面向开放域问答的问题分类技术研究进展

杨思春[1],戴新宇[2],陈家骏[2]
(1.安徽工业大学计算机科学与技术学院,安徽马鞍山 243032;2.南京大学计算机软件新技术国家重点实验室,江苏南京 210023)

摘 要: 开放域问答是当前自然语言处理和信息检索领域的研究热点,作为开放域问答系统的重要组成部分,问题分类可以缩小答案的搜索空间并决定答案的选择策略.近年来,基于机器学习的问题分类技术受到广泛的关注,相关研究表明问题分类的准确性直接影响问答系统的整体性能.本文从分类体系与数据集、特征提取、分类器设计、性能评测等层面,总结了问题分类技术近年的主要研究成果.重点分析了各种基于监督学习的问题分类方法的特点和不足,讨论了核方法、半监督学习、主动学习、迁移学习等在问题分类中的应用,同时对问题分类技术未来研究动向进行了展望.

关键词: 开放域问答; 问题分类; 机器学习; 特征提取; 分类器设计
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2015)08-1627-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.08.024

Advances in Question Classification for Open-Domain Question Answering

YANG Si-chun[1], DAI Xin-yu[2], CHEN Jia-jun[2]
(1. *School of Computer Science and Technology, Anhui University of Technology. Maanshan, Anhui* 243002, *China*;
2. *State Key Laboratory for Novel Software Technology, Nanjing University. Nanjing, Jiangsu* 210023, *China*)

Abstract: Open-domain question answering is becoming a hot topic in the fields of natural language processing and information retrieval. Question classification, as an important component of question answering, has shown its significant influence on the overall performance of question answering systems. It can help reduce the search space and choose the exact search strategy to find answers. In this paper, we present a through overview of the state-of-the-art approaches to question classification, in terms of category/dataset, feature extraction, classification methods and performance metrics. Firstly, we give a detailed analysis of the supervised learning based question classification approaches. Then, we introduce some related work on question classification, such as kernel methods, semi-supervised learning methods, active learning and transfer learning methods, and so on. Finally, we give some possible research directions on question classification.

Key words: open-domain question answering; question classification; machine learning; feature extraction; classifier design

1 引言

开放域问答(open-domain question answering)是当前自然语言处理和信息检索领域倍受关注的研究方向,它允许用户以自然语言形式进行提问并返回简洁、准确的答案.自 1999 年 TREC(Text REtreival Conference)会议引入问答系统评测后,吸引了众多研究机构开展问答技术的研究[1].特别是 2011 年 IBM 的 Watson 系统在著名的智力竞猜电视节目 Jeopardy 中一举击败两位冠军选手以后,更是表现了问答系统超强的问题回答能力.

总体上可以将问答系统分为问题分析、文档检索和答案抽取三个部分[2,3].作为问题分析的第一步,问题分类(question classification)通过确定问题的目标答案类型,为后续答案抽取提供语义约束.问题分类对提升问答系统的整体性能具有特别重要的意义,Moldovan 等[4]的实验结果表明由于问题分类不准确导致答案出错的影响最大.

对于问题分类,由于问句所包含的词汇很少,决定了问题分类比文本分类更难[5].早期大多基于规则方法对问题进行分类[6],但是规则的维护和更新非常困难,在应用领域改变时移植性和适应性也较差.

随着机器学习在相关领域的成功应用,基于机器学习的问题分类方法占据越来越重要的角色[5~7],该方法使研究者把主要精力放在分类模型的选择和分类特征

收稿日期:2014-07-21;修回日期:2015-02-10;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61472183,No.61170181)

的提取上.近年来在自然语言处理、信息检索和机器学习等领域的国际顶级会议上多次出现问题分类方面的研究论文.由此不仅看出 QA 研究人员对问题分类的重视程度,而且也反映了机器学习在问题分类任务中的重要应用.

本文从分类体系与数据集、特征提取、分类器设计、性能评测等层面,总结了问题分类技术近年的主要研究成果.重点分析了各种基于监督学习的问题分类方法的特点和不足,讨论了核方法、半监督学习、主动学习、迁移学习等在问题分类中的应用,同时对问题分类技术未来研究动向进行了展望.

2 分类体系与数据集

2.1 分类体系

分类体系是问题分类的依据,分类体系设计是否恰当直接关系到问题分类的性能.在历年 TREC QA 评测中,绝大多数机构都采用自己的分类体系.针对英文问题分类,Li 等[5]定义的分类体系(如表 1 所示)最为典型,该分类体系具有较好的覆盖面,比较适合开放域问题[6~9].但是,由于该分类体系过多关注 factoid 问题,Bu 等[10]将其称为基于内容(content-based)的分类体系,同时提出一种基于功能(function-based)的分类体系,包含 fact、list、reason、solution、definition 和 navigation 等 6 个类别.

针对中文问题分类,研究人员大多参照哈尔滨工业大学社会计算和信息检索中心提出的中文问题分类体系(如表 2 所示).另外,邱锡鹏等[11]从问题形式和答案类型两方面建立了一套中文问题分类体系,刘小明等[12]采用领域本体的层次结构和实体模型作为中文问题分类体系.

表 1 UIUC 英文问题分类体系

大类	小类
ABBR	abbreviation, expansion
DESC	definition, description, manner, reason
ENTY	animal, body, color, creation, currency, disease, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUM	description, group, individual
LOC	city, country, mountain, other, state
NUM	code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight

从参加 TREC QA 评测所采用的问题分类体系来看,问题类别比较倾向于较小粒度[5].因为小粒度问题类别在语义信息的表现方面更加丰富,可以为后续的答案抽取提供更加有效的语义约束.但是太小的类别粒度反而增加问题分类的难度,进而也影响最终的答案生成.Suzuki 等[13]度量了基于四种类别粒度的问题分类精度,结果发现从第一到第四层次,分类精度从 95% 下降到 75%.分类体系到底包含多少类别,是采用平行(flat)还是层次(hierarchical)结构,需要考虑特定问答系统的性能需求.

表 2 哈工大中文问题分类体系

大类	小类
DES	abbr, definition, expression, judge, manner, meaning, other, reason
HUM	desc, list, organization, person
LOC	addres, city, continent, country, island, lake, list, mountain, ocean, other, province, river
NUM	area, code, count, distance, frequency, list, money, multiple, order, percent, period, range, speed, temperature, weight, other
OBJ	animal, body, clothing, color, culture, currency, disease, event, food, instrument, language, list, material, microbe, other, plant, religion, social, sports, substance, term, universe
TIME	age, day, holiday, list, month, other, range, season, solarterm, time, week, year

现有问题分类体系大多基于问题类别有助于答案类型判定来设计的,实际上问题类别还应有助于选择答案生成的方法.Wu 等[14]从问题分类与答案生成方法、问题分类与答案语义约束两个不同的视角,提出一种新的问题分类体系,该分类体系平行地给出了方法(method)类别体系和语义(semantic)类别体系.

现有问题分类体系都需要预定义问题类别,Pinchak 等[15]指出了这种静态分类体系的缺点:(1)有些问题不能与预定义问题类别匹配;(2)预定义问题类别的粒度不好把握.为此,Mikhailian 等[16]提出一种静态和动态相结合的问题分类体系,对于显式的答案类型,采用静态的 AP(asking point)问题类别,对于隐式的答案类型,则采用动态的 EAT(expected answer type)问题类别.

2.2 数据集

基于机器学习的问题分类需要一定规模的数据集来训练学习模型.对于英文问题分类,以 UIUC 问题集(http://12r.cs.uiuc.edu/~cogcomp/data/ qa/qc)最有代表性.其中,训练问题主要来自 USC[6]、TREC 8 和 TREC 9,测试问题全部来自 TREC 10.由于具有较好的覆盖面,该问题集已经成为当前英文问题分类研究的公用数据集[7~9].但是,UIUC 问题集的类别分布不是非常均衡.Li 等[6]以取自 TREC 8 和 TREC9、USC 以及人工构造的问题组成训练集,使 UIUC 问题集中问题类别的分布更加趋于均衡.

对于中文问题分类,目前还没有标准问题集.现有文献大多采用哈尔滨工业大学社会计算和信息检索研

究中心、中国科学院自动化所提供的问题集[17~19],也有研究人员通过翻译 TREC QA 或 UIUC 问题集来构造中文问题集.

3 特征提取

特征提取是基于机器学习的问题分类中最为重要的环节,好的特征能够显著地改善问题分类的性能.以下从词法、句法和语义等层面,分别介绍问题分类所需的各种特征信息.

3.1 词法特征

词袋(bag-of-words)是问题分类的最基本特征,常用的有 unigram, bi-gram 和 tri-gram[14].虽然 bi-gram 和 tri-gram 同时考虑了词序信息,但 Huang 等[8]的实验结果表明,总体上 bi-gram 和 tri-gram 的分类精度要比 unigram 低很多.因此,一般很少使用 bi-gram 和 tri-gram.

单纯基于词袋特征进行问题分类的精度并不理想.Zhang 等[7]利用 SVM 分类器在 UIUC 数据集上的实验结果表明词袋特征的小类分类精度仅为 79.2%.这是因为(1)没有区分问句中每个词对问题分类的贡献;(2)没有考虑问句的句法、语义信息.但是,词袋特征是其他所有特征的基础.

除词袋特征以外,Huang 等[8]还利用了问句的数字(digit)、小写(lower case)、大写(upper case)及其混合形式(mixed)等词形特征.另外,Blunsom 等[20]还使用了问句的长度特征.

3.2 句法特征

词性(part-of-speech)是最简单的句法特征.虽然相关研究表明单独的词性特征对问题分类的贡献很小,但是一般都把它作为基本特征之一.

疑问词是问题分类时可以利用的重要句法特征,有时直接根据某些疑问词就可以判定问题的类别.Huang 等[8]的实验结果表明疑问词对问题分类有较大的贡献.关于疑问词提取,可以根据疑问词词表或机器学习方法[21].

除疑问词以外,疑问词附近的词,特别是疑问词右边的名词性质的词,对问题分类也具有重要的作用,这类词称为中心词(head word)或焦点词(focus word).Huang 等[8]的实验结果表明在疑问词的基础上引入中心词以后,小类分类精度由 46.8% 上升至 82%.Silva 等[22]和 Loni 等[23]的实验结果也验证了中心词在问题分类中的重要作用.关于中心词或焦点词的提取,Huang 等[8]和 Li 等[24]基于浅层句法分析和启发式规则提取中心词,Zhang 等[21]基于条件随机域(CRFs)进行焦点词识别.另外,为了使焦点词具有更强的问题类别区分能力,Bunescu 等[25]还将焦点词扩展为名词短语的形式.

除上述句法特征以外,Li 等[6]还将疑问词后面第一个名词短语或动词短语即中心语块(head chunk)作为特征,Nguyen 等[26]将问句的句法结构子树作为特征,文勖等[18]将问句主干(主谓宾)和疑问词及其附属成分作为特征,李鑫等[27]将问句中的依存关系作为特征.

这里要说明一点的是,由于现有文献大多通过句法分析来获取问句句法特征,因此实际特征提取的处理开销较大.另外,句法分析器的性能也会影响问题分类效果.

3.3 语义特征

利用词法和句法特征进行问题分类可以获得较高的大类分类精度,但是对于语义要求更高的小类分类,还需要进一步提取相应的问句语义特征.

词汇语义是可以直接利用的最简单语义特征.英文问题分类一般利用 WordNet[28]作为词汇语义知识资源.Li 等[6]将问句中每个词在 WordNet 中的所有词义及其上位词(hypernym)作为语义特征.但是,直接利用所有词的 WordNet 语义可能会引入噪声.为此,Huang 等[8]提取中心词的 WordNet 上位词作为分类特征,李鑫等[27]将名词的 WordNet 词义及其上位词,以及中心动词(head verb)的同义词集作为分类特征.

对于中文问题分类,一般利用 HowNet[29]作为词汇语义知识资源.余正涛等[30]选取 HowNet 词义及其同义词词义作为问题分类特征,孙景广等[19]以疑问意向词的 HowNet 首义原作为问题分类特征.

命名实体(named entity)也可以作为问题分类的语义特征.Li 等[5]验证了命名实体对问题分类有一定的贡献.但是,由于命名实体的类型数量以及在问句中的分布均很少,一般认为命名实体对问题分类的贡献很小.

Li 等[5]还利用了类别关联词(semantically related words)作为问题分类的语义特征,但是由于类别关联词是手工构造的,因此特征提取的成本非常昂贵.为了避免人工获取,Metzler 等[31]给出一种利用 WordNet 自动获取类别关联词的方法.

与问题分类所利用的句法特征相比,在语义特征的获取方面还有很大的探索空间.在语义知识的来源方面,不仅仅局限于 WordNet、HowNet 等,还可以利用 Wikipedia 等其他语义知识资源.例如,Ray 等[32]同时利用 WordNet 和 Wikipedia 来获取问句中的其他语义特征.

表 3 对现有文献中不同词法、句法和语义特征的使用方式等进行了比较.

表 3 不同词法、句法和语义特征的应用

词法特征	使用方式	分类性能	备注
词袋(Unigram)	最常用	较好	可单独使用
词袋(Bi-gram)	较少用	一般	可单独使用
词袋(Tri-gram)	很少用	一般	可单独使用
词形(数字、大小写及其混合)	较少用	不明显	须与词袋等其他特征组合使用
句子长度	较少用	不明显	须与词袋等其他特征组合使用
句法特征	**使用方式**	**分类性能**	**备注**
词性	常用	较好	一般与词袋等其他特征组合使用
疑问词	常用	较好	可单独使用
中心(焦点)词	常用	较好	可单独使用
(中心)语块	较常用	不明显	一般与词袋等其他特征组合使用
(依存)句法结构	较常用	较好	一般与词袋等其他特征组合使用
语义特征	**使用方式**	**分类性能**	**备注**
词义	常用	较好	一般与词袋等其他特征组合使用
词义(上位词)	较常用	较好	一般与词袋等其他特征组合使用
词义(同义词)	较常用	较好	一般与词袋等其他特征组合使用
命名实体	常用	不明显	须与词袋等其他特征组合使用
类别相关词	较少用	较好	可单独使用

4 分类器设计

4.1 分类器结构

分类器结构设计是否合理对问题分类有较大的影响,在相同特征的情况下,不同结构的分类器往往会导致分类结果差异很大.分类器结构通常有平行和层次式两种,层次式结构首先进行简单的大类分类,再依次向分类难度逐渐增加的小类展开.例如 Li 等[5]、Metzler 等[31]、李方涛等[33]定义的问题分类器都是层次式的.实验结果表明,层次分类的精度要好于平行分类的精度,各层采用不同类型的分类器要好于采用相同类型的分类器[5].

对多个分类器输出的结果进行融合是问题分类器结构发展的一个方向.例如,Bu 等[10]提出一种基于 MLN(markov logic network)的问题分类器,将基于规则的方法和基于统计的方法进行合一;Silva 等[22]以预定义的问题分类规则作为特征构造基于学习的问题分类器;李鑫等[27]基于错误驱动集成规则学习方法 TBL 和统计方法 SVM;张志昌等[34]根据问题结构(疑问词和焦点词)的不同分别选择不同的分类器(最近邻和 SVM);Banerjee 等[35]基于投票(voting)机制对 Naive Bayes, Kernel Naive Bayes, Rule Induction 和 Decision Tree 等分类器进行集成.

4.2 学习算法

问题分类学习算法常用的有最近邻、朴素贝叶斯、决策树、SNoW、最大熵和 SVM 等.其中,最近邻、朴素贝叶斯、决策树等都是机器学习领域常用的基本学习算法,但是将它们用于问题分类时性能并不理想[7].这里只介绍分类性能较好的 SNoW、最大熵和 SVM 等学习算法.

(1)SNoW 它是对原始 Winnow 算法的一种改进,可以高速处理大数据量、多类别问题.UIUC 大学开发、共享了 SNoW 算法的工具包(http://l2r.cs.uiuc.edu/~danr/snow.htm).Li 等[5]将该学习算法成功应用于问题分类,并获得了 84.2%的小类分类精度.

(2)最大熵 又称指数模型(exponential)、对数线性模型(log-linear).最大熵是目前问题分类研究中仅次于 SVM 的学习模型之一.Manning 等[36]利用最大熵模型进行问题分类,获得了 89.0%的小类分类精度.

(3)支持向量机 是基于 Vapnik[37]提出的统计学习原理构建的学习模型,在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势[38].关于 SVM 的数学模型,车万翔的博士论文[39]对其作了非常通俗的解释和分析,此处不再赘述.

当前问题分类大多采用 SVM 和最大熵模型.最大熵模型较之 SVM 有更快的训练速度,但是不便于使用复杂的结构特征.SVM 理论上对噪声数据和不相关特征具有鲁棒性,其分类的准确率较高.Zhang 等[7]的研究工作表明,在使用相同特征的情况下,SVM 算法的性能要明显高于其他学习算法.

表 4 对现有文献中不同学习算法的优缺点、分类性能等进行总结.

表 4 不同分类器的特点及性能分析

分类器	优点	缺点	分类性能
NN	算法简单	效率低	一般
NB	算法简单,误差率低	特征相互独立条件难以满足	一般
DT	算法简单,效率高	过拟合	一般
SNoW	训练快,精度高	所需特征数量很大	较好
MEM	训练快,精度高	存在数据稀疏问题	较好
SVM	精度高	特征数量很大时训练较慢	较好

4.3 核方法

针对问题分类时利用 SVM 的标准核函数表示结构化特征并不适合,已有文献将核方法(kernel method)引

入问题分类.核方法是近年机器学习领域的研究热点,它通过定义高效的核函数以替代特征向量方法的向量内积运算,并隐藏了低维线性不可分空间映射到高维线性可分空间的具体细节.

Haussler[40]最早提出用于离散结构的卷积核(convolution kernel).这里,介绍卷积树核(convolution tree kernel),它以训练集中所有句法树的子树片段作为特征向量的组成,但是不需要显式抽取这些子树片段,通过树核函数就可以隐式地实现两个向量的内积计算.

Zhang 等[7]利用树核 SVM 获得的大类分类精度与 Li 等[5]仅差 1 个百分点,充分体现了树核函数对问题分类的重要贡献;Moschitti 等[41]设计的浅层语义树核获得了 91.8%的大类分类精度,高于 Zhang 等[7]利用改进的 Collins 树核的分类精度,也接近于 Li 等[6]基于特征向量方法的分类精度.

由于核函数具有良好的复合特性,可以通过对树核和线性核进行复合操作,以实现更加高效的问题分类.例如,Liu 等[42]将线性核与句法依存树核进行复合.也可以将不同的树核进行复合,或者将树核与串核进行复合.

实际应用核方法进行问题分类时,需要根据所处理问题本身的特点来设计相应的核函数.例如,Moschitti 等[43]在对 Jeopardy! 问题进行分类时,发现 ST 树核(sub tree)的分类精度远低于 PT 树核(partial tree)的分类精度.

4.4　半监督学习

对基于监督学习(supervised learning)的问题分类,由于人工标注训练数据的成本很高,严重影响了分类精度的进一步提升.因为未标注数据比较容易获得,半监督学习(semi-supervised learning)已成为机器学习领域的一个研究热点.

在机器学习领域,已经提出了各种半监督学习方法.其中,Blum 等[44]提出的协同学习(co-training)是一种典型的半监督学习方法.为解决标准 co-training 算法处理标记置信度时出现的时间问题和不稳定问题,Zhou 等[45]进一步提出了一个简单、高效的 tri-training 算法.

半监督学习可以有效解决当前问题分类所面临的数据标注困难.Yu 等[46]提出一种基于 co-training 的问题分类方法,进一步提升了问题分类的精度;Tri 等[47]应用改进的 tri-training 算法进行问题分类,在相同分类器和特征下比基于监督学习的问题分类提高 3 个多百分点.

4.5　其他学习范式

现有问题分类方法需要利用所提取的大量特征来训练分类模型,这种学习范式称为被动学习(passive learning).主动学习(active learning)可以在不借助复杂特征的情况下,选取那些最有价值的样本实例来提升学习的效果.Mishra 等[48]提出一种基于主动学习的问题分类方法,分类精度比基于 co-training 的问题分类方法高 4.8 个百分点;Marinˇciˇc 等[49]提出一种基于领域模型和主动学习的问题分类方法,分类精度比基于被动学习的问题分类方法高 10 个百分点.

在机器学习领域,通过迁移学习(transfer learning)可以在训练集和测试集数据不满足同领域同分布的情况下,将一个领域中的知识迁移学习到另一个相关的领域.基于迁移学习实现问题分类时,对领域适应性问题的解决最为关键.Su 等[50]基于迁移学习并利用不同领域的训练数据实现问题分类,进一步改善了问题分类的性能.

另外,随着机器学习领域研究的不断深入,其他一些学习范式如集成学习[35]、增量式学习[51]等也开始应用于问题分类,并取得了一定的分类效果.

5　性能评测

5.1　评价标准

对于问题分类的性能评价,通常采用分类准确率(accuracy)作为标准,其定义如下:

$$\text{accuracy} = \frac{\text{被正确分类问题数}}{\text{总的问题数}} \tag{1}$$

对于衡量单个类别上的分类准确率,需要同时给出其分类精度 P(precision)、召回率 R(recall)和 $F1$ 值($F1$-score),具体定义如下:

$$\text{precision} = \frac{\text{该类被正确分类问题数}}{\text{该类总的问题数}} \tag{2}$$

$$\text{recall} = \frac{\text{该类被正确分类问题数}}{\text{该类被预测的问问题}} \tag{3}$$

$$F1 = \frac{2PR}{P+R} \tag{4}$$

对于衡量所有类别上的分类精度,其准确率与召回率实际上是一样的.但是,对于衡量单个类别上的分类精度,其准确率与召回率不一样,因为该类中的问题可能被分类到其他不同的类别.

对于一个问题可以属于几个不同问题类别的情况,实际预测结果只要属于其中某一类别就可以认为该问题被正确分类.为此,Li 等[5]采用了两种不同的评价标准(P_1和 $P_{\leqslant 5}$).其中,P_1是指对问题进行分类时每个问题只能赋予 1 个问题类别;$P_{\leqslant 5}$是指对问题进行分类时允许为每个问题赋予最多 5 个问题类别.以下是$P_{\leqslant 5}$的具体定义:

假设 $k_i(k_i \leqslant 5)$为第 i 个问题所属的类别标签数,并按照类别标签密度值递减排序.首先定义 I_{ij},

$$I_{ij}=\begin{cases}1, \text{若第 } i \text{ 个问题的正确类别排在第 } j \text{ 个位置}\\ 0, \text{否则}\end{cases} \tag{5}$$

基于 I_{ij},再分别定义 P_1和 $P_{\leqslant 5}$如下,

$$P_1 = \sum_{i=1}^{m} I_{i1}/m \tag{6}$$

$$P_5 = \sum_{i=1}^{m}\sum_{j=1}^{k_i} I_{ij}/m \tag{7}$$

这里,m 是测试集所含的问题数.

采用评价标准 $P_{\leqslant 5}$的好处是,可以使问题的分类结果更好地服务于问答系统的后续其他处理阶段. *Li* 等[5]的实验结果表明,允许一个问题属于几个不同问题类别不仅没有导致后续答案生成出错,反而确保了系统对候选答案的正确选择.

5.2 现有工作的比较

表 5[52]、表 6 分别对近年来英文问题分类和中文问题分类的一些代表性工作进行了比较.

表 5 英文问题分类工作的比较

	分类器	特征集	大类精度	小类精度
(Li et al,2002)	SNoW	*U + P + HC + NE + R*	91.0%	84.2%
(Zhang et al,2003)	Tree kernel SVM	*U + NG*	90.0%	
(Metzler et al,2005)	RBF kernel SVM	*U + B + H + HY*	90.2%	83.6%
(Krishnan et al,2005)	Linear SVM	*U + B + T + IS + HY*	94.2%	88.0%
(Li et al,2006)	SNoW	*U + P + HC + NE + R + S*		89.3%
(Blunsom et al,2006)	MEM	*U + B + T + P + H + NE + more*	92.6%	86.6%
(Merkel et al,2007)	Language Model	*U + B*		80.8%
(Li et al,2008)	SVM + CRF	*U + L + P + H + HY + NE + S*		85.6%
(Pan et al,2008)	Semantic tree kernel SVM	*U + NE + S + IH*	94.0%	
(Huang et al,2008)	MEM	*U + WH + WS + H + HY + IH*	93.6%	89.0%
(Huang et al,2008)	Linear SVM	*U + WH + WS + H + HY + IH*	93.4%	89.2%
(Silva et al,2011)	Linear SVM	*U + H + HY + IH*	95.0%	90.8%
(Loni et al,2011)	Linear SVM	*U + B + WS + H + HY + R*	93.6%	89.0%

表 6 中文问题分类工作的比较

	数据集	分类器	特征集	大类精度	小类精度
(张宇等,2005)	训练集 3300 测试集 980	Bayes	词、词性		72.4% (65 小类)
(余正涛等,2005)	训练集 1200 测试集 300	SVM	词、词性、语块、命名实体、词义、同义词、类别关联词	88.7% (6 大类)	
(文勖等,2006)	训练集 5265 测试集 1300	Bayes	主干词、疑问词及附属	86.62% (7 大类)	71.92% (60 小类)
(孙景广等,2007)	训练集 4316 测试集 1297	MEM	疑问词、句法结构、疑问意向词及其《知网》首义原	92.18% (7 大类)	83.86% (60 小类)
(张志昌等,2009)	训练集 36472 测试集 1157	NN + SVM	非停用词词性及词义、疑问词及词义、焦点词及词义	96.11% (6 大类)	89.97% (75 小类)
(李茹等,2009)[53]	训练集 1341 测试集 670	MEM	疑问词、框架元素及中心词、框架名-疑问词框架元素	91.38% (7 大类)	83.2% (73 小类)
(余正涛等,2010)	训练集 21025 测试集 2337	Co-forest	高频词	88.9% (5 大类)	78.2% (23 小类)
(杨思春等,2014)[54]	训练集 4966 测试集 1300	SVM	基本特征及其词袋绑定特征		84.7% (77 小类)

对于英文问题分类,由表 5 可以看出:

(1)在分类器选择方面,大多采用 SVM,也有采用 MEM 和 SNoW 分类器.实验结果表明,SVM 的分类精度要明显高于其他分类器.

(2)在特征提取方面,除词袋以外,还利用中心词、疑问词、类别相关词、中心语块等特征,而且分类精度高的几乎都利用了这几类特征.

(3)核方法能有效获取问句的句法、语义结构信息,但是受现阶段自然语言处理技术所限,目前这方面的研究工作还比较初步.

对于中文问题分类,由表 6 可以看出:

(1)大多采用 SVM 和 MEM 分类器,并利用词性、词义、命名实体、中心词(焦点词,疑问意向词,主干词)、疑问词、句法结构等其他特征.

(2)目前中文问题分类主要还是基于监督学习.将核方法、半监督学习、主动学习、迁移学习等应用到中文问题分类的研究工作还非常少见.

(3)与英文问题分类相比,由于中文自然语言处理技术所限及相应语义知识资源的缺乏,现阶段中文问题分类精度普遍较低,特别在小类分类上.

6　总结与展望

本文从分类体系与数据集、特征提取、分类器设计、性能评测等层面总结了问题分类技术近年的主要研究成果,重点分析了各种基于监督学习的问题分类方法的特点和不足,讨论了核方法、半监督学习、主动学习、迁移学习等在问题分类中的应用.

在各种问题分类方法中,基于监督学习的问题分类方法占据主流并取得了较高的分类精度,核方法、半监督学习、主动学习、迁移学习等问题分类方法也取得了一定的分类效果.但是,随着开放域问答面临更加复杂的应用场景,问题分类研究还需要在以下几个方面不断深化和拓展:

(1)高效的问句特征提取技术研究

特征提取对基于机器学习的问题分类至关重要,现有研究大多基于自然语言处理技术获取问句特征.这种特征提取方式不仅处理开销大,而且还会受到自然语言处理技术的限制.对于中文问题分类,由于现阶段中文自然语言处理技术的不成熟,这种特征提取方式对问题分类性能的影响更为明显.目前在自然语言处理领域已有研究人员基于深度学习(deep learning)来获取文本特征[55,56].可以尝试将深度学习用于问题分类中的问句特征提取,在一定程度上克服现有特征提取技术的不足.

(2)融合多种学习范式的问题分类研究

在当前基于机器学习的各种问题分类方法中,基于监督学习的问题分类方法占据主流,核方法、半监督学习、主动学习、迁移学习等在问题分类中的应用还比较初步.下一步可以考虑结合各种学习范式的优点,在传统的监督学习的基础上,融合其他学习范式,并根据待分类问题的特点动态地选取合适的学习模型.

(3)面向真实问题的问题分类研究

受近年诸多 QA 评测(TREC, CLEF 和 NTCIR 等)的影响,现有问题分类研究大多基于人工编辑处理的简单问题,这些问题不能代表互联网上海量的真实问题.对于中文问题分类,由于中文语言表达的多样性和复杂性,实际所面临的问题更为复杂.为了使问题分类真正有助于问答系统的答案生成,下一步需要深入开展面向真实问题的问题分类研究.

(4)面向社区问答系统的问题分类研究

近年来,社区问答(community question answer-ing)[57]逐渐成为开放域问答的研究热点.在社区问答中对问题进行分类有助于用户更加便捷地获取问题的解答,但是现有问答社区中的问题都是按照预定义的分类体系组织的[58],在用户无法确定问题类别时需要系统自动地对其归类.由于要考虑分类体系的动态更新[59]和复杂的多句子级问题的处理[60]等,实现社区问答的问题分类要面临更多的技术挑战.目前这方面研究工作还非常少见.

(5)面向知识库问答系统的中文问题分类研究

针对知识库问答系统[61],中文问题分类比英文问题分类面临更高的语义要求.为了进一步提升中文问题分类的性能,需要在 HowNet 等现有语义知识资源的基础上,通过结合知识图谱[62](knowledge graph)等其他语义知识资源,来获取更为丰富的语义特征信息.目前知识图谱已开始应用于搜索引擎等领域,但是由于现阶段支持中文的知识图谱很少,构建中文知识图谱是一项重要的基础性工作.

参考文献

[1] 张志昌,张宇,等.开放域问答技术研究进展[J].电子学报,2009,37(5):1058 – 1069.
Zhang Zhichang, Zhang Yu, et al. Advances in open-domain question answering[J]. Acta Electronica Sinica, 2009, 37(5): 1058 – 1069. (in Chinese)

[2] 范士喜,王晓龙.面向真实环境的问句分析方法[J].电子学报,2010,38(5):113 – 1135.
Fan Shixi, Wang Xiaolong. Real environment oriented question analyzing[J]. Acta Electronica Sinica, 2010, 38(5): 1131 – 1135. (in Chinese)

[3] 高明霞,刘椿年.基于约束的自然语言问题到 OWL 的语义映射方法研究[J].电子学报,2007,35(8):1598 – 1602.
Gao Mingxia, Liu Chunnian. A constraints-based semantic mapping method from natural language questions to OWL[J]. Acta Electronica Sinica, 2007, 35(8): 1598 – 1602. (in Chinese)

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

[5] Li X, et al. Learning question classifiers[A]. Proc of the 19th International Conference on Computational Linguistics[C]. Taipei, China: Association for Computational Linguistics, 2002. 1 – 7.

[6] Li X, et al. Learning question classifiers: the role of seman-tic

information[J]. Natural Language Engineering, 2006, 12 (3): 229 – 249.

[7] Zhang D, et al. Question classification using support vector machines[A]. Proc of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Toronto, Canada: ACM, 2003. 26 – 32.

[8] Huang Z, et al. Question classification using head words and their hypernyms[A]. Proc of the 2008 Conference on Empirical Methods on Natural Language Processing[C]. Honolulu, USA: Association for Computational Linguistics, 2008. 927 – 936.

[9] Huang Z, et al. Investigation of question classifier in ques-tion answering[A]. Proc of the 2009 Conference on Empirical Methods in Natural Language Processing[C]. Singapore: Association for Computational Linguistics, 2009. 543 – 550.

[10] Bu F, et al. Function-based question classification for general QA[A]. Proc of the 2010 Conference on Empirical Methods in Natural Language Processing[C]. Massachusetts, USA: Association for Computational Linguistics, 2010. 1119 – 1128.

[11] 邱锡鹏,缪有栋,等.基于主动学习的中文问题分类数据集构建[J].哈尔滨工业大学学报,2012,44(5):125 – 128.
Qiu Xipeng, Miao Youdong, et al. Constructing Chinese question classification dataset with active learning[J]. Journal of Harbin Institute of Technology, 2012, 44(5): 125 – 128. (in Chinese)

[12] 刘小明,樊孝忠,等.一种结合本体和焦点的问题分类方法[J].北京理工大学学报,2012,32(5):498 – 502.
LiuXiaoming, Fan Xiaozhong, et al. A question classification method combining domain ontology and question focus[J]. Transactions of Beijing Institute of Technology, 2012, 32 (5): 498 – 502. (in Chinese)

[13] Suzuki J, et al. Hierarchical directed acyclic graph kernel: methods for structured natural language data[A]. Proc of the 41th Annual Meeting of the Association for Computational Linguistics[C]. Sapporo, Japan: Association for Computational Linguistics, 2003. 32 – 39.

[14] Wu Y Z, et al. Chinese question classification from approach and semantic views[A]. Proc of the Second Asia Information Retrieval Symposium[C]. Jeju Island, Korea: Asia Information Retrieval Society, 2005. 485 – 490.

[15] Pinchak C, et al. A probabilistic answer type model[A]. Proc of the 11th Conference of the European Chapter of the Association for Computational Linguistics[C]. Trento, Italy: Association for Computational Linguistics, 2006. 393 – 400.

[16] Mikhailian A, et al. Learning foci for question answering over topic maps[A]. Proc of the ACL-IJCNLP 2009 Conference [C]. Suntec, Singapore: Association for Computational Linguistics, Asian Federation of Natural Language Processing, 2009. 325 – 328.

[17] 张宇,刘挺,等.基于改进贝叶斯模型的问题分类[J].中文信息学报,2005,19(2):100 – 105.
Zhang Yu, Liu Ting, et al. Modified bayesian model based question classification[J]. Journal of Chinese Information Processing, 2005, 19(2): 100 – 105. (in Chinese)

[18] 文勖,张宇,等.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.
Wen Xu, Zhang Yu, et al. Syntactic structure parsing based Chinese question classification[J]. Journal of Chinese Information Processing, 2006, 20(2): 33 – 39. (in Chinese)

[19] 孙景广,蔡东风,等.基于《知网》的中文问题自动分类[J].中文信息学报,2007,21(1):90 – 96.
SunJingguang, Cai Dongfeng, et al. HowNet Based Chinese Question Automatic Classification[J]. Journal of Chinese Information Processing, 2007, 21(1): 90 – 96. (in Chinese)

[20] Blunsom P, et al. Question classification with loglinear models [A]. Proc of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. New York, USA: ACM, 2006. 615 – 616.

[21] Zhang Z, et al. Automatic recognition of focus and interrogative word in Chinese question for classification[J]. Computer and Information Science, 2010, 3(1): 168 – 174.

[22] Silva J, et al. From symbolic to sub-symbolic information in question classification [J]. Artificial Intelligence Review, 2011, 35(2): 137 – 154.

[23] Loni B, et al. Question classification with weighted combination of lexical, syntactical and semantic features[A]. Proc of the 14th International Conference of Text, Speech and Dialog [C]. Pilsen, Czech Republic: Springer, 2011. 243 – 250.

[24] Li F T, et al. Classifying what-type questions by head noun tagging[A]. Proc of the 22th International Conference on Computational Linguistics[C]. Manchester, UK: Association for Computational Ling uistics, 2008. 481 – 488.

[25] Bunescu R, et al. Towards a general model of answer typing: question focus identification[J]. Special issue: Natural Language Processing and its Applications. Research in Com-puting Science, 2010, 46: 231 – 242.

[26] Nguyen M L, et al. Subtree mining for question classification problem[A]. Proc of the 20th International Joint Conference on Artificial Intelligence[C]. Hyderabad, India: AA AI Press, 2007: 1695 – 1700.

[27] 李鑫,黄萱菁,等.基于错误驱动算法组合分类器及其在问题分类中的应用[J].计算机研究与发展,2008,45(3):535 – 541.
Li Xin, Huang Xuanjing, et al. Combined multiple classifiers based on TBL algorithm and their application in question classification[J]. Journal of Computer Research and Development, 2008, 45(3): 535 – 541. (in Chinese)

[28] Fellbaum C. WordNet: an electronic lexical database[M]. Cambridge: MIT Press, 1998.

[29] 董振东,董强.《知网》[EB/OL]. http://www.keenage.com/html/c-index.html,2014-07-16.

[30] 余正涛,樊孝忠,等.基于支持向量机的汉语问题分类[J].华南理工大学学报(自然科学版),2005,33(9):25-29.
Yu Zhengtao, Fan Xiaozhong, et al. Chinese question classifier based on support vector machine[J]. Journal of South China University of Technology(Natural Science Edition), 2005, 33(9):25-29.(in Chinese)

[31] Metzler D, et al. Analysis of statistical question classification for factbased questions[J]. Information Retrieval, 20 05, 8:481-504.

[32] Ray S K, et al. A semantic approach for question classification using WordNet and Wikipedia[J]. Pattern Recognition Letters, 2010, 31(13):1935-1943.

[33] 李方涛,张显,等.一种新的层次化结构问题分类器[J].中文信息学报,2008,22(1):93-98.
Li Fangtao, Zhang Xian, et al. A novel hierarchical structure question classifier[J]. Journal of Chinese Information Processing, 2008, 22(1):93-98.(in Chinese)

[34] 张志昌,张 宇,等.基于线索词识别和训练集扩展的中文问题分类[J].高技术通讯,2009,19(2):111-118.
Zhang Zhichang, Zhang Yu, et al. Chinese question classification based on identification of cue words and extension of training set[J]. Chinese High Technology Letters, 2009, 19(2):111-118.(in Chinese)

[35] Banerjee S, et al. An empirical study of combining multiple models in Bengali question classification[A]. Proc of the International Joint Conference on Natural Language Processing[C]. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013. 892-896.

[36] Manning C D, et al. Optimization, maxent models, and conditional estimation without magic[A]. Proc of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology[C]. Tutorial, Edmonton, Canada: Association for Computational Linguistics, 2003. 8-8.

[37] Vapnik V. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.

[38] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.
Zhang Xuegong. Introduction to statistical learning theory and support vector machines[J]. Acta Automatica Sinica, 2000, 26(1):32-42.(in Chinese)

[39] 车万翔.基于核方法的语义角色标注研究[D].哈尔滨:哈尔滨工业大学,2008.

[40] Haussler D. Convolution kernels on discrete structures[R]. Santa Cruz: Department of Computer Science, University of California in Santa Cruz, 1999.

[41] Moschitti A, et al. Exploiting syntactic and shallow semantic kernels for question answer classification[A]. Proc of the 45th Annual Meeting on Association for Computational Linguistics[C]. Prague, Czech Republic: Association for Computational Linguistics, 2007. 776-783.

[42] Liu L, et al. Chinese question classification based on ques-tion property kernel[J]. International Journal of Machine Learning and Cybernetics, 2013, 11:1-8.

[43] Moschitti A, et al. Using syntactic and semantic structural kernels for classifying definition questions in Jeopardy![A]. Proc of the 2011 Conference on Empirical Methods in Natural Language Processing[C]. Edinburgh, Scotland, UK: Asso-ciation for Computational Linguistics, 2011. 712-724.

[44] Blum A, et al. Combining labeled and unlabeled data with co-training[A]. Proc of the 11th Annual Conference on Computational Learning Theory[C]. Wisconsin, USA: ACM, 1998. 92-100.

[45] Zhou Z H, et al. Tri-training: exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11):1529-1541.

[46] Yu Z T, et al. Question classification based on cotraining style semi-supervised learning[J]. Pattern Recognition Letters, 2010, 31:1975-1980.

[47] Tri T N, et al. Using semi-supervised learning for question classification[J]. Information and Media Technologies. 2008, 3(1):112-130.

[48] Mishra T, et al. Qme!: A speech-based question-answering system on mobile devices[A]. Proc of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics[C]. Los Angeles, USA: Association for Computational Linguistics, 2010. 55 63.

[49] Marin-i-D, et al. Question classification with active learning[A]. Proc of 15th International Conference on Text, Speech and Dialogue[C]. LNCS 7499, Springer, 2012. 673-680.

[50] Su L, et al. Domain adaptation for question classification[J]. Journal of Computational Information Systems, 2011, 7(9):32 61-3267.

[51] 李鹏,王晓龙,等.一种基于粗糙集增量式规则学习的问题分类方法研究[J].电子信息学报,2008,30(5):1127-1130.
Li Peng, Wang Xiaolong, et al. Question classification with incremental rule learning algorithm based on rough set[J]. Journal of Electronics and Information Technology, 2008, 30(5):1127-1130.(in Chinese)

[52] Loni B. A survey of state-of-the-art methods on question classification[R]. Delft: TU Delft Repository, 2011.

[53] 李茹,宋小香,等.基于汉语框架网的中文问题分类[J].计算机工程与应用,2009,45(31):111-115.
Li Ru, Song Xiaoxiang, et al. Chinese question classification

based on Chinese FrameNet[J]. Computer Engineering and Applications, 2009, 45(31): 111 – 115. (in Chinese)

[54] 杨思春,高超,等.基于差异性和重要性的问句特征组合[J].电子学报,2014,42(5):918 – 924.
Yang Sichun, Gao Chao, et al. Combining Features of Question Based on Diversity and Importance[J]. Acta Electronica Sinica, 2014, 42(5): 918 – 924. (in Chinese)

[55] Collobert R, et al. Natural language processing (almost) from scratch[J]. The Jounal of Machine Learning Research, 2011, 12: 2493 – 2537.

[56] Zeng D, et al. Relation classification via convolutional deep neural network[A]. Proc of the 25th International Con ference on Computational Linguistics[C]. Dubin, Ireland: Association for Computational Linguistics, 2014. 2335 – 2344.

[57] Zhou G, et al. Exploiting bilingual translation for question retrieval in community-based question answering[A]. Proc of the 23th International Conference on Computational Ling-uistics[C]. Mumbai, India: Association for Computational Linguistics, 2012. 3153 – 3170.

[58] Zhou G, et al. Group non-negative matrix factorization with natural categories for question retrieval in community question answer archives[A]. Proc of the 25th International Conference on Computational Linguistics[C]. Dublin, Ireland: Association for Computational Linguistics, 2014. 89 – 98.

[59] Singh A, et al. CQC: Classifying questions in CQA websites [A]. Proc of the 20th ACM International Conference on Information and Knowledge Management[C]. Glasgow, Scotland, UK: ACM, 2011. 2033 – 2036.

[60] Chan W, et al. Community answer summarization for multi-sentence question with group L1 regularization[A]. Proc of the 50th Annual Meeting of the Association for Computational Linguistics[C]. Jeju, Korea: Association for Computational Linguistics, 2012. 582 – 591.

[61] Bao J, et al. Knowledge-based question answering as machine translation[A]. Proc of the 52th Annual Meeting of the Association for Computational Linguistics[C]. Baltimore, Maryland: Association for Computational Linguistics, 2014. 967 – 976.

[62] Zhao S, et al. Tailor knowledge graph for query under-standing: linking intent topics by propagation[A]. Proc of the 2014 Conference on Empirical Methods in Natural Language Processing[C]. Doha, Qatar: Association for Computational Linguistics, 2014. 1070 – 1080.

作者简介

杨思春 男,1970 年生于安徽六安.博士,副教授.研究方向为自然语言处理、自动问答.
E-mail: yangsc@nlp.nju.edu.cn

戴新宇(通信作者) 男,1979 年生于江苏盱眙.博士,副教授.研究方向为自然语言处理、机器翻译.
E-mail: daixinyu@nlp.nju.edu.cn