结合词向量和Bootstrapping的领域实体上下位关系获取与组织

马晓军[1] 郭剑毅[1,2] 线岩团[1,2] 毛存礼[1,2] 严 馨[1,2] 余正涛[1,2]
(昆明理工大学信息工程与自动化学院 昆明 650500)[1]
(昆明理工大学智能信息处理重点实验室 昆明 650500)[2]

摘 要 实体上下位关系是构建领域知识图谱不可或缺的一种重要的语义关系,传统抽取上下位关系的方法大多不考虑关系的组织。提出一种结合词向量和Bootstrapping的方法来实现领域实体上下位关系的获取与组织。首先,选取旅游领域的种子语料集;然后,采用基于词向量的相似度计算方法对种子集中包含的上下位关系模式进行聚类,筛选出置信度高的模式并对未标注语料进行上下位关系识别,得到候选关系实例,同时选择置信度高的关系实例加入到种子集中,进行下一轮的迭代,直到得到所有的关系实例;最后,根据领域实体上下位关系对的向量偏移并结合领域实体层级关系的特点,采用映射的学习方法进行领域实体层级关系组织。实验结果表明,与传统的方法相比,所提方法的F值提高了近10%。

关键词 上下位关系,关系抽取,Bootstrapping方法,词向量,映射学习,层级关系组织
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.01.010

Entity Hyponymy Acquisition and Organization Combining Word Embedding and Bootstrapping in Special Domain

MA Xiao-jun[1] GUO Jian-yi[1,2] XIAN Yan-tuan[1,2] MAO Cun-li[1,2] YAN Xin[1,2] YU Zheng-tao[1,2]
(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)[1]
(The Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming 650500, China)[2]

Abstract The semantic relation of entity hypomypy is important to build the domain knowledge graphs. The organization of hierarchical relations is not considered in the traditional method of extracting hyponymy. A method of extracting and organizing the entity hyponymy in the specific field was proposed in this paper, which combines the word embedding and bootstrapping method. Firstly, the tourism corpus is selected as seed corpus, then the hyponymy patterns included in the seed corpus are clustered based on the method of word embedding similarity. Thus, the patterns of high-confidence level are filtrated which is used to identify hyponymy in the unlabeled corpus. After that, the high-confidence instances of relation are obtained which are selected to put in the seed sets. And the next iteration is performed until all the instances of relation are obtained. Finally, the mapping learning methods are applied to conduct the hierarchical relation of domain entity based on the character of the entity of domain hierarchical relations and the vector-deviation of the hyponymy pairs of the entity. The experimental results show that the proposed method improves the F-value by 10% compared with the traditional method.

Keywords Hyponymy relation, Relation extraction, Bootstrapping method, Word embedding, Projection learning, Hierarchical relation organization

实体上下位关系是构建领域知识图谱不可或缺的一种重要的语义关系。上下位关系的定义一般采用Miller[1]的定义,即若X是Y的实例或子集,实体X是Y的下位词,Y是X的上位词,则X和Y之间具有上下位关系。目前针对上下位关系的研究主要有以下几类方法:

(1)基于模式匹配或规则的方法。该方法以Hearst[3]的研究为代表,主要根据特定语言的使用习惯,将人工设置的多种匹配模式用在大规模语料中以获取上下位关系;早期的上

到稿日期:2017-03-03 返修日期:2017-06-16 本文受国家自然科学基金(61562052,61363044,61472168)资助。
马晓军(1991—),男,硕士生,主要研究方向为信息抽取;**郭剑毅**(1964—),女,硕士,教授,主要研究方向为自然语言处理,E-mail:giade86@hotmail.com(通信作者);**线岩团**(1981—),男,博士生,讲师,主要研究方向为机器学习;**毛存礼**(1977—),男,博士,副教授,主要研究方向为自然语言处理;**严 馨**(1968—),女,硕士,副教授,主要研究方向为自然语言处理;**余正涛**(1970—),男,博士,教授,主要研究方向为自然语言处理和机器翻译。

下位关系抽取主要用这类方法；Mann[4] 和 Fleischman[5] 等使用部分的语言模式来抽取专有名词的下位词子集；Ando 等[6] 借鉴 Hearst 等人的思想使用句法模式从日语语料中抽取上下位关系，并将结果与关联概念词典进行比较，其准确率达到了 63%。在中文方面，刘磊[7] 提出了基于“是一个”模式的下位实体获取方法，利用半自动获取的词典和句型对“是一个”模式进行分析，然后根据不同的规则，分别获取下位实体。这种方法的准确率较高，但需要人工确定模式。

(2)基于词典和知识库的方法。该方法主要根据一些现有人工构建的词汇词典或者在线百科等知识库中含有的同义、反义等语义信息来获取实体之间的上下位关系，如 Nakaya 等[8] 使用 WordNet 来获取英文实体间的分类关系；Sumida 等[9] 借助英文版维基百科进行上下位关系抽取，主要对维基百科中文章内容的分层结构进行分析，然后抽取下位词；Suchanek[10] 使用维基百科中的分类信息对语义关系进行抽取；在中文方面，董振东编写了一部通用领域词典 HowNet，在词典中用义原树来描述词汇之间的关系；范庆虎[11] 采用基于《中文概念词典》、百科资源和百度相关搜索的方法进行下位词的发现，在召回率和准确率上比单一的百科效果更好。这种方法的准确率较高，但需要领域词典作为基础。

(3)基于机器学习的方法。这种方法的主要思想是将上下位关系抽取转化为一个分类问题，按照有无标注好的训练语料可以分为有监督、无监督和半监督的方法。其中，有监督的方法需要人工标注大量语料，费时费力，但效果好，如Caraballo等[12] 利用连接词和同位语获取名词，通过上下文中名词的连接关系或同位关系构造名词特征向量，通过聚类得到名词间的上下位关系；Boella 等[13] 运用依存句法构建语义模型，通过 SVM 进行分类来抽取上下位关系。无监督的方法不需要人工进行语料标注，但是现阶段其准确率不高，性能较差，如 Etzioni 等[14] 采用开放的信息抽取技术从海量数据中抽取关系信息，通过聚类算法获取关系类型。半监督的机器学习由于只需要标注少量语料，就可以达到理想的结果，因此成为了目前主流的方法。基于半监督的自扩展模式生成方法能够兼顾精度和性能要求，以少量人工标注的上下位关系实体对为种子，采用迭代的方式从语料中获取关系模式，利用模式再获取新的关系实例。Fang[15] 和 Kozareva[16] 等采用半监督的方法从 Web 中自动提取给定词的下位词，通过自举(Bootstrapping)的思想来抽取上下位关系；半监督的机器学习方法不仅对语料的依赖程度低，而且可以获取较好的准确率。

另外，传统基于模式和语义词典的方法很难用于计算两个词语的相似度，获取到的大量实体对往往在语义上并不相似，如以下例句：

句子 1 “荨麻，别名蚕人草、咬人草，荨麻科植物”。

句子 2 “金雕，俗称为鹫雕、金鹫、黑翅雕、洁白雕(幼鸟)等，是一种性情凶猛、体态雄伟的猛禽”。

采用相同模式可从句子 1 获取上下位实体对(荨麻，植物)，从句子 2 获取上下位实体对(金雕，猛禽)。虽然这两个句子实例的模式相似，但其在语义上不相似，一种属于植物的实体对，一种属于动物的实体对。如何将语义或语法相似度相近的上下位实体对聚类在一起，是上下位关系获取与组织研究的难点和重点。而借助词向量的方法，通过计算实体间的语法和语义的线性相似度，可以解决上述问题。实体的语法和语义的线性相似度，是指实体的语法相似度和语义相似度可以进行近似的线性计算。下面给出语义相似度的示例：

$$X_{\text{king}} - X_{\text{man}} \approx X_{\text{queen}} - X_{\text{woman}}$$

其中，X_{king} 表示词 king 的词向量。词向量的方法是通过训练将每个词映射成 K 维的实数向量，通过计算 cos 相似度、欧氏距离来判断词与词之间的相似度，最后通过聚类方法得到具有相同语义或语法相似度相近的上下位实体对。因此，利用词向量的方法可以很好地描述语义相似度信息。

先通过 Bootstrapping 机器学习得到分类效果较好的分类器，同时利用词向量技术对旅游领域的语料进行词向量模型训练，基于词向量计算相似度并对上下位关系进行模式聚类；然后对未标注的语料进行上下位关系识别；最后采用映射矩阵的方法进行领域实体上下位关系的自动组织。

1 领域实体上下位关系获取的原理和框架

本文所提方法的框图如图 1 所示。该方法的层次结构主要包括 5 个部分：词向量模型训练、种子集获取、抽取模式的生成、候选关系实例的获取和筛选以及映射学习。

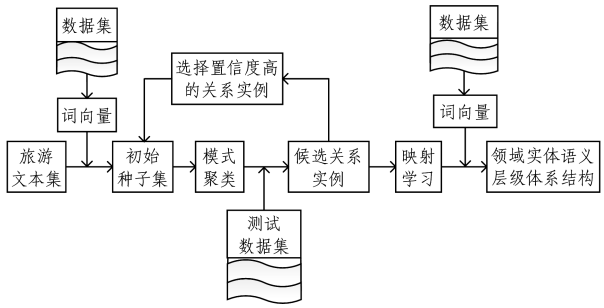

图 1 领域实体上下位关系的原理框图

Fig. 1 Principle framework of domain entity hyponymy relation

1.1 词向量模型训练

Mikolov[17] 于 2013 年提出了两个最著名的对数线性模型：Skip-gram 和 CBOW 模型。这两个模型都是用来对大规模语料库进行向量转化的。在识别词语之间的语义关系方面，Skip-gram 模型有着更好的效果。我们利用网络爬虫在互联网上爬取了大量的旅游领域数据，并将其作为词向量模型的训练语料，经过预处理后的语料规模约有 700M；然后选择 Google 的开源工具包 word2vec[18]，采用 Skip-gram 模型进行词向量模型训练，窗口大小为 5，向量维数为 200。

1.2 种子集的获取

首先对文本进行预处理；然后以句子为单位对文本进行切分，并进行人工的实体标注；最后对处理过的文档进行扫描，筛选出同时含有两个以上领域实体的句子。选取第一个实体前的词(BEF)、两个实体间的词(BET)和第二个实体后的词(AFT)作为特征上下文。

对于 BEF 上下文，采用浅层启发式的方法进行关系模式

的选取。这种模式只考虑动词的介导关系,若两个实体间不存在动词,则选择两个实体间所有的词作为 BEF 上下文。

每一种上下文文本在去除停用词和形容词后,剩余的每个词都转化为单独的词向量,然后进行简单的组合得到特征向量。这种向量组合的方式的作用在 Mikolov 等人于 2013 年的研究中得到了证明。

本文采用 3 个向量的组合来表示任意关系实例 i,即 V_{BEF},V_{BET} 和 V_{AFT}。

例如下面的句子:

云南盛产的野生菌包括松茸菌、牛肝菌等。

可以抽取关系实例:

$V_{BEF}=E(\text{“云南”})+E(\text{“盛产”})$

$V_{BET}=E(\text{“包括”})$

$V_{AFT}=E(\text{“牛肝菌”})$

其中,$E(x)$是词 x 的向量形式。

1.3　抽取模式的生成

使用上文获得的种子关系实例,采用 Single-pass 聚类的方法生成抽取模式。每个聚类结果都包含一类层级关系实例集,并且以关系实例的 3 种上下文向量表示。

聚类算法的描述如算法 1 所示,算法的输入是种子关系实例的列表,输出是关系模式集,并且定义第一个实例属于第一个新的空簇。然后,遍历种子实例列表,计算任意种子实例 i_n 与每个聚类簇 Cl_j 的相似度。种子实例集 i_n 与聚类簇 Cl_j 之间的相似度计算公式为 $Sim(i_n,Cl_j)$,实例间的相似度计算公式如式(1)所示:

$$\begin{aligned}Sim&=(S_n,S_j)\\&=\alpha\cdot\cos(BEF_i,BEF_j)+\beta\cdot\cos(BET_i,BET_j)+\\&\quad\gamma\cdot\cos(AFT_i,AFT_j)\end{aligned}\tag{1}$$

其中,参数 α,β 和 γ 是向量的权重。具体的 Single-pass[19] 聚类算法如算法 1 所示。

算法 1　Single-pass 聚类算法

输入:种子实例 Instances=$\{i_1,i_2,\cdots,i_n\}$

输出:关系模式 Patterns=$\{\}$

　　$Cl_1=\{i_1\}$

　　Patterns=$\{Cl_1\}$

For $i_n\in$ Instances do

For $Cl_j\in$ Pattrens do

If $Sim(i_n,Cl_j)\geqslant T_{sim}$ then

　$Cl_j=Cl_j\cup\{i_n\}$

else $Cl_j=\{i_n\}$

　Patterns=Patterns$\cup Cl_m$

为了防止错误模式被加入到模式集中,本文采用打分的方式进行模式的筛选。根据模式抽取的关系实例对模式进行打分,每个模式抽取的实例包括 3 个类别:P(Positive),N(Negative)和 U(Unknown)。若得到的关系实例中的两个实体的关系与种子集中的相同,则定义该关系实例是积极的,将其加入到集合 P;若该关系实例间实体对的关系与种子集的某个关系实例的实体对关系相矛盾,则定义该关系实例是消极的,将其加入到集合 N;若该关系实例间实体对的关系不是种子集中的一部分,则定义该关系实例是未知的,将其加入到集合 U。模式 P 的置信度计算公式如式(2)所示:

$$Conf_p(p)=\frac{|P|}{|P|+W_n\cdot|N|+W_u\cdot|U|}\tag{2}$$

其中,W_n 和 W_u 分别是消极的和未知的关系实例对应的权重。关系实例的置信度是根据该关系实例与抽取它的模式之间的相似度分数来计算的。模式的置信度权重计算公式如式(3)所示:

$$Conf(i)=1-\prod_{j=0}^{|\zeta|}(1-Conf_p(\zeta_j)\times Sim(C_i,\zeta_j))\tag{3}$$

其中,ζ 是抽取关系实例 i 的模式,C_i 是关系实例 i 对应的上下文。每个实例对应的置信度阈值 T_i 在下一次迭代中将被作为种子。

1.4　候选关系实例的获取

在得到抽取模式后,采用算法 2 进行候选关系实例的获取。再次扫描未标注的文档,得到所有与种子集中的关系实例的语义类型相同的段落。对于每一个段落,关系实例 i 的生成过程见 1.2 节,其与所有之前生成的抽取模式的相似度都可以计算出来。如果关系实例 i 和某个模式 Cl_j 的相似度大于或等于阈值 T_{sim},那么关系实例 i 就被认为是一个候选实例,而且模式 Cl_j 的置信度将会被更新。将与关系实例 i 相似度最高的模式记为 P_{best},相对应的相似度分数记为 Sim_{best},并且将这些数据作为中间的输出保存到 Results 中。采用 Bootstrapping 的方式进行迭代,不断更新 Results 和 P_{best}。每次迭代后,根据式(1)的分数对得到的候选关系实例进行排序。若关系实例的得分大于或等于阈值 T_i,则将这个关系实例加入到对应的种子集中,作为 Bootstrapping 算法下一次迭代的依据。经过所有的迭代和筛选得到最终的候选关系实例集,候选关系实例获取如算法 2 所示。

算法 2　新关系实例获取

输入:候选句子集 Sentence S=$\{S_1,S_2,\cdots,S_n\}$,关系模式集 Pattern P=$\{Cl_1,Cl_2,\cdots,Cl_n\}$

输出:候选关系实例集 Results

　For $S_i\in$ Sentences do

　　i=instances(S_i)

　　$Sim_{best}=0$

　　P_{best}=null

　For $Cl_j\in$ patterns do

　If $Sim(i,Cl_i)\geqslant T_{sim}$ then

　　Conf(Cl_i)

　If $Sim(i,Cl_i)\geqslant Sim_{best}$

　　$Sim_{best}=Sim$

　$P_{best}=Cl_i$

1.5　映射学习和层级关系组织

1.5.1　映射学习

Mikolov 等人于 2013 年的研究证明,经过词向量模型训练的词语间存在着一些语言学的规律,包含着大量的句法和语义关系。例如,$v(\text{king})-v(\text{queen})\approx v(\text{man})-v(\text{woman})$,这表明词向量偏移确实可以表征两个词对之间的语义关系。Fu 等人[20] 的研究发现,这些特性同样适用于一些上下位关系。我们在旅游领域也做了一些简单的实验,随机地选择一些具

有上下位关系的实体对，然后计算它们之间的嵌入偏移，并以此估算它们之间的相似性，实验结果如表 1 所列。

表 1 上下位词对间的词向量偏移

Table 1 Word embedding offsets on a sample of hyponymy word-pairs

序号	Examples
1	V(古城)−V(大理古城)≈V(雪山)−V(玉龙雪山)
2	V(茶叶)−V(普洱茶)≈V(野生菌)−V(鸡枞)
3	V(古城)−V(大理古城)≠V(野生菌)−V(鸡枞)

由表 1 中的前两组例子可以看出，实体对之间的上下位关系也可以通过嵌入偏移来表示。然而，在第三组实验中，“古城”到“大理古城”和“野生菌”到“鸡枞”之间的嵌入偏移差别还是比较大的，这也表明实体词对之间的上下位关系很难用简单的矢量偏差来精确表示。通过实验验证发现，不同类型的上下位关系在空间中的分布差别很大，为了解决这个问题，我们采用映射矩阵的方法来识别实体词之间的上下位关系。

映射学习的思想是，假设所有的词都可以通过过渡矩阵映射到它们的上位词。根据前面的结论，将训练数据中所有的上下位关系实体对(x,y)表示为它们的向量偏移：$y-x$，然后进行聚类，则聚类后每个聚类簇中的实体对有相似的上下位关系类型。本文采用 K-means 聚类的方法进行聚类，原始数据集为 $S=\{S_1,S_2,\cdots,S_i\}$，其中，S_i 是实体对(x,y)的向量偏移：$y-x$，且$S_i\in\mathbb{R}^n$，聚成 k 个簇，具体过程如下。

(1)从数据集 S 中随机选择 k 个聚类质心点 $S_1,S_2,\cdots,S_k\in\mathbb{R}^n$；

(2)分别计算剩下的元素到 k 个聚类簇的相异度，使 $\arg\min_m\sum_{i=1}^{k}\sum_{s_j\in m_i}\|s_j-u_i\|^2$ 的值最小，其中 u_i 表示分类簇 C_i 的平均值；

(3)根据聚类结果重新计算 k 个聚类簇的质心；

(4)重复步骤(2)、步骤(3)，直到聚类结果不再变化，输出结果。

例如，当给定实体 x 和它的上位词 y 时，存在矩阵 Φ_k 使得 $y=\Phi_k x$。由于为所有的上下位关系对间的映射都计算过渡矩阵 Φ_k 是困难的，因此采用式(4)计算 Φ 的近似值，本文采用梯度下降算法进行最优化求解。

$$\Phi_k^*=\arg\min_{\Phi_k}\frac{1}{N_k}\sum_{(x,y)\in C_k}\|\Phi_k x-y\|^2 \tag{4}$$

其中，N_k 是聚类簇集 C_k 第 k 个聚类簇中实体对的数量。

为了学习映射矩阵，人工构建了小规模的领域知识库作为映射矩阵的训练数据。本文在深入分析领域属性及行业属性的基础上，人工定义了领域知识体系，收集了领域相关概念种子集合，并利用网络百科[21]的资源构建了小规模的领域知识库。

通过分析发现，互动百科的分类体系相对比较规范，因此本文采用互动百科的分类体系作为基础，构建出包含 10000 个领域实体的旅游领域知识库。部分的旅游领域知识库的语义层次结构图如图 2 所示。

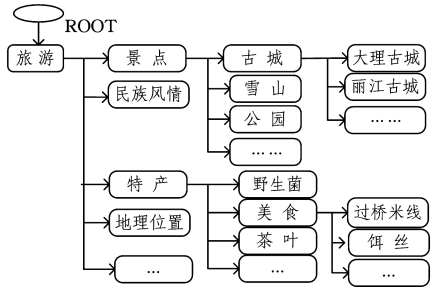

图 2 部分旅游领域知识库的语义层次结构图

Fig. 2 Semantic hierarchical structure diagram of knowledge base in part of tourism fields

1.5.2 层级关系组织

实体上下位关系组织的主要任务就是给定某个实体及其上位词列表，并自动地构建它们的语义层级结构，如图 3 所示。可以把实体语义层级结构看作是一个有向图 G，节点代表实体，边代表实体间的层级关系，并且这种层级关系具有不对称性和传递性，如式(5)所示：

$$\begin{aligned}&\forall x,y\in L:x\xrightarrow{H}y\Rightarrow\neg(y\xrightarrow{H}x)\\&\forall x,y,z\in L:(x\xrightarrow{H}z\wedge z\xrightarrow{H}y)\Rightarrow x\xrightarrow{H}y\end{aligned} \tag{5}$$

其中，L 表示某个实体的上下位词列表，x,y,z 是列表 L 中的上位词，$\xrightarrow{H}$代表上下位关系。x,y,z 是某个实体的上位词，因此 G 是一个有向无环图。

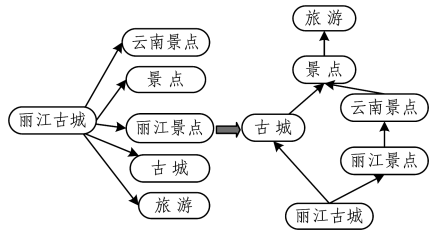

图 3 领域实体的语义层次体系构建示例

Fig. 3 An example of constructing semantic hierarchy system in tourism field

通过对训练数据的聚类和相应的映射训练，可以判断给定的两个实体是否存在上下位关系。当给定两个实体 x 和 y，根据实体的词向量偏移：$y-x$，可以在聚类簇 C_k 中找到与该偏差最近的聚类簇中心，而且还可以得到映射的过渡矩阵 Φ_k。若满足条件 $\Phi_k x$ 和 y 之间的欧氏距离 $d(\Phi_k x,y)$必须小于某个阈值 δ，则认为 y 是 x 的上位词，如式(6)所示：

$$d(\Phi_k x,y)=\|\Phi_k x-y\|^2<\delta \tag{6}$$

最终得到的结果就是如图 3 所示的一个有向无环图。

2 实验设计与结果分析

2.1 实验语料及评测指标

本文的实验语料是从旅游网站和百科词条上爬取的旅游领域文本，共有 30000 多篇。采用人工标注的方式选择 200 条关系实例并将其作为种子集，将剩余的语料经过预处理后作为未标注的语料集进行自扩展的迭代。映射学习的训练数据是人工构建的旅游领域知识库，抽取大约 1780 组实体上下位关系对作为训练语料进行映射矩阵参数训练。

为了对领域实体关系的自动组织性能进行评价,随机地选择 210 个领域实体和它们的上位词进行人工标注。然后将标注的数据分为 10 份,其中 8 份作为训练数据,2 份作为测试数据,采用交叉验证的方法进行实验。

2.2　实验相关参数设置

实验的过程包括词向量训练、语料预处理、种子集获取、Bootstrapping 迭代、候选实例获取和映射矩阵学习等。词向量训练采用 Google 的开源工具包 word2vec 实现,窗口大小设置为 5,向量维度为 200 维。语料的预处理过程采用开源的工具包 Ansj 完成,包括分词、词性标注、去停用词和命名实体识别等过程。选择的特征包括 BEF,BET 和 AFT 上下文。种子集的获取由人工标注完成,共包含 200 个种子实例,并将其作为算法的输入。将参数 W_n 和 W_u 分别设置为 0.1 和 2,阈值 T_{sim} 和 T_t 的取值范围是[0.5,1],Bootstrapping 的迭代次数为 4。本文采用最常用的准确率 P、召回率 R 和查全率 F 作为评测标准,计算公式如下:

$$准确率(P)=\frac{正确抽取的关系实例总数}{抽取的所有关系实例总数}\times 100\%$$

$$召回率(R)=\frac{正确抽取的关系实例总数}{标准结果中的关系实例总数}\times 100\%$$

$$查全率(F)=\frac{2PR}{P+R}\times 100\%$$

2.3　实验设计

为了验证本文所提方法的可行性,本文设置了以下 4 组实验:

实验 1　通过设置不同的权重参数 α,β 和 γ 来验证 BEF,BET 和 AFT 3 类特征对领域实体上下位关系抽取性能的影响。

实验 2　比较本文的方法、Snowball 算法和传统的机器学习分类算法对领域实体上下位关系的抽取性能。

实验 3　验证领域知识库对领域上下位关系组织的影响。

实验 4　本文提出的方法与基于规则的方法、基于 CRF 的方法在领域实体上下位体系构建过程中的效果比较。

2.4　实验结果与分析

实验 1　为了验证 3 种特征对领域实体上下位关系抽取性能的影响,本文分别选取两种不同的权重参数,如 Conf1 和 Conf2 所示:

Conf1:$\alpha=0.1,\beta=0.8,\gamma=0.1$

Conf2:$\alpha=0.2,\beta=0.6,\gamma=0.2$

其中,Conf1 只包含两个实体间词(BET)的上下文,Conf2 包含所有的 3 类特征上下文信息。这里的准确率、召回率和查全率是 TOP5 模式下的平均值。实验结果如表 2 所列。

表 2　不同特征对领域实体上下位关系抽取性能的影响/%

Table 2　Influence of different features on the performance of hyponym relation extraction between domain entities/%

参数	P	R	F
Conf1	85.8	70.2	77.2
Conf2	79.4	63.5	70.6

由表 2 的实验数据可知,对于大多数类型的上下位关系模式,Conf2 参数设置所取得的召回率都要低于 Conf1 参数。通过对实验结果进行分析发现,其主要的原因在于 BEF 和 AFT 的上下文数据太过稀疏,包含了很多对实体对间关系没有贡献的词语。实验结果表明,实体对间的上下文词语对实体对的上下位关系识别有着更重要的作用。

实验 2　为了验证本文提出的方法的可行性,在相同的实验数据集上进行对比实验。选择模式聚类结果的 TOP5 进行实验,实验结果如表 3 所列。

表 3　不同上下位关系抽取方法的比较/%

Table 3　Comparison of different extraction methods of hyponymy/%

方法	模式	P	R	F
Snowball	Top1	97.2	75.4	84.9
	Top2	79.8	72.3	75.9
	Top3	77.1	64.9	70.5
	Top4	73.3	42.4	53.7
	Top5	67.5	56.4	61.5
本文方法	Top1	96.9	84.8	90.4
	Top2	85.6	78.4	81.8
	Top3	81.3	72.6	76.7
	Top4	83.1	64.5	72.6
	Top5	67.9	57.2	62.1

由表 3 的实验数据可知,对于模式聚类结果的 TOP5,与 Snowball 算法相比,本文提出的方法都取得了比较好的 F 值。而且对于某些关系模式,甚至取得了比 Snowball 高出 20%的 F 值。实验结果表明,使用词向量模型可以表示语言学的语义特征,提高了实体上下位关系的效果。

实验 3　为了验证领域知识库对层级关系体系构建的影响,实验分别在加入领域知识库和不加人领域知识库的两种情况下进行,实验结果如表 4 所列。

表 4　领域知识库对领域实体上下位关系组织的影响/%

Table 4　Influence of domain knowledge base on the hyponymy relation organization/%

方法	P	R	F
词向量	75.3	67.5	69.3
词向量+知识库	78.3	79.8	79.0

由表 4 可知,在加入领域知识库进行约束的情况下,方法的召回率有了很大的提高,证明领域知识库对领域实体的上下位关系组织有着重要作用。

实验 4　为了验证本文方法的可行性,将所提方法与基于规则的方法、基于 CRF 的方法进行比较,实验结果如表 5 所列。

表 5　领域实体上下位关系的识别结果/%

Table 5　Identified result of hyponymy relation in domain entity/%

方法	P	R	F
基于规则的方法	84.4	48.9	61.9
基于 CRF 的方法	75.1	72.4	73.7
本文方法	78.2	79.8	79.0

由表 5 可知,与基于规则的方法相比,本文提出的基于词向量的方法虽然在准确率上稍低,但在召回率上远远超过基于规则的方法。而与基于层叠条件随机场的方法相比,本文提出的方法在准确率和召回率上都有所提高。实验结果表明了本文提出的方法在领域实体层级体系自动构建任务中的可行性。

结束语 本文提出了一种新的结合词向量和 Bootstrapping 的领域实体上下位关系获取方法，包括语料预处理、词向量运算、初始种子集选择、模式抽取、Bootstrapping 迭代、候选关系实例的获取以及领域实体层级关系自动组织。从实验结果可以看出，本文所提方法在领域实体上下位关系抽取和层级关系组织方面都取得了不错的效果，证明了该方法的可行性。目前本文构建的旅游领域知识库还不够完善，接下来将继续完善领域知识库，将该方法移植到其他特定领域以进行分析对比。

参 考 文 献

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

[2] SHEN D R. SKM:A Schema Matching Model Based on Schema Structure and Known Matching Knowledge[J]. Journal of Software,2009,20(2):327-338.

[3] HEARST,MARTI A. Automatic acquisition of hyponyms from large text corpora[C]// Conference on Computational Linguistics. 1992:539-545.

[4] MANN G S. Fine-grained proper noun ontologies for question answering[C]// The Workshop on Building & Using Semantic Networks. Association for Computational Linguistics,2003.

[5] FLEISCHMAN M,HOVY E,ECHIHABI A,et al. Offline strategies for online question answering: answering questions before they are asked[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics,2003:1-7.

[6] ANDO M,SEKINGE S,ISHIZAKI S. Automatic Extraction of Hyponyms from Newpaper Using Lexicosyntactic Pattern[J]. Ipsj Sig Notes,2004,2003:77-82.

[7] LIU L,CAO C G,WGNG H T,et al. A Method of Hyponym Acquisition Based on" isa" Pattern[J]. Computer Science,2006,33(9):146-151. (in Chinese)
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.

[8] NAKAYA N,KUREMATSU M,YAMAGUCHI T. A Domain Ontology Development Environment Using a MRD and Text Corpus[J]. Casopís Lékar Ceskych,2002,128(37):1166-1169.

[9] SUMIDA A,TORISAWA K. Hacking Wikipedia for hyponymy relations acquisition [C] // International Joint Conference on Natural Language Processing. 2008.

[10] SUCHANEK FM,KASNECI G,WEIKUM G. Yago:A core of semantic knowledge unifying wordnet and wikipedia[C]// Proceedings of the Third International Joint Conference on Natural Language Processing. 2008:883-888.

[11] FAN Q H,ZAN H Y,CHAI Y M,et al. hyponym discovery of multiple resource fusion[J]. Computer Engineering and Design,2013,34(12):4310-4315. (in Chinese)
范庆虎，昝红英，柴玉梅，等. 多资源融合的下位词发现[J]. 计算机工程与设计，2013,34(12):4310-4315.

[12] CARABALLO S A. Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text[C]// Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. 1999:120-126.

[13] BOELLA G,CARO L D. Extracting Definitions and Hypernym Relations Relying on Syntactic Dependencies and Support Vector Machines[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013:532-537.

[14] ETZIONI O,BANKO M,SODERLAND S,et al. Open information extraction from the web[C]// International Joint Conference on Artifical Intelligence. Morgan Kaufmann Publishers Inc.,2007:68-74.

[15] FANG N M,NON-MEMBER C Y,MEMBER F R. Hyponym extraction from the web by bootstrapping[J]. IEEJ Transactions on Electrical & Electronic Engineering,2012,7(7):62-68.

[16] KOZAREVA Z,HOVY E. A semi-supervised method to learn and construct taxonomies using the web[C]// Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2010:1110-1118.

[17] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science,2013,23(1):1301-1306.

[18] GOLDBERG Y,LEVY O. word2vec Explained:deriving Mikolov et al.'s negative-sampling word-embedding method [J/OL]. https://arxiv.org/abs/1402.3722.

[19] BENNETT J,GROUT R,PEBAY P,et al. Numerically stable, single—pass, parallel statistics algorithms[C]// IEEE International Conference on Cluster Computing and Workshops. 2009:1-8.

[20] FU R J,QIN B,LIU T. Exploiting multiple sources for open-domain hypernym discover[C]// EMNLP. 2013:1224-1234.

[21] WANG P,HU J,ZENG H J,et al. Improving Text Classification by Using Encyclopedia Knowledge[C]// IEEE International Conference on Data Mining. IEEE,2007:332-341.

(上接第 54 页)

[27] YE F,LI Y N. Group multi-attribute decision model to partner selection in the formation of virtual enterprise under incomplete information[J]. Expert Systems with Applications,2009,36(5):9350-9357.

[28] XU Z S. Group decision making based on multiple types of linguistic preference relations[J]. Information Sciences,2008,178(2):452-467.

[29] XU Z S. A multi-attribute group decision making method based on term indices in linguistic evaluation scales[J]. Journal of Systems and Engineering,2005,20(1):84-88. (in Chinese)
徐泽水. 基于语言标度中术语指标的多属性群决策法[J]. 系统工程学报，2005,20(1):84-88.

[30] PANG J F,SONG P. A research on evaluation and selection of materials suppliers for large coal enterprises[J]. East China Economic Management,2015,29(2):117-122. (in Chinese)
庞继芳，宋鹏. 面向大型煤炭企业的物资供应商评价与选择研究[J]. 华东经济管理，2015,29(2):117-122.