文章编号: 1003-0077(2007)01-0090-06

基于知网的中文问题自动分类

孙景广,蔡东风,吕德新,董燕举

(沈阳航空工业学院 自然语言处理研究室,辽宁 沈阳 110034)

摘 要: 问答系统应能用准确、简洁的答案回答用户用自然语言提出的问题。问题分类是问答系统所要处理的第一步,分类结果的正确率直接影响后续工作的进行。本文提出了一种使用知网作为语义资源选取分类特征,并使用最大熵模型进行分类的新方法。该方法以问题的疑问词、句法结构、疑问意向词、疑问意向词在知网中的首义原作为分类特征。实验结果表明,在知网中选取的首义原能很好的表达问题焦点词的语义信息,可作为问题分类的一个主要特征。该方法能显著地提高问题分类的精度,大类和小类的分类精度分别达到了 92.18 %和 83.86 %。

关键词: 计算机应用;中文信息处理;问答系统;问题分类;知网;最大熵模型;分类特征

中图分类号: TP391 **文献标识码**: A

HowNet Based Chinese Question Automatic Classification

SUN Jing-guang, CAI Dong-feng, LV De-xin, DONG Yan-ju

(Natural Language Processing Laboratory, Shengyang Institute of Aeronautical Engineering, Shenyang, Liaoning 110034, China)

Abstract: Question answering system can provides a precise and concise answer to a natural language query. Question classification is the first task of Question Answering System, and the precision of question classification has great effect on the subsequent processes. In this paper, we present a new method on feature extraction which uses HowNet as semantic resource, and use Maximum Entropy Model to realize it. We choose the interrogative words, syntax structure, question focus words and their first sememes as classification feature. The experiment result show that the first sememes in HowNet can express the main meaning of the question focus words, it can be as an important feature. This method can improve the precision of question classification: the classification precision of coarse classes and fine classes reaches 92.18 % and 83.86 % respectively.

Key words: computer application; Chinese information processing; question answering system; question classification; HowNet; maximum entropy model; classification feature;

1 引言

问答系统(Question Answering System)集知识表示、信息检索、自然语言处理于一体,比传统的基于关键字检索的搜索引擎更加方便、快捷、高效[1],能更好的满足用户的检索需求,近年来受到国际上的广泛关注,成为一个新的热点研究课题。

问答系统一般分为问题理解、信息检索、答案抽取以及答案验证几部分,几乎所有的问答系统在问题理解阶段都有问题分类这一过程。问题分类就是对于给定的问题,根据问题的答案类型把该问题映射到给定的语义类别中。例如,对于问题"中华人民共和国是哪一年成立的?",问题分类后得到的预期答案类型应该是"时间_年"。问题分类是问答系统所要处理的第一步,分类结果的正确率直接影响后续工作的进行。因此,问题分类是问题理解非常关键的一步。

对英文问题自动分类的研究最初采用了基于规则的方法,后来文献[2]又提出了采用 SVM(支持向

收稿日期: 2006-07-30 **定稿日期**: 2006-10-12

基金项目: 国家航空基金(05J54011);辽宁省自然科学基金(20042004)

作者简介: 孙景广(1981—),男,硕士生,主要研究方向为自然语言处理。

量机)进行分类的方法。文献[3,4]分别采用层次分类思想和 SNoW(Sparse Network of Winnow)分类器进行问题分类。文献[5]采用语义词典(WordNet)进行问题分类,也取得了不错的分类效果。

对中文问题自动分类的研究还不是很多,主要有复旦大学和哈尔滨工业大学,它们分别采用了 SVM 算法和改进的贝叶斯模型进行问题分类。后者对大类和小类的分类准确率分别达到了 86.62 % 和 71.92 %[6]。

本文提出了一种采用知网(HowNet)作为语义资源选取分类特征,并且使用最大熵模型(Maximum Entropy,ME)进行分类的新方法,取得了较好的实验结果。本文第二、第三部分将分别简要介绍知网和最大熵模型的有关内容;第四部分介绍本文所采用的问题分类体系;第五部分详细介绍了本文的分类特征的选取方法;第六部分给出具体的实验结果及错误分析;第七部分是总结和展望;第八部分是致谢。

2　知网

知网(HowNet)是一个以汉英双语来表示概念与概念之间以及概念的属性之间关系的知识库[7]。知网将客观世界中的词汇所代表的概念分为四大类:实体、事件、属性、属性值,并通过义原来标注概念。在知网中,义原是最基本的、不易于再分割的意义的最小单位,共包含了大约 2 200 个义原。义原间存在 8 种关系:上下位关系、同义关系、反义关系、对义关系、属性 —宿主关系、部件 —整体关系、材料 —成品关系、事件 —角色关系。这些义原以上下位关系为主干,形成树状结构分别存放于相应的义类文件中。

在知网中对于概念的定义采用知识描述语言(Knowledge Database Mark-up Language,KDML)来描述。KDML 对概念的定义采用 DEF 语义表达式,DEF 描述了词语详尽的语义特征,如:

生日: DEF = {time| 时间: TimeSect = {day| 日},{ComeToWorld| 问世:time = {}}}

词语在知网中的首义原是指该词语在 DEF 定义中出现的第一个义原,例如,"生日"的首义原就是"time| 时间"。它能较好地表达出该词语所对应概念的主要语义信息。

3　最大熵模型

最大熵(Maximum Entropy,ME)模型是一个比较成熟的统计模型,适合于分类问题的解决。目前该模型已成功应用于自然语言处理的多个领域,如:文本分类、词性标注、组块识别等等[8,9]。其基本思想是:对未知的不做任何假设,也就是当模型与训练数据中的已知约束条件一致且没有其他信息可利用时,它所估计的概率分布应该尽可能地接近均匀分布,即具有最大的熵。下面以基于答案类型的问题分类为例简要介绍一下最大熵模型的原理。

对于特定的问题样本集合 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$,其中每一个 $x_i (0 < i < n+1)$ 表示一个问题中所具有的词、语义、句法结构等特征信息,$y_i (0 < i < n+1)$ 表示问题的答案类型。在给定 T 和与之相关的约束条件下,存在一个唯一的概率模型 $P(y|x)$,其熵值最大,并且可以证明 $P(y|x)$ 的取值符合下面的指数模型:

$$p(y \mid x) = Z(x) \exp\left(\sum_i \lambda_i f_i(x, y) \right)$$

$$Z(x) = \frac{1}{\exp\left(\sum_i \lambda_i f_i(x, y) \right)}$$

其中 f_i 是一个特征函数,它的取值是 0 或 1。λ_i 是模型的参数,它表征了 f_i 对于模型的重要程度。$Z(x)$ 在 x 一定的情况下为一个泛化常数。上式使模型由求概率值转化为求参数值 λ_i,一般的估计方法是 Darroch 和 Ratcliff[10] 的通用迭代算法(Generalized Iterative Scaling,GIS),来得到具有最大熵分布的所有参数值 λ_i。

4　问题分类体系

由于对中文问题的分类还没有一个统一的分类体系,同时为了便于比较实验结果,本文采用了哈尔滨工业大学信息检索研究室提出的中文问题分类体系[6]。表 1 给出了该分类体系的具体内容,将所有问题分为 7 个大类,每个大类根据实际情况又定义了一些小类,共 60 小类。

表 1　本文采用的问题分类体系

大类(Coarse)	小类(Fine)
人物(HUM)	特定人物 团体机构 人物描述 人物列举 人物其他
地点(LOC)	星球 城市 大陆 国家 省 河流 湖泊 山脉 大洋 岛屿 地点列举 地址 地点其他
数字(NUM)	号码 数量 价格 百分比 距离 重量 温度 年龄 面积 频率 速度 范围 顺序 数字列举 数字其他

续表

大类(Coarse)	小类(Fine)
时间(TIME)	年 月 日 时间 时间范围 时间列举 时间其他
实体(OBJ)	动物 植物 食物 颜色 货币 语言文字 物质 机械 交通工具 宗教 娱乐 实体列举 实体其他
描述(DES)	简写 意义 方法 原因 定义 描述其他
未知(Unknown)	未知

5 问题分类特征的选取

本文一共选取了四种分类特征: 1. 问题疑问词(IW) 2. 句法结构(SS) 3. 疑问意向词(QFW) 4. 疑问意向词在知网中的首义原(FS)。在下面的具体介绍中以问题 Q: "CNN 第一次广播是什么时候?"为例进行说明。

5.1 问题疑问词的选取(IW)

中文疑问句中的疑问词包含着非常重要和明确的问题分类信息,是所有分类特征中最重要的特征,因此正确抽取疑问词非常关键。首先,我们将经常使用的一些疑问词,例如"什么"、"为什么"、"怎么样"、"谁"等建立一个疑问词词表 T。然后,对给定的问题进行分词和词性标注处理,例如得到"CNN/nx 第一/m 次/q 广播/vn 是/v 什么/r 时候/n",再选取其中标记为"/r"的词到 T 中进行查找,从而确定问题 Q 的疑问词是"什么"。另外,为了减少数据稀疏,对一些同义疑问词进行了转化处理,例如:"何时"、"何地"实际上和"什么时候"、"什么地方"完全等价,经过转化,统一选取"什么"作为疑问词。实验结果证明,这样的转化处理收到了较好的效果,对于"问题的句法结构"特征的正确抽取也有一定帮助。

5.2 句法结构的选取(SS)

汉语中疑问词是"什么"的问题比较常见,且涉及的问题种类较多,分类起来有一定难度。因此,本文对这类问题进行了特殊处理。我们发现这类问题在表达上有一些比较固定的句法结构,并且可以作为问题的分类特征。

我们认为疑问词以及疑问词附近的具有名词特性的词和动词含有重要的信息,把词性为"n、nx、ng、vn"等的各类名词统一标记为"n",把动词和疑问词分别标记为"v"和"r"。由于"的"字结构比较常见,把"的"字的词性标记为"D"。

通过以上标记,含有疑问词"什么"的问题的句法结构最终可以标记为 nvrD 的一个不同组合。另外,对于问题"什么是'禽流感'",分词后得到"什么/r 是/v '/w 禽流感/n '/w",标记后的句法结构是"rvn";对于另一个问题"什么是 pH 值",分词后得到"什么/r 是/v pH/nx 值/n",同样能够得到它的句法结构为"rvnn",但我们认为这两个问题的句法结构应该是一样的,所以把这两种句法结构规约为"rvn"。经过类似的规约,最后我们选取了 12 种具有代表性的句法结构作为分类特征。

5.3 疑问意向词的选取(QFW)

疑问意向词[11]是表达"问题问的到底是什么"这样一个含义的概念。关于疑问意向词目前还没有明确的定义。一般认为,用户的疑问意图就是要得到一个未知信息,也就是问题中最能体现答案类型的词。比如例句中的"时候"。

根据汉语句子的表达习惯,在问题疑问词附近的词更能表达整个句子所要表达的语义信息,对于问题分类常常具有更加重要的作用,特别是其中具有名词特性的词,也就是上面标记为"n"的词。

对于分类问题,我们发现通常疑问词右边标记为"n"的词所表达的语义信息比疑问词左边标记为"n"的词更丰富和有效。于是,我们提出了如下疑问意向词选取方法。

1. 选取疑问词右边标记为"n"的词作为疑问意向词,并最多选取两个。如果疑问词的右边没有标记为"n"的词,则转到第 2 步;

2. 在疑问词左边选取标记为"n"的词作为疑问意向词,并最多选取两个。

实验发现,如果存在有多个标记为"n"的词,并非选择的越多越好,选取的过多反而会增加很多干扰信息,产生噪声。因此,本文对于某一问题最多选取两个疑问意向词作为分类特征。

5.4 疑问意向词在知网中的首义原的选取(FS)

由于自然语言的丰富性,通常一个问题可以用多种不同的问句形式表达。例如,上述例句也可以改写为"CNN 第一次广播是什么日期?",或"什么时候 CNN 开始了第一次广播?"以及"CNN 的首次广播始于哪一年?"等形式,既可以是选词用词的不同,

也可以是句法结构的不同,但它们在语义上是相同或相近的,从问题分类的角度来看,应该属于同一类别。通过进一步的分析,我们发现不同问句之所以属于同一类别,常常是由于不同问句中的疑问意向词具有相同或类似的语义,在知网中它们对应概念的 DEF 定义中具有相同的首义原。如下所示,上例中的疑问意向词"时候"、"日期"、"年"的 DEF 首义原都是"时间"。

时候: DEF = {time| 时间}

日期: DEF = {time| 时间: TimeSect = {day| 日}}

年: DEF = {time| 时间: TimeSect = {year| 年}}

由此可见,疑问意向词的 DEF 首义原可以作为问题分类的一个语义特征。2005 版知网中对 81 447 个词汇进行了语义描述,定义了 157 185 个概念记录。应该说这个规模能够较好地覆盖目前开放域问答系统中出现在问题中的词汇。所以我们选取了知网作为语义资源,选取疑问意向词的 DEF 首义原作为问题分类的一个重要特征。

那么,对于给定的疑问意向词,如何才能自动地获取其 DEF 的首义原呢? 如果疑问意向词在知网中只有一个 DEF 定义,那么就可以从 DEF 中直接获取;如果疑问意向词是多义词,对应多个 DEF 定义时,原则上就需要先对疑问意向词进行语义识别,识别出在特定问题中的语义和对应的 DEF。众所周知,对于一般情况下的词汇语义识别(WSD)或标注问题,是自然语言处理中至今还未解决的著名难题。幸好,在问题分类中,疑问意向词多数为单义或少义的名词,使问题得到一定程度的简化。本文首先对每个大类和小类分类选取正确的分类义原,然后,对有多个 DEF 定义的词,直接选取和分类义原一致的 DEF 的首义原作为分类特征,一般来说,一个词的 DEF 定义只属于唯一的大类和小类分类。对于其他情况,我们目前只是简单地按知网中 DEF 的排列顺序进行选取。虽然以上方法不够精确,也过于简单,但实验结果表明是可行且有效的。

6　实验结果及错误分析

6.1　问题集

本实验使用了哈尔滨工业大学信息检索研究室和中科院自动化研究所模式识别国家重点实验室提供的问题集,并且按文献[6]同样的方法将问题集划分为训练集和问题集,最后经去重整理后得到的句子分布情况如表 2,可以看出各种问题的分布大致均匀。

表 2　训练语料和测试语料中的问题分布

大类(Coarse)	训练集问题数目	测试集问题数目
人物(HUM)	320	179
地点(LOC)	876	352
数字(NUM)	1 062	238
时间(TIME)	619	148
实体(OBJ)	982	225
描述(DES)	457	155
合　计	4 316	1 297

6.2　问题分类流程图

本文在总结已有方法的基础上,采用最大熵模型进行问题分类,大类和小类的训练和测试可以并行进行,具体的问题分类流程如下图。

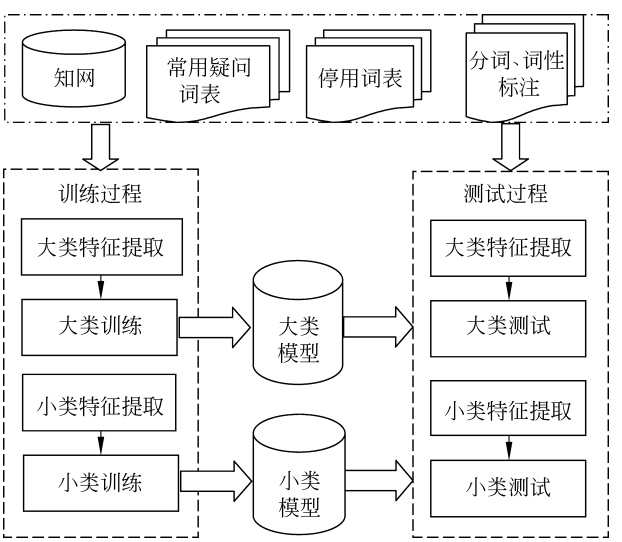

图 1　问题分类流程示意图

6.3　评价标准

对大类和小类的分类准确率采用了下面的公式进行评价:

$$\text{分类准确率} = \frac{\text{测试集中正确分类的问题数}}{\text{测试集中总的问题数}} \times 100\ \%$$

6.4　实验结果

本文使用最大熵模型进行问题分类,把问句的

疑问词(IW)、句法结构(SS)、疑问意向词(QFW)、疑问意向词在知网中的首义原(FS)作为分类特征,针对不同的特征组合进行了比较实验,得到的结果如表 3 所示。

表 3 选用不同的分类特征时得到的结果 ①

分类特征 / 准确率	IW	IW + SS	IW + QFW	IW + FS	IW + SS + QFW + FS
7 大类准确率	69.24 %	71.84 %	77.42 %	85.56 %	92.18 %
60 小类准确率	40.31 %	43.53 %	69.73 %	81.36 %	83.86 %

6.5 实验结果分析

由表 3 可以看出,当选取不同的特征时,采用最大熵模型的分类结果有明显的不同。当综合使用所有特征时,取得了较好的实验结果,7 个大类的准确率达到了最高值 92.18 %,60 个小类的准确率达到了 83.86 %。实验表明,通过进一步优化对首义原的选取,综合使用多种分类特征能使分类准确率得到进一步的提高。

通过对实验中的错误问题进行分析后发现,主要由以下原因造成:

第一:分词和词性标注造成的错误。这样会给疑问意向词和其首义原的特征选择带来错误。例如"什么花象征爱情",分词后得到"什么/r 花/v 象征/n 爱情/n",其中非常重要的疑问意向词"花"的词性应该是名词,但被错误地标注为动词"v",从而错误地选择了"象征"和"爱情"作为疑问意向词。

第二:训练集中存在的错误,部分问题分类标准不一致。例如:

(1) TIME_OTHER 电灯是什么时候发明的?

(2) LOC_OTHER 第一次使用青霉素是在什么时候?

(3) TIME_MONTH 美人蕉什么时候开花?

对于第一个和第二个问题,其实应该属于同一种分类中。对于第三个问题,如果理解为询问的是美人蕉开花的具体月份应该没有问题,如果理解为只是询问一个大体的表述时间,那么它的分类属于 TIME_OTHER 也可以。类似的问题会降低整个问题分类的准确率。

第三:训练集不能覆盖所有提问方式,当测试集中出现新的问题类型,很难正确分类。

例如:DES_REASON 图灵以什么闻名于世?

由于训练集中没有相似的表达方式,所以造成错误。

第四:由于中文问题的表达方式中存在着省略的现象,很难判断省略的是什么内容。

例如:NUM_CODE 急救中心怎么拨?

这种句子的省略很严重,仅根据疑问词无法判断,根据选择出来的疑问意向词不但无法判断,反而会造成噪声,更不利于特征的选择。这种问题现阶段还无法处理。

7 总结和展望

从实验结果可以看出,本文采用的最大熵模型表现出了较好的性能。本文主要运用知网作为语义资源,从语义的角度进行特征选取,最终使 7 个大类的准确率达到 92.18 %,60 个小类的准确率达到 83.86 %,比同类实验结果[6] 分别提高了 5.54 %和 11.94 %。

目前,疑问意向词以及疑问意向词在知网中的首义原的选取方法还可以进一步改进。并可以考虑加入知网提供的其它语义信息,进一步提高问题分类的准确率。

8 致谢

本文使用了哈尔滨工业大学信息检索研究室和中科院自动化研究所模式识别国家重点实验室提供的问题集。在此,对他们表示诚挚的感谢!

参考文献:

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

[2] Dell Zhang, Wee Sun Lee. Question classification using

① 由于对特征获取方法进行了一些改进,本实验结果比学生会议时发表的实验结果有所提高。

support vector machines[A]. In:the 26th ACM SIGIR [C]. 2003.

[3] Xin li ,Dan Roth. Learning Question classification using support vector machines[A]. In: the 26th ACM SIGIR [C]. 2003.

[4] Carlson ,C. Cumby ,J. Rosen ,etal. The SNoW learning architecture [A]. In: UIUCDCS-R-99-2101, UIUC Computer Science Department[C] ,2004 ,451-458.

[5] Xin Li, Dan Roth. The Role of Semantic Information in Learning Question Classifiers[A]. In: First International Joint Conference on Natural Language Processing[C] ,2004 ,451-458.

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

[7] 董振东 ,董强. 知网. http://www.keenage.com/zhiwang/c_zhiwang.html.

[8] 李荣陆 ,王建会 ,陈晓云 ,等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展 ,2005 ,42 (1): 94-101.

[9] R Adwait. A maximum entropy model for Part-of-Speech tagging[A]. In: Proceedings of the Empirical Methods in Natural Language Processing Conference [C]. Philadelphia, USA. 1996.

[10] Darroch, J. N, Ratcliff, D. Generalized Iterative Scaling for Log-Linear models [J]. Annals of Mathematical Statistics 1972, 43(5) :1470-1480.

[11] 吕德新. 中文自动问答系统中问题理解技术的研究[D]. 沈阳航空工业学院硕士论文 ,2006 年 3 月.

书　讯(合订本)

2006 年《中文信息学报》合订本已出 ,还有少量过刊合订本 ,详细定价如下:

出版年份	定价(元)	出版年份	定价(元)
1997	30	2002	55
1998	30	2003	55
1999	55	2004	65
2000	55	2005	70
2001	55	2006	85

愿购者(邮购需加 10 %的邮资费) ,请按以下地址汇款:

邮编:100080　　通信地址: 北京 8718 信箱《中文信息学报》编辑部

电话:010-62562916　　E-mail: cips@iscas.ac.cn