10/29/2019

Today’s Goal:

1. Get inter-rater reliability metric for Tangram HHH corpus
   1. Collected Summer 2019 from Johnson, Coker, and Avery
   2. What is the interrater reliability metric?
      1. Metric of homogeneity between different raters
      2. Intra-rater reliability – metric of homogeneity within (or for the same) rater
      3. Metrics: Joint probability agreement, kappa statistics (Cohen’s kappa for 2 or ***Fleiss kappa*** > 2), Correlation coefficients, Intra-class correlation coefficient
   3. What tools can one use to get it?
      1. NVivo
   4. What are the results from this dataset to use as a baseline?
      1. I just used a pseudo-WER (count assertations, deletions, removals) / total words
2. **Why** do you need the inter-rater reliability. To extract the utterances that have human agreement to put into the “lexical text classification algorithm” as a “global and realiable” corpus
3. This I did
   1. Use ELAN 🡪 export tab-delimited transcriptions 🡪 Got this from .eaf files
4. ISSUES:
   1. Imbalance in transcript utterance count
   2. Avery (11,961) – Coker( 11,969) – Johnson (11, 943)
   3. So Johnson has 18 less than Avery and 26 less than Coker and
   4. Coker has 8 less than Avery
   5. Working on understanding why

10/30/2019

Today’s Goal:

1. Keep working on the Inter-rater reliability metric
2. Follow-up/Think about metrics for Interruption Timing Comparison

To do:

1. Find all the common utterances between raters (process)

11/6/2019

Goal(s):

* Get an “agreeability” metric for Tangram\_HHH data collection
* Use transcripts with agreeability over a threshold for “Lexical Interruption” input
  + Agreement strategy: 2 out of 3 only
  + What percentage of certain words are used by ALL INTERRUPTERS?
  + What percentage of words are used by INDIVIDUALS?
  + Can we get a similar metric

Work

Process the transcriptions

1. All lower case
2. Get rid of special characters
3. Get rid of special characters
4. Cat = cats (get ride of plural)

Accomplishments

1. Did everything in the “Work” section

11/7/2019

Objectvive

1. Get the transcriptions with 100% agreemtn
   1. So 2 /3 agree on all the words
2. Put these transcriptins into the following groups
   1. Tb vs nontb
   2. Int vs non int
   3. Tb vs int
   4. H vs c
   5. Data Groups [ALL, Team, Person]
   6. Maybe Histograms 1st

Work

1. Got WER code and found an agreement metric
   1. Take the WER between two annotations and if the 1 – WER is 100%, usable transcription (Use of 92% or 7% decrease of the transcriptions (11,943 ) 🡪 11,069). 11,943 is common across all transcribers and 11,069 is count of 100% agreement between 2 or more transcribers.
   2. Get the average agreement between transcribers and take an average across all (approximately 96% agreement)

Accomplishment’s

1. N/A

11/8/2019

Objectives

1. Graphs to produce (Interrater Reliability for Tangram Transcriptions)
   1. Utterance Length vs Avg. Agreement (no correlation) \*\*
   2. Who agreed more? (Avery-John, John-Coker, Coker-Avery) – bar graph
   3. Average agreement (pie chart)
   4. Overall agreement (> 100%, >95%, >90%)
2. Lexical Analysis of Tangram HH dataset
   1. Dataset 2 of 3 100% agreement on Tangram\_HH utterances
   2. Research Question (1): If there a significant improvement in the identification of task boundaries in comparing: all data vs. team data vs individual
      1. Hypothesis (11): Significant difference in individual utterances as a predictor of a task boundary because communication is “individualized (people do not consider their interlocutor when using confirmation cues). This could be contradicted by Hyp (2).
      2. Hypothesis (12): There may be a trade-off between the amount of train-test and the performance of each model (all, team, individual). All > Individual > team and the performance may correlate with the train data
   3. Research Question (2): At what level are we able to better discriminate interruption decisions vs non interruption decisions based on lexical information (all, interrupter, team, individual). ALL: Is there a general model that best discriminates these two classed? INTERRUPTER: Does the interrupter listen for certain lexical cues to discriminate a time to interrupt vs not ; TEAM: Is there some common used with a team that helps the interrupter know when to disseminate interruptions based on the collaborative team communication? INDIVIDUAL: Do individuals use different language that discriminates when an interruption was send vs not
      1. Hypothesis (21): There is a significant different in discrimination of when an interruption as sent vs not because interrupters have individual differences that may not being generated in the ALL model. May still have the data amount vs accuracy issue.
      2. Hypothesis (22): Difference from RQ 1 because the interruption times are based on the interruptions perception of the primary task. Whereas in RQ 1 , the lexical content within the context of the task boundary is with respect to the primary task and from the participants’’ use of language to illustrate they are finished with a task\*\*
   4. Research Question (3): At what level are we able to discriminate human vs computer interruption decisions (all, interrupter, participants, team)
      1. Same as RQ 3
   5. Research Question (4): At what level (all, interruption, partcipants, team) can we discriminate TB vs INT decisions. Is there a difference in the lexical content preceding a task boundary and human interruption decisions
      1. Based on previous work that shows that 52% of the interruption decision humans make precede task boundaries, this may not be easily discriminated but useful to explore to see if there is additional content prior to interruption times that are different from task boundaries.
      2. Potential errors\*\* Incorrect logging for PTT and TB timestampes (experimental flaw), but also collaborative nature of a task boundary (e.g. p1: got it, p2: ok) so we log ok as task boundary utterance. (SOLUTION: Use two/three step back analysis window of task boundary for “ok” only)
3. COULD POTENTIALLY LOOK AT WHO THE ITNERRUPTION WAS INTENDED FOR

Work

1. Code to put files into ALL, TEAMS, SPK, and INTERRUPTER
2. Stopped to process data for CC-IT (1330)
   1. Results – there are some bugs in the data processing so I will wait until Wednesday

Accomplishments

1. Processing text of Tangramm\_HHH dataset (validated by ASR, Subject validation, and 3 transcribers)
2. Code: N:/Peters/Tangram/scr2018/HumanInterruptionStrategies/LexInterruptionClassification/TextClassification/TextClassificationDataPrep.

12 November 2019

Objectives

1. Change textclassification.m

Work

Accomplishments