微博数据关键词抽取

任务描述

1.给定一段微博文本(字符串格式,可以是一篇长文章,也可以是短摘要)作为输入,提取文章中的关键词作为输出. 文本中的关键词具有一定的主观性,因此得到输出的结果后与数据集中已给出的Ground Truth结果作对比,(建议使用的指标是F1-score),得到输出的指标,指标越高则说明结果越精确
2.根据训练文件和验证文件跑出关键词提取的模型,并使用test测试集测试指标

数据集描述

1.各个src文件(测试集,训练集,验证集)中给定微博提取的文本,每单行指定单条微博
2.各个trg文件(测试集,训练集,验证集)中给定文本对应的关键词,每单行对应单条微博的关键词(GROUND TRUTH),通常单条微博的关键词在一个到两个左右

数据集目录中包含六个文件:

train_src.txt 和 train_trg.txt: 训练集的输入微博文本和输出关键词,占总数的80%
test_src.txt和test_trg.txt:测试集的输入微博文本和输出关键词,占总数的10%
valid_src.txt和valid_trg.txt:验证集的输入微博文本和输出关键词,占总数的10%

测试指标

1. F1@1:针对top-1的结果计算F1 score
得到输出的关键词序列后,获取其中的第一个关键词和GT中的第一个关键词作对比,从而计算出f1 score
2. F1@3:针对top-3的结果计算F1 score
得到关键词序列中的前三关键词与GT中的关键词做对比(不关心这三个关键词的相互顺序)
3. MAP（mean average precision）平均精度均值(参考 https://www.jianshu.com/p/82be426f776e)

参考指标值

Model	Twitter			Weibo		
	F1@1	F1@3	MAP	F1@1	F1@3	MAP
Baselines						
MAJORITY	9.36	11.85	15.22	4.16	3.31	5.47
TF-IDF	1.16	1.14	1.89	1.90	1.51	2.46
TEXTRANK	1.73	1.94	1.89	0.18	0.49	0.57
KEA	0.50	0.56	0.50	0.20	0.20	0.20
State of the arts						
SEQ-TAG	22.79±0.3	12.27±0.2	22.44±0.3	16.34±0.2	8.99±0.1	16.53±0.3
SEQ2SEQ	34.10±0.5	26.01±0.3	41.11±0.3	28.17±1.7	20.59±0.9	34.19±1.7
SEQ2SEQ-COPY	36.60±1.1	26.79±0.5	43.12±1.2	32.01±0.3	22.69±0.2	38.01±0.1
SEQ2SEQ-CORR	34.97±0.8	26.13±0.4	41.64±0.5	31.64±0.7	22.24±0.5	37.47±0.8
TG-NET	-	-	-	-	-	-
Our model	**38.49**±0.3	**27.84**±0.0	**45.12**±0.2	**34.99**±0.3	**24.42**±0.2	**41.29**±0.4