高技术通讯 2009 年 第 19 卷 第 2 期:111 ~ 118

doi: 10.3772/j.issn.1002-0470.2009.02.001

基于线索词识别和训练集扩展的中文问题分类①

张志昌② 张 宇 刘 挺 李 生

(哈尔滨工业大学计算机学院信息检索研究室 哈尔滨 150001)

摘 要 针对问题分类的数据稀疏问题,提出了一种以疑问词和焦点词为关键线索的中文事实型问题分类方法。该方法首先自动识别用户提出的问题中的疑问词和焦点词,若疑问词和焦点词存在,则用最近邻模型进行分类,而对没有用最近邻方法分类的其他问题,则用支持向量机(SVM)模型进行分类。训练 SVM 模型时,从 Web 上自动获取新问题来对训练集进行扩展,最近邻方法只利用线索词词义距离进行类别判断。实验表明,这种按照问题结构的不同而选择不同分类器的方法,在性能上要优于单一分类方法;词义距离的应用和训练集自动扩展改善了训练数据的稀疏,提高了分类性能。

关键词 问题分类, 焦点词, 词义距离, 训练集扩展

0 引 言

问答技术在信息检索领域中已成为重要的研究方向,其目标是根据用户用自然语言提出的问题,从大规模文档集合中提取出所提问题的正确答案。现有的问答系统一般由问题分析、包含答案的文档和文档片段检索、答案抽取三部分组成[1],其中问题分析的任务主要是确定问题的预期答案语义类别,如问题"哪国人口最多?"是问国家名;问题"中国人口有多少?"是问一个数量。确定问题的预期答案语义类别,可看作是按照某个语义类别对问题进行分类,目的是为答案的抽取提供约束,使抽取模块只判断属于特定语义类别的答案候选。根据文献[2]的分析,问答系统中 36.4% 的答案错误是由问题分类的错误产生的。

目前问题分类方法有待进一步研究。基于规则的方法[3,4]需要编写大量的规则,人工的负担太重,因此不属于研究的重点。基于统计的机器学习方法,一般都是应用 SNoW[5,6]、支持向量机(SVM)[7]、Naïve Bayes[8]、最大熵[9]等算法,在已标注的训练问题集上学习分类模型,但由于用人工方式去收集问题、标注类别,既耗时又耗力,因此获取足够规模的训练问题并不容易。

问题分类可以看作是特殊的文本分类。相对于文本,问题中确定类别的特征很少,分类效果对单项特征的缺失非常敏感,因此训练数据的稀疏问题对分类效果的影响更为明显;同时,问题中其它对分类没有作用的噪音词,会给分类器带来很大干扰。因此,提高问题分类效果的一种途径是选择真正对分类有用的特征,如文献[6]将组块(chunk)、词性、词在 WordNet 中的词义、命名实体、问题类别的语义相关词作为特征,其中,语义相关词以手工方式获得。另外一种途径是在某些特征维度上补充训练数据,以克服训练数据的稀疏,如文献[10]从 Web 上获取问题类别相关的类别特征词,获取时,用训练问题的答案作为关键词进行检索,获得包含该答案的句子,然后从这些句子中选择其它动词、名词、形容词作为问题所属类别的相关特征词。但在 Web 检索结果中,和答案共同出现在一个句子中的词汇非常混乱、分散,简单地选择具有特定词性的词并不完善,因此该方法的整体性能不高。

针对这种情况,本文提出一种结合最近邻分类和 SVM 分类的中文事实型问题分类方法,对不同结构的问题选择不同的分类器。这种结合方法对一个测试问题只用一个分类器进行分类,相对于多分类器的并行竞争或投票组合等方法,能在保证分类性能的前提下,使问题分类在实际应用中有较好的时间效率。

① 863 计划(2006AA01Z145)和国家自然科学基金(60435020,60503072)资助项目。
② 男,1976 年生,博士生;研究方向:问答技术,文本检索;联系人,E-mail: pangzhang@gmail.com
(收稿日期:2007-12-07)

1 方法概述

本文只解决事实型问题的分类,这类问题的提问是基于事实的,且答案比较简短,一般是一个命名实体或者短语[11],如“美国总统是谁?”,“珠穆朗玛峰有多高?”等。事实型问题的提问可分为两种形式:

(1) 问题对预期答案的类型描述明显。如“科技竞争力第一名的国家是哪个?”,通过名词“国家”和疑问词“哪个”来表述该问题的答案类型是国家名称。

(2) 问题对预期答案的类型描述不明显。如“我国进入世界 500 强的有哪些?”,“高山贪污了多少?”,其中没有明显表达问题所预期答案类型的词汇,而疑问词所涵盖的答案类型也不够具体。

把问题中对表达答案类型具有重要意向信息的名词称为问题焦点词(focus word)[12]。对第(1)种形式的问题,如果能够确定问题的疑问词和焦点词,且焦点词没有歧义,则可确定问题的类别。对第(2)种形式的问题,尽管省略了焦点词,但存在和焦点词语义相关的其它词,这些词和疑问词一起确定了问题的类别。对第(1)种形式的问题,如果能在训练集中找出和测试问题在疑问词和焦点词上最相似的问题,则可将该训练问题的类别作为分类结果。而最近邻分类方法能很直接地体现这种分类思路,不需要提前训练,而且只有两个特征的相似度计算,实现上更容易,因此我们选择用最近邻方法对这种形式的问题进行分类。

对没有被最近邻方法分类的剩余问题,我们利用 SVM 模型进行分类,这些问题中有一些存在焦点词,但没有被系统标识出来,还有些是不存在明显的焦点词,因此分类时需要利用更多的其它特征。对这些问题选择 SVM 模型分类的原因是,在文本分类中,相对其它模型,SVM 一般表现出最好的性能[13]。

对这两种形式的问题进行分类时,需要解决的一个关键问题是训练数据稀疏问题。如对第(1)种形式的问题“科技竞争力第一名的国家是哪个?”,如果训练集中不存在疑问词和焦点词分别为“哪个”和“国家”的训练实例,则分类器很难对该问题进行正确分类。如对第(2)种形式的问题“我国进入世界 500 强的有哪些?”,如果训练集中没有关于“世界 500 强”的训练实例,则该问题的类别很难区分。

因此,对于第(1)种问题,我们在判断和待分类问题在疑问词和焦点词上最相似的训练问题时,不是用词汇本身,而是用词在《同义词词林(扩展版)》(http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm)中的词义距离,以解决测试问题中的焦点词在训练集中不存在的情况。对于第(2)种问题以及其它没有被最近邻方法分类的问题,从 Web 上获取大量新问题来补充人工标注的训练集合来训练得到 SVM 模型进行分类。获取时,需要利用人工标注训练集中各个类别的疑问词和焦点词。下面的算法描述了对测试问题进行分类的过程。

算法 1 以疑问词和焦点词为线索的结构化问题分类。

步骤 1 利用训练问题的疑问词和焦点词,从 Web 上获取新的训练问题,得到扩展训练集。

步骤 2 在人工标注训练集和扩展训练集上训练 SVM 模型。

步骤 3 对测试集中的每个问题,按如下处理:

(1) 自动识别问题中的疑问词和焦点词;

(2) 若疑问词和焦点词存在,则用基于 < 疑问词,焦点词 > 的词义最近邻方法分类;否则,用 SVM 方法分类。

其中,最近邻方法只用到了疑问词和焦点词特征,去除了问题中的其它干扰信息,分类特征不是直接使用词,而是使用词义距离;而在训练 SVM 模型时,针对人工构造更大规模训练数据的困难,利用人工标注训练集中问题的疑问词和焦点词及其它们的词义,从 Web 上获取大量的问句作为新的训练问题来扩展训练集规模。这样,线索词词义距离的使用和训练集的自动扩展有效克服了训练数据的稀疏问题。

2 基于线索词词义距离的最近邻分类

2.1 线索词的自动识别

把问题中的疑问词和焦点词统称为线索词,通过训练条件随机域(CRF)模型来对线索词进行自动标注。为了训练模型和确定模型所需的特征,将训练问题集中的全部问题都进行了人工的疑问词和焦点词标注,并分割为训练集和测试集两部分,通过实验确定选择上下文为 5 的滑动窗口,并选用词、词性、修饰词及其词性、依存关系特征。另外实验发现,对这些特征的组合能小幅度地提高模型的标注性能。具体的特征选择如下:

(1) 词 N-gram 特征、词性 N-gram 特征:捕获当前位置的上下文信息,包括 Unigram 和 Bigram 特征。

(2) 句法修饰词的 Unigram 特征、修饰词词性的

N-gram 特征、与修饰词之间的依存关系 N-gram 特征:捕获当前位置的依存结构信息。

(3) 组合特征。CRF 模型可利用丰富的彼此重叠的特征增强系统的描述能力,因此对上述特征进行了组合,得到组合特征,以进一步提高疑问词和焦点词的识别性能,包括当前词/词性组合、词性/修饰词词性/依存关系组合、修饰词词性/依存关系组合。

2.2 最近邻方法分类

如果问题中存在明显的疑问词和焦点词,则可确定该问题的答案语义类别。这样,根据一个训练集 T,对一个问题 q 进行分类时,如果能在 T 中找到一个问题 d,使得 d 和 q 在疑问词和焦点词上相同,则 d 的类别就是 q 的类别。

但是,由于训练数据规模有限,如果按照词汇匹配,可能在训练集合 T 中找不到一个问题 d,使得 d 和 q 在疑问词和焦点词的词汇形式上相同。因此,我们利用基于词义距离的 < 疑问词, 焦点词 > 二元组相似度计算,获取和测试问题最相似的训练问题。由于只用两个分类特征,并且在分类时需计算词义之间的相似度,因此选择最近邻法作为分类这类问题的方法,通过计算测试问题和所有训练问题在疑问词和焦点词上的词义相似度,找出与测试问题有最大相似度值的训练问题,将它的类别作为分类类别。

本文采用《同义词词林(扩展版)》作为知识资源来计算词义相似度,该资源按照词义从粗到细的义元划分,设计了一个 5 层的词义代码义元树,词义形式为“$L_1L_2L_3L_4L_5L_6$”,其中,“L_i”($i=1,\cdots,5$)代表该词在第 i 层上的词义,“L_6”是对该词义的附加标识。因此,如果两个词的词义在第 j 层($1\leqslant j\leqslant 6$)上属于相同的同义类别,则这两个词义在对应不同层的 L_i ($1\leqslant i\leqslant j$)上相同。如“教授”的词义“Ae13A10 #”,“教师”的词义“Ae13A01 =”,“研究生”的词义“Ae13B13 #”。其中“教授”和“教师”在前四层上相同,而“教授”和“研究生”在前三层上相同。如图 1。

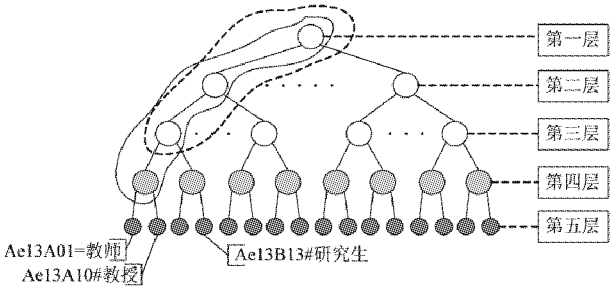

图 1 《同义词词林(扩展版)》层次结构

把两个词的词义相似度定义为两个词义在义元树中从根到叶子的公共路径数,在词义代码串上的表现就是最长的公共前缀子串长度。而一个词在某个集合中的最近邻就是词义相似度最大的词,即在词义代码上有最长公共前缀子串的词。由于词在第一层和第二层上可能存在歧义,我们在寻找最近邻时,要求最近邻的词义至少要在第三层上相同,根据词义代码的设计形式,即要求两个词的词义代码公共前缀子串长度必须要大于或者等于 4。下面的算法描述了基于线索词词义距离的最近邻分类方法。

算法 2 基于线索词词义的最近邻分类。

步骤 1 对测试问题 q 和训练问题 $d_i\in$ 训练集 T ($i=1,\cdots,n$),用哈工大信息检索研究室的词义消歧模块进行词义消歧,然后获取 q 和 d_i 在《同义词词林(扩展版)》中的词义代码。

步骤 2 利用训练得到的 CRF 模型,对测试问题 q 进行疑问词和焦点词的自动识别。若疑问词和焦点词不存在,则返回空;否则,得到疑问词和焦点词词义二元组 $\langle q\_qSense, \_fSense\rangle$。

步骤 3 在训练集合 T 中寻找问题 d_j,满足:

$$d_j = \arg\max_{1\leqslant i\leqslant n} Sim(q,d_i) \text{ 且 } Sim(q,d_j)\geqslant 4 \quad (1)$$

$$\begin{aligned} &Sim(q,d_i)\\ &= Sim(\langle q\_qSense, q\_fSense\rangle, \langle qSense_i, fSense_i\rangle)\\ &= \begin{cases} 0, \text{ if } q\_qSense \neq qSense_i \text{ or } \langle qSense_i, fSense_i\rangle \\ \qquad\qquad \text{not exist in } d_i \\ Sim(q\_fSense, fSense_i), \text{ otherwise} \end{cases}\\ &= \begin{cases} 0, \text{ if } q\_qSense \neq qSensei \text{ or } \langle qSense_i, fSense_i\rangle \\ \qquad\qquad \text{not exist in } d_i \\ \text{CommonLevelPrefix}(q\_fSense, fSense_i), \text{ otherwise} \end{cases} \end{aligned} \quad (2)$$

步骤 4 若满足条件的 d_j 存在,则把 d_j 的类别作为 q 的类别返回;否则,返回空。

上述公式中,q_qSense、q_fSense 分别是测试问题 q 的疑问词词义代码、焦点词词义代码,$qSense_i$、$fSense_i$ 分别是训练问题 d_i 的疑问词词义代码和焦点词词义代码。CommonLevelPrefix(q_fSense, $fSense_i$)是计算词义代码 q_fSense 和 $fSense_i$ 在词义义元树中从根到叶子节点的公共路径。

3 基于训练集自动扩展的 SVM 分类

3.1 训练集的自动扩展

利用各种统计机器学习方法进行分类时,训练

数据的规模是影响分类模型的重要因素。由于人工方法收集问题、标注问题类别在效率上的局限性,需要研究自动收集训练数据的方法。Web 上存在各种类型的海量数据,其中包含有各种形式的问题,但这些问题一般都没有类别标注。

而另一方面,事实型问题中无歧义的 < 疑问词,焦点词 > 可以确定问题的语义类型,基于这个判断,本文用基于线索词词义距离的最近邻方法进行问题的分类。进一步地,用人工标注训练集中无歧义的疑问词和焦点词作为关键词,从 Web 上检索包含这两个词的问句,将满足条件的问句作为新的训练问题,其类别为关键词所在原始训练问题的类别。下面的算法描述了训练集自动扩展的方法。

算法 3 利用 Web 进行训练集自动扩展。

步骤 1 对人工标注类别为 $Class_i$ 的训练问题 $d_i \in$ 训练集 $T(i = 1,\cdots,n)$,若该问题存在疑问词 $qword_i$ 和焦点词 $focus_i$,则对问题 d_i 进行词义消歧,获取 $qword_i$ 和 $focus_i$ 在《同义词词林(扩展版)》中的词义 $qwordSense_i$ 和 $focusSense_i$。

步骤 2 根据焦点词 $focus_i$ 的词义 $focusSense_i$,在《同义词词林(扩展版)》中获取焦点词 $focus_i$ 的所有同义词,得到同义词集 $FocusSet$,利用 $focus\_ext_j \in FocusSet(j = 1,\cdots,l)$ 和 $qword_i$ 作为查询关键词,用搜索引擎进行检索,从检索结果的摘要(Snippets)中获得 m 个包含该关键词的扩展问句集合 S。

步骤 3 对每一个扩展问题 $web\_q_j \in S(j = 1,\cdots, m)$ 进行分词、词义消歧,并用 3.1 节所述的方法进行疑问词和焦点词标注。若该问题满足如下条件,则其类别为 $Class_i$,且为有效的扩展训练问题,如果在现有扩展训练问题集合 $Web\_TrainingSet$ 中不存在该扩展问题 web_q_j,则将其加入 $Web\_TrainingSet$。条件(1)在标注结果中,疑问词为 $qword_i$,焦点词属于 $FocusSet$。条件(2)词义消歧后的结果中,$qword_i$ 的词义为 $qwordSense_i$,焦点词的词义为 $focusSense_i$。

步骤 4 若人工标注训练集中,仍有未处理的问题,则返回步骤 1;否则,算法结束,返回结果 $Web\_TrainingSet$。

用搜索引擎进行检索时,以疑问词 $qword$ 和焦点词 $focus$ 作为关键词,在返回的检索结果中,包含 $qword$ 和 $focus$ 的问句,其疑问词和焦点词并不一定是 $qword$ 和 $focus$。如关键词为“哪个”和“公司”,检索结果中有“房地产公司哪个部门最重要?”,但该问题的焦点词并不是“公司”。因此,由算法的步骤 3 来保证获取的新问题的焦点词和疑问词和查询关键词相同。

对训练集扩展的目的是去补充人工标注集。例如,人工标注训练集中有疑问词为“哪些”和焦点词为“企业”的训练问题,我们用“哪些”和“企业”作为查询输入,从 Web 上可以获取到如下问题:“中国有哪些企业进入了世界 500 强?”,“请问,在福州开发区有哪些世界 500 强投资的企业?”等,这些问题可以加入到训练集中。这样,当有测试问题“我国属于世界 500 强的有哪些?”时,由于扩展后的训练集中已经包含了“世界 500 强”等特征,分类器对该问题的分类要更容易。

3.2 特征选择

在训练 SVM 模型时,首先对问题进行分词、词性标注、词义消歧。如对问题“澳大利亚最大的城市是什么”,分词和词性标注结果为:“澳大利亚/ns 最/d 大/a 的/ue 城市/n 是/vx 什么/rs”;词义消歧的结果为:“澳大利亚/Di02 最/Ka02 大/Ea03 的/Kd01 城市/Cb25 是/Ja01 什么/Ba10”,其中的词义是《同义词词林(扩展版)》词义义元树中的第三层代码。我们选择问题中的所有不属于非停用词的词、词性、词义,问题的疑问词、焦点词作为分类特征,并通过实验,比较这些特征对分类效果的影响。

4 实验结果与分析

4.1 实验数据与评价方法

本文采用和文献[9,10]相同的训练和测试问题集,该问题集的分类体系中包含 7 个粗类,细分为 60 个细类。为了使分类结果能更好地帮助答案抽取,我们对粗类进行了更细的划分,使细类数达到 84 个,其中,涉及事实性问题的粗类别有 6 个,包含的细类别有 75 个。最终的事实型问题类别体系、训练集和测试集中的问题分布见表 1,其中训练集问题总数为 4123,测试集问题总数为 1157。

参考文献[7]的评价标准,本文用粗类上的准确率($Accuracy\_C$)和粗类中全部细类整体上的准确率($Accuracy\_F$)来评价算法的分类性能。

$$Accuracy\_C = \frac{\sum_F Test_F \text{ 中被分类为粗类 } C \text{ 下任意细类的问题数}}{\sum_F \text{粗类 } C \text{ 下细类 } F \text{ 的测试集 } Test_F \text{ 问题数}} \times 100\% \quad (3)$$

表 1　事实型问题的类别体系及训练与测试语料的分布情况

粗　类	训练数	测试数	包含的细类
人物(HUM)	261	173	特定人物 团体机构 其它
地点(LOC)	936	390	星球 城市 大陆 国家 省 河流 湖泊 山脉 大洋 岛屿 建筑 地址 其它
数值(NUM)	1076	244	温度 面积 体积 重量 速度 频率 距离 钱数 数目 顺序 倍数 百分比 号码 时间长度 范围 其它
时间(TIME)	598	153	年 月 日 季节 年代 星期 节气 节假日 时间 时间范围 其它
实体(OBJ)	1242	194	物质 动物 植物 微生物 身体部位 材料 用品 衣物 食物 货币 票据 语言 事件 疾病 艺术品 服务 文字作品 学科 计划 法律 头衔 职业 符号 奖励 罪刑 民族 权利义务 颜色 宗教 运动 术语 其它
未知(UN)	10	3	未知

$$Accuracy\_F = \frac{\sum_F Test_F \text{ 被正确分类为细类 } F \text{ 的问题数}}{\sum_F \text{粗类 } C \text{ 下细类 } F \text{ 的测试集 } Test_F \text{ 问题数}} \times 100\% \quad (4)$$

4.2　词义最近邻方法实验结果

本文使用开源的 CRF ++ 工具(http://crfpp.sourceforge.net),将 4213 个训练问题集在每个细类别上随机选择 10%作为测试集,90%作为训练集,在这个训练集上训练得到 CRF 标注模型。测试结果中,焦点词识别的准确率为 94.58,疑问词识别的准确率为 99.39%,较高的识别性能为后续使用疑问词和焦点词作为分类选择条件提供了基础。根据算法 2,在全部测试集上用疑问词和焦点词词义最近邻方法进行分类的性能见表 5。在全部 1157 个测试问题中,识别出疑问词和焦点词、并应用该方法进行了分类的问题数目是 945 个,占全部测试问题的 81.68%。表 2 中的准确率是在最近邻方法已分类的问题中,分类正确的问题数所占的比例。

表 2　疑问词和焦点词词义最近邻方法的分类性能

类别	被分类数	被分类数占问题总数(%)	粗类分类正确数	细类分类正确数	*Accurancy _ C* (%)	*Accurancy _ F* (%)
HUM	134	77.46	131	130	97.76	97.01
LOC	342	87.69	330	311	96.49	90.94
NUM	194	79.51	194	179	100.00	92.27
OBJ	136	70.10	126	123	92.65	90.44
TIME	139	90.85	136	136	97.84	97.84
UN	0	0	0	0	0	0
Total	945	81.68	917	879	97.04	93.02

对 66 个细类别分类错误的问题进行分析,错误的原因主要是:(1)在训练集和测试问题上的词义消歧结果错误,如将“电阻的单位是什么”中的“单位”解释为“机关、机构”,这会导致在寻找词义最近邻时发生错误,所占比例为 41.82%。(2)焦点词识别错误,如将“NASA 是哪个国际组织”的焦点词识别为“国际”,所占比例为 36.06%。(3) 方法本身的局限,所占比例为 22.12%。其中,第(3)种错误主要发生在各个粗类别下的“其它(OTHER)”细类和其它细类之间。在一个粗类下,“其它(OTHER)”细类包含了其它具体细类均没有覆盖的问题,因此相对其它细类别,该类别的范围确定比较模糊,导致在该细类别上的分类精确率较低。

4.3　选择不同特征时的 SVM 分类实验结果

选择 LIBSVM[14] 工具,用人工标注的全部训练集训练 SVM 模型。训练和分类时,将问题表示成向量的形式:$(x_1, x_2, \cdots, x_n)$,其中第 i 维上的特征 $x_i \in \{0,1\}$,表示该特征是否在问题中出现。在选择词、词性、词义特征时,没有过滤停用词。在选择不同特征时的全部粗类和全部细类的整体准确率对比见表 3。

表 3　SVM 模型在选择不同特征时在全部测试集(1157 个问题)上的分类性能比较(%)

特征	*Accuracy _ C*	*Accuracy _ F*
问题中的所有非停用词	88.58	76.90
+ 问题中所有非停用词的词性	89.87	76.90
+ 问题中所有非停用词的词义	91.00	79.49
+ 问题的疑问词	91.08	79.49
+ 问题的疑问词词义	91.08	79.50
+ 问题的焦点词	92.21	81.74
+ 问题的焦点词词义	92.64	84.34

实验结果表明,在特征中依次加入问题中所有非停用词的词性、词义时,都会提高 SVM 分类效果。另外,将问题的疑问词、焦点词及其它们的词义在特征向量的特定维上明确指定,会显著提高分类性能。

4.4 训练集自动扩展后的 SVM 分类实验结果

依算法 3,从 Web 上获取新问题对人工标注的训练集进行扩展。本文对扩展到不同训练集规模时的分类效果进行了比较。训练和分类时,依据 4.3 节的实验,选择问题中的所有词、词性、词义、疑问词及其词义、焦点词及其词义作为特征。对训练集为全部人工标注训练集(表示为 Base,包含训练问题数为 4213 个)以及自动扩展到 8540 个问题、14785 个问题、36472 个问题、46975 个问题、51098 个问题时的分类准确率,在图 2 和图 3 中进行了对比,其中图 2 是对单个粗类(HUM、LOC、NUM、OBJ、TIME)和全部粗类(total)上的准确率(*Accuracy _ C*)进行比较,图 3 是对不同粗类下的细类整体准确率和全部细类整体(total)准确率(*Accuracy _ F*)进行比较。由于"未知(UN)"类的准确率均没有变化,故在图中没有显示。

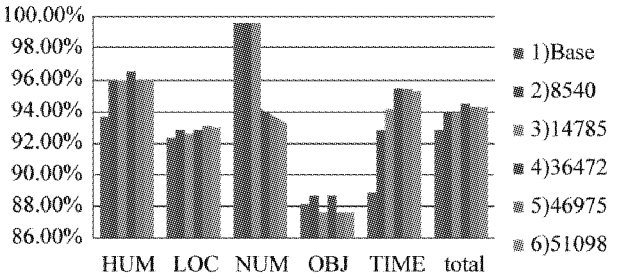

图 2 训练集扩展到不同规模时在全部测试集上的粗类准确率(*Accuracy _ C*)对比

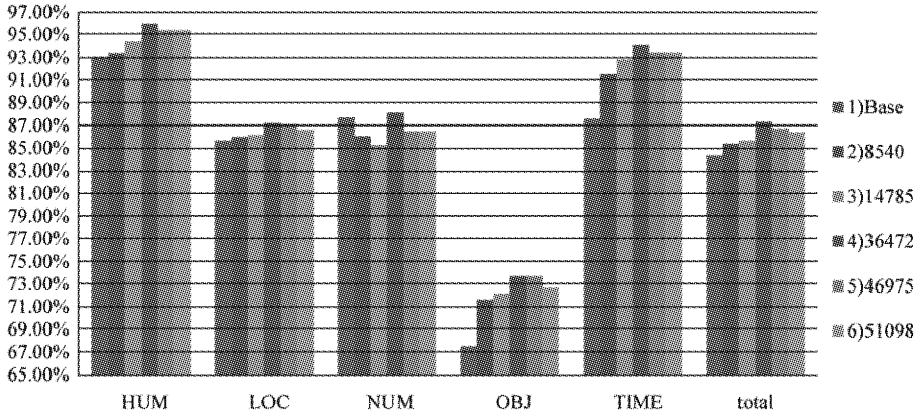

图 3 训练集扩展到不同规模时在全部测试集上的细类整体准确率(*Accuracy _ F*)对比

实验结果表明,将训练集进行自动扩展后,SVM 在全部粗类和细类测试集上的分类效果总体上均有提高。当训练集扩大到 30000 到 40000 个时,分类的性能总体达到最大值,但当扩展到更大规模时,分类的效果会有所下降,一个主要原因是从 Web 上获取的问题存在一定噪音,当问题规模持续增加时,噪音对分类器的影响会更加明显。

4.5 结合方法实验结果

依算法 1,根据问题结构的不同来选择不同分类器,即首先用词义的最近邻方法对存在线索词的问题进行分类,而对剩余问题,用 SVM 模型进行分类。显然,结合后要比单纯应用最近邻方法性能要好。对单纯用 SVM 的分类性能、最近邻和 SVM 结合后的分类性能以及相对于单纯用 SVM 的性能增/减幅,在表 4 中进行了对比。

SVM 分类器在分类中一般表现出更好的性能。但从实验中"只用 SVM(未扩展)"和"NN + SVM(未扩展)"的结果对比看出,与只用 SVM 方法比较,即使在 SVM 模型中也添加了疑问词和焦点词线索特征,按照是否存在线索词而选择不同分类器的结合方法,在粗类别和细类别上的整体分类性能均有明显提高。这也表明,在文本分类中多分类器组合一般要优于单分类器的情况,在问题分类中也适用。

从实验结果中也可以看出,将训练集扩展到 36472 个问题时,训练得到的 SVM 模型和最近邻方法结合分类性能要优于未扩展时的结合分类性能;从效果上看,对于只用一种方法时性能较低的类别,两种方法结合后的性能提升更为明显。这表明训练集的自动扩展改善了训练集的数据稀疏问题。

表 4　最近邻结合 SVM 的分类方法与只用 SVM 的分类性能比较(%)

类别	只用 SVM(未扩展)		NN + SVM(未扩展)		NN + SVM(扩展到 36472 个)	
	Accuracy _ C	*Accuracy _ F*	*Accuracy _ C*	*Accuracy _ F*	*Accuracy _ C*	*Accuracy _ F*
HUM	93.64	87.7	96.53(+2.89)	95.95(+8.25)	98.26(+4.62)	97.69(+9.99)
LOC	92.30	93.06	94.87(+2.57)	88.71(-4.35)	96.41(+4.11)	90.51(-2.55)
NUM	99.59	87.70	99.59(0.00)	88.11(+0.41)	100.00(+0.41)	89.34(+1.64)
OBJ	88.14	67.52	87.11(-1.03)	75.77(+8.25)	87.63(-0.51)	78.80(+9.28)
TIME	88.88	87.58	98.03(+9.15)	97.38(+9.80)	98.04(+9.16)	97.38(+9.8)
UN	50.00	50.00	50.00(0.00)	50.00(0.00)	50(0.00)	50(0.00)
total	92.82	84.34	95.15(+2.33)	88.58(+4.24)	96.11(+3.29)	89.97(+5.63)

5　结　论

本文针对中文事实型问题分类中的训练数据稀疏问题,根据不同问题的结构差异,提出了一种以疑问词和焦点词为关键线索的结构化方法。相对于问题分类上的其它工作,本文的主要贡献包括:

(1) 提出了以疑问词和焦点词为线索,综合最近邻和 SVM 模型的一种结构化的分类方法。其中,最近邻方法只使用疑问词和焦点词特征,而去掉了其它干扰信息,从而提高了分类性能。而在使用 SVM 模型时,不仅利用了问题中全部非停用词的词特征、词性特征、词义特征,同时也显示利用了疑问词及其词义特征、焦点词及其词义特征。

(2) 在用最近邻方法分类时,使用疑问词和焦点词在《同义词词林(扩展版)》义元树中的词义距离,而不是在疑问词和焦点词上进行直接的词匹配,因此可改善训练数据的稀疏问题。用 SVM 模型分类时,用人工标注训练集的疑问词和焦点词作为种子,从 Web 上获取大量确定类型的问句作为扩展的训练问题,进一步改善训练数据稀疏问题。

在将来的工作中,我们将继续改进 CRF 模型标注问题中疑问词和焦点词的性能;同时,在训练问题集自动扩展方面,研究对属于噪音的扩展问题的自动去除方法,以使训练集的扩展对分类性能的提高更有效;另外,多种分类器的并行组合和投票等方法在问题分类上的性能,也是需要进一步研究的问题。

参考文献

[1] Hirschman L, Gaizauskas R. Natural language question answering: the view from here. *Natural Language Engineering*, 2001,7(4): 275-300

[2] Moldovan D, Pasca M, Harabagiu S, et al. Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems*, 2003, 21(2): 133-154

[3] Hovy E, Gerber L, Hermjakob U, et al. Toward semantics-based answer pinpointing. In: Proceedings of the First International Conference on Human Language Technology Research (HLT 2001). San Diego, CA, USA: Association for Computational Linguistics, 2001. 1-7

[4] Brill E, Dumais S, Banko M. An analysis of the AskMSR question-answering system. In: Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). Philadelphia, PA, USA: Association for Computational Linguistics, 2002. 257-264

[5] Li X, Roth D. Learning question classifiers. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING'02). Taipei, Taiwan: Association for Computational Linguistics, 2002. 556-562

[6] Li X, Roth D. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 2006, 12(3): 229-249

[7] Zhang D, Lee W S. Question classication using support vector machines. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR 2003). Toronto, Canada, 2003. 26-32

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

[9] 孙景广,蔡东风,吕德新等. 基于知网的中文问题自动分类. 中文信息学报, 2007, 21(1): 90-95

[10] Lin S J, Lu W H. Learning question focus and semantically related features from Web search results for Chinese question classification. In: Proceedings of the 3rd Asia Information Retrieval Symposium (AIRS). Singapore: Springer, 2006. 284-296

[11] Voorhees E M. Overview of the TREC 2003 question answering track. In: Proceedings of the 12th Text REtrieval Conference(TREC 2003). Gaithersburg: National Institute of Standards and Technology (NIST), 2004. 54-68

[12] Diekema A, Liu X, Chen J, et al. Question answering: CNLP at the TREC-9 question answering track. In: Proceedings of the 9th Text REtrieval Conference (TREC-9). Gaithersburg, Maryland, USA: National Institute of Standards and Technology (NIST), 2000. 412-421

[13] 苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展. 软件学报, 2006, 17(9): 1848-1859

[14] Chang C C, Lin C J. LIBSVM——a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Chinese question classification based on identification of cue words and extension of training set

Zhang Zhichang, Zhang Yu, Liu Ting, Li Sheng

(Information Retrieval Laboratory of Computer Science & Technology School,
Harbin Institute of Technology, Harbin 150001)

Abstract

In view of the data sparseness problem in question classification, the paper proposes an approach for classifying Chinese factoid questions using interrogative and focus words as the key cues. The approach first identifies interrogative and focus words in the questions raised by users autometically and classifies the questions using the nearest neighbor model if the cue words exist, and then, classifies other questions using the support vector machine (SVM) model. The training set of SVM is extended automatically with the questions mined from Web when training the SVM model, while for the nearest neighbor model, only using the sense distance of the cue words for classification judgment. The experimental results show that the approach, selecting different classifiers according to question structure, outperforms the single classification model, and the problem of training data sparseness is alleviated using the word sense distance and the extension of training set, thus the classification performance is improved.

Key words: question classification, focus word, word sense distance, extension of training set