深层语境化词语表示

马修·彼得斯†，马克·诺伊曼†，莫希特·艾耶†，马特·加德纳†，
{ MatthWp, MARKN, MOHITI, Mattg } AeleNay.ORG

Christopher Clark*，肯顿李*，Luke Zettlemoyer†*
{ccDang-Kutnl, LSZ}@ C.Walntut.EDU

†艾伦人工智能研究所
*华盛顿大学Paul G. Allen计算机科学与工程学院

摘要

我们引入了一种新的深层语境化词语表示，它模拟了（1）单词使用的复杂特征（例如，语法和语义），以及（2）这些用法如何在语言上下文中变化（即模型多义词）。我们的单词向量是深度双向语言模型（biLM）的内部状态的学习函数，它在大型文本语料库中预先训练。我们表明，这些表示可以很容易地添加到现有模型中，并显着改善六个具有挑战性的NLP问题的技术发展水平，包括问答，文本蕴涵和情感分析。我们还提供了一个分析，表明暴露预训练网络的深层内部是至关重要的，允许下游模型混合不同类型的半监督信号。

1 介绍

预先训练的单词表示（Mikolov等。，2013; Pennington等。，2014）是许多神经语言理解模型的关键组成部分。但是，学习高质量的表示可能具有挑战性。它们应该理想地模拟（1）单词使用的复杂特征（例如，语法和语义），以及（2）这些用途如何在语言上下文中变化（即，模型多义词）。在本文中，我们介绍了一种新型的深层语境化词语表示，可直接解决这两个难题，可以轻松地集成到现有模型中，并在一系列具有挑战性的语言理解问题的每个考虑案例中显着改善现有技术水平。

我们的表示与传统的单词类型嵌入不同，因为每个标记被赋予一个表示，该表示是整个输入句子的函数。我们使用从双向LSTM派生的向量，该双向LSTM通过耦合语言训练大文本语料库中的语言模型（LM）目标。出于这个原因，我们称它们为ELMo（嵌入式语言模型）表示。与以前学习情境化单词向量的方法不同（彼得斯等人。，2017; 麦肯 等。，2017），ELMo表示是深层的，因为它们是biLM的所有内部层的函数。更具体地说，我们学习了每个结束任务的每个输入字上方堆叠的矢量的线性组合，这显着提高了仅使用顶部LSTM层的性能。

以这种方式组合内部状态允许非常丰富的单词表示。使用内在评估，我们表明较高级别的LSTM状态捕获了词义的上下文相关方面（例如，它们可以在没有修改的情况下使用以在监督的词义检测消歧任务上很好地执行），而较低级别的状态模拟语法的方面（例如，它们可用于进行词性标注）。同时暴露所有这些信号是非常有益的，允许学习模型选择对每个最终任务最有用的半监督类型。

大量实验表明ELMo表示在实践中非常有效。我们首先表明，它们可以很容易地添加到现有模型中，用于六种不同且具有挑战性的语言理解问题，包括文本蕴涵，问答和情感分析。单独添加ELMo表示在每种情况下都显着改善了现有技术水平，包括高达20％的相对误差减少。对于可以进行直接比较的任务，ELMo优于CoVe（麦肯 等。，2017），使用神经机器翻译编码器计算情境化表示。最后，对ELMo和CoVe的分析表明，深层表征优于表现

那些只来自LSTM的顶层。我们训练有素的模型和代码是公开的，我们希望ELMo能够为许多其他NLP问题提供类似的收益。[1]

2 相关工作

由于它们能够从大规模未标记文本中捕获单词的句法和语义信息，所以预训练的单词向量（Turian等。，2010; Mikolov等。，2013; Pennington等。，2014）是大多数最先进的NLP架构的标准组件，包括问答（刘等人。，2017），文字蕴涵（陈等人。，2017）和语义角色标记（他等人。，2017）.然而，这些用于学习单词向量的方法仅允许每个单词的单个上下文无关表示。

以前提出的方法通过用子词信息丰富它们来克服传统词向量的一些缺点（例如，Wieting等。，2016; Bojanowski等。，2017）或为每个词义学习单独的向量（例如，Neelakantan等。，2014）.我们的方法还通过使用字符卷积从子字单元中受益，并且我们无缝地将多感知信息结合到下游任务中，而无需明确地训练以预测预定义的感知类。

最近的其他工作也侧重于学习依赖于上下文的表示。上下文2VECMelamud等。，2016）使用双向长短期记忆（LSTM;Hochreiter和Schmidhuber，1997）围绕一个枢轴词编码上下文。用于学习上下文嵌入的其他方法包括表示中的枢轴词本身，并且用监督神经机器翻译（MT）系统（CoVe;麦肯等人。，2017）或无监督的语言模型（彼得斯等人。，2017）.这两种方法都受益于大型数据集，尽管MT方法受到并行语料库大小的限制。在本文中，我们充分利用了丰富的单语数据，并在大约3000万句话的语料库中训练我们的biLM（Chelba等人。，2014）.我们还将这些方法概括为深层的上下文表示，我们在各种不同的NLP任务中表现出良好的工作效果。

[1] http://allennlp.org/elmo

先前的工作还表明，深度biRNN的不同层编码不同类型的信息。例如，在深度LSTM的较低级别引入多任务语法监督（例如，词性标签）可以提高更高级别任务的整体性能，例如依赖性解析（Hashimoto等人。，2017）或CCG超级标记（Søgaard和Goldberg，2016）.在基于RNN的编码器 - 解码器机器翻译系统中，Belinkov等。（2017）表明在2层LSTM编码器的第一层学习的表示更好地预测POS标签然后第二层。最后，LSTM的顶层用于编码单词上下文（Melamud等。，2016已被证明可以学习单词意义的表示。我们表明，类似的信号也是由ELMo表示的修改语言模型目标引起的，并且学习混合这些不同类型的半监督的下游任务的模型是非常有益的。

戴和勒（2015）和Ramachandran等。（2017）预训练编码器 - 解码器对使用语言模型和序列自动编码器，然后微调任务特定监督。相反，在使用未标记的数据预训练biLM之后，我们修复权重并添加额外的任务特定模型容量，从而允许我们在下游训练数据大小决定较小的监督模型的情况下利用大的，丰富的和通用的biLM表示。

3 ELMo：语言模型的嵌入

与最常用的词嵌入不同（钢笔- nington等。，2014），ELMo单词表示是整个输入句子的功能，如本节所述。它们是在具有字符卷积的两层biLM之上计算的（Sec。3.1），作为内部网络状态的线性函数（Sec。3.2）.这种设置允许我们进行半监督学习，其中biLM是大规模预训练的（Sec。3.4并且很容易融入广泛的现有神经NLP架构（Sec。3.3）.

3.1 双向语言模型

给定N个令牌的序列，（t_1，t_2，...，t_n），前向语言模型通过对以下事项的概率进行建模来计算序列的概率。

ken t_k 给出历史 $(t_1, \ldots, t_{k-1})$：

$$p(t_1, t_2, \ldots, t_n) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1}).$$

最新的最先进的神经语言模型（Jozefowicz等人。，2016; Melis等人。，2017; Merity等人。，2017）计算与上下文无关的令牌表示 x_k^{lm}（通过令牌嵌入或CNN over characters）然后通过L层前向LSTM的前提。在每个位置k，每个LSTM层输出依赖于上下文的表示 $\overrightarrow{h}_{k,j}^{lm}$ 其中j = 1,...，L。顶层LSTM输出 $\overrightarrow{h}_{k,L}^{lm}$ 用于预测下一个代币 t_{k+1} 具有Softmax层。

后向LM类似于前向LM，除了它反过来在序列上运行，预测给定未来背景的前一个标记：

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N).$$

它可以以类似于前向LM的方式实现，每个后向LSTM层j在L层深度模型中产生表示 $\overleftarrow{h}_{k,j}^{lm}$ 给出 $(t_{k+1}, \ldots, t_n)$。

biLM结合了前向和后向LM。我们的公式共同最大化了前后方向的对数可能性：

$$\sum_{k=1}^{N} \left(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_n; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \right).$$

我们在前向和后向上绑定令牌表示（Θ_x）和Softmax层（Θ_s）的参数，同时在每个方向上保持LSTM的单独参数。总的来说，这个公式类似于彼得斯等人。(2017)，除了我们在方向之间共享一些权重而不是使用完全独立的参数。在下一节中，我们将介绍一种学习单词表示的新方法，这种方法是biLM层的线性组合，从而脱离了以前的工作。

3.2 埃尔莫

ELMo是biLM中的中间层表示的任务特定组合。对于每个标记t_k，一个L层biLM计算一组2L + 1表示

$$R_k = \{x_k^{lm}, \overrightarrow{h}_{k,j}^{lm}, \overleftarrow{h}_{k,j}^{lm} \mid j = 1, \ldots, L\} = \{h_{k,j}^{lm} \mid j = 0, \ldots, L\},$$

其中$h_{k,0}^{lm}$是令牌层和$h_{k,j}^{lm} = [\overrightarrow{h}_{k,j}^{lm}; \overleftarrow{h}_{k,j}^{lm}]$，对于每个biLSTM层。

为了包含在下游模型中，ELMo将R中的所有层折叠成单个载体，$ELMo_k = E(R_k; \Theta_e)$。在最简单的情况下，ELMo只选择顶层$E(R_k) = h_{k,L}^{lm}$，如TagLM（彼得斯等人。，2017）和CoVe（麦克-Cann等人。，2017）。更一般地说，我们计算a所有biLM层的任务特定权重：

$$\text{ELMo}_k^{\text{任务}} = E(R_k; \Theta^{\text{任务}}) = \gamma^{\text{任务}} \sum_{j=0}^{L} s_j^{\text{任务}} h_{k,j}^{LM}. \quad (1)$$

在 $s^{\text{任务}}$ 是softmax标准化的权重和标量参数$\gamma^{\text{任务}}$允许任务模型缩放整个ELMo向量。γ具有实用性有助于优化过程的重要性（有关详细信息，请参阅补充材料）。考虑到每个biLM层的激活具有不同的分布，在某些情况下它也有助于应用层标准化（Ba等人。，2016）在加权之前到每个biLM层。

3.3 将biLM用于受监督的NLP任务

给定预先训练的biLM和目标NLP任务的监督架构，使用biLM来改进任务模型是一个简单的过程。我们只需运行biLM并记录每个单词的所有图层表示。然后，我们让结束任务模型学习这些表示的线性组合，如下所述。

首先考虑没有biLM的监督模型的最低层。大多数受监督的NLP模型在最低层共享一个共同的架构，允许我们以一致，统一的方式添加ELMo。给定一系列令牌$(t_1, \ldots, t_n)$，使用预先训练的字嵌入和可选的字符，为每个令牌位置形成与上下文无关的令牌表示x_k是标准的。基于陈述。然后，该模型形成上下文敏感表示h_k，通常使用双向RNN，CNN或前馈网络。

要将ELMo添加到监督模型，我们首先冻结biLM的权重然后

连接ELMo矢量ELMo^{任务} $_k$ 同 $\mathbf{x}_k$ 并传递ELMo增强表示[x_k;ELMo^{任务}]进入任务RNN。对于某些任务（例如，SNLI，SQuAD），我们通过在任务RNN的输出处包括ELMo，通过引入另一组输出特定线性权重并用[h_k替换h_k来观察进一步的改进;毛毛^{任务}。由于监督模型的其余部分保持不变，这些添加可以在更复杂的神经模型的$_k$背景下发生。例如，参见Sec中的SNLI实验。4 其中双重关注层遵循biLSTM，或者是共聚模型在biLSTM之上分层的共参考分辨率实验。

最后，我们发现向ELMo添加适量的辍学是有益的（Srivastava等。，2014并且在某些情况下，通过在损耗中加入λw^{2} 来规范ELMo权重。这对ELMo重量施加了归纳偏差，以保持接近所有biLM层的平均值。

3.4 预先训练的双向语言模型架构

本文中预先训练的biLM类似于中的体系结构Jo'zefowicz 等 人 。 (2016 ） 和 Kim 等 人 。(2015），但修改为支持两个方向的联合训练并在LSTM层之间添加残余连接。我们专注于这项工作中的大规模biLMs彼得斯等人。(2017)强调了将biLM用于仅前向LM和大规模训练的重要性。

为了平衡整体语言模型的困惑与下游任务的模型大小和计算要求，同时保持纯粹的基于字符的输入表示，我们将单个最佳模型CNN-BIG-LSTM 中 的 所 有 嵌 入 和 隐 藏 维 度 减 半 。Jo'zefowicz等人。 (2016).最终模型使用L = 2个biLSTM层，其具有4096个单元和512个维度投影以及从第一层到第二层的剩余连接。上下文不敏感类型表示使用2048个字符的n-gram 卷 积 滤 波 器 ， 后 跟 两 个 公 路 层（Srivastava等。，2015并且线性投影低至512表示。因此，biLM为每个输入令牌提供三层表示，包括由于纯字符输入而在训练集外部的表示。相反，传统的单词嵌入方法仅为固定词汇表中的标记提供一层表示。

在1B Word上训练了10个时代基准（Chelba等人。，2014），平均前向和后向困惑为39.7，而前向CNN-BIG-LSTM为30.0。通常，我们发现前向和后向的困惑大致相等，后向值略低。

预训练后，biLM可以计算任何任务的表示。在某些情况下，在特定于域的数据上微调biLM会导致困惑度显着下降和下游任务性能的提高。这可以被视为biLM的一种域转移。因此，在大多数情况下，我们在下游任务中使用了微调的biLM。有关详细信息，请参阅补充材料

4 评估

表1 显示了ELMo在各种六种基准NLP任务中的性能。在所考虑的每项任务中，简单地添加ELMo都会建立一种新的最先进的结果，与强基础模型相比，相对误差降低了6-20％。这是跨多种模型体系结构和语言理解任务的非常通用的结果。在本节的其余部分，我们提供了各个任务结果的高级草图;请参阅补充材料以获取完整的实验细节。

回答斯坦福问题答疑数据集（SQuAD）的问题（Rajpurkar等人。，2016）包含100K +众群来源的questionanswer对，其中答案是给定维基百科段落中的跨度。我们的基线模型（克拉克 和加德纳, 2017）是双向注意流模型的改进版本Seo等人。（比法达夫;2017).它在双向关注组件之后添加了一个自我关注层，简化了一些池化操作，并将LSTM替换为门控循环单元（GRU;卓 等。，2014).将ELMo添加到基线模型后，测试集F_1 从81.1％提高了4.7％至85.8％，相对于基线的相对误差降低了24.9％，并将整体单一模型的最新技术提高了1.4 ％。一个11人的合奏团将F_1 推向87.4，这是提交给排行榜时的最新技术水平。[2] ELMo增加4.7％也显着大于将CoVe增加到基线模型后增加1.8％（麦肯等人。，2017).

[2]截至2017年11月17日。

任务	以前的sota		我们的底线	埃尔莫底线	增加（绝对/相对）
队	刘等人。(2017)	84.4	81.1	85.8	4.7 / 24.9%
斯恩利	陈等人。(2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	他等人。(2017)	81.7	81.4	84.6	3.2 / 17.2%
科里夫	李等人。(2017)	67.2	67.2	70.4	3.2 / 9.8%
纳	彼得斯等人。(2017)	91.93±0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	麦肯等人。(2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

表1：ELMo增强神经模型与六个基准NLP任务中最先进的单一模型基线的测试集比较。性能指标因任务而异 - SNLI和SST-5的准确度;F_1 用于SQuAD，SRL和NER;Coref的平均F_1 。由于NER和SST-5的测试尺寸较小，我们使用不同的随机种子报告了五次运行的平均值和标准差。“增加”列列出了基线的绝对和相对改进。

文本蕴涵文本蕴涵是在给定“前提”的情况下确定“假设”是否为真的任务。斯坦福自然语言推理（SNLI）语料库（鲍曼 等。，2015）提供大约550K假设/前提对。我们的基线，来自的ESIM序列模型陈等人。(2017)，使用biL-STM对前提和假设进行编码，然后是矩阵关注层，局部推理层，另一个biLSTM推理组合层，最后是输出层之前的池化操作。总的来说，将ELMo添加到ESIM模型可以使五个随机种子的准确度平均提高0.7%。五人合奏将整体精确度提升至89.3%，超过之前合奏的88.9%（龚等人。，2018).

语义角色标记语义角色标记（SRL）系统模拟句子的谓词 - 参数结构，通常被描述为回答“谁对谁做了什么”。他等人。 (2017）模拟SRL作为BIO标记问题，并使用8层深双向，前向和后向交错，周和徐(2015).如表所示1，当将ELMo添加到重新实现时他等人。 (2017单模型测试集F_1 从81.4%跃升至84.6% - 这是OntoNotes 基准测试的最新技术水平（Pradhan等人。，2013），甚至比之前的最佳合奏结果提高了1.2%。

核心参考解决方案核心参考解决方案是在文本中聚类提及相同底层现实世界实体的任务。我们的基线模型是基于端到端跨度的神经模型李等人。(2017).它使用biLSTM和注意机制首先计算跨度表示，然后应用softmax提到排名模型来查找共同参与链。在我们使用CoNLL 2012共享任务中的OntoNotes共同注释进行的实验中（Pradhan等人。，2012），加入ELMo将平均F_1 从67.2提高到70.4，提高了3.2%，建立了一种新的技术水平，再次比之前的最佳整体结果提高了1.6% F_1。

命名实体提取CoNLL 2003 NER任务（桑和梅尔德，2003）由来自路透社RCV1语料库的新闻专线组成，其标记有四种不同的实体类型（PER，LOC，ORG，MISC）。遵循最新的先进系统（Lample等。，2016; 彼得斯等人。，2017），基线模型使用预训练的字嵌入，基于字符的CNN表示，两个biLSTM层和一个条件随机场（CRF）丢失（Lafferty等。，2001），相近Collobert等。 (2011).如表所示1，我们的ELMo增强型biLSTM-CRF在五次运行中平均达到92.22%F_1 。我们的系统与以前的技术水平之间的关键区别彼得斯等人。 (2017）是我们允许任务模型学习所有biLM层的加权平均值，而Pe- ters等。(2017）仅使用顶部biLM层。如第二节所示。5.1使用所有图层而不是最后一层可以提高多个任务的性能。

情绪分析斯坦福情绪树库中的细粒度情绪分类任务（SST-5;Socher等。，2013）涉及选择五个标签中的一个（从非常消极到非常正面）来描述电影评论中的句子。句子包含各种语言现象，如习语和复杂的语法

任务	底线	最后一个	所有图层	
			λ=1	λ=0.001
队	80.8	84.7	85.0	**85.2**
斯恩利	88.1	89.1	89.3	**89.5**
SRL	81.6	84.1	84.6	**84.8**

表2：SQUAD，SNLI和SRL的开发集性能比较使用biLM的所有层（具有不同的正则化强度λ的选择）到顶层。

任务	输入只要	输入&产量	产量只要
小队	85.1	**85.6**	84.8
SRL	88.9	**89.5**	88.7
	84.7	84.3	80.9

表3：在监督模型中的不同位置包含ELMo时，SQUAD，SNLI和SRL的开发集性能。

诸如否定之类的抽象结构，模型难以学习。我们的基线模型是来自的自我分类网络（BCN）麦肯等人。(2017)，当使用CoVe嵌入增强时，它也保持了先前的最新结果。在BCN模型中用ELMo代替CoVe导致比现有技术提高1.0%的绝对精度。

5 分析

本节提供消融分析，以验证我们的主要声明，并阐明ELMo表示的一些有趣方面。秒。5.1 表明在下游任务中使用深层上下文表示可以提高与仅使用顶层的先前工作相比的性能，无论它们是从biLM还是MT编码器生成，并且ELMo表示提供最佳的整体性能。秒。5.3 探索了在biLM中捕获的不同类型的上下文信息，并使用两个内在评估来表明语法信息在较低层更好地表示，而语义信息被捕获到更高层，与MT编码器一致。它还表明我们的biLM始终提供比CoVe更丰富的表示。此外，我们分析了ELMo包含在任务模型中的敏感度（Sec。5.2），训练集大小（Sec。5.4），并在任务中可视化ELMo学习的权重（Sec。5.5).

5.1 替代层加权方案

等式有很多替代方案1 用于组合biLM层。以前关于上下文表示的工作仅使用最后一层，无论是来自biLM（彼得斯等人。，2017）或MT编码器（CoVe;麦肯等人。，2017).正则化参数λ的选择也很重要，因为诸如λ = 1的大值有效地将加权函数减少到层上的简单平均值，而较小的值（例如，λ = 0.001）允许层权重变化。

表2 比较了SQUAD，SNLI和SRL的这些替代方案。包括来自所有层的表示提高了整体性能而不仅仅使用最后一层，并且包括来自最后一层的上下文表示提高了基线上的性能。例如，在SQuAD的情况下，仅使用最后一个biLM层可将开发F_1 比基线提高3.9%。平均所有biLM层而不是仅使用最后一层提高F_1 另外0.3%（比较“最后仅”到λ = 1列），并允许任务模型学习单个层权重提高F_1 另一个0.2%（λ = 1对λ = 0.001）。在大多数情况下使用ELMo优选小λ，但是对于NER，具有较小训练集的任务，结果对λ不敏感（未示出）。

整体趋势与CoVe类似，但与基线相比增幅较小。对于SNLI，平均所有具有λ = 1的层，与仅使用最后一层相比，将开发精度从88.2提高到88.7%。SRL F_1 增加了0.1%的边际对于λ= 1的情况，82.2与仅使用最后一层相比。

5.2 哪里包括ELMo？

本文中的所有任务体系结构都只包含字嵌入作为最低层biRNN的输入。但是，我们发现在任务特定体系结构中将biMNN输出包含在ELMo中可以改善某些任务的整体结果。如表所示3，包括SNLI和SQUAD输入和输出层的ELMo仅改善输入层，但对于SRL（和共同参考分辨率，未显示），当它仅包含在输入层时，性能最高。对此结果的一种可能解释是SNLI和SQuAD体系结构都在biRNN之后使用关注层，因此在该层引入ELMo允许模型直接参与biLM的内部表示。在SRL案例中，

	资源	最近的邻居
手套	玩	玩，游戏，游戏，播放，球员，戏剧，播放器，播放，足球，多人游戏
贝尔	奇科鲁伊斯在阿鲁西克的比赛中表现出色地滚球	Kieffer是该组中唯一的一名大三学生，因其能够击中传球以及全面出色的表现而受到表彰。
	{Olivia De Havilland签约为Garson做一个百老汇戏剧。..}	{...他们是那些被赋予了肥胖角色的演员一个成功的游戏，并有足够的才能，以充分的轻描淡写地填补角色。

表4：使用GloVe“播放”的最近邻居和来自biLM的上下文嵌入。

模型	F_1
WordNet第一感测基线	65.9
Raganato等。(2017a)	69.9
Iacobacci等。(2016)	**70.1**
CoVe，第一层	59.4
CoVe，第二层	*64.7*
biLM，第一层	67.4
biLM，第二层	*69.0*

表5：所有单词细粒WSD F_1。对于CoVe和biLM，我们报告第一层和第二层biLSTM的分数。

模型	加。
Collobert等。(2011)	97.3
马和霍维(2016)	97.6
凌等人。(2015)	**97.8**
CoVe，第一层	*93.3*
CoVe，第二层	92.8
biLM，第一层	*97.3*
biLM，第二层	96.8

表6：PTB的测试集POS标记准确度。对于CoVe和biLM，我们报告第一层和第二层biLSTM的分数。

任务特定的上下文表示可能比来自biLM的表示更重要。

5.3 biLM的陈述捕获了哪些信息？

由于添加ELMo仅改善了单词向量上的任务性能，因此biLM的上下文表示必须编码通常对于未在单词向量中捕获的NLP任务有用的信息。直观地说，biLM必须使用它们的上下文来消除单词的含义。考虑“游戏”，一个高度多义的词。表顶部4 使用GloVe向量列出“play”的最近邻居。它们分布在几个词性（例如，“播放”，“播放”作为动词，“玩家”，“游戏”作为名词），但集中在与“游戏”有关的体育相关的感官中。相比之下，底部的两行显示来自SemCor数据集的最近邻句（见下文），使用源语句中的“play”的biLM上下文表示。在这些情况下，biLM能够消除源句中的词性和词义。

可以使用a来量化这些观察结果类似于上下文表征的内在评价Belinkov等。(2017)。为了隔离由biLM编码的信息，表示用于直接对细粒度词义消歧（WSD）任务和POS标记任务进行预测。使用这种方法，还可以与CoVe和每个单独的层进行比较。

词义消歧给定一个句子，我们可以使用biLM表示来使用简单的1-最近邻法来预测目标词的意义，类似于米拉姆 等。 (2016)。为此，我们首先使用biLM来计算Sem-Cor 3.0中所有单词的表示，我们的训练语料库（米勒等人。，1994），然后取每个意义的平均表示。在测试时，我们再次使用biLM来计算给定目标词的表示，并从训练集中获取最近邻感，从而恢复到WordNet中针对训练期间未观察到的引理的第一感觉。

表5 使用来自的评估框架比较WSD结果Raganato等。 (2017b）在同一套四个测试集中拉加纳托 等。(2017a)。总的来说，biLM顶层代表

怨言的F_1为69.0，并且在WSD然后第一层更好。这与使用手工制作功能的最先进的WSD专用监督模型相比具有竞争力（Iacobacci等。，2016）和任务特定的biLSTM，也使用辅助粗粒度语义标签和POS标签进行训练（Raganato等。，2017a）.CoVe biLSTM层遵循与biLM相似的模式（第二层与第一层相比具有更高的整体性能）；然而，我们的biLM优于CoVe biLSTM，后者追踪WordNet的第一感觉基线。

POS标记为了检查biLM是否捕获基本语法，我们使用上下文表示作为线性分类器的输入，该分类器使用Penn Treebank（PTB）的华尔街日报部分预测POS标签（马库斯等人。，1993）.由于线性分类器仅添加少量模型容量，因此这是对biLM表示的直接测试。与WSD类似，biLM表示与精心调整的，任务特定的biLSTM竞争（凌等人。，2015; 马和 霍维，2016）.然而，与WSD不同，使用第一个biLM层的准确度高于顶层，与多任务训练中深度biL-STM的结果一致（Søgaard和Gold- 伯格，2016; Hashimoto等人。，2017）和MT（是- linkov等。，2017）.CoVe POS标记精度遵循与biLM相同的模式，就像WSD一样，biLM实现了比CoVe编码器更高的精度。

对监督任务的影响总之，这些实验证实biLM中的不同层代表不同类型的信息，并解释为什么包括所有biLM层对于下游任务中的最高性能是重要的。此外，biLM的表示比CoVe中的表示更易于转移到WSD和POS标记，这有助于说明为什么ELMo在下游任务中优于CoVe。

5.4 样品效率

将ELMo添加到模型中可以显着提高样本效率，无论是在参数更新次数方面，还是达到最先进的性能和整体训练集大小。例如，在没有ELMo的486个训练时期之后，SRL模型达到最大发展F_1。添加ELMo后，模型在第10纪元超过了基线最大值，达到所需更新数量的相对减少了98％

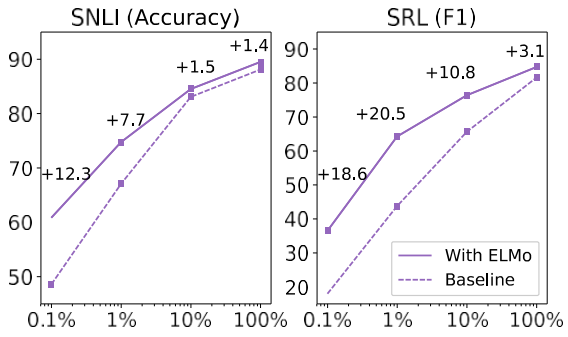

图1：SNLI和SRL的基线与ELMo性能的比较，因为训练集大小在0.1％到100％之间变化。

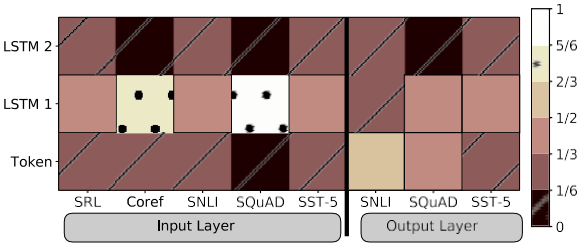

图2：跨任务和ELMo位置的softmax标准化biLM层权重的可视化。用水平线画出小于1/3的标准化重量，然后用大于2/3的标准重量标记。

相同的表现水平。

此外，ELMo增强型号比没有ELMo的型号更有效地使用较小的训练集。图1比较了有和没有ELMo的基线模型的性能，因为完整训练集的百分比在0.1％到100％之间变化。ELMo的改进对于较小的训练集来说是最大的，并且显着减少了达到给定性能水平所需的训练数据量。在SRL情况下，具有1％训练集的ELMo模型具有与具有10％训练集的基线模型大致相同的F_1。

5.5 学习权重的可视化

图2显示了softmax标准化的学习层权重。在输入层，任务模型支持第一个biLSTM层。对于共享和SQUAD，这是非常受欢迎的，但其他任务的分布不那么高。输出层权重相对平衡，略低于较低层。

6 结论

我们已经介绍了一种从biLM学习高质量深度上下文相关表示的一般方法，并且在将ELMo应用于广泛的NLP任务时显示出很大的改进。通过消融和其他对照实验，我们还确认了biLM层有效地编码关于单词上下文的不同类型的句法和语义信息，并且使用所有层提高了整体任务性能。

参考

Jimmy Ba, Ryan Kiros和Geoffrey E. Hinton。2016年
图层规范化。CoRR abs / 1607.06450。

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad和James R. Glass。2017.神经机器翻译模型对形态学有什么了解？在ACL中。

Piotr Bojanowski , Edouard Grave , Armand Joulin和Tomas Mikolov。2017.使用子字信息丰富单词向量。TACL 5：135-146。

Samuel R. Bowman, Gabor Angeli, Christopher Potts和Christopher D. Manning。2015.用于学习自然语言推理的大型注释语料库。在2015年自然语言处理经验方法会议（EMNLP）的会议记录中。计算语言学协会。

Ciprian Chelba , Tomas Mikolov , Mike Schuster , Qi Ge , Thorsten Brants , Phillipp Koehn和Tony Robinson。2014.用于衡量统计语言建模进展的十亿字基准。在INTERSPEECH。

钱倩，朱晓丹，凌振华，司伟，慧江，戴安娜·卡斯彭。2017.增强自然语言推理的lstm。在ACL中。

Jason Chiu和Eric Nichols。2016.具有双向LSTM-CNN的命名实体识别。在TACL。

Kyunghyun Cho , Bart van Merrienboer , Dzmitry Bah-danau和Yoshua Bengio。关于神经机器翻译的特性：编码器 - 解码器方法。在SSST @ EMNLP。

克里斯托弗克拉克和马修加德纳。2017.简单有效的多段阅读理解。CoRR abs / 1710.10723。

凯文克拉克和克里斯托弗D.曼宁。2016年。针对提及排名共指模型的深度强化学习。在EMNLP中。

Ronan Collobert, Jason Weston, Le'on Bottou, Michael Karlen, Koray Kavukcuoglu和Pavel P. Kuksa。2011.自然语言处理（几乎）从头开始。在JMLR中。

Andrew M. Dai和Quoc V. Le。2015.半监督序列学习。在NIPS。

Greg Durrett和Dan Klein。2013年。在共识解析中轻松胜利和艰难战斗。在EMNLP中。

Yarin Gal和Zoubin Ghahramani。2016.在递归神经网络中丢失的理论基础应用。在NIPS。

Yichen Gong, Heng Luo和Jian Zhang。2018.交互空间的自然语言推理。在ICLR。

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka和Richard Socher。2017.一个联合的多任务模型：为多个nlp任务发展神经网络。在EMNLP 2017中。

Luheng He, Kenton Lee, Mike Lewis和Luke S. Zettlemoyer。2017.深层语义角色标签：什么有效，什么是下一步。在ACL中。

Sepp Hochreiter和Ju¨rgenSchmidhuber。1997.长期短暂记忆。神经计算9。

Ignacio Iacobacci, Mohammad Taher Pilehvar和Roberto Navigli。2016.用于词义消歧的嵌入：评估研究。在ACL中。

Rafal Jo'zefowicz , Oriol Vinyals , Mike Schuster, Noam Shazeer和Yonghui Wu。2016.探索语言建模的极限。CoRR abs / 1602.02410。

Rafal Jo'zefowicz , Wojciech Zaremba 和 Ilya Sutskever。2015.对经常性网络架构的实证探索。在ICML中。

Yoon Kim , Yacine Jernite , David Sontag 和 Alexander M Rush。2015.字符感知神经语言模型。在2016年AAAI。

Diederik P. Kingma和Jimmy Ba。亚当：随机优化的一种方法。在ICLR。

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer , Ishaan Gulrajani James Bradbury, Victor Zhong, Romain Paulus和Richard Socher。2016.问我任何事情：用于自然语言处理的动态内存网络。在ICML中。

John D. Lafferty, Andrew McCallum和Fernando Pereira。条件随机场：用于分割和标记序列数据的概率模型。在ICML中。

Guillaume Lample , Miguel Ballesteros , Sandeep Subramanian , Kazuya Kawakami 和 Chris Dyer。2016.命名实体识别的神经架构。在NAACL-HLT中。

Kenton Lee, Luheng He, Mike Lewis和Luke S. Zettlemoyer。2017.端到端神经共指消解。在EMNLP中。

Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Lu'ısMarujo和TiagoLu'ıs。2015.在表单中查找功能：用于开放词汇表单词表示的组合字符模型。在EMNLP中。

刘晓东，沉从龙，凯文杜，高剑锋。2017.机器阅读理解的随机答案网络。arXiv preprint arXiv：1712.03556。

马学哲和Eduard H. Hovy。2016.通过双向LSTM-CNNs-CRF进行端到端序列标记。在ACL中。

Mitchell P. Marcus, Beatrice Santorini和Mary Ann Marcinkiewicz。建立一个大的注释英语语料库：宾州树库。计算语言学19：313-330。

Bryan McCann, James Bradbury, Caiming Xiong和Richard Socher。2017.Learnned in translation：Contextualized word vectors。在NIPS 2017中。

Oren Melamud, Jacob Goldberger和Ido Dagan。2016. context2vec：学习使用双向lstm的通用上下文嵌入。在CoNLL。

Ga'bor Melis, Chris Dyer和Phil Blunsom。2017.关于神经语言模型评估的最新进展。CoRR abs / 1707.05589。

Stephen Merity, Nitish Shirish Keskar和Richard Socher。2017.规范和优化lstm语言模型。CoRR abs / 1708.02182。

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado和Jeff Dean。2013.单词和短语的分布式表示及其组合性。在NIPS。

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock和Robert G. Thomas。1994.使用语义一致性进行感觉识别。在HLT。

Tsendsuren Munkhdalai和Hong Yu。2017.用于文本理解的神经树索引器。在EACL。

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos和Andrew McCallum。2014.向量空间中每个单词的多个嵌入的高效非参数估计。在EMNLP中。

Martha Palmer, Paul Kingsbury和Daniel Gildea。命题库：一个带注释的语义角色语料库。计算语言学31：71-106。

Jeffrey Pennington, Richard Socher和Christopher D. Manning。手套：单词表示的全球载体。在EMNLP中。

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula和Russell Power。2017.使用双向语言模型的半监督序列标记。在ACL中。

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, AndersBjo¨rkelund, Olga Uryupina, Yuchen Zhang和Zhi Zhong。2013.使用ontonotes进行强大的语言分析。在CoNLL。

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina和Yuchen Zhang。2012. Conll-2012共享任务：在ontonotes中建模多语言无限制共指。在EMNLP-CoNLL共享任务中。

Alessandro Raganato, Claudio Delli Bovi和Roberto Navigli。2017A。词义消歧的神经序列学习模型。在EMNLP中。

Alessandro Raganato, Jose Camacho-Collados和Roberto Navigli。2017b。词义消歧：统一的评估框架和实证比较。在EACL。

Pranav Rajpurkar, 张健, Konstantin Lopyrev和Percy Liang。2016.Squad：100,000多个机器理解文本的问题。在EMNLP中。

Prajit Ramachandran, Peter Liu和Quoc Le。2017.改进序列以使用未标记的数据进行序列学习。在EMNLP中。

Erik F. Tjong Kim Sang和Fien De Meulder。2003. CoNLL-2003共享任务简介：与语言无关的命名实体识别。在CoNLL。

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi和Hannaneh Hajishirzi。2017.双向关注流程，用于机器理解。在ICLR。

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng和Christopher Potts。2013.针对情感树库的语义组合的递归深度模型。在EMNLP中。

AndersSøgaard和Yoav Goldberg。2016.深层多任务学习，在较低层次监督低级别任务。在ACL 2016中。

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever和Ruslan Salakhutdi-nov。辍学：一种防止神经网络过度拟合的简单方法。机器学习研究期刊15：1929-1958。

Rupesh Kumar Srivastava, Klaus Greff和Ju¨rgenSchmidhuber。2015.培训非常深入的网络。在NIPS。

Joseph P. Turian, Lev-Arie Ratinov和Yoshua Ben-gio。Word表示：半监督学习的简单通用方法。在ACL中。

王文辉，杨楠，傅福伟，鲍宝昌，周明。2017.门控自我匹配网络，用于阅读理解和问答。在ACL中。

John Wieting，Mohit Bansal，Kevin Gimpel和Karen Livescu。2016.Charagram：通过字符n-gram嵌入单词和句子。在EMNLP中。

Sam Wiseman，Alexander M. Rush和Stuart M. Shieber。2016.学习共识解析的全局功能。在HLT-NAACL中。

马修D.泽勒。Adadelta：一种自适应学习率方法。CoRR abs / 1212.5701。

周杰和魏旭。2015年。使用递归神经网络进行语义角色标记的端到端学习。在ACL中。

周鹏，齐振宇，郑顺聪，徐家明，鲍红云，徐波。2016.通过将双向lstm与二维最大池相结合，改进了文本分类。在COLING。

A 伴随深层语境化词表示的补充材料

本补充包含模型架构，训练程序和超参数选择的详细信息，用于剖面中最先进的模型4.

所有单独的模型在最低层共享一个共同的体系结构，在几层堆叠的RNN下面有一个与上下文无关的令牌表示 - 除了使用GRU的SQuAD模型之外的每种情况都有LSTM。

A.1 微调biLM

如第二节所述。3.4，对任务特定数据的biLM进行微调通常会导致困难的显着下降。为了对给定任务进行微调，监督标签暂时被忽略，biLM在训练分组中针对一个时期进行了微调，并在开发分割上进行了评估。经过微调后，biLM权重在任务训练期间得到修复。

表7 列出所考虑任务的开发集困惑。在除CoNLL 2012之外的所有情况下，微调都会导致困难的大幅改善，例如SNLI从72.1升至16.8。

微调对监督性能的影响取决于任务。在SNLI的情况下，微调biLM使我们的单一最佳模型的开发精度从88.9%提高到89.5%。然而，对于情绪分类，无论是否使用微调的biLM，开发设定精度大致相同。

A.2 γ在方程中的重要性。 (1)

方程中的γ参数。 (1由于biLM内部表示与任务特定表示之间的不同分布，因此对于辅助优化具有实际重要性。在Sec的最后一个案例中，这一点尤为重要。5.1.如果没有此参数，则SNLI的最后一个案例表现不佳（远低于基线），并且对于SRL，培训完全失败。

A.3 文本蕴涵

我们的基线SNLI模型是来自的ESIM序列模型陈等人。 (2017).在最初的实施之后，我们对所有LSTM和前馈层使用了300个维度，并且在训练期间修复了预训练的300维GloVe嵌入。为了正规化，我们

数据集	之前调音	后调音
斯恩利	72.1	16.8
第2012卷（COREF/SRL）	92.3	-
2003号（NER）	103.2	46.3
小平方	99.1	43.5
队	158.2	52.0
海湾	131.5	78.6

表7：针对各种数据集的训练集上的一个时期的微调之前和之后的发展集困惑（越低越好）。报告的是前向和后向困惑的平均值。

增加了50％的变异辍学率（加尔和加赫拉 - 摩尼，2016）到每个LSTM层的输入和50％的丢失（Srivastava等。，2014）在最后两个完全连接的层的输入处。所有前馈层都使用ReLU激活。使用Adam优化参数（金马和巴，2015）梯度范数剪裁为5.0，初始学习率为0.0004，每次开发集的准确度在随后的时期没有增加，减少一半。批量大小为32。

最好的ELMo配置将ELMo向量添加到最低层LSTM的输入和输出，使用（1）层标准化，λ= 0.001。由于ELMo模型中参数数量的增加，我们在注意层之后向所有递归和前馈权重矩阵以及50％的丢失添加了具有正则化系数0.0001的$£^2$ 正则化。

表8 将我们系统的测试集准确性与以前发布的系统进总体而言，将ELMo添加到ESIM模型后，准确度提高了0.7％，建立了88.7％的新单一模型，五人组合将整体精度提高到89.3％。

A.4 问题回答

我们的QA模型是该模型的简化版本克拉克和加德纳 (2017).它通过连接每个令牌的区分大小写来嵌入令牌
300维GloVe字矢量（彭宁 吨等。，2014）使用卷积神经网络生成的字符派生嵌入，然后对学习字符嵌入进行最大池化。令牌嵌入通过共享的双向GRU，然后是BiDAF的双向注意机制（Seo等人。，2017).增强的

模型	加。
基于特征（鲍曼等人。，2015)	78.2
迪因龚等人。，2018)	88.0
BCN+CHAR+COVE麦肯等人。，2017)	88.1
埃辛陈等人。，2017)	88.0
ESIM + TelelSTM陈等人。，2017)	88.6
ESIM_ELMo	**88.7** ± 0.17
DIIN合奏（龚等人。，2018)	88.9
ESIM + ELMo合奏	**89.3**

表8：SNLI测试集精度。[3]单个模型结果占据该部分，整体结果位于底部。

然后，文本向量通过具有ReLU激活的线性层，使用GRU的残余自我关注层，随后是应用上下文的相同注意机制，以及具有ReLU激活的另一线性层。最后，结果通过线性层提供，以预测答案的开始和结束标记。

在输入GRU和线性层之前以0.2的速率使用变化压差。GRU使用90维，线性层使用180维。我们使用批量大小为45的Adadelta优化模型。在测试时，我们使用权重的指数移动平均值，并将输出范围限制为最多17个。我们不会在训练期间更新单词向量。

将没有层规范化的ELMo添加到上下文GRU层的输入和输出并且使ELMo权重不规则化（$\lambda = 0$）时，性能最高。

表9 从2017年11月17日我们提交系统时，比较了SQuAD排行榜的测试集结果。总体而言，我们提交的单一模型和集合结果最高，将之前的单一模型结果（SAN）提高了1.4%F_1，基线提高了4.2%。一个11人的合奏推动F_1达到87.4%，比之前的合奏增加了1.0%。

A.5 语义角色标签

我们的基线SRL模型是（（）的精确重新实现他等人。，2017).使用100维向量表示的串联来表示单词，使用GloVe初始化（彭宁 吨等。，2014）和使用100维em-表示的二进制，每单词谓词特征寝具。然后，该200维标记表示通过具有300维隐藏尺寸的8层“交错”biLSTM，其中LSTM层的方向每层交替。这个深LSTM使用公路连接（Srivastava等。，2015）层与变分复发辍学（加尔和 加哈拉尼，2016).然后使用最终密集层然后通过softmax激活来投射该深度表示以在所有可能的标签上形成分布。标签由PropBank的语义角色组成（帕尔默等人。，2005）增加了BIO标签方案来表示参数跨度。在训练期间，我们使用Adadelta最小化标签序列的负对数可能性，学习率为1.0且$\rho = 0.95$（蔡勒，2012).在测试时，我们执行Viterbi解码以使用BIO约束强制执行有效跨度。所有LSTM隐藏层都添加了10%的变差丢失。如果它们的值超过1.0，则会剪裁渐变。模型训练500个时期或直到验证F1没有改善200个时期，以较早者为准。预训练的GloVe向量在训练期间进行了微调。最终的致密层和所有LSTM的所有单元被初始化为正交。对于所有LSTM，遗忘门偏置初始化为1，所有其他门初始化为0，如（Jo'zefowicz等人。，2015).

表10 比较我们的ELMo增强实现的测试集F1分数（他等人。，2017）以前的结果。我们的单一模型得分84.6 F1代表了CONLL 2012语义角色标签任务的最新结果，超过之前的单一模型结果2.9 F1和5模型合奏1.2 F1。

A.6 共同决议

我们的基线共参数模型是来自的端到端神经模型李等人。(2017）所有hy-

[3]可以在以下找到综合比较https://NLP.St.Fr.EdU/Studis/SNLI/

模型	相对长度单位	F_1
BiDAFSeo等人。，2017)	68.0	77.3
BiDAF +自我注意	72.1	81.1
dcn_	75.1	83.1
雷格拉索	75.8	83.3
融合网	76.0	83.9
R-网王等人。，2017)	76.5	84.3
桑(刘等人。，2017)	76.8	84.4
BIDAF +自我关注+ ELMO	**78.6**	**85.8**
DCn+系综	78.9	86.0
融合网综合体	79.0	86.0
互动AoA Reader + Ensemble	79.1	86.5
BIDAF +自我关注+ ELMO集成	**81.0**	**87.4**

表9：SQuAD的测试集结果，显示完全匹配（EM）和F_1。表格的上半部分包含单个模型结果，底部有合奏。提供参考资料。

模型	F_1
Pradhan等人。(2013)	77.5
周和徐(2015)	81.3
他等人。(2017），单身	81.7
他等人。(2017），合奏	83.4
他等人。(2017），我们的impl。	81.4
他等人。(2017+埃尔莫	**84.6**

表10：SRL CoNLL 2012测试集F_1。

模型	平均F_1
Durrett和Klein(2013)	60.3
Wiseman等人。(2016)	64.2
克拉克和曼宁(2016)	65.7
李等人。(2017）（单）	67.2
李等人。(2017）（合奏）	68.8
李等人。(2017+埃尔莫	**70.4**

表11：来自CoNLL 2012共享任务的测试集上的共参考分辨率平均值F_1。

完全遵循原始实现的参数。

最佳配置将ELMo添加到最低层biLSTM的输入并使用（对biLM层进行加权）（1没有任何正则化（λ= 0）或层归一化。ELMo表示中添加了50%的辍学率。

表11 将我们的结果与之前公布的结果进总的来说，我们通过3.2%的平均F_1改进了单一模型的最新技术，我们的单一模型结果将之前的整体效果提高了1.6%F_1。除了biLSTM输入之外，将ELMo添加到biLSTM的输出还将F_1减少大约0.7%（未示出）。

A.7 命名实体识别

我们的基线NER模型连接了50维预训练的塞纳矢量（Collobert等。，2011）具有基于CNN字符的表示。字符表示使用16维字符嵌入和128个宽度为3个字符的卷积滤波器，ReLU激活和最大池化。令牌表示通过两个biLSTM层，第一个具有200个隐藏单元，第二个具有100个隐藏单元，最终密集层和softmax层之前。在训练期间，我们使用CRF丢失并在测试时使用Viterbi算法执行解码，同时确保输出标签序列有效。

变差输出被添加到两个biLSTM层的输入中。在训练期间，如果他们的£^{2} 范数超过5.0并且使用Adam以恒定学习率0.001更新参数，则重新调整梯度。经过预先训练的塞纳嵌入物在训练期间进行了微调。我们在开发集上使用早期停止，并使用不同的随机种子报告五次运行的平均测试集分数。

ELMo已添加到最低层任务biLSTM的输入中。由于CoNLL 2003 NER数据集相对较小，我们通过将 $\lambda = 0.1$设置为约束可训练层权重来实现最佳性能（1).

表12 将我们的ELMo增强型biLSTM-CRF标记器的测试集F_1 得分与之前的结果进行比较。总体而言，我们系统的92.22%F_1 建立了一种新的先进技术。与之相比彼得斯等人。(2017），使用表示法

模型	F_1 ± std
Collobert等。 (2011)♣	89.59
Lample等。 (2016)	90.94
马和霍维 (2016)	91.2
邱和尼科尔斯(2016)♣,♦	91.62±0.33
彼得斯等人	91.93±0.19
	92.22±0.10

表12：CoNLL 2003 NER任务的测试集F_1。带♣的型号包括地名索引和使用♦的地名录
火车和发展部门都进行了培训。

模型	加。
DMN库马尔等人。，2016)	52.1
LSTM美国有线电视新闻网周等人。，2016)	52.4
NTI（NTI）Munkhdalai和Yu，2017)	53.1
BCN+CHAR+COVE麦肯等人。，2017)	53.7
BCN_ELMo	**54.7**

表13：SST-5的测试集精度。

来自biLM的所有层提供了适度的改进。

A.8 情绪分类

我们使用几乎相同的双重分类网络架构麦肯等人。（2017），除了用更简单的前馈网络替换最终的最大网络，该网络由两个带有丢失的ReLu层组成。具有批量标准化最大化网络的BCN模型在我们的实验中达到了显着降低的验证准确度，尽管我们的实施与实施之间可能存在差异。麦肯等人。(2017).为了匹配CoVe训练设置，我们只训练包含四个或更多令牌的短语。我们为biLSTM使用300-d隐藏状态，并使用Adam优化模型参数（金马和巴，2015）使用0.0001的学习率。可训练的biLM层权重由$\lambda = 0.001$正则化，并且我们将ELMo添加到biLSTM的输入和输出;输出ELMo向量用第二个biLSTM计算并连接到输入。