Compositional Distributional Semantics

L98 Lecture 11

Guy Emerson

What I'll Cover...

- Recap: Distributional semantics

What I'll Cover…

- Recap: Distributional semantics
- Compositionality vs. context dependence

What I'll Cover...

- Recap: Distributional semantics
- Compositionality vs. context dependence
- Composition in vector space models

What I'll Cover...

- Recap: Distributional semantics
- Compositionality vs. context dependence
- Composition in vector space models
- Truth-conditional distributional semantics (my work)

Distributional Hypothesis/Hypotheses

- "Similar words appear in similar contexts"
  (For history and discussion, see: Sahlgren, 2008, "The distributional hypothesis")

Distributional Hypothesis/Hypotheses

- “Similar words appear in similar contexts”
  (For history and discussion, see: Sahlgren, 2008, “The distributional hypothesis”)

- “The contexts in which an expression appears give us information about its meaning”
  (Emerson, 2020, “What are the Goals of Distributional Semantics?”)

Distributional Semantics

… being hurt by another	horse	especially if some rider …
… beaten by a better	horse	at the distance on …
… from these studies that	horses	reared with other horses …
… horses reared with other	horses	in a free and …
… ‘Is that all your	horse	gets to eat?’ in …
… cache of cattle and	horse	bones, while from the …
… was a sterling good	horse,	especially at Ascot, but …
… way as a domestic	horse	that it is stabled …
… 1790 – that is, one	horse	or two cows for …
… as coarse as a	horse	’s tail straying from …

Distributional Semantics

… being hurt by another	horse	especially if some rider …
… beaten by a better	horse	at the distance on …
… from these studies that	horses	reared with other horses …
… horses reared with other	horses	in a free and …
… ‘Is that all your	horse	gets to eat?’ in …
… cache of cattle and	horse	bones, while from the …
… was a sterling good	horse,	especially at Ascot, but …
… way as a domestic	horse	that it is stabled …
… 1790 – that is, one	horse	or two cows for …
… as coarse as a	horse	’s tail straying from …

Distributional Semantics

… being hurt by another	horse	especially if some rider …
… beaten by a better	horse	at the distance on …
… from these studies that	horses	reared with other horses …
… horses reared with other	horses	in a free and …
… ‘Is that all your	horse	gets to eat?’ in …
… cache of cattle and	horse	bones, while from the …
… was a sterling good	horse,	especially at Ascot, but …
… way as a domestic	horse	that it is stabled …
… 1790 – that is, one	horse	or two cows for …
… as coarse as a	horse	’s tail straying from …

Distributional Semantics

… being hurt by another	horse	especially if some rider …
… beaten by a better	horse	at the distance on …
… from these studies that	horses	reared with other horses …
… horses reared with other	horses	in a free and …
… ‘Is that all your	horse	gets to eat?’ in …
… cache of cattle and	horse	bones, while from the …
… was a sterling good	horse,	especially at Ascot, but …
… way as a domestic	horse	that it is stabled …
… 1790 – that is, one	horse	or two cows for …
… as coarse as a	horse	’s tail straying from …

Distributional Semantics

- The context of a word gives us information about its meaning

Distributional Semantics

- The context of a word gives us information about its meaning
- What should the model learn?
- How can the model learn it?

Distributional Semantics

- The context of a word gives us information about its meaning
- What should the model learn?
  - Mainstream NLP: vectors
- How can the model learn it?

Distributional Semantics

- The context of a word gives us information about its meaning
- What should the model learn?
  - Mainstream NLP: vectors
- How can the model learn it?
  - Skip-gram, Transformers, …

Compositionality

- Principle of Compositionality: “The meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined” (Partee, 1984)

Compositionality

- Principle of Compositionality: “The meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined” (Partee, 1984)

- (What is allowed in a meaning representation?)

- (What is allowed in a composition function?)

Compositionality

- Productivity: new combinations are immediately understandable

Compositionality

- Productivity: new combinations are immediately understandable
- Productivity requires compositionality

Context Dependence

- The meaning of one part depends on another?
  - {big, small} {elephant, dog, ant, ...}

Context Dependence

- The meaning of one part depends on another?
  - {big, small} {elephant, dog, ant, ...}
  - Productive! Compositional!

Context Dependence

- The meaning of one part depends on another?
  - {big, small} {elephant, dog, ant, ...}
  - Productive! Compositional!
  - Lexical semantics harder than it might seem...

Context Dependence

- The meaning of one part depends on another?
  - {big, small} {elephant, dog, ant, ...}
  - Productive! Compositional!
  - Lexical semantics harder than it might seem...
- Idioms not compositional (or only semi-compositional):
  - Big Bang (early universe)
  - Big Apple (New York)

Contextualisation vs. Composition

- Contextualised representations
  - One per token (e.g. BERT vectors)

Contextualisation vs. Composition

- Contextualised representations
  - One per token (e.g. BERT vectors)
- Compositional representations
  - One for whole expression (e.g. semantic graph)

Vector Composition

- Goal: derive one vector for an expression

Vector Composition

- Goal: derive one vector for an expression
- Option 1: explicit operation
- Option 2: neural network

Vector Operations

- Vector addition surprisingly effective

Vector Operations

- Vector addition surprisingly effective
  - But insensitive to word order

Vector Operations

- Vector addition surprisingly effective
  - But insensitive to word order
- (Multiplication is addition of log vectors)

Vector Operations

- Vector addition surprisingly effective
  - But insensitive to word order
- (Multiplication is addition of log vectors)
- Tensors have compositional structure
  (Coecke et al., 2010; Baroni et al., 2013)
  - … but high-order tensors are hard to learn

Neural Composition

- Neural net:
  - Sequence of vectors as input
  - One vector as output

Neural Composition

- Neural net:
  - Sequence of vectors as input
  - One vector as output
- Many architectures...
  - LSTM
  - Tree-LSTM
  - DIORA (Deep Inside-Outside Recursive Autoencoder)
  - RNNG (Recursive Neural Network Grammar)
  - ...

Distinct Meanings

- Every fluffy dog barked
- Every dog that barked is fluffy

Distinct Meanings

- Every fluffy dog barked
- Every dog that barked is fluffy

- The fluffy dog barked
- The dog that barked is fluffy

Distinct Meanings

- Every fluffy dog barked
- Every dog that barked is fluffy

- The fluffy dog barked
- The dog that barked is fluffy

- Every student passed the exam
- No student failed the exam

Combinatorics of Composition

- A {dog, cat, ...} {saw, chased, ...} a {dog, cat, ...}

Combinatorics of Composition

- A {dog, cat, ...} {saw, chased, ...} a {dog, cat, ...}
- A {dog, cat, ...} {saw, chased, ...} a {dog, cat, ...} which {saw, chased, ...} a {dog, cat, ...}

Combinatorics of Composition

- A {dog, cat, …} {saw, chased, …} a {dog, cat, …}
- A {dog, cat, …} {saw, chased, …} a {dog, cat, …} which {saw, chased, …} a {dog, cat, …}
- Exponential growth in distinct meanings

Limits of Vector Representations

- Exponential growth in distinct meanings

Limits of Vector Representations

- Exponential growth in distinct meanings
- For vectors, one of the following must hold:
  - Magnitudes grows exponentially
  - Sensitive to arbitrarily small changes
  - Lossy

Limits of Vector Representations

- Exponential growth in distinct meanings
- For vectors, one of the following must hold:
  - Magnitudes grows exponentially
  - Sensitive to arbitrarily small changes
  - Lossy
- Lossy representations can still be useful,
  but don’t give us a full semantic theory

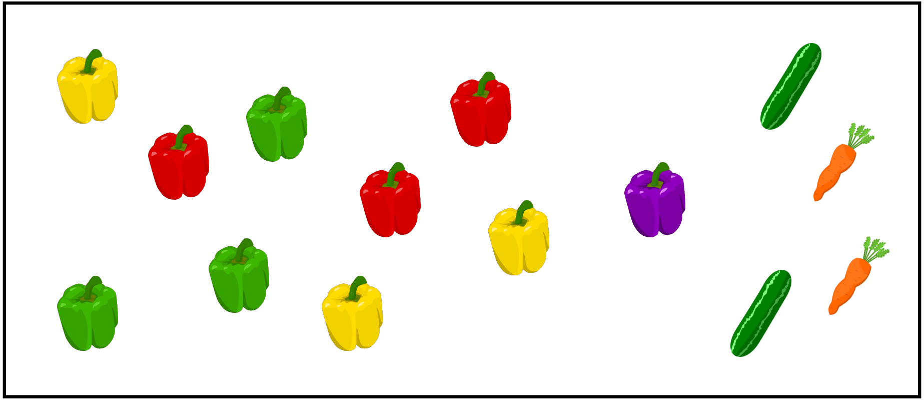
Compositional Semantics

- Compositional by design:
  - Predicate logic with lambda calculus
  - Semantic graphs with graph rewriting

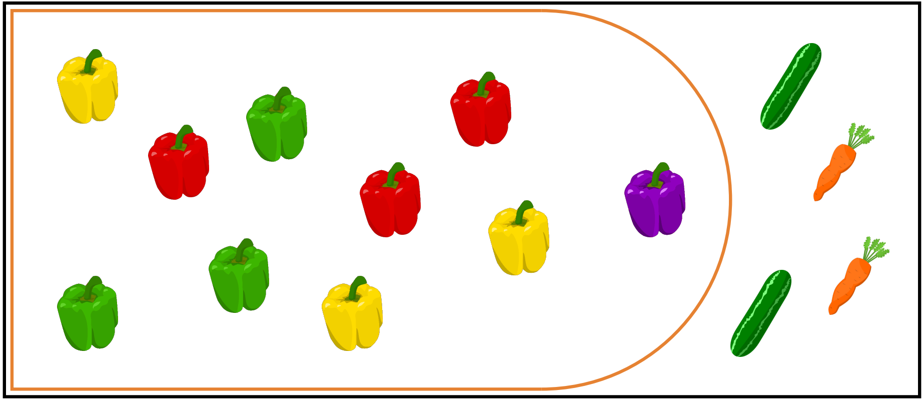
Compositional Semantics

- Compositional by design:
  - Predicate logic with lambda calculus
  - Semantic graphs with graph rewriting
- Bring compositional structure into distributional semantics?

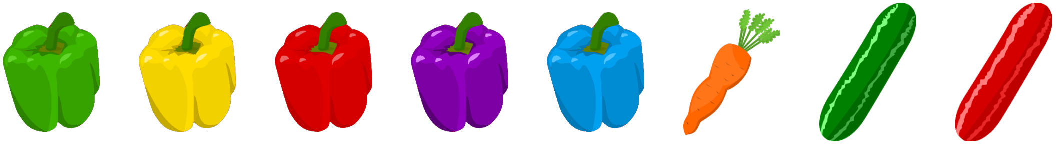
Truth-Conditional Semantics



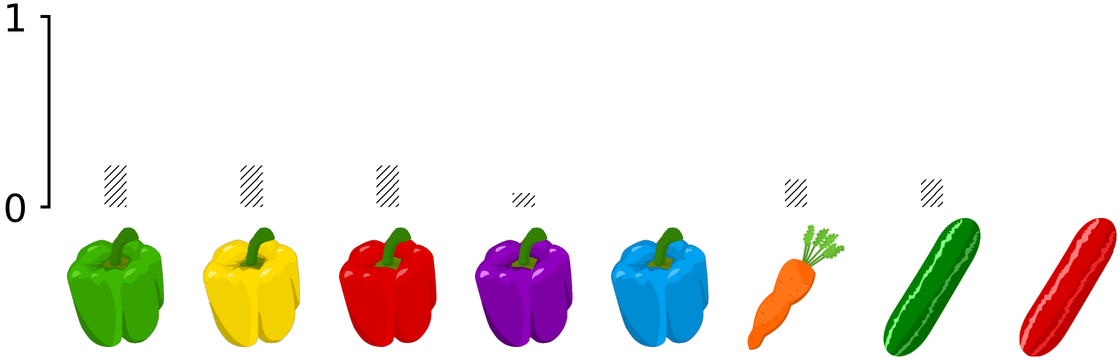
Truth-Conditional Semantics



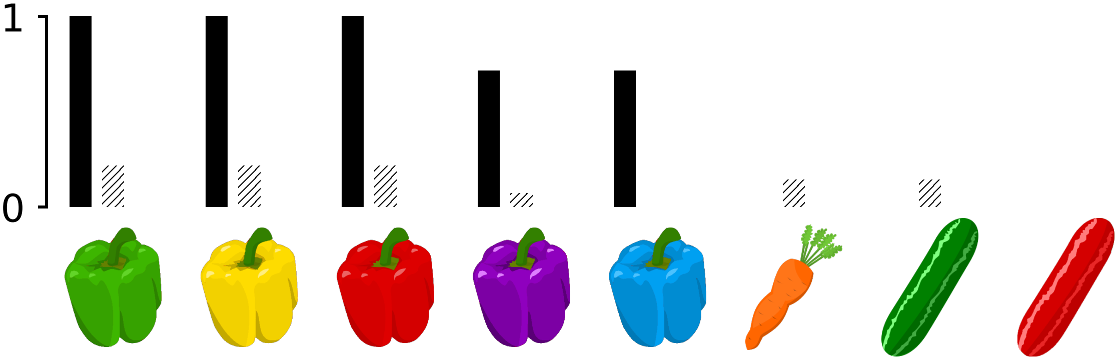
Truth-Conditional Functions



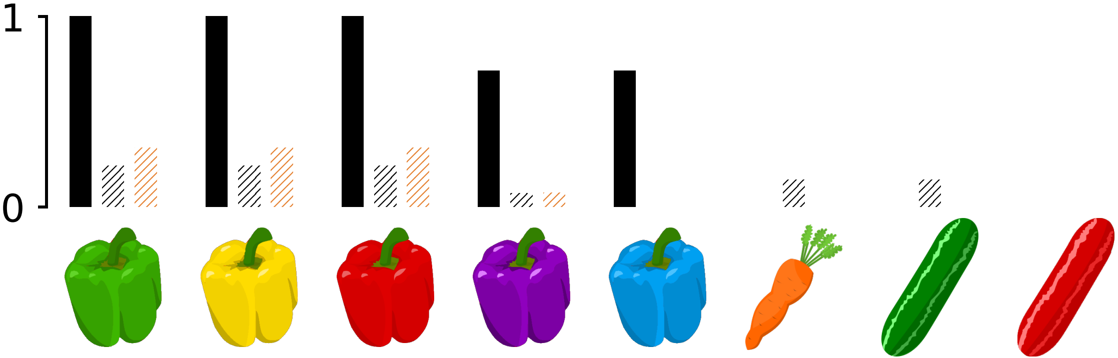
Truth-Conditional Functions



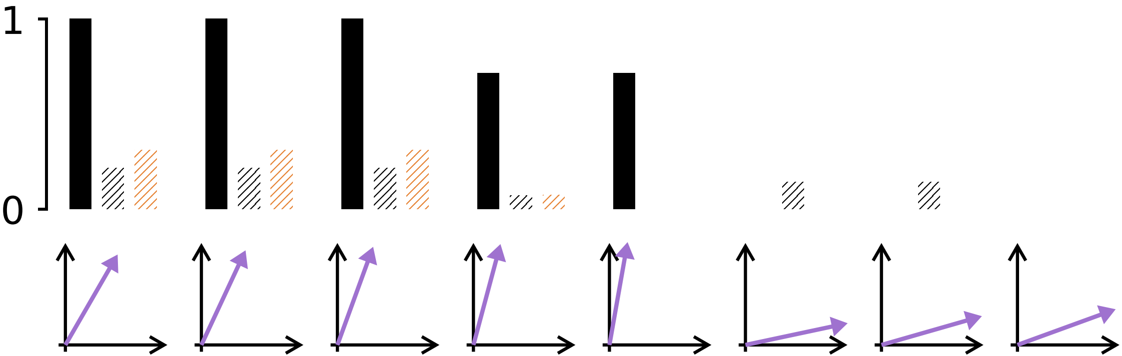
Truth-Conditional Functions



Truth-Conditional Functions



Truth-Conditional Functions



Summary of What's New

- Pixie: entity representation
- Word meanings as functions:
  pixie $\mapsto$ probability of truth

Situation Semantics

x

pepper(x)

Situation Semantics

x

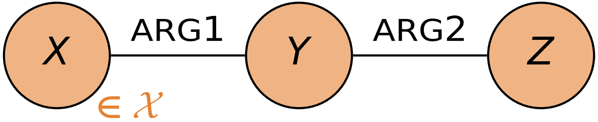
pepper(x)
vegetable(x)
animal(x)
dog(x)
cat(x)

Situation Semantics

$$x \xleftarrow{\text{ARG1}} y \xrightarrow{\text{ARG2}} z$$

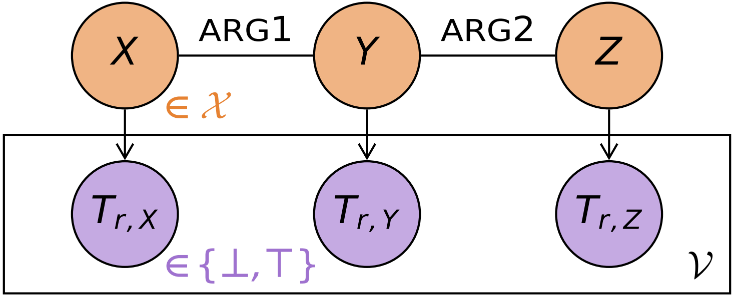
dog(x)	chase(y)	cat(z)
animal(x)	pursue(y)	animal(z)
chase(x)	dog(y)	chase(z)
pursue(x)	cat(y)	pursue(z)
cat(x)	animal(y)	dog(z)

Probabilistic Situation Semantics



dog(X)	chase(Y)	cat(Z)
animal(X)	pursue(Y)	animal(Z)
chase(X)	dog(Y)	chase(Z)
pursue(X)	cat(Y)	pursue(Z)
cat(X)	animal(Y)	dog(Z)

Probabilistic Situation Semantics



Probabilistic Situation Semantics

- World model: $\mathbb{P}(x, y, z)$
  (Joint distribution of pixie-valued random variables)

- Lexical model: $\mathbb{P}(t_{r,X} | x)$
  (Conditional distribution of truth-valued random variables, given a pixie)

Distributional Semantics

- What should the model learn?
- How can the model learn it?

Distributional Semantics

- What should the model learn?
  - Probabilistic situation semantics
- How can the model learn it?

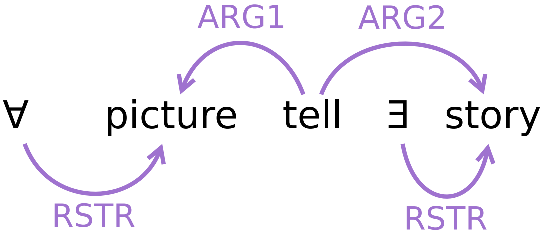
Distributional Semantics

- What should the model learn?
  - Probabilistic situation semantics
- How can the model learn it?
  - Probabilistic graphical model
  - Data: semantic dependency graphs

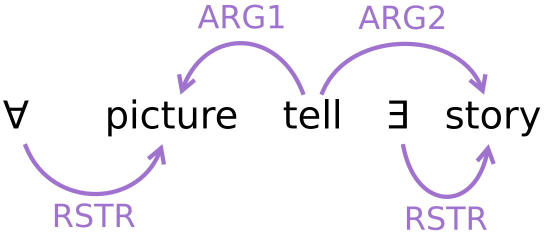
Dependency Minimal Recursion Semantics

Every picture tells a story

Dependency Minimal Recursion Semantics

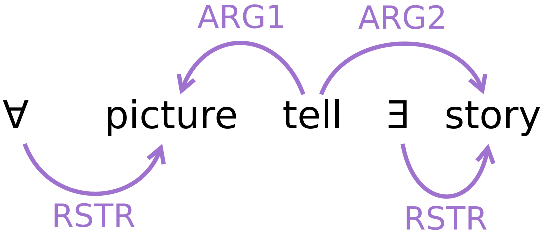


Dependency Minimal Recursion Semantics



$$\forall x \exists y \exists z \ \text{picture}(x) \Rightarrow \big[\,\text{story}(z) \wedge \text{tell}(y)$$
$$\wedge\ \text{ARG1}(y, x) \wedge \text{ARG2}(y, z)\,\big]$$

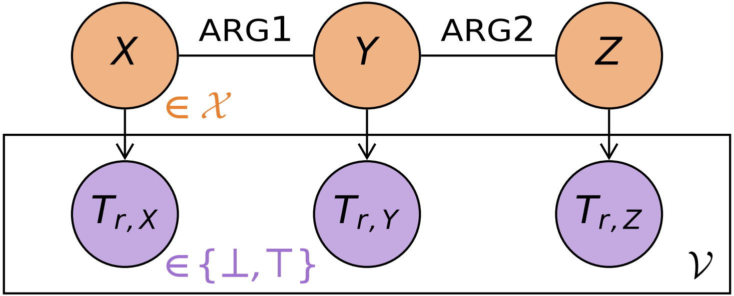
Dependency Minimal Recursion Semantics



$$\forall x \exists y \exists z \; \text{picture}(x) \Rightarrow \big[\, \text{story}(z) \wedge \text{tell}(y) \\ \wedge \text{ARG1}(y, x) \wedge \text{ARG2}(y, z) \,\big]$$

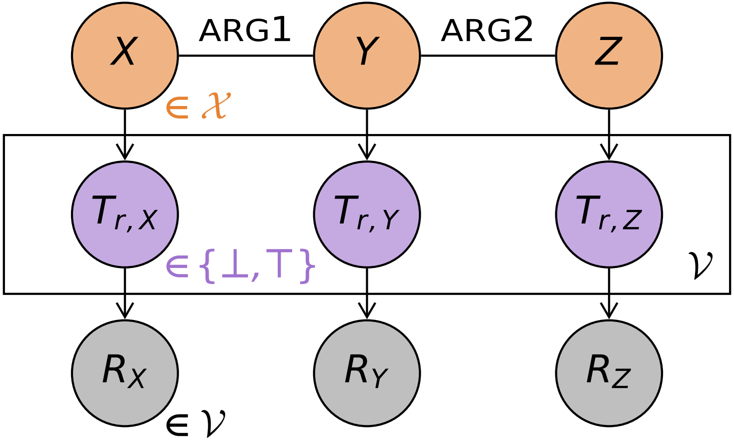
- See: "Linguists Who Use Probabilistic Models Love Them: Quantification in Functional Distributional Semantics" (PaM2020)

Functional Distributional Semantics



$\text{dog} \xleftarrow{\text{ARG1}} \text{chase} \xrightarrow{\text{ARG2}} \text{cat}$

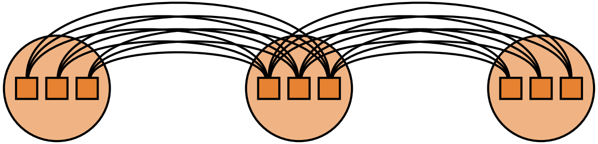
Functional Distributional Semantics



Functional Distributional Semantics

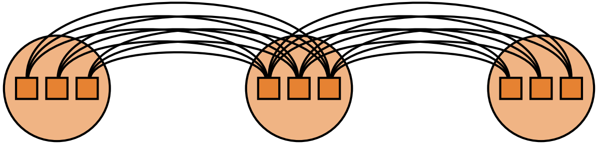
- Latent situation semantics
  - World model: $\mathbb{P}(x, y, z)$
  - Lexical model: $\mathbb{P}(t_{r,X} \mid x)$
- Observed DMRS graphs
  - Extended lexical model: $\mathbb{P}(r_X \mid x) \propto \mathbb{P}(t_{r,X} \mid x)$
    (For simplicity, probability of utterance assumed proportional to probability of truth)

World Model



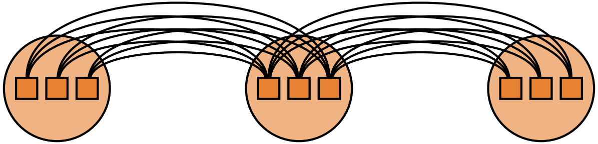
- Restricted Boltzmann Machine (binary vectors)

World Model



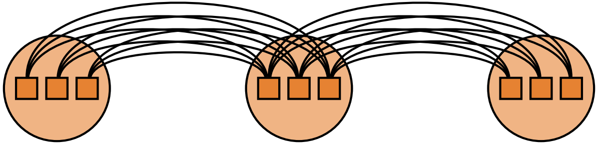
- Restricted Boltzmann Machine (binary vectors)
- Fabiani 2021 MPhil project: prototype version using real-valued vectors and Gaussian distributions

World Model



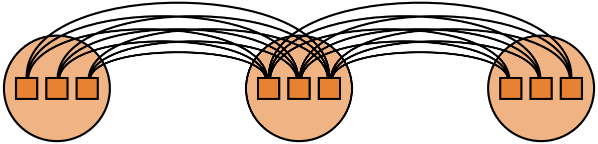
- Restricted Boltzmann Machine (binary vectors)

World Model



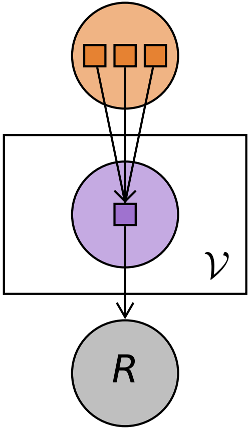
- Restricted Boltzmann Machine (binary vectors)
- $\mathbb{P}(s) \propto \exp\left(-E(s)\right)$

World Model



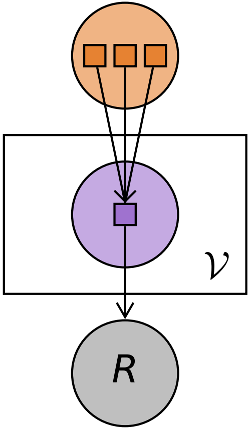
- Restricted Boltzmann Machine (binary vectors)
- $\mathbb{P}(s) \propto \exp\left(\sum_{x \xrightarrow{L} y \text{ in } s} w_{ij}^{(L)} x_i y_j \right)$

Lexical Model



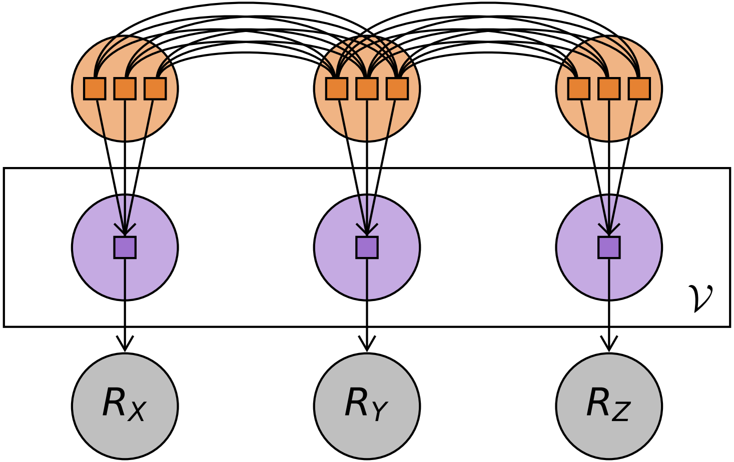
- Feedforward networks
- $\mathbb{P}\left(t^{(r,X)}\,\middle|\,x\right) = \sigma\left(v_i^{(r)} x_i\right)$

Lexical Model



- Feedforward networks
- $\mathbb{P}\left(t^{(r,X)} \middle| x\right) = \sigma\left(v_i^{(r)} x_i\right)$
- $\mathbb{P}\left(r^{(X)} \middle| x\right) \propto \mathbb{P}\left(t^{(r,X)} \middle| x\right)$

Functional Distributional Semantics



Gradient Descent

$$\frac{\partial}{\partial\theta}\log\mathbb{P}(g) = \mathbb{E}_{s|g}\left[\frac{\partial}{\partial\theta}\log\mathbb{P}(s)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial\theta}\log\mathbb{P}(g\,|\,s)\right]$$

Gradient Descent

$$\frac{\partial}{\partial \theta} \log \mathbb{P}(g) = \mathbb{E}_{s|g}\left[\frac{\partial}{\partial \theta} \log \mathbb{P}(s)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial \theta} \log \mathbb{P}(g \mid s)\right]$$

Gradient Descent

$$\frac{\partial}{\partial\theta}\log\mathbb{P}(g) = \mathbb{E}_{s|g}\left[\frac{\partial}{\partial\theta}\log\mathbb{P}(s)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial\theta}\log\mathbb{P}(g\,|\,s)\right]$$

- Latent variables necessary but inconvenient

Gradient Descent

$$\frac{\partial}{\partial\theta}\log\mathbb{P}(g) = \mathbb{E}_{s|g}\left[\frac{\partial}{\partial\theta}\log\mathbb{P}(s)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial\theta}\log\mathbb{P}(g\,|\,s)\right]$$

- Latent variables necessary but inconvenient
- Need approximation (e.g. amortised variational inference: train neural net to do it)

Pixie Autoencoder (Emerson, 2020)

- Generative model (world model & lexical model)
- Inference network (approximate inference)

Pixie Autoencoder (Emerson, 2020)

- Generative model (world model & lexical model)
- Inference network (approximate inference)
- Unique selling point:
  - Truth-conditional distributional semantics

Training Needs Graphs

- Training needs dependency graphs, not raw text

Training Needs Graphs

- Training needs dependency graphs, not raw text
- WikiWoods
  - English Wikipedia, parsed into DMRS graphs
  - 31 million graphs (after preprocessing)

Training Needs Graphs

- Training needs dependency graphs, not raw text
- WikiWoods
  - English Wikipedia, parsed into DMRS graphs
  - 31 million graphs (after preprocessing)
  - (So far, only verbs with ARG1 & ARG2 nouns)

Sanity Check: Lexical Similarity

- Lexical similarity: given two words (out of context), how similar are they?

Sanity Check: Lexical Similarity

- Lexical similarity: given two words (out of context), how similar are they?
- Competitive with state of the art
- Can distinguish similarity (*mouse*, *rat*) from relatedness (*law*, *lawyer*)

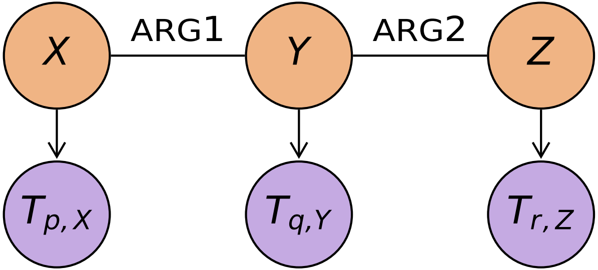
Similarity in Context (GS2011)

- Controlled semantic evaluation
- Starts to use expressiveness of functional model

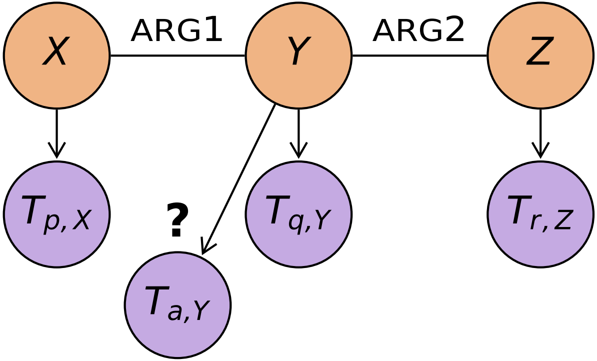
Similarity in Context (GS2011)

student	write	name
student	spell	name
scholar	write	book
scholar	spell	book

Pixie Autoencoder for GS2011

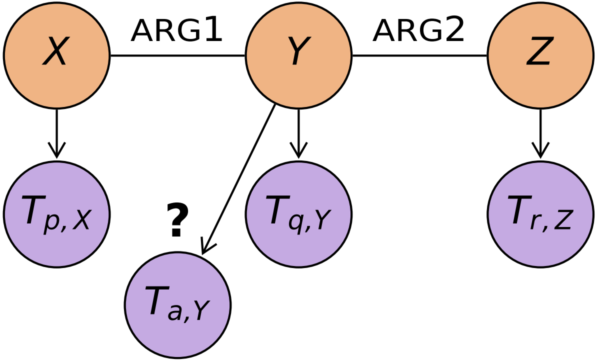


Pixie Autoencoder for GS2011



$$\mathbb{P}\left(t_{a,Y} \,\middle|\, t_{p,X}, t_{q,Y}, t_{r,Z}\right)$$

Pixie Autoencoder for GS2011



$$\mathbb{P}\left(t_{spell,Y} \,\middle|\, t_{student,X}, t_{write,Y}, t_{name,Z}\right)$$

BERT for GS2011

Pseudo-logical form: (employer provide training)

- “an employer **provides** training .”
- “employer **provides** training .”
- “an employer **provides** a training .”
- “a employer **provides** a training .”
- “employers **provide** training .”
- “employers **provide** trainings .”
- “training is **provided** by an employer .”
- “trainings are **provided** by employers .”
- ...

GS2011 Results

Model	Correlation
Skip-gram (vector addition)	.348
BERT (with tuned template strings)	.446
Pixie Autoencoder	.504

- Smaller model, less data, better performance

RELPRON Dataset (Rimell et al., 2016)

- Controlled semantic evaluation
- Starts to use expressiveness of functional model

RELPRON Dataset (Rimell et al., 2016)

- Controlled semantic evaluation
- Starts to use expressiveness of functional model
- Large gap between human performance (~100%) and state of the art (~50%)

RELPRON Dataset (Rimell et al., 2016)

telescope	*device that astronomers use*
telescope	*device that detects planets*
saw	*device that cuts wood*
philosopher	*person that defends rationalism*
survivor	*person that helicopter saves*
farming	*activity that soil supports*
...	*...*

RELPRON Dataset (Rimell et al., 2016)

telescope	*device that astronomers use*
	device that detects planets
	device that cuts wood
	person that defends rationalism
	person that helicopter saves
	activity that soil supports
	...

RELPRON Dataset (Rimell et al., 2016)

saw	*device that astronomers use*
	device that detects planets
	device that cuts wood
	person that defends rationalism
	person that helicopter saves
	activity that soil supports
	…

RELPRON Dataset (Rimell et al., 2016)

philosopher	*device that astronomers use*
	device that detects planets
	device that cuts wood
	person that defends rationalism
	person that helicopter saves
	activity that soil supports
	...

RELPRON Dataset (Rimell et al., 2016)

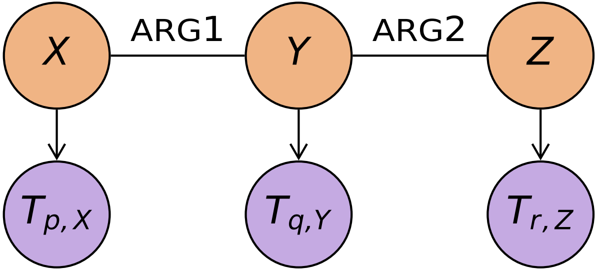
soil	*device that astronomers use*
	device that detects planets
	device that cuts wood
	person that defends rationalism
	person that helicopter saves
	activity that soil supports
	...

RELPRON Dataset (Rimell et al., 2016)

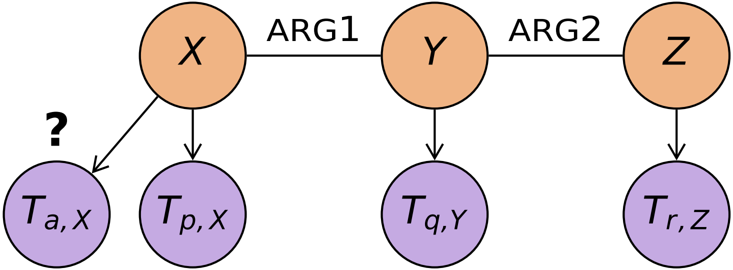
soil

device that astronomers use
device that detects planets
device that cuts wood
person that defends rationalism
person that helicopter saves
activity that soil supports
...

Logical Inference for RELPRON

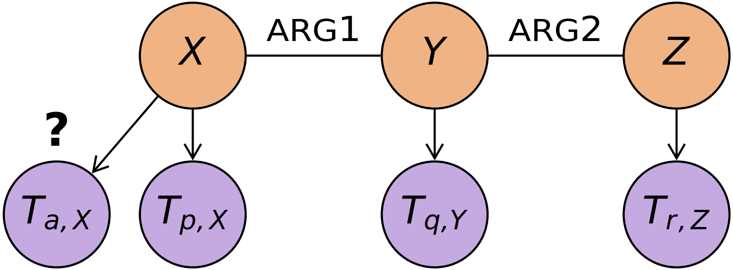


Logical Inference for RELPRON



$$\mathbb{P}\left(t_{a,X} \,\middle|\, t_{p,X}, t_{q,Y}, t_{r,Z}\right)$$

Logical Inference for RELPRON



$$\mathbb{P}\left(t_{a,X} \,\middle|\, t_{p,X}, t_{q,Y}, t_{r,Z}\right)$$

$$\mathbb{P}\left(t_{philosopher,X} \,\middle|\, t_{person,X}, t_{defend,Y}, t_{rationalism,Z}\right)$$

BERT for RELPRON

Pseudo-logical form: (person that defend rationalism)

- “A person that defends rationalism is a **[MASK]** .”
- “Person that defends rationalism is **[MASK]** .”
- “A person that defends a rationalism is a **[MASK]** .”
- “People that defend rationalisms are **[MASK]** .”
- “A **[MASK]** is a person that defends rationalism .”
- “A **[MASK]** is a person that defends a rationalism .”
- “A **person** that defends rationalism .”
- “A **person** that defends a rationalism .”
- ...

RELPRON Results

Model	MAP
Simp. Prac. Lex. Func. (Rimell et al., 2016)	.497
Dependency vectors (Czarnowska et al., 2019)	.439
Word2Vec	.474
BERT (with carefully tuned template strings)	.186
BERT & Word2Vec ensemble	.479
Pixie Autoencoder	.189
Pixie Autoencoder & Word2Vec ensemble	.489

RELPRON Results

Model	MAP
Simp. Prac. Lex. Func. (Rimell et al., 2016)	.497
Dependency vectors (Czarnowska et al., 2019)	.439
Word2Vec	.474
BERT (with carefully tuned template strings)	**.186**
BERT & Word2Vec ensemble	.479
Pixie Autoencoder	**.189**
Pixie Autoencoder & Word2Vec ensemble	.489

RELPRON Results

Model	MAP
Simp. Prac. Lex. Func. (Rimell et al., 2016)	.497
Dependency vectors (Czarnowska et al., 2019)	.439
Word2Vec	**.474**
BERT (with carefully tuned template strings)	.186
BERT & Word2Vec ensemble	**.479**
Pixie Autoencoder	.189
Pixie Autoencoder & Word2Vec ensemble	**.489**

RELPRON Conclusion

- Pixie Autoencoder compared to BERT:
  - More data efficient (1.2% no. tokens)
  - Doesn’t require tuning to apply
  - More “different” from Word2Vec

RELPRON Conclusion

- Pixie Autoencoder compared to BERT:
  - More data efficient (1.2% no. tokens)
  - Doesn’t require tuning to apply
  - More “different” from Word2Vec
- Word2Vec still state of the art
  - Error analysis: good at relatedness
  - Need “topic” in world model?

Ongoing/Future Work

- Joint learning with grounded data
- Joint learning with lexical resources
- More efficient model (continuous pixies)
- Latent variable for “topic”
- Correlated truth values (for pragmatics)
- Deeper networks (for polysemy)
- Semi-compositional idioms
- More general logical inferences

Ongoing/Future Work

- Joint learning with grounded data
- Joint learning with lexical resources
- More efficient model (continuous pixies)
- Latent variable for "topic"
- Correlated truth values (for pragmatics)
- Deeper networks (for polysemy)
- Semi-compositional idioms
- More general logical inferences

Joint Learning with Grounded Data

- Fundamental distinction between words and entities

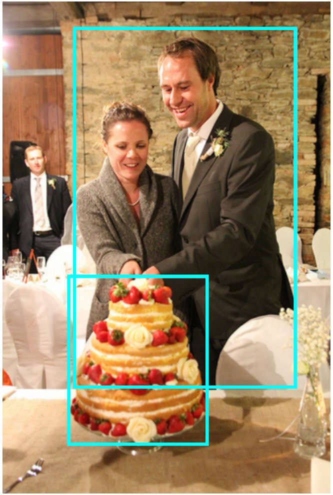
Joint Learning with Grounded Data

- Fundamental distinction between words and entities
- Vector space models:
  - Early fusion, late fusion, cross-modal maps...

Joint Learning with Grounded Data

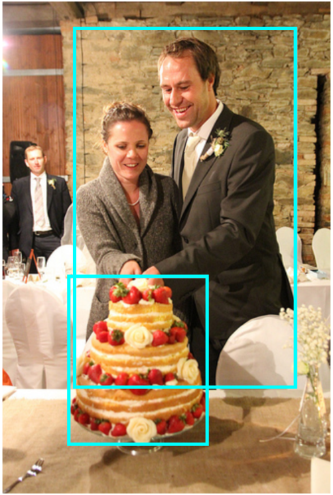
- Fundamental distinction between words and entities
- Vector space models:
  - Early fusion, late fusion, cross-modal maps…
- Functional Distributional Semantics:
  - Text → pixies are latent
  - Grounded data → pixies are observed

Visual Genome Dataset

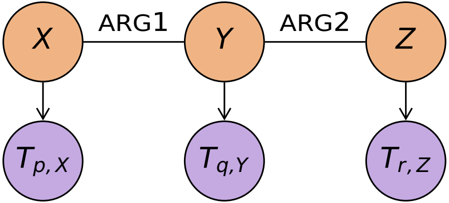


"couple cutting cake"

Visual Genome Dataset



"couple cutting cake"



Visual Genome Semantics

- Liu 2021 MPhil project: learn functional model from Visual Genome
- (Not joint learning… yet)

Visual Genome Semantics

- Liu 2021 MPhil project: learn functional model from Visual Genome
- (Not joint learning… yet)

Model	MEN	SL999	GS2011	RELPRON
VG-count (Herbelot, 2020)	.336	.224	.063	.038
EVA (Herbelot, 2020)	.543	.390	.068	.032
Functional	.639	.431	.171	.117

Conclusion

- Distributional semantics: more than just similarity
- Compositionality $\neq$ Context dependence
- Vectors: useful but have fundamental limitations
- Truth-conditional distributional semantics: feasible, but long road ahead!