NLP - PROJECT

TOPIC AND FOCUS IDENTIFICATION

TEAM: Whatever_it_takes

20171083 - Parijat Jain
20171104 - Eesha Dutta

For English

* We begin with reading papers, theory, and understanding the basic linguistic concepts behind Topic and Focus. Also read ~~upon~~ centering theory and how it is applied for co-reference/anaphora resolution and topic and focus identification
Also read/understand tutorials of Snorkel.

* We then scrape the articles on which we will be doing the analysis. As discussed (will be further instructed), we will be scraping news/NPTEL articles (around 20) and then will study the data and apply the above learnt concepts to label the data. (Strongly supervised)

* Along with this we also use 'some' knowledge base and write some Labelling functions to train our Snorkel Model and then we will use the model to generate automatically labelled data. (Weakly supervised)

* With the combination of this ~~strongly~~ & weakly supervised annotated data, we then use existing tools / implement some of the research papers to generate results on test data.
Here we will make a model and feed test data to obtain results

Dipti Misra 5/10/19

Some of the doubts!.

1) Is this a machine learing based on rule based approach?
2) If this is ML approach, will 20 articles suffice?
3) If this is rule based then why are we using snorkel (as given in the problem definition) to generate ~~test~~ train data?
4) will the train data / articles be provided to us?
5) Can you please suggest any papers / tutorial or reading material to ease into the theory.

For Hindi

* The basic approach of reading and understanding will be same.

But since there are complications involved in developing models ~~and~~ so the basic aim for Hindi will be to annotate / label the data (articles) and generate a dataset.