# 技术部语义系统26周汇报模板（20160620-20160624）

**一．本周主要工作内容如下：**

1. 清洗网络爬虫爬取的门户网站数据，字幕数据，问答数据;
2. 对聚熵的chat or not的数据，针对于儿童类型的数据进行抽取；
3. 基于固定模式的控制语句的语句模式设计和合成控制命令；
4. 基于合成语句的实体发现算法尝试，爬虫配合爬取更多实体相关数据；
5. 技术部工作梳理和工作部署。

**二．【详细进度】如下：**

|  |  |  |  |
| --- | --- | --- | --- |
| **任务** | **本周工作总结** | **下周工作计划** | **备注** |
| 清洗网络爬虫数据 | **1) 问答类数据**  a）问题提取，目前已经清洗出7500+多句问题，具体的问答对数据的条数；  b) 网络爬虫的问答对提取工作继续进行。  **2)其它数据**  a) 动画片/动漫字幕清洗，目前清洗150+个字幕，得到12.5万条数据。；  b) Childes数据库中中文数据清洗，得到8.5万条。  一共大概约4MB的数据，数据量比较少。  **数据存放地：**\\192.168.1.102\ShareFiles\语音语义\语料\qieyin语料\自然语言语料\网络资源清洗 | 1）继续尝试获取更多的动画片的字幕数据。 | 对于语义系统而言，字幕这种自由对话比百科类的语句效果更好 |
| **筛选聚熵的chat or not的儿童控制语句** | 主要以“动画片”、“儿歌”、“儿童故事”、“动漫”、“儿童视频”关键词进行筛选， 并且去掉"怎么"、"什么"、"怎么样"、"喜欢看什么"、"哪"、"吗"、"是什么"、"有什么"等带有聊天性质的词语的句子；一共得到1.36万条句子。  **数据存放地：**\\192.168.1.102\ShareFiles\语音语义\语料\qieyin语料\聚熵ChatOrNot筛选 |  |  |
| **基于模式的控制语句合成** | 1. 设计控制命令模板 2. 按照资源名称合成控制命令语句   **数据存放地：**\\192.168.1.102\ShareFiles\语音语义\语料\qieyin语料\基于模板的合成语料 |  | Emma |
| **实体发现算法** | 基于CRF的实体发现算法研发 |  | Emma |

**三．遇到的问题汇总**

1 数据资源少，需要更多增加一些获取数据的渠道，只靠爬虫和人工编辑还是比较慢。