Autoencoding Pixies

Amortised Variational Inference with Graph Convolutions for Functional Distributional Semantics

Guy Emerson

What I’ll Cover…

- Meanings as *functions*, not vectors

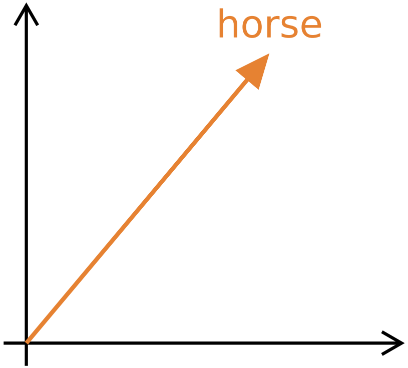
What I’ll Cover...

- Meanings as *functions*, not vectors
- Logically interpretable model

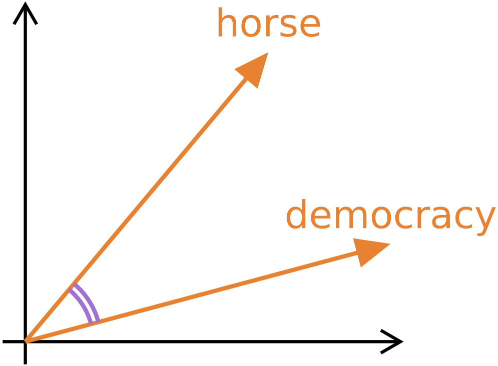
What I’ll Cover…

- Meanings as *functions*, not vectors
- Logically interpretable model
- Outperforms BERT at semantics

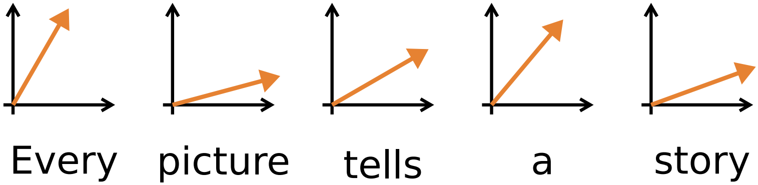
Vector Space Models



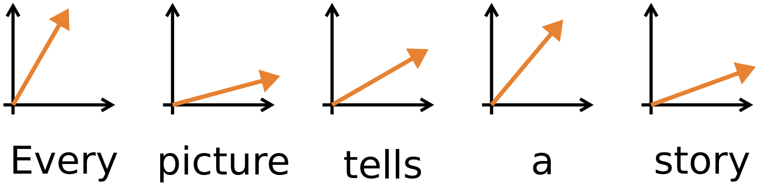
Vector Space Models



Vector Space Models for Sentences?

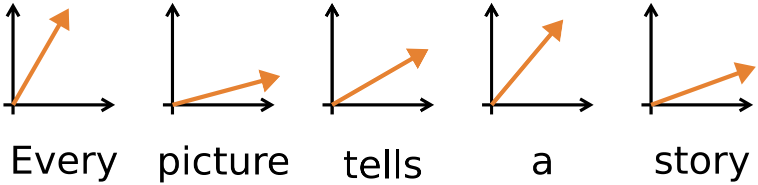


Vector Space Models for Sentences?



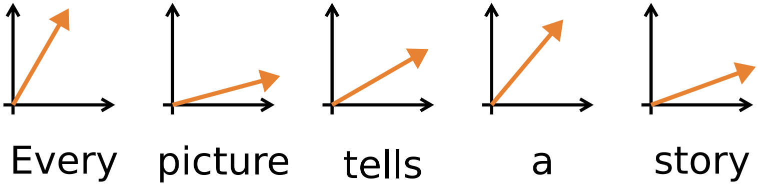
- Composition? Logic?

Vector Space Models for Sentences?



- Composition? Logic?
- Long history of attempts…
  - See: “What are the Goals of Distributional Semantics?”

Vector Space Models for Sentences?



- Composition? Logic?
- Long history of attempts…
- Rethink fundamentals → why vectors?

Words are not Entities

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

Words are not Entities

- Fundamental distinction between:
  - Words
  - Entities they refer to
- Meaning as a function over entities

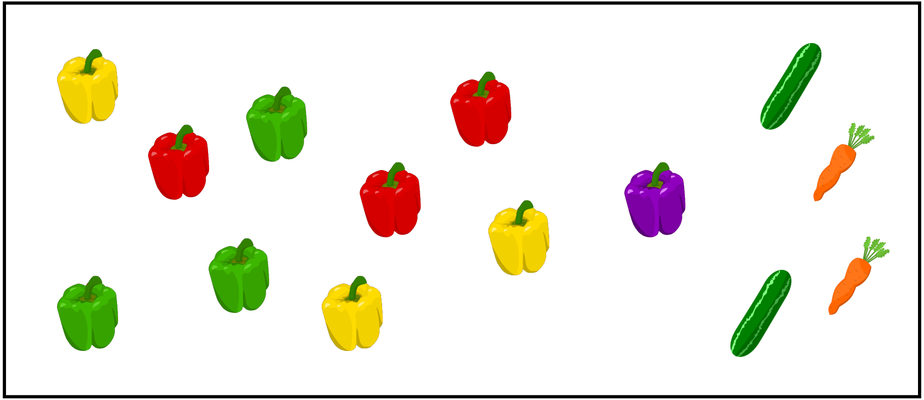
Overview

- *Functions*, not vectors
- Probabilistic graphical model

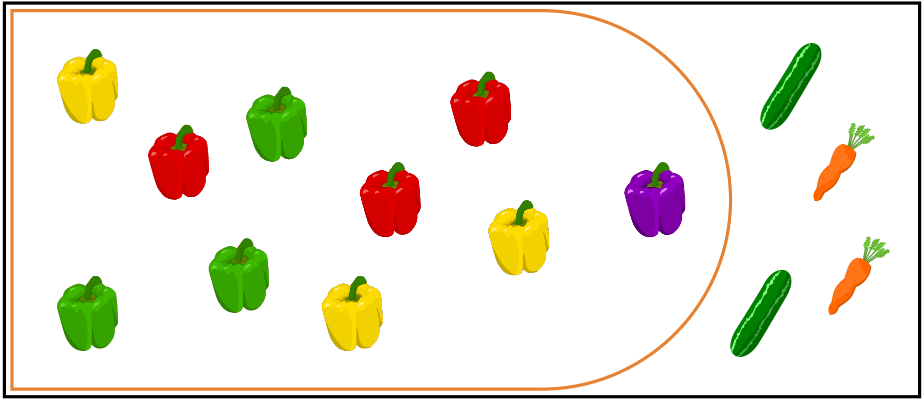
Overview

- *Functions*, not vectors
- Probabilistic graphical model
- NEW: Amortised variational inference
- NEW: Experimental results

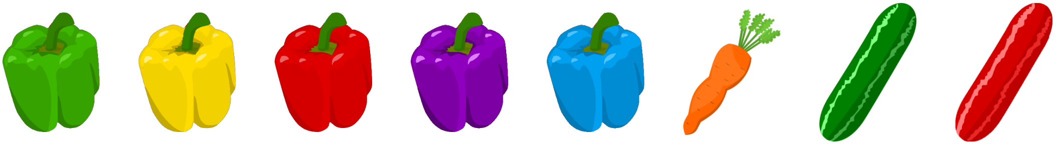
Truth-Conditional Semantics



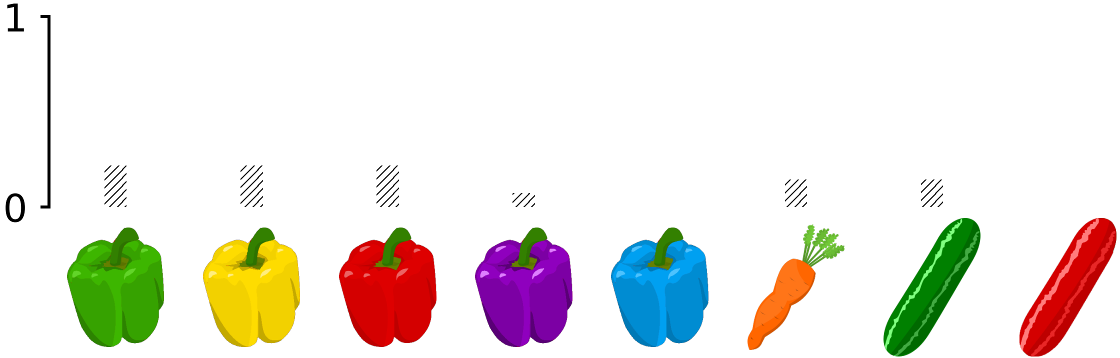
Truth-Conditional Semantics



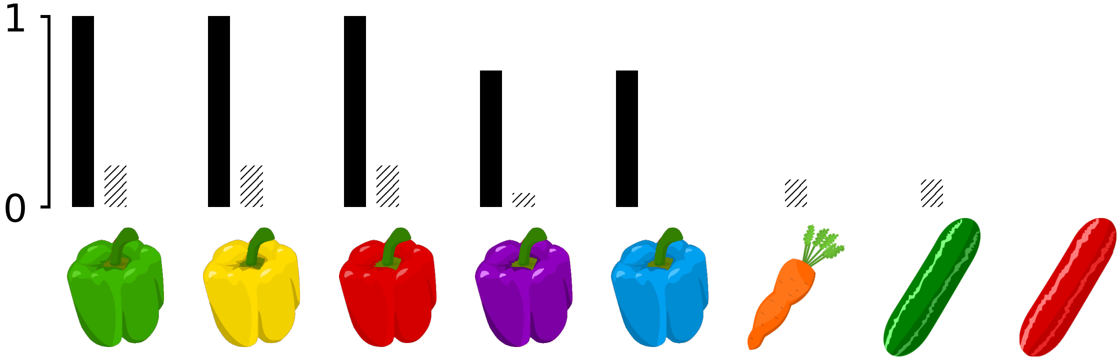
Truth-Conditional Functions



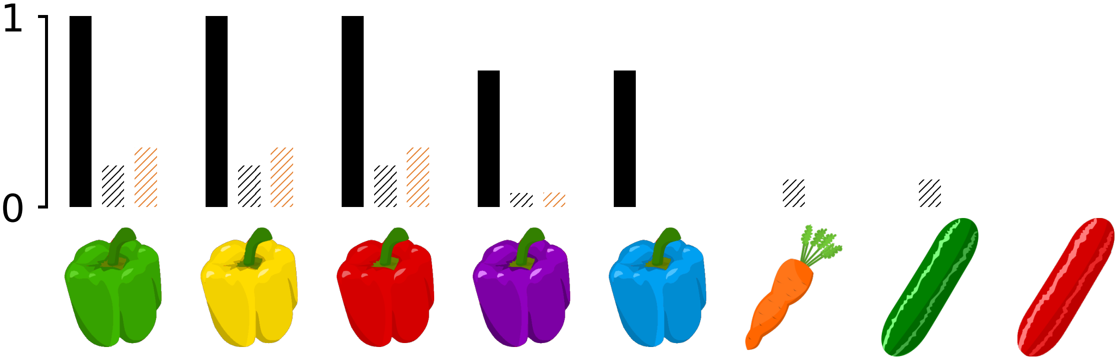
Truth-Conditional Functions



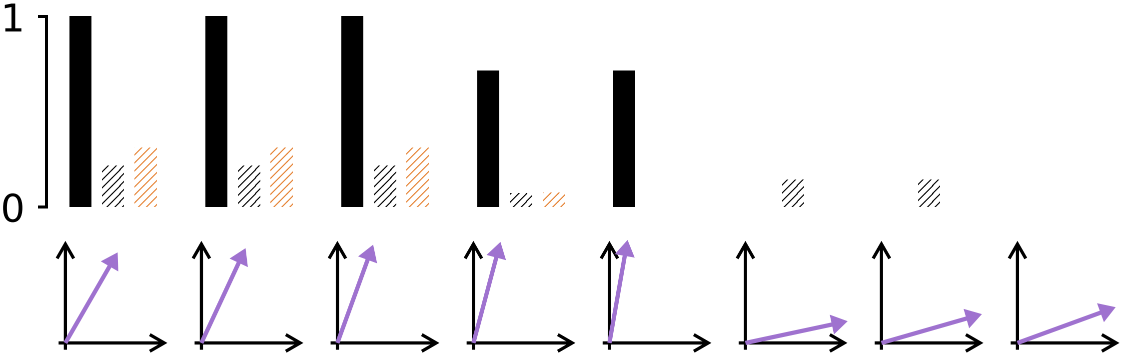
Truth-Conditional Functions



Truth-Conditional Functions



Truth-Conditional Functions



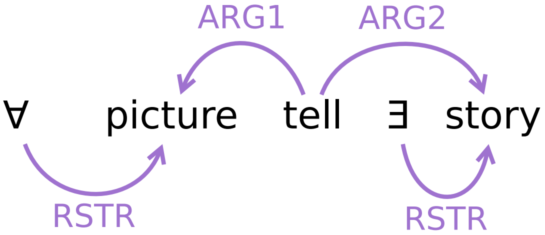
Summary So Far

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

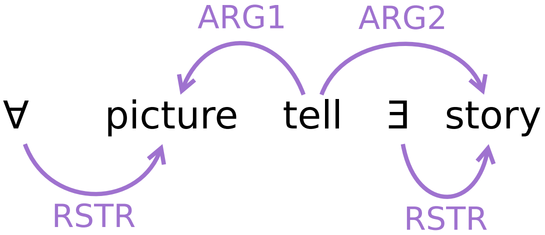
Sentences as Graphs (DMRS)

Every picture tells a story

Sentences as Graphs (DMRS)

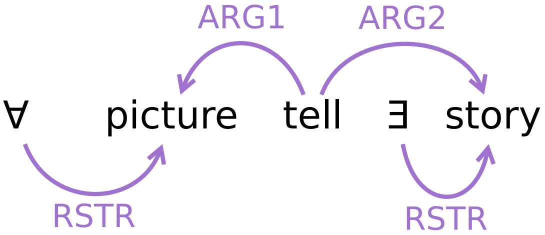


Sentences as Graphs (DMRS)



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

Sentences as Graphs (DMRS)



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

- 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

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

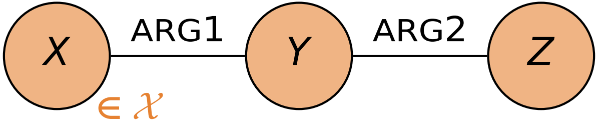
dog(x) chase(y) cat(z)

Functional Distributional Semantics

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

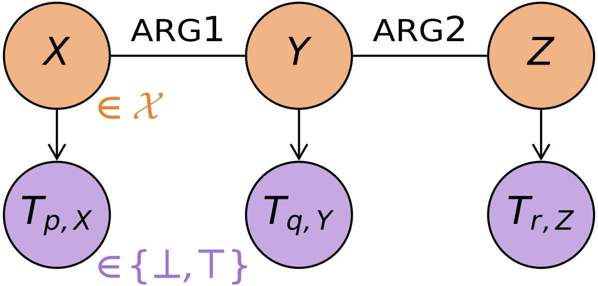
$$p(x) \qquad\qquad q(y) \qquad\qquad r(z)$$

Functional Distributional Semantics

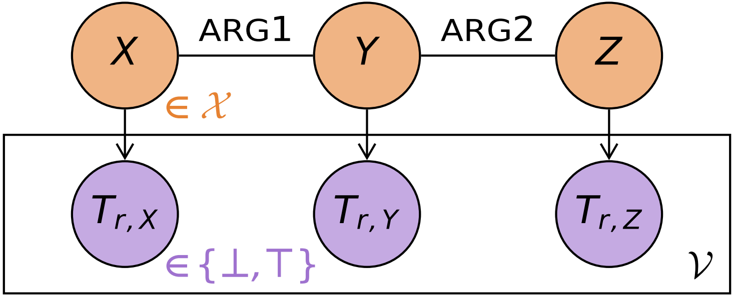


$p(X)$ $q(Y)$ $r(Z)$

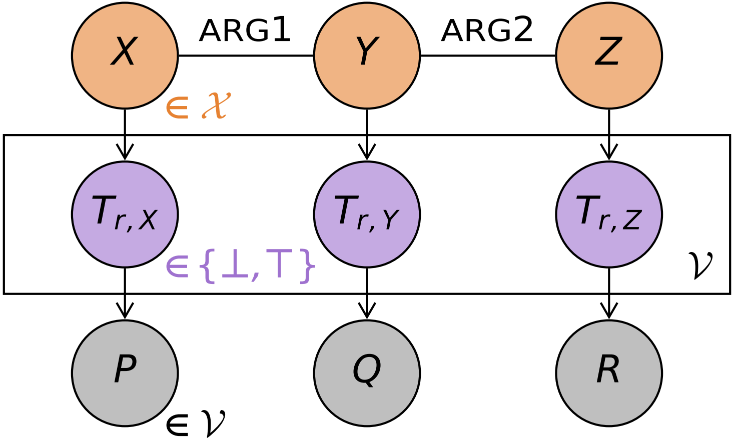
Functional Distributional Semantics



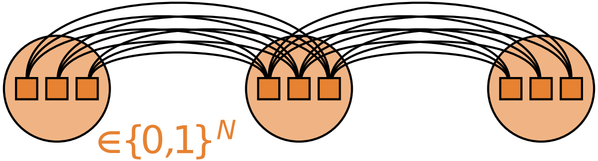
Functional Distributional Semantics



Functional Distributional Semantics



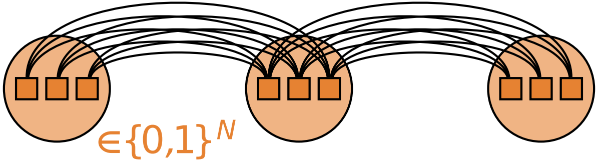
World Model



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

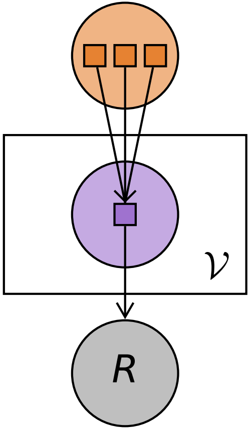
- $\mathbb{P}(s) \propto \exp\left(-E(s)\right)$

World Model



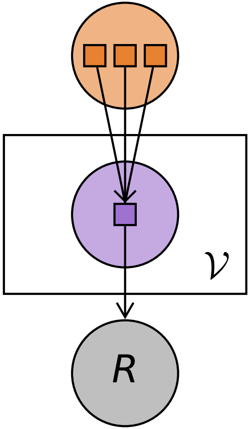
- Cardinality Restricted Boltzmann Machine (CaRBM; Swersky et al., 2012)
- $\mathbb{P}(s) \propto \exp\left(\sum_{x \xrightarrow{L} y \text{ in } s} w_{ij}^{(L)} x_i y_j \right)$

Lexical Model



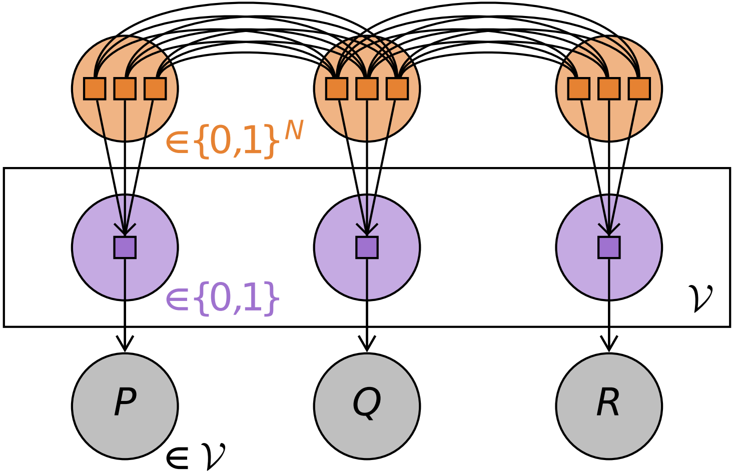
- Feedforward networks
- $t^{(r)}(x) = \sigma\left(v_i^{(r)} x_i\right)$

Lexical Model



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

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

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

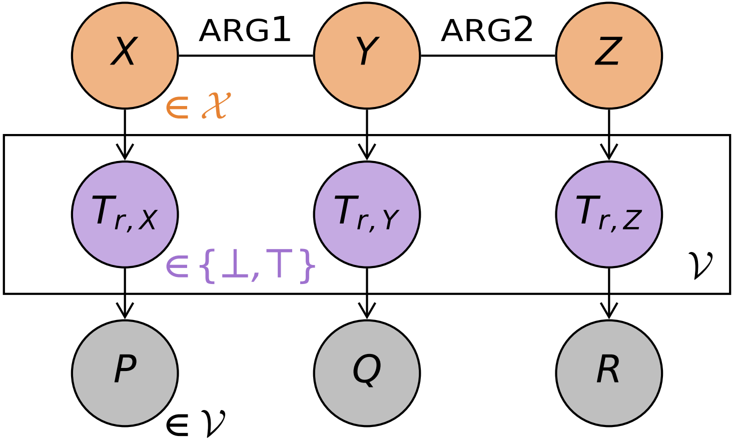
- Latent variables necessary but inconvenient

Gradient Descent

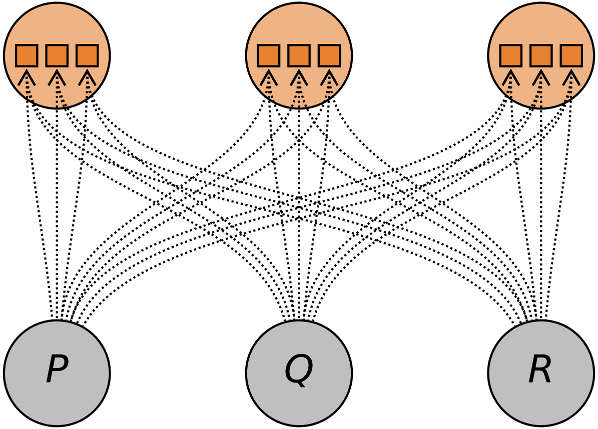
$$\frac{\partial}{\partial \theta} \log \mathbb{P}(g) = \left(\mathbb{E}_{s|g} - \mathbb{E}_{s} \right) \left[\frac{\partial}{\partial \theta} \left(-E(s) \right) \right] + \mathbb{E}_{s|g} \left[\frac{\partial}{\partial \theta} \log \mathbb{P}(g \mid s) \right]$$

- 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

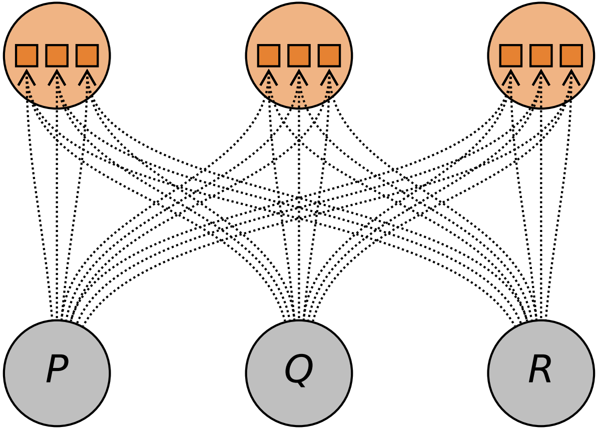
- Variational distribution must be optimised *for each input graph*

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

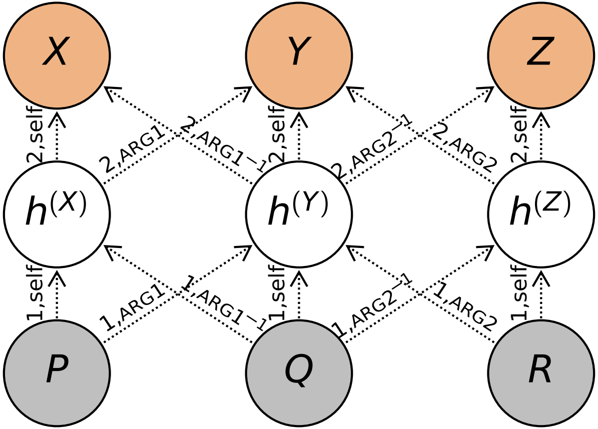
Amortised Variational Inference

- 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 $\mathbb{E}_s$…

Pixie Autoencoder

- Generative model & inference network

Pixie Autoencoder

- Generative model & inference network
- NLP interest:
  - Truth-conditional distributional semantics

Pixie Autoencoder

- Generative model & inference network
- NLP interest:
  - Truth-conditional distributional semantics
- General ML interest:
  - Efficient inference for latent variables

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)

Similarity in Context (GS2011)

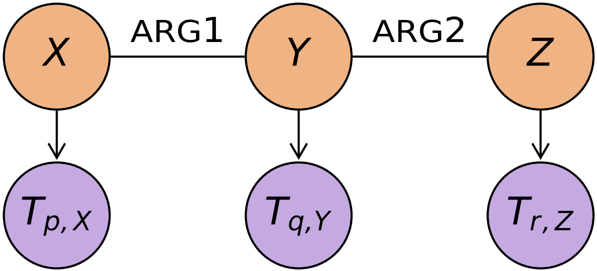
student	write	name
student	spell	name
scholar	write	book
scholar	spell	book

BERT for GS2011

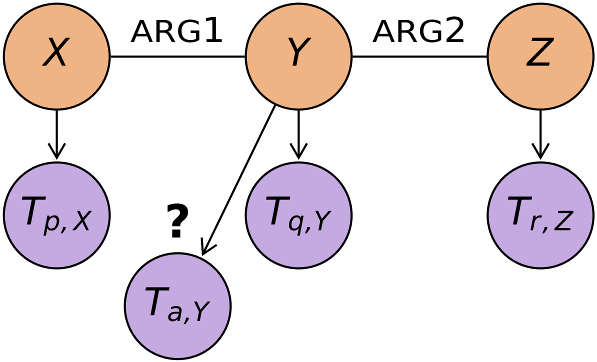
Pseudo-logical form: (employer provide training)

- “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 .”
- ...

Pixie Autoencoder for GS2011



Pixie Autoencoder for GS2011



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

GS2011 Results

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

GS2011 Results

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

- Smaller model, less data, better performance

GS2011 Results

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

- Smaller model, less data, better performance
- More results in the paper!

Summary

- Meanings: functions
- Sentences: graphs
- Inference: graph convolutions
- Logic: useful

Linguists who use Probabilistic Models Love Them

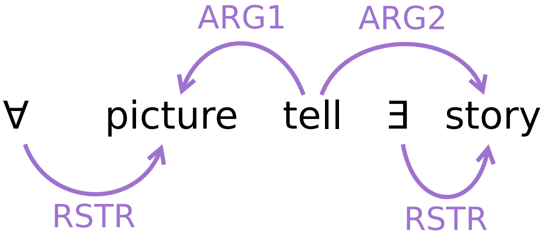
Quantification in Functional Distributional Semantics

Guy Emerson

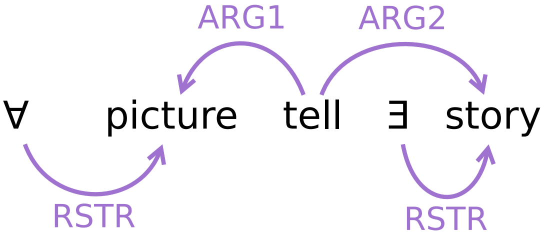
Sentences as Graphs (DMRS)

Every picture tells a story

Sentences as Graphs (DMRS)



Sentences as Graphs (DMRS)



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

Overview

- Probabilistic quantification
- Generic quantification
- Bonus: donkey anaphora

Generalised Quantifier Theory

- A quantifier has a *restriction* $\mathcal{R}$ and *body* $\mathcal{B}$

Generalised Quantifier Theory

- A quantifier has a *restriction* $\mathcal{R}$ and *body* $\mathcal{B}$

- For example:
  - *Some dog barked.*
  - *Every dog barked.*
  - *No dog barked.*
  - *Most dog barked.*

Generalised Quantifier Theory

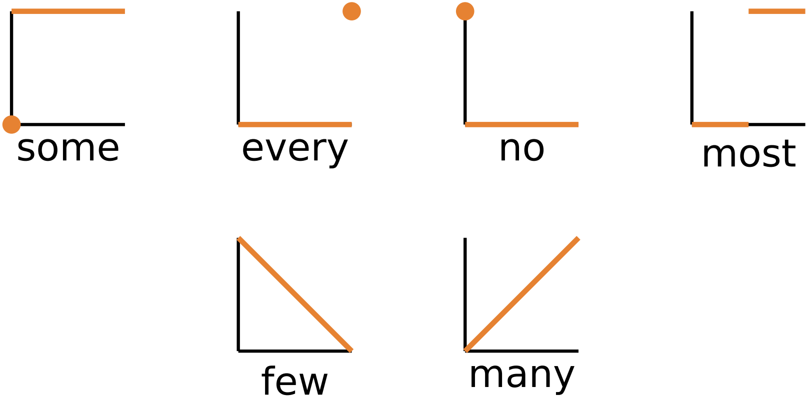
- A quantifier has a *restriction* $\mathcal{R}$ and *body* $\mathcal{B}$
- Truth defined in terms of sizes of sets:
  - *Some*: $|\mathcal{R} \cap \mathcal{B}| > 1$
  - *Every*: $|\mathcal{R} \cap \mathcal{B}| = |\mathcal{R}|$
  - *No*: $|\mathcal{R} \cap \mathcal{B}| = 0$
  - *Most*: $|\mathcal{R} \cap \mathcal{B}| > \frac{1}{2}|\mathcal{R}|$

Probabilistic Quantifiers

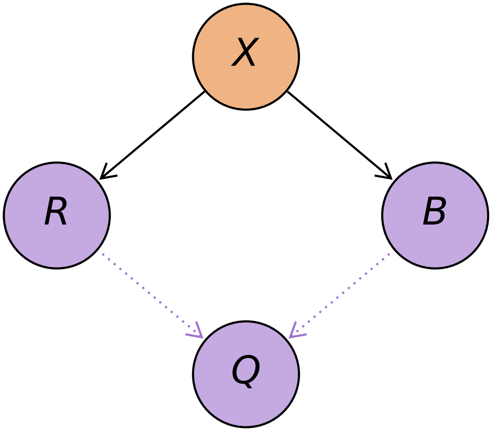
- $\mathbb{P}(B \mid R) = \dfrac{\mathbb{P}(R, B)}{\mathbb{P}(R)}$

- Truth defined in terms of probabilities:
  - *Some*: $\mathbb{P}(B \mid R) > 0$
  - *Every*: $\mathbb{P}(B \mid R) = 1$
  - *No*: $\mathbb{P}(B \mid R) = 0$
  - *Most*: $\mathbb{P}(B \mid R) > \frac{1}{2}$

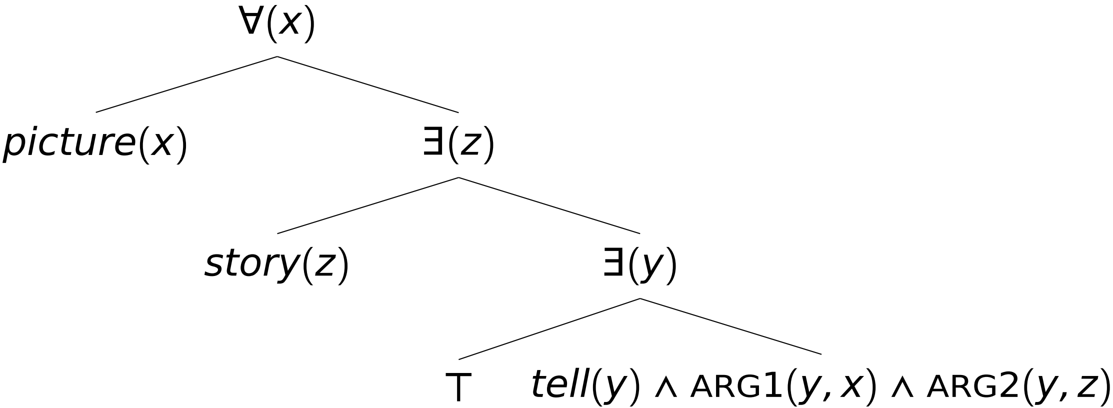
Probabilistic Quantifiers



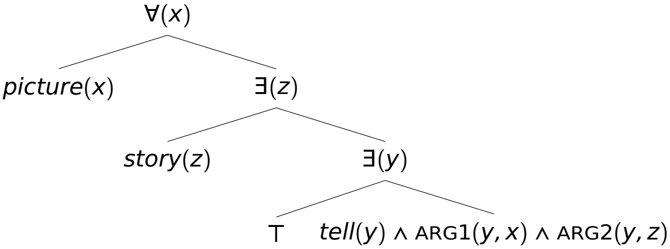
Probabilistic Quantifiers

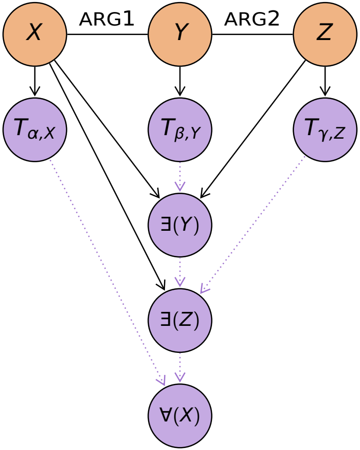


Scope Trees


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

Probabilistic Scope Trees





Generics

- Dogs bark
- Ducks lay eggs
- Mosquitoes carry malaria

Generic Puzzle

- Generics vs. classical quantifiers:
  - Harder to define mathematically
  - Easier for children to acquire

Generic Puzzle

- Generics vs. classical quantifiers:
  - Harder to define mathematically
  - Easier for children to acquire
- Proposal: computationally simpler

Rational Speech Acts

- Communication as a cooperative game:
  - Speaker knows something; listener does not
  - Speaker chooses to say something
  - Listener must infer what the speaker knows

Rational Speech Acts

- Communication as a cooperative game:
  - Speaker knows something; listener does not
  - Speaker chooses to say something
  - Listener must infer what the speaker knows
  - Inference as Bayesian inference

Rational Speech Acts

- Communication as a cooperative game:
  - Literal listener: infer based on truth

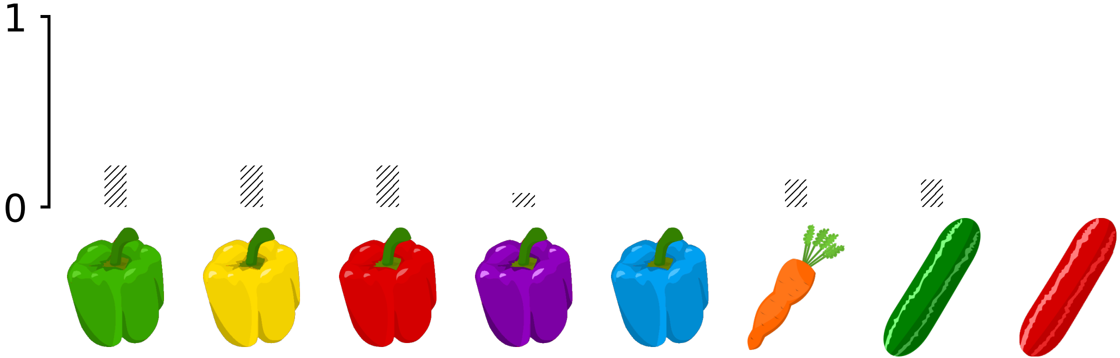
Rational Speech Acts

- Communication as a cooperative game:
  - Literal listener: infer based on truth
  - Pragmatic speaker: optimise choice for literal listener

Rational Speech Acts

- Communication as a cooperative game:
  - Literal listener: infer based on truth
  - Pragmatic speaker: optimise choice for literal listener
  - Pragmatic listener: infer based on pragmatic speaker

Rational Speech Acts



RSA for Generics (Tessler and Goodman, 2019)

- Semantically simple
  - Increasing ratio, increasing probability
  - $\mathbb{P}(Q) = \mathbb{P}(B \mid R)$

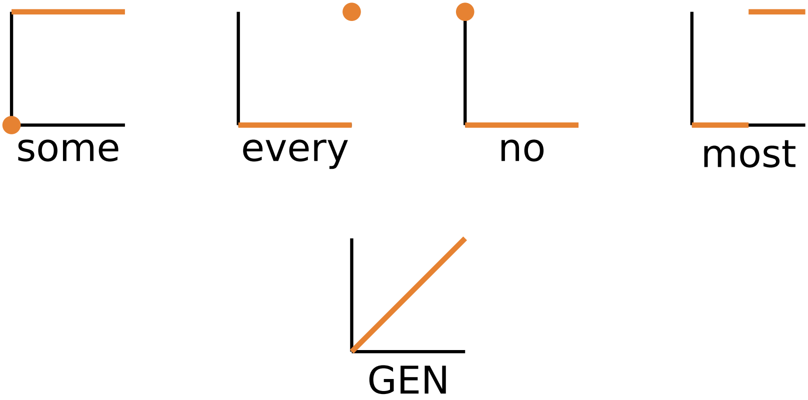
RSA for Generics (Tessler and Goodman, 2019)

- Semantically simple
  - Increasing ratio, increasing probability
  - $\mathbb{P}(Q) = \mathbb{P}(B \mid R)$
- Pragmatically dependent on prior knowledge
  - Dogs bark
  - Ducks lay eggs
  - Mosquitoes carry malaria

Generic Puzzle

- Generics vs. classical quantifiers:
  - Harder to define mathematically
  - Easier for children to acquire
- Proposal: computationally simpler

Probabilistic Quantifiers



Computational Cost of Quantification

- Classical quantifiers are sensitive to probabilities being exactly 0 or 1

Computational Cost of Quantification

- Classical quantifiers are sensitive to probabilities being exactly 0 or 1
  - A vague predicate has to be seen as a distribution over precise predicates
  - Summing over this distribution is expensive

Computational Cost of Quantification

- Classical quantifiers are sensitive to probabilities being exactly 0 or 1
  - A vague predicate has to be seen as a distribution over precise predicates
  - Summing over this distribution is expensive
- GEN doesn't need precise predicates

Computational Cost of Quantification

- Classical quantifiers are sensitive to probabilities being exactly 0 or 1
  - A vague predicate has to be seen as a distribution over precise predicates
  - Summing over this distribution is expensive
- GEN doesn’t need precise predicates
  - GEN can be lazy! Easier to compute!

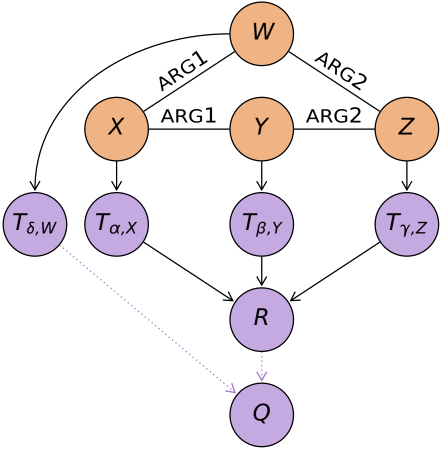
Bonus: Donkey Anaphora

- Every farmer who owns a donkey feeds it

Bonus: Donkey Anaphora

- Every farmer who owns a donkey feeds it
- Farmers who own donkeys feed them
- Linguists who use probabilistic models love them
- Mosquitoes which bite birds infect them with malaria

Bonus: Donkey Anaphora



Summary

- Quantification: conditional probability
- Generics: lazy probabilistic quantification
- Donkey anaphora: generic quantification

Classical Donkeys