知识分析比较实验

I. 知识抽取范围比较

本实验的目标在于比较KG方法抽取的知识内容与文本查询的方法有何异同。

比较对象

Converter

当前诊断方法中的KBConverter

文本查询

去年通过匹配文本实现的诊断内核。

数据集

- 知识数据集
  - 《诊断学》第一篇“常见症状”所有章节
  - 《内科学》所有章节
- 病例
  - k份随机抽取的省医病例

实验流程

1. 根据当前标记数据构造KG。根据电子版医书将拆分的章节导入ES。
2. 构造以下三类症状集：
   - [x] 随机症状集。记为$\mathbb{R}_k^n$，其中n表示症状集的大小，取值范围为：n，k表示此类症状集的数量。
   - [x] 病例主诉集。记为$\mathbb{C}_k$，其中k的含义与上述相同。此类集合中只包含病例内的主诉。
   - [x] 病例病史集。记为$\mathbb{H}_k$，其中k的含义与上述相同。此类集合中包含病例的主诉与现病史。**其中出现情况为“否”的不包含在$\mathbb{H}_k$内。**
3. 将所有症状集分别输入Converter和文本查询算法，得出每一组输入查到的症状和疾病集合。
4. 统计每一类症状集对应结果的评价指标。

评价指标

设由Converter得到的实体集合（包含疾病和症状）为$\mathbb{G}$，由文本匹配得到的为$\mathbb{D}$。则测量以下指标：

1. 集合大小。简单比较提取知识的覆盖面。
   - $|\mathbb{G}|$
   - $|\mathbb{D}|$

2. 集合相似度。比较提取知识的相似程度。
    - $$J = \frac{|\mathbb{G} \cap \mathbb{D}|}{|\mathbb{G} \cup \mathbb{D}|}$$
3. 集合覆盖率。结合集合相似度比较两个结果的大致包含关系。如果两个集合的J较小，而其中一个集合的F较大，则可以得出集合的包含关系。
    - $$F_{\mathbb{G}} = \frac{|\mathbb{G} \cap \mathbb{D}|}{|\mathbb{G}|}$$
    - $$F_{\mathbb{D}} = \frac{|\mathbb{G} \cap \mathbb{D}|}{|\mathbb{D}|}$$

II. 疾病收敛能力比较

比较对象

Converter

当前诊断方法中的KBConverter

文本查询

去年通过匹配文本实现的诊断内核。

数据集

- 知识数据集
    - 《诊断学》第一篇“常见症状”所有章节
    - 《内科学》所有章节
- 病例
    - k份随机抽取的省医病例

实验流程

1. 根据当前标记数据构造KG。根据电子版医书将拆分的章节导入ES。
2. 构造病例病史集，记为$\mathbb{H}'_k$，其中k的含义与上述相同。此类集合中包含病例的主诉与现病史。**其中出现情况为“否”的也应当包含在$\mathbb{H}'_k$内。**
3. 将所有症状集分别输入Converter和文本查询算法，得出每一组输入查到的症状和疾病集合。
4. 统计每一类症状集对应结果的评价指标。

评价指标

设由Converter得到的疾病集合为$\mathbb{G}$，由文本匹配得到的为$\mathbb{D}$。则测量以下指标：

1. 集合大小。简单比较提取知识的覆盖面。
    - $$|\mathbb{G}|$$

- $$|\mathbb{D}|$$

2. 集合相似度。比较提取知识的相似程度。

   - $$J = \frac{|\mathbb{G} \cap \mathbb{D}|}{|\mathbb{G} \cup \mathbb{D}|}$$

3. 集合覆盖率。结合集合相似度比较两个结果的大致包含关系。如果两个集合的J较小，而其中一个集合的F较大，则可以得出集合的包含关系。

   - $$F_{\mathbb{G}} = \frac{|\mathbb{G} \cap \mathbb{D}|}{|\mathbb{G}|}$$
   - $$F_{\mathbb{D}} = \frac{|\mathbb{G} \cap \mathbb{D}|}{|\mathbb{D}|}$$