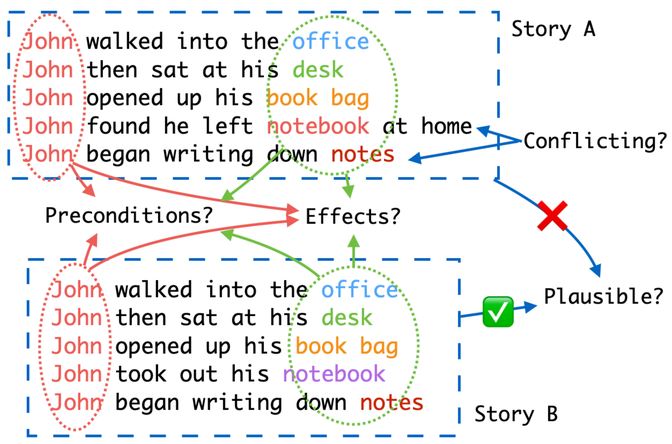
Coalescing Global and Local Information for Procedural Text Understanding

Kaixin Ma[1], Filip Ilievski[2], Jonathan Francis[13], Eric Nyberg[1], Alessandro Oltramari[3]

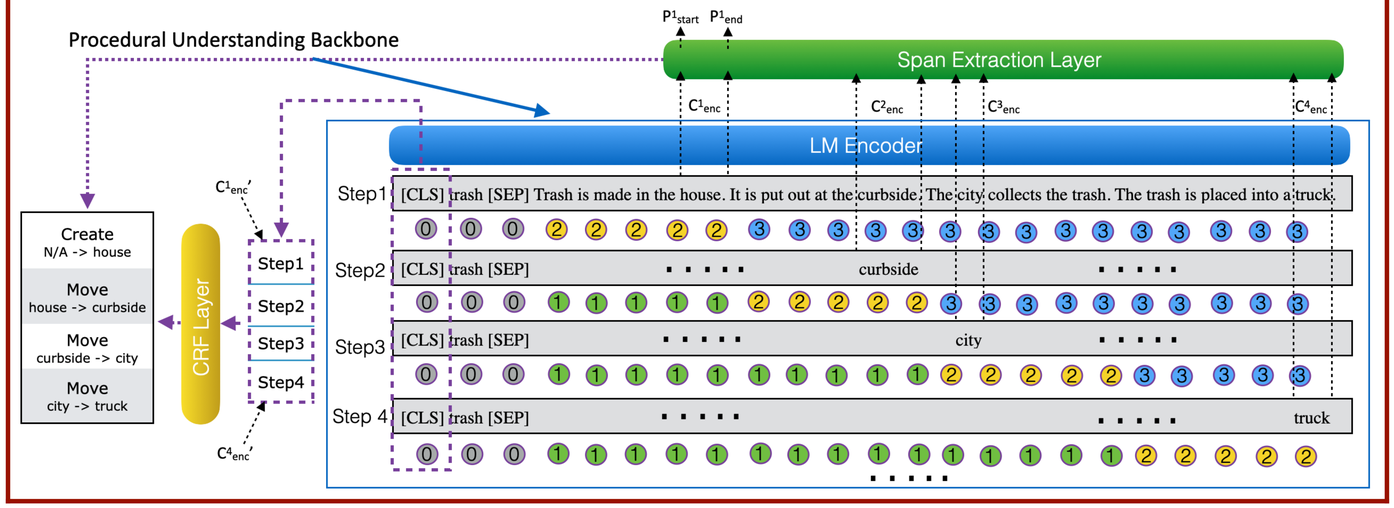
[1]Language Technologies Institute, Carnegie Mellon University
[2]Information Science Institute, University of Southern California
[3]Human Machine Collaboration, Bosch Research Pittsburgh

Introduction & Summary



- Procedural text understanding requires the model to track entity states across timestep, which is an integral part of fine-grained story understanding
- Procedural text modeling requires entity-specific, timestep-aware input representations considering whole context, and dependency of outputs across timesteps.
- We present CGLI, which effectively incorporate all important global and local modeling aspects, leading to SOTA results on two tasks ProPara and TRIP.

CGLI Model



Results & Analysis

Table 1: ProPara test set results. Modeling: E=entity, T=timestep-specific, GC=global context, GO=global outputs.

	Modeling				Sentence-level					Document-level		
Model	E	T	GC	GO	Cat1	Cat2	Cat3	Macro^{avg}	Micro^{avg}	P	R	F1
ProLocal (Dalvi et al., 2018)	Y	Y	N	N	62.7	30.5	10.4	34.5	34.0	**81.7**	36.8	50.7
ProGlobal (Dalvi et al., 2018)	Y	Y	Y	N	63.0	36.4	35.9	45.1	45.4	61.7	48.8	51.9
ProStruct (Tandon et al., 2018)	Y	Y	N	Y	-	-	-	-	-	74.3	43.0	54.5
KG-MRC (Das et al., 2018)	N	Y	N	N	62.9	40.0	38.2	47.0	46.6	64.5	50.7	56.8
NCET (Gupta and Durrett)	N	N	Y	Y	73.7	47.1	41.0	53.9	54.0	67.1	58.5	62.5
IEN (Tang et al., 2020)	N	N	Y	Y	71.8	47.6	40.5	53.3	53.0	69.8	56.3	62.3
DynaPro (Amini et al., 2020)	Y	Y	N	N	72.4	49.3	44.5	55.4	55.5	75.2	58.0	65.5
TSLM (2021)	Y	Y	Y	N	78.8	56.8	40.9	58.8	58.4	68.4	68.9	68.6
KOALA (Zhang et al., 2021)	N	N	Y	Y	78.5	53.3	41.3	57.7	57.5	77.7	64.4	70.4
CGLI (Ours)	Y	Y	Y	Y	80.3	60.5	**48.3**	**63.0**	**62.7**	74.9	**70.0**	72.4
CGLI (Ours) + Data Augmentation	Y	Y	Y	Y	**80.8**	**60.7**	46.8	62.8	62.4	75.7	**70.0**	**72.7**

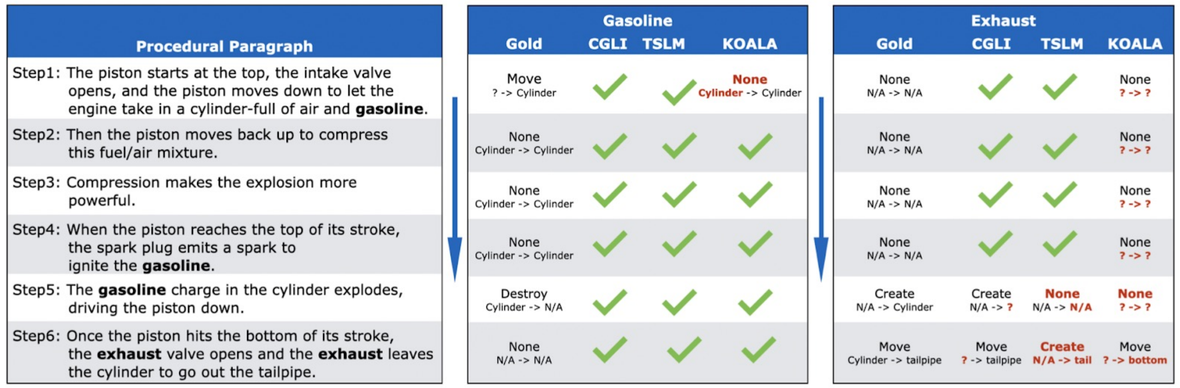


Figure 5: Example predictions on ProPara from three models for two entities. Red font indicate errors.

- CGLI achieves high precision and high recall, by considering all four model aspects
- The gains over the baselines are mainly from the harder-to-answer categories

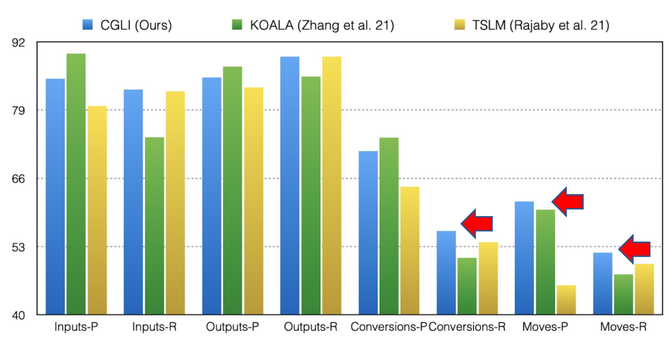


Figure 4: Document-level evaluation on ProPara test set, split by precision (P) and recall (R) per category (Inputs, Outputs, Conversions, Moves).

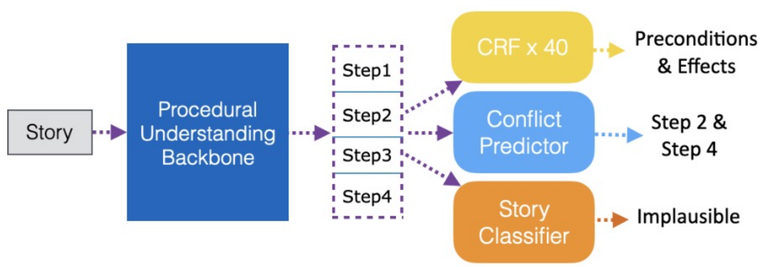


Table 4: Results on the TRIP dataset. The F1 scores of last two columns are Macro averages of 20 attributes.

Model	Accuracy	Consistency	Verifiability	Precondition F1	Effect F1
TRIP-RoBERTa (Storks et al., 2021)	73.2	19.1	9.1	51.3	49.3
CGLI (Ours)	93.4(±1.5)	76.3(±1.7)	24.8(±1.6)	70.8(±1.8)	74.9(±1.7)
CGLI (Ours) No CRF	**94.1**(±0.7)	**77.3**(±1.0)	**28.0**(±2.5)	**72.1**(±1.6)	**75.6**(±1.6)

- CRF may not be helpful for modeling implausible stories
- Future work can look into improving the commonsense ability of the model

Table 5: Error Examples on TRIP. The conflicting pairs are marked with *, and the entity of interest with *italic*.

Ann washed her hair in the bathtub.
Ann used the hair dryer to get ready to go out.
Ann applied deodorant to her armpits.
**Ann* put her pants on.
- (Effects, is wet), Pred: False, Gold: Irrelevant
*Ann ironed her *pants* before going out.
- (Preconditions, is wet), Pred: True, Gold: Irrelevant
*John forgot his *notebook* at home.
- (Effects, location), Pred: Moved, Gold: Irrelevant
John sat at his desk.
John opened up his book bag.
* John took out his *notebook*.
- (Preconditions, location),
- Pred: Picked up, Gold: Taken out of container
John began writing down notes.