Textual Time Travel: A Temporally Informed Approach to Theory of Mind

Akshatha Arodi
Mila/McGill University
Montreal, QC, Canada
akshatha.arodi@mail.mcgill.ca

Jackie Chi Kit Cheung
Mila/McGill University
Montreal, QC, Canada
jcheung@cs.mcgill.ca

Abstract

Natural language processing systems such as dialogue agents should be able to reason about other people's beliefs, intentions and desires. This capability, called *theory of mind* (ToM), is crucial, as it allows a model to predict and interpret the needs of users based on their mental states. A recent line of research evaluates the ToM capability of existing memory-augmented neural models through question-answering. These models perform poorly on false belief tasks where beliefs differ from reality, especially when the dataset contains distracting sentences. In this paper, we propose a new temporally informed approach for improving the ToM capability of memory-augmented neural models. Our model incorporates priors about the entities' minds and tracks their mental states as they evolve over time through an extended passage. It then responds to queries through textual time travel—i.e., by accessing the stored memory of an earlier time step. We evaluate our model on ToM datasets and find that this approach improves performance, particularly by correcting the predicted mental states to match the false belief.

1 Introduction

Humans have evolved social intelligence to reinforce cooperation in society (Brüne and Brüne-Cohrs, 2006). In human interactions, understanding another person's mental states — for example, their purpose, intention, knowledge, belief, thoughts, doubts, likes and needs — is a critical step in correctly interpreting or predicting their behavior (Yott and Poulin-Dubois, 2016; Premack and Woodruff, 1978). This ability to attribute mental states to oneself and to others, called *theory of mind* (ToM), becomes increasingly necessary for natural language processing (NLP) systems as they integrate into modern society. Acquiring ToM capabilities permits more accurate responses in several situations. For example, it allows for disambiguating a difficult query by correctly deducing the true needs of the user based on their mental state, thereby providing the missing piece in solving inference and reasoning tasks. Figure 1 shows a scenario with two people interacting with an intelligent virtual assistant. In this example, the assistant (Bot) uses theory of mind to reason about Ella's belief about the location of the umbrella, which differs from reality. Clearly, demonstrating

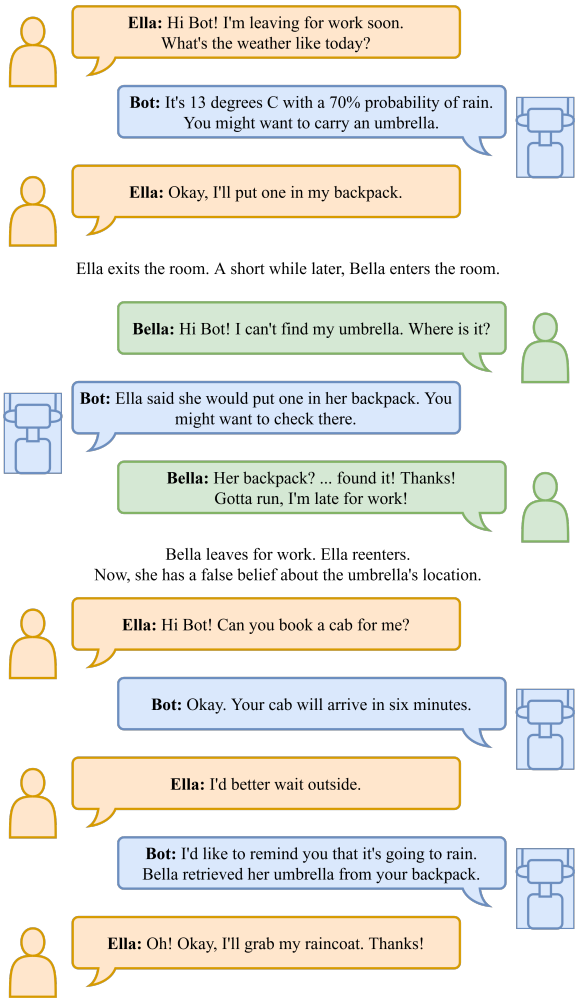

Figure 1: Ella, Bella and the umbrella—an example of how theory of mind matters for a virtual assistant application.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4162–4172
November 7–11, 2021.

social intelligence is a major barrier that needs to be crossed in our march towards the applied end-goal of creating NLP systems that blend seamlessly into the human world (Weston et al., 2016; Bisk et al., 2020).

A recent line of work evaluates the theory-of-mind capability of memory-augmented neural models via a question answering task (Grant et al., 2017; Nematzadeh et al., 2018; Le et al., 2019). This task consists of stories with multiple entities that interact with each other in a synthetic environment, followed by a question about the entities' beliefs. Memory augmented neural models like EntNet (Henaff et al., 2017), that are successful at solving reasoning tasks such as bAbi (Weston et al., 2016) by tracking *world* states, perform poorly at *false belief* tasks where the *mental* states of the entities do not match the world states. Moreover, these models are sensitive to distracting sentences, which decrease their accuracy.

In this paper we propose a new model, that we call *Textual Time Travel*, to correctly track the mental states of the entities when they have a false belief. Our key insight is to incorporate priors about the entities' minds in our model and add a temporal dimension to the neural model's memory. This allows us to track the changes in the mental states with time as the story progresses.[1] We also aim to have an *interpretable* model of the world and the mental states of the entities. Our temporally-informed neural model allows us to visualize how it tracks changes in these states. We find that our model does indeed track the mental states of the entities, and with additional supervision it can provide correct responses when the entities have false beliefs, thus improving performance.

2 Related Work

Theory of mind has been extensively researched by the psychology community over the last few decades (Baron-Cohen, 1997; Flavell, 2004). Generally, false belief tasks are used to test the ToM capabilities of children and animals (Premack and Woodruff, 1978; Wimmer and Perner, 1983; Leslie and Frith, 1988; Heyes, 1998; Wellman, 2014). In developmental psychology, the famous Sally-Anne test (Baron-Cohen et al., 1985) is widely adopted to assess a child's ability to attribute false beliefs to others.

[1] Our code and dataset can be found at: *Textual Time Travel*.

2.1 False Belief Tests

The Sally-Anne test (Baron-Cohen et al., 1985) evaluates the ability of children to reason about others' false beliefs. In this test, the child observes two dolls, Sally and Anne. Sally first places a marble into her basket and leaves the scene. The marble is then transferred by Anne and hidden in her box. When Sally returns, the child is asked, "Where will Sally look for her marble?". To pass the test, the child has to understand that Sally does not know that the marble is in the box, and thus has a *false belief* about the location of the marble. The child is also asked *memory* and *reality* questions: "Where was the marble in the beginning?", and "Where is the marble really?"

The ice-cream-van test (Perner and Wimmer, 1985) aims to evaluate *beliefs about beliefs*, i.e., second-order beliefs in children. In this test, the child observes John and Mary, who see an ice-cream van in a park. The ice-cream man tells them that he will be in the park all afternoon, and they make plans to get ice cream later in the day. Mary leaves the park alone, and the ice-cream man, after a change of plans, tells John that he is going to the church. On the way to the church, the ice-cream man happens to run into Mary, and he fills her in about his updated plans. The child is asked "Where does John think Mary goes to get ice-cream?". The child has to recognise that John doesn't know that Mary knows the ice-cream van's location, and therefore has a false belief about Mary's belief. The child is also asked corresponding control questions on memory, reality and first-order false belief to verify the understanding of the environment.

2.2 ToM Through Question Answering

Based on the Sally-Anne test, Grant et al. (2017) propose a question answering task to evaluate the theory-of-mind capabilities of neural models. Nematzadeh et al. (2018) extend this work and include a second-order false belief task based on the ice-cream van experiments of Perner and Wimmer (1985) and propose an artificial dataset called ToM. This dataset is similar to bAbi (Weston et al., 2016) in that it consists of stories with multiple entities interacting with each other in a synthetic environment. Le et al. (2019) attempted to alleviate the biases of the ToM dataset with ToMi, an improved dataset and evaluation method.

Previous work showed that memory-augmented neural models such as the End-to-End Memory Net-

work (Sukhbaatar et al., 2015), the Recurrent Entity Network (EntNet) (Henaff et al., 2017), and the Relation Network (Santoro et al., 2017) perform well on bAbi tasks and poorly on the ToM and ToMi datasets, especially with distracting sentences. Of all of these memory-augmented neural networks, EntNet showed the most promising results on ToMi and performed best on false belief tasks.

In this paper, we extend this body of work on theory of mind with a temporally-informed memory-augmented neural model. We build heuristics into the ToMi dataset to enable our textual time travel, and evaluate our model on this dataset.

3 Theory of Mind Datasets

The ToM and ToMi datasets contain multiple stories with a structure similar to the Sally-Anne and ice-cream van tests, but with a variety of entities and objects. The ToM dataset (Nematzadeh et al., 2018) follows a strict template to generate the stories and has a simple random sentence as a distractor. The ToMi dataset (Le et al., 2019) aims to addresses the dataset biases of ToM by breaking the strict event sequence. It adds actions involving unrelated entities and randomizes the order of events in the story. Each story has two main entities, an object of interest, a main location where the events take place, and two containers for the object. The story begins with the introduction of the two entities, the object and their locations. ToMi consists of 1000 stories of 3 story types:

1. **True belief**: After the introductions, entity A moves the object from container1 to container2 in the presence of entity B. In this type of story, both entity A's and entity B's mental states match the world state, so they both have a true belief.

2. **First-order false belief**: After the introductions, entity A leaves the main location. Entity B moves the object from container1 to container2. In this type of story, entity A has false belief about the location of the object.

3. **Second-order false belief**: After the introductions, entity A leaves the main location. Entity B moves the object from container1 to container2 and leaves the location. Entity A reenters the main location and is now aware of the final location of the object. In this type of story, entity B has a false belief about entity A's belief.

1. William entered the bedroom.
2. Jackson loves the strawberry.
3. Jackson entered the bedroom.
4. Logan entered the living-room.
5. Logan exited the living-room.
6. The apple is in the red-envelope.
7. Jackson exited the bedroom.
8. William moved the apple to the green-basket.
Q1. Where was the apple at the beginning?
Q2. Where will William look for the apple?
Q3. Where does William think that Jackson searches for the apple?
Q4. Where is the apple really?
Q5. Where will Jackson look for the apple?
Q6. Where does Jackson think that William searches for the apple?

Table 1: ToMi example of a first-order false belief story. Here lines 2, 4, and 5 contain distracting sentences. The questions are, Q1: Memory, From William's perspective - Q2: First-order, Q3: Second-order, Q4: Reality, From Jackson's perspective - Q5: First-order, Q6: Second-order.

Each story is followed by a question. The model must provide an answer based on the mental states of the entities. In this task, the mental states refer to the beliefs of an entity about the locations of entities and objects in the artificial environment. Each story is included in the dataset six times, once for each question that is associated with it, distributed by type as follows:

- Two first-order questions about an entity's belief about the environment.
- Two second-order questions about an entity's belief about the belief of the other entity.
- One memory and one reality question.

Table 1 presents an example of a ToMi story. While inspecting the failure cases of our model, we observed an inconsistency in the gold standard answers (true labels) to second-order belief questions in the ToMi dataset (i.e., the true labels were incorrect for some particular types of second-order false-belief tasks). We corrected this inconsistency and re-generated the ToMi dataset with 1000 stories.[2] We evaluate our model on the corrected dataset.

[2]Section A in the appendix provides more details about the errors and our process for generating the fixed dataset.

4 Modelling ToM

4.1 The ToMi Task

Each story S consists of T sentences. For each time step $t \in T$, let s_t be the sentence embedding of the story at t. Given a query q, the answer $y(q)$ is a distribution over all vocabulary words V in the corpus; i.e., the set of all stories.

4.2 The Textual Time Travel Model

Our model has two stages, as illustrated by Figure 2. The first stage is a memory updater that stores a snapshot of the memory contents at each time step, from which we can derive the mental states of entities (LHS of Figure 2, Section 4.2.1). In the second stage, we use what we call the textual time travel mechanism: the model predicts a timestep of interest and retrieves the contents of the memory cells at that time step in order to predict an answer to a query (RHS of Figure 2).

We will present a heuristic rule-based (Section 4.2.2) and a learning-based (Section 4.2.3) version of the textual time travel mechanism. Both are derived from the following basic assumptions about the entities' minds:

- **Local entity perception** An entity is aware of the objects and other entities in their current location only, and not elsewhere.
- **Recency assumption** An entity assumes that the most recently available information that it has access to about an object or another entity is correct.
- **Reciprocity assumption** Entities assume that other entities also behave according to the local perception and recency assumptions above.

Whereas the heuristic method directly hard-codes a time travel mechanism based on these assumptions, the learned method learns to select an appropriate time step through training in order to implement these assumptions.

4.2.1 Memory Updater

The starting point of our model is the memory-augmented neural architecture of Henaff et al. (2017) (EntNet). Our model contains a fixed number of dynamic memory cells. At any time step t, each cell j has a key w_j, a value $h_j^{(t)}$ and a gate $g_j^{(t)}$. The gate $g_j^{(t)}$ controls the memory value updates using the key w_j and the previous value $h_j^{(t-1)}$. We let $M^{(t)}$ denote the memory contents at time step t. The memory updater has three main components.

The *input encoder* generates the sentence embedding $s^{(t)}$ at time t as a weighted sum of the word embeddings e_k of the sentence. Here, e_k represents the embedding of the k^{th} word in a sentence. The weights f_k are shared across time steps and are learned jointly with the other parameters.

$$s^{(t)} = \sum_i f_i \odot e_i$$

The *dynamic memory module* implements a gating mechanism for the memory cells. The gate $g_j^{(t)}$ is activated if $s^{(t)}$ matches the value $h_j^{(t-1)}$ or the key w_j of cell j. The new value $\tilde{h_j}^{(t)}$ of the gate is a weighted sum of $h_j^{(t-1)}$, w_j, and s_t. The value is updated if the gate is activated. Then, the new value is normalized. At each time step t we store the mental states, represented by the memory contents $m_j^{(t)}$.

$$\begin{aligned}
g_j^{(t)} &\leftarrow \sigma(s_t^T h_j^{(t-1)} + s_t^T w_j) \\
\tilde{h_j}^{(t)} &\leftarrow \phi(U h_j^{(t-1)} + V w_j + W s_t) \\
h_j^{(t)} &\leftarrow h_j^{(t-1)} + g_j \odot \tilde{h_j}^{(t)} \\
h_j^{(t)} &\leftarrow \frac{h_j^{(t)}}{||h_j^{(t)}||} \\
m_j^{(t)} &\leftarrow h_j^{(t)}
\end{aligned}$$

Here U, V, and W are the parameters of the model, σ is a sigmoid, and ϕ is a ReLU activation function.

The *output module* generates a distribution y over all the vocabulary words in the corpus. Given a question q, it computes an initial distribution $p_j^{[1]}$ over all memory cells. Additionally, our model computes a second distribution $p_j^{[2]}$ over all keys. These are then passed through a non-linearity to generate y.

$$\begin{aligned}
p_j^{[1]} &= \text{Softmax}(q^T h_j) \\
p_j^{[2]} &= \text{Softmax}(q^T w_j) \\
u^{[1]} &= \sum_j p_j^{[1]} h_j \\
u^{[2]} &= \sum_j p_j^{[2]} h_j \\
y &= R\phi(q + H^{[1]} u^{[1]} + H^{[2]} u^{[2]})
\end{aligned}$$

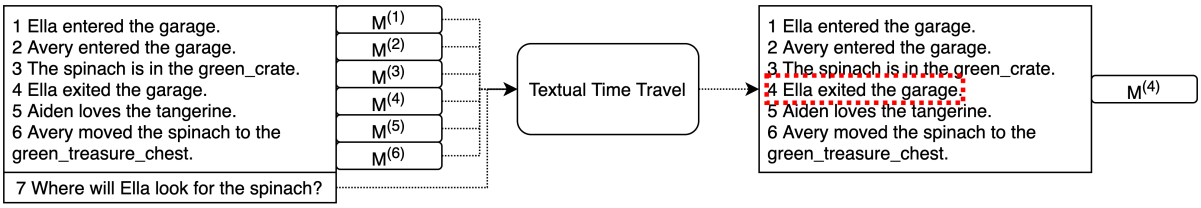

Figure 2: Textual Time Travel example. $M^{(t)}$ contains the mental states of the entities at time t.

Here, $H^{[1]}$, $H^{[2]}$ and R are the parameters. By default, the memory cells h_j are drawn from the memory contents at the last time step. In the textual time travel models below, they will actually be drawn from the time step that is predicted by the time travel mechanism.

4.2.2 Heuristic Textual Time Travel

This mechanism implements the assumptions in Section 4.2 and allows the model to go back in (textual) time to the correct time step and access the stored memory cells at that time step to answer a query. For a given question q, we define an attention a_q over all time steps T, that allows us to choose the correct time step t and fetch the relevant memory contents $m_j^{(t)}$ at that time step.

Specifically, for each entity, the model computes a *final exit time*. If the entity exits the location and does not reenter, this value is equal to the time step at which the exit takes place. If the entity never exits, or exits and reenters, this value is equal to the final time step in the story. Then, for first-order queries, the heuristic chooses the final exit time of the entity in the query. For second-order queries, it chooses the earlier of the two exit times between the two entities. The final response is generated by retrieving the values $h_j^{(t)}$ from the stored memory $m_j^{(t)}$ at the selected time step.

Table 2 shows an example for second-order false belief story. In this case, the model returns time step 6, since that is the last time Noah and Logan are in the main location, and Logan is unaware of Noah's mental states after he leaves the scene.

4.2.3 Learned Textual Time Travel

With this method, we train the model to obtain a distribution a_q over the time steps. Based on the local entity perception assumption, we first train the model to predict the locations of the entities in the world, then use that information to predict the time step for time travel (Figure 3). We introduce the following three labels for entity location prediction:

1. Noah entered the lounge.
2. Olivia entered the lounge.
3. Logan entered the lounge.
4. Olivia exited the lounge.
5. The spinach is in the blue-crate.
6. Logan exited the lounge.
7. Noah moved the spinach to the blue-suitcase
8. Where does Logan think that Noah searches for the spinach?

Table 2: Heuristic Textual Time Travel example of a second-order false belief story. Time step 6 is the last time step in which Logan and Noah were at the lounge.

1. UNKNOWN: The location of the entity is not known at this time. This indicates that either the entity has not yet been introduced or the entity has exited their last known location.

2. MAINLOCATION: The entity is introduced and is present at the main location at this time.

3. ALTERNATELOCATION: The entity has left the main location and is at a new location.

We augment the ToMi dataset to include questions about the location of the entities as shown in Table 3. Specifically, we add "What is the location of <entity>?" questions and answers to random time steps at random positions in the training data.

Given a question q about the mental state of entity A, we generate an auxiliary question, $\tilde{q}$, that asks "What is the location of A?" at each time step. Based on the prediction to $\tilde{q}$, we assign a location label to the entity. Table 4 shows an example of this process, tracking the location label of the entity Abigail from UNKNOWN to MAINLOCATION (i.e., the porch), then later to ALTERNATELOCATION (i.e., the hall).

Then, we generate an attention distribution a_q based on the location label of the entity. Using the attention a_q, the model attempts to return to

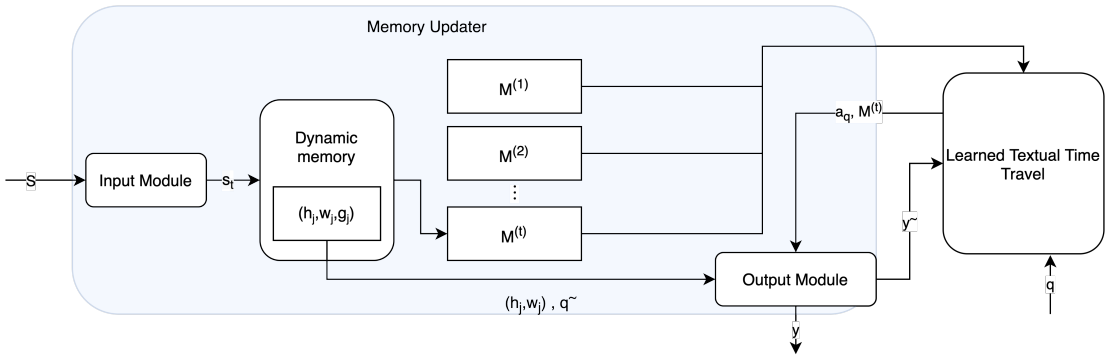

Figure 3: Learned Textual Time Travel model architecture

the time step t at which the entity exits the location containing the object, and retrieves the stored memory $m_j^{(t)}$ and the corresponding value $h_j^{(t)}$ to generate the final output using the output module.

1. Jacob entered the patio.
2. Jayden entered the patio.
3. Isabella entered the patio.
4. The pumpkin is in the red-box.
5. Jayden moved the pumpkin to the red-container.
6. Jacob exited the patio.
7. What is the location of Jacob? UNKNOWN
8. What is the location of Jayden? MAINLOCATION(patio)
9. What is the location of Isabella? MAINLOCATION(patio)

Table 3: The training data is updated with questions about the locations of the entities. Here, after Jacob exits the patio, we are not aware of his location, so we label it as UNKNOWN.

5 Experiments

As described above, to train the Learned Textual Time Travel model we add "What is the location of <entity>?" questions at various timesteps of several stories extracted from the training data. We then randomly sample 1000 of these stories for the training dataset. We initialize the PReLU slopes to 1, and initialize all the other weights with values drawn from a gaussian distribution with mean zero and standard deviation 0.1. We set the key w_j to the word embeddings of all the named entities in the dataset, and the memory cell contents $h_j^{(0)}$ are initialised with w_j. We remark that initializing w_j with GloVe (Pennington et al., 2014) pre-trained word embeddings did not improve performance. We use the Adam optimizer (Kingma and Ba, 2015) with a batchsize of 32, and clip the gradients at 40. We start with an initial learning rate of 0.01 and we halve the learning rate after every 25 epochs. We train the models on the corrected ToMi dataset for 200 epochs.

5.1 Evaluation measures and baseline

We evaluate our model using the accuracy score. We report accuracy based on the belief of the entities in the question (either true belief or false belief) and on each question category; namely, memory, reality, first-order and second-order. We evaluated the same set of models discussed in Nematzadeh et al. (2018) and Le et al. (2019) on the corrected ToMi dataset, namely the End-to-End Memory Network (Sukhbaatar et al., 2015), the Multiple Observer Model (Grant et al., 2017), the Recurrent Entity Network (EntNet) (Henaff et al., 2017), and the Relation Network (Santoro et al., 2017). We found that all of these models performed poorly on the false belief questions of the corrected dataset, and EntNet performed the best out of these, both overall and specifically on the false belief tasks. We therefore chose to compare our results against EntNet as our baseline.

6 Results

Table 5 shows the performance of (a) EntNet, (b) the Heuristic Textual Time Travel model, and (c) the Learned Textual Time Travel model. In terms of overall accuracy, both Textual Time Travel models outperform the baseline, and this difference is statistically significant in both cases ($p < 10^{-7}$, two-tailed Wilcoxon signed-rank). All three models achieve perfect accuracy on the memory and

1. Oliver dislikes the kitchen.	What is the location of Abigail?	UNKNOWN
2. Carter entered the porch.	What is the location of Abigail?	UNKNOWN
3. Abigail entered the porch.	What is the location of Abigail?	MAINLOC(porch)
4. The potato is in the green-suitcase.	What is the location of Abigail?	MAINLOC(porch)
5. Abigail exited the porch.	What is the location of Abigail?	UNKNOWN
6. Abigail entered the hall.	What is the location of Abigail?	ALTLOC(hall)
7. Carter moved the potato to the green-envelope.	What is the location of Abigail?	ALTLOC(hall)
8. Oliver entered the hall.	What is the location of Abigail?	ALTLOC(hall)
9. Where will Abigail look for the potato?		

Table 4: Learned Textual time travel: We predict the location of the entity at each time step, $\tilde{y}$, and visit the time step where the entity was last at the main location based on the label. Here, the last time Abigail was at the porch is at time step 5. So, the model predicts the response, y, by attending to $M^{(5)}$.

Model	Belief type	Row Avg.	Memory	Reality	First-order	Second-order
EntNet	Overall	90	100	100	87	84
	True Belief	87	-	-	92	79
	False Belief	83	-	-	68	89
Heuristic Textual Time Travel	Overall	**93**	100	100	91	**87**
	True Belief	86	-	-	91	78
	False Belief	94	-	-	94	95
Learned Textual Time Travel	Overall	92	100	100	**92**	85
	True Belief	85	-	-	91	75
	False Belief	95	-	-	98	94

Table 5: Model performance on the corrected ToMi dataset in terms of accuracy (%). The 'Row Accuracy' column reports the row-wise average of the accuracy scores.

reality questions. In particular, the modest decline in performance on true belief tasks is outweighed by the pronounced increase on false-belief tasks, from an average accuracy of 83% for the baseline to 94% and 95% respectively for the Heuristic and Learned models. The Learned Time Travel model outperforms the baseline with a 30% increase in the accuracy on first-order false belief task.

Note that there are about twice as many true-belief questions as false-belief ones. This explains the slight increase in the overall accuracy despite the large increase in performance on false-belief questions.

These results validate our hypothesis that building in assumptions about the evolution of entities' mental states into a model improves performance on theory-of-mind questions.

7 Discussion

In this section, we describe an ablation study and analyze the model by considering its internal memory gate activations, and by examining its errors.

Model	First	Second
EntNet	87	84
Memory Updater	86	85

Table 6: Accuracy (%) of the memory updater without textual time travel on first-order and second-order questions.

7.1 Ablation

In order to demonstrate that the performance gains of our models are due to the time travel mechanisms and not other model changes, we perform an ablation study by removing the time travel component of our models. This corresponds to the default model of the Memory Updater in Section 4.2.1. Table 6 shows that the first-order and second-order results of the Memory Updater is comparable to the EntNet baseline. Both models get an overall accuracy of 90%, and memory and reality accuracy of 100%, on the corrected ToMi dataset. This indicates that the textual time travel mechanism is indeed responsible for the observed improvements in performance.

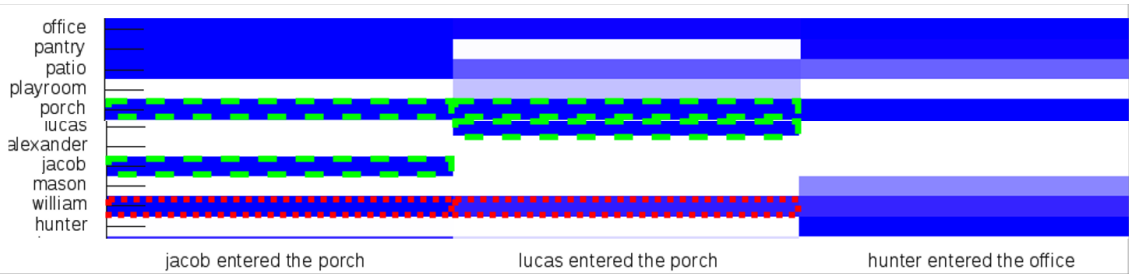

Figure 4: Visualizations of gate activations of various memory cells. Green boxes (dashed lines) indicate correct activations; red boxes (dotted lines) indicate incorrect ones.

1. Evelyn entered the bathroom.	What is the location of spinach?	unknown
2. Ella entered the hallway.	What is the location of spinach?	hallway
3. Avery entered the hallway.	What is the location of spinach?	hallway
4. Avery hates the strawberry.	What is the location of spinach?	unknown
5. The spinach is in the red-bucket.	What is the location of spinach?	unknown
6. Ella exited the hallway.	What is the location of spinach?	unknown
7. Avery moved the spinach to the suitcase.	What is the location of spinach?	unknown
1. Evelyn entered the bathroom.	Where is the spinach?	bathtub
2. Ella entered the hallway.	Where is the spinach?	bathtub
3. Avery entered the hallway.	Where is the spinach?	bathtub
4. Avery hates the strawberry.	Where is the spinach?	bathtub
5. The spinach is in the red-bucket.	Where is the spinach?	red-bucket
6. Ella exited the hallway.	Where is the spinach?	red-bucket
7. Avery moved the spinach to the suitcase.	Where is the spinach?	suitcase

Table 7: Analysis of Learned Textual Time Travel model after replacing "Ella" with "spinach" in the 'location' questions.

7.2 Memory activations

Analyzing the gate activations of memory cells in a story could give us insight as to how the model is storing information from a story. We would for example expect that the locations and entities being mentioned in a sentence would be activated when processing that sentence. Figure 4 presents a visualization of the gate activations of memory cells of the memory updater at each timestep of a selected story. The X-axis lists the sentences at every timestep, and the Y-axis represents the memory cells and their respective keys. The color intensity of each shaded cell corresponds to its gate activation value (darker = higher weight). Observe that the model correctly activates the gates of the entities in the sentence as expected (green boxes in Figure 4). However, the model also activates unrelated entities (red boxes in Figure 4). We remark that this provides a potential explanation for the negative effects observed on introducing distracting sentences, and that this merits further exploration in future work.

7.3 Incorrect predictions

Table 7 shows a case where the Learned Textual Time Travel model makes different predictions about the location of an object ("spinach") depending on how the question is phrased. In particular the model seems to be tracking the location of the entity "Ella" rather than the spinach. It appears that the model has learned a surface-level association between the question type "What is the location of <entity>?" and the target entity type being tracked, which is likely a result of how we train the model. While this issue does not impact our model's results on the ToMi dataset, it does show that these neural models are easily affected by dataset biases, and care must be taken to ensure that they learn associations that are useful for answering ToM questions in a target test environment.

7.4 Convoluted examples

For some cases in the dataset, the model must have additional knowledge to answer the question cor-

1. Jacob entered the porch.
2. Lucas entered the porch.
3. Hunter entered the office.
4. The eggplant is in the red-suitcase.
5. Lucas exited the porch.
6. Lucas entered the office.
7. Jacob moved the eggplant to the green-bottle.
8. Where will Lucas look for the eggplant?

Table 8: Some stories require the model to connect several sentences in order to resolve the entity location and identify the false belief.

rectly. For example, in Table 8, it is unclear at first whether the red-suitcase containing the eggplant is in the office or the porch. Resolving its location requires the observation that if Jacob moved the eggplant, then its original container, i.e. the red-suitcase, should be at the same location as Jacob. This example, and others like it, are not exclusively testing the ToM capabilities of the model, as they require the model to understand spatial relationships, perform pragmatic reasoning and show common sense.

8 Conclusion

ToM is an important capability that NLP systems need to acquire in order to have human-like reasoning abilities. Understanding and predicting the mental states of others will help in comprehending their intentions and needs, and thereby generate better responses in interactive systems like dialogue agents. In this paper, we attempt to improve the ToM abilities of memory-augmented neural models by building priors about the entities' minds and performing textual time travel (i.e., retrieving the metal states of entities from earlier timesteps). We find that our Heuristic and Learned Textual Time Travel approaches improve performance, particularly on false belief tasks.

Starting from synthetic datasets like ToMi is necessary because they allow the development and testing of new techniques in controlled environments. These datasets act as prerequisites for new models to pass, and models that fail on these are unlikely to scale to real world data. In a naturalistic setting like QA and dialogue, it is much harder to find instances involving false beliefs automatically. More importantly, identifying and controlling for confounding variables in naturalistic data could be more difficult. For example, issues such as recency or lexical overlap might result in Clever Hans phenomena as shown for NLI (McCoy et al., 2019). Demonstrating that our approach works for ToMi is a first step towards building models with complete ToM capabilities. Theory of Mind is a complex problem and we believe that we can make progress by gradually increasing the complexities of the ToM tasks in a controlled setting. This work is a part of a bottom-up process for solving ToM. To this end, our approach adds the missing piece of incorporating mental-state tracking along the time axis. With this prerequisite met and barrier crossed, we can move towards tackling other challenges in ToM in the future.

Acknowledgements

This work is supported by the Canada CIFAR AI Chair program and the Natural Sciences and Engineering Research Council of Canada. We are grateful to Kristine Onishi for the initial discussions.

References

Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Martin Brüne and Ute Brüne-Cohrs. 2006. Theory of mind—evolution, ontogeny, brain mechanisms and psychopathology. *Neuroscience & Biobehavioral Reviews*, 30(4):437–455.

John H Flavell. 2004. Theory-of-mind development: Retrospect and prospect. *Merrill-Palmer Quarterly (1982-)*, pages 274–290.

Erin Grant, Aida Nematzadeh, and Thomas L. Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, CogSci 2017, London, UK, 16-29 July 2017*. cognitivesciencesociety.org.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world

state with recurrent entity networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Cecilia M Heyes. 1998. Theory of mind in nonhuman primates. *Behavioral and brain sciences*, 21(1):101–114.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Alan M Leslie and Uta Frith. 1988. Autistic children's understanding of seeing, knowing and believing. *British Journal of Developmental Psychology*, 6(4):315–324.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Josef Perner and Heinz Wimmer. 1985. "john thinks that mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3):437–471.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4967–4976.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2440–2448.

Henry M Wellman. 2014. *Making minds: How theory of mind develops*. Oxford University Press.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Jessica Yott and Diane Poulin-Dubois. 2016. Are infants' theory-of-mind abilities well integrated? implicit understanding of intentions, desires, and beliefs. *Journal of Cognition and Development*, 17(5):683–698.

A The ToMi Dataset (and Fixed Dataset)

We found irregularities in some second-order questions in the ToMi dataset. Tables 9 and 10 show examples of these cases.

1. Liam entered the master-bedroom.
2. Chloe entered the master-bedroom.
3. Hunter entered the master-bedroom.
4. Chloe exited the master-bedroom.
5. The pineapple is in the green-cupboard.
6. Liam exited the master-bedroom.
7. Hunter moved the pineapple to the blue-pantry.
8. Where does Liam think that Hunter searches for the pineapple?
Given Answer: blue-pantry
Correct Answer: green-cupboard

Table 9: Liam exits the master-bedroom before the move. He is not aware of the final location of the pineapple. So, the answer should be green-cupboard.

Several false belief questions were incorrectly classified as true belief, leading to an unexpectedly large count for the true belief questions. The code to generate this dataset maintains an oracle and a

1. Aria likes the melon.
2. Aria entered the pantry.
3. Oliver entered the pantry.
4. Noah entered the pantry.
5. The melon is in the blue-bathtub.
6. Noah exited the pantry.
7. Oliver exited the pantry.
8. Aria moved the melon to the red-drawer.
9. Noah entered the kitchen.
10. Where does Noah think that Aria searches for the melon?
Given Answer: red-drawer
Correct Answer: blue-bathtub

Table 10: Noah exits the pantry before Aria moves the melon. Noah reenters a different location, so, he is still unaware of the final location of the melon. The answer should be blue-bathtub.

map of direct and indirect beliefs. In two particular scenarios, the oracle was not updated to reflect the entity beliefs:

1. If the agent exits before the move.
2. If the agent enters a different location.

We corrected these instances and re-generated the dataset with 1000 stories.

The dataset contains templates in English for each of the "entry", "exit", "move" and "noise" actions. The train, validation and test sets contain 5994 questions each. Table 11 shows the distribution of the questions.

Belief	First-order	Second-order
True belief	1571	958
False belief	424	1036

Table 11: Distribution of questions in test dataset.

B Training details

We train our model using a GTX 1080Ti GPU. Our hyperparameter search includes the following ranges, which were chosen manually.

- `nhop`: $\{1, 3, 5\}$
- `batchsize`: $\{8, 16, 32\}$
- `memslots`: $\{10\}$
- `sdt`: $\{0.01, 0.001\}$
- `runs`: $\{1\}$
- `embedding`: $\{icmul, bow, GloVe\}$

The training time for EntNet, Heuristic Time Travel and Learned Time Travel models was about 120 minutes each. The models were trained for 200 epochs each and has 87147 parameters.