文章编号：1003-0077(2019)10-0010-08

基于远程监督的关系抽取研究综述

白 龙，靳小龙，席鹏弼，程学旗

（1. 中国科学院 计算技术研究所 中国科学院网络数据科学与技术重点实验室，北京 100190；
2. 中国科学院大学 计算机与控制学院，北京 100190）

摘 要：关系抽取作为信息抽取的一项关键技术，在知识库自动构建、问答系统等领域有着极为重要的意义，一直以来受到人们的关注。远程监督关系抽取技术通过外部知识库作为监督源，自动对语料库进行标注，能够大量节省人工标注成本，因而受到了研究者们的重视。该文针对远程监督关系抽取技术做了较为系统性的梳理，将已有方法分为基于概率图的、基于矩阵补全的和基于嵌入的三大类，并且对其当前面临的挑战进行了探讨，最后总结并展望了远程监督关系抽取技术未来的发展。

关键词：远程监督；关系抽取；信息抽取

中图分类号：TP391 **文献标识码**：A

A Survey on Distant Supervision Based Relation Extraction

BAI Long, JIN Xiaolong, XI Pengbi, CHENG Xueqi

(1. CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: As a key technique of information extraction, relation extraction is of great importance to many tasks such as automatic knowledge base construction and question answering systems. Distant supervision for relation extraction uses an external knowledge base as supervision signals to automatically label corpus, which can reduce the high cost of manual labelling. This paper presents a systematic survey to distantly supervised relation extraction. It classifies the existing methods into three types, including probabilistic graph-based, matrix completion-based and embedding-based ones. This paper also discusses the challenges and the future research directions of distantly supervised relation extraction.

Keywords: distant supervision; relation extraction; information extraction

0 引言

信息抽取是自然语言处理领域的一个子领域，它的目标是从非结构化数据中挖掘结构化信息。关系抽取是信息抽取的一项关键技术，其目的是挖掘实体之间存在的语义关系。关系抽取对于知识库自动构建、问答系统等领域有着极为重要的意义。根据论元个数不同，关系一般可以分为二元关系和多元关系，目前知识库中最常见的是二元关系，这也是目前关系抽取领域主要研究的关系类型。二元关系建立在两个实体之上，表达了两个实体之间存在的某种语义联系，一般称这两个实体为头实体和尾实体，加上关系，则构成一个三元组。关系抽取算法一般以三元组列表作为输出。

现有的关系抽取方法可以分为 4 类，分别是有监督关系抽取、半监督关系抽取、远程监督关系抽取和无监督关系抽取。有监督关系抽取方法将关系抽取建模为分类问题：针对句子 s，若有实体对$\langle e_1, e_2\rangle$在句子 s 中出现，则对这个实体对进行分类，判断它

收稿日期：2019-02-17 **定稿日期**：2019-06-20

基金项目：国家重点研发计划(2016YFB1000902)；国家自然科学基金(61772501, 61572473, 61572469, 91646120)

们属于哪种关系。然而，有监督方法需要大量有标注数据，这使得获取训练数据的成本较高。半监督关系抽取方法同时使用少量有标注数据和大量无标注数据，从而降低算法对于有标注数据的依赖性，代表性的半监督关系抽取方法有基于自举的方法[1]、基于主动学习的方法[2]和基于标签传播的方法[3]等。基于远程监督的关系抽取方法由 Mintz 等[4]提出，通过外部知识库代替人对语料进行标注，从而可以低成本地获取大量有标注数据，进而通过分类方法进行关系抽取。无监督的关系抽取一般用于待抽取关系未知的开放领域。开放领域关系抽取的概念由 Banko 等[5]提出，通过对语料的句法模式进行学习，从而识别出关系短语和相关的参数，无须标注数据便可以抽取实体间的关系。

由于远程监督关系抽取方法能够极大地减少标注成本，因而近年来受到了研究者们的关注。然而我们观察到，在远程监督关系抽取方面，目前尚缺乏较为系统的梳理。为此，本文将对远程监督关系抽取当前的方法进行归纳总结，并展望该任务的未来。本文首先将对远程监督关系抽取方法进行简要介绍，然后详细说明几种主要方法，接着叙述当前远程监督关系抽取的一些挑战，最后对该任务未来的研究方向进行展望。

1　远程监督关系抽取方法简介

远程监督关系抽取由 Mintz 等[4]首先提出，Mintz 等人认为这是有别于有监督、半监督和无监督关系抽取的第 4 类方法，且兼取了其他 3 类方法的优点。其主要假设是：假如两个实体之间存在某种关系，那么所有这两个实体共现的句子都有可能表达这种关系。根据这一假设，Mintz 等使用维基百科数据对无标注语料进行自动标注，并使用分类方法求解关系抽取问题。

Riedel 等[6]将 Mintz 等[4]提出的假设称为“远程监督假设(distant supervision assumption)”，并认为这一假设过强，从而提出了“至少一次假设(at-least-once assumption)”，该假设表述如下：若两个实体之间存在某种关系，那么在所有这两个实体共现的句子中，至少有一句表达了这种关系。在此假设下，Riedel 等将远程监督关系抽取建模为多实例学习(multi-instance learning)问题，将一个实体对共现的所有句子聚合成一个句袋(bag)，并对句袋进行分类。换言之，多实例学习的方法只关注句袋体现了实体对之间的哪些关系，而并不关注每个句子表达了哪种关系。Riedel 等人认为，这样做有 3 个优点：第一，与实践更加契合；第二，能聚合各处的证据来更好地判断关系是否成立；第三，简化了机器学习任务。

Hoffmann 等[7]与 Surdeanu 等[8]观察到，在一个实体对上可能有不止一种关系成立，因此在多实例学习的基础上引入了多标签学习方法，Surdeanu 等人将此总结为多实例多标签学习(multi-instance multi-label learning)。

远程监督关系抽取方法需要将一个无标注语料库对齐到已知的知识库。Mintz 等[4]将维基百科词条页面对齐到 Freebase[9]；Riedel 等将纽约时报语料库[10]对齐到 Freebase(以下称为 NYT 数据集)；Surdeanu 等将 KBP-2010[11]、KBP-2011[12]评测任务中给定的文档对齐到相应的知识库(以下称为 KBP 数据集)；Zeng 等[13]将纽约时报语料库对齐到 Wikidata[14]。其中，最为常用的数据集是 Riedel 等将纽约时报对齐到 Freebase 形成的 NYT 数据集。

在模型评估方面，Mintz 等设计了两部分评估：自动评估和人工评估。需要人工评估的原因是知识库是不完备的，因此可能存在预测正确的关系由于不在知识库中而被误判为预测错误的情况。Mintz 等采用“准确率—召回率”曲线来作为模型性能的评价指标。根据 Lin 等的实验结果，当前一些常用的基准模型的性能如图 1 所示。其中，Mintz 模型由 Mintz 等[4]提出，MultiR 模型由 Hoffmann 等[7]提出，MIML 模型由 Surdeanu 等[8]提出，CNN＋ATT 和 PCNN＋ATT 模型皆由 Lin 等[15]提出。

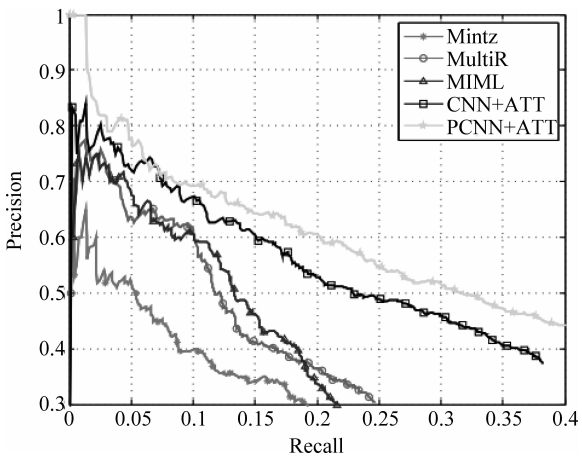

图 1　P-R 曲线图[15]

当前远程监督关系抽取方法大体可以分为 3

类，分别是基于概率图的方法、基于矩阵补全的方法以及基于嵌入的方法。以下3节将分别介绍3种方法的代表性工作，以及各自所面临的挑战。

2 基于概率图的方法

本节主要介绍基于概率图的方法。基于概率图的方法将句袋和句子的标签视为隐变量，将关系抽取视为对隐变量赋值的过程。Riedel等[6]提出的模型中，句袋(文中称为关系relation)隐变量为Y(文中称为关系变量relation variable)，其取值为$y\in R$，R为所有关系类型组成的集合，用以表示句袋的标签；句袋中每个句子(文中称为关系指涉relation mention)隐变量为Z_i(文中称为关系指涉变量relation mention variable)，其取值$z_i\in\{0,1\}$，用以表示该句是否真实表达了句袋标签Y所对应关系，其模型结构如图2所示。Hoffmann等[7]、Surdeanu等[8]将该问题扩展为多标签分类问题。两者将句袋的隐变量扩展为$|R|$个，每个隐变量Y^r的取值为$y^r\in\{0,1\}$，其中$r\in R$，用以表示该句袋在每个关系上是否成立；句袋中每个句子的隐变量为Z_i，其取值为$z_i\in R$，用以表示每个句子所表达的关系。Hoffmann等设计的模型MultiR结构如图3所示，Surdeanu等设计的模型MIML结构如图4所示。

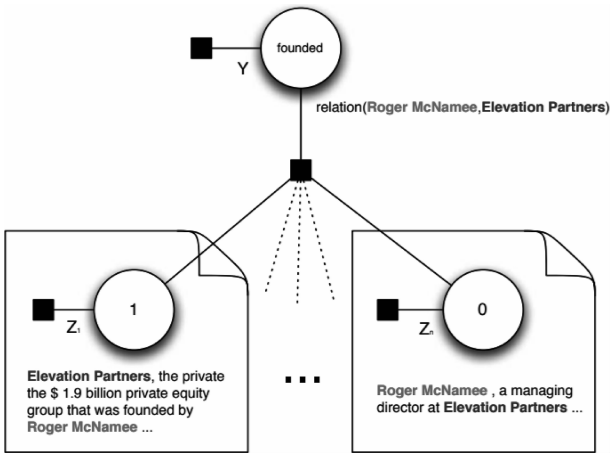

图2 因子图模型示例

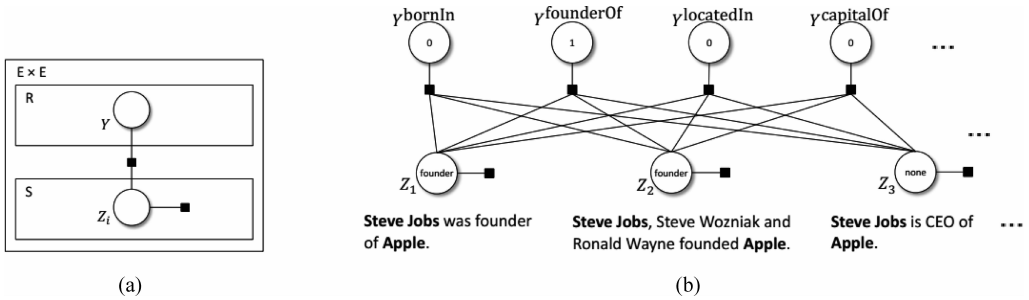

图3 (a)MultiR盘式记法，(b)实体对为〈Steve Jobs，Apple〉的示例

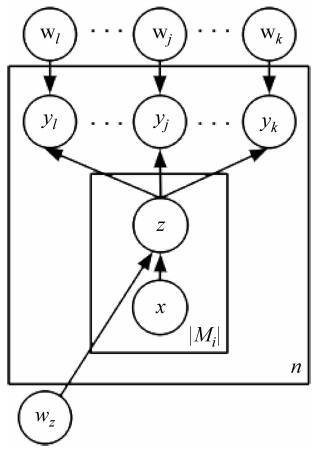

图4 MIML的盘式记法

Surdeanu等人提到，MIML模型与MultiR模型的区别主要有两点。第一，MultiR直接通过取并集的方式聚合每个实例的标签，而MIML模型通过句袋层面的分类器捕捉不同标签之间的依赖关系；第二，MultiR采用类感知机风格的参数更新策略，而MIML在贝叶斯框架内进行训练。Surdeanu认为这两点是MIML性能优于MultiR的原因。

在此基础上，又有许多研究者通过不同方法进行了改进。由于远程监督与有监督方法最大的不同就是错误标注所带来的噪声，大部分工作都关注如何降噪，Takamatsu等[16]、Min等[17]、Xu等[18]、Ritter等[19]分别提出了不同的模型。其中Takamatsu提出了一种生成模型，用以建模自动标注的模式，从而发现其中的错误标注；Min等、Xu等关注知识库不完备所产生的伪反例，其中Min等[17]认为训练集中的句袋标签为观测标签，并加入了隐变量l，用以表示句袋的真实标签，Xu等[18]在训练集中采用排序学习的方法检索到相似文档，从而改善训练集的标注质量；Ritter等[19]的模型同时关注伪正例和伪

反例两类噪声，并将其统称为“缺失数据（missing data）”，他们采用了一种软约束的方法，允许句子标签和句袋标签在一定程度上可以不一致，从而缓解了缺失数据的问题。

基于概率图的方法当前存在的最大问题是，模型依赖于人工定义的特征函数。此类模型往往存在高准确率、低召回率的问题，当我们综合考虑准确率和召回率时，基于概率图的模型就存在一定的局限性。其次，特征需要使用其他自然语言处理工具抽取，例如，句法分析和依存分析工具。此类工具在分析时产生的错误会向后传递，影响预测精度[20]。

3　基于矩阵补全的方法

本节将主要介绍基于矩阵补全的方法。Fan 等[21]首先提出了使用矩阵补全的方法处理远程监督关系抽取问题。该方法将训练样本、测试样本拼成一个矩阵 Z，矩阵分为 4 块，左上块是训练样本的特征 X_{train}，右上块是训练样本的标签 Y_{train}，左下块是测试样本的特征 X_{test}，右下块是未知的测试样本标签 Y_{test}，该方法的目的就是补全矩阵的右下块，如式(1)所示。

$$\mathrm{Z}=\begin{pmatrix}X_{\text{train}} & Y_{\text{train}}\\ X_{\text{test}} & Y_{\text{test}}\end{pmatrix}=\mathrm{Z}^{*}+\mathrm{E} \tag{1}$$

在远程监督关系抽取问题中，特征、标签都含有噪声，因此 Fan 等认为观测到的矩阵 Z 是由一个低秩矩阵 Z^{*} 加上一个噪声矩阵 E 所形成的，如图 5 所示。图 5 最左边的矩阵代表观测矩阵，其左上角为训练样本的特征，右上角为训练样本的标签，左下角为测试样本的特征，需要预测的是右下角测试样本的标签。

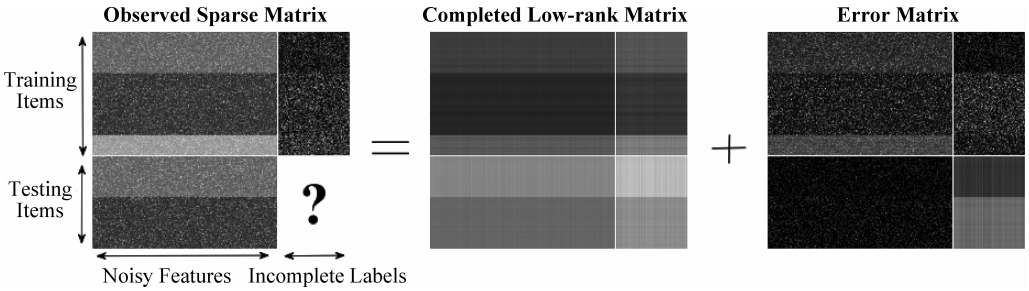

图 5　矩阵补全示意图

Fan 等人将远程监督关系抽取问题转化为：在观测到 X_{train}，Y_{train}，X_{test} 后补全 Y_{test}，使得矩阵 Z 的秩最小化。由于矩阵秩最小化的问题是困难的，他们采取了最小化核范数作为替代，达到了较好的效果。

Zhang 等[22]在此基础上提出了基于非参数贝叶斯的模型，该模型分为 3 个部分，第一部分建模噪声，将噪声 ε 自发地聚为 K 类；第二部分是矩阵最小化的秩，将低秩矩阵 Z^{*} 分解为两个矩阵 U 和 V，通过概率方法求得；第三部分，在求得 U，V，ε 的基础上，对测试集标签 Y_{test} 进行预测。该方法的优点在于使用非参数方法避免了大量的调参。

Fan[21]等人提到，基于矩阵补全的方法的缺点在于，假如新来了一批测试数据，该方法必须重新构建矩阵 Z 进行补全，不能如其他类型的方法一样增量地进行抽取。

4　基于嵌入的方法

基于嵌入的方法将词映射到向量空间中，通过向量运算来表达词与词之间存在的语义关联。该方法最早由 Weston 等[23]提出，通过对“文档—关系”的相似度打分来预测实体对之间的关系，他们同时将 Bordes 等[24]提出的 TransE 模型引入关系抽取问题中，将知识库的信息用向量表达，增强了关系抽取模型的性能。Weston 等认为句子的嵌入向量是其包含的所有词的嵌入向量之和，这可以视为传统词袋模型与嵌入模型的结合。该方法并没有将词的位置信息考虑在内，使用的模型也偏浅层，无法挖掘词与词之间更加深层的依存关系。

Zeng 等[20]首先将深度学习模型用于远程监督关系抽取。他们在卷积神经网络的基础上提出了 PCNN 模型，其模型如图 6 所示。在输入方面，Zeng 等不仅考虑了词本身的语义信息，还考虑了每个词相对其他实体词的位置信息；在池化层方面，Zeng 等采用了分段池化的方法，避免了对隐层节点的过度削减，从而保留了更加细粒度的信息，这一模型也被之后的研究者广泛采纳。同时，Zeng 等还提出了基于 PCNN 模型的多实例学习方法，他们认

为,一个句袋的预测值,是其中每个句子的预测值中 最大的一个。

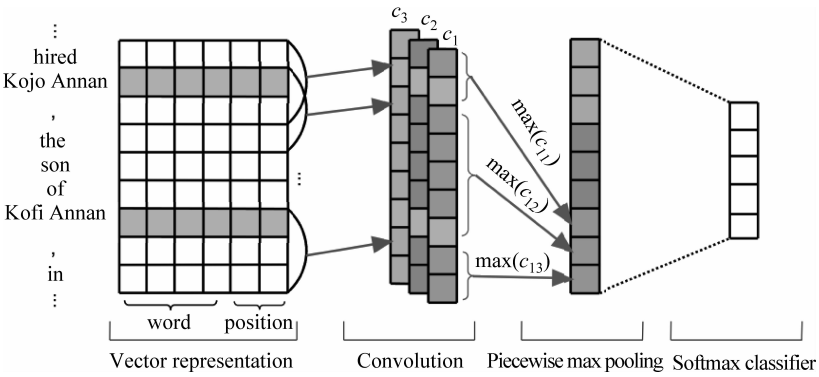

图 6 PCNN 结构图

Lin 等[15]在 Zeng 等的基础上提出了基于注意力机制的多实例学习策略,其模型如图 7 所示。他们使用关系向量作为注意力,对一个句袋中的每个句子分配权重,表示在该关系上该句的重要程度,并对句子向量进行加权求和,得到句袋嵌入向量,并直接对句袋进行多分类。在测试时,由于不知道句袋的真实关系,因此要穷举每一个关系用以计算注意力,在最终汇总预测概率时,也是选取了对每一个关系的最大预测概率值。

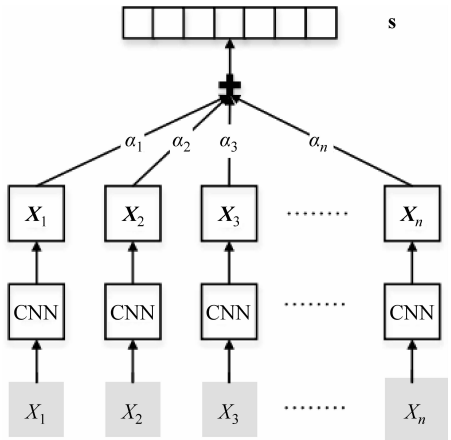

图 7 注意力模型结构图

自从强化学习在围棋领域上取得了突破性的进展,越来越多的研究者试图将强化学习应用在其他领域。在远程监督关系抽取方面,也有这方面的尝试。Feng 等[25]、Zeng 等[26]分别提出了两种基于强化学习的远程监督关系抽取模型。如图 8 所示,Feng 的模型有两个部件,分别是实例选择器(instance selector)和关系分类器(relation classifier),实例选择器对每个句子 x_i 执行动作 a_i,选择是否将 x_i 放入训练集;分类器对样本进行预测,并根据预测结果将回报(reward)回传给实例选择器,帮助实例选择器更好地选择训练样本。需要注意的是,虽然 Feng 等保留了多实例学习里句袋(bag)的概念,但是这是为了使训练过程更有效率,保留更多的反馈信息。关系分类器的输入实际上仍是句子而非句袋,所以该模型可以用于句子级的关系抽取。Zeng 等从另一个角度使用了实例选择器,他们将“至少一次”假设进行了重新表述:在对句袋进行关系预测时,句袋是“无关系”当且仅当每个句子的预测标签都是“无关系”,否则句袋的关系类型由其句子所表达。Zeng 等通过强化学习实现了对“无关系”和其他关系类型的有区分的预测,其模型如图 9 所示。

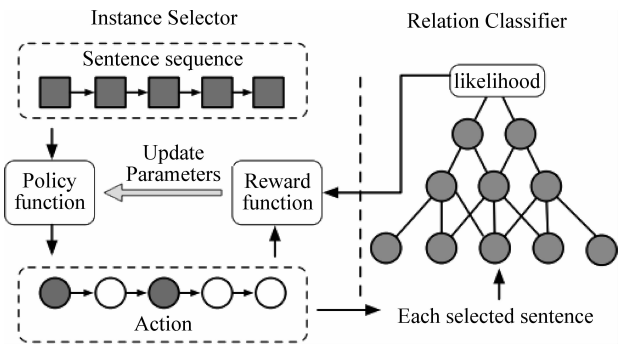

图 8 Feng 等的强化学习模型

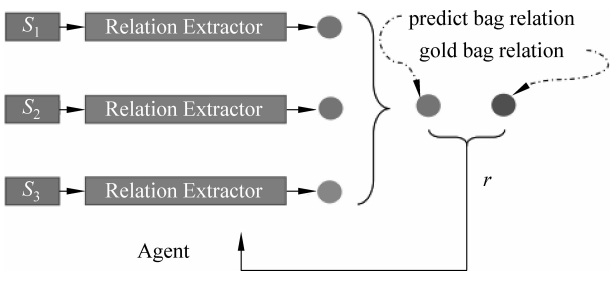

图 9 Zeng 等的强化学习模型

5　远程监督关系抽取的挑战

5.1　错误标注

错误标注(wrong labeling)问题是远程监督关系抽取最主要的问题。由于句子的标签是通过知识库自动标注的，因此在标注结果中混入了大量的错误标注。目前应对错误标注问题大致有两种方法，一种常用的方法是实例选择方法，即通过降低错误标注实例的权重(或者将其移出训练集)降低噪声；另一种方法是由 Fan 等[21]、Luo 等[27]使用的噪声建模方法，通过建模噪声产生的过程，来还原真实的标签。

在错误标注问题中，有两类特殊的错误标注值得注意。其中一类是由“至少一次假设”失效所造成的错误标注。即针对一个实体对，虽然有实体对共现的句子，但这些句子无一体现了两实体之间的关系。“至少一次假设”认为，当句子聚合成为句袋时，每种关系至少对应着一个句子，而该假设失效所造成的错误标注实例，可以认为是一种句袋级别的伪正例。Takamatsu 等[16]、Ritter 等[19]、Luo 等[27]都提到了这一现象，并提出了自己的模型以应对该问题。另外，Zeng 等[13]提出了融合关系路径预测的关系抽取模型，若当前模型不足以预测实体对$\langle h,t\rangle$的关系，但是对$\langle h,e\rangle$，$\langle e,t\rangle$预测置信度较高时，可以用后者辅助前者的预测。虽然 Zeng 等[13]并未提及，但我们认为这一模型也可能有助于缓解伪正例的问题。

另一类错误标注是由知识库不完备所造成的伪反例，即某个实体对在句子中确实表明了某种关系，但由于知识库中不存在该信息，因此机器标注时将该句标为无关系。该问题在训练集和测试集中都有出现，并且会较大程度地影响测试结果。由于伪反例的出现，测试集的标签实际上是不准确的，这也是 Mintz 等[4]在自动评估之后还要进行手动评估的原因。Ritter 等[19]的模型也试图解决这一问题，他们将伪正例、伪反例合称为“缺失数据(missing data)”，伪正例是在语料库方面缺失数据，伪反例是在知识库方面缺失数据。

这两类错误标注问题都具有相当大的挑战性，而且对于模型的性能较难评估。

5.2　其他挑战

远程监督关系抽取问题常常被建模为一种多标签分类问题，因此多标签分类问题的策略也会应用于远程监督关系抽取问题，例如，标签之间的相关性。Feng 等[28]、Ye 等[29]分别用不同的模型建模了关系之间的相关性。此外，Zeng 等[13]提出的融合关系路径预测的关系抽取模型，也可以视为用另外两个关系去预测某个关系的置信度。

实体—关系联合抽取模型也在远程监督关系抽取问题中得到应用，如 Zheng 等[30]采用了一种新的序列标注模型，同时标注实体和关系；Ren 等[31]提出了 CoType 模型，同时对实体的细粒度类别和关系进行抽取。

此外，将关系抽取扩展到多元、多句关系抽取也是一个新的方向。Quirk 和 Poon[32]、Peng 等[33]在此方面做了一些探索，然而该模型也只是针对一种关系类型进行抽取，如何扩展到多种不同类型的关系抽取，仍有待研究。在多句关系抽取问题中，如何进行语料库的标注也成为了一个新的问题，Quirk 和 Poon[32]给出了一种启发式的标注规则，更加有效的标注规则也有待进一步研究。

6　总结与展望

基于远程监督的关系抽取方法，能够用较低的成本获取大量的训练数据，近年来越来越受到研究者的关注。本文从远程监督关系抽取方法面临的几方面挑战出发，大致梳理了前人针对这些挑战所提出的各种方法。

回顾过去工作，我们可以看到：

(1) 自从深度学习模型被引入这一领域之后，已取得了相当大的成功，成为最受研究者青睐的方法，对于深度学习模型的研究会是今后一段时间研究者们非常关注的问题。

(2) 强化学习在其他领域的成功也受到了本领域研究者们的关注，今后也许会有更多的工作基于强化学习模型。

(3) 除了应对自动标注错误产生的伪正例之外，如何应对“至少一次假设”失效产生的伪正例和知识库不完备产生的伪反例，仍然是一个难点，有待更多人进行更深入的研究。

(4) 在研究各种降噪方法之外，研究者们还试图扩展远程监督关系抽取技术的应用场景，开始尝试跨句的关系抽取和多元关系抽取。

(5) 对于远程监督关系抽取模型的评估方法，仍有不尽如人意的地方，由于测试集存在错误标注的问题，自动评估无法较好地展现模型的性能，而人工评估

代价较大,更好的评估方法和评估指标仍有待探索。

关系抽取作为信息抽取技术的一环,一直以来受到了相当广泛的关注。远程监督作为一种可以低成本获取标注数据的方法,也受到了人们的重视。本文详细阐述了解决远程监督关系抽取问题的一些具有代表性的方法,对其研究现状与挑战进行了总结,并在此基础上,对未来的研究方向进行了分析和展望。

参考文献

[1] Brin S. Extracting patterns and relations from the World Wide Web[C]//Proceedings of International Workshop on The World Wide Web and Databases. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999: 172-183.

[2] Sun A. Grishman R. Active learning for relation type extension with local and global data views[C]//Proceedings of the 21st ACM international conference on Information and knowledge management. Maui, Hawaii, USA: ACM, 2012: 1105-1112.

[3] Chen J, Ji D, Tan C L, et al. Relation extraction using label propagation based semi-supervised learning [C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006: 129-136.

[4] 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: Association for Computational Linguistics, 2009: 1003-1011.

[5] 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: Morgan Kaufmann Publishers Inc., 2007: 2670-2676.

[6] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[C]//Proceedings of the Springer Berlin Heidelberg, Heidelberg, 2010: 148-163.

[7] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, 2011: 541-550.

[8] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction [C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: Association for Computational Linguistics, 2012: 455-465.

[9] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. Vancouver, Canada: ACM, 2008: 1247-1250.

[10] Sandhaus E. The new york times annotated corpus [J]. Linguistic Data Consortium, Philadelphia, 2008. 6(12): e26752.

[11] Ji H, Grishman R, Dang H T, et al. Overview of the TAC 2010 knowledge base population track[C]//Proceedings of the Third Text Analysis Conference (TAC 2010). 2010: 3-13.

[12] Ji H, Grishman R, Dang H. Overview of the TAC2011 Knowledge Base Population Track[C]// Proceedings Text Analysis Conference. 2011.

[13] Zeng W, Lin Y, Liu Z, et al. Incorporating relation paths in neural relation extraction[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 1768-1777.

[14] Vrande D. Wikidata: a free collaborative knowledgebase[J]. Commun. ACM, 2014. 57(10): 78-85.

[15] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics Berlin, Germany: Association for Computational Linguistics, 2016: 2124-2133.

[16] Takamatsu S, Sato I, Nakagawa H. Reducing wrong labels in distant supervision for relation extraction [C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea: Association for Computational Linguistics, 2012: 721-729.

[17] Min B, Grishman R, Wan L, et al. Distant supervision for relation extraction with an incomplete knowledge base[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013: 777-782.

[18] Xu W, Hoffmann R, Zhao L, et al. Filling knowledge base gaps for distant supervision of relation extraction[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: Association for Computational Linguistics, 2013: 665-670.

[19] Ritter A, Zettlemoyer L, Mausam, et al, Modeling missing data in distant supervision for information Extraction[J]. Transactions of the Association for Computational Linguistics, 2013, 1: 367-378.

[20] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015: 1753-1762.

[21] Fan M, Zhao D, Zhou Q, et al. Distant supervision for relation extraction with matrix completion[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland: Association for Computational Linguistics, 2014: 839-849.

[22] Zhang Q, Wang H. Noise-Clustered distant supervision for relation extraction: A nonparametric bayesian perspective[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 1808-1813.

[23] Weston J, Bordes A, Yakhnenko O, et al. Connecting language and knowledge bases with embedding models for relation extraction[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013: 1366-1371.

[24] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of Advances in neural information processing systems. 2013: 2787-2795.

[25] Feng J, Huang M, Zhao L, et al. Reinforcement learning for relation classification from noisy data [C]//Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence. AAAI Press, 2018.

[26] Zeng X, He S, Liu K, et al. Large scaled relation extraction with reinforcement learning[C]//Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence. AAAI Press, 2018.

[27] Luo B, Feng Y, Wang Z, et al. Learning with noise: enhance distantly supervised relation extraction with Dynamic Transition Matrix[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017: 430-439.

[28] Feng X, Guo J, Qin B, et al. Effective deep memory networks for distant supervised relation extraction [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI Press, 2017: 4002-4008.

[29] Ye H, Chao W, Luo Z, et al. Jointly extracting relations with class ties via effective deep ranking[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1810-1820.

[30] Zheng S, Wang F, Bao H, et al. Joint extraction of entities and relations based on a novel tagging Scheme [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1227-1236.

[31] Ren X, Wu Z, He W, et al. CoType: Joint extraction of typed entities and relations with knowledge bases[C]//Proceedings of the 26th International Conference on World Wide Web. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017: 1015-1024.

[32] Quirk C, Poon H. Distant supervision for relation extraction beyond the sentence boundary[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain: Association for Computational Linguistics, 2017: 1171-1182.

[33] Peng N, Poon H, Quirk C, et al, Cross-sentence n-ary relation extraction with graph LSTMs [J]. Transactions of the Association for Computational Linguistics, 2017,5: 101-115.

白龙(1993—)，通信作者，博士研究生，主要研究领域为关系抽取，知识图谱。
E-mail: bailong18b@ict.ac.cn

席鹏弼(1981—)，博士研究生，工程师，主要研究领域为 P2P 网络、自然语言处理、知识图谱。
E-mail: xipengbi@ict.ac.cn

靳小龙(1976—)，博士，研究员，主要研究领域为大数据知识工程、知识计算、知识图谱。
E-mail: jinxiaolong@ict.ac.cn