ORIGINAL ARTICLE

Automatic Code Summarization Using Abbreviation Expansion and Subword Segmentation

Yu-Guo Liang[1] | **Gui-Sheng Fan**[1] | **Hui-Qun Yu**[1] | **Ming-Chen Li**[1] | **Zi-Jie Huang**[2,1]

[1]School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China.

[2]Shanghai Key Laboratory of Computer Software Testing and Evaluating, Shanghai Development Center of Computer Software Technology, Shanghai, China.

Correspondence
Gui-Sheng Fan, School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China.
Email:gsfan@ecust.edu.cn
Hui-Qun Yu, School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China.
Email: yhq@ecust.edu.cn

Funding information
This work was partially supported by the National Natural Science Foundation of China (No. 62372174), the Computational Biology Program of Shanghai Science and Technology Commission (No. 23JS1400600), the Capacity Building Project of Local Universities Science and Technology Commission of Shanghai Municipality (No. 22010504100), the Research Programme of National Engineering Laboratory for Big Data Distribution and Exchange Technologies (No. 2021-GYHLW-01007), and the Shanghai 2024 Science and Technology Innovation Action Plan Star Cultivation (Sailing Program, No. 24YF2719900 and 24YF2720000).

Abstract Automatic code summarization refers to generating concise natural language descriptions for code snippets. It is vital for improving the efficiency of program understanding among software developers and maintainers. Despite the impressive strides made by deep learning-based methods, limitations still exist in their ability to understand and model semantic information due to the unique nature of programming languages. We propose two methods to boost code summarization models: context-based abbreviation expansion and unigram language model-based subword segmentation. We use heuristics to expand abbreviations within identifiers, reducing semantic ambiguity and improving the language alignment of code summarization models. Furthermore, we leverage subword segmentation to tokenize code into finer subword sequences, providing more semantic information during training and inference, thereby enhancing program understanding. These methods are model-agnostic and can be readily integrated into existing automatic code summarization approaches. Experiments conducted on two widely used Java code summarization datasets demonstrated the effectiveness of our approach. Specifically, by fusing original and modified code representations into the Transformer model, our Semantic Enhanced Transformer for Code Summarization (SETCS) serves as a robust semantic-level baseline. By simply modifying the datasets, our methods achieved performance improvements of up to 7.3%, 10.0%, 6.7%, and 3.2% for representative code summarization models in terms of *BLEU-4*, *METEOR*, *ROUGE-L* and *SIDE*, respectively.

KEYWORDS
Automatic Code Summarization, Code Abbreviation Expansion, Subword Segmentation, Program Understanding, Deep Learning

1 | INTRODUCTION

Program understanding is essential to software development and maintenance [1]. The presence of high-quality natural language descriptions for code can significantly enhance the readability and understandability of the program, thereby boosting the work efficiency of software development and maintenance personnel [2]. Automatic code summarization, as a task of automatically generating corresponding functional descriptions for code, is currently a hot research topic in the field of program understanding [3, 4].

As advances in deep learning techniques and the enrichment of open-sourced code summarization corpora, data-driven deep learning methods have significantly improved the efficiency and quality of auto-generated summaries. [5] pioneered the integration of deep neural networks in automatic code summarization, employing the sequence-to-sequence (Seq2Seq) model within the end-to-end NMT framework to generate code summaries. Since the Transformer [6] emerged in recent years has advantages in representing long sequences, researchers have continuously proposed advanced code summarization frameworks based on this prevailing model. Most deep learning based automatic code summarization approaches draw inspiration from NMT solutions in NLP, and concentrate on exploring the relationship between code-related semantic as well as structural information and natural language descriptions [4].

Pre-trained code models, which build upon the architectures of existing deep learning models, are initially trained on extensive multi-language datasets and subsequently fine-tuned on smaller, task-specific datasets. These models leverage elaborated pre-training tasks to obtain universal code representation suitable for multiple programming languages. This makes them versatile for various downstream software engineering tasks, including automatic code summarization. Similarly, these models borrow key concepts from pre-trained language models in the NLP field, with a primary focus on designing innovative pretraining tasks that accommodate the unique characteristics of code [7].

TABLE 1 A code snippet containing abbreviations and identifiers that does not comply with naming conventions.

Function ID	36110318

Code	`public void load(URL u){`
	`  FileCacheSeekableStream s = new FileCacheSeekableStream(`
	`  u.openStream());`
	`  load(s);`
	`  imgname = u.toString();`
	`  ......`
	`}`

Summary	Loads the image from a URL .

Although deep learning based automatic code summarization approaches have achieved impressive results, we discover that existing code summarization models are still facing difficulties in understanding and modeling complex information contained in code. For instance, Table 1 presents a Java code snippet (part of the code is truncated for the sake of brevity) and the corresponding summary description in the Funcom dataset [8], where information of the abbreviated formal parameter 'u' is reflected in the summary. Since Java is a strongly typed language, the type 'URL' of the formal parameter in this example may aid models in generating an accurate summary to some extent. However, basic data types like 'int' and 'char' in other code snippets can offer limited information, making it challenging for these models. This necessitates the conversion of abbreviations nested in source code, particularly in identifiers, into

corresponding full terms, which is the goal of the code abbreviation expansion task. Code abbreviation expansion is able to enhance both the understandability of source code and the accuracy of natural language analysis techniques [9]. Ideally, the uncertainty of abbreviations' semantic information can be eliminated by means of code abbreviation expansion, which not only helps code summarization models better understand codes but enables them to focus on critical identifiers themselves rather than their types, fostering better text alignment between programming and natural language. Exploratory experiments suggest that an increase in code abbreviations deteriorates the performance of a code summarization model. Therefore, this paper's primary objective is to investigate whether code abbreviation expansion is capable of improving the performance of code summarization models.

Moreover, the out-of-vocabulary (OOV) issue is another challenge in automatic code summarization [10, 11]. This problem usually arises when the model encounters identifiers that it has not seen during training, therefore, they are not included in its vocabulary. To mitigate this issue, current code summarization approaches split code and summary sequences into individual words using predefined split functions based on the CamelCase and snake_case naming conventions [8, 12, 13]. For example, if the 'imgname' identifier included in the code snippet appears infrequently across the dataset, it may not be included in the model's vocabulary. In such cases, during both model training and inference stages, the identifier would be replaced by a special symbol (usually denoted as <unk>), representing an unknown word. This replacement leads to the loss of critical information because the model cannot learn the semantic meaning of it. However, even if the identifier is frequent enough to be included in the vocabulary, it can still be challenging for code summarization models to understand its actual meaning and generate an accurate summary. This is because the traditional naming convention-based split functions can not split 'imgname' into the more meaningful tokens 'img' and 'name'. As a result, the model might struggle to generate the corresponding summary 'image'.

Although subword segmentation methods, initially developed for NMT, have effectively addressed the OOV problem and have been widely adopted in pre-trained language models, these methods have yet to be considered in automatic code summarization approaches. Existing pre-trained code models have directly utilized subword algorithms from referenced pre-trained language models, without making necessary adjustments to accommodate the unique characteristics of code [7]. As a result, their usefulness in addressing the aforementioned challenges is limited. Consequently, the second aim of this paper is to explore how to effectively employ subword segmentation algorithms to tokenize words that traditional functions fail to split, and to validate their effectiveness in code summarization models. The main contributions of our work include:

- We propose the use of code abbreviation expansion to weaken the negative impact of abbreviations on program understanding and strengthen the language alignment ability of code summarization models. A series of context-based heuristic algorithms are adopted to expand abbreviations nested in code snippets of Java code summarization datasets.
- We introduce the unigram subword segmentation algorithm to expose more semantic information and further enhance the program understanding performance of code summarization models. Code-specific tokenizers are developed to tokenize code-summary pairs into more granular and semantically preserved subword sequences.
- We present a framework Semantic Enhanced Transformer for Code Summarization (SETCS) to better leverage the semantic information introduced by above methods. A robust baseline is designed by fusing embeddings of original and newly generated subtoken sequences, allowing for effective capture of critical information.
- To the best of our knowledge, this is the first work that incorporates code abbreviation expansion and subword segmentation into the automatic code summarization task. These methods are model-agnostic and can be easily integrated into existing automatic code summarization approaches. Experiments conducted on two widely evaluated datasets demonstrate the effectiveness of our proposed methods.

The remainder of this paper is structured as follows. Section 2 summarizes related work. Section 3 details our proposed methods. The experimental setup and results are explained and analyzed in Section 4 and Section 5, respectively. Following that, some threats to validity are presented in Section **??**. Finally, we conclude the paper and discuss potential avenues for future research in Section 6.

2 | RELATED WORK

2.1 | Automatic Code Summarization

Automatic code summarization approaches focus on leveraging code-related information to generate high-quality summary descriptions. Based on the type of information leveraged, existing research can be divided into two categories.

Structure-Driven Code Summarization Models. Hu et al. [14] first proposed a method of using the abstract syntax tree (AST) representation of source code to improve the performance of the code summarization model. Subsequent works tried to adopt, optimize AST, or introduce more advanced structural information, such as combined usage of AST and serialized code [8, 12, 15, 16], fine-grained split ASTs [17, 18], and utilization of code property graph [19], multi-view graph [20], dataflow graph [21], as well as heterogeneous code graph [22].

Semantic-Driven Code Summarization Models. TL-CodeSum [23] and API2Com [24] demonstrated the effectiveness of application programming interface (API) information for code summarization. DMACOS [25] exploited the deliberation network and adopted method name prediction as an auxiliary training task to improve the quality of generated summaries. Li et al. [26] utilized multi-task joint learning to incorporate action word prediction into code summarization models. Both Rencos [27] and Re2Com [28] combined traditional information retrieval techniques with deep neural networks to exploit the information contained in retrieved similar code snippets or corresponding summaries. MLCS [29] and MPCos [30] designed meta-learning frameworks for the automatic code summarization task in different scenarios, among which the key idea is to use similar code samples to obtain specific summary generators optimized for each target code snippet.

Existing pre-trained code models can also be classified into the above two categories according to different types of model input in the pre-training stage. For example, in addition to source code, GraphCodeBERT [31] and SPT-Code [32] took control flow graph and AST as additional code-related structural input respectively, while CodeBERT [33], CodeT5 [34] and PLBART [35] took code-related semantic information such as summaries and posts as additional model inputs.

Both code abbreviation expansion and subword segmentation methods introduced in this paper fall into the second category, as the former method utilizes related identifiers to expand abbreviations nested in the source code and the latter method assists in code summarization models by exposing more semantic information included in the code snippet.

2.2 | Code Abbreviation Expansion

Due to the limitations of abbreviation dictionaries and general English dictionaries, more advanced approaches for code abbreviation expansion focus on contextual information of abbreviations, including comments, methods, classes, and projects. In addition, most researchers generally adopt certain predefined matching rules to find potential expansions by identifying different types of abbreviations. According to our survey, a series of works made by Jiang et al. [36, 37, 38, 39] in recent years have significantly improved the recall and precision scores of the code abbreviation

expansion task in multiple open-source applications.

Literature [36] used the semantic relationships between software entities and construct knowledge graphs for entities, semantically related entities, and their relationships to obtain full terms of abbreviations in software entities. Literature [37] designed a series of heuristic methods utilizing specific fine-grained context to expand the abbreviations in both formal and actual parameters. In response to the question of whether target abbreviations should be replaced with the corresponding full names, literature [38] proposed an automatic decision-making tool for abbreviation expansion. On the basis of [36], literature [39] further proposed an automatic identifier abbreviation expansion method that leverages the semantic relationship between software entities and migration expansion within the same application.

To expand abbreviations nested source codes of code summarization datasets, we re-implement and refine three heuristic algorithms so that abbreviations in identifiers such as parameters and variable names can be expanded as much as possible. These algorithms have been proved to be highly precise when tested across a range of well known open-source projects [37].

2.3 | Subword Segmentation

Byte pair encoding (BPE) [40] is a data compression technology and the original idea is to iteratively replace the most frequently occurred byte pairs in a sequence with a single, unused byte. It was later adopted by Sennrich et al. [41] to solve the OOV problem in the NMT task and became the dominant method for subword segmentation. By continuously merging frequently occurred character pairs or sequences, BPE can retain the most frequently occurred subwords in the process of segmenting rare words. It is worth to note that both CodeBERT and CodeT5 adopt the tokenizer of Roberta [42], which is a pretrained language model that utilizes this algorithm.

Similarly, the WordPiece algorithm [43] also starts from a small vocabulary and continuously learns the merging rules during the training of a tokenizer. The difference is that WordPiece prioritizes character pairs with lower frequencies in each part of the vocabulary, and it does not use merging rules learned in the training stage but looks for the longest subword from the vocabulary for segmentation in the tokenization stage.

Contrary to the above two methods, the Unigram algorithm [44] continuously removes unnecessary words from a large vocabulary until the desired vocabulary size is reached. In addition, both BPE and WordPiece segment sentences or words into unique subword sequences, while Unigram is capable to produce multiple subword segmentation results based on probability.

To ensure the selection of the most suitable result from tokenized subword candidates, we employ the Unigram algorithm to train code-specific tokenizers for each code summarization dataset, aiming to preserve the original semantic information of the data samples to the greatest extent possible.

3 | METHODS

Figure 1 shows the flowchart of our approach. Initially, we extract code snippets and corresponding summaries from source code files. Subsequently, these codes are parsed into Abstract Syntax Trees (ASTs), enabling the extraction of key information to assist in expanding nested abbreviations within the code. Following this, we utilize the subword segmentation algorithm to train a tokenizer based on words in the new corpus, which comprises sequences of expanded codes and original summaries. Ultimately, tokenized codes and corresponding summaries are used to train a code summarization model.

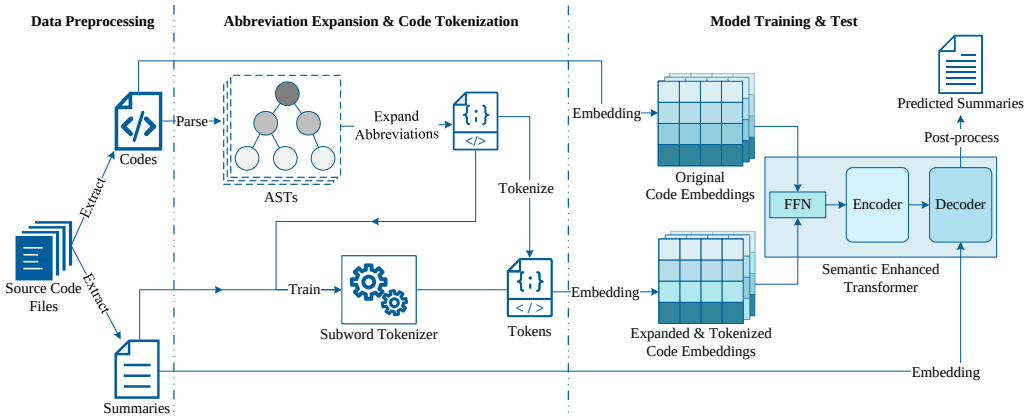


FIGURE 1 Flowchart of our approach.

The method of fusing the embeddings of both source and modified code using a Feature Fusion Network (FFN) is not strictly necessary, as the expanded and tokenized code can be directly used to train a new code summarization model. However, the technique of feature fusion is significant and has been employed in many automatic code summarization approaches. To better leverage the critical semantic information introduced by methods proposed in this paper, we further present a new encoder-decoder-based model, namely the Semantic Enhanced Transformer for Code Summarization (SETCS).

3.1 | Context-based Code Abbreviation Expansion

Figure 2 illustrates the AST corresponding to the code snippet shown in Table 1, while only part of the key attributes and values are displayed for brevity. Non-terminal nodes in the AST represent various attributes, such as parameters, name, and body of the method declaration. Terminal nodes represent values of related attributes, such as identifiers and keywords contained in the code snippet. In the process of parsing source codes into ASTs, four sets of auxiliary information for each code snippet are extracted and stored:

1) Method ID, project ID, method name, called methods, and passed actual parameters.

2) Formal parameters as well as their types, split parameters, and involved abbreviations.

3) Parameters and their types within the method, split parameters, and involved abbreviations.

4) Variables and their types within the method, split variables, and involved abbreviation.

The method name, actual parameters passed in the called methods, and types of formal parameters are used as reference words for expanding abbreviations involved in split formal parameters. Types of parameters and variables are used as the reference words to expand abbreviations involved in split parameters and variables, respectively. The method ID and project ID are used to locate specific methods in the project when expanding abbreviations. For example, in the illustrated AST, the method name of 'load' (extracted name of the method declaration) and formal parameter's type of 'URL' (extracted reference type of the formal parameter) will be used to expand the abbreviation 'u' in the formal parameter; the variable's type of 'FileCacheSeekableStream' (extracted reference type of the local variable declaration) will be utilized to expand abbreviation of 's' in the variable name.

Note that before identifying abbreviations, corresponding identifiers are split using a traditional predefined split function, which splits identifiers based on naming conventions and converts all split words to lowercase. For example,

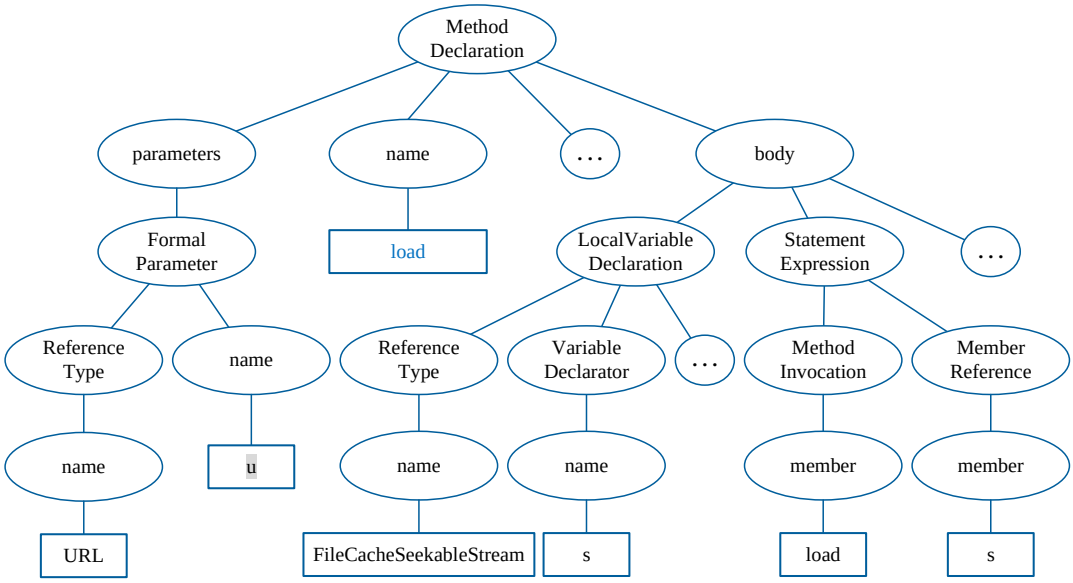


FIGURE 2 Illustration of AST.

either 'fileName' or 'file_name' would be split into 'file' and 'name'. In addition, all abbreviation expansion algorithms utilize the function to split reference words. Code abbreviation expansion algorithms are shown as follows:

Algorithm 1: Acronym Expansion

Input: abbreviation, reference word

Output: expansion candidate

```
1 words ← split(reference)
2 initialCharacters = ''
3 for each word in words do
4     initialCharacters = initialCharacters + word[0]
5 if abbreviation equals initialCharacters then
6     expansion = initialCharacters
7 return expansion
```

For longer identifiers that are composed of multiple words, developers often select the initial characters of each word as an abbreviation during programming, and such an abbreviation form is termed as acronym. For example, the identifier of 'timePerFrame' may be abbreviated as 'tpf'. When expanding such kind of abbreviations, the initial characters of each split word are extracted (lines 1-4) and used to compare with the abbreviation, if the abbreviation and combinations of initial characters are the same, the split word will be considered as the expansion candidate of the abbreviation (lines 5-6). It should be noted that the abbreviation may be equivalent to the initial characters of part split words, such as the situation of abbreviating 'setKeystoreFilename' as 'kf', instead of 'skf'. To leverage method names to expand abbreviations present in formal parameters, this case is also considered during the implementation of the algorithm.

Algorithm 2: Prefix Abbreviation Expansion

```
Input: abbreviation, reference word
Output: expansion candidates
1  words ← split(reference)
2  candidates = []
3  for each word in words do
4      if word starts with and not equals abbreviation then
5          candidates ← candidates ∪ word
6  if len(candidates) > 0 then
7      expansion ← candidates[0]
8      for each candidate in candidates do
9          if len(candidate) < len(expansion) then
10             expansion = candidate
11 return expansion
```

Prefix abbreviations are commonly found in identifier definition statements, among which 'String str' is the most typical example. The idea of expanding these abbreviations is to find split words that begin with the abbreviation but are not exactly equivalent to it in the process of splitting the reference word and add them to the set of expansion candidates (lines 1-5). Since basic forms of words are usually short, the shortest one is selected as the final expansion candidate of the abbreviation if multiple candidate expansions are obtained (lines 6-10).

Algorithm 3: Dropped-letters Abbreviation Expansion

```
Input: abbreviation, reference word
Output: expansion candidates
1  words ← split(reference)
2  candidates = []
3  for each word in words do
4      i ← 0
5      for j in range(len(word)) do
6          if abbreviation[i] equals word[j] then
7              i ← i + 1
8              if i equals len(abbreviation) then
9                  candidates ← candidates ∪ word
10                 break
11 if len(candidates) > 0 then
12     expansion ← candidates[0]
13     for each candidate in candidates do
14         if len(candidate) < len(expansion) then
15             expansion = candidate
16 return expansion
```

The term of 'idx' is a common dropped letters abbreviation, and 'index' is usually its full name. In the process

of splitting the reference word, every split word and each character in the abbreviation are compared sequentially, and if a split word (lines 8-9) contains all the characters of the abbreviation, it is appended to the list of expansion candidates. Then the next split word and each character in the abbreviation are compared again until all split words are traversed. The code logic of lines 11-15 is the same as lines 6-10 in algorithm 2, where the shortest word in the list of expansion candidates is finally selected, while the purpose of which is to avoid introducing extraneous long words that contain abbreviations. Considering that this algorithm is prone to generate erroneous expansion results for single-letter abbreviations, in practice, the length of input abbreviations is limited to more than 1.

In summary, context information such as parameter types, method names, and actual parameters passed into called methods are utilized as reference words for formal parameter abbreviations in a specific method. Subsequently, the most frequent expansion candidate obtained by the three expansion algorithms is selected as the final choice. For abbreviations contained in parameters and variables within the method body, expansion candidates obtained by acronym and prefix abbreviation expansion algorithms are favored based on their types. While the overall approaches of the three abbreviation expansion algorithms described above are generally consistent with that of [37], the main distinction arises from the original study's focus on expanding abbreviations in parameters and evaluation on 9 open-source projects, compared to our need to expand abbreviations nested in both parameters and variables within datasets containing approximately 4.7k and 0.5M projects, respectively. Consequently, in our implementation, we encounter more specific scenarios, such as the discovery of 'setKeystoreFilename' during expanding acronyms, and address these issues to balance precision and recall as effectively as possible. More detailed information can be found in our open-source code.

3.2 | Unigram-based Subword Segmentation

As shown in Figure 1, after obtaining the expanded code snippets, the new corpus's word collection obtained by the traditional split method is deemed as the initial vocabulary; then a code-specific tokenizer is trained by leveraging the unigram subword segmentation algorithm, which based on the unigram language model; finally, the tokenizer is utilized to tokenize all code-summary pairs into more fine-grained subword sequences before they are fed into the code summarization model.

In the context of automatic code summarization, the unigram subword segmentation algorithm aims to segment code sequences and their corresponding summary sequences into subword units, considering subword-level probabilities. The algorithm follows the steps outlined below:

For a pair of code sequence C and summary sequence S in the new corpus D, let $c = (c_1, \ldots, c_x)$ and $s = (s_1, \ldots, s_y)$ correspond to subword sequences for C and S, respectively. The unigram language model assumes that each subword appears independently, so the occurrence probability of a subwords sequence $c = (c_1, \ldots, c_x)$ can be formalized as product of each subword's occurrence probability:

$$\begin{aligned} P(c) &= \prod_{i=1}^{x} p(c_i) \\ \forall_i c_i \in \mathcal{V}, &\sum_{i=1}^{|\mathcal{V}|} p(c_i) = 1 \end{aligned} \tag{1}$$

where V is the pre-determined initial vocabulary. Let $T(C)$ represent the set of segmentation candidates for C, then the most likely segmentation sequence can be formulated as:

$$c^* = \underset{c \in \mathcal{T}(C)}{\operatorname{argmax}} P(c) \tag{2}$$

After that, the expectation maximization (EM) algorithm is used to maximize the following marginal likelihood $\mathcal{L}$, and estimate the occurrence probability of subwords in the form of hidden variables $P(c_i)$.

$$\mathcal{L} = \sum_{j=1}^{|\mathcal{D}|} \log\left(P\left(C^{(j)}\right)\right) = \sum_{j=1}^{|\mathcal{D}|} \log\left(\sum_{c \in \mathcal{T}(c^j)} P(c)\right) \tag{3}$$

where $\mathcal{D} = \left\{\left\langle C^{(j)}, S^{(j)}\right\rangle\right\}_{j=1}^{|\mathcal{D}|} = \left\{\left\langle c^{(j)}, s^{(j)}\right\rangle\right\}_{j=1}^{|\mathcal{D}|}$ represents the new code-summary corpus, and $|D|$ is the size of the corpus.

Finally, following steps are iterated over until the desired vocabulary size $|V|$ is reached:

1) Maintain a fixed vocabulary and use the EM algorithm to optimize $P(c)$.

2) Calculate the loss ℓ_i for each subword c_i, where ℓ_i represents the change in the loss value of $\mathcal{L}$ when ℓ_i is removed from the current vocabulary.

3) Sort all subwords according to ℓ_i and retain the top n% of subwords.

Note that high-frequency basic words, including single characters and keywords in the programming language, should always be kept in the vocabulary to prevent issues of OOV and over-fine-grained tokenization, so that critical semantic information in initial sequences can be preserved as much as possible. Finally, a vocabulary that contains subword tokens and their corresponding occurrence probabilities is obtained, and the trained tokenizer utilizes Equation (2) to generate the most likely subword sequences c^* and s^* for each pair of C and S based on the final vocabulary.

In practice, the tokenizer is used to tokenize each word in the target sequence sequentially. If a word can be represented by a combination of multiple tokens in the vocabulary, it will be tokenized based on (1) whether the tokenized subword is included in the pre-split word set of code and the sequence in the same method, which gives subword candidates occurring in somewhere of the same method higher priority; (2) the number of tokens after tokenization, which means shorter subword candidates would be a priority. Eventually, the semantically preserved and/or shortest tokenization result from the Top-k subword combination candidates will be selected.

By leveraging the vocabulary that includes characters, common subwords, and words, rare words in almost all codes and summaries can be properly tokenized. Most importantly, the fine-grained and semantically preserved subword representation exposes more meaningful information, which is expected to further improve the performance of the code summarization model.

3.3 | Semantic Enhanced Transformer for Code Summarization

Figure 3 shows the framework of Semantic Enhanced Transformer for Code Summarization (SETCS). Similar to most code summarization models, SETCS utilizes the encoder-decoder framework, and adopts the Transformer model as backbone. Both encoder and decoder of the model are stacked with N identical layers, and each layer contains several sublayers. Specially, SETCS takes both original and modified code sequences as input of the encoder, while only original summary sequences are fed into the decoder. Besides, the relative positional encoding [45], instead of Transformer's default positional encoding mechanism, is used to leverage representations of relative positions between

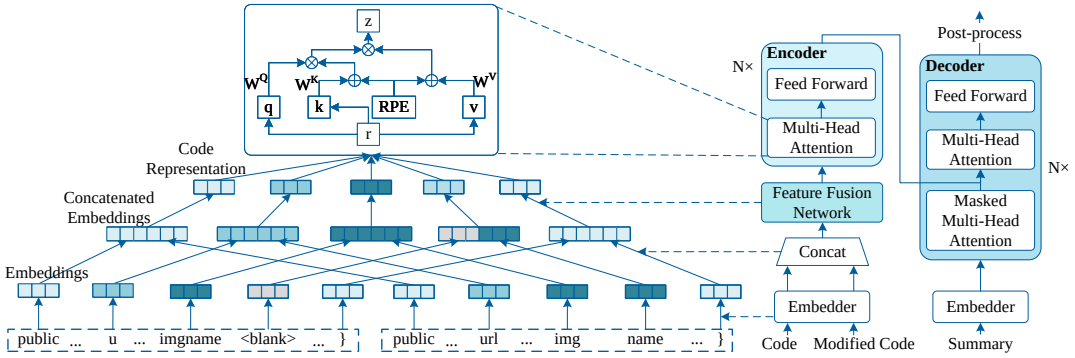


FIGURE 3 Framework of SETCS.

elements of input sequences effectively.

Given that the modified code sequence is typically longer than the original one, after obtaining the optimal subword sequence c^* of a source code sequence C, we insert special <pad> tokens into the original code sequence, making its length equal to the modified code sequence. This operation aims to align these two sequences precisely and avoid improper concatenation in the latter. Using a predefined embedder class, these sequences are converted into dense vector representations that capture the lexical information of both the original and modified code. Following that, embeddings of C and c^* are generated and concatenated together:

$$e_C = \text{concat}\left(e_{c'}, e_{c^*}\right) \tag{4}$$

where $e_{c'}$ and e_{c^*} represent embedding of original and tokenized code sequence separately. The word embeddings shown in 3 are actually stacked together, similar to the representations shown in Figure 1. However, we have separated them for better understanding.

To obtain the code representation r_c that fuses features of both input sequences, the concatenated code embedding e_C is sent into a customized network consisting of a Linear layer and a ReLU activation followed:

$$r_c = \max\left(0, e_C W^e + b^e\right) \tag{5}$$

where W^e and b^e are learnable parameters in the form of matrix and vector, respectively. After that, r_c is fed into the encoder that is composed of a multi-head self-attention sublayer and a feed-forward sublayer.

The multi-head self-attention sublayer consists of h heads to keep the model focused on information at different locations in the input representation. Each head performs the self-attention function in parallel and computes an output sequence $z = (z_1, \ldots, z_x)$ for the input representation of code, $r_c = (r_1, \ldots, r_x)$ of x elements:

$$z_i = \sum_{m=1}^{x} \alpha_{mn} \left(r_m W^V + p^V_{mn}\right) \tag{6}$$

where $r_m \in \mathbb{R}^{d_r}$, $z_i \in \mathbb{R}^{d_z}$. The involved weight coefficient α_{mn} can be formulated as:

$$\alpha_{mn} = \frac{\exp\left(e_{mn}\right)}{\sum_{o=1}^{x} \exp\left(e_{mo}\right)} \tag{7}$$

where e_{mn} is computed via a scaled dot-product attention:

$$e_{mn} = \frac{r_m W^Q \left(r_n W^K + p_{mn}^K\right)^T}{\sqrt{d_z}} \tag{8}$$

The parameter matrices W^Q, W^K, $W^V \in \mathbb{R}^{d_r \times d_z}$ are unique per sublayer and head. The encoding vectors p_{mn}^V, $p_{mn}^K \in \mathbb{R}^{d_z}$ include the relative position information between the input elements r_m and r_n.

Similarly, the outputs of each head are then concatenated together and fed into the feedforward network sublayer. The only difference between our customized feature fusion network and the feedforward layer is that the latter consists of an additional linear transformation:

$$FeedForward(Z) = \max\left(0, ZW^1 + b^1\right) W^2 + b^2 \tag{9}$$

where W^1, W^2, b^1, b^2 are trainable parameters, and Z represents output of the multi-head self-attention sublayer. Note that each sublayer in the model is followed by a residual connection and layer normalization, which are omitted from Figure 3 for brevity.

Compared to the encoder, each layer of the decoder contains an additional masked multi-head self-attention sublayer. This sublayer is designed to prevent the model from seeing future information during the prediction of the next word. It achieves this by applying a mask to the part of the summary sequence that comes after the current word to be predicted. This ensures that the model's attention is focused only on the known part of the sequence during the training phase. After passing through the multi-head self-attention sublayer, the token representations are passed through a feedforward sublayer. Each token representation in the target summary sequence is generated sequentially, with each token's generation based on the current encoding state and the outputs generated for the previous tokens. This process allows the model to build up a context for the current prediction. Finally, the output of the decoder is passed through a softmax activation function. This function maps the raw model output to a probability distribution over the possible next tokens, making it possible to select the most likely next token for the summary.

4 | EXPERIMENTAL SETUP

4.1 | Datasets

Given the indispensable role of project information in code abbreviation expansion, we exclude the dataset open-sourced by Hu et al. [14], even though it is relatively small in scale and has been more widely evaluated, due to its lack of project information. Instead, we conduct experiments using the Funcom dataset [8] and the Java portion of the CodeSearchNet corpus [46], henceforth referred to as CSN-Java.

The CSN corpus, sourced from the GitHub open-source repository, comprises code snippets and corresponding

summary descriptions across six programming languages. Among them, CSN-Java contains approximately 4.7k samples from nearly 0.5M projects. The Funcom dataset, originated from the Sourcerer repository open-sourced by Lopes et al. [47], consists of 2.1M Java samples from around 29k projects, as preprocessed by LeClair et al. [8].

Despite the preliminary filtering of these two code summarization datasets, we observed a significant number of low-quality samples. These could negatively impact or inflate the evaluation results of code summarization models [8, 48]. As a result, we remove samples that meet any of the following conditions during the extraction of code and summaries from source code files.

1) The code cannot be parsed, or it is not recognized as a method declaration. This step is necessary for the process of code abbreviation expansion.

2) The length of the split code or summary sequence is less than three. Most of these samples contain fragmented information with very limited meaning.

3) The summary is identified as Self-Admitted Technical Debt (SATD). These summaries are consisted of meaningless contents such as TODO/Fixme.

4) The summary includes auto-generated phrases such as 'auto generated' or 'generated by', which is usually associated with auto-generated code that need to be removed according to previous studies [8, 12, 46].

5) The contents of the summary are identical, occur more than 300 times, and do not relate to the actual functionality of the corresponding code.

6) The code is an exact or near duplicate, which may inflate model evaluation results [48].

TABLE 2 Statistics of code-summary pairs, parsed identifiers, split identifiers, identified abbreviations, and expanded abbreviations in two datasets.

Dataset	Partition	Code-Summary Pairs	Parsed Identifiers	Split Identifiers	Identified Abbreviations	Expanded Abbreviations
CSN-Java	Train	368,224	12,996,895	22,424,406	3,620,121	602,310
	Valid	16,846	602,239	1,028,123	187,108	30,129
	Test	16,746	595,283	994,543	137,407	26,048
	Total	401,816	14,194,417	24,447,072	3,944,636	658,487
Funcom	Train	1,371,687	16,896,844	28,956,036	4,368,006	931,854
	Valid	86,165	1,077,001	1,850,750	271,223	59,134
	Test	81,642	1,022,339	1,753,158	259,266	61,124
	Total	1,539,494	18,996,184	32,559,944	4,898,495	1,052,112

In the process of dataset filtering, we use the javalang [1] library to parse the code, the SATD detection tool [2] to identify SATDs, and the Near-Duplicate Code Detector [3] to detect cloned codes, respectively. Refer to LeClair et al. [8], both filtered datasets are partitioned into training, validation, and test set by project, maintaining a ratio of 90:5:5. The third column in Table 2 shows the number of code-summary pairs in two filtered datasets. For clarify, these filtered dataset are referred to as the original dataset used in subsequent experiments.

[1] https://github.com/c2nes/javalang

[2] https://github.com/Tbabm/SATDDetector-Core

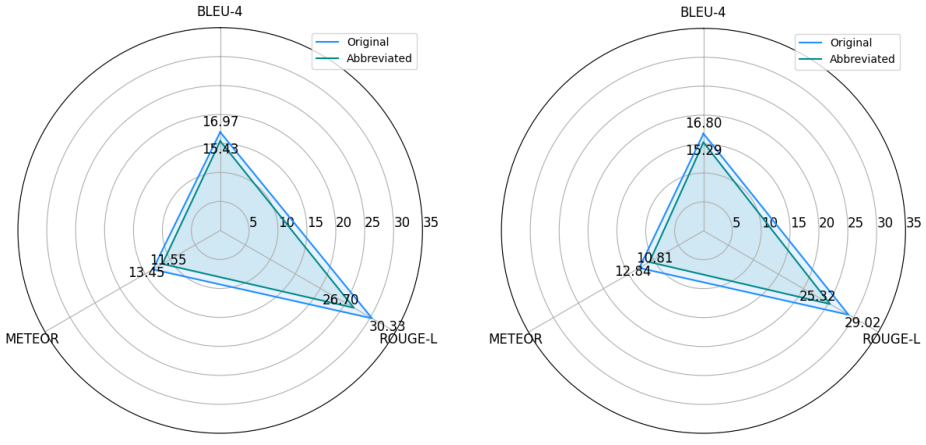
[3] https://github.com/microsoft/near-duplicate-code-detector

4.2 | Exploratory Experiments

To investigate the potential adverse effects of abbreviations in code on code summarization models, we conduct exploratory experiments by actively augmented the prevalence of abbreviations in the code. We then observe the resultant changes in model performance on both the original and abbreviated datasets. This allowed us to assess the impact of abbreviation-rich code on the effectiveness of code summarization.

Specifically, we crawl open-source Java projects with over 20 stars from GitHub and extract parameters, variables, and their corresponding types from the parsed code. If a specific parameter or variable was identified as an acronym, prefix, or dropped-letters abbreviation of its corresponding type, the parameter or variable and it's type will be added to the expansion-abbreviation library. Ultimately, we obtain a library containing 5956 pairs of expansion and abbreviation. Using this library, we replace identifiers in the CSN-Java dataset that match the expansions with corresponding abbreviations. To minimize the influence of manually introduced abbreviations on the original semantic meaning of code in the dataset, identical identifiers in a code snippet will be replaced with the same predetermined abbreviation. If an identifier can be replaced with multiple different abbreviations, it will be randomly replaced with an abbreviation that does not duplicate existing identifiers in the current code snippet. Subsequently, we train and test two representative code summarization models, Seq2Seq and Transformer, on both the original and abbreviated datasets, and evaluate the models' performance using common evaluation metrics, namely BLEU-4, METEOR, and ROUGE-L. Detailed information regarding the models and evaluation metrics used in the experiments will be provided in Section 4.4 and Section 4.5, respectively.

The changes in evaluation metrics for Seq2Seq and Transformer models on the original and abbreviated datasets are depicted in Figure 4. It is evident from the results that increasing the proportion of abbreviations in the dataset negatively impacts the performance of code summarization models. Both models exhibit a decrease of approximately 1.5, 2, and 3.5 points in the BLEU-4, METEOR, and ROUGE-L metrics, respectively, when more abbreviations are introduced into the dataset. These findings suggest the potential for enhancing the performance of code summarization models by minimizing the presence of abbreviations in the datasets.



(a) *Seq2Seq* (b) *Transformer*

FIGURE 4 Radar map showing performance degradation of models on original and abbreviated CSN-Java datasets.

4.3 | Preliminary Experiments

Studies in the code abbreviation expansion domain define a word as an abbreviation if it if not found in an English dictionary [37, 49]. We employ the PyEnchant [4] library to identify abbreviations from split identifiers. Specifically, words not included in the 'en_US' dictionary of the enchant library are considered abbreviations. Additionally, single letters, with the exception of 'a', are also treated as abbreviations to complement the identification results. The last four columns in Table 2 show the number of parsed identifiers, split identifiers, identified abbreviations, and expanded abbreviations in two datasets respectively. It can be found that more than 25% of identifiers contain abbreviations. After leveraging abbreviation expansion algorithms, about 21% of the abbreviations in the Funcom data set are expanded, while this percentage in CSN-Java is approximately 17%. We attribute the difference to: (1) Compared with the Fucnom dataset, each code snippet in CSN-Java contains a larger number of abbreviations on average (about 3 to 10), indicating that there is substantial room for exploration in abbreviation expansion for this dataset. (2) The projects in CSN-Java contain partial methods, which means that only a fraction of the full method implementation is present in the dataset. Consequently, the amount of context information available for expanding abbreviations is inherently limited.

Given that the precision of abbreviation expansion directly or indirectly affects the performance of code summarization models in subsequent experiments, we randomly sampled 1000 expanded abbreviations from two datasets for manual evaluation. Specifically, we found two cases of expansion errors:

1) The term abbreviation is contained within the reference word. For example, 'uri' typically refers to the Uniform Resource Identifier. However, due to the presence of 'Security' in the method name 'getSecurityProtocol', the split 'security', as a reference word, was incorrectly interpreted by the Dropped Letters expansion algorithm as the full name of the abbreviated parameter 'uri'.

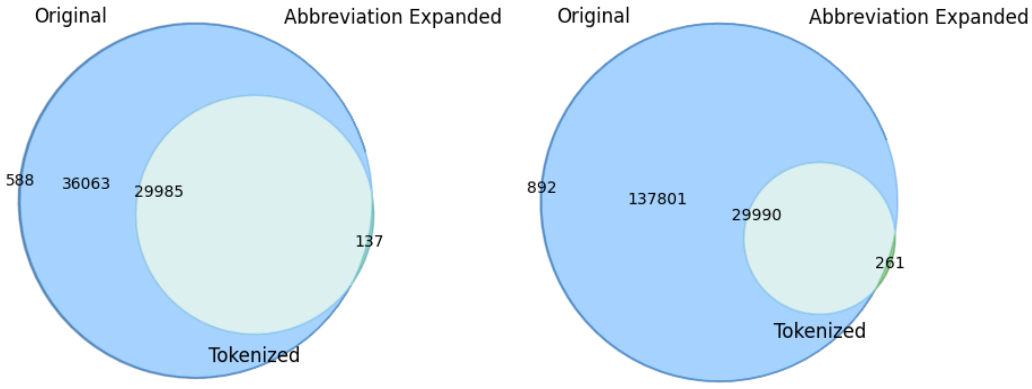
2) There are multiple expansion candidates in the reference words. For example, when expanding the abbreviated parameter 'p' using the Acronym expansion algorithm, the 'player' from the parameter type 'PlayerPreferences' was initially identified and determined as its expansion. However, based on the context of the function, the expansion corresponding to abbreviation 'p' should be 'preferences', or more precisely, 'player preferences'.

Overall, heuristic-based acronym expansion algorithms cannot achieve perfect precision and are susceptible to the influence of developer abbreviation habits. The two types of expansion errors mentioned above are unavoidable. Fortunately, both cases are rare (one case for each type found in 1000 manually evaluated samples), and in most times, developers use abbreviations that include the initials of all words in parameter or variable types, which are correctly expanded by the utilized algorithms.

During the training of the tokenizer, we set the expected vocabulary size to 30k, and retain the top 90% subwords at the end of each iteration. In the process of dataset tokenization, the final tokenization result is selected from the Top-9 candidate subword combinations for both datasets. More detailed information about determining the 'k' value will be discussed in Section 5.3.

To prevent data leakage, we construct the initial vocabulary using only split code and summary words from the training and validation sets. When tokenizing words in the test set, we select the final tokenization results by referring only to the split words from the code.

[4] https://pyenchant.github.io/pyenchant



(a) *CSN-Java* (b) *Funcom*

FIGURE 5 Venn diagram showing statistics of shared and unique tokens for original, abbreviation expanded, and tokenized results of two datasets.

The distributions of shared and unique tokens for original, abbreviation expanded, and tokenized datasets are shown in Figure 5. The outermost navy blue, adjacent dodger blue, and innermost light cyan circles in the venn diagram represent the unique token distribution of the original, abbreviation expanded, and ULM tokenized datasets, respectively. Numbers in the middle represent the quantity of shared tokens of datasets in different status, where we can find that trained tokenizers effectively limit vocabulary size of tokenized datasets to less than 30K; numbers on the leftmost part (colored in navy blue) and rightmost part (colored in light cyan) indicate the quantity of unique tokens in the original and ULM tokenized datasets respectively. The unique tokens in both ULM tokenized datasets are subsequences of longer numerical sequences. In addition, basic numeric tokens of 0-9 are also included in vocabularies to guarantee all fresh numbers appearing in the test set can be properly tokenized via existing numeric tokens. It is worth noting that code abbreviation expansion also reduces the number of unique tokens in original datasets to some extent. Even in small quantities, these eliminated tokens are usually relatively important abbreviated identifiers as mentioned earlier. If we don't expand these abbreviations, they will be generally identified as the <unk> symbols due to the low occurrence frequency. However, they will likely be tokenized into longer character sequences by trained tokenizers after introducing the subword segmentation algorithm. Both circumstances may result in the loss of critical information. Therefore, we believe that it is necessary to perform abbreviation expansion before training and adopting the tokenizer. Results of ablation experiments (Section 5.2) and example analysis (Section 5.4) on the Transformer baseline will demonstrate the effectiveness of abbreviation expansion as well as its usefulness in combining with the introduced Unigram-based subword segmentation method.

4.4 | Baseline Models

To verify the effectiveness of our proposed methods, we conduct experiments using four representative code summarization models:

Seq2Seq: A classical open-sourced NMT framework [50], based on recurrent neural network (RNN) and equipped with an attention mechanism. Specifically, this baseline uses LSTM [51] to generate summaries for given code snippets and is adopted by Rencos [27], Re2Com [28], MLCS [29] as model backbone.

Transformer: The vanilla Transformer [6] model incorporated with relative positional encoding mechanism. Specifically, it has been employed by NCS [13], API2Com [24], SiT [20], AST-Trans [16] and the framework of SETCS presented in this paper.

NCS [13]: An enhanced Transformer designed for code summarization that utilizes both relative positional encoding and copying mechanism [52] for the first time. The copying mechanism enables the Transformer to generate words from the vocabulary and copy from the input source code.

MLCS [29]: A state-of-the-art code summarization framework based on meta-learning and code retrieval. By optimizing a unique code summarizer for each target code snippet knowledge learned from the retrieved similar examples, MLCS was able to outperform typical deep-learning models and retrieval-based neural models.

It is worth noting that since both code summarization datasets came from open-source communities, pre-trained code models typically utilize larger-scale open-source corpora for pre-training, these models should have encountered test samples from the datasets used in our study during the pre-training stage. Therefore, we excluded these models from the baselines to avoid threats of pre-training technique and data leakage to the internal validity of this study.

Referring to prior works [13, 18, 29, 53], we limit the maximum input and output lengths for all models to 150 and 30, correspondingly. Meanwhile, we set the batch size, vocabulary size, maximum training epochs, and beam size to 64, 30K, 30, and 4, respectively. The best model for code summarization is determined based on the BLEU scores from the validation set, and the training process will be halted if there is no enhancement in the BLEU score over 10 successive epochs. All experiments are conducted on a Linux server, which is equipped with a NVIDIA Tesla P40 GPU. The duration of experiments executed on the CSN-Java dataset is less than a day, while those performed on the Funcom dataset typically require approximately three days.

4.5 | Evaluation Metrics

The commonly adopted evaluation metrics, BLEU [54], METEOR [55], and ROUGE [56], are predicated on the same underlying scenario. Specifically, for each candidate text, which is the prediction result generated by the trained model, there exists a corresponding reference text within the dataset, typically a reference summary authored by the developer. The computation of these evaluation metrics are fundamentally based on precision and recall scores:

$$P_n = \frac{\text{gram}_n(\text{ pred , ref })}{\text{gram}_n(\text{ pred })}, R_n = \frac{\text{gram}_n(\text{ pred , ref })}{\text{gram}_n(\text{ pred })} \tag{10}$$

where pred, ref, and $gram_n$ refers to the candidate text, reference text, and the overlapping n-grams, respectively.

The BLEU metric highlights precision, which calculates the geometric average of $gram_n$ matches between pred and ref:

$$\text{BLEU} = \sigma \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log P_n\right) \tag{11}$$

The classical BLEU-4 is calculated by $gram_4$.

The METEOR metric further considers recall, word form, and synonym matching, which creates unigram alignment between pred and ref, while longer $gram_n$ alignment is prioritized in this stage.

$$\text{METEOR} = \sigma \cdot \frac{P_n R_n}{(1-\alpha) R_n + \alpha P_n} \tag{12}$$

where α is the default parameter used for evaluation.

Note that the penalty factor σ differs in different evaluation metrics. The ROUGE metric calculates $gram_n$ between pred and ref. The calculation formula can be expressed as:

$$\text{ROUGE} = \frac{2 P_n R_n}{R_n + P_n} \tag{13}$$

The widely used ROUGE-L is calculated based on the longest common sequence.

However, the above-mentioned metrics primarily focus on evaluating textual similarity between candidate and reference texts, which may penalize semantically equivalent texts that differ in wording. To complement these metrics and capture the extent to which the candidate text aligns with the semantics of the corresponding code snippet, we also adopt the newly proposed SIDE metric [57], which has been shown to align well with human assessment. This metric measures the cosine similarity between embeddings of the candidate text and the corresponding code sequence:

$$\text{SIDE} = \cos\left(e_{pred}, e_C\right) \tag{14}$$

where e refers to embedding generated by a fine-tuned MPNet [58] model via contrastive learning.

In all subsequent experiments, we employ the BLEU-4, METEOR, ROUGE-L and SIDE metrics to evaluate the quality of the summaries generated by the code summarization models, with higher metric scores representing better quality of generated summaries. For fair comparison, model predictions as well as ground-truth references before and after tokenization are used for calculation, and the mean score is deemed as the final result for each evaluation metric.

5 | ANALYSIS OF EXPERIMENTAL RESULTS

For simplicity, this section adopts CAE and ULM to represent Code Abbreviation Expansion and ULM-based subword segmentation, respectively. In addition, best results of each metric in tables are boldfaced.

5.1 | Method Validation

Experimental results of SETCS compared with baselines and improvements of the baselines after adopting CAE and ULM on two datasets are shown in Table 3 and 4 respectively.

As shown in Table 3, compared to the Transformer baseline, the proposed SETCS, which further harnesses the critical semantic information provided by both CAE and ULM, yields an improvement of over 2 absolute points across almost all evaluation metrics on both the CSN-Java and Funcom datasets. As suggested by Roy et al. [59], this assures systematic enhancements in summarization quality, implying that our proposed methods, in conjunction with

TABLE 3 Experimental results of SETCS and baselines on two datasets.

Model	CSN-Java				Funcom			
	BLEU-4	METEOR	ROUGE-L	SIDE	BLEU-4	METEOR	ROUGE-L	SIDE
Seq2Seq	16.87	13.37	30.24	83.62	25.79	17.44	38.58	85.71
Transformer	16.65	12.76	28.92	83.55	25.11	17.31	37.60	84.06
MLCS	18.17	12.71	30.66	84.64	27.15	18.34	40.34	86.91
NCS	**18.22**	13.41	**31.72**	**85.86**	**27.81**	18.82	41.07	87.80
SETCS	18.14	**13.96**	31.66	85.78	**27.81**	**19.62**	**41.72**	**88.28**

TABLE 4 Improvements of baselines after adopting both CAE and ULM on two datasets.

Model	CSN-Java				Funcom			
	BLEU-4	METEOR	ROUGE-L	SIDE	BLEU-4	METEOR	ROUGE-L	SIDE
Seq2Seq	16.87	13.37	30.24	83.62	25.79	17.44	38.58	85.71
Seq2Seq w/ Both	17.40 (+3.1%)	13.67 (+2.2%)	30.72 (+1.6%)	84.57 (+1.1%)	26.26 (+1.8%)	18.35 (+5.2%)	39.54 (+2.5%)	86.25 (+0.6%)
Transformer	16.65	12.76	28.92	83.55	25.11	17.31	37.60	84.06
Transformer w/ Both	17.60 **(+5.7%)**	**14.04** **(+10.0%)**	30.82 **(+6.6%)**	84.83 **(+1.5%)**	26.95 **(+7.3%)**	18.51 **(+6.9%)**	40.12 **(+6.7%)**	86.72 **(+3.2%)**
MLCS	18.17	12.71	30.66	84.64	27.15	18.34	40.34	86.91
MLCS w/ Both	18.45 (+1.5%)	12.98 (+2.1%)	31.29 (+2.1%)	85.35 (+0.8%)	27.96 (+3.0%)	18.88 (+2.9%)	41.29 (+2.4%)	87.94 (+1.2%)
NCS	18.22	13.41	31.72	85.86	27.81	18.82	41.07	87.80
NCS w/ Both	**18.51** (+1.2%)	13.99 (+4.3%)	**32.25** (+1.7%)	**85.99** (+0.2%)	**28.02** (+0.7%)	**19.10** (+1.6%)	**41.31** (+0.6%)	**88.04** (+0.3%)

the feature fusion approach, could be effectively employed in other code summarization models that utilize a similar framework to SETCS. Notably, the NCS model, despite being proposed earlier, still outperforms the state-of-the-art MLCS and other baseline models that merely leverage code-related semantic information on both datasets. Besides, the improvement of SETCS over NCS is less significant, underscoring the potent potential of the copying mechanism. Nonetheless, the primary focus of this study is to validate the effectiveness and applicability of CAE and ULM on existing code summarization models, rather than proposing a new state-of-the-art model. More importantly, SETCS could serve as a robust baseline or backbone for future studies on two well-curated datasets.

Experimental results in Table 4 demonstrate that the performance of all code summarization models improves with the adoption of our proposed methods. Specifically, the following conclusions can be drawn:

1) Compared to the smaller CSN-Java dataset, the overall performance improvement of all baseline models on the Funcom dataset is more significant. Taking the prevailing Transformer model as an example, after adopting CAE and ULM, it can achieve score improvements of 7.3%, 6.9%, 6.7%, and 3.2% in terms of BLEU-4, METEOR, ROUGE-L, and SIDE, respectively. More significantly, collaboratively utilizing both methods could yield 10.0% performance gain for Transformer regarding the METEOR metric on the CSN-Java dataset, which enables the baseline comparable to SETCS and the improved NCS.

2) In comparison to the other three metrics, the majority of models exhibit relatively larger absolute score gains with respect to the ROUGE-L metric on both datasets. We attribute this phenomenon to the extension of the reference summary by ULM, coupled with the more granular subword representation. This enables the model to capture more semantic information and contributes to the observed significant improvement.

3) Overall, the NCS model exhibits the least performance improvement following the adoption of the proposed methods. This outcome is reasonable given that the multiple identical expansion results introduced by CAE could potentially interfere with the copying mechanism employed by NCS. Furthermore, both methods, particularly ULM, might increase the code length. Any content that exceeds the maximum code length limitation is truncated during the stages of model training and inference, which could lead to the loss of crucial information.

5.2 | Ablation Experiments

Table 5 presents the experimental results of the Transformer and SETCS models after incorporating CAE, ULM, and both methods, separately, on two different datasets. The primary distinction between the two sets of ablation experiments lies in the fact that only the datasets are modified in the first set of experiments, whereas in the latter set, modifications are also made to the models. Besides, performance of the Transformer model can be seen as the ablation result of SETCS without the feature fusion network.

In ablation experiments results of the first group, it is clearly that both methods can improve the performance of Transformer to various degrees, among which ULM plays a more important role. The combination of two methods could bring further improvements in terms of almost all textual similarity-based evaluation metrics, where the minor degradation of the METEOR and ROUGE-L metrics on CSN-Java can be neglected as difference of the absolute value is less than 0.1. Notably, on the CSN-Java dataset, for both Transformer and SETCS, the proposed ULM could bring about the best improvement for the semantic similarity-based SIDE metric. In fact, compared with the traditional split method, code summarization models adopting ULM tokenizers have better evaluation results on reference summaries whatever before and after tokenization. Moreover, the experimental results of 'Transformer w/ Both' with 30k vocabulary on the Funcom dataset are still better than that with a 50k vocabulary. All these findings further prove that CAE and ULM can effectively introduce and expose more critical semantic information, which plays a key role in improving model's performance. In addition, when testing ULM tokenizers in preliminary experiments, we found that tokeniz-

TABLE 5 Ablation experiment results of Transformer and SETCS on two datasets.

Model	CSN-Java				Funcom			
	BLEU-4	METEOR	ROUGE-L	SIDE	BLEU-4	METEOR	ROUGE-L	SIDE
Transformer	16.65	12.76	28.92	83.55	25.11	17.31	37.60	84.06
Transformer w/ CAE	17.13	13.38	30.20	84.98	25.47	17.50	38.14	84.31
Transformer w/ ULM	17.42	**14.13**	**30.87**	**85.36**	25.86	17.87	38.79	85.08
Transformer w/ Both	**17.60**	14.04	30.82	84.83	**26.95**	**18.51**	**40.12**	**86.72**
SETCS w/o CAE	17.84	13.73	31.19	**85.89**	27.79	**19.66**	41.68	88.15
SETCS w/o ULM	17.96	**14.00**	31.61	85.73	27.71	19.64	41.65	88.14
SETCS	**18.14**	13.96	**31.66**	85.78	**27.81**	19.62	**41.72**	**88.28**

ers trained with a smaller vocabulary will tokenize most nouns in plural forms, resulting in substantial score gains in terms of the ROUGE-L metric and decreased performance on other metrics, which indicates that the granularity of subword segmentation is not the finer the better. Therefore, when training a tokenizer for the code summarization model, factors such as size of the desired vocabulary and length limitations of model's input and output should be comprehensively considered.

For the second group of ablation experiments' results, it's interesting that CAE plays a more significant role in improving performance on the CSN-Java dataset, while ULM plays a more significant role in the Funcom dataset. Each of the two methods significantly boosts the performance of all metrics compared to the Transformer baseline, which indicate the effectiveness of the feature fusion network equipped by SETCS. However, the collaboration of the two methods yield relatively fewer improvements across most evaluation metrics, which contradicts the earlier findings. We speculate that the customized network operated in SETCS is capable of learning more specific transformations but struggles with learning complex patterns when both methods are combined. More specifically, the modification of code snippets introduced by CAE is fixed in most circumstances as its algorithms are predefined to expand abbreviations for parameters or variables in very specific places, while modifications brought by ULM are randomly distributed in different locations of the code. In short, this phenomenon can be attributed to the limitations of the feature fusion strategy employed by SETCS, and more effective approaches are yet to be discovered. Actually, we have explored many other feature fusion strategies but reaped relatively fewer improvements compared to method presented in this paper. These tested strategies include concatenating embeddings of both original and modified code sequences from another dimension, concatenating embeddings of original and differences between both code sequences, concatenating both code representations directly, and utilizing different customized networks when transforming concatenated embeddings to code representations. Therefore, we leave this challenge for future research. For the purpose of better illustration and broader applicability, experiments in the subsequent sections are conducted on the Transformer baseline.

5.3 | ULM Tuning & Comparison

In order to determine the appropriate 'k' value for the Top-k subword combination candidates, as discussed in Section 3.2, we carry out experiments using the 'Transformer w/ ULM' model on CSN-Java, with 'k' values ranging from 1 to 13 and the span set to 2. Additionally, we conduct comparative experiments to further examine the effects of the

introduced ULM algorithm against basic subword segmentation algorithms. The choice to perform these experiments on CSN-Java instead of Funcom is primarily driven by considerations of time efficiency.

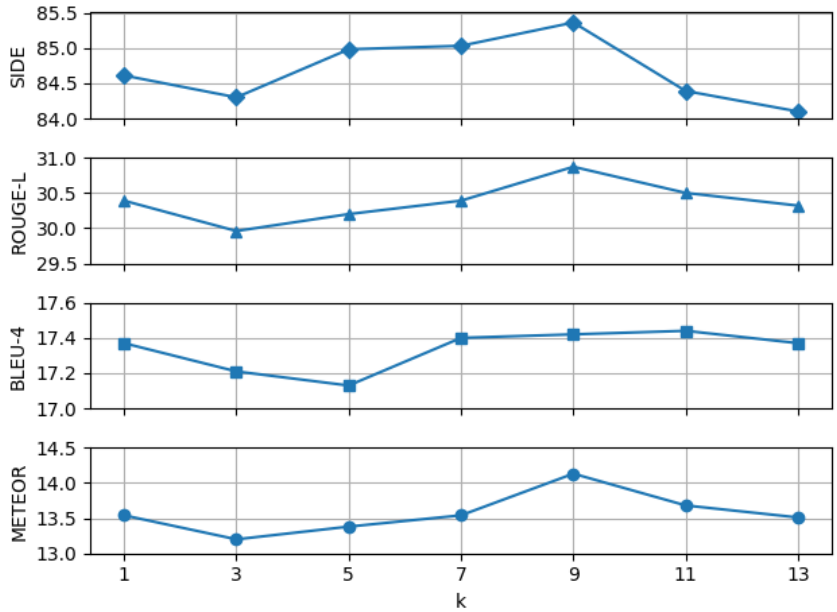


FIGURE 6 Experimental results of 'Transformer w/ ULM' with different k values on CSN-Java.

TABLE 6 Experimental results of Transformer with different subword segmentation algorithms on CSN-Java.

Model	BLEU-4	METEOR	ROUGE-L	SIDE
Transformer	16.65	12.76	28.92	83.55
Transformer w/ BPE_Basic	17.21	13.21	30.12	84.59
Transformer w/ ULM_Basic	17.15	13.56	30.53	84.62
Transformer w/ ULM_Top-9	**17.42**	**14.13**	**30.87**	**85.36**

Figure 6 displays the changing curves of four evaluation metrics, where the trend of all curves goes down, up, and then down. Table 6 shows experimental results regarding different subword segmentation algorithms, where results of the Transformer baseline, Transformer with the basic BPE algorithm, Transformer with the basic ULM algorithm, and Transformer with the introduced ULM algorithm are listed from up to down. The essential algorithms operated by both basic and introduced ULM are the same, but the latter further optimized the training and tokenization procedures of code-specific tokenizers to obtain semantic-preserved results. Besides, the basic WordPiece algorithm is not involved since it is not open-sourced. The 'k' value of the introduced ULM algorithm is set to 9 in all experiments on CSN-Java, as the overall performance of 'Transformer w/ ULM' by selecting tokenization results from Top-9 candidates is proved to be the best.

Overall, all subword segmentation algorithms could significantly improve the performance of Transformer in terms all metrics, which is expected. Specifically, the difference between each pair of subword segmentation algorithms is relatively small in terms of BLEU-4, but differences are obvious when it comes to other three metrics. The tactic of selecting most semantic-preserved tokenization results from Top-k subword combination candidates introduced in this paper is proved to be more effective compared with the direct adoption of basic ULM algorithm, which performance is sightly inferior to the introduced ULM with k set to 1. To sum up, the introduction of subword segmentation algorithms can bring about remarkable improvements for code summarizations models, and the performance could be further upgraded if more code-related semantic information can be preserved.

5.4 | Example Analysis

Table 7 illustrates two examples from Funcom. The last four rows of the table list generated summaries of the Transformer model before and after using the proposed method(s).

TABLE 7 Two illustrative examples from the Funcom dataset.

Function ID	27906163	44895355
Code	`public SummaryItem getSummaryItemForMsg(int msgNumber){ return (SummaryItem) summaryItems.get(msgNumber-1); }`	`public void setFeedbacktype (String feedbacktype){ setPropertyString(QTI_RDFS+ "feedbacktype)", feedbacktype)); }`
Summary	return the summary item info for a particular message number .	sets the feedbacktype to the given string.
Transformer	returns the summary item for the given msg number .	sets the name of the qti rdfs property.
Transformer w/ CAE	returns the summary item for a given message number .	sets the <unk>property.
Transformer w/ ULM	returns the summary item for the given message number .	sets the feedback type property.
Transformer w/ Both	returns the summary item for the given message number .	sets the feedback type .

For the first code snippet, after using CAE to expand the abbreviation 'msg' nested in the formal parameter 'msgNumber' to 'message' Transformer accurately generates the corresponding summaries for the expanded formal parameter 'message number'. It is interesting that ULM also enables the model to generate the correct summary for the abbreviated formal parameter. We speculate the code summarization model has the potential to generate the corresponding full names for corresponding abbreviations, and semantic information exposed by trained tokenizers convinces the model that the full name of abbreviation 'msg' in the code should be 'message'. In other words, both methods effectively enhanced the ability of language alignment for code summarization models.

When it comes to the second code snippet, although the formal parameter 'feedbacktype' appears multiple times in the code, it is still being identified as <unk> due to its overall low frequency in the dataset, which is reflected in the summary generated by 'Transformer w/ CAE'. Instead of generating <unk> with a relative small probability, the vanilla Transformer finally chose 'qti rdfs' as the summary, which appears in the code but has nothing to do with the actual functionality of the code. After tokenizing 'feedbacktype' into 'feedback type' using the Unigram subword algorithm, the model correctly understood its meaning and accurately generated a corresponding summary for it.

In summary, the methods proposed in this paper improve the performance of the code summarization model at the semantic level, and the two methods complement each other. Code abbreviation expansion eliminates some rare words. It also avoids the unigram subword algorithm tokenizing them into overlong subwords. The subword algorithm can expose more abbreviation information. If the abbreviation 'img' nested in the identifier 'imgname' contained in the code snippet of Table 1 is accurately tokenized and expanded, code summarization models will be more likely to generate the correct summary 'image' for the code. Therefore, the subword segmentation algorithm also has practical implications for the study of abbreviation expansion, and proposing more advanced techniques to combine the copying mechanism with methods proposed in this paper is worthy of further exploration as well.

6 | CONCLUSION AND FUTURE WORK

In this paper, we propose two methods to enhance the semantic performance of code summarization models. By expanding abbreviations within identifiers, we eliminate the uncertainty of the corresponding semantic information and allow the model to focus more on the identifiers themselves rather than their types. Moreover, by leveraging the Unigram subword segmentation algorithm, we train code-specific tokenizers to tokenize code into more granular subword sequences, which enables the code summarization model to capture more critical information during training and inference stages. Experimental results from three typical code summarization models and the presented SETCS on two datasets demonstrate the effectiveness of our proposed methods.

Future works include:

1) Incorporate advanced feature fusion techniques into SETCS to unlock the full potential of our proposed methods, or employ the framework to verify other automatic code summarization approaches at either the semantic or structural level.

2) Explore further how expanding code abbreviations in different proportions and types impacts the performance of code summarization models, and how the performance is influenced by different subword segmentation algorithms with varying vocabulary sizes.

3) Apply the methods proposed in this paper to pre-trained code models and other program understanding or generation tasks, particularly in conjunction with prompt learning [60] or meta-learning techniques. This could potentially enhance the efficiency and performance of these models and tasks.

To facilitate future research, we have made datasets used in experiments, as well as the source code of SETCS, publicly available at https://github.com/Hugo-Liang/SETCS.

CONFLICT OF INTEREST

The authors have no competing interests to declare that are relevant to the content of this paper.

DATA AVAILABILITY STATEMENT

The authors declare that the data supporting the findings of this study are available within the paper.

ORCID

Yuguo Liang https://orcid.org/0009-0002-8738-2891

REFERENCES

[1] Storey MA. Theories, methods and tools in program comprehension: Past, present and future. In: 13th International Workshop on Program Comprehension; 2005. p. 181–191.

[2] He H. Understanding Source Code Comments at Large-Scale. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering New York, NY, USA: Association for Computing Machinery; 2019. p. 1217–1219.

[3] Moreno L, Marcus A. Automatic Software Summarization: The State of the Art. In: Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings New York, NY, USA: Association for Computing Machinery; 2018. p. 530–531.

[4] Rai S, Belwal RC, Gupta A. A Review on Source Code Documentation. ACM Trans Intell Syst Technol 2022 jun;13(5).

[5] Iyer S, Konstas I, Cheung A, Zettlemoyer L. Summarizing Source Code using a Neural Attention Model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics Berlin, Germany: Association for Computational Linguistics; 2016. p. 2073–2083.

[6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6000–6010.

[7] Niu C, Li C, Luo B, Ng V. Deep Learning Meets Software Engineering: A Survey on Pre-Trained Models of Source Code. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence International Joint Conferences on Artificial Intelligence Organization; 2022. p. 5546–5555.

[8] LeClair A, Jiang S, McMillan C. A neural model for generating natural language summaries of program subroutines. In: Proceedings of the 41st International Conference on Software Engineering IEEE Press; 2019. p. 795–806.

[9] Newman CD, Decker MJ, Alsuhaibani RS, Peruma A, Kaushik D, Hill E. An Empirical Study of Abbreviations and Expansions in Software Artifacts. In: 2019 IEEE International Conference on Software Maintenance and Evolution; 2019. p. 269–279.

[10] Sharma R, Chen F, Fard F. LAMNER: Code comment generation using character language model and named entity recognition. In: Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension New York, NY, USA: Association for Computing Machinery; 2022. p. 48–59.

[11] Cheng W, Hu P, Wei S, Mo R. Keyword-Guided Abstractive Code Summarization via Incorporating Structural and Contextual Information. Inf Softw Technol 2022 oct;150(C).

[12] Hu X, Li G, Xia X, Lo D, Jin Z. Deep Code Comment Generation with Hybrid Lexical and Syntactical Information. Empirical Software Engineering 2020 may;25(3):2179–2217.

[13] Ahmad W, Chakraborty S, Ray B, Chang KW. A Transformer-based Approach for Source Code Summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics Online: Association for Computational Linguistics; 2020. p. 4998–5007.

[14] Hu X, Li G, Xia X, Lo D, Jin Z. Deep code comment generation. In: Proceedings of the 26th Conference on Program Comprehension New York, NY, USA: Association for Computing Machinery; 2018. p. 200–210.

[15] Zhou Z, Yu H, Fan G, Huang Z, Yang X. Summarizing Source Code with Hierarchical Code Representation. Inf Softw Technol 2022 mar;143(C).

[16] Tang Z, Shen X, Li C, Ge J, Huang L, Zhu Z, et al. AST-trans: Code summarization with efficient tree-structured attention. In: Proceedings of the 44th International Conference on Software Engineering New York, NY, USA: Association for Computing Machinery; 2022. p. 150–162.

[17] Zhang J, Wang X, Zhang H, Sun H, Wang K, Liu X. A novel neural source code representation based on abstract syntax tree. In: Proceedings of the 41st International Conference on Software Engineering IEEE Press; 2019. p. 783–794.

[18] Lin C, Ouyang Z, Zhuang J, Chen J, Li H, Wu R. Improving Code Summarization with Block-wise Abstract Syntax Tree Splitting. In: 2021 IEEE/ACM 29th International Conference on Program Comprehension; 2021. p. 184–195.

[19] Liu S, Chen Y, Xie X, Siow JK, Liu Y. Retrieval-Augmented Generation for Code Summarization via Hybrid GNN. In: International Conference on Learning Representations; 2021. .

[20] Wu H, Zhao H, Zhang M. Code Summarization with Structure-induced Transformer. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 Online: Association for Computational Linguistics; 2021. p. 1078–1090.

[21] Gao S, Gao C, He Y, Zeng J, Nie L, Xia X, et al. Code Structure–Guided Transformer for Source Code Summarization. ACM Trans Softw Eng Methodol 2023 feb;32(1).

[22] Guo J, Liu J, Liu X, Li L. Summarizing source code through heterogeneous feature fusion and extraction. Information Fusion 2024;103:102058.

[23] Hu X, Li G, Xia X, Lo D, Lu S, Jin Z. Summarizing Source Code with Transferred API Knowledge. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence AAAI Press; 2018. p. 2269–2275.

[24] Shahbazi R, Sharma R, Fard FH. API2Com: On the Improvement of Automatically Generated Code Comments Using API Documentations. In: 2021 IEEE/ACM 29th International Conference on Program Comprehension; 2021. p. 411–421.

[25] Xie R, Ye W, Sun J, Zhang S. Exploiting Method Names to Improve Code Summarization: A Deliberation Multi-Task Learning Approach. In: 2021 IEEE/ACM 29th International Conference on Program Comprehension; 2021. p. 138–148.

[26] Li M, Yu H, Fan G, Zhou Z, Huang Z. Enhancing code summarization with action word prediction. Neurocomputing 2024;563:126777.

[27] Zhang J, Wang X, Zhang H, Sun H, Liu X. Retrieval-Based Neural Source Code Summarization. New York, NY, USA: Association for Computing Machinery; 2020. p. 1385–1397.

[28] Wei B, Li Y, Li G, Xia X, Jin Z. Retrieve and Refine: Exemplar-Based Neural Comment Generation. In: Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering New York, NY, USA: Association for Computing Machinery; 2021. p. 349–360.

[29] Zhou Z, Yu H, Fan G, Huang Z, Yang K. Towards Retrieval-Based Neural Code Summarization: A Meta-Learning Approach. IEEE Transactions on Software Engineering 2023;49(4):3008–3031.

[30] Xie R, Hu T, Ye W, Zhang S. Low-Resources Project-Specific Code Summarization. In: Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering New York, NY, USA: Association for Computing Machinery; 2023. .

[31] Guo D, Ren S, Lu S, Feng Z, Tang D, LIU S, et al. GraphCodeBERT: Pre-training Code Representations with Data Flow. In: International Conference on Learning Representations; 2021. .

[32] Niu C, Li C, Ng V, Ge J, Huang L, Luo B. SPT-Code: Sequence-to-Sequence Pre-Training for Learning Source Code Representations. In: 2022 IEEE/ACM 44th International Conference on Software Engineering; 2022. p. 01–13.

[33] Feng Z, Guo D, Tang D, Duan N, Feng X, Gong M, et al. CodeBERT: A Pre-Trained Model for Programming and Natural Languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020 Online: Association for Computational Linguistics; 2020. p. 1536–1547.

[34] Wang Y, Wang W, Joty S, Hoi SCH. CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 8696–8708.

[35] Ahmad W, Chakraborty S, Ray B, Chang KW. Unified Pre-training for Program Understanding and Generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Online: Association for Computational Linguistics; 2021. p. 2655–2668.

[36] Jiang Y, Liu H, Zhang L. Semantic Relation Based Expansion of Abbreviations. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering New York, NY, USA: Association for Computing Machinery; 2019. p. 131–141.

[37] Jiang Y, Liu H, Zhu J, Zhang L. Automatic and Accurate Expansion of Abbreviations in Parameters. IEEE Transactions on Software Engineering 2020;46(7):732–747.

[38] Jiang Y, Liu H, Zhang Y, Niu N, Zhao Y, Zhang L. Which Abbreviations Should Be Expanded? In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering New York, NY, USA: Association for Computing Machinery; 2021. p. 578–589.

[39] Jiang Y, Liu H, Jin J, Zhang L. Automated Expansion of Abbreviations Based on Semantic Relation and Transfer Expansion. IEEE Transactions on Software Engineering 2022;48(2):519–537.

[40] Gage P. A New Algorithm for Data Compression. C Users J 1994 feb;12(2):23–38.

[41] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics Berlin, Germany: Association for Computational Linguistics; 2016. p. 1715–1725.

[42] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach; 2019.

[43] Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al., Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation; 2016.

[44] Kudo T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics Melbourne, Australia: Association for Computational Linguistics; 2018. p. 66–75.

[45] Shaw P, Uszkoreit J, Vaswani A. Self-Attention with Relative Position Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 464–468.

[46] Husain H, Wu HH, Gazit T, Allamanis M, Brockschmidt M, CodeSearchNet Challenge: Evaluating the State of Semantic Code Search; 2020.

[47] Lopes C, Bajracharya S, Ossher J, Baldi P, UCI Source Code Data Sets; 2010.

[48] Allamanis M. The Adverse Effects of Code Duplication in Machine Learning Models of Code. In: Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software New York, NY, USA: Association for Computing Machinery; 2019. p. 143–153.

[49] Di Martino S, Maggio V, Corazza A. LINSEN: An efficient approach to split identifiers and expand abbreviations. In: Proceedings of the 2012 IEEE International Conference on Software Maintenance USA: IEEE Computer Society; 2012. p. 233–242.

[50] Klein G, Kim Y, Deng Y, Senellart J, Rush A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In: Proceedings of ACL 2017, System Demonstrations Vancouver, Canada: Association for Computational Linguistics; 2017. p. 67–72.

[51] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation 1997;9(8):1735–1780.

[52] See A, Liu PJ, Manning CD. Get To The Point: Summarization with Pointer-Generator Networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics Vancouver, Canada: Association for Computational Linguistics; 2017. p. 1073–1083.

[53] Wei B, Li G, Xia X, Fu Z, Jin Z. In: Code generation as a dual task of code summarization Red Hook, NY, USA: Curran Associates Inc.; 2019. .

[54] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics USA: Association for Computational Linguistics; 2002. p. 311–318.

[55] Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization Ann Arbor, Michigan: Association for Computational Linguistics; 2005. p. 65–72.

[56] Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74–81.

[57] Mastropaolo A, Ciniselli M, Penta MD, Bavota G. Evaluating Code Summarization Techniques: A New Metric and an Empirical Characterization. In: 2024 IEEE/ACM 46th International Conference on Software Engineering Los Alamitos, CA, USA: IEEE Computer Society; 2024. p. 1002–1002.

[58] Song K, Tan X, Qin T, Lu J, Liu TY. MPNet: Masked and permuted pre-training for language understanding. In: Proceedings of the 34th International Conference on Neural Information Processing Systems Red Hook, NY, USA: Curran Associates Inc.; 2020. .

[59] Roy D, Fakhoury S, Arnaoudova V. Reassessing Automatic Evaluation Metrics for Code Summarization Tasks. In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering New York, NY, USA: Association for Computing Machinery; 2021. p. 1105–1116.

[60] Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Comput Surv 2023 jan;55(9).