Suggesting Accurate Method and Class Names

Miltiadis Allamanis† Earl T. Barr‡ Christian Bird⋆ Charles Sutton†
†School of Informatics
University of Edinburgh
Edinburgh, EH8 9AB, UK
{m.allamanis, csutton}@ed.ac.uk
dsutton@ucl.ac.uk
cbird@microsoft.com
‡Dept. of Computer Science
University College London
London, UK
e.barr@ucl.ac.uk

ABSTRACT

Descriptive names are a vital part of readable, and hence maintainable, code. Recent progress on automatically suggesting names for local variables tantalizes with the prospect of replicating that success with method and class names. However, suggesting names for methods and classes is much more difficult. This is because good method and class names need to be functionally descriptive, but suggesting such names requires that the model goes beyond local context. We introduce a neural probabilistic language model for source code that is specifically designed for the method naming problem. Our model learns which names are semantically similar by assigning them to locations, called embeddings, in a high-dimensional continuous space, in such a way that names with similar embeddings tend to be used in similar contexts. These embeddings seem to contain semantic information about tokens, even though they are learned only from statistical co-occurrences of tokens. Furthermore, we introduce a variant of our model that is, to our knowledge, the first that can propose neologisms, names that have not appeared in the training corpus. We obtain state of the art results on the method, class, and even the simpler variable naming tasks. More broadly, the continuous embeddings that are learned by our model have the potential for wide application within software engineering.

Categories and Subject Descriptors:
D.2.3 [Software Engineering]: Coding Tools and Techniques
General Terms: Algorithms
Keywords: Coding conventions, naturalness of software

“You shall know a word by the company it keeps.” — J. R. Firth

1. INTRODUCTION

Language starts with names. While programming, developers must name variables, parameters, functions, classes, and files. They strive to choose names that are meaningful and conventional, i.e., consistent with other names used in related contexts in their code base. Indeed, leading industrial experts, including Beck [9], McConnell [34], and Martin [33], have stressed the importance of identifier naming in software. Finding good names for programming language constructs is difficult; poor names make code harder to understand and maintain [29, 50, 30, 7]. Empirical evidence suggests that poor names lead to software defects [13, 1]. Code maintenance exacerbates the difficulty of finding good names, because the appropriateness of a name changes over time: an excellent choice, at the time a construct is introduced, can degrade into a poor name, as when a variable is used in new context or a function’s semantics changes.

Names of methods and classes are particularly important, and can be difficult to choose. Høst et al. eloquently captured their importance: “Methods are the smallest named units of aggregated behavior in most conventional programming languages and hence the cornerstone of abstraction” [26]. Semantically distinct method names are the basic tools for reasoning about program behaviour. Programmers directly think in terms of these names and their compositions, since a programmer chose them for the units into which the programmer decomposed a problem. Moreover, method names can be hard to change, especially when they are used in an API. When published in a popular library, method naming decisions are especially rigid and poor names can doom a project to irrelevance.

In this paper, we suggest that modern statistical tools allow us to automatically suggest descriptive, idiomatic method and class names to programmers. We tackle the method naming problem: the problem of inferring a method’s name from its body (or a class from its methods). As developers spend approximately half of their development time trying to understand and comprehend code during maintenance alone [17], any progress toward solving the method naming problem will improve the comprehensibility of code [49] leading to an increase programmer productivity [24].

In previous work, we introduced the NATURALIZE framework [2], which learns the coding conventions used in a code base and tackles one naming problem programmers face — that of naming variables — by exploiting the “naturalness” or predictability of code [25]. However, the method naming problem is much more difficult than the variable naming problem, because the appropriateness of method and class names depends not solely on their uses but also on their internal structure — their body or their set of methods. An adequate name must describe not just what the method is, but what it does. Variable names, by contrast, can often be predicted solely from a few tokens of local context; for example, it is easy to predict the variable name that follows the tokens for ( int. Because method and class names must be functionally descriptive, they often have rich internal structure: method names are often verb phrases and class names are often noun phrases. But this means that method and class names are often neologisms, that is, names not seen in the training corpus. Existing probabilistic models of source code, including the n-gram models used in NATURALIZE, cannot suggest neologisms. These aspects of the method naming problem severely exacerbate the data sparsity problem faced by all probabilistic language models, because addressing them by building models that consider more context necessarily means that any individual context will be observed less often. Therefore, the method naming problem requires models that...
private void createDefaultShader () {
    String vertexShader = "literal_1";
    String fragmentShader = "literal_2";
    shader = new ShaderProgram(vertexShader, fragmentShader);
    if (shader.isCompiled() == false)
        throw new IllegalArgumentException("" + shader.getLog());
}

Figure 1: A method from libgdx's CameraGroupStrategy. A programmer named it; automatically naming it requires inventing a neologism, a very hard inference problem. Our subtoken model understands that its name should start with create and suggests createShaders.

Our suggestion model can be embedded within a variety of tools to support code development and code review. During development, suppose that the developer is adding a method or a class to an existing project. After writing the body, the developer may be unsure if the name she chose is descriptive and conventional within the project. Our model suggests alternative names from patterns it learned from other methods in the project. During code review, our model can highlight those names to which our model assigns a low score. In either case, the system has two phases: a training phase, which takes as input a training set of source files (e.g. the current revision of the project) and returns a neural network model that can suggest names; and a testing or deployment phase, in which the input is a trained neural network and the source code of a method or class, and the output is a ranked list of suggested names. Any suggestion system has the potential to suffer from what we have called the “Clippy effect” [2], in which too many low quality suggestions alienate the user. To prevent this, our suggestion model also returns a numeric score that reflects its degree of confidence in its suggestion; practical tools would only make a suggestion to the user if the confidence were sufficiently high.

2. NEURAL CONTEXT MODELS OF CODE

In this section, we introduce four language models of code, starting with the n-gram model to build intuition. Then we introduce
neural probabilistic language modelling and follow it with two novel models that, specifically designed for method naming, refine the underlying neural model: our logbilinear context model, which adds context and features, and subtoken context model, which adds subtokens and can be used to generate neologisms.

Language models (LM) are probability distributions over strings of a language. These models assume that we are trying to predict a token $t$ given a sequence of other tokens $c = (c_0, c_1, \ldots, c_N)$ that we call the context. LMs are very general; for example, if the goal is to sequentially predict every token in a file, as a n-gram model does, then we can take $t = y_m$ and $c = (y_{m-n+1}, y_{m-n+2}, \ldots, y_{m-1})$. Alternatively, for the method naming problem, we can take $t$ to be the identifier token in the declaration that names the function, and $c$ to be a sequence that contains all identifiers in the function body. Obviously, we cannot store a probability value for every possible context, so we must make simplifying assumptions to make the modeling tractable. Different LMs make different simplifying assumptions.

2.1 Background

To build intuition, we begin by reviewing the n-gram LM, which is a standard technique in NLP and speech processing, and which has become increasingly popular in software engineering [25, 41, 2]. The n-gram model assumes that all of the information required to predict the next token is contained within the previous $n-1$ tokens i.e. $P(y_1 \ldots y_M) = \prod_{m=1}^{M} P(y_m | y_{m-1} \ldots y_{m-n+1})$. To specify this model we need (in principle) a table of $V^n$ numbers, where $V$ is the number of possible lexemes, that specifies the conditional probabilities for each possible n-gram. These are the parameters of the model that we learn from data.

There is a large literature on methods for training these models [16], which basically revolve around counting the proportion of times that token $y_m$ follows $y_{m-1} \ldots y_{m-n+1}$. However, even when $n = 4$ or $n = 5$, we cannot expect to estimate the counts of all n-grams reliably, as the number of possible n-grams is exponential in $n$. Therefore, smoothing methods are employed, which generally modify the count of a rare n-gram $y_1 \ldots y_n$ to make it more similar to the count of a shorter suffix $y_2 \ldots y_n$, whose frequency we can estimate more reliably. This procedure involves the implicit assumption that two contexts are most similar if they share a long suffix. Such a procedure can be used to generate neologisms.

Logbilinear models Neural LMs [10] address the challenge that the simple n-gram model has by making similar predictions for similar contexts. They predict the next token $y_m$ using a neural network that takes the previous tokens as input. This allows the network to flexibly learn which tokens, like int, provide much information about the immediately following token, and which tokens, like the semicolon ‘;’, provide very little. Unlike a n-gram model, a neural LM makes it easy to add general long-distance features of the context into the prediction — we simply add them as additional inputs to the neural net. In our work, we focus on a simple type of neural LM that has been effective in practice, namely, the log-bilinear LM [37] (LBL). We start with a general treatment of loglinear models considering models of the form

$$P(t|c) = \frac{\exp(s_g(t,c))}{\sum \exp(s_g(t',c))}. \quad (1)$$

Intuitively, $s_g$ is a function that indicates how much the model likes to see both $t$ and $c$ together, the exp function maps this to be always positive, and the denominator ensures that the result is a probability distribution. This choice is very general. For example, if $s_g$ is a linear function of the features in $c$, then the discriminative model is simply a logistic regression.

Logbilinear models learn a map from every possible target $t$ to a vector $q_t \in \mathbb{R}^D$, and from each context $c$ to a vector $r_c \in \mathbb{R}^D$. We interpret these as locations of each context and each target lexeme in a $D$-dimensional space; these locations are called embeddings. The model predicts that the token $t$ is more likely to appear in context $c$ if the embedding $q_t$ of the token is similar to that $r_c$ of the context. To encode this in the model, we choose

$$s_g(t,c) = r_c^T q_t + b_t, \quad (2)$$

where $b_t$ is a scalar bias which represents how commonly $t$ occurs regardless of the context. To understand this equation intuitively, note that, if the vectors $r_c$ and $q_t$ have norm 1, then their dot product is simply the cosine of the angle between them. So $s_g$, and hence $p(t|c)$, is larger if either vector has a large norm, if $b_t$ is large, or if $r_c$ and $q_t$ have a small angle between them, that is, if they are more similar according to the commonly used cosine similarity metric.

To complete this description, we define the maps $t \mapsto q_t$ and $c \mapsto r_c$. For the targets $t$, the most common choice is to simply include the vector $q_t$ for every $t$ as a parameter of the model. That is, the training procedure has the freedom to learn an arbitrary map between $t$ and $q_t$. For the contexts $c$, this choice is not possible, as there are too many possible contexts. Instead, a common choice [31, 39] is to represent the embedding $r_c$ of a context as the sum of embeddings of the tokens within it, that is,

$$r_c = \sum_{i=1}^{C} C_i r_{c_i}, \quad (3)$$

where $r_{c_i} \in \mathbb{R}^D$ is a vector for each lexeme $c_i$ that is included in the model parameters. The variable $i$ indexes every token in the context $c$, so if the same lexeme occurs multiple times in $c$, then it appears multiple times in the sum. The matrix $C_i$ is a diagonal matrix that serves as a scaling factor depending on the position of a lexeme within the context. This allows, for example, a lexeme’s influence on $c_i$’s position to depend on how close it is to the target. The $D$ non-zero values in $C_i$ for each $c$ are also included in the model parameters. Each lexeme $v$ has two embeddings: an embedding $q_t$ for when it is used as a target and an embedding $r_c$ for when it appears in the context.

To summarize, logbilinear models make the assumption that every token and every context can be mapped in a $D$-dimensional space. There are two kinds of embedding vectors: those directly learned (i.e. the parameters of the model) and those computed from the parameters of the model. To indicate this distinction, we place a hat on $r_c$ to indicate that it is computed from the model parameters, whereas we write $q_t$ without a hat to indicate that it is a parameter vector that is learned directly by the training procedure. These models can also be viewed as a three-layer neural network, in which the input layer encodes all of the lexemes in $c$ using a 1-of-$V$ encoding, the hidden layer outputs the vectors $r_c$, for each token in the context, and the output layer computes the score functions $s_g(t,c)$ and passes them to a softmax nonlinearity. For details on the neural network representation, see Bengio et al. [10].

To learn these parameters, it has recently been shown [39, 38] that an alternative to the maximum likelihood method called noise contrastive estimation (NCE) [21] is effective. NCE measures how well the model $p(t|c)$ can distinguish the real data in the training set from “fantasy data” that is generated from a simple noise distribution. At a high level, this can be viewed as a black box alternative to maximum likelihood that measures how well the model fits the training data. We optimize the model parameters using stochastic gradient descent. We employ NCE for all models in this paper.

2.2 Logbilinear Context Models of Code

Now we present a new neural network, a novel LBL LM for code, which we call a logbilinear context model. The key idea
is that loglinear models make it especially easy to exploit long-distance information; e.g., when predicting the name of a method, it is useful to take into account all of the identifiers that appear in the method body. We model long-distance context via a set of feature functions, such as "Whether any variable in the current method is named addCount", "Whether the return type of the current method is int," and so on. The loglinear context model combines these features with the local context.

As before, suppose that we are trying to predict a code token \( t \) given a sequence of context tokens \( c = (c_0, c_1, \ldots, c_N) \). We assume that \( c \) contains all of the other tokens in the file that are relevant for predicting \( t \); e.g., tokens from the body of the method that \( t \) names. The tokens in \( c \) that are nearest to the target \( t \) are treated specially. Suppose that \( t \) occurs in position \( i \) of the file, that is, if the file is the token sequence \( t_1, t_2, \ldots, t_i, t \), then \( t = t_i \). Then the local context is the set of tokens that occur within \( K \) positions of \( i \), that is, the set \( \{ t_{i-K} \} \) for \( -K \leq k \leq K, k \neq 0 \). The local context includes tokens that occur both before and after \( t \).

The overall form of the context model will follow the generic form in (1) and (2), except that the context representation \( \mathbf{f}_c \) is defined differently. In the context model, we define \( \mathbf{f}_c \) using two different types of context: local and global. First, the local context is handled in a very similar way to the loglinear L.M. Each possible lexeme \( v \) is assigned to a vector \( \mathbf{r}_v \in \mathbb{R}^D \), and, for each token \( t \) that occurs within \( K \) tokens of \( t \) in the file, we add its representation \( \mathbf{r}_v \) into the context representation.

The global context is handled using a set of features. Each feature is a binary function based on the context tokens \( c \), such as the examples described at the beginning of this section. Formally, each feature \( f \) maps a \( c \) value to either 0 or 1. Maddison and Tarlow [31] use a similar idea to represent features of a syntactic context, that is, a node in an AST. Here, we extend this idea to incorporate arbitrary features of long-distance context tokens \( c \). The first column of Table 4 presents the full list of features that we use in this work.

To learn an embedding, we assign each feature function to a single vector in the continuous space, in the same way as we did for tokens. Mathematically, let \( F \) be the set of all features in the model, and let \( \mathbf{F}_c \), for a context \( c \), be the set of all features \( f \) with \( f(c) = 1 \). Then for each feature \( f \in F \), we learn an embedding \( \mathbf{r}_f \in \mathbb{R}^D \), which is included as a parameter to the model in exactly the same way that \( \mathbf{r}_v \) was for the language modeling case.

Now, we can formally define a context model of code as a probability distribution \( P(t | c) \) that follows the form (1) and (2), where

\[
\mathbf{f}_c = \sum_{f \in \mathbf{F}_c} \mathbf{r}_f + \sum_{k \leq |c|} \mathbf{C}_k \mathbf{r}_{c+k+1}
\]

for each token \( t \), for \( i \in I_t \), then average them. When we do this, we carefully rename all occurrences of \( t \) to a special token called SELFF to remove \( t \) from its own context.

**2.3 Subtoken Context Models of Code**

A limitation of all of the previous models is that they are unable to predict neologisms, that is, unseen identifier names that have not been used in the training set. The reason for this is that we allow the map from a lexeme \( v \) to its embedding \( \mathbf{r}_v \) to be arbitrary (i.e., without learning a functional form for the relationship), so we have no basis to assign continuous vectors to identifier names that have not been observed. In this section, we sidestep this problem by exploiting the internal structure of identifier names, resulting in a new model which we call a subtoken context model.

\[1\] Note that \( k \) can be positive or negative, so that in general \( C_{-2} \neq C_2 \).
The subtoken context model exploits the fact that identifier names are often formed by concatenating words in a phrase, such as `getLocation` or `setContentLengthHeader`. We call each of the smaller words in an identifier a subtoken. We split identifier names into subtokens based on camel case and underscores, resulting in a set of subtokens that we use to compose new identifiers. To do this, we exploit the summation trick we used in $r_{\text{context}}$. Recall that we constructed this vector as a sum of embedding vectors for particular features in the context. Here, we define the embedding of a target vector to be the sum of the embeddings of its subtokens.

Let $r$ be the token that we are trying to predict from a context $c$. As in the context model, $c$ can contain tokens before and after $r$, and tokens from the global context. In the subtoken model, we additionally suppose that $i$ is split up into a sequence of $M$ subtokens, that is, $i = s_1 s_2 \ldots s_M$, where $s_M$ is always a special END subtoken that signifies the end of the subtoken sequence. That is, the context model now needs to predict a sequence of subtokens in order to predict a full identifier. We begin by breaking up the prediction one subtoken at a time, using the chain rule of probability:

$$P(s_1 s_2 \ldots s_M | c) = \prod_{m=1}^{M} P(s_m | s_{1 \ldots m-1}, c).$$

Then, we model the probability $P(s_m | s_{1 \ldots m-1}, c)$ of the next subtoken $s_m$ given all of the previous ones and the context. Since preliminary experiments with an $n$-gram version of a subtoken model showed that $n$-grams did not yield good results, we employ a bilinear model

$$P(s_m | s_1 \ldots s_{m-1}, c) = \frac{\exp(s_g(s_m, s_1 \ldots s_{m-1}, c))}{\sum_c \exp(s_g(s_m, s_1 \ldots s_{m-1}, c))}.$$  \hspace{1cm} (5)

As before, $s_g(s_1 s_2 \ldots s_{m-1}, c)$ can be interpreted as a score, which can be positive or negative and indicates how much the model “likes” to see the subtoken $s_m$, given the previous subtokens and the context. The exponential functions and the denominator are a mathematical device to convert the score into a probability distribution.

We choose a bilinear form for $s_g$, with the difference being that in addition to tokens having embedding vectors, subtokens have embeddings as well. Mathematically, we define the score as

$$s_g(s_m, s_1 \ldots s_{m-1}, c) = \hat{r}_{\text{context}}^T \hat{D}_m q_{s_m} + b_{s_m},$$  \hspace{1cm} (6)

where $q_{s_m} \in \mathbb{R}^D$ is an embedding for the subtoken $s_m$, and $\hat{r}_{\text{context}}$ is a continuous vector that represents the previous subtokens and the context. To define a continuous representation $\hat{r}_{\text{context}}$ of the context, we break this down further into a sum of other embedding features as

$$\hat{r}_{\text{context}} = \hat{r}_{\text{context}^+} + \hat{r}_{\text{context}^-}.$$  \hspace{1cm} (7)

In other words, the continuous representation of the context breaks down into a sum of two vectors: the first term $\hat{r}_{\text{context}^+}$ represents the effect of the surrounding tokens $c$ — both local and global — and is defined exactly as in the context model via (4).

The new aspect is how we model the effect of the previous subtokens $s_1 \ldots s_{m-1}$ in the second term $\hat{r}_{\text{context}^-}$. We handle this by assigning each subtoken $s$ a second embedding vector $r_s \in \mathbb{R}^D$ that represents its influence when used as a previous subtoken; we call this a history embedding. We weight these vectors by a diagonal matrix $C_{m}^{-} - k$ to allow the model to learn that subtokens have decaying influence the farther they are from the token that is being predicted. Putting this all together, we define

$$\hat{r}_{\text{context}^-} = \sum_{i=1}^{M} C_{m-i}^{-} r_{s_{m-i}}.$$  \hspace{1cm} (8)

This completes the definition of the subtoken context model. To sum up, the parameters of the subtoken context model are (a) the target embeddings $q_{s_m}$ for each subtoken $s_m$ that occurs in the data, (b) the history embeddings $r_s$ for each subtoken $s$, (c) the diagonal weight matrices $C_{m}^{-}$ for $m = 1, 2, \ldots, M$ that represent the effect of distance on the subtoken history (we use $M = 3$, yielding a 4-gram-like model on subtokens) and the parameters that we carried over from the logbilinear context model: (d) the local context embeddings $r_f$ for each token $f$ that appears in the context, (e) the local context weight matrices $C_{-k}$ and $C_k$ for $-K \leq k \leq K, k \neq 0$, and (f) the feature embeddings $r_f$ for each feature $f(c)$ of the global context. We estimate all of these parameters from the training corpus.

Although this may seem like a large number of parameters, this is typical for language models, e.g., consider the $V^5$ parameters, if $V$ is the number of lexemes, required by a 5-gram language model. How can we handle so many parameters? The reason is simple: in the era of vast, publicly available source code repositories like GitHub and Bitbucket, code scarcity is a thing of the past.

Generating Neologisms

A final question is “Given the context $c$, how do we find the lexeme $r$ that maximizes $P(r|c)$?”. Previous models could answer this question simply by looping over all possible lexemes in the model, but this is impossible for a subtoken model, because there are infinitely many possible neologisms. So we employ beam search (see Russell and Norvig [44] for details) to find the $B$ tokens (i.e., subtoken sequences) with the highest probability.

2.4 Source Code Features for Context Models

In this section, we describe the features we use to capture global context. Identifying software measures and features that effectively capture semantic properties like comprehensibility or bug-proneness is a seminal software engineering problem that we do not tackle in this paper. Here, we have selected measures and features heavily used in the literature and industry. For instance, control flow is indisputably important; we selected Cyclomatic complexity, despite its correlation with code size, to measure it. The first column of Table 4 defines the features we used in this work. In the table, “Variable Type” tracks whether the type is generic, its type after erasure, and, if the type is an array, its size. “Contained Methods” and “Sibling Methods” exclude method overloads and recursion.

The features of a target token are its target features; we assign a $r_f$ vector to each of them; this vector is added in the left summation of Equation 4 if a feature’s indicator function $f$ returns 1 for a particular token. Although features are binary, we describe some — like the modifiers of a declaration, the node type of an AST, etc. — as categorical. All categorical features are converted into binary using a 1-of-K encoding. For methods, we include Cyclomatic complexity, clipping it to 10 and treating it as categorical. When features do not make sense for a particular token, like the Cyclomatic complexity of a variable, the feature’s function simply returns zero.

3. METHODOLOGY

The core challenge of solving the method naming problem from code is data sparsity. Our guiding intuition is that source code contains rich structure that can alleviate the sparsity problem. We therefore pose the following question: How can we better maximally exploit the structure inherent to source code? This question in turn leads us to the research questions:

**RQ1.** Can we identify and extract long and short-range context features of identifiers for naming?

**RQ2.** Do identifiers contain exploitable substructure?

Answering both of these questions in the affirmative, we turn our attention to exploiting the resulting naming information; here, we ask if this new information is sufficiently rich to allow us to accurately suggest names. More concretely:

**RQ3.** Can we accurately suggest method declaration names, looking only at the context of the declared method?

**RQ4.** Can we do the same for class (i.e. type) names?
In reference to the first two research questions we describe the features that we use and how we capture substructure in the next section. We then definitively answer these research questions by comparing our approach with previous techniques for suggesting variable names on a broad software corpus. There is little to no research that tackles the second two research questions to compare against. Nonetheless, we use an n-gram model as a point of comparison for naming methods and classes to demonstrate the performance of our approach as that model has performed the best for variable naming in the past and we hypothesize is reasonable for these new naming tasks. The rest of this section describes our experimental setup and methodology.

Data We picked the top active Java GitHub projects on January 22nd 2015. We obtained the most popular projects by taking the sum of the z-scores of the number of watchers and forks of each project, using the GitHub Archive. Starting from the top project, we selected the top 20 projects excluding projects that were in a domain that was previously selected. We also included only projects with more than 50 collaborators and more than 500 commits. The projects along with short descriptions are shown in Table 1. We used this procedure to select a mature, active, and diverse corpus with large development teams. Finally, we split the files uniformly into a training (70%) and a test (30%) set.

Methodology We train all models on the train sets formed over the files of each project. To evaluate the models, for each of the test files and for each variable (all identifiers that resolve to the same symbol), method declaration or class declaration, we compute the features and context of the location of the identifier and ask the model to predict the actual target token the developer used (which is unknown to the model), as in Allamanis et al. [2]. In the use cases we consider (see Section 1), models that are deployed with a “confidence filter”, that is, the model will only present a suggestion when the probability of the top ranked name is above some threshold. This is to avoid annoying the user with low-quality suggestions. To reflect this in our evaluation, we measure the degree to which the quality of the results changes as a function of the threshold. Rather than reporting the threshold, which is not directly interpretable, we instead report the suggestion frequency, which is the percentage of names in the test set for which the model decides to make a prediction for a given threshold.

To measure the quality of a suggestion, we compute the F1 score and the accuracy for the retrieval over the subtokens of each correct token. Thus, all methods are given partial credit if the predicted name is not an exact match but shares subtokens with the correct name. F1 is the harmonic mean of precision and recall and is a standard measure [32] because it conserves: as a harmonic mean, its value is influenced most by the lowest of precision and recall. We also compute the accuracy of each prediction: a prediction is correct when the model predicts exactly (exact match) the actual token. When computing the F1 score for suggestion rank \( k > 1 \), we pick the precision, recall, and F1 of the rank \( l \leq k \) that results in the highest F1 score.

Because this evaluation focuses on popular projects, the results may not reflect performance on a low quality project in which many names are poor. For such projects, we recommend training on a different project that has high quality names, but leave evaluating this approach to future work. Alternatively, one could argue that, because we measure whether the model can reproduce existing names, the evaluation is too harsh: if a predicted name does not match the existing name, it could be equally good, or even an improvement. Nonetheless, matching existing names in high quality projects, as we do, still provides evidence of overall suggestion quality, and, compared to a user study, an automatic evaluation has the great advantage that it allows efficient comparison of a larger number of different methods.

Finally, during training, we substitute rare identifiers, subtokens and features (i.e. those seen less than two times in the training data) with a special UNK token. During testing, when any of the models suggests the UNK token, we do not make any suggestions; that is, the UNK token indicates that the model expects a neologism that it cannot predict. For the subtoken model, during testing, we may produce suggestions that contain UNK subtokens. In the unlikely case that a context token \( t_{i+k} = t_i \) (i.e. the same token), we replace \( t_{i+k} \) with a special $\text{SELF}$ token. This makes sure that the context of the model includes no information about the target token.

Training Parameters We used learning rate 0.07, \( D = 50 \), minibatch size 100, dropout 20% [48], generated 10 distractors for each sample for each epoch and trained for a maximum of 25 epochs picking the parameters that achieved the maximum likelihood in a held out validation set (the 10% of the training data). The context size was set to \( K = 6 \) and subtoken context size was set to \( M = 3 \). Before the training started, parameters were initialized around 0 with uniform additive noise (scaled by \( 10^{-4} \)). The bias parameters \( b \) were initialized such that \( P(|v|) \) matches the empirical (unigram) distribution of the tokens (or subtokens for the subtoken model). All the hyperparameters except for \( D \) were tuned using Bayesian optimization on bigbluebutton for method declarations. The parameter \( D \) is special in that as we increase it, the performance of each model increases monotonically (assuming a good validation set), with diminishing returns. Also, an increase in \( D \) increases the computational complexity of training and testing each model. We picked \( D = 50 \) that resulted in a good trade-off of the computational complexity vs. performance.

4. IDENTIFIER REPRESENTATION

First we evaluate our model qualitatively, by visualizing its output. All of the models that we have described assign tokens, features, and subtokens to embeddings, which are locations in a \( D \)-dimensional continuous space. These locations have been selected by the training procedure to explain statistical properties of tokens, but it does not necessarily follow that the embeddings capture anything about the semantics of names. To explore this question, we examine qualitatively whether names that appear semantically similar to us are assigned to similar embeddings, by visualizing the continuous embeddings assigned to names from a few projects. This raises the immediate difficulty of how to visualize vectors in a \( D = 50 \) dimensional space. Fortunately, there is a rich literature in statistics and machine learning about dimensionality reduction methods that map high dimensional vectors to two-dimensional vectors while preserving important properties of the original space. There are various ideas behind such techniques, such as preserving distances or angles.

<table>
<thead>
<tr>
<th>Name</th>
<th>Git SHA</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>elasticsearch</td>
<td>d3e1019</td>
<td>REST Search Engine</td>
</tr>
<tr>
<td>Android-Universal-Image-Loader</td>
<td>19c3e39</td>
<td>Android Library</td>
</tr>
<tr>
<td>spring-framework</td>
<td>2bf6b41</td>
<td>Application Framework</td>
</tr>
<tr>
<td>libgdx</td>
<td>789a50b</td>
<td>Game Dev Framework</td>
</tr>
<tr>
<td>storm</td>
<td>bc54a8e</td>
<td>Distributed Computation</td>
</tr>
<tr>
<td>zxing</td>
<td>71d8595</td>
<td>Barcode image processing</td>
</tr>
<tr>
<td>netty</td>
<td>3b1f15e</td>
<td>Network App Framework</td>
</tr>
<tr>
<td>platform_frames_base</td>
<td>f19176f</td>
<td>Android Base Framework</td>
</tr>
<tr>
<td>bigbluebutton</td>
<td>02bc78c</td>
<td>Web Conferencing</td>
</tr>
<tr>
<td>junit</td>
<td>c730003</td>
<td>Testing Framework</td>
</tr>
<tr>
<td>rxjava</td>
<td>cf5ae70</td>
<td>Reactive JVM extensions</td>
</tr>
<tr>
<td>retrofit</td>
<td>af40044</td>
<td>REST client</td>
</tr>
<tr>
<td>clojure</td>
<td>f437b85</td>
<td>Programming Language</td>
</tr>
<tr>
<td>dropwizard</td>
<td>741a161</td>
<td>RESTful web server</td>
</tr>
<tr>
<td>okhtp</td>
<td>0a197f6</td>
<td>HTTP/SPDY client</td>
</tr>
<tr>
<td>presto</td>
<td>600e4f7</td>
<td>Distributed SQL engine</td>
</tr>
<tr>
<td>metrics</td>
<td>4984f6b</td>
<td>Metrics Framework</td>
</tr>
<tr>
<td>spring-boot</td>
<td>b54a3a</td>
<td>App Framework Wrapper</td>
</tr>
<tr>
<td>bukkit</td>
<td>f210234</td>
<td>Minecraft Mod API</td>
</tr>
<tr>
<td>nokgiri</td>
<td>a93c2d6</td>
<td>HTML/XML/CSS parser</td>
</tr>
</tbody>
</table>

Table 1: Evaluation projects (Java). Ordered by popularity.
between nearby points, or minimizing the distance between each point and its image in the 2D space. Classical techniques for dimensionality reduction include principal components analysis (PCA) and multidimensional scaling. We will also employ a more modern method called t-SNE [51].

Figure 3 displays the vectors assigned to a few method names from a typical project (elasticsearch). Each point represents the \( q \) vector of the indicated token. To interpret this, recall that the model uses the \( q \) vectors to predict whether token \( t \) will occur in particular context. Therefore, tokens \( t \) and \( t' \) with similar vectors \( q_t \) and \( q_{t'} \) are tokens that the model expects to occur in similar contexts. These embeddings were generated from the log bilinear context model — that is, without using subtokens — so the model has no information about which tokens are textually similar. Rather, the only information that the model can exploit is the contexts in which the tokens are used. Despite this, we notice that many of the names which are grouped together seem to have similar functions. For example, there is a group of \texttt{assertXXX} methods on the left hand side. Especially striking is the clump of construction methods on the right-hand side \texttt{newDoubleArray}, \texttt{newIntArray}, \texttt{newLongArray}, and so on. It is also telling that near this clump, the names \texttt{grow} and \texttt{resize} are also close together. Analysis reveals that these names do indeed seem to name methods of different classes that seem to have similar functionality. Our previous work [2] indicates that developers often prefer such entities to have consistent names.

Additionally we examine the nearest neighbors of tokens in the \( D \)-dimensional space. This type of analysis avoids the risk, inherent in any dimensionality reduction method, that important information is lost in the projection from \( D \) dimensions to 2D. Table 2 shows some identifiers on a different project, clojure, for each identifier giving a list of other identifiers that are nearest in the continuous space. The nearest neighbors of a token \( t \) are those tokens \( v \) such that the inner product of the embeddings, that is, \( q_t^\top q_v \), is maximized. We choose this measure because it most closely matches the notion of similarity in the model. Again, we are using the log bilinear context model without subtoken information. We again see that the nearest neighbors in the continuous space seem to have similar semantic function such as the triple \texttt{fieldName\_meth\_name}, and \texttt{className} or the names \texttt{returnType}, \texttt{typ}, and \texttt{type}.

Table 3 takes this analysis a bit further for the subtoken model. This table shows the “nearest nearest neighbors”: those pairs of tokens or subtokens that are closest in the embedding space out of all possible pairs of tokens. On the left column, we see pairs of close neighbors from the feature-based bilinear context model without subtokens. These contain many similar pairs, such as \texttt{width} and \texttt{height}. It is striking how many of these pairs contain similar subtokens \textit{even though this model does not contain subtokens}. Moving to the subtoken model, the right column of Table 3 shows pairs of subtokens that are closest in the embedding space. The model learns that pairs like numerals, \texttt{Max} and \texttt{Max}, and \texttt{width} should be placed near to each other in the continuous space. This is further evidence that the model is learning semantic similarities given only statistical relationships between tokens.

We can also attempt to be a bit more specific in our analysis. In this we are inspired by Mikolov et al. [35], who noticed that adding together two of their embeddings of natural language words often yielded a compositional semantics — \textit{e.g.} \texttt{embedding(“Paris”) + embedding(“Vietnam”)} yielded a vector whose nearest neighbor was the embedding of “Hanoi”. To attempt something similar for source code, we consider semantic relationships that \textit{pairs} of identifiers have with each other.

For Figures 4 and 5, we project the \( D \)-dimensional embeddings to 2D using PCA rather than t-SNE. Although a technical point, this is important. Unlike t-SNE, PCA is a linear method, that is, the mapping between the \( D \)-dimensional points and the 2D points is linear. Therefore, if groups of points are separated by a plane in the 2D space, then we know that they are separated by a plane in the higher-dimensional space as well. Figure 4 shows the embeddings of all pairs of setter and getter methods for the project netty. The subtoken model did not generate these models, so the model cannot cluster these tokens based on textual similarity. Nevertheless, we find that getter and setter tokens are reasonably well separated in the continuous space, because they are used in similar contexts. In Figure 5, we match pairs of variable names in the libgdx project in which one name (the “plural name”) equals another name (the “singular name”) plus the character \texttt{a}. The Java convention that \texttt{Collection} objects are often named by plural variable names motivates this choice. Although this mapping is more noisy than the last, we still see that plural names tend to appear on the left side of the figure, and singular names on the right.

From this exploration we conclude that the continuous locations of each name seem to be capturing semantic regularities. Readers who wish to explore further can view the embeddings at \url{http://groups.inf.ed.ac.uk/cup/naturalize/}.

Even though the continuous embeddings are learned from context alone, these visualizations suggest that these embeddings also contain, to some extent, \textit{semantic} information about which identifiers are similar. This suggests that local and global context do provide information that can be represented and exploited, that is, semantically similar names are used in similar contexts. This is evidence pointing towards an affirmative answer to RQ1. It is especially striking that we have consistently found that nearby tokens in the continuous space tend to share subtokens, even when the model does not include subtoken information. The right column of Table 3 reinforces this point since it shows that, when we do use the subtoken model, nearby pairs of subtokens in the continuous space seem to be meaningfully related. This provides some evidence for an affirmative answer to RQ2.

Finally, it can be objected that this type of analysis is necessarily subjective. When backed and validated by quantitative analysis,
5. EVALUATION

In this section, we quantitatively evaluate the performance of the neural model on the data set (Table 1) answering all the RQs.

**Variable Naming**  Renaming variables and method invocations has been previously shown [2] to achieve good performance using n-gram LMs. Figure 6 shows the performance of the baseline n-gram model along with the performance of the other neural models for variable names. For low frequency of suggestions (high confidence decisions), the neural models outperform the n-gram-based suggestions. This is expected since such models perform better than plain n-gram models in NLP [39]. Additionally, the features give a substantial performance increase over the models that lack features.

The subtoken model performs worse compared to the token-level model for suggestion frequencies higher than 6%. This is to be expected, since the subtoken model has to make a sequence of increasingly uncertain decisions, predicting each subtoken sequentially, increasing the possibility of making a mistake at some point. For suggestion frequencies lower than 6% the performance of the subtoken model is slightly better compared to the token-level model, thanks to its ability to generate novel identifiers. Thus, we positively answer RQ1 and RQ2.

We computed Table 4 over only three classes because of the cost of retraining the model one feature at a time. Looking at Table 4 for variable names one may see how each feature affects the performance of the models over the baseline neural model with no features at rank $k = 5^7$. First, we observe that the features help mostly at high suggestion frequencies. This is due to the fact that for high-confidence (low suggestion frequency) decision the models are already good at predicting those names. Additionally, combining all the features yields a performance increase, suggesting that for variable names, only the combination of the features gives sufficiently better information about variable naming.

**Method Declaration Naming Accuracy**  We now attempt to use the neural model for suggesting method names, using only features available during the declaration of a method. Surprisingly, the neural model is exceptionally good at predicting method declaration names. Figure 7a shows the performance of the models on all method declarations excluding any method declarations that are method overrides. We exclude overrides so as to avoid giving the models credit for predicting easy names like toString. When we include overrides, the performance of all models improves. To exclude method overrides, we remove methods that contain the @Override annotation as well as those methods that we can statically determine as method overrides.

The graphs in Figures 7a show that the neural models are substantially better at suggesting method names, compared to the n-gram language model. Adding features increases the performance of the models, indicating that the model is able to use non-local context to make better predictions. Naturally, the performance degrades handles 5 objects and because short term memory is usually $7 \pm 2$ is the size of human short term memory.
Table 4: Absolute increase in performance for each type of feature compared to the normal and subtoken models with no features at 5% suggestion frequency and at 20% suggestion frequency for rank \( k = 5 \). Averages from clojure, elasticsearch and libgdex, chosen uniformly at random from our corpus. If a model does not produce suggestions at a given frequency, it is not counted in the average. The “vocabulary” of an identifier (e.g. method name) are all the subtokens of that identifier.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Absolute F1 Increase (%)</th>
<th>Absolute Accuracy Increase (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple @5%</td>
<td>@20%</td>
</tr>
<tr>
<td>Variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AST Ancestors</td>
<td>-0.3</td>
<td>-1.0</td>
</tr>
<tr>
<td>Method, Class, Superclass and Interface Subtokens</td>
<td>-1.2</td>
<td>-0.1</td>
</tr>
<tr>
<td>Declaration Modifiers</td>
<td>-0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>Variable Type</td>
<td>-0.2</td>
<td>-0.3</td>
</tr>
<tr>
<td>All</td>
<td>1.2</td>
<td>-4.8</td>
</tr>
<tr>
<td>Method Declarations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AST Ancestors</td>
<td>3.6</td>
<td>1.3</td>
</tr>
<tr>
<td>Cyclomatic Complexity</td>
<td>4.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Fields Subtokens</td>
<td>9.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>Class, Superclass and Interfaces Subtokens</td>
<td>11.7</td>
<td>7.8</td>
</tr>
<tr>
<td>Method Implementation Subtokens</td>
<td>7.9</td>
<td>-4.2</td>
</tr>
<tr>
<td>Declaration Modifiers</td>
<td>1.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Return Type</td>
<td>10.0</td>
<td>21.7</td>
</tr>
<tr>
<td>Sibling Methods</td>
<td>8.8</td>
<td>-4.9</td>
</tr>
<tr>
<td>Number of Arguments</td>
<td>7.7</td>
<td>0.4</td>
</tr>
<tr>
<td>Thrown Exceptions</td>
<td>6.4</td>
<td>20.8</td>
</tr>
<tr>
<td>All</td>
<td>17.0</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Figure 6: Evaluation of single point suggestions for variables at rank \( k = 1 \) averaged across all projects. The “features” and “no-features” models lack sufficient confidence to make suggestions at the higher suggestion frequencies.

slightly as the prediction confidence decreases. Interestingly, the token-level models are unable to make any suggestions beyond a suggestion frequency of 15%. For all other tokens, the token-level methods return the special UNk token, indicating that the models expect a neologism which they cannot predict. In contrast, the subtoken models sustain a good F1 score, even for large suggestion frequencies. This is due to the fact that the subtoken models learn naming conventions at the subtoken level, capturing linguistic patterns [5] such as that specific functions may contain various subtokens e.g. `get`, `set`, `has`, `is`.

Table 4 shows a full list of the effect that each feature has on the performance of the neural models at rank \( k = 5 \). As expected, the return type, the subtokens of the class where the method is declared in and the subtokens of the variables and method invocations inside that method provide the most substantial performance increases.

Based on these results, we conclude that we are able to suggest accurate method names and that our suggestions are better than previous approaches. We therefore answer RQ3 in the affirmative.

Class Definitions Accuracy In the previous section, the performance of the neural model on suggesting names for method declarations was shown. In this section, we evaluate the neural model when making suggestions for class definitions. Figure 7b shows the performance of the n-gram language model and the neural models for class name definitions. In contrast to the previous models, the token-level models cannot make any suggestions, always suggesting the UNk token. However, the subtoken model is able to make suggestions even at high suggestion frequencies maintaining an F1 score of more than 40% outperforming the n-gram model.

Thanks to the ability of the subtoken model to suggest neologisms the subtoken-level model is able to suggest class definition names that it has never seen before, with a good F1 score. Table 4 shows that the subtokens of the superclass and interfaces that a type is implementing are informative about the name of the class. Additionally, when combining all the available features, we get a significant increase in F1 score. Thus, we answer yes to RQ4 as well; we are able to suggest accurate type (class) names.

6. RELATED WORK

Naming In Software Engineering Naming in code has achieved a fair amount of research attention. There has been prior research into identifying poorly named artifacts. Host and Østvold [26] developed a technique for automatically inferring naming rules for methods based on the return type, control flow, and parameters of methods. Using these rules they found and reported “naming bugs” by identifying methods whose names contained rule violations. Arnaudova et al. presented a catalog of “linguistic anti-patterns” in code that lead to developers misunderstanding code and built a detector of such anti-patterns [5]. Binkley used part of speech tagging to find field names that violate accepted patterns, e.g. the field `create_mp4` begins with a verb and implies an action which is a common pattern for a method rather than a field [11]. Our work is complementary, as we make suggestions for names when naming bugs are found, anti-patterns occur, or naming rules are violated.

De Lucio et al. attempted to automatically name source code artifacts using LSI and LDA and found that this approach doesn’t work as well as simpler methods such as using words from class and...
method names [18]. Many studies of naming have also been conducted giving us insight into its importance. Butler et al. found that “flawed” identifier names (those that violate naming conventions or do not follow coding practice guidelines) are related to certain types of defects [14]. Later they also examined the most frequent grammatical structures of method names using part of speech tagging [15]. Lawrie et al. [29] and Takang et al. [50] both conducted empirical studies and concluded that the quality of identifier names in code have a profound effect on program comprehension. Liblit et al. explored how names in code “combine together to form larger phrases that convey additional meaning about the code.” [30]. Arnaoudova et al. [6] studied identifier renamings, showing that naming is an important part of software construction. Additionally, in a survey of 94 developers, they found that about 68% of developers think that recommending identifiers would be useful. These studies highlight the importance of our work, by being able to suggest quality names or parts of names.

As method and class names are expected to indicate their semantics, they can be viewed as a special case of code summarization. Haiduc et al. showed that NL text summarization does not work well for code [23] and such techniques must be adapted to be effective. They later developed summaries that are used to improve comprehension [22]. Sridhara et al. used idioms and structure in the code of methods to generate high level abstract summaries. While they don’t suggest method names, they discuss how their approach may be extended to provide them [47]. Sridhara also showed how to generate code summaries appropriate for comments within the code (e.g. as method headers) [46, 45]. For more work in this area, Eddy et al. provide a survey of code summarization methods [20]. We note that most studies and approaches in this area focus on names of variables, fields, and methods. Although some examine all identifiers in the code, we are unaware of any work that focuses on type (class) names as we do.

Language Models In Software Engineering

Probabilistic models of source code have been applied in software engineering. Hindle et al. and Ngyuen et al. [25, 41] used n-gram models to improve code autocompletion. Allamanis and Sutton [3] present an application of code n-gram models at scale. Maddison and Tarlow [31] built a more sophisticated generative model of source code using log-bilinear models that reflects the syntactic structure of the code. Although the machine learning principles we use are similar, their model differs significantly from ours, because their purpose is to build models that generate source code rather than improve existing code. In other words, our model is discriminative rather than generative. Mou et al. [40] use a convolutional neural network to classify code from programming competition problems. Karaivanov et al. [27] combine LMs with static program analysis to suggest method calls and fill-in gaps. Other applications of probabilistic source code models are extracting code idioms [4] and code migration [27]. Closely related to this work is our previous work where we infer formatting and naming conventions [2] using n-gram LMs to suggest natural renamings. Raychev et al. [43] present a discriminative probabilistic model to predict types and names of variables in JavaScript. In contrast, our current work introduces a log-bilinear model that greatly improves on the n-gram LM, especially on method and class naming, proposing neologisms by taking into account subtokens and non-local context.

Other Applications of Neural Log-bilinear Models

Neural log-bilinear models have been used in NLP for LMs [37, 39] and describing images with NL [28]. Log-bilinear models have been shown in NLP to produce semantically consistent and interesting vector space representations (embeddings). Notable systems include word2vec [35, 36] and GloVe [42]. In contrast to these approaches, we use a rich notion of non-local context by incorporating features specific to source code while we produce similar vector space models for method names, variables and types. Additionally, we present a novel sub-token model. Related to our subtoken model is the work of Botha and Blunsom [12] that integrate compositional morphological representations of words into a log-bilinear LM but the morphological features are only used in the context of an LM.

7. CONCLUSION

We introduced the method naming problem, that of automatically determining a functionally descriptive name of a method or class. Previous work on automatically assigning names [2, 43] focuses on local variables, and relies on relatively local context. Naming methods is more difficult because it requires integrating non-local information from the body of the method or class. We presented a first solution using a log-bilinear neural language model, which includes feature functions that capture long-distance context, and a subtoken model that can predict neologisms, names that did not appear in the training set. The model embeds each token into a high dimensional continuous space.

Continuous embeddings of identifiers have many other potential applications in software engineering, such as rejecting commits whose names violate project conventions; exploration of linguistic anti-patterns, such as a getter starting with get [5] and feature localization. Finally, a problem similar to method naming arises in NLP, namely the problem of generating a headline from the text of an article [8, 19]. It is possible that models similar to ours could shed light on that problem as well.

8. ACKNOWLEDGEMENTS

This work was supported by Microsoft Research through its PhD Scholarship Programme. Charles Sutton was supported by the Engineering and Physical Sciences Research Council [grant number EP/K024043/1].
9. REFERENCES


