(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116992022 A
(43)申请公布日 2023.11.03

(21)申请号 202310484570.1

(22)申请日 2023.04.28

(30)优先权数据
102022204296.4 2022.05.02 DE

(71)申请人 罗伯特·博世有限公司
地址 德国斯图加特

(72)发明人 L·朗格 J·施特根
H·阿德尔-武

(74)专利代理机构 中国专利代理(香港)有限公司 72001
专利代理师 刘晗曦 刘春元

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 40/30(2020.01)
G06N 3/0455(2023.01)
G06N 3/0464(2023.01)
G06N 3/044(2023.01)
G06N 3/08(2023.01)
G06N 5/02(2023.01)

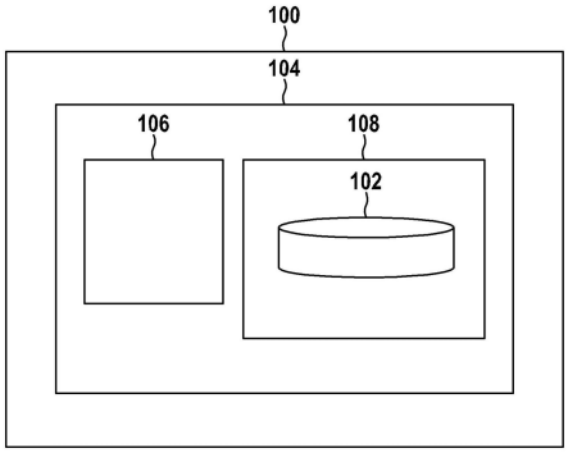
权利要求书2页 说明书8页 附图2页

(54)发明名称

用于处理来自非结构化文本的时间表达的设备和方法

(57)摘要

用于处理来自特别是非结构化文本的时间表达以填充知识数据库(102)的方法和设备,其中,在文本中确定时间表达,其中根据文本而确定时间表达的类型,并且其中所述时间表达和所述类型被映射到所述时间表达的与上下文无关的表示中的时间表达的值的预测。



CN 116992022 A

1.用于处理来自特别是非结构化文本的时间表达以填充知识数据库(102)的计算机实现的方法,其特征在于,在文本中确定时间表达(304),其中根据文本而确定所述时间表达的类型(306),并且其中所述时间表达和所述类型被映射到所述时间表达的与上下文无关的表示中的时间表达的值的预测(308)。

2.根据权利要求1所述的方法,其特征在于,模型被设计成,将所述时间表达和所述类型映射到所述预测,其中,所述时间表达和所述类型利用所述模型被映射到所述预测(308)。

3.根据权利要求1或2所述的方法,其特征在于,提供文本(300),所述文本包括所述时间表达,其中针对所述文本而确定或预给定参考日期(302),其中,根据与上下文无关的表示以及根据所述参考日期而确定时间属性(310),并且其中用所述时间属性填充所述知识数据库(312)。

4.根据权利要求2或3中任一项所述的方法,其特征在于,提供时间表达(202),其中提供与上下文无关的表示(204),并且其中根据所述时间表达和所述与上下文无关的表示来训练所述模型,以将所述时间表达映射到针对与上下文无关的表示的预测(208)。

5.根据权利要求4所述的方法,其特征在于,所述与上下文无关的表示包括多个插槽,其中通过所述插槽的内容与上下文无关地表示所述时间表达,其中至少一个插槽被掩码(206),并且其中所述模型根据时间表达被训练用于,将所述时间表达映射到针对所述至少一个被掩码的插槽的被掩码的内容的预测(208)。

6.根据权利要求5所述的方法,其特征在于,在第一训练阶段中有至少一个插槽未被掩码,并且在所述第一训练阶段中有至少一个插槽被掩码,所述模型在所述第一训练阶段中被训练用于,将所述时间表达和所述至少一个未被掩码的插槽的未被掩码的内容映射到针对所述至少一个被掩码的插槽的被掩码的内容的预测(208)。

7.根据权利要求6所述的方法,其特征在于,在第二训练阶段中特别是掩码所有插槽,在所述第一训练阶段中所训练的模型在所述第二训练阶段中被训练用于,将所述时间表达映射到针对所述插槽的内容的预测(208)。

8.根据权利要求4至7中任一项所述的方法,其特征在于,在多次迭代中训练所述模型,其中在第一次迭代中被掩码(206)的插槽要比在第二次迭代中更少,特别是比在所述第一次迭代之后执行的第二次迭代中更少。

9.根据前述权利要求中任一项所述的方法,其特征在于,利用所述模型确定数字或数值或运算符,特别是算术运算符,或字母,或单词成分,或来自预给定的词汇表的单词或占位符作为预测(208、308)。

10.根据权利要求9所述的方法,其特征在于,利用所述模型,为插槽的每个被掩码的内容确定词汇表上的概率分布,并且根据所述概率分布而确定所述插槽的内容(208、308)。

11.根据前述权利要求中任一项所述的方法,其特征在于,特别是取决于文本而基于规则地,从包含所述时间表达的文本中确定文本片段(304),并且其中根据包含所述时间表达的文本片段而确定利用模型的预测(308)。

12.根据前述权利要求中任一项所述的方法,其特征在于,特别是取决于文本而基于规则地,从包含所述时间表达的文本中确定文本片段(202),并且其中根据包含所述时间表达的文本片段而训练所述模型。

13.用于处理来自特别是非结构化文本的时间表达以填充知识数据库(102)的设备(100),其特征在于,所述设备(100)包括被设计用于执行根据权利要求1至12中任一项所述的方法的计算装置。

14.计算机程序,其特征在于,所述计算机程序包括计算机可读指令,当由计算机执行所述计算机可读指令时,运行根据权利要求1至12中任一项所述的方法。

用于处理来自非结构化文本的时间表达的设备和方法

技术领域

[0001] 本发明涉及用于处理来自非结构化文本的时间表达(temporaler Ausdrück)以填充知识数据库的设备和方法。

背景技术

[0002] 为了填充知识数据库可以设置:从非结构化文本中提取并且规范化(Normalisierung)时间表达。

[0003] 这两个子任务可以一起被解决,例如在Chang与Manning 2012年的“SUTime:A library for recognizing and normalizing time expressions”(https://aclanthology.org/L12-1122/)或Strötgen与Gertz 2013年的“Multilingual and cross-domain temporal tagging”(https://link.springer,com/article/10.1007/s10579-012-9179-y)中所描述的那样。

[0004] 这两个子任务可以彼此独立地执行,例如在Bethard 2013年的“A Synchronous Context Free Grammar for Time Normalization”(https://aclanthology.org/D13-1078/)或在Lee等人2014年的“Context-dependent Semantic Parsing for Time Expressions”(https://aclanthology.org/P14-1135/)中所描述的那样。

[0005] 这些规范化方法以及在Ning等人2018年的“CogCompTime:A Tool for Understanding Time in Natural Language”(https://aclanthology.org/D18-2013/)以及在Ding等人2021年的“Automatic rule generation for time expression normalization”(https://aclanthology.org/2021.findings-emnlp.269)所描述的规范化方法都是基于规则系统。

发明内容

[0006] 根据独立权利要求1所述的用于对来自非结构化文本的时间表达进行规范化以填充知识数据库的计算机实现的方法包括:用于对时间表达进行规范化的基于深度学习的方法。

[0007] 与基于规则的规范化相比,基于深度学习的用于规范化的方法的一个优势是:可以迁移到(Transfer)新的语言和文本流派(Textgenre),否则就只有通过大量耗费通过创建新规则才能涵盖(abdecken)所述语言和文本类型。

[0008] 与使用固定规则进行规范化相比,利用基于深度学习的用于规范化的方法,也可以对静态(statisch)规则系统未涵盖的时间表达进行规范化。这意味着:基于深度学习的用于规范化的方法与基于规则的方法相比在文本中的干扰源方面更灵活,所述干扰源例如是较小的正字错误和可以从上下文推断出其含义的未知词。

[0009] 基于深度学习的用于规范化的方法可以与不同的提取模型结合使用,而无需对其进行显式(explizit)训练。可以使用基于规则的方法或使用神经网络在示例性提取模型中进行提取。所述提取可以由用户进行,其中由用户自己为要规范化的表达设定边界

(Grenze)。

[0010] 用于处理特别是来自非结构化文本的时间表达以填充知识数据库的方法规定:在文本中确定时间表达,其中根据文本确定时间表达的类型,并且其中所述时间表达和类型被映射到所述时间表达的与上下文无关的表示中的所述时间表达的值的预测。时间表达和/或类型例如基于规则来确定或利用为此训练的神经网络根据文本来确定。

[0011] 优选地,模型被设计成,将时间表达和类型映射到所述预测,其中,所述时间表达和类型利用所述模型被映射到预测。例如,使用为此训练的转换器(Transformer)作为模型来确定时间表达的值,其中所述转换器被设计、即训练成,将所述时间表达和类型映射到预测。代替转换器,也可以设置为此训练的另一种模型,例如深度学习模型,特别是递归神经网络或卷积神经网络,其被设计、即训练成,将所述时间表达和类型映射到预测。

[0012] 优选地,提供文本,所述文本包括时间表达,其中针对所述文本而确定或预给定参考日期,其中,根据与上下文无关的表示以及根据所述参考日期而确定时间属性,并且其中,用所述时间属性填充知识数据库。由此,使用在该文本预给定的时间上下文中所锚定的时间属性来填充知识库。

[0013] 优选地,提供时间表达,其中提供与上下文无关的表示,并且其中根据时间表达和所述与上下文无关的表示来训练模型,以将时间表达映射到针对与上下文无关的表示的预测。由此产生的转换器在文本中的干扰源方面是灵活的,所述干扰源例如是正字错误和可以从上下文推断出其含义的未知词。产生的模型、例如转换器或深度学习模型以及对于与上下文无关的表示的确定能应用于多种语言。

[0014] 与上下文无关的表示优选地包括多个插槽(Slots),其中通过所述插槽的内容与上下文无关地表示所述时间表达,其中至少一个插槽被掩码(maskieren),并且其中所述模型根据时间表达被训练用于,将所述时间表达映射到针对所述至少一个被掩码的插槽的被掩码的内容的预测。由此,非常有效地训练所述模型,因为相比于与上下文相关的表示,需要较少的训练数据。

[0015] 在第一训练阶段中有至少一个插槽未被掩码,并且在所述第一训练阶段中有至少一个插槽被掩码,所述模型优选地在所述第一训练阶段中被训练用于,将所述时间表达和所述至少一个未被掩码的插槽的未被掩码的内容映射到针对所述至少一个被掩码的插槽的被掩码的内容的预测。由此而非常有效地训练所述模型。

[0016] 在第二训练阶段中特别是掩码所有插槽,优选地,在第一训练阶段中所训练的模型在第二训练阶段中被训练用于,将所述时间表达映射到针对所述插槽的内容的预测。由此,该模型得到了非常有效的训练,并学习了与上下文无关的表示的完整预测。

[0017] 优选地,在多次迭代中训练该模型,其中在第一次迭代中被掩码的插槽要比在第二次迭代中更少,特别是比在第一次迭代之后执行的第二次迭代中更少。由此,该模型非常有效地得以训练,尤其是当只存在少量训练数据、例如各种语言的训练数据时。

[0018] 优选地,利用该模型确定数字或数值或运算符,特别是算术运算符,或字母,或单词成分,或来自预给定的词汇表的单词或占位符作为针对与上下文无关的表示的插槽的预测。不同的词汇表可用于不同的与上下文无关的表示。所述词汇表是根据所期望的与上下文无关的表示预给定的。因此,通过该模型进行的预测是非常可靠的。

[0019] 利用该模型,优选地为插槽的每个被掩码的内容确定词汇表上的概率分布,并且

根据所述概率分布而确定插槽的内容。由此可以非常有效地训练模型，并可以基于概率分布进行进一步的分析或替代性预测。

[0020] 可以规定，特别是取决于文本而基于规则地从包含时间表达的文本中确定文本片段，并且其中根据包含时间表达的文本片段来确定利用模型的预测。例如，文本片段是一个句子。所述预测基于其上下文中的时间表达。这附加地改进了预测。

[0021] 可以规定，特别是取决于文本而基于规则地从包含时间表达的文本中确定文本片段，并且其中根据包含时间表达的文本片段训练所述模型。

[0022] 一种用于处理来自特别是非结构化文本的时间表达以填充知识数据库的设备包括：被设计用于执行该方法的计算装置。该设备具有与该方法相对应的优点。

[0023] 包括计算机可读指令的计算机程序具有与之相应的优点，其中当由计算机执行所述计算机可读指令时运行该方法。

附图说明

[0024] 在下面的描述和附图中可以得到另外的有利实施方式。其中：

[0025] 图1示出了用于处理来自非结构化文本的时间表达以填充知识数据库的设备的示意图，

[0026] 图2示出了用于处理来自非结构化文本的时间表达以填充知识数据库的方法中的训练步骤，

[0027] 图3示出了用于处理来自非结构化文本的时间表达以填充知识数据库的方法中的推理步骤。

具体实施方式

[0028] 在图1中，示意性地示出了用于处理来自特别是非结构化文本的时间表达以填充知识数据库102的设备100。

[0029] 针对时间表达的示例是例如日期说明，例如2022年1月和时间说明，例如13点，以及时间段，例如5个月。

[0030] 针对其中规定了知识数据库102填充的应用情况的示例是：创建或扩展知识图(英文：knowledge base(知识库))或具有时间属性的数据库。例如，可以根据这些时间表达来确定所述时间表达中这些条目(Eintrag)的有效期(Gültigkeitszeiträume) ，并且可以将代表它们的事件分配给特定的时间点。

[0031] 所述设备100包括计算装置104，其被设计成执行如下所述的方法，所述方法用于处理来自非结构化文本的时间表达以填充知识数据库102。

[0032] 计算装置104例如包括至少一个处理器106及至少一个存储器108。所述至少一个存储器108在本示例中设计用于储存知识数据库102。在示例中，所述至少一个存储器108被设计为存储指令，例如计算机程序，当由至少一个处理器106执行所述指令时运行该方法。

[0033] 在下文中，以时间表达的提取和随后的规范化为例来描述该方法。提取和规范化是“Temporal Tagging(时间标记)”的部分。

[0034] 在示例中，所述提取包括：分类(Typisierung)。给定文本的情况下，提取的任务是找到时间表达的边界。这通常是所谓的字符偏移量(Zeichen-Offset)，其通过两个数值来

设定文本内的表达的清晰边界。例如,对于长度为11个字符的表达,从第20个字符开始到第30个字符(含该字符在内)结束。

[0035] 例如,所述提取是基于规则进行的。例如,如果表达根据预给定规则而与针对时间表达的模式(Muster)相一致,则该表达被识别为时间表达。

[0036] 例如,所述提取是使用深度学习方法执行的,特别是通过解决序列标记(Sequence Labeling)任务或Seq2Seq问题执行的。

[0037] 在Laparra等人2018年的"From Characters to Time Intervals:New Paradigms for Evaluation and Neural Parsing of Time Normalizations"(https://aclanthology.org/Q18-1025/)以及在Lange等人2020年的"Adversarial Alignment of Multilingual Models for Extracting Temporal Expressions from Text"(https://aclanthology.org/2020.repl4nlp-1.14/)中描述了通过解决序列标记任务来进行提取。

[0038] 以包含时间表达的文本为例来描述该方法。

[0039] 在示例中,使用HeidelTime执行分类。例如,在Strötgen与Gertz2015年的"A Baseline Temporal Tagger for all Languages"(https://aclanthology.org/D15-1063/)中描述了HeidelTime。

[0040] 时间表达的分类在例如四个类(DATE、TIME、DURATION、SET)之一中进行。可以设置其他更多或更少的类。

[0041] 在示例中,时间表达的分类随着时间表达的提取而深入(eingehend)进行。还可以设置时间表达的提取和随后的分类。

[0042] 在示例中,针对每个时间表达,其在文本中的注释界限(Annotationsgrenzen)和类型被确定作为提取和分类的输出。

[0043] 例如,对于示例文本片段"Der Eta Aquariiden Meteorschauer wird im Mai erscheinen."(水瓶座艾塔流星雨将在五月出现),例如提取时间表达"Mai(五月)"并使用DATE类型被分类。

[0044] 在文本片段中,时间表达"五月"在示例中使用HeidelTime提取,并在文本片段本身中使用DATE类型进行注释。生成的经注释的文本片段是"Der Eta Aquariiden Meteorschauer wird im<TIMEX3type＝"DATE">Mai</TIMEX3>erscheinen."(水瓶座艾塔流星雨将在<TIMEX3 type＝"DATE">五月</TIMEX3>出现)。

[0045] 在规范化过程中,时间表达(在示例中为五月)由时间表达的与上下文无关的表示进行补充。在示例中,与上下文无关的表示与TIMEX 3表达中的时间表达一起使用,而不是与文本段中的时间表达一起使用。对于该示例,由此得出文本片段:

[0046] "Der Eta Aquariiden Meteorschauer wird im<TIMEX3 type＝"DATE"value＝"UNDEF-year-05">Mai</TIMEX3>erscheinen."(水瓶座艾塔流星雨将在<TIMEX3 type＝"DATE"value＝"UNDEF-year-05">五月</TIMEX3>出现)。

[0047] TIMEX3例如在James PustejovSky等人的"The specification language TimeML",In The language of time:a reader,第545-557页,Oxford University Press中被描述。

[0048] 在示例中,在HeidelTime中给时间表达"Mai(五月)"分配时间表达kuRepr＝"UNDEF-year-05"的与上下文无关的表示kuRepr,其中所述时间表达"Mai(五月)"的字母M在

文本中的位置为:START＝42,而其字母i在文本中的位置为:END＝44。

[0049] 在示例中,针对与上下文无关的表示kuRepr设置了TIMEX 3表达中的一个值。

[0050] 在示例中,使用包含TIMEX 3表达的文本片段作为针对规范化的输入参量。

[0051] 在示例中,所述规范化包括掩码语言建模(Masked Language Modeling)。这意味着:所述规范化被建模为掩码语言建模任务。例如,所述掩码语言建模包括:根据例如BERT等语言模型的转换器。

[0052] 例如,在Devlin等人2019年的“BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding”(https://aclanthology.org/N19-1423/)中描述了BERT。

[0053] 代替转换器,也可以设置为此而训练的另一种模型,例如深度学习模型,特别是递归神经网络或卷积神经网络。

[0054] 在掩码语言建模任务中,用掩码来对与上下文无关的表示中的特定区域进行掩码,并且转换器被训练以填充掩码。

[0055] 在转换器的训练中,转换器被训练用于,针对时间表达,根据该时间表达而确定对其与上下文无关的表示的预测。

[0056] 在示例中,针对利用转换器进行的与上下文无关的表示的预测,使用多个插槽。

[0057] 在示例中,通过插槽的内容表示:TIMEX 3表达中的值中的与上下文无关的表示kuRepr。

[0058] 在该示例中,插槽的内容包括与上下文无关的表示的一部分或占位符。在示例中,插槽具有预给定的、不可改变的长度。在示例中,插槽具有相同的长度。也可以定义不同长度的插槽。在示例中,这些插槽分别具有不同的语义。

[0059] 例如,对于日期说明,针对具有两位数的日说明和具有两位数的月说明设置各一个插槽,而针对年说明则设置两个插槽。例如,对于年说明,针对一年和十年设置一个两位数的插槽,针对百年和千年说明设置一个两位数的插槽。示例中定义了插槽的顺序。插槽的内容根据其对时间表达的重要性被明确分配。例如,小于一天的时间说明存储在三个另外的插槽中,一个具有两位数的小时说明,一个具有两位数的分钟说明,一个具有两位数的秒说明。可以规定:设置至少一个另外的插槽。所述至少一个另外的插槽的内容可以是函数调用(Funktionsaufruf)。函数调用可以由诸如“加”或“减”的运算符来标识。至少一个其他插槽的内容可以是针对函数调用的自变量(Argument)。所述至少一个另外的插槽的内容可以包含附加信息。所述附加信息例如是作为公元前或公元后的日期说明的标识,或者作为时间段或公共假期的说明的至少一个其他插槽的标识。例如,内容BC或AD是针对公元前或公元后的年说明的。例如,四个插槽的内容是:针对后天的this-day-plus-2或者针对昨天的this-day-minus-1。例如,插槽的内容表示假期。没有为其他表达、例如UNDEF设置插槽,因为在处理包含特定词、例如“下个”或“最后”的时间表达时会隐式(implizit)假定此表达。

[0060] 所述内容和插槽到与上下文无关的表达的部分的预给定分配允许重建与上下文无关的表达。

[0061] 未使用的插槽在示例中配备有占位符。未使用的插槽是对其而言在与上下文无关的表示中不包含任何值的插槽,或者对其而言转换器预测出在与上下文无关的表示中不包含任何值的插槽。

[0062] 在示例中使用了十一个插槽。可以规定使用其他数量的插槽。转换器被设计用于，根据时间表达预测插槽的内容。在示例中，转换器被设计用于，根据包含所述时间表达的文本片段来预测插槽的内容。在示例中，转换器被设计用于，根据包含TIMEX 3表达的文本片段预测插槽的内容。

[0063] 在训练中，用掩码来对一个或多个插槽进行掩码。转换器被训练以确定针对一个被掩码的插槽或多个被掩码的插槽的尽可能一致的预测。

[0064] 在该示例中，带有TIMEX 3表达的文本片段被输入到转换器中，并计算出关于针对相应掩码的可能用途的概率分布作为输出。在示例中，转换器包括权重，例如具有权重的人工神经网络。在示例中，反向传播被用于适配转换器的权重。

[0065] 在示例中，为了规范化时间表达，其与上下文无关的表示利用转换器而被预测。可以规定：使用转换器来确定TIMEX3表达中的具体值，即与上下文无关的表示kuRepr。在示例中，对于时间表达“Mai”(五月)，例如预测针对月份说明的插槽的内容“05”，其他插槽的内容为占位符。在这种情况下，内容“05”基于规则地被映射到与上下文无关的表示kuRepr＝″UNDEF-year-05″。

[0066] 针对所述时间表达所预测的与上下文无关的表示在示例中是规范化的表达。

[0067] 在示例中，知识库的填充包括两个子方面：

[0068] 1.如果时间表达是文本中的相对表达，则它被锚定到预给定的参考日期。例如，时间表达“昨天”是利用预给定的参考日期而有意义地得以确定的。

[0069] 2.如果时间表达是来自文本的隐式表达，比如平安夜，则通过语义知识将其解析为12月24日晚上。

[0070] 在该示例中，知识数据库包括实体和用于将实体相互链接、优选成对地链接的关系。知识数据库的一个示例是知识图，它将实体表示为节点，而将关系表示为节点之间的边。知识数据库在一个示例中包括多于100个实体。知识数据库在一个示例中包括多于1000个实体。知识数据库在一个示例中包括多于10000个实体。知识数据库在一个示例中包括多于100个关系。知识数据库在一个示例中包括多于1000个关系。知识数据库在一个示例中包括多于10000个关系。

[0071] 在示例中，文本包括至少一个时间表达。在一个示例中，文本包括多于100个时间表达。在一个示例中，文本包括多于1000个时间表达。在一个示例中，文本包括多于10000个时间表达。在示例中，文本包括参考日期。在一个示例中，文本包括多于100个参考数据。在一个示例中，文本包括多于1000个参考数据。在一个示例中，文本包括多于10000个参考数据。

[0072] 用于处理来自特别是非结构化文本的时间表达以填充知识数据库102的方法包括训练步骤和推理步骤。训练步骤的目标是：训练转换器以针对文本中的一个时间表达或针对文本中的多个时间表达而预测其与上下文无关的表示。推理步骤的目标是：使用该转换器用于针对文本中的一个时间表达或针对文本中的多个时间表达而预测其与上下文无关的表示。

[0073] 填充知识库的目标是：使用该与上下文无关的表示或这些与上下文无关的表示，用于将一个或多个实体或一个或多个关系或针对它们的属性添加到知识数据库中，其中所述与上下文无关的表示是通过文本中的时间表达的推理步骤所确定的。

[0074] 这些训练步骤可以独立于推理步骤来执行。特别是可以使用已经训练过的转换器独立于训练步骤执行推理步骤。

[0075] 训练步骤在图2中被示出。

[0076] 在步骤202中，提供时间表达。

[0077] 在示例中，从包括时间表达的文本中确定文本片段，例如句子。

[0078] 所述时间表达例如在文本中基于规则来确定。

[0079] 在步骤204中，提供与上下文无关的表示kuRepr。

[0080] 在示例中，所述与上下文无关的表示kuRepr包括多个插槽。时间表达通过插槽的内容与上下文无关地被表示。

[0081] 在示例中，提供了文本片段，其包含针对时间表达的TIMEX 3表达及其与上下文无关的表示。

[0082] 在步骤206中，与上下文无关的表示kuRepr的插槽中的至少一个被掩码。

[0083] 在示例中，转换器在多次迭代中得以训练。与在第二迭代中相比，在第一迭代中掩码更少插槽。在示例中，第二次迭代在第一次迭代之后执行。

[0084] 在步骤208中，与上下文无关的表示kuRepr和时间表达利用转换器被映射到针对与上下文无关的表示kuRepr的预测。在示例中，包含时间表达的文本片段利用转换器被映射到预测。在示例中，包含针对时间表达的TIMEX 3表达的文本片段利用转换器被映射到预测。

[0085] 可以规定：在第一训练阶段至少有一个插槽未被掩码并且至少有一个插槽被掩码。在第一训练阶段中，时间表达和至少一个未被掩码的插槽的内容利用转换器被映射到对至少一个被掩码的插槽的内容的预测。

[0086] 可以规定：在第二训练阶段中，尤其是所有插槽都被掩码。在第二训练阶段中，使用在第一训练阶段中所训练的转换器将时间表达映射到对插槽的内容的预测。

[0087] 在示例中，利用转换器而确定数字或数值或运算符，特别是算术运算符，或字母，或单词成分，或来自预给定的词汇表的单词或占位符作为预测。

[0088] 文本片段“Der Eta Aquariiden Meteorschauer wird im<TIMEX3type＝″DATE″value＝″UNDEF-year-05″>Mai</TIMEX3>erscheinen.”(水瓶座艾塔流星雨将在<TIMEX3type＝″DATE″value＝″UNDEF-year-05″>五月</TIMEX3>出现)在示例中包括与上下文无关的表示kuRepr＝″UNDEF-year-05″。在该示例中，转换器被训练用于，针对为此所定义的插槽通过转换器的针对该插槽所设置的输出端而预测数字“05”作为内容。

[0089] 以下说明针对其他预测的示例。

[0090] 例如，文本片段“…后天…”通过与上下文无关的表示“UNDEF-this-day-PLUS-2”来表示，其中针对所述表示，利用转换器而在所述转换器的针对第一插槽所设置的输出端预测单词“this”作为内容，在所述转换器的针对第二插槽所设置的输出端预测单词“day”作为内容，在所述转换器的针对第三插槽所设置的输出端预测运算符PLUS作为内容，在所述转换器的针对第四插槽所设置的输出端预测数字“2”作为内容。

[0091] 在示例中，利用转换器对于插槽的每个被掩码内容确定词汇表上的概率分布以及根据所述概率分布而确定该插槽的内容。

[0092] 如果与上下文无关的表示包括多个插槽，则使用转换器将与上下文无关的表示和

时间表达映射到对至少一个被掩码的插槽的内容的预测。

[0093] 步骤202至208在示例中的训练时期中针对来自文本的多个时间表达来执行。

[0094] 在步骤210中，确定转换器的至少一个权重。例如，根据针对与上下文无关的表示的预测与所述与上下文无关的表示的偏差来确定权重。如果与上下文无关的表示包括多个插槽，则根据对插槽中的至少一个插槽的被掩码的内容的预测与所述至少一个插槽的被掩码的内容的偏差来确定权重。在示例中，所述权重是通过反向传播根据大量偏差所确定的，其中这些偏差是针对来自文本的大量时间表达所确定的。

[0095] 在示例中，针对多个训练时期执行步骤210。

[0096] 推理步骤如图3所示。为了进行推理，转换器被设计为将时间表达和类型映射到所述时间表达的与上下文无关的表示的预测。

[0097] 在步骤300中，提供包括时间表达的文本。

[0098] 在步骤302中，为所述文本确定或预给定参考日期。例如，参考日期是从文本中确定的，特别是基于规则地确定。

[0099] 在步骤304中，特别是基于规则地在文本中确定时间表达。

[0100] 在示例中，特别是基于规则地根据文本而确定来自包括所述时间表达的文本的文本片段。

[0101] 在步骤306中，时间表达的类型特别是基于规则地根据文本而确定。

[0102] 在步骤308中，使用转换器将时间表达和类型映射到针对与上下文无关的表示的预测。

[0103] 在示例中，利用所述转换器而确定数字或数值或运算符，特别是算术运算符，或字母，或单词成分，或来自预给定的词汇表的单词或占位符作为预测。

[0104] 在示例中，对于每个插槽，利用所述转换器而确定词汇表上的概率分布以及根据所述概率分布而确定其内容。

[0105] 在示例中，利用转换器根据包含时间表达的文本片段来确定预测。

[0106] 在步骤310中，根据与上下文无关的表示以及根据参考数据而确定时间属性。

[0107] 在步骤312中，知识数据库被填充有时间属性。

[0108] 接着执行步骤304。

[0109] 在示例中，执行步骤304至312，直到已处理文本中存在的时间表达。

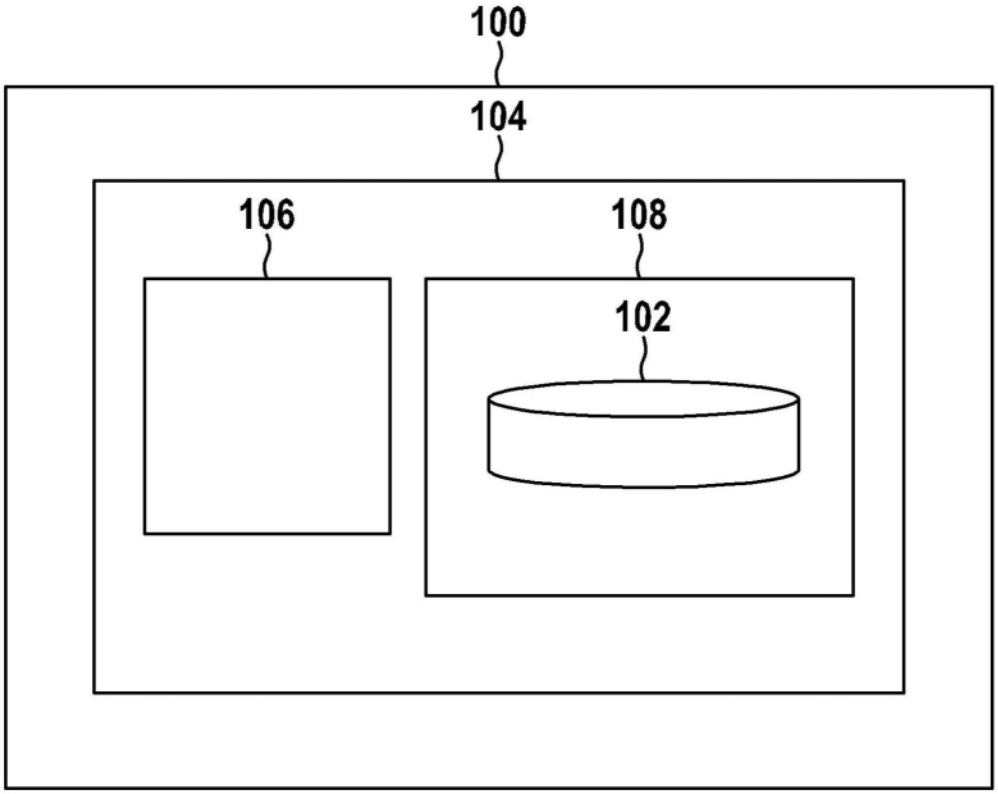


图1

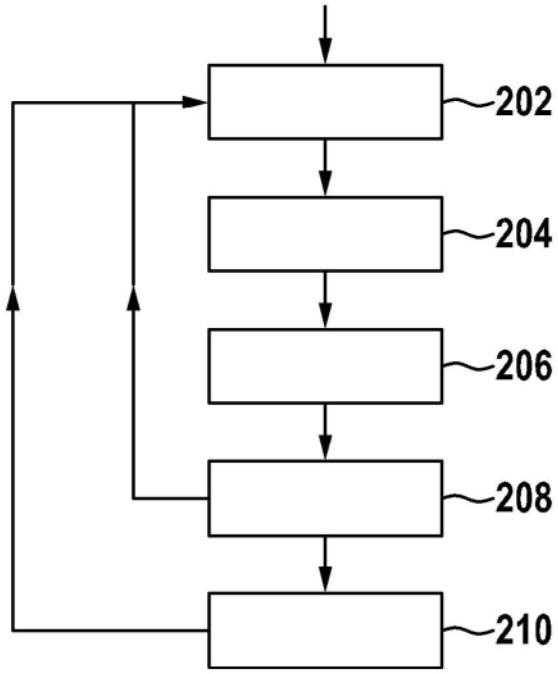


图2

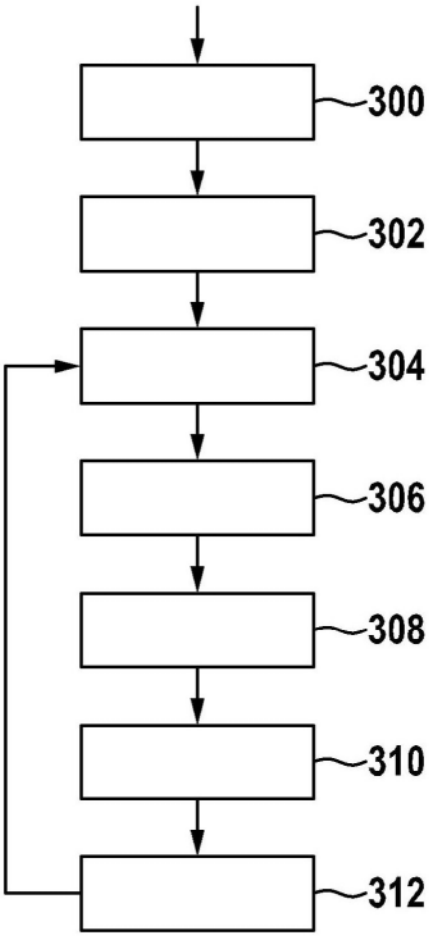


图3