INFO 259 HW_3 Interpretation

I think we cannot fully trust the results as trends in ACL. First, I observed that some of the categories were not getting any traction until recent years, such as ETHICS and GENERATION, so there were less data points for the CNN model to learn. The performance of these two categories were bad in the confusion matrix. While I look at the more popular categories, such MT, SENTSEM and MULTILING, I found that they were predicted more accurately according to the confusion matrix. Second, I find the categories with more inaccurate predictions such as MLCLASS, RESOURCES and QA usually have overlaps with other topics, which making the prediction performance not very good. Given my experience annotating the dataset in HW1, many articles could possibly belong to more than 1 category, but we have to select one that we find most appropriate. When we are unsure, we tend to choose the topic that is more commonly seen or have higher coverage compared to categories newly seen in academic journals. This is partly the reason for bias for predicting these articles to the most frequent categories, such as MT, SENTSEM, MULTILING, QA, APPLICATIONS and DIALOGUE. Thus, when we only have one label for each article, the prediction results for the dataset are not showing the whole picture of trends in ACL community. It may be more appropriate to have multiple labels (e.g. top 3 relevant labels) for the articles to show the trends.