开源大数据智能处理方法及应用

黎梦雪、李明兰、刘海鲸、张雪

联系人：张雪 手机号：18810350251 邮箱：zhangxue2015@iscas.ac.cn

基于众包的中文酒店评论论辩语料库的构建

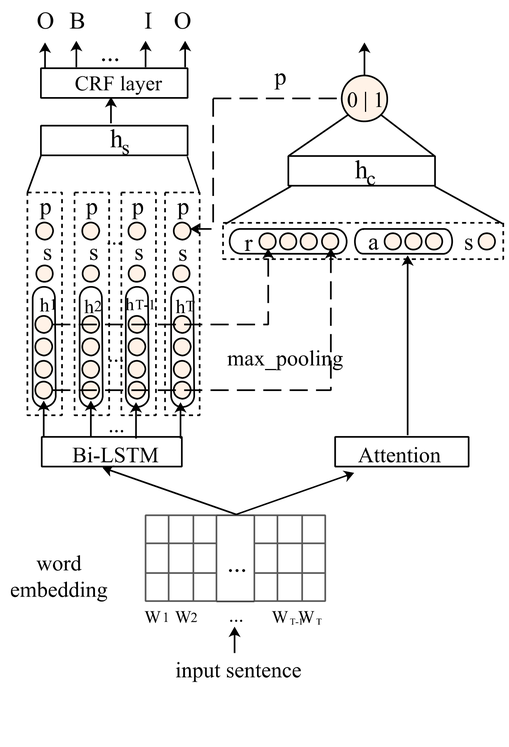
研究背景：论辩挖掘是一项研究如何从自然语言文本中自动提取出论辩结构的技术。将AM应用于用户评论中有着巨大的前景：应用于推荐系统中，通过分析用户评论给出可解释的推荐；应用于舆情监控中，可以辅助决策者了解舆论变化的趋势及原因。然而，论辩挖掘作为一项比较新的技术，由于论辩结构标注的争议性本质，目前面临的最大困难就是标注语料库不足的问题，面向论辩挖掘的中文标注语料库更是尚未出现。

语料库简介：使用众包技术建立了第一个面向论辩挖掘的中文酒店评论语料库，克服了传统聘请专家标注建库方式成本高、速度慢、规模小的缺点。该语料库一共包含了4814个论辩组件，3243个情感极性的标注以及411个论辩关系的标注，且此语料库的标注质量可与专家标注的语料库质量相当。

论辩组件识别的联合RNN模型

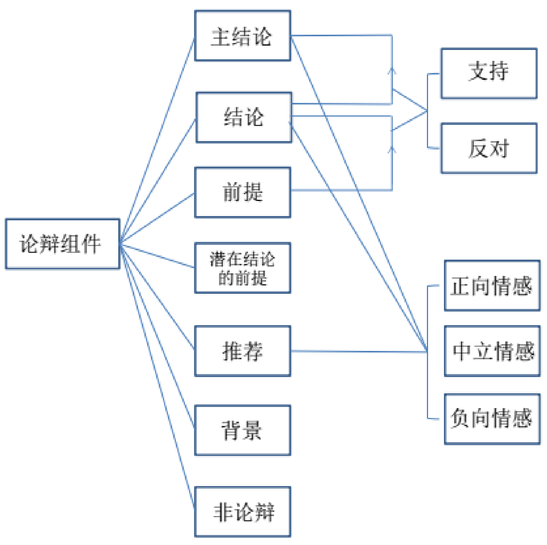
模型介绍：将句子是否包含论辩组件这一分类问题和句子中组件详细边界的序列标注问题进行联合建模。

应用场景：论辩文本挖掘中的论辩组件识别问题，同时也适用于其他序列标注任务。



- 双向LSTM模型自动学习输入文本的隐含语义特征；
- 通过最大池化层获得句子的特征表示r；
- 通过Attention注意力机制对输入的词语向量学习不同的权重，得到输入语句的加权表示a；
- 将这些特征表示和结构化信息s连接，通过softmax层得到句子的分类结果；
- 句子的分类结果与双向LSTM的隐层表示一起用于条件随机场模型中，从而实现句子的序列标注任务；
- 损失函数由序列标注问题和分类问题的损失函数联合表示。

基于论辩特征的用户评论质量的预测与分析



用户评论论辩结构示意图

模型介绍：本工作从论辩的角度对用户评论质量进行了预测，通过定义用户评论论辩结构、设计论辩特征进而对用户评论中的推理论证信息进行建模。

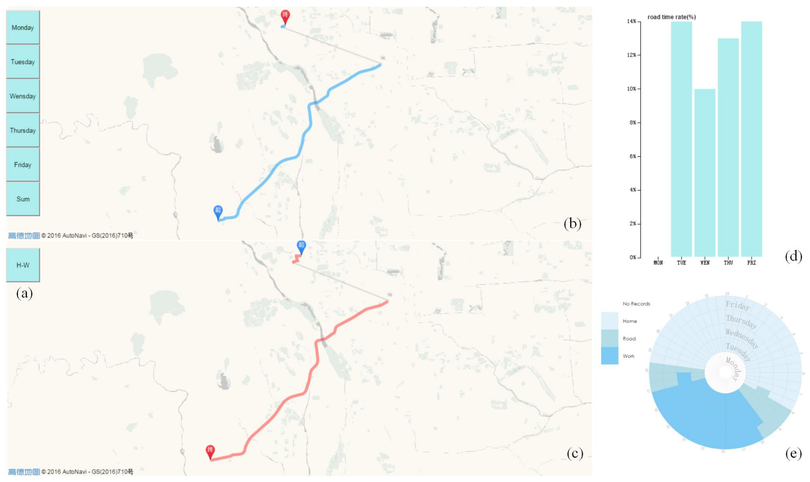
	Accuracy	F1-score	AUC
STR	0.600	0.450	0.500
STR+AF	0.604	0.607	0.599
UGR	0.697	0.646	0.627
UGR+AF	**0.718**	**0.717**	**0.706**
GALC	0.621	0.579	0.560
GALC+AF	0.647	0.649	0.640
INQUIRER	0.533	0.517	0.493
INQUIRER+AF	0.657	0.659	0.651

AF与不同Baseline结果对比表

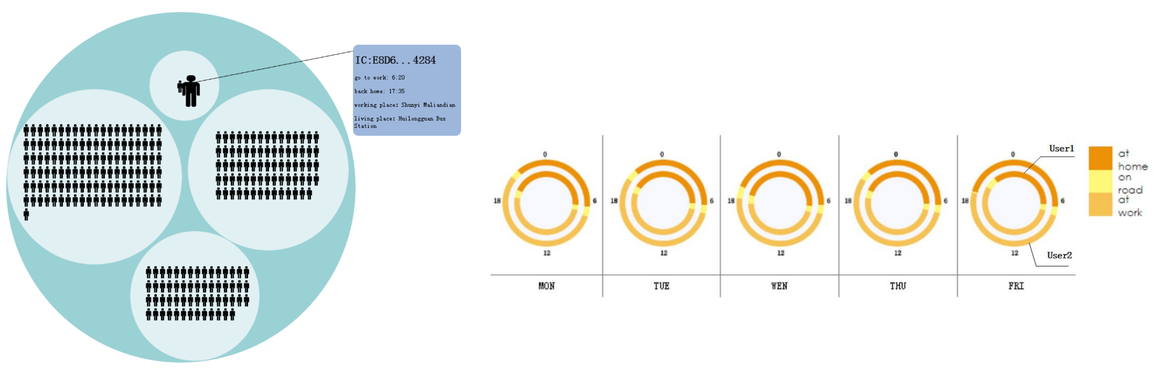
实验结果：左图是本文的实验结果列表。实验结果表明：结合本文提出的基于论辩特征的方法(AF)，现有的四种Baseline方法的预测性能均有了显著的提升。

结论：基于论辩的方法通过对评论中的推理论证信息建模显著提升了预测性能，用户评论中的推理论证信息对于评论质量的预测有积极的影响。

基于出行行为的可视分析系统



系统介绍：可视化乘客出行信息，一方面补全乘客刷卡数据中的缺失信息，另一方面辅助判断乘客的职住分布，进一步可以从城市空间的宏观层面分析城市的功能区分布以及“多中心性”的特征。柱状图与同心圆用以分析乘客时间分配情况。



算法介绍：利用K-means算法对城市中人群进行聚类，通过分析职住分布以及上下班时间，辅助找到城市中存在的“熟悉的陌生人”。上图中左边代表对居住在北京某小区的居民进行聚类的结果，左边代表某两个职住区域相近的乘客的乘车时间表。