ggplot with factors

Francisco Guzman

May 12, 2016

In this recipe, we will learn:

- How to load data from a text file (reinforce)
- How to create summaries with ddply (reinforce)
- How to plot bar plots with ggplot
- How to plot different views using facets
- How to save your graph as a pdf (reinforce)

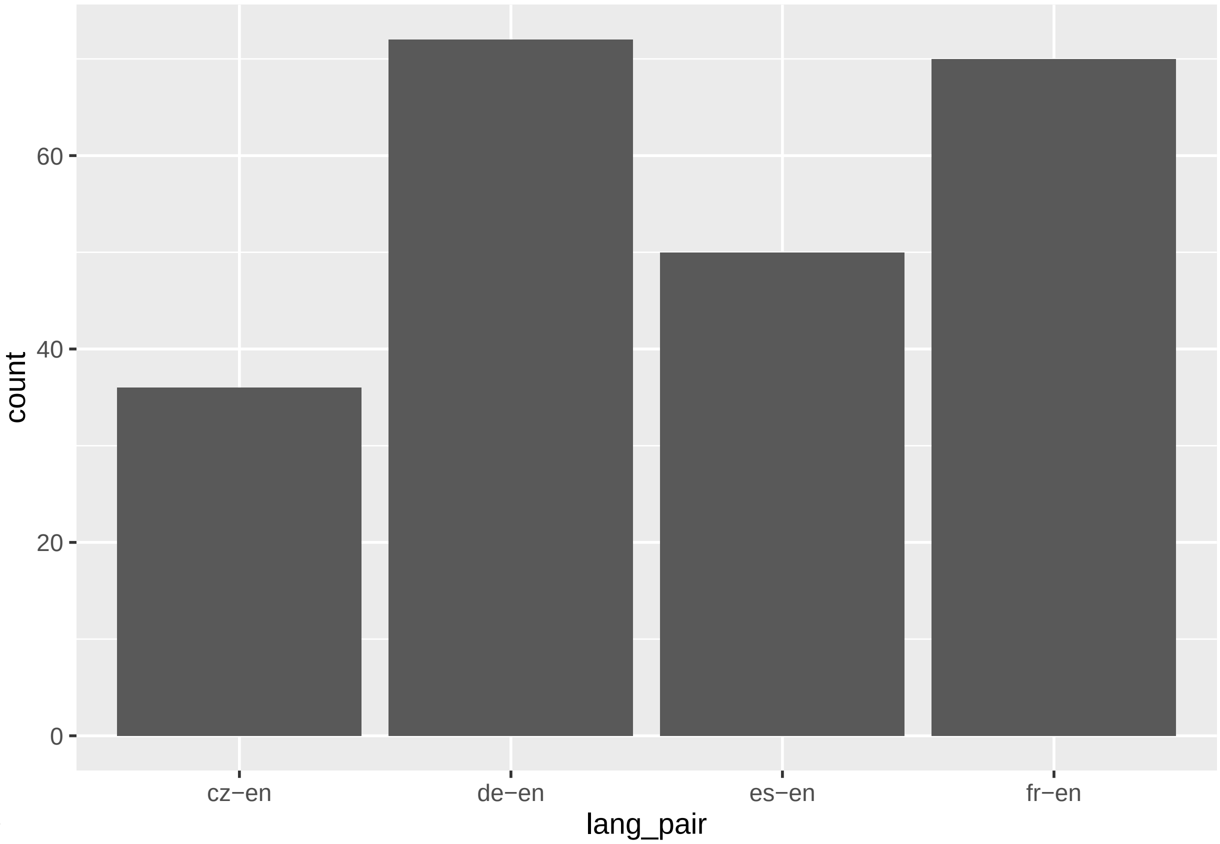
0.1 Load data

```
df <-read.table("data/mte_metrics.dat",header=FALSE,col.names=c("metric","lang_pair","testset","system"
# The data looks a bit different than in the previous exercise
head(df)
```

```
##   metric lang_pair             testset                         system
## 1 DR_LEX     cz-en newssyscombtest2011                      bbn-combo
## 2 DR_LEX     cz-en newssyscombtest2011 cmu-heafield-combo-contrastive
## 3 DR_LEX     cz-en newssyscombtest2011             cmu-heafield-combo
## 4 DR_LEX     cz-en newssyscombtest2011                cst-contrastive
## 5 DR_LEX     cz-en newssyscombtest2011                            cst
## 6 DR_LEX     cz-en newssyscombtest2011           cu-bojar-contrastive
##       score
## 1 0.5486888
## 2 0.5381331
## 3 0.5414202
## 4 0.4709361
## 5 0.4659940
## 6 0.5049184
```

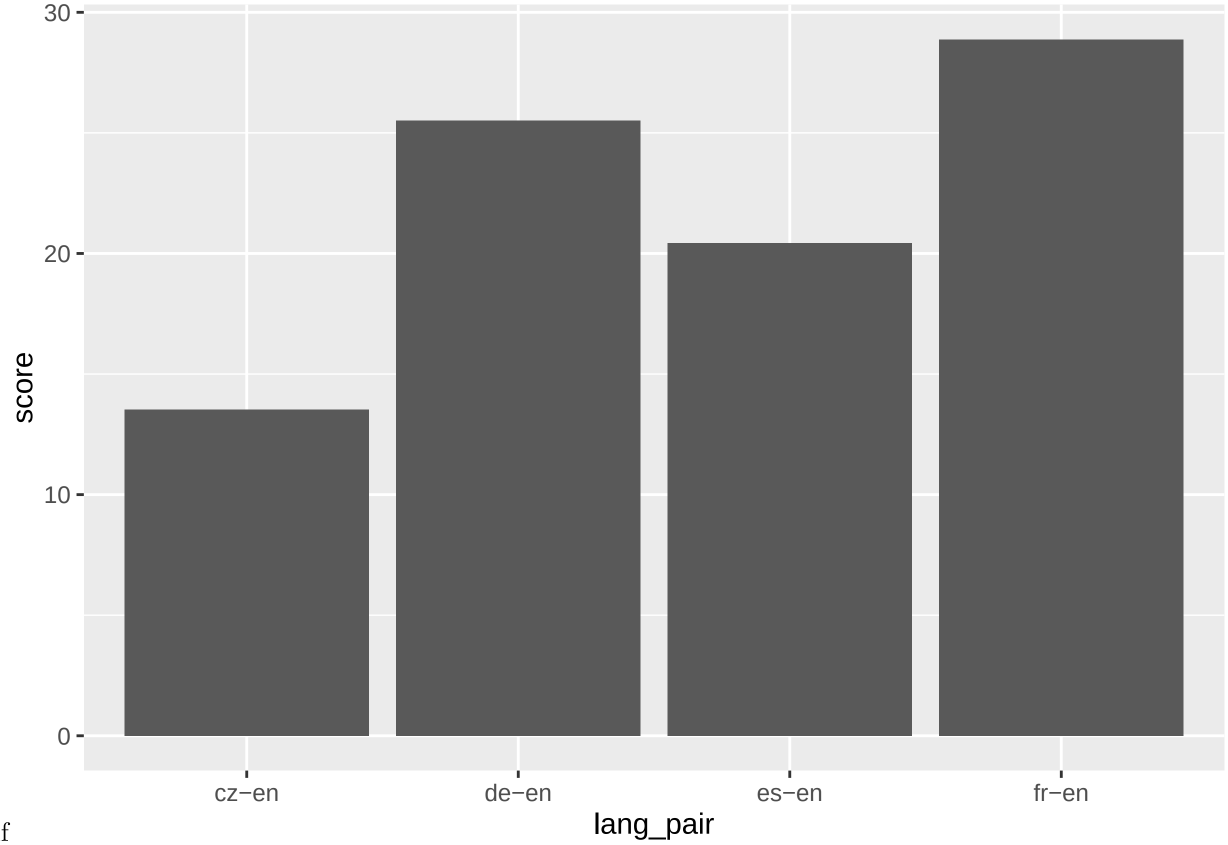
1 Bar plots with ggplot

```
# load ggplot2
library(ggplot2)

# If we just use the geom bar over the original data, it will create a frequency distribution (histogra
ggplot(df,aes(x=lang_pair))+geom_bar()
```



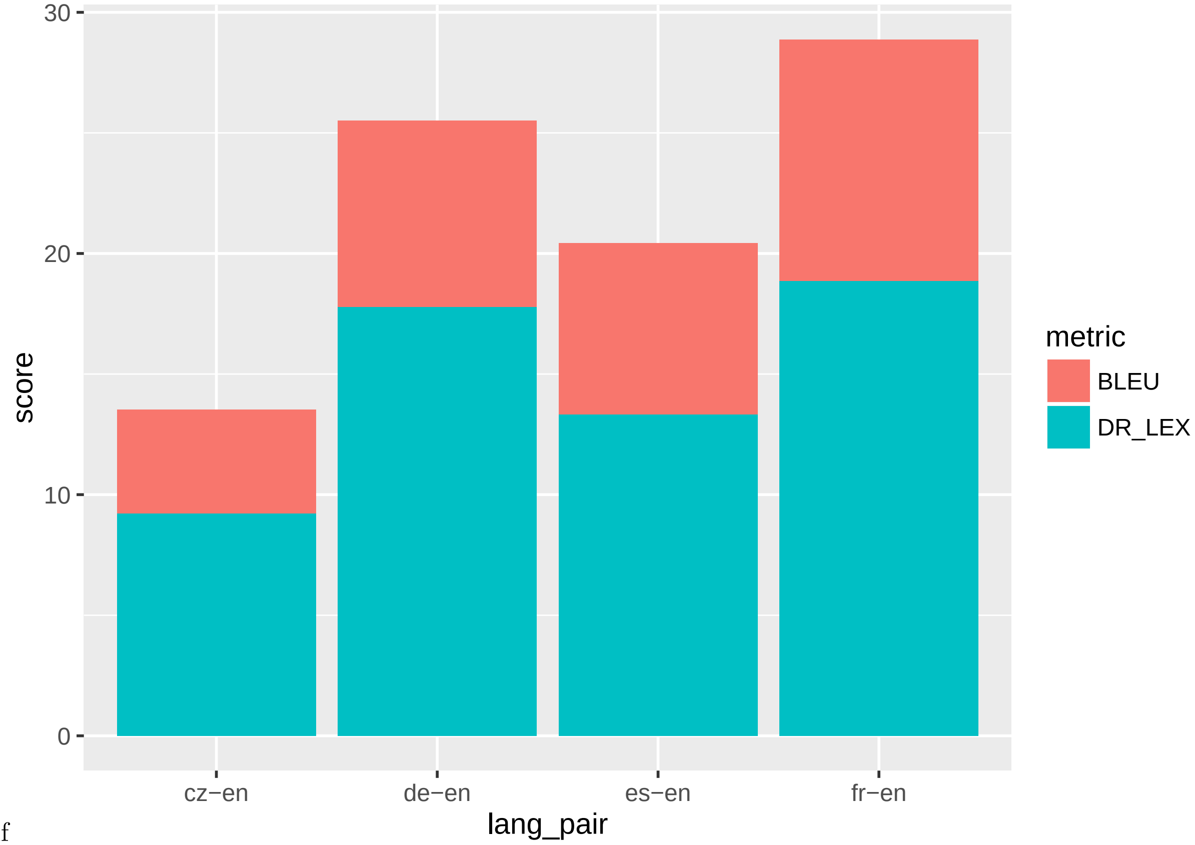
plot-1.pdf

```
# But this is not very useful because we want a comparison of scores between BLEU and DR_LEX

#Let's add scores into the picture
ggplot(df,aes(x=lang_pair,y=score))+geom_bar(stat="identity") # we need to tell ggplot that we want to
```



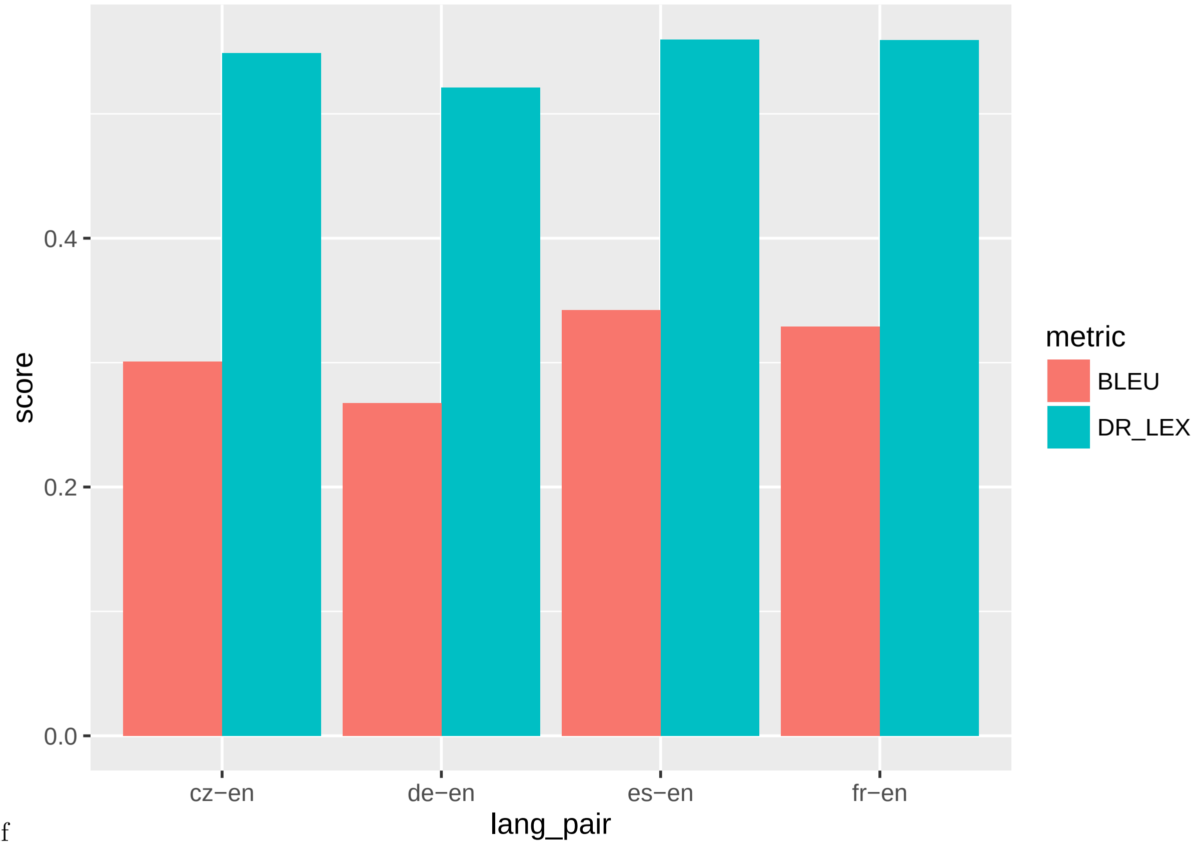
plot-2.pdf

```
#Still this picture is not useful, because we can't compare the two metrics
#Let's color the bars
ggplot(df,aes(x=lang_pair,y=score,fill=metric))+geom_bar(stat="identity")
```



plot-3.pdf

```
#This is not a good graph because it is "stacking the graphs". Let's put them side by side

ggplot(df,aes(x=lang_pair,y=score,fill=metric))+geom_bar(stat="identity", position="dodge")
```



plot-4.pdf

```
# This is more useful!
```

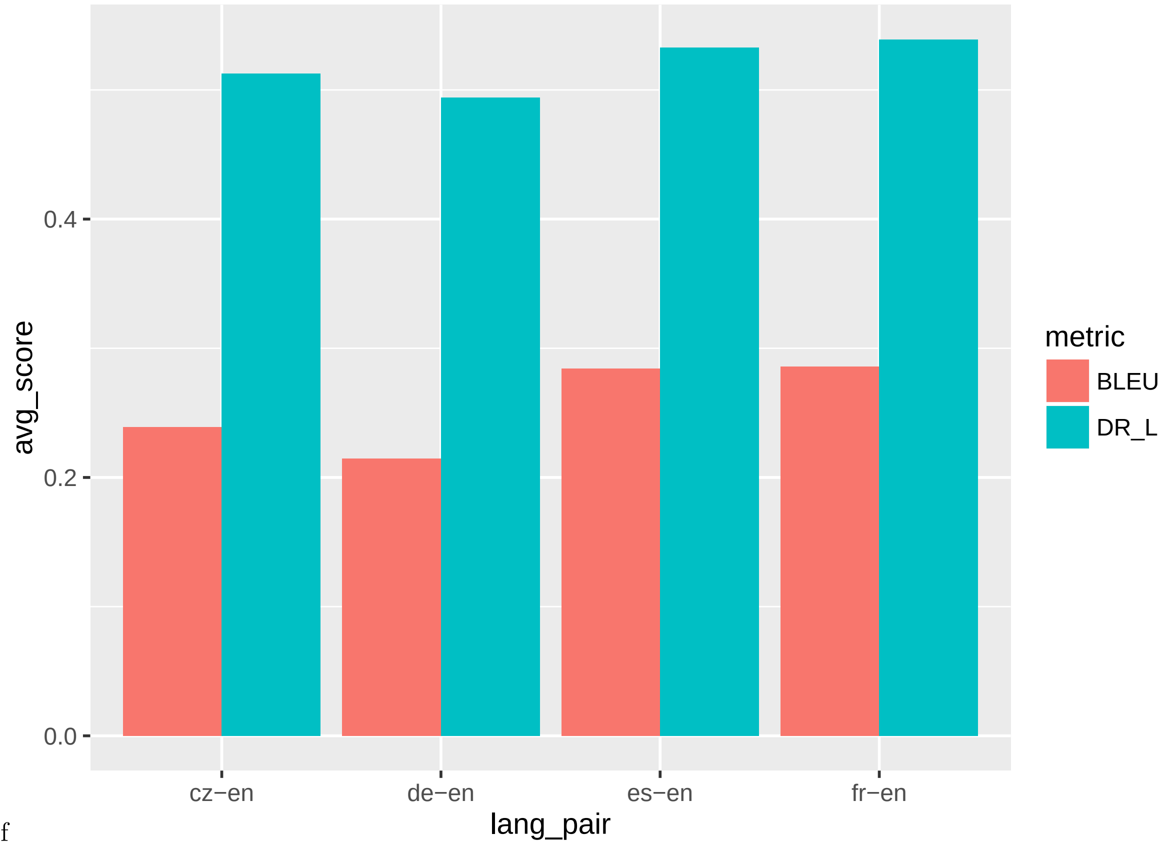
2 Method 2: using ddply

We first prepare a useful summary

```
#we use ddply to obtain average scores by metric and language pair
library(plyr)
summary<-ddply(df,.(lang_pair,metric),summarize,avg_score=mean(score))
print(summary)
```

```
##   lang_pair metric avg_score
## 1     cz-en   BLEU 0.2391139
## 2     cz-en DR_LEX 0.5126081
## 3     de-en   BLEU 0.2144891
## 4     de-en DR_LEX 0.4939107
## 5     es-en   BLEU 0.2844944
## 6     es-en DR_LEX 0.5327323
## 7     fr-en   BLEU 0.2858628
## 8     fr-en DR_LEX 0.5389791
```

```
ggplot(summary,aes(x=lang_pair,y=avg_score,fill=metric ))+geom_bar(stat="identity",position="dodge")
```

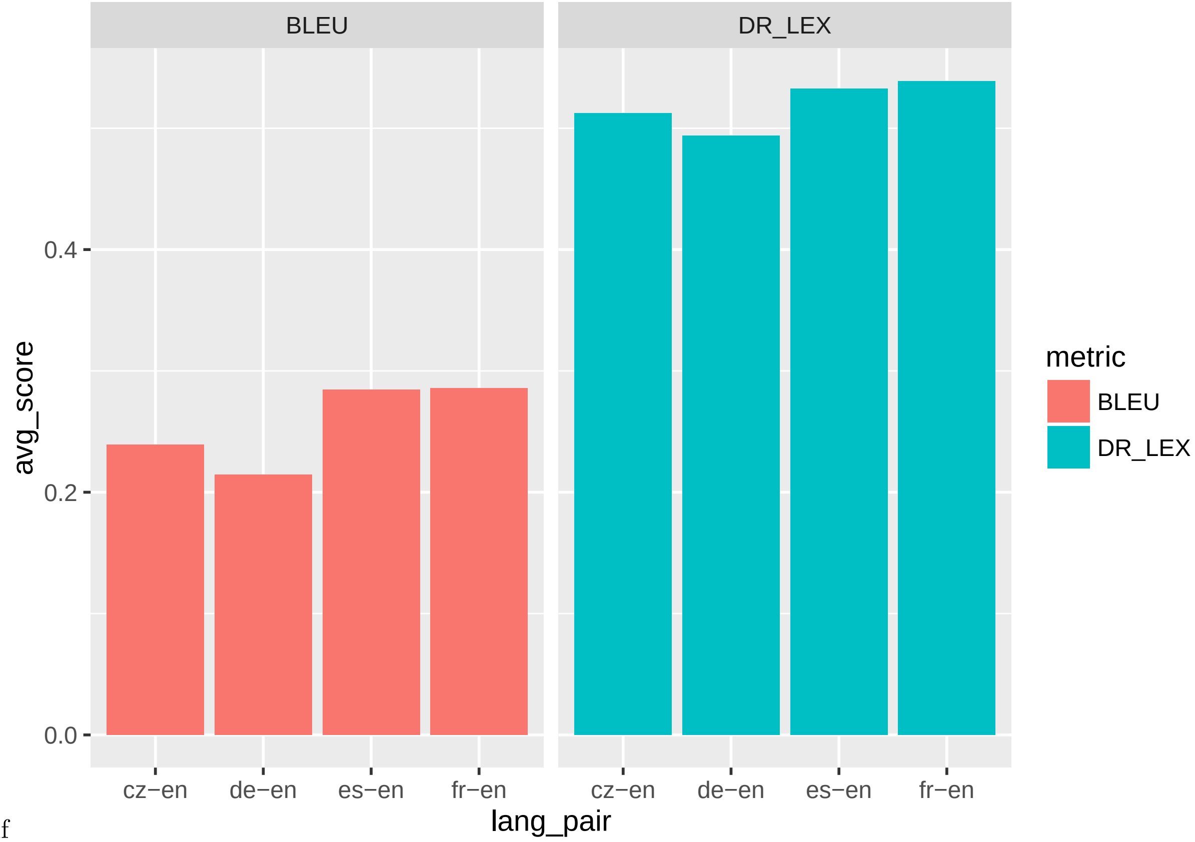


plot over summaries-1.pdf

3 Using facets

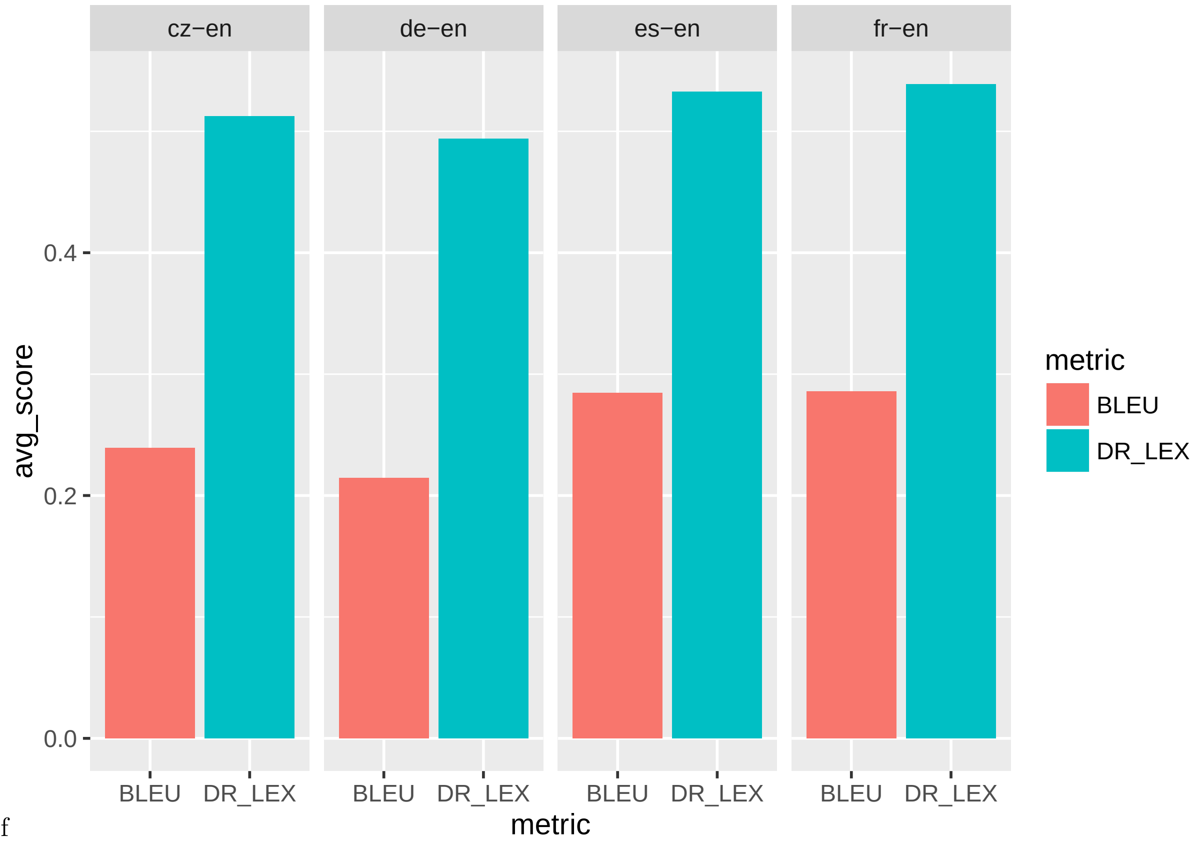
Now, instead of “dodging” the bars, we could plot them side by side using facets

```
ggplot(summary,aes(x=lang_pair,y=avg_score,fill=metric))+geom_bar(stat="identity")+
  facet_grid(.~metric)  #this wil facet horizontally using the factor  metric
```



the metric-1.pdf

```
ggplot(summary,aes(x=metric,y=avg_score,fill=metric))+geom_bar(stat="identity")+
  facet_grid(.~lang_pair)
```



the lang pair-1.pdf

4 Set titles, and print

```
my_plot<- ggplot(summary,aes(x=metric,y=avg_score,fill=metric))+geom_bar(stat="identity",width=0.9)+  fa

pdf("img/my_second_plot.pdf",height=5,width=7)
print(my_plot)
dev.off()
```

```
## RStudioGD
##         2
```