245 Assignment 2

Shriya Sravani Y
UCSC
sy4@ucsc.edu

February 05, 2025

1 Abstract

This assignment is based on dataset exploration and annotation. The two datasets used here are Ubuntu Dialog corpus and MultidoGO.

2 Introduction

The assignment focuses on dataset exploration and annotation. In this assignment, I analyzed two dialogue datasets, Ubuntu Dialog Corpus and MultiDoGo, by performing manual annotation using the SWBD-DAMSL annotation scheme. The datasets were manually labeled according to the utterances, and were used to measure the inter-annotator agreement using Cohen's Kappa (for pairwise comparisons) and Fleiss' Kappa (for overall agreement among all annotators).

3 Datasets

3.1 Ubuntu Dialog Corpus

- It is a large collection of technical dialogues related to Ubuntu troubleshooting.
- The conversations were more goal-oriented and technical.
- The structure is relatively consistent, with many problem-solution patterns.

3.2 MultiDoGo

- It is a multi-domain goal-oriented dialogue dataset and contains various real-world task-oriented conversations.
- The dialogue topics were more diverse, including customer support, reservations, and software queries.

4 Annotation Process

Labels were normalized by converting them to lowercase and stripping whitespace to ensure consistency. The annotation scheme used was SWBD-DAMSL, which categorizes dialogue utterances based on discourse functions.

5 Inter-Annotator Agreement Calculation

We computed:

- **Cohen's Kappa** for pairwise comparisons between annotators.Cohen's Kappa is used to measure agreement between two annotators while accounting for the agreement occurring by chance. It is calculated as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

  where:

  - P_o is the observed agreement (proportion of times the annotators agree).
  - P_e is the expected agreement due to chance.

  Cohen's Kappa is useful in our assignment because we have pairwise agreements (e.g., Kiara vs Jack, Kiara vs Shriya, Jack vs Shriya), and we need a metric that adjusts for agreement occurring randomly.

- **Fleiss' Kappa** to assess agreement across all three annotators.Fleiss' Kappa is an extension of Cohen's Kappa that applies to multiple annotators. It measures the overall agreement among N annotators across k categories. The formula is:

$$\kappa = \frac{P - P_e}{1 - P_e} \tag{2}$$

  where:

  - P is the mean observed agreement across all annotators.
  - P_e is the mean expected agreement due to chance.

  Fleiss' Kappa is used in this assignment because we have three annotators labeling the same utterances, and we need a single metric to quantify their agreement.

6 Comparative Analysis of the Corpora

6.1 Objective Observations

- The Ubuntu dataset showed higher agreement across annotators, with a Fleiss' Kappa score of 0.36, compared to 0.22 for MultiDoGo.

Dataset	Cohen's Kappa (Kiara vs Jack)	Cohen's Kappa (Kiara vs Shriya)	Cohen's Kappa (Jack vs Shriya)	Fleiss' Kappa (All Annotators)
MultiDoGo	0.30	0.21	0.33	0.22
Ubuntu	0.36	0.42	0.51	0.36

Table 1: Inter-Annotator Agreement Results

- MultiDoGo had lower agreement, likely due to the diversity of domains and more varied dialogue structures.
- Pairwise agreement was higher in Ubuntu, indicating more consistent interpretation of utterances.

An example where all annotators agreed in the Ubuntu dataset was "do I have to manually delete the files it created" and an example where there was disagreement was "yes it was complaining i didnt have lpd."

6.2 Subjective Observations

I found the Ubuntu dataset easier to label because of its structured format and simpler text.

The diverse and varied nature of conversations in MultiDoGo made it harder to annotate as a text could have more than 1 annotation. It was more realistic too. There were more repeated phrases in the dataset.

MultiDoGo had dialogues of different context and was more broader in terms of topics covered, whereas the Ubuntu dataset was more technical in nature and was overall focused on one or two topics.

Based on the annotations made by my team mates, Kiara and Jack, as well as the inter-annotator values, I will say that the Ubuntu Dialog corpus will be better to use to train a tech support chatbot due to its consistency. For a multi-domain dialogue system, the MultidoGO dataset is a better option for more robust performance.

7 Improvement Recommendations

Some methods to improve the results can be to provide clearer definitions and more examples for each label.

In the annotations, some texts were given more than 1 tag, which could have also reduced the inter-annotator values. If these ambiguous categories are reduced, it can avoid confusion.

8 Conclusion

This assignment highlights the differences in annotation agreement across two dialogue datasets. The Ubuntu Dialog Corpus appears to be more consistent for annotation and may be preferable for training a dialogue system in a technical support setting. On the

other hand, MultiDoGo presents challenges due to its diversity but remains valuable for multi-domain dialogue research. Further refinements to the annotation process could enhance agreement and improve dataset reliability for training dialogue models.