

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 11 月 26 日 (26.11.2020)



(10) 国际公布号
WO 2020/233261 A1

- (51) 国际专利分类号:
G06F 16/33 (2019.01)
- (21) 国际申请号: PCT/CN2020/083591
- (22) 国际申请日: 2020 年 4 月 7 日 (07.04.2020)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201910629843.0 2019年7月12日 (12.07.2019) CN
- (71) 申请人: 之江实验室(ZHEJIANG LAB) [CN/CN];
中国浙江省杭州市余杭区文一西路 1818 号, Zhejiang 310023 (CN)。
- (72) 发明人: 李劲松(LI, Jingsong); 中国浙江省杭州市余杭区文一西路 1818 号, Zhejiang 310023 (CN)。

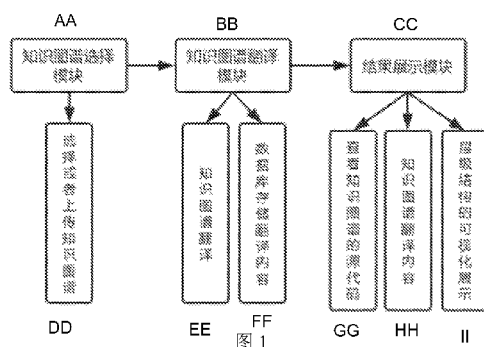
尚勇(SHANG, Yong); 中国浙江省杭州市余杭区文一西路 1818 号, Zhejiang 310023 (CN)。 吕可伟(LYU, Kewei); 中国浙江省杭州市余杭区文一西路 1818 号, Zhejiang 310023 (CN)。 周天舒(ZHOU, Tianshu); 中国浙江省杭州市余杭区文一西路 1818 号, Zhejiang 310023 (CN)。

(74) 代理人: 杭州求是专利事务所有限公司(HANGZHOU QIUSHI PATENT OFFICE CO., LTD.); 中国浙江省杭州市西湖区玉古路 147 号 黄鸿年科技综合大楼 215 室 / 刘静, Zhejiang 310013 (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,

(54) **Title:** NATURAL LANGUAGE GENERATION-BASED KNOWLEDGE GRAPH UNDERSTANDING ASSISTANCE SYSTEM

(54) 发明名称: 一种基于自然语言生成技术的知识图谱辅助理解系统



- AA Knowledge graph selection module
BB Knowledge graph translation module
CC Results display module
DD Select or upload knowledge graph
EE Knowledge graph translation
FF Database storage of translation content
GG View source code of knowledge graph
HH Knowledge graph translation content
II Layer-structured visual display

(57) **Abstract:** A natural language generation-based knowledge graph understanding assistance system. The system comprises a knowledge graph selection module, a knowledge graph translation module, and a results display module. The system uses natural language generation technology to convert a knowledge graph into a natural language text, so that prior to using a knowledge graph pertaining to a relevant domain, a domain expert can gain an accurate, in-depth, and full understanding of the knowledge graph even without having an understanding of the knowledge graph source code and software. In addition, each short sentence is associated with source code

GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, 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, 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))。
- 在修改权利要求的期限届满之前进行, 在收到该修改后将重新公布(细则48.2(h))。
- 根据申请人的请求, 在条约第21条(2)(a)所规定的期限届满之前进行。

corresponding to a knowledge graph; hence, if redundancies or erroneous information are found to be present in a knowledge graph, a correction thereto may be made promptly. Further, the invention may be applied broadly. A visualization method is utilized in the invention to further accelerate a domain expert's understanding of a knowledge graph.

(57) 摘要: 一种基于自然语言生成技术的知识图谱辅助理解系统, 该系统包括知识图谱选择模块、知识图谱翻译模块和结果展示模块; 利用自然语言生成技术将知识图谱转换为自然语言文本, 方便领域专家在不了解知识图谱的源代码和软件的基础上, 在使用知识图谱前, 对于其领域的知识图谱有准确、深入且全面的了解。同时每一个短句与知识图谱本身对应的源代码相关联, 如果发现知识图谱存在的冗余和错误信息, 可以及时对其进行纠正, 且该方法通用性较强。利用可视化方法进一步加快领域专家对于知识图谱的理解。

一种基于自然语言生成技术的知识图谱辅助理解系统

技术领域

本发明涉及知识图谱技术领域，尤其涉及一种基于自然语言生成技术的知识图谱辅助理解系统。

背景技术

知识图谱是一个语义知识库，它通常采用主语-谓语-宾语的三元组形式表示一个知识点，相比较于本体对于逻辑和语义的严苛要求，知识图谱强调了弱语义和弱逻辑，因此在学术界和工业界，知识图谱得到了较好地推广，以谷歌为首的大型互联网公司纷纷开始研究知识图谱以提高搜索的质量。根据 2014 年的报告显示，目前谷歌的知识图谱已经搜集到了超过 16 亿的事实，其中 2.71 亿个事实的真实性被认为超过 90%。在 2016 年 5 月的 Google 搜索中，知识图谱大约回答了该月 1000 亿次搜索中的三分之一的问题。

自然语言生成技术是自然语言处理技术的其中一大技术。不同于自然语言理解，自然语言生成技术关注的是计算机如何以自然语言文本来表达给定的含义、思想等。对于知识图谱，尤其是特定领域的知识图谱，其对于实际应用的知识图谱的准确性要求非常高，例如医学相关的知识图谱，其知识图谱的质量严重关系到整个系统的准确性。然而构建知识图谱的编程语言和本体一样，主要是 RDF (Resource Description Framework, 资源描述框架)、OWL (Web Ontology Language, 网络本体语言)，采用的软件主要是斯坦福大学开发的 Protégé 等。这些语言和软件专业性强，如果没有经过长期学习和培训，非相关人员很难理解其具体含义。同时以 OWL 和 RDF 存储的知识点是无序的，同一内容相关的知识点存储在程序的不同部位，这进一步加大了领域专家直接理解知识图谱的源代码的难度。知识图谱大多由计算机行业从业者建立，但使用者则是该知识图谱内容相关的领域的学者和专家，两者的不匹配性导致了领域专家无法理解知识图谱的内容，只能通过使用来进一步完善知识图谱，而不能提前对知识图谱的内容有直观的理解并进行改善。这间接导致了知识图谱质量的不稳定性以及相同内容的知识图谱二次开发现象的严重性。有学者在 2017 年随机抽取了美国国家生物医学本体中心中的 200 个生物医学相关的本体，发现在其相应的设计文档中，只有 17 个得到了专家们的正式评估。

许多领域的知识图谱在使用前需要领域专家对其表示内容进行深入而全面的了解，以便保障其在实际使用过程中的准确性。但知识图谱的相关语言和软件专业性强，相同主题的知识分布零散，领域专家很难在短期内掌握并了解它们。目前辅助理解知识图谱的软件大多

是通过搜索，以可视化的手段呈现不同知识节点的关联，这样呈现的知识是局部知识，且没有涉及到知识图谱本身。同时这些方法都是在知识图谱使用过程中来发现其存在的问题，而没有在其使用之前对其进行全面的了解和评估。

发明内容

本发明的目的是针对于目前知识图谱质量把控不足、领域专家难以理解领域相关的知识图谱的基础上，提出一种基于自然语言生成技术的知识图谱辅助理解系统，本发明根据现实需要，利用自然语言生成技术根据知识图谱的内容生成自然语言文本，并对同一主题的短句进行句子的适当聚合，并且每一个句子都与知识图谱中程序语言一一对应，以方便领域专家在较短时间内对知识图谱进行快速、全面和深入的了解，进行知识图谱质量的把控。

本发明是通过以下技术方案来实现：一种基于自然语言生成技术的知识图谱辅助理解系统，该系统包括知识图谱选择模块、知识图谱翻译模块和结果展示模块；

所述知识图谱选择模块用于获取符合 RDF 或 OWL 语法规则的目标知识图谱；

所述知识图谱翻译模块：首先提取目标知识图谱的三元组，对提取的三元组进行字符串分割得到三个动态数组：主语数组、谓语数组、宾语数组，三者之间具有一一对应关系，然后通过嵌套循环，利用 `simplenlg` 工具对主语、谓语和宾语进行组装，形成一句完整的短句；同时对于主语-谓语-宾语的关系存在一对一对多和一对多对多的情况，在谓语数组和宾语数组中加入特殊字符进行标识，以确定该谓语对应某一主语以及该宾语对应某一主语和某一谓语，然后在嵌套循环中判断这些特殊字符以确定主语、谓语、宾语的对应关系，利用 `simplenlg` 工具对对应的主语、谓语和宾语进行组装，形成一个完整的长句；其中注解部分对应的三元组不单独成句，而是作为补充其他句子的注解信息；然后将目标知识图谱翻译成短句和长句，句子得到进一步规范后被存入本地数据库（可以采用 MySQL 数据库）中，并从主语、谓语、宾语三个动态数组中选出类与子类、类与实例关系的内容组装成 JSON 格式的文件。

所述结果展示模块从本地数据库调取目标知识图谱的翻译内容（即短句和长句），将翻译内容以及目标知识图谱的源文件（RDF（资源描述框架），OWL（网络本体语言））进行共同展示，同时获取 JSON 格式的文件，通过可视化工具（可以采用 D3 工具）绘制树状图，对知识图谱中的类与子类以及类与实例的层级结构进行可视化展示。

进一步地，所述知识图谱选择模块获取目标知识图谱的途径包括两种：

途径一：从开源知识图谱数据库（该系统应用生物医学领域的知识图谱辅助理解时，开源知识图谱数据库可以选择美国国家生物医学本体中心（National Center for Biomedical Ontology, NCBO））中爬取符合 RDF 或 OWL 语法规则的知识图谱，对爬取的知识图谱通过知识图谱翻译模块进行翻译，将翻译结果存入本地数据库中；当用于在系统中搜索某一主题

的知识图谱时，输入名称与知识图谱的英文名称进行相似度计算，按相似度从大到小进行排序，得到待选目标知识图谱；

途径二：用户上传符合 RDF 或 OWL 语法规则的知识图谱作为目标知识图谱。

进一步地，获取目标知识图谱的途径一中，相似度判断系数采用 Jaccard 相似系数（杰卡德系数），它常用于比较有限样本集之间的相似性与差异性，Jaccard 系数值越大，样本相似度越高。

将用户输入名称的概念集合记为 C_1 ，知识图谱英文名称的概念集合记为 C_2 ，则两者之间的 Jaccard 相似系数 $J(C_1, C_2)$ 为：

$$J(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} = \frac{|C_1 \cap C_2|}{|C_1| + |C_2| - |C_1 \cap C_2|}, \quad J(C_1, C_2) \in [0, 1]$$

如果 C_1 和 C_2 完全相同，则 $J(C_1, C_2)$ 值为 1；每次搜索结果按照相似度大小排序，呈现相似度较高的 N 个结果， N 为用户自定义。

进一步地，所述知识图谱翻译模块中提取目标知识图谱的三元组的步骤具体如下：利用 SPARQL（SPARQL Protocol and RDF Query Language，SPARQL 协议和 RDF 查询语言）提取目标知识图谱中全部知识点（类、实例、对象属性、数据属性、注解等）对应的主语、谓语和宾语，并将其编码成资源描述框架的三元组（RDF Triple）。

进一步地，所述知识图谱翻译模块中目标知识图谱的短句生成步骤具体如下：首先对获取的三元组进行字符串分割，获得主语、谓语和宾语的名称，构建三个动态数组。在短句生成中，由于主语、谓语和宾语关系是一一对一的关系，所以通过嵌套循环，将对应的主语、谓语、宾语利用 Simplenlg 直接组装成短句即可。

进一步地，所述知识图谱翻译模块中目标知识图谱的长句生成的步骤具体如下：首先对获取的三元组进行字符串分割，获得主语、谓语和宾语的名称，构建三个动态数组。在长句生成中，考虑到一个主语可以对应多个谓语，而每个谓语又可以对应多个宾语，因此在谓语数组中，不同主语对应的谓语之间用特殊标识符进行标记；在宾语数组中，不同主语对应的不同谓语的宾语之间采用另一特殊标识符进行标记，实现主语、谓语、宾语的一一对应关系，然后采用嵌套循环，对这些特殊标识符进行判断，将对应的主语、谓语和宾语利用 Simplenlg 进行组装。其中同一主语不同谓语各自构成一个句子，同一主语的所有句子构成一个段落，不同宾语之间用连接词（和、或）进行连接。

进一步地，所述知识图谱翻译模块中目标知识图谱的补充句子的注解信息步骤具体如下：首先对谓语数组进行循环，如果谓语为“comment”（表示宾语为主语的注释），则将对应的主语和宾语提取，形成一个新的动态数组-注解数组，其中奇数下标的数组元素存储主语，偶数下

标的数组元素存储宾语。然后进行主语数组、谓语数组和宾语数组的嵌套循环，判断主语、宾语是否在注解数组中，如果存在，则在该主语或者宾语后面添加括号，括号内为其注解，然后判断谓语，如果谓语不是“comment”，则进行组装，否则不组装。

进一步地，所述知识图谱翻译模块中目标知识图谱的短句和长句插入数据库的步骤具体如下：利用 JDBC（Java DataBase Connectivity）API 进行数据库的连接，首先创建存放翻译结果的数据库以及数据表，定义表名、表字段、确认主键等，然后将知识图谱的英文名称与数据库中存储的名称进行匹配，如果该知识图谱的翻译结果已经存在于知识库中，则不进行插入操作，如果没有，则将生成的短句数组和长句数组添加到数据表中。

进一步地，所述结果展示模块中翻译内容和源文件展示的具体步骤如下：在网页界面中选定目标知识图谱后，利用 ajax 从数据库中调取该知识图谱对应的所有翻译内容显示在界面上，并且从本地服务器读取目标知识图谱的源文件，共同显示在界面中。

进一步地，所述结果展示模块中可视化展示的具体步骤如下：在网页界面中选定目标知识图谱后，利用 ajax 获取后端相应的 JSON 格式的文件，绘制树状图；在树状图中，每个节点表示主语或者宾语，每个节点通过连线与其他关联的节点进行连接。

本发明的有益效果是：本发明利用自然语言生成技术将知识图谱转换为自然语言文本，方便领域专家在不了解知识图谱的源代码和软件的基础上，在使用知识图谱前，对于其领域的知识图谱有准确、深入且全面的了解。同时每一个短句与知识图谱本身对应的源代码相关联，如果发现知识图谱存在的冗余和错误信息，可以及时对其进行纠正，且该方法通用性较强。本发明利用可视化方法进一步加快领域专家对于知识图谱的理解。

附图说明

- 图 1 为本发明基于自然语言生成技术的知识图谱辅助理解系统的结构框图；
- 图 2 为本发明基于自然语言生成技术的知识图谱辅助理解系统的实现流程图；
- 图 3 为本发明知识图谱翻译模块的自然语言生成流程图；
- 图 4 为某一知识图谱的部分源代码示意图；
- 图 5 为利用自然语言技术生成的短句示意图；
- 图 6 为利用自然语言技术生成的长句示意图；
- 图 7 为类与子类的树状图。

具体实施方式

下面结合附图和具体实施例对本发明作进一步详细说明。

如图 1、2 所示，本发明提供一种基于自然语言生成技术的知识图谱辅助理解系统，包括知识图谱选择模块、知识图谱翻译模块和结果展示模块；

一、知识图谱选择模块

知识图谱选择模块用于获取符合 RDF 或 OWL 语法规则的目标知识图谱；获取目标知识图谱的途径包括两种：

途径一：从开源知识图谱数据库（该系统应用生物医学领域的知识图谱辅助理解时，开源知识图谱数据库可以选择美国国家生物医学本体中心（National Center for Biomedical Ontology, NCBO））中爬取符合 RDF 或 OWL 语法规则的知识图谱，对爬取的知识图谱通过知识图谱翻译模块进行翻译，将翻译结果存入本地数据库中；当用于在系统中搜索某一主题的知识图谱时，输入名称与知识图谱的英文名称进行相似度计算，按相似度从大到小进行排序，得到待选目标知识图谱；

相似度判断系数可以采用 Jaccard 相似系数（杰卡德系数），它常用于比较有限样本集之间的相似性与差异性，Jaccard 系数值越大，样本相似度越高。

将用户输入名称的概念集合记为 C_1 ，知识图谱英文名称的概念集合记为 C_2 ，则两者之间的 Jaccard 相似系数 $J(C_1, C_2)$ 为：

$$J(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} = \frac{|C_1 \cap C_2|}{|C_1| + |C_2| - |C_1 \cap C_2|}, \quad J(C_1, C_2) \in [0, 1]$$

如果 C_1 和 C_2 完全相同，则 $J(C_1, C_2)$ 值为 1；每次搜索结果按照相似度大小排序，呈现相似度较高的 N 个结果， N 为用户自定义， N 可以设置为 15。

途径二：用户上传符合 RDF 或 OWL 语法规则的知识图谱作为目标知识图谱。

二、知识图谱翻译模块

具体流程如图 3 所示，首先提取目标知识图谱的三元组，对提取的三元组进行字符串分割得到三个动态数组：主语数组、谓语数组、宾语数组，三者之间具有一一对应关系，然后通过嵌套循环，利用 `simplenlg` 工具对主语、谓语和宾语进行组装，形成一句完整的短句；同时对于主语-谓语-宾语的关系存在一对一对多和一对多对多的情况，在谓语数组和宾语数组中加入特殊字符进行标识，以确定该谓语对应某一主语以及该宾语对应某一主语和某一谓语，然后在嵌套循环中判断这些特殊字符以确定主语、谓语、宾语的对应关系，利用 `simplenlg` 工具对对应的主语、谓语和宾语进行组装，形成一个完整的长句；其中注解部分对应的三元组不单独成句，而是作为补充其他句子的注解信息；然后将目标知识图谱翻译成短句和长句，同时生成的句子需要进一步规范，例如句子开头的英文字母大写，部分名称添加超链接等。规范后的句子被插入本地数据库中，并从主语、谓语、宾语三个动态数组中选出类与子类、类与实例关系的内容组装成 JSON 格式的文件。本地数据库可采用 MySQL 数据库，MySQL 是目前比较流行的一个开源的关系型数据库管理系统，它可以将数据保存在不同的表中，而

不是把数据全部放在一个的仓库里，这样就增加了速度。

提取目标知识图谱的三元组的步骤具体如下：利用 SPARQL（SPARQL Protocol and RDF Query Language，SPARQL 协议和 RDF 查询语言）提取目标知识图谱中全部知识点（类、实例、对象属性、数据属性、注解等）对应的主语、谓语和宾语，并将其编码成资源描述框架的三元组（RDF Triple）。

目标知识图谱的短句生成步骤具体如下：首先对获取的三元组进行字符串分割，获得主语、谓语和宾语的名称，构建三个动态数组。在短句生成中，由于主语、谓语和宾语关系是一一对一的关系，所以通过嵌套循环，将对应的主语、谓语、宾语利用 `Simplenlg` 直接组装成短句即可。

目标知识图谱的长句生成的步骤具体如下：首先对获取的三元组进行字符串分割，获得主语、谓语和宾语的名称，构建三个动态数组。在长句生成中，考虑到一个主语可以对应多个谓语，而每个谓语又可以对应多个宾语，因此在谓语数组中，不同主语对应的谓语之间用特殊标识符进行标记；在宾语数组中，不同主语对应的不同谓语的宾语之间采用另一特殊标识符进行标记，实现主语、谓语、宾语的一一对应关系，然后采用嵌套循环，对这些特殊标识符进行判断，将对应的主语、谓语和宾语利用 `Simplenlg` 进行组装。其中同一主语不同谓语各自构成一个句子，同一主语的所有句子构成一个段落，不同宾语之间用连接词（和、或）进行连接。

目标知识图谱的补充句子的注解信息步骤具体如下：首先对谓语数组进行循环，如果谓语为“comment”（表示宾语为主语的注释），则将对应的主语和宾语提取，形成一个新的动态数组-注解数组，其中奇数下标的数组元素存储主语，偶数下标的数组元素存储宾语。然后进行主语数组、谓语数组和宾语数组的嵌套循环，判断主语、宾语是否在注解数组中，如果存在，则在该主语或者宾语后面添加括号，括号内为其注解，然后判断谓语，如果谓语不是“comment”，则进行组装，否则不组装。

目标知识图谱的短句和长句插入数据库的步骤具体如下：利用 JDBC（Java DataBase Connectivity）API，实现 Java 与数据库的连接，首先创建存放翻译结果的数据库以及数据表，定义表名、表字段、确认主键等，然后将知识图谱的英文名称与数据库中存储的名称进行匹配，如果该知识图谱的翻译结果已经存在于知识库中，则不进行插入操作，如果没有，则将生成的短句数组和长句数组添加到数据表中。

三、结果展示模块

结果展示分为三部分。当网页端选中目标知识图谱或者在网站上传目标知识图谱时，通过 `ajax` 会将该文件或者参数递交给后端，文件被传入后端后，其源代码会显示在网页上，并

自动进行自然语言生成，生成的结果会被插入数据库中，然后从数据库中读取相关内容显示在网页端。同时系统从主语、谓语、宾语三个动态数组中选出类与子类、类与实例关系的内容组装成 JSON 格式的文件，传送到前端，并利用可视化工具 D3 绘制树状图，显示其主要的层级结构。以美国生物医学本体中心公开的一个有关慢性肾病的知识图谱为例，其运行结果如图 4-7 所示，图 7 展示了树状图的部分内容。

利用本发明系统，把目标知识图谱上传到网站上或者在网站上选择库中的知识图谱后，系统会自动查询知识图谱里的相关内容，分割字符串，将 RDF Triple 翻译成短句和长句，并进一步规范句式，最后将生成的文本展示给领域专家，其中每一条句子都与其知识图谱的源代码相对应。同时系统对知识图谱中重要的类与子类、类与实例关系以树状图的形式呈现，帮助专家快速理解和掌握该知识图谱的内容和信息，以便在短时间内进行质量的把控。

以上仅为本发明的实施实例，并非用于限定本发明的保护范围。凡在本发明的精神和原则之内，不经过创造性劳动所作的任何修改、等同替换、改进等，均包含在本发明的保护范围内。

权 利 要 求 书

1、一种基于自然语言生成技术的知识图谱辅助理解系统，其特征在于，该系统包括知识图谱选择模块、知识图谱翻译模块和结果展示模块；

所述知识图谱选择模块用于获取符合 RDF 或 OWL 语法规则的目标知识图谱；

所述知识图谱翻译模块：首先提取目标知识图谱的三元组，对提取的三元组进行字符串分割得到三个动态数组：主语数组、谓语数组、宾语数组，三者之间具有一一对应关系，然后通过嵌套循环，利用 simplenlg 工具对主语、谓语和宾语进行组装，形成一句完整的短句；同时对于主语-谓语-宾语的关系存在一对一对多和一对多对多的情况，在谓语数组和宾语数组中加入特殊字符进行标识，以确定该谓语对应某一主语以及该宾语对应某一主语和某一谓语，然后在嵌套循环中判断这些特殊字符以确定主语、谓语、宾语的对应关系，利用 simplenlg 工具对对应的主语、谓语和宾语进行组装，形成一个完整的长句；其中注解部分对应的三元组不单独成句，而是作为补充其他句子的注解信息；然后将目标知识图谱翻译成短句和长句，句子得到进一步规范后被存入本地数据库中，并从主语、谓语、宾语三个动态数组中选出类与子类、类与实例关系的内容组装成 JSON 格式的文件；

所述结果展示模块从本地数据库调取目标知识图谱的翻译内容，将翻译内容以及目标知识图谱的源文件进行共同展示，同时获取 JSON 格式的文件，通过可视化工具绘制树状图，对知识图谱中的类与子类以及类与实例的层级结构进行可视化展示。

2、根据权利要求 1 所述的一种基于自然语言生成技术的知识图谱辅助理解系统，其特征在于，所述知识图谱选择模块获取目标知识图谱的途径包括两种：

途径一：从开源知识图谱数据库中爬取符合 RDF 或 OWL 语法规则的知识图谱，对爬取的知识图谱通过知识图谱翻译模块进行翻译，将翻译结果存入本地数据库中；当用于在系统中搜索某一主题的知识图谱时，输入名称与知识图谱的英文名称进行相似度计算，按相似度从大到小进行排序，得到待选目标知识图谱；

途径二：用户上传符合 RDF 或 OWL 语法规则的知识图谱作为目标知识图谱。

3、根据权利要求 2 所述的一种基于自然语言生成技术的知识图谱辅助理解系统，其特征在于，获取目标知识图谱的途径一中，相似度判断系数采用 Jaccard 相似系数。

将用户输入名称的概念集合记为 C_1 ，知识图谱英文名称的概念集合记为 C_2 ，则两者之间的 Jaccard 相似系数 $J(C_1, C_2)$ 为：

$$J(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} = \frac{|C_1 \cap C_2|}{|C_1| + |C_2| - |C_1 \cap C_2|}, \quad J(C_1, C_2) \in [0, 1]$$

如果 C_1 和 C_2 完全相同, 则 $J(C_1, C_2)$ 值为 1; 每次搜索结果按照相似度大小排序。

4、根据权利要求 1 所述的一种基于自然语言生成技术的知识图谱辅助理解系统, 其特征在于, 所述知识图谱翻译模块中提取目标知识图谱的三元组的步骤具体如下: 利用 SPARQL 提取目标知识图谱中全部知识点(类、实例、对象属性、数据属性、注解等)对应的主语、谓语和宾语, 并将其编码成资源描述框架的三元组。

5、根据权利要求 1 所述的一种基于自然语言生成技术的知识图谱辅助理解系统, 其特征在于, 所述知识图谱翻译模块中目标知识图谱的短句生成步骤具体如下: 首先对获取的三元组进行字符串分割, 获得主语、谓语和宾语的名称, 构建三个动态数组。在短句生成中, 由于主语、谓语和宾语关系是一一对一的关系, 所以通过嵌套循环, 将对应的主语、谓语、宾语利用 Simplenlg 直接组装成短句即可。

6、根据权利要求 1 所述的一种基于自然语言生成技术的知识图谱辅助理解系统, 其特征在于, 所述知识图谱翻译模块中目标知识图谱的长句生成的步骤具体如下: 首先对获取的三元组进行字符串分割, 获得主语、谓语和宾语的名称, 构建三个动态数组。在长句生成中, 考虑到一个主语可以对应多个谓语, 而每个谓语又可以对应多个宾语, 因此在谓语数组中, 不同主语对应的谓语之间用特殊标识符进行标记; 在宾语数组中, 不同主语对应的不同谓语的宾语之间采用另一特殊标识符进行标记, 实现主语、谓语、宾语的一一对应关系, 然后采用嵌套循环, 对这些特殊标识符进行判断, 将对应的主语、谓语和宾语利用 Simplenlg 进行组装。其中同一主语不同谓语各自构成一个句子, 同一主语的所有句子构成一个段落, 不同宾语之间用连接词进行连接。

7、根据权利要求 1 所述的一种基于自然语言生成技术的知识图谱辅助理解系统, 其特征在于, 所述知识图谱翻译模块中目标知识图谱的补充句子的注解信息步骤具体如下: 首先对谓语数组进行循环, 如果谓语为“comment”(表示宾语为主语的注释), 则将对应的主语和宾语提取, 形成一个新的动态数组-注解数组, 其中奇数下标的数组元素存储主语, 偶数下标的数组元素存储宾语。然后进行主语数组、谓语数组和宾语数组的嵌套循环, 判断主语、宾语是否在注解数组中, 如果存在, 则在该主语或者宾语后面添加括号, 括号内为其注解, 然后判断谓语, 如果谓语不是“comment”, 则进行组装, 否则不组装。

8、根据权利要求 1 所述的一种基于自然语言生成技术的知识图谱辅助理解系统, 其特征在于, 所述知识图谱翻译模块中目标知识图谱的短句和长句插入数据库的步骤具体如下: 利用 JDBC API 进行数据库的连接, 首先创建存放翻译结果的数据库以及数据表, 定义表名、表字段、确认主键等, 然后将知识图谱的英文名称与数据库中存储的名称进行匹配, 如果该知识图谱的翻译结果已经存在于本地数据库中, 则不进行插入操作, 如果没有, 则将生成的

短句数组和长句数组添加到数据表中。

9、根据权利要求 1 所述的一种基于自然语言生成技术的知识图谱辅助理解系统，其特征在于，所述结果展示模块中翻译内容和源文件展示的具体步骤如下：在网页界面中选定目标知识图谱后，利用 ajax 从数据库中调取该知识图谱对应的所有翻译内容显示在界面上，并且从本地服务器读取目标知识图谱的源文件，共同显示在界面中。

10、根据权利要求 1 所述的一种基于自然语言生成技术的知识图谱辅助理解系统，其特征在于，所述结果展示模块中可视化展示的具体步骤如下：在网页界面中选定目标知识图谱后，利用 ajax 获取后端相应的 JSON 格式的文件，绘制树状图；在树状图中，每个节点表示主语或者宾语，每个节点通过连线与其他关联的节点进行连接。

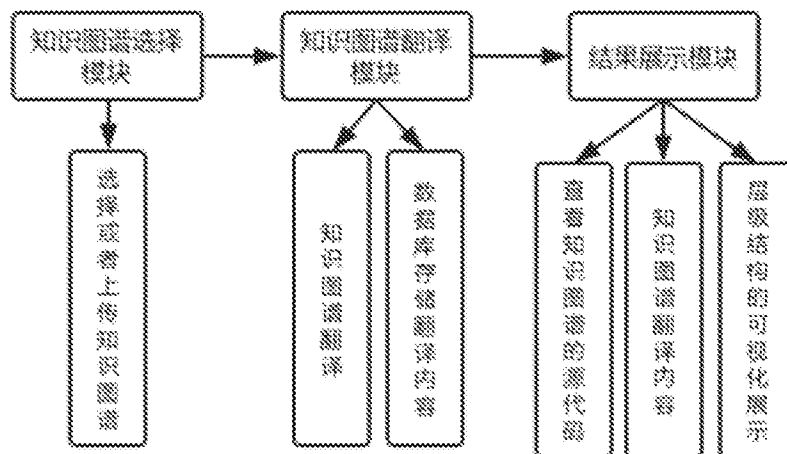


图 1

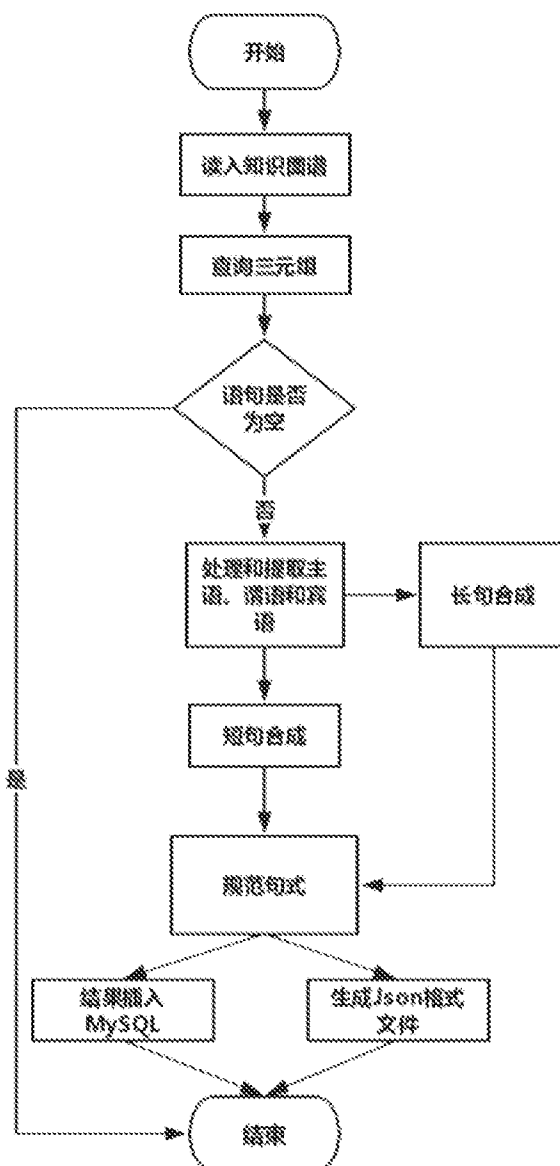


图 2

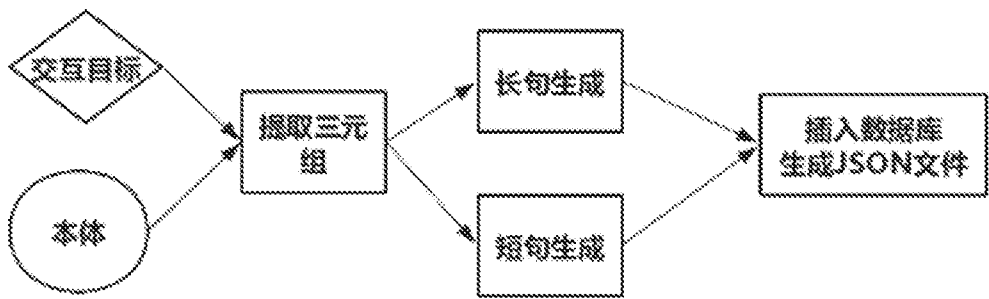


图 3

```
<owl:Class rdf:about="http://clininf.eu/ckdo#24h_urinary_protein_excretion">
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >CKD defined as &gt; 30mg in 24 hours, confirmed over at least 3 months
</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Class rdf:about="http://clininf.eu/ckdo#Proteinuria_estimation"/>
  </rdfs:subClassOf>
</owl:Class>
```

图 4

Proteinuria estimation includes 24h urinary protein excretion(CKD defined as > 30mg in 24 hours, confirmed over at least 3 months).

图 5

The location of Schloss Volrad is German region.Schloss Volrad produces Schloss Volrad Trockenbeerenauslese Riesling.

图 6

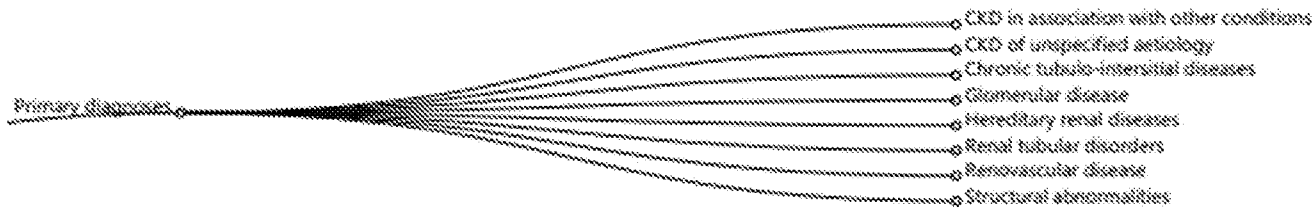


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/083591

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/33(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, IEEE: 图谱, 树状图, 主语, 谓语, 宾语, 短句, 长句, 注释, 注解, RDF, OWL, simplenlg, JSON, icon, atlas, dendrogram, subject, predicate, object, short, long, annotation, note

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110347798 A (ZHIJIANG LAB) 18 October 2019 (2019-10-18) claims 1-10	1-10
A	CN 109033260 A (TIANJIN UNIVERSITY) 18 December 2018 (2018-12-18) claim 1, description paragraphs [0002]-[0084]	1-10
A	CN 108959613 A (HANGZHOU DIANZI UNIVERSITY) 07 December 2018 (2018-12-07) entire document	1-10
A	CN 103020148 A (FUDAN UNIVERSITY) 03 April 2013 (2013-04-03) entire document	1-10
A	WO 2019118256 A1 (MICROSOFT TECHNOLOGY LICENSING, LLC) 20 June 2019 (2019-06-20) entire document	1-10
A	WO 2019103727 A1 (GOOGLE LLC) 31 May 2019 (2019-05-31) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

23 June 2020

Date of mailing of the international search report

09 July 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/083591

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110347798	A	18 October 2019	None			
CN	109033260	A	18 December 2018	None			
CN	108959613	A	07 December 2018	None			
CN	103020148	A	03 April 2013	None			
WO	2019118256	A1	20 June 2019	CN	110019471	A	16 July 2019
WO	2019103727	A1	31 May 2019	CN	110741389	A	31 January 2020
				US	2020118009	A1	16 April 2020
				EP	3622444	A1	18 March 2020

国际检索报告

国际申请号

PCT/CN2020/083591

A. 主题的分类

G06F 16/33 (2019.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNPAT, WPI, EPDOC, CNKI, IEEE: 图谱, 树状图, 主语, 谓语, 宾语, 短句, 长句, 注释, 注解, RDF, OWL, simplenlg, JSON, icon, atlas, dendrogram, subject, predicate, object, short, long, annotation, note

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
PX	CN 110347798 A (之江实验室) 2019年 10月 18日 (2019 - 10 - 18) 权利要求1-10	1-10
A	CN 109033260 A (天津大学) 2018年 12月 18日 (2018 - 12 - 18) 权利要求1, 说明书第[0002]-[0084]段	1-10
A	CN 108959613 A (杭州电子科技大学) 2018年 12月 7日 (2018 - 12 - 07) 全文	1-10
A	CN 103020148 A (复旦大学) 2013年 4月 3日 (2013 - 04 - 03) 全文	1-10
A	WO 2019118256 A1 (MICROSOFT TECHNOLOGY LICENSING, LLC) 2019年 6月 20日 (2019 - 06 - 20) 全文	1-10
A	WO 2019103727 A1 (GOOGLE LLC) 2019年 5月 31日 (2019 - 05 - 31) 全文	1-10

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2020年 6月 23日

国际检索报告邮寄日期

2020年 7月 9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

胡贝贝

电话号码 86-(10)-53961415

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/083591

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	110347798	A	2019年 10月 18日	无	
CN	109033260	A	2018年 12月 18日	无	
CN	108959613	A	2018年 12月 7日	无	
CN	103020148	A	2013年 4月 3日	无	
WO	2019118256	A1	2019年 6月 20日	CN 110019471 A	2019年 7月 16日
WO	2019103727	A1	2019年 5月 31日	CN 110741389 A	2020年 1月 31日
				US 2020118009 A1	2020年 4月 16日
				EP 3622444 A1	2020年 3月 18日