# Summarizing Source Code using a Neural Attention Model

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Luke Zettlemoyer
University of Washington, Seattle, USA

## Overview

Our task is to generate high-level summaries of the function of programming source code.

| | | |
|---|---|---|
| C# | `var x = "FundList[10].Amount";`<br>`int xIndex = Convert.ToInt32(`<br>`  Regex.Match(x,@"\d+").Value);` | Identify the number in given string |
| | `foreach (string parentText in xml.parent){`<br>`   TreeNode parent = new TreeNode();`<br>`   foreach (string childText in xml.child){`<br>`      TreeNode child = new TreeNode();`<br>`     parent.Nodes.Add(child); }}` | Adding childs to a tree node dynamically in C# |
| | `string url = baseUrl +`<br>`  "/api/Entry/SendEmail?emailId=" + emailId;`<br>`WebRequest req = WebRequest.Create(url);`<br>`req.Method = "GET";`<br>`req.BeginGetResponse(null, null);` | Execute a get request on a web server and receive the response asynchronously |
| SQL | `SELECT * FROM table`<br>`ORDER BY RANDOM() LIMIT 10;` | Select random rows from mysql table |
| | `SELECT GROUP_CONCAT(`<br>`  CONCAT_WS(',', PlayerId, R1))`<br>`FROM (`<br>`  SELECT PlayerId, SUM(Rank=1) R1`<br>`  FROM Result GROUP BY PlayerId)` | Get sum of group values based on condition and concatenate them into a string |

These summaries have many SE applications:

Code Search    Documentation    Code Navigation

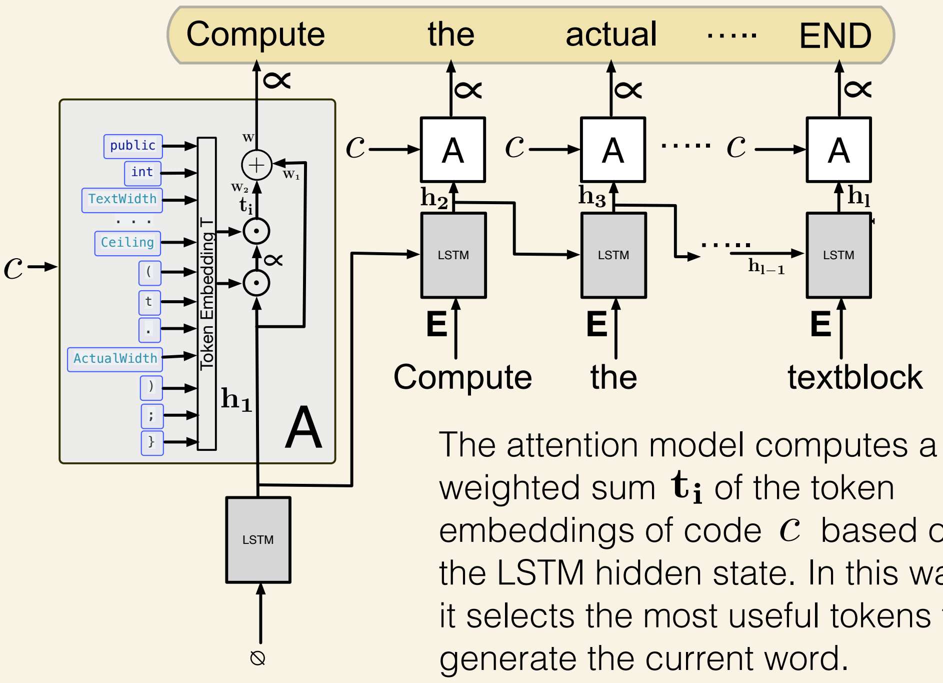
## Neural Attention Model

We use an end-to-end model that jointly performs content selection using an attention mechanism, and surface realization using Long Short Term Memory networks.

We model the conditional next-word probability as:

$$p(n_i|n_1,\dots,n_{i-1}) \propto \mathbf{W}\ \tanh(\mathbf{W_1 h_i} + \mathbf{W_2 t_i})$$

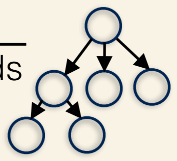
$\mathbf{h_i}$ is the hidden state of the LSTM cell at the time step i



The attention model computes a weighted sum $\mathbf{t_i}$ of the token embeddings of code $c$ based on the LSTM hidden state. In this way, it selects the most useful tokens to generate the current word.

## Future Work

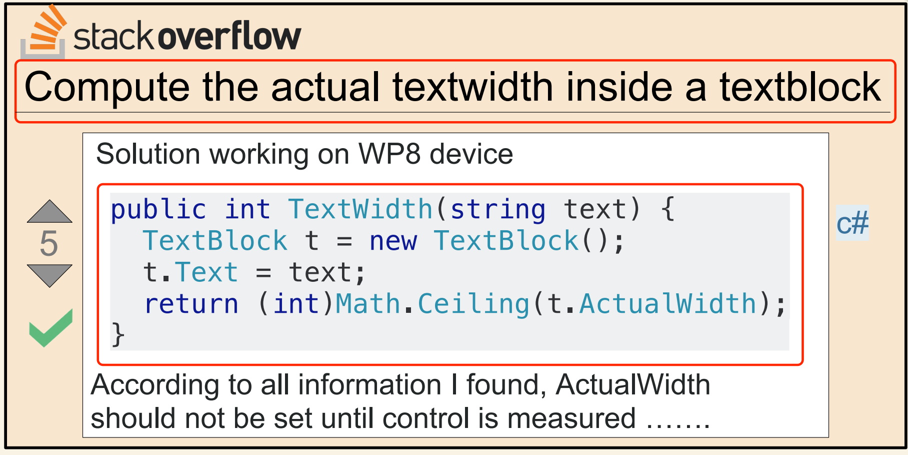
1. Generation models using tree to sequence methods
2. Discovery and explanation of code idioms
3. Using language-code models for code synthesis



## Code Summarization Dataset

NEW

We create a new dataset from programming QA websites containing 66K C# and 33K SQL examples.



stackoverflow

Compute the actual textwidth inside a textblock

Solution working on WP8 device

```
public int TextWidth(string text) {
  TextBlock t = new TextBlock();
  t.Text = text;
  return (int)Math.Ceiling(t.ActualWidth);
}
```

According to all information I found, ActualWidth should not be set until control is measured …….

Code snippets in this dataset are non-trivial:

| | | | | | |
|---|---|---|---|---|---|
| Loops | > 20% | > 2 Functions | 50% | \|Code\| | 38 |
| Conditionals | > 22% | > 2 Statements | 45% | \|Summary\| | 12 |

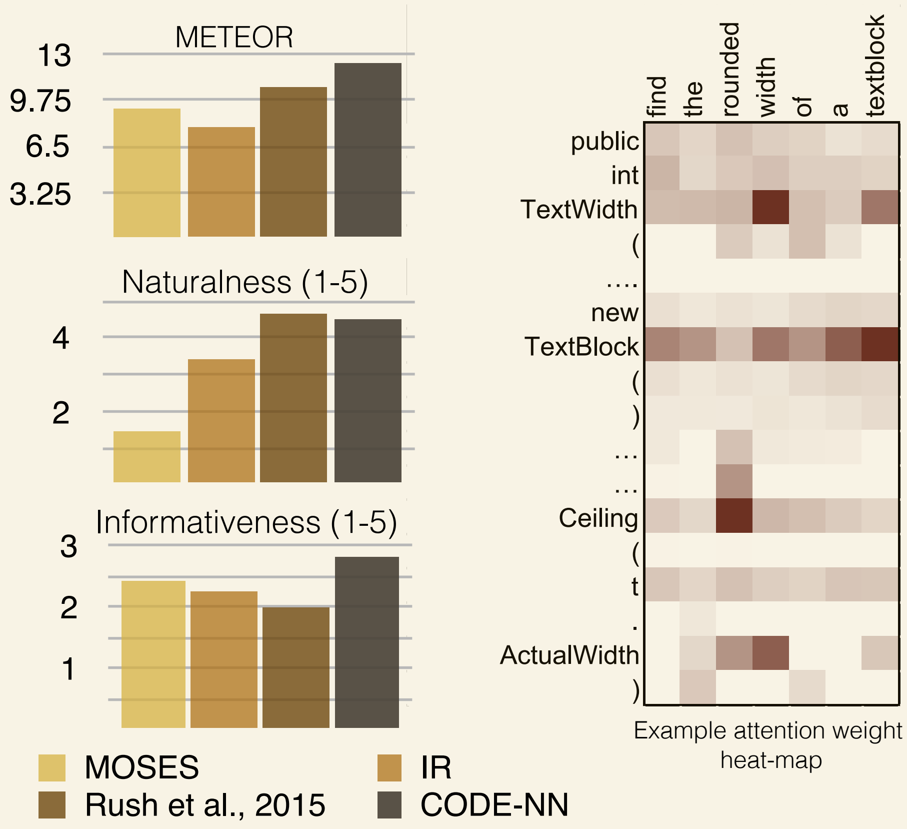
### Human Annotations

We gather 2 additional references for 200 code snippets for more accurate development and testing.

Data/Code at: https://github.com/sriniiyer/codenn

## Experiments

Our model beats competitive baselines on summarization metrics and human evaluations.



Example attention weight heat-map

### Example Outputs

| | |
|---|---|
| C# | How to convert string to int? |
| | How to get all child nodes in TreeView? |
| | How to call a URL from a web api post ? |
| SQL | How to get random rows from a mysql database? |
| | How to get the sum of a column in a single query? |