

Глубинное обучение для работы с кодом

Корягин Никита

Генерация кода (Code Synthesis)

Постановка задачи

Given a list of citations counts, where each citation is a nonnegative integer, **write a function h_index that outputs the h-index**. The h-index is the largest number h such that h papers have each least h citations.

$[3,0,6,1,4] \rightarrow 3$
 $[1,4,1,4,2,1,3,5,6] \rightarrow 4$
 $[1, 0] \rightarrow 1$
 $[1000,500,500,250,100,100,100,100,75,50,30,20,15,15,10,5,2,1] \rightarrow 15$



```
def h_index(counts):  
    n = len(counts)  
    if n > 0:  
        counts.sort()  
        counts.reverse()  
        h = 0  
        while (h < n and  
               counts[h]-1>=h):  
            h += 1  
        return h  
    else:  
        return 0
```

(may be one type of input or both)

Генерация кода (Code Synthesis)

Какие бывают датасеты

CONCODE dataset

```
public class SimpleVector implements Serializable {  
    double[] vecElements;  
    double[] weights;  
  
    NL Query: Adds a scalar to this vector in place.  
    Code to be generated automatically:  
    public void add(final double arg0) {  
        for (int i = 0; i < vecElements.length; i++) {  
            vecElements[i] += arg0;  
        }  
    }  
  
    NL Query: Increment this vector  
    Code to be generated automatically:  
    public void inc() {  
        this.add(1);  
    }  
}
```

Размер: 100К Источник: GitHub

- Язык Java
- Есть контекст
- Содержит базовые операции
- Много плохо размеченных сэмплов

APPS dataset

Problem

H-Index

Given a list of citations counts, where each citation is a nonnegative integer, write a function `h_index` that outputs the h-index. The h-index is the largest number h such that h papers have each least h citations.

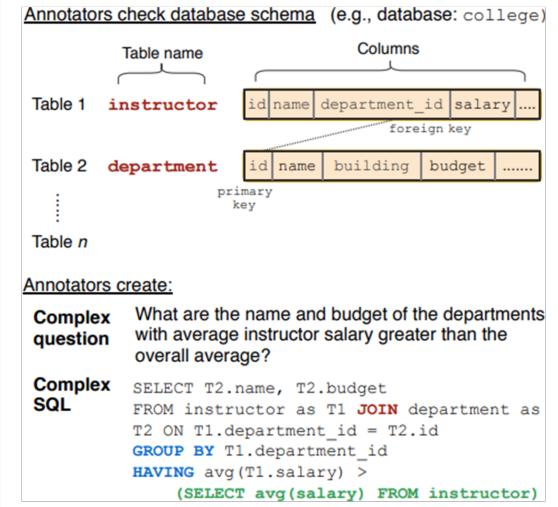
Example:

Input: [3,0,6,1,4]
Output: 3

Generated Code

```
def h_index(counts):  
    n = len(counts)  
    if n > 0:  
        counts.sort()  
        counts.reverse()  
        h = 0  
        while (h < n and  
               counts[h]-1>=h):  
            h += 1  
        return h  
    else:  
        return 0
```

Spider dataset



Размер: 10К задач / 260К программ
Источник: сайт типа Codeforces

- Язык Python
- Есть контекст
- Три уровня сложности

Размер: 10К запросов / 200 баз данных

- Язык SQL
- Есть контекст(схема)
- Базовые операции

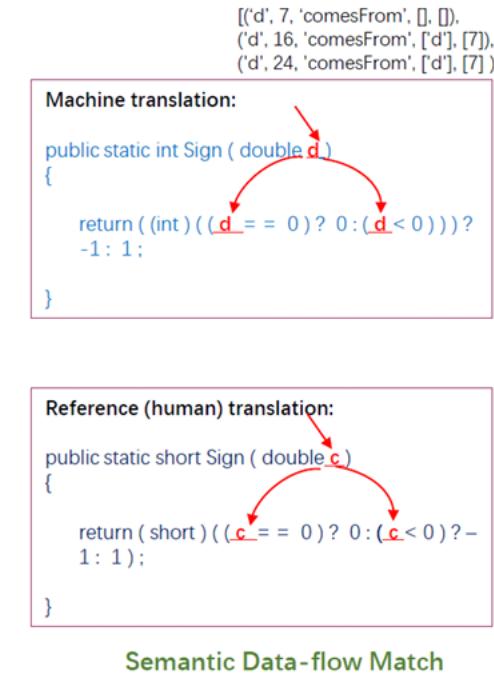
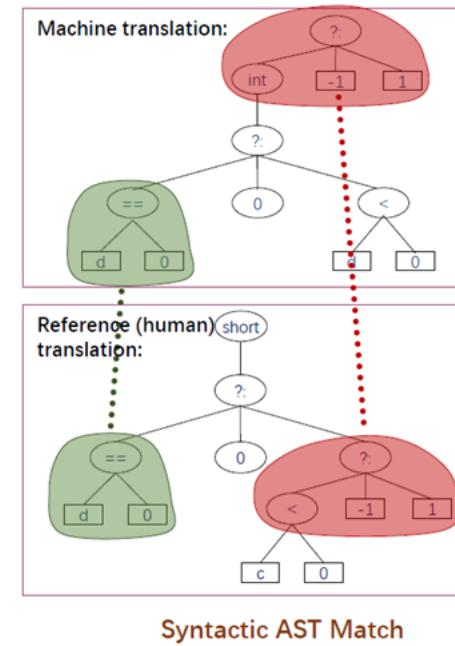
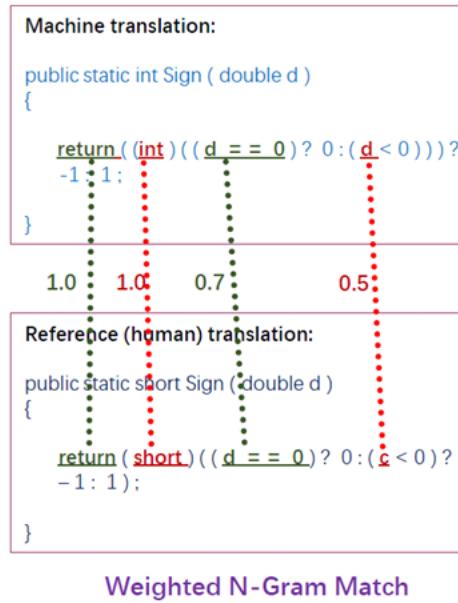
Генерация кода (Code Synthesis)

Какие бывают метрики

- **Exact match (EM):** сравниваем посимвольно – очевидно, плохая метрика
- **BLEU:** сравниваем код как текст с помощью n-грамм. Тоже плохая метрика
- **CodeBLEU:** сравниваем код на уровне токенов, абстрактного синтаксического дерева и потока информации (DataFlow). Хорошая метрика
- **Запуск кода на тестах:** лучшая метрика, но мало тестов

Генерация кода (Code Synthesis)

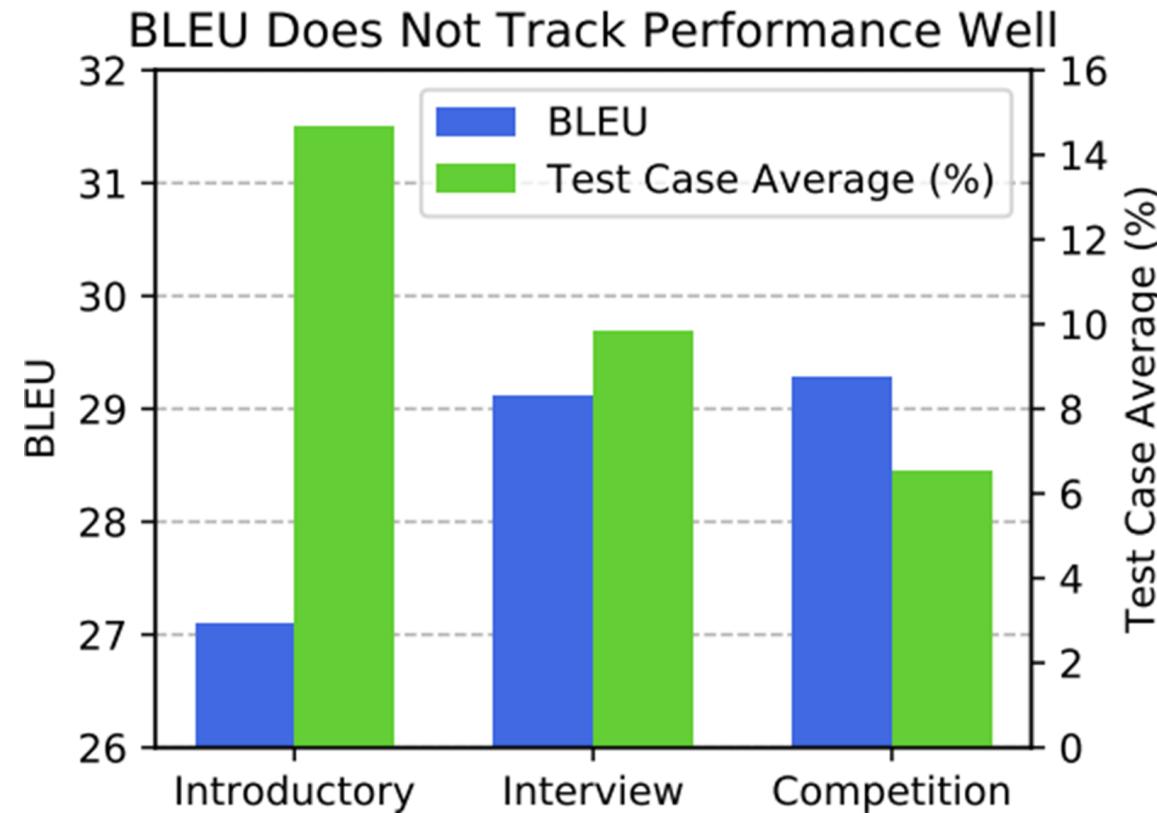
CodeBLEU



$$\text{CodeBLEU} = \alpha \cdot \text{N - Gram Match (BLEU)} + \beta \cdot \text{Weighted N-Gram Match} + \gamma \cdot \text{Syntactic AST Match} + \delta \cdot \text{Semantic Data-flow Match}$$

Генерация кода (Code Synthesis)

BLEU плохо показывает качество программы



Генерация кода (Code Synthesis)

Нерешенные задачи

Результаты GPT на APPS

Model	Test Case Average					Strict Accuracy			
	Introductory	Interview	Competitive	Average		Introductory	Interview	Competition	Average
GPT-2 0.1B	5.64	6.93	4.37	6.16		1.00	0.33	0.00	0.40
GPT-2 1.5B	7.40	9.11	5.05	7.96		1.30	0.70	0.00	0.68
GPT-Neo 2.7B	14.68	9.85	6.54	10.15		3.90	0.57	0.00	1.12
GPT-3 175B	0.57	0.65	0.21	0.55		0.20	0.03	0.00	0.06

Предобучена на простом тексте(без fine-tuning)

Предобучены на тексте и дообучены на коде с GitHub и fine-tuned на APPS

Code2Vec

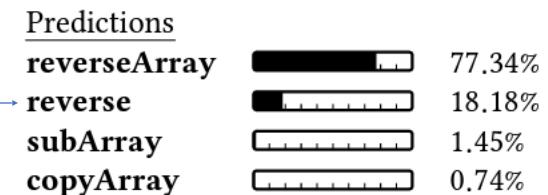
Идея

Что хотим: code embeddings – похожие куски кода семантически
кода отображаются в похожие вектора

Как будем это делать: обучим модель, которая по коду функции
определяет ее название

Главная идея: даем на вход модели синтаксические пути

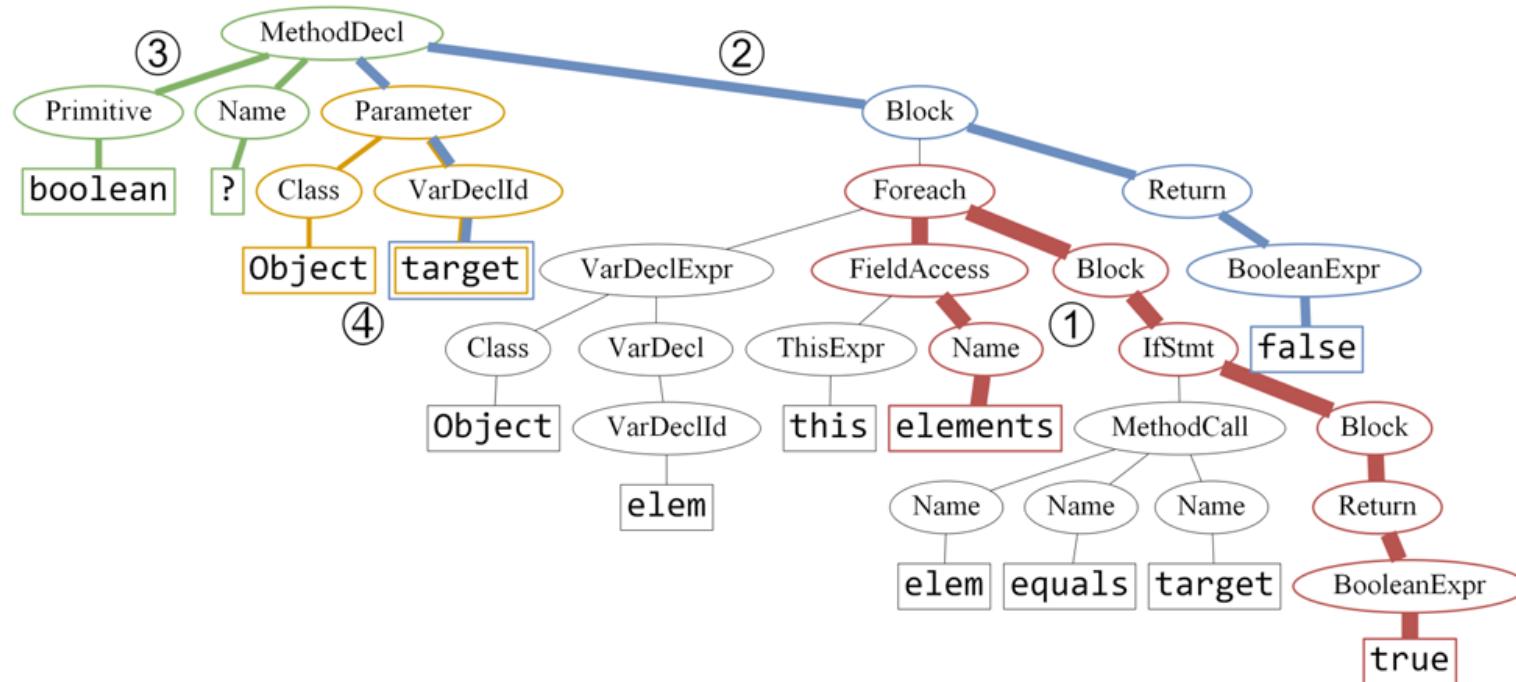
```
String[] f(final String[] array) {  
    final String[] newArray = new String[array.length];  
    for (int index = 0; index < array.length; index++) {  
        newArray[array.length - index - 1] = array[index];  
    }  
    return newArray;  
}
```



Code2Vec

Пути

Тройки вида (начальный лист, путь, конечный лист) – path-context



Красный path-context = (elements, Name↑FieldAccess↑Foreach↓Block↓IfStmt↓Block↓Return↓BooleanExpr, true)

Code2Vec

Почему пути

- Легко представить код в виде путей
- Модели будет легче различать похожие по синтаксису, но разные по семантике куски кода
- Можно интерпретировать результат

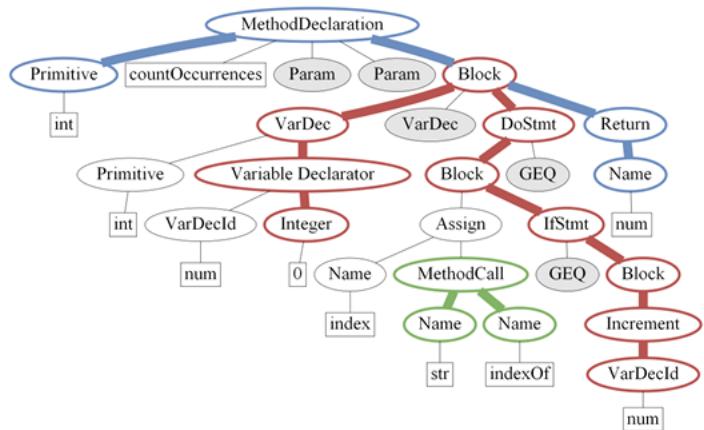
Code2Vec

Почему пути

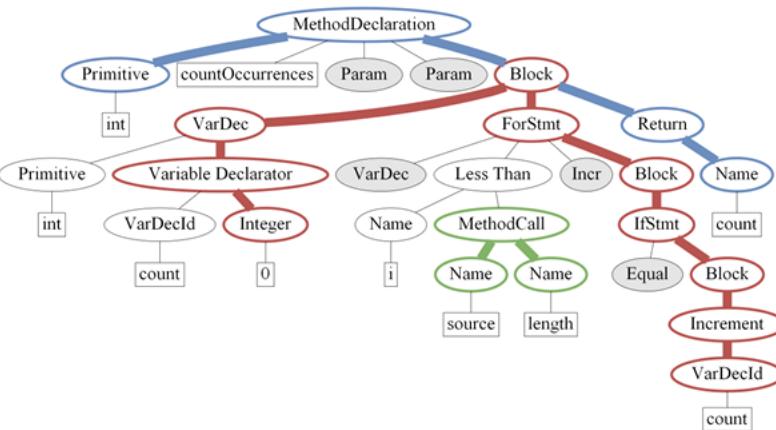
```
int countOccurrences(String str, char ch) {  
    int num = 0;  
    int index = -1;  
    do {  
        index = str.indexOf(ch, index + 1);  
        if (index >= 0) {  
            num++;  
        }  
    } while (index >= 0);  
    return num;  
}
```

```
int countOccurrences(String source, char value) {  
    int count = 0;  
    for (int i = 0; i < source.length(); i++) {  
        if (source.charAt(i) == value) {  
            count++;  
        }  
    }  
    return count;  
}
```

(a)



(b)

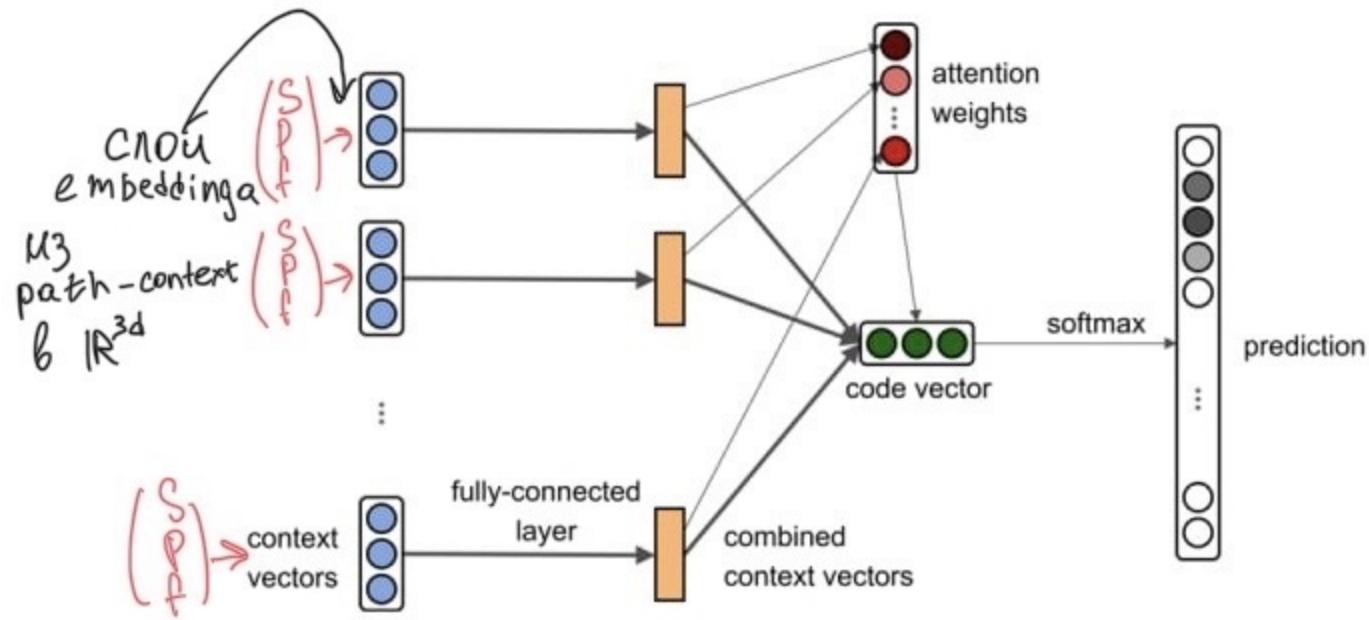


(c)

(d)

Code2Vec

Модель



По слоям:

Первый embedding слой
переводит все path-context в
context vector размерности 3d
Тут обучаем матрицу перевода
вершины в вектор и матрицу
перевода пути в вектор

Применяем FC
слой(одинаковый!) для
каждого context вектора

Применяем функцию
активации и attention слой

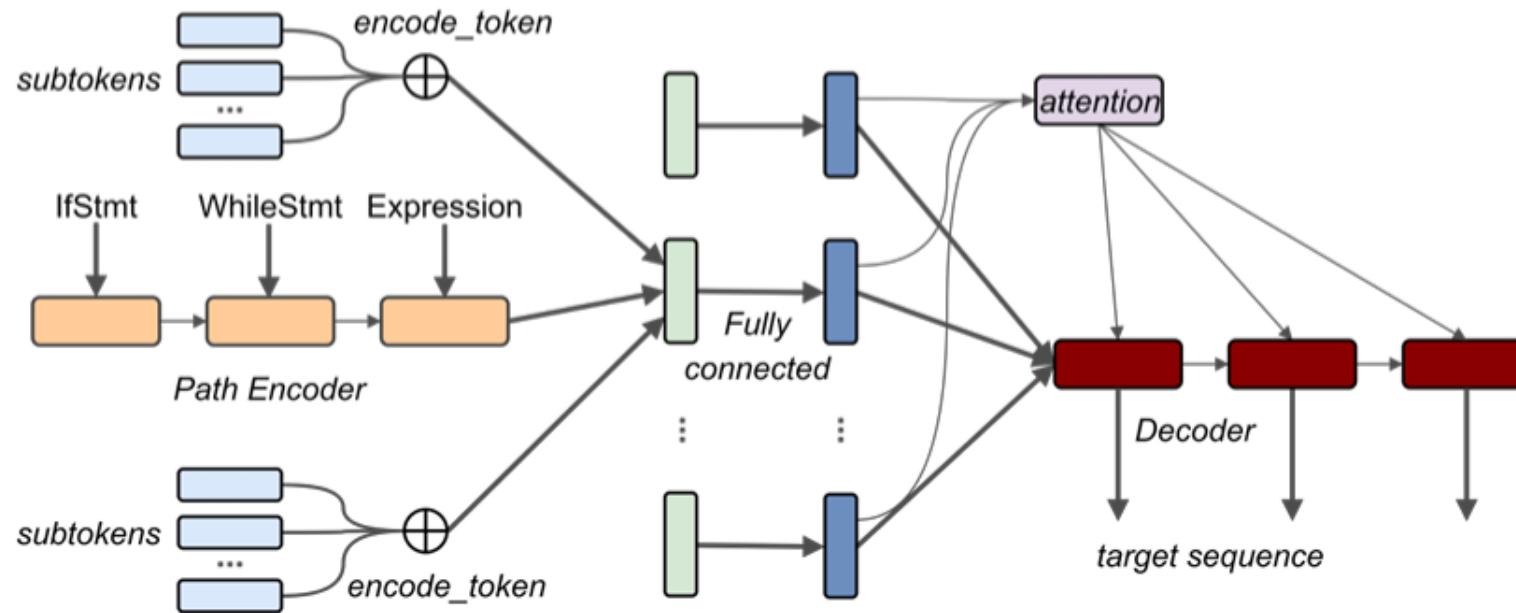
$$\text{attention weight } \alpha_i = \frac{\exp(\tilde{c}_i^T \cdot a)}{\sum_{j=1}^n \exp(\tilde{c}_j^T \cdot a)}$$

$$\text{code vector } v = \sum_{i=1}^n \alpha_i \cdot \tilde{c}_i$$

Code2Seq

Модель

Идеально то же, что и Code2Vec



Применения Code2Seq

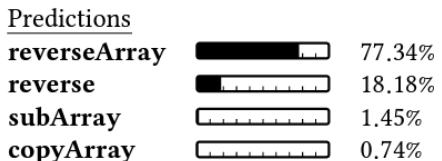
Code Summarization

Хотим дать название функции по ее телу

Dataset CodeSearchNet

2КК функций с названиями и аннотациями
(Source:GitHub)

```
String[] f(final String[] array) {
    final String[] newArray = new String[array.length];
    for (int index = 0; index < array.length; index++) {
        newArray[array.length - index - 1] = array[index];
    }
    return newArray;
}
```



Code Translation

Хотим перевести код с одного языка на другой

Dataset CodeXGLUE

10К функций, каждая из которых написана на Java и C#

```
public class JavaCode
{
    public static void main(String []args)
    {
        System.out.println("Hello, World!");
    }
}
```

```
using System;
public class JavaCode
{
    public static void Main(String[] args)
    {
        Console.WriteLine("Hello, World!");
    }
}
```