基于差异性和重要性的问句特征组合

杨思春[1,2],高 超[3],戴新宇[1],尹存燕[1],陈家骏[1]
(1.南京大学计算机软件新技术国家重点实验室,江苏南京 210023;2.安徽工业大学计算机学院,安徽马鞍山 243002;
3.安徽工程大学机电学院,安徽芜湖 241000)

摘 要: 在问答系统问句分类研究中,对问句特征进行组合有助于构造高效的问句分类器.针对当前问句分类中的特征组合问题,提出一种基于差异性和重要性的特征组合(Diversity and Importance based Feature Combination,DIFC)方法.通过计算待组合特征与当前特征组合的错分差异度和正分差异度,以及待组合特征本身的重要度,从候选特征集中动态获取优化的特征组合.在哈工大中文问句集上对词袋绑定特征进行组合的实验结果表明,与其他特征组合方法相比,DIFC方法灵活高效,准确率更高.

关键词: 问句分类;特征组合;差异性;重要性
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2014)05-0918-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.05.013

Combining Features of Question Based on Diversity and Importance

YANG Si-chun[1,2], GAO Chao[3], DAI Xin-yu[1], YIN Cun-yan[1], CHEN Jia-jun[1]
(1. *State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu* 210023, *China*; 2. *School of Computer Science, Anhui University of Technology, Maanshan, Anhui* 243002, *China*; 3. *College of Mechanical & Electrical Engineering, Anhui Polytechnic University, Wuhu, Anhui* 241000, *China*)

Abstract: In research on question classification in question answering system, combining features can greatly help construct efficient question classifier. In order to deal with the problem of low performance of existing methods, a new method of diversity and importance based feature combination(DIFC) is proposed. By calculating the diversity between candidate feature and current combination for error and correct classification respectively, and the importance of candidate feature, features can be dynamically selected from candidate feature set. The experimental results of bag-of-words binding features on the HIT Chinese question set show that, compared with other methods, the new method is flexible and efficient, and gets more optimal feature combination.

Key words: question answering system; question classification; feature combination; diversity; importance

1 引言

自动问答(Question Answering,QA)[1]是当前自然语言处理和信息检索领域的一个研究热点.问答系统通常包括问句分析、文档检索和答案抽取等三个关键处理模块[2,3].作为问句分析的第一步,问句分类(Question Classification)通过确定问句的期望答案语义类别,有效地过滤不相关的候选答案,进而提升问答系统的性能.问句分类可以看成是一种特殊的文本分类,但是,由于问句中所包含的词汇信息很少,因此需要对问句作更深层次的句法、语义分析才能获得较高的分类精度.

当前问句分类研究大多基于机器学习的方法,获取丰富的特征信息有助于构造高效的问句分类器[4~8].在利用各种特征训练问句分类器时,现有文献大多基于单个特征本身的贡献对这些特征进行递进式组合.我们将这种方法称为基于重要性的特征组合(Importance based Feature Combination,IFC).文献[8]提出一种启发式的特征组合(Heuristic Feature Combination,HFC)算法,但是,HFC在每轮组合过程中均以分类精度来衡量所有候选特征的预组合效果,在特征数量较大的情况下需要花费很长的时间.因此,如何以较高的效率和准确率实现问句特征组合,对于进一步提升当前问句分类的性能具有

收稿日期:2013-04-17;修回日期:2013-08-17;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61170181);江苏省自然科学基金(No.BK2011192);安徽省高校省级自然科学研究重点项目(No.KJ2011A048)

重要的意义.

文献[9]对分类器集成中的分类器选择方法进行了比较和分析,实际上,当每个分类器只含有一个特征时,可以把分类器的选择看成是特征组合过程中的特征子集选择.为此,本文借鉴文献[10]中的分类器互补性度量,引入错分差异度、正分差异度、重要度等概念,给出一种基于差异性和重要性的特征组合(Diversity and Importance based Feature Combination, DIFC)方法.通过定量计算待组合特征与当前特征组合之间的差异性,以及待组合特征本身的重要性,从候选问句特征集中动态获取优化的特征组合.实验结果表明,与 IFC 等其他特征组合方法相比,所提出的方法灵活高效,准确率更高.

2　特征提取

2.1　基本特征

我们利用 LTP 平台(http://ir.hit.edu.cn/demo/ltp)对中文问句进行句法语义分析,并提取词袋(Bag-of-Words, BOW),词性(Part-of-Speech, POS),词义(Word Sense, WS),命名实体(Named Entity, NE),依存关系(Dependency Relation, DR)以及核心词(Core Word, CW)等基本特征.

图 1 是问句"中国哪一条河流经过的省份最多?"的分析结果,所提取的 BOW 特征为{中国,哪,一,条,河流,经过,的,省份,最,多,?},POS 特征为{ns,r,m,q,n,p,u,n,d,a,wp},NE 特征为{(中国,Ns,S),(一,Nm,B),(条,Nm,E)},WS 和 WS'特征分别为{Di02,Ka35,Dn04,Dn08,Be05,Kb 06,Kd01,Di02,Ka02,Dn05,-1}和{Di,Ka,Dn,Be,Kb,Kd,Di,-1}(WS 和 WS'分别是 3 层和 2 层词义编码),DR 特征为{ATT, QUN, DE, SBV, ADV, HED},CW 特征为{条,河流,的,省份,多}.

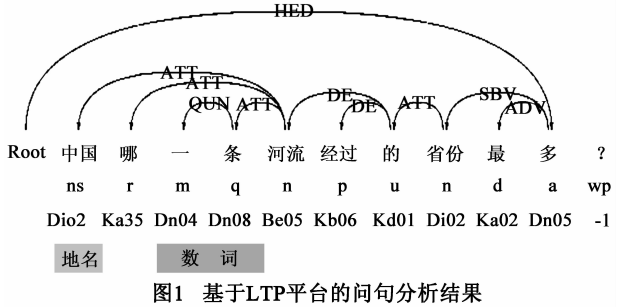


图1　基于LTP平台的问句分析结果

2.2　词袋绑定特征

文献[8]引入词袋绑定操作将词性、词义等基本特征分别对应到每个特定的词,进而生成一类新的问句特征.这样做的好处是可以利用所获取特征与词汇之间的语义联系,以很小的额外处理开销获得一类潜在的问句特征.本文在其基础上进一步将多个基本特征同时对应到每个特定的词,以形成多重词袋绑定(Multiple Bag-of-Words Binding, MBWB)特征.另外,我们还借鉴文献[7]的做法,将问句中的主、谓、宾、疑问词及其附属成分定义为主干词词袋(Trunk_BOW, T_BOW)(前面的词袋特征称为全词词袋(All_BOW, A_BOW),并通过将多个基本特征同时绑定到 T_BOW,以形成 T_MBWB 特征(前面的 MBWB 特征记为 A_MBWB 特征).我们假定每一特征对分类起正或负作用(即可以提高或者降低分类精度),并采用类似于图的深度搜索方法来生成 MBWB 特征.从基本的词袋特征开始,不断绑定新的特征,如果绑定某个特征以后导致分类精度下降,就回溯到上一次绑定的特征.具体的过程如下:

步骤 1　将每个基本特征 Fi 分别与 A_BOW(或 T_BOW)进行绑定,以形成 1 重绑定特征 A/Fi(或 T/Fi).

步骤 2　若 1 重绑定特征中精度最高的 A/Fj(或 T/Fj)高于 A_BOW(或 T_BOW)的精度,则将除 Fj 以外的其他基本特征与 A/Fj(或 T/Fj)绑定,以形成 2 重绑定特征.

步骤 3　若 2 重绑定特征中精度最高的 A/Fj/Fk(或 T/Fj/Fk)高于 A/Fj(或 T/Fj)的精度,则将除 Fj、Fk 以外的其他基本特征与 A/Fj/Fk(或 T/Fj/Fk)绑定,以形成 3 重绑定特征.

步骤 4　如此反复,直至 $n+1$ 重绑定特征中最高精度不高于 n 重绑定特征中最高精度.

3　基于差异性和重要性的特征组合

为了充分利用上述词袋绑定特征,本文通过定量计算待组合特征与当前特征组合之间的差异性,以及待组合特征本身的重要性,在基本特征的基础上进一步组合词袋绑定特征.

3.1　相关定义

本文借鉴文献[10]中的分类器互补性定义,来度量待组合特征之间的差异性,并将差异性的定义由仅考虑样本集被错误分类时的差异(以下称错分差异),扩展为同时考虑样本集被正确分类时的差异(以下称正分差异).这是因为,实际分类时不仅两个特征的错误分类样本有区别,而且正确分类样本也有区别.对于错分差异,通常认为差异越大则互补性越强,而对于正分差异,差异越小则互补性越强.具体定义如下:

设待组合特征 F_i 与当前特征组合 C_k,非空样本集 S 被 F_i 和 C_k 错误分类的样本集分别为 S_i^E 和 S_k^E,非空样本集 S 被 F_i 和 C_k 正确分类的样本集分别为 S_i^C 和 S_k^C,则 F_i 和 C_k 在样本集 S 上的错分差异度 err_div(F_i, C_k)和正分差异度 cor_div(F_i, C_k)分别定义为式(1)和式(2):

$$err\_div(F_i, C_k) = \frac{|S_i^E \cup S_k^E| - |S_i^E \cap S_k^E|}{|S_i^E \cup S_k^E|} \quad (1)$$

$$cor\_div(F_i, Ck) = \frac{|S_i^C \cup S_k^C| - |S_i^C \cap S_k^C|}{|S_i^C \cup S_k^C|} \quad (2)$$

与文献[10]中的定义不同的是,式(1)、式(2)的分母分别为$|S_i^E \cup S_k^E|$和$|S_i^C \cup S_k^C|$(而不是$|S|$).这里,主要为了区分错分差异和正分差异(当式(1)、式(2)的分母均取$|S|$时,实际上两式完全相同)

本文将错分差异度和正分差异度的差值定义为总体差异度.设待组合特征 F_i 与当前特征组合 C_k 的错分差异度为 $err\_div(F_i, C_k)$,正分差异度为 $cor\_div(F_i, C_k)$,则 F_i 和 C_k 在样本集 S 上的总体差异度 $div(F_i, C_k)$定义为:

$$div(F_i, C_k) = |err\_div(F_i, C_k) - cor\_div(F_i, C_k)| \quad (3)$$

总体差异度越大,则特征之间的互补性越强;反之,则互补性越弱.

文献[10]在定义分类器的互补性时仅考虑了分类器的差异性,本文在定义特征的互补性时还考虑了待组合特征本身的重要性.我们认为:虽然差异性是需要考虑的一个主要因素,但重要性也同样不可忽视.本文以待组合特征本身的分类精度来度量其重要性,即 F_i 相对于当前特征组合 C_k 的重要度 $imp(F_i)$定义为:

$$imp(F_i) = \frac{|S_i^E|}{|S|} \quad (4)$$

在分别计算待组合特征 F_i 与当前特征组合 C_k 在样本集 S 上的总体差异度 $div(F_i, C_k)$,以及待组合特征 F_i 在样本集 S 上的重要度 $imp(F_i)$以后,我们将 F_i 和 C_k 在样本集 S 上的互补度 $com(F_i, C_k)$定义为 $div(F_i, C_k)$和 $imp(F_i)$的加权平均:

$$com(F_i, C_k) = \frac{k_1 * div(F_i, C_k) + k_2 * imp(F_i)}{k_1 + k_2} \quad (5)$$

这里,$k1$、$k2$ 分别为差异度和重要度的权重参数.

3.2 DIFC 特征组合算法

根据上述定义,我们提出一种基于差异性和重要性的组合算法(如算法 1 所示).首先计算侯选特征与初始特征组合的互补度,选择互补度最高的侯选特征;若加入该选特征以后所得特征组合的精度增加,则计算去除该特征以后的侯选特征集合中每个侯选征与当前特征组合的互补度,并选择互补度最高的侯选特征;重复上述步骤,直到当前特征组合的精度不再增加为止.

算法 1 DIFC 特征组合算法

Input:候选特征集合 F

Output:最优或次优特征组合 C

1) $F = \{f_0\ f_1, \cdots, f_n\}$; $C = \phi$;

2)将候选特征集合 F 中选取分类精度最高的特征记作 f_0;

3) $F = F\{f_0\}$; $C = C \cup \{f_0\}$

4)计算 F 中每个特征与当前特征组合 C 的互补度,将互补度最大的特征记作 f_1;

5) $C = C \cup \{f_1\}$;

6)若 C 的分类精度小于 $C - f_1|$ 的分类精度,则 $C = C - \{f_1\}$,算法结束;

7)否则 $F = F - \{f_1\}$;

8)重复 4) - 7).

与基于重要性的特征组合方法相比,由于同时考虑了特征之间的差异性以及特征本身的重要性,该算法可以确保获得更加优化的特征组合;与文献[8]中的特征组合方法相比,该算法每次循环时不再依据实际分类精度,而是借助差异度、重要度等定量计算对每个候选特征进行初步筛选,因而显著提高了效率;另外,该算法在每轮的候选特征筛选以后,还实际验证所得特征组合的分类精度,以确保所得特征组合具有较高的分类性能.

4 实验结果与分析

4.1 实验数据与评价方法

本文选用哈工大社会计算和信息检索研究中心提供的中文问句集作为实验数据,共有 6266 个问句(4966 个为训练集,1300 个为测试集),共分为 6 个大类和 77 个小类.

我们以 Liblinear - 1.4(http://www.csie.ntu.edu.tw/~cjlin/liblinear/)作为分类器,在问句集(训练集 + 测试集)上做 10 倍交叉验证来评价问句分类的性能.实验硬件配置为 1.73GHz CPU,0.99 G 内存和 120G 硬盘,开发环境为 Komodo IDE 8.0.2 - 78971 和 python 3.2.

4.2 基本特征和词袋绑定特征

本文以 A_BOW、POS、NE、WS、WS'、DR、CW 以及 T_BOW 作为基本特征,并按照 2.2 中的绑定方法自动生成 A_MBWB 特征和 T_MBWB 特征.表 1、表 2 给出所有的 A_MBW B 特征和 T_MBWB 特征的分类精度.

可以看出,绝大部分词袋绑定特征(特别是对应 POS、NE 和 DR 的)的精度都比原来的对应基本特征有较大幅度的提高,这说明词袋绑定特征对于问句分类也具有重要的贡献.

4.3 DIFC 特征组合

为初步验证算法 1 的有效性,我们在所有基本特征(以下记为 Base)的基础上,分别组合单个的词袋绑定特征.表 3、表 4 给出了组合单个词袋绑定特征时的分类精度、差异度、重要度以及互补度(求互补度时参数 $k1 = k2 = 5$).

表 1　A_MBWB 特征的分类精度

待绑定特征	分类精度/%					
	POS(17.8104)	NE(11.8098)	WS(64.0919)	WS'(39.1318)	DR(11.8417)	CW(67.8264)
A_BOW(80.1947)						
A/	80.1149	80.45	79.9394	80.067	76.189	73.9068
A/NE/	80.5298		80.2585	80.4022	76.3645	64.826
A/NE/POS/			80.1628	80.3383	76.157	64.9697

表 2　T_MBWB 特征的分类精度

待绑定特征	分类精度/%					
	POS(17.8104)	NE(11.8098)	WS(64.0919)	WS'(39.1318)	DR(11.8417)	CW(67.8264)
T_BOW(72.9492)						
T/	72.8375	73.2844	72.4386	72.5662	69.135	58.8094
T/NE/	73.3323		72.9173	72.9812	70.0128	59.2563
T/NE/POS/			72.9652	73.0131	69.917	59.2882

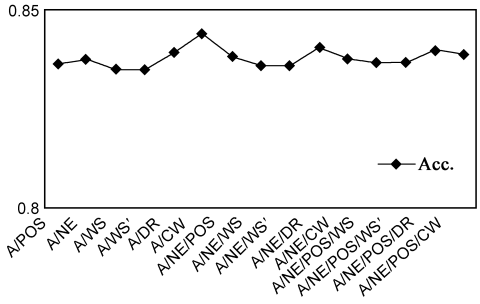


图2　Base+A_MBWB的精度变化

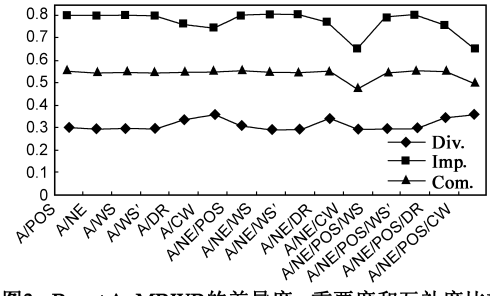


图3　Base+A_MBWB的差异度、重要度和互补度比较

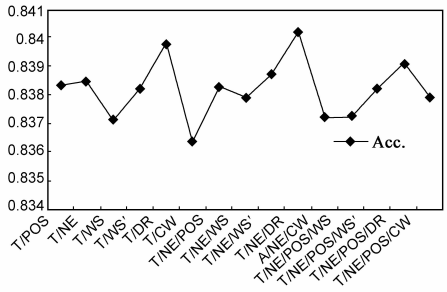


图4　Base+T_MBWB的精度变化

从表 3、表 4 可以看出,在 Base 基础上分别组合单个 A_MBWB、T_MBWB 特征以后,所得分类精度均获得了一定程度的提高.

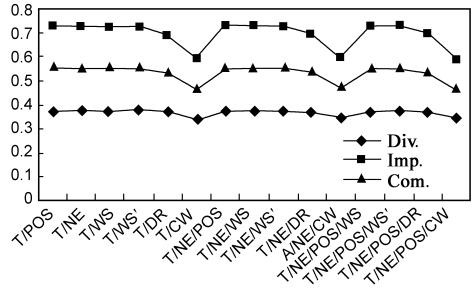


图5　Base+T_MBWB的差异度、重要度和互补度比较

图 2 ~ 图 3、图 4 ~ 图 5 分别对 Base 与单个 A_MBWB、T_MBWB 特征组合时的分类精度与对应的差异度、重要度以及互补度进行了比较.

从图 2 ~ 图 3、图 4 ~ 图 5 可以看出,整体上“Div.”与“Acc.”的变化趋势很接近,而“Imp.”与“Acc.”则差别很大.这表明特征组合时需要首先考虑特征之间的差异性.但是,“Div.”与“Acc.只是整体上很接近,其中也会出现个别“异常点”.例如在图 3 中,“Acc.”在“A/NE/POS/CW”位置的取值低于前一位置的取值,而“Div.”在“A/NE/POS/CW”位置的取值却高于前一位置的取值.为了对这个“异常点”进行修正,在同时考虑重要性以后,发现“Div.”与“Acc.”在这两个位置的变化趋势完全一致,而且相对于“Div.”,“Com.”总体上与“Acc.”也更加一致.

表 5 给出在 base 基础上组合多个词袋绑定特征的实验结果.“Ci,A/”列表示当前特征组合 Ci 与待组合的 A_MBWB 特征的互补度,加粗表示本轮的最高互补度,下划线表示预组合精度低于当前组合精度,“Ci,T/”列表示 Ci 与待组合的 T_MBWB 特征的互补度,“/POS”等

行表示待组合的 A_MBWB 或 T_MBWB 特征,“Acc.(%)”行表示互补度最高的特征组合的精度.

可以看出,经过 8 次循环以后所得特征组合的分类精度达到 84.73%.比 Base 高出 1.31 个百分点,比表 3、表 4 中的最高精度也高出 0.35 个百分点,进一步说明了 DIFC 方法的有效性.

表 3 Base 与单个 A_MBWB 特征进行组合

Base(83.42%)+	Acc.(%)	Div.	Imp.	Com.
A/POS	83.63	0.304	0.801	0.553
A/NE	83.74	0.294	0.805	0.549
A/WS	83.48	0.295	0.799	0.547
A/WS′	83.47	0.295	0.801	0.548
A/DR	83.89	0.335	0.762	0.549
A/CW	84.38	0.362	0.739	0.550
A/NE/POS	83.8	0.302	0.805	0.554
A/NE/WS	83.58	0.292	0.803	0.547
A/NE/WS′	83.56	0.295	0.804	0.549
A/NE/DR	84.04	0.341	0.764	0.552
A/NE/CW	83.75	0.292	0.648	0.470
A/NE/POS/WS	83.66	0.297	0.802	0.549
A/NE/POS/WS′	83.66	0.299	0.803	0.551
A/NE/POS/DR	83.98	0.340	0.762	0.551
A/NE/POS/CW	83.85	0.361	0.649	0.505

4.4 参数对算法性能的影响

DIFC 算法的性能可能会随公式(5)中参数 $k1$ 和 $k2$ 的影响,为此,取值多组($k1$,$k2$)进行实验.首先保持 $k2$ 不变,逐渐增加 $k1$,直至其与 $k2$ 相等;然后保持 $k1$ 不变,逐渐减少 $k2$.表 6 从循环次数、预组合次数、最终组合的精度等方面比较了 $k1$、$k2$ 对 DIFC 算法性能的影响.

表 4 Base 与单个 T_MBWB 特征进行组合

Base(83.42%)+	Acc.(%)	Div.	Imp.	Com.
T/POS	83.83	0.378	0.728	0.553
T/NE	83.85	0.375	0.733	0.554
T/WS	83.71	0.378	0.724	0.551
T/WS′	83.82	0.377	0.726	0.551
T/DR	83.98	0.372	0.691	0.532
T/CW	83.63	0.338	0.588	0.463
T/NE/POS	83.83	0.376	0.733	0.555
T/NE/WS	83.79	0.375	0.729	0.552
T/NE/WS′	83.87	0.376	0.729	0.553
T/NE/DR	84.02	0.373	0.700	0.537
T/NE/CW	83.72	0.341	0.593	0.467
T/NE/POS/WS	83.72	0.375	0.729	0.552
T/NE/POS/WS′	83.82	0.375	0.730	0.552
T/NE/POS/DR	83.91	0.372	0.699	0.536
T/NE/POS/CW	83.79	0.341	0.593	0.467

表 5 Base 与多个 A_MBWB 特征、T_MBWB 特征进行组合($k1=5$,$k2=5$)

	C0,A/	C1,A/	C2,A/	C2,T/	C3,T/	C3,A/	C4,A/	C4,T/
/POS	0.553	0.546	**0.555**	0.556	0.553	**0.559**		0.554
/NE	0.549	0.541	0.549	**0.559**		0.554	0.549	
/WS	0.547	0.543	0.551	0.555	0.552	0.556	0.549	0.553
/WS’	0.548	0.543	0.551	0.555	0.552	0.556	0.550	0.553
/DR	0.549	0.551	0.552	0.537	0.536	0.551	0.549	0.536
/CW	0.550	**0.553**		0.469	0.469			0.469
/NE/POS	**0.554**			0.558	**0.554**			0.469
/NE/WS	0.547	0.541	0.547	0.556	0.553	0.552	0.548	**0.555**
/NE/WS’	0.549	0.543	0.548	0.557	0.553	0.554	0.549	0.554
/NE/DR	0.552	0.552	0.553	0.542	0.540	0.552	**0.552**	0.541
/NE/CW	0.470	0.507	0.509	0.473	0.472	0.509	0.510	0.473
/NE/POS/WS	0.549	0.541	0.547	0.556	0.553	0.552	0.547	0.554
/NE/POS/WS’	0.551	0.542	0.548	0.556	0.553	0.554	0.549	0.554
/NE/POS/DR	0.551	0.550	0.552	0.541	0.539	0.551	0.550	0.539
/NE/POS/CW	0.505	0.507	0.510	0.473	0.473	0.511	0.511	0.474
Acc.(%)	83.80	84.58	84.52	84.68	84.47	**84.73**	84.34	84.41

表 6　权重 $k1$,$k2$ 对 DIFC 算法的影响

$k1$,$k2$ 取值	循环次数	预组合次数	最终组合的精度(%)
$k1=3, k2=5$	9	9	84.07
$k1=4, k2=5$	6	6	84.39
$k1=5, k2=5$	8	8	84.73
$k1=5, k2=4$	8	8	84.73
$k1=5, k2=3$	6	6	84.49

可以看出,随着 $k1$ 的增加,所得特征组合的精度也在增加,当($k1$,$k2$)取值(5,4)时所得特征组合取得了 84.73% 的最高分类精度.但是,随着 k 的进一步增加,当($k1$,$k2$)取值(5,3)时所得特征组合的精度又开始下降.这说明总体上差异性比重要性更加有利于描述特征的互补性,但是当 $k1$ 增加到一定程度以后,又会因为差异性过分消弱了重要性的地位而导致特征之间的互补性变弱.因此,实际应用该算法时,需要根据实验结果设置合适的参数 $k1$ 和 $k2$.

4.5　数据集大小对算法性能的影响

为了验证数据集大小对算法性能的影响,我们将原训练样本和测试样本的数量均减少 10%,然后按照 DIFC 算法对问句特征进行组合.这里,权重参数取 $k1=k2=5$.

实验结果表明,训练样本和测试样本的数量均减少了 10% 以后,虽然所得特征组合的最高精度有所降低(为 84.5485%),但与组合单个 A_MBWB 特征或 T_MBWB 特征的精度相比,提高了 0.73 到 1.18 个百分点,这说明本文方法在样本减少的情况下仍然具有优越性.

4.6　与其他特征组合方法的比较

为进一步验证 DIFC 的有效性,我们将 DIFC($k1=5, k2=4$)与 IFC、基于差异性的特征组合(Diversity based Feature Combination,DFC)进行对比,结果如表 7 所示.这里的 DFC 分为仅考虑错分差异的 DFC、仅考虑正分差异的 DFC 以及同时考虑错分差异和正分差异的 DFC.

表 7　DIFC 与 IFC、DFC 的性能比较

方法	循环次数	预组合次数	最终组合的精度(%)
DIFC($k1=5, k2=4$)	8	8	84.73
IFC	6	6	84.12
DFC(仅考虑错分差异)	3	3	83.75
DFC(仅考虑正分差异)	5	5	84.20
DFC	6	6	84.46

从最终组合的精度看,DIFC 均优于 IFC 和 DFC,说明同时考虑差异性和重要性更加有助于描述特征的互补性;对三种类型 DFC,前面两个实现效率较高,但从最终组合的精度看,它们均低于同时考虑错分差异和正分差异的 DFC,这说明在特征的互补性描述方面,综合考虑错分差异和正分差异要优于单纯考虑错分差异或正分差异;另外,仅考虑错分差异的 DFC 也低于仅考虑正分差异的 DFC,说明在特征的互补性描述方面正分差异要优于错分差异.

本文还实验了文献[8]中的 HFC 方法,实验结果表明虽然 HFC 获得了与 DIFC 同样的最高精度,但由于每次循环均需要对所有的候选特征进行预组合训练,5 轮组合共进行了 30 + 29 + 28 + 27 + 26 = 140 次预组合,因此执行效率远远低于 DIFC.

5　结论

针对当前问句分类中的特征组合问题,提出一种基于差异性和重要性的特征组合方法.主要工作在于:(1)借鉴文献[10]中分类器的差异性来度量特征之间的差异性,并将差异性的定义扩展为同时考虑错分差异和正分差异;(2)在定义特征之间的互补性时,同时考虑了特征之间的差异性以及特征本身的重要性;(3)提出一种基于差异性和重要性的的特征组合(DIFC)算法,以实现从候选特征集中动态获取优化的特征组合.

DIFC 算法采取的是一种序列前向选择(SFS)搜索策略,本质上属于贪心算法,因此容易陷入局部最优.下一步将尝试采用遗传算法等其他优化方法,避免 DIFC 算法陷入局部最优.

致谢:

感谢审稿专家对本文提出的修改意见,感谢哈工大社会计算和信息检索研究中心提供的相关处理模块和语料资源!

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):1131 – 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] Li X, et al. Learning question classifiers[A]. Proceedings of the

19th International Conference on Computational Linguistics (COLING2002)[C]. Taipei: Association for Computational Linguistics, 2002.1 – 7.

[5] Zhang D, et al. Question classification using support vector machines[A]. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)[C]. Toronto, Canada: ACM, 2003.26 – 32.

[6] Huang Z H, et al. Investigation of question classifier in question answering[A]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP2009)[C]. Singapore: Association for Computational Linguistics, 2009.543 – 550.

[7] 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)

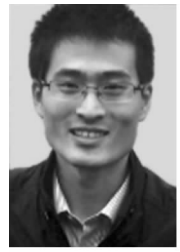
[8] 杨思春,高超,等.融合基本特征和词袋绑定特征的问句特征模型[J].中文信息学报,2012,27(5):46 – 52.
Yang Sichun, Gao Chao, et al. A feature model integrating basic features and bag-of-words binding features[J]. Journal of Chinese Information Processing, 2012, 27(5): 46 – 52. (in Chinese)

[9] Ruta D, et al. Classifier selection for majority voting[J]. Information Fusion, 2005, 6(1): 63 – 81.

[10] 郝红卫,王志彬,等.分类器的动态选择与循环集成方法[J].自动化学报,2011,37(11):1290 – 1295.
Hao Hongwei, Wang Zhibin, et al. Dynamic selection and circulating combination for multiple classifier systems[J]. Acta Automatica Sinica, 2011, 37(11): 1290 – 1295. (in Chinese)

作者简介

杨思春 男,1970 年生于安徽六安.博士生,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



高 超 男,1986 年生于河南开封.助教,研究方向为自然语言处理、自动问答.

陈家骏(通信作者) 男,1963 年生于南京,教授,博士生导师,主要研究方向为自然语言处理、机器翻译、软件工程.