**System Description - CWI SemEval 2016**

Team Name: **JU\_NLP**

***System 1 Name:*** **Random Forest Based System**

Description: We developed two supervised systems, which rely on the following features. We used features like number of hypernyms, hyponyms, synsets, syllables in a particular word. We identified the stopword, Named-Entity, and its part of speech tag (POS) as features. We also used a corpus of 5000 most used words in English Language[[1]](#footnote-1) to disambiguate whether a word is comprehensible based on its regular usage. We trained Random Forest classifier based on the above features.

Finally, we used post-processing technique to identify the complex scientific words adhering to biology, physics, and geography using the corpus collected from the web. We also used the dictionary of Non-English words to mark the words as complex. Again, we identified named entity to identify the named entity and these words were tagged as non-complex.

***System 2 Name:*** **Naïve Bayes Based System**

Description: In the above system, we just changed the classifier to Naïve Bayes.

1. <http://www.wordfrequency.info/free.asp> [↑](#footnote-ref-1)