命名实体识别的神经架构

Guillaume Lample*Miguel Ballesteros**
Sandeep Subramanian*Kazuya Kawakami*Chris Dyer*
*卡内基梅隆大学*NLP 集团，庞贝法布拉大学
{glample^sandeeps^kkawakam^cdyerjgcs.cmu.edu，
miguel.ballesteros@upf.edu

摘要

最先进的命名实体识别系统在很大程度上依赖手工制作的特征和特定领域的知识，以便从现有的小型、受监督的培训语料库中有效地学习。在本文中，我们介绍了两种新的神经体系结构-一种基于双向 LSTM 和条件随机字段的神经体系结构，另一种是使用基于转换的方法来构造和标记段，这种方法是由移位减少解析器启发的。我们的模型依赖于关于单词的两个信息来源：从监督语料库中学习的基于字符的单词表示和从无注释语料库中学习的无监督单词表示。我们的模型在四种语言的 NER 中获得了最先进的性能，而不使用任何特定语言的知识或资源，如地名录。[1]

1 导言

命名实体识别(NER)是一个具有挑战性的学习问题。一方面，在大多数语言和领域，只有很少的监督培训数据可用。另一方面，对可以是名称的词的种类限制很少，因此从这个小样本的数据中推广是困难的。因此，精心构建的正字形特征和语言特定的知识资源，如地名录，被广泛用于解决这一任务。不幸的是，特定于语言的资源和特性在新语言和新领域的开发成本很高，使 NER 成为适应的挑战。无监督学习从无注释语料库提供了一种替代策略，以获得更好的推广从少量的监督。然而，即使是广泛依赖无监督特征的系统(Collobert 等人，2011 年；Turian 等人，2010 年；Lin 和 Wu，2009 年；Ando 和 Zhang，2005 年 b，除其他外)也利用这些功能来增加而不是取代手工设计的功能（例如，关于某一特定语言的资本化模式和字符类的知识)和专门知识资源(例如，地名录）。

在本文中，我们提出了 NER 的神经架构，它除了少量的监督训练数据和未标记的语料库之外，不使用特定于语言的资源或特征。我们的模型旨在捕捉两种直觉。首先，由于名称通常由多个令牌组成，因此对每个令牌的标记决策进行联合推理是很重要的。我们比较两个模型，

(i) 具有顺序条件随机层的双向 LSTM(LSTM-CRF；§2)和

(ii) 一种新的模型，它使用基于转换的解析的算法来构造和标记输入句子的块，其状态由堆栈 LSTM 表示(S-LSTM；§3)。第二，“作为一个名字”的令牌级别证据包括两个正字法证据（被标记为名字的单词看起来像什么？） 和分布证据（被标记的单词往往发生在语料库中？）。为了捕获正射灵敏度，我们使用基于字符的词表示模型(Ling 等人，2015b)来捕获分布灵敏度，我们将这些表示与分布表示相结合(Mikolov 等人，2013b)。我们的单词表示结合了这两者，辍学训练被用来鼓励模型学习信任两种证据来源（§4）。

英语、荷兰语、德语和西班牙语的实验表明，我们能够获得状态

1LSTM-CRF 和 Stack-LSTMNER 的实现可在 https://github.com/glample/tagger 和 https://github.com/clab/stack-lstm-ner 上获得

NAACL-HLT 2016 的会议记录，第 260-270 页，
加利福尼亚州圣地亚哥，2016 年 6 月 12 日至 17 日。

最先进的 NER 性能与 LSTM-CRF 模型在荷兰，德语和西班牙语，非常接近最先进的英语，没有任何手工设计的功能或地名录（第 5 节）。基于过渡的算法同样超过了以前在几种语言中发布的最佳结果，尽管它的性能不如 LSTM-CRF 模型。

2 LSTM-CRF 模型

我们简要描述了 LSTM 和 CRF，并提出了一种混合标记体系结构。 这种结构类似于 Collobert 等人提出的结构。（2011 年）和 Huang 等人。（2015）。

2.1 LSTM

递归神经网络(RNNs)是一种基于序列数据的神经网络。 它们以一个向量序列(xi，X2，Xn)作为输入，并返回另一个序列(hi，h2，..表示输入中每一步的序列的一些信息。虽然 RNNs 在理论上可以学习长期的依赖关系，但在实践中它们没有这样做，而且往往偏向于序列中的最近输入(Bengio 等人，1994 年)。长期短期记忆网络(LSTMs)已经被设计成通过包含一个记忆单元来解决这个问题，并且已经被证明可以捕获长期依赖关系。他们使用几个门来控制输入给内存单元的比例，以及从以前的状态到忘记的比例(Hochreiter 和 Schmidhuber，1997 年)。我们使用以下实现：每个单

它=一个(WxiXt+Whihti+W。iCt_i+b《)CT=（1-它）OCTI+

它 是 OTANH(WX)。*Xt+Whcht-i+b*。 不

=a(WxoXt+Woht-i+WcoCT+b。=tanh(Ct) 我们认为其中 a 是元素级乙状结肠函数，0 是元素级乘积。

对于包含 n 个单词的给定句子(xi，X2，...，Xn)，每个单词表示为 d 维向量，LSTM 计算左边的表示 ht

=(易，Y2，-•，YN)，

我们把它的分数定义为

nn

美国(X，Y)

=^~？ 阿依，伊+我£尸山

我=0 我=我

词 t 的句子上下文。当然，生成正确上下文 ht 的表示也应该添加有用的信息。这可以使用第二个 LSTM 来实现，该 LSTM 反向读取相同的序列。我们将前者称为前向 LSTM，后者称为后向 LSTM。 这是两个不同的网络，具有不同的参数。这种向前和向后的 LSTM 对被称为双向 LSTM(Graves 和 Schmidhuber，2005)。

使用该模型的单词的表示是通过连接其左右上下文表示 ht=[ht；ht]来获得的。 这些表示有效地包括上下文中单词的表示，这对于许多标记应用程序是有用的。

2.2 CRF 标签模型

一个非常 simple 一 but 令人惊讶的 effective 一 tagging 模型是使用 h/s 作为特征，为每个输出 yt 做出独立的标记决策(Ling 等人，2015b)。 尽管该模型在像 POS 标记这样的简单问题上取得了成功，但当输出标签之间有很强的依赖关系时，它的独立分类决策是有限的。NER 是这样的任务之一，因为标记可解释序列的“语法”施加了几个硬约束(例如，I-PER 不能遵循 B-LOC；详见§2.4)，这是不可能用独立假设建模的。

因此，我们没有独立建模标记决策，而是使用条件随机字段联合建模它们(Lafferty 等人，2001 年)。 为输入句.

P 是双向 LSTM 网络输出的分数矩阵。其中 k 是不同标记的数目，PI，j 对应于 j 的分数 th 我的标签 th 句子中的单词。 一系列的预测

其中 A 是一个转换分数的矩阵，使得 AI，j 表示从标记 i 到标记 j 的转换的分数 哟和你 *n* 是句子的开始和结束标记，我们将其添加到可能的标记集合中。因此，A 是大小为 k+2 的平方矩阵。

在所有可能的标签序列上的 Softmax 产生序列 y：的概率

$$p(y|x)=\frac{e^{\text{美国}(X,\ Y)}}{\text{前*)}}$$

在训练过程中，我们最大限度地提高了正确标签序列的对数概率：

$$\text{日志}(p(y|X)=s(X,\ y)\ \text{日志}\left|\pounds e^{\underset{\text{是的 Yx}}{\text{美国}}(X,\ Y)}\right| /$$

$$=(X,\ y)\ \underset{\text{是的}}{\text{logadds}}(X,\ y), \quad (1)$$

其中 Yx 表示句子 X 的所有可能的标记序列(甚至那些不验证 IOB 格式的序列)。从上面的公式中，很明显，我们鼓励我们的网络产生一个有效的输出标签序列。在解码时，我们预测得到：给出的最大分数的输出序列

$$=\underset{\text{是的}}{\text{argmaxs}}(X,\ y)。 \quad (2)$$

由于我们只建模输出之间的 Bigram 交互，所以在 Eq 中的求和。 方程中的最大后验序列 y*。 可以使用动态编程计算 2。

2.3 参数化和培训

与每个标记决策相关的分数(即 PI，y)被定义为在双向 LSTM 一下计算的 Wordin-上下文嵌入之间的点积，与 Ling 等人的 POS 标记模型完全相同。 (2015b)和这些都与 bigram 兼容性分数(即 Ay‘s)相结合。此架构如图 1 所示.. 圆圈代表观察到的变量，钻石是父母的确定性函数，双圆是随机变量。

因此，该模型的参数是 Bigram 兼容性分数 A 的矩阵，以及产生矩阵 P 的参数，即双向 LSTM 的参数、线性特征权重和单词嵌入。如第 2.2 部分，让 XI 表示句子中每个单词的单词嵌入序列，并作为它们的关联标记。我们回到讨论如何在第 4 节中建模嵌入 xi。单词嵌入序列作为双向 LSTM 的输入给出，它返回每个

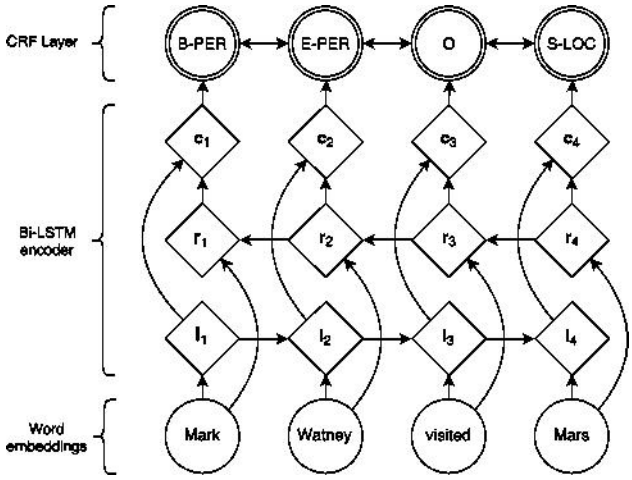

图 1：网络的主要体系结构。 字嵌入被赋予双向 LSTM。 l表示单词 i 及其左上下文，ri 表示单词 i 及其右上下文。 将这两个向量连接起来，在其上下文 Ci 中产生单词 i 的表示。

单词的左和右上下文的表示，如 2.1 所解释的。

这些表示被连接(c)，并线性地投影到一个层上，其大小等于不同标签的数量。 我们没有使用这个层的 Softmax 输出，而是使用前面描述的 CRF 来考虑相邻标签，从而产生每个单词 YI 的最终预测。此外，我们还观察到，在 c 表和 CRF 层之间添加一个隐藏层略微改善了我们的结果。用这个模型报告的所有结果都包含了这个额外的层。参数被训练以最大化方程。在一个注释的语料库中观察到的 NER 标签序列中的 1 个，给出了观察到的单词。

2.4 标签计划

命名实体识别的任务是为句子中的每个单词分配一个命名实体标签。单个命名实体可以跨越句子中的几个标记。句子通常以 IOB 格式表示（内部、外部、开始），如果令牌是命名实体的开始，则每个令牌都被标记为 B 标签，如果令牌位于命名实体中，而不是命名实体中的第一个令牌，则 Ilabel，或者 O。 然而，我们决定使用 IOBES 标记方案，IOB 的一个变体通常用于命名实体识别，它编码关于单例实体(S)的信息，并显式标记命名实体(E)的结束。使用这个方案，以高度自信的 Ilabel 标记一个单词，将后续单词的选择缩小到 Ilabel 或 Elabel，然而，IOB 方案只能确定后续单词不能是另一个标签的内部。 拉蒂诺夫和罗斯（2009 年）和戴等人。（2015）表明，使用像 IOBES 这样更具表

现力的标记方案可以略微提高模型性能。然而，我们没有观察到比 IOB 标记方案有明显的改进。

3 基于过渡的春京模型

作为上一节讨论的 LSTM-CRF 的替代方案，我们探索了一种新的体系结构，该体系结构使用类似于基于转换的依赖解析的算法对输入序列进行分块和标记。该模型直接构造多个令牌名称的表示(例如，Mark Watney 的名称被组合成一个表示)。

该模型依赖于堆栈数据结构来逐步构造输入的块。为了获得用于预测后续行动的堆栈的表示，我们使用 Dyer 等人提出的堆栈-LSTM。（2015），其中 LSTM 被一个“堆栈指针”增强。当顺序 LSTM 模型序列从左到右时，堆栈 LSTM 允许嵌入一堆对象，这些对象既被添加到（使用推送操作）中，又被从(使用 POP 操作)中删除。这允许 Stack-LSTM 像一个堆栈一样工作，它维护其内容的“摘要嵌入”。为了简单起见，我们将此模型称为 Stack-LSTM 或 S-LSTM 模型..

最后，我们请感兴趣的读者参考原始论文(Dyer 等人，2015 年)，以获得关于 StackLSTM 模型的详细信息，因为在本文中，我们只是通过下面一节中提出的一种新的基于过渡的算法使用相同的体系结构。

3.1 春京算法

我们设计了一个过渡清单，如图 2 所示，它受基于过渡的解析器的启发，特别是 Nivre（2004）的弧形标准解析器。在该算法中，我们利用两个堆栈（指定的输出和堆栈分别表示已完成的块和划痕空间）和一个缓冲区，其中包含尚未处理的单词。转换库存包含以下转换：SHIFT 转换将一个字从缓冲区移动到堆栈，OUT 转换将一个字从缓冲区直接移动到输出堆栈，而 REDUCE(Y)转换从堆栈顶部弹出所有项目，创建一个“块”，将其标记为标签 y，并将此块的表示推到输出堆栈上。当堆栈和缓冲区都为空时，算法完成。该算法如图 2 所示，它显示了处理 MarkWatney 访问火星的句子所需的操作序列。

该模型通过定义每个时间步骤上动作的概率分布来参数化，给定堆栈、缓冲区和输出的当前内容以及所采取动作的历史。跟随 Dyer 等人。（2015），我们使用堆栈 LSTM 来计算每一个的固定维数嵌入，并将它们连接起来以获得完整的算法状态。此表示用于定义在每个时间步骤中可以采取的可能行动的分布。该模型被训练以最大化参考动作序列的条件概率（从标记训练语料库中提取），给出输入句子。为了在测试时标记新的输入序列，贪婪地选择最大概率动作，直到算法达到终止状态。虽然这不能保证找到全局最优，但在实践中是有效的。由于每个令牌要么直接移动到输出（1 个动作)，要么首先移动到堆栈，然后是输出(2 个动作），因此长度 n 序列的动作总数最大为 2n。

值得注意的是，该算法的性质.

出去	史塔克特	布弗特	行动	出去	斯塔克特+	布弗特	片段
o	*s*	*(u，u)，b*	换班	o	(u，u)，S	b	
o	$(u)^{u)}$,. ■，(v，v)，SB		减少(y)	■■，v，ry)，O	s	b	(u.v，y)
o	s	(u，u)，b	出去	g(u，$_{0}$)，o	s	b	_

图 2：Stack-LSTM 模型的转换，指示应用的操作和结果状态。粗体符号表示（学习的）单词和关系的嵌入，脚本符号表示相应的单词和关系。

过渡时期	产出	垃圾	缓冲	分部
	[]	[]	[马克，沃特尼，去过火星]	
换班	[]	[马克]	[沃特尼，参观过，火星]	
换班	[]	[马克，沃特尼]	[访问过，火星]	
减少（每）	[（马克·沃特尼）-珀]	[]	[访问过，火星]	（马克·沃特尼）-珀
出去	[（马克·沃特尼）-佩尔，参观]	[]	[火星]	
SHI FT	[（马克·沃特尼）-佩尔，参观]	[火星]	[]	
减少(LOC)	[(Mark Watney)-PER，参观过，(Mars)-LOC]	[]	[]	（火星）-LOC

图 3：Mark Watney 使用 Stack-LSTM 模型访问火星的过渡序列。

模型使它与所使用的标记方案无关，因为它直接预测标记块。

3.2 代表标记春克

当执行 REDUCE(Y)操作时，该算法将标记序列（连同它们的向量嵌入）作为单个完成的块从堆栈转移到输出缓冲区。为了计算这个序列的嵌入，我们在其组成令牌的嵌入上运行一个双向 LSTM，以及一个表示被识别块类型的令牌(即 y)。此函数给出为 g(u，v，ry)，其中 ry 是标签类型的学习嵌入。因此，输出缓冲区包含生成的每个标记块的单个向量表示，而不管其长度如何。

4 输入单词嵌入

我们两个模型的输入层都是单个单词的向量表示。从有限的 NER 训练数据中学习单词类型的独立表示是一个难题：参数太多，无法可靠地估计。由于许多语言都有正字法或形态学证据表明某物是一个名字（或不是一个名字），所以我们想要对单词拼写敏感的表示。因此，我们使用一个模型，从它们由（4.1）组成的字符的表示中构造单词的表示。我们的第二个直觉是，名称，可能是相当不同的，出现在规则的上下文中的大语料库。因此，我们使用嵌入

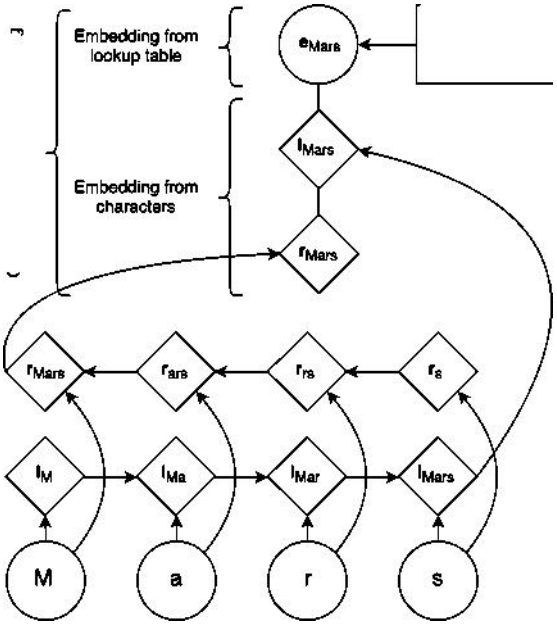

图 4：“火星”一词的字符嵌入是给双向 LSTM 的。我们将它们的最后输出连接到从查找表的嵌入，以获得此字的表示。

从一个对语序敏感的大语料库中学习的 dings（4.2）。最后，为了防止模型过于依赖于一个表示或另一个表示，我们使用辍学训练，并发现这对于良好的泛化性能至关重要（4.3）。

4.1 基于字符的单词模型

我们的工作与以前大多数方法的一个重要区别是，我们学习字符级特征，同时训练，而不是手工工程前缀和后缀信息的单词。学习字符级嵌入具有特定于任务和领域的学习表示的优点。它们被发现对于形态丰富的语言和处理词汇外问题很有用，例如词性标注和语言建模(Ling 等人，2015 年 b)或依赖解析(Ballesteros 等人，2015 年)。

图 4 描述了我们的体系结构，以从其字符中生成一个单词嵌入。随机初始化的字符查找表包含每个字符的嵌入。将单词中每个字符对应的字符嵌入直接和反向顺序给出一个向前和向后的 LSTM。从其字符派生的单词的嵌入是从双向 LSTM 中连接其前向和后向表示。然后，这个字符级表示与字查找表中的字级表示连接。在测试过程中，查找表中没有嵌入的单词被映射到 UNK 嵌入。为了训练 UNK 嵌入，我们用概率为 0.5 的 UNK 嵌入替换单例。在我们所有的实验中，向前和向后字符 LSTM 的隐藏维数各为 25，这导致了我们基于字符的单词表示为维度 50。

递归模型，如 RNNs 和 LSTM，能够编码非常长的序列，然而，它们的表示偏向于它们最近的输入。因此，我们期望前向 LSTM 的最终表示是单词后缀的精确表示，而后向 LSTM 的最终状态是其前缀的更好表示。 最值得注意的替代 approaches 一如卷积 networks 一 have 被提出从其字符中学习单词的表示(Zhang 等人，2015 年；Kim 等人，2015 年)。 然而，卷积网的设计是为了发现它们输入的位置不变特征。虽然这适用于许多问题，例如图像识别（猫可以出现在图片的任何地方)，但我们认为重要的信息是位置依赖的(例如，前缀和后缀编码的信息与茎不同），使 LSTM 成为一个先验的更好的函数类来建模单词与其字符之间的关系。

4.2 预先训练的嵌入

如 Collobert 等人案。（2011），我们使用预先训练的单词嵌入来初始化查找表。我们观察到使用预先训练的单词嵌入在随机初始化的单词上的显著改进。嵌入使用 skip-n-gram 进行预训练(Ling 等人，2015a)，这是 Word2vec 的一个变体(Mikolov 等人，2013a)，解释了语序。 这些嵌入是在训练期间微调的..

西班牙语、荷兰语、德语和英语的单词嵌入分别使用西班牙语 Gigaword 版本 3、Leipzig 语料库收集、2010 年机器翻译讲习班的德语单语培训数据和英语 Gigaword 版本 4(删除了 LA 时报和纽约时报的部分)进行培训。[2] 我们使用嵌入维度为 100 的英语，64 的其他语言，最小单词频率截止 4，窗口大小为 8。

4.3 辍学培训

最初的实验表明，字符级嵌入并没有提高我们的整体性能，当与预先训练的单词表示一起使用时。为了鼓励模型依赖于这两种表示，我们使用辍学训练(HintonEtal，2012)，在图 1 中双向 LSTM 的输入之前，将辍学掩码应用于最终嵌入层。我们观察到在使用辍学后，我们的模型的性能有了显著的改善（见表 5）。

5 实验

本节介绍了我们用来训练模型的方法、我们在各种任务上获得的结果以及我们的网络配置对模型性能的影响。

5.1 培训

对于这两个模型，我们使用反向传播算法对我们的网络进行训练，在每个训练示例上更新我们的参数，每次一个，使用随机梯度下降(SGD)，学习速率为 0.01，梯度裁剪为 5.0。已经提出了几种提高 SGD 性能的方法，如 Adadelta(Zeiler，2012 年)或 Adam(Kingma 和 Ba，2014 年)。 虽然我们使用这些方法观察到更快的收敛，但没有一种方法的性能和 SGD 的梯度裁剪。

我们的 LSTM-CRF 模型使用单层的向前和向后的 LSTM，其尺寸设置为 100。 调整这个维度对模型性能没有显著影响。 我们把辍学率定为 0.5。 使用更高的速率会对我们的结果产生负面影响，而较小的速率会导致更长的训练时间。

堆栈-LSTM 模型为每个堆栈使用两层维度 100。组合函数中使用的动作的嵌入每个有 16 个维度，输出嵌入为维度 20。我们对不同的辍学率进行了实验，并使用每种语言的最佳辍学率报告了分数。[3] 这是一个贪婪的模型，它应用局部最优操作，直到整个句子被处理，可以通过波束搜索(Zhang 和 Clark，2011 年)或探索训练(Ballesteros 等人，2016 年)获得进一步的改进。

5.2 数据集

我们在不同的数据集上测试我们的模型以进行命名实体识别。 为了证明我们的模型能够推广到不同的语

[2](Graff，2011 年；Biemann 等人，2007 年；Callison-Burch 等人，2010 年；Parker 等人，2009 年)

英语(D=0.2)、德语、西班牙语和荷兰语(D=0.3)

言，我们在 CoNLL-2002 和 CoNLL-2003 数据集(TjongKimSang，2002；TjongKimSang 和 DeMoulder，2003)上给出了包含英文、西班牙文、德文和荷兰文独立命名实体标签的结果。所有数据集都包含四种不同类型的命名实体：不属于前三类中任何一类的位置、人员、组织和杂项实体。虽然 POS 标签可用于所有数据集，但我们没有将它们包含在我们的模型中。我们没有执行任何数据集预处理，除了在英文 NER 数据集中用零替换每个数字。

5.3 结果

表 1 显示了我们与其他模型的英文命名实体识别的比较。为了使我们的模型和其他模型之间的比较公平，我们报告了其他模型的分数，无论是否使用外部标记数据，如地名录和知识库。我们的模型不使用地名录或任何外部标记的资源。这项任务报告的最佳分数是罗等人。（2015）。他们通过联合建模 NER 和实体链接任务获得了 91.2 的 F1(Hoffart 等人，2011 年)。他们的模型使用了许多手工设计的功能，包括拼写功能、Word Net 集群、Brown 集群、POS 标签、块标签以及词干和外部知识库，如 Freebase 和 Wikipedia。我们的 LSTM-CRF 模型优于所有其他系统，包括使用外部标记数据的系统，如地名录。除了 Chiu 和 Nichols（2015）提出的模型外，我们的 StackLSTM 模型也优于所有以前不包含外部特征的模型。

与其他模型相比，表 2、表 3 和表 4 分别给出了德语、荷兰语和西班牙语的 NER 结果。在这三种语言上，LSTM-CRF 模型明显优于所有以前的方法，包括使用外部标记数据的方法。唯一的例外是荷兰语，其中 Gillick 等人的模型。（2015 年）通过利用来自其他 NER 数据集的信息可以表现得更好。与不使用外部数据的系统相比，Stack-LSTM 还一致地显示了 statethe-art（或接近）结果。

正如我们在表格中所看到的，Stack-LSTM 模型更依赖于基于字符的表示来实现竞争性能；我们假设 LSTM-CRF 模型需要更少的正交信息，因为它从双向 LSTM 中获得更多的上下文信息；然而，Stack-LSTM 模型逐个消耗单词，并且它只是在对单词进行分块时依赖单词表示。

5.4 网络架构

我们的模型有几个组件，我们可以调整，以了解它们对整体性能的影响。我们探讨了 CRF、字符级表示、预训练对我们的影响

模式	Fi
Collobertetal，（2011）*	89.59
林和吴（2009）	83.78
林和吴（2009）*	90.90
Huang 等人 （2015）*	90.10
Passos 等人 （2014）	90.05
Passos 等人 （2014）*	90.90
罗等人 （2015 年）*+加沙	89.9
罗等人 （2015 年）*+公报+链接	**91.2**
邱和尼科尔斯（2015 年）	90.69
邱和尼科尔斯（2015）*	90.77
LSTM-CRF（无炭）	90.20
LSTM-CRF	**90.94**
S-LSTM(无 char)	87.96
S-LSTM	90.33

表 1：英语 NER 结果(CoNLL-2003 测试集)。*表示使用外部标记数据训练的模型

模式	Fi
Florian 等人 （2003）*	72.41
安藤和张(2005a)	75.27
Qi 等人 （2009）	75.72
Gillick 等人。（2015）	72.08
Gillick 等人。（2015）*	76.22
LSTM-CRF 没有字符	75.06
LSTM-CRF	**78.76**
S-LSTM-没有字符	65.87
S-LSTM	75.66

表 2：德国 NER 结果(CoNLL-2003 测试集)。*独立使用外部标记数据训练的 Cates 模型

模式	Fi
Carreras 等人。（2002）	77.05
Nothman 等人 （2013）	78.6
Gillick 等人。（2015）	78.08
Gillick 等人。（2015）	**82.84**
LSTM-CRF-没有字符	73.14
LSTM-CRF	**81.74**
S-LSTM-没有字符	69.90
S-LSTM	79.88

表 3：荷兰 NER(CoNLL-2002 测试集)。*表示模式使用外部标记数据训练 ELS

模式	Fi
Carreras 等人。（2002）*	81.39
桑托斯和吉马拉斯（2015 年）	82.21
Gillick 等人。（2015）	81.83
Gillick 等人。（2015）*	82.95
LSTM-CRF-没有字符	83.44
LSTM-CRF	**85.75**
S-LSTM-没有字符	79.46
S-LSTM	83.93

表 4：西班牙 NER(CoNLL-2002 测试集)。*表示模式使用外部标记数据训练 ELS

我们的 LSTMCRF 模型上有单词嵌入和辍学。我们观察到，预训练我们的单词嵌入给了我们最大的改善，整体性能的+7.31 在 Fi。通用报告格式层使我们增加了+1.79，而使用辍学导致差异+1.17，最后学习字符级单词嵌入导致增加约+0.74。对于 Stack-LSTM，我

们进行了一组类似的实验.. 表 5 给出了不同体系结构的结果。

模式	变式	Fi
LSTM	查尔+辍学+前训练	89.15
LSTM-CRF	查尔+辍学	83.63
LSTM-CRF	预处理	88.39
LSTM-CRF	预处理+焦炭	89.77
LSTM-CRF	预培训+辍学	90.20
LSTM-CRF	预训练+辍学+焦炭	**90.94**
S-LSTM	查尔+辍学	80.88
S-LSTM	预处理	86.67
S-LSTM	预处理+焦炭	89.32
S-LSTM	预培训+辍学	87.96
S-LSTM	预训练+辍学+焦炭	90.33

表 5：英语 NER 结果与我们的模型，使用不同

安特配置。“预处理”指的是包括预处理的模型
经过训练的单词嵌入，“char”指的是包括在内的模型
基于字符的单词建模，“辍学”指的是模型
包括辍学率。

6 相关工作

在 CoNLL-2002 共享任务中，Carreras 等人。（2002 年）在荷兰语和西班牙语方面取得了最好的成绩，结和 Yarowsky（1999；2002)提出了半监督引导算法，通过共同训练字符级(字内)和令牌级(上下文）特征来识别命名实体。 Eisenstein 等人（2011 年）使用贝叶斯非 parametrics 在几乎无监督的环境中构建命名实体数据库。 拉蒂诺夫和罗斯（2009）定量比较了 NER 的几种方法，并使用正则化的平均感知器和聚合上下文信息建立了自己的监督模型。

最后，目前人们对使用基于字母表示的 NER 模型很感兴趣。Gillick 等人。将序列标记任务建模为序列学习问题，并将基于字符的表示纳入其编码器模型。邱和尼科尔斯（2015）使用了一种类似于我们的体系结构，但相反，使用 CNN 来学习字符级特征，其方式类似于 Santos 和 Guimaraes（2015）的工作。

7 结论

本文提出了两种神经结构的序列标记，提供了最好的 NER 结果在标准评估设置，甚至与模型使用外部资源，如地名录。

我们的模型的一个关键方面是，它们通过一个简单的 CRF 体系结构来建模输出标签依赖关系，或者使用基于转换的算法显式地构造和标记输入的块。单词表示对于成功也是至关重要的：我们使用预先训练的单词表示和“基于字符”的表示来捕获形态和正交信息。为了防止学习者过度依赖一个表示类，使用辍学。合了几种小型的固定深度决策树。 明年，在 CoNLL-2003 共享任务中，Florian 等人。（2003）结合四种不同分类器的输出，获得了德语的最佳成绩。Qi 等人（2009）之后，通过在大量未标记语料库上进行无监督学习，通过神经网络改进了这一点。

以前已经为 NER 提出了其他几种神经结构。例如，Collobert 等人。（2011 年）在一系列单词嵌入上使用 CNN，上面有 CRF 层。 这可以被认为是我们的第一个模型，没有字符级嵌入，双向 LSTM 被 CNN 所取代。 最近，Huang 等人。（2015）提出了一个类似于 LSTM-CRF 的模型，但使用手工编写的拼写功能。 周和徐（2015）也使用了类似的模型，并将其适应于语义角色标记任务。 林和吴（2009）使用具有 L 正则化的线性链式 CRF，他们添加了从 Web 数据和拼写特征中提取的短语聚类特征。Passos 等人（2014 年）还使用了具有拼写特征和地名录的线性链通用报告格式。

像我们这样的语言独立 NER 模型在过去也被提出。库塞赞

致谢

这项工作部分由国防部高级研究计划局(DARPA)信息创新办公室(I2O)根据 DARPA/I2O 根据合同编号发布的紧急事故低资源语言(LORELEI)计划赞助。 人力资源 0011-15-c-0114。根据合同编号 FP7-ICT-610411（项目），Miguel Ballesteros 得到欧盟委员会的支持

多传感器)和 H2020-RIA-645012(项目 KRISTINA)。

参考资料

日久保田安藤和通张.. 2005a。 从多个任务和未标记数据中学习预测结构的框架。 机器学习研究杂志，6：1817-1853。

日久保田安藤和通张.. 2005b. 学习预测结构。 JMLR，6：1817-1853。

米格尔·巴列斯特罗斯，克里斯·戴尔和诺亚·A·史密斯。2015. 通过建模字符而不是使用 LSTMS 的单词来改进基于过渡的依赖关系解析。 在 EMNLP 的诉讼中。

米格尔·巴列斯特罗斯，尤夫·戈尔德格，克里斯·戴尔和诺亚·A·史密斯。 2016. 探索训练提高了贪婪的堆栈-LSTM 解析器。 在 ar Xiv：1603.03793 中。

Yoshua Bengio、Patrice Simard 和 Paolo Frasconi。1994. 学习具有梯度下降的长期依赖关系是困难的.. 神经网络，IEEE 交易，5（2）：157-166。

Chris Biemann， Gerhard Heyer， Uwe Quasthoff 和

Matthias Richter。 2007. 莱比锡语料库收集-标准大小的单语语料库。 语料库语言学论文集。

克里斯·卡里森-伯奇，菲利普·科恩，克里斯多夫·蒙兹，凯·彼得森，马克·普齐博基和奥马尔·菲·扎伊丹。 2010. 2010 年统计机器翻译和机器翻译指标联合研讨会的结果。在第五次统计机器翻译和度量匹配联合研讨会论文集，第 17-53 页。 计算语言学协会。

Xavier Carreras、Lluis Marquez 和 Lluis Padr6。2002. 命名实体提取使用 adaboost，第 6 次自然语言学习会议的记录。 8 月，31：1-4。

杰森·PC 邱和埃里克·尼科尔斯。 2015. 具有双向 lstm-cnns 的命名实体识别..*阿尔西夫预印阿尔西夫：151L08308*。

Ronan Collobert、Jason Weston、Leion Bottou、Michael Karlen、Koray Kavukcuoglu 和 Pavel Kuksa。2011. 自然语言处理（几乎）从头开始。机器学习研究杂志，12：24932537。

Silviu Cucerzan 和 David Yarowsky。 1999. 语言独立命名实体识别结合形态和上下文证据。 1999 年 SIGDAT 关于 EMNLP 和 VLC 的联合会议记录，第 90-99 页。

Silviu Cucerzan 和 David Yarowsky。 2002. 语言独立的 NER 使用统一的内部和上下文证据模型。 在第六次自然语言学习会议的记录中-卷 20，第 1-4 页。 计算语言学协会。

戴宏杰，赖宝亭，张永春，蔡德宗。 2015. 利用有代表性的标记方案和细粒度标记增强化合物和药物名称识别。*化学信息学杂志，7(Suppl.1)：S14*。

克里斯·戴尔、米格尔·巴列斯特罗斯、王玲、奥斯汀·马修斯和诺亚·A·史密斯。 2015. 基于过渡的依赖解析与堆栈长的短期内存。 在检察官办公室。 ACL。

雅各布·艾森斯坦，泰亚诺，威廉·W·科恩，诺亚·A·史密斯，埃里克·P·兴。 2011. 贝叶斯非参数化命名实体的结构化数据库。在 NLP 的第一次无监督学习研讨会论文集中，第 2-12 页。 计算语言学协会。

拉杜·弗洛里安，阿部·艾提契亚，红艳静，张彤。2003. 通过分类器组合命名实体识别 .. 在 HLT-NAACL2003-Volume4 第七次自然语言学习会议记录中，第 168-171 页。 计算语言学协会。

丹·吉里克、克里夫·布伦克、奥里奥尔·维尼亚尔斯和阿玛纳格·苏布拉曼亚。 2015. 多语言处理从字节。 阿尔西夫预印阿尔西夫：1512.00103。

大卫·格拉夫。 2011. 西班牙 gigaword 第三版 (ldc2011t12).. 语言数据联合会，宾夕法尼亚大学，费城，宾夕法尼亚州。

亚历克斯·格雷夫斯和吉尔根·施密杜伯。 2005. 双向 LSTM 网络的帧音素分类。在检察官办公室。ijcnn。

杰弗里·E·辛顿、尼蒂什·斯里瓦斯塔瓦、亚历克斯·克里哲夫斯基、伊利亚·苏茨克维尔和鲁斯兰·R·萨拉胡特迪诺夫。 2012. 通过防止特征检测器的共适应来改进神经网络。*阿尔西夫预印阿尔西夫：1207.0580*。

Sepp Hochreiter 和 Jiirgen Schmidhuber。 1997. 长的短期记忆.. 神经计算，9（8）：1735-1780。

Johannes Hoffart、Mohamed Amir Yosef、Ilaria Bordino、Hagen Furstenau、 Manfred Pinkal、 Marc Spaniol、 Bilyana Taneva、 Stefan Thater 和 Gerhard Weikum。 2011. 文本中命名实体的鲁棒消歧。 《自然语言处理经验方法会议记录》，第 782-792 页。 计算语言学协会。

黄志恒，魏旭，凯宇。 2015. 双向 LSTM-CRF 模型进行序列标注.. Co RRabs/1508.01991。

Yoon Kim、Yacine Jernite、David Sontag 和 Alexander M.Rush。 2015. 感知角色的神经语言模型。 共同 RR，abs/1508.06615。

狄德里克·金马和吉米·巴。 2014. 亚当：一种随机优化的方法。 阿尔西夫预印阿尔西夫：1412.6980。

约翰·莱弗蒂、安德鲁·麦克·卡勒姆和费尔南多·坎·佩雷拉。 2001. 条件随机字段：分割和标记序列数据的概率模型。 在检察官办公室。 ICML。

德康林和吴晓云。 2009. 短语聚类用于鉴别学习。 在第 47 届 ACL 年会和第四届 AFNLP 自然语言处理国际联席会议的会议记录：第 2 卷，第 1030-1038 页。 计算语言学协会。

王玲，林楚成，尤利亚·茨韦科夫，西尔维奥·阿米尔，拉蒙·费尔南德斯·阿斯迪略，克里斯·戴尔，艾伦·W·布莱克，伊莎贝尔·特兰科索。 2015a. 并不是所有的上下文都是平等的：更好的单词表示和可变的注意。 在检察官办公室。 艾蒙普。

王玲，蒂亚戈·路易斯，路易斯·马鲁霍，拉姆 6n·费尔南德斯·阿斯迪略，西尔维奥·阿米尔，克里斯·戴尔，艾伦·W·布莱克，伊莎贝尔·特兰科索。2015b. 形式上的寻找函数：开放词汇词表示的组合字符模型。 自然语言处理经验方法会议论文集。

罗刚，黄小江，林钦耀，聂再青。 2015. 联合命名实体识别和消歧。 在检察官办公室。 艾蒙普。

托马斯·米科洛夫、陈凯、格雷格·科拉多和杰弗里·迪安。 2013a。 向量空间中单词表示的有效估计。*阿尔西夫预印阿尔西夫：1301.3781*。

托马斯·米科洛夫、伊利亚·苏茨克弗、陈凯、格雷格·S·科拉多和杰夫·迪安。 2013b. 单词和短语的分布式表示及其组成性。 在检察官办公室。 尼普斯。

Joakim Nivre。 2004. 确定性依赖解析中的增量。 在增量解析：将工程和认知结合在一起的研讨会论文集中。

乔尔·诺斯曼、尼基·林格兰、威尔·拉德福德、塔拉·墨菲和詹姆斯·R·柯兰。 2013. 从维基百科学习多语言命名实体识别。*人工智能，194：151-175*。

罗伯特·帕克，大卫·格拉夫，孔俊波，柯晨，和和崎茂达。 2009. 英语 gigaword 第四版(ldc2009t13).. 语言数据联合会，宾夕法尼亚大学，费城，宾夕法尼亚州。

亚历山大·帕索斯、维内特·库马尔和安德鲁·麦克·卡勒姆。 2014. 词汇注入短语嵌入

用于命名实体分辨率的 Dings。阿尔西夫预印阿尔西夫：1404.5367。

齐燕军、罗南·科洛伯特、帕维尔·库克萨、科雷·卡武库格鲁和杰森·韦斯顿。 2009. 将标记和未标记数据与单词类分布学习相结合。在第 18 届 ACM 信息和知识管理会议记录中，第 1737-1740 页。 ACM。

列夫·拉蒂诺夫和丹·罗斯。 2009. 命名实体识别中的设计挑战和误解。第十三届计算自然语言学习会议记录，第 147-155 页。 计算语言学协会。

Cicero Nogueira dos Santos 和 Victor Guimaraes。 2015. 用神经字符嵌入增强命名实体识别。*ar Xivprint ar Xiv：1505.05008*。

Erik F.Tjong Kim Sang 和 Fien De Meulder。 2003. 介绍 conll-2003 共享任务：语言独立命名实体识别。在检察官办公室。 Co NLL。

Erik F.Tjong Kim Sang。 2002. 介绍 conll-2002 共享任务：语言独立命名实体识别。在检察官办公室。Co NLL。

约瑟夫·图里安，列夫·拉蒂诺夫，和尤斯华·本戈。2010. 单词表示：一种简单而通用的半监督学习方法。在检察官办公室。 ACL。

Matthew D Zeiler。 2012. 自适应学习速率方法。阿尔西夫预印阿尔西夫：1212.5701。

张岳和克拉克。 2011. 利用广义感知器和波束搜索进行句法处理。 *计算语言学，37（1）*。

张向，赵俊波，和燕乐村。 2015. 用于文本分类的字符级卷积网络。 神经信息处理系统的进展，第 649-657 页。

周杰和魏旭。 2015. 使用递归神经网络对语义角色标记的端到端学习。 计算语言学协会年会论文集。