数据筛选说明

当前要做的人工筛选分为两个部分：**判断语法点正确与否**、**检查负样本构建是否合理**。

1. 数据集说明

	source	target	source得分	label得分	op
0	种植物立刻这种植物在这个季节长得很快，经过短短一个星期，它就长满了这面墙，叶子很厚，绿绿的。				[[(3, 4, '立刻'), '词类副词时间副词', 'S', ('null', 'null', '在')], [(0, 0, '种'), '词类代词指示代词', 'M', ('null',
1	如此种植标这种植物在这个季节长得很快，经过短短一个星期，它就长满了这面墙，叶子很厚，绿绿的。				[[(5, 6, '有时'), '词类副词时间副词', 'S', ('null', 'null', '在')], [(0, 1, '如此'), '词类代词指示代词', 'S', ('null'
2	这种植物立这种植物在这个季节长得很快，经过短短一个星期，它就长满了这面墙，叶子很厚，绿绿的。				[[(4, 4, 'None'), '词类副词时间副词', 'M', (26, 26, '在')], [(28, 28, 'None'), '词类代词指示代词', 'M', (33, 33, '这'
3	这种植物立这种植物在这个季节长得很快，经过短短一个星期，它就长满了这面墙，叶子很厚，绿绿的。				[[(4, 4, '这'), '词类副词时间副词', 'S', ('null', 'null', '在')], [(4, 4, '这'), '词类代词指示代词', 'M', (30, 30, '1
4	这种植物立这种植物在这个季节长得很快，经过短短一个星期，它就长满了这面墙，叶子很厚，绿绿的。				[[(4, 4, '这'), '词类副词时间副词', 'M', ('null', 'null', '在')], [(27, 27, 'None'), '词类代词指示代词', 'M', (33, 3
5	他年轻很，他很年轻，可是遇到问题很冷静，比相同年龄的人更成熟。				[[(1, 1, 'None'), '词类副词程度副词', 'M', (3, 3, '很')], [(15, 15, '和'), '词类介词引出对象', 'S', ('null', 'null',
6	很他年轻是他很年轻，可是遇到问题很冷静，比相同年龄的人更成熟。				[[(0, 0, '很'), '词类副词程度副词', 'M', (2, 2, 'None')], [(10, 10, '的'), '词类助词结构助词', 'M', (22, 22, 'None')
7	他年轻，可他很年轻，可是遇到问题很冷静，比相同年龄的人更成熟。				[[(1, 1, '年'), '词类副词程度副词', 'M', ('null', 'null', '很')], [(14, 14, '相'), '词类介词引出对象', 'M', ('null',
8	他更真年长他很年轻，可是遇到问题很冷静，比相同年龄的人更成熟。				[[(1, 1, '更'), '词类副词程度副词', 'S', ('null', 'null', '很')], [(21, 21, '所'), '词类助词结构助词', 'S', ('null',
9	管理是一门管理是一门艺术，例如批评人的时候要考虑用正确的方法，让被批评的人不觉得难受，而且能愿觉到是在帮助他。				[[(38, 39, '如果'), '词类介词连接分句或句子', 'S', ('null', 'null', '而且')], [(45, 45, '帮'), '词类副词时间副词', 'M
10	管理是一门管理是一门艺术，例如批评人的时候要考虑用正确的方法，让被批评的人不觉得难受，而且能愿觉到是在帮助他。				[[(38, 38, '但'), '词类介词连接分句或句子', 'S', ('null', 'null', '而且')], [(44, 45, '仍旧'), '词类副词时间副词', 'S
11	管理是一门管理是一门艺术，例如批评人的时候要考虑用正确的方法，让被批评的人不觉得难受，而且能愿觉到是在帮助他。				[[(7, 8, '而且'), '词类介词连接分句或句子', 'M', (40, 40, 'None')], [(45, 46, '刚刚'), '词类副词时间副词', 'S', ('nul
12	管理是一门管理是一门艺术，例如批评人的时候要考虑用正确的方法，让被批评的人不觉得难受，而且能愿觉到是在帮助他。				[[(39, 39, '且'), '词类介词连接分句或句子', 'M', (40, 40, '而')], [(19, 19, '在'), '词类副词时间副词', 'M', (46, 46,
13	管理是一门管理是一门艺术，例如批评人的时候要考虑用正确的方法，让被批评的人不觉得难受，而且能愿觉到是在帮助他。				[[(38, 38, '能'), '词类介词连接分句或句子', 'M', ('null', 'null', '而且')], [(43, 43, '已'), '词类副词时间副词', 'S',
14	这样做，只这样做，只能暂时解决问题，所以要想完全解决这个难题，还需要找更好的办法。				[[(14, 14, '要'), '句子的类型复句因果复句', 'M', ('null', 'null', '以')], [(25, 25, '再'), '词类副词关联副词', 'S', (
15	这样做，只这样做，只能暂时解决问题，所以要想完全解决这个难题，还需要找更好的办法。				[[(13, 13, '以'), '句子的类型复句因果复句', 'M', (14, 14, '所')], [(26, 26, '也'), '词类副词关联副词', 'S', ('null',
16	还这样做，这样做，只能暂时解决问题，所以要想完全解决这个难题，还需要找更好的办法。				[[(15, 15, '要'), '句子的类型复句因果复句', 'M', ('null', 'null', '以')], [(0, 0, '还'), '词类副词关联副词', 'M', (26
17	这样做，只这样做，只能暂时解决问题，所以要想完全解决这个难题，还需要找更好的办法。				[[(13, 13, '也'), '句子的类型复句因果复句', 'S', ('null', 'null', '所以')], [(25, 25, '就'), '词类副词关联副词', 'S',

如上图所示，对数据集各项做如下说明：

（1）**source**：构造的**错误样本**，每个样本有**1–5**个错误点；

（2）**target**：**正确样本**；

（3）**source得分**：待人工标注的**负样本得分**；

人工评判的分数有4个档次，分别为：

0：该负样本构造不合理、或语句过分不通畅、或没有错误，舍弃；

1：该负样本不太合理；

2：该负样本比较合理；

3：该负样本构造合理，保留。

（4）**label得分**：待人工标注的**语法点标签得分**；

人工评判的分数有3个档次，分别为：

0：该语法点判断错误；

1：不确定该语法点是否正确；

2：该语法点判断正确。

（5）**op**：样本相应的错误位置、类型和修正值；

op数组含义：**[（错误起始index，错误终止index，待修改字符），标签，错误类型，（修改起始index，修改终止index，修改值）]**

错误类型说明：　'**W**'：Word Ordering Errors，**错序**；

'**M**'：Missing Words，**遗漏**；

'**S**'：Word Selection，**误用**。

下图给出相应例子：

source：种植物正这个季节长得很快，经过短短一个星期，它就长满了这面墙，叶子很厚，绿绿的。
target：这种植物在这个季节长得很快，经过短短一个星期，它就长满了这面墙，叶子很厚，绿绿的。
[(0, 0, '种'), 词类代词指示代词, 'M', (null, null, "这")]
[(3, 3, '正'), 词类副词时间副词, 'S', (null, null, "在")]
source：刚才听广播说明天能会下大雨，足球比赛怕恐要推迟了。
target：刚才听广播说明天可能会下大雨，足球比赛恐怕要推迟了。
[(8, 8, '能'), 词类动词能愿动词，'M', (null, null, '可')]
[(18, 18, '怕'), 词类副词情态副词，'W', (19, 19，'恐')]
source：刚才恐怕听广播说明天可能会下大雨，足球比赛要推迟了。
target：刚才听广播说明天可能会下大雨，足球比赛恐怕要推迟了。
[(2, 3, '恐怕'), 词类副词情态副词，'W', (21, 21, 'None')]

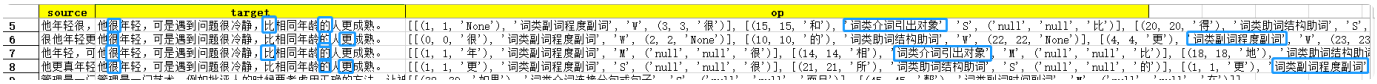
2. 语法点检查

根据op和label定位target中的语法点，并判断该语法点是否正确，**并在“label得分”一栏给定相应分数**。

如，下表所示的样例中，将“在”字对应的语法点错误匹配成“词类副词时间副词”，与实际不符，则其对应“label得分”应为0。

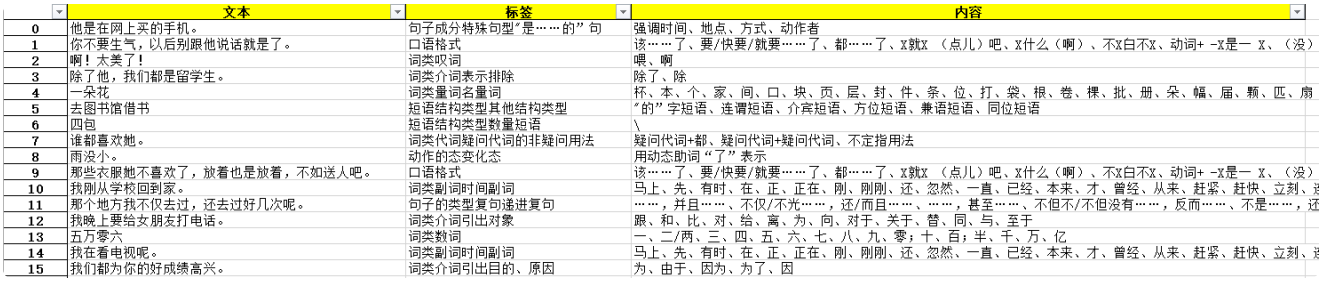
source	target	label得分	op
明天就可以有时网上查成绩不敢当，我估计这次考得不坏。	明天就可以在网上查成绩了，我估计这次考得不坏。	0	[[(12, 14, '不敢当'), '句子的类型复句紧缩复句', 'S', ('null', 'null', '了')], [(5, 6, '有时'), '词类副词时间副词', 'S', ('null', 'null', '在')]]

同时，还需注意！！！对于同一个正样本target，与它相匹配的每个负样本并非都对应着相同的语法点！！！不能仅通过判断该target第一个样本的语法点正确（错误） 就认定它的其余样本语法点均都正确（错误）！！！ *（如下图所示，5、7行对应“比”这个语法点进行操作，而6、8行则对“更”的语法点进行操作）*。



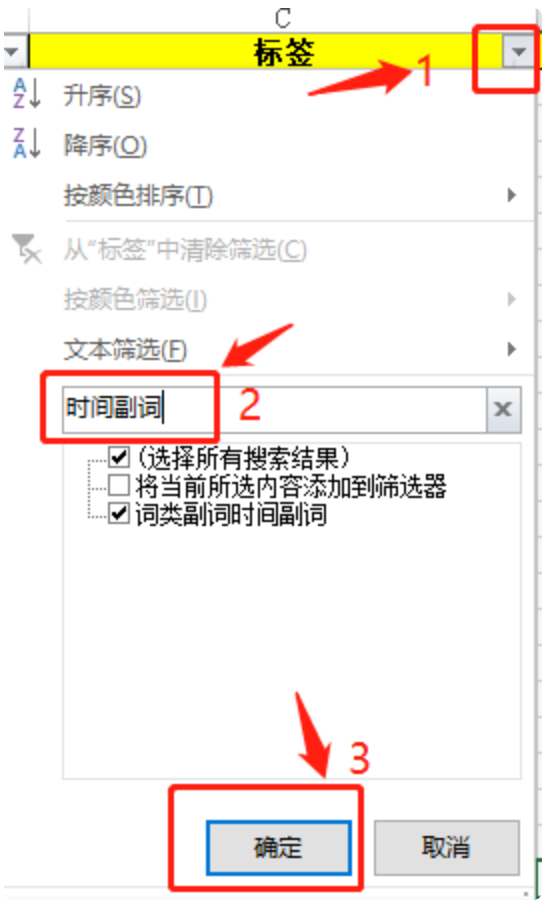
PS：附带给了 *参考.xlsx* 文件，在不确定某语法点是否正确时可对照该文件进行判断，或者上网查询。

如下图所示，*参考.xlsx* 文档给出了一系列 **语法点匹配正确的例句**（“文本”列），“内容”列展示了所有匹配该语法点的关键词。



	文本	标签	内容
0	他是在网上买的手机。	句子成分特殊句型“是……的”句	强调时间、地点、方式、动作者
1	你不要生气，以后别跟他说话就是了。	口语格式	该……了、要/快要/就要……了、都……了、X就X （点儿）吧、X什么（啊）、不X白不X、动词+ -X是一 X、（没）
2	啊！太美了！	词类叹词	喂、啊
3	除了他，我们都是留学生。	词类介词表示排除	除了、除
4	一朵花	词类量词名量词	杯、本、个、家、间、口、块、页、层、封、件、条、位、打、袋、根、卷、棵、批、册、朵、幅、届、颗、匹、扇
5	去图书馆借书	短语结构类型其他结构类型	“的”字短语、连谓短语、介宾短语、方位短语、兼语短语、同位短语
6	四包	短语结构类型数量短语	\
7	谁都喜欢她。	词类代词疑问代词的非疑问用法	疑问代词+都、疑问代词+疑问代词、不定指用法
8	雨没小。	动作的态变化态	用动态助词“了”表示
9	那些衣服她不喜欢了，放着也是放着，不如送人吧。	口语格式	该……了、要/快要/就要……了、都……了、X就X （点儿）吧、X什么（啊）、不X白不X、动词+ -X是一 X、（没）
10	我刚从学校回到家。	词类副词时间副词	马上、先、有时、在、正、正在、刚、刚刚、还、忽然、一直、已经、本来、才、曾经、从来、赶紧、赶快、立刻、
11	那个地方我不仅去过，还去过好几次呢。	句子的类型复句递进复句	……，并且……、不仅/不光……，还/而且……、……，甚至……、不但不/不但没有……，反而……、不是……，还
12	我晚上要给女朋友打电话。	词类介词引出对象	跟、和、比、对、给、离、为、向、对于、关于、替、同、与、至于
13	五万零六	词类数词	一、二/两、三、四、五、六、七、八、九、零；十、百；半、千、万、亿
14	我在看电视呢。	词类副词时间副词	马上、先、有时、在、正、正在、刚、刚刚、还、忽然、一直、已经、本来、才、曾经、从来、赶紧、赶快、立刻、
15	我们都为你的好成绩高兴。	词类介词引出目的、原因	为、由于、因为、为了、因

查询时，可通过excel“筛选”快速查找所需语法点，如下图所示。



3. 负样本筛选

观察source，筛掉其中**构造不合理的**、生活中**不可能犯的**、或是**正确没有语病**的样本，**并在“source得分”一栏给定相应分数**。

如下表中的样例，给出了对部分负样本的打分（打分的主观性比较强，但尽可能确保0分和3分的“稳定”）

source	target	source得分	op
这样做，只能暂时解决问题，所要想完全解决这个难题，再需要找更好的办法。	这样做，只能暂时解决问题，所以要想完全解决这个难题，还需要找更好的办法。	3	[[(14, 14, '要'), '句子的类型复句因果复句', 'M', ('null', 'null', '以')], [(25, 25, '再'), '词类副词关联副词', 'S', ('null', 'null', '还')]]
这样做，只能暂时解决问题，以所要想完全解	这样做，只能暂时解决问题，所以要想完全解决这	3	[[(13, 13, '以'), '句子的类型复句因果复句', 'W', (14, 14, '所')], [(26, 26,

决这个难题，也需要找更好的办法。	个难题，还需要找更好的办法。		也'), '词类副词关联副词', 'S', ('null', 'null', '还')]]
还这样做，只能暂时解决问题，所要想完全解决这个难题，需要找更好的办法。	这样做，只能暂时解决问题，所以要想完全解决这个难题，还需要找更好的办法。	1	[[(15, 15, '要'), '句子的类型复句因果复句', 'M', ('null', 'null', '以')], [(0, 0, '还'), '词类副词关联副词', 'W', (26, 26, 'None')]]
这样做，只能暂时解决问题，也要想完全解决这个难题，就需要找更好的办法。	这样做，只能暂时解决问题，所以要想完全解决这个难题，还需要找更好的办法。	1	[[(13, 13, '也'), '句子的类型复句因果复句', 'S', ('null', 'null', '所以')], [(25, 25, '就'), '词类副词关联副词', 'S', ('null', 'null', '还')]]
他年轻很，可是遇到问题很冷静，和相同年龄得人更成熟。	他很年轻，可是遇到问题很冷静，比相同年龄的人更成熟。	3	[[(1, 1, 'None'), '词类副词程度副词', 'W', (3, 3, '很')], [(15, 15, '和'), '词类介词引出对象', 'S', ('null', 'null', '比')], [(20, 20, '得'), '词类助词结构助词', 'S', ('null', 'null', '的')]]
很他年轻更，可是遇到的问题很冷静，比相同年龄人成熟。	他很年轻，可是遇到问题很冷静，比相同年龄的人更成熟。	0	[[(0, 0, '很'), '词类副词程度副词', 'W', (2, 2, 'None')], [(10, 10, '的'), '词类助词结构助词', 'W', (22, 22, 'None')], [(4, 4, '更'), '词类副词程度副词', 'W', (23, 23, 'None')]]
他年轻，可是遇到问题很冷静，相同年龄地人更成熟。	他很年轻，可是遇到问题很冷静，比相同年龄的人更成熟。	3	[[(1, 1, '年'), '词类副词程度副词', 'M', ('null', 'null', '很')], [(14, 14, '相'), '词类介词引出对象', 'M', ('null', 'null', '比')], [(18, 18, '地'), '词类助词结构助词', 'S', ('null', 'null', '的')]]
他更真年轻，可是遇到问题很冷静，比相同年龄所人成熟。	他很年轻，可是遇到问题很冷静，比相同年龄的人更成熟。	0	[[(1, 1, '更'), '词类副词程度副词', 'S', ('null', 'null', '很')], [(21, 21, '所'), '词类助词结构助词', 'S', ('null', 'null', '的')], [(1, 1, '更'), '词类副词程度副词', 'W', (23, 23, 'None')]]

同时，在人工进行评估时，我们***应尽量从标准的、语法规范的角度去评判***（如“相同年龄人”这种表达在生活中确实会出现，似乎也不构成一个错误的点。但从更规范的语法角度去理解，“年龄”和“人”之间缺少助词，不能直接相连，应改成“相同年龄的人”。）

PS：在进行数据筛选的同时，顺便**检查一下其op是否正确**（有很少一部分样本会有点bug=.=），有错误的地方人工修正一下~。

比如下表样例1，对比source和op后发现*[(0, 1, '多久'), '词类代词疑问代词', 'S', ('null', 'null', '谁')]*这个错误修正有误（究其原因，是因为在构造负样本时先将“谁”替换成“多久”，而后续在对“多”进行操作时错误定位到了“多久”而非“最多”上），对于这种构造出错的样本，直接删去这一行source、target、op所有数据。

又如样例2所示，它对“又脏又乱”中的2个“又”交换了位置，虽然*[(4, 4, '又'), '固定格式', 'W', (6, 6, '又')]*这个错误修正是对的，但它没有实质性表现和意义，故将其修改成样例3所示（删去对应多余操作）。

id	source	target	op
1	久在规定的时间内接到的球最多，谁就赢了比赛，所以这种游戏十分简单。	谁在规定的时间内接到的球最多，谁就赢了比赛，所以这种游戏十分简单。	[[(0, 1, '多久'), '词类代词疑问代词', 'S', ('null', 'null', '谁')], [(0, 0, '久'), '词类代词疑问代词', 'M', ('null', 'null', '多')]]
2	那个房间又脏又乱，星期六我去打扫了、整理一下。	那个房间又脏又乱，星期六我去打扫、整理了一下。	[[(4, 4, '又'), '固定格式', 'W', (6, 6, '又')], [(16, 16, '了'), '句子的类型复句紧缩复句', 'W', (20, 20, 'None')]]
3	那个房间又脏又乱，星期六我去打扫了、整理一下。	那个房间又脏又乱，星期六我去打扫、整理了一下。	[[(16, 16, '了'), '句子的类型复句紧缩复句', 'W', (20, 20, 'None')]]