Automated similarity detection: identifying duplicated requirements

Motger de la Encarnación, Quim
Universitat Oberta de Catalunya
Universitat Politècnica de Catalunya

Abstract

Natural Language Processing (NLP) is one of the main areas in Artificial Intelligence (AI) with a large potential in different applications involving all kinds of documents and textual items. NLP applies to the process of developing partial representations of features and rules of natural language based on its textual information, which includes both syntactic and semantic knowledge [1]. The main purpose of this technology is to use this representative knowledge in order to apply automated analysis and generation of text units, such as comprehensive sentences or full documents.

One of the areas of application of NLP is the Requirements Engineering (RE) field. RE is the set of activities and processes of software engineering focused on the development, analysis, communication and management of a set of requirements that describes the features of a system [2]. Software development experience in recent years proves that managing and maintaining large sets of requirements have become critical issues. This problem is even more challenging due to the large amount of data and the dimensions that these projects are dealing with nowadays [3]. Whether the analysis and the evaluation of requirements is a tedious, time-consuming task, it is critical that they are carried out with both accuracy and efficiency in any software development project.

Between the main problems of RE we can highlight the detection and management of duplicated requirements [4]. If ignored, these duplicated items may lead to redundancy in the textual information of a project and therefore this may lead to the duplicity of tasks, which are critical issues from the project management perspective. Moreover, the automation of this process and the standardized usage of specific, accurate tools are still at a state-of-the-art stage. On the one hand it is difficult to find open source tools and frameworks providing generic, adaptive solutions for duplicate detection, and most of them are addressed to a very specific casuistic or use case [5]. On the other hand, similarity detection algorithms are highly tightened to the quality of the data used for the detection process [6].

This is the starting point of this master thesis: a proposal to apply automated requirement similarity detection, using artificial intelligence techniques, for the detection of duplicates between project requirements. The main goals of this research can be classified in three phases. First of all, to perform a deep evaluation of the state of the art in the field of similarity detection in natural language texts. This analysis must be based on the study of different similarity detection algorithms, as well as to comprehend and to understand the main features of each described algorithm, its strengths, its weaknesses, and the main areas of application. Second of all, to develop and to implement a subset of similarity detection algorithms that can be applied to a requirements' dataset of a real use case. Finally, the analysis in terms of both accuracy and performance of the algorithms, insisting not only in the importance of reliable and accurate results but also in the efficiency of the algorithms.

This thesis will be developed within the OpenReq project [7], an EU Horizon 2020 project which main goal is *"to build an intelligent recommendation and decision system for community-driven requirements engineering"*.

Keywords— artificial intelligence, requirements engineering, machine learning, natural language processing, textual similarity, similarity detection

References

[1] Collobert, Ronan, Jason Weston, León Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research, http://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf.

[2] Jeremy Dick, Elizabeth Hull, Ken Jackson, Requirements Engineering, Springer, pp. 7–9, ISBN 978-3-319-61073-3

[3] Kasauli, Rashidah, Grischa Liebel, Eric Knauss, Swathi Gopakumar, and Benjamin Kanagwa. Requirements Engineering Challenges in Large-Scale Agile System Development - IEEE Conference Publication. https://ieeexplore.ieee.org/document/8049141.

[4] Natt och Dag, J., Regnell, B., Carlshamre, P. et al, A Feasibility Study of Automated Natural Language Requirements Analysis in Market-Driven Development - Requirements Eng (2002) 7: 20, https://doi.org/10.1007/s007660200002

[5] Tung Khuat, Nguyen Hung, Le Thi My Hanh, A Comparison of Algorithms used to measure the Similarity between two documents. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), https://pdfs.semanticscholar.org/43f8/027780d2694331ca373c57f9a2ace509a7b6.pdf

[6] Yu Huang, Fei Chiang, Refining Duplicate Detection for Improved Data, http://ceur-ws.org/Vol-2038/paper3.pdf Quality

[7] Requirements Engineering - Tools and Solutions Offered by OpenReq, OpenReq, https://openreq.eu/.