

零标注冷启动构建领域图谱

秦海龙、Ben Lin、彭滢、穆啸天、李维
上海弘玑信息技术有限公司



简介

知识图谱是积累和存储领域场景中实体及关系等知识情报的常见方式。在具有原生文本大数据的领域，自然语言抽取挖掘技术支持自动构建并不断更新动态的知识图谱。在无需依赖标注数据的情况下，通过新词发现和语言结构基础上的解析抽取，跨领域多层深度解析器（Multi-Level Deep Parser）可以快速领域化，赋能领域知识图谱的构建和应用。

本项目介绍跨领域解析器领域化的具体步骤和创新，落地到刑事判决书文本语料库，展示了刑事案件场景的零标注冷启动知识图谱构建的方法。所谓零标注并非没有明确规定的抽取目标。抽取目标是领域用户提供的字段需求描述，每个字段通常提供一两个实例作为示意。开发集是未经任何标注的原数据，开发人员借助语言解析结构的支持，对照客户需求快速编码具有结构泛化能力的领域化抽取规则。60 多目标字段多抽取规则，初版开发上线时间仅仅 10 个人天。抽取系统的调试稳定直到赋能图谱的自动构建，再加 10 个人天。这样的有限资源投入所获得的冷启动快速领域化效果，得益于强大的跨领域解析引擎和平台，以及大数据领域化词典的自动习得。

图谱涵盖 60 多个字段的关键信息，包括百万级的四个主要入口实体：被告人、判决机构、罪名、判决地。除了实体描述字段的丰富信息外，联系图谱实体的关键实体关系包含判决（连接判决机构与被告人）、共犯（连接多个同案被告人）、罪名（连接罪名与被告人）、判决机构地所属省份（连接省份与其所辖机构）。系统清晰地展现了刑事判决场景中重要的行为主体和行为关系，为刑事判决场景中自然语言处理相关的查询、推理、自动问答等任务提供了优质的垂直领域知识库。本系统的创新之处在于开拓了一种有效的冷启动构建领域知识图谱的算法流程，克服了学习系统依赖大量标注数据的知识瓶颈。

结论

领域知识图谱自动构建的最大挑战是缺乏大规模标注数据可供机器学习或深度神经网络训练。这种知识瓶颈是阻碍领域知识图谱规模化和动态更新的根本问题。领域场景一旦变化，情报类型和结构随之改变，一切必须重新来过。垂直领域的现场，用户往往只能提供简单的样例表达需求。组织大量人力标注每一个领域场景的数据显然不是可持续的规模化领域转移方案。利用规则进行领域信息抽取也有自己的知识瓶颈。抽取规则的编制虽然不必依赖大量的标注数据，但在缺乏语言结构深层解析的情况下，也是非常耗时耗力的工程。本项目的最大特点就是探索出一种可以复制和规模化的领域移植方法，利用跨领域的语言深层解析引擎，借助词典的领域化习得，快速赋能领域化的抽取挖掘，从而有效克服了领域知识图谱构建的知识瓶颈。

表格名称	表格字段
行为人	行为人 ID、姓名、类型、出生年月日、地址、民族、学历、工作
判决机构	判决机构 ID、判决机构名称、类型
罪名	罪名 ID、罪名名称
犯罪关系	行为人 ID、罪名 ID
共犯关系	行为人 ID、共犯 ID
判决关系	判决机构 ID、行为人 ID
判决机构归属地	判决机构 ID、判决机构所在地 ID

刑事判决书知识图谱实体与关系

系统描述

本项目采用了一套基于计算语言学的多层次语言解析器对判决书进行解析。在 60 层解析的结构支持下，建立了基于结构的三级刑事判决书抽取模块计 6 层，提取出判决书中 64 个相关字段和关系。

系统对 248,112 万判决书数据进行解析和抽取，平均每小时可以解析 191.4 个判决书，并对每个判决书的内容进行篇章信息融合。之后得到了一个相对完整的刑事判决知识图谱，该图谱包含行为人、判决机构、罪名、判决机构所在地四种关键实体共计 41.9 万个节点，包含判决、共犯、侵犯罪名、判决机构所属地四个实体关系共计 91.8 万链接。

跨领域解析引擎面对不同领域文本往往出现算法性能大幅下降的问题，从而难以有效支持领域的信息抽取。因此，解析领域化是“解析支持抽取”架构成功的关键。本项目的跨领域解析器是语言学家编制，不依赖标注数据。多个项目的领域化实践表明，解析器对于不同领域文本的适应性良好，算法性能下降的压倒性主因是领域词汇，而不是结构规则。只要顺利对接领域词汇到系统跨领域核心词汇及其本体知识库，性能可以接近专家的解析水平（实验精确率和召回率均达 90% 以上）。

多层次解析引擎中有三个领域化的自然接口，使得三层领域抽取模块可以无缝对接核心引擎的管道系统，N 元组规则放在查询典模块后，适合做非常具体的表层规则。解析引擎支持的 N 元组规则可以利用引擎内部的特征系统（包括部分《知网》的本体特征知识库），可做基于知识的泛化。浅层抽取规则放在浅层解析后，可以得到短语组块的结构支持，特别是专名（NE）和数据单位（DE，如时间）。第三个接口涉及深层解析后的子图抽取模块。有逻辑语义结构的支持，一条子图规则可以达到语言表层几十乃至上百个 N 元组模式的效力。

语言解析器对刑事判决书的解析抽取结果以 Json 字符串的形式返回，其中属性（Pnorm）字段为抽取标签，抽出的短语是属性值（Text）字段。系统会根据这些字段生成一个关于刑事判决书的 Record。其中每个 Record 可能包含上部分提到的 64 种信息类型。

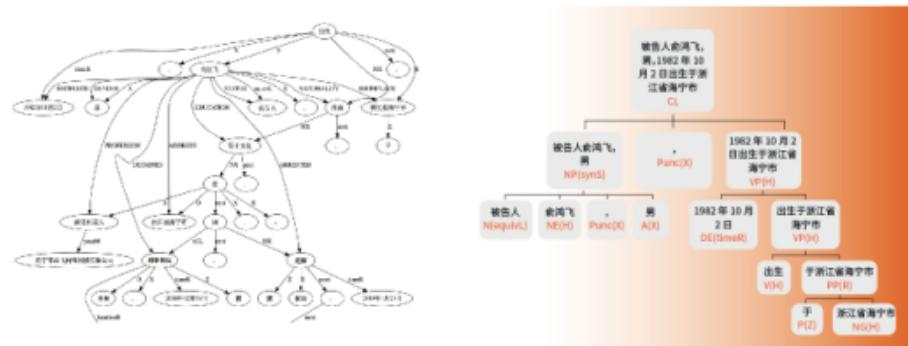
在解析引擎质量实验中，系统在测试集中随机选择了 4 份判决书，引擎识别出了 2,199 个依存关系。最终得到精确率（Precision）96.1%，召回率（Recall）89.6%，F 值（F1-Score）92.7%。

在关键字段抽取精确率实验中，系统在测试集中随机选择了 10 份判决书，对抽取结果做人工检验。对于缺乏标注的场景，精确率测试只需要人工查验系统抽取结果，无需通读原文，测试比较快速。在抽取出的 3,526 个关键字段中，查验有 3,405 个是正确的，精确率为 96.6%。

本项目属于零标注冷启动的图谱构建系统，召回实验由于缺乏标注数据难以扩大规模。抽取召回率实验中，测试员针对用户要求的第一批 34 个核心字段，每个字段对随机测试集中的 5 份判决书通读以求字段出现与否，逐字段查证 34 次，测得召回率为 96.7%，在共计出现的 1,346 个核心字段中漏抓了 44 个。

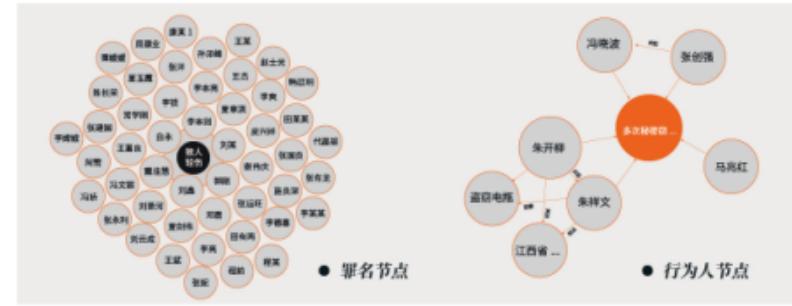
在信息融合阶段，系统还构建了被告人的共犯关系。在处理同一个判决书时，系统会将每个被告人单独作为一个实体，并分别抽取不同被告人的身份资料，并建立这些被告人之间的共犯关系。

本系统采用 Neo4j 作为知识图谱的展示工具，其中包括四种实体（行为人、罪名、判决机构、判决机构所在地）为中心的视角，全面的展示了实体之间的重要关系。



● 内部逻辑语义解析：依存结构图图示(部分)

● 内部句法解析：短语结构树图示(部分)



● 罪名节点

● 行为人节点