一种基于弱监督学习的论坛帖子对话行为分类方法*

孙承杰，林磊，刘秉权
（哈尔滨工业大学，黑龙江 哈尔滨 150001）

摘要：论坛帖子对话行为分类可以明确每个帖子在当前线索中的角色，有助于重构论坛线索中的对话关系，提高论坛信息检索的效果。本文提出了一种基于弱监督学习的论坛帖子对话行为分类方法，把帖子的对话行为分类作为线索的序列标注问题来解决。该方法的特点只要指定合理的特征约束，就可以训练对话行为分类模型。方法在 CNET 和 edX 数据集上的分类精确率达到 75.6%和 60.7%，优于有监督的条件随机域方法。

关键词：弱监督学习；特征约束；对话行为分类；论坛线索结构分析
中图分类号：TP391　　**文献标识码**：A

A Lightly Supervised Learning Method for Forum Posts Dialogue Act Classification

Sun Chengjie, Lin Lei, Liu Bingquan
(Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Dialogue act classification for online forum post can indicate the role of a post in a thread, which is helpful for reconstruct the conversation relation in a thread and to improve the performance of forum retrieval. This paper proposed a lightly supervised learning method for online forum post dialogue act classification and it cast posts dialogue act classification as sequential labeling problem for threads. The proposed approach can lean the model for dialogue act classification with feature constrains and unlabeled data. It achieved an accuracy of 75.6% and 60.7% in CNET data set and edX data set respectively, which are better that the performances of supervised conditional random fields method.

Key words: lightly supervised learning; feature constrains; dialogue act classification; forum thread structure analysis.

1 引言

随着 Web2.0 技术的发展，现实社会中的各种知识和活动正大量被移植到互联网上，如各种社交网络、论坛和在线教育等。其中论坛是一种重要的交流形式和信息载体，它被广泛应用于在线客户服务、在线社区和在线教育中。某些经过多年发展的论坛中已经积累了丰富的知识，这些知识一般网站中是很难找到的，这使得论坛成为一个独特而重要的知识宝藏。但是由于论坛是一个自由交流的交互性平台，因而其中包含了太多的噪音。海量信息和包含其中的噪音让论坛用户越来越难找到他们需要的信息。

论坛中的每个线索可以看作是一个对话过程，每个帖子对应着提问、回答和确认等不同的对话行为。论坛帖子对话行为分类可以看作是论坛线索结构分析的子任务。论坛线索结构分析可以把按时间顺序线性排列的帖子变成按对话关系排列的树形结构，从而提高论坛信息的访问效率，如文献[1]表明论坛结构分析可以提高针对论坛信息的检索系统的效果。因此，对论坛的线索结构进行分析有重要意义。在论坛线索结构之上，可以进行问答

*收稿日期：　　定稿日期：
基金项目：国家自然科学基金（61100094, 61300114）

对抽取[2]、基于不同级别的论坛检索[3]和专家发现[4]等研究。

不同功能的论坛，其对话行为可以有不同的类别划分，本文主要针对为用户解答问题的论坛的线索。目前，论坛帖子对话行为分类主要采用的还是有监督的机器学习方法，这类方法因为需要标注训练数据，因而成本较高，可移植性较差。本文的主要贡献是提出了一种基于弱监督学习的论坛帖子对话行为分类方法，该方法可以利用由先验知识指定的特征约束来进行机器学习模型参数的训练，具有很好的移植性。在 CNET 和 edX 数据集上的实验结果验证了本文提出方法的有效性。

2 相关工作

文献[5,6]把论坛帖子对话行为分类作为论坛结构分析的子任务，对比了条件随机域(CRF)模型、SVM-HMM 和最大熵模型在不同特征集上的分类效果，实验结果表明 CRF 模型能够更好地利用帖子所在的上下文特征，效果较好。文献[7]提出了 threadCRF 模型来寻找一个线索中帖子间的 reply-to 关系，把一个线索从线性结构转换成树状结构，但没有对 reply-to 关系的类型进行区分。对话行为分类还常被用到对话摘要，电子邮件分析[8]和短消息分析等应用中。目前论坛帖子对话行为分类主要采用的还是有监督的机器学习方法，无监督的方法研究较少。

论坛帖子对话行为分类与贴子所在线索的类型相关，文献[9]利用机器学习的方法对一个在线教育论坛中的线索进行了分类，具体的类别包括公告、问题和解释等。主题信息对论坛对话行为分类也具有一定帮助，属于同一个主题的帖子更有可能形成对话关系。论坛是一种交互式异步对话方式，一个线索中经常会包含多个主题，文献[10]利用论坛结构和 LDA 模型对论坛中的线索进行主题分割和标注。

由于在线论坛中蕴含着丰富的知识，因此针对在线论坛的信息抽取也吸引了很多研究者，跟本文比较相关的研究是问答对信息的抽取。如文献[11]研究了论坛中的问答对抽取问题，提出了基于模式匹配的问题识别和基于图传播方法的答案识别方法。文献[12]分析了在答案识别过程中文本相似度特征的作用，并提出了很多非文本特征。

除了细粒度的论坛信息抽取，还有很多研究者从宏观上研究论坛数据。文献[13]通过对大规模在线教育(MOOC)中的论坛数据分析学生的学习投入程度。Anderson 等人[14] 利用 Stack Overflow 论坛上的数据进行问题的回答速度与答案质量之间的关系分析、答案和问题的影响力预测等研究。微观上的线索结构分析也可以为宏观分析提供特征，使宏观分析的结论更有说服力。

3 在线论坛帖子对话行为分类

3.1 任务定义

一个线索里的帖子组成了一个对话过程，这个过程中的每个帖子可以对应到特定类别的对话行为。假设$\mathbf{F}=\{\mathbf{T}_0,\ \mathbf{T}_1,\ldots,\mathbf{T}_n\}$ 表示一个论坛中所有的线索集合；每个线索$\mathbf{T}$由按时间顺序排列的$\mathbf{m}$个帖子 $\{\mathbf{p}_0,\ \mathbf{p}_1,\ldots,\mathbf{p}_{m-1}\}$ 组成。论坛帖子对话行为分析的目标是为每个帖子$\mathbf{p}_i$指定一个对话行为类别标记$\mathbf{c}_i$。本文采用的论坛帖子对话行为类别标记集共包含 5 个大类，12 个小类，如表 1 所示。每类标记的具体含义可以参考文献[5]。

3.2 特征

常用于论坛帖子对话行为分类的特征主要有 4 类：词特征，帖子在线索中的结构特征，语义特征和发帖人特征。本文主要用到的每种特征的详细描述如下。

词特征(Word Feature)是指利用帖子中出现的词来表示帖子。本文采用 TFIDF 值进行特征选择，选取了不同数量的词特征来进行实验。

表 1 对话行为类别标记

Category	Sub-category	Category	Sub-category
Question	Question-question	Answer	Answer-answer
	Question-add		Answer-add
	Question-correction		Answer-confirmation
	Question-confirmation		Answer-correction
Resolution	Resolution		Answer-objection
Reproduction	Reproduction	Other	Other

帖子在线索中的结构特征(Structure Feature)包含两种：1）帖子的作者是否是帖子所在线索的发起者(Initiator)；2）帖子在线索中的位置(Position)。这些特征跟对话行为比较相关，如线索的发起者所写的帖子的对话行为更可能是 Question 类别的。

本文的语义特征包括简单语义特征和语义相似度特征。帖子的简单语义特征(Post Characteristic Feature)有 3 种，分别是帖子含有 URL 链接、问号和叹号的数量。这些特征的类别区分性也比较强。如根据经验，URL 链接经常出现在 Answer 类别的行为中。语义相似度特征主要包括两种：1）贴子标题之间的语义相似度特征（TitleSim）；2）贴子内容的语义相似度特征（PostSim）。本文采用基于词频的余弦相似度作为语义相似度的度量。一个帖子的 TitleSim 特征的值为位于该帖子之前并且与其具有最大标题相似度的帖子与当前帖子的相对位置。PostSim 的定义与此类似。

发帖人特征(UserProfile)是指帖子的作者所具有的特征。比如该发帖人的权威性，发帖人已经发表的帖子的类别分布等。本文采用了发帖人的 PageRank 值来表示发帖人特征。利用回帖关系，所有的发帖人可以形成一个有向图。利用这个图，就可以计算出每个发帖人的 PageRank 值。PageRank 值大的发帖人更愿意回答别人的问题。

4 方法

4.1 弱监督学习与广义期望准则

弱监督学习是介于无监督学习和半监督学习之间的一类学习方法。它可以在没有标注样本的情况下，利用先验知识和未标注样本对机器学习的模型进行参数估计。利用先验知识来进行机器学习有很多框架，本文采用的是基于广义期望准则(Generalized Expectation Criteria)的框架。广义期望准则框架是由 McCallum 在 2007 年提出的[15][16]，适合与判别式模型结合进行弱监督学习，如文献[17]使用基于最大熵模型的广义期望准则来进行文本分类。

为了把先验知识引入到机器学习模型的参数估计过程中，广义期望准则通常表现为机器学习模型目标函数中的一项，该项可以定义不同的函数S表示模型在某些特征上的期望的偏好。公式(1)中的函数S定义了用 KL 距离来衡量某些特征$\emptyset(x, y_U)$的经验分布$\tilde{\emptyset}$和模型分布$E_{p(y_U|x;\theta)}[\emptyset(x, y_U)]$之间的距离。常用的函数$S$的定义还有欧式距离和最小平方误差等。

$$S(E_{p(y_U|x;\theta)}[\emptyset(x, y_U)]) = -D_{KL}(\tilde{\emptyset}||E_{p(y_U|x;\theta)}[\emptyset(x, y_U)]) \quad (1)$$

广义期望准则所需要的先验知识可以由以下方式获取：领域专家人工指定；通过特征

标注获取（相对于样本标注，特征标注可以减少标注的工作量）；已有的标注数据中获取。缺少标注数据是进入新领域时经常碰到的情形。很多情况下，相近的领域可能已有标注数据。虽然两个领域并不完全相同，但是存在某些同样的特征，这些特征的约束可以从已有的标注数据中获取，然后用于指导新领域的模型学习。

4.2 基于广义期望准则的条件随机域模型

由于广义期望准则只是定义了特征约束和模型期望之间的数值函数，并没有涉及具体的模型，因此需要和具体的机器学习模型相结合来解决实际问题。在线论坛帖子的对话行为类别受其所在的线索的对话历史影响，因此对一个帖子的对话行为分类必须考虑其所在的线索。这样，论坛帖子对话行为分类问题就转化成为一个论坛线索的序列标注问题，因此条件随机域模型是比较合适的选择。所以，本文采用了基于广义期望准则的条件随机域模型(GE-CRF)，模型的目标函数如公式(2)所示。

$$O(\boldsymbol{\theta}) = \log p(y_L|x;\boldsymbol{\theta}) + S(E_{p(y_U|x;\boldsymbol{\theta})}[\emptyset(x,y_U)]) + \log p(\boldsymbol{\theta}) \quad (2)$$

在公式(2)中，$\boldsymbol{\theta}$ 是条件随机域模型的参数，$\log p(\boldsymbol{\theta})$是正则化项，用来约束$\boldsymbol{\theta}$ 的大小。$\mathrm{Log}\, p(y_L|x;\boldsymbol{\theta})$用来计算标注数据的对数似然度，在没有标注数据的情况下，可以去掉这一项。GE-CRF 模型可以采用梯度下降方法来求解参数，本文使用了 Mallet 工具包[†]来实现求解过程。

4.3 特征约束

本文主要采用了从已有的标注数据中获取特征约束的方式。采用这种方式虽然增加了对标注数据的要求，但是依然可以说明本文提出弱监督学习方法的有效性，而且便于提高本文结果的可重复性。在实际应用中，为了减少对训练数据的依赖，可以采用专家指定的方式获取特征约束。表 2 给出了本文所用的特征约束示例。表 2 中的第 1 列表示特征名字，如 1@Initiator 表示该帖子的作者是帖子所在线索的发起者；表 2 中的第 2 列表示第 1 列的特征名字对应的特征对应的特征约束，其实质是每个特征在每个对话行为类别上的概率分布。如第 1 行第 2 列表示如果某个论坛帖子具有 1@Initiator 特征，那么它是 Question-question 类别的概率为 0.45，是 Question-add 类别的概率为 0.3。

表 2 特征约束示例

特征	特征约束
1@Initiator	Question-question:045 Question-add:0.3 Resolution:0.15 Reproduction:0.05 …
0@Initiator	Question-question:0.1 Answer-answer:0.4 Resolution:0.1 Other:0.1 …
01@Postion	Question-question:0.9 Resolution:0.0 Reproduction:0.0 Other:0.1 …

为每个类别都指定准确的概率分布是很难做到的，GE-CRF 并不要求完全准确的概率分布约束，只要是一个大概的估计，甚至可以指定一个概率范围。此外，GE-CRF 也不需

[†] http://mallet.cs.umass.edu/

要为每个特征的每个类别都指定约束，只需要给出那些比较容易估计的特征约束就可以了，因此在表 2 中第 2 列的特征约束只给出了几个类别的概率分布。这些性质极大地降低了 GE-CRF 的使用难度。

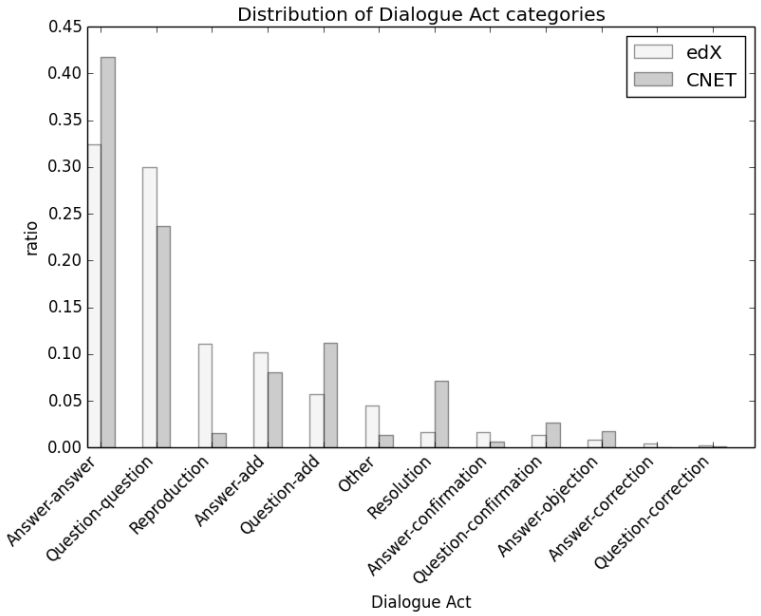
5 实验与结果分析

5.1 实验数据集

本文采用了两个数据集来进行实验。分别是 CNET 数据集和 edX 数据集。CNET 数据集的数据来自 CNET 论坛[‡]，包含 320 个线索，1332 个帖子[5]。数据集的标注采用了表 1 中的类别体系，其中数量最多的类别标记是 Answer-answer，占 40.3%，各个类别的具体数量分布如图 1 所示。从图 1 中可以看出，各个类别的数量分布极不平衡，很多类别的数量比较少， Answer-correction 类别甚至都没有出现。

edX 数据集来自 MOOC 网站 edX[§]上 2013 年春季课程“7.00x: Introduction to Biology – The Secret of Life”的课程论坛。共包含 561 个线索，1977 个帖子。该数据集是采用 Amazon 的 Mechanical Turk 用众包的方式进行标注的。标注集与 CNET 数据集相同，数量最多的类别标记也是 Answer-answer，占 31.9%。

图 1 CNET 和 edX 数据集各类别数量分布



5.2 实验设置

为了便于与他人工作比较，在计算实验结果时，本文采用了按照线索数量划分的 10-fold 交叉验证的平均结果。评价指标采用了整体精确率(Accuracy)。对于全部测试样本而言，分类的整体精确率与整体微平均 F 值(Micro-F)是相等的，因此本文的结果可以直接与文献[5]

‡ http://forums.cnet.com/
§ https://www.edx.org/

中的对话行为分类结果对比。

与有监督学习的 CRF 模型相比，GE-CRF 的训练过程不需要标注样本，只需要有特征约束就可以了。为了在方便在训练过程中构造 GE-CRF 模型训练所需的标记转移矩阵，本文随机给出了每个论坛帖子对应的对话行为类别。

本文利用最小平方差损失函数(L2)作为广义期望项的得分函数。采用最小平方差损失函数的好处是在指定特征约束时，不需要对所有的类别进行指定，这对于标记比较多的任务来说非常方便。如果采用 KL 距离，则需要为每个特征对应的所有类别指定特征约束。本文采用了 Mallet 工具包中实现的 GE-CRF 模型。

实验的任务主要有 3 个：1)测试不同特征组合的分类效果；2）比较 GE-CRF 和其他方法的分类效果；3）测试 GE-CRF 方法在不同数据集上的效果。为了完成任务 1，实验中采用了不同的特征组合方式，每种特征组合的具体实验结果如表 3 所示。表 3 中的实验都是在 CNET 数据集上进行的。表 3 中的词特征是根据 TFIDF 值进行特征选择的。由于很难直接判断某个词的对话行为类别偏好，因此没有对词特征指定特征约束。

5.3 实验结果与分析

从表 3 中可以看出，结构特征对的分类效果最为明显；语义相似度对分类效果有提升作用；没有特征约束的词特征对分类效果也有帮助。词特征对分类效果的提升说明了 GE-CRF 模型在训练过程中，可以利用已有的特征约束，自动优化没有约束的特征，使他们发挥作用。表 3 中，发帖人特征没有对分类效果起到促进作用，可能是因为 CNET 数据集中的帖子数量较少，因而算出的发帖人的 PageRank 值不够准确。

表 3 CNET 数据集上不同特征组合的对话行为分类实验结果对比

特征		Accuracy
结构特征		73.9%
结构特征+简单语义特征		73.9%
结构特征+简单语义特征+语义相似度特征		75.6%
结构特征+简单语义特征+语义相似度特征+发帖人特征		75.1%
结构特征+简单语义特征	+ top 20 word features	74.5%
	+ top 50 word features	74.7%
	+ top 100 word features	74.3%
	+ top 2000 word features	75.4%
	+ top 5000 word features	76.1%

为了评价 GE-CRF 模型的效果，本文对比了在采用同样特征集时，不同方法的实验结果。表 4 中前两行是两种基准方法：第 1 行对应采用大数投票(Majority voting)的方法，把所有的类别都分成 Answer-answer；第 2 行是一种基于帖子在线索中的位置(Position-conditioned baseline)的分类方法，把每个线索中的第 1 个帖子分成 Question-question，把其他所有的帖子都分类成 Answer-answer。从表 4 中可以看出，GE-CRF 的分类效果超过了两个基准方法，甚至好于有监督的 CRF 模型的分类效果。

在 edX 数据集上，机器学习的方法的效果和第二种基线方法相差不大，分析可能的原因有：1）edX 数据集是采用众包方式标注的，标注质量不够高。经过与专家标注的少量数

据比较，kappa 值只有 0.51[**]。2）MOOC 上的论坛的学习者背景多样化，这种多样化使 MOOC 论坛表达方式和用词习惯比较多样化，因而较难分析。

表 4 不同方法的对话行为分类实验结果对比

方法	特征	CNET	edX
Majority voting baseline		40.3%	31.9%
Position-conditioned baseline		64.1%	60.3%
Supervised CRF	结构特征+简单语义特征+语义相似度特征	73.9%	60.5%
GE-CRF		75.6%	60.7%

6 结论与未来工作

本文提出了一种基于弱监督学习的在线论坛帖子对话行为分类方法。该方法以条件随机域模型为基础，可以利用多种方式获得特征约束，具有很好的可移植性。本文测试并分析了不同组合的特征分类效果。在 CNET 和 edX 两个数据集上的实验结果显示了本文提出的基于最大期望准则的弱监督学习方法的有效性。

本文只是利用弱监督学习方法对一个线索中的帖子的对话行为进行了分类，还没有确定每个帖子的链接目标。因此，未来工作包括如何利用弱监督学习寻找每个帖子的链接目标。此外，本文采用的特征约束比较简单，只考虑了单个特征的类别分布约束，探索更复杂的特征约束表示方法，也是未来的工作之一。

参考文献

[1] WANG L, KIM S, BALDWIN T. The Utility of Discourse Structure in Forum Thread Retrieval[C]//Proceedings of 9th Asia Information Retrieval Societies Conference. 2013: 284–295.

[2] 王宝勋, 刘秉权, 孙承杰, 王晓龙孙林. 基于论坛话题段落划分的答案识别[J]. 自动化学报, 2013, 39(1): 11–20.

[3] SEO J, CROFT W, SMITH D. Online community search using thread structure[C]//Proceedings of the 18th ACM conference on Information and knowledge management. 2009: 1907–1910.

[4] RIAHI F, ZOLAKTAF Z, SHAFIEI M, et al. Finding expert users in community question answering[C]//Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion. New York, New York, USA: ACM Press, 2012(i): 791–798.

[5] KIM S, WANG L, BALDWIN T. Tagging and linking web forum posts[C]//Proceedings of the Fourteenth Conference on Computational Natural Language Learning. 2010: 192–202.

[6] WANG L, LUI M, KIM S N, et al. Predicting thread discourse structure over technical web forums[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011: 13–25.

[7] WANG H, WANG C, ZHAI C, et al. Learning online discussion structures by conditional random fields[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011: 435–444.

[8] LAMPERT A, DALE R, PARIS C. The nature of requests and commitments in email messages[C]//Proceedings of the AAAI 2008 Workshop on Enhanced Messaging. 2008: 42–47.

[9] LIN F-R, HSIEH L-S, CHUANG F-T. Discovering genres of online discussion threads via text mining[J]. Computers & Education, Elsevier Ltd, 2009, 52(2): 481–495.

[10] JOTY S, CARENINI G, NG R T. Topic Segmentation and Labeling in Asynchronous Conversations[J]. Journal of Artificial Intelligence Research, 2013, 47: 521–573.

[11] CONG G, WANG L, LIN C-Y, et al. Finding question-answer pairs from online forums[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York, New York, USA: ACM Press, 2008: 467–474.

[**] Kappa 值是通过计算每类标记的 Kappa 值然后通过平均得到的。

[12] GANGADHAR R, KAR R. Does Similarity Matter?? The Case of Answer Extraction from Technical Discussion Forums[C]//Proceedings of COLING 2012: Posters. 2012, 1(December): 175–184.
[13] RAMESH A, GOLDWASSER D. Modeling Learner Engagement in MOOCs using Probabilistic Soft Logic[C]//NIPS Workshop on Data Driven Education. 2013: 1–7.
[14] ANDERSON A, HUTTENLOCHER D, KLEINBERG J. Discovering Value from Community Activity on Focused Question Answering Sites?: A Case Study of Stack Overflow[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012: 850–858.
[15] MCCALLUM A, MANN G, DRUCK G. Generalized expectation criteria[R]. 2007.
[16] MANN G, MCCALLUM A. Generalized expectation criteria for semi-supervised learning with weakly labeled data[J]. The Journal of Machine Learning Research, 2010(11): 955–984.
[17] DRUCK G, MANN G, MCCALLUM A. Learning from labeled features using generalized expectation criteria[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 2008: 595–602.

作者简介：孙承杰（出生年 1980），男，讲师，主要研究领域为信息抽取、推荐系统。Email: cjsun@insun.hit.edu.cn；林磊（出生年 1970），男，副教授，主要研究领域为计算广告学、生物信息学。 Email: linl@insun.hit.edu.cn；刘秉权（出生年 1970），男，副教授，主要研究领域为自然语言处理、智能人机接口。Email: liubq@insun.hit.edu.cn。