EMNLP 2022

hyadata lab | AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE | האוניברסיטה העברית בירושלים THE HEBREW UNIVERSITY OF JERUSALEM

# Breakpoint Transformers for Modeling and Tracking Intermediate Beliefs

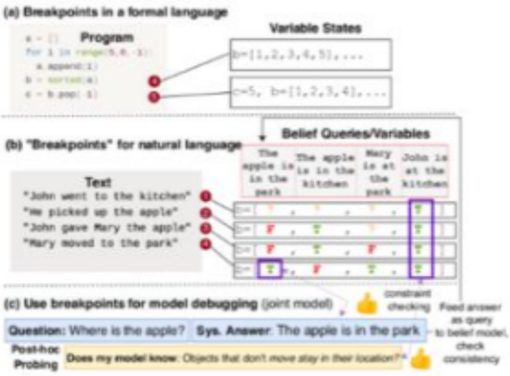

Kyle Richardson[2], Ronen Tamari[1*], Oren Sultan[1], Reut Tsarfaty[3], Dafna Shahaf[1] and Ashish Sabharwal[2]

[1]Hebrew University of Jerusalem, Israel, [2]Allen Institute of Artificial Intelligence, Seattle, USA, [3]Bar-Ilan University, Israel

* Work begun during internship at AI2 ❄ Equal contribution

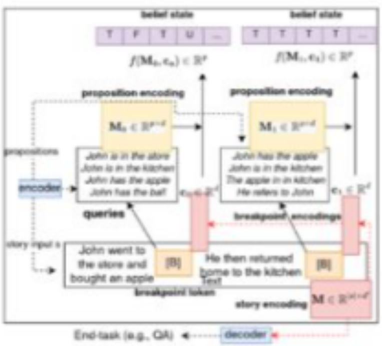

## Tracking intermediate Beliefs: Motivation

- Wish your language model (LM) had **breakpoints** you could inspect to probe its intermediate semantic representations?
- Breakpoints in programming are vital for code interpretability: **allow inspection** of program state at **intermediate points throughout execution**
- We develop a new idea of "natural language breakpoints" that can be used to probe LM encodings of input texts




## Breakpoint Transformers (BPTs): Modeling Approach

- Breakpoints are simply a special token **[B]** inserted after each sentence
- Breakpoint encoding can then be queried against natural language proposition *p* to obtain *{T,F,?}* prediction
- Represent summary of model "beliefs" at that point in text

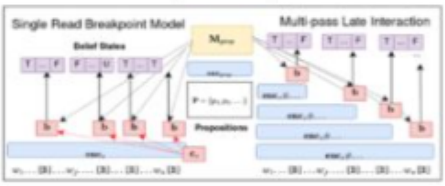



**Optimization**

Semantic logic loss translates to cross-entropy loss:

$$\bigwedge_{b_j \in B^{(i)}} \bigwedge_{p_k \in \mathbf{P}_j^{(i)}} Y_{j,k}^{(i)}(b_j, p_k) = \sum_{j=1}^{m} \sum_{k=1}^{t} -\log \Pr[y_{j,k}]$$

- BPT compared against standard late-interaction baseline (Sentence Transformers*) - BPT single read enables efficient scaling




*Reimers & Gureyvich (2019)

## Datasets: Annotating Intermediate State

| Task | Example Stories | Breakpoint Propositions |
|---|---|---|
| Relational Reasoning (CLUTRR) | John is the brother of Susan [B]₁ Susan's mother is Janice [B]₂ ... | P₁ : [ "Susan is the sister of John" true, "Susan is the sister-in-law of Janice" false, "Janice is the mother of John" unk ] P₂ : [ "Janice is the mother of John" true, "John is the father of Janice" false, ...] |
| Story Understanding (bAbI) | John moved to the kitchen [B]₁ He picked up an apple [B]₂ John then gave the apple to Mary [B]₃ ... | P₁ : [ "John has the apple" false, "John is in the kitchen" true, ...] P₂ : [ "John has the apple" true, "John is in the kitchen" ...] ... P₃ : [ "John has the apple" false, "Mary has the apple" true ] |
| Commonsense (TRIP) | Tom dropped his radio ...carpet. [B]₁ The radio broke ... [B]₂ Tom turned on the radio ... [B]₃ ... | P₂ : [ "radio is in pieces" true, "radio is powered" false], ... P₃ : [ "radio was powered" true ] |

bAbI, CLUTRR: synthetic | TRIP (Storks et. al, 2021): human-authored

## Experimental Results

**CLUTRR**

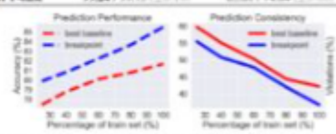
- BPTs show improved prediction acc., consistency and training efficiency

Proposition Prediction

| Model | Dev / Test Set + (std) (Acc %) |
|---|---|
| Majority Baseline | 44.60 / 41.60 |
| BiLSTM (Multi-pass) | 60.36 / 58.59 (±0.24) |
| T5-large (Multi-pass) | 81.41 / 81.94 (±0.17) |
| **BPT-large** | **85.16** / **85.24** (±0.34) |

Question-answering, **dev** / **test** + (std), (**EM** Acc %)

| Model | i.i.d | **generalization** |
|---|---|---|
| FT-T5-base | 99.00 / 99.78 (±0.15) | 84.19 / **75.13** (±0.94) |
| FT-Bart-base | 98.65 / 98.94 (±0.74) | 83.21 / 70.42 (±1.33) |
| **BPT-base** | **99.24** / 99.75 (±0.16) | 85.61 / 74.34 (±0.86) |




**bAbI**

- BPTs can accurately predict hundreds of relations across long stories jointly (compared to **multi-pass** baseline)

| Model | i.i.d Prop.% | i.i.d QA% | hard QA QA% |
|---|---|---|---|
| Majority | 65.87 | – | – |
| FT-T5-base (QA) | – | 97.29 ([illegible]) | 69.09 ([illegible]) |
| FT-Bart-base | – | 97.87 ([illegible]) | 67.21 ([illegible]) |
| BiLSTM (Multi) | 80.2 ([illegible]) | – | – |
| T5-base (Multi) | 99.1 ([illegible]) | – | – |
| **BPT-base** | 98.5 ([illegible]) | – | – |
| **BPT-base + QA** | 98.5 ([illegible]) | 94.9 ([illegible]) | **76.81** ([illegible]) |

**TRIP**


- BPT show up to 20-30% improvement against RoBERTA-based (RoB) approach of Storks et. al (2021)
- BPT needed no additional arch. adaptation, RoB tailored arch specifically for TRIP

| Split | Model | Task 1 *(Plaus.)* | Task 2 *(Consist.)* | Task 3 *(Verif.)* |
|---|---|---|---|---|
| Dev | RoB | 73.6 | 22.4 | 10.6 |
| | **BPT-base** | **81.99** (±0.91) | **58.07** (±0.76) | **36.44** (±0.53) |
| Test | RoB | 72.9 | 19.1 | 9.1 |
| | **BPT-base** | **80.55** (±1.20) | **53.83** (±1.65) | **32.37** (±0.27) |

*example figure from Storks et. al (2021)

## Discussion

- BPTs are modular extension of Transformers: added to existing models without harming performance
- BPTs improve model interpretability, easily applicable to narrative/procedural text understanding tasks
- Limitations & future work:
  - Systematic generalization: BPTs inherit limitations of pre-trained LMs
  - Causal relation between breakpoints and generated outputs is unclear, can possibly be enhanced by new joint consistency losses

**Code and experiments available on GitHub!** https://github.com/allenai/situation_modeling