# 多语知识库的构建思路的初步设想

## 研究现况和相关技术

有很多种可以划分检索模型的概念，一个个人认为比较不错的划分形式，是可以将跨语言信息检索模型划分为布尔模型、向量空间模型、概率模型、统计语言模型和本体模型。其中统计语言模型和本体模型是研究较多和相对来说更加成熟的部分（因为考虑到实际情况，所以我也推荐我们在这两个方面研究入手）。因为，在此之前有一定本体的相关概念知识，所以先研究本体模型。

目前，大多数的跨语言本体，其构建都是基于WordNet，较为熟悉的实际应用有中国知网。多语言本体是各类跨语言系统的处理桥梁，在检索和消除词语的歧义上尤其重要。

王进提出的Onto-CLIR模型，是利用本体来刻画不同语言中对应的领域知识，解决从查询语言到检索语言之间转化过程中出现的语义损失和曲解等问题（基于本体的跨语言信息检索模型\_王进）。吴丹在Cindor系统基础上，设计了包含双语本体库、索引库、检索库主题三部分构成的中英跨语言信息检索模型（本体在跨语言信息检索中的应用机制研究\_吴丹）。还有很多不错的例子，就不一一列举了，可以在文件夹／多语知识库构建／多语言技术 中查看相关资料。

## 于本系统中的应用

在本系统的实际构建中，所面对的情况又与大多数的情况所不同。前面所说的实际构建中，都是面对普遍的领域，而本系统是专一面对专家信息，所以具有更强的领域性与专一性。所以，在建立多语言本体的时候，我们不好直接使用WordNet也是基于以上原因。所以，考虑是否要在WordNet的基础上，结合我们实际的词语信息情况，生成本系统专用的多语言本体。

个人觉得前期目标不宜太大，以支持中英双语为宜。

在先前了解本体知识的时候，熟悉了RDF，但是在自然语言领域上，RDF并不能满足实际的需求，所以用以替代，拟采用OWL设计多语言本体。在OWL的三个子语言上，现阶段觉得OWL DL最为合适。

使用OWL的话，需要建立起中英双语之间语义的对应关系。现成的一个不错的对应方式是通过爬取知网现有的中英对应概念（基于HowNet的语义检索模型的设计与实现—\_祁智苗）。还有一种不太好实现，但是对于以后真正扩展为多语而不是单单的中英双语而言，更好的设计是通过设置对象属性在多种语言的概念分类体系中建立联系，从而实现概念映射的方法（面向跨语言信息检索的蒙汉语义词典构建\_塔娜）。嗯，个人在了解相关知识的时候，产生了一个想法，就是现在在自然语言语义理解与处理上，比较好的方式是通过上下文来判断实际语义，那么，对于我们这个专家系统，对应相关语义的词语其实都可以在相关论文的摘要中出现，那么对应摘要的中英文，是不是可以很好的匹配中英词语的实际语义。而且优点是，匹配的语义减少了很多不相关的匹配，更符合系统。缺点上，个人觉得会不会不如直接爬取知网的效果好。

还有一个想法，大概是受了之前系统元模型的影响。就是将系统中的专家划分为不同领域，然后针对每一个领域建立一个独自的多语言本体。然后，对针所有的不同领域的多语言本体，在其上建立一个对应各个领域的多语言本体的本体。在用户进行查询的时候，先将其对应到其所在领域的多语言本体，再在此本体上进行查询。示意图如下：

专家知识库

领域多语言本体 n

领域多语言本体 2

领域多语言本体 1

总体多语言本体

## 不足与更多思考

因为现阶段还处于摸索期，也不能一下子拿出行之有效的方案，只能根据他人已有的研究和自己的合理猜测进行思考。

目前，上述的方法，其实我是都不满意的，因为上面的方法都是总结现有的一些研究技术，即都是自上而下。其考虑的出发点都是，需要查询的东西都是已经固定的、大量的、已有的，他们要做的就是我做一个东西去匹配这个现有的东西。但是，对于我们这个系统，我们其实自上而下、自下而上两个方面可以同时考虑，更重要的是，我们需要查询的库，他本身是有本体的，那么我们其实应该有更加合理高效的方法，来做这件事情。就是多语言本体，如何与专家知识库的本体结合。

所有，接下来的研究目标就是这样。

知识图谱小组研究周报

2017.11.26