**报告题目：中文自然语言处理亚符号特征研究**

**内容简介：**

自然语言处理是人工智能、机器学习、语言学等的交叉学科，目标是让计算机能够在一定程度上理解人类语言进而进行相关任务，如信息抽取、数据挖掘、信息检索、问答系统、机器翻译等。本报告主要讨论分词、词性标注这两个中文自然语言处理的基础步骤。现有机器学习模型所使用的符号特征存在数据稀疏等问题，且效果受限于有限的人工标注的数据集。本报告研究使用实数向量这一亚符号形式表示的中文词汇特征，其可从海量的不需要人工标注的数据中获取，进一步提高了分词、词性标注的效果，并可能应用于其它中文自然语言处理任务中。

**报告人简介：**

张开旭，2012年于清华大学计算机科学与技术系获得博士学位。研究兴趣为机器学习、中文自然语言处理。目前为厦门大学博士后，申请获得博士后基金课题《基于深度神经网络的自然语言亚符号特征研究》。