1 Introduction

My project proposal is to apply information extraction techniques to capture compatibility relationships between software components that exists in the form of unstructured text on online forums. My goal is to focus on the Stack Overflow Q&A site. Therefore, related work here is presented in the context of information extraction techniques with a focus on the Software Engineering domain. In section 2, relevant work related to mining software engineering knowledge is presented. Specifically, work related to extracting information from Stack Overflow is reviewed. In section 3, work related to applications of relation extraction is discussed. First, relation extraction in the Biomedical domain is presented since it is a well known research area for application of relation extraction techniques. Secondly, relation extraction in the context of ontology construction in software engineering is presented. Work pertaining to Software Engineering relation extraction from Stack Overflow is then discussed. Section 4 provides a summary and lists the challenges that need to be solved.

2 Text mining software engineering knowledge

NLP techniques have been used to mine software engineering knowledge from various sources Al Omran and Treude (2017). For example, Treude et al. (2015) present a task extraction technique for software documentation. A document corpus is analyzed and passages are detected that describe how to perform a particular task. Dependency parsing is used in the task identification process. Stack Overflow Q&A site has been studied as a useful resource to mine software engineering information. Barua et al. (2014) use LDA topic modeling technique to discover the main topics present in developer discussions. In Chen and Xing (2016), association rule mining and community detection techniques are used to mine the technology landscape from Stack Overflow tags. The landscape is presented as a type of graphical structure called a Technology Associate Network. In Treude and Robillard (2016), the authors aim to augment API documentation with relevant insight sentences extracted from Stack Overflow. Insight sentences are text that contain additional useful information not contained in the API documentation. A supervised machine learning approach is used employing many features from the Stack Overflow post itself to generate the insight sentences.

3 Relation extraction applications

3.1 Relation extraction in the Biomedical domain

In Zhou et al. (2014), a survey of relation extraction techniques in the Biomedical domain is presented. The task of capturing Protein to Protein interaction (PPI) is modeled as a binary relation extraction task. Protein name tagging is first done and then trigger words for relations between the proteins are identified. In rule based approaches for PPI detection, regular expressions are used over words or POS tags to identify relations between protein entities. In machine learning based approaches, syntactic and semantic features are used to detect relations between entities. Similar approaches used in PPI could be applied in the software engineering domain.

3.2 Relation extraction in the Software Engineering domain

There have been works that are related to developing a taxonomy for the software engineering domain, as compared to general purpose taxonomies
like WordNet. In Zhu et al. (2015), the authors use Stack Overflow tags as vocabularies for software engineering domain and identify subsumption relations between tags.

The most relevant work for my project is perhaps Zhao et al. (2017). This work is related to applying information extraction techniques from online content in the software engineering domain such as Stack Overflow. This work is related to discover domain specific concepts and extracting relation triples from content of web pages. The evaluation is performed on content from Stack Overflow and a knowledge graph is built with concepts and relation triples from the software engineering domain. Here, a dependency parser is used to extract dependencies of the Noun and Verb phrases from sentences. Rule-based chunking is used to generate candidate relation triples. This is accomplished in two parts. First, the tokens in the sentence are chunked to nouns and verbs according to the POS and then the dependencies are used to chunk the nouns and verbs to relation triples. From the candidate set of relation triples, features are extracted and an SVM classifier is used to retrieve the relation triples with higher software engineering domain relevance.

4 Summary

While my proposal shares some similarities with Zhao et al. (2017) in performing relation extraction from Stack Overflow, the goals are different. Zhao et al. (2017) use an open information extraction technique to discover any domain relevant relations. My goal is to extract specific “compatibility” relationship types between software component entities. So far, the literature indicates such an exercise has not been attempted. In the proposed high level approach for my project, the first step is to perform Named Entity Recognition of software components and then detect compatibility relations between the entities. As seen, NER can be accomplished through rule-based or machine learning approaches by leveraging “tag” information in Stack Overflow and relevant taxonomies. The main challenge lies in the Relation Extraction task. The following are the potential limitations that need to be overcome:

- There are no standard trigger words for “compatibility” relation identification and there is a lack of relevant annotated relation corpora for the software engineering domain.

Here, my plan is to first explore manual annotation and then possibly use bootstrapping methods.

- The relations between components may not be binary and may not occur within the same sentence. This can potentially be addressed by using coreference resolution, but Zhou et al. (2014) states that this is a well known problem in NLP and therefore its usefulness remains to be tested.

References


