(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2020 年 12 月 10 日 (10.12.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/244065 A1

(51) 国际专利分类号:
G06F 40/20 (2020.01)

(21) 国际申请号: PCT/CN2019/102462

(22) 国际申请日: 2019 年 8 月 26 日 (26.08.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910483399.6 2019年6月4日 (04.06.2019) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518033 (CN)。

(72) 发明人: 陈闽川 (CHEN, Minchuan); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518033 (CN)。 马骏 (MA, Jun); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518033 (CN)。 王少军 (WANG, Shaojun); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518033 (CN)。

(74) 代理人: 北京市京大律师事务所 (BEIJING JINGDA LAW FIRM); 中国北京市海淀区海淀中街16号7层2单元703, Beijing 100080 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,

(54) **Title:** CHARACTER VECTOR DEFINITION METHOD, APPARATUS AND DEVICE BASED ON ARTIFICIAL INTELLIGENCE, AND STORAGE MEDIUM

(54) **发明名称**: 基于人工智能的字向量定义方法、装置、设备及存储介质

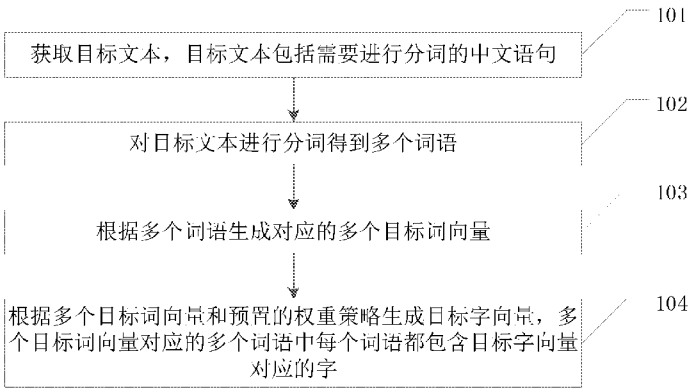

图 1

101 Acquire target text, wherein the target text comprises a Chinese statement needing to be subjected to word segmentation
102 Perform word segmentation on the target text to obtain a plurality of words
103 Generate a plurality of corresponding target word vectors according to the plurality of words
104 Generate target character vectors according to the plurality of target word vectors and a preset weight policy, wherein each word from among the plurality of words corresponding to the plurality of target word vectors includes characters corresponding to the target character vectors

(57) **Abstract:** The present application relates to the technical field of artificial intelligence, and particularly relates to the field of word segmentation models. Disclosed are a character vector definition method, apparatus and device based on artificial intelligence, and a storage medium, which are used for giving a plurality of different meanings to each character in different words by taking a single character as the minimum structure, thereby improving the accuracy of the meaning of a character vector in a sentence. The method of the present application comprises: acquiring target text, wherein the target text comprises a Chinese statement needing to be subjected to word segmentation; performing word segmentation on the target text to obtain a plurality of words; generating a plurality of

WO 2020/244065 A1

[见续页]

GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

corresponding target word vectors according to the plurality of words; and generating target character vectors according to the plurality of target word vectors and a preset weight policy, wherein each word from among the plurality of words corresponding to the plurality of target word vectors includes characters corresponding to the target character vectors.

(57) 摘要：本申请涉及人工智能技术领域，尤其涉及分词模型领域，公开了一种基于人工智能的字向量定义方法、装置、设备及存储介质，用于将单个字作为最小结构，给每个字在不同的词语中赋予多个不同的含义，增加了字向量在句子中含义的准确性。本申请方法包括：获取目标文本，目标文本包括需要进行分词的中文语句；对目标文本进行分词得到多个词语；根据多个词语生成对应的多个目标词向量；根据多个目标词向量和预置的权重策略生成目标字向量，多个目标词向量对应的多个词语中每个词语都包含目标字向量对应的字。

基于人工智能的字向量定义方法、装置、设备及存储介质

本申请要求于2019年6月4日提交中国专利局、申请号为201910483399.6、发明名称为“基于人工智能的字向量定义方法、装置、设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在申请中。

技术领域

本申请涉及分词模型领域，尤其涉及一种基于人工智能的字向量定义方法、装置、设备及存储介质。

背景技术

随着人工智能技术的快速发展，深度学习技术的突飞猛进，为自然语言处理（natural language processing，NLP）提供了强大的工具。其中，深度学习为自然语言处理带来的最令人兴奋的突破是词向量技术和字向量技术。词向量技术是将词转化成为稠密向量，并且对于相似的词，其对应的词向量也相近。在自然语言处理应用中，词向量和字向量作为深度学习模型的特征进行输入。因此，最终模型的效果很大程度上取决于词向量和字向量的效果。

目前在具体实践中，字向量与词向量的定义是相对于独立的，在做实体识别的时候大部分采用字向量，在做文本分类，主题提取的时候大部分采用词向量的方式来进行识别。词向量的缺点是数量庞大而且在实体抽取中，容易出现小词不对而导致大词不对的问题。字向量的缺点在于，单个的词可能会有多完全不相关的意思，比如“老人”中的‘老’和“老子”中的‘老’，他的字向量只能表示一个意思和我们理解的不符合。

发明人意识到现有方案中，对于同一个字，当应用到词中时，单个的词中的字义可能会与词义完全不相关，字向量表意单一。

发明内容

本申请提供了一种基于人工智能的字向量定义方法、装置、设备及存储介质，用于将单个字作为最小结构，考虑字向量和词向量的结合，给每个字在不同的词语中赋予多个不同的含义，增加了字向量在句子中含义的准确性。

本申请实施例的第一方面提供一种基于人工智能的字向量定义方法，包括：获取目标文本，所述目标文本包括需要进行分词的中文语句；对所述目标文本进行分词得到多个词语；根据所述多个词语生成对应的多个目标词向量；根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标

字向量对应的字；将每个目标字向量作为模型参数输入到长短时记忆网络 LSTM 和条件随机场 CRF 模型中，对所述中文语句进行实体识别生成预测分词。

可选的，在本申请实施例第一方面的第一种实现方式中，所述根据所述多个词语生成对应的多个目标词向量包括：将所述多个词语输入到预置的算法模型中；将每个词语映射成 K 维词语向量，其中，K 为大于 0 的整数；计算得到每个词语向量之间的距离；根据所述每个词语向量之间的距离确定每个词语向量之间的语义相似度；根据所述每个词语向量之间的语义相似度确定与所述多个词语中目标词语的语义相似度最高的向量为目标词向量；确定多个目标词向量，每个目标词向量对应一个词语。

可选的，在本申请实施例第一方面的第二种实现方式中，在所述将所述多个词语输入到预置的算法模型中之前，所述方法还包括：随机生成一个词向量矩阵，每一个行对应一个单词的向量；在所述词向量矩阵确定一个目标单词，并从所述词向量矩阵中提取周边单词的词向量；计算所述周边单词的词向量的均值向量；将所述均值向量输入到预置的逻辑回归模型中进行训练；生成预置的算法模型，所述预置的算法模型输出的概率向量与所述目标单词的 one-hot 编码向量相匹配。

可选的，在本申请实施例第一方面的第三种实现方式中，所述计算得到每个词语向量之间的距离包括：在所述词语向量中确定第一向量和第二向量；计算所述第一向量和所述第二向量之间的余弦值，满足公式：$Sim(D_1,D_2)=\cos\theta=\frac{\sum_{k=1}^{n}W_{1k}W_{2k}}{\sqrt{(\sum_{k=1}^{n}W_{1k}^2)(\sum_{k=1}^{n}W_{2k}^2)}}$，其中，$D_1$、$D_2$ 分别表示第一向量和第二向量，W_{1k} 表示第一向量的权重，W_{2k} 表示第二向量的权重，第一向量和第二向量都包括 N 个特征值，$1\le k\le N$；将所述余弦值确定为所述第一向量和所述第二向量的距离。

可选的，在本申请实施例第一方面的第四种实现方式中，所述根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字包括：获取预置的权重策略，所述预置的权重策略包含每个词向量的权重值；确定所述多个目标词向量中每个词向量的目标权重值；根据所述多个目标词向量和所述多个目标词向量中各个目标权重值生成目标字向量。

可选的，在本申请实施例第一方面的第五种实现方式中，所述根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字之后，所述方法还包括：将每个目标字向量作为模型参数输

入到长短时记忆网络 LSTM 和条件随机场 CRF 模型中，对所述中文语句进行实体识别生成预测分词。

可选的，在本申请实施例第一方面的第六种实现方式中，所述将每个目标字向量作为模型参数输入到长短时记忆网络 LSTM 和条件随机场 CRF 模型中，对所述中文语句进行实体识别生成预测分词包括：将每个目标字向量作为模型参数输入到 LSTM 和 CRF 模型中；确定每个目标字向量在字空间中的位置；根据所述每个目标字向量在字空间中的位置对所述中文语句进行实体识别，生成预测分词。

本申请实施例的第二方面提供了一种基于人工智能的字向量定义装置，包括：获取单元，用于获取目标文本，所述目标文本包括需要进行分词的中文语句；分词单元，用于对所述目标文本进行分词得到多个词语；第一生成单元，用于根据所述多个词语生成对应的多个目标词向量；第二生成单元，用于根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字。

可选的，在本申请实施例第二方面的第一种实现方式中，第一生成单元包括：第一输入模块，用于将所述多个词语输入到预置的算法模型中；映射模块，用于将每个词语映射成 K 维词语向量，其中，K 为大于 0 的整数；第一计算模块，用于计算得到每个词语向量之间的距离；第一确定模块，用于根据所述每个词语向量之间的距离确定每个词语向量之间的语义相似度；第二确定模块，用于根据所述每个词语向量之间的语义相似度确定与所述多个词语中目标词语的语义相似度最高的向量为目标词向量；第三确定模块，用于确定多个目标词向量，每个目标词向量对应一个词语。

可选的，在本申请实施例第二方面的第二种实现方式中，第一生成单元还包括：第一生成模块，用于随机生成一个词向量矩阵，每一个行对应一个单词的向量；提取模块，用于在所述词向量矩阵确定一个目标单词，并从所述词向量矩阵中提取周边单词的词向量；第二计算模块，用于计算所述周边单词的词向量的均值向量；训练模块，用于将所述均值向量输入到预置的逻辑回归模型中进行训练；第二生成模块，用于生成预置的算法模型，所述预置的算法模型输出的概率向量与所述目标单词的 one-hot 编码向量相匹配。

可选的，在本申请实施例第二方面的第三种实现方式中，第一计算模块具体用于：在所述词语向量中确定第一向量和第二向量；计算所述第一向量和所述第二向量之间的余弦

值，满足公式：$Sim(D_1,D_2)=\cos\theta=\frac{\sum_{k=1}^{n}W_{1k}W_{2k}}{\sqrt{(\sum_{k=1}^{n}W_{1k}^2)(\sum_{k=1}^{n}W_{2k}^2)}}$，其中，$D_1$、$D_2$分别表示第一向量和第二向量，$W_{1k}$表示第一向量的权重，$W_{2k}$表示第二向量的权重，第一向量和第二向量都包括 N 个特征值，$1\le k\le N$；将所述余弦值确定为所述第一向量和所述第二向量的距离。

可选的，在本申请实施例第二方面的第四种实现方式中，第二生成单元具体用于：获取预置的权重策略，所述预置的权重策略包含每个词向量的权重值；确定所述多个目标词向量中每个词向量的目标权重值；根据所述多个目标词向量和所述多个目标词向量中各个目标权重值生成目标字向量。

可选的，在本申请实施例第二方面的第五种实现方式中，基于人工智能的字向量定义装置还包括：第三生成单元，用于将每个目标字向量作为模型参数输入到长短时记忆网络 LSTM 和条件随机场 CRF 模型中，对所述中文语句进行实体识别生成预测分词。

可选的，在本申请实施例第二方面的第六种实现方式中，第三生成单元包括：第二输入模块，用于将每个目标字向量作为模型参数输入到 LSTM 和 CRF 模型中；第四确定模块，用于确定每个目标字向量在字空间中的位置；识别生成模块，用于根据所述每个目标字向量在字空间中的位置对所述中文语句进行实体识别，生成预测分词。

本申请实施例的第三方面提供了一种基于人工智能的字向量定义设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现上述任一实施方式所述的基于人工智能的字向量定义方法。

本申请实施例的第四方面提供了一种非易失性计算机可读存储介质，包括指令，当所述指令在计算机上运行时，使得计算机执行上述任一实施方式所述的基于人工智能的字向量定义方法的步骤。

本申请实施例提供的技术方案中，获取目标文本，目标文本包括需要进行分词的中文语句；对目标文本进行分词得到多个词语；根据多个词语生成对应的多个目标词向量；根据多个目标词向量和预置的权重策略生成目标字向量，多个目标词向量对应的多个词语中每个词语都包含目标字向量对应的字。本申请实施例，将单个字作为最小结构，考虑字向量和词向量的结合，给每个字在不同的词语中赋予多个不同的含义，增加了字向量在句子中含义的准确性，进而提高了中文分词的效率。

附图说明

图 1 为本申请实施例中基于人工智能的字向量定义方法的一个实施例示意图；

图 2 为本申请实施例中基于人工智能的字向量定义方法的另一个实施例示意图；

图 3 为本申请实施例中基于人工智能的字向量定义装置的一个实施例示意图；

图 4 为本申请实施例中基于人工智能的字向量定义装置的另一个实施例示意图；

图 5 为本申请实施例中基于人工智能的字向量定义设备的一个实施例示意图。

具体实施方式

本申请提供了一种基于人工智能的字向量定义方法、装置、设备及存储介质，用于将单个字作为最小结构，考虑字向量和词向量的结合，给每个字在不同的词语中赋予多个不同的含义，增加了每个字在句子中含义的准确性，提高了中文分词的效率。

为了使本技术领域的人员更好地理解本申请方案，下面将结合本申请实施例中的附图，对本申请实施例进行描述。

本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等（如果存在）是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。此外，术语“包括”或“具有”及其任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

请参阅图 1，本申请实施例提供的基于人工智能的字向量定义方法的流程图，具体包括：

101、获取目标文本，目标文本包括需要进行分词的中文语句。

服务器获取目标文本，该目标文本包括需要进行分词的中文语句。例如，中文语句可以为“我喜欢苹果”，或者是“工程师傅打电脑”等，本申请实施例以“工程师傅打电脑”作为中文语句进行说明。

可以理解的是，本申请的执行主体可以为基于人工智能的字向量定义装置，还可以是终端或者服务器，具体此处不做限定。本申请以服务器为执行主体为例进行说明。

102、对目标文本进行分词得到多个词语。

服务器利用预置的分词工具，例如，结巴中文分词工具、HanLP 等工具，对目标文本

进行分词得到多个词语。例如，目标文本是“我是一名算法工程师”，那么，可以分别得到“我”、“是”、“一名”、“算法”、“工程师”等五个词语。

例如，对目标文本“工程师傅打电脑”进行分词，采用 3-Gram 或 2-Gram 分词，发现文本前面有“工程师”、“程师”，后面有“师傅”、“师傅打”，总共有“工程师”、“程师”、“师傅”、“师傅打”。

103、根据多个词语生成对应的多个目标词向量。

服务器根据多个词语定义对应的多个目标词向量。具体的，服务器将多个词语输入到预置的算法模型中；服务器将每个词语映射成 K 维词语向量，其中，K 为大于 0 的整数；服务器计算得到每个词语向量之间的距离；服务器根据每个词语向量之间的距离确定每个词语向量之间的语义相似度；服务器根据每个词语向量之间的语义相似度确定与多个词语中目标词语的语义相似度最高的向量为目标词向量；服务器确定多个目标词向量，每个目标词向量对应一个词语。

例如，服务器通过预置的模型，将上面识别到的四个词输入到预置的模型中，在预置的模型中匹配到了“工程师”，“师傅”这两个词，确定词向量“工程师”和词向量“师傅”。

需要说明的是，生成的预置的模型，需要确保每个词向量的有效性。如果在词向量中只有“工程”，“师傅”，当我们遇到一个句子为“工程师是一个好职业”时，预置的模型预测出的结果只能和已有的词向量相匹配，如果继续做实体识别，就会得到“工程（B） 师（E） 是” 割裂了“工程师”这个整体，虽然也可以得到结果，但是信息缺被割裂了。

在向量空间模型中，目标文本泛指各种机器可读的记录。文本用 D（Document）表示，特征项用 T（Term）表示，T 是指出现在文档 D 中且能够代表该文档内容的基本语言单位，主要是由词或者短语构成，文本可以用特征项集表示为 D(T1，T2，…，Tn)，其中 Tk 是特征项，$1 \le k \le N$。例如一篇文档中有 a、b、c、d 四个特征项，那么这篇文档就可以表示为 D(a，b，c，d)。对含有 n 个特征项的文本而言，通常会给每个特征项赋予一定的权重表示其重要程度。即 D＝D(T1，W1；T2，W2；…，Tn，Wn)，简记为 D＝D(W1，W2，…，Wn)，我们把它叫做文本 D 的向量表示。

104、根据多个目标词向量和预置的权重策略生成目标字向量，多个目标词向量对应的多个词语中每个词语都包含目标字向量对应的字。

服务器根据对应的多个词向量和预置的权重策略生成目标字向量，其中，多个词向量对应的多个词语中每个词语都包含该字向量对应的字。例如，对“师”字进行定义字向量，

则现在“师”的字向量为（师的字向量+师傅的词向量+工程师的词向量）/3，得到“师”的目标字向量。

需要说明的是，为了反映字对于周围环境影响而导致字向量的变化情况，除了上述例子中取平均值的方法，丰富了字向量的属性，还可以考虑用注意力机制 attention 为每一个字都自己计算一个适合权重然后加权平均，可参考现有技术，具体此处不再赘述。

本申请实施例，获取目标文本，目标文本包括需要进行分词的中文语句；对目标文本进行分词得到多个词语；根据多个词语生成对应的多个目标词向量；根据多个目标词向量和预置的权重策略生成目标字向量，多个目标词向量对应的多个词语中每个词语都包含目标字向量对应的字。将单个字作为最小结构，考虑字向量和词向量的结合，给每个字在不同的词语中赋予多个不同的含义，增加了字向量在句子中含义的准确性。

请参阅图 2，本申请实施例提供的基于人工智能的字向量定义方法的另一个流程图，具体包括：

201、获取目标文本，目标文本包括需要进行分词的中文语句。

服务器获取目标文本，该目标文本包括需要进行分词的中文语句。例如，中文语句可以为“我喜欢苹果”，或者是“工程师傅打电脑”等，本申请实施例以“工程师傅打电脑”作为中文语句进行说明。

可以理解的是，本申请的执行主体可以为基于人工智能的字向量定义装置，还可以是终端或者服务器，具体此处不做限定。本申请以服务器为执行主体为例进行说明。

202、对目标文本进行分词得到多个词语。

服务器利用预置的分词工具，例如，结巴中文分词工具、HanLP 等工具，对目标文本进行分词得到多个词语。例如，目标文本是“我是一名算法工程师”，那么，可以分别得到“我”、“是”、“一名”、“算法”、“工程师”等五个词语。

例如，对目标文本“工程师傅打电脑”进行分词，采用 3-Gram 或 2-Gram 分词，发现文本前面有“工程师”、“程师”，后面有“师傅”、“师傅打”，总共有“工程师”、“程师”、“师傅”、“师傅打”。

203、根据多个词语生成对应的多个目标词向量。

服务器根据多个词语定义对应的多个目标词向量。具体的，服务器将多个词语输入到预置的算法模型中；服务器将每个词语映射成 K 维词语向量，其中，K 为大于 0 的整数；服务器计算得到每个词语向量之间的距离；服务器根据每个词语向量之间的距离确定每个

词语向量之间的语义相似度；服务器根据每个词语向量之间的语义相似度确定与多个词语中目标词语的语义相似度最高的向量为目标词向量；服务器确定多个目标词向量，每个目标词向量对应一个词语。

例如，服务器通过预置的模型，将上面识别到的四个词输入到预置的模型中，在预置的模型中匹配到了“工程师”,“师傅”这两个词，确定词向量“工程师”和词向量“师傅”。

需要说明的是，生成的预置的模型，需要确保每个词向量的有效性。如果在词向量中只有“工程”,“师傅”，当我们遇到一个句子为“工程师是一个好职业”时，预置的模型预测出的结果只能和已有的词向量相匹配，如果继续做实体识别，就会得到“工程（B） 师（E） 是” 割裂了“工程师”这个整体，虽然也可以得到结果，但是信息缺被割裂了。

在向量空间模型中，目标文本泛指各种机器可读的记录。文本用 D（Document）表示，特征项用 T（Term）表示，T 是指出现在文档 D 中且能够代表该文档内容的基本语言单位，主要是由词或者短语构成，文本可以用特征项集表示为 D(T1，T2，…，Tn)，其中 Tk 是特征项， 。例如一篇文档中有 a、b、c、d 四个特征项，那么这篇文档就可以表示为 D(a，b，c，d)。对含有 n 个特征项的文本而言，通常会给每个特征项赋予一定的权重表示其重要程度。即 D＝D(T1，W1；T2，W2；…，Tn，Wn)，简记为 D＝D(W1，W2，…，Wn)，我们把它叫做文本 D 的向量表示。其中 Wk 是 Tk 的权重，$1 \le k \le N$。在上述例子中，假设 a、b、c、d 的权重分别为 30，20，20，10，那么该文本的向量可以表示为 D(30，20，20，10)。在向量空间模型中，两个文本 D1 和 D2 之间的内容相关度 Sim(D1，D2)常用向量之间夹角的余弦值表示。

例如，文本 D1 的特征项为 a，b，c，d，权值分别为 30，20，20，10，文本 C1 的特征项为 a，c，d，e，权值分别为 40，30，20，10，则 D1 的向量表示为 D1(30,20,20,10,0),C1 的向量表示为 C1（40，0，30，20，10），则计算出来的文本 D1 与文本 C1 相似度是 0.86。

204、根据多个目标词向量和预置的权重策略生成目标字向量，多个目标词向量对应的多个词语中每个词语都包含目标字向量对应的字。

服务器根据对应的多个词向量和预置的权重策略生成目标字向量，其中，多个词向量对应的多个词语中每个词语都包含该字向量对应的字。例如，对“师”字进行定义字向量，则现在“师”的字向量为（师的字向量+师傅的词向量+工程师的词向量）/3，得到“师”的目标字向量。

需要说明的是，为了反映字对于周围环境影响而导致字向量的变化情况，除了上述例

子中取平均值的方法，丰富了字向量的属性，还可以考虑用注意力机制 attention 为每一个字都自己计算一个适合权重然后加权平均，可参考现有技术，具体此处不再赘述。

205、将每个目标字向量作为模型参数输入到长短时记忆网络 LSTM 和条件随机场 CRF 模型中，对中文语句进行实体识别生成预测分词。

服务器将每个目标字向量作为模型参数输入到长短时记忆网络 LSTM 和条件随机场 CRF 模型中，对中文语句进行实体识别生成预测分词。具体的，服务器将每个目标字向量作为模型参数输入到 LSTM 和 CRF 模型中；服务器确定每个目标字向量在字空间中的位置；服务器根据每个目标字向量在字空间中的位置对中文语句进行实体识别，生成预测分词。

可选的，根据每个目标字向量在字空间中的位置对中文语句进行实体识别，生成预测分词的过程具体包括：

将中文语句表示为 $\mathrm{X}=(\mathrm{x1},\mathrm{x2},...,\mathrm{x}n)$，根据预置的公式生成标签序列 $\mathrm{y}=(\mathrm{y1},\mathrm{y2},...,\mathrm{y}n)$，$\mathrm{y1}$ 和 $\mathrm{y}n$ 分别表示中文语句的开始标签和结束标签，预置的公式为：$s(X|y)=\sum_{i=0}^{n}A_{yi,yi+1}+\sum_{i=0}^{n}P_{i,yi}$；其中，P 为双向 LSTM 的输出经过全连接层后映射的分数矩阵，$\mathrm{P}_{i,j}$ 表示中文语句中第 i 个字对应的第 j 个标签的分数，其维度为 n×k，k 为标签的个数，A 表示分词标签的转移矩阵，$\mathrm{A}_{i,j}$ 表示标签 i 与标签 j 之间的过渡分数，$1\le i\le k,1\le j\le k$；确定分词标签序列的概率 p，满足公式：$p(y|X)=\frac{e^{s(X|y)}}{\sum_{\tilde{y}\in YX}e^{s(X|\tilde{y})}}$；计算概率 p 的损失函数：$-\log(p(y|X))=-s(X|y)+\log(\sum_{\tilde{y}\in YX}e^{s(X|\tilde{y})})$，其中，$\mathrm{Y}_{\mathrm{X}}$ 表示中文语句 X 所有的标签序列；根据损失函数确定分数最高的标签作为预测分词，预测分词满足：$y^{*}=\arg\max_{\tilde{y}\in YX}s(X|\tilde{y})$。

本申请实施例，获取目标文本，目标文本包括需要进行分词的中文语句；对目标文本进行分词得到多个词语；根据多个词语生成对应的多个目标词向量；根据多个目标词向量和预置的权重策略生成目标字向量，多个目标词向量对应的多个词语中每个词语都包含目标字向量对应的字。将单个字作为最小结构，考虑字向量和词向量的结合，给每个字在不

同的词语中赋予多个不同的含义，增加了字向量在句子中含义的准确性，进而提高了中文分词的效率。

上面对本申请实施例中基于人工智能的字向量定义方法进行了描述，下面对本申请实施例中基于人工智能的字向量定义装置进行描述，请参阅图 3，本申请实施例中基于人工智能的字向量定义装置的一个实施例包括：

获取单元 301，用于获取目标文本，所述目标文本包括需要进行分词的中文语句；

分词单元 302，用于对所述目标文本进行分词得到多个词语；

第一生成单元 303，用于根据所述多个词语生成对应的多个目标词向量；

第二生成单元 304，用于根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字。

本申请实施例，获取目标文本，目标文本包括需要进行分词的中文语句；对目标文本进行分词得到多个词语；根据多个词语生成对应的多个目标词向量；根据多个目标词向量和预置的权重策略生成目标字向量，多个目标词向量对应的多个词语中每个词语都包含目标字向量对应的字。将单个字作为最小结构，考虑字向量和词向量的结合，给每个字在不同的词语中赋予多个不同的含义，增加了字向量在句子中含义的准确性。

请参阅图 4，本申请实施例中基于人工智能的字向量定义装置的另一个实施例包括：

获取单元 301，用于获取目标文本，所述目标文本包括需要进行分词的中文语句；

分词单元 302，用于对所述目标文本进行分词得到多个词语；

第一生成单元 303，用于根据所述多个词语生成对应的多个目标词向量；

第二生成单元 304，用于根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字。

可选的，第一生成单元 303 包括：

第一输入模块 30301，用于将所述多个词语输入到预置的算法模型中；

映射模块 30302，用于将每个词语映射成 K 维词语向量，其中，K 为大于 0 的整数；

第一计算模块 30303，用于计算得到每个词语向量之间的距离；第一确定模块，用于根据所述每个词语向量之间的距离确定每个词语向量之间的语义相似度；

第二确定模块 30304，用于根据所述每个词语向量之间的语义相似度确定与所述多个词语中目标词语的语义相似度最高的向量为目标词向量；

第三确定模块 30305，用于确定多个目标词向量，每个目标词向量对应一个词语。

可选的，第一生成单元 303 还包括：

第一生成模块 30306，用于随机生成一个词向量矩阵，每一个行对应一个单词的向量；

提取模块 30307，用于在所述词向量矩阵确定一个目标单词，并从所述词向量矩阵中提取周边单词的词向量；

第二计算模块 30308，用于计算所述周边单词的词向量的均值向量；

训练模块 30309，用于将所述均值向量输入到预置的逻辑回归模型中进行训练；

第二生成模块 30310，用于生成预置的算法模型，所述预置的算法模型输出的概率向量与所述目标单词的 one-hot 编码向量相匹配。

可选的，第一计算模块 30303 具体用于：

在所述词语向量中确定第一向量和第二向量；计算所述第一向量和所述第二向量之间的余弦值，满足公式：$Sim(D_1,D_2)=\cos\theta=\frac{\sum_{k=1}^{n}W_{1k}W_{2k}}{\sqrt{(\sum_{k=1}^{n}W_{1k}^2)(\sum_{k=1}^{n}W_{2k}^2)}}$，其中，$D_1$、$D_2$ 分别表示第一向量和第二向量，W_{1k} 表示第一向量的权重，W_{2k} 表示第二向量的权重，第一向量和第二向量都包括 N 个特征值，$1\le k\le N$；将所述余弦值确定为所述第一向量和所述第二向量的距离。

可选的，第二生成单元 304 具体用于：

获取预置的权重策略，所述预置的权重策略包含每个词向量的权重值；确定所述多个目标词向量中每个词向量的目标权重值；根据所述多个目标词向量和所述多个目标词向量中各个目标权重值生成目标字向量。

可选的，基于人工智能的字向量定义装置还包括：

第三生成单元 305，用于将每个目标字向量作为模型参数输入到长短时记忆网络 LSTM 和条件随机场 CRF 模型中，对所述中文语句进行实体识别生成预测分词。

可选的，第三生成单元 305 具体包括：

第二输入模块 3051，用于将每个目标字向量作为模型参数输入到 LSTM 和 CRF 模型中；

第四确定模块 3052，用于确定每个目标字向量在字空间中的位置；

识别生成模块 3053，用于根据所述每个目标字向量在字空间中的位置对所述中文语句进行实体识别，生成预测分词。

本申请实施例，获取目标文本，目标文本包括需要进行分词的中文语句；对目标文本

进行分词得到多个词语；根据多个词语生成对应的多个目标词向量；根据多个目标词向量和预置的权重策略生成目标字向量，多个目标词向量对应的多个词语中每个词语都包含目标字向量对应的字。将单个字作为最小结构，考虑字向量和词向量的结合，给每个字在不同的词语中赋予多个不同的含义，增加了字向量在句子中含义的准确性，进而提高了中文分词的效率。

上面图 3 至图 4 从模块化功能实体的角度对本申请实施例中的基于人工智能的字向量定义装置进行详细描述，下面从硬件处理的角度对本申请实施例中基于人工智能的字向量定义设备进行详细描述。

图 5 是本申请实施例提供的一种基于人工智能的字向量定义设备的结构示意图，该基于人工智能的字向量定义设备 500 可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上处理器（central processing units，CPU）501（例如，一个或一个以上处理器）和存储器 509，一个或一个以上存储应用程序 507 或数据 506 的存储介质 508（例如一个或一个以上海量存储设备）。其中，存储器 509 和存储介质 508 可以是短暂存储或持久存储。存储在存储介质 508 的程序可以包括一个或一个以上模块（图示没标出），每个模块可以包括对基于人工智能的字向量定义设备中的一系列指令操作。更进一步地，处理器 501 可以设置为与存储介质 508 通信，在基于人工智能的字向量定义设备 500 上执行存储介质 508 中的一系列指令操作。

基于人工智能的字向量定义设备 500 还可以包括一个或一个以上电源 502，一个或一个以上有线或无线网络接口 503，一个或一个以上输入输出接口 504，和/或，一个或一个以上操作系统 505，例如 Windows Serve，Mac OS X，Unix，Linux，FreeBSD 等等。本领域技术人员可以理解，图 5 中示出的基于人工智能的字向量定义设备结构并不构成对基于人工智能的字向量定义设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。处理器 501 可以执行上述实施例中获取单元 301、分词单元 302、第一生成单元 303、第二生成单元 304 和第三生成单元 305 的功能。

下面结合图 5 对基于人工智能的字向量定义设备的各个构成部件进行具体的介绍：

处理器 501 是基于人工智能的字向量定义设备的控制中心，可以按照设置的基于人工智能的字向量定义方法进行处理。处理器 501 利用各种接口和线路连接整个基于人工智能的字向量定义设备的各个部分，通过运行或执行存储在存储器 509 内的软件程序和/或模块，以及调用存储在存储器 509 内的数据，执行基于人工智能的字向量定义设备的各种功

能和处理数据，将报文域中的不可读标签转换为可读标签，从而实现对报文中的应用场景的快速识别。存储介质 508 和存储器 509 都是存储数据的载体，本申请实施例中，存储介质 508 可以是指储存容量较小，但速度快的内存储器，而存储器 509 可以是储存容量大，但储存速度慢的外存储器。

存储器 509 可用于存储软件程序以及模块，处理器 501 通过运行存储在存储器 509 的软件程序以及模块，从而执行基于人工智能的字向量定义设备 500 的各种功能应用以及数据处理。存储器 509 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（比如根据多个目标词向量和预置的权重策略生成目标字向量）等；存储数据区可存储根据基于人工智能的字向量定义设备的使用所创建的数据（比如多个目标词向量）等。此外，存储器 509 可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在本申请实施例中提供的基于人工智能的字向量定义方法程序和接收到的数据流存储在存储器中，当需要使用时，处理器 501 从存储器 509 中调用。

本申请还提供一种非易失性计算机可读存储介质，包括指令，当所述指令在计算机上运行时，使得计算机执行如下基于人工智能的字向量定义方法的步骤：

获取目标文本，所述目标文本包括需要进行分词的中文语句；

对所述目标文本进行分词得到多个词语；

根据所述多个词语生成对应的多个目标词向量；

根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字。

在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线（例如同轴电缆、光纤、双绞线）或无线（例如红外、无线、微波等）方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存储的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质，（例如，软盘、硬盘、磁带）、光介质（例如，光盘）、或者半导体介质（例如固态硬盘（solid state disk，

SSD））等。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

权 利 要 求

1、一种基于人工智能的字向量定义方法，包括：

获取目标文本，所述目标文本包括需要进行分词的中文语句；

对所述目标文本进行分词得到多个词语；

根据所述多个词语生成对应的多个目标词向量；

根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字。

2、根据权利要求 1 所述的基于人工智能的字向量定义方法，所述根据所述多个词语生成对应的多个目标词向量包括：

将所述多个词语输入到预置的算法模型中；

将每个词语映射成 K 维词语向量，其中，K 为大于 0 的整数；

计算得到每个词语向量之间的距离；

根据所述每个词语向量之间的距离确定每个词语向量之间的语义相似度；

根据所述每个词语向量之间的语义相似度确定与所述多个词语中目标词语的语义相似度最高的向量为目标词向量；

确定多个目标词向量，每个目标词向量对应一个词语。

3、根据权利要求 2 所述的基于人工智能的字向量定义方法，在所述将所述多个词语输入到预置的算法模型中之前，所述方法还包括：

随机生成一个词向量矩阵，每一个行对应一个单词的向量；

在所述词向量矩阵确定一个目标单词，并从所述词向量矩阵中提取周边单词的词向量；

计算所述周边单词的词向量的均值向量；

将所述均值向量输入到预置的逻辑回归模型中进行训练；

生成预置的算法模型，所述预置的算法模型输出的概率向量与所述目标单词的 one-hot 编码向量相匹配。

4、根据权利要求 2 所述的基于人工智能的字向量定义方法，所述计算得到每个词语向量之间的距离包括：

在所述词语向量中确定第一向量和第二向量；

计算所述第一向量和所述第二向量之间的余弦值，满足公式：

$$Sim(D_1,D_2)=\cos\theta=\frac{\sum_{k=1}^{n}W_{1k}W_{2k}}{\sqrt{(\sum_{k=1}^{n}W_{1k}^2)(\sum_{k=1}^{n}W_{2k}^2)}}$$，其中，D_1、D_2分别表示第一向量和第二向量，W_{1k}表示第一向量的权重，W_{2k}表示第二向量的权重，第一向量和第二向量都包括 N 个特征值，$1\le k\le N$；

将所述余弦值确定为所述第一向量和所述第二向量的距离。

5、根据权利要求 1 所述的基于人工智能的字向量定义方法，所述根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字包括：

获取预置的权重策略，所述预置的权重策略包含每个词向量的权重值；

确定所述多个目标词向量中每个词向量的目标权重值；

根据所述多个目标词向量和所述多个目标词向量中各个目标权重值生成目标字向量。

6、根据权利要求 1-5 中任一所述的基于人工智能的字向量定义方法，所述根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字之后，所述方法还包括：

将每个目标字向量作为模型参数输入到长短时记忆网络 LSTM 和条件随机场 CRF 模型中，对所述中文语句进行实体识别生成预测分词。

7、根据权利要求 6 所述的基于人工智能的字向量定义方法，所述将每个目标字向量作为模型参数输入到长短时记忆网络 LSTM 和条件随机场 CRF 模型中，对所述中文语句进行实体识别生成预测分词包括：

将每个目标字向量作为模型参数输入到 LSTM 和 CRF 模型中；

确定每个目标字向量在字空间中的位置；

根据所述每个目标字向量在字空间中的位置对所述中文语句进行实体识别，生成预测分词。

8、一种基于人工智能的字向量定义装置，包括：

获取单元，用于获取目标文本，所述目标文本包括需要进行分词的中文语句；

分词单元，用于对所述目标文本进行分词得到多个词语；

第一生成单元，用于根据所述多个词语生成对应的多个目标词向量；

第二生成单元，用于根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字。

9、根据权利要求 8 所述的基于人工智能的字向量定义装置，第一生成单元包括：

第一输入模块，用于将所述多个词语输入到预置的算法模型中；

映射模块，用于将每个词语映射成 K 维词语向量，其中，K 为大于 0 的整数；

第一计算模块，用于计算得到每个词语向量之间的距离；

第一确定模块，用于根据所述每个词语向量之间的距离确定每个词语向量之间的语义相似度；

第二确定模块，用于根据所述每个词语向量之间的语义相似度确定与所述多个词语中目标词语的语义相似度最高的向量为目标词向量；

第三确定模块，用于确定多个目标词向量，每个目标词向量对应一个词语。

10、根据权利要求 9 所述的基于人工智能的字向量定义装置，第一生成单元还包括：

第一生成模块，用于随机生成一个词向量矩阵，每一个行对应一个单词的向量；

提取模块，用于在所述词向量矩阵确定一个目标单词，并从所述词向量矩阵中提取周边单词的词向量；

第二计算模块，用于计算所述周边单词的词向量的均值向量；

训练模块，用于将所述均值向量输入到预置的逻辑回归模型中进行训练；

第二生成模块，用于生成预置的算法模型，所述预置的算法模型输出的概率向量与所述目标单词的 one-hot 编码向量相匹配。

11、根据权利要求 9 所述的基于人工智能的字向量定义装置，第一计算模块具体用于：

在所述词语向量中确定第一向量和第二向量；

计算所述第一向量和所述第二向量之间的余弦值，满足公式：

$$Sim(D_1,D_2)=\cos\theta=\frac{\sum_{k=1}^{n}W_{1k}W_{2k}}{\sqrt{(\sum_{k=1}^{n}W_{1k}^2)(\sum_{k=1}^{n}W_{2k}^2)}}$$，其中，D_1、D_2分别表示第一向量和第二向量，W_{1k}表示第一向量的权重，W_{2k}表示第二向量的权重，第一向量和第二向量都包括 N 个特征值，$1\le k\le N$；

将所述余弦值确定为所述第一向量和所述第二向量的距离。

12、根据权利要求 8 所述的基于人工智能的字向量定义装置，第二生成单元具体用于：

获取预置的权重策略，所述预置的权重策略包含每个词向量的权重值；

确定所述多个目标词向量中每个词向量的目标权重值；

根据所述多个目标词向量和所述多个目标词向量中各个目标权重值生成目标字向量。

13、根据权利要求 8-12 中任一所述的基于人工智能的字向量定义装置，基于人工智能的字向量定义装置还包括：

第三生成单元，用于将每个目标字向量作为模型参数输入到长短时记忆网络 LSTM 和条件随机场 CRF 模型中，对所述中文语句进行实体识别生成预测分词。

14、根据权利要求 13 所述的基于人工智能的字向量定义装置，第三生成单元包括：

第二输入模块，用于将每个目标字向量作为模型参数输入到 LSTM 和 CRF 模型中；

第四确定模块，用于确定每个目标字向量在字空间中的位置；

识别生成模块，用于根据所述每个目标字向量在字空间中的位置对所述中文语句进行实体识别，生成预测分词。

15、一种基于人工智能的字向量定义设备，其特征在于，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如下步骤：

获取目标文本，所述目标文本包括需要进行分词的中文语句；

对所述目标文本进行分词得到多个词语；

根据所述多个词语生成对应的多个目标词向量；

根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字。

16、根据权利要求 15 所述的基于人工智能的字向量定义设备，所述处理器具体用于执行以下步骤：

将所述多个词语输入到预置的算法模型中；

将每个词语映射成 K 维词语向量，其中，K 为大于 0 的整数；

计算得到每个词语向量之间的距离；

根据所述每个词语向量之间的距离确定每个词语向量之间的语义相似度；

根据所述每个词语向量之间的语义相似度确定与所述多个词语中目标词语的语义相似度最高的向量为目标词向量；

确定多个目标词向量，每个目标词向量对应一个词语。

17、根据权利要求 16 所述的基于人工智能的字向量定义设备，所述处理器具体用于执行以下步骤：

随机生成一个词向量矩阵，每一个行对应一个单词的向量；

在所述词向量矩阵确定一个目标单词，并从所述词向量矩阵中提取周边单词的词向量；

计算所述周边单词的词向量的均值向量；

将所述均值向量输入到预置的逻辑回归模型中进行训练；

生成预置的算法模型，所述预置的算法模型输出的概率向量与所述目标单词的 one-hot 编码向量相匹配。

18、根据权利要求 16 所述的基于人工智能的字向量定义设备，所述处理器具体用于执行以下步骤：

在所述词语向量中确定第一向量和第二向量；

计算所述第一向量和所述第二向量之间的余弦值，满足公式：

$Sim(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{n} W_{1k} W_{2k}}{\sqrt{(\sum_{k=1}^{n} W_{1k}^2)(\sum_{k=1}^{n} W_{2k}^2)}}$，其中，$D_1$、$D_2$分别表示第一向量和第二向量，$W_{1k}$表示第一向量的权重，$W_{2k}$表示第二向量的权重，第一向量和第二向量都包括 N 个特征值，$1 \le k \le N$；

将所述余弦值确定为所述第一向量和所述第二向量的距离。

19、根据权利要求 15 所述的基于人工智能的字向量定义设备，所述处理器具体用于执行以下步骤：

获取预置的权重策略，所述预置的权重策略包含每个词向量的权重值；

确定所述多个目标词向量中每个词向量的目标权重值；

根据所述多个目标词向量和所述多个目标词向量中各个目标权重值生成目标字向量。

20、一种非易失性计算机可读存储介质，包括指令，当所述指令在计算机上运行时，使得计算机执行如下步骤：

获取目标文本，所述目标文本包括需要进行分词的中文语句；

对所述目标文本进行分词得到多个词语；

根据所述多个词语生成对应的多个目标词向量；

根据所述多个目标词向量和预置的权重策略生成目标字向量，所述多个目标词向量对应的多个词语中每个词语都包含所述目标字向量对应的字。

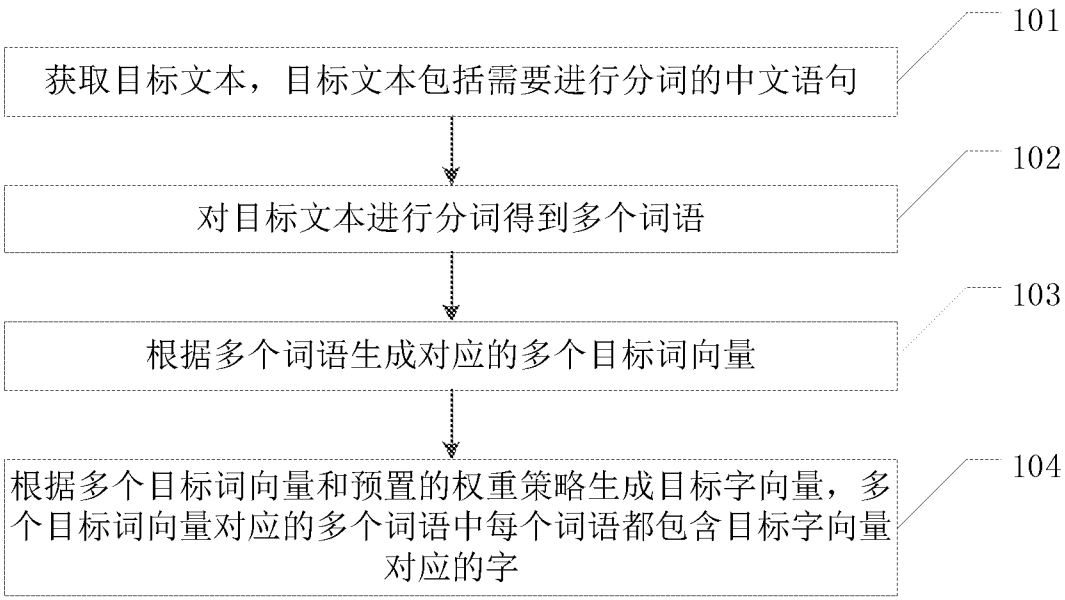
101
获取目标文本，目标文本包括需要进行分词的中文语句
102
对目标文本进行分词得到多个词语
103
根据多个词语生成对应的多个目标词向量
104
根据多个目标词向量和预置的权重策略生成目标字向量，多个目标词向量对应的多个词语中每个词语都包含目标字向量对应的字

图 1

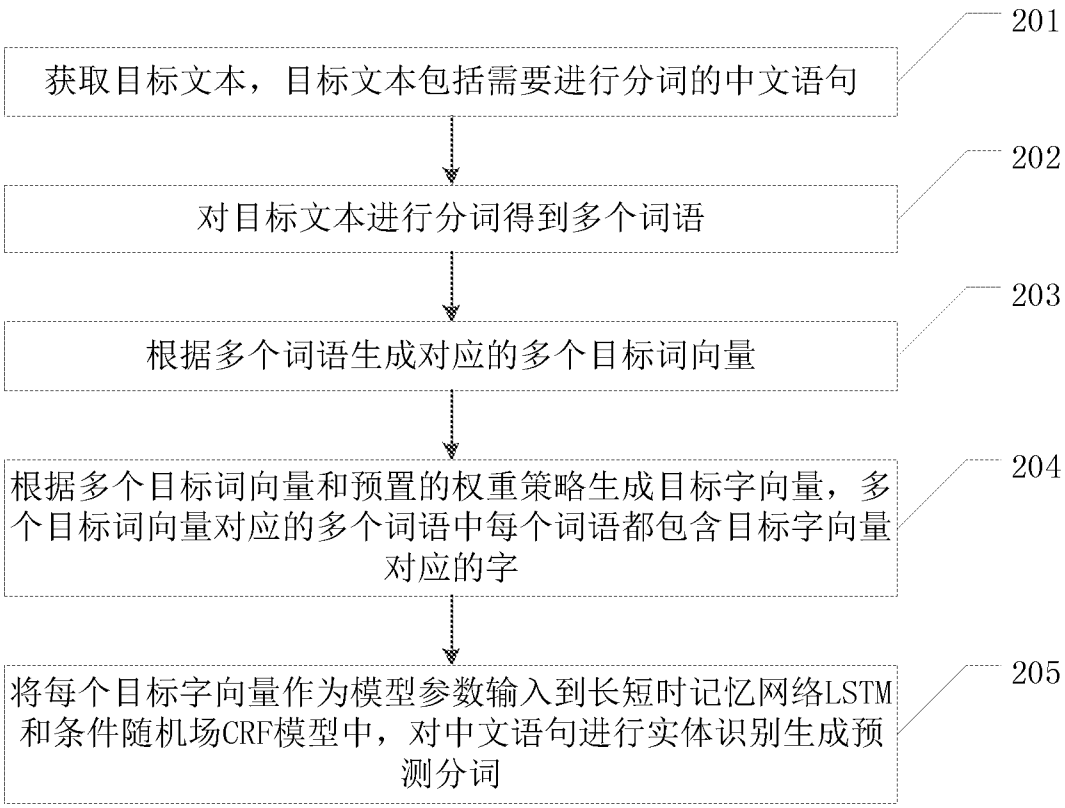
201
获取目标文本，目标文本包括需要进行分词的中文语句
202
对目标文本进行分词得到多个词语
203
根据多个词语生成对应的多个目标词向量
204
根据多个目标词向量和预置的权重策略生成目标字向量，多个目标词向量对应的多个词语中每个词语都包含目标字向量对应的字
205
将每个目标字向量作为模型参数输入到长短时记忆网络LSTM和条件随机场CRF模型中，对中文语句进行实体识别生成预测分词

图 2

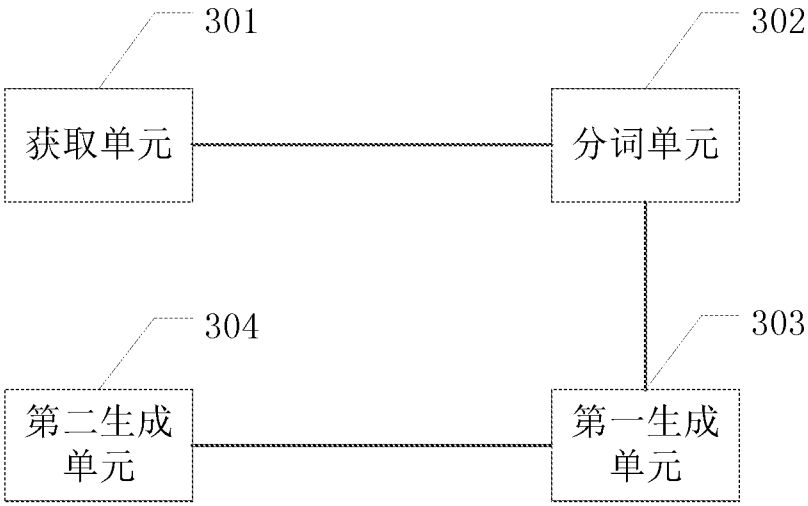

图 3

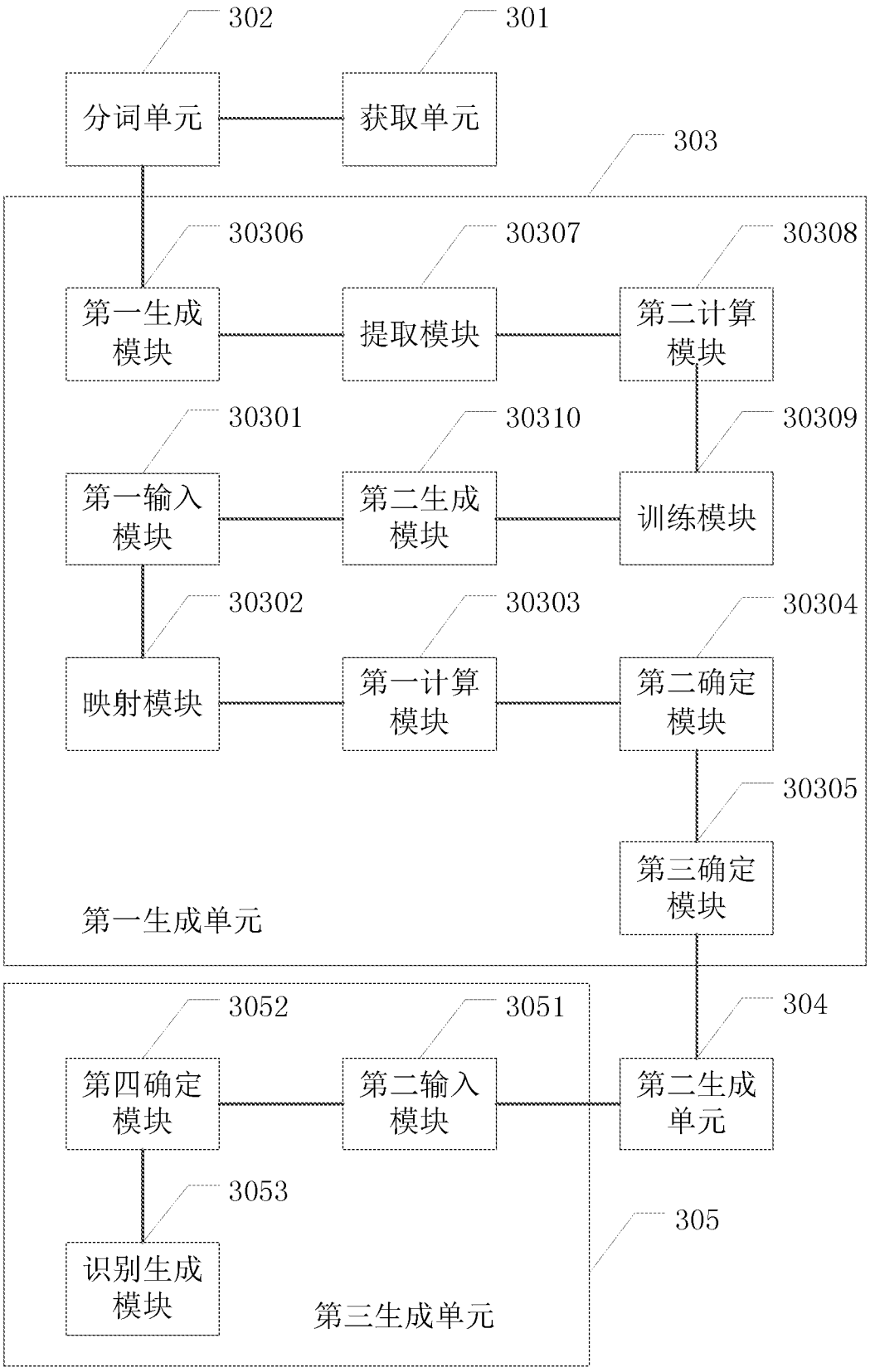
302
301
分词单元
获取单元
303
30306
30307
30308
第一生成模块
提取模块
第二计算模块
30301
30310
30309
第一输入模块
第二生成模块
训练模块
30302
30303
30304
映射模块
第一计算模块
第二确定模块
30305
第三确定模块
第一生成单元
3052
3051
304
第四确定模块
第二输入模块
第二生成单元
3053
305
识别生成模块
第三生成单元

图 4

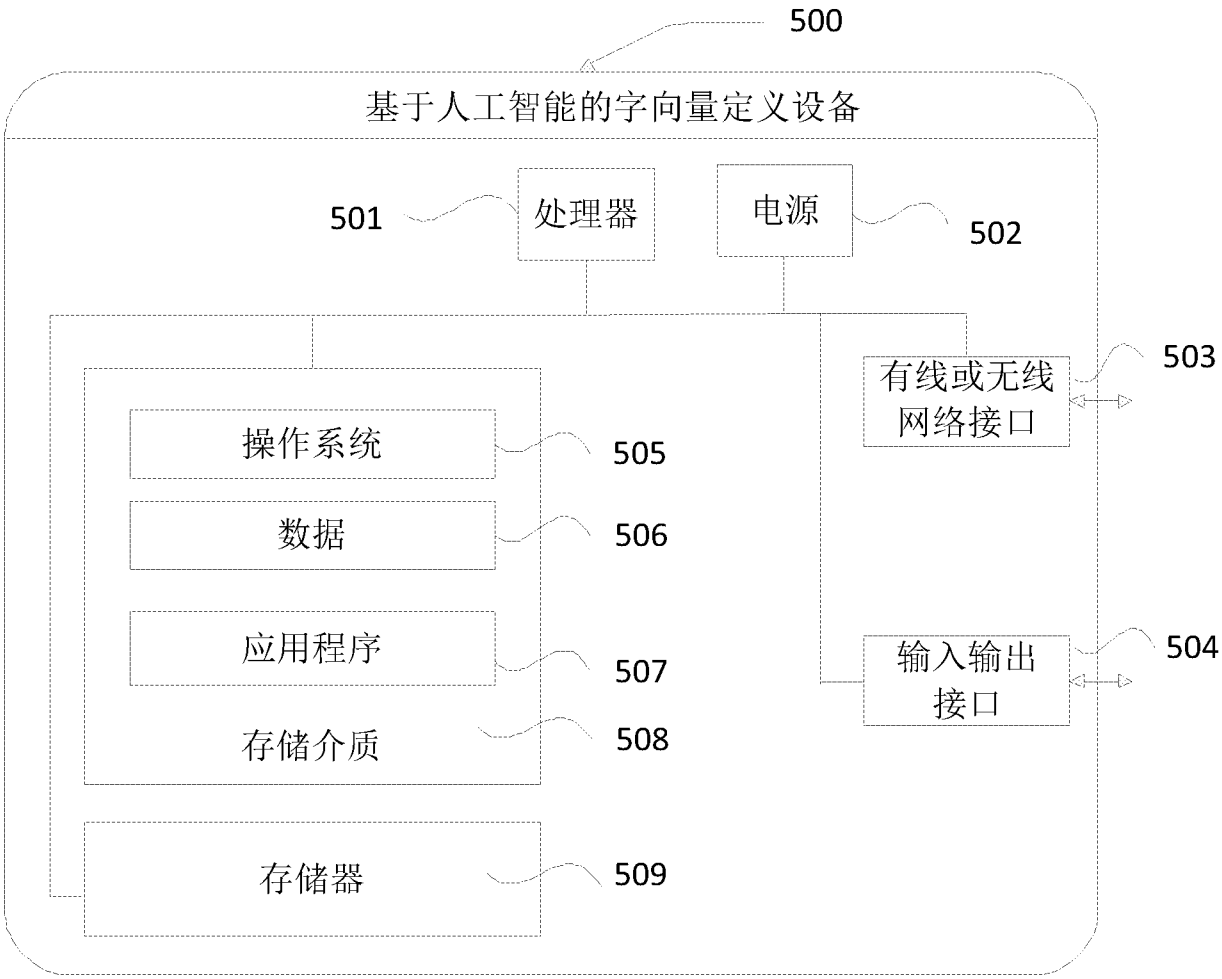
500
基于人工智能的字向量定义设备
501
处理器
电源
502
503
有线或无线
网络接口
操作系统
505
数据
506
应用程序
507
存储介质
508
输入输出
接口
504
存储器
509

图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/102462

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/20(2020.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F, G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNKI, CNPAT, GOOGLE, BAIDU: 概率, 匹配, 最小单元, 最小结构, 语义单元, 字, 加权, 权重, 目标, 词向量, 字向量, 词语, 均值, 平均, 分词, 矩阵, probability, match, unit, minimum, minimal, smallest, word, character, weight, weigh, target, vector, average, dismember, partition, division, matrix

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	李伟康 等 (LI, Weikang et al.). "深度学习中汉语字向量和词向量结合方式探究 (Combination Methods of Chinese Character and Word Embeddings in Deep Learning)" *中文信息学报 (Journal of Chinese Information Processing)*, Vol. 31, No. 6, 30 November 2017 (2017-11-30), section 3	1-20
Y	CN 107688604 A (ALIBABA GROUP HOLDING LIMITED) 13 February 2018 (2018-02-13) claims 2-5	1-20
Y	CN 109063035 A (HARBIN INSTITUTE OF TECHNOLOGY) 21 December 2018 (2018-12-21) claim 2	3, 10, 17
A	CN 107273355 A (DALIAN UNIVERSITY OF TECHNOLOGY) 20 October 2017 (2017-10-20) entire document	1-20
A	US 2017091318 A1 (KABUSHIKI KAISHA TOSHIBA) 30 March 2017 (2017-03-30) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
26 February 2020	**04 March 2020**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/102462

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 107688604 A	13 February 2018	HK 1248866 A0	19 October 2018
CN 109063035 A	21 December 2018	None	
CN 107273355 A	20 October 2017	None	
US 2017091318 A1	30 March 2017	CN 106557460 A	05 April 2017
		JP 2017068833 A	06 April 2017

国际检索报告

国际申请号
PCT/CN2019/102462

A. 主题的分类

G06F 40/20(2020.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F, G06K

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

WPI, EPODOC, CNKI, CNPAT, GOOGLE, BAIDU:概率，匹配，最小单元，最小结构，语义单元，字，加权，权重，目标，词向量，字向量，词语，均值，平均，分词，矩阵，probability, match, unit, minimum, minimal, smallest, word, character, weight, weigh, target, vector, average, dismember, partition, division, matrix

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	李伟康 等. "深度学习中汉语字向量和词向量结合方式探究" 中文信息学报, 第31卷，第6期，2017年 11月 30日 (2017-11-30)， 第3节	1-20
Y	CN 107688604 A (阿里巴巴集团控股有限公司) 2018年 2月 13日 (2018-02-13) 权利要求2-5	1-20
Y	CN 109063035 A (哈尔滨工业大学) 2018年 12月 21日 (2018-12-21) 权利要求2	3， 10， 17
A	CN 107273355 A (大连理工大学) 2017年 10月 20日 (2017-10-20) 全文	1-20
A	US 2017091318 A1 (KABUSHIKI KAISHA TOSHIBA) 2017年 3月 30日 (2017-03-30) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2020年 2月 26日	2020年 3月 4日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 李萍 电话号码 53961602

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/102462

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	107688604	A	2018年 2月 13日	HK	1248866	A0	2018年 10月 19日
CN	109063035	A	2018年 12月 21日		无		
CN	107273355	A	2017年 10月 20日		无		
US	2017091318	A1	2017年 3月 30日	CN	106557460	A	2017年 4月 5日
				JP	2017068833	A	2017年 4月 6日