

基于句法结构特征分析及分类技术的答案提取算法

胡宝顺¹⁾ 王大玲²⁾ 于 戈²⁾ 马 婷²⁾

¹⁾(东北大学软件学院计算机科学与技术系 沈阳 110004)

²⁾(东北大学信息科学与工程学院计算机软件与理论研究所 沈阳 110004)

摘 要 由于中文自然语言处理的特点和困难以及相应的语言处理基础资源的相对缺乏,使得国外一些成熟技术和研究成果不能直接应用到中文问答系统中。为此,针对中文事实型问答系统,提出一种新的基于句法结构特征分析及分类技术的答案提取算法,该方法将答案提取问题看成是候选答案的分类问题,即将候选答案分类为正确和错误两类。首先,该方法根据与问题类型所对应的候选答案的类型信息,从文本片断中提取出候选答案及其在句子中的简单特征和句法结构特征;然后利用这些特征训练分类器;最后用训练得到的分类器判别候选答案是否为正确答案。针对中文事实性问题,该方法与目前典型的基于模式匹配的中文答案提取算法相比,准确率提升 6.2%, MRR 提升 9.7%。

关键词 句法依存分析;分类;答案提取;中文问答系统;事实性问题

中图法分类号 TP391

An Answer Extraction Algorithm Based on Syntax Structure Feature Parsing and Classification

HU Bao-Shun¹⁾ WANG Da-Ling²⁾ YU Ge²⁾ MA Ting²⁾

¹⁾(Department of Computer Science and Technology, Software Collage, Northeastern University, Shenyang 110004)

²⁾(Institute of Computer Software and Theory, School of Information Science and Engineering, Northeastern University, Shenyang 110004)

Abstract Due to the feature and difficulty of Chinese natural language processing and the lack of related resources, some foreign mature techniques can not be applied in Chinese Question Answering (QA) system. For the Chinese factoid QA system, a new answer extraction method based on syntax structure feature parsing and classification is presented in this paper. With the method, the answer extraction is regarded as candidate answer classification problem, i. e. candidate answers are classified into correct and incorrect answer. According to the part-of-speech information of candidate answers corresponding to question types, the candidate answers and their features (both simple and syntactic) in sentences from snippets are firstly extracted. Then these features are used to train the classifier. Finally, the trained classifier is used to distinguish whether the candidate answer is correct or not. For Chinese factoid questions, comparing to currently typical pattern matching based answer extraction algorithm, the new method improves precision by 6.2% and MRR by 9.7%.

Keywords syntax dependency parsing; classification; answer extraction; Chinese Question Answering (QA) system; factoid questions

1 引言

随着互联网的普及,搜索引擎已经成为人们快速查找信息和资源的重要手段.但目前的搜索引擎主要采用基于关键字的查询,而关键字的简单组合不能明确表述用户的查询意图,这一问题已成为制约搜索引擎性能提高的瓶颈之一.问答式检索系统(简称问答系统)正是为克服传统搜索引擎的这一弊端应运而生的.与基于关键字的传统搜索引擎不同,问答系统允许用户以自然语言形式提问,并将准确简短的答案、而非大量的相关文本和网页返回给用户.比如:用户提问“第三届亚洲政党国际会议是由哪个政党主办的?”,问答系统就可以将“中国共产党”的答案返回给用户.因此可以说,问答系统是更高效、更人性化的新一代搜索引擎.同时也是集自然语言处理、信息检索、信息抽取、机器学习等多学科技术于一体的复杂系统.

一般来说,问答系统主要包括问题分析、信息检索和答案提取 3 个部分.其中,问题分析的主要工作包括确定问题类型和提取问题中的关键字等;信息检索部分的任务是利用问题关键字生成查询条件,然后利用文档库或提交给 Web 搜索引擎进行检索,返回相关的文档或段落;答案提取部分的任务则是从候选的文档或段落中提取出正确答案.作为问答系统中的一个关键环节,答案提取部分性能的优劣直接影响整个问答系统的性能.Moldovan^[1]等人关于问答系统错误的分析结果表明,约 18.7% 的回答错误是由诸如候选答案识别错误、答案排序错误等导致的.因此,答案提取算法的研究对提高问答系统整体性能具有重要的意义.

近几年来,国外很多科研院所和著名公司如 IBM、Microsoft、ISI、MIT、University of Cambridge 等都积极投入到问答技术的研究中,多个问答系统评测平台如 TREC、NTCIR、CLEF 的成功举办也极大地推动了该领域的快速发展.目前,国外已经有一些相对成熟的问答系统问世,同时也不乏研究人员提出了很多效果理想的答案提取算法.同时近些年,国内从事问答系统相关研究的机构不断增加,其中中国科学院自动化研究所、哈尔滨工业大学、复旦大学、清华大学和沈阳航空工业学院等都在该领域做了很多研究工作^[2-4].但相对而言,中文问答技术的研究尚处于初级阶段,与国外存在较大差距.一方面,由于中文自然语言处理的特点和困难,目前这方

面的各种底层技术还不够成熟和完善;另一方面,相应的语言处理基础资源如知识库、语料库等也相对缺乏,这使得国外一些成熟技术和研究成果不能直接应用到中文问答系统中.基于此,本文提出一种应用于中文问答系统的基于句法结构特征分析及分类技术的答案提取算法.

本文第 2 节简单介绍答案提取算法的相关研究工作;第 3 节简要介绍我们提出的算法的总体实现步骤;第 4 节论述提取句子句法特征时应用的关键技术:基于句法依存分析的路径相似度计算;第 5 节阐述候选答案的特征提取及分类问题;第 6 节给出实验的具体步骤和实验结果;第 7 节是总结和展望.

2 相关工作

目前中文问答系统的答案提取算法主要包括 3 类:(1) 基于信息检索和信息抽取的问答技术^[5-7];(2) 基于模式匹配的问答技术^[2,8-13];(3) 基于机器学习的答案提取技术^[3-4].

文献[6]描述了一个典型的基于信息抽取的答案提取算法.该算法的主要思想是,在信息检索模块返回的前几个相关的句子的基础上,进行更细化(fine-grained)的命名实体识别,将问题类型对应的命名实体作为候选答案,然后将与匹配词距离最近的候选答案作为正确答案.该文献提出的系统的整体性能良好,但是仅就答案提取而言,算法显得有些简单,且仅使用了匹配词与候选答案词的距离这一个特征.

文献[13]提出了一种基于表面文本模式(surface text pattern)匹配的答案提取算法.该算法首先人工标注问题的标准答案;然后根据搜索引擎检索含有问题中的焦点词和正确答案的句子,利用广义后缀树(generalized suffix tree)算法提取出这些句子的公共字符串;对公共字符串经过过滤和准确性评估后,将保留下的字符串中的焦点词和标准答案替换为插槽词(slot word)生成答案模板;最后利用答案模板来进行答案提取.该方法在问题类型为询问生日、发明者、发现者、定义、成名原因、地点时有很好的效果.但是解答其他的问题类型时性能不佳,且不能处理焦点词和正确答案之间长距离的依存关系.

文献[2]中提出了一种基于无监督学习的问答模式抽取技术,并通过实验证明应用问答模式提取答案是有用的.该算法无需用户提供〈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 and 水平聚类等步骤完成该类型提问的答案模式的学习. 该算法存在的问题是:需要对问题类型进行详细的划分,针对每一个问题类型均需要从互联网中学习相应的问答模式. 该算法针对只有一个“提问焦点词”的问题的性能较好. 针对有多个必需限定词的问题,该算法只能通过增加问题模式类型来解决. 如“中国最长的河流是什么?”,该问题中的“中国”和“最长”均为必需的限定词. 按照该算法,问题的“提问焦点词”为“河流”. 这就导致了该算法的扩展性受到制约.

Sun^[3]将答案句子提取问题视为分类问题,即将候选答案句子分类为正确或是错误. 他们通过提取问题和候选答案句子的特征训练最大熵模型,然后利用得到的模型提取答案,并通过实验证明该方法的有效性. 受到该文章的启发,我们提出了本文这个基于分类技术的答案提取算法. 我们的方法与 Sun 的方法的不同之处在于,我们的方法可以直接提取出精确的答案词,而不是答案所在的句子.

文献[4]中提出了一种基于实例的答案提取算法. 该算法利用问题及其对应的正确答案句子、错误答案句子和正确答案词中提取得到的特征作为分类算法最大熵模型(maximum entropy model)的训练特征. 该文章主要提取了以下 3 个特征:(1) 查询词与句子的匹配情况;(2) 问题句子中的词与句子中的词的匹配情况;(3) 疑问词与句子中的词的匹配情况,即句子中是否含有与问题答案相同词性的词. 以上 3 个特征均为布尔型值,即“真”(TRUE)或者“假”(FALSE). 该文章仅对地点(国家)和实体(语言)型问题进行了性能测试,没有与其他答案提取算法进行性能对比实验. 该算法提取的分类特征比较简单,且均为布尔类型. 没有考虑词之间的语义特征,所以在分类性能上将会受到一定的制约.

3 基于分类技术的答案提取算法

因为本文的重点是答案提取算法,问题分析和信息检索非本文的重点,所以我们将问题类型信息视为已知信息. 对于信息检索模块,我们简单地使用 Google 搜索引擎检索得到的文本片断(snippet)作为答案提取的来源.

3.1 生成查询词

生成查询词是文本片断检索的基础. 我们借鉴了文献[7]中系统的查询词生成算法并加以改进,具

体算法如下:

(1) 根据问题集,生成一个疑问词列表. 疑问词为形如:“谁”、“哪”、“什么”等等的词;

(2) 对问题进行分词和词性标注,将问题中出现的疑问词及其后面的量词或数量词均作为疑问词剔除;如:“哪一年”这样的由疑问词和数量词构成的词将作为疑问词被剔除;

(3) 去除停用词. 如“的”、“在”、“于”等等. 同时去除介词、助词和标点符号;

(4) 将剩余的词作为关键词,构成查询条件(关键词之间简单地以空格分隔,构成一个“布尔或”查询).

3.2 训练分类器

训练分类器的目的在于:找出候选答案所在的句子的特征与候选答案是否为正确答案的一种潜在的映射关系,是实现候选答案分类的基础,具体实现步骤如下:

(1) 将上面生成的查询条件提交给 Google 搜索引擎,保存检索返回的前 100 个文本片断;

(2) 根据问题的类型,利用命名实体(人名、地名、机构名、时间词、数量词)识别技术,识别出与问题类型对应的命名实体作为候选答案,然后计算候选答案在所在句子中的各个特征值,最后根据问题对应的标准答案,给候选答案加上类别标签(0:候选答案为非正确答案;1:候选答案为正确答案);

(3) 重复执行上面两个步骤,得到候选答案训练样本集,从而可以利用相应的分类器训练算法,训练得到用于分类的分类模型.

3.3 答案提取

答案提取是我们最后的目标,具体步骤如下:

(1) 将问题查询词提交给搜索引擎,取得搜索引擎返回的前 30 个文本片断;

(2) 根据问题类型,识别出每个文本片断中的候选答案,并计算候选答案所在句子的各特征值;

(3) 利用训练好的分类器,预测各个候选答案的分类,并返回前 5 个结果.

4 基于句法依存分析的路径相似度计算

本节详细阐述提取句子句法特征时所要使用的关键技术:基于句法依存分析的路径相似度计算.

4.1 句法依存分析

句法分析(parsing)是自然语言处理领域研究的关键问题之一,属于浅层语义分析中的重要内容,

在机器翻译、信息抽取和自动问答等多个领域中有着广泛而重要的应用,而基于依存语法的句法分析(简称句法依存分析)是目前句法分析的主要方法之一.

依存语法是 1959 年由法国语言学家 Tesiniere 在其著作《结构句法基础》一书中提出的. 此语法的核心思想是:句子中述语动词是支配其它成分的中心,而它本身却不受其它任何成分的支配,所有的受支配成分都以某种依存关系从属于其支配者. 依存语法的句法结构的主要元素是依存关系(dependen-

cy relationship),即句子中词对的二元关系,其中一个记为核心词(head),另一个记为依存词(dependent). 依存关系反映的是核心词和依存词之间语义上的依赖关系.

对于事实性问题“中国是在哪一年恢复了在联合国的合法席位?”,利用哈尔滨工业大学信息检索研究室提供的汉语句法依存分析器进行解析的结果如图 1 所示.

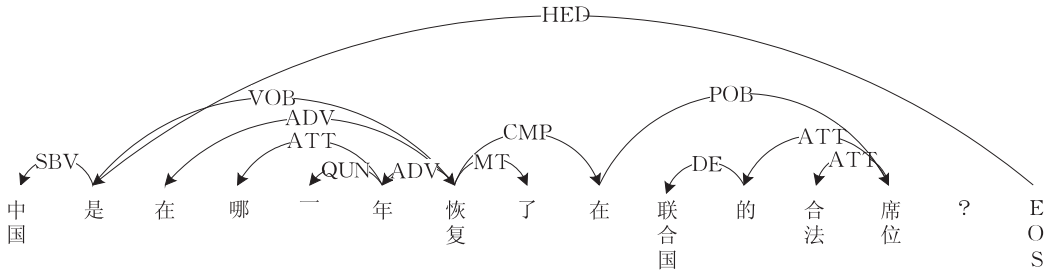


图 1 句法依存分析的一个例子

句子进行解析后得到的结果,我们将其简称为依存树,其中的词称为依存树的结点. 如果两个词之间有弧相连,则表示它们之间存在依存关系. 弧的方向是由核心词指向依存词,弧上的标记表示依存关系的类型. 在哈尔滨工业大学自然语言实验室开发的句法依存分析器中,依存关系类型共有 24 种^[14],如表 1 所示.

表 1 哈尔滨工业大学的依存分析器中定义的 24 种句法关系

关系	描述	关系	描述
CNJ	关联结构	BA	“把”字结构
IS	独立结构	BEI	“被”字结构
HED	核心	DEI	“得”字结构
RAD	后附加关系	DE	“的”字结构
POB	介宾关系	DI	“地”字结构
VV	连动结构	SIM	比拟关系
LAD	前附加关系	QUN	数量关系
COO	并列关系	APP	同位关系
ATT	定中关系	DC	依存分句
VOB	动宾关系	MT	语态结构
CMP	动补结构	SBV	主谓关系
IC	独立分句	ADV	状中结构

4.2 路径提取

我们将依存树中两个结点 W_1 和 W_n 之间的“路径”定义为:从 W_1 开始,到 W_n 结束中间所经过的一系列依存关系和词(不包括开始词和结束词),我们可以将其表示为下面的表达式 $Path_{1,n} := \langle pos_1 : Rel_1 : pos_2 \leftarrow W_2 \rightarrow pos_2 : Rel_2 : pos_3 \leftarrow \dots \rightarrow pos_{n-1} : Rel_{n-1} : pos_n \rangle$,其中 $Rel_i (1 \leq i \leq n-1)$ 表示其中的

第 i 个单一的依存关系; pos_j 表示第 j 个词的词性; \rightarrow 或者 \leftarrow 表示依存关系的方向; W_i 表示第 i 个词;因路径不包含开始词和结束词,所以在上面的定义表示中没有出现 W_1 和 W_n . 也就是说,关系路径描述依存树中两个结点之间依存关系和词的一个遍历. 虽然在对句子进行解析时,其依存关系都是有向的,但由于在问题和候选答案句子中,某个词语作为支配者和被支配者的角色经常是可以互换的,因此,我们在提取关系路径时将其方向忽略.

路径两端的位置被称作插槽(slot),而在此位置上的具体填充词被称为插槽词(slot word). 在一个插槽中可能出现不同的填充词,插槽词不属于路径的一部分.

提取得到的路径可以用于衡量问题句子和候选答案句子之间的相似度. 因为目前直接利用句子的依存树解析结果进行句子间的相似度计算比较困难,所以我们从依存树中提取出线性的路径. 每一个依存关系链接表示一个直接的语义关系,而一个关系路径允许我们表示两个词语之间非直接的语义关系. 因为路径是整个句子的一部分,所以可以通过不同句子间对应的关系路径的相似度来计算出句子间的相似度.

关系路径定义和提取方法参考了文献[8]中的内容,但提取关系路径时所进行的预处理和提取规则有所不同,具体规则如下:

规则 1. 对依存关系树中多个连续的时间词进行合并,使其成为一个时间词,同时剔除各个时间词

内部的依存关系,只保留其与外部的依存关系.例如,句子“1984 年 4 月 26 日至 5 月 1 日,美国总统

罗纳德·里根访问中国。”的原始依存树解析结果如图 2 所示.

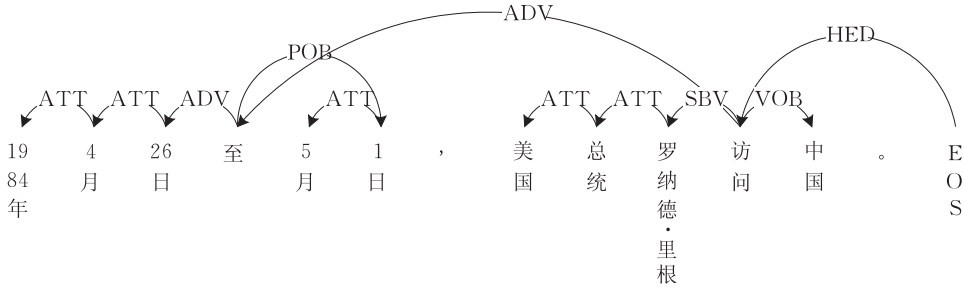


图 2 依存树中年/月/日被分开

从图 2 可见,经过分词后,一个完整的时间词“1984 年 4 月 26 日至 5 月 1 日”被分成“1984 年/4 月/26 日/至/5 月/1 日”.且从图 2 可以看出,这个完整的时间词与其他词之间只有一个依存关系“访问→ADV→至”.我们将时间词合并,将时间词内部的关系“ATT”、“ADV”和“POB”合并,只保留“至”与“访问”的依存关系.经过处理后,我们得到图 3 的解析结果.

作为“名字”和“我”之间的依存关系(如图 5 所示).

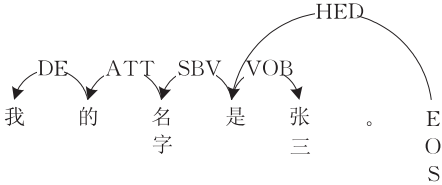


图 4 “我的名字是张三”的原依存树

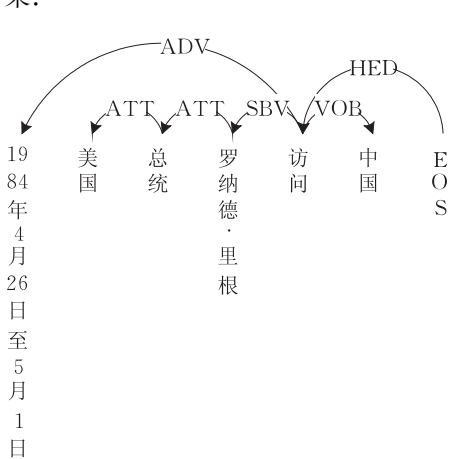


图 3 年/月/日合并后的依存树

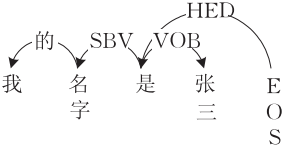


图 5 “我的名字是张三”删掉“的”字后的依存树

类似地,对“地”字结构和“得”字结构,也进行同样的处理.

规则 3. 根据词性标注,将一些虚词删除,对依存关系进行变换.方法与上面的步骤类似,如果被删除的词没有父节点,那么直接删除该词.表 2 中列举了被删除虚词的词性,词性标注标准使用的是哈尔滨工业大学分词模块的词性标注标准.

规则 2. 对“的”字结构(依存关系为“DE”)、“地”字结构(依存关系为“DI”)、“得”字结构(依存关系为“DEI”)3 个依存关系进行变换.删除句子中出现的“的”、“地”和“得”(3 个字必须均为助词词性,且对应依存关系须是“DE”,“DI”,“DEI”),同时将删掉的助词的父节点直接指向删掉的助词的孩子节点,同时将依存关系修改为被删掉的词.如果该助词的父亲节点不唯一,不能进行转换.例如句子“我的名字是张三。”的原始依存树解析结果如图 4 所示.

从图 4 可以看出,“的”字的父亲节点为“名字”,它的孩子节点为“我”,并且只有一个父亲节点.符合我们的转换要求,我们将其转换,即将助词“的”删掉,直接连接“的”所关联的两个词“名字”和“我”.同时,将“的”

规则 4. 规定一个关系路径中至少包括两个结点.

规则 5. 关系路径中的依存关系数量最多为 9 个,即关系路径中包含的词个数最多为 8 个.这是因为在实验中发现,我们所使用的依存句法分析器对于长距离的词之间的依存关系分析不够准确.另外,对路径的长度进行限定,也可提升系统提取路径的效率.

规则 6. 限定关系路径两端的插槽词必须为实词.具体的词性如表 3 所示.

表 2 将被删除的词的词性

词性标注	e	wp	u	p	c	o
含义	叹词	标点符号	助词	介词	连词	拟声词

表 3 关系路径两端的插槽词的词性

a	d	m	n	nd	nh	ni	nl	ns	nt	nz	q	r	v
形容词	副词	数词	普通名词	方位名词	人名	机构名	处所名词	地名	时间词	其他专名	量词	代词	动词

规则 7. 不提取与已提取的路径的词顺序相反的路径. 这里我们在提取关系路径的时候, 与文献[5]中的规则不同.

规则 8. 仿照规则 1, 将分词时疑问词被分开的词进行合并. 同时将疑问词的词性标注为“iw”, 我们新追加的词性, 不是词性标注标准中的标准词性.

例如: “中国是在哪一年恢复了在联合国的合法

席位?”的原始依存分析结果如图 1 所示. 该句子中“哪一年”为 1 个完整的疑问词, 但分词时被分开为“哪”、“一”、“年”3 个词, 且除了这 3 个词之间的“内部关系”外, 只有“年”与其他词有 1 个“外部关系”, 所以可以将这 3 个词合并, 合并后的依存分析结果如图 6 所示.

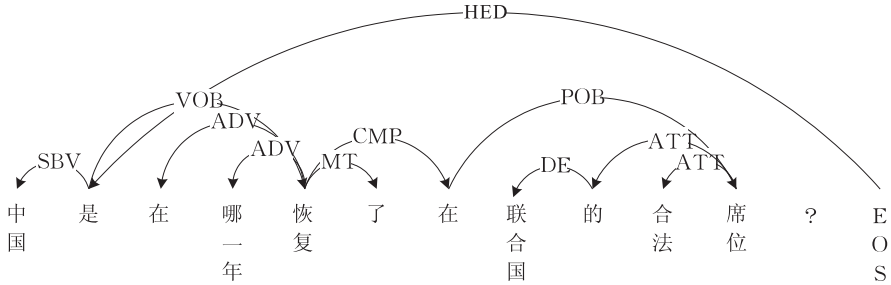


图 6 合并疑问词后的依存树

下面举 1 个利用我们的关系路径提取规则提取路径的例子. 例如从问题“中国是在哪一年恢复了在联合国的合法席位?”中提取得到的关键词之间的路径为:

- (1) (中国)ns:SBV:v←是→v:VOB:v(恢复);
- (2) (中国)ns:SBV:v←是→v:VOB:v→恢复→v:在:n(席位);
- (3) (中国)ns:SBV:v←是→v:VOB:v→恢复→v:在:n→席位→n:的:ni(联合国);
- (4) (中国)ns:SBV:v←是→v:VOB:v→恢复→v:在:n→席位→n:ATT:a(合法);
- (5) (恢复)v:在:n(席位);
- (6) (恢复)v:在:n→席位→n:的:ni(联合国);
- (7) (恢复)v:在:n→席位→n:ATT:a(合法);
- (8) (联合国)ni:的:n(席位);
- (9) (联合国)ni:的:n←席位→n:ATT:a(合法);
- (10) (合法)a:ATT:n(席位).

疑问词与关键词之间的路径为:

- (1) (中国)ns:SBV:v←是→v:VOB:v→恢复→v:ADV:iw(哪一年);
- (2) (是)v:VOB:v→恢复→v:ADV:iw(哪一年);
- (3) (哪一年)iw:ADV:v(恢复);

- (4) (哪一年)iw:ADV:v←恢复→v:在:n(席位);

- (5) (哪一年)iw:ADV:v←恢复→v:在:n→席位→n:的:ni(联合国);

- (6) (哪一年)iw:ADV:v←恢复→v:在:n→席位→n:ATT:a(合法).

4.3 路径间相似度的计算

这一节将具体介绍如何计算两个路径间的相似度. 我们首先分别找到路径各自所对应的最相似的“语料路径”(从文本语料中提取得到的路径), 然后利用文献[8]中提出的算法计算路径间的相似度.

计算一对路径的相似度计算公式如下:

$$\text{simP}(p_q, p_s) = \text{simL}(p_q, p\text{Inf}_q) \times \text{simI}(p\text{Inf}_q, p\text{Inf}_s) \times \text{simL}(p_s, p\text{Inf}_s) \quad (1)$$

其中, p_q 与 p_s 表示两个路径. $p\text{Inf}_q$ 表示与 p_q 最相似的某个语料路径; 类似地, $p\text{Inf}_s$ 表示与 p_s 最相似的某个语料路径.

$\text{simL}(p_q, p\text{Inf}_q)$ 表示利用 Lucene 全文检索工具得到的 p_q 和 $p\text{Inf}_q$ 的相似度; 类似地, $\text{simL}(p_s, p\text{Inf}_s)$ 表示利用 Lucene 全文检索工具得到的 p_s 和 $p\text{Inf}_s$ 的相似度; 如果 p_q 和 $p\text{Inf}_q$ 完全相同, 那么 $\text{simL}(p_q, p\text{Inf}_q)=1$; 类似地, 如果 p_s 和 $p\text{Inf}_s$ 完全相同, 则 $\text{simL}(p_s, p\text{Inf}_s)=1$.

$\text{simI}(p\text{Inf}_q, p\text{Inf}_s)$ 表示路径 $p\text{Inf}_q$ 与 $p\text{Inf}_s$

的相似度. 计算方法使用文献[8]中提出的互信息计算方法,我们将在下一节对该计算方法进行概要的阐述.

这里 $simL(p_i, pInf_i)$ 的计算方法是简单将路径中的依存关系和词看作“词袋”,利用 Lucene 的默认检索算法(向量空间模型算法的一种变形)计算得到的相似度.

这里用上面问题例子中提取得到的路径“ns:SBV:v←是→v:VOB:v”介绍如何使用 Lucene 全文检索工具检索与该路径最相似的语料路径.

首先,将该路径进行“分词”.即将该路径按照“→”或者“←”进行分词.经过分词后将得到下面的字符串:“ns:SBV:v”、“是”和“v:VOB:v”.根据这些字符串生成的查询关键词为“ns:SBV:v 是 v:VOB:v”,即将这些被分开的“词”以空格为分隔,在 Lucene 的检索语法中,这些关键词之间是“或”的关系.

然后,利用预先建立好的语料路径字符串的 Lucene 索引检索得到最相似的路径.建立索引时“分词”方法与上面描述的分词方法相同.

4.4 语料路径相似度计算

文献[8]提出了一种有效的“推理规则”提取算法.“推理规则”是指两个语义非常相近的路径,在它们出现的上下文中通常可以进行互换.文献[5]提出的算法的一个显著优点是可以根据大规模语料自动提取出句子的路径,并利用互信息来计算路径间的

SlotX:

层次	1	6.52309;	定义	1	5.78463;
环境	2	4.46362;	建筑物	1	5.9377;
模式	1	5.5289;	摩擦力	1	8.065;

SlotY:

产品	1	3.71832;	多样性	1	7.40526;
化学	1	5.47494;	环境	1	3.64685;
金属	1	4.86052;	巨头	1	9.40908;

上面例子中的第一列数字表示该行的插槽词在该插槽出现的频数.第二列数字表示该插槽与该词的互信息,即该词与该插槽连接的紧密程度.插槽词与对应的插槽的互信息的计算公式为

$$mi(p, Slot, w) = \log \frac{P(p, Slot, w)}{P(Slot)P(p|Slot)P(w|Slot)}$$
(2)

这里 p 表示一个路径, $Slot$ 为 $SlotX$ 或者 $SlotY$, w 是一个插槽词, $P(x)$ 表示概率.在计算时,利用频数来计算相应的概率.

利用 $|p, SlotX, w|$ 表示三元组 $(p, SlotX, w)$

相似度.从而可以得到“推理规则”.因为我们提出的基于依存关系的片段检索方法中使用了该算法计算一对路径间相似度,所以下面将对该算法进行简单介绍.

文献[8]提取推理规则时,使用了扩展的分布假设(extended distributional hypothesis).该假设来源于分布假设(distributional hypothesis),是 Harris 在 1985 提出的,其主要思想为:“倾向于在相同的上下文中出现的词,倾向于具有相似的含义.”.对该假设进行扩展后,得到了文献[8]中使用的扩展的分布假设:“如果两个路径倾向于出现在相同的上下文中,那么这两个路径的含义相似.”.对于路径来说,它的上下文即路径两端的插槽词.即:如果两个路径倾向于含有越多的相同的插槽词,则这两个路径在语义上就更相似.

为了计算路径之间的相似度,需要收集路径的插槽词的出现频数.对于一个给定的路径实例 p ,如果 p 连接了两个词 w_1 和 w_2 ,那么增加两个三元组 $\langle p, SlotX, w_1 \rangle$ 和 $\langle p, SlotY, w_2 \rangle$ 的频数.称 $\langle SlotX, w_1 \rangle$ 和 $\langle SlotY, w_2 \rangle$ 是 p 的特征.直觉上,两个路径共享越多的特征,它们就越相似.

可以利用三元组数据库来统计路径的特征的频数.

下面列出了一个路径的三元组数据情况:
n:SBV:v←有利于→v:VOB:n

方面	1	4.36296;	含量	1	4.95354;
局面	1	7.35556;	力	1	5.51887;
肉品	1	8.58684;	水流	1	6.58845

方面	1	4.16996;	公众	1	7.14094;
角度	2	6.36831;	结构	1	3.59716;
霉菌	1	8.7318;	体	1	4.72731

的出现频数; $|p, SlotX, *|$ 表示 $\sum_{w \in W} |p, SlotX, w|$, 即路径 p 的 $SlotX$ 的所有插槽词的频数和; $|*, *, *|$ 表示 $\sum_{p, Slot, w} |p, Slot, w|$, 即所有路径所有插槽对应的所有词的频数和.那么一个三元组 $(p, SlotX, w)$ 的互信息可以用下面的公式计算:

$$mi(p, Slot, w) = \log \frac{\frac{|p, Slot, w|}{|*, *, *|}}{\frac{|*, Slot, *|}{|*, *, *|} \frac{|p, Slot, *|}{|*, Slot, *|} \frac{|*, Slot, w|}{|*, Slot, *|}} =$$

$$\log \frac{|p, Slot, w| \times |*, Slot, *|}{|p, Slot, *| \times |*, Slot, w|} \quad (3)$$

当三元组数据库创建之后,两个路径之间的相似度就可以用与计算两个词相似度相同的方式进行计算。一对插槽 $Slot_1=(p_1, s)$ 和 $Slot_2=(p_2, s)$ 的相似度利用下面的公式进行计算:

$$sim(Slot_1, Slot_2) = \frac{\sum_{w \in T(p_1, s) \cap T(p_2, s)} mi(p_1, s, w) + mi(p_2, s, w)}{\sum_{w \in T(p_1, s)} mi(p_1, s, w) + \sum_{w \in T(p_2, s)} mi(p_2, s, w)} \quad (4)$$

这里 p_1 和 p_2 表示两个路径, s 表示一个插槽, $T(p_i, s)$ 表示路径 p_i 的 s 插槽词的集合。

两个路径 p_1 和 p_2 的相似度利用 $SlotX$ 和 $SlotY$ 相似度的几何平均值来定义:

$$S(p_1, p_2) = \sqrt{sim(SlotX_1, SlotX_2) \times sim(SlotY_1, SlotY_2)} \quad (5)$$

其中, $SlotX_i$ 或者 $SlotY_i$ 分别是路径 p_i 的两个插槽。

我们在实验中使用 863 文本语料作为提取路径的语料库,它包含了 3599 个文本文件,大小为 20.4MB,使用上面 4.2 节提出的更适合汉语特点的路径提取规则,我们从该语料中提取了 13992467 个路径(在后续章节中,我们将这些路径称为“语料路径”)用于计算一对路径间的相似度。

5 候选答案的特征提取、约简及分类

5.1 特征约简

特征约简可以理解为,在保证分类或决策能力不变的条件下,删除冗余特征。粗糙集(rough set)理论是一种具有模糊边界的数据挖掘方法,主要用于特征约简、决策规则生成以及预测等方面。著名的 Rosetta(<http://www.idi.ntnu.no/~aleks/rosetta/>)就是一个基于粗糙集理论框架的表格逻辑数据工具,它提供了多种数据预处理功能如决策表补齐、决策表离散化等及其算法,同时提供了粗糙集中常见的约简和规则的获取算法,支持从数据预处理到预测和分析规则的全过程,是一个很好的粗糙集理论软件和实验平台。

我们在实验中利用 Rosetta 软件对预想的 32 个特征进行约简,最后保留了 18 个,保留的特征将在 5.2 节详细介绍。由于本文用到的特征都是连续型实数,所以在特征约简前先要做数据离散化处理,我们选择的离散化算法为 Entropy/MDL,并选择遗传算法(genetic algorithm)用于特征约简。

5.2 候选答案所在句子的特征提取

通常情况下,一个句子可以完整地表达一个较完整的事实。所以我们假设候选答案所在的句子的相关特征能够有效地反映出该候选答案是否为正确答案。特征主要分为数量类、距离类、顺序类和句法结构类特征。下面结合具体例子详细介绍。

例:自然语言问题为:“中国是在哪一年恢复了在联合国的合法席位?”,查询词为:“中国 恢复 联合国 合法 席位”,得到的文本片断为(查询词用黑体标出,候选答案用黑斜体标出):

〈片断〉:李肇星撰文纪念**中国恢复联合国合法席位** 35 周年——
1971 年 10 月 25 日,第二十六届联合国大会以压倒性多数通过第 2758 号决议,决定**恢复新中国在联合国的合法席位**。这个决议草案是由阿尔巴尼亚、阿尔及利亚、缅甸、锡兰(今斯里兰卡)、古巴、赤道几内亚、几内亚、伊拉克、马里、毛里塔尼亚、尼泊尔、巴基斯坦、...

经过分词和词性标注后的结果为(/ * 表示词性标注):

〈问题〉:中国/ns 是/v 在/p 哪/r 一年/mq 恢复/v 了/u 在/p 联合国/nt 的/u 合法/v 席位/n ?/w
〈片断〉:李肇星/nr 撰文/v 纪念/v 中国/ns 恢复/v 联合国/nt 合法/v 席位/n 35 周年/mq -/w -/w
1971 年 10 月 25 日/t ,/w 第二十六届/mq 联合国大会/nz 以/p 压倒性/b 多数/mq 通过/p 第 2758 号/mq 决议/n ,/w 决定/v 恢复/v 新/a 中国/ns 在/p 联合国/nt 的/u 合法/v 席位/n . /w 这个/r 决议/n 草案/n 是/v 由/p 阿尔巴尼亚/ns ,/w 阿尔及利亚/ns ,/w 缅甸/ns ,/w 锡兰/ns (/w 今/n 斯里兰卡/ns)/w ,/w 古巴/ns ,/w 赤道几内亚/ns ,/w 几内亚/ns ,/w 伊拉克/ns ,/w 马里/ns ,/w 毛里塔尼亚/ns ,/w 尼泊尔/ns ,/w 巴基斯坦/ns /w .../w

此处我们利用一些启发式规则将连续的时间词以及表示日期范围的词合并,因为它们结合在一起才表示一个完整的时间。例如:“1984 年 4 月 26 日至 5 月 1 日”直接分词后的结果为:“1984 年/t 4 月/t 26 日/t 至/p 5 月/t 1 日/t”,“/t”是时间词的词性标注、“/p”表示介词。经过我们的合并处理后,将得到“1984 年 4 月 26 日至 5 月 1 日/t”。这个例子所匹配的启发式规则为 $\langle /t | m, /p, /t | n \rangle \rightarrow \langle /t \rangle$, 其中 $\langle /t | m, /p, /t | n \rangle$ 表示词性标注向量, m 和 n 表示时间词的个数,合并处理后将得到“1984 年 4 月 26 日至 5 月 1 日/t”。

上面这个例子中,查询关键词的个数为 5;得到的文本片断的句子数为 3(文本片断的标题也当作一个句子);问题类型为时间类型,该类型对应的答案应为表示时间的词。在上述例子中只有“1971 年 10 月 25 日”的词性符合要求,即候选答案词只有一个,提取的特征为第二个句子所含有的特征。

5.2.1 数量类特征

数量类特征包括:

(1) 句子中匹配词的个数占查询词个数的比例. 该特征反映该候选答案所在的句子与问题在词汇匹配层面的相似度.

针对上面的例子, 句子中出现的匹配词数为 5, 所以可得此特征的数值为: $5/5=1$.

(2) 匹配的名词(动词、数词或数量词)的个数占查询词中名词(动词、数词或数量词)个数的比例. 该特征反映指定词性的词与问题在词汇匹配层面的相似度.

针对上面的例子, 查询词中的名词为: “中国”、“联合国”、“席位”, 共 3 个. 因句子中出现了所有的查询词中的名词, 所以匹配的名词的个数占查询词中名词个数的比例为: $3/3=1$. 同理, 查询词中的动词为: “恢复”、“合法”, 共 2 个, 匹配的动词的个数占查询词中动词个数的比例为: $2/2=1$. 查询词中没有出现数词或数量词, 所以此项特征值为 0.

5.2.2 距离类特征

距离类特征包括:

(1) 候选答案及其前面出现的第一个匹配名词(动词、数词或数量词)间的距离占句子长度的比例. 该特征可反映候选答案与前面指定词性的匹配词的距离远近, 同时也可表示前面是否出现了指定词性的匹配词.

针对上面的例子, 候选答案前没出现匹配的名词、动词、数词或数量词, 所以这三个特征值均为 0.

(2) 候选答案及其后面出现的第一个匹配名词(动词、数词或数量词)间的距离占句子长度的比例. 该特征可反映候选答案与后面指定词性的匹配词的距离远近, 同时也可表示前面是否出现了指定词性的匹配词.

上述两个距离类特征中的距离、句子长度以及下面特征中将会提到的窗口宽度等都是按照分词处理后词(包含标点符号)的个数来计算的. 上述例子的句子长度为 21, 候选答案后面出现的第一个匹配名词为“中国”, 其与候选答案间的距离占句子长度的比例为: $(15-1)/21=0.6667$. 类似地, 候选答案及其后面出现的第一个匹配动词间的距离占句子长度的比例为: $(13-1)/21=0.5714$.

(3) 匹配词紧密度特征. 此特征反映的是句子中所有匹配词的出现的紧密程度, 计算公式如下:

匹配词紧密度 = 包含所有匹配词的最小窗口宽度 / 窗口中的匹配词的总数 (6)

其中, 匹配词总数的计算允许同一个匹配词出现多次. 对于上面的例子, 匹配词的窗口宽度为 8, 匹配词的频数为 5, 所以此项特征值为 $8/5=1.6$.

(4) 候选答案与最近一个匹配词间的距离. 我们借鉴了文献[6]中的答案提取算法, 提取了该特征.

针对上面的例子, 与候选答案词距离最近的词为“恢复”, 距离为 13.

(5) 候选答案与各匹配词间的距离平均值. 我们借鉴了文献[7]中的答案提取算法, 提取了该特征.

针对上面的例子, 该特征值为 $(13-1+15-1+17-1+19-1+20-1)/5=15.8$.

5.2.3 顺序类特征

顺序类特征包括:

(1) 句子词序列和问题词序列的顺序相似度特征. 此特征考察句子中的匹配词和候选答案出现的顺序是否与其在问题的顺序相同, 并用顺序相同的词数占查询词数的比例来度量. 我们把候选答案看作与问题中的疑问词相匹配. 将句子中的匹配词和候选答案组成一个词序列, 记为句子词序列; 同时将问题中查询词和疑问词组成另一个词序列, 记为问题词序列. 计算公式为

顺序相似度 = 句子词序列中与问题词序列中出现顺序相同的词的个数 / 问题词序列中的词的个数 (7)

对于上面的例子, 句子中的“1971 年 10 月 25 日 恢复 联合国 合法 席位”的出现顺序与问题中的“哪一年 恢复 联合国 合法 席位”的出现顺序相同, 出现顺序相同的词数为 5 个, 所以该特征值为 $5/6=0.8333$.

(2) 匹配词序列与查询词序列的顺序相似度特征. 此特征与“句子词序列和问题词序列的顺序相似度特征”类似, 只是不再考虑句子中的候选答案和问题中的疑问词. 计算公式如下:

顺序相似度 = 匹配词序列与查询词序列中出现顺序相同的词的个数 / 查询词序列中的词的个数 (8)

对于上面的例子, 句子中的“恢复 联合国 合法 席位”的出现顺序与问题中的“恢复 联合国 合法 席位”的出现顺序相同, 与问题中出现顺序相同的匹配词的个数为 4 个, 所以该特征值为 $4/5=0.80$.

(3) 各匹配词与候选答案的顺序关系与各查询词和疑问词的顺序关系相同的个数占查询词个数的比例. 此特征中, 同样将候选答案看作与问题中的疑问词相匹配. 计算方法为: 将句子中的候选答案与各匹配词按其顺序关系分别构成序偶, 记为候选答案

序偶集合;将问题中的疑问词与各查询词按其顺序关系分别构成序偶,记为疑问词序偶集合.然后考察候选答案序偶集合与疑问词序偶集合中相匹配的序偶的个数,将得到的数值除以查询词个数即得到此特征的值.此特征考察句子中的词的顺序和问题中词的顺序的相似程度.因为通常情况下,位置关系相同的个数越多,则表示候选答案所在的句子与问题有更高的相似度.

针对上面的例子,原始问题中“中国”出现在疑问词“哪一年”的前面.但是在候选答案所在的句子中,“中国”出现在候选答案词的后面,所以与原始问题中的位置关系不一致.除了匹配词“中国”与候选答案词的位置关系与问题中的不一致外,其余的匹配词与候选答案词的位置关系与原始问题中的均一致.所以该特征值为 $4/5=0.80$.

5.2.4 句法结构类特征

在本文的第4节我们已经阐述了如何计算两个路径之间的相似度.在本章节中我们将把路径间的相似度作为特征用于候选答案的分类.为了能更简洁地描述句法结构类特征,下面作一些简单的定义:

问句路径.从问题句子中提取得到的路径;

句子路径.从候选答案所在的句子中提取得到的路径.

(1)问句和句子的匹配词之间的路径的相似路径对数量平均值.该特征反映了问句与候选答案所在的句子之间的句法相似程度.计算方法如下:

针对句子中的每两个匹配词,分别在问句和句子中提取路径;

按照4.3节的算法,计算问句路径与句子路径的相似度,如果相似度值大于0,则记为1,表示问句路径与句子路径相似;否则为0.如果在句子中某个匹配词出现多次,那么将得到多个句子路径.此时,利用问句路径与多个句子路径的相似度的最大值来确定问句路径与句子路径是否相似.

对于上面两个步骤得到的问句路径与句子路径相似的数量,求平均值即得到了该特征值.

针对上面的例子,首先对候选答案所在的句子进行依存树解析,并且从问句和候选答案所在的句子中提取匹配词之间的路径.问句路径用Q作为标号的前缀,句子路径用S作为标号的前缀.共有10组如表4.

表 4 匹配词之间的路径相似度实例

编号	路径	相似度
Q1	(恢复)v;在:n→席位→n;的:ni(联合国)	
S1.1	(联合国)ni:ATT:n←大会←n;SBV:v←决定→v;VOB:v(恢复)	0
S1.2	(恢复)v;VOB:v←决定→v;在:n→席位→n;的:ni(联合国)	32.45
Q2	(中国)ns;SBV:v←是→v;VOB:v→恢复→v;在:n→席位→n;的:ni(联合国)	
S2.1	(联合国)ni:ATT:n←大会←n;SBV:v←决定→v;VOB:ns(中国)	0
S2.2	(中国)ns:VOB:v←决定→v;在:n→席位→n;的:ni(联合国)	0
Q3	(联合国)ni;的:n(席位)	
S3.1	(联合国)ni:ATT:n←大会←n;SBV:v←决定→v;在:n(席位)	0
S3.2	(联合国)ni;的:n(席位)	151.61
Q4	(联合国)ni;的:n←席位→n;ATT:a(合法)	
S4.1	(联合国)ni:ATT:n←大会←n;SBV:v←决定→v;在:n→席位→n;ATT:a(合法)	0
S4.2	(联合国)ni;的:n←席位→n;ATT:a(合法)	17.43
Q5	(中国)ns;SBV:v←是→v;VOB:v(恢复)	
S5	(恢复)v;VOB:v←决定→v;VOB:ns(中国)	23.16
Q6	(恢复)v;在:n(席位)	
S6	(恢复)v;VOB:v←决定→v;在:n(席位)	40.32
Q7	(恢复)v;在:n→席位→n;ATT:a(合法)	
S7	(恢复)v;VOB:v←决定→v;在:n→席位→n;ATT:a(合法)	0
Q8	(中国)ns;SBV:v←是→v;VOB:v→恢复→v;在:n(席位)	
S8	(中国)ns:VOB:v←决定→v;在:n(席位)	0
Q9	(中国)ns;SBV:v←是→v;VOB:v→恢复→v;在:n→席位→n;ATT:a(合法)	
S9	(中国)ns:VOB:v←决定→v;在:n→席位→n;ATT:a(合法)	19.87
Q10	(合法)a:ATT:n(席位)	
S10	(合法)a:ATT:n(席位)	6751.95

针对表4中的10组路径相似度数据,可知相似的路径对数量为7,可得该特征的值 $7/10=0.7$.

(2)问句中疑问词与匹配词之间的路径与句子

中候选答案与匹配词之间的路径的相似度评分平均值.该特征主要衡量候选答案词与匹配词的句法结构与问句中疑问词与匹配词的句法结构的相似程

度. 在计算相似度时需要补充说明的是, 因为疑问词的词性被我们标注为“iw”, 但是在计算路径相似度时, 会将其转换为该疑问词对应的答案词的词性. 如该例子中的疑问词为询问时间, 故在计算相似度时,

会将其转换为“nt”(时间词的词性).
仿照上面的计算方法, 得到如表 5 所示的 5 组路径数据.

表 5 疑问词与匹配词之间的路径相似度实例

编号	路径	相似度
Q1	(中国)ns;SBV:v←是→v;VOB:v→恢复→v;ADV:iw(哪一年)	0
S1	(1971 年 10 月 25 日)nt;ADV:v←决定→v;VOB:ns(中国)	
Q2	(哪一年)iw;ADV:v(恢复)	
S2	(1971 年 10 月 25 日)nt;ADV:v←决定→v;VOB:v(恢复)	
Q3	(哪一年)iw;ADV:v←恢复→v;在:n(席位)	45.18
S3.1	(1971 年 10 月 25 日)nt;ADV:v←决定→v;在:n(席位)	
Q4	(哪一年)iw;ADV:v←恢复→v;在:n→席位→n;的:ni(联合国)	
S4.1	(1971 年 10 月 25 日)nt;ADV:v←决定→v;SBV:n→大会→n;ATT:ni(联合国)	
S4.2	(1971 年 10 月 25 日)nt;ADV:v←决定→v;在:n→席位→n;的:ni(联合国)	51.47
Q5	(哪一年)iw;ADV:v←恢复→v;在:n→席位→n;ATT:a(合法)	
S5	(1971 年 10 月 25 日)nt;ADV:v←决定→v;在:n→席位→n;ATT:a(合法)	
		19.83

针对表 5 中的 5 组路径相似度数据, 可知相似的路径对数量为 3, 可得该特征的值为 $3/5=0.6$.

5.3 应用分类器分类候选答案

因为我们把答案提取问题看成是候选答案的分类问题, 所以选取一个合适的分类器是一个首要而关键的任务. 为了考察不同分类器针对我们提取特征的分类性能差别, 我们选取了 3 种分类器: BP 人工神经网络(ANN)、朴素贝叶斯(NB)和支持向量机(SVM).

神经网络适用于需要较多符号表示的问题, 如决策树学习任务, 且 Shavlik 等(1991)和 Weiss 以及 Kapouleas(1989)通过实验评估证明人工神经网络和决策树学习算法通常产生精度大体相当的结果^[15], 更重要的是, 神经网络学习的实例可以是任何实数构成的向量, 这正符合我们对分类器的要求.

朴素贝叶斯分类是基于统计学的分类方法, 被广泛用于文本分类领域. 在大规模的训练数据情况下, 有非常高的准确率和执行速度. 该分类算法建立在各个属性值相互独立的前提下, 即在属性之间, 假定其不存在依赖关系. 该分类器针对属性值为离散值的情况, 可以直接利用基于频数的方法来计算概率. 针对属性值为连续值的情况, 通常假定该属性服从高斯分布, 利用高斯分布的概率函数计算相应的概率. 我们提取的特征属性均为实数(连续值), 所以应用该分类器前, 我们首先对属性进行基于熵的离散化处理, 离散化处理所使用的方法与 5.1 节相同.

支持向量机也是一种在文本分类领域有较好性能的分类器. 该方法在样本规模较小的前提下, 与现有的各种分类器相比, 在许多应用领域有更好的性

能. 最初的 SVM 特别适合两类目标分类问题, 支持属性为连续值的分类问题, 这正符合我们的应用对分类器的要求. SVM 的性能受核参数和误差惩罚参数的影响. 我们在实验过程中, 针对每个问题类型的训练样本, 采用交叉验证的方法均查找了最优的核参数和误差惩罚参数.

6 实 验

实验主要分为两部分, 第一部分是候选答案分类实验, 考察 3 种分类器对答案分类的贡献以及句法结构类特征分析技术对分类器的作用; 第二部分是答案提取算法的对比实验, 考察本文提出的基于句法结构类特征分析和分类技术的答案提取算法的优势.

6.1 实验数据和相关工具

实验所使用的问题集是哈尔滨工业大学信息检索研究室的问答系统问题集(http://ir.hit.edu.cn/demo/ltip/Sharing_Plan.htm), 问题集已经过问题分类处理和人工校对, 含有问题类型信息. 由于此问题集中有些问题并没有确定的答案或者答案目前还存在争议(例如“人们最喜欢攀登哪座山?”), 我们选择了部分有确定答案的问题作为实验数据集.

中文分词工具使用的是天津海量科技发展有限公司开发的海量智能分词(研究版)(<http://www.hylanda.com/cgi-bin/download/download.asp?id=8>), 其分词速度快, 且准确率较高、稳定性好, 可以进行分词、词性标注和命名实体识别. 检索 Google 文本片断(snippet)所使用的工具是 Google 实验室

提供的 Google SOAP Search API (<http://code.google.com/apis/soapsearch/>). 依存树解析软件使用的是哈尔滨工业大学信息检索研究室语言技术平台程序库的依存句法分析模块 (http://ir.hit.edu.cn/demo/ltip/Sharing_Plan.htm). 人工神经网络的训练和仿真工具使用的是 MATLAB R2007a 的神经网络工具箱. 我们使用 SVM 算法的开源实现 libsvm (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>) 作为实验中使用的 SVM 分类程序. libsvm 是台湾大学林智仁博士等开发设计的易于使用、快速有效的通用 SVM 软件包, 可以解决分类问题、回归问题以及分布估计等问题. 它不仅提供了 LIBSVM 的 C++ 语言的算法源代码, 还提供了 Python、Java、R、MATLAB、Perl、Ruby、LabVIEW 以及 C#.net 等各种语言的接口.

6.2 候选答案分类实验

各问题类型中问题数量及其候选答案的数量的具体分布情况如表 6 所示.

表 6 训练和测试问题集				
问题类别	训练问题数	测试问题数	候选答案训练样本数量	候选答案测试样本数量
时间 (TIME)	45	30	9451	6030
人物 (HUM)	42	29	8417	5797
数字 (NUM)	43	30	9004	6109
组织 (ORG)	42	25	8820	5099
地点 (LOC)	44	28	9243	5615
实体 (OBJ)	41	26	8391	5362

每个问题对应 100 个 Google 返回的文本片段. 候选答案均是从文本片段中提取出来的, 且经过人工校对后都有一个类标签, 1 表示是正确答案, 0 表示是错误答案.

我们使用准确率 (*Precision*)、召回率 (*Recall*) 和 F_{β} -score 来评估分类的准确率. 它们的计算公式

如式(9)、式(10)和式(11)所示.

$$\text{准确率} (Precision) = \frac{\text{分类正确的候选答案数量}}{\text{总候选答案数量}}$$

(9)

$$\text{召回率} (Recall) = \frac{\text{总正确答案数量}}{\text{总正确候选答案数量}}$$

(10)

$$F_{\beta}\text{-score} = \frac{(\beta^2 + 1) \cdot Recall \cdot Precision}{\beta^2 \cdot Recall + Precision}$$

(11)

对于式(11), 在实验中 β 的取值均为 1. 为了考察 5.2.4 节提出的句法结构类特征对于分类性能的贡献, 我们分别应用朴素贝叶斯、BP 人工神经网络、支持向量机分类方法各做了两方面的实验: 一是仅使用 5.2.1、5.2.2 和 5.2.3 节的特征, 其结果分别记为 Bayes-1、BP-1 和 SVM-1; 二是利用 5.2 节提出的全部的特征 (即包括句法结构类特征) 对候选答案进行分类, 其结果分别记为 Bayes-2、BP-2 和 SVM-2. 其准确率、召回率和 F_{β} -score 比较分别如图 7~9 所示. 仅从 3 个图中的 Bayes-1、BP-1 和 SVM-1 可以看出, 不使用句法结构类特征的情况下, 分类的 F_{β} -score 的均值在 69%~76% 之间, 说明分类性能已经比较令人满意. 而 3 个图中 Bayes-1、BP-1 和 SVM-1 与 Bayes-2、BP-2 和 SVM-2 对照的平均值可以看出, 添加了 5.2.4 节中的句法结构类特征后, 无论使用任何分类算法 F_{β} -score 的绝对数值有约 5% 的提升. 这在原有性能就比较好的情况下, 是比较难能可贵的. 经过我们分析, 添加了句法依存特征后, 性能提升不很明显的原因在于: (1) Google API 返回的片段中的句子长度较短, 且有时完整的句子被截断, 影响了路径的提取和相似度的计算; (2) 多数问题生成的查询关键词的个数较少, 造成句法结构类特征的数据有些稀疏.

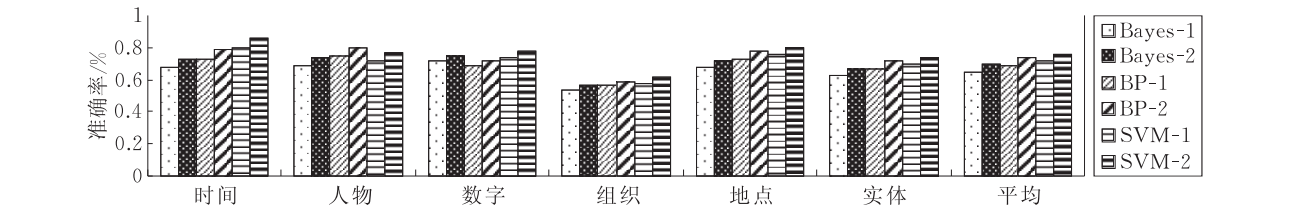


图 7 使用与不使用句法结构类特征候选答案分类的准确率比较

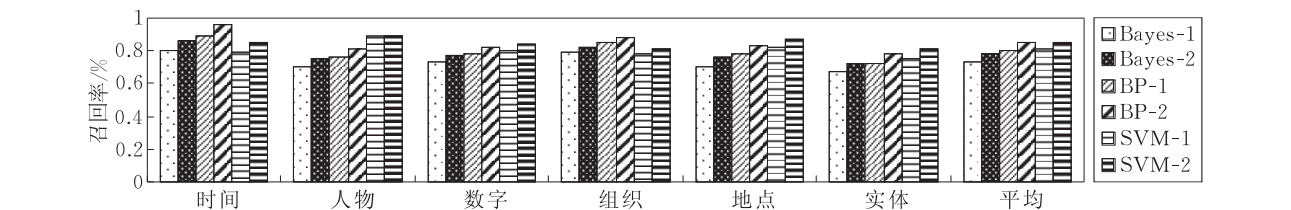


图 8 使用与不使用句法结构类特征候选答案分类的召回率比较

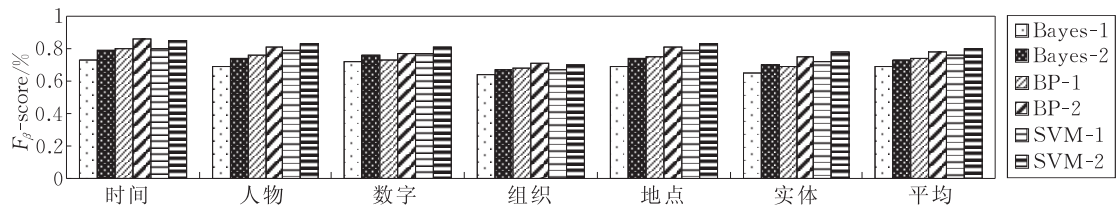


图 9 使用与不使用句法结构类特征候选答案分类的 F_{β} -score 比较

从图 7~9 还可以看出,3 种分类方法的分类性能相差不大,朴素贝叶斯分类性能稍差些,支持向量机分类方法性能最好.对于使用句法结构类特征的 3 种分类方法,其平均准确率在 69%~76%之间,平均召回率在 78%~85%之间, F_{β} -score 的均值在 73%~80%之间.另外,从图中看出各问题类别之间的实验数据差别不大,说明针对不同的问题类型分类性能的稳定性较好.

经我们分析,朴素贝叶斯分类性能稍差的主要原因有:(1)属性之间的独立性不能完全满足.如 5.2.1 节的两个数量类特征有一定的相关性.(2)数据经过离散化处理,丢失了部分信息.而对于支持向量机分类器,由于其在样本规模较小情况下的优势,使其性能略高于 BP 人工神经网络.

在下面与其他答案提取方法的对比实验中,我们选择支持向量机分类器,用它代表我们提出的基于句法结构类特征分析和分类技术的答案提取算法,并记为 SS&C-Based.

6.3 与其他答案提取算法的对比实验

在这一节里我们将针对不同的答案提取算法进行性能对比评测实验.我们选取的基准答案提取算法为文献[6]中提出的基于信息检索和信息抽取的答案提取技术.对比算法为文献[4]中提出的基于分类的答案提取技术和文献[2]中提到的基于模式匹配的答案提取技术.

我们对实验结果评价的时候采用准确率和平均排序倒数(Mean Reciprocal Rank, MRR)两个评测标准.公式分别为式(12)和式(13):

准确率 = 回答正确的问题数 / 问题总数 (12)

$$MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{\text{正确答案在系统给出的排序结果中的位置}} \quad (13)$$

式(13)中 N 表示测试集中的问题个数.即对每个问题而言,把标准答案在被评价系统给出结果中的排序取倒数作为它的准确度,再对所有的问题取平均.如果正确答案存在于系统给出的排序结果中的多个位置,以排序最高的位置计算;如果正确答案不在系统给出的排序结果中,则该问题的准确率为 0.

实验中的文本片断检索部分,使用 Google 搜索引擎返回的含有正确答案的前 10 个文本片断作为答案提取的来源文档.每个问题返回 5 个答案.

表 7 所示为本文中提出的基于句法结构类特征分析和分类技术的答案提取算法 SS&C-Based(即 6.2 节中的 SVM-2)与其他答案提取算法的性能对比评测数据.其中的 BaseLine 表示基于信息检索和信息抽取的答案提取技术;MEM-Based 表示文献[4]中提出的基于实例的答案提取算法;SYP-Based 为文献[2]中提出的基于无监督学习的问答模式抽取技术.表中的 Pre 表示准确率的值.

从表 7 的数据可以看出,本文提出的算法与文献[2]中提出的算法相比,性能有一定的提升.我们认为原因在于我们的算法不仅考虑了句子中包含的多种简单特征,更考虑了句法结构类的特征,且句法结构相似度的计算是利用大规模的语料的上下文来进行评估的,而文献[2]算法的聚类相似度计算公式稍显简单;而且就算法的实现难度和问题类型的扩展性而言(见第 2 节相关工作),我们的算法更具有优势.

表 7 多种答案提取算法的性能评测数据

问题类别		时间(TIME)	人物(HUM)	数字(NUM)	组织(ORG)	地点(LOC)	实体(OBJ)	平均值	比基准提高程度/%
BaseLine	Pre	0.300	0.379	0.267	0.280	0.393	0.346	0.327	-
	MRR	0.331	0.402	0.295	0.314	0.418	0.351	0.352	-
MEM-Based	Pre	0.400	0.483	0.333	0.360	0.536	0.423	0.422	29.0
	MRR	0.439	0.505	0.365	0.429	0.571	0.446	0.459	30.5
SYP-Based	Pre	0.667	0.621	0.500	0.520	0.643	0.577	0.588	79.5
	MRR	0.692	0.664	0.591	0.595	0.671	0.623	0.639	81.7
SS&C-Based	Pre	0.600	0.690	0.567	0.560	0.714	0.615	0.624	90.6
	MRR	0.653	0.738	0.651	0.670	0.776	0.721	0.701	99.3

MEM-Based 算法的性能在多种算法中居于中间水平,比基准算法性能提高约 30%。而我们提出的算法(SS&C-Based)要比基准算法性能提高约 91%。这说明,我们提出的算法是有效的,且可大幅度改善中文问答系统的答案提取性能。

7 结论与展望

中文表达方式灵活多样,句法结构复杂,这给问答系统的答案提取带来了极大的挑战。利用句法依存树解析工具,可以有效地提取出句子的句法结构和答案模板,且一些工作^[2,8-13]已经证明了该方法应用到问答系统中的可行性和有效性。但如果仅仅从句法结构分析入手去解决答案提取问题,一些重要的限定信息将丢失,且限制了答案提取算法的应用范围。

本文提出了一种新的基于句法结构类特征分析和分类技术的答案提取算法,将答案提取问题视为分类问题。首先根据问题类型,利用命名实体识别技术,将文本片断中的候选答案识别出来。然后利用候选答案所在的句子的数量类、距离类、顺序类和句法结构类特征训练分类器,最后将训练得到的分类模型用于答案提取。实验结果证明,我们的方法的性能优于目前典型的算法。

虽然取得了一定的成果,但是对实验使用的片段检索部分以及训练测试样本,我们仅使用了 Google API 返回的相关片段。Google API 的片段含有的句子较少,且有时完整的句子被断开,不利于句法结构的提取。下一步我们将在更多的片段检索算法上进行相关的实验。另外,5.2.4 节中描述的句法结构类特征依赖于语料路径(见 4.4 节)的规模,我们下一步将在更大规模的语料上进行实验,评估算法的性能。

参 考 文 献

- [1] Moldovan D I, Pasca M, Harabagiu S M, Surdeanu M. Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems (TOIS)*, 2003, 21(2): 133-154
- [2] Wu You-Zheng, Zhao Jun, Xu Bo. Unsupervised answer pattern acquisition. *Journal of Chinese Information Processing*, 2007, 21(2): 69-76(in Chinese)
(吴友政, 赵军, 徐波. 基于无监督学习的问答模式抽取技术. *中文信息学报*, 2007, 21(2): 69-76)
- [3] Sun Ang, Jiang Ming-Hu, Ma Yan-Jun. A maximum entropy model based answer extraction for Chinese question answering. *Fuzzy Systems and Knowledge Discovery*, 2006, 4223: 1239-1248
- [4] Sun Ang, Jiang Ming-Hu, Ma Yan-Jun. An instance-based approach for pinpointing answers in Chinese question answering//*Proceedings of the 8th International Conference on Signal Processing (ICSP2006)*. Beijing, China. 2006
- [5] Ittycheriah, Roukos S. IBM's statistical question answering system -TREC 11//Voorhees E M, Buckland Lori P. *NIST Special Publication: SP 500-251*. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, Maryland, 2002
- [6] Yang Hui, Chua Tat-Seng. The integration of lexical knowledge and external resources for question answering//Voorhees E M, Buckland Lori P. *NIST Special Publication: SP 500-251*. Department of Commerce, National Institute of Standards and Technology. Gaithersburg, Maryland, 2002
- [7] Li Xiao-Yan, Croft W Bruce. Evaluating question-answering techniques in Chinese//*Proceedings of the 1st International Conference on Human Language Technology Research*. Morristown, NJ, USA, 2001: 1-6
- [8] Lin De-Kang, Pantel Patrick. Discovery of inference rules for question-answering. *Natural Language Engineering*, 2001, 7(4): 343-360
- [9] Zhang Dell, Lee Wee Sun. Web based pattern mining and matching approach to question answering//Voorhees E M, Buckland Lori P. *NIST Special Publication: SP 500-251*. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, Maryland, 2002
- [10] Brill Eric, Lin Jimmy, Banko Michele, Dumais Susan, Ng Andrew. Data-intensive question answering//Voorhees E M, Harman D K. *NIST Special Publication 500-250*. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, Maryland, 2001: 363-370
- [11] Soubbotin M M, Soubbotin S M. Patterns of potential answer expressions as clues to the right answers//Voorhees E M, Harman D K. *NIST Special Publication 500-250*. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, Maryland, 2001: 293-302
- [12] Soubbotin M M, Soubbotin S M. Use of patterns for detection of likely answer strings: A systematic approach//Voorhees E M, Buckland Lori P. *NIST Special Publication: SP 500-251*. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, Maryland, 2002
- [13] Ravichandran D, Hovy E. Learning surface text patterns for a question answering system//Eugene Charniak, Dekang Lin. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, USA, 2002: 41-47
- [14] Liu Ting, Ma Jin-Shan, Li Sheng. Building a dependency treebank for improving Chinese parser. *Journal of Chinese Language and Computing*, 2006, 16(4): 207-224

[15] Mitchell Tom M. Machine Learning. Beijing: China Machine Press, 2003: 62-62(in Chinese)

(Mitchell Tom M 著, 曾华军, 张银奎等译. 机器学习. 北京: 机械工业出版社, 2003: 62-62)



HU Bao-Shun, born in 1981, Master. His main research interest is information retrieval.

WANG Da-Ling, born in 1962, Ph. D. , professor. Her main research interest is search engine technology.

YU Ge, born in 1962, Ph. D. , professor, P. D. supervisor. His main research interests include database and relevant technology.

MA Ting, born in 1981, master. Her main research interest is text mining.

Background

The problem discussed in this paper belongs to answer extraction in Question Answering System which is relevant to search engine technology. Recent years, Question Answering System is becoming a focus researching area. But the first contest on Chinese Question Answering System was held in 2005(NTCIR5). Currently the performance of Chinese QA is not well enough for users' requirements.

This paper presents a new answer extraction method for Chinese QA. Comparing to the typical answer extraction method, the performance has been improved.

The research in this paper belongs to the National Natural Science Foundation of China "Research of Deduction Model for Users' Motivation Orienting to New Generation Search Engine" grant No. 60573090.

New generation search engine has such characteristics as interactive searching, classific navigation, accurately related

querying, and rapid updating. Moreover, it pays attention to personalized and intelligent services. So studying on related technologies with new generation search engine such as personalized search, users' model, QA system, users' behavior and motivation, is helpful for improving the quality of search engines.

Now, the research group has achieved some productions. More than 40 papers have been published in "Chinese Journal of Computer", "Journal of Software", "Journal of Computer Research and Development", "Lecture Notes in Computer Science", and other Journals, and WISE, WAIM, APWeb, NDBC, DBAT, and other conferences.

QA system is a more effective and personalized new generation search engine, and the research content of this paper is an important part of new generation engine research.