基于多阶段和分层方法的言语行为分类研究*

李嘉[1,2,3]，张朋柱[2]，李欣苗[4]
(1. 华东理工大学商学院，上海 200237；
2. 上海交通大学安泰经济与管理学院，上海 200052；
3. 中国航空无线电电子研究所，上海 200233；
4. 上海财经大学信息管理与工程学院，上海 200433)

摘　要　在面向在线研讨的言语行为分类研究中，前指发言类型(即与当前发言形成回复关系的前一条发言的言语行为类别)是一个非常重要的特征。但是由于在测试集上前指发言类型是未知的，因此如何合理利用前指发言的类型信息就成为一个非常关键的问题。本文以E-learning语料为例，证实了发言类型的高低层次和分类器的运行顺序对分类效果都有较为显著的影响，并提出了一个基于多阶段分层的分类方法，可以给出一个合理的前指发言层次和分类器运行顺序。在盲测集上的运行结果证实了这组优选的参数能够稳定一致地提高E-learning语料上各言语行为类别的分类效果。

关键词　多阶段，分层体系，言语行为，基于转换的学习，支持向量机
中图分类号　TP311.13

1　背景

在面向在线研讨的言语行为分类研究中，许多学者认为前指发言类型(即与当前发言形成回复关系的前一条发言的言语行为类别，previous SA)是一个非常重要的特征[1,2]。例如，如果前一条发言的言语行为是问题，那么下一个发言的言语行为是回答的可能性就很大。但是由于在测试集上前指发言类型是未知的，因此如何合理利用前指发言的类型信息就成为一个非常关键的问题。

一个值得关注的现象是很多言语行为分类体系都是分层的，比如Gerassimenko等发展的Estonian Dialogue Corpus (EDiC)[3]和Su Nam Kim发展的针对技术论坛的言语行为分类体系[4]。在分层体系中，一个高层言语行为类别可能会包含若干个低层言语行为类别。例如，Question可能进一步包括wh-question、how-question、yes-no-question等几种类型。通常高层类型的分类准确率要比低层类别高，这是因为高层类别包含更少的类型数量，因此错分的概率会降低。因此，本文关注的第一个问题是，通过引入分层的言语行为分类体系，让前指发言的类型依赖于高层类型(意味着更加抽象的类型信息和更加准确的分类结果)是否会比低层类型(意味着更加具体的类型信息但是更加不准确的分类结果)更好。

* 基金项目：国家自然科学基金青年项目“面向在线群体研讨的自动化辅助方法研究”(项目编号：71001038)，中央高校基本科研业务费专项资金资助“群体研讨文本的自动摘要与研讨态势可视化研究”(项目编号：WN1022003)，2010年上海市“两新”组织和社会建设调研课题“主动把握社会问题舆情态势专题研究：基于信息检索和自然语言处理的方法”和国家自然科学基金青年项目“面向任务的开放式团队创新协同理论与方法研究”(项目编号：71001059)。
通信作者：李嘉，男，1980年12月生，汉族，湖南湘乡人，讲师(博士后)。E-mail：jiali@ecust.edu.cn。

现在很多机器学习算法(如支持向量机)本质上都是一个二元分类器,因此需要为每一个类型分别构造一个分类器,而不能同时标记所有的类型。由于分类器必须串行工作,因此前一个分类器输出结果的精度就会影响第二个分类器的分类效果。另外,各类别对其他类别的依赖程度也不一样,无依赖的分类器应该先于有依赖的分类器运行。因此本文关注的第二个问题是分类器的运行顺序问题。

综上所述,本文研究前指发言类型信息应用于言语行为分类时遇到的两个关键问题如下。

Q1:前指发言类型的高低层次是否对分类结果有影响?如果有,如何优选?

Q2:分类器的运行顺序是否对分类结果有影响?如果是,如何优选分类器运行顺序?

为了回答这两个研究问题,本文首先以一个 E-learning 语料为例构造了一个分层的言语行为分类体系,然后设计了一组实验,来检验前指发言类型的高低层次和分类器的运行顺序对分类效果的影响,并进一步提出了一种基于多阶段分层的方法来确定合适的前指发言层次和分类器运行顺序。最后,在盲测集上检验了所提方法的效果。

2 相关研究

2.1 分层的言语行为分类体系

有很多言语行为分类体系都具有分层的特性。例如 Gerassimenko 等发展的 Estonian Dialogue Corpus (EDiC)[3] 包括两层标注,高层说明发言的一般类型(ritual, questions/answers, directive, additional information, repair 等),低层则在更多的细节上描述发言。每个高层类型可能有好几个子类型,例如 rituals 可以进一步分成 greeting, thanking, apologizing 等,questions/answers 则可以进一步分成 wh-questions, open and closed yes/no questions, refusal to answer, yes/no answers 等。在低层标注中,总共有 126 个标签,各类别在大小上差别很大。

Su Nam Kim[4] 发展了针对技术论坛的言语行为分类体系。这个类别包括两个大的分类(QUESTION, ANSWER)和 3 个单独类(RESOLUTION, REPRODUCTION 和 OTHER)。进一步的,QUESTION 又包含 4 个子类(QUESTION, ADD, CONFIRMATION 和 CORRECTION),而 ANSWER 进一步包含 5 个子类(ANSWER, ADD, CONFIRMATION, CORRECTION 和 OBJECTION)。除此之外,Joty 等[5] 也发展了一套描述网络论坛的言语行为分类体系,包括陈述(statement)、礼貌用语(polite mechanism)、一般疑问(yes-no question)、行为动机(action motivator)、特殊疑问(wh-question)、接收回复(accept response)、开放式问题(open-ended question)、致谢(acknowledge and appreciate)、从句问题(or-clause question)、拒绝回复(reject response)、不确定回复(uncertain response)和口头提问(rhetorical question)共 12 个类别。

2.2 基于转换的学习

基于转换的学习(transformation-based learning, TBL)首先由 Brill[6] 引入。由于它克服了传统人工获取规则的局限,自动地从训练语料库中学习反映语言学知识规则,因此迅速在中文分词[7]、实体识别[8] 和停顿指数预测[9] 中得到应用并获得成功。TBL 算法基于一个规则集,通过将这个规则集依次应用于数据来将一些标签转换为另一些标签。

规则是由有监管训练的方式得到的。给定一个标记过的训练语料,首先从模板产生所有可能的

规则,然后用迭代的方式来选择最高:每一轮中那些能够使精度提高最快的规则被选中。这个过程一直持续到某一个停止条件得到满足,而最常见的停止规则是引入任何新的规则对目标函数都没有明显的改善。

Samuel 等在 1998 年将基于转换学习的方法用于解决会话行为标记(dialogue act tagging)问题[10]。Samuel 等使用的是在 Reithinger 和 Klesen(1997)[11]的研究中所用过的语料,定义了 18 种不同的会话行为(dialogue act),训练集包括 143 个对话共 2701 条发言,测试集包括 20 个对话共 328 条发言。为了提高算法的效率,他们提出用蒙特卡罗法来随机选择规则,并取得了良好的效果。Samuel 等报告在测试集上的分类准确率为 75.12%,最高准确率为 77.44%。进一步的,Samuel 等[10]将结合蒙特卡洛优化的 TBL 算法应用于 VerbMobil 语料来识别言语行为类别。除了常规使用的 TBL 特征(如临近的发言和言语行为类别)外,他们还是用发言者信息、标点信息、特征词等作为特征。他们取得了 75%的精度(precision)。

2.3 支持向量机

支持向量机[12,13]是一种比较好的实现了结构风险最小化思想的方法,它的机器学习策略是保持经验风险值固定而最小化致信范围。支持向量机通过核函数将向量映射到一个更高维的空间,在这个空间里建立有一个最大间隔超平面来将两类样本点分开。

Ravi 和 Kim[14]使用了支持向量机对 PedaBot 语料进行了言语行为分类。他们使用的类别比较简单,只包含问题类(QC)和回答类(AC),在 QC 和 AC 上分别获得了 83%和 73%的准确率。Cohen 等[15]进行了根据发送者的意图(如请求开会、传递信息等)对邮件进行言语行为分类的研究,并且对比了机器学习方法。研究结果显示,对于 E-mail 分类问题,SVM(支持向量机)在 Proposal、Commitment 和 Meet 类别上效果最好,而 DT(决策树)在 Request、Directive 和 Commissive 类别上效果更好。类似的,Surendran 和 Levow[16]结合线性支持向量机和隐马尔可夫模型在 MapTask 语料上进行言语行为分类,并且获得了比前人研究更好的结果。

3 言语行为分类体系

为了充分利用分层言语行为分类的优势,本文以 E-learning 语料为例构造一个包含高层和低层两个层次的言语行为分类体系。本案例所使用的语料数据来自一个叫做 PedaBot 的在线研讨系统,在这个研讨系统内学生可以讨论与课堂内容相关的技术问题。PedaBot 系统是 phpBB 系统的一个变种,用户登入 PedaBot 系统后,可以像正常的论坛一样发表帖子。帖子按照回复关系组织成 thread。用户在发言的时候有两种选择:创建一个新的主题或者回复一个已经存在的帖子。这样在一个主题内,用户发布的帖子就以树状的形式组织起来,除了根节点以外,其余每个帖子都有唯一的父节点。

在进行正式工作之前,语料中涉及系统问题的帖子(如报告磁盘空间配额不够)和幽默类的帖子被全部删除,因为它们和教学无关。我们聘请了 1 位本科生、3 位硕士生和 1 位博士生作为专家,经过若干次充分的小组讨论之后,最终给出的言语行为分类体系总结在表 1 中,包括 5 大类 18 小类。

表 1 分层的言语行为分类体系(括号中为科恩 Kappa 值)

高层类别	低层类别	描述
QUESTION (0.94)	QWHAT (0.92)	一类诸如"what is something"的提问,问题可以包括定义、概念等,也包括询问时间、地点和原因等信息的提问。不包括询问建议和如何做的问题。注意,"what to do"这类问题应该属于 QHOW,而不属于 QWHAT
	QHOW (0.95)	询问如何做某项任务,或者如何才能完成某一目标的提问。也包括询问建议或可选方案的提问。通常是一个关于任务或过程的提问
	QCONF (0.89)	证实某个主意或看法提问。也包括询问是否被允许做某事的提问,例如"Are we allowed to...?","Should we...?","Is it ok to...?","I was wondering if..."等
	QSURV (0.96)	询问其他人是否也有同样的问题(issue)
ISSUE (0.88)	UNCLEAR (0.91)	发言者认为对某事不清楚或者糊涂。发言者认为某事缺乏理解或者不可理解
	HAVE_PROBLEM (0.78)	发言者认为自己遇到了一些问题或者感觉很奇怪
ANSWER (0.86)	PROP (0.65)	发言者发表的看法、主意、建议或主张,但是对自己发表的意见不是很肯定。例如"I think...","It should probably...","I guess...","It seems...","My understanding is..."等。发言者可能会也可能不会陈述对他所提看法、主意、建议或主张的不确定程度。通常来说,RPOP 的发言者是学生
	INST (0.75)	发言者给出的解决一个问题的建议或方法。发言者通常对自己发表的意见非常肯定。也包括老师或助教的指导
	YES_A (0.92)	对一个 yes/no 提问的肯定回答。如果回答表达了一种赞同的态度,那么应该被看做 agree,而不应该是 YES_A
	NO_A (0.85)	对一个 yes/no 问题的否定回答。如果回答表达了一种不赞同的态度,那么应该被看做 disagree,而不应该是 NO_A
	HINT_Q (1.0)	用提问的方式给出解决问题的线索或暗示
POS_ACK (0.87)	OK (1.0)	对前面的发言表示认可
	THANK (1.0)	对于前面发言者给出的建议致谢。注意:这里不包括在提问的时候期望回答的致谢
	PRAISE (0.81)	通常是老师对某个学生回答的评价。通常出现在一个帖子或句子的开始,或者紧跟在一个人名的后面
	AGREE (0.75)	发言者赞同前一个发言的观点。这里的前一个发言可能是提问或回答,对前一个发言的引用可能是显式的,也可能是隐式的
	IT_WORKS (0.93)	对建议的方案奏效的陈述
NEG_ACK (0.85)	AGREE_OBJ (0.81)	发言者对前一个帖子部分赞同
	DISAGREE (0.89)	发言者反对前一个发言的观点。这里的前一个发言可能是提问或回答,对前一个发言的引用可能是显式的,也可能是隐式的

有很多统计指标可以用来确定评估者间信度,本研究采用科恩 Kappa。假设两个评估者把 N 个东西分到 C 个互斥类,科恩 Kappa 就是测量两个评估者间一致程度[17],即

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

式中,$p_0 = \sum_i p_{ii}$,称为观测一致率;$p_e = \sum_i p_{i.}\, p_{.i}$,称为期望一致率,即两次检验结果由于偶然机会

所造成的一致率。其中，$p_{i.}=R_i/N$，$p_{.i}=C_i/N$，R_i 和 C_i 分别为第 i 个格点所对的行合计和列合计，N 为总列数。当两个评估者完全一致时，$p_0=1$，此时 Kappa 值为 1。当观测一致率大于期望一致率时，Kappa 值为正数，且 Kappa 值越大，说明一致性越好。当观察一致率小于期望一致率时，Kappa 值为负数，这种情况一般来说比较少见。根据边缘概率的计算，Kappa 值应在 −1～1 之间。一般认为 Kappa 值大于 0.75 时一致性较好，Kappa 介于 0.4～0.75 之间时一致性一般，Kappa 小于 0.4 时一致性较差。

标注者间信度（科恩 Kappa 值）标注在每个类别后面的括号里。所有高层言语行为类别的 Kappa 值都超过了 0.85，因此可以认为高层类别是定义良好的。所有低层言语行为类别的 Kappa 值除 PROP 外都超过了 0.7，并且 PROP 的 Kappa 值也接近 0.7，因此可以认为低层类别是定义良好的。其中 QHOW、QHOW、QSURV、UNCLEAR、YES_A、HINT_Q、OK、THANK 和 IT_WORKS 的 Kappa 值超过了 0.9，QCONF、NO_A、PRAISE、AGREE_OBJ、DISAGREE 的 Kappa 值超过了 0.8，HAVE_PROBLEM、AGREE 的 Kappa 值超过了 0.7。Kappa 值在一定程度上反映了一个言语行为类别分类的难易程度。总共标记了 1 116 条发言，其中 2/3 用于训练集，1/3 用于盲测集。

4 特征选取

在归纳前人研究的基础上，根据语料特征，本文提出了六个用于言语行为分类的特征（见表 2）。

表 2 言语行为分类所用特征

特征名称	描述	例子
F1：n 元语法及其位置	unigram，bigram，trigram，two-unigram，three-unigram 和 unigram-bigram 及出现的位置	“where”，“where is”，“where is my”，“what … ?”，“what is … ?”，begin，end，begin & end，other
F2：发言在 thread 中的位置	是不是第一个帖子、是不是最后一个帖子等	第一个帖子、最后一个帖子、既是第一个又是最后一个帖子
F3：前一个发言的言语行为类型	当前发言回复的帖子的言语行为类型	QUESTION、ISSUE、QWHAT、QHOW
F4：发言者类型	关于发言者身份的信息	学生（S）、老师（I）
F5：发言者变换	前一个发言和后一个发言的发言者是不是同一个人	变换（Y）和不变换（N）
F6：发言长度	发言包含字符数的多少	Short（1～5 个字）、Medium（5～30 个字）和 Long（30 个字以上）

（1）n 元语法及其位置

本研究中的 n 元语法包括 unigram，bigram，trigram，two-unigram，three-unigram 和 unigram-bigram。unigram，bigram 和 trigram 是自然语言处理中最常用的特征，因此在本研究中也把它们作为特征。同时还注意到，不连续的词（短语）的组合对于识别言语行为非常有效，例如将“what”和“?”联合起来就是识别 QWHAT 很好的标识。但是由于在语料中“what”和“?”是分开标注的，仅仅用 unigram，bigram 和 trigram 都无法抓住这一特征。因此引入了 two-unigram，three-unigram 和 unigram-bigram 来抓住同一句子内不同特征短语之间交叉产生的特征。典型的例子有“where|_|?”，“does|_|?”等。

仅仅有 n 元语法还不够，因为 n 元语法处于不同位置时，具有不同的含义。例如，当“thank”位于

句子开头时很可能具有 THANK 的言语行为关系，但是位于句子的其他位置（如中间或结尾）时则一般没有 THANK 的言语行为关系。在本研究中区分两种 n 元语法位置，第一种是 n-gram 在句子中的位置；第二种是 n-gram 所在的句子在整个帖子中的位置。

（2）发言在主题中的位置

指示发言在主题中所处的位置，如是否第一个题的第一个帖子，是否主题的最后一个帖子等。位置对识别某些言语行为类别比较有效。例如，主题中的第一个帖子一般都是问题类（QUESTION）的言语行为，而 THANK 或 OK 类的言语行为一般都出现在靠近主题结尾的位置，用来结束一个主题的讨论。

（3）前指发言的言语行为类别

指示当前发言回复的帖子的言语行为类型。例如，如果一个帖子包含问题类（QUESTION）的言语行为，那么下一个帖子的言语行为是回答类（ANSWER）可能性就很大。

（4）发言者信息

关于发言者身份的信息。一个学生给的回答通常都不是很肯定，因此是 PROP 的可能性较大；而老师给的回答通常是 INST 的可能性很大，并且 PRAISE 通常是老师用来表扬学生的，学生的发言不大可能包含言语行为类别 PRAISE。在本研究中，仅仅区别学生（S）和老师（I）。

（5）发言者变换

指示前一个发言和后一个发言的发言者是不是同一个人。本研究识别两种发言者变换的情况：变换（Y）和不变换（N）。

（6）发言长度

指示发言包含字符数的多少。本研究将发言长度分为 Short（1～5 个字）、Medium（5～30 个字）和 Long（30 个字以上）。

5 基于分阶段和分层的言语行为分类方法

基于分阶段和分层的言语行为分类方法本质上是一个通过实验优选结构参数的方法。该方法通过设计和执行一组实验来观察各结构参数对分类结果的影响程度，并通过不断剪枝来确定最优的结构参数。假设有一个分层的言语行为类别体系，一共有 m 个高层言语行为类别和 n 个低层言语行为类别。对于一个给定的高层言语行为类别 $SAH_i, i=1,\cdots,m$，有 n_i 个对应的低层言语行为类别 $\{SAL_j \mid j=1,\cdots,n_i\}$，并且 $\sum_{i=1}^{m} n_i = n$。基于分阶段和分层的言语行为分类方法包含以下三个步骤。

（1）确定每一个高层类别和低层类别的分类难度。这一步的目的是获得各类别分类的难易程度，以便尽量让容易分类的分类器先运行。

（2）分别确定高层言语行为分类和低层言语行为分类时的最优前指层次。这一步的目的是通过探索语料数据的结构特点，获得最优前指层次信息，以便让分类器充分利用分层言语行为体系的好处。

（3）用基于转换的学习方法获得不同言语行为类别之间的依赖关系，并结合（1）的结果，从而让准确和无依赖的分类器先运行，让不准确和有依赖的分类器后运行。

为了避免分类器串行工作时分类准确程度的干扰，以上实验在测试集上运行时都是从金标（golden tag）而不是从分类器的输出（learned tag）获取 previous SA 信息。下面通过一个例子来说明这一方法的使用过程。

5.1 参数选择的实验设计

为了优选分类器运行顺序和前指发言的层次,本文设计了一个2×2的实验(如表3所示)。这个实验对分类的言语行为层次以及previous SA的层次这两个变量进行组合。本研究使用10-fold cross-validation的方法来测试每一组模型的优劣,交差验证得分(F_1值)最高的一组对应的模型就是最优模型。

表3 参数选择的实验设计

分 组	对高层/低层类别分类	Previous SA的层次
第1组	高层类别	高层类别
第2组	高层类别	低层类别
第3组	低层类别	高层类别
第4组	低层类别	低层类别

根据两类错误的定义,记tp(true positive)为所有被标记为真的发言中实际为真的发言数,tn(true negative)为所有被标记为假的发言中实际为假的发言数,fp(false positive)为所有被标记为真的发言中实际为假的发言数,fn(false negative)为所有被标记为假的发言中实际为真的发言数。那么精度、召回率的定义如式(2)和式(3)所示。

$$\text{precision}=\frac{\text{tp}}{\text{tp}+\text{fp}} \tag{2}$$

$$\text{recall}=\frac{\text{tp}}{\text{tp}+\text{fn}} \tag{3}$$

F_1值是精度和召回率的加权平均,其定义如式(4)所示。

$$F_1=2\cdot(\text{precision}\cdot\text{recall})/(\text{precision}+\text{recall}) \tag{4}$$

机器学习算法采用支持向量机模型,选择LibSVM软件包(http://www.csie.ntu.edu.tw/~cjlin/libsvm/)。LibSVM支持很多种工作方式,根据任务的特点,选择C-SVC(C-Support Vector Classification),核函数选择径向基函数(RBF)。该模型包含两个参数:C和γ。C是错误项惩罚参数,用来调节目标函数中模型结构复杂度与经验风险之间的比重,因此又称为调整因子。γ对应径向基函数中需要确定的参数,它反映了训练数据的范围或分布,又称为宽度参数。给定训练数据后,通过LibSVM自带的grid search脚本来获得最优参数C和γ。

5.2 各类别分类的难易程度

各言语行为分类器的平均分类效果(F_1值)如表4所示。由于表4的内容较多,为了方便读者阅读,将表4中的数据画成柱状图(见图1)。

表4 言语行为各类别的分类难易程度

高层类别	F_1值	低层类别	F_1值
QUESTION	0.927 8	QWHAT	0.733 3
		QHOW	0.65
		QCONF	0.746 5
		QSURV	0.714 3

续表

高层类别	F_1 值	低层类别	F_1 值
ISSUE	0.684 2	UNCLEAR	0.748 9
		HAVE_PROBLEM	0.529 05
ANSWER	0.776 2	PROP	0.513 25
		INST	0.614 95
		YES_A	0.798
		NO_A	0.388 9
		HINT_Q	1
ACK_POS	0.569 2	OK	0
		THANK	0.784 6
		PRAISE	0
		AGREE	0.419 9
		IT_WORKS	0
ACK_NEG	0	AGREE_OBJ	0
		DISAGREE	0

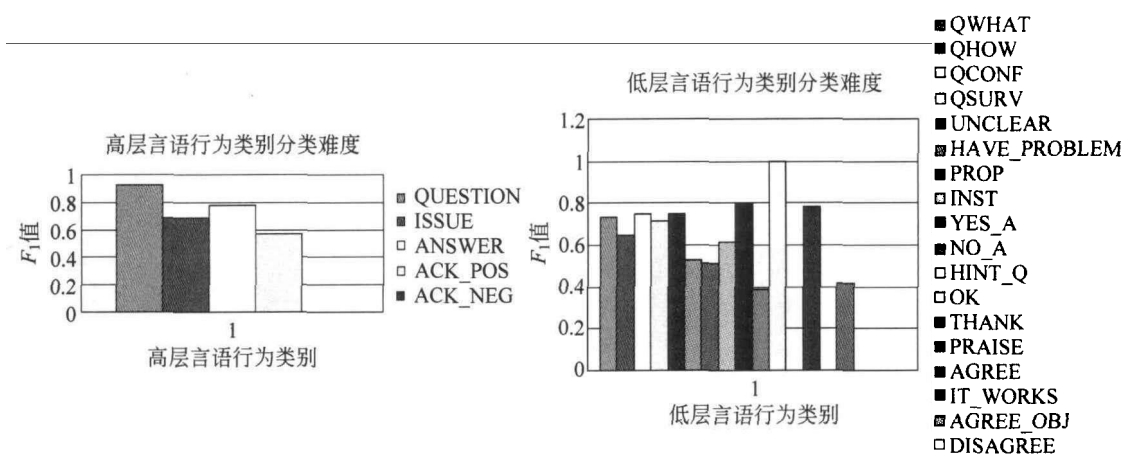

图 1 言语行为各类别分类难度统计图

从表 4 中可以看到，首先最容易分类的高层类别是 QUESTION(F_1=0.927 8)，这是一个非常不错的结果。QUESTION 容易分类，是因为 QUESTION 通常包含问号，并且通常整个 thread 的第一个 post 就包含 QUESTION(这是 PedaBot 语料特点决定的)。其次最容易分类的高层类别是 ANSWER(F_1=0.776 2)，这是因为 ANSWER 数量庞大，通常跟在 QUESTION 后面的就是 ANSWER。ISSUE(F_1=0.684 2)的分类效果是基本满意的。ACK_POS 的平均 F_1 值为 0.569 2，属于不太容易分类的高层类别。ACK_NEG 由于样本数量严重不够，因此不能正确分类。

对于低层言语行为类别，YES_A、HINT_Q、QCONF、QWHAT 的平均 F_1 值都超过了 0.7，算容易分类的低层类别。YES_A 和 HINT_Q 比较容易分类主要是这两种言语行为类型的变化形式较少，训练集中出现的情况基本都覆盖了测试集中出现的情况。QCONF 和 QWHAT 比较容易分类是因为这两种言语行为类型的训练样本较多，而且变化形式也相对较少。THANK、UNCLEAR、QHOW 的平均 F_1 值虽然没有 0.7，但是超过了 0.6，算比较容易分类的低层类别。这些类别的样本数

比起 QCONF、QWHAT 来有所减少，而且变化形式相对增加。

INST、PROP、AGREE、HAVE_PROBLEM、NO_A 属于比较难分类的低层类别。这些类别，有的是因为变化形式太多（如 INST 和 PROP），而有的则是因为训练集中样本数量太少（如 AGREE、HAVE_PROBLEM 和 NO_A）。NO_A 本来看上去应该是一个容易分类的类别，但是不幸的是说 NO 的方式太多（很多委婉的方式）。OK、PRAISE、IT_WORKS、AGREE_OBJ 和 DISAGREE 由于样本数量严重不够，因此不能正确分类。

5.3 前指发言的层次

表 3 中各组配置下的言语行为分类结果如表 5 所示。表 5 中将第 1 组和第 2 组的结果进行比较，第 3 组和第 4 组的结果进行比较，并将效果较好的用黑体显示。表 5 的结果显示，第 1 组全面好于或等于第 2 组。因此可以认为，对高层言语行为分类 previous SA 使用高层类别更为合理。这个结论是符合常识的，因为高层类别的分类不需要引用更加细致的底层类别。

表 5 不同的前指发言层次下言语行为分类结果（F_1值）

高层言语行为分类			低层言语行为分类		
类别	第 1 组（前指高层）	第 2 组（前指低层）	类别	第 3 组（前指高层）	第 4 组（前指低层）
QUESTION	0.927 8	0.927 8	QWHAT	0.733 3	0.733 3
			QHOW	0.5	0.8
			QCONF	0.746 5	0.746 5
			QSURV	0.714 3	0.714 3
ISSUE	0.684 2	0.684 2	UNCLEAR	0.777 7	0.72
			HAVE_PROBLEM	0.558 1	0.5
ANSWER	0.820 9	0.731 5	PROP	0.553 8	0.472 7
			INST	0.635 3	0.594 6
			YES_A	0.818 2	0.777 8
			NO_A	0.444 4	0.333 3
			HINT_Q	1	1
ACK_POS	0.638 3	0.5	OK	0	0
			THANK	0.8	0.769 2
			PRAISE	0	0
			AGREE	0.476 2	0.363 6
			IT_WORKS	0	0
ACK_NEG	0	0	AGREE_OBJ	0	0
			DISAGREE	0	0

对比第 3 组和第 4 组的结果，发现除了 QHOW 以外，第 3 组全面好于或等于第 4 组。这意味着对低层言语行为分类仍然是 previous SA 使用高层类别更为合理。这意味着使用粗糙的高层类别作为前指发言的类型会获得更好的效果。

使用低层言语行为类别作为 previous SA 效果不好的原因，可能是容易让模型过分依赖某一特定的低层言语行为类别。例如，QWHAT 可能导致一个 PROP 的回答，QHOW 也可能导致一个 PROP 的回答，如果使用低层次言语行为类别（如 QWHAT 和 QHOW）作为 previous SA，则需要两条规则才能捕捉到这一信息，增加了模型的复杂程度，从而降低其泛化能力；而如果使用高层言语行为类别

(如 QUESTION)作为 previous SA,则只需要一条规则就可以捕捉到这一信息,并且增加了支持这一规则的实例数量。有鉴于此,在后续的研究中无论对高层还是低层类别进行分类时,都使用高层类别作为 previous SA。

5.4 确定分类器运行顺序

通过在模板中包含前指发言类型的槽,基于转换的学习算法输出的规则可以包含对前指发言类型的引用[18]。例如,下面的例子就是基于转换的学习算法生成的一条规则:

```
Rule1: : IF cue - phrase = ("I am guessing" )
&previous SA = QUESTION
=> PROP
```

这个规则说明如果包含短语"I am guessing",并且前一个发言的类型是 QUESTION,那么这条发言就被判断为 PROP。在这条规则中包含了对前指发言类型 QUESTION 的引用,当基于转换学习算法输出一条这样的规则时,就认为 PROP 类型对 QUESTION 有一次依赖。

通过基于转换学习输出的规则集,以及各类别分类的难易程度,可以画出高层言语行为类别的路径依赖图(如图 2 所示)。图中的椭圆表示言语行为类别,连接椭圆的边表示言语行为关系间的依赖关系,而边上标注的数字则表示类别依赖关系在规则集中出现的次数。同时,还标明了每个高层类别分类的难易程度(F_1值)。例如,在图中可以看到从 QUESTION 有一个指向 ANSWER 的边,并且标有数字 3。这说明 ANSWER 依赖于 QUESTION(或者说 QUESTION 必须先于 ANSWER 被识别),并且这种依赖关系在规则集中出现了 3 次。图 2 中没有孤立点,除了 ACK_NEG 存在自身依赖外,没有其他任何循环依赖现象(即两个不同的言语行为类别相互依赖)。因此可以较容易地确定高层类别分类为如下顺序:QUESTION→ISSUE→ANSWER→ACK_POS→ACK_NEG。

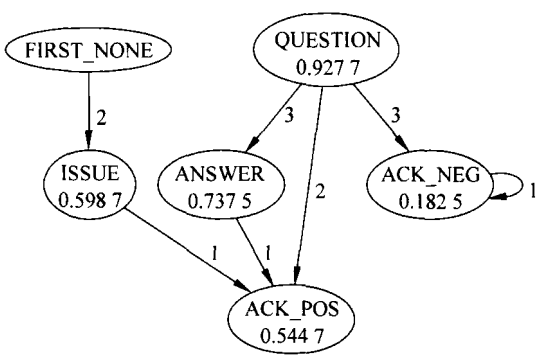

图 2 高层言语行为类别的路径依赖关系图

由于我们决定在低层言语行为分类中引用高层言语行为类别,因此在这里不再给出低层言语行为类别的路径依赖图。实际运行时让高层言语行为分类器先全部运行完,再运行低层言语行为类别分类器。低层类别之间在运行顺序上没有要求。

6 系统评估

为了验证本文所提的基于多阶段和分层的言语行为分类方法的有效性,本研究在盲测集上将多阶段分层法与多个基准方法进行比较。盲测集是在前面的训练中从未见过的数据,占整个语料的大

约1/3。对于高层言语行为分类，比较以下两种基准算法：①不使用 previous SA；②高层 previous SA+随机顺序(即 previous SA 选择高层言语行为类型，但是分类器运行顺序按照随机方式进行)。类似的，对于低层言语行为分类，比较以下两种基准算法：(1)不使用 previous SA；③低层 previous SA+随机顺序(即 previous SA 选择低层言语行为类型，但是分类器运行顺序按照随机方式进行)。机器学习算法采用支持向量机模型。表6展示了几种不同参数配置下的分类结果。

表6 不同的前指发言层次下言语行为分类结果(F_1 值)

高层言语行为分类				低层言语行为分类			
类别	不使用 previous SA	高层 previous SA+随机顺序	多阶段分层法	类别	不使用 previous SA	低层 previous SA+随机顺序	多阶段分层法
QUESTION	0.927 8	0.927 8	0.927 8	QWHAT	0.733 3	0.733 3	0.733 3
				QHOW	0.8	0.8	0.8
				QCONF	0.746 5	0.746 5	0.746 5
				QSURV	0.7	0.7	0.7
ISSUE	0.684 2	0.684 2	0.684 2	UNCLEAR	0.761 9	0.761 9	0.761 9
				HAVE_PROBLEM	0.524 1	0.524 1	0.56
ANSWER	0.721 2	0.789 2	0.820 9	PROP	0.432 6	0.473 9	0.507 9
				INST	0.522 7	0.602 1	0.631 6
				YES_A	0.652 6	0.7	0.814 8
				NO_A	0.2	0.333	0.4
				HINT_Q	1.0	1.0	1.0
ACK_POS	0.452 1	0.527 3	0.6	OK	0.333 3	0.5	0.666 7
				THANK	0.6	0.736 1	0.8
				PRAISE	0	0	0
				AGREE	0.363 6	0.363 6	0.363 6
				IT_WORKS	0	0	0
ACK_NEG	0	0	0	AGREE_OBJ	0	0	0
				DISAGREE	0	0	0

从表6所示的结果可以看出：

(1) 言语行为分类中前指言语行为类型对提高言语行为分类的效果具有重要作用。通过加入前指发言类型这一特征，可以提高 ANSWER 和 ACK_POS(及其对应低层类别)的分类效果。HINT_Q 和 AGREE 是其中的特例，因为它的样本数量特别少，而且不依靠前指发言类型来识别。对于 QUESTION 和 ISSUE(及其对应的低层类别)，引入前指发言类型这一特征对提高发言效果没有帮助。这是因为这些类别通常处于路径依赖关系图的起始位置，因此不依赖于其他任何类别。

(2) 在 E-learning 的语料上，通过构造合理的分层言语行为分类体系，前指发言这一特征使用高层类别比使用低层类别能获得更好的分类效果。由于目前很多言语行为分类体系都是分层的，这一发现为提高分类效果提供了一种思路。

(3) 分类器的运行顺序对于提高分类效果也有重要影响。基于多阶段分层法确定的运行顺序可以获得比随机顺序相同或更好的分类效果。

7 结论

为了解决言语行为分类中应用前指发言类型的问题，本研究提出了一种基于多阶段和分层方法的言语行为分类算法。首先构造一个分层的言语行为分类体系，然后用一个例子说明了基于分阶段和分层方法来选择前指发言类型和分类器运行顺序的方法。最后评估了所提方法的有效性。

本研究在 E-learning 语料上验证了前指发言类型这一特征用于言语行为分类的有效性，发现前指发言类型的高低层次和分类器的运行顺序对分类效果都有较为显著的影响。本研究所提的基于多阶段分层的方法可以给出一个优选的前指发言层次和分类器运行顺序。在盲测集上的运行结果证实了这组优选的参数能够稳定一致地提高 E-learning 语料上各言语行为类别的分类效果。

在本研究中，有一些类别（如 PRAISE、IT_WORKS、AGREE_OBJ、DISAGREE 等）由于样本数量太少，严重影响了分类的效果。对于这类样本数量特别少的类别，需要谨慎处理。如果这个类别有特殊的语言学或管理上的意义，那么是需要保留的；否则就可以考虑将它们进行合并以增加实例数量。

当然，本研究也有一定的局限性。例如，系统评估时使用了在盲测集上一次运行的简单方法，因此减少了评估样本的数量，并且得到的结论不具有严格的统计推断意义。将来的研究可以考虑采用类似 boot-strapping 的方法，采用随机抽样的原则来反复产生训练集和测试集（如产生 50 组训练集和测试集）。在每一组训练集和测试集上运行本文所述的方法，最终的性能是在这 50 组测试集上的平均性能。boot-strapping 方法得到的结论具有严格的统计推断意义，但是对于本研究而言由于非常费时而难以完成，因此可以作为后续研究考虑的一个内容。

言语行为自动分类研究对于对话系统、机器翻译和自动问答系统中问题理解和意图判断具有重要的意义。言语行为的自动分类对于机器理解和刻画研讨势态尤为重要。如果机器能够辨别出发言的言语行为分类体系，就等于理解了用户基本的谈话意图，就可以据此来刻画研讨势态并做出适当的反应。因此，未来的研究可以通过和具体的应用领域相结合，解决很多管理上的问题。例如，通过识别出电子会议中用户发言的言语行为，就可以自动判断会议的研讨态势（如共识点、争议点、分歧点[19]），对会议进行干涉[20]，甚至自动生成会议摘要。另一个研究方向是考虑将变参数贝叶斯模型（如 Dirichlet 过程）应用于言语行为分类。由于这类算法是非监督学习算法，不需要人工构建言语行为分类体系[21]并标注语料，因此具有很大的吸引力。

参考文献

[1] Crook N, Granell R, Pulman S. Unsupervised classification of dialogue acts using a Dirichlet process mixture model [C]. The 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL 2009), Association for Computational Linguistics, 2009.

[2] Feng D H, Shaw E, Kim J, Hovy E. An intelligent Discussion-Bot for answering student queries in threaded discussions [C]. Proceedings of the 11th international conference on Intelligent User Interfaces. Sydney, Australia, ACM Press, 2006.

[3] Gerassimenko O, Hennoste T, Koit M, R bis A, Strandson K, Valdisoo M, Vutt E. Annotated dialogue corpus as a language resource: An experience of building the Estonian dialogue corpus [C]. Proceedings of the 1st Baltic Conference on Human Language Technologies. Riga, Latvia, 2004.

[4] Kim S N, Wang L, Baldwin T. Tagging and linking web forum posts [C]. Proceedings of the Fourteenth

Conference on Computational Natural Language Learning. Uppsala, Sweden, Association for Computational Linguistics Press,2010.

[5] Joty S, Carenini G, Lin C Y. Unsupervised modeling of dialog acts in asynchronous conversations [C]. Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, catalonia (Spain),2011.

[6] Brill E. A corpus-based approach to language learning [D]. Department of Computer and Information Science, University of Pennsylvania. 1993,

[7] 夏新松，肖建国. 一种新的错误驱动学习方法在中文分词中的应用 [J]. 计算机科学,2006,33(3): 160-164.

[8] Zhou Y, Huang C,Gao J,Wu L. Transformation based Chinese entity detection and tracking [C]. Proceedings of the 2nd International Joint Conference on Natural Language Processing,2005.

[9] 赵永贞，刘挺,王志伟,陈惠鹏,邵艳秋. 汉语文语转换系统中停顿指数的自动标注 [J]. 中文信息学报,2004, 18(5): 48-55.

[10] Samuel K, Carberry S, Vijay-Shanker K. Dialogue act tagging with transformation-based learning [C]. Proceedings of COLING/ACL'98,ASSOCIATION FOR COMPUTATIONAL LINGUISTICS,1998.

[11] Reithinger N, Klesen M. Dialogue act classification using language models [C]. 5th European Conference on Speech Communication and Technology. Rhodes,Greece,ISCA,1997.

[12] Vapnik V N. 统计学习理论的本质 [M]. 北京：清华大学出版社,2000.

[13] Cristianini N, Shawe-Taylor J. 支持向量机导论 [M]. 北京：电子工业出版社,2004.

[14] Ravi S, Kim J. Profiling student interactions in threaded discussions with speech act classifiers [C]. Proceedings of AI in Education Conference,2007.

[15] Cohen W W, Carvalho V R,Mitchell T M. Learning to classify email into "speech acts" [C]. 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004). Barcelona,Spain,2004.

[16] Surendran D, Levow G A. Dialog act tagging with support vector machines and hidden Markov models [C]. Ninth International Conference on Spoken Language Processing (INTERSPEECH-2006). Pittsburgh, Pennsylvania,2006.

[17] Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit [J]. Psychological bulletin,1968,70(4): 213.

[18] 李嘉，张朋柱,邓莎莎,原海英. 基于多阶段转换学习的群体研讨文本言语行为分类 [J]. 系统管理学报,2012, (01): 126-132.

[19] 蒋御柱,张朋柱,张兴学. 群体研讨支持系统中的智能可视化研究 [J]. 管理科学学报,2009,12(3): 1-11+43.

[20] 李嘉,张朋柱,邓莎莎,原海英. 群体支持系统中的自动主持人研究 [J]. 管理科学学报,2010,13(12): 34-45.

[21] 李嘉,张朋柱,李欣苗. 面向在线群体研讨的言语行为分类体系设计框架研究 [J]. 现代图书情报技术,2012,(2): 1-9.

Research on Speech Act Classification Based on Multi-phase and Hierarchical Approach

LI Jia,ZHANG Pengzhu,LI Xinmiao

(1. Ease China University of Science and Technology,Shanghai 200237,China
2. Shanghai Jiao Tong University,Shanghai 200052,China
3. Chinese Aeronautical Radio Electronics Research Institute,Shanghai 200233,China
4. Shanghai University of Finance and Economics,Shanghai 200433,China)

Abstract Previous Speech Act (SA) is considered as an important feature in speech act classification. However,how to effectively use previous SA is a critical challenge because it's unknown on test corpus. This paper illustrates that both the level of previous SA and the sequence of running SA classifiers have significant impact on classification effect,and

thus proposes a multi-phase and hierarchical based approach which will suggest a good level of previous SA and a good order of running classifiers. Results on blind test corpus demonstrate that the parameters selected by our approach could steadily increase the classification accuracy for each category.

Key words multi-phase, hierarchical taxonomy, speech act, transformation-based learning, support vector machine

作者简介

李嘉，男，1980 年 12 月生，汉族，湖南湘乡人，讲师(中国航空无线电电子研究所博士后)。主要研究领域：群决策支持系统、信息检索、自然语言处理、数据可视化。在《Journal of the American Society for Information Science and Technology》、《管理科学学报》、《系统管理学报》等杂志上发表论文 10 余篇，会议论文 10 余篇。

张朋柱，男，上海交通大学管理学院管理信息系统系主任、责任教授、博士生(后)导师、中国系统工程学会理事，国际信息系统协会中国分会常务理事。研究领域包括决策与创新支持系统、电子政务、金融信息系统。在国际学术期刊发表论文 10 余篇，在国内重要学术期刊发表论文 60 余篇。曾获上海市科技进步二等奖、陕西省高等学校科技进步一等奖和国家教委科技进步三等奖。

李欣苗，女，副教授。研究方向为团队创新支持系统、群体决策支持系统。在《管理科学学报》、《系统工程理论与实践》、《系统管理学报》上发表论文 10 余篇。