基于多任务迭代学习的论辩挖掘方法

廖祥文[1),2),3)] 陈泽泽[1),2),3)] 桂 林[1)] 程学旗[4)] 陈国龙[1),2),3)]

[1)](福州大学数学与计算机科学学院 福州 350116)

[2)](福建省网络计算与智能信息处理重点实验室(福州大学) 福州 350116)

[3)](数字福建金融大数据研究所 福州 350116)

[4)](中国科学院网络数据科学与技术重点实验室,中国科学院计算技术研究所 北京 100190)

摘 要 论辩挖掘可分为论点边界的检测、论点类型的识别、论点关系的抽取三个子任务.现有的工作大多数对子任务分别建模研究,忽略了三个子任务之间的关联信息,导致性能低下.另外,还有部分的工作采用流水线模型把三个子任务进行联合建模,由于流水线模型仍然是独立的看待每个子任务,为每个子任务训练单独的模型,存在错误传播的问题,且在训练过程中产生了冗余信息.因此,本文提出了一种基于多任务迭代学习的论辩挖掘方法.该方法将论辩挖掘三个任务并行地联合在一起学习,首先通过深度卷积神经网络(CNN)和高速神经网络(Highway Network),获得文本字符和词级别的浅层共享参数表示;然后输入双向长短时记忆循环神经网络(Bi-LSTM),利用论辩挖掘三个任务之间的关联信息进行同时训练,不仅可以避免错误传播,而且能够克服冗余信息的产生;最后,联结三个任务的 Bi-LSTM 网络输出作为下一次迭代的输入,来提高模型的性能.实验采用了德国 UKP 实验室公开的学生论文数据集,实验结果表明:与目前最好的基准方法对比,该方法的准确率指标提高了 2.74%,"$F1$(100%)"和"$F1$(50%)"指标分别提高了 1.05%和 1.19%,很好地验证了该方法的有效性.

关键词 多任务学习;论辩挖掘;迭代模型;深度学习;卷积神经网络

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2019.01524

An Argumentation Mining Method Based on Multi-Task Iterative Learning

LIAO Xiang-Wen[1),2),3)] CHEN Ze-Ze[1),2),3)] GUI Lin[1)] CHENG Xue-Qi[4)] CHEN Guo-Long[1),2),3)]

[1)](*College of Mathematics and Computer Science, Fuzhou University, Fuzhou* 350116)

[2)](*Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing (Fuzhou University), Fuzhou* 350116)

[3)](*Digital Fujian Institute of Financial Big Data, Fuzhou* 350116)

[4)](*CAS Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

Abstract Argumentation mining has recently become a hot topic in the field of data mining and natural language processing. Its main task is automatic identification of argumentative structures in persuasive essays so as to help people better understand the massive text information. A persuasive essay usually consists of a series of argument components. The types of argument components are generally classified into claims or premises, and the types of relationship between argument components are commonly classified into support or attack. Argumentation mining typically contains three consecutive subtasks, i.e., (1) Argument component boundary detection (ACBD Task), which involves separating argument component from non-argumentative text units and identifying the argument component boundaries; (2) Argument component identification

收稿日期:2018-04-26;在线出版日期:2018-11-30.本课题得到国家自然科学基金项目(61772135,U1605251)、中国科学院网络数据科学与技术重点实验室开放基金课题(CASNDST201708,CASNDST201606)、可信分布式计算与服务教育部重点实验室主任基金(2017KF01)资助.**廖祥文**,博士,副教授,中国计算机学会(CCF)高级会员,研究方向为文本倾向性检索与挖掘.E-mail:liaoxw@fzu.edu.cn.**陈泽泽**,硕士研究生,研究方向为文本倾向性检索与挖掘.**桂 林**,博士,研究方向为自然语言处理.**程学旗**,博士,研究员,博士生导师,中国计算机学会(CCF)会员,研究领域为网络科学、网络信息安全、互联网数据挖掘.**陈国龙**,博士,教授,博士生导师,研究领域为人工智能与网络安全.

(ACI Task), whose goal is to classify argument components into different types, such as claims or premises; (3) Argument component relation identification (RI Task), which aims to identify the relationship type between argument components, such as support or attack. Recently, many researchers have proposed a series of argumentation mining models and made brilliant improvement. However, most of the existing approaches mainly focus on modeling each subtask and ignore the correlation information among the three subtasks, resulting in low performance. In addition, some of the approaches utilize pipeline methods to jointly model three subtasks. The pipeline methods still consider each subtask independently, and train separated models for each subtask, which could lead to error propagation and redundant information in the training process. More specifically, the error of argument component boundary recognition module affects the following argument component classification performance. Similarly, the error of argument component classification also influences the performance of argument component relation identification. To solve these problems above, we propose a multi-task iterative learning method which assumes that tags predicting for one task could be useful feature for other tasks, and joints three subtasks in parallel to learn together for argumentation mining. Firstly, we obtain the shallow shared parameters of the text character and word level by utilizing the deep Convolutional Neural Network (CNN) and the highway network. And then, the Bi-directional LSTM neural network is trained to solve three subtasks at the same time to avoid error propagation. In the training process, the correlation information among each subtask is used to overcome the generation of redundant information. Finally, the output of three subtasks is concatenated as the input for the next iteration to improve the performance. Multi-Task Learning (MTL) is an important machine learning mechanism and improves the generalization performance by learning a task together with other related tasks. Our model based on MTL could iterative utilize predicting tags' distribution of each task explicitly. Experimental results on student essays published by the UKP laboratory in Germany show that, compared to the state-of-the-art models, our model improve 2.74% on accuracy, 1.05% on "$F1(100\%)$" and 1.19% on "$F1(50\%)$", which verify the validity of our model. Besides, results also show that the performance of multi-task learning is better than single task learning.

Keywords　multi-task learning; argumentation mining; iterator model; deep learning; convolution neural network

1　引　言

随着互联网技术和社交媒体的快速发展，用户产生了大量的观点评论等主观性数据，对这些主观性数据的研究蕴含了巨大商业价值和学术价值. 论辩挖掘(Argumentation Mining)旨在研究如何从主观性数据中自动地识别论点并抽取论点关系，以满足信息化背景下人们对信息检索和信息抽取的更高需求[1]，正逐渐成为情感分析领域的研究热点. 它可以广泛地应用在司法[2]、人文与教育[3]、用户生成内容[4]等领域，为人们提供便捷的自动化工具.

论辩挖掘中的论点部件(Argument Component)是人们用来说服听众接受某种特定观点的基本单位[5]，通常一个主要论点(Major Claim)由多个主张(Claim)组成，而一个主张由多个前提(Premise)来支持它. 图 1 所示的是一段学生论文的例子，论点部

① [Living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet.]$_{主张}$ ② [One who is living overseas will of course struggle with loneliness, living away from family and friends]$_{前提1}$ but ③ [those difficulties will turn into valuable experiences in the following steps of life.]$_{前提2}$ Moreover, ④ [the one will learn living without depending on anyone else]$_{前提3}$

图 1　已标记的学生论文数据集[1]样例

件①是这段论文的主张，论点部件②、③、④是这段论文的前提.并且论点部件②与论点部件①的主张有着攻击(Attack)关系，论点部件③与论点部件②的也是攻击关系，而论点部件④与论点部件①是支持(Support)关系，最后形成了图2所示的整段论辩文本结构图.

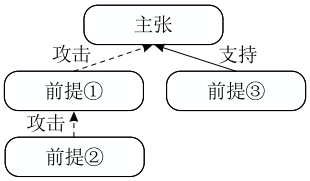

图2 样例文本的论辩结构图

论辩挖掘主要分为3个任务[6]，包括(1)论点边界的检测(Argument Component Boundary Detection，ACBD)，即从论点无关的文本中分离有论辩性的文本并检测论点的边界[7-8]；(2)论点类型的识别(Argument Component Identification，ACI)，即识别论点的类型，通常论点类型划分为主要主张、主张、前提[9-12]；以及(3)论点关系的抽取(Argument Component Relation Identification，RI)，即抽取论点之间的关系，通常把论点关系划分为支持和反对[10,13].

目前在论辩挖掘的研究中，大多数的工作主要侧重研究论辩挖掘的一个子任务，为每个子任务训练独立的模型，这些方法主要分为两大类：(1)基于机器学习的方法.通过提取文本中词法、语义、句法结构、情态动词和动词时态等特征，训练多项式朴素贝叶斯[8]、C4.5决策树[7]、支持向量机[1]等二元或多类分类器来进行论点边界分割，论点类型分类以及论点关系抽取，这些方法十分依赖于手工特征的设计；(2)基于深度学习的方法.通过训练递归神经网络模型[14]来进行论点边界的检测，利用循环神经网络对论点的类型进行分类[15].这些方法大多以句子为单位进行标注，只利用了文本中的局部信息解决论辩挖掘的一个或两个问题.另外，以词为单位的联合序列标注方法[1,16-17]，可以利用文本中上下文的长期依赖信息，对三个子任务进行联合训练，取得了较好的性能优势，主要分成两大类：(1)基于流水线(Pipeline Method)的方法[1,17]：主要有整数线性规划模型(Integer Linear Programming，ILP)，它首先通过使用支持向量机(SVM)、条件随机场(CRF)等方法，独立串行地训练三个子任务的分类模型，最后定义一个整数线性规划函数进行全局最优化求解任务的标签预测结果.流水线方法由于论点类型识别的错误会影响到论点关系的抽取，存在错误传播的问题.另外，这种方法将识别出来的论点进行两两配对，之后进行论点关系分类，产生了论点关系对的冗余信息；(2)基于深度学习的方法：其中有Bi-LSTM-CNNs-CRF序列标注模型[16]，将三个子任务的标签拼接成一个整体，训练神经网络模型来预测总体的标签分布，这种方法依然没有利用任务之间的关联信息.

针对上述问题，本文提出了一种基于多任务迭代学习的论辩挖掘方法，该方法假设论辩挖掘三个子任务之间是相互关联的，不是各自独立的子任务，一个任务的标签预测结果可以作为预测其它论辩挖掘子任务标签的有效特征.模型使用基于词级别的BIO标注方法[16]，迭代地利用每个子任务的标签分布，通过提取字符和词级别的特征表示构成共享参数层，并行的进行模型训练学习，并且在预测模型中融入了任务相关的特征.该模型不独立看待每个子任务，不仅使得每个子任务的标签预测结果相互学习，有利于减少错误传播的概率，并且避免了由于无关论点的两两配对，产生的冗余信息.

本文采用德国UKP实验室公开的学生论文数据集进行实验[1]，结果表明与BiLSTM-CNN-CRF、StagBLCC、LSTM-ER和ILP等基准方法对比，本文模型不仅在预测论辩挖掘三个子任务标签整体准确率指标上取得了最优的效果，在论点类型识别这个任务上，“C-$F1$(100%)”和“C-$F1$(50%)”评价指标分别提高了0.39%和1.05%；在论点关系抽取任务上“R-$F1$(100%)”和“R-$F1$(50%)”指标上提高了1.26%和1.18%；在论点边界检测任务上，$F1$值超过90.0%，达到了92.2%.更进一步地，本文验证了不同迭代次数的实验结果，发现随着迭代次数的增加，模型的性能越来越好，很好的证明了本文所提模型迭代学习的有效性.

本文第2节为相关的工作；第3节为问题描述与动机；第4节提出本文的模型；第5节介绍实验数据集；第6节为实验，通过与基准实验的对比验证本文方法的有效性，并对实验结果进行分析；第7节为结束语.

2 相关工作

论辩挖掘已成为当前研究的热点，大多数的工作是基于每个子任务单独建模研究，对于任务一，从论文无关文本中分离有论辩性的文本并检测其边界，通常被看做是一个二分类的问题，作为论辩挖掘流水线任务的第一步，传统的机器学习方法大多集

中在特征设计上，Moens 等人[8]在 Araucaria 语料集[18]中，通过提取单词对，文本统计，动词论辩性语句关键指示词特征，训练多项式朴素贝叶斯和最大熵模型作为分类器来分类论辩性和非论辩性语句，取得了最好性能为 73.75%的准确率. Florou 等人[7]基于标点、情态动词和动词时态等功能特征，使用 C4.5 决策树学习算法[19]作为分类器. Li 等人[14]把任务一看成一个序列标注的问题，训练一个不需要依赖于特征的递归神经网络模型来解决论点边界检测的问题.

对于任务二，论点部件类型的识别，Teufel[12]等人假设任务一已经完成，即假设已经从文本中准确地提取了论点部件，他们将每个论点句子划分为主张、结果和目的等七种类型，通过提取文本中的结构、词性、语法等特征，训练朴素贝叶斯模型来预测有论辩性句子的论点类型. Rooney 等人[11]使用基于自然语言处理的核方法来进行论点部件的分类，不需要任何启发式的特征. Feng 等人[9]的方法是基于从论点部件类型的互信息中提取特征，因此它需要预先知道论点的类型. Laha 等人[15]最先将基于神经网络模型来应用于论辩挖掘研究，使用两个循环神经网络对论点的类型进行分类. Gao 等人[20]把论点类型识别看成序贯决策问题，提出了一个基于强化学习的方法解决任务二.

对于任务三，论点关系类型的检测，Palau[10]等人根据法律领域的文档，手动创建上下文无关语法(CFG)来检测论点关系的类型，这种方法不具有通用性，无法应用在其它领域的文档上. Cabrio 等人[21]通过结合文本中的蕴含关系来预测论点部件之间的关系. Stab 等人[6]把任务三看成一个二分类任务，使用传统机器学习的方法，提取文本中的结构、词法、语义、指示词等特征，训练分类器，预测主张-前提论点关系类型是支持或者是攻击. Peldszus 等人[22]使用最小生成树算法通过计算论点之间的关系形成整体文章的整体结构，从而预测论点的关系类型.

目前联合论辩挖掘三个子任务一起做的模型通常是采用流水线的方法，Persing 等人[17]最先提出了使用基于流水线方法的端到端模型，输入未标注的文本，通过流水线模型解决三个子问题，最终输出标注的文本. 类似的，Stab 等人[1]，首先为每个子任务训练独立的模型，然后定义一个整数线性规划模型(ILP)进行全局最优化求解. Eger 等人[16]把论辩挖掘看成基于词级别的依赖解析和序列标注问题，将三个任务的标签融合在一起，对每个单词进行标注，并使用序列标注问题中的经典的双向 LSTM-CNNs-CRF 模型[23]预测论辩挖掘中标签结果. Niculae 等人[24]提出了一种不需要构建树结构就能从文档中抽取论辩关系的因子图的方法. 同时，Potash 等人[25]基于 Pointer 网络，提出了一个联合模型同时解决论点类型分类和抽取论点关系两个任务.

论辩挖掘被广泛应用于许多领域，Moens 等人[8,10]将论辩挖掘应用于法律决策；Kirschner 等人[26]将论辩挖掘研究方法用来分析科研论文的文档摘要；Boltužić 等人[27]应用于文本的观点挖掘，在教育领域；Somasundaran 等人[28]将论辩挖掘应用于论文的自动评分系统；Zhang 等人[29]将论辩挖掘应用于写作辅助系统，以及 Florou 等人[7]将论辩挖掘应用于支持政府制定政策.

3　问题描述与动机

3.1　问题描述

主观性数据的自然语言文本通常是由一系列论点通过一定的结构化关系组成，如图 2 所示的是图 1 中学生论文样例构成的论辩结构图. 论辩挖掘就是研究如何从主观性数据文本中自动地识别论点，判断论点的类型并抽取它们之间的关系. 实质上是一个序列标注问题.

论辩挖掘问题形式化定义描述如下：给定一篇文本 $X=\{x_1,x_2,x_3,x_4,\cdots\}$ 和类别标签集合 $y=\{y_1,y_2,y_3,y_4,\cdots\}$，其中 x_i 表示文本中的一个单词，每个 x_i 都跟标签 y_i 关联[16]. 利用算法模型，将文本中的每个单词 x_i 映射成一个类别标签 y_i，即 $X\rightarrow Y$. 算法模型的输入是一篇主观性文本，输出是论辩挖掘三个任务对应的标签 y_i，y_i 定义下：

$$\begin{aligned} y_i=\{(b,t,d,s)\,|\,&b\in\{\mathrm{B,I,O}\},t\in\{\mathrm{P,C,MC},\perp\},\\ &d\in\{\cdots,-2,-1,1,2,\cdots,\perp\},\\ &s\in\{\mathrm{Supp,Att,For,Ag},\perp\}\} \end{aligned} \tag{1}$$

标签 y_i 包含了 4 个元组(b,t,d,s)，其中 b，即为任务一的标签，使用 BIO 标记的方法，O 表示论点无关的单词，B 表示句子中论点开始的单词，I 表示论点句子中间部分的单词. t 表示论点类型，MC 表示主要主张(Major Claim)，即作者对文章主题提出的中心立场；C 表示主张(Claim)，即对主要主张(MC)某一个方面提出的一个观点；P 表示前提，即为主张(C)或者其它前提(P)提供支持或者反对的论据；(b,t)组成任务二的标签. d 表示当前论点部件与它相关的论点距离. s 表示论点关系的类型，其中 Supp 和 Att 分别表示前提与主张之间的支持和攻击关

系,For 和 Ag 分别表示主张(C)与主要主张(MC)之间的赞同和反对的关系,(d,s)组成任务三的标签.同时定义了一个特殊的符号$\perp$表示该类型属性为空,比如,当一个单词属于论点无关时,它显然没有论点类型,也无论点关系.表 1 中给出了学生论文数据集的标注样例.

表 1 学生论文数据集标注样例

Living **B,C,⊥,For**	And I,C,⊥,For	studying I,C,⊥,For	overseas I,C,⊥,For	is I,C,⊥,For	an I,C,⊥,For	irreplaceable I,C,⊥,For	experience I,C,⊥,For
when I,C,⊥,For	it I,C,⊥,For	comes I,C,⊥,For	to I,C,⊥,For	learn I,C,⊥,For	standing I,C,⊥,For	on I,C,⊥,For	your I,C,⊥,For
own I,C,⊥,For	feet I,C,⊥,For	. O	**One** **B,P,−1,Att**	who B,P,−1,Att	is I,P,−1,Att	living I,P,−1,Att	overseas I,P,−1,Att
will I,P,−1,Att	of I,P,−1,Att	course I,P,−1,Att	struggle I,P,−1,Att	with I,P,−1,Att	loneliness I,P,−1,Att	, I,P,−1,Att	living I,P,−1,Att
away I,P,−1,Att	from I,P,−1,Att	family I,P,−1,Att	and I,P,−1,Att	friends I,P,−1,Att			

3.2 模型动机

多任务学习是一种重要的机器学习模型,它能够通过与其它相关任务共享参数层和特征一起学习来提高模型的泛化性能[30].而论辩挖掘的三个子任务之间是有关联的相关任务,例如,论点类型为前提或者主张比论点类型为无关论点更有可能是攻击或者是支持的论点关系类型.

在序列标注问题中使用一个任务的预测标签来改善相关任务的性能,称为堆叠序列学习(Stacked Sequence Learning)[31].因此,本文基于如下假设:论辩挖掘中一个子任务的预测标签能够作为有效特征,来改善其它子任务标签的性能.

4 模型建立

4.1 基于多任务迭代学习的论辩挖掘方法

本文通过引入多任务迭代学习方法来解决论辩挖掘中的三个子任务.

如图 3 所示,对于给定的输入的文本序列 $\boldsymbol{x}$,多任务迭代学习模型预测第 i 个论辩挖掘子任务的标签分布 $\boldsymbol{y}^{(i)}$ 型的输入主要包括三个部分:

(1) $h^{(\text{shared})}$:迭代模型底层的通用参数,如图 5 所示,它通过 CNN 和高速神经网络,从数据中提取不同子任务的共同特征,并在模型中被所有任务共享的参数.

(2) $\boldsymbol{y}=\boldsymbol{y}^{(1)}+\boldsymbol{y}^{(2)}+\boldsymbol{y}^{(3)}$:位于迭代模型高层,联结上一次迭代的三个相关论辩挖掘任务的标签分布参数 $\boldsymbol{y}$.

(3) $f^{(\text{shared})}$:迭代模型中论辩挖掘每个子任务的特征表示,如表 2 所示,包括文本结构、文本语义等特征表示.

我们将 $h^{(\text{shared})}$ 和 $\boldsymbol{y}$ 联结起来作为双向长短时循环记忆神经网络(Bi-LSTM)的输入,如图 6 所示,论辩挖掘三个任务共享 Bi-LSTM,$h^{(\text{shared})}$ 与 Bi-LSTM 在每次多任务迭代训练学习过程中相互分离,并且在每次迭代训练预测标签过程中,我们将任务一论点边界检测任务的标签分布输出联结为任务二论点类型识别任务的神经网络输入,将论点类型识别任务的标签输出联结为任务三论点关系抽取任务的神经网络输入.在迭代训练的预测模型中,考虑三个子任务不同的特点,我们分别抽取每个子任务的特征构成 $f^{(\text{shared})}$ 来预测最后的标签结果.三个子任务的特征如表 2 所示,主要包括文本结构、文本语义等特征.Stab 等人[1]所提出的论辩挖掘模型中同样使用这些特征并取得了较好的效果.

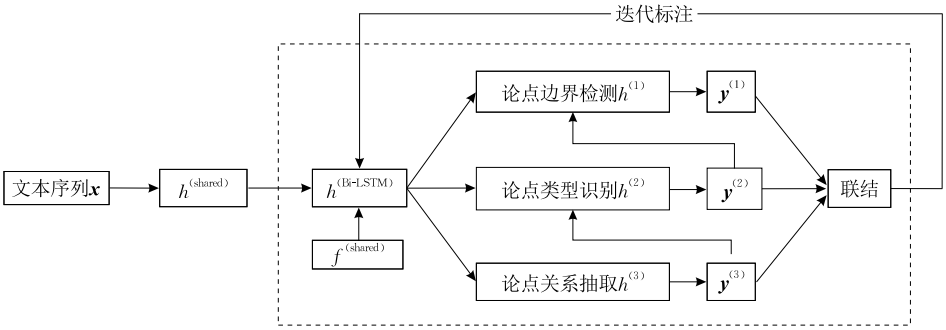

图 3 论辩挖掘的迭代学习标注模型

表 2　三个子任务的特征抽取

任务	特征描述
论点边界检测	单词是否是句子的开头或者结尾 单词位于整篇文本，段落，句子相对绝对位置 单词与句号，逗号，分号等标点符号的距离 词性特征(POS)
论点类型识别	单词是否为指示词以及指示词类型 单词是否共享同一个名词或动词短语 动词的时态以及是否为情态动词
论点关系抽取	论点部件之间是否有共享名词以及数量 论点部件是否在同个句子或者段落中 论点部件是否是在段落的开头或结尾 论点部件之间的距离长度

多任务迭代学习模型能够利用模型前一次迭代的所有任务标签分布作为下一次迭代的输入特征，对于每个任务来说，前一次迭代中所有任务的标签分布可以通过标签的交互信息来修改下一步中预测错误的标签结果. 同时，通过使用双向长短时记忆网络(Bi-LSTM)，该模型将标签的交互扩展到句子级别.

为了确保每次迭代预测的结果与真实的标签相接近，在每一步的迭代结果中定义了一个损失函数 $cost$，如方程(1)所示：

$$cost=\frac{1}{T}\sum_{i=1}^{T}L(\boldsymbol{y}_t,\boldsymbol{y}_*) \tag{2}$$

$$L(\boldsymbol{y}_t,\boldsymbol{y}_*)=\frac{1}{M}\sum_{i=1}^{M}\alpha_m\widetilde{L}(\boldsymbol{y}_t^{(m)},\boldsymbol{y}^{(m)*}) \tag{3}$$

其中，$\boldsymbol{y}_t$ 是第 t 次迭代的预测标签分布，$\boldsymbol{y}_*$ 是真实的标签结果，T 是迭代的总次数，也称为递归迭代层的长度，M 是相关任务的数量，α_m 表示第 m 个任务的权重，L 是交叉熵函数.

最终的预测的结果是所有预测标签分布的平均值，如方程(3)所示：

$$\boldsymbol{y}^{(m)}=\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{y}_t^{(m)} \tag{4}$$

在论辩挖掘序列标注模型中，本文构建了一个由 CNNs-Highway-LSTM 组成的神经网络序列标注模型. 模型首先由字符和词级别的 CNN 来捕捉文本的特征表示. 随后，将两个 CNN 提取得到的特征输入高速神经网络中，目的在于通过高速神经网络中的转换门(transform gate)来过滤有价值的特征. 然后，过滤后的特征作为多任务学习底层框架中的共享表示，输入至 Bi-LSTM 网络中进行训练. 最后，Bi-LSTM 输出相关任务的表示与底层的共享表示联结在一起，进行迭代学习.

4.2　基于 CNN 的论辩挖掘文本表示

本文采用了基于 CNN 的词级别和字符级别的论辩挖掘文本表示，该模型非常适用于形态丰富的语言文本中，能够从论辩挖掘文本中获取到丰富的词素、语义和形态等特征，为下一步的实验打下基础.

4.2.1　基于 CNN 的词级别表示

词级别的卷积神经网络，我们扩展使用了 Kim 等人[32]用来解决序列标注问题的卷积神经网络.

如图 4 所示，卷积神经网络(CNN)输入为文本

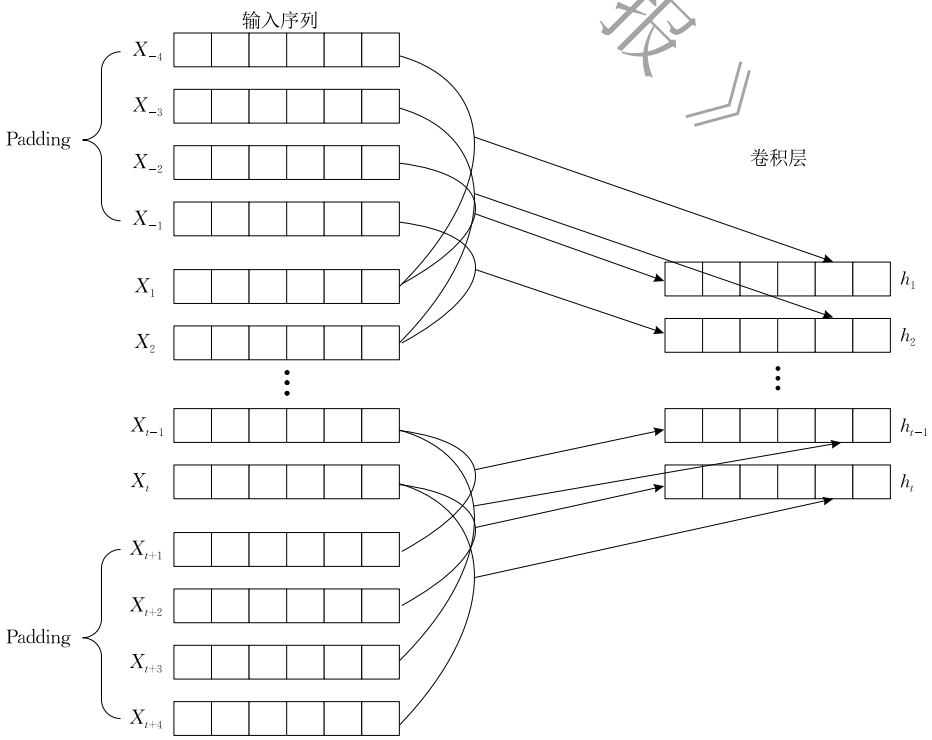

图 4　基于 CNN 的论辩挖掘文本表示

序列 $x=[x_1,x_2,\cdots,x_n]$，按照文本句子中单词的顺序，每一行都是一个由 d 维向量表示的单词，CNN 输出为序列 $C=[c_1,c_2,\cdots,c_n]$，C 表示输入每个单词的特征，n 表示输入序列的最大长度. 我们在 x 之间使用窄卷积和一个宽度为 k 的卷积核 $W\in R^{(d\times k)}$，并且将 $\left\lfloor\frac{k}{2}\right\rfloor$ 和 $\left\lfloor\frac{k-1}{2}\right\rfloor$ 作为填充向量填充到序列的头部和尾部，以便保证输入序列的长度在卷积层后不会发生改变.

$$c_i=f(W^T\cdot x_{\left(i-\left\lfloor\frac{k}{2}\right\rfloor\right):\left(i+\left\lfloor\frac{k-1}{2}\right\rfloor\right)}+b) \tag{4}$$

其中，c_i 卷积后的输出结果，f 是非线性激活函数，b 是偏差，$x_{i:j}$ 表示序列中第 i 个到第 j 个单词. 在输入论辩挖掘的文本序列中，滑动三种不同长度 $k=3,5$ 和 7 的卷积核 W 去获取多个局部上下文特征向量，最后这些多维度的特征被联结为局部特征.

4.2.2 基于 CNN 的字符级别表示

基于卷积神经网络的字符的表示已经被证明是从单词的字符中抽取形态特征有效的方法[33]. 与基于 CNN 的词级别表示类似，当给定一个单词，我们将它的字符嵌入到卷积神经网络层得到特征映射，接着通过池化层进行 max-over-time pooling 操作，从特征映射中捕捉重要的特征，经过池化层的输出就是单词的字符表示向量，最后将字符表示向量与词向量联合作为卷积神经网络的输入.

4.3 基于高速神经网络的特征过滤

在我们的实验中，如果只有词级别和字符级别的文本表示，实验性能无法达到最优. 为了更好地从 CNN 的字符和词级别的表示中提取出有效的特征，我们在卷积神经网络层之后紧接一个高速神经网络(Highway Networks)层[34]，如图 5 所示. 高速神经网络通过增加 transform 门和 carry 门来控制数据的比例，用于过滤出文本中的重要特征，具体实现如下：

$$r_i=c_{\left(i-\left\lfloor\frac{k}{2}\right\rfloor\right):\left(i+\left\lfloor\frac{k-1}{2}\right\rfloor\right)} \tag{6}$$

$$\breve{c}_i'=f(W_C^T\cdot r_i+b_C) \tag{7}$$

$$t_i=\sigma(c_i\cdot W_T+b_T) \tag{8}$$

$$g_i=1-t \tag{9}$$

$$c_i'=t_i\odot\breve{c}_i'+g_i\odot c_i \tag{10}$$

其中，$c_{i:j}$ 表示论辩挖掘文本序列中第 i 个单词到第 j 个单词的卷积结果，f 是非线性激励函数，W_C，b_C，W_T，b_T 是线性变换参数，t_i 是高速神经网络中的 transform 门，g_i 是高速神经网络的 carry 门，高速神经网络允许一部分的 c_i 在通过卷积变换后输出的结果不发现改变.

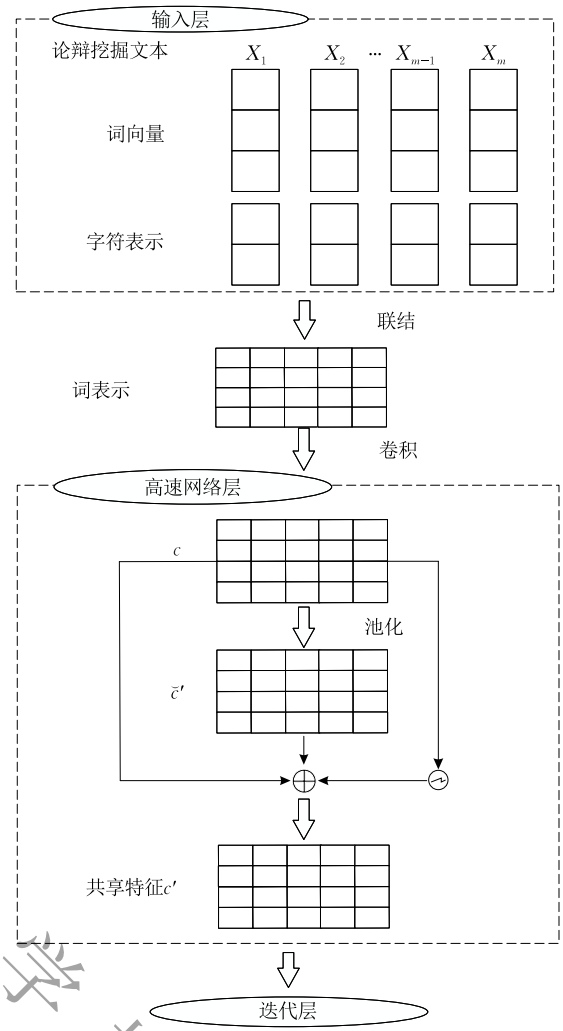

图 5 多任务学习迭代模型共享底层

4.4 基于 Bi-LSTM 模型的标注方法

在论辩挖掘问题的主观性文本中，文本中上下文信息蕴含着十分重要的特征. 因此我们将高速神经网络层输出的共享特征用做长短时记忆网络(LSTM)[35]输入，LTSM 网络通过维护三个门限来控制信息是否被遗忘或者是传送到下一步中，从而解决自然语言文本中长期依赖的问题. 具体实现如下：

$$I_t=\sigma(W_{SI}s_t+W_{HI}h_{t-1}+W_{CI}c_{t-1}+b_I) \tag{11}$$

$$F_t=\sigma(W_{SF}s_t+W_{HF}h_{t-1}+W_{CF}c_{t-1}+b_F) \tag{12}$$

$$c_i=F_i\odot c_{t-1}+I_t\odot c_t^f \tag{13}$$

$$c_t^f=\tanh(W_{SC}s_t+W_{HC}h_{t-1}+b_C) \tag{15}$$

$$o_t=\sigma(W_{SO}s_t+W_{HO}h_{t-1}+W_{CO}c_{t-1}+b_O) \tag{16}$$

$$h_t=o_i\odot\tanh(c_t) \tag{17}$$

其中，σ 为 sigmoid 激活函数，h_i 是高速神经网络输出层的第 i 个单词，$\odot$ 是点积.

在论辩挖掘这个序列标注的问题上，通过使用一个双向 LSTM 神经网络(Bi-LSTM)[36]来捕捉论

辩挖掘文本中的上下文信息. Bi-LSTM 神经网络在前向和后向传播过程中，使用两个隐藏状态 h' 和 h'' 分别取捕捉文本中“过去”和“未来”的信息，在神经网络的输出，将两个隐藏状态联合起来作为最后神经网络的输出结果.

4.5　模型求解

多任务迭代学习的论辩挖掘模型，由图 5 所示的共享底层和图 6 所示的迭代框架组成. 共享底层的由字符和词级别的卷积神经网络(CNN)和高速神经网络(Highway Networks)构成，它与迭代框架中的 LSTM 网络，一起组成论辩挖掘的基本标注模型 CNNs-Highway-LSTM. 论辩挖掘三个子任务在训练时一起共享 CNNs-Highway-LSTM 组成的网络结构，Bi-LSTM 在迭代框架中，被用来更好地捕获论辩挖掘文本中上下文的依赖信息. 如图 6 所示，在训练过程中，我们将任务一论点边界检测任务的标签分布输出，联结为任务二论点类型识别任务的神经网络输入，将论点类型识别任务的标签输出，联结为任务三论点关系抽取任务的神经网络输入，并加入论辩挖掘任务相关的特征. 在迭代模型的每次迭代中，随机选择一个任务并根据任务特定的目标更新模型，重复执行算法 1，直到达到训练模型的最大 epoch 次数. 值得注意的是，网络中每次迭代的参数不共享. Bi-LSTM 输出的 $h^{(i)}$ 均是任务相关的参数. 训练模型的算法如下：

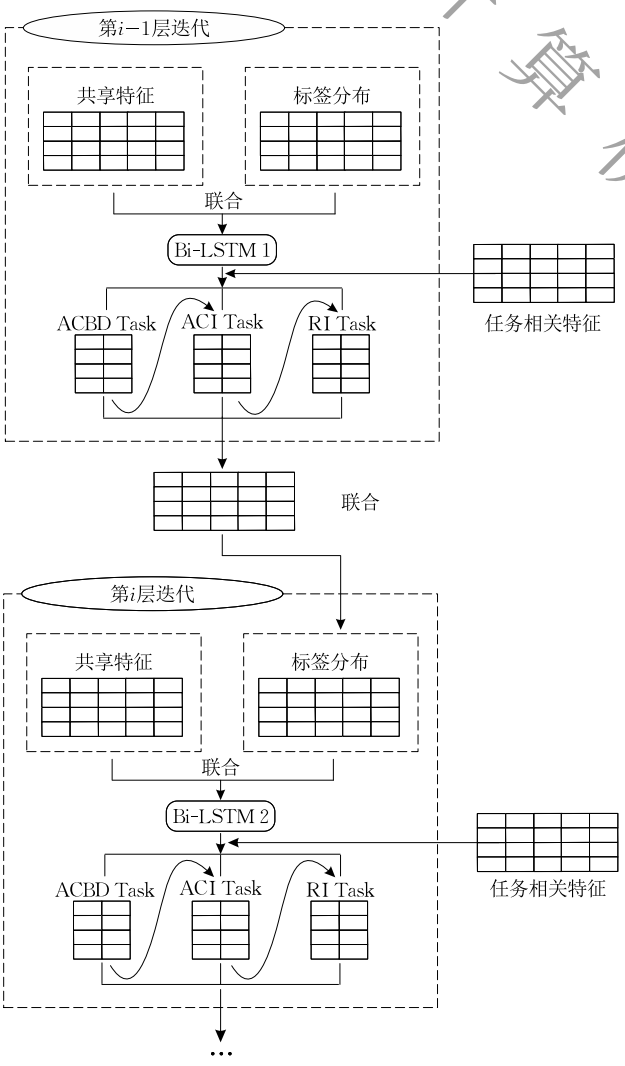

图 6　多任务学习迭代模型迭代框架

算法 1.　多任务迭代学习模型的训练算法.

输入：论辩挖掘三个子任务的训练数据集序列 $X=\{(X_1,X_2,\cdots)_m\}_{m=1}^3$ 和标签 $y^*=\{(y_1^*,y_2^*,\cdots)_m\}_{m=1}^3$

输出：给定论辩挖掘三个任务序列的预测标签 $y=\{(y_1,y_2,\cdots)_m\}_{m=1}^3$

1. 初始化模型参数 P
2. WHILE $t\leq T$（T 是总迭代次数）DO
3.　FOR 每个子任务 $m\leq M$（M 是任务总数）DO
4.　　从第 m 个任务中随机选取一批训练数据 b_m
5.　　计算 b_m 的 loss 值 L_m
6.　　根据 L_m 使用 Adam 方法计算 ∇p_m 梯度下降
7.　END FOR
8.　计算平均梯度 $\nabla p=\frac{1}{|M|}\sum_{i=1}^{M}\nabla p_m$
9.　根据 ∇p 更新模型的参数 p
10. END WHILE

5　数据集描述

本文采用德国 UKP 实验室公开的学生论文数据集进行实验[11]，这个数据集随机地从 essayforum① 论坛中挑选 402 篇学生英文论文，每篇论文包含一个主题. essayforum 是一个能够为不同类型的观点性文本提供书写反馈的在线论坛. 例如，学生用户可以根据论坛中的不同的主题，如表 1 样例中的“living and studying overseas” 在海外学习和生活好不好为主题，发表自己的观点，进行写作，并在线提交他们的论文. 专家会针对论文提供反馈意见. 数据集包含 7116 个句子，由 147271 个单词组成.

实验数据集中训练集和测试集的划分如表 3(a) 所示，将 402 篇论文中的 322 篇划分成训练集，80 篇划分成测试集. 数据集的对每个单词进行标注.

表 3(a)　数据集统计

	训练集	测试集
论文总数	322	80
段落总数	1786	449
单词数量	118648	29538

① https://essayforum.com/

数据集具有以下的特殊结构,(1)论点类型为主要主张(MC)的论点与其它的论点没有关联关系;(2)主张(C)总是关联全部的主要主张(MC);(3)每个主张(C)至少关联一个前提(P)或者其它主张(C).

数据集中,无关论点的单词的数量有47174个,包含了1631个句子,占总数的32.2%,表3(b)展示训练集和测试集分类标注的结果.总体而言,有751个单词为主要主张(MC),1506个单词为主张(Claim),3832个单词为前提(Premise).论点之间有5338个关系,其中大部分是支持关系(>90%).

表3(b) 训练集和测试集标签分布统计

类别	训练集	测试集
论点边界分类		
Arg-B	4823(4.1%)	1266(4.3%)
Arg-I	75053(63.6%)	18655(63.6%)
Arg-O	38071(32.3%)	9403(32.1%)
论点类型分类		
主要主张	598(12.4%)	153(12.1%)
主张	1202(24.9%)	304(24.0%)
前提	3023(62.7%)	809(63.9%)
论点关系分类		
支持	3820(90.4%)	1021(91.7%)
反对	405(9.6%)	92(8.3%)

6 实 验

6.1 实验环境

实验环境为Ubuntu 14.04.1,四块GeForce GTX 1080 Ti显卡,共44 GB显存,Intel(R) Xeon(R) CPU E5-2620,32 GB,Python 2.7.13,TensorFlow-GPU(0.12.1).

6.2 实验对比模型

将基准方法与本文的方法在相同的数据,实验选取了以下对比模型:

(1)ILP(Integer Linear Programming)模型[1].该模型基于特征的选择,模型首先选择文本中的结构,词法语法和上下文等特征对论辩挖掘的三个任务,分别通过支持向量机,条件随机场方法构造分类器进行分类标注,之后定义了一个带有约束条件的目标方程,对分类器的结果进行全局调优.

(2)LSTM-ER模型[16].该模型基于端到端的神经网络模型,联合了实体和树结构的关系信息,对文本中的命名实体和关系进行抽取,模型的实体检测是使用BiLSTM-CRF(BLC)标记模型,关系抽取则是实现一个神经网络用来预测检测到的实体之间的关系.这个关系抽取模块能够充分地使用依赖关系树中的信息.为了在让LSTM-ER模型适应论辩挖掘模型的学生论文数据集,本文编码了三种命名实体(前提P,主张C,主要主张MC),四种关系类型(支持Support,攻击Attack,赞同For,反对Against).

(3)$\text{Stag}_{\text{BLCC}}$模型[23].这个模型首先使用字符级卷积神经网络获得词的表示;之后,将词表示和训练完成的词向量联结起来,输入到Bi-LSTM网络中,得到每个状态的表示;最后,将Bi-LSTM的输出结果输入条件随机场(CRF)层,最终预测结果.

(4)Stag_{BL}多任务学习模型[16].这个模型将式(1)中的y看成多任务学习中的主要任务,将y中论点类型识别任务的(b,t)标签和论点关系抽取(d,s)标签看做辅助任务.

(5)LSTM-CRF-MTL多任务学习模型[37].这个模型是多任务学习的传统方法模型,将LSTM-CRF作为多任务学习的基础神经网络模型,任务之间共享一个通用的表示层,并单独为论辩挖掘三个子任务训练三个不同的LSTM神经网络.

(6)Joint RNN Model[14]模型.该模型利用递归深度神经网络来解决论辩挖掘的论点边界检测任务.

(7)HAs-augmented RL模型[20],该模型利用论辩挖掘文本中上下文的特殊语境信息,通过强化学习的方法,将论点部件类型识别任务看做序贯决策问题进行建模.

6.3 评价指标

为了评估本文提出模型的有效性,采用了以下评价指标进行实验:

(1)准确率(Acc).论辩挖掘任务中,三个任务分类正确的样本总数除以所有样本的总数.准确率越高,模型分类性能越好,其式子为

$$Acc=\frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i}{}_{y_j^*=y_j}1}{\sum_{i=1}^{M}N_i}.$$

N_i为第i个任务的测试集大小,y_j^*表示第j个样本的预测标签,y_j表示其正确的标签,M为任务的总数.

(2)$F1$.类似Eger等人[16],本文使用真阳性TP,假阳性FP,假阴性FN,真阴性TN,来计算模型分类结果的$F1$值,$F1=\frac{2TP}{2TP+FP+FN}$,对于预测论点部件识别抽取的性能,Persing等人[17]定义了一个"α匹配"的概念,比如当α为100%时,

表示预测的标签结果与真实的标签结果完全一致. 当 α 取值为 50%，表示预测标签中至少有 50% 的部分与真实的标签相匹配. 本文将这些分布称为 *C*-*F*1(100%)和 *C*-*F*1(50%). 类似的，对于论点关系的类型，定义为 *R*-*F*1(100%)和 *R*-*F*1(50%). 显然 *R*-*F*1 的值取决于 *C*-*F*1 的值，因为预测正确的论点关系类型必须要以预测为正确的论点类型为基础. 同时，定义了两个“α 匹配”分别 100% 和 50% 的全局 *F*1 值，其式子如下：

$$F1(100\%)=\frac{2\times C\text{-}F1(100\%)\times R\text{-}F1(100\%)}{C\text{-}F1(100\%)+R\text{-}F1(100\%)},$$

$$F1(50\%)=\frac{2\times C\text{-}F1(50\%)\times R\text{-}F1(50\%)}{C\text{-}F1(50\%)+R\text{-}F1(50\%)}.$$

6.4　实验参数设置

神经网络训练通过反向传播算法进行训练，并使用 Adam 梯度下降法[38]更新神经网络模型参数. 在本文的实验中，字符嵌入的随机初始化，维度设置为 64 维，并在训练过程中使用 Find-Tuned 方法进行调整. 与 Ma 等人[23]一样的，我们使用 Stanfords GloVe[39]中的 100 维向量作为本文的词向量.

在实验中，我们使用线性整流 ReLu 函数作为模型的激活函数，Adam 梯度下降法学习率初始化设置为 0.01，dropout 率设置为 0.2，l_2 正则化为 1E-5，最小的 batch 为 100. 本文网络中所有的参数均通过在[−0.1，0.1]的正态分布进行初始化. 字符级别的 CNN 网络的过滤窗户大小设置为 1，3，5 并且每个过滤窗口带有 30 个特征映射，词级别的 CNN 网络的过滤窗口大小设置为 1，3，5 并且每个特征窗口带有 128 个特征映射. 前向和后向传播的 LSTM 网络层的维数被设置为 128 维.

6.5　实验结果分析

我们首先分析了多任务迭代学习方法中的迭代次数以及网络结构部件对实验结果的影响；然后，将本文所提出的方法与其它论辩挖掘联合模型的实验结果进行对比来验证方法的有效性；最后，与传统的多任务学习模型进行对比，从而说明本文所提出的多任务迭代学习模型在解决论文挖掘问题上的优势.

6.5.1　迭代次数对实验结果的影响

为了验证循环迭代训练方法的有效性，本文在实验中，设置 15 组不同的迭代次数(分布取 1、2、3、4、5、6、7、8、9、10、11、12、13、14、15 次)，对比记录不同迭代次数下，在学生论文数据集上的三个任务的实验结果，结果如表 4 所示.

表 4　迭代次数对总性能的影响

迭代次数	*Acc*	*F*1
T=1	61.51	46.83
T=2	63.41	51.44
T=3	63.70	51.65
T=4	64.26	52.77
T=5	64.31	54.06
T=6	64.33	54.70
T=7	64.34	55.29
T=8	64.35	55.85
T=9	64.38	56.18
T=10	**64.41**	**56.47**
T=11	64.36	56.42
T=12	64.34	56.12
T=13	64.31	55.24
T=14	64.27	54.02
T=15	64.26	54.00

在训练过程中，在相同的参数设置和网络初始权值条件下进行训练. 在数据集上，每个迭代次数的设置采用 10 次实验的平均结果进行综合评价. 具体结果如表 4 所示，可以发现，当 *T* 小于等于 10 时，迭代模型在论辩挖掘三个任务上的总体准确率 *Acc* 和 *F*1 值随着迭代次数的增加不断提高. 当迭代次数为 10 次时，实验结果取得最好性能. 当迭代次数大于 10 次，随着迭代次数的增加，迭代模型的实验性能逐步降低. 因此，本文在接下来的实验中选取最优的迭代参数为 10 次. 当迭代次数为 1 次时，模型转化为传统的多任务学习神经网络模型，此刻的模型无法有效地利用论辩挖掘相关子任务之间关联的信息，*Acc* 和 *F*1 值均为最低值. 由此可见，模型的迭代的次数对论辩挖掘任务的总体性能有着很大的影响.

更加进一步的，我们分别考察迭代次数对论辩挖掘三个任务的影响. 如图 7 所示，随着迭代次数的增加，论点类型识别任务的 *C*-*F*1(100%)和 *C*-*F*1

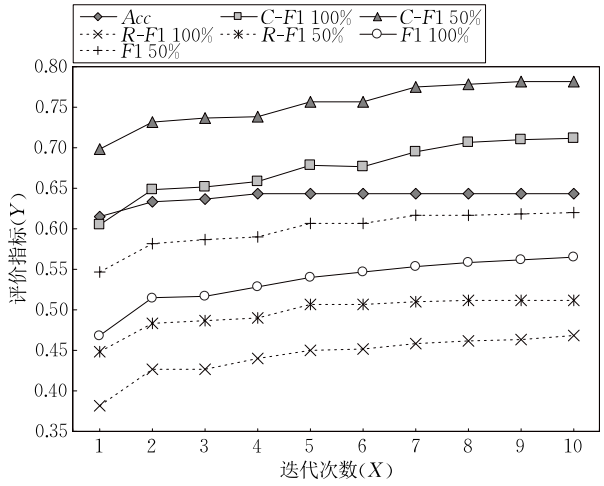

图 7　迭代次数对论辩挖掘子任务的影响

(50%),论点关系抽取任务的 R-$F1$(100%)和 R-$F1$(100%),三个任务总体的 $F1$(100%)和 $F1$(50%)都呈现线性提高.说明迭代模型能够有效地降低标签的错误传播,是一种有效且分类能力好的神经网络训练方法.

在迭代训练的过程中,本文模型将第 $i-1$ 次迭代预测错误的标签,在第 i 次迭代时被校正为正确的标签.例如,表 1 中的样例在第 1 次迭代的时候,单词"One"的预测标签结果是"$(B,P,-1,Supp)$",此时,任务一的标签(B)和任务二的标签(B,P)均预测正确,而任务三的标签$(B,-1,Att)$被错误地预测为$(B,-1,Supp)$;当经过第 2 次的迭代之后,预测标签变为$(B,P,-1,Att)$,可以发现,此时任务三被校正为了正确的标签$(B,-1,Att)$.说明通过本文提出的多任务迭代学习方法,模型学习到了论辩挖掘子任务之间潜在的关联信息使得评价结果表现得更好.

6.5.2 网络结构对实验结果的影响

为了测试模型中每层网络结构部件对实验结果的影响,我们单独地移除模型中的网络结构部件进行实验,实验包括在模型中去除字符级别 CNN 表示层,词级别 CNN 表示层,高速神经网络层以及 Bi-LSTM 层.实验结果如表 5 所示,可以发现,用于捕获上下文信息的 Bi-LSTM 层对实验结果影响最大,其次是字符级别的 CNN 表示层,字符级别 CNN 表示层能够提升模型 1.29%的准确率,并且词级别的 CNN 层和高速神经网络层都让模型的实验性能得到进一步的提升.由此表明,模型中的每一个网络结构部件对提升实验效果有着重要作用.

表 5 模型中网络结构部件对实验结果的影响

模型	Acc
完整的网络结构	**64.41**
去除字符级别 CNN 表示层	63.12(−1.29)
去除词级别 CNN 表示层	63.75(−0.66)
去除高速神经网络层	64.06(−0.35)
去除 Bi-LSTM 层	62.68(−1.73)

6.5.3 本文方法与其它联合模型的实验效果对比

为了验证多任务迭代学习训练方法联合解决论辩挖掘三个子任务的有效性,将本文方法和现有的论辩挖掘联合方法进行了对比,对比实验结果如表 6 所示,表格从上到下的方法分别是 ILP 模型、LSTM-ER 模型、$\text{Stag}_{\text{BLCC}}$模型以及本文的方法.

表 6 论辩挖掘联合模型在学生论文数据集的实验结果

	Acc	C-$F1$		R-$F1$		$F1$	
		100%	50%	100%	50%	100%	50%
ILP	60.32	62.61	73.35	34.74	44.29	44.68	55.23
LSTM-ER	61.67	70.83	77.19	45.52	50.05	55.42	60.73
$\text{Stag}_{\text{BLCC}}$	59.34	66.69	74.08	39.83	44.02	49.87	55.22
本文方法	**64.41**	**71.22**	**78.24**	**46.78**	**51.23**	**56.47**	**61.92**

从表 6 中,我们可以看出,各个基准方法整体的实验性能都明显低于本文提出的方法,本文所提的方法在准确率"Acc"上达到了 64.41%,比 ILP 流水线模型提升了 4.09%,比 LSTM-ER 模型提升了 2.74%,这表明多任务迭代学习方法优于传统的流水线方法.此外,在论点类型识别任务中,本文所提的方法,"C-$F1$(100%)"和"C-$F1$(50%)"两个评价指标比串行的 ILP 模型分别提高 6.61%和 4.89%;并且在论辩挖掘关系抽取任务中,"R-$F1$(100%)"和"R-$F1$(50%)"的评价指标上也比 ILP 流水线模型提高了 12.04%和 6.94%,表明多任务迭代学习模型能够有效减少传统流水线模型中由于论点类型分类错误而导致的接下来论点关系抽取的错误.综上所述,与论辩挖掘联合模型的对比实验中,可以看出,本文提出的方法在 7 个评价指标上均明显高于现有的方法,从而验证了多任务迭代学习模型方法在联合解决论辩挖掘三个任务时的有效性.

为了体现本文所提的方法实验效果的显著性,我们采用 T 检验的方法重复实验 10 次,本文方法以 $p\text{-}value<0.01$ 显著性优势压倒其它模型,说明采用多任务迭代学习的论辩挖掘方法起到了效果.

6.5.4 本文方法与其他多任务方法的实验效果对比

为了验证本文所提的多任务迭代方法比其他的多任务学习方法在解决论辩挖掘任务上具有更好的泛化效果.我们将所提的方法与相关的多任务方法在上述的 7 个评价指标上进行比较.对比方法包括 Stag_{BL}模型[16],LSTM-CRF-MTL 模型[37],具体结果如表 7 所示,可以看到,我们的方法相较于前两个多任务学习方法取得更好的效果.需要说明的是这三个方法的任务学习形式,Stag_{BL}模型是借助辅助任务学习的形式,将式(1)中的 y 看成多任务学习

中的主要任务(Main Task)，将 y 中论点类型识别任务的(b,t)标签和论点关系抽取(d,s)标签看做辅助任务(Auxiliary Tasks)，通过辅助任务来使得主要任务在多任务学习中受益. 而 LSTM-CRF-MTL 模型和本文所提的方法是借助多任务学习中的联合学习(Joint Learning)的形式，联合论辩挖掘的三个相关任务一起学习来提升模型的泛化能力. 在这样的情形下，本文提出的方法在 *Acc*、"*C*-*F*1(100%/50%)"、"*R*-*F*1(100%/50%)" 和 "*F*1(100%/50%)" 这 7 个评价指标上显著地高于其他多任务的方法. 从而验证了本文所提出的方法在泛化三个相关任务潜在的关联信息具有更大的优势.

表 7　多任务模型在学生论文数据集的实验结果

	Acc	*C*-*F*1		*R*-*F*1		*F*1	
		100%	50%	100%	50%	100%	50%
Stag_{BL}	50.33	54.58	67.66	30.22	40.30	38.90	50.51
LSTM-CRF-MTL	53.59	66.87	73.13	33.82	37.02	44.92	49.16
本文方法	**64.41**	**71.22**	**78.24**	**46.78**	**51.23**	**56.47**	**61.92**

6.5.5　论辩挖掘独立任务的实验结果分析

为了测试多任务迭代学习模型在论辩挖掘每个子任务上的效果，本文将所提的方法独立地对比论辩挖掘三个子任务的基准方法. 首先，论辩挖掘论点边界检测作为论辩挖掘的第一项任务，起着至关重要的作用. 为了验证本文提出的方法在论点边界检测任务的有效性，我们对比了传统的机器学习方法条件随机场(Conditional Random Field，CRF)[1]，以及神经网络模型 Bi-LSTM-CRF[16]，Joint RNN Model[14]. 具体的如表 8 所示，可以看到，最好的效果是采用 Joint RNN Model 在 *F*1 值上达到了 0.873. 但是，不管是传统的机器学习方法，还是神经网络的方法 *F*1 值指标上都未达到 0.90. 而本文所提的方法 *F*1 值达到了 0.922，超过了 0.90，取得了最优的效果. 并且本文所提的方法在精确率(*Precision*)、召回率(*Recall*)以及 BIO 标签 3 个类别的 *F*1 值指标上全部优于其他方法，具体的在 BIO 标签三分类测试上，"B"，"I"，"O"的 *F*1 值分别为 0.881，0.942 和 0.881，尤其"B"标签的 *F*1 值比 Joint RNN Model 高出 4.2%，证明了本文提出的模型能够有效的解决论点边界检测任务.

表 8　边界检测任务的实验结果

比较的方法	*F*1	*P*	*R*	*F*1-B	*F*1-I	*F*1-O
CRF	0.867	0.873	0.861	0.809	0.934	0.857
Bi-LSTM-CRF	0.860	0.868	0.853	0.832	0.919	0.828
Joint RNN Model	0.873	0.893	0.857	0.839	0.931	0.848
本文方法	**0.922**	**0.924**	**0.921**	**0.881**	**0.942**	**0.881**

其次，论辩挖掘的第二项任务是论辩部件类型的识别，我们对比了在该任务中被最广泛使用的 SVM 方法，以及基于强化学习的 HAs-augmented RL 模型. 实验结果如表 9 所示，可以看出，本文的方法与传统机器学习标准的 SVM 方法相比，预测的性能有很大的提升，同时与目前公认预测效果较好的 HAs-augmented RL 模型，也有着较大的进步，特别地，在"*F*1-MC"评价指标上，比 SVM 方法提升了 35.3%，比 HAs-augmented RL 提高了 19.1%，这表明了本文的模型在论点部件类型识别任务上具有很大的优势.

表 9　论点部件类型识别任务的实验结果

比较的方法	*F*1	*F*1-MC	*F*1-C	*F*1-P	*F*1-Oth
SVM	0.649	0.377	0.462	0.792	0.965
HAs-RL	0.692	0.539	0.459	0.845	0.925
本文方法	**0.767**	**0.730**	**0.581**	**0.881**	**0.880**

另外，在论点关系抽取任务上，如表 10 所示展示了本文方法与 SVM、ILP 算法在 *F*1 值上的表现，从实验结果得知，本文的提出的方法在 *F*1 值上达到了 0.773，并在"*F*1-Supp"和"*F*1-Att"指标上均优于其他模型，这表明了本文的模型在论点关系类型抽取任务上也具有良好的性能.

表 10　论点关系抽取任务的实验结果

比较的方法	*F*1	*F*1-Supp	*F*1-Att
SVM	0.702	0.946	0.456
ILP	0.752	0.913	0.591
本文方法	**0.773**	**0.934**	**0.613**

综上分析，多任务迭代学习方法训练的神经网络能够有效的利用论辩挖掘三个任务之间的关联信息，互相促进学校，提升泛化效果，使得论辩挖掘的三个任务都达到更好的效果.

7　结束语

本文提出了一个基于多任务迭代学习的论辩挖掘方法，首先，本文将论辩挖掘的三个任务看成有关

联信息的相关任务，改进多任务学习算法，在模型融入迭代的思想，迭代学习不同子任务，将任务一论点边界检测任务的标签分布输出联结为任务二论点类型识别任务的神经网络输入，将论点类型识别任务的标签输出联结为任务三论点关系抽取任务的神经网络输入.该模型很好地解决了论辩挖掘传统流水线方法中，错误传播的问题.其次，模型在词表示中融入了字符级的表示，提取了论辩挖掘文本中字符级的表示特征.最后在预测模型中，融合任务相关的文本结构，文本语义等共享特征.通过与现有方法对比实验表明，本文提出的方法能够有效的解决论辩挖掘的三个任务，并验证了论辩挖掘三个任务之间的关联信息，通过相互学习，能有效得提高模型的泛化效果.在接下来的工作中，我们将继续对论辩挖掘的三个子任务之间的潜在关系进行挖掘，建立性能更优秀的模型，以期进一步地提高标注质量.

参 考 文 献

[1] Stab C, Gurevych I. Parsing argumentation structures in persuasive essays. Computational Linguistics, 2017, 43(3): 619-659

[2] Mochales R, Moens M-F. Argumentation mining. Artificial Intelligence and Law, 2011, 19(1): 1-22

[3] Wachsmuth H, Al Khatib K, Stein B. Using argument mining to assess the argumentation quality of essays//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, 2016: 1680-1691

[4] Habernal I, Gurevych I. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany, 2016: 1589-1599

[5] Eckle-Kohler J, Kluge R, Gurevych I. On the role of discourse markers for discriminating claims and premises in argumentative discourse//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 2236-2242

[6] Stab C, Gurevych I. Identifying argumentative discourse structures in persuasive essays//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, 2014: 46-56

[7] Florou E, Konstantopoulos S, Koukourikos A, et al. Argument extraction for supporting public policy formulation//Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Sofia, Bulgaria, 2013: 49-54

[8] Moens M F, Boiy E, Palau R M, et al. Automatic detection of arguments in legal texts//Proceedings of the 11th International Conference on Artificial Intelligence and Law. Stanford, USA, 2007: 225-230

[9] Feng V W, Hirst G. Classifying arguments by scheme//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics. Portland, USA, 2011: 987-996

[10] Palau R M, Moens M F. Argumentation mining: The detection, classification and structure of arguments in text//Proceedings of the 12th International Conference on Artificial Intelligence and Law. Barcelona, Spain, 2009: 98-107

[11] Rooney N, Wang H, Browne F. Applying kernel methods to argumentation mining//Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference. Florida, USA, 2012: 272-275

[12] Teufel S. Argumentative Zoning: Information Extraction from Scientific Text [Ph.D. dissertation]. University of Edinburgh, Edinburgh, UK, 1999

[13] Wyner A, Mochalespalau R, Moens M F, et al. Approaches to text mining arguments from legal cases//Proceedings of the Semantic Processing of Legal Texts. Berlin, Germany, 2010: 60-79

[14] Li Minglan, et al. Joint RNN model for argument component boundary detection//Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2017). Banff, Canada, 2017: 57-62

[15] Laha A, Raykar V. An empirical evaluation of various deep learning architectures for bi-sequence classification tasks//Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics. Osaka, Japan, 2016: 2762-2773

[16] Eger S, Daxenberger J, Gurevych I. Neural end-to-end learning for computational argumentation mining//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 11-22

[17] Persing I, Ng V. End-to-end argumentation mining in student essays//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 1384-1394

[18] Reed C, Palau R M, Rowe G, et al. Language resources for studying argument//Proceedings of the International Conference on Language Resources and Evaluation, Lrec 2008. Marrakech, Morocco, 2008: 2613-2618

[19] Quinlan J R. C4.5: Programs for Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers, 1993

[20] Gao Y, et al. Reinforcement learning based argument component detection. arXiv preprint arXiv:1702.06239, 2017

[21] Cabrio E, Villata S. Combining textual entailment and argumentation theory for supporting online debates interactions //Proceedings of the Meeting of the Association for Computational Linguistics: Short Papers. Jeju Island, Korea, 2012: 208-212

[22] Peldszus A, Stede M. An annotated corpus of argumentative microtexts//Proceedings of the 1st Conference on Argumentation. Lisbon, Portugal, 2015: 801-815

[23] Ma Xuezhe, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany, 2016: 1064-1074

[24] Niculae V, Park J, Cardie C. Argument mining with structured SVMs and RNNs//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 985-995

[25] Potash P, Romanov A, Rumshisky A. Here's my point: Argumentation mining with pointer networks. arXiv preprint arXiv:1612.08994, 2016

[26] Kirschner C, Eckle-Kohler J, Gurevych I. Linking the thoughts: Analysis of argumentation structures in scientific publications//Proceedings of the Workshop on Argumentation Mining. Denver, USA, 2015: 1-11

[27] Boltužić F, Šnajder J. Back up your stance: Recognizing arguments in online discussions//Proceedings of the First Workshop on Argumentation Mining. Baltimore, USA, 2014: 49-58

[28] Somasundaran S, et al. Evaluating argumentative and narrative essays using graphs//Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, 2016: 1568-1578

[29] Zhang F, Litman D. Using context to predict the purpose of argumentative writing revisions//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 1424-1430

[30] Caruana R. Multitask learning. Machine Learning, 1997, 28(1): 41-75

[31] Hollingshead K, Roark B. Pipeline iteration//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic, 2007: 952-959

[32] Kim Y. Convolutional neural networks for sentence classification //Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). Doha, Qatar, 2014: 1746-1751

[33] Kim Y, Jernite Y, Sontag D, et al. Character-aware neural language models//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 2741-2749

[34] Srivastava R K, Greff K, Schmidhuber J. Training very deep networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 2377-2385

[35] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[36] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 6645-6649

[37] Reimers N, Gurevych I. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 338-348

[38] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014

[39] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1532-1543

LIAO Xiang-Wen, Ph.D., associate professor. His research interests include text orientation retrieval and mining.

CHEN Ze-Ze, M. S. candidate. His research interests include text orientation retrieval and mining.

GUI Lin, Ph. D. His research interest is natural language processing.

CHENG Xue-Qi, Ph. D., professor, Ph. D. supervisor. His research interests include network science, network information security, Web data mining.

CHEN Guo-Long, Ph. D., professor, Ph. D. supervisor. His research interests include artificial intelligence and network security.

Background

With the rapid development of internet technology and social media. People generate a large amount of subjective data such as opinion and commentary. It has enormous commercial and academic value to study these perspective

documents. Argumentation Mining (AM) aims at automatically recognizing argument components and extracting the relationship between them in perspective documents in order to establish new intelligent systems for facilitating information access, writing skills acquisition and text summarization. It has recently become a research hot topic in the field of sentiment analysis. It can be widely used in the judicial, humanistic education user-generated content and other fields to provide people with convenient automated tools.

Argumentation mining typically consists of three consecutive subtasks, i. e., argument component boundary detection, argument component identification, argument component relation identification. Most existing work addressed one of the argumentation mining subtasks separately. Few work solved all of the them by joint model. The joint model usually used pipeline method that trained model independently for each subtask. However, these methods ignored the relationship between the tasks. Moreover, it inevitably existed error propagation and redundant information. Thus, this paper proposed a multi-task iterative learning method for argumentation mining, which could iteratively utilize predicting tags' distribution of each task explicitly to take advantage of the relationship information between the related tasks, reduce error propagation and redundant information generation.

This work is supported by the National Natural Science Foundation of China (Grants No. 61772135 and No. U1605251), the Open Project of Key Laboratory of Network Data Science & Technology of Chinese Academy of Sciences (No. CASNDST201708 and No. CASNDST201606) and the Directors Project Fund of Key Laboratory of Trustworthy Distributed Computing and Service (BUPT) Ministry of Education (No. 2017KF01).