(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107451433 B
(45)授权公告日 2020.05.22

(21)申请号 201710499053.6

(22)申请日 2017.06.27

(65)同一申请的已公布的文献号
申请公布号 CN 107451433 A

(43)申请公布日 2017.12.08

(73)专利权人 中国科学院信息工程研究所
地址 100093 北京市海淀区闵庄路甲89号

(72)发明人 柳厅文 李全刚 李柢颖 亚静 时金桥 谭建龙

(74)专利代理机构 北京君尚知识产权代理有限公司 11200
代理人 邵可声

(51)Int.Cl.
G06F 21/16(2013.01)
G06F 40/211(2020.01)
G06F 40/284(2020.01)
G06N 3/04(2006.01)

(56)对比文件
CN 106886516 A,2017.06.23,
CN 106886516 A,2017.06.23,
CN 106569998 A,2017.04.19,
CN 106570179 A,2017.04.19,
CN 106682220 A,2017.05.17,

审查员 陈玲

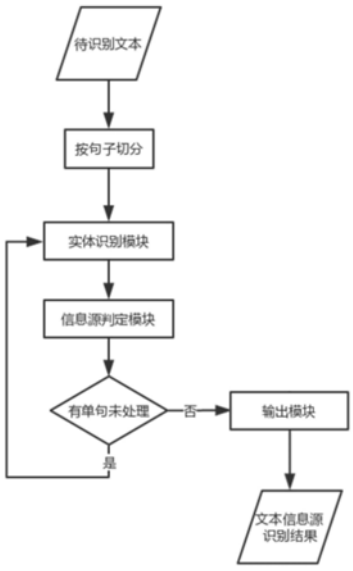
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种基于文本内容的信息源识别方法与装置

(57)摘要

本发明提供一种基于文本内容的信息源识别方法,适用于非结构化的文本,即自由文本,包括以下步骤:将输入的文本按句子切分并分词;识别出各句子中包含的类型为信息源的实体;如所述实体为其所在句子的信息源,则将其作为一信息源实体;整合各句子得到的信息源实体,作为文本信息识别结果。可以不依赖于网页结构化信息,不依赖于人工特征提取,通过分析文本内容,自动识别非结构化文本的信息源。同时提供对应实现上述方法的装置。



CN 107451433 B

1.一种基于文本内容的信息源识别方法,包括以下步骤:

将输入的文本按句子切分并分词;

识别出各句子中包含的类型为信息源的实体,包括:

用基于百度百科训练word2vec得到的词向量表示句子中的每个词;

在模型训练和测试时,输入为词向量序列,输出为与词向量序列等长的标签序列;

标签采用2tag方法,若词属于媒体名称指示词,则对应标签为‘1’,否则标签为‘0’;

网络结构包括输入层,双向LSTM层和输出层;

将多个标签为‘1’的词拼接起来,得到所在句子的候选信息源实体;通过采用基于CNN的句子分类方法,判定类型为信息源的实体是否为其所在句子的信息源,如所述实体为其所在句子的信息源,则将其作为一信息源实体,判定类型为信息源的实体是否为其所在句子的信息源包括:

首先需要将识别出的候选信息源实体合并为一个词,其次每个词的向量包含n维的词向量和m维的位置向量,每个词的向量长度为n+m;

对每个词的词向量表示拼接位置向量,然后输入到一卷积神经网络结构中;

依据网络输出的分类结果,判定该实体是否为其所在句子的信息源;

整合各句子得到的信息源实体,作为文本信息识别结果。

2.如权利要求1所述的基于文本内容的信息源识别方法,其特征在于,对输入的文本按句子切分并分词时,设定一预设句子长度,并通过截取或补充的方式使各句子长度与该预设句子长度一致。

3.如权利要求1所述的基于文本内容的信息源识别方法,其特征在于,通过采用基于双向LSTM的实体识别方法,各句子中包含的类型为信息源的实体。

4.如权利要求1所述的基于文本内容的信息源识别方法,其特征在于,所述卷积神经网络结构包括输入层,卷积层,最大池化层,全连接层及输出层,网络的输出为0或1的分类结果。

5.如权利要求1所述的基于文本内容的信息源识别方法,其特征在于,依据预先设定的文本最大信息源个数整合各句子得到的信息源实体,若信息源实体数量超过所述文本最大信息源个数,优先选取所在句子位置在前的信息源实体。

一种基于文本内容的信息源识别方法与装置

技术领域

[0001] 本发明涉及自然语言处理领域，尤其涉及一种基于文本内容的信息源识别方法与装置。

背景技术

[0002] 信息源作为动态信息的载体，是实施动态监测的重要基础保障，识别文本的信息源，可以用来构建信息源知识库，支撑领域动态信息获取。另一方面，文本中包含更多信息源往往意味着更强的参考性，更高的文本质量，利用文本信息源识别，可以进行文本过滤，从大量文本中筛选出有价值的信息。已有的关于信息源的研究多集中在信息源的特征和分类，信息源挖掘和体系构建，信息源发现等方向，具体到信息源识别的研究较少，仅在阐述实验过程中提及实现方法。已有的报文信息源的识别有基于规则等方法。而对网络信息源的识别主要针对结构化页面，基于链接关系，如网页的内链信息，社交网络的链接引用，或基于来源标注，如门户网站的转载标注，得到文本的信息源。

[0003] 网络文本信息量大，信息传播快，信息获取成本低廉，已成为重要信息源渠道。同时，由于互联网版权法规的不健全、操作难度大、违法成本低以及运作习惯等，各网站间的相互引用繁杂而混乱，且缺乏明显的引用标志。已有的信息源识别方法，仅依赖页面的链接关系或结构化信息标注，无法处理大量的非结构化页面的信息源识别。而基于规则的方法需要人工提取大量特征，工作量大，且领域间通用性差，不能满足实际的处理需求。

发明内容

[0004] 针对上述问题，本发明的目的在于提供一种基于文本内容的信息源识别方法及装置，可以不依赖于网页结构化信息，不依赖于人工特征提取，通过分析文本内容，自动识别非结构化文本的信息源。

[0005] 为达上述目的，本发明采取的技术方案是：

[0006] 一种基于文本内容的信息源识别方法，适用于非结构化的文本，即自由文本，包括以下步骤：

[0007] 将输入的文本按句子切分并分词；

[0008] 识别出各句子中包含的类型为信息源的实体；

[0009] 如所述实体为其所在句子的信息源，则将其作为一信息源实体；

[0010] 整合各句子得到的信息源实体，作为文本信息识别结果。

[0011] 进一步地，对输入的文本按句子切分并分词时，设定一预设句子长度(句子长度指词的数量)，并通过截取或补充的方式使各句子长度与该预设句子长度一致。

[0012] 进一步地，所述类型为信息源的实体为属于媒体名称指示词的词构成的序列。

[0013] 进一步地，通过采用基于双向LSTM的实体识别方法，各句子中包含的类型为信息源的实体。

[0014] 进一步地，通过下述步骤识别出各句子中包含的类型为信息源的实体：

[0015] 用基于百度百科训练word2vec得到的词向量表示句子中的每个词；

[0016] 在模型训练和测试时，输入为词向量序列，输出为与词向量序列等长的标签序列；

[0017] 标签采用2tag方法，若词属于媒体名称指示词，则对应标签为‘1’，否则标签为‘0’；

[0018] 网络结构包括输入层，双向LSTM层和输出层；

[0019] 依据测试数据得到的序列结果将多个标签为‘1’的词拼接起来，即为句子中包含的类型为信息源的实体。

[0020] 进一步地，通过采用基于CNN的句子分类方法，判定类型为信息源的实体是否为其所在句子的信息源。

[0021] 进一步地，判定类型为信息源的实体是否为其所在句子的信息源包括：

[0022] 将类型为信息源的实体合并为一个词

[0023] 对每个词的词向量表示拼接位置向量，然后输入到一卷积神经网络结构中；

[0024] 依据网络输出的测试数据的分类结果，判定该实体是否为其所在句子的信息源。

[0025] 进一步地，每个词的词向量包含n维的语义向量和m维的位置向量，每个词的向量长度为n+m。

[0026] 进一步地，所述卷积神经网络结构包括输入层，卷积层，最大池化层，全连接层及输出层，网络的输出为0或1的分类结果。

[0027] 进一步地，依据预先设定的文本最大信息源个数整合各句子得到的信息源实体，若信息源实体数量超过所述文本最大信息源个数，优先选取所在句子位置在前的信息源实体。

[0028] 一种基于文本内容的信息源识别装置，包括：

[0029] 文本预处理模块，用以将输入的文本按句子切分并分词；

[0030] 实体识别模块，用以识别出各句子中包含的类型为信息源的实体；

[0031] 信息源判定模块，用以判定所述实体是否为其所在句子的信息源，如是，则将其作为一信息源实体；

[0032] 输出模块，用以整合各句子得到的信息源实体，作为文本信息识别结果。

[0033] 具体而言，本发明可以基于文本内容识别其信息源，即判定文本描述内容是否引用自其它机构或网站，如果是，给出引用的结构或网站实体名。通过本发明提供的方法及装置分析文本内容识别信息源，能够避免现有方法识别文本信息源时对网页结构化信息的依赖和对人工提取特征的依赖，提出了基于文本内容的自动化信息源识别方法；并且采用实体识别和句子判定相结合的方法，充分利用了信息源实体内部特征和句式结构特征，不需要大量人工干预，有效解决了基于文本内容的信息源识别问题。

附图说明

[0034] 图1是本发明一实施例中基于文本内容的信息源识别方法的数据处理流程图

具体实施方式

[0035] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整的描述。

[0036] 如图1所示，在一实施例中，提供了一种基于文本内容的信息源识别方法，适用于非结构化的文本，即自由文本，包括以下步骤：

[0037] 将输入的文本按句子切分并分词；即将输入的文本按句子切分并分词，逐句进行处理。

[0038] 识别出各句子中包含的类型为信息源的实体；即识别句子是否包含类型为信息源的实体，如果是，得到句子及其候选信息源实体，进行下一步处理；如果否，跳过步骤3)继续进行下一个句子的处理。

[0039] 如所述实体为其所在句子的信息源，则将其作为一信息源实体；判断候选信息源实体是否是其所在句子的信息源，如果是，将该候选信息源作为一个信息源实体；

[0040] 整合各句子得到的信息源实体，作为文本信息识别结果。综合逐句识别出的信息源实体，得到并输出文本信息源识别结果。

[0041] 对应实现上述方法的装置，包括：

[0042] 文本预处理模块，用以将输入的文本按句子切分并分词；

[0043] 实体识别模块，用以识别出各句子中包含的类型为信息源的实体；

[0044] 信息源判定模块，用以判定所述实体是否为其所在句子的信息源，如是，则将其作为一信息源实体；

[0045] 输出模块，用以整合各句子得到的信息源实体，作为文本信息识别结果。

[0046] 具体说明各方法步骤及实现模块：

[0047] 上述文本预处理模块，主要指对文本按句子切分，并且对句子分词，设定句子最大长度(句子长度指词的数量)，超出截取，不足补齐。

[0048] 上述实体识别模块，抽取属于机构或网站名(媒体名称指示词)的词构成的序列，作为候选信息源实体。采用基于双向LSTM的Seq2seq的方法，用基于百度百科训练word2vec得到的词向量表示句子中的每个词。在模型训练和测试时，输入为词向量序列，输出为标签序列，与词向量序列等长。标签采用2tag方法，若词属于机构或网站名，则对应标签为‘1’，否则标签为‘0’。网络结构包括输入层，双向LSTM层和输出层。最后依据测试数据得到的序列结果将多个标签为‘1’的词拼接起来，即为所在句子的候选信息源实体。采用了基于深度学习的实体识别方式，不需要人工制定规则，相比于基于词匹配的规则方法，花费时间短，效果好，领域间可迁移性强。

[0049] 上述信息源判定模块，采用基于CNN(卷积神经网络)的分类方法，即给定句子及对应候选信息源实体，判定候选实体是否为该句信息源。网络的输入同样为词向量序列，不同于实体识别模块，首先需要将识别出的候选信息源实体合并为一个词，其次每个词的向量包含n维的语义向量(word2vec训练得到的词向量)和m维的位置向量(计算每个词到候选信息源实体的距离并将其向量化)，即每个词的向量长度为n+m。网络结构包括输入层，卷积层，最大池化层，全连接层及输出层。网络的输出为0或1的分类结果。依据测试数据的分类结果，若输出结果为1，即判定为正例，可以认为该实体为句子的信息源实体。该方法借鉴了句子关系分类方法，对每个词的词向量表示拼接位置向量，然后输入到通用的卷积神经网络结构中，实现对词语与句子间关系的判定。

[0050] 上述输出模块，需要对逐句得到的信息源实体整合。即依据预先设定的每篇文本最大信息源个数处理，若识别的实体数量超过最大个数，优先选取所在句子位置在前的实

体。

[0051] 下面提供一实际案例，该案例具体说明了本发明提供对军事领域，某军事论坛的帖子内容进行信息源识别的过程。

[0052] 文本预处理，将输入的文本按句子切分并分词，预设的句子最大长度为50。如句子“据英国《简氏防务周刊》11月1日报道称，中国第40艘江岛级(056/056A型)护卫舰10月28日在广州黄埔造船厂下水。”分词得到“['据','英国','《','简氏','防务','周刊','》','11','月','1','日','报道','称',',','中国','第','40','艘','江岛','级','(','056','/','056','A型',')','护卫舰','10','月','28','日','在','广州','黄埔','造船厂','下水','。']”，长度为37，则在句子后面填充13个‘PADDING’作为填充词，得到词向量序列。如果长度超过50，从前到后截取50词即可。

[0053] 实体识别模块，采用基于双向LSTM的Seq2seq的方法。查词向量表(百度百科语料训练word2vec得到的词向量)表示句子中的每个词，词向量维度为50，则句子可以表示为50*40维的向量。将向量输入得到训练好的双向LSTM网络，得到序列标注结果。依据测试数据得到的序列结果将多个标签为‘1’的词拼接起来，即为所在句子的候选信息源实体。对例句可以得到’简氏’,’防务’,’周刊’三个词对应的标签为‘1’，将三个词拼接起来，得到候选信息源实体“简氏防务周刊”。

[0054] 信息源判定模块，采用基于CNN的分类方法。例句词向量序列为['据','英国','《','简氏防务周刊','》','11','月','1','日','报道','称',',','中国','第','40','艘','江岛','级','(','056','/','056','A型',')','护卫舰','10','月','28','日','在','广州','黄埔','造船厂','下水','。']”，此时句子长度为35，填充15个‘PADDING’，句子序列长度为50。计算每个词到信息源实体“简氏防务周刊”的距离并转换为10维向量。拼接每个词的语义向量50维，得到60*40维的向量。输入训练好的CNN神经网络，得到结果为1，即可以认为该实体为句子的信息源实体。

[0055] 输出模块，依照文本长度，预先设定每篇文本的信息源个数上限为3，取一篇文本识别出的前3个信息源实体组合，即为最终识别结果。

[0056] 显然，所描述的实施例仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

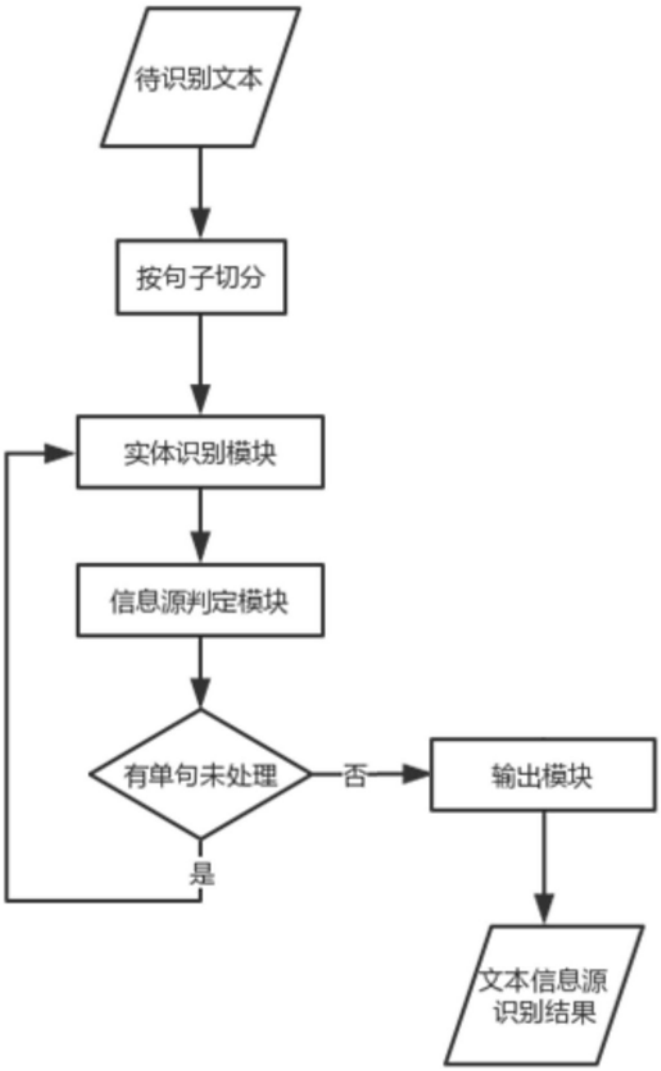


图1