基于模式学习的形式化答案抽取技术与置信度评价方法

李 鹏[1], 乔佩利[1], 王晓龙[2], 王宝勋[2]
(1. 哈尔滨理工大学计算机科学与技术学院, 黑龙江哈尔滨 150001;
2. 哈尔滨工业大学计算机科学与技术学院, 黑龙江哈尔滨 150001)

摘 要: 本文提出一种基于模式学习的形式化答案抽取方法, 区别于传统基于特征的答案抽取方法, 通过问题模式和答案模式的自动匹配, 直接获取问题答案. 本文通过机器学习的方法自动生成用于答案抽取的形式化模板, 克服了人工方法费时、费力以及覆盖率低等问题. 本文创造性地采用逻辑回归的方法对所学习到的模式进行置信度评价. 对比实验表明, 本文的方法取得了比较好的答案抽取效果. 本文方法实际应用于国际 TREC QA 评测, 评测结果证明本文的方法与传统基于特征的答案抽取方法具有很好的互补性.

关键词: 模式学习; 问答系统; 答案抽取; 置信度; 逻辑回归
中图分类号: TP391.2 文献标识码: A 文章编号: 0372-2112 (2008) 12-2339-05

Formalized Answer Extraction Based on Pattern Learning and Confidence Estimation

LI Peng[1], QIAO Pei-li[1], WANG Xiao-long[2], WANG Bao-xun[2]
(1. *College of Computer Science and Technology, Harbin University of Science and Technology. Heilongjiang, Harbin* 150001, *China*;
2. *School of Computer Science and Technology, Harbin Institute of Technology, Heilongjiang, Harbin* 150001, *China*)

Abstract: This paper presents a method of formalized answer extraction based on pattern learning. Our method extracts answers directly by matching the question patterns and the answer patterns automatically, which is different from the other traditional feature-based ones. The method gains formalized patterns automatically by using a machine learning strategy. Therefore, huge work load and low coverage of manually methods are avoided. Creatively, we apply logistic regression to estimate the confidence of the learned patterns. The contrast experiments give our method encouraging results, and it achieves fair and objective test in TREC QA track, the result of which shows that the method is a supplementary to the traditional feature based answer-extracting approaches.

Key words: pattern learning; question answering system; answer extraction; confidence; logistic regression

1 引言

答案抽取是信息抽取研究的一个子领域, 也是问答系统的核心组成部分, 它是问答系统区别于通常意义下文本检索系统的标志[1]. 随着多年来检索技术的发展与提高, 答案抽取技术已经成为影响问答系统最终效果的决定性因素. 因此, 答案抽取技术的研究越来越受到研究人员的重视, 而基于特征的排序或分类的方法成为了近几年来答案抽取技术的主流, 如: 神经网络[2]、最大熵[3]、支持向量机[4]、逻辑回归[5]等. 但是, 由于自然语言处理中语义处理技术的进展比较缓慢, 因此语义特征没有纳入到特征体系中, 使得基于特征的答案抽取效果达到了一个瓶颈. 如何在现有技术水平下提高答案抽取的准确率是问答系统研究的核心问题. 从理论上对比分析, 基于模式的形式化答案抽取与基于特征的答案抽取从原理上存在本质的区别, 这种方法上的差异使得两种答案抽取方法存在互补的可能性大大提高.

基于模式的形式化答案抽取方法通过寻找答案模式直接抽取完整的问题答案, 是最自然的答案抽取方法[6]. 这种答案抽取方法最早采用人工的方法进行模式创建[7]. 人工方法的优点是所构建的模式抽取准确率非常高, 但其缺点也十分突出, 人工构建模式库需要消耗

收稿日期: 2007-06-18; 修回日期: 2008-07-23
基金项目: 国家自然科学基金重点项目(No. 60435020); 国家 863 高技术研究发展计划重点项目(No. 2006AA01Z197)

巨大的人力资源, 费时、费力, 并且问题覆盖率低, 适应性差. 因此, 采用机器学习的方法自动构建用于抽取答案的形式化模板是形式化抽取方法的发展方向, 近年来也成为答案抽取技术的研究热点[8]. 但遗憾的是, 基于模式的形式化答案抽取在效果依然没有超越基于特征的抽取方法. 本文分析主要存在以下几点缺陷: (1) 问题模式的覆盖度太低; (2) 模式标记的不可靠性; (3) 模式置信度难以评价.

2 基于模式学习的形式化答案抽取技术

2.1 模式标记与分析

采用基于模式的抽取方法就要制定一套模式标记, 以便生成形式化模板. 目前的形式化答案抽取系统中问题模式和答案模式大多采用统一的标记集合, 并且大多使用强约束标记[8]. 这种标记集合的特点是标记种类比较多, 其中还包含语义信息标记. 从理论上讲, 这种标记体系适应基于模式的形式化抽取方法, 它可以提高模式的确定性和准确性, 是先进的模式标记体系. 但遗憾的是, 以现在自然语言处理领域许多基础性研究的发展水平来看, 采用这种强约束标记体系反倒会使系统性能下降, 这种现象在文献[6]中初步论述和验证.

本文从两个方面来分析这种现象发生的原因. 一方面, 标记本身的识别困难是造成这种现象的一大原因. 在强约束标记体系中存在一些语义信息标记, 如: 组块标记、语义成分标记等. 然而, 在自然语言处理技术目前的发展水平下, 这些语义信息识别的准确率并不足以达到应用的水平, 错误率比较高. 标记本身的识别错误会造成模式学习过程中构成错误的模式形式, 使模式本身就出现错误. 因此, 本文中使用的标记都是在实际应用中识别准确率比较高的可靠标记, 如词性(POS)、名词短语等. 其目的是使模式产生标记错误的可能性尽可能降低.

另一方面, 更为重要的是在形式化答案抽取中, 问题标记成分的识别与答案标记成分的识别难度存在比较大的差距. 两者相比之下, 问题标记的识别更加困难. 这是由于两者不同的语言环境造成的巨大分布差异. 答案所处的语言环境通常是普通的文本语言环境, 而代表问题的问句是一个比较独立的语言环境, 其特点是信息量少, 语言成分比较多, 这种差异造成了问题标记与答案标记识别的不同分布性. 众所周知, 目前绝大多数主流的识别工具都是采用机器学习的方法进行训练, 训练语料大多采用 Penn Treebank 一类已经标注好的大量文本资源, 这种识别工具对与其训练语料同分布的文本识别效果最好, 不同分布的文本识别效果将会大打折扣. 例如, 命名实体标记的识别, 它在答案所处的文本环境中的识别准确率是比较高的, 而在问句中的识别效果最多相当于查表的方法, 识别工具的其他特征(如上下文特征)根本不起作用, 甚至会起到相反作用. 由于这种识别效果的巨大差异, 使得问题和答案中本是同一个标记成分, 却识别成不同的两种标记(即使在信息抽取中经常应用的"将错就错"的策略也失去了意义), 这是影响形式化答案抽取效果非常重要的因素. 本文对问题模式和答案模式采用不同的标记集合来解决这个问题.

2.2 模式学习与答案抽取

本文用于答案抽取的所有模式都采用机器学习的方法自动获取, 用于训练的语料使用一些基本识别工具自动进行标注, 这样避免了使用大量的人力和专家资源, 也使得构建比较全面大量的训练语料库成为可能, 解决了形式化答案抽取学习语料匮乏的问题. 模式学习与评价流程如图 1 所示.

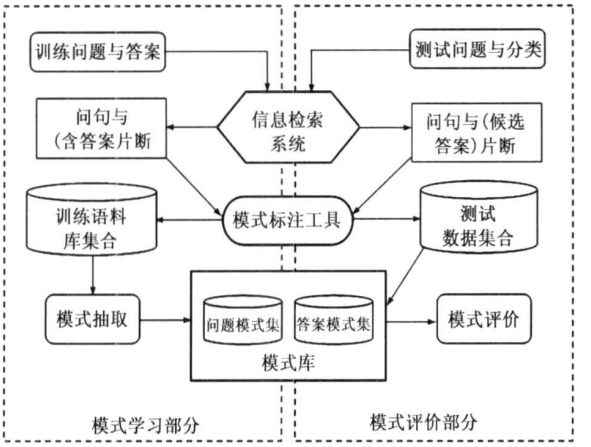

图1 模式学习与评价流程图

模式学习步骤:

(1) 检索查询: 对用于训练的问题, 将问句中的关键词和答案联合送入检索系统(如: Google, AskJeeve 等)进行查询, 抽取包含问题答案的句子片断

(2) 语料库构建: 对用于训练的问句和检索到的句子片断(包含答案)使用标注工具进行模式标记的标注, 最终形成用于学习形式化模板所需要的语料库

(3) 模式抽取: 对训练语料进行泛化处理. 问句用问题标记进行替换, 而包含答案的句子片断采用答案标记进行替换. 泛化处理后生成问题模式库和相对应的答案模式库, 所有问题模式按基于答案类型的分类体系进行类别划分, 每个问题模式对应所学到的答案模式.

模式评价步骤:

(1) 检索查询: 对用于测试的问题, 将问句中的关键词与候选答案类型联合送入检索系统(如: Google, AskJeeve 等)进行查询, 抽取包含问题候选类型答案的句子片断

(2) 模式标注: 对用于测试的问句和检索到的句子片断(包含候选答案)使用标注工具进行模式标记的标

注, 并经泛化处理形成与问题模式和答案模式相对应的形式化模板

(3) 答案抽取: 将泛化处理后的测试问题模式与学习到的问题模式库中的问题模式进行匹配, 匹配到的问题模式将其相对应的答案模式与测试问题的候选答案模式相匹配, 从而进行形式化抽取答案.

(4) 模式评价: 计算答案模式的覆盖度与准确率, 通过逻辑回归的方法进行模式置信度的综合评价.

3　基于逻辑回归的模式置信度评价方法

3.1　答案模式的覆盖度与准确率计算

模式的覆盖度是模式的一个重要指标, 它代表了一个模式实际应用中适用范围的大小. 在文献[6]中把覆盖度作为模式性能的唯一评价标准, 并采用人工的方法进行评价. 本文认为这种评价具有片面性, 一个模式的覆盖度大只能代表它适用范围大, 比较符合人类常用的表达方式, 但不能代表它一定准确; 相反, 若某一个模式虽不常见, 但很可能只要出现抽取的答案就很准确. 所以, 不能仅从覆盖度来判断模式置信度的大小. 因此, 本文认为覆盖度仅仅是模式性能的一个重要指标之一, 它不能表示模式的准确性.

本文采用自动的方法来计算模式覆盖度, 克服人工方法所消耗的巨大人力资源, 使大规模模式学习与评价成为可能. 通过统计训练语料中每一个问题模式所对应的答案模式在抽取时所出现的次数, 再进行归一化计算就可以得到每个答案模式的覆盖度.

设问题模式 Q_j 具有 K 个答案模式 $P_i(i=1, 2, \cdots, K)$, 而 $N(P_i)$ 代表第 i 个模式训练中出现的次数, 则答案模式 P_i 的覆盖度 $F(P_i)$ 的计算公式为

$$F(P_i)=\frac{N(P_i)}{\sum_{i=1}^{K}N(P_i)} \tag{1}$$

模式的准确率是模式的另一个重要指标, 它代表了一个模式抽取答案的准确性. 在文献[8]中把准确率作为模式性能的唯一评价标准, 我们认为这种评价同样不全面. 我们举个例子来说明, 假设有两个答案模式, 在实际测试中模式 1 应用次数为两次, 对、错比例为 1∶1, 而模式 2 应用次数为 20 次, 对、错比例为 10∶10. 两种模式仅仅从准确率来计算是完全相等, 但直觉告诉我们这两个模式应该有很大差异. 这正是忽略了模式覆盖度的影响仅考虑准确率因素而造成的偏差

设 P_i 是问题模式 Q_j 的一个答案模式, 在测试数据下进行统计, 模式 P_i 被匹配的次数为 $m(P_i)$, 其中匹配抽取到正确答案的次数 $n(P_i)$, 则答案模式 P_i 的准确率 $Z(P_i)$ 为

$$Z(P_i)=\frac{n(P_i)}{m(P_i)} \tag{2}$$

3.2　基于逻辑回归的置信度评价方法

模式置信度是对模式可信程度的评价, 它代表了模式的整体综合性能. 本文采用逻辑回归的方法综合考虑模式覆盖度与准确率两个方面计算模式的置信度.

回归分析是统计学中最重要的分支学科之一, 其研究的主要问题就是如何利用两个变量 X、Y 的观察值(样本)来确定它们之间的内在关联. 考虑传统的多元线性回归分析, 响应变量(response variable) Y 和自变量(explanatory variables) X 之间的关系可用下列方程来描述:

$$y=\alpha+\beta_1 x_1+\cdots+\beta_p x_p \tag{3}$$

式中 α, β_1, …, β_p 是回归系数; x 为 n 组观测值 x_{ip} 的数据矩阵, y 为由观测值 y_i 组成的向量. 对于二分类因变量的回归估计($Y\in[0, 1]$), 使用该方程计算时, 常会出现 $Y>1$ 或 $Y<0$ 的不合理情形, 因此, 对二分类因变量的分析应采用非线性函数. 为此, 对 Y 作对数单位转换, 即取 $\text{logit}\,Y=\ln(Y/1-Y)$ 作为应变量, 得到

$$\text{logit}\,Y=\ln(Y/1-Y)=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k \tag{4}$$

这样 $\ln(Y/1-Y)$ 可取负无穷到正无穷的任何数值, 而 Y 的值则限制在 0 到 1 之间, 上式经过变换后得到 Logistic 回归方程:

$$Y=\frac{\exp(\alpha+\beta_1 x_1+\cdots+\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\cdots+\beta_k x_k)} \tag{5}$$

该公式即为 logistic 回归公式, 是普通多元线性回归的推广. 但是与线性回归不同, logistic 回归是一种非线性模型, 因而回归系数的估计通常采用最大似然估计法. 给定一个由 n 个样本组成的集合, 其应变量观测值为 y_1, …, y_n, 设 $p_i=P(y_i=1/x_i)$ 为给定 x_i 的条件下得到结果 $y_i=1$ 的条件概率; 而同样条件下得到结果 $y_i=0$ 的条件概率为 $P(y_i=0/x_i)=1-p_i$, 则 n 个观测的似然函数可以表示为各个边际分布的乘积:

$$L(\theta)=\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{(1-y_i)} \tag{6}$$

最大似然估计就是使这一似然函数的值最大的基础上对参数 α 和 β 的值的估计. 使似然函数 $L(\theta)$ 最大化的实际过程就是一个迭代计算的过程, 当迭代到情况改善得很小时, 即第 k 步和第 $k+1$ 步的情况基本一致时, 迭代停止. 模型自动确定了输入特征的权值, 即这样的权值配备能够有效地提高模型的预测能力. 可以证明, 在随机样本条件下, logistic 模型的最大似然估计具有一致性、渐进有效性、和渐进正态性.

在模式置信度评价中, 模式覆盖度和准确率作为回归分析中的两个响应变量, 由逻辑回归公式(5)得出模式的置信度(Pattern Confidence, $C(P_i)$)的计算公式:

$$C(P_i)=\frac{\exp(\theta+\omega_1 F(P_i)+\omega_2 Z(P_i))}{1+\exp(\theta+\omega_1 F(P_i)+\omega_2 Z(P_i))} \tag{7}$$

其中, 公式中的权值 θ, ω_1 以及 ω_2 通过极大似然估计

的方法由公式(5)、(6)通过计算软件(如, SPSS)计算得出.

4 实验与性能分析

4.1 对比实验与性能分析

本文为了验证基于模式学习的形式化答案抽取技术的有效性, 采用对比实验的方法, 将本文的方法与一些主流的基于特征的方法进行对比. 用于对比实验的数据集是包含所有问题类别的 500 个基于实例的问题, 它们主要是由近几年来 TREC QA 评测中的 Factoid 问题和来自其它途径相类似的问题经过一定筛选所组成(各个问题类别具体组成见表 1). 每个问题对应 50 个相关的检索文档, 并保证其中至少有一个文档中包含有正确答案(这主要是保证我们的对比实验不受除了答案抽取以外, 问答系统其它组成部分的影响).

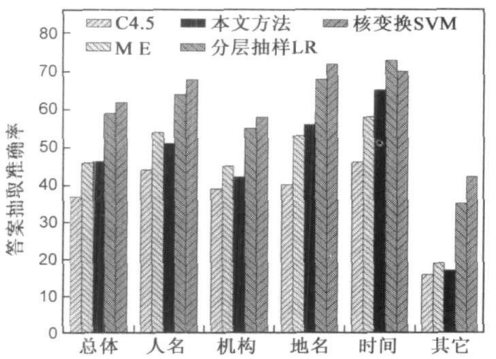

图2 对比实验结果

表 1 形式化答案抽取实验结果

问题数	问题类型	问题覆盖度		答案抽取准确率	
100	人名	64	64%	51	79.7%
100	地名	67	67%	56	83.6%
100	机构	54	54%	42	77.8%
100	时间	72	72%	65	90.3%
100	其它类别	29	29%	17	58.6%
500	总体	286	57.2%	231/286 (231/500)	80.8% (46.2%)

从以上的实验数据可以看出, 本文的方法在答案抽取的应用中取得了比较好的效果. 在与 C4.5、最大熵(ME)、分层抽样逻辑回归以及核变换 SVM 这四种方法的比较结果显示, 本文的方法优于前两者, 不如后两者. 但是, 后两种方法十分复杂, 而本文的方法简单自然, 并且在效果上优于普通的基于特征的答案抽取方法.

从表 1 可以看出, 本文的方法有其自身的特点, 首先是 57.2% 的问题覆盖度虽然在同类形式化答案抽取的报道中属于比较高的成绩, 但依然是影响最终抽取效果的最大因素, 并且各种不同的问题类型, 问题覆盖度差异比较大, 分布不均衡. 其次, 从答案抽取的准确率来看, 只要问题模式相匹配, 则准确率还是比较高, 达到了 80.8%. 最后, 某些问题类型的抽取效果还比较低, 这说明形式化答案抽取对一些问题类别的答案抽取效果还有待进一步提高.

4.2 实际参加 TREC QA 评测结果

本文的方法应用于 InsunQA05 和 InsunQA06 两个系统中作为答案抽取模块的组成部分. 两个问答系统中采用了不同的基于特征的答案抽取方法, InsunQA05 系统采用的是基于分层抽样逻辑回归的答案抽取方法, InsunQA06 系统采用的是基于核变换的 SVM 答案抽取方法, 但两个系统都采用了本文的形式化答案抽取方法. 我们对两年评测的数据进行了自评, 以检验两种答案抽取方法在系统中发挥的作用与效果. 由于 TREC QA 任务采用的是十分严格的封闭式评测, 并且评价标准复杂. 因此, 我们的赛后自评结果与真实评测中官方成绩略有差异, 自评结果如表 2 所示.

表 2 TREC QA 任务自评测试结果

	TREC 2005 QA 任务测试集			TREC 2006 QA 任务测试集		
答案抽取方法	聚类分层抽样逻辑回归	基于模式学习的形式化答案抽取	两种方法融合	基于重采样与核变换的 SVM	基于模式学习的形式化答案抽取	两种方法融合
答案抽取准确率	26.52%	17.96%	32.87%	28.04%	19.85%	35.24%

5 结论

本文提出了一种基于模式学习的形式化答案抽取方法, 实现了形式化模板的自动学习与获取, 避免了手工制定模式费时、费力, 适应性差以及覆盖度低等缺陷. 本文通过采用可靠的模式标记, 并针对问题模式与答案模式提出了两种不同的标记集合, 提高了问题模式的覆盖度. 本文创造性地采用逻辑回归的方法从模式的覆盖度和准确率两方面综合评价模式的置信度. 实验结果表明本文的方法简单而有效, 特别是在问题模式匹配上的情况下, 可以达到比较高的抽取准确率, 并且本文的答案抽取方法与传统基于特征的答案抽取方法具有很好的互补性.

参考文献:

[1] John O' Connor. Retrieval of answer sentences and answer-figures from papers by text searching[J]. Information Processing & Management, 1975, 11(5/7): 155 – 164.

[2] Marius A Pasca. High-performance, open-domain question answering from large text collections[D]. USA: University of Southern Methodist, 2001.

[3] Abraham Ittycheriah. Trainable question answering systems

[D]. USA: The State University of New Jersey, 2001.

[4] Jun Suzuki, Yutaka Sasaki, Eisaku Maeda. SVM answer selection for open-domain question answering[A]. 19th International Conference on Computational Linguistics (Coling-2002) [C]. Taipei: Howard International House, 2002. 974- 980.

[5] Peng Li, Yi Guan, Xiao-long Wang. Answer extraction based on system similarity model and stratified sampling logistic regression in rare data [J]. International Journal of Computer Science and Network Security, 2006, 6(3): 189- 196.

[6] Glenda Anaya. ANSFORM: Answer formulation for question-answering [D]. Canada: Concordia University of Canada, 2002.

[7] Soubbotin M and Soubbotin S. Patterns of potential answer expressions as clues to the right answers [A]. In Proceedings of the Tenth Text Retrieval Conference (TREC 10) [C]. Maryland: IEEE Press, 2001. 293- 302.

[8] 杜永萍, 黄萱菁, 吴立德. 模式学习在 QA 系统中的有效实现[J]. 计算机研究与发展, 2006, 43(3): 449- 455.
Du Yongping, Huang Xuanjing, Wu Lide. Effectively implementing a pattern learning method in the question answering system [J]. Journal of Computer Research and Development, 2006, 43(3): 449- 455. (in Chinese)

作者简介:

李　鹏　男, 1978 年生于黑龙江哈尔滨. 哈尔滨理工大学计算机科学与技术学院 副教授. 2007 年获得哈尔滨工业大学计算机科学与技术学院博士学位. 研究方向为问答系统、机器学习、网络信息处理.
E-mail: pli@insun.hit.edu.cn

王晓龙　男, 1955 年生于黑龙江哈尔滨. 哈尔滨工业大学计算机科学与技术学院教授, 博士生导师. 研究方向为人工智能、机器学习、自然语言处理.

乔佩利　男, 1951 年生于黑龙江哈尔滨. 哈尔滨理工大学计算机科学与技术学院教授, 院长. 1974 年毕业于上海复旦大学计算机专业. 研究方向为网络与信息安全、生产计划调度与智能算法、信息处理等.