CCKS 2020 评测

任务1：新冠知识图谱构建与问答

子任务1：东南大学认知智能研究所
子任务2：哈尔滨工业大学社会计算与信息检索研究中心
子任务3：浙江大学知识引擎实验室 & 华为云
子任务4：北京大学王选计算机研究所数据管理实验室

新冠开放图谱竞赛任务简介

OpenKG

Task1 **新冠百科知识图谱类型推断：**围绕新冠百科知识图谱构建中的**实体类型推断**（Entity Type Inference）展开，从实体**百科页面**出发，从给定的数据推断相关实体的类型。

Task2 **新冠概念图谱的上下位关系预测：**利用自动挖掘的手段从网络文本中采集的细粒度上位概念词，实体和上位词之间以及不同上位词之间复杂的层次关系，自动准确的构建**细粒度的上下位层次关系**

Task3 **新冠科研抗病毒药物图谱的链接预测：**依据抗病毒药物图谱Schema及知识图谱的实体、实体属性、实体之间的关系，预测新的两个实体的关系，以进行关系预测，如**药物和病毒的靶向作用、蛋白间的交互作用**等。

Task4 **新冠知识图谱问答评测：**面向新型冠状病毒构造了针对**健康、医药、疾病防控**等特定主旨的问答数据，输入中文问题后，期望问答系统从给定知识库中选择若干实体或属性值作为该问题的答案，并且既能处理百科类的**浅层**问题，也能处理具备一定领域知识（如流行疾病等）的**较深层**问题。

相关报名网站：https://www.biendata.com/competition/

CCKS·2020

CCKS·2020

子任务1-背景介绍

- **类型信息**在知识库中具有非常高的价值，实体类型推断的研究一直是领域的热点。然而，大量类型信息以非结构化文本形式呈现于网络页面中，文本处理难度大，抽取结果同时保证高准确度和覆盖率仍然是个极大的挑战。

- 针对实体的通用类型推断，近年来已有若干解决方案，如使用统计机器学习方法及利用外部知识（通向其他数据源的链接或文本信息）等。

子任务1-简介

- 给定实体集合和7个有效类型，对实体的类型进行推断，其中实体包含相关实体和噪音实体（NoneType）。

type.txt:

```
1. 病毒
2. 细菌
3. 疾病
4. 药物
5. 医学专科
6. 检查科目
7. 症状
```

输出样例

```
1. 烟草花叶病毒	病毒
2. 大肠杆菌	细菌
3. 艾滋病	疾病
4. 盐酸西普利嗪	药物
5. 内科	医学专科
6. 太阳	NoneType
```

评测本身**不限制**各参赛队伍使用的模型、算法和技术。

子任务1-评测指标

- 本任务采用精确率（Precision, P）、召回率（Recall, R）、F1值（F1-measure, F1）来评估效果。

- 设 A 为参赛队伍输出文件中的所有实体-类型对的集合，G 为评测方标注文件中的所有实体-类型对的集合，相关计算公式如下：

$$P = \frac{|A \cap G|}{|A|} \qquad R = \frac{|A \cap G|}{|G|} \qquad F1 = \frac{2PR}{P+R}$$

子任务1-评测结果汇总

名次	队名	单位	队员列表	指导老师
第一名	DeepBlueAI	深兰科技（上海）有限公司	罗志鹏、潘春光、张欢	
第二名	清博大数据	清博大数据	夏茂晋、余强、关宇航、任星凯	王欢
第三名	TMAIL	腾讯医疗AI实验室	吴喆、葛岫、吴贤	吴贤
第三名 (并列)	华资AI	广州华资软件技术有限公司	彭本、庄允贵、张士松、周玉晨	周云、翁庄明

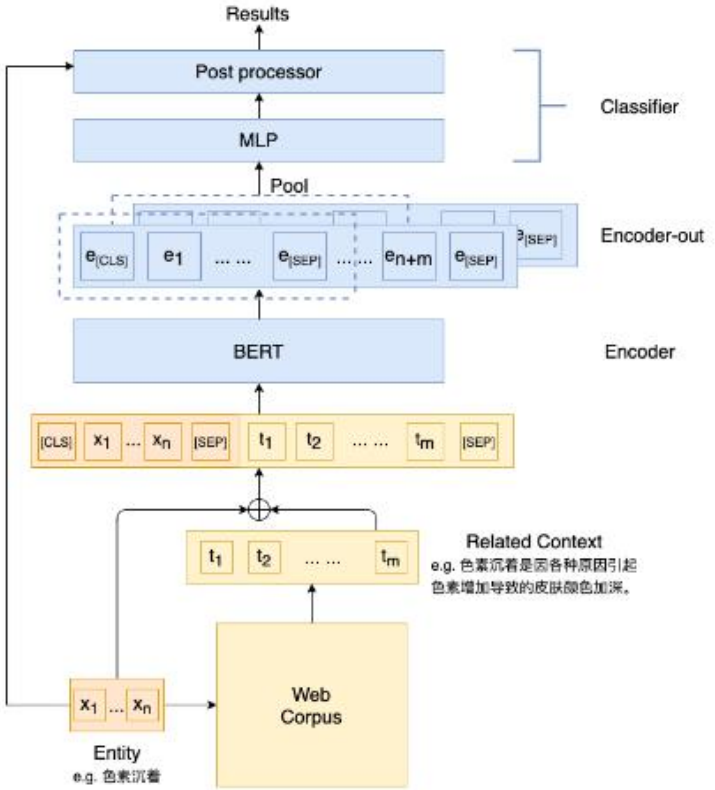
子任务1-评测结果总结

主要步骤

实体信息补充

预训练模型编码

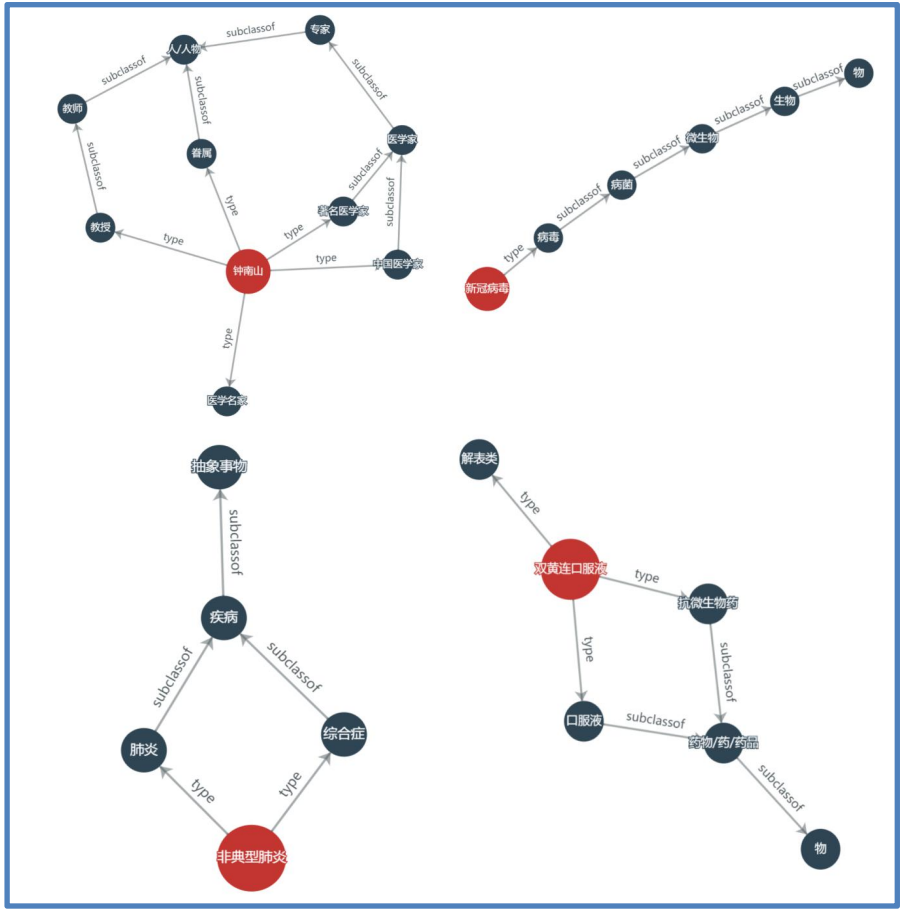
分类器预测



核心整体框架图(by TMAIL)

子任务2-背景介绍

- 在如今的信息化时代，互联网中实体类别多样化，且粒度更细并具有层次，相对于类别有限的传统命名实体，人们开始将目光转向开放域实体挖掘。
- 传统的知识图谱对实体的概念类别体系定义非常有限，如：ACE-2007将实体分为7大类、45小类，Yosef（2013）将实体分为505类，上述将命名实体的类别进行人为的定义。
- 对于互联网中的海量实体很难由人工预先定义出一个完备的类别体系，基于此，可由网络中动态的获得实体的概念类别。



子任务2-简介

- **输入**

  entity.txt：实体列表。
  concept.txt：概念（类型）列表。

- **输出**

  entity_concept.txt：实体-概念之间的类型关系。
  concept_concept.txt：概念-概念之间的上下位关系，前者是后者的子概念。

组织者：

哈尔滨工业大学社会计算与信息检索研究中心

组织人：

张裕舟、佘琪星、张景润、朱文轩、刘铭、秦兵、刘挺

类型预测任务

- **输入样例**

  entity.txt:
  钟南山
  新冠病毒

  concept.txt:
  病毒
  细菌
  疾病
  药物
  医学专科
  检查科目
  症状
  …

- **输出样例**

  entity_concept.txt:
  钟南山　著名医学家
  钟南山　教授
  新冠病毒　病毒

  concept_concept.txt:
  著名医学家　医学家
  医学家　专家
  专家　人物
  教授　教师
  教师　人物
  病毒　微生物
  微生物　生物

概念上下位关系预测任务

子任务2-评测指标

- **任务类型**：无监督任务
- **训练集**：不提供训练集。
- **测试集**：2万实体，1千概念
- **标注方式**：人工标注。
- **初赛/复赛**：40% ：60%

本任务采用**精确率**（Precision，P）、**召回率**（Recall，R）、**F1值**（F1-measure，F1）来评估效果。

对于Entity-Concept（实体-概念）类型关系，设E为测试集中的实体集合（注：测试集是公开的实体集的子集），A_i为选手对E中第i个实体给出的类型集合，G_i为第i个实体的正确类型集合，相关计算公式如下：

$$\text{Macro Precision(EC)} = \frac{1}{|E|}\sum_{i=1}^{|E|} P_i, \qquad P_i = \frac{|A_i \cap G_i|}{|A_i|}$$

$$\text{Macro Recall(EC)} = \frac{1}{|E|}\sum_{i=1}^{|E|} R_i, \qquad R_i = \frac{|A_i \cap G_i|}{|G_i|}$$

$$\text{Averaged F1(EC)} = \frac{1}{|E|}\sum_{i=1}^{|E|}\frac{2P_iR_i}{P_i + R_i}$$

对于Concept-Concept（概念-概念）上下位关系，设A为参赛队伍提交文件中的所有概念-概念对的集合，G为评测方标注文件中的所有概念-概念对的集合，相关计算公式如下：

$$P(CC) = \frac{|A \cap G|}{|A|} \qquad R(CC) = \frac{|A \cap G|}{|G|} \qquad F1(CC) = \frac{2PR}{P+R}$$

最终得分 $F1(\text{overall}) = (\text{Averaged F1(EC)} + F1(CC)) / 2$

子任务2-评测结果汇总

名次	队名	单位	队员列表	指导老师
第一名	UPSIDE-DOWN	国网信通产业集团福建亿榕信息技术有限公司	苏江文、王秋琳、宋立华、闫丽飞、李建华	
第二名	Slaxes_G	华中科技大学	钟嘉伦、孙昊海、刘宇航、韩旭	何琨
第三名	Leerumor	北京美团	李如寐，王思睿，张鸿志	

子任务2-评测结果总结

Part1：基于外部数据的上下位关系种子发现

- 根据表1进行上下位关系识别模式的扩充
- 基于模式的新冠知识图谱概念数据集的上下位关系种子发现
- 基于《大词林》的上下位关系种子补充

Part2：基于 BERT-Attention-Bi-LSTM 的上下位关系分类

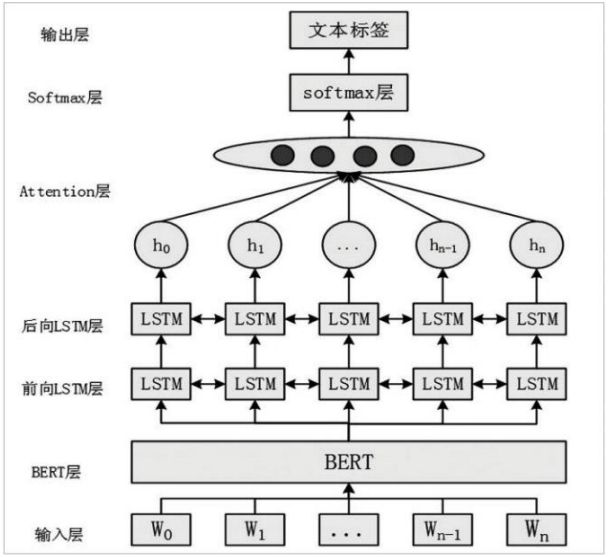


图 1：BERT-Attention-Bi-LSTM 模型

Part3：基于概念后缀词的规则匹配结果扩充及合并

```
if ent[-1] == '病':
    map_entity_concept[ent].append("疾病")
elif ent[-1] == '炎':
    map_entity_concept[ent].append("疾病")

elif ent[-1] == '素':
    map_entity_concept[ent].append("药物")
    map_entity_concept[ent].append("药品")
elif ent[-1] == '菌':
    map_entity_concept[ent].append("微生物")
    map_entity_concept[ent].append("生物")
    map_entity_concept[ent].append("细菌")
elif ent[-1] == '属':
    map_entity_concept[ent].append("科学")
    map_entity_concept[ent].append("自然科学")
    map_entity_concept[ent].append("生物")
elif ent[-3:] == '综合症':
    map_entity_concept[ent].append("综合症")
    map_entity_concept[ent].append("疾病")
elif ent[-1:] == '症':
    map_entity_concept[ent].append("疾病")
elif ent[-1:] == '散':
    map_entity_concept[ent].extend(['中药', '药品', '药物'])
elif ent[-1:] == '膏':
    map_entity_concept[ent].extend(['中药', '药品', '药物'])
```

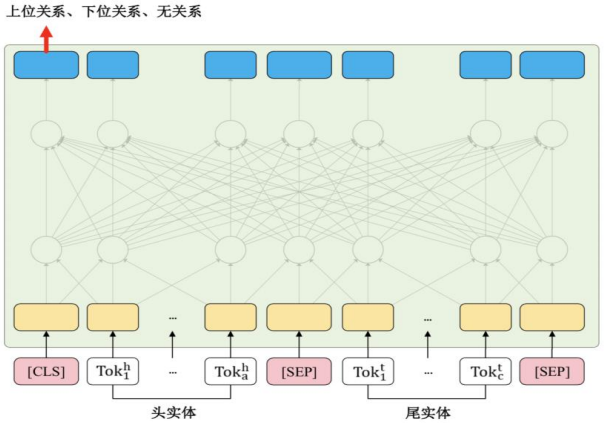
图 2：基于概念后缀词判定的上下位关系发现

子任务2-评测结果总结

Part1：数据集构造

- 依据中文词向量数据库给比赛中的概念词构建词嵌入表示
- 对于不在中文词向量词库中的概念词设计规则
- 根据概念词汇的词向量，使用K-Means进行聚类
- 依据聚类的结果，设计规则提取并构造一部分上下位关系的数据样本

Part2：预训练模型



说明：“头实体”与“尾实体”具有先后顺序，即标签是对称的

Part3：模型融合

- 用不同的随机种子以及不同的取样策略分别训练模型
- 采用模型融合的方式得出最终结果
- 采用人工规则补充提交的数据

表 2. 不同训练方式的实验结果

训练方式	提交得分（F1）
开源数据	0.372
开源数据 +Bert	0.430
开源数据 +Bert+ 融合	0.477
开源数据 +Bert+ 融合 + 规则	**0.482**

子任务2-评测结果总结

Part1：爬虫模块，即数据集构造

Part2：基于Bert的上下位关系识别模

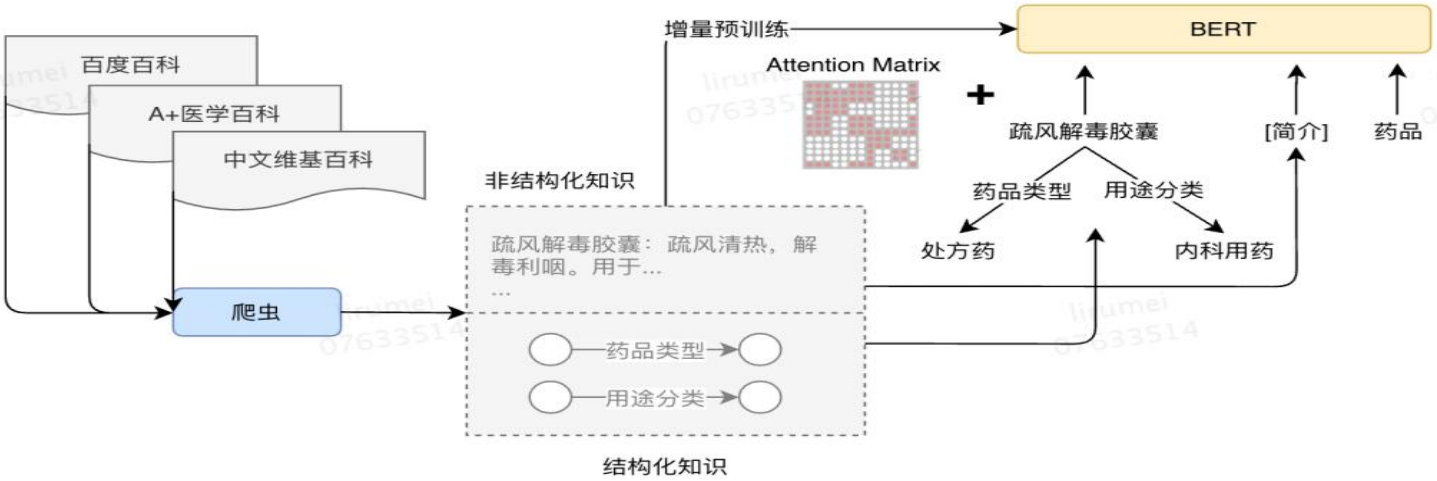


Fig. 1. 基于知识增强的上下位识别系统流程图

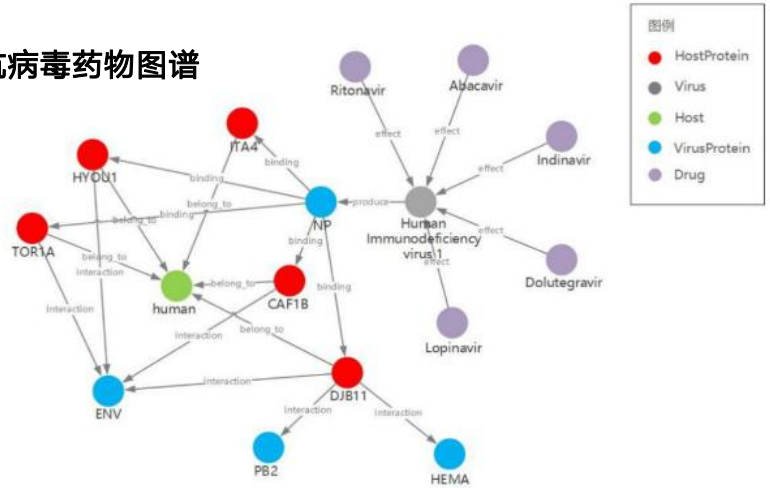
- **获取监督数据的两个主要方法：**
  - 通过百度百科的简介与结构化卡片获取上位词标签
  - 通过匹配获取上位词标签

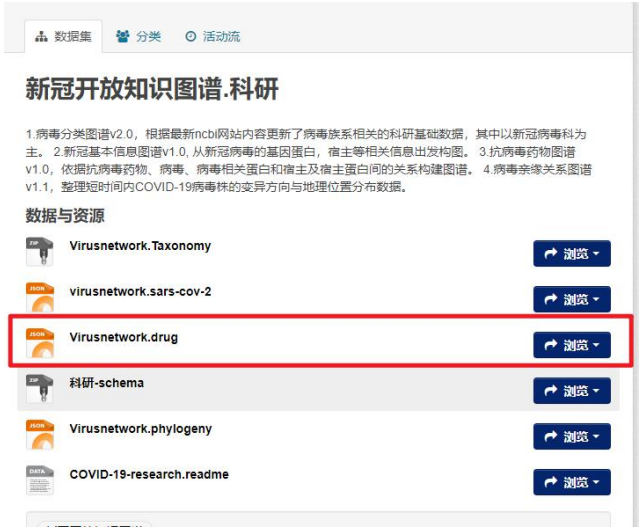
子任务3-背景介绍

病毒的核心
- 核酸分子（RNA或DNA）
- 蛋白质外壳

- 抗病毒药物可以靶向病毒复制的某个环节，起到抵抗病毒保护机体的作用。

抗病毒药物图谱





http://www.openkg.cn/dataset/covid-19-research

- 利用AI算法对抗病毒药物和病毒的靶向作用、病毒蛋白和宿主蛋白的交互作用等进行预测。

子任务3-简介

- **输入**

  schema.json：定义了知识图谱的实体类型（Entity）、实体属性名（Attribution key）和实体间的关系（Relationship）。

  entities.json：**实体列表**，即病毒、药物、宿主蛋白、病毒蛋白等实体。

  attrs.json：**实体属性列表**，即病毒的类型、药物的类型等等。

  relationships.json：**实体与实体之间的关系列表**，即病毒-药物作用，病毒-病毒蛋白关系，病毒蛋白-宿主蛋白的关系等。

  link_prediction.json：包含待评测的头实体或尾实体和关系组成的这样一个实体关系对。

- **输出**

  result.txt：针对每个部分缺失的实体关系对所预测的top10缺失实体队列集合（按可能性从高到低排序）。**选手需要从队列中删除训练集已有链接的实体数据，防止占用并浪费队列位置。**

schema.json：

```
{
    "entity_type": ["virus", "drug", ……],
    "attrs": {
        "virus": {"name": "string", "class": "string", ……},
        "drug": {"name": "string", "indication": "string",  "drug_type": "string",  ……},
        ……
    },
        "relationships": [["drug", "effect", "virus"], ["virus", "produce", "virusProtein"],
    ["HostProtein", "interaction", "virusProtein"], ……],
}
```

entities.json：

```
{
        'virus': ["Human papillomavirus 9", "Human adenovirus 28", "Equine arteritis virus
    Bucyrus", ...],
}
```

relaionships.json

```
{
    "relationships": ["Equine arteritis virus Bucyrus", "produce", "RPOA"]
}
```

link_prediction.json

```
{
    "relationships":
    [
            ("Equine arteritis virus Bucyrus", "produce", "?"),
            ("?", "produce", "RPOA"),
              ……
    ]
}
```

组织者：

浙江大学知识引擎实验室 & 华为云

组织人：

陈卓、王鹏、卢栋才

子任务3-评测指标

- 数据规模

实体：7844

三元组数：40000 <数据集全集>

第一阶段测试数据三元组数：4256

第二阶段测试数据三元组数：5000

- 评价指标：MRR（Measurement Result Recording）

MRR (Mean reciprocal rank，平均倒数排名)：对于一个 query，若第一个正确答案排在第 n 位，则 MRR 得分就是 $1/n$。

$$MMR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

其中，Q 为样本 query 的集合（*link_prediction.json*），|Q|表示 query 的个数。

$$\frac{1}{rank_i} = \begin{cases} \frac{1}{i} & (目标实体在第\,i\,个结果中命中) \\ 0 & (目标实体在所有结果中未命中) \end{cases}$$

每轮最多一次命中

子任务3-评测结果汇总

名次	队名	单位	队员列表	指导老师
第一名	First Blood	武汉大学计算机学院	王维川、谢志文、赵焜松	刘进
第二名	chill	腾讯(深圳)科技有限公司	周煜、魏琪康	何琨
第三名	百万调参大师队	北京交通大学	贾婷、杨玉霞、卢熙	周雪忠，杨扩

子任务4-背景介绍

- 本任务属于中文知识图谱自然语言问答任务，简称 CKBQA （Chinese Knowledge Base Question Answering）。即输入一句中文问题，问答系统从给定知识库中选择若干实体或属性值作为该问题的答案。
- 知识图谱问答在当前互联网信息爆炸、人工智能盛行的时代十分有战略价值和研究意义

典型中文知识图谱

手机语音助手

子任务4-简介

- **输入**

  task1-4_train_2020.txt：**训练集**，包含4000条问题及其对应的SPARQL和答案。

  task1-4_valid_2020.txt：**验证集**，包含1529条问题及其对应的SPARQL和答案。

  task1-4_test_2020.txt：**测试集**，包含1500条自然语言问题。

  pkubase.rar：**知识库文件**，其中知识库源文件pkubase-complete.txt包括超过6000万条关系；schema.txt包含类型和谓词之间的上下位关系信息；mention2ent.txt包含部分实体链接信息。

- **输出**

  result.txt：**测试集问题在知识库中对应的答案**，每行对应一个问题，多个答案之间使用 \t 进行分隔

组织者：

北京大学王选计算机研究所数据管理实验室

组织人：

邹磊、胡森、林殷年

- **输入样例**

  q1:抗疫英雄钟南山院士发表过哪些文章？

  q2:N95 口罩到底哪里好？

  q3:新冠病毒的命名组织是什么时候成立的？

- **输出样例**

  <慢性阻塞性肺疾病在中国>\t<感冒后咳嗽敏感性及气道神经源性炎症改变>\t<连花清瘟胶囊体外抗甲型流感病毒的实验研究>

  "气密性好"\t"过滤效率高"\t"防飞沫"

  "1948 年 4 月 7 日"

输入输出样例

```
q2:《湖上草》是谁的诗?
select ?x where { ?x <主要作品> <湖上草>. }
<柳如是_（明末“秦淮八艳”之一）>

q3:龙卷风的英文名是什么?
select ?x where { <龙卷风_（一种自然天气现象）> <外文名> ?x. }
"Tornado"

q4:新加坡的水域率是多少?
select ?x where { <新加坡> <水域率> ?x. }
"1.444%"

q5:商朝在哪场战役中走向覆灭?
select ?x where { <商朝> <灭亡> ?x. }
<牧野之战>
```

训练集示例

子任务4-评测指标

- **任务类型**：有监督任务
- **数据集**：6000万条关系
- **训练集**：4500条问题
- **测试集**：1500条问题
- **标注方式**：人工标注。

本任务的评价指标包括宏观准确率(Macro Precision)，宏观召回率(Macro RecalI)，Averaged F1值。最终排名以Averaged F1值为基准。设Q为问题集合，A_i为选手对第i个问题给出的答案集合，G_i为第i个问题的标准答案集合，相关计算公式如下：

$$\text{Macro Precision} = \frac{1}{|Q|}\sum_{i=1}^{|Q|} P_i, \quad P_i = \frac{|A_i \cap G_i|}{|A_i|}$$

$$\text{Macro Recall} = \frac{1}{|Q|}\sum_{i=1}^{|Q|} R_i, \quad R_i = \frac{|A_i \cap G_i|}{|G_i|}$$

$$\text{Averaged F1} = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{2P_i R_i}{P_i + R_i}$$

子任务4-评测结果汇总

名次	队名	单位	队员列表	指导老师
第一名	Artemis	KingSoft AI	汪洲、候依宁、汪美玲	李长亮
第二名	see	美团搜索与NLP部	张鸿志、李如寐、王思睿、黄江华	无
第三名	MiQa	小米公司人工智能部	代文、刘岩、刘惠文、吕荣荣	陈帅

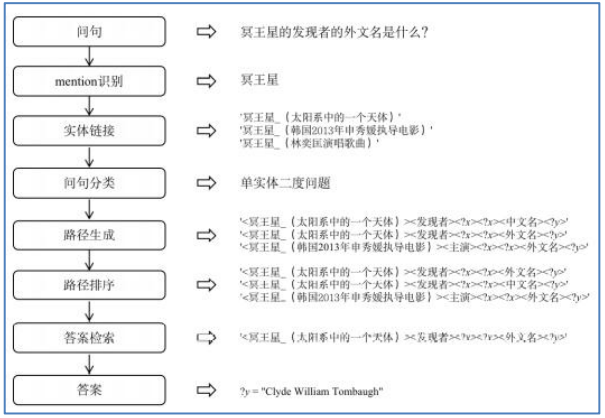
子任务4-评测结果总结

Rank.1

基于特征融合的中文知识库问答方法

汪洲 侯依宁 汪美玲 李长亮 *

AI Lab, KingSoft Corp, Beijing, China
{wangzhou1, houyining, wangmeiling1, lichangliang }@kingsoft.com

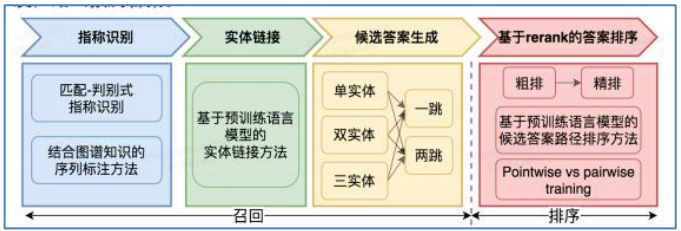


Rank.2

基于预训练语言模型的检索-匹配式知识图谱问答系统

张鸿志，李如寐，王思睿，黄江华

美团, 北京市朝阳区 100020
{zhanghongzhi03,lirumei,wangsirui,huangjianghua}@meituan.com

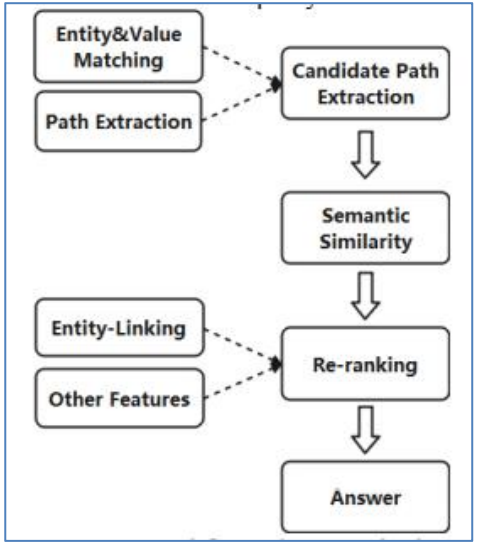


Rank.3

An Integrated Path Formulation Method for Open Domain Question Answering over Knowledge Base

Wen Dai, Huiwen Liu, Yan Liu, Rongrong Lv, Shuai Chen

Xiaomi Corporation, AI
{daiwen, liuhuiwen, liuyan15, lvrongrong, chenshuai3}@xiaomi.com



总结

- **预训练模型**：排名靠前的队伍均采用预训练模型，无论是针对实体的表示还是针对文本的表示，预训练模型均呈现了较好的效果。

- **训练集构建**：无监督任务下利用规则构建训练集。

- **实体增强表示**：词向量表示能力有效，可以通过回标方式获取实体的上下文表示增强实体的表示。

THANK YOU

请扫码关注OpenKG,查看更多行业内容