Annotation Task 1

Topic segmentation & labeling

Group 19:

- Parth Rajkumar Sonkhia - UBIT: psonkhia
- Rahul Yadav - UBIT: rahulyad
- Romina Marazzato Sparano - UBIT: rominama
- Sandeep Kunusoth - UBIT: skunusot
- Smarana Shrikant Pankanti - UBIT: smaranas
- Sungwook Sohn - UBIT: ssohn3

I) Segmenting and labeling topics (5p)

1. In your experience segmenting the text into topics, how clear were topic boundaries? What information did you use to help you identify boundaries? What was hard about identifying boundaries?

Topics boundaries were not always clear because of missing data in the transcription, context missing in the segments presented, swift changes between topics, and lack of clarity in transitions.

We paid attention to keywords related to a stated topic. For example, when they started talking about TV Shows, we considered this a hyper-topic and the names specific shows as separate topics. For a single TV show, different aspects were annotated further such as cancellation, preferences, watching, schedule conflict, etc. We also used agreement words such as yes, uh-huh, yeah, as clues for interlocutors staying on topic.

Identifying boundaries was difficult for the reasons listed above, so we made decisions such as basing an annotation on the majority of the topic in the exchange, even when more than one topic was present.

2. How much overlap in topic segmentation did you find there was across annotators?

We found some overlap, specially when specific shows were mentioned, topic boundaries were clearer. When topics overlapped, some annotators decided on shorter sections and some decided on longer sections, which is an indication that topics boundaries were less clear.

3. To what extent do you think topics can be found in adjacent segments? Do you think this could vary across different text types?

Adjacent segments show topic overlap in our corpus. When the sentence contains a topic, we can use the topic named to identify topics in the next sentence. For instance, when a broad topic was initiated, we could identify subtopics after, as in TV Shows, and specific shows (Golden Girls, Home, etc.)

This meant overlap in subsequent turns where a swift change of topic led us to annotate the topic as different. The text type can certainly affect topic overlap. Register can make an impact in using clearer outlines, whereas informal text or speech leads to fuzzier boundaries.

4. What criteria did you use to come up with labels for topics?

We used hypernymy and hyponymy to label topics such as TV Show and then the name of the show. Additionally, we used frames to identify aspects related to the main topic, such as watching, schedule conflict, preference, etc.

At the start, we starting to describe topics at length, such as "discussing task topic - TV shows." We quickly realized that we could clean up the labels and use the hypernym alone instead. We would need to go back to our data and do this cleaning, so for the example above we would use "TV shows." We therefore decided to use the same lengthy version to make the clean-up easier at the end.

5. Did you all come up with similar labels? For which topics did you find the most/least agreement?

We were able to agree on topic labels. For annotations including different topics, we had to negotiate a bit to settle on a specific label. Before starting the labeling, we discussed a strategy, which allowed us to move forward rather quickly. For example, we decided to change topics for specific shows versus keeping them under the main topic of TV Shows.

The most agreement was found on labeling segments about specific shows; least agreement was for labeling longer segments that covered multiple topics.

II) Annotation and Machine Learning models (4p)

1. If we were to set up a human-in-the-loop model, where the model queries an annotator about specific annotation instances, how could a statistical model help annotators make decisions?

A statistical model can help annotators in providing probabilities for each possible annotation to be mapped to a topic, obtaining input from other annotators, finding agreement based on the number of keywords overlapping from different annotators.

2. What do you think are the biggest challenges for annotating language datasets for ML models?

For different annotators, the level of specificity may be an issue. This is in part due to ambiguity, as annotators may interpret text in different ways. Also, annotations involve a lot of subjectivity as different opinions may fit a given segment. If annotations are too broad, applying to many segments, the model will not be able to discriminate at a level of specificity needed for good performance.

On the other hand, disagreement or extreme specificity can lead to wide variation in the training data. Wide variation in topic labels would translate to low probabilities and, thus, a poor performing model.

Another issue is time as annotating large language datasets can be time-consuming and labor-intensive. Annotators may need to be trained intensively to annotate accurately.

Bias can also be an issue as annotators may introduce their own views and preferences.

3. What kind of model do you think would benefit from the data you annotated?

Our annotations can help a topic classification model, a topic summarization model, a recommender model, a segmentation model.

4. How would you (partially) automate these tasks?

We can segment interventions using break words and calculating semantic similarity between sentences.

After segmentation, we can use keyword extraction. If we have a huge dataset, we can name entity recognition to tag a segment with a specific topic label.

III) Annotation interface (2p)

1. Discuss advantages and disadvantages about the way the tasks were set up. You may consider:

a. An alternate setup for segmentation in which longer sections of the conversation would be shown per screen. Would this be beneficial? How would you decide how many turns to show? What would be possible effects on annotators?

b. An alternate setup for segmentation in which the task is other than a binary decision task (same vs new/different topic). What could that task be?

We could show the segments at hand in context by showing in the same window the preceding and following segments. Annotators can use the context to guide the segmenting and labeling processes.

The binary decision task could be diversified into including:

- Same topic
- New/different topic
- Subtopic

We discussed the possibility of including a Random or Skip Segment category for removing, during this stage, segments unrelated to the task at hand so as to provide a cleaner dataset for labeling.

2. How would you improve the annotation interface?

- During segmentation, we could show the segments at hand in context by showing in the same window the preceding and following segments.
- During labeling, we could show the segment at hand in context by showing in the same window the preceding labeled segment.
- If the segments are large, we can show a fragment of the previous segment and its label to help in deciding the label for the segment at hand.
- We can add an autocomplete feature to reuse topic labels.
- The labels can be saved in a nested structure to show the relation between Topics and Subtopics.
- At the end of the task or after saving, we could add an option to review, and edit if necessary, the annotations.
- We can provide the ability to review group members' work and provide feedback.

The authors confirm contribution to the paper as follows:

Parth, Rahul, Romina, Sandeep, Smarana, Sungwook.