**PROJECT PROPOSAL**

**SEQUENTIAL LABELING TASK FOR NATUAL LANGUAGE REQUIREMENTS**

Natural language requirements are often incomplete and inconsistent. It is important to address the incompleteness and inconsistency in requirements, as such requirements might result in poor quality of end product. In particular, with respect to safety-critical systems, poor quality of requirements might result in hazardous and catastrophic events. To address incompleteness, inconsistency and abnormal behaviors caused because of them, we use Causal Component Model (CCM) [1]. CCM needs all natural language requirements to be converted into CCM specifications which are in form of rules. Every Rules is structured as Cause: Component(currentstate) -> Component(nextstate). So, CCM has three domain elements, components, their states and causes for state transition between components.

In order to automate the modeling process, it is essential to identify the above mentioned domain elements, i.e., components, states and causes, in natural language requirements documents. Since components, states and causes are largely independent label types, classification of these domain elements can be treated as distinct classification tasks. Since, components, states and causes can be phrases or have multiple words, we use IOB (Inside, Outside and Beginning) tags to label data, that is every word has a label which can be B or I or O. For example: Pulse Generator is part of pacemaker. If we need to identify components in this statement, the labels are as follows:

|  |  |
| --- | --- |
| Word | Label |
| Pulse | B |
| Generator | I |
| is | O |
| a | O |
| part | O |
| of | O |
| pacemaker | B |

However, using current LSTM with RNN is not able to find IOB tags in sequence. It is necessary to consider the previous predicted label along with the previous words information. Also, it is found most of the existing pre-trained word embeddings do not consider the parts of speech information. So generating word embeddings with respect to POS tags might help to improve accuracy of predictions.

So, I propose a sequence labeling tasks using LSTM RNN which considering previous label as well their state information that considers previous and next words. It will aid in successful identification of domain elements of CCM and will aid in automated model-generation for requirements analysis.

I would like to perform above task along with cross-validation in neural networks and publish a paper in ICSE 2018.

References:

1. D. Aceituna and H. Do, "Exposing the susceptibility of off-nominal behaviors in reactive system requirements," 2015 IEEE 23rd International Requirements Engineering Conference (RE), Ottawa, ON, 2015, pp. 136-145.