南京大学本科毕业论文（设计）中期检查表

|  |  |  |
| --- | --- | --- |
| 论文题目  题 目 | 中文：基于历时词嵌入模型的语义演变研究 | |
| 外文：Analysing Lexical Semantic Change Using Temporal Word Embeddings | |
| 学号：191220154 | | 姓名：张涵之 |
| 所在院系：计算机科学与技术系 | | 专业：计算机科学与技术 |
| 指导教师：黄书剑 | | 职称：副教授 |
| 计划完成时间：2023/5/17 | | |
| 论文（设计）的进度计划：  1-2周：阅读相关论文和源码，选择合适的模型（主要考虑不同时间片上训练的词向量如何对齐到同一个向量空间）。收集和整理人民日报新闻数据。  3-4周：考虑到人民日报语料规模较小，时间跨度较短（相对其他在英文语料上进行的研究），选定Di Carlo, V., Bianchi, F., & Palmonari, M.的TWEC（Temporal Word Embeddings with a Compass）模型（该模型无需先训练再进行对齐，而是通过预训练的一个compass来保证不同时间片的词向量自动处于同一空间）。  编写代码并首先用英文语料进行了测试。继续收集和整理人民日报新闻数据。  5-6周：进行数据预处理（按年份划分，对齐格式，分词，去停用词）。  训练得到词向量，筛选出可能存在语义变化的词汇进行语言学分析。  7-8周：尝试其他contextualized模型（不一定完成，取决于难度和工作量）。  如果完成，将训练结果与之前的一般word embedding模型进行对比。  9-10周：写论文。 | | |
| 已经完成的内容：  选定模型，收集和整理人民日报新闻数据，编写代码，数据预处理，在划分好的时间片上训练词向量。与之前的思路（从词义演变的语言学论文中找出词语，用词向量来验证语义变化）不同，尝试了对不同时间片上达到一定词频（暂定为出现次数≥100）的词语取交集，计算每个词在1946-1949和2020-2022两个时间片上的余弦相似度，从小到大排序，然后再根据排序结果筛选可能存在语义变化的词进行语言学分析（如分类为词性改变、由本义到引申义、搭配对象改变、褒贬程度改变等）。分析中除了最近邻（语义空间中最接近的5或10个词语），还试图采用可视化工具展示词义的变迁（目前，降维的效果还不太好）。 | | |
| 指导教师意见：（对目前完成情况的指导性意见，不少于100字）  指导教师签字：  年 月 日 | | |