安全类文章的多文本分类系统的设计与实现

吴习沫，朱广宇，张　雷

（华北计算机系统工程研究所，北京 100083）

摘　要： 目前安全类网站信息的分类标签各不相同，没有统一分类标准，使安全类网站无法准确地向用户展示特定类别的安全信息。面对大量的安全类网站的技术类文章信息，用户需要花费大量的时间来识别文本类别。因此，设计一个多文本分类系统对于提高安全类网站的用户体验和使用效率具有重要意义。开发了一套基于 CNN 和 LSTM 混合模型的安全类文章多文本分类系统，本系统采用基于 Scrapy 框架的网络爬虫，该网络爬虫支持定制化配置提取不同布局的页面数据，支持数据持久化存储。并在 CNN 和 LSTM 混合模型基础上设计实现了多文本自动标注模块，实现了网站安全类信息的自动分类，相对传统的 CNN 和 LSTM 模型分类准确率分别提升 1.79%和 1.54%，F1 值分别提升 1.02%和 0.32%。

关键词： 深度学习；文本分类；爬虫；系统

中图分类号： TP391.1　　**文献标识码：** A　　**DOI：** 10.19358/j.issn.2096-5133.2020.07.009

引用格式： 吴习沫，朱广宇，张雷. 安全类文章的多文本分类系统的设计与实现[J].信息技术与网络安全，2020，39(7)：52-56，60.

Design and implementation of multi-text classification system for security articles

Wu Ximo，Zhu Guangyu，Zhang Lei

(North China Institute of Computer Systems Engineering，Beijing 100083，China)

Absrtact： At present，the classification labels of security website information are different，and there is no unified classification standard，so that security websites cannot accurately display specific types of security information to users. Faced with a large number of technical article information of security websites，users need to spend a lot of time to identify text categories. So，it's significant to design a multi-text classification system to advance the user experience and make use of security websites' efficiency. This paper develops a security text multi-text classification system based on a hybrid model of CNN and LSTM. Based on the Scrapy framework，a web crawler，which supports both customized configuration to extract page data in different layouts and data persistence storage，is used in this system. Based on the mixed model of CNN and LSTM，a multi-text automatic labeling module is designed and implemented to realize the automatic classification of website security information. The rate of classification accuracy has increased by 1.79% and 1.54% in comparison with the traditional CNN and LSTM models respectively. Meanwhile，the F1 value has increased by 1.02% and 0.32%.

Key words： in-depth learning；text categorization；crawler；system

0 引言

互联网已成为信息传播的普遍途径，然而，由于互联网中的冗余信息过多，各网站提供的标签没有统一的分类标准，使得整合某一特定类的文章信息所消耗的时间成本和人力成本增加。但目前为止，针对网络安全类网站的技术类文章，还没有一套系统能够很好地解决上述对应问题。

为迅速掌握最新的网络安全信息，本文设计并实现了基于 CNN 和 LSTM 混合模型的安全类文章多文本分类系统，该系统从多种来源收集安全类技术文本，并将它们以特定格式汇总，自动标记汇总后的文章内容。就信息收集而言，系统主要采集近

一年的安全类技术文本，收集的目标内容主要包括文章内容和网页自带的标签，对于各网站自定义的文章标签，可作为多标签的一部分，供用户参考。安全类文本与普通文本对比需要由多个标签对其进行标记分类处理。因此安全类文本的分类要难于普通文本分类处理。

面向网络安全数据高并发的安全类网站，本文设计和实现了信息采集模块，该模块主要实现了基于 Scrapy 框架的分布式爬虫程序设计，完成了多个安全类网站技术类文章的文本信息数据采集。

本文设计并实现了信息分类模块，它负责对所获得的数据进行预处理、文本表示以及文本分类，其中文本分类模块具体提出了一种基于 CNN 和 LSTM 的混合分类模型，它综合了 CNN 与 LSTM 的优点，提高了模型的特征提取能力。实验结果表明，基于 CNN 和 LSTM 的混合分类模型达到了比较高的准确率，CNN 和 LSTM 的混合模型的准确率为 91.99%。CNN－LSTM 与 CNN、LSTM 相比分类准确率提高了 1.79%和 1.54%。

1 相关工作

文本分词是中文文本预处理过程中的一个重要环节，分词技术是把由字构成的句子按语义划分为由单词组成的句子。由于网络语言表现形式自由、语言不规范、内容多样化等特点，传统分词算法难以充分提取特征，此外，网络新词的创造相对于传统词汇具有一定的变异性[1]。为了克服上述问题，本文应用 Bi－LSTM－CRF 模型[2]对待分类文本进行中文分词。神经网络[3]通过自动学习从样本数据中提取文本特征，成功地克服了传统人工提取方法的不足，具有广泛的应用范围和广泛的适用性。在自然语言处理方面谷歌提出的 Word2Vec[4]算法在文本表示方面十分突出，该算法使得文本表示更加精准而被广泛使用，本文通过结合使用 Word2Vec 字向量和词向量完成了中文文本的分词和向量化表示。

HOCHREITER S 等人[5]提出短期记忆网络(Long Short Term Memory，LSTM)，它是一种时间递归神经网络，该神经网络近期被 AlexGraves 进行了改进。LSTM 模型一般用于多文本分类实验，该模型的优点是可以利用文本中的上下文信息和特征进行非线性拟合，从而实现时间序列的处理和预测。基于该模型设计的系统，已经在聊天机器人、文档摘要等相关领域被广泛使用[6]。

KIM Y[7]首次将 CNN 模型应用在了序列文本问题上，在效果上和 LSTM 模型互有所长，更多的研究发现，CNN 更擅长提取数据不同尺度的局部相关特征，且在处理效率上相较于循环神经网络有大大提高。两者单独使用已经在文本分类问题上取得一定效果。

Lai Siwei 等[8]提出在文本分类问题中同时使用 CNN 和 LSTM 的混合模型，这两种模型在文本分类中分别发挥了 CNN 模型和 LSTM 模型各自的优势，故分类试验的准确率将显著高于两种模型中任意一种单独使用的准确率。

2 系统设计

本系统包括图 1 所示的两个功能模块：信息采集和信息分类。

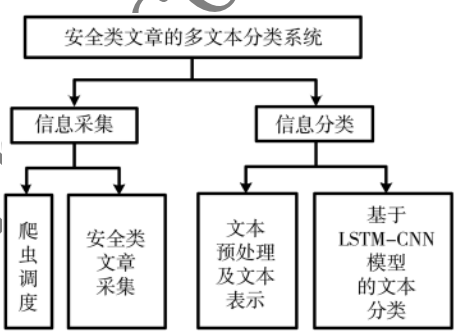

图 1 系统组成图

信息采集模块包含爬虫调度子模块和安全类文章采集子模块，实现了基于 Scrapy 框架对多个安全类网站进行文本信息采集的相关功能。

信息分类模块包括文本预处理及文本表示子模块和基于 LSTM－CNN 模型的文本分类子模块。

2.1 信息采集

信息采集模块具体功能包括：获取当前爬虫任务状态，对站点信息进行更新和配置查询模块状态从而进行信息采集。

为实现该爬虫的上述功能，如图 2 所示，设计并实现了两个交互页面，分别是配置子页面和爬虫管理子页面，其中页面交互功能通过信息处理层和业务逻辑层实现。

配置子页面用于接收具体采集属性信息，为便于后续消息预处理功能模块解析，每类请求都有相应的固定格式。该爬行器通过对子页集群爬虫的启动和停止进行配置，从而实现了对集群爬行的实时动态监控。其中，Handler 层的主要任务是验证和处

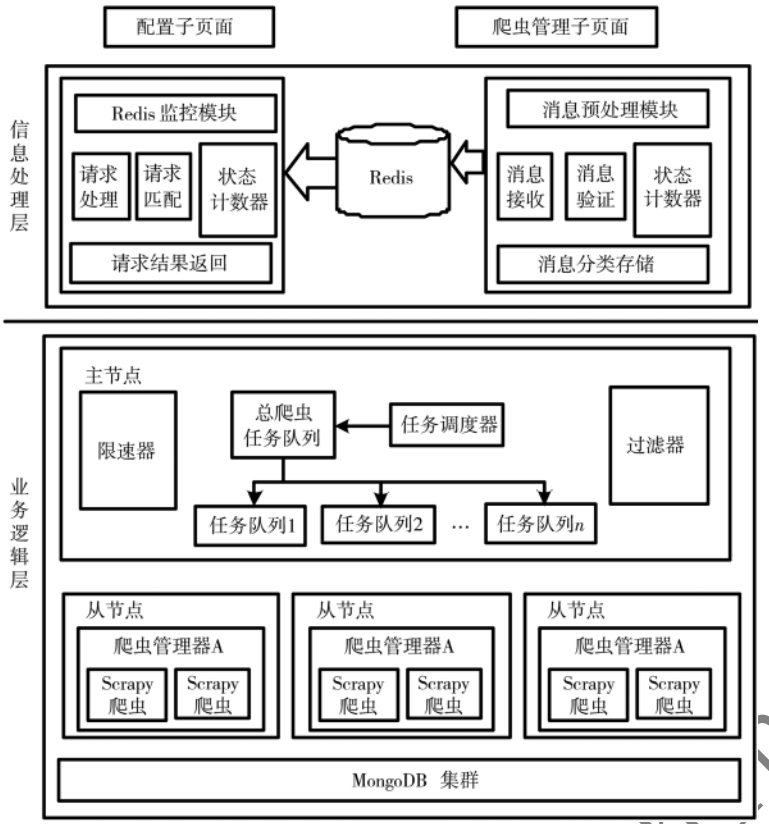

图 2 爬虫架构图

理页面发送的请求消息；MessagePression 模块通过自带的计数器功能对爬虫请求类型进行计数分析处理；通过对文本信息的请求验证，将符合队列信息的内容放入到爬虫请求队列中，Redis 的循环检测是由 Redisvisition 功能模块执行相关的检查处理，处理后启动响应插件，以完成对请求的处理解析，增加 Redis 监控数量可以显著提高请求响应速度和可靠性[9]。

2.2 信息分类

本文提出的信息分类模块主要包括文本预处理及文本表示子模块和基于 LSTM－CNN 模型的文本分类子模块，实现了三个主要功能：文本预处理，文本表示，训练分类模型。

2.2.1 文本预处理及文本表示

文本预处理包括文本清洗、分词和文本表示。文本清洗主要有去停用词，去除非文本符号等。分词阶段，本文使用了基于 LSTM 模型和字向量的的 Bi－LSTM－CRF 中文分词模型[2]进行分词，通过对典型语料数据集的测试结果表明[10]，对于未收录词的分词准确率，该模型明显好于传统的分词方法。

第二步结合 Bi－LSTM－CRF 的输出结果，对中文文本进行分词和和索引化，并使用 Word2Vec 词嵌入矩阵转化为词向量形式。词向量为文本在高维空间的分布式表示，如图 3 所示。文本向量化训练模型 Word2Vec 将文本中的每个单词映射成一个固定维度的向量，这些单词的向量组合到一起形成词向量空间。

2.2.2 基于 CNN 和 LSTM 的文本分类

对于文本分类，若每一类的关键字确定、稳定，使用正则匹配准确率较高。然而，若文本分类界限无明显区分，关键词混淆、不明确，正则关键字匹配分类效果较差。本文的自动化分类是基于深度学习的分类模型，采用基于 CNN 和 LSTM 的混合模型，对安全类文本进行分类。利用 K－MaxPooling 方法提取更多的特征用于分类，在分类器层使用 Softmax 函数来计算每一类的概率。基于 CNN 和 LSTM，本文设计并实现了一个多文本分类模型，其总体结构如图 4 所示，模型分为 9 个层次：输入层、Embedding 层、双向 LSTM 层、拼接层、CNN 层、K－MaxPooling 层、全连接层、Softmax 层、输出层。

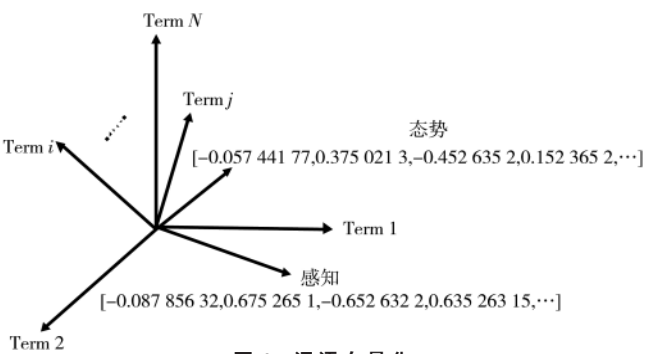

图 3 词语向量化

第一层为输入层，输入经过文本分词和索引化后的文本序列。

第二层为 Embedding 层，作用是将输入层传递来的数据中的每个数字索引转化为对应的词向量[11]。

第三层为 Bi－LSTM 层，如图 5 所示，主要负责提取句子向量的上下文信息。相较于单层的 LSTM，增加了对语句逆向特征的提取能力。为了防止过拟合，结合使用了 L2 正则和 dropout 方法。

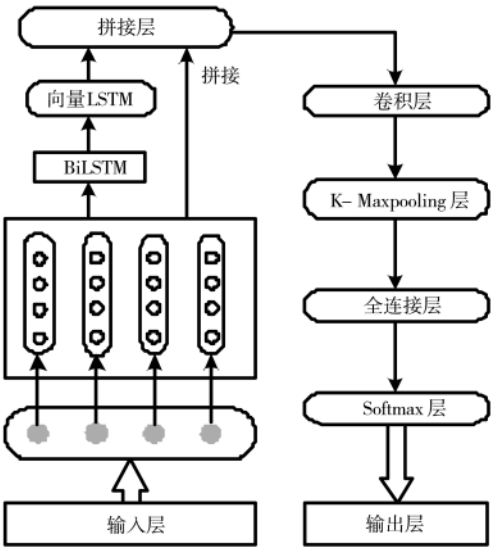

图 4 神经网络结构架构图

第四层为拼接层，函数主要用于拼接双向 LSTM 层和 Embedding 层所输出的特征向量和词向量。在 Embedding 层中将 LSTM 层的特征向量与特征提取和原始向量相结合，可以使处理的文本更丰富，原始信息更完整。在 CNN 层中加入最终得到的信息，可以有效地提高 CNN 层次特征表达的能力。

第五层为卷积层，作用是通过卷积操作提取特征之间的局部特征。

第六层为 K-MaxPooling 层，主要从卷积层提取出多个最大特征值，即提取出最重要的最大特征值数量的信息。

第七层为全连接层，作用是特征降维。

第八层为分类器，通过 Softmax 函数将文本特征进行类别分类。

第九层为输出层，主要负责输出结果数据。

3 实验分析

3.1 数据集

本实验数据集来源如表 1 所示，数据采集层采集多个安全类网站文本信息，采集内容属性具体包括：文章标题、作者、文章具体内容、原网页自带标签、文章发表日期。数据量为 1 万条。标注方式为人工标注。

表 1 数据集来源

网站网址	数据量/条
https://www.freebuf.com/	4 500
https://ti.qianxin.com/blog/	170
http://www.cjzzc.com/	1 680
https://bbs.2cto.com/	1 000
https://www.anquanke.com/	1 600
https://xz.aliyun.com/	1 050

3.2 基准模型分类实验结果

如表 2 所示，文章内容为采集模块输出的安全类文章的内容，人工标注的分类结果作为目标分类结果，为 CNN-LSTM 模型分类结果。

表 3 所示为不同模型的实验结果，表格显示 CNN-LSTM 模型比传统的 CNN、LST 模型更准确。

3.3 实验的参数设置

梯度更新采用动量梯度下降算法，相比原始梯度下降算法，收敛速度更快，效果更好，公式如下：

$$V_t=\beta * V_{t-1}+(1-\beta) * \nabla W_t \quad (1)$$

$$W_t=W_{t-1}-\mathrm{lr} * V_t \quad (2)$$

实验使用 Glove.Twitter.42B.300d 向量作为英文词向量表示，Glove.微博.300d 向量作为中文字向量表示，相关主要超参数初始化如表 4 所示。

4 结论

本文设计并实现了安全类文章多文本分类系

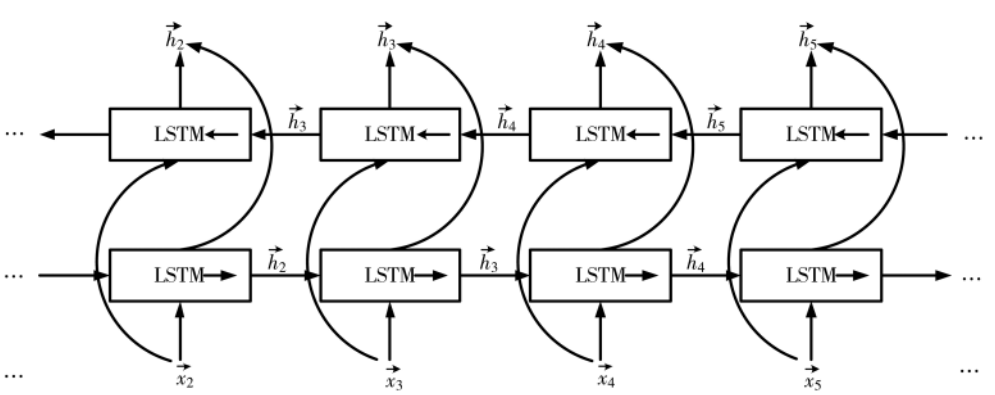

图 5 BiLSTM 网络结构图

表 2 文本内容及分类结果

文章内容	人工标注	CNN-LSTM
“漏洞发现于 DEF CON 会议的前几天，当时我在旧金山闲了几天，并准备和朋友们开车到拉斯维加斯参加 DEF CON 大会，凑巧就是在这放松的几天当中，我发现了该漏洞…”	Web 安全 漏洞	Web 安全 漏洞
“后面又使用 OOB（out of band）带外通信技术，结合 DNSlog 平台进行手注。速度是略微有点提高，但是还是觉得应该有更好的自动化方法…”	工具 漏洞	工具 漏洞
“Linux 库文件劫持这种案例在今年的 9 月份遇到过相应的案例，当时的情况是有台服务器不断向个可疑 IP 发包，尝试建立连接…”	系统安全 安全态势	系统安全
“Sparrow-wifi 本质上一款针对下一代 2.4GHz 和 5GHz 的 WiFi 频谱感知工具，它不仅提供了 GUI 图形化用户界面，而且功能更加全面，可以代替类似 inSSIDer 和 linssid 之类的 Linux 工具。在其最完整的使用场景下…”	工具 无线安全 战略战策	工具 无线安全 战略战策
“这三个恶意应用被伪装成摄影和文件管理工具。根据其中一款应用的证书信息推测这些应用自 2019 年 3 月起就一直处于活跃状态…”	网络安全 态势感知	网络安全 态势感知
…	…	…

表 3 不同模型的实验结果对比

分类方法	准确率/%	召回率/%	F1/%
CNN	90.2	84.01	87.0
LSTM	90.45	85.12	87.7
CNN-LSTM	91.99	84.38	88.0

表 4 实验主要超参数设置

参数名称	参数值
learning rate	0.001
L2	0.000 1
dropout 概率	0.2
LSTM 隐层大小	150
CNN 卷积核宽度	2*600
CNN 卷积核数	4

统，主要包含信息采集模块和信息分类模块，首先通过信息采集模块对安全类网站进行大量的信息采集，将采集到的文本数据存入数据库，然后通过信息分类模块对存入数据库中的文本进行了预处理及文本表示，具体包括数据清洗、文本分词、文本表示等，将文本表示后得到的词向量分别输入 CNN 模型、LSTM 模型和 CNN-LSTM 模型这 3 个文本分类模型进行多文本分类实验。实验结果表明，CNN-LSTM 混合模型的准确率和 F1 值分别达到 91.99% 和 88.02%，均优于 CNN 和 LSTM 模型。

参考文献

[1] REN P, CHEN Z, REN Z, et al. Leveraging contextual sentence relations for extractive summarization using a neural attention model[C]. Proc. of the 40th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2017.

[2] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer Science, 2015.

[3] WANG P, QIAN Y, SOONG F K, et al. Part-of-speech tagging with bidirectional long short-term memory recurrent neural network[J]. Computer Science, 2015.

[4] MIKOLOV T, CORRADO G, CHEN K, et al. Efficient estimation of word representations in vector space[C]. Proceedings of the International Conference on Learning Representations (ICLR 2013), 2013.

[5] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[6] CHANG P, GALLEY M, MANNING C. Optimizing Chinese word segmentation for machine translation performance[C]. Proceedings of the Third Workshop on Statistical Machine Translation, 2008.

[7] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv: 1408.5882V2, 2014.

[8] 黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 21(3): 8-19.

[9] 黄昌宁. 中文信息处理的分词问题[J]. 语言文字应用, 1997(1): 72-78.

[10] WU A, JIANG Z. Word segmentation in sentence analysis[C]. Proceedings of the 1998 International Conference on Chinese Information Processing, 1998.

[11] ZHAO H, HUANG C N, LI M, et al. An improved Chinese word segmentation system with conditional random field[J]. Proceedings of the Fifth Sighan

（下转第 60 页）

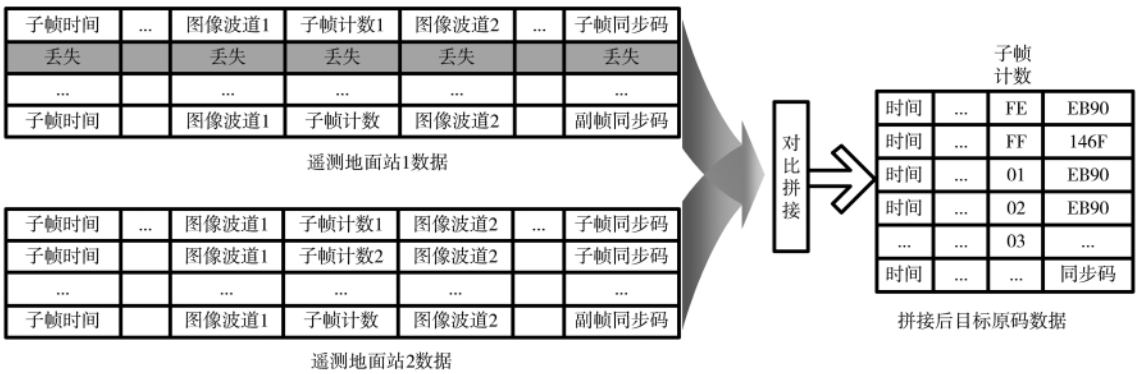

图 4 原始数据拼接示意图

IRIG106 标准，目前在标准第 7 章“下行链路数据包遥测”，定义了一种将其第 10 章数据包、TmNS (Telemetry Network Standard)数据包和以太网数据包融入 PCM 流的方法[7]，形成了多数据流(PCM+网络+视频)混合遥测解决方案。比如法国 Zodiac Aerospace 公司 MDR 数据记录器，符合标准第 10 章所制定的标准数据可以在进行存储的同时进行实时处理，也可以在任务结束后对数据进行事后处理，可以记录多种数据格式，以及进行模块化设计。目前多种模块支持视频/音频、模拟信号、数字量、总线和网络等数据的采集记录，视频及数据存储灵活性大大提高。

4 结束语

实时图像的获取在各类飞行试验中将发挥重要作用，随着关键舱段内外摄像头的增多以及高清图像的应用，视频信息将占用大量的频谱资源，遥测 PCM 数据流数据单向传输和点对点传输的局限性也越发明显。由于频率资源的进一步紧张，带宽利用率更高的 SOQPSK-TG、Multi-h CPM 体制也将陆续被推出，网络化遥测也具备了可实施的应用标准[7-8]，这些新技术的采用将显著提升遥测实时传输能力。

参考文献

[1] 李艳华，李凉海，谌明，等.现代航天遥测技术[M].北京：中国宇航出版社，2018.
[2] GJB5825-2006 靶场试验遥测图像传输要求[S].2006.
[3] 邢达波，李铁林，艾波.机载 LVDS 视频传输技术[J].中国科技信息，2015(10)：23-24.
[4] 巫福胜，姚冉中.遥测数据中红外图像信号提取技术研究[J].信息通信，2016(4)：28-29.
[5] 卢长海，张梦堃.某型遥测地面站实时处理系统设计与实现[J].信息技术与网络安全，2019，38(10)：45-47.
[6] 张军，张潇潇，李圳峰.多站遥测数据精确拼接方法设计与实现[J].遥测遥控，2017，38(4)：23-24.
[7] 李凉海.从 IRIG106 遥测标准看遥测新发展[J].遥测遥控，2017，38(1)：1-4.
[8] 李宏伟，张华栋，山鹏.基于美军新版 IRIG106标准的靶场遥测技术分析[J].飞航导弹，2013(8)：54-57.

(收稿日期：2020-04-25)

作者简介：

卢长海(1977-)，男，硕士，工程师，主要研究方向：遥测遥控。

(上接第 56 页)

Workshop on Chinese Language Processing，2006：162-165.

(收稿日期：2020-04-06)

作者简介：

吴习沫(1994-)，女，硕士研究生，主要研究方向：自然语言处理、机器学习。

朱广宇(1981-)，男，硕士，高级工程师，主要研究方向：信息安全。

张雷(1985-)，男，硕士，工程师，主要研究方向：社交网络、自然语言处理、工控安全。

版权声明