文章编号：1003-0077(2007)02-0069-08

基于无监督学习的问答模式抽取技术

吴友政，赵 军，徐 波

（中国科学院 自动化研究所 模式识别国家重点实验室，北京 100080）

摘 要：本文提出了一种基于无监督学习算法的问答模式抽取技术从互联网上抽取应用于汉语问答系统的答案模式。该算法可以避免有监督学习算法的不足，它无需用户提供＜提问，答案＞对作为训练集，只需用户提供每种提问类型两个或以上的提问实例，算法即可通过 Web 检索、主题划分、模式提取、垂直聚类和水平聚类等步骤完成该类型提问的答案模式的学习。实验结果表明，论文提出的无监督问答模式学习方法是有效的，基于模式匹配的答案抽取技术能够较大幅度地提高汉语问答系统的性能。

关键词：人工智能；自然语言处理；汉语问答系统；问答模式；机器学习

中图分类号：TP391　　**文献标识码**：A

Unsupervised Answer Pattern Acquisition

WU You-zheng, ZHAO Jun, XU Bo

(National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100080, China)

Abstract: The paper presents an unsupervised learning algorithm to learn answer pattern for answer extraction module of Chinese Question Answering (QA). Given two or more questions of one question type, the algorithm can learn the corresponding answer patterns from internet via web search, topic segmentation, pattern extraction, vertical clustering and horizontal clustering, etc. The experimental results show that the performance of pattern-based answer extraction of Chinese QA is improved significantly.

Key words: artificial intelligence; natural language processing; Chinese question answering; answer pattern; machine learning

1 引言与研究动机

由于自然语言本身的灵活性和多变性，对同一语义往往存在不同的表述，这使得对问答技术研究面临许多困难。在 TREC（Text REtrieval Conference）评测的推动下，人们已经提出了很多解决方法[1~5]。但目前的自动问答技术仍然还不成熟。

为解决语言的灵活性和多变性，最直接的方法就是把提问和答案句都表示成统一的语义表示形式，然后进行匹配。然而现阶段，自然语言处理的各种底层技术仍不完善和不成熟，对文本进行深层分析，从语义层面来处理语言的灵活性和多变性是一件十分艰难的任务。于是，人们提出了基于字符表层的文本分析技术。实际上，语言的灵活性和多样性在一定程度上是可以通过基于字符表层的文本分析技术来获取，模式匹配技术即是这种方法的典型代表。已经有一些英文问答系统采用了这种技术，并在 TREC 评测中获得了很好的成绩[4]。本文同样希望通过模式匹配技术来转化中文问答系统答案抽取的难度，把从语义层面进行答案抽取的过程变成模式的匹配过程。

将模式匹配技术应用于问答系统的代表性工作有 Ravichandran[1]、Soubbotin[4]、Du[6]、Dumais[7] 和 Zhang[8] 等。

Soubbotin[4] 完全采用人工编写规则的方法获取

收稿日期：2006-04-05　定稿日期：2006-07-03

基金项目：国家自然科学基金资助项目(60372016)；北京市自然科学基金资助项目(40052027)

作者简介：吴友政(1976—)，男，博士生，主要研究领域为自然语言处理，信息检索，问答技术等。

问答模式。这种方法代价昂贵,劳动强度大,速度慢;且模式的扩大很困难,算法可移植性差。

所以,近年来问答模式的获取方法逐渐从人工组织的方法向机器学习的方法转变。2002 年 Ravichandran[1] 提出了通过有监督机器学习从网络文本中自动提取 6 种,即 BIRTHYEAR, INVEN-TOR, DISCOVER, DEFINITION, WHYFAMOUS LOCATION 等提问类型的答案模式。例如 INVENTOR 类型提问答案的一个模式为:" the 〈ANSWER〉 was invented by 〈NAME〉"。其中,ANWER 和 NAME 分别表示提问关键词和答案。这种方法使用用户提供的〈提问,答案〉对作为训练语料进行 Web 搜索,在对 AltaVista 返回的前 1 000 篇文章进行后处理后,采用后缀树模型(Suffix Tree)提取字符表层模式。因此,它是一种有监督的机器学习的方法。此外,字符表层模式的缺点是无法解决 ANWER 和 NAME 之间的长距离依存关系以及缺乏良好的泛化性。

Du 等人[6] 于 2004 年提出的问答系统的答案模式学习方法类似于 Ravichandran 方法,也是一种基于有监督的机器学习方法,不同之处在于提问分类和模式的表示两个方面。Du 首先把提问关键词定义为 4 大类(Q_Foucs, Q_NameEntity, Q_Verb, Q_BNP),然后对不同类型的提问学习其答案句的模式。例如" What Q_BeVerb Q_Focus in Q_LCN "提问类型的一个答案模式为:" ,〈A〉 Q_BeVerb Q_Focus in Q_LCN "。

通过前述分析可以发现,Ravichandran 和 Du 的方法均属于有监督的机器学习算法,算法的性能在很大程度上依赖于用户提供的〈提问,答案〉对。然而,由于答案表示形式的多样性,用户很难提供答案的所有可能出现形式。这在一定程度上影响着有监督问答模式学习算法的性能。例如,提问' 毛泽东同志出生地是哪里?",它的答案可能是' 湖南 "、" 湖南省 "、" 韶山冲 "、" 韶山 "和' 韶山市上屋场 "等等。

对此,本文提出了一种基于无监督学习算法的问答模式抽取技术,从互联网上抽取应用于汉语问答系统的答案模式。该方法和 Ravichandran, Du 等人工作的不同是:本文提出的无监督学习算法无需用户提供〈提问,答案〉对作为训练语料,只需用户提供每种提问类型两个或以上的提问实例,算法即可通过 Web 检索、主题划分、模式提取、垂直聚类和水平聚类等步骤完成该类型提问的答案模式的学习。所以,本文的算法可以很好的避免有监督学习算法的缺点:即因用户很难提供尽可能多的提问答案而造成算法性能的下降。

在开放测试语料上的实验结果表明:本文提出的基于无监督的问答模式抽取方法是有效的,能够较大幅度地提高汉语问答系统的性能。其中,基于字符表层模式的答案抽取系统性能较 Baseline 提高约 9.0%,基于句法模式的答案抽取系统性能较 Baseline 提高约 14.0%。

2 基于无监督的问答模式学习算法

本节以 BOOKAUTHOR 类型提问的答案模式抽取为例来说明无监督答案模式学习算法的整个流程,并详细介绍算法的四个核心模块:主题划分、模式抽取、垂直聚类和水平聚类。

算法输入:BOOKAUTHOR 类型提问的 2 个提问实例

Q1:《平凡的世界》是谁写的?

Q2:《西厢记》的作者是谁?

操作步骤:

1. 关键词提取和分类

   其主要任务是对提问句进行分词、命名实体识别、提取查询关键词以及对查询关键词进行分类。其中,分词和命名实体识别采用 NlprCSegTagNer[12] 工具;关键词分类目前采用的是人工方法。例如 Q1 的查询 Query1 = {平凡的世界/Q_FOUCS, 写/Q_I};Q2 的查询 Query2 = {西厢记/ Q_ FOCUS, 作者/Q_I},其中 Q_FOUCS 表示提问的焦点词,Q_I 表示除提问焦点词之外的其他查询词。

2. 查询词扩展

   查询扩展的目的是为了提高 Web 检索的召回率,本文使用《同义词词林》对 Q_I 类型的查询关键词进行扩展。

3. Web 检索

   提交查询 Query1 和 Query2 到 Google 搜索引擎,提取 Google 返回的前 1 000 个相关网页的片段,分别标记每个查询网页片段集合为 D1 和 D2。

4. 句子切分和检索

   剔除 D1,D2 中的 Html 标记和其他标记,并分别对其进行句子切分;使用查询 Query1 和 Query2 进行基于语言模型的句子检索;保留同时包括 Q_FOUCS 和答案类型实体(对于提问 Q1 和 Q2,答案类型实体是人名)的所有句子,标记句子集合为 S1 和 S2。

5. 句子聚类

   对句子集合 S1 和 S2 进行' 一个句子多个主题 "聚类,对 S1 执行聚类后,类别包括 S1.1, S1.2, …, S1.M,对 S2 执行聚类后,类别包括 S2.1, S2.2, …, S2.N。

6. 模式提取

从 S1 = {S1.1, S1.2, …, S1.M} 和 S2 = {S2.1, S2.2, …, S2.N} 的每个主题中分别提取字符表层模式和句法模式，并对应地标记为 P1 = {P1.1, P1.2, …, P1.M} 和 P2 = {P2.1, P2.2, …, P2.N}，其中 P1.i 和 P2.j 分别表示 P1 和 P2 中的某个模式集合。

7. 垂直聚类

如果 P1 = {P1.1, P1.2, …, P1.M} 中的某几个主题的模式集合具有较高的相似度，并且相似度超过阈值 V1，则进行垂直聚类；同样地，如果 P2 = {P2.1, P2.2, …, P2.N} 中的某几个主题的模式集合具有较高的相似度，并且超过阈值 V1，也进行垂直聚类。垂直聚类后，P1 和 P2 的模式集合将分别变为 P1′ = {P1′.1, P1′.2, …, P1′.m} 和 P2′ = {P2′.1, P2′.2, …, P2′.n}，其中 m 和 n 分别表示垂直聚类后的模式数目。

8. 水平聚类和答案模式识别

如果垂直聚类后的 P1′ = {P1′.1, P1′.2, …, P1′.m} 和垂直聚类后的 P2′ = {P2′.1, P2′.2, …, P2′.n} 中的某几个主题具有较高的相似度，且相似度超过阈值 H1，则进行水平聚类。水平聚类后的模式集合标记为 P = {P.1, P.2, …, P.K}，K 表示水平聚类后的模式数目。经过垂直聚类和水平聚类后，如果 P = {P.1, P.2, …, P.K} 中的某个模式集合 P.k 是由最多的原始模式集合组成的，则该模式集合 P.k 即为该类提问的答案模式集合。

9. 模式评测

使用文献[1]中的方法评价 P.k 中每个模式的准确率，以便在答案抽取阶段按照模式准确率从最高到底进行匹配。

算法输出：句法模式的部分实例

1.00：Q_FOCUS ←是→处女作→ANSWER

1.00：Q_FOCUS ←创作→ANSWER

1.00：Q_FOCUS ←成名作→ANSWER

0.50：Q_FOCUS ←有→作品→ANSWER

0.33：Q_FOCUS←是→代表作→ANSWER

2.1　主题划分

句子检索返回的 TopN 个结果实际上是一系列和提问相关的句子集合，这些句子或者围绕提问的不同侧面展开，或者描述和提问相关但不相同的主题。所以，提问的句子检索结果是由和提问相关的不同侧面的信息组成的，这些不同侧面信息分别代表一个主题。因此，我们应该把检索结果按照主题重新组织起来，这就是主题划分。主题划分后，我们假定同一个主题中的句子表达了相同的意思。如果主题划分的性能达到一定得水平，就可以在此基础上进行模式的抽取。可以看出，主题划分是本文模式抽取的基础。对此，本文提出了一种高效的主题划分方法“一个句子多个主题”，该方法的核心是根据候选答案进行主题划分。下面通过举例来说明。

例如提问 Q3：谁发明了电话？句子检索的结果及其所包含的候选答案参见表 1 所示。

表 1　句子检索结果及其包含的候选答案

编号	句子内容	句子涉及的主题
S51	1876 年 3 月 10 日贝尔发明电话	贝尔
S52	维·西门子发明了电机，贝尔发明电话，爱迪生发明电灯。	维·西门子 贝尔 爱迪生
S53	最近在纪念这一重要发明时，“移动电话之父”马丁·库珀再次成为公众焦点。	马丁·库珀
S54	1876 年，发明家贝尔发明了电话。	贝尔
S55	接着，1876 年，美国科学家贝尔发明了电话；1879 年美国科学家爱迪生发明了电灯。	贝尔 爱迪生
S56	1876 年 3 月 7 日，贝尔成为电话发明的专利人。	贝尔
S57	贝尔不仅发明了电话，还成功地建立了自己的公司推广电话。	贝尔
S58	在首只移动电话投入使用 30 年以后，其发明人库珀仍梦想着未来电话技术实现之日到来。	库珀
S59	库珀表示，消费者采纳移动电话的速度之快令他意外，但移动电话的普及率还没有达到无所不在，这让他有些失望。	库珀
S510	英国发明家斯蒂芬·福肖将移动电话的所有电子元件设计在一张纸一样厚薄的芯片上。	斯蒂芬·福肖

“一个句子多个主题”主题划分思想可以归纳为下面两点：

1. 如果一个句子包含 M 个不同候选答案，则该句可以属于描述了 M 个不同的类别。

例如，表 1 中的句子 S55 就描述了“贝尔发明电话”和“爱迪生发明电灯”两个不同的主题，属于“贝尔”和“爱迪生”两个类别。

2. 不同的句子，如果包含的候选答案实指同一个实体，则它们属于一个类别。

例如，表 1 中的句子 S54 和 S55 均包含主题“贝尔发明电话”，同属于“贝尔”这一类别。基于上

述思想,表 1 的主题划分结果如表 2 所示。

表 2 “一个句子多个主题”主题聚类结果示意

主 题	涉及主题的句子
贝尔	S51 S52 S54 S55 S56 S57 S58
维·西门子	S52
爱迪生	S52 S55
马丁·库珀/库珀	S53 S58 S59
斯蒂芬·福肖	S510

2.2 模式提取

模式的表示主要存在字符表层模式和句法模式两种形式。字符表层模式是根据词语在文本中出现的先后位置进行提取,主要存在两个缺点:(1)由于无法处理长距离的依存关系,导致模式在一定程度上和训练语料相关性大,缺乏泛化能力;(2)由于仅根据锚点词确定模式的长度,使得模式的完整性受到限制。例如,从提问‘孙中山是哪一年出生的?”聚类结果“孙中山/PER 名/Vg 文/Ng,/w 字/Vg 逸仙/PER,/w 1866 年/TIM 生于/v 广东香山县/LOC”中提取的模式‘Q_FOCUS 名 文,字 逸仙,ANSWER”就存在这两方面的问题。表 3 给出了字符表层模式的部分实例。

表 3 字符表层模式的部分实例

Q_FOCUS,为 元代 著名 的 戏曲家 ANSWER
ANSWER 的 代表 作品 Q_FOCUS
ANSWER 的 处女作,小说 Q_FOCUS
ANSWER 成名作:Q_FOCUS
ANSWER 的 小说 Q_FOCUS
Q_FOCUS 的 作者 ANSWER
……

表 4 句法模式的部分实例

Q_FOCUS ← 是 → 处女作 → ANSWER
Q_FOCUS ← 代表作 → ANSWER
Q_FOCUS ← 是→ 作品 → ANSWER
Q_FOCUS ← 小说 → ANSWER
……

为了避免字符表层模式的这一缺点,本文尝试在依存句法分析树的基础上进行答案模式的抽取,这就是句法模式。因此,依存句法分析不仅可以反映出句子中各成分之间的语义修饰关系,而且它可以获得长距离的搭配,跟句子成分的物理位置无关。本文使用的汉语依存句法分析工具仅仅给出词语之间存在依存关系,并没有具体指出关系的种类,其准确率在 80% 左右。表 4 给出了句法模式的部分实例,其中的箭头是从被依存词语指向依存词语。

2.3 垂直聚类

提问的答案存在多种不同的表述形式,例如 BIRTHDATE 类型的提问‘甘地是何时出生的?”,其答案的表述就包括:1869 年,1869 年 10 月,1869 年 10 月 2 日,一八六九年十月二日等等。所以,在有监督的机器学习算法中,用户必须提供该提问答案尽可能多的表述以提高答案模式的召回率。而本文提出的基于无监督汉语问答系统的问答模式学习算法是通过垂直聚类实现的。因为聚类的过程是在一个提问的检索结果内进行的,不涉及到其他提问实例,所以称之为垂直聚类。

垂直聚类是基于这样的假设:如果某个提问中的某几个聚类结果都包含该提问的答案,那么从这些聚类中抽取的模式集合应该存在一定程度的相似度,当相似度大于某个阈值 V1 时,应该合并这些相似的模式集合。

例如,从提问 Q4‘《悲惨世界》的作者是谁?”的某 2 个聚类结果中提取的模式集合分别如表 5 和 6 所示。

表 5 从提问‘《悲惨世界》的作者是谁?”中提取的第 i 个模式聚类集合

< ClusterNo > 雨果 </ClusterNo >
Q_FOCUS 改 自 法国 作家 ANSWER
Q_FOCUS ANSWER
Q_FOCUS 作者 :ANSWER
ANSWER 的 主要 作品 有 Q_FOCUS
ANSWER 的 小说 Q_FOCUS
ANSWER 代表作 Q_FOCUS
……

表 6 从提问‘《悲惨世界》的作者是谁?”中提取的第 j 个模式聚类集合

< ClusterNo > 维克多·雨果 </ ClusterNo >
ANSWER 的 长篇小说 Q_FOCUS
Q_FOCUS 是 法国 大 文豪 ANSWER
ANSWER 作品集 含 Q_FOCUS
Q_FOCUS ANSWER
……

很显然,这两个模式集合都是该提问类型的答案模式,该把它们聚为一类。本文采用的垂直聚类

相似度计算公式如公式(1) ~ (2)所示。

$$sim(VC_i, VC_j) = \sum sim(VC_{im}, VC_{jn}) \quad (1)$$

$$sim(VC_{im}, VC_{jn}) = \begin{cases} 1 & if\ VC_{im} = VC_{jn} \\ 0 & else \end{cases} \quad (2)$$

其中，VC_i和 VC_j分别表示第 i 和第 j 个模式集合；VC_{im}和 VC_{jn}分别表示第 i 个模式集合中的第 m 个模式和第 j 个模式集合中的第 n 个模式。

2.4　水平聚类和答案模式识别

通过垂直聚类可以获得关于某个提问的很多模式集合，但此时并不知道到底哪个模式集合才是提问答案的模式集合。对此，本文提出了水平聚类方法。因为聚类的过程是在不同提问的检索结果内进行的，所以称之为水平聚类。

水平聚类的基本思想是：从某类型提问的 2 个提问实例中学习到的所有模式集合中，如果存在提问 1 中的某个模式集合和提问 2 种的某个模式集合具有较高的相似度，即大于阈值 H1，则这两个模式集合应该合并。

经过垂直聚类和水平聚类后，如果某个模式集合是由最多的原始模式集合组成的，则该模式集合即是该类型提问答案的模式集合。

例如，从提问 Q4“《悲惨世界》的作者是谁？”的一个聚类结果中提取的模式集合如表 7 所示，从提问 Q5“《平凡的世界》是谁的作品？”的一个聚类结果中提取的模式集合如表 8 所示。

表 7　从 Q4 中提取的一个模式集合

< ClusterNo > 维克多·雨果 </ ClusterNo >
Q_FOCUS ANSWER
Q_FOCUS 作者 ：ANSWER
ANSWER 的 主要 作品 有 Q_FOCUS
ANSWER 的 小说 Q_FOCUS
ANSWER 代表作 Q_FOCUS
ANSWER 的 长篇小说 Q_FOCUS
ANSWER 作品集 含 Q_FOCUS
……
</ Cluster >

表 8　从 Q5 中提取的一个模式集合

< ClusterNo > 路遥 </ ClusterNo >
Q_FOCUS 作者 ：ANSWER
Q_FOCUS 作者 ANSWER
Q_FOCUS ANSWER
ANSWER 的 小说 Q_FOCUS
ANSWER 的 Q_FOCUS
ANSWER-Q_FOCUS
……
</ Cluster >

由于 Q4 和 Q5 是两个相同提问类型的提问实例，因此，如果这两个模式聚类集合都是各自提问的正确模式集合，则它们之间应该具有较高的相似度；如果有一个模式集合不是对应提问的模式集合，则它们的相似度会比较低。很显然，表 7 和表 8 都是 BOOKAUTHOR 类型提问的答案模式集合，应该合并。本文采用的水平聚类相似度计算公式如公式(3) ~ (4)所示。

$$sim(HC_i, HC_j) = \sum sim(HC_{im}, HC_{jn}) \quad (3)$$

$$sim(HC_{im}, HC_{jn}) = \begin{cases} 1 & if\ HC_{im} = HC_{jn} \\ 0 & else \end{cases} \quad (4)$$

其中，HC_i和 HC_j分别表示第 i 和第 j 个模式集合；HC_{im}和 HC_{jn}分别表示第 i 个模式集合中的第 m 个模式和第 j 个模式集合中的第 n 个模式。

3　试验结果与分析

本文主要针对下列提问类型（如表 9 所示）进行答案模式的自学习，并把学习到的模式应用于汉语问答系统中以验证模式的性能。表 9 中的训练提问和测试提问是从 EPCQA 评测平台[14]中随机抽取的，且只包括答案类型是命名实体[13]（人名、地名、机构名、时间词、数量词共 5 大类）的提问。其中，训练提问数是指无监督自学习时使用的提问实例个数，共 72 个；测试提问数是指测试阶段的提问实例个数，共 178 个。

表 9　学习模式的提问类型

提问类型	训练提问数	测试提问数	提问类型	训练提问数	测试提问数
INVENTOR	4	13	ADDRESS	7	51
BOOKAUTHOR	4	20	BIRTHPLACE	11	19
PERSONNICKNAME	3	13	BIRTHTIME	5	7

续表

提问类型	训练提问数	测试提问数	提问类型	训练提问数	测试提问数
OLDNAME	5	5	DEATHTIME	5	1
JOBPOSITION	7	3	HOLIDAY	7	7
LOCATIONNICKNAME	2	24	LENGTH	5	5
CAPITAL	3	5	POPULATION	4	5

3.1 基于模式匹配的答案提取系统

通过主题划分、模式抽取,垂直聚类和水平聚类后,本文提取的各类型提问的答案模式情况如表 10 所示。其中,SUP 和 SYP 分别表示字符表层模式和句法模式。

表 10 各类型提问答案模式的数量

提问类型	SUP	SYP	提问类型	SUP	SYP
INVENTOR	148	137	ADDRESS	68	83
BOOKAUTHOR	141	132	BIRTHPLACE	205	322
PERSONNICKNAME	134	153	BIRTHTIME	43	22
OLDNAME	150	94	DEATHTIME	18	13
JOBPOSITION	233	237	HOLIDAY	128	176
LOCATIONNICKNAME	107	31	LENGTH	77	144
CAPITAL	108	191	POPULATION	36	45
SUM	1596	1780			

需要说明的是,本文提出的无监督问答模式学习方法虽然不需要用户提供提问的答案,但用户给定的提问对模式的学习还是有影响的。所以在选择初始提问时,应该尽量避免选择比较生疏的提问实例。比如在学习 BIRTHTIME 类提问的答案模式,尽量选择比较著名的人作为模式学习的提问。对于模式学习用的提问数量,当然是越多越好。

接下来的实验是将学习到的各种类型提问答案的模式应用于汉语问答系统中,同时使用准确率评测指标(P)验证其性能,如公式(5)所示。

$$P = \frac{\text{系统正确回答的提问数}}{\text{所有待测试的提问数}} \times 100\% \quad (5)$$

3.1.1 基于检索的答案抽取系统

本文采用基于语言模型检索的句子检索算法[11]的答案抽取系统为 Baseline。表 11 给出了 Baseline 系统的准确率性能。

3.1.2 基于模式匹配的答案抽取系统

本实验的目的是为了验证基于无监督的模式学习算法的性能,并分别对字符表层模式(SUP)和句法模式(SYP)进行对比实验。表 12 是基于标准语言模型句子检索算法的两种模式答案抽取系统的准确率性能对比。

表 11 基于检索的答案抽取中文问答系统准确率性能

提问类型	准确率	提问类型	准确率
INVENTOR	30.8%	ADDRESS	52.6%
BOOKAUTHOR	55.0%	BIRTHPLACE	80.0%
PERSONNICKNAME	53.8%	BIRTHTIME	14.3%
OLDNAME	20.0%	DEATHTIME	0.00%
JOBPOSITION	66.7%	HOLIDAY	42.9%
LOCATIONNICKNAME	29.2%	LENGTH	0.00%
CAPITAL	13.7%	POPULATION	20.0%
SUM	32.6%		

对比表 11 和表 12 发现,相对于基于语言模型检索算法的答案抽取系统,基于字符表层模式的答案抽取系统性能和基于句法模式的答案抽取系统性能都得到了较大幅度的提高,其提高幅度分别约为 9.0% 和 14.0%。此外,基于句法模式的系统性能比字符表层模式的系统性能有所提高,其提高幅度约为 4.6%,这和 2.2 节分析是相符合的,即句法模式在应用于中文问答系统系统时,比字符表层模式

更适合。

表 12　基于 SUP 答案抽取系统和 SYP 答案抽取系统的性能对比

提问类型	SUP	SYP	提问类型	SUP	SYP
INVENTOR	15.4%	30.8%	ADDRESS	31.6%	21.1%
BOOKAUTHOR	60.0%	85.0%	BIRTHPLACE	80.0%	80.0%
PERSONNICKNAME	100.0%	84.6%	BIRTHTIME	14.3%	42.9%
OLDNAME	60.0%	60.0%	DEATHTIME	100.0%	100.0%
JOBPOSITION	66.7%	66.7%	HOLIDAY	100.0%	85.7%
LOCATIONNICKNAME	29.2%	37.5%	LENGTH	20.0%	40.0%
CAPITAL	33.3%	33.3%	POPULATION	20.0%	20.0%
SUM	41.6%	47.2%			

4　结论与展望

语言本身的灵活性和多变性常导致问答系统的提问和答案的不匹配。然而，从语义层面对这些这一现象进行分析到目前为止还是一件十分艰难的任务。所以，本文希望通过模式匹配技术解决这一问题。

对此，本文提出了一种基于无监督的学习算法从互联网中学习应用于汉语问答系统的问答模式。该方法和有监督机器学习算法的不同在于：无监督学习算法无需用户提供〈提问，答案〉对，只需用户对每种提问类型提供两个或以上的提问实例，算法即可通过 Web 检索、主题划分、模式提取、垂直聚类和水平聚类等步骤完成该类型提问的答案模式的学习。

在测试语料上的实验结果表明：本文提出的无监督问答模式自学习的方法是有效的，能够较大幅度地提高汉语问答系统的答案抽取性能。

然而，为了主题划分的方便，本文只对答案类型是命名实体类型提问的答案模式抽取进行了研究。对于答案类型为非命名实体的提问，例如提问‘ ftp 的中文全称是什么？”，将在下一步的工作中展开。

此外，垂直聚类和水平聚类是无监督学习算法的两大核心技术，通过它们可以自动过滤非问答模式的集合，提炼出提问答案的模式集合，但目前的聚类相似度算法(1) ~ (4)还过于简单，下一步工作将对这一部分进行更加深入的研究。

参考文献：

[1] Deepak Ravichandran, Eduard Hovy. Learning Surface Text Patterns for a Question Answering[A]. In：Proceeding of the ACL2002 Conference[C]. Philadelphia, PA, July, 2002.

[2] Dekang Lin, Patrick Pantel. Discovery of Inference Rules for Question Answering[J]. In：Natural Language Engineering, volume 7, 343-360.

[3] Hui Yang, Tat-Seng Chua. The Integration of Lexical Knowledge and External Resources for Question Answering[A]. In：the Eleventh Text REtrieval Conference [C]. Maryland：USA, 2002. 155-161.

[4] M. M. Soubbotin, S. M. Soubbotin. Use of Patterns for Detection of Likely Answer Strings：A Systematic Approach[A]. In：the Eleventh Text Retrieval Conference [C]. Gaithersburg, Maryland：November 2002.

[5] Moldovan, D., Harabagio, S., Girju, R., et al. LCC Tools for Question Answering[A]. NIST Special Publication：SP 500-251 The Eleventh Text Retrieval Conference [C].

[6] Yongping Du, Xuanjing Huang, Xin Li, Lide Wu. A Novel Pattern Learning Method for Open Domain Question Answering [A]. In：the Proceedings of IJCNLP2004 [C]. Sanya：China.

[7] Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin and Andrew Ng. Web Question Answering：Is More Always Better？[A] In：the Proceeding of 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Tampere, Finland, 2002.

[8] Dell Zhang, Wee Sun Lee. Web Based Pattern Mining and Matching Approach to Question Answering[A]. In: the Proceeding of TREC-11 [C]. Gaithersburg, MD, 2002.

[9] Regina Barzilay and Noemie Elhadad. Sentence Alignment for Monolingual Comparable Corpora[A]. In: the Proceedings of EMNLP2003 [C]. Sapporo, Japan. 25-32.

[10] Y. Shinyama, S. Sekine, K. Sudo, R. Grishman. Automatic Paraphrase Acquisition from News Articles[A]. In: the Proceedings of Human Language Technology Conference[C]. San Diego, USA, 2002.

[11] J. Ponte, W. Bruce Croft. A Language Modeling Approach to Information Retrieval[A]. In: the Proceedings of ACM SIGIR 1998[C]. 1998. 275-281.

[12] Youzheng Wu, Jun Zhao, Bo Xu. Chinese Named Entity Recognition Model Based on Multiple Features[A]. In: Proceedings of HLT/EMNLP 2005[C]. October 6-8, Vancouver, B. C., Canada. 427-434.

[13] Youzheng Wu, Jun Zhao, Bo Xu. Chinese Question Classification from Approach and Semantic View[A]. In: Proceedings of the 2nd Asia Information Retrieval Symposium (AIRS2005)[C]. LNCS 3689, Jeju Island, Korea. October 13-15, 2005. 485-490.

[14] 吴友政, 赵军, 段湘煜, 等. 构建汉语问答评测平台[A]. 第一届全国信息检索与内容安全学术会议[C]. 上海, 2004. 315-323.