从在线聊天语料中识别并抽取需求

Detection of Hidden Feature Requests from Massive Chat Messages via Deep Siamese Network

论文阅读

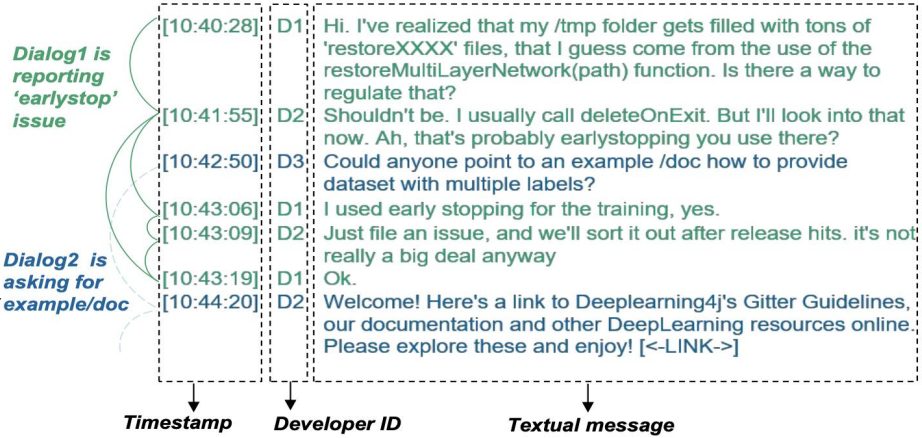
学院：计算机科学与技术学院　专业：软件工程

学生：庄毅非　学号：3200105872

论文目标

1.在线交流平台发展迅速，其中蕴藏的需求信息尚待挖掘

2.现有的信息挖掘技术对于从在线交流平台产生的语料表现欠佳，需要一种高效的挖掘算法。



图一 DL4J开发团队IRC的实时交流记录截图

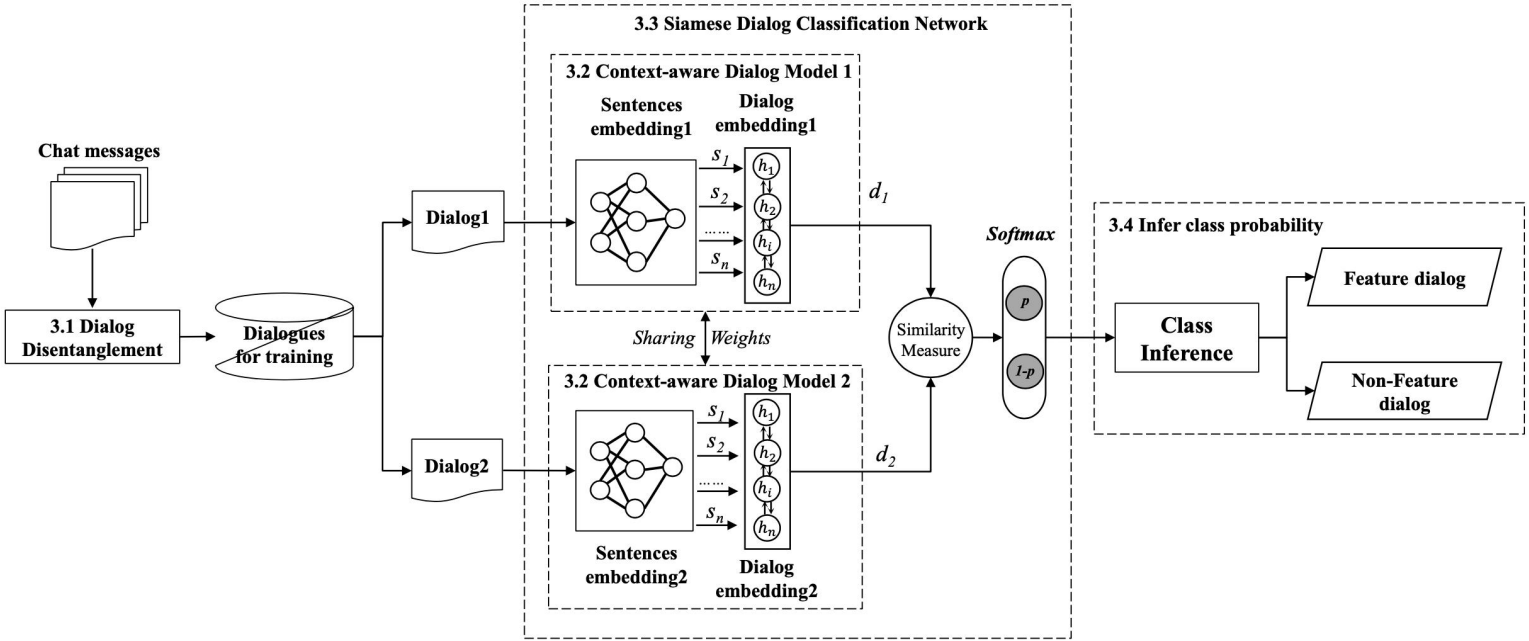
论文解决的问题

1.文本分析困难
解决方法：context-awaredialogmodel

2.数据集不平衡，有效信息比例低，需要大量标识数据
解决方法：siamesenetwork

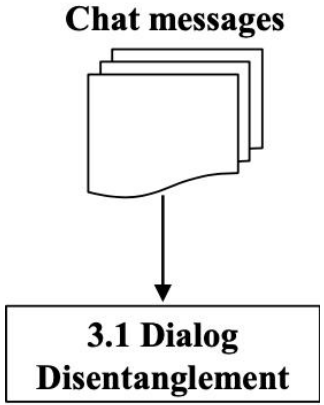
3.需要对语料中的无效信息进行甄别
解决方法：dialogdisentanglementmodel

FR-Miner架构图

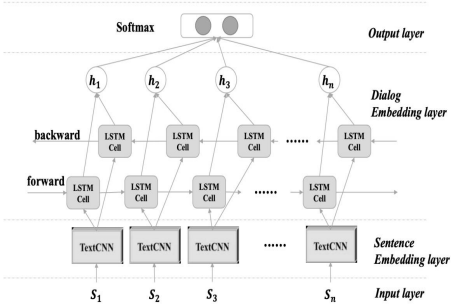

3.3 Siamese Dialog Classification Network
3.2 Context-aware Dialog Model 1
Sentences embedding1
Dialog embedding1
Chat messages
3.1 Dialog Disentanglement
Dialogues for training
Dialog1
Dialog2
Sharing Weights
3.2 Context-aware Dialog Model 2
Sentences embedding2
Dialog embedding2
Similarity Measure
Softmax
p
1-p
3.4 Infer class probability
Class Inference
Feature dialog
Non-Feature dialog

FR-Miner的实现步骤（一、二）

一、在进行对话分析之前，首先需要将聊天记录分割为多个单独的信息流，FR-Miner利用Kummerfeld等人提供的对话分割相关的语料库实现了对话分割。最终FR-Miner以较好的性能达到了74.9%的精度和79.7%的召回率。

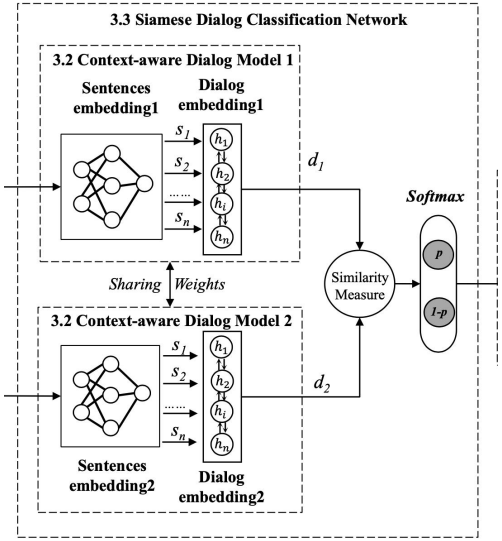


二、为了高效的进行文本语义提取，FR-Miner构建了四层模型，分别是输入层、句子嵌入层、对话嵌入层和输出层。通过在不同层次的迭代处理，最终将对话转换为能够被BiLSTM模型利用的token令牌，并最终通过BiLSTM模型获得对话上下文信息



FR-Miner的实现步骤（三、四）

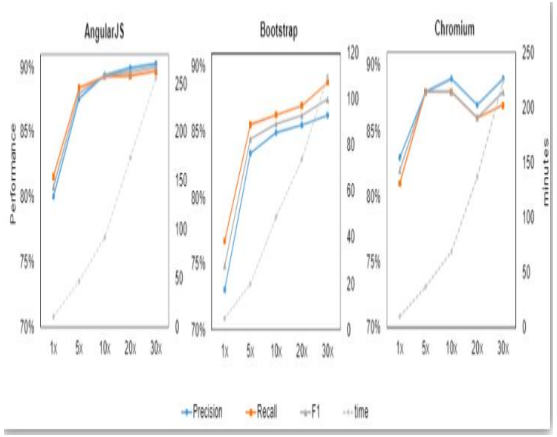
三、为了解决语料库大以及有效信息比例较低所带来的需要大量标识数据的问题，FR-Miner使用孪生神经网络对数据集进行训练，通过充分利用数据对间的对称关系增加训练样本量，从而解决训练样本过少所带来的性能不佳的问题。



四、为了推断一个对话是否属于一个featurerequest，FR-Miner将一个典型的非featurerequest的对话和要判断的对话作为孪生神经网络的两个输入，如果输出为真，那么说明要判断的对话也不属于featurerequest，否则其属于一个featurerequest。

模型效果

方法 \ 效果		AngularJS			Bootstrap			Chromium		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
本文方法	FRMiner	**90.28%**	**89.73%**	**90.00%**	**86.28%**	**88.78%**	**87.52%**	**89.00%**	**87.00%**	**88.00%**
	p-FRMiner	31.71%	54.17%	40.00%	50.00%	47.80%	48.98%	14.00%	44.00%	20.00%
已有研究方法	CNC	7.70%	44.44%	13.13%	16.38%	34.21%	22.13%	9.56%	67.00%	16.73%
	FRA	13.67%	80.33%	23.35%	23.00%	48.67%	31.00%	12.00%	81.00%	20.00%
文本分类方法	NB	20.00%	27.67%	22.33%	25.67%	62.00%	36.00%	14.33%	44.33%	21.00%
	GBDT	36.00%	22.33%	27.33%	41.67%	35.67%	38.33%	9.33%	7.33%	8.00%
	RF	52.67%	11.00%	16.33%	57.00%	29.00%	38.33%	0.00%	0.00%	NA
	FT	23.33%	5.33%	8.67%	57.67%	29.00%	38.33%	38.00%	9.10%	15.00%



方法 \ 效果		AngularJS			Bootstrap			Chromium		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
本文方法	FRMiner	**85.23%**	**86.56%**	**85.89%**	**86.84%**	**85.89%**	**86.37%**	**85.87%**	**86.81%**	**86.34%**
	p-FRMiner	31.03%	50.00%	38.30%	27.56%	69.08%	39.40%	16.00%	50.00%	24.24%
已有研究方法	CNC	7.70%	44.44%	13.13%	16.38%	34.21%	22.13%	9.56%	67.00%	16.73%
	FRA	13.67%	80.33%	23.35%	23.00%	48.67%	31.00%	12.00%	81.00%	20.00%
文本分类方法	NB	16.00%	75.00%	26.00%	27.00%	36.00%	31.00%	7.00%	26.00%	12.00%
	GBDT	18.00%	14.00%	16.00%	30.00%	11.00%	16.00%	20.00%	19.00%	19.00%
	RF	28.00%	14.00%	19.00%	37.00%	9.00%	15.00%	12.00%	26.00%	16.00%
	FT	32.00%	19.00%	24.00%	43.00%	13.00%	20.00%	19.00%	11.00%	14.00%

FRMiner和基线方法对比

FRMiner效果明显更好，因为判断两个对话类别是否相似要比对单个对话进行分类更加容易，孪生神经网络的应用也保证了模型在低数据量下的表现。

FRMiner和p-FRMiner对比结果

和不应用孪生神经网络的p-FRMiner相比，应用孪生神经网络的FRMiner表现明显更好。

跨项目实验结果

FRMiner在提取开发者在不同社区的需求表达习惯，并将其泛化到其他领域这方面有较好的性能

总结

通过使用基于孪生神经网络的FRMiner，我们能够对开源社区使用的在线聊天工具所产生的语料数据集进行分析。从中提取出用户和开发者针对对项目的bug以及feature 所提出的意见，从而准确把握需求。

THANKS!