Learning to parse developer documentation

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

In this work, we propose a regular expression (RE) synthesizer for entity linking in source code and API documentation which incorporates semantic and relational features from the surrounding context. We demonstrate the effectiveness of our synthesizer on a link prediction task using a corpus of Java code and developer documentation, and demonstrate the effectiveness of model-based representation learning for interpretable link prediction in the source-to-source and doc-to-doc setting.

1 Introduction

Semantic information plays a key role in natural and programming languages. In addition to its syntax, source code contains a rich denotational and operational semantics [Henkel et al., 2018]. To effectively reason about code in semantically similar but syntactically diverse settings requires models which incorporate features from the call graph [Gu et al., 2016, 2018, Liu et al., 2019] and surrounding typing context [Allamanis et al., 2017]. Many semantic features, such as data and control flow [Si et al., 2018] can be represented using a directed acyclic graph (DAG), which admits linear-time solutions to a variety of graph problems, including topological sorting, single-source shortest path and reachability queries.

The field of natural language has also developed a rich set of graph-based representations, including Reddy et al. [2016]'s and other typed attribute grammars which can be used to reason about syntactic and semantic relations between natural language entities. In the pointer network architecture, Vinyals et al. [2015b,a] emphasize the importance of constructing permutation-invariant representations and show SOTA improvements in semantic labeling tasks from dependency parsing [Ma et al., 2018], named-entity recognition [Lample et al., 2016], and coreference resolution where sequence-based techniques often struggle. Pointer networks have been recently extended with a copy-mechanism [Li et al., 2017] to handle out-of-vocabulary code tokens.

These tools are useful for study source code and natural language artifacts, such as documentation [Yang et al., 2016] and link-based [Zhang and Chen, 2018] analysis. Entity alignment in doc-to-doc (D2D) and source-to-source (S2S) is a straightforward application of existing link prediction [Zhang and Chen, 2018] and code embedding [Gu et al., 2018] techniques, but examples of cross-domain applications remain scarce. Robillard and Chhetri [2015], Robillard et al. [2017] first explore the task of suggesting reference API docs from source code using human feedback. Prior work also studies the relationship between comments and code entities [Iyer et al., 2018, Panthaplackel et al., 2020] using machine learning, but only within source code.

Maintainers of popular software projects often publish web-based developer docs, typically in markup languages like HTML or Markdown [Terrasa et al., 2018]. These documents contain natural language sentences, markup, and hyperlinks to other documents and source code

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

artifacts. Both the document and link graph contain important semantic information. The markup describes the text in relation to other entities in the document hierarchy [Yang et al., 2016], while the link graph describes relationships between related documents or source code entities. To compensate for the sparsity of hyperlinks between code and documentation, new techniques are likely required.

Unlike natural languages where polysemy is a common phenomenon [Ganea et al., 2016], most non-trivial tokens in source code are unique, even in vast corpora. While the frequency of out-of-vocabulary (OOV) tokens presents a significant challenge for language modeling, it is an auspicious property for code search, where two lexical matches almost always refer to a single entity. A skilled programmer develops an instinct for quickly locating entities in software projects. Suppose we are given the string `AbstractSingletonFactoryBean`. We observe it has the following properties:

1. The string is camel-case, indicating it refers to an entity in a camel-case language.
2. The string contains the substring `Bean`, a common token in the Java language.
3. The string begins with a capital letter, indicating it refers to a class or interface.

Developers often use a tool called `grep` to locate files, which accepts queries written in the regular expression (regex) language, a domain specific language for string parsing and text search. Skilled `grep` users are able to rapidly construct an RE which retrieves the target entity with high probability whilst omitting irrelevant results. Assuming this entity exists on our filesystem, we can simply execute the following command to locate it:

```
$ grep -rE --include *.java "(class|interface) AbstractSingletonFactoryBean" .
```

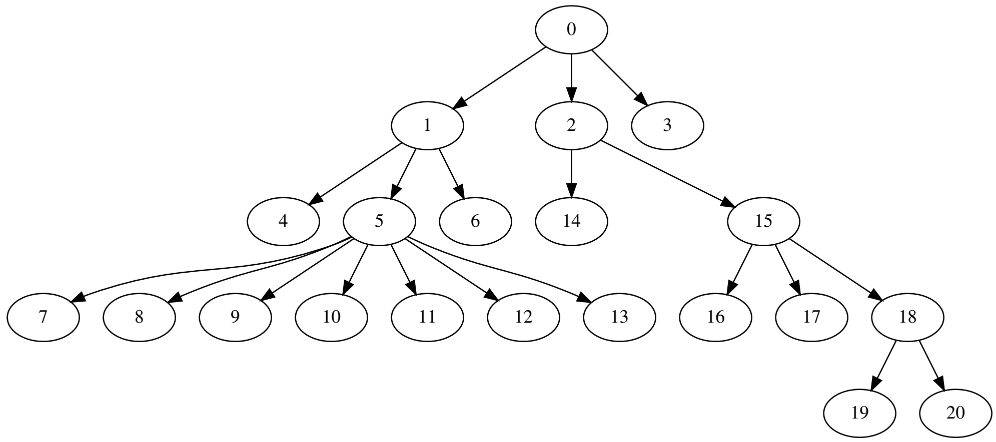
We hypothesize there exists a short regex which uniquely retrieves any named entity (assuming it exists) in a naturally occurring corpus of software artifacts. Given a named entity and its surrounding context, our goal is to synthesize an RE which selects only the link target, and as few other artifacts from the corpus as possible.

2 Dataset

Java, one of the most prolific programming languages on GitHub, is a statically typed language with a high volume of API documentation. Offering a variety of tools for source code [Parr, 2013, Hosseini and Brusilovsky, 2013, Kovalenko et al., 2019] and natural language [Manning et al., 2014, Grella and Cangialosi, 2018] parsing, it is a convenient language for both analysis and implementation. Our dataset consists of Java repositories on GitHub, and their accompanying developer documents. All projects in our dataset have a collection of source code files and multiple related repositories on GitHub.

We construct two datasets consisting of naturally-occurring links in developer docs, and a surrogate set of links constructed by matching lexical tokens in developer docs and source code. Our target is recovery of ground truth links in the test set and surrogate links in the lexical matching graph. We evaluate our approach on both D2D and C2C link retrieval, as well as precision and recall on the surrogate link relations.

Our data consists of two complementary datasets: abstract syntax trees (AST)s collected from Java source code and developer documentation. We use the astminer [Kovalenko et al., 2019] library to parse Java code, jsoup [Hedley, 2009] to parse HTML and Stanford's CoreNLP [Manning et al., 2014] library to parse dependency graphs from developer docs. Consider the following AST, parsed from the Eclipse Collections Java project:



The AST depicted above was generated by parsing the following code snippet:

```
public void lastKey_throws() {
    new ImmutableTreeMap<>(new TreeSortedMap<>()).lastKey();
}
```

Now consider the following dependency graph, taken from a Javadoc in the same project:

"The advantages to using this container over a `Map<K, Collection<V>>` is that all of the handling of the value collection can be done automatically. It also allows implementations to further specialize in how duplicate values will be handled. Value collections with list semantics would allow duplicate values for a key, while those implementing set semantics would not."

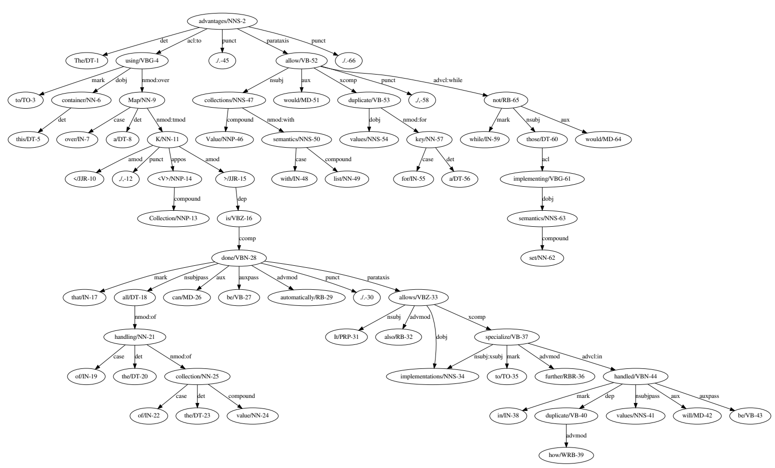


Figure 1: Dependency graph parsed from a JavaDoc comment, shown above.

Our goal is to connect these two graphs using a common model for source code and natural language. Absent any explicit `@link` or `@see` annotations, in order to relate these two graphs, we must somehow infer the shared semantic entities, which we can do for a subset using a simple lexical matching procedure, described in the following section.

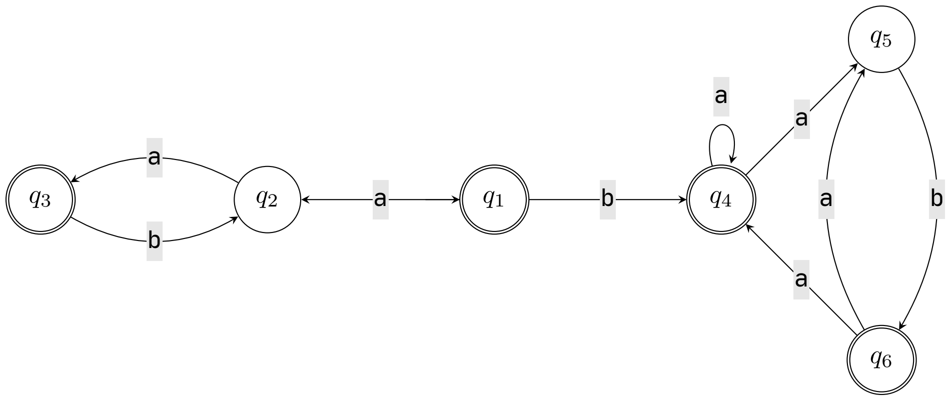


Figure 2: NFA corresponding to the RE `(a(ab)*)*(ba)*`.

3 Method

Let $\Sigma = \{$`A`, `a`, ..., `Z`, `z`, `0`, `1`, ..., `9`, `$`, `^`$\}$. Our language $J_<$ has the following productions:

$$\langle exp\rangle ::= \langle exp\rangle \circ \langle exp\rangle \;\mid\; \langle exp\rangle \texttt{|} \langle exp\rangle \;\mid\; \texttt{(}\langle exp\rangle\texttt{)} \;\mid\; \alpha \in \Sigma \;\mid\; \langle exp\rangle\texttt{*} \;\mid\; \texttt{!}\langle exp\rangle \;\mid\; \texttt{.} \tag{1}$$

An expression in $J_<$ is an RE, reducible to a non-deterministic finite automaton (NFA) using the Glushkov [1961] algorithm as shown in Figure 2 (independently discovered by McNaughton-Yamada-Thompson). NFA are reducible to both deterministic finite automata (DFA) using the powerset construction [Rabin and Scott, 1959] and regular expressions using Arden's Lemma [Arden, 1961]. Regular expressions can also be converted directly to DFA as described by Brzozowski [1964] and Berry and Sethi [1986].

Formally, an NFA is a 5-tuple $\langle Q, \Sigma, \Delta, q_0, F\rangle$, where Q is a finite set of states, Σ is the alphabet, $\Delta : Q \times (\Sigma \cup \{\epsilon\}) \rightarrow P(Q)$ is the transition function, $q_0 \in Q$ is the initial state and $F \subseteq Q$ are the terminal states. An NFA can be represented as a directed graph whose adjacency matrix is defined by the transition function, with edge labels representing symbols from the alphabet and binary node labels indicating whether the node is a terminal or nonterminal state.

We pose the problem as a few-shot generative modeling task, where the input is a node embedding consisting of the query text and local graph context, and the output is a regular expression retrieving the link target with high precision.

Let $\mathcal{D}$ be a document graph, constructed by semantically parsing the document's contents, and neighboring documents from the link graph. Let $\mathcal{T}$ be a code token, corresponding to a node in the document graph $\mathcal{D}$. Following prior work in few-shot meta-learning on graphs, we adapt the training procedure suggested by Bose et al. [2019] to pretrain a meta learner $\mathcal{M}$ on the following objective:

$$\mathcal{L}_G = \mathbb{E}_{q_\phi}[\log p(A^{train}|Z)] - KL[q_\phi(Z|X, A^{train})||p(z)] \tag{2}$$

Our synthesizer is trained on a node embedding from the local graph context, a semantic graph parsed from the parent document and neighboring documents in the link graph. Instead of directly performing link prediction, we train our generative model to output an NFA. We then compare precision and recall over the meta-test set.

Brzozowski [1964] defines the derivative of a language, $\mathcal{Q}$ with respect to a string `t` as follows:

$$\frac{\partial}{\partial \texttt{t}}\mathcal{Q} = \{s|\texttt{t}s \in \mathcal{Q}\} \tag{3}$$

If we interpret the operators | and $\circ$ from Equation 1 as $+$ and $\times$ respectively, we recover the standard rules from differential calculus:

$$\frac{\partial \mathtt{t}}{\partial \mathtt{t}} = \{\varnothing\} \tag{4}$$

$$\frac{\partial \lambda}{\partial \mathtt{t}} = \varnothing \tag{5}$$

$$\frac{\partial \mathcal{Q}*}{\partial \mathtt{t}} = \frac{\partial \mathcal{Q}}{\partial \mathtt{t}} \mathcal{Q}* \tag{6}$$

$$\frac{\partial \neg \mathcal{Q}}{\partial \mathtt{t}} = \neg \frac{\partial \mathcal{Q}}{\partial \mathtt{t}} \tag{7}$$

$$\frac{\partial \mathcal{Q} \circ \mathcal{S}}{\partial \mathtt{t}} = \frac{\partial \mathcal{Q}}{\partial \mathtt{t}} \circ \mathcal{S} \cup \frac{\partial \mathcal{S}}{\partial \mathtt{t}} \circ \mathcal{Q} \tag{8}$$

Given some RE corresponding to a regular language $\mathcal{R}$[1], this tells us how symbolic changes to the RE will change the language it recognizes. Suppose we have a corpus $\mathcal{C}$ and a set of target documents $\mathcal{D}^* \subseteq \mathcal{C}$. We define a loss on $\mathcal{R}$, $\mathcal{L}_\mathcal{R} : \langle \mathcal{R}, \mathcal{C}, \mathcal{D}^* \rangle \mapsto \mathbb{R}$ as follows:

$$\mathcal{L}(\mathcal{R}, \mathcal{C}, \mathcal{D}^*) = \mathbb{E}_{\mathcal{D}^* \sim \mathcal{C}} \frac{|\overbrace{\mathcal{R} \cap \mathcal{C}}^{P}| + |\overbrace{\mathcal{C} \setminus \mathcal{R}}^{N}|}{|\underbrace{\mathcal{R} \cap \mathcal{D}^*}_{TP}| + |\underbrace{\mathcal{C} \setminus \mathcal{D}^* \setminus \mathcal{R}}_{TN}|} \tag{9}$$

Given some set of strings we want to recognize, this ratio tells us how many documents from the corpus $\mathcal{C}$ does the regex accept $|\mathcal{R} \cap \mathcal{C}|$ and reject $|\mathcal{C} \setminus \mathcal{R}|$, over how many documents it should accept and reject. We need this ratio to be as low as possible for effective retrieval.

Suppose we have a function $\mathcal{G}_\theta : \mathbb{R}^n \times \mathbb{R}^{|\theta|} \rightarrow \mathcal{R}$, which takes a contextual entity embedding $\mathbb{R}^n$, a set of parameters θ, and returns a regular expression $\mathcal{R}$. Our goal is to minimize $\mathcal{L}(\mathcal{G}_\theta, \mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is a set of unlabeled context embeddings and $\mathbf{Y}$ is the ground truth link target. To compute $\nabla_\theta \mathcal{L}(\mathcal{G})$, we need $\nabla_\Sigma \mathcal{R}$, the vector of partial derivatives of $\mathcal{R}$ with respect to every symbol in the alphabet, Σ. Brzozowski [1964] shows us how to backpropogate loss through $\mathcal{R}$, into parameter space $\mathbb{R}^{|\theta|}$.

4 Preliminary Results

We compare precision on our benchmark using top-K rank retrieval with respect to various baselines. For instance, to compute the count-search baseline, we search our corpus for the hyperlink anchor text, and rank the resulting documents by frequency of the anchor text. Results are shown in Figure 3.

[1] Hereafter, we use $\mathcal{R}$ to denote the regex and the language it recognizes interchangeably.

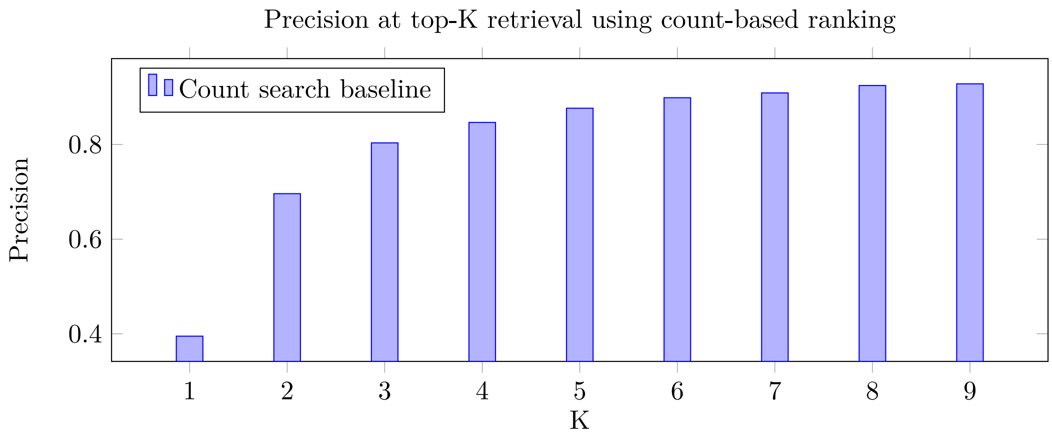


Figure 3: Results for count-based lexical matching baseline.

5 Discussion

Preliminary results indicate 40% of all links in our dataset point to the document in which the anchor text occurs most frequently in the corpus, and 93% of all links refer to documents where the ground truth link occurs in the top-10 results ranked by frequency of the anchor text. In future work, we will compare performance against a neural network using a simple word-embedding approach, trained to minimize cosine distance between the source document and target document's context, using a set of candidate links returned by the count-search baseline. Finally, we will compare our approach against the count-search baseline.

We will need to train a synthesizer using our loss function over the RE. . Furthermore, we will need to construct the semantic graph embedding over the surrounding context. This will require some additional consideration.

References

Miltiadis Allamanis, Marc Brockschmidt, and Mahmoud Khademi. Learning to represent programs with graphs. *arXiv preprint arXiv:1711.00740*, 2017. URL https://arxiv.org/pdf/1711.00740.pdf.

Dean N Arden. Delayed-logic and finite-state machines. In *2nd Annual Symposium on Switching Circuit Theory and Logical Design (SWCT 1961)*, pages 133–151. IEEE, 1961.

Gerard Berry and Ravi Sethi. From regular expressions to deterministic automata. *Theoretical computer science*, 48:117–126, 1986. URL https://core.ac.uk/reader/82374120.

Avishek Joey Bose, Ankit Jain, Piero Molino, and William L Hamilton. Meta-graph: Few shot link prediction via meta learning. *arXiv preprint arXiv:1912.09867*, 2019. URL https://arxiv.org/pdf/1912.09867.pdf.

Janusz A Brzozowski. Derivatives of regular expressions. *Journal of the ACM (JACM)*, 11 (4):481–494, 1964. URL http://maveric.uwaterloo.ca/reports/1964_JACM_Brzozowski.pdf.

Octavian-Eugen Ganea, Marina Ganea, Aurelien Lucchi, Carsten Eickhoff, and Thomas Hofmann. Probabilistic bag-of-hyperlinks model for entity linking. In *Proceedings of the 25th International Conference on World Wide Web*, pages 927–938, 2016. URL https://arxiv.org/pdf/1509.02301.pdf.

Victor Mikhaylovich Glushkov. The abstract theory of automata. *Russian Mathematical Surveys*, 16(5):1, 1961.

Matteo Grella and Simone Cangialosi. Non-projective dependency parsing via latent heads representation LHR. *arXiv preprint arXiv:1802.02116*, 2018. URL https://arxiv.org/pdf/1802.02116.pdf.

Xiaodong Gu, Hongyu Zhang, Dongmei Zhang, and Sunghun Kim. Deep API learning. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, pages 631–642, 2016. URL https://arxiv.org/pdf/1605.08535.pdf.

Xiaodong Gu, Hongyu Zhang, and Sunghun Kim. Deep code search. In *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*, pages 933–944. IEEE, 2018. URL https://guxd.github.io/papers/deepcs.pdf.

Jonathan Hedley. jsoup: Java HTML parser. 2009. URL https://jsoup.org.

Jordan Henkel, Shuvendu K Lahiri, Ben Liblit, and Thomas Reps. Code vectors: Understanding programs through embedded abstracted symbolic traces. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 163–174, 2018. URL https://arxiv.org/pdf/1803.06686.pdf.

Roya Hosseini and Peter Brusilovsky. JavaParser: A fine-grain concept indexing tool for Java problems. In *CEUR Workshop Proceedings*, volume 1009, pages 60–63. University of Pittsburgh, 2013.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. Mapping language to code in programmatic context. *arXiv preprint arXiv:1808.09588*, 2018. URL https://arxiv.org/pdf/1808.09588.pdf.

Vladimir Kovalenko, Egor Bogomolov, Timofey Bryksin, and Alberto Bacchelli. PathMiner: a library for mining of path-based representations of code. In *Proceedings of the 16th International Conference on Mining Software Repositories*, pages 13–17. IEEE Press, 2019. URL https://github.com/JetBrains-Research/astminer.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016. URL https://arxiv.org/pdf/1603.01360.pdf.

Jian Li, Yue Wang, Michael R Lyu, and Irwin King. Code completion with neural attention and pointer networks. *arXiv preprint arXiv:1711.09573*, 2017. URL https://www.ijcai.org/Proceedings/2018/0578.pdf.

Bohong Liu, Tao Wang, Xunhui Zhang, Qiang Fan, Gang Yin, and Jinsheng Deng. A neural-network based code summarization approach by using source code and its call dependencies. In *Proceedings of the 11th Asia-Pacific Symposium on Internetware*, Internetware '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450377010. doi: 10.1145/3361242.3362774. URL https://doi.org/10.1145/3361242.3362774.

Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. Stack-pointer networks for dependency parsing. *arXiv preprint arXiv:1805.01087*, 2018. URL https://arxiv.org/pdf/1805.01087.pdf.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford coreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014. URL https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf.

Sheena Panthaplackel, Milos Gligoric, Raymond J. Mooney, and Junyi Jessy Li. Associating natural language comment and source code entities. In *AAAI*, 2020. URL https://arxiv.org/pdf/1912.06728.pdf.

Terence Parr. *The definitive ANTLR 4 reference*. Pragmatic Bookshelf, 2013.

Michael O Rabin and Dana Scott. Finite automata and their decision problems. *IBM journal of research and development*, 3(2):114–125, 1959.

Siva Reddy, Oscar Täckström, Michael Collins, Tom Kwiatkowski, Dipanjan Das, Mark Steedman, and Mirella Lapata. Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics*, 4: 127–140, 2016. URL https://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00088.

Martin P Robillard and Yam B Chhetri. Recommending reference API documentation. *Empirical Software Engineering*, 20(6):1558–1586, 2015. URL https://www.cs.mcgill.ca/~martin/papers/cr2014a.pdf.

Martin P Robillard, Andrian Marcus, Christoph Treude, Gabriele Bavota, Oscar Chaparro, Neil Ernst, Marco Aurélio Gerosa, Michael Godfrey, Michele Lanza, Mario Linares-Vásquez, et al. On-demand developer documentation. In *2017 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 479–483. IEEE, 2017.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12(Sep):2539–2561, 2011.

Xujie Si, Hanjun Dai, Mukund Raghothaman, Mayur Naik, and Le Song. Learning loop invariants for program verification. In *Advances in Neural Information Processing Systems*, pages 7751–7762, 2018. URL https://papers.nips.cc/paper/8001-learning-loop-invariants-for-program-verification.pdf.

Alexandre Terrasa, Jean Privat, and Guy Tremblay. Using natural language processing for documentation assist. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015a. URL `https://arxiv.org/pdf/1511.06391.pdf`.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in neural information processing systems*, pages 2692–2700, 2015b. URL `https://arxiv.org/pdf/1506.03134.pdf`.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016. URL `https://www.cs.cmu.edu/~./hovy/papers/16HLT-hierarchical-attention-networks.pdf`.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, pages 5165–5175, 2018. URL `https://papers.nips.cc/paper/7763-link-prediction-based-on-graph-neural-networks.pdf`.