Published in NAACL'2021

Time-Stamped Language Model: Teaching Language Models to Understand the Flow of Events

Hossein Rajaby Faghihi
Michigan State University
rajabyfa@msu.edu

Parisa Kordjamshidi
Michigan State University
kordjams@msu.edu

Abstract

Tracking entities throughout a procedure described in a text is challenging due to the dynamic nature of the world described in the process. Firstly, we propose to formulate this task as a question answering problem. This enables us to use pre-trained transformer-based language models on other QA benchmarks by adapting those to the procedural text understanding. Secondly, since the transformer-based language models cannot encode the flow of events by themselves, we propose a Time-Stamped Language Model (TSLM model) to encode event information in LMs architecture by introducing the timestamp encoding. Our model evaluated on the Propara dataset shows improvements on the published state-of-the-art results with a 3.1% increase in F1 score. Moreover, our model yields better results on the location prediction task on the NPN-Cooking dataset. This result indicates that our approach is effective for procedural text understanding in general.

1 Introduction

A procedural text such as a recipe or an instruction usually describes the interaction between multiple entities and their attribute changes at each step of a process. For example, the photosynthesis procedure can contain steps such as 1. *Roots absorb water from soil*; 2. *The water flows to the leaf*; 3. *Light from the sun and CO2 enter the leaf*; 4. *The water, light, and CO2 combine into a mixture*; 5. *Mixture forms sugar*. Procedural text understanding is a machine reading comprehension task defined on procedural texts. Answering questions such as "what is the location of the mixture at step 4", in the above example, requires tracking entities' interactions to predict their attributes at each step (Dalvi et al., 2018; Bosselut et al., 2018). This is quite challenging due to the dynamic nature of the entities' attributes in the context.

Transformer-based language models have shown promising results on multi-hop or single-hop question answering benchmarks such as HotpotQA (Yang et al., 2018), SQuAD (Rajpurkar et al., 2016), and Drop (Dua et al., 2019). However, it is hard to expect LMs to understand the flow of events and pay attention to the time in the procedure (e.g., step 4) without extra modeling efforts.

In recent research, different approaches are taken to address procedural reasoning based on language models using QA formulations. Following the intuition that attributes of entities can be retrieved based on the current and previous steps, DynaPro (Amini et al., 2020) modifies the input to only contain those sentences in the input at each time. This will provide a different input to the model based on each question to help it detect changes after adding each step. KG-MRC (Das et al., 2018) also generates a dynamic knowledge graph at each step to answer the questions. However, this intuition is contradicted in some scenarios such as detecting inputs of the process. For instance, the answer to the question "Where is light as step 0?" is "Sun", even if it is not mentioned in the first sentence of the process. Inputs are entities that are not created in the process.

The architecture of the QA transformer-based LMs is very similar to the traditional attention mechanism. Other methods such as ProLocal (Dalvi et al., 2018) and ProGlobal (Dalvi et al., 2018) have structured this task by finding the attention of each entity to the text at each step. To be sensitive to the changes at each step, ProLocal manually changes the model's input by removing all steps except the one related to the question. ProGlobal computes attention to the whole context while adding a distance value. Distance value is computed for each token based on its distance to the direct mention of the entity at each step.

The current language models convey rich linguistic knowledge and can serve as a strong basis for

solving various NLP tasks (Liu et al., 2019; Devlin et al., 2019; Yang et al., 2019). That is why most of the state-of-the-art models on procedural reasoning are also built based on current language models (Amini et al., 2020; Gupta and Durrett, 2019). Following the same idea, we investigate the challenges that current models are facing for dealing with procedural text and propose a new approach for feeding the procedural information into LMs in a way that the LM-based QA models are aware of the taken steps and can answer the questions related to each specific step in the procedure.

We propose the Time-Stamped Language model (TSLM model), which uses timestamp embedding to encode past, current, and future time of events as a part of the input to the model. TSLM utilizes timestamp embedding to answer differently to the same question and context based on different steps of the process. As we do not change the portion of the input manually, our approach enables us to benefit from the pre-trained LMs on other QA benchmarks by using their parameters to initialize our model and adapt their architecture by introducing a new embedding type. Here, we use RoBERTa (Liu et al., 2019) as our baseline language model.

We evaluate our model on two benchmarks, Propara (Dalvi et al., 2018) and NPN-Cooking (Bosselut et al., 2018). Propara contains procedural paragraphs describing a series of events with detailed annotations of the entities along with their status and location. NPN-Cooking contains cooking recipes annotated with their ingredients and their changes after each step in criteria such as location, cleanliness, and temperature.

TSLM differs from previous research as its primary focus is on using pre-trained QA models and integrating the flow of events in the global representation of the text rather than manually changing the part of the input fed to the model at each step. TSLM outperforms the state-of-the-art models in nearly all metrics of two different evaluations defined on the Propara dataset. Results show a 3.1% F1 score improvement and a 10.4% improvement in recall. TSLM also achieves the state-of-the-art result on the location accuracy on the NPN-Cooking location change prediction task by a margin of 1.55%. In summary, our contribution is as follows:

- We propose Time-Stamped Language Model (TSLM model) to encode the meaning of past, present, and future steps in processing a procedural text in language models.
- Our proposal enables procedural text understanding models to benefit from pre-trained LM-based QA models on general-domain QA benchmarks.
- TSLM outperforms the state-of-the-art models on the Propara benchmark on both document-level and sentence-level evaluations. TSLM improves the performance state-of-the-art models on the location prediction task of the NPN-Cooking (Bosselut et al., 2018) benchmark.
- Improving over two different procedural text understanding benchmarks suggests that our approach is effective, in general, for solving the problems that require the integration of the flow of events in a process.

2 Problem Definition

An example of a procedural text is shown in Table 1. The example is taken from the Propara (Dalvi et al., 2018) dataset and shows the photosynthesis procedure. At each row, the first column is list of the sentences, each of which forms one step of the procedure. The second column contains the number of the step in the process and the rest are the entities interacting in the process and their location at each step. The location of entities at step 0 is their initial location, which is not affected by this process. If an entity has a known or unknown location (specified by "?") at step 0, we call it an input.

The procedural text understanding task is defined as follows. Given a procedure p containing a list of n sentences $P = \{s_1, ...s_n\}$, an entity e and a time step t_i, we find L, the location of that entity and specify the status S of that entity. Status S is one value in a predefined set $S = \{\text{non-existence, unknown-location, known-location}\}$. location L is a span of text in the procedure that is specified with its beginning and end token. We formulate the task as finding function F that maps each triplet of entity, procedure and time step to a pair of entity location and status: $(S, L) = F(e, P, t_i)$

3 Proposed Procedural Reasoning Model

3.1 QA Setting

To predict the status and the location of entities at each step, we model F with a question answering

		Participants				
Paragraph	State number	Water	Light	CO2	Mixture	Sugar
(Before the process starts)	State 0	Soil	Sun	?	-	-
Roots absorb water from soil	State 1	Root	Sun	?	-	-
The water flows to the leaf	State 2	Leaf	Sun	?	-	-
Light from the sun and CO2 enter the leaf	State 3	Leaf	Leaf	Leaf	-	-
The water, light, and CO2 combine into a mixture	State 4	-	-	-	Leaf	-
Mixture forms sugar	State 5	-	-	-	-	Leaf

Table 1: An example of procedural text and its annotations from the Propara dataset (Dalvi et al., 2018). "-" means entity does not exist. "?" means the location of entity is unknown.

setting. For each entity e, we form the input Q_e as follows:

$$Q_e = \text{[CLS] Where is } e\text{? [SEP] } s_1 \text{ [SEP] } s_2 \text{ [SEP] ..., } s_n \text{ [SEP]} \quad (1)$$

Although Q_e is not a step-dependent representation and does not incorporate any different information for each step, our mapping function needs to generate different answers for the question "Where is entity e?" based on each step of the procedure. For instance, consider the example in Table 1 and the question "where is water?", our model should generate different answers at four different steps. The answer will be “root”, “leaf”, “leaf”, “non-existence” for steps 1 to 4, respectively.

To model this, we create pairs of (Q_e, t_i) for each $i \in \{0, 1, ..., n\}$. For each pair, Q_e is timestamped according to t_i using $Timestamp(.)$ function described in Sec. 3.2 and mapped to an updated step-dependent representation, $Q_e^{t_i} = Timestamp(Q_e, t_i)$.

The updated input representation is fed to a language model (here ROBERTA) to obtain the step-dependent entity representation, $R_e^{t_i}$, as shown in Equation 2. We discuss the special case of $i = 0$ in more details in Sec. 3.2.

$$R_e^{t_i} = RoBERTa(Q_e^{t_i}) \quad (2)$$

We use the step-dependent entity representation, $R_e^{t_i}$, and forward it to another mapping function $g(.)$ to obtain the location and status of the entity e in the output. In particular the output includes the following three vectors, a vector representing the predictions of entity status S, another vector for each token's probability of being the start of the location span L, and a third vector carrying the probability of each word being the last token of the location span. The outputs of the model are computed according to the Equation 3.

$$(status, Start\_prob, End\_prob) = g(R_e^{t_i}) \quad (3)$$

where R_e is the tokens' representations output of RoBERTa (Liu et al., 2019), and $g(.)$ is a function we apply on the token representations to get the final predictions. We will discuss each part of the model separately in the following sections.

3.2 Timestamp Embedding

The timestamp embedding adds the step information to the input Q_e to be considered in the attention mechanism. The step attention is designed to distinguish between current (what is happening now), past (what has happened before), and future (what has not yet happened) information.

We use the mapping function $Timestamp(.)$ from the pair (Q_e, t_i) to add a number along with each token in Q_e and retrieve the step-dependent input $Q_e^{t_i}$ as shown in Figure 1. The Mapping function $Timestamp(.)$ integrates past, current, and future representations to all of the tokens related to each part. $Timestamp(.)$ function assigns number 1 for past, 2 for current, and 3 for future tokens in the paragraph by considering one step of the process as the current event. These values are used to compute an embedding vector for each token, which will be added to its initial representation as shown in Figure 2. The special number 0 is assigned to the question tokens, which are not part of the process timeline. For predicting State 0 (The inputs of the process), we set all the paragraph information as the current step.

3.3 Status classification

To predict the entities' status, we apply a linear classification module on top of the $[CLS]$ token representation in R_e as shown in Equation 4.

$$Attribute = Softmax(W^T Re_{[C]}) \quad (4)$$

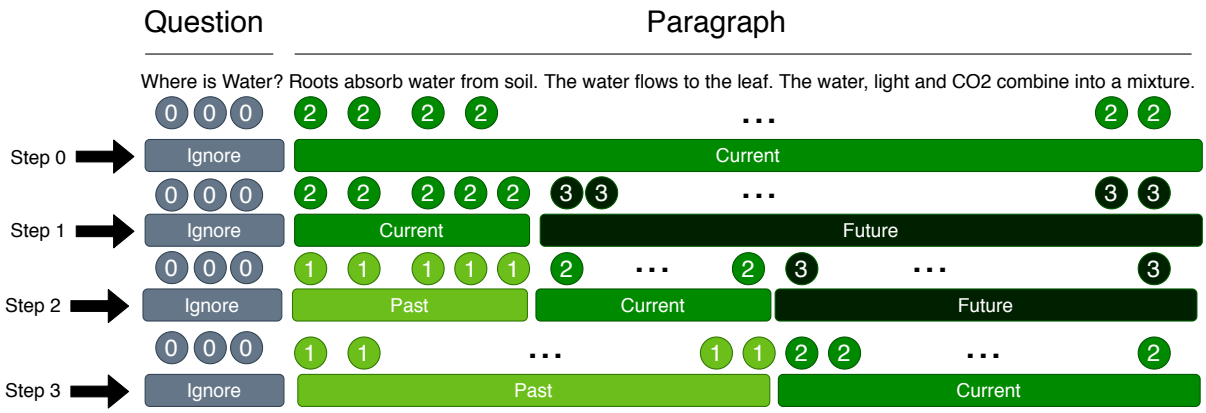

Figure 1: An example of timestamp embedding in a procedural text. The question is always ignored with value "0". At each step i, the tokens from that step are paired with "current" value, tokens from steps 0 to i are paired with "past" value, and the tokens from step i to last step are paired with the "future" value.

where $Re_{[C]}$ is the representation of the $[CLS]$ token which is the first token in R_e.

3.4 Span prediction

We predict a location span for each entity for each step of the process as shown in Equation 5, we follow the popular approach of selecting start/end tokens to detect a span of the text as the final answer. We compute the probability of each token being the start or the end of the answer span. If the index with the highest probability to be the start token is $token_{start}$ and for the end token is $token_{end}$, the answer location will be $Location = P[token_{start} : token_{end}]$.

$$\begin{aligned} \text{Start\_prob} &= \text{Softmax}(W_{\text{start}}^T R_e^{t_i}) \\ \text{End\_prob} &= \text{Softmax}(W_{\text{end}}^T R_e^{t_i}) \\ \text{token}_{\text{start}} &= \arg\max_i(\text{Start\_prob}) \\ \text{token}_{\text{end}} &= \arg\max_i(\text{End\_prob}) \end{aligned} \quad (5)$$

3.5 Training

We use the cross-entropy loss function to train the model. At each prediction for entity e at timestamp t_i, we compute one loss value $loss_{attribute}$ regarding the status prediction and one loss value $loss_{location}$ for the span selection. The variable $loss_{location}$ is the summation of the losses of the start token and the end token prediction, $loss_{location} = loss_{location_{start}} + loss_{location_{end}}$. The final loss of entity e at time t_i is computed as in Equation 6.

$$Loss_i^e = loss_{(i,attribute)}^e + loss_{(i,location)}^e \quad (6)$$

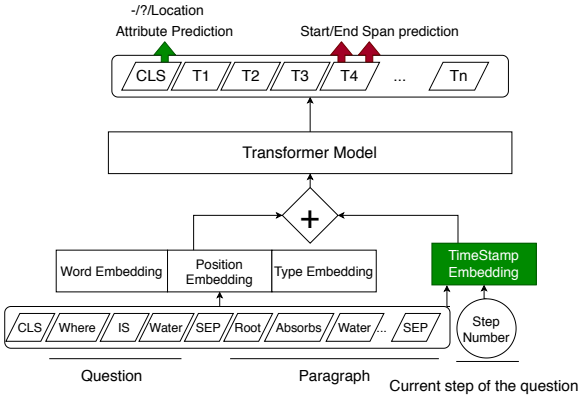

Figure 2: An overview of the proposed model. The "Timestamp Embedding" module is introduced in this work and the rest of the modules are taken from basic language model architecture.

3.6 Inference

At inference time, we apply two different post-processing rules on the outputs of the model. First, we impose that the final selected location answer should be a noun phrase in the original procedure. Considering that a location span is a noun phrase, we limit the model to do a $softmax$ over tokens of noun phrases in the paragraph to select the start and end tokens. Second, we apply consistency rules to make sure that our predicted status of entities are consistent. We define the two following rules:

- An entity can not be created if it has been already destroyed : *if $S_e^{t_i}$ is "non-existence" and $S_e^{t_{i+1}}$ is unknown or known location, then for every step j, if $S_e^{t_j}$ is unknown or known location and $S_e^{t_{j+1}}$ is "non-existence", then $i < j$.*

- An entity cannot be created/destroyed twice in a process: *if $S_e^{t_j}$ and $S_e^{t_i}$ are both "-", $S_e^{t_{j+1}}$ and $S^{t_{i+1}}$ are both either known or unknown location, then $i = j$.*

$S_e^{t_i}$ is the status of entity e at step t_i of the process.

We do not apply an optimization/search algorithm to find the best assignment over the predictions according to the defined constraints. The constraints are only applied based on the order of the steps to ensure that the later predictions are consistent with the ones made before.

4 Experiments

4.1 Datasets

Propara (Dalvi et al., 2018): This dataset was created as a benchmark for procedural text understanding to track entities at each step of a process. Propara contains 488 paragraphs and 3,300 sentences with annotations that are provided by crowdworkers. The annotations (81,000) are the location of entities at each step of the process. The location can be either the name of the location, unknown location, or specified as non-existence.

NPN-Cooking (Bosselut et al., 2018): This is a benchmark containing textual cooking instructions. Annotators have specified ingredients of these recipes and explained the recipe using different changes happening on each ingredient at each step of the instructions. These changes are reported in categories such as location, temperature, cleanliness, and shape. We evaluate our model on the location prediction task of this benchmark, which is the hardest task due to having more than 260 candidate answers. We do not use the candidates to find the locations in our setting; Instead, we find a span of the text as the final location answer. This is a relatively harder setting but more flexible and generalizable than the classification setting.

4.2 Implementation Details

We use SGD optimizer implemented by Pytorch (Paszke et al., 2017) to update the model parameters. The learning rate for the Propara implementation is set to $3 - e4$ and is updated by a scheduler with a 0.5 coefficient every 50 steps. We use $1 - e6$ as the learning rate and a scheduler with 0.5 coefficient to update the parameters every ten steps on the NPN-Cooking implementation. The implementation code is publicly available at GitHub[1].

Step	Entity	Action	Before	After
1	Water	Move	Root	Leaf
2	Water	Destroy	Leaf	-
1	Sugar	Create	-	Leaf
2	Sugar	None	Leaf	Leaf

Table 2: A sample table to evaluate the Propara document-level task.

We use RoBERTa (Liu et al., 2019) question answering architecture provided by HuggingFace (Wolf et al., 2019). RoBERTa is pretrained with SQuAD (Rajpurkar et al., 2016) and used as our base language model to compute the token representations. Our model executes batches containing an entity at every step and makes updates based on the average loss of entities per procedure. The network parameters are updated after executing one whole example. The implementation code will be publicly available on GitHub after acceptance.

4.3 Evaluation

Sentence-level evaluation is introduced in (Dalvi et al., 2018) for Propara dataset. This evaluation focuses on the following three categories.

- **Cat1** Is e created (destroyed/moved) during the process?
- **Cat2** When is e created (destroyed/moved) during the process?
- **Cat3** Where is e created (destroyed/moved from or to) during the process?

Document-level evaluation is a more comprehensive evaluation process and introduced later in (Tandon et al., 2018) for Propara benchmark. Currently, this is the default evaluation in the Propara leaderboard containing four criteria:

- **What are the Inputs?** Which entities existed before the process began and do not exist after the process ends.
- **What are the Outputs?** Which entities got created during the process?
- **What are the Conversions?** Which entities got converted to other entities?
- **What are the Moves?** Which entities moved from one location to another?

The document-level evaluation requires models to reformat their predictions in a tabular format as shown in Table 2. At each row of this table, for each entity at a specific step, we can see the action

[1] https://github.com/HLR/TSLM

	Sentence-level					Document-level		
Model	Cat1	Cat2	Cat3	Macro-Avg	Micro-Avg	P	R	F1
ProLocal (Dalvi et al., 2018)	62.7	30.5	10.4	34.5	34.0	**77.4**	22.9	35.3
ProGlobal (Dalvi et al., 2018)	63.0	36.4	35.9	45.1	45.4	46.7	52.4	49.4
EntNet (Henaff et al., 2017)	51.6	18.8	7.8	26.1	26.0	50.2	33.5	40.2
QRN (Seo et al., 2017)	52.4	15.5	10.9	26.3	26.5	55.5	31.3	40.0
KG-MRC (Das et al., 2018)	62.9	40.0	38.2	47.0	46.6	64.5	50.7	56.8
NCET (Gupta and Durrett, 2019)	73.7	47.1	41.0	53.9	54.0	67.1	58.5	62.5
XPAD (Dalvi et al., 2019)	-	-	-	-	-	70.5	45.3	55.2
ProStruct (Tandon et al., 2018)	-	-	-	-	-	74.3	43.0	54.5
DYNAPRO (Amini et al., 2020)	72.4	49.3	**44.5**	55.4	55.5	75.2	58.0	65.5
TSLM (Our Model)	**78.81**	**56.798**	40.9	**58.83**	**58.37**	68.4	**68.9**	**68.6**

Table 3: Results from sentence-level and document-level evaluation On Propara. Ci evaluations are defined Section 4.3.

applied on that entity, the location of that entity before that step, and the location of the entity after that step. Action takes values from a predefined set including, "None", "Create", "Move", and "Destroy". The exact action can be specified based on the before and after locations.

We have to process our (Status S, Location L) predictions at each step to generate a similar tabular format as in Table 2. We define r_e^i as a row in this table which stores the predictions related to entity e at step t_i. To fill this row, we first process the status predictions. If the status prediction S is either "-" or "?", we fill those values directly in the after location column. The before location column value of r_e^i is always equal to the after location column value of r_e^{i-1}. If the status is predicted to be a "Known Location", we fill the predicted location span L into the after location column of r_e^i.

The action column is filled based on the data provided in before and after locations columns. If the before location is/isn't "-" and after location is not/is "-", then the action is "Create"/"Destroy". If the before and after locations are equal, then the action is "None" and if the before and after locations are both spans and are different from each other, the action is "Move".

NPN-Cooking location change: We evaluate our model on the NPN-Cooking benchmark by computing the accuracy of the predicted locations at steps where the locations of ingredients change. We use the portion of the data that has been annotated by the location changes to train and evaluate our model. In this evaluation, we do not use the status prediction part of our proposed TSLM model. Since training our model on the whole training set takes a very long time (around 20 hours per iteration), we use a reduced number of samples for training. This is a practice that is also used in other prior work (Das et al., 2018).

4.4 Results

The performance of our model on Propara dataset (Dalvi et al., 2018) is quantified in Table 3. Results show that our model improves the SOTA by a 3.1% margin in the F1 score and improves the Recall metric with 10.4% on the document-level evaluation. On the sentence-level evaluation, we outperform SOTA models with a 5.11% in Cat1, and 7.49% in Cat2 and by a 3.4% margin in the macro-average. We report Table 3 without considering the consistency rules and evaluate the effect of those in the ablation study in Sec. 4.5.

In Table 5, we report a more detailed quantified analysis of TSLM model's performance based on each different criteria defined in the document-level evaluation. Table 5 shows that our model performs best on detecting the procedure's outputs and performs worst on detecting the moves. Detecting moves is essentially hard for TSLM as it is predicting outputs based on the whole paragraph at once. Outperforming SOTA results on the input and output detection suggests that TSLM model can understand the interactions between entities and detect the entities which exist before the process begins. The detection of input entities is one of the weak aspects of the previous research that we improve here.

A recent unpublished research (Zhang et al., 2021) reports better results than our model. However, their primary focus is on common-sense rea-

Model	Accuracy	Training Samples	Prediction task
NPN-cooking (Bosselut et al., 2018)	51.3	$\sim 83,000$ (all data)	Classification
KG-MRC (Das et al., 2018)	51.6	$\sim 10,000$	Span Prediction
DynaPro (Amini et al., 2020)	62.9	$\sim 83,000$ (all data)	Classification
TSLM (Our Model)	63.73	$\sim 10,000$	Span Prediction
	64.45	$\sim 15,000$	Span Prediction

Table 4: Results on the NPN-Cooking benchmark. Both class prediction and span prediction tasks are the same but use two different settings, one selects among candidates, and the other chooses a span from the recipe. However, each model has used a different setting and a different portion of the training data. The information of the data splits was not available that makes a fair comparison hard.

Criteria	Precision	Recall	F1
Inputs	89.8	71.3	79.5
Outputs	85.6	91.4	88.4
Conversions	57.7	56.7	57.2
Moves	40.5	56	47

Table 5: Detailed analysis of TSLM performance on the Propara test set on four criteria defined in the document-level evaluation.

soning and their goal is orthogonal to our main focus in proposing TSLM model. Such approaches can be later integrated with TSLM to benefit from common-sense knowledge on solving the Propara dataset.

The reason that TSLM performs better at *recall* and worse at *precision* is that our model looks at the global context, which increases the recall and lowers the precision when local information is strongly important. The same phenomenon (better recall) is observed in ProGlobal, which also considers global information as we do, compared to ProLocal.

Table 4 shows our results on the NPN-Cooking benchmark for the location prediction task. Results are computed by only considering the steps that contain a location change and are reported by computing the accuracy of predicting those changes. Our results show that TSLM outperforms the SOTA models with a 1.55% margin on accuracy even after training on 15,000 training samples. To be comparable with the KG-MRC (Das et al., 2018) experiment on NPN-Cooking which is only trained on 10k samples, we report the performance of our model trained on the same number of samples, where TSLM gets a 12.1% improvement over the performance of KG-MRC (Das et al., 2018).

4.5 Ablation Study

To evaluate the importance of each module one at a time, we report the performance of the TSLM by removing the noun-phrase filtering at inference, the consistency rules, timestamp embedding, SQuAD (Rajpurkar et al., 2016) pre-training, and by replacing RoBERTa (Liu et al., 2019) with BERT (Devlin et al., 2019). These variations are evaluated on the development set of the Propara dataset and reported in Table 6. As stated before and shown in Table 6, it is impossible to remove the timestamp embedding as that is the only part of the model enabling changes in the answer at each step. Hence, by removing that, the model cannot converge and yields a 25% decrease on the F1 score. The simple consistency and span filtering rules are relatively easy to be learned by the model based on the available data, therefore adding those does not affect the final performance of the model.

TSLM$_{\text{BERT}}$ experiment is designed to ensure a fair comparison with previous research (Amini et al., 2020) which has used BERT as their base language model. The comparison of *TSLM$_{BERT}$* to *-SQuAD Pre-training* and *- Timestamp Embedding* in Table 6 indicates that using RoBERTa instead of BERT is not as much important as our main proposal (using Time-stamp encoding) in TSLM model. Also, TSLM$_{\text{BERT}}$ achieves 66.7% F1 score on the Propara test set, which is 1.2% better than the current SOTA performance.

By removing the SQuAD pre-training phase, the model performance drops with a 10.6% in the F1 score. This indicates that despite the difference between the procedural text understanding and the general MRC tasks, it is quite beneficial to design methods that can transfer knowledge from other QA data sources to help with procedural reasoning. This is crucial as annotating procedural texts is relatively more expensive and time-consuming.

Model	P	R	F1
$\text{TSLM}_{\text{RoBERTa}}$	72.9	74.1	**73.5**
- constraints	73.8	73.3	**73.5**
- noun-phrase filtering	73.5	73.3	73.4
- SQuAD Pre-training	78.8	52.2	62.8
- Timestamp Embedding	94.6	32.6	48.5
$\text{TSLM}_{\text{BERT}}$	69.2	73.5	71.3

Table 6: Ablation study results on the development set of the Propara document-level task. "- constraints", "- Span filtering", and "- Timestamp Encoding" shows our model performance while removing those modules. -*SQuAD Pre-training* is when we do not pre-train our base language model on SQuAD. $\text{TSLM}_{\text{BERT}}$ is when we use BERT as the base language model.

5 Discussion

We provide more samples to support our hypothesis in solving the procedural reasoning task and answer some of the main questions about the ideas presented in TSLM model.

Why is the whole context important? The main intuition behind TSLM is that the whole context, not just previous information, matters in reasoning over a process. Here, we provide some samples from Propara to show why this intuition is correct. Consider this partial paragraph, "Step i: With enough time the pressure builds up greatly. Step $i+1$: The resulting volcano may explode.". Looking at the annotated status and location, the "volcano" is being created at Step i without even being mentioned in that step. This is only detectable if we look at the next step saying "The resulting Volcano...".

As another example, consider this partial paragraph: "Step i: Dead plants form layers called peat. ... Step $i+3$: Pressure squeezes water out of the peat.". The annotation indicates that the location of "water" is being changed to "peat" at step i, which is only possible to detect if the model is aware of the following steps indicating that the water comes out of the peat.

Positional Embedding VS Time-stamp encoding: As mentioned before the whole context (future and past events) is essential for procedural reasoning at a specific step. However, the reasoning should focus on one step at a time, given the whole context. While positional encoding encodes the order of information at the token-level for reasoning over the entire text, we need another level of encoding to specify the steps' positions (boundaries) and, more importantly, to indicate the step that the model should focus on when answering a question.

Advantages/Disadvantages of TSLM model: TSLM integrates higher-level information into the token representations. This higher-level information can come from event-sequence (time of events), sentence-level, or any other higher source than the token-level information. The first advantage of TSLM is that it enables designing a model which is aware of the whole context, while previous methods had to customize the input at each step to only contain the information of earlier steps. Furthermore, using TSLM enables us to use pretrained QA models on other datasets without requiring us to retrain them with the added time-stamped encoding. One main disadvantage of TSLM model, which is natural due to the larger context setting in this model, is not being sensitive to local changes, which is consistent with the observation in the comparison between ProGlobal and ProLocal models.

6 Related Works

ScoNe (Long et al., 2016), NPN-Cooking (Bosselut et al., 2018), bAbI (Weston et al., 2015), ProcessBank (Berant et al., 2014), and Propara (Dalvi et al., 2018) are benchmarks proposed to evaluate models on procedural text understanding. Processbank (Berant et al., 2014) contains procedural paragraphs mainly concentrated on extracting arguments and relations for the events rather than tracking the states of entities. ScoNe (Long et al., 2016) aims to handle co-reference in a procedural text expressed about a simulated environment. bAbI (Weston et al., 2015) is a simpler machine-generated textual dataset containing multiple procedural tasks such as motion tracking, which has encouraged the community to develop neural network models supporting explicit modeling of memories (Sukhbaatar et al., 2015; Santoro et al., 2018) and gated recurrent models (Cho et al., 2014; Henaff et al., 2017). NPN-Cooking (Bosselut et al., 2018) contains recipes annotated with the state changes of ingredients on criteria such as location, temperature, and composition. Propara (Dalvi et al., 2018) provides procedural paragraphs and detailed annotations of entity locations and the status of their existence at each step of a process.

Inspired by Propara and NPN-Cooking benchmarks, recent research has focused on tracking entities in a procedural text. Query Reduction Net-

works (QRN) (Seo et al., 2017) performs gated propagation of a hidden state vector at each step. Neural Process Network (NPN) (Bosselut et al., 2018) computes the state changes at each step by looking at the predicted actions and involved entities. Prolocal (Dalvi et al., 2018) predicts locations and status changes locally based on each sentence and then globally propagates the predictions using a persistence rule. Proglobal (Dalvi et al., 2018) predicts the status changes and locations over the whole paragraph using distance values at each step and predicts current status based on current representation and the predictions of the previous step. ProStruct (Tandon et al., 2018) aims to integrate manually extracted rules or knowledge-base information on VerbNet (Schuler, 2005) as constraints to inject common-sense into the model. KG-MRC (Das et al., 2018) uses a dynamic knowledge graph of entities over time and predicts locations with spans of the text by utilizing reading comprehension models. Ncet (Gupta and Durrett, 2019) updates entities representation based on each sentence and connects sentences together with an LSTM. To ensure the consistency of predictions, Ncet uses a neural CRF over the changing entity representations. XPAD (Dalvi et al., 2019) is also proposed to make dependency graphs on the Propara dataset to explain the dependencies of events over time. Most recently, DynaPro (Amini et al., 2020) feeds an incremental input to pre-trained LMs' question answering architecture to predict entity status and transitions jointly.

TSLM differs from recent research, as we propose a simple, straightforward, and effective technique to make our model benefit from pre-trained LMs on general MRC tasks and yet enhance their ability to operate on procedural text understanding. We explicitly inject past, current, and future timestamps into the language models input and implicitly train the model to understand the events' flow rather than manually feeding different portions of the context at each step. Procedural reasoning has also been pursued within the multi-modality domain (Yagcioglu et al., 2018; Rajaby Faghihi et al., 2020; Amac et al., 2019) which has additional challenges of aligning the representation spaces of different modalities.

7 Conclusion

We proposed the Time-Stamped Language Model (TSLM model), a novel approach based on a simple and effective idea, which enables pre-trained QA models to process procedural texts and produce different outputs based on each step to track entities and their changes. TSLM utilizes a timestamp function that causes the attention modules in the transformer-based LM architecture to incorporate past, current, and future information by computing a timestamp embedding for each input token. Our experiments show a 3.1% improvement on the F1 score and a 10.4% improvement over the Recall metric on Propara Dataset. Our model further outperforms the state-of-the-art models with a 1.55% margin in the NPN-Cooking dataset accuracy for the location prediction task.

As a future direction, it is worth investigating how common-sense knowledge can be integrated with the TSLM setting by augmenting the process context using external sources of related domain knowledge. We also intend to investigate the effectiveness of our approach on similar tasks on other domains and benchmarks. As another future direction, it can be effective to apply an inference algorithm to impose the global consistency constraints over joint predictions in procedural reasoning instead of using naive post-processing rules.

Acknowledgements

This project is partially funded by National Science Foundation (NSF) CAREER Award #2028626 and the Office of Naval Research (ONR) grant #N00014-20-1-2005.

References

Mustafa Sercan Amac, Semih Yagcioglu, Aykut Erdem, and Erkut Erdem. 2019. Procedural reasoning networks for understanding multimodal procedures. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 441–451.

Aida Amini, Antoine Bosselut, Bhavana Dalvi Mishra, Yejin Choi, and Hannaneh Hajishirzi. 2020. Procedural reading comprehension with attribute-aware context flow. In *Proceedings of the Conference on Automated Knowledge Base Construction (AKBC)*.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510.

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2018. Simulating action dynamics with neural process networks. In *Proceedings of the 6th International Conference for Learning Representations (ICLR)*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1595–1604, New Orleans, Louisiana. Association for Computational Linguistics.

Bhavana Dalvi, Niket Tandon, Antoine Bosselut, Wen-tau Yih, and Peter Clark. 2019. Everything happens for a reason: Discovering the purpose of actions in procedural text. 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 4496–4505, Hong Kong, China. Association for Computational Linguistics.

Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, and Andrew McCallum. 2018. Building dynamic knowledge graphs from text using machine reading comprehension. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Aditya Gupta and Greg Durrett. 2019. Tracking discrete and continuous entity state for process understanding. In *Proceedings of the Third Workshop on Structured Prediction for NLP*, pages 7–12, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Reginald Long, Panupong Pasupat, and Percy Liang. 2016. Simpler context-dependent logical forms via model projections. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1456–1465, Berlin, Germany. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Hossein Rajaby Faghihi, Roshanak Mirzaee, Sudarshan Paliwal, and Parisa Kordjamshidi. 2020. Latent alignment of procedural concepts in multimodal recipes. In *Proceedings of the First Workshop on Advances in Language and Vision Research*, pages 26–31.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. 2018. Relational recurrent neural networks. In *Advances in neural information processing systems*, pages 7299–7310.

Karin Kipper Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Min Joon Seo, Sewon Min, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Query-reduction networks for question answering. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 2440–2448. Curran Associates, Inc.

Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 57–66, Brussels, Belgium. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. Cite arxiv:1502.05698.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. In *EMNLP*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhihan Zhang, Xiubo Geng, Tao Qin, Yunfang Wu, and Daxin Jiang. 2021. Knowledge-aware procedural text understanding with multi-stage training. In *Proceedings of the Web Conference 2021*.