

(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 7 月 2 日 (02.07.2020)



(10) 国际公布号
WO 2020/133358 A1

(51) 国际专利分类号:
G06F 17/28 (2006.01)

(21) 国际申请号: PCT/CN2018/125358

(22) 国际申请日: 2018 年 12 月 29 日 (29.12.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 深圳市优必选科技有限公司 (UBTECH ROBOTICS CORP) [CN/CN]; 中国广东省深圳市南山区学苑大道 1001 号南山智园 C1 栋 16、22 楼, Guangdong 518000 (CN)。

(72) 发明人: 熊友军 (XIONG, Youjun); 中国广东省深圳市南山区学苑大道 1001 号南山智园 C1 栋 16、22 楼, Guangdong 518000 (CN)。 熊为星 (XIONG,

Weixing); 中国广东省深圳市南山区学苑大道 1001 号南山智园 C1 栋 16、22 楼, Guangdong 518000 (CN)。 廖洪涛 (LIAO, Hongtao); 中国广东省深圳市南山区学苑大道 1001 号南山智园 C1 栋 16、22 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳中细软知识产权代理有限公司 (SHENZHEN CIPRUN INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国广东省深圳市南山区粤海街道粤兴四道 1 号中山大学深圳产学研大楼 11 楼, Guangdong 518000 (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, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,

(54) Title: CHAT CORPUS CLEANING METHOD, APPARATUS, COMPUTER DEVICE AND STORAGE MEDIUM

(54) 发明名称: 聊天语料的清洗方法、装置、计算机设备和存储介质

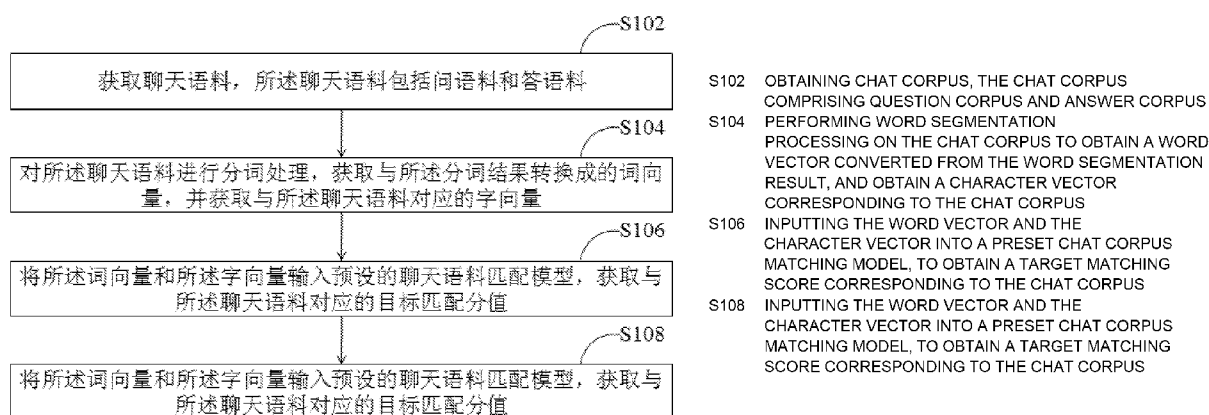


图 1

(57) Abstract: Provided are a chat corpus cleaning method, an apparatus, a computer device and a storage medium, comprising: obtaining chat corpus, the chat corpus comprising question corpus and answer corpus (S102); performing word segmentation processing on the chat corpus to obtain a word vector converted from the word segmentation result, and obtain a character vector corresponding to the chat corpus (S104); inputting the word vector and the character vector into a preset chat corpus matching model, to obtain a target matching score corresponding to the chat corpus (S106); and cleaning the chat corpus according to the target matching score. Through the above manner, the chat corpus can be automatically cleaned to improve the quality of the chat corpus, thereby improving the accuracy of subsequent model training.

(57) 摘要: 一种聊天语料的清洗方法、装置、计算机设备及存储介质, 包括: 获取聊天语料, 所述聊天语料包括问语料和答语料 (S102); 对所述聊天语料进行分词处理, 获取与所述分词结果转换成的词向量, 并获取与所述聊天语料对应的字向量 (S104); 将所述词向量和所述字向量输入预设的聊天语料匹配模型, 获取与所述聊天语料对应的目标匹配分值 (S106); 根据所述目标匹配分值对所述聊天语料进行清洗。通过上述方式, 能够自动对聊天语料进行清洗, 提高聊天语料的质量, 从而提高后续模型训练的准确性。

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))。

聊天语料的清洗方法、装置、计算机设备和存储介质

技术领域

本发明涉及计算机技术领域和深度学习技术领域，尤其涉及一种聊天语料的清洗方法、装置、计算机设备和存储介质。

背景技术

智能机器人聊天一直是人工领域的主要研究方向，如何通过深度学习等方法使得智能聊天机器人像人一样自如的进行聊天，例如，在产品售后部门中作为智能客服。在目前的智能聊天机器人的训练过程中，不管是检索式还是生成式，均需要闲聊语料来机器人进行训练。

对智能聊天机器人进行问答训练需要大量的闲聊语料，目前大量的闲聊语料来自于网上的开源材料，但是这些闲聊语料普遍存在质量不高的情况，需要对这些闲聊语料进行清洗。而采用人工筛选的方式需要专业的人员对闲聊语料进行标注，不仅耗费人力、效率低下，还可能因为标注人员的水平和理解的不同导致结果的准确性不足，从而导致最终训练语料的质量不高。

发明内容

基于此，有必要针对上述问题，提出一种对聊天语料清洗效率高的聊天语料的清洗方法、装置、计算机设备和存储介质。

在本发明的第一方面，提供了一种聊天语料的清洗方法，所述方法包括：
获取聊天语料，所述聊天语料包括问语料和答语料；

对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并获取与所述聊天语料对应的字向量；

将所述词向量和所述字向量输入预设的聊天语料匹配模型，获取与所述聊天语料对应的目标匹配分值；

根据所述目标匹配分值对所述聊天语料进行清洗。

在本发明的第二方面，还提供了一种聊天语料的清洗装置，包括：

聊天语料获取模块，用于获取聊天语料，所述聊天语料包括问语料和答语料；

聊天语料处理模块，用于对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并获取与所述聊天语料对应的字向量；

模型计算模块，用于将所述词向量和所述字向量输入预设的聊天语料匹配模型，获取与所述聊天语料对应的目标匹配分值；

语料清洗模块，用于根据所述目标匹配分值对所述聊天语料进行清洗。

在本发明的第三方面，提供了一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述计算机程序被所述处理器执行时，使得所述处理器执行以下步骤：

获取聊天语料，所述聊天语料包括问语料和答语料；

对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并获取与所述聊天语料对应的字向量；

将所述词向量和所述字向量输入预设的聊天语料匹配模型，获取与所述聊天语料对应的目标匹配分值；

根据所述目标匹配分值对所述聊天语料进行清洗。

在本发明的第四方面，提供了一种计算机可读存储介质，存储有计算机程序，所述计算机程序被处理器执行时，使得所述处理器执行以下步骤：

获取聊天语料，所述聊天语料包括问语料和答语料；

对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并获取与所述聊天语料对应的字向量；

将所述词向量和所述字向量输入预设的聊天语料匹配模型，获取与所述聊天语料对应的目标匹配分值；

根据所述目标匹配分值对所述聊天语料进行清洗。

实施本发明实施例，将具有如下有益效果：

本发明提出了一种聊天语料的清洗方法、装置、计算机设备和存储介质，

首先获取待清洗的聊天语料，每一条聊天语料包含了对应的问题和回复，对问题、回复对应的语料进行处理，分别转换成对应的词向量以及字向量，然后根据训练好的聊天语料匹配模型计算问题、回复对应的目标匹配分值，从而来判断当前聊天语料之间是否是匹配的，是否需要进行清洗。也就是说，对于原始

5 获取的聊天语料，可以根据聊天语料匹配模型进行自动的清洗，不再需要人工逐条聊天语料进行标注，省去了大量的人工操作时间，在一定程度上减少了成本花销。并且，采用上述聊天语料的清洗方法，避免了人工操作的认为错误，也在一定程度上提高了聊天语料清洗的准确性。

进一步的，在本实施例中，在对聊天语料匹配模型进行训练以及计算聊天

10 语料之间的目标匹配分值的过程中，同时考虑了聊天语料对应的词向量和自向量，在最大程度上保留了词向量与字向量的特征，提高了在聊天语料匹配模型中特征提取的有效性，从而提高了聊天语料清洗的准确性。

附图说明

15 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

其中：

- 20 图 1 为一个实施例中聊天语料的清洗方法的实现流程示意图；
图 2 为一个实施例中聊天语料的清洗方法的实现流程示意图；
图 3 为一个实施例中聊天语料匹配模型训练的实现流程示意图；
图 4 为一个实施例中问答对语料构建示意图；
图 5 为一个实施例中聊天语料匹配模型示意图；
25 图 6 为一个实施例中聊天语料的清洗方法的实现流程示意图；
图 7 为一个实施例中聊天语料的清洗装置的结构框图；
图 8 为一个实施例中计算机设备的结构框图。

具体实施方式

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、完整地描述，显然，所描述的实施例仅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作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。

如图 1 所示，在一个实施例中，提供了一种聊天语料的清洗方法，本发明实施例所述的聊天语料的清洗方法的执行主体可以是服务器，当然本发明实施例所述的聊天语料的清洗方法的执行主体还可以是其他终端设备，例如，机器人设备。

具体的，如图 1 所示，上述聊天语料的清洗方法包括如下步骤：

步骤 S102：获取聊天语料，所述聊天语料包括问语料和答语料。

聊天语料为从网络或其他途径获取的未经清洗的闲聊语料，其中，每条聊天语料包括一个问句（问语料）和一个答句（答语料）。例如，相应的聊天语料为若干个问答对，如（问题 1，回复 1），（问题 2，回复 2），……

需要说明的是，在本实施例中，在对聊天语料进行具体的清洗之前，还需要对聊天语料进行预处理，主要是针对原始的聊天语料中可能存在的不规范性，如，去掉反复重复的标点符号（如，一个问句后出现大量的问号，在这种情况下，仅保留其中的一个），再例如，去掉含有包情包的聊天语料，去掉聊天语料中含有的空格，过滤到敏感信息（如政治敏感词及色情暴力等词）。也就是说，经过针对聊天语料的预处理之后，可以去掉部分质量不高的聊天语料，提高后续聊天语料清洗的效率和准确性。

进一步的，在本实施例中，对聊天语料还需要进一步的改写，例如，去标点、取空格、英文大小写转换、去停用词等，以去掉对语义理解无关的字符，避免对后续聊天语料的清洗过程的准确性的影响。

步骤 S104：对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并获取与所述聊天语料对应的字向量。

在本实施例中,对聊天语料进行处理获取对应的词向量或字向量的过程中,既需要获取与聊天语料对应的词向量,也需要获取与聊天语料对应的字向量,在后续的模式匹配计算的过程中,既考虑词向量的特性,也保留字向量的特性,从而提高聊天语料清洗的准确性。

5 具体的,获取与聊天语料对应的词向量的过程具体如下:对聊天语料中的问语料或答语料进行分词处理,然后对于已经分词处理完毕的聊天语料,进一步的将每一个词转换成对应的词向量。

其中,采用随机初始化词向量的方式正态分布随机初始化为 300 维的向量,然后根据分词结果将每一个切割的词转换成对应的 300 维的词向量。

10 获取聊天语料对应的字向量的过程具体如下:对聊天语料中的问语料和答语料按字进行切割(如果是英文,按照字符进行切割),然后将每一个字/字符转换成对应的字向量。

其中,采用随机初始化字向量的方式正态分布随机初始化为 300 维的向量,然后将每一个切割的字或字符转换成对应的 300 维的字向量。

15 在本实施例中,因为中文、英文或者长度的问题,可能导致不同的词向量或字向量的长度不一致,为了后续向量或矩阵计算的方便,在本实施例中,还需要按照预设的长度阈值对相应的词向量和字向量进行长度的改写。即,按照预设的长度阈值,对所有的词向量或字向量进行截断或补齐操作。例如,设定一个用户咨询问题的最大长度,利用该长度值对聊天语料对应的词向量和字向量进行截断或补齐操作,对聊天语料进行规范化处理,得到规范化之后的问语料(对应的词向量、字向量)以及答语料(对应的词向量、字向量)。

20

在一个具体的实施例中,令 q_{1w}, q_{2w} 为聊天语料中的问语料和问语料对应的词向量, q_{1c}, q_{2c} 为聊天语料中的问语料和问语料对应的字向量,即:

$$q_{1w} = (x_{1w}, x_{2w}, x_{3w}, \dots, x_{mw})$$

$$q_{2w} = (y_{1w}, y_{2w}, y_{3w}, \dots, y_{nw})$$

$$q_{1c} = (x_{1c}, x_{2c}, x_{3c}, \dots, x_{pc})$$

25

$$q_{2c} = (y_{1c}, y_{2c}, y_{3c}, \dots, y_{qc})$$

其中 m 、 n 为问语料和问语料的词向量长度，此处 $m=n$ ， p 、 q 为问语料和问语料的字向量的长度，此处 $p=q$ 。

步骤 S106：将所述词向量和所述字向量输入预设的聊天语料匹配模型，获取与
5 所述聊天语料对应的目标匹配分值。

步骤 S108：根据所述目标匹配分值对所述聊天语料进行清洗。

在本实施例中，聊天语料匹配模型为根据深度学习模型构建的对聊天语料之间是否匹配进行评估预测的模型，输入为聊天语料中的问语料对应的词/字向量、答语料对应的词/字向量，输出为该问语料与答语料之间的目标匹配分值。

10 在本步骤中，获取聊天语料对应的目标匹配分值之后，即可根据该目标匹配分值进行清洗。例如，如图 2 所示，在目标匹配分值大于或等于预设的匹配阈值的情况下，确定聊天语料需要进行保留，反之，在目标匹配分值小于预设的匹配阈值的情况下，确定将该聊天语料丢弃。

具体的，在一个具体的实施例中，通过聊天语料匹配模型获取聊天语料对
15 应的目标匹配分值的过程具体如下：

步骤 S1062：按照预设的叉乘函数分别对所述聊天语料对应的词向量和字向量进行叉乘处理，根据预设的映射函数获取叉乘处理结果的预设数量的映射向量，所述映射向量包括映射词向量和映射字向量；按照预设的融合算法对所述映射词向量和映射字向量进行融合处理，对融合后的结果进行特征提取，
20 获取与所述聊天语料对应的第一目标匹配分值；

步骤 S1064：分别对所述聊天语料对应的词向量和字向量进行特征提取，按照预设的融合算法对特征提取后的词向量和字向量进行融合操作，将融合结果输入预设的投影层，获取与所述聊天语料对应的第二目标匹配分值；

步骤 S1066：按照预设的匹配拼接算法，根据所述第一目标匹配分值和所述
25 第二目标匹配分值，计算与所述聊天语料对应的目标匹配分值。

具体的，步骤 S1062 中：

$$z_w^{(0)} = f(\text{relu}(q_{1w} \otimes q_{2w}))$$

$$z_c^{(0)} = f(\text{relu}(q_{1c} \otimes q_{2c}))$$

其中, $q_{1w} \otimes q_{2w}$ 为对 q_{1w}, q_{2w} 进行叉乘处理, relu 为预设的激活函数进行激活, f 表示映射函数 (Mapping 函数), 在这里 Mapping 函数为挑选出叉乘之后的前 K 个值 (TopK, 例如, $K=10$ 或 $K=30$)。从而获取相应的词向量/字向量进行叉乘处理之后获取出关键的前 K 个值, 且输入问题也转换成固定长度的问题或回

复。

然后将 TopK 后词向量和字向量进行合并和融合, 即为:

$$z^{(0)} = (z_w^{(0)} \oplus z_c^{(0)})$$

在将词向量与字向量融合之后, 对融合的结果进行特征提取:

$$z^{(l)} = \text{relu}(W^{(l-1)} z^{(l-1)} + b^{(l-1)}), \quad l=1,2,\dots,L$$

其中, $W^{(l-1)}, b^{(l-1)}$ 为投影中对应层的权重参数矩阵及偏置向量, 是通过模型训练得到的; 并且, 通过预设的激活函数进行激活, 输出即为与所述聊天语料对应的第一目标匹配分值。

步骤 S1064 中:

首先对聊天语料对应的词向量、字向量进行特征提取:

$$g_w = \text{relu}(Uq_{1w} + b_w)$$

$$g_v = \text{relu}(Vq_{1v} + b_v)$$

其中 U, V 分别为词向量及字向量的权重矩阵, b_w, b_v 为词向量及字向量的偏置向量, 并且采用激活函数 relu 进行激活。

在激活之后, 按照预设的融合算法对特征提取后的词向量和字向量进行融合操作:

$$g = (g_w \oplus g_v)$$

在融合之后对融合结果进行投影:

$$h = \text{relu}(W^{(h)} g + b^{(h)})$$

其中, $W^{(h)}, b^{(h)}$ 为投影层的权重矩阵及偏置向量。

也即，将融合结果输入一个投影层（project 层），输出即为与所述聊天语料对应的第二目标匹配分值。

在步骤 S1062 及 S1064 中，获取到了与聊天语料对应的第一目标匹配分值 $z^{(l)}$ 和第二目标匹配分值 h ，然后即可根据第一目标匹配分值和第二目标匹配分值计算整体的目标匹配分值 e 。

具体的，在步骤 S1066 中，按照预设的匹配拼接算法，根据所述第一目标匹配分值和所述第二目标匹配分值，计算与所述聊天语料对应的目标匹配分值；也即，根据公式

$$e = (z^{(L)} \oplus h)$$

10 计算步骤 S102 中的聊天语料对应的目标匹配分值 e 。

在本实施例中，可以根据过滤的程度或次数的不同，设定不同的匹配阈值，例如，第一次过滤情况下，可以将匹配阈值设置为 0.5，在经过多次清洗之后，可以将匹配阈值逐步提高，将最后一次或最终的清洗过程中的匹配阈值设置为 0.9。

15 进一步的，在本实施例中，还需要对上述聊天语料匹配模型进行模型训练、验证，然后再进行具体的聊天语料的清洗。

在一个具体的实施例中，如图 3 所示，上述聊天语料的清洗方法还包括如下步骤：

步骤 S202：获取训练语料，根据所述训练语料构建问答对话料；

20 步骤 S204：根据所述问答对话料对预设的聊天语料匹配模型进行训练，获取训练完成的聊天语料匹配模型。

训练语料为获取到的闲聊语料之后经过预处理之后的聊天语料，与前述聊天语料可以为同一语料。在本实施例中，在获取到训练语料之后，构建相应的问答对话料，该问答对话料为符合聊天语料匹配模型训练数据的形式。

25 具体的，通过以下方式构建正负样本对。如问题 1-回复 1，问题 2-回复 2，问题 3-回复 3，以上三条闲聊语料，可构成 6 个问答对，具体可如图 4 所示，为：（问题 1，回复 1，回复 2）、（问题 1，回复 1，回复 3）、（问题 2，回复 2，

回复 1)、(问题 2, 回复 2, 回复 3)、(问题 3, 回复 3, 回复 1)、(问题 3, 回复 3, 回复 2) 等问答对话料, 其中(问题 1, 回复 1, 回复 2) 该问答对表示为问题 1 与回复 1 的匹配度比问题 1 与回复 2 的匹配度高。

进一步的, 在本实施例中, 在构建训练语料时, 根据 8:1:1 的比例构建训练
5 样本、验证样本、测试样本, 从而完成整个聊天语料匹配模型的训练。

也就是说, 上述问答对话料包括训练问语料、第一答语料和第二答语料; 则对应的转换成词向量和字向量的方式与前述步骤 S102-S108 一致, 则相应的词向量包括问语料词向量、第一答语料词向量、第二答语料词向量, 相应的字向量包括问语料字向量、第一答语料字向量、第二答语料字向量。

10 将构建完成的训练语料对应的问答对话料, 输入如图 5 所示的聊天语料匹配模型中, 并获取与所述问语料、第一答语料对应的第一评估分值和与问语料、第二答语料匹配的第二评估分值。然后将第一评估分值、第二评估分值与真实结果(第一评估分值大于第二评估分值)进行比对, 从而完成对于模型的训练。

具体的, 具体的聊天语料匹配模型的计算过程如下:

15 令 q_{1w} 为聊天语料中的问语料对应的词向量, q_{2w} 为聊天语料中的第一答语料对应的词向量, q_{3w} 为聊天语料中的第二答语料对应的词向量; q_{1c} 为聊天语料中的问语料对应的字向量, q_{2c} 为聊天语料中的第一答语料对应的字向量, q_{3c} 为聊天语料中的第二答语料对应的字向量。

在对聊天语料匹配模型进行训练的过程中, 分别计算与所述问语料、第一
20 答语料对应的第一评估分值 $e(q_1, q_2)$ 和与问语料、第二答语料匹配的第二评估分值 $e(q_1, q_3)$ 。

然后按照预设的损失函数计算损失值 L , 具体可以使用 hinge-loss 损失函数:

$$L(q_1, q_2, q_3; \Theta) = \max(0, \text{margin} - s(q_1, q_2) + s(q_1, q_3))$$

其中 margin 为正反间样本相似间距(本实施例中可以将 margin 设为 1), $e(q_1, q_2)$
25 表示 q_1, q_2 输入到聊天语料匹配模型计算的结果值, $e(q_1, q_3)$ 为将 q_1, q_3 输入到聊天语料匹配模型计算的结果值, Θ 为当前的给定参数。

根据损失值进行梯度更新, 即可完成模型的训练, 为了加快模型训练的速

度我们选用了 Adam 算法来完成梯度的更新。最后，保存模型以及其中的参数，并对模型进行更新。

需要说明的是，在第一次模型训练时，其中的参数的具体值可以采用初始值，例如，可以通过正态分布随机初始化的参数进行计算，并且，在每一轮模型训练完毕之后，对模型中的参数进行更新和迭代，以进行下一步的聊天语料的清洗，即执行前述步骤 S102-S108。

在一个具体的实施例中，在通过步骤 S202-S204 对聊天语料匹配模型完成训练之后，即可根据该聊天语料匹配模型对聊天语料进行清洗，并且，对于经过第一轮清洗之后的聊天语料，可以进一步的作为聊天语料匹配模型的训练语料，并对聊天语料匹配模型再进行一轮模型训练，再据此对聊天语料进行再次的清洗。并且，在此循环过程中，不断的减少匹配阈值，例如，第一次清洗过程中，匹配阈值取 0.5，并按照匹配阈值为 0.5、0.6、0.7、0.8、0.9 逐步进行过滤，完成最终的清洗工作，并将最后一轮聊天语料经过聊天语料匹配模型输出的目标匹配分值与 0.9 进行相比，在目标匹配分值大于或等于 0.9 的情况下，保留相应的聊天语料，否则，将相应的聊天语料进行过滤。具体可参考图 6 所示。

在本实施例中，经过聊天语料匹配模型的不断的训练、语料清洗的循环，通过反复操作进行无监督的清洗聊天语料，大量的节省了人工清洗聊天语料的时间，并且能保证聊天语料清洗之后的质量，提高了后续对智能聊天机器人进行训练的准确性。

如图 7 所示，提供了一种聊天语料的清洗装置，具体包括：

聊天语料获取模块 102，用于获取聊天语料，所述聊天语料包括问语料和答语料；

聊天语料处理模块 104，用于对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并获取与所述聊天语料对应的字向量；

模型计算模块 106，用于将所述词向量和所述字向量输入预设的聊天语料匹配模型，获取与所述聊天语料对应的目标匹配分值；

语料清洗模块 108，用于根据所述目标匹配分值对所述聊天语料进行清洗。

上述聊天语料的清洗装置，首先获取待清洗的聊天语料，每一条聊天语料包含了对应的问题和回复，对问题、回复对应的语料进行处理，分别转换成对应的词向量以及字向量，然后根据训练好的聊天语料匹配模型计算问题、回复对应的目标匹配分值，从而来判断当前聊天语料之间是否是匹配的，是否需要
5 进行清洗。也就是说，对于原始获取的聊天语料，可以根据聊天语料匹配模型进行自动的清洗，不再需要人工逐条聊天语料进行标注，省去了大量的人工操作时间，在一定程度上减少了成本花销。并且，采用上述聊天语料的清洗方法，避免了人工操作的认为错误，也在一定程度上提高了聊天语料清洗的准确性。

10 进一步的，在本实施例中，在对聊天语料匹配模型进行训练以及计算聊天语料之间的目标匹配分值的过程中，同时考虑了聊天语料对应的词向量和自向量，在最大程度上保留了词向量与字向量的特征，提高了在聊天语料匹配模型中特征提取的有效性，从而提高了聊天语料清晰的准确性。

在其中一个实施例中，模型计算模块 106 还用于：按照预设的叉乘函数分
15 别对所述聊天语料对应的词向量和字向量进行叉乘处理，根据预设的映射函数获取叉乘处理结果的预设数量的映射向量，所述映射向量包括映射词向量和映射字向量；按照预设的融合算法对所述映射词向量和映射字向量进行融合处理，对融合后的结果进行特征提取，获取与所述聊天语料对应的第一目标匹配分值；分别对所述聊天语料对应的词向量和字向量进行特征提取，按照预设的融合算
20 法对特征提取后的词向量和字向量进行融合操作，将融合结果输入预设的投影层，获取与所述聊天语料对应的第二目标匹配分值；按照预设的匹配拼接算法，根据所述第一目标匹配分值和所述第二目标匹配分值，计算与所述聊天语料对应的目标匹配分值。

在其中一个实施例中，如图 7 所示，上述装置还包括向量改写模块 110，
25 用于按照预设的第一长度阈值对所述词向量进行长度改写；按照预设的第二长度阈值对所述字向量进行长度改写。

在其中一个实施例中，语料清洗模块 108 还用于判断所述目标匹配分值是

否大于或等于预设的匹配阈值；在所述目标匹配分值小于所述匹配阈值的情况下，对所述聊天语料进行清洗。

在其中一个实施例中，如图 7 所示，聊天语料的清洗装置还包括模型训练模块 112，用于：

5 获取训练语料，根据所述训练语料构建问答对话料；

根据所述问答对话料对预设的聊天语料匹配模型进行训练，获取训练完成的聊天语料匹配模型。

10 在其中一个实施例中，所述问答对话料包括训练问语料、第一答语料和第二答语料；所述词向量包括问语料词向量、第一答语料词向量、第二答语料词向量；模型训练模块 112 还用于根据预设的聊天语料匹配模型对所述问答对话料进行评估预测，得到与所述训练问语料、第一训练答语料对应的第一评估分值和与训练问语料、第二训练答语料匹配的第二评估分值；按照预设的损失函数以所述第一评估分值、第二评估分值为输入，输出对应的损失值；按照预设的迭代算法对所述损失值进行更新迭代，并对所述聊天语料匹配模型进行更新。

15 在其中一个实施例中，模型训练模块 112 还用于以所述清洗完成的聊天语料作为训练语料，对所述聊天语料匹配模型进行训练，获取所述训练完成的聊天语料匹配模型。

图 8 示出了一个实施例中计算机设备的内部结构图。该计算机设备可以是服务器，也可以是机器人。如图 8 所示，该计算机设备包括通过系统总线连接的处理器、存储器和网络接口。其中，存储器包括非易失性存储介质和内存储器。该计算机设备的非易失性存储介质存储有操作系统，还可存储有计算机程序，该计算机程序被处理器执行时，可使得处理器实现聊天语料的清洗方法。该内存储器中也可储存有计算机程序，该计算机程序被处理器执行时，可使得处理器执行聊天语料的清洗方法。本领域技术人员可以理解，图 8 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

在一个实施例中，本申请提供的聊天语料的清洗方法可以实现为一种计算机程序的形式，计算机程序可在如图 8 所示的计算机设备上运行。计算机设备的存储器中可存储组成聊天语料的清洗装置的各个程序模板。

一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，

5 所述计算机程序被所述处理器执行时，使得所述处理器执行如下步骤：

获取聊天语料，所述聊天语料包括问语料和答语料；

对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并获取与所述聊天语料对应的字向量；

10 将所述词向量和所述字向量输入预设的聊天语料匹配模型，获取与所述聊天语料对应的目标匹配分值；

根据所述目标匹配分值对所述聊天语料进行清洗。

在一个实施例中，提出了一种计算机可读存储介质，存储有计算机程序，所述计算机程序被处理器执行时，使得所述处理器执行以下步骤：

获取聊天语料，所述聊天语料包括问语料和答语料；

15 对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并获取与所述聊天语料对应的字向量；

将所述词向量和所述字向量输入预设的聊天语料匹配模型，获取与所述聊天语料对应的目标匹配分值；

根据所述目标匹配分值对所述聊天语料进行清洗。

20 需要说明的是，上述聊天语料的清洗方法、聊天语料的清洗装置、计算机设备及计算机可读存储介质属于一个总的发明构思，聊天语料的清洗方法、聊天语料的清洗装置、计算机设备及计算机可读存储介质实施例中的内容可相互适用。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，
25 是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一非易失性计算机可读存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据

库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程 ROM(PROM)、电可编程 ROM(EPROM)、电可擦除可编程 ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限，RAM 以多种形式可得，
5 诸如静态 RAM(SRAM)、动态 RAM(DRAM)、同步 DRAM(SDRAM)、双数据率 SDRAM(DDRSDRAM)、增强型 SDRAM(ESDRAM)、同步链路(Synchlink) DRAM(SLDRAM)、存储器总线(Rambus)直接 RAM(RDRAM)、直接存储器总线动态 RAM(DRDRAM)、以及存储器总线动态 RAM(RDRAM)等。

以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述
10 实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对本申请专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和
15 改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

权利要求

1、一种聊天语料的清洗方法，其特征在于，所述方法包括：

获取聊天语料，所述聊天语料包括问语料和答语料；

对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并

5 获取与所述聊天语料对应的字向量；

将所述词向量和所述字向量输入预设的聊天语料匹配模型，获取与所述聊天语料对应的目标匹配分值；

根据所述目标匹配分值对所述聊天语料进行清洗。

2、根据权利要求 1 所述的聊天语料的清洗方法，其特征在于，所述将所述
10 词向量和所述字向量输入预设的聊天语料匹配模型，获取与所述聊天语料对应的目标匹配分值，还包括：

按照预设的叉乘函数分别对所述聊天语料对应的词向量和字向量进行叉乘处理，根据预设的映射函数获取叉乘处理结果的预设数量的映射向量，所述映射向量包括映射词向量和映射字向量；按照预设的融合算法对所述映射词向量
15 和映射字向量进行融合处理，对融合后的结果进行特征提取，获取与所述聊天语料对应的第一目标匹配分值；

分别对所述聊天语料对应的词向量和字向量进行特征提取，按照预设的融合算法对特征提取后的词向量和字向量进行融合操作，将融合结果输入预设的投影层，获取与所述聊天语料对应的第二目标匹配分值；

20 按照预设的匹配拼接算法，根据所述第一目标匹配分值和所述第二目标匹配分值，计算与所述聊天语料对应的目标匹配分值。

3、根据权利要求 1 所述的聊天语料的清洗方法，其特征在于，所述对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并获取与所述聊天语料对应的字向量之后，还包括：

25 按照预设的第一长度阈值对所述词向量进行长度改写；

按照预设的第二长度阈值对所述字向量进行长度改写。

4、根据权利要求 1 所述的聊天语料的清洗方法，其特征在于，所述根据所

述目标匹配分值对所述聊天语料进行清洗，还包括：

判断所述目标匹配分值是否大于或等于预设的匹配阈值；

在所述目标匹配分值小于所述匹配阈值的情况下，对所述聊天语料进行清洗。

5 5、根据权利要求1所述的聊天语料的清洗方法，其特征在于，所述方法还包括：

获取训练语料，根据所述训练语料构建问答对语料；

根据所述问答对语料对预设的聊天语料匹配模型进行训练，获取训练完成的聊天语料匹配模型。

10 6、根据权利要求5所述的聊天语料的清洗方法，其特征在于，所述问答对语料包括训练问语料、第一训练答语料和第二训练答语料；

所述根据所述问答对语料对预设的聊天语料匹配模型进行训练，还包括：

根据预设的聊天语料匹配模型对所述问答对语料进行评估预测，得到与所述训练问语料、第一训练答语料对应的第一评估分值和与训练问语料、第二训练答语料匹配的第二评估分值；

15 按照预设的损失函数以所述第一评估分值、第二评估分值为输入，输出对应的损失值；

按照预设的迭代算法对所述损失值进行更新迭代，并对所述聊天语料匹配模型进行更新。

20 7、根据权利要求5所述的聊天语料的清洗方法，其特征在于，所述根据所述目标匹配分值对所述聊天语料进行清洗之后，还包括：

以所述清洗完成的聊天语料作为训练语料，对所述聊天语料匹配模型进行训练，获取所述训练完成的聊天语料匹配模型。

8、一种聊天语料的清洗装置，其特征在于，所述装置包括：

25 聊天语料获取模块，用于获取聊天语料，所述聊天语料包括问语料和答语料；

聊天语料处理模块，用于对所述聊天语料进行分词处理，获取与所述分词结果转换成的词向量，并获取与所述聊天语料对应的字向量；

模型计算模块，用于将所述词向量和所述字向量输入预设的聊天语料匹配模型，获取与所述聊天语料对应的目标匹配分值；

5 语料清洗模块，用于根据所述目标匹配分值对所述聊天语料进行清洗。

9、一种计算机设备，其特征在于，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求 1 至 7 任一项所述聊天语料的清洗方法的步骤。

10 10、一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求 1 至 7 任一项所述聊天语料的清洗方法的步骤。

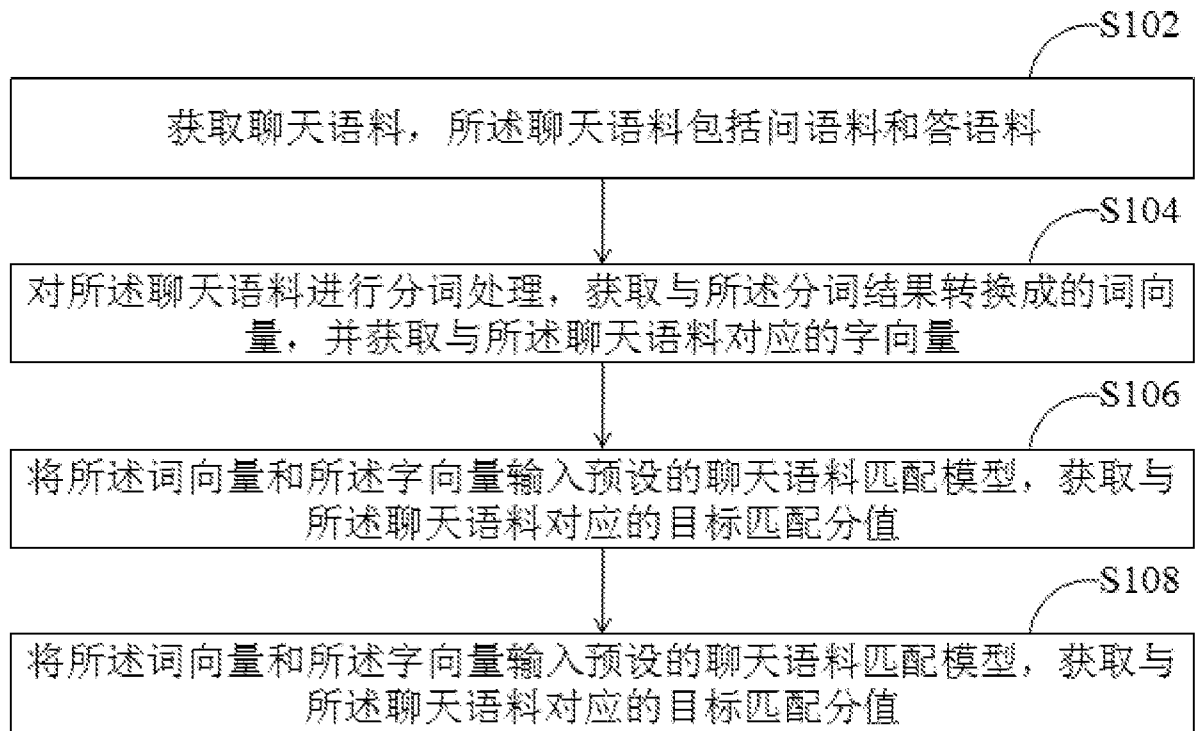


图 1

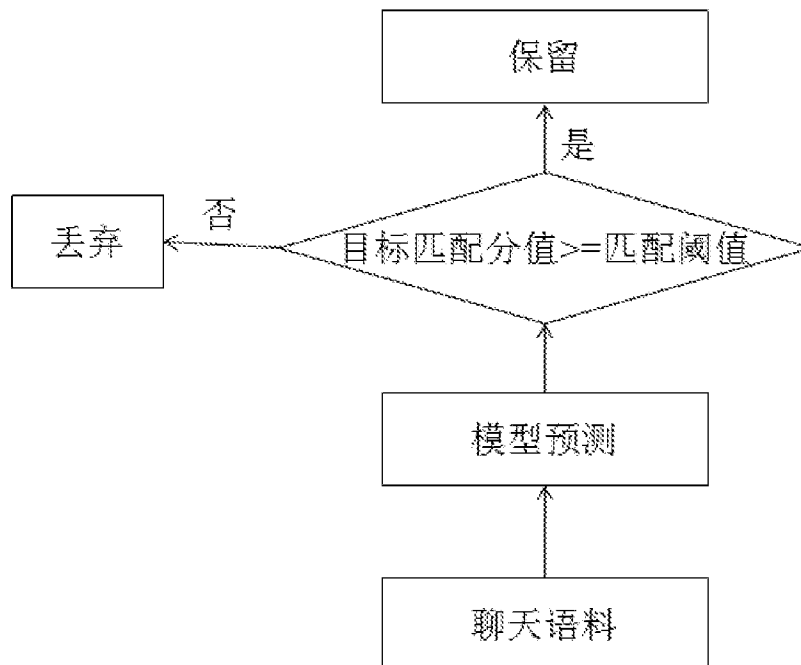


图 2

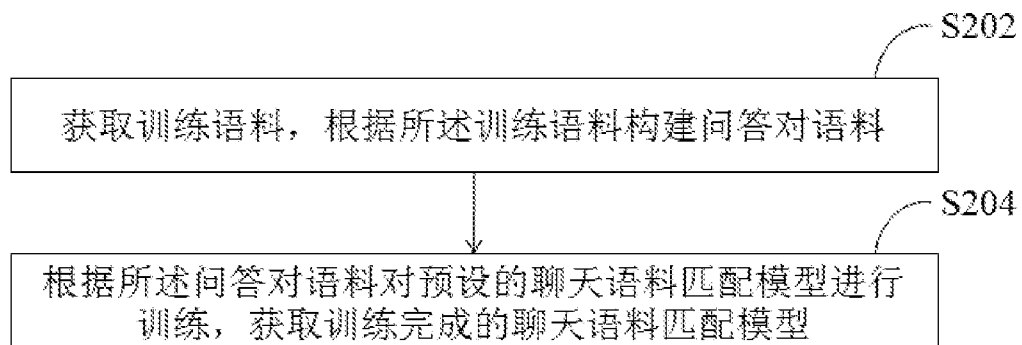


图 3

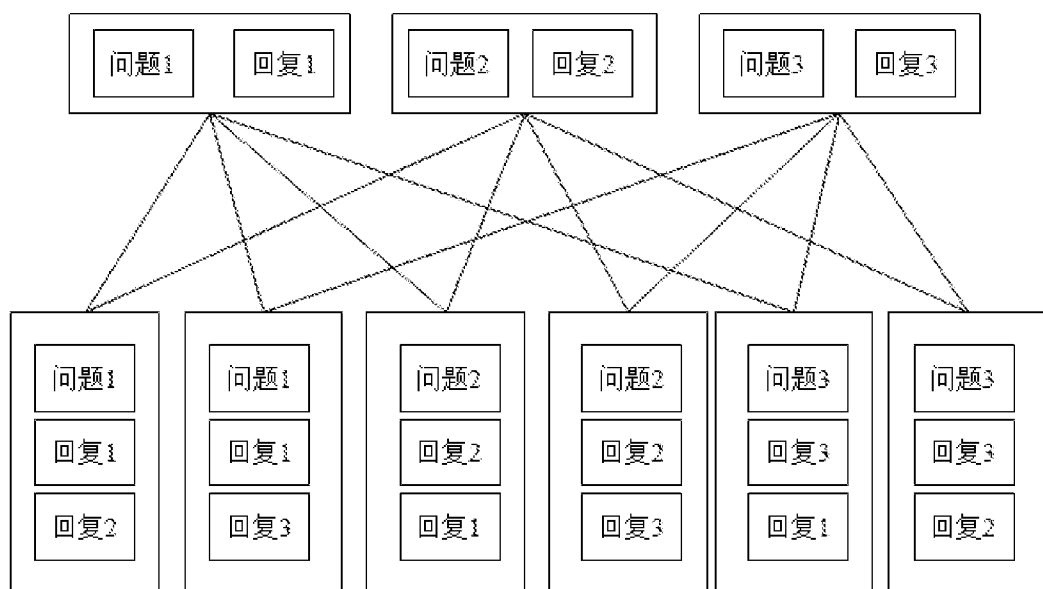


图 4

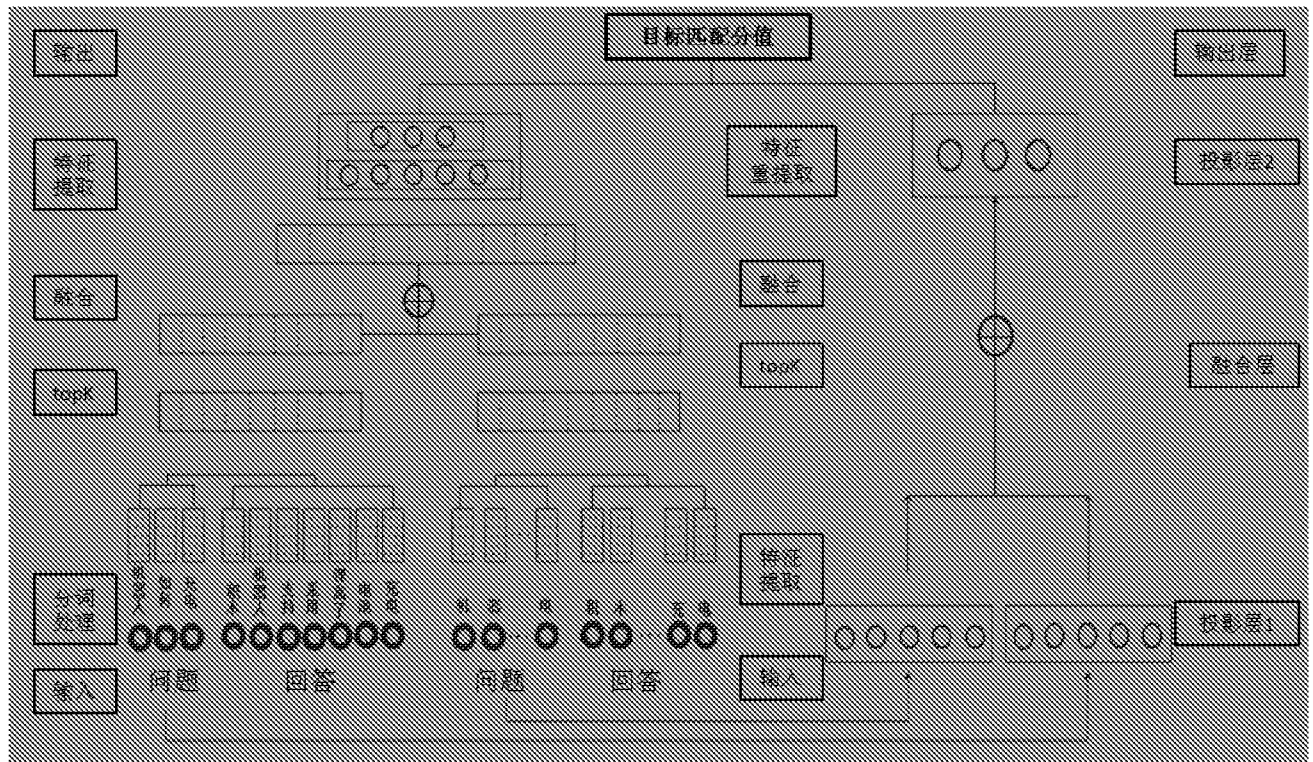


图 5

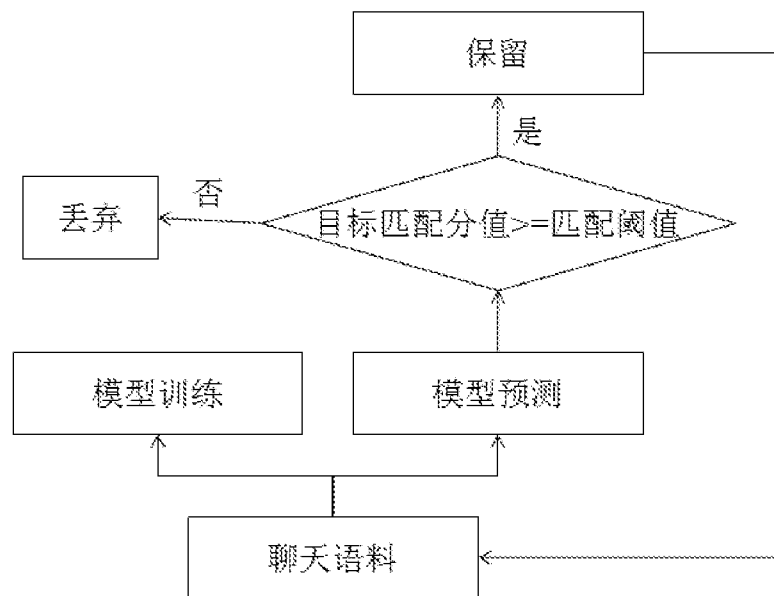


图 6

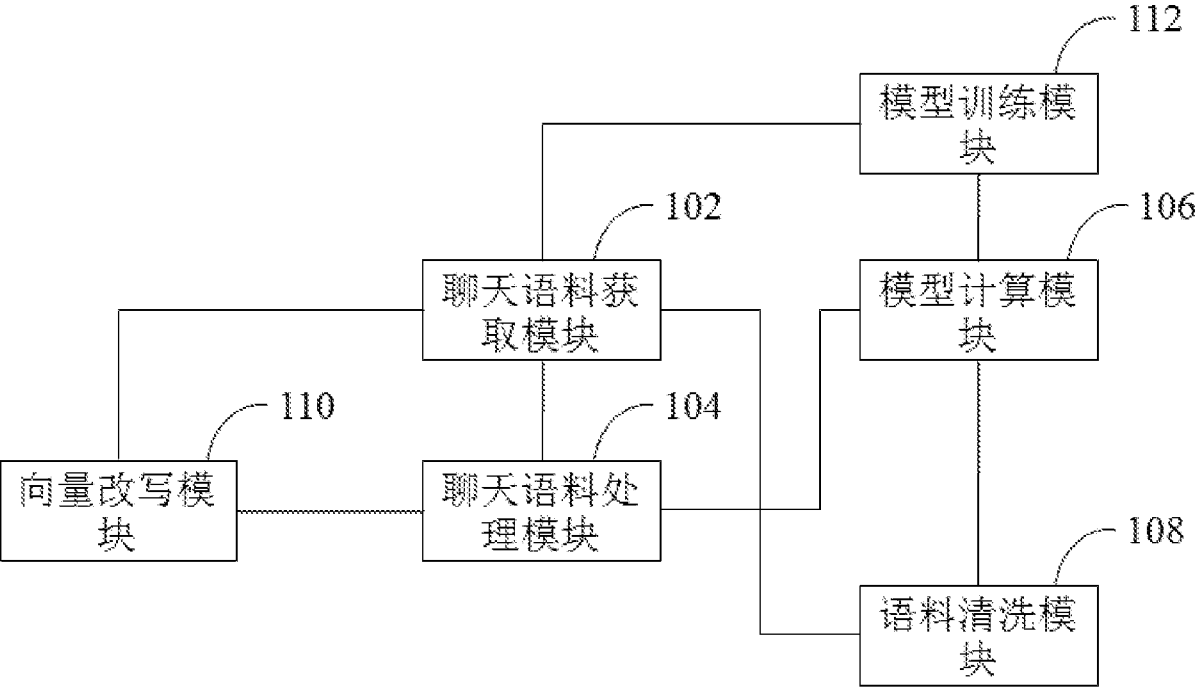


图 7

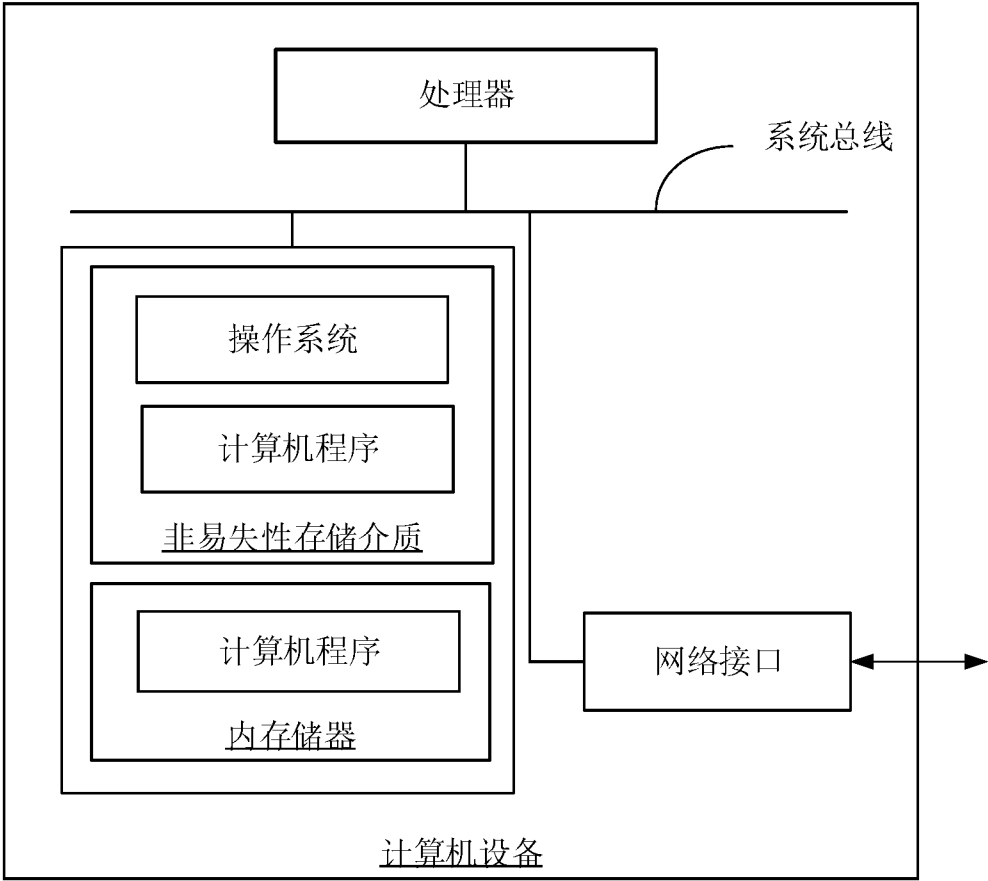


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/125358

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/28(2006.01)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

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)

CNPAT, EPODOC, WPI, CNKI, IEEE: 字向量, 语料, 清洗, 词向量, 匹配, 神经网络, 分值, character, word, vector, corpus, clean, match, neural network, score

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 108170853 A (GUANGDONG HATVEST TECHNOLOGY DEVELOPMENT CO., LTD.) 15 June 2018 (2018-06-15) description, paragraphs [0005]-[0039], [0126] and [0127]	1-10
A	CN 107833629 A (XIAMEN UNIVERSITY) 23 March 2018 (2018-03-23) entire document	1-10
A	CN 108766581 A (SHANGHAI PUBLIC HEALTH CLINICAL CENTER) 06 November 2018 (2018-11-06) entire document	1-10
A	US 2016026634 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 28 January 2016 (2016-01-28) entire document	1-10



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

12 September 2019

Date of mailing of the international search report

27 September 2019

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/CN2018/125358

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	108170853	A	15 June 2018	None			
CN	107833629	A	23 March 2018	None			
CN	108766581	A	06 November 2018	None			
US	2016026634	A1	28 January 2016	US	2018005117	A1	04 January 2018
				US	9754207	B2	05 September 2017
				US	10169706	B2	01 January 2019

国际检索报告

国际申请号

PCT/CN2018/125358

A. 主题的分类

G06F 17/28 (2006.01) 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

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))

CNPAT, EPDOC, WPI, CNKI, IEEE: 字向量, 语料, 清洗, 词向量, 匹配, 神经网络, 分值, character, word, vector, corpus, clean, match, neural network, score

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
X	CN 108170853 A (广东惠禾科技发展有限公司) 2018年 6月 15日 (2018 - 06 - 15) 说明书第[0005]-[0039]、[0126]-[0127]段	1-10
A	CN 107833629 A (厦门大学) 2018年 3月 23日 (2018 - 03 - 23) 全文	1-10
A	CN 108766581 A (上海市公共卫生临床中心) 2018年 11月 6日 (2018 - 11 - 06) 全文	1-10
A	US 2016026634 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2016年 1月 28日 (2016 - 01 - 28) 全文	1-10

☐ 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

2019年 9月 12日

国际检索报告邮寄日期

2019年 9月 27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

马丽莉

电话号码 86-(10)-53961409

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/125358

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	108170853	A	2018年 6月 15日	无			
CN	107833629	A	2018年 3月 23日	无			
CN	108766581	A	2018年 11月 6日	无			
US	2016026634	A1	2016年 1月 28日	US	2018005117	A1	2018年 1月 4日
				US	9754207	B2	2017年 9月 5日
				US	10169706	B2	2019年 1月 1日

表 PCT/ISA/210 (同族专利附件) (2015年1月)