TRANSLATOR Co-occurrence baseline calculations

Introduction

This document gives descriptive statistics on our pilot relation annotation project. It also establishes baseline performance of co-occurrence for the six relations. These will be used as benchmarks in development of relation extraction applications during the next funding period.

Author: Kevin Cohen

Calculate figures of merit for all relations

```
##                          relation yes  no anaphora entity speculation process
## 1    has_symptom(disease, symptom)  53 102        3      0           0      NA
## 2     causesNoMike(gene, disease) 288  18       15     12           6      NA
## 3       causesMike(gene, disease) 329  18       15     12           6      NA
## 4             regulates(gene, gene)   0 148        3      1           6       8
## 5         regulates(chemical, gene)  NA  NA       NA     NA          NA      NA
## 6             treats(drug, disease) 142 100        4    145          33       4
## 7 located(gene product, location) 157 116        7      4          29       7
##   other.relation segmentation increase decrease direction.unspecified
## 1             36            4        0        0                     1
## 2              2            9        0        0                     1
## 3              2            9        0        0                     1
## 4             12           50       11        7                    36
## 5             NA           NA       NA       NA                    NA
## 6             37           36        0        0                     0
## 7             63          193        0        0                     0


##                          relation yes  no anaphora entity speculation process
## 1    has_symptom(disease, symptom)  53 102        3      0           0      NA
## 2     causesNoMike(gene, disease) 288  18       15     12           6      NA
## 3       causesMike(gene, disease) 329  18       15     12           6      NA
## 4             regulates(gene, gene)   0 148        3      1           6       8
## 5         regulates(chemical, gene)  NA  NA       NA     NA          NA      NA
## 6             treats(drug, disease) 142 100        4    145          33       4
## 7 located(gene product, location) 157 116        7      4          29       7
##   other.relation segmentation increase decrease direction.unspecified precision
## 1             36            4        0        0                     1        NA
## 2              2            9        0        0                     1        NA
## 3              2            9        0        0                     1        NA
## 4             12           50       11        7                    36        NA
## 5             NA           NA       NA       NA                    NA        NA
## 6             37           36        0        0                     0        NA
## 7             63          193        0        0                     0        NA
##   recall f_measure
## 1     NA        NA
```
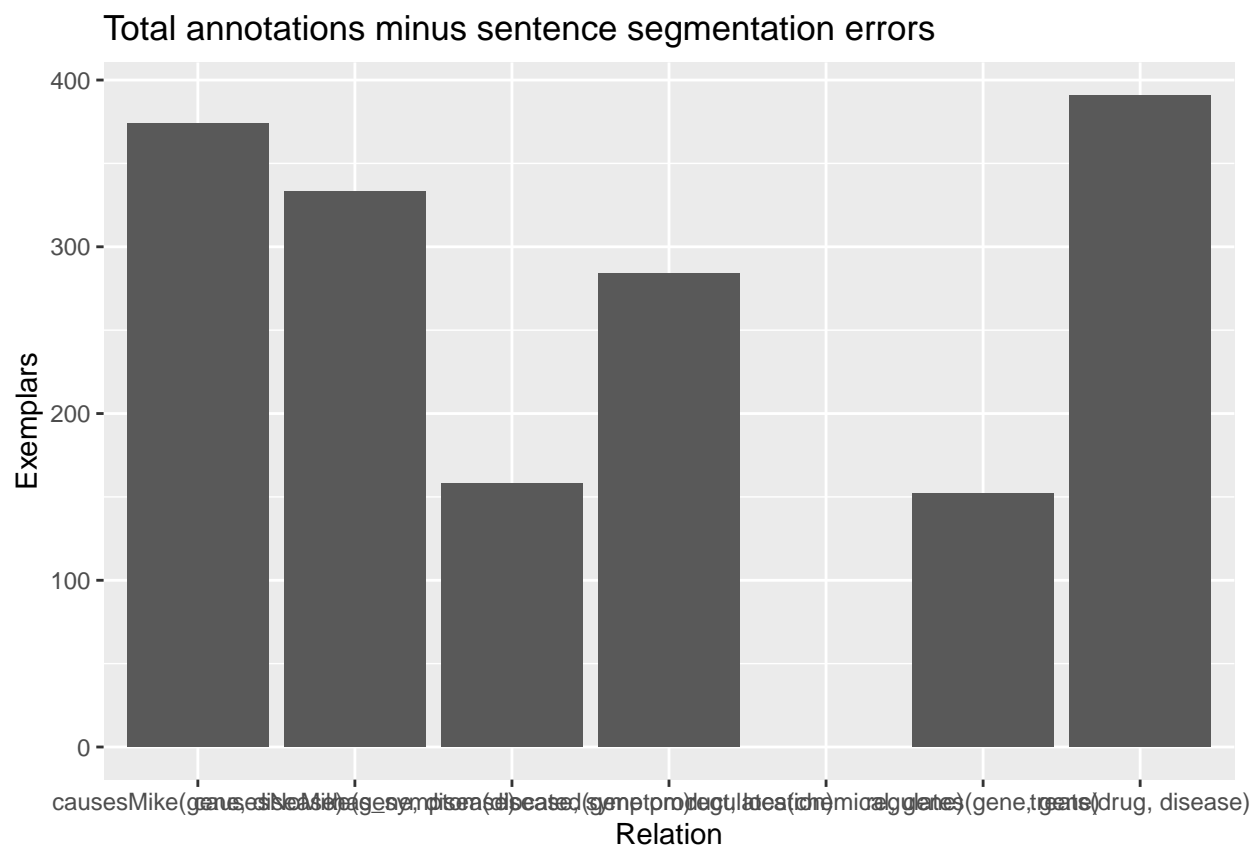
```
## 2      NA         NA
## 3      NA         NA
## 4      NA         NA
## 5      NA         NA
## 6      NA         NA
## 7      NA         NA
```

Descriptive statistics of dataset

TODO: plot balance per relation

```
## Warning: Removed 1 rows containing missing values (position_stack).
```

Total annotations minus sentence segmentation errors

Plot values for figures of merit

```
## Warning: package 'cowplot' was built under R version 4.0.2
```

```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```
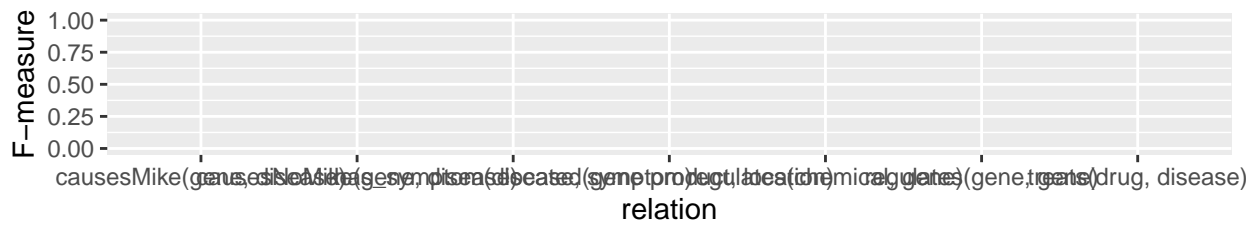
```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```

A   Baseline by relation: Precision
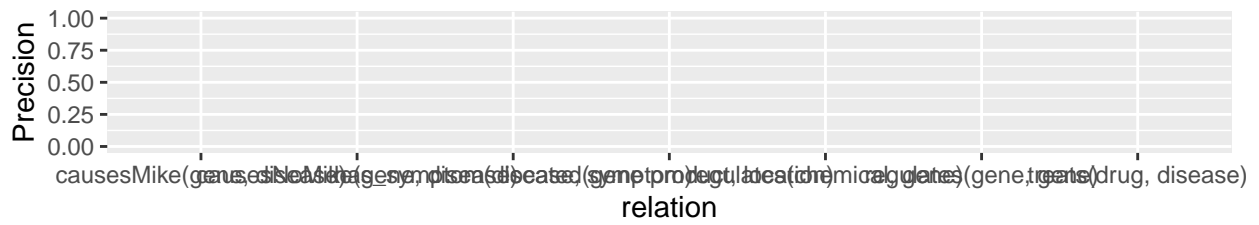
B   Baseline by relation: Recall
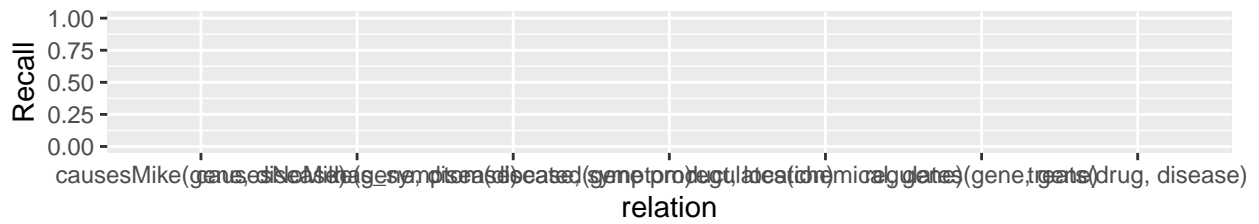
C   Baseline by relation: F−measure


```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```

Variability across relations: Precision


```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).

## Warning: Removed 7 rows containing non-finite values (stat_boxplot).

## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```

A Variability across relations: Precis **B** Variability across relations: Recall

C Variability across relations: F−measure