nl2bash

Expanding and Improving Existing Datasets

Oisin Doherty (oisind),
Abhinav Gottumukkala (anak4569),
Hans Jorgensen (thehans),
Lauren Martini (lmartini)

Motivation

- Bash can be difficult to learn
  - Commands are often unintuitive
- Even experienced users often have to look commands up

Example:

English: “Copy the first line from each text file in this directory into new_file.txt”

Bash: head -n1 *.txt > new_file.txt

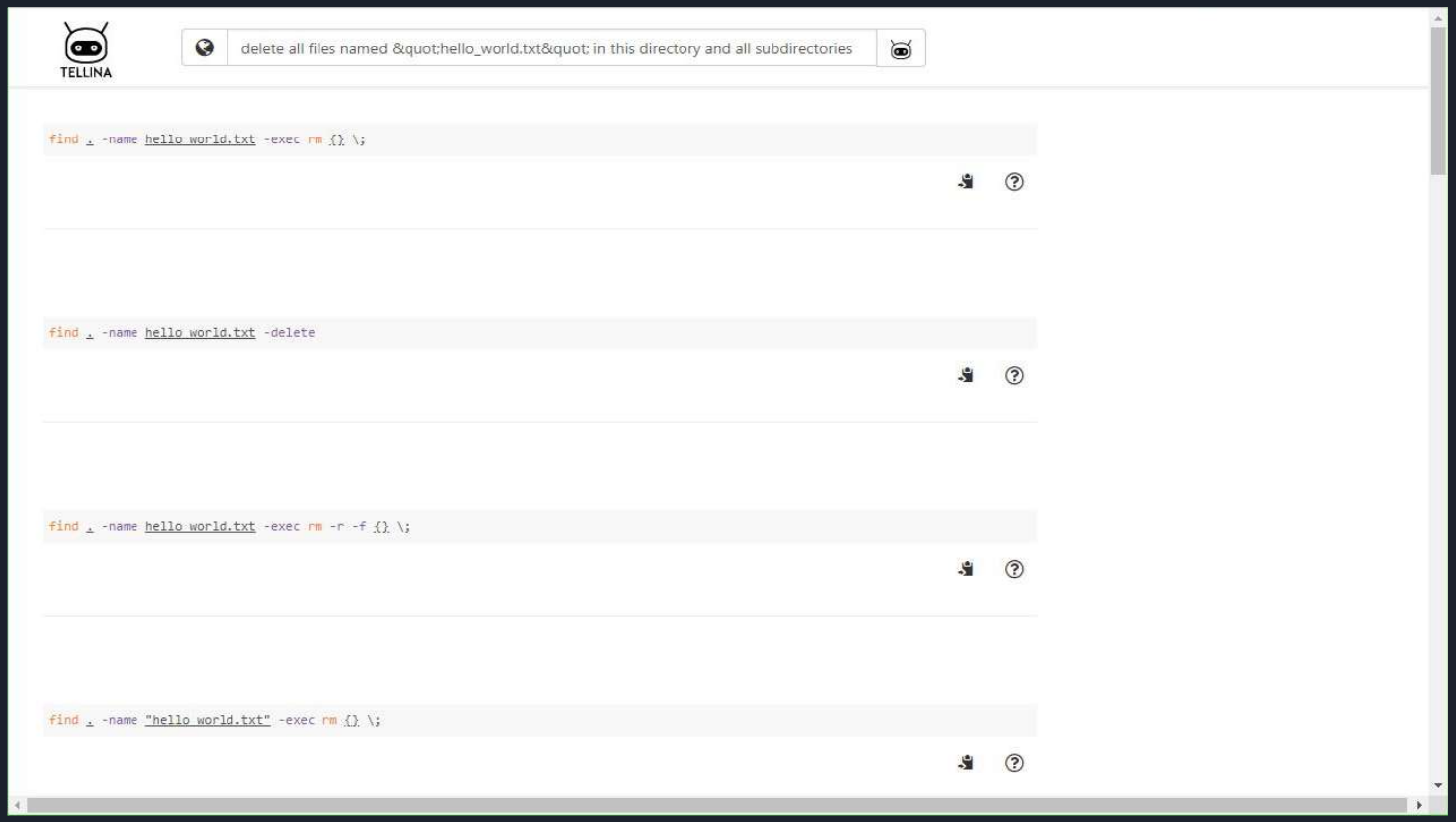
Existing Architecture

- Tellina
  - Uses machine learning for natural language processing
  - Accuracy shown to correlate with its training dataset size
  - Currently not very accurate:

English: “Remove the first line from each text file in this directory”

Tellina Output: find . -type f -exec grep California {} \; -exec rm {} \;

Tellina Interface



How Tellina's Current Dataset was Generated

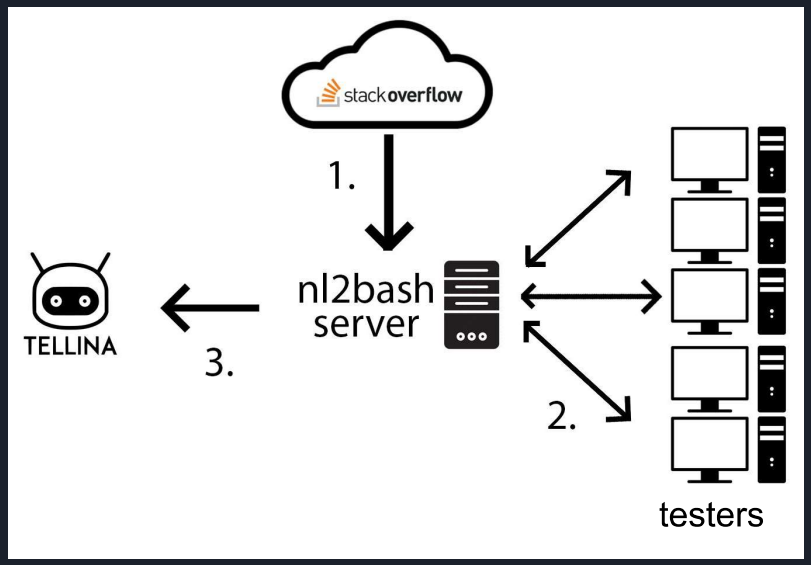
- Automated cleaning and filtering
  - Remove syntactically incorrect commands
  - Remove out-of-scope commands
  - Correct spelling errors in English commands

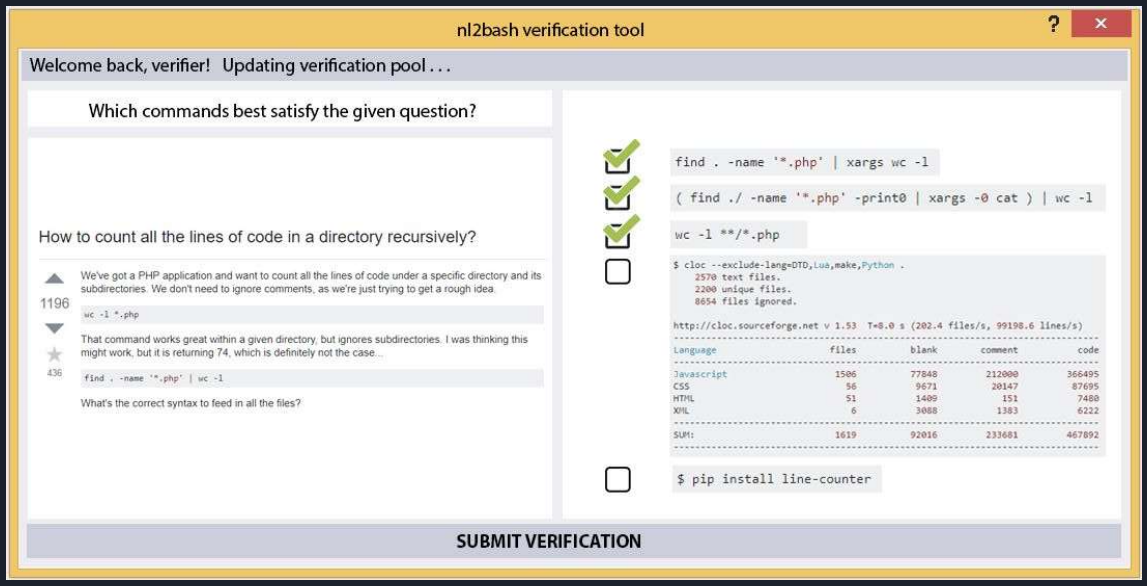
- Data collected by hand → 12,000 cleaned commands given descriptions generated by experts

Current Directions and Goals

- Main goal: **Improve the accuracy of Tellina**
- Directions:
  - **Expand Tellina's dataset**
    - Scrape resources such as Github and StackOverflow for English/Bash command pairs
    - Clean and verify this data - crowdsourcing
  - **Improve Tellina's dataset**
    - Develop better cleaning and verification tools to use on the existing dataset - crowdsourcing

Proposed Architecture and Interface





Proposed Implementation

- Web scraping tool to get bash commands and nearby descriptions (not expert written descriptions)
- Build on data cleaning tools to
  - Match Tellina's spec (no redirection, control flow, etc)
- UI for Testers (like mockup)
- Setup personal Tellina server using their repo's instructions
  - System to automate updating the model with new data

Evaluation

- Metrics
  - Translation accuracy - percentage of English/Bash pairs marked as correct
  - Comparisons against old models - Seq2Seq and CopyNet
- Goal:
  - More accurate translation than the current version of Tellina (exact stats to beat detailed in paper)

Conclusion

Improve and expand Tellina's dataset

↓

Increase the accuracy of Tellina's command pair generation

↓

Make it easier for people to learn and use Bash

Thank you! Any questions?

References

1. Lin, X. V., Wang, C., Zettlemoyer, L., & Ernst, M. D. (2018). NL2Bash: A Corpus and Semantic Parser for Natural Language Interface to the Linux Operating System. arXiv preprint arXiv:1802.08979.

2. Lin, X. V., Wang, C., Pang, D., Vu, K., & Ernst, M. D. (2017). Program synthesis from natural language using recurrent neural networks (Vol. 2). Technical Report UW-CSE-17-03-01, University of Washington Department of Computer Science and Engineering, Seattle, WA, USA.