第 41 卷 第 2 期 计 算 机 工 程 2015 年 2 月
Vol. 41 No. 2 Computer Engineering February 2015

·人工智能及识别技术· 文章编号：1000-3428(2015)02-0145-06 文献标识码：A 中图分类号：TP18

结合语境与布朗聚类特征的上下位关系验证

张志昌，陈松毅，刘 鑫，马慧芳

（西北师范大学计算机科学与工程学院，兰州 730070）

摘 要： 对海量文本语料进行上下位语义关系自动抽取是自然语言处理的重要内容，利用简单模式匹配方法抽取得到候选上下位关系后，对其进行验证过滤是难点问题。为此，分别通过对词汇语境相似度与布朗聚类相似度计算，提出一种结合语境相似度和布朗聚类相似度特征对候选下位词集合进行聚类的上下位关系验证方法。通过对少量已标注训练语料的语境相似度和布朗聚类相似度进行计算，得到验证模型和 2 种相似度的结合权重系数。该方法无需借助现有的词汇关系词典和知识库，可对上下位关系抽取结果进行有效过滤。在 CCF NLP&2012 词汇语义关系评测语料上进行实验，结果表明，与模式匹配和上下文比较等方法相比，该方法可使 F 值指标得到明显提升。

关键词： 上下位关系；语境相似度；布朗聚类相似度；点互信息；模式匹配；聚类验证

中文引用格式： 张志昌，陈松毅，刘 鑫，等. 结合语境与布朗聚类特征的上下位关系验证[J]. 计算机工程，2015，41(2)：145-150.

英文引用格式： Zhang Zhichang，Chen Songyi，Liu Xin，et al. Hyponymy Relation Validation Combined with Context and Brown Clustering Feature[J]. Computer Engineering，2015，41(2)：145-150.

Hyponymy Relation Validation Combined with Context and Brown Clustering Feature

ZHANG Zhichang，CHEN Songyi，LIU Xin，MA Huifang

（School of Computer Science and Engineering，Northwest Normal University，Lanzhou 730070，China）

【Abstract】 Hyponymy has many important applications in the field of Natural Language Processing（NLP） and the automatic extraction of hyponym relation from massive text datasets is naturally one of important NLP research tasks. The emphasis and difficult point of the research is how to validate a hyponym which is extracted with simple pattern matching method is really correct. By calculating the context feature similarity（*SimCF*） and Brown clustering similarity（*SimBrown*），this paper proposes a novel approach of hyponymy validation. It applies a clustering on hyponym candidates，and the clustering similarity feature is obtained by combining *SimCF* and *SimBrown*. The combination coefficient of two kinds of similarity is derived based on the *SimCF*s and *SimBrown*s between all labeled training words and their hyponyms. The model can filter roughly extraction results without any existed lexical relation dictionary or knowledge base. Evaluation on CCF NLP&CC2012 word semantic relation corpus shows that the proposed approach in this paper significantly improves the F measure value compared with other approaches including pattern matching and simple context comparison.

【Key words】 hyponymy relation；context similarity；Brown clustering similarity；Point Mutual Information（PMI）；pattern matching；clustering validation

DOI： 10.3969/j.issn.1000-3428.2015.02.028

1 概述

词汇上下位关系是指词汇概念之间在语义上的从属关系，即给定概念 *A* 和 *B*，若 *A* 的外延包含 *B* 的外延，则认为 *A* 和 *B* 具有上下位关系，即 *A* 是 *B* 的上位概念，*B* 是 *A* 的下位概念，这种关系也被称作

基金项目： 国家自然科学基金资助项目（61163039，61163036，61363058）；西北师范大学青年教师科研能力提升计划基金资助项目（NWNU-LKQN-10-2）。

作者简介： 张志昌（1976－），男，副教授、博士，主研方向：自然语言处理，Web 挖掘；陈松毅、刘 鑫，硕士研究生；马慧芳，副教授、博士。

收稿日期： 2014-03-04 **修回日期：** 2014-04-03 **E-mail：** zzc@nwnu.edu.cn

“is-a”关系,记作 *ISA*(*B*,*A*)。例如,“中国是一个国家”,则“国家”是“中国”的上位概念,即 *ISA*(中国,国家)。这种语义上的词汇上下位关系在本体知识库构建、机器翻译、自动问答等自然语言领域的相关应用中起着重要的作用。自文献[1]开始,已有很多关于上下位关系自动抽取的研究。但多数抽取方法都面临一个重要问题:如何验证抽取到的一组候选上下位关系词汇实例是否真正属于同一个语义类,即候选上下位关系的验证问题[2-3]。

本文提出一种基于统计并且无指导的词汇上下位关系验证方法。利用简单的模式匹配方法获得候选的词汇上下文关系后,通过计算词汇语境相似度和布朗聚类的相似度,将两者进行结合作为新的相似度特征,通过对上位词的全部候选下位词进行 K-means 聚类来对候选上下位关系进行验证和选择。

2 相关研究

对已有的研究成果进行总结,可将词汇上下位关系自动抽取的方法大致分为以下 3 类:

(1) 基于模式匹配的方法

该方法以文献[1]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在大语料中进行匹配来获取上下位关系。例如:设置模式形如“*B* is a A”,“*B* is a kind of *A*”,“*B*, *C* and other *A*”等(中文模式如:“*B* 是一个/类/种 *A*”等)。该方法有不同的变体,如文献[4]使用模式自举方法,而文献[5]使用了词性模板。该方法实现简单,并且模式的形式符合语言使用习惯,容易理解。但由于模式是由人来构造,模式的形式单一,只能覆盖部分词汇的表达形式,因此存在稀疏性问题[6],导致系统的准确率和召回率相对偏低。

(2) 基于语义词典、知识库的方法

目前广泛使用的语义词典、在线百科等知识库中都含有同义、反义、上下位关系等语义信息(英文有 WordNet[7], Wikipedia, Freebase 等, 中文有 HowNet、百度百科、互动百科等)。许多语义关系的抽取研究借助于此类语义词典、知识库所包含的语义信息[6,8]。但由于此类语义词典的构建多由人工参与,耗时耗力,因此往往其知识覆盖范围非常有限,且实时性较弱,无法及时体现最新的语言现象。

(3) 基于统计的方法

这类方法基于统计思想,通过机器学习方法构建语义模型,应用分类等数据挖掘技术计算不同概念之间的相关程度来获取上下位关系。文献[9]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,文献[10]运用了一种非线性概率模型,文献[11]构建了概念空间,并运用了潜在语义分析。该类方法越来越普遍地使用在语义关系抽取任务中。该类方法普遍基于以下假设:语义相似的概念出现在相似的上下文之中。

针对已有方法的特点和不足,本文提出一种基于统计并且无指导的词汇上下位关系验证方法,该方法和已有方法的区别在于:(1)利用无指导的聚类方法对上下位关系进行验证选择;(2)将聚类所用的相似度特征在传统的语境相似度的基础上结合了词汇的布朗聚类相似度。

3 候选上下位关系的获取

借鉴文献[12]方法,本文对候选上下位关系的获取方法进行了扩展,其实质是一种改良的基于模式匹配的方法。根据中文语法特点构造表 1 中的模式,然后利用搜索引擎索抽取大量能够匹配该模式的上下位候选上下位关系词对。在表 1 中,模式 1 为基本模式,模式 2 ~ 模式 4 为模式 1 的扩展模式,即通过扩展模式对基本模式获得的抽取结果进行自举扩展。

表 1 上下位关系抽取模式

模式	模式内容	举例
1	等[上位词]	汽车、火车、飞机等[交通工具]
2	[下位词]等[上位词]	[飞机]等[交通工具]
3	[下位词]等	[飞机]等
4	[下位词]	[飞机]

抽取算法如下:

输入 上位词 *C*,阈值 *R*(本文设 $R=5$)

输出 与 *C* 对应的实例集合 $IS=[I_1,I_2,\cdots,I_n]$

步骤 1 通过模式 1 在搜索引擎中进行查询获得支持句。从而获得候选上下位关系词对,放入集合 IS^*。

步骤 2 对于集合 IS^* 的每个元素,分别根据模式 2 ~ 模式 4 构造相应的查询字符串,获得相应的扩展支持句,并从中获得扩展后的上下位关系候选词对。

步骤 3 统计扩展词的出现次数,将出现次数大于阈值 *R* 次的词语放入集合 *IS* 中。

步骤 4 重复步骤 2 ~ 步骤 3,直到扩展词数量不再明显增加。

通过抽取算法可以获得一定数量的候选上下位关系。通过实验可知,对获取结果进行自举扩展对召回率有较大的提高,但同时又增加了错误结果的数量,准确率大大降低。因此,为有效提高准确率,本文提出一种基于语境特征与布朗聚类相结合的上下位关系验证方法,用于对模式匹配的结果进行验证过滤。

4 语境与布朗聚类特征结合的关系验证

将词汇的语境相似度特征和布朗聚类相似度特

征结合起来,通过聚类进行词汇的上下位关系验证,也是基于分布假设,即语义相似的概念出现在相似的上下文中。根据聚类理论:同一类别中的对象相似度较高,而不同类别中的对象相似度较小。同理,在候选上下位关系中,具有相同类别候选词的相似度较高,反之,相似度较低。

基于上述分析,本文将 K-means 聚类作为候选上下位关系的验证方法。在聚类过程中所使用的相似度分别为语境相似度、布朗聚类相似度和两者加权调和平均结合之后的相似度。

4.1　语境相似度特征

每个实体词在自然文本中都有各自的使用环境,即语境。语境即言语环境,分为狭义和广义2 种。狭义的语境是指书面语的上下文或口语的前言后语所形成的言语环境。后者则是指言语表达时的具体环境(既可指具体场合、也可指社会环境)。本文使用的词汇语境是指前者,即自然文本中的上下文信息。例如,“国家”一词常常出现在“举办”、“经济”等语境词之中,“中国”和“国家”有着相似的语境,但“中国人”跟“国家”的语境就有很大区别。如果可以获得概念的语境信息,就可以利用该信息对相应的上下位关系进行验证,从而过滤错误结果。

鉴于点互信息(Point Mutual Information,PMI)能较好地反映词汇与特征之间的共现关系,本文采用点互信息来选择和衡量词的语境特征及其权重,对词的语境信息进行量化建模。词汇 w_i 与上下文语境特征 f_j 之间的点互信息定义为:

$$PMI(w_i,f_j)=\mathrm{lb}\frac{P(w_i,f_j)}{P(w_i)P(f_j)} \tag{1}$$

其中,$P(w_i,f_j)$是词 w_i 和上下文语境特征 f_j 的共现概率;$P(w_i)$ 和 $P(f_j)$ 分别是词的出现概率,它们均可从语料库中用最大似然估计得到。

首先,通过点互信息值构造出目标词的语境特征词集合。本文通过对大量文本语料进行统计,取得与目标词互信息值最大的前20 个词,并将这些词作为目标词的语境特征词,记作 $CF(T)$。$CF(T)$是一个词集合,例如,“体育运动”的语境特征词如表 2 所示。

表 2　“体育运动”的语境特征词集合

词语	语境特征词集合
体育运动	国家 体育 田径 游泳 体校 武术 游戏 管理 后备 类型 篮球 中国 运动员 足球 参加 运动 发展 学校 批准 成立

根据向量空间模型可以构造该词的语境特征向量 $\boldsymbol{T}_{cf}=(w_{1,t},w_{2,t},\cdots,w_{N,t})$,其中,权重值 $w_{N,t}$ 为在目标词和第 n 维上的语境特征词之间的点互信息值;N 为词汇表中的词量。本文通过计算 2 个语境特征向量的余弦相似度值来得到两词之间的语境相似度,即:

$$SimCF(A,B)=\frac{T_{cf}(A)T_{cf}(B)}{|T_{cf}(A)||T_{cf}(B)|} \tag{2}$$

4.2　布朗聚类相似度特征

聚类方法是数据挖掘中通过特征进行无监督分类的有效方法。本文首先使用布朗聚类计算出各个候选下位词的前缀编码[13],得到候选词间布朗相似度,然后使用 K-Means 聚类方法进行多次聚类,通过计算上位词与每个候选词子集的距离,选择距离更近的一个,即可达到上下位关系验证的目的。

布朗聚类算法是文献[13]提出的一种基于纯文本的以词为处理单位的聚类算法。该方法用于分析未标注的大语料词汇聚合分布情况,并根据词分布相似度对词进行聚类。

定义分类器 C,$C:V\rightarrow\{1,2,\cdots,k\}$ 表示 C 将 V 中的词划分为 k 类,其中,V 为词汇表。

布朗聚类模型定义如下:

$$P(w_1,w_2,\cdots,w_n)=\prod_{i=1}^{n}e(w_i\mid C(w_i))\cdot q(C(w_i)\mid C(w_{i-1})) \tag{3}$$

其中,$w_1,w_2,\cdots,w_n$ 是自然句词序列;e 表示在 w_i 的分类下产生词 w_i 的概率;q 表示 w_{i-1}出现后接 w_i 的概率,即:

$$e(w_i\mid C(w_i))=\frac{Count(w_i,C(w_i))}{Count(C(w_i))} \tag{4}$$

$$q(C(w_i)\mid C(w_{i-1}))=\frac{Count(C(w_i),C(w_{i-1}))}{Count(C(w_{i-1}))} \tag{5}$$

根据以上定义,将分类器评价函数定义为:

$$Quality(C)=\sum_{i=1}^{n}\mathrm{lb}e(w_i\mid C(w_{i-1}))\cdot q(C(w_i)\mid C(w_{i-1}))=\sum_{c=1}^{n}\sum_{c^*}^{n}p(c,c^*)\cdot lb\frac{p(c,c^*)}{p(c)\cdot p(c^*)}+G \tag{6}$$

其中,G 为常数。

通过对语料进行布朗聚类分析,可得每个词的前缀编码(记为 $M(word)$),在此基础上可构造一颗分类树。根据分布假设可以知,具有相似前缀码的词的语义相似度较高,即分享同一个节点的词的语义相似度较高。所以,对于每一个从模式支持句中获得的候选上下位关系候选,本文使用候选词之间的布朗聚类相似度 $SimBrown(A,B)$进行验证过滤。

定义 A,B 节点距离为 $NodeDis(A,B)$:

$$NodeDis(A,B)=Len(M(A))+Len(M(B))-2Len(BLSS(M(A),M(B))) \tag{7}$$

其中,$BLSS(M(A),M(B))$表示 A 和 B 前缀码从根

开始的最长连续公共子序列;$Len(S)$代表序列长度。

通过节点距离,本文定义两节点布朗聚类相似度为:

$$SimBrown(A,B)=1-\frac{NodeDis(A,B)}{Len(M(A))+Len(M(B))} \tag{8}$$

4.3 语境和布朗聚类结合的相似度特征

除了利用语境相似度($SimCF$)和布朗聚类相似度($SimBrown$)作为 K-means 聚类的相似度特征,对候选上下位关系进行聚类验证之外,本文提出一种基于 2 种相似度相结合的新的相似度特征计算方法。该方法采用加权调和平均的方式结合了语境、布朗 2 种相似度。具体的结合公式如下:

$$Similarity(A,B)=\frac{(\alpha^2+1)\cdot SimCF(A,B)\cdot SimBrown(A,B)}{\alpha^2 SimCF(A,B)+SimBrown(A,B)} \tag{9}$$

其中,α 是结合系数。

通过式(9)计算出的候选上下位关系相似度值越高,则目标候选上下位关系属于正确关系的概率也就越大。所以,参数 α 优化过程的实质为使得 $\sum_{All}Similarity(A,B)$ 最大化的过程。通过训练可知,$\alpha=0.595$ 时获得最佳效果。

以结合相似度为例,选择上位词“主食”和其候选下位词,如表 3 所示。

表 3 “主食”的候选下位词集合

上位词	下位词
主食	米饭 黑芝麻 面粉 八宝粥 婚俗 小麦 赤豆 面包 青豆 馒头 核桃 面条

将候选下位词集合中的所有词基于式(9)所得的相似度进行 K-means 聚类(本文取 $K=2$),可得到如图 1 所示的散点图,该图体现点间距离的聚合关系,其坐标无实义。

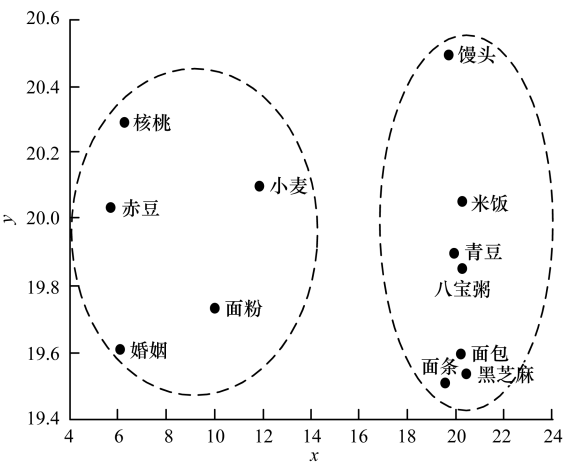

图 1 “主食”的候选下位词集合聚类散点图

从图 1 可得候选词集合的 2 个子集。定义上位词与候选词子集距离如下:

$$Dis(A,[B_1,B_2,\cdots,B_n])=\frac{\sum_{i=1}^{n}Similarity(A,B_i)}{n} \tag{10}$$

其中,$[B_1,B_2,\cdots,B_n]$是上位词 A 的下位词集合。

通过式(10)计算上位词与每个候选词子集的距离,选择距离更近的一个,并对结果进行多次迭代过滤,即可达到候选上下位关系验证过滤的目的。

5 实验结果与分析

5.1 评测语料与评价标准

本文采用 CCF NLP&CC 2012 语义关系识别标准评测集作为词汇上下位关系验证方法的训练和评测语料。该评测集包含 256 个上位词和分别与之对应的 5 718 个下位词。评测集的数据来源包括普通词典、百科词条、叙词表等多种资源。词汇的词性包括普通名词和专有名词。评测集格式如表 4 所示。

表 4 CCF NLP&CC 2012 标准评测集中“庙号”的下位词

上位词	下位词
庙号	玄宗 睿宗 神宗 孝宗 太祖 太宗 肃宗 宪宗 世祖 谥号 显宗 高祖

本文将评测集等分为训练集和测试集 2 个部分,每部分各有 128 个上位词,分别用于结合权重系数 α 的确定训练和方法的验证测试。

评测方法使用 CCF NLP&CC 2012 语义关系识别中的评测方法[14]。对抽取到的候选中文词汇上下位关系进行验证过滤,然后对结果采用准确率(Precision)、召回率(Recall)和 F 值(F-measure)3 个评测指标进行评价。

5.2 权重系数 α 的确定

单个上位词与其对应下位词相似度计算公式如下:

$$Q_1=Similarity(A,[B_1,B_2,\cdots,B_n])=\sum_{i=1}^{n}\frac{(\alpha^2+1)\cdot SimCF(A,B_i)\cdot SimBrown(A,B_i)}{\alpha^2 SimCF(A,B_i)+SimBrown(A,B_i)} \tag{11}$$

其中,Q_k 为 k 个词相似度。

在计算出训练集中所有 128 个上位词与它们分别对应的下位词的语境特征相似度、布朗聚类相似度之后,可计算出训练集中所有上位词与其对应下位词的结合相似度值,公式如下:

$$Q_{128}=\sum_{k=1}^{128}Similarity(A_k,[B_{k,1},B_{k,2},\cdots,B_{k,n}])=\sum_{k=1}^{128}\sum_{i=1}^{n}\frac{(\alpha^2+1)\cdot SimCF(A_k,B_{k,i})\cdot SimBrown(A_k,B_{k,i})}{\alpha^2 SimCF(A_k,B_{k,i})+SimBrown(A_k,B_{k,i})} \tag{12}$$

根据训练集计算出所有的 $SimCF(A_k, B_{k,i})$ 和 $SimBrown(A_k, B_{k,i})$ 之后,式(12)就成为了关于 α 的函数。依照上文分析,为使 Q_{128} 最大,对该函数求导,且令 $Q' = 0$,所获得的极值点,即最优的 α 值:

令 $X_{ik} = SimCF(A_k, B_{ik})$, $Y_{ik} = SimBrown(A_k, B_{ik})$,则有:

$$Q'_{128} = \left(\sum_{k=1}^{128} \sum_{i=1}^{n} \frac{(\alpha^2+1) \cdot SimCF(A_k, B_{k,i}) SimBrown(A_k, B_{k,i})}{\alpha^2 SimCF(A_k, B_{k,i}) + SimBrown(A_k, B_{k,i})} \right)' =$$

$$\left(\sum_{k=1}^{128} \sum_{i=1}^{n} \frac{(\alpha^2+1) \cdot X_{ik} \cdot Y_{ik}}{X_{ik}\alpha^2 + Y_{ik}} \right)' =$$

$$\sum_{k=1}^{128} \sum_{i=1}^{n} \frac{2X_{ik} \cdot Y_{ik}(X_{ik} \cdot \alpha^2 + Y_{ik}) - 2X_{ik}\alpha \cdot X_{ik}Y_{ik}(\alpha^2+1)}{(X_{ik}\alpha^2 + Y_{ik})^2} =$$

$$\sum_{k=1}^{128} \sum_{i=1}^{n} \frac{2X_{ik}^2 Y_{ik}\alpha^3 + 2X_{ik}Y_{ik}^2\alpha - 2X_{ik}^2 Y_{ik}\alpha^3 - 2X_{ik}^2 Y_{ik}\alpha)}{(X_{ik}\alpha^2 + Y_{ik})^2} =$$

$$\sum_{k=1}^{128} \sum_{i=1}^{n} \frac{2\alpha X_{ik}Y_{ik}(Y_{ik} - X_{ik})}{(X_{ik}\alpha^2 + Y_{ik})^2} \quad (13)$$

根据训练集数据(128 个上位词),计算可得 $\alpha = 0.595$。

5.3　结果分析

综合上述方法,对 CCF NLP&CC 2012 语义关系识别评测集中的上位词(即测试集中的 128 个词)做上下位关系抽取。本文使用搜狗实验室 2012 年发布的全网新闻数据和搜狐新闻数据(http://www.sogou.com/labs/resources.html)作为下位词抽取的主要数据来源,以训练语境特征模型和布朗聚类模型。该数据集为 2012 年 6 月 –2012 年 7 月国内、国际、体育、社会、娱乐等 18 个频道的新闻数据,共包含 2 623 521 篇文档。

首先用模式匹配和模式自举的方法抽取下位词,对获得的候选上下位关系集合进行性能评测,评测结果如表 5 所示。

表 5　基于模式匹配的上下位关系抽取结果

方法	准确率	召回率	F 值
模式匹配	0.138 7	0.444 8	0.198 0
模式匹配 + 自举	0.108 2	0.548 8	0.159 0

由表 5 可以看出,通过模式匹配抽取上下位关系的方法可以获得较多的候选结果,获得相对较高的召回率,但准确率很低。通过对抽取结果进行进一步的自举扩展,召回率方面获得了约 10% 的提升,但准确率进一步下降。说明模式自举扩展方法在提升召回率的同时使得错误结果数量也同时增大。

在模式匹配方法获取的候选结果基础上,本文分别使用语境特征相似度聚类验证方法、布朗聚类相似度聚类验证方法和二者结合的相似度特征聚类方法,对测试集进行验证过滤,不同方法的性能对比如表 6 所示。

表 6　不同上下位关系验证方法的性能对比

方法	准确率	召回率	F 值
语境特征	0.517 2	0.363 1	0.426 6
布朗聚类特征	0.653 3	0.436 9	0.523 6
语境特征 + 布朗聚类特征	0.735 1	0.491 6	0.589 2

从表 6 可知,对候选上下位关系分别进行基于语境特征相似度的聚类验证和基于布朗聚类相似度的聚类验证,抽取结果的准确率和 F 值均获得较大幅度的提升。但 2 种相似度特征结合后获得了比单一特征方法更好的效果,即证明了结合语境相似度和布朗聚类相似度为特征的上下位关系聚类验证方法的有效性。

将本文方法与其他参与了 CCF NLP&CC 2012 语义关系识别评测的中科院声学所等 5 种系统[15]进行比较,不同系统的方法性能比较情况如表 7 所示。

表 7　本文方法与其他系统的方法评测结果对比

方法	准确率	召回率	F 值
系统 1	0.804 1	0.122 5	0.212 6
系统 2	0.642 1	0.088 0	0.154 8
系统 3	0.721 9	0.464 6	0.565 3
系统 4	0.669 6	0.354 3	0.463 4
系统 5	0.637 0	0.503 3	0.562 3
本文方法	0.735 1	0.491 6	0.589 2

由表 7 可见,在参与该次评测的所有方法中,方法 5 所使用方法在 F 值上取到了较好的性能,该方法主要是通过使用维基百科和百度百科等现有的开放语义资源,并结合模板匹配和复合词拆解的方法得到了较高的准确率和召回率。可以认为该方法是一种基于现有知识词库和在线百科的上下位关系抽取方法。而本文所提出的方法无需借助现有上下位关系词库和在线百科,同样达到了较好的性能。

另外,由于测评结果根据 CCF NLP&CC 2012 语义关系所使用的标准评测集所判定,但该标准评测集中所包含的上下位关系相对有限,从而导致结果测试指标普遍偏低。例如,“传输协议”一词在标准集中的下位词集合与本文抽取结果对比如图 2所示。由图 2 可见,在使用本文抽取方法获得的结果中只有“网络传输协议”一词出现在标准评测集中,而根据人工评测,本文方法抽取到了更多的正确结果。因此,CCF NLP&&CC 2012 的评测集对上

下位关系的覆盖并不完备。鉴于此,笔者对本文方法抽取结果进行了人工评测(仅计算准确率),评测结果如表 8 所示。

传输层协议
电子邮件传输协议
实时传输协议
文件传输协议
安全传输协议
普通文件传输协议
超文本传输协议
网络传输协议
邮件传输协议

(a)标准评测集

mfc ftps rtp 吉以太网传输协议 https rtcp 互联网传输协议 hftp 数字电视传输协议 http xmppsip mmsh zmodem sftp rtmp xmodem 光纤传输协议 GSM SPX iSCSI **网络传输协议** 视频传输协议 以太网传输协议 rtsp CIFS GPS Infiniband POP3 SMPT IPX FTPS RTMP 数据传输协议 ftp JMS TCP RTP SCP NFS DDN RS485 socket COPS CDMA RTSP ISO CSMA 多播传输协议 网络电视传输协议 TLS

(b)本文抽取结果

图 2 “传输协议”在标准评测集中的下位词集合与本文抽取结果对比

表 8 人工评测结果

方法	准确率
模式抽取	0.375 5
模式抽取 + 自举	0.209 4
语境特征	0.552 3
布朗聚类特征	0.693 9
语境特征 + 布朗聚类特征	0.754 7

6 结束语

词汇的上下位关系在自然语言处理领域有着重要的应用价值。本文提出一种结合语境相似度特征和布朗聚类相似度特征的词汇上下位关系聚类验证方法,该方法在模式匹配方法抽取结果的基础上对上下位关系进行验证过滤。在 CCF NLP&CC 2012 评测语料上的实验结果表明,该方法实现简单,同时可取得较好的效果。

本文方法的不足在于语境特征提取过程和布朗聚类过程所需时间较长,且由于中文普遍存在的分词(词组)问题也对结果有较大的影响。下一步将尝试使用更高效的上下位词抽取方法,并结合有监督的自动分类方法对候选上下位关系进行是否为上下位关系的分类判断,以进一步优化验证效果。

参考文献

[1] Hearst M. Automatic Acquisition of Hyponyms from Large Text Corpora[C]//Proceedings of COLING'92. New York, USA:[s. n.],1992:539-545.

[2] Kozareva Z, Riloff E, Hovy E. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs[C]//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Columbus, USA:[s. n.],2008:1048-1056.

[3] Kozareva Z, Hovy E. A Semi-supervised Method to Learn and Construct Taxonomies Using the Web[C]//Proceedings of EMNLP'10. Boston, USA:[s. n.], 2010:1110-1118.

[4] Zhang Chunxia, Jiang Peng. Automatic Extraction of Definitions[C]//Proceedings of ICCSIT'09. Beijing, China:[s. n.],2009:364-368.

[5] Westerhout E. Definition Extraction Using Linguistic and Structural Features [C]//Proceedings of the 1st Workshop on Definition Extraction. Borovets, Bulgaria:[s. n.],2009:61-67.

[6] Akiba T, Sakai T. Japanese Hyponymy Extraction Based on a Term Similarity Graph[R]. Tokyo, Japan: IPSJ SIG, Technical Reprot:2011-IFAT-104,2011.

[7] Miller G A. WordNet:A Lexical Database for English[J]. Communications of the ACM,1995,38(11):39-41.

[8] Suchanek F M, Kasneci G, Weikum G. Yago: A Large Ontology from Wikipedia and WordNet[J]. Web Semantics: Science, Services and Agents on the World Wide Web,2008,6(3):203-217.

[9] Boella G, di Caro L. 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. Sofia, Bulgaria:[s. n.], 2013:532-537.

[10] Zhang Fan, Shi Shuming, Liu Jing, et al. Nonlinear Evidence Fusion and Propagation for Hyponymy Relation Mining[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, USA:[s. n.],2011,1159-1168.

[11] 刘 磊,曹存根,张春霞,等. 概念空间中上下位关系的意义识别研究[J]. 计算机学报,2009,32(8):1-14.

[12] Wang R C, Cohen W W. Automatic Set Instance Extraction Using Web[C]//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain:[s. n.],2009:101-110.

[13] Brown P F, Pietra V J D, de Souza P V. Class-based n-gram Models of Natural Language[J]. Computational Linguistics,1992,18(4):467-480.

[14] CCF NLP&CC2012 语义关系识别标准评测集[EB/OL]. [2014-02-14]. http://tcci.ccf.org.cn/conference/2012.

[15] CCF NLP&CC2012 语义关系评测结果[EB/OL]. [2014-02-14]. http://tcci.ccf.org.cn/conference/2012/dldoc/2012 语义关系评测结果.pdf.

编辑 金胡考