



(12)发明专利

(10)授权公告号 CN 105393248 B

(45)授权公告日 2019.01.08

(21)申请号 201480036652.9

大竹清敬

(22)申请日 2014.05.15

(74)专利代理机构 中科专利商标代理有限责任
公司 11021

(65)同一申请的已公布的文献号
申请公布号 CN 105393248 A

代理人 薛凯

(43)申请公布日 2016.03.09

(51)Int.Cl.
G06F 16/33(2019.01)

(30)优先权数据
2013-134649 2013.06.27 JP

(85)PCT国际申请进入国家阶段日
2015.12.25

(86)PCT国际申请的申请数据
PCT/JP2014/062913 2014.05.15

(87)PCT国际申请的公布数据
W02014/208213 JA 2014.12.31

(73)专利权人 国立研究开发法人情报通信研究
机构
地址 日本国东京都

(56)对比文件
CN 101872349 A,2010.10.27,
CN 101377777 A,2009.03.04,
JP 2009116662 B2,2009.05.20,
Jong-Hoon Oh 等.“Why Question
Answering using Sentiment Analysis and
Word Classes”.《Proceedings of the 2012
Joint Conference on Empirical Methods in
Natural Language Processing and
Computational Natural Language Learning》
.2012,第368-378页.

审查员 滕冲

(72)发明人 吴钟勋 鸟泽健太郎 桥本力
佐野大树 史蒂恩·德萨哲

权利要求书3页 说明书14页 附图11页

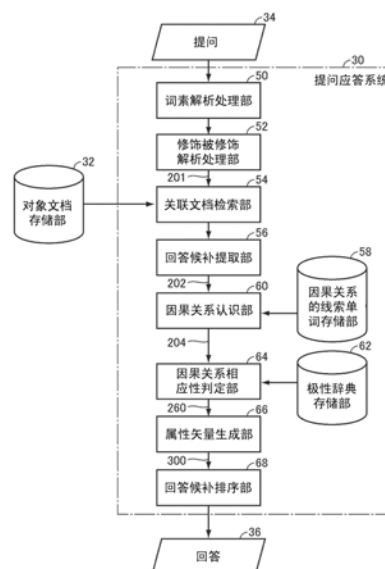
(54)发明名称

非事实类提问应答系统以及方法

(57)摘要

提供能适当地回答非事实类提问的提问应答系统。提问应答系统(30)包括:关联文档检索部(54),其响应于接受到提问,从对象文档存储部(32)取出回答候补;回答候补提取部(56),其从回答候补中提取有力的回答候补;因果关系认识部(60),其认识包含于所提取的回答候补中的因果关系表现;因果关系相应性判定部(64),其判定认识出的因果关系作为回答是否相称;属性矢量生成部(66),其使用到因果关系相应性判定部(64)为止的处理结果,生成与提问、和检索到的回答候补的组合相关的属性矢量;和属性矢量生成部(66),其在被给予属性矢量时,对成为生成该属性矢量的基础的提问和回答候补算出表示该回答候补是针对该提问的正确的回答的程

度的评分,将评分最高的回答候补作为针对提问的回答输出。



1. 一种非事实类的提问应答系统,接受非事实类的提问的输入,生成针对该提问的回答,与存储计算机可读的多个文档的文档存储单元连接而使用,

候补检索单元,其响应于接受到提问的输入,从所述文档存储单元取出针对该提问的回答候补;

属性生成单元,其响应于由所述候补检索单元检索到回答候补,对所述提问、和由所述候补检索单元检索到的各个回答候补的组合生成给定的属性的集合;和

回答选择单元,其在被给予由所述属性生成单元生成的所述属性的集合时,对成为生成该属性的集合的基础的提问和回答候补算出表示该回答候补是针对该提问的正确的回答的程度的评分,将评分最高的回答候补作为针对所述提问的回答输出,

所述属性生成单元包括:

第1属性算出单元,其对所述提问以及各回答候补算出根据各自的词素信息以及语法信息求得的第1属性的组;和

第2属性算出单元,其选择各所述回答候补中的因果关系表现中作为给出针对所述提问的回答相应的因果关系表现,算出从该因果关系表现得到的第2属性的组,

所述属性的集合包含所述第1属性的组和所述第2属性的组,

所述第2属性算出单元包括:

线索表现确定单元,其在各所述回答候补中确定成为用于因果关系表现的确定的线索的表现;

因果关系表现确定单元,其确定:通过在各所述回答候补中由所述线索表现确定单元所确定的线索表现联系起来的、由因果关系的原因部分和结果部分构成的因果关系表现;

一致判定单元,其判定包含于所述确定的因果关系表现的结果部分中的名词和修饰该名词的谓语的极性的组合、以及包含于所述提问中的名词和修饰该名词的谓语的极性的组合是否一致;

相应性判定单元,其使用所述一致判定单元的判定结果、和各所述回答候补与所述提问间的单词匹配以及修饰被修饰树的子树匹配,来判定在各所述回答候补中由所述因果关系表现确定单元确定的因果关系表现作为针对所述提问的回答是否相应;和

将表示所述相应性判定单元的判定结果的信息作为所述第2属性的组的属性而输出的单元。

2. 根据权利要求1所述的非事实类的提问应答系统,其中,

所述候补检索单元包括:

分类单元,其响应于接受到所述提问的输入,将该提问分类为给定的多个类型的任意者;

句分割单元,其响应于接受到所述提问的输入,将存储于所述文档存储单元的各文档分割为由连续的1个或多个句构成的句的组,并输出;和

对由所述句分割单元输出的各句的组,根据遵循所述分类单元的分类结果确定的基准来算出评分、将该评分上位的句的组作为所述回答候补而输出的单元。

3. 根据权利要求2所述的非事实类的提问应答系统,其中,

所述第2属性算出单元还包括:将由所述因果关系表现确定单元确定的因果关系表现的单词的n-gram、该因果关系表现中的修饰被修饰树的子树的n-gram、或者包含于该因果

关系表现的谓语的极性和修饰该谓语的名词的意义的组合、或它们的任意的组合作为所述第2属性的组的属性而输出的单元。

4. 根据权利要求2或3所述的非事实类的提问应答系统,其中,

所述第2属性算出单元还包括:将由所述因果关系表现确定单元确定的因果关系表现当中的结果部分和所述提问共有的单词、修饰被修饰树的子树、活性的谓语的数量、或者非活性的谓语的数量、或它们的任意的组作为所述第2属性的组的属性而输出的单元。

5. 根据权利要求1所述的非事实类的提问应答系统,其中,

所述第2属性算出单元还包括:将由所述因果关系表现确定单元确定的因果关系表现的单词的n-gram、该因果关系表现中的修饰被修饰树的子树的n-gram、或者包含于该因果关系表现的谓语的极性和修饰该谓语的名词的意义的组合、或它们的任意的组合作为所述第2属性的组的属性而输出的单元。

6. 根据权利要求1或5所述的非事实类的提问应答系统,其中,

所述第2属性算出单元还包括:将由所述因果关系表现确定单元确定的因果关系表现当中的结果部分和所述提问共有的单词、修饰被修饰树的子树、活性的谓语的数量、或者非活性的谓语的数量、或它们的任意的组作为所述第2属性的组的属性而输出的单元。

7. 一种非事实类的提问应答方法,接受非事实类的提问的输入,生成针对该提问的回答,使用存储计算机可读的多个文档的文档存储单元执行,

所述非事实类的提问应答方法包括:

候补检索步骤,响应于接受到提问的输入,从所述文档存储单元取出针对该提问的回答候补;

属性生成步骤,响应于在所述候补检索步骤中检索到回答候补,对所述提问、和在所述候补检索步骤中检索到的各个回答候补的组合,生成给定的属性的集合;和

回答选择步骤,响应于被给予在所述属性生成步骤中生成的所述属性的集合,对成为生成该属性的集合的基础的提问和回答候补算出表示该回答候补是针对该提问的正确的回答的程度的评分,将评分最高的回答候补作为针对所述提问的回答输出,

所述属性生成步骤包括:

第1属性算出步骤,其对所述提问以及各回答候补算出根据各自的词素信息以及语法信息而求得的第1属性的组;和

第2属性算出步骤,选择各所述回答候补中的因果关系表现中作为给出针对所述提问的回答相应的因果关系表现,算出从该因果关系表现得到的第2属性的组,

所述属性的集合包含所述第1属性的组和所述第2属性的组,

所述第2属性算出步骤包括:

线索表现确定步骤,其在各所述回答候补中确定成为用于因果关系表现的确定的线索的表现;

因果关系表现确定步骤,其确定:通过在各所述回答候补中在所述线索表现确定步骤中确定的线索表现联系起来的、由因果关系的原因部分和结果部分构成的因果关系表现;

一致判定步骤,其判定包含于所述确定的因果关系表现的结果部分中的名词和修饰该名词的谓语的极性的组合、以及包含于所述提问中的名词和修饰该名词的谓语的极性的组合是否一致;

相应性判定步骤,其使用所述一致判定步骤中的判定结果、和各所述回答候补与所述提问间的单词匹配以及修饰被修饰树的子树匹配,来判定在各所述回答候补中在所述因果关系表现确定步骤中确定的因果关系表现作为针对所述提问的回答是否相应;和

将表示所述相应性判定步骤中的判定结果的信息作为所述第2属性的组的属性而输出的步骤。

非事实类提问应答系统以及方法

技术领域

[0001] 本发明涉及利用了计算机以及网络的提问应答系统,特别涉及针对询问某件事情的理由等的所谓的为什么型提问等被称作非事实(non-factoid)类提问的提问的应答系统。

背景技术

[0002] 关于针对某些事实的提问应答(QA)的研究最近取得大的进步。当前,IBM公司的Watson、Apple公司的Siri等提问应答系统的改善特别突出,一部分已经开始商用利用。另外,这些以外的企业也在推进同样的开发。对于与事实相关的提问据说其精度为85%程度。

[0003] 另一方面,在不是与事实相关的提问,而是从大量的文档提取针对像「为什么会得癌症?」那样询问发生某些事态的理由的提问的回答即为什么型提问应答的情况下,则被认识为是非常困难的任务。上述的IBM公司以及Apple公司的产品也未应对为什么型提问。

[0004] 关于该问题,例如后述的非专利文献1公开了如下技法:在所谓的信息检索的技法中,特别在信息检索的查询中加进「理由」这样的单词,来从大量的文档中检索包含针对给出的为什么型提问的回答的一段。非专利文献2公开了如下技术:将人工制作的表征原因和结果的单词对的数据库、或表征理由的句法型式的数据库中出现的单词对以及型式用作属性,通过有指导学习来确定回答。

[0005] 另一方面,和这些不同,在非专利文献3中公开了如下技术:将文本的结构性特征、即局部句法树、单词的意义上的归类、评价表现等的意义上的特征用作属性,通过有指导学习来确定回答。

[0006] 现有技术文献

[0007] 非专利文献

[0008] 非专利文献1:Masaki Murata,Sachiyo Tsukawaki,Toshiyuki Kanamaru,QingMa,and Hitoshi Isahara.A system for answeringnon-factoid Japanese questions by using passage retrieval weighted based on type of answer.In Proceedings of NTCIR-6,2007.

[0009] 非专利文献2:Ryuichiro Higashinaka and Hideki Isozaki.Corpus-based questionanswering for why-questions.In Proceedings of IJCNLP,pages418-425., 2008.

[0010] 非专利文献3:Jong-Hoon Oh,Kentaro Torisawa,Chikara Hashimoto,Takuya Kawada,Stijn De Saeger,Jun'ichi Kazama and Yiou Wang.WhyQuestion Answering using Sentiment Analysis and Word Classes,In Proceedings ofEMNLP-CoNLL 2012, pp.368-378.

[0011] 发明的概要

[0012] 发明要解决的课题

[0013] 上述任意的现有技术虽然都有用,但在使任务的性质充分反映在所利用的属性以

及其他信息中的意义上,认为依然有改善的余地。特别在使该任务的性质得到反映的手法中,更有效率地利用因果关系的。因果关系是重要的这样的问题在本技术领域得到广泛的认识,但在现有技术中,有作为因果关系仅关注名词短语间的因果关系这样的问题。并且由于认为名词短语间的因果关系原则上出现在1句中,因此仅利用在同一句内的名词短语间的因果关系。

[0014] 然而,实际上因果关系并非仅有名词短语间的因果关系。有时在名词短语间以及名词短语与动词短语间都存在因果关系。进而,有跨2句出现因果关系的情况,这一点在过去一直置之不理。总而言之,出现在文档数据内的因果关系的形式是多样的,在为什么型提问应答系统中,需要在认识这样的多样的形式的因果关系的基础上再将其信息用在回答提取中。然而,过去对这样的多样性一直没有关注。

[0015] 进而,在要基于这样的多样的形式的因果关系来得到提问的回答的情况下,有如何处置与如此的因果关系相关的信息尚不明确这样的问题。

发明内容

[0016] 为此本发明的目的在于,提供能通过合适地处置出现在文档内的因果关系的多样的表现,来对非事实类提问给出合适的回答的非事实类提问应答系统。

[0017] 用于解决课题的手段

[0018] 本发明的第1局面所涉及的非事实类的提问应答系统是接受非事实类的提问的输入、生成针对该提问的回答的系统,与存储计算机可读的多个文档的文档存储单元连接而使用。该系统包括:候补检索单元,其响应于接受到提问的输入,从文档存储单元取出针对该提问的回答候补;属性生成单元,其响应于由候补检索单元检索到回答候补,对提问、和由候补检索单元检索到的各个回答候补的组合生成给定的属性的集合;和回答选择单元,其在被给予由属性生成单元生成的属性的集合时,对成为生成该属性的集合的基础的提问和回答候补算出表示该回答候补是针对该提问的正确的回答的程度的评分,将评分最高的回答候补作为针对提问的回答输出。属性生成单元包括:第1属性算出单元,其对提问以及各回答候补算出根据各自的词素信息(morpheme information)以及语法信息(syntactic information)求取的第1属性的组;和第2属性算出单元,其选择各回答候补中的因果关系表现中作为给出针对提问的回答相应的因果关系表现,算出从该因果关系表现得到的第2属性的组。属性的集合包含第1属性的组和第2属性的组。

[0019] 根据该非事实类的提问应答系统,能将回答候补中的因果关系表现作为给出针对提问的回答的因果关系表现是否相应用作第2组的属性。通过在选择回答候补的基础上对该因果关系表现进行考虑,能使提问的表现和因果关系的表现的含意关系反映在回答的选择中。其结果,能提供通过合适地处置出现在文档内中的因果关系的多样的表现来对非事实类提问给出合适的回答的非事实类提问应答系统。

[0020] 优选地,候补检索单元包括:分类单元,其响应于接受到提问的输入,将该提问分类为给定的多个类型的任意者;句分割单元,其响应于接受到提问的输入,将存储于文档存储单元的各文档分割为由连续的1个或多个句构成的句的组,并输出;和对由句分割单元输出的各句的组根据遵循分类单元的分类结果确定的基准来算出评分、将该评分上位的句的组作为回答候补而输出的单元。

[0021] 更优选地,第2属性算出单元包括:线索表现确定单元,其在各回答候补中确定成为用于因果关系表现的确定的线索的表现;因果关系表现确定单元,其确定通过在各回答候补中线索表现确定单元所确定的线索表现连接起来的由因果关系的原因部分和结果部分给出的因果关系表现;一致判定单元,其判定包含于确定的因果关系表现的结果部分中的名词和修饰该名词的谓语的极性的组合、以及包含于提问中的名词的意义和修饰该名词的谓语的极性的组合是否一致;相应性判定单元,其使用一致判定单元的判定结果、和各回答候补与提问间的单词匹配以及修饰被修饰树的子树匹配来判定在各回答候补中由因果关系表现确定单元确定的因果关系表现作为针对提问的回答是否相应;和将表示相应性判定单元的判定结果的信息作为第2属性的组的属性输出的单元。

[0022] 也可以让第2属性算出单元还包括:将由因果关系表现确定单元确定的因果关系表现的单词的n-gram、该因果关系表现中的修饰被修饰树的子树的n-gram、或者包含于该因果关系表现中的谓语的极性和修饰该谓语的名词的意义组合、或它们的任意的组合作为第2属性的组的属性输出的单元。

[0023] 也可以让第2属性算出单元还包括:将由因果关系表现确定单元确定的因果关系表现当中的结果部分和提问共有的单词、修饰被修饰树的子树、活性的谓语的数量、或者非活性的谓语的数量、或它们的任意的组合作为第2属性的组的属性输出的单元。

[0024] 本发明的第2方面所涉及的计算机程序使计算机作为上述任意的非事实类的提问应答系统的全部单元发挥功能。

[0025] 本发明的另一方面所涉及的非事实类的提问应答方法接受非事实类的提问的输入,生成针对该提问的回答。该方法使用存储计算机可读的多个文档的文档存储单元执行。该方法包括:候补检索步骤,响应于接受到提问的输入,从文档存储单元取出针对该提问的回答候补;属性生成步骤,响应于在候补检索步骤检索到回答候补,对提问、和在候补检索步骤检索到的各个回答候补的组合生成给定的属性的集合;和回答选择步骤,响应于被给予在属性生成步骤生成的属性的集合,对成为生成该属性的集合的基础的提问和回答候补算出表示该回答候补是针对该提问的正确的回答的程度的评分,将评分最高的回答候补作为针对提问的回答输出。属性生成步骤包括:第1属性算出步骤,对提问以及各回答候补算出根据各自的词素信息以及语法信息求取的第1属性的组;和第2属性算出步骤,选择各回答候补中的因果关系表现中作为给出针对提问的回答相应的因果关系表现,算出从该因果关系表现得到的第2属性的组。属性的集合包含第1属性的组和第2属性的组。

附图说明

[0026] 图1是本发明的1个实施方式所涉及的非事实类提问应答系统的概略框图。

[0027] 图2是用于说明从修饰被修饰树中的子树得到的n-gram的图。

[0028] 图3是用于说明因果关系的形式的图。

[0029] 图4是用于说明针对提问的回答的相应性的图。

[0030] 图5是用于说明在图1所示的系统中选择回答候补的方法的示意图。

[0031] 图6是表示回答候补的列表例的图。

[0032] 图7是表示从回答候补中确定表示因果关系的线索的单词的方法的图。

[0033] 图8是说明谓语的活性/非活性的图。

[0034] 图9是表示图1所示的系统当中因果关系认识部的构成的框图。

[0035] 图10是表示图1所示的因果关系相应性判定部的构成的框图。

[0036] 图11是表示图1所示的回答候补排序部的构成的框图。

[0037] 图12是表示本发明的1个实施方式所使用的属性当中、和现有技术所使用的属性相同的属性的图。

[0038] 图13是用于说明本发明的实施方式带来的效果的图表。

[0039] 图14是表示用于实现本发明的1个实施方式所涉及的提问应答系统的计算机的外观的示意图。

[0040] 图15是说明图14所示的计算机的硬件构成的框图。

具体实施方式

[0041] 在以下的说明以及附图中,对同一部件标注同一参考编号。因此不再反复对它们的详细的说明。

[0042] [基本的思路]

[0043] 本申请发明的发明者关注于每当从大量的文档检索针对非事实类提问的回答时,因果关系、即表征某事态的原因和结果的意义上的关系是有效的这一点。根据该观点,认识存在于为了检索回答而准备的文本中的因果关系的表现,通过使用了合适的属性的有指导学习来从这些文本中确定针对提问的回答。用于此的步骤的概略如以下那样汇总。

[0044] (1) 综合地认识以各文本中的1句内的名词短语、动词短语以及节等表现的因果关系、以及跨连续的2个以上的句而表现的因果关系等多样的形式的因果关系。采用的手法的详细在后面叙述

[0045] (2) 多数情况下,因果关系当中的结果部分成为提问的回答。进而,多数情况下,结果部分当中作为提问的回答而相称的部分和提问在意义上类似。在本实施方式中,关注这样的特性,取出在上述(1)得到的因果关系当中作为回答相称的因果关系。关于其手法也在后面叙述。特别是该手法的特征在于,根据结果部分的谓语的「活性/非活性」这样意义上的极性来认识含意关系,利用该含意关系这一点。

[0046] (3) 采用合适地表征通过上述(1)以及(2)的手法得到的因果关系的属性,通过机器学习手法来确定回答。

[0047] 上面记载的(1)以及(2)在为什么型提问应答的性能提升中都是重要的。这是因为,从以多样的形式表现的因果关系中认识作为提问的回答而相称的因果关系直接关系到回答提取的精度提升。在现有技术中,不管是利用多样的形式的因果关系,还是利用根据谓语的意义上的极性而得到的含意关系,都没有研讨过。

[0048] 在本实施方式中,使用已有的信息检索技术从大量的文档中收集回答候补。综合认识出现在收集的回答候补中的多样的形式的因果关系,进而通过基于「活性/非活性」这样意义上的极性的含意关系认识来确定作为提问的回答而相称的因果关系的信息。其结果,能有效地运用存在于文档中的成为针对为什么型提问的合适的回答的句。具体使用以下的步骤。

[0049] (A) 通过已有的信息检索技术,从大量的文本确定成为回答的候补的少数的语句。只要是能收集认为与提问有关联的文档的手法即可,能使用任何手法。

[0050] (B) 接下来,如以下那样从这些回答候补的文本,认识在生成针对提问的回答上有效的因果关系。以合适的属性表征该因果关系的结果部分,用在有指导学习中,从而评价作为回答候补的结果部分。将该评价高的采用为回答。具体的步骤如以下那样。

[0051] (B1) 从回答候补的文本找到「～ため(由于～)」、「～ので(因为～)」、「～からだ(因为是～)」、「～の原因は(的原因是～)」等因果关系的线索表现(单词)。这些表现是将因果关系的原因部分和结果部分联系起来的表現。因此,能通过提取这些表现来确定因果关系的原因部分和结果部分。通过对线索表现下工夫,不仅能确定原因部分和结果部分出现在1句内的因果关系,还能确定原因部分和结果部分在不同的句内的因果关系。

[0052] 参考图3,作为因果关系的形式,典型地有:1) 句内的名词短语间的因果关系;2) 句内的名词短语与动词短语间的因果关系;3) 句内的节与动词短语间的因果关系;以及4) 跨2句而出现的因果关系。

[0053] 关于它们当中的第1个因果关系,在图3所示的例句中,通过「が原因となる(成为原因)」这样的表现84将成为原因的名词短语80、和成为结果的名词短语82联系起来。在第2个因果关系中,通过「が原因で(为原因)」这样的表现90将成为原因的名词短语86和成为结果的动词短语88联系起来。在第3个因果关系中,通过「ため(由于)」这样的表现96将由节构成的原因部分92和成为结果的由动词短语构成的结果部分94联系起来。在第4个示例中,通过「が原因で(为原因)」这样的表现98将成为原因的第1个语句102和成为结果的第2个语句内的动词短语100联系起来。

[0054] 即,都是以特定的表现将因果关系的成为原因的部分和成为结果的部分联系起来。

[0055] (B2) 将在上述(B1)得到的因果关系当中作为为什么型提问的回答而相称的因果关系以外的因果关系丢弃。在此,关注因果关系的结果部分和为什么型提问在意义上类似这一点,通过1) 单词匹配、2) 修饰被修饰解析结果(修饰被修饰树)中的子树匹配、然后3) 根据谓语的活性/非活性这样意义上的极性而认识的含意关系的匹配,来进行因果关系的结果部分是否作为为什么型提问的回答而相称的判断。

[0056] 参考图4,设想提问为「なぜ津波が起こるのですか?(为什么出现海啸)」的情况。图4的2个回答候补都表征因果关系,原因部分92以及110和结果部分94以及112分别通过作为因果关系的线索表现的「ため(由于)」这样表现96以及114而联系起来。即便如此,认为第1个候补作为回答相称,而第2个候补不相称。这是因为,第1个结果部分94的表现「津波が発生します。(发生海啸)」和提问句的表现「津波が起こる(出现海啸)」表征类似的意义,与此相对,第2个结果部分112的表现「津波を弱める効果がある(有削弱海啸的效果)」表征和提问句的表现相反的意义。在实施方式中,需要将这样的相异合适地表现为属性。

[0057] (B3) 从通过上述(B1)和(B2)得到的因果关系作成用于确定回答的有指导学习用的属性。在此,使用1) 回答候补中的因果关系的存在的有无、2) 以n-gram等表现的因果关系的内容、还有3) 提问和因果关系的结果部分所共有的单词、子树以及含意关系的数量等用作属性。最后,通过使用这些属性进行了学习的有指导学习器来确定回答。

[0058] 通过认识具有多样的形式的因果关系,并运用作为提问的回答而相称的因果关系的信息,能更加精度良好地确定为什么型提问的回答。

[0059] [构成]

[0060] 参考图1,本实施方式所涉及的提问应答系统30接受为什么型提问等的非事实类的提问34,从存储成为回答搜索的对象的大量的文档的对象文档存储部32提取具有作为针对提问34的回答而相称的结果部分的因果关系表现,作为回答36而输出。在本实施方式中,为了易于理解以下的说明,在对象文档存储部32存放对各句预先分解为词素串并进行修饰被修饰解析从而附加了修饰被修饰解析结果的文档。

[0061] 提问应答系统30包括:词素解析处理部50,其接受提问34并进行词素解析处理;和修饰被修饰解析处理部52,其对由词素解析处理部50输出的词素串进行修饰被修饰解析、并输出修饰被修饰树201。修饰被修饰解析处理部52所输出的修饰被修饰树201在用于回答确定的机器学习中被用作属性。

[0062] 作为基于修饰被修饰树201的属性之一,有子树的n-gram。参考图2对其进行说明。现在修饰被修饰树201包含子树70。子树70具有4个节点A、B、C以及D。在这些单词当中假设将D设为对因果关系而言重要的单词。这种情况下,作为从子树70生成的n-gram,考虑由图2(B1)的子树构成的Tri-gram72、和由图2(B2)的子树构成的Bi-gram74。当然,若有紧接节点D的节点E(未图示),则也能将B-D-E以及C-D-E这样的Tri-gram作为属性采用。以上是和从子树得到的n-gram相关的简单的说明。另外,n-gram并不限定于Tri-gram或Bi-gram。

[0063] 再度参考图1,提问应答系统30进一步包括:关联文档检索部54,其使用修饰被修饰解析处理部52所输出的修饰被修饰树、和从提问34得到的信息,利用已有的信息检索手法从对象文档存储部32中检索、提取认为与提问34关联的文档;和回答候补提取部56,其用于从关联文档检索部54所提取的文档中,也是利用已有的手法提取多个回答候补202。

[0064] 在关联文档检索部54进行的文档的检索中,能利用已有的各种手法。例如能应用将从提问提取的内容词作为关键词的文档检索的手法。或者,也可以使用非专利文献1的第3.2节所记载的、提取与提问关联的文档的技术。在本实施方式中,使用被称作Solr(<http://lucene.apache.org/solr/>)的开源软件。

[0065] 回答候补提取部56从包含于由关联文档检索部54检索出的文档的语句中提取将连续的5个语句作为1组的回答候补。如图5所示那样,回答候补提取部56从1个文档中,如最初的5句的组120、第2个5句的组122、…、倒数第2个5句的组130以及最后的5句的组132那样提取多个5句的组。在这些5句的组中,相邻的2个组共有4个语句。例如最初的5句的组120和第2个5句的组122共有第2句-第5句。之所以如此将5个句设为1组是因为,是因为有因果关系的原因部分和结果部分存在于不同的句中的可能性。在本实施方式中,回答候补提取部56进一步以与提问34的比较来评价这样从各文档得到的众多的5句的组,选择有力的5句的组作为回答候补。在本实施方式中,将选择的组的数量设为20个。在回答候补的提取中也能利用各种手法。在本实施方式中使用非专利文献1所记载的手法。

[0066] 非专利文献1所记载的方法若简单进行说明,则如以下那样。首先,将非事实类提问分类为6个类型的任意者。这些类型如以下那样。

[0067] (1) 与定义相关的提问:寻求针对事项的定义。

[0068] (2) 与理由相关的提问:寻求针对事实的理由。

[0069] (3) 与方法相关的提问:寻求与方法相关的说明。

[0070] (4) 与程度相关的提问:寻求与某些事项的程度相关的说明。

[0071] (5) 与变化相关的提问:寻求与变化的事物相关的说明。

[0072] (6)与细节相关的提问:寻求对与一连串的事态关联的细节的说明。

[0073] 分类通过提问句包含疑问词、或者包含特定的形容词或副词来进行。例如若提问句包含「なぜ(为什么)」,则能判定为该提问与理由相关。

[0074] 接下来,在回答候补中提取包含对应于提问的类型而确定的特定的语句、和包含于提问中的单词的回答候补。这时在本实施方式中,算出各回答候补的评分,基于评分选择上位的回答候补。评分的算出方法的详细记载于非专利文献1。总而言之,使用包含于提问中的单词当中、出现在全对象文档中的频度少的单词越多则回答候补则评分越高这样的评分算出方即可。在图6示出由回答候补提取部56提取的回答候补的示例。

[0075] 再度参考图1,提问应答系统30还包括:因果关系认识部60,其对由回答候补提取部56提取的上位的各个回答候补202确定包含于其中的因果关系的表现,输出在因果关系的原因部分和结果部分附加了标签的回答候补204;因果关系的线索单词存储部58,其存储因果关系认识部60确定回答候补中的因果关系时参考的成为因果关系的线索的单词;和因果关系相应性判定部64,其对由因果关系认识部60对在各回答候补中找到的各个因果关系(如果有的话)判定该因果关系是否作为针对提问34的回答而相称(因果关系相应性),输出附加了表示因果关系的相应性的相应性标记的回答候补260。

[0076] 提问应答系统30还包括:极性辞典存储部62,其存储在因果关系相应性判定部64判定因果关系的相应性时使用的谓语的极性;属性矢量生成部66,其使用由到因果关系相应性判定部64为止的各处理部对各回答候补和提问得到的信息,针对各回答候补生成给定的属性矢量300;和回答候补排序部68,其使用由属性矢量生成部66对各回答候补生成的属性矢量300、和预先通过机器学习而学习完毕的模型来对各回答候补分配评分并进行排序,将最上位的回答候补作为针对提问34的回答36输出。

[0077] 存放于图1所示的因果关系的线索单词存储部58的因果关系的线索单词,例如是「～ため(由于～)」、「～ので(因为～)」、「～理由は(理由是)」、「～原因となる(原因是)」、「～ことから(的缘故)」、「～ことで(因而)」、「～からだ(因为是～)」等。在本实施方式中,作为该线索单词以人工收集由251个单词(句)构成的表现,存储在因果关系的线索单词存储部58。例如图7所示的回答候补1中的「ため(由于)」这样的单词160、回答候补2中的「ため(由于)」这样的单词162都相当于线索单词。

[0078] 因果关系认识部60在回答候补中寻找因果关系的线索单词,若有线索单词,则基于该单词来确定通过单词联系起来的因果关系表现的原因部分和结果部分。在该确定中使用通过机器学习而预先学习完毕的CRF(Conditional Random Fields:条件随机场)模型。

[0079] 参考图9,因果关系认识部60包括:线索单词确定部220,其在输入的各回答候补中确定存储于因果关系的线索单词存储部58的单词;预先学习完毕的CRF模型222,其在确定了回答候补中的线索单词时,确定通过该单词联系起来的因果关系表现的原因部分和结果部分;和因果关系表现确定部224,其通过使用由线索单词确定部220确定的线索单词和回答候补来参考CRF模型222,来对回答候补赋予表示通过线索单词联系起来的因果关系表现的原因部分的开始位置以及结束位置的标签、和表示结果部分的开始位置以及结束位置的标签,作为回答候补204而输出。

[0080] CRF模型222如前述那样预先学习完毕,使得若给出线索单词和回答候补,就确定由该线索单词表征的因果关系表现的原因部分和结果部分。参考图9,CRF模型222的学习处

理部200包括:学习用文本存储部230,其是与提问应答系统30独立的装置,存储学习用文本;签注处理部232,其对基于存储于因果关系的线索单词存储部58的线索单词而在学习用文本中找到的各线索单词进行在其原因部分和结果部分人工附加签注的签注处理,来生成学习用数据;学习用数据存储部234,其存储签注处理部232所输出的学习用数据;和机器学习处理部236,其使用存储于学习用数据存储部234的学习用数据来进行CRF模型222的学习。

[0081] 因果关系相应性判定部64如以下所述那样,将从回答候补提取的结果部分的记载和提问句进行对照,来判定回答候补的结果部分(以下仅称作「结果部分」)是否作为提问句的回答而相称。

[0082] 如参考图4已经说明过的那样,并不是说在回答候补认识到的因果关系表现作为针对提问的回答全都相称。多数情况下,作为针对提问的回答而相称的因果关系表现,其结果部分和提问相似。为此在本实施方式中,使用在提问与结果部分间的(1)单词匹配以及(2)子树匹配,来判定因果关系表现是否作为针对提问的回答而相称。

[0083] (1) 单词匹配

[0084] 单词匹配,涉及提问和结果部分是否共有、共有何种程度的数量的内容词(包含名词以及动词)。可以说共有的内容词的数量越多,则提问和结果部分越相似。

[0085] (2) 子树匹配

[0086] 子树匹配是句法结构上的子树的匹配,具体地,关于图2(B1)以及图2(B2)所示那样的子树,能看到在提问的子树和结果部分的子树间有几个共通的部分。更具体地,能测算从两者的子树得到的单词n-gram在何种程度上共通。

[0087] 但仅以单词匹配以及子树匹配,有时难以判定因果关系是否作为给出针对提问的回答的因果关系而相称。例如如图4所示的示例那样,即使是因果关系的线索单词共通的因果关系,包含于两者中的单词共通,有时也并非作为给出针对提问的回答而相称的因果关系。即,在图4所示的示例中,提问的「津波が起こる(出现海啸)」和作为「回答」而示出的语句的「津波が発生する(发生海啸)」,根据「津波が発生するとき(发生海啸时)」必然可以说「津波が起こる(出现海啸)」这样的意义,能说是处于含意关系。与此相对,提问的「津波が起こる(出现海啸)」和作为「并非回答」而示出的语句的「津波を弱める(削弱海啸)」则不处于含意关系。前者作为回答而相称,但后者不相称。

[0088] 为了如此判定回答候补是否作为针对提问的回答而相称,需要关于提问中的表现和回答候补中的表现有无含意关系的认识,这是困难的任务。为此在本实施方式中,使用谓语的「极性」这样的概念来判定该含意关系。如前述那样,作为谓语的极性考虑「活性」以及「非活性」这2个。所谓谓语的极性为活性,是指如「Xを引き起こす(引起X)」、「Xが発生する(发生X)」、以及「Xが起こる(出现X)」等那样、以包含变量X的表现型式来表示使由变量X表征的事物的功能等成为「开启」(使功能等活性化)。例如符合「Xを引き起こす(引起X)」、「Xを使う(使用X)」、「Xを買う(购买X)」、「Xを進行させる(使X进展)」、「Xを輸入する(进口X)」、「Xが増える(增加X)」等。这些都是发挥由X表征的事物的功能的表现。所谓谓语的活性为非活性,是指以包含变量X的表现型式来表示使由变量X表征的事物的功能等成为「关闭」(使功能等非活性化)。例如符合「Xを防ぐ(防止X)」、「Xを捨てる(丢弃X)」、「Xが減る(减少X)」、「Xを破壊する(破坏X)」、「Xが不可能になる(X变得不可能)」等。

[0089] 参考图8,若说前述的图4的示例,「津波が起こる(出现海啸)」(名词+が+起こる(出现))和「津波が発生する(发生海啸)」(名词+が+発生する(发生)),都在名词「津波(海啸)」和修饰「津波(海啸)」的谓语(起こる(出现)、発生する(发生))的极性为活性的点上共通。另一方面,「津波が起こる(出现海啸)(名词+が+起こる(出现))」和「津波を弱める(削弱海啸)」(名词+を+弱める(削弱)),在名词「津波(海啸)」上共通,但前者的「起こる(出现)」的活性、和后者的「弱める(削弱)」为非活性,在这点上不同。如此,将谓语的极性基于其意义分类为活性/非活性,仅将提问句的名词以及修饰该名词的谓语的极性的组合、和结果部分的名词以及修饰该名词的谓语的极性的组合一致的表现认识为表征含意关系的表现。能利用该性质来判定因果关系是否作为针对提问的回答而相称。

[0090] 图1所示的因果关系相应性判定部64,使用这样的手法来判定各回答候补是否作为给出针对提问的回答的因果关系而相应。

[0091] 参考图10,因果关系相应性判定部64包括:谓语极性判定部270,其对因果关系认识部60所输出的各个回答候补204,使用存储于极性辞典存储部62的极性辞典,来判定包含于检测到的因果关系表现的结果部中的谓语的极性;一致判定部272,其使用由谓语极性判定部270判定的极性,来判定包含于结果部中的名词以及修饰该名词的谓语的极性的组合、和包含于提问中的名词以及修饰该名词的谓语的极性的组合是否一致;和相应性标记设定部274,其按照提问与回答候补的结果部间的单词匹配以及子树匹配的结果、和一致判定部272的判定结果,将表示回答候补是否作为针对提问的回答而相称的相应性标记置位或复位后附加在回答候补,并作为回答候补260输出。

[0092] 参考图11,回答候补排序部68包括:基于SVM的评分部302,其用预先学习完毕的SVM(Support Vector Machine,支持向量机)评价由属性矢量生成部66生成的属性矢量300,输出表示作为针对提问的回答的合格性的评分;评分存储部304,其将由评分部302对各候补输出的评分和各候补一起存储;和最上位候补输出部306,其将存储于评分存储部304的全部回答候补按评分的降序排序,将具有最高评分的回答候补的因果关系表现的结果部分作为针对提问的回答36输出。

[0093] 在图12中作为一览表示出由属性矢量生成部66生成的属性的一览当中的和本申请发明没有直接关系的属性。这些属性在非专利文献3所记载的系统中使用。即,将作为属性矢量生成部66所生成的属性矢量的要素的属性如非专利文献3记载的那样,在现有技术中划分为根据词素解析和修饰被修饰解析处理的结果算出的属性的组、和与以下所述的因果关系关联而算出的属性的组这2组。

[0094] 在本实施方式中,作为和图12所述的属性不同的组的属性,使用以下。

[0095] (1) 因果关系的有无

[0096] 是回答候补是否包含判定为与回答相称的因果关系的二值特性。该值通过因果关系相应性判定部64的输出而得到。

[0097] (2) 因果关系的内容

[0098] 是回答候补内的因果关系的单词以及子树的n-gram、和谓语的活性/非活性的意义极性。它们通过回答候补的词素串、修饰被修饰树、以及因果关系相应性判定部64的输出内的谓语极性判定部270(图10)而得到。

[0099] (3) 因果关系的可信度

[0100] 判定为与回答相称的因果关系表现的结果部分和提问共有的单词、子树、或活性/非活性极性的数量或它们的任意的组合。它们能根据词素解析处理部50以及修饰被修饰解析处理部52的输出、因果关系认识部60的输出、各回答候补的词素串以及子树等生成。

[0101] 评分部302的学习如以下那样进行。

[0102] 参考图11的右侧,和提问应答系统30的学习处理部200分开准备评分部302的学习处理部290。该学习处理部290包括存储学习用的文档的学习用文档存储部310。学习用文档包含多个学习用数据集。各学习用数据集容纳提问、和成为针对该提问的回答候补的表征因果关系的多个语句。对各句附加表示是否应将包含于该句中的因果关系表现的结果部作为同一学习用数据集的提问的回答的指导数据。

[0103] 学习处理部290还包括:属性矢量生成部312,其对存储于学习用文档存储部310的各学习用数据集的各个因果关系,通过进行与提问应答系统30同样的处理来算出与属性矢量生成部66相同的属性,生成、输出附加了指导数据的学习用的属性矢量;学习用数据存储部314,其存储由从属性矢量生成部312输出的学习用的属性矢量构成的学习用数据;和SVM学习部316,其使用存放于学习用数据存储部314的学习用数据,来进行评分部302的SVM的学习处理。

[0104] 通过使用与实际的提问应答系统相同的属性来进行图11所示的评分部302的学习,评分部302在被给予从各回答候补得到的属性矢量时,能对该回答候补作为回答合适到何种程度进行评分,并输出。

[0105] [动作]

[0106] 以上说明了结构的提问应答系统30如以下那样进行动作。首先说明为了提问应答系统30进行动作的准备阶段中的提问应答系统30的动作。参考图1,在对象文档存储部32中预先存储众多成为检索回答的对象的文档。对包含于各文档中的各个语句进行词素解析和修饰被修饰解析,附加修饰被修饰树的信息。

[0107] 在因果关系的线索单词存储部58中,预先存储为了确定因果关系表现的原因部分和结果部分而使用的表现即线索单词(或语句)。另外,在极性辞典存储部62中存放关于众多的谓语记述其谓语的极性(活性/非活性)的极性辞典。

[0108] 参考图9,在学习处理部200的学习用文本存储部230中预先存储学习用的文本。由用签注处理部232对该各个文本人工赋予表示存储于因果关系的线索单词存储部58的线索单词的位置、和通过该线索单词联系起来的因果关系表现的原因部分和结果部分的范围的签注。将被赋予签注的语句作为学习用数据存储于学习用数据存储部234。机器学习处理部236使用该学习数据来进行CRF模型222的学习。

[0109] 参考图11,在评分部302的学习处理部290中,在学习用文档存储部310中存储多个学习用数据集。各学习用数据集包含提问、和成为针对该提问的回答候补的表征因果关系的多个语句。对各句附加表示是否应将包含于该句中的因果关系表现的结果部作为同一学习用数据集的提问的回答的指导数据。学习处理部290的属性矢量生成部312对存储于学习用文档存储部310的各学习用数据集的各个因果关系,通过进行与提问应答系统30同样的处理来算出与属性矢量生成部66相同的属性,通过附加指导数据来生成学习用的属性矢量,并输出到学习用数据存储部314。学习用数据存储部314存储由从属性矢量生成部312输出的学习用的属性矢量构成的学习用数据。SVM学习部316,使用容纳于学习用数据存储部

314的学习用数据,来进行评分部302的SVM的学习处理。其结果,评分部302在被给予从各回答候补得到的属性矢量时,能将该回答候补作为回答合适到何种程度作为评分而输出。通过以上,提问应答系统30能针对提问34输出合适的回答36。

[0110] 参考图1,在将提问34给予提问应答系统30时,词素解析处理部50对提问34进行词素解析,修饰被修饰解析处理部52进一步进行提问的修饰被修饰解析,将修饰被修饰树201附加给提问34,输出到关联文档检索部54。

[0111] 关联文档检索部54使用从修饰被修饰解析处理部52给予的信息,寻求认为与提问34有关联的文档,检索对象文档存储部32并进行提取。

[0112] 回答候补提取部56根据由关联文档检索部54提取的各文档,如参考图5说明的那样,每5句作为1组地生成众多的回答候补202。回答候补提取部56进一步从提取的回答候补202中选择认为是与提问34有关联的回答候补202。

[0113] 接下来,因果关系认识部60如以下那样进行动作。参考图9,线索单词确定部220对各回答候补202判定是否有存储于因果关系的线索单词存储部58的线索单词。在没有线索单词的情况下丢弃该候补,仅将有线索单词的回答候补给予因果关系表现确定部224。

[0114] 因果关系表现确定部224对从线索单词确定部220给予的各个回答候补,用CRF模型222来估计通过其中的线索单词联系起来的因果关系表现的原因部分和结果部分,附加表示各自的开始位置以及结束位置的标签后作为回答候补204输出。

[0115] 参考图10,接受到回答候补204的因果关系相应性判定部64的谓语极性判定部270,参考极性辞典存储部62来判定包含于回答候补204中的因果关系表现的结果部分的谓语的极性,将其附加在回答候补204后输出到一致判定部272。一致判定部272判定回答候补的结果部分的名词以及修饰该名词的谓语的极性的组合、和包含于提问中的名词以及修饰该名词的谓语的极性的组合是否一致,输出其结果。相应性标记设定部274按照一致判定部272的输出、和提问与回答候补的结果部的单词匹配以及子树匹配的结果,将相应性标记置位或复位,将该标记附加在回答候补后作为回答候补260输出。

[0116] 图1所示的属性矢量生成部66,给予附加在回答候补260中的各种信息来生成在图11所示的回答候补排序部68的评分部302中所用的属性矢量,给予评分部302。

[0117] 图11所示的评分部302使用通过学习处理部290学习完毕的SVM,算出针对被给予的属性矢量的评分。评分存储部304将对各回答候补算出的评分和回答候补一起存储。最上位候补输出部306在针对全部回答候补的评分的算出结束时,选择评分最高的回答候补,作为回答36输出。

[0118] [实验]

[0119] 为了研究为什么型提问应答系统中的因果关系属性的影响,研究对提问使用各手法得到的最上位回答的精度。将其结果在下面的表1示出。

[0120] [表1]

[0121] 表1

[0122]

手法	最上位回答的精度
非专利文献1	22.2
非专利文献3	37.4

仅因果关系属性	27.8
非专利文献3+因果关系属性	41.8

[0123] 非专利文献1的手法相当于上述实施方式中在回答候补提取部56的手法。非专利文献3在现有技术的手法当中得到最高的精度,为了和上述实施方式的比较而列举。

[0124] 根据该结果,根据上述实施方式所涉及的手法(使用在非专利文献3的属性中加入因果关系属性的技术对回答候补进行排序),最上位回答的精度大幅超过现有技术中精度最高的非专利文献3的技术。

[0125] 进而,为了研究使用上述实施方式所涉及的手法、非专利文献3所涉及的手法而得到的回答的可靠性,关于各个手法,将提问按其回答的精度从高到低的顺序排列,研究在上位25%的点的回答的精度。在图13示出其结果。

[0126] 参考图13,根据非专利文献3的手法的情况下得到的回答的精度每个顺位的值(图表322),在上位25%的点(图13中由一点划线326示出)的回答的精度为62%。与此相对,在上述实施方式所涉及的手法(图表320)中,该值为83%。为了参考,在图13中还示出在排序中仅用因果关系的情况下的图表324。

[0127] 如以上那样,根据上述实施方式,通过除了现有技术所使用的用于回答候补的排序的属性以外,再使用因果关系属性,能大幅提高针对非事实类的提问的回答的精度。因此,通过使用上述那样的因果关系属性,能得到针对非事实类的提问给出合适的回答的提问应答系统。

[0128] [计算机的实现]

[0129] 上述实施方式所涉及的提问应答系统30能通过计算机硬件、和在该计算机硬件上执行的计算机程序实现。图14表示该计算机系统330的外观,图15表示计算机系统330的内部构成。

[0130] 参考图14,该计算机系统330包括:具有存储器端口352以及DVD(Digital Versatile Disc,数字多功能盘)驱动器350的计算机340、键盘346、鼠标348、和监视器342。

[0131] 参考图15,计算机340除了存储器端口352以及DVD驱动器350以外,还包括:CPU(中央处理装置)356、CPU356、与存储器端口352以及DVD驱动器350连接的总线366、存储引导程序等的只读存储器(ROM)358、与总线366并存储程序命令、系统程序以及工作数据等的随机存取存储器(RAM)36、和硬盘354。计算机系统330还包括提供连接到能与其他终端通信的网络368的网络接口(I/F)344。

[0132] 用于使计算机系统330作为上述的实施方式所涉及的提问应答系统30的各功能部发挥功能的计算机程序,存储于装备在DVD驱动器350或存储器端口352的DVD362或者可移动存储器364中,进而转发给硬盘354。或者,程序也可以经过网络368发送给计算机340,存储在硬盘354。程序在执行时载入到RAM360。也可以从DVD362、从可移动存储器364或经由网络368直接将程序载入到RAM360。

[0133] 该程序包含用于使计算机340作为上述实施方式所涉及的提问应答系统30的各功能部发挥功能的多个命令所构成的命令串。通过在计算机340上动作的操作系统、或第三方的程序或安装在计算机340的各种编程工具包或程序库,提供使计算机340进行该动作所需要的若干基本功能。因此,该程序自身不一定非要包含为了实现在实施方式的系统以及方法而需要的全部功能。该程序仅包含通过以命令当中为了得到所期望的结果而进行控制的做

法调用合适的功能或编程工具包内的合适的程序工具、来实现作为上述的系统的功能的命令即可。当然,也可以仅用程序提供需要的全部功能。

[0134] 在上述实施方式中,对象文档存储部32、因果关系的线索单词存储部58、极性辞典存储部62、学习用文本存储部230、学习用数据存储部234以及314、CRF模型222、学习用文档存储部310、评分存储部304等通过RAM360或硬盘354实现。也可以将这些值进一步存储在USB存储器等可移动存储器364中,也可以经由网络368等通信介质发送给其他计算机。

[0135] 上述实施方式所涉及的系统,能作为装入已有的音声提问应答系统的1个模块使用。能在将非事实类提问作为输入给予时,为了从Web等的文本以高精度确定回答而利用上述实施方式的系统。上述系统并不限于提问应答系统,还能利用作为检索引擎以及信息分析系统的模块来向用户提供有用的信息。

[0136] 在上述实施方式中,在图1的关联文档检索部54中使用开源软件Solr。但本发明并不限于这样的实施方式。例如,也可以使用将出现在文档中的因果关系作为线索的手法,还可以使用将出现在文档中的因果关系、和出现在因果关系中的谓语的活性/非活性的组合作为线索的手法。一般,也可以使用将提问和文档使用它们分别包含的单词以多维的矢量的形式进行表现,通过计算矢量间的类似度来检索包含与提问类似的内容的文档,进一步从其中以上述的因果关系为线索锁定回答候补的手法。用于检索与提问关联的文档的软件工具不管有偿还是无偿都大量存在。在此,由于需要对提问高速做出回答,因此期望不是接受到提问后从各文档的开头扫描到末尾的所谓全文检索型,而是使用对预先收集好的大量的文档附加了索引的索引型。

[0137] 另外在上述实施方式中,作为非事实类提问设想为什么型提问,为了得到回答候补而将出现在文档中的因果关系作为线索。但本发明并不限于这样的实施方式。例如在属于同一非事实类提问的「怎样做苦瓜杂炒?」这样的怎样型提问的情况下,为了得到回答候补,也可以将「在苦瓜杂炒中放入猪肉或午餐肉、鸡蛋」这样的材料关系、「将苦瓜劈成两瓣,去掉当中的籽,切薄片」这样的步骤关系等意义上的关系作为线索。

[0138] 另外在上述实施方式中,在回答的选择中使用出现在提问中的名词与谓语的极性的组合、和出现在回答候补中名词与谓语的极性的组合是否一致,作为极性,使用谓语是否为活性(是否表示以包含变量X的表现型式让由变量X表征的事物的功能等「开启」(使功能等活性化))。但本发明并不限于这样的实施方式。还能用其他的极性。例如作为极性还能使用「good/bad」。在疾病的情况下,「病気がひどくなる(疾病变得严重)」这样的表现的「ひどくなる(变得严重)」,在活性/非活性这样的极性是「活性」,但在「good/bad」这样的极性中是「bad」。如此,作为谓语的极性不仅能使用活性/非活性,还能使用其他性质。另外,在上面的示例中,如「ひどくなる(变得严重)」为「活性」且「bad」那样,还能将多种类的极性的组合在回答选择时用作属性。

[0139] 进而在上述实施方式中,为了使理解容易,以「名词+谓语」的组合在提问以及结果部内仅存在1个的情况为例。但本发明并不限于这样的实施方式。例如在「タバコがガンを引き起こす(吸烟引起癌症)」这样的情况下,作为「名词+谓语」的组合,能得到「タバコが引き起こす(吸烟引起)」和「ガンを引き起こす(引起癌症)」这2个表现。可以仅将它们当中的特定的表现在属性生成时使用(例如仅采用名词相当于谓语的主部的表现,或者仅采用相当于宾语的表现,或者仅采用包含于提问中的名词和谓语的关系一致的表现),还可以将

全部表现在属性生成时使用。

[0140] 本次公开的实施方式仅是例示,本发明并不受上述的实施方式的限制。本发明的范围在参酌发明的详细的说明的记载的基础上,由权利要求的各权项示出,包含在与那里所记载的字句等同的意义以及范围内的全部变更。

[0141] 产业上的利用可能性

[0142] 本发明能运用在利用计算机以及网络对所谓的为什么型提问等的非事实类提问给出回答的提问应答服务、以及这些服务中所利用的装置中。

[0143] 标号的说明

[0144]	30	提问应答系统
[0145]	32	对象文档存储部
[0146]	34	提问
[0147]	36	回答
[0148]	50	词素解析处理部
[0149]	52	修饰被修饰解析处理部
[0150]	54	关联文档检索部
[0151]	56	回答候补提取部
[0152]	58	因果关系的线索单词存储部
[0153]	60	因果关系认识部
[0154]	62	极性辞典存储部
[0155]	64	因果关系相应性判定部
[0156]	66	属性矢量生成部
[0157]	68	回答候补排序部
[0158]	200	学习处理部
[0159]	201	修饰被修饰树
[0160]	202、204、260	回答候补
[0161]	220	线索单词确定部
[0162]	222	CRF模型
[0163]	224	因果关系表现确定部
[0164]	270	谓语句极性判定部
[0165]	272	一致判定部
[0166]	274	相应性标记设定部
[0167]	300	属性矢量
[0168]	302	评分部
[0169]	306	最上位候补输出部

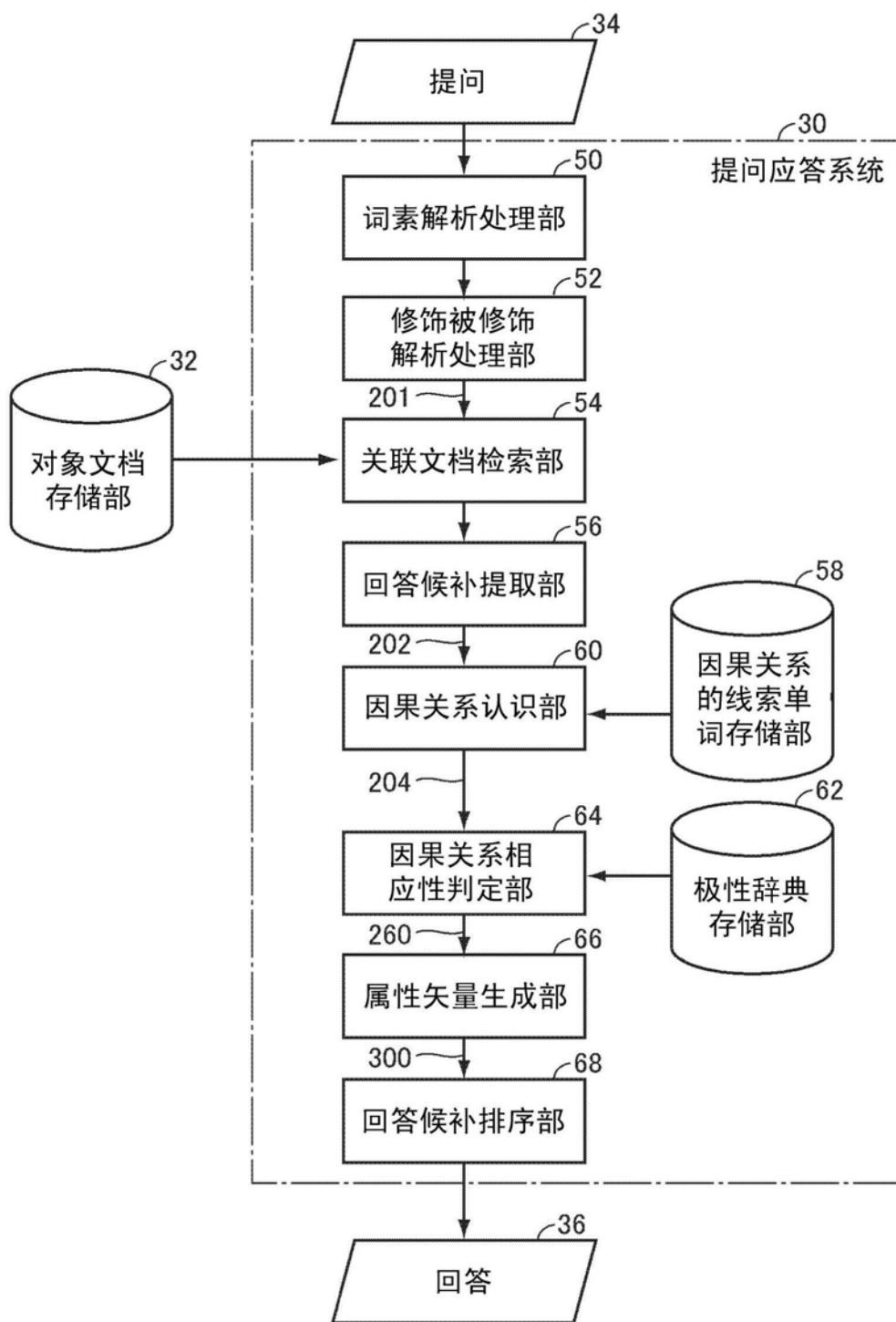


图1

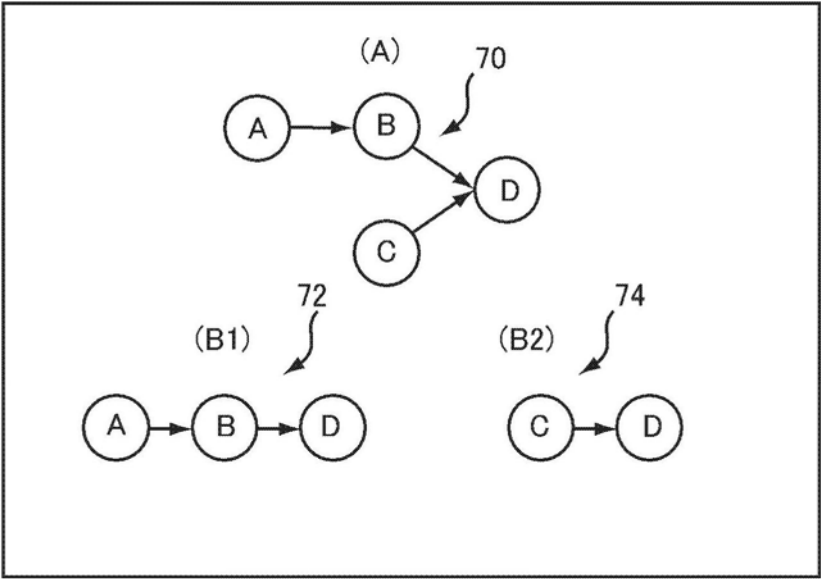


图2

因果关系的形式	表现例
句内的名词短语间的因果关系	地震が原因となる津波 80 84 82 (地震成为原因的海啸)
句内的名词短语和动词短语间的因果关系	海底地震が原因で津波が起きる。 86 90 88 (以海底地震为原因引发海啸。)
句内的节和动词短语间的因果关系	地震で陥没した海底に海水が流れ込んで水位 格差が起こるため津波が発生します。 92 96 94 (由于海水流入因地震下沉的海底 而出现水位高低差，因此发生海啸。)
跨2句出现的因果关系	地震が海底で起きた場合、海底の動きに伴って 海水面が盛り上がり、一気に崩れていきます。 これが原因で津波が発生します。 98 100 102 (在海底引发地震的情况下，伴随海底的运动而 海平面涌起，一口气崩溃。以此为原因发生海啸。)

图3

提问：（为什么出现海啸？）	
回答	地震で陥没した海底に海水が流れ込んで水位 格差が起こるため津波が発生します。 92 96 94 （由于海水流入因地震下沉的海底 而出现水位高低差，因此发生海啸。）
不是回答	海岸林は津波のエネルギーを減少させ、水位 速度を下げるため、津波を弱める効果がある。 110 114 112 由于海岸林减少海啸的能量，降低 水位速度，因此有削弱海啸的效果。

图4

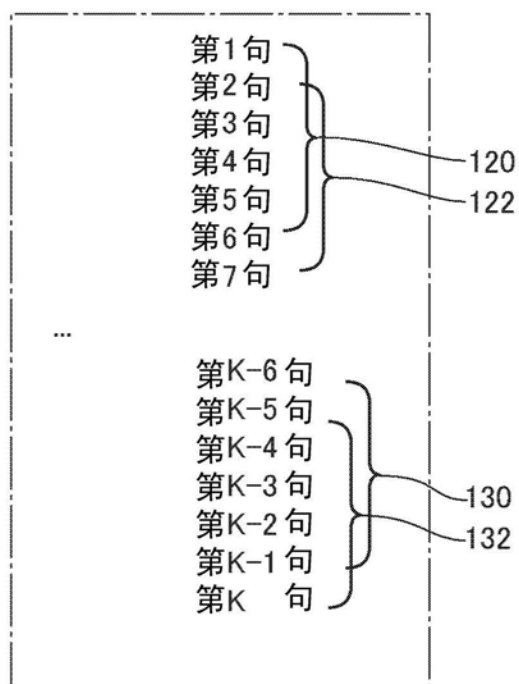


图5

202

- 回答候補 1 …ことが分かっています。地震で陥没した海底に海水が流れ込んで水位格差が起こるため津波が発生します。したがって…
(获知……。由于海水流入因地震下沉的海底而出现水位高低差，因此发生海啸。因此……)
- 回答候補 2 …が植林されている。海岸林は津波のエネルギーを減少させ、水位速度を下げるため、津波を弱める効果がある。しかし…
(……进行植树。由于海岸林减少海啸的能量，降低水位速度，因此有削弱海啸的效果。但是……)
- 回答候補 3 …が発生することが懸念されます。地震が海底で起きた場合、海底の動きに伴って海水面が盛り上がり、一気に崩れていきます。これが原因で津波が発生します。このような津波が…
(起、発生……。在海底引发地震的情况下，海水面隆起，一口气崩溃。海啸……)

图6

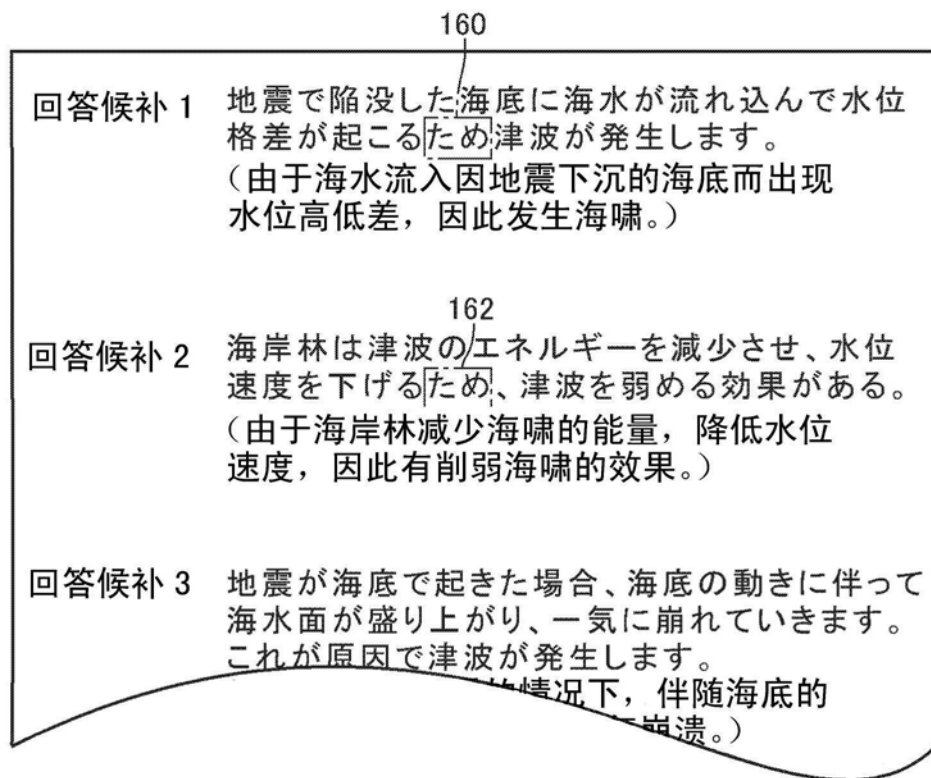


图7

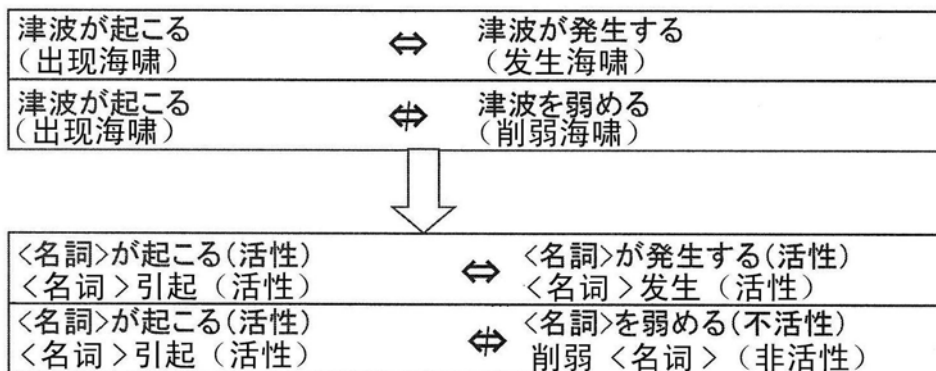


图8

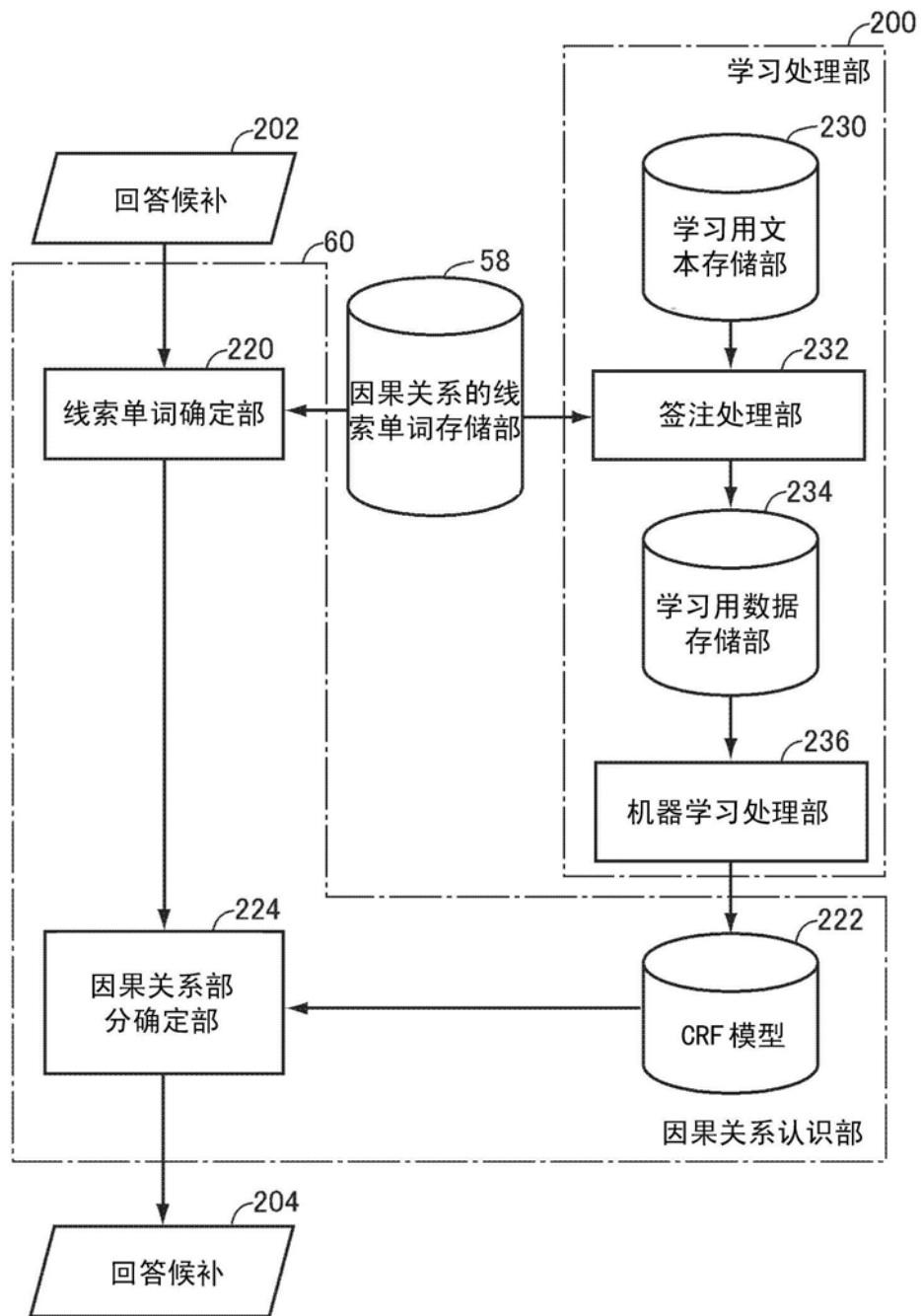


图9

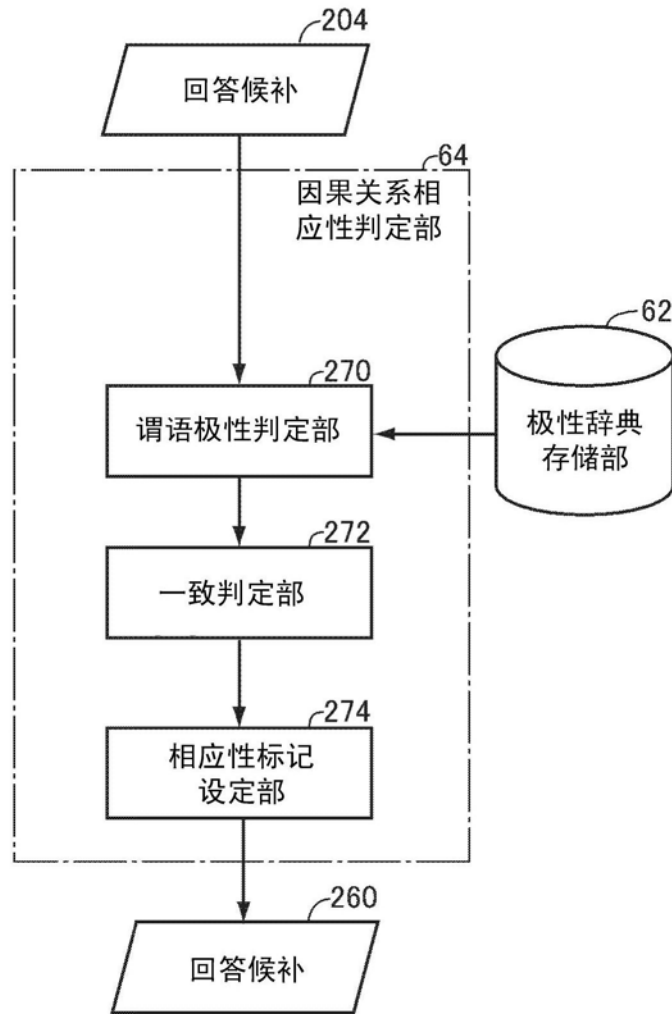


图10

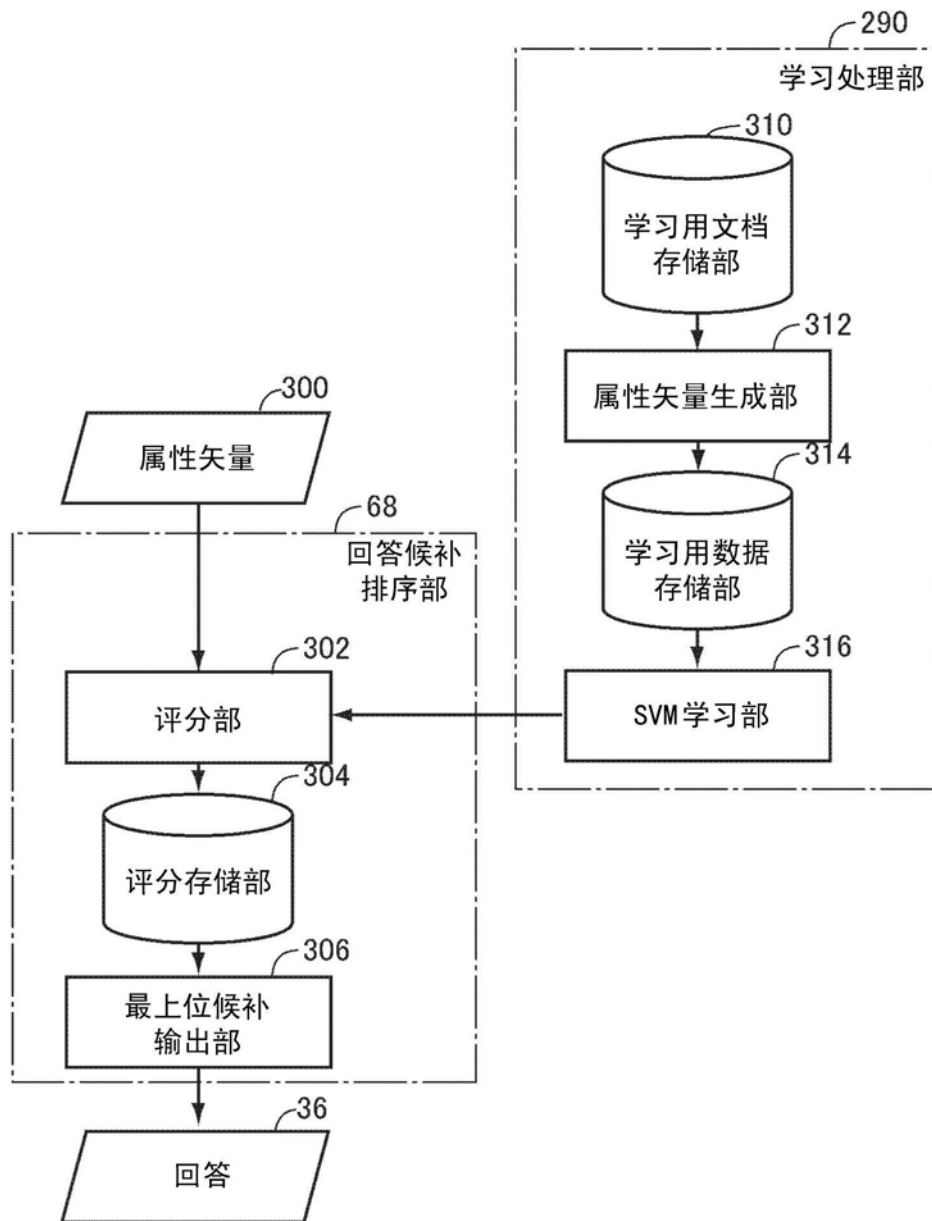


图11

MSA1	提问以及其回答中的词素 n-gram、句节 n-gram 以及语法结构依赖 n-gram，其中 n 为 1 到 3。提问中的 n-gram 和回答中的 n-gram 区别。
MSA2	回答候选补中的 MSA1 的 n-gram 当中、包含提问中的词的 n-gram。
MSA3	MSA1 的 n-gram 当中、包含理由、原因以及要因等成为线索的单词的 n-gram。提问中的这些 n-gram 和回答候选补中的这些 n-gram 区别。
MSA4	提问中的词中存在于回答候选补中的词相对于提问中的词的总数的比例。
SWC1	存在于提问、和其回答候选补中的单词类 n-gram。提问中的 n-gram 和回答候选补中的 n-gram 区别。
SWC2	回答候选补的 SA@W1 的 n-gram 当中、原始的 MSA1 的 n-gram 包含提问中的词的 n-gram。
SA@W1	提问和其回答候选补当中的单词的极性的 n-gram。提问中的 n-gram 和回答候选补中的 n-gram 区别。
SA@W2	回答候选补的 SA@W1 的 n-gram 当中、原始的 MSA 的 n-gram 包含提问中的词的任意者的 n-gram。
SA@W3	提问和其回答候选补中的类 / 极性连接 n-gram。提问中的 n-gram 和回答候选补中的 n-gram 区别。
SA@W4	回答候选补的 SA@W1 的 n-gram 当中、原始的 MSA1 的 n-gram 包含提问中的任意的词的 n-gram
SA@P1	在提问中的评价短语与回答候选补中的评价短语间极性是否一致。若这样的评价短语的任意的任意者一致则为 1，若不一致则为 0。
SA@P2	SA@P1 为 1 那样的评价短语的任意的极性（正或负）
SA@P3	将评价短语的词素 n-gram、句节 n-gram 以及语法依赖的 n-gram 和短语极性结合。 其中 n 为 1 到 3。提问中的这些 n-gram 和回答候选补中的这些 n-gram 区别。
SA@P4	包含提问中的词的回答候选补的 SA@P3 的 n-gram。
SA@P5	回答候选补中的位于具有评价短语的句中提问句中的词的数量相对于提问中的全部词的数量之比。
SA@P6	将评价短语中的单词类 n-gram 和评价短语的极性组合。提问中的 n-gram 和回答候选补中的 n-gram 区别。
SA@P7	原始的 MSA1 的 n-gram 包含提问中的任意的词的回答候选补中的 SA@P6 的 n-gram。
SA@P8	将提问中以及回答候选补中的评价短语中的类 / 极性结合 n-gram、和该评价短语的短语极性组合。提问中的 n-gram 和回答候选补中的 n-gram 区别。
SA@P9	回答候选补的 SA@P8 的 n-gram 当中、原始的 MSA1 的 n-gram 包含提问中的任意的词的 n-gram
SA@P10	SA@P6 的 n-gram 的对，即来自提问中的评价短语的 n-gram、和来自其回答候选补中的评价短语的 n-gram 的组合。其中，这 2 个评价短语评价的方向必须相同。

图12

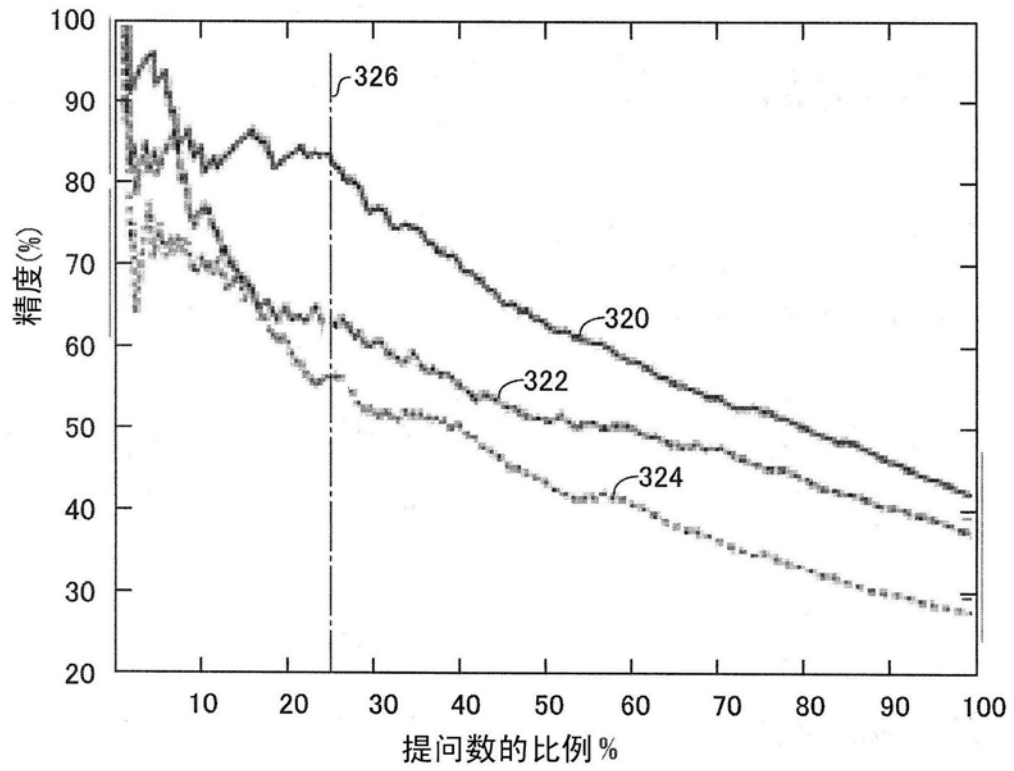


图13

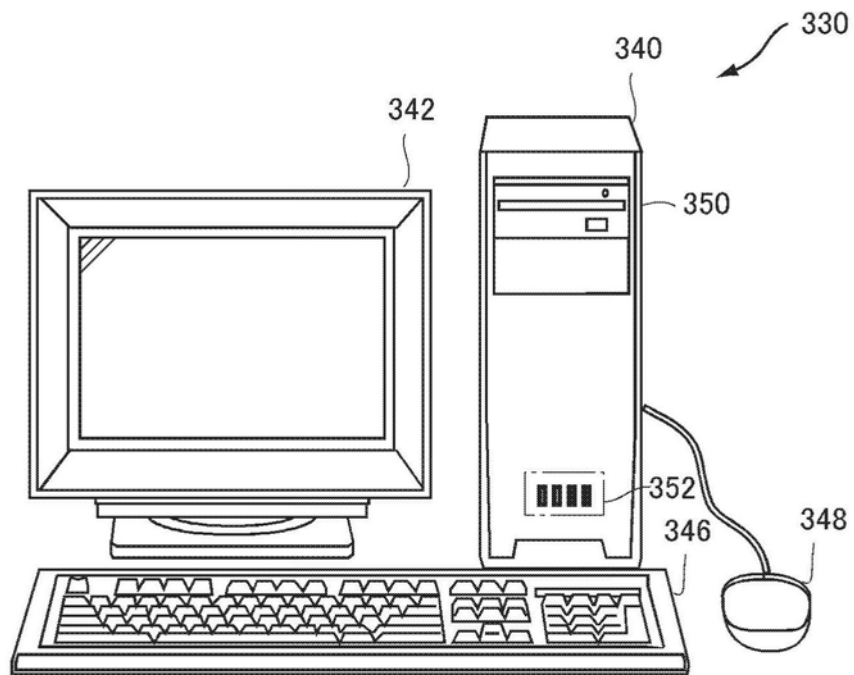


图14

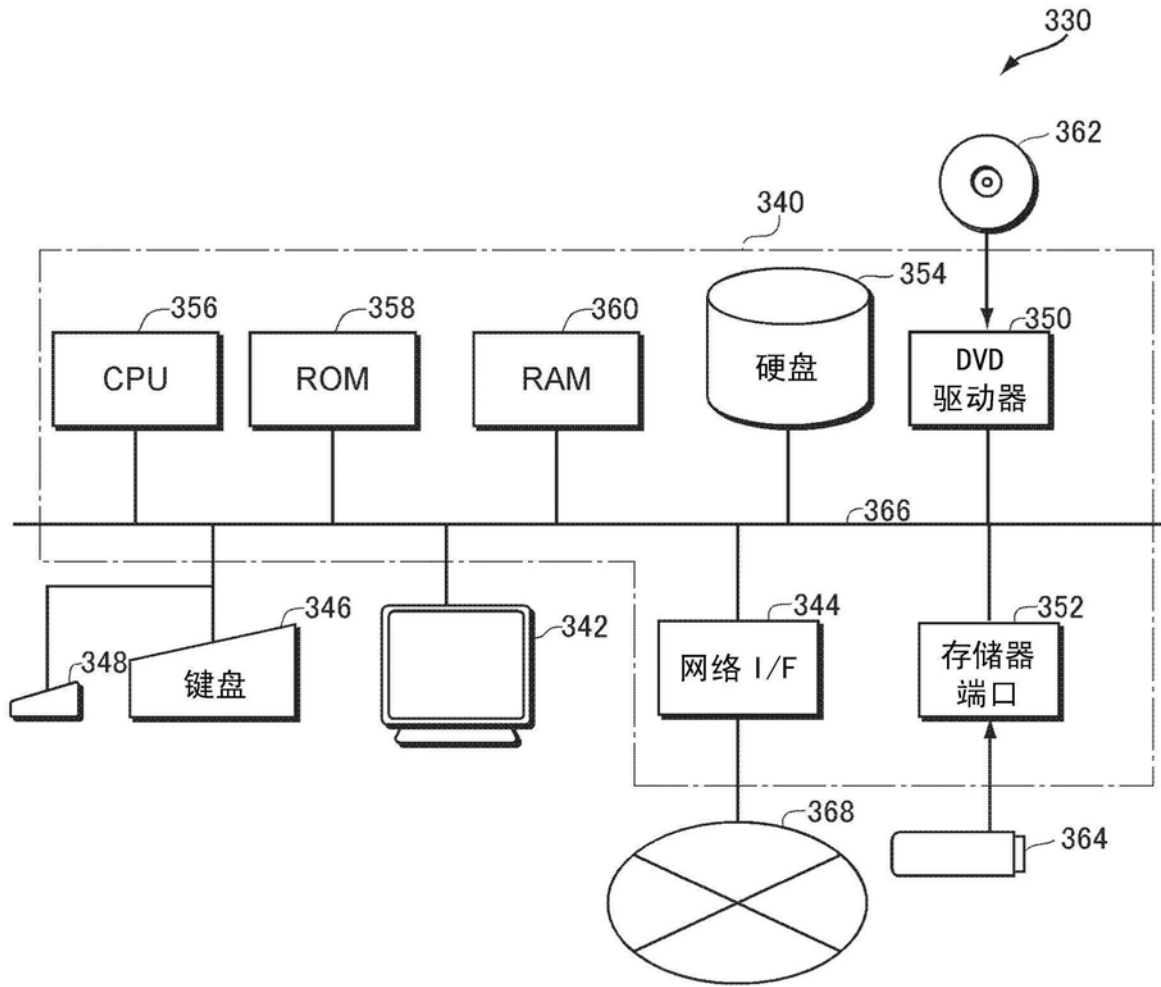


图15