5G消息中不良文本消息识别技术研究

杜雪涛，张晨，杜刚，王红雨

（中国移动通信集团设计院有限公司，北京 100080）

【摘　要】　5G消息通信能力极大增强，其中单条文本消息可达千字，消息的即时交互性更强，并且不良消息中常充斥大量变体字，这给现有的不良文本消息识别技术带来挑战。本文调研了关键词策略匹配技术和人工智能模型两种不良文本识别方法，总结了它们的优缺点，并针对上述挑战提出了考虑关键词上下文区间约束的策略匹配技术、跨消息关键词匹配技术和无监督分词技术，可以显著增强5G消息中不良文本识别的效果。

【关键词】　不良文本识别；策略匹配；关键词匹配；分词技术

1　引言

5G消息引入了更强的文本消息能力，单条消息可发送上千字，交互性更强，而且不良文本消息中常夹杂大量变体关键词[1]。这些给现有的不良文本识别方案带来了如下挑战：

（1）长文本挑战：关键词组合策略通过4、5个关键词判定一条文本是否违规，但5G消息文本长度可达数千字。仅通过少量关键词判定长文本是否违规的准确率将大幅降低，因为命中关键词组合策略的关键词在消息中可能相距字符个数很多，并不属于同一个上下文环境，这将直接影响关键词组合策略识别不良信息的准确率[2]。

（2）跨消息匹配挑战：由于5G文本消息交互性增强，不良信息发送者可以通过将一条消息拆分为多条消息的方法将敏感关键词进行分割发送，如敏感词为“赌博”，其可以将“赌”作为第一条消息的结尾，“博”作为第二条消息的开头，只有将两条消息联合起来分析才能够有效对不良信息进行识别。

（3）关键词变体挑战：一些不良信息发送者采用同音字、形近字、特殊符号等方式表达敏感的关键词，从而逃避关键词审查，这些关键词变体使得关键字组合策略失效。同时，由于包含变体的不良信息无法有效分词，也会导致人工智能模型识别效果明显下降[3]。

本文针对上述三个挑战分别提出了应对方案。具体地，针对长文本挑战提出了一种约束文本上下文区间的关键词组合策略匹配方案；针对跨消息匹配挑战提出了一种高效的跨消息匹配方法；针对关键词变体挑战提出了一种无监督分词技术，提取更丰富的变体特征用以深度学习。

2　约束关键词上下文区间的策略匹配

关键词在消息中出现位置可使用AC自动机（Aho-Corasick automation，AC自动机）获得[4]。关键词的位置指文本的开头到该关键词的第一个字符之间的字符数量。如图1所示，“娱乐城”距离文本的开头有两个字符，故位置为2，同理“现金”的位置为9。将两个关键词的位置相减得到关键词之间的距离，将此距离定义为这两个关键词的上下文区间大小。本文通过限定策略关键词在文本中出现的上下文区间大小来约束关键词所在的上下文。

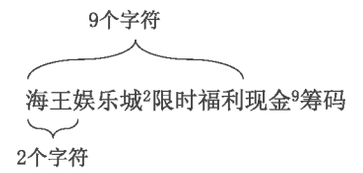

图1　关键词位置及其上下文区间示意图

2.1　策略二叉树表示

为了解析复杂的关键词策略嵌套逻辑，需将策略

表达式转换为逻辑嵌套的二叉树结构。如图2所示，任何一个策略可以转换为对应的二叉树结构。其中，二叉树的分支结点为逻辑运算符，叶子结点为关键词。具体地，可通过对二叉树进行后序遍历并结合堆栈实现转换，类似于中缀表达式借助两个栈进行求值的过程，不再赘述。

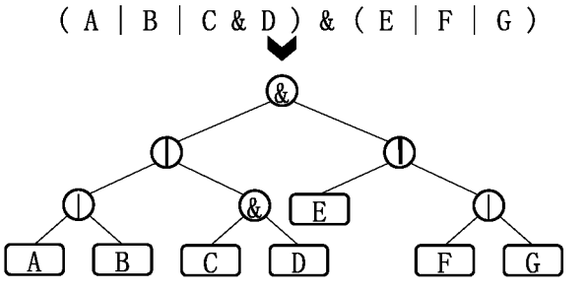

图2　策略二叉树

2.2 考虑关键词上下文区间约束的策略匹配

首先使用AC自动机获取到一条文本中包含的关键词及其位置信息，然后筛选出包含这些关键词的策略，对于每一条策略可以使用如下步骤完成考虑关键词上下文区间的策略匹配。下面结合图3详细说明。

（1）将策略使用二叉树结构表示。当叶子结点的关键词是匹配的关键词时，叶子结点加入True及其所在文本中的位置信息；否则，加入False。图3为策略“（A&B&D|E&F）&（G|H|J）”的二叉树表示。假设文本命中了策略关键词A、B、D、E、F、H、J，未命中关键词G，则A、B、D、E、F、H、J结点处为True，右上角的数字为关键词在文本中出现的位置；G结点处为False。

（2）后序遍历二叉树。当分支结点是“或”逻辑时，该分支结点的取值为其孩子结点布尔值取“或”。同时，可将孩子结点的上下文区间信息全部拷贝到分支结点。如图3中情况1，分支结点有两个上下文区间分别是其两个孩子结点的上下文区间。当分支结点的孩子结点为叶子结点时，可以构造新的上下文区间，区间的开始和结束相等，如图3中情况2。当分支结点是“与”逻辑时，该分支结点的取值为其孩子结点布尔值取“与”，若计算后为False，则该分支结点的上下文区间为空。若计算为True，则该分支结点的上下文区间为其左孩子的上下文区间跟右孩子的上下文区间两两组合后所得的最小覆盖区间，如图3中情况3、4所示。将计算得到的上下文区间与设定最大上下文阈值进行比较，仅保留小于阈值的上下文区间。若没有区间满足要求，则分支结点设置为False。

经过上述步骤，可以查找出满足上下文区间长度限定的关键词策略，同时还可以输出上下文区间对应的文

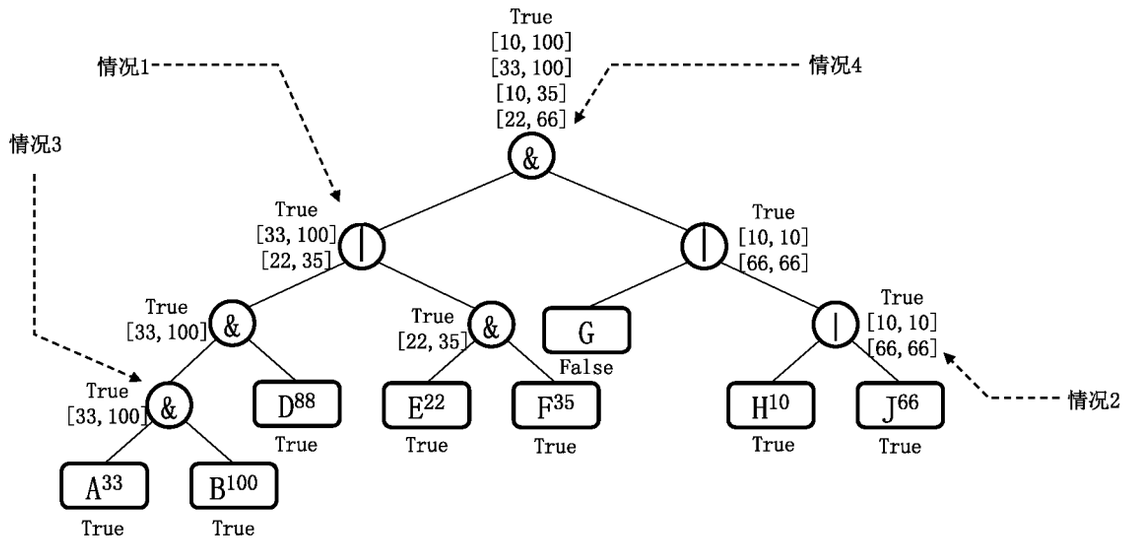

图3　策略细匹配过程示意图

本内容。使用提出的策略匹配方法，在10000条文本消息上的匹配准确率相比传统的匹配方法提高了6%，长文本消息越多提高越多，并且可输出指定长度的匹配证据，方便进行人工审核，大幅提高不良文本审核效率。

3 跨消息关键词匹配技术

3.1 跨消息缓存

为了在多条消息之间进行关键词组合匹配，本文提出首先将多条消息拼接为一条消息再进行匹配。若直接将多条消息拼接为单条消息，则消息的发送者和发送时间信息将丢失。为了保留相应信息，可在拼接前，给每条消息添加头信息，记录消息的发送者、时间等。具体地，可使用“###”作为头信息的开始和结束标志，如两条信息分别为“contentA”和“contentB”，发送者分别为C和D，则两条信息加入头信息后变为“###C###contentA”和“###D###contentB”，然后合并两条消息为“###C###contentA ###D###contentB”。使用正则表达式可将合并消息还原为原始多条消息，在进行消息合并时，可以设定最大缓存消息数量。当消息达到最大缓存消息数量后，可将最开头的消息删除，并在尾部追加新的消息[5]。

对于加入消息头的消息，若原始两条消息内容衔接处恰好命中关键词则会被遗漏，如用户A将不良信息“content”拆分为两条消息“cont”和“ent”发送，加入消息头后变为“###A###cont###A###ent”，直接在此合并消息上使用传统的关键词匹配将导致无法命中关键词“content”。为此，本文提出了一种能够自动忽略消息头的关键词匹配算法。

3.2 跨消息AC自动机匹配模型

标准的AC自动机通过建立关键词索引提高关键词匹配效率，但其不能直接用于添加了头信息的消息匹配，因为其无法自动忽略拼接消息中的头信息。本文改进了AC自动机，从而使其可以自动忽略拼接消息中的头信息。

如图4所示，左侧为关键词库构成的AC自动机，右侧为识别消息头的AC自动机。其中，0状态转变为关键词集AC自动机中的任意某个状态（图4中例子为状态16），（1）、（2）失败链接改为指向关键词集AC自动机的0状态。

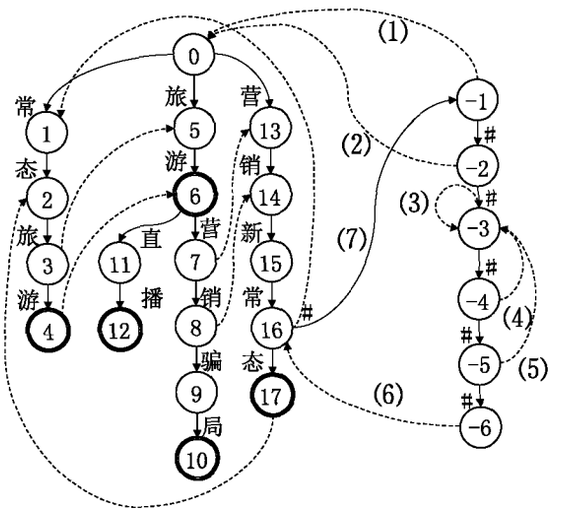

图4 改进的AC自动机模型

例如，当消息为“营销新常###A###态”时，其中“营销新常”可将自动机状态迁移到状态16，16状态中输入“#”会直接转入识别消息头的AC自动机。当消息头识别完成进入状态-6，-6状态输入任何字符会自动转回状态16，此时再输入“态”，则可以将自动机转入状态17，进而成功匹配了“营销新常态”这个关键词。但当消息为“营销新常#态”时，由于消息中的“#”并非消息头标记，自动机会进入-1状态后通过失败指针（1）返回0状态，自动机不会匹配“营销新常态”这个关键词。

传统的AC自动机是一个静态模型，当通过关键词集构建好自动机后，自动机的结构不会发生变化。本文将传统AC自动机动态化，即当自动机处于不同状态时，其结构也相应发生改变。具体的，在构建AC自动机时，仅需要构建一个消息头识别自动机，当自动机处于不同的状态时，只需要将图4中的连线（6）、（7）移动到自动机当前的状态后，再开始接收下一个字符。如图4中，自动机处于状态16，连线（7）从状态16指向状态-1，连线（6）指向状态16，当自动机处于13时，仅需将连线（7）从状态13指向状态-1、连线（6）指向状态13即可。通过如上自动机的动态变化，整个模型不再需要为每个状态都分配一个全新的消息头识别自动

机，仅需要重用一个消息头识别自动机即可，大幅降低了自动机的复杂度和内存消耗。在状态转移时，仅需要修改两个连接的起止点即可，引入的额外开销极小。

4 无监督分词技术

不良消息中常将关键词使用变体词表示，且一个关键词通常对应多种变体，使用传统的分词技术[6]无法有效对这些词语进行分词，不理想的分词结果会极大影响后续的文本分类模型的效果。为了能够自动发现不良消息中的变体关键词，本文提出了一种基于概率统计的无监督分词技术，该技术可根据不良消息中字间连接度来判断是否分割两字。

4.1 字间连接度学习

字“A”与字“B”的连接度定义为词组“AB”在语料库中出现的频次，字间连接度表示两个字之间连接的紧密程度，连接度越低，则分词时两字之间更可能被切分。计算字间连接度时需要大量的语料信息，这些语料信息无需标注，故整个计算字间连接度的过程是无监督的。

4.2 基于字间连接度的分词

给定一个句子，可根据字间连接度对该句子进行分词。首先查询得到句子中每两个相邻字之间的字间连接度信息，形成一条字间连接度序列。基于该序列，可以使用如下递归算法实现句子的分词：

（1）当句子的长度小于4（可配置）个字时，直接返回句子本身；当句子长度大于4个字时，选择字间连接度序列中连接度最小的元素，并定位到句子的具体位置进行一次切分，将句子分为左右两个子句。

（2）对左右两个子句子重复步骤（1），直至不再产生任何的切分操作。

通过上述算法可知，整个分词过程就是不断寻找句间连接度最低的点进行句子的分割，直至句子小于一定长度为止。整个过程可以使用递归算法实现，但在对长文本进行分词时容易出现递归次数过多产生堆栈溢出问题，为了进一步解决堆栈溢出问题，本文提出了一种基于Treap[7]的实现方案。

Treap是一种融合了二叉排序树和堆的数据结构[7]。如图5所示，其本身是一颗二叉排序树，即中序遍历该二叉树时，其结点中的关键值将以递增的顺序被访问。Treap不同之处在于其每个结点出了保存关键值以外还携带了一个优先级的数据属性，这个优先级属性在Treap中满足堆的性质，即每个结点的父结点的优先级都小于其孩子结点。本文将句间连接度所在句子位置作为关键值，将句间连接度作为优先级构建Treap，则最小句间连接度的切分点恰好对应Treap的根结点，Treap的任意分支结点也是潜在的分词切分点，只需要按层次遍历该Treap的各个分支结点中的关键值信息进行分词，直至分词长度小于预设值为止。使用提出的分词技术，在14000条变体消息上的分词准确率相比传统分词方法提高了10%。

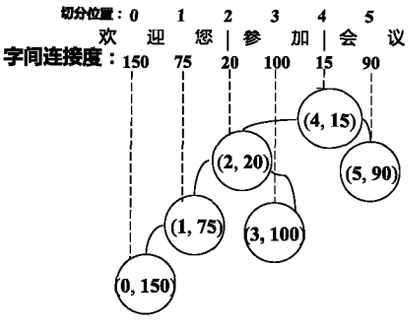

图5 Treap树示意图

5 结束语

本文通过分析5G消息中不良文本消息的特点，提出了相应的解决方案。为了应对长文本匹配挑战，提出了一种约束关键词上下文的关键词组合策略模型，该策略模型不但要求文本命中策略指定的关键词，同时还需要保证关键词出现在指定的长度的上下文中，该方法可以有效地降低关键词组合策略误匹配长文本的问题。为了应对跨消息匹配挑战，本文提出了基于AC自动机的跨消息匹配算法。该算法能够在进行消息缓存的同时保留消息的额外元信息，如发送者、发送时间等，有效地解决了关键词被拆分到多条消息中发送的问题。为了应对关键词变体挑战，本文还提出了一种无监督的分词方法，可有效提升变体词的分词效果，进而能够提升人工智能模型对文本的分类效果。

参考文献：

[1] Haiying Shen, Ze Li. Leveraging Social Networks for Effective Spam Filtering[J]. IEEE Transactions on Computers, 2014,63(11): 2743-2759.

[2] 侯整风 , 杨波 , 朱晓玲 . 一种适合中文的多模式匹配算法 [J]. 计算机科学 , 2013,40(11): 117-121.

[3] 张毓 , 陈军清 . 基于深度特征语义学习模型的垃圾短信文本聚类研究 [J]. 现代计算机 , 2018(7): 17-21.

[4] 陈新驰 , 韩建民 , 贾洞 . 基于 AC 自动机的多模式匹配算法 FACA[J]. 计算机工程 , 2012,38(11): 173-176.

[5] 侯整风 , 杨波 , 朱晓玲 . 一种节约内存的中文多模式匹配算法 [J]. 微型机与应用 , 2013,32(13): 53-57.

[6] 何莘 , 王琬芜 . 自然语言检索中的中文分词技术研究进展及应用 [J]. 情报科学 , 2008,26(5): 787-791.

[7] 刘毅 . 关于 Treap 数据结构问题的研究 [J]. 计算机应用与软件 , 2005(8): 36-38. ★

作者简介

杜雪涛：教授级高级工程师，现任职于中国移动通信集团设计院有限公司。

张晨：高级工程师，现任职于中国移动通信集团设计院有限公司。

杜刚：高级工程师，现任职于中国移动通信集团设计院有限公司。

王红雨：助理工程师，现任职于中国移动通信集团设计院有限公司。