Automated similarity detection: identifying duplicated requirements *PEC2 - Project development*

Quim Motger de la Encarnacin

Universitat Oberta de Catalunya
Universitat Politecnica de Catalunya

1 Introduction

This project is the first review report of the master thesis titled *"Automated similarity detection: identifying duplicated requirements"*. The following sections summarize the status of the planned tasks and goals of the project in correlation with the current status of their development. Additionally, all changes and deviations from the original plan depicted in previous deliverables are described and justified in this report. Finally we enumerate the list of partial results obtained until the date of this report.

2 Project development status

As an introduction to the project development status, it can be stated that the project has not suffered from major deviations from its scope perspective (see sec. 2.2 for more details). As the tasks work plan depicted in the previous deliverable, development work has been focused on satisfying each one of the specific goals. This has allowed to follow track of goal achievement from the beginning of the thesis development with enough detail to avoid major deviations (see sec. 3 for more details about work plan).

We focus now on two topics related to the project development status: first, we enumerate and evaluate the satisfaction status of each of the specific goals listed in the work plan; second, we justify the absence of project goals changes using the previous evaluation as a prove.

2.1 Goal achievement and expected results

In order to provide an exhaustive evaluation of the specific goals, we enumerate all specific objectives and we use a 3-tag model to determine its achievement status:

1. *Not achieved*: the objective has not been addressed yet and there are not available results for its satisfaction
2. *Partially achieved*: the objective is partially satisfied, meaning that there is important work and partial results but there are still remaining tasks for its complete achievement
3. *Achieved*: the objective has been completely achieved and there are partial results to justify it

For each goal we give further details about its status and its expected development until the next report. Notice that this is a general overview from a goal perspective: all details related to specific tasks are extended in section 3.

O1.1. To study the current status of similarity detection in the RE field from a general point of view.

- *Status:* Achieved
- *Description:* the state of the art review has been fully achieved and its only pending to be fully documented in the master thesis final deliverable. Currently a set of provisional artifacts (i.e. lists of publication references, summaries, schemes...) are available and are being used as an input for development tasks.

O1.2. To review and to enumerate similarity detection techniques/algorithms, and to be specific the ML and NLP techniques that represent the state-of-the-art of the field.

- *Status:* Achieved
- *Description:* as part of the state of the art, and as a step of the systematic review carried out (see 3.2 for further details) this literature review has been used to identify the state of the art from 3 main fields: from a general algorithmic solution view; from a set of NLP measures and preprocessing techniques view; and from a ML classification view. As a result is has been possible to identify and to study the most representative techniques from the areas of interest.

O1.3. To identify potential suitable algorithm candidates for the master thesis and the use case to be validated with.

- *Status:* Achieved
- *Description:* two different solutions have been chosen to develop and to evaluate in this master thesis. These two approaches have proven to be some of the most integrative and up-to-date solutions from similarity detection in natural language, and additionally have a reason of interest in its applicability to requirements similarity detection.

O2.1. To elaborate a development proposal for the implementation of the selected algorithms.

- *Status:* Partially achieved
- *Description:* currently the development of the project is limited to an initial PoC version of one of the two algorithms. However this version is fully usable and it is only pending to be improved by applying additional optimization techniques suggested in some research papers to improve its results. Partial results are already available for real use case data. Therefore we can conclude that this goal has been partially achieved, pending to finish the implementation of the first algorithm and to develop the second one.

O2.2. To integrate the algorithms with a unique tool to use and to test the different similarity detection scenarios.

- *Status:* Partially achieved
- *Description:* due to development requirements and the platforms and technologies required for developing the project in Java, which has been the programming language selected for this purpose, there is already an integrative tool prepared to work with different algorithms. It basically consists on a RESTful API developed with Spring which focuses on CRUD functionalities for requirement items and separated methods for applying the different similarity techniques. When objective *O2.1* is fulfilled with the second algorithm, it can be easily coupled in this integrative tool.

O3.1 To evaluate the requirements input data of the algorithms, in order to guarantee a comprehensive analysis of the results.

- *Status:* Partially achieved
- *Description:* Import/export functionalities, adaptation of algorithms...

O3.2 To optimize and to adapt the algorithms based on the use case requirements.

- *Status:* Partially achieved
- *Description:* same as above

O3.3 To analyze and to prepare a use case dataset for all scenarios (i.e. all the different similarity detection algorithms).

- *Status:* Partially achieved
- *Description:* same as above

O3.4 To carry out the experiments using the developed algorithms.

- *Status:* Not achieved
- *Description:*

O3.5 To perform a comparative analysis between algorithms.

- *Status:* Not achieved
- *Description:*

O3.6 To extract conclusions in terms of reliability of the results and performance of the algorithms.

- *Status:* Not achieved
- *Description:*