Do Transformers understand time?

Shawn Jain, Yonatan Bisk, Hamid Palangi, Jianfeng Gao

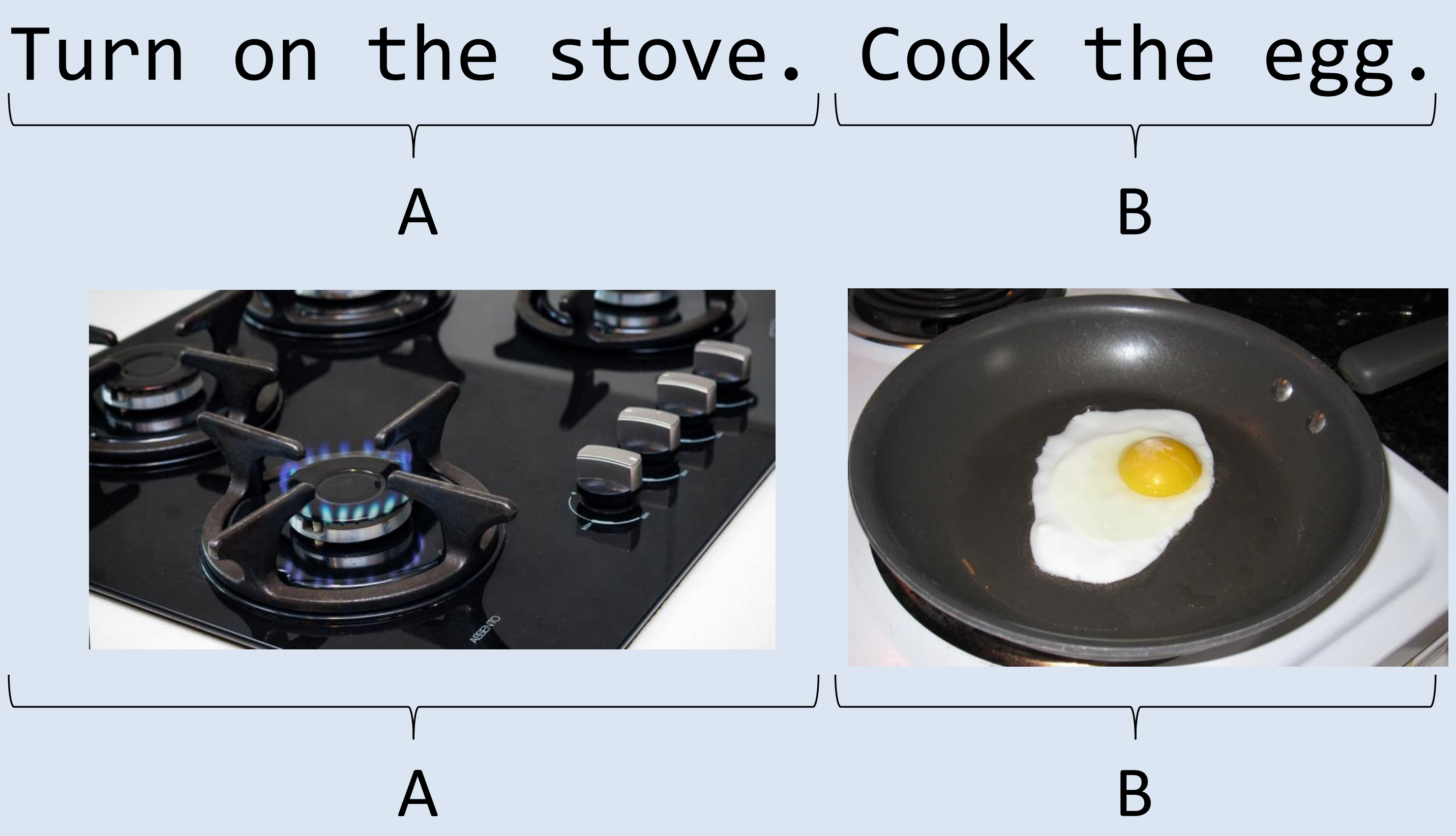
Microsoft Research

jains@microsoft.com

Motivation

Temporal reasoning is critical for AI models.

Can you order these?



`Turn on the stove.` `Cook the egg.`

A B

A B

Easy for humans, hard for models!

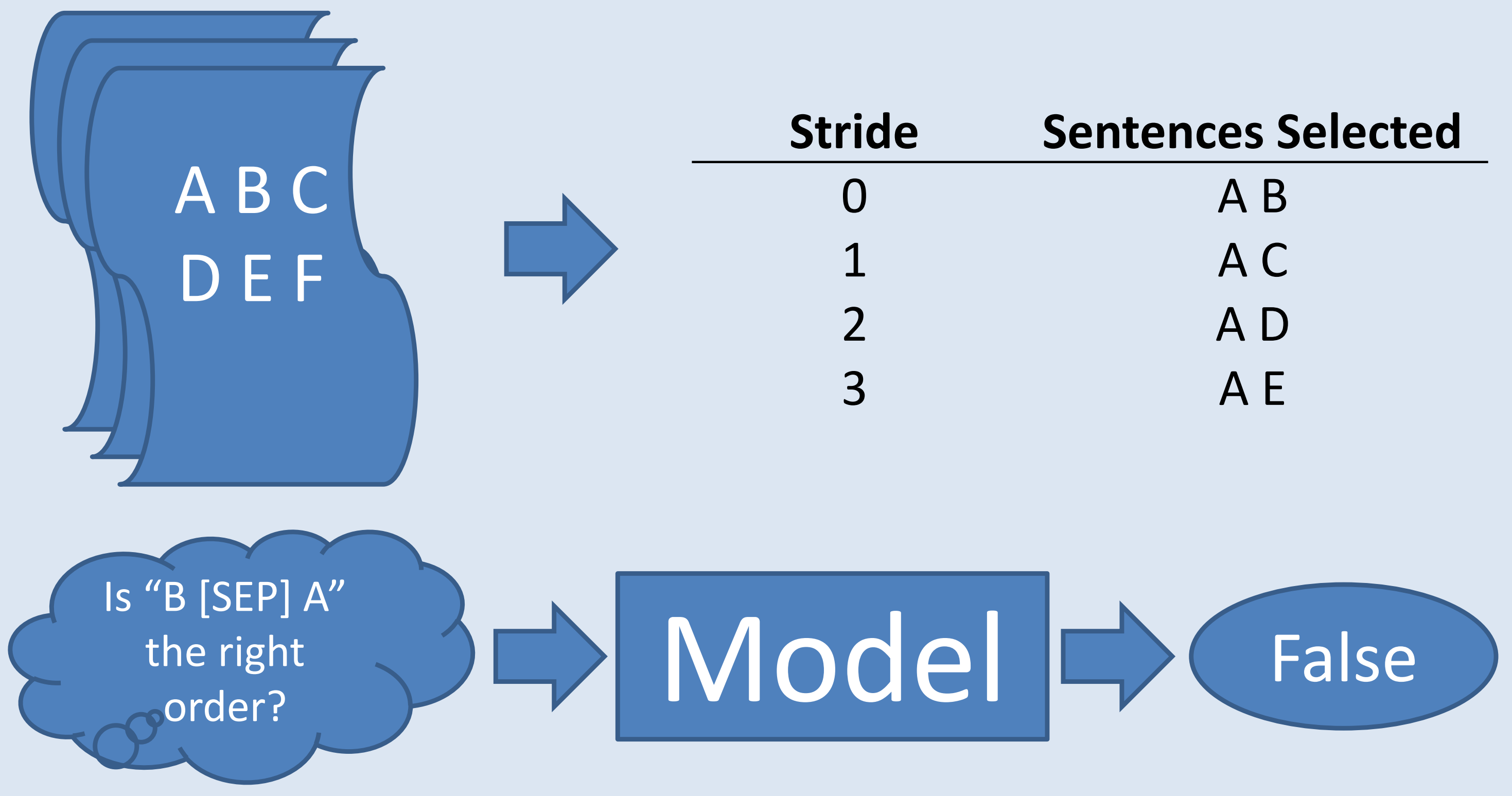
Research Questions

How well can Transformer models do temporal reasoning?

How can we define temporal reasoning?

Methods

- **Strided** Sentence Order Prediction (SOP) as a metric for temporal reasoning. Adapted from the pretraining literature (Lan et. al.).



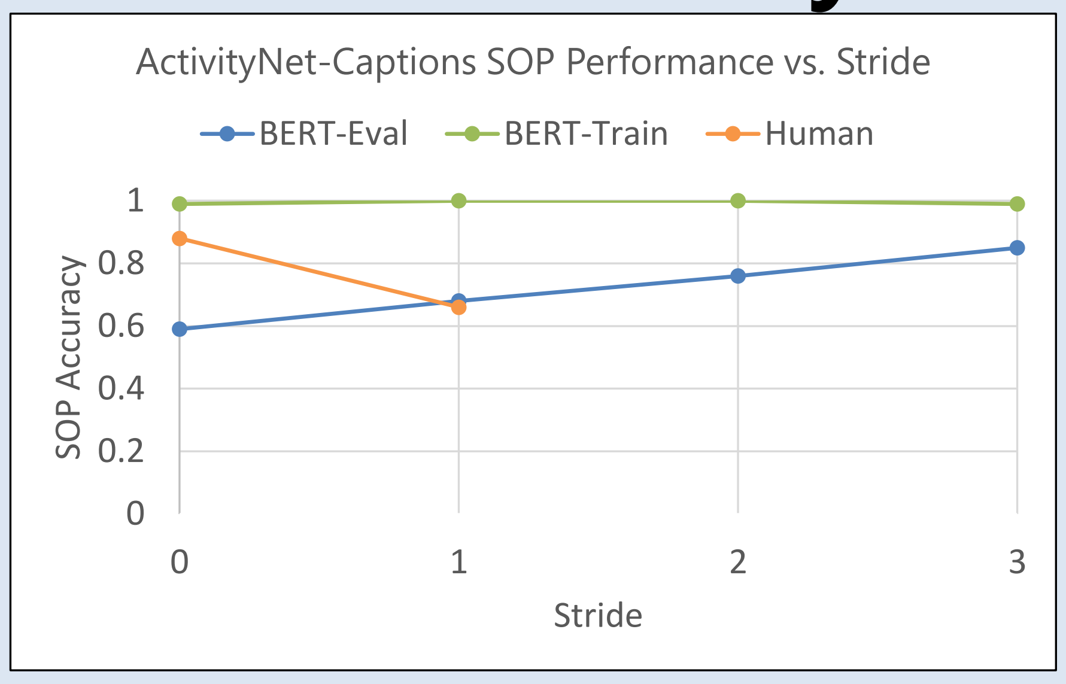
Stride	Sentences Selected
0	A B
1	A C
2	A D
3	A E

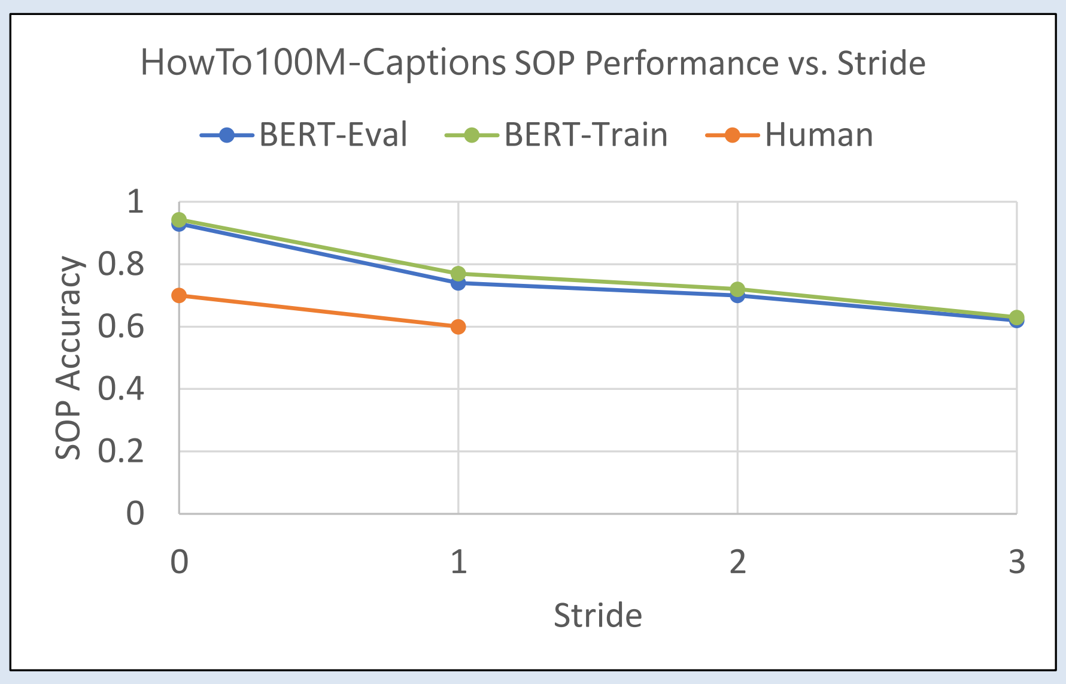
- From the pretrained checkpoints, train and evaluate the models using SOP as the training objective.
- Compare SOP performance on the evaluation set as stride increases.
- Trained and evaluated on
  - ActivityNet-Captions – Human generated
  - HowTo100M-Captions – ASR on YouTube videos

Hypothesis

- Expect to see SOP performance decrease with stride.
- *Rationale*: As stride increases, ordering sentences ostensibly requires temporal understanding, not just simple context clues like content matching.
- Assumption: The two sentences provide enough “temporal context” to solve the strided SOP task.

Results and Analysis





- Performance-Stride relationship depends on dataset.
- Model not learning the true task.
- Instead it learns some other underlying patterns in the data and provides a degenerate solution.

ActivityNet-Captions

Error Type	Example	Stride
Success	The man drops the weight to the ground. The man lifts the weight over his head.	0
Need Visual Grounding	The director of the race gives an interview as people pass behind him. The people are seen running the marathon.	0
Unrelated Concepts	A man walks holding a paper and pen. A gym ##nast stands on his arms.	3
GT doesn’t make sense	He is scrap ##ping off the excess wax from the ski. He uses the buffer on the ski again.	0

HowTo100M-Captions

Error Type	Example	Stride
Success	mom was told me you should cut the. like a couple of inches I don't know why	1
Success - Repeated words	straight in in kinda even with you know enjoys the j ##ois ##t sometimes people get. enjoys the j ##ois ##t sometimes people get	0
Need Visual Grounding	brown that's what we want them they’re. not gonna get as dark as your regular	1
Unrelated Concepts	We're gonna mix that together make sure that it's not c ##lump ##y , you've actually. some ha ##m and cheese	0
GT doesn’t make sense	mystical powers so as you can see. the power of magic	1

Conclusions

- **“Temporal Context” for ordering is difficult to define. This breaks our previous assumption.**
- HowTo100M-Captions is particularly difficult for humans, because of the noisy ASR-generated captions.
- ActivityNet-Captions dataset still requires visual grounding.
- Vision-Language datasets (e.g. captions) do not provide enough text-only temporal context.

Future Work

- Redefine sentences to include *n*-seconds of captions.
- Order three (or more) sentences instead of two sentences – redefine strided SOP training objective.
- Develop a model that is stride invariant.
- Can a single model do temporal reasoning with varying input strides?
- Temporal reasoning in the Vision-Language domain. Use the datasets’ visual features and a Vision-Language model.

Microsoft AI Residency Program