基于 NLP 的知识抽取系统架构研究*

化柏林

(中国科学技术信息研究所 北京 100038)

【摘要】 在参考自然语言处理平台及知识抽取系统的系统结构的基础上,提出一个基于 NLP 的知识抽取系统的详细设计方案。自然语言处理过程包括分词、词性标注、句法分析、语义分析等 8 大模块;知识抽取过程包括论文类型分析、篇章结构分析、知识抽取、知识表示 4 大模块。通过对基于 NLP 的知识抽取系统架构的研究,明确自然语言处理与知识抽取的关系,分析出知识抽取的系统流程及关键技术。

【关键词】 自然语言处理 知识抽取 文献分析 内容分析 系统架构 关键技术

【分类号】 G35 TP391

Architecture of Knowledge Extraction Based on NLP

Hua Bolin

(*Institute of Scientific and Technical Information of China*, *Beijing 100038*, *China*)

【**Abstract**】 Based on the studies of system architecture of NLP platform and knowledge extraction system, the author brings forth a detailed resolution on how to design a knowledge extraction system based on NLP. NLP technique includes eight modules, such as segmentation, part - of speech tag, syntactic analysis and semantic analysis. Knowledge extraction includes four modules, such as documents type analysis, discourse analysis, knowledge extraction and knowledge representation. Research on system architecture of knowledge extraction based on NLP is beneficial to not only find relations between NLP and knowledge extraction, but also analyze system flow and critical technology of knowledge extraction.

【**Keywords**】 Natural Language Processing(NLP) Knowledge extraction Document analysis Content analysis System architecture Critical technology

1 引 言

知识抽取是指把蕴含于文本文献中的知识经识别、理解、筛选、格式化,把文献的每个知识点抽取出来,以格式化的形式存入知识库。知识抽取是知识获取的有效途径之一,也是知识工程的关键技术之一。本研究旨在探讨知识获取中的知识抽取的模式与方法,研究基于自然语言处理(Natural Language Processing,NLP)的知识抽取模式与方法,尝试运用 NLP 技术,在经过分词、词性标注、句法分析、语义分析等过程后从科学文献的语段中抽取知识,然后把用自然语言描述的句子通过知识表示转换成计算机可理解的形式,并存入知识库中。

为实现上述目标,需要解决以下几个关键技术问题:面向知识抽取的自然语言处理平台;学术论文正文内容元数据方案以及论文写作结构、写作手法的分析与总结;文献中蕴含知识与知识库中的知识单元匹配判定;根据知识的类型自动确定知识抽取模式;为平台的上层应用系统设计良好的可扩展的系统接口。

2 国内外知识抽取相关系统的架构设计

知识抽取的来源主要有结构化文本、半结构化文本、非结构化文本。结构化文本包括词典、主题词表、本体、大百科全书等;半结构化文本主要是指标记文本,包括 HTML 标记文本与 XML 标记文本;非结构化文本主要指图书、论文等传统文献。这些文本按前期的标注程度不同又可分为原始文本、粗标注文本和全标注文本。知识抽取的理论模型支撑有粗糙集、遗传算法、神经网络、潜在语义标引等。知识抽取的过程或多或少地用到 NLP 处理技术,用于支撑这些分析的资源包括词典、规则库、常识知识库、领域知识库。下面简单介绍两个较为典型的

收稿日期:2007 - 07 - 04

收修改稿日期:2007 - 07 - 18

* 本文系中国科学技术信息研究所预研基金项目"知识抽取系统架构与关键技术研究"(项目编号:YY2006018)的研究成果之一。

知识抽取系统的体系结构。

2.1 基于知识库的半自动知识抽取

荷兰特温特大学的 Plinius 是一个人工参与的辅助知识获取系统。该系统从短的自然语言文本(如文献的标题、摘要等字段)中抽取知识。运用词典把自然语言实体翻译成概念实体,用背景知识库建立新的知识库[1]。首先对文献进行预处理,然后对处理过的文献进行语言处理,最终的目标是获得集成的知识库。语言的处理是核心,语言处理时需要很多的资源,包括语法、词典、本体知识库、背景知识库等,语言处理过程中有专家的介入。其系统结构如图 1 所示:

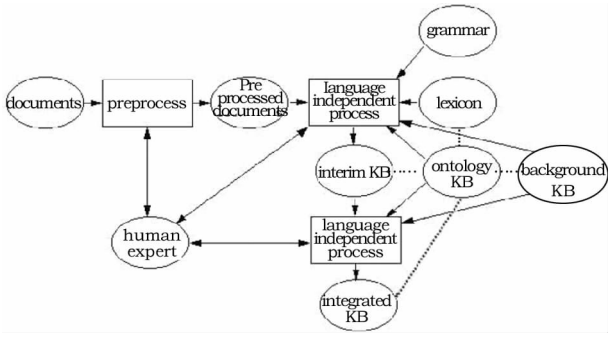

图 1 Plinius 知识半自动知识抽取系统结构图[1]

2.2 基于本体的知识抽取系统架构

中国台湾国立清华大学的人工智慧研究室运用智能代理技术从大规模生物文献中直接抽取知识。该系统包括 3 大处理过程:

(1)从 PubMed 中选取文献,用叙词表(包括 WordNet、医学主题词表 MeSH、基因本体 GO)标注文献中的术语,然后用模式匹配对术语进行行义项标注。

(2)用模式规则把词合成短语,用 Mimipar 对句子进行句法分析得到依存树。

(3)把句型语法映射成领域本体的语义结构,然后从文献中抽取知识[2]。其系统结构如图 2 所示。

在所有的知识抽取系统中,都特别强调资源的重要性,包括词典、规则、本体等。而处理过程中,自然语言的处理过程必不可少,包括自动分词、句法分析、语义标注等,有的系统进行深层语言处理,有的进行浅层语言处理。自然语言处理技术的运用与本体资源的支撑在大多数知识抽取系统中都发挥着重要作用。

3 知识抽取系统的总体架构设计

在参考其他自然语言处理平台与知识抽取系统架构的基础上,笔者设计出一套基于 NLP 的知识抽取的系统架构。在系统架构下,详细设计其处理流程,选择合适算法并逐个模块进行小规模实验。

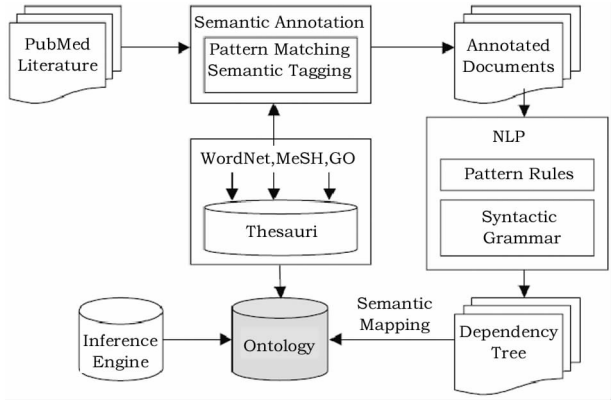

图 2 基于本体的医学知识抽取系统结构图[2]

3.1 知识抽取系统总体设计思路

本研究严格按照语言处理的层面和语言构成的单位,由小到大、由浅入深进行分析。从词、句到段、章逐步加大分析粒度,从语形分析、语法分析到语义分析、语用分析逐步加深分析层面,最终实现对文献的内容理解并抽取知识。不回避每一个处理步骤,每一个处理步骤选取最优化的算法,大多数步骤都会进行适当的消歧。基于自然语言处理的知识抽取系统构架如图 3 所示:

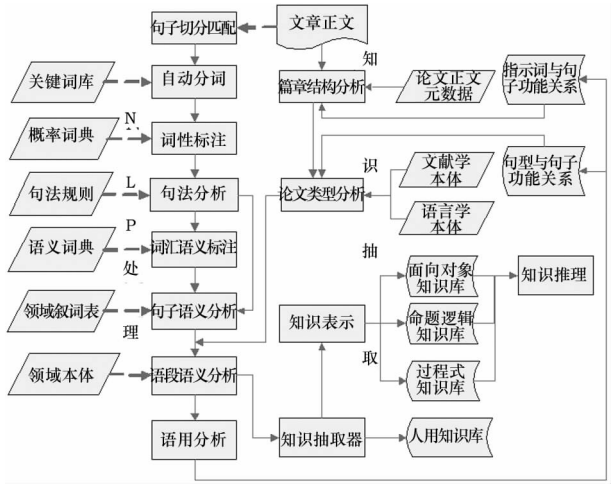

图 3 基于 NLP 的知识抽取系统架构

3.2 知识抽取系统的系统架构

知识抽取系统分为两大部分:一部分是自然语言处理,另一部分是知识抽取。自然语言处理主要从语言的角度对相关内容进行分析,包括句子切分、自动分词、词性标注、词义标注、句法分析、句义分析、语段分析及语用分析 8 大模块,其中前 4 个模块是基础,句法分析与句义分析是核心,语段分析与语用分析是扩展。在这 8 个模块的运行过程中,需要关键词库、概率词典、语义词典、句法规则、领域叙词表与领域本体 6 类资源的支撑。知识

抽取过程主要在自然语言处理的基础上,从文献表述的角度进行识别与抽取,主要包括论文类型分析、篇章结构分析、知识抽取、知识表示 4 大模块,其中前两个模块是基础,知识抽取是核心,知识表示是扩展。在这 4 个模块的运行过程中,需要论文正文元数据、指示词与句子功能关系、句型与句子功能关系、文献学本体以及语言学本体 5 类资源的支撑。

3.3 知识抽取系统的软件结构设计

基于 NLP 的知识抽取系统,设计模式采用 MVC,面向对象程序设计采用 Java 进行系统实现;面向对象数据库采用 ObjectStore,关系数据库采用 Oracle;自动分词采用最大向量匹配算法,词性标注采用最大概率算法,语法分析采用 LR 分析算法,语义分析采用谓词逻辑;系统接口采用 XML。

4 知识抽取中的 NLP 技术

4.1 切分句子与自动分词模块

利用标点符号与段落标记把文章的正文切分成句子,然后到句子库进行匹配分析,滤掉学术抄袭与科学引用的句子,得到可能含有新知识的句子。利用停用词把句子粗分成若干个待分析串,利用本领域的关键词词库使用嵌套的逆向最大向量进行切分[3]。对于有切分歧义的切分结果使用概率词典用最大概率法进行消歧,选出最优的切分结果。

4.2 词性标注模块

经过分词以后,就可以进行词性标注。只有词性标注以后,才可以进行后续的词汇语义标注和句法分析。使用隐马尔科夫模型和规则相结合的方法进行词性标注[4]。学术论文中,实词主要由名词和动词构成,形容词和副词使用甚少。连词、否定词、程度词在文中的作用较大,在进行语义分析时,其作用甚至超过名词与动词。对于具有唯一词性的词优先标记,然后利用规则对相邻词进行标记,剩余的词再利用隐马尔科夫模型进行标注。分析结果得到的规则可以添加到词性标注规则库,以提高后续准确率。

4.3 词汇语义标注模块

词性标注以后,就可以进行词义标注。有些词具有很多义项,因此需要根据语境、语义词典等信息标记出每个词在句子中的语义。词义标注采用互信息[5]与义类词典[6]相结合的方法。学术论文的写作有着一定的规律可循,语言的搭配也存在着某种模式,因此可以采用互信息进行词义标注。由于学术论文都有着明确的文献主题与分类,即使是跨学科的研究一般也不会超过三四个分类号,因此使用义类词典法进行语义标注的消歧是个不错的选择。

4.4 句法分析模块

分词、词性标注、词义标注都属于词汇的层面。在以句子为操作重点的知识抽取中,句法分析必不可少。句法理论模型有很多,句法分析算法也有很多。本系统以乔姆斯基的转换生成理论为基础,用词汇功能语法(Lexical Functional Grammar,LFG)作为语法模型范例,从成分结构与功能结构两个层面进行分析[7]。两种结构分别以树图与框图形式呈现,在线性表达与图形表达之间建立映射关系。在转向语义分析时,如果表达强调主题概念,则使用 HPSG 模型(中心语驱动的短语结构语法,Head - driven Phrase Structure Grammar)[8];如果表达以动词操作为主,则使用 LFG 语法的功能结构,因为在功能结构里已存在类似于谓词逻辑的表示方式,向逻辑语义转换较为容易。两种语法之间根据需要互相转换,不同结构之间进行转换与映射。句法分析的算法采用 LR(Left - to - Right)分析。

这一系列过程具有严格的承接关系,不经过分词就无法进行词性标记,不经过词性标记就不可能进行句法分析。但反过来亦有影响,进行词性标记时可能对分词结果进行回溯分析,句法分析时也可能对词性标注进行回溯消歧。因此,这是一系列的技术,只有充分运用这些语言处理的关键技术,才有可能理解自然语言文本,从而进行知识抽取。

5 基于 NLP 的知识抽取

不同的知识类型应该采用不同的知识表示方式,温有奎教授总结了 10 种知识类型[9]。对于静态概念及概念之间关系用面向对象形式来表示,对命题型问题用一阶逻辑来表示,对于系统流程和实验流程等过程性知识用脚本表示法。

5.1 文献内容解析与模式判别模块

科学文献有综述型、实验型、过程型等类型,不同的类型有着不同的写作结构(篇章结构),不同类型的文章有着不同的写作手法(句型与语用),每个问题都有着几种常用的句型。现在的元数据只是描述文献辅助信息,如作者、题名、出处等,并没有深入到内容进行描述。本研究旨在对学术论文的写作结构、写作手法、句型结构等进行规律性的总结,建立定义、分类、发展历史、关键技术、应用前景、发展趋势等内容元数据。通过小标题以及线索词等信息判断识别出这些内容。

5.2 语段分析与语用分析模块

语段分析主要根据关联词和主题概念确定段内句子

间的关系。“但是、例如、而且”等关联词在分析句间关系中有着重要作用。另外,根据实词也能确定句间关系,例如,不同句子的中心语概念之间存在上下位关系,那么这两个句子之间很有可能是递进的关系,从研究的角度来讲就是细化、深入。然后进行语用分析,确定句型与句子的功能关系、指示词与句子的功能关系,并在此基础上实现对论文类型分析和篇章结构分析。

5.3 知识抽取模式选择模块

在经过了以上的系列分析以后,能够确定知识在文献中的表述方式及表述类型。根据不同类型的论文内容元数据,选择不同的知识抽取模式。如对定义的抽取只是从句子的线性表达中抽取,对分类的抽取要借助于主题词表、概念体系结构等,强调分析句子之间的关系,分清概念之间的上下位关系等。抽取的知识既可以按主题进行分类存储,亦可按知识的结构及表现形式分类存储。

5.4 知识抽取与映射模块

知识抽取与映射模块是整个系统的核心。其关键是从语段中抽取知识,这种知识存到数据库里供人使用。然后把用自然语言描述的句子通过知识表示转换成计算机可理解的形式,如面向对象知识库、产生式规则知识库、过程式知识库[10],不同的知识采取不同的知识表示方式。基本概念采取面向对象知识库,概念之间的关系采用语义网表示,论点、结论等采用命题逻辑表示方式,处理流程、实验过程等采用产生式表示方式。有了这些知识,就可以运用知识推理进行知识创新,这种创新包括关联规则挖掘、模糊逻辑推理等方法。

知识抽取平台一旦建立起来,就可以进行知识抽取。实现从文献中自动抽取知识,就可以构建知识库,从而实施知识工程了。从文献中抽取知识构建知识库是知识抽取平台的最主要应用。通过抽取建立的知识库也可以应用于自动问答、智能检索、机器翻译、专家系统等。

6 结 语

总体上,本实验采取小规模实验、经过分析获取资源、然后追加到资源库,改进分析算法,进而进行大规模实验。总体上采取“边分析(文本)、边抽取(知识)、边改进(算法)、边建设(资源)”的技术路线。

支撑知识抽取的自然语言处理要经过一系列的复杂过程,这一系列的过程有着严格的承接关系,只有充分运用这些语言处理的关键技术,才有可能理解自然语言文本,从而进行知识抽取。自然语言处理是知识抽取的基础,但每一步处理过程中都存在着诸多困难,如分词中的组合型切分歧义、句法分析中的介词短语附着问题、语段分析中的指代词所指、首语省略等问题都是自然语言处理的难点。建立一个通用的自然语言处理平台,在经过完整的分析过程之后再进行知识抽取,还有诸多细节问题需要处理,处理的准确率与普适度有待于提高。

参考文献:

[1] Jionghua Ji. Semi - automatic Ontology - based Knowledge Extraction and Verification From Unstructured Document[D]. State University System of Florida,2000.

[2] Von - Wun Soo, Hsiang - Yuen Yeh, Shih - Neng Lin,et al. Ontology - based Knowledge Extraction from Semantic Annotated Biological Literature[C]. The Ninth Conference on Artificial Intelligence and Applications, 2004.

[3] 化柏林,赵亮. 知识抽取中的嵌套向量分词技术[J]. 现代图书情报技术,2007(7):50 - 53.

[4] 刘开瑛. 中文文本自动分词和标注[M]. 北京:商务印书馆,2000.

[5] Brown P F,Della Pietra S A,Della Pietra V J,et al. Word - sense Disambiguation Using Statistical Methods[EB/OL]. [2007 - 07 - 05]. http://acl. ldc. upenn. edu/P/P91/P91 - 1034. pdf.

[6] Yarowsky D. Decision List for Lexical Ambiguity Resolution:Application to Accent Restoration in Spanish and Freneh[C]. Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics,Las Cruces,NM,1994.

[7] Kaplan,Ronald M. The Formal Architecture of Lexical - Functional Grammar[J]. Journal of Information Science and Engineering, 1989,5:305 - 322.

[8] Jean - Pierre Koenig. Book Reviews: Head - driven Phrase Structure Grammar and German in Head - driven Phrase Structure Grammar[EB/OL]. [2007 - 07 - 06]. http://acl. ldc. upenn. edu/J/J96/J96 - 1005. pdf.

[9] 温有奎,温浩,徐端颐,等. 基于知识元的文本知识标引[J]. 情报学报,2006(3):282 - 288.

[10] John F. Sowa. 知识表示(英文版)[M]. 北京:机械工业出版社,2003.

(作者 E - mail:huabolin@ istic. ac. cn)