基于双向语言模型的半监督序列标注

摘要：

从未标注文本中学习到的预训练词向量已经成为 NLP 任务神经网络结构的重要组成部分。但是大多数情况下，现在的循环神经网络还是从极少的标注数据中学习上下文相关的表示。所以我们这篇论文研究一种通用的半监督学习方法，将从双向语言模型中预训练出来的词向量加到 NLP 系统中，把它应用到序列标注任务中。我们在两个任务上做实验：NER 和 chunking。

介绍：

这篇论文我们探讨一种半监督学习方法，不需要额外的标注数据。我们使用一个神经语言模型，在大量未标注数据上训练，计算出每个位置上下文的编码，然后应用到半监督的标注模型中。

我们主要贡献是证明了 LM 训练出的上下文相关表示在半监督标注模型中是很有用的。

第二个贡献是多使用一个后向的 LM 效果更好。同时我们发现没必要针对某个领域数据来专门训练。

TagLM：

基本的序列标注模型

双向语言模型

两者结合

实验：

CoNLL 2003 NER

CoNLL 2000 chunking

预训练语言模型

训练

实验结果

分析

- 怎样使用 LM？
- 使用哪种 LM 重要吗？
- 针对特定任务的 RNN 的重要性？
- 数据规模？
- 参数个数？
- LM 跨领域吗？

相关工作：

未标注数据

神经语言模型

解释 RNN 状态

其他序列标注模型

结论：

我们提出了一种简单、通用的半监督方法，使用预训练的神经语言模型，来给序列标注模型增加上下文表示。

我们的方法在 NER 和 chunking 任务上比其他的方法都要好。

我们发现多使用一个后向的 LM 效果更好。

我们发现，即使 LM 在不同领域数据上训练，或者标注模型在大量标注数据上训练，效果依然有很大提升。