文章编号：1003-0077（2018）00-0000-00

基于卷积神经网络与篇章结构的足球新闻自动生成方法*

刘茂福，齐乔松，胡慧君+

（1. 武汉科技大学计算机科学与技术学院，武汉 430065，中国；2. 智能信息处理与实时工业系统湖北省重点实验室，武汉 430065，中国）

摘要：当前的足球比赛新闻通常是由专家或记者手工撰写的，足球比赛新闻的手工写作费时而且低效。随着在线直播平台与社交媒体的流行，体育网络直播脚本大幅增加；但网络直播脚本通常只记载一场比赛的流水，具有冗长且重点模糊特性，不适宜于赛后直接阅读。为了解决以上问题，在比赛之后，可以基于直播脚本撰写和发布足球比赛新闻。因此，本文提出一种从网络直播脚本直接生成足球比赛新闻的方法。该方法基于卷积神经网络和足球新闻篇章结构，从足球比赛过程中的多个时间段提取出已发生的重要事件，进而抽取相关句子来生成足球新闻，同时，此方法还会针对比赛评价生成一个简短总结。实验结果表明，使用本文提出的方法从网络直播脚本生成足球新闻是可行的。

关键词：足球新闻生成；卷积神经网络；篇章结构；句子抽取；句子生成

中图分类号：TP391 **文献标识码：**A

The Automatic Generation of Football News Based on Convolutional Neural Networks and Discourse Structure

LIU Maofu, QI Qiaosong, HU Huijun+

(1. College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China; 2. Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430065, China)

Abstract: The football news is usually written by experts or journalists, and however, the handwriting of the football news is time-consuming and inefficient. With the popularity of live platforms and social media, the live broadcast script from sport webcast has increased dramatically. But the live broadcast script usually only records the playing information of a match, which is lengthy and vague, and not suitable for direct reading after the match. In order to solve the above problems, after the football match, we can write and publish the news of football match based on live broadcast script. Therefore, this paper proposes a method of directly generating news from football live broadcast script. The method is based on the convolution neural network and the structure of football news text, and it can locate important events from multiple periods in the football match, and then extract relevant sentences to generate football news. Moreover, this method will also generate a brief summary to the match comments. The experimental results show that it is feasible to use the proposed method in this paper to generate news of football match from the live broadcast script.

Key words: Football news generation; Convolution neural network; Discourse structure; Sentence extraction; Sentence generation

1 引言

足球赛事直播正处于蓬勃发展阶段，但受限于生活与工作的快节奏，足球爱好者无法拥有充足的时间观看所有足球赛事直播或重播，取而代之的是通过阅读一篇简短的足球新闻，

*基金项目：国家社会科学基金重大项目（11&ZD189）；湖北省教育厅科学技术研究计划项目（B2016010）；湖北省教育厅人文社会科学研究项目（17Y018）；武汉市科学技术计划项目（2016060101010047）
作者简介：刘茂福（1977－），男，博士，教授，博导，研究方向为计算语言学、自然语言处理；齐乔松(1993-－)，男，硕士，研究方向为自然语言处理；胡慧君（1976－），女，博士，副教授，研究方向为智能计算，+通信作者 E-mail: huhuijun@wust.edu.cn。

获取比赛中发生的事情。然而，时至今日，足球新闻依然由专家或记者手工撰写，耗时费力。因而，采用相关信息抽取、自然语言处理等技术，从体育赛事直播脚本自动生成足球新闻，显得尤为重要。本文选择足球领域的网络直播脚本作为语料，试图提出一个基于卷积神经网络与篇章结构的足球新闻自动生成方法，尝试取代手工撰写新闻的方式。

本文提出的足球新闻生成方法是从一组网络直播脚本中抽取并生成句子。在传统的文档摘要中，重要的词或句子在文档中会更为频繁的出现，但在一场体育比赛的网络直播脚本中，尤其是足球之类的比赛，激动人心的时刻往往十分罕见并且难以复现，而高频的词语或者句子则通常更为平淡。当一个足球迷在阅读一篇足球新闻时，一些关键的信息也许会给他留下深刻的印象，比如进球、球星表现、禁区内激烈对抗、双方球队概述。这就要求新闻生成方法能够从网络直播脚本中找出关键信息，并且抽取出包含这些关键信息的句子。由于足球新闻存在着上述特性，本文利用这种特性作为评价时的辅助指标，以期取得更好的摘要效果。

文档摘要生成技术在专业领域发展迅速，Wang 等人采取了统计方法为中文新闻文章生成单文本摘要[1]。林莉媛等人提出了一种基于评论质量的多文本情感摘要[2]。李培等人基于斯坦纳树使用最小权重支配集在既定的微博数据集上自动生成故事线[3]。

本文使用基于文本句子分类的抽取模型预选出包含重要事件的句子。目前，文本分类技术已有成熟的体系，段旭磊等人使用句向量计算微博文本相似度[4]；吕超镇等人基于文档主题生成模型特征扩展的方法解决短文本分类任务[5]；陈宇等人使用基于差分演化优化极端学习机的分类算法解决林业信息分类任务[6]。

随着计算机性能发展与语料增多，近些年来深度学习在自然语言处理领域发展迅速，各种新颖的研究方法百花齐放。Wan 等人使用抽取多个候选句并对候选句排序的方法来生成跨语种的文本摘要[7]。Cao 等人使用基于神经网络的抽象摘要方法来取代抽取式摘要，得到了忠实原文的摘要结果 [8]。Cao 等人使用文本分类的方法来解决多文档摘要中数据匮乏的问题[9]。Bahdanau 等人最早提出注意力机制在自然语言处理领域的应用，并将其应用于基于编解码模型的机器翻译，取得了突破性的成果[10]。Sabour 等人提出了改进卷积神经网络的思路，使用新的胶囊网络来做分类研究[11]。

近些年来，很多知名媒体企业在非通用领域都先后研发了自动写作机器人，美联社半自动化写作机器人 WordSmith 以财经题材为主生成文章[1]，财经领域的特殊性要求文章内容要以数据挖掘结果为主，文章中会出现大量的数字，对前后文逻辑关联性要求较高。洛杉矶时报的写作机器人 Quakebot 则是以实时发布地震消息为主[2]，曾经在洛杉矶地震三分钟后就发布了新闻，这种应对突发情况的业务场景对自动生成技术的实时性与准确性要求比较高，生成文章也应该相对简明扼要，突出重点。而在国内也有成熟的文章自动生成产品，如腾讯的 Dreamwriter[3]、阿里巴巴的 DT 稿王[4]、今日头条的 Xiaomingbot[5]等。这些写作机器人都对特定的场景有着特定的要求，这种差异化也使得特定领域的文章自动生成技术呈现出百花齐放的局面。

针对足球新闻，本文提出了一个基于卷积神经网络与篇章结构的足球新闻自动生成方法。首先人工对直播文本中的句子进行抽取性标注，基于统计结果抽取文本中含有的人工特征，根据标注结果与人工特征的性质对特征进行处理。使用了词向量、人工特征与标注结果训练卷积神经网络分类模型，使用分类模型预测句子是否应该被抽取；另一方面使用训练集中的数据统计文件来生成关于队伍和球员表现总结的句子。最终这些句子将会被按照训练集结果中篇章结构来重新组合，最终生成一篇足球新闻。

---

[1] https://automatedinsights.com/case-studies/associated-press
[2] http://knowledge.wharton.upenn.edu/article/will-robot-journalists-replace-humanl-ones/
[3] http://tech.qq.com/dreamwriter.htm
[4] http://writingmaster.cn/
[5] http://www.wanfangdata.com.cn/details/detail.do?_type=perio&id=zgcmkhj201609002

2 总体框架

本文中提出的足球新闻自动生成方法主要包含数据预处理、特征与分类、规则与统计文件、篇章结构四个模块，如图 1 所示。

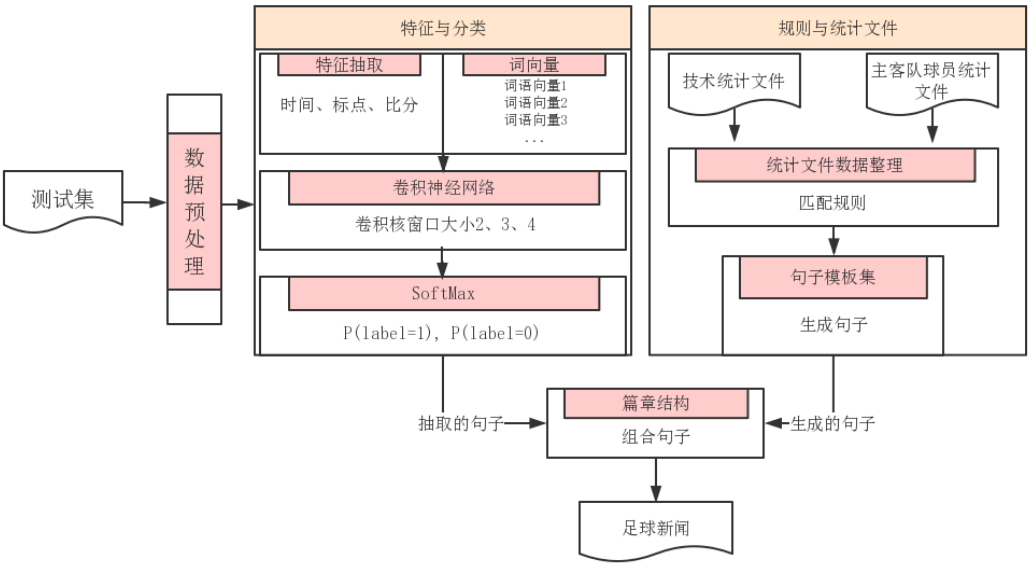

图 1　方法框架图

在数据预处理部分，该方法的主要工作是中文分词以及去停用词。本文选择了中科院的中文分词工具[6]和哈尔滨工业大学的停用词表[7]。

在特征与分类模块中，本文从测试集中抽取时间、标点与比分三类句子特征，时间特征是指一个句子出现的时间点；标点特征反映了一个句子中包含的标点情况；进球特征是指一个句子是否包含进球信息。在词向量特征模块中，本文使用一系列的词向量来表示一个句子。随后，该方法使用卷积神经网络 CNN（Convolutional Neural Networks）模型，对输入的人工特征与词向量特征加以预测。CNN 的预测结果为 SoftMax 二分类，两个结果分别表示输入句子应该被抽取（label=1）和不应该被抽取（label=0）的概率，当$P_{label=1} > P_{label=0}$时，表示抽取当前句子。

在规则与统计文件模块中，该方法通过处理统计统计文件和主客队球员统计文件，得到了双方球队和球员有关进攻与防守的数据。同时在句子生成模块中，本文使用了这些数据中的进球次数和控球率来衡量一个球队的进攻和控球能力，并且使用模板为球队生成一个简短的比赛评价；同时，该方法也为有进球的球员和扑救次数较多的守门员生成评价。

在篇章结构模块中，该方法将会遵照训练集的结果文件格式，把生成的句子置于文章的开头部分，再把抽取后筛选出的句子按照时间顺序来排序，置于文章开头部分之后的位置。本文将得到的句子按照篇章结构重新组合后，最终得到一篇足球新闻。

3 足球新闻自动生成

3.1 篇章结构

根据训练集结果文件中的足球新闻，本文把一篇足球新闻的篇章结构划分成如下四个部分。

[6]中科院分词工具：http://ictclas.nlpir.org/downloads
[7]哈工大停用词表：https://github.com/uk9921/StopWords

（1）时间、比赛和队伍

足球新闻的第一部分一般会说明比赛时间、地点、场次以及球队双方历史对阵情况等，如例 1 所示。

例 1：

北京时间 2 月 3 日凌晨 3:45，英超第 24 轮一场焦点战，阿森纳主场出战南普顿。

在直播脚本文件中，比赛时间与场次可以在其开头部分被找到，而球队数据则一般在其尾部。本文综合这些数据后，可以生成足球新闻第一部分内容。

（2）球员和全场比赛概述

在第二个部分，足球新闻将会展示比赛情况和表现杰出的球员，下面的例 2 说明了这个部分的内容。

例 2：

本场比赛中，双方均有多次破门的机会，且在球权上争夺激烈。伯拉姆-迪乌夫为斯托克城奉献了 1 粒进球。特劳雷为切尔西奉献了 1 粒进球。切尔西队门将库尔图瓦表现神勇，全场没收了 4 次射门。联赛首回合的交锋，切尔西也在客场 0：1 不敌对手。

例 2 中对比赛的评价总结来自比赛技术统计文件，对球员的总结则生成记录了主客队球员统计文件。表 1 和表 2 说明了这两种文件的数据格式和内容。

表 1　技术统计文件示例

切尔西	项目	桑德兰
17	总射门	11
12	犯规	12
67.50%	控球率	32.50%

表 2　主客队球员统计文件示例

位置	球员名	出场	时间	进球	助攻	威胁球	射门	扑救
门将	库尔图瓦	首发	90′	0	0	0	0	2
中场	威廉	首发	90‘	0	1	4	3	0
前锋	科斯塔	首发	76‘	0	0	0	2	0

（3）直播精彩时刻

足球新闻中的第三个部分是比赛中的主体部分，这个部分记录了直播中发生的所有重要时刻，表 3 给出了直播脚本的数据格式。

表 3　直播脚本的部分内容

内容	时间	比分
佩德罗将球一拨，左脚抽射，打进！！！	上半场 14‘	2-0
小法拿球内切，斜塞禁区找插入的威廉	上半场 15‘	2-0
球被回防的范安霍尔特抢先出脚碰出底线，角球	上半场 15‘	2-0

（4）双方出场名单

足球新闻的最后部分是双方球队的阵容，阵容中包含有双方出场球员、球员编号、球员出场时间等信息。如例 3 所示。

例 3：

阿森纳首发（4-2-3-1）：33-切赫；24-贝莱林；20-弗拉米尼（85’，科奎林）……

3.2 特征工程

本文在特征的不同垂直领域上，考虑到特征直观上的合理性，使用了时间、标点与比分

三类特征，并且通过数据统计对这些特征进行了合理性验证。其中，本文统计了训练集中足球新闻的句子与时间分布，不同时间占比如图 2 所示。

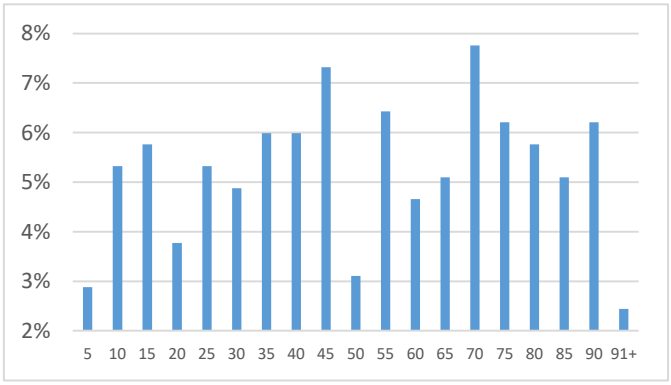

图 2　足球新闻中句子占比与时间分布

从图中可以看出，时间与句子的重要性呈现出明显的相关性，例如上下半场快结束的时候，句子的重要性会持续增高。基于 CNN 的分类没有直接使用到时序信息，引入时间特征可以一定程度上弥补时序信息的不足。训练数据中的时间是连续单位，而实质上时间本身的四则运算是没有实际含义的（如第 1 分钟与第 2 分钟本质上应该是并列的，而非（“$2-1=1$”的关系），本文将时间特征离散化，并表示为编码形式[12]，如表 4 所示。

表 4　部分时间特征表

时间/min	特征表示
[1-5]	(1，0，0，0，0，0，…)
[6-10]	(0，1，0，0，0，0，…)
[11-15]	(0，0，1，0，0，0，…)

根据足球领域的经验，进球往往是一场比赛中最重要的时刻。因此本文使用与上一句相比，本句比分是否发生变化来表示比分特征，对其进行编码表示，如表 5 所示。

表 5　比分特征表

比分是否变化	特征表示
是	(1，0)
否	(0，1)

本文统计了在直播文本中标注结果为“抽取”的句子中标点符号的分布，计算其 TF-IDF 值，如公式（1）、（2）与（3）所示。

$$\mathrm{TF}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

$$\mathrm{IDF}=\log\frac{|D|}{D_i+1} \tag{2}$$

$$\mathrm{TF\_IDF}=\mathrm{TF}\times\mathrm{IDF} \tag{3}$$

其中，$n_{i,j}$代表抽取句子中包含某种标点符号的数量，$\sum_k n_{k,j}$表示抽取句子中包含的标点符号总数量，$|D|$代表需要抽取的句子总数目，D_i代表需要抽取的句子中包含某标点符号的句子数。本文统计了上述三种标点符号的 TF-IDF 值，其结果如表 6 所示。

表 6　标点与 TF-IDF

标点符号	TF-IDF
逗号（，）	-0.00663
句号（。）	-0.000132
感叹号（！）	0.735758

由公式（1）、（2）与（3）可知，TF-IDF 值衡量了待抽取句子中标点符号的重要性，它们的值反映了标点符号的类型与句子抽取与否的相关程度，句子中标点符号与句子是否需要被抽取的相关性越高，其 TF-IDF 值也越高。从表中可以发现，逗号和句号与句子抽取几

乎不具有相关性，而感叹号对句子抽取影响较为显著。感叹号一般表示较为强烈的情感，跟进球、扑救、射失等行为相关，情绪激烈的语句往往包含多个感叹号，因此本文使用包含感叹号的个数来表达这类特征。本文将感叹号的个数转化为离散化向量，使用独热编码表示为（0，0，0，0，1）的形式，如表 7 所示。

表 7　符号特征表示

感叹号个数	特征向量
0	（0，0，0，0，1）
1	（0，0，0，1，0）
2	（0，0，1，0，0）
3	（0，1，0，0，0）
>3	（1，0，0，0，0）

3.3 基于卷积神经网络的句子抽取与生成

为了增强句子的可读性和连贯性，本文使用了一系列的句子模板。下面的表 8 在细节上描述了本文中的模板匹配策略。

表 8　模板匹配的一些例子

类别	关系	模板
总射门	主队的 70%>客队	本场比赛中，主队进攻意识强烈
射正次数	主队>4 且 客队<4	主队（并且）有多次破门的机会
	主队>4 且 客队>4	双方均有多次破门机会
控球率	主队的 50%>客队	主队（并且）在本场比赛中展现出惊人的控球能力
	主队的 70%>客队	主队（并且）在本场比赛中的控球能力略胜一筹
	其它	双方在球权上争夺激烈

本文使用模板来生成篇章结构的第一和第三部分；篇章结构的第二部分则使用基于 CNN 的句子抽取方法，使用 CNN 主要具有三方面的优势。

（1）CNN 可以很好地把以句子为单位的特征与句子在同一个层次结合；

（2）训练集规模较小的情况下，CNN 可以支撑端到端的文本深度生成模型；

（3）基于排序的句子抽取方法需要指定合理的抽取阈值，而 CNN 可以将其转换为基于二分类的句子抽取。

在卷积神经网络模型中，本文使用窗口宽度分别为 2、3、4 的三类卷积，每类卷积的数量都为 64 个。在词向量的输入层，本文将一个句子表示为词向量维度与句子长度的矩阵。为了保证神经网络的结构不变，本文固定句子长度为 20，当句子长度超出 20 时截断，少于 20 时则填充空格符号的词向量。由于需要保证词向量的语义完整性，在卷积层本文使用卷积核纬度与词向量维度需要保持一致，本文词向量维度取值 300。300×n 大小的卷积核与词向量表示的句子运算后，结果为(20-n)。本文将时间、标点与得分这三维特征加入到池化层，保证特征在维度与层次上的一致性。最终池化层输出结果会被连接，经过全连接层与 SoftMax 输出分类结果，输出结果的两维分别对应着分类为抽取与不抽取的概率值。本文的 CNN 结构与特征输入如图 3 所示。

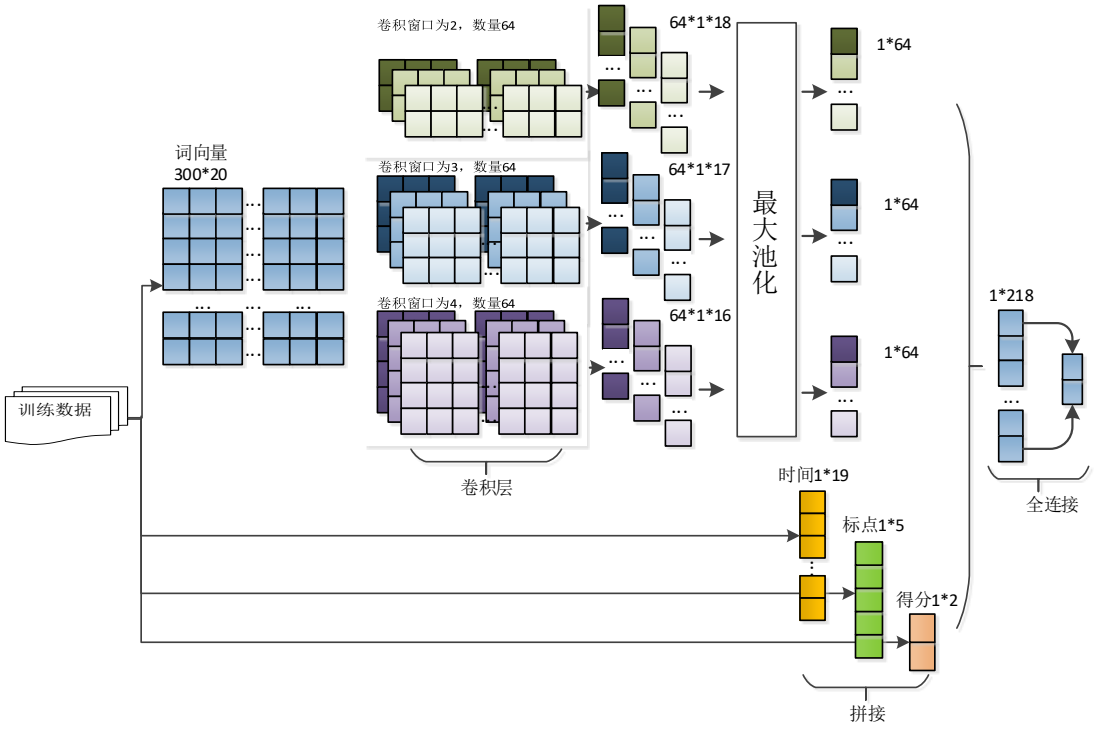

图 3　CNN 结构与特征输入

在模型超参数的选择上，本文尝试使用多种经验范围内的超参数组合，最终使用的学习率为 0.001，本文使用 RELU 激活函数，优化器选择 Adam Optimizer，优化目标方程使用 L2 正则项，正则项系数为 0.005，本文在全连接层设置 dropout 设置为 0.4。本文将 batch-size 设置为 128，训练时对全数据集循环 10 次，训练的终止条件为连续多轮损失函数不再下降活着达到全数据集循环 10 次。

本文的基于卷积神经网络句子抽取算法如下所示。

算法 1 基于卷积神经网络句子抽取算法

输入：足球直播文本

输出：从足球直播文本中抽取的句子序列

1. 将直播文本中的句子进行分词、去停用词等预处理操作。统计直播文本句子的长度分布，确定句子截取长度为 20；

2. 句子中超出长度 20 的部分截断，少于 20 的部分填充空格；

3. 提取人工特征时间、标点与得分，对特征进行离散化以及 one-hot 编码表示；

4. 输入一个句子，将一个句子表示为维度与句子长度的矩阵，将矩阵与对应大小的卷积核进行卷积运算；

5. 将句子卷积结果与人工特征池化；

6. 将池化的结果拼接，进行 SoftMax 层后输出句子二分类的概率，获得抽取结果；

7. 若句子预测为抽取，在输出序列中添加此句子；

8. 全量数据集循环 10 次或者损失不再降低，输出从足球直播文本中抽取的句子序列，反之则重复 2、3、4、5、6、7 步骤。

4　实验结果

4.1 实验设置

足球比赛门户网站通常会在比赛刚开始时滚动播放直播文本并且同步更新比赛技术统计数据，最后会在比赛结束之后给出一篇足球比赛新闻。不同的门户网站会有各自业务需求，因此这些网站之间给出的直播文本与足球比赛新闻的长度与文风差距都比较大，在门户网站爬取高质量数据较为困难。本文的实验语料训练集共有 50 组，实验的测试集共有 30 组。这些数据是由专业的足球新闻报道者撰写，可以认为是高质量的数据。其中每组数据中包含 250 左右个句子，训练集共包含了大约 12500 个句子，测试集共包含了大约 7500 个句子。本文对这些句子的二分类进行了人工标注，标注后正负样本比例大约为 1:5，为了平衡正负样本比例，本文对正样本采用了上采样，采样后的比例大约为 3:5，采样后的训练数据大约

有 16600 条句子。

在对比实验方面，本文使用了基于规则的方法[13]作为对比实验，同时，得到“基于规则”、“基于规则与篇章结构”、“基于 CNN”与“基于 CNN 与篇章结构”四种实验系统结果。

“基于规则”匹配句子中表示“禁区”的关键词。如果匹配成功，则进一步匹配句子中的其它敏感词，如：“手球”、“越位”等。如果句子满足两次匹配，或者当前句子所在的时间内比分发生变化，则抽取该句子。“基于规则与篇章结构”的方法综合了“基于规则”的抽取结果与模板匹配生成的句子；“基于 CNN 分类与篇章结构”综合了“基于 CNN 分类”与模版匹配生成的句子。

本文使用 ROUGE[14]作为自动评估方法，使用 ROUGE-N、F1-Measure 作为评价指标。其基本思想是将模型生成的摘要与参考摘要的 n 元组贡献统计量作为评判依据，主要考察文本生成结果的充分性与忠实性。

4.2 实验结果分析

在自动评估方面，使用 ROUGE 工具包，用 ROUGE-N 的 F1-Measure 作为评价指标，表 9 是本文方法在 30 组测试集上的评测结果。

表 9　评测结果

	Recall	0.57137		Recall	0.24771
ROUGE-1	Precision	0.57479	**ROUGE-2**	Precision	0.25121
	F-Measure	0.57307		F-Measure	0.24945
	Recall	0.10741		Recall	0.25461
ROUGE-3	Precision	0.10733	**ROUGE-SU4**	Precision	0.25373
	F-Measure	0.10737		F-Measure	0.25417

由表中的 ROUGE 结果可知，本文方法得到的结果精准率略大于召回率，说明本文中的方法在精度上略优于覆盖度。这是因为本文标注的训练数据正负样本比例不均衡，通常来说一场足球比赛中，直播脚本中的句子会有上百条，而足球新闻中需要的句子只有几十条，因此在标注直播脚本中句子时，负样本（不抽取）的比例要大于正样本（抽取），这会更倾向与预测出负样本类别，从而导致覆盖率降低，精准率升高。

本文对比基于卷积神经网络与规则的抽取模型，结果在同样 ROUGE 环境下的评测结果如图 4 所示。

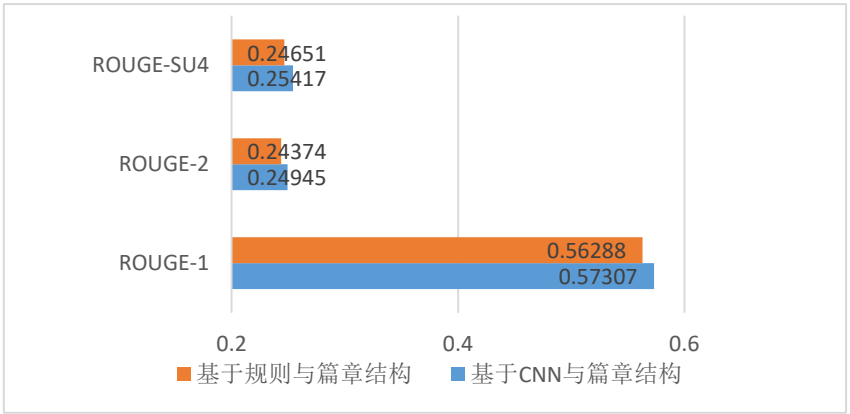

图 4　评测结果对比

由图 4 可知，本文方法在各项结果上均优于基于规则的生成方法。这是因为本文合理的使用了文本特征，将时间、标点与比分三类特征加入到模型的训练过程，并且 CNN 具有较好拟合能力；而本文中使用的规则方法是从有限的数据中加上先验知识总结出来的，仅仅匹配了进球、关键词与敏感词，难以拟合复杂的文本句式。因此，本文中使用 CNN 分类代替规则的抽取方法是有效的。本文中“基于规则”、“基于规则与篇章结构”、“基于 CNN”、“基于 CNN 与篇章结构”的对比结果如表 10 所示。

表 10　有无篇章结构的评测结果对比

	ROUGE-1	**ROUGE-2**	**ROUGE-SU4**
基于规则	0.55475	0.23728	0.23843
基于规则与篇章结构	**0.56288**	**0.24374**	**0.24651**
基于 CNN	0.56311	0.24142	0.24897
基于 CNN 与篇章结构	**0.57307**	**0.24945**	**0.25417**

由表 10 可知，采用篇章结构的实验结果略高于不使用篇章结构的结果。本文在生成方法中使用篇章结构可以让足球新闻层次鲜明，可阅读性强，其对评测结果的提升主要体现在篇章结构上的重复部分，如“时间”和“出场阵容”等部分。本文中使用卷积神经网络与篇章结构方法的生成结果如例 4 所示。

例 4：

北京时间 2 月 6 日英超联赛第 25 轮中最重要的一场比赛曼城主场对阵莱斯特。本场比赛中，曼城进攻意识强烈，同时在本场比赛中展现出了惊人的控球能力。胡特、马赫雷斯为莱斯特奉献了 3 粒进球……

第 2 分钟，鹅卵石左路的传中，乔哈特飞身双龙出海将球击出禁区。马赫雷斯右路得球，假动作后搓球再加速突向底线被德尔夫绊倒。

第 3 分钟，莱斯特获得一个右肋部的任意球机会。瓦尔迪一晃，马赫雷斯低平球送球门前，抢点的胡特近距离将球打进球门！。

第 10 分钟，福布斯传中，费尔南迪奥倒地将球破坏出禁区。瓦尔迪禁区里一打三，德尔夫将球解围。胡特后场得球，被斯特林逼抢下带球出了边线。

……

双方出场名单：曼城（4231）：1-哈特；5-萨巴莱塔……

从例 4 中可以看出，本文中的方法成功的抽取出射门、进球以及禁区内的攻防等信息。在第二分钟时，抽取出禁区内的防守和进攻队员被绊倒的信息；在第三分钟，抽取出胡特进球得分的信息；第九分钟，抽取出乔哈特禁区内封球的信息；第十分钟，抽取出瓦尔迪禁区里一打三，德尔夫将球解围的信息。

生成结果包含了比赛的基本信息、比赛中的精彩片段和双方的阵容。基于篇章结构的方法可以使得足球新闻结构更加明显，重点更为突出，增强了可读性。

5 结论与展望

本文提出了一种基于卷积神经网络与篇章结构的足球新闻自动生成方法，该方法基于卷积神经网络抽取句子，基于模板生成句子，将获得的句子按照篇章结构要求来排列从而得到最终结果。实验结果表明，本文的足球新闻生成方法具有良好的效果，可以从直播脚本中较

精准地抽取并生成符合大众常识的关键句子。

本文中的方法依然有提升的空间，一方面，可以通过拓宽训练集数据，加入规则为不同类型的体育比赛制定不同的足球新闻生成方法；另一方面，随着网络直播脚本规范化的发展，可以给每一条直播语句增加标签，从而使得语句所描述事件的特性更加明确。另外，还可以拓充新闻信息的图片和视频，使生成的足球新闻图文并茂且信息饱满，易于读者阅读。

参 考 文 献

[1] Wang J, Yang J. Statistical Single-document Summarization for Chinese News Articles[C]// IEEE Computer Society, 2012, 183-188.

[2] 林莉媛, 王中卿, 李寿山等. 基于评论质量的多文档文本情感摘要[J]. 中文信息学报, 2015, 29(4): 33-39.

[3] 李培, 翁伟, 林琛. 中文微博故事线生成方法[J]. 中文信息学报, 2016, 30(3): 143-151.

[4] 段旭磊, 张仰森, 孙祎卓. 微博文本的句向量表示及相似度计算方法研究[J]. 计算机工程, 2017, 43(5): 143-148.

[5] 吕超镇, 姬东鸿, 吴飞飞. 基于 LDA 特征扩展的短文本分类[J]. 计算机工程与应用, 2015, 51(4): 123-127.

[6] 陈宇, 王明月, 许莉薇. 基于 DE-ELM 的林业信息文本分类算法[J]. 计算机工程与设计, 2015, 36(9): 2412-2431.

[7] Wan X, Luo F, Sun X, et al. Cross-language document summarization via extraction and ranking of multiple summaries[J]// Knowledge and Information Systems, 2018: 1-19.

[8] Cao Z, Wei F, Li W, et al. Faithful to the original: Fact aware neural abstractive summarization[C]// Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2018.

[9] Cao Z, Li W, Li S, et al. Improving Multi-Document Summarization via Text Classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2017: 3053-3059.

[10] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]// Computer Science, 2014.

[11] Sabour S, Frosst N, Hinton G. Dynamic Routing Between Capsules[C]// Advances in Neural Information Processing Systems. 2017: 3859-3869.

[12] Harris D, Harris S. Digital Design and Computer Architecture, Second Edition[M]// Digital design and computer architecture, Chian Machine Press. 2014: 289-361.

[13] Liu M, Qi Q, Hu H, et al. Sports News Generation from Live Webcast Scripts Based on Rules and Templates[C]// Proceedings of the 5th CCF Conference on Natural Language Processing and Chinese Computing, 2016: 876-884.