文章编号:1007-1423(2015)03-0018-04 DOI:10.3969/j.issn.1007-1423.2015.03.005

关系抽取研究综述

母克东,万琪

(四川大学计算机学院,成都 610065)

摘要:
信息抽取、自然语言理解、信息检索等应用需要更好地理解两个实体之间的语义关系,对关系抽取进行概况总结。将关系抽取划分为两个阶段研究:特定领域的传统关系抽取和开放领域的关系抽取。并对关系抽取从抽取算法、评估指标和未来发展趋势三个部分对关系抽取系统进行系统的分析总结。

关键词:
关系抽取;机器学习;信息抽取;开放关系抽取

0 引言

随着大数据的不断发展,海量信息以半结构或者纯原始文本的形式展现给信息使用者,如何采用自然语言处理和数据挖掘相关技术从中帮助用户获取有价值的信息,是当代计算机研究技术迫切的需求。因此,信息抽取技术应运而生,信息抽取的主要目的是从自然语言文本中抽取指定的实体(Entity)、关系(Relation)、事件(Event)等事实信息。信息抽取技术可以经过一些列处理把文本中蕴含的无规律化信息转化成结构化的信息存储到数据库中,方便用户快速获取急需的信息,而关系抽取(Relation Extraction)是信息抽取的一个重要子任务,首次于1998年在MUC[1]会议正式提出,主要任务是确定两个实体之间的语义关系。实体关系抽取技术已经被广泛应用到信息检索(information extraction)、基因疾病关系挖掘(gene-disease)、蛋白质交互作用(protein-protein)等众多应用领域。

1 关系抽取

实体间的关系可被形式化描述为关系三元组<E1,R,E2>,其中E1和E2是实体类型,R是关系描述类型。实体关系抽取的主要目的是把无结构的自然语言文本中所蕴含的实体语义关系挖掘出来,整理成三元组<E1,R,E2>存储在数据库中,供进一步分析利用或查询。当前主流关系抽取研究主要朝着2个方向进行:面向领域的传统关系抽取(Traditional Relation Extraction,TRE)和开放领域的关系抽取(Open Relation Extraction,ORE)。

1.1 特定领域的传统关系抽取

(1)基于规则的方法

基于规则的方法需要提前定义能够描述两个实体所在结构的规则,Aone等人[2]通过对语料文本特点总结,邀请知识领域专家编写文本关系描述规则从而抽取关系实例。Humphreys等人[3]首先对句子进行句法树分析,从而手工构造一系列复杂的规则识别实体之间的语义关系。此方法要求规则构建者对领域的背景和特点有深入的了解,缺点是人工参与量大大,难以移植到其他领域。

(2)基于机器学习的方法

目前基于机器学习的实体关系抽取的研究主要集中于以下三类方法:有指导方法、半指导、无指导的方法。

①有指导的关系抽取(Supervised Approaches)

有指导方法将关系抽取看作一个分类问题,即通过2个实体的一系列特征来判断该实体对是否属于提前定义好的关系类型。这类方法一般需要人工标注足够多的数据作为训练语料库,然后抽取能描述刻画关

系表达的上下文特征，利用不同的分类模型对关系实例进行学习判别，对新来的实体关系样例进行关系类型预测。其算法框架如图 1 所示。

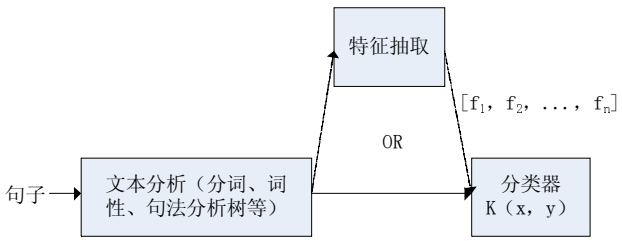

图 1　有指导方法框架

基于特征向量抽取以及基于核函数的方法是当前实体关系抽取领域最流行的基于有指导的方法。

基于特征向量抽取的方法主要从关系实例实体的上下文信息、词性、句法等信息中抽取一系列特征 $[f_1, f_2, \cdots, f_n]$ 训练一个分类器（朴素贝叶斯、支撑向量机、最大熵等），从而完成关系抽取任务。Kambhatla 等人[4]首次采用最大熵分类器对关系抽取进行建模，考虑实体上下文信息、句法分析树、依存关系在内的多种特征，结果表明实体上下文丰富的语言特征对关系表达具有丰富的价值，为后续关系抽取奠定了基础。Jiang 等人[5]对从各种信息中抽取特征进行了系统性的研究和描述，根据自然语言处理技术复杂度不同，将特征按照不同的维度划分为不同的子空间，实验结果表明这种划分在一定程度上能有效提升关系结果的准确率。董静等人[6]结合中文语料库的特点，将实体关系划分为包含实体关系抽取子任务以及非包含实体关系抽取子任务，采用不同的句法特征、词汇特征等信息，在条件随机场模型下，在 ACE2007 语料库中进行实验，取得较好的抽取效果。

基于核函数的方法是指利用核函数直接计算两个实例之间的相似度来训练关系分类模型。最核心的一步是如何设计计算两个实例（X,Y）相似度的核函数 K（X,Y）。Bunescu 等人[7]对短语句法和依存句法上的核函数进行深入的研究。Zhang M 等人[8]和 Zhou GD 等人[9]利用两个实体间最短路径封闭树（Shortest Path Enclosed Tree），考虑不同层面语义关系特征，定义了基于树的卷积核（Convolution Tree Kernel），并综合考虑谓词上下文，实验结果表明在关系抽取任务中使用卷积核函数可以得到更好的性能。

②半指导的关系抽取（Semi-supervised Approaches）

半指导的关系抽取方法是从关系种子（Seed）进行自举（Bootstrapping），在一定包含种子实例的文本语料库中抽取实体之间的关系。典型工作有 DIPRE[10]、Snowball[11]、KnowItAll[12]。该方法优点在于不需要训练语料，从而可以有效地减少对标注语料的依赖和人工参与，而且能获得很高的准确率，并且能自动扩展到大规模语料的任务中，目前广泛被使用。缺点在于，对初识种子的依赖程度很敏感，必须要具有一定的代表性和一般性。该方法目前研究重点在于如何获取可信度较高的新关系实例和抽取模板。

③无指导的关系抽取（Unsupervised Approaches）

无指导的关系抽取一种自底向上的信息抽取策略，直接从大规模的文本数据集出发，假设拥有相同关系类型的实体对，可以通过相似的上下文信息来表达刻画，可以通过聚类（Cluster）的方法来自动抽取其上下文集合来刻画实体对的语义关系。Hasegawa 等人[13]利用前面的假设信息，通过对 2 个实体之间的文本信息聚类，类簇集合来表达关系类别，结果表明聚类方法在关系抽取中具有很好的可行性。Zhang 等人[14]利用浅层句法树（shallow parsing tree）来表达关系，利用自顶向下层次聚类算法，自己定义句法树之间的相似度函数，从而获取关系抽取结果。无指导方法优点在于不依赖当前实体关系类型定义体系，从而方便算法进行跨领域的移植，缺点在于该方法产生的聚类结果很依赖语料库的质量，并且很多结果并没有实际的意义，难以定义合适的类别给类簇，另外，该方法对低频的实体对处理能力有限，往往还需要进行人工筛选，准确性和完整性没有统一的评价标准。

1.2　开放领域的关系抽取

开放领域关系抽取使用两个实体上下文中的一些词语来描述实体之间的语义关系，从而避免构建关系类型体系。主要任务是从文本中抽取关系三元组（实体 1，关系指示词，实体 2），其中关系指示词是指上下文中能够描述实体对语义关系的词或词序列。Banko 等人[15]最早提出开放式关系抽取（ORE）的概念，利用启发式规则和简单的句法特征训练分类器的 TextRunner 系统。Wu 等人[16]提出 WOE 系统，使用维基百科中信息框来标注关系抽取语料。Yao 等人[17]认为一个关系模板可

以描述不同的关系样例，提出了基于 LDA 的关系模板聚类方法构建关系类型体系。

2 关系抽取的评价体系

对于传统的关系抽取研究一般是在某个具体的领域语料定义多个关系类别，对每个子类别进行评估或者对多个类别进行评价评估。针对整个关系结果，可以通过计算对应的准确率(Precision)、召回率(Recall)和 F_1 度量值来衡量抽取结果，其对应的公式如下：

$$Precision=\frac{\sum_i r_i}{\sum_i t_i}\times 100\%$$

$$Recall=\frac{\sum_i r_i}{\sum_i a_i}\times 100\%$$

$$F_1=\frac{2\times Precision\times Recall}{Precision\times Recall}$$

其中 r_i 表示正确识别的第 i 个类别的实例数目；t_i 被识别成第 i 类的关系实例数目；a_i 实际上是第 i 类关系的实例数目。

对于开放关系抽取，一般通过考察抽取关系的准确性来评价系统性能。综合考虑算法的时间复杂度(运行时间)和空间复杂度。

3 未来关系抽取发展趋势

3.1 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两个实体之间的二元关系抽取，但不是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。

3.2 面向知识库构建的关系抽取

当前主流思想是采用远距离监督(Distant Supervision)方法，即利用已有知识库(FreeBase、维基百科等)蕴含的潜在的关系信息作为背景，并训练出一个潜在的关系分类抽取模型，在大规模未标注的语料上获取带有一定可信软关系类标的关系实例，从而补充已有知识库。

3.3 领域自适应的关系抽取

目前的研究工作主要面向特定的关系类型或者特定领域，使用特定的语料库，很难做到领域自动迁移，所以，是否可以搞一套领域自适应的关系抽取研究框架，系统可以自动发现关系类型、挖掘关系描述模式、抽取实体对。或者在已有领域标注语料库基础上，使用迁移学习(transfer learning)的方法推广到其他领域。

4 结语

综上所述，经过多年的发展，关系抽取的相关理论和方法已经越来越完善，从最开始的基于规则的匹配到后面的基于机器学习的方法，到现在流行的开放领域关系抽取。关系抽取已经变成机器学习和人工智能的重要研究方向，其关注点已从特定领域、特定类型的关系分类转变为面向 Web 大规模语料的开放实体关系自动发现。随着关系抽取技术进一步发展，将对大数据处理、QA 系统、本体自动构建、医学信息学等领域产生深远的作用。

参考文献：

[1]Automatic Content Extraction(ACE) Evaluation[EB/OL]. [2013-06-24]. http://www.itl.nist.gov/iad/mig//tests/ace/

[2]Aone C, Halverson L, Hampton T, et al. SRA: Description of the IE2 System Used for MUC-7[C]. (MUC-7),1998

[3]Kambhatla N. Combining Lexical, Syntactic, Semantic Features with Maximum Entropy Models for Extracting Relations[C] ACL 2004

[4]Humphreys K, Gaizauskas R, Azzam S, et al. University of Sheffield: Description of the LaSIE-II System as Used for MUC-7[C]. In: Proceedings of the 7th Message Understanding Conference (MUC-7),1998

[5]Jiang J, Zhai C X. A Systematic Exploration of the Feature Space for Relation Extraction[C]. NAACL-HLT'07. 2007:113~120

[6]董静, 孙乐, 冯元勇, 等. 中文实体关系抽取中的特征选择研究[J]. 中文信息学报, 2007,21(4): 80~85

[7]Bunescu R C, Mooney R J. A Shortest Path Dependency Kernel for Relation Extraction[C].ACL,2005: 724~731

[8]Zhang M, Zhang J, Su J, et al. A Composite Kernel to Extract Relations Between Entities with Both Flat and Structured Features[C]. ACL, 2006: 825~832

[9] Zhou G D, Zhang M, Ji D H, et al. Tree Kernel-based Relation Extraction with Context-Sensitive Structured Parse Tree Information

[C]. EMNLP/CoNLL-2007.2007:728~736
[10]Brin S. Extracting Patterns and Relations from the World Wide Web[C]. In: Proceedings of International Workshop on the World Wide Web and Databases. London, UK: Springer-Verlag, 1999: 172~183
[11]Agichtein E, Gravano L. Snowball: Extracting Relations from Large Plain-text Collections[C]. In: Proceedings of the 5th ACM Conference on Digital Libraries. ACM, 2000:85~94
[12]Etzioni O, Cafarella M, Downey D, et al. Unsupervised Named-entity Extraction from the Web: An Experimental Study[J]. Artificial Intelligence, 2005,165(1): 91~134
[13]Hasegawa T, Sekine S, Grishman R. Discovering Relations Among Named Entities from Large Corpora[C].ACL 2004
[14]Zhang M, Su J, Wang D, et al. Discovering Relations Between Named Entities from a Large Raw Corpus Using Tree Similarity-based Clustering[C]. IJCNLP'05.Berlin, Heidelberg: Springer-Verlag, 2005: 378~389
[15]Banko M. Open Information Extraction for the Web[D]. University of Washington,2009
[16]Wu F, Weld D S. Open information extraction using Wikipedia. ACL '10. 2010:118~127
[17]Yao L, Riedel S, McCallum A. Unsupervised Relation Discovery with Sense Disambiguation. ACL'12. 2012: 712~720

作者简介：
母克东(1989-)，**男**，**四川南充人**，**硕士研究生**，**讲师**，**研究方向为数据挖掘与自然语言处理**
万琪(1991-)，**男**，**湖北荆门人**，**硕士研究生**，**研究方向为数据挖掘与自然语言处理**
收稿日期：2014-12-09　　**修稿日期**：2014-12-29

Survey of the Research on Relation Extraction

MU Ke-dong，WAN Qi

(School of Computer Science，Sichuang University，Chengdu 610065)

Abstract：
Many applications in natural language understanding, information extraction, information retrieval require an understanding of the semantic relations between entities. Carries on the summary to the relation extraction. There are two paradigms extracting the relation-ship between two entities: the Traditional Relation Extraction and the Open Relation Extraction. Makes detailed introduction and analysis of the algorithm of relation extraction, evaluation indicators and the future of the relation extraction system.
Keywords：
Relation Extraction; Information Extraction; Machine Learning; Open Relation Extraction