Dependency-based Siamese long short-term memory network for learning sentence representations

中文标题：基于依存句法分析的句子表示研究

在自然语言处理领域中，语言表示是至关重要的。一个优秀的语言模型能够大大提升文本理解和信息抽取等自然语言处理任务的效率。目前短文本表示(如单词或者短语)工作已经取得了巨大进步，诸如word2vec等短文本建模方法已经相当成熟可靠，并且被广泛用于各种自然语言处理任务。然而，长文本(句子、段落或者篇章)表示的研究由于句子等长文本相对于单词和短语有着更为复杂的结构，仍存在巨大挑战。由于LSTM在处理序列上的优势，它经常被用于对句子的建模，但是我们认为标准的LSTM模型获取的句子表示对句子的各个组成部分都“一视同仁”，没有对句子的结构信息(主谓宾)给予足够的重视。而这些被忽略的结构信息可能对句子表示有着更加重要的影响。基于这样的思想，我们提出了基于依存句法分析的D-LSTM（Dependency-based Siamese long short-term memory）。



该模型通过一个预训练好的依存句法分析器对句子进行依存关系分析，并结合原始句子输入得到句子的辅助表示，然后通过LSTM得到句子的基本表示，最终通过引入权重调节因子辅助表示和基本表示在句子表示中的重要程度，进而获得完整的句子表示。与传统的LSTM得到的句子表示相比，通过D-LSTM得到的句子表示能更加突出句子的主要语义。

D-LSTM的性能评估在SICK（a sentence involving a compositional knowledge）数据集上进行，该数据集9927个句子对，其中5000对用于训练，4927对用于测试。每个句子对都有一个相似度评分以及一个类别标签（共有3类）。我们在2个自然语言处理任务中对模型进行了评估，分别是语义相关度评分（semantic relatedness scoring）和蕴含关系分类（entailment classification）。语义相关度评分任务的评价指标为皮尔逊相关性系数，蕴含关系分类的评价指标为分类准确率。实验表明，D-LSTM在上述两个自然语言处理任务中都比SICK公布的最优模型的性能好。

1. Wenhao Zhu, Tengjun Yao, Jianyue Ni, Baogang Wei, Zhiguo Lu. Dependency-based Siamese Long Short-Term Memory Network for Learning Sentence Representations [J]. Plos One, 2018.