智源学者成果展示——自然语言处理

作者　万小军（北京大学王选计算机研究所）

2020年6月

项目负责人：万小军

单位 / 团队：北京大学王选计算机研究所

成果名称：基于抽象语义表示的文本生成

成果简介：

抽象语义表示（AMR）将一个句子的语义抽象为一个单根有向无环图，很好地解决了论元共享问题。基于 AMR 的文本生成（AMR2Text）任务的目的是产生一个自然语言语句，该语句对输入的 AMR 图所编码的含义进行语言表达。图 1 展示了一个 AMR 图样例及其对应的句子。由于 AMR 在刻画句子语义时会剥离句子中谓词的时态、单复数，省略一些连接词等，句法结构信息也丢弃，因此它不再简单地与原始句子一一对应。同时，由于重入节点的存在，图结构的复杂性也大大增加，捕获基于图的数据中存储的复杂结构信息并非易事。这些特点和因素使得这一文本生成任务十分有挑战性。

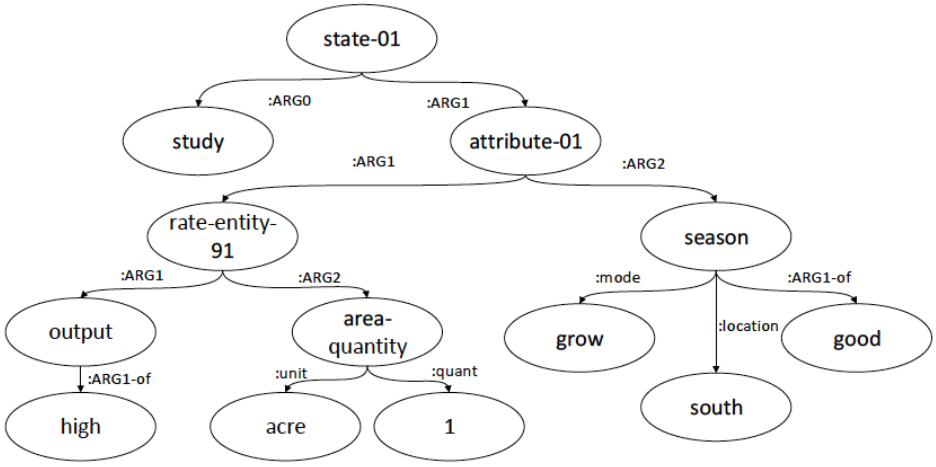

图 1　AMR 图样例（对应的句子为“The study stated that the high output per acre was attributed to a good growing season in the south”）

目前主流的 AMR 到语句生成的解决方法为采用图到序列的转换模型（如 GGNN、GraphLSTM、GCN 等），这些模型通过聚合相邻节点的信息来学习节点表示，并通

过多层编码网络的连接将局部信息传递到未直接相邻的节点，或者将未直接相邻节点对之间的路径编码作为它们的隐式语义关系，将其和显示语义关系信息共同聚合到节点表示中。这些图到序列模型仍然面临着生成文本中缺失信息和生成短语的语义与给定 AMR 图中语义关系不一致等问题。为了解决上述问题，申请人团队提出了一个图转换网络 Graph Transformer。基于 Transformer，Graph Transformer 的编码器和解码器都是由堆叠的注意力层所构成。图编码器直接以 AMR 图作为输入，使用了图注意力机制来聚合相邻节点的语义信息，从而进行节点表示的学习。堆叠的注意力层使得图注意力机制捕捉到的局部信息可以向远距离的节点传递，从而将全局信息也捕捉到节点的表示中。与 GAT 做法不同，Graph Transformer 为每个节点学习两个表示（也就是头表示和尾表示），并对出边关系和入边关系分别进行图注意力计算。这种做法使得模型可以在不同的空间对待与处理出边关系和入边关系，以更好地捕捉它们在语义信息上的特征。句子解码器通过自注意力机制来刻画已解码的文本信息，并通过注意力来利用图编码器学到的节点表示信息，同时使用了拷贝机制解决数据稀疏问题。所提出的 Graph Transformer 模型能够明显减少生成文本中信息缺失与语义不一致的问题。相关成果发表于自然语言处理领域顶级刊物 TACL，并受邀在 ACL 2020 进行宣讲。考虑到 AMR 图中节点之间关系的多样性，团队进一步提出 Heterogeneous Graph Transformer，该模型基于 AMR 图按照不同关系类型构建多个子图，并基于每个子图进行图转换器编码，并对结果进行深度融合。该模型在 LDC2015E86 和 LDC2017T10 两个基准数据集上分别取得了 31.84 与 34.10 的 BLEU 得分，代表了当前该任务上的 SOTA 水平。相关工作已经被自然语言处理领域顶级会议 ACL 2020 录用。

相关论文：

Tianming Wang, **Xiaojun Wan** and Hanqi Jin. AMR-To-Text Generation with Graph Transformer. *Transactions of the Association for Computational Linguistics (TACL)*.

Shaowei Yao, Tianming Wang and **Xiaojun Wan**. Heterogeneous Graph Transformer for Graph-to-Sequence Learning. **ACL** 2020.

Beijing Academy of Artificial Intelligence

微信关注
北京智源人工智能研究院