DOI: 10.11992/tis.201805006
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20181223.1553.006.html

面向中文开放领域的多元实体关系抽取研究

姚贤明[1], 甘健侯[2], 徐坚[1]
（1. 曲靖师范学院 信息工程学院, 云南 曲靖 655011; 2. 云南师范大学 民族教育信息化教育部重点实验室, 云南 昆明 650500）

摘　要: 针对当前中文开放领域多元实体关系抽取研究较少的情况, 借鉴国外已有的研究成果, 结合中文自身的特点, 提出了中文领域多元实体关系抽取的方法。该方法以句法分析结果的根节点作为入口, 迭代地获取所有谓语的主语、宾语及其定语成分, 再利用句法分析结果对这些成分进行完善, 最终获取句子中的多个实体之间的语义关系。该方法被应用在不同的领域并进行了对比分析, 实验结果表明: 其具有一定的参考价值。另外, 对实验数据进行了详细的分析, 归纳了错误的主要情形, 为今后的研究工作指明了方向。

关键词: 中文、开放域; 多元实体关系; 依存句法分析; 句法结构; 关系抽取; 语义关系; 主谓宾
中图分类号: TP311　　**文献标志码**: A　　**文章编号**: 1673−4785(2019)03−0597−08

中文引用格式: 姚贤明, 甘健侯, 徐坚. 面向中文开放领域的多元实体关系抽取研究[J]. 智能系统学报, 2019, 14(3): 597–604.
英文引用格式: YAO Xianming, GAN Jianhou, XU Jian. Chinese open domain oriented n-ary entity relation extraction[J]. CAAI transactions on intelligent systems, 2019, 14(3): 597–604.

Chinese open domain oriented n-ary entity relation extraction

YAO Xianming[1], GAN Jianhou[2], XU Jian[1]
(1. School of Information Engineering, Qujing Normal University, Qujing 655011, China; 2. Key Laboratory of Educational Informatization for Nationalities (YNNU), Ministry of Education, Kunming 650500, China)

Abstract: In view of the scant research conducted regarding n-ary entity relation extraction in the Chinese open domain, in this paper, we propose a method for performing n-ary entity relation extraction in the Chinese domain based on existing research conducted abroad and Chinese characteristics. Starting with the root node of syntactic analysis, we obtain the subject, object, and attributive components of all the predicates. Then, we use the syntactic analysis result to perfect these elements and, finally, obtain the semantic relations of the n-ary entity. For comparative analysis, we applied the proposed method to different domains. The experimental results demonstrate its reference value. In addition, we analyzed the experimental data in detail and have summarized the main errors, which indicate the direction for future research.

Keywords: Chinese open domain; n-ary entity relation; dependency syntax analysis; semantic structure; relation extraction; semantic relation; subject predicate object

实体关系抽取是指从文本中抽取实体与实体之间, 实体与数值表达式之间的语义关系, 这种语义关系体现了二者之间的相互作用[1]。例如“邓兆祥游览庐山”, 其中“邓兆祥”与“庐山”之间存在“游览”关系[2]。

收稿日期: 2018−05−07.　**网络出版日期**: 2018−12−25.
基金项目: 国家自然科学基金项目 (61562093); 云南省应用基础研究计划重点项目 (2016FA024).
通信作者: 徐坚. E-mail: qjncxj@126.com.

实体关系抽取任务最早在 1989 年的 MUC 评测会议中被提出, 在 ACE、TAC 等一系列评测会议的推动下, 获得了长足的发展, 陆续提出了基于规则的[3-4]、基于支持向量机等有监督的[5-6]和基于聚类等无监督[7-8]实体关系获取方法[9], 本文称这些方法为传统方法。传统方法主要是面向特定领域, 预先定义了实体类型和关系类型, 通过人工标注训练数据提交给机器学习算法自动学习分

类规则，从而实现文本中实体关系的自动识别。

随着互联网的飞速发展，海量多源异构信息构成了互联网的主体，机器翻译、问答系统和知识库等应用系统的发展也逐渐面向互联网开展相关研究，传统方法已经不能满足现实的需求，因为在当前环境中，实体类型、实体关系都是未知的，虽然有部分文章提出了上百种实体类型[10-11]，对相关研究也产生了积极影响，但是仍然不能满足现实中千变万化的需求，因此开放域实体关系抽取任务被提出[9]。

开放域实体关系抽取的发展正在经历 2 个阶段[12-13]：二元实体关系抽取、多元实体关系抽取。

二元实体关系抽取主要以抽取动词为主，通常是从一句完整的语句中抽取到一对实体之间的关系。以 TextRunner[14]、KnowItAll[15]、WOE[16]、和 Reverb[17-18]等为代表的系统已推动了二元实体关系抽取接近成熟。采用的方法主要包括远程监督 (distant supervision) 和有监督的方法。远程监督[19-20]利用百科信息框的结构化信息对非结构化文本进行自动标注，训练识别模型，通过一定的技巧 (trick) 能达到较好的效果，该方法降低了人工标注语料的繁重负担；有监督的方法仍然以支持向量机等方法为主，但是在特征选择方面，通常选择句法、依存关系等具有领域通用性的特征[21-22]，从而使其模型具有跨领域能力。

多元实体关系指的是语句中多个实体之间存在的不同语义关系，因此多元实体关系抽取的任务是抽取这些实体之间的语义关系。相对于二元实体关系抽取来说，该任务具有更大的挑战性。目前，多元实体关系的抽取还处于探索阶段。文献[23]在构建 Kraken 系统的过程中，给出了多元实体关系抽取的基本思路如下：

1) 检测事件短语。Kraken 将动词、修饰词和介词视为事件。

2) 检测实体中心词。Kraken 从事件短语出发，根据 nsubject 等依存关系找到实体中心词。

3) 检测实体全称。Kraken 从实体中心词出发，递归地查找所有向下连接的实体词。

最终，Kraken 将实体全称和事件短语组合成三元组，并将其视为抽取到的实体关系。以句子“Doublethink, a word that was coined by Orwell in the novel 1984, describes a fictional concept.”为例，使用该方法可获得 3 个实体间的语义关系：

关系 1：(Doublethink, was coined, by Orwell)，关系 2：(Doublethink, was coined, in the novel 1984)，关系 3：(Doublethink, describes, a fictional concept)[23]。

从上面的结果可看出，相对于二元实体关系抽取仅仅只能抽取一对实体之间的语义关系而言，多元实体关系抽取能够抽取到更多的实体之间的关系。在英文中，多元实体关系占据了 40% 的所有实体关系[24]，因此，多元实体关系的抽取是实体关系抽取中一项十分重要的工作，而这也是今后实体关系发展的一个重要方向。目前，在英文的多元实体关系抽取方面已经取得了初步的研究成果[25-29]。

在中文领域，多元实体关系抽取方面目前鲜有提及，主要的工作集中在二元实体关系抽取[30-31]。本文以 Kraken 系统提供的方法为基础，结合中文自身的特点，提出了基于依存语法的开放域多元实体关系抽取方法，本文将该方法应用于民族、自然科学、法律、经济、人文历史 5 个领域以验证该方法的有效性，实验结果表明，该方法具有一定的参考价值。

1　中文多元实体关系抽取

在英文的实体关系抽取中，主要以谓语作为实体之间关系的指示词，因此在中文的关系抽取中沿用了该方法[32]。文献[2]以谓语作为关键字，构建上下文特征，训练识别器，实现了旅游领域的实体关系抽取，但是仍属于有监督的方法，而且针对的是二元关系抽取。文献[33]利用依存分析结果，结合启发规则实现三元组的抽取。文献[34]以句法分析结果作为基础，以动词为中心，抽取主谓宾结构，同时给出了句子中存在多个连续动词的复杂情况下，抽取主谓宾结构的解决方案，但是该文献没有详细给出存在零指代的情况下获取主语的方法。总体而言，在中文实体关系抽取方面，仍然缺乏针对复杂中文句子结构的有效实体关系抽取方法，在多元实体关系抽取方面更是缺乏相关研究。

从中文句法结构来看，主语、谓语和宾语构成了句子的主体，是描述事实的基本组成单元。语句可以是由一个主谓宾构成的简单句子，也可以是由多个主谓宾构成的复杂语句，复杂语句以动词作为事件链，表述了实体 (主语、宾语) 之间的语义关系。

以语句“1937 年 6 月 4 日，周恩来第一次登上庐山，入住仙岩旅馆，同蒋介石进行国共第二次合作谈判。”[2]为例，其中包含了 3 个连续事件：

登上→入住→进行, 对应的实体关系分别为: (周恩来, 登上, 庐山)、(周恩来, 入住, 仙岩旅馆) 和 (周恩来, 进行, 国共第二次合作谈判)。

对于语句“到 1910 年的时候, 美国科学家摩尔根, 他研究果蝇的遗传规律的时候发现, 遗传信息是位于染色体上面, 所以知道染色体跟遗传有非常大的关系”, 该句子的句法结构分析结果如图 1 和图 2 所示 (限于篇幅, 本文将句法分析结果分割成为两部分, 两部分的首尾以词语“发现”作为连接点)。从图中的结果可以看出, 各个单句之间不完全是以动词为主的连续链结构 (COO), 也包括以宾语 (VOB) 为主的连续链结构。直观上看, 可得出如下 2 个重要事实: 事实 1, (遗传信息, 位于, 染色体上面); 事实 2, (染色体, 跟遗传有, 关系)。另外更为重要的是, 这两个事实都是“德国科学家摩尔根”“发现”而“知道”的, 因此, 这里还存在另外一层实体与事实之间的关系 (德国科学家摩尔根, 发现, 事实 1) 和 (德国科学家摩尔根, 知道, 事实 2), 展开即为 (德国科学家摩尔根, 发现, (遗传信息, 位于, 染色体上面)) 和 (德国科学家摩尔根, 知道, (染色体, 跟遗传有, 关系))。除此之外, 该句中还存在另外一个实体关系 (德国科学家摩尔根, 研究, 果蝇遗传规律), 只是该实体关系隐藏在偏正结构中。

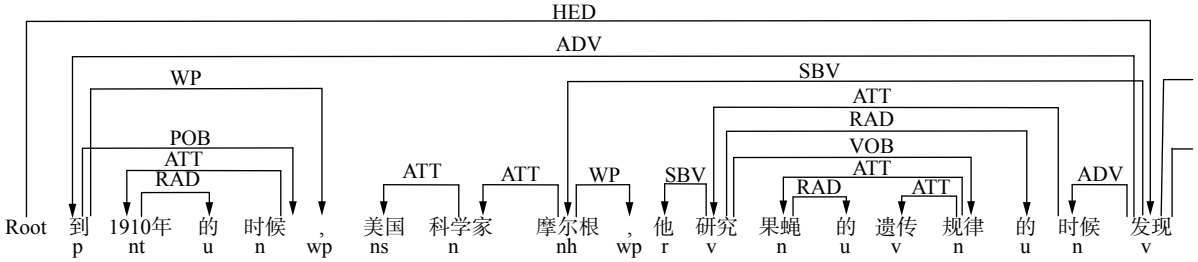

图 1　句子依存句法分析结果 (第 1 部分)

Fig. 1　Dependency parsing analysis result for example sentence (part 1)

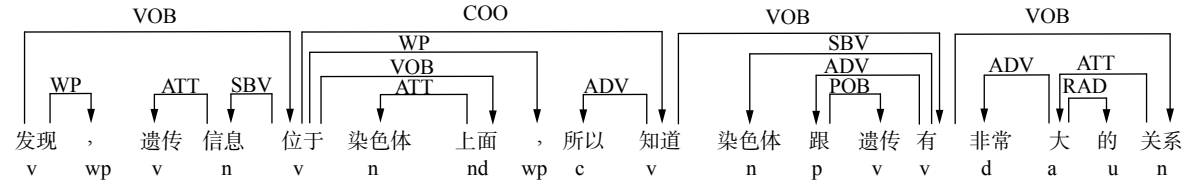

图 2　例句“到 1910 年的时候……”依存句法分析结果 (第 2 部分)

Fig. 2　Dependency parsing analysis result for sentence “Dao4 1910 nian2 de1 shi2 hou4” (part 2)

通过以上的分析可知, 实体关系在句子中呈现以下 3 个特点:

1) 实体关系在谓语上表现为连续链结构。所有连续的事件依照出现的先后顺序以 COO 相互连接。

2) 实体关系在实体本身也可能存在蕴含关系。作为实体关系中的实体, 可能为一实体名称, 也可能为一事件, 或者该事件本身又是一种迭代的结构。

3) 存在一些游离状态的实体关系。这些关系以松耦合的形式构成句子的一部分。

基于以上特点, 本文提出了基于依存语法的开放域多元实体关系抽取方法。该方法以哈工大 LTP 平台的句法分析结果作为依据, 抽取句子中以主谓宾结构为代表的实体关系, 具体算法步骤为:

1) 句法分析。将句子提交到 LTP 平台获取句法分析结果。

2) 事件链获取。获取句法分析结果中的 root 节点作为入口, 查找与该节点以 LTP 中定义的事件关系 (COO, IS) 相连接的动词并添加到动词集合。

3) 主谓宾获取。依据 LTP 定义的主语角色 (SBV) 和宾语角色 (VOB, IOB, FOB), 查找每个动词的主语和宾语集合 (可能存在多个主语和宾语并列的情形)。如果连接的角色是名词则将其添加到对应的主语或宾语集合, 并查找其他并列的主语或宾语; 如果连接的角色是动词, 则以该动词作为 root 节点, 并跳转到 2)。

4) 实体关系获取。将每次循环过程中获取到的主谓宾添加到实体关系集合中, 如果存在主语或宾语并列的情况, 则进行组合之后添加到实体关系集合中。

5) 顺序扫描句子中所有尚未在上述步骤中查

找到的动词，将其作为 root 节点，跳转到 2)。

6) 主语填充。利用一定的规则，对实体关系集合中缺乏主语 (零指代) 的主谓宾组合填充其主语。

7) 获取主语和宾语的定语部分。依据 LTP 平台定义的属性角色 (ATT) 获取主语和宾语的定语部分。

8) 输出所有以主谓宾形式表示的实体关系。

该算法的核心思想就是根据事件关系顺序和递归地查找所有的主谓宾结构，然后获取实体的修饰成分。值得注意的是，由于实体关系之间存在蕴含关系，一个事实可能为另外一个事实的成分，需要采用迭代的方式来获取，3) 中最后获取 root 节点就体现了该过程。该算法能够有效地获取句子中复杂的多元实体关系，避免无意义的实体关系对抽取结果的影响。

2　实验结果分析

作为目前比较新的研究方向，多元实体关系抽取尚缺乏权威的评测数据，在中文领域中目前亦如此。为了获得更加客观公正的测试结果，同时也为了验证算法的跨领域抽取能力，本文选取了历史、经济、民族、科技、法律 5 个领域的文本进行测试。其中经济和科技的文本属于口述性文本，民族领域文本来源于百度百科，法律文本则来自于法律条款，民族和法律领域的文本相对来说更加标准规范。

本文从这些领域文本中选取了部分具有代表性的句子作为评测数据，总共包含 167 个句子，其中包含多个实体关系的句子总数为 149 个，客观存在的实体关系总数为 408 对，抽取到的正确实体关系数量为 214 对，由此可见，该方法获取到的数量远大于二元实体关系抽取。

为了对具体领域的抽取效果有更直观的印象，本文采用信息抽取中常用的指标对系统性能进行评估，即正确率、召回率和 F 值。3 个指标的数据来源于上述选取的 167 个句子。每个指标在具体每个领域中的性能表现如表 1 所示。

表 1　本文算法在不同领域中的表现

Table 1　Performance of algorithm in this paper in different domains　%

指标	历史	经济	民族	科技	法律	平均
准确率	50	70	69	68	77	67
召回率	29	67	59	70	60	57
F 值	37	69	63	69	68	61

从表 1 中数据可以看出，总体的指标达到了 60% 左右，取得了一定的效果，也证明了本文中的方法具有一定可行性。在历史领域的文本中性能较差，但在其他领域中都有不俗的表现，而且在不同领域中的表现相对比较稳定，说明该方法具有一定的跨领域能力。

表 2 中列出了本文与其他文献开放域实体关系抽取的性能对比结果。其中，ZORE 是文献[35]中提出的 ZORE 系统，使用句法分析结果抽取中文开放域实体关系，与本文采用的方法类似，该系统的准确率等指标是性能最佳情况下的表现，该文献也是较早研究中文开放域实体关系抽取的工作之一；UnCORE 是哈工大秦兵教授在文献[30]中提出的面向大规模网络文本的无指导中文开放式实体关系抽取模型，在该文献中给出了正确率，但是因为文本规模较大，无法统计召回率，因此相关指标没有给出；Kraken 是文献[23]在英文领域抽取多元实体关系的性能表现，这也是英文多元实体关系抽取研究最早的工作之一。

表 2　与其他开放域实体关系抽取系统性能对比

Table 2　Comparisons with other open domain entity relation extraction systems　%

指标	ZORE	UnCORE	Kraken	本文
准确率	76.8	80	68	67
召回率	28.9		68	57
F 值	42		68	61

从表 2 中的数据可以看出，与 ZORE 相比，本文的召回率更高，体现出本文从文本中抽取到的实体关系数量更丰富，对于复杂句式效果更好，同时 F 值也更高。与 UnCORE 系统相比，本文的准确率不高，但是 UnCORE 系统是在大规模文本环境下运行的，数据的冗余性使得准确率得以提升，而召回率和 F 值这些指标却无法统计。与 Kraken 系统相比，本文所有的指标略有小幅下降，但是作为在中文领域中的一种尝试，本文得到这样的运行表现证明该方法具有一定的参考价值。

本文对实体关系抽取在不同领域错误的原因进行分析，对抽取到的实体关系的错误部分与未抽取到的实体关系进行了统计，将错误的原因大致分成 6 种情形，具体每种错误在不同领域中的占比如表 3 所示。

表 3　本文算法在不同领域中出现错误的原因及占比统计

Table 3　Case of errors and its proportion in different domain with method used in this paper　%

序号	错误情形	错误描述	历史	经济	民族	科技	法律	占比
1	情形 1	动词词性	13.25	18.52	17.34	14.49	61.19	27.30
2	情形 2	动词相邻	11.92	10.19	1.45	4.35	1.53	6.62
3	情形 3	实体在辅助结构中	49.01	41.67	31.79	50.72	2.29	32.96
4	情形 4	主语填充	7.28	8.33	3.18	2.90	0.57	4.60
5	情形 5	并列结构	14.57	12.04	26.01	5.80	28.68	19.41
6	情形 6	其他	3.97	9.26	20.23	21.74	5.74	9.10

从表 3 可看出, 导致错误的原因比较集中, 主要包括情形 1、情形 3 和情形 5, 占比总和达到了 79.67%, 这也为今后的工作指明了方向。对于每种错误的分析如下:

情形 1　动词词性导致的错误, 名动词被标注为动词, 导致名词性短语难以正确识别。本文使用的分词工具为哈尔滨工业大学 LTP 语言技术平台[36]本地工具包 (LTP4J), 工具中动词只包含一种类型“v”, 该词性分类体系与北京理工大学的 NLPIR[37]采用的计算所汉语词性标记集不同, 后者将动词 (v) 分成了 9 种 (vd、vn、vshi、vyou、vf、vx、vi、vl、vg) 类型, 每种类型的动词的作用更加清晰。本文使用 LTP 平台的主要原因是该平台具有句法分析、依存分析等功能, 同时该平台的分词能力在本文所使用的语料中表现更佳。

LTP 平台对句子“三线建设, 是 1964 年在毛泽东同志和中共中央的决策下进行的一场以战备为中心的经济建设战略”的词性标注结果为: “三线/j 建设/v, /wp 是/v 1964 年/nt……”, 从本例中可以看出本句的主语为“三线建设”, 但是由于“建设”的词性为“v”, 根据本文算法, 会继续寻找其主语, 从而导致主语“三线建设”识别失败, 但是如果将其标注为动名词“vn”, 则可以有效地提取到该主语。从表 2 中可以看出, 该问题导致抽取失败的占比达到了 27.30%, 其影响非常大。本文曾尝试使用 NLPIR 对该问题进行修复, 但是由于分词结果不同, 因此效果不佳。

情形 2　动词相邻, 在位置上前后紧密连接。以句子“毛泽东所说的‘屁股’, 是指基础工业”, 其分词结果为“……, /wp 是/v 指/v 基础/n 工业/v”。直观上说, “是指”可作为本句中的谓语, 然而由于在句法分析结果中二者是以 VOB 连接, 因此会以情形 1 中相同的方式进行处理, 从而导致错误的发生。

情形 3　实体词 (主语或宾语等) 在句子的附加结构中。以句子“从公元前 21 世纪以后, 相继出现了夏、商、西周几个王朝”为例, 本例中包含实体关系 (公元前 21 世纪以后, 出现, 夏王朝), 此处时间“公元前 21 世纪以后”虽然不是主语, 但是作为时间修饰成分, 同样也描述了基本的事实, 因此可作为实体关系而被抽取。但是在该句中, “从公元前 21 世纪以后”是作为 ADV 类型的状中结构存在, 本文采用的方法无法抽取到实体“公元前 21 世纪以后”这种时间类型的实体词。该情形是广泛存在于多元实体关系抽取中的问题, 在错误中的总占比为 32.96%, 同时由于本文之前尚未定义该类型实体关系的抽取规则, 几乎所有的实体关系都没有被检测到, 因此增加此类实体抽取规则将在很大程度上提升召回率。

情形 4　省略了主语情况下, 主语的自动填充结果带来的错误。在中文行文中, 省略语与指代是广泛存在的现象, 在实体关系抽取结果中占据非常大的比例。本文采用了简单的规则来弥补此问题: 在缺乏主语或存在代词的情况下, 向前一个语言片段寻找主语实体词, 将找到的第一个主语作为被省略的主语或代词的实体词, 如果在一个句子中前面位置找不到实体词, 则向后寻找。例如: 语句“汉族是中国的主体民族, 是上古时期黄帝和炎帝部落的后裔”, 该句第二个语言片段表达的是“汉族是上古时期黄帝和炎帝部落的后裔”, 但是“汉族”本身是前一个语句的主语, 通过本文的主语填充规则可以轻松地获得事实 (汉族, 是, 上古时期黄帝部落后裔) 和 (汉族, 是, 上古时期炎帝部落后裔)。该方法有效地降低了主语被省略的情况对实体抽取的影响。但是由于该方法过于简单, 也带来了一些错误, 如找到错误的主语, 或主语找不到的情况。

情形 5　实体词存在并列的情况。实体词并列出现的现象在文本中是广泛存在的, 存在几个并列关系就存在几种事实, 而本文的抽取规则尚

未完整地考虑到所有并列的情形，因此并列关系的存在对抽取结果产生了较大的影响。以句子“佤族主要分布云南省西南部的西盟、沧源、澜沧、孟连、双江、耿马、永德、镇康等县”为例，本句包含多个事实：(佤族，分布，云南省西南部西盟县)、(佤族，分布，云南省西南部沧源县)……(佤族，分布，云南省西南部镇康县)，总共 8 个事实，而本文的方法只能抽取到 (佤族，分布，云南省西南部西盟县) 这个事实，其余的 7 个事实则被忽略掉。由此可见，对并列结构中实体关系的抽取会极大地降低召回率。从表 2 中的数据也可以看出，其在总的错误中占比达到了 19.41%，因此提升空间是巨大的。

情形 6　其他原因，包括由于句子边界识别、未登录词、句子结构复杂等原因而导致的无法识别的情形。该情形在民族与科技领域中存在的情况比较常见。

3　结束语

作为在中文开放领域中多元实体关系抽取的一种尝试，本文从依存语法的角度出发，通过对句法分析的结果进行分析，抽取以主谓宾结构为代表的多元实体关系，并获得了一定的成效。同时本文对实验结果进行了分析，总结了导致抽取失败的 5 种主要情形，这也为今后的研究工作指明了方向。另外，本文只获取了主语和宾语的定语部分，但是对补语和状语没有进行抽取，这使得部分抽取结果理解比较困难，因此还需要进一步优化算法。

参考文献：

[1] CHINCHOR N, MARSH E. MUC-7 information extraction task definition (version 5.1)[C]//Proceedings of the Seventh Message Understanding Conference. Fairfax, Virginia, USA, 1998.

[2] 甘丽新, 万常选, 刘德喜, 等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展, 2016, 53(2): 284–302.
GAN Lixin, WAN Changxuan, LIU Dexi, et al. Chinese named entity relation extraction based on syntactic and semantic features[J]. Journal of computer research and development, 2016, 53(2): 284–302.

[3] JAYRAM T S, KRISHNAMURTHY R, RAGHAVAN S, et al. AVATAR information extraction system[J]. IEEE data engineering bulletin, 2006, 29: 1–9.

[4] SHEN W, DOAN A H, NAUGHTON J F, et al. Declarative information extraction using datalog with embedded extraction predicates[C]//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, Austria, 2007: 1033–1044.

[5] ZHAO Shubin, GRISHMAN R. Extracting relations with integrated information using kernel methods[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor, Michigan, 2005: 419–426.

[6] TRATZ S, HOVY E. ISI: automatic classification of relations between nominals using a maximum entropy classifier[C]//Proceedings of the 5th International Workshop on Semantic Evaluation. Los Angeles, California, 2010: 222–225.

[7] HASEGAWA T, SEKINE S, GRISHMAN R. Discovering relations among named entities from large corpora[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Barcelona, Spain, 2004: Article No.415.

[8] SHINYAMA Y, SEKINE S. Preemptive Information extraction using unrestricted relation discovery[C]//Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. New York, USA, 2006: 304–311.

[9] BANKO M, CAFARELLA M J, SODERLAND S, et al. Open information extraction from the web[C]//Proceedings of the 20th International Joint Conference on Artifical Intelligence. Hyderabad, India, 2007: 2670–2676.

[10] SEKINE S, SUDO K, NOBATA C. Extended named entity hierarchy[C]//Proceedings of the 3rd International Conference on Language Resources and Evaluation. New York, USA, 2002: 1818–1824.

[11] LING Xiao, WELD D S. Fine-grained entity recognition[C]//Proceedings of the 26th Conference on Advancement of Artificial Intelligence. Toronto, Canada, 2012: 94–100.

[12] 杨博, 蔡东风, 杨华. 开放式信息抽取研究进展[J]. 中文信息学报, 2014, 28(4): 1–11.
YANG Bo, CAI Dongfeng, YANG Hua. Progress in open information extraction[J]. Journal of Chinese information processing, 2014, 28(4): 1–11.

[13] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, 45(4): 589–606.
XU Zenglin, SHENG Yongpan, HE Lirong, et al. Review on knowledge graph techniques[J]. Journal of University of Electronic Science and Technology of China, 2016, 45(4): 589–606.

[14] YATES A, CAFARELLA M, BANKO M, et al. TextRunner: open information extraction on the web[C]//Proceedings ofHuman Language Technologies: the Annual

Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Rochester, New York, USA, 2007: 25–26.

[15] ETZIONI O, CAFARELLA M, DOWNEY D, et al. Unsupervised named-entity extraction from the Web: an experimental study[J]. Artificial intelligence, 2005, 165(1): 91–134.

[16] WU Fei, WELD D S. Open information extraction using Wikipedia[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, USA, 2010: 118–127.

[17] FADER A, SODERLAND S, ETZIONI O. Identifying relations for open information extraction[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, United Kingdom, 2011: 1535–1545.

[18] ETZIONI O, FADER A, CHRISTENSEN J, et al. Open information extraction: the second generation[C]//Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. Barcelona, Catalonia, Spain, 2011: 3–10.

[19] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the Afnlp. Suntec, Singapore, 2009: 1003–1011.

[20] 李杨. 中文开放式实体关系抽取研究与实现[D]. 成都: 电子科技大学, 2016.
LI Yang. Research and implementation of Chinese open entity relation extraction[D]. Chengdu: University of Electronic Science and Technology of China, 2016.

[21] BANKO M, CAFARELLA M J, SODERLAND S, et al. Open information extraction from the web[C]// International Joint Conference on Artifical Intelligence, New York, USA, 2007: 2670-2676.

[22] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. Barcelona, Spain, 2004: Article No.22.

[23] AKBIK A, LÖSER A. KrakeN: N-ary facts in open information extraction[C]//Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction. Montreal, Canada, 2012: 52–56.

[24] CHRISTENSEN J, MAUSAM, SODERLAND S, et al. An analysis of open information extraction based on semantic role labeling[C]//Proceedings of the Sixth International Conference on Knowledge Capture. Banff, Alberta, Canada, 2011: 113–120.

[25] HOFFART J, SUCHANEK F M, BERBERICH K, et al. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial intelligence, 2013, 194: 28–61.

[26] LING Xiao, WELD D S. Temporal information extraction[C]//Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. Atlanta, Georgia, USA, 2010: 1385–1390.

[27] WEIKUM G, NTARMOS N, SPANIOL M, et al. Longitudinal analytics on web archive data: it's about time![C]//Proceedings of the Fifth Biennial Conference on Innovative Data Systems Research. Asilomar, USA, 2011: 199–202.

[28] YADAV R, TANDAN S R. N-ary relation approach for open domain question answering system based on information extraction through world wide web[J]. International journal of engineering and applied sciences (IJEAS), 2015, 2(5): 141–144.

[29] BERRAHOU S L, BUCHE P, DIBIE J, et al. Xart: discovery of correlated arguments of n-ary relations in text[J]. Expert systems with applications, 2017, 73: 115–124.

[30] 秦兵, 刘安安, 刘挺. 无指导的中文开放式实体关系抽取[J]. 计算机研究与发展, 2015, 52(5): 1029–1035.
QIN Bing, LIU Anan, LIU Ting. Unsupervised chinese open entity relation extraction[J]. Journal of computer research and development, 2015, 52(5): 1029–1035.

[31] 赵军, 刘康, 周光有, 等. 开放式文本信息抽取[J]. 中文信息学报, 2011, 25(6): 98–110.
ZHAO Jun, LIU Kang, ZHOU Guangyou, et al. Open information extraction[J]. Journal of Chinese information processing, 2011, 25(6): 98–110.

[32] 王岁花, 赵爱玲, 马巍巍. 从 Web 中提取中文本体非分类关系的方法[J]. 计算机工程与设计, 2010, 31(2): 451–454.
WANG Suihua, ZHAO Ailing, MA Weiwei. Approach to extracting non-taxonomic relationships for Chinese ontology from web[J]. Computer Engineering and Design, 2010, 31(2): 451–454.

[33] 李明耀, 杨静. 基于依存分析的开放式中文实体关系抽取方法[J]. 计算机工程, 2016, 42(6): 201–207.

LI Mingyao, YANG Jing. Open Chinese entity relation extraction method based on dependency parsing[J]. Computer engineering, 2016, 42(6): 201–207.

[34] 古凌岚, 孙素云. 基于语义依存的中文本体非分类关系抽取方法[J]. 计算机工程与设计, 2012, 33(4): 1676–1681.
GU Linglan, SUN Suyun. Approach to Chinese ontology non-taxonomic relation extraction based on semantic dependency[J]. Computer engineering and design, 2012, 33(4): 1676–1681.

[35] QIU Likun, ZHANG Yue. ZORE: a syntax-based system for Chinese open relation extraction[C]//Processing of Conference on Empirical Methods in Natural Language. Doha, Qatar, 2014: 1870–1880.

[36] Che W, Li Z, Liu T. LTP: A Chinese Language Technology Platform[C] // Proceedings of the Coling 2010, Demonstrations. 2010, Beijing, China, 2010: 13-16.

[37] Zhang H, Yu H, Xiong D, et al. HHMM-based Chinese lexical analyzer ICTCLAS[J]. SIGHAN' 03 Proceedings of the second SIGHAN workshop on Chinese language processing, 2003, 17: 184-187.

作者简介：

姚贤明，男，1984 年生，讲师，主要研究方向为本体、知识图谱和问答系统。参与国家和省级科研项目 10 余项。发表学术论文 6 篇。

甘健侯，男，1976 年生，教授，博士生导师，主要研究方向为智能信息处理、计算机教育。主持国家自然科学基金项目 3 项，国家软科学项目 3 项，中央财政专项项目 1 项，教育部项目 2 项。发表学术论文 50 余篇，被 SCI、EI、CSSCI 检索 20 余篇，出版专著 3 部。

徐坚，男，1977 年生，副教授，主要研究方向为知识图谱、自然语言处理、教育信息化。发表学术论文 40 余篇，出版教材 4 部。