USC Institute for Creative Technologies

University of Southern California

Distributed Vector Representations of Words in Sigma

Volkan Ustun, Paul S. Rosenbloom, Kenji Sagae, and Abram Demski

8.4.2014

The work depicted here was sponsored by the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

Distributed Vector Representation or Word Embedding

- Simple yet general approach to integrating large amounts of diverse knowledge while yielding natural measures of similarity
- Assign long (e.g., 1000) random vectors to words & concepts

0.60665036	- 0.5666231	0.41830373	- 0.5400135	0.61649907	0.02903163	0.16481042

. . .

- Evolve "better" vectors from experience with usage
  - Co-occurring words, n-grams, phonetic structure, visual features, …
- Degree of similarity is a function of distance in vector space
  - For richer language models, simple forms of analogy, …

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

- Long history in cognitive science (particularly neural networks)
  - More recently an important thread in machine learning
  - Started to appear in a few cognitive architectures

Our Hypothesis

Sigma can efficiently and effectively support a *distributed vector representation* that enables implicit learning of the meanings of words and concepts from large but shallow information resources

Distributed Vector Representations in Sigma (DVRS)

Ordering Context

The AGI conferences encourage interdisciplinary research based on different understandings of intelligence, and exploring different approaches.

Ordering Context Vector

Lexical Vector

$$o(k) = \sum_{i=1}^{4} s(i) * e(k+i)$$

$$l(k) = l(k) + \widehat{c(k)} + \widehat{o(k)}$$

where $j \neq 0$ and $0 < (k+j) \leq n$

DVRS and BEAGLE

- DVRS is inspired by BEAGLE*
  - Both utilize environmental and lexical vectors
  - Both capture context and ordering information
  - Skip-grams rather than n-grams for ordering information
    - Fixed random sequence vectors
    - Point-wise multiplication as the binding operation rather than circular convolution

*Bound Encoding of the Aggregate Language Environment (BEAGLE)
Jones and Mewhort (2007). "Representing word meaning and order information in a composite holographic lexicon". *Psychological Review*. 114(1). 1-37

Sample Results from an External Simulator

Training data is enwik8 -> First 10^8 bytes of the English Wikipedia dump from 2006.
~12.6M words

language

Context	Ordering	Composite
spoken	cycle	languages
languages	society	vocabulary
speakers	islands	dialect
linguistic	industry	dialects
speak	era	syntax

film

Context	Ordering	Composite
director	movie	movie
directed	german	documentary
starring	standard	studio
films	game	films
movie	french	movies

Assessment of DVRS

- Word2Vec's Semantic-Syntactic Word Relationship Test Set*
  - "What is the word that is similar to *small* in the same sense as *biggest* is similar to *big*?"
    - $V = (I_{biggest} - I_{big}) + I_{small}$
  - or "Which word is the most similar to *Paris* in the way *Germany* is similar to *Berlin*?"
    - $V = (I_{germany} - I_{berlin}) + I_{paris}$

* https://code.google.com/p/word2vec/

Accuracy on Semantic-Syntactic Word Relationship Test Set

	Vector Size	Semantic	Syntactic	Overall
Co-occurrence only	1024	33.7 (31.1)	18.8 (18.6)	25.3 (24.3)
3-Skip-Bigram only	1024	2.7 (2.5)	5.0 (4.9)	4.0 (3.8)
3-Skip-bigram composite	512	29.8 (27.5)	18.5 (18.3)	23.4 (22.4)
3-Skip-bigram composite	1024	32.7 (30.2)	19.2 (18.9)	25.1 (24.0)
3-Skip-bigram composite	1536	34.6 (31.9)	20.1 (19.9)	26.4 (25.3)
3-Skip-bigram composite	2048	34.3 (31.7)	20.1 (19.9)	26.3 (25.2)

Word2Vec
19.3%

Sigma's Goals and DVRS

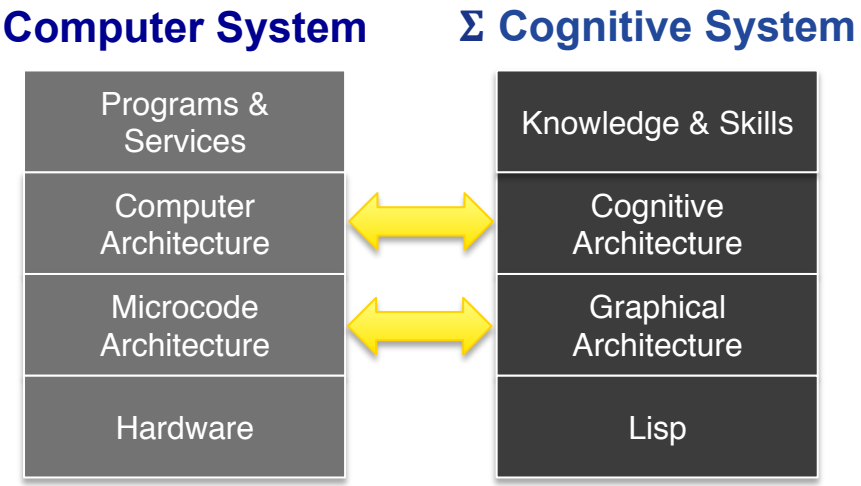
- A new breed of cognitive architecture that is
  - *Grand unified*
    - Expanding to distributed representations
  - *Functionally elegant*
    - Distributed representations and reasoning based on current Sigma
  - *Sufficiently efficient*
    - Fast enough for anticipated applications *
- For virtual humans, AGIs and intelligent robots
  - Bridging between speech and language and cognition

Overall Progress on Sigma

- Memory [ICCM 10]
  - Procedural (rule)
  - Declarative (semantic/episodic) [CogSci 14]
  - Constraint
  - Distributed vectors [AGI 14a]
- Problem solving
  - Preference based decisions [AGI 11]
  - Impasse-driven reflection [AGI 13]
  - Decision-theoretic (POMDP) [BICA 11b]
  - Theory of Mind [AGI 13, AGI 14b]
- Learning [ICCM 13]
  - Concept (supervised/unsupervised)
  - Episodic [CogSci 14]
  - Reinforcement [AGI 12a, AGI 14b]
  - Action/transition models [AGI 12a]
  - Models of other agents [AGI 14b]
  - Perceptual (including maps in SLAM)
- Mental imagery [BICA 11a; AGI 12b]
  - 1-3D continuous imagery buffer
  - Object transformation
  - Feature & relationship detection
- Perception
  - Object recognition (CRFs) [BICA 11b]
  - Isolated word recognition (HMMs)
  - Localization [BICA 11b]
- Natural language
  - Question answering (selection)
  - Word sense disambiguation [ICCM 13]
  - Part of speech tagging [ICCM 13]
- Graph integration [BICA 11b]
  - CRF + Localization + POMDP
- Optimization [ICCM 12]

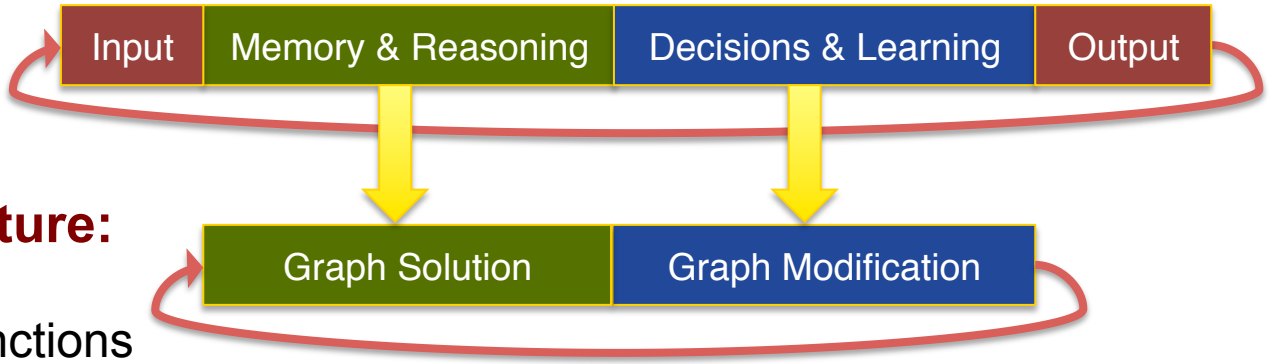
The Structure of Sigma

- Constructed in layers
  - In analogy to computer systems



Cognitive Architecture:

Predicates
Conditionals
Nested control structure



Graphical Architecture:

Graphical models
Piecewise-linear functions
Gradient-descent learning

```
CONDITIONAL Concept-Prior
   Conditions: Object(s,O1)
   Condacts: Concept(O1,c)
```

Walker	Table	Dog	Human
.1	.3	.5	.1

Predicates & Conditionals

- *Predicates* specify relations among typed arguments
  - `(predicate 'concept :arguments '((id id) (value type %)))`
  - Types may be *symbolic* or *numeric* (*discrete* or *continuous*)
- Each induces a segment of *working memory* (WM)
- Perception predicates also induce a segment of *perceptual buffer*
- *Conditionals* define *long-term memory* (LTM) and *basic reasoning*
  - Deep blending of traditional rules and probabilistic networks
- Comprise a *name* plus *predicate patterns* and an optional *function*
  - Patterns may include *constant tests* and *variables*
  - Patterns may be *conditions*, *actions* or *condacts*
  - Functions are *n*D piecewise continuous (linear) functions

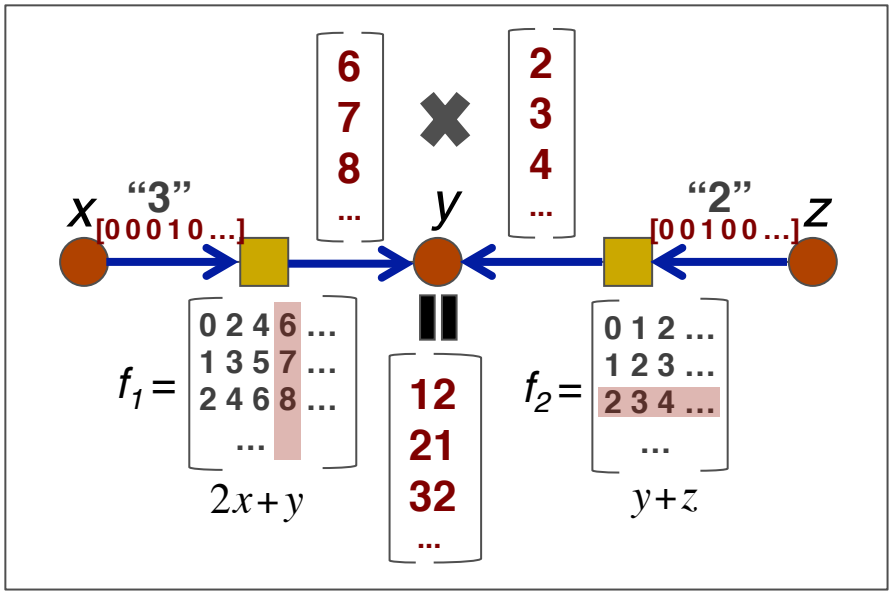
*y**x*	[0,10>	[10,25>	[25,50>
[0,5>	0	.2*y*	0
[5,15>	.5*x*	1	.1+.2*x*+.4*y*

Summary Product Algorithm

- Compute variable marginals (or mode of entire graph)
- Pass messages on links and process at nodes
  - Messages are distributions over link variables (starting w/ *evidence*)
  - At variable nodes messages are combined via *pointwise product*
  - At factor nodes do products, and summarize out unneeded variables:

$$m(y) = \int_x m(x) \times f_1(x,y)$$

$$f(x,y,z) = y^2 + yz + 2yx + 2xz$$
$$= (2x+y)(y+z) = f_1(x,y) f_2(y,z)$$



DVR in Sigma

- Vectors are discrete piecewise-constant functions

0.60665036	-0.5666231	-0.4183037	0.54001356	-0.6164990	0.02903163	0.16481042

- Sum-product algorithm manipulates (× & +) vectors
- Gradient-descent evolves lexical representations

Conditional for Context

w
1
0
1

✖

w \ d				
0.66	0.14	0.92	0.17	0.14
0.43	0.1	0.17	0.53	0.53
0.01	0.71	0.77	0.08	0.53
0.51	0.54	0.70	0.81	0.94

=

w \ d				
0.66	0.14	0.92	0.17	0.14
		0		
0.51	0.54	0.70	0.81	0.94

Summarization

d				
1.17	0.68	1.62	0.98	1.08

L2 Normalization

d				
0.46	0.27	0.63	0.38	0.42

CONDITIONAL *Co-occurence*

Conditions: Co-occuring-Words(word:*w*)

Actions: Context-Vector(distributed:*d*)

Function(*w*,*d*): *environmental-vectors*

$$c(k) = \sum_{i=1}^{n} e(i), where\ i \neq k$$

Conditionals for Ordering Information

```
CONDITIONAL Skip-gram
   Conditions: Skip-Gram-Words(word:w position:p)
               Environmental-Vectors(word:w distributed:d)
   Actions: Skip-Gram-Matrix(distributed:d position:p)

CONDITIONAL Ordering
   Conditions: Skip-Gram-Matrix(distributed:d position:p)
   Actions: Ordering-Vector(distributed:d)
   Function(p,d): *sequence-vectors*
```

Ordering Vector

$$o(k) = \sum_{j=-4}^{4} s(j) .* e(k+j)$$

$$where\ j \neq 0\ and\ 0 < (k+j) \leq n$$

Conditionals for Meaning/Lexical Vector

```
CONDITIONAL Context
  Conditions: Context-Vector(distributed:d)
              Current(word:w)
  Actions: Meaning-Vector(word:w distributed:d)

CONDITIONAL Ordering
  Conditions: Ordering-Vector(distributed:d)
              Current(word:w)
  Actions: Meaning-Vector(word:w distributed:d)
```

Lexical Vector

$$l(k)_t = l(k)_{t-1} + \widehat{c(k)} + \widehat{o(k)}$$

Gradient Descent Gradient via Action Combination

Sigma Results

- Capital common countries subset of the Word2Vec test data.
  - Vector dimension is 100
  - 506 test instances – 46 distinct capitals and countries
    - enwik8 has 65086 distinct words (28532 entries) co-occurring with the common capitals and countries

- 35.2% in DVR in Sigma vs. [26.1%,43.1] DVR in External Simulator

Conclusions

- DVRS is fast in the external simulator
- Accuracy on Semantic-Syntactic Word Relationship Test Set is as good as Word2Vec when both trained on a comparable relatively small corpus
- It fits naturally into Sigma; however, more is necessary for requisite efficiency and effectiveness
  - Revise and/or augment Sigma's function representation for efficiency with large (non-sparse) discrete vectors
  - Enable negative values in summary product and gradient descent
    - To enable use of all quadrants of vector space

Current State & Future Work

- Further progress
  - DVR in Sigma
    - Attuned Sigma more to explicit vector predicates
  - DVR in External Simulator
    - Running larger data sets and more comprehensive comparisons
    - Applying DVRS to a sentence classification task

- Future work
  - Further optimizations
  - Bridging between speech and language and cognition
  - Pervasive use for analogy and semantic memory