软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(10):2931−2947 [doi: 10.13328/j.cnki.jos.005552]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种准确而高效的领域知识图谱构建方法*

杨玉基, 许 斌, 胡家威, 仝美涵, 张 鹏, 郑 莉

(清华大学 计算机科学与技术系 知识工程实验室,北京 100084)

通讯作者: 杨玉基, E-mail: yangyujiyyj@gmail.com

摘 要: 作为语义网的数据支撑,知识图谱在知识问答、语义搜索等领域起着至关重要的作用,一直以来也是研究领域和工程领域的一个热点问题,但是,构建一个质量较高、规模较大的知识图谱往往需要花费巨大的人力和时间成本.如何平衡准确率和效率、快速地构建出一个高质量的领域知识图谱,是知识工程领域的一个重要挑战.对领域知识图谱构建方法进行了系统研究,提出了一种准确、高效的领域知识图谱构建方法——“四步法”,将该方法应用到中国基础教育九门学科知识图谱的构建中,在较短时间内构建出了准确率较高的学科知识图谱,证明了该方法构建领域知识图谱的有效性.以地理学科知识图谱为例,使用“四步法”共得到 67 万个实例、1 421 万条三元组,其中,标注数据的学科知识覆盖率和知识准确率均在 99%以上.

关键词: 语义网;知识图谱;本体;语义标注;实体集扩充;关系抽取

中图法分类号: TP18

中文引用格式: 杨玉基,许斌,胡家威,仝美涵,张鹏,郑莉.一种准确而高效的领域知识图谱构建方法.软件学报,2018,29(10):2931−2947. http://www.jos.org.cn/1000-9825/5552.htm

英文引用格式: Yang YJ, Xu B, Hu JW, Tong MH, Zhang P, Zheng L. Accurate and efficient method for constructing domain knowledge graph. Ruan Jian Xue Bao/Journal of Software, 2018,29(10):2931−2947 (in Chinese). http://www.jos.org.cn/1000-9825/5552.htm

Accurate and Efficient Method for Constructing Domain Knowledge Graph

YANG Yu-Ji, XU Bin, HU Jia-Wei, TONG Mei-Han, ZHANG Peng, ZHENG Li

(Knowledge Engineering Group, Department of Computer and Sciences, Tsinghua University, Beijing 100084, China)

Abstract: In supporting semantic Web, knowledge graphs have played a vital role in many areas such as knowledge QA and semantic search. Therefore, they have become a hot topic in the field of research and engineering. However, it is often costly to build a large-scale knowledge graph with high accuracy. How to balance the accuracy and efficiency, and quickly build a high-quality domain knowledge graph, is a big challenge in the field of knowledge engineering. This paper engages a systematic study on the construction of domain knowledge graphs, and puts forward an accurate and efficient method of constructing domain knowledge graphs as “four-steps”. This method has been applied to the construction of knowledge graphs of nine subjects in the k12 education of China, and the nine subject knowledge graphs have been developed with high accuracy, which demonstrates that the new method is effective. For example, the geographical knowledge graph, which is constructed using the “four-steps” method, has 670 thousand instances and 14.21 million triples. And as part of it, the annotation data’s knowledge coverage and knowledge accuracy are both above 99%.

Key words: semantic Web; knowledge graph; ontology; semantic annotation; entity set expansion; relation extraction

1998 年,互联网的创始人 Berners-Lee 最先提出了“语义网(semantic Web)”的概念[1],其核心思想是:在网页

* 基金项目: 国家高技术研究发展计划(863)(2015AA015401)
Foundation item: National High Technology Research and Development Plan of China (2015AA015401)
本文由“本体工程与知识图谱”专题特约编辑漆桂林教授推荐.
收稿时间: 2017-07-22; 修改时间: 2017-11-08; 采用时间: 2018-01-24; jos 在线出版时间: 2018-02-08
CNKI 网络优先出版: 2018-02-08 11:55:49, http://kns.cnki.net/kcms/detail/11.2560.TP.20180208.1155.008.html

数据中添加能够被计算机所理解的语义信息,从而提升机器的理解能力.作为语义网的数据支撑,知识图谱成为研究领域和工程领域的热点问题.知识图谱是一个巨大的知识网络,网络中的节点表示实体,节点之间的边表示实体和实体之间的关系,实体包含概念和实例两种,每个实体还有很多(属性-值)对来描述实体的内在特性.例如:“中国”和“俄罗斯”是实例,也是“国家”,而“国家”是一个概念;“中国”和“俄罗斯”有着“毗邻”的关系,“中国”和“俄罗斯”也都有各自的“人口数量”“面积”等属性.上述知识都可以通过(主语-谓语-宾语)的形式来描述,这种形式被称为三元组,也被称为事实.以上事实在知识图谱中就可以表示为:

(中国-类型-国家)
(俄罗斯-类型-国家)
(中国-毗邻-俄罗斯)
(中国-面积-9 634 057 平方公里)
(中国-人口数量-13.8 亿(2016 年))
(俄罗斯-面积-17 098 242 平方公里)
(俄罗斯-人口数量-1.4 亿(2016 年))

知识图谱可以分为不限领域的知识图谱(通用知识图谱)和限定领域的知识图谱(领域知识图谱)两种.通用知识图谱有很多,包括研究领域的 DBpedia[2]、YAGO[3]、Freebase[4]等和工程领域的 Google 的 Knowledge Graph[5]、百度的“知心”(http://baike.baidu.com/view/10972128.htm)、搜狗的“知立方”(http://baike.baidu.com/view/9645207.htm)等.领域知识图谱也有很多被构建出来,例如地理信息领域知识图谱 Geonames(http://www.geonames.org/ontology)、“天眼查”(https://www.tianyancha.com)的企业领域知识图谱等.

知识图谱的构建往往需要付出很大的代价.由于当前的自然语言处理方法还不够完善,完全自动化的构建方式难以得到较为准确的知识图谱,例如,DBPedia、YAGO 等都存在有较多错误;而完全人工构建的方法虽然保证了准确性,但却需要花费巨大的人力和时间成本,完全人工构建较大规模的知识图谱几乎不可能.因此,如何协调准确率和效率、平衡自动化方法和人工参与,以最高效的方式构建出最准确的知识图谱,是目前构建知识图谱需要解决的一大难题.

本文主要基于以上难题提出了系统性的解决办法——“四步法”,4 个步骤分别是:

步骤 1:领域本体构建;

步骤 2:众包半自动语义标注;

步骤 3:外源数据补全;

步骤 4:信息抽取.

本体构建是指构建出知识图谱的本体结构,本体结构可以理解为知识图谱的框架.众包半自动语义标注指的是将文本页面众包给多个标注者,根据步骤 1 构建好的本体,利用语义标注工具标注得到高质量的标注数据.外源数据补全是指将其他来源的结构化程度较好的数据按照本体结构处理后,与标注数据整合在一起.而信息抽取则是针对知识图谱中较为稀疏的实体或者关系,从文本中进行大规模的抽取和补充.步骤 1、步骤 2 是知识图谱的骨架部分,是基础,也是核心.两个步骤相互迭代,本体构建指导标注,标注中遇到的新的情况又可以反向改进本体结构.步骤 1、步骤 2 保证了知识图谱的准确性.步骤 3、步骤 4 是知识图谱的血肉部分.在步骤 1、步骤 2 得到的高质量标注数据的基础上进行有针对性的、可控的扩充和补全,保证了知识图谱的覆盖率和构建的高效性.步骤 3、步骤 4 也是相互迭代的关系,步骤 4 可以利用步骤 3 中得到的关系和实体从文本中进行信息抽取,步骤 3 也可以利用步骤 4 中抽取出的新的实体和关系,将其他来源的结构化数据中的相关知识补充到知识图谱中.

以上 4 个步骤能够充分利用领域内高质量的专业资料和海量的互联网数据,高效地构建出准确率较高的实际可用的领域知识图谱.我们还使用此方法构建出了面向基础教育的地理学科的知识图谱,实例数量 67 万、三元组数量 1 421 万,其中,标注数据的知识覆盖率和知识准确率达 99%以上.

本文的主要贡献如下.

- 提出了一种准确、高效地构建领域知识图谱的方法——“四步法”,并用“四步法”构建出了一个面向基础教育的高质量的地理学科知识图谱,验证了“四步法”的有效性;
- 构建的面向基础教育的高质量的地理学科知识图谱,为基于地理学科知识图谱的应用系统(知识问答与高考答题)提供语义数据支撑;
- 研究实现的众包半自动语义标注工具可以在标注三元组的时候很好地兼顾质量和效率,同时可以用于完善本体结构.

本文第1节是相关领域的研究综述.第2节是地理学科知识图谱构建,详细地介绍用“四步法”构建地理学科知识图谱的整个过程.第3节是实验,介绍众包半自动语义标注、实体集扩充和关系抽取的相关实验和效果,以及地理学科知识图谱的数量统计信息.第4节是结论.

1 相关研究综述

本节对知识图谱构建过程中的主要挑战进行介绍,包括本体构建、语义标注和信息抽取这3个部分.

1.1 本体构建

1993年,Gruber[6]将本体定义为“一种概念化的精确的规格说明”.1998年,Studer[7]进一步扩充了本体的概念,将其定义为“共享概念模型的明确形式化规范说明”.简而言之,本体主要是用来描述某个领域内的概念和概念之间的关系,使得它们在共享的范围内具有大家共同认可的、明确的、唯一的定义.所以,本体具有共享化、明确化、概念化和形式化这4个基本特征.

本体构建的过程相当繁琐,而且构建过程往往因各自领域和具体工程的不同而有所不同[8].但是大家公认的是,在领域本体的构建过程中,需要相关领域专家的协作与指导[9−11].一般而言,本体构建通常有人工、自动和半自动这3种构建方法.

- 人工构建本体的方法通常是由大量的领域专家相互协作完成,例如WordNet[12].常见的人工构建本体的方法主要有Skeletal法[13](又称骨架法)、TOVE法[14]、SENSUS法[15]、Methontology法[16]、Ontology Development 101法[17](又称七步法)等;
- 自动构建本体通常也称为本体学习,其目标在于利用知识获取技术、机器学习技术以及统计技术等从数据资源中自动地获取本体知识,从而降低本体构建的成本.杜小勇[18]根据数据源的结构化程度(结构化、半结构化、非结构化)以及本体学习对象的层次(概念、关系、公理)将本体学习问题划分为9类子问题,并详细分析了这9类子问题的研究进展.在此之上,还进一步介绍和比较了现有的本体学习工具.在自动构建本体方面,目前还极少有方法能够得到覆盖率和准确率都表现良好的本体[19];
- 半自动构建本体介于人工构建本体和自动构建本体之间,对于大多数领域而言,完全自动化地构建本体是难以实现的,所以在自动构建本体的过程中,通常还需要在用户的指导下进行.

本文采用的是半自动构建本体的方法,使用统计方法和无监督方法得到本体知识,结合其他知识图谱的本体知识,在专家的指导下构建出了本体,并在众包半自动语义标注过程中进行了完善.

1.2 语义标注

语义标注是指对原始数据做标记,使其包含一定的语义信息,这样不仅人可以理解,而且机器也能够理解.

语义标注的研究主要包括利用本体技术和自然语言处理等技术来进行语义标注的算法研究和应用研究[20−28].

根据语义标注结果的存储方式,语义标注可以分为两类[25]:嵌入式存储和独立存储.嵌入式存储方式是指将标注结果嵌入在原始网页中,标注格式可以是JSON-LD[29]、MicroData、RDFa[30]等,例如Google的结构化数据标记辅助工具[31];独立存储方式是指将标注结果保存在外部存储中,可以保存到文件中,也可以保存到数据库中,例如开源语义标注工具Pundit[32,33],它可以对任何网页进行标注,标注结果将保存在标注系统后端的数据库中.

根据语义标注的自动化程度,语义标注可以分为 3 类[24]:手工标注、半自动标注和自动标注.手工标注是指标注人员直接将语义数据写入到标注文档中,最典型的是 Semantic Wiki,即 Wiki 的语义版本.它的实现方式是在编辑 Wiki 页面时插入一些语义数据,使得 Wiki 系统能够解析这些语义数据,从而提供更加便捷的浏览和更加智能的检索.其他的手动标注工具还有 SHOE Knowledge[34]、OntoMat Annotizer[35]等.半自动标注是由标注人员指定网页或者网页中的文本片段,然后由标注人员选择合适的本体概念(或属性)或者由系统自动显示可选的本体概念(或属性),最后生成并保存语义标注结果.典型代表就是由 W3C 主导的 Annotea[36]项目,它是最早的基于 RDF 的语义标注项目,该项目实现了一个半自动语义标注工具 Amaya.标注人员可以通过添加标注模板来完成半自动标注,但是由于标注工作必须由人工在客户端软件中完成,因此 Annotea 并不适合大规模的网页语义标注.其他的半自动标注工具还有 SMORE[37]、Pundit 等.自动标注是指标注工具可以按照预定的规则自动产生并保存语义标注信息.典型代表有 AeroDAML[38],它把常见的概念和关系映射到 DAML+OIL 本体中的类和属性,并采用自然语言处理和信息抽取方法从网页文档中自动生成 DAML 标记的知识.其他的自动标注工具还有 MnM[39]等.研究表明,采用自动标注方式虽然提高了标注速度,但其标注质量很难得到保证[21].

本文采用的是基于 Pundit 开发的众包半自动语义标注工具,可以满足众包标注过程中的标注审核、标注溯源、共指消解、数据存储等需要,极大地提升了众包标注的效率.

1.3 信息抽取

信息抽取包括实体抽取、关系抽取和属性抽取.

实体抽取也称命名实体识别,是从文本中自动识别出命名实体.与之相似的任务是实体集扩充,该任务指的是根据种子实体集,从文本中抽取出相同类别的新实体.对于实体集扩充,Bootstrapping 方法是最直接的想法[40],此方法根据种子实体从文本中抽取出特征模板,然后利用这些模板从文本中抽取出新的实体,再根据新实体从文本中抽取新的特征模板.反复迭代此过程,便可以抽取出目标概念下大量的新实体.这种方法最大的问题是语义漂移[41],即:随着迭代次数的增加,扩充的新实体会逐渐偏移原来的类别.

关系抽取指的是从文本中抽取出实体和实体之间的关系,这样才能将零散的实体联系起来.关系抽取的算法可以分为基于规则的方法和基于机器学习的方法.基于规则的方法需要人工制定较多规则且难以全面;基于机器学习的方法又可以分为有监督、半监督和无监督这 3 类:有监督的方法需要大量质量较好的标注数据,半监督的方法需要少量标注数据,无监督的方法不需要标注数据.使用有监督的方法进行关系抽取,可以看作是多分类问题,预先将每个关系定义为一个类别,然后将句子中实体之间的关系划分到预先定义的类别中;较多采用的半监督算法有 Bootstrapping 算法、协同训练算法和标注传播算法[42],其中,Bootstrapping 方法中的远程监督方法目前最受学界关注,该方法首先将每种关系的少量三元组作为种子集,然后回标出同时包含种子集中三元组的两个实体的句子作为表征这一关系的训练数据,再从文本中找出符合这一关系的其他句子,这些句子中的实体和实体之间则很有可能也满足这一关系.该方法最大的问题是训练数据噪声,即:包含两个实体的句子的语义关系可能会有很多种,但是回标的时候所有句子被当成了一种语义关系.本文对有监督、半监督和无监督的方法都进行了相关实验.

属性抽取则是从文本中抽取出实体的属性信息,例如实体“中国”的“面积”、“人口数量”等属性.由于可以将实体的属性视为实体和属性间的一种名词性关系,因此也可以将属性抽取问题视为关系抽取问题[43],故以下属性抽取也归为关系抽取的范畴,不再分开表述.

2 地理学科知识图谱构建

本节将以面向基础教育的地理学科知识图谱的构建为例,详细介绍使用“四步法”构建领域知识图谱的过程,构建框架如图 1 所示.其中,4 个步骤的具体内容如下.

(1) 领域本体构建:基于地理学科权威的教材教辅资料,利用无监督的 OpenIE 方法和相关统计方法,参考其他知识图谱的本体结构,结合地理学科领域专家和一线教师的指导意见,完成面向基础教育领域的地理学科本体构建;

(2) 众包半自动语义标注:将地理学科教材教辅电子化后得到的文本作为标注对象,并以地理学科领域本体为标注依据,使用语义标注系统进行半自动语义标注,形成标注数据,且在此过程中完善地理学科领域本体;

(3) 外源数据补全:按照地理学科知识图谱的本体结构,对结构化的外部数据源进行相应的处理之后,得到外源数据,作为地理学科知识图谱的重要部分;

(4) 信息抽取:利用标注数据中的数据作为训练数据,按照地理学科知识图谱的本体结构,采用有监督、半监督和无监督的方法从百度百科等互联网文本中抽取实体和关系,得到扩充数据.

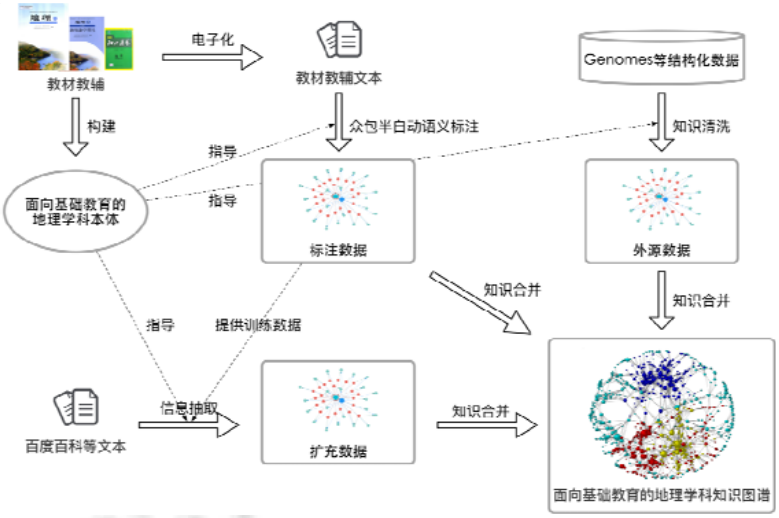

Fig.1 Construction route of geographical knowledge graph

图 1 地理学科知识图谱的构建路线

2.1 本体构建

对于基础教育领域的学科本体,覆盖率和准确率是非常重要的评价指标.在当前中文本体自动构建技术还不成熟的情况下,我们结合基础教育领域的特点,利用本体学习和统计学习等方法得到的本体知识,结合其他知识图谱的本体知识,在专家的指导下构建出了地理学科领域本体.

2.1.1 归纳领域概念

领域的核心概念对应的是本体中的类(owl:Class),每个核心概念都对应着许多实例,例如“国家”这个概念下就对应着“中国”“美国”“俄罗斯”等实例.我们采用 3 种方法来得到地理学科领域的核心概念.

(1) 用统计方法得到领域术语,再从领域术语中得到领域核心概念.

领域的核心概念是领域术语的子集合,因此可以用获取领域中术语的方法来获得领域概念.

从理论上讲,领域中的重要术语需要满足两个基本条件.(1) 术语在领域相关文档中出现的频率相对较高;(2) 术语在领域相关文档中出现的频率远高于在普通文档中出现的频率.

结合重要术语的两个基本条件可以看出,领域中的重要术语和文档集合中的关键词非常类似.所以利用相关统计学理论和文本挖掘技术,可以对领域中重要术语的归纳起到一定的辅助作用,进而大大缩小重要术语的查找范围.

TF-IDF[44]算法和 TextRank[45]算法是关键词提取研究领域中两个最基本的算法.TF-IDF 算法是一种统计方法,它的作用是评估一个词语对于一个语料库中的其中一份文档的重要程度.该算法的核心思想是:一个词语的重要性随着它在文档中出现的次数成正比地增加;但同时,会随着它在语料库中出现的频率成反比地下降.

TF-IDF 算法综合考虑了词语出现的频率、位置和密度等因素,但是它没有对整篇文档中相互有联系的词语进行综合考虑,而 TextRank 算法恰恰考虑到了词语之间的关系,并对词语的重要程度进行分配.

TextRank 算法基于 Google 的 PageRank[46]算法,其核心思想与 PageRank 算法相同:如果将网络中的节点看作是词语,那么在词语网络中词语的重要程度取决于与它相连的词语(指定窗口内的词语)给它的投票数目,而票的权重取决于该词语自己的票数.

以上两种算法得到的关键词确实有不少是地理学科的核心概念,例如“国家”“城市”“河流”“海洋”“地形”等等.而有些词语虽然出现频率较高,但是是实例而不是概念,例如“温带”和“中国”.还有些词语虽然出现频繁,但是与地理学科并不具有强相关性,例如“清单”“现象”“年代”“产生”“作用”等.

(2) 参考质量较高的知识图谱或数据源.

我们主要参考了 Schema.org(https://schema.org.cn)、DBpedia(http://mappings.dbpedia.org/server/ontology/classes/)和 Geonames.Schema.org 是一套包含语义信息的被各大搜索引擎所支持的 html 标签的词汇表,这里的标签等同于概念;DBpedia 是一个大规模的通用知识图谱,因此它也有一套概念体系;Geonames 中的每个地名都有对应的 featurecode 信息,而这些 featurecode 对应的就是概念.

(3) 众包半自动语义标注步骤中进行完善补充.

以上方法得到的核心概念的归纳整理需要参照本体构建的两个基本原则,即,本体中类的设计应当秉承独立性和共享性原则[10].前者指的是这个类可以独立存在,不依赖于特定的领域;后者指的是类是可以共享的,即有被复用的可能和必要.此外,本体中包含的类的数目应该尽可能地最小化,尽可能地去除冗余的类.根据上面的原则,我们最终确定了地理学科的核心概念.如图 2 展示的是位于“地理事实”概念下的核心概念(每个节点是一个概念,节点之间的连线表示父概念(靠近中心的节点)和子概念(远离中心的节点)的关系).

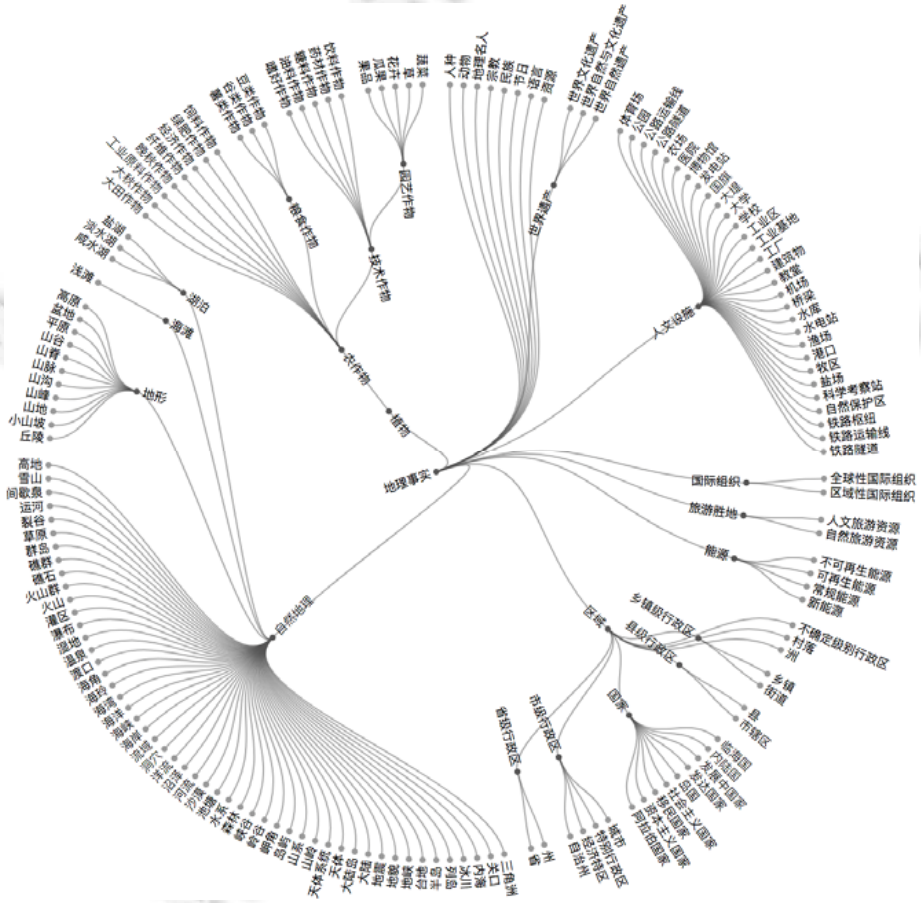

Fig.2 Concepts in geographical ontology (part)

图 2 地理学科本体中的核心概念(部分)

2.1.2 定义领域关系及其约束

关系是本体的核心基本要素,它是对领域中的概念、实例之间的相互作用的描述.关系直接决定了本体知识图谱的知识丰富程度以及基于知识图谱构建的其他应用系统的功能范围.关系学习是本体学习中的一个重要部分,我们主要通过 4 种方法来定义关系.

(1) 利用 OpenIE 方法对地理学科领域文本进行无监督的开放关系抽取,再从中找到有意义的关系.优点是无需标注语料或其他预处理直接就可以抽取原始文本,缺点是抽取出的结果大多数是像(船员们-历经-千辛万苦)这样无意义的关系.我们将从 27 本地理教材教辅中得到的 63 567 个句子用 Stanford 的 OpenIE 工具(https://stanfordnlp.github.io/CoreNLP/openie.html)处理后,得到了 112 145 个带有打分值的关系抽取结果,并在此基础上对关系进行过滤;

(2) 参考质量较高的知识图谱或数据源 Wikidata(https://www.wikidata.org/wiki/Wikidata:Main_Page)和 Schema.org.Wikidata 是维基百科的结构化版本,也可以理解为是一个大规模的通用知识图谱,我们主要参考了地理概念相关的关系列表(https://www.wikidata.org/wiki/Wikidata:List_of_properties/Geographical_feature);Schema.org 在每个标签下有其对应的众多关系,我们也主要参考了地理领域相关的标签,例如“城市”标签(https://schema.org.cn/City);

(3) 根据核心概念和百科信息框来确定关系.每个核心概念下都有很多实例,大多数实例在百科中都有对应的信息框.通过整合同一概念下多个实例的信息框信息,便可以得到该概念下较为重要的关系.例如“国家”概念下有“中国”“美国”“俄罗斯”等国家,这些国家的信息框中都包含了“面积”“人口数量”等关系,那么这些关系就比较重要;

(4) 众包半自动语义标注过程中补充新的关系.在众包半自动语义标注过程中,如果发现有新的关系无法用已有的关系表达时,便说明这是一个新的关系需要补充.

根据上述方法,我们最终整理得到了 400 多个关系,对于每个关系都有详细的描述,包括关系的名称、描述、URI 形式、定义域和值域等内容,表 1 是地理学科的“特征”关系的详细描述.

Table 1 Description of a relation called “characteristic”

表 1 “特征”关系的描述

名称	描述	值
Label	关系的名称	特征
Description	关系的描述	地理常用属性:特点,特性
URI	关系的 URI	http://edukb.org/knowledge/property/geo#tezheng
Domain	关系的定义域	[“http://www.w3.org/2002/07/owl#NamedIndividual”]
Range	关系的值域	[“http://www.w3.org/2000/01/rdf-schema#Literal”]
Type	关系的类型	数值属性

2.1.3 本体检查

目前,学术界研究学者公认,在构建领域本体的过程中需要领域专家的参与和协作,所以在完成了前面的本体构建的两个步骤之后,我们特别邀请了北京市具有丰富教学经验和教材分析能力的地理学科专家和一线教师来指导我们对本体进行检查和评估.根据专家的指导意见,我们修改和完善后得到最终的地理学科领域本体.地理专家和一线教师的指导主要包括两方面:一是核心概念的结构是否合理,例如专家建议总的可以分为地理概念、地理事实、地理方法和地理原理;二是每个概念的合理性、必要性以及相似概念间的辨析,例如专家建议要有“地理名人”这个概念.

2.2 众包半自动语义标注

标注数据是地理学科知识图谱的基础和重点,我们采取的是众包半自动语义标注的方式来保证质量和效率.标注的数据来源是 HTML 格式的教材教辅文本.

基于领域本体的语义标注是指在领域本体的指导下,从文档中抽取出结构化知识的过程,即,将文档中的纯文本知识用 RDF 语言描述出来.语义标注的过程通常可以包含两种标注.

(1) 类型标注:将文档中与本体中概念相对应的词语标记出来,并将该词语作为概念所对应的实例;

(2) 关系标注:找出实例之间存在的与本体中关系相对应的关系,关系标注可以丰富实例的内在信息.标注时,通常将实例及实例间的关系表示为三元组的形式(E_1,R,E_2),其中,R 是实例 E_1 和 E_2 之间的关系.

从对比结果中,我们可以总结出语义标注系统作为知识图谱构建的关键系统,其主要需求包括以下几点.

(1) 标注依据:语义标注系统提供的是基于本体的语义标注功能,所以它必须要能够导入一个或多个本体描述文件,或者采用包含本体信息的文件进行配置,这样,语义标注系统才有了基本的标注依据;

(2) 标注对象:语义标注系统一般都支持对文本文件或者静态网页文件的标注,目前,大多数的教材教辅书籍数据都存放在静态网页文件中,所以语义标注系统需要支持对静态网页文件的标注功能;

(3) 标注方式:语义标注系统必须能提供基本的标注功能,包括类型标注和关系标注.同时,考虑到教材教辅书籍数据中存在着大量的图片也需要进行标注,所以语义标注系统还要能够支持图片标注的功能;

(4) 本体语言:目前,大多数的语义标注工具都只支持 RDF(S)、DAML+OIL、XML 等本体语言中的某个或者某几个,而对 W3C 推荐的最新的本体描述语言 OWL 支持较少.所以为了更好地使用不同的本体语言,语义标注系统应该能够支持目前主流的本体语言,例如 RDF(S)和 OWL.

除了上面 4 个基本需求之外,结合我们构建的地理学科知识图谱目标,我们认为,以下需求对于语义标注系统同样重要.

(1) 协同式标注:出现时间较早的语义标注系统一般都是 C/S 模式的,不仅需要标注人员安装客户端,而且软件配置和语义标注过程都不太方便.随着互联网的发展,基于 B/S 模式的语义标注系统逐渐出现,因为它可以很方便地支持大量标注人员的协同式标注,显著地提高了标注速度;

(2) 标注审核:标注系统应该具有一定的用户权限控制.简单情况下,用户主要包括标注人员和审核人员两种,其中,标注人员只能编辑和删除自己的标注记录,而审核人员可以编辑和删除当前页面所有标注人员的标注记录;

(3) 标注溯源:对于任何一条由页面标注而产生的知识,在生成对应知识的同时需要保存将来能够追溯到具体的标注来源这个元数据信息.通常,标注溯源都是采用 XPointer[47]技术来实现的,XPointer 是一种根据数据在 XML 文件中的位置、字符内容或者属性值等特性对数据进行定位的语言;

(4) 标注数据存储:标注数据的存储也是需要重点考虑的问题,目前已有不少出色的 RDF 数据库可供选择.其中,Sesame[48]数据库是一个开源项目,它不仅架构简单易于部署,而且功能完善易于操作.它实现了一个通用的 RDF 数据管理框架,并提供了相应的编程接口,以便于集成不同的存储系统、推理和查询引擎等;

(5) 共指消解:实例的共指问题是标注网页数据的过程中经常出现的问题,它是指在不同的网页文档中出现了相同的实例,例如,多个文档中都出现了地名实例“中国”.为了避免产生多个重复冗余的实例,标注工具应该具有实例查询的能力,这样,当遇到相同实例时,可以选择已经存在的实例进行标注,从而避免重新生成新的实例造成的实例共指问题.

针对以上几点需求,结合当前要构建的地理学科知识图谱目标,我们提出相应的语义标注的架构,如图 3 所示.在地理学科本体和资源管理系统的基础上,利用语义标注系统,通过标注人员的标注产生标注数据库,最终清洗导出到标注数据中.

我们研发的语义标注系统在原 Pundit 系统功能的基础上添加了许多新的功能,例如,

(1) 标注审核权限控制:为了保证标注结果的准确率,对标注人员的标注结果进行审核是语义标注系统的一个核心需求;

(2) 添加本体描述文件作为标注系统的配置:原系统虽然支持自定义属性和领域词表,但是它并不直接支持将本体描述文件作为标注系统的配置,也不支持根据网页数据类型灵活地改变属性和领域词表的功能.所以,为了方便语义标注系统能够自适应不同领域的本体配置,我们添加了此项功能;

(3) 自定义新建实例和搜索实例.当实例的名称在页面内容中不存在时,如何创建新的实例?考虑到在标

注中有此需求,故而我们添加此功能.而搜索实例则是为了减少实例共指的问题,即,标注人员在进行关系标注时可能并不确定某个实例是否已经在标注数据中存在了.自定义新建实例的功能和实例搜索功能一般是联动的,首先是进行实例搜索,如果搜索到相应的实例就直接使用;如果没有搜索到相应的实例则可以新建实例.

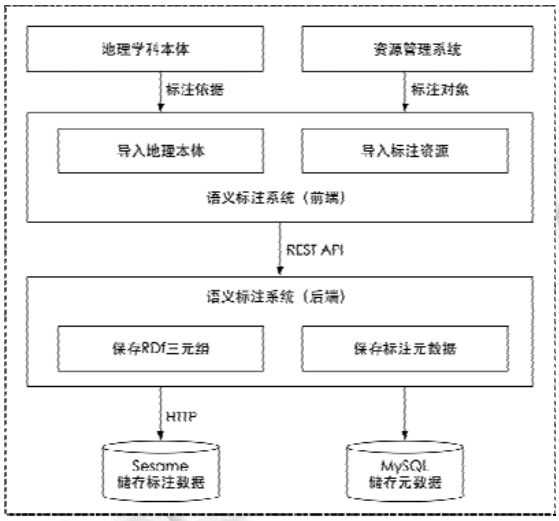

Fig.3 Semantic annotation architecture

图 3 语义标注的架构

知识清单是课本内容的凝练,考虑到标注所有课本花费的时间和人力成本较大,主要对知识覆盖率满足需求的初高中知识清单进行了众包标注(如图 4 所示).基本思路是:标注管理员确定好标注方案,由标注人员进行标注,标注管理员再进行审核.经过标注人员的标注和标注管理员的审核,我们获得了一个准确率和覆盖率都接近 100%的标注数据,同时也自下而上地完善、确定了本体,为后续外源数据补全和信息抽取打下了良好的基础.

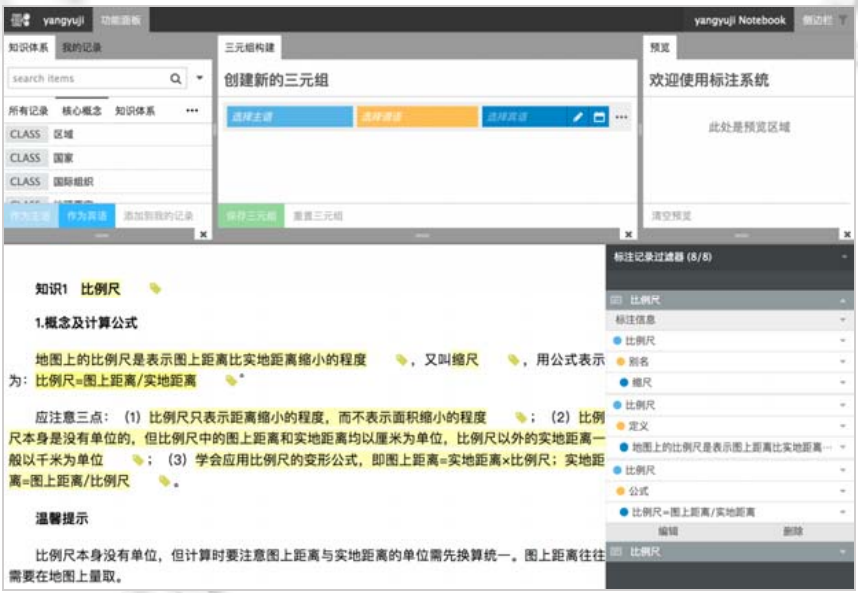

Fig.4 Semi-Automatic crowdsourcing semantic annotation system

图 4 半自动众包语义标注系统

2.3 外源数据补全

外源数据指的是外部数据源按照地理学科领域本体结构处理后得到的与标注数据结构一致的 RDF 数据.外部数据源一般是互联网上公开的知识图谱或其他结构化程度较好的网站,特点是数据量大、结构较好.以下介绍地理学科知识图谱中的 3 个外部数据源.

2.3.1 Geonames

Geonames 是地理信息领域较为权威的一个知识图谱,包含超过 1 000 万条的地理地名信息,数据准确率高.主要是英文数据,较重要的地名会有其他语言的名称(label),例如含有中文名称的地名有 61 万多个.每个地名信息有 19 个属性信息(部分属性可为空)[**].部分属性信息可以直接作为知识图谱中的三元组事实,例如经度(longitude);部分属性信息需要按照本体结构进行处理,例如我们将特征码(feature code)属性信息处理后作为实例和概念间关系;将一级行政区划码(admin1 code)、二级行政区划码(admin2 code)等属性信息处理后作为地名之间的上下位关系.

2.3.2 百度百科信息框

百度百科信息框是领域知识图谱扩充三元组事实较好的来源,在第 2.2 节中众包语义标注和第 2.4.1 节中实体集扩充步骤得到的实例的基础上,我们通过以下步骤得到高质量的三元组.

(1) 获取实例和关系名集合.对每个概念 c,我们用 $E=\{e_1,\ldots,e_N\}$表示它的实例集合,对每一个实例 e_i,我们都爬取该实例对应的百度百科信息框,得到所有信息框中的关系名集合 $R=\{r_1,\ldots,r_M\}$,集合大小为 M;

(2) 连边.如果实例 e_i 的信息框中含有 r_j,则将 e_i 与 r_j 之间边的权重设置为 1;如果不含,则设置为 0.为了避免出现图稀疏现象,我们加上了实例和实例、关系名和关系名之间的边.对于实例和实例连边的操作,首先为每一个实例设置一个关系名向量 V,向量的维度等同于关系名集合的大小 M.如果关系名 r_k 存在于该实例的信息框中,则设置为 1;若不存在,则将该位置设置为 0.进而可以得到实例和实例关系名向量之间的余弦相似度,作为实例和实例之间边的权重.同理,可以给每个关系名设置一个实例向量,进而得到关系名和关系名之间的余弦相似度作为它们之间边的权重;

(3) 迭代计算.我们采用了一种图强化算法[49]进行迭代,迭代计算后,便可得到每个概念下的实例和关系名典型度排序;

(4) 将典型度高的关系名及其值信息加入到知识图谱中.

值得一提的是:上述步骤还有检查知识图谱中实例分类错误的作用,如果步骤(3)中得到某概念下的实例典型度较低,则很有可能是分类错误的实例.

2.3.3 中国行政区划信息

中国行政区划在地理学科中较为重要,为此,我们从国家统计局网站(http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2016/)得到了中国行政区划精确到乡镇层级的信息,主要是行政区划之间的上下位关系.因为是完全结构化的呈现方式,因此,数据按照本体结构处理后直接加入知识图谱.

2.4 信息抽取

扩充数据指的是利用之前得到的标注数据和外源数据,运用机器学习等方法从文本中抽取的 RDF 三元组数据.扩充数据是地理学科知识图谱的重要组成部分.我们使用的文本语料是《中国大百科全书》中的《世界地理》卷、《中国地理》卷、《地理学》卷(以下简称中国大百科文本)和百度百科维基百科文本(以下简称百度维基文本).两部分语料各有特点,中国大百科文本数量虽少质量却很高,百度维基文本质量一般但数量却很大.

2.4.1 实体集扩充

我们想要根据知识图谱中每个概念的实体集进行扩充.使用的方法是词向量.词向量最早是由 Hinton 于

** http://download.geonames.org/export/dump/readme.txt. 19 个属性名依次为 geonameid, name, asciiname, alternatenames, latitude, longitude, feature class, feature code, country code, cc2, admin1 code, admin2 code, admin3 code, admin4 code, population, elevation, dem, timezone, modification date.

1986 年提出,又称为概念的分布式表达[50].词向量的主要作用是通过大量词语语料的训练,将每个词语映射到一个固定维度的向量,从而可以根据两个词语的向量之间的余弦距离来刻画两个词语的语义相关性.使用最为广泛的方法是 Google 团队的 Word2Vec 方法[51,52].

事实上,实体集扩充之后,应该还有一步实体消歧的操作,但是通用知识图谱存在着较多的歧义,领域知识图谱的歧义较少.例如,“苹果”既是水果,又是科技公司,但是几乎不存在某个领域知识图谱会同时包括科技公司和水果这两个概念.

2.4.2 关系抽取

我们采用了无监督、有监督和半监督这 3 种方法来进行关系抽取.

- 在无监督的方法中,我们使用了基于规则的方法和 LDA 模型.基于规则的方法中,我们为想要抽取的关系定义正则模板,然后从文本中抽取相应关系的文本描述;LDA 模型是一种无监督的机器学习技术,它可以用于识别文本中隐藏的关系类别信息,并且用词袋的方式来表示每类关系的特征;
- 在有监督的方法中,我们将知识图谱中已有的关系数据作为训练数据,从文本中抽取出相应的三元组.由于已有的数据仍然不够多,因此为防止过拟合,我们使用了较为简单的多层感知机;
- 在半监督的方法中,我们采用了基于多语言注意力机制的远程监督方法[53],通过利用多种语言之间具有一致性的信息,实现了比单语言更好的抽取效果.

3 实 验

本节将主要对半自动语义标注、实体集扩充和关系抽取进行相关实验,以验证“四步法”构建领域知识图谱的有效性.

3.1 半自动语义标注

地理学科的本体知识图谱是通过语义标注系统进行众包半自动语义标注完成的,而项目中数学学科的本体知识图谱则是采用手工标注完成的,数学学科的标注是两位学生采用自定义的标注格式对教辅书籍进行手工标注,并且标注之前他们需要在标注认识上达成一致,最终耗时 1 个月完成数学学科的基本知识图谱标注任务,经过后期数据处理之后,共得到近 3 000 个实例和近万条三元组.而地理学科的标注是 4 位学生在经过 1 次标注培训之后,耗时 1 周完成地理学科的基本知识图谱标注任务,共得到 2 400 多个实例和 30 000 多条三元组.对比可以看出:采用标注系统的半自动语义标注在标注速度上远远超过手工标注,而且大大降低了标注培训的时间以及后期标注结果处理的复杂度.

在标注质量上,我们采用准确率(P)、召回率(R)和 $F1$ 值来衡量.其中,N_1 表示标注人员标注出来的三元组总数,N_2 表示页面应该标注的三元组总数,N_3 表示人工标注出来的三元组中正确的总数.

表 2 列出了标注人员标注后,从地理学科标注文档中随机抽取 7 个标注文档,并对文档中的标注记录进行统计得到的结果.

Table 2 Results of the semi-automatic semantic annotation

表 2 半自动语义标注的标注结果

文档编号	1	2	3	4	5	6	7	总计
N_1	65	30	33	64	15	10	46	263
N_2	68	31	35	70	15	10	50	279
N_3	64	30	31	62	15	10	40	252
准确率 P(%)	98	100	94	97	100	100	87	96
召回率 R(%)	94	97	89	89	100	100	80	90
$F1$ 值(%)	96	98	91	93	100	100	83	93

从表 2 中可以看出,文档的标注 $F1$ 值大多数在 90%以上.通过对标注错误和漏标的三元组进行分析,我们发现,这些标注错误主要有以下 3 点原因所致.

(1) 标注人员标注时出现手误:标注过程需要标注人员从文档中选择合适的文本片段作为主语或者宾语,

标注人员在标注时可能手误而导致最后的文本片段多了或者少了某部分内容;

(2) 标注人员对知识的理解有误:标注出来的三元组依赖于标注人员对于相关知识的理解程度,如果标注人员对该知识的理解不正确,那么就容易导致一些错误的标注数据,而且这个错误会导致整个页面出现很多类似错误;

(3) 标注人员漏标注:标注人员在标注过程中由于理解错误或者不细心,会不可避免地漏掉一些内容,那么这些内容中的知识就漏标注了.

针对以上错误,标注管理员和标注人员对地理学科页面的标注进行了审核纠正.

3.2 实体集扩充

我们使用百度维基文本作为Word2Vec方法的训练语料,训练结束后输入一个词语,会得到这个词语的相似词语集.我们把知识图谱中某个概念下的 M 个实体作为输入,每个实体的相似词语取前 K 个,一共得到 $M\times K$ 个含有重复词语的集合.我们取重复次数$\geq N$ 的词语作为扩充的新实体.一个词语的重复次数越高,那么该词语映射到该概念下的新实体的可能性就越大.例如,表 3 是 M=3 时的几个例子(K=50,N=2).

Table 3 Results of instance expansion by Word2Vec method

表 3 Word2Vec 方法扩充实例的效果

核心概念	输入实体	扩充实体个数	扩充准确率(%)
河流	密西西比河 尼罗河 湄公河	10	70.0
湖泊	太湖 鄱阳湖 苏必利尔湖	10	60.0
山脉	黄山 泰山 云台山	10	90.0
城市	郑州 南京 青岛	28	96.4
大学	清华大学 复旦大学 南京大学	44	97.7

值得注意的是:如果概念之间的混淆度不高的话,扩充的实体的效果会很好,例如“城市”和“大学”;但如果核心概念之间的相关性大,效果就会不好,例如“河流”和“湖泊”都是“水域”的子概念,所以输出的结果中有很多是相互混淆的,也因此扩充准确率较低.原因是,相似概念的实体的上下文也是相似的.而 Word2Vec 方法就是根据上下文训练得到的词语的向量表示,故 Word2Vec 方法无法较好地区分出相似概念下的扩充实体.

3.3 关系抽取

3.3.1 无监督方法

我们使用 LDA 模型对中国大百科文本进行聚类分析,得到每种关系对应的特征词.表 4 展示了其中一些关系的特征词抽取效果,其中,“位于”和“毗邻”是严格的“关系”,其他的是“属性”.

Table 4 Results of extracting relations' characteristic words using LDA model

表 4 LDA 模型抽取关系特征词的效果

关系类别	特征词词集
地势	米 海拔 位于 公里 山地 山脉 平原 高原 盆地 地势 丘陵 东部 南北
气候	摄氏度 气候 降水量 毫米 降水 平均气温 温带 季节 季风气候
交通	公里 铁路 站 公路 机场 高速公路 交通 高速 建设 货物 运输 段 客运
动物	动物 保护 自然保护区 种类 丰富 湿地 生物 鸟类 鱼类 物种 资源
旅游资源	寺 山 旅游 风景区 瀑布 景观 洞 名胜区 著名 风光 道教 游客 景色
位于	属于 从属 包含 在 处于 处在 地处 坐落 所在地
毗邻	相邻 邻接 接壤 相接 挨着 国界 毗连 相连

接着,我们利用每种关系的特征词以及总结的正则式,使用 Bootstrapping 的方法从中国大百科文本中迭代抽取表达关系的文本.表 5 是不同策略下,基于中国地理大百科文本 3 000 个句子、20 个关系上测试的结果.

Table 5 Results of relation extraction with multiple strategies

表 5 多种策略关系抽取效果

策略名称	正确率(%)	召回率(%)
正则式	85.4	90.8
特征词	78.8	81.6
特征词+正则式	89.1	94.1

3.3.2 有监督方法

基于知识图谱中每种关系的训练数据,我们尝试用多层感知机对中国大百科文本来进行关系抽取,在将句子向量化时,采用了两种方式:(1) 选取句子中所有词语向量的均值作为句子的向量值;(2) 简单地将所有词语向量拼接在一起作为句子的向量值.我们选取了关系含义重叠度较小的 7 种关系,使用多层感知机的方法进行了实验.表 6 展示了训练文本的样式,表 7 展示了多层感知机对文本按照关系进行分类的效果.

Table 6 Examples of training data

表 6 训练文本的样式

待区分关系的文本	标记
地势低平,海拔只有 3 米	[1,0,0,0,0,0,0]
夏季酷热,8 月最高气温在 45 摄氏度以上	[0,1,0,0,0,0,0]
国内以陆上交通为主,有铁路 8 823 公里	[0,0,1,0,0,0,0]
动物有阿尔卑斯大角山羊和土拨鼠、山兔、小羚羊、雷鸟等	[0,0,0,1,0,0,0]
渴湖沿岸风景优美,旅游业颇盛	[0,0,0,0,1,0,0]
珠穆朗玛峰地处青藏高原	[0,0,0,0,0,1,0]
蒙古南部和中国接壤	[0,0,0,0,0,0,1]

Table 7 Results of relation classification using MLP

表 7 多层感知机关系分类的效果

模型	正确率(%)	召回率(%)
平均值法	64.6	70.4
拼接法	68.3	72.6

我们发现,多层感知机的效果远不如 LDA 和正则.原因是句子中的词语太多,拼接或者平均的操作均会导致关键词的信息被弱化,甚至被湮灭.而基于正则或特征词的方法直接抓住了关键词,虽然关键词可能有遗漏,但仍然可以保证覆盖到大多数正确的句子.

3.3.3 半监督方法

我们使用了清华大学自然语言处理实验室公开的中英双语关系抽取数据集(https://github.com/thunlp/MNRE)来做远程监督的关系抽取实验,选取了地理学科领域中的 49 种关系进行实验,其中,有 1 种特殊的关系 NA 表示实体之间没有任何关系.我们为中英文分别设置了训练集、验证集和测试集,详细信息见表 8.

Table 8 Statistics of the dataset

表 8 数据集的统计信息

数据集		关系数量	句子数量	事实数量
英文	训练集	49	983 930	34 147
	验证集		77 392	1 400
	测试集		156 480	2 758
中文	训练集	49	934 010	40 474
	验证集		79 835	1 400
	测试集		161 635	2 758

关系抽取方法使用的是基于多语言注意力机制的远程监督方法 MNRE[53],评估方法采用的是 held-out[54],即看知识图谱中已有的事实被关系抽取方法发现的情况,正确的事实被排序在前面的越多,那么在准确率/召回率曲线上,随着召回率的增加准确率就下降得越慢,关系抽取效果就越好.实验结果的准确率/召回率曲线如图 5

所示.

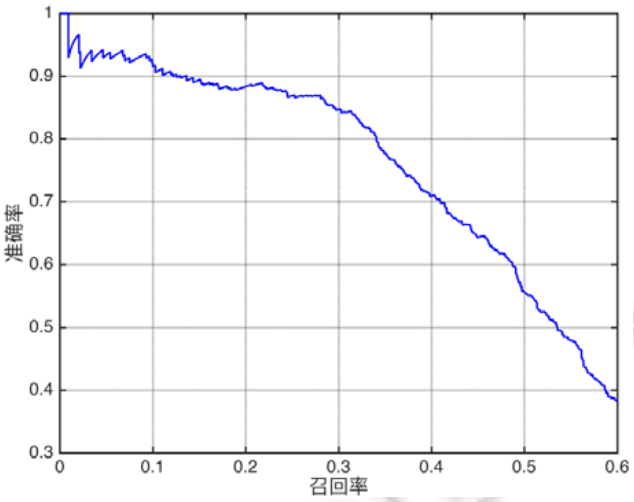

Fig.5 Aggregated precision/recall curves of MNRE

图 5 MNRE 方法的准确率/召回率曲线

3.4 实验讨论

众包半自动语义标注通过标注审核的方式协同标注,在提高效率的同时,保证了准确率.经地理学科专家检查,标注数据的知识覆盖率和知识准确率均达到了 99%以上.

外源数据的准确率和效率都很高,因为外部数据源的准确率高,结构较好,所以较易处理.

实体集扩充和关系抽取会引入错误的数据,因此为了保证知识图谱的高质量,需要对这些数据进行人工审核纠正.

综合上述方法,地理学科知识图谱准确率是较高的.由于引入了人工审核纠正,效率有所下降,但仍然是可以接受的.我们用“四步法”构建出的地理学科知识图谱包含 67 万个实例、1 421 万条 RDF 三元组(具体统计见表 9).

Table 9 Statistics of gepgraphical knowledge graph

表 9 地理学科知识图谱数据量统计

数据分类	来源	实例数量	RDF 数量
标注数据	教材教辅	2 438	30 600
外源数据	Geonames	616 282	13 757 146
	中国行政区划	45 025	224 149
	百度百科信息框	–	29 972
扩充数据	百度维基文本	4 490	156 647
	中国大百科文本	3 794	45 259
总计	–	672 029	14 243 773

4 结 论

领域知识图谱应用很广,构建难度却很大,自动化的方法尚不成熟,人工方法效率低下.本文提出的准确、高效地构建领域知识图谱的方法——“四步法”,可以很好地平衡自动化方法和人工参与,在效率可以接受的情况下实现很高的准确率.该方法的适用领域是对知识覆盖率和召回率要求较高的领域.例如本文中的地理学科知识图谱,作为基础教育学科的知识图谱,保证知识点完全覆盖是必需的.如果对知识覆盖率和召回率要求不严格的领域,可以考虑将众包语义标注替换为信息抽取和人工审核结合的方法来得到核心标注数据,同时降低时间和人力成本.

研究实现的众包半自动语义标注系统,在“四步法”中起着重要作用.相比于人工标注降低了标注难度;相比于自动标注等方法能够更好地保证标注质量,还可以在标注过程中修改完善本体结构.

本文还以地理学科知识图谱为例,详细介绍了“四步法”构建领域知识图谱的过程.希望能为其他研究者构建领域知识图谱提供一定的借鉴.

References:

[1] Bernerslee T, Hendler J, Lassila O. The semantic Web. Scientific American, 2001,284(5):34−43.

[2] Bizer C, Lehmann J, Kobilarov G, *et al*. DBpedia—A crystallization point for the Web of data. Web Semantics: Science, Services and Agents on the World Wide Web, 2009,7(3):154−165.

[3] Suchanek FM, Kasneci G, Weikum G. Yago: A core of semantic knowledge. In: Proc. of the 16th Int'l Conf. on World Wide Web. ACM Press, 2007. 697−706.

[4] Bollacker K, Evans C, Paritosh P, *et al*. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2008. 1247−1250.

[5] Singhal A. Introducing the knowledge graph: Things, not strings. Official Google Blog, 2012.

[6] Gruber TR. Towards principles for the design of ontologies used for knowledge sharing. Int'l Journal of Human-Computer Studies, 1993,43.

[7] Studer R, Benjamins VR, Fensel D. Knowledge engineering: Principles and methods. Data & Knowledge Engineering, 1998,25(1): 161−197.

[8] Du WH. A comparative study of ontology construction methods. Journal of Infomation, 2005,24(10):24−25 (in Chinese with English abstract).

[9] Shang XL. Comparative analysis of foreign ontology construction methods. Library and Information Service, 2012,56(4):116−119 (in Chinese with English abstract).

[10] Liu YS. Research of approaches and development tools in constructing ontology. Journal of Modern Infomation, 2009,29(9):17−24 (in Chinese with English abstract).

[11] Han J, Xiang Y. A survey on ontology building. Computer Applications and Software, 2007,24(9):21−23 (in Chinese with English abstract).

[12] Miller GA. WordNet: A lexical database for English. Communications of the ACM, 1995,38(11):39−41.

[13] Uschold M, King M. Towards a methodology for building ontologies. In: Proc. of the Workshop on Basic Ontological Issues in Knowledge. 1995,133(2):137−142.

[14] Fox MS. The TOVE project towards a common-sense model of the enterprise. In: Proc. of the Int'l Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE'92). Paderborn, 1992. 25−34.

[15] Swartout B, Patil R, Knight K, *et al*. Toward distributed use of large-scale ontologies. In: Proc. of the 10th Workshop on Knowledge Acquisition for Knowledge-Based Systems. 1996. 138−148.

[16] Fernández-López M, Gómez-Pérez A, Juristo N. METHONTOLOGY: From ontological art towards ontological engineering. In: Proc. of the AAAI'97. 1997.

[17] Noy NF, McGuinness DL. Ontology development 101: A guide to creating your first ontology. 2001. https://doi.org/10.1016/j.artmed.2004.01.014

[18] Du XY, Li M, Wang S. A survey on ontology learning research. Ruan Jian Xue Bao/Journal of Software, 2006,17(9):1837−1847 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1837.htm [doi: 10.1360/jos171837]

[19] Hu FH. Chinese knowledge graph construction method based on multiple data sources [Ph.D. Thesis]. Shanghai: East China University of Science and Technology, 2015 (in Chinese with English abstract).

[20] Qiu JP, Mu N, Lou W, *et al*. An analysis of the progress of semantic annotation at home and abroad. Information Studies: Theory & Application, 2014,37(5):12−16 (in Chinese with English abstract).

[21] Jing T, Zuo WL, Sun JG, *et al*. Semantic annotation of Chinese Web pages: From sentences to RDF representations. Journal of Computer Research and Development, 2008,45(7):1221−1231 (in Chinese with English abstract).

[22] Zou L, Liao SM. Comparison and analysis of semantic annotation tools based on ontology. Computer Application, 2004,24(S1): 328−330 (in Chinese with English abstract).

[23] Tao W, Li P, Liao SM. Analysis and summary of current ontology-based semantic annotation tools. Journal of Anhui University of Technology and Science, 2005,20(2):52−55 (in Chinese with English abstract).

[24] Yin CY, Bi Q, Wang CQ. Research on the characteristics of semantic annotation tools and its applicability. Information Studies: Theory & Application, 2014,37(12):111−116 (in Chinese with English abstract).

[25] Guo SY, Dou C, Chang Z. Review on semantic annotations of Web pages. Journal of Intelligence, 2015,(4):169−175 (in Chinese with English abstract).

[26] Kiryakov A, Popov B, Terziev I, *et al*. Semantic annotation, indexing, and retrieval. In: Proc. of the Semantic Web (ISWC 2003). Berlin, Heidelberg: Springer-Verlag, 2003. 484−499.

[27] Andrews P, Zaihrayeu I, Pane J. A classification of semantic annotation systems. Semantic Web, 2012,3(3):223−248.

[28] Reeve L, Han H. Survey of semantic annotation platforms. In: Proc. of the ACM Symp. on Applied Computing. 2005. 1634−1638.

[29] Sporny M, Longley D, Kellogg G, *et al*. JSON-LD 1.0. W3C Recommendation (2014-1-16). 2014.

[30] Adida B, Birbeck M, McCarron S, *et al*. RDFa in XHTML: Syntax and processing. W3C Recommendation. 2008.

[31] Structured Data Markup Helper. Webmasters, Google Inc. https://www.google.com/webmasters/markup-helper/

[32] Grassi M, Morbidoni C, Nucci M, *et al*. Pundit: Semantically structured annotations for Web contents and digital libraries. In: Proc. of the SDA. 2012. 49−60.

[33] Morbidonia C, Picciolib A. Pundit 2.0. 2015.

[34] Heflin J, Hendler J, Luke S. SHOE: A knowledge representation language for internet applications. 1999. https://www.researchgate.net/publication/2620999_SHOE_A_Knowledge_Representation_Language_for_Internet_Applications

[35] Petridis K, Anastasopoulos D, Saathoff C, *et al*. M-Ontomat-Annotizer: Image annotation linking ontologies and multimedia low-level features. In: Proc. of the Knowledge-Based Intelligent Information and Engineering Systems. Berlin, Heidelberg: Springer-Verlag, 2006. 633−640.

[36] Kahan J, Koivunen MR, Prud'Hommeaux E, *et al*. Annotea: An open RDF infrastructure for shared Web annotations. Computer Networks, 2002,39(5):589−608.

[37] Kalyanpur A, Hendler J, Parsia B, *et al*. SMORE-Semantic markup, ontology, and RDF editor. 2006. https://www.researchgate.net/publication/235138099_SMORE-semantic_markup_ontology_and_RDF_editor

[38] Kogut P, Holmes W. AeroDAML: Applying information extraction to generate DAML annotations from Web pages. In: Proc. of the Workshop Knowledge Markup & Semantic Annotation (K-CAP 2001). Victoria, 2001.

[39] Vargas-Vera M, Motta E, Domingue J, *et al*. MnM: Ontology driven semi-automatic and automatic support for semantic markup. In: Proc. of the Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. Berlin, Heidelberg: Springer-Verlag, 2002. 379−391.

[40] Gupta S, Manning CD. Improved pattern learning for bootstrapped entity extraction. In: Proc. of the CoNLL. 2014. 98−108.

[41] Curran JR, Murphy T, Scholz B. Minimising semantic drift with mutual exclusion bootstrapping. In: Proc. of the 10th Conf. of the Pacific Association for Computational Linguistics. 2007. 172−180.

[42] Zhang CY. The study of entity relation extraction algorithm [Ph.D. Thesis]. Beijing: Beijing University of Posts and Telecommunications, 2015 (in Chinese With English abstract).

[43] Liu Q, Li Y, Duan H, *et al*. Knowledge graph construction techniques. Journal of Computer Research and Development, 2016,53(3): 582−600 (in Chinese with English abstract).

[44] Salton G. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[45] Mihalcea R, Tarau P. TextRank: Bringing order into texts. In: Proc. of the Association for Computational Linguistics. 2004.

[46] Page L, Brin S, Motwani R, *et al*. The PageRank citation ranking: Bringing order to the Web. Stanford Digital Libraries Working Paper, 1998,9(1):1−14.

[47] DeRose S, Maler E, Daniel R. XML pointer language (XPointer). 2000. https://www.researchgate.net/publication/2771988_Xml_Pointer_Language_XPointer

[48] Broekstra J, Kampman A, Van Harmelen F. Sesame: A generic architecture for storing and querying RDF and RDF schema. In: Proc. of the Semantic Web (ISWC 2002). Berlin, Heidelberg: Springer-Verlag, 2002. 54−68.

[49] Zhang Z, Sun L, Han X. A joint model for entity set expansion and attribute extraction from Web search queries. In: Proc. of the AAAI. 2016. 3101−3107.

[50] Hinton GE. Learning distributed representations of concepts. In: Proc. of the 8th Annual Conf. of the Cognitive Science Society. 1986. 12.

[51] Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv Preprint arXiv:1301.3781, 2013.

[52] Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. In: Proc. of the Advances in Neural Information Processing Systems. 2013. 3111−3119.

[53] Lin Y, Liu Z, Sun M. Neural relation extraction with multi-lingual attention. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, Vol.1: Long Papers. 2017. 34−43.

[54] Mintz M, Bills S, Snow R, *et al*. Distant supervision for relation extraction without labeled data. In: Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int'l Joint Conf. on Natural Language Processing of the AFNLP, Vol.2. Association for Computational Linguistics, 2009. 1003−1011.

附中文参考文献:

[8] 杜文华.本体构建方法比较研究.情报杂志,2005,24(10):24−25.

[9] 尚新丽.国外本体构建方法比较分析.图书情报工作,2012,56(4):116−119.

[10] 刘宇松.本体构建方法和开发工具研究.现代情报,2009,29(9):17−24.

[11] 韩婕,向阳.本体构建研究综述.计算机应用与软件,2007,24(9):21−23.

[18] 杜小勇,李曼,王珊.本体学习研究综述.软件学报,2006,17(9):1837−1847. http://www.jos.org.cn/1000-9825/17/1837.htm [doi: 10.1360/jos171837]

[19] 胡芳槐.基于多种数据源的中文知识图谱构建方法研究[博士学位论文].上海:华东理工大学,2015.

[20] 邱均平,牟楠,楼雯,等.国内外语义标注研究进展分析.情报理论与实践,2014,37(5):12−16.

[21] 荆涛,左万利,孙吉贵,等.中文网页语义标注:由句子到 RDF 表示.计算机研究与发展,2008,45(7):1221−1231.

[22] 邹亮,廖述梅.基于本体的语义标注工具比较与分析.计算机应用,2004,24(S1):328−330.

[23] 陶皖,李平,廖述梅.当前基于本体的语义标注工具的分析.安徽工程科技学院学报:自然科学版,2005,20(2):52−55.

[24] 尹长余,毕强,王传清.语义标注工具的特征分析及其适用性研究.情报理论与实践,2014,37(12):111−116.

[25] 郭少友,窦畅,常桢.网页语义标注研究综述.情报杂志,2015,(4):169−175.

[42] 张春云.实体关系抽取算法研究[博士学位论文].北京:北京邮电大学,2015.

[43] 刘峤,李杨,段宏,等.知识图谱构建技术综述.计算机研究与发展,2016,53(3):582−600.

杨玉基(1994－),男,河南巩义人,硕士,主要研究领域为知识图谱,数据挖掘.

仝美涵(1995－),女,博士,主要研究领域为知识工程,信息抽取.

许斌(1973－),男,博士,副教授,博士生导师,CCF 高级会员,主要研究领域为知识图谱,数据挖掘,服务计算.

张鹏(1979－),男,工程师,CCF 专业会员,主要研究领域为知识图谱构建和应用,文本语义挖掘.

胡家威(1991－),男,工程师,主要研究领域为人工智能应用.

郑莉(1963－),女,教授,CCF 专业会员,主要研究领域为计算机应用.