



(51) International Patent Classification:

G06F 16/9032 (2019.01) G06F 16/33 (2019.01)

(21) International Application Number:

PCT/US2022/031699

(22) International Filing Date:

01 June 2022 (01.06.2022)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

202110771263.2 08 July 2021 (08.07.2021) CN

(71) Applicant: MICROSOFT TECHNOLOGY LICENSING, LLC [US/US]; One Microsoft Way, Redmond, Washington 98052-6399 (US).

(72) Inventors: WANG, Xueyun; Microsoft Technology Licensing, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). SHOU, Linjun; Microsoft Technology Licensing, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). GONG, Ming; Microsoft Technology Licensing, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). JIANG, Daxin; Microsoft Technology Licensing, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(74) Agent: CHATTERJEE, Aaron C. et al.; Microsoft Technology Licensing, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): 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, IQ, IR, IS, IT, JM, 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) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (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).

(54) Title: QUESTION GENERATION FROM QUERIES

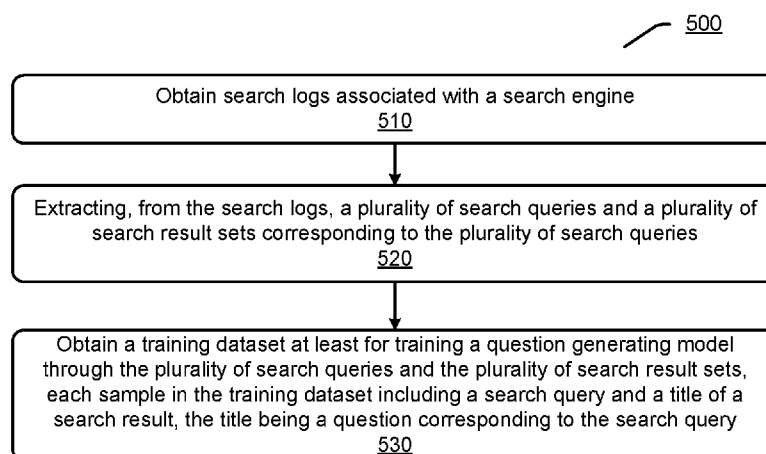


FIG 5

(57) Abstract: The present disclosure proposes a method, apparatus and computer program product for question generation from queries. Search logs associated with a search engine may be obtained. A plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries may be extracted from the search logs. A training dataset at least for training a question generating model may be obtained through the plurality of search queries and the plurality of search result sets, each sample in the training dataset including a search query and a title of a search result, the title being a question corresponding to the search query.

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*

QUESTION GENERATION FROM QUERIES

BACKGROUND

Natural Language Processing (NLP) is a technology that uses natural language to communicate with computers, which aims to enable computers to understand and employ natural language to achieve communications between humans and computers, thereby replacing humans to perform various tasks related to natural language. A task performed with NLP technology may be referred to as a NLP task. Examples of NLP tasks may include a Question Answering (QA) task, a Machine Reading Comprehension (MRC) task, a Query Passage (QP) relevance task, and a Query List (QList) relevance task, etc. A machine learning model may be trained with a training dataset for a specific NLP task. The trained machine learning model may be deployed to perform the NLP task.

SUMMARY

This Summary is provided to introduce a selection of concepts that are further described below in the Detailed Description. It is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter.

Embodiments of the present disclosure propose a method, apparatus and computer program product for question generation from queries. Search logs associated with a search engine may be obtained. A plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries may be extracted from the search logs. A training dataset at least for training a question generating model may be obtained through the plurality of search queries and the plurality of search result sets, each sample in the training dataset including a search query and a title of a search result, the title being a question corresponding to the search query.

It should be noted that the above one or more aspects comprise the features hereinafter fully described and particularly pointed out in the claims. The following description and the drawings set forth in detail certain illustrative features of the one or more aspects. These features are only indicative of the various ways in which the principles of various aspects may be employed, and this disclosure is intended to include all such aspects and their equivalents.

BRIEF DESCRIPTION OF THE DRAWINGS

The disclosed aspects will hereinafter be described in connection with the appended drawings that are provided to illustrate and not to limit the disclosed aspects.

FIG. 1 illustrates a schematic diagram including a set of exemplary queries 100a and a set of exemplary questions 100b corresponding to the set of queries.

FIG. 2 illustrates an exemplary process for question generation from queries according to an

embodiment of the present disclosure.

FIG. 3 illustrates an exemplary process for generating an initial training dataset according to an embodiment of the present disclosure.

FIG. 4 illustrates an exemplary process for removing a sample according to an embodiment of the present disclosure.

FIG. 5 is a flowchart of an exemplary method for question generation from queries according to an embodiment of the present disclosure.

FIG. 6 illustrates an exemplary apparatus for question generation from queries according to an embodiment of the present disclosure.

FIG. 7 illustrates an exemplary apparatus for question generation from queries according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

The present disclosure will now be discussed with reference to several example implementations. It is to be understood that these implementations are discussed only for enabling those skilled in the art to better understand and thus implement the embodiments of the present disclosure, rather than suggesting any limitations on the scope of the present disclosure.

Some NLP models that perform NLP tasks may perform corresponding processing based on an input query, to output a result for the query. As an example, a QA model used to perform a QA task may output an answer that can answer a query upon receiving the query from a user. As another example, a MRC model used to perform a MRC task may output a text segment in a text passage that can answer a query upon receiving the query and the text passage. A query provided to a NLP model may be informal, e.g., with grammar or spelling errors, with inexplicit intent, etc. For example, a query provided by a user to a NLP model may be "calculate density".

The query only includes two keywords instead of a complete sentence, so the intent is inexplicit.

Such a query may not be easily understood and processed by the NLP model, thus it is difficult to output an accurate result for the query.

Embodiments of the present disclosure propose question generation from queries. Herein, a query may refer to an original text segment provided by a user or other entity for obtaining specific information. The query may be informal, e.g., not following a correct grammar structure, with typing or spelling errors, with inexplicit intent, etc. For example, when the query includes only several keywords, its indicated doubt is ambiguous, and therefore the intent is inexplicit. Herein, a question may refer to a text segment that indicates a topic that requires an answer or explanation. The question is usually a complete sentence, e.g., an interrogative sentence. The question may be formal, e.g., following a correct grammar structure, with correct spelling, with an explicit intent, etc. Since a question has a form that is easily understood by

humans and machines, it may also be referred to as a natural language question. A question may be generated based on a query through a question generating model. The question generating model may be a machine learning model specially trained to generate a question from a query. The question generating model may generate a question suitable for being processed by a NLP model from an input query. For example, the question generating model may obtain a query originally input to the NLP model, and generate a question from the obtained query. The generated question may be provided to the NLP model. Since the question generated by the question generating model is formal, e.g., following a correct grammar structure, with correct spelling, with explicit intent, etc., the question is more easily understood and processed by the NLP model, thus a more accurate result may be obtained.

In an aspect, the embodiments of the present disclosure propose to obtain a training dataset for training a question generating model through search logs associated with a search engine. For example, a plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries may be extracted from the search logs. A search query may be a query in a search engine scenario, which may also be considered as an instance of a query. For example, a search query may refer to an original text segment provided by a user or other entity to a search engine to initiate a search for specific information. A search result set corresponding to a search query may include, e.g., one or more search results retrieved by a search engine based on the search query. Each search result may include a title. For example, in the case where the search result is a web page, the title of the search result may be a title of the web page. A title of a search result corresponding to a specific search query may be a question corresponding to the search query. For example, the search query may be "calculate density". A title of a web page retrieved based on the search query may be "How to calculate density?". The title "How to calculate density?" may be considered as a question corresponding to the search query "calculate density". Therefore, the search query and the title of the corresponding search result may be combined into a sample for training the question generating model. A search query may correspond to a search result set. Accordingly, a sample set corresponding to a search query may be obtained. There may be a plurality of search queries in the search logs. Accordingly, a plurality of sample sets corresponding to a plurality of search queries may be combined into an initial training dataset.

In another aspect, the embodiments of the present disclosure propose to filter an initial training dataset based on multiple factors to obtain a training dataset for training the question generating model. A first factor relates to an intent consistency between a title in a sample and a search query in that sample. For example, when filtering the initial training dataset, it may be determined whether intent of a title in each sample is consistent with intent of a search query in

the sample, and a sample in which the intent of the title is inconsistent with the intent of the search query may be removed. A second factor relates to a quality of a title in a sample. For example, when filtering the initial training dataset, for each sample, it may be determined whether a title meets a predetermined quality requirement based on a grammar structure, spelling, and intention explicitness of the title in the sample, and remove a sample in which the title does not meet the predetermined quality requirement. A third factor relates to a text similarity between a title in a sample and a search query in the sample. For example, when filtering the initial training dataset, it may be determined whether there is a text similarity between a title in a sample and a search query in the sample, and remove a sample in which there is a text similarity between the title and the search query. Through considering the three factors described above, samples that do not meet the above conditions may be removed from the initial training dataset, thereby a training dataset including high-quality samples that meet the requirements in terms of e.g., intent consistency, title quality, and text similarity, etc., may be obtained. In addition, the process described above is implemented based on a huge number of search logs in a completely automatic manner without manual intervention, thereby a training dataset including a large number of high-quality samples may be obtained. Such a training dataset may facilitate to train a question generating model with better performance.

In another aspect, the process of obtaining a training dataset according to the embodiments of the present disclosure is based on multiple languages. The training dataset obtaining process proposed by the embodiments of the present disclosure fully considers commonalities and characteristics of different languages, and therefore may be widely applied to an arbitrary language. That is, a training dataset for question generation based on any language may be obtained through the training dataset obtaining process proposed by the embodiments of the present disclosure. This is extremely beneficial for some languages for which training data is scarce or some uncommon languages. Further, in the case where different training datasets based on different languages are obtained, a question generating model may be trained with these training datasets, to obtain a question generating model with multi-lingual capabilities.

In another aspect, the embodiments of the present disclosure propose to perform data augmentation on a training dataset for a NLP task with a question generating model. For example, samples in training datasets for some NLP tasks may include queries. For each sample, a question may be generated from a query in the sample through a question generating model. The generated question may be combined with a part of the sample other than the query to form a new sample. The new sample may be added into the training dataset for the NLP task. In this way, the number of samples may be increased. In addition, since questions generated by the question generating model are formal and easy to be understood by humans and machines, the

quality of the samples has also been improved.

In another aspect, the embodiments of the present disclosure propose to further train a question generating model with a training dataset for a NLP task, to obtain a model for performing the NLP task. For example, the NLP task may be a Grammar Error Correction (GEC) task. The GEC task aims to convert an input text segment with grammar errors into a text segments without grammar errors. The question generating model may be further trained with a training dataset for the GEC task. The training dataset for the GEC task may include a plurality of samples, and each sample may include a text segment with grammar errors and a text segment without grammar errors. The question generating model that is further trained with the training dataset for the GEC task can perform the GEC task when it is actually deployed, e.g., it may convert an input text segment with grammar errors into a text segment without grammar errors.

In another aspect, the embodiments of the present disclosure propose to train other NLP models with a training dataset for a question generating model obtained according to the embodiment of the present disclosure. As an example, the training dataset for the question generating model may also be used to train a GEC model. Since the training dataset for the question generating model includes a large number of high-quality samples, when the training dataset is used to train another NLP model, data augmentation may be performed on a training dataset for the NLP model, e.g., improving the number and quality of samples, which facilitates to train an NLP model with better performance.

FIG. 1 illustrates a schematic diagram 100 including a set of exemplary queries 100a and a set of exemplary questions 100b corresponding to the set of queries. The set of exemplary queries 100a may include queries based on various languages. These queries may be informal, e.g., not following a correct grammar structure, with typing or spelling errors, with inexplicit intents, etc. Each question in the set of exemplary questions 100b may correspond to one query in the set of exemplary queries 100a. Each question may be, e.g., generated from a corresponding query by a question generating model according to the embodiments of the present disclosure. These questions may be formal, e.g., following a correct grammar structure, with correct spelling, with an explicit intent, etc.

A query 110 "sea blue" may be based on English. The query 110 only includes keywords instead of a complete sentence, and its intent is ambiguous. A question 112 corresponding to the query 110 may be "Why is the sea blue?". The question 112 clearly indicates that the question to be answered is "The reason why the sea is blue."

A query 120 "faire crepe (make crepe)" may be based on French. The spelling of "crepe" in the query 120 is wrong, it should be "crêpes". In addition, because of the lack of an article before "crepe", there is grammar error in the query 120. Moreover, the query 120 only includes

keywords instead of a complete sentence, and its intent is ambiguous. A question 122 corresponding to the query 120 may be "Comment faire des crêpes? (How to make crepes?)". The question 122 corrects the spelling and grammar errors in the query 120, and clearly indicates that the question to be answered is "a method for making crepes."

- 5 A query 130 "Körpertemperatur Babys (baby temperature)" may be based on German. Because of the lack of an article before "Babys", there is grammar error in the query 130. In addition, the query 130 only includes keywords instead of a complete sentence, and its intent is ambiguous. A question 132 corresponding to the query 130 may be "Wie misst man die Körpertemperatur eines Babys? (How to take the baby's temperature?)". The question 132 corrects the grammar error in the query 130, and clearly indicates that the question to be answered is "the method for taking the baby's temperature".

- A query 140 "人間の集中力持続時間 (The duration of human concentration)" may be based on Japanese. The query 140 only includes keywords instead of a complete sentence, and its intent is ambiguous. A question 142 corresponding to the query 140 may be "人間の集中力の持続時間はどれくらいですか? (How long is the duration of human concentration)". The question 142 clearly indicates that the question to be answered is "The length of the duration of human concentration."

- A query 150 "Прыжки с парашютом (sky diving)" may be based on Russian. The query 150 only includes keywords instead of a complete sentence, and its intent is ambiguous. A question 152 corresponding to the query 150 may be "Где он подходит для прыжков с парашютом? (Where is it suitable for sky diving?)". The question 152 clearly indicates that the question to be answered is "Locations suitable for sky diving."

- A query 160 "温室五月种植西红柿 (greenhouse May planting tomatoes)" may be based on Chinese. The query 160 does not follow a correct grammar structure, and its intent is also ambiguous. A question 162 corresponding to the query 160 may be "五月份是否可以在温室中种植西红柿? (Can tomatoes be planted in a greenhouse in May?)". The question 162 corrects the grammar error in the query 160 and clearly indicates that the question to be answered is "The feasibility of planting tomatoes in a greenhouse in May".

- It should be appreciated that the queries and questions shown in FIG. 1 are merely exemplary. The queries may also be based on other languages. The question corresponding to each query may also have other forms.

According to the embodiments of the present disclosure, a training dataset may be obtained through search logs associated with a search engine. The obtained training dataset may be at least for training a question generating model. FIG. 2 illustrates exemplary processes 200a and

200b for question generation from queries according to an embodiment of the present disclosure. In the process 200a, a training dataset may be obtained through search logs associated with a search engine. In the process 200a, a question generating model may be trained with the obtained training dataset. The trained question generating model may generate a question
5 suitable for being processed by a NLP model from an input query.

At 210, search logs associated with a search engine may be obtained. The search logs associated with the search engine may include a plurality of search queries. Each of the plurality of search queries may be a word, phrase, sentence, etc., input by a user or other entity into a search box of the search engine, to obtain specific information on the Internet. The search query may be
10 informal, e.g., not following a correct grammar structure, with typing or spelling errors, with inexplicit intent, etc. In addition, the search query may be based on an arbitrary language. For each search query, the search logs may include a search result set corresponding to the search query. Each search result in the search result set may be retrieved from the Internet by a search engine based on the search query. The search result set may also be based on an arbitrary
15 language.

At 220, a plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries may be extracted from the obtained search logs. One search result set in the plurality of search result sets may correspond to one search query in the plurality of search queries.

20 A training dataset at least for training a question generating model may be obtained through the extracted plurality of search queries and the plurality of search result sets. For example, an initial training dataset may be generated based on the extracted plurality of search queries and the plurality of search result sets; and the initial training dataset may be filtered to obtain a training dataset.

25 At 230, an initial training dataset 232 may be generated based on the plurality of search queries and the plurality of search result sets. A search result may include a title. For example, in the case where a search result is a web page, a title of the search result may be a title of the web page. The search result may also include an address. The address may be, e.g., a Uniform Resource Locator (URL). A title of a search result corresponding to a specific search query may
30 be a question corresponding to the search query. For example, the search query may be "calculate density". A title of a web page retrieved based on the search query may be "How to calculate density?". The title "How to calculate density?" may be considered as a question corresponding to the search query "calculate density". Therefore, the search query and the title of the corresponding search result may be combined into a sample for training the question
35 generating model. That is, a sample may include a search query and a title of a search result. For

each search query in the plurality of search queries, the search query may be combined with a title of each search result in a search result set corresponding to the search query, to obtain a sample set corresponding to the search query. Subsequently, a plurality of sample sets corresponding to the plurality of search queries may be combined into the initial training dataset 232. An exemplary process for generating an initial training dataset will be described later in conjunction with FIG. 3.

At 240, the initial training dataset 232 may be filtered to obtain a training dataset 242. When filtering the initial training dataset 232, for each sample in the initial training dataset 232, it may be determined whether the sample should be removed. An exemplary process for removing a sample will be described later in conjunction with FIG. 4.

Through the process 200a above, the training dataset 242 for training a question generating model is obtained. In the process 200a, the question generating model may be trained with the obtained training dataset.

A training dataset 250 in the process 200b may correspond to the training dataset 242 in the process 200a. A question generating model 260 may be trained with the training dataset 250, to obtain a trained question generating model 270. The question generating model may be a sequence-to-sequence model based on the encoder-decoder structure. An encoder and a decoder may be a Deep-Learning Neural Network (DNN) model, e.g., a Recurrent Neural Network (RNN) model, a Long Short-Term Memory (LSTM) model, a Transformer model, etc. The encoder may extract semantic features and syntactic features from a representation of an input query. The decoder may generate a natural language question based on the extracted semantic features and syntactic features.

The trained question generating model 270 is suitable for generating a question suitable for being processed by a NLP model from an input query. As an example, the NLP model may be a MRC model, a QA model, a QP relevance model, a QList relevance model, etc. A query 272 may be provided to the trained question generating model 270. For example, the query 272 may be a query originally input to a NLP model. The trained question generating model 270 may generate a question 274 from the query 272. The question 274 may be provided to the NLP model. Since the question 274 generated by the trained question generating model 270 is formal, e.g., following a correct grammar structure, with correct spelling, with explicit intent, the question 274 is more easily understood and processed by the NLP model, thus a more accurate result may be obtained.

Since the training dataset obtained according to the embodiments of the present disclosure, e.g., the training dataset 242 or 250, is obtained based on a huge number of search logs in an automatic and reliable manner, the training dataset may include a large number of high-quality

samples. Such a training dataset may facilitate to train a question generating model with good performance.

It should be appreciated that the process for question generation from queries described above in conjunction with FIG. 2 is merely exemplary. Depending on actual application requirements, the steps in the process for question generation from queries may be replaced or modified in any manner, and the process may comprise more or fewer steps. For example, in the process 200b, before training the question generating model 260 with the training dataset 250, some pre-training methods may be used to pre-train the question generating model 260 to improve the efficiency of model training. The pre-training methods may include, e.g., Masked Language Model (MLM), Next Sentence Prediction (NSP), etc. In addition, the specific order or hierarchy of the steps in the processes 200a and 200b are merely exemplary, and the process for question generation from queries may be performed in an order different from the described one.

FIG. 3 illustrates an exemplary process 300 for generating an initial training dataset according to an embodiment of the present disclosure. The process 300 may e.g., correspond to step 230 in the process 200a in FIG. 2. Through the process 300, an initial training dataset may be generated based on a plurality of search queries extracted from search logs and a plurality of search result sets corresponding to the plurality of search queries. For example, for each search query in the plurality of search queries, the search query may be combined with a title of each search result in a search result set corresponding to the search query, to obtain a sample set corresponding to the search query. Subsequently, a plurality of sample sets corresponding to the plurality of search queries may be combined into the initial training dataset.

A plurality of search queries extracted from search logs may be a search query 310-1 Q_1 , a search query 310-2 Q_2 , ..., a search query 310-M Q_M , where $M \geq 1$ denotes the number of search queries. The following takes the search query Q_1 as an example to illustrate an exemplary process of obtaining a sample set.

A search result set corresponding to the search query Q_1 extracted from the search logs may be a search result set 320-1 R_1 . The search result set R_1 may include, e.g., a search result 320-1-1 R_{11} , a search result 320-1-2 R_{12} , ..., a search result 320-1-N R_{1N} , where $N \geq 1$ denotes the number of search results. Each search result may include various types of information such as a title, an address, etc. Taking the search result being a web page as an example, the search result may include, e.g., a title of the web page, a URL of the web page, etc. The search result R_{11} may include, e.g., a title T_{11} , an address A_{11} , etc., the search result R_{12} may include, e.g., a title T_{12} , an address A_{12} , etc., the search result R_{1N} may include, e.g., a title T_{1N} , an address A_{1N} , etc.

According to the embodiments of the present disclosure, the search query Q_1 may be combined with a title of each search result in the search result set R_1 , to obtain a sample set 330-1 S_1

corresponding to the search query Q_1 . For example, the search query Q_1 may be combined with the title T_{11} of the search result R_{11} to obtain a sample 330-1-1 S_{11} <search query Q_1 , title T_{11} >; the search query Q_1 may be combined with the title T_{12} of the search result R_{12} to obtain a sample 330-1-2 S_{12} <search query Q_1 , title T_{12} >; the search query Q_1 may be combined with the title T_{IN} of the search result R_{IN} to obtain a sample 330-1-1 S_{IN} <search query Q_1 , title T_{IN} >. The sample S_{11} , the sample S_{12} , ..., the sample S_{IN} may form the sample set S_1 corresponding to the search query Q_1 .

An operation similar to the operation performed on the search query Q_1 may be performed on each search query of the search queries Q_2 to Q_M , so as to obtain a sample set corresponding to each search query. As an example, a search result set corresponding to the search query Q_2 extracted from the search logs may be the search result set 320-2 R_2 . The search query Q_2 may be combined with a title of each search result in the search result set R_2 , to obtain a sample set 330-2 S_2 corresponding to the search query Q_2 . As another example, a search result set corresponding to the search query Q_M extracted from the search logs may be the search result set 320-M R_M . The search query Q_M may be combined with a title of each search result in the search result set R_M , to obtain a sample set 330-M S_M corresponding to the search query Q_M .

Subsequently, the sample set S_1 , the sample set S_2 , ..., the sample set S_M may be combined to obtain an initial training dataset 340. The initial training dataset 340 may, e.g., correspond to the initial training dataset 232 in the process 200a in FIG. 2.

It should be appreciated that the search query Q_1 , the search query Q_2 , ..., the search query Q_M and/or the search result set R_1 , the search result set R_2 , ..., the search result set R_M in FIG. 3 may be based on an arbitrary language. In addition, it should be appreciated that process for generating the initial training dataset described above in conjunction with FIG. 3 is merely exemplary. Depending on actual application requirements, the steps in the process for generating the initial training dataset may be replaced or modified in any manner, and the process may comprise more or fewer steps. For example, when obtaining samples corresponding to each search query, in addition to combining the search query with the title of the corresponding search result, the search query may also be combined with both title and address of the corresponding search result. In addition, the specific order or hierarchy of the steps in the process 300 is merely exemplary, and the process for generating the initial training dataset may be performed in an order different from the described one.

FIG. 4 illustrates an exemplary process 400 for removing a sample according to an embodiment of the present disclosure. The process 400 may be performed for each sample in an initial training dataset generated based on a plurality of search queries and a plurality of search result sets extracted from search logs. For example, the process 400 may be performed for each sample

in the initial training dataset 232 in the process 200a in FIG. 2 or each sample in the initial training dataset 340 in FIG. 3. The following takes a sample 402 S as an example to illustrate an exemplary process for removing a sample. Through the process 400, it may be determined whether the sample S should be removed. The sample S may include, e.g., a search query Q and a title T.

At 404, it may be determined whether an address corresponding to the title T is accessed. The address corresponding to the title T may be an address of a search result corresponding to the title T, e.g., a URL. In an implementation, it may be determined whether the address corresponding to the title T is accessed through determining whether the address is clicked by users. A search result set retrieved based on the search query Q may be presented to users through a search result page. The search result page may include information about each search result, e.g., title, address, summary, etc. Generally speaking, a user will read titles of search results. If the user believes that a title of a search result has the same intent as a search query issued by the user, the user may visit an address corresponding to the title through, e.g., clicking on the address, to gain a deeper understanding of the search result. Therefore, it may be determined whether there is an intent consistency between the title T and the search query Q through determining whether the address corresponding to the title T is accessed.

If it is determined at 404 that the address corresponding to the title T is not accessed, there may be no intent consistency between the title T and the search query Q. In this case, the process 400 may proceed to 424, i.e., the sample S is removed.

If it is determined at 404 that the address corresponding to the title T has been accessed, there may be an intent consistency between the title T and the search query Q. In this case, the process 400 may proceed to 406.

According to the embodiments of the present disclosure, it may be determined whether the title T in the sample S meets a predetermined quality requirement, and in a case where it is determined that the title T does not meet the predetermined quality requirement, the sample S is removed. For example, it may be determined whether the title T meets the predetermined quality requirement based on at least one of a grammar structure, spelling, and intention explicitness of the title T. In an implementation, it may be determined whether the title T meets the predetermined quality requirement through determining whether the address corresponding to the title T is associated with a predetermined address. In another implementation, it may be determined whether the title T meets the predetermined quality requirement through determining whether the title T includes a question word.

For example, at 406, it may be determined whether the address corresponding to the title T is associated with a predetermined address. The predetermined address may be, e.g., an address of

a website that provides a Community Question Answering (CQA) service. As an example, a website that provides CQA service may be Quora, Zhihu, Guokr, etc. The website that provides CQA service has a good credibility. Generally speaking, a title of each page of these websites is good in terms of a grammar structure, spelling, and intention explicitness, therefore, it may be regarded as a high-quality question that meets the predetermined quality requirement.

If it is determined at 406 that the address corresponding to title T is not associated with the predetermined address, e.g., the address of the website that provides the CQA service, then the title T may not be a title of a web page in the website that provides the CQA service, thus it may not be a high-quality question. In this case, the process 400 may proceed to 424, i.e., the sample S is removed.

If it is determined at 406 that the address corresponding to the title T is associated with the predetermined address, e.g., the address of the website that provides the CQA service, then the title T may be the title of the web page in the website that provides the CQA service, thus it may be a high-quality question. In this case, the process 400 may proceed to 408.

At 408, it may be determined whether the title T includes a question word. A question word may be a word used to construct an interrogative sentence. As an example, a question word may include, e.g., why, what, where, how, how to, etc. The question is usually in the form of interrogative sentence. The question word may clearly express inquiry intent, and is a necessary component for constituting the interrogative sentence. Therefore, a high-quality question should include a question word.

If it is determined at 408 that the title T does not include a question word, the title T may not be a high-quality question. In this case, the process 400 may proceed to 424, i.e., the sample S is removed.

If it is determined at 408 that the title T includes a question word, the title T may be a high-quality question. In this case, the process 400 may proceed to 410.

For some languages, e.g., English, French, German, etc., the question word is usually located at the beginning of the question sentence. Preferably, when the title T is based on a predetermined language such as English, French, German, etc., it may be determined whether the beginning of the title T includes a question word. If it is determined that the beginning of the title T does not include a question word, the title T may not be a high-quality question. In this case, the process 400 may proceed to 424, i.e., the sample S is removed. If it is determined that the title T includes a question word, the title T may be a high-quality question. In this case, the process 400 may proceed to 410.

At 410, it may be determined whether the title T includes a predetermined word. The predetermined word may be a word that is unrelated to semantics of the title T. For example, a

title of a web page from a specific website may contain words related to a name of the website, e.g., words related to the name of the website are attached to the beginning or the end of the title. Such words are unrelated to the semantics of the title T. In the case where the sample in which a corresponding address of a title included in the sample is not associated with the predetermined address has been removed through the step 406, the sample involved in the step 410 may be a sample in which the corresponding address of the included in the sample is associated with the predetermined address. It may be determined, through checking titles of web pages of a website corresponding to the predetermined address, what words these titles usually include that are not related to the semantics of the titles, and these words may be used as the predetermined words.

According to the embodiments of the present disclosure, in a case where the title T of the sample S includes a predetermined word, the sample S may be updated through deleting the predetermined word. For example, if it is determined at 410 that the title T includes a predetermined word, the process may proceed to 412, i.e., the predetermined word is deleted.

If it is determined at 410 that the title T does not include a predetermined word, the process 400 may proceed to 414. It has been determined at 404 in the above whether there is an intent consistency between the title T and the search query Q through determining whether the address corresponding to the title T is accessed. According to the embodiments of the present disclosure, it may be further determined whether there is an intent consistency between the title T and the search query Q through an intent consistency model. For example, at 414, an intent consistency score between the title T and the search query Q may be predicted through an intent consistency model. The intent consistency model may be a machine learning model specially trained to predict an intent consistency score between a title T and a search query Q. At 416, it may be determined whether the intention consistency score predicted by the intention consistency model is above a consistency threshold.

If it is determined at 416 that the intent consistency score is not above the consistency threshold, it indicates that there may not be an intent consistency between the title T and the search query Q. In this case, the process 400 may proceed to 424, i.e., the sample S is removed.

If it is determined at 416 that the intent consistency score is above the consistency threshold, it indicates that there may be an intent consistency between the title T and the search query Q. In this case, the process 400 may proceed to 418.

According to the embodiments of the present disclosure, in order to ensure the effectiveness of the training dataset and improve the efficiency of model training, the title T and the search query Q included in each sample should be textually dissimilar. For example, when most of words contained in the title T are the same as some of words contained in the search query Q, the title T and the search query Q may be textually similar. Training a machine learning model, e.g., a

question generating model, with such samples will be inefficient, and the improvement of model performance will be limited. Therefore, when it is determined that there is a text similarity between the title T in the sample S and the search query Q in the sample S, the sample S may be removed.

5 For example, at 418, a text similarity score between the title T and the search query Q may be predicted through a text similarity model. The text similarity model may be a machine learning model specially trained to predict a text similarity score between a title T and a search query Q. As an example, the text similarity model may be a Jaccard similarity calculating model. At 420, it may be determined whether the text similarity score predicted by the text similarity model is
10 below a similarity threshold.

If it is determined at 420 that the text similarity score is not below the similarity threshold, it indicates that the title T and the search query Q may be textually similar. In this case, the process 400 may proceed to 424, i.e., the sample S is removed.

If it is determined at 420 that the text similarity score is below the similarity threshold, it
15 indicates that the title T and the search query Q may not be textually similar. In this case, the process 400 may proceed to 422, i.e., the sample S is retained.

In the process 400, it may be determined whether the sample S should be removed based on multiple factors. The first factor relates to the intent consistency between the title T in the sample S and the search query Q in the sample S. For example, it may be determined whether
20 there is an intent consistency between the title T and the search query Q through the step 404 and/or the steps 414 and 416. The second factor relates to the quality of the title T in the sample S. For example, it may be determined whether the title T meets the predetermined quality requirement through the step 406 and/or the step 408. The third factor relates to the text similarity between the title T in the sample S and the search query Q in the sample S. For
25 example, it may be determined whether there is the text similarity between the title T and the search query Q through the steps 418 and 420. Through considering the three factors described above, samples that do not meet the above conditions may be removed from the initial training dataset, thereby a training dataset including high-quality samples that meet the requirements in terms of intent consistency, title quality, and text similarity, etc., may be obtained. In addition,
30 since the training dataset is obtained based on a huge number of search logs in a completely automatic manner without manual intervention, the training dataset may include a large number of high-quality samples. Such a training dataset may facilitate to train a question generating model with better performance.

Through the process 400, it may be determined whether the sample S should be removed.
35 Generally speaking, when the search query Q or the title T included in the sample S is based on

some languages such as English, French, German, etc., the search query Q or the title T may have some specific formats, e.g., having uppercase letters, lowercase letters, etc. These specific formats usually have special meanings. Preferably, when the sample S is being processed, e.g., when the sample S is being processed through the process 400, the format of the search query Q or the title T included in the sample S may be retained, thereby enabling to more accurately determine whether the sample S should be removed.

It should be appreciated that the process 400 fully considers the commonalities and characteristics of different languages, and therefore may be widely applied to an arbitrary language. In addition, it should be appreciated that the process for removing the sample described above in conjunction with FIG. 4 is merely exemplary. Depending on actual application requirements, the steps in the process for removing the sample may be replaced or modified in any manner, and the process may comprise more or fewer steps. For example, in the process 400, the language on which the search query Q or the title T in the sample S is based may be predetermined, and one or more steps in the process 400 may be performed based on the determined language. For example, when the title T is based on a predetermined language such as English, French, German, etc., it may be determined whether the beginning of the title T includes a question word at 410, rather than determining whether the title T includes a question word. In addition, the specific order or hierarchy of the steps in the process 400 is merely exemplary, and the process for removing the sample may be performed in an order different from the described one. For example, it is also possible to perform the predetermined word deleting operation at 410 and 412 first, and then perform the operation at 408 to determine whether the title T includes a question word. In addition, it should be appreciated that the predetermined word deleting operations at 410 and 412 may update the sample S. In the case where the predetermined word in the title in at least one sample of the initial training dataset is deleted, the initial training dataset is also updated accordingly. Accordingly, the filtering operation performed on the initial training dataset, e.g., the sample removing operation involved in the process 400, may be performed on the updated initial training dataset.

As described above with reference to FIG. 2, a question generating model trained with a training dataset obtained according to the embodiments of the present disclosure is suitable for generating a question suitable for being processed by a NLP model from an input query. According to an embodiment of the present disclosure, a question generating model is also suitable for performing data augmentation on a training dataset for a NLP task. A sample in a training dataset for a NLP task may include a query. For example, a sample in a training dataset for a QA task, a MRC task, a QP relevance task, a QList relevance task, etc. may include a query. Taking the MRC task as an example, the task may find an answer segment for a specific

query from a given passage. Each sample in a training dataset for the MRC task may be, e.g., <query, passage, label of a start position of an answer segment, label of an end position of the answer segment>. For each sample, a question may be generated from a question the sample through a question generating model. The generated question may be combined with a part of the sample other than the query, e.g., "passage", "label of a start position of an answer segment", and "label of an end position of the answer segment" to form a new sample. The new sample may be, e.g., <question, passage, label of a start position of an answer segment, label of an end position of the answer segment>. The new sample may be added into the training dataset for the MRC task. In this way, the number of samples may be increased. In addition, since questions generated by the question generating model are high-quality questions, the quality of the samples has also been improved.

In addition, according to the embodiments of the present disclosure, a question generating model trained with a training dataset obtained according to the embodiments of the present disclosure is adapted to be further trained with a training dataset for a NLP task, so as to perform the NLP task. For example, the NLP task may be a GEC task. The GEC task aims to convert an input text segment with grammar errors into a text segment without grammar errors. The question generating model may be further trained with a training dataset for the GEC task. The training dataset for the GEC task may include a plurality of samples, and each sample may include a text segment with grammar errors and a text segment without grammar errors. The question generating model that is further trained with the training dataset for the GEC task can perform the GEC task when it is actually deployed, e.g., it may convert an input text segment with grammar errors into a text segment without grammar errors.

In addition, as described above with reference to FIG. 2, a training dataset for a question generating task obtained according to the embodiments of the present disclosure may be for training the question generating model. However, the embodiments of the present disclosure are not limited to this. The training dataset for the question generating task obtained according to the embodiments of the present disclosure may be further adapted to train a NLP model other than the question generating model. Since the training dataset for the question generating task is obtained based on a huge number of search logs in an automatic and reliable manner, the training dataset may include a large number of high-quality samples. Data augmentation may be performed on training datasets for other NLP models with the training dataset for the question generating task, to improve the sample quality and the sample number of the training dataset, thereby training an NLP model with better performance. As an example, a sample in a training dataset for the question generating task may include a query and a question, where the query may be a text segment that does not follow a correct grammar structure, and the question may be

a text segment that follows a correct grammar structure. Thus, the training dataset for the question generating task may also be used to train the GEC model. In an implementation, the GEC model may be trained through two stages. In the first stage, the GEC model may be trained through a training dataset for the question generating task. In the second stage, the GEC model may be further trained through the training dataset for the GEC task. In another implementation, the training dataset for the question generating task and the training dataset for the GEC task may be combined into a comprehensive training dataset, and the GEC model may be trained with the comprehensive training dataset. Since the training dataset for the question generating task includes a large number of high-quality samples, the training dataset for the GEC model is augmented, which facilitates to train a GEC model with better performance.

FIG. 5 is a flowchart of an exemplary method 500 for question generation from queries according to an embodiment of the present disclosure.

At 510, search logs associated with a search engine may be obtained.

At 520, a plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries may be extracted from the search logs.

At 530, a training dataset at least for training a question generating model may be obtained through the plurality of search queries and the plurality of search result sets. Each sample in the training dataset may include a search query and a title of a search result. The title may be a question corresponding to the search query.

In an implementation, the obtaining a training dataset may comprise: generating an initial training dataset based on the plurality of search queries and the plurality of search result sets; and filtering the initial training dataset to obtain the training dataset.

The generating an initial training dataset may comprise: for each search query in the plurality of search queries, combining the search query with a title of each search result in a search result set corresponding to the search query, to obtain a sample set corresponding to the search query; and combining a plurality of sample sets corresponding to the plurality of search queries into the initial training dataset.

The filtering the initial training dataset may comprise, for each sample in the initial training dataset: determining whether there is an intent consistency between a title in the sample and a search query in the sample; and in response to determining that there is no intent consistency between the title and the search query, removing the sample.

The determining whether there is an intent consistency between a title and a search query may comprise: determining whether an address corresponding to the title is accessed.

The determining whether there is an intent consistency between a title and a search query may comprise: determining whether there is an intent consistency between the title and the search

query through an intent consistency model.

The filtering the initial training dataset may comprise, for each sample in the initial training dataset: determining whether a title in the sample meets a predetermined quality requirement; and in response to determining that the title does not meet the predetermined quality
5 requirement, removing the sample.

The determining whether a title meets a predetermined quality requirement may comprise: determining whether the title meets the predetermined quality requirement based on at least one of a grammar structure, spelling, and intention explicitness of the title.

The determining whether a title meets a predetermined quality requirement may comprise:
10 determining whether an address corresponding to the title is associated with a predetermined address.

The determining whether a title meets a predetermined quality requirement may comprise: determining whether the title includes a question word.

The title may be based on a predetermined language. The determining whether the title includes
15 a question word may comprise: determine whether a beginning of the title includes a question word.

The filtering the initial training dataset may comprise, for each sample in the initial training dataset: determining whether there is a text similarity between a title in the sample and a search query in the sample; and in response to determining that there is a text similarity between the
20 title and the search query, removing the sample.

The method 500 may further comprise: updating the initial training dataset through deleting a predetermined word in a title in at least one sample in the initial training dataset. The filtering the initial training dataset may comprise: filtering the updated initial training dataset.

In an implementation, the plurality of search queries and/or the plurality of search result sets
25 may be based on an arbitrary language.

In an implementation, the question generating model may be adapted to generate, from an inputted query, a question suitable for being processed by a NLP model.

In an implementation, the question generating model may be adapted to be further trained with a training dataset for a NLP task, so as to perform the NLP task.

30 In an implementation, the training dataset may be further adapted to train a NLP model other than the question generating model.

It should be appreciated that the method 500 may further comprise any step/process for question generation from queries according to the embodiments of the present disclosure as mentioned above.

35 FIG. 6 illustrates an exemplary apparatus 600 for question generation from queries according to

an embodiment of the present disclosure.

The apparatus 600 may comprise: a search log obtaining module 610, for obtaining search logs associated with a search engine; a query and result extracting module 620, for extracting, from the search logs, a plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries; and a training dataset obtaining module 630, for obtaining a training dataset at least for training a question generating model through the plurality of search queries and the plurality of search result sets, each sample in the training dataset including a search query and a title of a search result, the title being a question corresponding to the search query. Moreover, the apparatus 600 may further comprise any other modules configured for question generation from queries according to the embodiments of the present disclosure as mentioned above.

FIG. 7 illustrates an exemplary apparatus 700 for question generation from queries according to an embodiment of the present disclosure.

The apparatus 700 may comprise: at least one processor 710; and a memory 720 storing computer-executable instructions. The computer-executable instructions, when executed, may cause the at least one processor 710 to: obtain search logs associated with a search engine, extract, from the search logs, a plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries, and obtain a training dataset at least for training a question generating model through the plurality of search queries and the plurality of search result sets, each sample in the training dataset including a search query and a title of a search result, the title being a question corresponding to the search query.

In an implementation, the obtaining a training dataset may comprise: generating an initial training dataset based on the plurality of search queries and the plurality of search result sets; and filtering the initial training dataset to obtain the training dataset.

It should be appreciated that the processor 710 may further perform any other step/process of the method for question generation from queries according to the embodiments of the present disclosure as mentioned above.

The embodiments of the present disclosure propose a computer program product for question generation from queries, comprising a computer program that is executed by at least one processor for: obtaining search logs associated with a search engine; extracting, from the search logs, a plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries; and obtaining a training dataset at least for training a question generating model through the plurality of search queries and the plurality of search result sets, each sample in the training dataset including a search query and a title of a search result, the title being a question corresponding to the search query. In addition, the computer programs may

further be performed for implementing any other step/process of the methods for question generation from queries according to the embodiments of the present disclosure as mentioned above.

The embodiments of the present disclosure may be embodied in a non-transitory computer-readable medium. The non-transitory computer-readable medium may comprise instructions that, when executed, cause one or more processors to perform any operations of the methods for question generation from queries according to the embodiments of the present disclosure as mentioned above.

It should be appreciated that all the operations in the methods described above are merely exemplary, and the present disclosure is not limited to any operations in the methods or sequence orders of these operations, and should cover all other equivalents under the same or similar concepts. In addition, the articles “a” and “an” as used in this specification and the appended claims should generally be construed to mean “one” or “one or more” unless specified otherwise or clear from the context to be directed to a singular form.

It should also be appreciated that all the modules in the apparatuses described above may be implemented in various approaches. These modules may be implemented as hardware, software, or a combination thereof. Moreover, any of these modules may be further functionally divided into sub-modules or combined together.

Processors have been described in connection with various apparatuses and methods. These processors may be implemented using electronic hardware, computer software, or any combination thereof. Whether such processors are implemented as hardware or software will depend upon the particular application and overall design constraints imposed on the system. By way of example, a processor, any portion of a processor, or any combination of processors presented in the present disclosure may be implemented with a microprocessor, microcontroller, digital signal processor (DSP), a field-programmable gate array (FPGA), a programmable logic device (PLD), a state machine, gated logic, discrete hardware circuits, and other suitable processing components configured for performing the various functions described throughout the present disclosure. The functionality of a processor, any portion of a processor, or any combination of processors presented in the present disclosure may be implemented with software being executed by a microprocessor, microcontroller, DSP, or other suitable platform.

Software shall be construed broadly to mean instructions, instruction sets, code, code segments, program code, programs, subprograms, software modules, applications, software applications, software packages, routines, subroutines, objects, threads of execution, procedures, functions, etc. The software may reside on a computer-readable medium. A computer-readable medium may include, by way of example, memory such as a magnetic storage device (e.g., hard disk,

floppy disk, magnetic strip), an optical disk, a smart card, a flash memory device, random access memory (RAM), read only memory (ROM), programmable ROM (PROM), erasable PROM (EPROM), electrically erasable PROM (EEPROM), a register, or a removable disk. Although memory is shown separate from the processors in the various aspects presented throughout the present disclosure, the memory may be internal to the processors, e.g., cache or register.

The previous description is provided to enable any person skilled in the art to practice the various aspects described herein. Various modifications to these aspects will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other aspects. Thus, the claims are not intended to be limited to the aspects shown herein. All structural and functional equivalents to the elements of the various aspects described throughout the present disclosure that are known or later come to be known to those of ordinary skilled in the art are expressly incorporated herein and intended to be encompassed by the claims.

CLAIMS

1. A method for question generation from queries, comprising:
obtaining search logs associated with a search engine;
extracting, from the search logs, a plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries; and
obtaining a training dataset at least for training a question generating model through the plurality of search queries and the plurality of search result sets, each sample in the training dataset including a search query and a title of a search result, the title being a question corresponding to the search query.
2. The method of claim 1, wherein the obtaining a training dataset comprises:
generating an initial training dataset based on the plurality of search queries and the plurality of search result sets; and
filtering the initial training dataset to obtain the training dataset.
3. The method of claim 2, wherein the generating an initial training dataset comprises:
for each search query in the plurality of search queries, combining the search query with a title of each search result in a search result set corresponding to the search query, to obtain a sample set corresponding to the search query; and
combining a plurality of sample sets corresponding to the plurality of search queries into the initial training dataset.
4. The method of claim 2, wherein the filtering the initial training dataset comprises, for each sample in the initial training dataset:
determining whether there is an intent consistency between a title in the sample and a search query in the sample; and
in response to determining that there is no intent consistency between the title and the search query, removing the sample.
5. The method of claim 4, wherein the determining whether there is an intent consistency between a title and a search query comprises:
determining whether an address corresponding to the title is accessed.
6. The method of claim 4, wherein the determining whether there is an intent consistency between a title and a search query comprises:
determining whether there is an intent consistency between the title and the search query through an intent consistency model.
7. The method of claim 2, wherein the filtering the initial training dataset comprises, for each sample in the initial training dataset:
determining whether a title in the sample meets a predetermined quality requirement; and

in response to determining that the title does not meet the predetermined quality requirement, removing the sample.

8. The method of claim 7, wherein the determining whether a title meets a predetermined quality requirement comprises:

determining whether the title meets the predetermined quality requirement based on at least one of a grammar structure, spelling, and intention explicitness of the title.

9. The method of claim 7, wherein the determining whether a title meets a predetermined quality requirement comprises:

determining whether an address corresponding to the title is associated with a predetermined address.

10. The method of claim 7, wherein the determining whether a title meets a predetermined quality requirement comprises:

determining whether the title includes a question word.

11. The method of claim 2, wherein the filtering the initial training dataset comprises, for each sample in the initial training dataset:

determining whether there is a text similarity between a title in the sample and a search query in the sample; and

in response to determining that there is a text similarity between the title and the search query, removing the sample.

12. The method of claim 2, further comprising:

updating the initial training dataset through deleting a predetermined word in a title in at least one sample in the initial training dataset, and

wherein the filtering the initial training dataset comprises: filtering the updated initial training dataset.

13. The method of claim 1, wherein the question generating model is adapted to generate, from an inputted query, a question suitable for being processed by a Natural Language Processing (NLP) model.

14. An apparatus for question generation from queries, comprising:

at least one processor; and

a memory storing computer-executable instructions that, when executed, cause the at least one processor to:

obtain search logs associated with a search engine,

extract, from the search logs, a plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries, and

obtain a training dataset at least for training a question generating model through the

plurality of search queries and the plurality of search result sets, each sample in the training dataset including a search query and a title of a search result, the title being a question corresponding to the search query.

15. A computer program product for question generation from queries, comprising a computer program that is executed by at least one processor for:

obtaining search logs associated with a search engine;

extracting, from the search logs, a plurality of search queries and a plurality of search result sets corresponding to the plurality of search queries; and

obtaining a training dataset at least for training a question generating model through the plurality of search queries and the plurality of search result sets, each sample in the training dataset including a search query and a title of a search result, the title being a question corresponding to the search query.

1/5

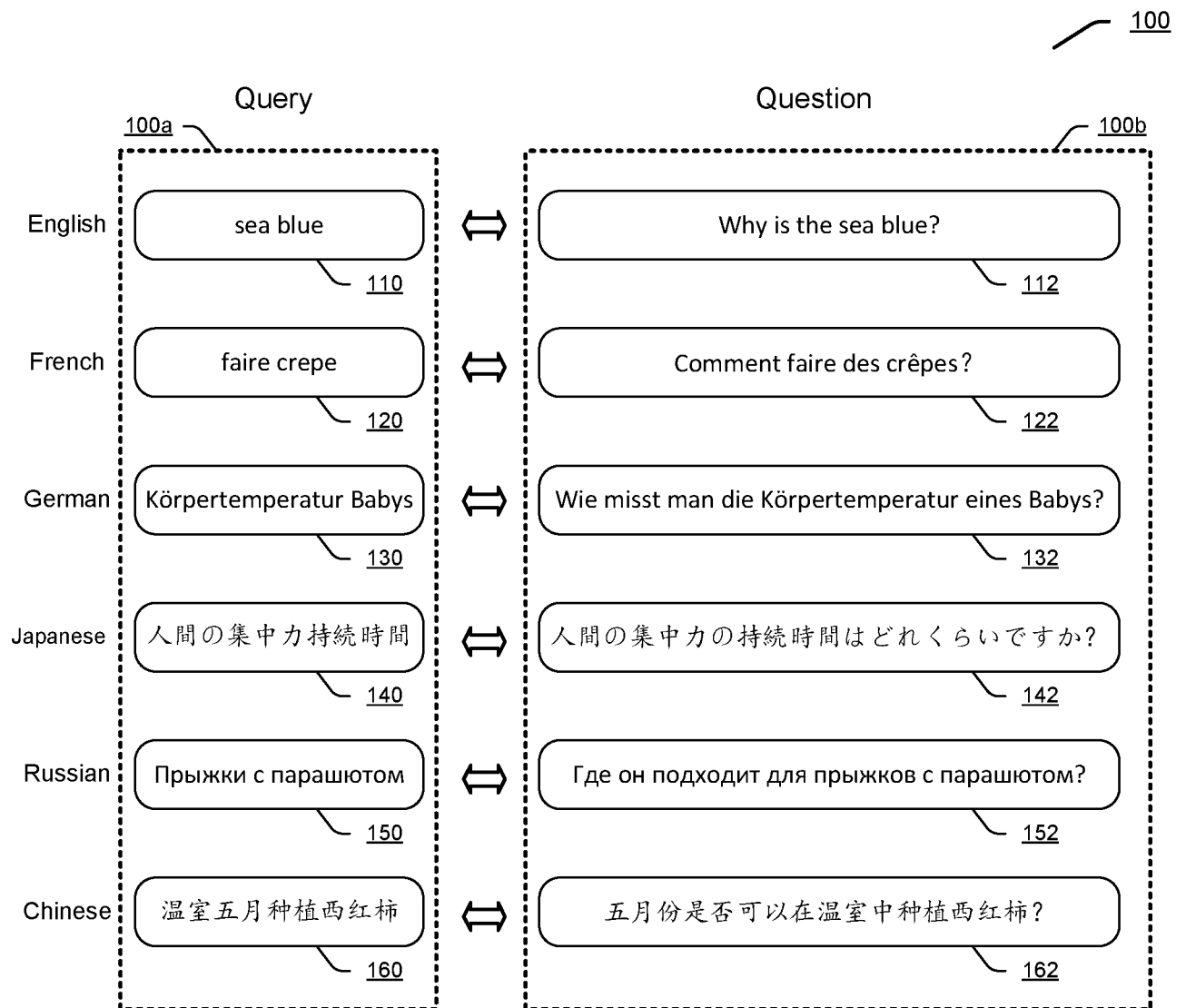


FIG 1

2/5

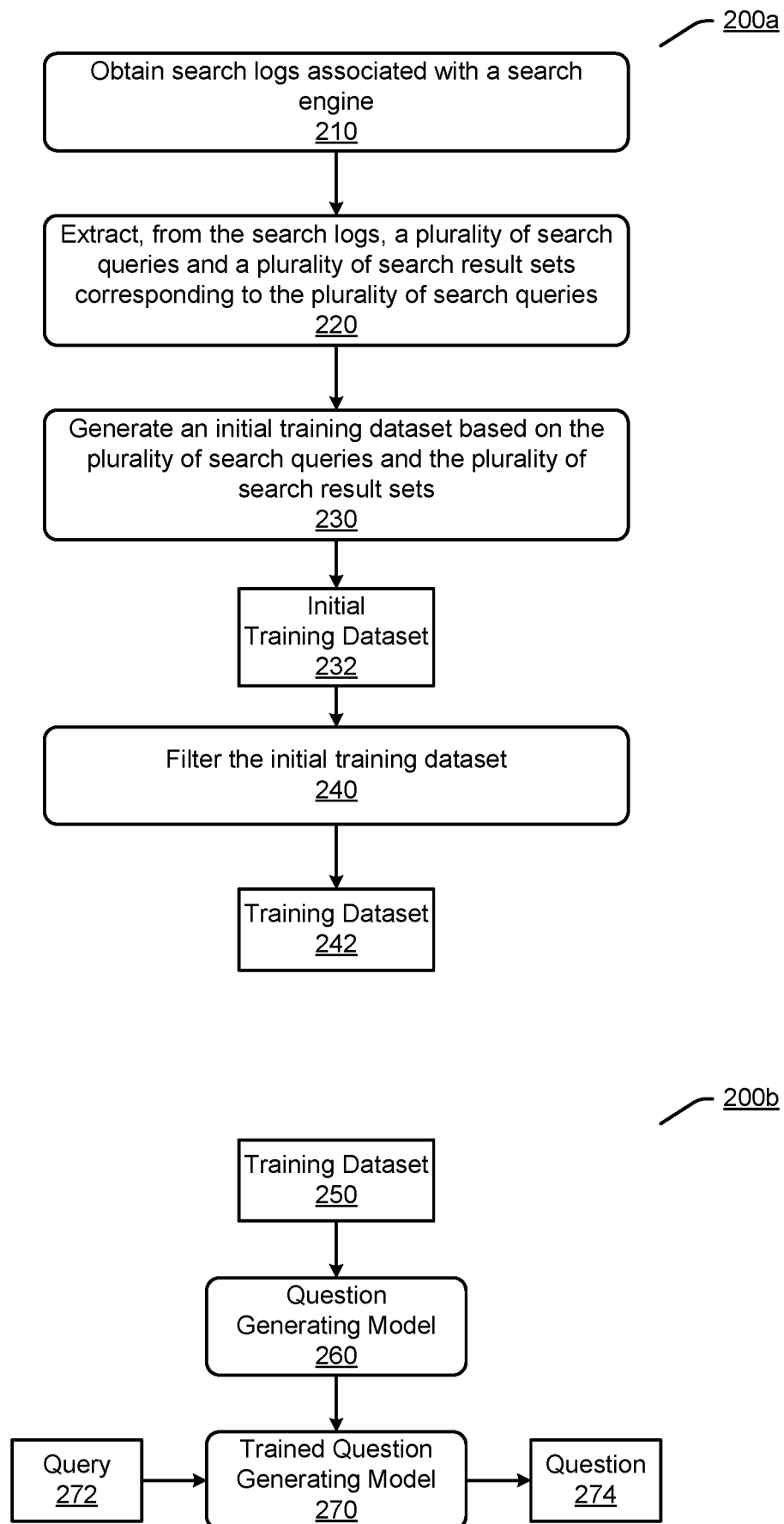


FIG 2

3/5

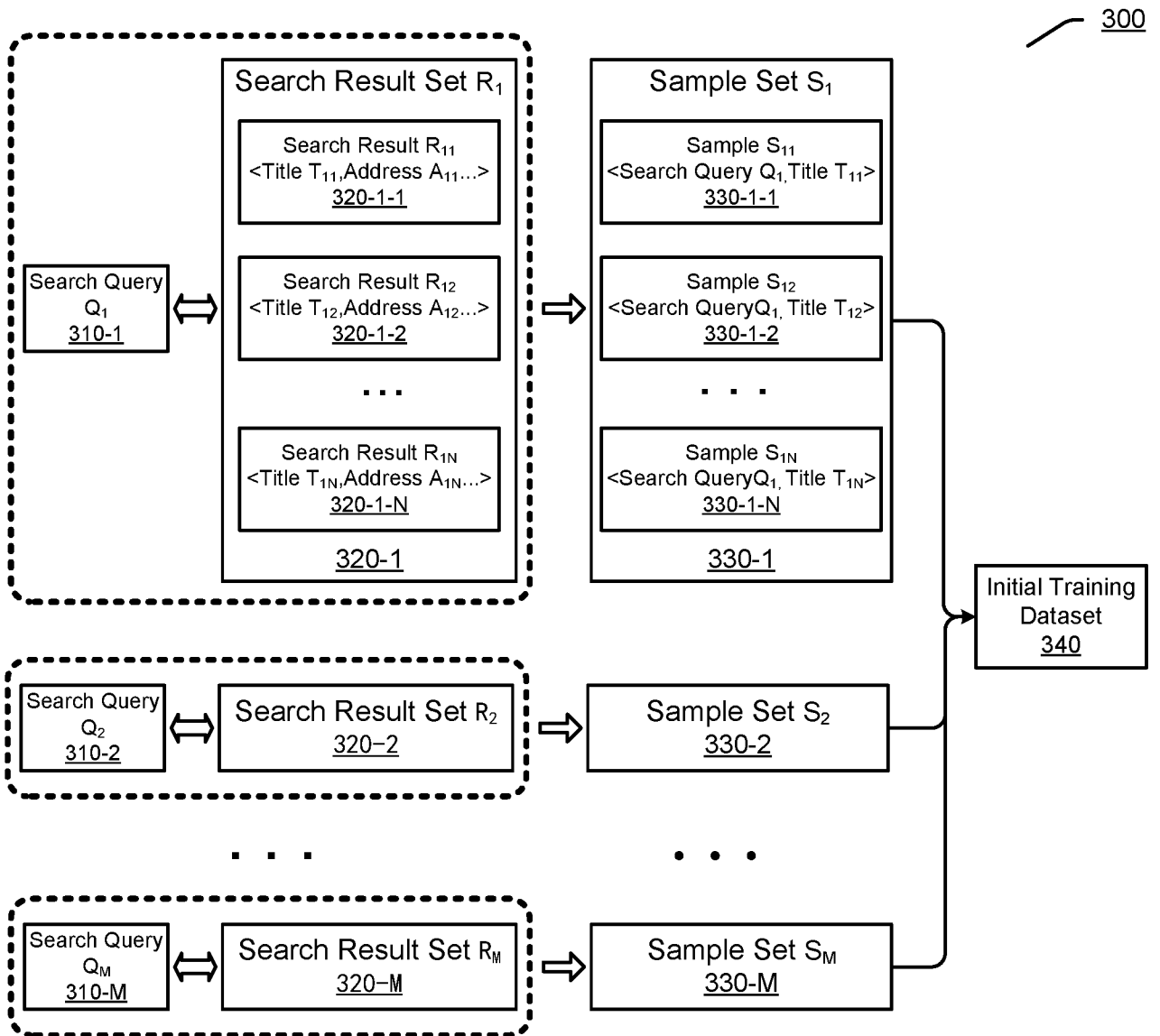


FIG 3

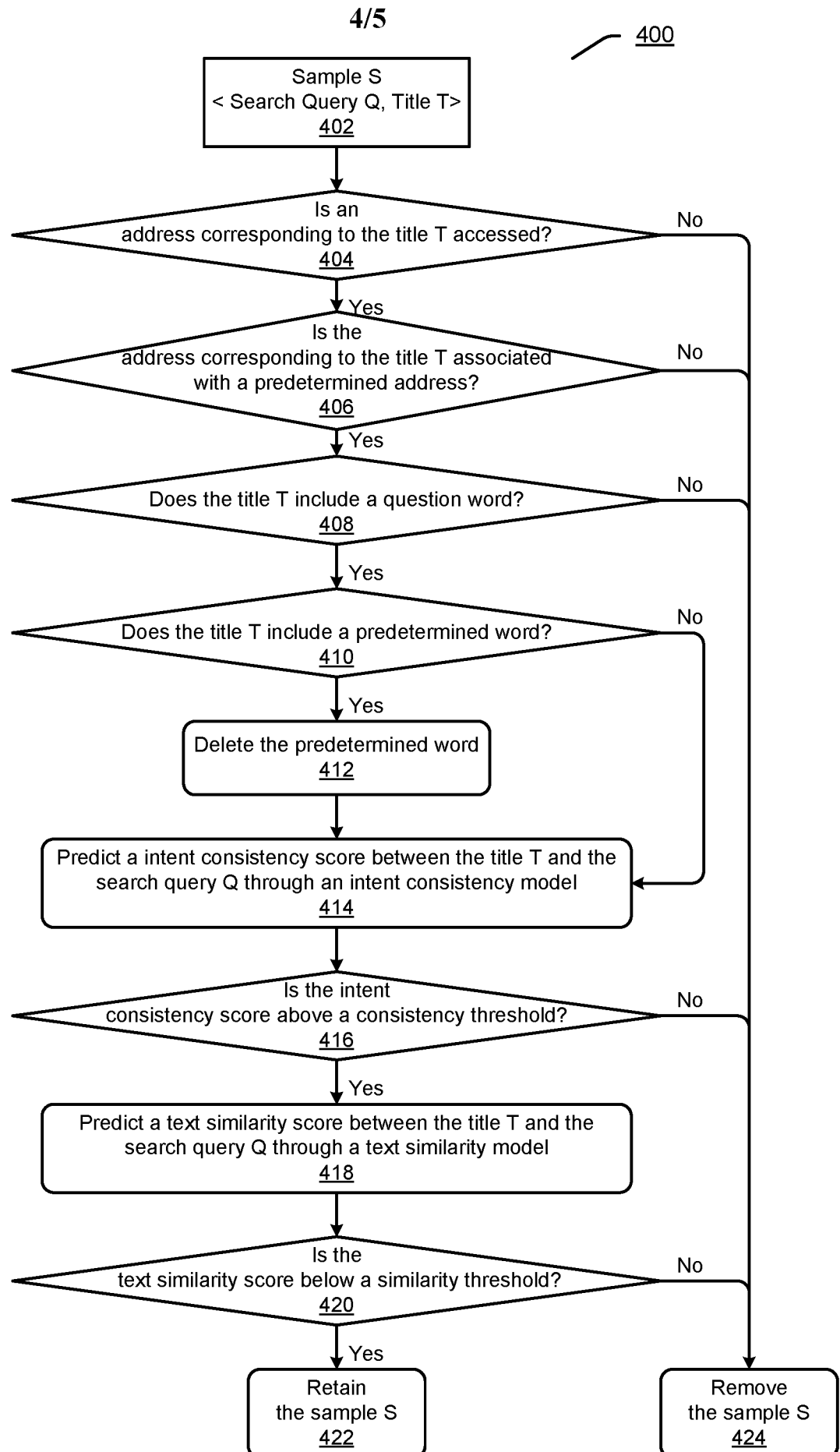


FIG 4

5/5

500

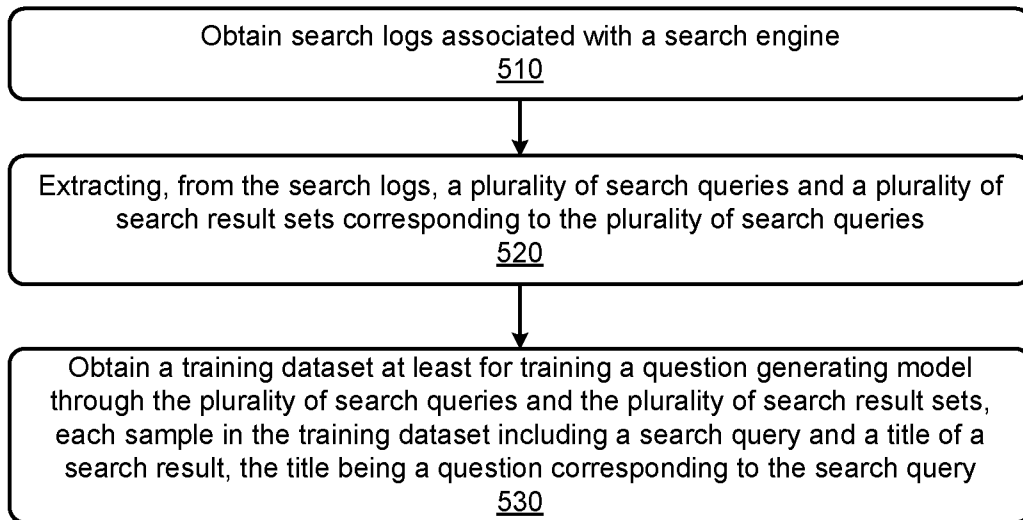


FIG 5

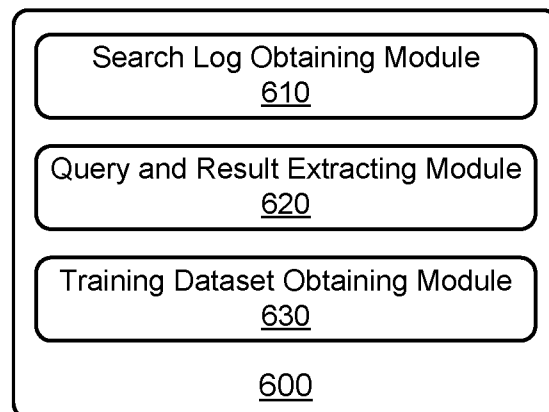


FIG 6

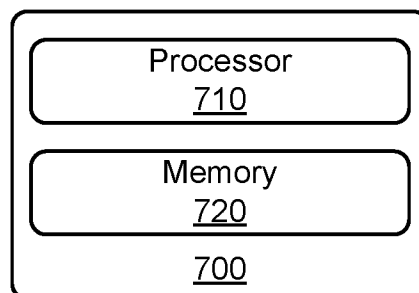


FIG 7

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/031699

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06F16/9032 G06F16/33
ADD.

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)

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 10 339 168 B2 (IBM [US]) 2 July 2019 (2019-07-02) section "Technical Field"; claim 1; figure 2 -----	1-15



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

22 August 2022

Date of mailing of the international search report

02/09/2022

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Thielemann, Benedikt

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2022/031699

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 10339168	B2	02-07-2019	NONE
