基于维基百科和模式聚类的实体关系抽取方法*

张苇如[1,2]，孙 乐[1]，韩先培[1]
[1]中国科学院 软件研究所，北京 100190
[2]中国科学院 研究生院，北京 100049
E-mail: weiru@nfs.iscas.ac.cn

摘 要：本文提出了一种基于维基百科和模式聚类的方法，旨在从开放文本中抽取高准确率的中文关系实体对。首次使用从人工标注知识体系知网到维基百科实体映射的方式获取关系实例，并且充分利用了维基百科的结构化特性，我们的方法很好地解决了实体识别的问题，生成了准确而显著的句子实例；进一步，提出了显著性假设和关键词假设，在此基础上构建基于关键词的分类及层次聚类算法，显著提升了模式的可信度。实验结果表明该方法有效提升了句子实例及模式的质量，获得了良好的抽取性能。

关键词：关系抽取；维基百科；模式聚类

An Entity Relation Extraction Method Based on Wikipedia and Pattern Clustering

Zhang Wei-ru[1,2], Sun Le[1], Han Xian-pei[1]
[1] Institute of Software, Chinese Academy of Sciences, Beijing 100190
[2]Graduate University of Chinese Academy of Sciences, Beijing 100049
E-mail: weiru@nfs.iscas.ac.cn

Abstract: This paper proposed a method based on Wikipedia and pattern clustering to extract Chinese entity relations of high accuracy from open text. We got relation instances by a mapping from HowNet to Wikipedia and made use of the structural characteristics of Wikipedia. Based on these, our method solved entity recognition problem and generated significant sentence instances. Furthermore, significance assumption and keyword assumption were proposed to support classification and hierarchy clustering algorithm for pattern reliability. The results showed that the method achieved a good performance by attaining high-quality seeds and patterns.

Keywords: relation extraction; Wikipedia; pattern clustering

1 引言

信息抽取研究技术是人们获取信息的有力工具，是应对信息爆炸带来的严重挑战的重要手段。信息抽取的目标是从无结构自然语言文本中提取计算机可理解的结构化信息，其中一种主要的结构化信息是实体语义关系。例如，实体间的上下位关系（is-a）、局部-整体关系(part-of)以及更具体的作者-作品关系(author-of)、首都-国家关系(capital-of)等。实体关系抽取是自动知识本体构建的基础，同时也可应用到问答、信息检索等多个领域。

传统方法通常采用自举迭代的模式匹配方法在自然语言文本上抽取关系实例。然而，这种方法有如下缺点：1）句子实例的获取过程依赖于实体识别的准确性，因为不正确的实体识别会导致错误的句子实例识别；2）包含关系实体对的句子不一定准确表示目标关系，这就导致构造的模式可能是不显著或错误的，并因此造成整个迭代过程的错误传播，影响抽取结果。

针对上述问题，本文利用维基百科生成句子实例。正如 Medelyan 等[1]所述，维基百科的结构特点恰好可以克服命名实体识别的问题。每一个维基页面代表一个概念即实体，在该页面中出现的其他实体，以超链接的方式标注。利用上述特性，就可以自动地识别出实体对，且共现的实体对通常都是显著的。同时，为了进一步提升模式的可信度，我们提出了模式的显著性假设和关键

* 本文相关研究得到国家自然科学基金(90920010)(60773027)以及国家重大科技专项经费(2010ZX01036-001-002-2)资助。

词假设。在此基础上，运用基于关键词的分类和层次聚类算法，对模式进行过滤和泛化。

本文选取五种不同的关系进行抽取实验。新实例对的抽取实验在中文维基百科及搜狗全网新闻语料库(Sogou_CA)上进行，大多数关系的抽取得到了良好的实验效果。

2 相关工作

早期的方法采用人工标注的数据作为初始输入，但这样做耗费大量的人力、时间成本。DIPRE 和 Snowball[2]使用少量的种子实例，通过自举迭代产生新的实例与模式。Ruiz-Casado 等[3]将 WordNet 中的关系实例映射到 Wikipedia 中，并抽取实例出现的上下文作为输入。实验总共抽取出 1200 个新的关系对，正确率在 61%~69%之间。Yan 等[4]不使用任何先验知识，提出一种无监督的关系抽取方法，并借助维基百科及互联网语料进行聚类。

在提升模式质量方面，DIPRE 和 Snowball 在每次迭代中，选取可信度高的实例加入正例集合，产生的模式在下一轮迭代中使用，其中，可信度高低由匹配到的模式数量决定。这种方法的问题在于对信息冗余的依赖性很大。由此，Pantel 和 Pennacchiotti[5]提出的 Espresso 方法采用了 Web Expansion 策略，在整个 Web 上获取冗余信息来挖掘大量的文本模式，并利用互信息判断模式和实例的可信度。该方法的潜在问题是，Web 上抽取的模式不一定能在当前领域的小数据集上可用。

3 关系抽取

本文提出的关系抽取策略共分为以下几个步骤：关系实例获取、句子实例获取、模式构建与挖掘、新实例抽取。

3.1 关系实例获取

语义关系挖掘是指，借助知网的概念描述体系，获取概念之间的潜在关系。除了本身所定义的 16 种显式关系外，知网中还存在大量隐式的语义关系。例如在下面的概念描述中：

CPU　　part|部件,%computer|电脑,heart|心

通过包含关系符号“%”，可以推断得知“CPU”是“电脑”的一部分。通过制定规则，我们从知网中挖掘出大量潜在的语义关系实体对。

3.2 句子实例获取

我们首先对维基百科数据进行如下处理：（1）繁体字符转换为简体；（2）关联重定向页面；（3）关联消歧页面；（4）对文本进行分词、词性标注及命名体识别。

为了实现实例映射，必须在维基百科中寻找 3.1 得到的实体对。沿用上例，实体对（CPU,电脑）的两个词条在维基百科中都存在，其中“CPU”有多个义项：

CPU是以下的简称：

- 中央处理器，是电脑的主要装置之一。（Central Processing Unit）
- 迦密柏雨中学，位于香港大埔的一间中学。（Carmel Pak U Secondary School）
- 中央政策组，一个香港政府部门，负责向行政长官等人提供意见。（Central Policy Unit）
- 哥伦比亚太平洋大学，一间位于美国加州的大学。（Columbia Pacific University）
- C.P.U.，一本香港电脑周刊，跨传媒集团出版。
- C.P.U.，一家香港运动服饰连锁店。

图 1　词条“CPU”的多个释义

我们采用一种基于字面匹配的算法 Lesk（Kilgarriff and Rosenzweig, 2005）进行消歧，选取维基百科中的释义及知网中的描述作为消歧上下文。经过消歧，“中央处理器”义项被选定。这种算法虽然简单，但是由于存在丰富的解释性上下文，因此可以得到理想的实体消歧效果。

接下来，抽取词条在对应文本中共现的句子作为句子实例。映射和消歧保证句子实例的准确性，而维基百科特有的超链接结构确保了句子内实体对之间关系模式的显著性。

3.3 模式挖掘

传统的启发式方法对句子实例进行词性标注，并用通配符替换实体对出现的位置来构建模式。例如："北京是中国的首都"的模式构建结果为"object 是/v target 的/u 首都/n"。

但是，上述方法构建的模式有如下缺点：1）通用性不足，需要泛化。例如，上述模式无法从句子实例"伯尼尔是欧洲联邦制国家瑞士的首都"中抽取出关系实例（伯尼尔，瑞士）。2）准确性不足，这是因为仅仅基于共现得到的句子实例通常包含大量噪声和反例。例如从"北京是中国政治文化的中心"中得到的模式并未表示北京和中国之间的 capital-of 关系。

为此，我们提出了下面两个假设：1 模式的显著性假设:表示一个关系的某种模式会在句子实例中出现多次；2 模式的关键词假设：模式通常以某个特定的关键词为核心。以上述两个假设为前提，本文利用基于关键词的分类与层次聚类方法，对模式进行过滤与泛化。

3.3.1 关键词的选择

基于关键词的分类方法能找到在语义及结构上都相似的模式。本文采用一种基于熵的特征选择[6]方法来确定关键词。其基本思想如下：假设 $P=\{p_1,p_2,\dots p_N\}$ 为所有模式的集合，$W=\{w_1,w_2,\dots w_M\}$ 为 P 中所有模式的词集合。利用如下公式计算 P 的熵值，其中，$S_{i,j}$ 为 p_i,p_j 之间的相似度函数 $S_{i,j}=\exp(-\alpha*D_{i,j})$, $D_{i,j}$ 是模式 p_i p_j 之间的距离，α 是一个正数，这里取值为 $-\frac{\ln 0.5}{\bar{D}}$：

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

从 P 集合中依次去掉 W 集合中的每个元素，计算得到$\{E_1,E_2,\dots E_M\}$,进行排序，E_i 值越大，则 w_i 越重要。我们选取对 E 值提升贡献最大的 K 个名词或动词作为目标关系的关键词。

3.3.2 基于关键词的过滤

对特定的目标关系，本文首先对其候选模式进行基于关键词的分类：按关键词的排序，包含该关键词的模式被归为一类，每个模式至多只能属于一个类。包含同一个关键词的模式也可能有多种。例如，对于关键词"首都"，可能存在"object 是/v target 的/u 首都/n"，"target 的/u 首都/n 是/v object"等多种模式。因此，针对每个基于关键词的类，我们对其内部样本再进行层次聚类，合并相似的模式，同时过滤出现频率较低的模式。具体算法如下：

输入：基于某个关键词的模式集合 $P=\{p_1,p_2\dots p_n\}$，阈值 t_1,t_2

输出：聚类后得到的簇 $Cluster=\{cluster_1,cluster_2\dots\}$;

Begin

初始化簇的集合 $Cluster=\{cluster_1,cluster_2\dots cluster_n\}$ ($cluster_i=\{p_i\}$ 1<=i<=n)

while $\min<=t_1$ **do**

if $\min_{cluster_i,cluster_j\in Cluster} dis_{complete\text{-}linkage}(cluster_i,cluster_j)<=t_1$ **then**

merge($cluster_i$, $cluster_j$)

for cluster ∈ Cluster **do**

if size(cluster) < t_2 **then** Cluster.remove(cluster)

return Cluster

End

距离的计算采用完全连锁（Complete-Linkage）即最大距离，保证一个簇中的模式两两之间都具有较高的相似性，其计算公式为：

$$\max(d(a,b):a\in A,b\in B) \quad (2)$$

其中，d(a,b)表示模式 a,b 的编辑距离。

3.3.3 基于编辑距离的泛化

本文在普通的编辑距离算法基础上，在 diff(i,j)中加入词性的影响，具体公式如下：

$$E(i,j)=\min\{1+E(i-1,j),\ 1+E(i,j-1),\ diff(i,j)+E(i-1,j-1)\}$$

$$\text{其中 diff}(i,j)=\begin{cases}0 & S[i]=T[j]\\ 0.5 & POS(S[i])=POS(T[j])\\ 1 & \text{else}\end{cases} \tag{3}$$

我们对 Ruiz-Casado[3]的方法作了改进。其基本思想是：在编辑距离矩阵中反向寻找编辑距离最短路径，同步进行保留、添加、删除、修改操作，最终得到泛化后的模式。假设有如下两个模式，泛化后的结果如箭头所示，图 2 展示了编辑距离计算及模式泛化的过程：

object/nx 是/v 位于/v target/nx 南部/f 的/u 一个/NUM 州/n

object/nx 是/v target/nx 西北部/f 的/u 一个/NUM 地区/n

⟹ object/nx 是/v * target/nx （南部|西北部）/f 的/u 一个/NUM （州|地区）/n

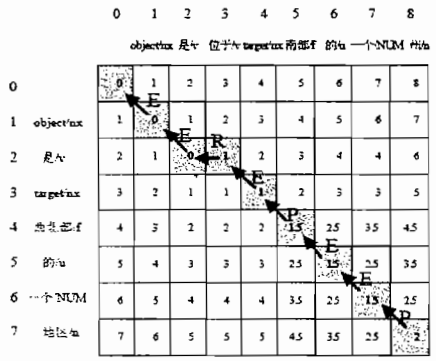

图 2 编辑距离的一个计算示例。其中‘E’表示保留该位置字符，‘R’表示删除，‘I’表示增加，‘U’表示更新，‘P’表示保留词性

3.4 新实例抽取

最后，我们利用生成的模式进行新关系实例的抽取。对象可以是维基百科也可以是其他语料。一旦任意模式匹配成功，object 和 target 所对应的实体就被作为目标关系的一个实例抽取。

4 实验

我们使用中文维基百科 2011 年 1 月 10 日的版本作为实验数据(包含 714,682 个词条)，另外还加入了搜狗全网新闻语料（Sogou_CA）。实验过程中，使用中国科学院自动化所开发的 NlprCSegTagNer2.0 对语料进行预处理。

4.1 关系实例映射

我们选取了首都-国家(capital-of)、位置关系(located-in)、整体-部分关系(part-of)、材料-成品关系(production)、属性-宿主关系(attribute-host)以及出生日期关系(was-born-on)作为目标关系，结果如表 1 所示。前五种关系从知网中挖掘而来，括号内是关系实例成功映射到维基百科中的比率。出生日期关系直接从维基百科的分类信息中得到。

表 1 关系实例挖掘与映射数量

	capital-of	located-in	part-of	production	attribute-host	was-born-on
挖掘数量	167	1112	3413	600	3101	9830
映射数量	153 (92%)	1015 (91%)	1084 (32%)	239 (40%)	726 (23%)	—

由表 1 数据，可以得知：1. 知网及维基百科的结构化信息中蕴含了大量的潜在关系实例，为抽取提供了良好的初始样本；2. 知网和维基百科的数据存在大量冗余，维基百科几乎覆盖了知网中的命名实体；3. 命名实体消歧可以很好地实现映射：对于有歧义的词，我们使用的消歧算法达到 82%的正确率，且大多错误是由于维基百科本身不存在与之对应的概念造成的。

4.2 模式性能验证

为了检验模式的性能，我们利用维基百科的结构化信息得到句子实例，进一步地通过基于关键词的分类与层次聚类算法过滤和泛化模式。表 2 是首都-国家关系(capital-of)的模式，其中，1-3 是使用我们提出的方法获得的模式，4-8 由 Ruiz-Casado 提出的一种简单的层次聚类算法生成：

表 2 capital-of 关系的泛化模式

			Pattern
基于关键词的分类与层次聚类		1	object/LOC （是|为/v） * target/LOC （的/u） 首都/n
		2	target/LOC （的/u） 首都/n （为/v|是/v） object/LOC
		3	target/LOC 定|迁/v 都/Ng object/LOC
简单的层次聚类	大阈值 (>=1.5)	4	target/LOC * * * （为/v|是/v） object/LOC
		5	object/LOC * * * * * target/LOC
	小阈值 (<1.5)	6	target/LOC 首都/n object/LOC
		7	target/LOC 首都/n 是/v object/LOC
		8	object/LOC 是/v target/LOC 的/u 首都/n

从表 2 可以看出，失去了关键词的依托，层次聚类终止条件的阈值较大可能造成过度泛化，包含大量反例与不显著模式，影响准确性；而小阈值对模式筛选苛刻，造成遗漏，降低了通用性。结果表明，我们提出的方法算法能有效地进行过滤与泛化，得到通用且准确的模式。

4.3 新实例抽取

最后，我们利用得到的过滤和泛化后的模式在维基百科与搜狗全网新闻语料上进行新实例的抽取。选取准确率作为抽取效果的评价指标：$\text{Precision} = \frac{\text{准确的实例个数}}{\text{抽取的总实例个数}}$。评价过程如下：对任意目标关系，随机选取大小为 1000 的实例子集（小于 1000 则全部选取），由一名标注者手工标注其准确性。为了确保标注的质量，从标注集中再选取大小为 200 的子集，由第二名标注者标注，并使用 kappa 系数评价两次标注的一致性。结果显示，κ 值大于 0.8，表明两次标注一致性很高，证明了评价的有效性。

表 3 中，capital-of 关系抽取的新实例个数分别占 49%，57%；was-born-on 分别占 73%, 99%。其他关系在两个数据集上抽取的关系对基本完全不同于种子数据，即所占比率达到 100%。

表 3 抽取效果

		capital-of	located-in	part-of	production	was-born-on
Wikipedia	Instances	219	5212	7433	75	4981
	Precision	93%	87%	53%	13%	84%
Sogou_CA	Instances	202	1662	13172	121	6513
	Precision	94%	77%	48%	15%	87%
All	Instances	282	6603	20601	176	11462
	Precision	91%	82%	49%	14%	85%

表4是利用表2中不同的模式对capital-of关系进行抽取得到的结果：

表4 使用不同的模式过滤泛化策略抽取得到的capital-of关系实例

	Instances	Precision
基于关键词的分类与层次聚类	282	91%
Ruiz-Casado使用的简单的层次聚类（小阈值）	223	89%

分析实验结果，我们可以得到如下结论：

（1）本文提出的方法能够取得良好的关系抽取性能

对于capital-of, located-in, was-born-on，无论是在维基百科数据集，还是在更为冗余的Web语料中,提出的方法都抽取出大规模高质量的关系实例。通过分析，我们认为高准确率得益于高质量的模式。这也就证明了利用维基百科作为句子实例的来源能有效地提高实体识别的准确性和显著性，而基于关键词的过滤与泛化策略能在模式的可信度和覆盖度之间取得好的平衡，如表4所示。尤其是对capital-of这种具体的关系类型，该方法非常有效，准确率在90%以上。少量错误源于特殊的上下文语境，例如“许多人认为悉尼是澳大利亚的首都”。

（2）基于自举的模式构建方法并不适用于所有的关系抽取

实验结果中，Production关系抽取的准确率和数量都不理想。据我们所知，目前针对这种关系的相关研究也都没有取得比较好的实验效果。究其原因，这类关系很少在文本中以实例共现的形式出现，因此无法生成适当的模式用以抽取实例。

5 结论及展望

本文提出了一种基于维基百科的抽取策略，旨在从开放文本中抽取高准确率的中文关系实体对。具体来说，我们首次提出了从人工标注知识体系知网到维基百科实体映射的方法获取关系实例，通过充分利用维基百科的结构化特性，该方法很好地解决了实体识别的问题，产生了准确而显著的句子实例；进一步，提出了模式的显著性假设和关键词假设，在此基础上构建了基于关键词的分类及层次聚类算法，显著提升了模式的可信度。实验结果表明我们的方法无论是在数量上还是准确率上都达到了良好的关系抽取性能。

在下一步工作中，我们希望对关系实例进行进一步挖掘，以提升抽取准确率并获取更多语义信息。另外，我们还将尝试抽取其他类别的关系实例。

参 考 文 献

[1] O. Medelyan, D. Milne, C. Legg and I. H. Witten. Mining Meaning from Wikipedia. International Journal of Human-Computer Studies, 2009.

[2] E. Agichtein, L. Gravano: Snowball: Extracting Relations from Large Plain-Text Collections. ACM DL 2000: 85-94.

[3] M. Ruiz- Casado, E. Alfonseca, P. Castells. Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia. In NLDB, 2006, 67-79.

[4] Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang, M. Ishizuka. Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. ACL and AFNLP, 2009.

[5] P. Pantel, M. Pennacchiotti. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. ACL, 2006, 113-120.

[6] J. X. Chen, D. H. Ji, C. L. Tan, Z. Y. Niu. Unsupervised Feature Selection for Relation Extraction. IJCNLP, 2005.

[7] 王刚. 自动抽取维基百科文本中的语义关系. 上海交通大学硕士学位论文. 2008.

[8] 董振东, 董强.《知网》, http://www.keenage.com.