NLP.TM | 命名实体识别基线 BiLSTM+CRF（上）

原创 机智的叉烧 CS的陋室 2019-07-11

心如止水

Ice paper - 心如止水

【NLP.TM】

本人有关自然语言处理和文本挖掘方面的学习和笔记，欢迎大家关注。

往期回顾：

- NLP.TM | tensorflow做基础的文本分类
- NLP.TM | Keras做基本的文本分类
- NLP.TM | 再看word2vector
- NLP.TM | GloVe模型及其Python实现
- NLP.TM | 我的NLP学习之路

命名实体识别是继文本分类之后的一个重要任务。在语言学方面，分词、词性标注、句法分析等，在工业应用方面，则有实体抽取等，其实都用到了命名实体识别技术，本文将介绍命名实体识别任务以及其重要的基线模型BiLSTM+CRF。

另外由于文章太长，所以我分为两块，理论和思路我放一篇，实现我放另一篇，本文是上篇。

想提前看我的代码的同学可以在下面的链接找到，拉到最下点击阅读原文也可以：

https://gitee.com/chashaozgr/noteLibrary/tree/master/nlp_trial/ner/src/bilstm_crf

懒人目录

- 命名实体识别问题剖析
- BiLSTM+CRF 模型

先修知识

本文涉及的理论知识非常多，为了保证下面的文章能看的更加流畅，建议大家确认下面内容是否至少能达到理解水平：

- 中文分词与文本预处理
- 自然语言处理下的Embedding，如word2vec等
- RNN系的深度学习模型结构，尤其是RNN的改进版LSTM以及其双向结果BiLSTM
- 条件随机场

命名实体识别概述

有关命名实体识别，其实我在之前的文章中提到过：

NLP.TM | 信息抽取

往简单的说，就是让机器在一段文本中，例如一篇新闻，找出对应的一个或者多个语义项，如人名、地点等，下面是一个非常简单的例子：

这篇新闻中，有时间地点，甚至是一些有关地震的具体数据都已经被标出，同时按照了时间、地点、地震数据等方面分类标出，这种任务就被成为命名实体识别。

下面，我们就来剖析这个问题，从而尝试找到这个问题的解决方案。

命名实体识别问题剖析

在《高效能人士的七个习惯》中提到一个非常重要的习惯，就是"以终为始"，即以最终的目标作为自己的努力方向和出发点，那我们也用这个思路，来看看命名实体识别的最终目标的底层展示是什么样的。

> 最 后 2 5 米 **曾 启 亮** 以 惊 人 的 爆 发 力 冲 刺
>
> O O O O O **B**nr M**nr E_nr** O O O O O O O O O

这是一句描述运动员冲刺过程的语句，上面一行是具体文本，下面一行是标注结果，这里通过"Bnr Mnr E_nr"的方式标注了"曾启亮"这个人名，而其他部分并不是人名，所以就都标为"O"。这里的B/M/E表示的是具体一个实体的起点、中间与终点，nr表示人名。

上面就是我们的目标，给定上面一行文本，我们要提出一种方法标注出其中的实体与非实体部分，那么抽象的，我们其实可以把它当做是一个"序列标注/分类"问题，即当做一种带有序列信息的有监督学习任务来完成，再换句话，就是一种特殊的分类问题，且这个分类并且只有一个输出，而是一串连续输出，每个分类其实都对应一个子分类问题。

把上面一句话抽象出来，其实就是给定一个序列{X(i)}，需要生成一串连续的分类结果{Y(i)}。

如此一来其实我们很容易能想到《统计学习方法》中的两个模型，也是这本书第一版里面难度最高的两个模型——HMM和CRF，确实如此，尤其是CRF，其实很长一段时间在该问题下是重要的基线模型，甚至在多个语言类工具包下都有所应用。

此处，不对这两个方法赘述啦，《统计学习方法》里面写的详细，另外推荐一本书《百面机器学习》，里面对概率图模型的解释比较浅显易懂，也可以辅助大家理解，至于用来解决命名实体识别问题，可以参照下面的文章：

https://blog.csdn.net/lilong117194/article/details/83106711

BiLSTM+CRF

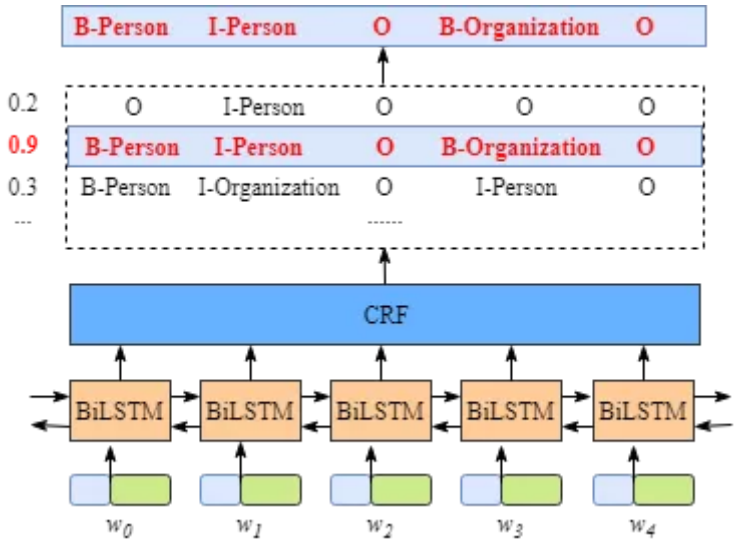
诚然CRF是一个非常优秀的模型，但是江山代有才人出，在深度学习逐渐普及下，更好地模型被引入，BiLSTM+CRF是成了目前重要的基线模型。下面是模型对应的论文原文（至少个人认为这应该是吧，网上对源论文的说法不一，有问题欢迎大家讨论）

Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.

这里有论文翻译：

https://blog.csdn.net/Elvira521yan/article/details/88415512

当然的，对于这个模型的解读，我觉得这张图应该是最为经典的，我就拿这张图来谈吧。



对于输入后的文本，我们用深度学习体系下的常规操作embedding将文本转化，然后输入BiLSTM中，此处的BiLSTM实质上就是两个独立反方向传播的LSTM，此处需要输出每一个输出节点的信息，正反向信息的输出经过拼接后，放入CRF中进行计算，最终得到序列标注结果。

当然的，这个也是深度学习后接机器学习的重要成功案例。

具体实现敬请期待下一篇文章，想看代码可以点击阅读原文查看~

阅读原文

喜欢此内容的人还喜欢