An Open Dataset of Abbreviations and Expansions

Christian D. Newman*, Michael J. Decker§, Reem S. AlSuhaibani†
Dishant Kaushik*, Anthony Peruma*, Emily Hill ‡
*Rochester Institute of Technology, Rochester, NY, USA
†Kent State University, Kent, OH, USA, Prince Sultan University, Riyadh, Saudi Arabia
‡ Drew University, New Jersey, USA
§ Bowling Green State University, Bowling Green, OH, USA
cnewman@se.rit.edu, mdecke@bgsu.edu, ralsuhai@kent.edu, dxk3597@rit.edu, axp6201@rit.edu, ehill1@drew.edu

***Abstract*—We present a data set of abbreviations and expansions, derived from a set of five open source systems, for use by the research and development communities.**

***Index Terms*—Abbreviation Expansion, Program Comprehension**

I. Introduction

Abbreviations are a type a non-dictionary term; i.e., terms that have no dictionary definition. Developers regularly use abbreviations in the code as shorthand for concepts that they are familiar with. Unfortunately, not all developers will be familiar with any given abbreviation. Additionally, tools that perform analysis using the natural language found in code may not have a way of dealing with these abbreviations appropriately (e.g., by expanding them). Thus, abbreviation expansion tools are important.

However, even with these tools, it is often difficult to find the appropriate expansion for a given abbreviation because a given abbreviation may have multiple expansions and the correct expansion may not appear in the same location (e.g., file, project, etc) as its abbreviation. The data we present in this artifact aims to support research into and construction of tools to expand abbreviations.

II. Dataset and Code

The dataset, derived from a previous study [1], is split into a collection of comma-separated values (CSV) files. An example of the data is given in Table I. There are three columns; the first is the original identifier in which the abbreviation was found. The second column is a parenthesized list of each abbreviation and expansion found in the original identifier. Each abbreviation:expansion entry in this list is separated by a dash (-). The third column is the original identifier but split into its individual terms. Additionally, every abbreviation found in the original identifier is separated from the other terms by parenthesis. This makes it easy to map the abbreviation found in the second column to its location in the original identifier if required.

There are other types of non-dictionary terms aside from abbreviations. When we ran into a term that is a non-dictionary term but not an abbreviation, we put it in one of two other categories. All categories used in this dataset are presented in Table II. If a word or symbol was not an abbreviation, then its expansion includes the category as part of its string (e.g., x:mathvariable, a:distinguisher). The entire set is derived from five open source systems varying in size, age, and language. These systems, along with the number of abbreviation-expansion pairs (in parenthesis) they contain, are: Open Office (143), Wycheproof (156), Enscript (156), Telegram (164), and Kdevelop (242) for a total of 861 abbreviation-expansion pairs.

TABLE I: Example CSV dataset shown in table form

Original Identifier	(Abbreviation:Expansion)	Split Identifier
locationIdx	(idx:Index)	location (Idx)
fl	(fl:flock)	(fl)
sem_name	(sem:semaphore)	(sem) name
ThreadInfo	(info:Information)	Thread (Info)
actWin	(act:activation-win:window)	(act) (Win)
dbusInterface	(db:database-us:user)	(db) (us) Interface

TABLE II: Categories Used to Classify Non-dictionary Terms

Category	Definition	Example
Abbreviation	Sequence of one or more letters which represent a longer word or phrase.	Pub $\rightarrow Public$ Cfg $\rightarrow Configure$ Kv $\rightarrow Keyvalue$
Distinguisher	Word or symbol whose only purpose is to avoid name-collision at compile-time.	int x, x1, x2; The numbers 1 and 2 are distinguishers
Math Variable	Sequence of one or more letters representing a mathematical concept which cannot be summarized with a word or simple phrase.	YScale: Y is a math variable AnimateToX: X is a math variable

III. Where to Obtain Dataset and Code

The dataset is available through github via the following link: https://github.com/SCANL/AbbreviationArtifact-ICSME2019. Additionally, we encourage others to submit pull requests with their own abbreviation:expansion sets for the research and development communities to take advantage of.

References

[1] C. D. Newman, M. J. Decker, R. S. AlSuhaibani, A. Peruma, D. Kaushik, and E. Hill, "An empirical study of abbreviations and expansions in software artifacts," in *Proceedings of the 35th IEEE International Conference on Software Maintenance and Evolution*, IEEE, 2019.