**PROJECT PROPOSALS**

**Note**: I am proposing three different ideas. Due to time constraints of workshop, it might not be possible to implement all the three proposals. I have written proposals in decreasing order of importance. So, if I need to choose only one project out of three, I would choose the first one.

**PROPOSAL 1:**

**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, we 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.

**PROPOSAL 2:**

**CO-REFERENCE RESOLUTION USING RECURRENT NEURAL NETWORK FOR REQUIREMENTS ANALYSIS**

It is often the case, in case of natural language requirements, there is inconsistency in using the terminology. In a given document, two different nouns can refer to the same component of the system. For example, in pacemaker requirements document, device and pulse generator refer to the same component. In order to automate the process of forming rules from requirements, it is essential to know if any two entities identified as component might refer to same component of the system. It is the same with acronyms and their full forms, which often happens in case of states of CCM model [1], a model used for model-driven requirements analysis.

In order to perform nominal as wells as pro-nominal co-reference resolution, we propose a co-reference resolution technique using LSTM recurrent neural network, where it holds states of previous sentences and words. It is essential to keep the track of some of previous sentences and words as the word mentioned in it sentence 1 might be the same component being addressed by pronoun in sentence 2 and 3. So, considering state information with respect to the sentences and words with in a sentence will be essential to perform better co-reference resolution.

**PROPOSAL 3:**

**SEMANTIC RELATIONS USING RECURRENT NEURAL NETWORK FOR AUTOMATED RULE GENERATION FOR REQUIREMENTS ANALYSIS**

Unlike determining relation between two words in a sentence, in order to generate a rule for Causal Component Model (CCM) [1], a model used for model-driven requirements analysis, semantic relations between domain elements of CCM, which can be a single word or multiple words or phrases is a difficult task. In order to find semantic relations, first we need to determine components, states and causes in a requirements document. Once the domain elements are determined.

We propose a LSTM recurrent neural network considering previous sentences as well as previous words with in current sentence is trained to determine relation of domain elements. In case of CCM, the relations that are of utmost importance to find are component-state relation, startstate-endstate relation, cause-state relation, component-cause relation, cause-cause relationship (there can be two or more causes referring to same transition).

So, by developing semantic relations using RNN, it will be beneficial to consider relations not only between words in a document but also words with phrases.

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.