Automatic Word Embeddings-Based Glossary Term Extraction from Large-Sized Software Requirements

Siba Mishra and Arpit Sharma

Department of Electrical Enginnering and Computer Science
Indian Institute of Science Education and Research
Bhopal, Madhya Pradesh, India
{sibam,arpit@iiserb.ac.in}

Abstract. [**Context and Motivation**] Requirements glossary defines specialized and technical terms used in a requirements document. Requirements glossary helps in improving the quality and understandability of requirements documents. [**Question/Problem**] Manual extraction of glossary terms from a large body of requirements is an expensive and time-consuming task. This paper proposes a fundamentally new approach for automated extraction of glossary terms from large-sized requirements documents. [**Principal Ideas/Result**] Firstly, our technique extracts the candidate glossary terms by applying text chunking. Next, we apply a novel word embeddings based semantic filter for reducing the number of candidate glossary terms. Since word embeddings are very effective in identifying terms that are semantically very similar, this filter ensures that only domain-specific terms are present in the final set of glossary terms. We create a domain-specific reference corpus by Wikipedia crawling and use it for computing the semantic similarity scores of candidate glossary terms. We apply our technique to a large-sized requirements document, i.e., CrowdRE dataset with around 3000 crowd-generated requirements for smart home applications. Semantic filtering reduces the number of glossary terms by 92.7%. To evaluate the quality of our extracted glossary terms we manually create the ground truth data from CrowdRE dataset and use it for computing precision and recall. Additionally, we also compute the requirements coverage of these extracted glossary terms. [**Contributions**] Our detailed experiments show that word embeddings based semantic filtering can be very useful for extracting glossary terms from a large body of requirements.

Keywords: Requirements Engineering · Natural Language Processing · Word Embeddings · Term Extraction · Semantic Filter.

1 Introduction

Requirements are the basis for every project, defining what the stakeholders in a potential new system need from it, and also what the system must do in order to satisfy that need [8, 19]. All subsequent steps in software development are