245 Assignment 2

This assignment is about exploring some datasets, by doing some annotation. Make a group of 3 people. Everyone in the team should annotate the items so that you can calculate inter-annotator agreement.

Here the two datasets:

Ubuntu Dialog Corpus

<https://www.kaggle.com/datasets/rtatman/ubuntu-dialogue-corpus?resource=download>

MultidoGO  
Address of data should be: <https://raw.githubusercontent.com/awslabs/multi-domain-goal-oriented-dialogues-dataset/master/data/unannotated/software.tsv>

Here is the SWBD-DAMSL annotation scheme to use for annotation:

<https://web.stanford.edu/~jurafsky/ws97/manual.august1.html>

  Annotate the conversations shown in the table below.

You will need to download Unbuntu data, this will result in a folder named “dialogs” table shows directory structure inside of /dialogs

|  |  |  |
| --- | --- | --- |
| **Conversation ID** | **# turns** | **Dataset** |
| acs-00ad5526-bb84-4148-ac30-19afca56c12b-1 | 22 | MultidoGO |
| acs-629291d6-ca2b-4b63-aaf7-012ef4f602c9-1 | 14 | MultidoGO |
| acs-b5a66197-127c-4bd1-8946-8813b9cbdeb8-1 | 14 | MultidoGO |
| acs-1d1dbfa6-4e53-44cc-b748-5e3a44b92508-1 | 13 | MultidoGO |
| acs-4463829b-6d87-477a-a882-d6cd78a6056b-1 | 15 | MultidoGE |
| **Total MultidoGO turns** | **70** | MultidoGO |
|  |  |  |
| dialogs/99/92.tsv | 99 | Unbuntu |
| **Total Unbuntu turns** | **99** | Unbuntu |
|  |  |  |

1. Calculate your inter-annotator agreement after reading this paper:  
   <https://aclanthology.org/J04-1005.pdf>
2. Write a report on your annotation results. Compare the corpora. Give both objective performance information and more subjective observations (e.g. one corpus was easier/harder to annotate, one corpus seemed more/less realistic, varied, etc.) Include your inter-annotator agreement for your group.  Which dataset do you think would be better for training a system in this domain? How could you improve your annotation results or process?