Learning meaning in a logically structured model

An introduction to Functional Distributional Semantics

Guy Emerson

What I’ll Cover...

- Meanings as *functions*

What I'll Cover...

- Meanings as *functions*
- Logically interpretable model

What I’ll Cover…

- Meanings as *functions*
- Logically interpretable model
- Outperforms BERT at semantics

What I'll Cover...

- Meanings as *functions*
- Logically interpretable model
- Outperforms BERT at semantics
- Clear path for multimodal learning

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
- Two questions:
  - What should the model learn?
  - How can the model learn it?

What should the model learn?

- Vectors?

What should the model learn?

- Vectors?
  - Long history of attempts…
  - See: “What are the goals of distributional semantics?” (ACL 2020)

What should the model learn?

- Vectors?
  - Long history of attempts…
  - See: “What are the goals of distributional semantics?” (ACL 2020)
- Back to fundamentals: truth-conditional semantics

Words are not Entities

- Fundamental distinction between:
  - Words
  - Entities they refer to

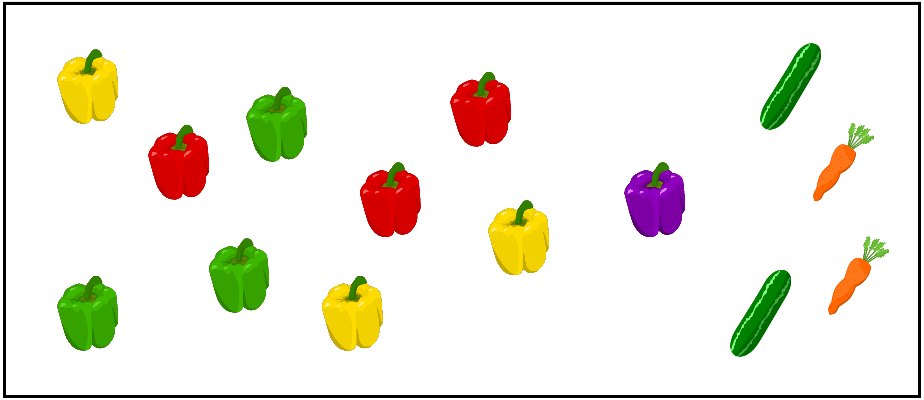
Words are not Entities

- Fundamental distinction between:
  - Words
  - Entities they refer to
- Important for discourse: anaphora resolution, question answering, dialogue processing...

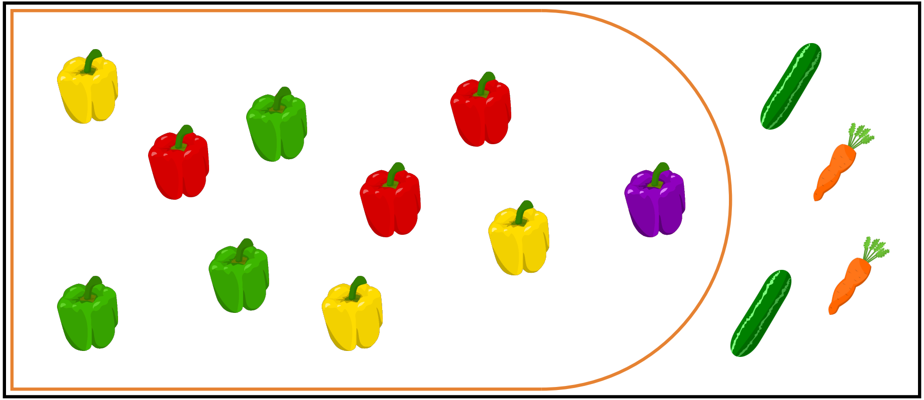
Words are not Entities

- Fundamental distinction between:
  - Words
  - Entities they refer to
- Important for discourse: anaphora resolution, question answering, dialogue processing…
- Meaning as a function over entities

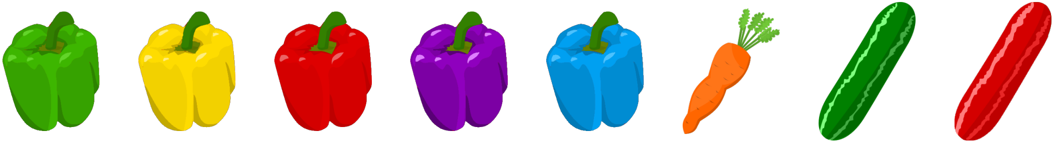
Truth-Conditional Semantics



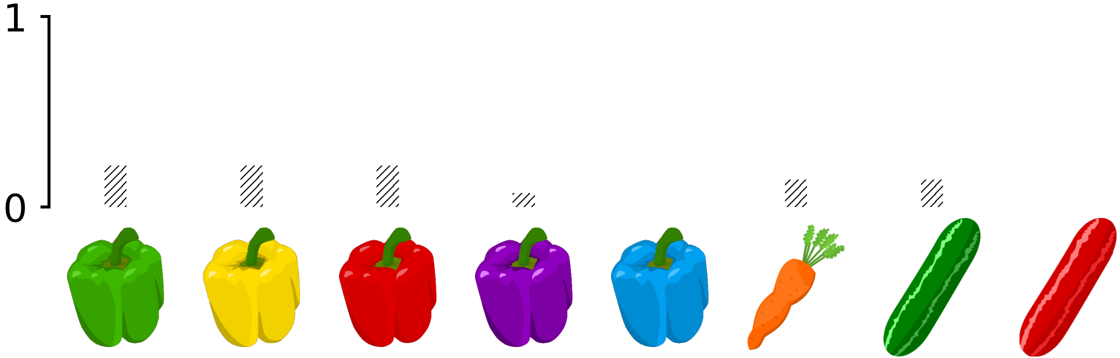
Truth-Conditional Semantics



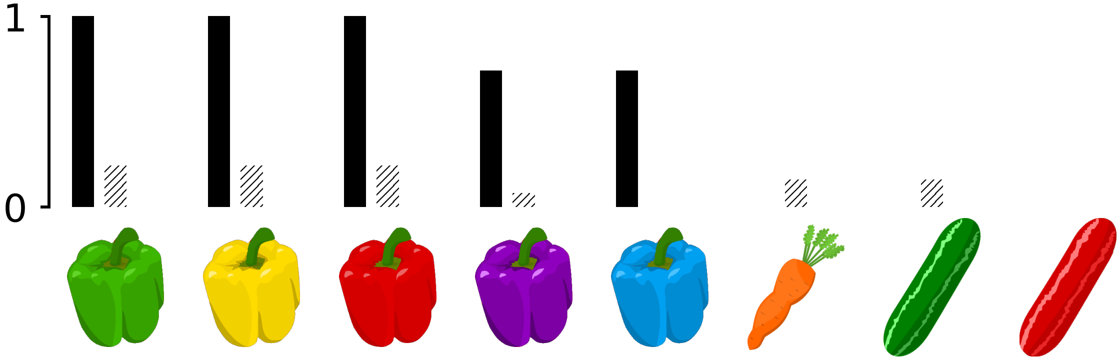
Truth-Conditional Functions



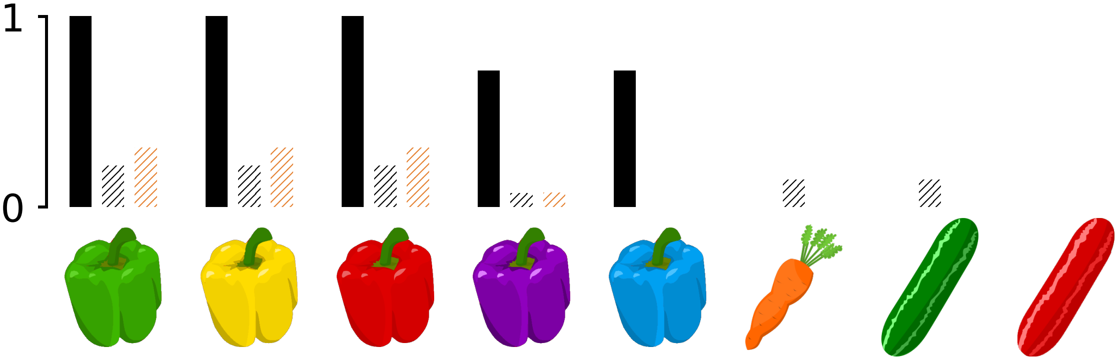
Truth-Conditional Functions



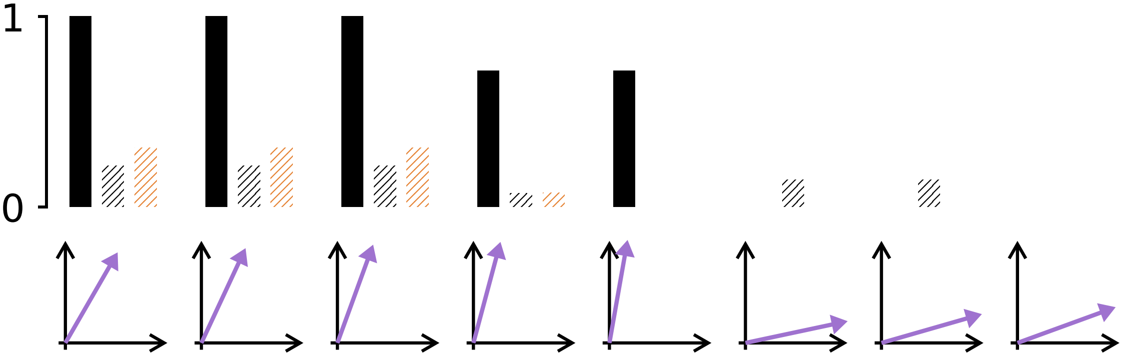
Truth-Conditional Functions



Truth-Conditional Functions



Truth-Conditional Functions



Summary of What's New

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

Summary of What's New

- Pixie: entity representation
- Word meanings as functions:
  pixie $\mapsto$ probability of truth
- (For deeper discussion, see: "Probabilistic Lexical Semantics: From Gaussian Embeddings to Bernoulli Fields", chapter in "Probabilistic Approaches to Linguistic Theory", 2022, CSLI Publications)

Situation Semantics

x

pepper(x)

Situation Semantics

x

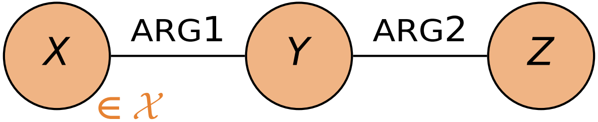
pepper(x)
vegetable(x)
animal(x)
dog(x)
cat(x)

Situation Semantics

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

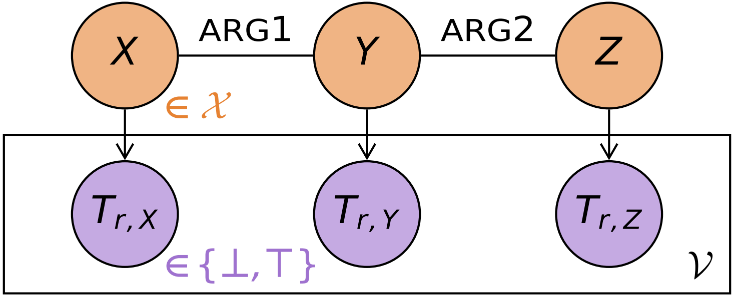
dog(x)	chase(y)	cat(z)
animal(x)	pursue(y)	animal(z)
chase(x)	dog(y)	chase(z)
pursue(x)	cat(y)	pursue(z)
cat(x)	animal(y)	dog(z)

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

Semantic Goals

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

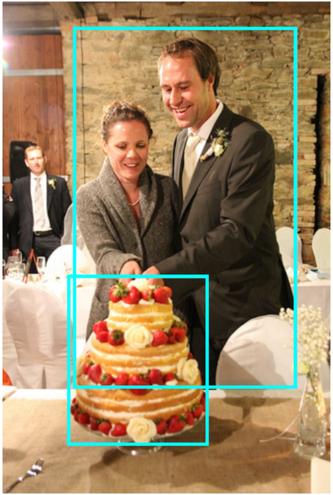
Semantic Goals

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

Semantic Goals

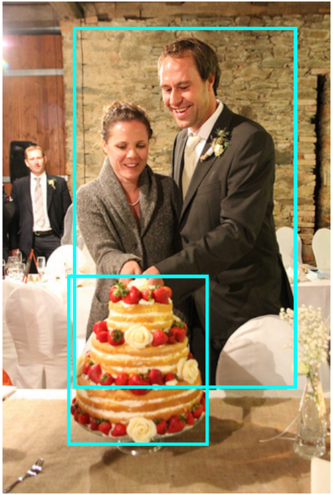
- What should the model learn?
  - Probabilistic situation semantics
- How can the model learn it?
  - Probabilistic graphical model
  - Data: annotated images

Visual Genome Dataset

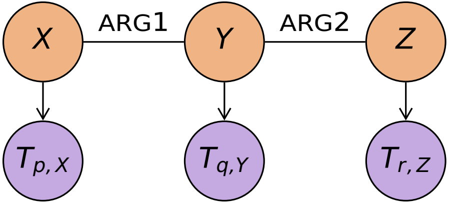


"couple cutting cake"

Visual Genome Dataset



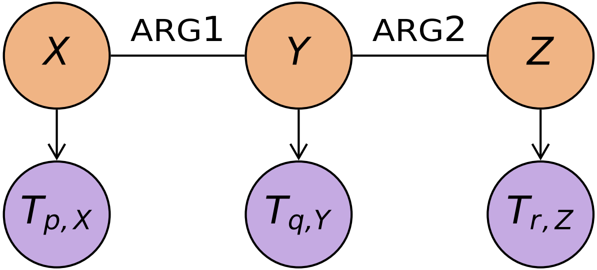
"couple cutting cake"



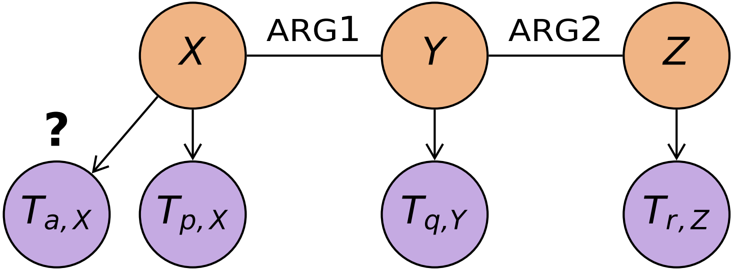
Liu and Emerson (2022)

- Image preprocessing: pixies given by pre-trained ResNet101
- World model: $\mathbb{P}(x, y, z)$ Gaussian
- Lexical model: $\mathbb{P}(t_{r,X} \mid x)$ one-layer sigmoid

Logical Reasoning with Latent Entities

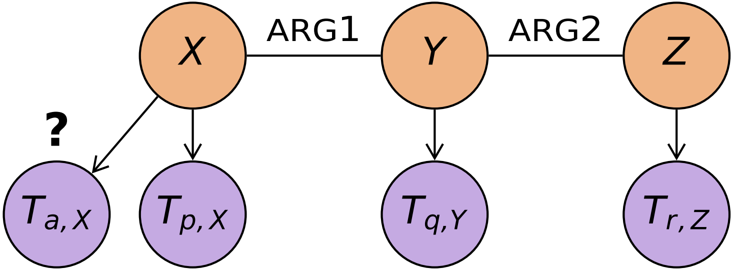


Logical Reasoning with Latent Entities



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

Logical Reasoning with Latent Entities



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

$$\mathbb{P}\left(t_{horse,X} \,|\, t_{animal,X}, t_{has,Y}, t_{tail,Z}\right)$$

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?

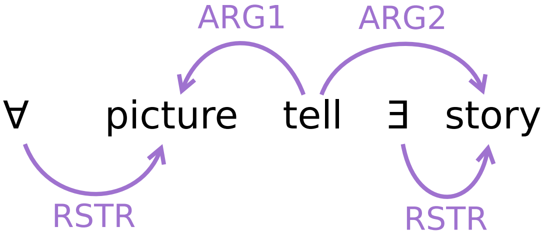
Distributional Semantics

- What should the model learn?
  - Probabilistic situation semantics
- How can the model learn it?
  - Probabilistic graphical model
    (all pixies are latent!)
  - Data: semantic dependency graphs

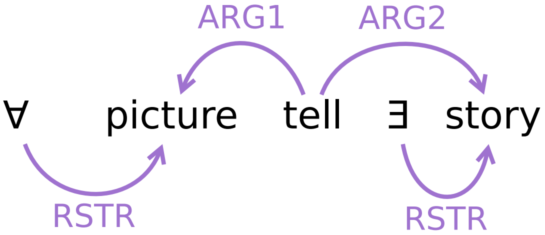
Dependency Minimal Recursion Semantics

Every picture tells a story

Dependency Minimal Recursion Semantics

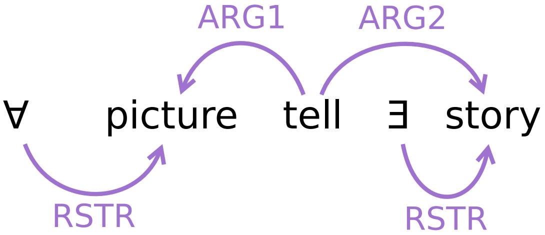


Dependency Minimal Recursion Semantics



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

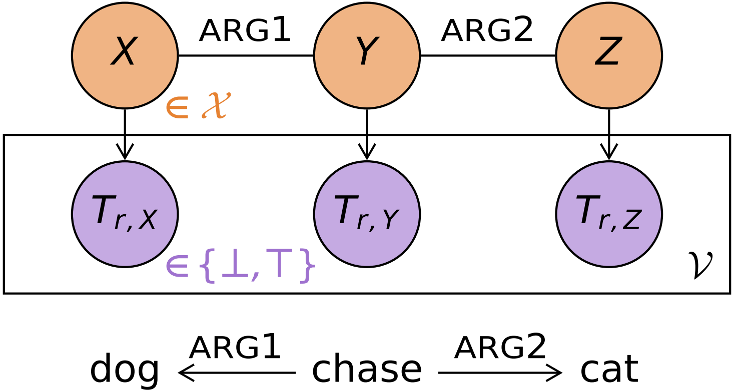
Dependency Minimal Recursion Semantics



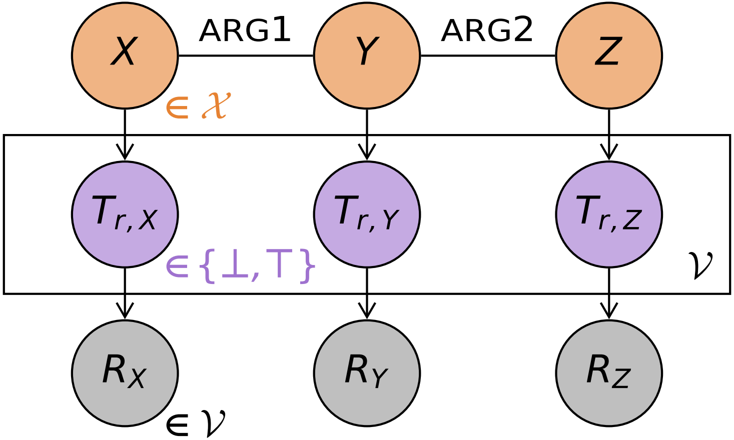
$$\forall x \exists y \exists z \; \text{picture}(x) \Rightarrow \left[\text{story}(z) \wedge \text{tell}(y) \right.$$
$$\left. \wedge \; \text{ARG1}(y, x) \wedge \text{ARG2}(y, z) \right]$$

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

Functional Distributional Semantics



Functional Distributional Semantics

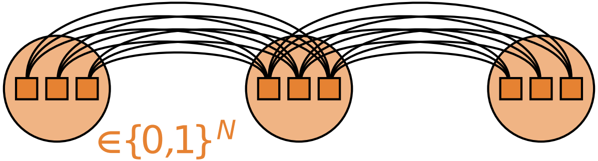


Functional Distributional Semantics

- Latent situation semantics
  - World model: $\mathbb{P}(x, y, z)$
  - Lexical model: $\mathbb{P}(t_{r,X} \mid x)$

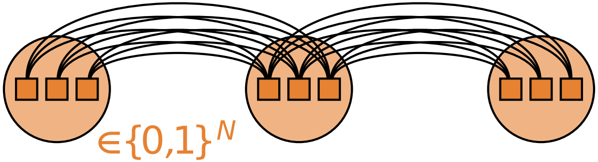
- Observed DMRS graphs
  - Extended lexical model: $\mathbb{P}(r_X \mid x) \propto \mathbb{P}(t_{r,X} \mid x)$
    (For simplicity, probability of utterance assumed proportional to probability of truth)

World Model



- Cardinality Restricted Boltzmann Machine
  (CaRBM; Swersky et al., 2012)

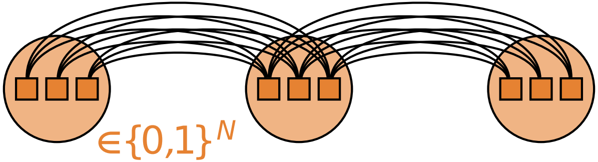
World Model



- Cardinality Restricted Boltzmann Machine
  (CaRBM; Swersky et al., 2012)

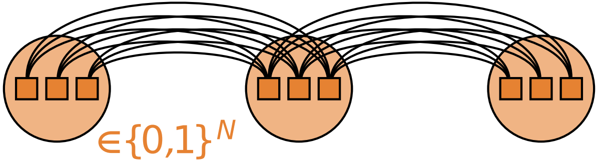
- (Gaussian MRF: work in progress, e.g. Fabiani, 2021)

World Model



- Cardinality Restricted Boltzmann Machine
  (CaRBM; Swersky et al., 2012)

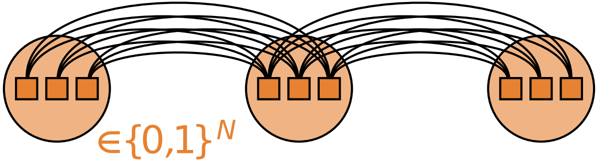
World Model



- Cardinality Restricted Boltzmann Machine
  (CaRBM; Swersky et al., 2012)

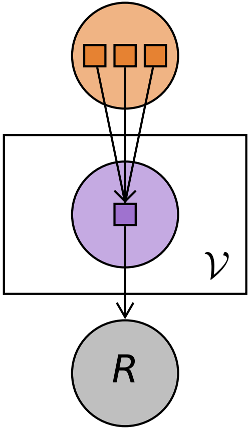
- $\mathbb{P}(s) \propto \exp\left(-E(s)\right)$

World Model



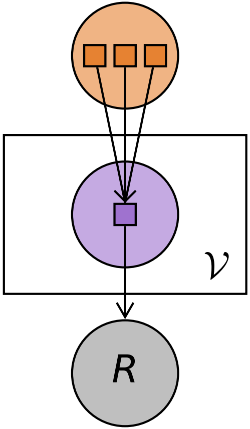
- Cardinality Restricted Boltzmann Machine
  (CaRBM; Swersky et al., 2012)
- $\mathbb{P}(s) \propto \exp\left(\sum_{x \xrightarrow{L} y \text{ in } s} w_{ij}^{(L)} x_i y_j\right)$

Lexical Model



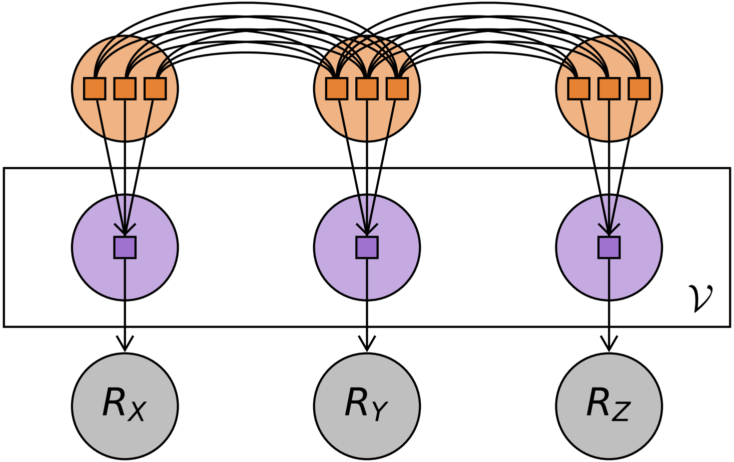
- Feedforward networks
- $\mathbb{P}\left(t^{(r,X)} \middle| x\right) = \sigma\left(v_i^{(r)} x_i\right)$

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

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

Gradient Descent

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

Gradient Descent

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

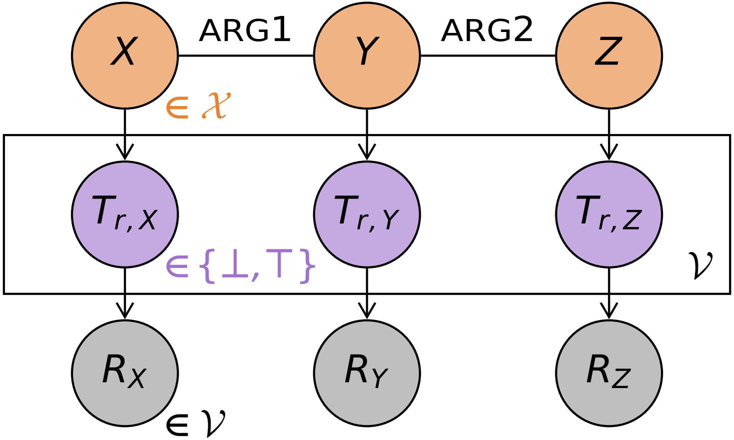
- Latent variables necessary but inconvenient

Gradient Descent

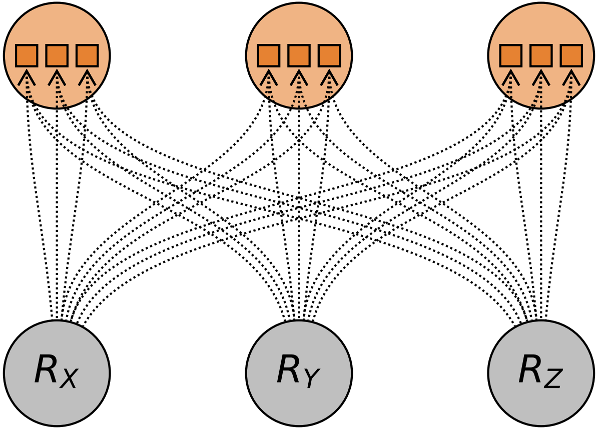
$$\frac{\partial}{\partial \theta} \log \mathbb{P}(g) = \left(\mathbb{E}_{s|g} - \mathbb{E}_s\right)\left[\frac{\partial}{\partial \theta}\left(-E(s)\right)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial \theta} \log \mathbb{P}(g \,|\, s)\right]$$

- Latent variables necessary but inconvenient
- Approximate distribution: variational inference
  (Jordan et al., 1999; Attias, 2000)

Functional Distributional Semantics



Variational Inference



Amortised Variational Inference

- Variational distribution must be optimised *for each input graph*

Amortised Variational Inference

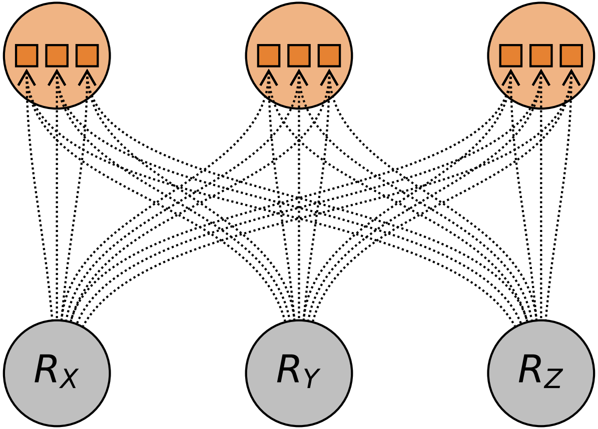
- Variational distribution must be optimised *for each input graph*

- Amortisation: train a network to predict the variational distribution (Kingma and Welling, 2014; Rezende et al., 2014; Mnih and Gregor, 2014)

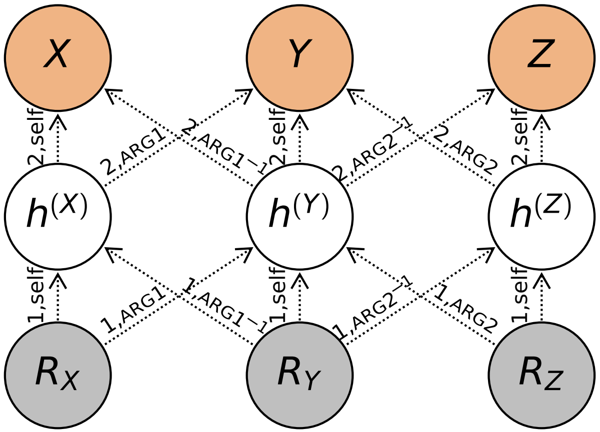
Amortised Variational Inference

- Variational distribution must be optimised *for each input graph*
- Amortisation: train a network to predict the variational distribution (Kingma and Welling, 2014; Rezende et al., 2014; Mnih and Gregor, 2014)
- Input graphs of different topologies: share network weights with graph convolutions (Duvenaud et al., 2015; Marcheggiani and Titov, 2017)

Variational Inference



Amortised Variational Inference



Amortised Variational Inference

$$\begin{aligned}
\frac{\partial}{\partial \phi} D(\mathbb{Q}|\mathbb{P}) = & -\frac{\partial}{\partial \phi} \mathbb{E}_{\mathbb{Q}(s)}\left[\log \mathbb{P}(s)\right] \\
& -\frac{\partial}{\partial \phi} \mathbb{E}_{\mathbb{Q}(s)}\left[\log \mathbb{P}\left(g \,|\, s\right)\right] \\
& -\frac{\partial}{\partial \phi} H(\mathbb{Q})
\end{aligned}$$

Gradient Descent

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

- Latent variables: amortised variational inference

Gradient Descent

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

- Latent variables: amortised variational inference
- Additional details... regularisation, dropout, β-VAE weighting, negative sampling, probit approximation, learning rate, warm start, soft constraints, belief propagation for $\mathbb{E}_s$...

Gradient Descent

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

- Latent variables: amortised variational inference
- See: "Autoencoding Pixies: Amortised Variational Inference with Graph Convolutions for Functional Distributional Semantics" (ACL 2020)

Pixie Autoencoder

- Generative model & inference network

Pixie Autoencoder

- Generative model & inference network
- 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)
  - (This talk: only verbs with ARG1 & ARG2 nouns; ongoing work: arbitrary graphs)

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*)

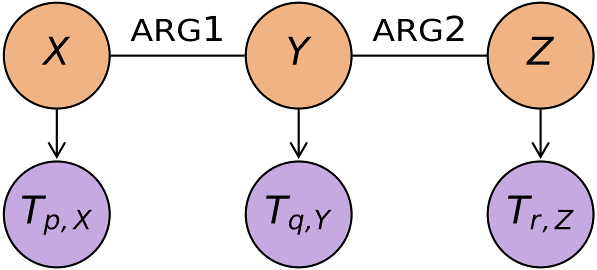
Similarity in Context (GS2011)

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

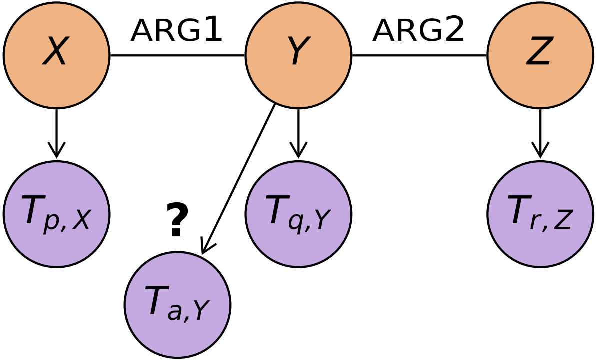
Similarity in Context (GS2011)

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

Pixie Autoencoder for GS2011

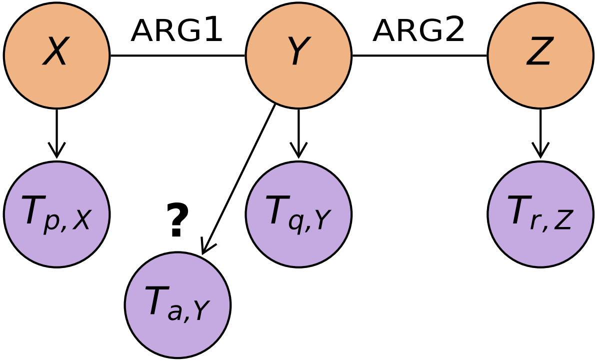


Pixie Autoencoder for GS2011



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

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)

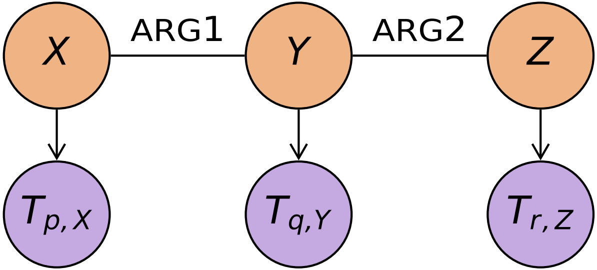
soil	*device that astronomers use*
	device that detects planets
	device that cuts wood
	person that defends rationalism
	person that helicopter saves
	activity that soil supports
	...

RELPRON Dataset (Rimell et al., 2016)

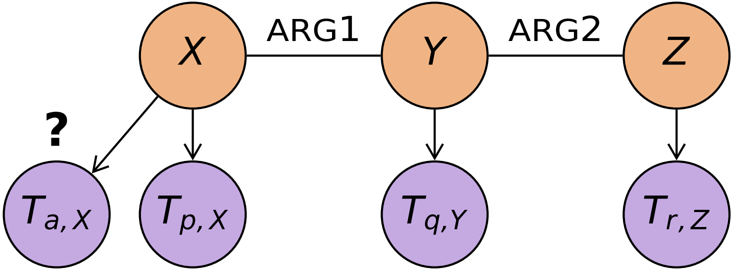
soil

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

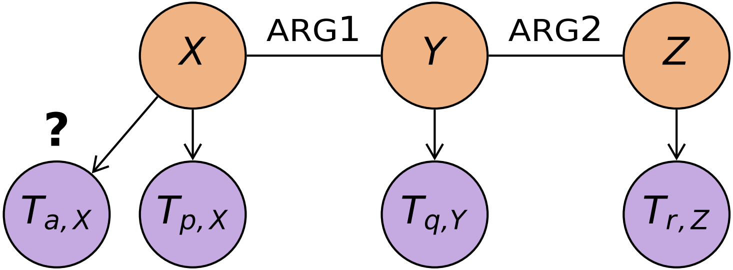


Logical Inference for RELPRON



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

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?

Visual Genome Semantics

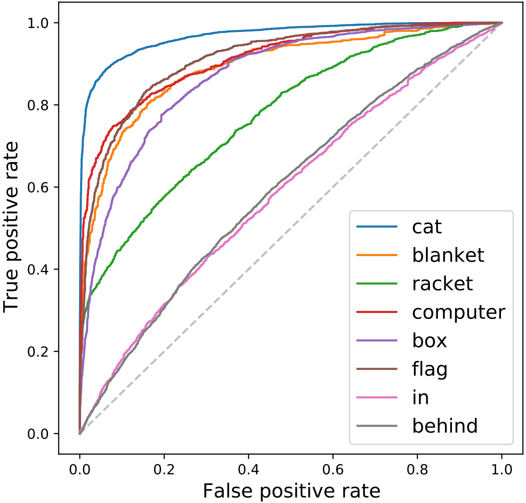
Model	MEN	SL999	GS2011	RELPRON
VG-count (Herbelot, 2020)	.336	.224	.063	.038
VG-retrieval	.420	.190	.072	.045
EVA (Herbelot, 2020)	.543	.390	.068	.032
Functional	.639	.431	.171	.117

Visual Genome Semantics

Model	MEN	SL999	GS2011	RELPRON
VG-count (Herbelot, 2020)	.336	.224	.063	.038
VG-retrieval	.420	.190	.072	.045
EVA (Herbelot, 2020)	.543	.390	.068	.032
Functional	.639	.431	.171	.117

- Truth-conditional structure helps generalisation

Classification accuracy per predicate



Visual Genome Summary

- Truth-conditional structure helps generalisation (even with a heavily simplified model!)
- Spatial relations are hard

Visual Genome Summary

- Truth-conditional structure helps generalisation (even with a heavily simplified model!)
- Spatial relations are hard
- Plausible path for joint learning...

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

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
  - Details need to be aligned...

Conclusion

- Meanings: functions
- Entities: (latent or observed) pixies
- Probabilistic logic: empirically useful