Title: System Description - CWI SemEval 2016

Team Name: BHASHA
System 1 Name: SVM
System 2 Name: DECISIONTREE

Description

Both these systems are based on conventional classification algorithms trained over 'cwi_training.txt' data. They comprise of three stage pipeline, namely Pre-processing, Classification (SVM and DT) and Post processing.

Pre-processing- This stage comprises of 4 preliminary checks that directly classify the words based on rules.

1) NER: All the named entities are classified as non-complex.
2) Non-English words: All the words which are non-English have been classified as complex (used various corpora available in nltk package, including wordnet and words list from UMich "http://www-personal.umich.edu/~jlawler/wordlist").
3) Word's length: Words with length less than 2 are classified as non-complex.
4) Parts-of-Speech tag: Words having pos-tags CD,DT and TO as per Upenn Tagset are classified as non-complex (nltk default tagger has been used).

Classification- Both systems were trained on the features described below.

1) Frequency of words at positions [-2,-1,0,1,2] in training data. To tackle the problems with boundary words, each sentence is restructured as START + START + SENTENCE + STOP + STOP. Also all the punctuations in the sentence are removed.
2) Postags of words at position [-2,-1,0,1,2]. The postag value is quantized as it's occurrence frequency in training and test set.
3) Number of synsets with same postag as the word in sentence and the ratio (#word / #sum of all such synsets).
4) Word position
5) Word length, Vowels and Consonants ratios, # Syllables, Stem size and number of n-grams of characters (a-z) (300 most commonly occurring uni-,bi-,tri- grams for SVM and only 26 characters for DT)

Post-processing- This stage comprises of final check to remove the highly probable classification errors.

1) Character bigrams (402 character bigrams were present in training set) which were not observed in training data but is present in test data and has occurrence frequency less than 0.00005 are classified as complex. (Ex: 'zt', 'kz' etc.)
2) Words with occurrence frequency above 0.1% and labelled as non-complex (if label available) in training data but vice versa by trained model, are classified non-complex.