

Generating Natural Language Proofs with Verifier-Guided Search Kaiyu Yang, Jia Deng, and Danqi Chen Department of Computer Science, Princeton University **Our Method: NLProofS** Experiments NLProofS (<u>Natural Language Proof Search</u>) **State-of-the-art on two benchmarks** RuleTaker ^[Tafjord et al. 2021]: Simple, synthetic profs A new method for stepwise proof generation Prover: Generate relevant steps conditioning on the hypothesis Verifier: Mitigate hallucination by training an independent network to check the proof steps Main Results on EntailmentBank Proof search: Use the prover/verifier to generate the final • Single-shot baselines: EntailmentWriter ^[Dalvi et al. 2021] proof with the optimal validity score Accuracies on Task 2 of EntailmentBank Prover EntailmentWriter IGRG MetGen NLProofS (ours) Finetune a T5 model to predict the next proof step *int2*: solar is a kind of renewable energy Generate multiple candidate steps via beam search sent2: solar is Hypothesis (*h*): sent4: solar is a kind of energy renewable h: solar is a kind of renewable 30 22.9 22.9 22.3 energy for heating homes 20 Supporting facts (*C*): 1-step partial proof The 2nd step *sent1*: homes are buildings int2: solar is a kind of sent2: solar is renewable int1: energy is used Steps Intermediates Overall leaves *sent3*: wind is a kind of energy renewable energy for heating homes *sent4*: solar is a kind of energy sent5: energy is used for heating buildings sent4: solar is a sent2: solar sent1: homes sent5: energy is used ent6: coal is nonrenewable kind of energy is renewable are buildings for heating buildings Accuracies on Task 2 of EntailmentBank Encode input/output as text sequences ■ w/o search ■ w/o search w/o stepwise ■ w/o verifier score NLProofS (full) hypothesis\$ = solar is a kind of renewable energy for heating homes sent2 & sent4 -> int2: solar is a kind of renewable energy spartial proof\$ = sent1 & sent5 -> int1: energy is used for heating homes 60 Verifier 50 40 Input: A proof step (multiple premises, one conclusion) 30 Output: A score in [0, 1] calculated by finetuning RoBERTa 20 Step scores are aggregated to calculate proof scores h: solar is a kind of renewable Steps Intermediates Leaves Overall energy for heating homes int1: energy is used for heating homes **0.7** *int2*: solar is a kind of Large Language Models w/ In-Context Learning **0.6** *int1*: energy is used for heating homes renewable energy *sent1*: homes *sent5*: energy is used • Single-shot proof generation with 7 in-context examples are buildings for heating buildings • GPT-3 and Codex cannot solve the task out of the box **1.0** *sent1*: homes **1.0** *sent5*: energy is used **1.0** *sent4*: solar is a **1.0** *sent2*: solar are buildings for heating buildings kind of energy is renewable

- **Proof Search**
- Initialization: a proof generated by the prover alone
- 2. Iteration: expand the graph iteratively
- Using steps proposed by the prover
- Checked by the verifier
- Average verifier/prover scores
- B. Extraction: proof tree with the best score

