

基于远程监督的多因子人物关系抽取模型

黄杨琛¹, 贾焰¹, 甘亮¹, 徐菁¹, 黄九鸣¹, 赫中翮²

(1. 国防科技大学计算机学院, 湖南 长沙 410073; 2. 湖南星汉数智科技有限公司知识图谱研发部, 湖南 长沙 410205)

摘 要: 针对远程监督的基本假设过强容易引入噪声数据的问题, 提出了一种可以对远程监督自动生成的训练数据去噪的人物实体关系抽取模型。在训练数据生成阶段, 通过多示例学习的思想和基于 TF-IDF 的关系指示词发现的方法对远程监督产生的数据进行去噪处理, 使训练数据达到人工标注质量。在模型分类器中, 提出采用词法特征和句法特征相结合的多因子特征作为关系特征向量用于分类器的学习。在大规模真实数据集上的实验结果表明, 所提模型结果优于同类型的关系抽取方法。

关键词: 关系抽取; 人物关系; 远程监督; 机器学习; 自然语言处理

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-436x.2018118

Multi-factor person entity relation extraction model based on distant supervision

HUANG Yangchen¹, JIA Yan¹, GAN Liang¹, XU Jing¹, HUANG Jiuming¹, HE Zhonghe²

1. College of Computer, National University of Defense Technology, Changsha 410073, China

2. KB R&D department, Hunan Singhand Intelligent Data Technology Co., Ltd., Changsha 410205, China

Abstract: Aiming at the problem that the basic assumption of distant supervision was too strong and easy to produce noise data, a model of the person entity relation extraction which could automatically filter the training data generated by distant supervision was proposed. For training data generation, the data produced by distant supervision would be filtered by multiple instance learning and the method of TF-IDF-based relation keyword detecting, which tried to make the training data has the manual annotation quality. Furthermore, the model combined lexical and syntactic features to extract the effective relation feature vector from two angles of words and semantics for classifier. The experiment results on large scale real-world datasets show that the proposed model outperforms other relation extraction methods which based on distant supervision.

Key words: relation extraction, person entity relation, distant supervision, machine learning, natural language processing

1 引言

在互联网产生的爆炸式增长的电子文本信息中, 大量人物实体以及他们之间的关系信息涵盖其中。面对如此多元异质的数据, 人们必须采用信息

抽取技术才能满足其从中快速获取有效信息的需求。关系抽取作为信息抽取的一项重要任务, 第一次正式提出是在 1998 年的第七届消息理解大会 (MUC, message understanding conference) 上^[1], 它是指从自然语言文本中发现和识别 2 个实体之间的

收稿日期: 2017-10-25; 修回日期: 2018-06-21

通信作者: 黄杨琛, huangyangchen13@nudt.edu.cn

基金项目: 国家重点研究发展计划基金资助项目 (No.2016QY03D0601, No.2016QY03D0603); 国家自然科学基金资助项目 (No.61502517); 湖南省重点研发计划基金资助项目 (No.2018GK2056)

Foundation Items: The National Key Research and Development Program of China (No.2016QY03D0601, No.2016QY03D0603), The National Natural Science Foundation of China (No.61502517), Key Research and Development Plan of Hunan Province (No.2018GK2056)

语义关系的过程^[2]。

实体关系抽取技术突破了传统的人工阅读、理解等方式来获得语义关系的限制,取而代之的是语义关系的自动查找和抽取^[3]。作为自然语言处理中的热门研究领域,实体关系抽取一直是信息抽取领域的重要方向。关系抽取的早期研究主要是通过人工建立语法和语义规则,然后通过模式匹配的方法来识别实体的关系^[4-6]。由于这些方法需要大量的人工处理和专业知识的后期准备,研究人员开始尝试机器学习方法。根据对标注数据的依赖程度,基于机器学习的关系抽取方法可分为有监督学习、半监督学习、远程监督学习及无监督学习的方法。有监督学习方法将关系抽取作为一个分类问题,根据训练数据设计有效的特征,然后构造各种分类模型,最后使用训练好的分类器来预测关系。在特征选择上,可以结合词汇、句法、语义等特征来训练关系分类器^[7],还可以加入语法分析树和依存关系树来形成特征向量^[8],此外,还有研究加入了关系特征词的位置信息特征来进行关系分类^[9]。另外,为了避免人工设计特征工程的缺陷,学者们开始利用神经网络结构来自动学习自然语言文本特征然后进行实体关系抽取^[10-12],这类深度学习方法也属于有监督学习方法。有监督的关系抽取系统准确率和召回率都很高,但是严重依赖于事先制定好的关系类型体系和标注数据集。尤其是深度学习的方法,由于神经网络本身的特点,需要大量的训练数据才能得到较好的分类网络模型。半监督学习方法主要采用 Bootstrapping^[13]、标签传播^[14]等方式来进行关系抽取。对于要抽取的关系,该方法首先手工设定若干种子实例,然后迭代地从数据中抽取关系对应的关系模板和更多的实例。与有监督学习方法相比,半监督学习方法可以大大减少学习过程中需要

的标注语料库的规模,但是初始种子集的选取问题以及迭代过程中噪声的干扰问题等会影响该方法的实际性能。而无监督^[15-16]的开放式关系抽取方法是假设拥有相同语义关系的实体对拥有相似的上下文信息,从而利用每个实体对应的上下文信息来代表该实体对的语义关系,并对所有实体对的语义关系进行聚类。无监督实体关系抽取不需要预先定义实体关系类型体系,具有领域无关性,这在处理海量开放领域数据时很有优势,但其聚类阈值难以事先确定,抽取结果的准确率较低,并且目前仍缺乏较客观的评价标准。

近年来,各种大型知识库(KB, knowledge base)如 Freebase^[17]、DBpedia^[18]、YAGO^[19]以及在线百科知识库已建成,这对于构造有监督机器学习方法的训练数据有极大的价值。Mintz 等^[20]于 2009 年首次在关系抽取领域提出远程监督(DS, distant supervision)的思想。远程监督方法假设如果 2 个实体在知识库中是有关系的,那么所有包含这 2 个实体的句子都将表达这种关系。基于远程监督的关系抽取自发地对齐自然语言文本和给定的知识库,然后使用对齐的结果产生弱标签训练数据来学习关系提取^[21-22]。图 1 是一个采用远程监督技术进行关系抽取的系统示例(本文出现的人名均为举例用,与实际无关)。在该系统中,首先通过远程监督技术在对齐自然语言文本和知识库时,将识别出的含有某人物实体对的句子标记为知识库中该实体对关系的弱标签数据,然后针对相关人物对的关系查询,系统通过将从句子中提取到的相关特征输入分类器中进行关系判断,最后通过分类结果中的关系概率大小来将正确的关系事实结果放入关系知识库中。这既解决了有监督方法过于依赖人工标记数据的问题,又在一定程度上避免了无监督方法准确率较低的问题。

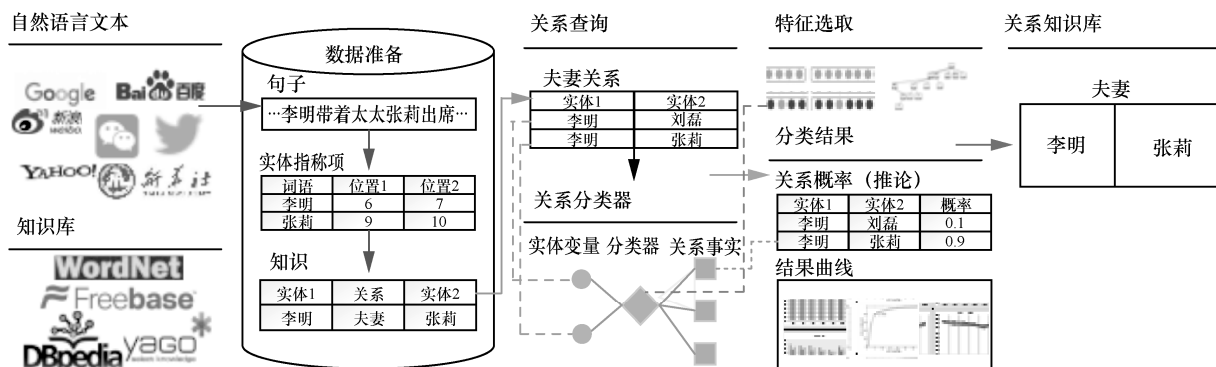


图 1 基于远程监督的关系抽取系统

但是, 远程监督的基本假设并不严谨, 在语料库中的实体对共现句中并不一定都能表达实体对在知识库中的关系。例如, “李明带领大家来到了新闻发布会现场, 张莉随后也出现在现场。”这个共现句在语义上并不能表达他们之间的“夫妻”关系事实。这种包含了实体对却不能提取到关系特征的句子属于远程监督方法产生的噪声数据, 应当将其过滤。目前, 关系抽取的研究主要集中在英文资源的处理上, 这主要是因为中文语料需要分词, 并且存在复杂的句式结构和隐含语义, 因此中文人物关系抽取更加困难。另外, 中文的知识库建设比较晚, 远程监督在中文语料的关系抽取中的研究还比较少。潘云等^[23]首次尝试利用中文互动百科在线资源构建中文的人物关系抽取系统, 采用的是标签传播算法训练模型, 得到 68% 左右的准确率, 但此方法并没有进行远程监督数据的去噪处理。黄蓓静等^[24]利用词向量及句子模式抽取、聚类及评分的方法, 对远程监督人物关系抽取过程得到的原始训练集中的噪声句子进行过滤, 达到对远程监督产生的训练集去噪的目的, 但是该方法所用的模式抽取方法可迁移性不好, 具有很强的领域特性。

基于以上研究的不足, 本文提出了一个针对中文文本的远程监督人物关系抽取模型。该模型的主要方法和贡献包括以下 3 个方面。

1) 通过远程监督技术自动产生标注数据集, 并且进一步地利用多示例学习思想以及本文提出的基于词频—逆文档频率 (TF-IDF, term frequency-inverse document frequency) 的过滤算法来获取更准确的训练数据。在不需要人工参与的情况下, 可以获得大量高质量的训练数据集。

2) 本文模型在训练过程的特征选择中, 综合考虑自然语言文本的多因子特征, 包括词法特征和句法依存分析产生的句法特征, 通过多因子特征向量各参数综合调优达到较好的分类效果。

3) 本文模型具有较好的可拓展性, 可适应新关系类型的抽取任务。即不需要人工干预标注的情况下, 任何新的人物关系的抽取任务都可以使用本文模型来快速实现。

2 主要方法与理论

2.1 远程监督

远程监督思想的首次提出是用于解决生物信

息领域问题的^[25], 经过发展, Mintz 等^[20]首次将远程监督方法应用到关系抽取的任务中。远程监督的思想主要是利用给定的知识库 D 和语料库 C , 其中, 知识库 D 中包含大量的关系三元组 (e_1, r, e_2) , 其中, $r \in R$ (R 为关系类型集合), (e_1, e_2) 是具有关系 r 的实体对。远程监督的方法将 D 与 C 对齐, 在 C 中任何包含知识库中的关系三元组中的一对实体的句子都认为能够表达 D 中三元组中实体对之间存在的关系事实。

例如, Mintz 等^[20]将 Freebase 作为结构化的知识库, 对于 Freebase 中关系实例的每个实体对, 他们将维基百科中所有包含这些实体对的句子找出来, 并且从中提取出相应的文本特征来训练关系分类器。但是这个远程监督的假设过于强烈, 会在训练数据中引入大量的噪声数据。此后, 一些研究^[26-27]通过多示例学习 (MIL, multiple instance learning) 来放松远程监督的假设。相较于有监督学习中输入的一系列被单独标注的示例, 在 MIL 中, 输入的是一系列被标注的“包”, 每个“包”都包括许多示例。当包中的所有示例都是负例时, 这个包会被标注为负包。而当包中至少含有一个正例时, 这个包会被标注为正包。当收到一系列被标注的包时, 分类器会学习出: 1) 归纳出一个类别概念以便正确标注个别示例; 2) 在归纳之外学习怎样去标注一个包。在多示例学习指导下的远程监督关系抽取模型中, 假设在实体对 (e_1, e_2) 的所有共现句中至少有一个共现句能够表示这 2 个实体间的关系。于是, 将同一个实体对的共现句放到同一个包中, 如果包中至少含有一个正例, 就标记为该关系的正例。而一个标记为负例的包中则全是不能表达该关系的句子。多示例学习的方法通过在训练过程中学习到同一个正例包中句子之间的互补性信息, 能够一定程度地缓解远程监督的噪声数据带来的错误标记的问题。

2.2 句法特征因子

句子的句法结构描述了句子中的短语结构、依存结构及其功能。依存结构分析是句法结构分析的一个重要方面, 它通过分析语言单位内成分之间的依存关系揭示其句法结构, 主张句子中核心谓词是支配其他成分的中心成分, 而其本身却不受其他任何成分的支配, 所有受支配的成分都以某种依存关系从属于支配者^[28]。对于关系抽取来说, 由于句子中的命名实体必定是作为一个短语结构出现在依

存结构中的,那么这种依存关系也必然会反映出相应实体之间的关系特征^[29]。

例如,“这是王磊与赵娟 11 岁的大女儿多多,大名叫王思南。”其中,文分词和句法分析结果如图 2 所示。从图 2 可以看出,人物实体“王磊”和关系词“女儿”存在着定中关系,关系词“女儿”与核心谓词“叫”存在着主谓关系,而核心谓词“叫”与人物实体“王思南”之间又存在着动宾关系,通过这样的句法依存分析可以发现,人物实体“王磊”与“王思南”都是依存于关系词“女儿”的。进一步,通过“王磊”与“赵娟”之间的并列关系,又可以得到人物实体“赵娟”与“王思南”之间与关系词“女儿”的依存关系。

与以上例句的分析结果类似,对依存句法分析的结果进行大量研究后发现,核心谓词对获取实体边界、承接实体关系起着关键作用。句子中命名实体分别与核心谓词、普通谓词的平均距离有明显差异。所以,在自然语言文本句子中,实体与核心谓词的距离也是实体之间的一种隐含关系特征。

3 基于远程监督的人物关系抽取模型设计

本节描述了基于远程监督的人物关系抽取模型,并分模块说明了从自然语言文本中自动抽取关系的过程。给定语料库 S , 人物关系抽取模型能够从其中含有人物对的句子 $s \in S$ 中自动抽取人物关系三元组 (e_1, r, e_2) , 如图 3 所示。该模型分为 3 个主要部分。

1) 远程监督模块:该模块通过对齐关系知识库和语料库中的自然语言文本生成弱标记数据。同时,本文模型中还采用了多示例学习的思想,将同一个人物实体对产生的所有弱标签关系实例放到同一个包中,通过同一关系的关系实例之间的信息互补性来提高后续分类的准确性。

2) 预处理模块:该模块完成 2 个功能。首先,对所有的弱标签数据进行词性标注、句法分析等自然语言处理操作,为后续的特征提取做准备。其次,则是针对远程监督产生的弱标签数据的正例包,通过基于 $TF-IDF$ 值的关系指示词发现的过滤算法进行去噪处理,以得到更加精准的正例数据用于关系分类器的训练。

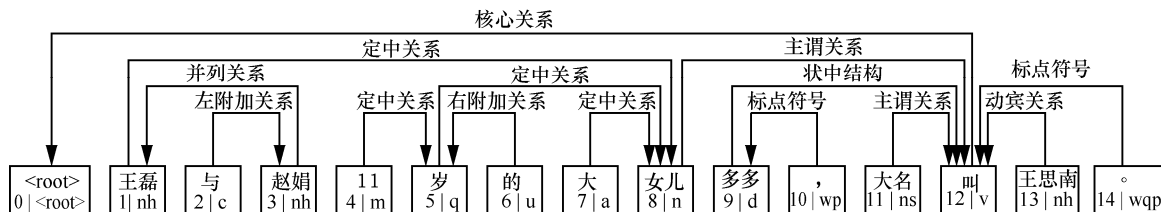


图 2 句法依存分析示例

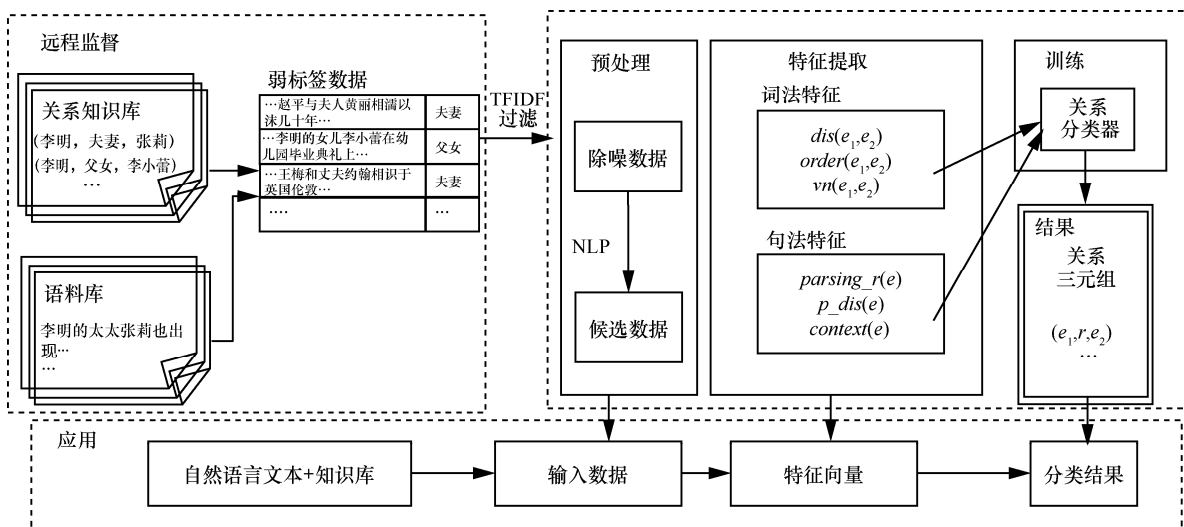


图 3 人物关系抽取模型框架

3) 特征提取模块: 该模块从语料库的自然语言文本中提取多因子特征向量, 分为词法特征和句法特征, 然后输入关系分类器中, 采用有监督的方法进行人物实体的关系分类。

接下来将从训练语料生成、实验数据去噪以及多因子特征向量3个部分来详细阐述人物关系抽取的关键过程。

3.1 训练语料生成

本文实验涉及2个部分的实验数据。关系知识库的数据是从650万个百度百科词条中直接爬取的2500万个中文三元组。语料库的自然语言文本语料为全网新闻数据, 其中, 涵盖了若干新闻站点在2012年6—7月国内、国际的新闻真实语料。

实验中使用远程监督技术构建了包含104593个句子的弱标签数据集。其中, 80%的弱标签数据(83675个句子)用作训练数据, 剩下的20%(20919个句子)用作测试数据。本文实验选择5种常见的人物关系进行实验, 分别为夫妻、父子、母子、兄弟、姐妹。表1展示了弱标签数据集的数据分布。

表1 弱标签数据集的数据分布

关系类型	编号	知识库中的三元组/个	训练集/个	测试集/个
夫妻	1	96593	22176	5543
父子	2	75304	19938	4984
母子	3	52047	16182	4046
兄弟	4	80691	10947	2735
姐妹	5	20785	14432	3611

3.2 实验数据去噪

得到远程监督产生的弱标签数据以后, 在进行关系抽取实验前, 应当对实验数据进行预处理。数据预处理的目的是通过自然语言工具对语料中的句子进行处理, 以得到词性标注、依存关系分析等结果。实验采用的是HanLP汉语言处理包来进行中文的自然语言处理, 其结果将作为特征表示以及向量生成的基础。

由于中文文本表达的多样性, 应当尽可能地挖掘出实验数据中对于关系分类有帮助的句子, 过滤掉其中的噪声。例如, 对于关系“夫妻”来说, 在自然语言文本当中含有“配偶”“夫妇”“妻子”“丈夫”等词语的句子对于训练关系“夫妻”

的抽取模型更有帮助。因此, 基于同义词词林扩展版构造了对应关系的关系词词典 $dic(r_word)$, 针对远程监督的强假设下产生的关系实例包中的弱标签正例实验数据, 进行了进一步的去噪处理。引入 $TF-IDF$ 来计算语料库中每一个句子分词后词语 w 的权值为

$$TI(w) = tf(w, s) \cdot idf(w, S) \\ = \frac{n_{ws}}{\sum_k n_{ws}} \cdot \lg \frac{|S|}{1 + |\{j: w \in s, s \in S\}|} \quad (1)$$

其中, n_{ws} 是词 w 在句子 s 中出现的次数, 而分母 $\sum_k n_{ws}$ 则是句子 s 中所有词的出现次数之和, 也是句子 s 的总词数, $|S|$ 是训练语料中的句子总数, $|\{j: w \in s, s \in S\}|$ 则是训练语料中所有包含词语 w 的句子数目。

由此可见, tf 值通过统计词语的归一化词频来反映词语的重要性。然而, 某些情况下, 一些通用的高频词语对于反映句子的主题并没有太大的作用, 反倒是一些频率较小的词更能表达句子的主题。所以, 还需要词语的 idf 值通过统计包含该词语的句子数与训练语料中句子总数的关系来体现词语的主题类别区分能力。因此, $TF-IDF$ 值通过综合 tf 值和 idf 值的大小, 可以判断当前词反映本句子主题类别的程度。基于此特性, 计算出句中词语 w 的 $TF-IDF$ 值之后, 保留其中值最大的3个词语放入集合 W 中, 并与相应的关系词词典 $dic(r_word)$ 匹配。

$$sen: W[i] \in sen = \begin{cases} \text{正例, } W[i] \in dic(r_word) \\ \text{过滤, 其他} \end{cases} \quad (2)$$

其中, sen 表示语料库中的句子, $W[i] \in sen$ 表示从句子 sen 中提取出的集合 W 中的元素。若匹配结果发现 $TF-IDF$ 值最大的3个词至少有一个在关系词典中, 则将包含该词的句子保留为该关系的正例, 否则此句子为噪声, 将其过滤。

3.3 多因子特征向量

在通过远程监督自动产生了弱标签的训练数据, 并且经过去噪处理后, 接下来从标记为正例和负例的自然语言句子中获得分类器的输入特征。自然语言文章结构一般有语素<词语<句子<段落几个层级。比词语更细粒度的语素特征, 表达语义特征不明显, 存在大量干扰杂音, 因此不选取。比句子

更高层的段落特征,目前,还没有好的方法标注段落特征,因此也未选取。本文模型中的分类器选择的多因子特征为词法特征因子和句法特征因子。词法因子是以词为对象,研究句子中词语的形成和用法,包括词法、词性以及词语的位置等信息。通过统计句子中的词法因子可以反映出句子的组织规律。而句法因子是以句子为对象,研究句子的构成和功能。句法分析将输入句子从序列形式变成树状结构,从而可以捕捉句子内部词语之间的搭配或修饰关系,得到句子的浅层语义分析结果。最后,形成了多因子特征向量 $feature=(dis, order, vn, parsing_r, p_dis, context)$ 。

3.3.1 词法因子向量

自然语言文本中,词语之间的位置、顺序以及词语的词性都能够反映出句子的重要信息。因此,从词语的维度出发,选择了以下词法因子构成特征向量。

1) 距离特征 (dis): 研究表明^[15], 距离更近的 2 个实体之间存在实体关系的可能性更大。因此,本文将 2 个人物实体在句子中的词距作为距离特征。

图 4 中的点 (5, 0.792 3) 表示 2 个实体之间的词距小于或等于 5 时的关系实例总数占总关系三元组数的 79.23%。从图 4 可以看出,开始阶段随着词距的增大,关系三元组的数目急剧增大。但是当词的数目超过 5 时,随着实体之间词距的增大,关系三元组数量的增加幅度越来越小。这也就说明了距离较近的 2 个实体更可能存在关系。

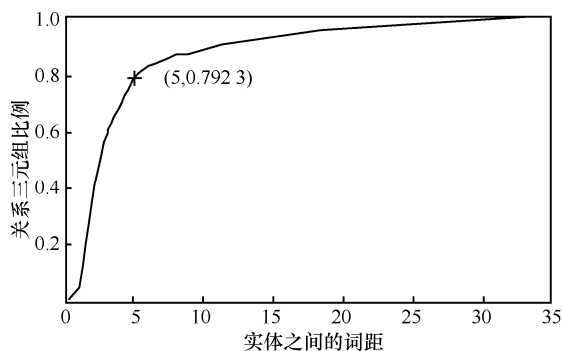


图 4 实体间的词距与三元组的数量关系

2) 相对位置特征 ($order$): 相对位置指的是 2 个人物实体 e_1 和 e_2 在句中的前后顺序。由于某些人物关系是单向性的,因此人物实体的相对位置对于实体对关系的判断十分重要。本文中,该特征的取

值分别为 1 (e_1 和 e_2 的前后顺序与知识库中的顺序一致)、0 (e_1 和 e_2 的前后顺序与知识库中的顺序不一致)。

3) 词性特征 (vn): 在人物关系抽取中,能够表示人物关系的词语通常是名词或动词。因此,对于语料库中句子特征的构建来说,名词和动词比其他词语更为重要。通过统计分词后句子中的动词和名词的数量,并进行归一化处理,从而衡量动词和名词对于人物关系判断的影响。

3.3.2 句法因子向量

从第 2 节的依存句法分析相关知识可以了解到,通过对句子进行依存分析所得结果的语块以及语块之间的依存关系,可以直接反映实体间的语义关系。因此,从句子的句法分析结果出发,选择以下句法因子构成特征向量。

1) 句法依存关系特征 ($parsing_r$): 由于人物实体将会作为短语结构出现在依存结构中,这些短语结构间的依存关系必然会反映出相应实体之间的关系特征。因此,通过获取实体对每个实体在句子中所属的句法关系依存值来反映人物实体间的关系。

2) 实体与核心谓词之间的距离特征 (p_dis): 根据命名实体识别和句法依存分析的结果,计算出人物实体与核心谓词之间的词距。

3) 实体上下文特征 ($context$): 实体的上下文可以直接反映句子的信息。从图 4 可以得知, 0.792 3 的关系实例都可以在实体之间得到。除此之外,实体对左右两侧的内容同样重要。因此,通过 $TF-IDF$ 计算所得到的权值,将人物实体对的上下文信息加入分类特征中。取 n 为实体对左右两侧的词语数目,从语料库中抽样统计得到表 2。从表 2 可以看出,随着 n 的增大,获得的信息也就更多,因此能在此范围内获得关系三元组的句子也就越多。但是 n 大于 2 以后的增幅并不明显,而 n 越大计算的开销则呈指数级增加。因此,本实验中取 $n=2$,即将句中人物实体的前 2 个词和后 2 个词的信息也加入分类特征。

表 2 关系三元组总数目与 n 的关系

n	数目/个
1	238
2	296
3	317
4	331
5	342

4 实验结果与分析

实验是为了表明本文提出的模型可以从中文自然语言文本中充分利用远程监督技术有效地提取人物实体间的关系。为此，本文基于以上语料处理及关系模型设计，最终采用准确率（precision，用 P 来表示）、召回率（recall，用 R 来表示）以及 F_1 值（ $F_1 = \frac{2PR}{P+R}$ ）来评估实验结果。

4.1 不同关系类型的关系抽取模型性能比较

为了验证人物社会关系抽取模型在不同关系上的性能，针对 5 种人物关系的抽取结果进行了对比。图 5 的纵坐标是每种关系抽取结果的 P 、 R 以及 F_1 值的数值大小，纵坐标是对应表 1 中的关系类型编号，其中，最后一组展示的是人物抽取模型的 3 项指标的平均值。从表 1 可以发现，知识库中含有关系“夫妻”的三元组是最多的，这就直接影响了实验的最终结果。实验结果显示，在所有关系中，关系“夫妻”的抽取模型的结果是最好的。同样地，关系“姐妹”在知识库中的三元组数量是最少的，其抽取结果也是所有关系中最差的。这是容易理解的，因为知识库中含有的知识越全面，能够从语料库的关系实例中学习到的该关系特征就会越多，那么对于后续关系识别的指导作用就越大。

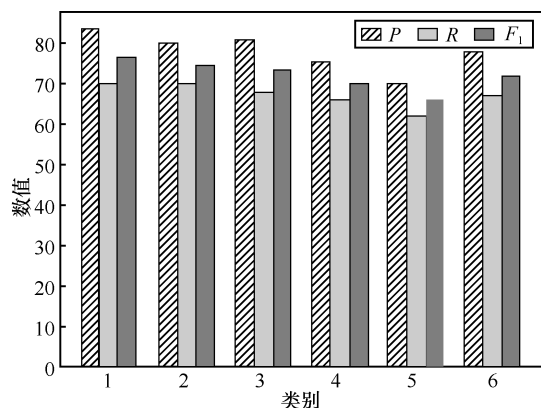


图 5 关系抽取模型的性能

另外，为了验证本文模型提出的针对关系正例包中的进一步除噪操作的有效性，本文实验还进行了有无除噪操作的关系抽取系统的性能对比。图 6 展示了除噪前后系统的性能对比，其中，横坐标同图 5，表示的是与表 1 对应的关系类型编号，最后一项表示的综合 5 种关系类型的 F_1 值

的平均值，纵坐标表示的是抽取结果的 F_1 值。从图 6 可以看到，进行了除噪操作的关系抽取系统的性能相较于未进行除噪操作的系统的性能有了进一步的提升，这证明本文提出的基于 TF-IDF 的关系指示词发现的去噪方法对于远程监督的关系抽取而言是有效的。

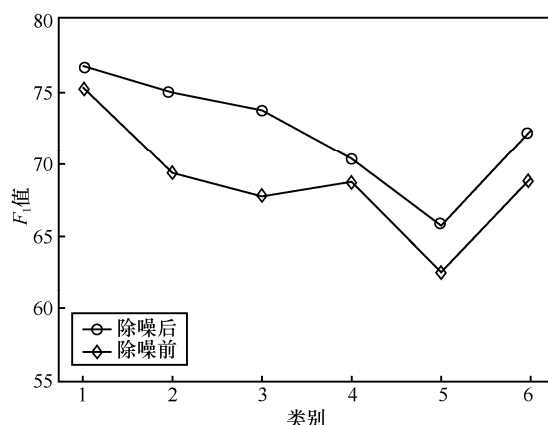


图 6 有无除噪操作的系统性能对比

4.2 各特征因子对抽取效果的影响

为了验证各特征因子对于人物社会关系抽取模型的有效性，对每一项特征因子对于关系抽取的作用都进行了因子叠加实验，共 6 组实验。第一个实验选取的特征是距离特征，第二个实验特征是在实验一的基础上加入了相对位置特征。依次类推，后一个实验是在前一个实验的基础上增加了一维特征。性能比较平均值如表 3 所示，其中， P 表示准确率， R 表示召回率， F_1 表示 F_1 值。

表 3 不同特征下模型性能比较

特征	P	R	F_1
<i>dis</i>	61.3%	57.5%	59.3%
+ <i>order</i>	65.1%	60.4%	62.2%
+ <i>vn</i>	66.4%	60.7%	63.4%
+ <i>parsing_r</i>	72.6%	64.1%	68.1%
+ <i>p_dis</i>	77.6%	65.8%	71.2%
+ <i>context</i>	79.1%	66.9%	72.5%

从表 3 结果来看，随着特征的增加，关系抽取的性能越来越好。这反映了分类器中输入的关系描述数据的信息越多，分类器的学习能力就会越好，那么在分类的时候就更容易获得

好的结果。但是, 仔细分析发现, 整个特征中, 动词和名词特征的加入对于关系抽取的性能提升作用并不是很大。经过分析认为, 可能仅仅统计句子中的动词、名词的数量并不能很好地反映某关系在句子中的存在。例如, “李明梦想着成为朗朗那样的人, 要知道朗朗毕业于美国柯蒂斯音乐学院, 他不仅是国际著名钢琴家, 还是联合国和平大使。”这句话中动词和名词的数量很多, 但是并不能反映 2 个人物实体之间的关系。另外, 还观察到, 当加入句子的依存句法分析特征以后, 关系抽取系统的性能得到了明显的提升。所以, 在人物关系抽取的任务中, 发现和理解自然语言文本中的语义才是抽取人物关系的关键。

4.3 新关系下的关系抽取模型性能

为了验证本文模型对于新关系的适应性, 在不需要其他任何额外操作的情况下, 利用以上的人物关系抽取模型进行新关系(“朋友”“同事”)的抽取实验。

新人物关系抽取结果如表 4 所示。从结果来看, 关系“朋友”的抽取结果要优于关系“同事”的抽取结果。经分析发现可能有以下原因: 1) 统计发现知识库中含关系“朋友”三元组为 56 557 组, 而关于“同事”的三元组为 48 623 组, 因此关系“朋友”能够从知识库中学到的关系特征更全面, 从而达到更好的关系抽取效果; 2) 关系“同事”的表述相对关系“朋友”的表述更加隐晦, 在很多句子中可能只是描述 2 个人物实体在同一机构中一起工作的事实, 而没有具体的关系描述词, 例如, “李明和赵西一起迈入公司的大门, 开启一天的工作。”在这句话中, 模型中的词法特征和句法特征实现的浅层的语义分析并不能得到 2 个人物实体是“同事”关系的有效特征。因此, 在关系分类器的学习过程中, 关系“同事”的无效训练数据可能更多, 因而抽取效果不太理想。

表 4 新关系的抽取性能比较

新关系	P	R	F_1
朋友	76.4%	63.1%	69.1%
同事	73.7%	61.2%	66.8%

4.4 相关研究对比

当前针对中文文本的基于远程监督的人物关

系抽取研究中, 还没有一个标准的数据集用于对比实验。但是, 从目前的研究现状来看, 中文知识库大多都是基于中文百科构建的, 语料库大多采用新闻数据构成, 这说明此情况下, 通过远程监督技术将知识库和语料库对齐所产生的训练数据并不会存在太大差异。因此, 在相似数据下, 可以将本文模型与其他相近研究进行对比。

标签传播模型(2015 年)^[23]和模式聚类模型(2017 年)^[24]都是在中文文本下, 基于百科知识库和新闻数据, 采用远程监督技术产生训练数据的人物关系抽取模型。不同的是, 标签传播模型采用的是基于人物对间相似度计算的标签传播方法来进行人物的关系抽取, 模式聚类模型则采用句子模式聚类的方法进行远程监督数据的过滤以后再进行人物关系的抽取。实验结果对比如表 5 所示。从实验结果来看, 标签传播模型的准确率明显低于后两者的准确率, 这说明对远程监督产生的弱标签训练数据进行过滤处理, 对人物关系抽取结果的准确性很有帮助。另外, 模式聚类模型采用的模式匹配方法在召回率上的结果不如本文模型, 这说明使用特征抽取的方法进行关系分类, 能在保证准确率的同时提高系统的召回率。综合对比发现, 在大规模真实数据集上的实验结果表明, 本文模型结果优于其他同类型的人物关系抽取模型。

表 5 相关研究对比

研究方法	P	R	F_1
标签传播模型	69.3%	—	—
模式聚类模型	74.2%	60.6%	64.6%
本文模型	79.1%	66.9%	72.5%

5 结束语

本文提出了一个针对中文文本自动生成训练数据的人物实体关系抽取模型。它首先利用远程监督技术产生弱标签数据集, 然后采用多示例学习的思想以及本文提出的基于 TF-IDF 的过滤算法获取更准确有效的训练数据, 最后利用多因子特征向量采用有监督的方法进行关系抽取。整个模型包含 3 个部分, 远程监督的模块、预处理模块和特征提取模块。本文将文本数据中的词法特征和句法特征综合考虑, 从文本词语结构和依存

句法上提取反映实体间关系的相关特征。在真实数据集上进行的实验表明, 基于关系指示词的过滤算法能有效提高远程监督关系抽取的准确率, 并且本文模型性能优于当前中文语料下的其他远程监督关系抽取系统, 另外, 本文模型还具有良好的新关系类型的适应性。

另一方面, 实验结果表明, 通过远程监督获得的弱标签数据的准确性在很大程度上影响着关系抽取的结果。此外, 源于自然语言处理技术的特征提取过程中可能产生的误差积累会导致关系抽取模型的性能很难提高。在未来, 将探讨如何进一步减少在远程监督过程中的错误标签的问题, 并且可以使用深度学习技术自动学习句子的相关特征。

参考文献:

- [1] CHINCHOR N, MARSH E. Muc-7 information extraction task definition[C]//The Seventh Message Understanding Conference (MUC-7), Appendices. 1998: 359-367.
- [2] CHE W, LIU T, LI S. Automatic entity relation extraction[J]. Journal of Chinese Information Processing, 2005, 19(2): 1-6.
- [3] 刘绍毓, 李弼程, 郭志刚, 等. 实体关系抽取研究综述[J]. 信息工程大学学报, 2016 (5): 541-547.
LIU S Y, LI B C, GUO Z G, et al. Review of entity relation extraction[J]. Journal of Information Engineering University, 2016(5): 541-547.
- [4] APPELT D E, HOBBS J R, BEAR J, et al. SRI International FASTUS system: MUC-6 test results and analysis[C]//The 6th Conference On Message Understanding. 1995: 237-248.
- [5] YANGARBER R, GRISHMAN R. NYU: description of the proteus/PET system as used for MUC-7 ST[C]//Seventh Message Understanding Conference (MUC-7). 1998.
- [6] AONE C, RAMOS S M. REES: a large-scale relation and event extraction system[C]//The Sixth Conference On Applied Natural Language Processing. 2000: 76-83.
- [7] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//The ACL 2004 On Interactive Poster And Demonstration Sessions. 2004: 22.
- [8] GUO D Z, JIAN S, JIE Z, et al. Exploring various knowledge in relation extraction[C]//The 43rd Annual Meeting on Association for Computational Linguistics. 2005: 427-434.
- [9] 宁海燕. 实体关系自动抽取技术的比较研究[D]. 哈尔滨: 哈尔滨工业大学, 2010.
NING H Y. Comparative study of automatic entity relation extraction[D]. Harbin: Harbin Institute of Technology, 2010.
- [10] WANG L, CAO Z, MELO D G, et al. Relation classification via multi-level attention CNNs[C]//The 54th Annual Meeting of the Association for Computational Linguistics. 2016: 1298-1307.
- [11] KUMAR S. A survey of deep learning methods for relation extraction[J]. arXiv preprint, arXiv:1705.03645, 2017.
- [12] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//The 54th Annual Meeting of the Association for Computational Linguistics. 2016: 207-212.
- [13] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning[C]//AAAI. 2010: 3.
- [14] 刘锦文, 许静, 张利萍, 等. 基于标签传播和主动学习的人物社会关系抽取[J]. 计算机工程, 2017, 34(2): 234-240.
LIU J W, XU J, ZHANG L P, et al. Personal social relation extraction based on label propagation and active learning[J]. Computer Engineering, 2017, 34(2): 234-240.
- [15] 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 B, LIU A A, LIU T. Unsupervised Chinese open entity relation extraction[J]. Journal of Computer Research and Development, 2015, 52(5): 1029-1035.
- [16] GASHTEOVSKI K, GEMULLA R, DEL C L. MinIE: minimizing facts in open information extraction[C]//The 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2630-2640.
- [17] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//The 2008 ACM SIGMOD International Conference on Management of Data. 2008: 1247-1250.
- [18] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: a nucleus for a web of open data[M]//The Semantic Web. 2007: 722-735.
- [19] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]//The 16th International Conference on World Wide Web. 2007: 697-706.
- [20] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]//The Joint Conference of the 47th Annual Meeting of the ACL and The 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009: 1003-1011.
- [21] JI G, LIU K, HE S, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions[C]//AAAI. 2017: 3060-3066.
- [22] PERSHINA M, MIN B, XU W, et al. Infusion of labeled data into distant supervision for relation extraction[C]//The 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 732-738.
- [23] 潘云, 布勒布丽汗·伊沙巴依, 杨静, 等. 利用中文在线资源的远程监督人物关系抽取[J]. 小型微型计算机系统, 2015, 4: 11.
PAN Y, BULEHULIHAN-YISHABAY, YANG J, et al. Distant supervised personal relation extraction using chinese online resource[J].

Journal of Chinese Mini-Micro Computer Systems, 2015, 4: 11.

- [24] 黄蓓静, 贺樑, 杨静. 远程监督人物关系抽取中的去噪研究[J]. 计算机应用与软件, 2017, 34(7): 11-18.
HUANG B J, HE L, YANG J. Research on noise reduction in distant supervised personal relation extraction[J]. Computer Application and Software, 2017, 34(7): 11-18.
- [25] CRAVEN M, KUMLIEN J. Constructing biological knowledge bases by extracting information from text sources[C]//ISMB. 1999: 77-86.
- [26] SURDEANU M, TIBSHIRANI J, NALLAPATI R, et al. Multi-instance multi-label learning for relation extraction[C]//The 2012 Joint Conference On Empirical Methods In Natural Language Processing And Computational Natural Language Learning. 2012: 455-465.
- [27] HOFFMANN R, ZHANG C, LING X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]//The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011: 541-550.
- [28] 胡宝顺, 王大玲, 于戈, 等. 基于句法结构特征分析及分类技术的答案提取算法[J]. 计算机学报, 2008, 31(4): 662-676.
HU B S, WANG D L, YU G, et al. An answer extraction algorithm based on syntax structure feature parsing and classification[J]. Chinese Journal of Computers, 2008, 31(4): 662-676.
- [29] 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.
GUO X Y, HE T T, HU X H, et al. Chinese named entity relation extraction based on syntactic and semantic features[J]. Journal of Chinese Information Processing, 2014, 28(6): 183-189.

[作者简介]



黄杨琛(1991-), 女, 湖南衡阳人, 国防科技大学博士生, 主要研究方向为数据挖掘、自然语言处理等。



贾焰(1960-), 女, 四川成都人, 国防科技大学教授、博士生导师, 主要研究方向为社交网络分析、信息安全等。

甘亮(1977-), 男, 江西樟树人, 国防科技大学讲师, 主要研究方向为知识工程、专家系统等。

徐菁(1989-), 女, 山东济南人, 国防科技大学博士生, 主要研究方向为知识工程、文本挖掘等。

黄九鸣(1981-), 男, 福建安溪人, 国防科技大学讲师, 主要研究方向为社交网络分析、信息安全等。

赫中翾(1991-), 男, 山西大同人, 湖南星汉数智科技有限公司工程师, 主要研究方向为信息抽取。