结合标签转移关系的多任务笑点识别方法

张童越[1], 张绍武[1], 徐博[1], 杨亮[1], 林鸿飞[1]†
1.大连理工大学/辽宁省大连市
zty9818@mail.dlut.edu.cn, zhangsw@dlut.edu.cn
xubo@dlut.edu.cn, liang@dlut.edu.cn
hflin@dlut.edu.cn

摘要

幽默在人类交流中扮演着重要角色，并大量存在于情景喜剧中。笑点（punchline）是情景喜剧实现幽默效果的形式之一，在情景喜剧笑点识别任务中，每条句子的标签代表该句是否为笑点，但是以往的笑点识别工作通常只通过建模上下文语义关系识别笑点，对标签的利用并不充分。为了充分利用标签序列中的信息，本文提出了一种新的识别方法，即结合条件随机场的单词级-句子级多任务学习模型，该模型在两方面进行了改进，首先将标签序列中相邻两个标签之间的转移关系看作幽默理论中不一致性的一种体现，并使用条件随机场学习这种转移关系，其次由于学习相邻标签之间的转移关系以及上下文语义关系均能够学习到铺垫和笑点之间的不一致性，两者之间存在相关性，为了使模型通过利用这种相关性提高笑点识别的效果，该模型引入了多任务学习方法，使用多任务学习方法同时学习每条句子的句义、组成每条句子的所有字符的词义，单词级别的标签转移关系以及句子级别的标签转移关系。本文在CCL2020“小牛杯”幽默计算—情景喜剧笑点识别评测任务的英文数据集上进行实验，结果表明，本文提出的方法比目前最好的方法提高了3.2%，在情景喜剧幽默笑点识别任务上取得了最好的效果，并通过消融实验证明了上述两方面改进的有效性。

关键词： 情感分析 ； 幽默计算 ； 多任务学习 ； 条件随机场

Multi-task punchlines recognition method combined with label transfer relationship

Tongyue Zhang[1], Shaowu Zhang[1], Bo Xu[1], Liang Yang[1], Hongfei Lin[1]†
1.Dalian University of Technology/Dalian City, Liaoning Province
zty9818@mail.dlut.edu.cn, zhangsw@dlut.edu.cn
xubo@dlut.edu.cn, liang@dlut.edu.cn
hflin@dlut.edu.cn

Abstract

Humor plays an important role in human communication and is abundant in sitcoms. Punchline is one of a form to achieve humorous effects in sitcoms. In the task of punchlines recognition, the label of each sentence represents whether the sentence is a punchline or not, but the existing punchlines recognition methods only recognize the punchline by modeling the contextual semantic relationship, therefore the use of tags is not sufficient. In order to make full use of the information in tag sequences, this paper proposes a new method named multi-task learning model combined with conditional random field on word-sentence level. Our model has been improved in two aspects. First, we regard the transfer relationship between two tags as a manifestation of inconsistency in humor theory, and we use the conditional random field to learn this

第二十届中国计算语言学大会论文集，第238页-第247页，呼和浩特，中国，2021年8月13日至15日。

transfer relationship. Secondly, learning the transfer relationship between adjacent tags and the contextual semantic relationship can both capture the inconsistency between the setup and punchline, there is a correlation between the two, in order to improve the effect of punchlines recognition by using this correlation, we introduce the multi-task learning method to learn the meaning of each sentence, the meaning of all the characters that make up each sentence, the label transfer relationship at the word level and the label transfer relationship at the sentence level. This paper conducts experiments on the English data set of CCL2020 "Mavericks Cup" humorous calculation-sitcom punchlines recognition and evaluation task. The results show that the method proposed in this paper is 3.2% higher than the current best method and achieve the best effect on the punchlines recognition task, and the ablation experiment proves the effectiveness of the two aspect of improvements.

Keywords: Sentiment analysis , humorous calculation , multi-task learning , conditional random field

1 引言

幽默是人类交流的重要组成部分，情景喜剧作为一种大量包含幽默元素的艺术形式深受人们的青睐，在情景喜剧中，笑点（punchline）是幽默的载体，它是对白中使人发笑的一个或多个句子，是情景喜剧具有幽默性的关键之一。情景喜剧的对白具有复杂的语境，识别对白中的笑点对于提高计算机识别人类幽默的能力有着重要作用，但是从大量的对白中识别出少量的笑点是一项有挑战性的任务。

情景喜剧的对白由大量的笑话组成，Bright等(1992)指出笑话一般是由笑点和笑点的上文铺垫（setup）所组成的，Raskin等(2012)认为"铺垫"的作用是为笑点提供上下文信息，笑点的作用是通过表达与铺垫相违背的语义来产生幽默的效果，这在幽默理论中被称为不一致性(Binsted et al.，2006)。在情景喜剧笑点识别任务中，笑点和非笑点在对白中的分布情况直观反映在句子标签的分布规律上，但是现有的笑点识别方法通常只通过建模铺垫和笑点之间的上下文语义关系分析不一致性并识别笑点，对标签的利用并不充分。为了弥补这一不足，本文引入了在标签层面的分析，将相邻标签之间的转移关系看作不一致性的一种体现。相邻标签之间的转移关系存在于单词级别和句子级别两个粒度上，例如句子A为"You fell Asleep !!"，句子A的下一句句子B为"There was no kangroo !"，句子A为铺垫，标签值为0，句子B为笑点，标签值为1，句子B能够产生和句子A的不一致性。在单词级别，将组成句子的字符（包括单词和标点符号）分为两类，正类代表该字符所在的句子能产生和铺垫的不一致性，标签值为1，其中字符的生成方式将在3.2节作详细介绍，因此句子A中所有字符均为负类，句子B中所有字符均为正类，即每个句子中相邻先后两个字符的标签不会是"1，0"或"0，1"，即标签不会在0和1之间发生转移。在句子级别，句子A和B之间的不一致性直观体现在句子的标签由0变为1的标签转移现象。为了挖掘在单词级别和句子级别同时存在的两种标签转移关系，考虑到条件随机场能够计算相邻标签的转移可能性大小，使用线性链条件随机场(CRF,Linear Chain Conditional Random Field)(Sutton and McCallum，2006)学习上述两种转移关系。

最后，由于学习相邻标签之间的转移关系以及上下文语义关系均能够学习到铺垫和笑点之间的不一致性，两者之间存在相关性，为了利用这种相关性提高模型的性能，本文采用多任务学习方法并定义了四个子任务，如表1所示，将子任务一作为主任务，其他作为副任务，子任务一和子任务三为仅学习词义或句义而不使用条件随机场挖掘标签序列中的信息，子任务二和子任务四为学习词义或句义的同时使用条件随机场挖掘标签序列中的信息，并同时对四个子任务进行学习。

本文的贡献分为以下两点：（1）提出了结合条件随机场的单词-句子级多任务学习方法.在情景喜剧笑点识别任务中，该方法将标签序列中相邻两个标签之间的转移看作幽默理论中的不

基金项目：国家自然科学基金(61632011，62076046，62076051)

一致性理论的一种体现，为了学习相邻标签之间的转移关系，本文结合了神经网络与条件随机场。（2）使用多任务学习方法同时学习四个子任务，并通过实验证明了多任务学习的有效性。

子任务	任务目标
任务一	在单词级别仅学习词义
任务二	在单词级别学习词义并分析标签转移关系
任务三	在句子级别仅学习句义
任务四	在句子级别学习句义并分析标签转移关系

表 1: 子任务描述

2 相关工作

幽默是人们日常交流的重要组成部分，随着机器学习和深度学习技术的发展，幽默识别受到了较多的关注。在以往的工作中，为了识别幽默，Mihalcea 和Strapparava(2005) 提出了四种特征，分别为头韵（西方诗歌的一种押韵形式）、反义词组，同义词组以及上下文特征，并使用这些特征识别文本是否幽默。Yang等(2015)设计了一种用于识别幽默的分类器，该分类器从不协调性、歧义、说话者和倾听者之间的影响以及发音特征四个层面挖掘幽默背后的语义结构。Morales和Zhai(2017) 利用了文本的背景信息，如维基百科的词条描述，并构建多种特征识别互联网上的幽默评论。以上都是基于特征工程的幽默识别方法，随着深度学习技术的发展，基于神经网络的方法同样取得了好的效果，Chen和Soo (2018)使用卷积神经网络进行幽默识别。Zhou(2020) 等指出双关语是实现幽默效果的方式之一，包括谐音相关和语义相关，并使用一种基于BERT和注意力机制的方法同时对单词的音素（根据语音的自然属性划分出来的最小语音单位）和语义进行建模。Xiaochao Fan等(2020)提出的模型结合了卷积神经网络、门控循环神经网络和注意力机制，该模型通过学习语音结构和语义表征的方法进行幽默识别。

笑点是幽默的重要表现形式之一，并广泛存在于对白中，为了识别笑点，以往的工作大多从铺垫和笑点组成的句子对这一角度解决问题。在以往的工作中，Xie等(2020)为了区分笑话和非笑话，使用预训练语言模型学习铺垫和笑点之间的语义关系，并将不一致性扩展到不确定度和惊喜度两个方面，从这两个方面对铺垫和笑点之间语义关系进行评估；受到不一致性理论的启发，Mihalcea等(2010)提出了一种通过计算铺垫和笑点之间的语义关系来识别笑点的方法。Andrew 和Cattle(2016)使用了五种不同的度量方法计算铺垫和笑点之间的语义相关度；Bertero和Fung(2016)使用长短期记忆网络建模铺垫和笑点之间的关系；Choube等(2020)通过融合来自文本、音频和视频三个模态的信息，提高了语义识别的准确性，同时使用门控循环单元进行上下文建模，进一步提高了笑点识别的效果。

综上所述，铺垫和笑点这一概念常被用于笑点识别工作中，但是以往的笑点识别工作通常只通过建模铺垫和笑点之间的上下文语义关系分析不一致性并识别笑点，忽视了标签序列中的信息，导致对标签信息的利用并不充分。

3 模型

3.1 模型概述

为了在标签层面分析铺垫和笑点之间的不一致性，本文提出的模型在单词级别和句子级别挖掘相邻标签之间的转移关系。为了提高模型拟合数据真实分布的能力，使用多任务学习方法使模型同时在标签和单纯语义两个层面进行学习，标签层面指在学习语义的同时使用条件随机场挖掘相邻字符或句子标签之间的转移关系，单纯语义层面指仅学习语义而不使用条件随机场挖掘相邻标签之间的转移关系。模型如图1所示。标签层面的损失函数为$loss_{crf}^{word}$和$loss_{crf}^{utt}$，分别对应任务二和任务四，单纯语义层面的损失函数为$loss^{word}$和$loss^{utt}$，分别对应任务一和任务三。任务一和任务二在单词级别上进行，任务一将共享层的输出送入全连接层以挖掘字符的词义，任务二将全连接层的输出送入条件随机场以学习相邻字符标签之间的转移关系。任务三和任务四在句子级别上进行，首先将共享层的输出中属于同一句子的所有字符的词向量取平均值，得到句向量，任务三将句向量送入全连接层以挖掘句子的句义，任务四将句向量送入另一个全连接层和条件随机场以学习相邻句子标签的转移关系。最后，将任务一的输出标签作为每

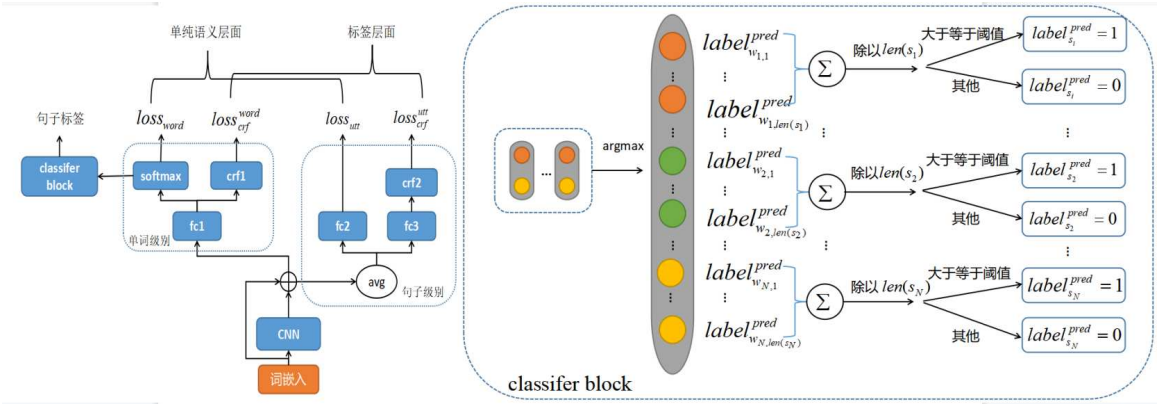

图 1: 结合标签转移关系的多任务笑点识别模型图

个字符的预测标签，并规定句子中被预测为“该字符所在的句子能够产生和铺垫的不一致性”的字符数量和该句字符总数的比值代表该句能够产生和铺垫的不一致性的置信度，当置信度大于一个阈值时，将该句识别为笑点。

3.2 词嵌入层

为了将输入文本转换为模型接受的向量形式，需要对输入文本进行词嵌入。模型的输入为一个由对白中N个句子拼接成的序列$S = [s_1, s_2, ..., s_N]$，使用预训练语言模型对输入序列进行词嵌入，现有的预训练模型包括BERT(Devlin et al., 2018)，XLNet(Yang et al., 2019)，Roberta(Liu et al., 2019)等，本框架使用Roberta作为词嵌入层，同时规定在向预训练模型输入数据时，策略如下：

$$\sum_{i=0}^{N} \text{len}(s_i) < \text{maxlen} \tag{1}$$

其中len(∗)是计算一个句子经过Roberta的分词处理后生成的字符总数的操作，maxlen为超参数，s_i代表输入序列中第i个句子，并且将s_i包含的字符表示为:$W = [w_{i,1}, w_{i,2}, ..., w_{i,\text{len}(s_i)}]$，其中$i \in R^N$，$w_{i,j}$为$s_i$的第j个字符，N表示一个batch包含的句子数量。将输入序列送词嵌入层，得到每个字符的词向量，将词嵌入层的输出定义为$R_{emb} = [R_1, ..., R_N]$，其中$R_i = [v_{i,1}, v_{i,2}, ..., v_{i,\text{len}(s_i)}]$，$v_{i,j} \in R^{b*l*d}$，$v_{i,j}$表示$s_i$的第j个字符所对应的词向量，b为输入的batch数量，l为s_i中包含的字符数量之和，d为Roberta隐藏层大小。

3.3 标签层面

本文将相邻标签之间的转移关系看作幽默理论中的不一致性理论的一种体现，为了在标签层面分析铺垫和笑点之间的不一致性，模型在单词级别和句子级别同时挖掘相邻标签之间的转移关系，本节详细介绍了如何在单词级别和句子级别上挖掘相邻标签之间的转移关系，以及如何设置共享层。

3.3.1 单词级别

在单词级别，标签层面和单纯词义层面共享一部分隐藏层，定义为$share^{word}$，被共享隐藏层的参数同时受到两个层面的影响，使模型学习到更多的信息。词嵌入层中的transformers能够忽略字符之间的距离，并学习句子中全部字符之间的特征依赖关系，为了学习到部分字符之间的局部特征依赖关系，考虑到卷积运算只对卷积核大小范围内的输入进行处理(Gu et al., 2018)，使用卷积神经网络处理输入字符序列的词嵌入向量，即图1中的“CNN”，并使用一个特征组合层将“CNN”的输出向量和词嵌入层的输出相加，公式如下：

$$R_{conv} = \text{ReLU}(\text{conv1d}(R_{emb})) \tag{2}$$

$$R^{word}_{comb} = R_{emb} + R_{conv} \tag{3}$$

其中，R_{emb}是词嵌入层的输出，R_{conv}是卷积神经网络“CNN”的输出，$R_{conv} \in R^{b\times l\times d}$，$R_{emb} \in R^{b\times l\times d}$，$R^{word}_{comb} \in R^{b\times l\times d}$，b为输入的batch数量，l为一个batch中包含的所有句子的字符总和，d为词嵌入层的隐藏层大小，ReLU(∗)代表relu激活函数(Glorot, Bordes, and Bengio, 2011)，conv1d(∗)代表一维卷积层，定义$R^{word}_{comb} = [c_{1,1}, ..., c_{1,len(s_1)}, ..., c_{N,1}, ..., c_{N,len(s_N)}]$，其中，$c_{i,j}$为输入序列的第i个句子的第j个字符的词向量。最后，使用图1中的全连接层“fc1”从R^{word}_{comb}中提取特征，并将共享层$share^{word}$设置为卷积神经网络“CNN”，特征组合层和全连接层“fc1”，共享层的输出被定义为：

$$R^{word}_{share} = \text{Linear1}(R^{word}_{comb}) \tag{4}$$

其中，Linear1(∗)为一层用于二分类的全连接层，即“fc1”。

为了学习相邻字符的标签之间的转移关系，将输出P_{crf1}定义为：

$$P_{crf1} = \text{CRF1}(R^{word}_{share}) \tag{5}$$

其中，CRF1(∗)为线性链条件随机场，P_{crf1}为线性链条件随机场对每个字符预测得到的标签。线性链条件随机场包含发射矩阵和转移矩阵，定义一个发射矩阵$E^{word} = R^{word}_{share}$，$E^{word}_{i,j}$表示输入序列中第i个字符对应的词向量中的第j个元素值，为输入序列定义一个转移矩阵$T^{word}_{2\times 2}$，$T^{word}_{i,j}$表示相邻两个字符的标签由i转移到j的可能性大小，转移矩阵内部的元素值被随机初始化，元素值在训练过程中被更新。定义得分函数：

$$\text{score}(X, y)_{word} = \sum_{i=0}^{n} T^{word}_{y_i, y_{i+1}} + \sum_{i=0}^{n} E^{word}_{i,y_i} \tag{6}$$

其中y_i为输入序列中第i个字符的标签，n为输入序列的长度，得分函数值越大，对输入序列的标签预测的合理性越高。

3.3.2 句子级别

在句子级别，标签层面和单纯词义层面同样共享一部分隐藏层，定义为$share^{utt}$，共享层$share^{utt}$包括单词级别的卷积神经网络“CNN”和特征组合层，以及图1中的取平均值操作“avg”。将共享层的输出定义为：$R^{utt}_{share} = (R^{utt}_{comb})_{avg}$，其中$R^{utt}_{comb} = R^{word}_{comb}$，$(R^{utt}_{comb})_{avg}$为计算$R^{utt}_{comb}$中属于同一个句子的所有字符的词向量的平均值的操作，即图1中的“avg”，$R^{utt}_{share} = [V_{s_1}, V_{s_2}, ..., V_{s_N}]$，$V_{s_i}$为输入序列中第i个句子的句向量，将共享层$share^{utt}$ 设置为卷积神经网络“CNN”，特征组合层以及取平均值操作“avg”。

在句子级别上，为了学习相邻句子的标签之间的转移关系，将输出定义为：

$$R_{uttcrf} = \text{Linear3}(R^{utt}_{share}) \tag{7}$$

$$P_{crf2} = \text{CRF2}(R_{uttcrf}) \tag{8}$$

Linear3(∗)为一层用于二分类的全连接层，即图1中的“fc3”，P_{crf2} 为条件随机场为每一个句子预测得到的标签。CRF2(∗)为线性链条件随机场，同样包含发射矩阵和转移矩阵，定义一个发射矩阵$E^{utt} = R^{utt}_{share}$，$E^{utt}_{i,j}$ 表示输入序列中第i个句子对应的句向量中的第j个元素值，为输入序列定义一个转移矩阵$T^{utt}_{2\times 2}$，其中$T^{utt}_{i\times j}$表示相邻句子的标签由i转移到j的的可能性大小，转移矩阵内部的元素值被随机初始化，元素值在训练过程中被更新。定义得分函数：

$$\text{score}(X, y)_{utt} = \sum_{i=0}^{n} T^{utt}_{y_i, y_{i+1}} + \sum_{i=0}^{n} E^{utt}_{i,y_i} \tag{9}$$

其中y_i为输入序列中第i个字符的标签，n为输入序列的长度，得分函数值越大，对输入序列的标签预测的合理性越高。

3.4 单纯语义层面

模型在单词级别和句子级别同时挖掘铺垫和笑点在单纯语义层面的关系，本节详细介绍了如何在单词级别挖掘词义和在句子级别挖掘句义，以及如何设置共享层。

3.4.1 单词级别

为了使模型学习到组成句子的字符之间的词义关系，在单词级别上，根据字符的分类结果对每个句子进行分类，将共享层$share^{word}$的输出R^{word}_{share}作为代表词义的特征，使用该特征对字符进行分类，公式为：

$$label^{pred}_{w_{i,j}} = \text{argmax}(\text{softmax}((R^{word}_{share})_{i,j})) \tag{10}$$

其中，$label^{pred}_{w_{i,j}}$为句子s_i的第j个字符的预测标签，$(R^{word}_{share})_{i,j}$表示R^{word}_{share}中第i个句子的第j个字符对应的词向量。定义当句子s_i的所有字符的预测标签之和占该句字符总数的比值大于一个阈值时，该句为笑点，否则不是笑点。

3.4.2 句子级别

为了使模型学习到对白中不同句子之间的句义关系，在句子级别上，将句子级别的共享层$share^{utt}$的输出R^{utt}_{share}送入一层全连接层，将输出定义为：

$$R_{utt} = \text{Linear2}(R^{utt}_{share}) \tag{11}$$

由于数据集中正样本数量远少于负样本数量，因此将二分类问题转化异常检测问题进行处理(Tax and Duin，1999)，其中Linear2(∗)代表用于异常检测的全连接层，即图1中的“fc2”。

3.5 训练和预测

为了使模型同时学习标签层面和单纯语义层面的信息，提高模型拟合真实数据分布的能力，使用多任务学习方法，同时对四个任务进行学习，提高模型的泛化性。任务一为学习字符的词义，损失函数采用交叉熵损失函数，并将损失函数定义为$loss^{word}$，由于数据集中只有每个句子的标签，因此将一个句子中所有字符的真实标签定义为该句的真实标签，字符的标签为1代表该字符所在的句子能够产生和上文的不一致性，否则相反。

任务二为学习不同类别字符的转移关系，损失函数为线性链条件随机场产生的损失函数，将损失函数定义为：

$$loss^{word}_{crf} = \frac{\sum_{i=1}^{batch}(-\text{score}(X_i, y^{true}_i)_{word} + \log(\sum_{y\in Y_{X_i}} e^{\text{score}(X_i,y)_{word}}))}{\sum_{i=1}^{batch}\text{len}(X_i)} \tag{12}$$

其中batch为每次输入模型的批次大小，$\text{score}(*)_{word}$为得分函数，X_i为第i个batch的输入字符序列，y^{true}_i为第i个batch中全部字符的真实标签序列，Y_{X_i}为预测得到的各种可能的标签，$Y_{X_i} \in \{0, 1\}$，len(∗)为计算X_i中包含的字符总数的操作。

任务三为学习句子的句义，由于数据集中存在正负样本数量不均衡问题，因此使用GHM损失函数(Li, Liu, and Wang，2019)，并将损失函数定义为$loss^{utt}$，GHM损失函数通过衡量一定梯度范围内的正负样本数量，使数量较多的类别对应的样本权重下降，让模型能够更多地关注数量较少的样本。

任务四为学习相邻句子的标签转移关系，损失函数定义为：

$$loss^{utt}_{crf} = \frac{\sum_{i=1}^{batch}(-\text{score}(X_i, y^{true}_i)_{utt} + \log(\sum_{y\in Y_{X_i}} e^{\text{score}(X_i,y)_{utt}}))}{\sum_{i=1}^{batch}\text{len}(X_i)} \tag{13}$$

其中batch为每次输入模型的批次大小，$\text{score}(*)_{utt}$为得分函数，X_i为第i个batch的输入句子序列，y^{true}_i为第i个batch中全部句子的真实标签序列，Y_{X_i}为预测得到的各种可能的标签，$Y_{X_i} \in \{0, 1\}$，len(∗)为计算X_i中包含的句子总数的操作。最后，将用于模型训练的损失函数定义为：

$$loss_{4task} = loss_{word} + loss^{word}_{crf} + loss_{utt} + loss^{utt}_{crf} + \lambda \mid\mid \theta \mid\mid^2 \tag{14}$$

其中，λ是权重衰减系数，θ是所有可训练参数。

		说话者	文本	句子标签
对白A	样本1	Joey	Come on, Lydia, you can do it.	0
	样本2	Joey	Push!	1
	样本3	Joey	Push 'em out, push 'em out, harder, harder.	1
	样本4	Joey	Push 'em out, push 'em out, way out!	1
	样本5	Joey	Let's get that ball and really move, hey, hey, ho, ho.	1
	样本6	Joey	Let'sI was justyeah, right.	0
	样本7	Joey	Push!	1

表 2: 对白结构示例

4 实验

4.1 数据集

本文在CCL2020"小牛杯"幽默计算—情景喜剧笑点识别评测任务的英文数据集上进行实验，该数据集来自电视剧《老友记》。在数据集中，根据场景的变换，将情景剧的对话结构分为对白（Dialogue）和句子（Utterance）两个层级。对白层级以一段独立的对白为单位，每段对白包含不同数量的样本，句子层级以一条样本为单位，每条样本为一个句子，如表2所示，对白A来自该数据集，包含7条样本，每条样本为一条句子。数据集中每条样本带有一个标签，标签为"0"表示该样本不是笑点，为"1"表示该样本是笑点，训练集包含7472条样本，550段对白，其中包括1773条笑点和5699条非笑点，测试集包含2096条样本，157段对白，如表3所示。

训练集				测试集	
正样本	负样本	样本数量	对白数量	样本数量	对白数量
1773	5699	7472	550	2096	157

表 3: 数据集规模

4.2 评价指标

情景喜剧被划分成若干段对白（Dialogue），每段对白由不同数量的句子（Utterance）组成，考虑到每段对白包含的句子数量不同，评测官方将F1+acc的值作为最终评价指标，该值的大小代表笑点识别效果的好坏[0]，其中"F1"为句子层级的F1分数，F1分数为所有正类样本的召回率（recall）和准确率（precision）的调和平均数；"acc"为对白层级的精确率，精确率的计算方法为首先得到每段对白的精确率，再对所有对白的精确率取平均值。

4.3 基线模型

基线模型如表3所示，第一类是自设计的基线模型，BERT代表使用预训练BERT模型对输入序列进行词嵌入，并使用全连接层和softmax函数为每个句子进行分类，TextCNN(Rakhlin，2016)和LSTM(Hochreiter and Schmidhuber，1997)被用于提取句向量中的幽默特征，bilstm使用双向LSTM从前到后和从后到前对输入句子序列的句向量进行编码，最后，所有自设计基线模型均使用预训练BERT作为词嵌入层。第二类是CCL2020任务3参赛队伍中在英文数据集上取得前六名的队伍所采用的方案[1]，第六名使用门控循环单元对句子进行上下文建模。第五名构建了四个基于长短期记忆网络，卷积神经网络和多头注意力机制的不同模型，并对四个模型进行模型融合。第四名使用了多种数据增强方法，并使用模型融合方法融合由不同数据训练得到的模型。第三名使用多任务学习方法同时对句子是否为笑点以及当前句子的说话者进行预测。第二名将命名实体识别的思想应用于笑点识别任务上来，通过判断一句话中所有预测标签为1的字符总数和该句长度的比例得到该句是否为笑点，本文的模型是在该方法上作出的改进。第一名使用预测下一句的方法对句子进行分类，在将数据输入到预训练模型时，输入策略为将待分类句的前十句和后两句作为输入的第一个序列，待分类句作为第二个序列。

[0]http://cips-cl.org/static/CCL2020/humorcomputation.html
[1]https://github.com/DUTIR-Emotion-Group/CCL2020-Humor-Computation

	基线模型	precision	recall	F1	acc	F1+acc
自设计模型	BERT	-	-	0.513	0.744	1.257
	TextCNN	-	-	0.519	0.756	1.275
	LSTM	-	-	0.510	0.745	1.255
	biLSTM	-	-	0.524	0.762	1.286
参赛队伍模型	第六名	-	-	0.527	0.763	1.290
	第五名	-	-	0.565	0.762	1.327
	第四名	-	-	0.555	0.772	1.327
	第三名	-	-	0.567	0.764	1.331
	第二名	-	-	0.576	0.778	1.354
	第一名	**0.599**	0.573	0.586	**0.783**	1.369
本文提出的模型		0.567	**0.700**	**0.627**	0.774	**1.401**

表 4: CCL2020任务3英文数据集实验结果

模型	F1	acc	F1+acc
结合条件随机场的单词-句子级多任务学习模型	**0.627**	0.774	**1.401**
w/o 任务二	0.597	0.781	1.378
w/o 任务三	0.617	0.770	1.387
w/o 任务四	0.625	0.768	1.393
w/o 任务四+任务二	0.590	0.777	1.367
w/o 任务四+任务二+任务三	0.581	**0.783**	1.364

表 5: 消融实验结果

4.4 实验细节

实验在Pytorch环境下进行。优化器AdamW。为了减少过拟合现象的发生，采用权重衰减策略，权重衰减系数为0.01。学习率为5×10^{-6}，同时采用学习率衰减策略，模型每处理300个batch学习率变为当前值的0.8倍。dropout设置为0.1。batchsize为1，即每次向模型输入由N条句子组成的单个序列，并规定输入序列经过预训练模型分词后的字符总数必须小于256。在词嵌入层使用的预训练模型为包含24个transformer(Vaswani et al., 2017)的Roberta（large），Roberta（large）的隐藏层大小为1024，并将最后一个transformer的输出向量作为词嵌入层的输出。卷积神经网络的卷积层层数为1，同时为了保证输出向量的维度与输入向量保持一致，卷积层的卷积核大小设置为3，步长设置为1，卷积核数量与词嵌入层的隐藏层大小相同。当threshold设置为0.3时模型的性能相对最佳。

4.5 实验结果分析

表4展示了本文提出的方法和基线模型的比较结果。在自设计的基线模型中，使用TextCNN挖掘句向量之间的短距离上下文语义关系，以及使用LSTM和双向LSTM（biLSTM）捕捉句向量之间的长距离上下文语义关系，虽然LSTM可以弥补TextCNN只能提取短距离特征这一不足，但是会更多地关注后输入的信息从而导致丢失部分信息，双向LSTM通过对输入序列进行从前到后和从后到前的编码解决了丢失信息的问题，但是由于挖掘幽默特征较为困难，而且数据集中正负样本数量差别大，因此使用传统神经网络作为分类器无法取得理想的识别效果。在参赛队伍使用的六种模型中，为了弥补传统神经网络在笑点识别工作中的不足，使用了BERT或基于BERT改进的预训练语言模型获得单词或句子向量，使用传统或人工设计的分类器提取语义特征并对句子进行分类，同时为了进一步提高笑点识别性能，第一名，第四名和第五名使用模型融合方法中的投票法提高输出结果的准确性，第三名使用多任务学习方法提高模型的泛化性。

和所有基线模型相比，本文提出的结合条件随机场的单词-句子级多任务学习模型将在标签层面的分析引入到笑点识别任务中，并使用多任务学习方法融合来自标签层面和单纯语义层面的信息，不同的子任务为模型训练提供了不同的噪声，噪声的存在能够提高模型拟合真实数据分布的能力，使模型从训练数据中学习到更具一般性的表征，提高模型的泛化性，超过了所有

基线模型中的最好模型。和所有基线模型相比，本文提出的模型在F1分数（F1）上比基线模型中最好的方法提高了4.1%；在精确率（acc）上比基线模型中的“第一名”低0.9%，原因是模型虽然提高了将正样本正确识别的概率，因此使正类的召回率（recall）比第一名高12.7%，但是将更多的负样本错误识别为正样本，因此正类的查准率（precision）比第一名低3.2%，导致精确率（acc）低于第一名。由于数据集由大量独立的对白组成，且每段对白包含的句子数量不同，只在句子层级或对白层级分析笑点识别效果无法准确判断模型的性能，因此遵循评测官方的规定，将F1分数和精确率之和“F1+acc”作为最终评价指标，本文提出的模型在“F1+acc”上比基线模型中最好的方法提高了3.2%，证明使用多任务学习方法将在标签层面的分析融入到情景喜剧笑点识别任务中能够有效提高笑点识别的性能。

为了更好地证明模型中各部分的作用，本文进行了消融实验研究，如表5所示，只删除任务二，任务三或任务四会使模型的性能分别下降2.3%，1.4%和0.8%，证明任务二对提高模型性能的贡献最大，删除任务四和任务二以及删除任务四任务二和任务三同样会导致模型的性能分别下降3.4%和3.7%，证明在情景喜剧笑点识别任务中，使用线性链条件随机场在单词级别学习相邻字符的标签转移关系以及在句子级别学习相邻句子的标签转移关系能使模型获得对提高笑点识别效果有用的信息，同时随着子任务数量的减少，模型的性能随之下降，证明多任务学习方法在笑点识别任务中的有效性以及学习标签序列中的信息和学习语义信息之间具有相关性。

5 结论

为了识别情景喜剧对白中的笑点，本文提出了结合条件随机场的单词级—句子级多任务学习模型，该模型将将标签序列中相邻两个标签之间的转移看作幽默理论中的不一致性理论的一种体现，并使用多任务学习方法同时在标签层面和单纯语义层面进行分析，通过和六个基线的对比实验以及消融实验证明了该模型在情景喜剧笑点识别工作中是有效的。在今后的工作中，将考虑通过将外部知识引入到情景喜剧笑点识别任务中提高模型的性能。

References

[1] Dario Bertero and Pascale Fung. “A long short-term memory framework for predicting humor in dialogues”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016, pp. 130–135.

[2] Kim Binsted et al. “Computational humor”. In: *IEEE Intelligent Systems* 21.2 (2006), pp. 59–69.

[3] William Bright. “International Encyclopedia”. In: *Psychology* 9 (1992), p. 151.

[4] Andrew Cattle and Xiaojuan Ma. “Effects of semantic relatedness between setups and punchlines in twitter hashtag games”. In: *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*. 2016, pp. 70–79.

[5] Peng-Yu Chen and Von-Wun Soo. “Humor recognition using deep learning”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 2018, pp. 113–117.

[6] Akshat Choube and Mohammad Soleymani. “Punchline Detection using Context-Aware Hierarchical Multimodal Fusion”. In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. 2020, pp. 675–679.

[7] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[8] Xiaochao Fan et al. “Phonetics and ambiguity comprehension gated attention network for humor recognition”. In: *Complexity* 2020 (2020).

[9] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. “Deep sparse rectifier neural networks”. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 315–323.

[10] Jiuxiang Gu et al. “Recent advances in convolutional neural networks”. In: *Pattern Recognition* 77 (2018), pp. 354–377.

[11] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[12] Buyu Li, Yu Liu, and Xiaogang Wang. “Gradient harmonized single-stage detector”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 8577–8584.

[13] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[14] Rada Mihalcea and Carlo Strapparava. “Making computers laugh: Investigations in automatic humor recognition”. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. 2005, pp. 531–538.

[15] Rada Mihalcea, Carlo Strapparava, and Stephen Pulman. “Computational models for incongruity detection in humour”. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. 2010, pp. 364–374.

[16] Alex Morales and ChengXiang Zhai. “Identifying humor in reviews using background text sources”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 492–501.

[17] A Rakhlin. “Convolutional Neural Networks for Sentence Classification”. In: *GitHub* (2016).

[18] Victor Raskin. *Semantic mechanisms of humor*. Vol. 24. Springer Science & Business Media, 2012.

[19] Charles Sutton and Andrew McCallum. “An introduction to conditional random fields for relational learning”. In: *Introduction to statistical relational learning* 2 (2006), pp. 93–128.

[20] David MJ Tax and Robert PW Duin. “Support vector domain description”. In: *Pattern recognition letters* 20.11-13 (1999), pp. 1191–1199.

[21] Ashish Vaswani et al. “Attention is all you need”. In: *arXiv preprint arXiv:1706.03762* (2017).

[22] Yubo Xie, Junze Li, and Pearl Pu. “Uncertainty and Surprisal Jointly Deliver the Punchline: Exploiting Incongruity-Based Features for Humor Recognition”. In: *arXiv preprint arXiv:2012.12007* (2020).

[23] Diyi Yang et al. “Humor recognition and humor anchor extraction”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 2367–2376.

[24] Zhilin Yang et al. “Xlnet: Generalized autoregressive pretraining for language understanding”. In: *arXiv preprint arXiv:1906.08237* (2019).

[25] Yichao Zhou et al. “” The Boating Store Had Its Best Sail Ever”: Pronunciation-attentive Contextualized Pun Recognition”. In: *arXiv preprint arXiv:2004.14457* (2020).