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

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
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

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>homes are buildings</td>
<td>solar is a kind of renewable energy for heating homes</td>
</tr>
<tr>
<td>energy is used for heating buildings</td>
<td></td>
</tr>
<tr>
<td>solar is a kind of energy</td>
<td></td>
</tr>
<tr>
<td>solar is renewable</td>
<td></td>
</tr>
</tbody>
</table>

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

Hypothesis ($h$):

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

Supporting facts ($C$):

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

\[ \text{...} \]

\[ \text{...} \]

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

Hypothesis \((h)\):

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

Supporting facts \((C)\):

\textit{sent}1: homes are buildings
\textit{sent}2: solar is renewable
\textit{sent}3: wind is a kind of energy
\textit{sent}4: solar is a kind of energy
\textit{sent}5: energy is used for heating buildings
\textit{sent}6: coal is nonrenewable
...
...

[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
Task: Proof Generation

Hypothesis ($h$):

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

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 ($T'$):

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

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

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

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

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

\[ \text{sent2: solar is renewable} \]

[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

[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

Generated by the model

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

Input

Output
Hypothesis ($h$):

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

Supporting facts ($C$):

- $sent_1$: homes are buildings
- $sent_2$: solar is renewable
- $sent_3$: wind is a kind of energy
- $sent_4$: solar is a kind of energy
- $sent_5$: energy is used for heating buildings
- $sent_6$: coal is nonrenewable
  ...

Proof tree ($T'$):

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

- $int_1$: energy is used for heating homes
- $int_2$: solar is a kind of renewable energy

- $sent_1$: homes are buildings
- $sent_5$: energy is used for heating buildings
- $sent_4$: solar is a kind of energy
- $sent_2$: solar is renewable
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

Input

Output

Generate the proof step by step
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
  - $sent5$: energy is used for heating buildings
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

Step 2

$int2$: solar is a kind of renewable energy

Step 3

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

sent1: homes are buildings
sent5: energy is used for heating buildings
sent4: solar is a kind of energy
sent2: solar is renewable
### Single-Shot vs. Stepwise Methods

<table>
<thead>
<tr>
<th>Generate the entire proof altogether</th>
<th>Generate the proof step by step</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRover [Saha et al. EMNLP 2020]</td>
<td>ProofWriter [Tafjord et al. Findings of ACL 2021]</td>
</tr>
<tr>
<td>EntailmentWriter [Dalvi et al. EMNLP 2021]</td>
<td>FaiRR [Sanyal et al. ACL 2022]</td>
</tr>
<tr>
<td>MetGen [Hong et al. Findings of NAACL 2022]</td>
<td></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):

- 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
...
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)
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)
Challenges in Generating **Valid** and **Relevant** Steps

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

![Diagram](image)

- **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
- **sent2**: solar is renewable
- **sent4**: solar is a kind of energy
- **sent5**: energy is used for heating buildings

[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

---

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

---

[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

![Diagram showing proof steps:]

- **sent1**: homes are buildings
- **sent2**: solar is renewable
- **send5**: energy is used for heating buildings
- **sent4**: solar is a kind of energy
- **int1**: energy is used for heating homes
- **h**: solar is a kind of renewable energy for heating homes
- **int2**: solar is a kind of renewable energy
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

---

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*

```
sent1: homes are buildings

int1: energy is used for heating homes

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

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
A new method for stepwise proof generation

Prover

Generate candidate proof steps

\[\begin{align*}
\text{sent1: homes are buildings} \\
\text{int1: energy is used for heating homes} \\
\text{solar is a kind of renewable energy for heating homes} \\
\text{sent4: solar is a kind of energy} \\
\text{sent5: energy is used for heating buildings} \\
\end{align*}\]
NLProofS: Natural Language Proof Search

- A new method for stepwise proof generation

Prover

Generate candidate proof steps

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

- A new method for stepwise proof generation

Prover

Verifier

Score the validity

\[ \cdots \quad 0.8 \quad 0.4 \quad 0.6 \]

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

Score the validity
NLProofS: Natural Language Proof Search

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

---

**Prover**

**Verifier**

---

**Training**

---

**sent1**: homes are buildings

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

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

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

**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
NLProofS: Natural Language Proof Search

- A new method for stepwise proof generation

Training

Proof search

Inference

Prover

Verifier

sent1: homes are buildings

sent4: solar is a kind of energy

int1: energy is used for heating homes

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

• A new method for stepwise proof generation

Prover

Verifier

Proof search

Training

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

0.6
0.8
0.4
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 ...

The 2nd step

**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
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$^{nd}$ step

$int2$: solar is a kind of renewable energy

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

• Finetune a T5 model to predict the next proof step

[Raffle et al. JMLR 2020] [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

...  

Encode input/output as text sequences

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

...
NLProofS: Natural Language Proof Search

Prover
Verifier
Proof search
Inference
Training

sent1: 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

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

0.6...0.4...0.8
NLProofS: Natural Language Proof Search
NLProofS: Natural Language Proof Search

Prover

Verifier

Training

Proof search

Inference

Finetune RoBERTa
[Liu et al. arXiv 2019]

$h$: 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

$sent4$: 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

Aggregate the step scores to across the entire proof

- **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

Proof search

Training

Prover

Verifier

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

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

Training

Proof search

Inference

**sent1**: homes are buildings

**sent4**: solar is a kind of renewable energy for heating homes

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

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

**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

Training

Verifier

Prover

Proof search

Inference

- **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

1. Initialization: a proof generated by the prover alone

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
1. Initialization: a proof generated by the prover alone
2. Iteration: expand the graph iteratively
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
- **sent3**: energy is used for heating buildings
- **sent4**: solar is a kind of energy
- **sent5**: energy is used for heating homes

- **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
sent2: solar is renewable
sent4: solar is a kind of energy
sent5: energy is used for heating buildings
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

---

Diagram:

- **Proof search**
  - **Initialization**:
    - **sent1**: homes are buildings
    - **sent5**: energy is used for heating buildings
  - **Iteration**:
    - **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 Graph**:

- **0.5**
  - **sent1**: homes are buildings
  - **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
   • Using steps proposed by the prover
   • Checked by the verifier
   • Average verifier/prover scores
3. Extraction: proof tree with best score

Proof search

1. homes are buildings
2. solar is renewable
3. energy is used for heating homes
4. solar is a kind of energy
5. energy is used for heating buildings
6. solar is a kind of renewable energy

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

[Tafjord et al. Findings of ACL 2021]
[Dalvi et al. EMNLP 2021]
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]

**Predicted proof tree**

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

**Ground truth proof tree**

- **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
  - **sent4:** solar is a kind of energy
- **int2:** solar is a kind of renewable energy
  - **sent2:** solar is renewable

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

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

- EntailmentWriter
- IGRG
- MetGen
- NLProofS (ours)

Leaves

Steps

Intermediates

Overall

[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

Single-shot

<table>
<thead>
<tr>
<th>Method</th>
<th>Leaves</th>
<th>Steps</th>
<th>Intermediates</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>EntailmentWriter</td>
<td>58.8</td>
<td>30.4</td>
<td>26.5</td>
<td>28.8</td>
</tr>
<tr>
<td>IGRG</td>
<td>48.6</td>
<td>32.7</td>
<td>22.3</td>
<td>24.4</td>
</tr>
<tr>
<td>MetGen</td>
<td>48.6</td>
<td>34.4</td>
<td>22.9</td>
<td>27.8</td>
</tr>
<tr>
<td>NLProofS (ours)</td>
<td>35.6</td>
<td>28.5</td>
<td>20.9</td>
<td>23.3</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>Method</th>
<th>Leaves</th>
<th>Steps</th>
<th>Intermediates</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>EntailmentWriter</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>28</td>
</tr>
<tr>
<td>NLProofS (ours)</td>
<td>56</td>
<td>34.4</td>
<td>33.3</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>Method</th>
<th>Leaves</th>
<th>Steps</th>
<th>Intermediates</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>EntailmentWriter</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>58.8</td>
<td>34.4</td>
<td>37.8</td>
<td>33.3</td>
</tr>
</tbody>
</table>

Source:
- [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)

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

Leaves: 58.8
Steps: 30.4
Intermediates: 37.8
Overall: 33.3

12.4
State-of-the-art Performance on EntailmentBank

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>EntailmentWriter</td>
<td>35.6</td>
<td>30.4</td>
<td>28.5</td>
<td>20.9</td>
</tr>
<tr>
<td>IGRG</td>
<td>48.6</td>
<td>34.4</td>
<td>32.7</td>
<td>22</td>
</tr>
<tr>
<td>MetGen</td>
<td>58.8</td>
<td>23.2</td>
<td>37.8</td>
<td>28</td>
</tr>
<tr>
<td>NLProofS (ours)</td>
<td>22.9</td>
<td>22.3</td>
<td>37.8</td>
<td>28</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></th>
<th>Leaves</th>
<th>Steps</th>
<th>Intermediates</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>EntailmentWriter</td>
<td>35.6</td>
<td>22.9</td>
<td>30.4</td>
<td>28.5</td>
</tr>
<tr>
<td>IGRG</td>
<td>48.6</td>
<td>22.3</td>
<td>34.4</td>
<td>32.7</td>
</tr>
<tr>
<td>MetGen</td>
<td>58.8</td>
<td>10.2</td>
<td>4.0</td>
<td>5.1</td>
</tr>
<tr>
<td>NLProofS (ours)</td>
<td>60.2</td>
<td>5.3</td>
<td>5.3</td>
<td>5.3</td>
</tr>
</tbody>
</table>

[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

- 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></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLProofS (full)</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>36.4</td>
<td>31.8</td>
</tr>
<tr>
<td>w/o search w/o stepwise</td>
<td>55.8</td>
<td>33.8</td>
<td>36.1</td>
<td>31.9</td>
</tr>
<tr>
<td>w/o verifier score</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ablations

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></td>
<td>56.5</td>
<td>33.7</td>
<td>36.4</td>
<td>31.8</td>
</tr>
<tr>
<td></td>
<td>55.8</td>
<td>29.7</td>
<td>32.2</td>
<td>27.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33.8</td>
<td>36.1</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 45.6 55.8
Steps: 34.4 33.7 29.7 33.8
Intermediates: 37.8 36.4 32.2 36.1
Overall: 33.3 31.8 27.1 31.9
Stepwise Generation Is Helpful

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>NLProofS (full)</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>36.4</td>
<td>31.8</td>
</tr>
<tr>
<td>w/o search w/o stepwise</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>
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

Leaves
Steps
Intermediates
Overall

58.8 56.5 55.8
45.6

34.4 33.7 33.8
29.7

37.8 36.4 36.1
32.2

33.3 31.8 27.1
31.9

Generating Natural Language Proofs with Verifier-Guided Search - Kaiyu Yang, Jia Deng, Danqi Chen
Validation accuracies on Task 2 of EntailmentBank

<table>
<thead>
<tr>
<th></th>
<th>NLProofS (1B parameters)</th>
<th>GPT-3 (175B parameters)</th>
<th>Codex (12B parameters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>56</td>
<td>38.4</td>
<td>41.3</td>
</tr>
<tr>
<td>Steps</td>
<td>15.3</td>
<td>12.3</td>
<td>22.3</td>
</tr>
<tr>
<td>Intermediates</td>
<td>19.8</td>
<td>14.6</td>
<td>23.2</td>
</tr>
<tr>
<td>Overall</td>
<td>12.3</td>
<td>14.4</td>
<td>37.1</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