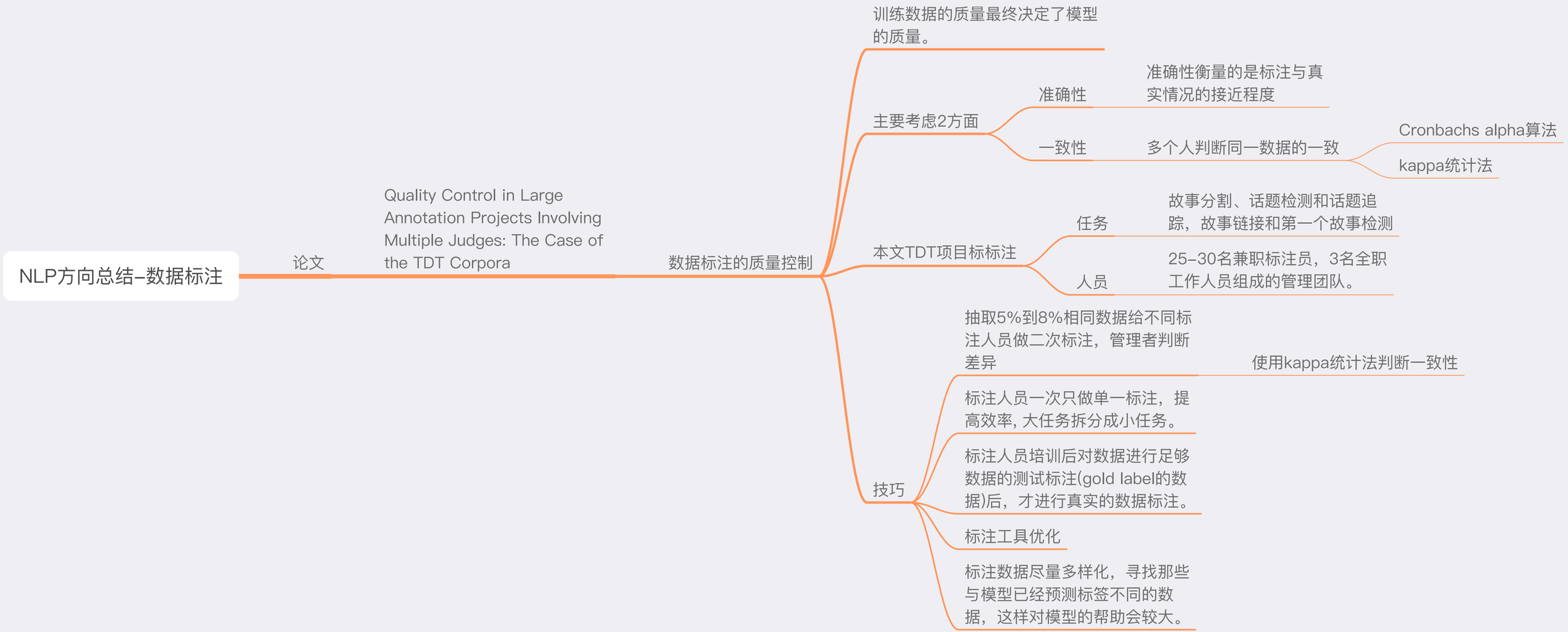
NLP方向总结–数据标注
论文
Quality Control in Large Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora
数据标注的质量控制
训练数据的质量最终决定了模型的质量。
主要考虑2方面
准确性
准确性衡量的是标注与真实情况的接近程度
一致性
多个人判断同一数据的一致
Cronbachs alpha算法
kappa统计法
本文TDT项目标标注
任务
故事分割、话题检测和话题追踪，故事链接和第一个故事检测
人员
25–30名兼职标注员，3名全职工作人员组成的管理团队。
技巧
抽取5%到8%相同数据给不同标注人员做二次标注，管理者判断差异
使用kappa统计法判断一致性
标注人员一次只做单一标注，提高效率，大任务拆分成小任务。
标注人员培训后对数据进行足够数据的测试标注(gold label的数据)后，才进行真实的数据标注。
标注工具优化
标注数据尽量多样化，寻找那些与模型已经预测标签不同的数据，这样对模型的帮助会较大。