

本体概念间上下位关系抽取研究

汤 青, 吕学强, 李 卓

(北京信息科技大学 网络文化与数字传播北京市重点实验室, 北京 100101)

摘 要: 以匹配“是一个”模式的句子作为研究对象, 提出了一种基于句法分析与规则匹配相结合的上下位关系抽取方法。针对结构复杂的句子中下位概念抽取困难的问题, 借助句法分析的结果, 给出了下位概念中心词的获取方法; 结合中心词位置信息和概念边界信息, 设计了合适的下位概念抽取规则。实验结果具有较高的准确率和召回率, 证实了方法的有效性。

关键词: 上下位关系; 本体概念; “是一个”模式; 句法分析; 概念中心词

中图分类号: TP391.4

文献标识码: A

文章编号: 1000-7180(2014)06-0068-04

Research on Domain Ontology Concept Hyponymy Relation Extraction

TANG Qing, LV Xue-qiang, LI Zhuo

(Beijing Key Laboratory of Internet Culture and Digital Dissemination Research,
Beijing Information Science and Technology University, Beijing 100101, China)

Abstract: This paper proposes a hyponymy extraction method based on parsing and rule matching, using sentences that meeting “is-a” pattern as research subjects. To solve the problem that difficult to get hyponym from complex sentences, this paper presents a method to get center concept based on parsing result. According to the position of center concept and hyponym boundaries, this paper designs a rule for hyponym extraction. Experiment result gets a higher accuracy rate and recall rate, confirming the validity of the method.

Key words: hyponymy; ontology concept; “is-a” pattern; parsing; center concept

1 引言

在自然语言处理领域, 本体概念的自动化抽取已经取得较大进展^[1], 本体概念间关系的分类理论研究也取得较大进步^[2]。随着 web 资源的爆炸式增长, 将网络上的海量数据形成一个互相关联的网络, 实现信息的高速运作与管理成为新的问题, 如何自动化抽取概念间关系就成为当前本体研究的重点。

本体概念间核心关系是上下位关系, 其研究主要以基于模式匹配的方法为主^[3]。其中, “是一个”模式是典型的上下位关系模式, 它可以简单理解为“A 是一个 B, 则 A 和 B 具有上下位关系”。本文基于该模式进行上下位关系抽取研究, 把上下位关系抽取

转化为句子中上位概念和对应下位概念的抽取问题, 实验取得较好结果。

2 “是一个”模式

概念间^[4]的上下位关系含两个元素: 上位概念和下位概念。借鉴 WordNet 对上下位关系进行定义^[5]。

定义 1 概念对具有上下位关系: 如果给定概念 x, y , 句子“ x 是一种 | 类 | 个 | 部 | ... | y ”是可以接受的, 则“是一种 | 类 | 个 | 部 | ”称为关系特征词, x 称为 y 的下位概念(hyponym), y 称为 x 的上位概念(hyerym), 上下位关系记为 $\text{hyponymy}(x, y)$ 。

收稿日期: 2013-08-30; 修回日期: 2013-10-22

基金项目: 国家自然科学基金项目(61171159, 61271304); 北京市教委科技发展计划重点项目暨北京市自然科学基金 B 类重点项目(KZ201311232037)

从定义中发现,关系特征词“是一个”作为搭建上下位关系的桥梁,在句子中起到主要组织者的作用.利用关系特征词的特点,设定上下位关系模式,便于抽取上下位关系.

2.1 “是一个”模式的获取

根据定义1可知,在“Troy是一款不错的CPU核心”句子中存在上下位关系 hyponymy (Troy, CPU 核心),“是一款”是关系特征词.类似关系特征词在语料上的词频统计结果的示例如表1所示.

表1 类似“是一个”的词语统计结果示例

是—* 频次	个	种	款	样	些	家	件	项	台
	1 151	746	184	59	49	49	38	36	35

形式上满足“是一个”的词语集合{是一种,是一块…}中,某些元素不能在概念间搭建上下位关系,称为关系伪特征词,它们的结构为“是+副词”;然而,另一些元素能在概念间搭建上下位关系,这类词的结构为“是+量词”,但它们与某些前缀词搭配后会产生歧义,导致上下位关系不能被接受.因此,需要消除伪关系特征词、前缀搭配歧义两方面因素对实验造成的影响.综合以上分析,基于有效的关系特征词定义“是一个”模式如下:

定义2 如果给定句子,能与以下模式结构相匹配,称该句子匹配“是一个”模式.

Define constant

{(量词)=<种|个|款|家|件|项|台|套|组|块|条|位|部|片|名|段|场|间|群|体|类|只|根|层|颗|支|笔…..|>

(前缀搭配)={ x | $x \notin$ {不,近似,貌似,好像,曾经,也许,不再,几乎就,并不,看起来,传说中…}}

Define Pattern

{模式:< * C1>< * ! 前缀搭配><是>—< ! 量词>< * C2> }

其中,“!”表示定义一个常量,“*”表示定义中可以出现任意次数,“|”表示或关系,“< * C1>”、“< * C2>”为任意字符串,“C1”和“C2”为概念,则S为匹配“是一个”模式的句子.

从语料中获得匹配“是一个”模式的句子,将抽取上下位关系的问题转化为抽取上位概念和对应下位概念的问题.关系特征词“是一个”在句子中充当谓语,上位概念和下位概念分别充当“是一个”的宾语和主语.上位概念是从句子中“是一个”和右侧最近的逗号(或句末)之间的内容中抽取,将这部分内容记为RS,下位概念是从句子中“是一个”左侧内容中抽取,将这部分内容记为LS.

2.2 “是一个”模式句型划分

在匹配“是一个”模式的句子S中,RS结构简单,故从RS中抽取上位概念比较容易.本文采用基于规则的方法^[3],规则为:(1)若RS中不含“的”字,则通过剥离RS获得上位概念;(2)若RS中含“的”字,则取RS中位置序号最大的“的”字到其末尾之间的内容,再对其剥离,获得上位概念.

在匹配“是一个”模式的句子中,LS可能含多个逗号,而含多个逗号的句子一般都含多个主谓结构或含多个概念,故从LS中抽取下位概念有一定的难度.根据LS中是否含逗号将句子划分为简单句和复杂句,分别采取不同的方法抽取下位概念.

定义3 给定匹配“是一个”模式的句子S,截取句子S中“是一个”左侧的内容(LS),如果LS中不含逗号,则S称为简单句,如果LS中含逗号,则S称为复杂句.

3 简单句中的下位概念抽取

概念边界的确定对于概念的准确抽取十分关键,而规则中存在的泛化问题导致难于准确判断本体概念的边界,因此需要构建一部剥离词典^[4]以辅助上下位概念的抽取.

3.1 剥离词典的构建

概念边界词相对比较固定,采用多策略的本体术语抽取方法^[6]抽取的概念和能匹配“是一个”模式的句子作为训练集,概念在句子中的边界词即为所要抽取的剥离词.剥离词不含名词成分的词,示例见表2.

表2 剥离词典元素示例

认为 表示 发现 说 看成 并非 充其量 终归 的确 任何 说明 由于 一款 而 也许 毕竟 本质上 可以 既有 诞生

剥离规则:从待剥离串左右边界出发,由外向内截取字符串,如果所截取字符串在剥离词典中存在,则将其从待剥离串中剥离;由外向内,层层剥离,直到待剥离串中不再含有剥离词.

3.2 基于规则的下位概念抽取

由定义可知,简单句只含一个主谓结构,主语是由所要抽取的下位概念充当,它可以是单个词语,也可以是具有并列关系的多个词语.根据对简单句的分析,发现它有两个特点:(1)句子结构简单;(2)概念左边界词以定语助词“的”字修饰.本文根据简单句中LS部分是否含定语助词“的”,分为两个支流,对应支流的下位概念抽取规则如表3所示.

表 3 简单句下位概念抽取规则

支流	LS 满足的条件	对应抽取规则
1	LS 含“的”字	将 LS 中“的”字以后的内容根据剥离规则进行剥离,剥离后剩余内容即为下位概念.
2	LS 不含“的”字	将 LS 根据剥离规则进行剥离,剥离后剩余内容即为下位概念.

4 复杂句中的下位概念抽取

根据复杂句定义,其特点有:(1)LS 中可能含多个主谓结构;(2)LS 中可能含多个概念;(3)LS 中部分概念之间存在着并列关系.举例如下:

例 1. 脚本是批处理的延伸,是一种程序.

例 2. Tomcat 和 IIS、Apache 等 Web 服务器一样,具有处理 HTML 页面的功能,另外它还是一种 Servlet 和 JSP 容器.

根据对例子的分析发现,仅采用语法规则进行表层的匹配不能准确抽取下位概念,需要深入句子结构进行分析.本文采用哈尔滨工业大学的语言技术平台 LTP^[7]对复杂句进行依存句法分析,过滤掉分析结果中的无关概念,获取“是一个”的主语成分进而抽取下位概念.

4.1 句法和规则相结合的下位概念中心词抽取

如果某个词或者是某些具有并列关系的词与“是一个”之间具有主谓关系,那么它们就是所要抽取的下位概念.故下位概念抽取主要借助 LTP 所定义的主谓关系(SBV)、并列关系(COO)和相似关系(SIM)等信息.经测试发现,LTP 对常见的句法关系(如主谓关系、并列关系)的分析较为准确.但 LTP 内嵌的分词系统将概念切割为碎片,故被 LTP 标记为主语的词语不能直接作为下位概念,它仅仅是下位概念的一部分,本文将之称为下位概念中心词,其定义如下:

定义 4 下位概念中心词:是下位概念或下位概念的子串.从句法分析的角度理解,即是被 LTP 分析为谓语“是一个”的主语成分的词,或是被 LTP 标记为与主语成分具有并列关系的词,以及与前述两种词具有同义关系的词.

例 1 的句法分析结果如图 1 所示.

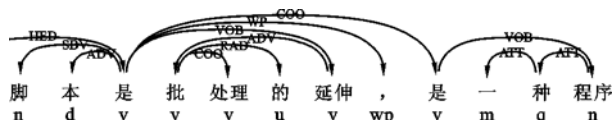


图 1 依存句法分析示例(例 1 句法分析结果)

例 1 中只有“脚”被标识为 SBV,帮助过滤了领域概念“批处理”,减少了其他概念对抽取结果的影响.其上下位关系的初步抽取结果为:[* 脚 *,程

序].

同样,例 2 中“Tomcat”和“它”均被标记为 SBV,“IIS”、“Apache”和“服务器”均被标记为 SIM.和标记为 SBV 的词语具有并列关系或相似关系的词在整个句子中也充当主语角色,例 2 中存在“多对一”上下位关系,其初步抽取结果为:[* Tomcat *, Servlet 和 JSP 容器],[* IIS *, Servlet 和 JSP 容器],[* 服务器 *, Servlet 和 JSP 容器].

综合上述分析,基于实验数据,设计下位概念中心词抽取的规则匹配算法为:

输入:[复杂句,复杂句法分析结果,上位概念];

输出:[下位概念中心词,复杂句,上位概念];

Step1. 被标记为 SBV 的指代词全部过滤.

Step2. 如果只有一个标记为 SBV 的词语,则保留.

Step3. 如果被标记为 SBV 的词为英文词汇,则保留,并过滤该句中其他标记为 SBV 的词.

Step 4. 如果句子不满足规则 3 中条件且存在多个词被标记 SBV 的词,且被标记为 SBV 的某个词和“是一个”之间不存在逗号,则保留,并将该句中存在的其他被标记为 SBV 的词过滤掉.

Step 5:当存在多个被标记 SBV 的词和“是一个”之间都存在逗号,则将位置序号最大的词保留.

Step 6. 判断 SBV 的词是否为指代词,如果被标记为 SBV 的词为指代词,将指代词左边距离自己最近且被标记为 SBV 的词保留;则将指代词右边距离它最近的被标记为 SBV 的词语保留.

Step 7. 如果存在标记为 SBV、COO 和 SIM 的词语,则判断标记为 COO 或 SIM 和标记为 SBV 的词之间的词语是否被标记为 HED、VOB 或 DE,如果有被标记,则认为标记为 COO 和 SIM 的词与主语之间没有并列关系并将其过滤,反之则保留.

Step 8. 保留下的即为下位概念中心词,输出结果.

4.2 融入中心词信息和同义特征词的下位概念抽取

下位概念中心词是下位概念的一部分,借助其位置信息,通过语法规则匹配算法确定概念边界,即可抽取下位概念.在语法规则匹配的过程中,部分句子中存在具有同义关系的多个下位概念,需要借助同义关系特征词词典辅助来抽取.为了描述方便,定义同义关系特征词概念如下:

定义 5 同义关系特征词:在句子中能把其右边的某个概念与其左边的某个概念之间建立起同义关系,即它是同义关系的表征.具体见表 4.

表4 同义关系特征词示例

同义关系特征词集合 $E = \{\text{又译为, 原名, 又称, 非正式中文名称, 又名, 被称为, 又译为, 指, 也叫, 又称为, 又称, 指, 全称, 又简称, 被叫做, 是, 叫, 即}\}$

借助上述同义关系特征词词典,设计基于中心词位置信息的下位概念抽取算法。根据结构特征将概念分为三类:纯英文概念,纯中文概念,中英文混合概念。纯英文概念获取相对容易,纯中文概念需要判断边界词,中英文概念在形式上一般是“英文+中文”的结构,而LTP对中英文概念所定位的主语中心语为中文部分。故下位概念抽取算法为:

输入:复杂句+该句中上位概念;

输出:[上位概念,下位概念]关系对;

Step1: 首先将句子分词,并标注词性;

Step2: 如果概念中心词为英文且其边界词为英文单词,则与边界英文词合并,作为下位概念;反之则概念中心词为下位概念;

Step3: 如果概念中心词被引号引入,引号里内容为下位概念,单引号优先,其次是双引号。如果不被引号引入,截取该中心词的最近左边逗号(或句首)和最近右边逗号(句尾)之间的内容,记为 P ;

Step4: 如果 P 中概念中心词右边的词语是同义关系特征词集合 E 中的任何一个元素,那么就用特定谓语元素将 P 分为两部分内容,分别记为 P_1 和 P_2 ;

Step5: 扫描 $P(P_1$ 或 $P_2)$ 中概念中心词的位置,设定距离 $P(P_1$ 或 $P_2)$ 首部位置距离为 K ;

Step6: 如果 $K \leq 6$,判断概念中心词前是否有“的”,如果有,则截取“的”和概念中心词之间的内容作为候选下位概念;

Step7: 如果 $K > 6$,判断概念中心词词右边边界,如果右边词汇为名词性的词,则概念中心词与其结合为候选下位概念,否则不结合;

Step8: 对候选下位概念采用3.1节中剥离规则进行剥离,获得下位概念;

Step9: 输出结果。

采用上述方法抽取的上下位关系结果中可能存在非领域概念关系对。为保证结果的准确度,用文献[6]中方法抽取的领域概念作为基准参考,采用最小公共子串(长度为2)对上下位关系抽取结果中的概念进行匹配,过滤掉非领域概念关系对,获得实验最终结果。

5 实验及评价

实验数据由百度百科的计算机领域和维基百科的计算机领域共750篇数据构成。从实验数据中提

取出2098个匹配“是一个”模式的句子,以它们作为研究对象。采用3.1节中方法创建剥离词典,词典中含有426个剥离词。采用文献[6]中方法抽取概念3050个,用于过滤非领域概念所在的关系对。

纯中文概念所在关系对、纯英文概念所在关系对以及中英文组合概念所在关系对等3类概念所在关系对能被正确的抽取。

6 结束语

本文借助句法工具获得概念中心词,并借助中心词位置信息抽取下位概念,从而抽取到上下位关系。这为本体其他关系的抽取开启了新的思路。结果中也存在个别错误,在以后的工作中,考虑对LTP分词系统加入用户词典提高句法分析的准确性,对上下位关系的可信度进行研究。

参考文献:

- [1] Lee C M, Huang C K, Tang K M, et al. Iterative machine-learning chinese term extraction[C]// Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries. Taipei, 2012:309-312.
- [2] 贾黎莉. Ontology 构建中概念间关系的研究[D]. 北京:中国农业科学院, 2007.
- [3] Liu L, Zhang S, Diao L, et al. Acquiring ISA relations from Chinese free text based on multiple patterns[C]// Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery. Shandong: Jinan, 2008:160-164.
- [4] 刘磊,曹存根,王海涛,等. 一种基于“是一个”模式的下位概念获取方法[J]. 计算机科学, 2006, 33(9):146-151.
- [5] Miller G. WordNet: an on-line lexical database[J]. International Journal of Lexicographer, 1990, 3(4):234-244.
- [6] 何琳. 基于多策略的领域本体术语抽取研究[J]. 情报学报, 2012, 31(8):798-804.
- [7] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: a chinese language technology platform[C]// Proceedings of the 23rd International Conference on Computational Linguistics. Israel, Haifa, 2010:13-16.

作者简介:



汤青女, (1986-), 硕士研究生。研究方向为中文信息处理。