基于语义链的检索在 QA 系统中的应用

张江涛 杜永萍
(北京工业大学计算机学院 北京 100124)

摘 要 自动问答系统以自然语言提出问题,并采用自然语言处理技术自动地将答案返回给用户。利用 WordNet 构建语义链,并将语义链用于问答系统。在面向 Web 的问答系统中,采用两种不同的计算文本相似度的方法对 Google 返回的 Snippets 按照相似度进行排序,对返回的第一个和前十个 Snippets 中包含答案片段的情况进行分析,与不使用语义链时的情况相比,使包含答案片段的准确率分别提高了 150%和 66.12%。对实验结果进行了显著性检验,在显著性水准 $\alpha=0.05$ 的条件下,得到 $p=0.000078$,使系统的准确率得到显著提高。

关键词 WordNet,语义链,问答系统
中图法分类号 TP391 **文献标识码** A

Retrieval Based on Lexical Chains in Q & A System

ZHANG Jiang-tao DU Yong-ping
(Institute of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract Question Answering System (QA) allows users to ask question in natural language. Using natural language processing technology, QA returns answer to the user automatically. The paper used WordNet to construct lexical chains, and used the lexical chain in a question answering system based on the Internet. Using two kinds of different way calculates similarity of Snippets download from Google to sort them according to similarity. When geting for the first one and first ten Snippets, the accuracy of the answer snippets is increased by 13.68% and 25.4%. We made significant inspection for the experiment results. In significant level scheffe post-hoc 0.05 conditions, $p=0.000078$ is got and the accuracy of the system is dramatically improved.

Keywords WordNet, Lexical chains, Question answering system

1 引言

智能问答系统(简称 QA 系统)是近年来自然语言处理研究中最受关注的课题之一[1]。智能问答系统是指根据用户以自然语言提出的问题找到正确答案[2],随着 Internet 的快速发展,人们获得信息的方式越来越多,快速获取资料并准确定位信息的愿望促进了问答系统的发展[3]。大规模文本处理技术的日趋成熟也成为推动问答系统实现的强大力量。在每年一度的文本信息检索会议(TREC)上,智能问答也成为备受关注的主题之一,越来越多的大学和研究机构参与了智能问答系统的研究[4]。

TREC-8,Cymfony 问答系统[5]利用了信息抽取系统——Textract,通过实体类别信息完成问答任务,该系统在评测中获得第一名。此后,在实体匹配的基础上,Cymfony 公司又利用基于链接结构匹配的方法来提高答案抽取的正确性,取得了不错的效果。TREC-9 SMU 的 FALCON 系统[6]采用的方法是循环检索,配以基于语义合一的算法,在 TREC-9 的评测中取得了优异的成绩。该系统很典型,同 TREC-8 的系统相比,系统规模上变化较大,同时开始引入语义分析来检验答案的正确性。TREC-10 InsightSoft-M 的 TextRoller 系统[7]根据问题的类别建立模块库,根据问句类型利用模板进行匹配。其除了使用正例模板外,还用到了反例模板,从候选答案集中排除不可能的答案。大量的模板使该系统在 TREC-10 的评测中取得了优异的成绩。该系统的成功表明,在问答系统中采用浅层自然语言处理技术也能获得较好的用户评价。TREC-11 LCC 的 PowerAnswer 系统[8]参加了 2002 年 TREC 问答系统评测——Main Task 测试,在 500 个问题的测试中,回答正确的有 415 个,不完全正确的有 8 个,未在文本集中找到答案的有 14 个,回答错误的有 63 个,置信度加权后评分高达 0.856 分,远远高于参加测试的其它系统的评分。该系统整合了大量自然语言处理的模块,将自然语言文本转化为逻辑表达式,通过逻辑推理机制来检验答案的正确性,检索答案时用到了多种知识库。

美国 ASKJeeves 公司的检索系统(http://www.askjeeves.com/)的最大的特点是允许用户用自然语言提问,该系统通过分析用户的提问并反问,即通过与用户交互辨别出用户的意图。这样用户能够充分表达检索需求,其比雅虎的关键字检索有了很大进步。

到稿日期:2012-05-07 返修日期:2012-09-24 本文受国家自然科学基金(60803086),北京市教委科技计划面上项目(KM200910005009)资助。
张江涛(1986—),男,主要研究方向为信息检索、自然语言处理,E-mail:tiandilinghuo@emails.bjut.edu.cn;杜永萍(1977—),女,博士,副教授,主要研究方向为信息检索、自然语言处理。

START 是全球第一个基于 Internet 的问答系统[9]，该系统主要采用基于知识库和信息检索的模式来获取答案，能够回答数以百万计的英文问题，主要包括与地理、电影、人物等相关方面的事实性问题。

AnswerBus(http://www.answerbus.com)系统在多语种问答系统上进行了一些尝试，可以回答英语、西班牙语、德语、意大利语等语言描述的问题，它使用 5 大搜索引擎搜索可能包含答案的网页并给出可能包含答案的句子。对于用户的每一次查询，系统将返回 5 个网页链接，同时给出 XML 和 txt 格式的候选答案。AnswerBus 在 TREC-8 的 200 个测试问题集中的正确率为 70.5%，这充分表明在 Web 上实现具有实用价值的问答系统是很有可能性的，AnswerBus 目前已成为众多研究者参考学习的重要对象。

国内也有不少大学和机构在进行问答系统的研究，以复旦大学、中国科学院、清华大学、大连理工大学、北京邮电大学、哈尔滨工业大学等为代表的越来越多的研究单位和研究团队加入到 TREC 的队伍中。复旦大学[10]参加了 2005 年的文本检索会议，并参与了问题问答、跨语言检索和文本过滤 3 个项目，取得了较好的成绩。中国科学院计算所从 2001 年 TREC-10 开始参加问答系统评测，主要采用段落评分、答案匹配等处理方法。中科院计算所正在进行大规模知识处理项目研究——National Knowledge Infrastructure（简称 NKI），该项目的一个具体应用就是 NKI 知识问答系统。该系统以 NKI 的知识库为基础，向用户提供各个领域的问答服务，特点是向用户提供准确的信息，并支持自然语言提问的方式。

目前，由于自然语言处理的复杂性和处理技术的局限性，使得计算机完全理解自然语言还是很困难的[11]。本文构建基于 WordNet 的语义链，并将其应用于抽取包含答案的文本片段。在使用语义链的情况下对 TREC-11 测试问题集随机抽取出的 307 个英文问题进行封闭测试，实现包含答案的文本片段的抽取。结果表明，语义链使系统的性能得到了显著提高，为问答系统提供了强有力的支持。

2 基于 WordNet 构建语义链

WordNet 最具特色之处是根据词义来组织词汇信息，是一部语义词典[12]。它是按照词汇的矩阵模型组织的，如表 1 所列。同义词集合(synonyms set)可以看作是词形(word form)之间具有中心角色的语义关系。

表 1 词汇矩阵

		词形			
		F_1	F_2	…	F_n
词义	M_1	E(1,1)	E(2,1)		
	M_2		E(2,2)		
	…				
	M_m				E(m,n)

表 1 指明了词汇矩阵的构想，F_1，F_2 … 表示多种词形，M_1，M_2 … 表示多种词义。当某列中有两个以上的元素时，表示该列对应的词形有多个词义，如表中 $E(2,1)$ 和 $E(2,2)$ 表示词形 F_2 有 M_1，M_2 两种词义。例如 spring，该单词作为名词时有春天、弹簧、泉眼等多种含义，作为动词时又有跳跃的含义。这种词汇矩阵的思想在 WordNet 中有充分的体现。

在 WordNet 中，不同的同义词集合之间通过某种关系(上下位关系、反义关系、部分整体关系)[13]联系起来构成了语义链，如图 1 所示。节点代表 WordNet 中的一个同义词集合，虚线所示的路径两端的节点在同一个同义词集中，实线所示的路径表示不同的同义词集之间以某种语义关系(上下位关系、反义关系、部分整体关系等)连接起来。节点旁边的单词表示该同义词集内容，图中以单词 book 为例，列出 book 的 3 个义项。

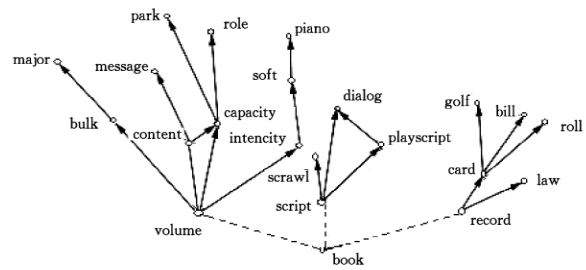

图 1 语义链示意图

相邻两个节点之间的路径长度设为 1，在图 1 中，称 book 为词源，扩展路径长度为 1 时，得到的词有 volume，script，record。扩展路径长度为 2 时，得到的词有 volume，script，record，bulk，content，capacity，intencity，scrawl 等。

可以看出，随着扩展的路径长度的增长，将会得到更多与词源 book 相关联的词，同时，路径长度越长，扩展出的词与词源之间的相似程度越低。不同的同义词集之间有不同的语义关系，根据其重要程度，赋予它们不同的权值，如表 2 所列。

表 2 语义关系权值表

语义关系	权重
同义词(同义词集)	0.9
上下位	0.7
相似关系	0.7
整体部分	0.5
反义	0.2

若词 A 和词 B 之间的路径长度为 n，路径上的权值分别为：$Weight_1$，$Weight_2$，……，$Weight_n$，利用

$$S_{A\text{-}B}=\frac{1}{n}\sum_{i=1}^{n}Ln(Weight_i) \tag{1}$$

计算词 A 和词 B 之间的相似度，其中 $S_{A\text{-}B}$ 为相似度。

若词 C 和词 H 之间有路径 C-D-……-G-H，假设 C 和 G 之间的路径长度为 $n-1$，C 和 H 之间路径长度为 n，由式(1)可得

$$S_{C\text{-}G}=\frac{1}{n-1}\sum_{i=1}^{n-1}Ln(Weight_i) \tag{2}$$

由式(2)推导

$$\begin{aligned}S_{C\text{-}H}&=\frac{1}{n}\sum_{i=1}^{n}Ln(Weight_i)\\&=\frac{1}{n}\left[\sum_{i=1}^{n-1}Ln(Weight_i)+Ln(Weight_n)\right]\\&=\frac{1}{n}\left[(n-1)S_{C\text{-}G}+Ln(Weight_n)\right]\end{aligned}$$

即

$$S_{C\text{-}H}=\frac{1}{n}\left[(n-1)S_{C\text{-}G}+Ln(Weight_n)\right] \tag{3}$$

且有 $S_{A\text{-}A}=1$

根据如上公式，语义链的构建算法描述如表 3 所列。

表 3　利用 WordNet 构建语义链算法

1）创建两个队列 DealingQ，FinishQ；
2）词源 W_1 入队列 DealingQ；i=1；输入最大扩展词数 max；
3）获取队列 DealingQ 的队首元素 W_i，在 WordNet 中获取 W_i 的孩子节点 $W_{i1}, W_{i2}, W_{i3}, \cdots, W_{ik}$
4）对于每个孩子节点，检查与其父节点的语义关系 $(W_i, W_{i1}), (W_i, W_{i2}), \cdots, (W_1, W_{ik})$
根据语义关系赋予权重 $Weight_{i1}, Weight_{21}, \cdots, Weight_{kl}$
利用式(3)计算各个孩子节点与词源的相似度，例如 $S_{W_{ik}-W_1}=\frac{1}{n}[(n-1)S_{W_i-W_1}+Ln(Weight_{ik})]$
分别将 $W_{i1}, W_{21}, \cdots, W_{ik}$ 入队列 DealingQ
将 W_i 入队列 FinishQ
从队列 DealingQ 中删除 W_i
i++；
若 FinishQ 中词个数已达到最大值 max；转到 6）
5）如果队列 DealingQ 非空，转到 3）
6）算法结束。

以上算法通过建立树状结构得到词源的扩展集合，W_1 为词源，W_1 与其自身的相似度为 1（其他词与词源的语义关系的权重均介于 0 到 1 之间），将 W_1 入队列。游标 i 初始化为 1，取队首元素 W_i，通过 WordNet 对 W_i 进行扩展，得到 $W_{i1}, W_{i2}, W_{i3}, \cdots, W_{ik}$。

称这些扩展出的词为 W_i 的孩子节点。利用式(3)计算各子节点与词源之间的相似度，并将这些孩子节点入队 DealingQ，重复以上过程，直到扩展出的词数量达到最大值，算法退出。

3　语义链在问答系统中的应用

3.1　问答系统基本框架

问答系统普遍采用如图 2 所示的基本框架。

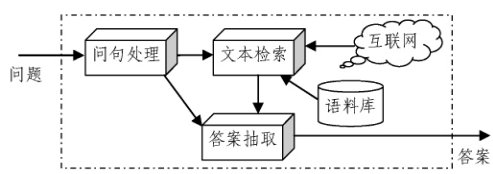

图 2　问答系统基本框架

框架主要由问句处理、文本检索、答案抽取 3 个模块组成。问句首先通过问句处理模块得到问句主干并检查问句类型；文本检索模块根据问句主干从语料库包括互联网获取文本或段落，如果数据来源是互联网，则获取语料难度更大，Web 信息资源呈现出海量化、非结构化的趋势，鉴于此，需要采取一定的信息抽取策略[14]，问答系统的答案信息源很广，可能的信息源还包括文本集合、数据库、分类目录、知识库、本体、常见问题解答等等[15]；最后答案抽取模块结合问句类型从文本段落中抽取答案。

3.2　基于句子相似度检索

本文抽取答案相关片段的方法是采用句子相似度匹配的方法。

在不使用语义链时，对每个英文问题的处理流程如下：

1）问句处理

在有些英语句子中，某些介词、代词、助词等停用词不表达实际意义，将其去除，例如本文测试问句中的 am，is，are，and，what，when 等，并将剩余的部分提取词干。将去停用词、取词干后的句子生成查询 Query。

2）面向 Web 检索

向 Google 提交 Query，并将 Google 返回的 Snippets 整理存盘。

3）相似度计算

计算 Query 与每条 Snippets 的相似度。根据 Snippets 中关键词（Query 中的词）出现的次数和平均距离计算每个 Snippets 与 Query 的相似度，即

$$S_i=\frac{Count_i}{L_i} \tag{4}$$

式中，$Count_i$ 为 Query 中的词在第 i 条 Snippets 中出现的总次数，$\bar{L}_i$ 为第 i 条 Snippets 中包含 Query 中出现的词的平均距离。S_i 为相似度。将 Snippets 按相似度降序排序。将排好序的 Snippets 分别返回第一个，前两个，……，前十个作为检索结果。

4）结果评价

分别计算返回第一个，前两个，……，前十个的 Snippets 时，能够包含答案的正确率。例如返回第一个时，检查该 Snippets 是否含有正确答案，返回前两个时，检查前两个是否含有正确答案，……，测试完所有问题后，系统在检索阶段的正确率＝结果中包含答案的问题数目/总问题数目。

3.3　基于语义链的检索

利用语义链抽取 Snippets 时，系统架构如图 3 所示。原始 Question 通过问句处理模块的词性标注，去停用词，取词干后，得到 Query（查询串）。向 Google 提交该 Query，返回若干 Snippets（文本片段），原始 Question 经过问句处理模块后提交给语义链构建模块，经过该模块处理后得到与 Query 对应的 Expend Words（扩展词集）。通过 Expend Words 计算 Query 与各个 Snippets 的相似度（见表 3），然后降序排列各个 Snippets，按照相似度从高到低依次检查各个 Snippets 是否包含答案。

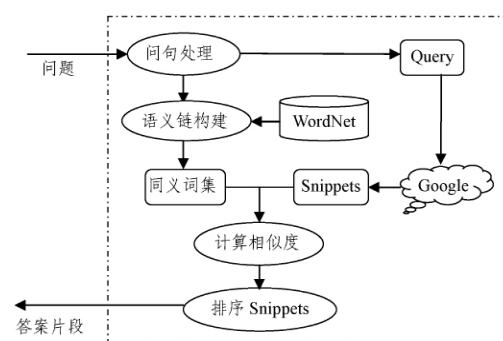

图 3　运用语义链获取答案相关 Snippets 过程示意图

在使用语义链时，加入语义链构建模块来提高检索的正确率。检索包含答案的 Snippets 可采用以下两种方案。

方案 1　使用语义链不考虑词间距检索。

利用式(5)对 Expend Words 中的每个词进行取词干操作。

$$S_i=\sum_{k=1}^{n} t_k/L_i \tag{5}$$

通过 Expend Words 计算 Query 与各个 Snippets 的相似度。其中 S_i 为 Query 与第 i 个 Snippets 的相似度，L_i 为第 i 个 Snippets 的长度（即词的个数），n 为 Expend Words 中词的个数，t_k 为 Expend Words 中第 k 个词在第 i 个 Snippets 中出现的次数，若 Expend Words 中所有的词在第 i 个 Snippets 中均未出现，则 $S_i=0$。

方案 2　使用语义链考虑词间距检。

对 Expend Words 中的每个词进行取词干操作。根据 Expend Words 中的词在 Snippets 中出现的次数和平均距离计算各个 Snippet 与 Query 的相似度。

$$S_i=(\sum_{k=1}^{n}t_k)/(\frac{1}{m}\sum_{j=1}^{m}l_j) \tag{6}$$

$$m=\sum_{k=1}^{n}t_k-1$$

式中，S_i 为 Query 与第 i 个 Snippets 的相似度；n 为 Expend Words 中词的数量；t_k 为 Expend Words 中第 k 个词在第 i 个 Snippets 中出现的次数，若 Expend Words 中的词在第 i 个 Snippets 中一次都未出现，则 $S_i=0$，若 Expend Words 中的词在第 i 个 Snippets 中仅出现一次，则将分母替换为 l_i，l_i 为第 i 个 Snippets 的长度（即词的个数）。否则 l_i 为 Snippets 中出现的第 j 个词与第 $j+1$ 个词之间的距离。

4　实验结果

TREC-11 测试问题集包含 3 种类型，共 500 个问题，其中事实型（Factoid）问题有 413 个，定义型（Definition）问题有 50 个，列举型（List）问题有 37 个。本实验对 TREC-11 随机抽取的 307 个问题进行测试，按照疑问词对其进行分类，其类型分布如图 4 所示。

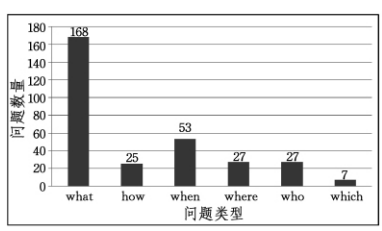

图 4　测试问题类型分布图

在不使用语义链的情况下，分别检测返回第一个，前两个，前三个，……，前十个 Snippets 时的正确率，结果都明显低于使用语义链时的情况。图 5 中，在只返回与问题相似度最高的一个 Snippets 时不使用语义链的正确率为 9.12%，使用语义链不考虑词间距时的正确率为 20.52%，使用语义链考虑词间距时的正确率可达到 22.80%，比不使用语义链时的正确率分别提高了 125%和 150%。而返回相似度最高的前十个 Snippets 时不使用语义链时的准确率为 38.43%，使用语义链不考虑词间距和考虑词间距时的准确率分别为 64.16%和 63.84%，比不使用语义链的情况正确率分别提高了66.95%和 66.12%。

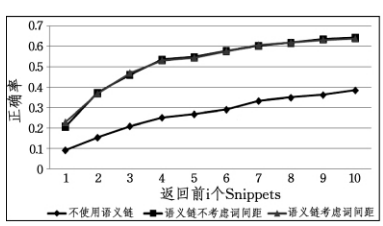

图 5　检索结果正确率曲线图

通过对 3 组实据进行单边 t 检验可以看出，语义链的使用使检索的正确率得到显著提高。t 检验结果如表 4 所列，其中 Route1 代表不使用语义链，Route2 代表使用语义链不考虑词间距，Route3 代表使用语义链并考虑词间距。

表 4　测试系统性能提高显著性的 t-test 结果

Route1 and Route2	Route1 and Route3	Route2 and Route3
t=7.59	t=8.01	t=0.07
p=0.00012	p=0.000078	p=0.4852

在以上的实验结果中不难看出，语义链在对相关片段的检索过程中起到了显著的促进作用，另外，在使用语义链时不考虑词间距（前文提到的计算相似度时的方案 1）和考虑词间距（前文提到的计算相似度时的方案 2）时的准确率相差不大，因为 Google 搜索返回的信息中词间距都是比较小的。

图 6 为使用语义链返回前十个 Snippets 时命中答案的问题类型分布规律图（考虑词间距与不考虑词间距分布规律相同）。横轴代表问题类型，纵轴代表检测成功问题中各类型问题出现的次数。

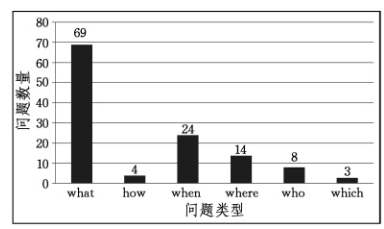

图 6　使用语义链检索返回前十个 Snippets 时命中问题类型分布图

从图 4 可以看出，在测试问题集中 where 和 who 类型的问题数量比值为 27/27=1，但在图 6 中 where 类型的问题找到答案的次数与 who 类型的问题找到答案次数的比例为 14/8=1.75。

可以看出，利用语义链时 where 类型的问题比 who 类型的问题检测成功的准确率更高。在图 4 中，when 类型的问题与 how 类型问题数量的比例为 53/25=2.12，在图 6 中，when 与 how 类型的问题找到答案次数的比例分别为 24/4=6，可以看出，when 类型问题比 how 类型问题更容易抽取到包含答案的文本片段。

结束语　本文利用对比英文句子相似度的方法来进行相关片段的检索，分别对比了不使用语义链和使用语义链时系统返回答案相关片段的准确率。通过对实验数据的分析可知，使用单边 t 检验，在显著性水准 $\alpha=0.05$ 的条件下，得到的 p 值为 0.000078，实验结果表明，利用英语语义链对问题进行扩展使问答系统中检索准确率得到了显著提高。

本文没有对 Google 搜索引擎返回的网页内容进行详细解析，只是分析了 Google 搜索返回的 Snippets 信息，以至于正确率没有达到理想的结果，我们下一步的工作是对 Google 搜索引擎返回的网页进行详细解析并抽取网页中包含的答案；另外本文只对英文问答系统进行了研究，并没有涉及到中文，这也将是我们下一步的工作。

参 考 文 献

[1] 康海燕，李飞娟，苏文杰. 基于问句表征的 Web 智能问答系统[J]. 北京信息科技大学学，2011，26(1)

[2] 崔桓，蔡东风，苗雪雷. 基于网络的中文问答系统及信息抽取算法研究[J]. 中文信息学报，2004，18(3)

[3] 刘亮亮，林乐宇. 基于查询模板的特定领域中文问答系统的研究与实现[J]. 江苏科技大学学报：自然科学版，2011，25(2)

[4] 曹志娟，李祖枢，刘朝涛. 自动问答系统中的问题理解研究[J]. 计算机科学，2005，32(11)：158-160

（下转第 300 页）

所示。

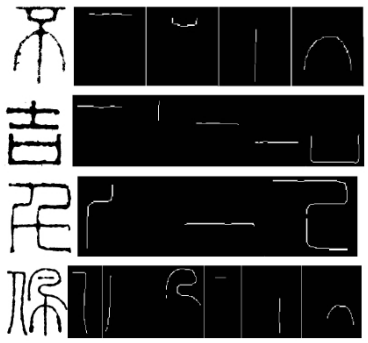

图 16　原印字和笔划分割的结果

依文献[11]中所述，以人工方式提取上述 300 个篆字的正确笔划作为比较依据，使用通用的指标来定量地衡量算法的效果。定义笔划分割的正确率，如式(3)所示：

$$\text{正确率}=\frac{\text{正确分割的样本数}}{\text{总测试样本数}}\times 100\% \tag{3}$$

式(3)中，若一个篆字的每一个基本笔划都能正确分割，则视该篆字样本为正确分割。将 300 个测试样本分成 3 组，每组的测试样本总数和正确分割的样本数如表 1 所列，得到平均笔划分割的正确率为 86.7%。

表 1　算法的测试结果

分组	总测试样本	正确分割数量	正确率
组 1	100	90	90%
组 2	100	83	83%
组 3	100	87	87%

该篆字笔划分割方法的特点在于去除了冗余的笔划交叉点，减少了细化变形的影响；子笔划的分割和笔划的组合按照同一个笔划在交叉处两端笔划走势不变的原则实现，符合篆字的笔划书写特征和构成规律。通过表 1 的统计结果，表明了该算法的有效性。

结束语　篆书作为一种汉字字体已经存在了两千多年，经历了从官方文字到中华民族独有的艺术表现形式的转变。千年来，其由于摆脱了官方文字的束缚，又加之不同篆刻家的不同艺术理解和诠释，字形结构变化复杂。利用传统的汉字识别技术难以实现篆字的准确识别。

本文在子笔划(或笔划基元)和被识别汉字之间增加了笔划的概念，将笔划这一显然具备汉字结构特征的对象层语义概念引入到汉字识别过程中来，意图改善篆字的识别性能。针对篆字的结构特征，采用模板匹配的方法得到交叉区域内正确的笔划组合，从而实现了笔划的自动分割提取，为篆字的识别奠定了基础。

参 考 文 献

[1] 曹忠升，苏哲文，王元珍，等. 基于模糊区域检测的手写汉字笔画提取方法[J]. 中国图象图形学报 A，2009，14(11)：2341-2348

[2] Bium H. A Transformation for Extracting New Descriptors of Models for the Perception of Speech and Visual Form[M]. Waithen-Dunn W，ed. US：MIT Press，1967

[3] Zeng Jia，Liu Zhi-qiang. Stroke Segmentation of Chinese Characters Using Markov Random Fields[C]// ICPR(1). 2006：868-871

[4] Zeng Jia，Liu Zhi-qiang. Type-2 Fuzzy Markov Random Fields and Their Application to Handwritten Chinese Character Recognition[J]. IEEE Transaction on Fuzzy system，2008，16(3)：747-760

[5] Cao R，Tan C L. A Model of Stroke Extraction from Chinese Character Images[C]// International Conference on Pattern Recognition (ICPR). Spain，2000：368-371

[6] Lau K K，Yuen P C，Tang Y Y. Stroke Extraction and Stroke Sequence Estimation on Signatures[C]// International Conference on Pattern Recognition (ICPR). 2002

[7] 孙晓红，张学东. 基于邻域特征的笔划交叉点提取算法的研究[J]. 计算机工程与设计，2008，29(19)：4985-4986，5058

[8] 张世辉. 一种新的基于距离的汉字笔画抽取方法[J]. 计算机工程，2003，29(14)：37-38

[9] 邵宏峰，罗予频. 一种基于 Delaunay 三角化的笔划分割算法[J]. 微计算机信息，2007，23(1)：269-271

[10] 康辉，李思莉. 脱机手写汉字中叉点精细化的改进算法[J]. 计算机工程，2006，32(20)：193-194，215

[11] 沈晓英. 篆书书写入门[M]. 太原：山西人民出版社，2002

(上接第 260 页)

[5] Srihari R，Li W. Information extraction supported question answering[C]// NIST. 1999，15

[6] Harabagiu S，Moldovan D，Pasca M，et al. Boosting knowledge for answer engines[C]// TREC. 2000

[7] Soubbotin M M. Patterns of potential answer expressionsas clues to the right ans-wers[C]// Text REtrieval Conference (TREC) TREC. 2002

[8] Moldovan D，Harabagiu S，Girju R，et al. LCC tools for question answering[C]// Text REtrieval Conference (TREC) TREC. 2002

[9] 于士涛，袁晓洁，师建兴，等. 一种 Web 问答系统中基于 XML 片段的语义项模型[J]. 计算机研究与发展，2007，3

[10] Niu Jun-yu，Sun Lin，Lou Lu-qun，et al. WIM at TREC 2005[C]// TREC. 2005

[11] 余正涛，邓锦辉，韩露，等. 受限域 FAQ 中文问答系统研究[J]. 计算机研究与发展，2007，2

[12] 王东睿，杨庚，陈蕾，等. 基于 WordNet 和 Kernel 方法的 Web 服务发现机制研究[J]. 计算机技术与发展，2010，20(12)

[13] 阮佳彬，杨育彬，林金杰，等. 基于本体词汇的三维模型语义检索[J]. 计算机科学，2009，36(2)

[14] 黄锋，吴华瑞. 一种自适应的 Web 信息抽取规则自动生成方法[J]. 广西师范大学学报：自然科学版，2011，3(1)

[15] 陈冰琦. 中英文双语问答系统中问句处理的研究[D]. 上海：上海交通大学，2004