自然语言处理课程实验报告

Koji Tadokoro
Shimokitazawa College

1. 背景

句法分析（syntactic parsing）的主要目标是给定一个句子，分析句子的句法成分信息，例如主谓宾定状补等成分。最终的目标是将词序列表示的句子转换成树状结构，从而有助于更准确地理解句子的含义，并辅助下游自然语言处理任务。例如，对于以下两个句子：

您转的这篇文章很无知。

您转这篇文章很无知。

虽然它们仅相差一个“的”字，但是表达的语义截然不同，这主要是因为两句话的主语不同：第一句话的主语是“文章”，而第二句话的主语是“转”这个行为。通过对这两句话进行句法分析，就可以准确地获知各自的主语，从而推导出不同的语义。

典型的句法结构表示方法包括短语结构（phrase structure）句法表示和依存结构（dependency structure）句法表示。他们的不同点在于依托的文法规则不一样。其中，短语结构句法表示依托于上下文无关文法（Context-Free Grammars, CFGs），属于一种层次性的表示方法。而依存结构句法表示依托于依存文法，依存语法认为词与词之间存在主从关系，这是一种二元不等价的关系。在句子中，如果一个词修饰另一个词，则称修饰词为从属词（dependent），被修饰的词语称为支配词（governor），两者之间的语法关系称为依存关系。

陈丹琦等在 2014 年首次提出基于神经网络的依存句法解析器 [1]，在保持解析准确度的同时极大地提高了解析速度，这次实验参考这篇文章采用前馈神经网络来处理依存句法解析任务。

2. 要点简述

本小节就阅读文章时和实验中遇到的一些要点做出和讨论。

依存句法解析. 依存句法解析任务的目标是分析出对于一条给定句子 S 的依存结构。依存解析器的输出是对应于句子 S 的一棵依存树，在依存树中，S 的词通过不同类型的依存边相互连接。形式化地说，依存解析器实现了一个从给定输入序列 $S = w_0 w_1 \cdots w_n$（w_0 是 ROOT）到与其对应的依存树 G 的映射。

具体来说，依存解析主要解决以下两个子问题：

- 学习：给定经过依存树标注的句子训练集 D，归纳出一个解析模型 M，模型 M 能够对新的句子进行解析。
- 解析：给定一个解析模型 M 和一条句子 S，根据模型 M 推导得到句子 S 的最优依存树 G。

基于转移的依存句法解析. 基于转移的依存句法解析（transition-based dependency parsing）基于状态机模型，状态机定义了一系列状态和可能发生的状态转移，以实现一个从输入语句到依存树的映射。这种方法将依存树（图）的构建过程转化成一个状态转移序列，通过转移动作推动状态的变化，而转移动作的选择本质上是一个分类问题，分类器从目前的状态信息中提取特征以预测决定选择哪种转移动作。

那么在基于转移的解析方法中，学习问题就是建立一个模型，可以根据过往的转移和状态信息预测状态机的下一个转移动作，解析问题就是在给定之前建立的模型的情况下，构造输入序列的最优转移序列。

确定性的基于转移的解析方法. Nivre 等人提出了一种确定性的基于转移的依存解析方法 [2, 3]。给定一个句子 $S = w_0 w_1 \cdots w_n$，一个状态可以用一个三元组 $c = (\sigma, \beta, A)$ 表示。其中，σ 代表栈，其中包含来源于句子 S 的单词 w_i；β 代表缓冲区，其中包含来源于句子 S 的单词 w_i；A 是一个依存关系的集合，依存关系形如 (w_i, r, w_j)，w_i 和 w_j 是来源于句子 S 中的单词，r 表示依存关系的类型。起始状态和终止状态定义如下：

- 起始状态：起始状态 c_0 形如 $([w_0]_\sigma, [w_1, \cdots, w_n]_\beta, \varnothing)$，此时栈 σ 中只有起始节点 ROOT，其他所有句子中的单词都在缓冲区 β 中，集合 A 中目前不存在任何依存关系，为空集。
- 终止状态：终止状态 c_{end} 形如 $([w_0]_\sigma, [\,]_\beta, A)$，此时缓冲区中没有单词，表明句子中的所有单词都经过了模型的处理，栈中只有起始节点 ROOT。

同时，定义了三种状态间的转移动作：移进（`SHIFT`）、左弧归约（`LEFT-ARC`）和右弧归约（`RIGHT-ARC`）。

基于神经网络的解析方法. 设计的状态机模型能够完成对句子的句法解析，但并没有指定模型靠什么进行决策。传统模型使用机器学习方法训练得到分类器，让分类器来决定每一步该如何进行状态转移，但是这类方法仍然存在不足：

- 稀疏性：传统方法的决策器使用的特征是高度稀疏的（比如 one-hot 向量以及 tf-idf 向量）。
- 不完整性：无法完美的量化句子所包含的所有信息；相较之后提出的方法，传统方法在这方面做的更差。
- 特征计算代价昂贵：决策器计算特征向量的过程需要花费相对较多的计算资源以及时间；在陈丹琦等所做的实验中，模型 95% 的运行时间都花在了构建特征上，而非预测 [1]。

陈丹琦等提出的神经依存解析模型 [1]较好地解决或缓解了这些不足，该模型与传统的基于转移的解析模型类似，仅在状态转移决策部分存在较大区别，文中选择了一个只有一层隐藏层的前馈神经网络作为选择转移动作的分类器。对于稀疏性，神经依存分析模型利用低维稠密的分布式向量对特征进行表示；对于不完整性，神经依存分析模型将单词、词性标签以及依存类别标签的分布式向量表示进行拼接作为神经网络模型的输入，这种方式所包含的信息量更大，也更加准确；对于传统方法特征计算代价昂贵的问题，神经网络方法表现出了相对更好的效率。总的来说，相比于传统的方法，神经依存模型的效果更好，同时计算速度更快。

3. 实验设置与实验结果

实验设置. 数据集采用 cs224n 课程作业提供的满足 CoNLL 格式的英文依存语料。

实验结果. 在验证集上达到的最高 UAS（Unlabeled Attachment Score）为 88.89%，在测试集上最高达到了 88.77% 的 UAS。

参考文献

[1] D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014. 1, 2

[2] J. Nivre. An efficient algorithm for projective dependency parsing. In *Proceedings of the Eighth International Conference on Parsing Technologies*, pages 149–160, Nancy, France, Apr. 2003. 2

[3] J. Nivre and M. Scholz. Deterministic dependency parsing of English text. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 64–70, Geneva, Switzerland, aug 23–aug 27 2004. COLING. 2