Inter-annotator Agreement vs. System Performance (F1 measure)

Mayla Boguslav

November 2016

1 Introduction

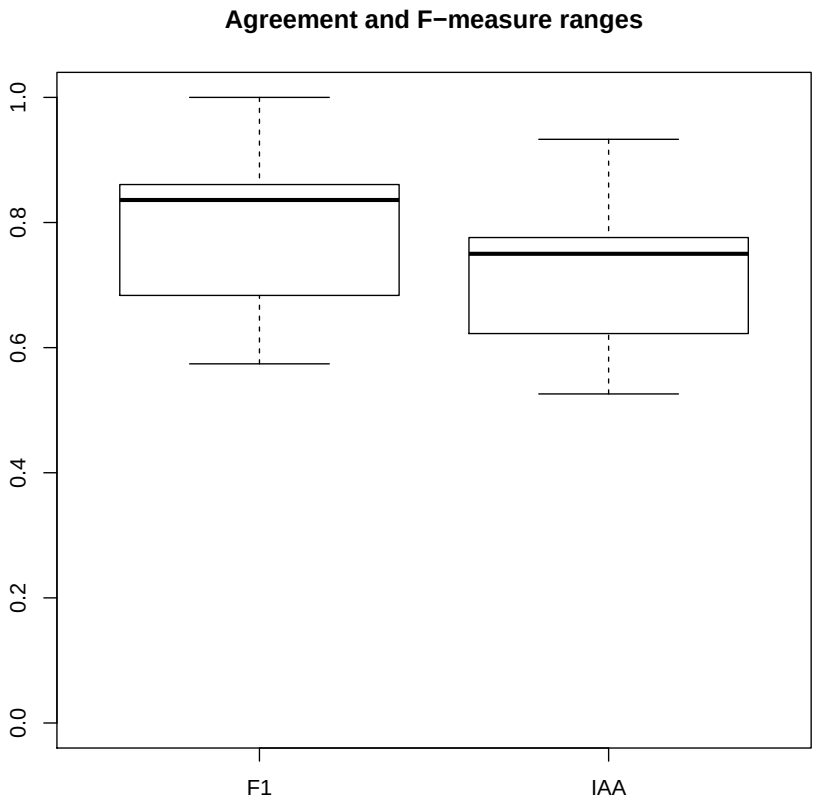
Abstract

To classify texts in natural language processing, we compute the agreement between annotators: do annotators classify texts the same? It is often thought the agreement between annotators is the upper limit on system performance: if humans cannot agree with each other about the classification more than some percentage of the time, then we do not expect a computer to do any better. We trace the logical positivist roots of the motivation for measuring inter-annotator agreement, trace the origins of the widely-held belief about the relationship between inter-annotator agreement and system performance, and present data on 6 systems that suggests that inter-annotator agreement is not in fact an upper bound, with evidence from the biomedical and general domains. Further, we found a significantly positive correlation between inter-annotator agreement and system performance. This will inform the ability of scientists to evaluate research, make funding decisions, and provide accurate information to the public about technology.

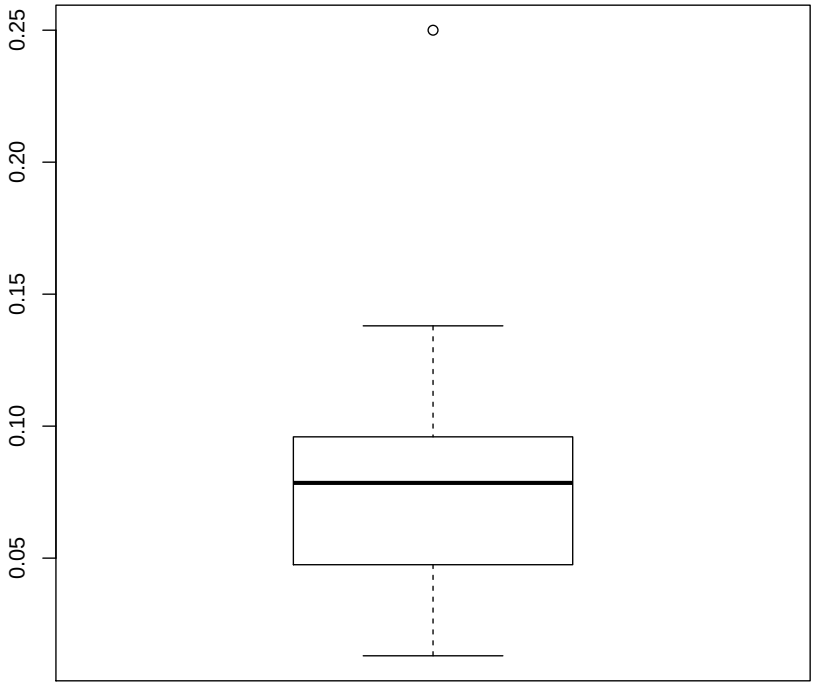
2 IAA < System Performance (F1 measure mostly)

```
iaa_F1GthanIAA <- c(0.5535, 0.5535, 0.5535, 0.617, 0.526,
                           0.69, 0.933, 0.819, 0.779, 0.815,
                           0.798, 0.773, 0.628, 0.75, 0.75,
                           0.75, 0.75, 0.75, 0.75, 0.75)

# F-measure, except two that are precision
system_F1GthanIAA <- c(0.5741, 0.6504, 0.6649, 0.63, 0.664,
                            0.725, 0.978, 0.875, 0.824, 0.87,
                            0.857, 0.823, 0.702, 0.839, 0.839,
                            0.845, 0.8644, 1, 0.845, 0.833)
```

```r
# Difference = F-measure minus IAA
difference_F1GthanIAA <- c(0.0206, 0.0969, 0.1114, 0.013, 0.138,
                           0.035, 0.045, 0.056, 0.045, 0.055,
                           0.059, 0.05, 0.074, 0.089, 0.089,
                           0.095, 0.1144, 0.25, 0.095, 0.083)
#Level of agreement
level_F1GthanIAA <- c(3,3,3,4,3,4,5,5,4,5,4,4,4,4,4,4,4,4,4,4)

#basic distributional aspects
labels1 <- c(rep("IAA", 20), rep("F1", 20))
iaa.and.f1_F1GthanIAA <- c(iaa_F1GthanIAA, system_F1GthanIAA)
boxplot(iaa.and.f1_F1GthanIAA~labels1, ylim=c(0, 1.0), main="Agreement and F-measure ranges"
```




```r
#Summary of the difference
boxplot(difference_F1GthanIAA)
```




```
#summary statistics
summary(iaa_F1GthanIAA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5260  0.6252  0.7500  0.7144  0.7745  0.9330

summary(system_F1GthanIAA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5741  0.6927  0.8360  0.7951  0.8588  1.0000

summary(difference_F1GthanIAA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01300 0.04875 0.07850 0.08071 0.09548 0.25000

#Shapiro-Wilk normality test
```

```r
#null hypothesis = the data is normally distributed (p >0.05)
#or not if p<0.05
shapiro.test(iaa_F1GthanIAA) #not normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iaa_F1GthanIAA
## W = 0.90252, p-value = 0.046
```

```r
shapiro.test(system_F1GthanIAA) #normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  system_F1GthanIAA
## W = 0.92209, p-value = 0.1087
```

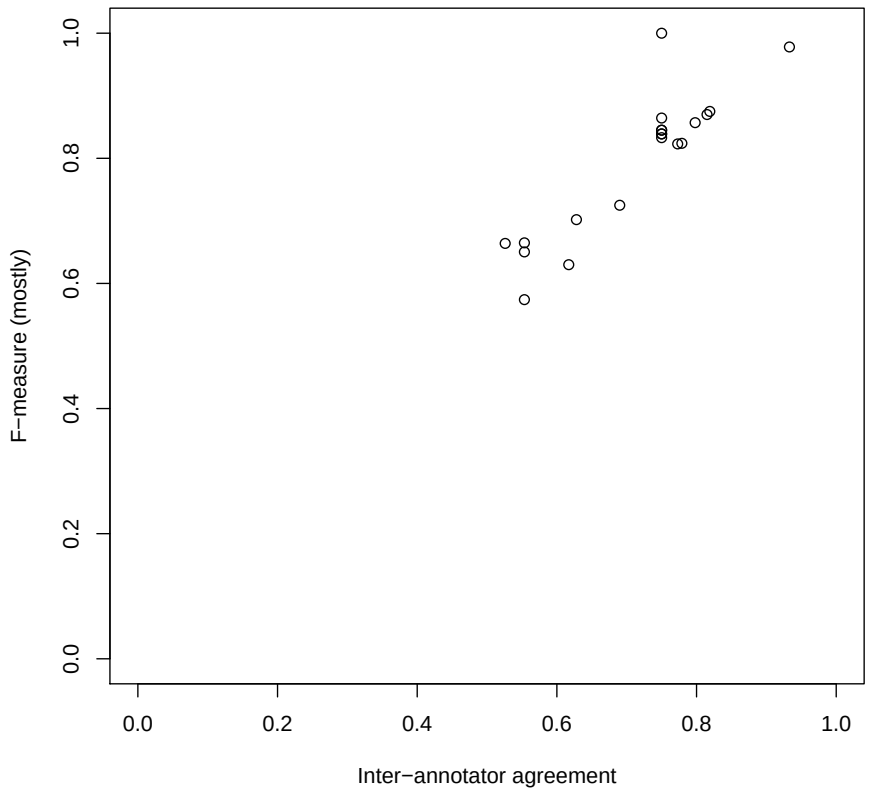
```r
shapiro.test(difference_F1GthanIAA) #not normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  difference_F1GthanIAA
## W = 0.84741, p-value = 0.004825
```

```r
shapiro.test(level_F1GthanIAA) #not normally distributed
```

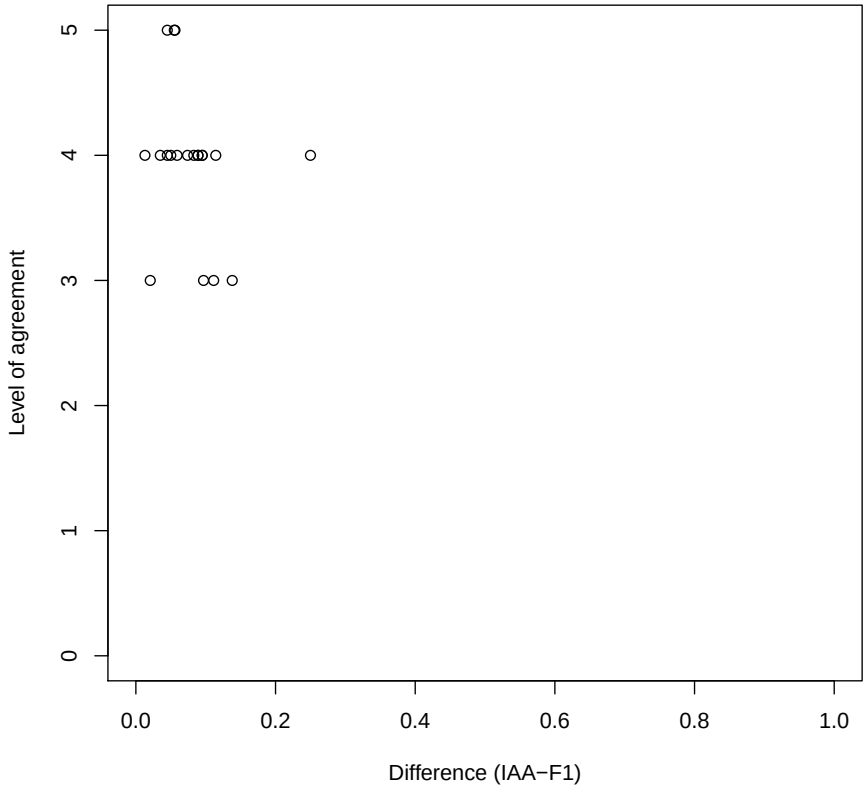
```
##
##  Shapiro-Wilk normality test
##
## data:  level_F1GthanIAA
## W = 0.76768, p-value = 0.000296
```

```r
#plot of IAA vs F measure
plot(iaa_F1GthanIAA, system_F1GthanIAA,
     xlim=c(0,1.0), ylim=c(0,1.0),
     xlab="Inter-annotator agreement",
     ylab="F-measure (mostly)")
```



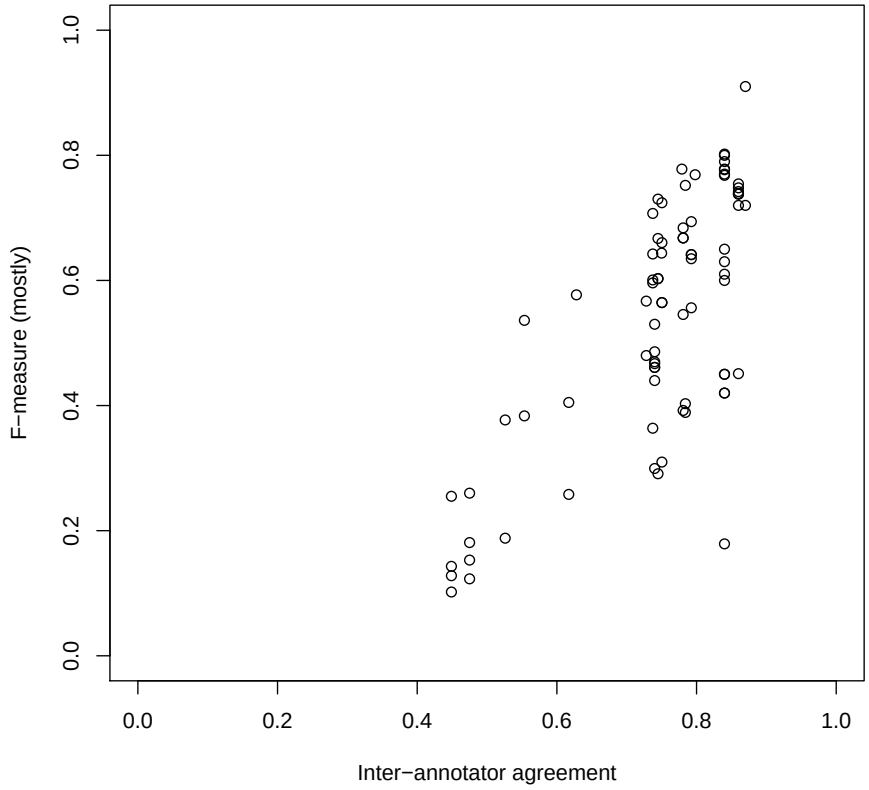

```
#Spearman's correlation
# positive = 0.8069756
cor(iaa_F1GthanIAA,system_F1GthanIAA, method ="spearman")

## [1] 0.8069756

# spearman's rank test
## positive correlation for iaa and system:
cor.test(iaa_F1GthanIAA,system_F1GthanIAA, alternative = 'greater',
         method="spearman", exact = TRUE,
         conf.level = 0.95, continuity = FALSE)

##
##  Spearman's rank correlation rho
##
## data:  iaa_F1GthanIAA and system_F1GthanIAA
```

```
## S = 256.72, p-value = 8.56e-06
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.8069756

## difference and level are not negatively correlated
cor.test(difference_F1GthanIAA, level_F1GthanIAA,
         alternative ='less',
         method="spearman", exact = TRUE,
         conf.level = 0.95, continuity = FALSE)

##
##  Spearman's rank correlation rho
##
## data:  difference_F1GthanIAA and level_F1GthanIAA
## S = 1798.6, p-value = 0.06381
## alternative hypothesis: true rho is less than 0
## sample estimates:
##        rho
## -0.3523232

plot(difference_F1GthanIAA, level_F1GthanIAA,
     xlim=c(0,1.0), ylim=c(0,5.0),
     xlab="Difference (IAA-F1)",
     ylab="Level of agreement")
```

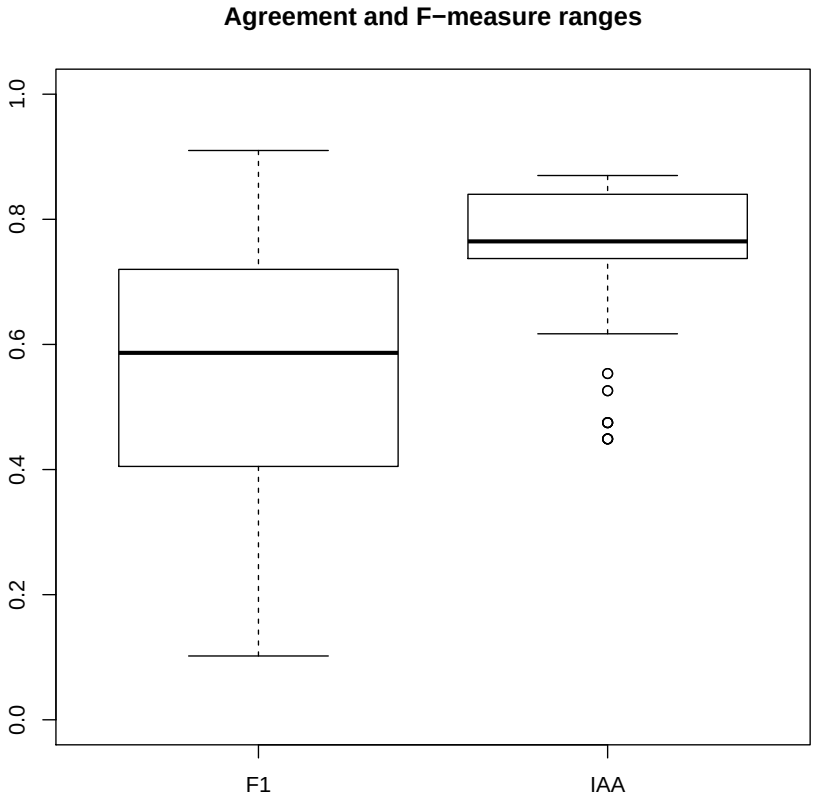


3 IAA > System Performance (F1 measure mostly)

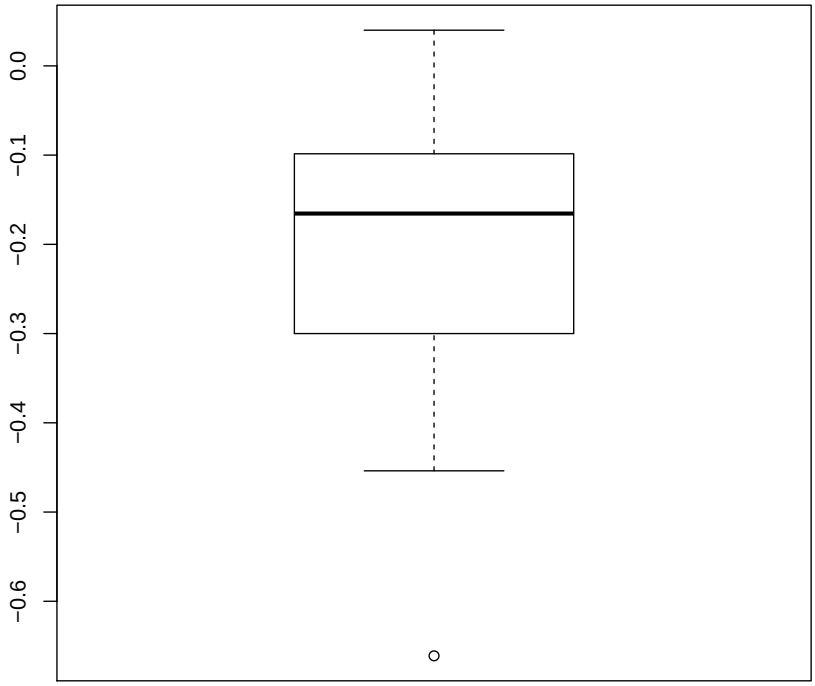
```
IAAGThanFmeasure <- read.table(file = 'IAAGthanFmeasure.csv',
                        header = TRUE, sep=',', dec='.')

plot(IAAGThanFmeasure$IAA, IAAGThanFmeasure$System,
     xlim=c(0,1.0), ylim=c(0,1.0),
     xlab="Inter-annotator agreement",
     ylab="F-measure (mostly)")
```




```
#basic distributional aspects
labels2 <- c(rep("IAA", 82), rep("F1", 82))
iaa.and.f1_IAAGthanF1 <- c(IAAGThanFmeasure$IAA,
            IAAGThanFmeasure$System)
boxplot(iaa.and.f1_IAAGthanF1~labels2, ylim=c(0, 1.0),
            main="Agreement and F-measure ranges")
```




```
#Summary of the difference
boxplot(IAAGThanFmeasure$Difference)
```




```
#summary statistics
summary(IAAGThanFmeasure$IAA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4490  0.7373  0.7647  0.7422  0.8400  0.8700

summary(IAAGThanFmeasure$System)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1020  0.4088  0.5865  0.5411  0.7168  0.9100

summary(IAAGThanFmeasure$Difference)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.6611 -0.2985 -0.1655 -0.2011 -0.1003  0.0400

#Shapiro-Wilk normality test
```

```r
#the data is normally distributed (p >0.05)
#or not if p<0.05
shapiro.test(IAAGThanFmeasure$IAA) #not normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  IAAGThanFmeasure$IAA
## W = 0.79957, p-value = 3.384e-09
```

```r
shapiro.test(IAAGThanFmeasure$System) # not normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  IAAGThanFmeasure$System
## W = 0.94439, p-value = 0.001425
```

```r
shapiro.test(IAAGThanFmeasure$Difference) # not normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  IAAGThanFmeasure$Difference
## W = 0.95359, p-value = 0.004845
```

```r
shapiro.test(IAAGThanFmeasure$Level) #not normally distributed
```

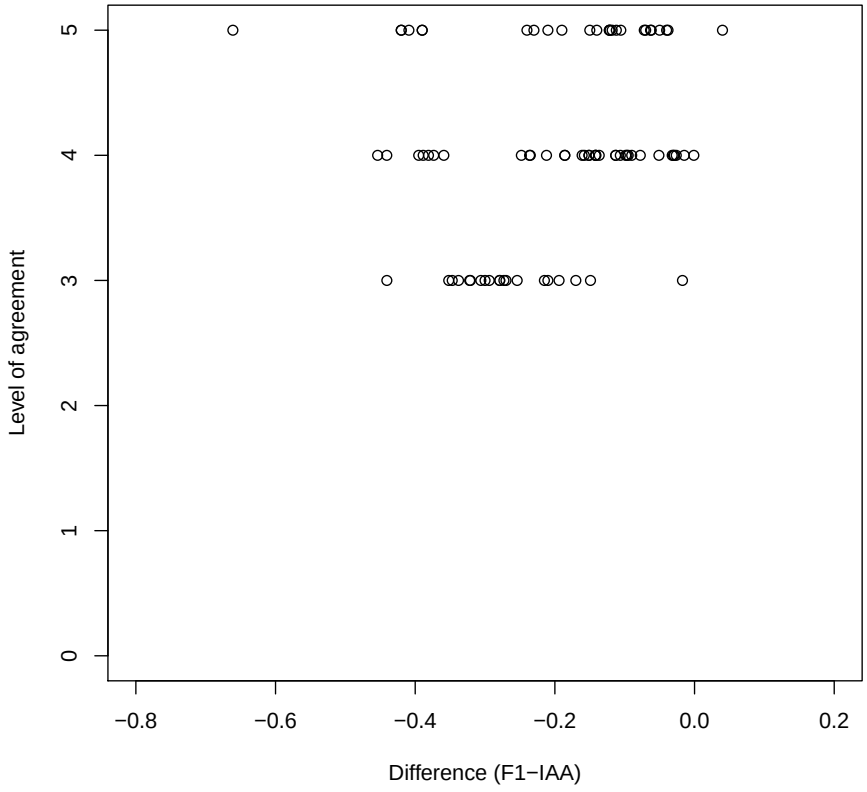
```
##
##  Shapiro-Wilk normality test
##
## data:  IAAGThanFmeasure$Level
## W = 0.80727, p-value = 5.626e-09
```

```r
#Spearman's correlation
cor(IAAGThanFmeasure$IAA, IAAGThanFmeasure$System,
         method ="spearman")
```

```
## [1] 0.6532331
```

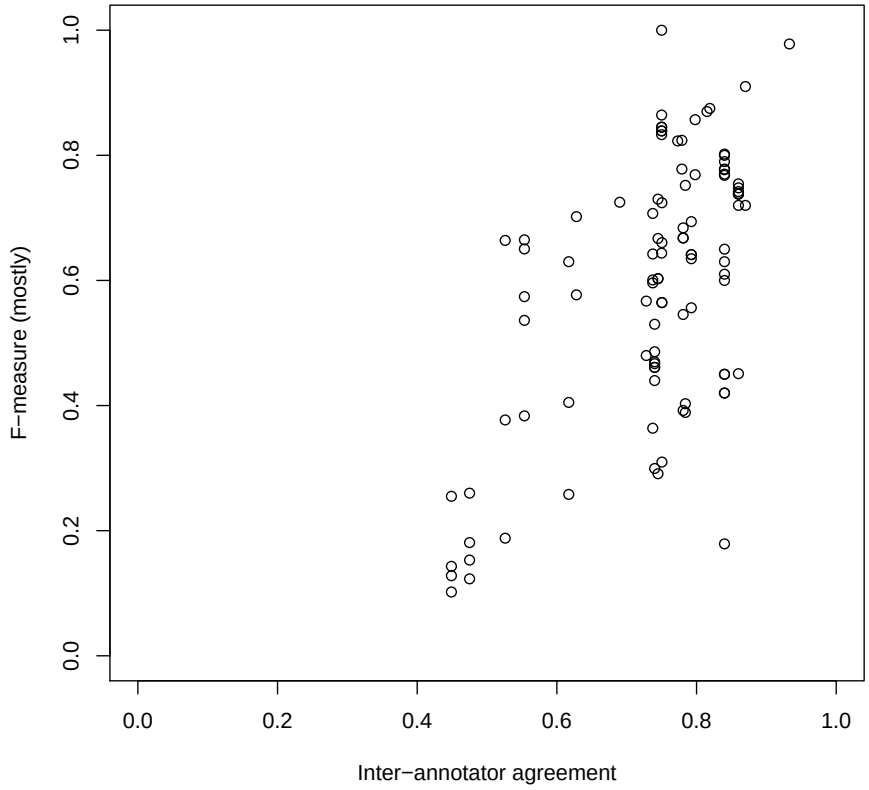
```r
# spearman's rank test
## positive correlation for iaa and system significant
cor.test(IAAGThanFmeasure$IAA, IAAGThanFmeasure$System,
         alternative = 'greater' , method="spearman", exact = TRUE,
         conf.level = 0.95, continuity = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  IAAGThanFmeasure$IAA and IAAGThanFmeasure$System
## S = 31861, p-value = 1.449e-11
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.6532331

## difference and level are significantly positively correlated
cor.test(IAAGThanFmeasure$Difference, IAAGThanFmeasure$Level,
         alternative = 'greater' , method="spearman", exact = TRUE,
         conf.level = 0.95, continuity = FALSE)

##
##  Spearman's rank correlation rho
##
## data:  IAAGThanFmeasure$Difference and IAAGThanFmeasure$Level
## S = 68147, p-value = 0.009563
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##      rho
## 0.258308

plot(IAAGThanFmeasure$Difference, IAAGThanFmeasure$Level,
     xlim=c(-0.8,0.2), ylim=c(0,5.0),
     xlab="Difference (F1-IAA)",
     ylab="Level of agreement")
```

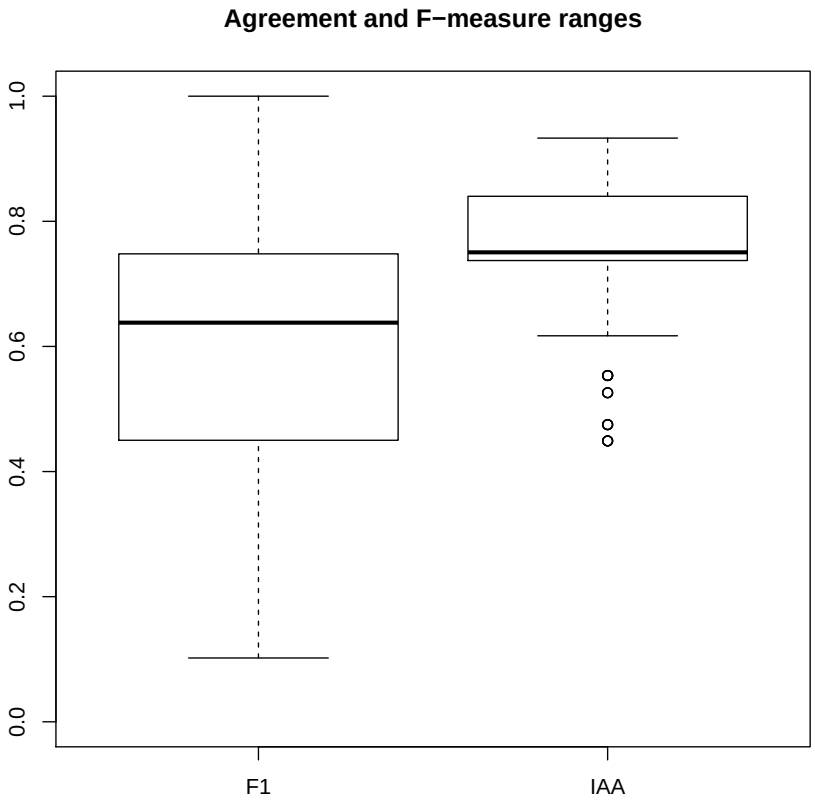


4 All Data Combined

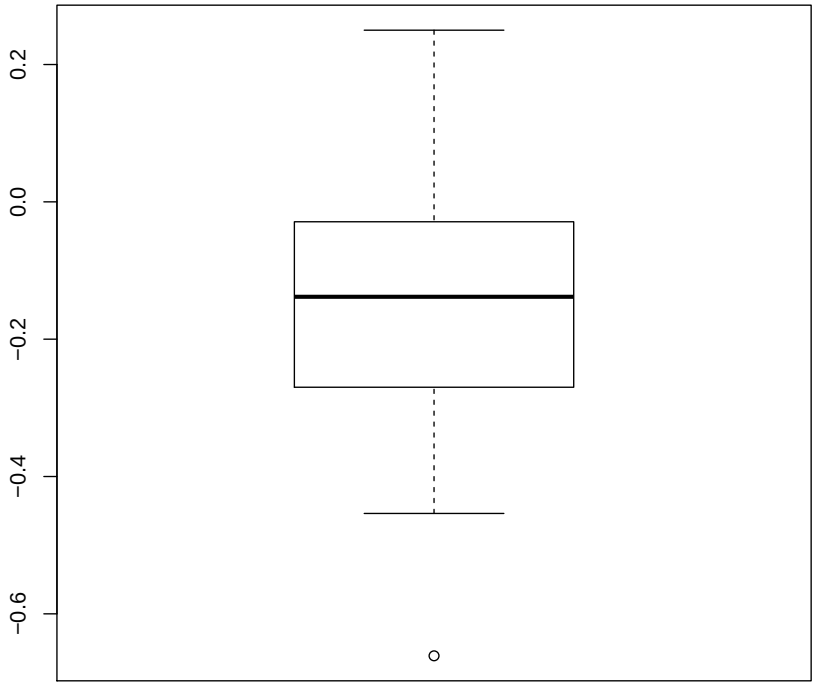
```
alldata <- read.table(file = 'All_IAA_Fmeasure.csv',
            header = TRUE, sep=',', dec='.')
plot(alldata$IAA, alldata$System,
     xlim=c(0,1.0), ylim=c(0,1.0),
     xlab="Inter-annotator agreement",
     ylab="F-measure (mostly)")
```




```
#basic distributional aspects
labels <- c(rep("IAA", 102), rep("F1", 102))
iaa.and.f1_alldata <- c(alldata$IAA, alldata$System)
boxplot(iaa.and.f1_alldata~labels, ylim=c(0, 1.0),
         main="Agreement and F-measure ranges")
```




```
#Summary of the difference
boxplot(alldata$Difference)
```




```
#summary statistics
summary(alldata$IAA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4490  0.7373  0.7504  0.7368  0.8400  0.9330

summary(alldata$System)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1020  0.4502  0.6380  0.5909  0.7465  1.0000

summary(alldata$Difference)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.6611 -0.2660 -0.1383 -0.1458 -0.0293  0.2500

#Shapiro-Wilk normality test
```

```r
#the data is normally distributed (p >0.05)
#or not if p<0.05
shapiro.test(alldata$IAA) #not normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alldata$IAA
## W = 0.84755, p-value = 7.518e-09
```

```r
shapiro.test(alldata$System) #not normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alldata$System
## W = 0.95872, p-value = 0.00289
```

```r
shapiro.test(alldata$Difference) #normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alldata$Difference
## W = 0.9836, p-value = 0.239
```

```r
shapiro.test(alldata$Level) #not normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alldata$Level
## W = 0.80806, p-value = 3.357e-10
```

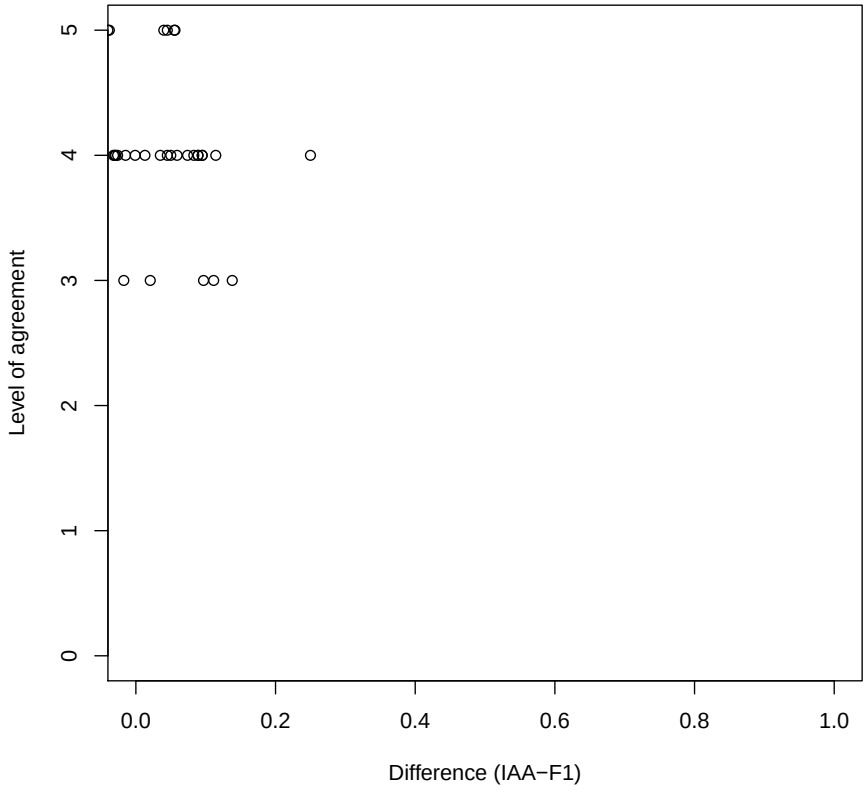
```r
#Spearman's correlation
cor(alldata$IAA, alldata$System, method ="spearman")
```

```
## [1] 0.5126484
```

```r
# spearman's rank test
## positive correlation for iaa and system significant
cor.test(alldata$IAA, alldata$System,
         alternative = 'greater' , method="spearman", exact = TRUE,
         conf.level = 0.95, continuity = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
```

```
## data:  alldata$IAA and alldata$System
## S = 86189, p-value = 1.81e-08
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.5126484

## difference and level are not negatively correlated
cor.test(alldata$Difference, alldata$Level,
         alternative = 'greater' , method="spearman", exact = TRUE,
         conf.level = 0.95, continuity = FALSE)

##
##  Spearman's rank correlation rho
##
## data:  alldata$Difference and alldata$Level
## S = 156300, p-value = 0.1223
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.1162316

plot(alldata$Difference, alldata$Level,
     xlim=c(0,1.0), ylim=c(0,5.0),
     xlab="Difference (IAA-F1)",
     ylab="Level of agreement")
```


Level of agreement
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
Difference (IAA−F1)