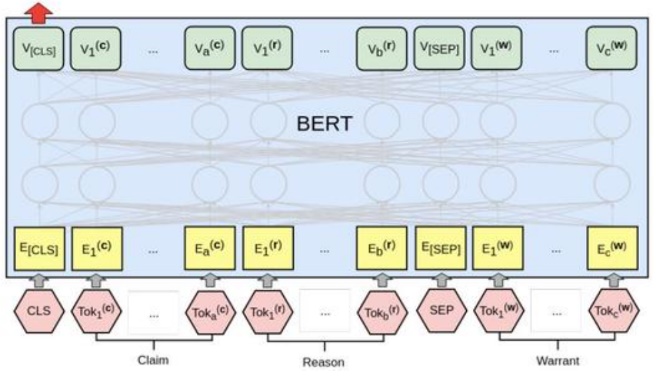
|  |
| --- |
| **两类命名实体识别算法** |
|  |
| **Xiao Chen,Dingzirui Wang** |
| Harbin Institute of Technology |
| {1180300121,1183710211}@stu.hit.edu.cn |
|  |
|  |

摘要

TODO

引言

随着互联网的迅速发展和普及，网络上的信息资源在变得越来越丰富的同时，也带来了信息过载的问题。如何从海量的文本资源中获取对用户有用的信息是目前亟待解决的一个问题。命名实体作为理解文本语义过程中的一种关键信息，在各种自然语言处理应用，如信息检索、自动文本摘要、自动问答、机器翻译和知识库构建等中起着至关重要的作用。

对于命名实体识别任务（Named Entity Recognition，NER），目前最常用的方法是把它们作为有监督的序列标注问题来求解。一般来说，监督数据越多，模型训练的效果就越好。但由于代价高，很难获取大规模的标注数据。因此，在这种情况下，人工设计的特征和特定领域的语言资源，被研究者广泛地运用在NER任务中。在新语言和新领域下开发特定的语言资源和特征的代价很大，这使得命名实体识别的研究也更具有挑战性。研究者开始研究如何从大量五标注语科学习词的表示方式来改进NER的性能。2011年，Collobert([2011](#Collobert2011))等是用于训练词向量初始化神经网络的词向量表，与使用随机初始化词向量模型相比，性能有了明显的提升。预训练词向量常用的方法是通过word2vec([Mikolov, 2013](#Mikolov2013))获得。word2vec虽然在一定程度上可以学习到词与词之间的一些相关信息，但是它训练得到的词向量在使用时与上下文是无关的。

图1：Bert作为编码器，线性层作为解码器

模型

Bert

如图[1](#图1)所示，该模型由两部分组成：一个BERT编码器和一个线性层解码器。

在喂给模型前，我们首先对句子进行word to piece操作（例如，将it's分解为it和's），然后根据词典中的序号，将每个piece序号化。

编码器会首先做embedding，将每个单词转换为一个768维的实向量。为了根据句子中各个单词的含义和关系来分析得到句中每个单词的标注，BERT中的transformer结构会综合每个单词的上下文信息，来生成反应单词深层信息的实向量。解码器会根据编码器的输出，得到一个维度为9的实向量，各个维度表示对应预测实体概率的对数值。

* + 1. 编码器

我们采用BERT模型来作为编码器，编码对话信息。

记编码器的输出为，其中为piece句子的长度，，每个在原单词的基础上，综合了句子的前后文信息。

注意到在喂给模型前，我们对每个单词进行了word to piece操作，但预测的结果应该是对每个单词的，所以我们在将编码器得到的

结果喂给解码器之前，应当先把组成该单词的piece的信息组合起来。具体来说，我们可以取各个piece的第一个、平均值或求和。在本文中取所有piece对应向量的第一个。

设每个word对应piece的角标范围为，其中为原句的长度。

则我们取作为编码器的输出，记作

* + 1. 解码器

考虑到BERT模型优秀的性能，我们决定直接采用线性层来得到最终的输出。

设为线性层的权重矩阵，则解码器的输出为。

由于词性标注任务可以看作是一个多分类认为，因此我们采用负对数似然损失NLLLoss来计算模型预测得到的结果和真实结果之间的误差。

为此，模型得到的应该是各个预测结果概率的对数，即。

相关工作

在早期的NER问题研究中，解决方法一般是基于词典和规则的，由于实现代价高且可扩展性差，所以得到的结果并不好。

传统机器学习方法求解NER一般基于HMM或CRF，其中尤其是CRF是NER问题目前的主流模型，它的目标函数不仅考虑输入的状态特征函数，而且还包含了标签转移特征函数，运用动态规划算法，来在给定输入序列求预测输出序列。但是，传统方法过于依赖人工特征，特征的设计需要很多的专家知识，特征选择的好坏更是直接影响到NER系统的性能。

近些年来，随着硬件计算能力的发展以及词的分布式表示的提出，神经网络方法夜奔引入到NER问题的求解中来，其主要的流程都是先将句子的token转化为稠密的embedding向量，随后将embedding序列送入到神经网络中，网络自动提取特征，最后再用Softmax来预测每个token的标签。

参考

R. Collobert, J. Weston, L.Bottou, et al. *Natural Language Processing (almost) from Scratch*. Journal of Machine Learning Research, 2011, 12(8): 2943-2537.

Mikolov T, Sutskever I, Chen K, et al. *Distributed representations of words and phrase and their compositionality*. Proc of NIPS. Cambridige: MIT Press. 2013:3111-3119.