信息系统学报
第2卷第2辑(总第3辑):68-77

China Journal of Information Systems
Vol.2 No.2, Oct. 2008: 68-77

从维基分类系统构建中文语义词典研究*

罗志成,马费成,吴晓东,宋倩倩
(武汉大学信息资源研究中心,武汉 430072)

摘 要 本文把维基百科的页面分类系统看成一个概念网络,根据网络的链接特征和句法分析匹配的方法来标注类目之间的语义关系,自动构建了一个大规模的包含大量父子关系与类-实例的语义词典。与人工判别实验的比较结果表明,关系自动识别方法取得了较好的识别率和准确率。

关键词 维基百科,语义词典,关系识别
中图分类号 TP182

1 序言

构建大规模的机器可读的通用知识库是信息检索的一个重要课题[1]。虽然目前统计学习方法占据信息检索的主流位置[2],但信息检索与自然语言处理的融合再次突出了知识库的重要性[3]。人工创建知识库的成本非常高,而且需要大量的时间和精力来维护[4]。另外,现有的大多数知识库都限定于特定领域,覆盖度有限——Cyc[5]和 WordNet[6]是少数的两个例外。中文的大型知识库尤为缺乏。针对这些问题,人们提出了各种本体学习的方法,即从文本学习得到语义词典或者本体。但是学习得到的本体通常非常小,大多还有领域依赖性,并且评价结果显示本体学习的效果比较差[7]。

本文通过利用维基百科来解决知识库的创建问题。维基百科[8]是一个由大量用户创造并具有高覆盖度的在线百科全书。不同于通常的本体学习方法,本文利用维基百科的分类系统所形成的概念网络作为一个半结构化的输入,并在此基础上构建语义词典。此分类系统中存在众多相关关系的概念对,但是概念之间的具体语义关系尚不确定。因此,创建语义词典的任务变成了识别 Is-a(父子关系)和 Not Is-a(非父子关系)关系。本文使用了基于维基百科分类系统网络的链接特征和句法分析匹配的方法对类目之间的关系进行标注。如此,我们得到了一个语义词典。在此基础上,针对一般语义词典中类与实例的混淆,我们进行了自动判断的探索性研究。

2 相关工作介绍与分析

2.1 语义词典

在广义的信息检索任务中,如问答系统、自动文摘和信息抽取等,自然语言处理和信息检索间的

* 国家自然科学基金项目(70773086)
通信作者:马费成,武汉大学信息管理学院教授,e-mail: fchma@whu.edu.cn

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

交互作用已经取得了很好的成果[3]。这些应用使得学者们再一次把关注的焦点放在语义词典和本体资源上。普林斯顿大学开发的 WordNet 是最有名的语义词典之一。目前因为大型通用本体的缺乏,所以很多研究者都把 WordNet 之类的语义词典作为轻量级的本体来使用。但实际上,WordNet 作为本体仍然存在若干缺陷,其中包括概念与实例之间的混淆[9],譬如 Organization(组织)这个概念的下位概念中,同时包含了 Company(公司)和 Red Cross(红十字会)。针对此批评,WordNet 的创建者 George Miller 通过人工判断,对 WordNet 的概念进行了类(Class)与实例(Instance)的区分[10]。

在中文语义词典的建设[11]方面,最知名的包括梅家驹先生的《同义词词林》、董振东先生的《知网》、北京大学计算语言所开发的《现代汉语语义词典》(SKCC)和山西大学构建的中文 FrameNet。这些词典已经应用到了信息检索的各种任务中,如自动文摘[12]等。Cimiano 等[13]指出,服务于自然语言处理的知识库应该具备以下几个特点:(1)领域独立,例如,有一个高的覆盖度,特别是在实例的层面;(2)及时,满足处理新信息的需要;(3)多语种,实现以语言独立的模式处理信息。针对这些要求,目前中文语义词典在及时更新和多语种方面都存在缺陷。

2.2 大众分类法

从 2004 年 5 月开始,维基百科允许用户为文章标引一定的类目,同时又为这些类目标引一定的父类目,从而构成一个页面与类目的分类体系,如图 1 所示。

维基分类系统与大众分类法(Folksonomy)有同有异。大众分类系统往往与传统的叙词表相比较[14]。叙词表中,所有的词语都是限定的,词语之间的连接构成了一定的层次结构;而在大众分类系统中,标签是随意添加的,标签的使用也无一定之规。在一般的提供大众分类服务的网站中,如"美味书签"(http://del.icio.us),其标签都是平行的,而没有构成层次结构。二者相比,叙词表存在用户参与不足、知识更新较慢的缺陷,而大众分类法存在缺乏层次结构、浏览不便的缺陷。

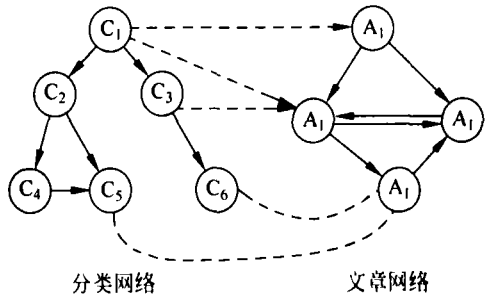

图 1 维基百科页面分类系统示意图

因此,为了结合两者优点,有的研究者利用大众分类网站的标签共现概率[15],或者计算被引向量的余弦相似度[16]来构建标签的层次结构。相对标签网络,维基页面分类系统是通过众多用户的协作编辑构建层次结构。从其产生过程来看,它结合了叙词表具有严格等级层次结构和大众分类系统中用户协作创造的优点。如果把每一个类目都视为一个概念,那么维基页面分类系统对应一个概念网络。这为进一步的研究提供了更多的便利。

2.3 维基百科页面分类系统

如图 1 所示,维基百科类目形成了一个图,并且与文章网络存在联系[17]。它满足所有 Cimiano 提出的要求[13]。但是,遗憾的是,维基百科分类系统中,类目之间关系是不明确的,不能形成一个完善的具有父子等级结构的语义词典。譬如,在类目"文学"下同时包含了子类目"儿童文学"和"文学体裁"。

对于维基百科页面分类系统,目前已经有一些研究,如 YAGO[18]等语义维基项目从分类系统中抽取语义信息。Strube 和 Ponzetto[19]在 2006 年提出,使用维基百科的分类系统作为概念网络来计算

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

词语之间的语义相关度，之后又在英文维基分类系统中，自动识别类目之间的 Is-a 和 Not Is-a 关系[20]。CÄacilia Zirn 等[21]在 Strube 和 Ponzetto 工作的基础上，区分了类目的种类：类或者实例。与上述研究不同，本文主要关注于中文维基分类系统，结合中文的特点，充分利用维基分类系统已有的网状结构，采用了一系列方法来从初始连接中识别 Is-a、Not Is-a 关系，区分了概念的类型(类或者实例)，最终初步构建了一个中文语义词典。

3 语义词典构建系统模型

根据上文分析，本文提出了一个语义词典构建的系统模型，如图 2 所示。下文对图中各个操作进行详细介绍。其中，第 3.2 节将详细说明语义关系识别的算法，第 3.3 节将详细说明类与实例识别的算法，第 4 节将介绍实验结果的评测。

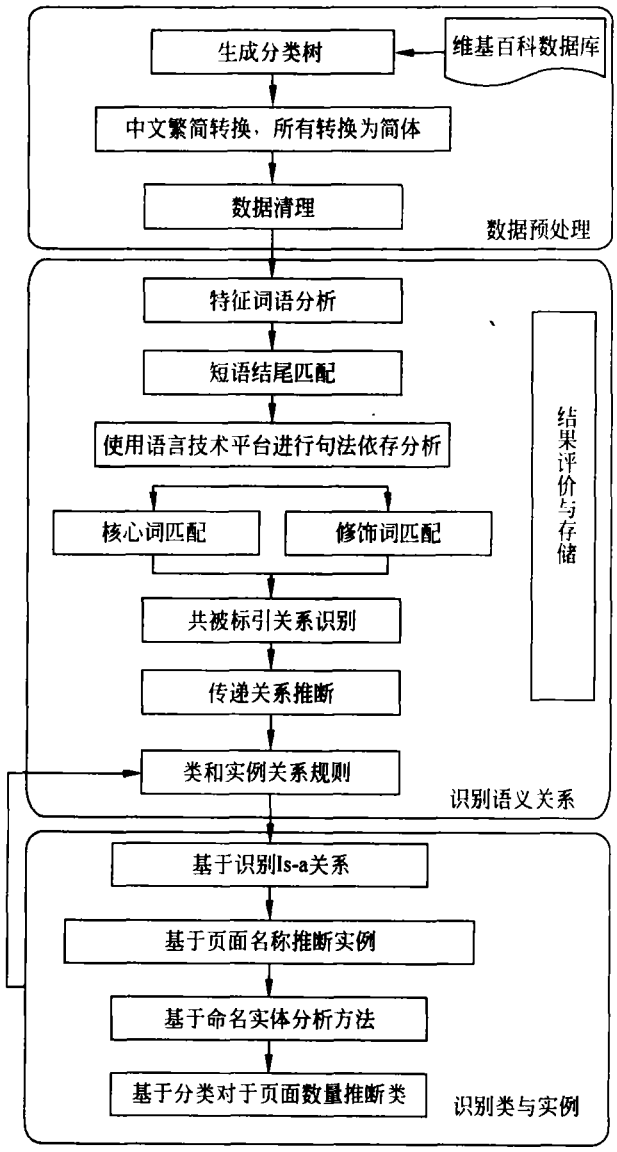

图 2 从维基分类系统构建语义词典模型图

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.1 数据预处理

3.1.1 生成类目网络

目前，维基百科定期免费提供各个语言版本的所有数据，放到网络上供人下载。其中，数据包中包括有 page. sql 文件和 categorylinks. sql 文件，前者记录了所有页面的基本信息，如：页面标题、命名空间、页面长度等，后者记录了各个页面标记的类目信息。虽然维基百科网站也提供了树状形式的类目网络供人浏览[22]，但是这个网页的数据存在如下缺陷：(1)由人工维护，所以数据更新比较缓慢；(2)迫于部分类目深度过深，所以这个网页把这些类目删除掉，如“生物分类树”；(3)由于采用简单的深度遍历算法将图输出为树状结构，所以部分类目深度失衡，例如“自然科学”类目就没有子类目。基于以上原因，我们编程实现了类目网络的自动生成。

3.1.2 中文繁简转换

目前中文存在两种书写系统——繁体中文与简体中文。一般来自中国台湾、香港、澳门的使用者使用正体中文(繁体中文)，来自中国大陆、新加坡、马来西亚的使用者则使用简体中文。作为一个全球华人共同创作的平台，中文维基百科发布的数据中，既有繁体形式的，也有简体形式的，甚至很多在同一篇文章中繁简夹杂。这给我们的抽取工作带来很大的不便：一方面，文本繁简混杂的问题使得我们不能用现有的基于单一文字模式的中文信息处理工具直接分析文本；另一方面，繁简夹杂必然使得我们的研究成果不能很好地得到利用。所以，我们在利用维基百科所提供的简繁对应词表基础上，借鉴 MediaWiki 1.4 的繁简转换功能的“用字模式”[23]，实现了繁简转换功能。

3.1.3 类目网络清理

在维基页面分类系统中，存在着若干为了方便管理而添加的元类目，例如：“维基百科站务”。因为这些类目所含语义信息较少，所以我们必须清理这些类目。我们剔除所有包含以下关键字的类目：维基，列表，模板，维基人，专题，分类，条目，小作品。另外，从数据库自动生成的类目网络中存在一些孤立点，我们将此类类目也全部清除。在清理之前，中文维基类目之间的直接连接数为 21 776 个，清理后的直接连接关系总数为 14 009 个。

3.2 识别父子关系

首先界定两个概念：下位词与“Is-a”关系(父子关系)。语言学家 Fromkin 和 Rodman[24]认为，下位词是一个一般化词语具体化之后的相关词语集合。例如，深红色、朱红、绯红色都是红色的下位词，而红色就是它们的上位词。同时，红色又是颜色的下位词。因此，下位关系也就是一般化术语(如多边形)和它的具体化实例(如三角形)之间的关系。

在计算机科学中，常常将此关系称为 Is-a 关系[25]。例如，用“红色 is a 颜色”来描述红色和颜色之间的下位关系。在知识表示和面向对象编程与设计中，在 A is a B 中，Is-a 表示类 A 是类 B 的子类，即 B 是 A 的父类。换言之，“A is a B”通常意味着概念 A 是概念 B 的具体化，概念 B 是概念 A 的一般化。举例而言，“水果”是“苹果”、“桔子”、“芒果”等概念的一般化。我们可以说，“苹果 is a 水果”。

下文对这些类目之间的 Is-a 和 Not Is-a 关系进行自动识别。文中规则 1～6 的提出基于两点考虑：(1)中文信息处理的现有研究成果，譬如汉语名词复合短语中核心词置后[26]；(2)针对英文维基百科分类系统，Ponzetto 等人通过采用特征词语分析、基于句法的方法、共被标引关系识别、传递关系

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

推断等方法构建了分类体系[20],其最终实验结果的识别率和准确率分别达到了80.6%和91.8%。需要指出的是,如无特殊说明,本文所说进行关系判断的两个类目,都是指两个直接相连的类目。下文用A、B和C来表示三个类目,其中A是B的父类目,B是C的父类目。

3.2.1 特征词语分析

在维基百科中,为了将大的类别进一步细化,往往会按照某种标准来进行划分,例如说,在"意大利人"类目下,包含了如下子类目:意大利人(以地区分类),意大利人(以职业分类)。其中,"以……分类"可以作为判断Is-a关系的特征词语。经过人工辨别分析,本文以"分类"和"区分"作为识别此类的特征词语。

3.2.2 短语结尾匹配

通过对维基类目标记的分析,我们发现很多标记都是名词复合短语。而汉语名词复合短语具有如下特点:短语的末尾词为整个短语的核心词,在功能上代表该短语,短语通过核心词与外部发生联系,限定成分被屏蔽[26]。这给我们两种概念类别判断的思路,一种是核心词匹配(将在下文具体介绍),另一种是对两个类目标记进行结尾匹配,也就是进行简单的字符串匹配。

3.2.3 基于句法的方法

如3.2.2中所述,我们考虑通过核心词对类目的语义关系进行判断。对于所有的类目标记,我们应用哈尔滨工业大学信息检索实验室提供的语言技术平台[27]进行处理。此平台中包含了多个自然语言处理工具,包括词法分析,命名实体识别,指代消解,句法依存分析等。其中,句法依存分析能够识别出句子或短语的各个成分,包括核心词(Head)和各种修饰词(Modifier),具体标记含义参见文献[26]。本文在句法依存分析结果上识别Is-a关系,并区分了Not Is-a关系。cate表示一个类目,level(cate)表示类目的层次,即当前类目到根类目的最短路径,head(cate)表示类目标记的核心词,modifier(cate)表示类目标记的修饰词。

规则一:核心词匹配

$$(head(A) = head(B)) \cap (level(B) - level(A) = 1) \rightarrow B\ Is\text{-}a\ A \quad (1)$$

即:如果两个类目标记的核心词相同,那么它们是Is-a关系,譬如说"哲学家"与"德国哲学家"。其中,类目A和B是两个直接相连的类目,并且B的层次比A要深一层,下文与此相同。

规则二:修饰词匹配

$$(head(A) = modifier(B)) \cap (level(B) - level(A) = 1) \rightarrow B\ Not\ Is\text{-}a\ A \quad (2)$$

即:如果类目A的核心词是B的修饰词,那么它们是Not Is-a关系,譬如说"律师"与"律师事务所"。

3.2.4 共标引关系识别

这种方法识别出标引了同一个页面的类目对。如果维基百科的用户给某个页面标引了两个类目,并且两个类目之间又存在标引与被标引关系,这往往意味着这个页面是两个不同的类目概念在不同粒度下的实例。例如,如图3所示,在维基页面分类系统中,"计算机总线"与"串行总线"同时标引了页面"USB",并且"计算机总线"同时标引了页面"串行总线"。那么我们可以根据传递性推断出一个类目蕴涵了另一个类目,例如图3中"计算机

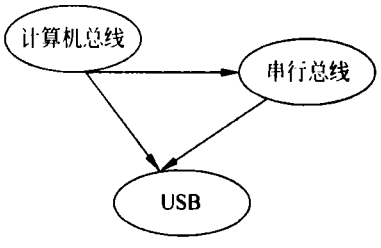

图3 共标引关系示意图

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

总线"蕴涵了"串行总线"。在维基百科中,类目和普通文章都属于页面,所以我们在评测过程中区分了这两种情况。

3.2.5 传递关系推断方法

有的时候,因为分词或者类目词语表述的原因,我们能够识别出 A 和 B 是 Is-a 关系,A 和 C 也是 Is-a 关系。因为 Is-a 关系具有传递性,所以我们给出如下规则:

规则三:传递关系推断

$$(C\ is\text{-}a\ A) \cap (B\ is\text{-}a\ A) \cap (level(C) - level(B) = 1) \rightarrow C\ Is\text{-}a\ B \quad (3)$$

其中,三个类目符合以下条件:B 是 A 的子类目,C 是 B 的子类目,即 A 和 B 直接相连,B 和 C 是直接相连。

3.2.6 类与实例推断父子关系

在下文 3.3 节中,本文根据若干语言学特征进行了类与实例的判别。通常而言,如果一个父类目是类,而子类目是实例的话,那么它们极有可能是父子关系。

规则四:类与实例规则

$$(A\ is\ Class) \cap (B\ isIns\ \tan ce) \cap (level(B) - level(A) = 1) \rightarrow B\ Is\text{-}a\ A \quad (4)$$

3.3 类和实例识别

3.3.1 区分类和实例的基本原则

语义词典在某些实际应用中可能被视为本体,所以我们有必要进一步区分一个类目是类还是实例。Miller 和 Hristea[10]针对 WordNet 的概念区分了类(Class)与实例(Instance),并提出实例的三个判别标准:(1)实例是名词;(2)实例是专有名词,因此它们都是大写的;(3)实例都是特指对象,因此实例不会再包含实例。对于中文而言,没有大小写形式差别,所以无法从形式上加以判断专有名词,但是可以依靠命名实体识别技术来判断专有名词。另外,根据 Miller 提出的第三条标准,笔者认为类是包含其类目的实体,而实例就是不包含其他类目的实体。

由于词语的多义性以及语境的复杂性,有些词语既可以看成类也可以看成实例,需要视具体语境而定。所以,类和实例的判别结果的评测很难通过人工进行判断,需要借助于其他的方法。在 3.2 节中,我们识别了类目之间的语义关系,在此基础上,本文给出识别类与实例的几种方法。

3.3.2 基于 Is-a 关系的分析方法

实例最重要的特点就是特指某一事物,不包含其他实例。对应到上文构建的语义词典,实例所对应的类目将不包含通过 Is-a 关系与之相连的子类目或者页面。反之,如果类目包含多个通过 Is-a 关系与之相连的子类目,那么此类目极有可能是类。基于上述假设,我们提出利用一种 Is-a 关系来识别类的方法,即规则五:

规则五:如果一个类目 A 至少有两个子类目 B、C,那么 A 是一个类。

$$(B\ is\text{-}a\ A) \cap (C\ is\text{-}a\ A) \rightarrow A\ is\ Class \quad (5)$$

因为上文识别的 Is-a 关系都是自动识别,可能会存在一定误差,所以我们部分调整规则五,提出规则六。

规则六:如果一个类目 A 只有一个子类目,但是子类目又包含至少两个子类目,则 A 是类。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$(B\ is\text{-}a\ A) \cap (C\ is\text{-}a\ B) \cap (D\ is\text{-}a\ B) \rightarrow A\ is\ Class \tag{6}$$

3.3.3 基于命名实体分析方法

实例对应于现实世界中的特定实体,也就是自然语言处理中的"命名实体"。所以,我们能够通过对类目进行命名实体识别分析来判断它是否是实例。因为维基百科类目名称都是名词性短语,所以我们只是判断短语的核心词是否为命名实体。此处,我们利用哈尔滨工业大学的语言技术平台进行词法分析,命名实体将标记为人名(nh)、组织(ni)、时间(nt)、日期(nr)、专有名词(nz)或地名(ns)。譬如,类目"哈利波特"所对应的核心词是"哈利波特",而其句法分析得到的词性是人名(nh),则认为类目"哈利波特"是实例。

3.3.4 基于同名维基条目推断方法

在维基百科中创建类目时,编辑者会看到这样的说明:文章应当放置到具有相同名称的类目中。所以文章会有一个名称相同的类目。这样我们就可以通过判断文章的名字来对类目的名字进行判断了。另外,因为每篇文章往往是对应一个具体独特的实体,所以和维基网页文章名字相同的类目是实例的可能性就比较大。譬如,类目"功利主义"下面有一个名称为"功利主义"的条目,则认为类目"功利主义"是实例。

4 实验结果和分析

4.1 总体评测思想

为了判断父子关系识别结果的准确度,我们设计了一个人工判断实验。我们从所有识别出的类目关系中随机抽取800个关系。为了保证所抽取类目对的覆盖面和随机性,我们按照顺序对所有类目对进行分组,每10对一组,然后从每一组中随机抽取一个类目对。如此获得测试集之后,我们让三位同学各自单独标引测试集中的类目对,即选择"父子关系"或者"非父子关系"。对于三位同学标引有争议的类目对,则由另外一个同学进行查准核实。这样得到一份"理想"关系结果集。然后将系统识别的关系与手工识别的"理想"关系作比较,通过计算平均识别率和准确率来评价系统识别关系的质量。对于实验结果,本文采用识别率和准确率两个指标来衡量。识别率是指所有识别出的关系在所有待识别关系中的比例,而准确率是指所有识别出的关系中识别正确的比例。

对于类与实例的识别结果,如前文指出,因为缺乏具体的语境,评测很难通过人工进行判断。所以我们考虑采用Bootstrap的方法,即把类与实例的识别结果作为判断父子关系的数据来源,通过评测这种方法识别父子关系的效果,从而间接评测类与实例的识别效果。

4.2 父子关系识别实验结果

为了检测以上各种概念类型识别算法的效果,我们计算了每种识别算法单独执行时得到的结果。各个算法识别出来的父子关系或者非父子关系总数如表1所示。

从表1中可以看出,本文所采用的各种方法在关系识别上都达到了较高的准确率,除了共标引规则中"两个类目同时标引另外一个类目"这种方法外,所有的识别方法的准确率都在90%以上。在综合各种方法得到的最终结果中,识别出的关系总数占所有关系总数的70.93%,准备率达到94.42%。其中识别的父子关系较多,占所有待识别关系的63.40%。这与非父子关系识别方法较少有关,本研

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

究中用到的非父子关系识别方法只有修饰词匹配方法。

表 1 维基类目关系类型识别结果表

关系识别算法	关系总数	识别结果	识别率/%	测试数量	准确数	准确率/%
所有类目	21 776		—	—	—	—
剔除管理类目和时间类目	14 009		—	—	—	—
特征词语分析	355	Is-a	2.53	26	25	96.15
核心词匹配	4 707	Is-a	33.60	417	408	97.84
短语结尾匹配	2 057	Is-a	14.68	186	179	96.24
修饰词匹配	1 606	Not Is-a	11.46	101	91	90.10
标引同一篇普通文章	3 071	Is-a	21.92	150	141	94.00
标引同一个类目	547	Is-a	3.90	33	28	84.85
传递推断	321	Is-a	2.29	18	17	94.44
类与实例推断	2 462	Is-a	17.57	118	108	91.53
识别父子关系总数	8 882		63.40	586	558	95.22
识别非父子关系总数	1 606		11.46	101	91	90.10
总共识别的关系总数	9 937		70.93	681	643	94.42
未识别出的关系总数	4 072		29.07	—	—	—

在各种方法中，识别率较高的方法包括“两个类目核心词匹配”、“两个类目标记了同一篇普通文章”及“两个类目分别是类和实例”。其中，核心词匹配方法在所有方法中识别率最高，并且准确率也最高，这充分说明了句法依存分析在语义分析中的价值。但是句法依存分析的准确度同样影响到实验结果。例如，在句法依存分析中，通常把动词视为句子的核心词。虽然在本研究中，类目都是短语，但是句法依存分析工具仍然把短语中的动词当做短语的核心，类目“效力德国球会的球员”被分析为“效力”。另外，部分句法分析结果产生错误，例如类目“清朝妃嫔”，分词之后结果为“清朝/nt 妃/v 嫔/n”，最终句法分析结果把“妃”作为核心词。这种分析的误差也造成修饰词匹配方法准确率下降。

4.3 类与实例识别实验结果

与 4.2 节类似，为了检测以上各种类和实例识别算法的效果，我们计算了每种识别算法单独执行时得到的结果。各个算法识别出来的类或者实例总数如表 2 所示。

表 2 类和实例识别结果表

识别算法	关系总数	识别结果	识别率/%
所有类目	14 009		—
基于识别的 Is-a 关系	1 539	类	10.99
基于命名实体分析方法	662	实例	4.73
基于同名维基条目推断方法	2 723	实例	19.44
所有识别的类目	4 924		35.15
未经识别的类目	9 085		64.85

从表 2 可以看出，所有识别出的类或者实例共有 4 924 个，识别率为 35.15%。其中包括 3 385 个实例，占总类目数量的 24.16%，另有 1 539 个类，占总类目的 10.99%。此实验识别率较低，这与每种判断方法各受到一定限制有关。例如，根据 4.2 中识别的 Is-a 关系来进行判断受到关系类型识别率的影响。虽然识别出来的类和实例的结果有限，但是它对于父子关系的判断却有较大的帮助。从表 1

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

可知,依据类与实例推断方法识别的结果为2 462,识别准确率为91.53%。

5 结论

本文提出了一种从维基页面分类系统中构建语义词典的方法。模型中根据中文维基百科分类系统的特点和项目的需要,运用了若干自然语言处理技术,提出了多种语义关系的识别算法,并建立了一个针对维基分类系统的语义词典构建系统。通过与人工判断语义关系方法的实验结果相比较,该语义词典构建系统在识别率和准确率上都取得了较好的效果。进一步的工作中,将考虑与现有的中文语义词典相融合,并且加入其他语言的链接,从而构建一个多语种的语义词典。

参考文献

[1] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval[M]. New York: ACM Press, 1999.

[2] 黄昌宁. 统计语言模型能做什么[J]. 语言文字应用, 2002, (1): 77-84.

[3] 王灿辉, 张敏, 马少平. 自然语言处理在信息检索中的应用综述[J]. 中文信息学报, 2007, 21(2): 35-45.

[4] Mohammad S, Hirst G. Distributional Measures as Proxies for Semantic Relatedness[EB/OL]. http://www.cs.toronto.edu/compling/Publications.

[5] Lenat D, Guha R. Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project [M]. Reading, Mass: Addison-Wesley, 1990.

[6] Fellbaum C(Eds.). WordNet: An Electronic Lexical Database[M]. Cambridge: The MIT Press, 1998.

[7] Buitelaar P, Cimiano P, Magnini B(Eds.). Ontology Learning from Text: Methods, Evaluation and Applications [M]. Amsterdam, the Netherlands: IOS Press, 2005.

[8] Wikipedia[EB/OL]. http://www.wikipedia.org/.

[9] Oltramari A, Gangemi A. Restructuring wordnet's top-level[J]. AI Magazine, 2002, 40(5): 235-244.

[10] Miller G, Hristea F. WordNet nouns: Classes and instances[J]. Computational Linguistics, 2006, 32(1): 1-3.

[11] 刘挺, 车万翔. 中文语义处理[EB/OL]. http://ir.hit.edu.cn/.

[12] CHEN Yanmin, LIU Bingquan, WANG Xiaolong. Automatic text summarization based on textual cohesion[J]. Journal of Electronics (China), 2007, 24(3): 338-346.

[13] Cimiano P, Pivk A, Schmidt-Thieme L. Learning taxonomic relations from heterogenous sources of evidence [C]. Buitelaar P, Cimiano P, Magnini B (Eds.), Ontology Learning from Text: Methods, Evaluation and Applications. Amsterdam, the Netherlands: IOS Press, 2005: 59-73.

[14] Voss J. Collaborative Thesaurus Tagging the Wikipedia Way[EB/OL]. http://arXiv.org/abs/cs/0604036.

[15] Schmitz P. Inducing ontology from flickr tag[C]. In Collaborative Web Tagging Workshop at WWW-2006, Edinburgh, Scotland, 2006: 210-214.

[16] Heymann P, Garcia-Molina H. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems[EB/OL]. http://labs.rightnow.com/colloquium/papers/tag_hier_mining.pdf.

[17] Zesch T, Gurevych I. Analysis of the wikipedia category graph for NLP applications [C]. Proc of the TextGraphs-2 Workshop, NAACL-HLT, Rochester, New York, 2007: 1-8.

[18] Suchanek F, Kasneci G, Weikum G. YAGO: A core of semantic knowledge unifying wordnet and wikipedia [C]. Proc of WWW-2007. Banff, Alberta, Canada, 2007: 697-706.

[19] Strube M, Ponzetto S. WikiRelate! Computing semantic relatedness using wikipedia[C]. Proc of AAAI-06, Boston: Mass, 2006: 1419-1424.

[20] Ponzetto S, Strube M. Deriving a large scale taxonomy from wikipedia [C]. Proc of the 22nd National Conference on Artificial Intelligence, Canada, 2007: 1440-1447.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[21] Zirn C, Nastase V, Strube M. Distinguishing between instances and classes in the wikipedia taxonomy[C]. Proc of the 5th European Semantic Web Conference, Tenerife, Spain, 2008: 376-387.
[22] Wikipedia Category Overview[EB/OL]. http://stats.wikimedia.org/EN/CategoryOverview_ZH_Complete.htm.
[23] Automatic Conversion between Simplified and Traditional Chinese[EB/OL]. http://meta.wikimedia.org/wiki/Automatic_conversion_between_simplified_and_traditional_Chinese.
[24] Fromkin V, Rodman R. Introduction to Language[M]. London: Holt, Rinehart and Winston, Inc., 1988.
[25] Brachman R, What IS-A is and isn't: An analysis of taxonomic links in semantic networks[J]. IEEE Computer, 1983, 16(10): 30-36.
[26] 马金山. 基于统计方法的汉语依存句法分析研究[D]. 哈尔滨: 哈尔滨工业大学, 2008.
[27] 语言技术平台 LTP. [EB/OL] http://ir.hit.edu.cn/.

Research on Building Chinese Ontology from Wikipedia Category System

LUO Zhicheng, MA Feicheng, WU Xiaodong & SONG Qianqian
(Center for Studies of Information Resources, Wuhan University, Wuhan 430072)

Abstract We take the category system in Wikipedia as a conceptual network. We label the semantic relations between categories using methods based on connectivity in the network and lexico-syntactic matching. As a result we are able to derive a Chinese ontology containing a large amount of subsumption, i.e. is-a, relations. Finally, an experimental was carried out which compared the human subject extraction results to our system result, and the recall and the precision showed that our model do a good job.

Key words Wikipedia, Chinese ontology, Hyponym discovery

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net