清华—中科汇联项目合作总结

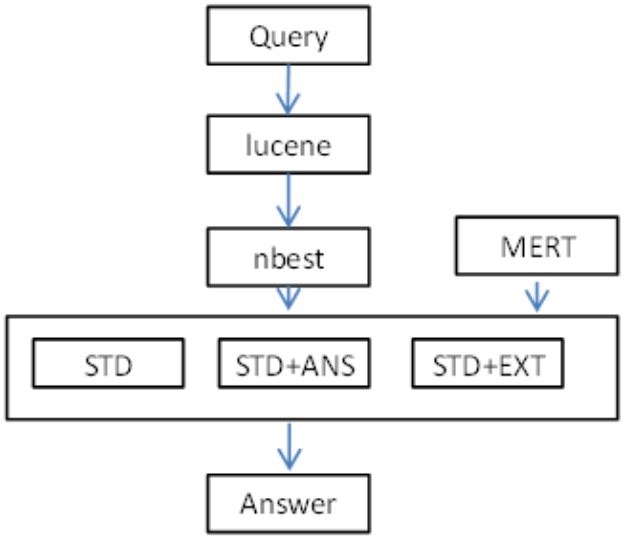
邢超 骆天一 刘荣
清华大学信息技术研究院语音与语言技术中心
北京中科汇联信息技术有限公司

目录

- 研究工作
  - 检索性能提升
  - 拼写检查
  - 知识图谱与反问
  - 线下训练模块
  - 部分调研工作
  - 排序学习
  - 错别字纠错
  - RNN自动古诗生成
  - RNN相似问句判别

研究工作—检索性能提升

- Lucene 检索性能的提高
  1. 关键词权重：NER,Parser
  2. 多字段组合的全维度检索：MERT做优化 提升2%

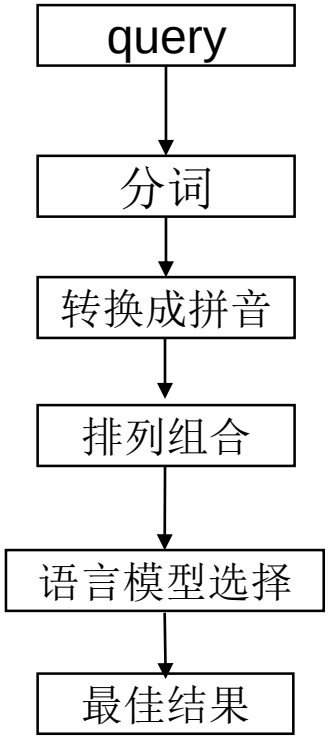


研究工作—拼写检查

- 基于语言模型与注音的拼写检查
  - 例子

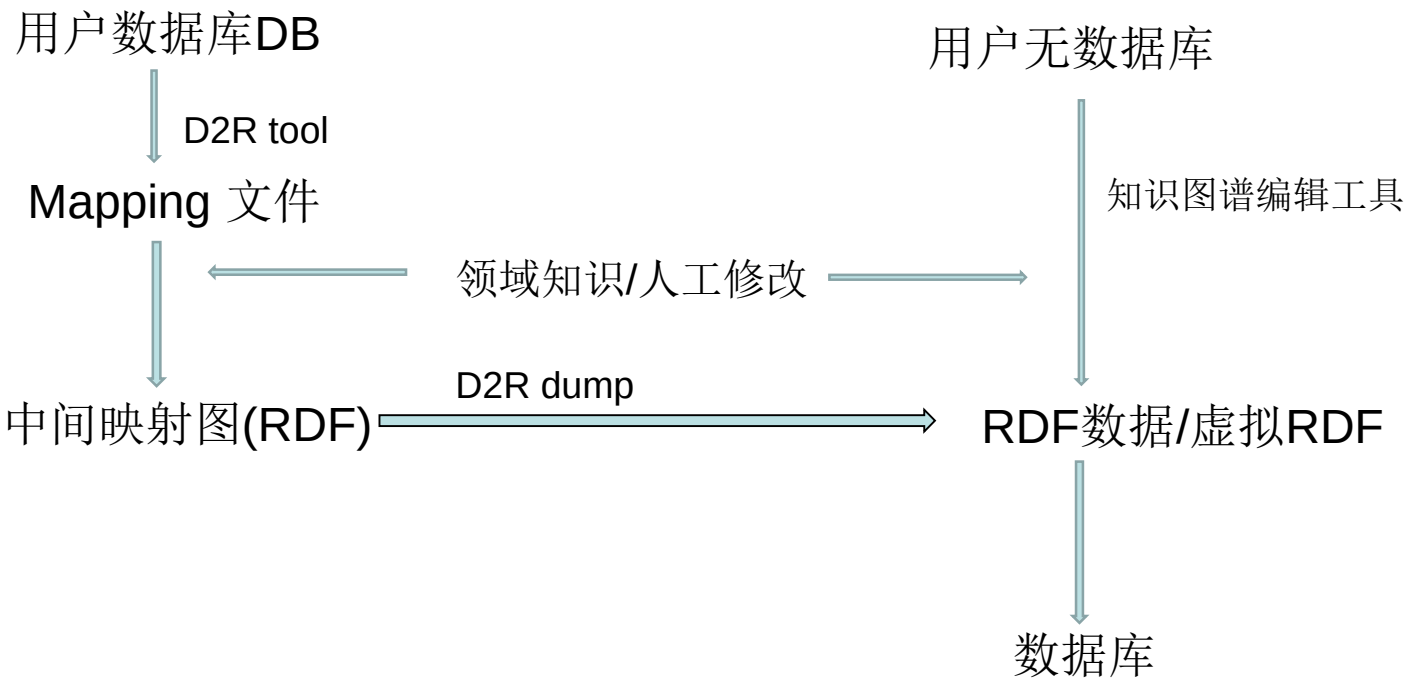
“如何办理虎口” => “如何办理户口”

“青华大学怎么走”=>”清华大学怎么走”



研究工作—知识图谱与反问

- 基于知识图谱的检索--DB



研究工作—知识图谱与反问

- 基于知识图谱的检索—检索
  - 渤海银行项目
  - 例子

电话是多少？

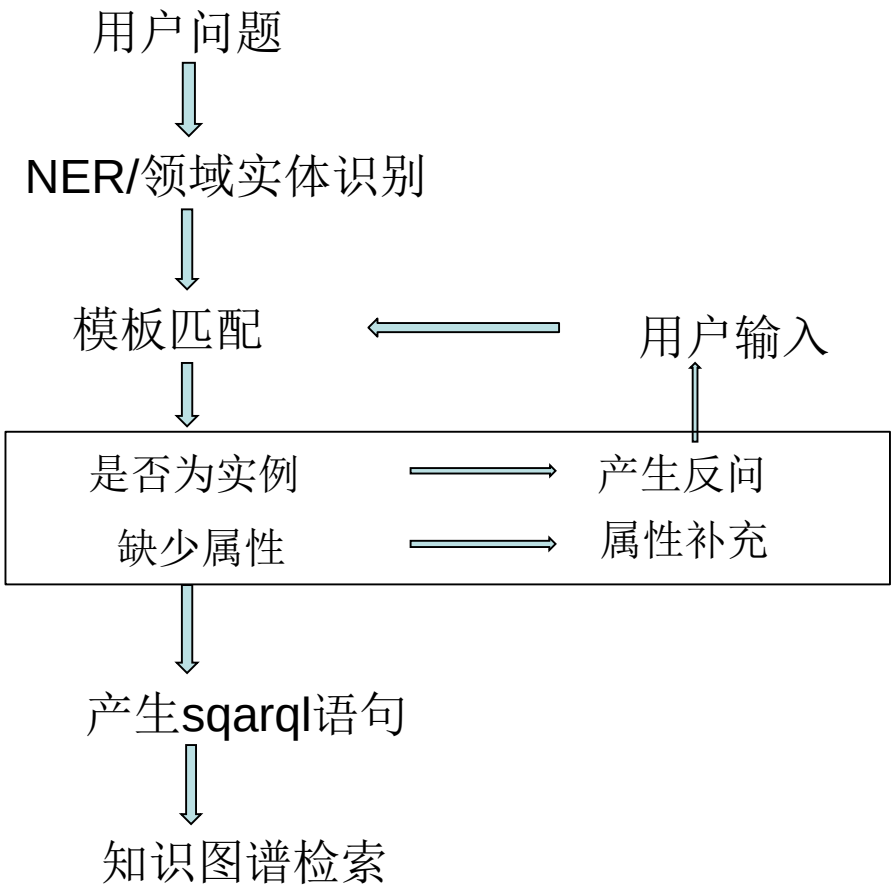
请问是哪个员工的？

小明

小明的电话******

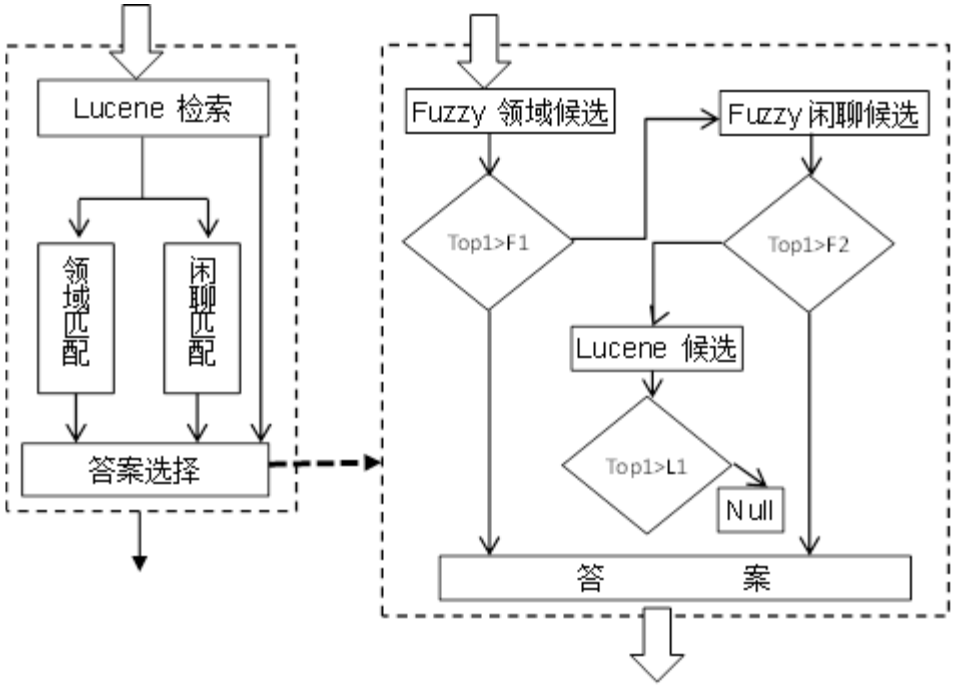
那小东的呢？

小东的电话******



研究工作—线下训练模型

- 线下训练模型
  - 遗传算法的模型训练 2%
  - 一种自动训练参数的模型
  - 一篇专利



研究工作—部分调研工作

- 调研工作
  - 句法分析工具
  - 语义归一化
  - 指代消歧

目录

研究工作—排序学习

- 离线学习
  - p@1 准确率： 61%->65%
  - 训练速度：最多提升1000倍
  - 论文发表：EMNLP 2015, 一篇长论文
- 在线学习
  - 用户可输入QA对进行系统调教->实时增加QA对
  - 用户对排名靠后结果点赞->实时排名提升

研究工作—错别字纠错

- 系统解决方案及功能
  - 大语料语言模型（全领域）+小语料语言模型（具体领域）
  - 支持国家领导人名字纠错
  - 支持错别字纠错对手动扩充
- 纠错效果（汇联的测试集）
  - 系统标出71处错误（实际错误6处，误报比约9/10）

我进期要申请美国。青华大学。李瑞坏是第十五届中央委员。

一个需要纠错的文本

纠错结果: **1 进期 null 9 青华大学 null 16 坏 环**

研究工作—RNN自动古诗生成

- 解决方案及功能
  - 利用RNN（循环神经网络）进行绝句自动生成
  - 支持五言绝句和七言绝句
  - 训练集：1000首五言绝句和1000首七言绝句
- 自动生成效果
  - 床前大风车 雨以此独甘 旧为玉所中 不为君恩开
  - 河曲回千里 鸣月隐世多 水水相苏乐 我以无金流
  - 风雪送春归 风水在征绯 至风有月人 今人几上情
- Demo

研究工作—RNN相似问句判别

- 解决方案及功能
  - 利用RNN（循环神经网络）进行相似问句判别
  - 系统用于自动模板扩充和问句相似度判别
  - 训练集：300多万QA对
- 相似度判别效果
  - Q:芜湖 的 计算机 软件 水平 考试 在 什么 地方 报名 ？

    A:你 可以 到 安徽 师大呀。

    相似度为：0.823901910035
  - Q:芜湖 的 计算机 软件 水平 考试 在 什么 地方 报名 ？

    A:1 . 找到 卖 你 电脑 发 的 磁盘 2 . 把 磁盘 放 进光 驱 3 . 按照 向导 安装

    相似度为：0.759943815081