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 \textit{functions}
- Logically interpretable model
- Outperforms BERT at semantics
- Clear path for multimodal learning
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

Important for discourse: anaphora resolution, question answering, dialogue processing...

Meaning as a function over entities
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 → probability of truth
Summary of What’s New

- Pixie: entity representation
- Word meanings as functions: pixie $\rightarrow$ 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

\[ x \]

\[ \text{pepper}(x) \]
Situation Semantics

x

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

\[ x \leftarrow \text{ARG1} \quad y \quad \text{ARG2} \rightarrow z \]

dog(x)  \hspace{1cm} \text{animal}(x)  \hspace{1cm} \text{chase}(y)  \hspace{1cm} \text{pursue}(y)  \hspace{1cm} \text{cat}(z)  \hspace{1cm} \text{animal}(z)
Probabilistic Situation Semantics

$\in \mathcal{X}$

- dog($X$)
- animal($X$)
- chase($X$)
- pursue($X$)
- cat($X$)

- chase($Y$)
- dog($Y$)
- cat($Y$)

- cat($Z$)
- animal($Z$)
- chase($Z$)
- pursue($Z$)
- dog($Z$)
Probabilistic Situation Semantics

\[ X \in \mathcal{X} \]

\[ \mathcal{X} \ni \mathcal{R} \]

\[ \mathcal{Y} \ni \mathcal{R} \]

\[ \mathcal{Z} \ni \mathcal{R} \]

\[ \mathcal{T}_{r,X} \]

\[ \mathcal{T}_{r,Y} \]

\[ \mathcal{T}_{r,Z} \]

\[ \mathcal{V} \ni \mathcal{R} \]

\[ \mathcal{V} \ni \mathcal{R} \]

\[ \mathcal{V} \ni \mathcal{R} \]
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)
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”
Image preprocessing: pixies given by pre-trained ResNet101

World model: $P(x, y, z)$ Gaussian

Lexical model: $P(t_{r,x} | x)$ one-layer sigmoid
Logical Reasoning with Latent Entities

\[ P(t_{\text{horse}}, X | t_{\text{animal}}, X, t_{\text{has}}, Y, t_{\text{tail}}, Z) \]
Logical Reasoning with Latent Entities

\[ \mathbb{P}(t_{a,X} | t_{p,X}, t_{q,Y}, t_{r,Z}) \]
Logical Reasoning with Latent Entities

\[ \mathbb{P}(t_{a,X} \mid t_{p,X}, t_{q,Y}, t_{r,Z}) \]

\[ \mathbb{P}(t_{\text{horse},X} \mid t_{\text{animal},X}, t_{\text{has},Y}, t_{\text{tail},Z}) \]
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
Every picture tells a story
Dependency Minimal Recursion Semantics

∀ picture tell ∃ story

\[\forall x \exists y \exists z \text{ picture}(x) \Rightarrow [\text{story}(z) \land \text{tell}(y) \land \text{ARG1}(y, x) \land \text{ARG2}(y, z)]\]
∀x∃y∃z picture(x) ⇒ [story(z) ∧ tell(y) ∧ ARG1(y, x) ∧ ARG2(y, z)]

Functional Distributional Semantics

$$x \in X$$

$$T_{r,X} \in \{\bot, T\}$$

$$T_{r,Y} \in \{\bot, T\}$$

$$T_{r,Z} \in \{\bot, T\}$$

Dog $\xleftarrow{\text{ARG1}}$ Chase $\xrightarrow{\text{ARG2}}$ Cat
Functional Distributional Semantics

\[ X \in \mathcal{X} \]

\[ T_{r, X} \]

\[ T_{r, Y} \]

\[ T_{r, Z} \]

\[ R_X \]

\[ R_Y \]

\[ R_Z \]

\[ \in \mathcal{V} \]

\[ \in \{ \bot, T \} \]
Latent situation semantics
- World model: $\mathbb{P}(x, y, z)$
- Lexical model: $\mathbb{P}(t_{r,x} | x)$

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

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

\[ w(L) \in \{0,1\}^N \]
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)

\[ P(s) \propto \exp(-E(s)) \]
World Model

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

- \( \mathbb{P}(s) \propto \exp \left( \sum_{x \rightarrow y \text{ in } s} w_{ij}^{(L)} x_i y_j \right) \)
Lexical Model

- Feedforward networks

\[ P(t^{(r,X)} | x) = \sigma(v_i^{(r)} x_i) \]
Lexical Model

- Feedforward networks
- \( \mathbb{P}(t^{(r,X)} | x) = \sigma(v_i x_i) \)
- \( \mathbb{P}(r^{(X)} | x) \propto \mathbb{P}(t^{(r,X)} | x) \)
Functional Distributional Semantics
Gradient Descent

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

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

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

- Latent variables necessary but inconvenient
\[
\frac{\partial}{\partial \theta} \log p(g) = \left( \mathbb{E}_{s|g} - \mathbb{E}_s \right) \left[ \frac{\partial}{\partial \theta} (-E(s)) \right] \\
+ \mathbb{E}_{s|g} \left[ \frac{\partial}{\partial \theta} \log p(g|s) \right]
\]

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

\[ X \in \mathcal{X} \]
\[ Y \in \{\bot, T\} \]
\[ T_{r,X} \]
\[ T_{r,Y} \]
\[ T_{r,Z} \]
\[ R_X \]
\[ R_Y \]
\[ R_Z \]
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

\[ h^{(X)} \xrightarrow{1,\text{self}} h^{(Y)} \xrightarrow{2,\text{ARG2}} h^{(Z)} \]

\[ h^{(Y)} \xrightarrow{2,\text{ARG1}} h^{(X)} \xrightarrow{1,\text{self}} h^{(Z)} \]

\[ h^{(Z)} \xrightarrow{2,\text{ARG2}} h^{(Y)} \xrightarrow{1,\text{self}} h^{(X)} \]
Amortised Variational Inference

\[ \frac{\partial}{\partial \phi} D(Q|\mathcal{P}) = - \frac{\partial}{\partial \phi} \mathbb{E}_{Q(s)} \left[ \log \mathbb{P}(s) \right] \\
- \frac{\partial}{\partial \phi} \mathbb{E}_{Q(s)} \left[ \log \mathbb{P}(g|s) \right] \\
- \frac{\partial}{\partial \phi} H(Q) \]
Gradient Descent

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

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

- Latent variables: amortised variational inference
Gradient Descent

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

- 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

\[
\frac{\partial}{\partial \theta} \log P(g) = \left( E_{s|g} - E_s \right) \left[ \frac{\partial}{\partial \theta} (-E(s)) \right] \\
+ E_{s|g} \left[ \frac{\partial}{\partial \theta} \log P(g | s) \right]
\]

- 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)
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)
• 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

X \rightarrow ARG1 \rightarrow Y \rightarrow ARG2 \rightarrow Z

Tp,X \rightarrow Tq,Y \rightarrow Tr,Z
Pixie Autoencoder for GS2011

$P(t_{a,Y} \mid t_{p,X}, t_{q,Y}, t_{r,Z})$
Pixie Autoencoder for GS2011

\[ P(t_{\text{spell}, Y} \mid t_{\text{student}, X}, t_{\text{write}, Y}, t_{\text{name}, Z}) \]
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

<table>
<thead>
<tr>
<th>Model</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skip-gram (vector addition)</td>
<td>.348</td>
</tr>
<tr>
<td>BERT (with tuned template strings)</td>
<td>.446</td>
</tr>
<tr>
<td>Pixie Autoencoder</td>
<td>.504</td>
</tr>
</tbody>
</table>

- 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%)
telescope    device that astronomers use
saw         device that cuts wood
philosopher  person that defends rationalism
survivor    person that helicopter saves
farming     activity that soil supports
...         ...
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
...
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
...

<table>
<thead>
<tr>
<th>soil</th>
</tr>
</thead>
<tbody>
<tr>
<td>device that astronomers use</td>
</tr>
<tr>
<td>device that detects planets</td>
</tr>
<tr>
<td>device that cuts wood</td>
</tr>
<tr>
<td>person that defends rationalism</td>
</tr>
<tr>
<td>person that helicopter saves</td>
</tr>
<tr>
<td>activity that soil supports</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
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

\[ X \xrightarrow{\text{ARG1}} T_{p,X} \xrightarrow{\text{ARG2}} Y \xrightarrow{T_{q,Y}} Z \xrightarrow{T_{r,Z}} \]

- \( T_{p,X} \): Title of philosopher, \( X \)
- \( T_{q,Y} \): Title of rationalism, \( Y \)
- \( T_{r,Z} \): Title of rationalism, \( Z \)
Logical Inference for RELPRON

\[ P(t_{a,X} \mid t_{p,X}, t_{q,Y}, t_{r,Z}) \]
Logical Inference for RELPRON

\[
P(t_{a,X} | t_{p,X}, t_{q,Y}, t_{r,Z})
\]

\[
P(t_{philosopher,X} | t_{person,X}, t_{defend,Y}, t_{rationalism,Z})
\]
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.”
- ...

BERT for RELPRON
## RELPRON Results

<table>
<thead>
<tr>
<th>Model</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simp. Prac. Lex. Func. (Rimell et al., 2016)</td>
<td>.497</td>
</tr>
<tr>
<td>Dependency vectors (Czarnowska et al., 2019)</td>
<td>.439</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>.474</td>
</tr>
<tr>
<td>BERT (with carefully tuned template strings)</td>
<td>.186</td>
</tr>
<tr>
<td>BERT &amp; Word2Vec ensemble</td>
<td>.479</td>
</tr>
<tr>
<td>Pixie Autoencoder</td>
<td>.189</td>
</tr>
<tr>
<td>Pixie Autoencoder &amp; Word2Vec ensemble</td>
<td>.489</td>
</tr>
</tbody>
</table>
## RELPRON Results

<table>
<thead>
<tr>
<th>Model</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simp. Prac. Lex. Func. (Rimell et al., 2016)</td>
<td>.497</td>
</tr>
<tr>
<td>Dependency vectors (Czarnowska et al., 2019)</td>
<td>.439</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>.474</td>
</tr>
<tr>
<td>BERT (with carefully tuned template strings)</td>
<td>.186</td>
</tr>
<tr>
<td>BERT &amp; Word2Vec ensemble</td>
<td>.479</td>
</tr>
<tr>
<td>Pixie Autoencoder</td>
<td>.189</td>
</tr>
<tr>
<td>Pixie Autoencoder &amp; Word2Vec ensemble</td>
<td>.489</td>
</tr>
</tbody>
</table>
## RELPRON Results

<table>
<thead>
<tr>
<th>Model</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simp. Prac. Lex. Func. (Rimell et al., 2016)</td>
<td>.497</td>
</tr>
<tr>
<td>Dependency vectors (Czarnowska et al., 2019)</td>
<td>.439</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>.474</td>
</tr>
<tr>
<td>BERT (with carefully tuned template strings)</td>
<td>.186</td>
</tr>
<tr>
<td>BERT &amp; Word2Vec ensemble</td>
<td>.479</td>
</tr>
<tr>
<td>Pixie Autoencoder</td>
<td>.189</td>
</tr>
<tr>
<td>Pixie Autoencoder &amp; Word2Vec ensemble</td>
<td>.489</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Model</th>
<th>MEN</th>
<th>SL999</th>
<th>GS2011</th>
<th>RELPRON</th>
</tr>
</thead>
<tbody>
<tr>
<td>VG-count (Herbelot, 2020)</td>
<td>.336</td>
<td>.224</td>
<td>.063</td>
<td>.038</td>
</tr>
<tr>
<td>VG-retrieval</td>
<td>.420</td>
<td>.190</td>
<td>.072</td>
<td>.045</td>
</tr>
<tr>
<td>EVA (Herbelot, 2020)</td>
<td>.543</td>
<td>.390</td>
<td>.068</td>
<td>.032</td>
</tr>
<tr>
<td>Functional</td>
<td>.639</td>
<td>.431</td>
<td>.171</td>
<td>.117</td>
</tr>
</tbody>
</table>

Truth-conditional structure helps generalisation.
## Visual Genome Semantics

<table>
<thead>
<tr>
<th>Model</th>
<th>MEN</th>
<th>SL999</th>
<th>GS2011</th>
<th>RELPRON</th>
</tr>
</thead>
<tbody>
<tr>
<td>VG-count (Herbelot, 2020)</td>
<td>.336</td>
<td>.224</td>
<td>.063</td>
<td>.038</td>
</tr>
<tr>
<td>VG-retrieval</td>
<td>.420</td>
<td>.190</td>
<td>.072</td>
<td>.045</td>
</tr>
<tr>
<td>EVA (Herbelot, 2020)</td>
<td>.543</td>
<td>.390</td>
<td>.068</td>
<td>.032</td>
</tr>
<tr>
<td>Functional</td>
<td>.639</td>
<td>.431</td>
<td>.171</td>
<td>.117</td>
</tr>
</tbody>
</table>

- Truth-conditional structure helps generalisation
Classification accuracy per predicate
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 $\rightarrow$ pixies are latent
  - Grounded data $\rightarrow$ 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...
Meanings: functions

Entities: (latent or observed) pixies

Probabilistic logic: empirically useful