**Combining Unsupervised Pre-training and Annotator Rationales to Improve Low-shot Text Classification**

会议：EMNLP2019

代码：<https://github.com/mihaela-bornea/low-shot-text-classification>

**摘要**

监督学习模型在低样本任务（即很少有标记数据可用于训练的任务）上通常表现不佳。 改善低射学习的一种重要方法是使用无监督的预训练神经模型。 另一种方法是通过收集注释者依据（支持标签注释的说明）来获得更丰富的监管。 在这项工作中，我们将这两种方法与两种新颖的方法相结合，以改善低射文本的分类：一种简单的词袋嵌入方法； 以及基于BERT模型的更复杂的上下文感知方法。 在两个英文文本分类数据集的实验中，我们证明了通过结合预训练和基本原理而获得的可观性能提升。 此外，我们对一系列训练集大小的研究表明，当只有几十个训练实例或更少训练实例，而更复杂的模型（例如BERT或CNN） 需要更多的培训数据才能发光。

**1引言**

当对大量标记数据进行训练时，监督的机器学习模型在从文本分类到图像识别的各个领域展示出令人印象深刻的性能（Yogatama等，2017； Krizhevsky等，2012）。但是，在许多领域中，很难获得带标签的数据。 这可能是由于注释成本，或者仅仅是因为在系统必须做出下一个分类决策之前，没有可用的实例可以注释。 例如，想象一下一个个性化的电子邮件标记器，可以在其中创建新的标记类型，然后随时将其用于用户注释。 如果只有很少的培训可用，则模型性能会急剧下降。

已经提出了各种方法来优化在少量数据上训练的机器学习模型，目的是针对给定的训练大小提高性能或减少需要收集的训练实例的数量。 文本分类（和其他NLP任务）最突出的方法之一是从无监督的数据或无监督的预培训中进行迁移学习。 随着预训练词嵌入的引入，该方法已被广泛采用（Mikolov等，2013; Joulin等，2017）。 单词类型表示随后被扩展为具有预训练的，上下文相关的表示，可在广泛的NLP任务中获得强大的结果（Melamud等人，2016; Peters等人，2018; Howard和Ruder，2018）。 BERT（Devlin et al。，2018）是这方面研究中的佼佼者，它是一种基于多层注意力的神经模型，已在大量的纯文本上进行了预训练，然后针对特定任务进行了微调。 使用BERT的系统已在各种NLP任务（包括短文本分类）上实现了最先进的性能。

训练带有很少标记实例的文本分类器的另一种方法是用注释原理来补充类标签。 文本的注释原理通常是文本实例中单词的特定子序列，注释者认为这些单词是支持标签分配的证据。 Zaidan等（2007）是第一个研究使用注释器原理来训练文本分类器的人。 他们获得了关于IMDB电影评论数据集的手动原理（Pang等，2002），并表明可以使用这些原理修改SVM分类器以获得更高的准确性。 电影评论摘录中的粗体文字“也许这部电影的最大要素是其不断的惊喜和不可预测性”，是该数据集中带有注释原理的一个示例。 Abedin等。 （2011年）使用支持向量机和分类原理显示了相似的结果在航空安全报告系统（ASRS）的航空安全报告中引起原因标识符。

在这项工作中，我们将无监督预训练的功能与基本原理的细粒度指导相结合，提出了两种新颖的lowshot文本分类方法。 两种方法都可以在输入中促进类似基本原理的功能，以改善分类结果。 第一种方法使用基本原理监督来将简单的词袋文本偏向更具区分性的特征。 在第二种方法中，我们展示了如何结合实例标签和基本原理对预训练的BERT模型进行微调。 对IMDB和ASRS文本分类数据集上的一系列训练集大小进行的调查显示，对于多达几十个实例的训练集，我们的“词袋”方法大大优于所有基线。 对于较大的尺寸，我们的基于BERT的模型和CNN基线表现最佳。我们的代码是公开可用的。

**2相关工作**

如引言中所述，Zaidan等人（2007年）是第一个显示如何为电影评论添加注释的手动推理（Pang等人，2002年）的工作。他们还研究了标注工作量的影响，发现在标准标签标注之上每个实例的一些基本原理的附加标注使标注时间大约增加了一倍。 当足够多的实例可供注释时，此信息对于评估获取注释者基本原理和获取标准标记实例之间的权衡非常重要。但是，我们注意到，在某些情况下，可能无法使用更多的未标记实例，然后获得注释者依据可能是提高分类器准确性的唯一方法。

与我们的词袋理论基础最紧密相关的模型是Sharma和Bilgic（2018）提出的模型（在5.2节中表示为RA-SVM）。 它通过以与所使用分类器类型无关的方式调整特征权重来偏置输入表示。 更具体地说，它只使用one-hot词特征，并且仅在训练期间对非理性词特征权重执行简单的固定折扣。 相比之下，我们的模型学习的是文本的基于理性的嵌入表示形式，并且无论是在训练还是在测试时间都使用了这种表示形式

与我们的基本原理BERT模型最相关的模型是Zhang等人提出的模型 （2016）（在5.2节中表示为RA-CNN）。 该模型使用带有预训练词嵌入的CNN句子编码器。 它遵循两步方法，即训练一个句子分类器来识别包含基本原理的句子，并在整体“加权平均句子”文档表示中对它们进行加权。然后训练第二个分类器，以基于该表示进行最终标签预测。与非合理化的CNN以及合理化的SVM模型相比，它们显示出更高的准确性。 对于这项工作，我们的主要贡献是引入了联合学习技术，该技术可以同时学习文本分类和基本单词识别，并使整个方法适应BERT。 我们的模型识别句子中的理论范围的能力对于解释性似乎也很有用。 Tepper等（2013年）还选择了类似于Zhang（2016年）等人的两步法。使用具有稀疏特征的更多传统分类器。 他们训练一个分类器来识别类似基本原理的文本，然后训练第二个分类器来仅考虑类似基本原理的文本来标记输入实例。

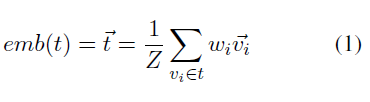
最后，小样本学习方法尝试使用通常被称为元学习的技术来学习如何仅用很少的标记实例来有效地训练新类别的分类器（Snell等，2017; Yu等，2018）。 Snell等（2017），例如，学习一个用于表示输入实例的嵌入函数，以便在暴露给少量带标签的示例后，可以有效地训练简单的最近原型分类器以识别新目标类的实例。 尽管在我们的工作中从很少的标记实例中学习了一个新类，但是前面的元学习过程依赖于聚集在许多紧密相关的类或任务上的大量标记实例，这是与我们低样本设置的重要区别。

**3偏倚词袋法**

在本节中，我们将描述我们提出的偏倚词袋法，该方法用于执行基于偏见的低射文本分类。

**3.1偏向原理的表示法**

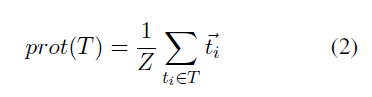
假设是单词v的向量表示法（一维或低维嵌入）。 我们将文本实例表示为其单词的L2归一化加权平均，从而提供文本嵌入功能：



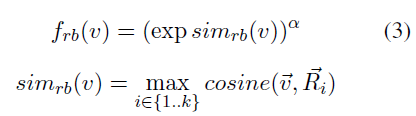
其中Z是归一化因子。

此嵌入是表示文本的标准词袋方法，其中通常使用权重（例如TF-IDF）将表示偏向于携带更重要信息的词。 我们假设，由于基本原理突出了对注释器分类决策的支持，因此将文本嵌入功能偏向基本原理类似的词会暴露出更具区分性的特征。 为此，给定一个估计单词v与训练数据集中带注释的基本原理文本之间相似度的基本原理偏差函数，则可以使用等式（1）计算基本原理偏差表示其中。 本节的其余部分描述了如何计算。

让文本实例集合T的原型表示形式为其文本实例向量的L2正则重心：



文本实例t和类标签c的基本原理是t的任何单词子序列r，都被注释为将c分配给t的实例。 类基本原理原型是为ci类标注的基本原理的原型表示形式，其中每个基本原理实例r使用配重均匀Eq（1）嵌入我们定义以下基本偏见函数：



也就是说，simrb（v）是单词向量与任何类基本原型的最大余弦相似度，而始终是simrb（v）的正数导数，并由超参数控制。更高的值意味着更大程度地放大了基本原理偏见的影响；= 0表示完全忽略了基本原理偏差。 有了这个基本偏见函数，我们现在可以使用Eq（1）计算文本实例的基本偏见表示形式。这些表示可以用作任何分类器的输入。

在早期的实验中，我们注意到目标领域中的许多非歧视性的常用词（例如电影，当然，男生）倾向于获得较高的基本偏见。 推测这是由于以下事实：这些数据集中的基本原理经常具有跨度，该跨度包括多个基本原理词以及其他词。 作为补偿，我们将类基本原理原型表示形式调整为：



Tall是所有类中所有训练实例的集合。 也就是说，我们从每个类基本原理原型中减去整个训练集原型矢量。 我们也对这个表示进行L2归一化。

图1是一个示例，说明了由我们的模型生成的实际偏差权重。 在这种情况下可以看出，最积极和最消极的指示词被“挑起”，如偏高的指示所示。 但是，还应注意，尽管该模型具有简单的序列不可知性，但它却错过了单词表达的情感极性反转。

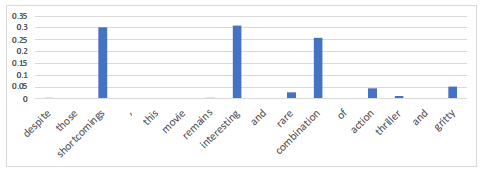
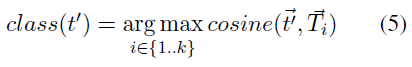


图1：在20个实例上训练的合理化模型所产生的偏差权重，来自示例电影评论中的摘录。

3.2分类器

可以训练各种类型的分类器，以在上一节中描述的合理化文本表示形式之上执行文本分类。 我们选择使用标准的SVM分类器以及简单的，最近的原型分类器进行实验，类似于先前的小样本学习工作中采用的方法（Snell等，2017）。 在训练和应用这些分类器之前，我们计算类基本原理原型以确定基本原理偏差函数frb。 然后我们使用这个函数来嵌入所有数据集文本实例，并将这些文本嵌入用作分类器的输入。 更具体地说，对于我们最近的原型分类器，给定一个训练的文本实例集合的集合；标记为类，我们用类原型表示每个类ci。 要分类一个新的文本实例t’，我们只需使用余弦相似度选择最接近的类原型：



我们的假设是，这种建议的方法在训练数据很少的情况下会比较好用，因为（a）在基本原理偏见的指导下，它可以更快地掌握单个歧视性单词（例如，对情感任务的好坏）; （b）结合使用预先训练的单词嵌入作为基础单词表示和基本原理偏差，将有助于泛化训练数据中未见的单词（例如，从best至good），并有助于有效的基本偏差相似度函数simrb（v） ; （c）与某些更复杂的神经模型（例如下一节中讨论的模型）相比，训练方案简单，需要学习的参数数量相对较少。

**4有偏见的BERT**

BERT（Devlin等人，2018）是一种上下文感知的神经网络模型，当用作各种受监督的NLP系统的基础时，已显示出出色的结果。 它的强大功能来自于在大量的纯文本上传递从无监督的培训中学到的上下文感知语言建模信息。 在本节中，我们说明如何使BERT适应于处理长文本，然后将基本原理监督与标准实例标签结合在一起。

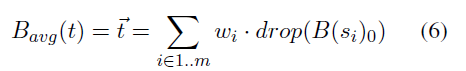
4.1 BERT的微调

BERT模型以单词令牌序列作为输入，为每个令牌输出上下文化的矢量编码，其中h0是整个序列的特殊编码。 该模型的预训练版本由Devlin（2018）等人训练。针对大文本集的语言建模和句子预测目标使得它们以这些编码捕获有价值的上下文信息。Devlin等（2018）应用了dropout，后面是一个简单的线性层，其中i> 0以执行标记级别分类，用于句子级别分类。 他们使用标准的交叉熵损失函数在特定于任务的标签上微调了整个模型。 通过这种方法，他们在各种NLP任务（例如命名实体识别（令牌级别）和单句情感分类（句子级别））上都取得了最新的成果。

4.2长文本的平均BERT分类器

尽管BERT被证明对短文本（例如单个句子）进行分类非常有效，但其计算复杂度是二次的，输入文本的长度导致文本长于几个句子的运行时间。 由于我们的输入文本可能很长，因此我们将它们拆分为句子，将BERT分别应用于每个句子，然后按以下方式平均输出。

给定一个包含句子集合的文本实例，我们将其编码为各个句子的加权平均值：



其中wi是加权平均值中使用的权重。我们使用线性层Lattn进行均匀权重实验，并以学习权重作为学习注意力。



最后，类似于原始的BERT方法，为了微调长文本实例的分类器，我们将线性层应用于编码的文本，并训练整个模型以使用损失函数预测目标标签：



其中lt是文本实例t的训练标签，CE是交叉熵损失函数。 图2说明了我们的体系结构。

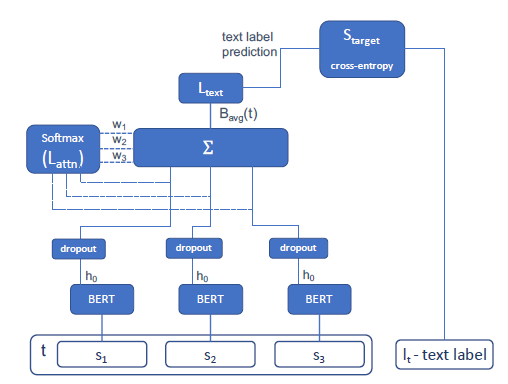


图2：平均BERT模型架构。 三个BERT框引用该模型的同一单个实例。 Lattn是Eq（7）的可选句子注意学习组件。

4.3基于偏见的BERT分类器

我们提出了两种方法来将合理性监督纳入我们的平均BERT文本分类器中。 第一种是基于直接的多任务学习，其中基本单词预测被认为是与主要文字标签预测任务共同训练的辅助任务。 第二个扩展了第一个，通过使用基本原理监督来估计每个句子的注意力。 然后，它使用该注意力（或重要性度量）从等式（6）中得出加权平均值。

简单的联合学习。 图3说明了添加到平均BERT的辅助模型。 训练相同的BERT模型以执行两个分类任务：（1）预测每个文本实例的标签（目标文本分类任务）； （2）为每个输入句子中的每个输入标记预测它是否是带注释的基本原理子序列的一部分（辅助任务）。 这是通过引入辅助损耗函数实现的：

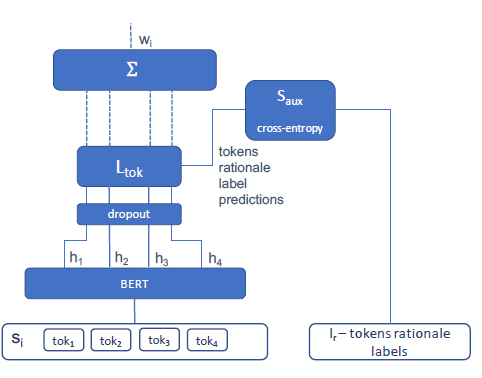
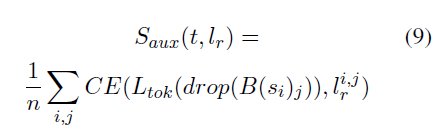


图3：基本原理偏差辅助模型组件。此处的BERT模型与用于主要目标任务的模型相同。wi是句子中的标记是基本原理的一部分的平均概率，可以在平均BERT中用作句子注意。



其中是句子i中标记j的理由标签，Ltok是用于学习对标记进行分类的线性层，n是文本t中标记的总数。 最后，每个输入文本t的多任务目标函数为：



我们的假设是，辅助任务将把有关输入单词相关性的较低级信息注入BERT模型，这将有助于提升主要任务分类器。

有偏见的关注。 图3中的虚线说明了我们提出的通过使用基本信号来学习如何识别输入文本中更重要的句子来扩展简单联合学习的方法。 为了训练该模型，我们在等式（10）中使用相同的多任务丢失功能。但是，这次，我们不是在平均的BERT中使用统一的权重，而是将句子的注意力权重视为其每个标记在理论范围内的概率的平均值。 这些权重用于计算等式（6）中文本输入的最终加权平均表示。在我们的实验中，我们发现最好不要将错误从Starget反向传播到令牌级基本原理分类器（即沿虚线）。 因此，仍然仅基于Saux训练基本原理分类器。

这里的假设是，该模型可以通过更明确地注入归纳性偏见而成功，该归纳性偏见是：对于最终分类任务而言，具有基本意义的句子比其他句子更重要。 在训练数据较少的情况下，引入更多的归纳偏差尤为重要。 但是，与上一节中介绍的词袋法相比，我们认为此处介绍的两种基于BERT的方法可能都面临着极少的训练数据挑战，因为底层模型和训练方案更加复杂。

图4显示了一个由我们的偏向注意的BERT模型计算出的句子注意和词理概率的示例。

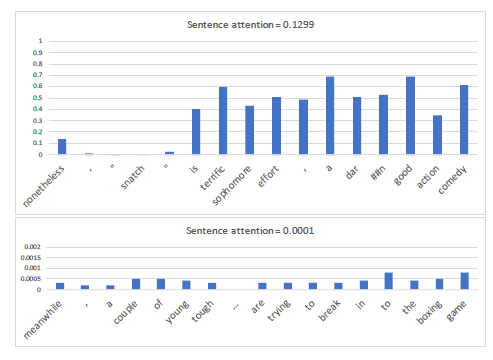


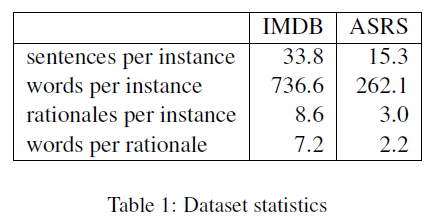
图4：根据我们偏爱的注意力BERT模型，在IMDB电影评论中获得最高关注度（顶部）和最低关注度（底部）的句子节选。 条形表示模型为作为基本原理一部分的每个单词分配的概率。

该模型具有上下文感知功能，它捕获单词的相关序列，而不仅仅是发现关键词。 这些句子和标记权重除了对分类任务有用之外，对模型的可解释性也很有用。

5实验设置

我们使用两个英文文本分类数据集进行了实验。 Zaidan等。 （2007年）添加了支持来自互联网电影数据库（IMDB）数据集的1800条电影评论中的正面/负面类别标签的基本原理（Pang等2002））。 我们使用900条评论进行培训，使用450条进行开发，使用450条进行测试，并且在每次分组中都保持了原来的正负平衡。

航空安全报告系统（ASRS）数据集包含航空安全事件的报告。我们使用了Abedin等人的版本（2011），培训集中有1233个报告，其中100个用于开发，1000个用于测试。 报告被标记为14种安全事件原因类别中的一种或多种。 为了创建可与IMDB数据集相媲美的二进制分类数据集，我们提取了一个平衡的报告子集，这些报告子集均标有“熟练程度”标签或“物理环境”标签（但不能同时包含两者）。 我们选择这两个特殊标签是因为它们的实例数量相似，并且属于数据集中最常见的标签之一。 这样就产生了386份训练报告和392份测试报告。 表1提供了有关这两个数据集的一些其他统计信息。



**5.1实验规程**

为了测试假设我们提出的合理化方法可以通过较少的训练数据改善分类的假设，我们进行了实验，将使用的训练数据量从2个实例更改为大约400个。对于每个训练量n，我们将n个训练实例采样为 训练模型并在整个开发/测试集上进行测试。 我们对该实验重复了30次（每n次），并报告了平均准确度。 在对IMDB和ASRS测试数据运行之前，我们仅基于对IMDB开发集的调整来固定所有参数（对ASRS数据未进行调整）。

**5.2比较方法**

RA-SVM是Sharma和Bilgic（2018）提出的合理性增强SVM，它通过在训练过程中以固定的折扣率对非理性字词特征进行折扣来使文本表示偏向于合理性。 所有单词在打折前均经过TF-IDF加权。 我们对IMDB开发集进行了不同的折扣试验，发现0.1给出了最佳结果。 虽然Sharma和Bilgic（2018）使用单字词表示法，但我们也尝试对其进行密集的字词嵌入，但这会降低性能。 我们还报告了不折不扣的结果-一个简单的SVM基线。

RA-CNN是Zhang等人在原理上增强的CNN。 （2016），在第2节中进行了描述。我们将作者的实现与经过预先训练的word2vec Google新闻单词嵌入一起使用，具有300个维度，可处理300万个单词。5我们在开发集上调整了批量大小，辍学率，学习率和时期数 。 我们还报告了没有合理权重的结果-一个简单的CNN基线。

RB-BOW-SVM和RB-BOW-PROTO是SVM和最接近原型的分类器，在我们提出的基本偏向词袋（BOW）文本表示形式上运行（第3节）。 我们调好了开发集上的超参数，两个分类器的最佳结果均为= 6。 BOW-PROTO是基于BOW的原型分类器，没有基本原理偏倚（= 0）。 我们使用与CNN和RA-CNN相同的word2vec嵌入尝试了单字词表示和单词嵌入。

AVG-BERT和ATTN-WAVG-BERT是我们建议的基于BERT的模型，分别使用统一的加权平均和基于注意力的句子加权平均值进行无理训练（第4.2节）。 RB-AVG-BERT是将基本原理预测作为辅助任务的联合模型，而RB-WAVG-BERT是也使用基本原理来估计句子级注意力的方法（第4.3节）。 对于以上所有内容，我们分别对IMDB和ASRS数据集使用了预先训练的模型“ bertbaseuncased”和“ bert-base-cased”。 我们使用默认的超参数，除了使用5e-6和10个纪元的学习率进行微调外，我们发现它们对于带光微调的AVG-BERT更好。 为了使计算图适合GPU内存（NVIDIA Tesla V100），我们修剪了长度超过48个单词的句子，并修剪了包含64个以上句子的文本。 我们使用BERT的PyTorch实现来实现我们的模型。

ULMFiT（Howard and Ruder，2018）是一种在多个数据集上实现最新文本分类结果的最新方法。 它分三个步骤进行训练：（1）在大型语料库（WikiText-103）上训练通用域递归神经网络语言模型； （2）不考虑类标签，将语言模型微调至目标任务的领域数据； （3）以学习到的微调语言模型的编码器为起点，微调任务的分类器。 我们在开发集上使用了ULMFiT的fast.ai7实现，调整批处理大小，辍学倍增因子，学习率和时期数。 ULMFiT当前没有合理化版本。

**6结果**

在本节中，我们将从调查IMDB开发集上提出的模型的不同变体的性能开始。 最有前途的设置然后在IMDB和ASRS测试集上进行评估。 由于我们的两个数据集都是二进制且平衡的，因此请注意随机基准精度为50％。

**6.1开发集调查**

为了证明我们的偏见词袋方法的各个组成部分的贡献，我们对开发集进行了一些消融实验，如图5所示。比较的变体包括：RB-BOW-PROTO：我们的 带有预训练词嵌入的完整方法； BOW-PROTO：无理由偏见； OH：RB-BOW-PROTO具有单词表示法，而不是词嵌入； NO-ADJUST：RB-BOW-PROTO，不通过减去通用原型向量而减去普通字（式（4））； 单：RBBOWPROTO仅具有一个代表所有基本原理的基本原理原型，而不是每个类的一个基本原理原型（等式（3））； 和INSTANCE：RB-BOW-PROTO，具有针对每个文本实例的基本原理原型，而不是针对每个类的实例。 可以看出，结合使用预训练和基本原理对于我们完整方法的成功至关重要。 首先，使用预训练词嵌入的重要性通过使用单热编码和少于60个实例的训练集时的性能显着下降而显而易见。 其次，BOW-PROTO在所有列车型号上的性能大幅下降都可以看出合理性偏见的重要性。 有了这两个组成部分，我们可以看到RB-BOW-PROTO能够在很少的训练实例的情况下，在50％的随机基准之上执行明显的工作。 我们注意到RBBOWSVM表现出与RB-BOW-PROTO类似的性能。 为简洁起见，我们在此处未提供详细的报告，但在测试结果部分中确实包含了RB-BOWSVM。

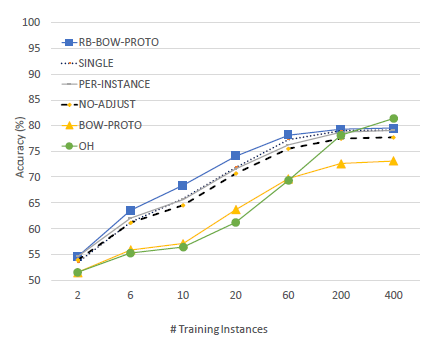


图5：在IMDB开发人员集上进行的烧蚀实验。



图6：IMDB开发人员集中基于BERT的模型。

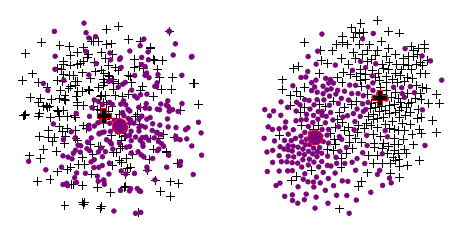


图7：没有理由偏见（左）和有（右）的词袋文字嵌入。 + / o标记是肯定/否定的开发人员文本实例。 较大的标记是在培训期间学习的正面/负面类原型。

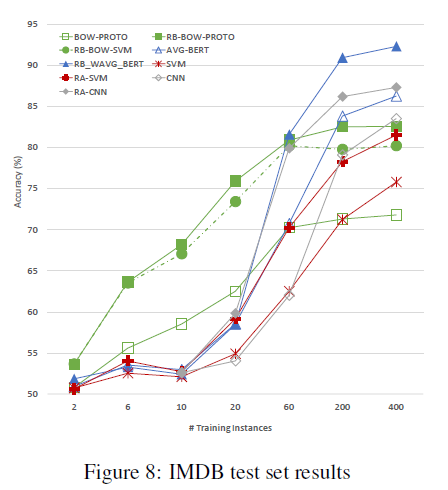
为了说明使用我们建议的基本偏向嵌入函数的潜在优点，我们在图7中显示了带有和不具有基本偏向的文本嵌入的二维图。我们从IMDB数据集中随机采样了大小为60的训练集并使用了它。 学习合理的嵌入功能。 我们绘制了从训练集以及IMDB开发集实例的嵌入中学到的正负类原型。 我们使用T-SNE将这些300维嵌入减少到可视化的两个维度。 合理化的表述似乎可以更好地区分积极实例和消极实例。 特别是，我们注意到类原型的表示之间的距离增加，并且开发集实例围绕它们各自的类原型的聚类更加清晰。对于此数据子集，使用基本偏向嵌入法时，最接近的原型分类器达到78.1％的准确性，而使用无偏表示法则为69.7％。

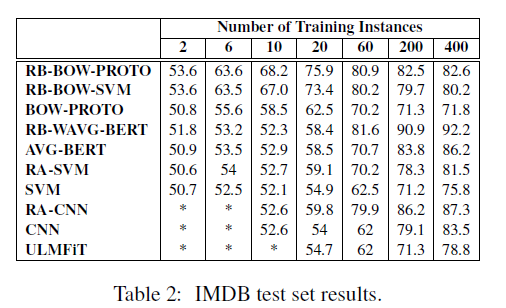
接下来，我们研究IMDB数据集上基于BERT的模型的性能。 图6显示了开发集上的结果。 首先，我们注意到，与简单的单词袋方法不同，所有这些模型都需要至少20个训练实例，然后才能在随机50％基准上有意义地进行改进。 另一方面，在200个或更多训练实例的情况下，所有基于BERT的模型都明显胜过词袋法。 更具体地说，我们看到使用基本原理监视的RB-AVG-BERT和RB-WAVG-BERT明显优于仅使用文本标签监视的AVG-BERT和ATTN-WAVG-BERT，最多可提高10个精度点。 在RB-AVG-BERT和RB-WAVG-BERT之间，后者具有更大的感性偏置，性能更强劲。

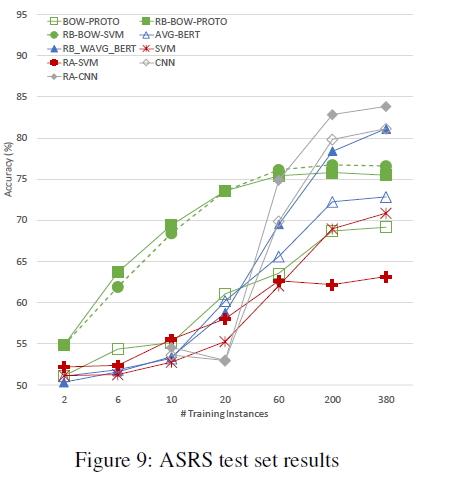
**6.2测试集结果**

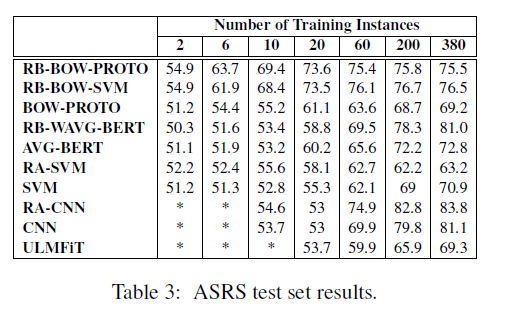
图8和表2显示了我们最佳的基本偏向方法及其非合理方法（基于开发集性能选择）的IMDB测试集结果，以及所有基线。8 字词模型RB-BOW-PROTO和RB-BOW-SVM结合了基本原理偏向和预训练的词嵌入，可大大超越所有其他基准，最多可为20多个绝对精度点提供训练 尺寸不超过20。 在这些特别小的数据集上，其他可感知基本原理的基准RASVM和RA-CNN效果不佳。 对于规模大于等于60的大型培训集，所有系统的性能均符合预期。 尤其是，基于CNN和BERT的更为复杂的模型在更多监督下可以显着提高性能。将预训练与基本原理相结合的RA-CNN和我们的RB-BERT都从基本原理监督中受益匪浅，绝对精度提高了30多个点。绝对而言，我们的BERT模型在此训练规模范围内胜过其CNN同行和所有其他基线。

最后，图9和表3显示了ASRS测试集的结果。 这里的趋势相似，显示了我们的基本原理偏向方法的强大优势。唯一的显着差异是，在此数据集中，CNN方法的性能优于基于BERT的方法的原理和非理性版本。 我们注意到，RA-CNN基线与我们的基线采用相似的方法，因为它还将预训练（预训练词嵌入）与基本原理相结合。 在特定数据集上，一种方法优于另一种方法的原因可能与预训练语料库与测试数据集之间的兼容性或对超参数调整的敏感性有关（我们未对ASRS开发集进行任何调整）。









**7结论**

在这项工作中，我们解决了有监督的机器学习中的一个重要挑战，即对大量标记训练数据的依赖。 我们证明，通过在各种方法中将无监督的预训练与注释器原理相结合，可以在低射文本分类中获得显着的性能提升。为此，我们提出了两种新颖的方法，它们可以在一系列列车尺寸上共同提供强大的结果。 我们使用各种基准，数据量和消融度进行了实验，以帮助了解哪种方法最适合不同数量的可用训练数据。 最值得注意的是，我们显示了带有预训练单词嵌入的简单单词袋方法最适合非常小的火车集，而基于预训练语言模型的更复杂方法则在有更多数据可用时更为出色。