Generating Natural Language Proofs with Verifier-Guided Search

Kaiyu Yang, Jia Deng, Danqi Chen
Reasoning in Natural Language

- homes are buildings
- energy is used for heating buildings
- solar is a kind of energy
- solar is renewable

Assumptions: homes are buildings, energy is used for heating buildings, solar is a kind of energy, solar is renewable

Conclusion: solar is a kind of renewable energy for heating homes
Reasoning in Natural Language

• Studied extensively in automated theorem proving
• **Remains challenging in natural language**
  • Fuzzy, imprecise, requiring implicit knowledge
  • No well-defined inference rules

homes are buildings
energy is used for heating buildings
solar is a kind of energy
solar is renewable

? = solar is a kind of renewable energy for heating homes

Assumptions Conclusion

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
Reasoning in Natural Language

- Studied extensively in automated theorem proving
- Remains challenging in natural language
  - Fuzzy, imprecise, requiring implicit knowledge
  - No well-defined inference rules

[Clark et al. IJCAI 2020]

LLMs for “soft” reasoning over natural language

Assumptions

- homes are buildings
- energy is used for heating buildings
- solar is a kind of energy
- solar is renewable

Conclusion

- solar is a kind of renewable energy for heating homes

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
Task: Proof Generation

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

$\text{sent1}$: homes are buildings
$\text{sent2}$: solar is renewable
$\text{sent3}$: wind is a kind of energy
$\text{sent4}$: solar is a kind of energy
$\text{sent5}$: energy is used for heating buildings
$\text{sent6}$: coal is nonrenewable
...
...

Input

[Dalvi et al. EMNLP 2021] Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
Task: Proof Generation

Hypothesis ($h$):
$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

- **sent1**: homes are buildings
- **sent2**: solar is renewable
- **sent3**: wind is a kind of energy
- **sent4**: solar is a kind of energy
- **sent5**: energy is used for heating buildings
- **sent6**: coal is nonrenewable
...
Task: Proof Generation

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

- **sent1**: homes are buildings
- **sent2**: solar is renewable
- **sent3**: wind is a kind of energy
- **sent4**: solar is a kind of energy
- **sent5**: energy is used for heating buildings
- **sent6**: coal is nonrenewable
  ...

Proof tree ($T'$):

- $h$: solar is a kind of renewable energy for heating homes
  - **int1**: energy is used for heating homes
    - **sent1**: homes are buildings
    - **sent5**: energy is used for heating buildings
  - **int2**: solar is a kind of renewable energy
    - **sent4**: solar is a kind of energy
    - **sent2**: solar is renewable

[Dalvi et al. EMNLP 2021]
Task: Proof Generation

Hypothesis ($h$):
$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

- sent1: homes are buildings
- sent2: solar is renewable
- sent3: wind is a kind of energy
- sent4: solar is a kind of energy
- sent5: energy is used for heating buildings
- sent6: coal is nonrenewable
...

Proof tree ($T'$):

- $h$: solar is a kind of renewable energy for heating homes
  - int1: energy is used for heating homes
    - sent1: homes are buildings
    - sent5: energy is used for heating buildings
  - int2: solar is a kind of renewable energy
    - sent4: solar is a kind of energy
    - sent2: solar is renewable

[Dalvi et al. EMNLP 2021]

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
Task: Proof Generation

Hypothesis ($h$):
$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):
- $sent1$: homes are buildings
- $sent2$: solar is renewable
- $sent3$: wind is a kind of energy
- $sent4$: solar is a kind of energy
- $sent5$: energy is used for heating buildings
- $sent6$: coal is nonrenewable

Proof tree ($T'$):
$h$: solar is a kind of renewable energy for heating homes

$int1$: energy is used for heating homes
$int2$: solar is a kind of renewable energy

$sent1$: homes are buildings
$sent5$: energy is used for heating buildings
$sent4$: solar is a kind of energy
$sent2$: solar is renewable

[Dalvi et al. EMNLP 2021]
Task: Proof Generation

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

$sent1$: homes are buildings
$sent2$: solar is renewable
$sent3$: wind is a kind of energy
$sent4$: solar is a kind of energy
$sent5$: energy is used for heating buildings
$sent6$: coal is nonrenewable
...
...

Proof tree ($T'$):

$h$: solar is a kind of renewable energy for heating homes

$int1$: energy is used for heating homes

$int2$: solar is a kind of renewable energy

$sent1$: homes are buildings
$sent5$: energy is used for heating buildings
$sent4$: solar is a kind of energy
$sent2$: solar is renewable

Generated by the model

[Dalvi et al. EMNLP 2021]
Single-Shot Methods

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

- **sent1**: homes are buildings
- **sent2**: solar is renewable
- **sent3**: wind is a kind of energy
- **sent4**: solar is a kind of energy
- **sent5**: energy is used for heating buildings
- **sent6**: coal is nonrenewable

... ...

Proof tree ($T'$):

$h$: solar is a kind of renewable energy for heating homes

*int1*: energy is used for heating homes

*int2*: solar is a kind of renewable energy

*sent1*: homes are buildings

*sent5*: energy is used for heating buildings

*sent4*: solar is a kind of energy

*sent2*: solar is renewable

Generating the entire proof altogether

---

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
Stepwise Methods

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

$\text{sent1}$: homes are buildings
$\text{sent2}$: solar is renewable
$\text{sent3}$: wind is a kind of energy
$\text{sent4}$: solar is a kind of energy
$\text{sent5}$: energy is used for heating buildings
$\text{sent6}$: coal is nonrenewable

... 

Proof tree ($T'$):

Step 1

$\text{int1}$: energy is used for heating homes

$\text{sent1}$: homes are buildings
$\text{sent5}$: energy is used for heating buildings

Input

Output

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
Stepwise Methods

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

$sent1$: homes are buildings
$sent2$: solar is renewable
$sent3$: wind is a kind of energy
$sent4$: solar is a kind of energy
$sent5$: energy is used for heating buildings
$sent6$: coal is nonrenewable
...
...

Proof tree ($T'$):

Step 1

$int1$: energy is used for heating homes

$sent1$: homes are buildings
$sent5$: energy is used for heating buildings

Step 2

$int2$: solar is a kind of renewable energy

$sent2$: solar is renewable
$sent4$: solar is a kind of energy

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
Stepwise Methods

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

- $sent1$: homes are buildings
- $sent2$: solar is renewable
- $sent3$: wind is a kind of energy
- $sent4$: solar is a kind of energy
- $sent5$: energy is used for heating buildings
- $sent6$: coal is nonrenewable

Proof tree ($T'$):

- $step1$: $int1$: energy is used for heating homes
- $step2$: $int2$: solar is a kind of renewable energy

$step3$: $h$: solar is a kind of renewable energy for heating homes

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
<table>
<thead>
<tr>
<th>Generate the entire proof altogether</th>
<th>Generate the proof step by step</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PRover</strong> [Saha et al. EMNLP 2020]</td>
<td><strong>ProofWriter</strong> [Tafjord et al. Findings of ACL 2021]</td>
</tr>
<tr>
<td><strong>EntailmentWriter</strong> [Dalvi et al. EMNLP 2021]</td>
<td><strong>FaiRR</strong> [Sanyal et al. ACL 2022]</td>
</tr>
<tr>
<td><strong>PRobr</strong> [Sun et al. Findings of ACL 2021]</td>
<td><strong>SCSearch</strong> [Bostrom et al. Findings of EMNLP 2022]</td>
</tr>
<tr>
<td><strong>MetGen</strong></td>
<td><strong>MetGen</strong> [Hong et al. Findings of NAACL 2022]</td>
</tr>
</tbody>
</table>

- Can better leverage compositionality and generalize to longer proofs
- Achieved limited success on challenging proofs authored by humans (e.g., EntailmentBank)
Stepwise Methods

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

$\text{sent1}$: homes are buildings
$\text{sent2}$: solar is renewable
$\text{sent3}$: wind is a kind of energy
$\text{sent4}$: solar is a kind of energy
$\text{sent5}$: energy is used for heating buildings
$\text{sent6}$: coal is nonrenewable
...
...

Proof tree ($T'$):

$\text{h: solar is a kind of renewable energy for heating homes}$

$\text{int1: energy is used for heating homes}$

$\text{int2: solar is a kind of renewable energy}$

$\text{sent1: homes are buildings}$

$\text{sent5: energy is used for heating buildings}$

$\text{sent4: solar is a kind of energy}$

$\text{sent2: solar is renewable}$


Generated by the model
Challenges in Generating **Valid** and **Relevant** Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)

Diagram:

- **h**: solar is a kind of renewable energy for heating homes
  - **int1**: energy is used for heating homes
    - **sent1**: homes are buildings
    - **sent5**: energy is used for heating buildings
  - **int2**: solar is a kind of renewable energy
    - **sent4**: solar is a kind of energy
    - **sent2**: solar is renewable

Challenges in Generating **Valid** and **Relevant** Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)
Challenges in Generating **Valid** and **Relevant** Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)

```
int1: homes are buildings that can be heated using energy

sent1: homes are buildings

int1: homes are buildings

int1: energy is used for heating homes

sent5: energy is used for heating buildings

int2: solar is a kind of renewable energy

sent4: solar is a kind of energy

sent2: solar is renewable

h: solar is a kind of renewable energy for heating homes
```
Challenges in Generating **Valid and Relevant** Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)
Challenges in Generating **Valid** and **Relevant** Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)
- The model hallucinates invalid steps

[Sanyal et al. ACL 2022]
[Bostrom et al. arXiv 2022]
Challenges in Generating **Valid** and **Relevant** Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)
- The model hallucinates invalid steps

[Sanyal et al. ACL 2022]
[Bostrom et al. arXiv 2022]
Challenges in Generating Valid and Relevant Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)
- The model hallucinates invalid steps

[Sanyal et al. ACL 2022]
[Bostrom et al. arXiv 2022]
Challenges in Generating **Valid** and **Relevant** Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)
- The model hallucinates invalid steps
- Existing stepwise methods
  - Struggle to generate valid and relevant steps
  - Underperform on human-authored proofs

---

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
Challenges in Generating **Valid** and **Relevant** Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)
- The model hallucinates invalid steps
- Existing stepwise methods
  - Struggle to generate valid and relevant steps
  - Underperform on human-authored proofs

**Our solution:** a new method for stepwise proof generation
NLProofS: **Natural Language Proof Search**

- A new method for **stepwise proof generation**

\[ h: \text{solar is a kind of renewable energy for heating homes} \]

\[ \text{int1: energy is used for heating homes} \]

\[ \text{sent1: homes are buildings} \]

\[ \text{sent5: energy is used for heating buildings} \]

\[ \text{sent4: solar is a kind of energy} \]

\[ \text{sent2: solar is renewable} \]
A new method for **stepwise proof generation**

**Prover**

*Generate candidate proof steps*

- **sent1**: homes are buildings
- **sent4**: solar is a kind of energy
- **sent5**: energy is used for heating buildings
- **int1**: energy is used for heating homes
- **int2**: solar is a kind of renewable energy
- **h**: solar is a kind of renewable energy for heating homes
NLProofS: Natural Language Proof Search

- A new method for stepwise proof generation

Prover

Generate candidate proof steps

$h$: solar is a kind of renewable energy for heating homes

$int1$: energy is used for heating homes

$int2$: solar is a kind of renewable energy

$sent4$: solar is a kind of energy

$sent5$: energy is used for heating buildings

$sent1$: homes are buildings
NLProofS: Natural Language Proof Search

- A new method for *stepwise proof generation*

Prover

Verifier

Score the validity

- **h**: solar is a kind of renewable energy for heating homes
- **int1**: energy is used for heating homes
- **sent1**: homes are buildings
- **int2**: solar is a kind of renewable energy
- **sent4**: solar is a kind of energy
- **sent5**: energy is used for heating buildings

Score the validity

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
NLProofS: Natural Language Proof Search

- A new method for stepwise proof generation

Prover

Verifier

Training

Prover

Verifier

Training

h: solar is a kind of renewable energy for heating homes

int1: energy is used for heating homes

sent1: homes are buildings

int2: solar is a kind of renewable energy

sent4: solar kind of energy

sent5: energy is used for heating buildings
NLProofS: Natural Language Proof Search

- A new method for **stepwise proof generation**

---

### Training

- **Prover**
- **Verifier**

### Inference

- **Proof search**

---

- **sent1**: homes are buildings
- **sent2**: solar is a kind of renewable energy
- **sent4**: solar is a kind of energy
- **sent5**: energy is used for heating buildings
- **int1**: energy is used for heating homes
- **int2**: solar is a kind of renewable energy
- **h**: solar is a kind of renewable energy for heating homes

---

**Figure**: Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
NLProofS: Natural Language Proof Search

- A new method for **stepwise proof generation**

---

**Generating Natural Language Proofs with Verifier-Guided Search** - Kaiyu Yang, Jia Deng, Danqi Chen
Stepwise Prover

Hypothesis \( (h) \):

\( h: \) solar is a kind of renewable energy for heating homes

Supporting facts \( (C) \):

\textit{sent1}: homes are buildings
\textit{sent2}: solar is renewable
\textit{sent3}: wind is a kind of energy
\textit{sent4}: solar is a kind of energy
\textit{sent5}: energy is used for heating buildings
\textit{sent6}: coal is nonrenewable
...
Stepwise Prover

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

- **sent1**: homes are buildings
- **sent2**: solar is renewable
- **sent3**: wind is a kind of energy
- **sent4**: solar is a kind of energy
- **sent5**: energy is used for heating buildings
- **sent6**: coal is nonrenewable
- ...

...
Stepwise Prover

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

- $sent1$: homes are buildings
- $sent2$: solar is renewable
- $sent3$: wind is a kind of energy
- $sent4$: solar is a kind of energy
- $sent5$: energy is used for heating buildings
- $sent6$: coal is nonrenewable

... 

... 

1-step partial proof

$int1$: energy is used for heating homes

$sent1$: homes are buildings

$sent5$: energy is used for heating buildings

The 2$^{nd}$ step

$int2$: solar is a kind of renewable energy

$sent4$: solar is a kind of energy

$sent2$: solar is renewable
Stepwise Prover

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

- $sent1$: homes are buildings
- $sent2$: solar is renewable
- $sent3$: wind is a kind of energy
- $sent4$: solar is a kind of energy
- $sent5$: energy is used for heating buildings
- $sent6$: coal is nonrenewable
  ...

1-step partial proof

- $int1$: energy is used for heating homes

The 2\textsuperscript{nd} step

- $int2$: solar is a kind of renewable energy

- $sent2$: solar is renewable
- $sent4$: solar is a kind of energy
- $sent5$: energy is used for heating buildings
- $sent6$: coal is nonrenewable

\[\text{[Raffle et al. JMLR 2020]}\]
\[\text{[Tafjord et al. Findings of ACL 2021]}\]
Stepwise Prover

Hypothesis ($h$):

$h$: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

- $sent1$: homes are buildings
- $sent2$: solar is renewable
- $sent3$: wind is a kind of energy
- $sent4$: solar is a kind of energy
- $sent5$: energy is used for heating buildings
- $sent6$: coal is nonrenewable

1-step partial proof:

$int1$: energy is used for heating homes

$sent1$: homes are buildings

$sent5$: energy is used for heating buildings

The 2nd step:

$int2$: solar is a kind of renewable energy

$sent4$: solar is a kind of energy

$sent2$: solar is renewable

Encode input/output as text sequences

$hypothesis$ = $sent2$: solar is a kind of renewable energy for heating homes;
$facts$ = $sent1$: homes are buildings $sent2$: solar is renewable ... ;
$partial-proof$ = $sent1$ & $sent5$ -> $int1$: energy is used for heating homes;

$T5$  

sent2 & $sent4$ -> $int2$: solar is a kind of renewable energy;
NLProofS: Natural Language Proof Search

Prover

Verifier

Proof search

Training

Inference

sent1: homes are buildings
int1: energy is used for heating homes
h: solar is a kind of renewable energy for heating homes

sent4: solar is a kind of energy
int2: solar is a kind of renewable energy

sent5: energy is used for heating buildings
NLProofS: **Natural Language Proof Search**

Prover

Verifier

Training

Proof search

Inference

**sent1:** homes are buildings

**sent4:** solar is a kind of renewable energy

**sent5:** energy is used for heating buildings

**int1:** energy is used for heating homes

**int2:** solar is a kind of renewable energy

**h:** solar is a kind of renewable energy for heating homes

Training

Inference
NLProofS: Natural Language Proof Search

Prover

Verifier

Training

Proof search

Inference

Finetune RoBERTa
[Liu et al. arXiv 2019]

0.6

sent1: homes are buildings

int1: energy is used for heating homes

h: solar is a kind of renewable energy for heating homes

sent4: solar is a kind of energy

sent5: energy is used for heating buildings

int2: solar is a kind of renewable energy

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
NLProofS: **Natural Language Proof Search**

Aggregate the step scores to across the entire proof:

- **sent1**: homes are buildings
- **int1**: energy is used for heating homes
- **sent4**: solar is a kind of energy
- **int2**: solar is a kind of renewable energy
- **sent5**: energy is used for heating buildings
- **h**: solar is a kind of renewable energy for heating homes

Prover

Verifier

Proof search

Training

Inference
NLProofS: Natural Language Proof Search

Training

Prover

Verifier

Proof search

Inference

sent1: homes are buildings

int1: energy is used for heating homes

sent4: solar is a kind of energy

int2: solar is a kind of renewable energy

sent5: energy is used for heating buildings

h: solar is a kind of renewable energy for heating homes
NLProofS: Natural Language Proof Search

- **Prover**
- **Verifier**
- **Proof search**

**Training**

- **sent1**: homes are buildings
- **sent2**: solar is a kind of renewable energy for heating homes
- **sent3**: energy is used for heating buildings
- **sent4**: solar is a kind of renewable energy
- **int1**: energy is used for heating homes
- **int2**: solar is a kind of renewable energy

**Inference**

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
1. Initialization: a proof generated by the prover alone

- \( h \): solar is a kind of renewable energy for heating homes
- \( int1 \): energy is used for heating homes
- \( int2 \): solar is a kind of renewable energy
- \( sent1 \): homes are buildings
- \( sent5 \): energy is used for heating buildings
- \( sent4 \): solar is a kind of energy
- \( sent2 \): solar is renewable
NLProofS: Natural Language Proof Search

1. Initialization: a proof generated by the prover alone
2. Iteration: expand the graph iteratively

Proof search

- **sent1**: homes are buildings
- **sent2**: solar is renewable
- **sent3**: solar is a kind of energy
- **sent4**: solar is a kind of renewable energy
- **sent5**: energy is used for heating buildings
- **int1**: energy is used for heating homes
- **int2**: solar is a kind of renewable energy
- **h**: solar is a kind of renewable energy for heating homes
NLProofS: Natural Language Proof Search

1. Initialization: a proof generated by the prover alone
2. Iteration: expand the graph iteratively

Proof search

- **h**: solar is a kind of renewable energy for heating homes
- **int1**: energy is used for heating homes
- **int2**: solar is a kind of renewable energy
- **sent1**: homes are buildings
- **sent5**: energy is used for heating buildings
- **sent4**: solar is a kind of energy
- **sent2**: solar is renewable
NLProofS: Natural Language Proof Search

1. Initialization: a proof generated by the prover alone
2. Iteration: expand the graph iteratively

- **sent1**: homes are buildings
- **sent2**: solar is renewable
- **sent4**: solar is a kind of energy
- **sent5**: energy is used for heating buildings
- **int1**: energy is used for heating homes
- **int2**: solar is a kind of renewable energy
- **h**: solar is a kind of renewable energy for heating homes

Proof search
NLProofS: Natural Language Proof Search

1. 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

Proof search

sent1: homes are buildings

h: solar is a kind of renewable energy for heating homes

int1: energy is used for heating homes

int2: solar is a kind of renewable energy

sent2: solar is renewable

sent4: solar is a kind of energy

sent5: energy is used for heating buildings

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
NLProofS: **Natural Language Proof Search**

1. **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
3. **Extraction:** proof tree with best score
Experiments

• Evaluate on two benchmarks
  • **RuleTaker**: Simple, synthetic proofs
  • **EntailmentBank**: ~2K challenging, human-written proofs

• State-of-the-art results on both

• Ablations highlight the importance of the verifier
Experiments

• Evaluate on two benchmarks
  • RuleTaker: Simple, synthetic proofs [Tafjord et al. Findings of ACL 2021]
  • EntailmentBank: ~2K challenging, human-written proofs [Dalvi et al. EMNLP 2021]
• State-of-the-art results on both
• Ablations highlight the importance of the verifier 25 supporting facts, including distractors
EntailmentBank: Evaluation Metrics

[Dalvi et al. EMNLP 2021]

- EntailmentBank’s four official metrics: Leaves, Steps, Intermediates, Overall

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
EntailmentBank: Evaluation Metrics

- EntailmentBank’s four official metrics: Leaves, Steps, Intermediates, Overall
- Based on heuristic matching between the nodes

[Dalvi et al. EMNLP 2021]
EntailmentBank: Evaluation Metrics

- EntailmentBank’s four official metrics: Leaves, Steps, Intermediates, Overall
- Based on heuristic matching between the nodes
- Limitations: Cannot handle correct predictions different from the ground truth

Predicted proof tree

Ground truth proof tree

[Dalvi et al. EMNLP 2021]
State-of-the-art Performance on EntailmentBank

### Accuracies on Task 2 of EntailmentBank

<table>
<thead>
<tr>
<th></th>
<th>EntailmentWriter</th>
<th>IGRG</th>
<th>MetGen</th>
<th>NLProofS (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>35.6</td>
<td>22.9</td>
<td></td>
<td>48.6</td>
</tr>
<tr>
<td></td>
<td>[Dalvi et al. EMNLP 2021]</td>
<td></td>
<td></td>
<td>[Ribeiro et al. Findings of NAACL 2022]</td>
</tr>
<tr>
<td>Steps</td>
<td>58.8</td>
<td>22.3</td>
<td>30.4</td>
<td>34.4</td>
</tr>
<tr>
<td></td>
<td>[Ribeiro et al. Findings of NAACL 2022]</td>
<td>[Hong et al. Findings of NAACL 2022]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediates</td>
<td>28.5</td>
<td>26.5</td>
<td>32.7</td>
<td>37.8</td>
</tr>
<tr>
<td>Overall</td>
<td>20.9</td>
<td>22</td>
<td>28</td>
<td>33.3</td>
</tr>
</tbody>
</table>

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
# State-of-the-art Performance on EntailmentBank

### Accuracies on Task 2 of EntailmentBank

<table>
<thead>
<tr>
<th>Method</th>
<th>Leaves</th>
<th>Steps</th>
<th>Intermediates</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EntailmentWriter</strong></td>
<td>35.6</td>
<td>22.9</td>
<td>28.5</td>
<td>20.9</td>
</tr>
<tr>
<td>IGRG</td>
<td>48.6</td>
<td>22.3</td>
<td>32.7</td>
<td>22</td>
</tr>
<tr>
<td>MetGen</td>
<td>58.8</td>
<td>30.4</td>
<td>37.8</td>
<td>33.3</td>
</tr>
<tr>
<td>NLProofS (ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Single-shot**

**[Dalvi et al. EMNLP 2021] [Ribeiro et al. Findings of NAACL 2022] [Hong et al. Findings of NAACL 2022]**

**Generating Natural Language Proofs with Verifier-Guided Search** - Kaiyu Yang, Jia Deng, Danqi Chen
State-of-the-art Performance on EntailmentBank

Accuracies on Task 2 of EntailmentBank

<table>
<thead>
<tr>
<th></th>
<th>EntailmentWriter</th>
<th>IGRG</th>
<th>MetGen</th>
<th>NLProofS (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>35.6</td>
<td>22.9</td>
<td>30.4</td>
<td>32.7</td>
</tr>
<tr>
<td>Steps</td>
<td>22.9</td>
<td>22.3</td>
<td>28.5</td>
<td>20.9</td>
</tr>
<tr>
<td>Intermediates</td>
<td>34.4</td>
<td>26.5</td>
<td>37.8</td>
<td>22</td>
</tr>
<tr>
<td>Overall</td>
<td>48.6</td>
<td>34.4</td>
<td>37.8</td>
<td>33.3</td>
</tr>
</tbody>
</table>

[Dalvi et al. EMNLP 2021] [Ribeiro et al. Findings of NAACL 2022] [Hong et al. Findings of NAACL 2022]
State-of-the-art Performance on EntailmentBank

Accuracies on Task 2 of EntailmentBank

- EntailmentWriter
- IGRG
- MetGen
- NLProofS (ours)

Leaves:
- EntailmentWriter: 58.8
- IGRG: 48.6
- MetGen: 48.6
- NLProofS: 35.6

Steps:
- EntailmentWriter: 48.6
- IGRG: 22.9
- MetGen: 22.3
- NLProofS: 22.9

Intermediates:
- EntailmentWriter: 30.4
- IGRG: 22.9
- MetGen: 28.5
- NLProofS: 28.5

Overall:
- EntailmentWriter: 34.4
- IGRG: 34.4
- MetGen: 32.7
- NLProofS: 37.8

References:

- Dalvi et al. EMNLP 2021
- Ribeiro et al. Findings of NAACL 2022
- Hong et al. Findings of NAACL 2022
- Dalvi et al. EMNLP 2021

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
State-of-the-art Performance on EntailmentBank

### Accuracies on Task 2 of EntailmentBank

<table>
<thead>
<tr>
<th></th>
<th>EntailmentWriter</th>
<th>IGRG</th>
<th>MetGen</th>
<th>NLProofS (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>35.6</td>
<td>48.6</td>
<td>58.8</td>
<td></td>
</tr>
<tr>
<td>Steps</td>
<td>22.9</td>
<td>22.9</td>
<td>30.4</td>
<td>34.4</td>
</tr>
<tr>
<td>Intermediates</td>
<td>28.5</td>
<td>26.5</td>
<td>32.7</td>
<td>37.8</td>
</tr>
<tr>
<td>Overall</td>
<td>20.9</td>
<td>22</td>
<td>28</td>
<td>33.3</td>
</tr>
</tbody>
</table>

[Dalvi et al. EMNLP 2021] [Ribeiro et al. Findings of NAACL 2022] [Hong et al. Findings of NAACL 2022]

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
State-of-the-art Performance on EntailmentBank

Accuracies on Task 2 of EntailmentBank

<table>
<thead>
<tr>
<th>Task</th>
<th>EntailmentWriter</th>
<th>IGRG</th>
<th>MetGen</th>
<th>NLProofS (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>35.6</td>
<td>22.9</td>
<td>22.9</td>
<td>23.2</td>
</tr>
<tr>
<td>Steps</td>
<td>48.6</td>
<td>30.4</td>
<td>34.4</td>
<td></td>
</tr>
<tr>
<td>Intermediates</td>
<td>28.5</td>
<td>26.5</td>
<td>32.7</td>
<td>37.8</td>
</tr>
<tr>
<td>Overall</td>
<td>20.9</td>
<td>22</td>
<td>28</td>
<td>33.3</td>
</tr>
</tbody>
</table>

[ Dalvi et al. EMNLP 2021 ]  [ Ribeiro et al. Findings of NAACL 2022 ]  [ Hong et al. Findings of NAACL 2022 ]
State-of-the-art Performance on EntailmentBank

Accuracies on Task 2 of EntailmentBank

- EntailmentWriter
- IGRG
- MetGen
- NLPinS (ours)

[Dalvi et al. EMNLP 2021] [Ribeiro et al. Findings of NAACL 2022] [Hong et al. Findings of NAACL 2022]
Ablations

Accuracies on Task 2 of EntailmentBank

<table>
<thead>
<tr>
<th></th>
<th>NLProofS (full)</th>
<th>w/o search</th>
<th>w/o search w/o stepwise</th>
<th>w/o verifier score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>58.8</td>
<td>56.5</td>
<td>45.6</td>
<td>55.8</td>
</tr>
<tr>
<td>Steps</td>
<td>34.4</td>
<td>33.7</td>
<td>29.7</td>
<td>33.8</td>
</tr>
<tr>
<td>Intermediates</td>
<td>37.8</td>
<td>36.4</td>
<td>32.2</td>
<td>36.1</td>
</tr>
<tr>
<td>Overall</td>
<td>33.3</td>
<td>31.8</td>
<td>27.1</td>
<td>31.9</td>
</tr>
</tbody>
</table>

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
### Ablations

#### Accuracies on Task 2 of EntailmentBank

<table>
<thead>
<tr>
<th></th>
<th>Leaves</th>
<th>Steps</th>
<th>Intermediates</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NLProofS (full)</strong></td>
<td>58.8</td>
<td>34.4</td>
<td>37.8</td>
<td>33.3</td>
</tr>
<tr>
<td>w/o search</td>
<td>56.5</td>
<td>33.7</td>
<td>32.2</td>
<td>31.8</td>
</tr>
<tr>
<td>w/o search w/o stepwise</td>
<td>45.6</td>
<td>29.7</td>
<td>36.1</td>
<td>27.1</td>
</tr>
<tr>
<td>w/o verifier score</td>
<td>55.8</td>
<td>33.8</td>
<td>36.4</td>
<td>31.9</td>
</tr>
</tbody>
</table>

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
Verifier-Guided Proof Search Is Helpful

Accuracies on Task 2 of EntailmentBank

- NLProofS (full)
- w/o search
- w/o search w/o stepwise
- w/o verifier score

Leaves: 58.8, 56.5, 55.8, 45.6
Steps: 34.4, 33.7, 29.7, 22.2
Intermediates: 37.8, 36.4, 32.2, 31.1
Overall: 33.3, 31.8, 27.1, 31.9
Stepwise Generation Is Helpful

Accuracies on Task 2 of EntailmentBank

<table>
<thead>
<tr>
<th></th>
<th>NLProofS (full)</th>
<th>w/o search</th>
<th>w/o search w/o stepwise</th>
<th>w/o verifier score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>58.8</td>
<td>56.5</td>
<td>45.6</td>
<td>55.8</td>
</tr>
<tr>
<td>Steps</td>
<td>34.4</td>
<td>33.7</td>
<td>29.7</td>
<td>33.8</td>
</tr>
<tr>
<td>Intermediates</td>
<td>37.8</td>
<td>36.4</td>
<td>32.2</td>
<td>36.1</td>
</tr>
<tr>
<td>Overall</td>
<td>33.3</td>
<td>31.8</td>
<td>27.1</td>
<td>31.9</td>
</tr>
</tbody>
</table>

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
The Verifier Is Necessary for Proof Search

Accuracies on Task 2 of EntailmentBank

- NLProofS (full)
- w/o search
- w/o search w/o stepwise
- w/o verifier score

<table>
<thead>
<tr>
<th></th>
<th>Leaves</th>
<th>Steps</th>
<th>Intermediates</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracies</td>
<td>58.8</td>
<td>34.4</td>
<td>37.8</td>
<td>33.3</td>
</tr>
<tr>
<td>NLProofS</td>
<td>56.5</td>
<td>33.7</td>
<td>36.4</td>
<td>31.8</td>
</tr>
<tr>
<td>w/o search</td>
<td>45.6</td>
<td>29.7</td>
<td>32.2</td>
<td>27.1</td>
</tr>
<tr>
<td>w/o verifier score</td>
<td>55.8</td>
<td>33.8</td>
<td>36.1</td>
<td>31.9</td>
</tr>
</tbody>
</table>
### Validation accuracies on Task 2 of EntailmentBank

<table>
<thead>
<tr>
<th></th>
<th>NLPProoFS (1B parameters)</th>
<th>GPT-3</th>
<th>Codex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>15.3</td>
<td>19.8</td>
<td></td>
</tr>
<tr>
<td>Steps</td>
<td>12.3</td>
<td>14.6</td>
<td></td>
</tr>
<tr>
<td>Intermediates</td>
<td>22.3</td>
<td>23.2</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>12.3</td>
<td>14.4</td>
<td></td>
</tr>
</tbody>
</table>

[Brown et al. NeurIPS 2020]  [Chen et al. arXiv 2021]
Key Takeaways

• The verifier is important
  • Prevent hallucinated generations
Key Takeaways

• The verifier is important
  • Prevent hallucinated generations
  • Also explored in other contexts, e.g., math word problems, code generation

[Cobbe et al. arXiv 2021] [Le and Wang et al. NeurIPS 2022]
Generating Natural Language Proofs with Verifier-Guided Search

Kaiyu Yang, Jia Deng, Danqi Chen

https://github.com/princeton-nlp/NLProofS