(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 December 2022 (29.12.2022)

WIPO | PCT

(10) International Publication Number
WO 2022/269369 A1

(51) International Patent Classification:
G06F 16/783 (2019.01) *G06F 7/00* (2006.01)
G06F 16/93 (2019.01)

(21) International Application Number:
PCT/IB2022/052469

(22) International Filing Date:
18 March 2022 (18.03.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
202141028709 25 June 2021 (25.06.2021) IN

(71) Applicant: **L&T TECHNOLOGY SERVICES LIMITED** [IN/IN]; DLF IT SEZ Park, 2nd Floor – Block 3, 1/124, Mount Poonamallee Road, Ramapuram, TamilNadu, Chennai 600089 (IN).

(72) Inventors: **MALVIYA, Ankit**; 1289, Betul Road Near Guru Shahab College, Chicholi, District, Madhya Pradesh, Betul 460330 (IN). **SINGH, Dr. MADHUSUDAN**; B-603, Ajmera Stone Park, 1st Cross, Neeladri Road, Electronic City - 1, Karnataka, Bangalore 560100 (IN). **BALARAMAN, Mridul**; B 206, SVS Palms 2, Chinnapanahalli Main Road, Dodanekundi, Karnataka, Bangalore 560037 (IN). **SRIVASTAVA, Prakhar**; 4-C, Jawahar Lal Nehru Road, Tagore Town, Uttar Pradesh, Prayagraj 211002 (IN).

(74) Agent: **VARGHESE, Robin**; DLF IT SEZ Park, 2nd Floor – Block 3, 1/124, Mount Poonamallee Road, Ramapuram, Chennai 600089 (IN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM,

(54) Title: A SYSTEM AND METHOD FOR EXAMINING RELEVANCY OF DOCUMENTS

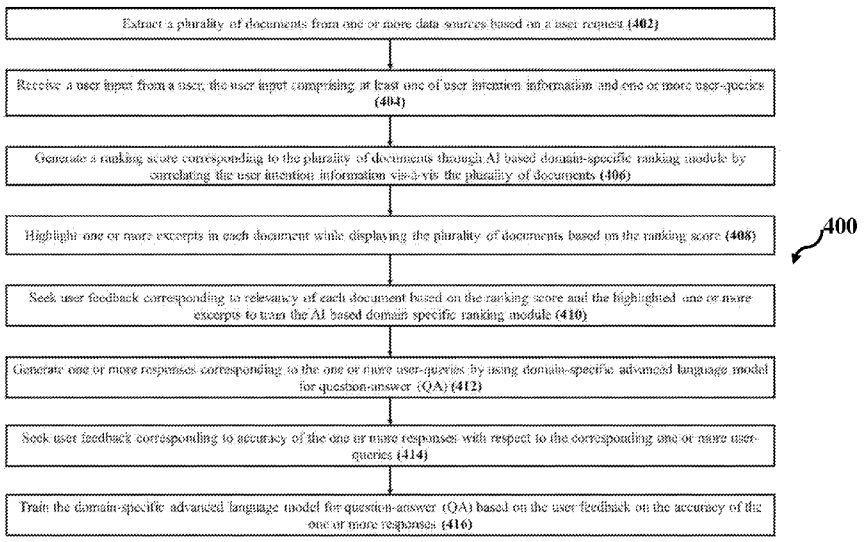

Figure 4

(57) **Abstract:** Disclosed herein is system 102 and method for examining relevancy of documents. The system 102, based on request from the user extracts documents from one or more data sources 210. The system 102 then obtains from the user, user intention information 108 and user queries 110. The system 102 then analyses each document with respect to user intention information 108 in order to determine a relevancy level of each document. The relevancy level is indicated in the form of a ranking score. The system ranks and displays the documents to the user in the order of their scores. The system 102 also highlights important excerpts from the documents and provides one or more responses 222 to the one or more user queries 110 submitted by the user for each and every document. Based on the received responses, user provides feedback for further training the system, thereby achieving better accuracy.

WO 2022/269369 A1

[Continued on next page]

AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, 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).

Declarations under Rule 4.17:

— *as to the identity of the inventor (Rule 4.17(i))*

Published:

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

— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

DESCRIPTION

TECHNICAL FIELD

[001] The present disclosure relates to a field of data analysis. More particularly, it relates to analyzing documents extracted from various sources so as to examine the relevancy of the documents.

BACKGROUND

[002] The following description includes information that may be useful in understanding the present disclosure. It is not an admission that any of the information provided herein is prior art or relevant to the present claims, or that any publication specifically or implicitly referenced is prior art.

[003] Creating a report for a research-based study requires extracting and analysing a large number of documents manually and is therefore a highly tedious task. An individual has to spend long time in finding relevant documents, and post finding the relevant documents, has to study each and every document to decide whether to include the document in the study or not. For instance, an individual wants to prepare a clinical evaluation report (CER) describing clinical evaluation of a medical device. For this purpose, the individual has to first identify clinical data from existing literature, clinical experience, clinical trials, or any combination of the three. The individual has to then appraise the data's relevance, applicability, quality, and significance and eventually articulate their conclusions in the CER, based on the data collected. Further, since these two steps are crucial for the quality of the generated report, these steps must be performed meticulously. Therefore, the whole process becomes highly time consuming and mentally exhausting when performed manually.

[004] There is therefore a need for a method and a system that reduces the manual effort involved in identification and appraisal of data and examining relevancy of documents.

SUMMARY

[005] The present disclosure overcomes one or more shortcomings of the prior art and provides additional advantages discussed throughout the present disclosure. Additional features and advantages are realized through the techniques of the present disclosure. Other embodiments and aspects of the disclosure are described in detail herein and are considered a part of the claimed disclosure.

[006] In one non-limiting embodiment of the present disclosure, a method for examining relevancy of documents is disclosed. The method comprises extracting a plurality of documents from one or more data sources based on a user request. The method further comprises receiving a user input from a user, the user input comprising at least one of user intention information and one or more user-queries. The method further comprises generating a ranking score for each of the plurality of documents through domain specific ranking module by correlating the user intention information vis-à-vis the plurality of documents, wherein the ranking score indicates a relevancy level of each of the plurality of documents with respect to the user intention information. The method further comprises highlighting one or more excerpts in each document while displaying the plurality of documents based on the ranking score. The method further comprises seeking user feedback corresponding to relevancy of each document based on the ranking score and the highlighted one or more excerpts to train the domain specific ranking module.

[007] In yet another non-limiting embodiment of the present disclosure, a system for examining relevancy of documents is disclosed. The system comprises an extraction unit configured to extract a plurality of documents from one or more data sources based on a user request. The system further comprises a receiving unit configured to receive a user input from a user, the user input comprising at least one of user intention information and one or more user-queries. The system further comprises a score generation unit configured to generate a ranking score corresponding to the plurality of documents through domain specific ranking module by correlating the user intention information vis-à-vis the plurality of documents, wherein each ranking score indicates relevancy level of a document with respect to the user intention information. The system further comprises a highlighting unit configured to highlight one or more excerpts in each document while displaying the plurality of documents based on the ranking score. The system further comprises a feedback unit configured to seek user feedback corresponding to relevancy of each document based on the ranking score and the highlighted one or more excerpts. The system further comprises a training unit configured to train the domain specific ranking module based on the user feedback.

[008] The foregoing summary is illustrative only and is not intended to be in any way limiting. In addition to the illustrative aspects, embodiments, and features described above, further aspects, embodiments, and features will become apparent by reference to the drawings and the following detailed description.

BRIEF DESCRITPION OF DRAWINGS

[009] The embodiments of the disclosure itself, as well as a preferred mode of use, further objectives and advantages thereof, will best be understood by reference to the following detailed description of an illustrative embodiment when read in conjunction with the accompanying drawings. One or more embodiments are now described, by way of example only, with reference to the accompanying drawings in which:

[0010] Figures 1A – 1C illustrates an environment **100** of a system for examining relevancy of documents, in accordance with an embodiment of the present disclosure;

[0011] Figure 2 illustrates a block diagram **200** of the system for examining relevancy of documents, in accordance with an embodiment of the present disclosure; and

[0012] Figure 3 illustrates an exemplary embodiment **300** illustrating a domain specific ranking module, in accordance with an embodiment of the present disclosure;

[0013] Figures 3A and 3B illustrate the dependency trees **300A** and **300B** corresponding to a first sentence and a second sentence, in accordance with an embodiment of the present disclosure;

[0014] Figures 3C and 3D illustrates attention matrices **300C** and **300D** corresponding to a first sentence and a second sentence respectively, in accordance with an embodiment of the present disclosure;

[0015] Figure 4 illustrates a flowchart **400** of a method for examining relevancy of documents, in accordance with an embodiment of the present disclosure.

[0016] The figures depict embodiments of the disclosure for purposes of illustration only. One skilled in the art will readily recognize from the following description that alternative embodiments of the structures and methods illustrated herein may be employed without departing from the principles of the disclosure described herein.

DETAILED DESCRIPTION

[0017] The foregoing has broadly outlined the features and technical advantages of the present disclosure in order that the detailed description of the disclosure that follows may be better understood. It should be appreciated by those skilled in the art that the conception and specific embodiment disclosed may be readily utilized as a basis for modifying or designing other structures for carrying out the same purposes of the present disclosure.

[0018] The novel features which are believed to be characteristic of the disclosure, both as to its organization and method of operation, together with further objects and advantages will be better understood from the following description when considered in connection with the

accompanying Figures. It is to be expressly understood, however, that each of the Figures is provided for the purpose of illustration and description only and is not intended as a definition of the limits of the present disclosure.

[0019] Disclosed herein is a system and method for examining relevancy of documents. The task of creating a research report, for instance, an evaluation report, a literature survey report, a qualitative report and a quantitative report, is a highly tedious and time-consuming task. An individual has to spend a long time to collect data that is relevant to his study. Post that, the individual has to analyze the collected data to examine data's relevance, applicability, quality, and significance. However, the first step of collecting the data and the subsequent step of analysing the data is highly tedious and may often take days to complete when undertaken manually. Since the tasks of collecting the data and analysing the data are very crucial with respect to the quality of the generated report, it becomes imperative that these tasks are performed meticulously. There is always a room for improvement since it is not humanly possible for an individual to analyze each and every document meticulously when performing the task manually.

[0020] The present disclosure provides a system that assists a user in the tasks of collecting and analysing the documents. The system, based on a request from the user, extracts documents from the data sources. The extracted documents are broadly based on the search that the user wants to conduct. The system then receives a user's intention behind performing the study along with certain queries that the user wishes to be answered by the system with respect to the extracted documents. The queries are domain-specific and can be customized by the user based on his/her needs and based on the domain of the study. For instance, a user who is working in a bio-medical domain would have different queries in comparison to a user who is working in a nanomaterials domain or transportation domain. The system then analyses each document with respect to user's intention to determine a relevancy level of each document. The relevancy level may be indicated in the form of a ranking score. The system ranks and displays the documents to the user in the order of their ranking scores. This allows the user to be view the documents in the order of relevancy. Further, the system also highlights important excerpts from the documents to help user to focus only on the highlighted excerpts instead of reading the entire document. Furthermore, the system also provides responses to the queries submitted by the user for each and every document. In this manner, it becomes comparatively easier for the user to decide whether to include or exclude a document from his study. The time taken to arrive at such a decision is greatly reduced as the user doesn't have to study the entire document

for examining its relevancy. The detailed working and explanation of the system is described in the subsequent paragraphs.

[0021] **Figure 1A-1C** show an exemplary environment **100** of a system for examining relevancy of documents, in accordance with an embodiment of the present disclosure. It must be noted by a skilled person that the exemplary environment **100** of **Figure 1A** is explained considering that a user wants to create a clinical evaluation report (CER) in the bio-medical domain. A person skilled in the art may further understand that the system **102** may also be implemented in various environments, other than as shown in **Figure 1A**.

[0022] The exemplary environment **100** is explained in conjunction with **Figure 2** that shows a block diagram **200** of a system **102**, in accordance with an embodiment of the present disclosure. Although the present disclosure is explained considering that the system **102** is implemented on a server, it may be understood that the system **102** may be implemented as a tool in a variety of computing systems, such as a laptop computer, a desktop computer, a notebook, a workstation, a mainframe computer, a server, a network server, a cloud-based computing environment.

[0023] In one implementation, the system **102** may comprise an I/O interface **202**, a processor **204**, a memory **206** and the units **208**. The memory **206** may be communicatively coupled to the processor **204** and the units **208**. Further, the memory **208** may store user intention information **108**, one or more queries **110**, one or more responses **222** and one or more tools **224**. The significance and use of each of the stored quantities is explained in the subsequent paragraphs. The processor **204** may be implemented as one or more microprocessors, microcomputers, microcontrollers, digital signal processors, central processing units, state machines, logic circuitries, and/or any devices that manipulate signals based on operational instructions. Among other capabilities, the processor **204** is configured to fetch and execute computer-readable instructions stored in the memory **206**. The I/O interface **202** may include a variety of software and hardware interfaces, for example, a web interface, a graphical user interface, and the like. The I/O interface **202** may enable the system **102** to communicate with other computing devices, such as web servers and external data servers (not shown). The I/O interface **202** may facilitate multiple communications within a wide variety of networks and protocol types, including wired networks, for example, LAN, cable, etc., and wireless networks, such as WLAN, cellular, or satellite. The I/O interface **202** may include one or more ports for connecting many devices to one another or to another server.

[0024] In one implementation, the units **208** may comprise an extraction unit **231**, a receiving unit **232**, a score generation unit **233**, a highlighting unit **234**, a response generation unit **235**, a feedback unit **236**, a training unit **237**, and a decision unit **238**. According to embodiments of the present disclosure, these units **231-238** may comprise hardware components like processor, microprocessor, microcontrollers, application-specific integrated circuit for performing various operations of the system **102**. It must be understood that the processor **204** may perform all the functions of the units **231-238** according to various embodiments of the present disclosure.

[0025] Referring to **Figure 1A**, the environment **100** shows the system **102** that extracts a plurality of documents from one or more data sources **210** based on a user request. In accordance with **Figure 2**, the extraction of documents is performed by the extraction unit **231**. In one embodiment, the user request comprises a set of keywords **104** pertaining to the plurality of documents. For instance, if the subject-matter of the CER that the user wants to create the report related to "quantitative determination of Access alpha-fetoprotein (AFP) in human serum as an aid to manage patients with non-seminomatous testicular cancer", the user provides a set of keywords **104** that may comprise "AFP", "human", "testicular cancer". The extraction unit **231** based on the set of keywords extracts the plurality of documents. In another embodiment, the user request comprises the user providing a plurality of unique IDs **106** pertaining to the plurality of documents. The plurality of unique IDs are provided to the extraction unit **231** that extracts the plurality of documents corresponding to the plurality of unique IDs **106**. In accordance with the exemplary environment **100** depicted in **Figure 1A**, the plurality of documents extracted is 250.

[0026] Once the plurality of documents are extracted, the receiving unit **232** receives a user input. In one embodiment, the user input comprises user intention information **108** indicating using natural language, an intention behind creating said CER. In accordance with the exemplary environment **100**, the user intention information **110** may read like: "*The Access alpha-fetoprotein (AFP) assay is a paramagnetic particle, chemiluminescent immunoassay for use with the Access Immunoassay Systems for the quantitative determination of AFP in Human-serum as an aid in management of patients with non-seminomatous testicular cancer. Maternal serum and amniotic fluid to aid in detection of fetal Open Neural Tube Defects (ONTD)*". The user input further comprises one or more user queries **110** pertaining to the plurality of documents as shown in **Figure 1A and Figure 1C**. For instance, the one or more user queries **110** may comprise questions such as "What is the intent of the document?", "What is the device used for?", "What is the nature of participants of the study?" etc. It must be noted by a skilled

person that the one or more user queries **110** are customizable based on the needs of the user. That is, the one or more user queries **110** can be edited or drafted based on the kind of study the user is undertaking and the domain the study belongs to. For instance, according to an exemplary environment **100** the study is of bio-medical domain, the one or more queries **110** are relevant to the bio-medical domain. However, a user whose study is in the domain of nano-materials may have different set of queries than mentioned above and can customize the queries based on his needs.

[0027] To examine the relevancy of the plurality of documents with respect to the user intention information **108**, the score generation unit **233** correlates user intention information **108** vis-à-vis the plurality of documents to generate a ranking score for each document. The ranking score indicates relevancy level of each document with respect to user intention information **108**. In accordance with the exemplary environment **100**, depicted in **Figure 1A**, the user intention information **108** is correlated vis-à-vis each of the 250 extracted documents.

[0028] To correlate the user intention information **108** vis-à-vis the plurality of documents, the score generation unit **233** employs a domain specific ranking module **300** as illustrated in **Figure 3**. The domain specific ranking module **300** is stored in the memory **206** as one of the one or more tools **224**. The details of the domain specific ranking module **300** is explained with the help of an example in the subsequent paragraphs.

[0029] At step **302** of the domain specific ranking module **300**, a sentence pair is input. The sentence pair comprises a first sentence derived from user intention information **110** and a second sentence derived from one of the plurality of documents. For instance, the sentence pair may comprise –

First Sentence – AFP is detected in patients with NSTC; and

Second Sentence – Elevated AFP revealed mixed germ cell tumours.

[0030] At step **304** of the domain specific ranking module **300**, a first dependency tree **300A** corresponding to the first sentence and a second dependency tree **300B** corresponding to the second sentence is generated. The first and the second dependency trees **300A, 300B** represent grammatical structure of the first sentence and the second sentence to uncover their syntactic structure as illustrated in **Figures 3A** and **3B**.

[0031] At step **306** of the domain specific ranking module **300**, a domain specific advanced language model is employed to encode a first set of tokens (or words) corresponding to the first sentence and a second set of tokens (or words) corresponding to the second sentence in domain-specific manner so as to not alter a meaning of the first sentence and the second sentence. The

working of the domain-specific advanced language model for the first sentence and the second sentence is illustrated in tables 1 and 2, respectively.

Table 1: Domain Specific Transformation for first sentence

First Sentence	AFP	Is	detected	In	Patients	With	NSTC	.
Tokens	'afp'	'is'	'detected'	'in'	'patients'	'with'	'nstc'	'.'
Encoding	28634	165	2490	121	568	190	3281	205

Table 2: Domain Specific Transformation for second sentence

First Sentence	Elevated	AFP	Revealed	Mixed	Germ	Cell	tumours	.
Tokens	'elevated'	'afp'	'revealed'	'mixed'	'germ'	'cell'	'tumours'	'.'
Encoding	5161	28634	2861	4055	2927	377	11571	205

[0032] However, if instead of domain-specific advanced language model, a normal language model is used, the tokens of the first sentence would be generated as – ['a', '##f' '##p', 'is', 'detected', 'in', 'patients', 'with', 'n', '##s', '##t', '##c', '.'"] and the tokens of the second sentence would be generated as - ['elevated', 'a', '##f', '##p', 'revealed', 'mixed', 'ge', '##rm', 'cell', 'tu', '##mour', '##s', '.']. Hence, the meaning of the first and the second sentence would be lost.

[0033] At step **308** of the domain specific ranking module **300**, a first attention matrix corresponding to the set of words of the first sentence and a second attention matrix corresponding to the set of words of the second sentence is created to establish explicit dependency relationship between the set of words of the first sentence and the set of words of the second sentence. The attention matrices corresponding to the set of words of first sentence and the set of words of the second sentence is depicted in **Figures 3C** and **3D** respectively, where the highlighted cells are with respect to the attention weights.

[0034] At step **310** of the domain specific ranking module **300**, mutual attention technique is employed to generate semantic representations corresponding to the first sentence and the second sentence. Mutual attention takes into consideration synonyms and near-synonyms of a word at different positions such that the similarity between words does not go unnoticed due to the positioning of words or due to the use of synonyms or near-synonyms. For instance, the first sentence "AFP is detected in patients with NSTC" can also be expressed as "NSTC

patients are detected with AFP levels" and the second sentence "Elevated AFP revealed mixed germ cell tumours" can also be expressed as "Mixed germ cell tumours are detected with high AFP".

[0035] At step **312** of the domain specific ranking module **300**, the set of words of the first sentence and the set of words of the second sentence are finetuned to generate semantically meaningful sentence representations or embeddings. The finetuning is accomplished by providing the set of words of the first sentence and the second sentence to a dual branch network with shared parameters to minimize a loss function. This dual branch network with shared parameters is pre-trained to minimize the loss function by providing the model with pairs of similarity-feedback-labelled sentences and minimizing the loss between the semantic representations of the sentences.

[0036] At step **314** of the domain specific ranking module **300**, ranking score is generated for the finetuned first sentence and the second sentence by employing a similarity measurement technique such as cosine similarity, Manhattan distance similarity, Euclidean distance similarity etc. It may however be noted by a skilled person that the similarity measurement techniques other than the ones mentioned above may also be used.

[0037] The procedure described in steps **302** – **314** is iteratively carried out to determine a ranking score for each of the plurality of documents in accordance with the user intention information **108**.

[0038] Once the ranking score for the plurality of documents is generated, the plurality of documents are displayed in order of their ranking scores as depicted in **Figure 1B**. In one embodiment, the document with highest ranking score may be displayed at the top. However, in another embodiment, the document with highest ranking score may be displayed at the bottom. Further, as shown in **Figure 1B**, each document is displayed with its unique ID and its ranking score. It may be noted by a skilled person that the ranking score depicted in Figure 1B is out of **100**, however, the ranking score can be generated or represented in other forms.

[0039] Further, for each displayed document, the highlighting unit **234** highlights one or more excerpts that are important in view of the user intention information **108**. It may be understood by a skilled person that one or more excerpts may comprise one or more paragraphs and/or one or more sentences. To highlight one or more important paragraphs, the domain specific ranking module **300** is employed by the highlighting unit **234** to identify a paragraph in a document with highest relevancy to the user intention information **108**. Further, to highlight one or more important sentences in an important paragraph, attention weights of respective tokens (or words) of the sentences of the important paragraph and the user intention information **108** are

calculated by the score generation unit **233**. Based on the attention weights, each token is assigned a score. Based on the scores, a start and an end of the important part of the most relevant paragraph is obtained which is highlighted as an important sentence by the highlighting unit **234**.

[0040] Highlighting of the one or more excerpts allows the user to readily identify the sections of a document that are most relevant to his/her intended purpose as defined by him/her in the user intention information **108**.

[0041] Further, as shown in **Figure 1C**, the system **102** allows a user to view one or more responses to the one or more user queries **110** provided by him/her by selecting a Response tab as shown in. The responses to the one or more user queries **110** is provided in **Figure 1B** by a response generation unit **235** that employs a domain specific advanced language model for question-answers (QA) stored as one or more tools **224** in the memory **206** of the system **102**. The generated one or more responses **222** are stored in the memory **206**. Further, based on the generated one or more responses, the feedback unit **236** allows the user to provide a feedback to the one or more responses. For instance, as shown in exemplary environment **100** of **Figure 1A**, the user is provided with an option to validate the response provided by the response generation unit **235** or change the response if he/she is not satisfied with the response provided or if the response generation unit was incapable of providing a response. Based on the feedback, the training unit **237**, re-trains the advanced language model for question-answer (QA) .

[0042] The system **102** further enables the user to select one or more relevant documents by providing a decision tab as shown in **Figure 1A**. The decision tab is controlled by the decision unit **238** and provides options to the user to either "Include" or "Exclude" a particular document. However, if the user is not able to decide, the decision unit **238** allows the user to revisit a particular document by providing "Cannot make a decision" option to the user.

[0043] Therefore, the system **102** through the interaction of its various components, makes it easier for the viewer to identify relevant documents with minimal time and effort, thereby enhancing user's experience. Further, the manner in which the ranking score for each of the plurality of documents is generated is highly extensive and accurate since the domain specific ranking module **300** employs a combination of various techniques.

[0044] **Figure 4** depicts a method **400** for examining relevancy of a plurality of documents, in accordance with an embodiment of the present disclosure.

[0045] As illustrated in **Figure 4**, the method **400** includes one or more blocks illustrating a method for managing information in the manufacturing plant. The method **400** may be described in the general context of computer executable instructions. Generally, computer

executable instructions may include routines, programs, objects, components, data structures, procedures, modules, and functions, which perform specific functions or implement specific abstract data types.

[0046] The order in which the method **400** is described is not intended to be construed as a limitation, and any number of the described method blocks may be combined in any order to implement the method. Additionally, individual blocks may be deleted from the methods without departing from the spirit and scope of the subject matter described.

[0047] At block **402**, the method **400** may include extracting a plurality of documents from one or more data sources **210** based on a user request.

[0048] At block **404**, the method **400** may include receiving a user input from a user, the user input comprising at least one of user intention information **108** and one or more user-queries **110**.

[0049] At block **406**, the method **400** may include generating a ranking score for each of the plurality of documents by correlating the user intention information **108** vis-à-vis the plurality of documents.

[0050] At block **408**, the method **400** may include highlighting one or more excerpts in each document while displaying the plurality of documents based on the ranking score.

[0051] At block **410**, the method **400** may include seeking user feedback corresponding to relevancy of each document based on the ranking score and the highlighted one or more excerpts to train the domain specific ranking module.

[0052] At block **412**, the method **400** may include generating one or more responses **222** corresponding to the one or more user-queries **110** by using domain-specific advanced language model for question-answer (QA).

[0053] At block **414**, the method **400** may include seeking user feedback corresponding to accuracy of the one or more responses **222** with respect to the corresponding one or more user-queries **110**.

[0054] At block **416**, the method **400** may include training the domain-specific advanced language model for question-answer (QA) based on the user feedback on the accuracy of the one or more responses **222**.

[0055] A description of an embodiment with several components in communication with each other does not imply that all such components are required. On the contrary, a variety of optional components are described to illustrate the wide variety of possible embodiments of the disclosure.

[0056] When a single device or article is described herein, it will be clear that more than one device/article (whether they cooperate) may be used in place of a single device/article. Similarly, where more than one device or article is described herein (whether they cooperate), it will be clear that a single device/article may be used in place of the more than one device or article or a different number of devices/articles may be used instead of the shown number of devices or programs. The functionality and/or the features of a device may be alternatively embodied by one or more other devices which are not explicitly described as having such functionality/features. Thus, other embodiments of the disclosure need not include the device itself.

[0057] Finally, the language used in the specification has been principally selected for readability and instructional purposes, and it may not have been selected to delineate or circumscribe the inventive subject matter. It is therefore intended that the scope of the disclosure be limited not by this detailed description, but rather by any claims that issue on an application based here on. Accordingly, the embodiments of the present disclosure are intended to be illustrative, but not limiting, of the scope of the disclosure, which is set forth in the following claims.

[0058] While various aspects and embodiments have been disclosed herein, other aspects and embodiments will be apparent to those skilled in the art. The various aspects and embodiments disclosed herein are for purposes of illustration and are not intended to be limiting, with the true scope and spirit being indicated by the following claims.

We Claim:

1. A method for examining relevancy of documents, wherein the method comprising:

extracting (402) a plurality of documents from one or more data sources (210) based on a user request;

receiving (404) a user input from a user, the user input comprising at least one of user intention information (108) and one or more user-queries (110);

generating (406) a ranking score for each of the plurality of documents through domain specific ranking module by correlating the user intention information (108) vis-à-vis the plurality of documents, wherein the ranking score indicates a relevancy level of each of the plurality of documents with respect to the user intention information (108);

highlighting (408) one or more excerpts in each document while displaying the plurality of documents based on the ranking score; and

seeking (410) user feedback corresponding to relevancy of each document based on the ranking score and the highlighted one or more excerpts to train the domain specific ranking module.

2. The method as claimed in claim 1, further comprising:

generating (412) one or more responses (222) corresponding to the one or more user-queries (110) by using domain-specific advanced language model for question-answer (QA)

seeking (414) user feedback corresponding to accuracy of the one or more responses (222) with respect to the corresponding one or more user-queries (110); and

training (416) the advanced language model for question-answer (QA) based on the user feedback to the accuracy of the one or more responses (222).

3. The method as claimed in claim 1, wherein the user request comprises:

a set of keywords (104) relevant to the plurality of documents; or

a plurality of unique IDs (106) pertaining to the plurality of documents.

4. The method as claimed in claim 1, wherein correlating the user intention information vis-à-vis the plurality of documents comprises:

generating (302, 304) a first dependency tree corresponding to a first sentence and a second dependency tree corresponding to a second sentence of a sentence pair;

encoding (306) a first set of tokens corresponding to the first sentence and a second set of tokens corresponding to the second sentence by employing a domain-specific advanced language model, wherein the first set of tokens correspond to a set of words in the first sentence and the second set of tokens correspond to a set of words in the second sentence,

generating (308) a first attention matrix corresponding to the set of words of the first sentence and a second attention matrix corresponding to the set of words of the second sentence to establish dependency relationship between set of words of the first sentence and the set of words of the second sentence;

generating (310) semantic representations corresponding to the first sentence and the second sentence by employing mutual attention;

finetuning (312) the set of words of the first sentence and the set of words in the second sentence by updating a plurality of weights in the network by using specific loss function to ensure that the generated embeddings are semantically relevant and can be differentiated with similarity measures to generate semantically meaningful sentence embeddings corresponding to the first sentence and the second sentence; and

generating (314) a ranking score for the finetuned first sentence and the second sentence by employing at least one similarity measurement technique.

5. The method as claimed in claim 1, wherein highlighting (408) the one or more excerpts in each document comprises:

highlighting at least one important paragraph in each document by:

generating a ranking score for each paragraph in each document with respect to the user intention information (108); and

selecting at least one paragraph from each document with highest ranking score with respect to the user intention information (108); and

highlighting at least one important sentence in the at least one important paragraph by:

calculating attention weights for the plurality of words in the at least one important paragraph and the user intention information (108);

assigning a score to each of the plurality of words; and

by obtaining a start point and an end point from the at least one important paragraph based on the maximum score.

6. A system for examining relevancy of documents, wherein the system comprises:

an extraction unit (231) configured to extract a plurality of documents from one or more data sources (210) based on a user request;

a receiving unit (232) configured to receive a user input from a user, the user input comprising at least one of user intention information (108) and one or more user-queries (110);

a score generation unit (233) configured to generate a ranking score corresponding to the plurality of documents through domain specific ranking module (300) by correlating the user intention information (108) vis-à-vis the plurality of documents, wherein each ranking score indicates relevancy level of a document with respect to the user intention information (108);

a highlighting unit (234) configured to highlight one or more excerpts in each document while displaying the plurality of documents based on the ranking score;

a feedback unit (236) configured to seek user feedback corresponding to relevancy of each document based on the ranking score and the highlighted one or more excerpts; and

a training unit (237) configured to train the domain specific ranking module (300) based on the user feedback.

7. The system as claimed in claim 7, further comprising:

a response generation unit (235) configured to generate one or more responses (222) corresponding to the one or more user-queries (110) by using advanced language model for question-answer (QA);

wherein:

the feedback unit (236) is further configured to seek user feedback corresponding to accuracy of the one or more responses (222) with respect to the corresponding one or more user-queries (110); and

the training unit (237) is further configured to train the advanced language model for question-answer (QA) based on the user feedback on the accuracy of the one or more responses (222).

8. The system as claimed in claim 7, wherein the user request comprises:

a set of keywords (104) relevant to the plurality of documents; or

a plurality of unique IDs (106) pertaining to the plurality of documents.

9. The system as claimed in claim 7, wherein to correlate the user intention information vis-à-vis the plurality of documents, the score generation unit (233) is further configured to:

generate a first dependency tree corresponding to a first sentence and a second dependency tree corresponding to a second sentence of a sentence pair;

encode a first set of tokens corresponding to the first sentence and a second set of tokens corresponding to the second sentence by employing a domain-specific advanced language model, wherein the first set of tokens correspond to a set of words in the first sentence and the second set of tokens correspond to a set of words in the second sentence,

generate a first attention matrix corresponding to the set of words of the first sentence and a second attention matrix corresponding to the set of words of the second sentence to establish dependency relationship between set of words of the first sentence and the set of words of the second sentence;

generate semantic representations corresponding to the first sentence and the second sentence by employing mutual attention;

finetune the set of words of the first sentence and the set of words in the second sentence by updating a plurality of weights in the network by using specific loss function to ensure that the generated embeddings are semantically relevant and can be differentiated with similarity measures to generate semantically meaningful sentence embeddings corresponding to the first sentence and the second sentence; and

generate a ranking score for the finetuned first sentence and the second sentence by employing at least one similarity measurement technique.

10. The system as claimed in claim 7, wherein to highlight the one or more excerpts in each document, the highlighting unit (234) is further configured to:

highlight at least one important paragraph in each document by:

generating a ranking score for each paragraph in each document with respect to the user intention information (108); and

selecting at least one paragraph from each document with highest ranking score with respect to the user intention information (108); and

highlight at least one important sentence in the at least one important paragraph by:

calculating attention weights for a plurality of words in the at least one important paragraph and the user intention information (108);

assigning a score to each of the plurality of words; and

obtaining a start point and an end point from the at least one important paragraph based on the score.

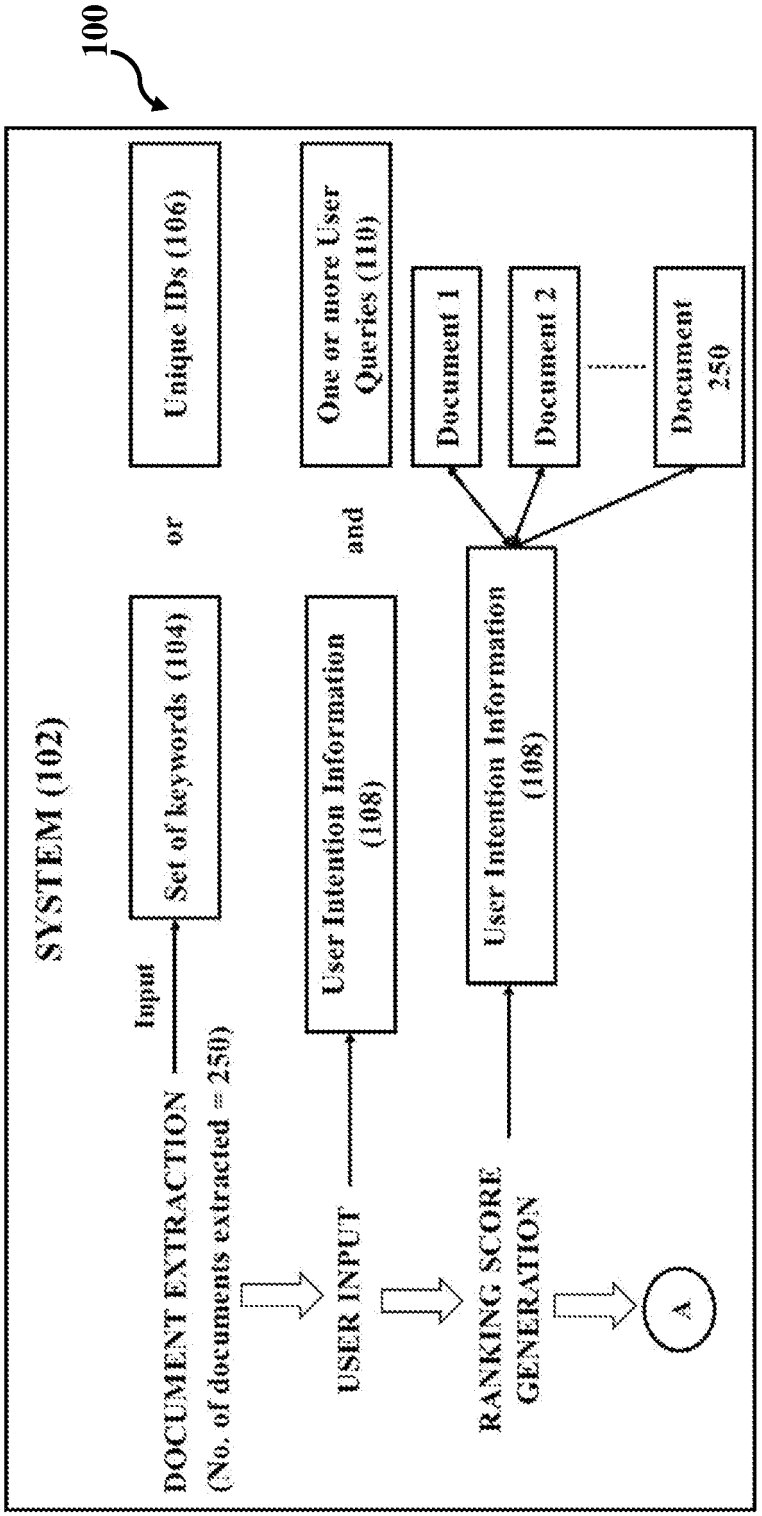
100
SYSTEM (102)
DOCUMENT EXTRACTION
(No. of documents extracted = 250)
Input
Set of keywords (104)
or
Unique IDs (106)
USER INPUT
User Intention Information (108)
and
One or more User Queries (110)
RANKING SCORE GENERATION
User Intention Information (108)
Document 1
Document 2
Document 250
A

Figure 1A

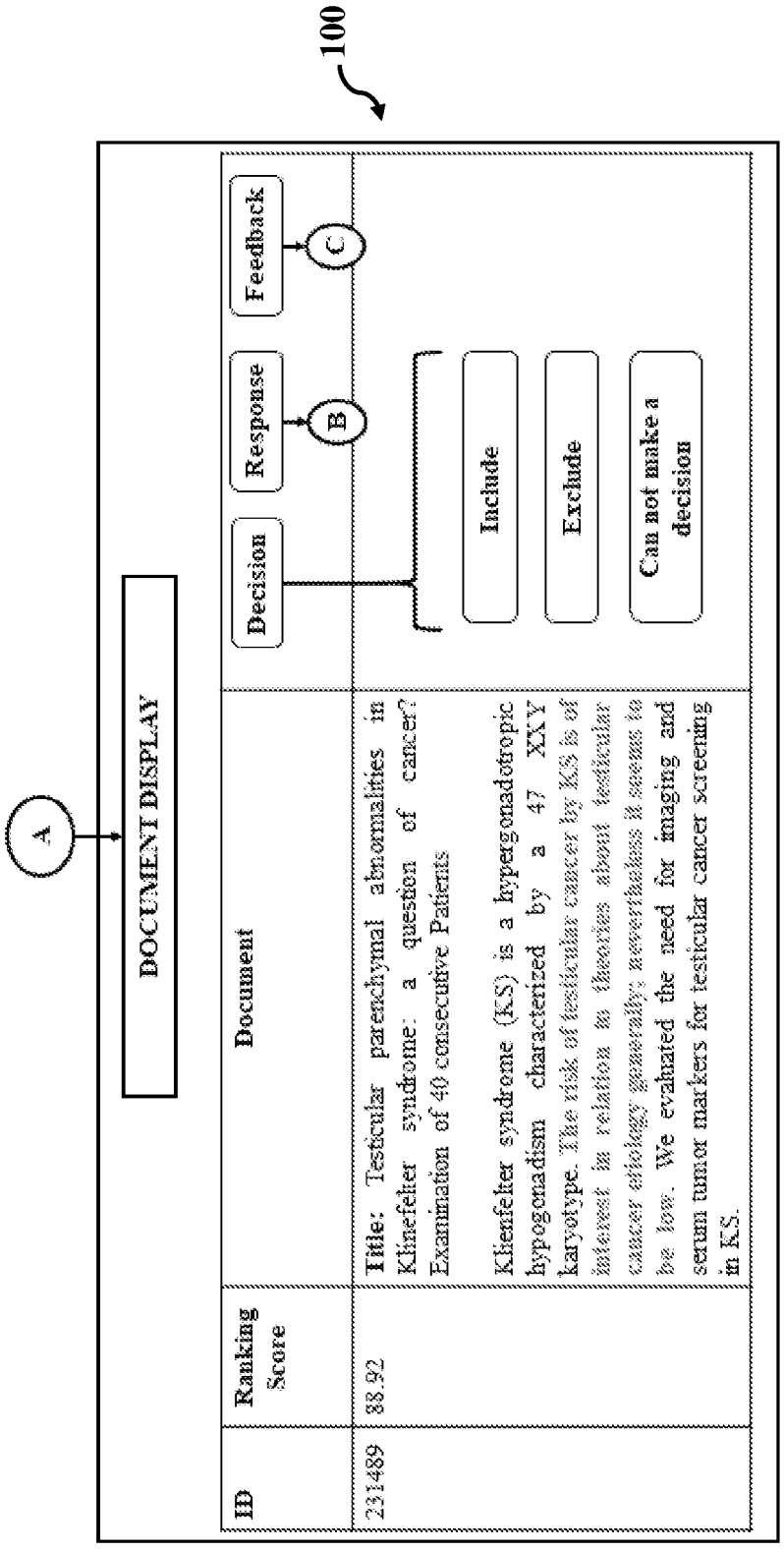
100
A
DOCUMENT DISPLAY
ID
Ranking Score
Document
Decision
Response
Feedback
B
C
231489
88.92
Title: Testicular parenchymal abnormalities in Klinefelter syndrome: a question of cancer? Examination of 40 consecutive Patients
Klienfelter syndrome (KS) is a hypergonadotropic hypogonadism characterized by a 47 XXY karyotype. The risk of testicular cancer by KS is of interest in relation to theories about testicular cancer etiology generally; nevertheless it seems to be low. We evaluated the need for imaging and serum tumor markers for testicular cancer screening in KS.
Include
Exclude
Can not make a decision

Figure 1B

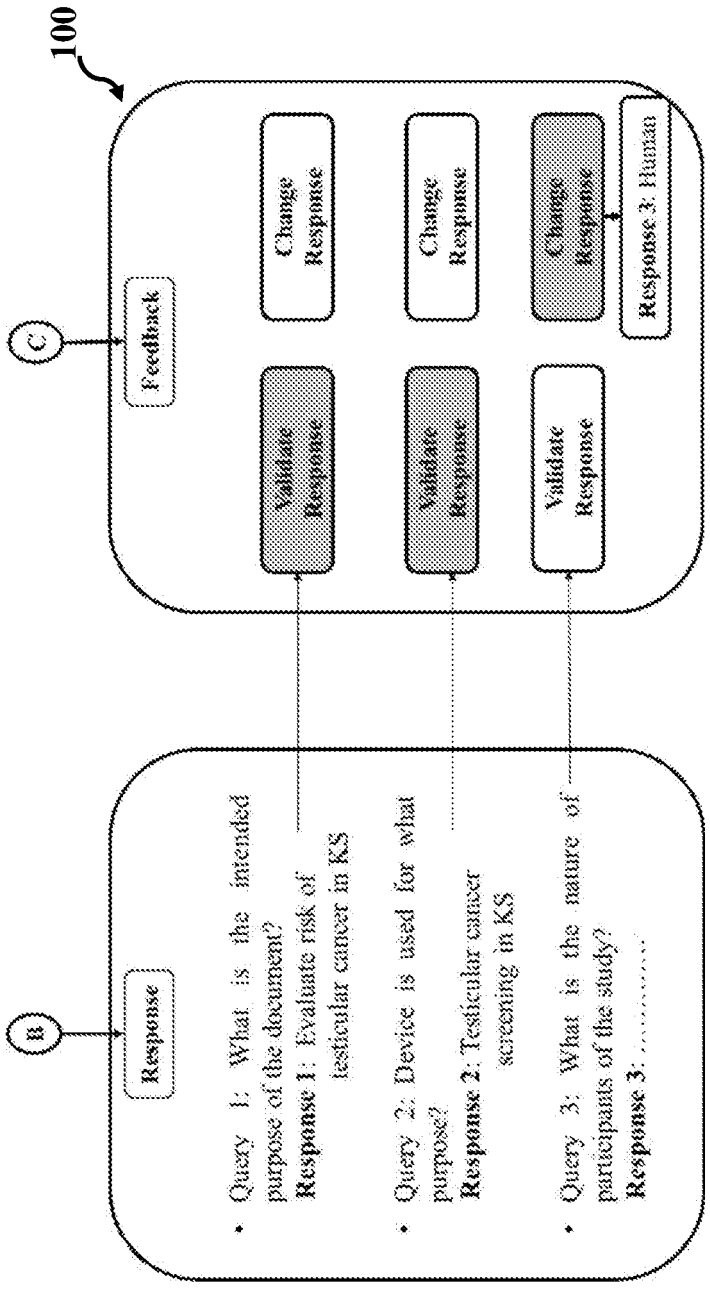
100
B
Response
• Query 1: What is the intended purpose of the document?
Response 1: Evaluate risk of testicular cancer in KS
• Query 2: Device is used for what purpose?
Response 2: Testicular cancer screening in KS
• Query 3: What is the nature of participants of the study?
Response 3: .....
C
Feedback
Validate Response
Validate Response
Validate Response
Change Response
Change Response
Change Response
Response 3: Human

Figure 1C

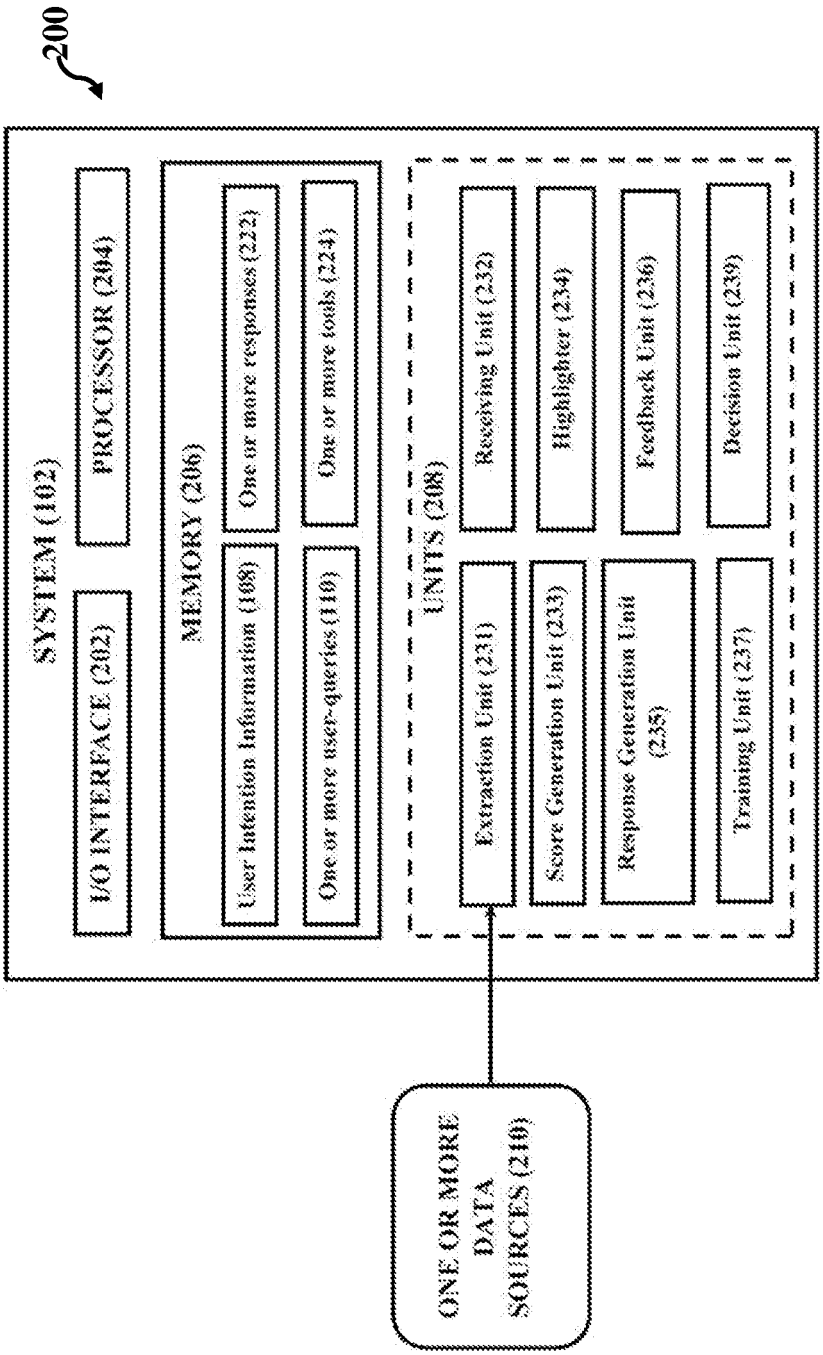
200
SYSTEM (102)
I/O INTERFACE (202)
PROCESSOR (204)
MEMORY (206)
User Intention Information (108)
One or more user-queries (110)
One or more responses (222)
One or more tools (224)
UNITS (208)
Extraction Unit (231)
Score Generation Unit (233)
Response Generation Unit (235)
Training Unit (237)
Receiving Unit (232)
Highlighter (234)
Feedback Unit (236)
Decision Unit (239)
ONE OR MORE DATA SOURCES (210)

Figure 2

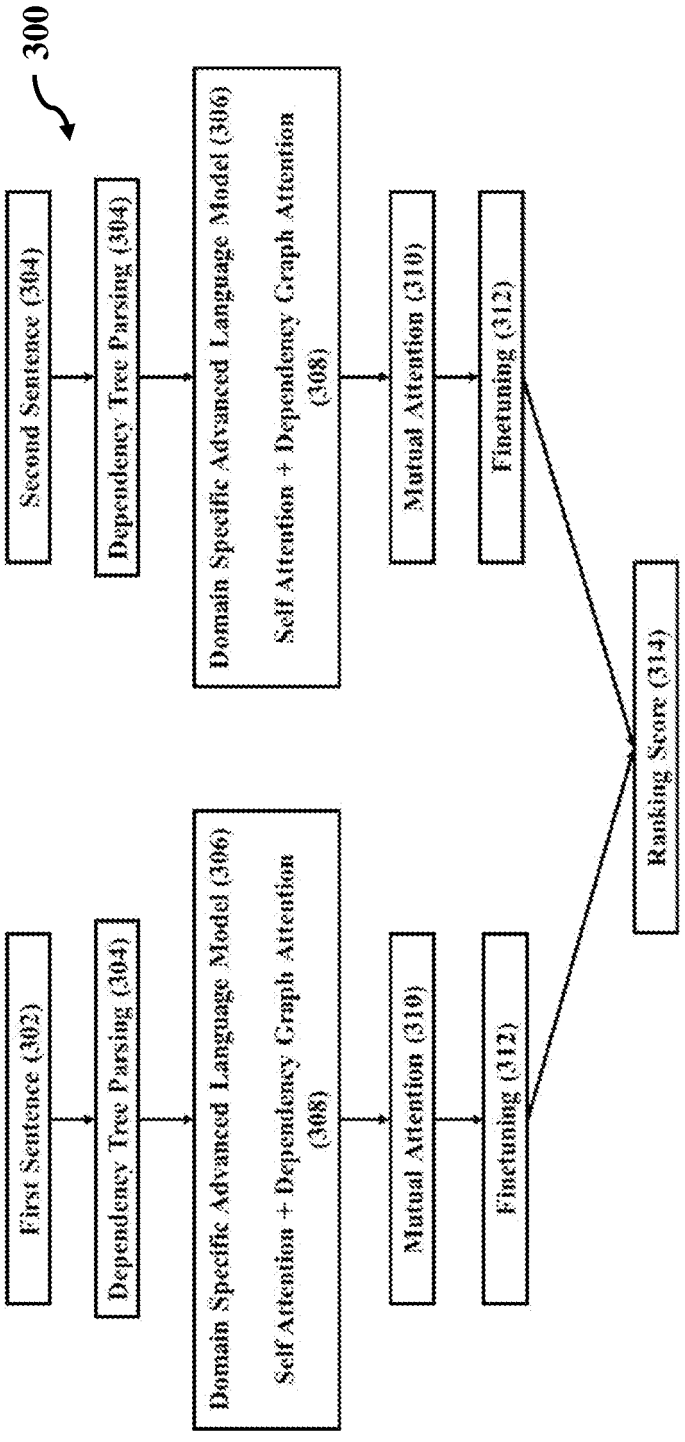
300
First Sentence (302)
Second Sentence (304)
Dependency Tree Parsing (304)
Dependency Tree Parsing (304)
Domain Specific Advanced Language Model (306)
Self Attention + Dependency Graph Attention (308)
Domain Specific Advanced Language Model (306)
Self Attention + Dependency Graph Attention (308)
Mutual Attention (310)
Mutual Attention (310)
Finetuning (312)
Finetuning (312)
Ranking Score (314)

Figure 3

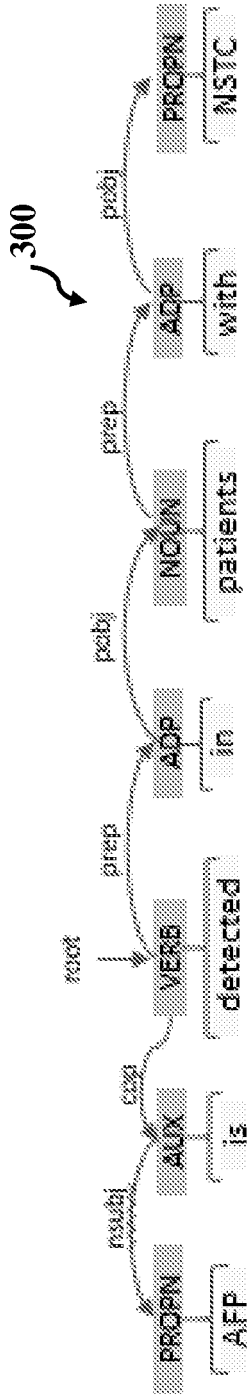

Figure 3

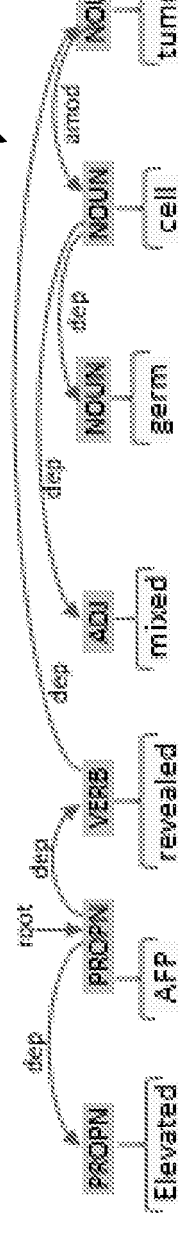

Figure 3B

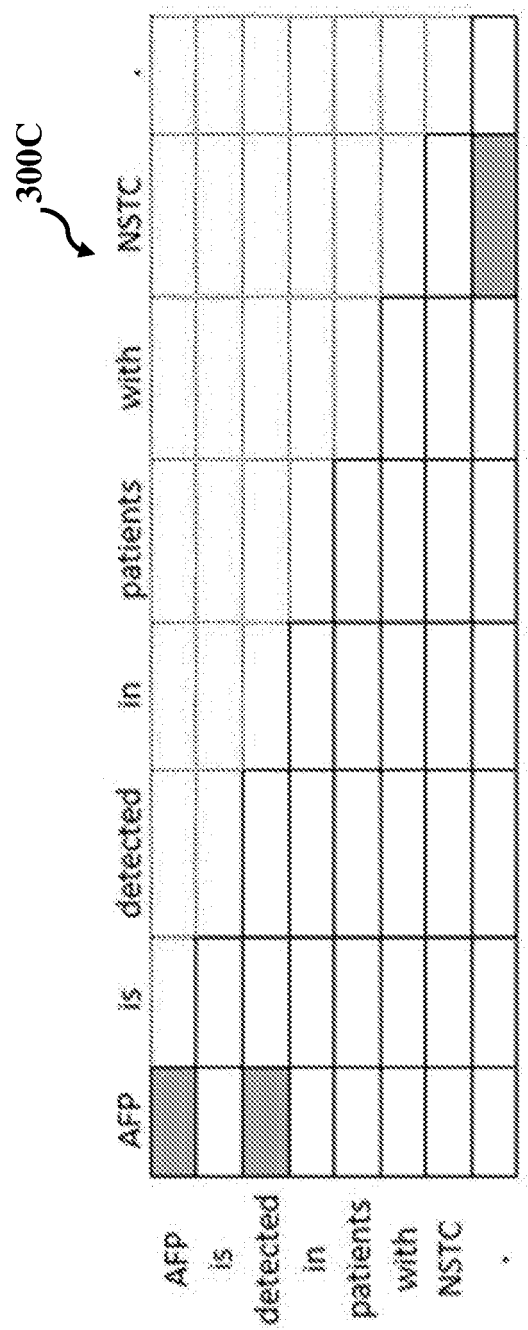
300C
AFP
is
detected
in
patients
with
NSTC
.
AFP
is
detected
in
patients
with
NSTC
.

Figure 3C

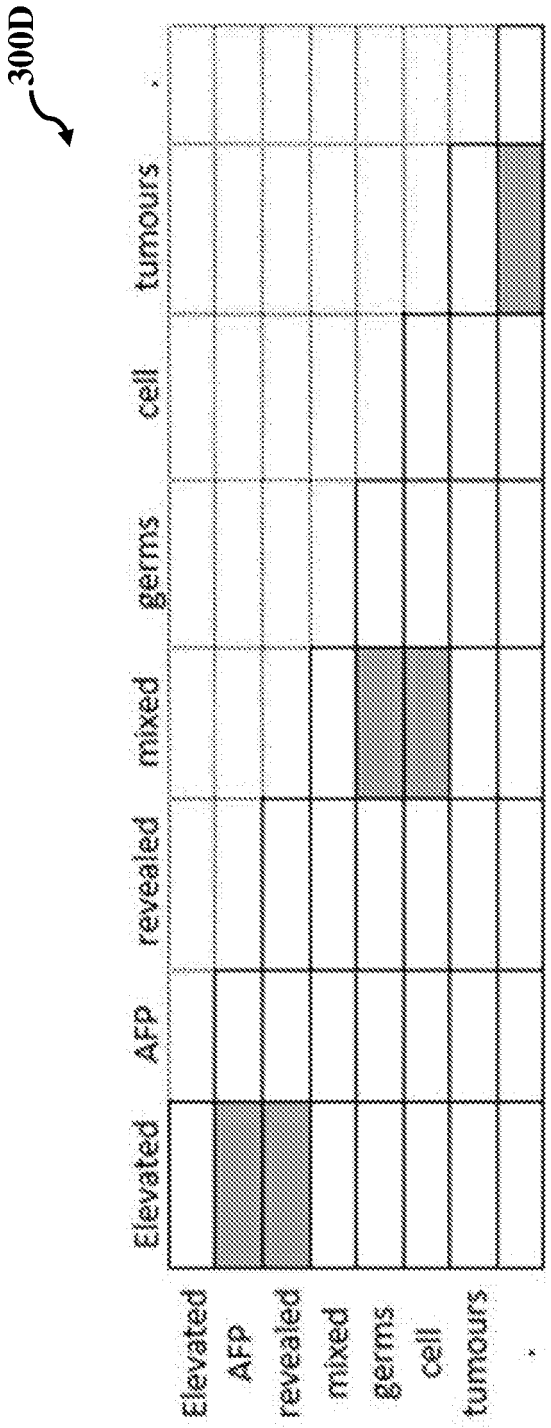
300D
Elevated
AFP
revealed
mixed
germs
cell
tumours
.
Elevated
AFP
revealed
mixed
germs
cell
tumours
.

Figure 3D

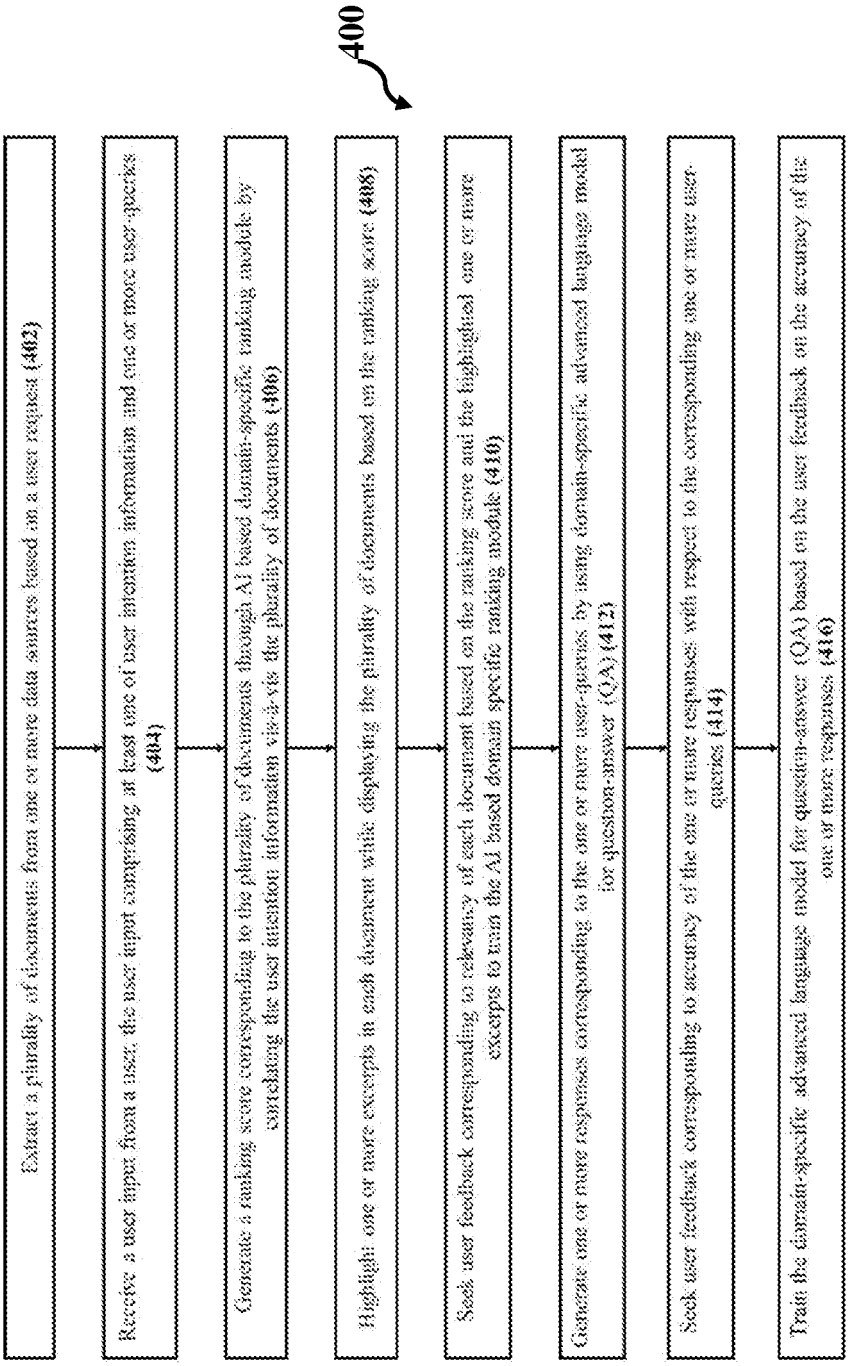
400
Extract a plurality of documents from one or more data sources based on a user request (402)
Receive a user input from a user, the user input comprising at least one of user intention information and one or more user-queries (404)
Generate a ranking score corresponding to the plurality of documents through AI based domain-specific ranking module by correlating the user intention information vis-à-vis the plurality of documents (406)
Highlight one or more excerpts in each document while displaying the plurality of documents based on the ranking score (408)
Seek user feedback corresponding to relevancy of each document based on the ranking score and the highlighted one or more excerpts to train the AI based domain specific ranking module (410)
Generate one or more responses corresponding to the one or more user-queries by using domain-specific advanced language model for question-answer (QA) (412)
Seek user feedback corresponding to accuracy of the one or more responses with respect to the corresponding one or more user-queries (414)
Train the domain-specific advanced language model for question-answer (QA) based on the user feedback on the accuracy of the one or more responses (416)

Figure 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2022/052469

A. CLASSIFICATION OF SUBJECT MATTER
G06F16/783,G06F16/93,G06F7/00 Version=2022.01

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)

Databases - PatSeer, IPO Internal Database
Keywords - relevancy document, highlight excerpts, rank, score

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KR20110027729A (BELENZON SHARON, ET AL.) 16 March 2011 (16-03-2011) {Whole Document specially Figures 3-6,11 with description}	1-10
Y	US9449080B1 (ZHANG GUANGSHENG) 20 September 2016 (20-09-2016) {Whole Document specially Columns 1-2}	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
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
30-06-2022	30-06-2022

Name and mailing address of the ISA/	Authorized officer
Indian Patent Office Plot No.32, Sector 14,Dwarka,New Delhi-110075 Facsimile No.	Saket Kumar Gupta Telephone No. +91-1125300200

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/IB2022/052469

Citation	Pub.Date	Family	Pub.Date
KR 20110027729 A	16-03-2011	CA 2727963 A1	30-12-2009
		EP 2321772 A2	18-05-2011
		JP 2011525673 A	22-09-2011
		US 2011093449 A1	21-04-2011
		WO 2009156987 A2	30-12-2009