问题分类中基于句法和语义信息的特征选择

袁晓洁,师建兴,宁 华,于士涛
YUAN Xiao-jie,SHI Jian-xing,NING Hua,YU Shi-tao

南开大学 信息技术科学学院,天津 300071
College of Information Technical Science,Nankai University,Tianjin 300071,China
E-mail:stein@mail.nankai.edu.cn

YUAN Xiao-jie,SHI Jian-xing,NING Hua,et al.Feature selection using syntactic and semantic information in question classification.Computer Engineering and Applications,2008,44(33):144-147.

Abstract: Question classification is a very important sub-module of question answering system,and the key lies in the feature selection.This paper proposes a new feature selection method based on syntactic and semantic information,using the question word,the main verb of the question,the dependency structure,the main noun and the top hypernym of the noun as features for classification.Evaluate the effect of feature selection using KNN and Naïve Bayes classifiers,and attain an expected result.In the predefined question taxonomy,the classification accurate reaches 82.2% and 83.7% respectively.It is better than the method using bag-of-words features.

Key words: question answering system;question classification;feature selection;dependency structure;hypernym

摘 要: 问题分类是问答系统中一个非常重要的子模块,其关键在于问题的特征选择。考虑了问题的句法信息和语义信息,提出了一种利用问题疑问词、依存关系、主要动词、中心名词和名词的最高上位词作为特征进行分类的新方法。实验中,采用 k-最邻近和朴素贝叶斯两种分类算法对该方法进行测试,结果表明了该方法具有较好的分类效果。在自定义的分类体系上,分别达到了82.2%和 83.7%的分类精度,性能高于基于 bag-of-words 的特征选择方法。

关键词: 问答系统;问题分类;特征选择;依存关系;上位词
DOI: 10.3778/j.issn.1002-8331.2008.33.045 **文章编号:** 1002-8331(2008)33-0144-04 **文献标识码:** A **中图分类号:** TP391

1 引言

随着 Internet 的不断发展与普及,各种网络服务应运而生。问答系统就是其中之一,它不同于传统的搜索引擎,主要利用自然语言的形式进行提问和回答,返回的结果更加精确,也更加符合用户需求,所以越来越受到关注。尤其是当 TREC[1]上面出现问答任务以后,问答系统逐渐成为了一个非常热门的研究方向。

问答系统一般包括三个主要部分:问题分析、信息检索和答案抽取[2]。问题分析的主要任务是对问题进行一系列预处理以及关键词提取,查询扩展等工作,其中一个重要的子模块就是问题分类。问题分类是指在事先定义好的问题类型中找到一个与该问题最相似的一个类别。可以表示为一种映射函数[3]:

$g:X\rightarrow\{c_1,c_2,\cdots,c_n\}$

其中 X 代表问题实例集合,$\{c_1,c_2,\cdots,c_n\}$是指问题类别集合,g 负责对输入的任何问题 $x\in X$,利用先验知识将其映射到类别集合中的某一个类别 c_i 中去。例如,对于问题“Where is the Kalahari desert ?”,经过问题分类,判断它是一个属于地点的问题,在后面的信息检索和答案提取模块中,缩小了搜索范围,同时也增大了抽取答案的精度。事实上,几乎所有的问答系统都包含问题分类模块,而问题分类结果的好坏在一定程度上也影响了整个问答系统的性能[4]。

针对问题分类,国内外有很多机构和大学都参与了研究,而大多数系统都采用了机器学习的方法,研究的重点主要集中在对问题的特征选择方法和分类模型上。比较著名的一个是 UIUC 大学提出的层次分类器[3,5],它主要选择词汇(word)、词性(POS)、语块(chunk)、命名实体(NE)、中心语块(head chunk)和相关词(related word)作为问题特征,基于 SNoW(Sparse Network of Winnow)结构进行分类,在大类上达到了 92.5%的分类精度。文献[6]利用 tree kernel 作为分类特征,而文献[7]选择 bag-of-word,依存关系,名词语义以及上位词作为分类特征,实验中它们都采用了支撑向量机(SVM)[8]分类模型,分类精度分别达到了 90.0%和 91.6%。

综合考虑以上系统,提出了一种新的问题特征选择方法,在句法信息上,提取了问题的主要动词以及与疑问词相关的依

基金项目: 微软亚洲研究院互联网服务科研基金(Microsoft Research Asia Internet Services in Academic Research Fund No.FY07-RES-OPP-116)。
作者简介: 袁晓洁(1963-),女,教授,博士生导师,主要研究方向为数据库管理、软件工程和信息检索;师建兴(1984-),男,硕士研究生,主要研究方向为数据库技术、问答系统;宁华(1984-),女,硕士研究生,主要研究方向为数据库技术、数据集成;于士涛(1981-),男,博士研究生,主要研究方向为机器学习、信息提取和 Web 信息检索。
收稿日期: 2007-12-17 **修回日期:** 2008-03-07

存关系作为特征,有效地减少了噪声,并以 XML 片段的形式来表示依存关系,能够无损的保存信息,同时有利于后期的计算。在语义信息上,又采用了问题的中心名词和它的最高上位词作为特征,提高了语义信息的区分度和精度。实验中分别基于 k-最邻近(k-Nearest Neighbor,kNN)和朴素贝叶斯(Naïve Bayes,NB)分类器对问题进行测试,取得了较为满意的分类效果。

2　分类体系

在介绍分类体系之前,先引出了一个概念——答案类型(Answer Type,AT),给出定义。

定义 1 答案类型(Answer Type,AT)。是指问题答案的类型,分为简短答案(short answer)类型和冗长答案(long answer)类型两种。属于 short answer 的问题,它们可以用一个词或者一个短语作为答案进行回答,类似于 TREC 上面提出 factoid 类型[9],例如询问时间、地点的问题。而属于 long answer 的问题,就是常说的描述性问题(descriptive question)[10],它们往往不能简单地用几个词或者短语作为答案,像一些以 Why、How 提问的问题,通常需要给出带有定义性或者描述性的答案进行回答。

从定义中可以看出,答案类型实际上对问答系统起到了一种指导性作用,它决定了将来抽取答案的目标形式:词或短语形式还是句子或段落形式。

文献[3]给出了一个层次分类体系,包含 6 个大类(coarse classes),50 个小类(fine classes),每个大类包含一些不重复的小类,而这些类别主要都是针对 TREC 上面的问题。在它的基础上,进行变换,提出了一种基于 Web 的问题分类体系,如表 1 所示。根据问题的语义信息,归纳出了 10 种问题类别,按照前面答案类型的定义划分成两个部分。可以看出,它是一个更加通用的分类体系,并不只是局限于 TREC 上面的问题类型。这样,在这个分类体系上定义的问题分类,就是将实例问题映射到这 10 个类别中去。为了简化后面的实验,假设某个问题只能属于其中的一个类别(最相似的类别),而不考虑问题的二义性问题。

表 1　问题分类体系

Answer Type	Question Classes
Short Answer	ENTITY
	HUMAN
	LOCATION
	NUMERIC
	TIME
Long Answer	DEFINITION
	DESCRIPTION
	MANNER
	REASON
	YNQ(yes-no)

3　问题特征选择

使用机器学习的方法对问题分类之前,首先需要对问题进行特征选择,决定哪些特征应该采用,哪些特征应该忽略[11]。然后表示成向量的形式,而向量的每一维就是问题的一个特征。一个典型的方法就是利用词袋(bag-of-words)模型提取特征,即选取问句中的所有词汇(term)作为特征项。然而,不同的 term(如疑问词)对分类结果的贡献应该是不相同的,所以它不能精确体现问题的特征。这里借鉴多种问题特征选择方法[5,7,12],同时考虑了问题的句法信息和语义信息,选择了问题的 5 种不同特征:

(1)疑问词(the question word)

(2)主要动词(the main verb of the question)

(3)依存关系(the dependency structure)

(4)中心名词(the first noun following the question word)

(5)名词的最高上位词(the top hypernym of the noun according to WordNet hierarchy)

其中,疑问词可以通过查询疑问词表获得,疑问词表是事先手工建立的,它包含所有像 what、why 等英语里面的所有疑问词。依存关系和主要动词是从问题的句法信息中获得,而中心名词和该词的最高上位词则是从语义信息中获得。下面分别介绍一下这些特征。

3.1　主要动词和依存关系

对问题进行句法分析,有很多现成的句法分析工具,其中一个比较著名的就是 Minipar[13]。它对输入的句子进行句法结构解析,核心是得到该句子的依存关系树(也叫句法树),当调用不同的接口时会返回不同形式的结果。例如对于问题"Where can I buy some furniture in Beijing?",Minipar 得到的句法树结果如图 1 所示。

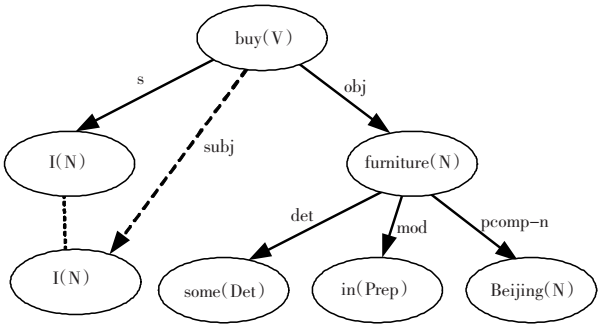

图 1　句法依存关系树

可以看出,一个问句中如果含有实义动词(有词义,不考虑助动词、系动词和情态动词,类似于陈述句里面的谓语动词),经过句法分析后得到的句法树中会出现是一棵以该动词为根节点的子树,如图 1 中以 buy 为根的子树。而且,一个问句通常只含有一个主要动词,所以,该问句的主要动词可以通过在它的句法树中搜索获得。

依存关系是句子中词项之间二元关系,一个记为核心词(head),另一个则记为依存词(dependent),它反映的是核心词和依存词之间语义上的依赖关系[4]。上面的句法树利用 Minipar 还可以表示为一系列的三元组,每一元组结构如下:

head pos:relation:pos dependent

其中,head、dependent 分别是核心词和依存词的词根(stem)形式,pos 是它们的词性,中间的 relation 则是核心词和依存词之间的依赖关系。最终结果如表 2 所示。

Minipar 给出的结果包括了句子中所有可能的依存关系。这其中有很多对分类结果影响不是很大,反而会带来消极作用,产生噪声,不适合作为问题特征。所以只选择其中包含疑问词的依存关系(核心词或者依存词是疑问词的关系对),作为最终的依存关系特征向量。

3.2　依存关系表示方案

在依存关系的表示方案上,考虑到 XML 技术的独特优势,

表 2 Minipar 句法分析结果

head	pos:relation:pos	Dependent
~	Q:whn:N	Where
~	Q:head:YNQ	~
~	YNQ:inv-aux:Aux	Can
~	YNQ:head:V	Buy
buy	V:s:N	I
buy	V:subj:N	I
buy	V:obj:N	Furniture
furniture	N:det:Det	Some
furniture	N:mod:Prep	In
in	Prep:pcomp-n:N	Beijing

采用 XML 片段来表示依存关系。因为它不仅能够精确无损失地体现这种关系,同时也方便后期的存储和计算,给出定义。

定义 2 依存关系 XML 模型(XML Schema)。指依存关系 XML 片段的 Schema 描述。每一个 Minipar 三元组的根节点是元素 triple,然后分别用元素 head、dependent、relationship 来表示核心词、依存词和依赖关系。

对于问题"Where can I buy some furniture in Beijing?",它的疑问词是 where,从 Minipar 结果中可以看出没有包含 where 的元组,所以它的依存关系特征被设为 null。

类似地,对于问题"What do rabbits eat?",根据定义 2,经过分析提取,得到的结果为:

```
<triples>
  <triple>
<head stem="eat" />
<dependent stem="what" />
<relationship name="obj" />
  <triple>
</triples>
```

3.3 中心名词和最高上位词

上面介绍的主要动词和依存关系,都是利用了问题的句法结构信息。在分析问题的语义信息时,我们考虑了问题的中心名词和该名词的最高上位词。中心名词这里是指,问句中出现在疑问词后的第一个名词,同样也是它的词根形式。

要获取问句的语义信息,可以借助工具 WordNet[11],它是一个针对英文的词库数据库(字典),包含了四种词性的词以及它们的语义等信息,分别是名词、动词、形容词和副词。在 WordNet 里面还定义了许多语义关系,像同义关系、反义关系、上下位关系、整体部分关系等等,这里只用到了其中的名词上位关系。上位关系可以看作是一种"is a"或者"is a kind of"的关系,它表现了名词之间的一种层级结构。例如"summer→season→abstraction"就是这样一种关系。可以看出,由于上位词的特点,名词在每一级都会得到一个在语义上更具有概括性的上位词。如果将这些上位词都作为问题的特征,显然会带入许多不必要的无用信息,一种好的解决办法就是只选择名词的最高上位词。

WordNet 里面的名词被组织成 11 种大的类别(如 entity,abstraction,psycho feature,activity…),这 11 个大类又可以细分成 25 在语义层次上较浅的基本类别(如{act,activity} {food} {possession} {animal,fauna} {group,grouping} {process} {artifact} …),显然这些基本类别相比于 11 个大类更加精确,对名词的区分度也更好一些。所以,选取这 25 个类别作为名词的最高上位词。这样,在得到问题的中心名词以后,就可以通过递归调用 WordNet 接口来搜索该名词的上位词,直到返回这 25 个类别中某种为止。例如问题"Where can I buy some furniture in Beijing?",它的中心名词是 furniture,得到的最高上位词则是 artifact。由于某些名词存在着多种语义,所以有时会出现一个名词对应多个上位词的情况。把这些上位词都作为分类特征的一项。

3.4 特征提取流程

通过上面的方法,就实现了问题的特征选择。整个流程可以概括为图 2 所示,对于实例问题,先要进行一系列的预处理工作,包括拼写自动更正、去停用词等等。然后进行特征提取,首先通过查询疑问词表提取出问题的疑问词,接着利用 Minipar 进行句法分析,获得依存关系和主要动词。之后,提取出问题的中心名词,从 WordNet 中找出它的最高上位词。在得到了问题这些特征以后,就可以将其表示成向量的形式,最后提交给分类器进行处理。分类器先要使用训练数据完成训练过程,然后对测试数据进行分类。

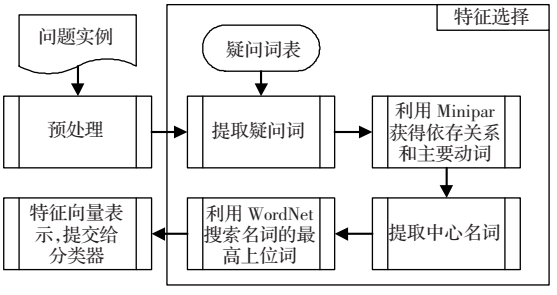

图 2 系统流程图

4 实验结果及分析

4.1 实验数据

实验数据部分使用了 UIUC 大学的语料,这里表示感谢,里面包括 5 500 个训练集和 500 个测试集。同时,又添加了部分从问答论坛上获得的语料,最终的实验数据包括 5 300 个训练语料和 600 个测试语料。训练语料被随机的分成五组,问题个数分别是 1 000,2 000,3 000,4 000 和 5 300 个。测试语料的分布情况如表 3 所示。

表 3 测试语料分布情况

问题类别	#
ENTITY	113
HUMAN	78
LOCATION	86
NUMERIC	72
TIME	54
DEFINITION	109
DESCRIPTION	32
MANNER	19
REASON	26
YNQ(yes-no)	11

4.2 评价指标

根据前面定义的分类体系,采用测试语料在 10 个问题类别上的分类精度(Accuracy),或者叫做分类准确率,对系统进行评价,分类精度定义如下:

$$Accurary=\frac{\#of\ correct\ predictions}{\#of\ predictions} \quad (1)$$

4.3 分类器

对于分类器的选择,kNN 和 Naïve Bayes 是两种传统的分类算法,它们在文本分类中表现了很好的性能。实验中分别采用这两种分类器进行测试。

kNN 分类算法是一种基于实例的模式识别方法。就是在训练集中寻找与待分类文档最相似的 K 个实例（最近的邻居），然后利用这 K 个近邻文档依次计算每类的权重，最后将输入文档分到权重最大的类别中去。其中,文档的相似度可以通过向量的余弦来计算。

Naïve Bayes 分类器是基于一个简单的假设：在给定目标值时属性之间相互条件独立。然后,利用贝叶斯公式,来找出使得目标文档概率值最大的一个类别,公式如下所示：

$$v_{NB}=\underset{c_j\in C}{\operatorname{argmax}}\ P(a_1|c_j)*\cdots*P(a_n|c_j)*P(c_j) \quad (2)$$

其中,v_{NB} 表示输出的目标值;c_j 指类别集合 C 中的某一类;a_1,a_2,…,a_n 分别表示文档 d 的 n 维特征向量;而 $P(a_i|c_j)$表示在类别 c_j 中特征项 a_i 的出现概率。

4.4 实验结果及分析

为了对比实验效果,设计了两种特征选择方法,一种是基于 bag-of-words 的特征选择方法，另一种则是采用提出的方法。图 3 和图 4 分别显示采用这两种不同的特征选择方法,kNN 和 Naïve Bayes 在测试语料上的分类精度曲线。

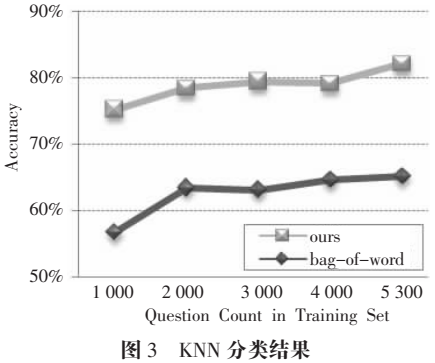

图 3　KNN 分类结果

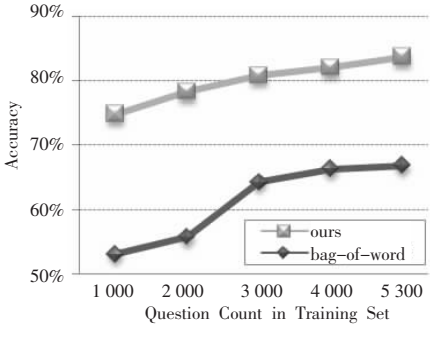

图 4　NB 分类结果

从图中可以看出,使用提出的问题特征选择方法进行问题分类,结果较好一些。说明这种方法更能精确地体现问题特征,具有更好的分类效果。另一方面,在分类器上,NB 分类器整体比 KNN 表现出了更好的性能，这可能是由于它们自身的特点以及测试环境决定的。

5 结论

提出了一种新的问题特征选择方法。在句法信息上,采用了问题主要动词以及和疑问词相关的依存关系作为特征,有效地减少了噪声,并以 XML 片段的形式表示,能够无损地保存信息,同时有利于后期的计算。在语义信息上,又提取了问题的主要名词和它的最高上位词作为特征向量,提高了语义信息的区分度和精度。实验中分别采用 KNN 和 Naïve Bayes 分类算法进行测试，结果表明该方法在问题分类上表现出了较好的效果,分类精度高于基于 bag-of-words 的特征选择方法。下一步研究的重点,要考虑如何更准确地选取问题特征信息,还应该分析比较更多的分类器,尝试对分类算法进行改进,使它们更能满足问题分类的需要。

参考文献:

[1] Voorhees E.Overview of the TREC 2003 question answering track[C]//Proceedings of the 12th Text REtrieval Conference (TREC 2003),2004:54-68.

[2] 郑实福,刘挺.自动问答综述[J].中文信息学报,2002,16(6):46-52.

[3] Li X,Roth D.Learning question classifiers[C]//proceeding of the 19th International Conference on Computational Linguistics (COLING'02),Taipei,2002:556-562.

[4] 文勖,张宇.基于句法结构分析的中文问题分类[J].中文信息学报.2006,20(2):33-39.

[5] Li Xin,Roth D,Small K.The role of semantic information in learning question classifiers[C]//Proceedings of the 1st International Joint Conference on Natural Language Processing.Cambridge University Press,2006,12(3):229-249.

[6] Zhang D,Lee Wee Sun.Question classification using support vector machines[C]//the 26th ACM SIGIR,2003.

[7] 李鑫,杜永萍.基于句法信息和语义信息的问题分类[C]//第一届全国信息检索与内容安全学术会议,2004:243-251.

[8] 田立,刘振丙.一种改进的线性 SVM[J].计算机工程与应用,2007,43(20):173-176.

[9] Soricut R,Brill E.Automatic question answering using the web:Beyond the Factoid[J].Information Retrieval,2006,9(2):191-206.

[10] Oh Hyo-Jung,Lee Chung-Hee,Kim Hyeon-Jin,et al.Descriptive question answering in encyclopedia [C]//Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions,Association for Computational Linguistics,2005:21-24.

[11] WordNet[EB/OL].http://wordnet.princeton.edu/.

[12] Jijkoun V,van Rantwijk J,Ahn D,et al.The University of Amsterdam at CLEF@QA 2006.In Working Notes CLEF,2006.

[13] Minipar[EB/OL].http://www.cs.ualberta.ca/~lindek/minipar.htm.