Natural Language Processing (NLP) Applied in Issue Trackers

4th International Workshop on
NLP for Software Engineering (NL4SE)
Lake Buena Vista, USA

Mathias Ellmann
Master of Science (M.Sc.)
Bachelor of Engineering (B.Eng.)

Research Method and Research Data

- Literature study of 35 - 40 papers
- Focus on the characteristics, syntactical, semantical and prediction analysis of duplicates
- Issue trackers are Eclipse Bugzilla, Mozilla and Open Office
- There are 225.000 bug reports in Eclipse and 420.000 bug reports in Mozilla
- 20% bugs in Eclipse and 30% bugs in Mozilla are duplicated

Duplicate Tasks in Eclipse Bugzilla

Characteristic Analysis

- Frequencies of Duplicates
  - In Sony Ericsson there are 10%, in Mozilla 30% and in Eclipse Bugzilla are 20% duplicate development task
- Content of Duplicates
  - There are same titles and descriptions
  - Task ids are very close to each other
  - Duplicates contain paraphrases and can often not be understood
- Reasons of Duplicates
  - Duplicate does not depend on the type of report (enhancement or defect)
  - There is no template available to formulate development tasks
- Creation and Closing time of duplicates
  - Duplicate were added after the same or a short period of time
  - Duplicates increase in the first year of a software project

Syntactic Analysis

- Terms & Paraphrases
  - Developers use synonyms and morpholic variations of words to describe duplicates
  - Developers use short forms of words as “config” and “configuration”
  - With a n-gram algorithm that evaluates a sequence of characters it is possible to identify duplicate Information & Concepts
  - Information co-occur e.g., in a description of a development task
  - Duplicate tasks can contain misspelled words
  - Duplicates can describe similar concepts and can be found by a n-gram algorithm or a Jaccard algorithm
- Similarities Between Duplicates
  - Similarity measurements have a similarity of up to 73.62% if they are located in the same product and component

Semantic Analysis

- Term Dependencies of Duplicates
  - There are similar concepts especially in system messages
  - There are paraphrases in duplicate development tasks
  - The semantic similarity of duplicates might depend on the chosen terms as class names, method names and others
- Context Dependencies of Duplicates
  - Terms are used that can be understood in the development context as RemoteFileAction and org.eclipse.ui.DefaultTextEditor
  - Similarities and topics (3G, alarm or others) can be identified by using the LDA algorithm
- Identifying a Semantic Similarity Between Duplicates
  - Measuring from a character level might be needed because of compound words: Out of memory Vs. OutOfMemoryError
  - Contextual information are also needed to identify duplicates
  - In comments are information located to identify duplicates

Prediction and Classification Analysis

- Feature Selection
  - Textual features can be extracted from the summary of a description
  - Categorical features can be extracted as the priority of a development task or the component name
  - Contextual features are needed that can be extracted from the architecture of a system
- Feature Extraction and Optimization
  - Measurement of frequencies of the information entities
  - Using a logarithmic function to lower the influence of large numbers of words that occur in the task descriptions
  - Using BM25F to weight the features
  - Using LDA algorithms to find duplicates with a k from 140-320
- Duplicate Prediction
  - Decision Tree Models outperformed support vector machines (SVN) and others to identify duplicate tasks

Takeaways from the Study

- Duplicate development tasks exist because of missing expertise and knowledge to use the right terms in the right development context
- Contextual Features are very useful when predicting duplicates when non-tuning features e.g., when using a gradient descent function
- Concepts let appear tasks as semantically similar
- Similar terms in a duplicate development tasks might be better understood when working in a development context
- There are several systems as Mylyn or others that aggregate the development context during the development of a software

➡ A system can compare the development context to find duplicates before they are posted in an issue tracker

Mathias Ellmann

University of Hamburg,
Germany

mathias.ellmann.cs@gmail.com

https://scholar.google.de/citations?user=JOo4WLgAAAAJ&hl=en

ACM Reference Format:
Mathias Ellmann. 2018. Natural Language Processing (NLP) Applied on Issue Trackers. In *Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering (NL4SE '18), November 4, 2018, Lake Buena Vista, FL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3283812.3283825