Project 4: Humor evaluation: judging funniess level of the dialogue in The Big Bang Theory

Hongfei Li(hl963), Shibo Zang(sz428)

1. Project Task

Commedy TV shows are playing a very special part of people's life nowadays. The key to the success of commedy Tv shows is the level of funnies of the themselves. Our task is to use natural language processing technics to evaluate comedy script's funnies level. We may evaluate the funnies level on a scale of 10.
One example from a popular show Big Bang Theory is the following

```
Raj: I don't like bugs, okay. They freak me out.
Sheldon: Interesting. You're afraid of insects and women. Ladybugs must render you catatonic.
```

Clearly the second sentence is the laugh point. Our system will look into the context of the laugh point and may give the laugh point a funniess level of 8/10.

The reason we choose to investigate in this is because we love watching Big Bang Theory and it is the TV series that is making the most money. We want to look into something that can be potentially useful and practical.

In sum, the input is text corpus selected from commedy show The Big Bang Theory.And the out put is the funny level of a certain sentence.

2. General Approach

The main task of our project is to use NLP technics to evaluate comedy script's funnies level. The general idea we propose to solve this task is in some sense similar to Professor Cristian Danescu-Niculescu-Mizil's work on judging politeness of corpus. We would like to break our task into several subtasks:

1.. Humor Category Classifcation:
Following examples illustrate the type of linguistic elements that underlie humor.

Example	Category
“What do you use to talk to an elephant? An elly-phone."	Phonological Similarity
“MIT stands for Mythical Ilogy."	Acronym multiple sense
“Infants don't enjoy infancy like adults do adultery."	Pun
“Of all the things I lost, I miss my mind the most."	Human centric vocabulary
“Money can't buy your friends, but you do get a better class of enemy"	Negative Orientation
“It was so cold last winter that I saw a lawyer with his hands in his own pockets."	Professional Communtities
“Wonderful weather we are having." (While the weather is terrible actually.)	Irony

2.. Predicting Funniess Level
We plan to apply machine learning algorithms to this task. Mainly there are three levels that needs to be considered: not funny, neutral, and funny. And with classification algorithm, we could first train the training dataset produced by turkers' annotation using bag of word (BOW) classifier or linguistically (Ling.) informed classifier. The BOW classifier is based on Support Vector Machine using a unigram feature representation. And the linguistically informed classifier is an SVM using the linguistic feature.

The reason we decide to divide training dataset into multiple categories before applying machine learning algorithm on the dataset directly is that we find there exists different linguistic relationship under each category. For example, following examples are scripts taken from some episode of The Big Bang Theory.

Phonological Similarity

```
Howard: Watch this, it’s really cool. Call Leonard Hofstadter.
Howard’s phone: Did you say, call Helen Boxleitner?
Howard: No. Call Leonard Hofstadter.
Howard’s phone: Did you say, call Temple Beth Sader.
Howard: No.
Leonard: Here, let me try. Call McFlono McFloonyloo. Heh-heh.
Howard’s phone: Calling Rajesh Koothrappali. (Raj’s phone rings).
Raj: Oh, that’s very impressive. And a little racist.`
```

Acronym multiple sense

```
Leonard: Why do they say AA?
Sheldon: Army Ants.
Leonard: Isn’t that confusing? AA might mean something else to certain people.
Sheldon: Why would a physics bowl team be called anodised aluminium?
Leonard: No, I meant…. never mind. Hey, check it out. I got you a Batman cookie jar!
```

Pun

```
Sheldon: Interesting! So it went beyond the mere fact of coitus to a blow by blow account, as
Amy: Pun intended?
Sheldon: I'm sorry, what pun?
```

Irony

```
Leonard: Hope you're hungry.
Sheldon: Interesting, a friendly sentiment in this country - a cruel taunt in the Sudan.
```

From the above dialogues, we take pun and irony as an example. We can see that the specific word matters while using pun to show humors, while ironic humor always relates to surrounding context. This informs us that we need to be aware of how different types of humor is created and how should we design classification method to recognize them.

3. Data and Data Annotation

We will get all transcrips of every epsiode of a commedy TV show such as Big Bang Theory, which can be found at *https://bigbangtrans.wordpress.com*. Mark all laugh point with a label as well as a funny score associated with a label <lp,9> using amazon's mechanical turk (9 means this laugh point has a funnies of 9 out 10). Since The score is subjective score, we may have 5 or more people rate the laugh point and take the average of them. Each turk is going to read through the transcript of one episode and rate the laughing points.

For example, for the following transcript which is the first couple of lines in season 1 episode 2 of Big Bang Theory. The label are added after the turk's work.

```
Leonard: There you go, Pad Thai, no peanuts.
Howard: But does it have peanut oil?
Leonard: Uh, I’m not sure, everyone keep an eye on Howard in case he starts to swell up. <lp,8
Sheldon: Since it’s not bee season, you can have my epinephrine. <lp, 7>
Raj: Are there any chopsticks?
Sheldon: You don’t need chopsticks, this is Thai food.
Leonard: Here we go.
Sheldon: Thailand has had the fork since the latter half of the nineteenth century. Interestin
```

In all the text corpus we need is a dialogue. The purpose is to find lingustic relationship between sentences to design a system to rate the humor level of specific language.

The project will require a huge amount of dataset, We could get these text corpus from Internet.

4. Methods and System Development

4.1 Preprocessing

The first thing we should think about in our system is the analysing scope. Treating each individual plot as an analysing part is resonable. In other words, we only consider the relationship between sentences within a plot while evaluating humor levels.

Thus, the preprocessing part is straightforward. First, we let turkers split one episode into multiple plots basing on the information appears in the show. For example, like The Big Bang Theory, there is a cut scene between two plots. What we need the turkers do is to divide the scripts into several parts basing on the time when cut scene is showing. Second, we will ignore those plots with very few dialogues, e.g. less than three senteces. At last, we will let the turkers make judgement on the scripts about how they think the humor levels is.

4.2 Categorization

Similar to the genre classification, we could implement the humor categorization method using machine learning approaches. For example, we could use Naive Bayes classifier to train a model to distinguish humor type. Since feature selection is critical to Naive Bayes classifier, we would like to different feature set that is currently proposed to implement the classifier we want.

4.3 Model Training

After proprocessing and categorization part, we will let turkers to rate the humor level of each sentence given a certain text corpus. With these ample training data set, what we need is to train another classifier to rate other text corpus accordingly. I recommend using SVM this time because there seems to be huge amount of features going in our system.

5. Implementation

We plan to use nltk library to implement word tokenization and POS tagging. However, most part of our system needs to be implemented from scratch since we doesn't find related and outstanding library that can be directly used in our project.

6. Evaluation

Since our the goal of project is to evaluate how funny a transcript is, we can evaluate the accuracy based on human inputs we gathered from the amazon mechanic turks. We are going to run a k-fold cross validation with Mean Absolute Error (MAE) to evaluate the accuracy of our system. MAE is a measure of the deviation of the scores given by our system from their true user-specified values. The equation of MAE is the following, where p is the score given by our system, and q is the averaged user ratings.

The lower MAE we get, the more likely our system is rating laughping points like a real human being.

7. Future Work

8. Individual Member Contribution

Hongfei Li: Part1, 3, 6, 7 Shibo Zang: Part2, 4, 5