

知识表示 2024 作业 2

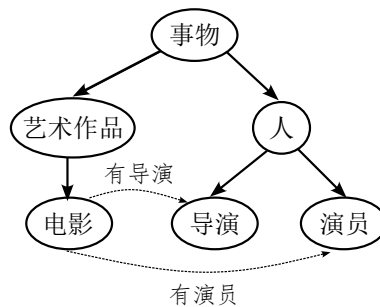
截止时间：6 月 12 日 23:45

提交文件格式：学号 _ 姓名 _hw2.pdf

授课教师：戴洪良 (hongldai@nuaa.edu.cn)

TA：赵春志 (chunzhizhao@nuaa.edu.cn)、杨静雯 (1918773632@qq.com)、吴炜 (3025384698@qq.com)

Q1. 本体



1.1 上图是一个简单的本体，请说明其中包含哪些概念间的关系。(5 分)

1.2 如在以上本体中加入电视剧这一概念，请说明它应与其他哪些概念建立关系，建立什么关系。(5 分)

Q2. 描述逻辑和语义网 (Semantic Web)

2.1 说明描述逻辑中的 TBox 和 ABox 是什么，并举例。(5 分)

2.2 请用 RDF 三元组表示如下内容 (10 分):

“有一个名叫**吴进**的演员，他出演了一部名为**流浪火星**的电影。”

使用以下相关 URI:

实体**吴进**: <http://www.ex.org/res/wj>

实体**流浪火星**: <http://www.ex.org/res/llhx>

类型**演员**: <http://www.ex.org/ontology/actor>

类型**电影**: <http://www.ex.org/ontology/movie>

谓词... 的类型是...: <http://www.ex.org/rel/type>

谓词... 的人名是...: <http://www.ex.org/att/personName>

谓词... 的电影名是...: <http://www.ex.org/att/movieTitle>

谓词... 出演了...: <http://www.ex.org/rel/actIn>

可用完整 RDF 三元组格式，如：

`<http://www.w3.org/EM/contact#me> <http://www.w3.org/pim/contact#fullName> "Eric Miller"` .
或 Turtle 表示。

Q3. 知识图谱概览

3.1 例举 5 个与知识图谱相关的技术，并对它们进行简要说明。(10 分)

Q4. 命名实体识别

4.1 说明如何用基于词典的方法识别文本中的实体，包括识别实体位置和类别。(5 分)

4.2 上述基于词典的方法有什么优缺点？(至少 3 个优点、2 个缺点) (5 分)

4.3 举例说明如何把 NER 转化为分类问题；并设计 NER 模型，说明模型各部分的功能和输入输出，给出模型训练的损失函数。(20 分)

4.4 现需要训练一个基于 BERT 的 NER 模型，但由于项目时间紧迫，只有少量人工标注（每个类别约 50 条）。另外，前期已为该 NER 任务实现了一个基于词典和规则的方法，有一定效果，但一般。现在希望对该已有方法以及少量的人工标注数据结合利用，训练效果更好的模型，请问可以尝试什么样的做法？(5 分)

Q5. 实体分类

5.1 请设计基于 BERT 模型实现普通细粒度实体分类的方法。说明提供给 BERT 的完整文本输入、如何计算各类别对应的分数、训练的损失函数、如何根据结果预测出多个类别标签。(15 分)

5.2 现有一细粒度实体分类任务，采用某种方法为其自动生成了一批训练样本。对每个训练样本 m_i ，记其正标签集合为 Y_i ，负标签集合为 \bar{Y}_i 。且在正标签集合 Y_i 中，已知一些标签是确定正确的，而其他标签则并不那么确定，我们用 $c(y)$ 表示对一个标签的确定性，其值为 0 或 1，如果 $c(y) = 1$ 则 y 是一个确定标签， $c(y) = 0$ 则 y 是一个不确定标签。我们希望在训练时，而赋予不确定标签较小的权重，从而减小它们的影响，请问为达到此目的可以怎样设计训练损失函数？(5 分)

Q6. 事件抽取

6.1 说明事件抽取的两个子任务，以及在 pipeline 方法中实现它们的做法（只需说明实现思路，无需给出具体模型）。(10 分)