大数据认知计算在内容安全管控中的应用

杜雪涛
中国移动通信集团设计院有限公司，北京 100080

摘要
通信网络中存在海量垃圾和不良信息，这些信息需要被阅读和理解，以便对其进行有效的特征提取和拦截封堵。基于人工分析的方法已经无法达到目的，需要使用基于大数据的认知计算技术代替人工进行海量的数据分析和理解，帮助人们制订内容安全管控策略。针对电信诈骗治理、不良消息治理、变体消息治理和不良网站治理4个方面遇到的实际问题，分别提出了大数据认知计算的解决方案，并给出了创新性实践的效果。实践表明，提出的解决方案能够快速发现不良信息，有效地提升内容管控质量。

关键词
大数据；认知计算；内容安全；诈骗识别

中图分类号：TP181　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2021060

Applications of big data cognitive computing in content security governance

DU Xuetao
China Mobile Group Design Institute Co., Ltd., Beijing 100080, China

Abstract
In the communication network, there is a mass of bad information needed to be read and understood to extract useful knowledge and features for governance. Methods based on manual analysis can not achieve this goal. It is necessary to adopt the big-data-based cognitive computing technology to help to understand massive data and customize content security strategies. Aiming at four practical problems including telecommunication fraud governance, bad message governance, variant message governance and bad website governance, the big data cognitive computing solutions were put forward, and the practical results were given. The results show that the solutions could find the bad information quickly, and improve the quality of the content security governance effectively.

Key words
big data, cognitive computing, content security, fraud detection

1 引言

随着人工智能技术在自然语言处理领域的突破性进展，使用计算机代替人类阅读和理解海量数据，帮助人们进行科学决策和方案制订成为可能。基于大数据的认知计算技术应运而生。随着该技术的不断成熟，其被应用到医疗、法律、教育和金融等多个领域，成为各行业的研究热点。

作为关键信息通信基础设施的运营者和维护者，运营商有义务对通信网络中传播的信息进行内容安全管控。随着信息传输速度日益加快，信息容量越来越大，信息变化速度越来越高，治理压力持续加大。面对海量数据，人工分析方法已经无法应对不良信息的快速演变。因此亟须引入基于大数据分析的认知计算技术，用其代替人工分析，自动总结最新不良信息的规律和知识，帮助内容安全管控人员快速对新型不良信息做出正确有效的响应。

虽然认知计算已经被广泛应用于多个领域，但其与内容安全治理相结合的场景尚不多见。本文讨论的内容安全治理特指不良文本内容。目前通信运营商治理不良文本内容的手段主要分为线上拦截和线下分析两种。在线上拦截中，可以配置关键词组合策略，对发送的不良文本消息进行实时拦截。在线下分析中，可以对海量数据进行大数据分析，最终实现两个目的：第一，发现线上分析无法识别的隐蔽不良文本消息，如诈骗信息与正常通信内容非常接近，很难通过定义关键词进行识别；第二，优化线上的关键词组合策略，发挥线上拦截系统的最大功效，如发现了更加精准高效的关键词，用其替换已有线上关键词。

围绕上述两个目的，本文将大数据认知计算技术创新性地应用到4个场景：诈骗信息识别与易感人群发现、不良关键词知识库构建、垃圾消息变体词自动发现以及不良域名拟态拓展。诈骗信息识别与易感人群发现是为了发现隐蔽诈骗信息，后面3个应用场景都是为了有效地优化线上关键词组合策略。其中，不良关键词知识库构建的目的是优化关键词本身以及关键词之间的布尔逻辑；垃圾消息变体词自动发现的目的是生成变体关键词策略，精准拦截变体垃圾信息；不良网站域名拟态扩展的目的是发现未知不良域名，以便将域名配置为关键词，对包含不良域名的不良文本进行精准拦截。

本文基于自然语言处理与机器学习技术提出了大数据认知计算在这4种内容安全治理问题中的解决方案，并结合案例分析展示了认知计算在内容安全治理中的实践效果。

2 应用场景1——诈骗信息识别与易感人群发现

2.1 问题背景

电信诈骗给用户带来了巨大的经济损失，其中诈骗消息是诈骗分子与受害者建立联系的重要环节。随着电信诈骗黑色产业链逐步成熟，诈骗日趋呈现专业化、精准化、隐蔽化的特点。诈骗分子通过购买黑产数据获得受害者个人信息，并在诈骗过程中准确说出受害者名字，冒充受害者的熟人，从而获得受害者的信任。不同于其他违法类信息，该类信息几乎不使用敏感词，使用文本分类技术很难将其与正常消息进行区分，误判率较高，治理效果不理

想。为了实现对该类信息的精准识别，需要使用技术手段对犯罪分子使用各种身份群发信息的行为（以下称为滥用身份行为）进行捕捉。为了实现这一目标，需要使用认知计算技术对海量非结构化信息内容进行精细化语义理解，识别其中的身份信息，并使用机器学习技术推断身份信息的归属。当发现大量身份信息附着在同一个发送者身上时，则该发送者可能是滥用称谓诈骗者。分析滥用称谓诈骗者的诈骗对象，可以得到电信诈骗易感人群。

2.2 基于大数据认知计算的解决方案

如**图1**所示，在识别滥用身份类诈骗时，首先需要使用命名实体识别技术对消息中的人名、组织机构名称、QQ号、微信号、抖音号等信息进行精准识别。关于命名实体识别的研究成果国内外已有很多[1-4]，最新的研究成果有基于BERT嵌入[5]、转移学习[6]、自注意力机制[7]等方法。一个命名实体可能代表了一种身份信息。当识别出身份信息后，还需要进一步推断身份信息属于消息发送者还是消息接收者。本文采用基于Transformer[8]的深度神经网络对身份信息的所有者进行推断，从而将不同的身份信息聚合到消息发送者和消息接收者上。选择Transformer主要有两个原因：第一，Transformer模型的多头自注意力网络能够自动学习输入文本中词语之间的任意距离的依赖关系；第二，Transformer模型的位置编码机制将词语的位置信息也融合到词嵌入中，这就保证称谓在开头或结尾时，模型的自注意力网络能够有效地感知位置信息，进而通过位置信息对称谓的归属进行准确的推断。

命名实体归属的推断通常需要考虑命名实体所在的上下文，如命名实体的前序词语为“尊敬的”，则显然该命名实体归属于消息接收者；而若命名实体的前序词语是“我是”，则归属于消息发送者。同时命名实体所在消息位置也直接影响了消息归属，如命名实体在消息开头则属于接收者，在消息结尾则属于发送者。在推断命名实体归属时，Transformer可以充分考虑消息中的每一个词对命名实体归属的影响，同时还可以通过位置编码技术考虑命名实体所在的位置信息，因此能够准确地推断出命名实体的归属。

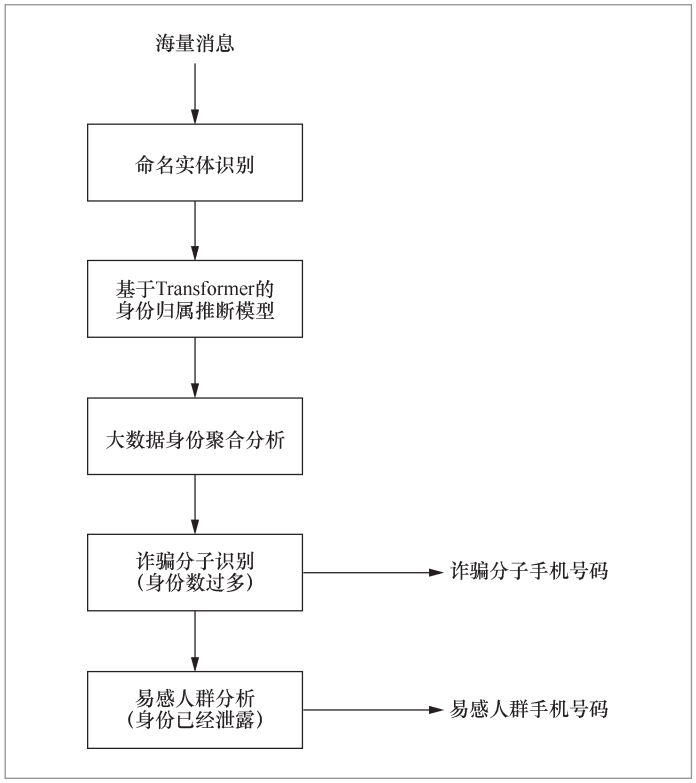

图1 滥用身份类诈骗的认知计算技术解决方案

可以使用图数据库对分析出的海量号码关联身份信息进行存储，并通过图计算，快速找到身份信息过多的消息发送者。一般情况下，当一个消息发送者使用的身份信息超过10个时，则可以判定消息发送者为诈骗分子。当一个消息发送者被判定为诈骗分子后，其所发送信息的接收者均为潜在的诈骗受害者。同时若信息中有

信息接收者的身份信息，则证明信息接收者的身份已经泄露，其还有可能被其他诈骗分子当作潜在的诈骗目标，属于电信诈骗的易感人群。针对该类易感人群，可重点进行反电信诈骗的宣传教育。

2.3 实践案例

图2是通过分析海量真实数据得到的滥用称谓诈骗示例，每个类型的示例消息为同一个号码发送。加粗的字段为算法识别出的称谓信息，为了保护个人信息，示例中的称谓信息已被模糊化。从消息内容可看出，消息的发送者称谓信息可能会出现在消息的开头、中间或结尾，模型都能够进行较好的称谓分辨。上述例子中每一种诈骗的发送者实际上都被模型赋予了20个以上的身份信息，此处限于篇幅仅各列出3个。

通过分析海量消息中的命名实体归属，将消息中的命名实体聚合到消息的发送者和接收者上，可以快速分析出滥用或伪造身份的诈骗消息发送行为。在实践中，该算法每天可发现滥用称谓类垃圾消息近百万条，治理成效显著。另外，由于该方法从诈骗分子伪造身份这一本质特征进行分析，并不依赖于具体的诈骗套路，故诈骗分子很难通过改变诈骗套路绕过该方法。

综上所述，通过使用大数据认知计算中的自然语言处理技术，提取海量非结构化文本中的命名实体，再通过机器学习技术使用Transformer模型学习如何推断命名实体属于消息发送者还是接收者，可以有效地将命名实体按照消息的发送者聚类，从而找到具有过多命名实体的消息发送者，进而确定诈骗分子的手机号码。

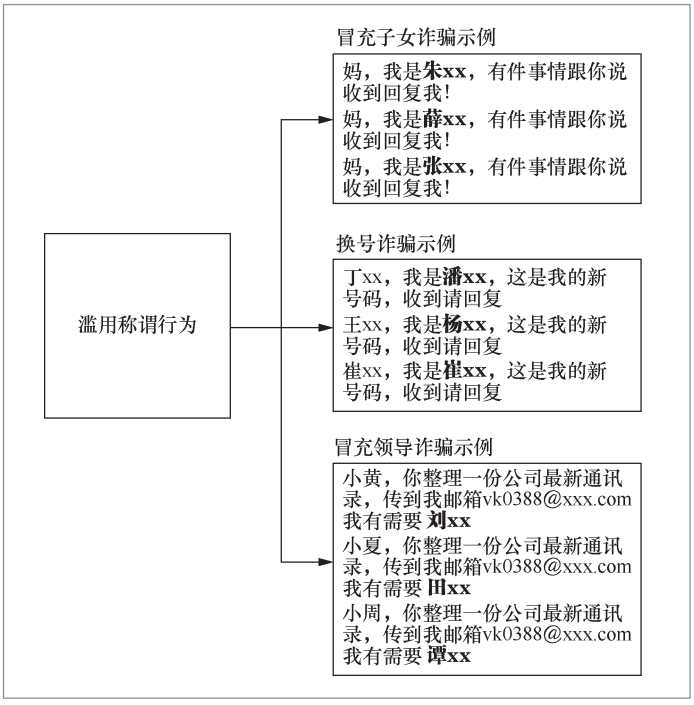

图 2 滥用称谓诈骗示例

3 应用场景2——不良关键词知识库构建

3.1 问题背景

运营商在进行不良文本消息治理时，通常使用关键词组合策略。关键词组合策略由一系列关键词和“与”“或”逻辑有机构成。当一条信息中包含策略定义的关键词且满足策略定义的逻辑组合时，该信息就会被判定为违规信息。关键词组合策略通常由人来定义。策略制订人员需要根据不同的不良文本消息特征定义不同的关键词组合策略，过程费时费力，且覆盖不全面。当策略数量达到上千条时，人工维护每一条策略的生命周期变得不可行。

此外，不同水平的策略制订人员制订的策略也存在较大的质量差距。普通策略制订人员在制订一条策略时往往聚焦于少量特定不良信息，只有有经验的策略制订人员才会进行策略的适度拓展，提高策略

泛化能力。通过大数据认知计算技术，将海量不良信息凝练成不良关键词知识库，可以帮助缺乏经验的策略制订人员进行适度的拓展发挥。

为了达到上述目的，需要使用认知计算技术分析海量非结构化垃圾文本消息，使用深度学习与自然语言处理技术自动挖掘垃圾文本中不良关键词之间的“共现”和“替代”关系，并形成知识库。具体地，具有替代关系的两个关键词经常在相同的语境中出现，如“美国”和“漂亮国”在政治类消息中共享相同的语境，可相互替代。若要自动发现具有替代关系的关键词，需要使用深度学习技术计算每个词语的上下文语境表示，并计算语境之间的相似度，相似度越大，则两个词语之间的替代性越强。替代关系可以帮助策略管理人员拓展现有策略的“或”逻辑。

具有共现关系的两个关键词经常在相同类型的消息中一同出现：如“代开”和“发票”经常在涉黑类消息中出现。在进行共现关系挖掘时，不但要考虑两个词语在消息中共同出现的概率，还需要考虑其对不良消息的判别作用，可以通过机器学习技术构建文本分类模型来评价不同词语共现特征对分类结果的影响，影响越大，则共现关系越强。策略管理人员可以通过共现关系拓展策略的“与”逻辑。

3.2 基于大数据认知计算的解决方案

关键词的属性信息中的类别倾向性和热度比较容易使用大数据统计的方法获得，统计关键词在相应类别下的频次即可。这里不再赘述。

关键词的替代关系可以通过基于词嵌入层的文本分类器来实现。词嵌入层可以将输入的关键词转化为稠密空间中的一个向量表达。当分类器进行训练时，词嵌入层将为不同词语的向量表达进行优化，使得不同类别倾向性的词语距离拉长，相同类别倾向性的词语距离缩短。当在特定类别下两个词语具有相互替代效果时，两个词语的距离非常接近。可使用两个向量的余弦距离量化关键词替代关系的强弱。带有词嵌入层的文本分类模型有很多。例如，Ge L H等人[9]通过词嵌入模型来优化文本分类性能；Liu Q等人[10]将面向特定领域的词嵌入模型用于文本分类；同时标准Transformer网络也包含词嵌入层，Shaheen Z等人[11]将Transformer应用于文本分类任务。另外，王玲[12]将词嵌入与长短期记忆（long short-term memory，LSTM）网络进行组合，形成分类器。对于短消息分类场景，任选一种结构较简单的包含词嵌入层的分类器即可满足要求。

关键词的共现关系可以使用基于卷积与注意力机制神经网络的分类器来实现。卷积窗口的大小决定了共现词语的个数。卷积特征图中的每一个元素代表了一种词语共现关系。这些共现关系对分类结果会有不同程度的影响，注意力层会将这些影响量化为权重。当分类器输入一条消息时，可以通过注意力矩阵权重找到与消息类别关联最紧密的词语共现关系。对每条消息都提取最重要的词语共现关系，并进行统计聚合。可以实现对关键词共现关系网络的快速挖掘。将卷积与注意力机制组合的分类器较丰富，如Du J C等人[13]提出了卷积循环注意力网络（convolutional recurrent attention network，CRAN）；Gao S等人[14]构建了一种层次化的卷积注意力网络，从词级和句子级两个层次对文档进行分类；Liu G等人[15]和Zheng J等人[16]将卷积网络、双向LSTM网络与注意力机制进行了不同的组合尝试，并获得了不错的效果；闫跃等人[17]使用多重注意力机制与卷积网络结合，形成文本分类器。对于消

息类短文本分类，采用卷积循环注意力网络已经足够。

综上所述，关键词的替代关系与共现关系需要训练一个同时包含词嵌入、卷积层和注意力层的神经网络。如**图3**所示，卷积循环注意力网络将词嵌入层通过卷积操作后输出到注意力层，注意力信号的每一个元素代表了一种词语共现关系。通过训练该模型得到词嵌入表达，同时在输入消息时得到消息中词语共现关系权重。这些信息可以用于计算关键词替代关系和共现关系。

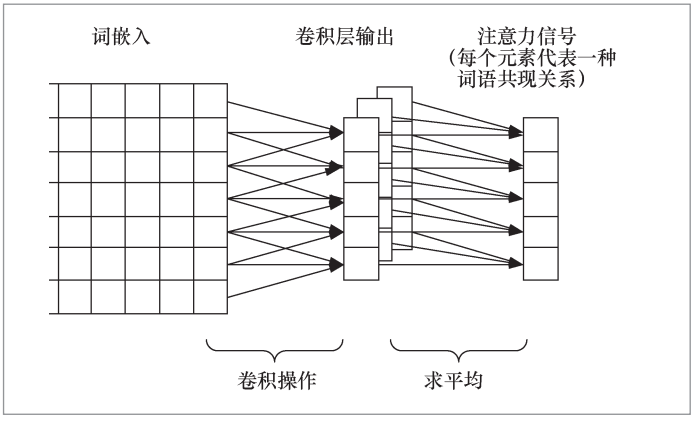

图 3 CRAN 核心网络结构

3.3 实践案例

图4展示了模型在真实短消息数据中的输出数据示例。当将海量消息输入卷积循环注意力网络后，通过观察注意力网络的最大权重可以得到每条消息最重要的共现关系。**图4中**案例使用的卷积窗口大小为3，因此共现关系表现为3个连续的词语共同出现的特征。通过统计海量消息的共现特征，可以得到右侧的知识库。知识库中的节点为共现特征库中的词，节点之间的边描述词之间的关系。**图4**中“全场”和“低至”出现频次较高，则可以构建两者之间的“共现”关系连接。通过进一步计算节点的词嵌入之间的余弦相似度，可以获得替代关系，如“元”和“折”两者的词嵌入较为接近，故二者存在替代关系。通过如上知识，可以生成策略“(元|折)&低至”，即“元”和“折”是“或”逻辑，二者与“低至”形成“与”逻辑。

策略制订人员和管理人员借助不良关键词知识库可以快速对最新的不良信息

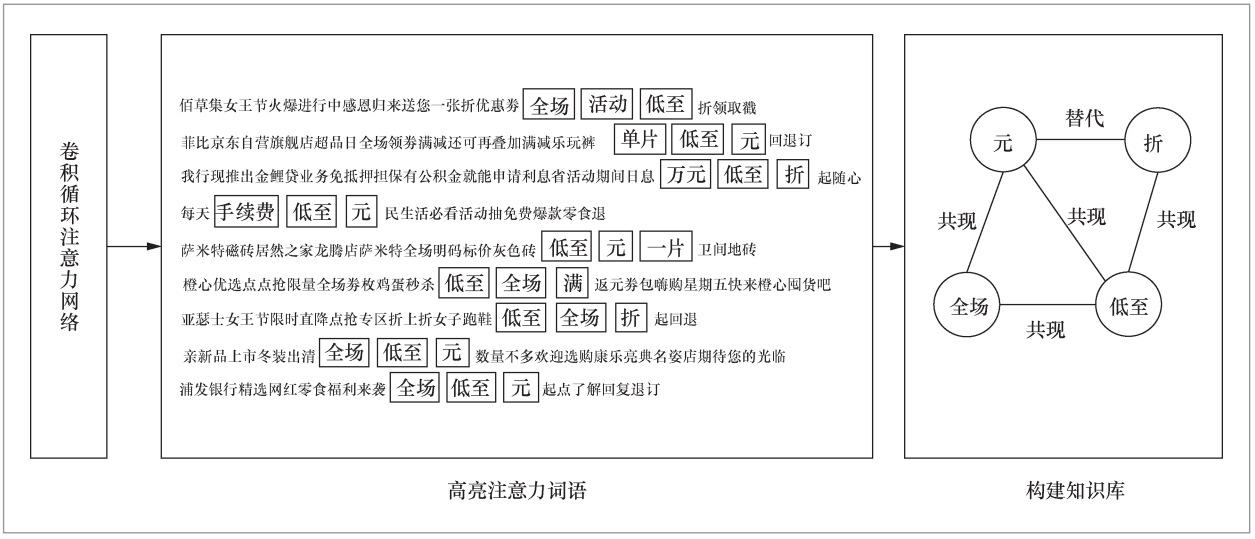

图 4 不良关键词知识库构建案例

提取关键词并形成策略，从而提高不良信息的识别质量。基于该知识库开发的策略查准优化功能能够平均提升策略查准率15%，基于该知识库开发策略查全优化功能能够平均提升策略贡献力10%。基于该知识库研发的策略自动优化流程能够大大提升策略制订人员应对新型不良信息的响应速度（由小时级别提升到分钟级别）。

综上所述，在使用大数据与认知计算前，将不良信息转化为关键词策略主要依靠人的智慧和经验，这些智慧和经验并没有外化为知识库作为长期的知识沉淀。本文提出了一种自动从海量数据中自动学习不良词语“替代”关系和“共现”关系的方法，并将学习到的关系构成不良关键词知识库，借助知识库可实现不良信息到关键词策略的自动转化。具体地，本文应用大数据认知计算中的机器学习技术对文本进行自动分类，模型选择包含词嵌入层、卷积层和注意力层的神经网络模型。在模型训练完毕后，可根据模型预测阶段得到的神经网络权重反推显著的不良词语“替代”和“共现”关系特征。将这些关系形成知识库可帮助策略制订和管理人员自动地完成从不良信息到关键词策略的高质量转化。

4 应用场景3——垃圾消息变体词自动发现

4.1 问题背景

随着运营商对垃圾消息的持续治理，垃圾消息发送者开始在消息中引入大量变体关键词，以规避关键词审查。变体关键词将敏感关键词中的字用同音字、形近字、拼音或拼音首字母、特殊符号等方式进行替换。不同于其他关键词，变体关键词几乎不会在正常消息中出现，因此及时准确发现变体关键词，并制订关键词策略可以高效、准确地实现变体垃圾消息拦截。

通常一个敏感关键词可以衍生出数十种甚至上百种变体，且变体会随时间不断变化。只有及时了解敏感关键词变体的发展变化情况，才能快速对最新关键词变体进行响应。但采用人工总结的方式很难实现上述目标，需要使用大数据认知计算技术自动分析海量垃圾信息，并理解和推断出其中包含的变体关键词。

具体地，在给定一条变体垃圾信息时，首先需要使用深度学习技术对变体垃圾信息的本体进行智能还原。该过程同时考虑变体消息中每个字的发音、字形和所处上下文，对每个字是否需要还原进行判断，若需要还原，则自动给出还原结果。如“菠菜网站”是“博彩网站”常用的变体消息，“菠菜”是否要还原为“博彩”首先要看“菠菜”本身的发音，其次还需要看其后面是否为“网站”。

在对变体消息进行还原后，可对还原后的消息进行敏感词分析，并在变体中反推出敏感词变体。如还原后，信息中“充值”可能在变体信息中是“冲值”，那么“冲值”为“充值”的变体关键词。通过分析海量变体消息，可以总结大量变体关键词，这些关键词大多不会在正常消息中出现，故可以将其配置为关键词策略以进行消息拦截。如策略“(枰部|坪部|評苔|苹苔|坪苔|呯部)”配置了“平台”这个关键词的各种变体。消息中只要包含其中一个变体，则会被立刻拦截。

4.2 基于大数据认知计算的解决方案

变体关键词推断的灵感来源于拼音输入法的实现方法。在拼音输入法中，给定拼音序列，输入法可以给出拼音序列

对应的最可能的中文句子。在拼音输入法功能中，拼音序列中每一个拼音最终对应输出的一个文字。这是一个典型的序列到序列的映射学习问题。可以使用LSTM、Transformer等深度学习模型实现映射学习。由于Transformer模型可以更好地处理长距离依赖关系，本文选用Transformer模型。具体地，Transformer可以从拼音序列中任何有帮助的位置来推断当前拼音对应的文字，其变体还原能力比LSTM更强，这种长距离拼音的推理对于变体还原任务非常重要，会直接影响变体还原的效果。

在给定变体消息时，首先将变体消息转换为拼音序列，再通过深度神经网络推理最可能的原始消息内容。通过对比还原后的消息与变体消息的差异，可以锁定消息中出现的变体关键词。变体消息中可能会有特殊符号，需要为特殊符号分配相应的发音。如给“+”分配发音“jia”。当特殊符号的发音不易确定时，可为其分配一个唯一的虚拟发音，如给“/”分配虚拟发音“zxg”（即“左斜杠”的拼音首字母，虚拟发音可任意指定）。同时，在变体消息中还会出现拼音本身或英文缩写，可以在转换拼音序列时直接保留，不做转换。

当消息中的关键词变体为同音变体时，将消息转化为拼音序列后，同音文字变体差异被消除，其完全转化为从拼音序列推测文本内容的任务，因此推测识别率较高。但当变体关键词为形近变体时，变体关键词的发音有可能与原始关键词不同，会干扰模型的推理。

为了解决这一问题，可以通过向输入拼音中加入智能干扰的方式增强模型的还原能力。此时，输入拼音序列中每个元素不再是一个拼音，而是多个拼音。其中一个拼音为正确拼音，其他拼音为干扰拼音。在训练模型时，可完全将不带变体关键词的消息作为训练数据，消息本身是模型期望的输出，消息的输入为带智能干扰的拼音序列。具体的智能干扰方式如下。

针对消息中的每一个字，需要生成n个拼音。其中一个拼音是该字本身的发音，其余拼音有如下生成规则：当该字有形近字，且拼音与该字不同时，则加入形近字的拼音，可以加入多个；当该字有相似的特殊符号可以表示时，加入特殊符号的拼音。如果上述两种干扰拼音都加入后仍不足n个，则考虑随机加入拼音。在模型进行预测时，可将输入变体消息的第一个字转为形近字拼音和特殊字符拼音，若不足n个拼音，则加入一个空拼音，使随机干扰尽可能变小。综上所述，通过在训练时增加更多随机干扰，模型可以在预测时有更强的还原能力。通过在预测时仅加入文字本身、形近字和形近特殊字符发音，不加入随机发音，可让模型专注于对这几类变体进行推理。

图5所示为一个对Transformer网络进行改造得到的变体消息还原网络。与标准Transformer网络不同，该网络在多头自注意力模块与嵌入层之间加入了拼音融合层。该层主要将干扰发音叠加到原始发音之上，使Transformer网络能够学习对抗这种干扰发音的叠加。

4.3 实践案例

图6所示为变体还原模型对6条真实垃圾消息的还原结果。其中，第1条消息中的“蕞篙”被成功恢复为“最高”，属于同音和形近字双重变体复原；第4条消息中的“筷③”被成功恢复为“快三”，包含了特殊字符的变体复原；第4条消息中的“蝉遰”被成功恢复为“单带”，属于形近不同音变体的复原。由此可见，模型能够支持对形近、同音、特殊字符变体的复原。

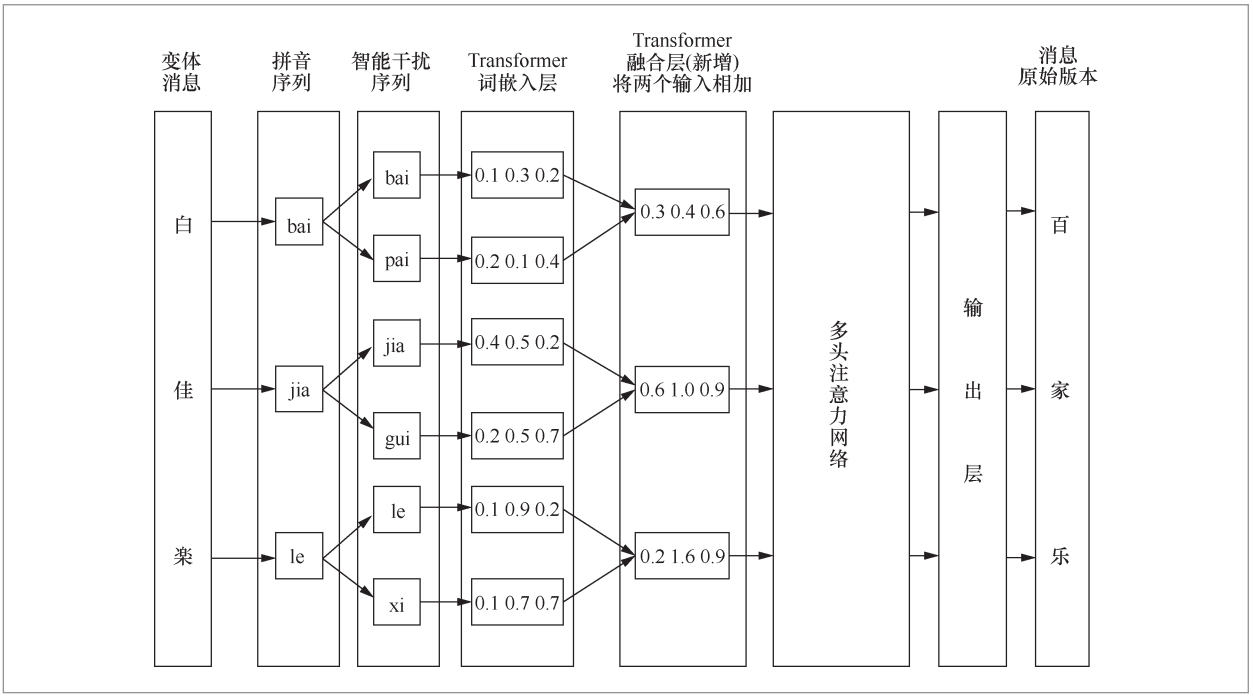

图 5 基于 Transformer 网络的变体消息还原网络

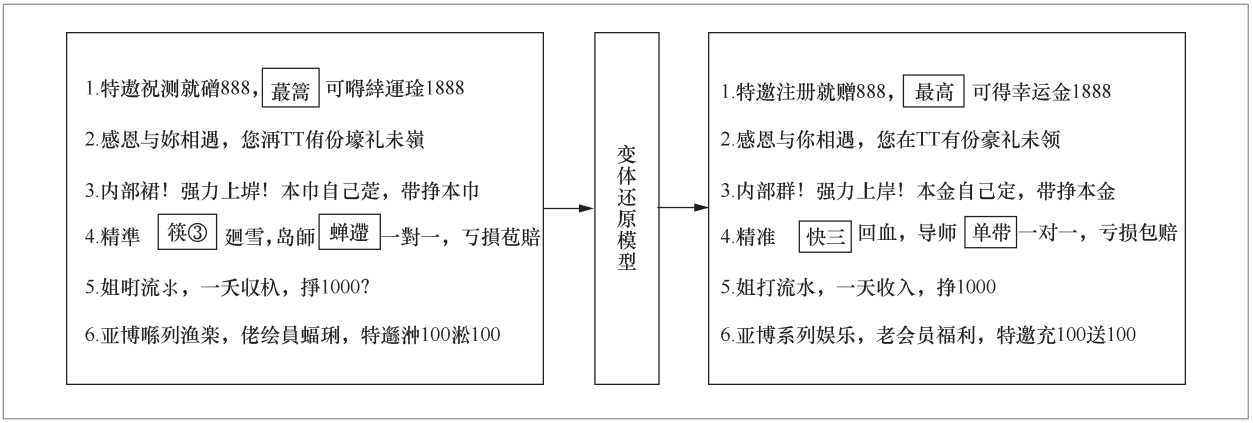

图 6 使用真实变体垃圾信息还原效果示例

表1为从**图6**的变体垃圾信息中自动提取的变体词列表。变体词通过对还原后的文本进行分词后反推而得。其中大部分变体词是同音变体词，这也符合真实的垃圾信息使用变体的情况。变体还原模型同时考虑了变体词的发音和其形近字的发音，故能够有效地对这些变体进行还原。此外，这些变体词在正常消息中几乎不可能出现，故可将这些变体词配置为关键词策略用于对变体垃圾信息进行快速拦截。

实践证明，使用变体还原模型可有效地还原垃圾消息中的大部分变体。通过比较还原前后的文本，可以快速定位敏感关键词的变体。通过该方法可迅速构建出不良关键词变体库，基于变体词库输出的变体关键词策略在实际应用中一周可以识别和拦截数十万条变体垃圾信息，有效地解决了变体垃圾消息的漏拦问题。

表1 从变体垃圾消息中提取的变体词

消息	变体	本体	消息	变体	本体	消息	变体	本体
1	特邀	特邀	3	上埠	上岸	4	苞賠	包赔
1	祝测	注册	3	本巾	本金	5	流氷	流水
1	蕞篙	最高	4	精準	精准	5	一天	一天
1	可嗬	可得	4	筷③	快三	5	収杕	收入
1	絳運琿	幸运金	4	廻雪	回血	6	喙列	系列
2	侑份	有份	4	岛師	导师	6	渔楽	娱乐
2	壕礼	豪礼	4	蝉遰	单带	6	佬绘員	老会员
2	未嶺	未领	4	一對一	一对一	6	蝠琍	福利
3	内部裙	内部群	4	亏損	亏损	6	特邋	特邀

综上所述，变体垃圾信息对垃圾信息的识别造成了巨大干扰，一些变体甚至可能会迷惑人的审核判断。本文利用大数据认知计算技术中的机器学习技术学习拼音序列到文字序列的正确转化。

5 应用场景4——不良域名拟态拓展

5.1 问题背景

开设赌博、色情网站在国内属于违法行为，因此不良网站的服务器通常不在国内，运营商无法对服务器直接进行处理，仅能对服务器的域名进行封堵。不良网站创建者为了规避封堵风险，会集中生成一批风格相近的域名，一些域名一旦被封，立刻切换域名，并不影响用户访问。

目前运营商发现不良域名的方法是分析用户访问域名本身是否具有不良特征、对应网站中的文本和图片信息是否包含敏感内容等。这些方法多是在用户发生访问行为后再进行网站识别的。一方面访问网站的事实已经发生，已经造成了一定的不良影响；另一方面封堵时并没有考虑被封网站可能有备用域名的问题，封堵不彻底。

一些有经验的不良网站审核员可以通过被封堵的不良网站域名规律推测出其他未知的不良网站域名，这样可以在网络中还没有出现用户访问该域名的记录的前提下发现这些不良网站，如已知“xx991.com”和“xx993.com”是不良域名，则很可能“992xx.com”也是一个不良域名。这些不良网站的规律千差万别，采用人工的方式很难全面总结。需要使用认知计算技术自动学习已知的不良网站域名特征，并自动模仿不良域名的表现形态，举一反三，生成形态相似的潜在不良域名。具体地，此过程主要涉及使用深度学习技术帮助人们自动学习和理解海量不良网站域名的格式特征、字符关联、字符与数字的组合特点，并根据学到的规则自动创造全新的符合规则的潜在不良域名。通过对生成的潜在不良域名进行内容分析，最终确认未知不良网站。

5.2 基于大数据认知计算的解决方案

为了实现不良网站的拟态拓展能力，可以使用双向LSTM模型对已知不良网站的构成特征进行学习。具体训练步骤是在给定不良网站域名中的任意n个字符后，预测不良网站域名的下一个字符。若模型能够在给定任意已知域名的任意n个连续字符后，都可以准确预测下一个字符，则代表模型

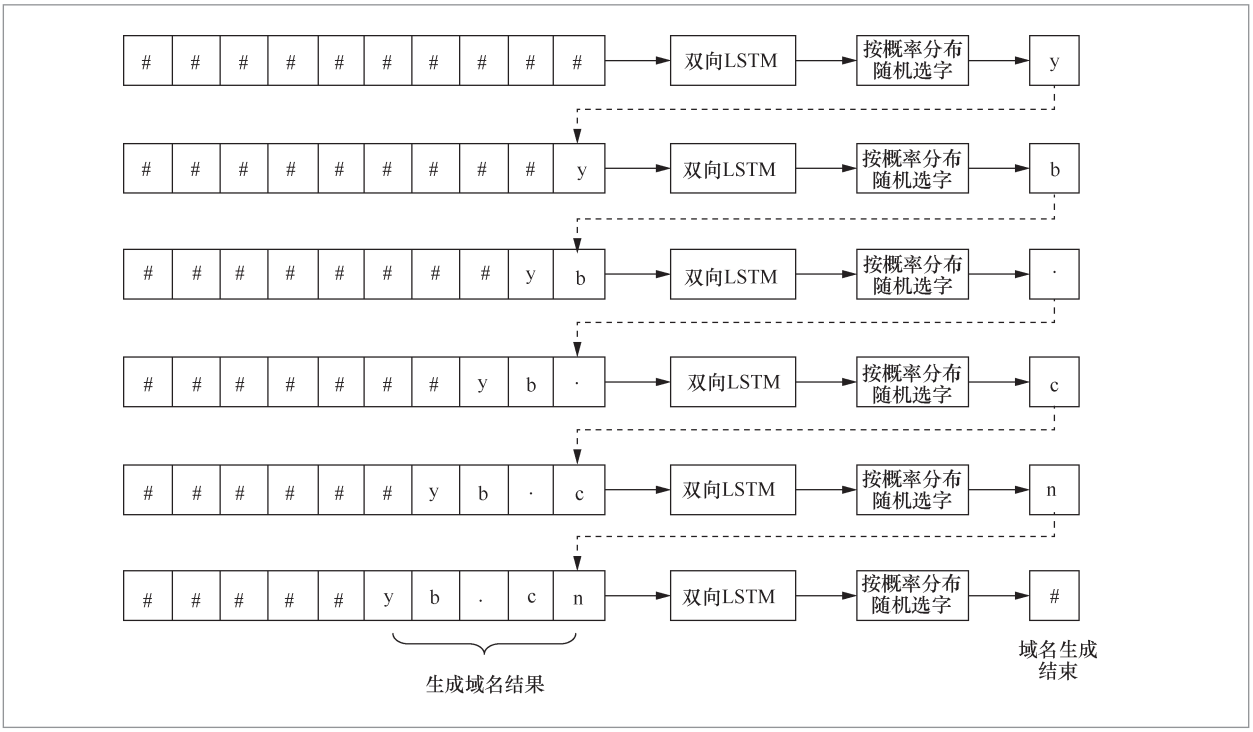

图 7 双向 LSTM 生成域名的过程示意图

已经充分学习了已知不良域名的字符构成特征，就可以进行相同形态域名的智能生成。

双向LSTM生成域名的过程如**图7**所示。在生成一个域名时，首先向模型中输入n个空字符（**图7**中为10个），则模型会输出域名的第一个字符，接下来将模型刚输出的字符加入输入，则输入变为$n-1$个空字符和最新输出的字符。将该输入再输入模型，模型会继续输出下一个字符。依此类推，不断将模型输出的字符加入输入中，则输入一直保存最近模型输出的连续n个字符，并不断输出下一个字符，直到输出空字符为止。此时一个域名生成完毕。

采用上述生成方法虽然可以得到形态相似的域名，但生成的域名较大概率为已知不良域名本身。为了让模型在模拟形态的基础上发挥自身的创造力，可以在生成下一个字符的过程中加入一些随机性，即并不总是选择推测概率最大的字符作为输出字符，而是按照推测的各种字符的出现概率进行随机选择，如**图7**所示。

除了使用双向LSTM模型，很多文本生成模型也可以完成域名生成的任务，数据的训练方法和文本的生成方法与双向LSTM模型相同。如许晓泓等人[18]使用Transformer模型完成从数据到文本的生成过程；Pawade D等人[19]使用字级别的RNN-LSTM生成文本；钱揖丽等人[20]提出了基于句子级LSTM编码的文本标题生成模型等。由于域名结构相对简单和简短，不太可能出现字符之间的长距离依赖，故采用双向LSTM已经足够实现域名的拟态拓展。

5.3 实践案例

从训练数据中找到所有包含“av”和“zy”两种模式的不良域名，并在模型生

成的不良域名中寻找上述两种特征，可以分析模型如何利用训练数据中的模式拓展生成域名。

图8为双向LSTM模型的训练数据模式与拓展数据模式。为了避免传播不良网站域名，**图8**中对不良网站域名进行了模糊化处理，“#”代表任意一个数字，“*”代表任意一个字符。如**图8**所示，双向LSTM模型不但可以模仿训练数据中的已有模式，还可以创造更多全新的域名模式。按照这些域名模式可以发现更多不良网站。将被确认为不良网站的域名新模式加入训练数据中，可以加强LSTM对新不良模式的学习，如此循环可以形成一个不良域名特征自动学习更新拓展的闭环。

研究发现，使用不良域名拟态拓展能力学习3 000个不良域名后，每生成10 000个不良域名，平均有大约18个域名是重复的，重复率为0.18%。通过使用爬虫进行内容验证，发现平均有2 032个域名是真实存在的，平均有876个域名为真实的不良域名。从生成域名到最终发现不良域名，转化率大约为8.76%。将不良域名拟态拓展能力应用于实际工作中，每天可以发现上千个活跃的未知色情、赌博类网站，使不良网站的封堵更加主动、彻底、高效。

综上所述，不良网站通常会注册风格相似的域名。人为观察已有不良域名特征预测未知不良域名工作量巨大，且仅能进行小范围的尝试。本文利用大数据认知计算技术中的自然语言生成能力，将域名信息看作一种自然语言，使用LSTM模型对海量不良域名构建语言模型，并实现了模仿不良域名特征拓展生成全新不良域名的能力。实践证明，该算法能够发现大量未知的不良域名，实现了不良域名的主动发现、事前发现。

6 结束语

通信运营商在进行内容安全管控的过程中遇到了诸多需要进行海量数据分析理解的问题。在使用大数据认知计算前，这些任务多采用人工分析的方法，数据处理能力有限，治理效率不高。大数据认知计

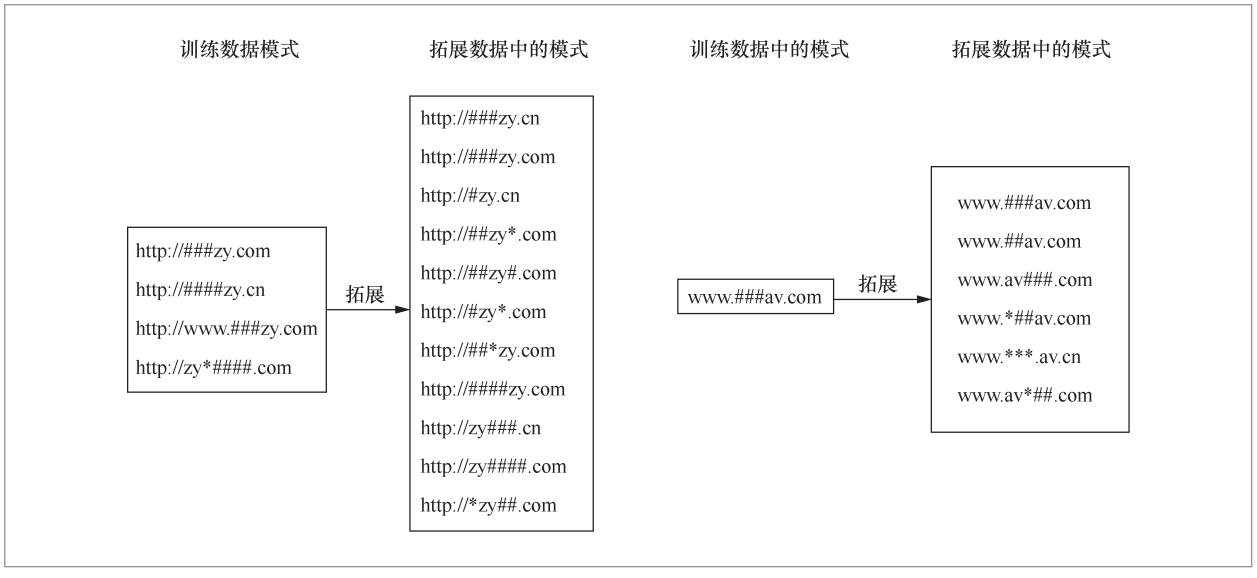

图 8 双向 LSTM 模型的训练数据模式与拓展数据模式

算技术可以帮助安全管控人员分析理解海量数据，发现更多不良信息，大幅提高不良信息的治理效率。本文从不良文本线下分析的两个目的入手，总结了大数据认知计算在诈骗信息识别与易感人群发现、不良关键词知识库构建、垃圾消息变体词自动发现、不良域名拟态拓展4个内容安全领域的创新性实践。

上述大数据创新实践方案有效地使用大数据认知计算替代了人工，帮助人们理解海量不良信息的关键内容，大力支撑了内容安全管控工作。实践研究证明，本文提出的应用方案能够帮助内容安全管控人员快速响应最新不良信息，全面有效提升整体管控质量。

参考文献:

[1] YADAV V, BETHARD S. A survey on recent advances in named entity recognition from deep learning models[J]. arXiv preprint, 2019, arXiv:1910.11470v1.

[2] LI J, SUN A X, HAN J L, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020: 1.

[3] GOYAL A, GUPTA V, KUMAR M. Recent named entity recognition and classification techniques: a systematic review[J]. Computer Science Review, 2018, 29: 21–43.

[4] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329–340.
LIU L, WANG D B. A review on named entity recognition[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(3): 329–340.

[5] 杨飘, 董文永. 基于BERT嵌入的中文命名实体识别方法[J]. 计算机工程, 2020, 46(4): 40–45, 52.
YANG P, DONG W Y. Chinese named entity recognition method based on BERT embedding[J]. Computer Engineering, 2020, 46(4): 40–45, 52.

[6] 周法国, 吴锡坤, 孙泰, 等. 基于转移学习的中文命名实体识别[J]. 计算机工程与应用, 2018, 54(5): 117–121.
ZHOU F G, WU X K, SUN T, et al. Chinese named entity recognition based on transformation learning[J]. Computer Engineering and Applications, 2018, 54(5): 117–121.

[7] 李明扬, 孔芳. 融入自注意力机制的社交媒体命名实体识别[J]. 清华大学学报(自然科学版), 2019, 59(6): 461–467.
LI M Y, KONG F. Combined self-attention mechanism for named entity recognition in social media[J]. Journal of Tsinghua University (Science and Technology), 2019, 59(6): 461–467.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000–6010.

[9] GE L H, MOH T S. Improving text classification with word embedding[C]// Proceedings of 2017 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2017: 1796–1805.

[10] LIU Q, HUANG H Y, GAO Y, et al. Task-oriented word embedding for text classification[C]//Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 2023–2032.

[11] SHAHEEN Z, WOHLGENANNT G, FILTZ E. Large scale legal text classification using transformer models[J]. arXiv preprint, 2020, arXiv:2010.12871.

[12] 王玲. 基于Word2Vec词嵌入和双向长短时记忆网络的文本分类实现[J]. 电子技术与软件工程, 2020(15): 70–71.
WANG L. Implementation of text classification based on Word2Vec word embedding and bidirectional long-term

and short-term memory network[J]. Electronic Technology & Software Engineering, 2020(15): 70-71.

[13] DU J C, GUI L, XU R F, et al. A convolutional attention model for text classification[M]//Natural Language Processing and Chinese Computing. Cham: Springer, 2018: 183-195.

[14] GAO S, RAMANATHAN A, TOURASSI G. Hierarchical convolutional attention networks for text classification[C]// Proceedings of the 3rd Workshop on Representation Learning for NLP. Stroudsburg: Association for Computational Linguistics, 2018: 11-23.

[15] LIU G, GUO J B. Bidirectional LSTM with attention mechanism and convolutional layer for text classification[J]. Neurocomputing, 2019, 337: 325-338.

[16] ZHENG J, ZHENG L M. A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification[J]. IEEE Access, 2019, 7: 106673-106685.

[17] 闫跃, 霍其润, 李天昊, 等. 融合多重注意力机制的卷积神经网络文本分类设计与实现[J]. 小型微型计算机系统, 2021, 42(2): 362-367.
YAN Y, HUO Q R, LI T H, et al. Design and implementation of text classification based on convolutional neural network with multiple attention mechanisms[J]. Journal of Chinese Computer Systems, 2021, 42(2): 362-367.

[18] 许晓泓, 何霆, 王华珍, 等. 结合Transformer模型与深度神经网络的数据到文本生成方法[J]. 重庆大学学报, 2020, 43(7): 91-100.
XU X H, HE T, WANG H Z, et al. Research on data-to-text generation based on transformer model and deep neural network[J]. Journal of Chongqing University, 2020, 43(7): 91-100.

[19] PAWADE D, SAKHAPARA A, JAIN M, et al. Story scrambler - automatic text generation using word level RNN-LSTM[J]. International Journal of Information Technology and Computer Science, 2018, 10(6): 44-53.

[20] 钱揖丽, 马雪雯. 基于句子级LSTM编码的文本标题生成[J]. 计算机应用与软件, 2021, 38(5): 190-195.
QIAN Y L, MA X W. Text headline generation based on sentence-level LSTM encoding[J]. Computer Applications and Software, 2021, 38(5): 190-195.

作者简介

杜雪涛(1973-), 女, 中国移动通信集团设计院有限公司网络规划与设计优化研发中心网信安全产品部教授级高级工程师, 主要从事网络与信息安全研究工作。

收稿日期: 2021-09-01