基于神经网络的体育新闻自动生成研究*

李浥尘，胡珀，王丽君

（华中师范大学，湖北省 武汉市 430079）

摘要：面向体育比赛的大规模直播脚本快速及时地反映了比赛的实时进程，但依靠体育新闻记者来据此人工撰写新闻报道往往耗时费力。鉴于此，本文提出了一种自动生成体育直播脚本所对应的体育新闻的神经网络模型，该模型在一定程度上避免了传统模型过于依赖人工选择特征的局限性，同时还能综合考虑脚本中句子级局部信息与全局信息以及句子和新闻内容间的语义关联性，从而实现联合建模下的体育新闻生成。在公开数据集上的实验结果验证了本文所提方法的可行性和有效性。此外，我们还尝试了基于规则和模板来自动生成体育新闻的标题以突显新闻正文的关键内容。

关键词：神经网络模型；直播脚本；体育新闻；新闻标题

Sports News Generation Based on Neural Networks

Abstract: The large-scale live scripts for sporting events quickly and timely reflect the real-time process of the game, but it is often time-consuming and laborious for a journalist to write sports news. In view of this, we propose a neural network model which automatically generates the sports news corresponding to the sports live scripts. The model avoids the limitation that the traditional model is too dependent on the hand-crafted features. Besides, it can also consider the script sentence-level local information and global information and semantic relevance between script sentences and corresponding news content in the scripts, which realize the formation of sports news under joint modeling. The experimental results on the open data set verify the feasibility and effectiveness of the proposed method. In addition, we also try to generate the title of sports news based on rules and templates to extract the key content of it.

Key Words: neural network model; live webcast script; sports news; news headline

1 引言

体育比赛直播脚本以体育比赛中的实况数据为信息源,以网络平台为媒介,以文字形式向广大体育爱好者及时转播比赛实况。由于文字直播方式能为暂时无法通过传统媒介观看体育比赛的网民提供另一种动态观赏比赛进程的方式,已逐渐成为基于视频的体育比赛直播的有益补充,获得大量网民的关注和参与。直播脚本通过文字方式来描绘一场比赛的进展情况，随着比赛的不断进行，直播脚本根据交锋双方的赛况实时地更新报道，增强用户体验，同时还为体育新闻记者提供第一手的报道信息,帮助他们在比赛结束后能据此撰写出高质量的体育新闻。

尽管当前众多与体育比赛有关的主流网站已纷纷推出实时的文字直播服务并持续更新赛况。然而截止目前为止，绝大多数基于直播脚本的体育新闻均由专业新闻记者手工撰写，耗时费力且效率低,如何根据直播脚本来自动生成体育新闻逐渐成为近年来 NLP 领域的研究热点之一，具有相当大的挑战性，其主要表现为以下两方面：

第一，直播脚本和体育新闻往往从不同的视角来描述同一场体育比赛。直播脚本实时更新，侧重于描述比赛进程中的各个细节。而体育新闻则更注重于提取出比赛中的关键部分，

* **基金项目：**国家自然科学基金青年基金项目（61402191）；华中师范大学中央高校基本科研业务费教育科学专项资助项目（CCNU16JYKX15）；国家语委科研项目（WT135-11）

作者简介：李浥尘（1993-），男，硕士研究生，主要研究领域为自然语言处理；胡珀（1980-），男，博士，副教授，通讯作者，主要研究方向为自然语言处理、自动文摘；王丽君（1992-），女，硕士研究生，主要研究领域为自然语言处理

辅以更简洁明快的方式报道，因此如何从直播脚本中抽取出“好”的句子作为新闻候选句将是需要解决的关键问题之一。第二，解决这个问题目前的方法大多采取基于人工特征选择的无监督或有监督机器学习方法，而这将在一定程度上限制对不同类型体育比赛或不同领域体育赛事新闻生成的泛化能力和灵活性。鉴于此，如何利用体育直播文本自身的特点及它与对应的体育新闻间的语义关联性来达到尽可能少的人工特征依赖及良好的领域泛化能力是当前迫切需要解决的难点问题，也是本文的研究动机所在。

本文提出了一种新的自动生成体育直播脚本所对应的体育新闻的神经网络模型，该模型在一定程度上避免了传统模型过于依赖人工选择特征的局限性，同时还能综合考虑脚本中句子级局部信息与全局信息以及句子和新闻内容间的语义关联性，实现联合建模下的更高质量的体育新闻生成。在本任务公开数据集上的初步实验结果验证了本文所提方法的可行性和有效性。

2 相关工作

本研究涉及的任务与自动文摘密切相关，自动文摘是自然语言处理中一个传统的研究领域，其应用对象主要集中在新闻和社交媒体。目前文摘的主流方法大致可分为两类：抽取式和生成式。现阶段，抽取式方法相对成熟和高效，因此在本研究中我们暂将该任务作为一个抽取式摘要问题。

绝大多数抽取式摘要方法基于无监督或有监督学习。在无监督学习方法中，基于特征的排序方法通常基于句子的语义或统计学特征，如词频、句子位置、线索词、标点词、词汇链、修辞结构、主题信息等[1][2]。基于聚类的方法通常从每个子主题中选择一个或多个具有最小冗余度和最大覆盖度的代表句构成摘要[3]。近年来，基于图模型的方法取得了较好的效果，LexRank[4]和 TextRank[5]则是采用诸如 PageRank 和 HITS 的代表性方法。

在有监督的摘要方法中，摘要往往被当作句子级的聚类、回归或序列标注任务求解，众多有监督的学习算法如隐马尔科夫模型[6]、支持向量回归[7]、因子图模型[8]等获得了广泛应用。然而，由于有监督的学习方法大多需要大量的标注数据，而这在很多情况下尤其是特定领域很难直接获取并利用。

本研究面向直播脚本的体育新闻自动生成可被视为一种特殊的自动文摘任务，目前这个领域的研究才刚刚开展，近期最相关的工作之一是利用传统句子特征以及任务特定特征来构建一个有监督的学习框架对体育脚本中的所有句子打分，然后结合 DPP(行列式点过程)算法去冗余和排序生成最终的体育新闻[9]。

其他相关工作大多集中在如何使用社交媒体如 Twitter 的状态更新数据来辅助生成体育赛事的新闻[10]或使用基于实体的信息来生成体育比赛摘要[11][12]。还有少量研究利用体育视频的集锦来生成体育比赛的梗概[13]。

纵观现有的研究，大多摘要方法适用于通用的新闻领域，尚未被有效应用于特定领域的摘要任务如体育新闻的自动生成。此外，绝大多数现有方法依赖于人工提取的小规模特征集，但由于依靠手工来选取特征往往耗时费力、泛化性弱、调节麻烦，因此需要提出新的方法来自动学习特征，提高体育新闻的生成质量。

深度学习近年来在诸多 NLP 任务中取得了显著进展，主要原因在于它能通过优化层叠模型自动学习更好的数据表征。一个基于查询的抽取式文本摘要系统将相关性和显著性两个方面合并考虑，利用深度学习可以自动学习句子和文档聚类的词嵌入，并且当查询给定之后，可以应用注意力机制来模拟人类阅读行为[14]。还有使用条件卷积神经网络来生成摘要，条件是卷积注意力模型，用来确保每一步生成词的时候都可以聚焦到合适的输入上。模型仅仅依赖于学习到的特征，并且很容易在大规模数据上进行端到端的训练[15]。

由此可见，利用深度学习网络来加强对直播脚本的分析，是相关研究领域的重要趋势之

一，这也是本文提出基于神经网络模型方案的研究动机。

3　基于神经网络的体育新闻自动生成方法

3.1　方法概述

为了避免人工提取特征，我们提出了一种通用的神经网络模型自动地从直播脚本中生成体育新闻。模型中综合考虑了脚本中句子级局部信息与全局信息以及句子和新闻内容间的语义关联性，从而实现联合建模下的体育新闻摘要生成。此外，我们还尝试了基于规则和模板来自动生成体育新闻的标题以突显新闻正文的关键内容。

图 1 描述了我们提出方法的基本流程。

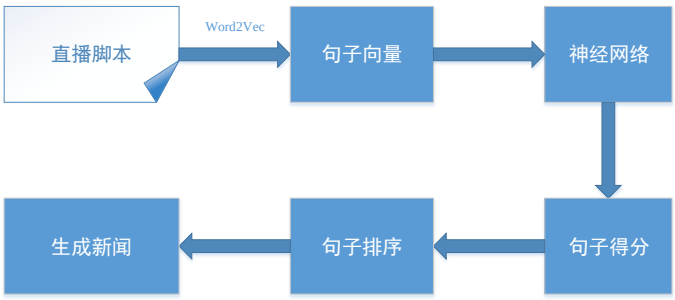


图 1：方法流程图

3.2　体育新闻正文内容的生成模型设计

图 2 显示了提出的神经网络模型。

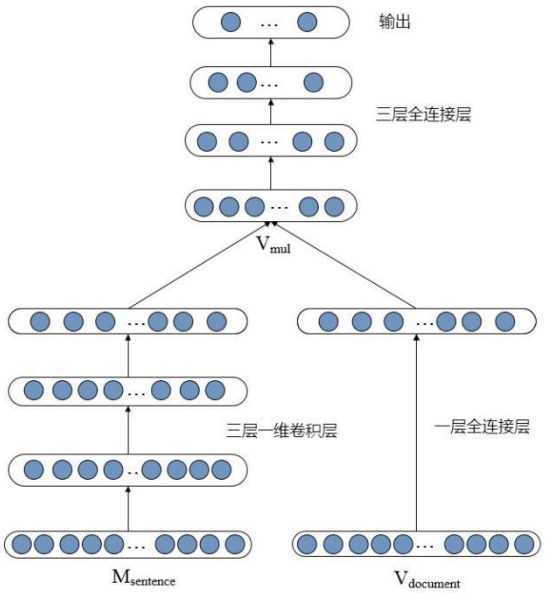


图 2：神经网络模型

在本研究中，每一个句子均被看作词的序列，word2Vec 模型用于实现词向量表示。然后用句子中所有的词向量构成句子的向量表示（即句子矩阵）。此外，为了评估句子对所训练的神经网络模型的重要性，我们先将直播脚本中的每个句子与给定训练集中对应的体育新闻中的所有句子进行相似度比较，然后选择最大值作为该句子的重要性得分。我们的假设是

如果利用提出的模型从直播脚本中抽取出来的句子都与体育新闻文本的相似度高，那么基于这些句子所生成的体育新闻将更接近于标准体育新闻，从而表明我们的模型效果更优。

定义 $M_{sentence}$ 和 $V_{document}$ 作为神经网络模型的输入。直播脚本中的每一个句子都有一个对应的矩阵表示 $M_{sentence}$ 。在实验中将每个词所表示的向量维度设置为 50。

$$M_{sentence}=(V_{w1},V_{w2},...,V_{wk}),W_k \in sentence,k \le 20$$

这里，W_k 表示句子的第 k 个词，V_{wk} 表示第 k 个词所表示的向量，句子所表示的矩阵由 20 个词的向量组成。

将 $M_{sentence}$ 转换为一个 1000 维的向量，$V_{document}$ 表示直播脚本中所有句子向量的总和。

$$V_{document}=\sum_{i=1}^{N} M_{sentence}^{(i)}, N=1,2,...,n$$

神经网络模型的输出定义为句子的重要性得分。

$$score=Max\{similarity(V_s,V_r)\}$$

V_s 表示直播脚本中的句子，V_r 表示新闻中的句子。在本方法中基于 gensim 模块计算句子的相似度，比较直播脚本中的每一个句子与体育新闻中的每一个句子的相似度，取直播脚本中句子所对应相似度的最大值作为模型的输出值。我们基于 tfidf 模型创建相似度矩阵，将句子表示成词的 tfidf 值拼接成的向量，计算向量的余弦相似度。由于每一篇直播脚本文档对应两篇体育新闻文档（163 新闻和 sina 新闻），我们分别计算直播脚本句子所对应的最大相似度得分，取两者平均数作为最终的输出值。

在实验中，我们也尝试基于 LSI 和 LDA 模型计算句子间的相似度，但是最后的实验结果表明基于 tfidf 模型得到的实验效果最好。此外，我们也试图将每一个句子看作一篇文档，计算该文档在所对应的体育新闻的 ROUGE-1 的 F 得分，但是计算结果显示绝大多数句子得分为 0，使得神经网络模型中的输出值过于稀疏，不利于模型的训练。

本研究神经网络模型的构建按照如下方式。首先，$V_{document}$ 添加一层全连接层得到一个 200 维度的向量。

$$X^{(l)}=f(W^{(l)} \bullet X^{(l-1)}+b^{(l)})$$

$W^{(l)}$ 表示第 l 层模型的参数，$X^{(l-1)}$ 表示第 l-1 层的输出，$b^{(l)}$ 为第 l 层的偏置矩阵，$X^{(l)}$ 表示经过全连接层得到的输出值。

然后 $M_{sentence}$ 添加三层一维卷积层得到一个矩阵并用扁平化函数将其转换为一个向量 $V_{sentence}$。

$$X^{(l)}=f\left(W^{(l)}\left[X_{i-s}^{(l-1)},...,X_{i}^{(l-1)},...,X_{i+s}^{(l-1)}\right]+b^{(l)}\right)$$

$W^{(l)}$ 为卷积层中的参数，$X_i^{(l-1)}$ 表示第 l-1 层的第 i 个输出，(2s+1)为卷积窗口的大小 $X_{i-s}^{(l-1)},\dots,X_i^{(l-1)},\dots,X_{i+s}^{(l-1)}$ 表示把卷积窗口内的向量进行拼接，$b^{(l)}$ 为第 l 层的偏置矩阵。卷积层中的局部连接和权值共享降低了参数量，使训练复杂度大大下降，并减轻了过拟合，同时权值共享还赋予了卷积网络对平移的容忍性。使用卷积层在提取特征的同时可以考虑到输入的上下文信息。

接着将两个向量每一维度相乘得到 V_{mul}，这样可以实现综合考虑到脚本中句子级的局部信息与全局信息。

$$V_{mul}=V_{sentence} \bullet V_{document}$$

最后，V_{mul} 添加三层全连接层得到最终的输出结果，基于句子与新闻内容间的语义关联性实现了联合建模下的体育新闻摘要生成。

$$output = sigmoid\left(W^{(l)} \bullet X^{(l-1)} + b^{(l)}\right)$$

这里，sigmoid 是常用的非线性激活函数，把输入的连续实值“压缩”到 0 和 1 之间，output 即为神经网络模型的输出值。output 在模型当中的意义即为直播脚本中的每一个句子与体育新闻中的每一个句子的最大相似度得分。

我们使用交叉熵函数作为损失函数，如下所示：

$$C = -\frac{1}{n}\sum_{x}\left[y\ln\left(output\right) + \left(1-y\right)\ln\left(1-output\right)\right]$$

y 表示标签值的大小，即上文中的 score 所表示的值，output 表示输出的结果值。

3.3 体育新闻标题生成

在本研究中，除了采用前面提出的方法自动生成体育新闻外，还尝试了基于规则和模板来自动生成体育新闻的标题以突显新闻正文的关键内容，基于模板与规则生成新闻标题，标题的构成由队伍名称、最终比分、重要球员表现三部分来构成。

从直播脚本中直接抽取出对阵双方的球队，球队名称集中出现在直播脚本的比赛介绍部分（未赛）和总结部分（完赛）。同时在比赛结束总结部分（完赛）也可提取出比赛的最终比分，由此可得队伍名称以及最终比分两部分内容。

例 1：

本次直播给大家带来的是 2015-2016 赛季欧冠小组赛第一轮，皇马主场和顿涅茨克矿工的比赛　未赛　　0-0

全场比赛结束，皇马主场 4-0 大胜顿涅茨克矿工，取得本赛季欧冠的开门红！！！ 完赛　4-0

在直播文本中的比分栏实时记录对阵双方的比分，若某一行内容出现变化，即表明在该时刻有球员进球。新闻标题中出现的重要球员往往是本场比赛中发挥出色的球员，因此在直播脚本中出现比分变化的句子当中提取出重要球员的名字，统计重要球员的进球数，并依据进球数量总结重要球员的比赛表现，由此可得重要球员表现部分的内容。

例 2：

本泽马，推射空门得手！！！　上半场 30　1-0

C 罗主罚，一蹴而就！！！　下半场 11　2-0

C 罗主罚，助跑，右脚劲射打球门左侧死角，皮亚托夫判断错了方向，3-0！！！　下半场 20　3-0

进球啦，4-0！！！C 罗的补射，上演帽子戏法，其中两个点球　　下半场 36　4-0

结合以上抽取出来的队伍名称、最终比分、重要球员表现三部分内容生成新闻标题，示例如下：

例 3：

皇马 4:0 顿涅茨克矿工，C 罗上演帽子戏法

4　实验与评估

4.1 实验设置

1）数据集

为了评估本文提出的方法在体育新闻自动生成任务上的可行性与有效性，我们采用由张建敏等于 2016 年首次构建并发布的本任务开放数据集（即 acl16_sports 数据集）[9]。该数据集共有 450 篇文档，其中 150 篇文档是直播脚本，另 300 篇文档是每篇直播脚本对应的网易和新浪体育新闻编辑所撰写的体育新闻。

2）评价指标

在本实验中，我们将数据集随机分为两个不同的部分，其中一部分包含 100 篇直播脚本与其对应的 200 篇体育新闻，它们被用作训练集，另一部分则被设为测试集。为了便于评估，使用 ROUGE-1.5.5 工具包[16]来比较基于脚本生成的体育新闻与新闻媒体记者撰写的体育新闻的内容重叠度。作为评价指标，我们报道了 ROUGE-1（R-1）和 ROUGE-2（R-2）的 F 指标分数。

3）比较方法

我们使用以下的主流摘要系统为基准，与我们提出的方法进行比较。这里，前三个系统是典型的无监督摘要方法，它们直接用于每个体育比赛的直播脚本，通过提取最重要的句子来生成新闻。RF + DPP 系统是专门针对该任务设计的有监督摘要系统，该系统利用直播脚本和对应的人工编辑新闻来自动学习生成体育新闻。

Centroid：是基于质心的摘要系统[17]，它计算文档中一个称为质心句的伪句子。质心句由 TFIDF 分数高于预定义阈值的词组成。通过基于不同特征的得分总和：包括句子与质心句的余弦相似度、位置权重以及句子与首句的余弦相似度来定义每个句子的得分。

LexRank：LexRank[4]基于句子图表示中特征向量的中心性概念来计算句子的重要性。在该模型中，使用基于句内余弦相似度的连接矩阵作为句子图表示的邻接矩阵。

ILP：整数线性规划（ILP）方法[18]将文档摘要看作组合优化的问题。ILP 模型通过最大化摘要中包含的二元组概念的频率权重的总和来选择句子。

RF + DPP：将此任务看作学习排序问题，在一个有监督学习的框架下通过计算文档的传统特征及特定任务的特征求解[9]。

4.2 实验结果

4.2.1 对比方法

表 1 给出了不同方法下的实验结果。

Method	R-1	R-2
Centroid	0.32508	0.08113
LexRank	0.31284	0.06159
ILP	0.32552	0.07285
RF+DPP	0.39391	0.11986
Our method	**0.42310***	**0.12351***

表 1：不同方法的实验结果

（*表示结果具有统计显著性意义）

从表 1 中可以看到，提出的基于神经网络模型的方法与传统的无监督和有监督方法相比，获得了更好的结果。

从实验结果可以看出，传统的文档摘要的方法应用于该任务效果并不好。Centroid 是一种基于中心点的句子抽取方法，它在赋予句子权重的过程中，综合考虑了句子级以及句子之间的特征，但是对于本任务而言，部分特征的设置并不合理。ILP 将摘要看做一个带约束的优化问题，同时进行句子抽取与冗余去除，非常适合解决多文档摘要问题，但是在实验中效果并不好。LexRank 方法通过句子间的相似性为多文档构建句图，使用 tf 与 idf 来衡量句子间的相似性，然而直播脚本的句子多以短句为主，实时描述比赛的进程，反映比赛的发展过程，所以句子之间的相似度并不高。

RF+DPP 模型将文档摘要的传统特征与任务的特有特征相结合，使用概率句子选择算法去除冗余句子。RF+DPP 模型的特征验证表明，两方面的特征均有利于摘要的生成，相比较而言，为该任务设定的特征如重要比赛事件、得分变化、重要球员等等在句子抽取的过程中影响力更大。RF+DPP 方法相较于传统的文档摘要方法取得了更好的效果，但是这个方法还是依赖于人工提取的一系列句子特征。

我们的方法能够达到最优效果，其主要来自于两个方面的原因：第一，我们的方法是有监督的学习方法，基于神经网络模型，以直播脚本中的句子与新闻中句子的相似度作为训练目标；第二，提出的方法综合考虑了脚本中句子级的局部信息与全局信息以及直播脚本中的句子与新闻内容间的语义关联性。此外，我们的模型没有使用任何人工提取的句子特征来生成相应的体育比赛新闻，不仅适用于生成足球比赛直播脚本对应的体育新闻，也适用于其他体育比赛的领域，具有更好的领域泛化能力。

4.2.2 错误分析

尽管实验结果表明我们的方法是可行且有效的，但结果集中仍然存在一些错误。

错误一：在直播脚本中充斥着大量短句子甚至噪音句子，有时几个连续的短句子描述了一个重要的事件，而当前的模型当中，往往不能将这部分句子抽取出来从而导致信息的缺失。

例 1：

皇马进球啦！！！　　上半场 30　　1-0
本泽马，推射空门得手！！！　　上半场　1-0

这些短句子描述的是比赛进程中的重要事件，然而它们没有被抽取出来是因为模型在计算这部分句子时，短句子在直播脚本文档中影响力较小且与体育新闻中的语句关联性较低。

错误二：在直播脚本中，比赛开始之前，会有大段的篇幅介绍比赛的背景知识，主要内容包括两支球队各自的风格特点、人员变换、球队对阵记录、近期状态等。研究发现在我们的模型中会抽取大量直播脚本中的未赛部分的句子，由于新闻的长度限制，导致生成的体育新闻中描述比赛重要事件的部分被压缩，没能提取出直播脚本中部分关键事件的信息。

例 2：

大家好 ， 欢迎收看新浪体育为您带来的英超第 7 轮 莱斯特 VS 阿森纳 未赛　　0-0
温格麾下的球队，技术华美，但精神力软弱，这是足坛公论 未赛　0-0
本赛季阿森纳是顺风球之王，对阵切尔西、水晶宫、纽卡斯尔、斯托克城、热刺 5 战，阿森纳先进球，5 战全部取胜 未赛 0-0
而对阵萨格勒布、切尔西、西汉姆三战，阿森纳先丢球，三战全部落败 未赛 0-0
本战阿森纳的对手莱切斯特则是本赛季英超“逆风球之王” 未赛　0-0
双方历史交锋 133 场，阿森纳 61 胜 44 平 29 负 未赛　0-0
阿森纳近 19 次对阵莱斯特城取得 11 胜 8 平保持不败 未赛 0-0

这些句子都是作为比赛的背景知识出现在直播脚本当中，然而我们的模型却将这些句子都抽取出来作为生成的体育新闻的内容。这部分句子与新闻的语义关联度较高并具有总结性意义，同时多以长句子为主，句子当中的词更容易同时出现在直播脚本的其他位置以及新闻的内容当中。

这两个问题在所提出的方法中尚没有得到很好的解决，我们将在后续的工作中重点解决。

5 总结与展望

本文研究如何从直播脚本中自动生成体育新闻，我们将此任务作为特殊的抽取型摘要问题，提出了一种基于神经网络的方法提高现有方法的泛化能力和灵活性。提出的方法不仅考虑到脚本中句子级的局部信息与全局信息，还考虑了句子与新闻内容间的语义关联性，从而实现联合建模下的体育新闻摘要生成。初步的实验结果验证了我们方法的有效性，在不使用任何人工提取特征的情况下，性能优于众多基准方法甚至是有监督学习的方法，达到了目前的最优实验效果。

在未来的工作中，我们将探索生成式摘要而不仅仅采用纯抽取型摘要的方法，自适应地学习适合不同领域的体育新闻模板，并通过引入注意力机制将脚本和新闻的多粒度上下文层次信息融入当前的神经网络模型中。

参考文献

[1] Luhn, H. P. (1969). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2):159-165.

[2] Lin, C.Y. and Eduard, H. (2000). The automated acquisition of topic signatures for text summarization. In Proceedings of the 17th Conference on Computational Linguistics (COLING 2000), pages 495-501, Association for Computational Linguistics, Stroudsburg, PA.

[3] Nomoto, T. and Matsumoto, Y. (2001). A new approach to unsupervised text summarization. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001), pages 26-34, ACM, New York, NY.

[4] Erkan, G. and Radev, D.R. (2004). LexPageRank: prestige in multi-document text summarization. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004).

[5] Mihalcea, R. and Tarau, P. (2004). TextRank: bringing order into texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)

[6] Conroy, J.M. and Oleary, D.P. (2001). Text summarization via hidden markov models. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001), pages 406-407,ACM, New York, NY.

[7] You, O.Y., Li, W.J., Li, S.J., and Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. Information Processing and Management, 47(2):227-237.

[8] Yang, Z., Cai, K.K., Tang, J., Zhang, L., Su, Z., and Li, J.Z. (2011). Social context summarization. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011), pages 255-264, ACM, New York, NY.

[9] Jianmin Zhang, Jin-ge Yao, and Xiaojun Wan. 2016. Towards Constructing Sports News from Live Text Commentary. In Proceedings of ACL 2016.

[10] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. Summarizing Sporting Events using Twitter. In Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, pages 189-198.

[11] Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. 2011. Content Selection from an Ontology

Based Knowledge Base for the Generation of Football Summaries. In Proceedings of the 13th European Workshop on Natural Language Generation, pages 72–81.

[12] Nadjet Bouayad-Agha, Gerard Casamayor, Simon Mille, and Leo Wanner. 2012. Perspective-oriented Generation of Football Match Summaries: Old Tasks, New Challenges. ACM Transactions on Speech and Language Processing (TSLP), 9(2):3.

[13] D.Tjondronegoro,Yi-Ping Phoebe Chen, and Binh Pham.2004. Highlights for More Complete Sports Video Summarization. page(s): 22-37.

[14] Ziqiang Cao, Wenjie Li, and Sujian Li. 2016. AttSum: Joint Learning of Focusing and Summarization with Neural Attention. In Proceedings of COLING 2016.

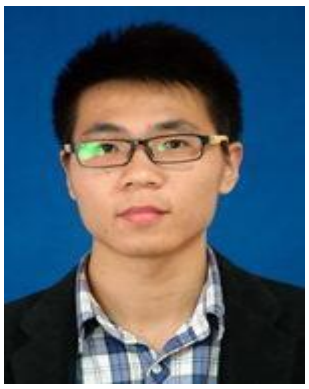
[15] Sumit Chopra,Michael Auli, and Alexander M.Rush. 2016. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In Proceedings of NAACL 2016.

[16] Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram cooccurrence statistics. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 71–78. Association for Computational Linguistics.

[17] Dragomir R Radev, Hongyan Jing, and Malgorzata Budzikowska.2000. Centroid-based summarization of multiple documents: sentence extraction,utility-based evaluation, and user studies. In Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization, pages 21–30. Association for Computational Linguistics.

[18]Dan Gillick, Benoit Favre,and Dilek Hakkani-Tur.2008. The icsi summarization system at tac 2008. In Proceedings of the Text Understanding Conference

作者简介：



李浥尘（1993--），男，硕士研究生，主要研究领域为自然语言处理。Email: 13720164837@163.com.

胡珀（1980--），男，博士，副教授，通讯作者，主要研究方向为自然语言处理、自动文摘。Email: phu@mail.ccnu.edu.cn.

王丽君（1992--），女，硕士研究生，主要研究领域为自然语言处理。Email: lijunWang06@163.com.