Concept learning as modeled via modern knowledge-base models

Zhiwei Li and Kelsey Moty

Categories are an important tool that allow us to make broad inferences about the world based on limited evidence. For example, they lead us to make generalizations from an individual (e.g., this individual dog barks) to the broader category (e.g., dogs bark)—and, in turn, we use them to make inferences about a novel individual exemplar based on its category membership (e.g., this poodle barks because it's a kind of dog). One question of interdisciplinary interest: how should we characterize the process of conceptual learning, both in humans and in machines? The present paper explores two computational approaches—the classic parallel distributed processing model and modern embedding models—as viable models of knowledge acquisition.

How do children learn about the world? A brief summary of the importance of concepts.

Children—and even infants (Graham, Kilbreath, & Welder, 2004)—use category membership as a basis for generalization, extending information from one individual to another of the same category, even when the they differ perceptually (Gelman & Markman, 1986). This process of category-based induction requires understanding that categories have fundamental similarities between members (e.g., all snakes don't have legs)—that is, one most hold an assumption of category homogeneity—while also acknowledging that they have important variation (e.g., only some snakes are poisonous; Gelman, 2003; Murphy, 2002). While categories tend to have a coherent and, at times, homogeneous structure, neither the similarity of its members nor the frequency of shared features are sufficient for learning concepts. Rather, concepts are tightly linked with our intuitive theories of the world, and we more readily learn concepts consistent with those theories (Keil, 1989; Murphy & Medin, 1985). Specifically, our theories play an important role in learning by imposing order on observed evidence and guiding our interpretations of it (see Carey, 1985; 2009).

A significant body of work demonstrates that children and infants are sensitive to patterns in their environment (e.g., Saffran, Aslin, & Newport, 1996, see Aslin & Newport, 2012 for review)—and can use these patterns to generate rules and update theories (Gopnik et al., 2001, Kushnir, Wellman, & Gelman, 2009,). Importantly, knowledge is *domain-specific*. People hold distinct theories for various domains (e.g., an intuitive theory of biology, Keil, 1994; of artifacts, Keil, Greig, & Kermer, 2007; of psychology, Wellman, 1990) that ascribe different causal mechanisms (e.g., natural selection applies to biological kinds but not artificial kinds) and appeal to different unobservable entities (e.g., DNA, beliefs). We use these theories to guide the inferences we make about novel kinds. Young children have a bias to view categories as largely homogeneous (e.g., Cimpian & Park, 2013; Moty & Brandone, submitted), resulting in the *overgeneralization* of properties (e.g., assuming all birds). But young kids also show some sensitivity to factors limiting generalizability, including category domain and property type (Brandone & Gelman, 2009; 2013)—and their appreciation for variability within categories increases with age (Gelman, 1988).

Parallel distributed process model account of knowledge acquisition.

In contrast to the theory-driven approach to conceptual development and knowledge acquisition described above, Rogers and McClelland (2004) argue for a computational account that does not posit domain-specific theories nor assumes any prior or innate knowledge on the part of the learner. Their parallel distributed processing (PDP) model (along with other models

within the same framework, e.g., Mareschal, French, & Quinn, 2000; Mayor & Plunkett, 2010) demonstrated that many of the characteristic behaviors of knowledge acquisition (e.g., overgeneralization, domain-specific category-based induction) can unfold naturally as a consequence of domain-general learning mechanisms.

Aim of the present project.

The Rogers and McClelland's (2004) PDP model accounts for a number of behaviors we see in during knowledge acquisition and early conceptual learning, including phenomenon like overgeneralization (e.g., children's initial assumption that penguins can fly). Yet this model is trained on a small amount of data (~20 entities) compared to modern knowledge bases with significantly larger datasets. It is unclear whether the Roger & McClelland model would scale to larger datasets. On the other hand, modern knowledge-base researchers (e.g., Bordes et al., 2013; Yang et al., 2014) have developed a variety of parallel processing models that can be successfully trained on large knowledge bases (e.g., WordNet with over 40K entities; Freebase 15K with 15K entities). Yet, they usually focus on the final performance of triplet completion without examining the learning process that led to the final result.

In this project, we re-implement the Rogers and McClelland model with some modification. We train it with both the original dataset used in Rogers and McClelland (2004) as well as a larger knowledge base dataset. Additionally, we implement a modern embedding model (Yang et al., 2014) and use similar methods employed in Rogers and McClelland (2004) to analyze the learning process of this model and draw parallel with human behaviors. We also develop new tasks like discovering new triplets and surprising facts for the modern embedding model to explore its "cognitive" capability. Finally, we explore whether incorporating additional human cognitive component—logical deduction—would improve the learning of knowledge base.

Method

Data

Representing knowledge as triplets.

We represent pieces of knowledge as triplets with <head - relation - tail>.

The datasets. The data used in this paper are derived from 2 sources:

1. *Simple species dataset.* We first extracted a data table of 8 species, with 36 relations each. This table is extracted from Appendix B, table B-1 in Rogers and McClelland (2004).
2. *Canadian mammals dataset.* The following dataset was scraped from the website for the FactGuru knowledge base about Canadian mammals (found here: http://www.site.uottawa.ca/~tcl/factguru1/animals/index.html). The dataset contains a hierarchy of 364 unique animal kinds. At the top, the category *mammal* contains all subordinate categories—and at the bottom, the hierarchy consists of categories as narrow as *adult*/*baby* (or *male/female*) groups within a species or subordinate categories of a species (e.g., *short-tailed shrew* or *pigmy shrew* as subordinates of *shrew*). When possible, continuous relations were translated to dichotomous categorical variables (e.g., "has gestation period of 8 months" converted to "has gestation period"); other continuous relations (e.g., "has weight of 8 kg") were dropped from the dataset. Relations that were

true of only a single animal kind were removed from the dataset. After the cleaning process, the final dataset had 2040 relations.

Negative sample generation. For each head-relation pair (h-r), we generate negative samples by appending false tails t_f that: (1) < h-r-t_f> is not in the dataset and (2) < h'-r-t_f> is in the dataset (in other words, r-t_f is a valid relation-tail pair). The ratio between negative and positive pairs is reportedly influential on training performance (Trouillon et al., 2016). Here we tentatively set it to at most 3 negative samples per correct triplet (some triplets have less because not enough legitimate t_f). This will be used in the training of embedding model.

Re-implementing the classic PDP model.

Network and parameters. We implemented the Rumelhart semantic memory model as specified in Figure 2.2 of Rogers and McClelland (2004). Specifically, head, relation and tails are represented in a one-hot manner: the head is connected to a representation layer consisting of the same number of units as the head layer, and then both the representation layer and the relation connect to a hidden layer with 15 units, which feedforward to the output tail layer. Each layer's output is rectified with ReLU function (instead of softmax as in the original model), but we didn't test whether this is a significant modification. Finally, the output layer had a softmax function to constrain the output between 0 and 1.

Training. We used the *pytorch* package to construct this model. RMSprop was used for optimizing the network. 300 episodes in total are included in the training data.

Modern embedding model.

Network and parameters. First appearing with the transE model developed by Bordes et al. (2013), one recent popular trend for knowledge base models is the *embedding model*. In these kinds of models, the head, relation and tail are transformed into an embedding vector space, and the knowledge triplets correspond to a set of linear algebra operations. Yang et al. (2014) summarized a number of possible vector and operation choices, and in this paper, we chose to explore the specific form referred to as "DistMulti model" in Yang et al. (2014). In this model, entities (i.e., both head and tails) are represented as a vector while each relation is represented as a diagonal matrix. A triplet is represented as a bilinear transformation: $Score(h\text{-}r\text{-}t) = \boldsymbol{v}h\ \boldsymbol{B}r\ \boldsymbol{v}t$. Correct triplets are assigned a value of 1, while false triplets are assigned a value of -1. In our model, we tested embeddings with different numbers of dimensions (between 10 and 20 dimensions) and ultimately used an embedding of 15 dimensions for both datasets given that it provided the best performance.

Training. We used the *pytorch* package to construct this model, and the Adam optimization algorithm to optimize the network. 1500 episodes are included in the training data.

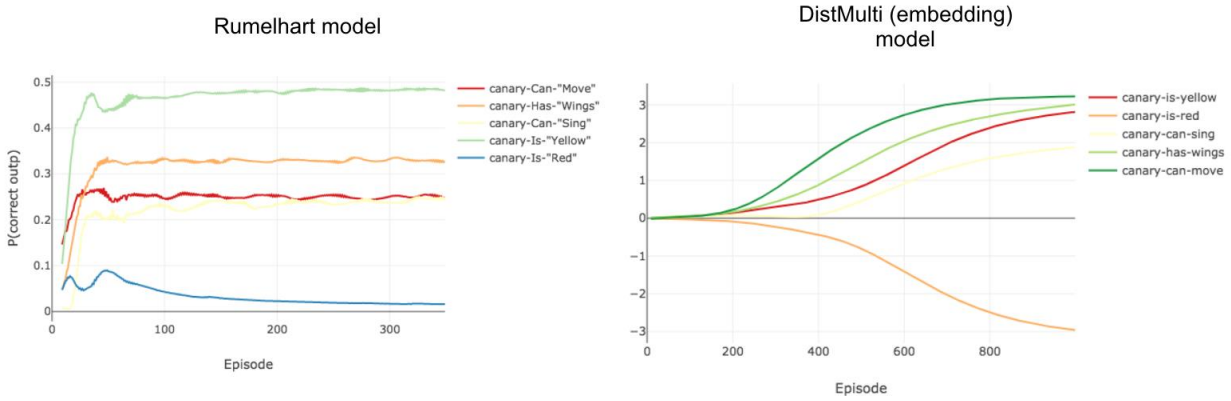
Results

Simple species dataset (the smaller dataset).

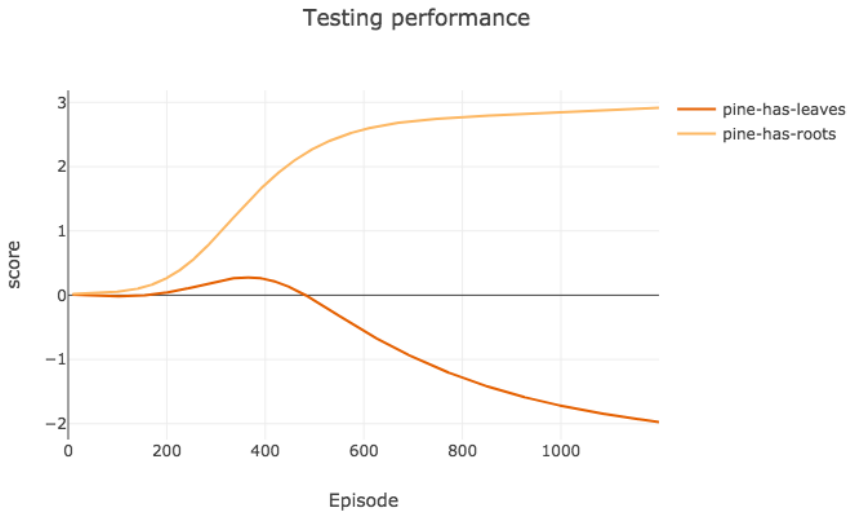
With this dataset, we first replicate the results from the classic PDP model, then observe whether the embedding model shows similar behaviors. Moreover, we introduced an inference task that has rarely been studied in the current knowledge base literature and developed algorithms to solve these tasks for the embedding model.

1. Learning process

First, we examined whether both models could learn to distinguish the correct and wrong triplets. For example, we tested the model output for triplets related to *canary*, adding a false triplet (i.e., "canary-is-red") as comparison. The results are averaged across 10 runs.

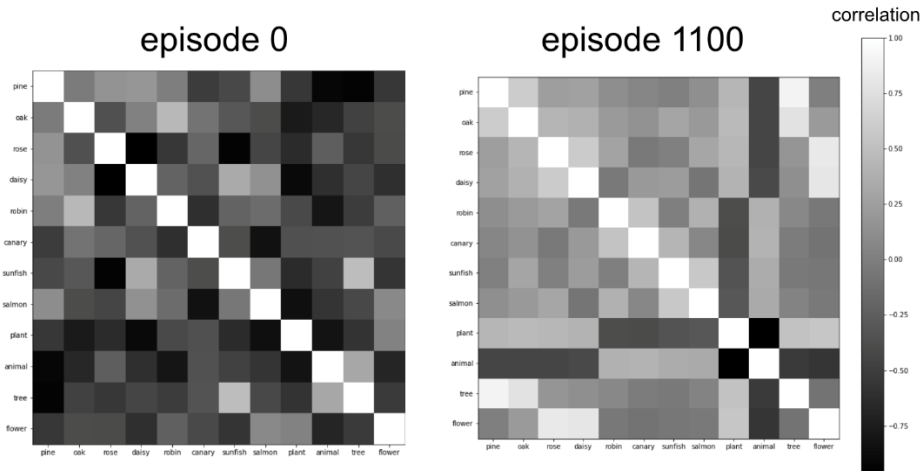


Rogers and McClelland (2004) observed an overgeneralization phenomenon for this data: "pine-has-leaves" is initially viewed by the model as correct because the other tree in the dataset (i.e., oak tree) does have leaves. This overgeneralization phenomenon was also present in our embedding model (see below). Early in learning, the model at first thought that pine do have leaves.



2. Emergence of clusters

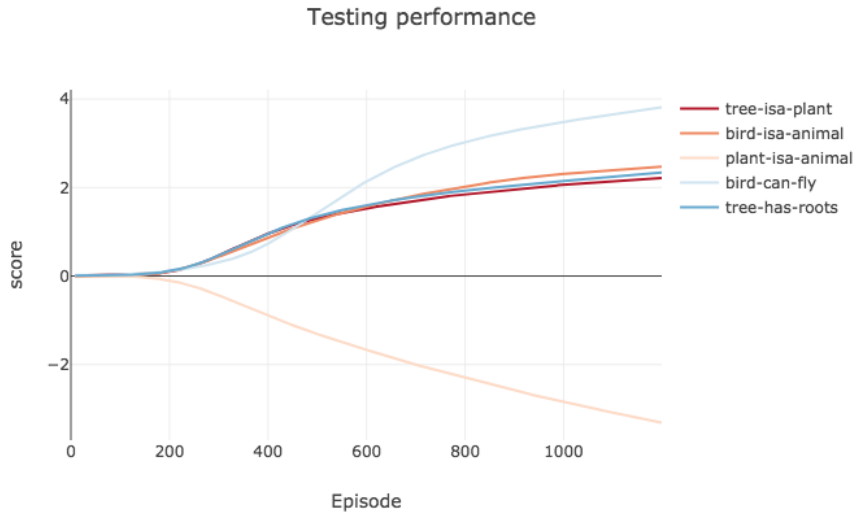
To check whether similar entities gradually develop similar embedding patterns and consequently whether a hierarchical structure emerges, we compared the correlation matrix of the embeddings from the first episode to those of 1100th episode.



We can see the clustering structure emerge among the eight living things. Moreover, higher-order concepts like plants and animals show strong dissimilarity from each other, reflecting the conceptual divide between the superordinate categories *plant* and *animal*. Note that the training data does not include triplets that explicitly describe the abstract concept *animal*. It is only included as a tail in statements like "fish-isA-animal"—yet the model learned to infer the properties of animals. In the next section, we further explore the inferences made by the embedding model.

3. Inference task

To examine what the embedding model can learn about abstract concepts that only appeared in the tails of "isA" relations, we tested an additional set of statements where the abstract concept appears in the tail (e.g., “tree-isA-plant”) that were not included in the training set and looked at the learning progress. As shown below, the embedding model is able to learn to assign correct judgements to these relations, e.g. it knows that "bird-isA-animal" is correct but that "plant-isA-animal" is wrong.



4. Automated discovery

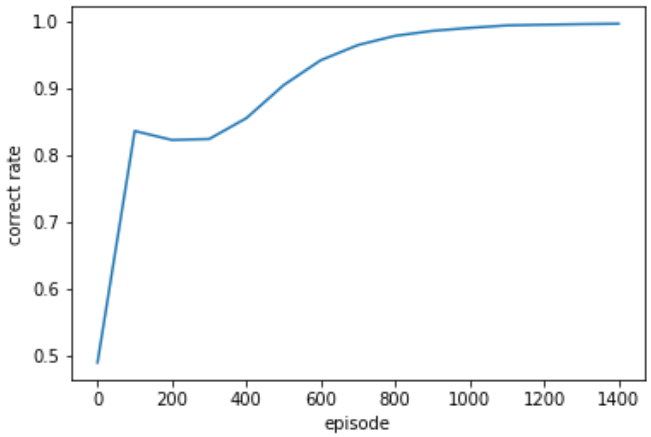
In the previous task, the novel triplets tested were provided by humans and approved by the model. Can our model discover novel knowledge by itself? To test this, we searched all the grammatical relations for a given entity and set a threshold to filter strong positive/negative judgements. With a threshold set at 4.5, our model correctly learned true triplets, such as “animal-has-skin” and “flower-is-pretty”, while also correctly learning that other triplets are false, e.g., “animal-has-roots”. While the vast majority of the triplets generated by this method are evaluated by the model in a manner consistent with human evaluation, the model had difficulty evaluating novel triplets that had “living thing” as the head. For example, it evaluated "livingthing-can-grow" as false despite being something that humans evaluate as true. This difficulty likely stems from the dataset does not include any examples of non-living things, thus, providing the model no point of comparison.

Novel triplet	**Model evaluation**	**Consistent with human evaluation**
livingthing-isa-livingthing	False	No
livingthing-is-living	False	No
livingthing-can-grow	False	No
plant-has-skin	False	Yes
plant-has-roots	True	Yes
animal-has-skin	True	Yes
animal-has-roots	False	Yes
tree-isa-tree	True	Yes
tree-is-pretty	False	Yes
tree-is-big	True	Yes
tree-has-bark	True	Yes
tree-has-branch	True	Yes
flower-isa-flower	True	Yes
flower-is-pretty	True	Yes
flower-has-leaves	True	Yes
flower-has-bark	False	Yes
flower-has-branch	False	Yes
flower-has-petals	True	Yes
bird-has-feathers	True	Yes
fish-has-gills	True	Yes
fish-has-scales	True	Yes

Canadian mammals dataset (the bigger dataset).

1. Learning process

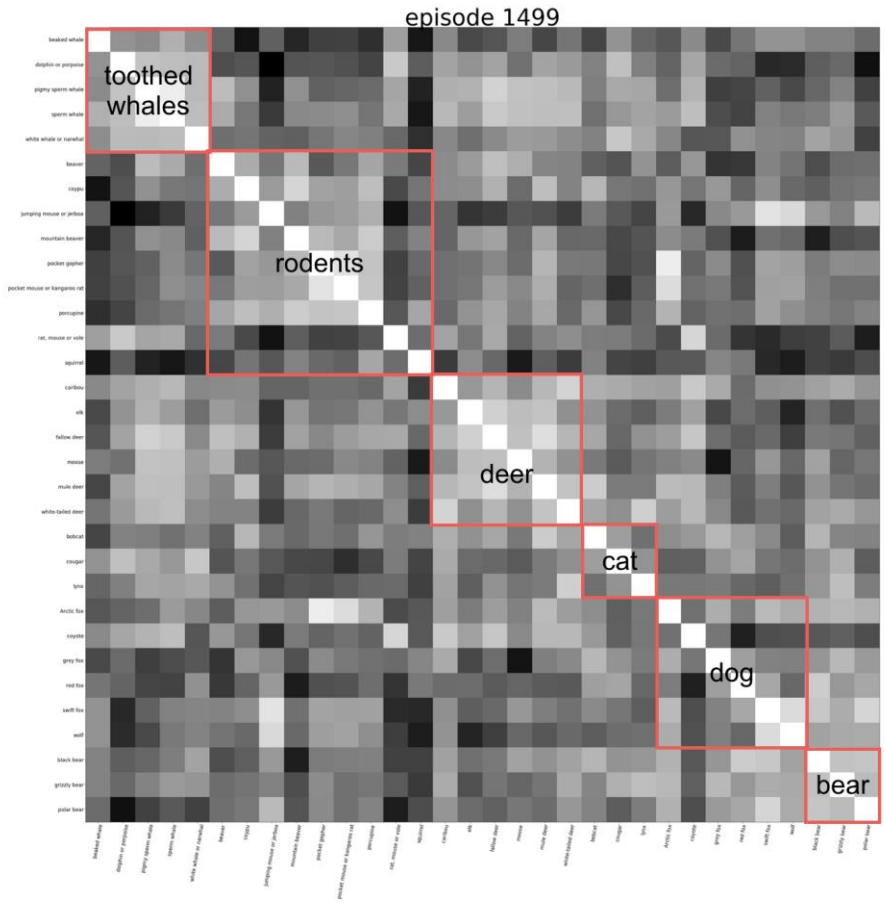
Our larger dataset includes 766 entities, many being very specific in nature (e.g., “baby long-tailed weasel”, “female little brown bat”, etc.). Thus, it is hard to pick representative triplets to demonstrate the training process, and we simply checked the correctness of across all triplets.



As shown in the graph above, the model does reach a rate of 100% correct at around 1200 episodes. It should be noted that we did not need to adjust the network or learning parameters used with the small dataset when examining this larger dataset. One possibility as to why we see the dip in the rate of correct at around episode 300 may be due to patterns of overgeneralization by the model. Further exploration would be needed to confirm this.

2. Emergence of clusters

For clarity, rather than showing clustering patterns for all 766 entities, we present 47 animals that belong to a subset of 6 higher-order concepts (*toothed whale*, *rodent*, *deer*, *cat*, *dog*, and *bear*).



After training, clusters emerge, but some mammal types show more homogeneity (e.g., toothed whales, deer) than others (e.g., rodents, dogs).

3. Automated discovery

We replicated the method we used for the smaller dataset to find new probable facts. We limited discovery to new concepts about higher-order categories (like those mentioned in the previous section). We set the threshold to 5 and found in total 8 new relations:
cloven-hoofed mammal-is eaten by-wolf, rodent-has habitat-farmlands, cat-chases away-opposite sex after mating, cat-eats-frog, dog-has part-short tail, dog-is a kind of-beaked whale, bear-eats-carrion, bear-lives in-hollow log. While the marjority of these make sense (e.g., "dog-has part-short tail"), some—perhaps only one—of these do not (e.g., "dog-is a kind of-beaked whale").

4. Surprising knowledge

Another component of knowledge acquisition that is empirically interesting is how people respond to learning something surprising. For example, people may find it surprising that bats can fly despite not being a kind of bird or insect. Learning something surprising like this may prompt further questions and be a source of curiosity in a way that an unsurprising fact would not. Most knowledge-base models are not capable of judging the surprising-ness of facts.

Here, we provide a possible procedure for identifying surprising facts through an example from our dataset. From 6 subcategories belonging to the category "dog", we chose the most atypical category exemplar (i.e. the concept with least correlation with other concepts)—which in this case, is the category "coyote" (as can be seen from the correlation matrix in section 2). Next, we took each relation associated with "coyote" and replaced "coyote" with its superordinate category "dog". Then, for each pair of triplets, we computed the difference of the prediction scores from the embedding model. Using this method, the three most potentially surprising facts were: "coyote-makes sound-howl", "coyote-makes sound-yelp", and "coyote-is eaten by-wolf".

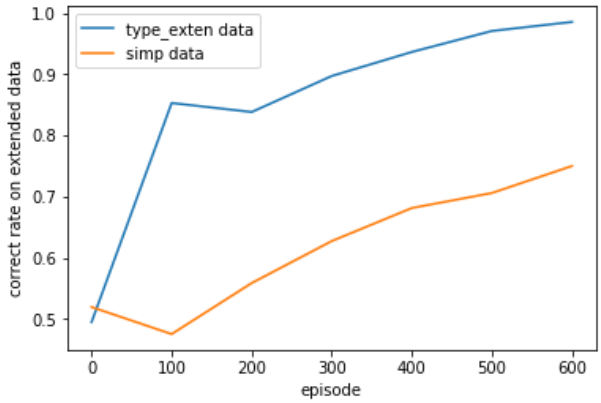
5. Integrating logic into knowledge base learning

At this point, our model is making inferences solely based on similarity within the embedding space. However, human learners can also exploit logic (e.g., syllogisms) to make inferences. For example, if they know that coyotes are a kind of dog and dogs are carnivorous, they likely infer that coyotes are also carnivorous.

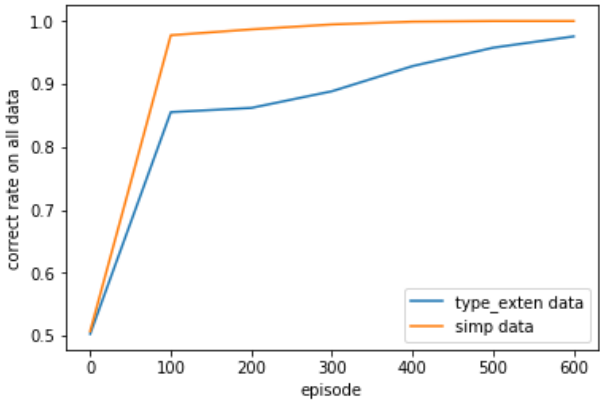
We designed an experiment to explore if adding the capability to employ logic will increase the model's ability to make inferences. If true, it provides evidence for the benefit of constructing superordinate concepts. Here, we focus on the "is a kind of" relation and teach the model the following syllogism:

if : "x - is a kind of - A" && if "A -r - t",
then : "x - r - t".

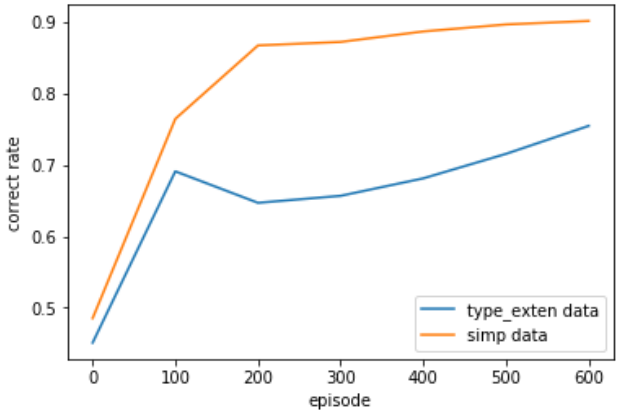
We incorporated this logic by extending our training dataset to include new triplets created by this rule and then trained the model on this extended dataset. In this first graph, we show the correctness on the extended dataset as a check to ensure that the model trained on the extended dataset does, in fact, learn the facts about subordinate categories better.



In this second graph, we show the correctness on each model's respective training data. As expected, the extended dataset slowed down the rate of learning.



Finally, we compared the correctness on the test dataset between the model trained on the original training dataset versus that trained on the new extended dataset. In both cases, the test dataset is not seen during training.



Contrary to our prediction, adding more data did not improve the model's performance to make inferences. One reason we may have found this result could be due to the simplicity of this dataset. A pattern throughout the Canadian mammal dataset is that as long as there are relations

like “coyote is a kind of dog" and “dog is carnivorous”, there is not an explicit relation for “coyote is carnivorous”. Thus, the kind of new data that will benefit most from our procedure won’t exist in our testing dataset. But this may not necessarily true in larger, less well-simplified knowledge bases, and is definitely not true to people’s daily experience. Therefore, testing on other types of datasets is needed to examine the efficacy of this logic-augmented training approach.

References

Aslin, R. N., & Newport, E. L. (2012). Statistical Learning. *Current Directions in Psychological Science, 21*(3), 170–176. http://doi.org/10.1177/0963721412436806

Bordes, A., Usunier, N., Weston, J., & Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-Relational Data. *Advances in NIPS, 26*, 2787–2795.

Brandone, A. C., & Gelman, S. A. (2009). Differences in preschoolers' and adults' use of generics about novel animals and artifacts: a window onto a conceptual divide. *Cognition, 110*(1), 1–22.

Brandone, A. C., & Gelman, S. A. (2013). Generic Language Use Reveals Domain Differences in Children's Expectations about Animal and Artifact Categories. *Cognitive Development, 28*(1), 63–75.

Carey, S. (1985). *Conceptual change in childhood*. MIT Press.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Cimpian, A., & Park, J. J. (2014). Tell me about Pangolins! Evidence that children are motivated to learn about kinds. *Journal of Experimental Psychology: General*, *143*(1), 46–55.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, *20*(1), 65–95. Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*(5), 620–629.

Graham, S. A., Kilbreath, C. S., & Welder, A. N. (2004). Thirteen-Month-Olds Rely on Shared Labels and Shape Similarity for Inductive Inferences. *Child Development*, *75*(2), 409–427.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. MIT Press.

Keil, F. C., Greif, M. L., & Kerner, R. S. (2007). A world apart: How concepts of the constructed world are different in representation and in development. In Margolis E & Laurence S (Eds.), *Creations of the mind: Theories of artifacts and their representation* (pp. 231–245). New York, NY: Oxford University Press.

Kushnir, T., Wellman, H. M., & Gelman, S. A. (2009). A self-agency bias in preschoolers' causal inferences. *Developmental Psychology*, *45*(2), 597–603.

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, *117*(1), 1–31.

Mareschal, D., & French, R. (2000). Mechanisms of Categorization in Infancy. *Infancy, 1*(1), 59–76.

Moty, K. & Brandone, A. C. (submitted). Beliefs about category homogeneity and variability in familiar and unfamiliar categories.

Murphy, G. L. (Gregory L. (2002). *The big book of concepts*. MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.), 274*(5294), 1926–8.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016, June). Complex embeddings for simple link prediction. In International Conference on Machine Learning (pp. 2071-2080).

Wellman, H. M. (1990). *The child's theory of mind*. MIT Press.

Yang, B., Yih, W. T., He, X., Gao, J., & Deng, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575.