(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 101599071 B
(45)授权公告日 2012.04.18

(21)申请号 200910063114.X

(22)申请日 2009.07.10

(73)专利权人 华中科技大学
地址 430074 湖北省武汉市洪山区珞瑜路1037号

(72)发明人 黄本雄 黄毅青 胡广 温杰

(74)专利代理机构 北京市德权律师事务所 11302
代理人 王建国

(51)Int.Cl.
G06F 17/30(2006.01)

审查员 王静

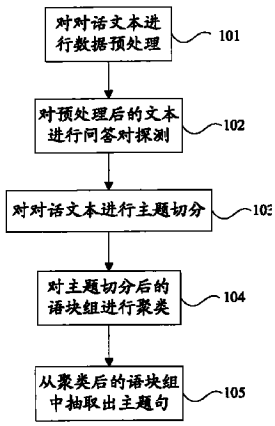
权利要求书 2 页 说明书 8 页 附图 4 页

(54)发明名称

对话文本主题的自动提取方法

(57)摘要

本发明公开了一种对话文本主题的自动提取方法，包括：对对话文本进行数据预处理，对预处理后的对话文本进行问答对探测；对所述对话文本进行主题切分，并对主题切分后的语块组进行聚类，从聚类后的语块组中抽取出主题句。采用本发明方法提取的对话文本主题更为准确，用户可以从提取出来的主题句中检索或发现感兴趣的对话记录，提高用户的体验。



CN 101599071 B

1. 一种对话文本主题的自动提取方法，其特征在于，包括：

对对话文本进行切词处理、词性标注、二次切分处理以及停用词处理，对预处理后的对话文本进行问答对探测；所述问答对探测具体包括：探测出对话文本中的问句；通过问句在对话文本中的位置，将两个问句之间的陈述语句列为答句候选集；在答句候选集中探测出对话文本中的每个问句相对应的答句；

对所述对话文本进行主题切分，并确定使用的聚类算法，根据相似性函数生成主题线索树，从而对主题切分后的语块组进行聚类，从聚类后的语块组中抽取出主题句；所述主题切分具体包括：将对话语句集作为输入，通过隐含语义概率模型获取词汇在对话文本中各个对话语句中的概率分布；根据所述概率分布，获取相邻句子间的语义相似度；比较各个相邻句子间的语义相似度和预设定的阈值范围，判定相邻的两个句子间是否为不同主题的切分点。

2. 根据权利要求 1 所述的方法，其特征在于，所述探测出对话文本中的问句具体包括：

选择识别问句的特征；

对准备用于训练集的句子进行人工手动标识句子类别；

基于所述选择的识别问句的特征，对用做训练集的句子提取出代表各个特征的值，记录下每个句子对应的特征值序列；

将训练集的每个句子的特征值序列和人工标识的句子类别共同作为分类器的输入，对分类器进行训练；

对准备用于测试集的句子进行人工手动标识句子类别；

根据所述训练集句子特征值的提取方法，记录下代表测试集中每个句子的特征值序列；

将测试集中抽取出的特征值序列和人工标识的句子类别共同作为分类器的输入，对分类器输出的分类结果的准确率进行评估，从而对选定的训练集、分类器和特征做相应的调整；

根据所述训练集句子特征值的提取方法，记录下代表待处理对话文本中每个句子的特征值序列；

将待处理对话文本抽取出的特征值序列作为分类器的输入，得到输出的分类结果。

3. 根据权利要求 2 所述的方法，其特征在于，所述识别问句的特征具体包括：

问句的高标识特征、输入的对话语句中词的个数，及句子中最前面的五个词的词性和句子中最后面的五个词的词性。

4. 根据权利要求 2 或 3 所述的方法，其特征在于，所述探测对话文本中的每个问句相对应的答句的方法具体包括：

选择识别答句的特征；

对准备用于训练集的句子进行人工手动标识句子类别；

从选定的训练集对话语句中抽取出代表答句特征的特征值序列；

将训练集中每个对话语句代表答句特征的特征值序列和人工标识的句子类别一同作为分类器的输入，对分类器进行训练；

对作为测试集的句子进行人工手动标识句子类别；

从作为测试集的对话语句中抽取出代表答句特征的特征值序列；

将测试集抽取出的特征值序列和人工标识的句子类别共同作为分类器的输入，对分类器输出的分类结果的准确率进行评估，从而对选定的训练集、分类器和特征做相应的调整；

根据所述训练集句子特征值的提取方法，记录下代表待处理对话文本中每个句子的特征值序列；

将待处理对话文本抽取出的特征值序列作为分类器的输入，得到输出的分类结果。

将探测出的每个问句及其相对应的答句合并到同一个对话语句，并进行标记。

5. 根据权利要求 4 所述的方法，其特征在于，所述答句的特征具体包括：

答句候选集中前五个词的词性标注和后五个词的词性标注；

答句候选集中的句子个数；

答句候选集中的答句与问句的距离；

答句候选集中的答句与问句的相似度。

6. 根据权利要求 5 所述的方法，其特征在于，所述生成主题线索树的方法具体包括：

将已进行主题切分的语块按照时间顺序进行排列；

第一个语块内容 Seg1 形成树的根节点，同时也形成树 T_1；

获取第二个语块内容 Seg2 与第一个树 T_1 的相似度 $Sim(Seg2, T_1)$，若 $Sim(Seg2, T_1) >$ 预定门限值 k，将 Seg2 加入树 T_1；否则，语块内容 Seg2 新建一个树 T_2；

获取第三个语块内容 Seg3 与前两棵树的相似度 $Sim(Seg3, T_1)$ 和 $Sim(Seg3, T_2)$，若 $Sim(Seg3, T_1) < Sim(Seg3, T_2)$ 且 $Sim(Seg3, T_2) >$ 预定门限值 k，则将语块内容 Seg3 加入树 T_2；若 $Sim(Seg3, T_1) < Sim(Seg3, T_2)$ 且 $Sim(Seg3, T_2) <$ 预定门限值 k，则由第三个语块内容 Seg3 新建一个树 T_3；并按相同方法处理对话文本中的所有语块；

其中，假定存在两个语块是 Seg_i 和 Seg_j，融入的语言特征用条件概率表示就是 $P(T(Seg_i, Seg_j) | Seg_i PPL, Seg_j PPF)$；对于给与的两个语块 Seg_i 和 Seg_j，定义一个函数 $T(Seg_i, Seg_j)$：$T(Seg_i, Seg_j) = \begin{cases} 1 \\ 0 \end{cases}$，

则获取语块与树之间的相似性函数为：

$$Sim(Seg, T) = \max_{i=1}^{m} \cos(Seg, Seg_i) * P(T(Seg_i, Seg_j) | Seg_i PPL, Seg_j PPF)$$。

7. 根据权利要求 6 所述的方法，其特征在于，所述抽取出主题句具体包括：

确定每个主题单元提取主题句的个数；

获取句子在主题单元中的贡献度；假设每个主题单元里含有 s 个句子，则获取主题线索树中当前句子 k 对主题单元的贡献度的方法为：$C_i = \sum_{i=1}^{S} |Sim(S_k, S_i) - 1|$

根据所述每个句子在主题单元中的贡献度，按照由大到小的顺序进行排序，取排名靠前的 N_i 个句子作为主题句；其中，$N_i = [\frac{N_{itree}}{3}]$，$N_{itree}$ 表示第 i 个主题线索树中包含的节点个数；N_i 表示的是第 i 个主题单元中需要提取的主题句个数；

将每个主题单元中合并的问答句提取出来，作为主题句。

对话文本主题的自动提取方法

技术领域

[0001] 本发明涉及计算机及通信技术领域，尤其涉及一种对话文本主题的自动提取方法。

背景技术

[0002] 网络通讯如今已成为了人们日常沟通的重要方式，为人们的交流提供了巨大的便利。同时，即时通信软件、网络留言板、电子邮件、网络会议等交流方式生成了大量的网络信息数据，这些数据与网页类型的数据有着本质的区别，它们以对话模式存在，其内容中蕴含着两个或多个参与者的观点和态度。因此网络对话数据中含有丰富的信息，能够给人们的工作和学习带来很大的帮助。例如，可以用于协助警察侦查疑犯的想法和行动，帮助心理医生了解病人的思考方式和辅助人类学家探究人类的行为模式等。但在海量数据中寻找有用数据需要相当大量的人力和时间，研究者希望结合计算机人工智能领域的一些方法，在海量对话数据中高效准确地获取重要的信息，因此基于对话文本的主题提取成为了近年来关注的热点。

[0003] 对话文本作为一种全新的信息资源，属于自然语言处理范畴。早期研究者们认为对对话文本的主题提取可以由普通文本的主题提取方法过渡而来。然而由于其在语言上的特点，用在普通文本的主题提取方法对对话文本发挥不了较好的效果。普通文本一般由一个作者编写，是具有逻辑合理、思维缜密、措辞得当、语句通顺、上下文联系紧密和主题脉络清晰等特点的书面语；对话一般由两个或多个参与者共同完成，是具有指代不明、语句缺省、大量问答句式存在和主题脉络混乱特点的口语。对于两种语言特点差异很大的语料，不能将普通文本的主题提取方法直接应用于对话文本的主题提取。

[0004] 目前，国内外针对对话文本的主题提取方法包括：

[0005] 1、基于机器学习的主题提取方法。机器学习的方法对选取特征集、训练集大小等都有一定的要求，需要多次测试比较，选择合适的模型、特征集、训练样本等。

[0006] 2、基于语义理解的主题提取方法。先提取出对话文本的句子中的名词或动词，依赖于 WordNet 知识库，找出它们在知识库里对应的概念集，计算句子间的语义相似度，在此基础上对对话文本中的句子进行排序，从而将排名靠前的句子视为主题句。该方法依赖于 WordNet 有一定的局限性，WordNet 中的词语毕竟也是有限的。特别是针对对话文本，其中包含的大部分是口语词汇，WordNet 很难全部囊括。

[0007] 3、融合语义和机器学习的主题提取方法。选取一些语义特征、词网、语料结构特征和词频等作为特征，从训练集中提取这些特征放入模型进行训练。

[0008] 4、基于统计的主题提取方法。将用于书面语文本的主题提取方法 tf*idf 统计方法做一些扩展，用于对话文本的主题提取。对对话文本中的词汇进行统计，从而对词进行评分，提取代表主题的词。这种简单的统计方法适合处理实时对话信息，其处理的速度较快。

[0009] 5、基于知识理解的主题提取方法。基于一个限定领域的知识理解系统对文本的语义进行“理解”，从而生成主题句。其应用于对话文本的主题提取的不足之处在于有领域限

制，而网络上的对话文本是开放领域的，需要人工编制大量的知识理解系统，可行性不高。

[0010] 但由于网络通讯对话文本的特点，对话中语句之间的词语相似度比较低，口语词汇很多，主题交织且组织结构混乱，导致应用以上几种方法提取出的主题词准确度不高。

发明内容

[0011] 有鉴于此，本发明的目的在于提供一种对话文本主题的自动提取方法，用于在对话文本中实现对话主题的自动提取。

[0012] 本发明的实施例提供了一种对话文本主题的自动提取方法，包括：

[0013] 对对话文本进行切词处理、词性标注、二次切分处理以及停用词处理，对预处理后的对话文本进行问答对探测；所述问答对探测具体包括：探测出对话文本中的问句；通过问句在对话文本中的位置，将两个问句之间的陈述语句列为答句候选集；在答句候选集中探测出对话文本中的每个问句相对应的答句；

[0014] 对所述对话文本进行主题切分，并确定使用的聚类算法，根据相似性函数生成主题线索树，从而对主题切分后的语块组进行聚类，从聚类后的语块组中抽取出主题句；所述主题切分具体包括：将对话语句集作为输入，通过隐含语义概率模型获取词汇在对话文本中各个对话语句中的概率分布；根据所述概率分布，获取相邻句子间的语义相似度；比较各个相邻句子间的语义相似度和预设定的阈值范围，判定相邻的两个句子间是否为不同主题的切分点。

[0015] 本发明实施例对对话文本，特别是针对网络通讯的对话文本，首先进行切词、词性标注等一系列数据预处理后，再从对话文本中找出所有的问答对，并将问句与相应的答句合并为同一句话；然后对对话文本进行主题切分，将属于不同主题且相邻的对话语句切分为不同的语块；最后对相邻且属于不同主题的语块组进行聚类，针对每个不同的主题从聚类后的语块组中抽取出主题句，使得提取出的主题具有较高的准确性。

附图说明

[0016] 图 1 是本实施例提供的对话文本主题自动提取的方法流程图；

[0017] 图 2 是本发明实施例中问句探测的原理图；

[0018] 图 3 是本发明实施例中问句探测方法的流程图；

[0019] 图 4 是本发明实施例中答句探测的原理图；

[0020] 图 5 是本发明实施例中答句探测方法的流程图；

[0021] 图 6 是本发明实施例中对对话文本进行主题切分的原理图；

[0022] 图 7 是本发明实施例中相邻句子间相似性计算的示意图；

[0023] 图 8 是本发明实施例中主题切分的可能结果示意图；

[0024] 图 9 是本发明实施例中构建的主题树示意图。

具体实施方式

[0025] 本发明实施例着重针对网络聊天对话形式的对话文本，总结出其有别于书面语文本的三个显著特点：对话文本中含有大量的问 - 答句式，不同主题的对话之间边界模糊，主题交织且组织结构混乱。针对这三个特点，本发明实施例对对话文本进行切词、词性标注等

一系列数据预处理后，再从对话文本中找出所有的问答对，并将问句与相应的答句合并为同一句话；然后对对话文本进行主题切分，将属于不同主题且相邻的对话语句切分为不同的语块；最后对相邻且属于不同主题的语块组进行聚类，针对每个不同的主题从聚类后的语块组中抽取出主题句，使得提取出的主题具有较高的准确性。

[0026] 为使本发明的目的、技术方案和优点更加清楚，下面结合附图对本发明作进一步的详细描述。

[0027] 图 1 是本实施例提供的对话文本主题自动提取的方法流程图，该流程包括以下步骤：

[0028] 步骤 101、对对话文本进行数据预处理。该数据预处理是指对聊天对话文本进行切词、词性标注、二次切分处理以及停用词处理的一系列工作。该对话文本是指用户双方的一次聊天对话内容，即用户从打开聊天窗口开始聊天到本次聊天结束关闭聊天窗口。

[0029] (1) 对对话文本进行切词处理与词性标注。

[0030] 在切词处理中，对中文和英文的切词有很大的区别，英文切词可以直接通过空格完成，而中文是紧凑排列的，需要通过专门的切词器进行切分。本实施例实现中文切词与词性标注功能采用的是中科院计算所研发的汉语词法分析系统 ICTCLAS。

[0031] (2) 对对话文本的二次切分处理。

[0032] 经过中文切词与词性标注后，句子被切分成了一个词集，由许多不同词性的词组成。如短语“自然语言理解”就会被切分为“自然 / 语言 / 理解”这三个词，但是这个短语所表达的意思与被切分为三个词后表达的意思是不一样的。

[0033] 按照 VSM(vector-space model，向量空间模型）理论，句子可以表示成 n 维空间向量，n 维表示的是对话语句的词条项数目，用 tf*idf 来计算对话语句在向量空间各个维度上的权重。如果将短语“自然语言理解”划分为“自然 / 语言 / 理解”三个词，就要用向量空间的 3 个维度表示，若一个句子中同时出现短语“自然语言理解”和“理解”一词的时候，词条“理解”的权重就明显变高，但事实上“理解”这个词在该句子中的权重应该与短语“自然语言理解”等同。

[0034] 为了避免上述情况的发生，采取的方法是在进行完切词处理后，再对句子进行二次切分处理。采用的方法是基于统计的方法，选取对话记录方面的语料库，统计两个词连续出现的共现概率，选取共现概率较高的词存入共存词集。在切词结束后，扫描一次共存词集，有匹配的词将其划归为短语。

[0035] 针对网络对话记录，会经常出现一些比较流行的短语。定期更新已有的共存词集，添加一些新出现的短语，可以使句子的切分达到更好的效果。

[0036] (3) 停用词处理。

[0037] 本实施例中所谓的停用词，指的是没有实意的虚词、类别色彩不强的词以及出现频率高但没有表意的词。编辑一个停用词表，对二次切分处理后的字词进行扫描，若判断为停用表里存在的字词，就对其标注为停用词。

[0038] 步骤 102、对预处理后的文本进行问答对探测。找出对话文本中的每个问句和其相应的答句，并将它们合并为同一句话。

[0039] 通过对对话文本进行分析，发现其含有大量的问 - 答对，且问 - 答对里面的内容包含着重要的交流信息。对话模式中往往通过多轮回的问答模式，对话双方对一个或多个

主题进行深入的探讨。所以本实施例中有一个关键的环节就是探测到对话文本中存在的问－答对，提取出的主题句信息中也会包含问－答对合并后的句子。

[0040] 本实施例针对数据预处理后的对话文本，利用机器学习的方法寻找出文本中存在的所有问句和可能存在的其相对应的答句，目的是将找出的每个问句和其对应的答句合并为一个句子，从而在提取主题句的时候可以将其整体提取出来，增强提取出主题句的可读性和全面性。

[0041] 本实施例采用的探测问答对的方法为：

[0042] 步骤 1021、探测出对话文本中的问句。

[0043] 本实施例是利用机器学习的方法探测对话文本中的问句。先通过对问句进行分析，先选定适合判断问句的一些特征；然后对准备用于训练集的句子手动标识句子类别，将从训练集句子中提取出的代表问句特征的特征值序列和人工标识的句子类别共同放入分类模型进行训练；再对作为测试集的句子手工标识句子类别，将从测试集中提取出的代表问句特征的特征值序列和手工标识的句子类别共同放入分类模型，从而得到分类模型输出结果的准确率，以便对选定的训练集、分类器和特征做相应的调整；最后对输入的新对话语句提取特征值，按照训练集提取特征的格式输入分类器，从而获得输出的分类结果。图 2 是问句探测的原理图。

[0044] 具体来说，本实施例采用的问句探测方法如图 3 所示，包括如下步骤：

[0045] 步骤 10211、选择识别问句的特征。

[0046] 对对话文本中问句的探测，分为两个层面。浅层的探测可以通过一些简单的特征来实现，如问号、疑问词、语气助词等，可以通过这些简单的特征判断出一些问句。但是网络聊天中是手写的对话文本，问号往往会被忽略。随机抽取 1000 条对话语料，有 37％省略了问号，11％的句子没有答句，还有 7％用陈述句的句型来表达问句。所以只用浅层探测方法是不充分的，需要使用其他特征识别问句。深层的探测是选择一些问句具有的隐性特征，如对话语句中词语的个数、语句前段和后段的词性顺序等。根据对话文本的特点，本实施例选择了如下特征作为分类问句的评判标准：

[0047] (1) 高标识特征，如问号、语气助词、问句疑问词、问句标识词（如“是不是”、“怎么样”等）；

[0048] (2) 输入的对话语句中词的个数；

[0049] (3) 句子中最前面的五个词的词性和句子中最后面的五个词的词性。

[0050] 步骤 10212、对准备用于训练集的句子进行人工手动标识句子类别。主要是标识经过数据预处理的训练集的句子是否为问句，从而将标识结果与训练集一起作为分类器的输入，对分类器进行训练。

[0051] 步骤 10213、基于步骤 10211 所选择的识别问句的特征，对用于训练集的句子的属性进行标识记录。首先判断句子中是否包含高标识特征，如果是的话，将高标识项对应的值置 1，不是则置 0；记录句子中词的个数，即通过步骤 101 的数据预处理切词后，记录下切分得到的句子中词的个数；记录下句子中前 5 个词和后 5 个词的词性标注。这样，就得到了训练集中的每个句子的特征值序列，每个特征值序列中包含 12 项特征在句中对应的值：是否包含高标识特征、句子中词的个数、前 5 个词和后 5 个词的词性。

[0052] 步骤 10214、将训练集的每个句子的特征值序列和人工标识句子类别共同作为分

类器的输入，对分类器进行训练。本实施例采用的分类器是朴素贝叶斯分类器，其功能就是将输入的句子分类为问句和非问句。在对分类器进行测试和正式的使用前，需要先对分类器进行训练，从而提高分类器的精度。训练集就是专门针对训练分类器而定义的句子样本集，对分类器的训练就是将训练集的每个句子的特征值序列和人工标识的句子类别共同作为分类器的输入，分类器通过对给与的输入和输出不断地学习，不断地完善分类器中的模型和参数，并通过测试集作为输入得到分类器输出结果的准确率，根据准确率的高低，再对选定的训练集、分类器和特征进行相应的调整。通过多次的训练和测试，来提高分类器的分类精确度。

[0053] 步骤 10215、将测试集中对话语句按步骤 10213 的方法，记录下代表其问句特征的特征值序列，将测试集句子的属性值序列和人工标识的句子类别共同作为分类器的输入，对分类器分类结果的准确率进行评估。通过训练集对分类器进行训练和测试集对分类器进行评估后，就要对待处理的对话文本进行问句的探测了。

[0054] 步骤 10216、将待处理的对话文本中抽取出的特征值序列作为分类器的输入，得到输出的分类结果。

[0055] 步骤 1022、通过问句在对话文本中的位置，将两个问句之间的陈述语句列为答句候选集。

[0056] 步骤 1023、在答句候选集中探测出对话文本中的每个问句相对应的答句。

[0057] 答句检测也是使用机器学习的方法，每个问句相对应的答句所存在的范围是当前问句和下一个问句之间的所有陈述句。答句探测的方法与问句探测相类似，图 4 是其原理图。

[0058] 答句探测的方法如图 5 所示，包括：

[0059] 步骤 10231、选择识别最佳答句的特征。

[0060] 根据对话文本的特点以及问句和对应答句的关联性，本实施例选择了如下特征作为判别答句的特征：

[0061] (1) 答句候选集中前五个词的词性标注和后五个词的词性标注；

[0062] (2) 答句候选集中的句子个数；

[0063] (3) 答句候选集中的答句与问句的距离；

[0064] (4) 答句候选集中的答句与问句的相似度。采用余弦相似度算法：

[0065] $$idf-\mathrm{mod}\,ified-\cos ine(x,y)=\frac{\sum_{w\in x,y} tf_{w,x}tf_{w,y}(idf_w)^2}{\sqrt{\sum_{x_i\in x}(tf_{x_i,x}idf_{x_i})^2}\sqrt{\sum_{y_i\in y}(tf_{y_i,y}idf_{y_i})^2}}$$

[0066] 步骤 10232、从选定的训练集对话语句中抽取出代表答句特征的特征值序列。

[0067] 与问句探测一样，答句探测同样采用训练集对分类器先进行训练，然后采用测试集来衡量分类器的分类准确性。最后对待处理对话文本进行分类的方法。根据前一步骤选定的识别答句的特征，将经过预处理的训练集对话语句输入，提取出每个特征所对应的特征值。每句对话语句都对应一个相应的特征值序列，特征值序列中包含 13 项，分别是该对话语句中前五个词和后五个词的词性、该句子所在的答句候选集中所包含的句子个数、该句子与问句的距离、该句子与问句的相似度。

[0068] 步骤 10233、将训练集中每个对话语句代表答句特征的特征值序列和人工标识的句子类别一同作为分类器的输入，对分类器进行训练。本实施例采用的分类器是 C4.5 决策树分类器，其功能就是将输入的句子分类为答句或非答句。

[0069] 步骤 10234、将测试集中的对话语句按步骤 10231 抽取出代表答句特征的特征值序列。并将测试集抽取出的特征值序列和人工标识的测试集句子类别作为分类器的输入，可获得分类器输出结果的准确率，从而对选定的训练集、分类器和特征进行相应的调整，使得分类器的分类准确度提高。

[0070] 步骤 10235、将待处理的答句候选集作为分类器的输入，可得到在答句候选集中与问句较适合的答句。

[0071] 步骤 10236、将寻找出的每个问句和其相对应的答句合并到同一个对话语句，并做出一定的标记。

[0072] 步骤 103、对对话文本进行主题切分。

[0073] 针对网络聊天的对话文本具有主题交织出现，各个主题之间边界模糊，组织结构混乱的特点，在抽取主题句之前，先对对话文本按照不同的主题进行切分，判别语句之间是否已经发生对话主题的偏移，识别出语义块边界，以便于对对话语句按主题进行聚类，可以更加精准的抽取主题句。

[0074] 将基于概率的主题模型思想应用于本实施例处理的对话文本，将对话文本看做是多个主题的随机组合，每个主题可以由词汇的概率分布来体现。基于这个思想，需要计算词汇在各个对话语句中的概率分布，从而计算各个相邻句子间的语义相似度，最后比较各相邻句间语义相似度与给定阈值的大小，从而确定主题切分点。

[0075] 图 6 是对对话文本进行主题切分的原理图，该方法包括 ：

[0076] 步骤 1031、将对话语句集作为输入，通过隐含语义概率模型计算得到词汇在对话文本中各个对话语句中的概率分布 $P(w|S_i)$，其中的隐含语义概率模型可以使用现有的潜在语义分析模型 PLSA、LDA 进行实现。

[0077] 步骤 1032、根据词汇在对话文本中各个对话语句中的概率分布 $P(w|S_i)$，计算相邻句子间的语义相似度，采用计算相似度的算法为 ：

[0078] $$Sim_{s_i,s_{i+1}} = \frac{\sum_{w\in W} P(w|S_i)P(w|S_{i+1})}{\sqrt{\sum_{w\in W} P(w|S_i)^2}\sqrt{\sum_{w\in W} P(w|S_{i+1})^2}}$$

[0079] 图 7 是相邻句子间相似性计算的示意图。

[0080] 步骤 1033、比较各个相邻句子间的语义相似度和给定的阈值范围，从而判定相邻的两个句子间是否为不同主题的切分点。

[0081] 步骤 104、对主题切分后的语块组进行聚类。

[0082] 在对话文本中可能存在这样的情况 ：聊天一方想对前一个话题进行一定的补充，在结束完当前话题后又去讨论前一个话题。但在这种情况下，若只对对话文本进行主题切分处理，对话文本会被切分为三个属于不同主题的语块，但事实上第一个主题和第三个主题同属一个主题，如图 8 所示。主题切分处理的不足在于只能将对话文本中相邻对话语句切分为不同主题，但不能确定非相邻语块为同一主题的情况。

[0083] 为了避免上述情况的发生，本实施例对主题切分处理进行了后续处理，使得属于

同一主题的语块能尽量聚类到一个对话文本组，从而提高抽取出的主题句的准确度。本实施例使用了一种融入语言特征的聚类算法对主题切分后的语块进行聚类处理。因为通过对大量对话文本的分析得知，在相邻两个语块之间存在着一些潜在的关联语言规则，选取关联语言特征融入聚类算法，能使聚类算法更加适用于对话文本。本实施例中融入的一个语言特征是指代特征，因为一般对话语句中代词的出现说明当前语句仍在讨论之前对话语句中说过的人或事。本实施例采用的聚类方法如下：

[0084] 步骤 1041、确定使用的聚类算法。

[0085] 假定存在两个语块是 Seg_i 和 Seg_j，融入的语言特征用条件概率表示就是 $P(T(Seg_i, Seg_j) | Seg_iPPL, Seg_jPPF)$。对于给与的两个语块 Seg_i 和 Seg_j，定义一个函数 $T(Seg_i, Seg_j)$：

[0086] $$T(Seg_i, Seg_j) = \begin{cases} 1 \\ 0 \end{cases}$$

[0087] 如果 Seg_i 和 Seg_j 属于同一个主题，计算式值为 1；否则，计算式值为 0。

[0088] 根据贝叶斯公式：

[0089] $$P(T(Seg_i, Seg_j) | Seg_iPPL, Seg_jPPF) = \frac{P(Seg_iPPL, M_jPPF | T(Seg_i, Seg_j) * P(T(Seg_i, Seg_j)))}{P(Seg_iPPL, Seg_jPPF)}$$

[0090] 计算式右边的参数估计是通过对训练数据做最大似然估计。

[0091] $$Sim(Seg, T) = \max_{i=1}^{m} \cos(Seg, Seg_i) * P(T(Seg_i, Seg_j) | Seg_iPPL, Seg_jPPF)$$

[0092] 该计算式是计算语块与建立的树之间的相似性函数。

[0093] 步骤 1042、根据相似性函数生成主题线索树。主题线索树是一种表示每一个对话语块归属的树形数据结构。

[0094] 通过语块与树之间的相似性函数作为判断当前语块是不是属于已建立的主题线索树或者一棵新树根节点的标准。以下是构建主题线索树的具体步骤：

[0095] 步骤 10421、将已进行主题切分的语块按照时间顺序进行排列。按时间排序的原因是对话主题的发展是一个时间延续的过程，从而可以判断后续语块是前面某个语块的顺承。

[0096] 步骤 10422、第一个语块内容 Seg1 形成树的根节点，同时也形成树 T_1。

[0097] 步骤 10423、处理第二个语块内容 Seg2，计算它与第一个树 T_1 的相似度 $Sim(Seg2, T_1)$。若 $Sim(Seg2, T_1)$ ＞预定门限值 k，将 Seg2 加入树 T_1；否则，语块内容 Seg2 新建一个树 T_2。

[0098] 步骤 10424、处理第三个语块内容 Seg3，分别计算它与前两棵树的相似度 $Sim(Seg3, T_1)$ 和 $Sim(Seg3, T_2)$，若 $Sim(Seg3, T_1) < Sim(Seg3, T_2)$ 且 $Sim(Seg3, T_2)$ ＞预定门限值 k（k 值根据实验结果选定），则将语块内容加入树 T_2；若 $Sim(Seg3, T_1) < Sim(Seg3, T_2)$ 且 $Sim(Seg3, T_2)$ ＜预定门限值 k，则由第三个语块内容 Seg3 新建一个树 T_3。依照 $\max_{i=1}^{m} \cos(Seg, Seg_i)$，可计算得到在当前语块所属的树下与当前语块相似度最大的语块 SegX，则当前语块 Seg3 为 SegX 的叶子节点。

[0099] 步骤 10425、之后的语块内容按照步骤 10424 中描述的方法分别进行处理，直至处理完文本中的所有语块。

[0100] 图 9 为按照以上方法构建的主题线索树。

[0101] 步骤 105、从聚类后的语块组中抽取出主题句。

[0102] 从构造的主题线索树的结构来看，已经将以时间序列排序的语块组划分为一个个的主题线索树。针对每一个主题线索树，可以将该树包含的所有语块组的对话语句看作为一个主题单元，从每个主题单元中抽取出最具代表性的一些句子作为主题句。具体方法包括：

[0103] 步骤 1051、确定每个主题单元提取主题句的个数。

[0104] $$N_i = [\frac{N_{itree}}{3}]$$

[0105] 计算式中 N_{itree} 表示第 i 个主题线索树中包含的节点个数；N_i 表示的是第 i 个主题单元中需要提取的主题句个数。

[0106] 步骤 1052、计算句子在主题单元中的贡献度。

[0107] 若主题单元中的一个句子与其他一些句子反映的是相同内容，则句子与其他句子的相似度高，若其与其他句子反映不同内容，则与其他句子的相似度低，则其对主题单元的贡献度大。假设每个主题单元里含有 s 个句子，计算主题线索树中当前句子 k 对主题单元的贡献度：

[0108] $$C_i = \sum_{i=1}^{S} |Sim(S_k, S_i) - 1|$$

[0109] 步骤 1053、通过计算式计算出每个句子对其所在的主题单元的贡献度，按照由大到小的顺序进行排序，取排名靠前的 N_i 个句子作为主题句。

[0110] 步骤 1054、将每个主题单元中分布的合并问答句都提取出来，作为一部分主题句。

[0111] 最后，用户可以从提取出来的主题句中检索或发现感兴趣的对话记录，提高用户的体验。

[0112] 总之，以上所述仅为本发明的较佳实施例而已，并非用于限定本发明的保护范围。

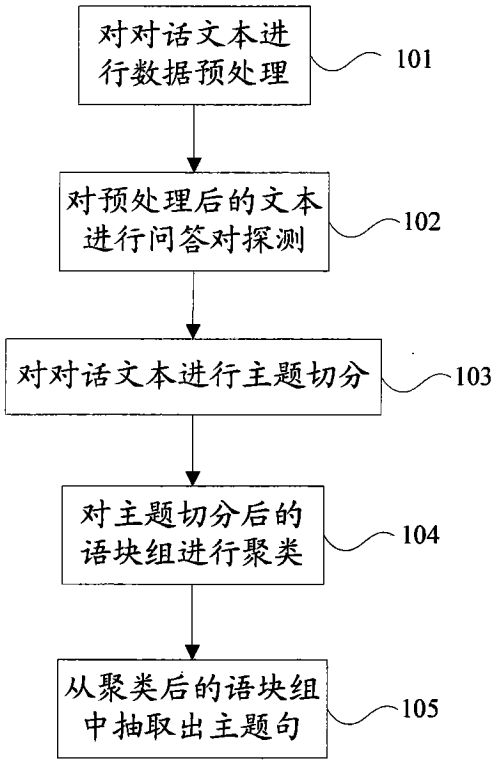

对对话文本进行数据预处理
101
对预处理后的文本进行问答对探测
102
对对话文本进行主题切分
103
对主题切分后的语块组进行聚类
104
从聚类后的语块组中抽取出主题句
105

图 1

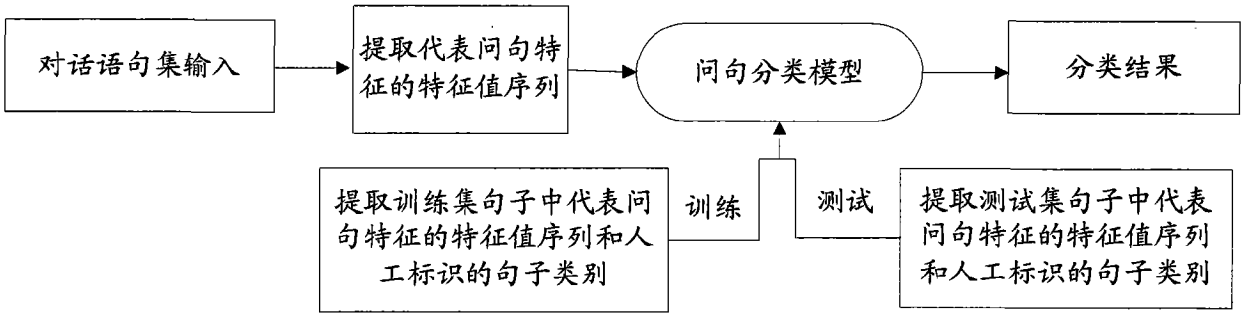

对话语句集输入
提取代表问句特征的特征值序列
问句分类模型
分类结果
提取训练集句子中代表问句特征的特征值序列和人工标识的句子类别
训练
测试
提取测试集句子中代表问句特征的特征值序列和人工标识的句子类别

图 2

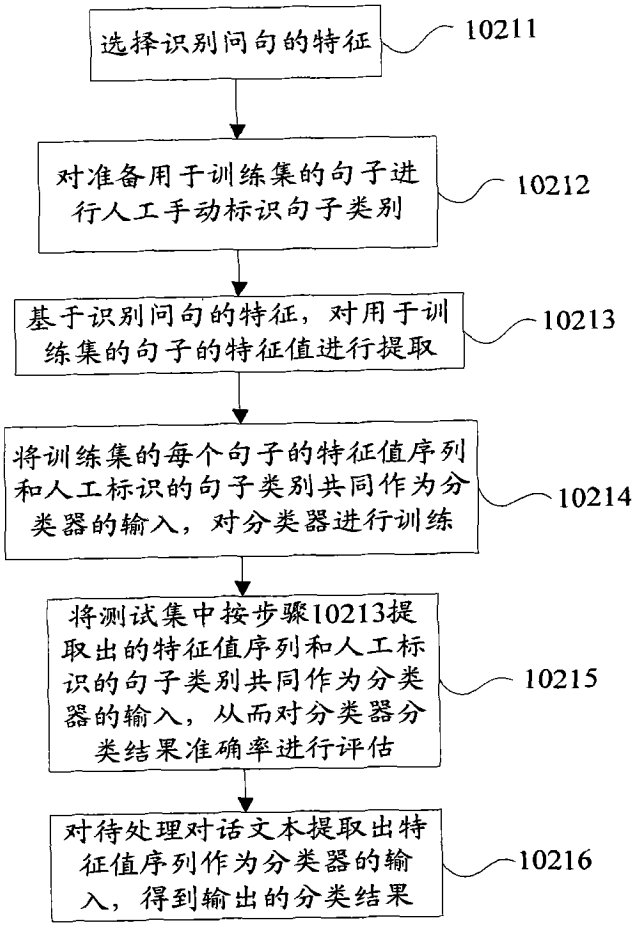

选择识别问句的特征
10211
对准备用于训练集的句子进行人工手动标识句子类别
10212
基于识别问句的特征，对用于训练集的句子的特征值进行提取
10213
将训练集的每个句子的特征值序列和人工标识的句子类别共同作为分类器的输入，对分类器进行训练
10214
将测试集中按步骤10213提取出的特征值序列和人工标识的句子类别共同作为分类器的输入，从而对分类器分类结果准确率进行评估
10215
对待处理对话文本提取出特征值序列作为分类器的输入，得到输出的分类结果
10216

图 3

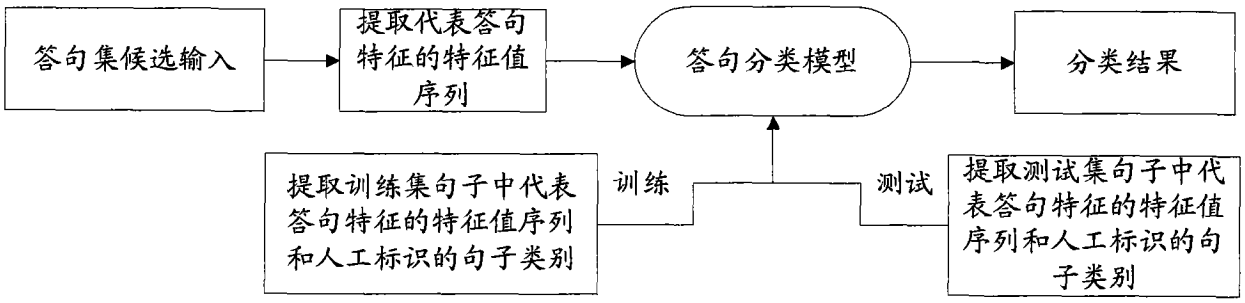

答句集候选输入
提取代表答句特征的特征值序列
答句分类模型
分类结果
提取训练集句子中代表答句特征的特征值序列和人工标识的句子类别
训练
测试
提取测试集句子中代表答句特征的特征值序列和人工标识的句子类别

图 4

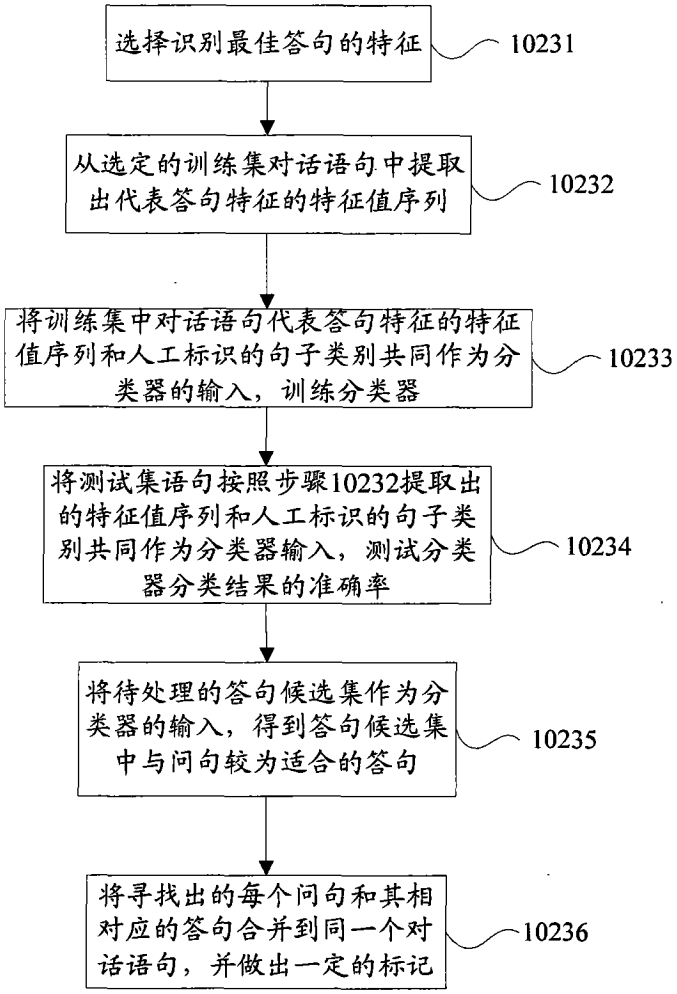

选择识别最佳答句的特征
10231
从选定的训练集对话语句中提取出代表答句特征的特征值序列
10232
将训练集中对话语句代表答句特征的特征值序列和人工标识的句子类别共同作为分类器的输入，训练分类器
10233
将测试集语句按照步骤10232提取出的特征值序列和人工标识的句子类别共同作为分类器输入，测试分类器分类结果的准确率
10234
将待处理的答句候选集作为分类器的输入，得到答句候选集中与问句较为适合的答句
10235
将寻找出的每个问句和其相对应的答句合并到同一个对话语句，并做出一定的标记
10236

图 5

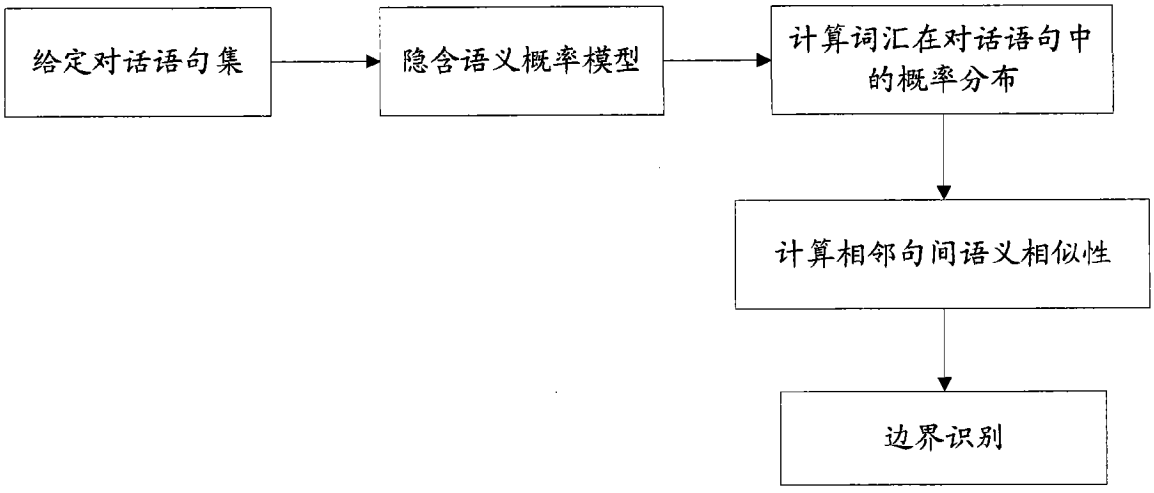

给定对话语句集
隐含语义概率模型
计算词汇在对话语句中的概率分布
计算相邻句间语义相似性
边界识别

图 6

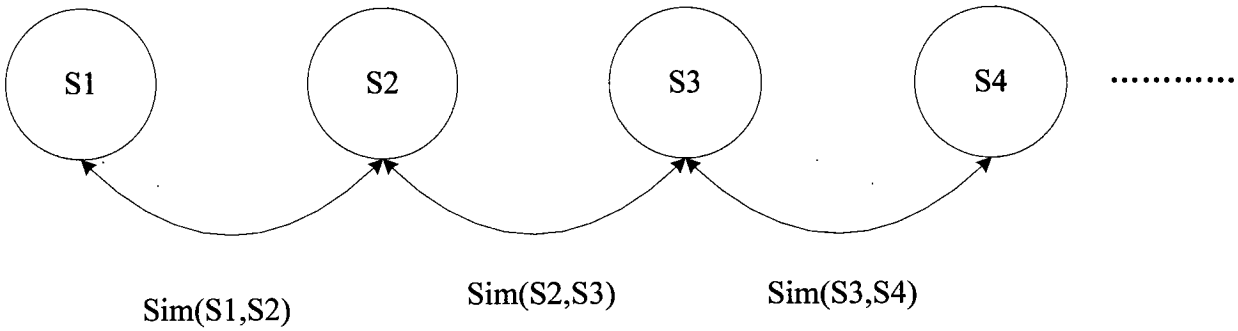

S1
S2
S3
S4
.....
Sim(S1,S2)
Sim(S2,S3)
Sim(S3,S4)

图 7

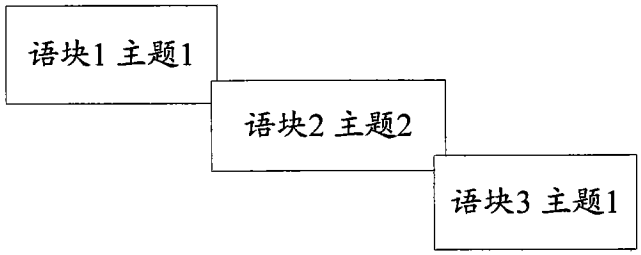

语块1 主题1
语块2 主题2
语块3 主题1

图 8

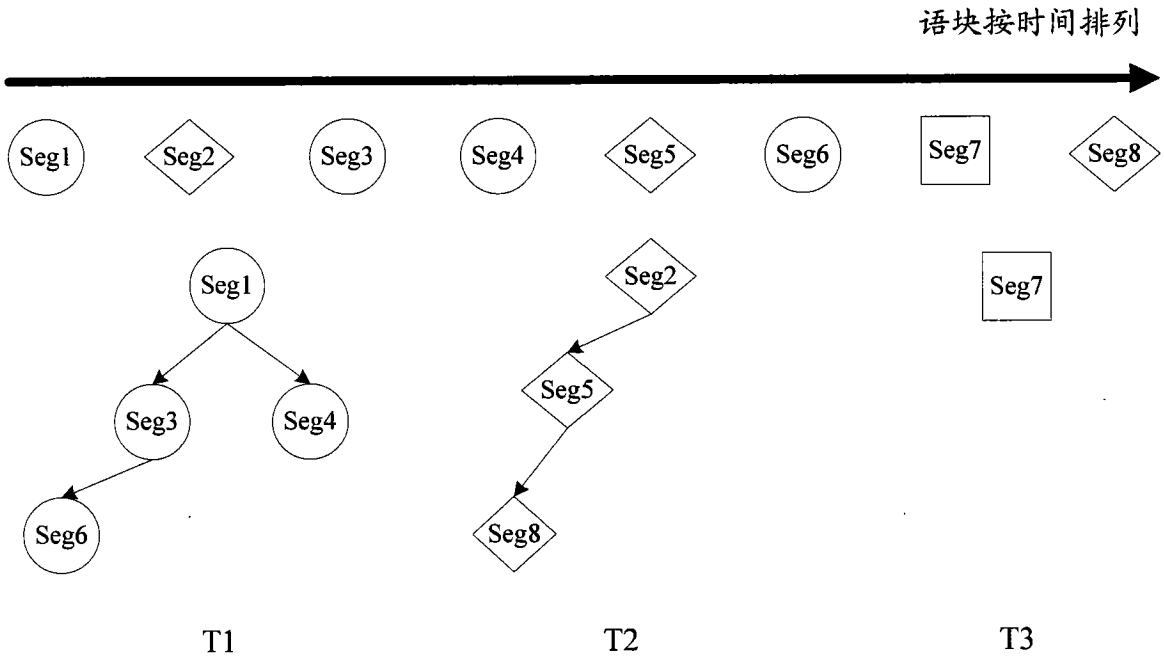

语块按时间排列
Seg1
Seg2
Seg3
Seg4
Seg5
Seg6
Seg7
Seg8
Seg1
Seg3
Seg4
Seg6
Seg2
Seg5
Seg8
Seg7
T1
T2
T3

图 9