基于弱监督学习的中文百科数据属性抽取

贾 真，杨 燕，何大可

(西南交通大学信息科学与技术学院 成都 610031)

【摘要】提出基于弱监督学习的属性抽取方法，利用知识库中已有结构化的属性信息自动获取训练语料，有效解决了训练语料不足问题。针对训练语料存在的噪声问题，提出基于关键词过滤的训练语料优化方法。提出*n*元模式特征提取方法，该特征能够缓解传统*n*-gram特征稀疏性问题。实验数据源来自互动百科，从互动百科信息盒中抽取结构化属性信息构建知识库，从百科条目文本中自动获取训练数据和测试数据。实验结果表明，关键词过滤能有效提高训练语料的质量，与传统*n*-gram特征相比，*n*元模式特征能够提高属性抽取的性能。

关 键 词 属性抽取; 特征提取; 关系抽取; 弱监督学习

中图分类号 TP391 文献标志码 A doi:10.3969/j.issn.1001-0548.2014.05.022

Attribute Extraction of Chinese Online Encyclopedia Based on Weakly Supervised Learning

JIA Zhen, YANG Yan, and HE Da-ke

(School of Information Science and Technology, Southwest Jiaotong University Chengdu 610031)

Abstract An attribute extraction method based on weakly supervised learning is proposed in the paper. The training corpus is automatically acquired from natural language texts by using structured attribute information from knowledgebase. To solve the problem that noise exists in the training corpus, an optimization method based on keywords filtering is proposed. *N*-pattern features extraction method is proposed which can relieve to some extent the data sparsity problem of traditional *n*-gram features. Experiment data are downloaded from Hudong Baike. Structured attribute information is extracted from infoboxes of Hudong Baike and used to construct knowledgebase. Training data and testing data are acquired from encyclopedia entry texts. Experiment results show that the method of keywords filtering can effectively improve the quality of training corpus, and achieve better performance of attribute extraction by using *n*-pattern features, compared with traditional *n*-gram features.

Key words attribute extraction; feature extraction; relation extraction; weakly supervised learning

属性抽取是关系抽取任务之一，在知识库构建、自动问答、信息检索等多个领域具有重要的应用价值。传统基于有监督学习的关系抽取需要大量人工标注的训练语料，面向海量的网络数据，人工标注几乎是不可能的。如何能够实现监督最小化，即不使用人工标注或减少人工标注来构建高性能的关系抽取系统是当前的研究热点。

弱监督学习(weakly supervised learning)[1]关系抽取是一种基于噪声训练数据的半监督方法，利用知识库中已有的关系实体对，从文本集中自动获得训练语料，有效解决了训练语料不足的问题。然而，这种基于实体对共现自动建立起来的训练语料含有大量的噪声。如知识库中存在关系<余锡渠，籍贯，澄海县>。从文本集中获得包含实体对“余锡渠，澄海县”的句子：“新中国成立后，余锡渠任澄海县第一任县长。”该句子并未描述“余锡渠”和“澄海县”是籍贯的关系。含有噪声的训练语料影响关系抽取模型的性能。在属性关系描述语句中通常以某个特定关键词为核心，本文提出利用关键词对训练数据进行过滤，去除噪声，提高训练数据的质量。本文将属性抽取看作分类问题，利用训练语料训练分类器，对测试数据中是否具有属性关系进行预测。分类器的性能取决于选择的特征是否能够最大程度地表达不同类别的差异，选择恰当的特征有助于训练出性能较好的分类器。传统句子分类往往采用*n*-gram特征，本文提出了一种*n*元模式(*n*-pattern)特

收稿日期：2014－02－24；修回日期：2014－07－08
基金项目：国家自然科学基金(61170111, 61202043, 61262058)
作者简介：贾真(1975－)，女，博士生，主要从事信息抽取、内容安全、知识工程方面的研究.

征。与*n*-gram不同，*n*元模式中的项既可以是词语，也可以是词性，如当词语是实体词或频次较低的词时可用词性代替。*n*元模式不仅能够捕捉词语之间的序列关系，体现语法习惯，也能够对词语进行泛化，进一步缓解数据稀疏性问题。

本文基于弱监督学习方法，以属性抽取展开研究。主要贡献有：1) 提出了基于关键词过滤的训练语料优化方法；2) 提出了*n*元模式特征提取方法，并与传统*n*-gram特征进行了比较；3) 以中文网络百科为数据源进行了人物属性抽取实验，验证了基于关键词过滤和*n*元模式特征对弱监督学习属性抽取效果的提升。

1　相关工作

基于弱监督学习的关系抽取最早由文献[1]提出，用于从学术文献的摘要中抽取蛋白质与基因之间的关系。文献[2]利用维基百科信息盒中结构化的<属性，属性值>二元组对维基百科条目文本的句子进行回标，自动获取属性抽取的训练语料，基于最大熵模型和CRF模型分别训练句子分类器和属性值抽取器。文献[3]分别将具有关系的实体对正例和反例作为查询请求，从搜索引擎查询结果中提取包含实体对的句子作为训练语料。文献[4]从Freebase中获取关系实体对，从维基百科条目文本中获取关系抽取训练数据。文献[5]把关系抽取和实体的种类综合考虑，利用实体的类别来过滤掉部分错误的关系。文献[6]认为如果两个实体之间存在某种关系，那么含有实体对的句子中至少有一个句子描述了该关系。文献[7]基于弱监督学习对TAC-KBP进行属性模板填充。文献[8]提出了利用协同训练方法强化弱监督关系抽取模型。文献[9]提出了MultiRu算法，一个多实例多标签分类模型，可为一个实体对预测多种关系。

2　弱监督学习属性抽取框架

弱监督学习属性抽取框架如图1所示。首先从知识库中获取<实体，属性值>二元组；从文本集中寻找含有该二元组的句子，建立训练语料；再训练分类器，对测试文本集中的句子进行预测。

该框架主要包括3个重要的因素：知识库、训练语料和分类器。

1) 知识库：弱监督学习属性抽取依赖于某领域的知识库，只要构建了知识库，并找到与知识库相契合的文本集合，就可以自动地进行抽取。

2) 训练语料：利用知识库中的<实体，属性值>二元组，在文本集中进行匹配，提取包含二元组的句子作为训练语料。与人工标注构建训练语料不同，训练语料是自动获取的，其中存在错误与噪声。

3) 分类器：利用传统有监督学习方法训练分类器。

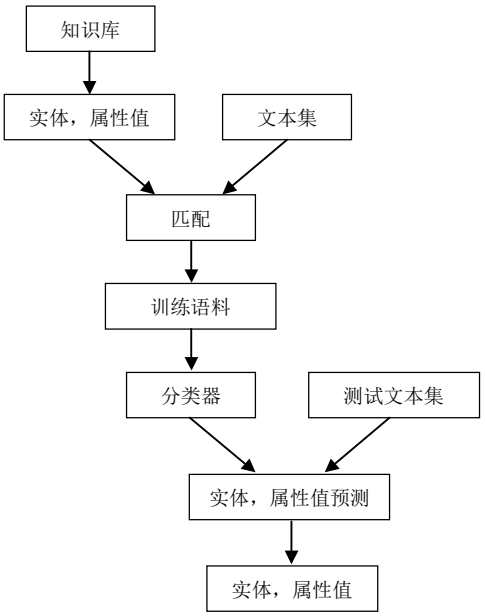

图1　弱监督学习属性抽取框架

3　弱监督学习属性抽取方法

3.1　知识库

弱监督学习属性抽取依赖于结构化的知识库。互动百科是目前最大的中文网络百科之一，互动百科的部分条目中存在人工创建的信息盒，信息盒中包含了大量半结构化的属性信息。从互动百科信息盒中抽取属性信息构建知识库。根据属性值的实体类型抽取结构化的<实体，属性，属性值>三元组。如，条目“钱学森”毕业院校信息盒中的内容为“上海交通大学机械工程系、北京清华大学留美公费生、美国麻省理工学院、美国加利福尼亚理工学院”。毕业院校属性值的实体类型是机构名，对信息盒内容进行分词、词性标注和实体标注[10]后为：“上海交通大学机械工程系/nt、/w 北京清华大学/ntu 留美/vi 公费生/n、/w 美国麻省理工学院/nt，/w 美国加利福尼亚理工学院/nt”。其中，nt、ntu 是机构名实体标注。该信息盒内容转化为 4 个结构化的属性关系三元组：<钱学森，毕业院校，上海交通大学机械工程系>、<钱学森，毕业院校，北京清华大学>、<钱学森，毕业院校，美国麻省理工学院>和<钱学森，毕业院校，美国加利福尼亚理工学院>。

3.2 训练语料

利用知识库中的实体和属性值对从百科条目文本中自动获取训练语料。属性描述语句中通常以某个特定的关键词为核心，如，描述“出生年月”属性，通常会有“生于、出生于、生”等词语。反之，不含有关键词的句子很可能没有描述属性信息。

本文采用一种基于熵的特征选择方法[11]提取关键词。将实体和属性值之间的文本组成集合 $S=\{s_1,s_2,\cdots,s_n\}$，且限定文本中的词语为名词和动词，过滤其他词性的词。$W=\{w_1,w_2,\cdots,w_m\}$ 是所有句子集合S中所有词的集合。利用式(1)计算S的熵值E：

$$E=-\sum_{i=1}^{n}\sum_{j=1}^{n}T_{i,j}\log T_{i,j}+(1-T_{i,j})\log(1-T_{i,j})) \quad (1)$$

式中，$T_{i,j}$ 为 s_i 和 s_j 之间的相似度函数，$T_{i,j}=\exp(-\alpha D_{i,j})$，$D_{i,j}$为$s_i$和$s_j$之间的欧式距离，$\alpha$ 是一个正数，取值为 $-\ln 0.5\sqrt{D}$ 。

从S集合中依次去掉W集合中的每个词语，计算得到 $\{E_1,E_2,\cdots,E_m\}$，选择对E值提升最大的前K个词语作为关键词。利用关键词对属性抽取的训练语料进行过滤。

3.3 分类器

把属性抽取看作一个分类问题，利用训练语句训练分类器，然后对测试文本集中的实体对进行预测。分类器性能的优劣往往取决于选择的特征是否能够最大程度地表达不同类别的差异，选择恰当的特征有助于训练出性能较好的分类器，实现不同类别的最优划分。句子分类常用的特征包括词法特征、句法特征和n-gram特征[8]。词法特征由句子中的词序列或词性序列构成，而句子中的语言描述过于具体，很难在其他的句子中再次出现，导致严重的数据稀疏性问题，也使得训练出的模型缺乏泛化能力。句法特征从句子的依存句法分析结果中获取，也存在词法特征中的数据稀疏性问题，且依赖于句法分析的效果，现有中文句法分析工具的准确率都不是很理想，导致句法特征不可靠。n-gram特征通常是文本中n个连续词或n个连续词性组成的序列，可以捕捉到局部范围内连续词语之间的序列关系，体现语法习惯，然而，当n较大时，n-gram特征也存在词法特征的数据稀疏性问题。

本文提出一种n元模式(n-pattern)特征。n-pattern是由文本中n个词语、词性标注或实体标注组成的序列。与n-gram不同，n-pattern中的项可以有间隔，当n-pattern中的词语是实体词或频次较低的词时，可以用词性代替。n-pattern的本质是文本子序列，能够捕捉词语之间的序列关系，体现语法习惯；n-pattern中的词可以不连续，实体词或低频词用词性进行泛化，使得n-pattern特征能够进一步缓解数据稀疏性问题。

如，从句子“sub 出任/v 北京大学/ntu obj”中提取词法特征、2-gram特征和2-pattern如表1所示。

表1 特征提取

特征类型	特征示例
词法特征	sub 出任 北京大学 obj sub v ntu obj
2-gram	<sub 出任>,<出任 北京大学>,<北京大学 obj>
2-pattern	<sub 出任>,<sub ntu>,<sub obj>,<出任 ntu>,<出任 obj>,<ntu obj>

表中，sub表示实体，obj表示属性值，北京大学的实体类别是大学机构名(标注为ntu[10])，在n元模式中用实体标注“ntu”代替。

从句子中提取n-pattern时，由于词语间距离越远，关联性越小，本文通过设定窗口W限制n-pattern的提取范围。n-pattern提取算法如下：

算法1 n-pattern提取算法。

输入：n；句子集合 $S=\{s_1,s_2,\cdots,s_n\}$；窗口N；最小词频次$f_{\min}$；实体标注集E。

输出：n-pattern集合$P^{(n)}=\{p_1^{(n)},p_2^{(n)},\cdots\}$。

对句子集合 S 中的词进行词频统计，建立词频表 $W=\{(w_1,f_1),(w_2,f_2)\cdots\}$。

$P^{(n)}\leftarrow\varnothing$
For $s_i\in S$ do
  $L\leftarrow$lengthof(s_i)
  For (j from1to L–N) do
    $P^{(n-1)}\leftarrow\varnothing$, Sequence$e\leftarrow\varnothing$
    $e\leftarrow$Select((w_j, t_j), E, W, $f_{\min}$)
    for (u from j+1 to j+N–1)
      Sequence $\leftarrow$ Select((w_u, t_u), E, W, $f_{\min}$)
    Endfor
    $P^{(n-1)}\leftarrow$Subsequence(n–1, Sequence)
    $P^{(n)}\leftarrow e\oplus P^{(n-1)}$
  Endfor
  Sequence$\leftarrow\varnothing$
  For (j from L–N+1 to L)
    Sequence$\leftarrow$Select((w_j, t_j), E, W, $f_{\min}$)
  Endfor
  $P^{(n)}\leftarrow$Subsequence(n, Sequence)
Endfor
Output($P^{(n)}$)

n-pattern提取算法包含以下主要步骤。

1) 对句子集合S中的词进行词频统计，得到词频表W。2) 创建一个空的*n*-pattern集合$P^{(n)}$。3) 对于每一个句子s_i，统计词语的个数，得到句子长度L。4) 窗口从句子第一个词开始逐词向后滑动，直到第L–N个词。5) 提取窗口中第一个项，当第一个项是实体或词频小于最小词频时，提取其词性标注。6) 提取窗口内其他项(词语、词性标注或实体标注)组成新的序列。7) 从序列中提取(n–1)元模式，并与窗口中第一项组成*n*-pattern。当窗口滑动到第(L–N+1)个词时，提取窗口内所有项组成*n*-pattern。8) Select的功能是从二元组(w, t)中提取词语、词性标注或实体标注，当$t \in E$时，提取实体标注t；当w的词频小于最小词频时，提取词性t。9) Subsequence算法的功能是从序列Sequence中提取n元子序列。10) *n*-pattern提取算法时间复杂度为O(n^3)。

4 实验与分析

4.1 数据集与预处理

从互动百科下载了约30万个人物条目页面，从条目页面提取出条目名、信息盒和正文内容作为数据源。

将职业、籍贯、出生年月、毕业院校、去世月和出生地信息盒的内容进行结构化处理，构建知识库。提取出的属性三元组个数约为363 000个。从正文中抽取含有人名实体和属性值对的句子作为训练语料。自然语言预处理工具使用西南交通大学中文分词平台[10]，分类器采用最大熵模型[12]。

4.2 关键词过滤

利用基于熵的特征选择方法提取关键词，并过滤掉不含关键词的句子。本文统计了关键词过滤前、后训练语料中未表达指定属性关系的句子，统计结果如表2所示。

表2　未表达属性关系的句子统计　　%

属性	过滤前	过滤后
毕业院校	28.75	14
出生年月	4.47	0.44
去世年月	21	4.5
出生地	36.25	9
籍贯	37.15	6
职业	12.8	8

从表2看出未用关键词过滤前，出生地、籍贯等训练语料中有较多句子没有表达对应的属性关系。利用关键词过滤后，未表达属性关系的句子数量明显减少，但仍存在一些噪声，这是因为训练数据是自动生成的，某些句子虽然含有关键词，但仍然没有表达对应的属性关系。如句子：

“出生才几个月，sub 就开始接拍某知名品牌 obj 地区广告。”

该句子虽然有关键词“出生”，但并没有表达sub出生在obj的关系。

毕业院校属性的训练语料用关键词过滤后仍有较多的噪声，是由于某些关键词不准确。如“学习”、“考入”等关键词有时并没有表达毕业关系。

4.3 属性抽取

本文共设计了两种不同的*n*-pattern特征：2-pattern(2np)和3-pattern(3np)。*n*-pattern提取的窗口W为3。此外，还设计了多种不同的*n*-gram特征：基于词序列的2-gram(2wg)和3-gram(3wg)、基于词性序列的2-gram(2pg)和3-gram(3pg)，以及词和词性组合的2-gram(2wpg)和3-gram(3wpg)。对实体和属性值之间的文本进行特征提取。用准确率(P)、召回率(R)和F值(F)评价属性抽取性能。

2-gram特征和2-pattern特征下属性抽取性能比较如表3所示。从表3看出，基于词序列的2-gram(2wg)特征优于基于词性序列的2-gram(2pg)特征，词和词性组合的2-gram(2wpg)特征优于基于词序列的2-gram(2wg)特征。本文提出的2-pattern(2np)特征除了出生地的召回率低于2pg特征和2wpg特征，出生年月的准确率略低于2wpg，其他属性的准确率和召回率均高于2-gram特征。2-pattern(2np)特征的6个属性抽取平均准确率、平均召回率和性能综合评价F值均高于2-gram特征。

3-gram特征和3-pattern特征下属性抽取性能比较如表4所示。

从表4看出，基于词性序列的3-gram(3pg)特征优于基于词序列的3-gram(3wg)特征，词和词性组合的3-gram(3wpg)特征优于基于词性序列的3-gram(3pg)特征。3-pattern(3np)特征的召回率除了籍贯低于3wpg特征，职业低于3wg特征外，其他属性均高于3-gram特征。3-pattern(3np)特征的准确率除了毕业院校低于3wg特征，出生年月低于3wpg特征外，其他属性均高于3-gram特征。6个属性抽取平均准确率、平均召回率和性能综合评价F值3-pattern(3np)特征均高于3-gram特征。从表3和表4看出3-gram特征性能较2-gram更低，说明多个连续词语导致的特征稀疏问题比较严重。出生地和籍贯两个属性的抽取性能较差。这是由于出生地和籍贯两个属性的训练语

料中存在较多相同的特征，类别之间的差异不明显，导致分类效果较差。

表3 2-gram特征与2-pattern特征的属性抽取性能 %

属性	2wg			2pg			2wpg			2np		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
毕业院校	84.9	91	87.8	63.8	72.6	67.9	87.6	86.4	87	90.7	91.6	91.1
出生年月	80	85.6	82.7	76.2	79.2	77.6	88.4	82.2	85.2	87.4	91.4	89.3
去世年月	89.6	86.2	87.9	68.2	70.4	69.3	83.2	88	85.5	92.1	92.8	92.5
出生地	44.5	48.8	46.6	45.5	54.6	49.4	45	56.4	50	52.7	54	53.4
籍贯	44.7	30	35.8	45.2	25.2	32.3	47.8	34.8	40.3	55.9	48.2	51.8
职业	69.2	77.8	73.3	68.8	70.2	69.5	78.9	82.4	80.6	83.4	87.4	85.4
平均	68.8	69.9	69	61.3	62	61	71.8	71.7	71.4	77	77.6	77.3

表4 3-gram特征与3-pattern特征的属性抽取性能 %

属性	3wg			3pg			3wpg			3np		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
毕业院校	90	68.2	77.6	70.6	71.6	71.1	78.2	75.2	76.7	88.6	87.4	88
出生年月	65.3	19.2	29.7	74.5	34.4	47.6	77.7	58.6	66.8	77.2	87.6	82.1
去世年月	86.6	45.2	59.4	75.4	60.2	67	73.6	78.2	75.8	87	82.8	84.8
出生地	29.9	18	22.4	40.4	38.8	39.6	42.9	48	45.3	48.8	57.6	52.8
籍贯	40.5	13.2	19.9	41.9	38.6	40.2	42.6	40	41.2	51.4	39.8	44.9
职业	24.3	85	37.8	38.2	70.4	49.5	58.5	68	62.9	76	74	75
平均	56.1	41.5	41.1	56.8	52.3	52.5	62.3	61.3	61.5	71.5	71.5	71.3

文献[8]提出采用协同训练方法提升有噪声训练数据下分类器的性能，本文与该方法进行比较。为了构造含有噪声的训练语料，本文在训练语料中加入噪声数据，正确的训练语料与噪声数据比例为4:1。每次迭代用投票策略选择40个句子加入到两个视图的训练数据中。6个属性抽取的平均准确率(P')、平均召回率(R')和平均F值(F')在不同迭代次数中变化情况如图2所示。

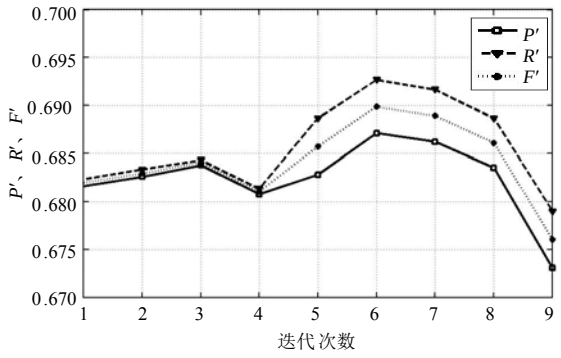

图2 基于协同训练的弱监督属性抽取性能

在第1、2、3次迭代时，分类器的性能逐渐提高，第4次迭代时，分类器性能略有下降，第5～6次迭代时，平均召回率有较大提高，总体性能有所提高，第7～9次迭代时，分类器性能逐渐下降。从图2看出，协同训练虽然能够在一定程度上提高分类器的性能，但是由于初始训练数据以及每次引入的训练数据中均含有噪声，平均F值最高为0.69，低于本文方法(77.6%)。

5 结 论

本文提出了基于弱监督学习的属性抽取方法，针对自动获取训练语料中含有较多的噪声和错误的问题，通过关键词过滤的方法获取较为可靠的训练数据。针对传统*n*-gram特征存在特征稀疏性问题，提出了一种较为灵活的*n*元模式特征，可以在句子中窗口范围内提取*n*元模式，并对*n*元模式中的词语进行泛化，进一步缓解了特征稀疏性问题。从互动百科采集实验数据集，实验结果验证了用关键词对训练语料进行过滤能够有效去除训练语料中的噪声，提高训练语料的质量，提升分类器的性能；与传统*n*-gram特征相比，*n*元模式特征能够提高属性抽取的性能。

在接下来的工作中，将继续研究弱监督学习关系抽取方法，深入研究噪声训练数据对关系抽取性能的影响，并研究如何利用集成学习方法提高属性抽取的性能。

参 考 文 献

[1] CRAVEN M, KUMLIEN J. Constructing biological knowledge bases by extracting information from text sources[C]//Proc of the 7th International Conference on Intelligent Systems for Molecular Biology. Heidelberg: AAAI, 1999: 77-86.

[2] WU F, WELD D S. Autonomously semantifying Wikipedia[C]//Proc of the sixteenth ACM Conference on Information and Knowledge Management. Lisbon, Portugal: ACM, 2007: 41-50.

[3] BUNESCU R C, MOONEY R J. Learning to extract relations from the web using minimal supervision[C]//Proc of the 45th Annual Meeting of the Association for Computational Linguistics. Prague: ACL, 2007.

[4] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]//Proc of the 47th Annual Meeting of the Association for Computational Linguistics. Singapore: ACL, 2009: 1003-1011.

[5] YAO L, RIEDEL S, MCCALLUM A. Collective cross document relation extraction without labeled data[C]//Proc of Empirical Methods in Natural Language Processing. Massachusetts: ACL, 2010: 1013-1023.

[6] RIEDEL S, YAO L, MCCALLUM A. Modeling relations and their mentions without labeled text[J]. Machine Learning and Knowledge Discovery in Databases, 2010(6323): 148-163.

[7] SURDEANU M, MCCLOSKY D, TIBSHIRANI J, et al. A simple distant supervision approach for the TAC-KBP slot filling task[C]//Proc of the TAC-KBP 2010 Workshop. Gaithersburg: Springer, 2010:1-5.

[8] 陈立玮, 冯岩松, 赵东岩. 基于弱监督学习的海量网络数据关系抽取[J]. 计算机研究与发展, 2013, 50(9): 1825-1835.
CHEN Li-wei, FENG Yan-song, ZHAO Dong-yan. Extracting relations from the web via weakly supervised learning[J]. Journal of Computer Research and Development, 2013, 50(9): 1825-1835.

[9] HOFFMANN R, ZHANG C, LING X, et al. Knowledgebased weak supervision for information extraction of overlapping relations[C]//Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: ACL, 2011: 541-550.

[10] 尹红风, 贾真, 李天瑞. 西南交通大学中文分词平台[EB/OL]. [2013-07-01]. http://ics.swjtu.edu.cn.
YIN Hong-feng, JIA Zhen, LI Tian-rui. Southwest Jiaotong University Chinese segmentation platform[EB/OL]. [2014-07-01]. http://ics.swjtu.edu.cn.

[11] CHEN J, JI D, TAN C L, et al. Unsupervised feature selection for relation extraction[C]//Proc of the 2nd International Joint Conference on Natural Language Processing. Korea: Springer, 2005.

[12] BERGER A, PIETRA V, PIETRA S. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 1996, 22(1): 39-71.

编 辑 税 红