Indian Institute of Information Technology (IIIT) Chittoor, Sricity
Questionnaire for BTP-Progress Evaluation
Wednesdays, 3:30 – 5pm
Spring 2016

Project Title: AD003 Automatic reply, suggestions for mobile based dialogues: Humourable - Quotable

Faculty: Dr. Amitava Das

Group Members: Akshay Kumar (201301001), Pooja Makula (201301014)

1. Describe what you have done in the last 3 weeks.

Ans.
-The main focus of these 3 weeks was to improve the accuracy of the classifier and we have obtained an accuracy of **70%** (initially it was 34%). - **FOR THE ENGLISH SENTENCES CORPUS**
-We have started including hindi-english code mix data in our corpus and have collected conversations of whatsapp and have collected a corpus of 4110 code mix sentences and have annotated them under the 11 speech-act classes.
- We have also annotated and corrected 4941 english sentences (from 3 rounds of bootstrapping.)
- We have also added additional features for testing like(wh-features, ? , senti-words, apology-features, thanking-features, appreciation features) to the english data.
-We have annotated a total 9051 sentences(total english + code mix data) and obtained an accuracy of 53% **(FOR ENGLISH + CODE MIXED DATA).**
- We have submitted an abstract for AACL conference.
- Started writing a paper targetting Coling Conference.

2. How many times did you meet / talk with your faculty / guide?

Ans. We meet our advisor Dr. Amitava Das atleast twice a week. On Monday we meet as individual groups and discuss the progress and milestones whereas on Thursdays all the groups meet along with Amitava sir and discuss what they have done.

3. How many papers / articles / technical materials have you read in the last 3 weeks?
Ans. We have stutied about sentiwords and have read papers on humour and quote and have read amitav sirs paper on code mixing and techniques to find the CMI (code mixing index).

4. Provide a brief summary of your learning?
Ans.
- We have used the senti-words as additional features (we considered only those features that have positive and negative sentiment) and observed that adding the additional features has increased the accuracy of the classifier from 62% to 70%.
- Also we have observed that after adding code-mixed data the accuracy has come down to 53%. (as there are additional hindi words that are added in the features.)

5. What development / programming / practical activity did you do in the last 3 weeks.
Ans. We have modified the code of the classifier (have written a code in java) to add the additional 5 features (wh-features, ? , senti-words, apology-features, thanking-features, appreciation features) and have trained the system for individual as well as combination of these features.
-We also had to modify the code of the classifier as we have reduced the number of classes(Speech Acts) [Previously 43 Speech acts, now we have reduced them to 11 from 19].
-We have also written a code in java that gives the distribution of speechacts in the annotated corpus.
- We have also written a code to extract the senti-words and use them as features.

6. How close/far are you from the milestone set by your Guide?
Ans. -Our main goal during these 3 weeks was to improve the accuracy of the classifier. Previously we have obtained an accuracy of 35.3846 % (WEKA analysis) but on bootstrapping 3-4 times also there was no significant increase in the accuracy. We have bootstapped 3 times and corrected and annotated around 4941 english sentences and 4110 hindi-english code mixed data.

- We have also added additional features for testing like(wh-features, ? , senti-words, apology-features, thanking-features, appreciation features) to the english data and have obtained an accuracy of **70%.**
- We have then combined the english and code-mixed data and trained the system for 9051 sentences and obtained an accuracy of 53% (includes code mix data).

7. What specific challenges are you facing/you faced in the last 3 weeks?
Ans. The main problem we have faced was annotating the huge data set. We have annotated 4941 English sentences and around 4110 hindi-english codemix data of whatsapp conversations.

8. Propose your plan for the next 3 weeks as agreed with your supervisor. It would be verified in the next round.
Ans. We would start implication analysis on our present data set containing both english data and code mixed whatsapp data and continue experimenting on our corpus containing both english and code-mix data.