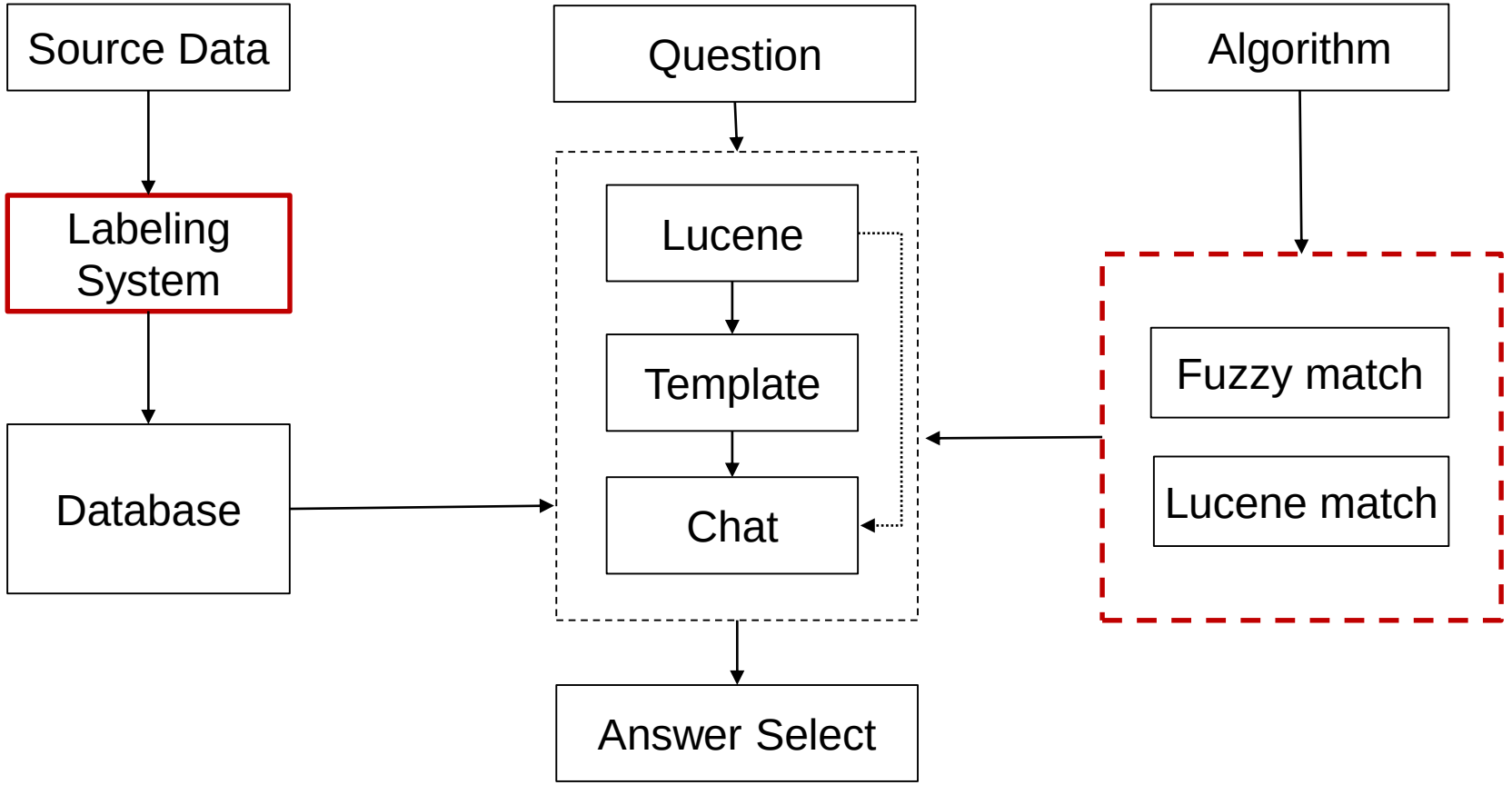
中科汇联问答系统v1

刘 荣
北京中科汇联信息技术有限公司
2014-10-8

目录

目前版本--框架


Source Data
Labeling System
Database
Question
Lucene
Template
Chat
Answer Select
Algorithm
Fuzzy match
Lucene match

问答系统—结构知识

- **机构化知识—三元组**
  1. 例子：木里县

     “如何办理身份证” ➔ 〈身份证 ， 办理 ，流程，木里县〉 ➔ 答案

     “哪里可以办理身份证” ➔ 〈身份证 ， 办理 ，地点/机构，木里县〉 ➔ 答案

     “办理残疾人证的收费标准” ➔ 〈残疾人证 ， 办理 ，费用，木里县〉 ➔ 答案

  2. 三元组

     实体事务：身份证，户口，企业。动作：办理，办，补办，注销..

     〈实体事务，动作，实体事务属性，地点〉：〈身份证 ， 办理，流程 ， 木里县〉
- **复杂结构知识**
  1. 例子：木里县

     “残疾人办理企业享有哪些优惠政策” ➔ 〈企业 ， 开办 ，**政策**，木里县〉

     〈**政策**，企业，残疾人〉

问答系统—结构知识

- **机构化知识技术**

  1.代码参考

  bookQA,tencentQA

  2.论文

  关于三元组的抽取及对应，有很多成熟的技术。

- **机构化知识优点**

  1. 可以精准回答，尤其对于业务知识比较固定。
  2. 可以快速应用于各个政府，减少后期的数据标注的工作量。

- **结构化知识缺点**

  1.前期需要定义三元组，工作量可能会比较大。

  2.需要相应的标注系统进行配合标注。

  3.对于复杂句式和知识外的问题，不能回答。

问答系统—模板扩充

- **自动模板扩充**

  1. 例子

  “如何办理身份证”➔ [请问]{如何}{办理}{身份证}[呢]

  扩展： [请问]{如何,怎么，该怎么}{办理，办}{身份证}[呢]

  {身份证}{如何，怎么}{办理}[呢]

  2. 技术

  可利用ngram,句法分析，词性标注

  3. 优点

  可以快速扩充模板，减少人力。并且可利用大数据知识。

问答系统—语义信息

- **语义信息**

  1. 关键词及词语权重

     业务词权重

     非业务词权重计算：TF-IDF,proME

  2. 词语相似度

     word2vec：需要重新训练对于不同词表

     同义词林，hownet：词语难以覆盖业务词

  3. 目前技术

     已经准备好相应模块，需要进行测试及改进

问答系统—标注系统

- **标注系统**
  1. 现在是否有原型

专利

- **准备的专利**

  1. GA 方法去优化文档系统

- **保护公司算法及系统**

  1. 开发过程中如何保护公司问答系统的算法，知识整理，标注系统，流程的专利。是否需要咨询相关专利公司

其它

- **代码规范**
  1. 代码规范文档
- **合作开发**
  1. 如何规范代码的合作开发