Chuan Ning, SID 450039506

1. Introduction

This report is made based on the collective annotation released on March 28th. In order to produce relevant insights of the data, I developed a program facilitating to calculate the quantitative results where the conclusions may drive from. In this section, I briefly present an overview of what and how the results are calculated.

Raw data source: 5046_w4_0328_annotations.csv
All relevant source code includes: AnnEditor.py, AnnAnalyst.py, InterCompare.py, InterCat.py
All generated files include: ann_edited.csv, individual results (e.g. id_cnin0770.csv), summarizing.csv, overall.csv
Deployment environment: Python 3.6 and MS Excel
For complete files and code, see: https://github.com/cnin0770/5046_w5_ass1

1.1 AnnEditor (figure 1)

Given the raw .csv file that collects all individual annotation, this script draws a table with two axes of Unikeys and annotated subjects to give a better view of the data.
In addition, it also summarises the Gold column which represents the most voted annotation for each subject. See the output file in figure 2.

1.2 AnnAnalyst function (figure 3)

The Analyst function calculated the most relevant statistics such as precision and recall on the individual base (each Unikey per se). The individual result contains the desired statistics of each category as well as an overview of the performance combining all 15 categories of that individual annotator. An example can be seen in figure 4 ("Overview" on the last row).

1.3 InterCompare function (figure 5)

This function aims to compare the performance among all annotators. It produces a table with two axes of Unikey and statistics (e.g. precision and recall).
For convenience purpose, Fleiss' Kappa is calculated in Analyst function and presented in the output file of this function. See figure 6.

1.4 InterCat function (figure 9)

In order to compare the difference performance across 15 categories, this function is designed to draw a table with dimensions of key statistics for each category. The statistics are calculated combining all 55 annotators. See figure 10 for its production.

2. Analysis

2.1 Confusion Matrix

A confusion matrix gives the accumulated amounts of the annotating task. It counts 4 base parameters:

- True Positive: positive agreement between Gold class and annotated results.
- True Negative: negative agreement between Gold class and annotated results.
- False Positive: annotator gives a positive answer while Gold class disagrees.
- False Negative: annotator gives a negative answer while Gold class disagrees.

2.2 Precision, Recall and F-score

Statistical results may be drawn from the confusion matrix are (e.g. figure 4):

- Precision: higher if more correct answers were achieved against the total annotated subjects.
- Recall: higher if more correct answers were achieved against the total subjects that ought to be selected out.
- F-score: comprehensively presents the correctness rate with the balance between Precision and Recall; here I used 1 as the weight, per se, Precision and Recall are weighted equally.

2.3 Cohen's Kappa

Cohen's Kappa measures the agreement rate between the Gold class and the individual performance. It takes the accuracy rate, which represents the observed agreement, and expected agreement into account.

2.4 Fleiss' Kappa

Fleiss' Kappa gathers all annotators' performance to draw a conclusion of the agreement degree among them despite what the Gold class or the correct answers are. Thus, it can only be calculated with the results from more than 2 annotators. In this case, it is from all 55 annotators.

2.5 Quantitative analysis

2.5.1 Individual performance

Precision and Recall give a straightforward indication of how was the annotating task performed by an individual student. F-score combines them both with a weighted balance. In contrast, Cohen's Kappa presents the results in a different aspect of the agreement.

Take the example of my annotation (Unikey "cnin0770", figure 4), the best-annotated categories are Education, Health, etc. where all subjects in these categories were founded. Accordingly, the four above statistics are 100% despite the difference of actual amounts of these categories.

Transport is the most poorly categorized task with the lowest Cohen rate of 56.14%. Consistently, the F-score is also being the lowest even though its Recall is not.

This disagreement of statistics can be explained by that F-score combines both Precision and Recall. While their weights stay the same, Recall does not overtake the Precision mathematically in this particular case. This reflects the consistency between Cohen's Kappa and F-score.

The same situation can be found in another example of Unikey "aari5136" (figure 7) in categories Sports and Trade and the collective summary (figure 6). Therefore, there might be a conclusion to be observed that F-score is often consistent with Cohen's Kappa while F-score can be changed with its F weight.

2.5.2 Collective performance

Fleiss' Kappa is the best indicator to show how the agreement was achieved between annotators. According to the calculation, it is rated 68.15%.

By the digits alone, it may show a relatively and comprehensively high degree of agreement in terms of categories, subjects, and annotator numbers.

However, it does not provide a comparison or guideline to show whether this conclusion is acceptable. In practice, there might be an issue to request a range of acceptance regarding this statistic from the NLP community or task-providers.

Another way to observe the collective performance is through all the individual results presented with their statistics (figure 6) and performance across categories (figure 10). It compares the annotators against each other in an easy understanding form; although when the dataset comes much larger, this might not be the best way to gain the overall view for human evaluators. Some observation can be gained:

- deli9372 and rnic7557 occupy the top of the correctness ranking.
- Many of the results are mathematically similar, such as aair5136, awon8465, saxu2211, yeli0406. This phenomenon may imply the commonly confused categories. However, through this report and the calculation, it is hard to direct the reader to the exact location of such annotated items.
- Sports is ranked the most correctly categorized annotation, 93%, while Resource is the least, 56.92%; F1-score shares the consistent view with Cohen's Kappa.
- In addition to above, "Other" category achieves 100% F1-score. This abnormal result is explained in next section.

3. Critiques

3.1 Human performance

In processing the data, an error occurred regarding a typo. It was eventually proved to be a lower case of the category name "error" instead of the formatted "Error" (figure 8). This took a while to find out while the fixing did not have any difficulty. The cause of this might be that the annotating task was carried out manually.

This experience in processing the data implies that human performance may introduce uncertainty that hard to predict.

3.2 Procedural flaws

It is noted that in summarizing the individual performance across all categories, False Positive always equals to False Negative (see rows "overview" in figures 4 and 7). It can be well explained in terms of math, that they added up to the same amount. However, this may lose the representativeness in some degree that two indicators always share the same number.

Another thing worth noting is that according to the Gold class, no subject is assigned to "Other" category. This causes *boundary issues* such as Recall and F-score result in error due to dividing by zero.

4. Summary

By processing the annotating data, simple and clear conclusions can be observed such as overview performance and individual performance in form of statistics.
The statistical indicators are inner-related and describe different aspects of the annotation task.
These conclusions are also affected by errors in human performance and potential systematically flaws.

5. Images and Charts

Please see following pages for the figures and the Gold class annotation (the most voted categories).

aari5136	tp	fn	fp	tn	precision	recall	f1	accuracy	marginalFals	marginalTrue	expectedAgr	cohensKappa
Health	1	0	0	149	100.00%	100.00%	100.00%	100.00%	98.67%	0.00%	98.68%	100.00%
Hospitality	0	2	0	148	100.00%	100.00%	100.00%	100.00%	100.00%	0.00%	100.00%	100.00%
Real estate	1	0	0	149	100.00%	100.00%	100.00%	100.00%	98.67%	0.00%	98.68%	100.00%
Resources	0	1	0	149	100.00%	100.00%	100.00%	100.00%	100.00%	0.00%	100.00%	100.00%
Other	0	0	2	148	100.00%	100.00%	100.00%	100.00%	100.00%	0.00%	100.00%	100.00%
SciTech	4	0	1	145	80.00%	100.00%	88.89%	99.33%	94.09%	0.09%	94.18%	88.55%
Education	11	2	1	136	91.67%	84.62%	88.00%	98.00%	84.03%	0.69%	84.72%	86.91%
Entertainme	31	6	1	112	96.88%	83.78%	89.86%	95.33%	59.26%	5.26%	64.52%	86.85%
Public admin	19	2	3	126	86.36%	90.48%	88.37%	96.67%	73.39%	2.05%	75.44%	86.43%
Sports	18	4	1	127	94.74%	81.82%	87.80%	96.67%	74.52%	1.86%	76.38%	85.89%
Trade	11	0	4	135	73.33%	100.00%	84.62%	97.33%	83.40%	0.73%	84.13%	83.19%
overview	124	26	26	2074	82.67%	82.67%	82.67%	97.69%	87.11%	0.44%	87.56%	81.43%
Finance	2	1	0	147	100.00%	66.67%	80.00%	99.33%	96.69%	0.03%	96.72%	79.67%
Transport	2	1	0	147	100.00%	66.67%	80.00%	99.33%	96.69%	0.03%	96.72%	79.67%
Error	19	1	10	120	65.52%	95.00%	77.55%	92.67%	69.91%	2.58%	72.49%	73.34%
Social	5	6	3	136	62.50%	45.45%	52.63%	94.00%	87.72%	0.39%	88.12%	49.51%

Figure 7

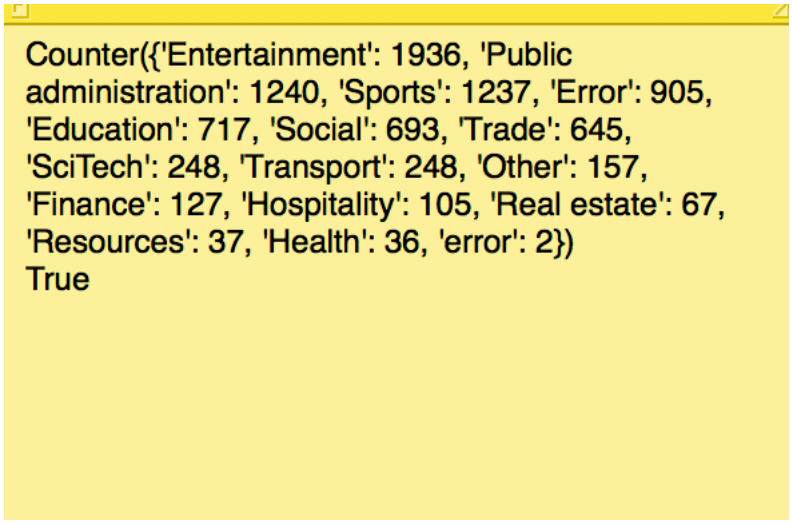

Figure 8

overall	precision	recall	f1	cohen
Education	89.20%	87.83%	88.51%	87.43%
Entertainme	96.63%	90.17%	93.29%	91.19%
Finance	80.65%	60.61%	69.20%	68.67%
Health	94.29%	60.00%	73.33%	73.19%
Hospitality	89.32%	83.64%	86.38%	86.21%
Public admin	77.11%	81.39%	79.19%	75.70%
Real estate	65.15%	78.18%	71.07%	70.86%
Resources	72.22%	47.27%	57.14%	56.92%
SciTech	67.21%	74.55%	70.69%	69.84%
Social	63.64%	71.74%	67.44%	64.70%
Sports	93.91%	94.30%	94.10%	93.09%
Trade	80.91%	84.79%	82.81%	81.41%
Transport	54.69%	81.21%	65.37%	64.52%
Other	100.00%	100.00%	100.00%	100.00%
Error	75.76%	61.09%	67.64%	63.27%

Figure 10