SimCSE模型实验结果报告

实验背景

项目初启动阶段，业务场景数据一般比较少，如何基于小样本数据集来提升模型效果具有很大的研究价值。

SimCSE利用无标注样本自监督 学习方式能取得堪比有监督模型的效果(注：原论文是基于英文数据集训练，无中文)，值得拿来解读并在中文Chinese-STS数据集进行结果复现。

实验目的

探讨对比模型SimCSE在小样本学习中的可行性。

实验数据

实验数据采用Chinese-STS和Chinese-SNLI两个数据集

1. Chinese-STS：数据格式为 seqA||seqB||label，label代表seqA和seqB的相似程度，有0-5共6个级别；
2. Chinese-SNLI：数据格式为 seqA||seqB||seqC；
3. 训练数据：SimCSE是自监督学习训练，无需标注数据，训练数据我们从Chinese-STS(去label)和Chinese-SNLI中随机选择1w数据来训练；
4. 验证数据和测试数据：验证数据和测试数据需要label来计算指标，所以都选自Chinese-STS，其中 验证数据大小1458对，测试数据1361对。

实验内容

1. 先验知识

模型评测指标：

- spearman相关系数，统计两个变量之间的相关性，简单说就是两者在变大或者变小的趋势上多大程度上步调保持一致。

2. 验证集、测试集先验指标

SimCSE模型训练是基于预训练语言模型，如bert、bert_wwm等，我们先看预训练模型在验证集、测试集上的表现。

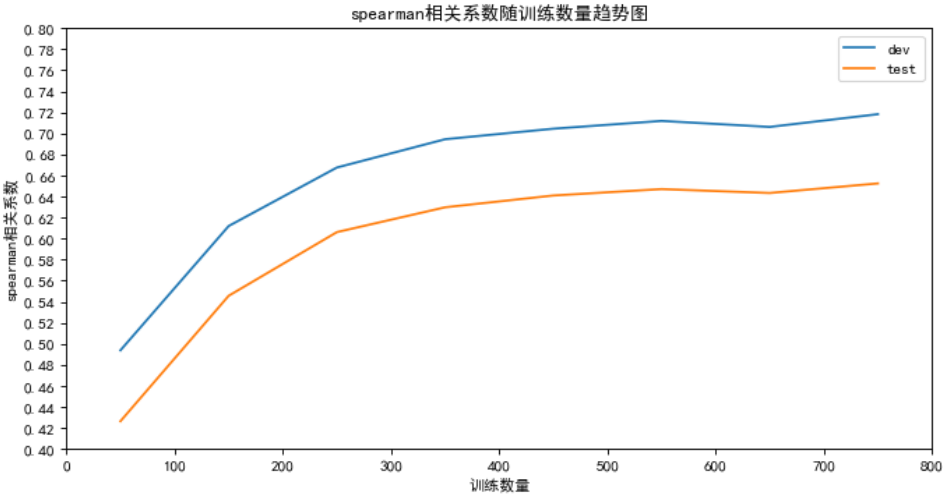
模型	STS-B dev	STS-B test
BERT-BASE	0.4228	0.3395
BERT-wwm	0.4465	0.3408
BERT-wwm-ext	0.3251	0.3172

实验结果：可以看出，预训练模型未经过SimCSE训练的情况下，在验证集、测试集表现效果都很差。

3. 模型在小样本数据集中的表现

模型参数：

1. 预训练模型: BERT-wwm-ext
2. 训练数据量: 50-800 以100的步长递增
3. epoch: 1
4. batch_size: 16
5. LR: 1e-5
6. dropout: 0.3
7. TMP_Coefficient: 0.05



实验结果：

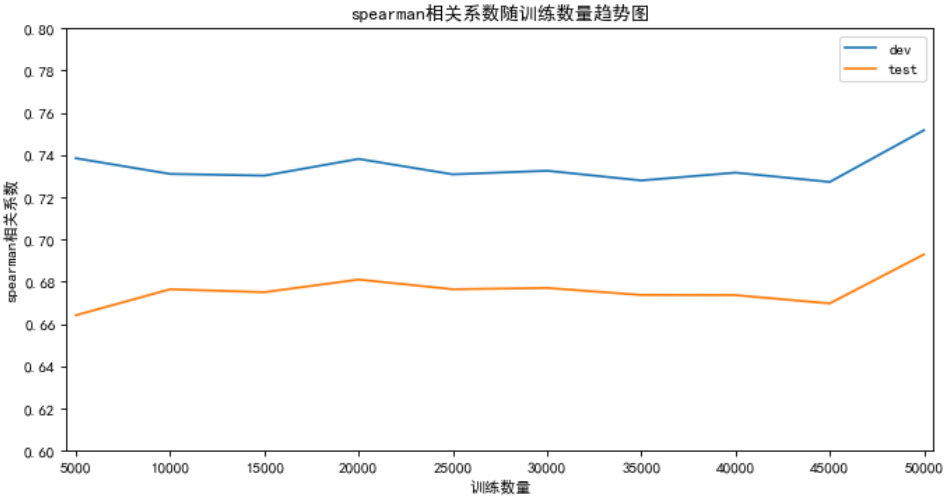
- 训练样本量小于400时，模型指标随着样本量大小急剧增加，大于400后趋于平稳；
- 证明模型在小样本数据量下有很好的学习能力，也很快遇到瓶颈，指标不会随着数据量增加继续增加。

4. 大数据量训练对模型指标的影响

数据才是王道，我们尝试下加大数据量，看看对模型指标的影响。

模型参数：

1. 预训练模型: BERT-wwm-ext
2. 训练数据量: 5000-50000 以5000的步长递增
3. epoch: 1
4. batch_size: 16
5. LR: 1e-5
6. dropout: 0.3
7. TMP_Coefficient: 0.05



实验结果： 可以看出较大数据集对模型提升效果甚微。

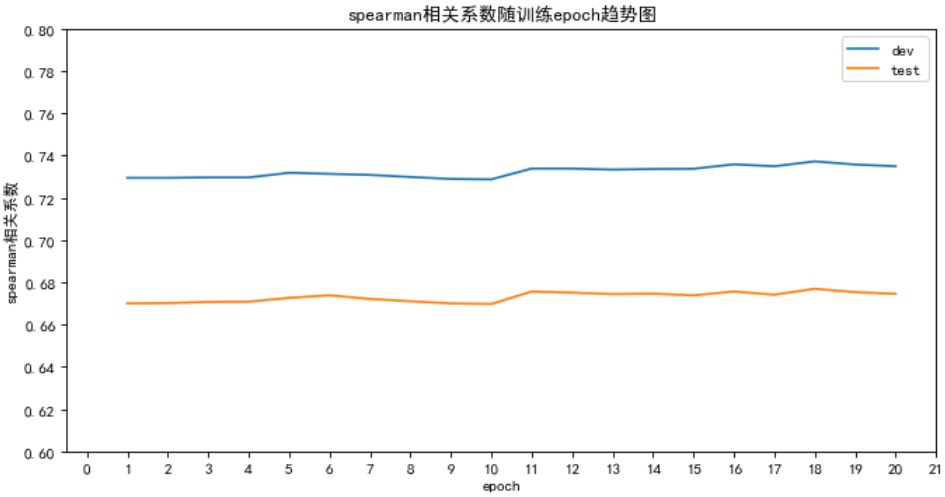
5. 模型指标随训练epoch趋势图

上面都只训练了1个epoch，我们尝试多训练几个epoch。

模型参数：

1. 预训练模型: BERT-wwm-ext
2. 训练数据量: 1w
3. epoch: 20
4. batch_size: 16
5. LR: 1e-5
6. dropout: 0.3

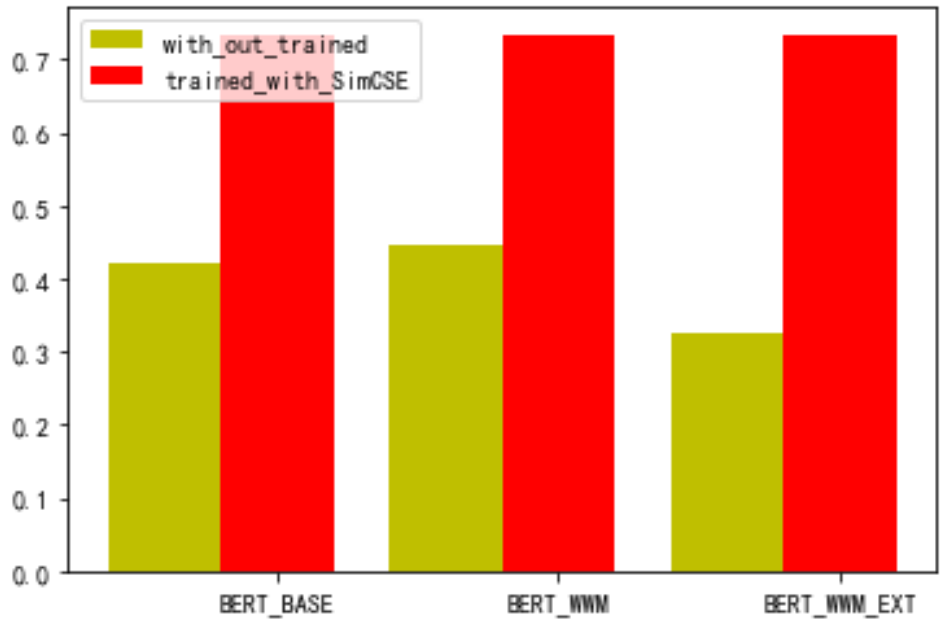
7. TMP_Coefficient: 0.05



实验结果：

- 增加epoch，类似增加数据量，对模型最终指标影响不大。

6. SimCSE训练效果一览



实验结果：

- 在预训练语言模型上，基于少量无标注数据用SimSEC进行自监督训练，在文本语义相似度任务上取得很好的提升效果。

实验结果

1、 SimCSE本质是一个句子表示模型，用来得到一个好的句子向量表示；

2、易训练，可基于少量的无标注样本自监督训练后用来提升文本语义相似任务，且能取得较好的效果；

3、场景应用：

1. FQA场景的问句召回，FQA一般存在大量的标准问句，如果利用模型实时召回，时间消耗巨大。SimCSE可先将问句转成向量进行召回并保证召回效果；
2. 对话系统中存在大量意图，存在多个意图模型，可通过SimCSE对意图进行预召回，减少意图识别计算量并保证召回效果；
3. 数据增强，某个类别标注样本缺乏，但存在大量未标注样本，可通过模型筛选出相似度高的数据，打上伪标签，供人工标注；
4. 意图OOV情况，待调研。