文章编号:1001 - 9081(2008)06 - 1575 - 03

基于互联网和 self-training 的中文问答模式学习

李志圣,孙越恒,何丕廉,侯越先
(天津大学 计算机科学与技术学院,天津 300072)
(lzs_jeff@tom.com)

摘 要:在已有的问答模式学习中,模式定义和候选答案评分偏于简单,而且学习过程依赖于人工标定语料。通过挖掘 Web 文本中动、名词序列的骨架模式,用以扩充模式定义;将 self-training 学习机制引入问答模式学习:用一对训练语料进行初始学习,通过互联网搜索,自动选择可靠程度较高的问答对,重新训练;扩充了启发规则,改进候选答案的评分方法。实验结果表明:所提出的问答模式学习方法能有效地提高中文问答系统的性能。

关键词:互联网;问答模式;self-training;机器学习
中图分类号: TP301 **文献标志码**:A

Chinese question answering pattern learning based on self-training mechanism and Web

LI Zhi-sheng, SUN Yue-heng, HE Pi-lian, HOU Yue-xian
(*College of Computer Science and Technology, Tianjin University, Tianjin* 300072, *China*)

Abstract: In the past, the learning for QA pattern relies on the labeled data, and the definition of pattern and the scoring method for the candidate answers are over simplified. The verb and noun sequence was extracted as the skeleton pattern to expand definition of QA pattern. In the learning process, a learning mechanism was established based on self-training. At first, the initial study was completed on a labeled QA pair, then the system would automatically select the reliable data for self training through searching in the Web while the system was running. The scoring method of the candidate answers was also improved by applying several heuristic rules. The experimental results show that the performance of Chinese QA system based on our method is improved significantly.

Key words: Web; QA pattern; self-training; machine learning

0 引言

由于自然语言处理的底层技术尚不成熟,从语义层面处理语言的灵活性和多变性是一件十分艰难的任务[1]。因此,以模式匹配技术为代表的基于字符表层的文本分析技术在问答系统中得到广泛应用,代表性工作有:

文献[2]采用人工编写规则的方法得到问答模式。人工编写规则的方法代价昂贵、模式扩充困难,因此,用机器学习的方法获取问答模式就成为研究的重点。

文献[3-5]等通过有监督学习从 Web 文本中自动提取答案模式。模式定义可简要描述为:在包含问题词和答案的实例文本串中,替换问题词和答案,形成类似于"前缀串 Q 中间串 A 后缀串"的模式(后文称为"朴素模式")。例如:对发明人类型问题,可能的模式有:"The Q was invented by A"。上述方法使用用户提供的〈问题词,答案〉对作为训练语料进行 Web 搜索。存在的问题是:朴素模式较难处理长距离语义关系以及缺乏良好的泛化性。

文献[1]提出由用户提供关于某问答类型的两个问题实例,依靠聚类方法获取该类问答的模式。该方法的实质是利用少量标注信息,进行半监督学习,它是对传统有监督学习方法的突破,但是,系统性能取决于人工提供的两个实例,缺乏进一步泛化的能力。

本文认为以往的问答模式学习在模式定义和学习机制上还存在改进的空间:

1)朴素模式简单直接,但是存在泛化能力弱、不能反映长距离语义关系和待学习模式空间过大等问题。虽然可以通过引入命名实体识别和语义分析技术来简化模式的目标空间,但将导致计算量的增加,同时,因语义分析技术本身不成熟,将导致错误率上升。

在 2.2 节中,本文尝试以文本中的动名词序列"骨架"为基础,定义新的问答模式,在模式的泛化能力和计算复杂性之间找到良好的折中。

2) self-training 方法[7]是新近出现的一种半监督学习方法,核心思想是先利用少量的人工标定数据,估计出系统初始的参数,系统在实际运行时,如果发现与人工标定数据相似度较高的未标定数据,系统将其作为"自动标定"数据加入到训练集中,重新训练,从而改进系统性能。self-training 方法在实体间语义关系学习[8]和主语常用模式学习[9]上均收到好的效果。

本文将 self-training 方法引入问答模式学习过程:通过少量的标定语料,获得初始的模式集,然后从互联网中挖掘可靠实例,进行重训练,从而扩充模式集、改进系统性能。

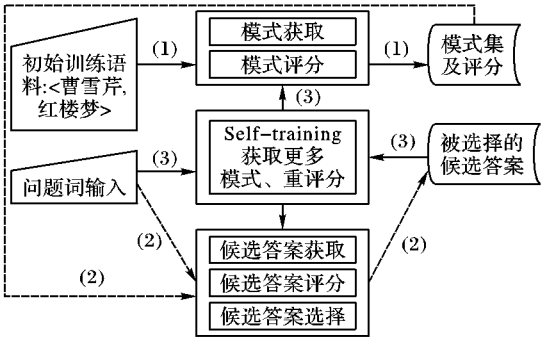
收稿日期:2007 - 12 - 14。
基金项目:国家自然科学基金资助项目(60603027);天津市应用基础研究计划资助项目(05YFJMJC11700)。
作者简介:李志圣(1977 -),男,江西上高人,博士研究生,主要研究方向:信息检索; 孙越恒(1974 -),男,山东烟台人,讲师,主要研究方向:信息检索、文本分类; 何丕廉(1943 -),男,天津人,教授,博士生导师,主要研究方向:人工智能、计算机辅助教育; 侯越先(1972 -),男,天津人,副教授,主要研究方向:机器学习、维数约简。

1 基于 self-training 方法的问答系统

下面以作者类型的问答模式获取为例,说明本系统的方法。

1.1 学习流程

学习流程如图1所示。



注:(1)初始训练,获得模式集;(2)获取候选答案及其评分;(3)重训练。

图1 学习流程

1.2 模式的学习

1.2.1 模式的定义

文献[6]发现在政治文本领域,敏感词汇的上下文呈现出某种积聚特性:敏感词汇总是和一些特定词汇联合出现,而与其他词汇无关。特定词汇一般以动、名词为主。这些特定词汇序列被获取后,将形成上下文的"骨架",实际上就是关于敏感词汇的上下文的合理模式。以"骨架"定义模式,具有简洁、泛化能力强、获取难度小等优良特性。

与之类似,问答模式学习中也可以获取问答文本之中的动、名词,形成"骨架"模式。但尚需考虑到有部分文本依靠非中文符号来表示问答语义。例如:"吴贯中 - - - 三国演义",用符号串" - - - "来指征语义关系。为此,本文定义了如下两类模式:

1)第一类模式是在中文文本中获取的动名词骨架模式,其构成为:

骨架模式::= 左端词汇 A 中间词汇序列 Q 右端词汇
|左端词汇 Q 中间词汇序列 A 右端词汇

左端词汇::= 动词|名词|*

右端词汇::= 动词|名词|*

中间词汇序列::=(动词|名词)中间词汇序列|动词|名词|*

其中:"|"意为或;"*"表示匹配任何词汇,它适合规定位置上无动词或者名词的情形。

2)第二类模式为第1节中定义的朴素模式。

1.2.2 模式获取和评分

以问题词 Q 和答案词 A 作为输入,获取模式串和评分的方法如下:

1)向百度提交问答对"Q A";

2)在返回的摘要中,提取出包含 Q 和 A 的句子,并除去末尾的分隔符。

3)在每个句子中,先抽取位于 Q 与 A 之前的文本串 T_1、之后的文本串 T_2、中间的文本串 T_3。用中科院计算所的 ICTCLAS 程序,对句子进行分词和词性标注,按照以下两种情形抽取模式:

a)若句子中不存在非中文符号,则获取 T_1 最末端的一个动词或名词,T_2 中最前端的一个动词或名词和 T_3 中所有动、名词。若 T_1 和 T_2 内不含有动词或名词,则在模式中记为"*"。依照第一类模式的定义进行存储。

b)若句子中存在非中文符号,则将 T_1、T_2 和 T_3 按第二类模式存储;

4)模式评分:若当前获取的模式已经存在,那么评分加1,如果不存在,那么评分设置为1。该评分方法基于一个观察:模式发生频度越大,则越可靠。

在实际训练时,先提供一个出现在较多页面中的问答对:"红楼梦 曹雪芹",提交到百度。表1中列出了获取的模式串。可以看到:模式:"* Q 作者 A *"的评分比"Q [清] A"的评分高得多。这表明:前者是最常见的模式,可信度较高。

训练后,获取的骨架模式有321个,朴素模式有113个。

表1 以"红楼梦 曹雪芹"为训练语料获取的部分模式串及评分

模式类型	模式	评分
骨架模式	* Q 作者 A *	96
骨架模式	* Q 作者是 A *	23
骨架模式	* A 是 中国 名著 Q 作者	1
朴素模式	Q《 A	39
朴素模式	Q [清] A	1

1.3 候选答案的查找和评分

当用户输入问题词 Q 后,系统将按如下方法查找候选答案,并进行评分:

1)向百度提交"Q + 关键词"的查询。

关键词是在模式集的动名词集中取频度最大的词。例如:在初始学习后,作者类问题的关键词为"作者"。利用上述查询方式,有助于提高查询结果的相关性。

2)在返回摘要中,寻找所有与模式集中模式匹配的答案候选。

3)对所有候选进行评分。

以往的文献,都基于候选答案频度进行评分。但由于网页噪声较大,应该发掘尽可能多的特征来改进候选的评分。本文在基于频度的启发规则外,扩充了3条启发规则,形成如下的规则集:

规则1 高频度的候选答案应比低频度的候选答案更可靠。

例如:表2第二行中,"曹雪芹"的频度比其他候选频度高很多。

规则2 正确的候选答案,通常在候选集中与较多的其他候选相似。

为简化计算,按如下方法定义候选的相似性:

如果候选 E 是候选 F 的子串,那么称 E 和 F 互为相似候选。例如,表2中与"曹雪芹"相似的候选有"是曹雪芹"、"不是曹雪芹"、"曹雪芹的身世"等。

表2 "红楼梦"作者的部分候选答案(查询百度前十个返回页面)

候选答案	出现频度	相似候选频度	模式数量
曹雪芹	47	325	7
是曹雪芹	2	106	2
不是曹雪芹	9	84	3
曹雪芹的身世	12	50	1
的朝代	11	0	1
介绍	11	3	1
之谜	3	6	1

规则 3　出现在更多模式中的候选答案,应该比出现在一个模式中的答案更可靠。

原因是正确的答案应该出现在多个网页中。由于不同的网页由不同的作者写出,它们具有较多模式。而不正确的候选通常来源于特定的网页,它们具有较少模式。表 2 第三行中,“曹雪芹”出现在 7 个模式中,错误的候选,如:“的朝代”,“介绍”和“之谜”仅仅出现在 1 个模式中。

规则 4　出现在具有较高评分的模式中的候选答案,应该比出现在较低评分的模式中答案更可靠。该规则在 2.2.2 节中已提及。

文献[9]在利用模式抽取基本名词词组时,采用的评分函数是对与候选同现的所有模式评分进行加总。与此类似,本文采用如下公式计算模式对候选答案的影响度:

$$score(F) = \sum_{i}^{N} (score(p_i)) \quad (1)$$

其中:F 为某候选答案,若 T_i 为 F 在第 i 次出现时的文本,则 p_i 指代 F 在 T_i 中的模式串,N 为 F 出现的总次数。

规则 2 指出:相似候选词条的评分对候选答案的评分有增益作用,为此扩充公式(1) 如下:

$$score(F) = \sum_{i}^{N} (score(p_i)) + \lambda \cdot score(F\_similar) \quad (2)$$

其中,λ 为相似候选答案评分的增益系数,它应介于 0 和 1 之间,实际设为 0.5。

1.4　self-training

1)当用户提交某个问题词 Q 进行查询时,在获得的候选答案集中,如果评分最高的候选答案,其分值比评分次高的候选答案超出 m 倍(阈值 m 实际被设为 3),那么该候选答案被认为是可靠答案。

2)使用该候选答案和源问题词作为训练对,根据 2.2.2 节的方法,重新训练,获取更多的模式,更新模式评分。

2　实验结果

2.1　自训练集与测试集

由于中文问答系统的评测缺乏统一的数据平台,因此,以往文献都缺乏与同行工作的对比。本文通过以下方法,对作者型问答寻求一个公开的数据集:

在蔚蓝网络书店的搜索页面(http://search.wl.cn/),在“出版社”后填入“人民出版社”,“出版时间”后填入“200501 至 200512”。选择排列方式为:按“上架时间”排序。在返回书目中,选择前 20 个为自训练集,后 20 个为测试集。

对于其他问答,本文在百度上采用对应的关键字搜索,然后在返回页面中随机挑选 20 对自训练集,20 对测试集。

2.2 节在作者型问答上,测试 self-training 机制和骨架模式对系统性能的改进效果,2.3 节针对更多问答类型,分析在自训练集上完成 self-training 后的系统的准确率,并与文献[1,10]作参照比较。

2.2　self-training 和骨架模式的作用

在作者型问题中,先使用“红楼梦 曹雪芹”进行初始训练,部分模式和评分列在表 1。

在 self-training 阶段,系统对自训练集中的 20 个书名,自动在互联网上查找答案,有 5 个候选被自动选择进行重训练。这 5 次选择都是正确的,这表明:选择方法是可靠的。

经过重训练后,得到了 543 个模式,评分被重新计算。由表 1 和表 3 对比可看出:表 1 中高分的模式在表 3 中评分有所增加,而噪声模式在重训练过程后,维持低分不变。

表 3　self-training 后的部分模式串及评分

模式	评分
* Q 作者 A *	110
* Q 作者是 A *	25
* A 是 中国 名著 Q 作者	1
Q《 A	170
Q[清] A	1

使用骨架模式对作者型问答的 20 对测试语料进行评测,准确率为 55%,而使用朴素模式进行评测,准确率为 35%,这说明:骨架模式在答案的定位和泛化能力方面都较朴素模式更出色。

2.3　准确率

文献[1]研究基于句法模式的问答学习系统,文献[10]研究基于语言模型检索的问答学习系统,表 4 中记录了各系统在 6 种问答类型上的准确率,可以看出:本系统在大部分问答类型评测中,都较其他系统准确率更高。

在本系统的评测中,首都类型问答的准确率高达 95%,这是由于首都类型问答的表达通常较规范,且问题词和答案词之间的相隔通常都很近,如:“中国首都北京”,“中国的首都是北京”等表达大量在网页中出现,因而准确率很高;与此相反,出生时间的表达方式多样,且问题词与答案间隔一般较大,不利于模式学习,因而出生时间类型问答的准确率较低。

表 4　不同方法的准确率比较　%

系统	作者	发明	地址	出生地	出生时间	首都
文献[10]系统	55.0	30.8	52.6	80.0	14.3	13.7
文献[1]系统	85.0	30.8	21.1	80.0	42.9	33.3
本系统(骨架模式+朴素模式)	65.0	55.0	60.0	85.0	45.0	95.0

表 5 分析 120 次测试的准确率与网页数量关系,可以看出:当包含问题词的网页数量越大,本方法的准确率越高。

表 5　包含问题词的网页数量和准确率的关系

网页数量	测试数量	找出正确答案数量	准确率/%
1 万以下	26	13	50.0
1 万~10 万	34	25	73.5
10 万~100 万	46	36	78.2
100 万以上	14	11	78.5
总计	120	85	70.8

3　结语

本文提出了一个基于 self-training 机制的问答模式获取和评估的方法,并实现了对六类问答模式的学习。本文将 self-training 机制运用到了问答模式学习中,从而使得系统对人工标定语料的依赖度减少到最小;其次,本文引入了骨架模式,以提高模式对答案的定位能力;再次,本文利用启发规则,改进了候选答案的评分机制。对比实验的结果表明:本文的方法能较好地改进问答模式的性能。

目前只对答案类型为人名、地点、时间等三类实体的问答进行了研究,对于更为复杂的问答,留待下一步研究。

(下转第 1581 页)

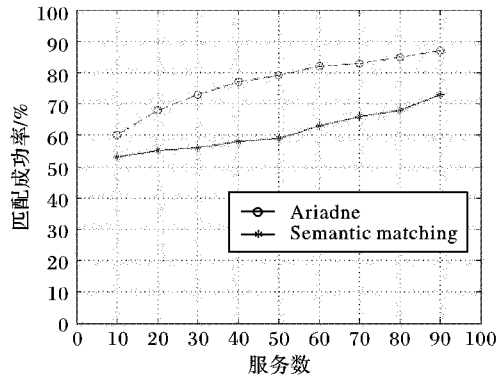


图 3　不同服务数的匹配成功率

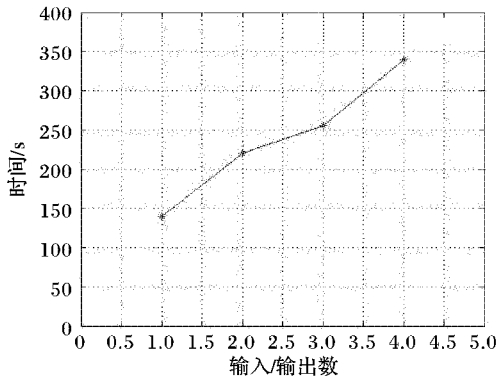


图 4　不同数量的输入/输出参数服务匹配时间

4　结语

本文在给出服务描述规范的基础上,通过服务的类型、服务的输入输出以及 QoS 参数进行服务的匹配,通过逐步过滤掉不符合服务请求的服务,来不断减少匹配的服务数;对匹配出的服务,通过匹配相似度的计算,从而找出最符合服务请求的服务。文中给出了相应的算法,并通过实验证明该算法能较好地发现所需要的服务性能。

本文算法没有根据服务的功能进行服务匹配,但同样可以经过扩展来按照服务的功能进行服务的匹配;该算法在 QoS 参数匹配中主要使用了数值参数进行服务匹配,在服务的发现过程中,其他质量类的参数比如安全QoS参数也是一个重要的方面,在本文的 QoS 参数匹配方式中,可以进行适当扩展,使其使用与其他 QoS 参数的服务匹配,这也是我们进一步的研究方向。

参考文献:

[1] ERL T. Service - oriented architecture: Concept, technology, and design [M]. Upper Saddle River, NJ, USA: Prentice Hall, 2005.

[2] Sun Microsystems. Jini architectural overview [EB/OL]. (1999 - 01) [2007 - 10 - 29]. http://www.sun.com/software/jini/whitepapers/architecture.pdf.

[3] OSGi[EB/OL]. [2007 - 11 - 06]. http://www.osgi.org/Main/HomePage

[4] GUTTMAN E, PERKINS C. IETF RFC 2608, Service location protocol, Version 2 [S/OL]. (1999 - 06) [2007 - 11 - 02]. www.ietf.org/rfc/rfc2608.txt.

[5] DOULKERIDIS C, LOUTAS N, VAZIRGIANNIS M. A system architecture for context-aware service discovery [C]// Proceedings of the First International Workshop on Context for Web Services (CWS 2005). Paris: ELSEVIER Press, 2005, 146(1): 101 - 116.

[6] MOKHTAR S B, KAUL A, GEORGANTAS N, *et al.* Towards efficient matching of semantic Web service capabilities [C]// Proceedings of the Workshop of Web Services Modeling and Testing (WS-MATE'06). Palermo: ELSEVIER Press, 2006: 137 - 152.

[7] MARTIN D, BURSTEIN M, HOBBS J, *et al.* OWL-S: Semantic Markup for Web Services [EB/OL]. (2004 - 11 - 22) [2007 - 11 - 03]. http://www.w3.org/Submission/OWL-S/.

[8] CHEN H, PERICH F, FININ T, *et al.* SOUPA: Standard Ontology for Ubiquitous and Pervasive Applications [C]// Proceedings of the 1st International Conference on Mobile and Ubiquitous Systems. Boston: Wiley-IEEE Press, 2004: 258 - 267.

[9] FaCT + + [EB/OL]. (2007 - 07 - 05) [2007 - 10 - 12]. http://owl.man.ac.uk/factplusplus/.

[10] CHIBOUT R, SAILHAN F, ISSARNY V. Ariadne user guide [EB/OL]. [2007 - 10 - 22]. http://www-rocq.inria.fr/arles/download/ariadne/doc/user-guide.pdf.

[11] LI LEI, HORROCKS I. A software framework for matchmaking based on semantic web Technology [C]// Proceedings of the Twelfth International World Wide Web Conference (WWW 2003). Budapest: ACM Press, 2003: 331 - 339.

(上接第 1577 页)

参考文献:

[1] 吴友政, 赵军, 徐波. 基于无监督学习的问答模式抽取技术[J]. 中文信息学报, 2007, 21(2): 69 - 76.

[2] SOUBBOTIN M M, SOUBBOTIN S M. Use of patterns for detection of likely answer strings: A systematic approach [C]// Proceedings of the 11th Text Retrieval Conference (TREC-11). Gaithersburg, Maryland: NIST Special Publication, 2002: 325 - 331.

[3] RAVICHANDRAN D, HOVY E. Learning surface text patterns for a question answering [C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002). Philadelphia, PA: Association for Computational Linguistics, 2002: 41 - 47.

[4] ZHANG D, LEE W S. Web based pattern mining and matching approach to question answering [C]// Proceedings of the 11th Text Retrieval Conference (TREC-11). Gaithersburg, Maryland: NIST, 02: 505 - 512.

[5] ROUSSINOV D, ROBLES J. Web question answering through automatically learned patterns [C]// Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL). New York: ACM Press, 2004: 347 - 348.

[6] 郑逢斌, 陈志国, 姜保庆, 等. 语义校对系统中的句子语义骨架模糊匹配算法[J]. 电子学报, 2003, 31(8): 1138 - 1140.

[7] ZHU XIAO - JUN. Semi - supervised learning literature survey, TR 1530 [R]. University of Wisconsin-Madison: Department of Computer Sciences, 2006.

[8] AGICHTEIN E, GRAVANO L. Snowball: Extracting relations from large plain-text collections [C]// Proceedings of the 5th ACM International Conference on Digital Libraries. New York: ACM Press, 2000: 85 - 94.

[9] RILOFF E, WIEBE J, WILSON T. Learning subjective nouns using extraction pattern bootstrapping [C]// Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 (CONLL). Morristown, NJ: Association for Computational Linguistics, 2003, 4: 25 - 32.

[10] PONTE J M, CROFT W B. A language modeling approach to information retrieval [C]// Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1998: 275 - 281.