The relational processing limits of classic and contemporary neural network models of language processing

Guillermo Puebla (guillermo.puebla@ed.ac.uk)
Department of Psychology University of Edinburgh, 7 George Square
Edinburgh, EH8 9JZ United Kingdom

Andrea E. Martin (andrea.martin@mpi.nl)
Max Planck Institute for Psycholinguistics, Wundtlaan 1
Nijmegen, 6525XD The Netherlands

Leonidas A. A. Doumas (alex.doumas@ed.ac.uk)
Department of Psychology University of Edinburgh, 7 George Square
Edinburgh, EH8 9JZ United Kingdom

Abstract

The ability of neural networks to perform relational reasoning is a matter of long-standing controversy. Recently, some researchers have argued that (1) classic PDP models can learn relational structure and (2) the successes of deep learning suggest that structured representations are unnecessary to explain human language. In this study we tested a classic PDP model and a contemporary deep learning model for text processing. Both models were trained to answer questions about stories based on the thematic roles that several concepts played on the stories. In three critical test we varied the statistical structure of new stories while keeping their relational structure intact with respect to the training data. Both models performed poorly in our tests. These results cast doubts on the suitability of traditional neural networks for explaining phenomena based on relational reasoning.

Keywords: relational reasoning; deep neural networks

Introduction

The ability to represent and reason in terms of the relations between objects plays a crucial role across human cognition (Halford, Wilson, & Phillips, 2010). Several computational models in cognitive science have sought to capture its main characteristics and development (for a review see, Gentner & Forbus, 2011).

These models differ in their representational assumptions. In the canonical view, relational reasoning entails using predicate representations. A predicate is an abstract structure that can be dynamically bound to an argument, specifying a set of properties about that argument (Doumas & Hummel, 2005). For example, *predator*(x) specifies a series of properties about the variable x (e.g., carnivore, hunts, etc.). Predicate representations have two main attributes. In the first place, predicates maintain role-filler independence in that at least some aspect of the semantic content of the predicate is invariant with respect to its arguments. For example, *predator*(fox) and *predator*(lynx) will specify the same set of properties (e.g., carnivore, hunts, etc.) about the objects fox and lynx. In the second place, predicates can be dynamically bound to arguments, namely, fillers can be assigned and reassigned to different roles as needed during processing. Models based on predicates successfully account for a wide variety of phenomena in the relational thinking literature (for a review see Forbus, Liang, & Rabkina, 2017).

By contrast, traditional Parallel Distributed Processing (PDP) models explicitly eschew structured representations (see, e.g., Rogers & McClelland, 2014). In these models representations are patterns of activation across a layer of units. These representations are unstructured because relational roles and objects are not independently represented, but instead are compressed together into a fixed-sized vector. Recently, Rogers and McClelland (2014) have proposed that the gestalt models of text comprehension (St. John, 1992; St. John & McClelland, 1990) exhibit successful effective role-to-filler binding. Some of this optimism is based on the achievements of deep learning architectures in natural language processing. For example, Rabovsky, Hansen, and McClelland (2018) argue that the success of Google's neural machine translation (GNMT) system (Wu et al., 2016) implies that structured representations are an obstacle to capturing the regularities of human language.

In the present study, we tested the Story Gestalt (SG) model (St. John, 1992) and a Sequence-to-Sequence with Attention (Seq2seq+Attention) model (Bahdanau, Cho, & Bengio, 2015)—the architecture behind the GNMT system—in a series of tasks requiring binding a number of concepts to several roles in a story.

Task overview

Our task, based on the original materials of St. John (1992), consists on answering questions about stories generated by a series of (5) scripts. All the scripts describe events as a sequence of propositions where several concepts play different thematic roles: agent-1, agent-2, topic, patient-theme, recipient-destination, location, manner and attribute. As an illustrative example, consider the Restaurant script (Table 1). This script describes an event where two people go to a restaurant. Each sentence of the Restaurant script defines fillers for some roles. To generate a specific instance of a

Restaurant script (i.e., a Restaurant story) the roles are given values corresponding to specific concepts. Table 2 presents an example of an instantiated Restaurant story in a pseudo-natural language format. Note that, as illustrated in Table 1, our scripts produce stories with no repeated topic concepts across propositions.

Table 1: Restaurant Script.

Script
1. [agent-1] and [agent-2] decided to go to restaurant
2. Restaurant quality [expensive/cheap]
3. Distance to restaurant [far/near]
4. [agent-1/agent-2] ordered [cheap-wine/expensive-wine]
5. [agent-1/agent-2] paid bill
6. [agent-1/agent-2] tipped waiter [big/small/not]
7. Waiter gave change to [agent-1/agent-2]
Concept restrictions
The roles agent-1 and agent-2 are never 'Lois' or 'Albert'
Deterministic rule
The quality of the restaurant determines the distance completely: $expensive \rightarrow far$, $cheap \rightarrow near$

Each script implements a tree structure where each node represents a proposition and each branch of the tree represents a story. The scripts also implement rules that specify the probability of transitioning from one node to another conditioned on the value of a character or location role. For example, a rule in the Restaurant script (see Table 1) specifies that if the restaurant is expensive, it will be located far away.

We had two training conditions. In the *concept restricted condition*, some character or object names were never used in specific scripts. For example, in the Restaurant stories the characters Lois and Albert were never used to fill the roles agent-1 or agent-2. In the *concept unrestricted condition* all concepts were used in all stories. Stories were generated as follows: (1) a script is chosen at random, (2) a sequence of propositions is generated by traversing the tree structure of a scrip and (3) character and vehicles names are given specific values (respecting the script's deterministic rule and the script's concept restrictions if necessary).

To get a criterion for each model's performance we designed a *baseline test*. We presented the models trained in the unrestricted condition with concept unrestricted stories and asked questions about them. The questions were the concepts filling the topic role. The correct answer was the full proposition in which the topic concept was involved. For example, if a proposition in a restaurant story stated that the "waiter gave change to Anne" and the model was asked about the "gave" proposition the correct answer was "waiter gave change to Anne". Because in our stories there was no repeated topics the correct answer was unequivocal. Table 2 presents an example of a Restaurant baseline story, its questions and correct answers.

Table 2: Example of a Baseline Story (Restaurant).

Story	
1. [Anne] and [Gary] decided to go to restaurant	
2. Restaurant quality [expensive]	
3. Distance to restaurant [far]	
4. [Anne] ordered [cheap-wine]	
5. [Anne] paid bill	
6. [Anne] tipped waiter [big]	
7. Waiter gave change to [Anne]	
Question	**Criteria**
decided	[Anne] and [Gary] decided to go to restaurant
quality	Restaurant quality [expensive]
distance	Distance to restaurant [far]
ordered	[Anne] ordered [cheap-wine]
paid	[Anne] paid bill
tipped	[Anne] tipped waiter [big]
gave	Waiter gave change to [Anne]

Models

Story gestalt model The SG model (St. John, 1992, see Figure 1) integrates a sequence of propositions into a distributed representation of a story, which is then used to answer questions about the story. The model represents all propositions in its input layer through 137 localist units coding for each possible filler of each role (e.g., there is a unit coding for Albert-agent and another unit coding for Albert-recipient). To represent a complete proposition, the units coding for the concept filling each role are activated. For example, a representation of the sentence Anne and Gary decided to go to the restaurant would consist of a vector of 137 units were the three units coding for Anne-agent, Gary-agent, decided-topic and restaurant-location are set to 1 and all other units are set to 0 (Figure 1A).

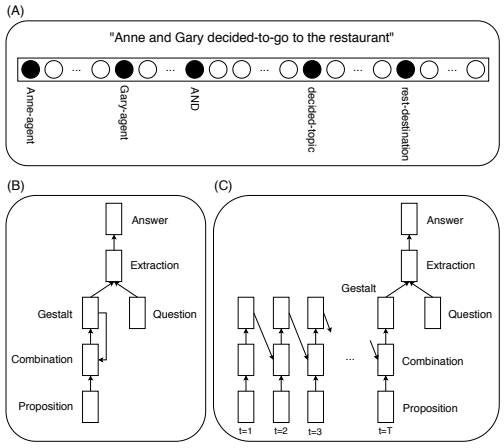

Figure 1: Story Gestalt model.

Seq2seq with attention model The Seq2seq+Attention model (Bahdanau et al., 2015, see Figure 2) is a deep neural

network architecture originally designed to solve translation problems. Typically, the source and target sentences have different lengths. In general, a Seq2seq model consist of an encoder network and a decoder network. Both are recurrent neural networks with their own independent time steps (*t* for the encoder and *t'* for the decoder in Figure 2B). The encoder transforms the input sequence into a sequence of fixed-size vectors and the decoder processes these transformed vectors to get the output sequence. Two important features this model are the use of word2vec representations for the input words (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and an attention mechanism that allows the model to selectively attend to different parts of the encoders output (Bahdanau et al., 2015).

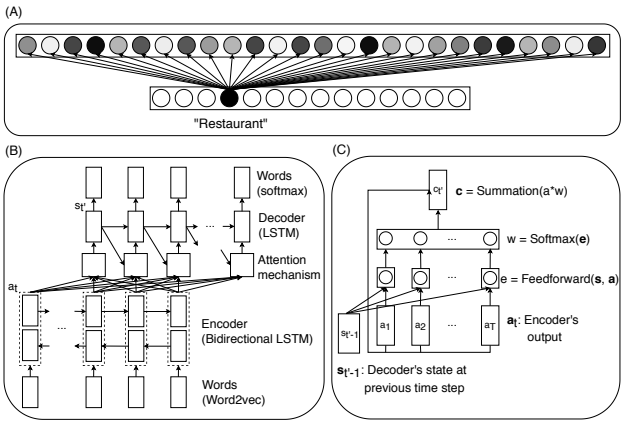

Figure 2: Seq2seq+Attention model.

Simulations

We designed three critical tests for the models. In our first test, termed *concept violation*, we trained the models in the concept restricted condition and then tested them with stories where the some roles were filled by the restricted concepts. For example, the concept Lois had never appeared as agent in any Restaurant story during the model's training (see Table 1). The model was then tested using a Restaurant story in which Lois appeared as agent by asking, for example, about the "tipped" proposition. The correct (role-based) answer was "Lois tipped waiter big". Note that, while the model was trained in stories where Lois appeared as an agent in other locations, and had been trained to output that someone tipped big with other agents, it had never been trained to output the exact proposition "Lois tipped waiter big".

In our second test, termed *correlation violation*, we presented the models trained in the concept unrestricted condition with stories where we inverted a perfect statistical regularity of the story script. For example, a rule in the Restaurant script establish that if the restaurant was cheap it was nearby and if it was expensive it was far away (see Table 1). To create a Restaurant correlation violation story, we switched the second term of the correlation (e.g., a cheap restaurant that was far away) and asked abut the "distance" proposition. The role-based answer was "The restaurant was far away", even though all cheap restaurants were close by during training.

In our third test, termed *shuffled propositions*, we presented the models trained in the concept unrestricted condition with stories where we randomized the order of the propositions. Recall that in our stories there are no repeated topic concepts. As a direct consequence, a role-based answer to a question should use the concepts of the proposition corresponding to each question to fill its roles, ignoring the ordering.

Training

We trained two versions of the SG model, one in 1,000,000 randomly generated concept restricted stories and another in 1,000,000 randomly generated concept unrestricted stories. We also trained two versions of the Seq2se2+Attention model, one in 500,000 randomly generated concept restricted stories and another in 500,000 randomly generated concept unrestricted stories. We used the Nadam optimization algorithm with default learning parameters.

Results

For each of our tests, we created a dataset of stories by generating 1,000,000 stories and saving all unique ones. Due to the combinatorics of concepts and scripts, these datasets had different sizes (baseline and shuffled sentences: 14,652, concept violation: 728, correlation violation: 14,647). For all tests we compared the proposition generated by the model with the role-based answer. We coded the answer as correct (with a value of 1) if the all the concept fillers in the answer corresponded to the concept fillers in the role-based answer and as a incorrect (with a value of 0) otherwise. Figure 3 shows the proportion of correct answers per test and model. As can be seen, both models performed well in our baseline test.

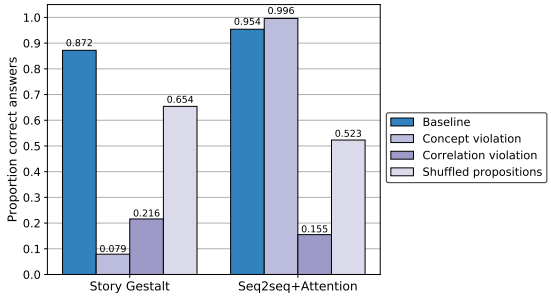

Figure 3: Results per test and model.

In our concept violation test the SG model almost invariably filled the roles of the restricted concepts with the most common concepts playing that role during training. For example, if it was presented with a story were the role agent-1 the restricted "Lois", the model tended to output answers where the agent-1 was any of the other unrestricted agents (e.g., "Barbara" or "Clement"). The Seq2seq+Attention model

performed significantly better at this test. The attention mechanism seems to allow this model to apply its word representations to previously unseen sequences of words.

Both models performed poorly in the correlation violation test. Such behavior would seem quite unnatural for a human reader as it would amount to, when presented with a proposition stating that a restaurant is close by, answering the question "where is the restaurant" by stating the restaurant is far away. Notably, the SG model achieved a higher accuracy than Seq2seq+Attention model in this test. We suspect that the same attention mechanism that allows the Seq2seq+Attention model to pass the concept violation test makes it even more likely to overfit to a perfect correlation in the dataset.

While our shuffled proposition test affected both models, the SG model performed significantly better than the Seq2seq+Attention model. We again hypothesize that the attention mechanism is the main reason for this difference in performance. Unfortunately, due to the length of our stories, taking out the attention mechanism yields the Seq2seq+Attention model unable to pass our baseline test, so we could not test our hypothesis directly.

Discussion

We tested the relational processing capabilities of a classic and a contemporary neural network model of text comprehension. In three critical tests we varied the statistical properties of the test stories while keeping their relational structure intact. Our results show clearly that these models are not using the relational information of the stories to answer the questions, but instead they are relying on the statistical regularities of the training dataset.

Our results are highly consistent with the findings of (Lake & Baroni, 2018), who found that sequence-to-sequence models failed at a command-to-action translation task that required composing the meaning of new commands formed by using known primitive concepts combined in ways unseen during training. Truly compositional behavior requires independent representations of objects and roles that can be bound together dynamically. A model that dynamically binds roles to fillers would easily pass our tests by filling the untrained concepts into the trained roles to answer the questions (see, Doumas & Hummel, 2005).

Interestingly, there has been a resurgence of interest on the binding problem in neural networks (Besold et al., 2017). Moreover, relational learning and reasoning have become a core topic on deep learning research (for a review, see Battaglia et al., 2018) with some deep learning architectures implementing elements traditionally associated with symbolic processing such as a content-addressable memory (e.g., Graves et al., 2016). Whether these non-traditional neural network architectures are capable of relational reasoning remains an open question. Our results suggest, however, that for a model to successfully account for all aspects of relational processing, it will need to implement a solution to the binding problem.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International conference on learning representations.*

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... Pascanu, R. (2018). *Relational inductive biases, deep learning, and graph networks.*

Besold, T. R., d'Avila Garcez, A., Bader, S., Bowman, H., Domingos, P., Hitzler, P., ... Zaverucha, G. (2017). *Neural-symbolic learning and reasoning: A survey and interpretation.*

Doumas, L., & Hummel, J. E. (2005). Approaches to modeling human mental representations: What works, what doesnt and why. *The Cambridge handbook of thinking and reasoning, ed. KJ Holyoak & RG Morrison*, 73–94.

Forbus, K. D., Liang, C., & Rabkina, I. (2017). Representation and computation in cognitive models. *Topics in cognitive science*, *9*(3), 694–718.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, *2*(3), 266–276.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... others (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471.

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in cognitive sciences*, *14*(11), 497–505.

Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th international conference on machine learning.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the n400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive science*, *38*(6), 1024–1077.

St. John, M. F. (1992). The story gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science*, *16*(2), 271–306.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial intelligence*, *46*(1-2), 217–257.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation.*