(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110738050 A
(43)申请公布日 2020.01.31

(21)申请号 201910984742.5

(22)申请日 2019.10.16

(71)申请人 北京小米智能科技有限公司
地址 100085 北京市海淀区清河朱房路临66号F栋2单元1层101-103室

(72)发明人 齐保元 史亮 鲁骁 唐可欣 王斌

(74)专利代理机构 北京名华博信知识产权代理有限公司 11453
代理人 白莹

(51)Int.Cl.
G06F 40/289(2020.01)
G06F 40/295(2020.01)

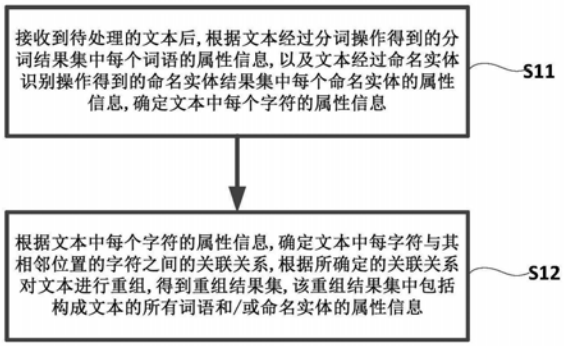
权利要求书4页 说明书14页 附图5页

(54)发明名称

基于分词和命名实体识别的文本重组方法及装置、介质

(57)摘要

本公开是关于一种基于分词和命名实体识别的文本重组方法及装置、介质，涉及自然语言处理领域。本公开提供的一种基于分词和命名实体识别的文本重组方法，包括：接收到待处理的文本后，根据文本中每个词语的属性信息，以及文本中每个命名实体的属性信息，确定文本中每个字符的属性信息；并根据每个字符的属性信息，确定每字符与其相邻位置的字符之间的关联关系，根据所述关联关系对所述文本进行重组，得到重组结果集，重组结果集中包括构成文本的所有词语和/或命名实体的属性信息。本公开的技术方案对待处理的文本进行重新组合与渲染，实现多样性标注的输出，再通过建立字符的属性信息中的权重，来实现便于用户习惯的阅读渲染方式。



CN 110738050 A

1.一种基于分词和命名实体识别的文本重组方法,其特征在于,包括:

接收到待处理的文本后,根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息,以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息,确定所述文本中每个字符的属性信息;

根据所述文本中每个字符的属性信息,确定所述文本中每字符与其相邻位置的字符之间的关联关系,根据所确定的关联关系对所述文本进行重组,得到重组结果集,所述重组结果集中包括构成所述文本的所有词语和/或命名实体的属性信息;

其中,重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

2.根据权利要求1所述的方法,其特征在于,所述分词结果集中每个词语的属性信息,所述命名实体结果集中每个命名实体的属性信息,以及所述文本中每个字符的属性信息,至少包括如下任一种或几种:内容,开始位置,结束位置,类型,标识,权重。

3.根据权利要求2所述的方法,其特征在于,

所述字符的属性信息中的权重,小于所述词语和命名实体的属性信息中的权重;

所述词语的属性信息中的权重小于或等于所述命名实体的属性信息中的权重。

4.根据权利要求2或3所述的方法,其特征在于,所述根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息,以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息,确定所述文本中每个字符的属性信息,包括:

基于所述命名实体结果集中每个命名实体的属性信息,以及所述文本中每个字符与所述命名实体结果集中每个命名实体在所述文本中的位置关系,设置所述文本中每个字符的属性信息;

基于所述分词结果集中每个词语的属性信息,以及所述文本中每个字符与所述分词结果集中每个词语在所述文本中的位置关系,设置所述文本中每个字符的属性信息;

其中,基于所述命名实体结果集和所述分词结果集,为同一字符设置出多条不相同的属性信息时,将多条不相同的属性信息中权重取值最大属性信息确定为该字符的属性信息。

5.根据权利要求1所述的方法,其特征在于,所述根据所述文本中每个字符的属性信息,确定所述文本中每字符与其相邻位置的字符之间的关联关系,根据所确定的关联关系对所述文本进行重组,得到重组结果集,包括:

将所述文本中所有字符依次与上一个字符的属性信息进行对比;

当对比结果为存在交集,则将当前字符与上一字符划分为同一词语或命名实体;

当对比结果为不存在交集,则确定上一词语或命名实体划分完毕,确定当前字符属于新的词语或命名实体;

按照上述方式遍历所述文本的所有字符后,得到所述文本的重组结果集。

6.根据权利要求1或5所述的方法,其特征在于,所述方法还包括:

接收到所述文本后,对所述文本,进行分词操作,得到一个或多个分词结果集,其中,每个分词结果集中包括构成所述文本的所有词语的属性信息。

7.根据权利要求6所述的方法,其特征在于,所述对所述文本,进行分词操作,包括:

利用分词工具对所述文本进行分词操作;

其中，利用不同的分词工具得到多个分词结果集。

8.根据权利要求6所述的方法，其特征在于，所述方法还包括：

接收到所述文本后，对所述文本，进行命名实体识别操作，得到一个或多个命名实体结果集；

其中，每个命名实体结果集中包括构成所述文本的所有命名实体的属性信息。

9.根据权利要求8所述的方法，其特征在于，对所述文本，进行命名实体识别操作，包括：

利用命名实体识别工具对所述文本进行命名实体识别操作；

其中，利用不同的命名实体识别工具得到多个分词结果集。

10.一种基于分词和命名实体识别的文本重组装置，其特征在于，包括：

字符分词和实体特征设置模块，用于接收到待处理的文本后，根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息，以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定所述文本中每个字符的属性信息；

文本字符的属性合并模块，用于根据所述文本中每个字符的属性信息，确定所述文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对所述文本进行重组，得到重组结果集，所述重组结果集中包括构成所述文本的所有词语和/或命名实体的属性信息，其中，重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

11.根据权利要求10所述的装置，其特征在于，所述分词结果集中每个词语的属性信息，所述命名实体结果集中每个命名实体的属性信息，以及所述文本中每个字符的属性信息，至少包括如下任一种或几种：

内容，开始位置，结束位置，类型，标识，权重。

12.根据权利要求11所述的装置，其特征在于，

所述字符的属性信息中的权重，小于所述词语和命名实体的属性信息中的权重；

所述词语的属性信息中的权重小于或等于所述命名实体的属性信息中的权重。

13.根据权利要求11或12所述的装置，其特征在于，所述字符分词和实体特征设置模块，根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息，以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定所述文本中每个字符的属性信息，包括：

基于所述命名实体结果集中每个命名实体的属性信息，以及所述文本中每个字符与所述命名实体结果集中每个命名实体在所述文本中的位置关系，设置所述文本中每个字符的属性信息；

基于所述分词结果集中每个词语的属性信息，以及所述文本中每个字符与所述分词结果集中每个词语在所述文本中的位置关系，设置所述文本中每个字符的属性信息；

其中，基于所述命名实体结果集和所述分词结果集，为同一字符设置出多条不相同的属性信息时，将多条不相同的属性信息中权重取值最大属性信息确定为该字符的属性信息。

14.根据权利要求10所述的装置，其特征在于，所述文本字符的属性合并模块，根据所述文本中每个字符的属性信息，确定所述输入文本中每字符与其相邻位置的字符之间的关

联关系，根据所确定的关联关系对所述文本进行重组，得到重组结果集，包括：

将所述文本中所有字符依次与上一个字符的属性信息进行对比；

当对比结果为存在交集，则将当前字符与上一字符划分为同一词语或命名实体；

当对比结果为不存在交集，则确定上一词语或命名实体划分完毕，确定当前字符属于新的词语或命名实体；

按照上述方式遍历所述文本的所有字符后，得到所述文本的重组结果集。

15.根据权利要求10或14所述的装置，其特征在于，所述装置还包括：

分词模块，用于接收到所述文本后，对所述文本，进行分词操作，得到一个或多个分词结果集，其中，每个分词结果集中包括构成所述文本的所有词语的属性信息。

16.根据权利要求15所述的装置，其特征在于，

所述分词模块，利用分词工具对所述文本进行分词操作；

其中，利用不同的分词工具得到多个分词结果集。

17.根据权利要求15所述的装置，其特征在于，所述装置还包括：

命名实体识别模块，用于接收到所述文本后，对所述文本，进行命名实体识别操作，得到一个或多个命名实体结果集，其中，每个命名实体结果集中包括构成所述文本的所有命名实体的属性信息。

18.根据权利要求17所述的装置，其特征在于，

所述命名实体识别模块，利用命名实体识别工具对所述文本进行命名实体识别操作；

其中，利用不同的命名实体识别工具得到多个命名实体结果集。

19.一种基于分词和命名实体识别的文本重组装置，其特征在于，包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器被配置为：

接收到待处理的文本后，根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息，以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定所述文本中每个字符的属性信息；

根据所述文本中每个字符的属性信息，确定所述文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对所述文本进行重组，得到重组结果集，所述重组结果集中包括构成所述文本的所有词语和/或命名实体的属性信息，其中，重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

20.一种非临时性计算机可读存储介质，当所述存储介质中的指令由移动终端的处理器执行时，使得移动终端能够执行一种基于分词和命名实体识别的文本重组方法，所述方法包括：

接收到待处理的文本后，根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息，以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定所述文本中每个字符的属性信息；

根据所述文本中每个字符的属性信息，确定所述文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对所述文本进行重组，得到重组结果集，所述重组结果集中包括构成所述文本的所有词语和/或命名实体的属性信息，其中，重组得到的每个词

语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

基于分词和命名实体识别的文本重组方法及装置、介质

技术领域

[0001] 本公开涉及自然语言处理领域，尤其是涉及一种基于分词和命名实体识别的文本重组方法及装置、介质。

背景技术

[0002] 命名实体识别(Named Entity Recognition,NER)可以从自然语言中抽取出人名、地名、机构名等实体，从而有助于用户的理解。目前，实现NER的方法大体分为如下三类：

[0003] 一、基于规则的方法。该方法主要利用手工编写的规则，将文本与规则进行匹配来识别出命名实体。

[0004] 二、基于特征模板的方法。该方法主要通过统计机器学习方法将NER视作序列标注任务，利用大规模语料来学习出标注模型，从而对句子的各个位置进行标注，识别出命名实体。

[0005] 三、基于神经网络的方法。该方法主要利用神经网络做为有效处理NER任务的模型，将token从离散one-hot表示映射到低维空间中成为稠密的embedding，随后将句子的embedding序列输入到循环神经网络(Recurrent Neural Networks,RNN)中，用神经网络自动提取特征，Softmax来预测每个token的标签，从而识别出命名实体。

[0006] 基于上述三类方法，相关技术中存在很多种命名实体识别的工具，这些工具为人类提供了多种获取实体的方式。但是，每种工具的标注集合不同、标注结果不一致，即其识别出的命名实体略有差别。

发明内容

[0007] 为克服相关技术中存在的问题，本公开提供一种基于分词和命名实体识别的文本重组方法及装置、介质。

[0008] 根据本公开实施例的第一方面，提供一种基于分词和命名实体识别的文本重组方法，包括：

[0009] 接收到待处理的文本后，根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息，以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定所述文本中每个字符的属性信息；

[0010] 根据所述文本中每个字符的属性信息，确定所述文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对所述文本进行重组，得到重组结果集，所述重组结果集中包括构成所述文本的所有词语和/或命名实体的属性信息；

[0011] 其中，重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

[0012] 可选地，上述方法中，所述分词结果集中每个词语的属性信息，所述命名实体结果集中每个命名实体的属性信息，以及所述文本中每个字符的属性信息，至少包括如下任一种或几种：内容，开始位置，结束位置，类型，标识，权重。

[0013] 可选地，上述方法中，所述字符的属性信息中的权重，小于所述词语和命名实体的属性信息中的权重；

[0014] 所述词语的属性信息中的权重小于或等于所述命名实体的属性信息中的权重。

[0015] 可选地，上述方法中，所述根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息，以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定所述文本中每个字符的属性信息，包括：

[0016] 基于所述命名实体结果集中每个命名实体的属性信息，以及所述文本中每个字符与所述命名实体结果集中每个命名实体在所述文本中的位置关系，设置所述文本中每个字符的属性信息；

[0017] 基于所述分词结果集中每个词语的属性信息，以及所述文本中每个字符与所述分词结果集中每个词语在所述文本中的位置关系，设置所述文本中每个字符的属性信息；

[0018] 其中，基于所述命名实体结果集和所述分词结果集，为同一字符设置出多条不相同的属性信息时，将多条不相同的属性信息中权重取值最大属性信息确定为该字符的属性信息。

[0019] 可选地，上述方法中，所述根据所述文本中每个字符的属性信息，确定所述文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对所述文本进行重组，得到重组结果集，包括：

[0020] 将所述文本中所有字符依次与上一个字符的属性信息进行对比；

[0021] 当对比结果为存在交集，则将当前字符与上一字符划分为同一词语或命名实体；

[0022] 当对比结果为不存在交集，则确定上一词语或命名实体划分完毕，确定当前字符属于新的词语或命名实体；

[0023] 按照上述方式遍历所述文本的所有字符后，得到所述文本的重组结果集。

[0024] 可选地，上述方法还包括：

[0025] 接收到所述文本后，对所述文本，进行分词操作，得到一个或多个分词结果集，其中，每个分词结果集中包括构成所述文本的所有词语的属性信息。

[0026] 可选地，上述方法中，所述对所述文本，进行分词操作，包括：

[0027] 利用分词工具对所述文本进行分词操作；

[0028] 其中，利用不同的分词工具得到多个分词结果集。

[0029] 可选地，上述方法还包括：

[0030] 接收到所述文本后，对所述文本，进行命名实体识别操作，得到一个或多个命名实体结果集；

[0031] 其中，每个命名实体结果集中包括构成所述文本的所有命名实体的属性信息。

[0032] 可选地，上述方法中，对所述文本，进行命名实体识别操作，包括：

[0033] 利用命名实体识别工具对所述文本进行命名实体识别操作；

[0034] 其中，利用不同的命名实体识别工具得到多个分词结果集。

[0035] 根据本公开实施例的第二方面，提供一种基于分词和命名实体识别的文本重组装置，包括：

[0036] 字符分词和实体特征设置模块，用于接收到待处理的文本后，根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息，以及所述文本经过命名实体识别操作

得到的命名实体结果集中每个命名实体的属性信息，确定所述文本中每个字符的属性信息；

[0037] 文本字符的属性合并模块，用于根据所述文本中每个字符的属性信息，确定所述文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对所述文本进行重组，得到重组结果集，所述重组结果集中包括构成所述文本的所有词语和/或命名实体的属性信息，其中，重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

[0038] 可选地，上述装置中，所述分词结果集中每个词语的属性信息，所述命名实体结果集中每个命名实体的属性信息，以及所述文本中每个字符的属性信息，至少包括如下任一种或几种：

[0039] 内容，开始位置，结束位置，类型，标识，权重。

[0040] 可选地，上述装置中，所述字符的属性信息中的权重，小于所述词语和命名实体的属性信息中的权重；

[0041] 所述词语的属性信息中的权重小于或等于所述命名实体的属性信息中的权重。

[0042] 可选地，上述装置中，所述字符分词和实体特征设置模块，根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息，以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定所述文本中每个字符的属性信息，包括：

[0043] 基于所述命名实体结果集中每个命名实体的属性信息，以及所述文本中每个字符与所述命名实体结果集中每个命名实体在所述文本中的位置关系，设置所述文本中每个字符的属性信息；

[0044] 基于所述分词结果集中每个词语的属性信息，以及所述文本中每个字符与所述分词结果集中每个词语在所述文本中的位置关系，设置所述文本中每个字符的属性信息；

[0045] 其中，基于所述命名实体结果集和所述分词结果集，为同一字符设置出多条不相同的属性信息时，将多条不相同的属性信息中权重取值最大属性信息确定为该字符的属性信息。

[0046] 可选地，上述装置中，所述文本字符的属性合并模块，根据所述文本中每个字符的属性信息，确定所述输入文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对所述文本进行重组，得到重组结果集，包括：

[0047] 将所述文本中所有字符依次与上一个字符的属性信息进行对比；

[0048] 当对比结果为存在交集，则将当前字符与上一字符划分为同一词语或命名实体；

[0049] 当对比结果为不存在交集，则确定上一词语或命名实体划分完毕，确定当前字符属于新的词语或命名实体；

[0050] 按照上述方式遍历所述文本的所有字符后，得到所述文本的重组结果集。

[0051] 可选地，上述装置还包括：

[0052] 分词模块，用于接收到所述文本后，对所述文本，进行分词操作，得到一个或多个分词结果集，其中，每个分词结果集中包括构成所述文本的所有词语的属性信息。

[0053] 可选地，上述装置中，所述分词模块，利用分词工具对所述文本进行分词操作；

[0054] 其中，利用不同的分词工具得到多个分词结果集。

[0055] 可选地，上述装置还包括：

[0056] 命名实体识别模块，用于接收到所述文本后，对所述文本，进行命名实体识别操作，得到一个或多个命名实体结果集，其中，每个命名实体结果集中包括构成所述文本的所有命名实体的属性信息。

[0057] 可选地，上述装置中，所述命名实体识别模块，利用命名实体识别工具对所述文本进行命名实体识别操作；

[0058] 其中，利用不同的命名实体识别工具得到多个命名实体结果集。

[0059] 根据本公开实施例的第三方面，提供一种基于分词和命名实体识别的文本重组装置，包括：

[0060] 处理器；

[0061] 用于存储处理器可执行指令的存储器；

[0062] 其中，所述处理器被配置为：

[0063] 接收到待处理的文本后，根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息，以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定所述文本中每个字符的属性信息；

[0064] 根据所述文本中每个字符的属性信息，确定所述文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对所述文本进行重组，得到重组结果集，所述重组结果集中包括构成所述文本的所有词语和/或命名实体的属性信息，其中，重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

[0065] 根据本公开实施例的第四方面，提供一种非临时性计算机可读存储介质，当所述存储介质中的指令由移动终端的处理器执行时，使得移动终端能够执行一种基于分词和命名实体识别的文本重组方法，所述方法包括：

[0066] 接收到待处理的文本后，根据所述文本经过分词操作得到的分词结果集中每个词语的属性信息，以及所述文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定所述文本中每个字符的属性信息；

[0067] 根据所述文本中每个字符的属性信息，确定所述文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对所述文本进行重组，得到重组结果集，所述重组结果集中包括构成所述文本的所有词语和/或命名实体的属性信息，其中，重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

[0068] 本公开的实施例提供的技术方案可以包括以下有益效果：

[0069] 本公开的技术方案基于分词和命名实体识别技术，对待处理的文本进行重新组合与渲染，实现多样性标注的输出，再通过建立字符的属性信息中的权重，来实现便于用户习惯的阅读渲染方式。

[0070] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本公开。

附图说明

[0071] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本发明的实施例，并与说明书一起用于解释本发明的原理。

[0072] 图1是根据一示例性实施例示出的一种基于分词和命名实体识别的文本重组方法的一种实施方式的流程图。

[0073] 图2是根据一示例性实施例示出的另一种基于分词和命名实体识别的文本重组方法的另一种实施方式的流程图。

[0074] 图3是根据一示例性实施例示出的确定待处理的文本中每个字符的属性信息的方法流程图。

[0075] 图4是根据一示例性实施例示出的根据所确定的关联关系对文本进行重组，得到重组结果集的一种实施方法的流程图。

[0076] 图5是根据一示例性实施例示出的根据重组结果集对文本进行渲染展现的一种实施方式的原理示意图。

[0077] 图6是根据一示例性实施例示出的一种基于分词和命名实体识别的文本重组装置的结构框图。

具体实施方式

[0078] 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

[0079] 图1是根据一示例性实施例示出的一种基于分词和命名实体识别的文本重组方法的一种实施方式的流程图。如图1所示，该方法包括以下步骤：

[0080] 步骤S11，接收到待处理的文本后，根据文本经过分词操作得到的分词结果集中每个词语的属性信息，以及文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定文本中每个字符的属性信息。

[0081] 本实施例中，可以是多维形式来表示分词结果集中每个词语的属性信息，命名实体结果集中每个命名实体的属性信息，以及文本中每个字符的属性信息，即属性信息至少包括如下任一种或几种：内容，开始位置，结束位置，类型，标识，权重。

[0082] 其中，内容指，经过分词操作得到的任意一个词语的具体内容，或者经过命名实体识别得到的任意一个命名实体的具体内容，或者任意一个字符的具体内容。

[0083] 开始位置指，上述内容在整个待处理的文本中起始的位置标记。

[0084] 结束位置指，上述内容在整个待处理的文本中结束的位置标记。

[0085] 本文对整个文本的位置标记方式不做限制，可以包括多种方式。例如，可以设置整个文本的初始的位置标记从0开始，也可以从任意整数开始。可以设置整个文本中每两个相邻的字符之间的位置标记的差值均为1，也可以设置整个文本中每两个相邻的字符之间的位置标记的差值均为5。只需要整个文本中两两相邻的字符之间的位置标记的差值为相同值即可。由此可见，本实施例根据开始位置对应的位置标记和结束位置的位置标记确定上述内容在整个文本中所在的位置。

[0086] 类型指,文本类型,可以是字符,词语或者命名实体的一种。

[0087] 标识指,文本在整个系统中的唯一标识,即通过此标识可以唯一确定一个字符或词语或命名实体等,可以采用UUID(Universally Unique Identifier,通用唯一识别码)的形式来设置此标识。

[0088] 权重指,上述内容的优先级别,一般权重与类型有关,不同类型的文本的优先级别不同。例如,按照一般习惯,字符类型的文本的权重最小,词语类型的文本的权重次小,命名实体类型的文本的权重最大。但也可以根据实际需求,修改不同类型的文本之间的权重关系。例如,设置字符类型的文本的权重最小,设置命名实体类型的文本的权重为次小,设置词语类型的文本的权重最大。另外,同一类型的文本,依据其使用的工具的权重不同,也可以设置不同的权重。例如,采用多种工具划分出了命名实体类型的文本,其中,第一工具的权重最大,第二工具的权重次大,第三工具的权重最小,则基于第一工具得到的命名实体的权重最大,基于第二工具得到命名实体的权重次大,基于第三工具得到命名实体的权重最小。一般工具的权重可以默认设置,也可以由用户设置。用户设置的工具的权重可以反映出用户对不同的工具的依赖程度有所不同。这样,在设置识别结果的权重时,可以更偏重于用户依赖的工具所识别出的结果,从而使得识别出的结果更贴近于用户的阅读习惯。

[0089] 上述步骤S11中,可以利用现有的分词工具进行分词操作。还可以利用现有的命名实体识别工具进行命名实体识别操作。

[0090] 步骤S12,根据文本中每个字符的属性信息,确定文本中每字符与其相邻位置的字符之间的关联关系,根据所确定的关联关系对文本进行重组,得到重组结果集,该重组结果集中包括构成文本的所有词语和/或命名实体的属性信息。

[0091] 其中,重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

[0092] 上述步骤S12中,获得重组结果集的过程可以参照如下操作:

[0093] 将文本中所有字符依次与上一个字符的属性信息进行对比;

[0094] 当对比结果为存在交集,则将当前字符与上一字符划分为同一词语或命名实体;

[0095] 当对比结果为不存在交集,则确定上一词语或命名实体划分完毕,确定当前字符属于新的词语或命名实体;

[0096] 按照上述方式遍历所述文本的所有字符后,得到待处理的文本的重组结果集。

[0097] 在上述步骤S12之后,还可以包括如下步骤:根据重组结果集对文本进行渲染展现。

[0098] 由上述实施例可以看出,本公开的技术方案利用了分词工具,以及命名实体识别工具等,基于这些工具之间所采用的标注方法不同,可以获得多种识别结果,将多种识别结果进行综合考量后再进行重组,可以提高重组的准确性和可靠性。

[0099] 图2是根据一示例性实施例示出的基于分词和命名实体识别的文本重组方法的另一种实施方式的流程图。在此实施例中,假设待处理的文本为一句话,“未来科技公司的CEO是不是张三”,并同时利用一种分词工具和两种不同的命名实体识别工具,进行文本重组,如图2所示,该文本重组过程包括以下步骤:

[0100] 在步骤S21中,接收到待处理的文本后,对待处理的文本进行分词操作,得到分词结果集,其中,每个分词结果集中包括构成待处理的文本的所有词语的属性信息。

[0101] 上述步骤S21中，得到的分词结果集可以是一个或多个，每个分词结果集中包括构成待处理的文本的所有词语的属性信息即可。

[0102] 其中，分词结果集中，可以以多维的形式保存每个词语的属性信息，例如每个词语包括如下所示各项属性信息：

[0103] {内容、开始位置、结束位置、类型＝“seg”、UUID、priority＝0}。

[0104] 本实施例以一个分词结果集为例进行说明，上述步骤S21可细分为如下两步骤：

[0105] 步骤S21a，可以使用本领域中公开使用的分词工具进行分词操作，以实现将待处理的文本的句子切分为多个词语。

[0106] 步骤S21b，设置分词操作得到的每个词语的属性信息。

[0107] 本实施例中，开始位置、结束位置为分词操作得到的词语在原始字符串(即待处理的文本)中第一个字符和最后一个字符的下标。

[0108] UUID为该词语在系统内的唯一的标识，此UUID需要保证不重复。

[0109] priority为词语的权重，一般需要在整个系统期间设置唯一。

[0110] 在本示例中，对于待处理的文本进行分词操作以后得到的分词结果集如表1所示，

[0111] 表1为分词结果集中所有词语的属性信息列表。

[0112]

内容	开始位置	结束位置	类型	UUID	Priority
未来	0	1	SEG	uuid0	0
科技	2	3	SEG	uuid1	0
公司	4	5	SEG	uuid2	0
的	6	6	SEG	uuid3	0
CEO	7	9	SEG	uuid4	0
是	10	10	SEG	uuid5	0
不是	11	12	SEG	uuid6	0
张三	13	14	SEG	uuid7	0

[0113] 从表1可以看出，本实施例中，整个待处理的文本的初始的位置标记从0开始标记，即整个待处理的文本中第一个字符的下标从0开始标记。整个文本中每两个相邻的字符之间的位置标记的差值均为1。即整个待处理的文本中每个字符的下标是其前一个字符的下标序号加1。

[0114] 例如，待处理的文本的分词结果集中，“科技”一词的开始位置为2，即在整个待处理的文本中“科技”一词起始的位置标记为2。结束位置为3，，即在整个待处理的文本中“科技”一词结束的位置标记为3。“科技”一词的UUID为“uuid1”。

[0115] 在步骤S22中，接收到待处理的文本后，对待处理的文本进行命名实体识别操作，得到命名实体结果集，其中，每个命名实体结果集中包括构成待处理的文本的所有命名实体的属性信息。

[0116] 上述步骤S22中，得到的命名实体结果集可以是一个或多个，每个命名实体结果集中包括构成待处理的文本的所有命名实体的属性信息即可。

[0117] 其中，命名实体结果集中，可以以多维的形式保存每个命名实体的属性信息，例如每个命名实体包括如下所示各项属性信息：

[0118] {内容、开始位置、结束位置、类型、UUID、priority$_y$＝#pri}；

[0119] 实施例以两个命名实体结果集为例进行说明，上述步骤S22可细分为如下两步骤：

[0120] 步骤S22a，可以使用本领域公开的命名实体识别工具，实现对待处理的文本中提取对应的命名实体的属性信息，包括但不限于命名实体的值、属性；

[0121] 步骤S22b，设置每个命名实体的识别结果的属性信息。其中，开始位置、结束位置为识别结果在原始字符串中第一个字符和最后一个字符的下标；UUID为唯一标示编号，需要保证不重复；#pri为针对识别出的命名实体的权重值。

[0122] 一般设置命名实体的#pri的值大于分词结果集中词语的权重的值，即表示本实施例中命名实体识别的结果比分词的结果更重要。但在其他应用场景中，也可以根据用户的需要，将分词结果集中词语的权重的值，设置大于命名实体结果集中命名实体的权重的值。另外，此权重值也可以按照不同的命名实体识别工具来设置，可以根据用户的偏好进行设置。

[0123] 例如，对于本实施例的待处理的文本，按照第一种命名实体识别工具进行识别以后得到的结果如表2所示。

[0124] 表2为第一种命名实体识别工具识别出的结果集中所有的命名实体的属性信息列表。

[0125]

内容	开始位置	结束位置	类型	UUID	Priority
未来科技公司	0	5	COMPANY	uuid10	1
的	6	6	\	uuid11	
CEO	7	9	TITLE	uuid12	1
是	10	10	\	uuid13	
不是	11	12	\	uuid14	
张三	13	14	PERSON	uuid15	1

[0126] 从表2可以看出，待处理的文本的第一种命名实体结果集中，“未来科技公司”是一个命名实体，其开始位置为0，即在整个待处理的文本中命名实体“未来科技公司”起始的位置标记为0。结束位置为5，即在整个待处理的文本中命名实体“未来科技公司”结束的位置标记为5。即命名实体“未来科技公司”的UUID为“uuid10”。

[0127] 再例如，对于本实施例的待处理的文本，按照第二种命名实体识别工具进行识别以后得到的结果如表3所示。

[0128] 表3为第二种命名实体识别工具识别出的结果集中所有的命名实体的属性信息列表。

[0129]

内容	开始位置	结束位置	类型	UUID	Priority
未来科技公司	0	5	COMPANY	uuid20	1
的	6	6	\	uuid21	
CEO	7	9	\	uuid22	
是	10	10	\	uuid23	
不是	11	12	\	uuid24	
张三	13	14	PERSON	uuid25	1

[0130] 从表3可以看出，使用第二种命名实体识别工具进行命名实体识别时，“CEO”这个词并没有被识别为“TITLE”。从这里可以看出，不同的命名实体识别工具识别的结果可能不

相同,因此,为了提高重组结果的可靠和准确性,可以利用多种不同的命名实体识别工具识别命名实体。

[0131] 另外,从上述步骤S21和S22的描述可以看出,本实施例对分词和命名实体识别的开始位置计数均从0开始,但是在其他应用场景中,也可以包括使用其它计数开始位置(例如,位置计数从1开始)或者以其它步长(例如每两字符之间的间距可以设置为2)的方式,只要能够保证位置的唯一性以及字符的可定位性,就可以认为与本实例中使用的方式是等同的。

[0132] 在其他应用场景中,上述步骤S21和步骤S22的顺序也可以调换,只要得到分词结果集和命名实体结果集即可。

[0133] 在步骤S23中,根据待处理的文本经过分词操作得到的分词结果集中每个词语的属性信息,以及待处理的文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息,确定待处理的文本中每个字符的属性信息。

[0134] 为了便于理解本实施例的技术方案,可以将上述步骤S23的操作分为几个步骤进行说明,图3所示即为本示例性实施例中确定待处理的文本中每个字符的属性信息的方法流程图,如图3所示,该方法包括以下步骤:

[0135] 在步骤S23a中,对待处理的文本的每个字符的属性信息进行初始化;

[0136] 本实施例中,每个字符的初始化的属性信息可以包括如下几项:

[0137] {内容、开始位置、结束位置、类型、UUID、priority};

[0138] 在本实施例中,字符的属性信息中的权重(priority),小于词语和命名实体的属性信息中的权重;词语的属性信息中的权重小于或等于命名实体的属性信息中的权重。

[0139] 可见,本实施例中,字符的权重最小,词语的权重次小,命名实体的权重最大。

[0140] 以本示例性实施例的待处理的文本为例,对于待处理的文本中第一字符“未”的初始化的属性信息为{“未”,0,0,[],[],“-1”},即对于字符“未”,其起始位置为0,结束位置为0,类型为空,UUID为空,priority为-1。

[0141] 在步骤S23b中,基于命名实体结果集中每个命名实体的属性信息,以及待处理的文本中每个字符与命名实体结果集中每个命名实体在待处理的文本中的位置关系,设置待处理的文本中每个字符的属性信息;

[0142] 上述步骤S23b的实现方式可以有多种,在本实施例中,以如下方式为例进行说明其具体操作:

[0143] 将待处理的文本的所有字符在NER结果集中依次进行遍历,对每个字符的属性信息进行设定。其中,针对NER结果集中任意一个命名实体N,如果char(字符)的下标关系满足要求(char.开始位置>=N.开始位置&&char.结束位置<=N.结束位置),则将命名实体N的UUID与字符的UUID进行附加、并将命名实体的类型与字符的类型进行附加、并将命名实体N的.priority赋值给char.priority,即设置char.priority为1。

[0144] 以本实施例的待处理的文本为例,按照上述操作处理后,对于待处理的文本的第一个字符“未”,经过第一种命名实体识别工具的NER的结果集,设置的属性信息为:

[0145] {“未”,0,0,[“COMANY”],[“UUID10”],1}。

[0146] 由于本实施例中使用了两种不同的命名实体识别工具,因此,本实施例中存在有两个NER的结果集,每个字符需要分别遍历这两个NER结果集,此时,按照步骤S23b的操作,

对于待处理的文本的第一个字符“未”,经过第二种命名实体识别工具的NER的结果集,设置的属性信息为:

[0147] {“未”,0,0,[“COMANY”],[“UUID20”],1}。

[0148] 在步骤S23c中,基于分词结果集中每个词语的属性信息,以及待处理的文本中每个字符与分词结果集中每个词语在待处理的文本中的位置关系,设置待处理的文本中每个字符的属性信息,其中,基于命名实体结果集和分词结果集,为同一字符设置出多条不相同的属性信息时,将多条不相同的属性信息中权重取值最大属性信息确定为该字符的属性信息。

[0149] 上述步骤S23c的实施方式可以有多种,在本实施例中,以如下方式为例进行说明其具体操作:

[0150] 将经过步骤S23b操作的所有字符在分词结果集中进行遍历,对于任意的分词结果集中的词语S,如果下标满足关系(char.开始位置>=S.开始位置&&char.结束位置<=S.结束位置),并且如果满足S.priority>char.priority,则将词语S的UUID与字符的UUID进行附加、并将词语S的类型与字符的类型进行附加、并将词语S的priority赋值给char priority,即设置char priority为0,最后将char的属性信息设置如下:

[0151] {内容、开始位置、结束位置、S类型、S的UUID、priority=0}。

[0152] 仍然以待处理的文本中第一个字符“未”为例说明,按照上述步骤S23b的操作,此时字符“未”的属性信息有两条,分别为{“未”,0,0,[“COMANY”],[“UUID10”],1},{“未”,0,0,[“COMANY”],[“UUID20”],1}。在本实施例中,由于分词结果集中词语的priority为0,而该字符的priority的当前值为1,即词语的权重小于已设置的字符的权重,因此不会将词语权重赋值给字符。即按照步骤S23c的操作后,字符“未”的属性信息仍然有两条,分别为{“未”,0,0,[“COMANY”],[“UUID10”],1},{“未”,0,0,[“COMANY”],[“UUID20”],1}。

[0153] 再例如,对于待处理的文本中的字符“不”,其不属于任何命名实体结果,因此,按照上述步骤S23b的操作后,字符“不”的属性信息为{“不”,11,11,[],[],0}。此时,再按照步骤S23c的操作时,由于字符“不”属于分词结果集中的词语,而词语的priority为0,大于该字符的priority当前值-1,因此要将词语的权重赋值给字符。即按照步骤S23c的操作后,字符“不”的属性信息为{“不”,11,11,[“SEG”],[“UUID6”],0}。

[0154] 由上文可以看出,本实施例中采用了两种命名实体识别工具,因此,经过步骤S23c的操作后,一些字符的属性信息的记录可能有两条,此时,可以根据这两种命名实体识别工具的权重,最终确定为字符保留一条属性信息。例如,如上文所述的字符“未”的属性信息有两条,分别为{“未”,0,0,[“COMANY”],[“UUID10”],1},{“未”,0,0,[“COMANY”],[“UUID20”],1},其中第一条属性信息是利用第一种命名实体识别工具得到的,第二条属性信息是利用第二种命名实体识别工具得到的。假设第一种命名实体识别工具的权重大于第二种命名实体识别工具的权重,则最终确定字符“未”的属性信息为{“未”,0,0,[“COMANY”],[“UUID10”],1}。

[0155] 在步骤S24中,根据待处理的文本中每个字符的属性信息,确定待处理的文本中每字符与其相邻位置的字符之间的关联关系,根据所确定的关联关系对待处理的文本进行重组,得到重组结果集。

[0156] 上述步骤S24的实现方式可以有多种,本实施例以其中一种方式为例里进行说明。

图4所示即为本示例性实施例中根据所确定的关联关系对待处理的文本进行重组，得到重组结果集的一种实施方法的流程图，如图4所示，该方法包括以下步骤：

[0157] 在步骤S24a中，定义一个结构体temp，此结构体的初始信息为空，即结构体temp的初始信息如下：

[0158] {[],0,0,[],["UUID0"],0}。

[0159] 在步骤S24b中，将待处理的文本的所有字符按照起始位置从小到大排序，按照排序的顺序将待处理的文本的字符依次与temp进行对比，判断字符的UUID与temp的UUID的交集是否为空，如果字符的UUID与temp的UUID的交集为空，执行步骤S24c，如果字符的UUID与temp的UUID的交集不为空，执行步骤S24d。

[0160] 在步骤S24c中，如果字符的UUID与temp的UUID的交集为空，说明当前字符与存储于temp中的上一字符不属于同一个词语或不属于同一个命名实体，也就是确定了上一词语或命名实体划分完毕，而当前字符属于新的词语或命名实体了，此时，将已存储在temp中的内容移动到重组结果集列表list中，将temp置空后将当前字符添加进temp，进入步骤S24e。

[0161] 在步骤S24d中，如果字符的UUID与temp的UUID的交集不为空，说明当前字符与存储于temp中的上一字符属于同一个词语或属于同一个命名实体，此时将该字符内容添加进temp，并将字符的属性信息与temp的属性信息进行合并，进入步骤S24e。

[0162] 在步骤S24e中，检查当前字符是否为待处理的文本的最后一个字符，如果当前字符是最后一个字符，说明全部字符已遍历完成，进入步骤S24f，如果当前字符不是最后一个字符，说明全部字符未遍历完，返回步骤S24b，检查下一个字符的UUID与temp的UUID的交集是否为空直到全部字符遍历完成。

[0163] 在步骤S24f中，得到待处理的文本的重组结果集。

[0164] 在步骤S25中，根据重组结果集对待处理的文本进行渲染展现。

[0165] 图5所示即为本示例性实施例中根据重组结果集对待处理的文本进行渲染展现的一种实施方式的原理示意图。该方法的实施原理如下：

[0166] 通过html中的css渲染，根据类型设置各个字符对应的配色，其中，权重越大透明度越低，不同命名实体工具得到的命名实体结果可设置为不同的配色。

[0167] 通过对步骤S24得到的重组结果列表，按照起始位置从小到大将重组结果输出，输出时按照每字符对应的配色进行输出。

[0168] 其中，可能存在位置重叠，因此渲染时可以使用遮罩的展现方式，从而可以将一个字符属于多个类型的情况展现出来。从图5可以看出，本公开的技术方案提供了层次化交叠的可视化展现方式，便于进行多种识别结果的展现。

[0169] 图6所示为一示例性实施例示出的一种基于分词和命名实体识别的文本重组装置的结构框图。如图6所示，该装置至少包括字符分词和实体特征设置模块601和文本字符的属性合并模块602。

[0170] 字符分词和实体特征设置模块601，用于接收到待处理的文本后，根据待处理的文本经过分词操作得到的分词结果集中每个词语的属性信息，以及待处理的文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定待处理的文本中每个字符的属性信息；

[0171] 本实施例中，分词结果集中每个词语的属性信息，命名实体结果集中每个命名实

体的属性信息，以及所述待处理的文本中每个字符的属性信息，都可以采用多维信息来表示，即属性信息可以包括如下任一种或几种信息：

[0172] 内容，开始位置，结束位置，类型，标识，权重。

[0173] 本实施例中，内容指，经过分词操作得到的任意一个词语的具体内容，或者经过命名实体识别得到的任意一个命名实体的具体内容，或者任意一个字符的具体内容

[0174] 开始位置指，上述内容在整个待处理的文本中起始的位置标记。

[0175] 结束位置指，上述内容在整个待处理的文本中结束的位置标记。

[0176] 本实施例根据开始位置对应的位置标记和结束位置对应的位置标记可以确定上述内容在整个待处理的文本中所在的位置。

[0177] 类型指，文本类型，可以是字符，词语或者命名实体的一种。

[0178] 标识指，文本在整个系统中的唯一标识，即通过此标识可以唯一确定一个字符或词语或命名实体等，可以采用UUID的形式来设置此标识。

[0179] 权重指，上述内容的优先级别，一般权重与类型有关，不同类型的文本的优先级别不同。一般不同类型的文本的权重需要保证不重复。例如，命名实体类型的权重值和词语类型的权重值不同。一般可以设置命名实体的权重的值大于分词结果集中词语的权重的值，即表示本实施例中命名实体识别的结果比分词的结果更重要。但在其他应用场景中，也可以根据用户的需要，将分词结果集中词语的权重的值，设置大于命名实体结果集中命名实体的权重的值。另外，对于命名实体的权重的值，也可以根据用户需要，按照用户对不同的命名实体识别工具的偏好，将不同的命名实体识别工具识别出的命名实体的权重的值设置为不一样的数值。例如，将用户偏好的命名实体识别工具识别出的命名实体的权重的值设置为，大于其他命名实体识别工具识别出的命名实体的权重的值。

[0180] 实际应用中，字符分词和实体特征设置模块601，确定待处理的文本中每个字符的属性信息的方式可以采用各种方式，本实施例举例说明其中一种实施式如下：

[0181] 字符分词和实体特征设置模块基于命名实体结果集中每个命名实体的属性信息，以及待处理的文本中每个字符与命名实体结果集中每个命名实体在待处理的文本中的位置关系，设置待处理的文本中每个字符的属性信息；

[0182] 并且，基于分词结果集中每个词语的属性信息，以及待处理的文本中每个字符与分词结果集中每个词语在待处理的文本中的位置关系，设置待处理的文本中每个字符的属性信息；

[0183] 其中，按照上述方式对待处理的文本中每个字符设置属性信息时，如果某一字符的属性信息有多条，则将这多条属性信息中权重最大的属性信息确定为该字符的属性信息。

[0184] 文本字符的属性合并模块602，用于根据待处理的文本中每个字符的属性信息，确定待处理的文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对待处理的文本进行重组，得到重组结果集，重组结果集中包括构成待处理的文本的所有词语和/或命名实体的属性信息，其中，重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

[0185] 实际应用中，上述文本字符的属性合并模块602，得到重组结果集的方式可以有多种，本实施例以其中一种实施方式为例进行说明。此种实施方式包括如下操作：

[0186] 将所述待处理的文本中所有字符依次与上一个字符的属性信息进行对比；

[0187] 当对比结果为存在交集，则将当前字符与上一字符划分为同一词语或命名实体；

[0188] 当对比结果为不存在交集，则确定上一词语或命名实体划分完毕，确定当前字符属于新的词语或命名实体；

[0189] 按照上述方式遍历待处理的文本的所有字符后，得到待处理的文本的重组结果集。

[0190] 另外，以上述装置的结构为基本结构，还可以增加分词模块和命名实体识别模块。

[0191] 其中，增加的分词模块，用于对待处理的文本，进行分词操作，得到一个或多个分词结果集，其中，每个分词结果集中包括构成待处理的文本的所有词语的属性信息。实际应用中，分词模块可以利用现有分词工具来实现。对于有多个分词结果集的场景，一般是采用多个不同的分词工具得到的。

[0192] 增加的命名实体识别模块，用于对待处理的文本，进行命名实体识别操作，得到一个或多个命名实体结果集，其中，每个命名实体结果集中包括构成待处理的文本的所有命名实体的属性信息。与分词模块的实现方式类似，命名实体识别模块，也可以利用现有的命名实体识别工具进行命名实体识别操作。其中，当命名实体识别模块采用了多个不同的命名实体识别工具时，可以得到多个命名实体结果集。

[0193] 除了分词模块和命名实体识别模块，在一些场景中，还可以增加结果渲染输出模块，该模块根据重组结果集对待处理的文本进行渲染展现。

[0194] 关于上述实施例中的装置，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

[0195] 一示例性实施例提供一种基于分词和命名实体识别的文本重组装置，包括处理器，以及用于存储处理器可执行指令的存储器；

[0196] 其中，处理器被配置为：

[0197] 接收到待处理的文本后，根据待处理的文本经过分词操作得到的分词结果集中每个词语的属性信息，以及待处理的文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定待处理的文本中每个字符的属性信息；

[0198] 根据待处理的文本中每个字符的属性信息，确定待处理的文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对待处理的文本进行重组，得到重组结果集，重组结果集中包括构成待处理的文本的所有词语和/或命名实体的属性信息，其中，重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

[0199] 关于上述实施例中的装置，其中处理器被配置执行的操作的具体方式已经在有关基于分词和命名实体识别的文本重组方法的实施例中进行了详细描述，此处将不做详细阐述说明。

[0200] 一示例性实施例提供一种非临时性计算机可读存储介质，当所述存储介质中的指令由移动终端的处理器执行时，使得移动终端能够执行一种基于分词和命名实体识别的文本重组方法，所述方法包括：

[0201] 接收到待处理的文本后，根据待处理的文本经过分词操作得到的分词结果集中每个词语的属性信息，以及待处理的文本经过命名实体识别操作得到的命名实体结果集中每

个命名实体的属性信息，确定待处理的文本中每个字符的属性信息；

[0202] 根据待处理的文本中每个字符的属性信息，确定待处理的文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对待处理的文本进行重组，得到重组结果集，所述重组结果集中包括构成待处理的文本的所有词语和/或命名实体的属性信息，其中，重组得到的每个词语或命名实体的属性信息是由组成该词语或命名实体的所有字符的属性信息合并得到的。

[0203] 关于上述实施例中的非临时性计算机可读存储介质，其中移动终端执行的一种基于分词和命名实体识别的文本重组方法的具体方式已经在有关基于分词和命名实体识别的文本重组方法的实施例中进行了详细描述，此处将不做详细阐述说明。

[0204] 本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本发明的其它实施方案。本申请旨在涵盖本发明的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本发明的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本发明的真正范围和精神由下面的权利要求指出。

[0205] 应当理解的是，本发明并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本发明的范围仅由所附的权利要求来限制。

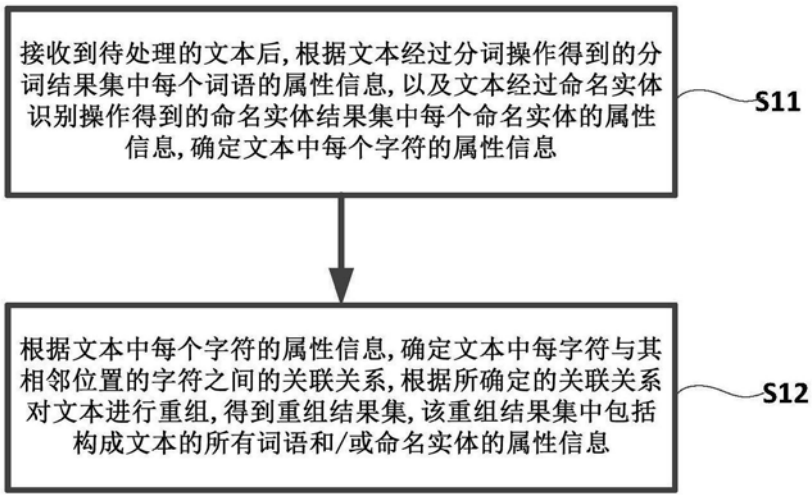

接收到待处理的文本后，根据文本经过分词操作得到的分词结果集中每个词语的属性信息，以及文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息，确定文本中每个字符的属性信息
S11
根据文本中每个字符的属性信息，确定文本中每字符与其相邻位置的字符之间的关联关系，根据所确定的关联关系对文本进行重组，得到重组结果集，该重组结果集中包括构成文本的所有词语和/或命名实体的属性信息
S12

图1

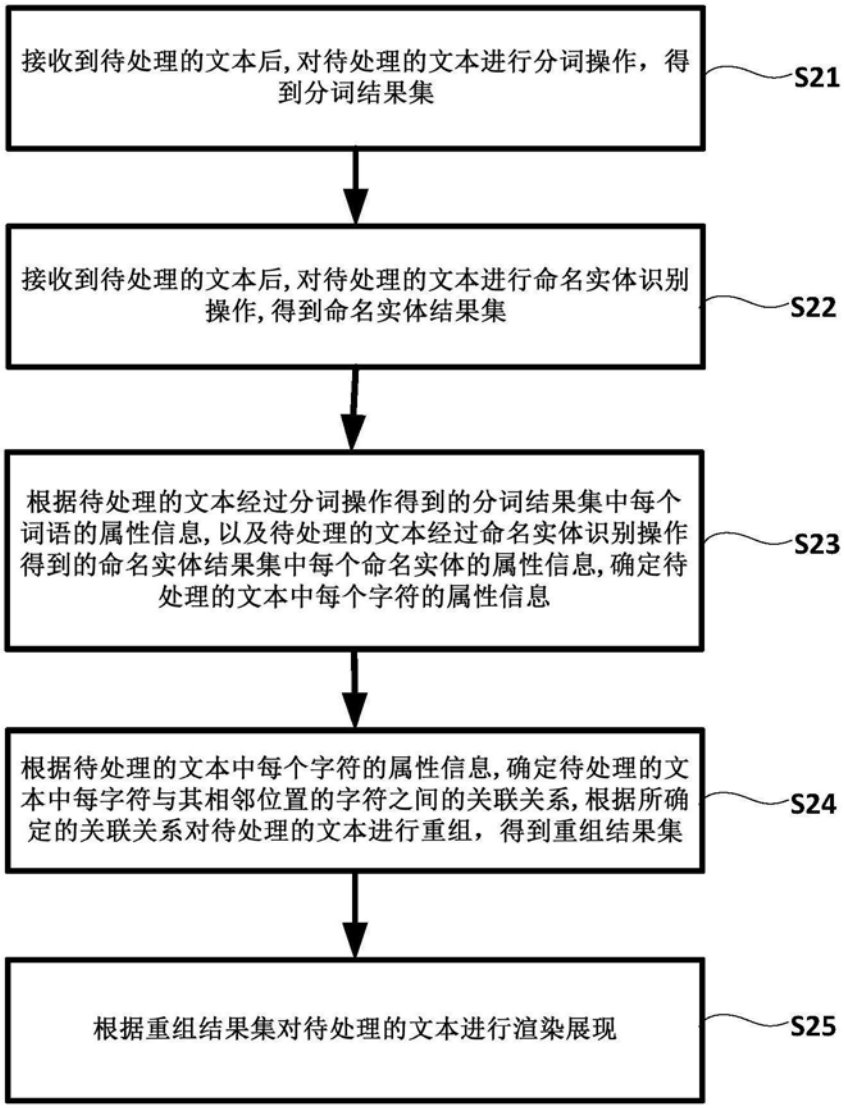

接收到待处理的文本后,对待处理的文本进行分词操作，得到分词结果集
S21
接收到待处理的文本后,对待处理的文本进行命名实体识别操作,得到命名实体结果集
S22
根据待处理的文本经过分词操作得到的分词结果集中每个词语的属性信息,以及待处理的文本经过命名实体识别操作得到的命名实体结果集中每个命名实体的属性信息,确定待处理的文本中每个字符的属性信息
S23
根据待处理的文本中每个字符的属性信息,确定待处理的文本中每字符与其相邻位置的字符之间的关联关系,根据所确定的关联关系对待处理的文本进行重组，得到重组结果集
S24
根据重组结果集对待处理的文本进行渲染展现
S25

图2

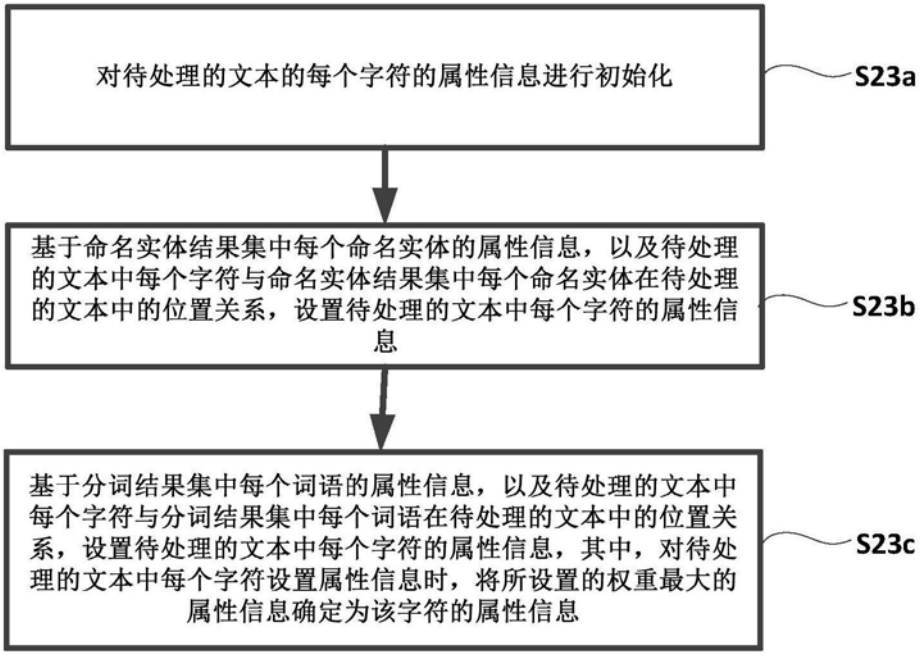

对待处理的文本的每个字符的属性信息进行初始化
S23a
基于命名实体结果集中每个命名实体的属性信息，以及待处理的文本中每个字符与命名实体结果集中每个命名实体在待处理的文本中的位置关系，设置待处理的文本中每个字符的属性信息
S23b
基于分词结果集中每个词语的属性信息，以及待处理的文本中每个字符与分词结果集中每个词语在待处理的文本中的位置关系，设置待处理的文本中每个字符的属性信息，其中，对待处理的文本中每个字符设置属性信息时，将所设置的权重最大的属性信息确定为该字符的属性信息
S23c

图3

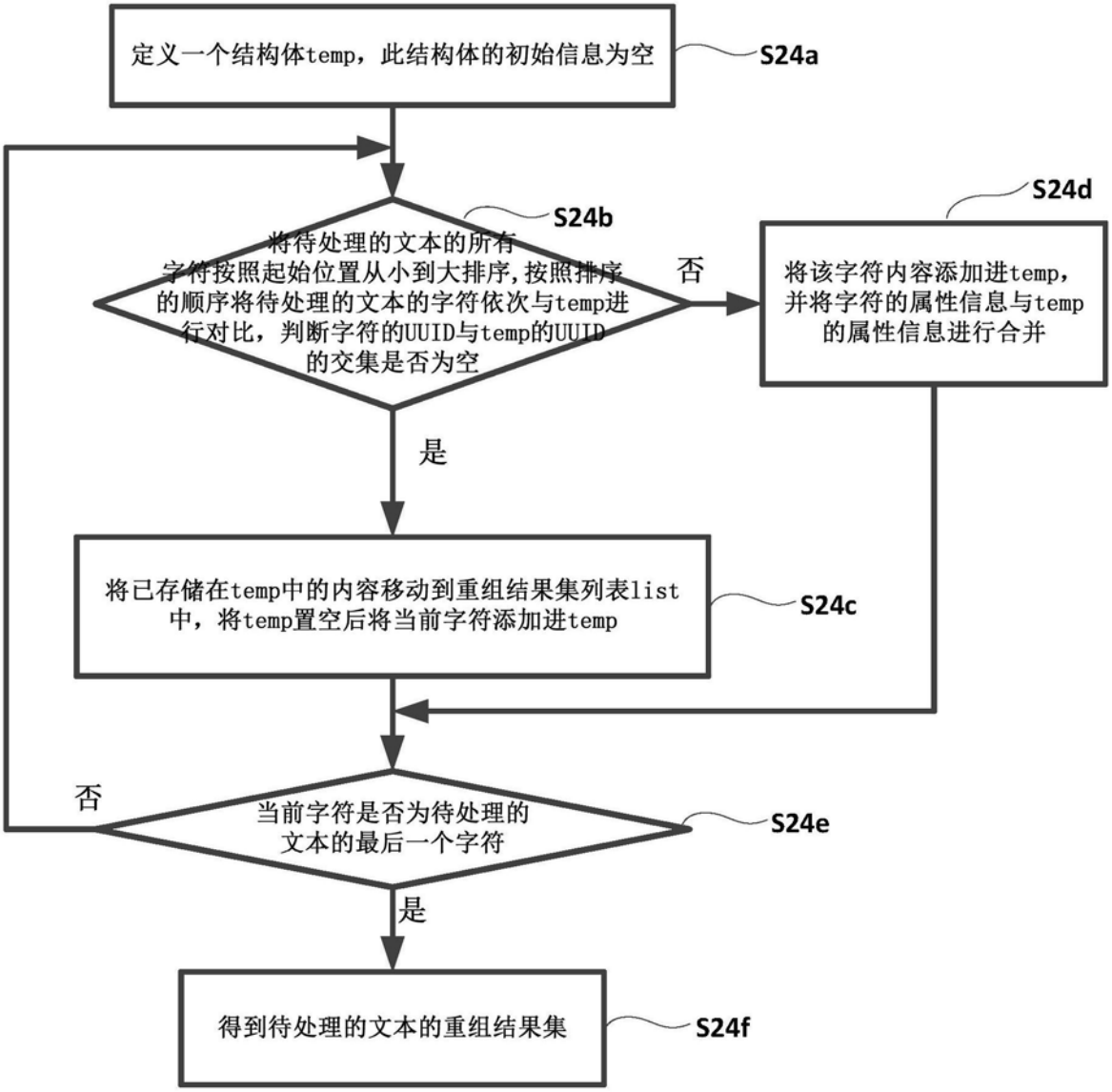

定义一个结构体temp，此结构体的初始信息为空
S24a
将待处理的文本的所有字符按照起始位置从小到大排序，按照排序的顺序将待处理的文本的字符依次与temp进行对比，判断字符的UUID与temp的UUID的交集是否为空
S24b
否
将该字符内容添加进temp，并将字符的属性信息与temp的属性信息进行合并
S24d
是
将已存储在temp中的内容移动到重组结果集列表list中，将temp置空后将当前字符添加进temp
S24c
否
当前字符是否为待处理的文本的最后一个字符
S24e
是
得到待处理的文本的重组结果集
S24f

图4

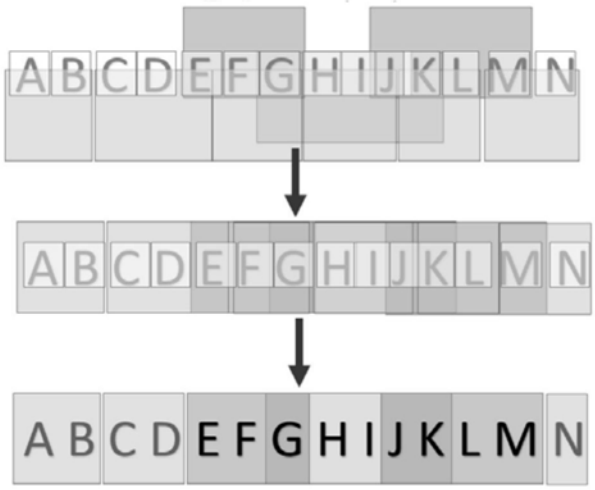

A B C D E F G H I J K L M N
A B C D E F G H I J K L M N
A B C D E F G H I J K L M N

图5

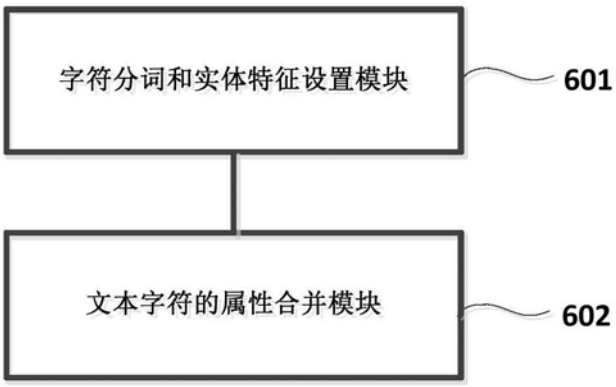

字符分词和实体特征设置模块
601
文本字符的属性合并模块
602

图6