This is a note for the paper: Shikhar Vashishth,Manik Bhandari, Incorporating Syntactic and Semantic Information in Word Embeddings using Graph Convolutional Networks(ACL 2019)

paper and code have been published by the author

问题背景

希望得到包含句法信息和语义信息的词嵌入向量。

句法信息：一个词在句子中的成分信息，比如作为主语、宾语还是谓语

语义信息：一个词的词义信息

传统引入句法信息的方式：将作为句子不同成分的同一个词看作多个词，比如“水”，可以作主语，可以作宾语，还可以做谓语动词。这需要先将句子中的词标注上依赖关系（比入主语、宾语、谓语、修饰语）。

这种句法上的上下文称为dependency context。

相对应的，利用词之间前后关系的上下文叫sequential context。

传统引入句法信息的方式会大大扩充词表（因为一个词变为了多个词），需要更大的语料库，同时程序的运行速度会变慢。作者将单词看作节点，句法依赖看作节点之间的边，利用图上的局部运算来使输出的词向量包含某种句法信息（这合理吗？？）。等价于我们没有扩充词表，而是将多余的关系转化为了单词节点之间的边。但是有一个问题，两个词之间可能有多种依赖关系，不能通过一条边简单描述，实际中对同一个依赖关系进行权值共享。

具体细节

SynGCN

输入是一句话
通过句法关系构建一张图，权重与边的关系有关。虽然句子不同导致每张图不同，但是边的关系就那么几种，同一关系权值共享，十分自然。
图的输入就是基于上下文训练好的每个词的embedding vector，一个词为一个节点。节点间的关系有：主语、宾语、状语、修饰语等
这样再进行加入门控的GCN训练，得到的输出就具有了语义信息。
加入门控可以突出相关的边，减弱错误标记或无关联的边的影响

GCN可以更好的提取全局语义信息

SemGCN

节点代表词，边代表语义关系，具体有四种：上位关系、下位关系、同义关系、反义关系

输入词向量用SynGCN训练出来的词向量效果更佳

评测的下游任务

相似词的词向量距离
概念分类
词汇类推
命名实体识别
QA
句子成分标注？
识别同一实体的不同表示