ICON: *I*nferring Temporal *Con*straints from Natural Language API Descriptions

Rahul Pandita[1], Kunal Taneja[2], Tao Xie[3], Laurie Williams[1], Teresa Tung[2]
[1]Department of Computer Science, North Carolina State University, Raleigh, NC, USA
[2]Accenture Technology Labs, San Jose, CA, USA
[3]Department of Computer Science, University of Illinois, Urbana-Champaign, IL, USA
rpandit@ncsu.edu, k.a.taneja@accenture.com, taoxie@illinois.edu, williams@csc.ncsu.edu, teresa.tung@accenture.com

ABSTRACT

Temporal constraints of an Application Programming Interface (API) are the allowed sequences of invocations of methods from the API. These constraints govern the secure and robust operation of client software using the API. Typically, these constraints are described in natural language text of API documents and thus cannot be used by existing verification tools which typically accept only formal constraints. Because manually writing temporal constraints based on API documents can be prohibitively time consuming and error prone, we propose ICON: a Natural Language Processing (NLP) based approach to automatically infer temporal constraints. In particular, our approach includes novel techniques to reduce the number of lexical tokens as a way to make API documents amenable to existing NLP techniques (that are currently designed to work on well written news articles). Our approach also includes a novel technique of identifying method references in the natural language text by building domain-dictionaries systematically from API documents and generic English dictionaries. To evaluate our approach, we apply ICON to infer temporal constraints from commonly used package `java.io` in the JDK API and from the `Amazon S3 REST API`. Our results show that ICON effectively identifies constraint sentences (from over 3900 API sentences) with the average precision, recall, and F-score of 65.0%, 72.2%, and 68.4%, respectively. Furthermore, ICON also achieves an accuracy of 70% in inferring 63 temporal constraints from these sentences.

Categories and Subject Descriptors

D.2.1 [**Software Engineering**]: Requirements/Specifications; F.3.1 [**LOGICS AND MEANINGS OF PROGRAMS**]: Specifying and Verifying and Reasoning about Programs—*Specification techniques*

General Terms

Specifications

Keywords

Temporal Specifications, NLP, API Documents

FSE '14 Hong Kong

1. INTRODUCTION

Temporal constraints [2] of an Application Programming Interface (API) are the allowed sequences of invocations of methods within the API. These constraints govern the secure and robust operation of client software using the API. If these constraints are formal (machine-readable such as linear temporal logic), they can be used as inputs to the formal analysis tools such as model checker and runtime verifiers in detecting the violations of these temporal constraints as defects [19].

Typically, these constraints are described in natural language text of API documents. Such documents are provided to client-code developers through an online access, or are shipped with the API code. For a method under consideration, an API document may describe both the constraints on that method parameters as well as the temporal constraints in terms of methods that must be invoked pre/post that method. We observed that an considerable portion (roughly 12%) of the sentences (that describe some sort of constraints) in `Amazon S3 REST API` documentations describe temporal constraints. However, existing verification tools which typically accept only formal constraints cannot use the natural language constraints from API documents.

One way of addressing the issue, is to manually convert the natural language API description into formal constraints. However, manually writing formal constraints based on natural language text in API documents can be prohibitively time consuming and error prone [27,39]. For instance, Wu et al. [39] report that it took one of the authors hours to browse the documentation of one method of a web service API, even before they attempted to formalize the constraints on the method. To reduce the manual effort, we propose ICON: a Natural Language Processing (NLP) based approach to automatically infer formal temporal constraints.

Although, existing approaches focus on inferring method pre-post conditions (in terms of parameter constraints) using either program analysis techniques [6,13–16] or NLP [25,39], temporal constraints are beyond these parameter constraints. Temporal constraints, in general focus on rules related to orchestration of methods within an API rather than focusing on the requirements on the input parameters of these methods.

Apart from method specifications, Zhong et al. [42] leverage machine learning and type information to infer constraints on resources based on phases of resource usage: creating, manipulating, and releasing. However, temporal constraints are not limited to resource usage. Furthermore, their approach fails to make the finer grained distinction on the ordering of the methods within a phase. For the description mentioned in previously, assuming that the method calls are associated with resource "part", both the methods referenced are logically associated with the creation phase of resource. Zhong et al. [42] approach fails to make a distinction in

ordering of such methods. In contrast our proposed approach leverages natural language processing to infer generic temporal constraints (not limited to resource related constraints) and also improves upon existing work for making finer grained distinction between ordering of methods within a phase of a resource.

However, inferring temporal constraints from the natural language text is challenging. There are existing challenges in NLP for software engineering domain namely *ambiguity*, *programming keywords*, and *semantic equivalence* identified in previous work [25]. In addition to these challenges, we have an added challenge of implicit method referencing: identifying the method referenced in the natural language text to infer temporal constraint. Recall the description sentence "You must initiate a multipart upload before you can upload any part." In this sentence, phrases "multipart upload" and "upload any part" refer to individual methods in the API. Identifying these phrases as method invocation instances requires domain dictionaries. Ad-hoc construction of such dictionaries is prohibitively time and resources intensive. To address this challenge, we propose to build domain-dictionaries systematically from API documents and generic English dictionaries.

In summary, the proposed work leverages natural language description of API's to infer temporal constraints of method invocations. As the proposed work analyzes API documents in natural language, it can be reused independent of the programming language of the API library. Additionally, our approach complements existing mining based approaches [5, 36, 38, 41] that partially address the problem by mining for common usage patterns among client software that use the API. The proposed work in general, makes the following contributions:

- An NLP based approach that infers temporal constraints of method invocations. To the best of our knowledge, our approach is the first one to apply NLP for the goal of inferring temporal constraints from API documents.
- A prototype implementation of our approach based on extending the Stanford Parser [17], which is a natural language parser to derive the grammatical structure of sentences. An open source implementation of the prototype is publicly available on our project website[1].
- An evaluation of proposed approach on commonly used package `java.io` from JDK API and `Amazon S3 REST API`.

The rest of the paper is organized as follows. Section 2 presents real world examples that motivate our approach. Section 3 discusses related work in this area. Section 4 presents the background on NLP techniques used in this work. Section 5 presents our approach. Section 6 presents evaluation of our approach. Section 7 presents a brief discussion and future work. Finally, Section 8 concludes.

2. MOTIVATING EXAMPLE

We next present a real world example to motivate our approach. In particular, through the example, we demonstrate that developers often ignore the temporal constraints of an API described in the documentation. We suspect the reason for this phenomena is that the documentation is often verbose and the information is distributed across various pages. For instance, the PDF version of the documentation for `Amazon S3 REST API`[2] spans 278 pages. Developers often may not have time (and/or patience) to go through all the documentation and may overlook some temporal constraints of the API, resulting in defective client applications that invoke API methods in sequences prohibited by documentation.

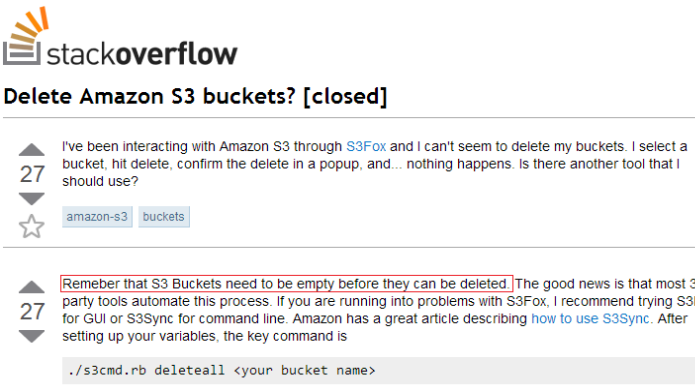

Figure 1: The Query posted on Stack Overflow forum regrading Amazon S3 REST API

Consider the question asked in *Stack Overflow* [3] as shown in Figure 1. Stack Overflow is an online question and answer forum for professional and enthusiast programmers. The query is about the delete functionality of a third-party software `S3Fox` to interact with `Amazon S3 REST API`. The inquisitor complains about an issue in the delete bucket functionality of the `S3Fox`. The `S3Fox` developers overlooked the constraints in `Amazon S3 REST API` developer documentation, causing the issue. The API document pertaining to the delete bucket functionality states that before deleting the bucket, the objects in the buckets must be deleted. *"All objects (including all objects versions and Delete Markers) in the bucket must be deleted before the bucket itself can be deleted"*. The issue was fixed. However, one of the forum responses contained a recommendation for the inquisitor to switch to another product. Customer dissatisfaction, such as was caused by this issue with the delete bucket functionality, can lead to a lose in revenue.

The preceding issue can be easily detected using formal analysis tools. For instance, a specification rule (temporal constraint) can be added to a static checker to verify the presence of a call to delete object functionality before the call to delete bucket functionality. In next section, we briefly discuss the related work pertinent to our proposed approach.

3. RELATED WORK

Our proposed approach touches a few research areas such as software verification, NLP in Software Engineering (SE), and document augmentation. We next discuss the relevant work pertinent to our proposed approach in these areas.

Formal Specification: Contracts have become a well-known mechanism for formally specifying functional behavior of the software. Contracts specify the behavior in terms of conditions that must hold before/after and during the execution of the method. A significant amount of work has been done in automated inference of contracts. Existing approaches use program analysis [8, 22, 37] to automatically infer contracts. However, recent studies [13, 26] demonstrate that a combination of developer-written and automatically extracted contracts is the most effective approach for formally specifying the constraints on an API. Since our approach infers temporal constraint from API documents, we believe our approach

[1] `https://sites.google.com/site/temporalspec`
[2] `http://awsdocs.s3.amazonaws.com/S3/latest/s3-api.pdf`
[3] `http://stackoverflow.com/`

can work in conjunction with existing approaches to infer a compressive set of specifications.

Furthermore, another set of approaches exist that infer code-contract-like specifications (such as behavioral model, algebraic specifications, and exception specifications) either dynamically [14–16] or statically [6, 13] from source code and binaries. In contrast, the approach presented in this report infers specifications from the natural language text in API documents, thus complementing these existing approaches when the source code or binaries of the API library is not available.

NLP in SE: NLP techniques are being increasingly applied in the SE domain. Tan et al. [32] were the first to apply Machine Learning (ML) and NLP on code comments to detect mismatches between the comments and the implementation. They rely on predefined rule templates targeted towards threading and lock related comments, thus limiting their scope both in terms of application area as well as language used in the comments. Furthermore, the constraints inferred by their approach are the restrictions imposed by the developer on the client code. In comparison, the temporal constraints inferred by our approach are the restriction imposed by the API library being used by the client code. Furthermore, approach presented in this report relies on generic natural language based templates thus relaxing the restriction on the style of the language used to describe specifications.

Zhong et al. [42] leverage machine learning and type information to infer constraints on resources based on phases of resource usage: creating, manipulating, and releasing. However, temporal constraints are not limited to resource usage thus Furthermore, the specifications inferred by their approach infers implicit specifications across multiple method descriptions, whereas the temporal constraints inferred by our approach infers explicit ordering information described in individual method descriptions. Furthermore, the performance of their approach is dependent on the quality of the training sets used for ML. In contrast, approach presented in this report is independent of such training set and thus can be easily extended to target respective problems addressed by them.

Xiao et al. [40] and Slankas et al. [31] use shallow parsing techniques to infer Access Control Policy (ACP) rules from natural language text in use cases. The use of shallow parsing techniques works satisfactorily on natural language texts in use cases, owing to well-formed structure of sentences in use case descriptions. In contrast, often the sentences in API documents are not well-formed. Additionally, their approaches do not deal with programming keywords or identifiers, which are often mixed within the method descriptions in API documents.

Most closely related work to the approach presented here is our previous work on inferring parameter constraints from method descriptions in the API documents [25]. Our proposed approach differs from the previous approach as follows. Our proposed approach addresses the problem of inferring temporal constraint, whereas the previous approach infers parameter constraints. Our proposed approach is a significant extension to the infrastructure used in the previous work in following areas. First, the proposed approach introduces hybrid shallow parsing that relies both on parts-of-speech tags as well as Stanford-typed dependencies to construct intermediate representation. Second, the proposed approach introduces a new heuristics namely frequent phrase reduction to deal with the domain specific and ill-formed sentences in API documents. Finally, the proposed approach leverages the concept of semantic graphs constructed from class and method names in API to automatically infer the implicit method references in a sentence.

Augmented Documentation: Improving the documentation related to a software API [11, 33] is another related field of research. Dekel and Herbsleb [11], were the first to create a tool namely eMoose, an Eclipse [4] based plug-in that allowed developers to create directives (way of marking the specification sentences) in the default API documentation. These directives are highlighted whenever they are displayed in the Eclipse environment. Lee et al. [19] improved upon their work by providing a formalism to the directives proposed by Dekel et al. [11], thus allowing tool-based verification. However, a developer has to manually annotate such directives. In contrast, our proposed approach both identifies the sentences pertaining to temporal constraints and infers the temporal constraints automatically.

In next section, we briefly introduce the NLP techniques used by our approach.

4. BACKGROUND

Natural language is well suited for human communication, but converting natural language into unambiguous specifications that can be processed and understood by computers is difficult. However, research advances [9, 10, 17, 18] have increased the accuracy of existing NLP techniques to annotate the grammatical structure of a sentence. These advances in NLP have inspired researchers/ practitioners [24, 25, 31, 35, 40] to adapt/apply NLP techniques to solve problems in SE domain.

In particular, this work proposes novel techniques on top of previously proposed techniques [24, 25] to demonstrate the effectiveness of applying NLP on API documents. We next briefly introduce the techniques used in this work that have been grouped into broad categories. We first introduce the core NLP techniques used in this work. We then introduce the SE specific NLP techniques proposed in previous work [24, 25] that are used in this work.

4.1 Core NLP techniques

Parts Of Speech (POS) tagging [17, 18]. Also known as *'word tagging'*, *'grammatical tagging'* and *'word-sense disambiguation'*, these techniques aim to identify the part of speech (such as noun, verbs, etc.), a particular word in a sentence belongs to. The most commonly used technique is to train a classification parser over a previously known data set. Current state of the art approaches have demonstrated to be effective in classifying POS tags for well written news articles.

Phrase and Clause Parsing. Also known as chunking, this technique divides a sentence into a constituent set of words (or phrases) that logically belong together (such as a Noun Phrase and Verb Phrase). Chunking thus further enhances the syntax of a sentence on top of POS tagging. Current state-of-the-art approaches can effectively classify phrases and clauses over well written news articles.

Typed Dependencies [9, 10]. The Stanford typed dependencies representation is designed to provide a simple description of grammatical relationships directed towards non-linguistics experts to perform NLP related tasks. It provides a hierarchical structure for the dependencies with precise definitions of what each dependency represents, thus facilitating machine based manipulation of natural language text.

4.2 SE specific NLP techniques

Noun Boosting [25]. Accurate annotation of POS tags in a sentence is fundamental to effectiveness of any advanced NLP technique. However, as mentioned previously, POS tagging works satisfactorily on well written news articles which does not necessary entail that the tagging works satisfactorily on domain specific text

[4] `http://www.eclipse.org/`

as well. Thus, noun boosting is a necessary precursor to application of POS tagging on domain specific text. In particular, with respect to API documents certain words have a different semantic meaning, in contrast to general linguistics that causes incorrect annotation of POS tags.

Consider the word `POST` for instance. The online Oxford dictionary[5] has eight different definition of word `POST`, and none of them describes `POST` as an HTTP method[6] supported by REST API. Thus existing POS tagging techniques fail to accurately annotate the POS tags of the sentences involving word `POST`.

Noun Boosting identifies such words from the sentences based on a domain-specific dictionaries, and annotates them appropriately. The annotation assists the POS tagger to accurately annotate the POS tags of the words thus in turn increasing accuracy of advanced NLP techniques such as chunking and typed dependency annotation.

Lexical Token Reduction [24]. These are a group of generic preprocessing heuristics to further improve the accuracy of core NLP techniques. The accuracy of core NLP techniques is inversely proportional to the number of lexical tokens in a sentence. Thus, the reduction in the number of lexical tokens greatly increases the accuracy of core NLP techniques. In particular, following heuristics have been used in previous work to achieve the desired reduction of lexical tokens:

- **Period Handling**. Besides marking the end of a sentence in simplistic English, the character period ('.') has other legal usages as well such as decimal representation (periods between numbers). Although legal, such usage hinder detection of sentence boundaries, thus causing core NLP techniques to return incorrect or imprecise results. The text is pre-processed by annotating these usages for accurate detection of sentence boundaries.

- **Named Entity Handling**. Sometimes a sequence of words correspond to the name of entities that have a specific meaning collectively. For instance, consider the phrases *"Amazon S3", "Amazon simple storage service"*, which are the names of the service. Further resolution of these phrases using grammatical syntax is unnecessary and would not bring forth any semantic value. Also these phrases contribute to length of a sentence that in turn negatively affects the accuracy of core NLP techniques. This heuristic annotates the phrase representing the name of the entities as a single lexical token.

- **Abbreviation Handling**. Natural-language sentences often consist of abbreviations mixed with text. This phenomenon can result in subsequent components to incorrectly parse a sentence. This heuristic finds such instances and annotates them as a single lexical unit. For example, text followed by abbreviations such as *"Access Control Lists (ACL)"* is treated as single lexical unit. Detecting such abbreviations is achieved by using the common structure of abbreviations and encoding such structures into regular expressions. Typically, regular expressions provide a reasonable approximation for handling abbreviations.

Intermediate-Representation Generation [24]. This technique accepts the syntax-annotated sentences and builds a First-Order-Logic(FOL) representation of the sentence. Earlier researches have shown the adequacy using FOL for NLP related analysis tasks [25, 29, 30]. In particular, WHYPER [24] demonstrates the effectiveness of this technique, by constructing an intermediate representation generator based on shallow parsing [3] techniques. The shallow parser itself is implemented as a sequence of cascading finite state machines based on the function of Stanford-typed dependencies [9, 10, 17, 18].

In next section we describe our proposed generic approach to infer constraints from API documents.

[5] `http://oxforddictionaries.com/us/definition/american_english/post?q=POST`

[6] In HTTP vocabulary `POST` means: *"Creates a new entry in the collection. The new entry's URI is assigned automatically and is usually returned by the operation"*

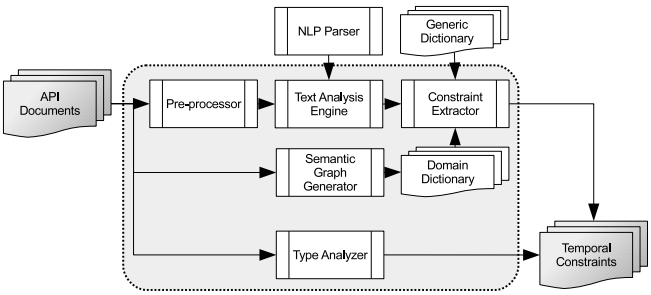

Figure 2: Overview of ICON approach

5. APPROACH OVERVIEW

We next present our approach for inferring temporal constraints from the method descriptions in API Documents. Figure 2 gives an overview of our approach. Our approach consists of five major components: a preprocessor, a text-analysis engine, a semantic graph generator, constraint extractor, and a type analyzer.

The preprocessor accepts API documents and preprocesses the sentences in the method description, such as annotating sentence boundaries and reducing lexical tokens. The text-analysis engine accepts the pre-processed sentences and annotates them using an NLP parser. The text-analysis engine further transforms the annotated sentences into the first-order-logic (FOL) representation. The semantic graph generator accepts the API documents and generates the semantic graphs that are leveraged by constraint extractor component. The type analyzer component infers temporal constraints encoded in the type system of a language by analyzing the API methods parameter and return types. Finally, the constraint extractor leverages the semantic graphs to infer temporal constraints from the FOL representation of a sentence. We next describe each component in detail.

5.1 Preprocessor

The preprocessor accepts the API documents and extracts method descriptions from it. In particular, the preprocessor extracts the following fields within method descriptions: 1) *Summary of the API method*, 2) *Summary and type information of parameters of the API method*, 3) *Summary and type information of return values of the method*, and 4) *Summary and type information of exceptions thrown by the methods*.

This step is required to extract the desired descriptive text from the API documents. In particular, different API documents may have different styles of presenting information to developers. This difference in style may also include the difference in the level of detail presented to developers. Our approach thus relies on only basic fields that are trivially available for API methods across different presentation styles.

After extracting desired information, the natural language text is further preprocessed to be analyzed by subsequent components.

The preprocessing steps are required to increase the accuracy of core NLP techniques (described in Section 4.1) that are used in the subsequent phases of ICON approach. In particular, the preprocessor first employs the noun boosting followed by heuristics listed under lexical token reduction, as introduced in Section 4.2.

Although the previous techniques and heuristics significantly lower the number of lexical tokens in a sentence, some sentences may still contain a considerable number of lexical tokens to overwhelm the POS tagger. To address this issue, we propose a new heuristic (*'Frequent Phrases Reduction'*) to further reduce the number of lexical tokens in a sentence by annotating frequent phrases as a single lexical unit.

In particular, we use n-gram [4] based approach to achieve this reduction. In the fields of computational linguistics, an n-gram is a contiguous sequence of n words from a given text. In statistical NLP and information theory, n-gram has been shown to be effective in computing the probability of occurrence of next word given a sequence of words. However, for our approach we use n-grams that always occur together in a given body of text.

To achieve *frequent phrases reduction*, we first calculate the most frequently occurring n-grams in the text body. In particular, we are interested in the n-grams of length four or greater to achieve a reasonable reduction. We chose four as a threshold for n-grams because we observed that bigrams (size two) and trigrams (size three) frequently resulted in change of semantics of the sentence in comparison to n-grams of size four or grater. We then prune the list of n-grams based on a subsumption. We consider an n-gram of length k (n_k) to subsume n-gram of length k-1 (n_{k-1}) iff n_{k-1} is a substring of (n_k) and the frequency of occurrence of n_{k-1} equals frequency of occurrence of n_k. Finally, we rank the list of n-grams based on the frequency of their occurrence in the text, and select top-k n-grams for reduction. For instance, *Amazon Simple Storage Service*, *an I/O Error Occurs*, and *end of the stream* are the examples of such n-grams detected by our approach.

Currently our prototype implementation works with online `Amazon S3 REST API` developer documentation and JDK API. However, almost all of the developer documents are provided online as structured webpages. Thus, current prototype implementation of preprocessor can be easily extended to extract the desired information from any API developer documents.

Additionally, in current implementation we have manually built the dictionaries for preprocessing using the glossary of terms collected from the websites pertaining to REST and Java API. We further leveraged the HTML style information in `Amazon S3 REST API` developer documentation to look for words that were highlighted in code like format. We further leveraged WordNet to maintain a static lookup table of shorthand words to aid named entity handling and abbreviation handling.

Finally, to achieve n-gram reduction we used Apache Lucene [20]. Apache Lucene is a high-performance, full-featured text-search-engine library written entirely in Java, facilitating scalable cross-platform full-text search.

5.2 NLP Parser

The NLP parser accepts the pre-processed documents and annotates every sentence within each document using core NLP techniques described in Section 4.1. From an implementation perspective, we chose the Stanford parser [21]. However, this component can be implemented using any other existing NLP libraries or approaches. In particular, we annotate each sentence with POS tags, named-entity annotations and Stanford-typed dependencies. For more details on these techniques and their application, please refer to [9, 10, 24, 25, 35].

```
-> deleted-VBN (root)
  -> objects-NNS (nsubjpass)
    -> All-DT (det)
    -> including-VBG (dep)
      -> object versions-NNS (pobj)
        -> all-DT (det)
        -> Delete Markers-NNS (conj_and)
      -> Delete Markers-NNS (pobj)
    -> bucket-NN (prep_in)
      -> the-DT (det)
  -> must-MD (aux)
  -> be-VB (auxpass)
  -> bucket-NN (prep_before)
    -> the-DT (det)
    -> deleted-VBN (rcmod)
      -> itself-PRP (nsubjpass)
      -> can-MD (aux)
      -> be-VB (auxpass)
```

Figure 3: Sentence annotated with Stanford dependencies

Next we use an example to illustrate the annotations added by the NLP Parser. Consider the sentence from the example section ***'All objects (including all object versions and Delete Markers) in the bucket must be deleted before the bucket itself can be deleted."***. Figure 3 shows the sentence annotated with Stanford-typed dependencies. The words in red are the names of dependencies connecting the actual words of the sentence (in black). Each word is followed by the Part-Of-Speech (POS) tag of the word (in green). For more details on Stanford-typed dependencies and POS tags, please refer to [9, 10].

5.3 Text Analysis Engine

The text analysis engine component accepts the annotated documents and creates an intermediate representation of each sentence. We define our representation as a tree structure that is essentially a FOL expression. Research literature provides evidence of the adequacy of using FOL for NLP related analysis tasks [24, 25, 29, 30].

In our representation, every node in the tree except for the leaf nodes is a predicate node. The leaf nodes represent the entities. The children of the predicate nodes are the participating entities in the relationship represented by the predicate. The first or the only child of a predicate node is the governing entity and the second child is the dependent entity. Together the governing entity, predicate and the dependent entity node form a tuple.

As described in Section 4.2 the intermediate representation generation technique is based on the principle of shallow parsing [3]. In particular, the intermediate-representation technique is implemented as a function of Stanford-typed dependencies [9, 10, 18], to leverage the semantic information encoded in Stanford-typed dependencies.

However, we observed that such implementation is overwhelmed by complex sentences. This limitation mandates the use of additional novel technique of *'Frequent Phrases Reduction'* in preprocessing phase. We further improve the accuracy of intermediate-representation generation by proposing a hybrid approach, i.e. taking into consideration both the POS tags as well as Stanford-typed dependencies. The POS tags which annotate the syntactical structure of a sentence are used to further simplify the constituent elements in a sentence. We then use the Stanford-typed dependencies that annotate the grammatical relationships between words to construct our FOL representation. Thus, the intermediate representation generator used in this work is a two phase process as opposed to previous work [24, 25]. We next describe these two phases:

POS Tags: We first parse a sentence based on the function of POS tags. In particular, we use semantic templates to logically break a sentences into smaller constituent sentences. For instance,

consider the sentence which are then accurately annotated by the underlying NLP Parser:

"All objects (including all object versions and Delete Markers) in the bucket must be deleted before the bucket itself can be deleted.".

The Stanford parser finds it difficult to annotate accurately the Stanford-typed dependencies of the sentence because of presence of different clauses acting on different subject-object pairs. As shown in figure 3 the word including is annotated with Stanford-typed dependencies "dep" that is a catch all dependency. We thus break down the sentence into two smaller tractable sentences:

"All objects in the bucket must be deleted before the bucket itself can be deleted."
"All objects including all object versions and Delete Markers."

Table 1 shows a list the semantic templates used in this phase. Column "Template" describes conditions where the template is applicable and Column "Summary" describes the action taken by our shallow parser when the template is applicable. All of these semantic templates are publicly available on our project website[7]. With respect to the previous example the template 3 *(A noun phrase followed by another noun/pronoun/verb phrase in brackets)* is applicable. Thus our shallow parser breaks the sentence into two individual sentences.

Stanford-typed Dependencies: This phase is equivalent to the intermediate-representation technique described in Section 4.2.

5.4 Constraint Extractor

This component accepts the FOL representation of the sentence from the previous component, then extracts the temporal constraints if present in a sentence. Constraint Extractor then classifies the sentence as a constraint sentence (containing temporal constraint) candidate based on following ordered set of rules:

1. The sentence is not from parameter summary or return variable summary. Typically such sentences describe pre-post conditions as opposed to temporal constraints this approach addresses.

2. The sentences contains modal modifiers such as "*can, could, may, must, should*" expressing necessity. Typically, presence of such modal modifier is a strong indicator of presence of constraints imposed by an API developer

3. If the sentence does not contain modal modifiers described previously, sentence must contain temporal modifier relationship, identified by Stanford-typed dependency parser. Typically, presence of temporal modifier is an indicator of presence of temporal information.

4. If rules 2 and 3 don't apply then the sentences should be a conditional sentence, identified by the presence of keywords such as "*if*" and "*whether*".

Once a candidate sentence is identified, this component selects an semantic graph. In particular, the semantic graph of the API class to which the candidate sentence belongs to selected. Since there is no concept of classes in REST API, semantic graphs of all resources described in the API are selected. A semantic graph constitutes the keyword representation of the classes and the corresponding applicable actions. Figure 4 shows a graph for `Object` resource in `Amazon S3 REST API`. The phrases in rounded rectangle are the actions applicable on `Object` resource. Section 5.5 further describes how these graphs are generated.

[7] `https://sites.google.com/site/temporalspec`

Algorithm 1 Action_Extractor

Input: K_Graph g, FOL_rep rep
Output: String $action$

```
1:  String action = φ
2:  List r_name_list = g.resource_Names
3:  FOL_rep r' = rep.findLeafContaining(r_nam_list)
4:  List actionList = g.actionList
5:  while (r'.hasParent) do
6:      if actionList.contains(r'.parent.predicate) then
7:          action = actionList.matching(r'.parent.predicate)
8:          break
9:      else
10:         if actionList.contains(r'.leftSibling.predicate) then
11:             action = actionList.matching(r'.leftSibling.predicate)
12:             break
13:         end if
14:     end if
15:     r' = r'.parent
16: end while
17: return action
```

Constraint extractor then uses the semantic graph to determine whether a candidate sentence is a constraint sentence and if so extract the action that should be performed prior to the method the sentence belongs to. Algorithm 1 describes this action extraction process.

Our algorithm systematically explores the FOL representation of the candidate sentence to determine if a sentence describes a temporal constraint. First, our algorithm attempts to locate the occurrence of class name or its synonym within the leaf nodes of the FOL representation of the sentence (Line 3). The method findLeafContaining(r_name_list) explores the FOL representation to find a leaf node that contains either the class name or one of its synonyms. In particular, we use WordNet [12] and Lemmatisation to deal with synonyms of a word in question to find appropriate matches. Once a leaf node is found, we systematically traverse the tree from the leaf node to the root, matching all parent predicates as well as immediate child predicates [Lines 5-16].

Our algorithm matches each of the traversed predicate with the actions associated with the class defined in semantic graph. Similar to matching entities, we also employ WordNet and Lemmatisation to deal with synonyms to find appropriate matches. If a match is found, then the matching action name is returned.

5.5 Semantic-Graph Generator

A key way of identifying reference to a method within the API in our proposed approach is the employment of a semantic graph of an API. In particular, we propose to initially infer such graphs from API documents. Manually creating a semantic graph is prohibitively time consuming and may be error prone. We thus employ a systematic methodology (proposed previously in [24]) to infer such semantic graphs from API documents that can potentially be automated. We first consider the name of the class for the API document in question. We then find the synonyms terms used refer to the class in question. The synonym terms are listed as by breaking down the camel-case notation in the class name. This list is further augmented by listing the name of the parent classes and implemented interfaces if any.

We then systematically inspect the member methods to identify actions applicable to the objects represented by the class. From the name of a public method (describing a possible action on the object), we extract verb phrases. The verb phrases are used as the associated actions applicable on the object. In case of REST API we first identified the resources and then listed REST actions on those resources as applicable actions. Figure 4 shows the graph for `Object` resource in REST API. The phrases in rounded rectangle

Table 1: Semantic Templates

S No.	Template	Summary
1.	Two sentences joined by a conjunction	Sentence is broken down into two individual sentences with the conjunction term serving as the connector between two.
2.	Two sentences joined by a ","	Sentence is broken down to individual independent sentences
3.	A noun phrase followed by another noun/pronoun/verb phrase in brackets	Two individual sentences are formed. The first sentence is the same as the parent sentence sans the noun/pronoun.verb phrase in bracket. The second sentence constitutes of the noun phrase followed by noun/pronoun/verb phrase without the brackets.
4.	A noun phrase by a conditional phrase in brackets	Two individual sentences are formed. The first sentence is the same as the parent sentence sans the conditional phrase in bracket. The second sentence constitutes of noun phrases followed by conditional in the bracket.
5.	A conditional phrase followed by a sentence	Two dependent sentences are formed. The first sentence constitutes the conditional phrase. The second sentence constitutes rest of the sentence.
6.	A sentence in which the parent verb phrase is over two child verb phrases joined by a conjunction	Two dependent sentences are formed where the dependency is the conjunction. The first sentence is formulated by removing conjunction and second child verb phrase. The second sentence is formulated by removing conjunction and first child verb phrase.

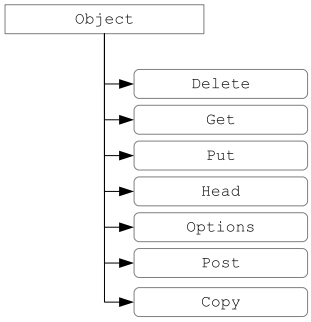

Figure 4: Semantic Graph for the `Object` related operations in Amazon S3 REST API

are the REST actions applicable on `Object` resource in `Amazon S3 REST API`.

5.6 Type Analysis

As mentioned earlier that some temporal constraints are enforced by the type system in typed Languages. For instances a method (m) accepting input parameter (i) of type (t) mandates that (at least one) method (m') be invoked whose return value is of type (t). To extend the temporal constraints inferred by the analyzing the natural language text, this component infers additional constraints that are encoded in the type system. Algorithm 2 lists the steps followed to infer type based temporal constraints.

The algorithm accepts the list of methods as an input produces a graph with the nodes representing methods in an API and the directed edges representing temporal constraints. First, an index is created based on the return types of the method (Line 2). Second, all methods in an API are added to an unconnected graph (Line 3-4). Then, for every public method in the input list, the algorithm checks the types of the input parameters and constructs and directed edge from all the methods whose return value have the same type to the method in question (Line 14- 20). The algorithm does not take into consideration the basic parameter types such as integer, string (Line 15). Additionally, an edge is created from the constructors of a class to the non static members methods of a class (Line 8 -13). The resultant graph is then returned by the algorithm.

Algorithm 2 Type_Sequence_Builder

Input: List $methodList$
Output: Graph seq_Graph
1: $Graph\ seq\_Graph\ =\ \phi$
2: $Map\ idx\ = createIdx(methodList)$
3: **for all** $Method\ mtd\ in\ methodList$ **do**
4: $\quad seq\_Graph.addVertex(mtd)$
5: **end for**
6: **for all** $Method\ mtd\ in\ methodList$ **do**
7: $\quad$ **if** $mtd.isPublic()$ **then**
8: $\quad\quad$ **if** $!mtd.isStatic()$ **then**
9: $\quad\quad\quad List\ preList\ =\ idx.query(mtd.declaringType)$
10: $\quad\quad\quad$ **for all** $Method\ mtd'\ in\ preList$ **do**
11: $\quad\quad\quad\quad seq\_Graph.addEdge(mtd', mtd)$
12: $\quad\quad\quad$ **end for**
13: $\quad\quad$ **end if**
14: $\quad\quad$ **for all** $Parameter\ param\ in\ mtd.getParameters()$ **do**
15: $\quad\quad\quad$ **if** $!isBasicType(param.Type)$ **then**
16: $\quad\quad\quad\quad List\ preList\ =\ idx.query(paramType)$
17: $\quad\quad\quad\quad$ **for all** $Method\ mtd'\ in\ preList$ **do**
18: $\quad\quad\quad\quad\quad seq\_Graph.addEdge(mtd', mtd)$
19: $\quad\quad\quad\quad$ **end for**
20: $\quad\quad\quad$ **end if**
21: $\quad\quad$ **end for**
22: $\quad$ **end if**
23: **end for**
24: **return** seq_Graph

The temporal constraints based on the type information can be extracted by querying the graph. The incoming edges to a node denoting a method represents the set of pre-requisite methods. The temporal constraint being, at least one of the pre-requisite methods must be invoked before invoking the method in question.

6. EVALUATION

We next present the evaluation we conducted to assess the effectiveness of ICON. In our evaluation, we address three main research questions:

- **RQ1**: What are the precision and recall of ICON in identifying temporal constraints from sentences written in natural language? Answer to this question quantifies the effectiveness of ICON in identifying constraint sentences.
- **RQ2**: What is the accuracy of ICON in inferring temporal constraints from constraint sentences in the API documents?

Answer to this question quantifies the effectiveness of ICON in inferring temporal constraints from constraint sentences.

- **RQ3**: What is the degree of the overlap between the temporal constraints inferred from natural language text in comparison to the typed-enforced temporal constraints?

6.1 Subjects

We used the API documents of the following two libraries as subjects for our evaluation.

- `Amazon S3 REST API` provides a REST based web services interface that can be used to store and retrieve data on the web. Furthermore, `Amazon S3` also empowers a developer with rich set of API methods to access a highly scalable, reliable, secure, fast, inexpensive infrastructure. `Amazon S3` is reported to store more than 2 trillion objects as of April 2013 and gets over 1.1 million requests per second at peak time [1].
- `java.io` : is one of a popular packages in `Java` programming language. The package provides APIs for system input and output through data streams, serialization and the file system, which are one of the fundamental functionalities provided by a programming language.

We chose `Amazon S3` and `java.io` APIs as our subjects because they are popular and contain decent documentation.

6.2 Experimental Setup.

We first manually annotated the sentences in the API documents of the two APIs. Two authors manually labeled each sentence in the API documentation as sentence containing temporal constraints or not. We used `cohen kappa` [7] score to statistically measure the inter-rater agreement. The `cohen kappa` score of the two authors was .66 (on a scale of 0 to 1), which denotes a statically significant agreement [7]. After the authors classified all the sentences, they discussed with each other to reach a consensus on the sentences they classified differently. We use these classified sentences as the golden set for calculating precision and recall.

To answer RQ1, we measure the number of true positives (TP), false positives (FP), true negative (TN), and false negatives (FN) in identifying the constraint sentences by ICON. We define constraint sentence as a sentence describing a temporal constraints. We define the TP, FP, TN, and FN of ICON as follows:

1. TP: A sentence correctly identified by ICON as constraint sentence.
2. FP: A sentence incorrectly identified by ICON as constraint sentence.
3. TN: A sentence correctly identified by ICON as not a constraint sentence.
4. FN: A sentence incorrectly identified by ICON as not a constraint sentence.

In statistical classification [23], `precision` is defined as a ratio of number of true positives to the total number of items reported to be true, `recall` is defined as a ratio of number of true positives to the total number of items that are true. `F-Score` is defined as the weighted harmonic mean of `precision` and `recall`. Higher value of `precision`, `recall`, and `F-Score` are indicative of higher quality of the constraint statements inferred using ICON. based on the calculation of TP, FP, TN, and FN of ICON defined previously we computed the `precision`, `recall`, and `F-Score` of ICON as follows:

$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$
$$F-Score = \frac{2XPrecisonXRecall}{Precision+Recall}$$

To answer RQ2, we manually verified the temporal constraints inferred from constraint sentences by ICON. However, we excluded the type-enforced temporal constraints inferred using Algorithm 2, described in Section 5. We excluded the type-enforced constraints because they are correct by construction and are by default enforced by modern IDE's such as eclipse. We then measure $accuracy$ of ICON as the ratio of the total number of temporal constraints that are correctly inferred by ICON to the total number of constraint sentences. Two authors independently verified the correctness of the temporal constraints inferred by ICON. We define the `accuracy` of ICON as the ratio of constraint sentences with correctly inferred temporal constraints to the total number of constraint sentences. Higher value of `accuracy` is indicative of effectiveness of ICON in inferring temporal constraints from constraint sentences.

To answer RQ3, we counted the overlap in the temporal constraints inferred by ICON from the natural language text in API documents to the type-enforced temporal constraints inferred using Algorithm 2, described in Section 5.

6.3 Results

6.3.1 RQ1: Effectiveness in Identifying Constraint Sentences

In this section, we quantify the effectiveness of ICON in identifying constraint sentences by answering RQ1. Table 2 shows the effectiveness of ICON in identifying constraint sentences. Column "API" lists the names of the subject API. Columns "Mtds" and "Sen" lists the number of methods and sentences in each subject API's. Column "Sen$_C$" list the number of manually identified constraint sentences. Column "Sen$_{ICON}$" lists the number of sentences identified by ICON as constraint sentences. Columns "TP", "FP", "TN", and "FN" represent the number of `true positives`, `false positives`, `true negatives`, and `false negatives`, respectively. Columns "P(%)", "R(%)", and "F$_S$(%)" list percentage values of `precision`, `recall`, and `F-score` respectively. Our results show that, out of 3,909 sentences, ICON effectively identifies constraint sentences with the average `precision`, `recall`, and `F-score` of 65.0%, 72.2%, and 68.4%, respectively.

We next present an example to illustrate how ICON incorrectly identifies a sentence as a constraint sentence (producing false positives). For instance, consider the sentence "*This is done by flushing the stream and then closing the underlying output stream.*" from `close` method description from `PrintStream` class. ICON incorrectly identifies the action "flush" being performed before the action "close". However ICON fails to make the distinction that it happens internally (enforced within the body) in the method. ICON, thus incorrectly identifies the sentence as a constraint sentence.

Another major source of FPs is the incorrect parsing of sentences by the underlying NLP infrastructure and/or inadequacy of generic dictionaries for synonym analysis. For instance, consider the sentence "*If this stream has an associated channel then the channel is closed as well.*" from the `close` method description from `FileOutputStream`. The sentence describes an effect that happens as a result of calling the `close` method and does not describe any temporal constraint. However, ICON annotates the sentence as a constraint sentence because underlying Wordnet dictionaries matches the word "has" as a synonym of "get". This in-

Table 2: Evaluation Results

API	Mtds	Sen	Sen_C	Sen_{ICON}	TP	FP	FN	P(%)	R(%)	F_S(%)	$Spec_{ICON}$	Acc(%)
java.io	662	2417	78	88	57	31	21	64.8	73.1	68.8	56	71.8
AMAZON S3 REST	51	1492	12	12	8	4	4	66.7	66.7	66.7	7	58.3
Total	713	3909	90	100	65	35	25	65.0*	72.2*	68.4*	63	70.0*

* Column average; Mtds: Total no. of Methods; Sen: Total no. of Sentences; Sen_C: Total no. of constraint Sentences; Sen_{ICON}: Total no. of constraint Sentences identified by ICON; TP: Total no. of True Positives; FP: Total no. of False Positives; FN: Total no. of False Negatives; P: Precision; R: Recall; F_S: F-Score; Acc: Accuracy $Spec_{ICON}$: Total no. of temporal constraint correctly identified by ICON;

correct matching in turn causes ICON to incorrectly annotate the sentence as constraint sentence because "has" is matched against (get) method in `FileOutputStream`. We observed 8 instances of previously described example in our results.

If we manually fixed the Wordnet dictionaries to not match "has" and "get" as synonyms, our precision is further increased to 70.8% effectively increasing the F-Score of ICON to 71.2%. Although an easy fix, we refrained from including such modifications for reporting the results to stay true to our proposed framework. In the future, we plan to investigate techniques to construct better domain dictionaries for software API.

We next present an example to illustrate how ICON fails identify a constraint sentence (producing false negative). False negatives are undesirable in the context of our problem domain because they can mislead the users of ICON into believing that no other temporal constraint exists in the API documents. Furthermore, an overwhelming number of false negatives works against the practicality of ICON. For instance, consider the sentence "*This implementation of the PUT operation creates a copy of an object that is already stored in Amazon S3.*" from `PUT Object-Copy` method description in `Amazon S3 REST API`. The sentence describes the constraint that the object must already be stored (invocation of `PUT Object`) before calling the current method. However, ICONcannot make the connection owing to the limitation of the semantic graphs that do not list "already stored" as a "valid operation" on object. In the future, we plan to investigate techniques to further improve knowledge graphs to infer such implicit constraints.

Another major source of false negatives (similar to reasons for false positives) is the incorrect parsing of sentences by the underlying NLP infrastructure. For instance, consider the sentence "*If any in-memory buffering is being done by the application (for example, by a BufferedOutputStream object), those buffers must be flushed into the FileDescriptor (for example, by invoking OutputStream.flush) before that data will be affected by sync.*" The sentence describes that the `OutputStream.flush()` must be invoked before invoking the current method if in-memory buffering is performed. However, the length and complexity in terms of number of clauses causes the underlying Stanford parser to inaccurately annotate the dependencies, which eventually results into incorrect classification.

Overall, a significant number of false positives and false negatives will be reduced as the current NLP research advances the underlying NLP infrastructure. Furthermore, use of domain specific dictionaries as opposed to generic dictionaries sed in current prototype implementation will further improve the precision and recall of ICON.

6.3.2 RQ2: Accuracy in Inferring Temporal Constraints

In this section, we evaluate the effectiveness of ICON in inferring temporal constraints from the identified constraint sentences from API documents.Table 2 shows the effectiveness of ICON in inferring temporal constraints from the identified constraint sentences. Column "API" lists the names of the subject API. Columns "Mtds" and "Sen" list the number of methods and sentences in each subject API's. Column "Sen_C" lists the number of manually identified constraint sentences. Column "$Spec_{ICON}$" lists the number of sentences with correctly inferred temporal constraints by ICON. Column "Acc(%)" list percentage values of accuracy. Our results show that, out of 90 manually identified constraint sentences, ICON correctly infers temporal constraints with the average accuracy of 70.0%.

We next present an example to illustrate how ICON incorrectly infers temporal constraints from a constraint sentence. Consider the sentence "*if the stream does not support seek, or if this input stream has been closed by invoking its close method, or an I/O error occurs.*" from `skip` method of `java.io.FilterInputStream` class. Although ICON correctly infers that method `close` cannot be called before current method, ICON incorrectly associates the phrase "support seek" with method `markSupported` in the class. The faulty association happens due to incorrect parsing of the sentence by the underlying NLP infrastructure. Such issues will be alleviated as the underlying NLP infrastructure improves.

Another, major cause of failure for ICON in inferring temporal constraints from sentences is the failure to identify the sentence as a constraint sentences at the first place (false negatives). Overall, accuracy of ICON can be significantly improved by lowering the false negative rate in identifying the constraint sentences.

6.3.3 RQ3: Comparison to Typed-Enforced Constraints

In this section, we compared the temporal constraints inferred from the natural language API descriptions to those enforced by the type-system (referred to as type-enforced constraint). The constraints that are enforced by the type-system can be enforced by IDEs. Hence, for such types of constraints, we do not require sophisticated techniques like ICON. For `java.io`, we define a type-enforced constraint as a constraint that mandates a method M accepting input parameter I of type T to be invoked after (at least one) a method M' whose return value is of type T. Since there are no types in REST APIs, for `Amazon S3`, we consider a constraint as a type-enforced constraint if the constraint is implicit in the `CRUD` semantic followed by REST operations. `CRUD` stands for resource manipulation semantic sequence create, retrieve, update, and delete. In particular, we consider a constraint as a type-enforced constraint, if the constraint mandates a DELETE, GET, or PUT operation on a resource to be invoked after a POST operation on the same resource.

To address this question, we manually inspect each of the constraints reported by ICON and classify it as a type-enforced constraint or a non type-enforced constraint. We observed that none of the constraints inferred by our ICON from natural language text were classified as a type-enforced constraint. Hence, the con-

straints detected by ICON are not trivial enough to be enforced by a type system.

6.4 Summary

In summary, we demonstrate that ICON effectively identifies constraint sentences (from over 3900 API sentences) with the average precision, recall, and F-score of 65.0%, 72.2%, and 68.4% respectively. Furthermore, we also show that ICON infers temporal constraints from the constraint sentences an average accuracy of 70%. Furthermore, also provide discussion that a false positives rate and false negatives rate can be further improved by improving the underlying NLP infrastructure. Finally, we provide a comparison of the temporal constraints inferred from natural language description against the temporal constraints enforced by a type system.

6.5 Threats to Validity

Threats to external validity primarily include the degree to which the subject documents used in our evaluations are representative of true practice. To minimize the threat, we used API documents of two different API's: JDK `java.io` and `Amazon S3 REST API` developer documentation. On one hand, Java is a widely used programming language and `java.io` and is one of the main packages. In contrast, `Amazon S3 REST API` developer documentation provides HTTP based access to online storage allowing developers the freedom to write clients applications in any programming language. Furthermore, the difference in the functionalities provided by the two API's also address the issue of over fitting our approach to a particular type of API. The threat can be further reduced by evaluating our approach on more subjects API's.

Threats to internal validity include the correctness of our prototype implementation in extracting temporal constraints and labeling a statement as a constraint statement. To reduce the threat, we manually inspected all the constraints inferred against the API method descriptions in our evaluation. Furthermore, we ensured that the results were individually verified and agreed upon by two authors, using the `cohen kappa` [7] score to statistically measure the inter-rater agreement.

7. DISCUSSION AND FUTURE WORK

Our approach serves as a way to formalize the description of constraints in the natural language texts of REST API documents (targeted towards generating code contracts), thus facilitating existing tools to process these specifications. We next discuss some of the limitations of the current implementation and our approach.

Validation of Method Descriptions. API documents can sometimes be misleading [28, 34], thus causes developers to write faulty client code. In future work, we plan to extend our approach to find documentation-implementation inconsistencies.

Inferring implicit constraints. The approach presented in this work only infers temporal constraints explicitly described in the method descriptions. However, there are instances where the constraints are implicit. For instance, consider the method description for `markSupported` method in `BufferInputSttream` class in Java, which states "*Test if this input stream supports* `mark`". For a developer it is easy to understand that method `markSupported` must be called before method `mark`. The approach presented in this work is unable to infer such constraints. In future work, we plan to investigate techniques to infer these implicit constraints.

Extending generic dictionaries. The use of generic dictionaries for software engineering related text is sometimes inadequate. For instance, Wordnet matched "has" as a synonym for the word "get". Although, valid for generic English, such instances cause our approach to incorrectly distinguish a constraint sentence from a regular sentence, or vice versa. In future work, we plan to investigate techniques to extend generic dictionaries for software engineering related text.

8. CONCLUSION

Although highly desirable, most API's do not have formal temporal constraints. In contrast, documentation of methods contain detailed specifications of the usage in natural language text. Manually writing formal specifications based on natural language text in API documents is prohibitively time consuming and error prone. To address this issue, we proposed a novel approach ICON to infer temporal constraints from natural language text of API documents. We applied ICON to infer temporal constraints from commonly used package `java.io` in the JDK API and from the Amazon S3 REST API. Our evaluation results show that ICON effectively identifies sentences describing temporal constraints with an average 65% precision and 72% recall, from more than 3900 sentences in subject API documents. Furthermore,ICON also achieved an accuracy of 70% in inferring 63 formal temporal constraints from these sentences.

9. REFERENCES

[1] Amazon S3 - Two Trillion Objects, 1.1 Million Requests / Second. `http://aws.typepad.com/aws/2013/04/amazon-s3-two-trillion-objects-11-million-requests-html`.

[2] T. Ball and S. K. Rajamani. The SLAM project: debugging system software via static analysis. In *ACM SIGPLAN Notices*, volume 37, pages 1–3. ACM, 2002.

[3] B. K. Boguraev. Towards finite-state analysis of lexical cohesion. In *Proc. FSMNLP*, 2000.

[4] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[5] R. P. Buse and W. Weimer. Synthesizing API usage examples. In *Proc. 34th ICSE*, pages 782–792, 2012.

[6] R. P. Buse and W. R. Weimer. Automatic documentation inference for exceptions. In *Proc. 17th ISSTA*, pages 273–282, 2008.

[7] J. Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

[8] C. Csallner, N. Tillmann, and Y. Smaragdakis. DySy: Dynamic symbolic execution for invariant inference. In *Proc. 30th ICSE*, pages 281–290, 2008.

[9] M. C. de Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *Proc. LREC*, 2006.

[10] M. C. de Marneffe and C. D. Manning. The stanford typed dependencies representation. In *Workshop COLING*, 2008.

[11] U. Dekel and J. D. Herbsleb. Improving API Documentation Usability with Knowledge Pushing. In *Proc. 31st ICSE*, pages 320–330, 2009.

[12] F. et al. *WordNet: an electronic lexical database*. Cambridge, Mass: MIT Press, 1998.

[13] C. Flanagan and K. R. M. Leino. Houdini, an annotation assistant for ESC/Java. In *Proc. 10th FME*, pages 500–517, 2001.

[14] C. Ghezzi, A. Mocci, and M. Monga. Synthesizing intensional behavior models by graph transformation. In *Proc. 31st ICSE*, pages 430–440, 2009.

[15] J. Henkel, C. Reichenbach, and A. Diwan. Discovering documentation for Java container classes. *IEEE Transactions on Software Engineering*, 33:526–543, 2007.

[16] J. Henkel, C. Reichenbach, and A. Diwan. Developing and debugging algebraic specifications for Java classes. *ACM Trans. Softw. Eng. Methodol.*, 17(3):14:1–14:37, 2008.

[17] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proc. 41st ACL*, pages 423–430, 2003.

[18] D. Klein and C. D. Manning. Fast exact inference with a factored model for natural language parsing. In *Proc. 15th NIPS*, pages 3 – 10, 2003.

[19] C. Lee, D. Jin, P. Meredith, and G. Rosu. Towards categorizing and formalizing the JDK API. *Technical Report http://hdl.handle.net/2142/30006, Department of Computer Science, University of Illinois at Urbana-Champaign*, 2012.

[20] Apache Lucene Core. `http://lucene.apache.org/core/`.

[21] C. Manning and H. Schutze. Foundations of statistical natural language processing. *The MIT Press*, 2001.

[22] J. W. Nimmer and M. D. Ernst. Automatic generation of program specifications. In *Proc. ISSTA*, pages 232–242, 2002.

[23] D. Olson. *Advanced data mining techniques*. Springer Verlag, 2008.

[24] R. Pandita, X. Xiao, W. Yang, W. Enck, and T. Xie. Whyper: towards automating risk assessment of mobile applications. In *Proc. 22nd USENIX conference on Security*, pages 527–542, 2013.

[25] R. Pandita, X. Xiao, H. Zhong, T. Xie, S. Oney, and A. Paradkar. Inferring method specifications from natural language API descriptions. In *Proc. 34th ICSE*, 2012.

[26] N. Polikarpova, I. Ciupa, and B. Meyer. A comparative study of programmer-written and automatically inferred contracts. In *Proc. 18th ISSTA*, pages 93–104, 2009.

[27] B. Rubinger and T. Bultan. Contracting the Facebook API. In *4th AV-WEB*, pages 61–72, 2010.

[28] C. Rubino-González and B. Liblit. Expect the unexpected: Error code mismatches between documentation and the real world. In *Proc. 9th PASTE*, pages 73–80, 2010.

[29] A. Sinha, A. M. Paradkar, P. Kumanan, and B. Boguraev. A linguistic analysis engine for natural language use case description and its application to dependability analysis in industrial use cases. In *Proc. DSN*, pages 327–336, 2009.

[30] A. Sinha, S. M. SuttonJr., and A. Paradkar. Text2test: Automated inspection of natural language use cases. In *Proc. ICST*, pages 155–164, 2010.

[31] J. Slankas and L. Williams. Access control policy extraction from unconstrained natural language text. In *Proc. PASSAT*, 2013.

[32] L. Tan, D. Yuan, G. Krishna, and Y. Zhou. /*icomment: bugs or bad comments?*/. In *21st SOSP*, pages 145–158, 2007.

[33] L. Tan, Y. Zhou, and Y. Padioleau. aComment: mining annotations from comments and code to detect interrupt related concurrency bugs. In *Proc. 33rd ICSE*, pages 11–20, 2012.

[34] S. H. Tan, D. Marinov, L. Tan, and G. T. Leavens. @tComment: Testing javadoc comments to detect comment-code inconsistencies. In *Proc. 5th ICST*, April 2012.

[35] S. Thummalapenta, S. Sinha, N. Singhania, and S. Chandra. Automating test automation. In *Proc. 34th ICSE*, pages 881–891, 2012.

[36] S. Thummalapenta and T. Xie. PARSEWeb: A programmer assistant for reusing open source code on the web. In *Proc. 22nd ASE*, pages 204–213, 2007.

[37] N. Tillmann, F. Chen, and W. Schulte. Discovering likely method specifications. In *Proc. 8th ICFEM*, pages 717–736, 2006.

[38] J. Wang, Y. Dang, H. Zhang, K. Chen, T. Xie, and D. Zhang. Mining succinct and high-coverage API usage patterns from source code. In *Proc. 10th Working Conference on MSR*, pages 319–328, 2013.

[39] Q. Wu, L. Wu, G. Liang, Q. Wang, T. Xie, and H. Mei. Inferring dependency constraints on parameters for web services. In *Proc. 22nd WWW*, pages 1421–1432, 2013.

[40] X. Xiao, A. Paradkar, S. Thummalapenta, and T. Xie. Automated extraction of security policies from natural-language software documents. In *Proc. 20th FSE*, pages 12:1–12:11, 2012.

[41] H. Zhong, T. Xie, L. Zhang, J. Pei, and H. Mei. Mapo: Mining and recommending API usage patterns. In *Pro. 23rd ECOOP*, pages 318–343, 2009.

[42] H. Zhong, L. Zhang, T. Xie, and H. Mei. Inferring resource specifications from natural language API documentation. In *Proc. 24th ASE*, pages 307–318, 2009.