Classical approach
  - Concept are defined by dictionary entries
  - Represented by features
  - Logic (boolean conditions) define combination of features
- Prototype theory
  - Extension of classical approach
  - Motivated by typicality of some class members
  - Prototype defined as a weighted combination of features
- Exemplar theory
  - Motivated by priming
  - Class is defined by a list of examples
- Theory-theory
- Knowledge-based approach

Conceptual Spaces to Metric Spaces

- Intro to lexical similarity features and metrics
- Method 1 (Low dim. metric spaces):
  1. start from similarities or co-occurrence counts
  2. perform multidimensional scaling
- Method 2 (Vector Space Models, LSA):
  1. define vector space model of features or co-occurrences
  2. perform PCA to go down to 200-300 dimensions
- Some examples:
  - Osgood 1957: from similarity to affective spaces
  - Method 1 makes sense for small homogeneous domains, e.g., mammals
  - Method 2 provides good results for synonymy and other tasks

The Math behind the Models

- Logic
  - Boolean logic
  - (First-order) predicate calculus and $\lambda-$calculus
  - Quantum logic
  - Non-monotonic logic
  - Statistical logic
- Models
  - Set theory
  - Metric spaces
    - Vector space models
  - Lattices and formal concept analysis
  - Graphical models
  - Neural networks
    - Topology constrained - Kohonen's maps
    - Harmonicity functions and attractors

Sets and Boolean Logic

Set Theory		Boolean Algebra	
Membership	$x \in A$	Predicate	$p(x \in A) = 1$
Complement	$\bar{A}$	Negation	NOT A
$\bar{A} = \{x : x \notin A\}$		$1 - p(x \in A) = p(x \notin A)$	
Intersection	$A \cap B$	Conjunction	A AND B
$A \cap B = \{x : x \in A, x \in B\}$		$p(x \in A).p(x \in B)$	
Union	$A \cup B$	Disjunction	A OR B

- e.g., $x \cap x = x$ is equivalent to $x^2 = x$ (idempotency)
- probabilistic generalization of boolean algebra, e.g.,
    - max-plus algebra, tropical semi-rings
- implement using finite-state machines

First-Order Predicate Calculus

- Augment predicate logic with quantification, i.e., $\forall$, $\exists$
- Reasoning about properties shared by many objects
  - use variables a, e.g., $\forall a(\text{Horse}(a) \rightarrow \text{Mammal}(a))$
- (Most) sentences can be interpreted using FOPC
  - e.g., $\exists x(\text{Person}(x) \wedge \forall y(\text{Time}(y) \rightarrow \text{Canfool}(x, y)))$
  - cannot account for beliefs or opinions (higher-order logic)
- $\lambda-$ calculus [Jurafsky and Martin, NLP]
  - express computation via variable binding and substitution
  - lexical rules augmented with semantics
  - verb rules acting on noun phrases ($\lambda-$ variables)
- Inference
  - excellent for deduction
  - poor at induction

Alternative Logic Systems I

- Fuzzy logic, many valued logic
- Quantum logic
  - failure of the distributive law $p \wedge (q \vee r) = (p \wedge q) \vee (p \wedge r)$
  - projections on a Hilbert space propositions about physical observables [John von Neumann]
  - vector space representation: if $p(x \in A)$, $q(x \in B)$, then $p \vee q = \text{prop}(x \in span\{base\{A\}, base\{B\}\})$
- Non-monotonic logic, e.g., inheritance nets [Touretsky 1986]
  - Monotonicity: if p can be inferred on S it can also be inferred on a superset of S
  - Cognitive logic non-monotonic, e.g., property generalization from a category to an atypical example (bird to chicken)

Alternative Logic Systems II

- Statistical logic
  - define harmonicity function on neural nets
  - energy minimization provides attractors
  - achieve inference by using attractor space (low dim.)
  - can emulate boolean logic, FOPCs
- Structured neural nets
  - topological constraints, e.g., Kohonen maps
  - deep neural nets, e.g., Google brain
    - trains layers sequentially (evidence from neurocognition)
    - alternating layers of correlational and max pooling
    - see `http://deeplearning.net/reading-list/`

Metric Spaces I

- Geometric properties of equidistance and betweenness can be extended to form metric spaces
- Metric spaces are sets with a distance function $d(.,.)$ that is real-valued, nonnegative and
  1. $d(x,y) = 0 \Leftrightarrow x = y$
  2. $d(x,y) = d(y,x)$
  3. $d(x,y) \leq d(x,z) + d(z,y)$
- Triangle inequality implies convergence, e.g.,
  - any convergent sequence in a metric space is a Cauchy sequence
- Relax conditions 1-3:
  1. pseudo-metrics, e.g., 2-D Euclidean distance in 3-D space
  2. quasi-metrics, e.g., time-travel on map
  3. semi-metrics, e.g., ultra-metrics (max), $\rho-$relaxed triangle inequality

Metric Spaces II

- Extension to normed spaces $\|.\|$
  - Vector spaces
  - $d(ax+b, ay+b) = ad(x,y)$, $a, b \in \mathcal{R}$, then:
  - $d(x,y) = \|x-y\|$
  - Minkowski distances, i.e., $d(x,y) = (\sum_i |\xi_i - \eta_i|^p)^{1/p}$
  - All norms give equivalent topologies
  - Bounded linear operators are matrices in finite dimensions
  - Balls are convex
  - Projections of points onto subspaces are convex regions

Metric Spaces III

- Extension to inner product spaces $< .,. >$
  - $< .,. >$ linear in 1st argument and conjugate linear in 2nd
  - $\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$, then:
  - $d(0,x) = \|x\| =< x, x >^{1/2}$
  - Euclidean distance
  - Orthogonality, orthogonal complement, orthonormal sets
  - Riesz representation of linear bounded functionals
  - Projections of points onto subspaces are unique (points)
- Critique of vector and inner product spaces
  - Orthogonality is defined as $< x, y >= 0$
  - Also $d(x,y) =< x-y, x-y >^{1/2}= (\|x\|^2 + \|y\|^2)^{1/2}$
  - Most words/concepts are far away, i.e., orthogonal
  - Thus these spaces inherently high-dimensional!!!

Vector Spaces and Boolean Logic

- Extended Boolean logic in information retrieval
- Negation
  - Belongs in $span\{\vec{a}, \vec{b}\}$, i.e., $\vec{a}$ NOT $\vec{b} = \vec{a} - \lambda\vec{b}$
  - Orthogonal to $\vec{b}$, i.e., $< \vec{a}$ NOT $\vec{b}, \vec{b} >= 0$
  - thus $\lambda = \frac{<\vec{a},\vec{b}>}{||b||^2}$
- Conjunction (AND) of subspaces
  - Intersection of subspaces (still a subspace)
- Disjunction (OR) of subspaces
  - Union of subspaces (not a subspace!)
  - Span of their union which is a subspace (a better option)!

Distributed Semantic Models (DSMs)

- Vector space models of contextual lexical similarity vectors
- Similarity metric is cosine similarity (norm. inner product)
  - Relation to Euclidean distance: $d_E^2(a, b) = 2 - 2 < a, b >$ (for normalized vectors)
- Inherently thousands of dimensions: 3K-100K
- Use various techniques: PCA, non-negative matrix factorization to reduce to 30-300 dimensions
- Compositional semantics, e.g., adj-noun pairs
  - linear operators (matrices) of adj. operating on noun
- Computationally useful, cognitively unnatural (induction?)
- Our proposal: Neighborhood-based DSMs
  - Hierarchical organization of low-dim spaces

Lattices and Formal Concept Analysis

- Ordered sets:
  - reflexivity: $x \preceq x$
  - antisymmetry: $x \preceq y$ and $y \preceq x$ implies $x = y$
  - transitivity: i$x \preceq y$ and $y \preceq x$ implies $x \preceq z$
- Define operators
  - Join: least upper bound
  - Meet: greatest lower bound
- Lattice: ordered set with a unique join and meet
- Containment relation $\subseteq$ forms a lattice for (some) sets with
  - Join operator being the set union $\cup$
  - Meet operator being the set intersection $\cap$
- Add negation
- Quantum logic as lattices

Conceptual Spaces

[Gardenfors 2000]

1. Meaning is a conceptual structure in cognitive systems
2. Conceptual structures are embodied (perceptual)
3. Semantic elements generated from geometric/topological structures
4. Cognitive models are image-schematic (not propositional), transferred via metaphoric/methonymic operations
5. Semantics is primary to syntax (not the other way around)
6. Concepts show prototype effects

Our Work [Iosif, Athanasopoulou, Potamianos]

- Neighborhood-based VSMs are consistent with
  - Two-tier architecture (connectionist and conceptual)
  - Account for asymmetry and triangle inequality violations
  - Account for category typically and priming effects
  - Perform very-well for semantic similarity tasks (both at the word-level and for noun-noun combinations)
- Hidden set multi-dimensional scaling
  - Assume that words are union of word senses (concepts)
  - Assume common sense set distance
  - Show that semantic spaces are only locally metric (due to word senses)
  - Split word into senses automatically to maximize metricity
  - Operate on neighborhoods rather than globally
  - Missing glue: how to combine conceptual subspaces

Conclusions

- Missing link between propositional and connectivist approach
- Structure and good geometrical properties needed
- Structure could be learned automatically
  - Attractors in neural net
  - Topologically constrained NN
  - Deep learning in NN
- Language can be codified as sub-spaces, operators and transformation in conceptual spaces
- Cognitive linguistics, cognitive logic, pre-metric spaces ...

Thank you