Report on Annotation Guidelines SANTA_6

Introduction

This report shows the quantitative inter-annotator agreement (IAA) that was achieved using one guideline. Each guideline has been used three times: By the authors of the guideline ("own"), by one other participant team ("foreign"), and by a paid student assistant at Stuttgart University ("student"). The document contains IAA calculations between all three annotations, but also between every pair of annotations. Therefore, you'll find four tables in the document. Each table contains the IAA values for each text in the test set (eight texts) as well as the arithmetic mean over all texts (columns 1 to 3). In addition, the columns 4 and 5 ("Gamma (mean)" and "Gamma (max)") show the mean and maximal values that have been achieved on the given text over all guidelines.

Intuition

IAA is typically calculated as a combination of two values: The observed and the expected agreement. Observed agreement is the amount of agreement we see in the annotations.

Expected agreement is a somewhat artificial notion and is calculated as a chance agreement, i.e., the amount of agreement we would get if every annotator throws dice. This gives us an insight into the difficulty of the annotation task. For instance, an annotation task with only two categories is much easier than one with ten categories. This is reflected in the chance agreement, which is higher in the former.

Both values are then combined into a single score, which you'll find in the column called "Gamma". The **best possible score** (perfect agreement) is 1.0. Scores below 0 (zero) indicate an agreement less than chance agreement (i.e., throwing a die would be better). Everything between 0 (zero) and 1 (one) is positive agreement (i.e., better than throwing a die).

Choices

There are different ways to calculate IAA. A decision on a suited metric is typically made when details about the annotation setup (guidelines) are known. To compare different guidelines, we opted for a metric called "Gamma" that makes very little assumptions on the nature of the annotation task. It is described in detail in Mathet, Widlöcher, and Métivier (2015).

nan values

Some of the tables contain several `nan` values in the gamma column. `nan` stands for "not a number" and is the return value of computation that doesn't work (e.g., division by zero, depending on the programming language). Most of the issues occur for the expected agreement, and then percolate to the gamma calculation. If the expected agreement is `nan`, gamma will also be `nan`. In some cases, the program has problems estimating the expected agreement, for instance if there is only a single annotation in a document (e.g., a single narrative level covering the entire text).

For the calculation of mean (and max) of gamma scores, `nan` scores have been excluded. We will investigate this further and hopefully send an update before the workshop.

Overall Inter-Annotator Agreement

Text	Observed	Expected	Gamma	Gamma (mean)	Gamma (max)
Buechner	2.183	2.344	0.069	0.070	0.225
Chekhov	3.036	2.885	-0.053	0.098	0.548
Henry	2.160	2.033	-0.062	0.055	0.369
Kafka	1.818	2.786	0.348	0.365	0.380
Kleist	1.532	2.607	0.412	0.214	0.487
Lagerloef	2.808	2.388	-0.176	0.069	0.444
Storm	1.971	2.054	0.040	0.121	0.341
Tieck	2.231	2.077	-0.075	0.047	0.313
Mean	2.217	2.397	0.063	0.130	

Own vs. Foreign

Text	Observed	Expected	Gamma	Gamma (mean)	Gamma (max)
Buechner	1.386	1.441	0.038	0.029	0.282
Chekhov	2.000	1.514	-0.321	0.031	0.625

Text	Observed	Expected	Gamma	Gamma (mean)	Gamma (max)
Henry	1.260	1.288	0.021	0.061	0.356
Kafka	1.176	1.698	0.307	0.337	1.000
Kleist	1.435	1.805	0.205	0.324	1.000
Lagerloef	2.500	1.926	-0.298	-0.049	0.622
Storm	1.682	1.632	-0.031	0.011	0.289
Tieck	1.847	1.581	-0.168	-0.004	0.321
Mean	1.661	1.611	-0.031	0.093	

Own vs. Students

Text	Observed	Expected	Gamma	Gamma (mean)	Gamma (max)
Buechner	1.991	1.746	-0.140	0.043	0.386
Chekhov	2.500	1.883	-0.327	0.003	0.465
Henry	1.752	1.700	-0.031	0.013	0.392
Kafka	2.251	2.094	-0.075	0.304	1.000
Kleist	1.245	1.822	0.317	0.168	0.504
Lagerloef	2.013	1.406	-0.432	0.221	1.000
Storm	1.712	1.704	-0.005	0.225	0.731
Tieck	1.702	1.643	-0.036	0.101	0.465
Mean	1.896	1.750	-0.091	0.135	

Foreign vs. Students

Text	Observed	Expected	Gamma	Gamma (mean)	Gamma (max)
Buechner	1.876	1.850	-0.014	-0.047	0.153
Chekhov	2.071	1.882	-0.101	0.051	0.435
Henry	1.661	1.493	-0.113	-0.039	0.292
Kafka	2.333	2.168	-0.076	0.216	0.556
Kleist	1.208	1.777	0.320	0.499	1.000
Lagerloef	1.842	1.811	-0.017	-0.034	0.420
Storm	1.504	1.500	-0.003	-0.010	0.228
Tieck	1.731	1.551	-0.116	-0.026	0.197
Mean	1.778	1.754	-0.015	0.076	

References

Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. "The Unified and Holistic Method Gamma ( ) for Inter-Annotator Agreement Measure and Alignment." *Computational Linguistics* 41 (3):437–79.