典型关系抽取系统的技术方法解析*

□ 徐健 / 中国科学院国家科学图书馆 北京 100190
中国科学院研究生院 北京 100049
中山大学资讯管理系 广州 510275
□ 张智雄 / 中国科学院国家科学图书馆 北京 100190

摘要：实体关系抽取是信息抽取领域中的一项重要任务。文章在对当前关系抽取的相关文献、系统和项目进行分析研究的基础上，将基于非结构化文本的实体关系抽取技术方法归纳为：以模式构造和匹配为主线进行关系抽取、以词典驱动关系抽取、运用机器学习算法进行关系抽取、借助Ontology进行关系抽取以及多种方法有机结合进行关系抽取。从技术应用特点、核心模块的实现细节以及系统评测结果等方面深入分析了典型的关系抽取系统，它们包括REES关系抽取系统、SVM关系抽取系统、T-Rex关系抽取系统、KMI语义网络门户的混合关系抽取系统，旨在为进一步构建实体关系抽取系统提供良好借鉴。该文为2008年第9期本期话题“知识抽取”的文章之一。

关键词：知识抽取，关系抽取，关系抽取方法，典型系统，数字图书馆
DOI:10.3772/j.issn.1673-2286.2008.09.002

1 引言

信息抽取任务在细节和可靠性上有不同的选择，但一般都包括两个普遍存在并且紧密关联的子任务：实体识别和关系抽取。实体识别通过实体抽取技术抽取各个知识要素。抽取出的知识要素以离散的形式存在，只能反映出文本中包含哪些实体，例如人、机构、地点等，却不能反映出知识要素之间的关系，例如机构与人之间的雇用关系、机构与地点之间的位置关系等，而关系抽取则是要解决这一难题。

关系抽取技术在很多领域具有应用价值。例如，在自动问答系统中，关系抽取技术能够实现自动地将相关问题和答案进行关联；在检索系统中，关系抽取技术使类似于“找出某个机构所有成员的出版物”这样的语义检索功能的实现成为可能；在本体学习过程中，关系抽取技术一方面可以帮助本体库增加更多的关系实例，另一方面能够通过发现新的实体间关系来丰富本体结构；在语义网标注任务中，关系抽取能够将语义网相关知识单元进行自动关联。

关系抽取技术路线经历了从模式、词典等简单方法到机器学习、基于Ontology的关系抽取等复杂方法，从基于分词、句法等匹配的浅表分析到基于语义的深层分析的发展过程。关系抽取性能正在逐步提高，技术也在不断进步和完善。尽管关系抽取技术还未达到普遍应用的成熟度，一些典型关系抽取原型系统的发展仍然值得我们关注。从这些原型系统，我们可以看到关系抽取技术的关键问题、发展趋势以及广泛的应用前景。本文的第2部分对关系抽取的技术路线进行总结。第3部分选取具有代表性的几个关系抽取系统，从关系抽取的技术特色方面进行了分析。

2 关系抽取的几种思路

通过长期探索和不懈努力，信息抽取领域的学者们已经提出一些关系抽取技术路线，并被应用在各种实验系统当中。这些技术路线所遵循的思路基本可以归纳为：以模式构造和匹配为主线进行关系抽取、以词典驱动关系抽取、运用机器学习算法进行关系抽取、借助Ontology进行关系抽取以及多种

* 本文受国家社会科学基金项目“从数字信息资源中实现知识抽取的理论和方法研究”（05BTQ006）和国家“十一五”科技支撑计划课题“网络科技信息监测与评价”（2006BAH03B05）的资金资助。

方法有机结合进行关系抽取。

（1）以模式构造和匹配为主线进行关系抽取

在关系抽取研究领域，普遍使用模式匹配的关系抽取方法。这类抽取方法通过运用语言学知识，在进行真正的关系抽取之前，人工或半自动地构造出若干融合语词特征、词性特征或语义特征的模式集合，并存储起来。在关系抽取过程中，将经过预处理的语句片段与先前准备好的模式集合中的模式进行匹配。匹配成功的语句片段就被认为具有对应模式的关系属性。

Douglas E. Appelt等人[1]在MUC-6上提出的FASTUS抽取系统中，通过引入“宏”的概念将各种领域依赖规则以一种具有扩展性的、通用的方式表达出来。用户只需要修改相应“宏”中的参数设置，就可以快速配置好特定领域任务的关系模式规则。Roman Yangarber等人[2]在MUC-7上提出的Proteus抽取系统采用了基于样本泛化的关系抽取模式构建方法。用户通过Proteus系统提供的模式构建界面，对含有某种关系的例句进行分析，识别出所含关系的要素，并将这些要素泛化，最后经用户确认存储经泛化表达的模式。

（2）以词典驱动关系抽取

模式的构造过程通常需要较大的努力来归纳语词、词性、语义等特征以提取某类关系的模板，新增加的关系类型往往需要较长时间来进行相应模式的获取。与之相比，以词典驱动关系抽取的思路使新关系类型的增加能够轻松实现。通过向词典添加对应的动词入口，就可以实现对新增加关系类型的抽取。用户不需要具备复杂的模式语言知识就可以轻松配置抽取系统。

Chinatsu Aone等人[3]在MUC-7上提出了一个快速、灵巧的大规模事件和关系抽取系统REES（Large-Scale Relation and Event Extraction System）。该系统采用的词典驱动方法，只要对每一个事件指示词（通常是动词）设置一个词典入口，就可以实现相应关系的抽取。词典入口可以设置该动词相关参数的句法和语义限制。

以词典驱动关系抽取方法的缺点也非常明显。它通常只能识别以动词为中心词的关系，而对于名词同位语之类的关系抽取就很难实现了。另外，使用这种方法无法对系统中没有对应词汇入口的新关系进行探测。

（3）运用机器学习算法进行关系抽取

运用机器学习算法进行关系抽取的方法是目前应用比较广泛的方法。这类方法虽然具体应用算法有所不同，但其本质上都是将关系抽取任务看成一个典型的分类问题。首先使用机器学习算法在经过人工标引的语料上构造分类器，然后将这个分类器用于未经标引的语料上，实现有限关系类别的判断。目前这类方法中使用最为广泛的是SVM方法。

Zhu Zhang[4]提出的基于SVM的弱监督关系分类系统应用SVM方法进行关系抽取。该系统的核心组件有两个：底层监督学习器和bootstrapping算法。底层监督学习器是一个支持向量分类器，它使用从当前可获得的已标注数据训练而来的模型，对未标记的数据进行分类。Bootstrapping算法则负责选择最有可能被正确标记的实例，并通过使用它们来增强已标记数据的训练效果。

（4）借助Ontology进行关系抽取

知识管理过程中，利用信息抽取技术抽取的实体以及实体间的关系来构建和丰富本体，是一种行之有效的方法。另一方面，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，也不失为一种行之有效的关系抽取方法。

José Iria等人[5-6]提出了一个基于本体的关系抽取通用软件框架——可训练关系抽取框架（Trainable Relation Extraction framework，简称T-Rex）。该框架的目的是要提供语义网自动化语义标注任务需要的灵活度。T-Rex最具特色的地方是它采用了规范的基于图的数据模型。该数据模型借助本体实现等级层次的表达结构，并允许以一致的方式任意链接子图，例如共指关系链接、语法关系链接、与HTML格式相关的链接等。通过对本体的定义和扩充，可以实现使用该多层次数据模型对于语料的多种特征集表达的一致性。

（5）多种方法有机结合进行关系抽取

在关系抽取研究的初期阶段，无论是以词典驱动进行关系抽取还是以模式构造和匹配进行关系抽取，都仅以一种抽取方法的应用作为整个关系抽取过程的核心。随着对关系抽取技术方法研究的不断深入，众多学者渐渐意识到单纯的一种抽取方法在识别特征和识别模式方面不可避免地具有局限性。

为了克服单一关系抽取方法的局限性，一些系统将多种现有关系抽取方法进行有机结合，将更多的关系识别特征加入到关系抽取过程中来。本文3.4小节对这类方法中具有代表性的KMI语义网门户的混合关系抽取系统进行了分析。

3 典型的关系抽取系统解析

在关系抽取技术的发展历程中，已经有很多关系抽取系统原型被设计和评测。这些系统在关系抽取的关键技术上进行了多方位的大胆尝试，对关系抽取技术的发展起到了重要的推动作用。我们选取了具有代表性的REES关系抽取系统、SVM关系抽取系统、T-Rex关系抽取系统以及KMI语义网门户的混合关系抽取系统，旨在通过对这些系统的解析，比较各种关系抽取技术在具体系统中灵活的应用方式。

3.1 REES关系抽取系统

Chinatsu Aone等人在MUC-7上提出了一个快速、灵活的大规模事件和关系抽取系统REES（Large-Scale Relation and Event Extraction System）。该系统采用的基于词典驱动的关系抽取方法旨在抽取尽可能多类型的关系和事件，同时人工介入的成本最小，准确率高。

在REES系统中，输入语料经过名称标识和名词短语标识阶段的处理，形成基于XML的输出。接着关系识别模块应用词典驱动模型，通过基于句法的一般模式来识别关系和事件。REES由3个主要组件构成：一个tagging组件，一个co-reference resolution模块以及一个模板生成模块，这三个模块依次相连，构成系统的主要框架。

REES提出了一种新颖的词典驱动方法来进行关系抽取。该方法需要对于每一个事件指示词设置一个词典入口，而这些词通常是动词。词典入口具体描述了该动词参数的句法和语义限制。例如，下面的词典入口对应动词“attack”。这个表达式指示出动词“attack”属于CONFLICT本体和ATTACK_TARGET类型。动词“attack”的第一个参数（ARG1_SEM）语义上是一个组织，地点，人物或物品，它在句法上是一个主语（ARG1_SYN）。第二个参数（ARG2_SEM）语义上是一个组织，地点、人物或物品，句法上是一个直接宾语。第三个参数（ARG3_SEM）语义上是一个武器，句法上是一个通过“with”引入的前置短语。

```
ATTACK { { {CATEGORY VERB}
{ONTOLOGY CONFLICT}
{TYPE ATTACK_TARGET}
{ARG1_SEM {ORGANIZATION LOCATION
PERSON ARTIFACT} }
{ARG1_SYN {SUBJECT} }
{ARG2_SEM {ORGANIZATION LOCATION
PERSON ARTIFACT} }
{ARG2_SYN {DO} }
{ARG3_SEM{WEAPON} }
{ARG3_SYN {WITH} } } }
```

通过类似这样的词汇入口支持，REES能够抽取出一般的关系、事件及其相关参数。

当前REES通过模块化的、可配置的、可升级的模式，能够处理100种关系和事件。REES系统使用从12个新闻源获取的文本进行了系统性能的评测，其中训练集为200个文本，测试集为208个文本。每个集合中对于每一种关系和事件包含至少3个样例。这些关系包括了MUC定义的关系和事件。对于关系而言REES系统的召回率达到了74%，准确率达到了74%，F测度达到了73.74%。

3.2 SVM关系抽取系统

密歇根大学的Gumwon Hong[7]在Zhu Zhang等人建立的系统基础上提出了一个基于SVM的关系抽取系统。

SVM是从统计学习理论发展而来的监督学习技术，它是由V. N. Vapnik[8]在COLT-92（Computational Learning Theory-92）上首次提出，从此迅速发展起来，目前已经在许多智能信息获取与处理领域都取得了成功的应用。运用该算法进行关系抽取的思路是：通过某种事先选择的非线性映射（核函数）将输入向量映射到一个高维特征空间，在这个空间中寻找最优分类超平面，使得它能够尽可能准确地将两类数据点分开，同时使分开的两类数据点距离分类面最远。

将SVM应用到关系抽取任务时，作为SVM输入的特征集的选取对于关系抽取的结果至关重

要。Gumwon Hong提出的SVM关系抽取系统定义的特征集包括：分词（Words）、词性标注（Part of speech）、实体类型（Entity type）、实体提及类型（Entity mention type）、块标签（Chunk tag）、语法功能标签（Grammatical function tag）、IOB链（IOB chain）、主要词路径（Head word Path）、距离（Distance）以及顺序（Order）。

该系统将关系抽取分为两个阶段：首先仅探测有无关系，然后对探测到的关系进行分类来抽取特定关系类型。之所以选择这样的两步骤抽取，是因为可以通过探测关系阶段来增加整体关系抽取任务的召回率。句子中每两个实体的组合就构成一个候选关系，因此在训练数据中存在相当多的否定关系实例。如果执行单一步骤的N+1分类，由于样例中各类实例数量不对称，特别是否定关系类的实例数量远多于其它关系类别的实例数量，分类器很有可能将测试实例识别为否定关系类，这将会降低系统整体性能。通过在多类分类阶段之前加入一个二类分类阶段，并使用候选关系中实体之间的距离阈值进行过滤，就能够去除大量不相关的候选关系，从而提高第二阶段N类分类的性能。

该系统使用由NIST（National Institute for Standards and Technology）提供的ACE2语料进行关系抽取评测。ACE语料包含了519篇文本文档，这些语料选自广播新闻节目、新闻报纸以及新闻报道。实验结果显示，该系统的准确率达到68.8%，召回率达到51.4%，F指数达到58.8%。

3.3 T-Rex关系抽取系统

José Iria等人提出了一个基于本体的关系抽取通用软件框架——可训练关系抽取框架（Trainable Relation Extraction framework，简称T-Rex）。由于T-Rex采用了参数化的插件结构，因此可以对多种基于不同抽取算法的插件进行集成和测试。

在许多IE系统中数据表达和算法是紧密结合的，而T-Rex采用了规范的基于图的数据模型，它被所有在框架中采用的算法所使用。T-Rex的数据模型允许以一致的形式表达子图之间的任意链接，例如共指链接、语法链接、相关HTML格式和用户提供的关系标注的链接。另外，通过提供潜在的能够在数据模型中获取和重用的特征，T-Rex促进了基于新的算法的快速原型。图1显示了T-Rex数据模型的一个例子。

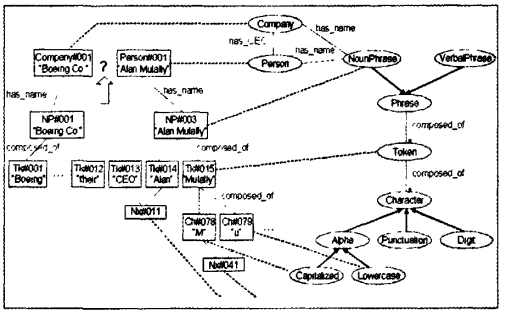

图1 T-Rex数据模型的例子[5]

图1的例子对应句子“Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results”。在上图表示的简单数据模型中，被预处理过的输入文本表示在左边，相关本体部分表示在右边。只有少量图的节点和边被显示出来。箭头边表示一般的“is_a”关系，虚线边表示“instance_of”关系。从该图可以看出，如果通过NLP技术已经判断出“Boeing Co.”属于“Company”类，而“Alan Mulelly”属于“Person”类，那么通过本体中“Company”类和“Person”类之间已存在的“has_CEO”关联可以推测在实例“Boeing Co.”和“Alan Mulelly”之间也可能存在“has_CEO”关系。

T-Rex数据模型的表示是等级化的，这意味着它能够将语料模型化到字符级、语词级、短语级、语句级和文档级等多个层次。另一方面，表达的一致性意味着它能够提供各种将语料特征化的方式。通过对本体的定义和扩充，可以实现使用该多层次数据模型对于语料的多种特征集表达的一致性。

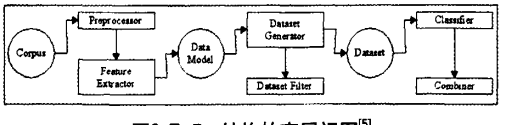

图2 T-Rex结构的高层视图[5]

T-Rex模块结构如图2所示。对于每一个组件类型，例如Processor、Classifier、Combiner，有若干个相应的插件被实现。因此，用户仅仅通过参

数定制就可以使T-Rex支持不同的信息抽取场景。T-Rex框架大致可分为两个子系统：处理子系统（由processors和feature extractors构成）和分类子系统（由classifiers，feature selection算法和predictions combiners构成）。这两个子系统的边界通过数据模型定义。处理子系统可以看成是框架的NLP依赖部分，在这里语料通过一个或多个自然语言处理工具被分析，分析输出的表达被填充到数据模型中。而分类子系统则被看成是框架的机器学习依赖部分，在这里一个或多个分类器运行在数据模型特征产生的数据集上。

3.4 KMI语义网门户的混合关系抽取系统

Lucia Specia和Enrico Motta[9]为KMI语义网门户（KMI Semantic Web Portal）开发了一个抽取语义关系的应用系统。该系统用来识别输入文本中实体间的语义关系，这包括已经存在于知识库的实体间关系、通过领域本体推理而得到的新关系和全新的关系。

该系统通过管道（pipeline）方式引入了解析器（parser）、词性标注器（part-of-speech tagger）、命名实体识别系统（named entity recognition system）、基于模式的分类器（pattern-based classification）以及词义辨析模块（word sense disambiguation models），并使用了领域本体、知识库以及词语数据库等资源。

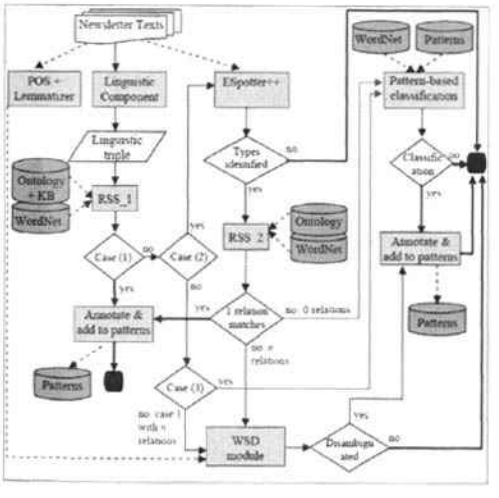

图3 基于混合关系抽取方法的系统框架[9]

图3显示了基于混合关系抽取方法的系统框架。该系统核心策略是将一个语言学三元组和他们对应的语义组件进行匹配。这不仅包括匹配关系，还包括匹配这些关系相关联的项。词项和概念的匹配通过一个领域本体和一个命名实体识别系统引导。关系识别依赖于在领域本体和词库中的知识，以及基于模式的分类和词义辨析模块。

该系统进行关系抽取的第一步是处理自然语言文本，识别语言学三元组（具有句法关系的三个元素的组，它能够指示潜在的相关语义关系）。在图3的结构中，这一步通过Aqualog设计的语言学组件模块和Gate来实现。当语言学三元组被识别出来，下一步就是检验在那个三元组中的动词表达是否传递了一个属于本体的实例之间的相对语义关系。这个阶段通过图3中的一系列模块表示。首先该系统试图通过使用一个经修改的Aqualog关系相似服务（Aqualog’s Relation Similarity Service，RSS）将语言学三元组和本体三元组进行匹配。RSS试图通过查找领域本体结构和存储在KB中的信息使语言学三元组具有语义。为了实现语言学三元组和本体三元组之间的匹配，除了精确查找之外，RSS还借助WordNet的同义词关系和先前的三元组匹配词典来实现部分匹配。

上述过程对于识别已经在领域本体中存在的关系是有效的。除此之外，该框架还可以通过模式匹配策略来识别新的关系。模式匹配策略是语义关系抽取的一个有效方式，但需要预先定义可能的关系。为了突破这个限制，该框架使用基于模式的分类模块（Pattern-based classification），它能够基于少量的初始模式来识别相似模式。该模块的原理是：将一小组相关SVO（Subject-Verb-Object）模式作为样例，调用种子模式，并使用基于WordNet的语义相似性测度来比较要分类的模式和相关的模式，并为两个模式之间的相似性打分。得分高于某个阈值，则可判断两种模式是相似的。在这种情况下，对于输入三元组的一个新的标注被产生，并被加入模式集。

该系统将新闻简报文本电子版作为输入。领域本体使用KMI-basic-portal-ontology，该本体的概念实例存储在KMI-basic-portal-kb。另外还用到WordNet词库和一个关系模式仓储。相关文献没有

报道该系统具体的关系抽取评测结果。

4 结语

本文中分析的四个关系抽取系统各自都具有一定的代表性。REES关系抽取系统对大量的关系和事件及其相关参数进行抽取，采用基于词典驱动的关系抽取方式减少了构建关系抽取模式时的人工介入成本；SVM关系抽取系统将各种词性、句法、语义特征作为SVM机器学习算法的输入，进行关系抽取；T-Rex框架通过采用参数化的插件结构和基于图的数据模型，实现了基于不同抽取算法的插件集成和测试以及各种特征的重用；KMI语义网门户关系抽取系统将多种关系抽取方法进行有机组合，实现已有关系和新关系的抽取。

从上述系统的分析我们可以看到，多种技术方法已被应用到关系抽取原型系统上，并取得了一定成果。虽然现阶段各种关系抽取系统的性能还有待提高（例如2007年ACE评测实体抽取的分数达到了82.9，而关系抽取分数最高只有33.4[10]），但其广阔的应用前景必然会成为关系抽取技术发展的强大动力，关系抽取技术的发展也将会受到越来越多的关注。

参考文献

[1] APPELT D E, HOBBS J R. SRI International FASTUS System: MUC-6 Test Results and Analysis [C/OL]//Message Understanding Conference. Proceedings of the 6th Message Understanding Conference(MUC-6) . Columbia, Maryland, 1995. [2008-07-01]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.52.246.

[2] YANGARBER R, GRISHMAN R. NYU: Description of the Proteus/PET System as Used for MUC-7 ST [C/OL]//Message Understanding Conference. Proceedings of the 6th Message Understanding Conference (MUC-7). Morgan Kaufman, 1998. [2008-07-01]. http://citeseerx.ist.psu.edu/viewdoc/summary?cid=268488.

[3] AONE C, RAMOS-SANTACRUZ M. Rees: A large-scale relation and event extraction system [C/OL]//Proceedings of the 6th Applied Natural Language Processing Conference, New York, 2000. [2008-07-01]. http://acl.ldc.upenn.edu/A/A00/A00-1011.pdf.

[4] ZHANG Zhu. Weakly-supervised relation classification for information extraction[C/OL]//Proceedings of the Thirteenth ACM conference on Information and knowledge management. Washington D.C, 2004. [2008-07-01]. http://portal.acm.org/citation.cfm?id=1031279.

[5] IRIA J. T-Rex: A Flexible Relation Extraction Framwork[C/OL]//Proceeding of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics(CLUK' 05). Manchester, January 2005. [2008-07-01]. http://eprints.aktors.org/396/01/cluk05.pdf.

[6] IRIA J, CIRAVEGNA F. Relation Extraction for Mining the Semantic Web[C/OL]//Proceedings Machine Learning for the Semantic Web Dagstuhl Seminar 05071. Dagstuhl, DE, 2005. [2008-07-01]. http://eprints.pascal-network.org/archive/00001957/01/dagstuhl.pdf.

[7] HONG GW. Relation Extraction Using Support Vector Machine[C/OL]//Second International Joint Conference, Jeju Island, Korea, 2005. [2008-07-01]. http://www.springerlink.com/content/bd654489albb1g6h/fulltext.pdf.

[8] VAPNIK V N. The nature of statistical learning theory[M/OL]. New York: Springer-Verlag New York, Inc., 1995:138-139[2008-07-01]. http://books.google.com/books?id=sna9BaxVbj8C&pg=PA243&hl=zh-CN&sig=ACfU3U2f-vGclLc0YGyal2hLHB8WwQeUyw&vq=%22the+convolution+of+two+functions+is+equal+to+the+product+of+the+Fourier+transforms+of+these+two%22&source=gbs_quotes_s&cad=2.

[9] SPECIA L, MOTTA E. A hybrid approach for extracting semantic relations from texts[C/OL]//Proceedings of the 2nd Workshop on Ontology Learning and Population. 2006. [2008-07-01]. http://www.dcs.shef.ac.uk/~lucia/publications/SpeciaMotta_OLP2-2006.pdf.

[10] ACE07. NIST 2007 Automatic Content Extraction Evaluation Official Results [EB/OL]. (2007-04-02)[2007-07-01]. http://www.nist.gov/speech/tests/ace/2007/doc/ace07_eval_official_results_20070402.html.

作者简介

徐健（1977-），男，中国科学院国家科学图书馆在读博士研究生，中山大学资讯管理系讲师，发文8篇。通讯地址：北京市海淀区中关村北四环西路33号，中国科学院国家科学图书馆 100190

张智雄（1971-），男，中国科学院国家科学图书馆研究馆员、博士生导师，发文60余篇。通讯地址：同上

The Technical Method Analysis of Typical Relation Extraction System

Xu Jian / National Science Library, Chinese Academy of Sciences, Beijing, 100190; Graduate School of the Chinese Academy of Sciences, Beijing, 100049; Department of Information Management, Sun Yat-Sen Univ., Guangzhou, 510275

Zhang Zhixiong / National Science Library, Chinese Academy of Sciences, Beijing, 100190

Abstract: Entity relation extraction is a very important task in information extraction domain. Based on the analysis of recent related literatures, systems and projects, this paper concludes the entity relation extraction methods as follows: using templates to extract relation, lexicon driven to extract relation, using machine learning algorithm to extract relation, using ontology to extract relation and using hybrid approach to extract relation. Several typical relation extraction system are analyzed on aspects of technical features, core modules implementation and evaluation results. They are REES relation extraction system, SVM relation extraction system, T-Rex relation extraction system and hybrid relation extraction system of KMI semantic web portal. The analysis of these systems can help to build more efficient entity relation extraction system in further step.

Keywords: Knowledge extraction, Entity relation extraction, Relation extraction methods, Typical system, Digital library

（收稿日期：2008-07-13；责任编辑：贾延霞）