NLP 245

Assignment 2

This assignment is about exploring some datasets, by doing some annotation. Find 1-2 people to partner with. You and your partner(s) should annotate the same items so that you can calculate inter-annotator agreement.

Here are two datasets:

Ubuntu Dialog Corpus
https://www.kaggle.com/datasets/rtatman/ubuntu-dialogue-corpus?resource=download

MultDoGO
https://github.com/awslabs/multi-domain-goal-oriented-dialogues-dataset/tree/master/data/unannotated
Use the software domain.

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

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

1) Each person should annotate 2 Unbuntu dialogs and 10 MultiDoGO dialogs. (Unbuntu dialogs are much longer than MultiDoGO dialogs)
2) Calculate your inter-annotator agreement after reading this paper: https://aclanthology.org/J04-1005.pdf
3) 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 and list who your partner(s) are. Which dataset do you think would be better for training a system in this domain? How could you improve your annotation results or process?

Due ~~18 April 2023~~
21 April 2023

Revised Instructions for item #1 of assignment:
Annotate the conversations shown in the table below with a total of approximately 200 turns for each dataset:

MultidoGO
Address of data should be:

https://raw.githubusercontent.com/awslabs/multi-domain-goal-oriented-dialogues-dataset/master/data/unannotated/software.tsv

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
acs-bde6434f-e72b-493e-8950-429c2a9de2b4-1	17	MultidoGO
acs-786b69fa-a2bd-4b27-99b5-8b2111cd1bc8-1	15	MultidoGO
acs-658105cf-4abe-4d69-81d0-5501cd6afa2b-1	18	MultidoGO
acs-4f6b78f1-f373-4943-9cdd-022d04f3d889-1	18	MultidoGO
acs-ac86e19d-1ab6-43af-8dfb-1477d0ba727f-1	19	MultidoGO
acs-88a7c284-3312-469c-9880-38275cca39e0-1	16	MultidoGO
acs-b858a682-e084-4dda-8804-1df04b44d870-1	20	MultidoGO
Total MultidoGO turns	**201**	MultidoGO
dialogs/99/92.tsv	62	Unbuntu
dialogs/99/9.tsv	79	Unbuntu
dialogs/99/87.tsv	55	Unbuntu
Total Unbuntu turns	**196**	Unbuntu

If you have already started the assignment and annotated other conversations than the ones listed in the table, I salute your initiative. You should keep whatever you have already done and add conversations from the table as needed to reach as close to 200 turns as possible for each dataset. Please list conversation IDs in your write up so I know what you have annotated.