Context2Name: A Deep Learning-Based Approach to Infer Natural Variable Names from Usage Contexts

Rohan Bavishi and Michael Pradel and Koushik Sen

Technical Report TUD-CS-2017-0296-1
TU Darmstadt, Department of Computer Science
March, 2018
We evaluate Context2Name JavaScript code and show that it successfully predicts 47.5% of all smaller and thus reduce the download time of, e.g., client-side web reusing names in different scopes. The resulting JavaScript files are gluing local names into short, cryptic ones, and by aggressively local identifier names with short, arbitrary, and meaningless names. before shipping their code. This process replaces all minification developer. Code with non-meaningful identifier names is particular functionality, can make a program extremely difficult to read for a formatting, such as indentation and white-spaces, while retaining discourages review and inspection. Combined with the removal of auto-encoder neural network to generate embeddings, which are much smaller in dimensionality, and retain the key features that are sufficient to categorize usage contexts. The RNN used for prediction can be trained much more efficiently with these embeddings.

To train our Context2Name approach, we leverage the huge amount of JavaScript code that is available online. Machine learning techniques exploiting this availability of source code have been used to solve a variety of development tasks, such as code completion [32], fixing syntactic errors [18], code clone detection [40], malware analysis [14], and even generating programs with specific constraints [34]. Deep Learning [22] is a fast-growing field in machine learning that has been very successful in natural language tasks, such as completion, translation and summarization. We show that this effectiveness propagates to the problem of inferring meaningful variable names as well.

We evaluate our technique on a large, publicly available corpus of JavaScript source code, wherein Context2Name is able to recover 47.5% of all minified identifiers exactly, i.e., our tool predicts the same name that the authors of the original programs had in mind. We also show that our approach is practical. It takes an average of 2.9 milliseconds to predict a name, or 110.7 milliseconds to process a file, on average.

A comparison with the state-of-the-art tools JSNice [31] and JNaughty [38] shows that our work achieves comparable accuracy while improving in efficiency. Evaluating all tools on the same
data set, along with a reasonable time limit for processing, Context2Name is able to recover about as many identifiers as JSNice and 8.1% more identifiers than JSNaughty. The main contribution of our work is the conceptually much simpler approach that will be easier to adapt to another programming language than JSNice. The reason is that our work does not perform any program analysis for feature extraction, but instead relies in neural networks to learn which parts of a token stream are relevant for predicting natural names.

In summary, this paper contributes the following:

- A deep learning-based framework to recover natural identifier names from minified JavaScript code.
- A technique for computing vector embeddings of the usage contexts of a variable. The technique makes minimal assumptions about the underlying programming language, and can be adapted to other languages and usage scenarios.
- Empirical evidence that the approach exactly recovers 47.5% of all minified names in a large, publicly available corpus of JavaScript code, which is comparable to existing approaches. The average time for processing each file is also well under a second.

The implementation of our approach, as well as all data to reproduce our results, will be made available.

2 OVERVIEW AND EXAMPLE
This section presents an informal outline of our approach and illustrates the main ideas with an example. The example is a piece of JavaScript code, shown in the upper-left corner of Figure 1. The code has been minified with UglifyJS, a popular minification tool that replaces all local variable names with short, meaningless, and arbitrarily chosen names. All globally visible names, such as http, and property names, such as responseText, are not minified to ensure that the code preserves the semantics of the original code. In practice, UglifyJS also removes unnecessary white-space and indentation, which we preserve here for clarity.

The minification makes it difficult to understand the meaning of individual variables and the overall purpose of the code. The goal of Context2Name is to recover meaningful and natural names that help developers not familiar with the code to reason about it. Ideally, the recovered names are those chosen by the developers in the original code. In general, predicting the exact same names as in the original code is not always possible. The reason is that, even though developers generally tend to write regular and "unsurprising" code [19], developers do not always choose the most natural identifier name, but sometimes choose a semantically equivalent name or a slightly misleading name. To achieve the goal of helping developers understand the code despite minification, it is therefore sufficient to recover some meaningful name, not necessarily the original name, for each local variable.

For the example, the upper-right corner of Figure 1 shows the code with names recovered by Context2Name. Five out of the six local names have been recovered exactly, i.e., as in the original code: req, url, method, async, and body. The remaining name, response, is also very similar in spirit to the name callback that was used by the author of the original code. Overall, the recovered code very much reflects the developer's intention, making it easy for another developer to reason about the code.

The key insight used by our approach to recover names is that the code surrounding different occurrences of a local variable offers a lot of information about the meaning of the variable. For example, the tokens surrounding i in Figure 1 include the keyword function, the global function name http, and the property names open, onreadystatechange, send, and responseText. An average JavaScript developer can infer from this information that i is likely to be an XMLHttpRequest object. A more difficult case is the minified name i. Its surrounding tokens include if, onreadystatechange, and responseText, which at first glance do not reveal much about the purpose of i. A skilled developer who has used XMLHttpRequest may learn from the fact that i is the third argument of open and that onreadystatechange is close to the if condition where i evaluates to true, while responseText is close to another if condition that involves i and a negative. All these features suggest that i toggles the asynchronous behavior of an HTTP requests.

Our approach essentially simulates this semantic reasoning based on prior knowledge of JavaScript code. However, instead of relying on an expert JavaScript developer, we exploit the power of advanced machine learning techniques and the availability of large corpora of code. The approach identifies regularities in the way identifier names occur in real code and uses these regularities to predict meaningful names.

Figure 1 illustrates the different steps taken by Context2Name during this process:

1. At first, a simple token-based analysis of the code extracts usage contexts for every minified name. Specifically, the analysis extracts one usage context for each occurrence of the name. The usage context consists the sequence of length 3 in this case of lexical tokens before and after the minified name.
2. The next step converts the set of usage contexts of a name into a vector representation. This representation is based on a standard one-hot encoding and yields a sparse, binary vector, where most elements are zero and only few elements are one.
3. To ensure the scalability and efficiency of the overall approach, the next step converts the sparse vectors into a dense vector representation. This step is fully automated by using a sequence auto-encoder [13], i.e., an semi-supervised neural network that searches an efficient encoding.
4. The dense vector representations are fed into a supervised machine learning model, a recurrent neural network (RNN). We train this model to predict a ranked list of meaningful names for a given vector representation of the usage contexts of a variable. A key insight is that the prediction depends only on the usage contexts of the variable, and not on the minified name.
5. Finally, Context2Name greedily selects for each minified variable the name predicted with maximum probability and outputs the code with recovered names.

An important conceptual benefit of Context2Name over the state of the art technique JSNice is that our approach makes very
3 APPROACH

In this section, we provide the technical details of our approach. The key idea is to approximate the semantic meaning of a variable or function using a sequence of lexical tokens surrounding its different points of usage in the code. We formally describe this notion of a usage summary in Section 3.1. Sections 3.2 and 3.3 then describe two neural networks: One network to reduce the usage summary into a usage context, and another network to predict a suitable name for the variable. A usage summary is composed of the different contexts in which an identifier is used. To extract the usage summary, we view the source code as a sequence of lexical tokens $T = \langle t_0, t_1, \ldots, t_{\text{len}} \rangle$. We drop some tokens, specifically dotted punctuators and round parentheses, from $T$ as we do not find them to be particularly useful. Let $N = \{ n_1, n_2, \ldots, n_k \} \subseteq T$ be the set of all local names in the code. Because a single name may independently occur in multiple scopes, variables or functions in different scopes having the same name have separate entries in $N$.

For constructing the context of each occurrence of an identifier, we define a helper function. Let $T[k]$ denote the $k_{th}$ token in the sequence $T$. We define a token projection function $g_T$ for $T$ as follows:

$$ g_T(k) = \begin{cases} T[k] & \text{if } 0 \leq k \leq \text{len} \land T[k] \notin N \\ 1 & \text{if } 0 \leq k \leq \text{len} \land T[k] \in N \\ \text{PAD} & \text{if } k < 0 \\ \text{PAD} & \text{if } k > \text{len} \end{cases} \quad (1) $$

The function replaces all local names in the code with a special 1D token. The reason is that these names are minified in the given code and therefore do not contribute any semantic information. Also note that $g_T$ returns the special padding token PAD when its argument is out of range. This case is useful for extracting the context of tokens that appear close to the beginning and end of the analyzed token sequence $T$.

Based on the helper function $g_T$, we now define the context of an occurrence of a local name. For each occurrence of a local name in $T$, we extract the $q$ preceding tokens and the $q$ following tokens...
into a the **context** c(t) of a token t in T:

\[
c(t) = c_{\text{prev}}(k - q), \ldots, c_{\text{prev}}(k - 1), \ldots, c_{\text{prev}}(k + q) > \text{ if } t = T[k]
\]

This local context captures the usage of a name at a particular code location. The hyper-parameter \( q \) can be configured to adjust the number of tokens extracted as context. We use \( q = 5 \) as a default in our evaluation.

Finally, for each local name \( n \), we concatenate the contexts for different usages of \( n \) in the code into a single sequence of tokens, which we call the **usage summary** of the name. We use up to \( l \) contexts per name to construct the usage summary, where \( l \) is another hyper-parameter that can be configured to adjust the size of summaries. If less than \( l \) contexts are available for a particular name, we pad the sequence with the special PAD tokens. If more than \( l \) contexts are available, we use the first \( l \) contexts. We use \( l = 5 \) as a default in our evaluation. We formally define the usage summary as a function \( U(n) \) where \( n \in \mathcal{N} \) as follows (\( \circ \) is the sequence concatenation operator):

\[
U(n) = c(t_1) \circ c(t_2) \circ \cdots \circ c(t_l)
\]

where \( t_i \forall 1 \leq i \leq l \) are occurrences of \( n \)

For our running example from Section 2, the lower left corner of Figure 1 show the usage contexts of variables \( e \) and \( n \) (with parameter \( q = 3 \) for space reasons).

The set of tokens in the usage contexts, along with their position relative to the variable occurrences, captures the syntactic context of the variable usages. The intuition behind **CONTEXT2NAME** is that this context is often sufficient to infer the meaning of the variable.

### 3.2 Learning Embeddings for Usage Summaries

After extracting usage summaries for each local name in the code, the next step is to summarize them into an efficient vector representation called **embedding**. The motivation for this step is twofold. First, to benefit from a machine model that predicts likely variable names, we need to convert the information extracted from source into a format suitable for machine learning. The neural network model we use here, as many other machine learning models, expects vectors of real numbers as its input. Second, the usage summaries are highly redundant, e.g., because the same kind of token occurs many times and because subsequences of tokens occur repeatedly. To ensure the scalability and efficiency of the overall approach, we compress usage summaries into an efficient vector representation.

One option to convert usage summaries into a compact vector representation would be to manually define a set of features and to create vectors that describe the presence or absence of these features. However, coming up with a suitable set of features that capture the semantics of identifier usages in JavaScript would require significant manual effort. Moreover, manually designed features would tie our approach to a specific programming language, and require additional manual effort to adapt it to another language.

Instead of manually defining how to compress the usage summaries, we use an auto-encoder. An auto-encoder is a supervised neural network model that learns to compress a given vector while preserving as much of the original information as possible. We train a sequence auto-encoder [13] that compresses each context in a usage summary into a compact vector, allowing us to represent the usage summary as the concatenation of these compact vectors.

The first step is to define an input vocabulary \( \mathcal{V}_{\text{inp}} \) of tokens that the network recognizes. We construct the input vocabulary by picking the most frequent \( |\mathcal{V}_{\text{inp}}| \) tokens (plus the special PAD token) across our training set of code. In our experiments, \( |\mathcal{V}_{\text{inp}}| = 4,096 \). All tokens that are not frequent enough to occur in \( \mathcal{V}_{\text{inp}} \) are represented by a special UNK token. The input vocabulary is then used to convert the tokens in the usage summary to their one-hot representations, i.e., binary vectors of size \( |\mathcal{V}_{\text{inp}}| \) with the \( i^{\text{th}} \) bit set for the \( i^{\text{th}} \) token in \( \mathcal{V}_{\text{inp}} \). The size of the one-hot representation of the context \( c(t) \) of a token \( t \) is \( s_c = |\mathcal{V}_{\text{inp}}| \cdot q \cdot 2 \) because the context contains \( q \) tokens before and \( q \) tokens after \( t \), each represented by \( |\mathcal{V}_{\text{inp}}| \) bits. For our running example, Figure 1 illustrates the one-hot vector representation of the usage contexts of variables \( e \) and \( n \).

The one-hot representation of contexts is highly redundant and we next describe how to compress it using a LSTM-based sequence auto-encoder. The auto-encoder can be thought of as two jointly learned functions. An encoder function \( \text{enc} : \{0, 1\}^{s_c} \rightarrow \mathbb{R}^E \) which maps the one-hot representation of a context to a real-valued vector of length \( E \), and a decoder function \( \text{dec} : \mathbb{R}^E \rightarrow \{0, 1\}^{s_c} \) that maps the real-valued vector back to a binary vector of the original length. The notation \( \mathbb{R}^E_{[0,1]} \) refers to a vector of length \( E \) of real-valued numbers in the range \([0, 1]\). The goal of training the auto-encoder is to minimize the difference between \( \text{dec}(\text{enc}(c)) = \tilde{c} \) and the original context vector \( c \). The two functions are trained in tandem to minimize this difference. Once trained, we use the intermediate vector returned by \( \text{enc} \) for a given context \( c \) as the embedding for \( c \). That is, the auto-encoder compresses the input vector corresponding to a context into an embeddings of length \( E \), where \( E \) is a hyper-parameter of our approach.

Our implementation of the auto-encoder consist of two jointly trained networks that represent the functions \( \text{enc} \) and \( \text{dec} \). The encoder network consists of a single LSTM layer (a class of recurrent neural network (RNN) models that maintains an internal hidden state and therefore is particularly well-suited to processing
sequences of inputs) with an hidden state of size $E$, which also corresponds to the size of the embedding vectors. The encoder network takes a sequence of $2 \cdot q$ one-hot encoded vectors, denoting a context, and produces an embedding vector of length $E$. The embedding vector is the final hidden state of the LSTM. We use $E = 80$ as a default value in our implementation. The decoder network consists of a layer that repeats the input $2 \cdot q$ times (the number of tokens in the context), and a single LSTM layer with a hidden state of size $|V_{inp}|$. Figure 2 provides an illustration of this network.

In summary, we use the encoder component of this network to convert a usage summary for a name $n \in \mathcal{N}$ into a sequence of embeddings:

$$< \text{enc}(c_1), \text{enc}(c_2), \ldots, \text{enc}(c_l) >$$

where $\mathcal{U}(n) = c_1 \circ c_2 \circ \cdots \circ c_l$ (4)

For the running example, the “Dense vector representation” part of Figure 1 shows the real-valued vector that results from concatenating the embeddings of each context in the usage summary of each variable.

### 3.3 Predicting Names from Usage Summaries

Based on the compactly represented usage summaries, we train a second neural network to predict the name of a variable given its usage summary. The intuition behind this idea is that the way a variable is used in code implicitly conveys sufficient knowledge about its meaning to predict a suitable name. We first define an output vocabulary $\mathcal{V}_{out}$ to choose names from. For our experiments, the vocabulary contains the 60,000 most frequent names encountered in our code corpus.

We then learn a function $P : \mathbb{R}^{E \cdot l} \rightarrow \mathbb{R}^{\mid \mathcal{V}_{out}}$. Given a sequence of embeddings that represent a usage summary as in Equation 4, we first reverse it as suggested in [36], so the embeddings constructed out of PAD tokens come first. We then apply the function which yields a probability distribution over the output vocabulary. This probability distribution can be interpreted as a ranked list of predicted names.

To learn function $P$, we use a recurrent neural network (RNN), i.e., a class of neural models that maintains an internal memory and therefore is particularly well-suited to processing sequences of inputs. The RNN we use consists of a single long short-term memory (LSTM) layer with a hidden state of size $h$ ($h = 3,500$ in our experiments), followed by a softmax layer, which returns a probability distribution. Figure 3 provides an illustration of this network.

The size of the output vocabulary $\mathcal{V}_{out}$ directly corresponds to the range of names our network can output as a prediction. That is, the larger the vocabulary, the higher is the accuracy of name recovery. The trade-off, however, is that the network size increases linearly with $|\mathcal{V}_{out}|$. The vocabulary size we choose for our experiments ($|\mathcal{V}_{out}| = 60,000$) strikes a balance between performance and precision.

In Figure 1, the lower right part shows the names predicted by Context2Name for variables $e$ and $n$. As in this example, several of the top-ranked names may convey the semantics of the minified variable.

### 3.4 Semantics-Preserving Recovery of Names

Our RNN-based predictor outputs a ranked list of possible names for each minified name. The final step is to map each minified name to a single predicted name. This mapping must preserve the semantics of the minified program. Specifically, the same name cannot be mapped as predictions to two different minified variables in the same scope, the predicted name cannot be a keyword, and the predicted name cannot overshadow a name from its parent scope if the name is used in one of its child scopes.

To recover names while respecting these constraints, we compose the ranked predictions for different variables into a single list and then use this list to greedily assign predictions. Algorithm 1 summarizes our approach for recovering names in a consistent and

---

**Algorithm 1 Semantics-Preserving Name Recovery**

1. procedure PredictNames(minifiedCode)
2. recoveredCode ← copy(minifiedCode)
3. minNames ← GetMINIFIEDNAMES(minifiedCode)
4. pQueue ← new PriorityQueue()
5. for all minName ∈ minNames do
6. (pred, prob) ← NextPrediction(minName)
7. pQueue.push((prob, minName, pred))
8. end for
9. while pQueue ≠ ∅ do
10. elem ← pQueue.pop()
11. minName ← elem.minName
12. pred ← elem.pred
13. if NoConflicts(minName, pred) then
14. Replace minName with pred in recoveredCode
15. end if
16. end while
17. return recoveredCode
18. end procedure
We have implemented Context2Name with their corresponding probabilities. To this end, we use a procedure PredictNames takes the minified code as input, makes a copy (line 2), and extracts all the minified names using the GetMinifiedNames procedure (line 3). The algorithm then initializes a priority queue, which will use the probabilities of name predictions, as returned by the RNN, as the key for sorting in descending order. The priority queue essentially tracks the minified names yet to be recovered. Lines 6-9 initialize the priority queue with the top predictions for every name along with their corresponding probabilities. To this end, we use a procedure NextPrediction, which for a given minified name, returns a pair where the first element is the next best prediction after the previous invocation for the same name, and the second element is its corresponding probability.

Lines 11-21 greedily replaces all minified names with predicted names, as provided by the priority queue, until all names have been replaced, i.e., until the queue becomes empty. At each iteration of the while loop on line 11, the algorithm pops the element with the highest probability from the priority queue (lines 12-14). Then, at line 15, a procedure NoConflicts checks whether the algorithm can replace the minified name with the predicted name without creating conflicts. Specifically, we check whether the name predicted has not already been assigned to a different variable in the same scope, is not a keyword, and does not overshadow a replaced name or global name of the parent scope that is referenced in a child scope. If the check passes, the algorithm replaces the minified name with the prediction in recoveredCode. Otherwise, we take the next prediction, and add it to the priority queue. After the loop ends, the algorithm finally returns the recovered code file.

For our running example, the upper right part of Figure 1 shows the code with the names inferred by Context2Name. Even though only five out of the local six names are predicted exactly as in the original code, the code is much more readable than its minified version.

4 EVALUATION

We have implemented Context2Name in Python using Keras\(^1\) as the deep learning framework. The implementation has 397 lines of Python code. We now present an experimental evaluation of our approach to demonstrate its effectiveness and applicability. Specifically, we attempt to answer the following research questions:

**RQ1:** How effective is our approach at predicting natural names for minified variables and function names in real-world JavaScript code?

**RQ2:** How does our approach compare to the current state-of-the-art, specifically JSNice [31] and JSNaughty [38]?

**RQ3:** Is the approach efficient enough to be practical, and does it scale well to large programs?

To answer these questions, we evaluate Context2Name with a large corpus of real-world JavaScript files. The corpus consists of disjoint sets of training files and validation files. For training, we minimize the training files with the popular UglifyJS\(^2\) tool and then train the approach to recover the original names. For validation, we give minified versions of the validation files to Context2Name and then measure the accuracy of recovering the original, unminified names. Although we use UglifyJS in our experiments, our technique is agnostic of the minifier being used, as long as it does not restructure or remove code. Specifically, our techniques works for all minifiers that only modify variable names, and that remove white-space and other terminals, such as semi-colons and commas. It is important to note that this accuracy metric gives a lower-bound of the effectiveness of any technique, as it is possible that the predicted names may be similar, or even better than the original names. A qualitative evaluation involving the manual inspection of suggested names is beyond the scope of this work.

4.1 JavaScript Corpus

We constructed our training and testing corpus using a publicly available data set\(^3\) of JavaScript programs [30]. The data set contains 150,000 non-minified JavaScript files: 100,000 files marked for training and 50,000 files marked for testing. Before our experiments, we cleaned the corpus by the following procedure: We removed 3,150 files that are common between the sets of training and testing files from the training files to ensure that the training and testing data sets are disjoint. We then removed duplicates from these sets, followed by the removal of files that cannot be processed by UglifyJS. Following the setup of the JSNice artifact, we also remove files that are very large (more than 131,072 characters) or that are already (mostly) minified. This is done to ensure that large or minified files do not skew the results. For fairness, we also remove files that cannot be processed by JSNice or JSNaughty due to implementation-level bugs. This entire process reduces the number of training and testing files to 64,750 and 33,229, respectively. A possible alternative to using fixed training and testing sets is k-fold cross validation. We follow the experimental setup of JSNice [31] and JSNaughty [38], which uses fixed training and testing sets.

We present some statistics about the testing corpus. The number of lines of code in the original source files (excluding comment and blank lines) in the validation corpus ranges from 1 to 7,239 (mean 169.2, median 61). The files contain between 0 to 1,388 unique (per-file, per-scope) local variable names (mean 38.4, median 12), between 1 to 1,401 unique variables (local + global) (mean 45.2, median 18), between 0 to 6,139 (mean 160.1, median 40) usages of local variable names, and between 1 and 6,208 (mean 191.5, median 61) usages of all variable names, i.e., both local and global names. Across all files, the number of unique local variable names is 1,277,558. The total number of usages of local variable names is 5,321,106, and the total number of usages of all variable names (global + local) is 6,364,368.

4.2 Training

For training Context2Name, we build usage summaries for all minified names in the minified files in the training set, and then train our model by using the original names provided by the source maps as the ground-truth. Across all training files, the total number of unique local variable names is 2,551,118 (i.e., the size of $N$), which also corresponds to the number of usage summaries used for training. When collecting unique local variable names, we exclude any names not minified by UglifyJS to preserve the semantics of the code, such as references to built-in functions and global variables.

---

1https://keras.io/
2Version 3.1.9, run with the -m (mangler) parameter
3http://www.srl.inf.ethz.ch/js150.php
Table 1: Hyperparameters and values for the evaluation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. $q$ of neighbours used on either side in contexts</td>
<td>5</td>
</tr>
<tr>
<td>No. $l$ of contexts used in usage summaries</td>
<td>5</td>
</tr>
<tr>
<td>Input vocabulary size $</td>
<td>V_{inp}</td>
</tr>
<tr>
<td>Output vocabulary size $</td>
<td>V_{out}</td>
</tr>
<tr>
<td>Embedding size $E$</td>
<td>80</td>
</tr>
<tr>
<td>Hidden layer size $h$</td>
<td>3,500</td>
</tr>
</tbody>
</table>

Table 2: Impact of input vocabulary size $|V_{inp}|$. The bold line is our default setting.

| $|V_{inp}|$ | Percentage of unique tokens covered | Percentage of tokens covered |
|----------|-----------------------------------|-----------------------------|
| 500      | 0.09                              | 93.39                       |
| 1,000    | 0.19                              | 94.53                       |
| 2,000    | 0.38                              | 95.46                       |
| 3,000    | 0.56                              | 95.92                       |
| 4,000    | 0.75                              | 96.23                       |
| 4,096    | 0.77                              | 96.26                       |
| 5,000    | 0.94                              | 96.46                       |
| 6,000    | 1.13                              | 96.64                       |
| 7,000    | 1.32                              | 96.79                       |
| 8,000    | 1.50                              | 96.91                       |
| 9,000    | 1.69                              | 97.02                       |
| 10,000   | 1.88                              | 97.11                       |

Table 3: Impact of output vocabulary size $|V_{out}|$. The bold line is our default setting.

| $|V_{out}|$ | Percentage of unique names covered | Percentage of names covered |
|----------|-----------------------------------|-----------------------------|
| 1,000    | 0.40                              | 63.19                       |
| 5,000    | 1.99                              | 75.07                       |
| 10,000   | 3.97                              | 79.48                       |
| 20,000   | 7.95                              | 83.82                       |
| 30,000   | 11.92                             | 86.38                       |
| 40,000   | 15.89                             | 88.16                       |
| 50,000   | 19.87                             | 89.56                       |
| 60,000   | 23.84                             | 90.74                       |
| 70,000   | 27.81                             | 91.62                       |
| 80,000   | 31.79                             | 92.41                       |
| 90,000   | 35.76                             | 93.19                       |
| 100,000  | 39.74                             | 93.98                       |


derection: Some tokens, such as 1D token in usage summaries. The right column in Table 2 shows the percentage of all tokens, seen across all usage summaries, that are present in the input vocabulary. The entries in Table 2 suggest that the input vocabulary, albeit very small in comparison to the set of all unique tokens, covers a very high fraction of all tokens (96.26% of the tokens from a total of 127,555,900 tokens for $|V_{inp}| = 4,096$).

Table 3 paints a similar picture for the output vocabulary. The middle column compares the relative size of the output vocabulary to the number of unique non-minified names (251,663) seen in our training set, and the right column shows the percentage of all names covered by the output vocabulary. The entries in bold correspond to the size we chose ($|V_{out}| = 60,000$). The conclusion is the same as for the input vocabulary: The output vocabulary is small but sufficiently rich (recognizing 90.74% of the names from a total of 2,551,118 local variables) to perform naming tasks satisfactorily.

4.3 Parameter Selection

The effectiveness and efficiency of CONTEXT2NAME depends on several hyperparameters. The values of these hyperparameters during the evaluation is provided in Table 1. We found these values to strike a balance between the size of the network and the amount of information used to predict names. In particular, the vocabulary sizes $|V_{inp}|$ and $|V_{out}|$ need to be set carefully. Large vocabularies may capture a lot of information and allow many more names to be predicted, but may severely affect performance. We select sizes that cover the majority of the tokens in the training set.

We construct both vocabularies using the training data set only, by picking the $|V_{inp}|$ most frequent tokens across all usage summaries for the input vocabulary, and the $|V_{out}|$ most frequent names for the output vocabulary. A key finding is that a relatively small number of tokens in the input vocabulary accounts for a significant fraction of all tokens across all usage summaries. The reason is that the frequencies of tokens follow a long-tail distribution: Some tokens, such as 1D and semi-colons are extremely frequent, whereas many other tokens, such as application-specific literal values, occur only rarely. The situation is similar for the output vocabulary, where a relatively small set of popular identifier names cover the large majority of all occurrences of identifiers.

Tables 2 and 3 show the impact of different vocabulary sizes. The middle column in Table 2 compares the relative size of the input vocabulary to the number of unique tokens (531,943) seen across all usage summaries in the training set. Note that tokens corresponding to minified names are replaced by the special 1D token in usage summaries. The right column in Table 2 shows the percentage of all tokens, seen across all usage summaries, that are present in the input vocabulary. The entries in Table 2 suggest that the input vocabulary, albeit very small in comparison to the set of all unique tokens, covers a very high fraction of all tokens (96.26% of the tokens from a total of 127,555,900 tokens for $|V_{inp}| = 4,096$).

4.4 Setup for JSNice and JSNaughty

To compare CONTEXT2NAME with the state-of-the-art, we train JSNice and JSNaughty on the same training set as ours, enabling an apples-to-apples comparison. We use the JSNice artifact4, and use the same parameters and arguments as suggested in the accompanying README file. For prediction tasks, we again use the same command line arguments as provided in the README, with an additional parameter to produce a source map from the minified file to the file recovered by JSNice. We then use this map to compute the accuracy.

For JSNaughty, we trained both their translation and language models, as well as the Nice2Predict framework5 [31] on our training set by following instructions provided by the authors. It is important to train a Nice2Predict instance because JSNaughty combines results from both their models and Nice2Predict to make the final prediction.

By default, we run each of three tools with a time limit of 10 minutes for processing a file. If a timeout occurs, we assume that

---

4http://files.srl.inf.ethz.ch/jsniceartifact/index.html
5http://www.nice2predict.org/
the respective tool failed to recover any names in this file. Imposing a time limit is reasonable for de-minification tools for two reasons. First, developers use such tools to save time while inspecting code, and waiting for a tool to finish defeats this purpose. Second, such tools usually have a web-interface (both JSNice and JSNaughty have a web interface) and response times are critical to satisfactorily serve a large number of users. We have not yet released a web interface for Context2Name to maintain anonymity, but will do so once this work is accepted. In practice, the 10-minute time limit affects only JSNaughty. For a full comparison, we also run JSNaughty without any time limit, where it takes over 40 minutes for some files.

4.5 Evaluation Criteria

The evaluation criterion for assessing the effectiveness of the approaches is accuracy, i.e., the ability to recover the original names from minified files. To measure accuracy, we run Context2Name, JSNice, and JSNaughty on each minified file in the validation corpus, and extract a mapping between the minified names and the corresponding predictions made by the tools. We then combine this mapping with the source maps produced by UglifyJS to define a mapping between the original names and the predicted names. Finally, we measure accuracy by computing the percentage of original names that a tool recovers correctly.

There are four variants of the “accuracy” metric, which differ in subtle ways. The metrics differ in two dimensions, as illustrated in Table 4. On the one hand, we can measure accuracy either for all variables and functions in the code, which includes global names, or only consider local names. Since minifying global names in a file may break the semantics of the code, minifiers, such as UglifyJS, do not modify these names. Hence, the task of recovering global names is trivial, as they are not minified at all. Arguably, both definitions of accuracy make sense, and therefore we consider both of them.

On the other hand, we can measure accuracy either per variable or per usage of a variable. For example, if a local variable foo is referenced three times in the same scope, then the per variable metric counts the prediction for foo once, whereas the per usage metric counts the prediction three times. If the variable foo appears twice in different scopes in the same file, then the per variable metric counts the prediction for foo twice. Again, both definitions make some sense, so we here consider both.

Together, these two dimensions yield four accuracy metrics, Local-Once, All-Once, Local-Repeat, All-Repeat, as shown in Table 4. Metric Local-Repeat can be seen as a weighted computation of metric Local-Once, in that correctly predicting a name for a more frequently used variable would give a higher score. Metric All-Repeat (used by JSNice [31]) allows us to directly gauge the similarity between the original, non-minified file and the recovered file. As a trivial baseline, we also compute a baseline accuracy, which gives the percentage of global names only. That is, the baseline accuracy effectively represents the accuracy of a tool that does not predict any names.

4.6 Results

4.6.1 Accuracy. Addressing RQ1 and RQ2, Table 5 provides the four accuracy metrics for all the three techniques. The Local-Once accuracy of Context2Name, i.e., when only unique occurrences of local variable names are considered, is 47.5%, only 0.8% lower than JSNice’s accuracy of 48.3%, and 8.1% higher than JSNaughty’s accuracy of 39.4%. When not imposing any time limit, JSNaughty performs better and reaches an accuracy of 55.3%. However, this increase comes at a high cost on efficiency (Section 4.6.3), which significantly reduces the practicability of the tool. Setting a stricter time limit of only one minute, the Local-Once accuracy of JSNaughty drops even further to only 8.9%. The main reason for JSNaughty’s sharp drop in accuracy when imposing a time limit is that it suffers from an inherent scalability problem, which the authors, on their GitHub page, attribute to the size of phrase-tables used in predicting names. For the experiments reported in the JSNaughty paper [38], the tool was run only on files with 100 lines or less.

4.6.2 Detailed Comparison. We also analyze the relationships between the sets of names recovered by the three approaches. In the Venn diagram shown in Figure 6a, each approach is represented by a colored circle. The percentages inside the colored regions represent the percentage of local minified names for which the original names are recovered correctly (i.e., based on the Local-Once metric). The sizes of the regions also reflect this percentage. It is easy to verify that the sum of percentages inside every circle is equal to the Local-Once accuracies presented in Table 5. Across all files, 5.3% of the total number of unique local variable names are recovered correctly only by Context2Name, 4.88% only by JSNice and 4.14% only by JSNaughty. 13.97% of the names are recovered only by Context2Name and JSNice and 22.42% of the names were recovered by all three tools. Figure 6b shows the results for running JSNaughty without any time limit. Again, each tool recovers some names that the other tools miss. Overall, the two diagrams suggest that all tools are complementary to each other to some extent.
Thus, it may be possible to build a tool that combines all the three approaches to yield accuracies upwards of 60%.

4.6.3 Efficiency. To address RQ3 about the efficiency and scalability of Context2Name, we measure the time needed for predicting names. We performed our experiments on a 32-core machine with four 2.40 GHz Intel Xeon processors running Ubuntu 16.04.1 64-bit, with 256GB RAM. We trained and hosted our model on a separate machine with one 4.20 GHz Intel i7 processor, with one Nvidia 1080Ti GPU, running Ubuntu 16.04.1 64-bit with 48 GB RAM. The embedding and prediction networks are trained separately, both for 5 epochs, and the time to train them was 1.5 days and 3 days, respectively, which is a one-time effort. For prediction tasks, we developed a client running on the first machine that queries the model hosted on the second. Our timing therefore is the sum of processing time on the first machine, and the time taken to query the model loaded on the GPU.

Table 4 shows the per-file timing statistics. Both Context2Name and JSNice perform comparably and are able to process most files in under half a second, making the techniques good candidates for online interactive tools. JSNaughty takes an average of 65 seconds. Thus, the improvement in accuracy of JSNaughty comes at a very significant penalty in efficiency, making it less suitable to be deployed in an interactive setting.

Overall, the results show that Context2Name performs comparably to the existing techniques with respect to accuracy, and that it outperforms the existing techniques, in particular JSNaughty, with respect to running time. Moreover, Context2Name complements existing techniques by predicting 5.3% of all names that are missed by both existing tools. These results are particularly remarkable given that Context2Name does not rely on any kind of manual and language-specific feature engineering, making it straightforward to apply the approach to another programming language.

5 RELATED WORK

5.1 Obfuscation

Program obfuscation has applications in protection of intellectual property [15], resistance against reverse-engineering and software tampering [6, 21, 39] as well as watermarking [33]. An obfuscator is basically a compiler, that takes an input program, and outputs a semantically equivalent program that is mostly unintelligible to a third party without access to the source. Barak et al. [8] showed that obfuscation in general is not realizable. Nonetheless a large body of work has been published on new techniques for obfuscation [12, 23], as it has proved to be practically useful.

Obfuscation is an attractive option in the domain of JavaScript programs as the code is shipped as source, allowing anyone to view and download the original code of the author. But excessive obfuscation can be detrimental to performance w.r.t. bandwidth usage and execution time. Obfuscators that increase the size of the program, or make it significantly slower are not particularly useful. Compatibility is also an issue, as many JavaScript programs rely on external APIs whose usage has to appear in the clear. Minifiers, such as UglifyJS, are an excellent compromise, as the resultant programs are much smaller, and variable renaming is a sufficient deterrence for most adversaries.

5.2 Deobfuscation

Deobfuscation techniques attempt to uncover various aspects of the semantics of the program, which has applications in reverse engineering and malware analysis. Most of the proposed techniques rely on static and dynamic analyses [11, 26, 37], which are more
adept at uncovering semantic information about the obfuscated program. Instead, our work focuses on producing friendlier summarizations in the form of variable names, which is useful in the domain of JavaScript programs. Statistical techniques that relate the semantics to the names used in the program are expected to perform better at this task.

5.3 Probabilistic Models for Code

Machine learning has been used in various domains such as natural language, computer vision with tremendous success. These successes have created a lot of interest in applying machine learning to programs, targeting a variety of development tasks such as code completion [10, 29, 32], fixing syntactic errors [9, 18], generating inputs for fuzz testing of programs [17, 24, 28], and generating programs or short code snippets to assist developers or test compilers [4, 34]. Deep learning techniques in particular have been used for code synthesis [7], detecting clones of code fragments [40], and malware detection and signature generation [14]. The core idea behind all these techniques is to exploit the abundance of available source code by mining useful patterns and correlating them to the value of the desired property. In our case, we try to correlate a name to a context comprised of its surrounding tokens, which may include identifiers that have not been minified (global names).

The two closest existing techniques are JSNice [31] and JSNaughty [38], and we show in Section 4 that our approach outperforms both. Conceptually, our work differs from JSNice by not relying on any program analysis to extract features that may be relevant for predicting variable names, but to instead let neural networks decide which parts of the syntactic context matter. A practical benefit of this design decision is that Context2Name is less language-dependent and can be adapted to another language easily. Compared to JSNaughty, an important difference concerns the efficiency and scalability to larger files. As discussed in Section 4.6.1, JSNaughty takes very long for files with only a few hundreds of lines code, which is why the evaluation of JSNaughty in [38] only considers files with up to 100 lines.

Statistical techniques which also use identifiers have been used to tackle a number of problems. JSNice [31], in addition to predicting names, also predicts types for variables in JavaScript. We believe our approach of using sequences of neighbouring lexical tokens can be extended to type prediction as well. Another approach [3] tries to predict concise yet descriptive names that summarize code snippets, which can then effectively serve as method names. Naturalize [1] is a framework that enforces a consistent naming style by providing better names. We expect our number of incorrect predictions to go down after using Naturalize on our testing corpus, as the number of “surprising” and inconsistent names would decrease.

Sequences of lexical tokens are also used by the SmartPaste framework [2] which is designed to assist a developer in extending an existing code base with snippets, by attempting to align the variables in the added snippet with the rest of the code. SmartPaste uses the lexical tokens around a variables in the snippet to compute the probability of it being replaced by some variable in the rest of the code. In addition to using lexically preceding tokens, they also use data flow relations to compute the relevant tokens surrounding the use of a variable. We expect to benefit from using data flow analysis as well.

Embeddings are popular in machine learning-based natural language processing, where they abstract words into vectors [25]. Beyond pure natural language texts, embeddings for code elements have been proposed, e.g., for API elements [27] and for terms that appear both in code and in natural language documentation [41]. Our embeddings differ from these approaches by serving a different purpose, namely to compress usage contexts while preserving their important features, instead of reasoning about individual program elements.

6 CONCLUSION

This paper addresses the problem of recovering meaningful identifier names in code that has been obfuscated by replacing all local names with short, arbitrary, and meaningless names. We present a deep learning-based approach that exploits and learns from the large amount of available code. The key idea is to predict from the syntactic usage context of a variable what meaning the variable has and to then assign a suitable name. To this end, we combine a lightweight static analysis, an auto-encoder neural network, and a recurrent neural network into a fully automated tool. An evaluation on a set of 150,000 JavaScript files shows that Context2Name clearly outperforms the two-state-of-the-art tools JSNice and JSNaughty. Our approach predicts 17% and 43% more names, respectively, than the existing tools, while taking only 2.9 milliseconds per predicted name, on average. Our work contributes a practical tool to help developers understand minified code and shows the power of deep learning to reason about code. Because the approach makes few assumptions about the analyzed programming language and uses a very lightweight, token-based program analysis, it will be easily applicable to other languages.

ACKNOWLEDGMENTS

This work was supported in part by NSF grants CCF-1409872 and CCF-1423645, by a gift from Fujitsu Laboratories of America, Inc., by the German Federal Ministry of Education and Research and by the Hessian Ministry of Science and the Arts within CRISP, by the German Research Foundation within the ConCys and Perf4JS projects, and by the Hessian LOEWE initiative within the Software-Factory 4.0 project.

REFERENCES
