11719 Assignment 1b

Seungwhan (Shane) Moon

Feb 8, 2014

1 Selection of the Model and Episodes

In this section, I describe how I selected the 2 episodes for Jim Gee Style analysis.

1.1 Model and Task

In the previous assignment, I have found out that the block HMM model significantly outperformed other models on the prediction task in the Heteroglossia framework. Figure 1 shows the results from the previous experiment with varying window sizes (k).

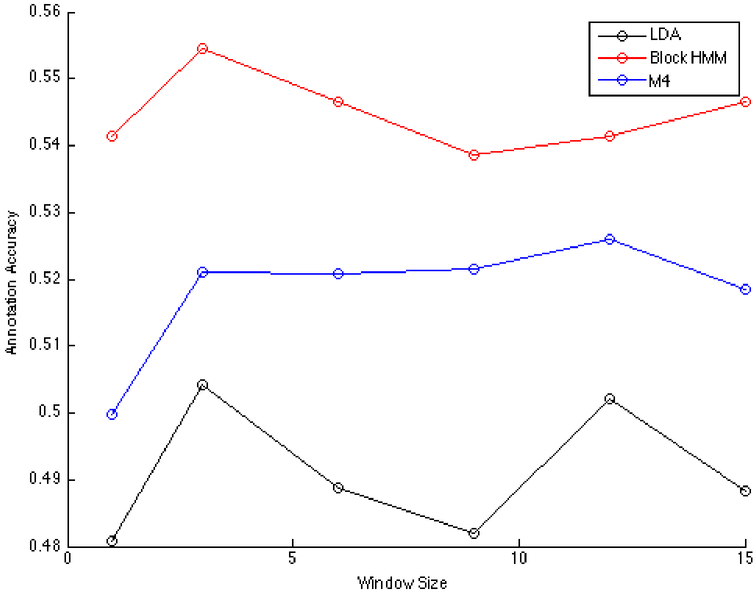

Figure 1: Experiment 2 (Heteroglossia framework). The Y-axis denotes the annotation accuracy (F1 score), whereas the X-axis denotes the window size (k) of the turns that we consider.

Based on this experiment, I choose the Block HMM model as the topic model and the annotation task in the Heteroglossia framework for ranking the episodes.

1.2 Ranking Episodes

I then perform the same experiment on each episode, treating each episode as a held-out test data. Each episode is tested by the same learner trained on the entire data, in order to reduce any variability related to the training data. Figure 2 shows the results.

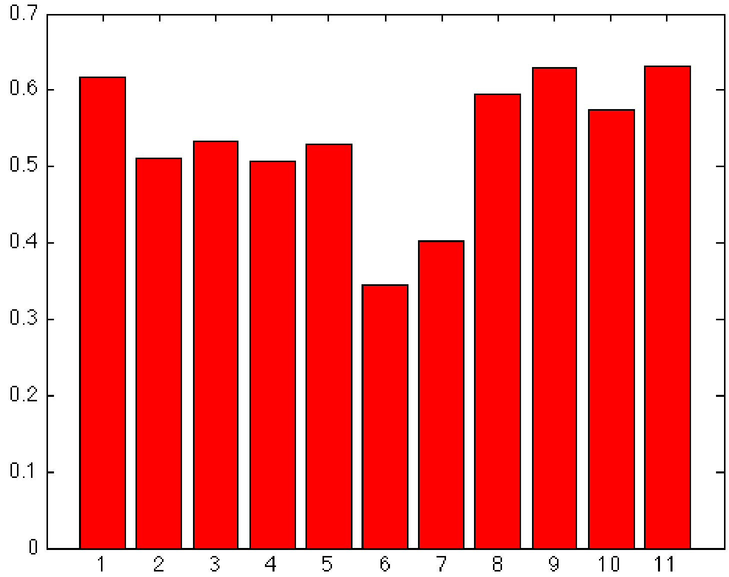

Figure 2: Annotation prediction task on each episode in the Heteroglossia framework. The Y-axis denotes the annotation accuracy (F1 score), and each bar represents each episode. ($k = 1$)

It can be seen from Figure 2 that the learner produces the best result on Episode 11 ($f_1 = 0.63$, lines: 820 - 964. *it turned out that Episode 11 actually has two episodes. The second part was not indicated with parent ID = -1.), and that it results in the worst performance on Episode 6 ($f_1 = 0.34$, lines: 297 - 372). This is a non-trivial difference, and thus I expect that Episode 6 would contain significantly more complex or subtle forms of language interactions when analyzed in the Jim Gee style. I test this hypotheses in the next section.

2 Jim Gee Style Analysis

My Jim Gee Style Analysis on Episode 11 and Episode 6 are attached in the appendix.

3 Reflections

I have hypothesized that the episode in which the classifier performed the best on the classification task would have a simpler and more obvious language style, thus making the Jim Gee style analysis easier.

It is interesting to note that Episode 6 (which the classifier had the worst performance in prediction task) was in general a much better class than Episode 11, and its students exchanged a lot of meaningful and constructive questions to each other (e.g. ”Okay, I think that when the distilled water is put into the glucose the glucose will get lighter and then well I don’t know what then.....”, ”wouldnt for B, the water will once again shrink because its mixing water with glucose again?”). In Episode 11 (the best performance), on the other hand, most of the conversations were about students identifying each other, or about students complaining about the materials of the class environment being confusing.

In order to figure out what comprised of the differences in the classification performance between the test on Episode 11 and the test on Episode 6, I delved more into the ground-truth label distribution of each

Table 1: Classification report on Episode 11 (lines: 820 - 964)

Label	precision	recall	f_1 score	support
M	0.56	0.58	0.57	48 (33%)
HE	0.28	0.90	0.43	10 (7%)
N/A	0.82	0.59	0.69	86 (60%)
avg / total	0.70	0.61	0.63	144

Table 2: Classification report on Episode 6 (lines: 297 - 372)

Label	precision	recall	f_1 score	support
M	0.43	0.35	0.38	26 (34%)
HE	0.08	0.75	0.15	4 (5%)
HC	0.00	0.00	0.00	0
N/A	0.71	0.22	0.34	45 (60%)
avg / total	0.58	0.29	0.34	75

episode and the classification performance per each class. Tables 1 and 2 show the results.

Note that the test on Episode 6 shows a significantly poorer recall performance on the prediction task of the label (N/A) than the test on Episode 11. I suspect that this is because Episode 11 contains significantly more trivial questions that are labeled as (N/A), namely the "who is xxx" questions. Episode 6, on the other hand, exchanged a number of more complex conversations, often involving richer exchanges of ideas, agreeing, answering, statement of information, encouragement, etc. The participants of Episode 6 seemed to be more respectful to the tutor and to each other.

The fact that the classifier failed to perform well (f_1 score = 0.34) on a "good" dataset (Episode 6) which contains richer components of the building blocks in the Jim Gee framework, discourages me from thinking that the topic models are finding good patterns about the text. Although I also believed at first on a conceptual level that there are some associations between language features and what LDA is doing, I conclude that I cannot confirm such associations at least with the experiment that I performed here.