计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190358
57(7): 1424-1448, 2020

实体关系抽取方法研究综述

李冬梅 张 扬 李东远 林丹琼
(北京林业大学信息学院 北京 100083)
(国家林业草原林业智能信息处理工程技术研究中心 北京 100083)
(lidongmei@bjfu.edu.cn)

Review of Entity Relation Extraction Methods

Li Dongmei, Zhang Yang, Li Dongyuan, and Lin Danqiong
(*School of Information Science and Technology, Beijing Forestry University, Beijing* 100083)
(*Engineering Research Center for Forestry-oriented Intelligent Information Processing, National Forestry and Grassland Administration, Beijing* 100083)

Abstract There is a phenomenon that information extraction has long been concerned by a lot of research works in the field of natural language processing. Information extraction mainly includes three sub-tasks: entity extraction, relation extraction and event extraction, among which relation extraction is the core mission and a great significant part of information extraction. Furthermore, the main goal of entity relation extraction is to identify and determine the specific relation between entity pairs from plenty of natural language texts, which provides fundamental support for intelligent retrieval, semantic analysis, etc, and improves both search efficiency and the automatic construction of the knowledge base. Then, we briefly expound the development of entity relation extraction and introduce several tools and evaluation systems of relation extraction in both Chinese and English. In addition, four main methods of entity relation extraction are mentioned in this paper, including traditional relation extraction methods, and other three methods respectively based on traditional machine learning, deep learning and open domain. What is more important is that we summarize the mainstream research methods and corresponding representative results in different historical stages, and conduct contrastive analysis concerning different entity relation extraction methods. In the end, we forecast the contents and trend of future research.

Key words natural language processing; entity relation extraction; machine learning; deep learning; open domain

摘 要 在自然语言处理领域,信息抽取一直以来受到人们的关注。信息抽取主要包括3项子任务:实体抽取、关系抽取和事件抽取,而关系抽取是信息抽取领域的核心任务和重要环节.实体关系抽取的主要目标是从自然语言文本中识别并判定实体对之间存在的特定关系,这为智能检索、语义分析等提供了基础支持,有助于提高搜索效率,促进知识库的自动构建.综合阐述了实体关系抽取的发展历史,介绍了常用的中文和英文关系抽取工具和评价体系.主要从4个方面展开介绍了实体关系抽取方法,包括:早期的传统关系抽取方法、基于传统机器学习、基于深度学习和基于开放领域的关系抽取方法,总结了在不同历史阶段的主流研究方法以及相应的代表性成果,并对各种实体关系抽取技术进行对比分析.最后,对实体关系抽取的未来重点研究内容和发展趋势进行了总结和展望.

收稿日期:2019-06-10;修回日期:2020-01-14
基金项目:国家自然科学基金项目(61772078);北京市重点研发计划项目(D171100001817003)
This work was supported by the National Natural Science Foundation of China (61772078) and the Key Research and Development Program of Beijing (D171100001817003).

关键词 自然语言处理;实体关系抽取;机器学习;深度学习;开放领域

中图法分类号 TP18; TP391

在大数据时代,如何从海量的无结构或半结构数据中抽取出有价值的信息,引起了众多研究者的关注,促使这一领域的研究者投入更多的精力进行研究,信息抽取技术应运而生.信息抽取主要包括3项子任务:实体抽取(entity extraction)、关系抽取(relation extraction)和事件抽取(event extraction).而关系抽取作为信息抽取中的关键一步,近年来也受到学术界和工业界的广泛关注.关系抽取将文本中的无结构化的信息转化为结构化的信息存储在知识库中,为之后的智能检索和语义分析提供了一定的支持和帮助.研究人员利用关系抽取技术,从无结构化的自然语言文本中抽取出格式统一的实体关系,便于海量数据的处理;将分析出的多个实体之间的语义关系和实体进行关联,促进了知识库的自动构建;对用户查询意图进行理解和分析,提高了搜索引擎的检索效率等.综上所述,关系抽取技术不仅具有理论意义,还具有十分广阔的应用前景.

历经MUC(Message Understanding Conference),ACE(Automatic Content Extraction),TAC(Text Analysis Conference),SemEval(Semantic Evaluation)会议和OpenIE(open information extraction)技术的20多年发展,关系抽取的理论和方法愈加完善.从最初的人工设计模式和词典进行关系抽取,发展到目前借助传统机器学习、深度学习技术进行关系抽取,从单一领域关系抽取发展到开放领域关系抽取.随着关系抽取的正确率和召回率在不断提高,关系抽取模型对不同领域的适应性也在不断加强.

目前,关系抽取主要基于一种语言文本.事实上,人类知识蕴藏于不同模态和类型的信息源中,我们需要探索如何利用多语言文本、图像和音频信息进行关系抽取.这一领域仍然存在一些比较实际的问题阻碍了关系抽取在实际中的应用,这包括已标注数据集的获取、关系抽取模型的构建、共指消解等问题.随着这些问题的进一步解决,关系抽取技术必然会在增强检索系统功能、语义标注、本体学习等领域得到广泛应用.

1 实体关系抽取综述

1.1 实体关系抽取的发展历史

1998年在MUC-7[1]会议上第1次正式提出实体关系抽取任务.当时,这一任务主要利用模板的方式抽取出实体之间的关系,抽取的关系模板主要有location_of,employee_of,manufacture_of这三大类.在关系抽取方面,该会议主要以商业活动内容为主题,通过人工构建知识工程的方法,针对英语完成关系分类.研究人员利用Linguistic Data Consortium提供的New York Times News Service Corpus训练集和测试集构建关系抽取模型,并完成模型的性能评估.

由于MUC会议停办,ACE[2]评测会议替代MUC会议,继续专门针对多源文本的自动抽取技术进行研究.ACE会议指出,实体关系定义的是实体之间显式或者隐式的语义联系,因此需要预先定义实体关系的类型,然后识别实体之间是否存在语义关系,进而判定属于哪一种预定义的关系类型.该会议预先定义了位置、机构、成员、整体-部分、人-社会五大类关系,主要使用机器学习(有监督、半监督)的方法,针对英语、阿拉伯语、西班牙语等语言完成关系抽取任务.此外,会议提供了一定规模的标注语料(ACE04,ACE05)供大家研究,这为后续的研究提供了便利和支持.

此后,ACE会议于2009年并入TAC会议,同时将关系抽取任务并入KBP[3](Knowledge Base Population)会议.TAC是一系列评估研讨会,旨在促进自然语言处理和相关应用的研究.KBP是人口知识库,旨在提高从文本自动填充知识库的能力.TAC和KBP会议提供的大规模开源知识库(TAC-KBP),极大地推动了面向知识库构建过程中的关系抽取技术的研究和发展.

继MUC和ACE会议之后,SemEval会议[4]在自然语言处理领域受到了广泛关注.SemEval会议的前身是在1997年由ACL-SIGLEX组织成立的Senseval.Senseval是国际权威的词义消歧评测会议,其潜在的目标是增进人们对词义与多义现象的理解.之后,除词义消歧之外,由于其他有关语义分析的任务也越来越多,因此Senseval委员会决定把评测名称改为SemEval,并于2007年组织了SemEval2007评测.该会议聚焦于句子级单元间的彼此联系、语句间的联系以及自然语言(情感分析、语义关系)等.SemEval会议定义了最初9种常见名词及其关系(原因-影响、仪器-机构、产品-生产者、含

量-包含者、实体-来源地、实体-目的地、部分-整体、成员-集合、行为-主题),采用传统机器学习或者深度学习的方法完成英语、中文等语言的词义、语义的消歧任务,最终对数据库中的关系种类进行扩充.此外,该会议提供了 SemEval-2010 Task 8 数据集,逐渐掀起了研究人员对实体关系抽取研究的高潮,发展成为规模空前、极具影响力的评测会议.

权威评测会议 MUC, ACE, TAC, SemEval 为传统的关系抽取提供了评测语料.这些领域由专家人工标注和构建的评测语料库具有较高的质量和公认的评价方式,因此有力地引导和推进了传统的关系抽取研究的发展,大幅度地提升了关系抽取性能.由于传统关系抽取基于特定领域、特定关系进行抽取,导致关系抽取这一任务耗时耗力,成本极高,同时不利于扩展语料类型.近年来,针对开放领域的实体关系抽取方法逐渐受到人们的广泛关注.

研究者利用 Wikipedia, HowNet, WordNet, FreeBase 等涵盖大规模事实性信息的知识库解决了语料获取困难的问题,为关系抽取任务提供了有效的数据支持.与传统的人工标注语料的方法相比较,基于 Web 开放语料的规模更宏大,涉及的领域更广阔,涵盖的关系类型也更丰富,并不需要事先对关系进行定义.为了解决互联网海量数据的文本挖掘和分析任务,越来越多的研究者开始研究 OpenIE 技术[5].而开放领域的实体关系抽取作为其中的重要子任务和关键技术,自然也受到了研究者的广泛关注.研究人员无需事先指定关系的定义方式,可以采用深度学习和模式匹配结合的方法,针对开放领域完成实体关系抽取任务.该类方法提高了关系模型的可移植性和扩展性,能够通过迁移学习(transfer learning)等方式应用于其他领域.实体关系抽取的研究趋势和关键会议如表 1 所示:

Table 1 History of Entity Relation Extraction

表 1 实体关系抽取的发展历史

Main Conferences Research Trends	Relation Definition Mode	Research Content	Research Method	Datasets
MUC	3-class Relation	Business Activity Content	Knowledge Engineering	New York Times, News Service Corpus
ACE	5-class Relation	5-class Relation Extraction	Machine Learning (Supervised, Semi-supervised)	ACE04, ACE05
TAC	Relational Set in Knowledge Base	Knowledge Base Construction	Machine Learning, Pattern Matching	TAC-KBP
SemEval	9-class Relation	Word Sense and Semantic Disambiguation, etc.	Machine Learning, Deep Learning	SemEval-2010 Task 8
OpenIE	Not Specified in Advance	Open Domain	Deep Learning, Pattern Matching	OpenIE

1.2 关系抽取定义

在自然语言处理领域,关系通常主要指代文本中实体之间的联系,如语法关系、语义关系等.通常将实体间的关系形式化地描述为关系三元组$\langle E_1, R, E_2\rangle$,其中 E_1 和 E_2 指的是实体类型,R 指的是关系描述类型.实体关系抽取的主要目的是从自然语言文本中识别并判定实体对之间存在的特定关系.文本经过命名实体识别、关系触发词识别 2 个数据预处理过程,将判定的三元组$\langle E_1, R, E_2\rangle$存储在数据库中,供进一步的分析或查询.

基于以上的定义,可以直观地将关系抽取任务分成 3 个关键的模块,即为命名实体识别和触发词识别 2 个预处理模块以及关系抽取模块.关系抽取系统框架如图 1 所示[6]:

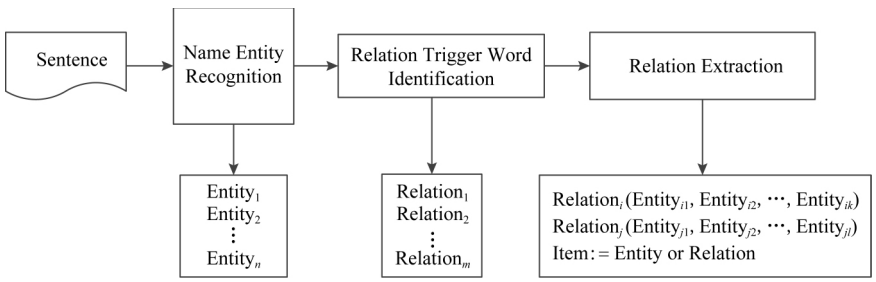


Fig. 1 The general framework of a relation extraction system

图 1 关系抽取系统框架

1) Name entity recognition,即命名实体识别,是指识别文本中具有特定意义的实体,主要包括人名、地名、机构名、专有名词等;

2) Relation trigger word identification,即关系触发词识别,是指对触发实体关系的词进行分类,识别出是触发词还是非触发词,判定抽取出的关系是正类还是负类;

3) Relation extraction,即关系抽取,是指从识别出的实体中抽取实体间的语义关系,如地点、雇员、产品等.

以句子"姚明出生于上海"为例,首先对句子进行预处理,识别出命名实体"姚明"和"上海",然后"出生于"作为关系触发词表明这2种实体之间可能存在某种关系,最后通过关系抽取模型的判定,得出2个实体之间存在着"地点"这一关系.

1.3 关系抽取特点

关系抽取是一个文本分类问题,相比于情感分类、新闻分类等其他任务,关系抽取主要有3个特点.

1) 领域众多,关系模型构建复杂.针对一个或者多个限定领域的关系抽取的研究时间较长,研究者投入的精力相对开发领域多,因此方法众多,技术成熟.由于限定了关系类别,可采用基于规则[7-13]、词典[14-17]以及本体[18-20]的方法,也可采用传统机器学习的有监督[21-35]、半监督[36-48]以及无监督[49-57]方法,深度学习的有监督[58-89]、远程监督[90-99]方法.这类方法的模型构建难度相对于开放领域难度较低,但是移植性和扩展性较差.而针对开放领域的关系抽取[100-117],由于关系类型多样且不确定,可以采用无监督和远程监督等方法.

2) 数据来源广泛,主要有结构化、半结构化、无结构3类.针对表格文档、数据库数据等结构化数据,方法众多,现通常采用深度学习相关[65]的方法等;针对纯文本的无结构数据,由于无法预料全部关系类型,一般采用以聚类为核心的无监督方法[50-55]等;而针对维基百科、百度百科等半结构化数据,通常采用半监督[52-53]和远程监督方法[92-93]等.

3) 关系种类繁多复杂,噪音数据无法避免.实体之间的关系多样,有一种或多种关系,早期方法主要针对一种关系(忽略重叠关系)进行抽取,这类方法忽略了实体间的多种关系,对实体间的潜在关系难以处理.近年来,图结构[73-77,88-89]逐渐应用于关系抽取领域,为关系重叠和实体重叠提供了新思路.而针对噪音数据,Bekoulis等人[86]发现少量对抗样本会避免模型过拟合,提出使用对抗训练提高模型的性能.

1.4 关系抽取常用工具

实体之间的关系一般用文本的句法特征和语义特征来表示,因此需要对文本进行分析.下面主要介绍国内外性能比较稳定且广受关注的文本分析工具.

1.4.1 英文关系抽取常用工具

1) NLTK(natural language toolkit)[118]

2009年宾夕法尼亚大学计算机和信息科学系实验室里开发了NLTK.NLTK是一个基于脚本语言Python开发的自然语言处理工具包,该工具包具有免费、开源等特点,并集成了中文分词、词形还原、文本分类以及语义推理等一系列文本处理技术,并涉及50多种语料库和词汇资源的交互界面,促进了研究人员对自然语言处理领域的开发和研究.

在关系抽取方面,研究人员通过该工具包提供的文本分析、文本分类等功能对文本进行预处理,进而对句子结构和语法特征进行分析,推断句子中实体之间是否存在的语义联系.

2) DeepDive[119]

2014年斯坦福大学发布了DeepDive.它是一种新型数据管理系统,可以在单个系统中解决提取、集成和预测问题.相对于其他关系抽取工具,DeepDive使研究者关注重点在实体关系之间的特征而不是具体的算法,这有效地减轻了研究者的工作负担.此外,DeepDive是一个性能良好的系统,使用机器学习消除各种形式的噪音和不精确数据.在科学领域,DeepDive抽取复杂知识的表现优于人类志愿者,特别是在实体关系抽取比赛中取得了较好的成绩.

3) Stanford CoreNLP[120]

2014年斯坦福大学自然语言处理研究小组在第52届国际计算语言学协会(The Association for Computational Linguistics, ACL)发布了一系列较为成熟的自然语言处理工具包Stanford CoreNLP.该工具包由众多语法分析工具集成,提供多种编程语言的接口,能实现对任意自然语言文本进行分析.该工具包为研究者提供了许多基础性的工具,如词性标记器(POS)、命名实体识别器(NER)、解析器、共参考分辨率系统、情感分析、自举模式学习和开放信息提取等.研究者利用这些工具包,可以根据短语和语法依赖来标记句子的结构、发现实体之间的关系、分析出句子所表达的情感等.

1.4.2 中文关系抽取常用工具

1) 中文分词工具

和英语等语言相比较,中文具有较大的差异,如中文词语之间没有空格,因此对中文的关系抽取任务首先需要进行中文分词.结巴分词(jieba)、清华分词(THULAC)、中国科学院计算技术研究所分词(NLPIR)、哈尔滨工业大学分词(LTP)等是国内常见中文分词的工具.这些工具对文本数据进行预处理,将字序列切分成具有语言含义的词序列,便于对中文领域的文本进行关系抽取.

2) LTP-Cloud[121]

2014 年哈尔滨工业大学联合科大讯飞公司共同推出了 LTP-Cloud.LTP-Cloud 以哈工大社会计算与信息检索研究中心研发的"语言技术平台(LTP)"为基础,为用户提供高效精准的中文自然语言处理云服务.LTP-Cloud 支持跨平台、跨语言编程等,并提供了一整套自底向上的丰富、高效、高精度的中文自然语言处理模块应用程序接口和可视化工具等.在实体关系抽取方面,研究人员利用该系统对中文文本进行分词、词性标注、命名实体识别等进行预处理,通过依存句法分析、语义角色标注和语义依存分析,抽取实体间存在的关系.

1.5 中文关系抽取的特殊性

面向中文文本的关系抽取起步较晚,而且中文与英文等语言相差较大.中文语料库的建立需要经过中文分词、词性标注和句法分析等预处理,并且在处理的过程中会存在很多错误,这就导致中文实体关系抽取的效果也略差于英文关系抽取.因此,中文领域的实体关系抽取研究具有较大的挑战性,主要存在 3 个特殊性:

1) 中文的单元词汇边界模糊,缺少英文文本中空格这样明确的分隔符,也没有明显的词形变换特征,因此容易造成许多边界歧义,从而加大了关系抽取的难度.

2) 中文触发词抽取难度较大,且数目过多.中文自然语言处理底层技术研究还不够成熟,导致错误的级联.如在长句子的句法分析上,ACE 语料中大量出现词语个数大于 30 的长句子,句法分析效果较差.此外,中文触发词数目过多,导致关系抽取召回率较低.通过对语料的分析发现,由于中文词汇表达的多义性,对同一类事件,中文触发词的个数要远大于英文.文献[123]统计表明在 ACE 语料里中文触发词个数比英文多 30%.

3) 中文存在多义性、句式复杂表达灵活、多省略等特点.不同领域中的同一个词语表示的意思并不一样,或者同一种语义可能存在多种表达形式.此外,由于互联网的快速发展,网络文本中的文字描述更加个性化,许多词语具有不同意义,中文命名实体在不同语境下被赋予了不同的意义(如高富帅、黑天鹅等),使得关系类型的识别更为困难.

1.6 关系抽取评价体系

针对特定领域的关系抽取的结果,一般通过计算对应的准确率(*Precision*)、召回率(*Recall*)和 *F*1 值来评价.其中,准确率是对于给定的测试数据集,分类器正确分类为正类的样本数与全部正类样本数之比;召回率则是对于给定的测试数据集,预测正确的正类与所有正类数据的比值;而 *F*1 值则是准确率和召回率的调和平均值,可以对系统的性能进行综合性的评价.对应的计算为

$$Precision=\frac{TP}{TP+FP}, \tag{1}$$

$$Recall=\frac{TP}{TP+FN}, \tag{2}$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall}, \tag{3}$$

其中,数据有 2 种类型:测试集数据和预测结果数据.对一批测试数据进行预测,一般可以将关系抽取的结果分成 4 种:

1) *TP*(true positive).原本是正类,预测结果为正类(正确预测为正类).

2) *FP*(false positive).原本是负类,预测结果为正类(错误预测为正类).

3) *TN*(true negative).原本是负类,预测结果为负类(正确预测为负类).

4) *FN*(false negative).原本是正类,预测结果为负类(错误预测为负类).

针对开放领域的关系抽取,目前还缺少公认的评测体系,一般通过考查抽取关系的准确性以及综合考虑算法的时间复杂度、空间复杂度等因素来评价关系抽取模型的性能.

2 实体关系抽取主要方法

本文以关系抽取的发展历程为主线,经过总结和整理将关系抽取的方法主要分为四大类,接着根据处理特点细分为若干种不同的子方法,并简要表示了各类方法之间的联系和区别.具体分类方法如图 2 所示:

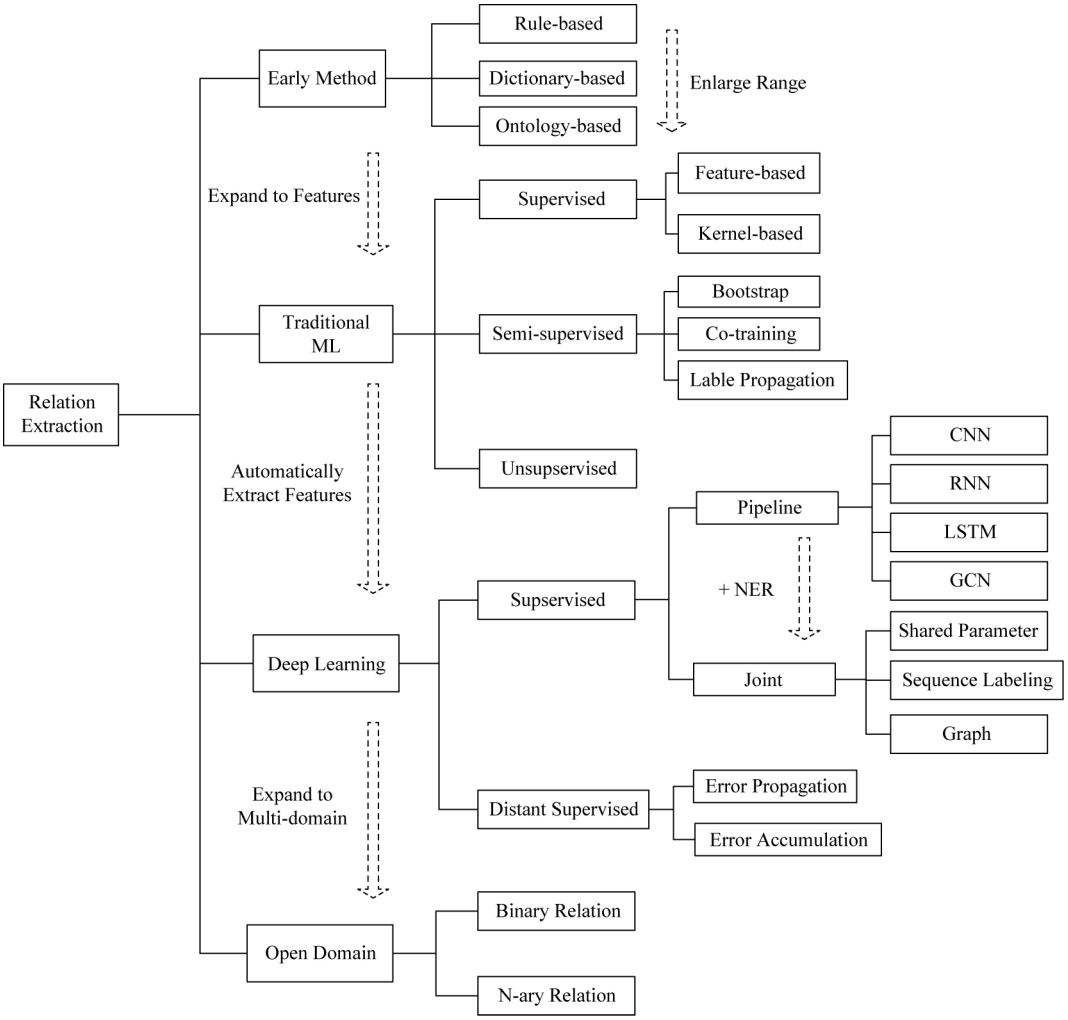


Fig. 2 The classification of relation extraction

图 2 关系抽取分类

2.1 早期的关系抽取方法

2.1.1 基于规则的关系抽取方法

早期的关系抽取方法主要是通过人工构造语法和语义规则.基于规则的方法需要运用语言学知识提前定义能够描述 2 个实体所在结构的规则,这些定义的规则主要由若干基于词语、词性或者语义的模式集合构成.在关系抽取的过程中,将已经预处理的语句片段与模式进行匹配判定,完成关系抽取的分类.

Aitken[7]借助自然语言数据并应用归纳逻辑编程(ILP)技术获得了信息提取规则,在包含 371 个句子的数据集中,$F1$ 值可达 66%.McDonald 等人[8]利用语义过滤和专家评估显示解析器在生物领域分别使用 NE,MC,PC 这 3 个系统进行生物相关的关系抽取,其中 PC 系统的平均 $F1$ 值为 69%,且相比使用最大斜率和枚举的方法分别减少 13%和 31%的错误率.Aone 等人[9]对语料文本的特点进行总结,邀请知识领域专家编写文本关系描述规则,从而在文本中抽取与规则匹配的关系实例.Humphreys 等人[10]首先对句子进行句法树分析,将分析的结果作为输入,并利用人工构造复杂的句法规则实现实体之间语义关系的识别.Fukumoto 等人[11]提出了 OKI 信息抽取系统,可以进行命名实体、模板化元素、模板化关系,其中关系抽取采用实体之间的谓词信息来判定 2 个实体之间的语义关系.中文领域基于规则的关系抽取起步较晚.邓擘等人[12]发现,相比于英文直接用模板匹配句子的方式,中文关系抽取方法的准确率和召回率低很多.因此它们在模式匹配的基础上引入了词汇、语义匹配技术对中文领域的实体关系进行抽取.实验结果表明,利用词汇、语义模式匹配的方法更适合于处理中文实体关系抽取任务,词汇语义模式匹配相比直接

匹配模式 $F1$ 值提高了近 30%.温春等人[13]提出一种扩展的关联规则方法用于抽取中文非分类关系,在利用普通关联规则抽取出非分类关系概念对后,通过语言学规则抽取相应的非分类关系名称.该方法克服了普通关联规则方法无法得出具体非分类关系名称的缺点,能够确定非分类关系的定义域和值域.

基于规则的关系抽取方法要求规则构建者(如语言学家等)对领域的背景和特点有深入的了解.在限定了领域以及语料的规模时,早期的关系抽取方法取得了一定的成就.而基于规则的关系抽取方法的缺点则是对跨领域的可移植性较差、人工标注成本较高以及召回率较低.这些基于规则的关系抽取方法所带来的困扰驱使研究者尝试跳出该方法的局限,转而使用基于词典等方法.

2.1.2 基于词典驱动的关系抽取方法

在基于词典驱动的关系抽取方法中,需要对词典进行扩充,通常只需新增指示实体关系类型的动词即可.该方法通过字符串匹配算法识别给定文本中的实体,并利用领域词典中的动词及其动词的关系结构判别关系类型,最终完成关系抽取任务.该方法以其简洁高效的特点曾经引起研究的热潮.

Aone 等人[14]基于大规模事件提出一种关系抽取方法,该方法具有开销小、准确率高的特点,在 39 种关系类型构成的测试数据集中,实验数据 $F1$ 可达 75.35%,相比于 McDonald 基于规则的方法提高 6.35%,而在多类型的关系和事件共同抽取时 $F1$ 值可达 73.95%;Temkin 等人[15]利用词典表示 2 个蛋白质之间关系的关系词,但是该关系词抽取方法的性能完全依赖于词典的质量和规模,而且需要耗费大量的人工;Neelakantan 等人[16]尝试利用包含 18 546 条标注句子的数据集来训练二元分类器,从未标注的大规模语料的候选实体中选择真实的实体,自动地构建相关实体类型的词典;文献[17]通过对文本分析发现,信息系统通常需要 2 个词典:语义词典和表示关系类型提取模式的词典,该文献提出了一种多级自举(bootstrapping)的方法,可以同时生成语义词典和提取模式.该方法将标注的文本和关系类别的种子词作为输入,采用多级自举的算法交替选择最佳的提取模式,通过不断迭代的方式扩充语义词典和关系抽取模板词典.

由于构建的词典均是以动词为关系抽取的核心依据,难以解决其他词的关系类型的抽取识别,而且灵活性较差.因此,研究者开始探索新的关系抽取方法.

2.1.3 基于本体的关系抽取方法

知识管理过程中,基于本体的方法利用信息抽取技术抽取出的实体以及实体间的关系来构建和丰富本体,借助已有的本体层次结构和其所描述的概念之间的关系来协助进行关系的抽取.

Iria[18]提出了可训练关系抽取的框架(trainable relation extraction framework, T-Rex),该框架是一个基于本体的关系抽取通用软件框架,可以自动灵活地对语义网进行语义标注,能够将语料模型化到字符级、语词级、短语级、语句级和文档级层次,实现对本体的定义和扩充.在足球领域,Schutz 等人[19]结合 DOLCE, SUMO, SEO 等本体构建了 Relext 系统,该系统能自动识别实体和实体之间的关系.Sabou 等人[20]提出自动选择和查询本体的 SCARLET 系统可以使用 2 种策略发现实体概念之间的关系:1)如果 2 个概念之间的关系已经被定义于单个本体中,则认为这 2 个概念之间有关系;2)以递归的方式跨实体发现 2 个概念之间的关系.

2.2 基于传统机器学习的抽取方法

基于传统机器学习的方法以统计语言模型为基础,研究思路明确,并采用相对简单的方法获得较好的效果.基于机器学习的实体关系抽取方法以数据是否被标注作为标准进行分类,主要集中于 3 类方法:有监督的关系抽取算法、半监督的关系抽取算法、无监督的关系抽取算法.机器学习的方法优点明显,能够明显提升结果的召回率,领域限制性弱于早期的 3 种关系抽取方法.

基于传统机器学习的关系抽取算法主要分为学习过程和预测过程 2 个主要部分,一般流程如图 3 所示.

1) 学习过程.采用训练样本,学习出关系抽取模型.

① Preprocessing,即预处理,将语料文本清洗成可以直接抽取的纯文本格式;

② Textual analysis,即文本分析,对文本的表示及其特征(POS, NER 等)进行选取;

③ Relation represention,即关系表示,即对实体之间的联系进行语义表示;

④ Relation extraction models,即关系抽取模型,基于关系表示构建分类模型.

2) 预测过程.利用学习过程获得的关系抽取模型对测试文本进行关系的预测和抽取.

一般预测过程和训练过程中的 Preprocessing, Textual analysis, Relation represention 步骤相同,不同在于 Relation decision,该步骤的具体工作为:

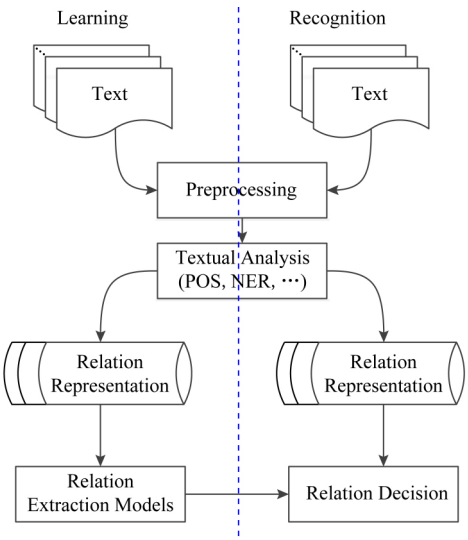


Fig. 3 General flow of relation extraction algorithm based on traditional machine learning

图 3 传统机器学习的关系抽取算法的一般流程

Relation decision,即关系判定,利用训练过程中得到的关系抽取模型对测试集数据中的实体之间的关系进行判定.

2.2.1 有监督的关系抽取方法

有监督的关系抽取方法将关系抽取任务看作分类问题.通常需要预先了解语料库中所有可能的目标关系的种类,并通过人工对数据进行标注,建立训练语料库.使用标注数据训练的分类器对新的候选实体及其关系进行预测、判断.

有监督的机器学习方法将一般的二元关系抽取视为分类问题:

$$F(S)=\begin{cases}+1, & S\text{ 为预测的目标关系},\\ -1, & \text{其他},\end{cases} \tag{4}$$

其中,$s=w_1,w_2,\cdots,e_1,\cdots,w_j,\cdots,e_2,\cdots,w_n$,即为包含实体关系的文本,$e_i$ 为实体类型,w_j 为关系触发词,F 为关系分类器.基于特征向量抽取以及基于核函数的方法是实体关系抽取方面中最流行的有监督的抽取方法.

1) 基于特征向量的抽取方法

基于特征向量抽取的方法主要从关系实例中提取一系列特征向量,主要有 3 种特征类型:词汇特征、句法特征、语义特征.研究者根据不同的特征类型,利用机器学习算法显式地将语料构造成特征向量这一形式,以此建立不同的分类模型,例如最大熵(max intropy, MI)、支持向量机(support vector machine, SVM)、朴素贝叶斯(naive Bayes, NB)、条件随机场(conditional random field, CRF)等.这一类机器学习算法相对简单,方便实体关系抽取任务的顺利完成.

基于特征向量的抽取方法的一般流程如下:

① 根据语料库的文本信息,选择合适的特征;

② 根据选取特征的重要程度,赋予特征不同的权重进行计算;

③ 选择合适的分类器训练特征向量,得到关系抽取模型.

Kambhatla[21]综合实体上下文信息、句法分析树、依存关系等多种特征,将词汇、句法和语义特征与最大熵模型相结合进行关系分类.该方法利用实体上下文丰富的语言特征有利于扩展关系表达的规模和质量,为后续关系抽取奠定了基础.Zhou 等人[22]的研究更为深入,他们借鉴 Kambhatla 的经验方法,融合了基本的文法分块(chunking)信息、半自动地收集特征(如名称列表、词汇列表等),利用支持向量机进行关系分类,利用具有 43 种子类型的测试数据集 ACE 进行评测,$F1$ 值可达 55.5%.

Sun 等人[23]融合上下文特征和 2 个实体间的长期相关性特征、实体顺序特征、实体间顺序特征以及标点符号特征,并混合朴素贝叶斯模型和投票感知模型(voted perceptron, VP)两种算法进行关系分类.

Jiang 等人[24]为了进一步提高关系抽取的准确性,系统地研究和分析了从各种信息中的抽取特征并进行了描述.该方法综合考虑了技术的复杂程度以及不同维度的特征,将特征划分成不同的子空间,结合条件随机场模型取得了较好的效果,利用包含 97 篇文档的 ACE 测试数据集(1 386 条句子约合 5 万个单词)进行评测,$F1$ 值可达 54.0%.

在中文领域,车万翔等人[25]结合实体类别、实体位置关系、前后词信息等,利用 Winnow 和 SVM 2 种机器学习算法进行训练和识别中文关系,实验表明,相对于 Winnow 算法,SVM 算法所需的运行时间较长,但当将窗口大小设置为 2 时,其平均召回率和平均 $F1$ 值分别提高约 2%和 1%.郭喜跃等人[26]以词法特征、实体原始特征为基础,融合依存句法关系、核心谓词和语义角色标柱等特征进行关系抽取,极大程度上提高了关系抽取方法的性能.高俊平等人[27]提出了一种基于关系推理模型的领域知识来演化关系抽取方法.实验结果表明,该方法相对于传统方法,考虑了深层句法特征,因此具有更高的准确性,更适合中文领域知识演化关系抽取.甘丽新等人[28]综合词法特征、实体特征、句法特征以及

语义特征等,丰富了实体间的关系特征.将 1998 年 1 月份的《人民日报》所有版面内容的 40 000 多条中文句子作为语料库,得到了 3.6 亿个二元实体对,扩大了中文实体关系库的规模.以"基本特征"和"基本特征+句法语义特征"2 种方法进行关系抽取,实验表明后者在准确率、召回率、$F1$ 值这 3 个评估指标中比前者分别提高 2.21%,7.83%,4.98%,分别可达 76.03%,79.85%,77.89%,提高效果十分明显.

2) 基于核函数的抽取方法

基于特征向量抽取的方法是显式地构造特征向量形式,而基于核函数的方法则是隐式地计算特征向量的内积.此类方法在输入句法结构树之后,直接利用核函数比较关系实例之间的结构相似性.基于核函数方法的关键在于设计出计算 2 个关系实例相似度的核函数.早期的核函数主要是序列核函数,这种方法综合关系实例特征向量的顺序和结构信息,具有较好的复合性能.基于核函数在一定程度上能提高分类的准确率,有利于指导和促进了实体关系抽取的研究和发展.

使用核函数方法来抽取实体关系一般流程如下:

① 合理选择解析结构(如语法树等)隐式地计算特征向量的内积;

② 合理选择基础核函数,之后考虑关系实例特征向量的顺序和结构信息,分析关系实例的相似性;

③ 充分利用各种特征,可以对多个核函数进行复合,以提高关系抽取任务的分类精度.

近年来,研究者将多种不同的核函数运用在英文领域的关系抽取任务中.Zelenco 等人[29]利用动态规划算法,首次在浅层解析树结构中应用核函数.该方法使用支持向量机和投票感知模型等方法进行关系抽取的分类任务;在 Zelenco 的基础之上,Culotta 等人[30]运用基于支持向量机的方法,融合依存树函数和知识库 WordNet,提出了扩展子树节点间的匹配算法,并使用数量少于 15%正类关系实例进行训练,相比 Zelenco 的方法 $F1$ 提高了 2%~3%;Zhou 等人[31]融合最短路径和卷积树核函数进行实体关系抽取,该方法考虑不同层面的语义关系特征,定义了基于树的卷积核,综合考虑了谓词上下文,最终完成了关系抽取任务;Zhang 等人[32]首次提出融合多个单一核函数的方法,利用复合核函数进行关系抽取任务.实验结果表明,复合核函数的表现比任何单一核函数实验效果更佳,其准确率、召回率、$F1$ 分别达到了 76.6%,67.0%,71.5%,但复合核函数容易产生过拟合现象,且计算复杂度较高.

在中文研究方面中,刘克彬[33]利用基于核函数的关系抽取方法自动地抽取中文实体关系.该方法在语义序列核函数的基础之上,结合 K-近邻算法(KNN),构造了关系分类器进行关系抽取;郭剑毅等人[34]改进了径向基核函数,并融合了多项式函数及卷积树核函数,利用向量离散化的矩阵训练关系抽取模型,实验表明改良的多核融合方法性能更优;虞欢欢等人[35]在卷积树核函数方法的基础上,以实体的语义信息作为树结构的结点进行扩展,使用 ACE RDC 2005 中文基准数据集(预处理后挑选了 532 个文档,总共有正类关系 7 630 个,负类关系 83 063 个)进行实验,在大类抽取中最佳 $F1$ 达到了 67.0%,能有效地对中文文本进行关系抽取.

基于核函数的方法以语料本身的结构信息为基础,比较结构化关系实例之间的相似性,完成关系抽取任务.该方法在一定程度上节省了构建高维特征的复杂工作,但在隐式计算的过程中容易产生噪声,而且运算速度较慢.关于基于特征向量和基于核函数的比较如表 2 所示:

Table 2 Comparison of Relation Extraction Methods Based on Supervised Machine Learning

表 2 有监督机器学习的关系抽取方法比较

Method	Feature Space	Representation	Key Factor	Speed
Feature-based	Context, Syntactic Tree, et al.	Explicit	Feature Vector	Faster
Kernel-based	Tree kernels, Convolution Kernels, et al.	Implied	Kernel	Slower

综上所述,有监督的机器学习关系方法在关系抽取任务中取得了较好的效果.然而有监督的机器学习方法依赖标注的语料资源库,必须进行大量的预处理工作,耗费大量人力,而且无法自动地进行关系抽取和扩展实体关系的类型.因此,越来越多的研究者开始利用较少的人工参与和标注语料资源的半监督方法进行关系抽取.

2.2.2 半监督的关系抽取方法

为了解决有监督的关系抽取方法在标注大量语料时所带来的高成本问题,学者开始研究利用少量的标注语料或数据库进行关系抽取任务,半监督的关系抽取方法应运而生.该方法利用少量标注数据和相关的学习算法,训练大量未标记的测试文本的语料库进行关系抽取.该方法不仅能有效地减少对标注语料的依赖和人工参与,而且性能较好,能自动

扩展到大规模语料的关系抽取任务中,广泛被研究者使用.半监督机器学习关系抽取的一般流程如图4所示[123]:

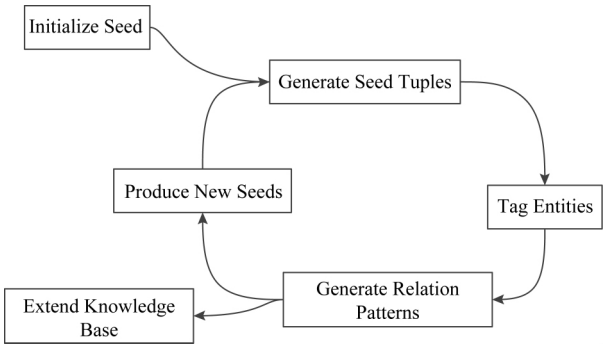


Fig. 4 The general process of semi-supervised relation extraction method

图4 半监督机器学习关系抽取方法一般流程

① Initialize seed,即初始种子,利用少量关系实例人工构造的初始种子集合.

② Generate seed tuples,即生成初始种子的关系三元组,由初始种子集合之间的实体关系产生,便于之后的实体的标识.

③ Tag entity,即标识实体,对文本进行预处理,利用知识库中的初始关系三元组识别训练文本中实体.

④ Generate relation patterns,即生成抽取模式,利用模式学习的方法,通过不断迭代,产生新的关系实例.

⑤ Produce new seeds,即产生新的种子,根据新的关系实例增加新的种子,不断扩充种子集合的规模.

⑥ Extend knowledge base,即扩展知识库,将新的关系实例扩展到知识库中.

目前,半监督的关系抽取方法主要有自举方法、协同训练(co-training)和标注传播(label propagation)等.

1) 自举方法

Brin[36]首次利用自举的方法构建了DIPRE系统进行关系抽取.他们首先确认少量的关系种子类型,通过不断迭代的方法自动地从大量训练语料库中获取抽取模板和新的关系实例;在Brin的基础上,Agichtein等人[37]设计的Snowball抽取系统完善了关系的描述方法,在最佳情况下可提高6.0%,达到了96.0%,提高了对新获取的关系实例评价方式的可信度;此外,Zhu等人[38]设计了StatSnowball抽取系统,用于识别人际关系,该系统基于Marka逻辑网络(Markov logic networks,MLNs),不断改进了Snowball系统的模板评价方式,在sent500数据集取得了86.9%的$F1$值,此外其准确率也相对Snowball系统提高了1.78%,达到了97.8%,进一步提高了关系抽取的性能;为减少产生错误模板,Carlson等人[39]约束了不同类别的抽取模板的范围,以15个种子实例和5个种子模式为基础,在3 570万唯一的文本关系模式中不断训练模型,提高了模型的性能和错误输出;语义漂移是自举方法的重要挑战之一,即将错误预测为正类的实体对加入到迭代过程中,最终影响关系抽取模型的性能,为此,Gupta等人[40]提出联合使用实体和模板种子,在迭代的过程中以平行且相互制约的方式扩展实体和模板,并引入高质量的相似性算法来判别模板;Qin等人[41]利用搜索引擎中设置的大型新闻标题句子,通过描述词与句子集之间的共现关系来评估实例的可靠性,然后通过模式历史匹配中的正负实例数来评估模式的可靠性.实验结果表明:迭代中使用的实例和模式的可靠性评估有效地提高了关系提取的准确性,其中CSEAL方法对130个实例的测试准确率率达到了97%,并提高了提取模式的质量.

在中文方面,何婷婷等人[42]提出了基于种子自扩展机制,利用自举的方法抽取1998年上半年纯文本《人民日报》语料的中文实体间的关系.实验表明:当上下文窗口大小设置为6时,对2615条候选命名实体对进行关系抽取,实验结果性能最佳,准确率、召回率及$F1$分别达到了83.1%,79.6%,81.3%.在地理领域,余丽等人[43]利用自举的方法,根据语料库中词语的特征来提取表示实体关系的关系指示词,其中准确率和召回率分别提高了5%和23%.

基于自举半监督式机器学习的方法借助高质量的初始关系种子,不依赖大规模的标注语料库,可以自动挖掘自然语言的部分词法特征.该方法有利于在缺乏大量标注语料中进行关系抽取任务.

2) 协同训练

协同训练是由Blum等人[44]提出的一种半监督机器学习算法,该方法利用2个分类器对同一个实例从不同角度进行关系分类.2个分类器相互学习、相互强化,不断提高关系抽取的性能,它被广泛应用在自然语言处理和信息检索领域中.Abney[45]提出了一种对Yarowsky算法改进的协同训练的评估方式,实验表明在完全独立的条件下,该算法能一定程度强化实体之间的较弱联系的关系表示;Zhang[46]

提出基于随机特征的 BootProject 算法.该算法基于协同训练的思想,用于对半监督的语义库进行关系分类,并在包含 5 260 条标注关系的语料集 ACE 中发现关系,其结果 $F1$ 均值可达到 70.9%.

3) 标注传播

标注传播算法是由 Zhu 等人[47]提出的,这是一种基于图的半监督机器学习方法,基本思路是用已标记节点的标签信息去预测未标记节点的标签信息.该算法将分类问题看作是标签在图上的传播,所有实体看作图中的节点,实体对之间的关系看作边.但是该方法的不确定性较高,不适合关系类别特别复杂的文本数据.Hoffmann 等人[48]采用多实例多标签(multi-instance multi-label)的方法,考虑关系抽取系统的重叠问题,其中 MULTIR 方法的准确率为 72.4%,召回率为 51.9%,$F1$ 值为 60.5%,进一步提高了关系抽取的性能.

对初始种子的选取,是半监督机器学习关系抽取算法的重点.此外,如何降低迭代过程中的噪声问题困扰着研究者.为了进一步提高关系抽取方法的性能,针对半监督的机器学习关系抽取算法依旧吸引着众多研究者的深入探索.然而在面向大规模的语料库的条件下,无法全部预知所有关系类型,这促使一些学者将研究的目光转向无监督的关系抽取方法.

2.2.3 无监督的关系抽取方法

事先确定关系类型是有监督和半监督机器学习的关系抽取方法的局限性之一,而在大规模的语料中无法预知所有的实体关系类型,研究者提出利用无监督机器学习的方法进行关系抽取.无监督的机器方法是自底向上从大规模的语料库中抽取实体之间的关系.该方法首先通过基于聚类(cluster)的思想将上下文信息相似性的实体对聚成一类,然后选取合适的词语标记关系,之后自动地抽取实体之间的语义关系.

2004 年 Hasegawa 等人[49]基于相同语义实体对具有相似的上下文语境的假设,首次提出使用无监督的机器学习的方法进行关系抽取.无监督的机器学习关系抽取一般流程如下:

① 获取命名实体识别及其上下文的信息;

② 聚类具有相似性的命名实体对;

③ 选择核心词汇标注各类的语义关系.

然而,该假设存在一些问题,如已经选取的聚类实体对之间可能包含多种关系.Rozenfeld 等人[50]提出在同一语料库中聚类实体对或者将具有多种关系的候选实体对剔除的方法完善了 Hasegawa 的假设,不仅如此,Rozenfeld 利用基于上下文特征的模式极大地提高了关系抽取的性能;Shinyama 等人[51]提出多层级聚类的方法抽取关系,该方法通过基础模板映射新的次生聚类(主要包含相同关系的实体对),在美国 12 家主流报纸中挑选了 2005-09-21—2005-11-21 两个月的文章,获得了 643 767 个基础模式和 7 990 中唯一的类型,不断扩展了实体关系库的规模;Davidov 等人[52]限定了概念词,利用 Google 搜索为知识背景,自动抽取与其相关的实体和语义关系.该方法无需提前预定义任何关系类型,在最佳的情况下准确率达到了 81.0%,召回率达到了 79%;Yan 等人[53]融合依存特征和浅层语法模板,利用聚类方法在大规模的语料库中抽取维基百科词条中的实体所有的语义关系;此外,为了进一步提高单层次聚类算法的性能,Bollegala 等人[54]分析聚类后的模板,发现了实体对之间的隐含语义关系,从候选的关系模板中筛选合适的抽取模板,扩展了实体关系的范围,在一定程度上提高了准确率和召回率;在医学专业领域,Rink 等人[55]以产生式模型为基础构建了无监督实体关系抽取框架,可以挖掘实体间的潜在关系信息,并在数据集 2010 i2b2/VA Challenge 验证了 RDM 方法的有效性,促进了无监督的机器学习方法在产业的应用.

在中文实体关系抽取方面,黄晨等人[56]基于卷积树核,结合句法树结构信息的特点,提出了一种新的无监督中文实体关系抽取方法.该方法采用最短路径表示结构化的关系实例,利用卷积树核函数实现关系的分层聚类;刘安安等人[57]首先使用实体之间的距离限制和关系指示词的位置限制获取候选关系三元组,然后采用全局排序和类型排序的方法来挖掘关系指示词,最后使用关系指示词和句式规则对关系三元组进行过滤.在获取大量关系三元组的同时,还保证了 80%以上平均准确率.

无监督的关系抽取方法无需事先人工定义实体关系的类型,可以方便地移植到别的领域,适合针对大规模地网络文本数据进行实体间的关系抽取.虽然无监督的实体关系抽取方法有效地减少了对标注语料的依赖和人工参与,但是仍然依赖于初始种子和语料库的质量,而且需要人工筛选低频的实体对.目前,无监督的关系抽取方法的研究热点之一是如何利用聚类的算法新增可信度较高的关系实例和抽取模板.

综上所述,3 种传统机器学习关系抽取方法各有所长,也各有所短.研究者充分利用各种算法的优势,进一步提升实体关系抽取的性能.3 种方法的相关比较如表 3 所示:

Table 3 Comparative Analysis of Relation Extraction Methods Based on Supervised, Semi-Supervised and Unsupervised

表 3 基于有监督、半监督、无监督关系抽取比较

EM	IM	DGA	DDC	MA	MD	MPI
Supervised	Classifying	Weakest	Strongest	① Full Use of Prior Knowledge ② For Specific Domain	① Need a Large Amount of Annotation Data ② Poor Portability	① Improve Features ② Improve Kernel
Semi-supervised	Classifying	Stronger	Stronger	① Need a Small Amount of Corpus ② For Open/Web Domain	Need a lot of Analysis and Post Processing	① Extend Model ② Reduce Noise
Unsupervised	Clustering	Strongest	Weakest	① No Need to Label Corpus ② For Large-scale Unlabeled Corpus	Need to Determine the Clustering Threshold in Advance	① Extend Feature ② Improve Clustering

Notes: EM(extraction method), IM(implementation method), DGA(domain generalization ability), DDC(degree of dependence on the corpus), MA(method advantages), MD(method disadvantages), MPI(method of performance improvement).

2.3 基于深度学习的关系抽取方法

由于传统机器学习的关系抽取方法选择的特征向量依赖于人工完成,也需要大量领域专业知识,而深度学习的关系抽取方法通过训练大量数据自动获得模型,不需要人工提取特征.2006 年 Hinton 等人[124]首次正式提出深度学习的概念.深度学习经过多年的发展,逐渐被研究者应用在实体关系抽取方面.目前,研究者大多对基于有监督和远程监督 2 种深度学习的关系抽取方法进行深入研究.此外,预训练模型 Bert(bidirectional encoder representation from transformers)[125]自 2018 年提出以来就备受关注,广泛应用于命名实体识别、关系抽取等多个领域.

2.3.1 有监督的关系抽取方法

有监督的深度学习关系抽取方法能解决经典方法中存在的人工特征选择、特征提取误差传播 2 大主要问题,将低层特征进行组合,形成更加抽象的高层特征,用来寻找数据的分布式特征表示.目前,有监督的关系抽取方法主要有流水线学习(pipeline)和联合学习(joint)两种.

1) 流水线学习

流水线学习方法是指在实体识别已经完成的基础上直接进行实体之间关系的抽取.早期的流水式学习方法主要采用卷积神经网络(convolutional neural networks, CNNs)和循环神经网络(recurrent neural networks, RNNs)两大类结构.其中,CNNs 多样性卷积核的特性有利于识别目标的结构特征,而 RNNs 能充分考虑长距离词之间的依赖性,其记忆功能有利于识别序列.随着深度学习的不断发展,研究者不断改进和完善 CNN 和 RNN 的方法,并产生了许多变体,如长短期记忆网络(long short-term memory, LSTM)、双向长短期记忆网络(bidirectional long short-term memory, Bi-LSTM)等,此外,随着图卷积神经网络(graph convolutional network, GCN)在自然语言处理领域的应用,GCN 也越来越多地用于挖掘和利用实体间的潜在信息,为解决关系重叠、实体重叠提供了新思路,从而进一步促进了关系抽取的发展.

① CNN

2014 年 Zeng 等人[58]首次使用 CNN 提取词级和句子级的特征,通过隐藏层和 softmax 层进行关系分类,提高了关系抽取模型的准确性;Liu 等人[59]在实体关系抽取方面使用简单的 CNN 模型,该模型主要由输入层、卷积层、池化层和 softmax 层组成,输入词向量和距离向量等原始数据进行实体关系抽取;为了消除了文本大小的任意性所带来的不便,Collobert 等人[60]利用设置大小固定的滑动窗口和在输入层和卷积层之上增添 max 层 2 种办法,提出了一种基于 CNN 的自然语言处理模型,方便处理多种任务;Nguyen 等人[61]设计了多种窗口尺寸的卷积核的 CNN 模型,能自动学习句子中的隐含特征,最大限度上减少了对外部工具包和资源的依赖;Santos 等人[62]使用逐对排序这一新的损失函数,有效地区分了关系类别;Xu 等人[63]融合卷积神经网络和最短依存路径的优势进行实体关系抽取,在公有数据集 SemEval-2010 Task 8 的评估结果中,$F1$ 值为 85.4%,相比于不使用最短依存路径的方法提高了 4.1%,验证了卷积神经网络和最短依存

路径结合的有效性；Ye 等人[64]基于关系类别之间的语义联系，利用 3 种级别的损失函数 AVE，ATT，Extended ATT，在包含 10 717 条标注样例的 SemEval-2010 Task 8 中进行模型评估，最佳情况下准确率、召回率、$F1$ 值分别达到了 83.7%，84.7%，84.1%，有效地提高了关系抽取方法的性能；Fan 等人[65]提出了一种最小监督关系提取的方法，该方法结合了学习表示和结构化学习的优点，并准确地预测了句子级别关系.通过在学习过程中明确推断缺失的数据，该方法可以实现一维 CNN 的大规模训练，同时缓解远程监管中固有的标签噪音问题.

在中文研究方面，孙建东等人[66]基于 COAE 2016 数据集的 988 条训练数据和 937 条测试数据，提出有效结合 SVM 和 CNN 算法可以用于中文实体关系的抽取方法.传统文本实体关系抽取算法多数是基于特征向量对单一实体对语句进行处理，缺少考虑文本语法结构及针对多对实体关系的抽取算法；基于此，高丹等人[67]提出一种基于 CNN 和改进核函数的多实体关系抽取技术，并在 25 463 份法律文书的实体关系抽取上，取得了较好的抽取效果和较高的计算效率.

② RNN

除 CNN 关系分类的方法外，Socher 等人[68]首先采用 RNN 的方法进行实体关系抽取.该方法利用循环神经网络对标注文本中的句子进行句法解析，经过不断迭代得到了句子的向量表示，有效地考虑了句子的句法结构；面对纯文本的实体关系抽取任务，Lin 等人[69]使用了一种多种语言的神经网络关系抽取框架，并在句子级别引入注意力机制(attention)，极大地减少了噪音句子的影响，有效地提高了跨语言的一致性和互补性.由于神经网络经常受到有限标记实例的限制，而且这些关系抽取模型是使用先进的架构和特征来实现最前沿的性能；Chen 等人[70]提出一种自我训练框架，并在该框架内构建具有多个语义异构嵌入的递归神经网络.该框架利用标记的、未标记的社交媒体数据集 THYME 实现关系抽取，并且具有较好的可扩展性和可移植性.

③ LSTM/Bi-LSTM

为了解决 RNN 在自然语言处理任务中出现的梯度消失和梯度爆炸带来的困扰，研究者使用性能更为强大的 LSTM.LSTM 是一种特殊的循环神经网络，最早是 Hochreiter，Schmidhuber 提出.2015 年 Xu 等人[71]提出基于 LSTM 的方法进行关系抽取，该方法以句法依存分析树的最短路径为基础，融合词向量、词性、WordNet 以及句法等特征，使用最大池化层、softmax 层等用于关系分类；Zhang 等人[72]使用了 Bi-LSTM 模型结合当前词语之前和词语之后的信息进行关系抽取，在最佳实验结果中相比于文献[58]的方法提高了 14.6%，证实了 Bi-LSTM 在关系抽取上具有有效性.

④ GCN

图神经网络最早由 Gori 等人[127]提出，应用于图结构数据的处理，经过不断发展，逐渐应用于自然语言处理领域.而图卷积神经网络能有效地表示实体间的关系，挖掘实体间的潜在特征，近年来受到了越来越多的关注.

Schlichtkrull 等人[73]提出使用关系图卷积神经网络(R-GCNs)在 2 个标准知识库上分别完成了链接预测和实体分类，其中链接预测抽取出了缺失的关系，实体分类补全了实体缺失的属性；为有效利用负类数据，Zhang 等人[74]提出一种扩展的图卷积神经网络，可以有效地平行处理任意依赖结构，便于对实体关系进行抽取.通过在数据集 TAC 和 SemVal-2010 Task 8 上的评估，其最佳的实验结果的准确率、召回率、$F1$ 值为 71.3%，65.4%，68.2%，该方法的性能优于序列标注和依赖神经网络.此外，作者还提出一种新的剪枝策略，对输入的树结构的信息，可以快速找到 2 个实体之间的最短路径；图神经网络是最有效的多跳(multi-hop)关系推理方法之一，Zhu 等人[75]提出一种基于自然语言语句生成图神经网络(GP-GNNs)参数的方法，使神经网络能够对无结构化文本输入进行关系推理；针对多元关系的抽取，Song 等人[76]提出了一种图状的 LSTM 模型，该模型使用并行状态模拟每个单词，通过消息的反复传递来丰富单词的状态值.该模型保留了原始图形结构，而且可以通过并行化的方式加速计算.不仅提高了模型的计算效率，也实现了对多元关系的抽取；为有效利用依赖树的有效信息，减少无用信息的干扰，Guo 等人[77]提出一种直接以全依赖树为输入的、基于注意力机制的图卷积网络模型.该模型是一种软剪枝(soft-pruning)的方法，能够有选择地自动学习对关系提取任务有用的相关子结构，支持跨句多元关系提取和大规模句级关系提取.

⑤ 混合抽取

为了进一步提高关系抽取模型的性能，一些研究者开始采取融合多种方法的方式进行关系抽取.2016 年 Miwa 等人[78]使用联合的方法，他们融合 Bi-LSTM 和 TreeLSTM 模型的优点对实体和句子

同时构建模型,分别在3个公有数据集ACE04,ACE05,SemVal-2010 Task8对关系抽取模型进行评估,有效地提高了实体关系抽取的性能;Zhou等人[79]提出一种基于注意力的Bi-LSTM,着重考虑词对关系分类的影响程度,该方法在只有单词向量的情况下,优于大多数当时的方法;Li等人[80]融合Bi-LSTM和CNN的特点,利用softmax函数来模拟目标实体之间的最短依赖路径(SDP),并用于临床关系提取的句子序列,在数据集2010 i2b2/VA的实验结果$F1$为74.34%,相比于不使用语义特征的方法提高2.5%;陈宇等人[81]提出一种基于DBN(deep belief nets)的关系抽取方法,通过将DNB与SVM和传统神经网络2种方法在ACE04数据集(包含221篇消息文本、10 228个实体和5 240个关系实例)进行了比较,$F1$值分别提高了1.26%和2.17%,达到了73.28%;召回率分别提高了3.59%和2.92%,达到了70.86%,验证了DBN方法的有效性.此外,DBN方法表明,字特征比词特征更适用于中文关系抽取任务,非常适用于基于高维空间特征的信息抽取任务.

流水线方法的实验结果相对良好,但容易产生错误传播,影响关系分类的有效性;将命名实体识别和关系抽取分开处理,容易忽视这2个子任务之间的联系,丢失的信息会影响抽取效果;另外,冗余信息也会对模型的性能产生较大的影响.为解决这些问题,研究人员尝试将命名实体识别和关系抽取融合成一个任务,进行联合学习.

2) 联合学习

联合学习方法有3种,包括基于参数共享的实体关系抽取方法、基于序列标注的实体关系抽取方法和基于图的实体关系抽取方法.

① 基于共享参数的方法

命名实体识别和关系抽取通过共享编码层在训练过程中产生的共享参数相互依赖,最终训练得到最佳的全局参数.因此,基于共享参数方法有效地改善了流水线方法中存在的错误累积传播问题和忽视2个子任务间关系依赖的问题,提高模型的鲁棒性.

2016年Miwa等人[82]首次利用循环神经网络、词序列以及依存树将命名实体识别和关系抽取作为一个任务进行实验,通过共享编码层的LSTM的获得最优的全局参数,在数据集ACE04,ACE05分别减少了5.7%和12.1%的错误率,在数据集SemEval-2010 Task 8的$F1$达到了84.4%.然而Miwa忽略了实体标签之间的长距离依赖关系,为此Zheng等人[83]将输入句子通过公用的Embedding层和Bi-LSTM层,分别使用一个LSTM进行命名实体识别和一个CNN进行关系抽取,该方法的$F1$达到了85.3%,相对Miwa提高了近1%.

② 基于序列标注的方法

由于基于共性参数的方法容易产生信息冗余,因此Zheng等人[84]将命名实体识别和实体关系抽取融合成一个序列标注问题,可以同时识别出实体和关系.该方法利用一个端到端的神经网络模型抽取出实体之间的关系三元组,减少了无效实体对模型的影响,提高了关系抽取的召回率和准确率,分别为72.4%和43.7%.为了充分利用实体间有多种关系,Bekoulis等人[85]将命名实体识别和关系抽取看作一个多头选择问题,可以表示实体间的多个关系;此外Bekoulis等人[86]还发现对模型加入轻微的扰动(对抗样本)可以使得WordEmbedding的质量更好,不仅提高了置信度还避免了模型过拟合,模型的性能大大提升.因此首次将对抗学习(adversarial training, AT)加入联合学习的过程中.实验结果表明,在4个公有数据集ACE04,CoNLL04,DREC,ADE的$F1$提高了0.4%~0.9%.

③ 基于图结构的方法

针对前2种方法无法解决的实体重叠、关系重叠问题,基于图结构的方法能有效得解决.Wang等人[87]发现生成标记序列后的合并三元组标签过程采用的就近组合无法解决关系重叠问题,因此提出一种新的基于图架构的联合学习模型.该方法不仅能有效解决关系重叠问题,而且使用偏执权重的损失函数强化了相关实体间的关联,实验结果的准确率、召回率及$F1$值分别为64.3%,42.1%,50.9%.此外,Fu等人[88]提出将图卷积神经网络用于联合学习,利用图的节点表示实体,边表示关系,有效地解决了关系重叠和实体重叠问题,不仅如此,还对边(关系)加入了权重,有效挖掘了实体对间的潜在特征,通过使用NYT和WebNLG数据集的评估,该方法在最佳情况下准确率、召回率及$F1$值可达63.9%,60.0%,61.9%,与文献[87]相比,召回率和$F1$分别提高17.9%和11.0%.

本文选取了几种经典的有监督关系抽取方法进行了综合比较,具体如表4所示.

深度学习的有监督方法能够自动地学习大量特征,避免人工选择特征,但对大量没有进行标记的数据,这种方法就显出其弊端.为了减少对大数据的标注的人工成本,研究者尝试使用远程监督的方法进行关系抽取.

Table 4 Comparison of Relation Extraction Methods Based on Supervised Learing

表 4 有监督学习关系抽取方法对比

Category	Method	Year	Datasets	$F1$/%	Keys
Pipeline	Ref[58]	2013	ACE04 ACE05	83.8	CNN, Codes synonym
	Ref[61]	2015	SemEval-2010 Task 8	82.2	CNN, Automatically learns features Independences on NLP tools
	Ref[68]	2012	SemEval-2010 Task 8	82.4	RNN, Long-term dependence
	Ref[71]	2015	SemEval-2010 Task 8	83.7	LSTM, Shortest dependency path
	Ref[72]	2015	SemEval-2010 Task 8	84.3	Bi-LSTM, Completes sequential information about all words
	Ref[77]	2019	TACRED PubMed-based	85.7	Attention Guided-GCN, N-ary relation extraction Automatically learns to select relevant sub-structures
Joint	Ref[82]	2016	ACE04 ACE05 SemEval-2010 Task 8	84.4	Shared parameters, Word sequence Dependency tree substructure
	Ref[84]	2017	NYT	85.3	Sequence labeling, Biased loss function
	Ref[87]	2018	NYT	50.9	Graph, Transition-based, Overlapping relations
	Ref[88]	2019	NYT、WebNLG	61.9	Graph, Overlapping relations, Interaction between relations

2.3.2 远程监督的关系抽取方法

针对海量无标记数据的处理,远程监督的实体关系抽取方法极大地减少了对人工的依赖,可以自动地抽取大量的实体对,从而扩大了知识库的规模.此外,远程监督的方法具有较强的可移植性,比较容易应用到其他领域.远程监督的基本假设是如果2个实体在已知知识库中存在着某种关系,那么涉及这2个实体的所有句子都会以某种方式表达这种关系.Mintz等人[89]首次在ACL会议上将远程监督方法应用于实体关系抽取的任务中.他们将新闻文本与知识图谱FreeBase进行中的实体进行对齐,并利用远程监督标注的数据提取文本特征,训练关系分类模型.

这类方法在数据标注过程会带来2个问题:噪音数据和抽取特征的误差传播.基于远程监督的基本假设,海量数据的实体对的关系会被错误标记,从而产生了噪音数据;由于利用自然语言处理工具抽取的特征也存在一定的误差,会引起特征的传播误差和错误积累.本文主要针对减少错误标签和错误传播问题对远程监督的关系抽取方法进行阐述.

1) 针对错误标签

由于在不同语境下同一对实体关系可能存在不同含义,为了减少因此而产生的错误关系标签,Alfonseca等人[90]利用FreeBase知识库对关系进行分层处理,以启发式的方式自动识别抽取表示关系的语义和词汇;由于利用启发式的规则标记实体关系时会产生一些错误标记,Takamatsu等人[91]提出一种产生式模型,用于模拟远程监督的启发式标记过程,使用903 000篇Wikipedia文章进行模型的训练,并使用400 000篇文章进行测试,实验结果的准确率、召回率和$F1$值分别为89.0%,83.2%,82.4%;为了解决Alfonseca提出的方法缺乏实体的知识背景问题,Ji等人[92]提出了一种在句子级别引入注意力机制的方法来抽取有效的实例,并通过FreeBase和Wikipedia不断地扩充实体的知识背景;之前大多方法对负类数据的利用率较低,Yu等人[93]提出结合从句子级远程监督和半监督集成学习的关系抽取方法,该方法减少了噪声数据,充分利用了负类数据.该方法首先使用远程监督对齐知识库和语料库,并生成关系实例集合,接着使用去噪算法消除关系实例集中的噪声并构建数据集.为了充分利用负类数据,该方法将所有正类数据和部分负类数据组成标注数据集,其余的负类数据组成未标注数据集.通过改进的半监督集成学习算法训练关系分类器的各项性能,然后进行关系实例的抽取.

此外,为了减少错误标签产生的噪音数据对关系抽取模型的影响,Wang等人[94]提出了一种无标签的远程监督方法;该方法只是使用了知识库中的关系类型,而由2个实体来具体确定关系类型,避免了知识库中的先验知识标签对当前关系类型判别造成影响,也无需使用外部降噪工具包,大大提高了关系抽取的效率和性能;为了进一步提高对数据的使用效率,Ru等人[95]使用Jaccard算法计算知识库中的关系短语与句子中2个实体之间的语义相似性,借此过滤错误的标签.该方法在减少错误标签的过程中,利用具有单词嵌入语义的Jaccard算法选择核心的依赖短语来表示句子中的候选关系,可以提取关系分类的特征,避免以前神经网络模型关系提

取的不相关术语序列引起的负面影响.在关系分类过程中,将 CNN 输入的核心依赖短语用于关系分类.实验结果表明,与使用原始远程监督数据的方法相比,使用过滤远程监督数据的方法在关系提取方面结果更佳,可以避免来自不相关术语的负面影响;为了突破距离对关系抽取模型性能的限制,Huang 等人[96]提出一种融合门控循环单元(gated recurrent unit, GRU)和注意力机制的远程监督关系抽取方法,该方法解决了传统深度模型的实体在长距离依赖性差和远程监督中容易产生错误标签的问题;实验结果表明,文献[89]的方法召回率在大于 0.2 时就开始迅速下降,而该方法在整个过程中都相对稳定,保证了模型的鲁棒性;此外,通过与文献[69]的方法进行比较,该方法的召回率平均提高 10%,能够充分利用整个句子的序列信息,更适合自然语言任务的处理.

2) 针对误差传播

Fan 等人[97]提出远程监督关系提取的本质是一个具有稀疏和噪声特征的不完整多标签的分类问题.针对该问题,Fan 使用特征标签矩阵的稀疏性来恢复潜在的低秩矩阵进行实体关系抽取;为了解决自然语言处理工具包提取问题带来的错误传播和错误积累问题,Zeng 等人[98]融合 CNN 和远程监督的方法,提出分段卷积神经网络(piecewise convolutional neural network, PCNN)用于实体关系抽取,并尝试将基于 CNN 的关系抽取模型扩展到远程监督数据上.该方法可以有效地减少了错误标签的传播和积累,在最佳情况下,准确率、召回率以及 $F1$ 值达到了 48.30%,29.52%,36.64%.

针对目前在中文领域实体-属性提取中模型的低性能,He 等人[99]提出了一种基于 Bi-LSTM 的远程监督关系抽取方法.首先,该方法使用 Infobox 的关系三元组获取百度百科的信息框,从互联网获取训练语料库,然后基于 Bi-LSTM 网络训练分类器.与经典方法相比,该方法在数据标注和特征提取方面是全自动的.该方法适用于高维空间的信息提取,与 SVM 算法相比,准确率提高了 12.1%,召回率提高了 1.21%,$F1$ 值提高了 5.9%,准确率和 $F1$ 值得到显著提高.

有监督的关系抽取方法借助人工标注的方法提高了关系抽取的准确性,但是需要耗费大量人力,其领域泛化能力和迁移性较差.远程监督的方法相对于有监督的方法极大地减少了人工成本,而且领域的迁移性较高.但是,远程监督的方法通过自动标注获得的数据集准确率较低,会影响整个关系抽取模型的性能.因此,目前的远程关系抽取模型的性能仍然和有监督的关系抽取模型有一定的差距,有较大的提升空间[127].基于深度学习的监督和远程监督方法抽取对比如表 5 所示:

Table 5 Comparison of Supervised and Distant Supervised Relation Extraction Based on Deep Learning

表 5 基于深度学习的有监督和远程监督实体关系抽取对比

Contrast Content	Supervised	Distant Supervised
Dataset Labeling Method	Manual Labeling	Remote Alignment Knowledge Base
Dataset Characteristics	Higher Accuracy, Smaller Noise Data	Lower Accuracy, Bigger Noise Data
Dataset Scale	Smaller	Bigger
Method Cost	Higher	Lower
Method Mobility	Worse	Better
Method Domain	Stronger	Weaker
Extraction Effect	Better	Worse

2.3.3 BERT

2018 年 Google AI Language 发布了 BERT 模型,该模型在 11 个 NLP 任务上的表现刷新了记录,在自然语言处理学界以及工业界都引起了不小的热议.BERT 的出现,彻底改变了预训练产生词向量和下游具体 NLP 任务的关系.

在关系抽取领域,应用 BERT 作预训练的关系抽取模型越来越多,如 Shi 等人[128]提出了一种基于 BERT 的简单模型,可用于关系抽取和语义角色标签.在 CoNLL05 数据集中,准确率、召回率和 $F1$ 值分别为 88.6%,89.0%,88.8%,相比于 baseline 方法分别提高了 1.0%,0.6%,0.7%;Shen 等人[129]借助 BERT 的强大性能对人际关系进行关系抽取,减少了噪音数据对关系模型的影响.此外,又使用了远程监督可以对大规模数据进行处理,在 CCKS 2019 eval Task3 IPRE 数据集的结果表明,该方法优于大多数人际关系抽取方法,$F1$ 值达到了 57.4%.

BERT 作为一个预训练语言表示模型,通过上下文全向的方式理解整个语句的语义,并将训练学到的知识(表示)用于关系抽取等领域.但 BERT 存在许多不足之处.

1) 不适合用于长文本.BERT 以基于注意力机制的转换器作为基础,不便于处理长文本,而关系抽取领域的文本中经常出现超过 30 个单词的长句,BERT 会对关系抽取的性能产生影响.针对长句子

的情况,可以另外设计一个深度的注意力机制,以便层级化的捕捉关系.

2) 易受到噪音数据的影响.BERT 适用于短文本,而短文本中若出现不规则表示、错别字等噪音数据,这不仅会对关系触发词的抽取造成一定的影响,而且在联合学习时进行命名实体识别阶段也会产生错误的积累和传播,最终导致模型的性能下降.

3) 无法较好地处理一词多义问题.虽然通过上下文能在一定程度上缓解一词多义的影响,但一词多义对 BERT 的原始输入中的词编码影响极大.从而进行关系抽取时容易产生错误标签,无法有效地使用关系标签等进行关系分类,降低模型的准确率、召回率、$F1$ 值.这需要加以一定的机制来解决一词多义的表示问题.

2.4 基于开放领域的关系抽取方法

由于传统关系抽取基于特定领域、特定关系进行抽取,导致关系抽取这一任务耗时耗力,成本极高,同时不利于扩展语料类型.近年来,针对开放领域的实体关系抽取方法逐渐受到人们的广泛关注.由于互联网不断发展,开放语料的规模不断扩大,并且包含的关系类型愈加复杂,研究者直接面向大多未经人工标注的开放语料进行关系抽取,有利于促进实体关系抽取的发展,而且具有更大的实际意义.

开放领域关系抽取的方法是信息抽取领域的新的研究方向.该关系抽取方法主要分为半监督和无监督 2 种,并结合语形特征和语义特征自动地在大规模非限定类型的语料库中进行关系抽取.开放领域关系抽取的方法无需事先人为制定关系类型,减轻了人工标注的负担,而由此设计的系统可移植性较强,极大地促进关系抽取的发展.

开放领域的关系抽取方法主要有 3 个流程:

1) 深层解析小规模的语料集,自动抽取实体间关系三元组,利用朴素贝叶斯分类器训练已标注可信和不可信的关系三元组构建关系表示模型;

2) 利用关系抽取模型并输入词性、序列等特征等数据,在训练好的分类器上进行大量网络文献的关系抽取,获取候选关系三元组;

3) 合并候选三元组,通过统计的方法计算各个关系三元组的可信度,并建立索引.

2.4.1 英文开放领域文本关系抽取方法

针对非限定领域的关系抽取,Sekine 曾尝试按需抽取的思路,利用浅层匹配的方法,自动构造简单模板进行关系抽取,并为之后的面向开放领域的关系抽取提供了新思路.早期研究人员主要针对二元关系进行抽取,包括先识别实体词和先识别关系词 2 种主要方法.随着人们对文本信息蕴含的深层次关系的研究,多元关系抽取也逐渐进入研究者的视野.

1) 二元关系抽取方法

① 先识别实体词的二元抽取方法

在早期的开放式信息抽取领域主要是针对实体词进行关系抽取.该阶段利用无语义的特征,自动地学习实体之间的关系,并构建好表示文本关系的模型.主要的信息抽取系统包括 TextRunner, WOE, PATTY 等.

2007 年 Washington 大学的人工智能研究组的 Banko 等人[5]正式提出了面向开放领域的信息抽取方法框架,并发布了开放领域的第 1 个信息抽取信息系统 TextRunner.该系统依赖少量的人工标记数据,通过自监督的学习方式训练了朴素贝叶斯模型,并进行实体关系分类.该系统在大规模开放的网页进行实体关系分类测试,取得了当时较为优秀的效果.随后,该系统融合线性条件随机场和马尔可夫逻辑模型,其性能不断得到了提高,这对关系抽取领域的发展起到了促进作用.

在 TextRunner 的基础之上,Wu 等人[100]设计开发了一种新颖的自监督学习的信息抽取系统 WOE.WOE 系统利用启发式规则训练维基百科网页信息框(Infobox)中的数据,自动地构建实体关系集.WOE 有 2 种运行模式:1)以词性标记为限制条件时,该系统的运行速度可比肩 TextRunner;2)以解析依赖关系为限制条件时,虽然抽取的速度将会减慢,但在极大地程度上提高了实体关系抽取的准确率和召回率.同时 WOE 系统中充分考虑了依存关系特征,实验结果表明:相对于 TextRunner 该方法的 $F1$ 平均值提高了 18%~34%,进一步大幅度提高了该系统的性能.

此外,Nakashole 等人[101]基于频繁项集挖掘算法提出了 PATTY 系统.该系统以模式为依据进行语义分类,构建了一个包容性的分类体系,在以 350 569 个模式构成的 Wikipedia 数据集中,对实体间的关系进行抽取,便于在大规模的语料库中表示实体间的二元关系.

2010 年 Yao 等人[102]充分结合了远程监督以及 Open IE 的优势提出了一种通用模型框架.该模型是一个涉及所有模式的并集,并且避免了对现有数据集的依赖.该模型利用矩阵分解的方法自学习到实体元组和关系的潜在特征,能有效地处理结构化

和无结构化数据.相对传统的分类方法,该模型的计算速度更快,学习效率更优,准确率更高,可扩展性更强.

② 先识别关系词的二元抽取方法

由于早期的关系抽取系统存在抽取的关系词不连贯以及关系词无法提供有效信息的问题.因此,之后面向开放领域的关系抽取开始转向先识别关系词,并深入地解析句子的语言成分进行关系抽取.该阶段比较引人注意的有 ReVerb,OLLIE,ClausIE 等.其中,ReVerb 主要以动词为核心,OLLIE 主要以名词和副词为核心,ClauseIE 主要以从句为核心.

2011 年 Fader 等人[103]深入分析了语法、词汇、语义等特征,设计了 ReVerb 系统.该系统有效减少了 TextRunner 系统和 WOE 系统所产生的错误关系三元组和无信息关系三元组.该系统使用浅层句法抽取较短的语句,而对于较长的语句则采用先识别关系词再识别实体的方法.实验结果表明,ReVerb 系统只需进行词性标注,并结合匹配的方法就能完成关系抽取的任务,有效地提高了关系抽取的准确率,在极大地程度上提高了关系抽取的性能,有力地促进了关系抽取方面的发展.

此外,Xavier 等人[104]提出了一种较为简单的方法挖掘名词与名词之间的关系以及形容词与形容词之间的关系.该方法首先识别名词或形容词及其属性,之后对识别的名词或形容词进行解析,接着自动地产生描述二元关系的三元组.该方法进一步增加了信息量,也提高了关系抽取的准确性.即对名词的属性进行抽取,使得信息量增多,抽取的准确性更高.Del 等人[105]通过对句子的结构进行分析,提出了 ClausIE 系统.该系统融合了句法模式学习、自学习算法、句子分解技术等的优势,将复杂语句分解成多个简单的语句,通过计算关系短语的相似度来对关系短语进行整合.

另外,Faruqui 等人[106]提出一种跨语言注释映射的方法,无需依赖语言包和解析目标语言,借助机器翻译就可以对多种语言进行关系抽取.在人工标注的 3 种语言(法语、印地语、俄语)进行关系抽取的实验结果表明,该开放领域抽取方法能够对维基百科 61 种语言进行关系抽取,具有较强的可移植性和扩展性;为了简化当前众多方法结构的复杂性,Song 等人[107]将实体间的语义信息转化成二进制结构,以便利用更少的时间提取更多的语义信息,高效地抽取关系三元组,并通过 SENT500 数据集测试,获得了 83.8%的 $F1$ 值.

2) 多元关系抽取方法

上述的关系抽取系统主要是针对二元关系的,Akbik 等人[108]提出针对多元关系进行抽取,设计开发了 KRAKEN 系统.该系统改进了 OIE 系统,可以对不同的关系类别进行多元关系抽取,挖掘了潜在的隐含关系,与传统方法针对特定领域进行关系抽取相比较,面向开放领域的关系抽取方法所获得准确率和召回率仍然比较低;Gamallo 等人[109]针对英语、西班牙语、葡萄牙语、加利西亚语等语种,利用一些制定的规则,采用依存分析的技术完成了关系抽取任务,取得了较好的效果,相对于 ReVerb 需要 27%的计算机 RAM,该系统只需 0.1%.Fossati 等人[110]利用语言的语义理论框架,实现了同时利用 T-Box 和 A-Box 填充知识库,完成了语义标注,最终对实体间的多元关系进行抽取.

2.4.2 中文开放领域文本关系抽取方法

1) 二元关系抽取方法

由于中文与英文存在较大的差距,因此针对英语的关系抽取系统无法直接对中文进行抽取.为了解决中文中缺省某些语言成分和倒序的问题,研究者发布了 CORE,ZORE,UnCORE 这 3 个面向开放领域的信息抽取系统.

考虑到中英文之间的差异,在面向中文开放领域的文本时,Petroni 等人[111]提出了 CORE 模型.该模型利用上下文信息进行矩阵分解可以获得关系三元组.该方法首先完成对分句、词性标注和特殊词的处理,之后对给定的语句利用 CKIP 解析器进行语法解析,最后通过识别中心关系词逐渐扩展去识别中心实体词.该方法有力地促进了面向开放的中文领域的关系抽取的研究和发展.

此外,ZORE 也是面向中文开放领域文本的关系抽取模型.ZORE 是由 Qiu 等人[112]在 2014 年提出,通过利用依存解析树识别候选实体关系三元组,采用双向传播算法迭代抽取实体关系三元组和语义模板.实验表明,该模型在对 5 MB 大小的 Wikipedia 中文构成的数据集进行关系抽取时,准确率取得了较好的成绩,达到了 76.8%.

通过对大规模开放的网络文本进行分析之后,哈尔滨工业大学的秦兵等人[113]发现实体之间的关系与实体之间的距离以及关系词的位置有较大关系.2015 年秦兵等人发布了以无监督的方式进行关系抽取的 UnCORE 系统.该系统首先对在网页上获得的大规模文本进行预处理,得到分词和标注好的词性等,接着通过约束实体之前的距离和关系词的

位置得到候选三元组,然后使用基于规则的排序(全局排序和类型排序)的算法获取关系指示词,最后采用构造好的规则和关系指示词对候选关系三元组进行过滤得到准确率较高的关系三元组.实验结果显示,该方法的平均准确率达到 80%,能有效地提取大量关系三元组,不断地扩充实体关系库.

除此之外,郭喜跃等人[114]采用半监督的方式在百科类的开放领域文本进行关系抽取,从不同方面对百度百科的信息框采用不同方法进行标注、筛选、整合,最终获得了质量较高的实体间的二元关系.该方法有效地减少了人工参与,提高了关系抽取的效率.文献[115]研究了基于无监督的中文开放领域的关系抽取,可以在没有任何人工标记数据集的情况下自动发现任意关系,建立了大规模语料库.通过将实体关系映射到依赖树,考虑到独特的中文语言特征,该文献提出了一种基于依赖语义规范形式的新型无监督中文开放领域的关系抽取模型.该模型对实体和关系之间的相对位置没有任何限制,通过抽取由动词或名词为媒介的关系,处理并行子句来提高关系抽取的性能.该方法在 4 个异构数据集上获得了稳定的性能,并获得了更好的准确率和召回率,分别为 83.76%,58.68%.

2) 多元关系抽取方法

以上所述的方法主要是针对二元关系进行中文文本抽取,李颖等人[116]基于依存分析的方法,提出了面向中文开放领域文本的多元实体关系抽取模型 N-COIE.该模型首先对中文文本进行词性标注和依存关系标注,然后在一定的约束条件下识别基本的名词短语,抽取候选实体关系多元组,最后通过过滤的方法扩充关系库.实验结果表明,该方法在面向大规模的中文领域开放的文本能够取得 81%的准确率.姚贤明等人[117]提出了中文领域多元实体关系抽取的方法.该方法以依存句法分析结果的根节点作为入口,迭代地获取所有与谓语相关联的主语、宾语及其定语成分,再利用依存句法分析结果来完善定语成分,最终获取句子中的多个实体之间的语义关系.

目前,面向大规模开放领域的关系抽取方法仍与特定领域的方法存在一定的差距,留给研究者一定的研究空间.面向开放领域的关系抽取仍然存在着一些难点,亟待解决:

1) 如何继续提高实体二元关系的准确率和召回率,进一步实现对实体间多元关系的抽取;

2) 如何继续深度挖掘实体间的隐含关系,进一步提高实体间关系的信息的有效利用;

3) 如何提出公认的评价体系,制定统一的评测标准.

3 关系抽取总结和未来发展趋势

3.1 关系抽取总结

关系抽取研究已历经 20 多年的发展,关系抽出的方法不断得到改进,关系抽取的模型性能不断得以提升,逐渐应用于知识图谱、文本摘要、机器翻译等领域.

早期的方法主要通过寻找文本的规律,制定一系列规则抽取关系,如基于规则、词典、本体的方法.该类抽取方法的准确率等评价指标较高,然而需要人工构造,其成本高昂,且处理的文本规模较小,为了突破早期方法的局限,研究人员将目光转向以特征等为基础的传统机器学习方法,如有监督学习的基于特征和核函数的方法,半监督学习的自举、协同训练、标注传播的方法以及无监督方法以聚类为核心的方法.但是传统机器学习的模型性能十分依赖人工标注特征数据的规模和数量,因此需要一个能自动地抽取特征的方法.深度学习具有自学习的特点,能够自动抽取特征,减少对人工的依赖,而且能抽取大规模文本数据.深度学习的方法主要有有监督和远程监督 2 种方法,其中有监督主要有流水线学习(如 CNN,RNN,LSTM,GCN 及其变体)和联合学习(如基于共享参数、序列标注、图)2 种,基于深度学习的方法极大地促进了关系抽取领域的发展.针对特定领域方法的模型性能良好,但其可扩展性和移植较差,因此针对开放领域的方法越来越吸引研究者的目光,但该类方法的模型性能还有待提高,此外还缺少公认的评价体系,需要进一步完善.

3.2 未来发展趋势

目前,实体关系抽取技术日渐成熟,但依然需要研究人员投入大量精力进行不断探索,通过对现有实体关系抽取研究工作进行总结,在以后的研究中可以从 5 个方面展开相关的研究.

1) 从二元关系抽取到多元关系抽取的转化.当前的关系抽取系统主要集中在 2 个实体之间的二元关系抽取,但并非所有的关系都是二元的,如有些关系实例需要考虑时间和地点等信息,所以会考虑更多的论元.李颖等人[116]提出的关系抽取模型 N-COIE 针对多元关系抽取,但该方法与二元关系抽取模型的抽取相比,在准确率和召回率上仍有较大的差距.如何根据上下文信息,识别跨越句子的多元

实体关系,提高关系抽取的准确率和智能化,这促使研究者不断投入更多的精力.

2) 开放领域的实体关系抽取的深入研究.目前的研究工作大多面向特定的关系类型或者特定领域,而使用特定的语料库,很难做到其他领域的自动迁移.虽然,一些研究者针对开放领域的关系抽取进行了研究,提出了一系列的方法用于实体关系抽取,然而这类方法和特定领域相比仍有一定的差距.如何不断提高系统的准确率、可移植性以及可扩展性,这都激励着研究人员投入更多的精力和时间,促进开放领域的实体关系抽取的发展.

3) 远程监督关系抽取方法得到不断改进.目前,由于远程监督的方法仍然存在错误标签和误差传播2个主要问题,研究者多是基于这些问题对深度学习的关系抽取模型加以改进.为了避免产生过多的错误标签,人们主要采用多示例、注意力机制的方法等方法减少噪音数据.而Qin等人[130]融合增强学习和远程监督方法的优点,不断地减少错误标签,进而降低负类数据对关系抽取模型的影响.针对误差传播的问题,研究者多是对句子的语义信息进行深入挖掘,而对句子语法信息却少有涉及.如何有效地解决远程监督产生的错误标签和误差传播,如何有效地融合语法和语义信息,这些吸引着研究者不断改进相关算法,不断提高深度学习方法的性能.

4) 深度学习有监督方法的性能提升.近年来,越来越多的研究人员关注于联合学习和基于图结构的抽取方法.联合学习将命名实体识别和关系抽取作为一个任务,减少了错误信息的积累和传播,也减少了冗余信息对模型的影响.而针对关系重叠和实体间潜在特征等问题,基于图结构的抽取方法提供了一些新的思路.然而这2种方法的性能还需进一步改进,不断促进信息抽取领域的发展.

5)工业级实体关系抽取系统的继续研发.关系抽取现已被广泛应用于智能搜索、智能问答、个性化推荐、内容分发、权限管理,人力资源管理等领域.通过对学术研究和市场需求进行深入地融合,不断提高实体关系抽取的可靠性、置信度、执行效率等,促进关系抽取模型的性能进一步得到提升,为人们的生活提供更多便利.

4 结束语

综上所述,关系抽取是自然语言处理领域的重要研究方向之一,其研究内容已从限定领域、限定类型的关系分类转变为面向互联网开放领域的实体关系自动发现.随着关系抽取技术进一步实现自动化,将对海量信息处理、智能问答、知识库自动构建等领域产生积极推动,具有广阔的应用前景.

参 考 文 献

[1] Chinchor N, Marsh E. MUC-7 information extraction task definition [C] //Proc of the 7th Message Understanding Conf (MUC-7). Berlin: Springer, 1998: 359-367

[2] Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction (ACE) program-tasks, data, and evaluation [C] //Proc of COLING 2014. Stroudsburg: ACL, 2014: 23-29

[3] McNamee P, Dang H T, Simpson H, et al. An evaluation of technologies for knowledge base population [C] //Proc of the Int Conf on Language Resources and Evaluation. Stroudsburg: ACL, 2010: 369-372

[4] Hendrickx I, Kim S N, Kozareva Z, et al. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals [C] //Proc of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Stroudsburg: ACL, 2009: 94-99

[5] Banko M, Cafarella M J, Soderland S, et al. Open information extraction from the Web [C] //Proc of Int Joint Conf on Artificial Intelligence. Amsterdam: Elsevier, 2007: 2670-2676

[6] Zhou Deyu, Zhong Dayou, Heulan Y. Biomedical relation extraction: From binary to complex [J]. Computational and Mathematical Methods in Medicine, 2014, 2014: 298473

[7] Aitken J S. Learning information extraction rules: An inductive logic programming approach [C] //Proc of External Credit Assessment Institution. Ohmsha: IOS, 2002: 355-359

[8] McDonald R, Pereira F, Kulick S, et al. Simple algorithms for complex relation extraction with applications to biomedical IE [C] //Proc of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL, 2005: 491-498

[9] Aone C, Halverson L, Hampton T, et al. SRA: Description of the IE2 system used for MUC-7 [C] //Proc of the 7th Message Understanding Conf (MUC-7). Stroudsburg: ACL, 1998 [2020-05-29]. https://www.researchgate.net/publication/2243565_Sra_Description_Of_The_Ie2_System_Used_for_MUC-7

[10] Humphreys K, Gaizauskas R, Azzam S, et al. University of Sheffield: Description of the laSIE-II system as used for MUC-7 [C] //Proc of the 7th Message Understanding Conf (MUC-7). Stroudsburg: ACL, 1998 [2020-05-29]. https://www.researchgate.net/publication/2550114_University_of_Sheffield_Description_of_the_LaSIE-II_system_as_used_for_MUC-7

[11] Fukumoto J, Masui F, Shimohata M, et al. OKI electric industry: Description of the OKI system as used for MUC-7 [C] //Proc of the 7th Message Understanding Conf. Stroudsburg: ACL, 1998 [2020-05-29]. https://core.ac.uk/display/21411891

[12] Deng Bo, Fan Xiaozhong, Yang Ligong. Entity relation extraction method using semantic pattern [J]. Computer Engineering, 2007, 33(10): 212-214 (in Chinese)
(邓擘, 樊孝忠, 杨立公. 用语义模式提取实体关系的方法[J]. 计算机工程, 2007, 33(10): 212-214)

[13] Wen Chun, Shi Zhaoxiang, Xin Yuan. Chinese non-taxonomic relation extraction based on extended association rule [J]. Computer Engineering, 2009, 35(24): 63-65 (in Chinese)
(温春, 石昭祥, 辛元. 基于扩展关联规则的中文非分类关系抽取[J]. 计算机工程, 2009, 35(24): 63-65)

[14] Aone C, Ramos M. Rees: A large-scale relation and event extraction system [C] //Proc of the 6th Applied Natural Language. Stroudsburg: ACL, 2000: 76-83

[15] Temkin J M, Gilder M R. Extraction of protein interaction information from unstructured text using a context-free grammar [J]. Bioinformatics, 2003, 19(16): 2046-2053

[16] Neelakantan A, Collins M. Learning dictionaries for named entity recognition using minimal supervision [J]. arXiv preprint, arXiv:1504. 06650, 2015

[17] Riloff E, Jones R. Learning dictionaries for information extraction by multi-level bootstrapping [C] //Proc of the Association for the Advance of Artificial Intelligence. Menlo Park, CA: AAAI, 1999: 474-479

[18] Iria J. T-rex: A flexible relation extraction framework [C] //Proc of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK'05). Stroudsburg: ACL, 2005 [2020-05-29]. https://www.researchgate.net/publication/228749479_T-rex_A_flexible_relation_extraction_framework

[19] Schutz A, Buitelaar P. Relext: A tool for relation extraction from text in ontology extension [C] //Proc of the Int Semantic Web Conf. Berlin: Springer, 2005: 593-606

[20] Sabou M, Daquin M, Motta E. SCARLET: Semantic relation discovery by harvesting online ontologies [C] //Proc of European Semantic Web Conf. Berlin: Springer, 2008: 854-858

[21] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations [C] //Proc of the ACL 2004 on Interactive Poster and Demonstration Sessions. Stroudsburg: ACL, 2004: 22-26

[22] Zhou Guodong, Su Jian, Zhang Jie, et al. Exploring various knowledge in relation extraction [C] //Proc of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL, 2005: 427-434

[23] Sun Xia, Dong Lehong. Feature-based approach to chinese term relation extraction [C] //Proc of the Int Conf on Signal Processing Systems. Piscataway, NJ: IEEE, 2009: 410-414

[24] Jiang Jing, Zhai Chengxiang. A systematic exploration of the feature space for relation extraction [C] ///Proc of the Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2007: 113-120

[25] Che Wanxiang, Liu Ting, Li Sheng. et al. Automatic entity relation extraction [J]. Journal of Chinese Information Processing, 2005, 19(2): 2-7 (in Chinese)
(车万翔, 刘挺, 李生. 实体关系自动抽取[J]. 中文信息学报, 2005, 19(2): 2-7)

[26] Guo Xiyue, He Tingting, Hu Xiaohua et al. Chinese named entity relation extraction based-on the syntactic and semantic features [J]. Journal of Chinese Information Processing, 2014, 28(6): 183-189 (in Chinese)
(郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报, 2014, 28(6): 183-189)

[27] Gao Junping, Zhang Hui, Zhao Xujian, et al. Evolutionary relation extraction for domain knowledge in Wikipedia [J]. Chinese Journal of Computers, 2016, 39(10): 2088-2101 (in Chinese)
(高俊平, 张晖, 赵旭剑, 等. 面向维基百科的领域知识演化关系抽取[J]. 计算机学报, 2016, 39(10): 2088-2101)

[28] Gan Xinli, 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 (in Chinese)
(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)

[29] Zelenco D, Aone C, Richardella A. Kernel methods for relation extraction [J]. Journal of Machine Learning Research, 2003, 3(2): 1083-1106

[30] Culotta A, Sorensen J. Dependency tree kernels for relation extraction [C] //Proc of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL, 2004: 423-433

[31] Zhou Guodong, Qian Longhua, Fan Jianxi. Tree kernel-based semantic relation extraction with rich syntactic and semantic information [J]. Information Sciences, 2010, 180(8): 1313-1325

[32] Zhang Xiaofeng, Gao Zhiqiang, Zhu Man. Kernel methods and its application in relation extraction [C] //Proc of the Int Conf on Computer Science and Service System (CSSS). Piscataway, NJ: IEEE, 2011: 1362-1365

[33] Liu Kebin. Implement of a kernel-based Chinese relation extraction system [J]. Journal of Computer Research and Development, 2007,44(8): 1406-1411 (in Chinese)
(刘克彬. 基于核函数中文关系自动抽取系统的实现[J]. 计算机研究与发展, 2007, 44(8): 1406-1411)

[34] Guo Jianyi, Chen Peng, Yu Zhengtao, et al. Domain specific Chinese semantic relation extraction based on composite kernel [J]. Journal of Chinese Information Processing, 2016, 30(1): 24-30 (in Chinese)

(郭剑毅,陈鹏,余正涛,等.基于多核融合的中文领域实体关系抽取[J].中文信息学报,2016,30(1):24-30)

[35] Yu Huanhuan, Qian Longhua, Zhou Guodong, et al. Chinese semantic relation extraction based on unified syntactic and entity semantic tree [J]. Journal of Chinese Information Processing, 2010, 24(5): 17-23 (in Chinese)
(虞欢欢,钱龙华,周国栋,等.基于合一句法和实体语义树的中文语义关系抽取[J].中文信息学报,2010,24(5):17-23)

[36] Brin S. Extracting patterns and relations from the world wide web [C] //Proc of the Int Workshop on The World Wide Web and Databases. Berlin: Springer, 1998: 172-183

[37] Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections [C] //Proc of the 5th ACM Conf on Digital Libraries. New York: ACM, 2000: 85-94

[38] Zhu Jun, Nie Zaiqing, Liu Xiaojing, et al. StatSnowball: A statistical approach to extracting entity relationships [C] //Proc of the 18th Int Conf on World Wide Web. New York: ACM, 2009: 101-110

[39] Carlson A, Betteridge J, Wang R C, et al Coupled semi-supervised learning for information extraction [C] //Proc of the 3rd ACM Int Conf on Web Search and Data Mining. New York: ACM, 2010: 101-110

[40] Gupta P, Roth B, Schütze H. Joint bootstrapping machines for high confidence relation extraction [J]. arXiv preprint, arXiv:1805.00254, 2018

[41] Qin Zhentao, Ye Feiyue. Research on reliability of instance and pattern in semi-supervised entity relation extraction [M]. Recent Developments in Intelligent Computing, Communication and Devices. Singapore: Springer, 2019: 377-385

[42] He Tingting, Xu Chao, Li Jing. et al Named entity relation extraction method based on seed self-expansion [J]. Computer Engineering, 2006, 32(21): 183-184 (in Chinese)
(何婷婷,徐超,李晶,等.基于种子自扩展的命名实体关系抽取方法[J].计算机工程,2006,32(21):183-184)

[43] Yu Li, Feng Lu, Liu Xiliang, et al. A bootstrapping based approach for open geo-entity relation extraction [J]. Acta Geodaetica et Cartographica Sinica, 2016, 45(5): 616-622

[44] Balcan F, Blum A, Yang Ke. Co-training and expansion: Towards bridging theory and practice [C] //Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: Massachusetts Institute of Technology, 2005: 89-96

[45] Abney S. Bootstrapping [C] //Proc of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2002: 360-367

[46] Zhang Zhu. Weakly-supervised relation classification for information extraction [C] //Proc of the 13th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2004: 581-588

[47] Zhu Xiaojin, Ghahramani Z, Lafferty J D. Semi-supervised learning using Gaussian fields and harmonic functions [C] //Proc of the 20th Int Conf on Machine Learning. New York: ACM, 2003: 912-919

[48] Hoffmann R, Zhang Congle, Ling Xiao, et al. Knowledge-based weak supervision for information extraction of overlapping relations [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2011: 541-550

[49] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora [C] //Proc of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL, 2004: 415-422

[50] Rozenfeld B, Feldman R. High-performance unsupervised relation extraction from large corpora [C] //Proc of the 6th Int Conf on Data Mining (ICDM'06). Piscataway, NJ: IEEE, 2006: 1032-1037

[51] Shinyama Y, Sekine S. Preemptive information extraction using unrestricted relation discovery [C] //Proc of Human Language Technology Conf of the North American Chapter of the Association of Computational Linguistics. Stroudsburg: ACL, 2006: 304-11

[52] Davidov D, Rappoport A, Koppel M. Fully unsupervised discovery of concept-specific relationships by web mining [C] //Proc of the 45th Annual Meeting of the Association of Computational Linguistics. Stroudsburg: ACL, 2007: 232-239

[53] Yan Yue, Okazaki N, Matsuo Y, et al. Unsupervised relation extraction by mining wikipedia texts using information from the web [C] //Proc of the Joint Conf of the 47th Annual Meeting of the ACL. Stroudsburg: ACL, 2009: 1021-1029

[54] Bollegala D T, Matsuo Y, Ishizuka M. Measuring the similarity between implicit semantic relations from the Web [C] //Proc of the 18th Int Conf on World Wide Web. New York: ACM, 2009: 651-660

[55] Rink B, Harabagiu S. A generative model for unsupervised discovery of relations and argument classes from clinical texts [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2011: 519-528

[56] Huang Chen, Qian Longhua, Zhou Guodong, et al. Research on unsupervised Chinese entity relation extraction based on convolution tree kernel [J]. Journal of Chinese Information Processing, 2010, 24(4): 11-17 (in Chinese)
(黄晨,钱龙华,周国栋,等.基于卷积树核的无指导中文实体关系抽取研究[J].中文信息学报,2010,24(4):11-17)

[57] Liu An'an, Qin Bing, Liu Ting. Unsupervised Chinese open entity relation extraction [J]. Journal of Computer Research and Development, 2015, 52(5): 1029-1035 (in Chinese)
(刘安安,秦兵,刘挺.无指导的开放式中文实体关系抽取[J].计算机研究与发展,2015,52(5):1029-1035)

[58] Zeng Daojian, Liu Kang, Lai Siwei, et al. Relation classification via convolutional deep neural network [C] //Proc of the 25th Int Conf on Computational Linguistics. Stroudsburg: ACL, 2014: 2335-2344

[59] Liu Chunyang, Sun Wenbo, Chao Wenhan, et al. Convolution neural network for relation extraction [C] //Proc of the Int Conf on Advanced Data Mining and Applications. Berlin: Springer, 2013: 231-242

[60] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. The Journal of Machine Learning Research, 2011, 12(12): 2493-2537

[61] Nguyen T H, Grishman R. Relation extraction: Perspective from convolutional neural networks [C] //Proc of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Stroudsburg: ACL, 2015: 39-48

[62] Santos C N, Xiang Bing, Zhou Bowen. Classifying relations by ranking with convolutional neural networks [J]. arXiv preprint, arXiv:1504.06580, 2015

[63] Xu Kun, Feng Yansong, Huang Songfang, et al. Semantic relation classification via convolutional neural networks with simple negative sampling [J]. arXiv preprint, arXiv: 1506.07650, 2015

[64] Ye Hai, Chao Wenhan, Luo Zhunchen, et al. Jointly extracting relations with class ties via effective deep ranking [J]. arXiv preprint, arXiv:1612.07602, 2016

[65] Fan Bai, Ritter A. structured minimally supervised learning for neural relation extraction [J]. arXiv preprint, arXiv: 1904.00118, 2019

[66] Sun Jiandong, Gu Xiusen, Li Yan, et al. Chinese entity relation extraction algorithms based on COAE2016 datasets [J]. Journal of Shandong University: Natural Science, 2017, 52(9): 7-12 (in Chinese)
(**孙建东, 顾秀森, 李彦, 等.基于** COAE2016 **数据集的中文实体关系抽取算法研究**[J]. **山东大学学报: 理学版**, 2017, 52(9): 7-12)

[67] Gao Dan, Peng Dunlu, Liu Cong. Entity relation extraction based on CNN in large-scale text data [J]. Journal of Chinese Computer Systems, 2018, 39(5): 1021-1026 (in Chinese)
(**高丹, 彭敦陆, 刘丛. 海量法律文书中基于** CNN **的实体关系抽取技术**[J]. **小型微型计算机系统**, 2018, 39(5): 1021-1026)

[68] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces [C] //Proc of the 2012 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: ACL, 2012: 1201-1211

[69] Lin Yankai, Shen Shiqi, Liu Zhiyuan, et al. Neural relation extraction with selective attention over instances [C] //Proc of the 54th Annual Meeting of the Association for Computational Linguistics Stroudsburg: ACL, 2016: 2124-2133

[70] Chen Lin, Miller T, Dligach D, et al. Self-training improves recurrent neural networks performance for temporal relation extraction [C] //Proc of the 9th Int Workshop on Health Text Mining and Information Analysis. Stroudsburg: ACL, 2018: 165-176

[71] Xu Yan, Mou Lili, Li Ge, et al. Classifying relations via long short term memory networks along shortest dependency paths [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 1785-1794

[72] Zhang Shu, Zheng Dequan, Hu Xinchen, et al. Bidirectional long short-term memory networks for relation classification [C] //Proc of the 29th Pacific Asia Conf on Language Information and Computation. Stroudsburg: ACL, 2015: 73-78

[73] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks [C] //Proc of the European Semantic Web Conf. Berlin: Springer, 2018: 593-607

[74] Zhang Yuhao, Qi Peng, Manning C D. Graph convolution over pruned dependency trees improves relation extraction [J]. arXiv preprint, arXiv:1809.10185, 2018

[75] Zhu Hao, Lin Yankai, Liu Zhiyuan, et al. Graph neural networks with generated parameters for relation extraction [J]. arXiv preprint, arXiv:1902.00756, 2019

[76] Song Linfeng, Zhang Yue, Wang Zhiguo, et al. N-ary relation extraction using graph state LSTM [J]. arXiv preprint, arXiv:1808.09101, 2018

[77] Guo Zhijiang, Zhang Yan, Lu Wei. Attention guided graph convolutional networks for relation extraction [J]. arXiv preprint, arXiv:1906.07510, 2019

[78] Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures [J]. arXiv preprint, arXiv: 1601.00770, 2016

[79] Zhou Peng, Shi Wei, Tian Jun, et al. Attention-based bidirectional long short-term memory networks for relation classification [C] //Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016: 207-212

[80] Li Zhiheng, Yang Zhihao, Shen Chen, et al. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text [J]. BMC Medical Informatics and Decision Making, 2019, 19(1): 43-50

[81] Chen Yu, Zheng Dequan, Zhao Tiejun. Chinese relation extraction based on deep belief nets [J]. Journal of Software, 2012, 23(10): 2572-2585 (in Chinese)
(**陈宇, 郑德权, 赵铁军. 基于** Deep Belief Nets **的中文名实体关系抽取**[J]. **软件学报**, 2012, 23(10): 2572-2585)

[82] Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures [J]. arXiv preprint, arXiv: 1601.00770, 2016

[83] Zheng Suncong, Hao Yuexing, Lu Dongyuan, et al. Joint entity and relation extraction based on a hybrid neural network [J]. Neurocomputing, 2017, 257(12): 59-66

[84] Zheng Suncong, Wang Feng, Bao Hongyun, et al. Joint extraction of entities and relations based on a novel tagging scheme [J]. arXiv preprint, arXiv: 1706.05075, 2017

[85] Bekoulis G, Deleu J, Demeester T, et al. Joint entity recognition and relation extraction as a multi-head selection problem [J]. Expert Systems with Applications, 2018, 114: 34-45

[86] Bekoulis G, Deleu J, Demeester T, et al. Adversarial training for multi-context joint entity and relation extraction [J]. arXiv preprint, arXiv:1808.06876, 2018

[87] Wang Shaolei, Zhang Yue, Che Wanxiang, et al. Joint extraction of entities and relations based on a novel graph scheme [C] //Proc of the Int Joint Conf on Artificial Intelligenc. Amsterdam: Elsevier, 2018: 4461-4467

[88] Fu Tujiu, Li Pengsuan, Ma Weiyun. Graphrel: Modeling text as relational graphs for joint entity and relation extraction [C] //Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 1409-1418

[89] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data [C] //Proc of the 47th Annual Meeting of the ACL. Stroudsburg: ACL, 2009: 1003-1011

[90] Alfonseca E, Filippova K, Delort J Y, et al. Pattern learning for relation extraction with a hierarchical topic model [C] //Proc of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2012: 54-59

[91] Takamatsu S, Sato I, Nakagawa H. Reducing wrong labels in distant supervision for relation extraction [C] //Proc of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2012: 721-729

[92] Ji Guoliang, Liu Kang, He Shizu, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 3060-3066

[93] Yu Xiaokang, Chen Ling, Guo Jing, et al. Relation extraction method combining clause level distant supervision and semi-supervised ensemble learning [J]. Pattern Recognition and Artificial Intelligence, 2017, 30(1): 54-63 (in Chinese)
(**余小康, 陈岭, 郭敬, 等. 结合从句级远程监督与半监督集成学习的关系抽取方法**[J]. **模式识别与人工智能**, 2017, 30(1): 54-63)

[94] Wang Guanying, Zhang Wen, Wang Ruoxu, et al. Label-free distant supervision for relation extraction via knowledge graph embedding [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 2246-2255

[95] Ru Chengsen, Tang Jintao, Li Shasha, et al. Using semantic similarity to reduce wrong labels in distant supervision for relation extraction [J]. Information Processing & Management, 2018, 54(4): 593-608

[96] Huang Zhaowei, Chang Liang, Bin Chenzhong, et al. Distant supervision relationship extraction based on GRU and attention mechanism [J]. Application Research of Computer, 2019, 36(10): 1-7 (in Chinese)
(**黄兆玮, 常亮, 宾辰忠, 等. 基于** GRU **和注意力机制的远程监督关系抽取**[J]. **计算机应用研究**, 2019, 36(10): 1-7)

[97] Fan Miao, Zhao Deli, Zhou Qiang, et al. Distant supervision for relation extraction with matrix completion [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014: 839-849

[98] Zeng Daojian, Dai Yuan, Li Feng, et al. Adversarial learning for distant supervised relation extraction [J]. Computers, Materials & Continua, 2018, 55(1): 121-136

[99] He Zhonghe, Zhou Zhongcheng, Gan Liang, et al. Chinese entity attributes extraction based on bidirectional LSTM networks [J]. Journal of Computational Science and Engineering, 2019, 18(1): 65-71

[100] Wu Fei, Weld D S. Open information extraction using Wikipedia [C] //Proc of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2010: 118-127

[101] Nakashole N, Weikum G, Suchanek F. PATTY: A taxonomy of relational patterns with semantic types [C] //Proc of the 2012 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: ACL, 2012: 1135-1145

[102] Yao Limin, Riedel S, Mccallum A. Collective cross-document relation extraction without labelled data [C] //Proc of the 2010 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2010: 1013-1023

[103] Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2011: 1535-1545

[104] Xavier C C, De lima V L S. Boosting open information extraction with noun-based relations [C] //Proc of the 9th Int Conf on Language Resources and Evaluation. Stroudsburg: ACL, 2014: 96-100

[105] Covro L D, Gemulla R. Clausie: Clause-based open information extraction [C] //Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013: 355-366

[106] Faruqui M, Kumar S. Multilingual open relation extraction using cross-lingual projection [J]. arXiv preprint, arXiv: 1503.06450, 2015

[107] Song Shengli, Sun Yulong, Di Qiang. Multiple order semantic relation extraction [J]. Neural Computing and Applications, 2019, 31: 4563-4576

[108] Akbik A, Löser A. Kraken: N-ary facts in open information extraction [C] //Proc of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction. Stroudsburg: ACL, 2012: 52-56

[109] Gamallo P, garcia M, Fernández-lanza S. Dependency-based open information extraction [C] //Proc of the Joint Workshop on Unsupervised and Semi-supervised Learning in NLP. Stroudsburg: ACL, 2012: 10-18

[110] Fossati M, Dorigatti E, Giuliano C. N-ary relation extraction for simultaneous T-Box and A-Box knowledge base augmentation [J]. Semantic Web, 2018, 9(4): 413-439

[111] Petroni F, Corro L, GEMULLA R. Core: Context-aware open relation extraction with factorization machines [C] // Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 1763-1773

[112] Qiu Likun, Zhang Yue. ZORE: A syntax-based system for Chinese open relation extraction [C] //Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1870-1880

[113] Qin Bing, Liu An'an, Liu Ting. Unsupervised Chinese open entity relation extraction [J]. Journal of Computer Research and Development, 2015, 52(5): 1029-1035 (in Chinese)
(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)

[114] Guo Xiyue, He Tingting. Leveraging Chinese encyclopedia for weakly supervised relation extraction [C] //Proc of the Joint Int Semantic Technology Conf. Berlin: Springer, 2015: 127-140

[115] Jia Shengbin, Li Maozhen, Xiang Yang. Chinese open relation extraction and knowledge base establishment [J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2018, 17(3): 15-22

[116] Li Ying, Hao Xiaoyan, Wang Yong. N-ary Chinese open entity-relation extraction [J]. Computer Science. 2017, 44 (S1): 80-83 (in Chinese)
(李颖, 郝晓燕, 王勇.中文开放式多元实体抽取[J]. 计算机科学, 2017, 44(增刊 1): 80-83)

[117] Yao Xianming, Gan Houjian, Xu Jian. Chinese open domain oriented n-ary entity relation extraction [J]. CAAI Transactions on Intelligent Systems, 2019, 14(3): 597-604 (in Chinese)
(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)

[118] Bird S, Klein E, Loper E. Natural language processing with Python: Analyzing text with the natural language toolkit [G] //Natural Language Toolkit. Sebastopol, CA: O'Reilly Media, 2009

[119] Stanford University. DeepDive: Version 0.8.0 [EB/OL]. [2019-04-28]. http://deepdive.stanford.edu/

[120] Manning C, Surdeanu M, Bauer J, et al. The Stanford CoreNLP natural language processing toolkit [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014: 55-60

[121] Harbin Institute of Technology. ITP-Cloud [EB/OL]. [2019-04-28]. http://www.ltp-cloud.com/

[122] Li Peifeng, Zhou Guodong, Zhu Qiaoming, et al. Employing compositional semantics and discourse consistency in Chinese event extraction [C] //Proc of the 2012 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: ACL, 2012: 1006-1016

[123] Bach N, Badaskar S. A review of Relation extraction [J]. Literature Review for Language and Statistics II, 2007, 2 (8): 25-36

[124] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507

[125] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint, arXiv:1810.04805, 2018

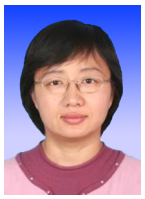
[126] Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains [C] //Proc of 2005 IEEE Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2005: 729-734

[127] E Honghai, Zhang Wenjing, Xiao Siqi, et al. A survey of entity relationship extraction based on deep learning [J]. Journal of Software, 2019, 30(6): 1793-1818 (in Chinese)
(鄂海红, 张文静, 肖思琪. 深度学习实体关系抽取研究综述[J]. 软件学报, 2019, 30(6): 1793-1818)

[128] Shi Peng, Lin Jimmy. Simple BERT models for relation extraction and semantic role labeling [J]. arXiv preprint, arXiv:1904.05255, 2019

[129] Shen Tielin, Wang Daling, Feng Shi, et al. BERT-based denoising and reconstructing data of distant supervision for relation extraction [J]. arXiv preprint, arXiv:1908.11337, 2019

[130] Qin Pengda, Xu Weiran, Wang Yang. Robust distant supervision relation extraction via deep reinforcement learning [J]. arXiv preprint, arXiv:1805.09927, 2018



Li Dongmei, born in 1972. PhD, associate professor. Her main research interests include natural language processing and knowledge graph.

Zhang Yang, born in 1995. Master candidate. His main research interests include natural language processing and knowledge graph.

Li Dongyuan, born in 1995. Master. His main research interests include machine learning and natural language processing.

Lin Danqiong, born in 1994. Master. Her main research interests include machine learning and natural language processing.