

比赛链接位置 <https://www.statmt.org/wmt20/quality-estimation-task.html>
点击 result 可以看到结果， 点击 submissions 得到当年比赛的各个队伍的提交文件

以 wmt20 ende qe 任务为例
打开 team NJUNLP 的提交文件， 里面是 predictions_mt.txt 文件
预测的内容是 gap mt gap mt gap ...的形式

wmt20 官方这个结果是 mt 和 gap 合在一起算的结果

English-Chinese

	Team	Codalab username	Run	Target MCC ▼	Target F1-BAD	Target F1-OK	Source MCC	Source F1-BAD	Source F1-OK
1	NiuTrans	NiuTrans		0.610	0.723	0.887	0.308	0.666	0.639
2	HW-TSC	yuriak		0.587	0.714	0.866	0.000	0.000	0.000
3	NICT Kyoto	Raphael_NICT		0.582	0.704	0.878	0.336	0.668	0.669
4	IST and Unbabel	ddk		0.575	0.706	0.850	0.287	0.705	0.410
5	IST and Unbabel	Joao1996		0.567	0.701	0.842	0.287	0.705	0.403
6	NJUNLP	cuiqu		0.551	0.672	0.877	0.000	0.000	0.000
7	Organisers	erickrf	BASELINE	0.509	0.658	0.849	0.270	0.682	0.547

下图是 wmt21 的最终结果， 可以看到 mt gap src 三个指标是分开算的， 也就是每年官方的计算方法可能有变动

statmt.org/wmt21/quality-estimation-task_results.html

English-German

top)

Words in MT

Team	Codalab username	Run	Rank	MCC ▼	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# model params
1	JHU-Microsoft	sding	3	0.523	0.599	0.907	0.543	6,863,178,235	484,431,872
2	HW-TSC		3.6	0.510	0.587	0.900	0.528	2,243,954,093	560,944,640
3	IST-Unbabel		3.8	0.466	0.551	0.914	0.504	2,294,887,576	569,368,609
4	Abulice		4.2	0.437	0.530	0.884	0.468	2,243,439,613	560,814,661
5	POSTECH	dammy	3	0.413	0.497	0.915	0.454	1,561,188,430	390,210,052
6	Organisers	fbain	BASELINE	3.4	0.370	0.455	0.911	1,142,441,796	281,297,685

GAPs in MT

Team	Codalab username	Run	Rank	MCC ▼	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# model params
1	HW-TSC		3.2	0.300	0.294	0.969	0.285	2,243,954,093	560,944,640
2	JHU-Microsoft	sding	3.4	0.256	0.266	0.985	0.262	6,863,178,235	484,431,872
3	IST-Unbabel		3.8	0.183	0.178	0.986	0.176	2,294,887,576	569,368,609
4	Organisers	fbain	BASELINE	2.8	0.116	0.098	0.986	1,142,441,796	281,297,685
5	POSTECH	dammy	3.8	0.110	0.124	0.982	0.122	1,561,188,430	390,210,052
6	Abulice		4	0.000	0.000	0.000	0.000	0	0

Words in SRC

Team	Codalab username	Run	Rank	MCC ▼	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# model params
1	HW-TSC		3.2	0.450	0.516	0.894	0.461	2,243,954,093	560,944,640
2	IST-Unbabel		3.8	0.404	0.483	0.921	0.445	2,294,887,576	569,368,609
3	Abulice		3.8	0.392	0.468	0.875	0.409	2,243,439,613	560,814,661
4	Organisers	fbain	BASELINE	2.8	0.322	0.393	0.924	1,142,441,796	281,297,685
5	POSTECH	dammy	3.4	0.320	0.395	0.922	0.364	1,561,188,430	390,210,052
6	JHU-Microsoft	sding	4	0.000	0.000	0.000	0.000	0	0

回到 wmt20enzh, njunlp 提交的结果是 predicions_mt.txt
对应的正确的结果是 all.tags 文档， 压缩包里有
(这个 all.tags 文档是我们自己算的， 不是当时官方公开的， 当时官方误传了部分测试集文件， 随后官方很快删了， 但被之前的师兄下载下来了。我们然后通过这个官方上传的文件计算出了 all.tags 文件， 所以这个并不是完全的标准答案， 和正确答案可能有一点点偏差)

比较提交结果 predictions_mt.txt 和 all.tags 文档

用的 python 文件 f1cal_mcc.py, 注意更改存的文档的位置, 运行这个 python 文件

```
def my_main():  
    system_file = "/home/zhangy/predictions_mt.txt"  
    gold_file = "/home/zhangy/all.tags"
```

得到的结果是

```
[0.8762554775662156, 0.6701126847927221]  
0.5871899106362256  
0.548433823258102  
finished
```

第一行是 f1ok, f1bad

第二行是 f1-mult=f1ok* f1bad

第三行是 mcc

下图是网页上官方计算的最终结果, 和我们计算的基本一致, 有一点误差, 都在 0.003 内
这是因为用于计算的 all.tags 是我们自己算的, 不完全准确

6	NJUNLP	cuiqu	0.551	0.672	0.877
---	--------	-------	-------	-------	-------