

Comparing Natural Language Embeddings for Libc Functions as Rich Labels

Bachelor defense

Ruben Triwari

Ludwig Maximilian University Munich

19, February 2025

Ruben Triwari 1 / 12

Outline

Motivation & Research Objective

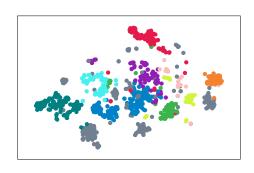
State Of The Art

Methodology

Results

Limitations

Conclusion & Future Work



Ruben Triwari $2\ /\ 12$

Motivation & Research Objective

"I had a lot of fun NLP-Model writing my bachelor thesis!"

- → Encoding natural language was a huge factor in recent nlp advancements
- → Information described as a vector can be used in many downstream task
- → That motivates encoding binary code and describing them as a vector
- → That motivates using NLP tools to encode binary code

Ruben Triwari 3 / 12

Motivation

Motivation & Research Objective

```
factorial(int):
                                                                                  push
                                                                                           rbp
                                                                                            rbp, rsp
                                                                                   mov
                                                                                           rbx
                                                                                   push
                                                                                   sub
                                                                                            rsp, 24
                                                                                            DWORD PTR [rbp-20], edi
                                                                                   mov
                                                                                            DWORD PTR [rbp-20]. 0
                                                                                   cmp
unsigned long long factorial(int n) {
                                                                                   ine
                                                                                            .L2
  // Base case: Factorial of 0 is 1
                                                                                   mov
                                                                                            eax, 1
  if (n == 0) {
                                                                                   jmp
                                                                                            .L3
    return 1:
                                                    Compiler
                                                                                 L2:
                                                                                            eax, DWORD PTR [rbp-20]
                                                                                   mov
  // Recursive case: n! = n * (n-1)!
                                                                                           rbx. eax
                                                                                   movsx
  return n * factorial(n - 1):
                                                                                            eax, DWORD PTR [rbp-20]
                                                                                  mov
                                                                                   sub
                                                                                            eax, 1
                                                                                   mov
                                                                                            edi. eax
                                                                                           factorial(int)
                                                                                   call
                                                                                   imul.
                                                                                           rax, rbx
                                                                                .L3:
                                                                                            rbx, QWORD PTR [rbp-8]
                                                                                  mov
                                                                                   leave
                                                                                   ret
```

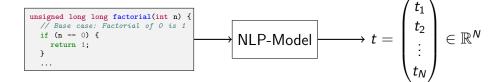
→ Compiler removes important inforamtion in natural language

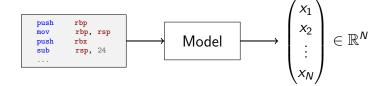
Ruben Triwari 4 / 12

Motivation & Research Objective

Motivation

Idea:





Ruben Triwari

Ruben Triwari

Ruben Triwari

Curry-Howard Correspondance

Ruben Triwari 8 / 12

Denotational Semantics

Ruben Triwari 9 / 12

HALLO

Ruben Triwari 10 / 12

HALLO

Ruben Triwari 11 / 12

Discussion

Ruben Triwari 12 / 13