清华大学深圳研究生院

计算智能实验室 周报

|  |  |  |
| --- | --- | --- |
| **姓名** | **研究方向** | **报告覆盖时间** |
| 李晨辉 | NLP（法律文本分类） | 2010. 2.26-2019.33 |
| **本周主要完成的工作** | | |
| 1. NLP领域基础知识学习； 2. 数据结构算法复习； | | |
| 1. **NLP领域基础知识学习**   自然语言特性有离散性（discrete）、组合性（compositional）、稀疏性（sparse）；深度学习在NLP方面的应用：   1. 嵌入层（embedding layer）将离散的符号映射为相对较为低纬的连续向量 ，向量之间的距离可以等价为单词（或句子等）之间的距离； 2. 两种网络  * 前馈网络(也叫‘多层感知机’‘Multi-Layer Perceptron’），它是一种输入大小固定，忽略输入顺序的网络；可应用于所有线性模型能用的地方，且具有非线性的特点； * 循环/递归网络： * 循环神经网络（RNN），适用于序列数据的特殊模型，允许打破“马尔可夫假设”（马尔可夫假设：当前事物的状态只与其之前的n个状态有关），不受制于数据的稀疏性。 * 递归网络：将循环网络的应用从序列扩展到树  1. 自然语言处理任务用到的特征类型：   直观特征包括：   * + - * 单独词特征：词元和词干、词典资源、分布信息       * 文本特征：词袋、权重、tf 、idf       * 上下文词特征：n-gram、位置       * 词关系特征  1. 特征应用案例：   主题分类：主要应用的是二元文法词袋；  作者归属：应用到特征集合，其中主要包括：功能词与代词词袋、词性词袋、 词性的二元、三元、四元文法词袋  命名实体识别（NER）：性质是序列分割任务，但通常被建模为序列标注任务；  核心特征也类似与词形标注任务的特征；例如焦点单词周围每侧大小为2窗口中的单词；单词共现度等。 | | |
| **二、**数据结构算法复习 | | |
| 本周主要对动态规划类问题进行学习：  问题1：  给出一根钢条长度及切割后每段价值，求最优切割方法（切割后总价值最大）  解决思路：   * 带备忘录自顶向下法   递归求解，对每一段是否切割都判断q=max(q,p[i]+CUT(p,n-i)),q为当前总价值，p为价值列表，n为钢条长度。但在求解的过程中记录已求解的子问题（CUT函数进入时先判断是否已有当前问题最优解，有则查表返回，没有则求解，返回之前将求解结果记录。）   * 自底向上法   从长度为1的钢条一直求到长度为n,求解过程中记录每个长度钢条的最优解。  问题2：  矩阵链乘，根据结合率求最优的计算顺序。以10x100,100x5,5x50的三个矩阵为例，应该先求10x100x5,再求10x5x50，共7500次乘法;否则100x5x50+10x100x50=75000;计算量差了10倍；  解决思路：  使用动态规划，矩阵Ai的规模为p(i-1)xp(i),m[i,j]记录Ai...Ak...Aj  链乘法运算的次数；将原问题变为更小规模的子问题求解，  m[i,j]=min(k=i+1 to j-1){m[i,k]+m[k+1,j]+p(i-1)p(k)p(j)}  采用自底向上的求解方法，依次将长度为2,3,n的所有矩阵链的乘法次数求出来，并记录划分位置（使得m[i,j]最小的k）。 | | |
| **下周工作主要安排**   1. 对深度学习在NLP方面的具体应用案例如具有哪些语言模型、wordEmbedding等进行学习； 2. 继续学习一些其他的数据结构； | | |