一种基于语义的上下位关系抽取方法

陈金栋 肖仰华
(复旦大学计算机科学学院 上海 200433)

摘 要 分类体系主要由上下位关系组成,传统的基于模板的上下位关系抽取方法分为两类:第一类方法只使用高质量的模板导致低召回率;第二类方法使用所有可用的模板导致低精度。根据模板的质量将其分为更细粒度的强句法模板和弱句法模板。为了提高弱模板的精度,将弱模板和概念/实体结合构建语义模板。结合强句法模板和语义模板,提出一套新颖的框架从语料中抽取上下位关系,具有高精度和召回率的特点。在中英文语料上进行的实验,实验结果证明了框架的有效性。

关键词 知识图谱 分类体系 关系抽取 上下位关系 句法模板

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.02.039

HYPERNYMY RELATION EXTRACTION BASED ON SEMANTICS

Chen Jindong Xiao Yanghua
(*School of Computer Science, Fudan University, Shanghai 200433, China*)

Abstract Taxonomies mainly compose of hypernymy relations. The traditional pattern-based methods for hypernymy relation extraction are usually divided into two categories. One only uses high-quality patterns, which causes low recall rate. The other uses all available patterns, which leads to low precision. According to the quality of patterns, it can be divided into more fine-grained strong syntactic pattern and weak syntactic pattern. In order to improve the accuracy of weak syntactic pattern, a semantic pattern was constructed by combining with concept/entity. Combining strong syntactic pattern with semantic pattern, we proposed a novel framework for extracting hypernymy relations from corpus with high precision and recall rate. Experiments on Chinese and English corpus demonstrate the effectiveness of the framework.

Keywords Knowledge graph Taxonomy Relation extraction Hypernymy relation Syntactic pattern

0 引 言

分类体系广泛应用于短文本分类[1]、Web 服务发现[2]、表示学习[3]等领域。分类体系包含实体、概念以及上下位关系,其中上下位关系也称为 isA 关系。本文用 hyponym(A,B)表示上下位关系,例如 hyponym(苹果,水果)表示“苹果”是“水果”的下位词,“水果”是“苹果”的上位词。

上下位关系抽取是大规模中文分类体系构建的重要方法之一。早期的分类体系比如 WordNet[4]是人工构建的,这种分类体系精度较高但是规模很小。因此,近期的研究工作都围绕在自动化分类体系构建。基于模板的方法是分类体系构建的主流方法之一。文献[5]利用人工定义的 Hearst 模板从文本中抽取上下位关系。为了进一步提高上下位关系的召回率,文献[6]提出了一套 bootstrapping 的框架,从文本中获取上下位关系。

大部分句法模板都面临了低质量或者低覆盖率的问题,高质量高覆盖率的模板非常少。因此,先前的工作使用高质量的模板来确保精度,同时采用 bootstrapping 的方式来提高召回率[6]。但是在 bootstrapping 过程中产生的模板质量较低,这导致了语义漂移的问题[7]。上述方法在中文上的效果比英文更差,因为中文的表达更加灵活,语法更加复杂,这导致中文高质量的模板非常少[8]。因此目前出名的分类体系都是英文

收稿日期:2018-09-15。国家重点研发计划项目(2017YFC0803700)。**陈金栋**,硕士,主研领域:知识图谱。**肖仰华**,教授。

的,如 WikiTaxonomy[9]、YAGO[10]、Probase[11],中文的高质量高覆盖率的分类体系几乎不存在。

本文主要针对众多低质量高覆盖率的句法模板,将这些模板称为弱句法模板。同时,将高质量高覆盖率的句法模板称为强句法模板。弱句法模板质量低的主要原因是缺乏语义信息,因此把它和实体或概念结合,设计了一种新模板——语义模板来获取更多高精度的上下位关系。例如,"NP 是 NP"是弱句法模板,其中 NP 表示名词短语。已知"北京"是一个实体,将上面提及的弱句法模板和实体结合得到语义模板"北京是 NP"。利用该语义模板,可以从句子"北京是中国首都"中获得 hyponym(北京,中国首都)。

基于强句法模板和语义模板,本文提出了一套新颖的迭代框架用于上下位关系抽取,强句法模板进行第一轮关系抽取,在迭代的过程中使用语义模板来抽取更多的上下位关系,这极大地提高了召回率。本文提出的方法能够克服传统 bootstrapping 方法中的语义漂移的问题,因为在迭代的过程使用了语义信息,能克服弱句法模板的低质量问题。本文在中英文数据集上进行实验,实验结果证明了方法的有效性和通用性。

1　相关研究

主流的上下位关系抽取方法可以分为三种:基于模板的方法、基于百科全书的方法和基于嵌入的方法。

基于模板的方法使用句法模板从文本中抽取上下位关系。文献[5]是第一个将句法模板用于上下位关系抽取,提出了一套自动化的上位词获取算法,利用 Hearst 模板从非结构化文本中获取上位词。文献[6]提出了一套迭代式算法从互联网数据中抽取上下位关系。该算法定义一些种子关系实例,利用它们获取新的句法模板,这些句法模板可用于抽取新的关系实例,重复执行上述步骤,直到没有新的关系实例产生为止。文献[12]使用搜索引擎发现匹配句法模板的句子并从中抽取上下位关系。文献[13]训练一个上下位关系分类器来发现有用的依赖路径,然后将分类器用在新的语料上识别新的上下位关系。Liu 等[14]提出了一套迭代抽取中文上下位关系方法,只用到了两个强句法模板,完全忽略了弱句法模板。Wu 等[11]提出了一套英文上下位关系抽取方法,构建了一个大规模的英文分类体系。上述方法没有严格区分高质量模板和低质量模板,都面临了低精度或低覆盖率的问题。

基于百科全书的方法从相对结构化的百科全书中抽取上下位关系。文献[9]以维基百科的种类系统为数据源,把它建模成一个语义网络,将语义网络中的关系分为上下位关系和非上下位关系。文献[10]将维基百科的种类系统中的概念映射到 WordNet 来获取大量的上下位关系。类似的方法也可用于中文,文献[8,15]使用相似的方法分别从中文维基百科和百度百科中抽取上下位关系。这种方法的精度较高,但是覆盖率较低。

基于嵌入的方法将单词或短语映射到一个隐式的向量空间,然后基于这些向量发现上下位关系。文献[16]基于词向量来获取上下位关系。文献[17]将语法规则也映射到隐式空间,为发现上下位关系提供更多的特征。但是这些模型的精度较低(80% 左右),这导致了此类方法不满足实际工程的需要。

2　句法模板和语义模板

本文目标是从文本中抽取上下位关系。在详细介绍本文提出的算法之前,先定义句法模板和语义模板。

2.1　句法模板

高质量的模板可以产生高精度的上下位关系,而低质量的模板倾向于产生低精度的上下位关系。因此,根据模板精度将其分为强句法模板和弱句法模板。

定义 1　模板 P 的精度定义如下:

$$pre(P)=\frac{\#\text{correct}}{\#\text{matched}} \tag{1}$$

式中:分母表示模板 P 从语料库中抽取的上下位关系数量;分子表示这些关系中是正确的上下位关系数量。

定义 2　给定一个模板精度阈值 γ,如果模板 P 满足 $pre(P)\geqslant\gamma$,则它是一个强句法模板;反之,它是一个弱句法模板。

$pre(P)$是针对特定语料库而言的,一般是从语料库中采样得到样本数据,在样本数据上评估得到 $pre(P)$。阈值 γ 的设置对于区分强弱句法模板至关重要,在设定阈值时需要考虑两点:第一,在不同语言上 γ 的设定是不同的,因为语言的差异,相同的句法模板在不同语言上的精度是不同的;第二,当期望得到高精度的上下位关系时,往往会将 γ 设置的比较高。

表 1 显示了中文中常用的 Hearst 句法模板。当 $\gamma=0.85$ 时,P_{syn1} 和 P_{syn2} 是强语法模板,P_{syn3} 和 P_{syn4} 是弱句法模板,其中精度是在每个模板抽取得到的 300 组上下位关系上评估得到的。一方面,强句法模板质量较高,可以产生高精度的上下位关系,但如果仅使用强句法模板,召回率太低。另一方面,弱句法模板可用于提升召回率,但弱句法模板产生的上下位关系精度

太低。为了平衡精度和召回率,本文设计了一种语义模板来解决此问题。

表 1 中文 Hearst 句法模板

ID	模板	精度
P_{syn1}	NP 是一(个/种…/本) NP	94.7%
P_{syn2}	NP{,NP}* 等 NP	92.0%
P_{syn3}	NP 是 NP	80.6%
P_{syn4}	NP 包括{,NP}* NP	78.4%

2.2 语义模板

本文先定义元语义模板,因为语义模板的定义依赖于元语义模板。

定义 3 元语义模板是由弱句法模板和一个概念占位符 $CON 或实体占位符 $ENT 组成。

如表 2 所示,元语义模板 P_{sem2} 和 P_{sem3} 分别是通过弱句法模板 P_{syn3} 结合概念占位符 $CON 和实体占位符 $ENT 构成的。基于元语义模板定义语义模板。

表 2 中文元语义模板

ID	模板
P_{sem1}	CON 包括{,NP}* NP
P_{sem2}	NP 是 CON
P_{sem3}	ENT 是 NP

定义 4 语义模板是由一个具体的概念或实体来实例化元语义模板中的概念或实体占位符产生的。

例如“水果包括{,NP}* NP”是一个语义模板,它是由概念“水果”替换元语义模板 P_{sem1} 中的概念占位符得到的。

3 框 架

基于强句法模板和语义模板,本文设计了一个迭代式抽取框架从文本中抽取上下位关系。基本思路是用强句法模板获取高精度的上下位关系,用语义模板来提升召回率同时保证上下位关系的精度。如图 1 所示,框架由两个主要部分组成:预备抽取和迭代抽取。在预备抽取中,使用一组固定的强句法模板来获得高精度的上下位关系。在迭代抽取中,使用语义模板来提升召回率,获取更多的上下位关系。迭代的动力来自上一次迭代中生成的新概念/实体。从不同模板生成的上下位关系的交集中得到新概念/实体,这确保了新概念/实体的质量。新概念/实体用于构造语义模板。在迭代中使用语义信息,因此解决了语义漂移的问题。表 3 总结了本文中使用的符号。

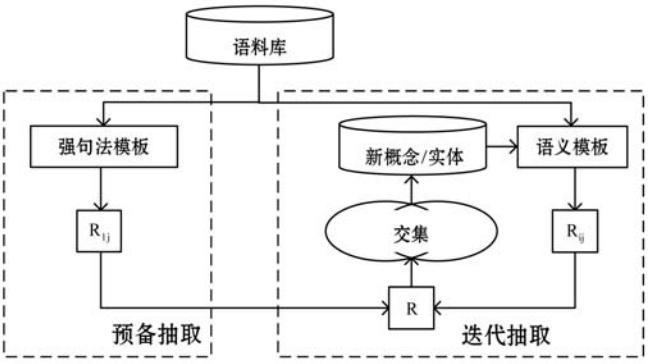

图 1 上下位关系抽取框架

表 3 符号

符号	意义
R	从语料中抽取得到的上下位关系集合
R_{ij}	第 i 轮迭代第 j 个模板抽取得到的上下位关系集合
R_{in}	不同模板抽取到的上下位关系的交集
S	语料库中包含的所有句子集合
s	语料库中的一个句子
P_{syn}	强句法模板集合
P_{sem}	语义模板集合
S_{con}/S_{ent}	已经发现高质量概念/实体集合
$S_{con}^{now}/S_{ent}^{now}$	从 R_{in} 中获取到的概念/实体集合
$S_{con}^{new}/S_{ent}^{new}$	当前这轮迭代中新发现的概念/实体集合

先介绍预备抽取,R 表示上下位关系集合,初始化为空集。R_{ij} 表示在第 i 轮迭代中第 j 个模板产生的上下位关系集合。对于第 j 个强句法模板,扫描语料库并找到与模板匹配的句子,通过 isAExtraction 模块获取上下位关系,加入到 R_{1j} 中。然后将 R_{1j} 合并到 R 中。

算法 1 详细描述了迭代抽取模块。S_{con} 和 S_{ent} 表示已经发现的高质量的概念和实体集合,初始化为空(第 1 行)。R 初始化为预备抽取阶段抽取得到的上下位关系集合(第 2 行)。在每一轮迭代中,对不同模板产生的关系做交集(第 4 行),这避免了单个模板产生的噪声关系,提高了语义模板中用到的概念和实体的质量。接下来计算新的上位词和下位词(第 5 ~ 8 行)并更新 S_{con} 和 S_{ent}(第 9,10 行)。然后使用新概念和新实体构建语义模板(第 11 行)。最后使用语义模板从句子中抽取上下位关系(第 12 ~ 20 行),这过程类似于预备抽取。重复上述步骤,当没有新的实体和概念产生时,终止算法。

算法 1 迭代抽取

输入:S,语料库中的句子

P_{sem},元语义模板

输出:R,上下位关系集合

$S_{con} \leftarrow \varnothing, S_{ent} \leftarrow \varnothing$
$R \leftarrow$预备抽取阶段产生的上下位关系
Repeat
  $R_{in} \leftarrow$对不同模板的上下位关系做交集
  $S_{con}^{now} \leftarrow$从 R_{in} 中获取概念
  $S_{ent}^{now} \leftarrow$从 R_{in} 中获取实体
  $S_{con}^{new} \leftarrow$获取属于 S_{con}^{now} 不属于 S_{con} 的概念
  $S_{ent}^{new} \leftarrow$获取属于 S_{ent}^{now} 不属于 S_{ent} 的实体
  $S_{con} \leftarrow S_{con} \cup S_{con}^{new}$
  $S_{ent} \leftarrow S_{ent} \cup S_{ent}^{new}$
  使用 $S_{con}^{new}, S_{ent}^{new}$ 构建语义模板
  foreach $p \in P_{sem}$ do
    foreach $s \in S$ do
      if s. match(p) then
        $X_{ent}^{s}, X_{con}^{s} \leftarrow$IsAExtraction($s, p$)
        把 hypernym(X_{ent}^{s}, X_{con}^{s}) 加入到 R_{ij}
      end
    end
    $R \leftarrow R \cup R_{ij}$
  end
Until 没有新概念和实体加入到 S_{con} 和 S_{ent}

上面介绍了 isAExtraction 模块,该模块用于从匹配到模板的句子中抽取上下位关系。经过观察,本文把模板分为两类,针对不同类别的模板使用不同的算法。

对于包含动词的模板,使用基于依赖路径(dependency path)的方法。首先要对句子进行依存句法分析,然后通过依赖路径获取上位词和下位词。例如给定一个匹配模板 P_{syn1} 句子“上海是一座城市”,“是”的词性为动词,“上海”和“是”之间是主谓关系,“是”和“城市”之间是动宾关系,通过依赖路径得到 hyponym(上海,城市)。

对于不包含动词的模板,使用基于功能词的方法。模板 P_{syn2} 中“等”就是一个功能词,发现上下位词往往在功能词的前后,可以直接通过正则表达式匹配的方式获取得到。例如给定匹配 P_{syn2} 的句子“中国、印度等国家”,上位词“国家”在功能词之后,下位词“中国”和“印度”在功能词之前,能够获得 hyponym(中国,国家)和 hyponym(印度,国家)。

4　实　验

维基百科是互联网上规模最大,最受欢迎的百科类网站,包含多种语言。为了验证本文提出的方法的有效性和通用性,本文在中文和英文维基百科语料库上进行实验。

4.1　实验一

本实验从中文维基百科语料库中抽取上下位关系。在互联网上下载中文维基百科语料库,它包含 948 835 个网页和 7 911 297 个句子。中文分词、词性标注和依存句法分析由开源中文语言处理平台 LTP[19] 提供。超参数 γ 凭经验设置为 0.85。两个强句法模板(表 1)和三个语义模板(表 2)分别用于预备抽取和迭代抽取。

为了估计使用本文方法抽取的上下位关系的精度,选取了不同领域的 30 个概念作为基准数据集。对于每个概念,随机选取它的 50 个实体或子概念并进行评估。5 名硕士生参与了实验评估,最终通过投票的方式确定最终结果。在其他信息抽取的研究工作也采用了和本文一样的评估方式[11]。图 2 显示了基准数据集上每个概念的上下位关系精度,平均精度为 94.8%,远远大于以前的中文关系抽取方法,如传统的基于模板的方法(78%)[14] 和基于嵌入的方法(约 80%)[17-18]。表 4 显示了基准数据集中的 10 个概念以及它们的典型实例。

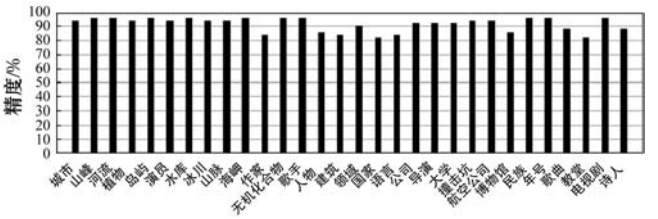

图 2　不同领域的上下位关系精度

表 4　10 个概念及其典型实体

概念	典型实体
城市	西雅图、乌鲁木齐、南昌
山峰	彼得峰、多尔峰、埃克斯波热山
河流	亚马孙河、俄亥俄州河、布哈河
植物	苹果、凤仙花、丁香
演员	许家荣、贾斯汀·亨利、杰森·方特
山脉	乌姆巴山脉、阿尔卑斯山脉、施内山脉
海岬	沃韦格角、波斯特角、拇指角
作家	琼瑶、任祥、穆勒
建筑	金茂大厦、关帝庙、维多利亚剧院
国家	美国、英国、巴西

图 3 显示了每轮迭代上下位关系的累积数量。在预备抽取(第 1 轮)中,抽取得到了 128 215 个上下位关系,这几乎是总关系的三分之一。在迭代抽取中(在第 1 轮之后),曲线在前几轮中快速增长,然后随着 bootstrapping 的过程收敛而饱和。最后,从中文维

基百科语料库中抽取了 327 370 个上下位关系。

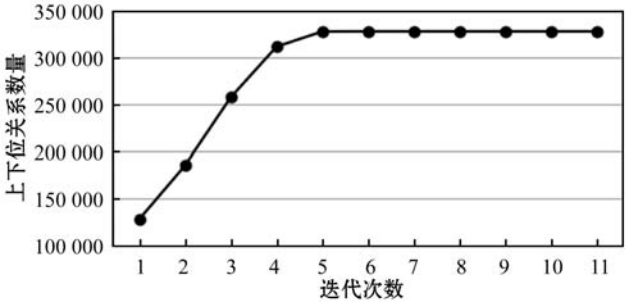

图 3 上下位关系数量随迭代次数的变化

图 4 显示了在每轮迭代在基准数据集上的精度，并且将本文提出的方法与目前最新的基于模板的迭代抽取方法 Probase 进行比较。在第一轮迭代中，本文的方法的精度是 92.4%，略低于 Probase 的 97.3%，因为 Probase 在第一轮迭代中只抽取高置信度的上下位关系。随着迭代的进行，本文方法的精度有所提高，因为在迭代抽取中使用了语义模板并考虑了语义信息，这种现象证明本文方法解决了语义漂移的问题。相反 Probase 的精度有所下降，这是由于错误的上下位关系作为先验知识用于指导下一轮上下位关系的抽取导致的[11]。最后，本文方法的精度超过了 Probase。

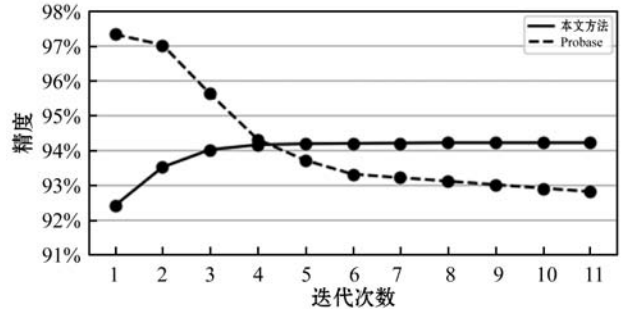

图 4 上下位关系精度随迭代次数的变化

从两个方面评估语义模板的有效性。1）精度：使用相同的评估方法来评估语义模板的精度，精度为 94.7%，和强句法模板的精度相近，远大于弱句法模板的精度；2）召回率：从中文语料库中获得了 320 K 上下位关系，其中约 62% 的上下位关系是由语义模板生成的，这极大地提高了召回率。因此，语义模板可用于获得更高精度的上下位关系。

将本文方法和以下方法进行对比：1）SP：只使用表 1 中的两个强句法模板；2）SP&WP：使用表 1 中的所有句法模板，包括两个强句法模板，两个弱句法模板；3）文献[14]：一个基于模板的迭代抽取方法，该方法没有将句法模板分为更细粒度的模板。

为了评估这些方法，从数据集中随机选择 1 000 个句子来计算精度、召回率和 F1 值，实验结果如表 5 所示。方法 SP 具有最高的精度但召回率低，因为它只使用高质量的模板。本文方法优于 SP 和 WP，因为将本文高质量模板与低质量模板区分开来。文献[14]只使用词汇特征而忽略了句法特征，因此精度低但召回率高。与这些方法相比，本文方法精度和召回率都相对较高，在指标 F1 值上取得了最好的效果。

表 5 评估结果

方法	精度	召回率	F1
SP	100.0%	18.8%	18.8%
SP&WP	81.5%	82.6%	82.0%
文献[14]	72.1%	91.7%	80.7%
本文方法	94.2%	81.2%	87.2%

4.2 实验二

本文框架中用到了句法模板，这些模板在其他语言中也存在，比如英语[5]、韩语[20]。因此，本文提出的方法也可以通过调整阈值 γ 用于其他语言。由于知识有限，只在英语上进行实验。从英语维基百科语料中抽取了 202 846 个和强句法模板匹配的句子，127 727 个和弱句法模板匹配的句子。依存句法分析使用的是斯坦福大学的 CoreNLP 工具。阈值根据经验设置为 0.90。三种强语义模板（表 6 中的前三种模板）和三种元语义模板（表 7）分别用于预备抽取和迭代抽取。

表 6 英文句法模板

模板	精度
NP is a NP	96.2%
such NP as {NP,}* NP	95.6%
NP such as {NP,}* NP	92.0%
NP {,NP}* and other NP	89.3%
NP including {NP,}* NP	81.7%

表 7 英文元语义模板

模板
$CON includes {NP,}* NP
NP {,NP}* and other $CON
$ENT {, $ENT}* and other NP

使用和实验一相同的评估方式，评估得到上下位关系的平均精度是 92.5%，优于之前的信息抽取框架 KnowItAll（平均 64%）[21]、NELL（74%）[7]、TextRunner（平均 80%）[22]，并与目前最新的方法 Probase（92.8%）相近。对召回率进行定性分析，与仅使用句法模板的 Probase 相比，本文方法的召回率高于 Probase，因为充分利用弱句法模板并将它们与实体/概念相结合，以构

建用于迭代抽取的语义模板。从该数据集中总共抽取到 320 199 上下位关系。

5 结　语

本文根据句法模板的质量,将其分成更细粒度的强句法模板和弱句法模板,并将语义信息融入弱句法模板来构建语义模板。基于强句法模版和语义模板提出了一套通用的、有效的上下位关系抽取框架,从文本中抽取上下位关系。从中文维基预料中抽取得到 32 万的上下位关系,精度超过 94%。本文方法具有高精度和高召回率的特点。此外它还可用于其他语言,只需要调整区分强弱句法模板的阈值。在中英文数据上进行了实验,实验结果证明了方法的有效性和通用性。

未来工作方向分为两部分:第一是将本文的框架用在更大规模的语料上进行上下位关系抽取来构建一个大规模高质量的中文分类体系;第二使用更多的弱句法模板,来进一步提高召回率。

参 考 文 献

[1] Wang J, Wang Z, Zhang D, et al. Combining knowledge with deep convolutional neural networks for short text classification[C]//Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017.

[2] 王真, 孙富春, 刘志友. 基于资源本体的 Web 服务发现与组合研究[J]. 计算机应用与软件, 2012, 29(3):191-194,244.

[3] Faruqui M, Dodge J, Jauhar S K, et al. Retrofitting word vectors to semantic lexicons[C]//Proceedings of NAACL 2015.

[4] Miller G A. WordNet: a lexical database for English[J]. Communications of the Acm, 1995, 38(11):39-41.

[5] Hearst M A. Automatic acquisition of hyponyms from large text corpora[C]//Proceedings of the 14th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1992: 539-545.

[6] Brin S. Extracting patterns and relations from the World Wide Web[M]//The World Wide Web and Databases. Springer Berlin Heidelberg, 1998:172-183.

[7] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning[C]//Twenty-Fourth AAAI Conference on Artificial Intelligence. AAAI Press, 2010:1306-1313.

[8] Lu W, Lou R, Dai H, et al. Taxonomy Induction from Chinese encyclopedias by combinatorial optimization[M]//Natural Language Processing and Chinese Computing. Springer International Publishing, 2015:299-312.

[9] Ponzetto S P, Strube M. WikiTaxonomy: A large scale knowledge resource[C]//Proceedings of the 18th European Conference on Artificial Intelligence, Patras, Greece, 21-25 July, 2008:751-752.

[10] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 697-706.

[11] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding[C]//Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012: 481-492.

[12] Evans R, Street S. A framework for named entity recognition in the open domain[J]. Recent Advances in Natural Language Processing III: Selected Papers from RANLP, 2003, 260(267-274): 110.

[13] Snow R. Learning syntactic patterns for automatic hypernym discovery[C]//Proceedings of the Neural Information Processing Systems, 2004:1297-1304.

[14] Liu L, Zhang S, Diao L, et al. An iterative method of extracting Chinese ISA relations for ontology learning[J]. Journal of Computers, 2010, 5(6):870-877.

[15] Li J, Wang C, He X, et al. User generated content oriented chinese taxonomy construction[C]//Asia-Pacific Web Conference. Springer, Cham, 2015: 623-634.

[16] Fu R, Guo J, Qin B, et al. Learning semantic hierarchies via word embeddings[C]//Meeting of the Association for Computational Linguistics. 2008:1199-1209.

[17] Wang C, Yan J, Zhou A, et al. Transductive non-linear learning for Chinese hypernym prediction[C]//Meeting of the Association for Computational Linguistics. 2017: 1394-1404.

[18] Che W, Li Z, Liu T. LTP: a Chinese language technology platform[C]//International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2010:13-16.

[19] Shin H J, Cho S Z. Pattern selection using the bias and variance of ensemble[J]. Journal of the Korean Institute of Industrial Engineers, 2002, 28(1):112-127.

[20] 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.

[21] Yates A, Cafarella M, Banko M, et al. TextRunner: open information extraction on the web[C]//Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2007:25-26.