Hans Journal of Data Mining 数据挖掘, 2020, 10(4), 282-302
Published Online October 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.104030

Hans汉斯

面向知识图谱的信息抽取

赵海霞[1,2]，李　磊[1,2]，吴信东[1,3*]，何　进[1,2]

[1]大数据知识工程教育部重点实验室(合肥工业大学)，安徽　合肥
[2]合肥工业大学计算机与信息学院，安徽　合肥
[3]明略科技集团，明略科学研究院，上海
Email: 1743022346@qq.com, lilei@hfut.edu.cn, *xwu@hfut.edu.cn, jinhe@mail.hfut.edu.cn

收稿日期：2020年9月16日；录用日期：2020年9月30日；发布日期：2020年10月13日

摘　要

随着大数据时代的到来，海量数据不断涌现，从中寻找有用信息，抽取对应知识的需求变得越来越强烈。针对该需求，知识图谱技术应运而生，并在实现知识互联的过程中日益发挥重要作用。信息抽取作为构建知识图谱的基础技术，实现了从大规模数据中获取结构化的命名实体及其属性或关联信息。同时，由于具有多样化的实现方法，扩充了信息抽取技术的应用领域和场景，也提升了对信息抽取技术研究的价值和必要性的认可度。本文首先以知识图谱的构建框架为背景。探讨信息抽取研究的意义；然后从MUC、ACE和ICDM三个国际测评会议的角度回顾信息抽取的发展历史；接着，基于面向限定域和开放域两个方面，介绍信息抽取的关键技术，包括实体抽取技术、关系抽取技术和属性抽取技术。

关键词

知识图谱，信息抽取，实体抽取，关系抽取，开放域

Knowledge Graph Oriented Information Extraction

Haixia Zhao[1,2], Lei Li[1,2], Xindong Wu[1,3*], Jin He[1,2]

[1]Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, Hefei Anhui
[2]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei Anhui
[3]Mininglamp Academy of Sciences, Mininglamp Technology, Shanghai
Email: 1743022346@qq.com, lilei@hfut.edu.cn, *xwu@hfut.edu.cn, jinhe@mail.hfut.edu.cn

Received: Sep. 16th, 2020; accepted: Sep. 30th, 2020; published: Oct. 13th, 2020

*通讯作者。

文章引用: 赵海霞, 李磊, 吴信东, 何进. 面向知识图谱的信息抽取[J]. 数据挖掘, 2020, 10(4): 282-302.
DOI: 10.12677/hjdm.2020.104030

Abstract

With the advent of the new era of big data, massive data constantly emerge. Therefore, the demand to find useful information and extract corresponding knowledge becomes intense. In response to this demand, knowledge graph technology came into being and has increasingly played an important role in achieving knowledge integration. Information extraction, as a basis for constructing knowledge graphs, obtains structured named entities with their attributes and relationships from large-scale data. This paper starts with the significance of information extraction in the context of knowledge graph construction. Then, from the viewpoints of the MUC, ACE, and ICDM conferences, this paper reviews the evolving history of information extraction. Next, this paper introduces closed domains and open domains oriented key technologies of information extraction, respectively, including entity extraction, relationship extraction and attribute extraction.

Keywords

Knowledge Graph, Information Extraction, Entity Extraction, Relationship Extraction, Open Domain

1. 引言

随着计算机技术和互联网的飞速发展和知识互联时代的到来，人们期寄着构建一个更加智能的、机器可理解可计算的万维网。知识图谱(Knowledge Graph)的概念逐渐出现在人们视野中。知识图谱在语义处理、开放处理等功能方面都显现出很强的能力，在智能推荐、问答和对话系统以及大数据分析和决策等应用中也体现出越来越重要的价值。知识图谱预计将在互联网知识互联的实现过程中起到中流砥柱的作用。

文献[1]给出了知识图谱的定义：知识图谱是一个用于描述物理世界中的概念及其联系的语义网络，它包含以下三个重要的因素：1) 概念。概念可以是实体、属性，也可以是一个事实，例如“一个人有两只手”。概念通常被描述为节点；2) 关系。关系是两个概念节点之间的语义联系，例如属性关系、拥有关系等；3) 概念和关系的背景知识。因为同一个概念和关系都有许多不同的表达方式，因此需要其背景知识作为提供查询的字典或者本体对多种表现形式进行连接。

知识图谱是知识工程在现今大数据阶段的一个标志性工具。知识工程是将人工智能的原理和方法[2]用于构建大规模知识库。知识工程创立者费根鲍姆(Feigenbaum)给出了知识工程的确切定义，即将知识集成到计算机系统从而完成只有特定领域专家才能完成的复杂任务[3]。知识工程从以图灵测试为代表的前知识工程开始，经历了以知识库、框架、推理机为核心的专家系统，Web1.0、群体智能 Web2.0 等发展阶段，随着 2012 年知识图谱概念的提出以及 Dbpedia、Freebase、YAGO 等知识库的建立，知识工程发展进入了一个新的发展阶段[4]，即大数据知识工程(BigKE)。大数据知识工程实现了对数据中的语义，包括隐含语义的挖掘，使数据成为了智慧数据(Smart Data)，其目标是自动或半自动地获取知识，融合碎片化知识，然后建立基于知识的系统[2]，最终达到为一众应用(例如，语义搜索系统、智能推荐系统、问答和对话系统以及大数据分析与决策)提供互联网智能知识服务的目的。

知识图谱的构建经历了人工构建和群体构建(众包)，现在自动构建技术成为了各个业界的研究热点[5]。知识图谱构建的两个基本构造是“实体－关系－实体”三元组和“实体－属性(值)”键值对的构建。实体通过它们之间的关系连接在一起形成图数据库[1]。知识图谱的构建从数据来源分类，可分为面向结构化数据、面向半结构化数据以及面向非结构化数据的知识图谱构建。本文主要介绍面向非结构化数据的知识图谱构建过程，以及应用的关键信息抽取技术。吴信东等人在文献[6]提出了大数据知识工程模型BigKE，实现了三层次的知识建模过程：首先对大数据进行三阶段处理，进行在线挖掘学习得到碎片化知识模型；接着对碎片化知识进行多个步骤的知识融合；最终实现以需求为导向的知识服务。因此，对应于 BigKE 提出的三层次过程，知识图谱的构建(Knowledge Graph Construction)技术按照自底向上的过程也包括三个层次：信息抽取(Information Extraction)、知识融合(Knowledge Fusion)和知识加工(Knowledge Processing) [7]。

基于大数据知识工程下知识图谱的构建，如吴信东等人在文献[8]提出的 HACE 定理所述，信息抽取可以描述为这样的一个过程：首先，第一阶段对大量孤立、模糊、复杂的动态非结构化数据进行初步处理和计算；然后，第二阶段对数据进行深层语义分析、用户隐私保护问题分析以及应用领域知识的结合分析；最后，第三阶段选择合适的挖掘算法和抽取技术进行数据抽取和融合[8]。通过将抽取得到的碎片化知识存入知识库的数据层和模式层，我们最终可以对数据形成本体化表达。这里的抽取技术又按照抽取过程分为实体抽取(Entity Extraction)、关系抽取(Relation Extraction)、属性抽取(Attribute Extraction)以及实体链接(Entity Linking)等[9] [10]。其中，实体抽取用于发现文本或者网页中的命名实体，并将其加入现有知识库中。关系抽取用于自动抽取实体之间存在的语义关系。属性抽取属于一种特殊的关系抽取。信息抽取的目标是自动化知识获取，即实现自动地从异构数据源中抽取实体、关系、属性等信息进而得到候选知识单元。

由于知识图谱的构建过程是通过以结构化形式描述客观世界中的概念、实体以及其关系开始的[11]，概念、实体、关系等信息提取的准确性对构建过程至关重要，信息丢失、冗余、重叠往往是知识图谱构建面临的最大挑战[1]。作为知识图谱构建的第一步，信息抽取是得到候选知识单元的关键。信息抽取的完整度、准确度直接显性影响后续知识图谱构建步骤的质量和效率以及最终知识图谱的质量。

面向知识图谱的信息抽取与传统信息抽取有很大区别。面向知识图谱的信息抽取大多面向开放域(Open Domain)而不再是限定领域(Closed Domain)。同时，随着维基百科(Wikipedia)等知识库的出现，知识图谱的数据源从有限的文本类型扩展为多源、异构、语义结构复杂的海量数据。因此，信息抽取的核心技术从单一的文本分析变为复杂的知识发现、知识链接等，并在新的应用场景和领域中对现有技术和实现方法提出了新的挑战问题。

信息抽取作为构建知识图谱的基础技术，实现了从大规模数据中获取结构化的命名实体及其属性或关联信息。同时，由于具有多样化的实现方法，扩充了信息抽取技术的应用领域和场景，也提升了对信息抽取技术研究的价值和必要性的认可度。

本文首先以知识图谱的构建框架为背景。探讨信息抽取研究的意义；然后从 MUC、ACE 和 ICDM 三个国际测评会议的角度回顾信息抽取的发展历史；接着，基于面向限定域和开放域两个方面，介绍信息抽取的关键技术，包括实体抽取技术、关系抽取技术和属性抽取技术。

2. 信息抽取研究的发展历史

2.1. 信息抽取相关概念

信息抽取系统是一种从大量信息源中迅速抛开无效信息找到有用信息的信息获取工具。关于信息抽

取的定义有以下几种。

定义 1 信息抽取的目标是从海量数据中，尤其是本文数据中，快速精准分析抽取出特定的事实信息(Factual Information)，将其转换成可理解可使用的结构化形式信息[12]，最后将条理的结构化信息存储在数据库中，等待下一步的分析利用。

定义 2 信息抽取是一种自动地从结构化(Structured Data)、半结构化(Semi-structured Data)或非结构化(Unstructured Data)数据中抽取概念、实体、事件，以及其相关的属性和之间的关联关系等结构化信息的技术[13]。

信息抽取带有一定的文本理解。可以看作深层的信息检索技术，也可以看作是简化的文本理解技术。信息抽取通常从两方面进行实现：一类是基于知识发现(Knowledge Discovery in Databases, KDD)和数据挖掘(Data Mining)的方法，通常处理结构化、半结构化的数据；另一类是基于自然语言处理(Natural Language Processing, NLP)和文本挖掘(Text Mining)的方法[12]，通常处理非结构化数据。信息抽取的具体方法可分为三类：第一类是基于规则(基于专家系统)的方法。主要在早期使用，使用人工编制规则，存在效率低，系统可移植性差等不可忽视的局限性；第二类是基于统计的方法，可在一定程度弥补第一类方法的缺点；第三类是基于机器学习的方法，，它大幅减少了人工干预，并具有处理新文本的能力，是目前常用的方法。

2.2. 信息抽取发展史

2.2.1. MUC 会议和 ACE 会议

到 20 世纪 80 年代末，由于消息理解系列会议(Message Understanding Conference, MUC)的召开，信息抽取技术开始飞速发展，逐渐进入蓬勃期，成为了自然语言处理领域的重要分支之一。

MUC 会议自 1987 年召开第一届起，一共进行了 7 届会议。会议由美国国防高级计划研究局 DARPA 资助，其主要目的是对信息抽取系统进行评测[14]，是典型的评测驱动会议。会前 MUC 组织会提供样例文本和抽取任务说明，参会单位进行信息抽取系统的开发。在会议召开时参会单位将对各自系统进行样例文本集合的测试，然后通过与手工标注结果进行对比，得到评测结果。最后在会议中对评测结果进行分享、交流、讨论。

MUC 会议在抽取任务中定义了模板、槽的填充规则以及模板填充机制，将信息抽取规定为模板填充的过程，模板填充即将抽取出的文本信息按照一定规则填入模板的相应槽中[12]。除此，会议还定义了一套完整的评价指标，由准确率(Precision)、召回率(Recall)、F1 值以及平均填充错误率(Error Per Response Fill, EPRF)等进行结果评价。

在会议的逐年开展过程中，信息抽取任务逐渐细化、复杂化：抽取模板由单一的扁平结构变为多个模板的嵌套结构；组成模板的槽，从 18 个、24 个到 47 个的逐渐增加；评测任务也在开始仅有的场景模板(Scenario Templates)填充任务上进行了命名实体识别(Named Entity Recognition)任务、共指消解(Coreference Resolution)、模板元素填充(Template Elements)、模板关系抽取和事件抽取等的任务扩充。

总之，MUC 会议的召开吸引了世界各地的研究者开始信息抽取系统的开发，在信息抽取研究的实践和理论方面都起到了极大的促进作用[15]，并确立了信息抽取的各种标准和规范，以及信息抽取技术的研究和发展方向。

继 MUC 之后，2000 年 12 月，由美国国家标准技术学会(NIST)、美国国家安全局(NSA)以及中央情报局(CIA)共同主管举办的自动内容抽取(Automatic Content Extraction, ACE)评测会议接着成为了信息抽取研究的又一巨大推动力，将信息抽取技术推向了一个新的高度。ACE 会议的研究内容是开发自动内容抽取技术，实现对不同来源的语言文本的自动处理，尤其对新闻语料中的实体、关系、事件进行自动识

别、抽取和描述。

和MUC相比，ACE不限定某个领域或场景[16]，增加了对系统跨文档处理(Cross-Document Processing)能力的评价，采用基于漏报和误报的评价体系。其中，“漏报”表示实际结果中存在而系统输出中没有；“误报”表示实际结果中不存在而系统输出中有。

2.2.2. ICDM2019 知识图谱比赛 KGC [6]

2019年IEEE国际数据挖掘大会ICDM (International Conference on Data Mining)举办了知识图谱构建比赛 KGC。该比赛由明略(Mininglamp)科学院和合肥工业大学主办，旨在对特定领域或多领域的非结构化文本进行自动知识图谱构建。该比赛的目的是生成类似人在阅读一段文字时的思维模式的知识图谱，因此比赛的评判由专家进行。比赛邀请了学位授予机构和工业实验室的团队参加，要求参与者首先设计模型，以文本作为输入，以知识图谱作为输出，从文本数据中提取知识三元组，并在比赛方提供的统一测试集上进行测试，若通过第一轮筛选，则进一步提供Web应用程序来可视化给定数据集的知识图谱。比赛规定知识图中的节点必须是文章中的实体词；链接必须是实体之间的关系词或属性；并且节点必须由原始文本中的单词或短语表示，且对同一单词的同义词进行合并。比赛的数据集是涵盖汽车工程、化妆品、公共安全和餐饮服务四个行业的300篇新闻短文本，其中120篇为专家预先进行手工标记的文章。

这个KGC比赛的新颖之处在于，没有为实体或关系预先提供任何类型的架构。除了ICDM 2019的KGC比赛，还涌现出了不少于信息抽取技术相关的国际学术会议，如国际信息和知识管理大会(International Conference on Information and Knowledge Management, CIKM)。

2.3. 性能衡量指标

在衡量信息抽取系统性能的指标中最常用的是准确率(Precision)跟召回率(Recall)。准确率指的是在抽取的所有结果中正确抽取结果所占的比例[17]；召回率指的是所有可能的抽取结果中正确抽取结果所占的比例[12]。通常两者的调和平均数F指数也常用于性能衡量，F指数的计算如下：

$$f-measure=\left(beta^2+1\right)\frac{presicion*recall}{\left(beta^2*precision\right)+recall}$$

其中beta是召回率和准确率的相对权重。beta的取值一般为1、1/2、2。当beta = 1/2时召回率的重要程度是准确率的2倍；当beta = 2时召回率的重要程度是准确率的一半；为1时两者则同等重要。

3. 信息抽取中的关键技术

3.1. 命名实体识别

3.1.1. 命名实体识别相关概念

除了一些众所周知的英文缩写，如IP、CPU、FDA，所有的英文缩写在文中第一次出现时都应该给出其全称。文章标题中尽量避免使用生僻的英文缩写。

实体(Entity)是世界上客观存在并可相互区分的对象或事物。实体根据其在现实世界中的自然划分，通常分为如下三大类七小类[18]：实体类包括人名、地名和机构名类三小类；时间类包括时间，日期两小类；数字类包括货币类和百分比类。

命名实体识别(Named Entity Recognition, NER)是信息抽取的第一步，是信息抽取中最为关键和重要的步骤。命名实体识别是从文本中识别出实体的命名指称。命名实体识别又称为“专名识别”、“实体抽取”[19]。实体识别包括两个步骤：实体边界识别和实体分类。边界识别的目的是判断字符串是否是一

个完整实体，实体分类将实体划分到预先设定的不同类别。命名实体识别可以看作是识别出表示命名实体的短语，并对其进行类型指定的过程。

实体识别通常与实体链接密不可分。实体识别负责指定实体类别，实体链接是将识别出的实体通过识别和消歧等步骤后与数据库中的实体进行对应。实体识别与链接将文本转换为结构化的、以实体为中心的语义表示形式，是问答系统、机器翻译、数据标注、句法分析的基础前提步骤[20]，是海量文本分析、知识图谱构建补全的“核心技术”之一。

3.1.2. 命名实体识别经典模型方法

命名实体识别技术方法分为基于规则、基于统计以及基于机器学习三类[21]。随着时代的变换更新，命名实体识别技术也在不断革新。从早期面向特定领域，逐渐发展为面向开放域(Open Domain)；从最初基于人工编写规则，使用启发式算法转变为基于条件随机场(Conditional Random Field, CRF)、最大熵(Maximum Entropy, ME)、K-最近邻(K-Nearest Neighbors)等统计机器学习的方法；从基于有监督学习逐渐变为弱监督学习，再到无监督机器学习方式。以下是一些经典的面向特定领域的实体识别方法：

1) 基于规则的实体识别方法

这类方法通常利用一组手工定义的规则，在文本中搜索与这些规则匹配的字符串，来抽取人名、地名、组织名等。其中，谢菲尔德大学提出了用于英语命名实体识别的 LaSIE-II 系统[22]较为经典。除此，文献[23]利用启发式算法与规则模板结合的方法首次实现了公司名称抽取系统对公司实体进行抽取。

这类方法依赖固定的词法(Lexical)、句法(Syntactic)和语义约束(Semantic Constraints)，准确率较高，但是需要依靠特定专家对特定领域的规则进行编写，存在领域性强，系统可移植性差等缺点。

2) 最大熵分类模型[24]

最大熵模型(Maximum Entropy)是一种概率估计模型，估计构建模型与已有训练集的效果相似度。其基本思想是选择创建一个模型使得其与给定的训练数据、训练样本产生效果尽可能一致。比如训练数据中命名实体前面的词为动词的概率为 50%，则最大熵模型得到的结果中命名实体前为动词的概率也要为 50%。最大熵模型的形式化描述如下：

$$c=\left\{p\in P\middle|p\left(f_i\right)=p'\left(f_i\right),i\in\left\{1,2,\cdots,n\right\}\right\}$$

其中，p' 表示样本经验分布，P 表示所有概率模型的集合[25]。

通过上述表达式可知，满足给定训练集的模型并不唯一，而最终寻找的是在约束条件下各种评价指标分布最均匀的模型，即最符合客观情况、具有最大熵的模型。

最大熵模型可以用于特征函数的生成、特征函数选取、参数估计，常应用于文本分类、数据挖掘、词性标注等问题。例如，MENE 系统采用最大熵模型实现英语命名实体的识别。MENE 使用和比较了多种特征，包括外部系统特征、分类字典特征等等，提高了系统的跨语言可移植性和系统性能，实现了将文档中的每个单词分类为人名、组织、位置、日期、时间、金钱价值、百分比或“以上都不是”。该系统可以用于 Internet 搜索引擎，机器翻译，文档自动索引，也可以作为处理更复杂的信息提取任务的基础[26]。

最大熵模型将实体识别的任务转换为子字符串的分类任务[11]。该模型的优点是结构紧凑，通用性较高，便于自然语言处理，但存在训练复杂度高，时间消耗和计算空间开销大等缺点[21]。

3) 隐马尔科夫模型

隐马尔可夫模型(Hidden Markov Model, HMM)是众多基于统计的模型中评价性能最佳的一种模型。

HMM 模型的基本思想就是给定观测序列(句子)，其数据是可以观测到的，通过捕获需要的状态转移信息，寻找观测值所对应的最佳状态序列(句子的标记序列) [26]，这类数据是隐藏的，无法直接观测。

HMM 模型采用了 Viterbi 算法[27]求取命名实体最佳标记序列(状态序列)，显著提高了模型的训练速度、识别效率，这是隐马尔可夫区别于其他模型的显著优势，但是 HMM 模型的准确率要比期望最大化(Expectation Maximization, EM)模型、CRF 模型低一些。因此 HMM 模型适用于实时性要求较高的场合，如语音识别、词性标注等领域。

HMM 由于其输出独立性假设，导致其不能考虑上下文的特征，限制了特征的选择。虽然之后提出了更为有效的最大熵马尔可夫模型(Maximum Entropy Markov Model, MEMM)，考虑了整个观察序列，但仍存在“标注偏置”(Label Bias)问题。

4) 条件随机场模型

Lafferty 等人[28]在 2001 年提出了条件随机场(Conditional Random Field, CRF)模型，它是一种判别式概率模型和一种序列分割及标记的可区分训练模型，其状态值取值的独立性不仅取决于临近的过去，也取决于未来，相对于 MEMM 和 HMM 更加有效。常用于分词、命名实体识别等预测问题。

CRF 模型将实体识别问题转化为序列标注问题。文献[29]提出并实现了从简短非正式的 Twitter 文章中进行命名实体识别，处理推文的命名实体识别所面临的信息不足以及训练数据不可用的挑战。文章提出在半监督的学习框架下，将 K 个最近邻(KNN)分类器与线性条件随机场(CRF)模型相结合，基于 KNN 的分类器进行预标记以收集整个推文中的全局粗略证据，而 CRF 模型进行顺序标记以捕获推文中编码的细粒度信息。

条件随机场模型为命名实体识别提供了一个特征灵活、全局最优的标注框架[21]，但是也存在收敛速度慢、训练时间长、依赖特征多的局限性。

5) 混合模型

基于规则的方法可移植性差，费时费力但是识别结果比较理想，基于统计机器学习的方法性能依赖于训练样本的规模，也出现了一定局限性。因此，出现了将两者相互结合的方法。Lin 等人[30]实现的是从自然语言文本中识别生物医学命名实体，提取生物医学信息。文章提出的识别方法分为两个阶段：先使用最大熵作为基础的机器学习方法；然后结合基于字典和基于规则的方法进行后处理，包括边界检测扩展和错误分类纠正。对 Medine 论文摘要的 GENIA 数据集进行了实体抽取测试，取得了较理想的结果，召回率和准确率都得到了提升。

6) 基于深度学习的方法

深度学习方法的训练是一个端对端的过程，无需人工定义相关特征[4]，其基本思想是使用训练数据学习对命名实体识别有用的特征，然后利用学习的特征在文本中进行命名实体识别。基于深度学习的方法主要有以下两类：

i. 神经网络－条件随机场架构(Neural Network-Conditional Random Field, NN-CRF) [31]

在这个架构中，卷积神经网络(Convolutional Neural Networks)/长短期记忆人工神经网络(Long Short-Term Memory, LSTM)用来学习每一个词的位置的向量表示，然后根据这个向量表示 NN-CRF 模型可以计算得到这个位置处的最佳标签。这类方法解决了实体识别的序列化标记问题。文献[32]提出了使用词向量表示特征的最简单、最有效的方法。文章[33]提出了一种半监督系统(以无监督的方式从大型语料库中学习单词表示，并使用这些单词表示作为有监督训练的输入特征，而不是使用手工制作的输入特征)，从 4 亿个 Twitter 微博中自动推断出的单词嵌入表示形式，作为系统输入，使用前馈神经网络(Feed Forward Neural Network, FFNN)进行分类，得到多种实体类别：公司、设施、地理位置、音乐艺术家、电影、人

物、产品。该篇论文提出的方法没有加入人工参与，专注于分布式单词表示，可以应用于不同的语料库，并且得到较好结果。最近，文献[34]提出了一种神经半马尔可夫(Neural Semi-Markov)结构的支持向量机模型，这是一种训练精度驱动的 NER 模型，该模型将实体抽取扩展到序列标记问题，引入了代价敏感学习(Cost-Sensitive Learning)来控制精度和召回率之间的折衷。

ii. 基于滑动窗口分类的方法

该方法使用神经网络学习句子中的每一个 N-Gram 的表示，然后预测该 N-Gram 是否是一个目标实体[4]。文献[35]实现了从科学文章中提取关键字短语并根据任务、材料、过程等方面对其进行分类的任务，该文章使用神经标记模型并引入基于图的半监督算法，将实体抽取归结为序列标记问题，对未标注的训练数据进行学习。这种方法对单一领域内、高数据量数据比使用跨域、小数据量数据具有更好的性能。

3.1.3. 面向开放域的实体抽取方法

在面向开放域的信息抽取中，信息来源不再是特定的知识领域，成为了全网信息，面向海量 Web 语料[36]。例如，KnowItAll 系统[37]处理的是规模庞大、具有异质性的 Web 语料库，例如 Twitter、Wikipedia 等。

开始研究人员采用人工方法进行实体识别和分类。例如，Sekine 等人[38]在 2002 年采用人工预定义实体分类体系首次展示了一个层次结构的命名实体分类框架，将全网的实体分为了 150 个种类。Ling 等人[39]在此基础上接着在 2012 年提出了 112 种的分类方法，该方法基于 Freebase 类型独特标记方法，先利用 Wikipedia 文件中的锚链接自动标记实体段，训练条件随机场模型，用来分割识别到的实体边界，接着采用自适应感知器算法实现对多类多标签实体的自动分类。

实体分类体系通过人工干预进行构建显得很是困难，因此，出现了通过统计机器学习方法从数据集抽取与当前类别实体具有相似上下文特征的实体，从而实现分类和聚类的方法。Jain 等人[40]提出并实例化了一种用于通过 web 搜索查询日志进行公开信息提取的新颖模型。该方法的处理对象是网页中的查询日志，通过应用基于模式的启发式方法和统计方法，使用无监督方法从搜索查询日志中提取实体，采用聚类算法对基于日志搜索得到的实体进行聚类，进而得到分类。这是一种面向开放域的无监督学习算法，该方法可以应用在协助搜索的关键字生成方面，例如搜索“手机”出现“华为”“小米”等建议。

由于传统统计模型需要进行大量语料标注、人工构造大量特征的局限性，出现了一些新方法，例如，使用基于半监督算法[41]、远距离监督算法[42]、基于海量数据冗余性[20]的自学习方法等来解决开放式实体抽取问题。面向开放域的实体抽取方法常应用于基于常识的新颖的问答系统[24]。

3.2. 关系抽取

命名实体识别是从文本中抽取特定实体，但仅孤立、离散的实体是无法得到语义结构无法满足应用需求的，这时候确立实体之间的关联关系显得更为重要。实体关系抽取是对已经识别出的实体进行预定义的关系识别，为更深层次的分析提供资源也是知识图谱构建的重要环节之一。

关系抽取是一种获取已经识别出的实体之间的语法或语义之间连接方式的技术。和命名实体识别类似，关系抽取中实体关系的类型也需要预先定义，例如人物之间的亲属关系、组织机构和地点之间的关系等等。

关系抽取的范围分为面向特定领域(Close Domain)、面向开放领域(Open Domain)以及联合推理三大类。面向特定领域的关系抽取方法和实体识别相似，前期主要使用基于模式匹配和基于词典驱动的方法，依靠人工编写抽取规则。随着人工构造规则低效性和领域局限性的明显化以及研究的深入，现在较多使用的两类方法是：基于机器学习(Machine Learning)的方法和基于本体(Ontology)的方法。其中，

基于机器学习的方法又分为有监督、弱监督和无监督三类。在面向开放域的关系抽取发展中出现了以OIE 系统为基础的多个系统，例如，WOE 系统、OIE ReVerb 系统、OILLIE 系统等，实现了动词、非动词的关系抽取和二元、多元的关系抽取。同时，为了解决隐含关系的抽取，产生了将面向开放域的关系抽取方法与传统面向特定领域的信息抽取方法相结合的联合推理的思想，是关系抽取方法上的一个巨大进步。

3.2.1. 基于机器学习(Machine Learning)的办法

基于机器学习的实体关系抽取方法的思想是：首先对人工标注的语料库进行不断学习不断训练，获取特定领域的信息抽取规则，接着利用机器学习算法进行关系识别。基于机器学习的实体关系抽取系统一定程度上可以处理新的文本，这是其区别于以往方法的最大优点。基于机器学习的方法根据是否需要人工标注训练集以及对标签的需求程度又分为有监督、弱监督和无监督三类[43]。

1) 有监督的学习方法

有监督的学习方法需要人工预先标注大量语料训练集以确保算法的有效性，然后对训练集进行不断学习获取信息抽取规则。关系抽取的有监督学习可以分为两大类：基于特征向量的方法和基于核的方法。

最早的有监督的学习方法是基于特征向量的学习方法。该方法将训练语料转换为特征向量形式，使用各种机器学习算法(最大熵模型(Maximum Entropy)、支持向量机(Support Vector Machine, SVM))为其构造分类器，从而对新数据进行分类和测试。基于特征向量的方法将信息抽取问题看作分类问题，对数据的正确分类即对信息的正确抽取。其研究重点是如何获取各种有效的词汇、语法和语义特征进行集成。Zhou 等人[44]使用支持向量机，运用了多种词汇、语法解析树、依存树特征，并且加入了各种语义信息，如 WordNet、名称列表 name list、分块短语信息等，实现了基于特征的关系提取，使用语言数据协会(Linguistic Data Consortium, LDC)[1]提供的 ACE 语料，抽取出了 ACE 2004 定义的 7 大类关系类型。这些基于有监督的学习方法发现实体的类别信息特征的提取有助于提高关系抽取性能。

基于核函数的方法以核函数理论为基础，以结构树为处理对象，通过直接计算两个离散对象(如语法结构树)之间的相似度来进行分类，不需要构造高维特征向量空间。核函数方法可以有效地利用句法树中的结构化信息，已成功应用于文本分类和生物信息学等问题。Liu 等人[45]借助 HowNet 提供的本体知识构造语义核函数，在开放数据集上对六类 ACE 定义的实体进行识别，准确率达到了 88%。Zhuang 等人[46]提出了使用卷积树核方法进行实体语义抽取，在关系的结构化信息中加入实体的语义信息，应用树裁剪策略，在减少冗余信息的同时扩充了原有的树结构，使之包含更丰富的实体语义信息。通过直接计算两个实体关系对象(即句法树)的相同子树的个数来比较相似度，也改善了实体语义关系识别抽取的效果。实验数据取自 ACE RDC 2004 中的 347 篇新闻报道，共有 4307 个关系实例，系统对 ACE 所定义的 7 个大类进行关系抽取实验。Zelenko 等人[47]在浅层句法分析树基础上定义了核函数，并设计了一个用于计算核函数的动态规划算法，然后通过支持向量机和表决感知器(Voted Perceptron)等分类算法来抽取实体语义关系，系统对 200 篇新闻文章(语料库包含来自不同新闻社和出版物(美联社，《华尔街日报》，《华盛顿邮报》，《洛杉矶时报》)进行处理，最终提取得到两种关系，“人员－隶属”关系(一个特定的人从属于一个特定的组织(如“小王是腾讯公司的程序开发工程师”中在人物“小王”和组织“腾讯公司”之间存在着人员－隶属关系)和“组织－位置”关系。

2) 弱监督的方法

弱监督学习方法又称为半监督学习，使用预先定义的关系类型和关系实例的种子来取代大量的人工

[1]https://www.ldc.upenn.edu/。

信息标注过程，减轻了对标签的依赖。在定义了适当的实体作为种子之后，利用机器学习方法，挖掘对应关系描述模式，通过模式匹配抽取新的关系实例。关系抽取的弱监督学习中基于 Bootstrap 算法、基于神经网络模型是经典的学习方法。

基于 Bootstrap 算法的半监督学习方法由 Carlson 等人[48]提出，该算法实现了自动实体关系建模，首先利用少量实例作为初始种子集合，通过 Pattern 方式迭代学习非结构文本以获取新实例，接着从新实例中继续学习并扩展 Pattern 集合。Wang 等人[49]以原始文本为输入，提出使用一个单一的模型、端到端联合识别边界、实体提及的类型和关系，使用了一种基于结构感知器的增量联合框架，利用有效的集束搜索进行实体和关系的抽取，该框架使用基于半马尔可夫链思想实现基于分段的解码算法。此后，Brin 等人[50]发布了 DIPRE 系统，该系统使用少量的种子模板，从网络上大量非结构文本中抽取实例，通过新的实例学习新的抽取模板，设计了一个永无止境学习者系统(Never-Ending Language Learner, NELL)，用来不间断抽取学习网络文本中信息到结构化知识库中，对数据库中的事实、知识不断扩充。NELL 主要学习的是两种类型的知识，一种是表示特定类别的词汇(比如，公司，家，学校)，另一种是表示特定关系的名词对(比如，表示所属关系的(小王，腾讯公司))。通过在前人抽取系统基础上进行大规模 Pattern 构建或完善对新抽取实例、新构建 Pattern 的描述限制，很多系统如 Snowball 系统[42]、NELL 系统[51]相继出现，推动了知识图谱的构建进度。

斯坦福大学(Stanford University)的 Mintz 等人[52]于 2009 提出基于远距离监督学习的无标注文本的关系抽取方法。该方法以 Freebase 为训练数据进行远距离监督学习，设计面向文本特征的分类器，是融合了有监督和无监督的信息抽取方法；何婷婷[53]提出了基于种子的自扩展命名实体关系抽取方法，选取有关系的命名实体对作为初始关系种子集合，通过弱监督学习扩展关系种子，接着计算关系种子和命名实体对之间的上下文相似度，进而抽取新的命名实体对。

3) 无监督的方法

无监督方法使用未经人工标注的训练文本集，通过实体对聚类的方法，构造分类器，给定实体间的关系。无监督学习主要利用语料中大量冗余信息进行聚类分析，进而得到实体间关系[5]。无监督方法既可以处理 web 文档也可以对文本文档进行处理。

无监督方法可以用来对 web 文档信息进行抽取。Kathrin [54]实现了基于无监督学习的 web 文档信息抽取，过程分为预处理、关系抽取和关系聚类三步；同样地，Etzioni 等人[37]实现了一个 web 信息抽取系统 KNOWITALL，通过无监督方法实现了高召回率(Recall)的信息抽取。

实体之间语义关系的抽取是 web 挖掘和自然语言处理，例如信息提取，关系检测和社交网络挖掘中各种任务的重要第一步。Hashimoto 等人[55]提出了一种词嵌入的方法对语义关系进行分类(监督学习)，词嵌入通过借助大型未标注语料库中特定关系的词汇特征来预测得到名词对中的特征，接着词嵌入用于构建特征向量，最终特征向量被训练成一个关系分类模型。Hashimoto 等人[55]使用原始 Wikipedia 文件中提取的 8000 万个句子作为训练数据进行词嵌入的预训练，最后将文本中的名词对之间的关系分为 9 个特定关系类(比如原因－结果、物质－来源)和 1 个其他关系类(例如，“养家糊口是人们努力赚钱的很大动力之一”中“养家糊口”－“赚钱”之间存在因果关系)。无监督方法也可以通过协同聚类算法实现。Bollegala 等人[56]提取了实体之间的语义关系，使用顺序联合聚类(co-clustering)算法，从未标记数据中提取大量有效关系，包括语义关系的双重关系(比如获取关系，房地产公司购买了一栋老洋房，同时可以表示为，老洋房被房地产公司收购)。该方法使用算法产生的聚类，训练了一个 L1 正则化逻辑回归模型识别用来描述聚类表达关系的模式[56]。其中提出的模型对 ENT 基准数据集中实体对之间的关系相似性进行了计算；对 SENT500 基准数据集的 500 个手动注释的句子中的四种语义关系进行了开放信息提取；以

及对包含 3500 万个节点的社交网络系统中 53 种不同的关系进行了识别和分类。

无监督方法可以用来对文本信息进行抽取。文献[57]通过将非结构化文本与知识库对齐来自动生成大量训练数据。文献[58]尝试将远程监督纳入文本处理中，以通过使语料和文本对齐来自动生成训练样本，从而提取特征训练分类器。

除了上述方法，Zhang 等人[45]提出了基于实例的无监督学习方法，能够对实体之间的雇佣关系、生产关系以及位置关系进行准确的识别；Ji 等人[59]提出了一个句子级别的注意力机制模型，该模型选择多个有效实例并充分利用知识库中的监督信息，使用传统 CNN 从 Freebase 或 Wikipedia 中抽取得到的实体特征信息来丰富实例的背景知识，提高实体表示。Qi 等人[45]使用 Riedel 2010 开发通过将 NYT 语料对齐知识库得到的数据进行实验。

4) 深度学习方法

深度学习方法在自然语言处理(NLP)和图像识别方面表现的性能非常强大，使得众多研究者将其应用于解决关系抽取的问题。深度网络的结构有很多种，如 RNN (Recurrent Neural Networks) [9]，CNNS (Convolutional Neural Networks) [3]，CNNs 和 RNNs 的结合结构[60] [61]以及 LSTMs (Long Short-Term Memories) [62]。基于神经网络模型不需要加入太多的特征，一般加入词向量特征、位置特征等就可以。Hsahimoto 等人[45]利用 Word Embedding 方法来学习给定标注预料中特定名词对应的上下文特征，将特征加入神经网络分类器中；JainPoon 等人[63]使用了用于关系提取的卷积神经网络(CNN)，针对不平衡语料库，自动从句子中学习特征并最大程度地减少对外部工具包和资源的依赖，从而摆脱了传统的复杂特征工程方法。该模型利用无监督框架自动训练词嵌入作为系统输入，模型使用预训练的词嵌入进行初始化，并优化词嵌入和位置嵌入作为模型参数，对句子中两个实体间的相对距离进行编码，并且提供了多种窗口大小的卷积过滤器，从而使网络适合于 n 元关系提取。从文本中提取实体对之间的语义关系可以用于信息抽取、知识库填充、问题解答等等。Zeng 等人[64]将分段卷积神经网络(PCNN)与多实例学习一起用于远程监督关系提取。此方法中，无需复杂的 NLP 预处理即可自动学习特征。Zhang 等人[65]提出了将 LSTM 序列模型与实体位置感知相结合的关系抽取神经序列模型，通过更好的监督数据和更合适的大容量模型的结合实现了更好的关系提取性能。

以上四种机器学习方法均可以对实体关系进行抽取。有监督的信息抽取方法需要预先人工标注大量语料集，对人工的依赖性较强，抽取的准确率较高，常常用来处理自然语言文本；弱监督学习减少了对标签的依赖，降低了对人工的依赖，其使用了预先定义的关系类型和关系实例的种子，实现了很多自动关系抽取模型，推动了知识图谱的构建进度；无监督方法使用的文本集不需要进行人工标注，它使用实体对聚类方法实现关系抽取。弱监督以及无监督学习常常用来处理规模大的 web 文本。深度学习方法通过引入神经网络模型进一步提升了关系抽取的自动化程度，并取得了更优秀的关系提取性能。

3.2.2. 基于本体(Ontology)的方法

基于本体的信息抽取技术，借助预定义的本体层次结构，可有效识别特定领域的概念、实体、关系等知识。本体可以看作一个呈树状结构的知识库模具，是同一领域内不同主体之间进行交流、连通的语义基础[66]。

本体的构建是信息抽取的基础，本体的构建方法也随着技术的发展逐渐从人工构建、半自动化构建向自动构建发展。人工构建本体由大量的领域专家相互协作完成，Swartout 等人[67]提出的循环获取法(CYC)，Nov 等人[68]提出的 Ontology Development 101 (七步法)都是人工构建的经典方法，其步骤包括确定领域范围、复用现有本体、列出概念术语、定义类与类之间的层次关系、定义属性之间关

系、定义属性的约束和创建实例。但是七步法存在主观性强，评价机制弱的缺陷，缺少科学管理和评价机制。

半自动化构建本体主要是利用相关领域内的专业词典、叙词表等专家知识从中抽取感兴趣的概念和关系，构建需要的实体[69]。这类方法复用了本体中的概念和关系带来了不同本体匹配的问题。

自动构建本体利用知识获取技术、机器学习方法以及统计的思想和技术从数据资源中自动获取本体知识。其具体方法分为基于语言规则和基于机器学习方法两类。基于语言规则的方法[70]，通过对自然域文本的分析，提取候选关系并将其映射到预定义的语义表示中实现本体的构建。这类方法中一个动词可以表示两个或多个概念之间的关系。但也存在以下缺点：1) 不会发现新的关系，只是发现已知关系实例；2) 本体构建的效果依赖于语义模式，因而需事先构建较完备的语义模式。另一类是基于统计分析的机器学习方法[71]，基于数据聚类对用于构建每个组的本体树的文档进行分组，使用模式树挖掘从部分本体树构建集成本体进行结构化的本体构建。其中，文档聚类主要通过潜在语义分析(Latent Semantic Analysis, LSA)和K-Means等检索关键字关系矩阵的方法来实现；本体构建主要通过形式概念分析和本体集成实现。机器学习方法比起基于规则的方法适用于范围更广的领域，构建的本体倾向于更好地描述概念间的关系，结构也更加复杂。但是，缺乏必要的语义逻辑基础，因此抽取概念关系松散且可信度无法得到很好的保证。信息抽取可以通过一个或者多个本体实现。Moreno [72]提出了在一个独立域中基于本体实现信息抽取的方法，应用面向分子生物学领域，对大肠杆菌信息进行抽取，建立大肠杆菌监管网络，所建设的系统对该领域科学论文的摘要和完整文献进行了测试，先设计领域本体，然后根据本体所包含的知识实现信息抽取。Li 等[73]人实现了基于农业本体的农业领域对结构化的 AJAX 数据的提取。Daya [74]提出了使用多个本体进行信息抽取，分别在子域的确定和子域的表达两种情况下使用多个本体，所实现的第一个基于多本体的系统是针对大学领域开发的，它使用两种专门针对子域的本体，语料库由 100 所大学，50 所来自北美和 50 所来自世界其他地区的网页组成文献。实现的第二个系统应用在恐怖袭击的领域和消息理解会议(MUC)使用的语料库实现子域的表达。

3.2.3. 基于开放域的关系抽取

随着大数据时代的来临，文本数据急剧增多，数据规模增大，传统的领域受限的、限制语义关系的信息抽取方法、知识表示结构出现了很大的局限性。之前的信息抽取方法面向的是特定数量的文本需要预先定义好的关系类别，领域知识也是由本体(Ontology)结构来表示，随着处理数据的海量化，本体构建越来越困难，抽取方法也开始出现问题。并且面向特定领域的抽取方法导致了信息抽取技术的难以普及和扩展，系统的可移植性差。

面向开放域的关系抽取技术直接利用语料库的中关系词汇进行实体关系分类建模，不再需要预先指定关系的分类，就可以实现数据分类。该方法成为了抽取模式上的一个巨大进步。开放式 IE 系统都采取标签－学习－提取三个步骤的方法：首先使用启发式或远距离监督方法自动标记句子；接着使用序列标记图形模型(例如 CRF)学习关系短语提取器；最后系统将一个句子作为输入，从句子中识别出参数，利用提取器将两个自变量之间的每个单词标记为关系短语的一部分或不作为关系短语的一部分。抽取器用于语料库中的连续句子，然后收集所得的抽取内容[11]。

华盛顿图灵中心的 Banko 等人[16] [75] [76]在 2007 年提出了面向开放领域的信息抽取框架(Open Information Extraction, OIE)，发布了基于自监督学习方式的开放信息抽取原型系统 TextRunner，标志着第一个 OIE 系统的问世。TextRunner (O-CRF)首先利用启发式规则来训练样本，然后采用二阶线性链条件随机场抽取器从开放式文本中自动抽取关系三元组[16]。TextRunner 可以自动抽取文本中大量实体关系，但是在准确率跟召回率方面不是很理想。

Wu 等人[77] 2010 年在 OIE 的基础上提出了基于 Wikipedia 的 WOE (Wikipedia-based Open Extractor)系统，将 Wikipedia 作为数据源利用维基百科网页信息框(Infobox)中的属性信息经自监督学习与相应语句匹配，自动构造实体关系训练集，然后从样本中抽取出关系独立的训练数据经自监督学习得到抽取器。WOE 系统实现了大批量构造高质量训练语料的方法，并且在准确率跟召回率方面都得到了改善，令人遗憾的是它速度方面出现了不足。Fader 等人[20]在 TextRunner 系统和 WOE 系统基础上引入了语法限制条件和字典约束，进行关系指示词的预识别，消除了不合理实体关系三元组的生成。

随着研究的进一步发展，出现了第二代 OIE 系统 ReVerb [20] [78]，基于通用句法和词法约束实现了关系短语识别器，处理的是随机抽取的英语句子，对其进行全面语言分析，使用动词表达句子中关系，抽取得到动词关系短语(例如，句子“Mr. Wang fought against Mr. Li，but finally lost the job”，系统将抽取出两组元组：(Mr. Wang, fought against, Mr. Li)和(Mr. Li, lost, the job))。Etzioni 等人[11]通过应用浅层句法约束和词性约束减少了无意义信息以及错误信息的产生，所设计的 Reverb 系统主要进行动词关系的抽取，先抽取满足约束的关系，然后依据临近原则确定左右实体。REVERB 支持学习选择偏好，获取常识知识，识别蕴含规则等等。

Mausam 等人[20]在第二代 OIE 基础上提出了支持非动词性关系抽取的 OILLIE (Open Language Learning for Information Extraction)系统，有效弥补了以往 OIE 系统抽取以动词为主而忽略名词形容词的缺陷，开始结合上下文全局分析而不是仅对语句局部分析、部分抽取，有效改善了自动抽取系统的召回率和准确率。McCallum 等人[75]提出了后期采用关系推理的方法，有效地提高了隐含语义关系的发现识别能力。

以上提到的抽取方法都是二元的开放式关系抽取。开放式的关系抽取按抽取关系的复杂程度可以分为二元和多元。Alan 等人[79]提出了基于 N 元关系模型的 OIE 系统，对除了常见二元实体关系的高阶多元实体关系进行识别；文献[79]在 OIE ReVerb 系统上提出了 KPAKEN 方法，通过输入 Stanford 的依存分析结果，经过检测事件短语、检测实体主导词、检测全部实体等步骤，实现了对任意英文语句中的 N 元实体关系的抽取。Del 等人[80]提出了一种新颖的基于条款的开放信息提取方法，称为 ClausIE，该方法从自然语言文本中提取关系及其参数，ClausIE 基于依赖性分析和一小组与域无关的词典，无需经过任何后处理即可逐句操作，并且不需要训练数据(无论是带标签的还是无标签的)。ClausIE 利用英语语法知识来首先检测输入句子中的从句，并随后根据其组成部分的语法功能识别每个从句的类型。根据此信息，ClausIE 能够生成高精度提取系统，在实验中使用了三个不同的数据集：包含手工标记的 500 句子的 Reverb 数据集；从 Wikipedia 页面中随机提取的 200 个句子；从《纽约时报》合集随机提取的 200 个随机句子。ClausIE 依据依存关系获取子句集合，并将其按类型灵活组合来抽取实体的 N 元关系。由于 N 元关系具有更加丰富的语义，因此由二元关系向 N 元关系的过渡是必然的，也是以后的研究发展方向。

随着理论研究的不断进行，更多面向开放域理论模型的出现，更优秀的知识表示结构的出现，更多研究成果正不断投入实践应用中，信息抽取研究正在不断取得进步，正在获得更大更开放的发展空间，为后续知识图谱的高质量构建提供了有力保障。

3.2.4. 联合推理

隐含关系抽取是关系抽取的一大难点。因此，为了挖掘文本中的隐含的深层语义信息，一些学者将面向开放域的关系抽取方法与传统面向特定领域(Close Domain)的信息抽取方法相结合，取长补短，提出了联合推理(Joint Inference)的概念[25]。JainPoon 等人[63]提出了一种完全联合方法。目前联合推理主要包括基于马尔科夫逻辑网和基于粗略至精细(Coarse-to-Fine)的本体推理两种。

1) 基于 Markov 逻辑网的逻辑推理

基于马尔可夫逻辑网 MLN (Markov Logic Network) [79] [81]的方法是联合推理关系抽取中的经典方法，该方法在 OIE 中加入了推理，将马尔可夫网络与一阶逻辑相结合，维护一个基于一阶逻辑的规则库，并对每一个逻辑规则附上权重，构建统计关系学习框架。其中马尔可夫逻辑是一种强大的新语言，将一阶逻辑与概率图形模型无缝结合[77]。MLN 的基本推理任务是寻找一个值从而使得可满足的子句的权值最大，即 MAP (Maximum A Posteriori)推理。MLN 可看作一种用一阶逻辑公式来实例化 Markov 网络的模板语言。该方法在语义角色标注、共指消解、文本蕴含、实体链接消歧等研究方面有很好的应用。

微软公司的人立方(Renlifang)项目基于该方法提出了 StatSnowball 模型[59]实现了自动生成或选择模板生成抽取器，从 web 挖掘实体关系，该模型在小型标记数据集和大规模 web 数据中都提现了较好的性能。该方法是一种基于无监督自学习的知识挖掘模型，可以抽取多种实体关系，并且可移植性强。人立方系统主要由以下几个应用：1) 搜索实体关系信息；2) 对话题相关人物进行排序；3) 检测某实体的受欢迎程度，并使用户可以浏览给定时间段内按其在网络上的知名度排名的不同类别的实体；4) 对人物进行排名。基于 StatSnowball 文献[82]提出了一种实体识别与关系抽取相结合的 ENTSum 模型，即将实体识别和关系抽取在一个模型中联合处理同时实现。该模型由扩展的 CFR 命名实体抽取模块和基于 StatSnowball 的 Bootstrapping 关系抽取模块组成，两个模块使用迭代方法相结合，实体识别可以利用关系抽取的模板语法特征和知识语义特征，使得两个模块准确率和召回率都得到了改善。文献[75] [83]提出了一种简易的 Markov 逻辑 TML (Tractable Markov Logic)。Banko 等人[78]提出了基于条件随机场的关系抽取模型(H-CRF)，根据目标数据集关系数量多少以及有无预定义的分类模型选择机器学习方法或开放域关系抽取方法。

2) 基于本体推理的联合推理

基于本体推理的联合推理面向开放域抽取方法形成的知识库基本上都是信息的基本存储并没有进行内容的规范和组织。为了使抽取结果形成的知识库成为真正的知识库，即能够推断文本深层含义进而从已有事实信息包含的隐含信息中推理出新的知识，能够为决策和问答所使用。研究者们提出了基于本体推理的信息抽取方法。

Zhang 等人[14]提出了 KOG 模型，该方法基于 MLN 联合推理，将 Wikipedia 的 Infobox 与 WordNet 相结合用于本体结构的构建，本体结构是“实体－属性－属性值”的结构，为 Wikipedia 的查询/专题浏览功能提供了辅助作用。Moro 等人[84]提出的 VELVET 方法利用联合推理以及本体平滑方法实现了最弱监督下实体关系的抽取，为结构化知识库的建立奠定了基础。Domingos 等人[85]将概率推理(Lifted Probabilistic Inference)与 Markov 相结合，提出了简易 Markov 逻辑(Tractable Markov Logic, TML)。在 TML 逻辑语言中，领域知识按照层次结构分为若干部分，各部分又按照所属事物类进一步分解为若干部分，以此类推，最终形成了一个层次化的类/局部结构。TML 被证明是目前最为丰富和高效的逻辑语言之一，可能将来在本体知识推理前进中起到推波助澜的作用。

另外一些学者提出了采用联合抽取模型的方法，典型成果如利用双层的 LSTM-RNN (长短期记忆－递归神经网络)模型通过神经网络进行分类模型的训练[64]联合推理结合了面向特定领域和面向开放域的方法，在许多方面展示出了优势。对于隐含关系的抽取和抽取阶段的平衡，联合推理方法显现出比主流开放式信息抽取方法更高的性能[86]。当前信息抽取技术多是顺序式抽取，即抽取过程分解为实体识别、关系抽取、属性抽取等连续的多个子任务再集成。这样的模式存在些缺陷，比如前一阶段无法识别的信息在后一阶段将不再被处理，从而出现了信息的缺失和不完整。前一阶段的错误信息结果将无法在后面

阶段进行修复，从而在所有阶段结束后大大增加了错误率的积累。此外顺序式处理方式使前面阶段无法使用后面阶段出现的有用特征，准确率和效率得到了限制。而联合推理方法不仅能够综合各个阶段，实现相互补充和促进，而且可以实现文本深层理解，实现隐含信息的自动推理。因此，联合处理的方法将成为之后的研究重点。

3.3. 属性抽取

属性抽取是为实体识别而服务的，属性可以很好的对实体进行刻画。实体的属性可以看作实体和属性值之间的名称性关系，因此实体属性抽取可以视为一种特殊的关系抽取。属性抽取的方法之一是从各类百科网站抽取结构化知识作为属性抽取的训练集，再将模型运用到开放域中的属性抽取[12]。例如，Domingos 等人[85]提出了基于规则与启发式算法的属性抽取方法，实现了从 Wikipedia 和 WordNet 的半结构网页中自动抽取相应属性名称与属性值，而且达到了很高的准确率。另一种方法是利用实体属性与属性值之间的关系模式直接从开放域的数据集上抽取实体属性[87]。Huang 等人[88]使用 DNN 架构的规则，模式和约束条件实现了从大量原始文件中提取给定实体的某些属性类型值即 Slot Filling (SF)的提取。

4. 信息抽取方法总结

信息抽取包括实体抽取、关系抽取、属性抽取等多个子任务。以下分别以应用领域、技术方法以及数据源为分类依据对提及的三个子任务分别进行了介绍。具体的方法和领域分类见表 1 和表 2。

Table 1. Models of NER
表 1. 实体识别方法

实体识别方法	策略	代表工作
开放域	半监督算法	[41]
	远距离监督算法	[42]
	两种实体分类方法	[38] [39]
特定领域	条件随机场	[21] [28] [29]
	最大熵	[24] [25] [26]
	隐马尔可夫模型	[27]
	神经网络条件随机场	[31] [32] [33] [34]
	滑动窗口	[35]

Table 2. Models of RE
表 2. 关系抽取方法

关系抽取方法	策略	代表工作
开放域	OIE 框架	[16] [75] [76]
	WOE 系统	[77]
	OIE ReVerb 系统	[11] [20] [78]
	OILLIE 系统	[20]
	N 元关系抽取	[79] [80]

Continued

联合推理	马尔科夫逻辑网	[59] [79] [81] [82]
	本体推理	[14] [84] [85]
特定领域	有监督学习	[44] [45] [46] [47]
	弱监督学习	[48] [49] [50] [52] [53]
	无监督学习	[54]-[59]
	基于本体	[67]-[74]

信息抽取方法根据处理信息来源的不同分为面向自然文本的信息抽取、面向 web 文本的抽取以及面向社交网络的信息抽取，如表 3。

Table 3. Source-based classification
表 3. 按处理对象分类

处理对象	代表工作
自然语言文本	[11] [23] [30] [35] [44] [46] [47] [74] [80]
Web 文本	[33] [37] [48] [52] [55] [59] [79]
社交网络文本	[56]

面向开放领域方法信息抽取方法应用范围广泛，可以很好的处理大规模数据，既可以处理自然语言文本，例如文献[80]提出的 ClausIE 模型，文献[11]提出的 REVERB 系统以及基于本体的系统[74]都是对文本进行信息抽取；又可以有效处理 web 文本，例如文献[79]提出 N 元关系抽取模型 KPAKEN 来对网络文本进行多元关系抽取。

在面向特定领域的信息抽取关系抽取方法中，基于有监督的抽取方法常用来处理自然语言文本，例如文献[47]提出基于核函数的系统，文献[46]提出使用卷积树核方法来对文本中的关系进行抽取，文献[44]使用了 ACE 语料作为输入来进行信息抽取，其数据规模较小，在人工标注预料训练集方面占有优势，通过学习训练集得到抽取规则因此准确率也较高；基于弱监督和无监督的抽取方法更多的用来处理大规模 web 数据，其减少了对于人工信息标注的需求，实现了对 Freebase、Wikipedia 等 web 文档的信息抽取，并且可以得到较准确的抽取效果，例如文献[37]基于无监督的机器学习方法提出 KNOWITALL 系统，对 web 文档进行实体和关系抽取，文献[48]基于弱监督机器学习方法 Bootstrap 对实体关系进行抽取，文献[52]使用 Freebase 为数据源进行基于远距离监督学习的无标注文本的关系抽取，文献[59]基于无监督方法提出的句子级别注意力级别模型，对 Freebase、Wikipedia 数据进行处理，文献[55]基于无监督方法提出的词嵌入方法处理 Wikipedia 文件中的信息。

在实体识别抽取中，基于规则以及基于统计的实体识别方法通常用来处理自然语言文本，其针对性强，准确率高，通常在人工标注下可以获得好的识别效果，例如文献[23]使用基于规则的方法实现了以公司名称为处理对象的，文献[79]将 K 最近邻(KNN)分类器与线性条件随机场(CRF)模型相结合实现了从简短非正式 Twitter 文章中进行命名实体识别，文献[30]使用混合模型将最大熵模型和基于规则的方法结合实现了从自然语言文本中识别生物医学命名实体；基于深度学习的方法无需人工定义相关特征通过训练数据自主学习有用特征然后利用特征进行命名实体识别，基于深度学习的方法既用来处理单领域自然文本，例如文献[35]以科学文章为处理对象使用神经标记模型实现从科研文章中提取关键字短语，深度学习也可以用来处理 web 数据例如文献[33]提出了一种半监督系统对 Twitter 微博进行

实体识别和分布式表示。

信息抽取的数据来源除了自然语言文本以及 web 文本这两种数据源外，社交网络数据也是一种丰富数据源。社交网络节点规模大且关系种类繁多，文献[56]提出了基于无监督方法使用顺序联合聚类算法对包含多个节点的社交网络中的多种关系进行抽取。

5. 结束语

本文首先根据知识图谱的概念、构建技术框架引出了信息抽取的概念，接着通过三个国际评测会议介绍了信息抽取的发展历史；后续详细介绍了信息抽取关键技术，包括实体抽取、关系抽取和属性抽取；最后分析了信息抽取的研究趋势。我们系统性分析了面向知识图谱信息抽取的常用方法，根据技术特点分为实体抽取、关系抽取以及属性抽取三类子任务。其中各个子任务根据其应用领域分为面向特定领域和面向开放域两种，根据其数据来源分为面向文本和面向 Web 两种。

在面向特定领域的情境下，信息抽取各个子任务的技术方法较成熟、经典，例如在实体抽取中常用 CRF、ME、HMM、NN-CRF 等基于统计的模型；在关系抽取中常使用基于监督、半监督或无监督的机器学习方法。

在面向开放领域的应用中，随着大数据时代、全网时代的到来，更多新的优秀的方法正在不断地涌现。具体地，在实体识别任务中，出现了一些基于自学习方法的实体分类模型，从而不再需要通过人工构造大量语料标注、大量的特征；在关系抽取中，出现了以 OIE 框架为基础的众多优秀系统，基本实现了各种词性间的关系抽取以及隐含关系的抽取。

基金项目

重点研发计划(YFB1000901)；国家自然科学基金(91746209 & 61673152)；教育部创新团队(IRT17R32)。

参考文献

[1] Wu, X.D., Wu, J., Fu, X.Y., Li, J.C., Zhou, P. and Jiang, X. (2019) Automatic Knowledge Graph Construction: A Report on the 2019 ICDM/ICBK Contest. 2019 *IEEE International Conference on Data Mining* (*ICDM*), Beijing, China, 8-11 November 2019, 1540-1545. https://doi.org/10.1109/ICDM.2019.00204

[2] Wu, X.D., He, J, Lu, R.Q., *et al*. (2016) From Big Data to Big Knowledge: HACE + BigKE. *Computer Science*, **42**, 3-6.

[3] Lin, Y.K., Shen, S.Q., Liu, Z.Y., *et al*. (2016) Neural Relation Extraction with Selective Attention over Instances. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*, Berlin, 7-12 August 2016, 2124-2133. https://doi.org/10.18653/v1/P16-1200

[4] China Chinese Information Society (2018) Language and Knowledge Computing Committee. Knowledge Graph Development Report. Higher Education Press, Beijing.

[5] Surdeanu, M., Tibshirani, J., Nallapati, R., *et al*. (2012) Multi-Instance Multi-Label Learning for Relation Extraction. *Proceedings of the* 2012 *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language*, Jeju Island, 12-14 July 2012, 455-465.

[6] Wu, X.D., *et al*. (2015) Knowledge Engineering with Big Data. *IEEE Intelligent Systems*, **30**, 46-55. https://doi.org/10.1109/MIS.2015.56

[7] Liu, Q., Li, Y., Duan, H., Liu, Y. and Qin, Z.G. (2016) Knowledge Graph Construction Techniques. *Journal of Computer Research and Development*, **53**, 582-600.

[8] Wu, X.D., Zhu, X.Q., Wu, G.Q., *et al*. (2014) Data Mining with Big Data. *IEEE Transactions on Knowledge and Data Engineering*, **26**, 97-107. https://doi.org/10.1109/TKDE.2013.109

[9] Socher, R., Huval, B., Manning, C., *et al*. (2012) Semantic Compositionality through Recursive Matrix-Vector Spaces. *Proceedings of the* 2012 *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, 12-14 July 2012, 1201-1211.

[10] Augenstein, I., Das, M., Riedel, S., *et al*. (2017) SemEval 2017 Task 10: ScienceIE—Extracting Keyphrases and Relations from Scientific Publications. CoRR abs/1704.02853.

[11] Etzioni, O., Fader, A., Christensen, J., *et al*. (2011) Open Information Extraction: The Second Generation. *Proceedings of the* 22*nd International Joint Conference on Artificial Intelligence*, Barcelona, July 2011, 3-10.

[12] Li, B.L., Chen, Y.Z. and Yu, S.W. (2003) Research on Information Extraction: A Survey. *Computer Engineering and Applications*, **39**, 1-5.

[13] Guo, X.Y. and He, T.T. (2014) Survey about Research on Information Extraction. *Computer Science*, **42**, 14-17.

[14] Zhang, C., Hoffmann, R. and Weld, D.S. (2012) Ontological Smoothing for Relation Extraction with Minimal Supervision. *Proceedings of the* 26*th AAAI Conference on Artificial Intelligence*, Toronto, Toronto, Ontario, Canada, 22-26 July 2012.

[15] Wikipedia (2019) Message Understanding Conference. https://en.wikipedia.org/wiki/Message_Understanding_Conference

[16] Banko, M., Cafarella, M.J., Soderland, S., *et al*. (2007) Open Information Extraction from the Web. *Proceedings of the* 20*th International Joint Conference on Artificial Intelligence*, Hyderabad, 6-12 January 2007, 2670-2676.

[17] Brin, S. (1998) Extracting Patterns and Relations from the World Wide Web. *Proceedings of Lecture Notes in Computer Science*, **1590**, 172-183. https://doi.org/10.1007/10704656_11

[18] Liu, L. and Wang, D.B. (2018) A Review on Named Entity Recognition. *Journal of the China Society for Scientific and Technical Information*, **37**, 329.

[19] Nadeau, D. and Sekine, S. (2007) A Survey of Named Entity Recognition and Classification. *Linguistics Investigations*, **30**, 3-26. https://doi.org/10.1075/li.30.1.03nad

[20] Fader, A., Soderland, S. and Etzioni, O. (2011) Identifying Relations for Open Information Extraction. *Proceedings of the* 2011 *Conference on Empirical Methods in Natural Language Processing*, *John McIntyre Conference Centre*, Edinburgh, 27-31 July 2011, 1535-1545.

[21] Sun, Z. and Wang, H.L. (2010) Overview on the Advance of the Research on Named Entity Recognition. *New Technology of Library and Information Service*, **26**, 42-47.

[22] Humphreys, K., Gaizauskas, R., Azzam, S., *et al*. (1998) University of Sheffield: Description of the LaSIE-II System as Used for MUC-7. *Proceedings of the* 7*th Message Understanding Conference*, Fairfax, 29 April-1 May 1998. https://www.aclweb.org/anthology/M98-1007/

[23] Rau, L.F. (1991) Extracting Company Names from Text. *Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications Piscataway*, Miami Beach, 24-28 February 1991, 29-32.

[24] RatnaParkhi, A. (1997) A Simple Introduction to Maximum Entropy Models for Natural Language Processing. Institute for Research in Cognitive Science, Technical Reports, University of Pennsylvania, Pennsylvania, 97-108.

[25] McCallum, A. (2009) Joint Inference for Netural Language Processing. P*roceedings of the* 13*th Conference on Computational Natural Language Learning* (*CoNLL*-2009), Boulder, June 2019, 1. https://doi.org/10.3115/1596374.1596376

[26] Zhang, X.Y., Wang, T. and Chen, H.W. (2005) Research on Named Entity Recognition. *Computer Science*, **32**, 44-48.

[27] Zhang, H.L. (2008) Visual C^{++} Digital Image Pattern Recognition Technology and Engineering Practice. People's Posts and Telecommunications Press, Beijing, 58-93.

[28] Lafferty, J., McCallum, A. and Pereira, F. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the* 18*th International Conference on Machine Learning*, Williamstown, 28 June-1 July 2001, 282-289.

[29] Liu, X.H., Zhang, S.D., Wei, F.R., *et al*. (2011) Recognizing Named Entities in Tweets. *Proceedings of the* 49*th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, Portland, Oregon, USA, 19-24 June 2011, 359-367.

[30] Lin, Y.F., Tsai, T., Chou, W.C., *et al*. (2004) A Maximum Entropy Approach to Biomedical Named Entity Recognition. *Proceedings of the* 4*th ACM SIGKDD Workshop on Data Mining in Bioinformatics*, Seattle, August 2004, 56-61.

[31] Xu, M.B., Jiang, H. and Sedtawut, W. (2016) A Fofe-Based Local Detection Approach for Named Entity Recognition and Mention Detection. *Computer Science*, *Computation and Language,* November 2016. arXiv:1611.00801v1 [cs.CL].

[32] Cherry, C. and Guo, H.Y. (2015) The Unreasonable Effectiveness of Word Representations for Twitter Named Entity Recognition. *Proceedings of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Denver, Colorado, USA, 31 May-5 June 2015, 735-745.

[33] Godin, F., Vandersmissen, B., Neve, W.D., *et al*. (2015) Multimedia Lab @ ACL W-NUT NER Shared Task: Named

Entity Recognition for Twitter Microposts Using Distributed Word Representations. *Proceedings of the Workshop on Noisy User-Generated Text*, Beijing, July 2015, 146-153. https://doi.org/10.18653/v1/W15-4322

[34] Arora, R., Tsai, C.T., Tsereteli, K., Kambadur, P. and Yang, Y. (2019) A Semi-Markov Structured Support Vector Machine Model for High-Precision Named Entity Recognition. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, 5862-5866. https://doi.org/10.18653/v1/P19-1587

[35] Yi, L., Mari, O. and Hannaneh, H. (2017) Scientific Information Extraction with Semi-Supervised Neural Tagging. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, September 2017 2641-2651.

[36] Zhao, J., Liu, K., Zhou, G.Y., *et al*. (2011) Open Information Extraction. *Journal of Chinese Information Processing*, **25**, 98-110.

[37] Etzioni, O., Cafarella, M., Downey, D., *et al*. (2005) Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, **165**, 91-134. https://doi.org/10.1016/j.artint.2005.03.001

[38] Sekine, S., Sudo, K. and Nobata, C. (2002) Extended Named Entity Hierarchy. *Proceedings of the* 3*rd Language Resources and Evaluation Conference*, New York, May 2002, 1818-1824.

[39] Xiao, L. and Weld, D.S. (2012) Fine-Grained Entity Recognition. *Proceedings of the* 26*th Conference on Association for the Advancement of Artificial Intelligence*, Menlo Park, 2012, Vol. 12, 94-100.

[40] Jain, A. and Pennacchiotti, M. (2010) Open Entity Extraction from Web Search Query Logs. *Proceedings of the 23th International Conference on computational Linguistics*, Stroudsburg, Beijing, August 2010, 210-518.

[41] Shi, B., Zhang, Z., Sun, L., *et al*. (2014) A Probabilistic Co-Bootstrapping Method for Entity Set Expansion. *Proceedings of the* 25*th International Conference on Computational Linguistics*: *Technical Papers*, Dublin, august 2014, 2280-2290.

[42] Agichtein, E. and Gravano, L. (2000) Snowball: Extracting Relations from Large Plain-Text Collections. *Proceedings of the* 5*th ACM Conference on Digital Libraries*, San Antonio, June 2010, 85-94. https://doi.org/10.1145/336597.336644

[43] Xie, D.P. and Chang, Q. (2020) View of Relation Extraction. *Application Research of Computers*, **7**, 1-5.

[44] Zhou, G.D., Su, J., Zhang, J., *et al*. (2005) Exploring Various Knowledge in Relation Extraction. *Proceedings of the* 43*rd Annual Meeting of the Association for Computational Linguistics*, Michigan, June 2015, 419-444.

[45] Qi, G.L., Gao, H. and Wu, T.X. (2017) The Research Advances of Knowledge Graph. *Technology Intelligence Engineering*, **3**, 4-25.

[46] Zhuang, C.L., Qian, L.H. and Zhou, G.D. (2009) Research on Tree Kernel-Based Entity Semantic Relation Extraction. *Journal of Chinese Information Processing*, **23**, 3. http://jcip.cipsc.org.cn/CN/abstract/abstract1128.shtml

[47] Zelenko, D., Aone, C. and Richardella, A. (2003) Kernel Methods for Relation Extraction. *The Journal of Machine Learning Research*, **3**, 1083-1106.

[48] Li, Q. and Ji, H. (2014) Incremental Joint Extraction of Entity Mentions and Relations. *Proceedings of the* 52*nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, 23-25 June 2014, 402-412. https://doi.org/10.3115/v1/P14-1038

[49] Whitelaw, C., Kehlenbeck, A., Petrovic, N., *et al*. (2008) Web-Scale Named Entity Recognition. P*roceedings of the* 17*th ACM Conference on Information and Knowledge Management*, Napa Valley, 26-30 October 2008, 123-132. https://doi.org/10.1145/1458082.1458102

[50] Carlson, A., Betteridge, J., Kisiel, B., *et al*. (2010) Towards an Architecture for Never-Ending Language Learning. *Proceedings in* 24*th AAAI Conference on Artificial Intelligence*, Atlanta Georgia, July 2010, 529-573.

[51] Mitchell, T. and Fredkin, E. (2014) Never-Ending Language Learning. 2014 *IEEE International Conference on Big Data* (*Big Data*), Washington DC, 27-30 October 2014, 1. https://doi.org/10.1109/BigData.2014.7004203

[52] Chang, X.L., Mi, X.M. and Muppala, J.K. (2013) Performance Evaluation of Artificial Intelligence Algorithms for Virtual Network Embedding. *Proceedings in Engineering Applications of Artificial Intelligence*, **26**, 2540-2550. https://doi.org/10.1016/j.engappai.2013.07.007

[53] He, T.T., Xu, C., Li, J., *et al*. (2006) Named Entity Relation Extraction Method Based on Seed Self-expansion. Proceedings in Computer Engineering, **32**, 183-184.

[54] Eichler, K., Hemsen, H. and Neumann, G. (2008) Unsupervised Relation Extraction from Web Documents. *Proceedings of the International Conference on Language Resources and Evaluation*, Marrakech, 26 May-1 June 2008.

[55] Hashimoto, K., Stenetorp, P., Miwa, M., *et al*, (2015) Task-Oriented Learning of Word Embeddings for Semantic Relation Classification. *Proceedings of the* 19*th Conference on Computational Natural Language Learning*, Beijing,

30-31 July 2015, 268-278. https://doi.org/10.18653/v1/K15-1027

[56] Bollegala, D.T., Matsuo, Y. and Ishizuka, M. (2010) Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web. *Proceedings of the* 19*th International Conference on World Wide Web*, *WWW* 2010, Raleigh, 26-30 April 2010, 151-160. https://doi.org/10.1145/1772690.1772707

[57] Quirk, C. and Poon, H. (2016) Distant Supervision for Relation Extraction beyond the Sentence Boundary. *Proceedings of the* 15*th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, 3-7 April 2017, 1171-1182.

[58] Mintz, M., Bills, S., Snow, R. and Jurafsky, D. (2009) Distant Supervision for Relation Extraction without Labeled Data. *Proceedings of the* 47*th Annual Meeting of the ACL and the* 4*th IJCNLP of the AFNLP*, Singapore, 2-7 August 2009, 1003-1011. https://doi.org/10.3115/1690219.1690287

[59] Ji, G.L., Liu, K., He, S.Z., *et al*. (2017) Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. *Proceedings of the* 31*st AAAI Conference on Artificial Intelligence*, San Francisco, 4-9 February 2017, 3060-3066.

[60] Guo, X.Y., Zhang, H., Yang, H.J., *et al*. (2019) A Single Attention-Based Combination of CNN and RNN for Relation Classification. *IEEE Access*, **7**, 12467-12475. https://doi.org/10.1109/ACCESS.2019.2891770

[61] Tran, V.H., Phi, V.T., Shindo, H., *et al*. (2019) Relation Classification Using Segment-Level Attention-Based CNN and Dependency-Based RNN. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 2-7 June 2019, 2793-2798. https://doi.org/10.18653/v1/N19-1286

[62] Zhou, P., Shi, W., Tian, J., *et al*. (2016) Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics* Berlin, 7-12 August 2016, 207-212. https://doi.org/10.18653/v1/P16-2034

[63] JainPoon, H. and Domingos, P. (2007) Joint Inference in Information Extraction. *Proceedings of the* 22*nd AAAI Conference on Artificial Intelligence*, Vancouver, 22-26 July 2007, 913-918.

[64] Zeng, D.J., Liu, K., Chen, Y.B., *et al*. (2015) Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, 17-21 September 2015, 1753-1762. https://doi.org/10.18653/v1/D15-1203

[65] Zhang, Y.H., Zhong, V. and Chen, D.Q. (2017) Position-Aware Attention and Supervised Data Improve Slot Filling. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 7-11 September, 2017, 35-45. https://doi.org/10.18653/v1/D17-1004

[66] Studer, R. (2008) Knowledge Engineering: Principles and Methods. *Data & Knowledge Engineering*, **25**, 161-197. https://doi.org/10.1016/S0169-023X(97)00056-6

[67] Swartout, B., Patil, R., Knight, K., *et al*. (1997) Toward Distributed Use of Large-Scale Ontologies. *Proceedings of AAAI*-97 *Spring Symposium on Ontological Engineering, Stanford University*, California, 1997, 138-148

[68] Noy, N.F. and McGuinness, D.L. (2001) Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, Stanford.

[69] Suryanto, H. and Compton, P. (2001) Discovery of Ontologies from Knowledge Bases. *Proceedings of ACM International Conference on Knowledge Capture*, Victoria, October 2001, 171-178. https://doi.org/10.1145/500737.500764

[70] Dahab, M.Y., Hassan, H.A. and Rafea, A. (2008) TextOntoEx: Automatic Ontology Construction from Natural English Text. *Expert Systems with Applications*, **34**, 1474-1480. https://doi.org/10.1016/j.eswa.2007.01.043

[71] Yu, Y.T. and Hsu, C.C. (2011) A Structured Ontology Construction by Using Data Clustering and Pattern Tree Mining. Proceedings of International Conference on Machine Learning and Cybernetics, Guilin, 10-13 July 2011, 45-50.

[72] Moreno, A., Isern, D. and López Fuentes, A.C. (2013) Ontology-Based Information Extraction of Regulatory Networks from Scientific Articles with Case Studies for Escherichia Coli. *Expert Systems with Applications*, **40**, 3266-3281. https://doi.org/10.1016/j.eswa.2012.12.090

[73] Li, C.X., Su, Y.R., Wang R.J., *et al*. (2012) Structured AJAX Data Extraction Based on Agricultural Ontology. *Journal of Integrative Agriculture*, **11**, 784-791. https://doi.org/10.1016/S2095-3119(12)60068-9

[74] Wimalasuriya, D.C. and Dou, D. (2009) Using Multiple Ontologies in Information Extraction. P*roceedings of the* 18*th ACM Conference on Information and Knowledge Management*, Hong Kong, 2-6 November 2009, 235-244. https://doi.org/10.1145/1645953.1645985

[75] Wu, F. and Weld, D.S. (2010) Open Information Extraction Using Wikipedia. *Proceedings of Annual Meeting of the Association for Computational Linguistics*, Uppsala, 11-16 July 2010, 118-127.

[76] Banko, M. and Etzioni, O. (2008) The Tradeoffs between Open and Traditional Relation Extraction. *Proceedings of the* 46*th Annual Meeting of the Association for Computational Linguistics*, Columbus, June 2008, 28-36.

[77] Domingos, P. and Lowd, D. (2009) Markov Logic: An Interface Layer for Artificial Intelligence. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, **3**, 155. https://doi.org/10.2200/S00206ED1V01Y200907AIM007

[78] Schmitzm, M., Bai, R., Soderiand, S., *et al*. (2014) Open Language Learning for Information Extraction. *Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, 12-14 July 2012, 523-534.

[79] Zhu, J., Nei, Z.Q., Liu, X.J., *et al*. (2009) StatSnowball: A Statistical Approach to Extracting Entity Relationships. *Proceedings of the* 18*th International Conference on World Wide Web*, Madrid, 20-24 April 2009, 101-110. https://doi.org/10.1145/1526709.1526724

[80] Del, C.L. and Gemulla, R. (2013) ClausIE: Clause-based Open Information Extraction. *Proceedings of the 22nd Internationa1 Conference on World Wide Web*, Rio de Janeiro, Brazil: WWW, 355-366.

[81] Miwa, M. and Bansal, M. (2016) End-to-End Relation Extraction Using LSTMs on Sequences and Tree Structures. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, 7-12 August 2016, 1105-1116. https://doi.org/10.18653/v1/P16-1105

[82] Liu, X.J. and Nen, H. (2010) People Summarization by Combining Named Entity Recognition and Relation Extraction. *Journal of Convergence Information Technology*, **5**, 233-241. https://doi.org/10.4156/jcit.vol5.issue10.30

[83] Suchanek, F.M., Kasneci, G. and Weikum, G. (2007) Yago: A Core of Semantic Knowledge. *Proceedings of the* 16*th International Conference on World Wide Web*, New York, May 2007, 697-706. https://doi.org/10.1145/1242572.1242667

[84] Moro, A. and Navigli, R. (2013) Integrating Syntactic and Semantic Analysis into the Open Information Extraction Paradigm. *Proceedings of the* 23*rd International Joint Conference on Artificial Intelligence*, Beijing, 3-9 August 2013 2148-2154.

[85] Domingos, P. and Webb, A. (2012) A Tractable First-Order Probabilistic Logic. *Proceedings of the* 26*th AAAI Conference on Artificial Intelligence*, Toronto, July 2012, 1902-1909.

[86] Xu, Z.L., Sheng, Y.P., He, L.R., *et al*. (2016) Review on Knowledge Graph Techniques. *Journal of University of Electronic Science and Technology of China*, **45**, 589-606.

[87] Wu, F. and Weld, D.S. (2007) Autonomously Semantifying Wikipedia. *Proceedings of the* 16*th ACM Conf on Information and Knowledge Management*, Lisbon, 6-8 November 2007, 41-50. https://doi.org/10.1145/1321440.1321449

[88] Huang, L., Sil, A., Ji, H., *et al*. (2017) Improving Slot Filling Performance with Attention Neural Networks on Dependency Structures. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 7-11 September 2017, 2588-2597. https://doi.org/10.18653/v1/D17-1274