## 基于本体的金融知识图谱自动化构建技术评测

达观数据 & 同济大学

金融研报是各类金融研究结构对宏观经济、金融、行业、产业链以及公司的研究报告。报告通常是有专业人员撰写，对宏观、行业和公司的数据信息搜集全面、研究深入，质量高，内容可靠。报告内容往往包含产业、经济、金融、政策、社会等多领域的数据与知识，是构建行业知识图谱非常关键的数据来源。同时，所构建出来的知识图谱可以服务于大金融行业、政府、研究机构以及研报所包括的各行业的公司，提供深度分析、辅助决策等支持，价值巨大。

另一方面，由于研报本身所容纳的数据与知识涉及面广泛，专业知识众多，不同的研究结构和专业认识对相同的内容的表达方式也会略有差异。这些特点导致了从研报自动化构建知识图谱困难重重，解决这些问题则能够极大促进自动化构建知识图谱方面的技术进步，具有重大的学术价值。

本评测任务参考TAC KBP中的Cold Start评测任务的方案，围绕金融研报知识图谱的自动化图谱构建所展开。评测从预定义图谱模式（Schema）和少量的种子知识图谱开始，从非结构化的文本数据中构建知识图谱。评测本身不限制各参赛队伍使用的模型、算法和技术。希望各参赛队伍发挥聪明才智，构建各类无监督、弱监督、远程监督、半监督等系统，迭代的实现知识图谱的自动化构建，共同促进知识图谱技术的进步。

### 输入输出

* 评测的输入：
	1. 知识图谱的图谱模式（Schema），即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）。
	2. 种子知识图谱，以三个json的文件提供，即实体列表，实体与实体之间的关系列表，实体属性列表
	3. 从研报抽出来的txt形式的文本
* 评测的输出：
	1. 三个json文件，为所构建出来的知识图谱的内容，包括实体列表、实体与实体之间的关系列表，实体属性列表，格式与输入的种子知识图谱一致。其内容为新增的数据，即构建结果图谱中，剔除了输入的种子知识图谱已有的数据。

### 例子

* 输入：
	+ schema.json
	+ entities.json
	+ attrs.json
	+ relationships.json
	+ yanbao1.txt, yanbao2.txt, ……
* 输出：
	+ new\_entities.json
	所有构建出来的实体列表，不同表示的实体需要做融合
	+ new\_attrs.json
	Schema发布的时候会同时标明属性的类型，部分属性类型会要求归一化，比如时间会要求归一化到YYYY-mm-dd。
	+ new\_relationships.json

其中：

* schema.json

{

 "entity\_type": ["机构", "产品", ……],

 "attrs": {

 "机构": {"地址": "string", "注册日期": "date", "上市日期": "date", ……},

 "产品": {"英文名称": "string", ……},

 ……

 }，

 "relationships": [["机构", "生产", "产品"], ["机构", "供应商", "机构"], ……],

}

* entities.json

{

 "公司": ["特斯拉", "松下", "LG", ...],

}

* attrs.json

{

 "attrs": [["新型冠状病毒", "别名", "2019-nCoV"], ...]

}

* relationships.json

{

 "relationships": ["松下", "生产", "锂离子动力电池"]

}

* yanbao1.txt

宁德时代将为特斯拉供应锂电池两年。根据协议，在有效供货期间内 (2020 年 7 月 1 日—2022 年 6 月 30 日)，特斯拉将根据自身需求以订单形式向宁德时代 采购锂离子动力电池。协议并未对特斯拉的采购数量进行强制约束。随着宁德时代的加入，特斯拉目前的动力电池合作伙伴增加至三家:松下、LG、宁德时代。

* yanbao2.txt

新型冠状病毒（2019-nCoV）抗体检测试剂盒（胶体金法），用于体外定性检测人血清、血浆和全血样本中新型冠状病毒抗体（IgM/IgG），为新冠肺炎的疑似患者、无症状患者、密切接触者，核酸检测阴性者提供快速、便捷的现场检测手段，是对核酸检测的补充或协同。但同时值得注意的是，该产品仅用作对新型冠状病毒核酸检测阴性疑似病例的补充检测指标或疑似病例诊断中与核酸检测协同使用，仅限医疗机构使用。

目前新冠病毒检测方法主要采用核酸检测和免疫学检测，不同方法学在使用的条件、检测灵敏度、检测速度、使用成本等方面存在不同特点。

新型冠状病毒仅含有 RNA 遗传物质，病毒中特异性 RNA 序列是识别该病毒的标志物。核酸检测是新冠确诊的金标准，可靠性好，但其使用条件相对苛刻，过程相对繁琐，检验流程耗时较长，成本相对较高。

免疫学方法省去了病毒核酸提取的步骤，大大降低了检测时间。在免疫学方法中，胶体金法由于操作便捷，不依赖仪器，小巧灵活，成本低廉，适合开发 POCT(即时检测)产品。但鉴于抗原/抗体检测试剂的特点与现状，目前其灵敏度和特异性有限，不能作为新冠肺炎确诊和排除的唯一依据，不适用于一般人群的筛查，仅可作为已有病毒核酸检测试剂的补充。尽管胶体金法产品存在上述局限，其灵活、即时检测、开展成本低的优点，决定了这一产品在医疗机构中可以与核酸法形成良好互补协同效应。

* yanbao3.txt

目前已有的新型冠状病毒检测方法主要有基于荧光 PCR 法的核酸检测，和对病毒感染后人体血清中的免疫球蛋白 IgM/IgG 检测的胶体金免疫层析法（POCT）和酶联免疫吸附测定（ELISA）检测。不同方法学在使用的条件、检测灵敏度、检测速度、使用成本等方面存在不同特点。

* new\_entities.json

{

 "公司": ["宁德时代", ...],

}

* new\_attrs.json

{

 "attrs": [["酶联免疫吸附测定", "别名", "ELISA"], ...]

}

* new\_relationships.json

{

 "relationships": ["宁德时代", "生产", "锂离子动力电池"]

}

### 数据描述

1. 金融研报知识图谱的图谱模式（Schema），即schema.json，包括：
	* 实体类型，10种实体类型，每种实体类型的说明
	* 实体类型的属性，部分实体类型拥有属性，给出属性名（key）及其说明
	* 实体与实体间的关系，约20种关系
2. 金融研报知识图谱的种子知识图谱，覆盖到图谱模式中的所有实体类型、属性和关系。包含三个json文件，即
	* entities.json
	所有的实体数据
	* attrs.json
	所有的属性三元组
	* relationships.json
	所有的关系三元组
3. 训练和验证数据
	* 来自财报内容的200份txt文件，即yanbao000.txt ~ yanbao199.txt
	* 从这200份txt文件中构建好的图谱内容，用以验证效果， 即：
		+ new\_entities.json
		+ new\_attrs.json
		+ new\_relationships.json
4. 测试数据
	* 来自财报内容的1000份txt文件，即yanbao000.txt ~ yanbao999.txt
	* 从这1000份txt文件中构建好的图谱内容，用以评测的最终效果测试， 即：
		+ new\_entities.json
		+ new\_attrs.json
		+ new\_relationships.json

### 评价指标

本次评测任务采用精确率（Precision, P）、召回率（Recall, R）、F1值（F1-measure, F1）来评估构建效果。

对于实体、属性三元组（属性，属性名，属性值）和关系三元组（实体，关系，实体），为三种类型的识别内容。对每一种识别内容，相关的定义如下：

识别内容的精确率 = 识别内容与标注相同的数量 / 识别内容总数量

识别内容的召回率 = 识别内容与标注相同的数量 / 标注内容总数量

识别内容的F1 = (识别内容的精确率 \* 识别内容的召回率)/(识别内容的精确率 + 识别内容的召回率)

最终结果F1定义如下：

F1 = (实体F1 + 2 \* 属性F1 + 2 \* 关系F1) / 5

### 评测交流平台

在训练及验证数据发布的同时会发布**任务交流平台**。要求所有参赛队伍至少有一名成员加入到交流平台中。相关的问题说明、解答和通知会在该平台中发布和交流。同时要求各参赛队伍间针对评测任务的交流也在该平台里进行。为公平起见，所有交流信息对所有参赛队伍公开。

### 任务提交指南

本次评测将依托biendata平台（[https://biendata.com/）](https://biendata.com/%EF%BC%89) 展开，请有意向的参赛队伍关注平台上的竞赛列表。

本次任务将采取刷榜的方式，验证集发布后，允许参赛队伍多次向平台提交结果，排名随时更新。每天每个参赛队伍允许提交2次。

提交的文件格式说明参见前述的输入输出 、 例子 和 数据说明 等章节。更加详细的数据说明和完整的示例会在任务发布在biendata时提供下载。

**外部知识库**

本次评测允许使用公开的外部知识库，私有知识库或者个人整理的知识库需要提前发布，对高质量的知识库建议发布到openkg上。

**预训练模型**
本次评测允许使用公开的各种预训练（pre-trained）模型，比如bert、albert、elmo等。对使用额外数据预训练的私有模型，需要提前公开发布。

**数据标注**
本次评测允许对评测中的非结构化文本自行标注部分数据用以模型训练等。标注的数据不超过所提供文本数量的30%。

**最终提交文件要求**：

对前20的参赛队伍，需要额外提交材料如下：

* 方法的描述文档（非评测论文，评测论文撰写要求见CCKS2020官网）。
* 可运行的自动化构建图谱的程序（包括模型）以及运行该程序相关的说明，用以验证构建自动化构建知识图谱的成绩与榜单上是否一致。
* 文档中需要说明所使用的外部知识库、预训练模型和数据标注的情况。

详细的提交要求、提交方法会在最终测试数据发布（7月10日）的同时发布。

### Baseline

Baseline会在准备好baseline的第一时间内给出，预计发布时间不会晚于评测任务正式发布后的两周的时间。

### 奖励

* 任务第一名5k，第二名3k，第三名2k，技术创新奖5k。
* 排名前10的队伍将获授精美参赛奖牌、证书。
* 比赛排名前20的选手将获得达观数据提供的全职（面向在职）和实习（面向在校生）的VIP通道，通过面试优先录用。

### 时间安排

* 评测任务发布：3月20日
* 报名时间：3月20日 — 7月10日
* 训练及验证数据发布：4月3日
* 公开排名榜(A)：4月3日-7月10日
* 测试数据发布：7月10日
* 最终排名榜(B)：7月10日-7月20日
* 最终测试结果：7月20日
* 评测论文提交：8月5日
* CCKS会议日期(评测报告及颁奖)：8月15日—18日

### 组织者

王文广 达观数据 wangwenguang@datagrand.com
王昊奋 同济大学 carter.whfcarter@gmail.com
郭翠翠 达观数据 guocuicui@datagrand.com