Language Models Are Poor Learners of Directional Inference

Tianyi Li, Mohammad Javad Hosseini, Sabine Weber, Mark Steedman | Findings of EMNLP 2022

Directionality in Predicate Entailments

- **John shopped at Tesco** |= **John went to Tesco**;
- **John went to Tesco** !|= **John shopped at Tesco**;
- Directionality is what separates predicate entailment from semantic similarity;
- Directionality is NOT subsumption: other than more general descriptions of the same event, directionality can hold between different events, where one is the pre-condition or consequence of another.

LM-Prompting in Entailment Detection

- Concatenate premise and hypothesis in a prompt sentence, feed into BERT / RoBERTa language model, use a sentence-classification head;
  - Example prompt: John shopped at Tesco, which means John went to Tesco.
- Language models are good paraphrase detectors, but they show limited evidence of learning directional predicate entailment, as our experiments discover.

Experiment setup

- We categorize entries in Levy / Holt dataset into the following **SUB-GROUPS** (in GREEN are entailments, in RED are non-entailments):
  - DirTrue: directional true-entailments, P |= Q, Q!|=P;
  - DirFalse: directional non-entailments, P !|= Q, Q |= P;
  - Paraphrases: symmetric paraphrases, P |= Q, Q |= P;
  - Unrelated: no entailment relations, P !|= Q, Q !|= P;
- We denote **SUBSETS** as two sub-groups contrasting each other:
  - Example: DirTrue – DirFalse (namely, the *Directional subset*)
- We train and evaluate S&S [1], a SOTA LM-prompting method, on each pair of sub-groups of LevyHolt, and compare their performances to sketch out a panorama of the LM's ability to learn various kinds of entailments.
- We compare between subsets with different random-precision-baselines, so we propose the metric of normalized AUC, instead of regular AUC with fixed thresholds.

The Directionality Triangles

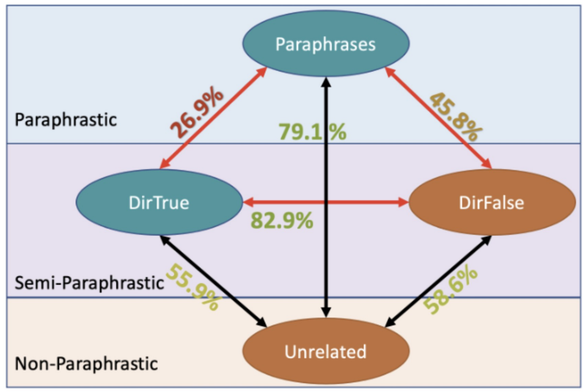

Figure 1: S&S models on mesh of pairs of sub-groups, results in AUC_{norm}.

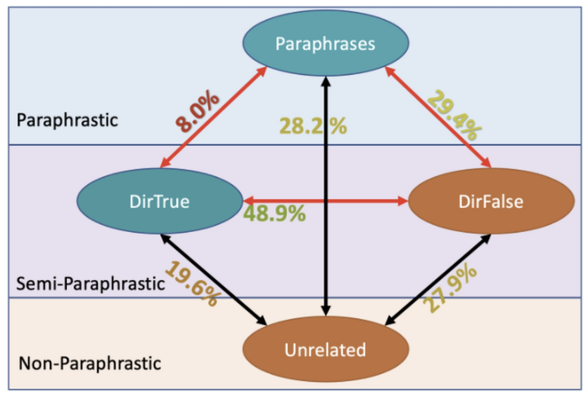

Figure 2: Hypothesis-only artefact baselines on mesh of pairs of sub-groups, results in AUC_{norm}.

- Above are meshes of subsets between each pair of sub-groups, in red triangles are the subsets where sensitivity to directionality is required in some way;
- As shown from the two figures, for the directional subset, coming with the anomalously high full-model AUC is the anomalously high AUC for the hypo-only model, and an anomalously low ratio between the two;
- This suggests dataset artefacts are at play here;
- Other evidence show that directionality is learnt poorly by the LM: for instance, the model is very poor at separating paraphrases and directional true-entailments, one holding on both sides, the other holding on only one side.

The BoOQA Dataset

- With the question of the LM's directional sensitivity in mind, we resort to an extrinsic evaluation outside the shadow of artefacts in Levy/Holt dataset.
- We present a Boolean Open QA dataset (BoOQA), for this evaluation. The dataset is based on the dataset from [2], but extended and refined on robustness.
- The BoOQA dataset is constructed from subsumption relations in WordNet, but is capable of testing for a wide range of directional entailments.
- We verify the quality of this dataset, and that it does not share the same set of artefacts as the Levy/Holt dataset, which we abandoned.

AUC_{norm} (%)	$BoOQA_{En}$	$BoOQA_{Zh}$
$LM_{unsupervised}$	15.9	30.6
$S\&S_{Full}$	25.6	23.1*
$\widehat{S\&S}_{Full}$	26.2	-
$S\&S_{Symmetric}$	25.1	-
$S\&S_{Directional}$	15.1	-
EG BInc	29.8	39.5
EG CNCE	**34.5**	-
EG EGT2	26.8	-

Table 4: Baselines on BoOQA test set in English and Chinese. All methods are taken out-of-the-box. "EG XX" are entailment graphs with various entailment scores as described below.

- From this extrinsic evaluation, it is confirmed that LM-prompting methods are almost indifferent to directional prompts, and are in general less charming than suggested by supervised results from Levy/Holt.

Key References

- [1] Martin Schmitt and Hinrich Schutze. Language Models for Lexical Inference in Context. EACL 2021.
- [2] Nick McKenna, Liane Guillou, Mohammad Javad Hosseini, Sander Bijl de Vroe, Mark Johnson, and Mark Steedman. Multivalent Entailment Graphs for Question Answering. EMNLP 2021.
- [3] Adam Poliak, Jason Naradowsky, Aparajita Haldar, 878 Rachel Rudinger, and Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. SEM 2022.