Modeling Complex Social Behavior: Topic Control

-Akash Ahuja

DATASET:

- 10 tagged conversations(xml files).
- Ground Truth given.

GOAL:

- To calculate the Topic Control ranks of participants based on automated modeling and comparing it with the ground truth data.

INDEX CONSTRUCTION:

Five indices of Topic Control were chosen and their method of extraction is as follows:

1. Named Entities:

- Stanford Named Entity Recognizer was used to tag the named entities in the dataset.
- 7 class caseless classifier was used to classify the named entities in following 7 tags: Person, Organization, Location, Date, Money, Percent & Time. This classifier was chosen to account for all entities starting with either case.
- The named entity use for each participant was computed using the XML DOM parser in Java.

2. Turn Length:

- Turn length is the average number of words spoken per turn by each participant.
- This was calculated with the help of “speaker” tags and turn text content.

3. Local Topic Introductions:

- Local topics are topics first introduced by some participant and subsequently mentioned by other participants.
- Local topics were calculated with the help of “topic” “speaker” and tags and xml dom parser.

4. **Turn Count:**
    - This is the number of turns per participant.
    - This index was calculated with the help of “turn” and “speaker tags” and xml dom parser.

5. **Cite-Score:**
    - Cite-score is the number of times each participant is cited by other participants.
    - This was calculated with the help of “link-to” tags and xml dom parser.

Methodology:

- All the indices for each dataset were calculated and participants given ranks for each index(higher rank means higher topic control)
- Correlation between ranks of participants based on each index and rank based on ground truth was found.
- My aim was to get good enough accuracy over all datasets using one standard weighting scheme instead of changing the weighting scheme for each dataset.
- So I calculated the average of correlations for each index across all datasets.
- Then I normalized the average correlation scores and considered normalized scores as weights.
- I then combined the scores to calculate weighted topic control scores.

Analysis of Results:

- Named Entity had the highest correlation with the ground truth data.
- Topic intro, cite-score and turn count came next showing almost comparable correlations.
- Turn length showed the least correlation.
- The following weights were given to indices:
  named entity: 0.25
  topic-intro:0.2
  cite-score:0.2
  turn-count:0.2
  turn-length:0.15
- These weights were multiplied with the participant ranks and sum of these products for each index gave weighted final scores.

Results and Accuracy:

- **The results for one of the datasets is:**

	Weighted score	Ground truth rank	Calculated rank
meg	6.25	1	1
george	5.3	2	2
mara	5.15	3	3
nick	4.8	4	4
amy	3.3	5	5
michelle	1.7	6	7
john	2.5	7	6

- **Accuracy: 71.4%**
- The same weighting scheme was applied to other datasets as well.
- Weighted scores of 7 datasets calculated.
- 100 % accuracy was obtained on two of the datasets.
- The average of accuracy over all seven datasets was found to be : **62.84%**
- All calculations regarding correlations and weights are shown in the excel sheet attached with the report.