(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
2022 年 2 月 24 日 (24.02.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/036616 A1

(51) 国际专利分类号:
G06F 16/30 (2019.01) *G06N 3/04* (2006.01)

(21) 国际申请号: PCT/CN2020/110151

(22) 国际申请日: 2020 年 8 月 20 日 (20.08.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 中山大学 (SUN YAT-SEN UNIVERSITY) [CN/CN]; 中国广东省广州市海珠区新港西路135号, Guangdong 510275 (CN)。

(72) 发明人: 余建兴 (YU, Jianxing); 中国广东省广州市海珠区新港西路135号, Guangdong 510275 (CN)。 王世祺 (WANG, Shiqi); 中国广东省广州市海珠区新港西路135号, Guangdong 510275 (CN)。 印鉴 (YIN, Jian); 中国广东省广州市海珠区新港西路135号, Guangdong 510275 (CN)。

(74) 代理人: 广州粤高专利商标代理有限公司 (YOGO PATENT AND TRADEMARK AGENCY LIMITED COMPANY); 中国广东省广州市天河区体育西路中石化大厦B塔4416室, Guangdong 510620 (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, IT, 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,

WO 2022/036616 A1

(54) **Title:** METHOD AND APPARATUS FOR GENERATING INFERENTIAL QUESTION ON BASIS OF LOW LABELED RESOURCE

(54) 发明名称: 一种基于低标注资源生成可推理问题的方法和装置

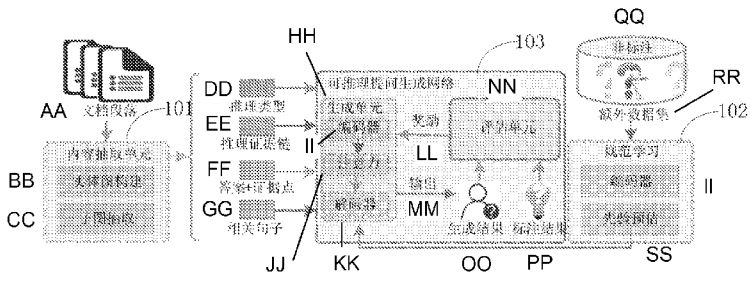

图 1

101 Content extraction unit
102 Standard learning
103 Inferential question generation network
AA Document paragraph
BB Entity diagram building
CC Sub-diagram extraction
DD Inference type
EE Inference evidence chain
FF Answer + evidence point
GG Related sentence
HH Generation unit
II Encoder
JJ Attention
KK Decoder
LL Reward
MM Output
NN Evaluation unit
OO Generation result
PP Labeling result
QQ Unlabeled
RR Additional data set
SS Prior estimation

(57) **Abstract:** Disclosed are a method and apparatus for generating an inferential question on the basis of a low labeled resource. The method comprises the following steps: S1, acquiring a labeled data set and an unlabeled data set, and establishing a question generation function; S2, building an entity diagram by taking entity words as nodes; S3, analyzing a relationship between the entity words of the entity diagram to connect the entity words to obtain a sub-diagram; S4, representing text and an inference chain as vectors, and then processing same by means of an attention mechanism to obtain a fusion vector of an input of step S5; S5, using the unlabeled data set to estimate a parameter for controlling an expression mode of a question, and using a probability distribution to perform calculation in order to generate the question; and S6, calculating a loss function index for the question, and if a preset condition is met, obtaining

[见续页]

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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))。

a final model, and ending calculation, otherwise, adjusting a model parameter, and then returning to step S4. The advantages of the present invention lie in that prior knowledge, such as an expression mode, is learned from unlabeled question data, and the syntax of a generated question is standardized, thereby effectively improving the readability of the generated question.

(57) **摘要**：本发明公开了一种基于低标注资源生成可推理问题的方法和装置，其中方法包括以下步骤：S1.获取标注数据集和非标注数据集，建立问题生成函数；S2.以实体词为节点构建实体图；S3.分析实体图的实体词之间的关系来连接实体词，得到子图；S4.将文本和推理链表示为向量，然后通过注意力机制处理为步骤S5的输入的融合向量；S5.利用非标注数据集来预估用于控制问题的表达模式的参数，采用概率分布进行计算，生成问题；S6.对问题计算损失函数指标，如果达到预设条件，则得到最终模型，结束计算；否则调整模型参数，返回步骤S4。本发明的优点在于，从未标注的提问数据中学习出表达模式等先验知识，规范所生成提问的句法，有效提升所生成提问的可读性。

一种基于低标注资源生成可推理问题的方法和装置

技术领域

本发明涉及人工智能领域，更具体地，涉及一种基于低标注资源生成可推理问题的方法和装置。

背景技术

机器阅读理解是人工智能和自然语言处理领域的研究热点，它以问答的方式来衡量机器对给定文本语义的理解能力。作为与之对偶的研究课题，问题生成（QG）致力于基于文本生成问题和与之对应的答案，能够提供训练数据来支撑问答模型的构建、生成用于教学的考题或习题、通过问题的方式来获得对话反馈等。已有的问题生成方法主要是针对简单问题，即学习单个句子中的词和问题单词间的对齐关系和映射关系，通过该关系来生成问题。

然而，目前缺乏对可推理问题的研究，而且当前方法也未能有效生成需要逻辑推理的问题；而推理是衡量机器的高级认知能力的重要指标，具有非常高的科研价值和产业应用价值。这种可推理的问题不但需要在句法语法上要正确，而且需要关联多个句子和段落中的词语和实体来推导出答案。传统的方法聚焦于研究文本和问题的映射关系，例如中国发明专利申请（公开号：CN103226562A，公开日期：2013 年 07 月 31 日）采用针对给定领域设定规则的方案，例如速度符号 v 的答案与轿车、货车、飞机等移动物体的问题相关联；中国发明专利申请（公开号：CN109726274A，公开日期：2019 年 05 月 07 日）则首先对文本的结构进行识别，根据结构类型选择对应的问题生成模型，对不同结构的文本进行针对性地生成问题的操作。上述方案缺乏对文本中实体和关系的细粒度建模，导致难以有效生成需要实体关系关联推理的问题。

另一方面，现有的问题生成方法大多需要大量的标注数据来训练，其中标注数据包括由文本、答案和问题组成的组合。例如，中国发明专利申请（公开号：CN101369265A，公开日期：2009 年 02 月 18 日）在对文本的结构进行识别后，在预先准备的词语数据库中搜索上述结构中被选中的词语的上位概念，对同样是预先准备的标签进行匹配，从而对词语进行语义标注，构建问题和答案。模型的性能直接受限于训练数据的规模。以往的研究表明，训练数据的规模与模型性能之间存在着近似对数的关联关系，即训练数据越多，模型性能一般越好。然而标

注过程非常耗费人力且昂贵，这限制了标注数据的规模，也同时限制了模型的性能。可以说，标注数据不足是在产业界和科研界普遍存在的难题。据文献调研所致，目前缺乏对在标注数据不足的情况下来做可推理问题生成的方法。

发明内容

本发明为了解决标注数据不足而未能充分地训练模型的难题，从非标注的问题中学习出先验的问题文本表达模式，并用于规范问题的生成，提升生成结果的通顺性和可读性，提供一种基于低标注资源生成可推理问题的方法和装置。

为解决上述技术问题，本发明的技术方案如下：

一种基于低标注资源生成可推理问题的方法，包括以下步骤：

S1. 获取标注数据集和非标注数据集，建立问题生成函数，其中，所述标注数据集的数据少于第二非标注数据集的数据，所述问题生成函数通过判断问题中的词与文本、答案和问题中所有的词相对应的概率，确定问题是否能够在文本中推理出答案；

S2. 从所述文本中识别出实体词，以实体词为节点构建实体图；

S3. 获取推理类型，针对推理类型分析所述实体图的实体词之间的关系，所述实体词之间的关系能够构成与推理类型对应的证据链，通过实体词之间的关系连接实体词，得到子图；

S4. 使用编码器通过编码处理将答案和证据链中的实体词的组合、推理类型、证据链相关的句子以及推理链以向量形式表示，然后通过注意力机制进行处理，在获取答案和句子之间的关联信息后，得到作为步骤 S5 的输入的融合向量；

S5. 使用规范学习单元通过隐含变量表征问题的单词片段及其上下文，并利用所述非标注数据集来预估用于控制问题的表达模式的参数，使用解码器获取所述步骤 S4 的融合向量，基于所述步骤 S1 的问题生成函数和用于控制问题的表达模式的参数，采用概率分布进行计算，生成能够在文本中推理出答案的问题；

S6. 采用训练文本和对应的训练问题，通过评估单元对步骤 S5 得到的问题进行评估计算，得到损失函数指标，如果达到预设损失函数计算迭代次数或者损失函数指标不再减少，其中预设损失函数计算迭代次数至少为2，则得到编码器、注意力机制和解码器作为生成器模型，结束计算；否则根据损失函数指标，对步骤 S4 的编码器和注意力机制、以及步骤 S5 的解码器的参数进行训练调整，返回步骤 S4。

优选地，在所述步骤 S1 中，标注数据集为$D_L=\{(B_i,A_i,Y_i)\}_{i=1}^{n}$，其中，$B$ 为文档段落，A 为答案，Y 为问题，n 为标注数据数量；非标注数据集为$D_U=\{Q_j\}_{j=1}^{\square}$，其中，$Q_j$ 为非标注问题，与标注数据问题$\{Y_i\}_{i=1}^{n}$具有相似的表达模式，非标注数据数量$\square > n$；

所述问题生成函数为以下公式：

$$\hat{Y}=\arg\max_{Y} p(Y\mid A,B)=\arg\max_{Y}\prod_{t=1}^{T} p(y_t\mid A,B,Y_{<t}) \qquad \text{公式 1；}$$

其中，B 代表文本，文本$B=(s_1,\ldots,s_I)$，A 代表根据文本内容获得的答案，答案$A=(a_1,\ldots,a_L)$，Y 代表生成的问题，问题$Y=(y_1,\ldots,y_T)$，$\hat{Y}$代表与文本 B 和答案 A 对应的问题；

其中，s_I表示文本 B 中第l^{th}个句子，I 代表文本 B 中的句子的总数，y_T表示问题中第l^{th}个词，T 代表问题中词的总数，a_L表示答案中第l^{th}个词，L 表示答案中词的总数；

其中，y_t 代表问题 Y 中的词，通过从概率分布 $p(\cdot)$ 中采样而获得，$Y_{<t}$代表问题 Y 中第 1 个到第 t-1 个的词。

优选地，所述步骤 S2 采用自然语言识别工具箱 CoreNLP 识别实体词并分析和记录实体词的属性。

优选地，在所述步骤 S2 中构建实体图的步骤中，对实体词进行比对并标记上关系标签，具体如下：

S201. 如果两个实体词共同出现在同一句子中，将所述两个实体词连接并标记上共同出现的关系标签；

S202. 如果两个实体词共同出现在同一段落的不同句子中，而且两个实体词通过词语级精确匹配计算得到的相似度值大于第一阈值，则将两个实体词连接并标记上句子级匹配的关系标签；

S203. 如果两个实体词共同出现在不同段落的不同句子中，而且两个实体词通过词语级精确匹配计算得到相似度值的大于第二阈值，则将两个实体词连接并标记上段落级匹配的关系标签；

S204. 如果两个实体词通过指代解析工具计算出具有相互引用指代的关系，则将两个实体词连接并标记上相互引用的关系标签。

优选地，第一阈值为 2/3，第二阈值为 2/3。

优选地，所述步骤 S3 中的推理类型包括线性推理类型、交集推理类型和比较推理类型。

优选地，在所述步骤 S3 中，针对线性推理类型，分析实体词之间的关系以及得到子图的具体过程是，遍历实体图并记录符合条件的关系标签，将所述符合条件的关系标签对应的实体词根据连接，得到子图以及由子图呈现的证据链，具体如下：

S3101：选择起始的实体词，通过递归地访问相邻的实体词，从实体图中检索出连接多个实体词的连续的路径；

S3102：统计路径上的关系标签，得到路径上的关系标签的总数；

S3103：判断路径是否符合给定条件，如果符合全部的给定条件，则输出路径，否则不进行操作，其中，给定条件包括：路径上的共同出现的关系标签大于 1；路径上的相互引用的关系标签大于 1；路径中不包括高频词，其中高频词为标注训练集统计出的频次排列前 5%的词；

S3104：重复所述步骤 S3101 至步骤 S3103 直至遍历实体图中全部的实体词，将输出的路径作为子图。

优选地，在所述步骤 S3 中，针对交集推理类型，分析实体词之间的关系以及得到子图的具体过程如下：

S3201：选择包含至少 2 个关系标签的实体词作为起始的实体词，通过递归地访问相邻的实体词，从实体图中检索出连接多个实体词的连续的路径；

S3202：统计路径上的关系标签，得到路径上的关系标签的总数；

S3203：判断路径是否符合给定条件，如果符合全部的给定条件，则输出路径，否则不进行操作，其中，给定条件包括：路径上的共同出现的关系标签大于 1；路径上的相互引用的关系标签大于 1；路径中不包括高频词，其中高频词为标注训练集统计出的频次排列前 5%的词；

S3204：重复所述步骤 S3201 至步骤 S3203 直至遍历实体图中全部的实体词，将输出的路径作为子图。

优选地，在所述步骤 S3 中，针对比较推理类型，分析实体词之间的关系以及得到子图的具体过程如下：

S3301：记录实体图中的全部的关系标签；

S3302：选择单个关系标签，将所述单个关系标签的两端实体词的属性与其余在步骤 S3301 得到的关系标签的两端实体词的属性逐一比对，如果比对的结果是一致时，将所述单个关系标签和比对的关系标签记录为关系对；

S3303：重复步骤 S3302 直至遍历全部的关系标签，将具有相同的关系标签的关系对通过关系标签连接成子图。

优选地，所述步骤 S4 中，编码器具体进行以下操作：

S401. 从所述步骤 S3 获得的子图的实体词筛选出答案词和证据点实体词，对文本中证据点实体词所在的全部的句子屏蔽部分答案词，其中，部分答案词为不属于比较推理类型的答案词；

S402. 对所述步骤 S401 中获得的答案词、证据点实体词以及推理链相关的句子，通过分布式向量词库，将答案词、证据点实体词和推理链相关的句子分别表示成答案向量、证据点实体向量和句子向量；

S403. 使用门控循环神经网络对所述步骤 S402 的答案向量、证据点实体向量和句子向量进行处理，通过句子向量生成第一具有上下文信息的词向量，将答案向量和证据点实体向量共同处理成实体向量，并且使用 N 层的图变换器将推理链处理成分布式向量；

S404. 基于注意力机制对句子向量进行处理；

S405. 基于答案感知的交互编码，对第一具有上下文信息的词向量和实体向量进行处理，计算并拼接答案向量和证据点实体向量整体的关联、答案向量和证据点实体向量的每个词累计向量的关联、以及答案向量和证据点实体向量的每个词最大向量的关联，得到第一答案信息感知的向量，将所述答案信息感知的向量输入到另一门控循环神经网络获得第二具有上下文信息的向量，将第一具有上下文信息的词向量和第二具有上下文信息的向量进行拼接，得到第二答案信息感知的向量；

S406. 对步骤 S402 至步骤 S405 得到的向量进行处理，得到基于可训练的参数的融合向量。

优选地，所述步骤 S5 中规范学习单元的计算过程具体如下：

S501. 基于马尔可夫神经网络模型，建立用于多次取样生成问题的单词的联合分布；

S502. 通过所述步骤 S501 的联合分布得到问题的单词后，基于双向门控循

环神经网络，建立将问题的单词表示成向量的函数；

S503. 通过反向传播算法获得问题的边际分布，最大化对数似然估计损失函数，从非标注数据学习编码器和解码器的参数；

S504. 通过维特比算法预测问题的状态序列并构成序列池，从序列池提取状态序列作为问题的表达模式，计算规范变量，其中包括问题的单词片段的状态信息和上下文信息。

优选地，所述步骤 S5 中解码器进行概率分布计算的具体过程如下：

S505. 基于复制机制生成问题的词，然后通过所述步骤 S504 得到的规范变量将所述问题的词进行组合，得到能够在文本中推理出答案的、句法表达适当的问题。

优选地，所述步骤 S6 的评估单元计算损失函数指标的具体过程包括以下步骤：

S601. 基于有监督方法和训练数据，通过最小化负交叉熵得到第一损失函数；

S602. 采用强化学习，将问题中的每个词依序逐个补充，在每次补充后，将当前得到的词作为部分序列进行评估打分，通过累计部分序列的损失函数，得到第二损失函数；

S603. 采用混合目标训练，将所述第一损失函数和基于强化学习的损失函数进行加权融合，得到输出的损失函数。

优选地，所述步骤 S602 中评估打分为分析基准输出问题和生成器输出问题，具体过程如下：

对于基准输出问题和生成器输出问题，分别计算所述部分序列的语法流畅度指标、问题的可解答性指标、以及语义关联度指标，将所述语法流畅度指标、问题的可解答性指标、以及语义关联度指标进行加权融合，得到基准输出问题的打分函数和生成器输出问题的打分函数，将基准输出问题的打分函数和生成器输出问题的打分函数相减，得到生成器输出问题的损失函数。

优选地，所述方法还包括在步骤 S6 结束后执行的步骤 S7，所述步骤 S7 包括评价性能的过程，采用 BLEU-4 指标、METEOR 指标和 ROUGE-L 指标评估所生成的问题的质量。

一种基于低标注资源生成可推理问题的装置，包括：输入模块、预处理模块、生成器模块和输出模块；

所述生成器模块包括编码器、规范学习单元、解码器和评估单元；

所述输入模块用于接收用户输入的文本；

所述预处理模块用于得到证据点实体词并构建子图；

所述编码器模块将文本、答案、证据点实体词、子图和推理类型进行编码并输出为向量；

所述规范学习单元模块表征问题的单词片段及其上下文，生成用于控制问题的表达模式的参数；

所述解码器基于问题生成函数和用于控制问题的表达模式的参数，生成能够在文本中推理出答案的问题；

所述评估单元模块对所述问题计算损失函数，根据损失函数对所述生成器模块的参数进行训练和调整，生成器模块重新生成问题，直到损失函数达到预设损失函数计算迭代次数或者不再减少，其中预设损失函数计算迭代次数至少为 2；

所述输出模块向用户输出生成器模块最后得到的问题。

与现有技术相比，本发明技术方案的有益效果是：

本发明首先从文本中抽取建立实体的关联图，通过分析实体词之间的关系识别出推理链，并利用推理链来引导结果的生成；在此基础上，为了在少量标注数据的情况下有效地训练模型，本发明从未标注的提问数据中学习出提问的表达模式等先验知识，并利用该先验知识来规范模型的生成结果，进而提升性能，从而利用未标注的提问数据含有丰富的提问表达模式和结构的特点，可以用来帮助提升所生成提问的可读性。

本发明充分利用非标注的数据来辅助提高对小规模标注数据的训练过程，有助于解决行业内普遍存在的标注训练数据短缺的问题。

附图说明

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。

图 1 是本发明的基于低标注资源生成可推理问题的方法流程示意图。

图 2 是本发明的编码器、解码器和评估单元示意图。

图 3 是本发明的基于低标注资源生成可推理问题的装置的结构示意图。

具体实施方式

附图仅用于示例性说明，不能理解为对本专利的限制；

为了更好说明本实施例，附图某些部件会有省略、放大或缩小，并不代表实际产品的尺寸；

对于本领域技术人员来说，附图中某些公知结构及其说明可能省略是可以理解的。

下面结合附图和实施例对本发明的技术方案做进一步的说明。

一种基于低标注资源生成可推理问题的方法，如图 1 和 2 所示，包括以下步骤：

S1. 获取标注数据集和非标注数据集，建立问题生成函数，其中，所述标注数据集的数据少于第二非标注数据集的数据，所述问题生成函数通过判断问题中的词与文本、答案和问题中所有的词相对应的概率，确定问题是否能够在文本中推理出答案；

S2. 从所述文本中识别出实体词，以实体词为节点构建实体图；

S3. 获取推理类型，针对推理类型分析所述实体图的实体词之间的关系，所述实体词之间的关系能够构成与推理类型对应的证据链，通过实体词之间的关系连接实体词，得到子图；

S4. 使用编码器通过编码处理将答案和证据链中的实体词的组合、推理类型、证据链相关的句子以及推理链以向量形式表示，然后通过注意力机制进行处理，在获取答案和句子之间的关联信息后，得到作为步骤 S5 的输入的融合向量；

S5. 使用规范学习单元通过隐含变量表征问题的单词片段及其上下文，并利用所述非标注数据集来预估用于控制问题的表达模式的参数，使用解码器获取所述步骤 S4 的融合向量，基于所述步骤 S1 的问题生成函数和用于控制问题的表达模式的参数，采用概率分布进行计算，生成能够在文本中推理出答案的问题；

S6. 采用训练文本和对应的训练问题，使用评估单元对步骤 S5 得到的问题进行评估计算，得到损失函数指标，如果达到预设损失函数计算迭代次数或者损失函数指标不再减少，其中预设损失函数计算迭代次数至少为2，则得到编码器、注意力机制和解码器作为生成器模型，结束计算；否则根据损失函数指标，对步骤 S4 的编码器和注意力机制、以及步骤 S5 的解码器的参数进行训练调整，返回步骤 S4。

在本实施例中，在所述步骤 S1 中，标注数据集为$D_L=\{(B_i,A_i,Y_i)\}_{i=1}^{n}$，其中，$B$ 为文档段落，A 为答案，Y 为问题，n 为标注数据数量；非标注数据集为$D_U=\{Q_j\}_{j=1}^{\square}$，其中，$Q_j$为非标注问题，与标注数据问题$\{Y_i\}_{i=1}^{n}$具有相似的表达模式，非标注数据数量$\square>n$；

所述问题生成函数为以下的公式（1）：

$$\hat{Y}=\arg\max_{Y} p(Y\mid A,B)=\arg\max_{Y}\prod_{t=1}^{T}p(y_t\mid A,B,Y_{<t}) \qquad \text{公式（1）；}$$

其中，B 代表文本，文本$B=(s_1,\ \mathbb{L},\ s_I)$，$A$ 代表根据文本内容获得的答案，答案$A=(a_1,\mathbb{L},a_L)$，Y 代表生成的问题，问题$Y=(y_1,\mathbb{L},y_T)$，$\hat{Y}$代表与文本 B 和答案 A 对应的问题；

其中，s_I表示文本 B 中第l^{th}个句子，I 代表文本 B 中的句子的总数，y_T表示问题中第l^{th}个词，T 代表问题中词的总数，a_L表示答案中第l^{th}个词，L 表示答案中词的总数；

其中，y_t 代表问题 Y 中的词，通过从概率分布$p(\cdot)$中采样而获得，$Y_{<t}$代表问题 Y 中第 1 个到第 t-1 个的词。

在本实施例中，所述步骤 S2 采用自然语言识别工具箱 CoreNLP 识别实体词并分析和记录实体词的属性。

在本实施例中，在所述步骤 S2 中构建实体图的步骤中，对实体词进行比对并标记上关系标签，具体如下：

S201. 如果两个实体词共同出现在同一句子中，将所述两个实体词连接并标记上共同出现的关系标签；

S202. 如果两个实体词共同出现在同一段落的不同句子中，而且两个实体词通过词语级精确匹配计算得到的相似度值大于第一阈值，则将两个实体词连接并标记上句子级匹配的关系标签；

S203. 如果两个实体词共同出现在不同段落的不同句子中，而且两个实体词通过词语级精确匹配计算得到相似度值的大于第二阈值，则将两个实体词连接并标记上段落级匹配的关系标签；

S204. 如果两个实体词通过指代解析工具计算出具有相互引用指代的关系，则将两个实体词连接并标记上相互引用的关系标签。

在本实施例中，第一阈值为 2/3，第二阈值为 2/3。

在本实施例中，所述步骤 S3 中的推理类型包括线性推理类型、交集推理类型和比较推理类型。

在本实施例中，在所述步骤 S3 中，针对线性推理类型，分析实体词之间的关系以及得到子图的具体过程是，遍历实体图并记录符合条件的关系标签，将所述符合条件的关系标签对应的实体词根据连接，得到子图以及由子图呈现的证据链，具体如下：

S3101：选择起始的实体词，通过递归地访问相邻的实体词，从实体图中检索出连接多个实体词的连续的路径；

S3102：统计路径上的关系标签，得到路径上的关系标签的总数；

S3103：判断路径是否符合给定条件，如果符合全部的给定条件，则输出路径，否则不进行操作，其中，给定条件包括：路径上的共同出现的关系标签大于 1；路径上的相互引用的关系标签大于 1；路径中不包括高频词，其中高频词为标注训练集统计出的频次排列前 5%的词；

S3104：重复所述步骤 S3101 至步骤 S3103 直至遍历实体图中全部的实体词，将输出的路径作为子图。

在本实施例中，在所述步骤 S3 中，针对交集推理类型，分析实体词之间的关系以及得到子图的具体过程如下：

S3201：选择包含至少 2 个关系标签的实体词作为起始的实体词，通过递归地访问相邻的实体词，从实体图中检索出连接多个实体词的连续的路径；

S3202：统计路径上的关系标签，得到路径上的关系标签的总数；

S3203：判断路径是否符合给定条件，如果符合全部的给定条件，则输出路径，否则不进行操作，其中，给定条件包括：路径上的共同出现的关系标签大于 1；路径上的相互引用的关系标签大于 1；路径中不包括高频词，其中高频词为标注训练集统计出的频次排列前 5%的词；

S3204：重复所述步骤 S3201 至步骤 S3203 直至遍历实体图中全部的实体词，将输出的路径作为子图。

在本实施例中，在所述步骤 S3 中，针对比较推理类型，分析实体词之间的关系以及得到子图的具体过程如下：

S3301：记录实体图中的全部的关系标签；

S3302：选择单个关系标签，将所述单个关系标签的两端实体词的属性与其余在步骤 S3301 得到的关系标签的两端实体词的属性逐一比对，如果比对的结果是一致时，将所述单个关系标签和比对的关系标签记录为关系对；

S3303：重复步骤 S3302 直至遍历全部的关系标签，将具有相同的关系标签的关系对通过关系标签连接成子图。

在本实施例中，所述步骤 S4 中，编码器具体进行以下操作：

S401. 从所述步骤 S3 获得的子图的实体词筛选出答案词、证据点实体词和推理链相关的句子，使用标记<UNK>来屏蔽文本中证据点实体词所在的全部的句子中的部分答案词，其中，部分答案词为不属于比较推理类型的答案词；

S402. 对所述步骤 S401 中获得的答案词、证据点实体词以及推理链相关的句子，通过分布式向量词库，将答案词、证据点实体词和推理链相关的句子分别表示成答案向量、证据点实体向量和句子向量；

具体的，对于文本类编码，采用 BERT 分布式向量词库，并且通过自然语言识别工具箱 CoreNLP 获取用于表示文本的语义和上下文关联关系的语言特征，包括：字符大小写、词性标签、命名实体标签和相互引用指代标签；但本发明不局限于此，可以根据需要引入其他的语言特征；

然后，基于上述选定的分布式向量词库和语言特征，通过基于神经网络的词分布式表示方法，将语言特征标记转换成对应的分布式向量，在具体的实施方式中，转换后的向量的维度分别为 3, 12, 8 和 3；通过把答案词、证据点实体词以及各类语言特征的向量进行拼接，可以获得增强型的文本分布式向量；

S403. 使用门控循环神经网络（GRU）对所述步骤 S402 的答案向量、证据点实体向量和句子向量进行处理，通过句子向量生成第一具有上下文信息的词向量，将答案向量和证据点实体向量共同处理成实体向量，并且使用 N 层的图变换器将推理链处理成分布式向量；

然后，通过双向的 GRU 来捕捉文本的上下文关联语义。GRU 编码器来源于文章("K. Cho, B.V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of EMNLP")。给定句子每个词的分布式向量，经过 GRU 的处理后能生成两类表示，包括：（a）

带上下文信息的词向量，对于句子中第 j^{th} 个词，可以表示成一个向量 $h_j^b=[\overleftarrow{h_j^b},\overrightarrow{h_j^b}]$，

$\overleftarrow{h_j^b}=GRU(e_j^b,\overleftarrow{h_{j+1}^b})$，$\overrightarrow{h_j^b}=GRU(e_j^b,\overrightarrow{h_{j-1}^b})$，其中 $\overleftarrow{h_j^b}$ 和 $\overrightarrow{h_j^b}$ 分别表示前向和后向 GRU 中第 j^{th} 个词对应的隐藏状态向量，e_j^b 表示这个词的分布式向量，符号[·;·]表示两个向量的拼接操作；（b）整体的编码，通过拼接开始和终止状态获得句子的整体表示 $s=[\overleftarrow{h_1^b},\overrightarrow{h_J^b}]$，其中 J 表示句子中词的总数。

类似地，答案和证据点实体一起可以表示成 $h^a=[\overleftarrow{h_1^a},\overrightarrow{h_O^a}]$，其中它们第 o^{th} 个词可表示成 $h_o^a=[\overleftarrow{h_o^a},\overrightarrow{h_o^a}]$ 向量；

对于推理链编码，为了捕捉链上的关联关系，使用 N 层的图变换器把推理链表示成分布式向量。该变换器来源于文章（“Petar Veli ˇ ckovi′ c, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Li`o, and Yoshua Bengio. 2018. Graph Attention Networks. In ICLR.”），适合捕捉图中各个节点间的关联关系。假定推理链上有□ 各节点，每个节点 v 通过其对应的实体词分布式向量来表示，即 $e_v=[\overleftarrow{h_j^b};\overrightarrow{h_{j+k}^b}]$,其中 $\overleftarrow{h_j^b}$ 是实体词的第一个单词对应的分布式向量，$\overrightarrow{h_{j+k}^b}$ 是最后的单词对应的分布式向量，k 表示实体词的单词数量。节点间的上下文通过对邻近节点做注意力加权融合获得，即 $h_v^g=e_v+\|_{n=1}^N\sum_{j\in□_v}a^n(e_v,e_j)W^ne_j$，其中‖表示向量间的拼接运算，$e_v$ 表示节点 v 的分布式表示向量，$□_v$ 表示节点 v 的邻近节点集合。$a^n(\cdot,\cdot)$ 是第 n^{th} 个注意力函数，函数如以下的公式（6）所示：

$$a^n(e_v,e_j)=\frac{\exp((W_ke_j)^TW_ee_v)}{\sum_{k\in□_v}\exp((W_ke_k)^TW_ee_v)} \qquad \text{公式（6）；}$$

其中，每个函数可以独立地学习出对应的权重，$W_k,W_e\in□^{d\times d}$。所得的点积结果通常通过对所有的边来做归一化，在实际中，为了减少这些点积求梯度的计算复杂度，本发明通过 $\frac{1}{\sqrt{d}}$ 来做归一化。

最后通过公式（7）聚合所有的节点，可以得到向量 c_g，具体如下：

$$c_g=s_t+\|_{n=1}^N\sum_{v\in□}a^n(s_t,h_v^g)W_g^nh_v^g$$
$$a^n(s_t,h_v^g)=\frac{\exp((W_hh_v^g)^TW_ds_t)}{\sum_{k\in□}\exp((W_hh_k^g)^TW_ds_t)} \qquad \text{公式（7）；}$$

其中，W_g^n，W_h，W_d 是可训练矩阵，□ 是推理链所有节点构成的集合。

S404. 基于注意力机制对句子向量进行处理；

具体的，为了能有效刻画句子中单词在语义上的长关联依赖，本发明使用自身注意力机制来进一步优化句子的分布式表示方式，即$[\hat{h}_j^b]_{j=1}^J = SelfAttn([h_j^b]_{j=1}^J)$。该机制来源于文章(“Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In Proceedings of the 55th ACL”)。具体地，给定句子的表示 H，该机制使用控制变量通过公式（8）来衡量句子内部各个单词之间的关联关系，具体如下：

$$
\begin{aligned}
u_j &= H\alpha_j \\
\alpha_j &= softmax(H^T W^u h_j^b) \\
f_j &= \tanh(W^f[h_j^b; u_j]) \qquad \text{公式(8)；} \\
g_j &= sigmoid(W^g[h_j^b; u_i]) \\
\hat{h}_j^b &= g_j \cdot f_j + (1-g_j)\cdot h_j
\end{aligned}
$$

其中，α_j 表示第 j 个单词 h_j^b 与句子 H 中其他单词的关联分数，u_j 表示第 j 个单词的上下文关联向量，h_j^b 根据 u_j 来更新为 f_j，由控制变量 g_j 来确定更新的部分。

S405. 基于答案感知的交互编码，对第一具有上下文信息的词向量和实体向量进行处理，计算并拼接答案向量和证据点实体向量整体的关联、答案向量和证据点实体向量的每个词累计向量的关联、以及答案向量和证据点实体向量的每个词最大向量的关联，得到第一答案信息感知的向量，将所述答案信息感知的向量输入到另一门控循环神经网络获得第二具有上下文信息的向量，将第一具有上下文信息的词向量和第二具有上下文信息的向量进行拼接，得到第二答案信息感知的向量；

具体的，答案感知的交互编码是$[h_j^{b'}]_{j=1}^J = MulPerFuse([\hat{h}_j^b]_{j=1}^J, [h_o^a]_{o=1}^O)$。给定句子 s 中第 j^{th} 个词的表示 h_j^b，以及答案和证据点的表示 $h_l^a|_{l=1}^L$， 通过函数 $f_m(\cdot)$ 来从多个维度捕捉它们的交互关联；本发明采用三个维度，包括整体关联，即计算 h_j^b 和答案和证据点整体的关联 $m_1 = f_m(h_j^b, h_O^a, W)$；累计关联，即计算 h_j^b 和答案和证据点各个词累计向量的关联 $m_2 = f_m(h_j^b, \sum_{o=1}^O h_o^a, W)$;；最大关联，计算 h_j^b 和答

案和证据点各个词最大向量的关联 $m_3 = f_m(h_j^b, \max_{o=1}^{O} h_o^a, W)$。而函数被定义为 $f_m(\mu, \nu, W) = \cos(W_k \odot \mu, W_k \odot \nu)$，其中 $\odot$ 表示向量间的点乘数学符号，W 表示权重矩阵，该矩阵的每列 W_k 表示对应关联维度的权重。通过拼接这些维度对应的关联向量，可以获得一个答案信息感知的向量 $m_j = [m_1; m_2; m_3]$，把该向量输入另一个双向门控循环神经网络(GRU)中来获得带上下文信息的向量 $h_j^m = [\overleftarrow{h_j^m}; \overrightarrow{h_j^m}]$。最后通过拼接获得针对句子第 j^{th} 个词的带答案信息感知的新向量 $h_j^{b'} = [\overleftarrow{h_j^b}; \overrightarrow{h_j^b}; \overleftarrow{h_j^m}; \overrightarrow{h_j^m}]$；

S406. 对步骤 S402 至步骤 S405 得到的向量进行处理，得到基于可训练的参数的融合向量；

具体的，通过公式（9）加权来融合以上的分布式表示向量，可以获得向量 c_t，其中 α_{tj} 是归一化后的注意力权重，a_{tk} 表示文本单词之间的对齐分数，s_t 表示生成出的第 t^{th} 个词对应的隐含变量，v, b, W_s, W_b 是可训练的参数，公式（9）如下所示：

$$
\begin{aligned}
c_t &= \sum_{j=1}^{J} \alpha_{tj} h_j^{b'} \\
\alpha_{tj} &= \exp(a_{tj}) / \sum_{k=1}^{J} \exp(a_{tk}) \qquad \text{公式（9）。} \\
a_{tk} &= v^T \tanh(W_s s_t + W_b h_j^{b'} + b)
\end{aligned}
$$

在本实施例中，所述步骤 S5 中规范学习单元的计算过程具体如下：

S501. 基于马尔可夫神经网络模型，建立用于多次取样生成问题 $[q_t]_{t=1}^{T} \in D_U$ 的单词的联合分布，如以下的公式（2）所示：

$$
\sum_{t=1}^{T'} p(q_{i(t-1)+1:i(t)} \mid z_t, l_t) \sum_{t=0}^{T'-1} p(z_{t+1}, l_{t+1} \mid z_t, l_t) \qquad \text{公式（2）；}
$$

其中，$p(z_{t+1}, l_{t+1} \mid z_t, l_t)$ 代表第 $(t+1)^{th}$ 个片段的状态变量和长度变量的转移概率，这些概率由前一个 t^{th} 状态来决定产生；在建立联合分布前先生成功能类似的单词片段，例如 $(q_{i(t-1)+1}, \cdots, q_{i(t)})$，其中 $i(\cdot)$ 是用于记录片段内单词下标的索引函数，其中第 t^{th} 个单词的下标为 $i(t) = \sum_{j=1}^{t} l_j$，$i(0) = 0$，$i(T') = T$；转移概率可以被分解为 $p(l_{t+1} \mid z_{t+1}) \times p(z_{t+1} \mid z_t)$，其中 $p(l_{t+1} \mid z_{t+1})$ 是关于片段最大长度 L 的均匀分

布，$p(z_{t+1}|z_t)$是关于片段状态的转移概率，如公式（3）所示：

$$p(z_{t+1}=j|z_t=i)=\frac{\exp(e_j^T e_i+b_{i,j})}{\sum_{k=1}^{K}\exp(e_k^T e_i+b_{i,k})} \quad \text{公式（3）；}$$

其中，$e_i,e_j,e_k\in\square^d$ 是片段状态 i, j, k 对应的分布式表示，$b_{i,j}$，$b_{i,k}$是标量偏置参数；在具体的实施方式中，$b_{i,j}$设置为负无穷大以避免自身迭代转移，因为相邻状态的提问片段在表达模式上通常扮演不同的语法或语义角色；

其中，$p(q_{i(t-1)+1:i(t)}|z_t,l_t)$为问题的单词的生成分布，被定义为所有提问的单词项生成概率的乘积，即$p(q_{i(t-1)+1}|z_t,l_t)\times\prod_{j=2}^{l_t}p(q_{i(t-1)+j}|q_{i(t-1)+j-1},z_t,l_t)$；

S502. 通过所述步骤 S501 的联合分布得到问题的单词后，基于双向门控循环神经网络，建立将问题的单词表示成向量的函数h_t^j，如公式（4）所示：

$$\begin{aligned}h_t^j&=GRU(h_t^{j-1},[e_{z_t};e_{q_{i(t-1)+j-1}}])\\ v_t^j&=g_{z_t}\square h_t^j\end{aligned} \quad \text{公式（4）；}$$

其中，$e_{q_{i(t-1)+j-1}}$和e_{z_t}分别表示在提问词和单词片段的分布式表示；$\square$表示按元素进行的乘法；g_{z_t}表示每个单词片段$[z_t]_{t=1}^K$对应的门控因子，该因子可通过学习获得；然后，我们通过相乘获得v_t^j，该参数捕获了单词片段上下文信息。通过 softmax 层把v_t^j输出各个提问单词的概率分布，即：

$$p(q_{i(t-1)+j}|q_{i(t-1)+j-1},z_t,l_t)=softmax(W_q v_t^j+b_q)；$$

其中，W_q和b_q通过训练获得的参数；

S503. 通过反向传播算法获得问题 Y 的边际分布$p(Y)$，如公式（5）所示：

$$\begin{aligned}&\beta_t(i)=p(q_{t+1:S}|h_t=i)=\sum_{j=1}^{K}\beta_t^*(j)p(z_{t+1}=j|z_t=i)\\ &\beta_t^*(j)=p(q_{t+1:S}|h_{t+1}=j)=\sum_{d=1}^{L}[\beta_{t+d}(j)p(d|j)p(q_{t+1:t+d}|j,d)]\\ &p(Y)=\sum_{j=1}^{K}\beta_0^*(j)p(h_1=j)\end{aligned} \quad \text{公式（5）；}$$

其中，其中$\beta_t(i)$表示第t^{th}个单词片段内的状态参数反向传播权重，$\beta_t^*(j)$表示第t^{th}个单词片段内的长度参数反向传播权重；h_t表示第t^{th}个提问 Y 的单词对应的分布式向量，初始状态为$\beta_S(i)=1,\forall i\in\{1,\cdots,K\}$；

在具体的实施方式中，为了更合理地学习出单词片段，本发明使用中文处理工具 CoreNLP 来识别提问文本的词性，本发明让模型在切分提问片段的时候尽

量不要破坏诸如动词短语（VP）和名词短语（NP）等句法成分；最后，通过反向传播算法来最大化对数似然估计损失函数，从非标注数据 D_U 学习编码器和解码器的参数；

S504. 通过维特比算法预测问题的状态序列并构成序列池，从序列池提取状态序列作为问题的表达模式，计算规范变量，其中包括问题的单词片段的状态信息和上下文信息；

具体的，本发明无偏地从序列池中抽样出一个状态序列 $[z_t]_{t=1}^{S}$ 作为提问的表达模式，其中，每个状态对应的片段长度参数 l_t 能够通过以上的 $p(l_t \mid z_t)$ 概率来计算得出；最后，本发明通过公式（4）计算出 v_m^k 来整合所有片段 $[z_t, l_t]_{t=1}^{S}$ 的状态信息和上下文信息；其中，$h_m^k = GRU(h_m^{k-1}, [e_{z_m}; e_{y_{t-1}}])$，变量 m 满足约束 $i(m-1) < t \le i(m)$，$k = t - i(m-1)$；v_m^k 捕捉了提问表达模式的有效信息，可以作为先验知识对应的参数去规范化提问的生成，其中 y_{t-1} 表示第 $(t-1)^{th}$ 个生成的提问单词。

在本实施例中，所述步骤 S5 中解码器进行概率分布计算的具体过程如下：

S505. 基于复制机制生成问题的词，然后通过所述步骤 S504 得到的规范变量将所述问题的词进行组合，得到能够在文本中推理出答案的、句法表达适当的问题；

具体的，基于上下文向量 c_t，本发明通过公式（10）的概率分布来生成提问的每个单词，具体如下：

$$\begin{aligned} p(y_t) &= p_g \cdot p_{voc}(y_t) + (1 - p_g) \cdot p_{copy}(y_t) \\ p_{voc}(y_t) &= Softmax(W_o[s_t; c_t; c_g; \rho] + b_o) \\ s_t &= GRU(s_{t-1}, v_m^k) \\ p_g &= Sigmoid(c_t, s_t, y_{t-1}) \end{aligned} \qquad \text{公式（10）；}$$

其中，推理链的分布式表示 c_g 可以引导生成器考虑推理的证据点逻辑关联，而规范变量 v_m^k 能促进模型生成语法和句法表达正确的提问。其中 ρ 是一个 1 维的向量来表示推理类型；W_o 和 b_o 表示可训练的参数；$p_{voc}(y_t)$ 表示生成提问单词的概率分布。为了解决无登录词的问题（即生成的词未在训练数据的词集合中出现），本发明采用复制机制，该机制来源于文章(“Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In Proceedings of the 54th ACL”)。这个机制能通过复制输入文本的词来

一定程度解决未登录词的问题，其中 $p_{copy}(\cdot)$ 表示复制词的概率分布， p_g 表示选择复制词或者生成词的控制变量。

在本实施例中，所述步骤 S6 的评估单元计算损失函数指标的具体过程包括以下步骤：

S601. 基于有监督方法和训练数据，通过最小化负交叉熵得到第一损失函数；

具体的，为了提升训练的收敛速度，本发明先使用有监督的方法基于训练数据 D_L 通过最小化负交叉熵来预先训练第一损失函数公式（11）的模型，具体如下：

$$L_{sl}=-\frac{1}{n}\sum_{i\in D_L}\sum_{t=1}^{T_i}\log p(y_{it}\,|\,Y_{i;<t},A_i,B_i,U_i) \qquad \text{公式（11）；}$$

为了加速收敛，本发明通过对提问 Y_i 运行 Viterbi 算法而不是采样来获得表达方式的规范变量 v_m^k， T_i 表示提问 Y_i 对应的单词个数；

S602. 采用强化学习，将问题中的每个词依序逐个补充，在每次补充后，将当前得到的词作为部分序列进行评估打分，通过累计部分序列的损失函数，得到第二损失函数；

具体的，考虑到传统的有监督学习存在硬匹配偏差和训练和测试之间的评估差异等不足，导致单纯依靠有监督学习并不一定能产生最优解；为了解决该问题，本发明借助于强化学习来微调模型，让模型更容易获得最优解；强化学习是业界广泛使用的一种训练方法，擅长于优化非连续函数的目标；本发明使用第二损失函数 $L_{rl}=-E_{Y^s\square\pi_\theta}[r(Y^s)]$，找出最佳的生成单词策略 π_θ 来最小化所生成提问 Y^s 对应的；其中，θ 是模型的参数集，分值函数 $r(Y)$ 通过指定指标来衡量模型输出的提问文本 Y^s 和标注提问 Y^* 之间的差异；

S603. 采用混合目标训练，将所述第一损失函数和基于强化学习的损失函数进行加权融合，得到输出的损失函数；

具体的，考虑到使用单一的损失函数有可能导致生成提问的可读性不强，为了解决该问题，本发明采用一个混合目标的损失函数来提升可读性，如以下的公式（13）所示：

$$L=\gamma L_{rl}+(1-\gamma)L_{sl} \qquad \text{公式（13）；}$$

其中，γ 是权重参数。

在具体的实施方式中，考虑到模型需要约束来逼近标注结果，来避免各类局

部最优的可能，强化学习的权重γ设置为0.3。

在本实施例中，所述步骤S602中评估打分为分析基准输出问题和生成器输出问题，具体过程如下：

对于基准输出问题和生成器输出问题，分别计算所述部分序列的语法流畅度指标、问题的可解答性指标、以及语义关联度指标，将所述语法流畅度指标、问题的可解答性指标、以及语义关联度指标进行加权融合，得到基准输出问题的打分函数和生成器输出问题的打分函数，将基准输出问题的打分函数和生成器输出问题的打分函数相减，得到生成器输出问题的损失函数；

每一种指标的评估方式具体如下：

(a)流畅度：本发明采用基于语言模型计算负困惑度的方式来衡量所生成的提问文本的流畅度，计算方式为$r_{ppl}(Y)=-2^{-\frac{1}{T}\sum_{t=1}^{T}\log_2 p_{LM}(y_t|Y_{<t})}$，来源于文章（"X. Zhang and M. Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In Proceedings of EMNLP"），在实际应用中能有效衡量生成文本的质量；

(b)可解答：本发明采用$QBLEU_4(Y^s,Y^*)$来衡量生成的提问的可解答性；具体地，准确率和召回率的计算方法分别是$P_{avg}=\sum_i w_i\frac{c(S_i)}{|l_i|}$和$R_{avg}=\sum_i w_i\frac{c(S_i)}{|r_i|}$，其中$i\in\{r,n,q,f\}$，$\sum_i w_i=1$，$|l_i|$，$|r_i|$分别表示属于$i^{th}$种类型的生成提问和标注提问单词数，r, n, q, f分别代表相关内容词、实体词、提问词和功能词；通过以下公式加权获可解答函数$QBLEU_4(\cdot,\cdot)=\delta Answerability+(1-\delta)BLEU_4$其中$Answerability=\frac{2P_{avg}\cdot R_{avg}}{P_{avg}+R_{avg}}$，$\delta$是权重参数；$BLEU_{n=4}$是匹配度函数，来源于文章（"K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2019. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th ACL"），通过计算文本对应子串的重叠度来衡量翻译文本和真实文本的匹配状况，即越多子串能匹配，分值越高；

(c)语义关联：考虑到问题表达方式的多样性，本发明奖励地提升与真实问题Y^*在分布式空间中高度相似的提问Y^s的分值；为了计算相似度，本发明采用词步长距离(WMD)，来源于文章（"H. Gong, S. Bhat, L. Wu, J. Xiong, and W. Hwu. 2019. 2019. Reinforcement Learning Based Text Style Transfer without Parallel

Training Corpus. In Proceedings of the 57th NAACL")，具有高效和鲁棒性很强的特点，用于计算两个文本在分布式空间中的语义相似度；通过生成文本的词语长度来正则化，就能获得语义关联指标的分值 $r_{sem}(Y) = -WMD(Y^s, Y^*) / Length(Y^*)$；其中 WMD(.) 函数计算公式如下 $WMD(Y^s, Y^*) = \min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij} \left\| A(x_i - x_j) \right\|_2^2$

$\forall i,j, \sum_{j=1}^{n} T_{ij} = \tilde{d}_i^a$ and $\sum_{i=1}^{n} T_{ij} = \tilde{d}_j^b$ ， $\tilde{d}^a = (w \circ d^a) / (w^T \circ d^a)$；

考虑到以上奖励函数不可微不可导，本发明使用自临界策略梯度训练算法来训练模型，该算法来源于("S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel 2017. Self-Critical Sequence Training for Image Captioning. In Proceedings of the CVPR")。具体地，该算法定义生成器输出问题的损失函数，如以下的公式（12）所示：

$$L_{rl} = (r(Y^b) - r(Y^s)) \sum_{t=1}^{T} \log p(y_t^s \mid A, B, Y_{<t}^s) \qquad \text{公式（12）；}$$

其中，Y^b 表示基准方法的输出序列结果，该基准方法通过一种局部最优的方式生成训练，即使用贪婪算法每次生成概率最大的词；Y^s 是生成器所输出的序列结果，每个词 y_t^s 通过采用公式（12）的概率值来获得；通过最小化该损失函数，就能优化模型，让其生成比基准方法分值更高的序列。

在本实施例中，所述方法还包括在步骤 S6 结束后执行的步骤 S7，所述步骤 S7 包括评价性能的过程，采用 BLEU-4 指标、METEOR 指标和 ROUGE-L 指标评估所生成的问题的质量；

具体的，考虑到机器阅读理解是提问生成的对偶任务，本发明使用可推理数据集 HotpotQA 进行实验，该数据集来源于文章("Z. Yang, P. Qi, S. Zhang, Y. Bengio, W.W. Cohen, R. Salakhutdinov, and C.D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 56th ACL")。，该数据集被分割成训练和测试集，分别有 9 万和 7 千个标注样本。本发明使用 10%的训练数据作为开发集来调优模型。每个样本由一个提问、答案和若干个段落组成。此外，本发明还收集了两个非标注的提问数据集，用于训练提问表达模式的先验知识，包括 ComplexWebQuestions 和 DROP，这两个数据集均为人工标注构建的可推理提问，但没有标注关联上对应的文档和答案。这

两个数据集规模分别为 3.5 万条和 9.7 万条，其中 ComplexWebQuestions 数据集来源于论文 (“Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In Proceedings of the 2018 NAACL”)；DROP 数据集来源于论文(“Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Proceedings of the 2019 NAACL”)。

本发明使用三种传统指标方法来衡量生成的提问的质量，包括 BLEU-4、METEOR 和 ROUGE-L。其中指标 BLEU-4 来源于论文(“Kishore Papineni, Salim Roukos, Todd Ward, and Wei- Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th ACL”)；METEOR 来源于论文(“Kishore Papineni, Salim Roukos, Todd Ward, and Wei- Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th ACL”)；ROUGE-L 来源于论文(“Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out”)。

实验结果表明，本发明生成提问的质量明显地优于传统方法。

一种基于低标注资源生成可推理问题的装置，如图 3 所示，包括：输入模块、预处理模块、生成器模块和输出模块；

所述生成器模块包括编码器、规范学习单元、解码器和评估单元；

所述输入模块用于接收用户输入的文本；

所述预处理模块用于得到证据点实体词并构建子图；

所述编码器模块将文本、答案、证据点实体词、子图和推理类型进行编码并输出为向量；

所述规范学习单元模块表征问题的单词片段及其上下文，生成用于控制问题的表达模式的参数；

所述解码器基于问题生成函数和用于控制问题的表达模式的参数，生成能够在文本中推理出答案的问题；

所述评估单元模块对所述问题计算损失函数，根据损失函数对所述生成器模块的参数进行训练和调整，生成器模块重新生成问题，直到损失函数达到预设损失函数计算迭代次数或者不再减少，其中预设损失函数计算迭代次数至少为 2；

所述输出模块向用户输出生成器模块最后得到的问题。

显然，本发明的上述实施例仅仅是为清楚地说明本发明所作的举例，而并非是对本发明的实施方式的限定。对于所属领域的普通技术人员来说，在上述说明的基础上还可以做出其它不同形式的变化或变动。这里无需也无法对所有的实施方式予以穷举。凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明权利要求的保护范围之内。

权 利 要 求 书

1. 一种基于低标注资源生成可推理问题的方法，其特征在于，包括以下步骤：

S1. 获取标注数据集和非标注数据集，建立问题生成函数，其中，所述标注数据集的数据少于第二非标注数据集的数据，所述问题生成函数通过判断问题中的词与文本、答案和问题中所有的词相对应的概率，确定问题是否能够在文本中推理出答案；

S2. 从所述文本中识别出实体词，以实体词为节点构建实体图；

S3. 获取推理类型，针对推理类型分析所述实体图的实体词之间的关系，所述实体词之间的关系能够构成与推理类型对应的证据链，通过实体词之间的关系连接实体词，得到子图；

S4. 使用编码器通过编码处理将答案和证据链中的实体词的组合、推理类型、证据链相关的句子以及推理链以向量形式表示，然后通过注意力机制进行处理，在获取答案和句子之间的关联信息后，得到作为步骤 S5 的输入的融合向量；

S5. 使用所述规范学习单元通过隐含变量表征问题的单词片段及其上下文，并利用所述非标注数据集来预估用于控制问题的表达模式的参数，使用解码器获取所述步骤 S4 的融合向量，基于所述步骤 S1 的问题生成函数和用于控制问题的表达模式的参数，采用概率分布进行计算，生成能够在文本中推理出答案的问题；

S6. 通过评估单元对步骤S5得到的问题进行评估计算，得到损失函数指标，如果达到预设损失函数计算迭代次数或者损失函数指标不再减少，其中预设损失函数计算迭代次数至少为 2，则得到编码器、注意力机制和解码器作为生成器模型，结束计算；否则根据损失函数指标，对步骤 S4 的编码器和注意力机制、以及步骤 S5 的解码器的参数进行训练调整，返回步骤 S4。

2. 根据权利要求 1 所述的基于低标注资源生成可推理问题的方法，其特征在于，在所述步骤 S1 中，标注数据集为 $D_L = \{(B_i, A_i, Y_i)\}_{i=1}^{n}$，其中，$B$ 为文档段落，A 为答案，Y 为问题，n 为标注数据数量；非标注数据集为 $D_U = \{Q_j\}_{j=1}^{□}$，其中，Q_j 为非标注问题，与标注数据问题 $\{Y_i\}_{i=1}^{n}$ 具有相似的表达模式，非标注数据数量□ $> n$；

所述问题生成函数为以下公式：

$$\hat{Y} = \arg\max_{Y} p(Y \mid A, B) = \arg\max_{Y} \prod_{t=1}^{T} p(y_t \mid A, B, Y_{<t}) \qquad \text{公式 1；}$$

其中，B 代表文本，文本$B=(s_1,\ldots,s_I)$，A 代表根据文本内容获得的答案，答案$A=(a_1,\ldots,a_L)$，Y 代表生成的问题，问题$Y=(y_1,\ldots,y_T)$，$\hat{Y}$代表与文本 B 和答案 A 对应的问题；

其中，s_I表示文本 B 中第l^{th}个句子，I 代表文本 B 中的句子的总数，y_T表示问题中第l^{th}个词，T 代表问题中词的总数，a_L表示答案中第l^{th}个词，L 表示答案中词的总数；

其中，y_t 代表问题 Y 中的词，通过从概率分布 $p(\cdot)$ 中采样而获得，$Y_{<t}$代表问题 Y 中第 1 个到第 t-1 个的词。

3. 根据权利要求 1 所述的自动生成可推理问答的方法，其特征在于，所述步骤 S2 采用自然语言识别工具箱 CoreNLP 识别实体词并分析和记录实体词的属性。

4. 根据权利要求 1 所述的自动生成可推理问答的方法，其特征在于，在所述步骤 S2 中构建实体图的步骤中，对实体词进行比对并标记上关系标签，具体如下：

S201. 如果两个实体词共同出现在同一句子中，将所述两个实体词连接并标记上共同出现的关系标签；

S202. 如果两个实体词共同出现在同一段落的不同句子中，而且两个实体词通过词语级精确匹配计算得到的相似度值大于第一阈值，则将两个实体词连接并标记上句子级匹配的关系标签；

S203. 如果两个实体词共同出现在不同段落的不同句子中，而且两个实体词通过词语级精确匹配计算得到相似度值的大于第二阈值，则将两个实体词连接并标记上段落级匹配的关系标签；

S204. 如果两个实体词通过指代解析工具计算出具有相互引用指代的关系，则将两个实体词连接并标记上相互引用的关系标签。

5. 根据权利要求 4 所述自动生成可推理问答的方法，其特征在于，第一阈值为 2/3，第二阈值为 2/3。

6. 根据权利要求 1 所述的自动生成可推理问答的方法，其特征在于，所述步骤 S3 中的推理类型包括线性推理类型、交集推理类型和比较推理类型。

7. 根据权利要求 4 和 6 所述的自动生成可推理问答的方法，其特征在于，

在所述步骤 S3 中，针对线性推理类型，分析实体词之间的关系以及得到子图的具体过程是，遍历实体图并记录符合条件的关系标签，将所述符合条件的关系标签对应的实体词根据连接，得到子图以及由子图呈现的证据链，具体如下：

S3101：选择起始的实体词，通过递归地访问相邻的实体词，从实体图中检索出连接多个实体词的连续的路径；

S3102：统计路径上的关系标签，得到路径上的关系标签的总数；

S3103：判断路径是否符合给定条件，如果符合全部的给定条件，则输出路径，否则不进行操作，其中，给定条件包括：路径上的共同出现的关系标签大于 1；路径上的相互引用的关系标签大于 1；路径中不包括高频词，其中高频词为标注训练集统计出的频次排列前 5%的词；

S3104：重复所述步骤 S3101 至步骤 S3103 直至遍历实体图中全部的实体词，将输出的路径作为子图。

8. 根据权利要求 4 和 6 所述的自动生成可推理问答的方法，其特征在于，在所述步骤 S3 中，针对交集推理类型，分析实体词之间的关系以及得到子图的具体过程如下：

S3201：选择包含至少 2 个关系标签的实体词作为起始的实体词，通过递归地访问相邻的实体词，从实体图中检索出连接多个实体词的连续的路径；

S3202：统计路径上的关系标签，得到路径上的关系标签的总数；

S3203：判断路径是否符合给定条件，如果符合全部的给定条件，则输出路径，否则不进行操作，其中，给定条件包括：路径上的共同出现的关系标签大于 1；路径上的相互引用的关系标签大于 1；路径中不包括高频词，其中高频词为标注训练集统计出的频次排列前 5%的词；

S3204：重复所述步骤 S3201 至步骤 S3203 直至遍历实体图中全部的实体词，将输出的路径作为子图。

9. 根据权利要求 3、4 和 6 所述的自动生成可推理问答的方法，其特征在于，在所述步骤 S3 中，针对比较推理类型，分析实体词之间的关系以及得到子图的具体过程如下：

S3301：记录实体图中的全部的关系标签；

S3302：选择单个关系标签，将所述单个关系标签的两端实体词的属性与其余在步骤 S3301 得到的关系标签的两端实体词的属性逐一比对，如果比对的结果

是一致时，将所述单个关系标签和比对的关系标签记录为关系对；

S3303：重复步骤 S3302 直至遍历全部的关系标签，将具有相同的关系标签的关系对通过关系标签连接成子图。

10. 根据权利要求 1 所述的自动生成可推理问答的方法，其特征在于，所述步骤 S4 中，编码器具体进行以下操作：

S401. 从所述步骤 S3 获得的子图的实体词筛选出答案词和证据点实体词，对文本中证据点实体词所在的全部的句子屏蔽部分答案词，其中，部分答案词为不属于比较推理类型的答案词；

S402. 对所述步骤 S401 中获得的答案词、证据点实体词以及推理链相关的句子，通过分布式向量词库，将答案词、证据点实体词和推理链相关的句子分别表示成答案向量、证据点实体向量和句子向量；

S403. 使用门控循环神经网络对所述步骤 S402 的答案向量、证据点实体向量和句子向量进行处理，通过句子向量生成第一具有上下文信息的词向量，将答案向量和证据点实体向量共同处理成实体向量，并且使用 N 层的图变换器将推理链处理成分布式向量；

S404. 基于注意力机制对句子向量进行处理；

S405. 基于答案感知的交互编码，对第一具有上下文信息的词向量和实体向量进行处理，计算并拼接答案向量和证据点实体向量整体的关联、答案向量和证据点实体向量的每个词累计向量的关联、以及答案向量和证据点实体向量的每个词最大向量的关联，得到第一答案信息感知的向量，将所述答案信息感知的向量输入到另一门控循环神经网络获得第二具有上下文信息的向量，将第一具有上下文信息的词向量和第二具有上下文信息的向量进行拼接，得到第二答案信息感知的向量；

S406. 对步骤 S402 至步骤 S405 得到的向量进行处理，得到基于可训练的参数的融合向量。

11. 根据权利要求 1 所述的自动生成可推理问答的方法，其特征在于，所述步骤 S5 中规范学习单元的计算过程具体如下：

S501. 基于马尔可夫神经网络模型，建立用于多次取样生成问题的单词的联合分布；

S502. 通过所述步骤 S501 的联合分布得到问题的单词后，基于双向门控循

环神经网络，建立将问题的单词表示成向量的函数；

S503. 通过反向传播算法获得问题的边际分布，最大化对数似然估计损失函数，从非标注数据学习编码器和解码器的参数；

S504. 通过维特比算法预测问题的状态序列并构成序列池，从序列池提取状态序列作为问题的表达模式，计算规范变量，其中包括问题的单词片段的状态信息和上下文信息。

12. 根据权利要求 10 和 11 所述的自动生成可推理问答的方法，其特征在于，所述步骤 S5 中解码器进行概率分布计算的具体过程如下：

S505. 基于复制机制生成问题的词，然后通过所述步骤 S504 得到的规范变量将所述问题的词进行组合，得到能够在文本中推理出答案的、句法表达适当的问题。

13. 根据权利要求 1 所述的自动生成可推理问答的方法，其特征在于，所述步骤 S6 的评估单元计算损失函数指标的具体过程包括以下步骤：

S601. 基于有监督方法和训练数据，通过最小化负交叉熵得到第一损失函数；

S602. 采用强化学习，将问题中的每个词依序逐个补充，在每次补充后，将当前得到的词作为部分序列进行评估打分，通过累计部分序列的损失函数，得到第二损失函数；

S603. 采用混合目标训练，将所述第一损失函数和基于强化学习的损失函数进行加权融合，得到输出的损失函数。

14. 根据权利要求 13 所述的自动生成可推理问答的方法，其特征在于，所述步骤 S602 中评估打分为分析基准输出问题和生成器输出问题，具体过程如下：

对于基准输出问题和生成器输出问题，分别计算所述部分序列的语法流畅度指标、问题的可解答性指标、以及语义关联度指标，将所述语法流畅度指标、问题的可解答性指标、以及语义关联度指标进行加权融合，得到基准输出问题的打分函数和生成器输出问题的打分函数，将基准输出问题的打分函数和生成器输出问题的打分函数相减，得到生成器输出问题的损失函数。

15. 根据权利要求 1 所述的自动生成可推理问答的方法，其特征在于，所述方法还包括在步骤 S6 结束后执行的步骤 S7，所述步骤 S7 包括评价性能的过程，采用 BLEU-4 指标、METEOR 指标和 ROUGE-L 指标评估所生成的问题的质量。

16. 一种基于低标注资源生成可推理问题的装置，其特征在于，包括：输入

模块、预处理模块、生成器模块和输出模块；

所述生成器模块包括编码器、规范学习单元、解码器和评估单元；

所述输入模块用于接收用户输入的文本；

所述预处理模块用于得到证据点实体词并构建子图；

所述编码器模块将文本、答案、证据点实体词、子图和推理类型进行编码并输出为向量；

所述规范学习单元模块表征问题的单词片段及其上下文，生成用于控制问题的表达模式的参数；

所述解码器基于问题生成函数和用于控制问题的表达模式的参数，生成能够在文本中推理出答案的问题；

所述评估单元模块对所述问题计算损失函数，根据损失函数对所述生成器模块的参数进行训练和调整，生成器模块重新生成问题，直到损失函数达到预设损失函数计算迭代次数或者不再减少，其中预设损失函数计算迭代次数至少为 2；

所述输出模块向用户输出生成器模块最后得到的问题。

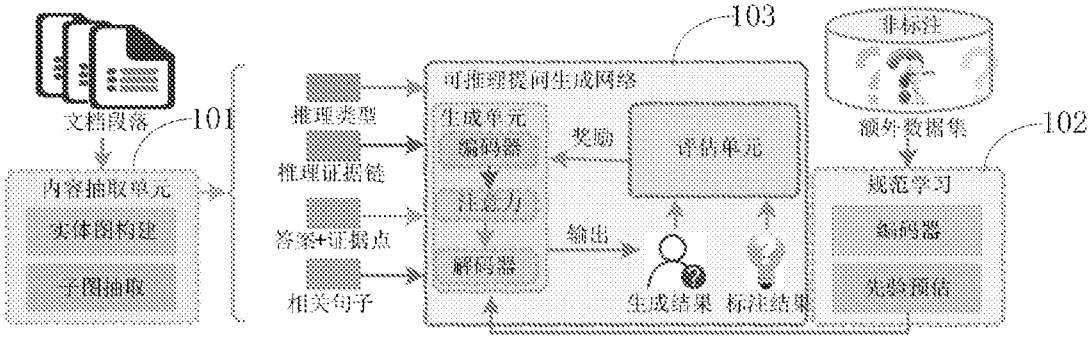
文档段落
101
内容抽取单元
实体图构建
子图抽取
推理类型
推理证据链
答案+证据点
相关句子
103
可推理提问生成网络
生成单元
编码器
注意力
解码器
奖励
评估单元
输出
生成结果
标注结果
非标注
额外数据集
102
规范学习
编码器
先验规范

图 1

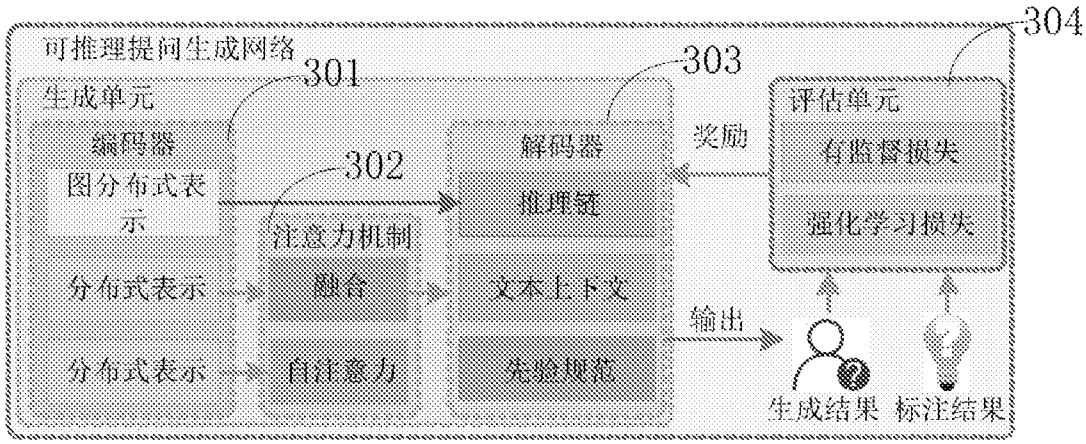
可推理提问生成网络
304
生成单元
301
303
编码器
图分布式表示
302
分布式表示
分布式表示
注意力机制
融合
自注意力
解码器
推理链
文本上下文
先验规范
奖励
评估单元
有监督损失
强化学习损失
输出
生成结果
标注结果

图 2

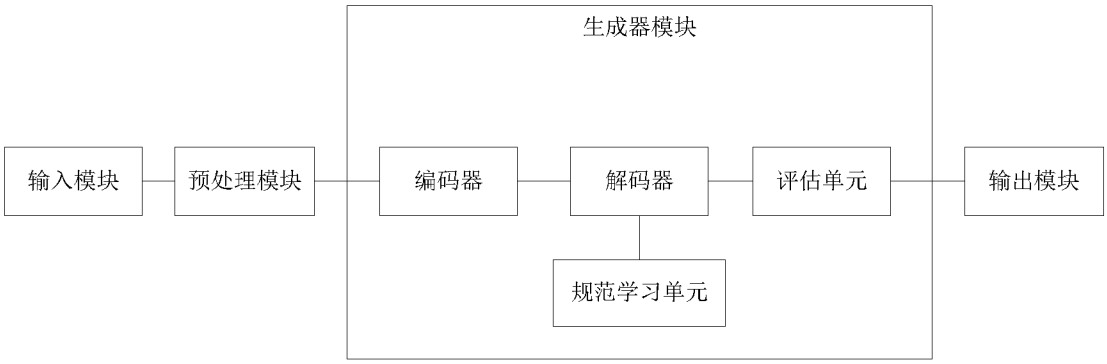

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/110151

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/30(2019.01)i; G06N 3/04(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 G06N

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)

CNABS; CNTXT; CNKI; SIPOABS; DWPI; USTXT; WOTXT; EPTXT: 提问, 关联, 数据集, 遍历, 标签, 标注, 推理, 关系, 证据链, 训练集, 少, 实体, 问题, 弱监督, 路径, 答案, 知识, 注意力, 无监督, 门控循环, 随机, 样本, question, relation, associate, sample, witness, proof, chain, lack, inference, marked, entity, supervised, weak, BERT, attention

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 109918489 A (SHANGHAI LEYAN INFORMATION TECHNOLOGY CO., LTD.) 21 June 2019 (2019-06-21) description, paragraphs [0039]-[0082], and figures 1-5	1-16
A	CN 111274814 A (ZHEJIANG UNIVERSITY) 12 June 2020 (2020-06-12) entire document	1-16
A	CN 111428490 A (BEIJING INSTITUTE OF TECHNOLOGY) 17 July 2020 (2020-07-17) entire document	1-16
A	CN 111125370 A (NANJING SINOVATIO TECHNOLOGY CO., LTD.) 08 May 2020 (2020-05-08) entire document	1-16
A	CN 110765269 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 07 February 2020 (2020-02-07) entire document	1-16

☑ 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
17 April 2021	**17 May 2021**
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

International application No.
PCT/CN2020/110151

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	曾宇涛 等 (ZENG, Yutao et al.). "基于多维信息融合的知识库问答实体链接 (Multi-Dimensional Information Integration Based Entity Linking for Knowledge Base Question Answering)" *模式识别与人工智能 (Pattern Recognition and Artificial Intelligence),* Vol. 32, No. 7, 15 July 2019 (2019-07-15), ISSN: 1003-6059, pp. 642-651	1-16

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2020/110151

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109918489 A	21 June 2019	CN 109918489 B	02 February 2021
CN 111274814 A	12 June 2020	None	
CN 111428490 A	17 July 2020	None	
CN 111125370 A	08 May 2020	None	
CN 110765269 A	07 February 2020	None	

国际检索报告

国际申请号
PCT/CN2020/110151

A. 主题的分类

G06F 16/30(2019.01)i; G06N 3/04(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 G06N

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）)

CNABS;CNTXT;CNKI;SIPOABS;DWPI;USTXT;WOTXT;EPTXT：提问，关联，数据集，遍历，标签，标注，推理，关系，证据链，训练集，少，实体，问题，弱监督，路径，答案，知识，注意力，无监督，门控循环，随机，样本，question，relation，associate，sample，witness，proof，chain，lack，inference，marked，entity，supervised，weak，BERT，attention

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
A	CN 109918489 A (上海乐言信息科技有限公司) 2019年 6月 21日 (2019-06-21) 说明书第[0039]-[0082]段，图1-5	1-16
A	CN 111274814 A (浙江大学) 2020年 6月 12日 (2020-06-12) 全文	1-16
A	CN 111428490 A (北京理工大学) 2020年 7月 17日 (2020-07-17) 全文	1-16
A	CN 111125370 A (南京中新赛克科技有限责任公司) 2020年 5月 8日 (2020-05-08) 全文	1-16
A	CN 110765269 A (华南理工大学) 2020年 2月 7日 (2020-02-07) 全文	1-16
A	曾宇涛 等. "基于多维信息融合的知识库问答实体链接" 模式识别与人工智能, 第32卷，第7期，2019年 7月 15日 (2019-07-15)， ISSN：1003-6059, 第642-651页	1-16

☐ 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
2021年 4月 17日	2021年 5月 17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 任洪潮 电话号码 （86-512）88995644

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/110151

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 109918489 A	2019年 6月 21日	CN 109918489 B	2021年 2月 2日
CN 111274814 A	2020年 6月 12日	无	
CN 111428490 A	2020年 7月 17日	无	
CN 111125370 A	2020年 5月 8日	无	
CN 110765269 A	2020年 2月 7日	无	