README

OpenKG-for-the-Lord-of-the-Rings

A python project for creating an openKG for the Lord of the Rings.

项目流程

本项目计划构建小说《魔戒》（The Lord of the Rings）中的人物、家族、种族的知识图谱。我们先从维基网站上抓取相关的三元组，构成初步的知识图谱。

网页数据获取

> extract.ipynb 中的代码负责抓取网页中的实体
> extract_relations.ipynb中的代码负责抓取实体对应的三元组

文件描述

clean_triplets/character_relation.txt: 通过抓取获得的三元组文件
clean_triplets/character_relation_reduced.txt: 用代码初步清洗后的三元组文件
clean_triplets/character_relation_reduced_manual.txt: 用人工再次清洗后的三元组文件

三元组的抓取

从网站魔戒中文维基魔戒中文维基中抓取需要的三元组信息。
首先从网页中抓取人物列表，网站已经进行了汇总：魔戒人物列表。
通过列表，我们得到了每个人物页面的url。以人物为最初的实体，在各个人物的页面上抓取相关关系。主要从页面的信息栏中提取关系数据，以人物图尔巩为例，其网页上的信息栏如下：

图尔巩	
Turgon	
基本信息	
别名	Turukáno
其他译名	特刚 图尔冈
头衔	刚多林之王 诺多至高王
统治时期	第一纪元116年 - 510年 （刚多林之王） 第一纪元492年 - 510年 （诺多至高王）
出生	双树纪元1300年，生于提力安
死亡	第一纪元510年，死于刚多林
势力信息	
种族	精灵
文化	诺多
家族	芬国盼家族
语言	辛达语（北方方言） 昆雅语
人物关系	
父亲	芬国盼
母亲	阿耐瑞
兄弟姐妹	兄长芬巩 妹妹阿瑞蒂尔 弟弟阿尔巩
配偶	埃兰葳
子嗣	伊缀尔
友善	芬罗德
仇视	魔苟斯

我们可以从中提取到如下三元组：

```
e:图尔巩  r:角色属于种族  e:精灵
e:图尔巩  r:角色属于家族  e:芬国盼家族
e:图尔巩  r:角色的父亲角色  e:芬国盼
e:图尔巩  r:角色的母亲角色  e:阿耐瑞
e:图尔巩  r:角色的兄弟姐妹角色  e:芬巩
e:图尔巩  r:角色的兄弟姐妹角色  e:阿瑞蒂尔
```

```
e:图尔巩　r:角色的兄弟姐妹角色　e:阿尔巩
e:图尔巩　r:角色的配偶角色　e:埃兰葳
e:图尔巩　r:角色的子嗣角色　e:伊缀尔
e:图尔巩　r:角色的友善角色　e:芬罗德
e:图尔巩　r:角色的仇视角色　e:魔苟斯
```

同时我们也抓取了角色的别名、头衔，用于之后的数据清洗。

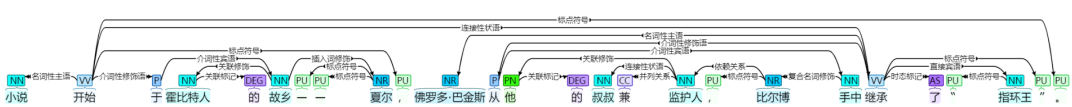
三元组的清洗

- **脚本清洗** 利用web_data_retrieval/clean_triplets/preprocess_relation_txt.py文件，对抓取的三元组web_data_retrieval/character_relation.txt进行初步清洗，得到web_data_retrieval/clean_triplets/character_relation_reduced.txt。例如
  - 过滤有“不明”，“无”等词的三元组
  - 替换“–”为“·”，“（页面不存在）”为“”（空字符串）等
- **实体提取** 利用web_data_retrieval/clean_triplets/extract_candidate_entities.py文件，根据初步清洗的三元组提取出其中的实体，以待后续处理
- **手工实体消减** 手工删除一些无效的实体，例如
  - 有男性后裔
  - 未出世
  - 一位北方辛达女子（其他版本）
- **手工实体替换** 手工替换一些等价的实体，例如
  - 将“克温（第一任）”替换为“克温”
  - 将“辛达”替换为“辛达族”
  - 将“哈烈丝一族”替换为“哈烈丝家族”
  - 将“多阿姆洛斯亲王家族”替换为“多阿姆洛斯家族”等
- **手工清洗** 根据实体替换和实体消减的结果对初步清洗的三元组web_data_retrieval/clean_triplets/character_relation_reduced.txt进行进一步清洗，得到最终的三元组web_data_retrieval/clean_triplets/character_relation_reduced_manual.txt。包括移除无效的三元组，将三元组中的实体进行等价替换等

深度关系抽取

小说文本处理

- **文本过滤与分句** 利用deep_relation_extraction/process_txt/text2sentences.py文件，读入小说文本deep_relation_extraction/process_txt/the_lord_of_the_ring.txt，过滤小说中的无效文本，对小说进行分句，得到分好的句子，并且将其储存在sentences.npy中。其中：
  - 文本过滤使用正则表达式匹配的方式过滤了书名、章节名、无意义的分隔符（如※），避免了后续不必要的存储和计算开销。
  - 分句减小了关系抽取模型每次要处理的文本规模，使模型能够进行高效的关系抽取。
- **分词** 利用deep_relation_extraction/process_txt/segment_words.py文件，读入句子deep_relation_extraction/process_txt/sentences.npy，对每个句子进行分词，并且识别出其中的人名，将分词结果储存在deep_relation_extraction/process_txt/tokens.json中。更具体地，我们使用HanLP的中文分词模型对句子进行分词和词性标注，标为“nr”的词被模型识别为人名。



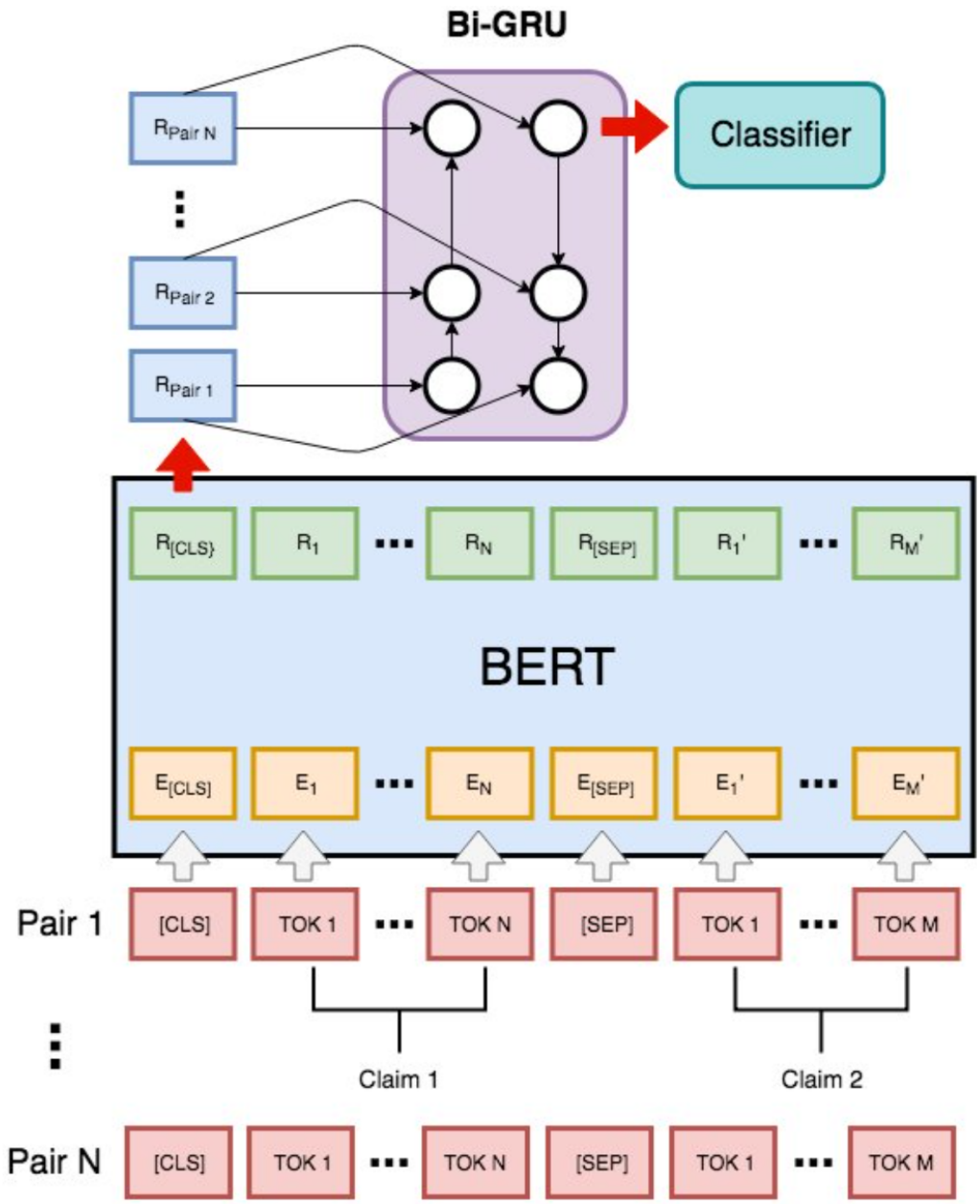
- **训练样本生成** 利用deep_relation_extraction/process_txt/generate_samples.py文件，读入分词结果deep_relation_extraction/process_txt/tokens.json，根据分词结果生成训练样本，写入samples.jsonl。训练样本格式如下：

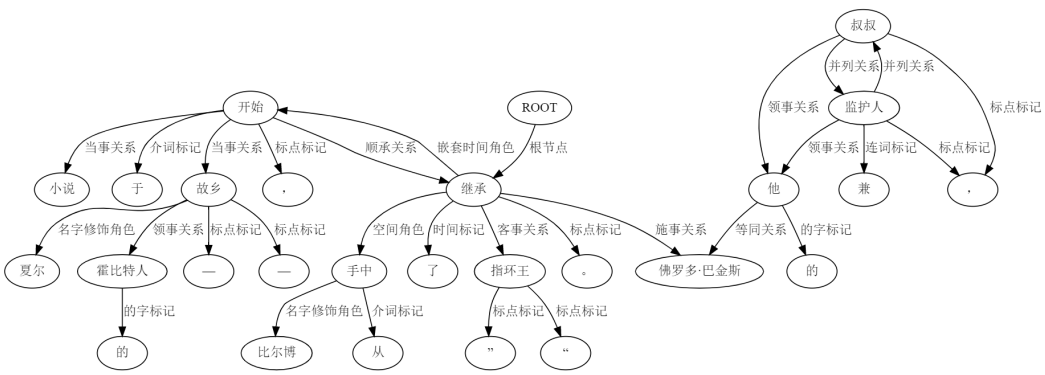
```
{
  "token": list of strings, the result of word segmentation
  "h":{
    "name": string, the subject of the relation,
    "pos": list of int, the position of the subject,
  },
  "t":{
    "name": string, the object of the relation,
    "pos": list of int, the position of the object,
  },
  "label": 0,
  "id": string, the id of the sample
}
```

  其他生成细节如下：
  - 若同一个人名在句子中出现多次，选择其第一次出现的位置作为输入
  - 若一个句子有三个及以上的人名，则每两个人名配对生成一个训练样本，共生成 C_k^2 个训练样本，其中 k 是人名数量

模型训练与关系抽取

由于网页获取的实体关系三元组数量有限，因此我们打算从纯小说文本中抽取出新的关系三元组。在此之前，我们要训练一个关系抽取模型来实现输入一段文本以及两个实体而获取两个实体间的关系。因此我们打算使用基于的中文BERT的实体关系多分类器来实现上述需求，如下图所示为网络架构。

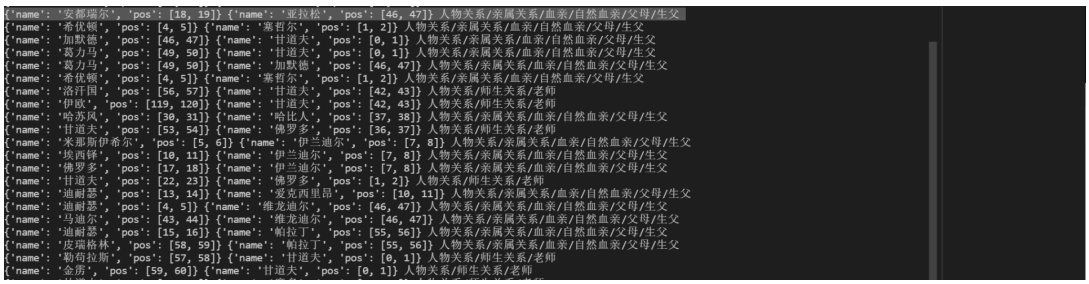




同时，我们采用预训练+微调的方式来训练完整的Bert模型，因此我们决定先采用数据规模较大的CCKS数据集（https://biendata.com/competition/ccks_2019_ipre/）中的数据集来预训练网络模型。通过预训练之后，模型具备了一定泛化的关系抽取能力。然后，我们利用已获取的魔戒关系三元组来微调已经预训练好的bert模型使其能够抽取到的具体相关人物关系。同时我们将BERT Chinese模型修改设置为多分类器，多分类器的个数对应着关系的数量。

训练模型的命令为：

```
python train_and_inference.py
```



通过预训练以及关系抽取，我们一共获得了有效实体关系三元组4854个。

知识图谱构建

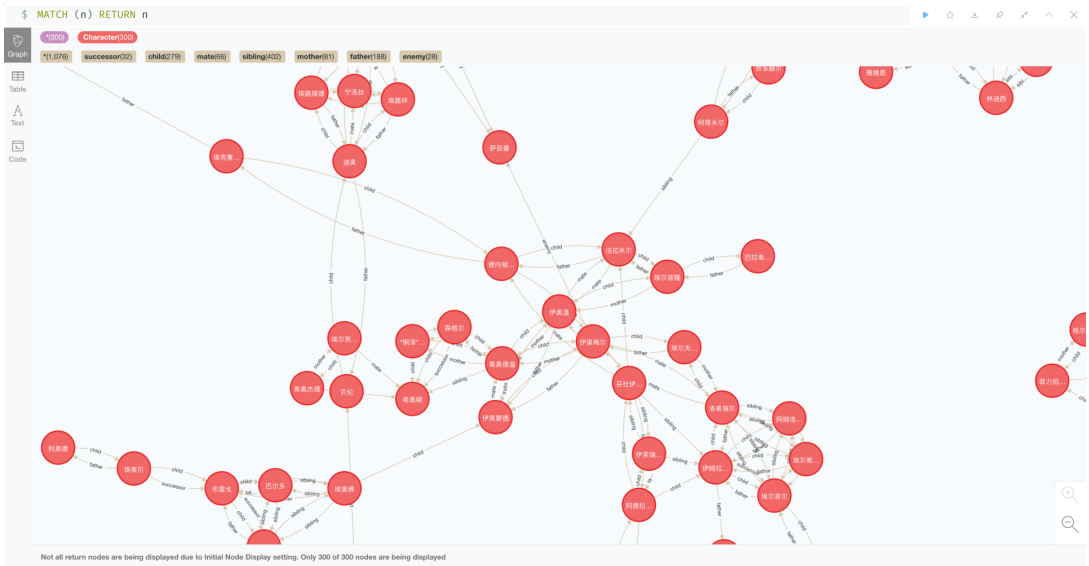
利用construct_kg_file/construct_json.py文件，根据网页上抓取和文本中抽取的三元组构建json文件，用于知识图谱的可视化和自动问答

可视化与问答

总体参考了项目《医疗知识图谱的构建与自动问答》。

知识图谱可视化

build_graph.py负责将知识图谱存入数据库中，我们使用neo4j完成。整体的存储规模如下：



知识图谱实体类型如下：

实体类型	中文含义	实体数量	举例
Culture	种族	66	西尔凡，乌鲁克族
Clan	家族	63	怒锤家族，博芬家族
Role	角色	1075	提理安，塔尔·苏瑞安

知识图谱实体关系如下：

实体关系类型	中文含义	关系数量	举例
belong_to_clan	角色属于家族	281	<陀思妥,属于,博芬家族>
belong_to_culture	角色属于种族	394	<铁尔哈,属于,矮人>
enemy	角色仇视角色	392	<莱格拉斯,仇视,索隆>
mate	角色的配偶角色	506	<美丽安,配偶,辛葛>
father	角色的父亲角色	432	<马迪尔,的父亲,沃隆迪尔>
mother	角色的母亲角色	323	<图林·图伦拔,的母亲,墨玟>
child	角色的子嗣角色	554	<瑟莱茵一世,的子嗣,梭林一世>
successor	角色的继承人角色	398	<欧洛菲尔,的继承人,瑟兰督伊>
siblings	角色的兄弟姐妹角色	2056	<欧玛,的兄弟姐妹,萨尔玛>

自动问答

- question_classifier.py负责将自然语言表述的分体分类成对应的问句类型；
- question_parser.py进行问句解析；
- answer_search.py负责搜索并组织答案语句；
- automatic_qa.py开启自动问答程序。

支持的问答类型如下：

问句类型	中文含义	问句举例
character_culture	角色属于哪个种族	乌妮是哪个种族的?
character_clan	角色属于哪个家族	萨多是哪个家族的?
character_enemy	角色的仇敌	赛洛斯有仇敌吗?
character_mate	角色的配偶	塔尔·瓦泥梅尔德的配偶是?
character_father	角色的父亲	佩烈格的父亲是?
character_mother	角色的母亲	泰理梅克塔的妈妈是?
character_child	角色的孩子	辛拉希恩有孩子吗?
character_successor	角色的继承人	马国尔的继承人是谁?
character_siblings	角色的兄弟姐妹	埃佛拉德·图克有兄弟姐妹吗?

问答结果展示：

****
您好，我是西瓜智能助手，为您解答《魔戒》相关问题～
****
问：雷金纳德·图克的种族是
答：雷金纳德·图克所属的种族有「 霍比特人 」
****
问：林迪西所属的家族
答：林迪西所属的家族是「 埃尔洛斯家族 」
****
问：图尔巩有敌人吗
答：图尔巩的死敌有「 魔苟斯 」
****
问：阿耐瑞的配偶是谁
答：阿耐瑞的配偶是「 芬国盼 」
****
问：塔尔·帕蓝提尔的父亲是
答：塔尔·帕蓝提尔的父亲是「 阿尔·基密佐尔 」
****
问：鲁道夫·博尔杰的妈妈是谁
答：鲁道夫·博尔杰的母亲是「 阿弗丽达·博尔杰 」
****
问：塔尔·卡尔马奇尔的孩子有谁
答：塔尔·卡尔马奇尔的子嗣有「 塔尔·阿尔达明，基密扎加尔 」
****
问：塔尔·阿塔那米尔有继承人吗
答：塔尔·阿塔那米尔的继承者是「 塔尔·安卡理蒙 」
****
问：罗洛·博芬的兄弟姐妹们有谁
答：罗洛·博芬的兄弟姐妹有「 雨果·博芬，乌福·博芬，普莉姆罗丝·绷腰带 」
****