

Robert Vacareanu, Marco A. Valenzuela-Escárcega, George C. G. Barbosa, Rebecca Sharp, Gus Hahn-Powell, Mihai Surdeanu

Computational Language Understanding Lab at the University of Arizona

1. Introduction

Summary

While deep learning approaches to information extraction have had many successes, they can be difficult to augment or maintain as needs shift. Rule-based methods, on the other hand, can be more easily modified. However, crafting rules requires expertise in linguistics and the domain of interest, making it infeasible for most users. Here we attempt to combine the advantages of these two directions while mitigating their drawbacks. We adapt recent advances from the adjacent field of program synthesis to information extraction, synthesizing rules from provided examples. We use a transformer-based architecture to guide an enumerative search, and show that this reduces the number of steps that need to be explored before a rule is found. Further, we show that *without training the synthesis algorithm on the specific domain*, our synthesized rules achieve state-of-the-art performance on the 1-shot scenario of a task that focuses on few-shot learning for relation classification, and competitive performance in the 5-shot scenario.

Proposed Method

- Enumerative Searcher + Neural Scorer
- The Neural Scorer gives the likelihood of a transition to be on the gold path
- The Enumerative Searcher repeatedly calls the Neural Scorer and chooses the most promising transition, while pruning the search space

Terminology

- Specification → (Sentence, Highlight) tuple
(She is an American actress and singer, is an American))
- Rules → An Odinson pattern ([lemma=be] [tag=DT] [word=American])
- Placeholder → □, our start symbol
- State → Current intermediate rule, together with what is matched

3. Evaluation

Results Few-Shot TACRED

model	5-way 1-shot	5-way 5-shot
Baseline	10.82 ± 0.01%	10.90 ± 0.01%
Sentence-Pair	10.19 ± 0.81%	-
Threshold	6.87 ± 0.48%	13.57 ± 0.46%
NAV	8.38 ± 0.80%	18.38 ± 2.01%
MNAV	12.39 ± 1.01%	30.04 ± 1.92%
Ours	15.40 ± 1.21%	24.16 ± 0.44%

Examples of Synthesized Rules

Span

ORGANIZATION, which is based in CITY

Synthesized surface rule:

[entity=organization] [tag=","] [tag=WDT] [tag=VBZ] [tag=VBN] [tag=IN] [entity=city]

Synthesized simplified syntax rule:

[entity=organization] [tag=NN] [word=based] [entity=city]

Span

ORGANIZATION, which represents ORGANIZATION

Synthesized surface rule:

[entity=organization] [tag=","] [word=which] [tag=VBZ] [entity=organization]

Synthesized simplified syntax rule:

[entity=organization] [lemma=represent] [entity=organization]

4. Conclusion

- We can generate explainable models directly without going through the hoops of training a neural architecture for classification
- Synthesis generalizes better than Neural Networks with little training data
- Our code and pre-trained models are available <https://github.com/clulab/releases/tree/master/lrec2022-odinsynth>

Disclosure

Marco A. Valenzuela-Escárcega, Gus Hahn-Powell, and Mihai Surdeanu declare a financial interest in lum.ai, which licenses the intellectual property involved in this research. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

2. Algorithm

