

Understanding Source Code through Machine Learning to Create Smart Software Engineering Tools

Miltos Allamanis, University of Edinburgh

March 13th, 2016



THE UNIVERSITY of EDINBURGH
informatics

Joint work with: *Charles Sutton (UoE), Earl T. Barr (UCL), Chris Bird (MSR), Daniel Tarlow (MSRC), Yi Wei (MSRC), Andrew D. Gordon (MSRC)*

My PhD is supported by

Microsoft®

Research

Developers implicitly embed **knowledge** in code that may be useful for the same or other projects.



GitHub

Atlassian
Bitbucket

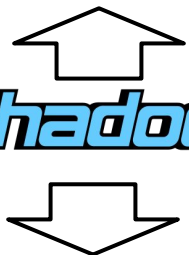
internal &
external
codebases

Mine the hidden knowledge to create **smart** software engineering tools.

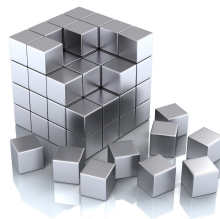
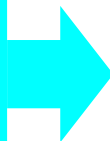
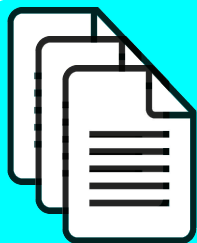
GitHub



hadoop



ACME
CORPORATION

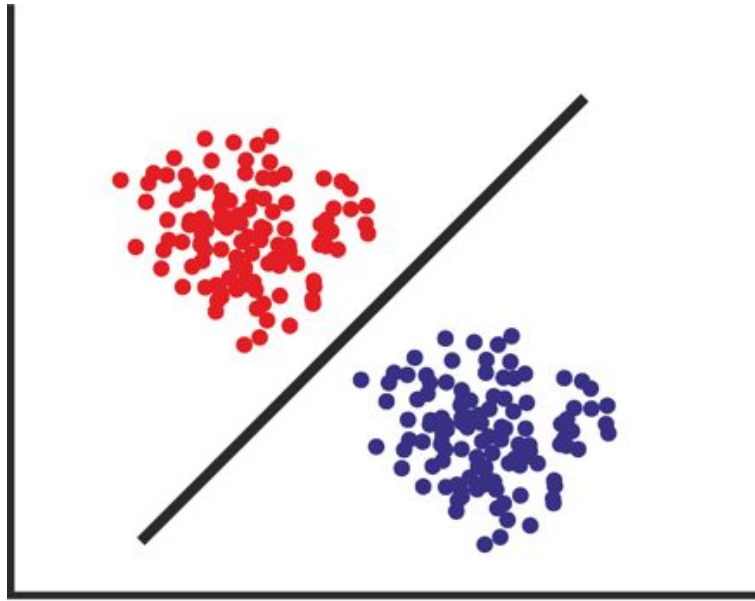


**Machine
Learning
Models of
Source Code**



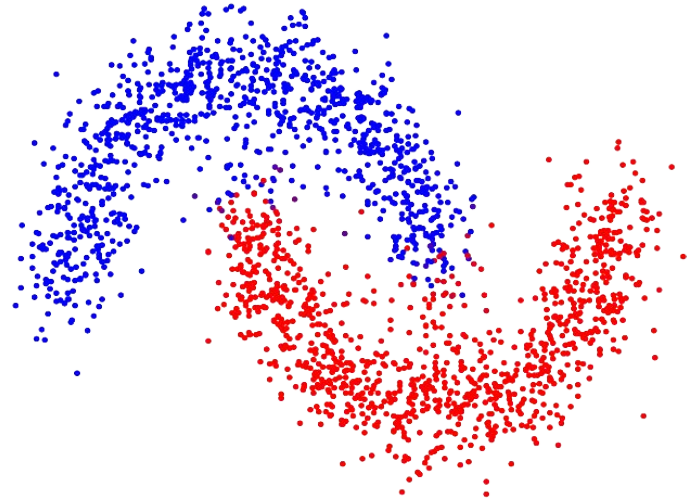
Software Engineer

A Spectrum of Problems for Machine Learning



Classification

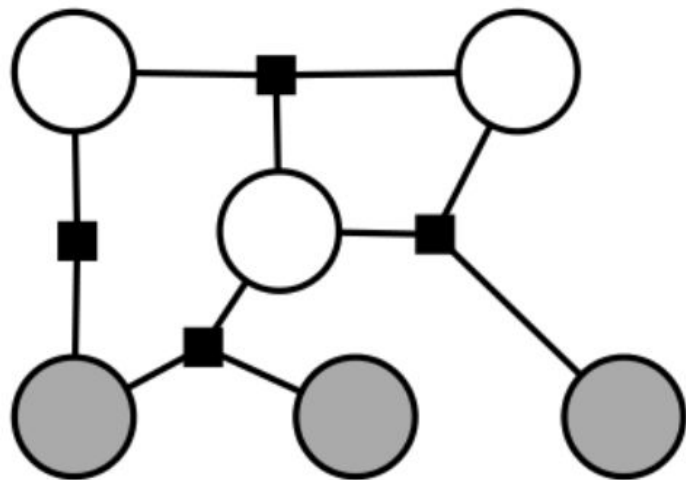
Supervised



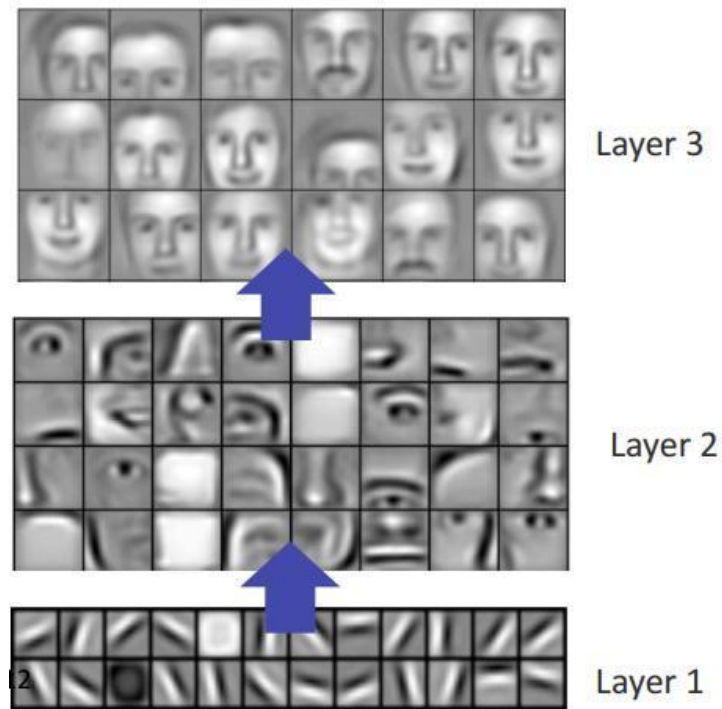
Clustering

Unsupervised

A Spectrum of Problems for Machine Learning

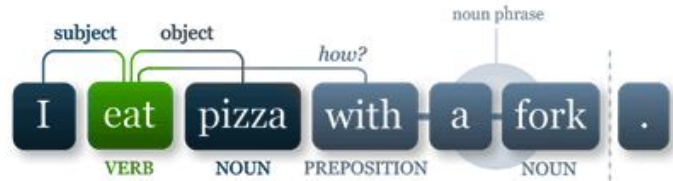


Joint Classification



Learning Features

Natural Language Processing with Machine Learning



- › **Resolve** language ambiguities with principled probabilistic models of language.
- › Learn model parameters from annotated corpora.

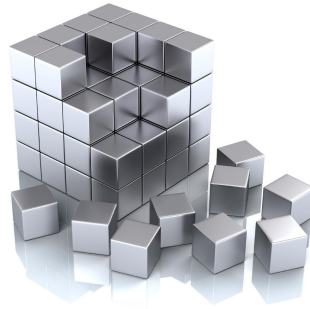
Natural Language Processing (NLP)



Some **Knowledge**
of **Linguistics**



Data: Corpora of
Text, Speech etc



**Models of Aspects
of Natural
Language**



Parsing

**Named Entity
Recognition**

**Machine
Translation**

....

Use Machine Learning to **model** aspects of a natural language.

Machine Learning Models of Source Code

“All models are wrong, some are useful” - George Box



Software Engineers



Codebases



Machine Learning Models
of Aspects of Source Code



Software Engineering
Tools

Language Models for Source Code

Assign a non-zero
probability to every
piece of valid code

Probabilities learned
from training corpus

```
for (int i = 0; i < nProperties; i++) {
    final List<TreeNode<TSGNode>> children = node
        .getChildrenByProperty().get(i);
    final int nChildren = children.size();
    ruleConsequent.nodes.add(Lists
        .<Integer> newArrayListWithCapacity(nChildren));
    for (int j = 0; j < nChildren; j++) {
        final int childNode = node.getChild(j, i).getData().nodeKey;
        ruleConsequent.nodes.get(i).add(childNode);
    }
}

return new CFGRule(rootId, ruleConsequent);
}

public BlockedPosteriorComputer getPosteriorComputer() {
    return samplePosteriorComputer;
}

public final CFGPrior getPrior() {
    return prior;
}

public void lockSamplerData() {
    prior.lockPrior();
    burninPosteriorComputer.getPrior().cfg = samplePosteriorComputer
        .getPrior().cfg;
}

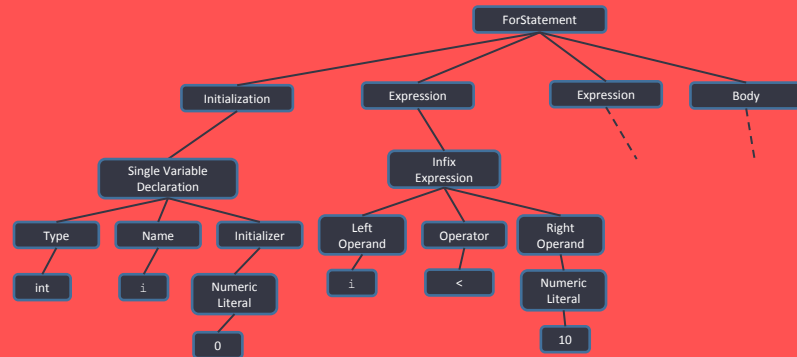
@Override
public void sampleAllTreesOnce(final int currentIteration,
    final int totalIterations, final AtomicBoolean stop) {
    final Thread termSignalHandler = new Thread() {
        @Override
        public void run() {
            stop.set(true);
        }
    };
};
```

Language Models of Source Code – Design Choices

Token-level Models

```
for (int i = 0; i < 10; i++){  
    Console.WriteLine(i);  
}
```

Syntactic Models



N-gram Language Models

$$P(t_0 \dots t_M) = \prod_{m=0}^M P(t_m | t_{m-1} \dots t_{m-n+1})$$

Parameters of ML Model

$$P(t_m | t_{m-1} \dots t_{m-n+1})$$

e.g. $P(\theta | \text{“for (int i =”})$



How n-gram models see code?

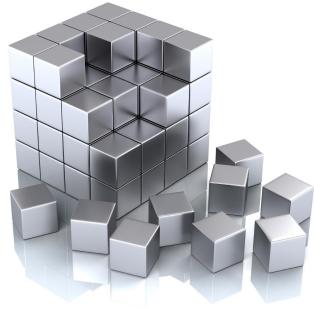
```
package org.cfclipse.cfml.snippets;

import org.rioproject.examples.logicdesigner.model.getState ( ) {
    cd1.Choreography;
import org.apache.thrift.protocol.TProtocolUtil.skip(iprot);
event.newLineCount == 3 )
    { case '|' :
        if ( rule.FireAllRulesCommand;

import org.apache.hadoop.conf.get(0, 0, newByteBuffer, 0,
count);
} switch ( classifierID ) {
    pd.getName() {
        cBondNeighborsB.get(MODULE).declaringType
            = (DEREnumerated)
```



Machine Learning



Machine Learning Model

Designed by humans



Model Parameters

Learned from data

Learn the parameters of the model from data.

Handle **uncertainty** and **noise**.

Learning Model Parameters

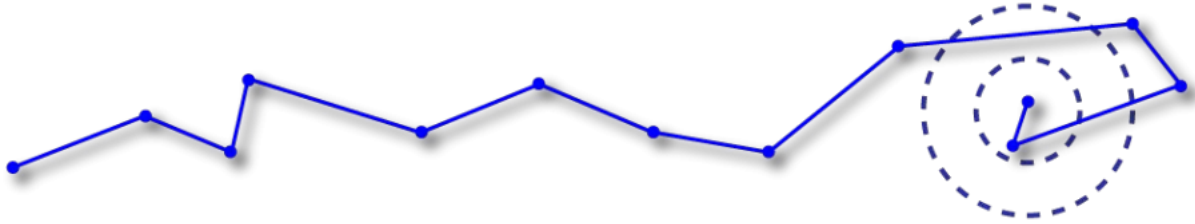
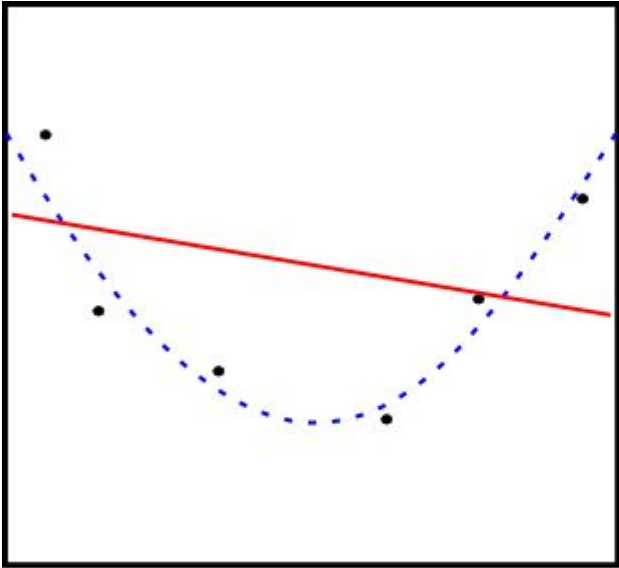


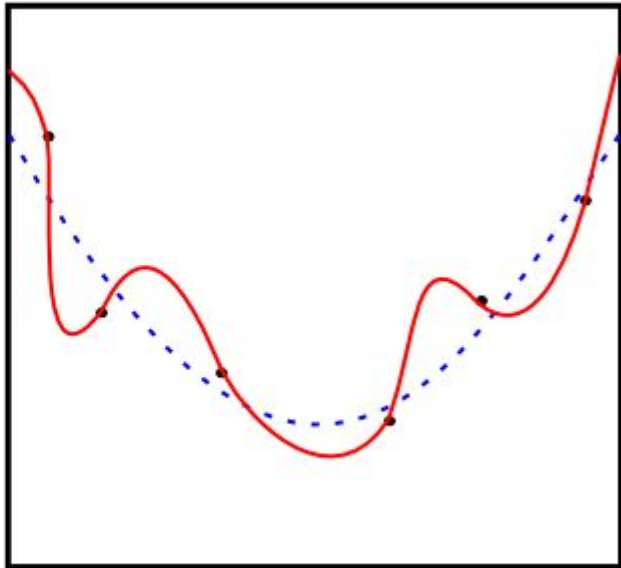
Image from marple.eeb.uconn.edu

- › **Optimize** objective function in training set
- › Use **computational methods** of optimization

Finding a good model



High Bias
Underfitting

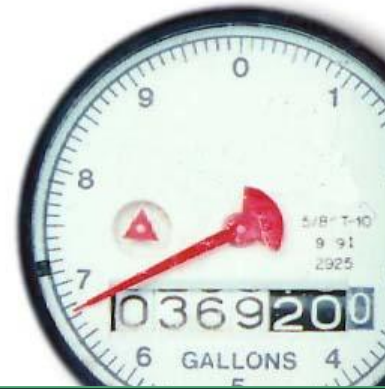


High Variance
Overfitting

Automatic Evaluation in Machine Learning

Imperfect measures of performance *such as*

- › Prediction Accuracy
- › Model Fit



- › **Quantify** performance in a **reproducible** manner
- › **Drive** improvement of systems in a **measurable** way

Source Code and Machine Learning

Coding Patterns

Mine & exploit common patterns

[*Hindle et al. 2012,*
Allamanis & Sutton 2014,
Allamanis et al. 2014, 2015]

Formal Methods

Probabilities over Search Space (e.g. Synthesis)

[*Ellis et al. 2015*]

Code & Text

Code search,
NL to Code

[*Yusuke, et al. 2015,*
Movshovitz-Attias & Cohen, 2013,
Allamanis et al. 2014]

Probabilistic Static Analyses

Probability Distribution of (Formal) Properties

[*Raychev et al. 2015,*
Mangal et al. 2015]

Runtime Traces

Infer Program Properties from Traces

[*Brockschmidt et al. 2014*
Yujia Li et al. 2015]



Outline

Learning Naming Conventions

› Lexical Patterns

Learning to Map Natural Language to Source Code

› Syntactic Patterns

```
class Foobar  
{  
    public List<Oof> DoSomething(Emails emails)  
{  
    var soapRequest = MakeSoapRequest(emails);  
    var xmlResult = IssueRequest(soapRequest);  
    var oofs = Go(xmlResult).ToList();  
    return oofs;  
}
```

***“Programs must be written for people to read,
and only incidentally for machines to execute.”***

- Abelson & Sussman, SICP, preface to the first edition

```
public IEnumerable<Oof> Go(String soap)  
{  
    namespace xns = "http://schemas.microsoft.com/soap";  
    ...  
}
```

Learning Naming Conventions

A coding convention is a syntactic constraint *beyond* those imposed by the language grammar.

The Importance of Coding Conventions

Code Review Discussions

Conventions	38%
<i>Naming</i>	24%
<i>Formatting</i>	9%

[Allamanis et al. FSE 2014]

Based on 169 *code reviews* with 1,093 discussion threads in Microsoft.

The Importance of Coding Conventions

Code Review Discussions

Conventions

38%

Naming

24%

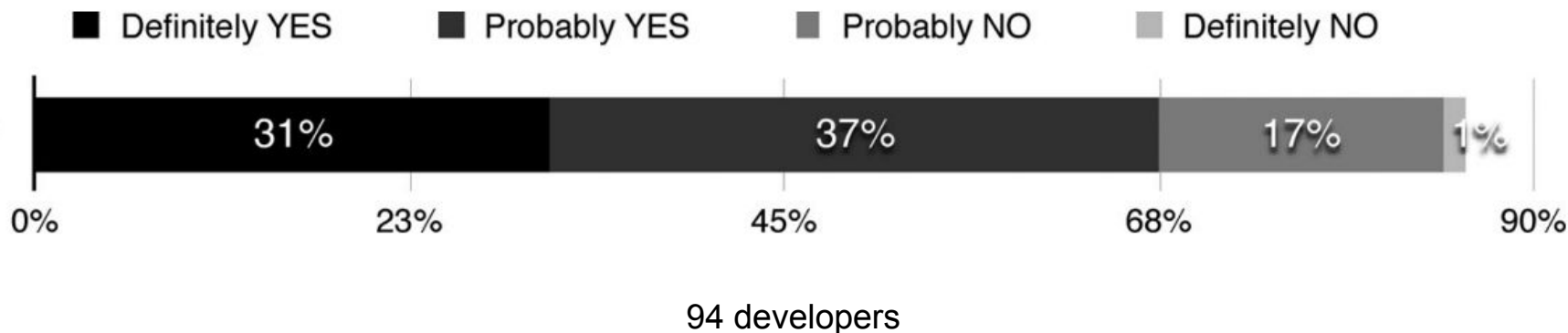
Formatting

9%

[Allamanis et al. FSE 2014]

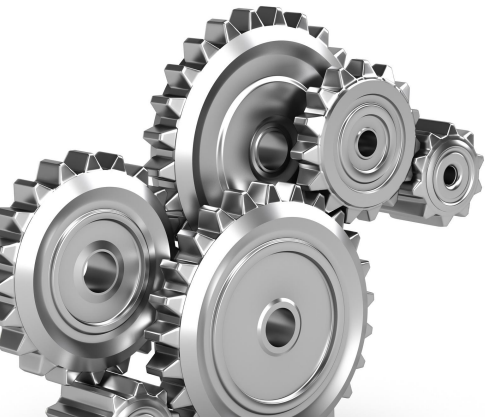
Based on 169 code reviews with 1,093 discussion threads in Microsoft.

Is recommending identifier renamings useful?



Arnaoudova, Venera, L. Eshkevari, Massimiliano Di Penta, Rocco Oliveto, Giuliano Antoniol, and Y. Gueheneuc. "REPENT: Analyzing the nature of identifier renamings." (2014)

A Machine Learning Perspective



A name reflects important aspects of code **functionality**.

Learning to name source code elements is a first step in *understanding* code through machine learning.

Suggestions for junit/src/test/java/junit/tests/runner/TextRunnerTest.java

```
public class TextRunnerTest extends TestCase {  
    void execTest(String testClass, boolean success) throws Exception {  
        ...  
        InputStream i = p.getInputStream();  
        while ((i.read()) != -1);  
        ...  
    }  
    ...  
}
```

Suggestions for junit/src/test/java/junit/tests/runner/TextRunnerTest.java

```
public class TextRunnerTest extends TestCase {  
    void execTest(String testClass, boolean success) throws Exception {  
        ...  
        InputStream i = p.getInputStream();  
        while ((i.read()) != -1);  
        ...  
    }  
    ...  
}
```

**automatically suggest
renamings**

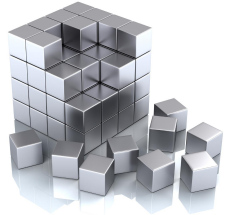


**Source Code
Language Model**

Suggestions for junit/src/test/java/junit/tests/runner/TextRunnerTest.java

```
public class TextRunnerTest extends TestCase {  
    void execTest(String testClass, boolean success) throws Exception {  
        ...  
        InputStream i = p.getInputStream();  
        while ((i.read()) != -1);  
        ...  
    }  
    ...  
}
```

automatically suggest renamings

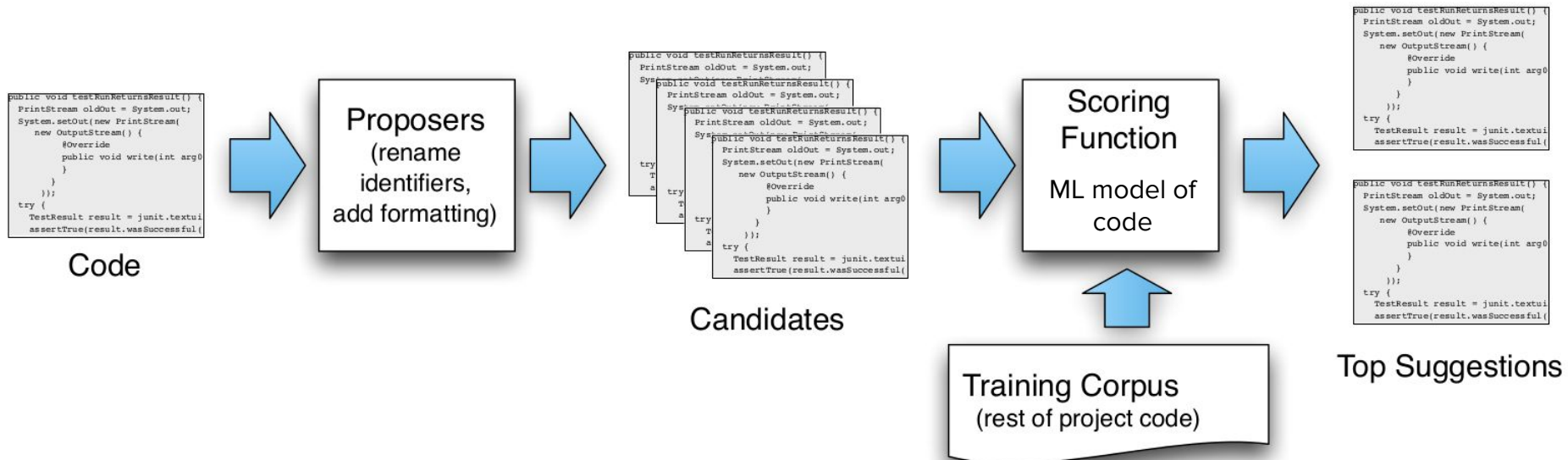


**Source Code
Language Model**

**Score by
naturalness
&
Threshold**

1. '**i**' (18.07%) -> {input (81.93%), }

Suggesting Names to Developers: The Naturalize Framework



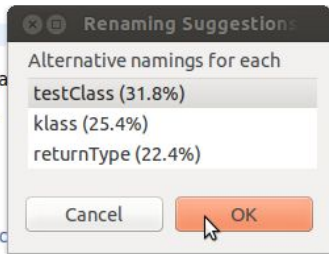
[Allamanis et al. FSE 2014, FSE 2015]

Naturalize Tools - devstyle



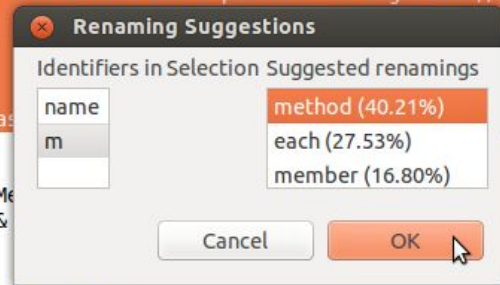
```
private Test testCaseForClass(Class<?> each) {
    if (TestCase.class.isAssignableFrom(each)) {
        return new TestSuite(each.asSubclass(TestCase.class));
    } else {
        return warning(each.getCanonicalName() + " is not a TestSuite");
    }
}

/**
 * Constructs a TestSuite from the given array of classes.
 *
 * @param testClasses the classes to be tested
 * @return a TestSuite containing the given classes
 */
public TestSuite createTestSuite(Class<?>... testClasses) {
    return new TestSuite(testClasses);
}
```



devstyle
suggests identifier
renamings

```
private void addTestMethod(Method m, List<String> names, Class<?> theClass) {
    String name = m.getName();
    if (names.contains(name)) {
        return;
    }
    if (!isPublicTestMethod(m)) {
        if (isTestMethod(m)) {
            addTest(warning("Test method isn't public: " + m.getName() + "(" + theClass.getName() + ")"));
        }
    }
    return;
}
names.add(name);
addTest(createTest(theClass, name));
}
```



```
private boolean isPublicTestMethod(Method m) {
    return isTestMethod(m) && m.isPublic();
}

private boolean isTestMethod(Method m) {
    return m.getParameterTypes().length == 0 && m.isAnnotationPresent(TestMethod.class);
}
```

mallamanis added some commits on 10 Feb

- Renamed the ImmutableBlobContainer container to blobContainer. ... af0f713
- Renamed XContentParser.Token named "t" to "token". ... 7d092c2
- Renamed ClusterBlocks variable named "blob" to "blobContainer". ... 64121e2

Javanna added **enhancement** **v1.2.0** **v2.0.0** label

Javanna self-assigned this on 7 Apr

Javanna commented on 7 Apr

Merged, thanks!

Javanna closed this on 7 Apr

mallamanis added some commits on 26 Feb

- Renamed "i" (15.46%) to "index" (31.34%). The naturalize tool detected ... 35301ce
- Renamed "value" (18.55%) to "scalar" (51.13%). The naturalize tool ... b522d52
- Minor changes in JavaDoc 46d4393
- Renamed "vector" (7.51%) to "point" (64.23%). The naturalize tool ... 5d6dff6
- Renamed "scale" (24.02%) to "scaleXY" (75.98%). The naturalize tool ... 33726e2

sinistersnare commented on 26 Feb

Very interesting project, and thanks for the contribution!

I've always wanted to spend a good amount of time using findbugs with libgdx, there's a ton of recommendations it provides.

badlogic commented on 26 Feb Owner

Wow, that's a pretty cool tool!

badlogic closed this on 26 Feb

badlogic reopened this on 26 Feb

badlogic merged commit **3d5040f** into `libgdx:master` from `mallamanis:master` on 26 Feb Revert

18 patches for 5 well known
open source projects:
14 accepted, 4 ignored

Method Naming Problem

```
1 private void                      () {  
2     String vertexShader = "literal_1";  
3     String fragmentShader = "literal_2";  
4     shader = new ShaderProgram(vertexShader,  
5         fragmentShader);  
6     if(shader.isCompiled() == false)  
7         throw new IllegalArgumentException(  
8             "literal_3" + shader.getLog());  
9 }
```

Method Naming Problem

```
1 private void                      () {  
2     String vertexShader = "literal_1";  
3     String fragmentShader = "literal_2";  
4     shader = new ShaderProgram(vertexShader,  
5         fragmentShader);  
6     if(shader.isCompiled() == false)  
7         throw new IllegalArgumentException(  
8             "literal_3" + shader.getLog());  
9 }
```

Names describe what it **does** not what it *is*

Models need to be “non-local”

Method Naming Problem

```
1 private void                      () {  
2     String vertexShader = "literal_1";  
3     String fragmentShader = "literal_2";  
4     shader = new ShaderProgram(vertexShader,  
5         fragmentShader);  
6     if(shader.isCompiled() == false)  
7         throw new IllegalArgumentException(  
8             "literal_3" + shader.getLog());  
9 }
```

Suggestions:

- create
- create?UNK?
- init
- createShader

Method Naming Problem

```
1 private void createDefaultShader () {  
2     String vertexShader = "literal_1";  
3     String fragmentShader = "literal_2";  
4     shader = new ShaderProgram(vertexShader,  
5         fragmentShader);  
6     if(shader.isCompiled() == false)  
7         throw new IllegalArgumentException(  
8             "literal_3" + shader.getLog());  
9 }
```

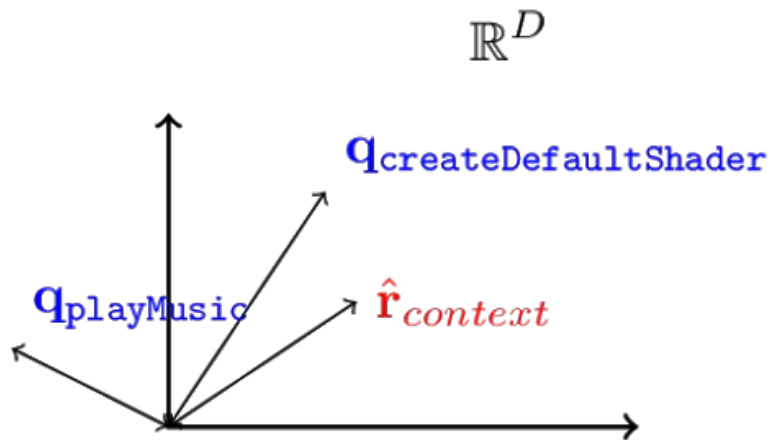
Suggestions:

- create
- create?UNK?
- init
- createShader

A Machine Learning Model of Names

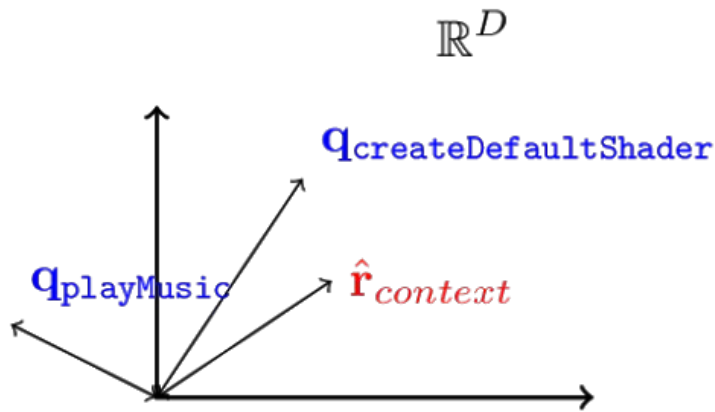
$$P(t|c) = \frac{\exp(s_{\theta}(t, c))}{\sum_{t'} \exp(s_{\theta}(t', c))}$$

Embedding Identifiers



$q_t \in \mathbb{R}^D$ are “embeddings” \therefore model parameters

Embedding Identifiers

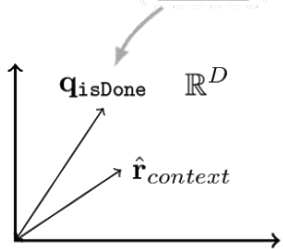


$$s_{\theta}(t, c) = \hat{\mathbf{r}}_{\text{context}}^{\top} q_t + b_t$$

$$P(t|c) = \frac{\exp(s_{\theta}(t, c))}{\sum_{t'} \exp(s_{\theta}(t', c))}$$

Neural Context Models of Source Code

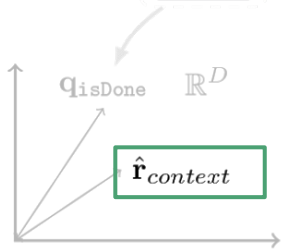
Variable: **isDone**



$$s_{\theta}(\cdot) = \hat{\mathbf{r}}_{context}^T \mathbf{q}_{isDone} + b_{isDone}$$

Neural Context Models of Source Code

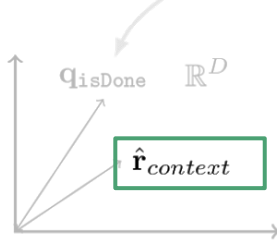
Variable: `isDone`



$$s_{\theta}(\cdot) = \hat{\mathbf{r}}_{context}^T \mathbf{q}_{isDone} + b_{isDone}$$

Neural Context Models of Source Code

Variable: `isDone`

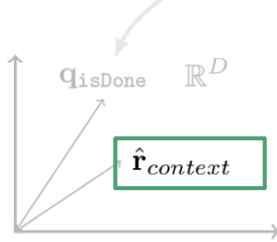


Features:
`boolean`, `in:MethodBody`, `final`

$$s_{\theta}(\cdot) = \hat{\mathbf{r}}_{context}^T \mathbf{q}_{isDone} + b_{isDone}$$

Neural Context Models of Source Code

Variable: `isDone`



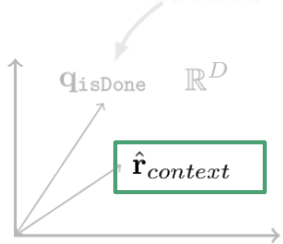
Features:
`boolean`, `in:MethodBody`, `final`

Contexts:
`while` `(! isDone)` `{`

$$s_{\theta}(\cdot) = \hat{\mathbf{r}}_{context}^T \mathbf{q}_{isDone} + b_{isDone}$$

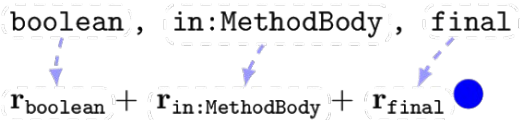
Neural Context Models of Source Code

Variable: `isDone`

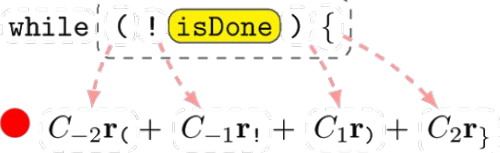


$$s_{\theta}(\cdot) = \hat{\mathbf{r}}_{context}^T \mathbf{q}_{isDone} + b_{isDone}$$

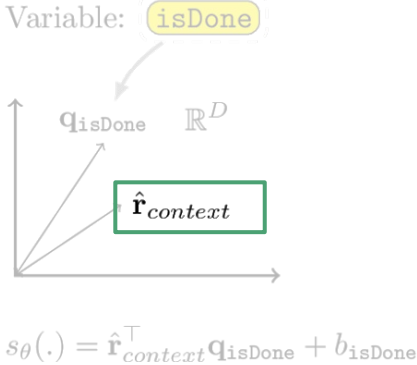
Features:



Contexts:



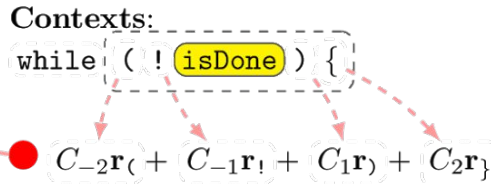
Neural Context Models of Source Code



Features:
`boolean`, `in:MethodBody`, `final`

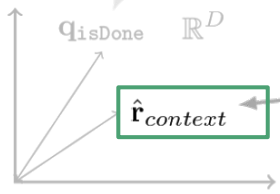
$\mathbf{r}_{boolean} + \mathbf{r}_{in:MethodBody} + \mathbf{r}_{final}$

$$\hat{\mathbf{r}}_{context} = \sum_{f \in F_{tc}} \mathbf{r}_f + \sum_{\forall k: K \geq |k| > 0} C_k \mathbf{r}_{t_{i+k}}$$



Neural Context Models of Source Code

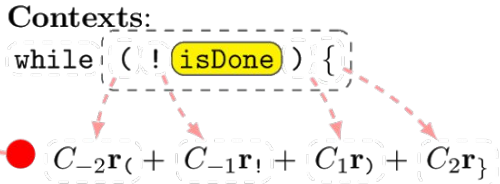
Variable: `isDone`



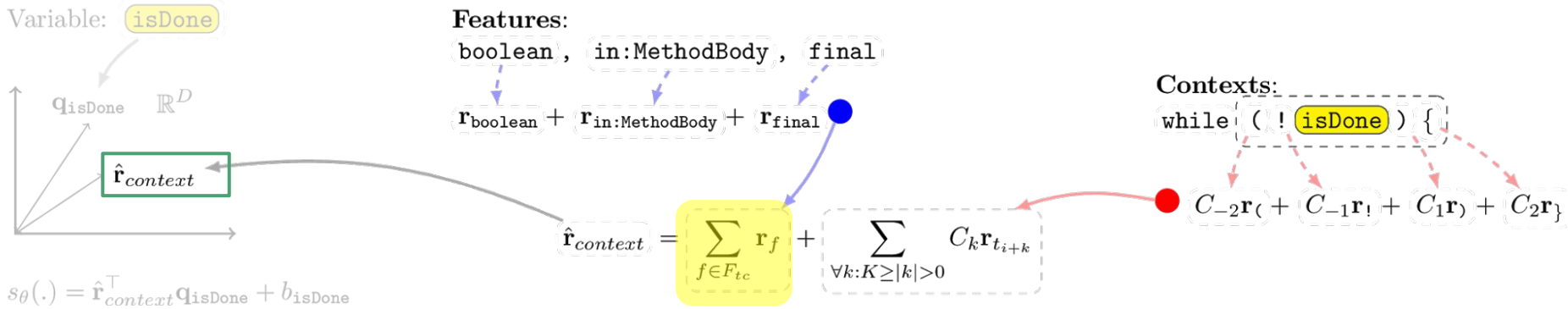
$$s_{\theta}(\cdot) = \hat{\mathbf{r}}_{context}^{\top} \mathbf{q}_{isDone} + b_{isDone}$$

Features:
`boolean`, `in:MethodBody`, `final`
 $\mathbf{r}_{boolean} + \mathbf{r}_{in:MethodBody} + \mathbf{r}_{final}$

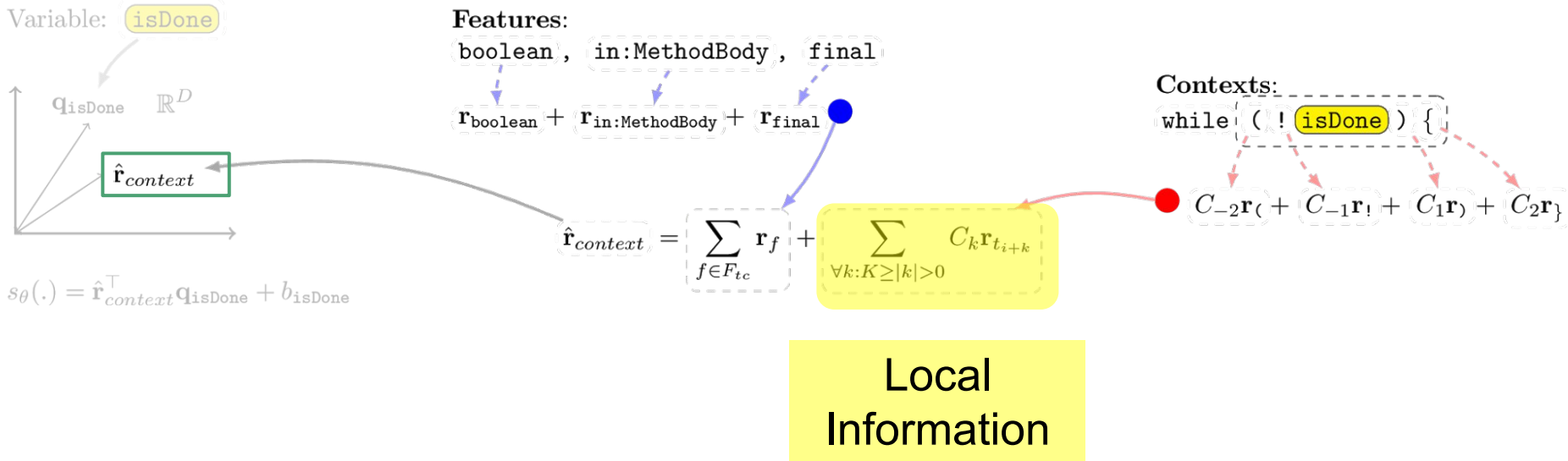
$$\hat{\mathbf{r}}_{context} = \sum_{f \in F_{tc}} \mathbf{r}_f + \sum_{\forall k: K \geq |k| > 0} C_k \mathbf{r}_{t_i+k}$$



Neural Context Models of Source Code

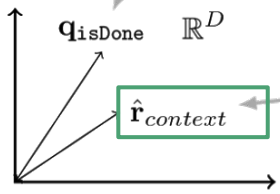


Neural Context Models of Source Code



Neural Context Models of Source Code

Variable: `isDone`



$$s_{\theta}(\cdot) = \hat{\mathbf{r}}_{context}^{\top} \mathbf{q}_{isDone} + b_{isDone}$$

Features:

`boolean`, `in:MethodBody`, `final`

$\mathbf{r}_{boolean} + \mathbf{r}_{in:MethodBody} + \mathbf{r}_{final}$

$$\hat{\mathbf{r}}_{context} = \sum_{f \in F_{tc}} \mathbf{r}_f + \sum_{\forall k: K \geq |k| > 0} C_k \mathbf{r}_{t_i+k}$$

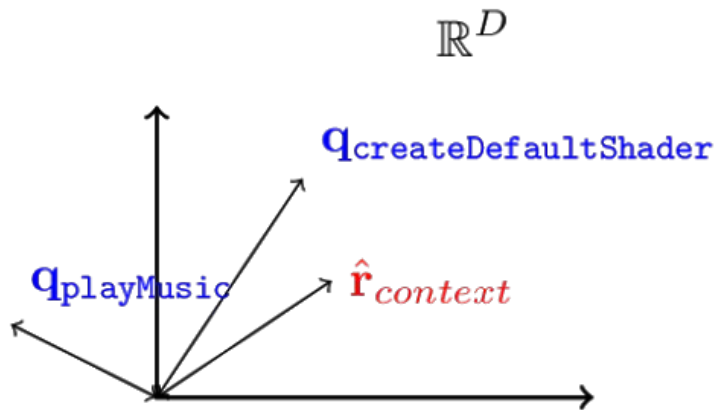
Contexts:

`while (! isDone) {`

$C_{-2}\mathbf{r}_{(} + C_{-1}\mathbf{r}_{!} + C_1\mathbf{r}_{)} + C_2\mathbf{r}_{\{}$




Embedding Identifiers



$$P(t|c) = \frac{\exp(s_{\theta}(t, c))}{\sum_{t'} \exp(s_{\theta}(t', c))}$$

Neologisms

neologism

[nɪˈɒlədʒɪz(ə)m] 

NOUN

a newly coined word or expression.

synonyms: new word · new expression · new term · new phrase · [coinage](#) · [More](#)

Powered by [OxfordDictionaries](#) · © Oxford University Press



Subtoken Context Models of Code



$$P(t_i | t_{i-1}, t_{i-2}, context)$$

Sequentially predict each subtoken given the context and the previous subtokens

Evaluation Methodology

Test File

```
ForkJoinTask<?> ■■■;  
if (task instanceof ForkJoinTask<?>) // avoid re-wrap  
    ■■■ = (ForkJoinTask<?>) task;  
else  
    ■■■ = new  
    ForkJoinTask.AdaptedRunnableAction(task);  
externalPush(■■■);
```



Suggestions

1. job (30%)
2. task (20%)
3. tsk (15%)

Evaluation on top 10 Java GitHub projects.

Perturb existing code and **retrieve** ground truth.

Evaluation Methodology

Test File

```
ForkJoinTask<?> █;
if (task instanceof ForkJoinTask<?>) // avoid re-wrap
    █ = (ForkJoinTask<?>) task;
else
    █ = new
ForkJoinTask.AdaptedRunnableAction(task);
externalPush(█);
```

```
ForkJoinTask<?> job;
if (task instanceof ForkJoinTask<?>) // avoid re-wrap
    job = (ForkJoinTask<?>) task;
else
    job = new
ForkJoinTask.AdaptedRunnableAction(task);
externalPush(job);
```

Suggestions

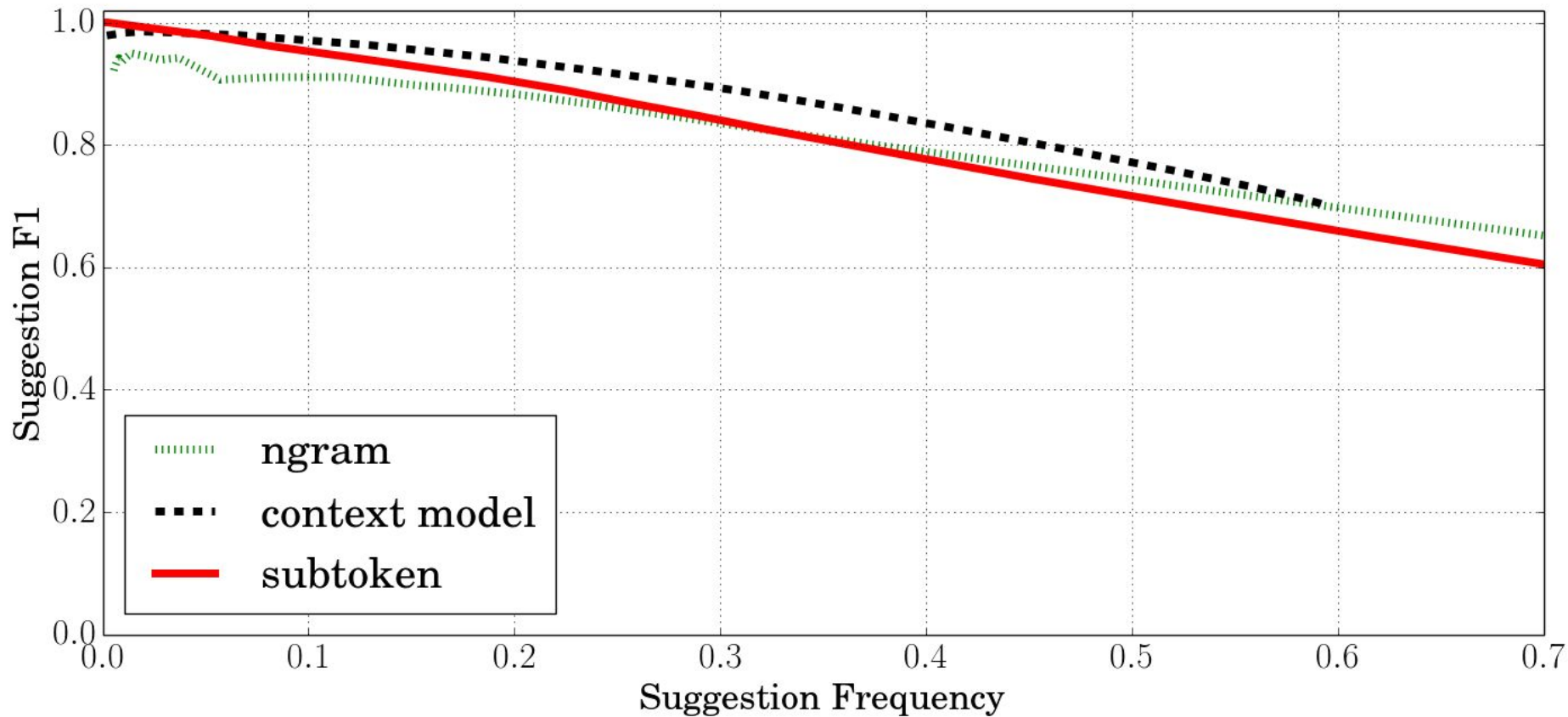
1. job (30%)
2. task (20%)
3. tsk (15%)

compare with
ground truth

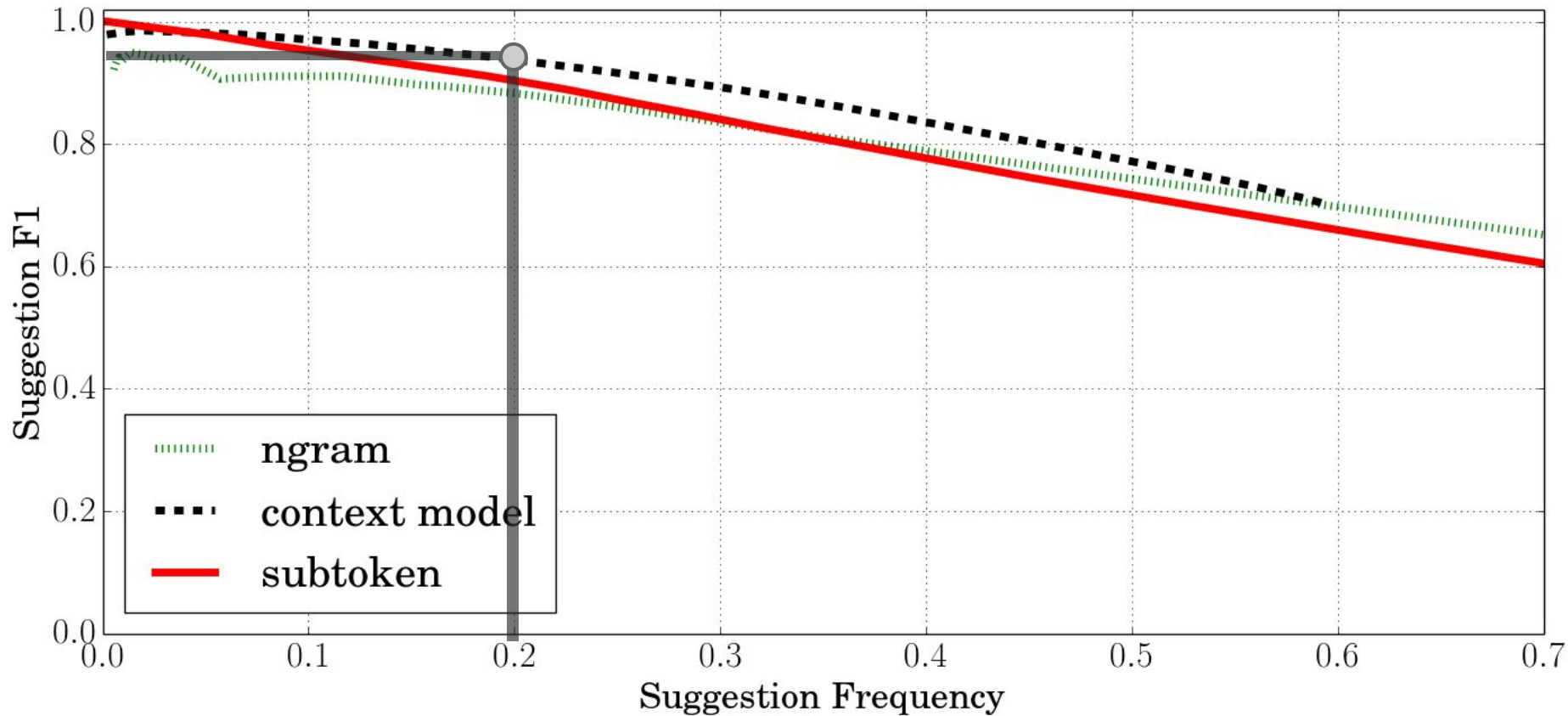
Evaluation on top 10 Java GitHub projects.

Perturb existing code and **retrieve** ground truth.

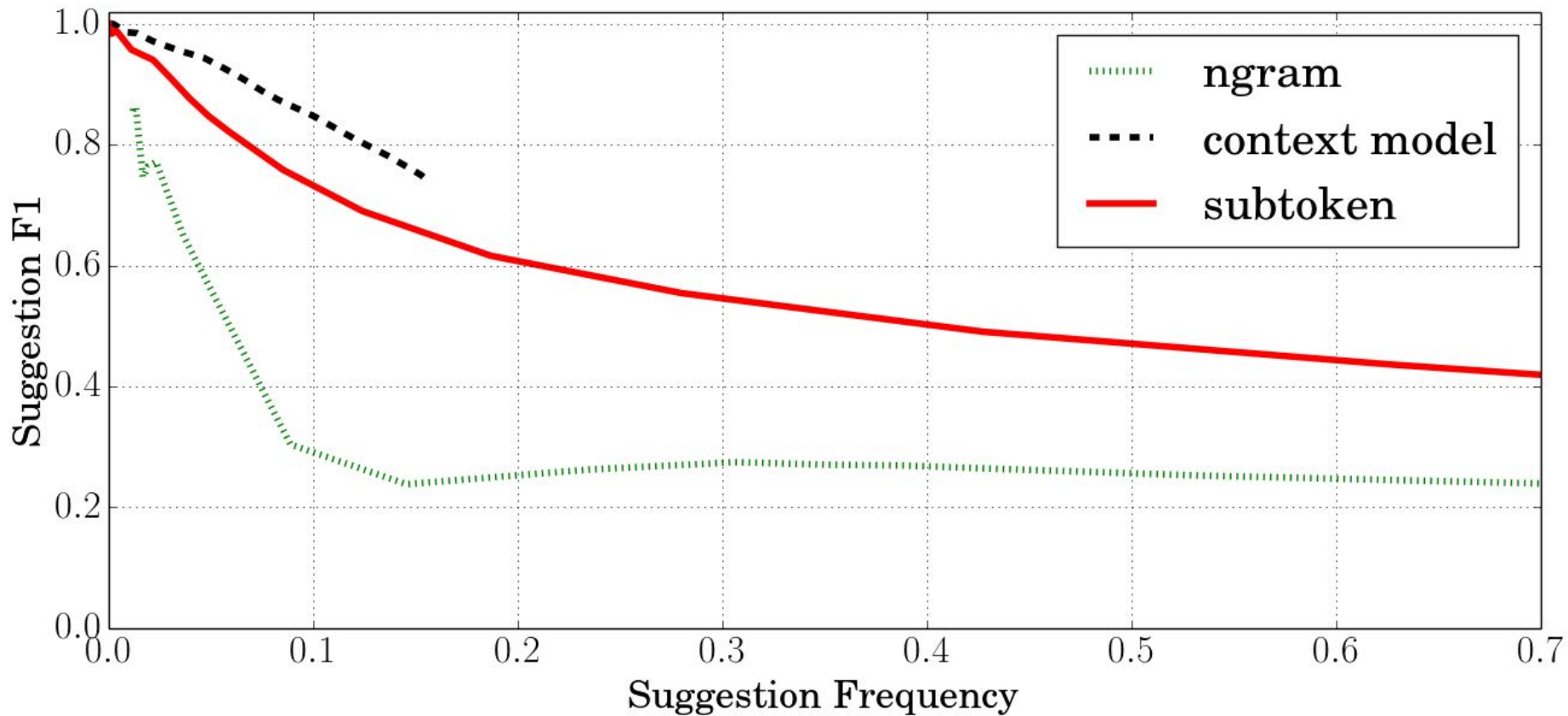
Suggesting Variable Names



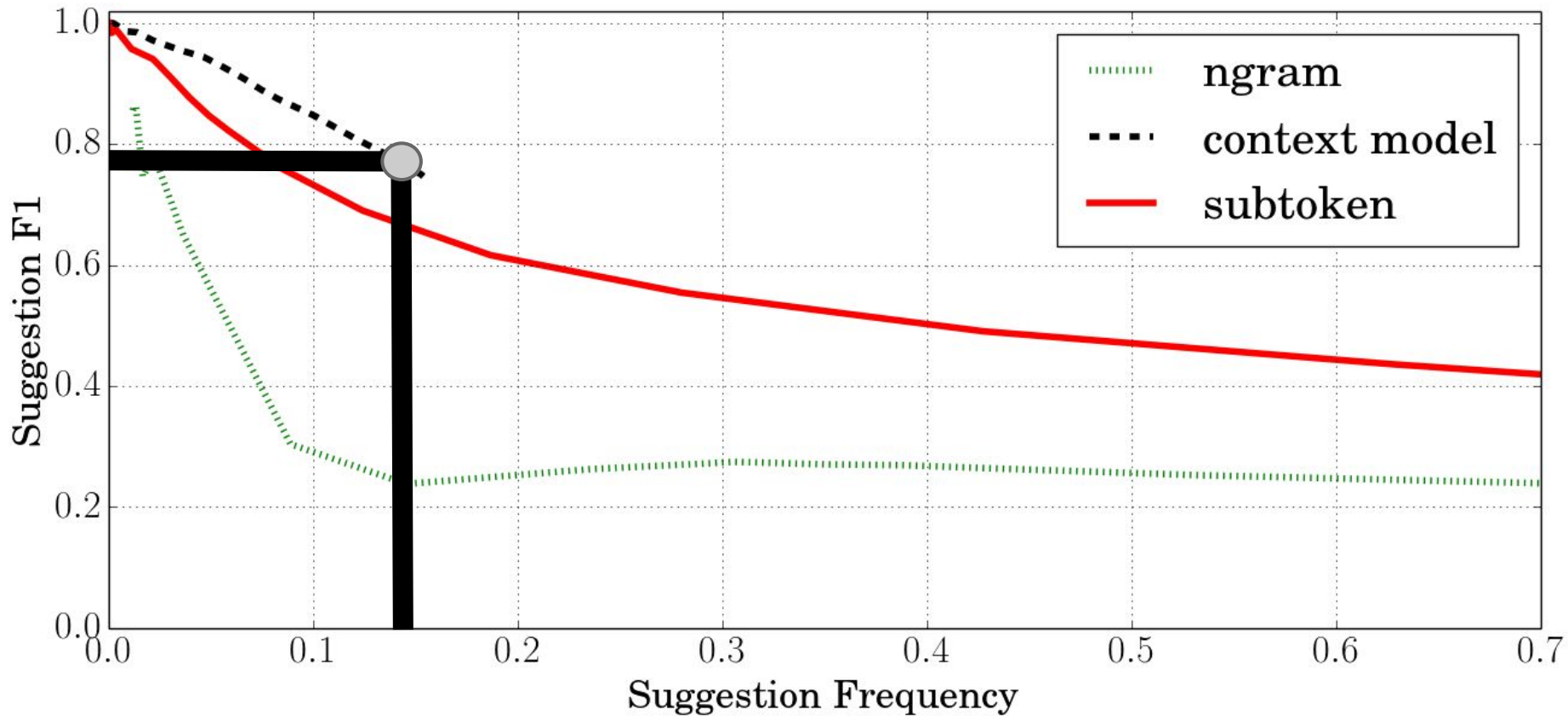
Suggesting Variable Names



Suggesting Method Names

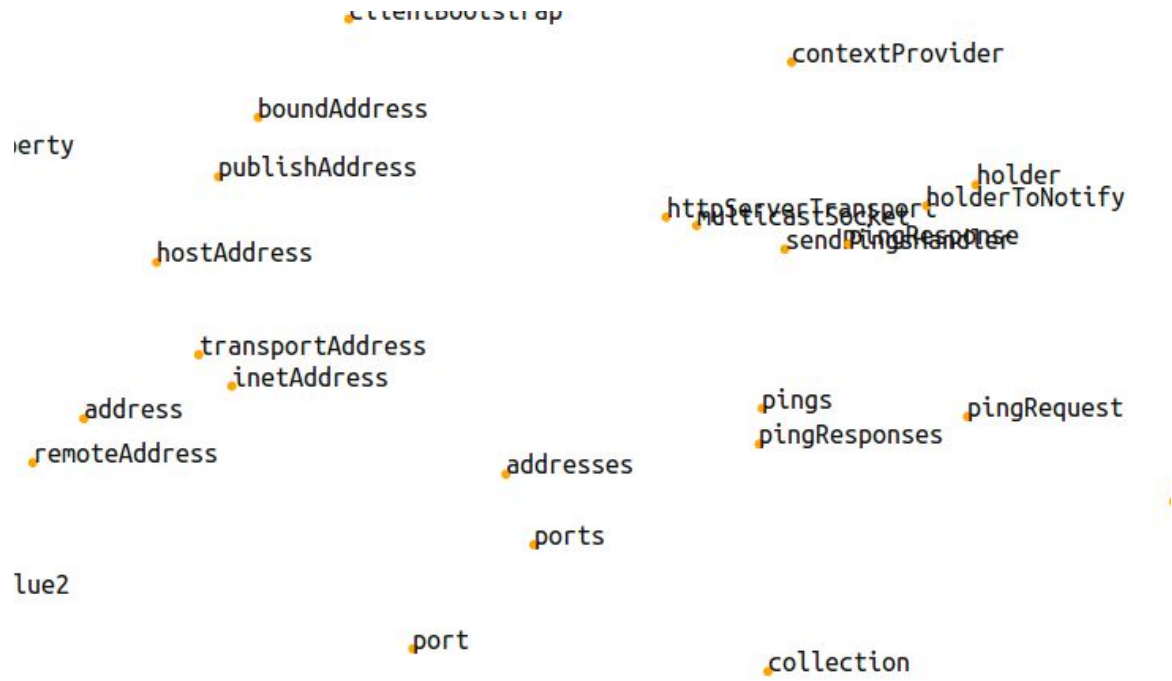


Suggesting Method Names



Embedding Visualization

<http://groups.inf.ed.ac.uk/cup/naturalize>



Learning to map natural language to source code

Work done in Microsoft Research - Cambridge

Joint work with Danny Tarlow, Yi Wei, Andy Gordon

```
def __init__(self, log_model, log_model_gradient, noise_sampler, noise_logp_model, number_of_distractors, batch_size, fudge_factor, adagrad_historical, step_size, dropout_random):
    self.log_model = log_model
    self.log_model_gradient = log_model_gradient
    self.noise_sampler = noise_sampler
    self.noise_logp_model = noise_logp_model
    self.number_of_distractors = number_of_distractors
    self.batch_size = batch_size
    self.fudge_factor = fudge_factor
    self.adagrad_historical = 0
    self.step_size = step_size
    self.dropout_random = dropout_random
```

```
def generate_distractors(self, current_batch_contexts):
    total_num_distractors = current_batch_contexts.shape[0]
    distractor_targets = np.zeros(total_num_distractors)
    distractor_logprobabilities = np.ones(total_num_distractors)
    distractor_contexts = np.zeros((current_batch_contexts.shape[0], total_num_distractors))
    for i in range(current_batch_contexts.shape[0]):
        context = current_batch_contexts[i, :]
        current_distractor_targets, current_distractor_logprobabilities = self.noise_sampler(context)
        frowids = i * self.number_of_distractors
        toids = frowids + self.number_of_distractors
        distractor_targets[frowids:toids] = current_distractor_targets
        distractor_logprobabilities[frowids:toids] = current_distractor_logprobabilities
    return distractor_contexts, distractor_logprobabilities, distractor_targets
```

```
def compute_data_expectation_gradient(self, current_batch_contexts):
    logprob_targets_given_context = self.log_model(current_batch_contexts)
    logprob_targets_for_noise_dist = self.noise_logp_model(current_batch_contexts)
    gradient_for_logmodel = self.log_model_gradient(current_batch_contexts, logprob_targets_given_context)
```

```
weight = np.log(self.number_of_distractors) + logprob_targets_for_noise_dist - logprob_targets_given_context
weight -> np.logaddexp(logprob_targets_for_noise_dist, logprob_targets_given_context)
weight = np.exp(weight) # to real space (from log-space)
data_exp_gradient = gradient_for_logmodel.dot(weight)
return data_exp_gradient
```

```
def compute_noise_expectation_gradient(self, theta, current_batch_contexts):
    noise_grad_for_data = self.log_model_gradient(current_batch_contexts, logprob_targets_for_noise_dist)
```

```
# Compute noise gradient contribution with less weight
temp = self.log_model(current_batch_contexts, distractor_targets)
temp -> np.logaddexp(temp, logprob_targets_for_noise_dist)
temp = np.exp(temp) # Now this contains the real-valued noise gradient contribution = noise_grad_for_data
```

Applications of Joint Models of Code & NL

Code Retrieval

NL Retrieval
for Source Code

and eventually code synthesis...

http://www.bing.com/search?c# get the first letter of each word in string and uppercase

WEB IMAGES VIDEOS MAPS NEWS MORE

bing c# get the first letter of each word in string and uppercase

MS Beta 691,000 RESULTS Any time

[c# - Make first letter of a string upper case - Stack Overflow](#)
stackoverflow.com/.../4135317/make-first-letter-of-a-string-upper-case
Resolved - Last updated: Sep 19, 2013 · 18 posts · First post: Nov 09, 2010
... the question in the link capitalizes the first letter of each word. ... the first letter of a string, even if the first ... of a string to be uppercase in C# ...

[c# - How can I uppercase the first letter of all words in ...](#) Jul 10, 2013 Resolved

[c# - Convert all first letter to upper case, rest lower ...](#) Sep 22, 2013 Resolved

[java - How to upper case every first letter of word in a ...](#) Oct 18, 2012 Resolved

[vb.net - Regex: How to Uppercase the first character of ...](#) Jan 22, 2014 Resolved

[C# Uppercase First Letter - Dot Net Perls](#)
www.dotnetperls.com/uppercase-first-letter
C# Uppercase First Letter. ... that uppercases the first letter. The first method here is ... method for uppercasing the first letter in each word in a string.

Code sample

```
static string UppercaseFirst(string s){  
    if (string.IsNullOrEmpty(s)){  
        return string.Empty;  
    }
```

See more on dotnetperls

[Uppercase first letter in each word of a string.-VBForums](#)
www.vbforums.com > ... > CodeBank - Visual Basic .NET
12 replies from February 2009 to March 2013
Here is very simple code to uppercase the first letter of each word in a string. ... word is being split in two, the first ... uppercase the first letter of each word ...

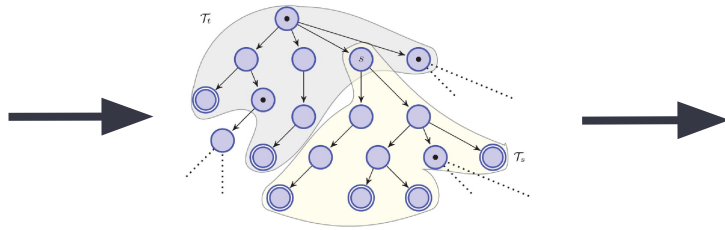
[Capitalize First Letter In Every Word In A String. - C# ...](#)
www.dreamincode.net > Dream.In.Code > Programming Help > C#
Capitalize first letter in every word in a string · Search: { static string UpperCase

Feedback

A Conditional Generative Model

*“get the first
letter of each
word in string
and uppercase”*

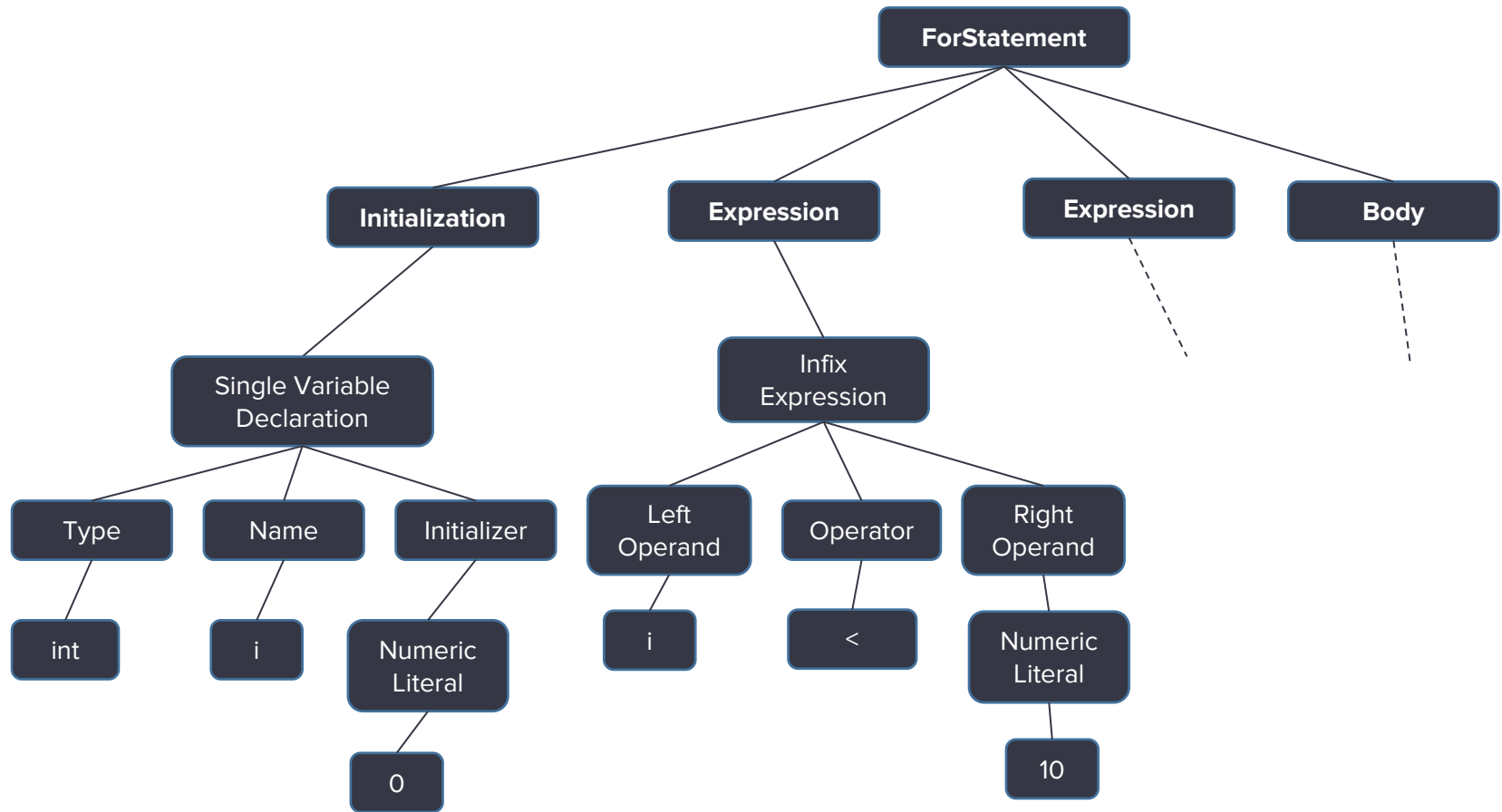
NL Query



Conditional
Generative
Model of Source Code

```
string s;  
string[] words = s.ToUpper().split(' ');  
string[] firstLetters = new string[words.Length];  
for (int i=0; i < words.Length; i++) {  
    firstLetters[i] = words.Substring(0,1);  
}
```

Synthesize/Score Code Snippet



Syntactic model of source code, *i.e.* model how AST is generated

Tree Generation Model: Context Free Grammars (CFG)

$$E \rightarrow E + E$$

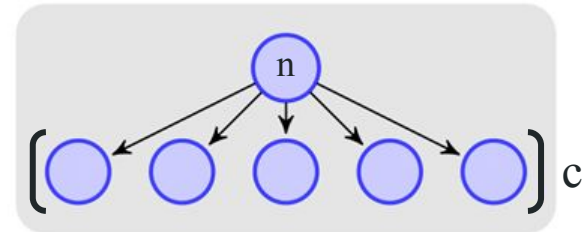
$$E \rightarrow T$$

$$F \rightarrow (E)$$

$$T \rightarrow F * F$$

$$T \rightarrow F$$

$$F \rightarrow id$$



Tree Generation Model: Probabilistic Context Free Grammars (PCFG)

$E \rightarrow E + E$ (prob 0.7)

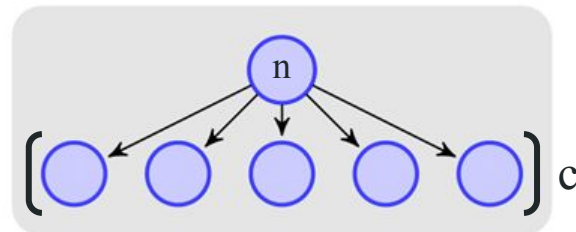
$E \rightarrow T$ (prob 0.3)

$F \rightarrow (E)$ (prob 0.1)

$T \rightarrow F * F$ (prob 0.6)

$T \rightarrow F$ (prob 0.4)

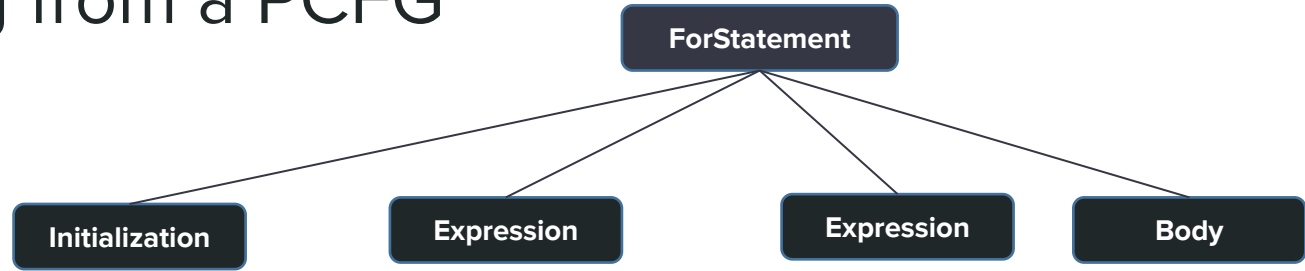
$F \rightarrow id$ (prob 0.9)



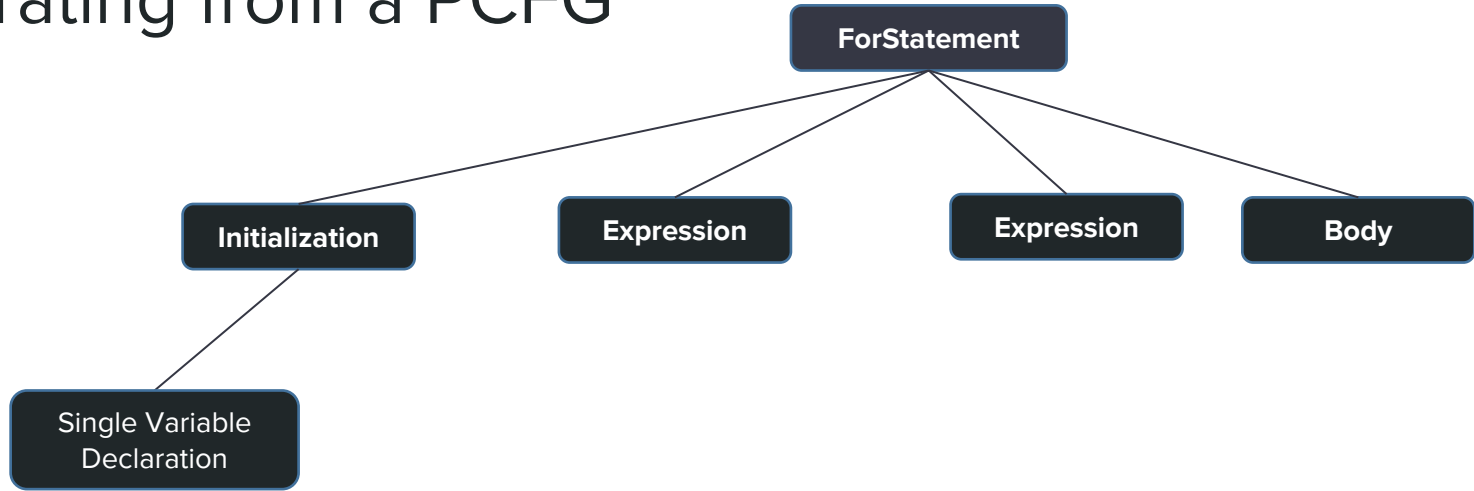
Generating from a PCFG

ForStatement

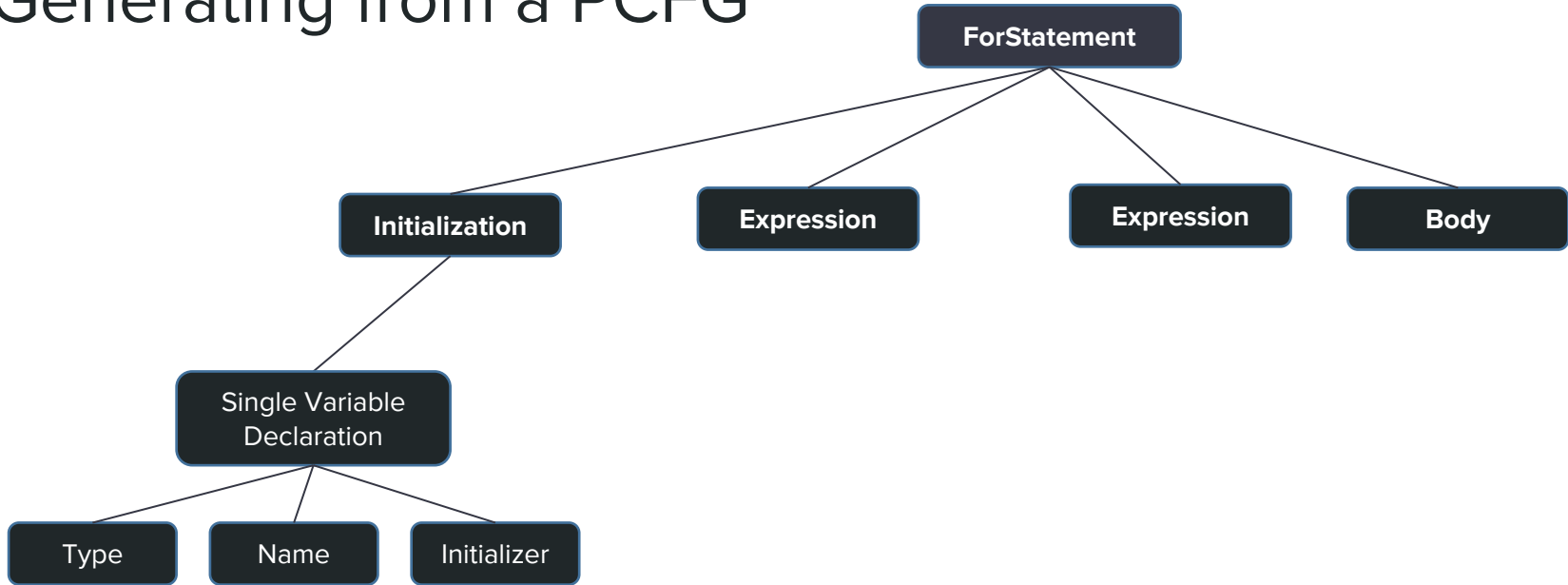
Generating from a PCFG



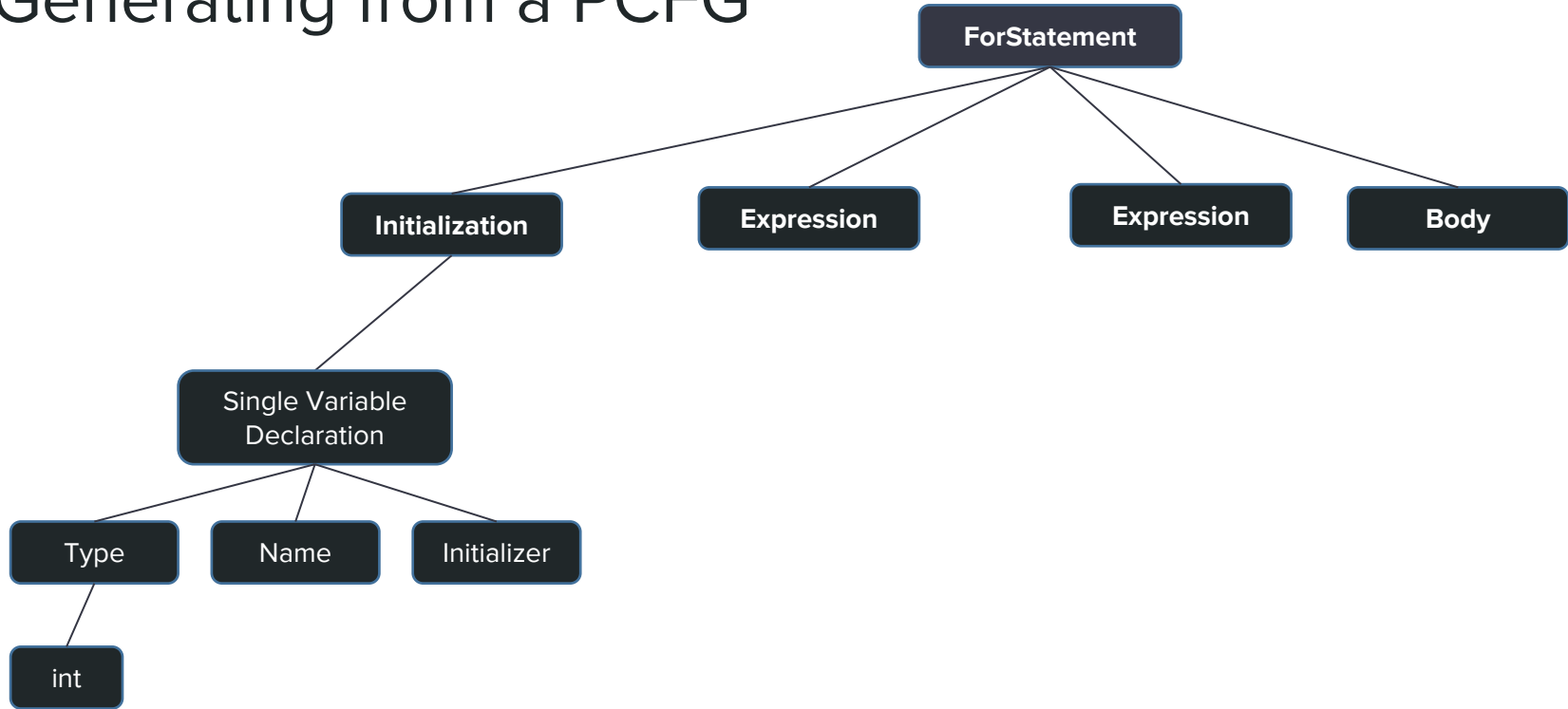
Generating from a PCFG



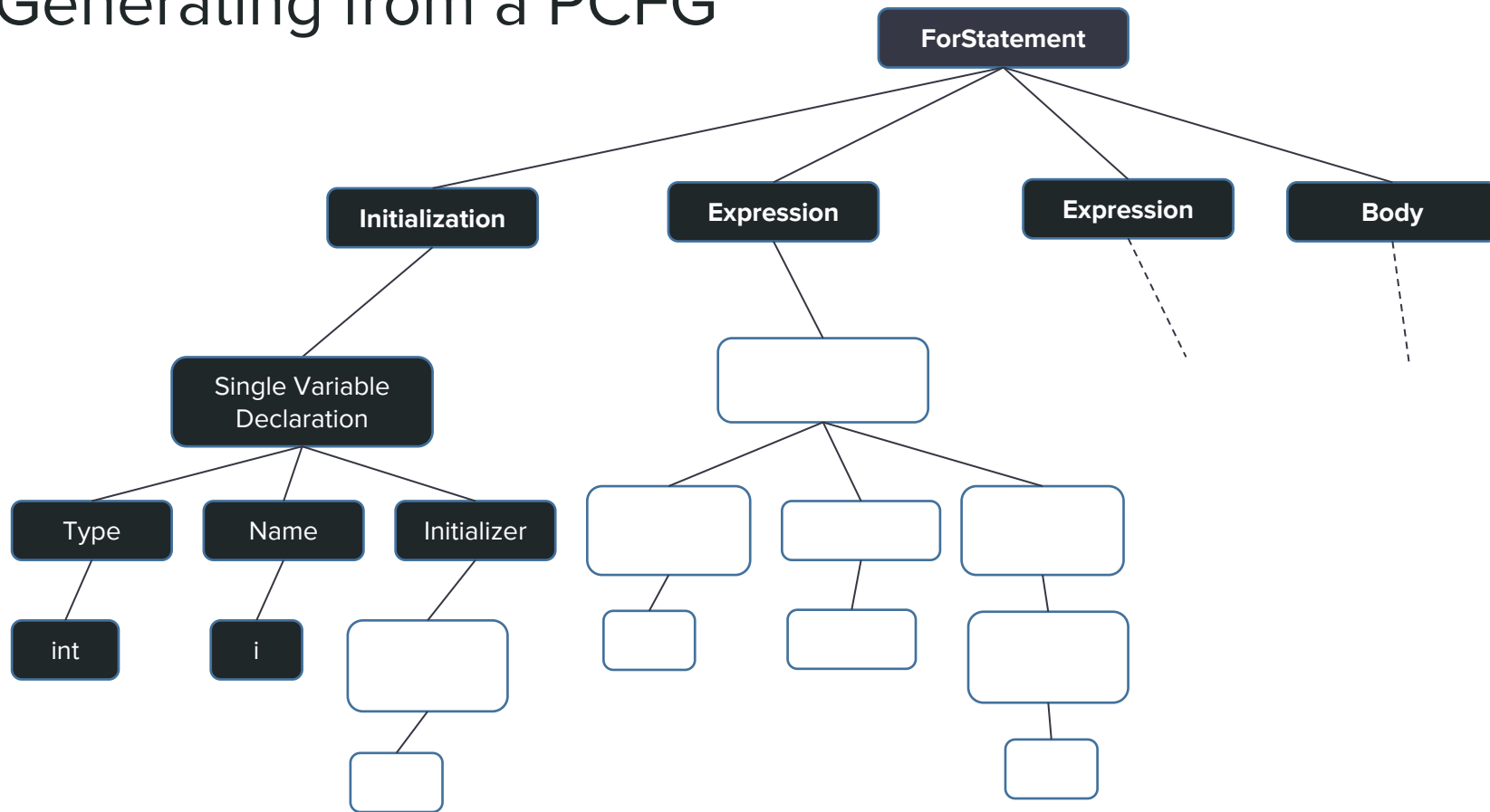
Generating from a PCFG



Generating from a PCFG



Generating from a PCFG



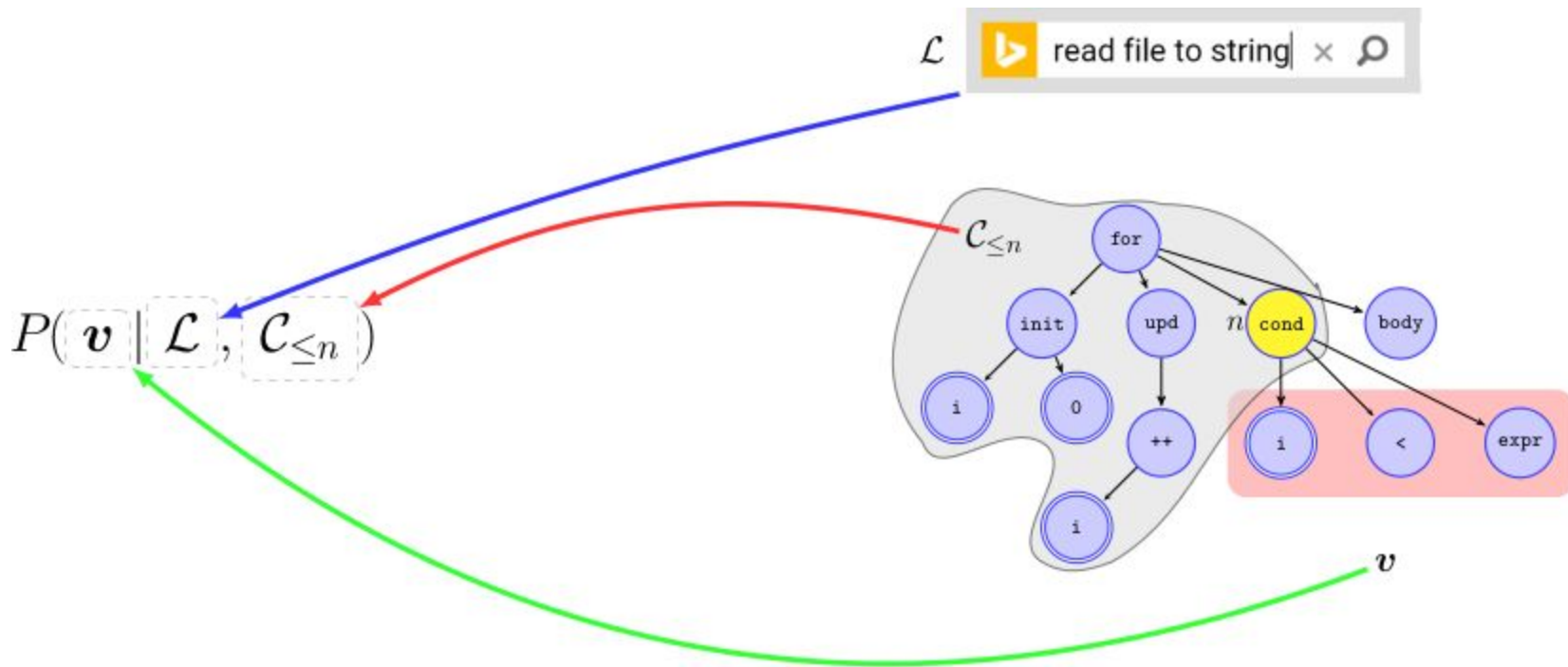
Conditional Generative Model of Source Code



Given natural language,
get a model that can
generate (probabilistically)
source code, *i.e.*

$P(\text{code} \mid \text{natural language})$

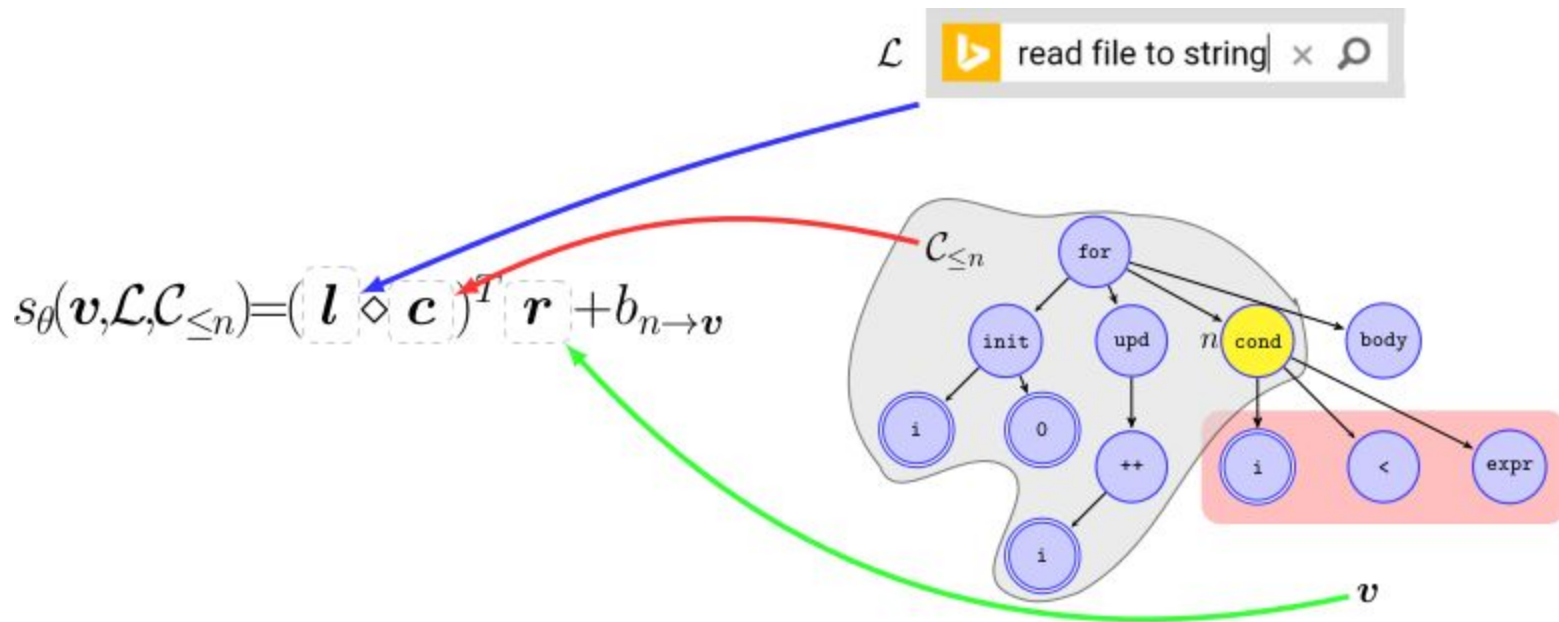
A Neural Log-Bilinear Bimodal Model of Code



Kiros, Ryan, Ruslan Salakhutdinov, and Rich Zemel. "Multimodal neural language models."

Maddison, Chris and Daniel Tarlow. "Structured generative models of natural source code."

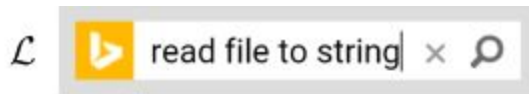
A Neural Log-Bilinear Bimodal Model of Code



Kiros, Ryan, Ruslan Salakhutdinov, and Rich Zemel. "Multimodal neural language models."

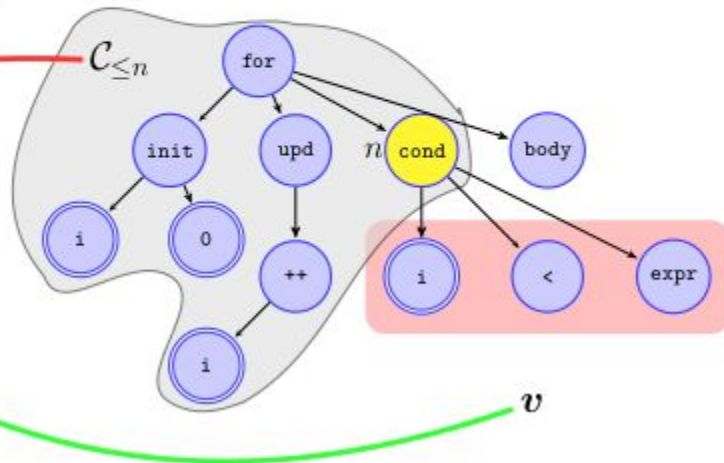
Maddison, Chris and Daniel Tarlow. "Structured generative models of natural source code."

A Neural Log-Bilinear Bimodal Model of Code



$$P(\mathbf{v} | \mathcal{L}, \mathcal{C}_{\leq n}) \propto \exp s_{\theta}(\mathbf{v}, \mathcal{L}, \mathcal{C}_{\leq n})$$

$$s_{\theta}(\mathbf{v}, \mathcal{L}, \mathcal{C}_{\leq n}) = (\mathbf{l} \diamond \mathbf{c})^T \mathbf{r} + b_{n \rightarrow v}$$



Kiros, Ryan, Ruslan Salakhutdinov, and Rich Zemel. "Multimodal neural language models."

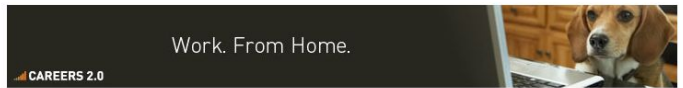
Maddison, Chris and Daniel Tarlow. "Structured generative models of natural source code."

StackOverflow Data & Augmenting Data with **bing** Queries



Stack Overflow is a question and answer site for professional and enthusiast programmers. It's 100% free, no registration required.

Make first letter of a string upper case



I have a `DetailsView` with a `TextBox` and I want the *input data* be saved always with the **FIRST LETTER IN CAPITAL**.

Example:

```
"red" --> "Red"
"red house" --> " Red house"
```

How can I achieve this **maximizing performance**?

C#

share | improve this question

edited Jul 9 at 14:35

asked Nov 9 '10 at 15:24

GibboK 9,262 25 144 283

19 Answers

active oldest votes

```
public static string FirstCharToUpper(string input)
{
    if (String.IsNullOrEmpty(input))
        throw new ArgumentException("ARGH!");
    return input.First().ToString().ToUpper() + String.Join("", input.Skip(1));
}
```

EDIT: This version is shorter. For a faster solution take a look at [Equiso's answer](#)

```
public static string FirstCharToUpper(string input)
{
    if (String.IsNullOrEmpty(input))
        throw new ArgumentException("ARGH!");
    return input.First().ToString().ToUpper() + input.Substring(1);
}
```

share | improve this answer

edited Jul 1 at 21:21

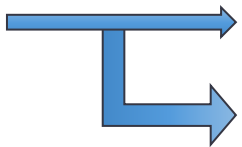
answered Dec 10 '10 at 5:12

Carlos Muñoz 6,238 3 28 57

Natural Language "Query"



Code Snippets



How do I enumerate an enum

1866

```
foreach (Suit suit in (Suit[]) Enum.GetValues(typeof(Suit)))  
{  
}
```



share | improve this answer

edited Nov 7 '13 at 0:00



Zain Rizvi
4,418 ● 8 ● 24 ● 57

answered Sep 19 '08 at 20:37



jop
32.9k ● 8 ● 36 ● 50

C# Configuration Manager . ConnectionStrings

21

Check your `machine.config`. If you only want your entry, you can add a `<clear />` element to the `<connectionStrings>` element like so ...



```
<connectionStrings>  
  <clear />  
  <add name="Target"  
    connectionString=  
      "server=MYSERVER; Database=MYDB; Integrated Security=SSPI;" />  
</connectionStrings>
```

share | improve this answer

edited Jul 16 '12 at 14:36



SteveC
2,783 ● 5 ● 33 ● 66

answered May 13 '10 at 15:36



Ryan Rinaldi
1,681 ● 8 ● 19

30K C# Questions
40K C# Snippets




How do I enumerate an enum

1866

```
foreach (Suit suit in (Suit[]) Enum.GetValues(typeof(Suit)))  
{  
}
```

share | improve this answer

edited Nov 7 '13 at 0:00
 Zain Rizvi
4,418 ● 8 ● 24 ● 57

answered Sep 19 '08 at 20:37
 jop
32.9k ● 8 ● 36 ● 50

<http://stackoverflow.com/questions/105372/how-do-i-enumerate-an-enum>

C# enum

C# foreach enum

C# enumerate enum

How to order enum values in C#

foreach enum C#

C# enumerate enumerations

C# enumeration

C# enumerator

C# foreach enum values

C# enumerate



40,092 C# Snippets
6,355,393 Natural Language Queries

Stack Overflow is a question and answer site for professional and enthusiast programmers. It's 100% registration required.

Make first letter of a string upper case

Work From Home.

CAREERS 2.0

I have a `DetailView` with a `TextBox` and I want the input data be saved always with the FIRST IN CAPITAL.

Example:

```
"red" -> "Red"
"red house" -> " Red house"
```

How can I achieve this **maximizing performance**?

share | improve this question | edited Jul 9 at 14:35 | asked Nov 9 '10 at 15:29

possible duplicate of [Capitalizing word in a string](#) – Bobby Nov 9 '10 at 15:29

questions corrected thanks – [Gibbok](#) Nov 9 '10 at 15:30

4 @Bobby: It's not a duplicate: the OP asks to capitalize the first letter of a string, the question in the link capitalizes the first letter of each word – [GvS](#) Nov 9 '10 at 16:31

1 @GvS: The first answer is very detailed and the first code-block is exactly what he is looking for. Also capitalizing every word and only the first word is just one step difference – Bobby Nov 9 '10 at 17:22

1 But you said "and I quote: "Make first letter of EACH WORD upper case". Therefore, why "red house house"? Why the "n" of "house" is not a capital letter? – [Guillermo Guzmán](#) Oct 30 '12 at 20:02

show 2 more comments

19 Answers

```
public static string FirstCharToUpper(string input)
{
    if (String.IsNullOrEmpty(input))
        throw new ArgumentException("ARGH!");
    return input.First().ToString().ToUpper() + String.Join("", input.Skip(1));
}

public static string FirstCharToUpper(string input)
{
    if (String.IsNullOrEmpty(input))
        throw new ArgumentException("ARGH!");
    return input.First().ToString().ToUpper() + input.Substring(1);
}
```

EDIT: This version is shorter. For a faster solution take a look at [Equis's answer](#)

share | improve this answer | edited Jul 1 at 21:21 | answered Dec 10 '11

6 that is a good one! – [Gibbok](#) Dec 16 '10 at 20:27

5 I like this solution! – [Eric Fick](#) Oct 21 '11 at 19:34

1 Because first parameter of `String.Join` is separator with which to join strings given with second parameter – [Dialekticus](#) Feb 10 '12 at 16:11

7 I really like your answer, but `var arr = input.ToCharArray(); arr[0] = Char.ToUpperInvariant(arr[0]); return new String(arr);` would probably gain some speed since you are creating less in objects (and especially you are skipping the `String.Join`). This of course depends on the length of string – [Hindenberg](#) Aug 26 '13 at 14:22

2 Awesome - Using `Linq` makes it very clear what this code does. – [Daniel James Dwyers](#) Nov 27 '13 at 14:22

show 4 more comments

CAREERS 2.0 by stackoverflow + Have projects on BitBucket? Import them easily to your profile

```
public string FirstLetterToUpper(string str)
{
    if (str == null)
        return null;
}
```

Performance Metric: Mean Reciprocal Rank

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{RANK}(i)}$$

Measures how well we rank the *correct* answer

Retrieval Evaluation - MRR Performance

Code Retrieval	Model	StackOverflow Test 1	StackOverflow Test 2
	NL+Code	0.18	0.17
	NL only	0.12	0.13

Query Retrieval	Model	StackOverflow Test 1	StackOverflow Test 2
	Multiplicative	0.43	0.41
	NL only	0.25	0.26

Test 1: Code snippets from training set with new natural language queries.

Test 2: New code snippets and new natural language queries.

Synthesis Samples

> **timespan day the week**

```
DateTime DateTime=DateTime.Now(0);
```

> **file exists on directory**

```
var path = new File(directory)
```

Synthesis Samples

> timespan day the week

```
DateTime DateTime=DateTime.Now(0);
```

```
foreach(string s in Days(new DateTime(2010, 1, 1), new DateTime(2010, 2, 1))) {  
    Console.WriteLine(s);  
}
```

> file exists on directory

```
var path = new File(directory)
```

```
System.IO.File.Exists(path)
```

Retrieval Sample



```
path=Path.GetFullPathInternal(path);  
new FileIOPermission(  
    FileIOPermissionAccess.Read,  
    new string[] { path },  
    false, false).Demand();  
flag = InternalExists(path);
```

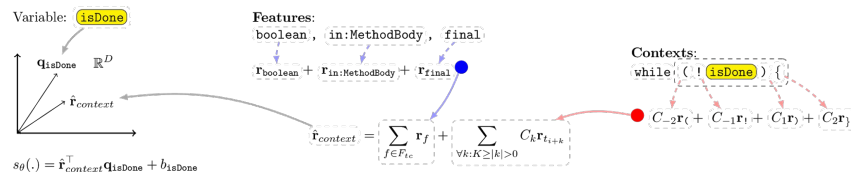
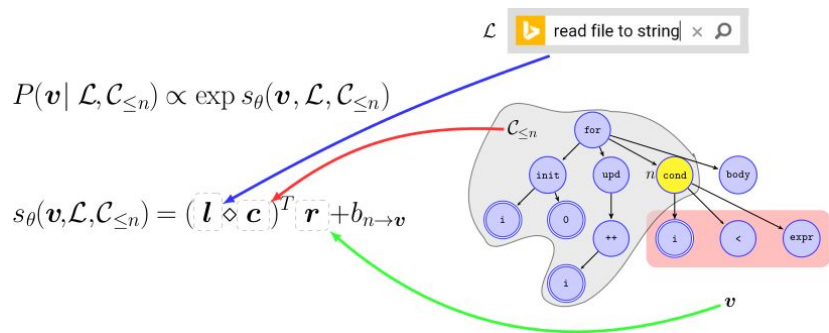
1. wpf get directory name from path
2. **determine a file exist on shared folder**
3. open file dialog class
4. create directory pathname
5. load binary file to variable

Challenges

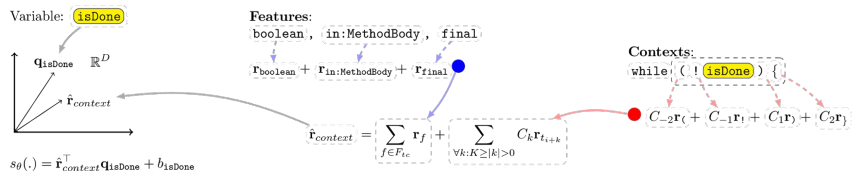
- › Code Representations in Machine Learning
- › Define Representative Evaluation Metrics for Software Engineering Tasks
- › Create Useful and Efficient Software Engineering Tools



Understanding Source Code through Machine Learning to Create Smart Software Engineering Tools



(end)



```

1 public void execute(Runnable task) {
2     if (task == null)
3         throw new NullPointerException();
4     ForkJoinTask<?> job;
5     if (task instanceof ForkJoinTask<?>) // avoid re-wrap
6         job = ((ForkJoinTask<?>) task);
7     else
8         job = new ForkJoinTask.AdaptedRunnableAction(task);
9     doSubmit(job);
10 }

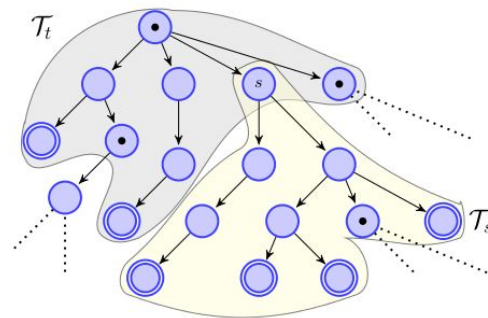
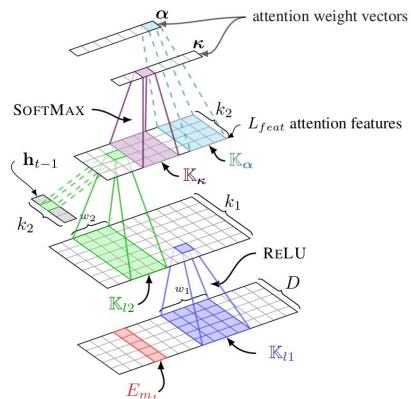
```

Learning Naming Conventions

Allamanis et al, 2014; 2015

n-gram LMs for Code

Allamanis et al, 2013



Learning to name source code

Mining Source Code Idioms

Allamanis and Sutton, 2014

The Sympathetic Uniqueness Principle

Rare names often usefully signify unusual functionality, and need to be preserved.

- Prune rare words
- Repurpose special UNK token
- Allows Naturalize to decide when it should not suggest

```
public void execute(Runnable task) {
    if (task == null)
        throw new NullPointerException();
    ForkJoinTask<?> job;
    if (task instanceof ForkJoinTask<?>) // avoid
        job = (ForkJoinTask<?>) task;
    else
        job = new ForkJoinTask.AdaptedRunnableActi
        externalPush(job);
}
```

Idioms vs. the Rest

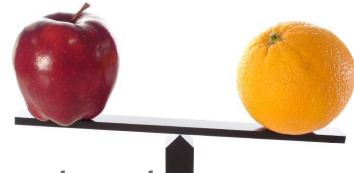
Code Clones copy-paste code fragments

- C. K. Roy et al. Comparison and evaluation of code clone detection techniques and tools: A qualitative approach. Science of Computer Programming, 2009.
- L. Jiang et al. Deckard: Scalable and accurate tree-based detection of code clones. ICSE 2007.
- H. A. Basit and S. Jarzabek. A data mining approach for detecting higher-level clones in software. IEEE Transactions on Software Engineering, 2009.

API Patterns usage patterns of methods

- T. T. Nguyen et al. Graph-based mining of multiple object usage patterns. ESEC/FSE 2009.
- J. Wang et al. Mining succinct and high-coverage API usage patterns from source code. MSR 2013.
- H. Zhong et al. MAPO: Mining and recommending API usage patterns. ECOOP, 2009.

Idioms syntactic code fragments



The Distributional Hypothesis

“You shall know a word by the company it keeps”.

John Rupert Firth, 1957

The Distributional Hypothesis

The ???????? is walking

“You shall know a word by the company it keeps”.

For IWESEP, in a way I'm inviting you as a ML/NLP person that can teach software engineering people the cool things you can do with ML/NLP techniques. I'm not sure if you see yourself this way, but I think a quick "intro to ML/NLP" for the first 1/3 or so, then "look at all the cool things you can do" for the second 2/3 would be one potential way to give the presentation.

Sanity Check:

String Manipulation Synthetic Data

```
var result = input_string.Split(' ').Select((string x) =>  
Double.parse(x)).Average();
```

each element parse double separated by a space and get mean

each element parse double separated by a space and get average

each element convert to double separated by a space and get mean

each element convert to double separated by a space and get average

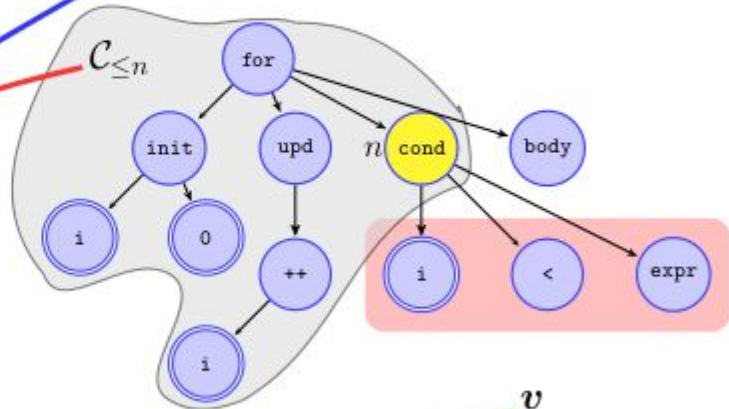
each element parse to double separated by a space and get mean

A Neural Log-Bilinear Bimodal Model of Code

$$\mathbf{c} = \sum_{j=1}^J \mathbf{H}_j \mathbf{c}_{\phi_j}$$



$$\mathbf{l} = \frac{1}{|\mathcal{L}|} \sum_{w \in \mathcal{L}} \mathbf{l}_w$$



$$s_{\theta}(\mathbf{v}, \mathcal{L}, \mathcal{C}_{\leq n}) = (\mathbf{l} \diamond \mathbf{c})^T \mathbf{r} + b_{n \rightarrow v}$$

Kiros, Ryan, Ruslan Salakhutdinov, and Rich Zemel. "Multimodal neural language models."

Maddison, Chris and Daniel Tarlow. "Structured generative models of natural source code."